Explanations
Week 2

CS329T

Stanford Spring 2021

Part II: Explanations
Goals

- Conceptual understanding of explanations
- Critically reason about tradeoffs for explanation methods
- View explanations as more than transparency, but rather a building block towards achieving trustworthy models

From the videos…

	Locally faithful only	**Local/global consistency**
Model-agnostic	LIME	Shapley Value (QII, SHAP)
Gradient-based (NNs)	Saliency maps	Aumann-Shapley values (IG, Influence-directed explanations)

How this relates to the learning objectives of the class:

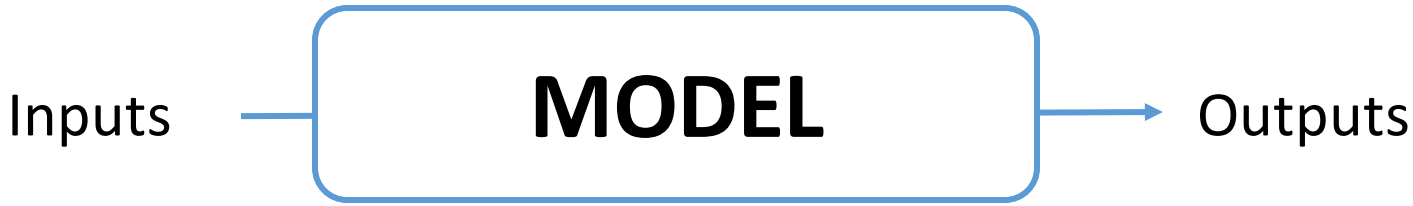
- Code LIME from scratch in HW2
- Reason about LIME, QII, SHAP in HW2
- Code saliency maps, IG from scratch in this week's lab
- Reason about gradient-based attribution strategies in HW2

Today at a glance

Taxonomy of explanations

- How to organize and evaluate explanations
- LIME, SHAP, QII, Integrated Gradients, Influence-directed Explanations
- How are these methods related

A Taxonomy of Explanations



A Taxonomy of Explanations

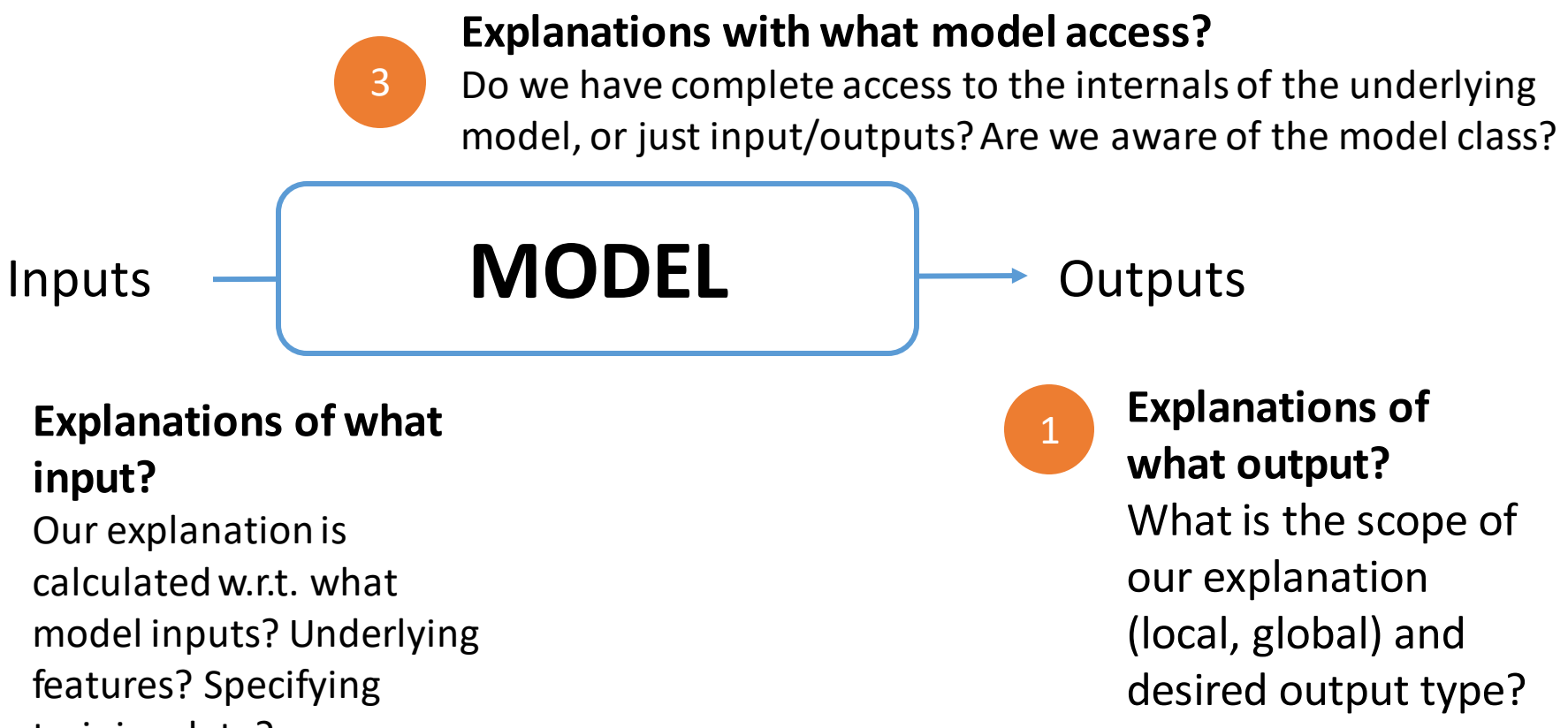
1 **Explanations of what output?**
What is the scope of our explanation (local, global) and desired output type?

A Taxonomy of Explanations

2 **Explanations of what input?**
Our explanation is calculated w.r.t. what model inputs? Underlying features? Specifying training data?

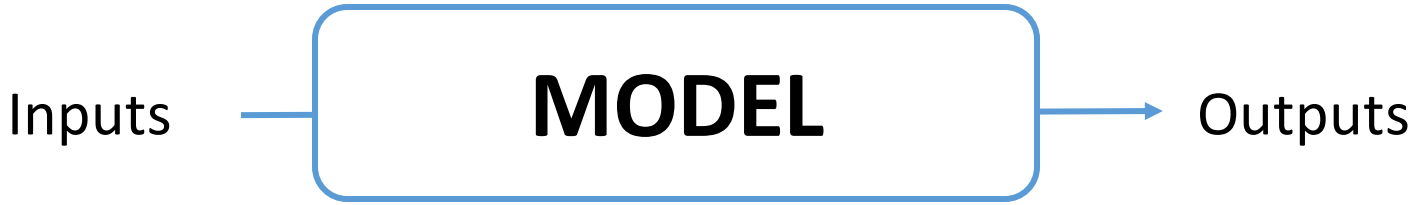
1 **Explanations of what output?**
What is the scope of our explanation (local, global) and desired output type?

A Taxonomy of Explanations



A Taxonomy of Explanations

3 **Explanations with what model access?**
Do we have complete access to the internals of the underlying model, or just input/outputs? Are we aware of the model class?



2 **Explanations of what input?**
Our explanation is calculated w.r.t. what model inputs? Underlying features? Specifying training data?

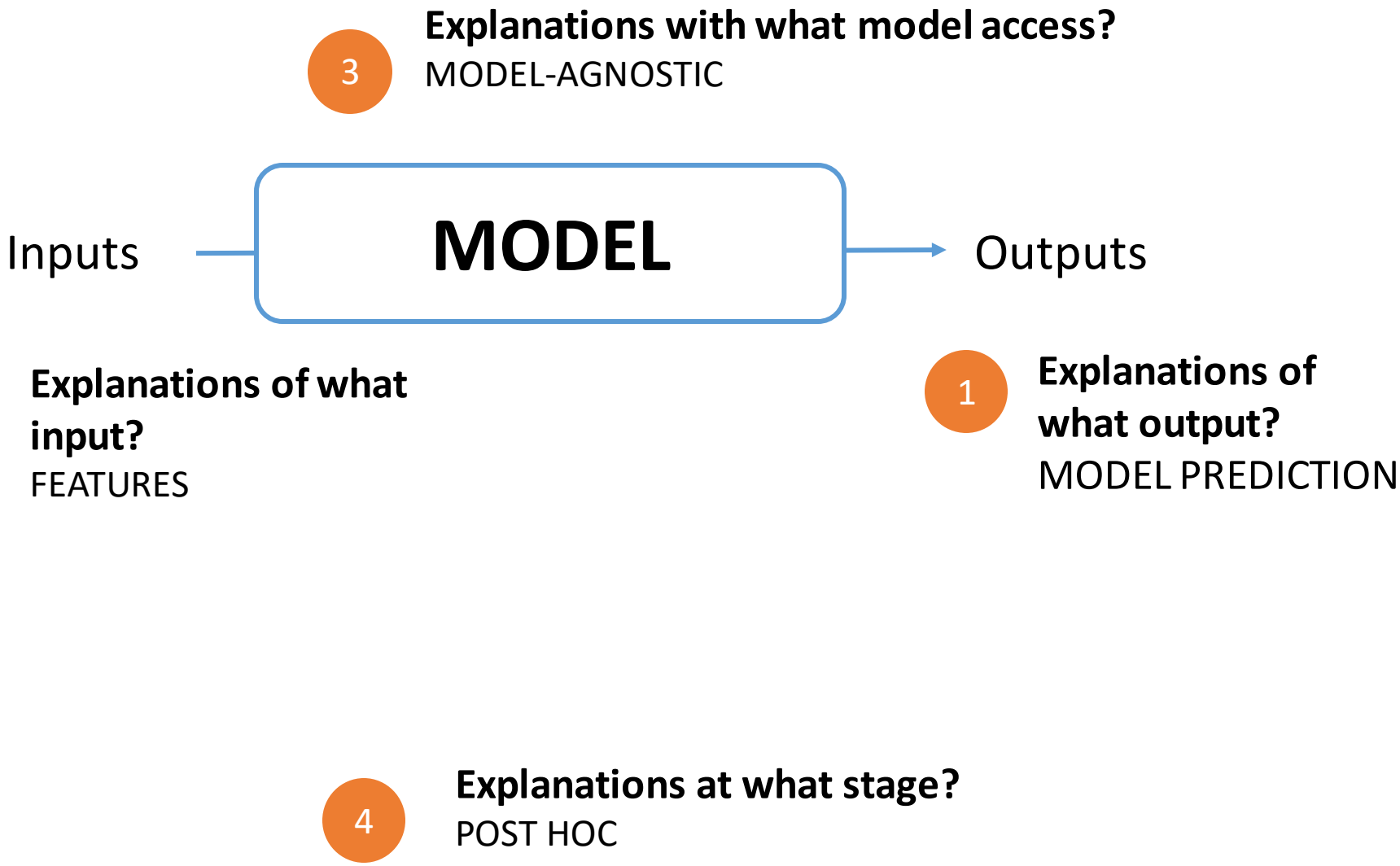
1 **Explanations of what output?**
What is the scope of our explanation (local, global) and desired output type?

4 **Explanations at what stage?**
"Inherently explainable" vs. "post-hoc" explainability.

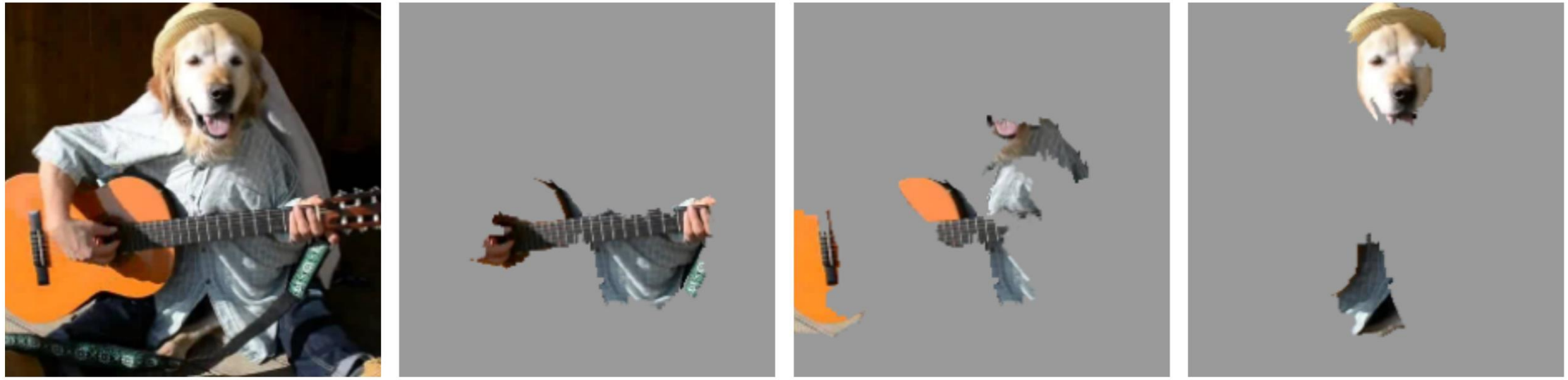
What do we want from an explanation method?

- Local agreement (on predictions)

A Taxonomy of Explanations: LIME



A Taxonomy of Explanations: LIME



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

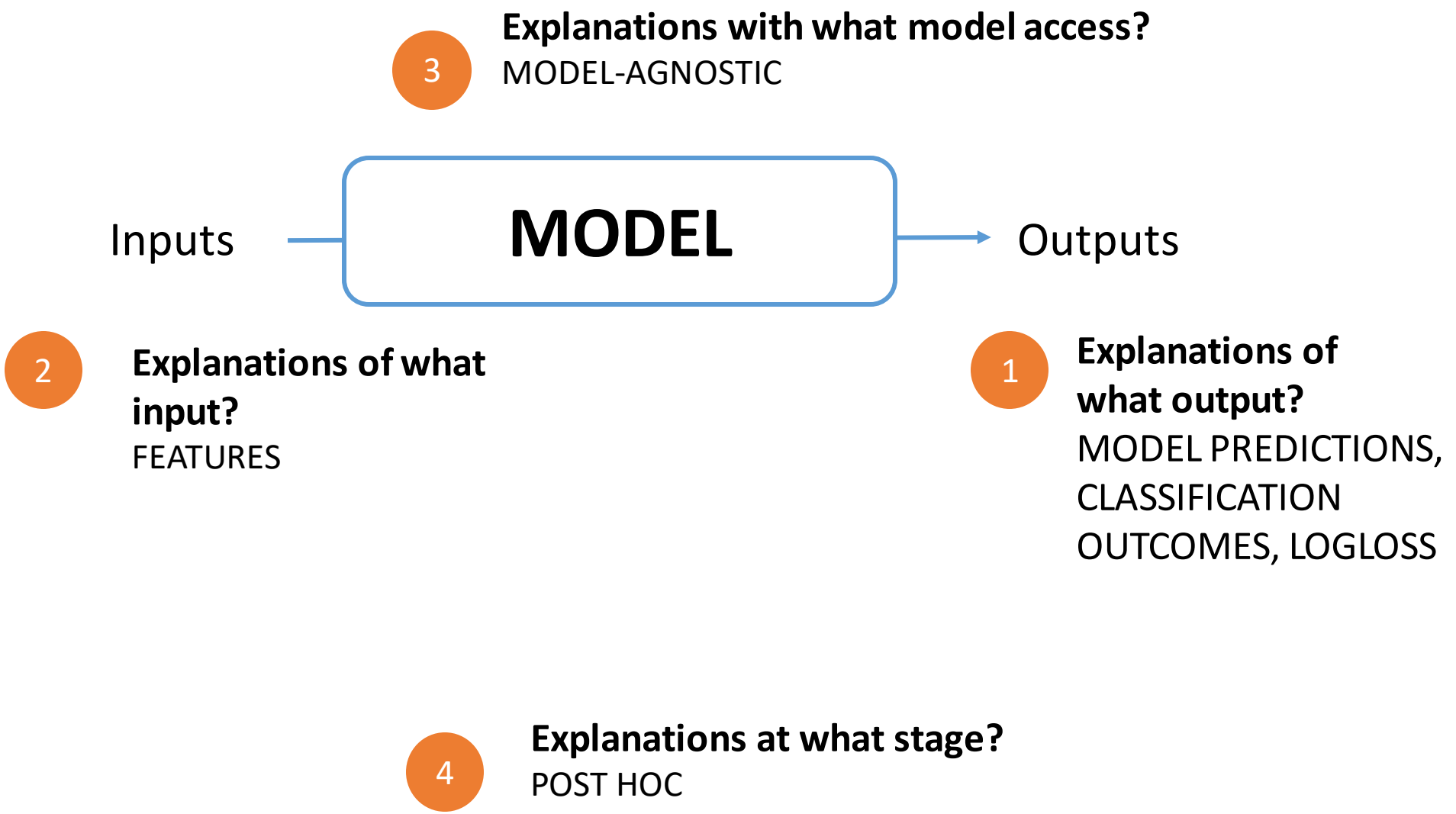
What do we want from an explanation method?

- Local agreement (on predictions)

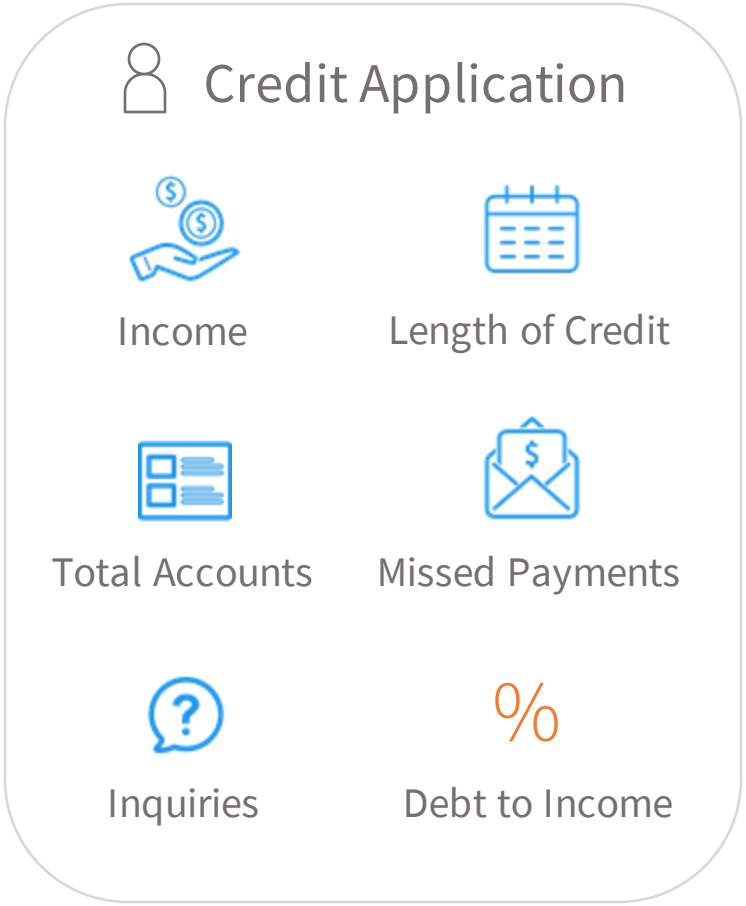
But also some other things:

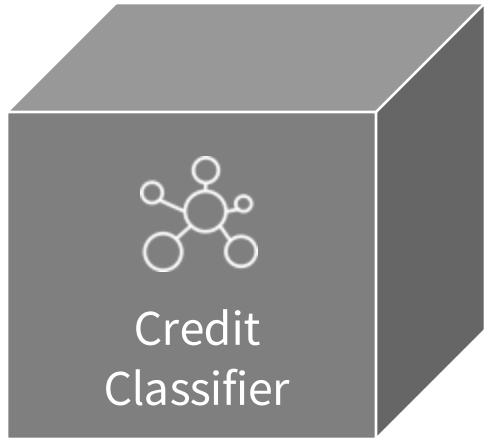
- Not too local (consistency between local and global)
- Causal
- Consistency (marginality principle)
- Completeness, dummy, etc...

A Taxonomy of Explanations: QII/SHAP



QII/SHAP


Credit Application
Income
Length of Credit
Total Accounts
Missed Payments
Inquiries
%
Debt to Income


Credit
Classifier


DENIED

QII/SHAP - Feature Importance

What we've highlighted

	Locally faithful only	**Local/global consistency**
Model-agnostic	LIME	Shapley Value (QII, SHAP)
Gradient-based (NNs)	?	?

What do we want from an explanation method?

- Local agreement (on predictions)

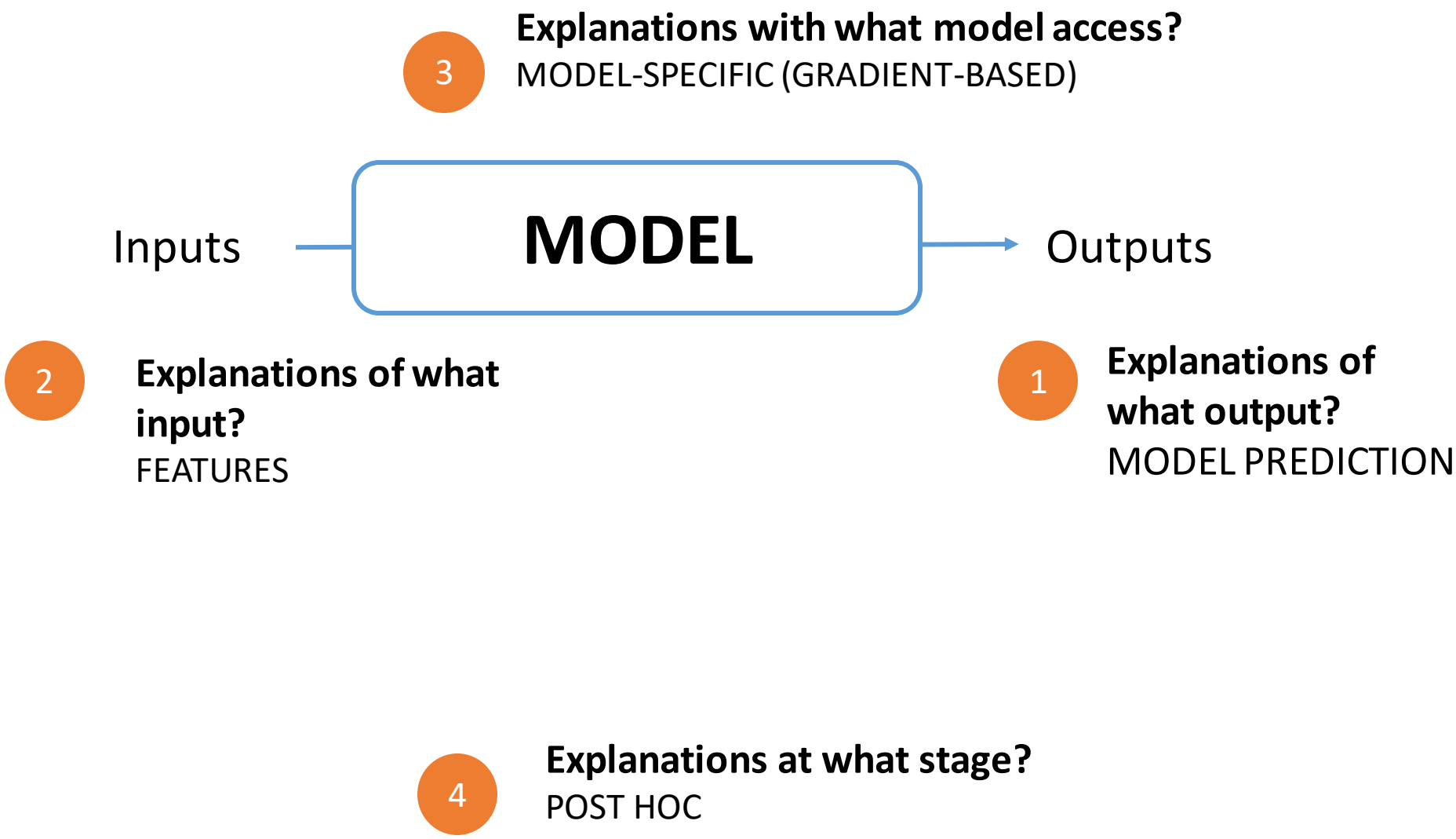
But also some other things:

- Too local (lacks consistency between global and global)
- Not causal
- Consistency (marginality principle)
- Completeness, dummy, etc...

What if the data isn't discrete/tabular?

- Works for continuous features / spaces
- Doesn't break for large models (e.g. Neural Nets)

A Taxonomy of Explanations: Integrated Gradients



Integrated Gradients

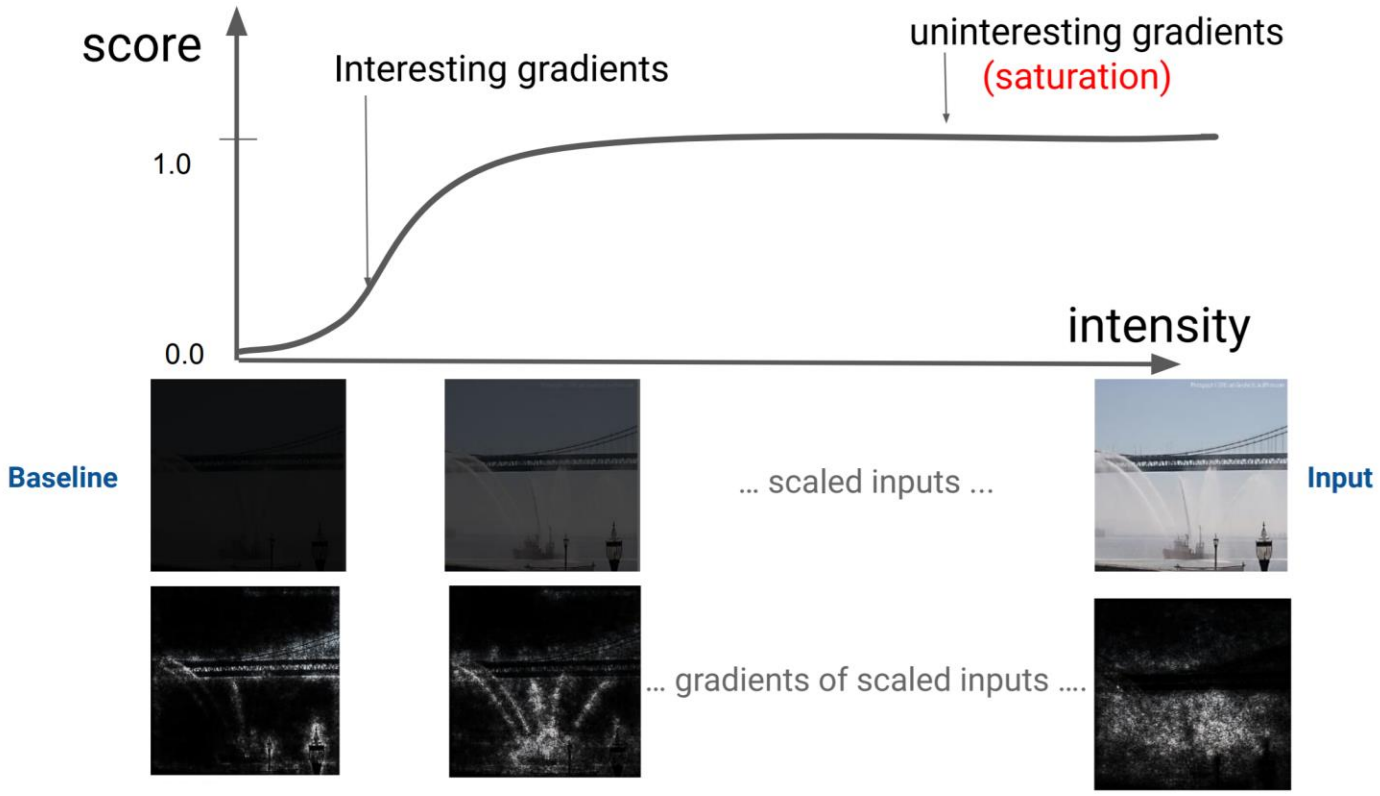
Original image

Integrated Gradients
(for label “clog”)

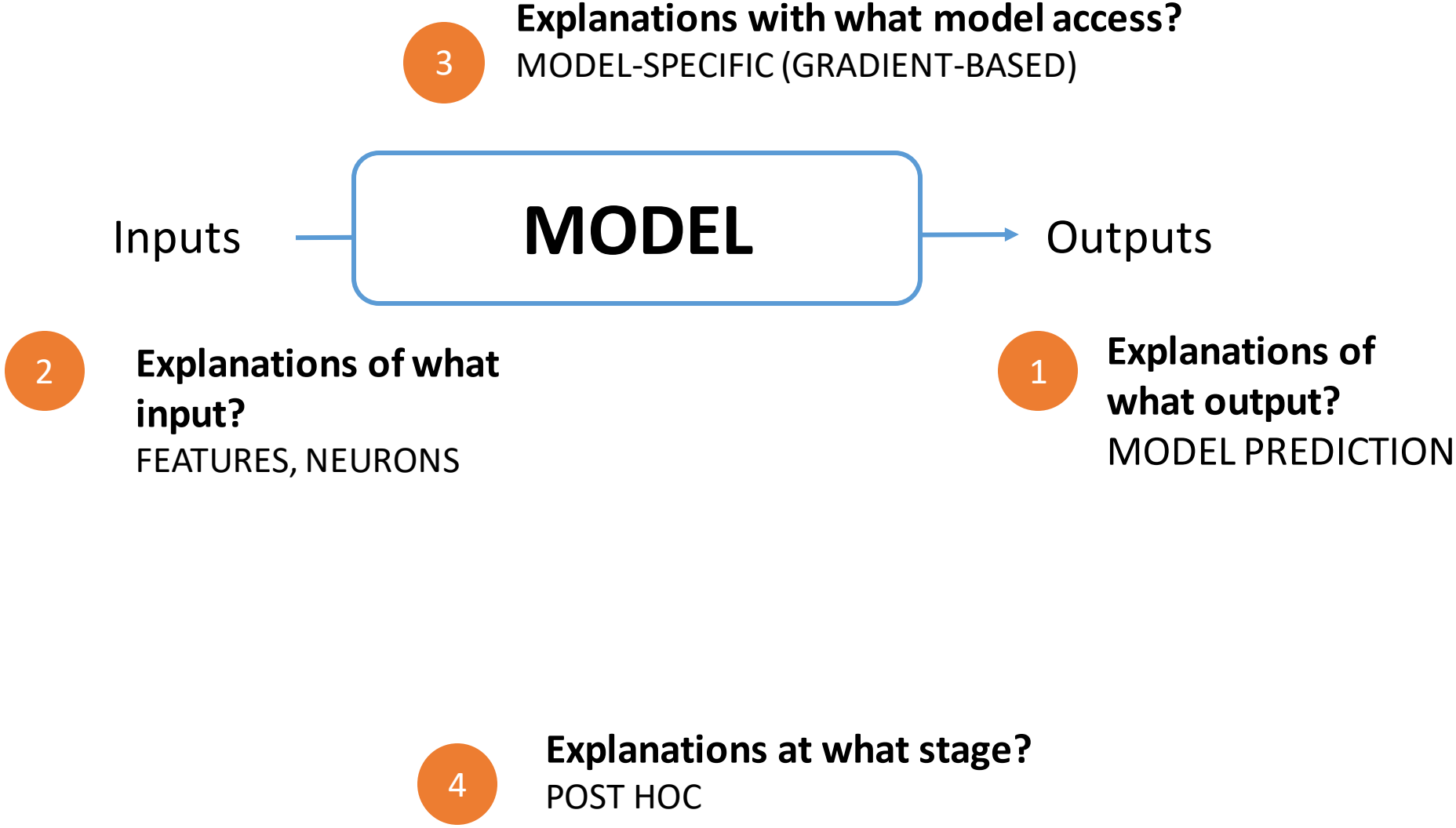
“Clog”

Shapley value properties in the NN world…

- Need to define a pixel's "average contribution" in the context of a *baseline* (e.g. all-black image)
- Integrate gradients along straight-line path from baseline to an input
- Connection to **Aumann-Shapley values**
  - extension of Shapley values for "infinite games" (e.g. a continuous feature space)



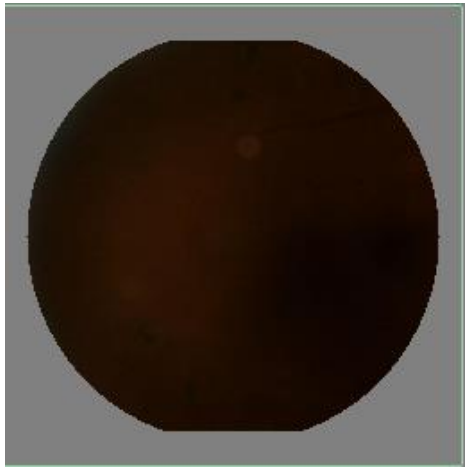
A Taxonomy of Explanations: Influence-Directed Explanations

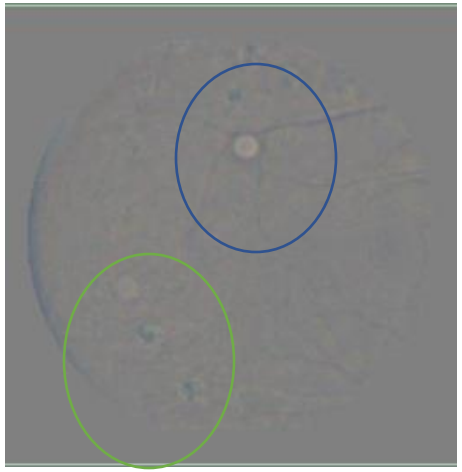


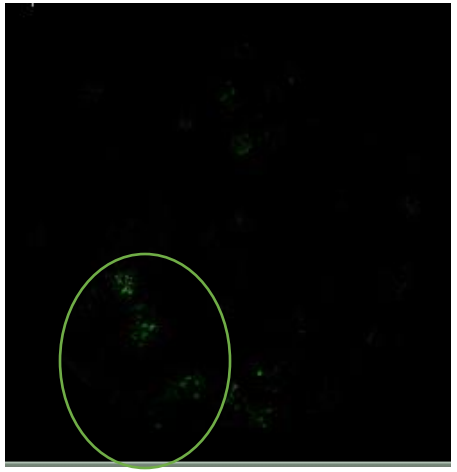
Why classified as diabetic retinopathy stage 5?

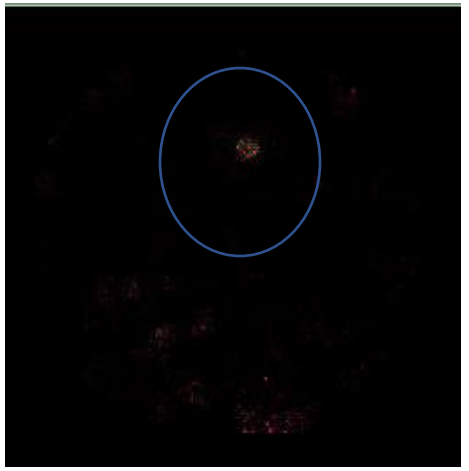
Inception network

Optic disk









Lesions

Why sports car instead of convertible?

VGG16 ImageNet model

Input image

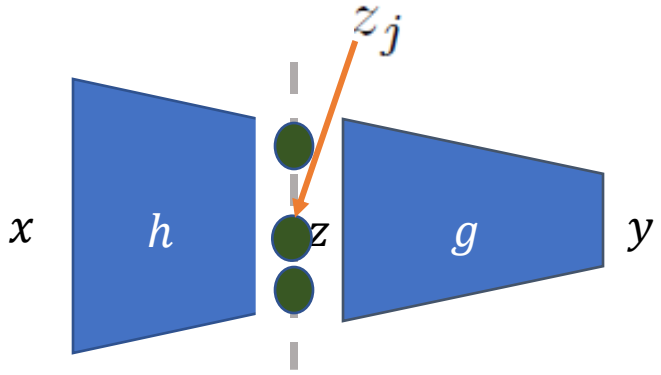
Influence-directed Explanation

Uncovers high-level concepts that generalize across input instances

Why Orlando Bloom?

Distributional influence

Influence = average gradient over distribution of interest



$$y = f(x) = g(h(x))$$

$$\chi_j^s(f, P) = \int_{\mathcal{X}} \left|\frac{\partial g}{\partial z_j}\right|_{h(\mathbf{x})} P(\mathbf{x}) d\mathbf{x}$$

Gradient

Weighted by probability of input x

For input x [note z = h(x)]

Theorem: Unique measure that satisfies a set of natural properties

Interpreting influential neurons

Depicts interpretation (visualization) of 3 most influential neurons

- Slice of VGG16 network: conv4_1
- Inputs drawn from distribution of interest: delta distribution
- Quantity of interest: class score for correct class

Interpreting influential neurons

Visualization method: Saliency maps [Simonyan et al. ICLR 2014]

- Compute gradient of neuron activation wrt input pixels
- Scale pixels of original image accordingly

What we've highlighted

	Locally faithful only	**Local/global consistency**
Model-agnostic	LIME	Shapley Value (QII, SHAP)
Gradient-based (NNs)	Saliency maps	Aumann-Shapley values (IG, Influence-directed explanations)

How this relates to the learning objectives of the class:

- Code LIME from scratch in HW2
- Reason about LIME, QII, SHAP in HW2
- Code saliency maps, IG from scratch in this week's lab
- Reason about gradient-based attribution strategies in HW2

What we've highlighted

	Locally faithful only	**Local/global consistency**
Model-agnostic	LIME	Shapley Value (QII, SHAP)
Gradient-based (NNs)	Saliency maps	Aumann-Shapley values (IG, Influence-directed explanations)

Learn more beyond this (if you're interested):

- Other NN explanation techniques (layerwise relevance propagation, guided backprop)
- Inherently interpretable explanations (generalized linear models, GAMs, decision trees)
- Lots of new explanation techniques & variants every day (browse NeurIPS, ICML, AAAI, etc.)

Questions for Discussion

- QII vs. SHAP?
- Baselines within Integrated Gradients
- IG vs. Influence-Directed Explanations
- What's special about Shapley values?
- Neurons vs. Features: the same?

Explanations: **a means, not an end.**

Explanations provide transparency into a model. But is that all?

How can we use these explanation techniques to improve the **overall quality of a model?** Think quality beyond just accuracy—fairness? Size or speed? Feature engineering? Others?

So how do we calculate these Shapley values?

- Clearly impractical to calculate them in closed form (exponential number of feature sets)
- Methods/frameworks to approximate (in this class): QII, SHAP

- Even once you fix the explanation method (e.g. QII or SHAP), lots of implementation choices that can dramatically affect the explanation result:
  - E.g. What are you explaining? Classification outcome, prediction, log-odds
  - You will see this more in HW2 when exploring NN attribution strategies

QII vs. SHAP: What's the difference?

- Sampling-based estimation (model-agnostic)
  - Interventional vs. Conditional approach
- Decomposing for tree-based models (model-specific)
  - Works for linear combinations of log-odds scores from individual trees
  - Additivity principle

What's with this "baseline" in IG?

- Reminder: we are using vision models as prototypical examples, but these NN explanation techniques could extend to most neural network types...
- What is a reasonable baseline in vision? Audio? Text? Tabular?
- As we saw in QII, comparison group is key—same applies here.
- If we take another path integral (not a straight line from baseline to input) does that break any of our axioms? Why?

IG vs. Influence-Directed Explanations: What's the difference?

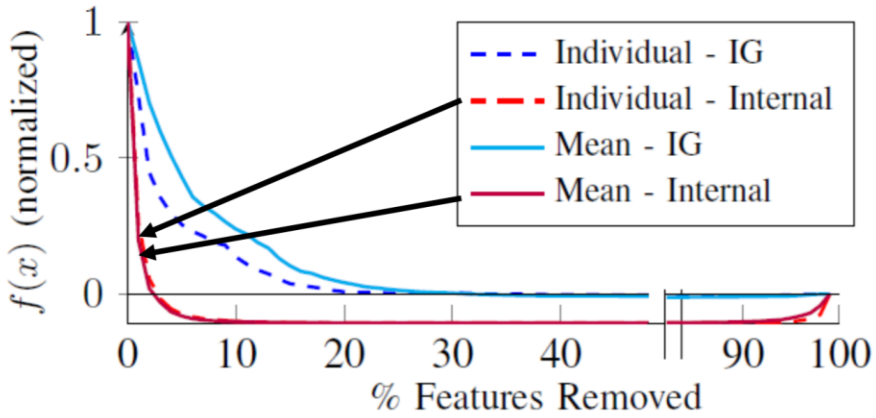
Think of Influence-Directed Explanations as taking the core of IG and generalizing it in a QII-like way: quantities of interest, and distribution of interests (or comparison groups)

Influence-Directed Explanations can be thought of as a **generalized framework** that extends IG by two directions:

- It abstracts the way the "explanation to what?" question—could be to the input features, to the input neurons of a layer
- It generalizes the notion of a "distribution of interest" and a "quantity of interest" that is trying to be explained
  - Specific DoI/QoI choices would collapse to IG and saliency maps.
  - Question: what DoI and QoI pair would correspond to integrated gradients?

Influence-Directed Explanations: Are neurons/features equivalent?

- Maybe from a mathematical perspective (just take a "cut" of the network at a particular layer—those neurons are "features" of a new NN that starts at that layer)
- Not so much in explanation behavior
  - Average most important neuron for a class is likely to also be important for each instance of the class
  - This doesn't apply to input features



Score for correct class drops rapidly as most influential neurons are turned off

Questions?

Appendix

What is so special about the Shapley value?

Consider a model with *p* features (players). The Shapley value is the *only* attribution method that satisfies the following desirable properties.

- Efficiency: $\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$
- Symmetry: Equal marginal contribution implies equal influence (redundant feature)
- Dummy: Zero marginal contribution implies zero influence (cloned feature)
- Monotonicity: Consistently higher marginal contribution yields higher influence (easy to compare scores)
- Additivity: For a setting with combined payouts (val and val+), Shapley values for a feature: $\phi_j + \phi_j^+$

What's with this "baseline" in IG?

And what is the analog of a "path" when we compare IG to SHAP/QII?

Note: computationally infeasible to approximately Shapley values well for NNs when the number of features is **huge.**

Remark 5. *If we allow averaging over the attributions from multiple paths, then are other methods that satisfy all the axioms in Theorem 1. In particular, there is the method by Shapley-Shubik (Shapley & Shubik, 1971) from the cost sharing literature, and used by (Lundberg & Lee, 2016; Datta et al., 2016) to compute feature attributions (though they were not studying deep networks). In this method, the attribution is the average of those from $n!$ extremal paths; here n is the number of features. Here each such path considers an ordering of the input features, and sequentially changes the input feature from its value at the baseline to its value at the input. This method yields attributions that are different from integrated gradients. If the function of interest is $min(x_1, x_2)$, the baseline is $x_1 = x_2 = 0$, and the input is $x_1 = 1$, $x_2 = 3$, then integrated gradients attributes the change in the function value entirely to the critical variable x_1, whereas Shapley-Shubik assigns attributions of $1/2$ each; it seems somewhat subjective to prefer one result over the other.*