

Data 146: Foundations for CPH

Interpretability

Irene Y. Chen

What would most increase your confidence in a readmission prediction algorithm?

A. An explanation listing its important features
B. Good performance at a second hospital
C. A clinician says its explanation makes sense
D. A visualization of its internal neurons



"Interpretability as a concept is underspecified… transparency to humans and post-hoc explanations [are] competing notions."

– Zack Lipton, "**The Mythos of Model Interpretability"**

"Interpretability can assist in qualitatively ascertaining whether other desiderata—such as fairness, privacy, reliability, robustness, causality, usability and trust—are met. For example, one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern."
- Finale Doshi-Velez, "Towards A Rigorous Science of Interpretable Machine Learning"

"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead"
- Cynthia Rudin, Nature Machine Intelligence 2019

“Interpretability as a concept is underspecified… transparency to humans and post-hoc explanations [are] competing notions.”

– Zack Lipton, “**The Mythos of Model Interpretability**”



“Interpretability can assist in qualitatively ascertaining whether other desiderata—such as fairness, privacy, reliability, robustness, causality, usability and trust—are met. For example, one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern.”
- Finale Doshi-Velez, “Towards A Rigorous Science of Interpretable Machine Learning”

“Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”
- Cynthia Rudin, Nature Machine Intelligence 2019

“Interpretability as a concept is underspecified… transparency to humans and post-hoc explanations [are] competing notions.”

– Zack Lipton, “**The Mythos of Model Interpretability”**

“Interpretability can assist in qualitatively ascertaining whether other desiderata—such as fairness, privacy, reliability, robustness, causality, usability and trust—are met. For example, one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern.”
- Finale Doshi-Velez, “Towards A Rigorous Science of Interpretable Machine Learning”



“Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”
- Cynthia Rudin, Nature Machine Intelligence 2019

Interpretability in AI public policy

- **EU AI Act**: Requires that high-risk systems be designed so that their operation is "sufficiently transparent."
  - Human Oversight (Article 14): A human must be able to "correctly interpret the AI system's output."
  - The "State of the Art" Requirement: Requires companies to follow "state-of-the-art" technical standards. In practice, industry standard for feature attribution is SHAP or LIME
- **Canada**: The Algorithmic Impact Assessment (AIA) is an assessment to determine if you are required to provide a "plain-language explanation" or a "technical explanation."
- **China**: Large platforms must file their algorithm's "basic principles" and "main running mechanisms" with the government.

Outline

- **Local methods** (15 mins)
- **Global methods** (20 mins)
- **Mechanistic methods** (15 mins)



How can we make Data 146 better for you?

Learning Objective: Understand criteria and formalization of interpretability

Interpretability Issues

- People understand simple models
  - George Miller, 7±2: “There seems to be some limitation built into us either by learning or by the design of our nervous systems, a limit that keeps our channel capacities in this general range.”
- What to do about complex models?
  - Combat overfitting
  - True complexity



Rashoman was a 1950’s Jidaigeki drama film

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2), 81–97.

Trust

- Critical for adoption of ML models
  - Case-specific prediction: local interpretability
    - Clinical decision-support
  - Confidence in model: global interpretability
    - Population health
  - Reverse-engineering models: mechanistic interpretability
    - Neural networks
- RCTs are flawed because of design constraints
- ML models are also flawed
  - Train and test samples often drawn from same population
  - Are results applicable?

Explanation – not a new idea!

- Mycin (1974) used backward-chaining rules to determine whether a patient had a bacterial infection that needed to be treated, and how best to treat
- Collection of several hundred rules, each of which encoded a relatively independent fact
- Certainty factors encoded a theory of uncertain reasoning (tantamount to very strong independence assumptions, leading to problems)
- Context mechanism to fill in implicit clauses in rules;
  patient→site→infection→culture→organism→ drug

RULE092

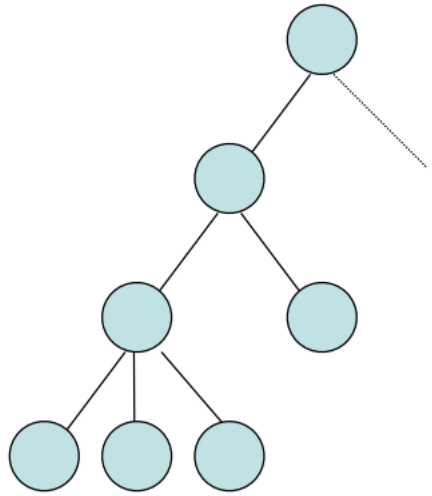
IF we have identified organisms for
    which treatment is indicated
THEN select a treatment that
    covers those organisms

RULE037

IF the organism
  1) stains gram positive
  2) has a coccus shape
  3) grows in chains
THEN
  There is suggestive evidence (.7)
    that the identify of the organism
    is streptococcus

Mycin (1794) in detail

- Dynamically generates an and/or tree via backward chaining
- To find out a fact
  - If there are rules that can conclude it, run them
  - Otherwise, ask the user
- To run a rule
  - Find out if the facts in the premises are true (enough)
  - If they all are, then assert the conclusion (with suitable certainty factor)
- This traces out the equivalent of a flowchart, but by generating it on the fly from underlying rules
  - Knowledge is always applied when relevant
  - Can answer questions about how/why it is working

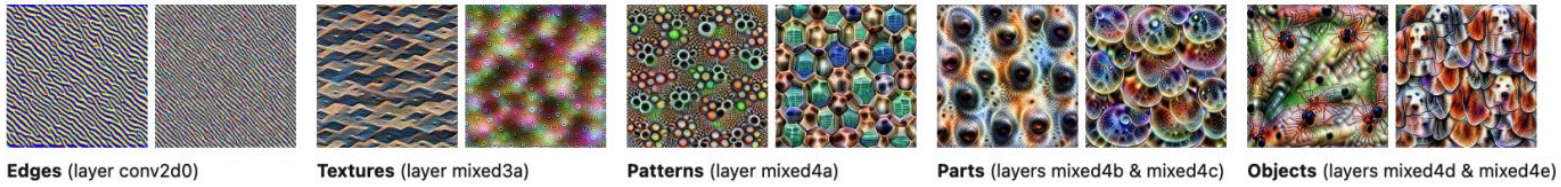


Local vs. Global interpretability

- Global interpretability – understand model as a whole
  - Will it work prospectively as intended? (dataset shift, label misspecification, label leakage)
  - What data was most useful? (find more signal of a similar type, form causal hypotheses, figure out how to simplify for deployment purposes)

Local vs. Global interpretability

- Global interpretability – understand model as a whole
  - For **any model**: do feature ablation. How does performance on held-out data change?
  - Ex. **linear models**: look at largest positive and negative weight features
  - Ex: **decision trees**: look at the top few splits
  - Ex. **deep models**: visualize specific filters



Olah et al., Feature Visualization: How neural networks build up their understanding of images, Distill 2017 https://distill.pub/2017/feature-visualization/

  - Increasingly more difficult as models become more complex…

Predicting Psychiatric Readmission with Bag-ofWords NLP model

- 470 of patient cohort (4687) were readmitted within 30 days for psychiatric diagnoses
  - 2977 readmitted with a nonpsychiatric diagnosis
  - 1240 not readmitted (only 26%)
- We built a model using demographics, comorbidity + 75 topics from LDA on notes
  - Top 1000 TF/IDF words from each patient's notes
  - Considerable overlap, but total vocabulary ~66K words
- SVM models to predict readmission
  - Baseline features
  - Baseline + top-1000 bag-of-words (really 66K)
  - Baseline + 75 topics

Table 3. Comparison of models with and without inclusion of LDA topics

Configuration	*AUC*	*Sensitivity*	*Specificity*
Baseline = age/gender/insurance/ Charlson	0.618	0.979	0.104
Baseline+top-1 words	0.654	—	—
Baseline+top-10 words	0.676	—	—
Baseline+top-100 words	0.682	—	—
Baseline+top-1000 words	0.682	0.213	0.945
Baseline+75 topics (no words)	0.784	0.752	0.634

Abbreviations: AUC, area under the curve; LDA, Latent Dirichlet Allocation.

Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro VM, McCoy TH, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational Psychiatry [Internet]. 2016 Oct;6(10):e921-5. Available from: http://dx.doi.org/10.1038/tp.2015.182

Example topics for MDD patients readmitted with a psychiatric diagnosis within 30 days

Terms	*Topic annotation*
*patient *alcohol withdrawal* depression *drinking* end ativan *etoh drinks* medications clinic inpatient diagnosis days hospital <*substance use treatment program name*> *use abuse* problem number	Alcohol
*mg daily discharge *anxiety klonopin seroquel clonazepam* admission wellbutrin given md lexapro date b signed night low admitted sustained hospitalization	Anxiety
**ideation suicidal mood decreased* hallucinations history *depressed depression thought* psychiatric energy denied sleep auditory appetite homicidal symptoms increased speech *thoughts*	Suicidality
**ect depression* treatment treatments dr mg course <*ECT physician name*> symptoms received medications prior improved decreased medication md trials tsh continued qhs	ECT
**weight eating* admission discharge hospital intake *loss* date hospitalization day dr week physical months *prozac food* increased md did *anorexia*	Anorexia
**seizure seizures* intact *eeg neurology* normal *temporal dilantin* head *bilaterally events activity* weakness sensation disorder tongue *neurologist brain* loss *tegretol*	Seizure
**therapist mother* program *father* disorder age school parents brother *abuse* treatment *relationship* outpatient college behavior partial plan currently *group personality*	Psychotherapy
*psychiatry *suicide overdose attempt transferred* depression *transfer* level *tylenol* hospital service *unit* normal floor screen *tox* room admission medical general	Overdose
**baby delivery bleeding vaginal breast feeding cesarean* weight ibuprofen *maternal newborn* available p fever *pregnancy* sex estimated *danger* gp	Postpartum
**psychotic thought* features *paranoid psychosis paranoia* symptoms psychiatric dose continued treatment mental cognitive memory *risperidone* people th somewhat interview affect	Psychosis

What are other ways to learn models with “good” global interpretability?

- Generalized additive models (GAMs)
- Falling rule lists
- Supersparse linear integer models
- …

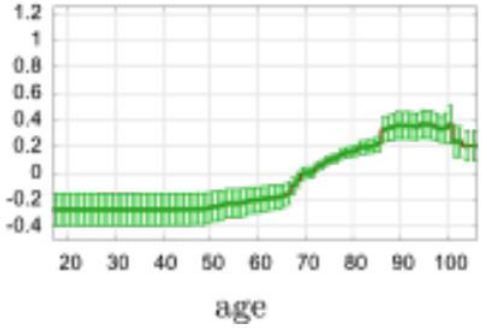
Generalized additive models (GAMs)

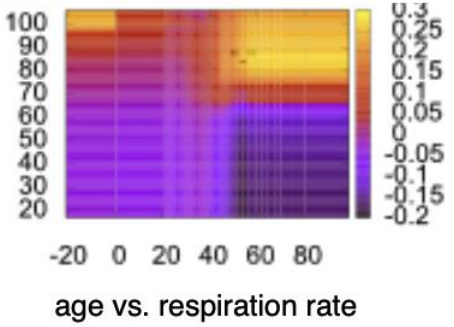
- GAMs with pairwise interactions have the form:

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

- g is the link function (e.g. logistic, for binary data), and E[f] = 0.

Model	Pneumonia	Readmission
Logistic Regression	0.8432	0.7523
GAM	0.8542	0.7795
GA^2M	0.8576	0.7833
Random Forests	0.8460	0.7671
LogitBoost	0.8493	0.7835





[Caruana et al., KDD '15]

Falling rule lists

- Ordered list of if-then rules where:
  1. It is a decision list, i.e. order matters
  2. Probability of outcome decreases monotonically

	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 F, Rudin C. Falling rule lists. AISTATS [Internet]. 2015; Available from: https://dblp.org/rec/conf/aistats/WangR15

Empirical Test: 30-Day Hospital Readmission

- 8,000 patients
- Features: “impaired mental status,” “difficult behavior,” “chronic pain,” “feels unsafe” and over 30 other features
- Mined rules with support ≥5%, no more than two conditions
- Expected length of decision list = 8
- Compared to SVM, Random Forest, Logistic Regression, CART. Inductive Logic Programming

Method	Mean AUROC (STD)
FRL	.80 (.02)
NF_FRL	.75 (.02)
NF_GRD	.75 (.02)
RF	.79 (.03)
SVM	.62 (.06)
Logreg	.82 (.02)
Cart	.52 (.01)

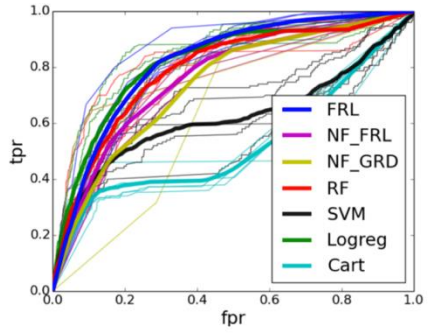


Figure 2: ROC curves for readmissions prediction.

Readmission Rule List

	Conditions		Probability	Support
IF	BedSores AND Noshow	THEN readmissions risk is:	33.25%	770
ELSE IF	PoorPrognosis AND MaxCare	THEN readmissions risk is:	28.42%	278
ELSE IF	PoorCondition AND Noshow	THEN readmissions risk is:	24.63%	337
ELSE IF	BedSores	THEN readmissions risk is:	19.81%	308
ELSE IF	NegativeIdeation AND Noshow	THEN readmissions risk is:	18.21%	291
ELSE IF	MaxCare	THEN readmissions risk is:	13.84%	477
ELSE IF	Noshow	THEN readmissions risk is:	6.00%	1127
ELSE IF	MoodProblems	THEN readmissions risk is:	4.45%	1325
ELSE		Readmissions risk is:	0.88%	3031

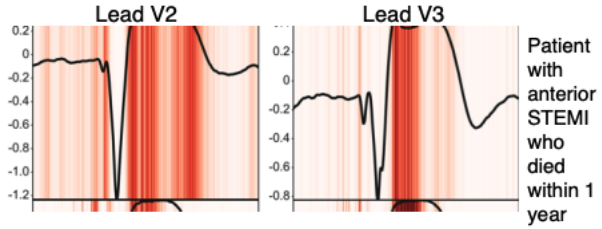
Table 2: Falling rule list for patients with no multiple readmissions history.

Local vs. Global interpretability

- **Local interpretability – understand predictions for individual data points (i.e., patients)**
  - Build trust in predictions; recognize errors due to model being poor, data point being an outlier, or engineering problems
  - Provide guidance to decision makers who may have additional information
  - **Explanations** that we described earlier, for Mycin, are an example of this

Local vs. Global interpretability

- **Local interpretability – understand predictions for individual data points (i.e., patients)**
  - Ex: **linear (bag of words) models**: look at highest weighted non-zero feature
  - Ex: **decision trees**: look at path to prediction for this patient
  - Ex: **deep models**: saliency maps and GradCAM



Gradient-CAM (Selvaraju et al., IJCV '19)

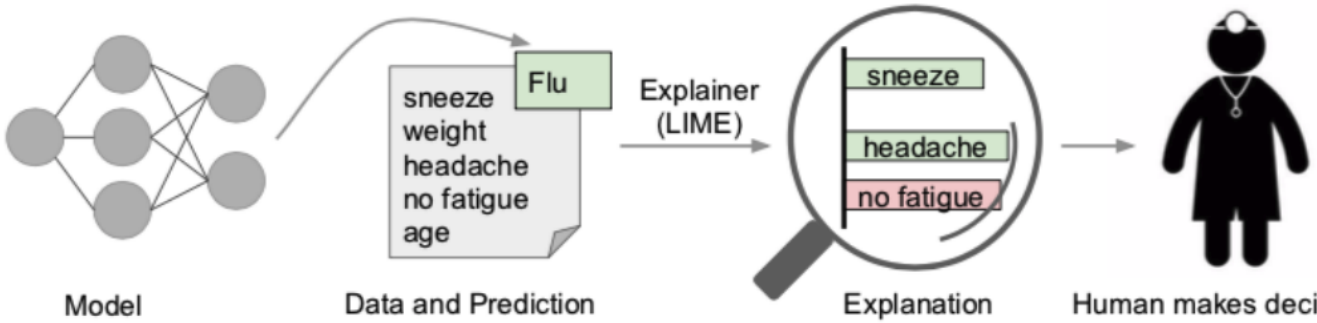
[Raghunath et al., Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network, Nature Medicine 2020]

- How can we do this more generally?

LIME slides developed from Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" (pp. 1135–1144). Presented at the the 22nd ACM SIGKDD International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2939672.2939778

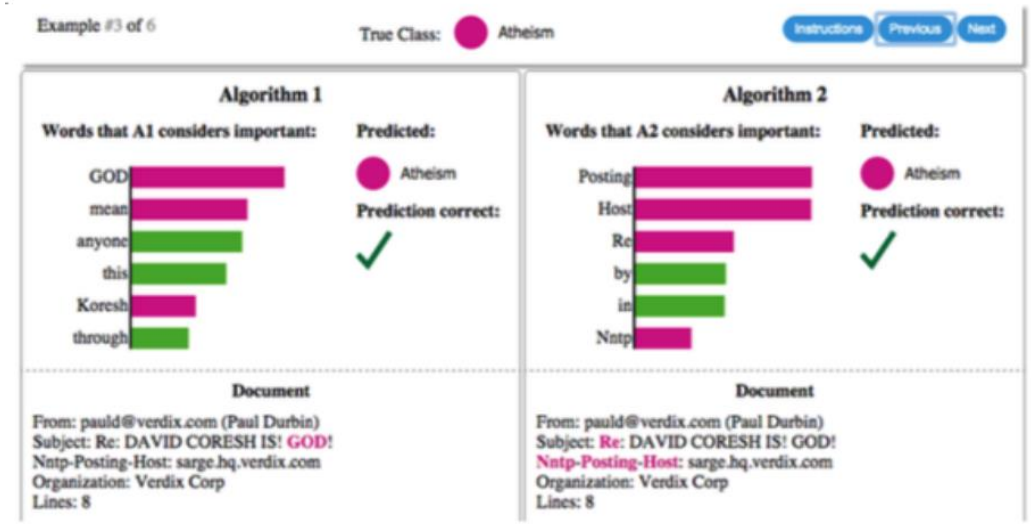
Model-agnostic Explanations

- A model predicts that a patient has the flu, and LIME highlights:
  - Sneeze and headache are portrayed as contributing to the "flu" prediction
  - "no fatigue" is evidence against it.
- With these, a doctor can make an informed decision about whether to trust the model's prediction.
- Approach helps detect data leakage, data set shift, using human expertise



Explanation of Cases may be Useful to Compare Models

- Predict whether a post is about “Christianity” or “Atheism”
- Algorithm 2 may be overall more accurate, but Algorithm 1 makes more sense, at least on this example.
- Again, relies on human expertise, which is much broader than any of our models



Desiderata for Explanations

- Interpretable — “provide qualitative understanding between the input variables and the response”
  - depends on audience
  - requires sparsity
  - features must make sense
    - e.g., eigenvectors in principal component analysis are not explainable features
- Local fidelity — “it must correspond to how the model behaves in the vicinity of the instance being predicted”
- Model-agnostic — “treat the original model as a black box”
  - *Is this really a good idea for all models?*

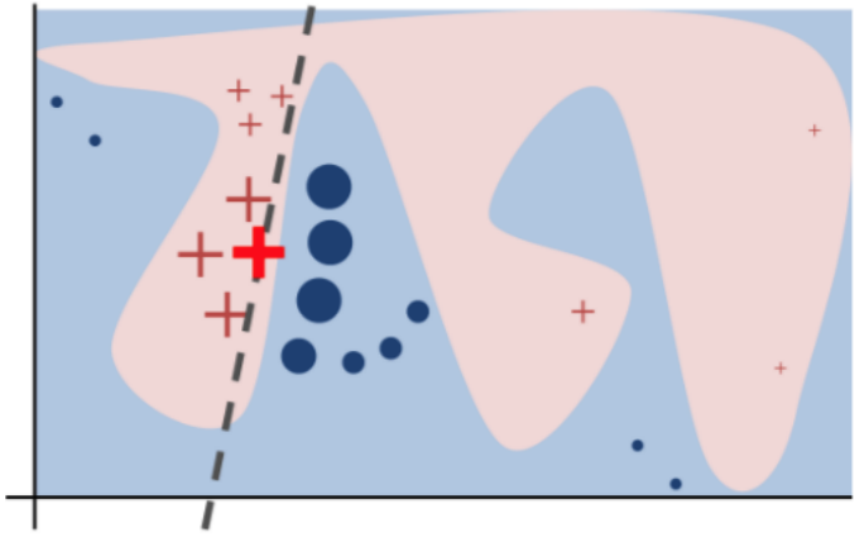
Sparse Linear Explanation

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

- Choose G to be the class of linear models such that $g(z') = w_g \cdot z'$
- Let $\pi_x(z) = \exp(-D(x,z)^2/\sigma^2)$ be an exponential kernel on some distance function *D* with width sigma $\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z)(f(z) - g(z'))^2$

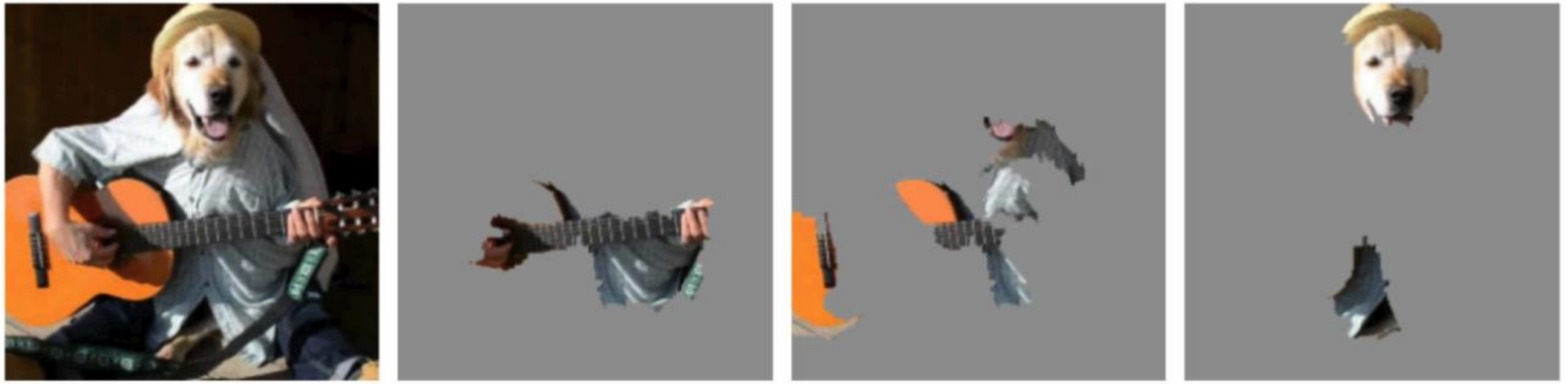


Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Apply to Image Interpretation

- Superpixel is a group of connected pixels with similar colors or gray levels
  - Image is segmented into super pixels
  - *K* is chosen as the number of superpixels to represent
- K-LASSO predicts label from superpixels, to select which *K* of them to use for explanation
- With N=5000, scikit-learn random forests with 1000 trees -> 3 sec
- Explaining Inception network results -> ~10 min

Apply to Image Interpretation



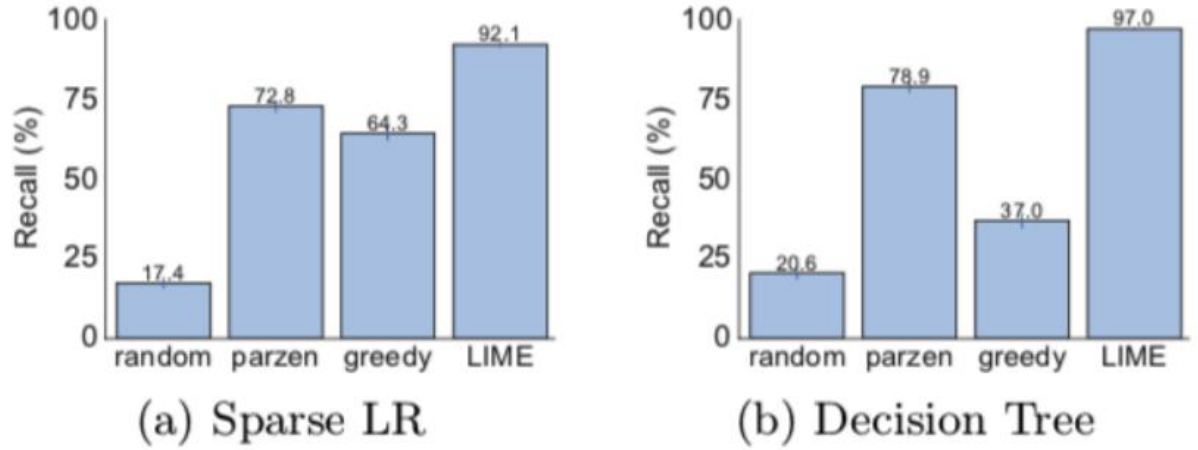
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

LIME Experiments

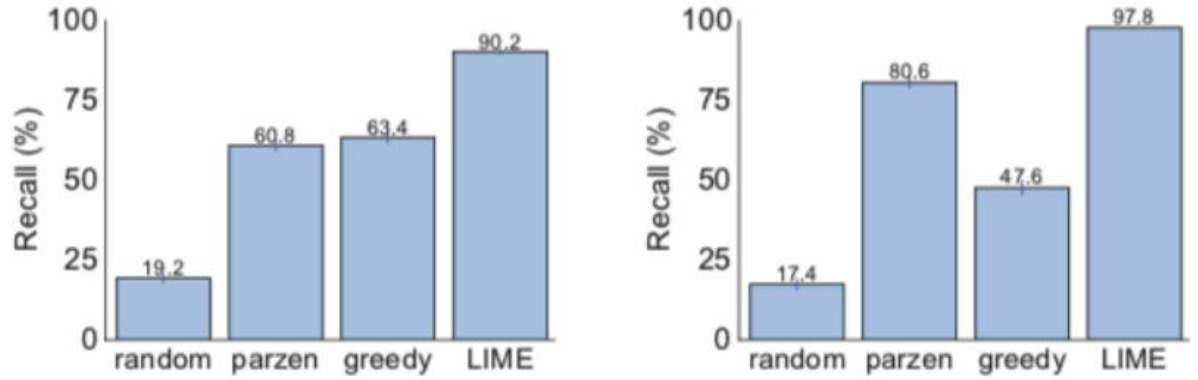
- Two sentiment analysis datasets (2000 instances, each; used 1600/400 test/train)
- Bag-of-words as features
- Models:
  - Decision Trees
  - Logistic Regression with L2 regularization
  - Nearest Neighbors
  - Support Vector Machines with RBF kernels
  - Random Forest (1000 trees) with word2vec embeddings
- K = 10

- **Random**: select features at random
- **Parzen**: features with largest global gradients
- **Greedy**: remove features contributing the most
- **LIME**: local interpretable model-agnostic explanations



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.



(a) Sparse LR (b) Decision Tree

Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Human Experiments

- Questions:
  - Can users choose which of two classifiers generalizes better
  - Based on the explanations, can users perform feature engineering to improve the model
  - Are users able to identify and describe classifier irregularities by looking at explanations
- “Christianity” vs. “Atheism” from 20-newsgroups dataset
  - known problems of data leakage from headers, …
  - trained original and “cleaned” classifiers for comparison
  - test set accuracy favors the “wrong” classifier!!!
- Separate test set of 819 web pages about these topics from http://dmoz-odp.org
- SVM with RBF kernels, trained on the 20-newsgroup data
- Mechanical Turk, 100 users, K=6 words, B=6 documents/Turk
  - in 2nd experiment, they are asked to remove word features they believe inappropriate

- Blue: Random pick
- Orange/Red: Submodular optimization using LIMe values

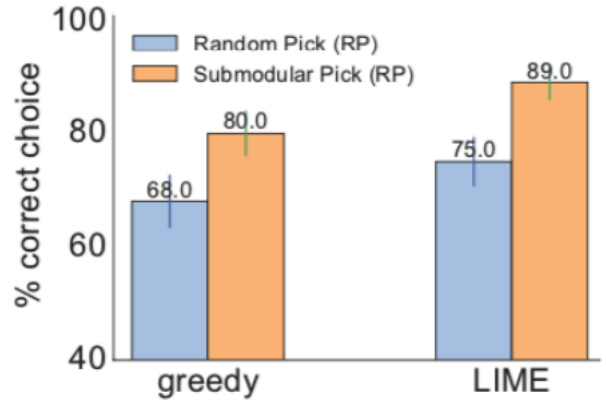


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

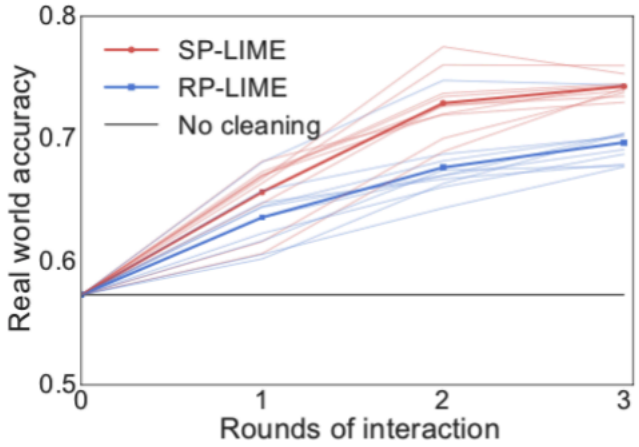


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

Can People Gain Insight from these Explanations?

- Trained a deliberately bad classifier between Wolf and Husky
  - All wolves in training set had snow in the picture, no huskies did
- Presented cases to graduate students with ML background
  - 10 balanced test predictions, with one husky in snow, one wolf not in snow
- Comparison between pre- and postexperiment trust and understanding



(a) Husky classified as wolf (b) Explanation

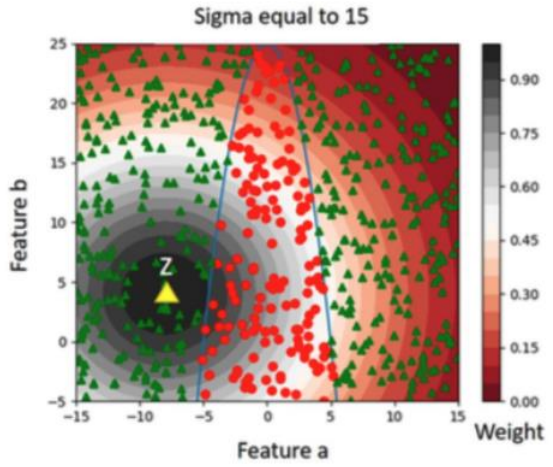
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

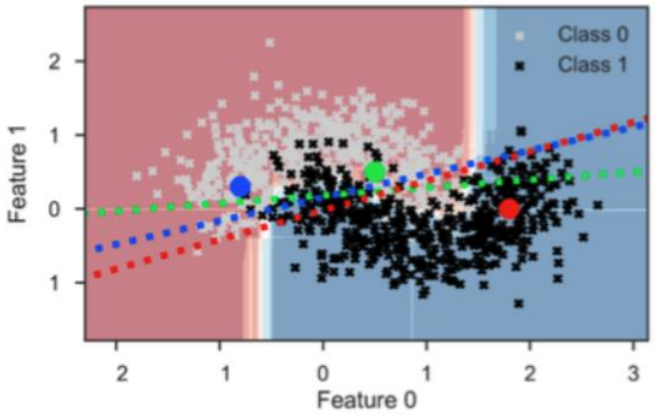
Table 2: "Husky vs Wolf" experiment results.

Critique of LIME

- Choice of σ is arbitrary and can lead to bad sampling
  - in implementation, often set to $0.75\sqrt{d}$
- It is important to tune the size of the neighborhood according to how far z is to the closest decision boundary



(a) A bad sampling scenario of LIME.



(b) Limitation of LIME spotted by Laugel et al. [14]

Adhikari, A., Tax, D. M. J., Satta, R., & Fath, M. (2018, December 21). Example and Feature importance-based Explanations for Black-box Machine Learning Models. arXiv.

Fooling LIME and SHAP

- Goal: Make a biased algorithm (you) while fooling post-hoc explanation methods like LIME and SHAP (adversary)

Slack et al, “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods” AIES 2020

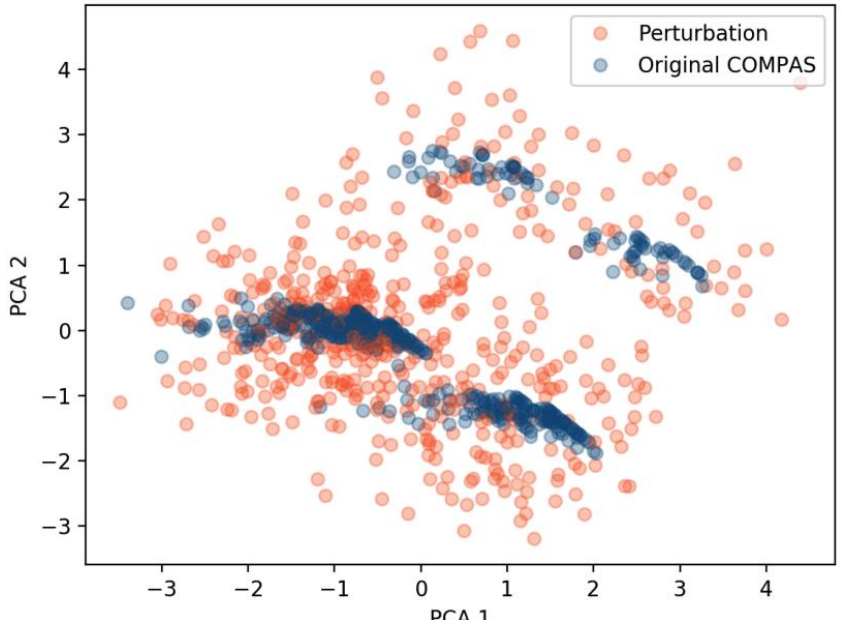
Fooling LIME and SHAP

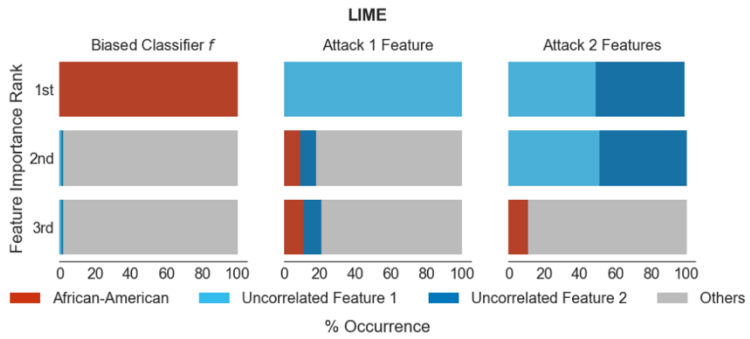
- Idea: Wrap the model in a protective scaffolding layer and control what the model puts out based on data input.
- If you can tell between real data input and SHAP explainability inputs, you can fool the interpretability method!

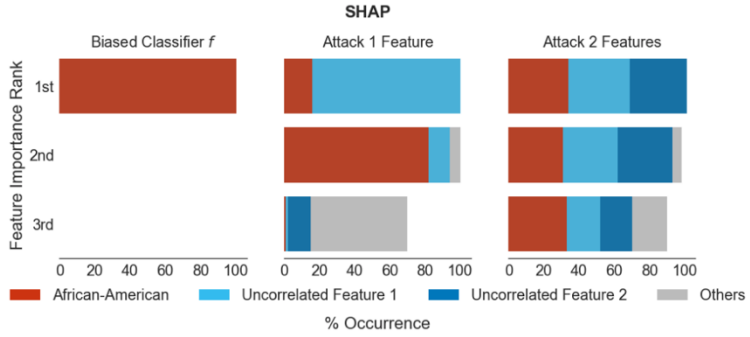
Slack et al, “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods” AIES 2020

Fooling LIME and SHAP

- Step 1: Train an out-of-distribution classifier to determine if incoming data point is from the actual data distribution or a synthetic data point
- Step 2: If data is real, use a real (biased) model. If data is perturbed, use an innocuous (fake) model
- LIME is particularly vulnerable, but SHAP was more resistant







Slack et al, “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods” AIES 2020

Counterfactual explanations

- Why did the treatment not work on the patient?
- Why was my loan rejected?
- Simplest approach:
  - Find the smallest change to the features that would change the prediction from rejected to approved
  - Note: (a) there may be many, (b) should be realistic

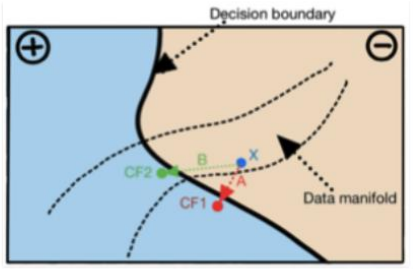


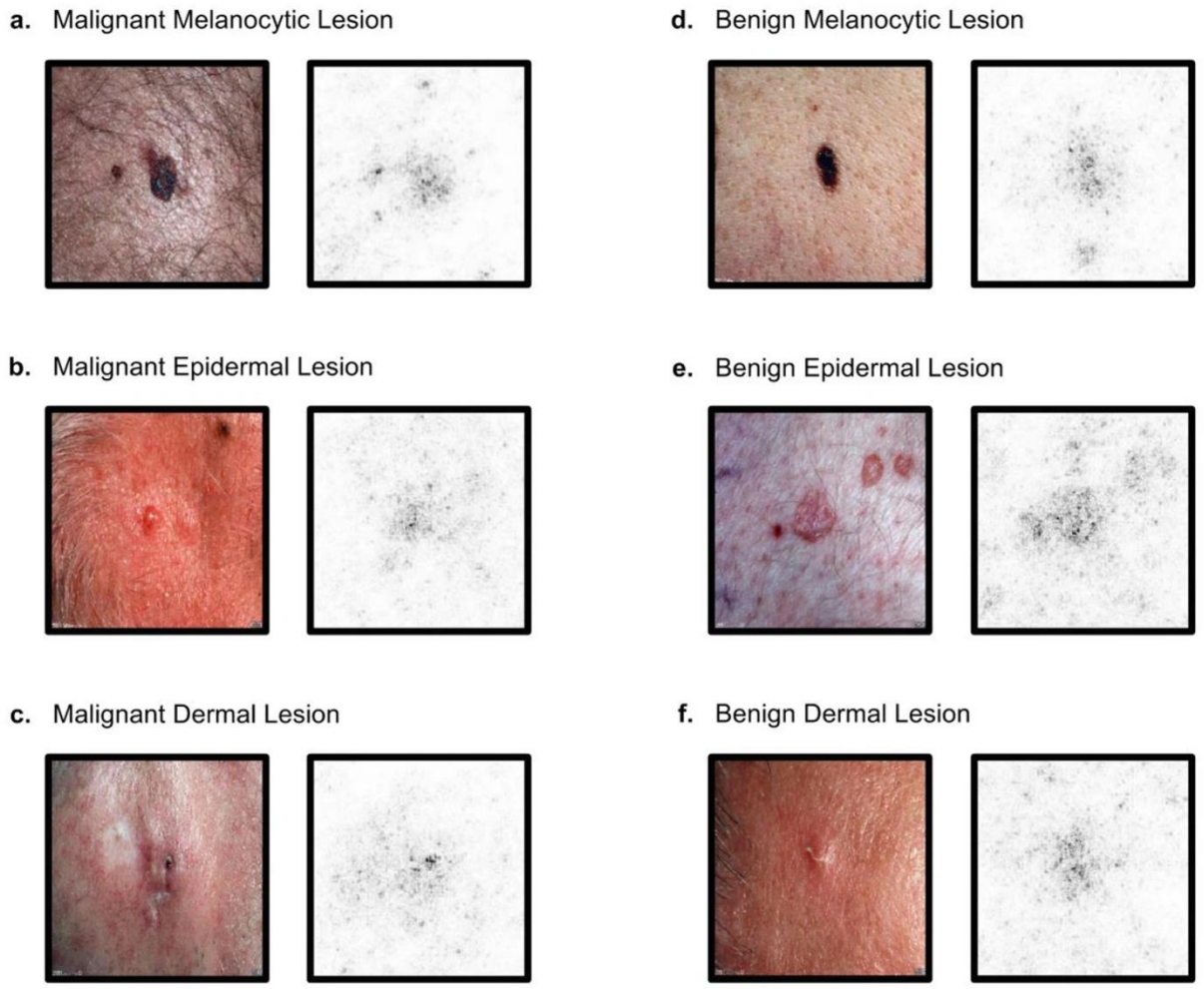
Figure 1: Two possible paths for a datapoint (shown in blue), originally classified in the negative class, to cross the decision boundary. The end points of both the paths (shown in red and green) are valid counterfactuals for the original point. Note that the red path is the shortest, whereas the green path adheres closely to the manifold of the training data, but is longer.

[Figure from: Verma et al., Counterfactual Explanations for Machine Learning: A Review, arXiv:2010.10596, 2020]

See also:
Karimi, Scholkopf, Valera. Algorithmic Recourse: from Counterfactual Explanations to Interventions. FAccT '21

Saliency maps as interpretability tools

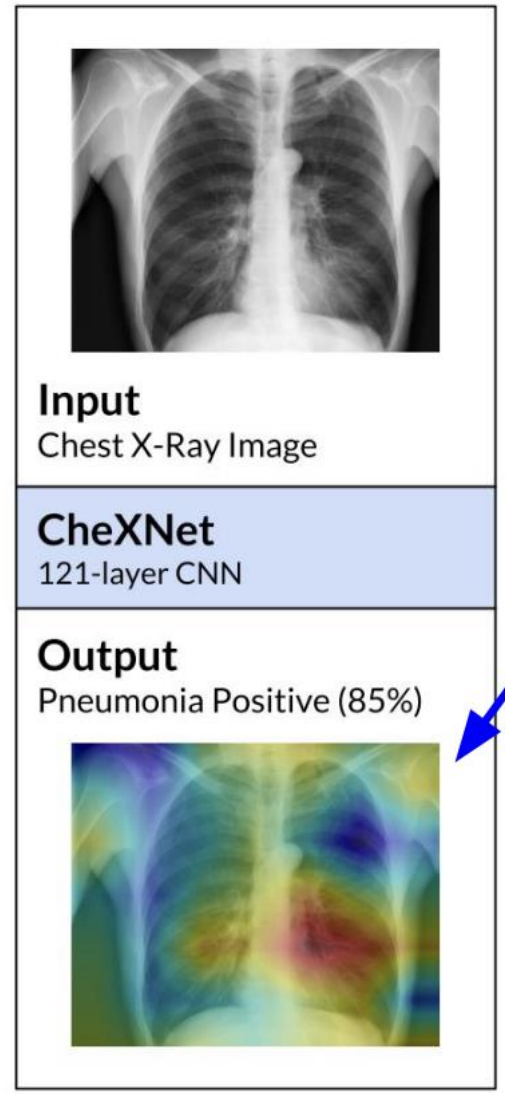
- Gradient-based explanations for neural networks
- Compute derivative with respect to the input
- Variants include GradCAM, which computes gradients with respect to the feature maps of a convolutional layer



Esteva et al, Dermatologist-level classification of skin cancer with deep neural networks, Nature 2017

Rajpurkar et al. 2017

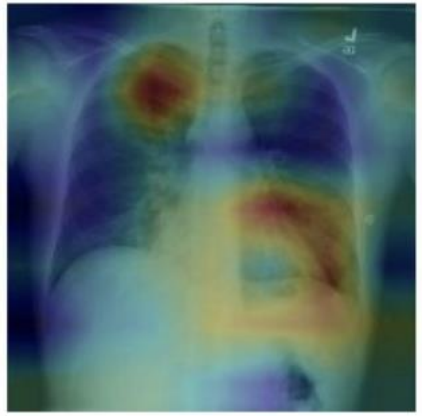
- Binary classification of pneumonia presence in chest X-rays
- Used ChestX-ray14 dataset with over 100,000 frontal X-ray images with 14 diseases
- 121-layer DenseNet CNN
- Compared algorithm performance with 4 radiologists
- Also applied algorithm to other diseases to surpass previous state-of-the-art on ChestX-ray14



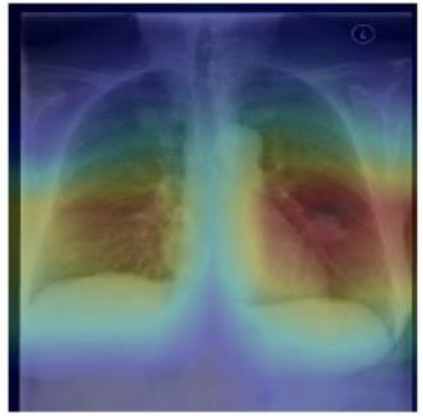
Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

Rajpurkar et al. 2017

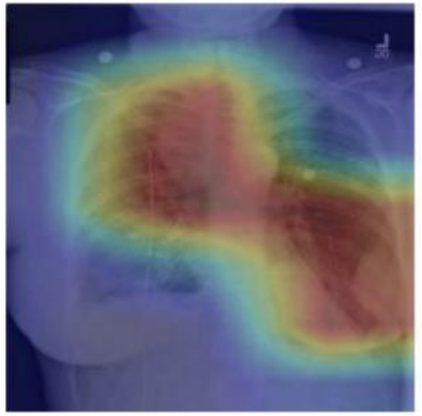
- Also showed CAM visualizations of predictions for other pathologies



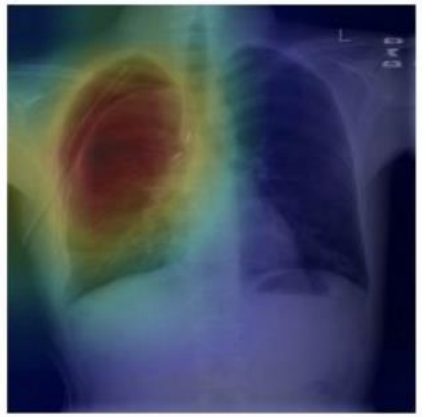
(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.



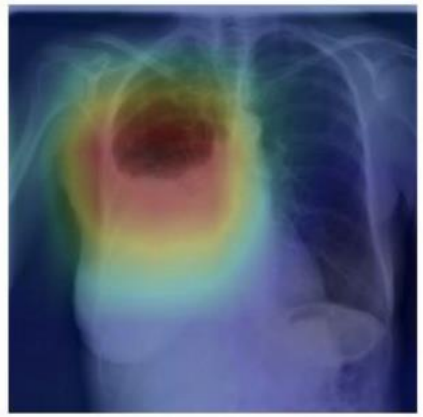
(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.



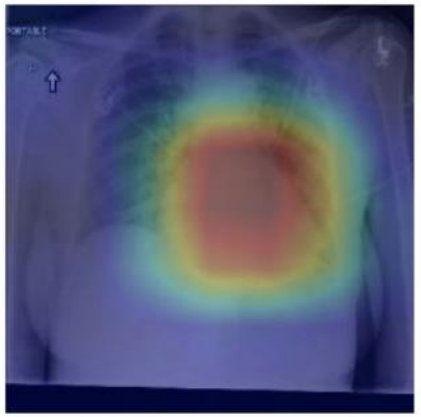
(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.



(d) Patient with a right-sided pneumothroax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).



(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.



(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

How to test if saliency maps are doing anything?

Saliency Map Sanity Checks

- Test 1: Compare saliency maps from trained models with saliency maps from models with **randomly initialized parameters**

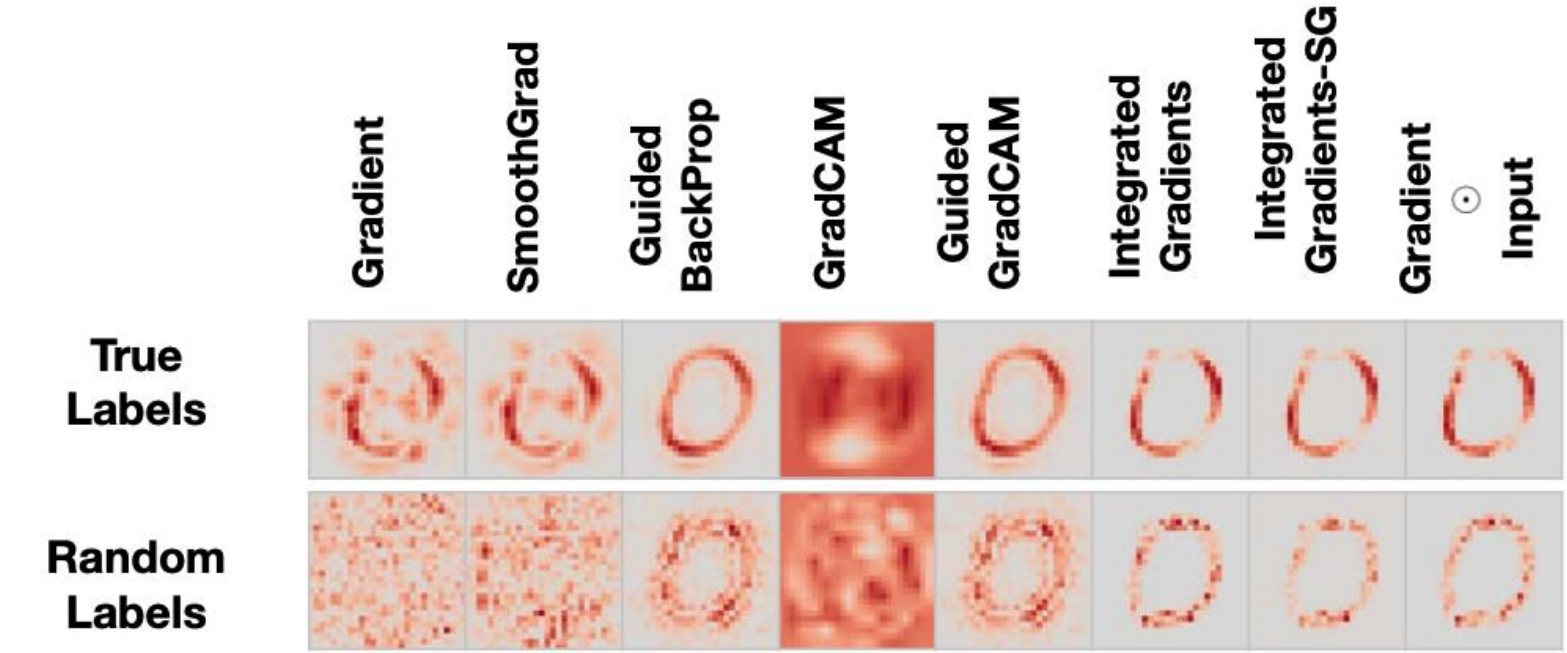
- Test 2: Compare saliency maps from trained models with saliency maps from models trained on data with **labels randomly flipped**

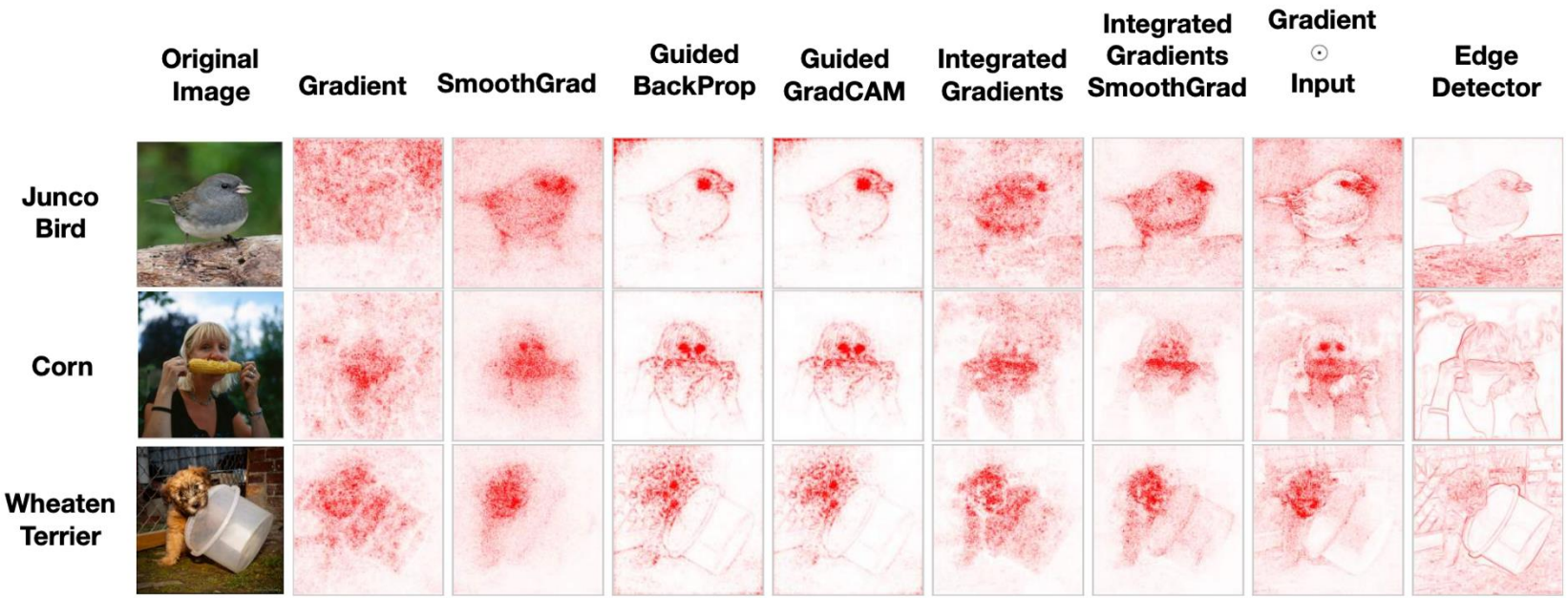
Adebayo et al, “Sanity Checks for Saliency Maps”, NeurIPS 2018

Option A

Option B

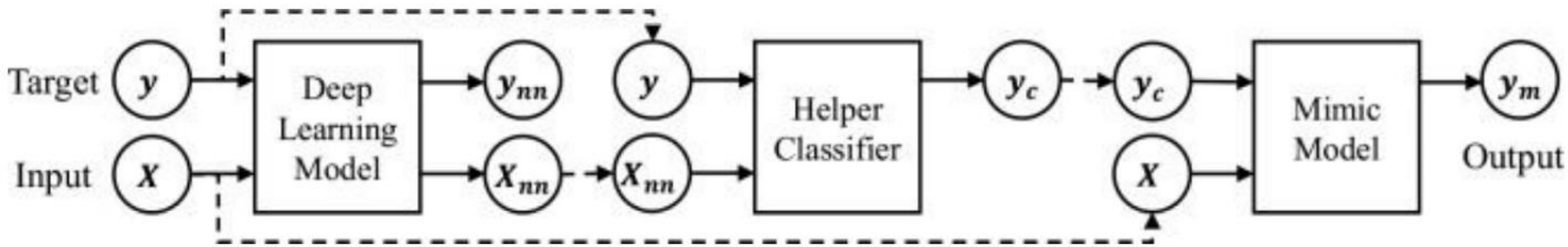
Gradient
SmoothGrad
Guided BackProp
GradCAM
Guided GradCAM
Integrated Gradients
Integrated Gradients-SG
Gradient ⊙ Input
True Labels
Random Labels


Gradient
SmoothGrad
Guided BackProp
GradCAM
Guided GradCAM
Integrated Gradients
Integrated Gradients-SG
Gradient ⊙ Input
True Labels
Random Labels


Original Image
Gradient
SmoothGrad
Guided BackProp
Guided GradCAM
Integrated Gradients
Integrated Gradients SmoothGrad
Gradient ⊙ Input
Edge Detector
Junco Bird
Corn
Wheaten Terrier

A different type of approach: “distill” a complex neural network to an interpretable decision tree

- Che et al. 2017: distill deep neural network for ICU outcome prediction into an interpretable gradient boosting trees model (called mimic model)
- Benefits of distillation: 1) DNN can learn to correct for errors and noise in training data; 2) classification probabilities from DNN give “soft labels” containing more information; 3) Mimic approach can also be seen as a regularization on more complex DNN



Che et al. Interpretable Deep Models for ICU Outcome Prediction, 2016.

A different type of approach: “distill” a complex neural network to an interpretable decision tree

Methods		MOR (Mortality)		VFD (Ventilator Free Days)	
		AUROC	AUPRC	AUROC	AUPRC
Baselines	SVM	0.6437 ± 0.024	0.3408 ± 0.034	0.7251 ± 0.023	0.7901 ± 0.019
	LR	0.6915 ± 0.027	0.3736 ± 0.038	0.7592 ± 0.021	0.8142 ± 0.019
	DT	0.6024 ± 0.013	0.4369 ± 0.016	0.5794 ± 0.022	0.7570 ± 0.012
	GBT	0.7196 ± 0.023	0.4171 ± 0.040	0.7528 ± 0.017	0.8037 ± 0.018
Deep Models	DNN	0.7266 ± 0.089	0.4117 ± 0.122	0.7752 ± 0.054	0.8341 ± 0.042
	GRU	0.7666 ± 0.063	0.4587 ± 0.104	0.7723 ± 0.053	0.8131 ± 0.058
	DNN + GRU	0.7813 ± 0.028	**0.4874 ± 0.051**	**0.7896 ± 0.019**	**0.8397 ± 0.018**
Best Mimic Model		**0.7898 ± 0.030**	0.4766 ± 0.050	**0.7889 ± 0.018**	0.8324 ± 0.016

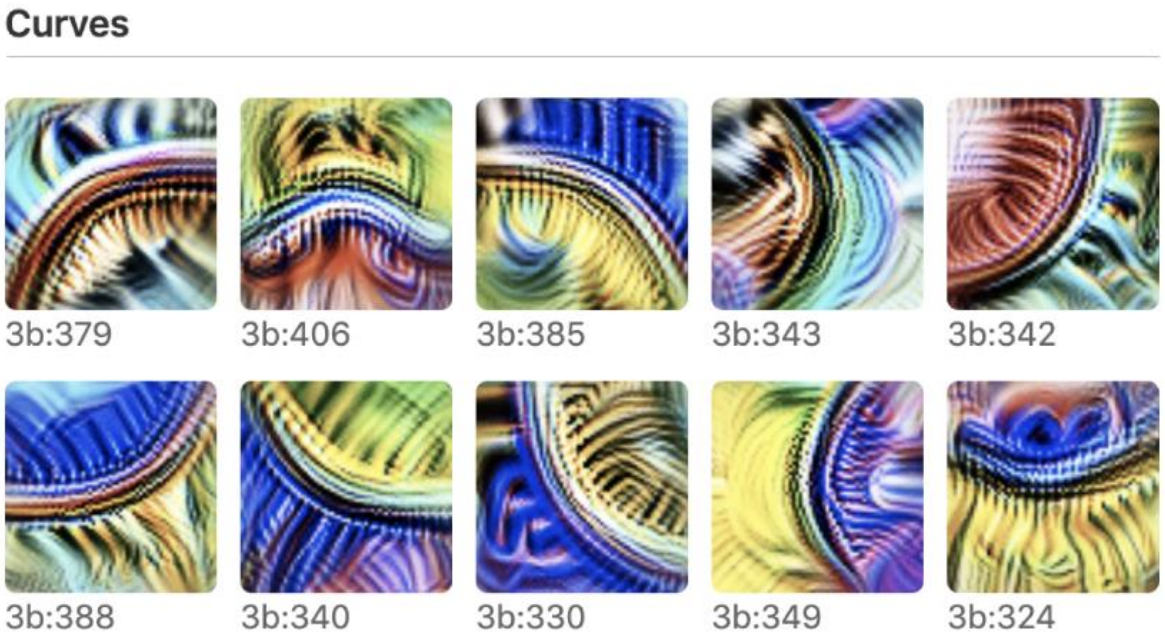
Che et al. Interpretable Deep Models for ICU Outcome Prediction, 2016.

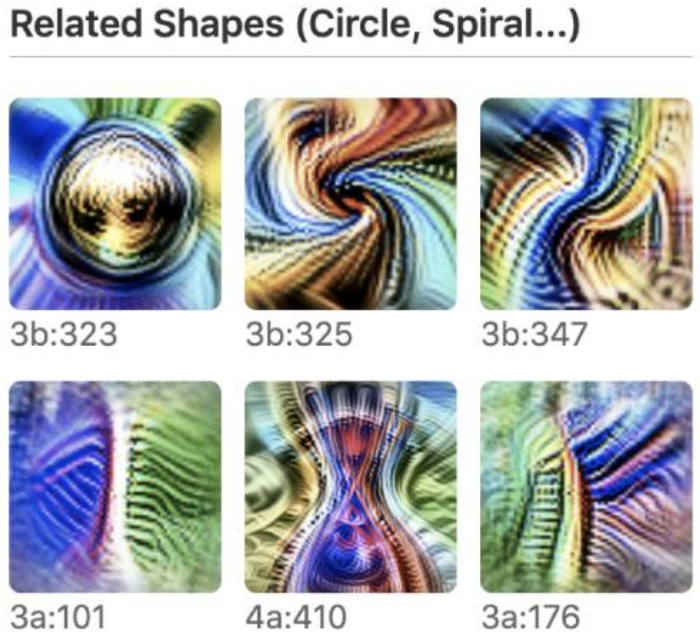
Mechanistic interpretability: Can we identify computational inside NNs?

- **Features** as defined as linear combinations of weights, which are the target of interest
  - Example: edge detectors
- **Circuits** are defined as the NN weights and linear combinations of the neurons
- **Intervention** relies on changing an internal component and measure effects

[Olah et al, “Zoom In: An Introduction to Circuits”, https://distill.pub/2020/circuits/zoom-in/]

Can we detect curves?





[Olah et al, “Zoom In: An Introduction to Circuits”, https://distill.pub/2020/circuits/zoom-in/]

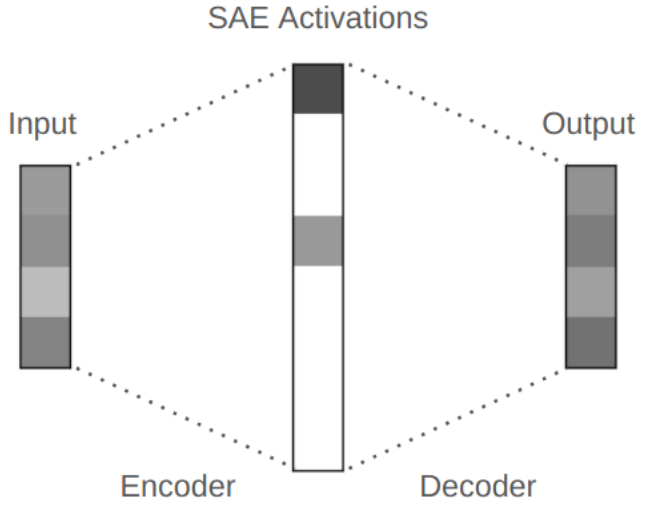
Can we detect curves?

1. **Feature visualization**: Generating images that use this combination of neurons does create images with curves
2. **Dataset examples:** Images that cause these neurons to fire do include curves
3. **Synthetic examples**: Generated curves (not straight lines or corners) cause neurons to fire
4. **Rotation**: Rotating one curve 90 degrees stops one neuron from firing and causes a related neuron to active, which together creates group of curve detectors
5. **Feature use:** Downstream circuits of neurons and edges seem to be curve-related: detecting circles, spirals, etc.

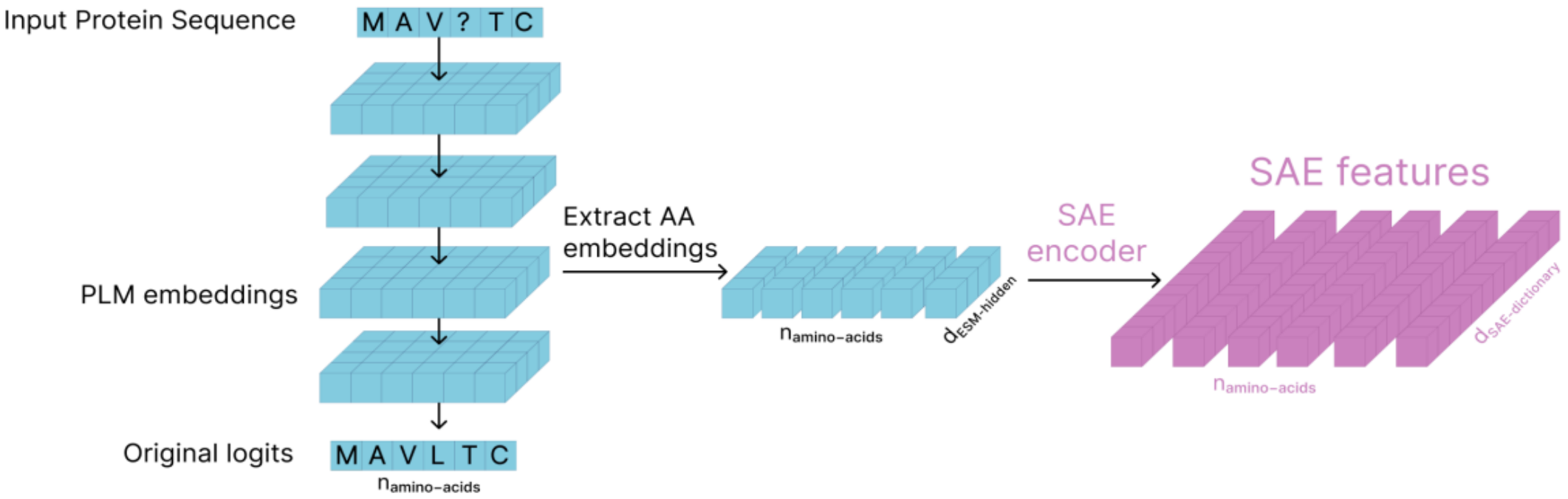
[Olah et al, “Zoom In: An Introduction to Circuits”, https://distill.pub/2020/circuits/zoom-in/]

Superposition motivates sparse autoencoders

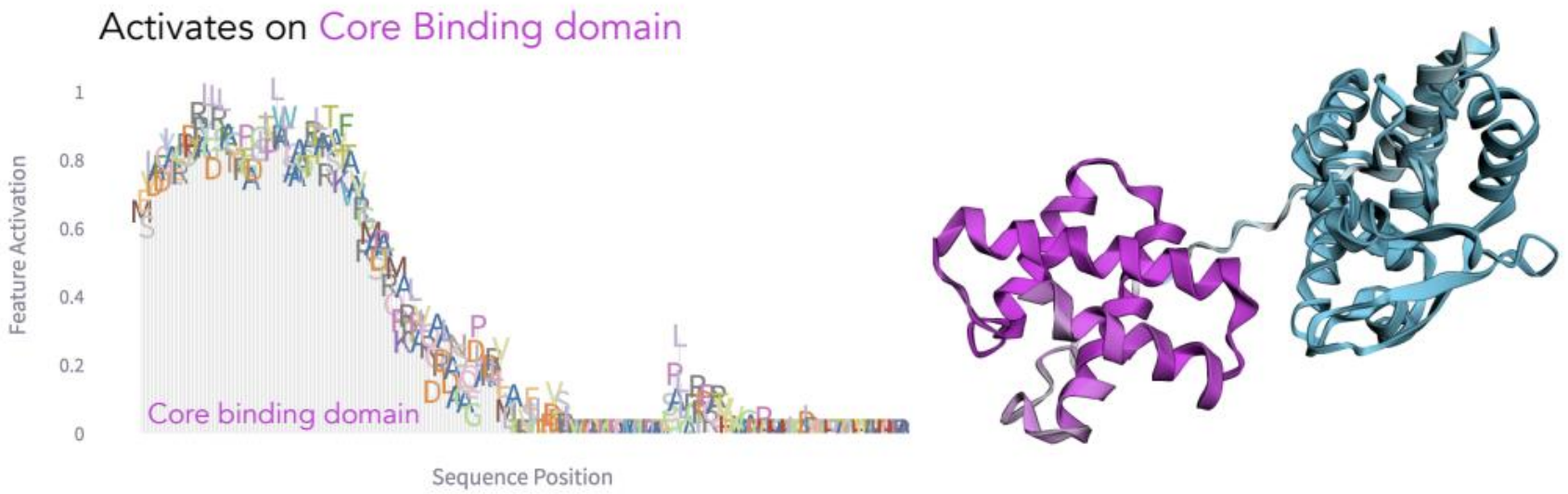
- Polysemantic neurons may fire and control multiple “human-perspective” features
  - Example: one neuron controls cat faces, fronts of cars, and cat legs
- A neuron can respond to multiple concepts, so features may occupy directions in activation space.



Example: Protein Models



Source: Simon and Zou, “InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders”





Source: Simon and Zou, “InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders”

Tests for mechanistic interpretability

1. **Fidelity**: Does the replacement reproduce model behavior?
2. **Semantic validity**: Does the feature consistently represent the claimed concept?
3. **Stability**: Does a similar feature recur across seeds and methods?
4. **Causal effect**: Does a controlled intervention change behavior as predicted?

Recent advances in LLM interpretability

1. **Explain after training:** SAEs and attribution graphs recover approximations of a dense network's computation.

2. **Train for interpretable circuits:** Weight-sparse networks seek simpler circuits from the outset, with capability tradeoffs.

Mitchell 2019: Model cards for Model Reporting

- Documentation accompanying trained models to detail performance characteristics

Model Card - Smiling Detection in Images

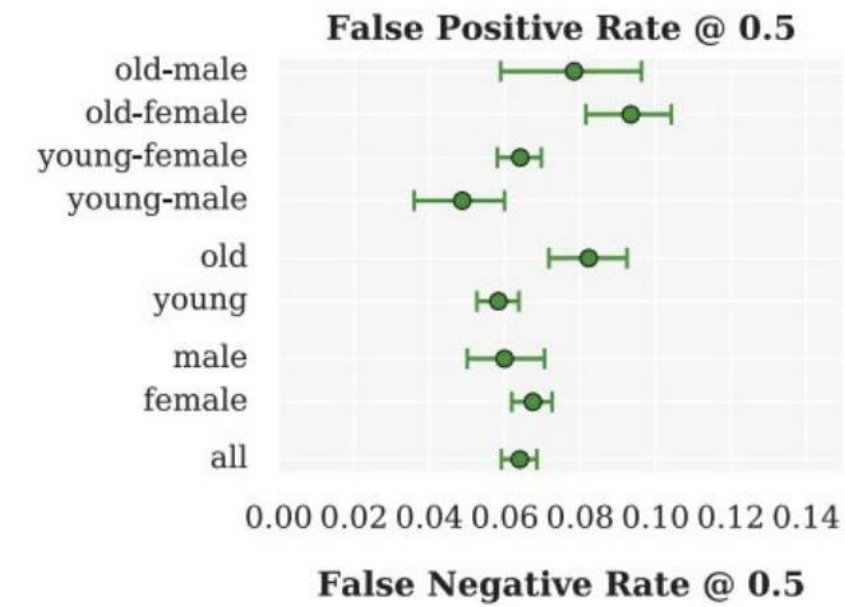
Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Quantitative Analyses



Mitchell et al. Model Cards for Model Reporting, 2019.

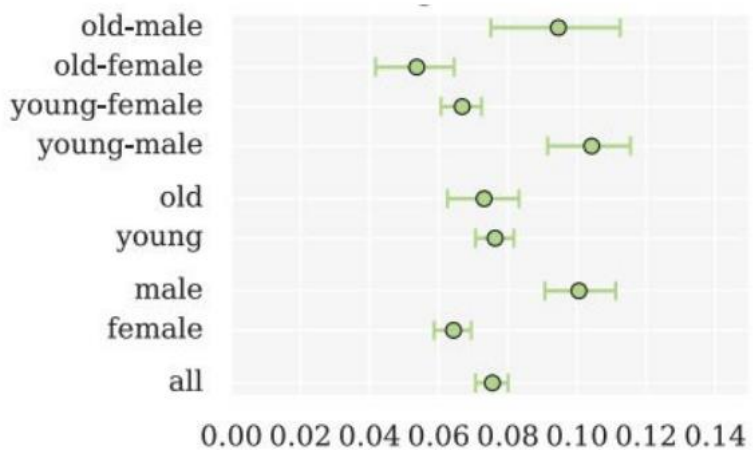
Mitchell 2019: Model cards for Model Reporting

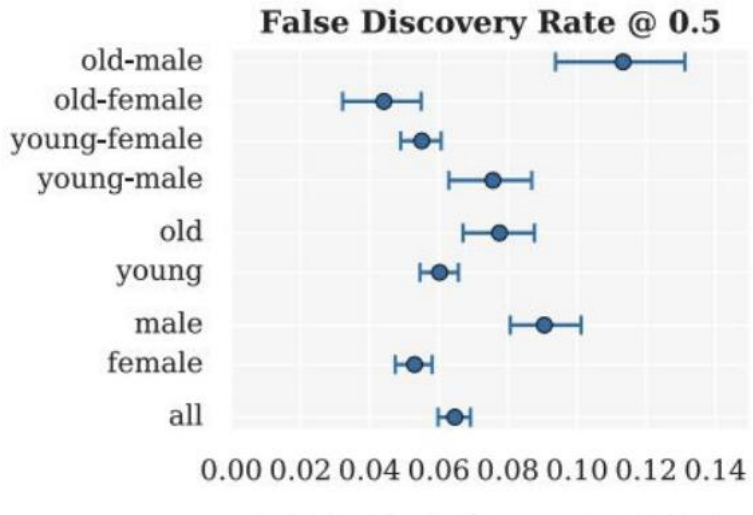
Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics corre-





Mitchell et al. Model Cards for Model Reporting, 2019.

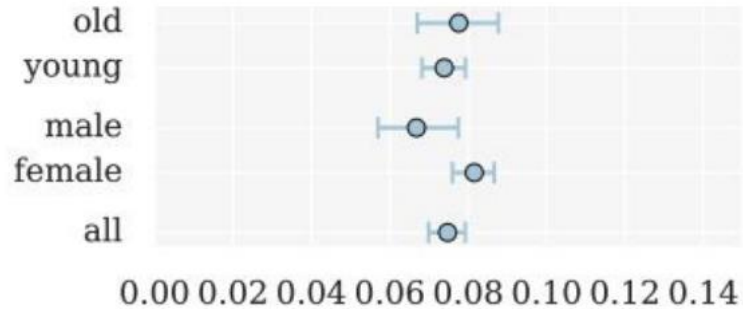
Mitchell 2019: Model cards for Model Reporting

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.



Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Mitchell et al. Model Cards for Model Reporting, 2019.

Gebru 2020: Datasheets for Datasets

A Database for Studying Face Recognition in Unconstrained Environments **Labeled Faces in the Wild**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Gebru et al. Datasheets for Datasets. 2020.

Summary

- **Local methods**
- **Global methods**
- **Mechanistic methods**



How can we make Data 146 better for you?

Next class: Data quality