Machine Learning for Healthcare

6.871, HST.956

Lecture 18: Interpretability

David Sontag

CSAIL

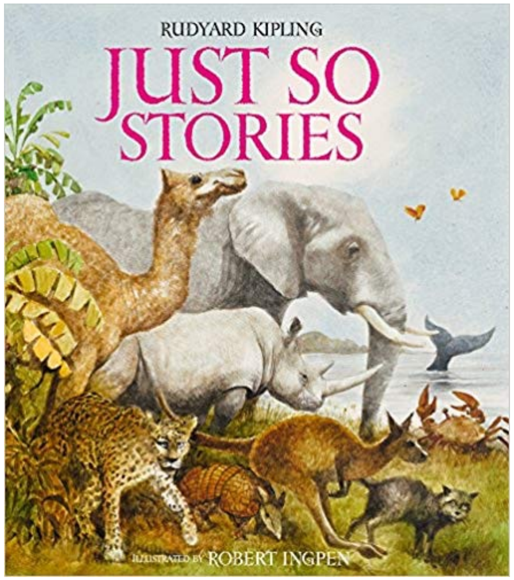
INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

Acknowledgement: Slides adapted from Peter Szolovits

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."
    - "… the number of chunks of information is constant for immediate memory. The span of immediate memory seems to be almost independent of the number of bits per chunk …"
  - Not surprising that one cannot "keep in mind" complex models
- What leads to complex models? And what to do about it?
  - Overfitting
    - Restrict model complexity; e.g., regularization
  - True complexity
    - Make up "just-so" stories that give a simplified explanation of how the complex model applies to specific cases
    - Trade off lower performance for simplicity of model



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
        - Clinical decision support
    - Confidence in model
        - Population health

- Recall what we've discussed of randomized controlled trials
    - Simplest cases (no comorbidities), smallest sample needed for significance test, shortest follow-up time
    - Results applied to very different populations

- Same concerns for ML models
    - Train and test samples often drawn from same population
    - Are results applicable elsewhere?

Explanation — Not a New Idea!
Mycin, 1975

- 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)

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

How Mycin Works

- 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

Explanation — Not a New Idea!
Mycin, 1975

- In light of the site from which the culture was obtained, and the method of collection, do you feel that a significant number of ORGANISM-1 were obtained?
  ****WHY**

- [1.0] It is important to find out whether there is therapeutically significant disease associated with this occurrence of ORGANISM-1
  It has already been established that:
  - [1.1] the site of the culture is not one of those which are normally sterile, and
  - [1.2] the method of collection is sterile
- Therefore, if:
  - [1.3] the organism has been observed in significant numbers
- Then: there is strongly suggestive evidence (.9) that there is therapeutically significant disease associated with this occurrence of the organism
- [Also : there is strongly suggestive evidence (.8) that the organism is not a contaminant]

Shortliffe, E. H., Davis, R., Axline, S. G., & Buchanan, B. G. (1975). Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, *8*, 303–320.

- **WHY
- [2.0] . . . in order to find out a locus of infection in PATIENT-1
  It has already been established that:
  - [2.1] the site of the culture is known, and
  - [2.2] the number of days since the specimen for the culture was obtained is less than 7
- Therefore, if:
  - [2.3] there is therapeutically significant disease associated with this occurrence of the organism
- Then: it is definite (1.0) that the site of the culture is a locus of infection in the patient

- ****WHY** [i.e., WHY is it important to find out a locus of infection in PATIENT-1?]
- [3.0] . . . in order to find out about an organism or class of organisms for which therapy should cover
  It has already been established that:
    - [3.1] the site of the culture is blood, and
    - [3.2] the stain of the organism is grampos, and
    - [3.3] the portal of entry of the organism is GI
- Therefore, if:
    - [3.4] [A- abdomen is a locus of infection in the patient, or B- pelvis is a locus of infection in the patient]
- Then : there is strongly suggestive evidence (.9) that enterobacteriaceae is a class of organisms for which therapy should cover

- ** **HOW 3.0** [i.e., HOW were you finding out about an organism or class of organisms for which therapy should cover?]
- [4.0] At that point RULE095 was being used.
  In addition :
  - [4.1] RULE162 was used to conclude that bacteroides is an organism for which therapy should cover, and
  - [4.2] RULE115 will be tried later, as well.

- **** HOW 4.1** [i.e., HOW was RULE162 used?]
- [5.0] We have already established that:
  - [5.1] the site of the culture is blood, and
  - [5.2] the portal of entry of the organism is GI, and
  - [5.3] the patient is a compromised host
- Therefore : it is definite (1.0) that bacteroides is an organism for which therapy should cover (RULE162)

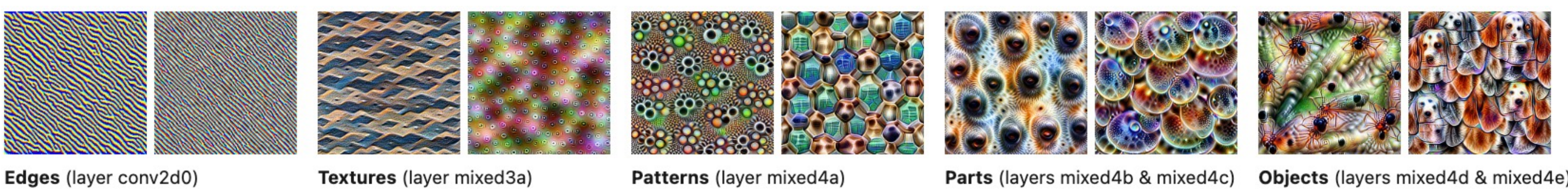
- **** DO YOU EVER PRESCRIBE CARBENICILLIN FOR PSEUDOMONAS INFECTIONS?**
- RELEVANT RULES: RULE064 RULE119 RULE120
  WHICH ONES DO YOU WISH TO SEE?
  ** RULE064
  IF: 1) IT IS KNOWN WITH CERTAINTY THAT THE IDENTITY OF THE ORGANISM IS PSEUDOMONAS, AND
  2) THE DRUG UNDER CONSIDERATION IS GENTAMICIN
- THEN: RECORD THE FOLLOWING AS A MORE APPROPRIATE THERAPY: GENTAMICIN-AND-CARBENICILLIN

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…*

Example of using global interpretability to debug ML setup

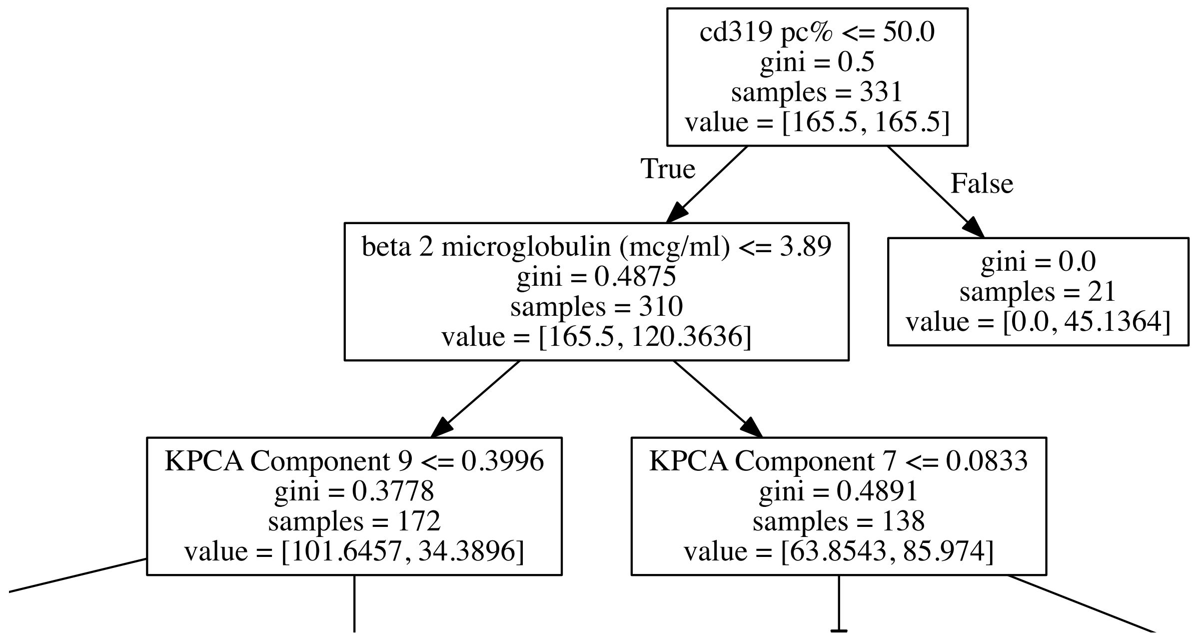
- In 2018, I submitted a paper using the Multiple Myeloma Research Foundation's IA9 data release. *Great results*

Table 3: **Predicting Mortality**

Method	1 Yr Full	1 Yr ISS-FISH	2 Yr Full	2 Yr ISS-FISH
LR	0.66 ± 0.1	$\boxed{0.62 \pm 0.14}$	0.8 ± 0.08	$\boxed{0.69 \pm 0.1}$
LR-B-PCA	0.66 ± 0.1	0.61 ± 0.13	0.79 ± 0.08	0.65 ± 0.11
LR-T-PCA	0.68 ± 0.1	0.61 ± 0.14	0.8 ± 0.08	0.65 ± 0.11
RF	0.65 ± 0.09	0.63 ± 0.12	0.82 ± 0.08	0.73 ± 0.09
RF-B-PCA	0.69 ± 0.11	0.63 ± 0.12	0.83 ± 0.08	0.73 ± 0.09
RF-T-PCA	$\boxed{0.72 \pm 0.1}$	0.64 ± 0.12	$\boxed{0.85 \pm 0.08}$	0.72 ± 0.09

Example of using global interpretability to debug ML setup

- Curious to see why “full” feature set with random forests so much better, so looked at one decision tree:



- Surprised to see cd319% at the top, but after discussing with clinical collaborator, concluded it is reasonable

Example of using global interpretability to debug ML setup

- 3 months later, new release of data (IA11) is available and I ask students to reproduce results

Big differences!

Old results (IA9):

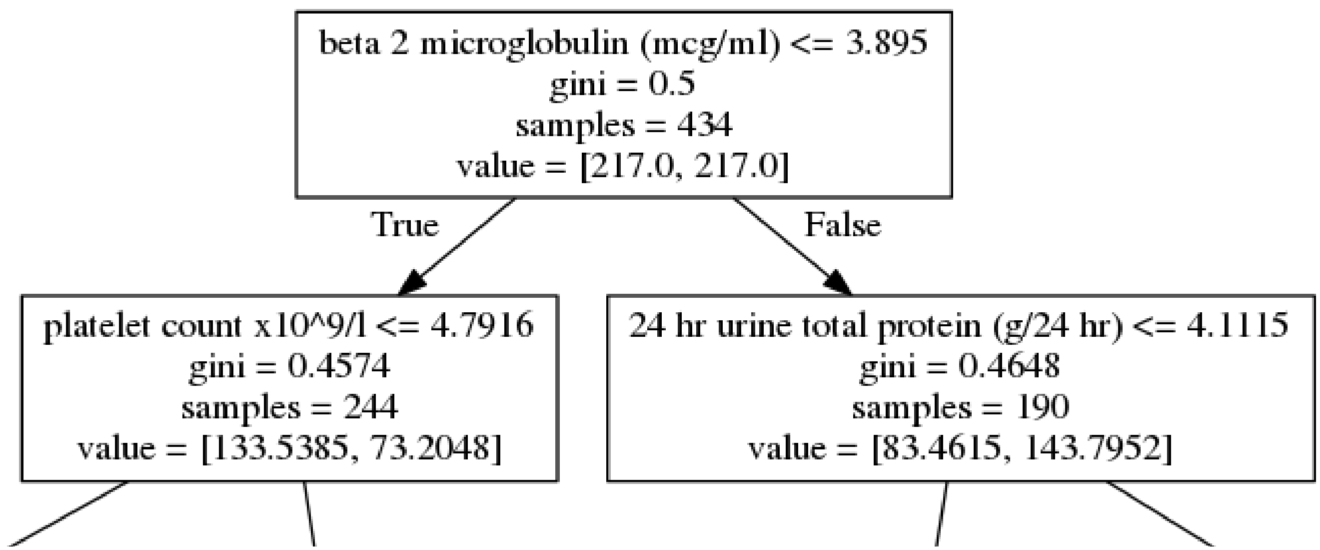
Method	1 Yr Full	1 Yr ISS-FISH	2 Yr Full	2 Yr ISS-FISH
LR	0.66 ± 0.1	0.62 ± 0.14	0.8 ± 0.08	0.69 ± 0.1
LR-B-PCA	0.66 ± 0.1	0.61 ± 0.13	0.79 ± 0.08	0.65 ± 0.11
LR-T-PCA	0.68 ± 0.1	0.61 ± 0.14	0.8 ± 0.08	0.65 ± 0.11
RF	0.65 ± 0.09	0.63 ± 0.12	0.82 ± 0.08	0.73 ± 0.09
RF-B-PCA	0.69 ± 0.11	0.63 ± 0.12	0.83 ± 0.08	0.73 ± 0.09
RF-T-PCA	0.72 ± 0.1	0.64 ± 0.12	0.85 ± 0.08	0.72 ± 0.09

New results (IA11):

Models	1 Yr Full	1 Yr ISS-FISH	2 Yr Full	2 Yr ISS-FISH
LR	0.68 ± 0.09	0.65 ± 0.14	0.76 ± 0.08	0.7 ± 0.09
LR-B-PCA	0.68 ± 0.1	0.65 ± 0.13	0.75 ± 0.08	0.67 ± 0.09
LR-T-PCA	0.69 ± 0.09	0.64 ± 0.13	0.77 ± 0.07	0.66 ± 0.09
RF	0.63 ± 0.1	0.63 ± 0.11	0.75 ± 0.08	0.73 ± 0.08
RF-B-PCA	0.66 ± 0.1	0.64 ± 0.11	0.76 ± 0.08	0.72 ± 0.08
RF-T-PCA	0.78 ± 0.08	0.64 ± 0.11	0.77 ± 0.08	0.72 ± 0.08

Example of using global interpretability to debug ML setup

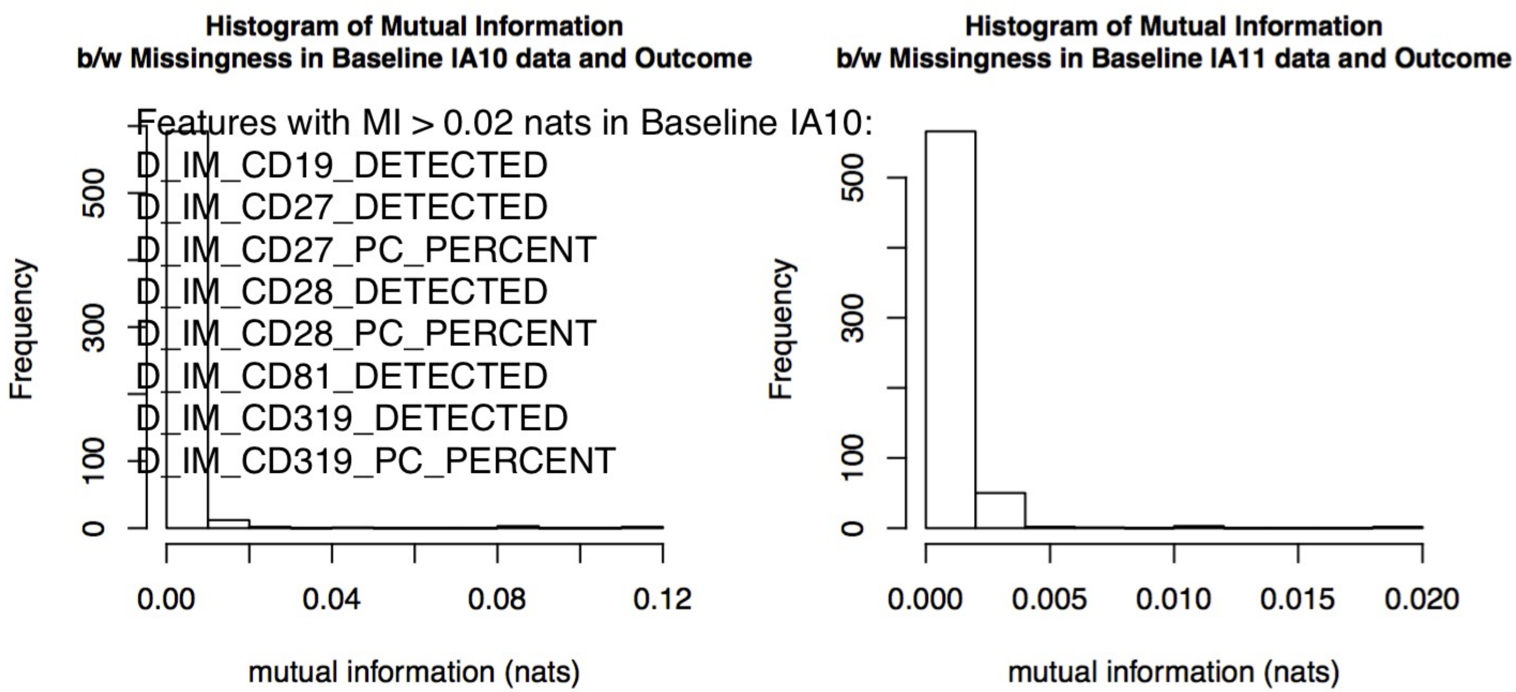
- 3 months later, new release of data (IA11) is available and I ask students to reproduce results



- Cd319% no longer shows up as a top predictor!
- What happened!?

Example of using global interpretability to debug ML setup

- After digging deeper, we realized that what was predictive originally was the feature Cd319% being missing, and moreover that this was correlated with the outcome (i.e. label leakage!)



[Figure credit: Rebecca Boiarsky]

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

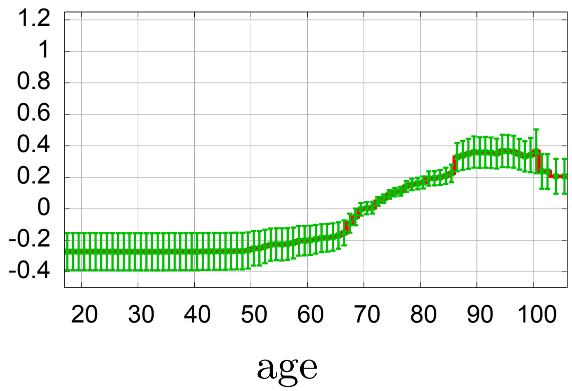
Generalized additive models (GAMs)

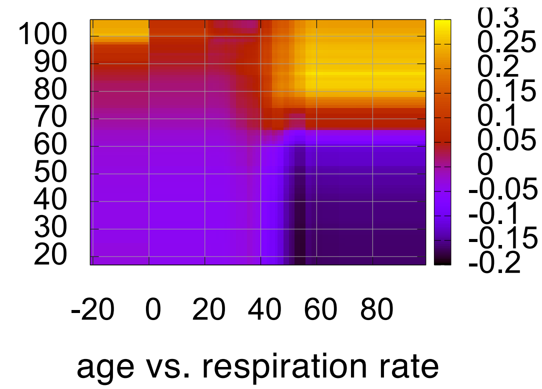
- GAMs with pairwise interactions have the form:

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

- *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
EI		EN malignancy risk is	63.40%	153
EI		EN malignancy risk is	39.68%	63
EI		EN malignancy risk is	26.09%	46
EI		EN malignancy risk is	10.38%	366

r mammographic mass dataset.

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)

Table 3: AUROC values for readmission data

[Wang & Rudin, AISTATS ‘15]

Supersparse linear integer models

- Learn **linear** model where:
  1. Coefficients are all integer
  2. As sparse as possible

Training objective:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_i \boldsymbol{\lambda}^T \boldsymbol{x}_i \leq 0\right] + C_0 \|\boldsymbol{\lambda}\|_0 + \epsilon \|\boldsymbol{\lambda}\|_1 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L}. \end{aligned}$$

PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1.	*age* ≥ 60	4 points	$\cdots\cdots$
2.	*hypertension*	4 points	$+ \quad \cdots\cdots$
3.	*body mass index* ≥ 30	2 points	$+ \quad \cdots\cdots$
4.	*body mass index* ≥ 40	2 points	$+ \quad \cdots\cdots$
5.	*female*	-6 points	$+ \quad \cdots\cdots$
	ADD POINTS FROM ROWS 1 – 5	**SCORE**	$= \quad \cdots\cdots$

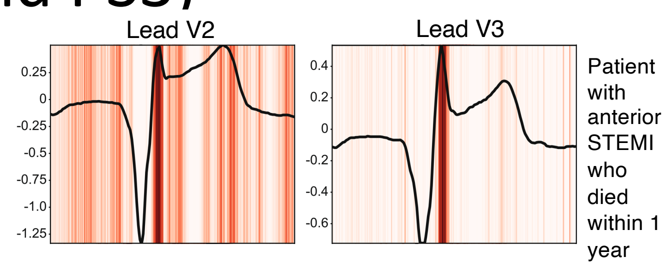
[Ustun & Rudin, ML ‘16]

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 (as in lectures 5 & 8, and PS3)

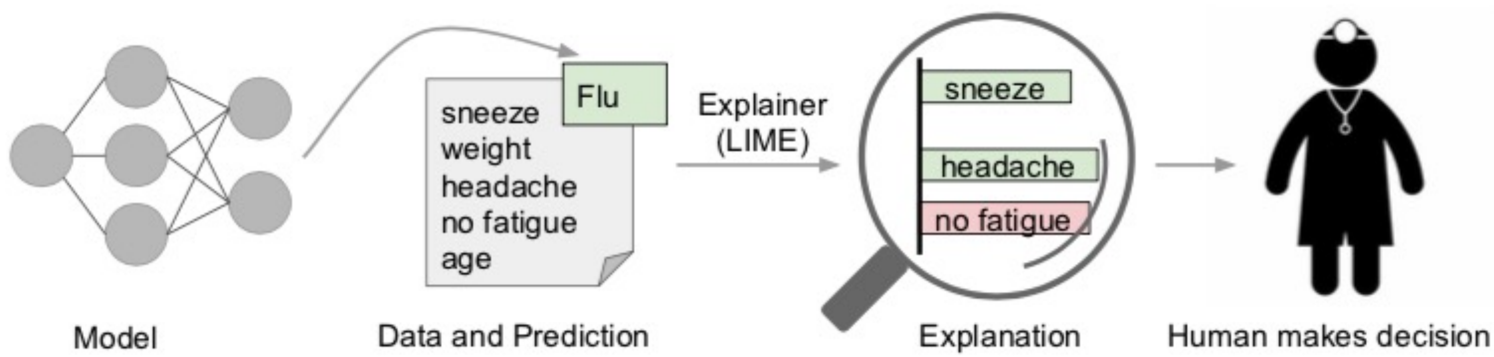


Gradient-CAM (Selvaraju et al., IJCV '19)

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

  - *How can we do this more generally?*

Model-agnostic Explanations

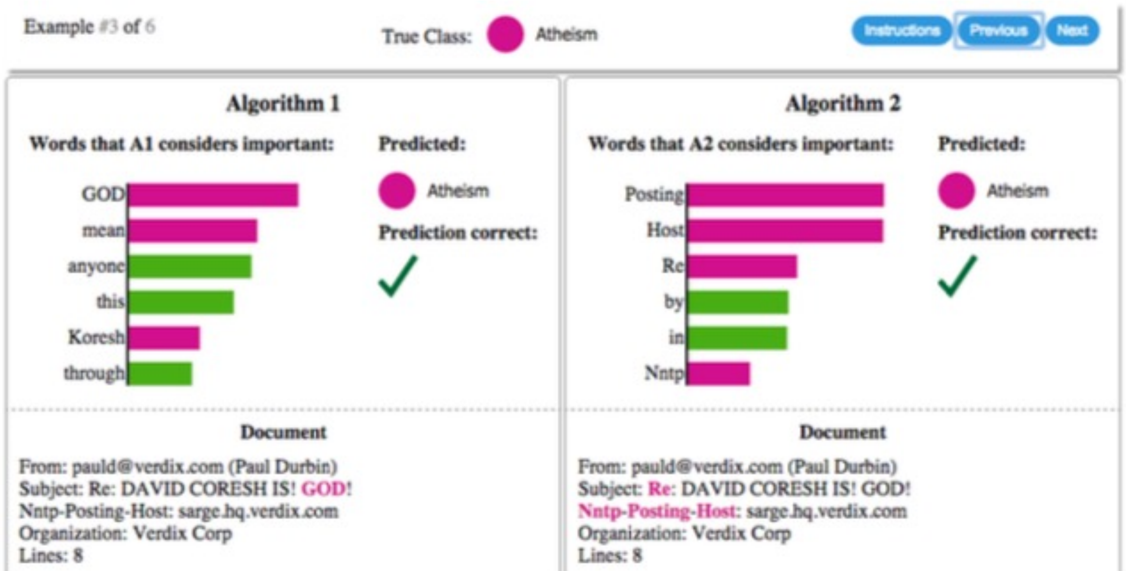


- A model predicts that a patient has the flu, and LIME highlights:
  - Sneeze and headache are portrayed as contributing to the “flu” prediction
  - “no fatigue” is evidence against it.
- With these, a doctor can make an informed decision about whether to trust the model’s prediction.

- *Approach helps detect data leakage, data set shift, using human expertise*

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

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.

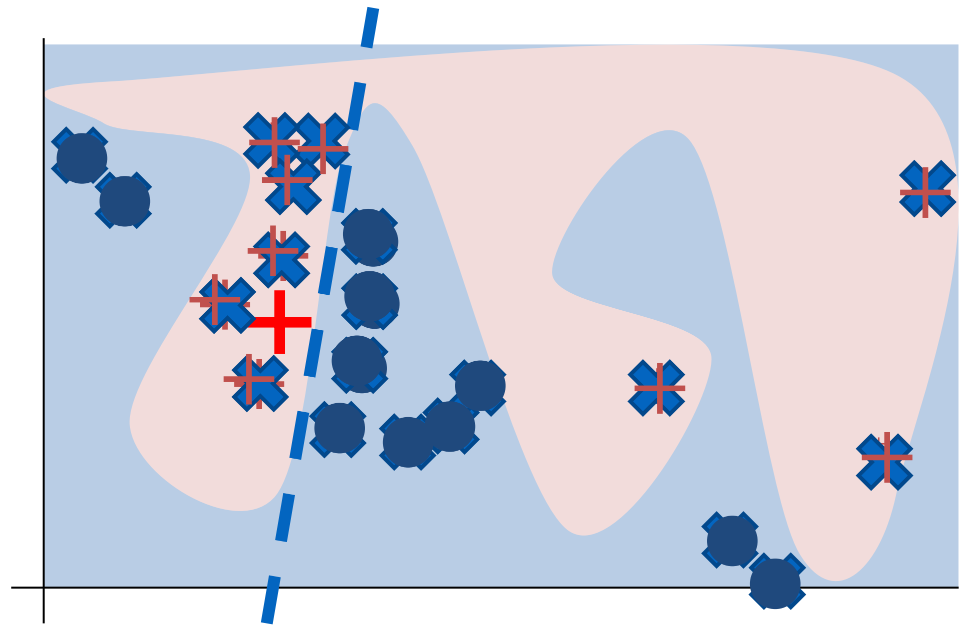
- *Again, relies on human expertise, which is much broader than any of our models*

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?*

LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples
5. Use simple model to explain



(Slide credit: Marco Tulio Ribeiro)

[Ribeiro et al., KDD '16]

How to Make Interpretable Models

- If the original data are $x \in \mathbb{R}^d$, define a new set of variables, $x' \in \{0,1\}^{d'}$ that can serve as the interpretable representation of the data
- An *explanation* is a model $g \in G$ where *G* is the class of interpretable models
    - E.g., linear models, additive scores, decision trees, falling rule lists, …
    - The domain of *g* is $\{0,1\}^{d'}$, i.e., the interpretable representation of the data
- The *complexity* of a model is $\Omega(g)$
    - E.g., depth of a decision tree, number of non-zero weights in a linear model
- The full model is $f : \mathbb{R}^d \to \mathbb{R}$
    - E.g., for classification, *f* is probability that *x* belongs to a certain class
- $\pi_x(z)$ is a proximity measure of how close *z* is to *x*, thus defining a locality around *x*
- Let $\mathcal{L}(f, g, \pi_x)$ be a measure of how *un*faithful *g* is to *f* in the locality defined by π_x
- Then

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

is the best explanatory model for *x* given our choices for $\{\mathcal{L}, \pi_x, \Omega\}$

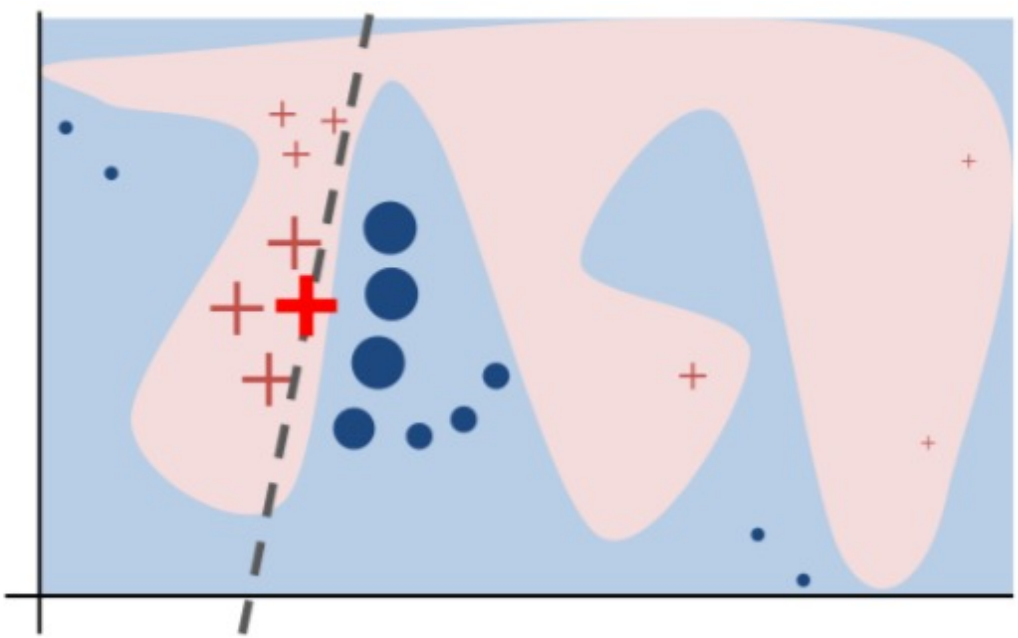
Sparse Linear Explanation

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 σ
  - E.g., cosine distance for bag-of-words, L2 distance or DICE for images
  - Below, z' is the sampled point, nearby to x, and z (a function of z') is the same point in the original space:

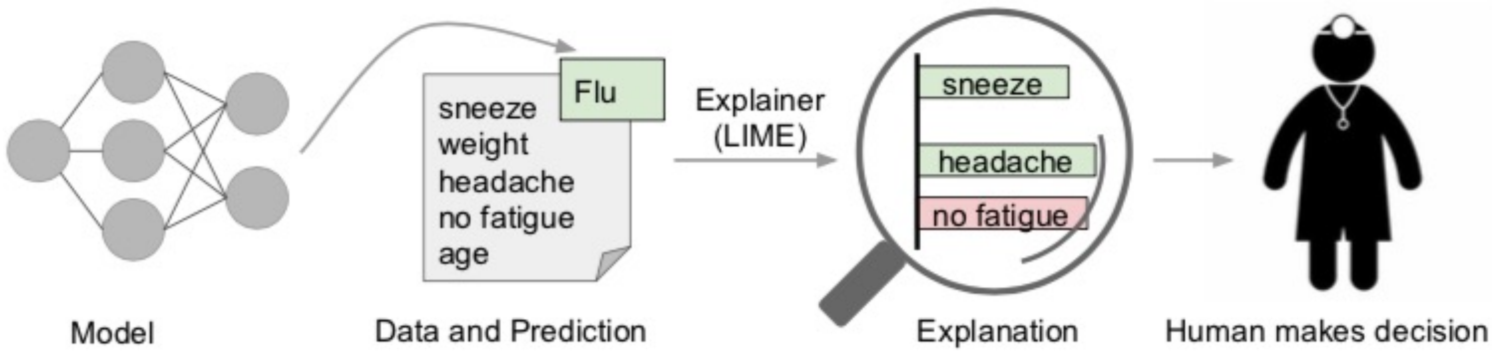
$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z)(f(z) - g(z'))^2$$



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 Text Classification

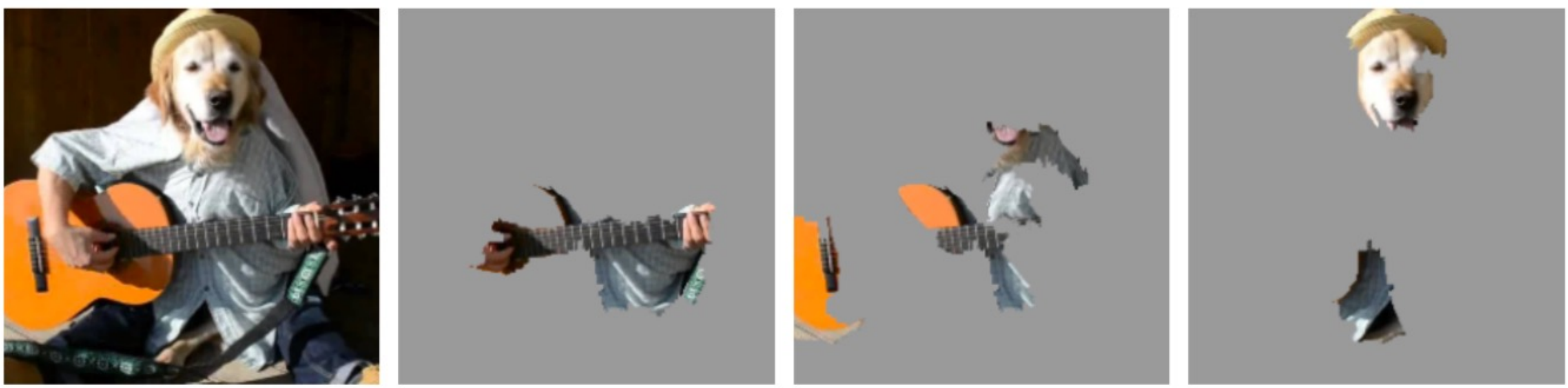
- Bag of words representation, cosine distance for π_x
- Choose K as a limit on the number of words in an explanation



- When sampling data points, subsample words from the original document *x*

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



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

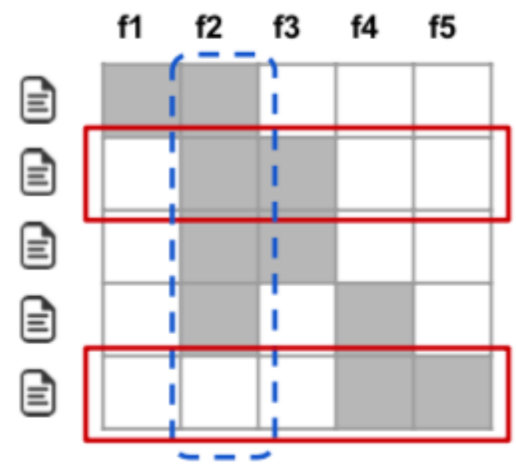
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Choosing a Suite of Examples to Explain

- Choose a diverse, comprehensive set of *B* examples to explain
- WHY?

Choosing a Suite of Examples to Explain

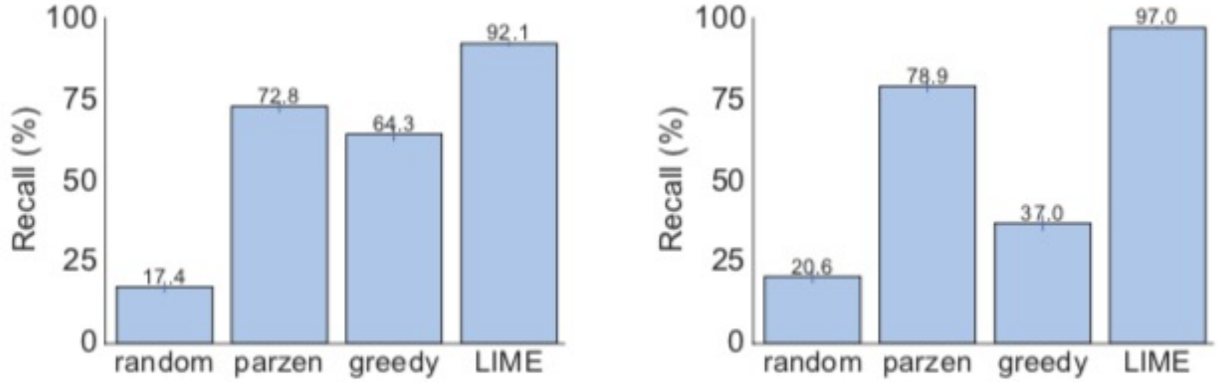
- Choose a diverse, comprehensive set of *B* examples to explain
- Given explanations for a set of instances $X(|X| = n)$, consider the $n \times d'$ *explanation matrix* $\mathcal{W}$ whose rows are examples and columns are features
  - Each entry gives the local importance of that feature for that example
  - For linear models, for instance $x_i, g_i = \xi(x_i)$, set $\mathcal{W}_{ij} = |w_{g_{ij}}|$
    - recall that $g(z') = w_g \cdot z'$
  - I_j is a measure of *global* importance of that feature
    - $I_j = \sqrt{\sum_{i=1}^{n} \mathcal{W}_{ij}}$ for text
    - more difficult for superpixels because they don't recur over different instances



- Use greedy algorithm to maximize marginal coverage (submodular optimization)

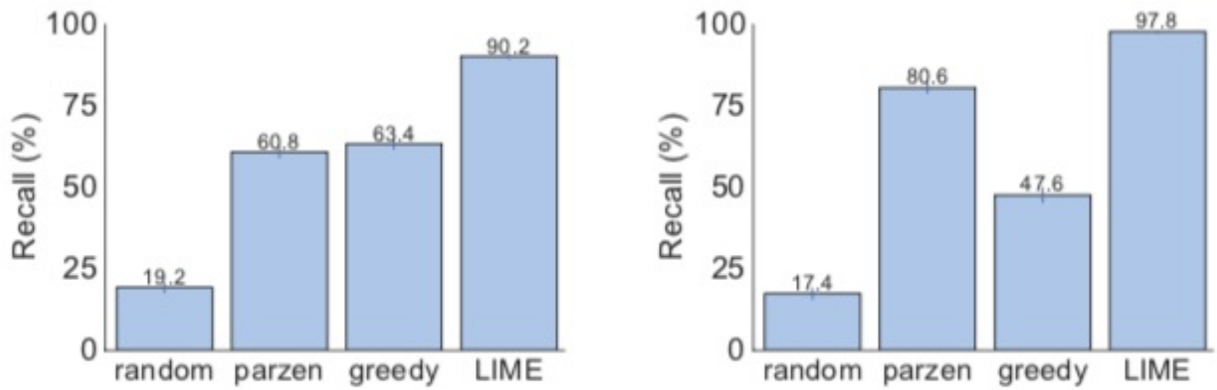
LIME Experiments

- 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
- $K = 10$



(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

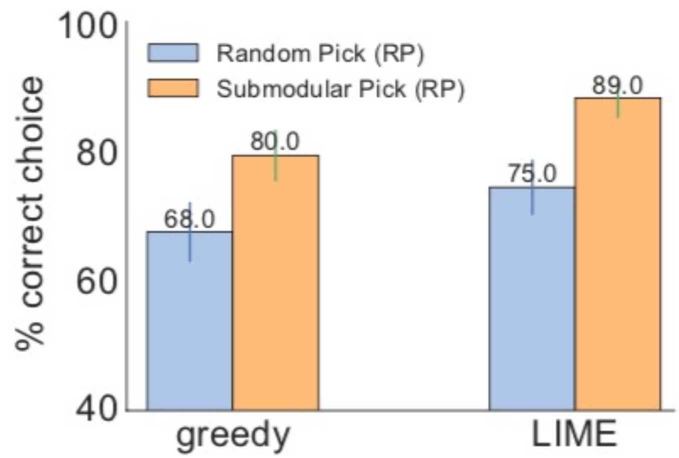


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

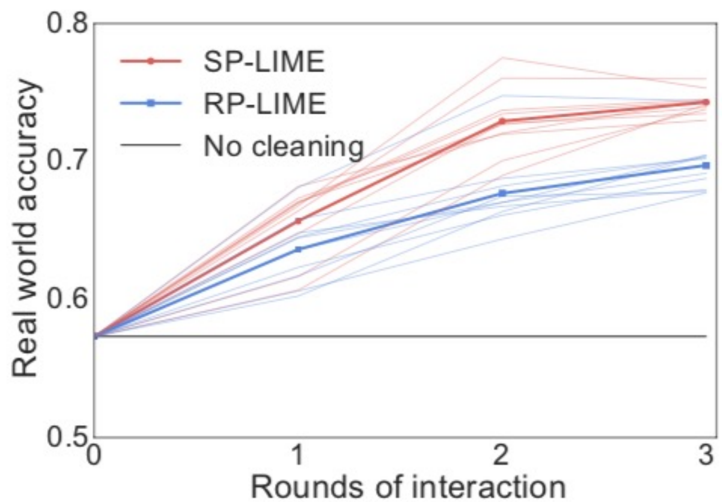
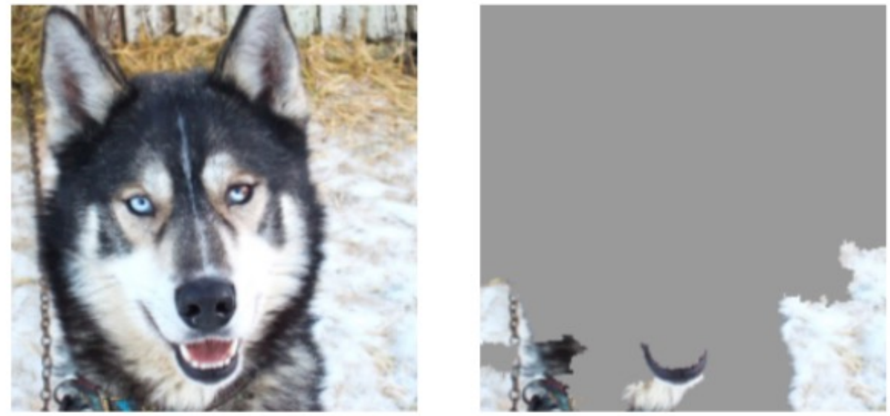


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

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 post-experiment trust and understanding



(a) Husky classified as wolf (b) Explanation

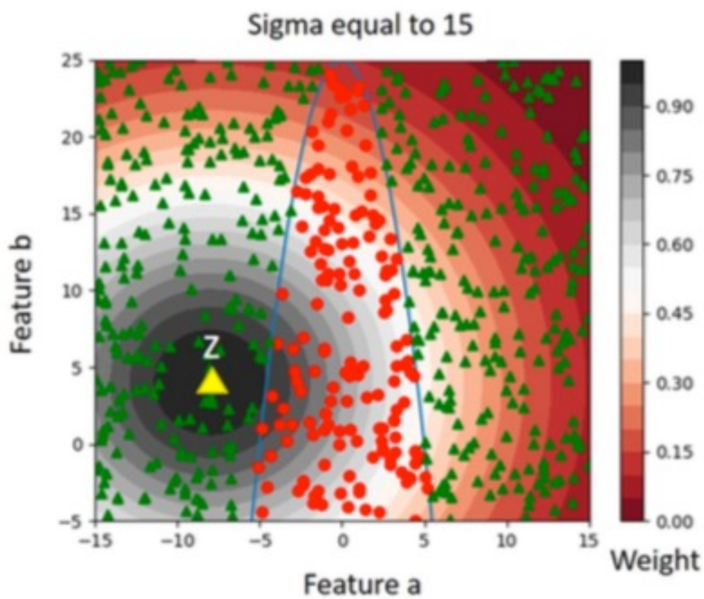
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	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

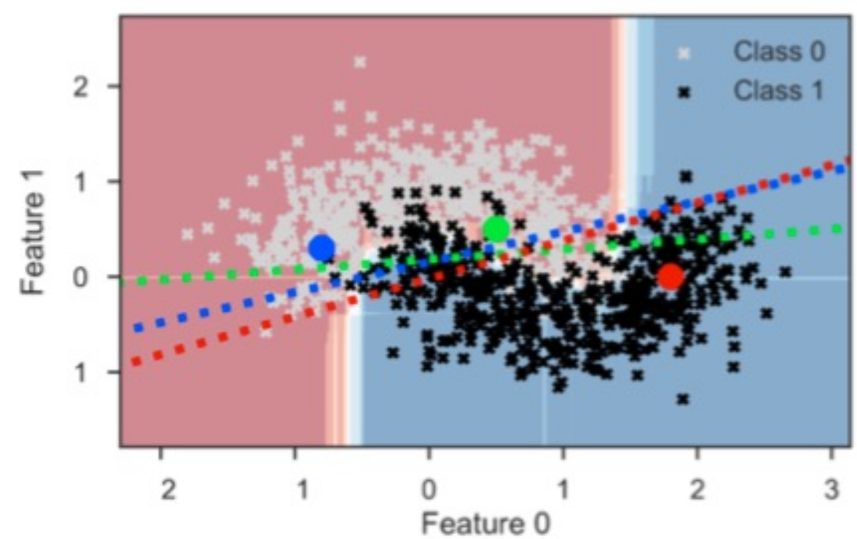
Table 2: "Husky vs Wolf" experiment results.

Critique of LIME

- Choice of σ (size of neighborhood) 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.

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*

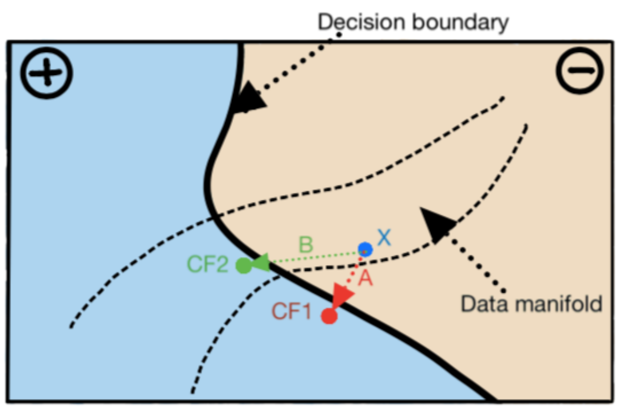


Figure 1: Two possible paths for a datapoint (shown in blue), originally classified in the negative class, to cross the decision boundary. The end points of both the paths (shown in red and green) are valid counterfactuals for the original point. Note that the red path is the shortest, whereas the green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

[Molnar, Interpretable Machine Learning: A guide for Making Black Box Models Explainable, 2022]

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*

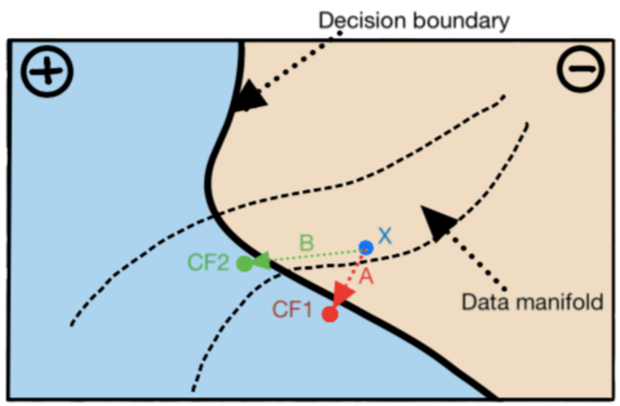


Figure 1: Two possible paths for a datapoint (shown in blue), originally classified in the negative class, to cross the decision boundary. The end points of both the paths (shown in red and green) are valid counterfactuals for the original point. Note that the red path is the shortest, whereas the green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

Can we constrain model class to give an explanation as part of prediction?

Can Attention Models in Deep Learning Serve as Explanations?

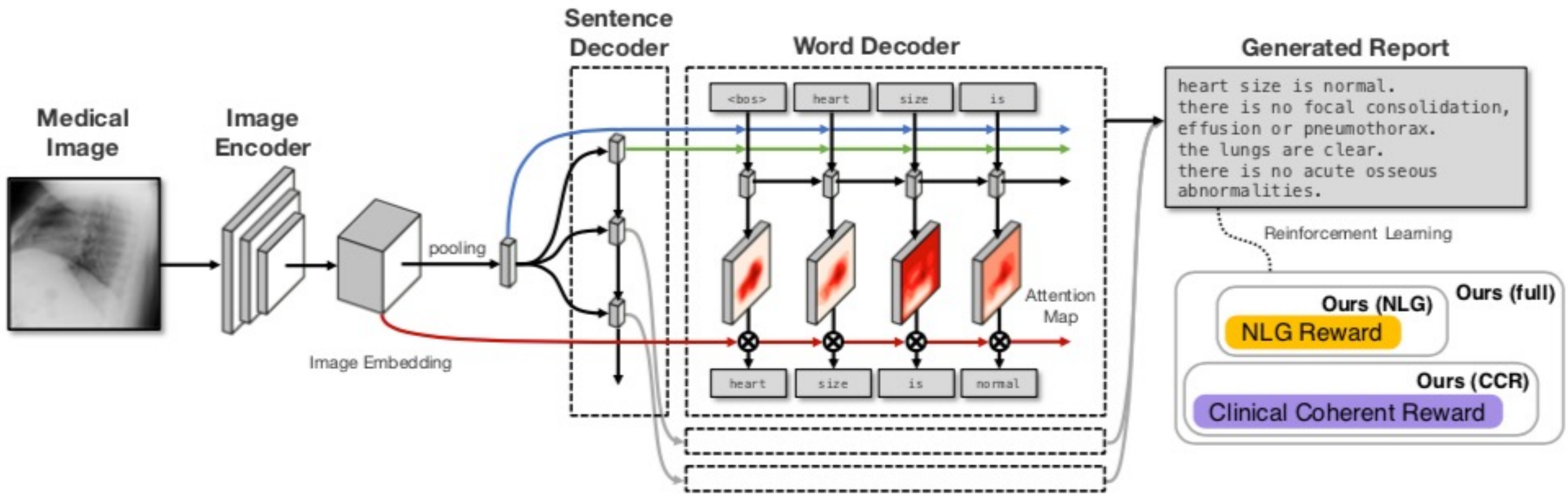
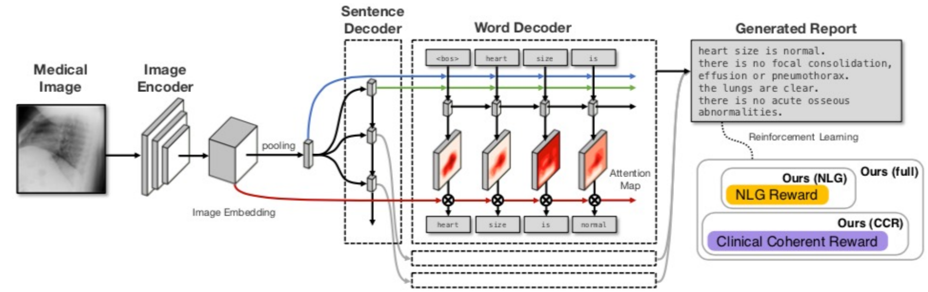


Figure 2: **The model for our proposed *Clinically Coherent Reward*.** Images are first encoded into image embedding maps, and a sentence decoder takes the pooled embedding to recurrently generate topics for sentences. The word decoder then generates the sequence from the topic with attention on the original images. NLG reward, clinically coherent reward, or combined, can then be applied as the reward for reinforcement policy learning.

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., & Ghassemi, M. (2019, April 4). Clinically Accurate Chest X-Ray Report Generation. arXiv.



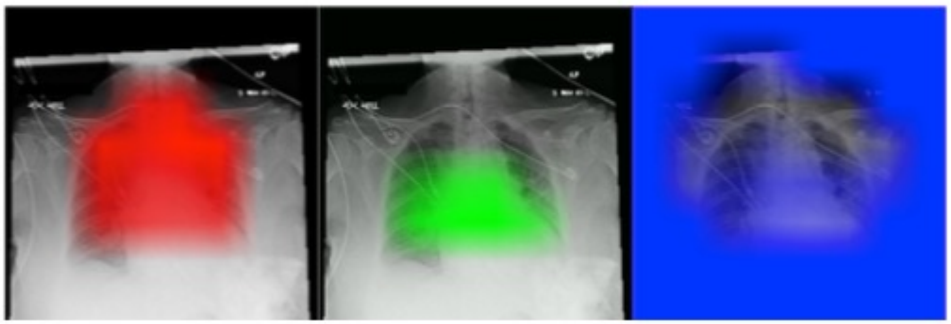
- Image encoder (CNN)
  - Spacial image features $\boldsymbol{V} = \{\boldsymbol{v}\}_{k=1}^{K}$
    - computed by fully connected layer on pre-global-pooling layer of CNN
- Sentence decoder (RNN/LSTM) uses image features
  - $\boldsymbol{h}_i, \boldsymbol{m}_i = \mathrm{LSTM}(\bar{\boldsymbol{v}}; \boldsymbol{h}_{i-1}, \boldsymbol{m}_{i-1})$
  - topic vector and stop signal $\boldsymbol{\tau}_i = \mathrm{ReLU}(\boldsymbol{W}_\tau^T \boldsymbol{h}_i + \boldsymbol{b}_\tau), \quad u_i = \sigma(\boldsymbol{w}_u^T \boldsymbol{h}_i + b_u)$
- Word decoder (RNN/LSTM)
  - Uses $\bar{\boldsymbol{v}}$, $\boldsymbol{\tau}$, and embedding of previous word generated
  - Word is sampled from either conditional probability or overall corpus probability
- Reinforcement learning to favor most readable and clinically correct output
  - Use CheXpert annotations for 12 diagnoses: pos, neg, uncertain, absent
- Hack: remove duplicate generated sentences

Ground Truth

cardiomegaly is moderate. bibasilar atelectasis is mild. there is no pneumothorax. a lower cervical spinal fusion is partially visualized. healed right rib fractures are incidentally noted.

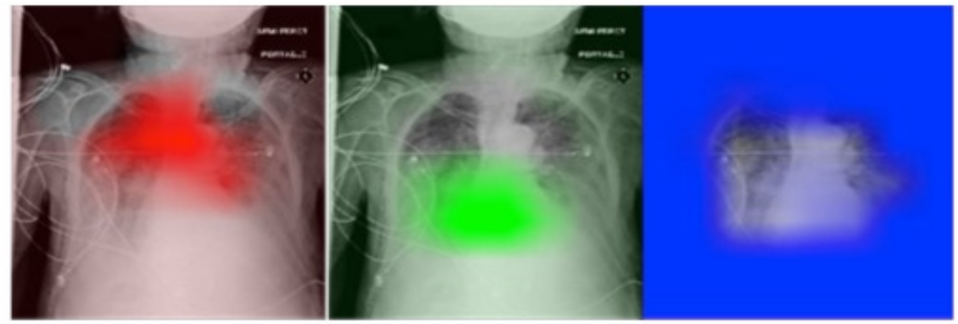
TieNet	Ours (full)
ap portable upright view of the chest. there is no focal consolidation, effusion, or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact.	pa and lateral views of the chest. there is mild enlargement of the cardiac silhouette. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.

Attention Map Identified Relevant Parts of the Image



ap upright and lateral views of the chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.

(a)



as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.

(b)

Figure 3: **Visualization of the generated report and image attention maps.** Different words are underlined with its corresponding attention map shown in the same color.

Attention is not Explanation

But

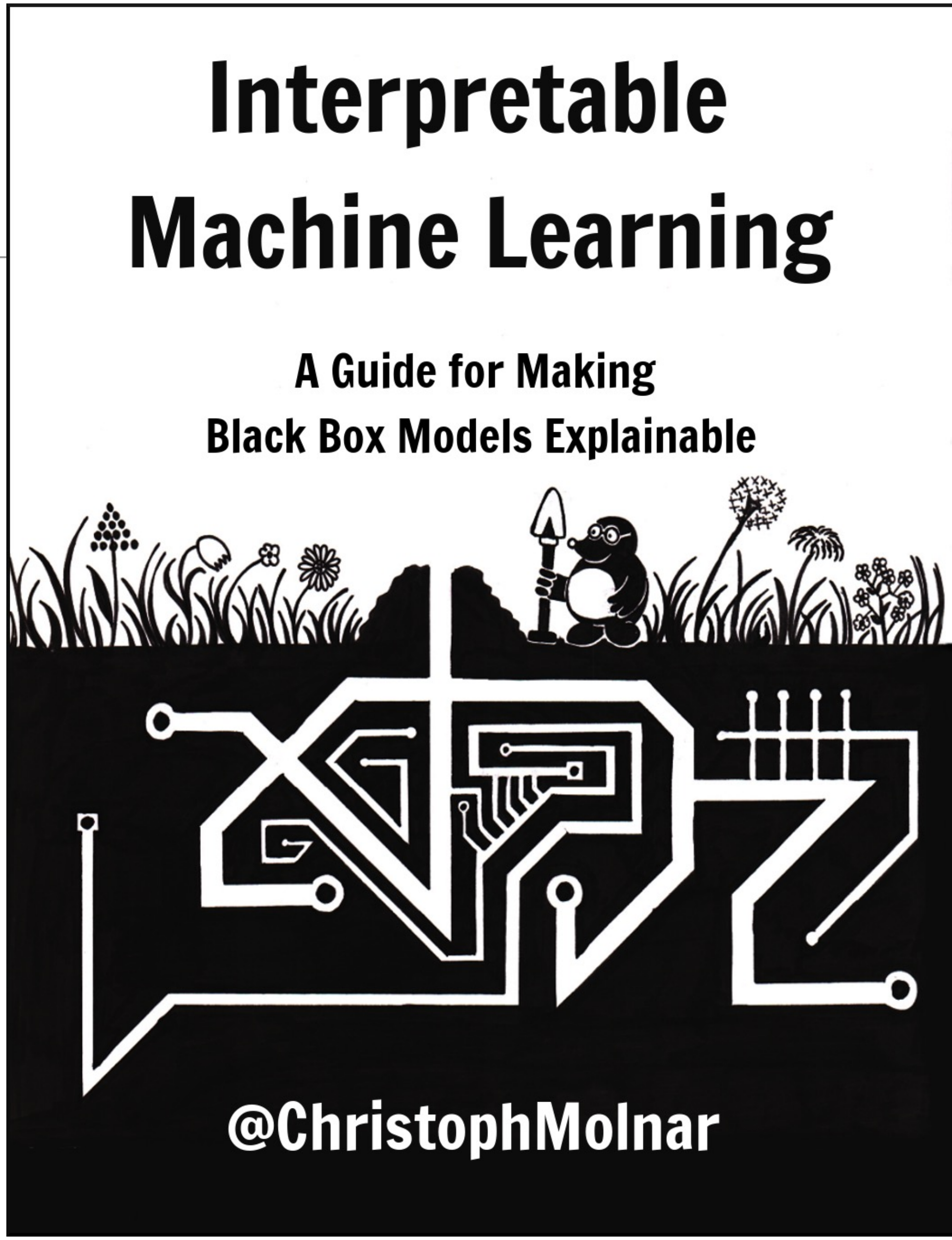
Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

- “assumption that the input units (e.g., words) accorded high attention weights are responsible for model outputs”
- Desiderata if *attention* actually is to give insight into how a DNN operates
  - Attention weights should correlate with feature importance measures (e.g., gradient-based measures)
  - Alternative (or counterfactual) attention weight configurations ought to yield corresponding changes in prediction

- Mixed results, though the study has been criticized for methodology
  - “evidence that correlation between intuitive feature importance measures (including gradient and feature erasure approaches) and learned attention weights is weak”
  - counterfactual attention distributions — which would tell a different story about why a model made the prediction that it did — often have no effect on model output

Jain, S., & Wallace, B. C. (2019, February 26). Attention is not Explanation. arXiv.



Also, see work by faculty here in Boston….

Hima Lakkaraju (Harvard)

Finale Doshi (Harvard)

Manish Raghavan (MIT)

Byron Wallace (Northeastern)

https://christophm.github.io/interpretable-ml-book/