Explaining by removing: a unified framework for model explanation

Ian Covert

UNIVERSITY *of* WASHINGTON

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Outline

1. **Motivation**
2. A unified framework
3. Example methods
4. Implications

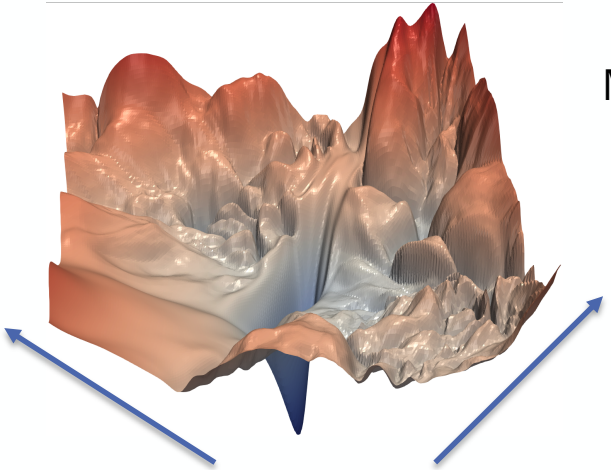
Modern machine learning

- ML/AI becoming increasingly widespread
- **Black-box models** are now predominant
  - GBTs, RFs, DNNs, CNNs, LSTMs, transformers, etc.
- Various concerns about **model transparency**

Explainable AI

XAI is difficult

- ML models are **highly complex**



(Showing just 2 dimensions)

Model prediction landscape

How to visualize all dimensions?

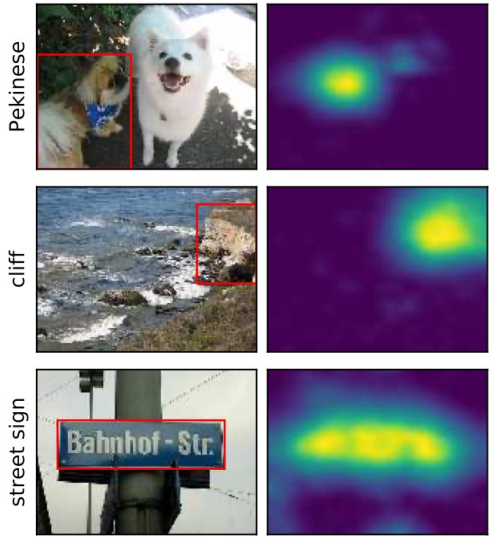
How to summarize for a user?

Very active XAI field

- Impossible to summarize models **fully**, but many approaches say **something** about they work
- Methods include…
  - SHAP, LIME, SAGE, Occlusion, DeepLift, SmoothGrad, Integrated Gradients, GradCAM, CXPlain, L2X, INVASE, Meaningful Perturbations, Extremal Perturbations, RISE, TCAV, Guided Backprop, Excitation Backprop, IME, QII, PredDiff, MIR, Permutation Tests, LRP, FIDO-CA, Masking Model, Expected Gradients, LossSHAP, Shapley Effects, MP2-G, Saliency Maps, PDPs, ICEs, TreeSHAP

Model explanation examples

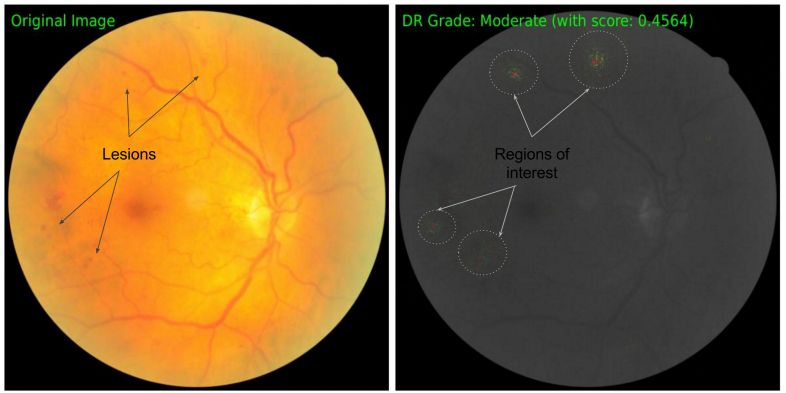
- Interpreting image classifiers



Fong & Vedaldi, 2017

Model explanation examples (cont.)

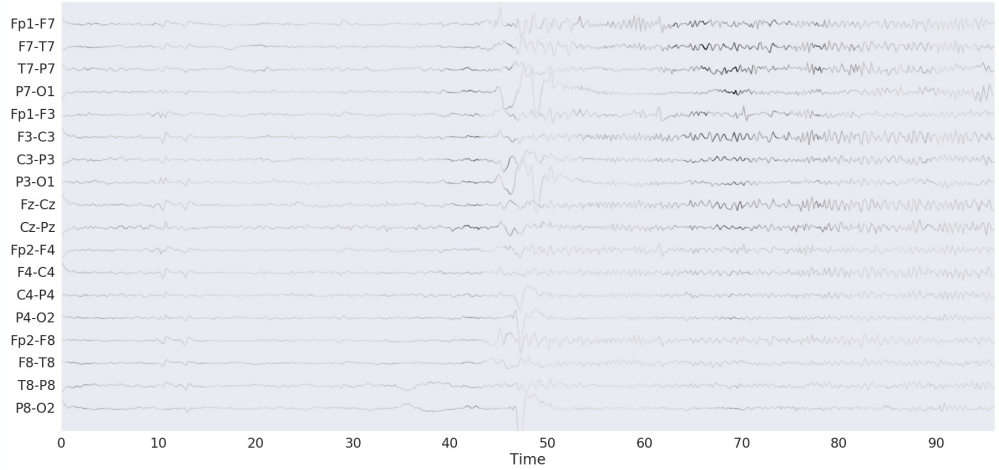
- Identifying regions of medical interest



Sundararajan et al., 2017

Model explanation examples (cont.)

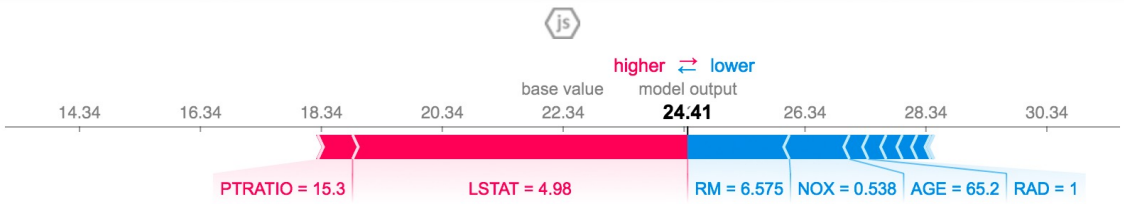
- Visualizing seizure prediction models



Covert et al., 2019

Model explanation examples (cont.)

- Decomposing tabular model predictions



Lundberg & Lee, 2017

State of the field

- **Significant progress**, and some widely used tools (SHAP, LIME, IntGrad)
- However, XAI has many open problems
- Today's focus:
  - Building theoretical foundations
  - Connecting XAI with human psychology
  - Unifying/consolidating approaches

High-level topics, often overlooked in existing work

Our recent paper

- Explaining by removing: a unified framework for model explanation **(JMLR 2021)**
- With Scott Lundberg and Su-In Lee

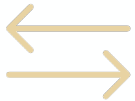
Questions?

Outline

1. Motivation
2. **A unified framework**
3. Example methods
4. Implications

Explaining by removing

A **unifying theory** that describes 25+ methods

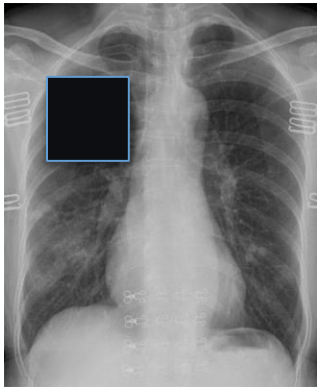


Monolithic algorithms → **interchangeable choices**

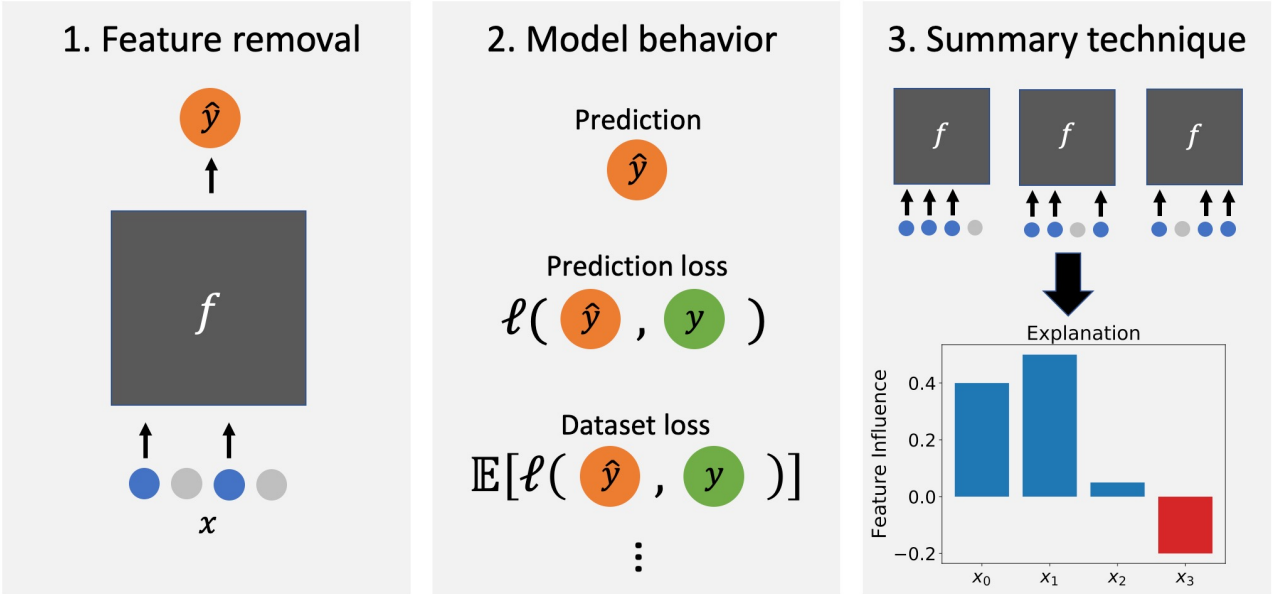
One key idea about how to explain ML models

Removal-based explanations

- A new class of methods
- Methods that **remove features** and observe the impact on the model
- Intuition: blocking part of a medical image to understand a diagnosis

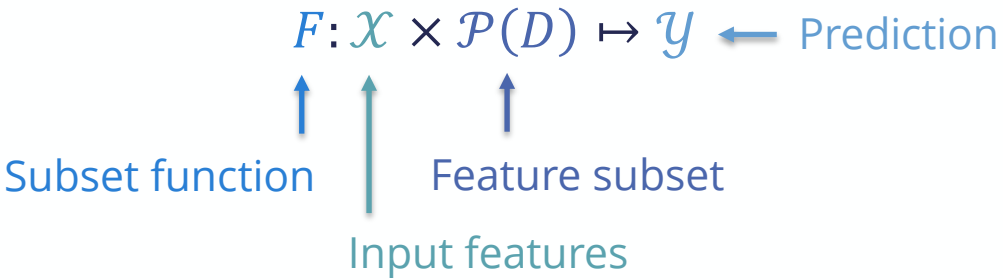


A simple framework



A mathematical view

1. Feature removal strategy



A mathematical view

1. Feature removal strategy

$$F: \mathcal{X} \times \mathcal{P}(D) \mapsto \mathcal{Y}$$

2. Model behavior to explain

u: $\mathcal{P}(D) \mapsto \mathbb{R}$ ← Associated value

Cooperative game

Subset (of features)

A mathematical view

1. Feature removal strategy

$$F: \mathcal{X} \times \mathcal{P}(D) \mapsto \mathcal{Y}$$

2. Model behavior to explain

$$u: \mathcal{P}(D) \mapsto \mathbb{R}$$

3. Summary technique

Explanation mapping → $E: \mathcal{U} \mapsto \mathbb{R}^d$ or $E: \mathcal{U} \mapsto \mathcal{P}(D)$

Cooperative game — Feature attributions — Influential subset

Outline

1. Motivation
2. A unified framework
3. **Example methods**
4. Implications

Existing methods

- Our framework describes a **significant portion** of the XAI literature
  - SHAP, LIME, permutations tests, Meaningful Perturbations, etc.
- Removal-based explanations can be distilled down to their **three choices**

LIME

1. **Feature removal:** default values
2. **Model behavior:** individual prediction
3. **Summary technique:** fit linear model

Ribeiro, Singh & Guestrin, KDD 2016

SHAP

1. **Feature removal:** marginalize out
2. **Model behavior:** individual prediction *(same as LIME)*
3. **Summary technique:** Shapley values

Lundberg & Lee, NeurIPS 2017

SAGE

1. **Feature removal:** marginalize out *(same as SHAP)*
2. **Model behavior:** dataset loss
3. **Summary technique:** Shapley values *(same as SHAP)*

Covert, Lundberg & Lee, NeurIPS 2020

Permutation test

1. **Feature removal:** marginalize out *(same as SAGE)*
2. **Model behavior:** dataset loss *(same as SAGE)*
3. **Summary technique:** remove individual

Breiman, 2001

A unified framework

Table 1: Choices made by existing removal-based explanations.

Method	Removal	Behavior	Summary
IME (2009)	Separate models	Prediction	Shapley value
IME (2010)	Marginalize (uniform)	Prediction	Shapley value
QII	Marginalize (marginals product)	Prediction	Shapley value
SHAP	Marginalize (conditional/marginal)	Prediction	Shapley value
KernelSHAP	Marginalize (marginal)	Prediction	Shapley value
TreeSHAP	Tree distribution	Prediction	Shapley value
LossSHAP	Marginalize (conditional)	Prediction loss	Shapley value
SAGE	Marginalize (conditional)	Dataset loss (label)	Shapley value
Shapley Net Effects	Separate models (linear)	Dataset loss (label)	Shapley value
SPVIM	Separate models	Dataset loss (label)	Shapley value
Shapley Effects	Marginalize (conditional)	Dataset loss (output)	Shapley value
Permutation Test	Marginalize (marginal)	Dataset loss (label)	Remove individual
Conditional Perm. Test	Marginalize (conditional)	Dataset loss (label)	Remove individual
Feature Ablation (LOCO)	Separate models	Dataset loss (label)	Remove individual
Univariate Predictors	Separate models	Dataset loss (label)	Include individual
L2X	Surrogate	Prediction loss (output)	High-value subset
REAL-X	Surrogate	Prediction loss (output)	High-value subset
INVASE	Missingness during training	Prediction mean loss	High-value subset
LIME (Images)	Default values	Prediction	Linear model
LIME (Tabular)	Marginalize (replacement dist.)	Prediction	Linear model
PredDiff	Marginalize (conditional)	Prediction	Remove individual
Occlusion	Zeros	Prediction	Remove individual
CXPlain	Zeros	Prediction loss	Remove individual
RISE	Zeros	Prediction	Mean when included
MM	Default values	Prediction	Partitioned subsets
MIR	Extend pixel values	Prediction	High-value subset
MP	Blurring	Prediction	Low-value subset
EP	Blurring	Prediction	High-value subset
FIDO-CA	Generative model	Prediction	High-value subset

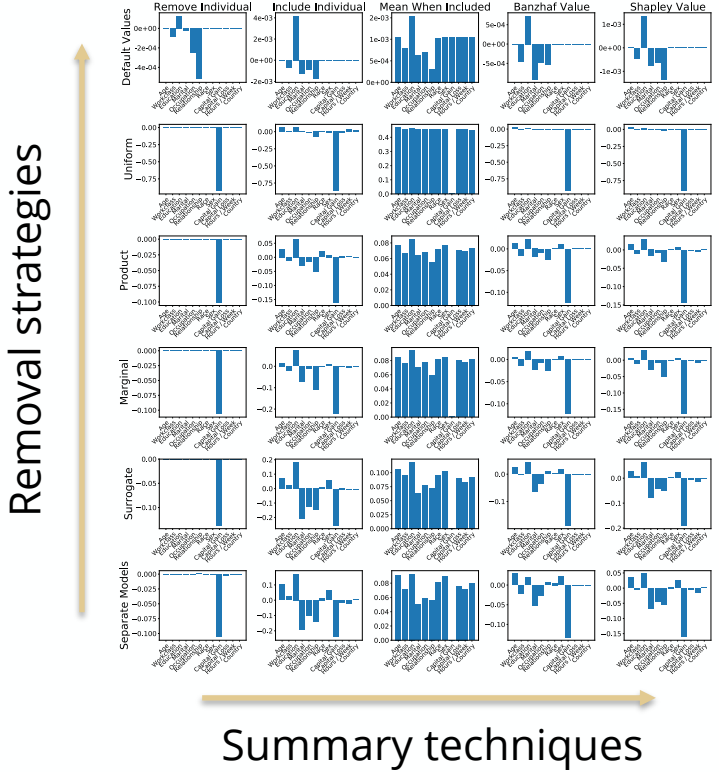
- 25+ methods
- Local and global methods
- Feature selection and feature attribution
- State-of-the-art methods are constructed using **interchangeable choices**

New removal-based explanations

Brand-new methods can be designed by specifying **new choices**

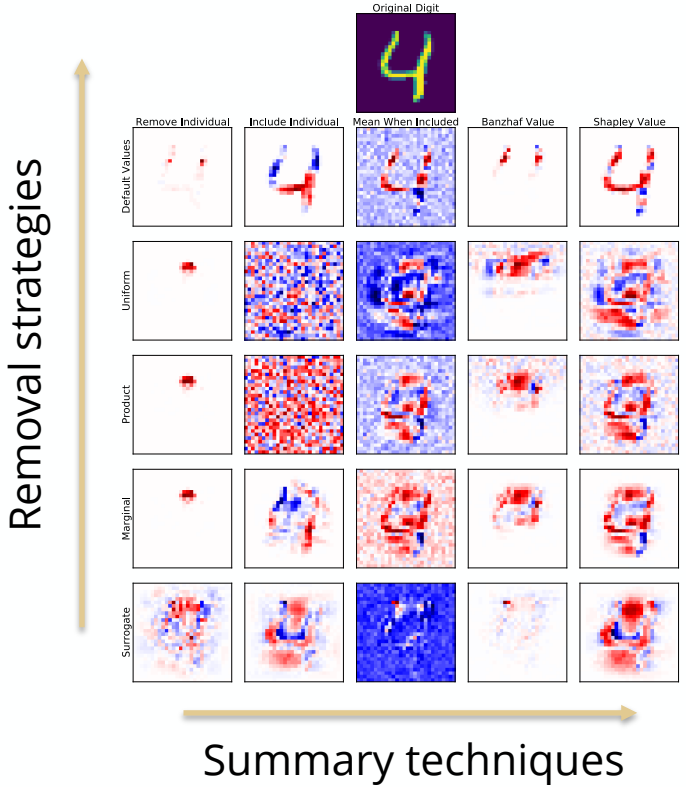
Hybrid methods can be designed by **mixing and matching** choices

Mixing and matching



- Tested many explanations on census income dataset
- Different combinations of **feature removal** and **summary techniques**

Mixing and matching (cont.)



- Ran similar experiment for MNIST digit recognition
- See paper for more details and metrics

Outline

1. Motivation
2. A unified framework
3. Example methods
4. **Implications**

Connections with related fields

- Our framework reveals underlying connections with several related fields
  - Cognitive psychology
  - Cooperative game theory
  - Information theory
- Connections apply to **all** removal-based explanations
  - New insights for many methods!

Cognitive psychology connections

- Feature removal is a form of **counterfactual reasoning**
  - Removing features is a **subtractive counterfactual**
  - Consider model's behavior when certain information is not observed
- Connections with **norm theory** and the **downhill rule**
  - Precedent in psychology suggests these methods are meaningful to **humans**, not just mathematicians

Game-theoretic explanations

- Every method has an underlying **cooperative game**

$$v: 2^d \mapsto \mathbb{R}$$

- XAI has often reinvented or borrowed ideas from game theory (e.g., Shapley value)

Game-theoretic explanations (cont.)

Summarization	Methods	Related To
Shapley value	Shapley Net Effects, IME, QII, SHAP (TreeSHAP, KernelSHAP, LossSHAP), Shapley Effects, SAGE	Shapley value, probabilistic values, modeling cooperative games
Mean value when included	RISE	Banzhaf value, probabilistic values, modeling cooperative games
Remove/include individual players	Occlusion, PredDiff, CXPlain, permutation tests, univariate predictors, feature ablation (LOCO)	Probabilistic values, modeling cooperative games
Linear model	LIME	Shapley value, Banzhaf value, modeling cooperative games
High/low value coalitions	MP, EP, MIR, MM, L2X, INVASE, REAL-X, FIDO-CA	Maximum/minimum excess

Game-theoretic explanations (cont.)

- The **Shapley** and **Banzhaf** values satisfy many desirable properties
  - SHAP, SAGE, Shapley Effects, etc.
- There is probably more to learn from game theory!

Information-theoretic explanations

- Do model explanations relate to the information contained in each feature?
  - In a rigorous sense, **yes!**
- We identify connections with **information theory**
  - A subfield of mathematics

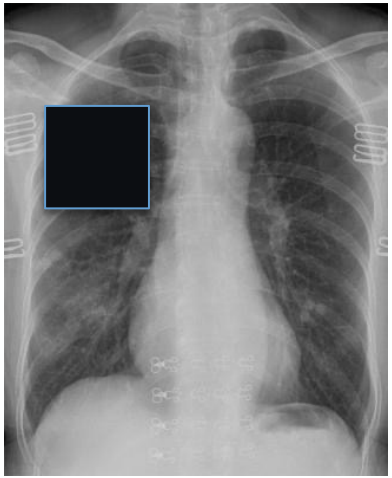
Info theory (cont.)

- Methods must remove features using their **conditional distribution** (the first choice)
  - Intuitively, account for all available information
- Chest X-Ray example
- Formally:

$$F(x_S) = \mathbb{E}[f(X)|X_S = x_S]$$
$$= \int p(x_{\bar{S}}|x_S) f(x_S, x_{\bar{S}})$$

Weighting alternative values using conditional dist.

Model prediction



Info theory (cont.)

Model Behavior	Set Function	Methods	Related To
Prediction	u_x	Occlusion, MIR, MM, IME, QII, LIME, MP, EP, FIDO-CA, RISE, SHAP, KernelSHAP, TreeSHAP	Conditional probability, conditional expectation
Prediction loss	v_{xy}	LossSHAP, CXPlain	Pointwise mutual information
Prediction mean loss	v_x	INVASE	KL divergence with conditional distribution
Dataset loss	v	Permutation tests, univariate predictors, feature ablation (LOCO), Shapley Net Effects, SAGE	Mutual information (with label)
Prediction loss (output)	w_x	L2X, REAL-X	KL divergence with full model output
Dataset loss (output)	w	Shapley Effects	Mutual information (with output)

Wrapping up

Implications for users & orgs.

- You aren't just choosing between LIME and SHAP
  - Can easily create **hybrid** or **brand-new** methods
  - Many options → must consider what's best for **your use-case**
- Removal-based explanations are well-suited to many, but **not all** XAI problems
  - ☑ Uncovering informative features
  - ☑ Identifying hidden biases
  - ☒ Recommending changes to achieve alternate outcomes

Thanks!