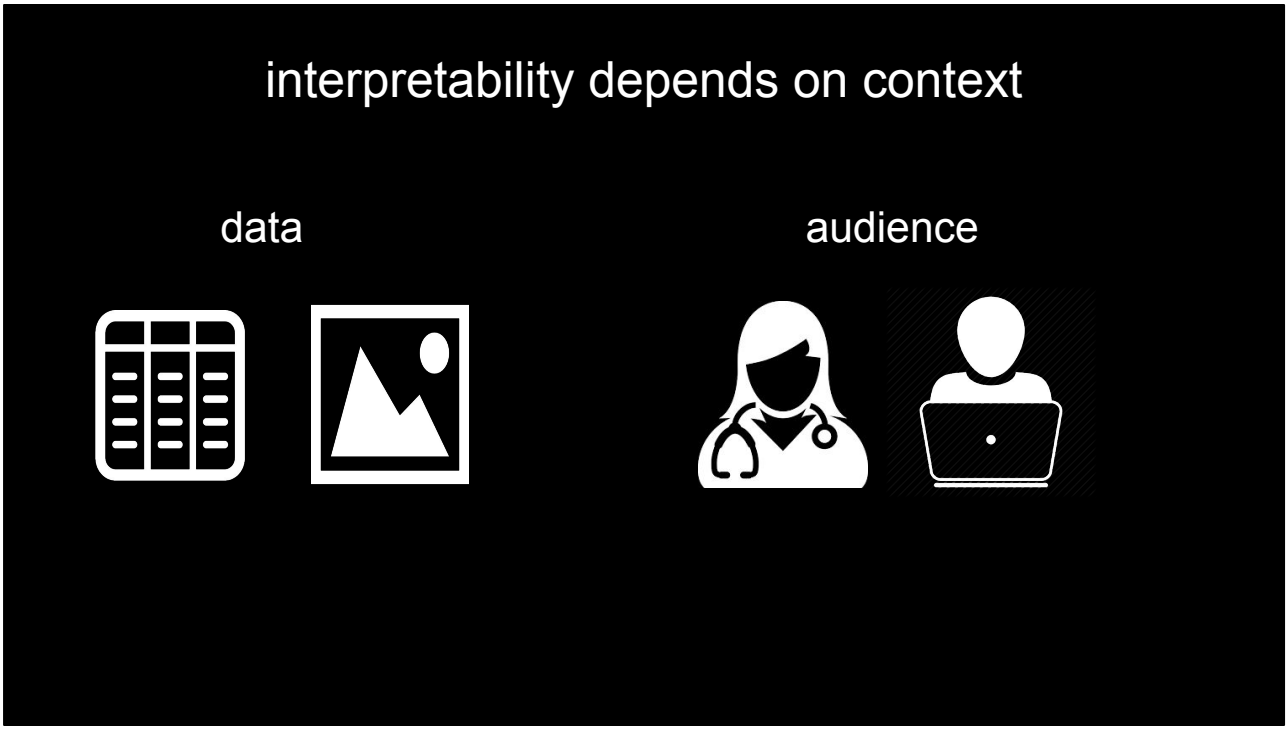
Github repo with accompanying notebooks

Some resources

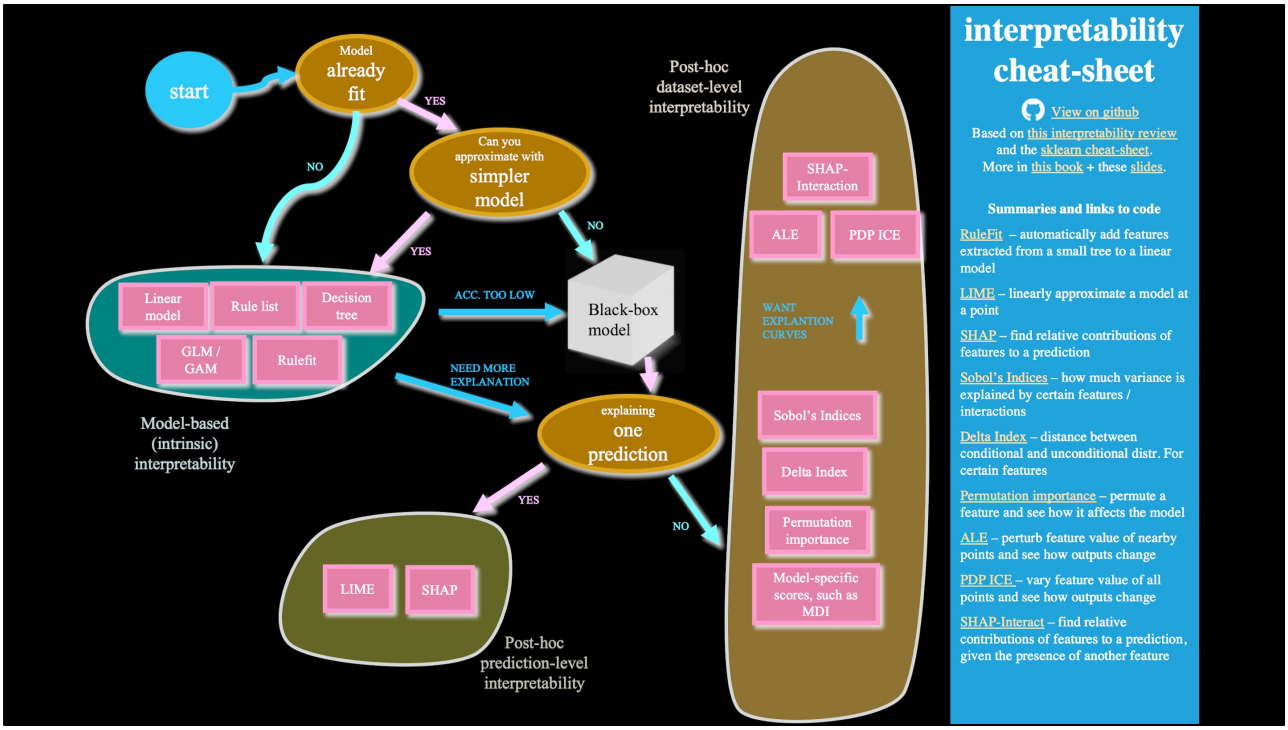
- book on interpretable machine learning
- high-level review on interpretable machine learning
- review on black-blox explanation methods
- review on variable importance



Interpretability is generally difficult to define. Murdoch et al 2019 defines interpretable machine as follows: “the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data. Here, we view knowledge as being relevant if it provides insight for a particular audience into a chosen domain problem”

overview

1. how can we build a simple model?
2. which features are globally important?
3. which features are locally important?



This is a little ugly, but the original version is nicer and has links to code in the right-hand panel.

How can we build a **simple model**?

This has become an increasingly important area of research, as having a simpler (or transparent model) is preferable to doing posthoc interpretability on a black-box model. One recent paper drives this point home very clearly.

example 1: sparse integer linear model

1.	Any cEEG Pattern with Frequency **2 Hz**	1 point	⋯
2.	**E**pileptiform Discharges	1 point	+ ⋯
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	+ ⋯
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	+ ⋯
5.	Prior **S**eizure	1 point	+ ⋯
6.	**B**rief Rhythmic Discharges	**2** points	+ ⋯
		SCORE	= ⋯

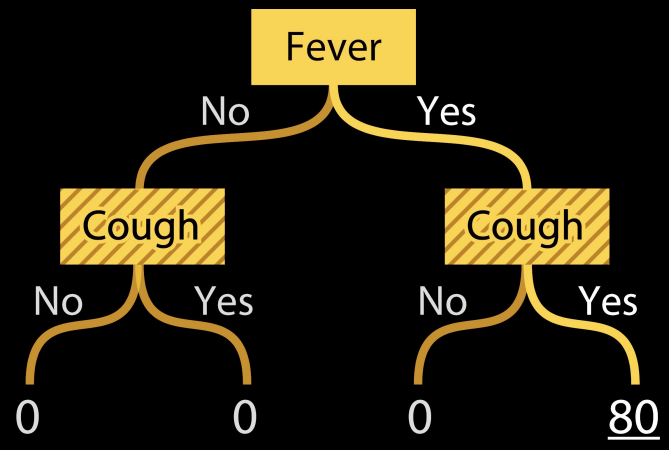
SCORE	0	1	2	3	4	5	6+
RISK	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

struck et al. 2017

These types of models are often made by doctors themselves but using machine learning we can identify such a model with high predictive accuracy. Note that sparsity and enforcing coefficients to be integers makes the fitting pro

example 2: optimal classification tree

$$\min \quad \frac{1}{\hat{L}} \sum_{t \in \mathcal{T}_L} L_t + \alpha \sum_{t \in \mathcal{T}_B} d_t$$



bertsimas & dunn 2017

Very short decision trees can be quite interpretable, but often don't yield very high predictive accuracy. Optimal classification trees are a type of model which attempts to explicitly learn short trees with high predictive accuracy. The key insight is to do a joint optimization over all splits of the tree, rather than the greedy optimization used by standard CART decision trees.

example 3: bayesian rule list

if hemiplegia **and** age $>$ 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age $>$ 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age $\leq$ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

letham et al. 2015

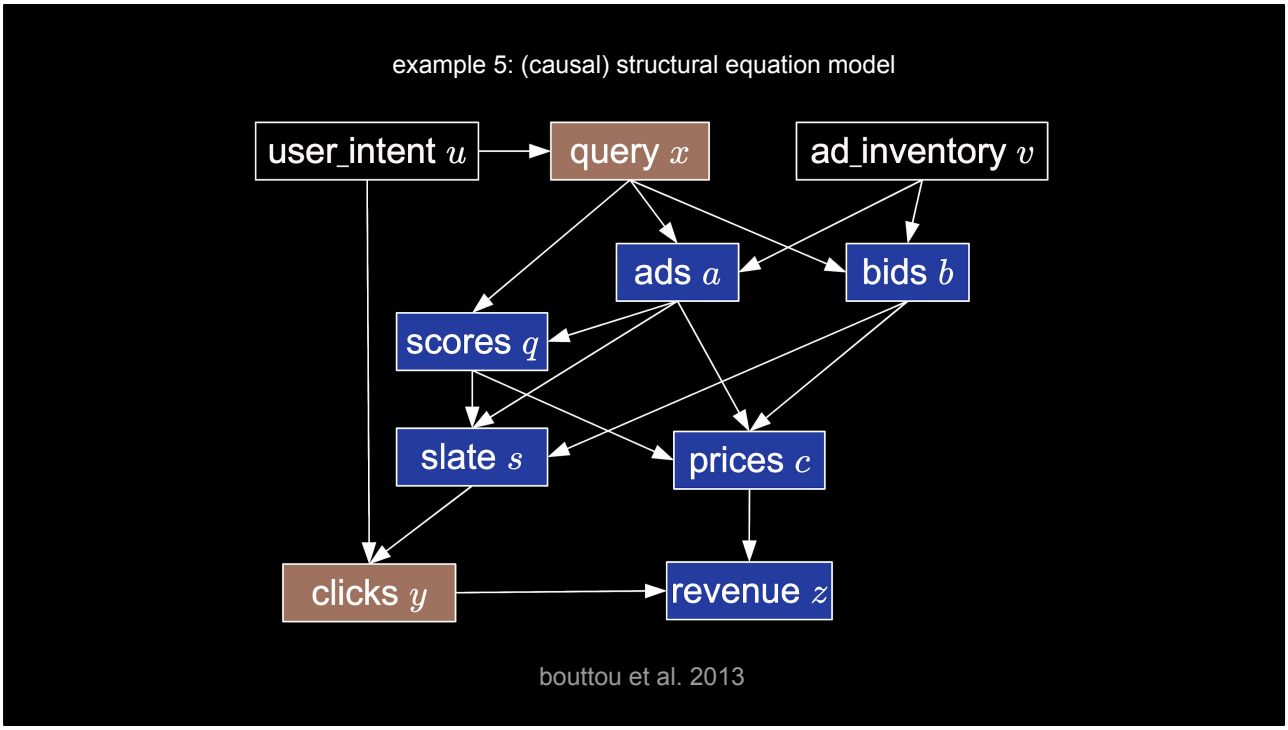
Bayesian rule lists attempt to learn a model of this form, again doing a joint optimization over the rules to try and make the list as short as possible while maintaining high predictive accuracy.

example 4: rulefit

Description	Weight
days_since_2011 > 111 & weathersit in (“GOOD”, “MISTY”)	793
37.25 <= hum <= 90	-20
temp > 13 & days_since_2011 > 554	676
4 <= windspeed <= 24	-41
days_since_2011 > 428 & temp > 5	366

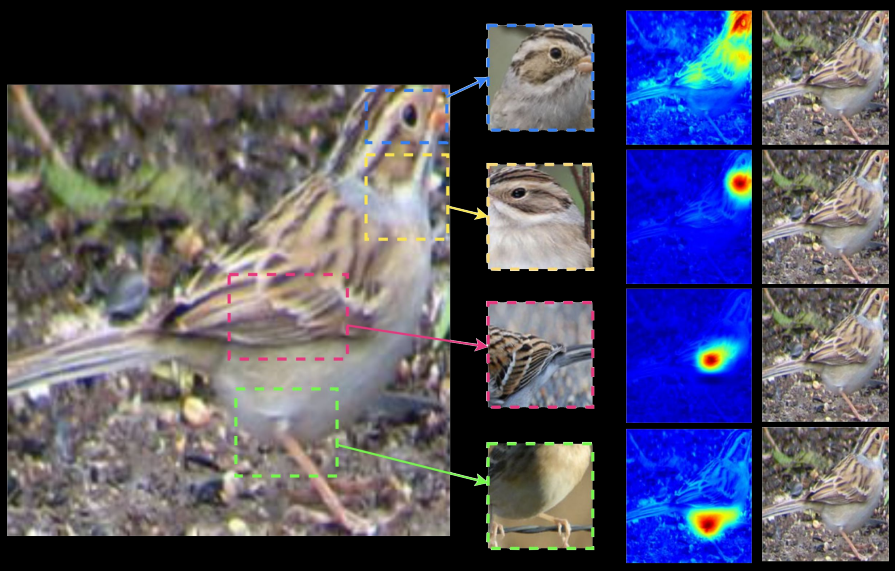
molnar et al. 2019

Rulefit attempts to learn a sparse linear model on simple, interpretable features which are automatically selected from the data. These features are extracted by taking paths out of a short (here depth 2) decision tree.



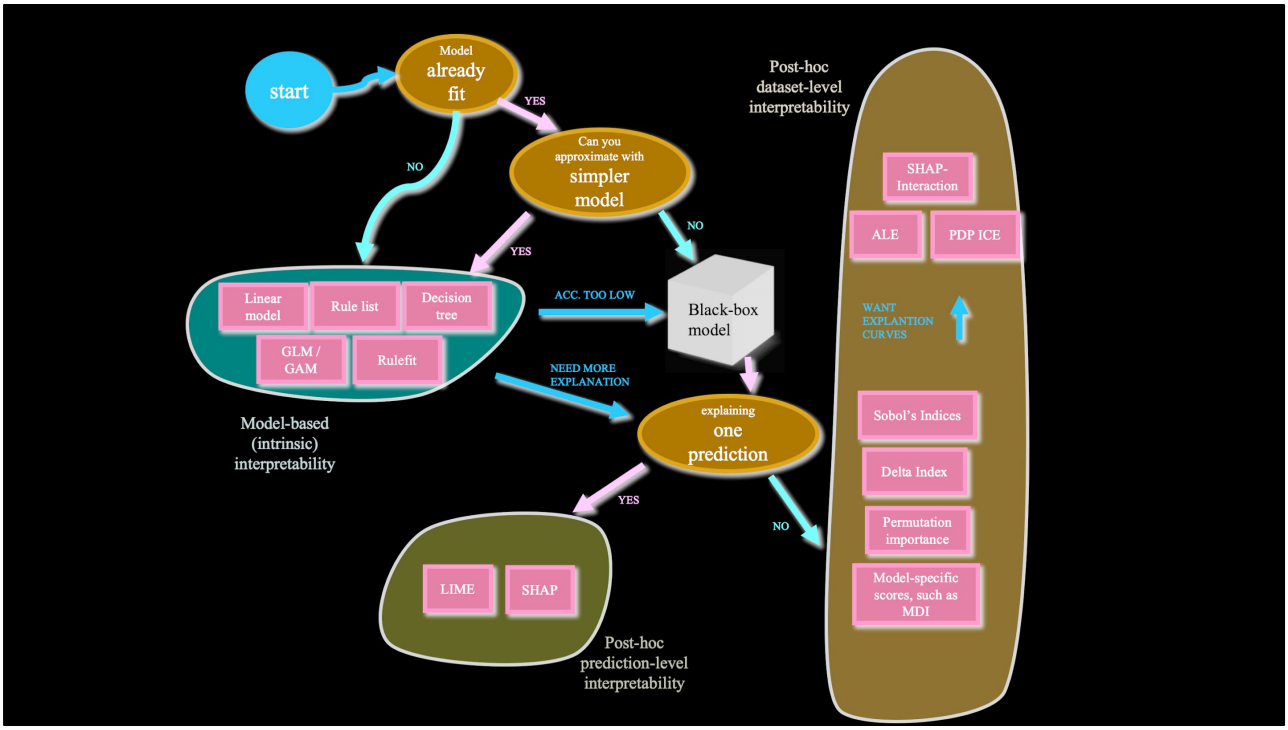
Structural equation models require one to explicitly figure out what causal relationships generate the observed data. Once fit, they allow us to ask causal questions about interventions and counterfactuals.

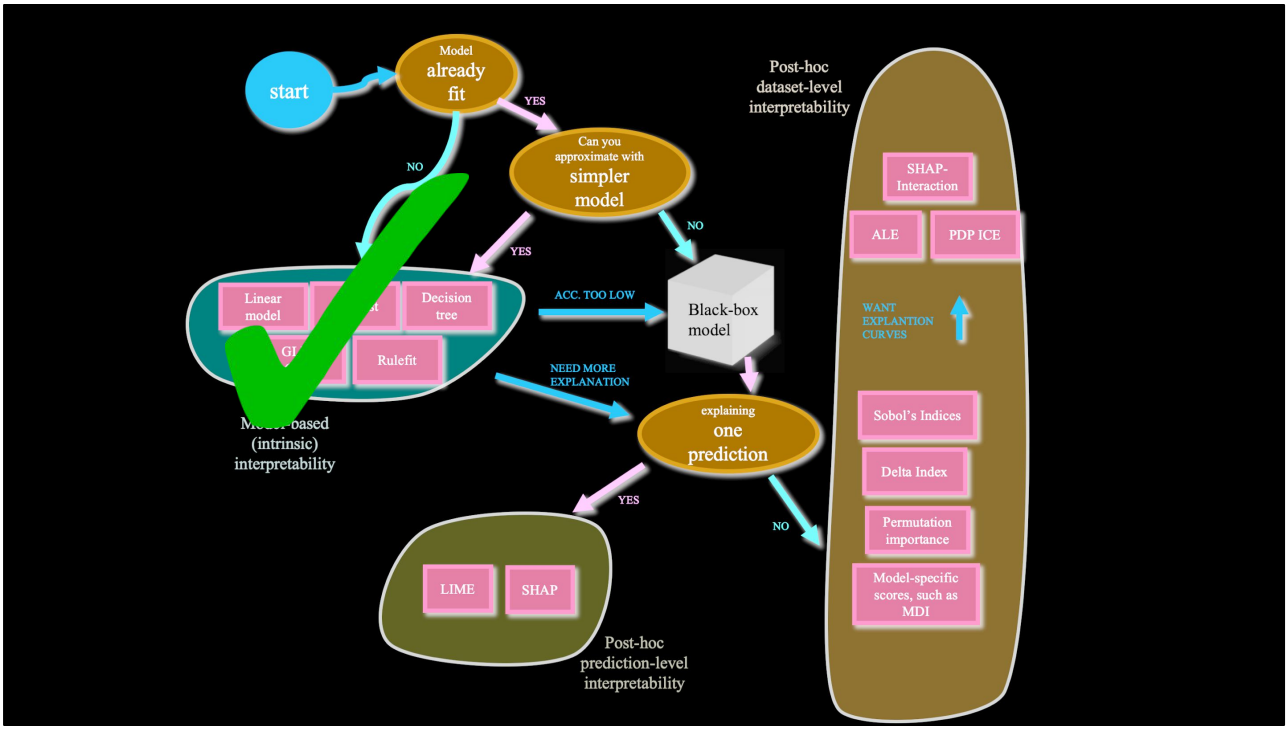
example 6: prototypical neural networks



chen et al. 2018

This line of work aims to help make more transparent neural networks by using prototypes. The network explains which part of the image it uses to make its decision (heatmaps) and which prototype that part is closest to.


start
Model already fit
YES
NO
Can you approximate with simpler model
NO
YES
Linear model
Rule list
Decision tree
GLM / GAM
Rulefit
Model-based (intrinsic) interpretability
ACC. TOO LOW
Black-box model
NEED MORE EXPLANATION
explaining one prediction
YES
NO
LIME
SHAP
Post-hoc prediction-level interpretability
Post-hoc dataset-level interpretability
SHAP-Interaction
ALE
PDP ICE
WANT EXPLANTION CURVES
Sobol's Indices
Delta Index
Permutation importance
Model-specific scores, such as MDI

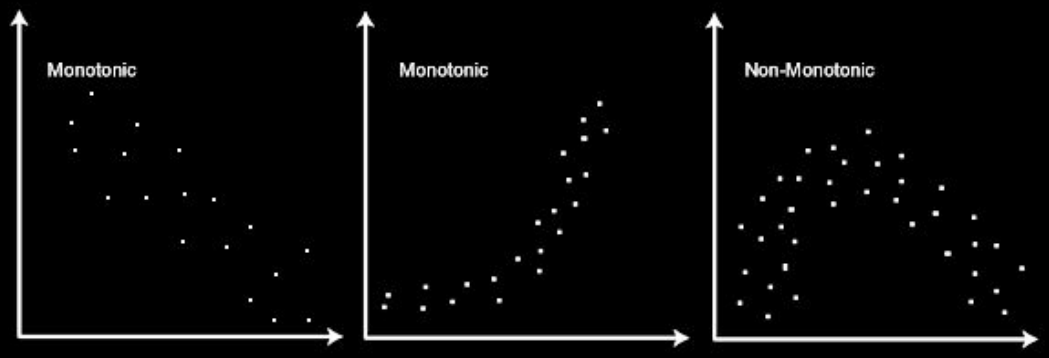

start
Model already fit
YES
NO
Can you approximate with simpler model
YES
NO
Linear model
Decision tree
Rulefit
ACC. TOO LOW
Black-box model
NEED MORE EXPLANATION
explaining one prediction
YES
NO
LIME
SHAP
Post-hoc prediction-level interpretability
(intrinsic) interpretability
Post-hoc dataset-level interpretability
SHAP-Interaction
ALE
PDP ICE
WANT EXPLANTION CURVES
Sobol's Indices
Delta Index
Permutation importance
Model-specific scores, such as MDI

Which features are **globally important**?

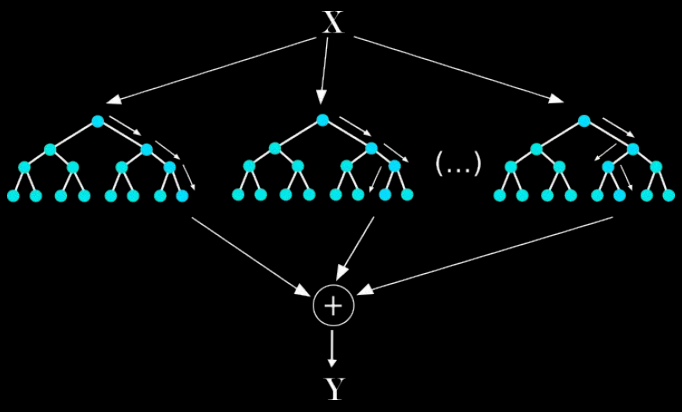
Here, globally important means we are assigning importance over the dataset, not just for an individual point.

global linear feature importances

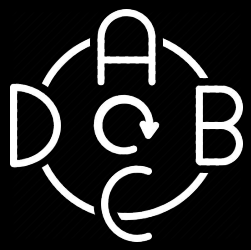
- no model: (rank) correlation, partial correlation
- linear / logistic: coefficients (caution with categorical variables)



- tree / tree ensembles: (normalized) total impurity reduction by a feature
- neural network / nonlinear svm: None



permutation importance (breiman 2001)



finding important variables: no interactions
screening unimportant variables: use interactions
only permutation importance requires a model

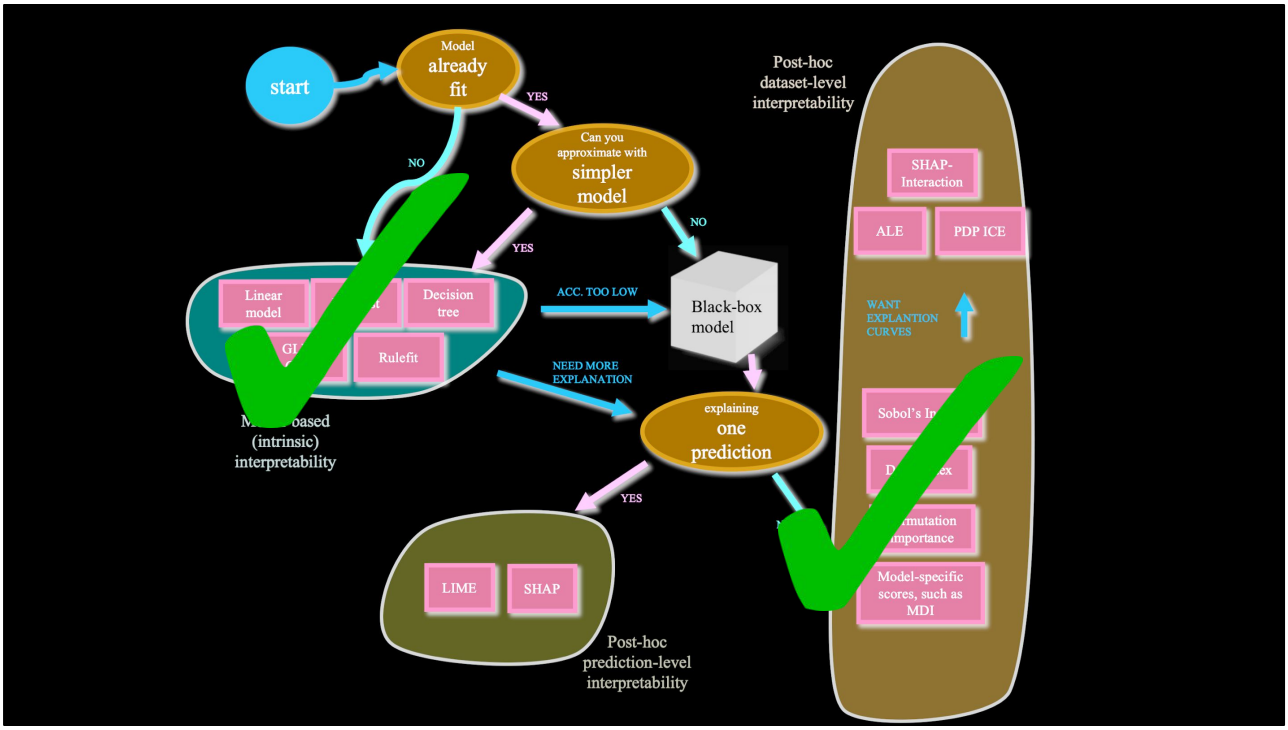
delta index (borgonovo 2007)

$$\delta_i = \frac{1}{2}\mathbb{E}\int |f_Y(y) - f_{Y|X_i}(y)|dy$$

sobol's indices (sobol 1993)

$$Y = g(\mathbf{X}) = g_0 + \sum_i g_i(X_i) + \sum_i \sum_{j>i} g_{ij}(X_i, X_j) + \cdots + g_{1,2,\ldots,n}$$

$$g_0 = \mathbf{E}(Y),\ g_i = \mathbf{E}(Y|X_i) - g_0,\ g_{ij} = \mathbf{E}(Y|X_i, X_j) - g_i - g_j - g_0$$


start
Model already fit
YES
NO
Can you approximate with simpler model
YES
NO
Linear model
Decision tree
Rulefit
ACC. TOO LOW
Black-box model
NEED MORE EXPLANATION
explaining one prediction
YES
LIME
SHAP
Post-hoc prediction-level interpretability
(intrinsic) interpretability
Post-hoc dataset-level interpretability
SHAP-Interaction
ALE
PDP ICE
WANT EXPLANTION CURVES
Model-specific scores, such as MDI

How does the model use different features?

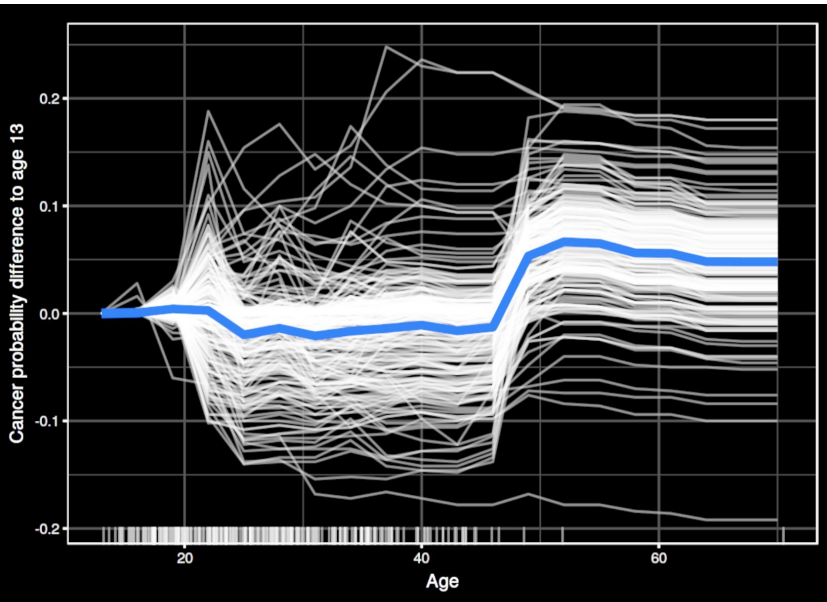
Here, we consider importance as it varies over the dataset, not just a scalar value for each feature.

PDP ICE Plot

friedman 2001

ALE Plot

apley 2016



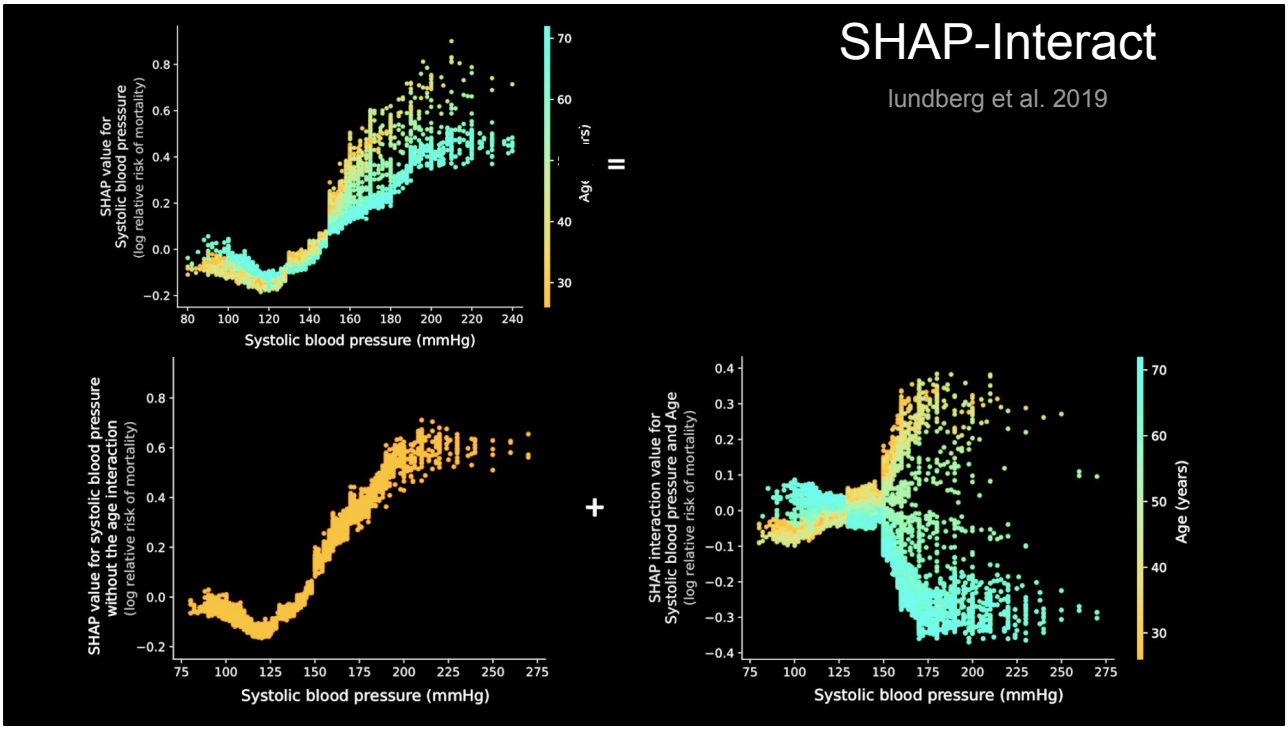
molnar 2019

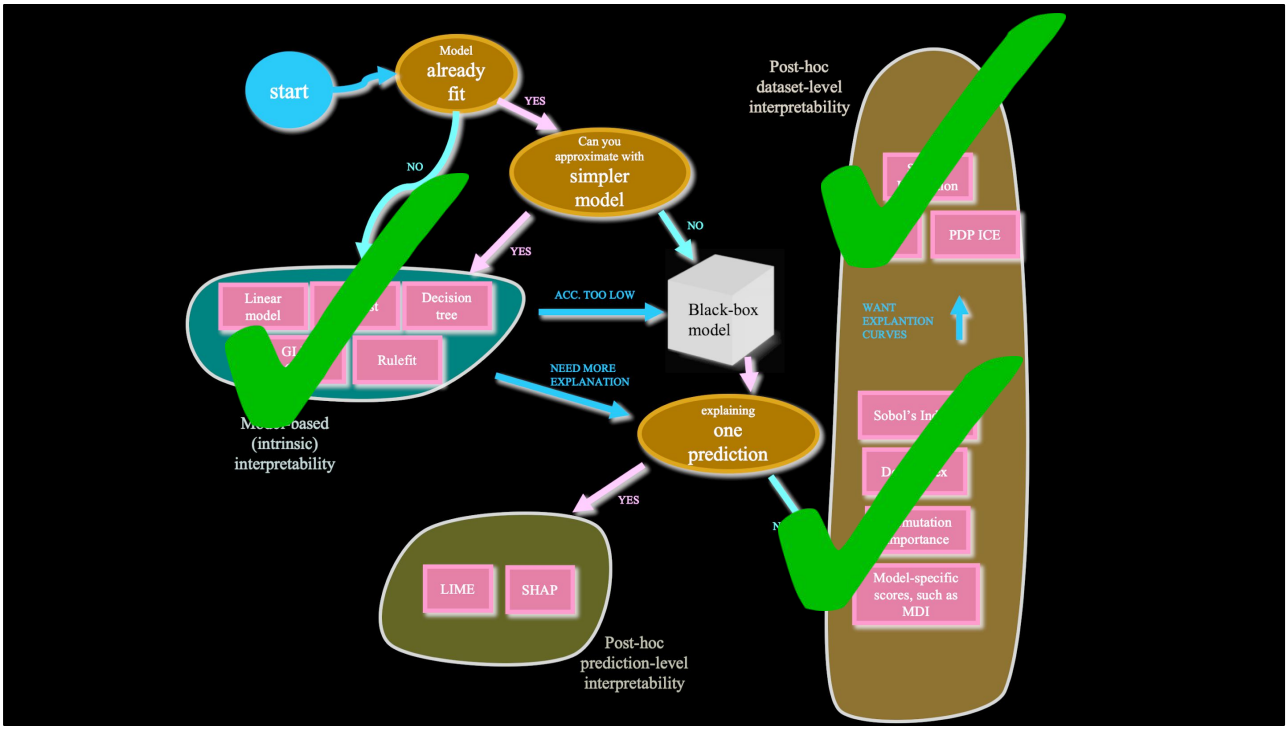
SHAP-Interact

lundberg et al. 2019

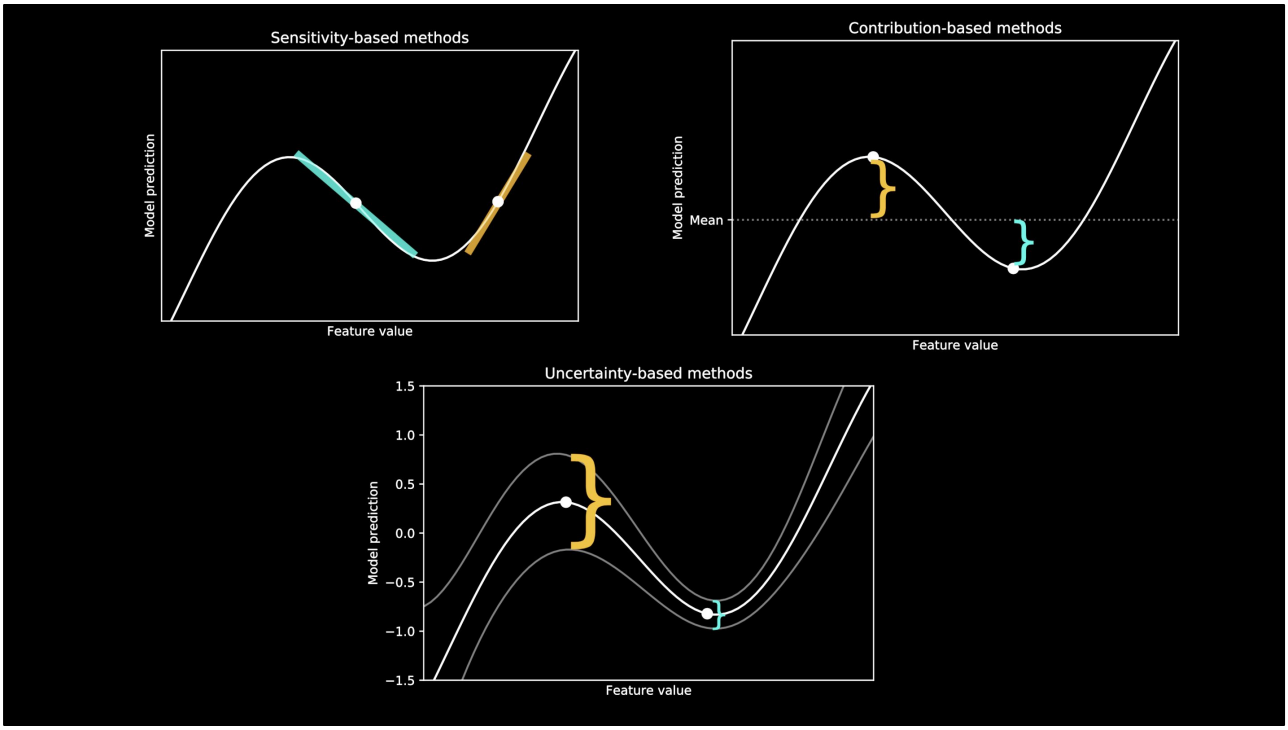
$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M-1)!} \nabla_{ij}(S)$$

$$\begin{aligned}\nabla_{ij}(S) &= f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \\ &= f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - [f_x(S \cup \{i\}) - f_x(S)]\end{aligned}$$

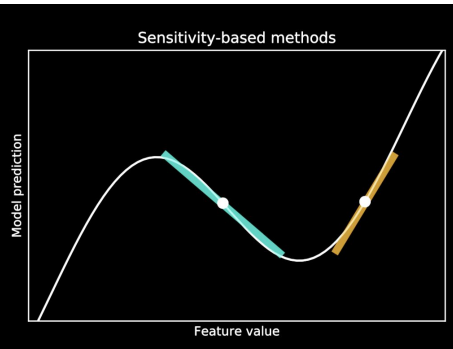

SHAP-Interact
lundberg et al. 2019
SHAP value for
Systolic blood presssure
(log relative risk of mortality)
Systolic blood pressure (mmHg)
=
SHAP value for systolic blood pressure
without the age interaction
(log relative risk of mortality)
Systolic blood pressure (mmHg)
+
SHAP interaction value for
Systolic blood pressure and Age
(log relative risk of mortality)
Systolic blood pressure (mmHg)
Age (years)

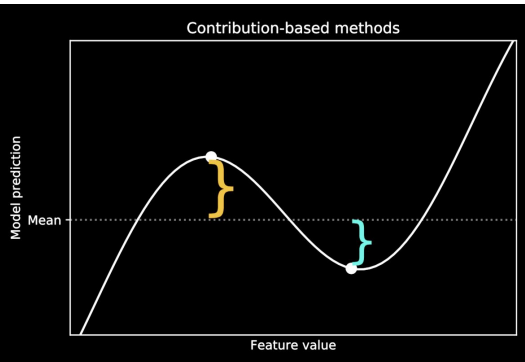

start
Model already fit
YES
NO
Can you approximate with simpler model
NO
YES
Linear model
Decision tree
Rulefit
ACC. TOO LOW
Black-box model
NEED MORE EXPLANATION
explaining one prediction
YES
LIME
SHAP
Post-hoc prediction-level interpretability
(intrinsic) interpretability
Post-hoc dataset-level interpretability
PDP ICE
WANT EXPLANTION CURVES
Model-specific scores, such as MDI

How can we understand
one prediction?



White points are the points we are trying to explain. Most existing post-hoc local explanation methods can fit into one of these categories.





LIME (ribeiro et al. 2016)

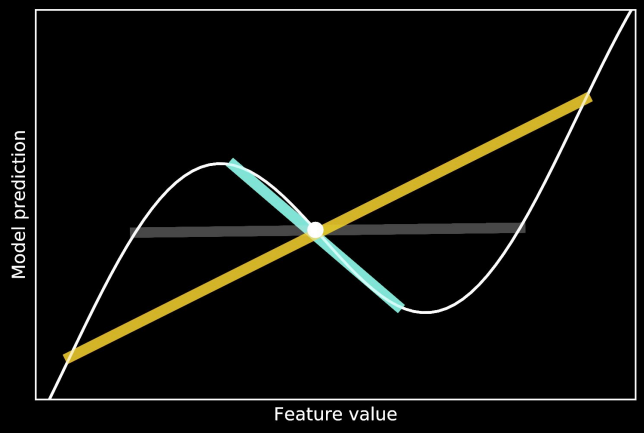
$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

SHAP (lundberg & lee, 2017)

$$\phi_i = \mathrm{C} \sum_{S \subseteq F \setminus \{i\}} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

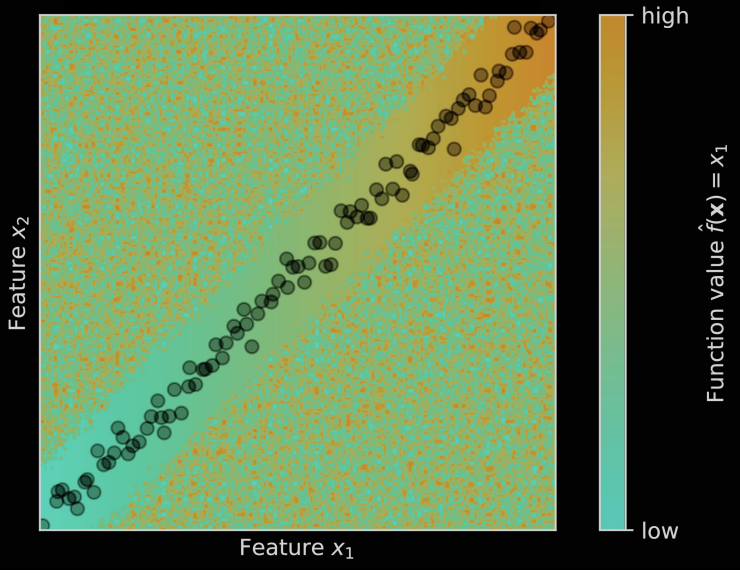
Sampling
the problem with local explanations



Local explanations are very sensitive to the region which they are explaining. The function at the white point has a different local derivative depending on the window in which it is calculated.

Sampling 2: conditional sampling can introduce spurious attribution for interactions

Let $f(\mathbf{x}) = X_1$

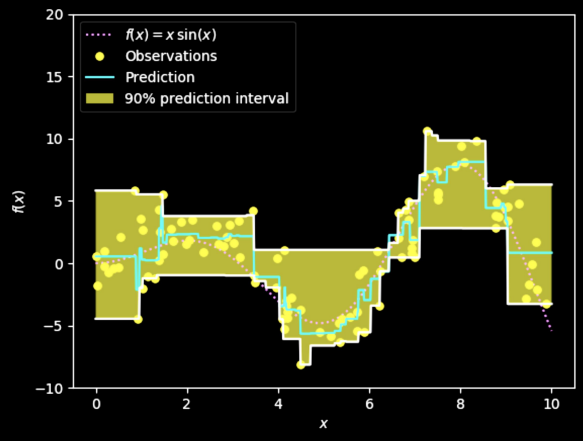


Name	Sampling			Generation		Output Type
	E	D	Type[a]	Model	Objectives	
Trepan [4]	Pop.	Pop.	P&R	Decision tree	Complexity, Fidelity	Decision tree
BETA [15]	Pop.	Pop.	S&D	Rule-based model	Interpretability, Fidelity Unambiguity	Rule-based global model
GoldenEye* [11]	Pop.	Pop.	P&R	Permutation	Fidelity, Interaction size	Important features interactions
VIN [13]	Pop.	Pop.	P&D	Permutation	ANOVA projection	Important features interactions
PDP ICE [14]	Pop.	Pop.	P&D	NA	NA	Dependence plot for one feature
QII [5]	Indiv. or or group	Pop.	P&D	NA	NA	Var. importance
Anchors [23]	$\{x_e\}$	Pop.	P&R	Rule-based model	Fidelity, Complexity, Generality	Rule-based local model
LIME [22]	$\{x_e\}$	$\emptyset$	P&R	Linear model	Fidelity, Complexity	Var. importance
Shapley [30]	$\{x_e\}$	Pop.	P&D	NA	NA	Var. importance
LEMNA [10]	$\{x_e\}$	$\emptyset$	P&R	Mixture of linear models	Fidelity, Complexity	Var. importance
Local Gradient [2]	$\{x_e\}$	Pop.	NA	Parzen window	Fidelity	Directions of highest slope
Counter-factuals [28]	$\{x_e\}$	Pop.	NA	Small deviation	Target output, Distance input	Example-based

henin & metayer 2019

easy, effective uncertainty

- ensemble uncertainty
- quantile loss prediction interval
- bayesian methods



There are a variety of methods for obtaining uncertainty estimates on predictions.

What else is out there?

- Influence functions - find points which highly influenced a model (koh & liang 2017)
- TCAV - see if representations of certain points learned by a DNN are linearly separable (kim et al. 2017)
- MMD Critic - find a few points which summarize classes (kim et al. 2016)
- ACD - hierarchical interpretations for DNNs (singh et al. 2019)