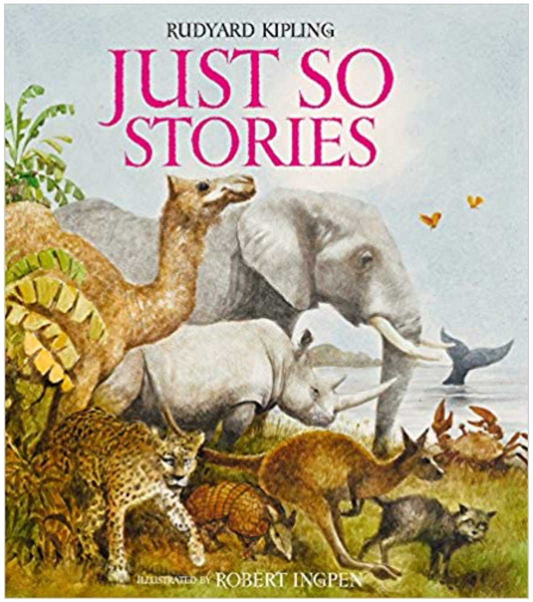
Interpretability

February 27, 2020

Massachusetts Institute of Technology

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 my critique 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)
- 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

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

Explanations from a Backward-Chaining Rule Interpreter

- 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
-

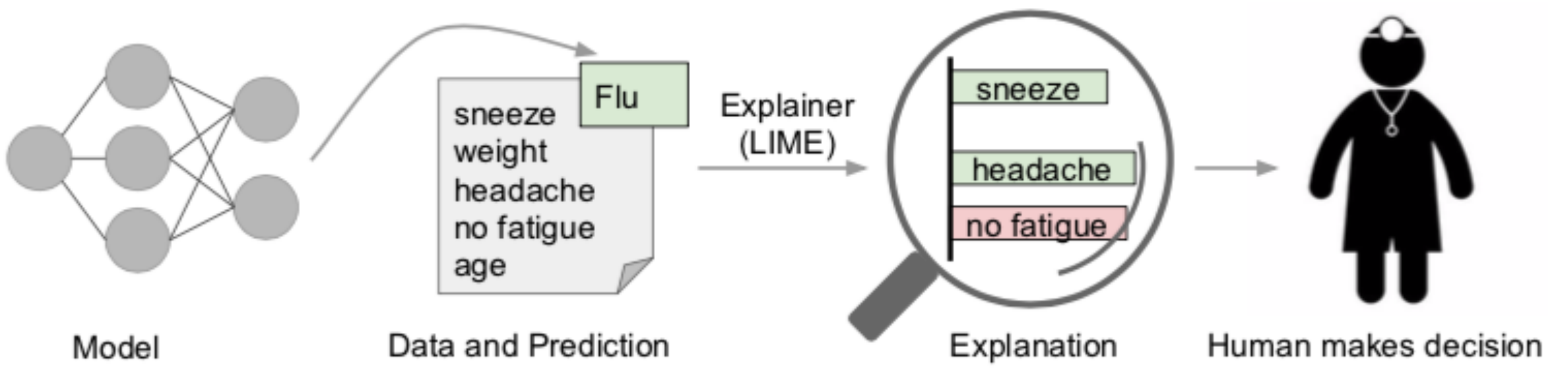
- ****WHY** [i.e., WHY is it important to find out a locus of infection in PATIENT-1?]
- [3.0] . . . in order to find out about an organism or class of organisms for which therapy should cover
  It has already been established that:
  - [3.1] the site of the culture is blood, and
  - [3.2] the stain 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 Interpretable Model-agnostic Explanations (LIME)

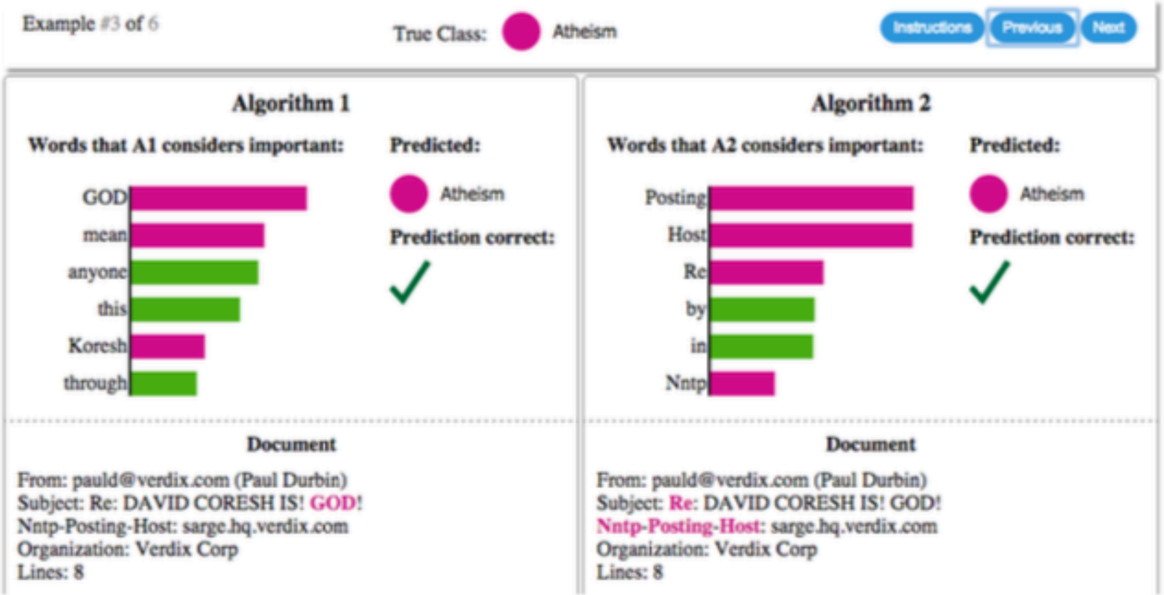


- A model predicts that a patient has the flu, and LIME highlights:
  - Sneeze and headache are portrayed as contributing to the “flu” prediction
  - “no fatigue” is evidence against it.
- With these, a doctor can make an informed decision about whether to trust the model’s prediction.

- *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



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

- 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?*

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* if to *f* in the locality defined by π_x
- Then

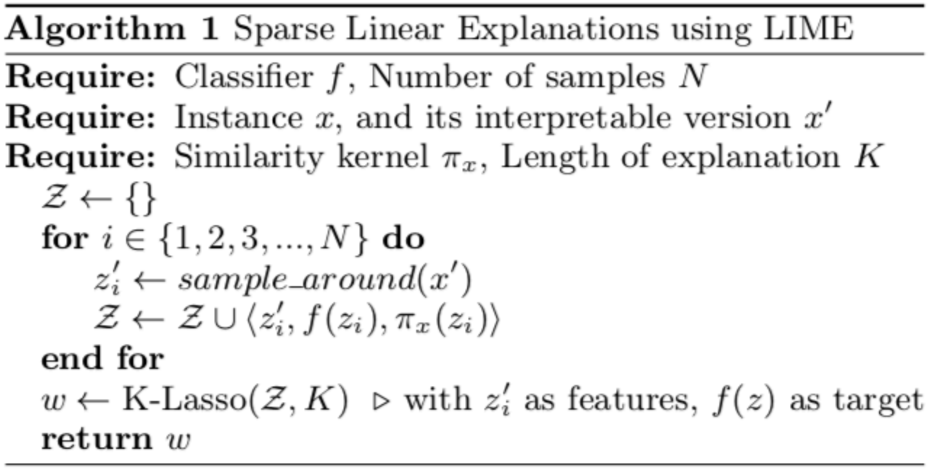
$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

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

Use Sampling to Generate Data in a Local Neighborhood

- Goal is model-agnostic explanation capability
  - Thus, cannot rely on knowing anything about the model *f*
- To explain the model's result around the interpretable point *x'*,
  - sample in the interpretable representation space to get a set of points $z' \in \{0,1\}^{d'}$ to create a dataset $\mathcal{Z}$ of perturbed samples
  - recover sample $z \in \mathbb{R}^d$ and compute $f(z)$ as the label for $z \in \mathcal{Z}$
  - optimize $\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$, weighting contributions of *z* by $\pi_x(z)$

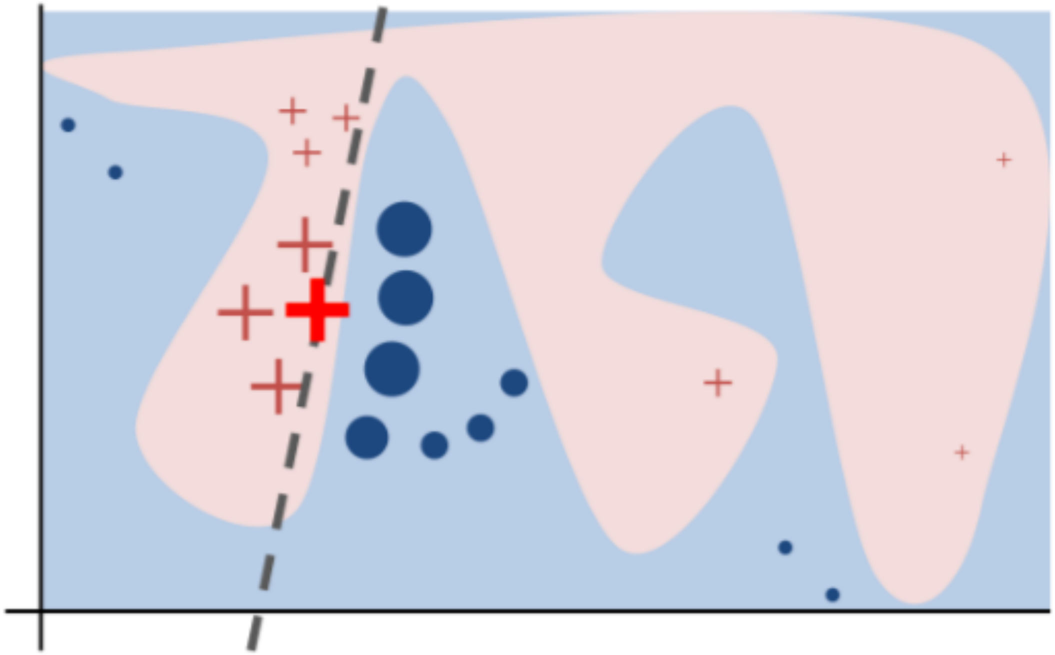
Sparse Linear Explanation



```
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 $\pi_x$, 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$(\mathcal{Z}, K)$  $\triangleright$ 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

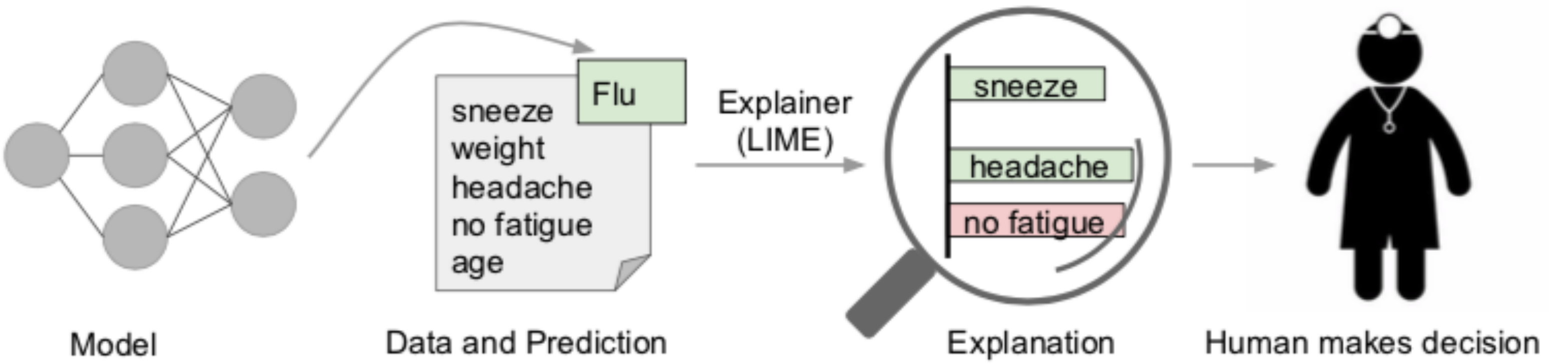
$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z)(f(z) - g(z'))^2$$



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.

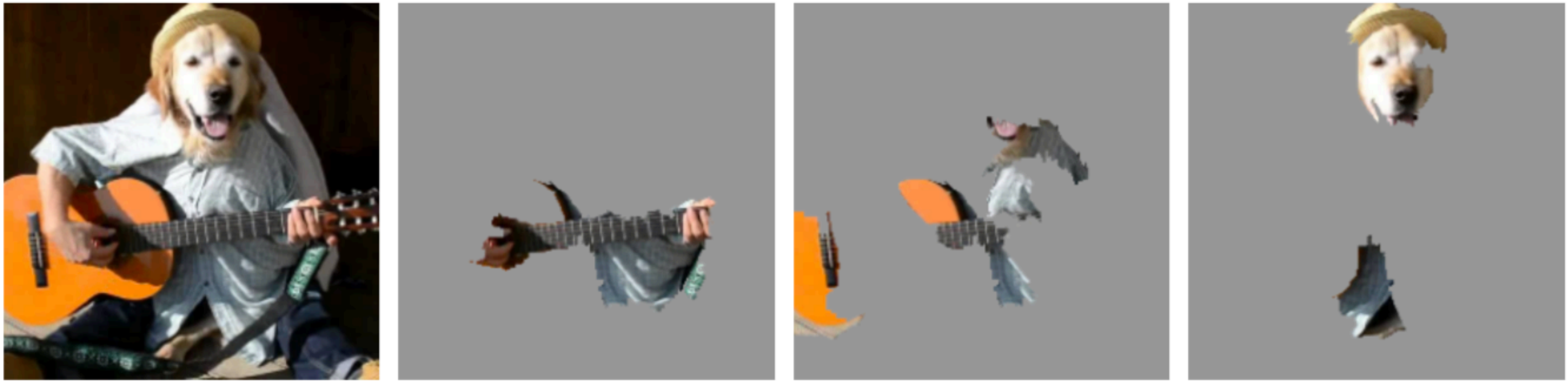
Apply to Text Classification

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



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

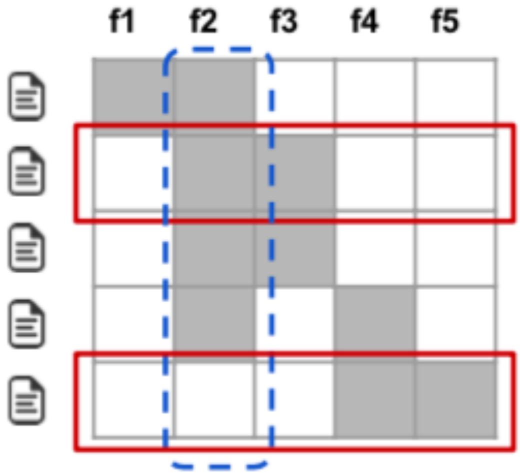


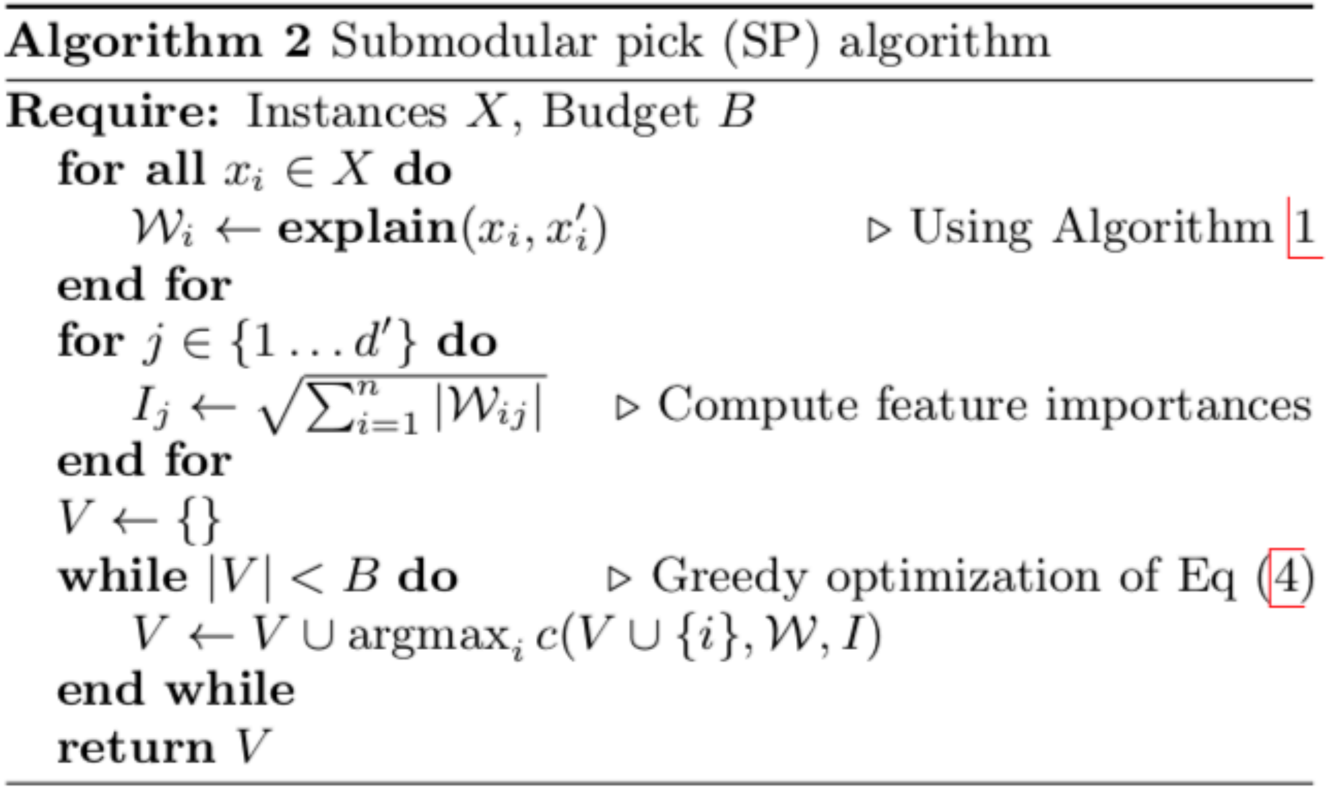
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

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
- 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 to superpixels because they don't recur over different instances





```
Algorithm 2 Submodular pick (SP) algorithm
Require: Instances $X$, Budget $B$
  for all $x_i \in X$ do
    $\mathcal{W}_i \leftarrow$ explain$(x_i, x'_i)$          ▷ Using Algorithm 1
  end for
  for $j \in \{1 \ldots d'\}$ do
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$     ▷ Compute feature importances
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ do            ▷ Greedy optimization of Eq (4)
    $V \leftarrow V \cup \text{argmax}_i\, c(V \cup \{i\}, \mathcal{W}, I)$
  end while
  return $V$
```

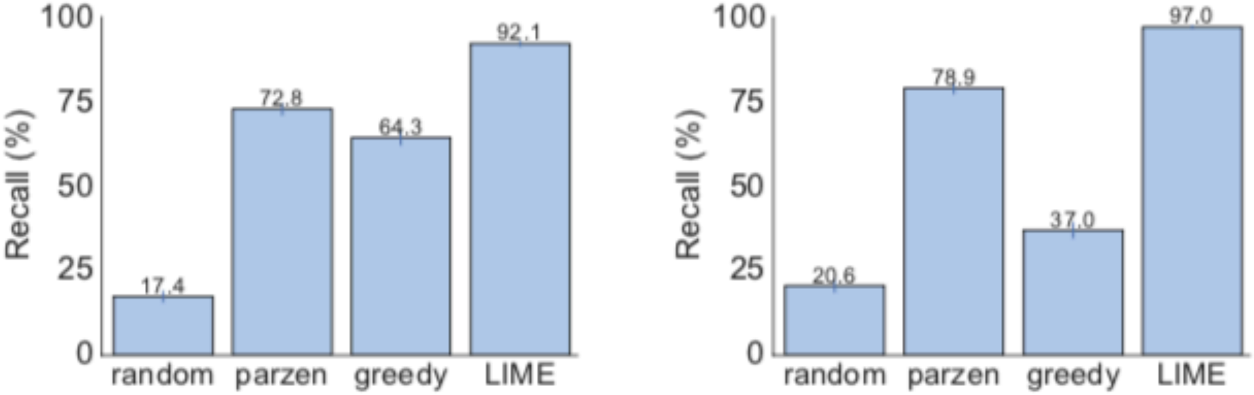
$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$$

$$\text{Pick}(\mathcal{W}, I) = \underset{V, |V| \leq B}{\arg\max}\, c(V, \mathcal{W}, I)$$

Choosing *i* that maximizes marginal coverage $c(V \cup \{i\}, \mathcal{W}, I) - c(V, \mathcal{W}, I)$ approximates optimum

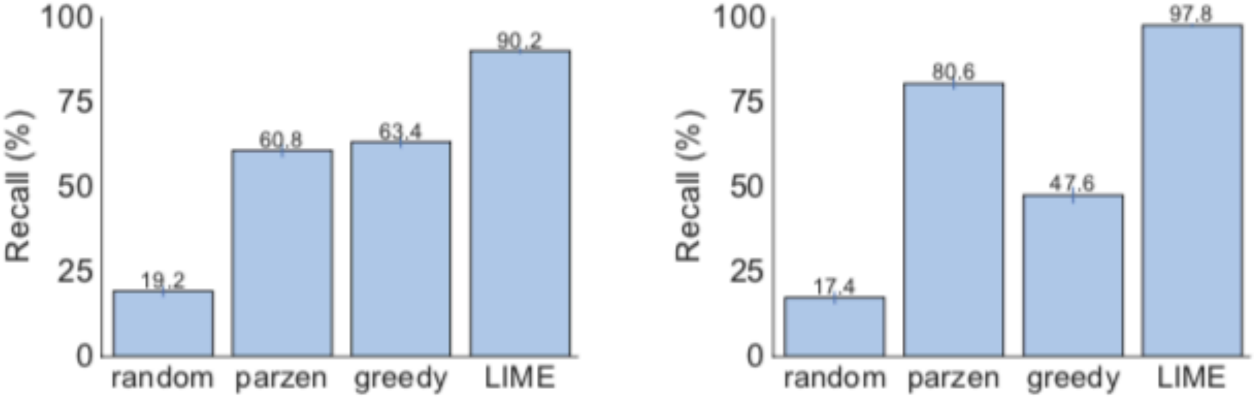
LIME Experiments

- 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

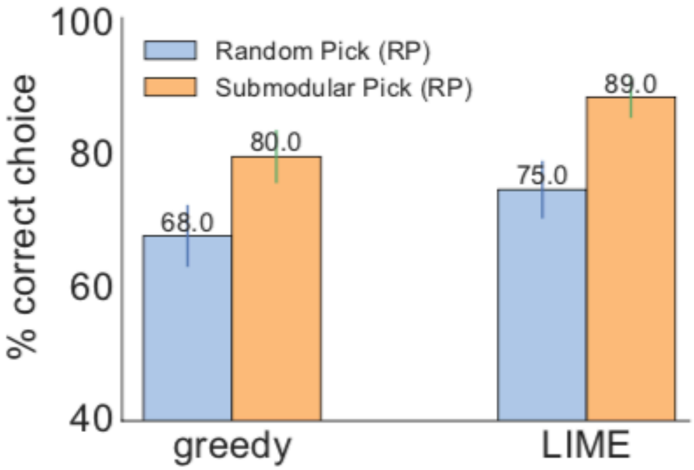


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

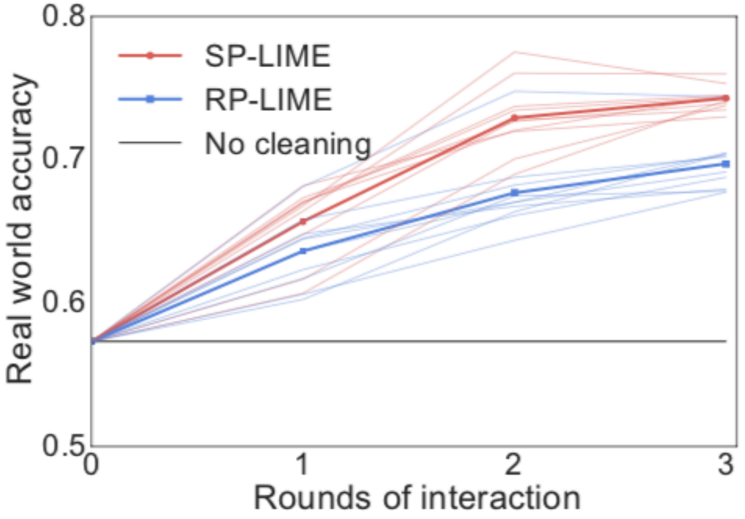
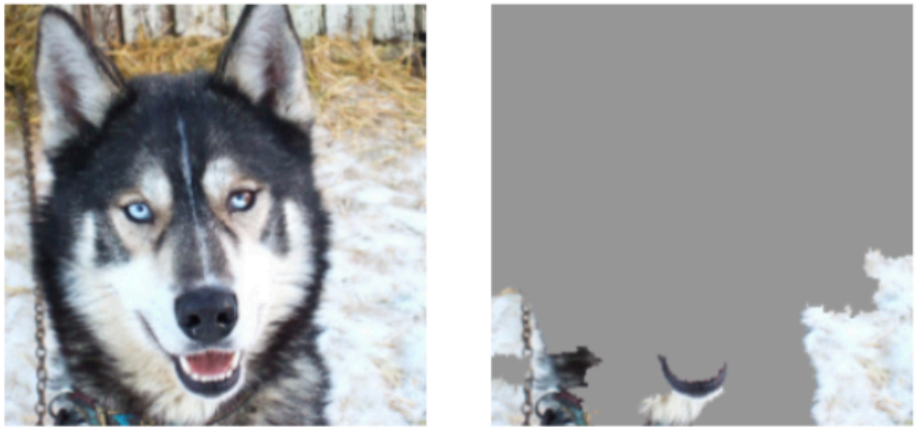


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

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

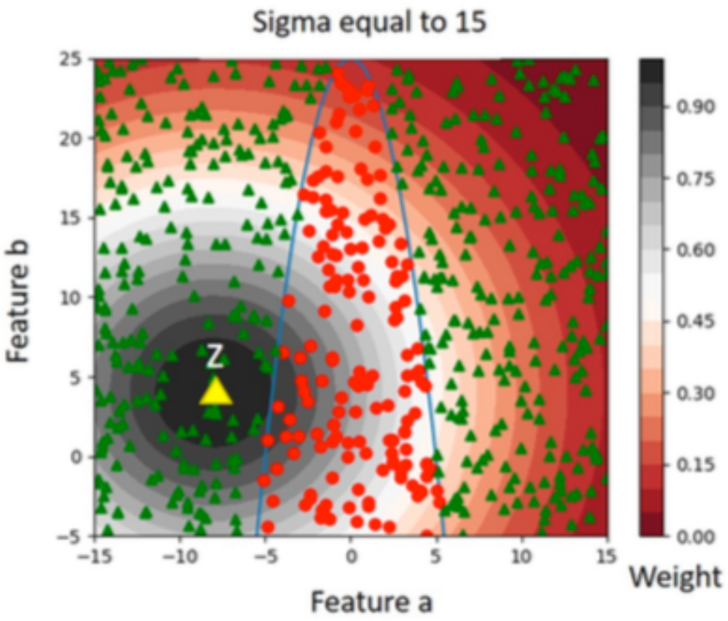
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	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

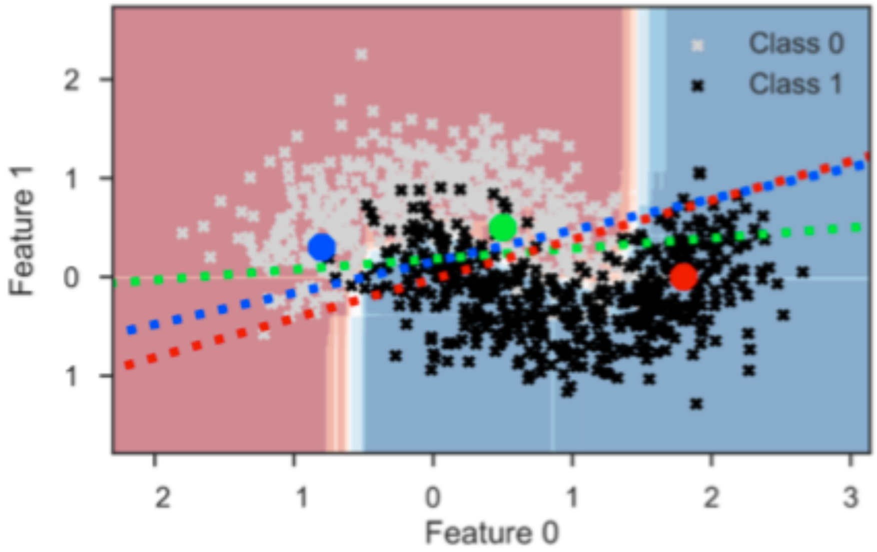
Table 2: "Husky vs Wolf" experiment results.

Critique of LIME

- Choice of σ 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 neighbourhood 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]

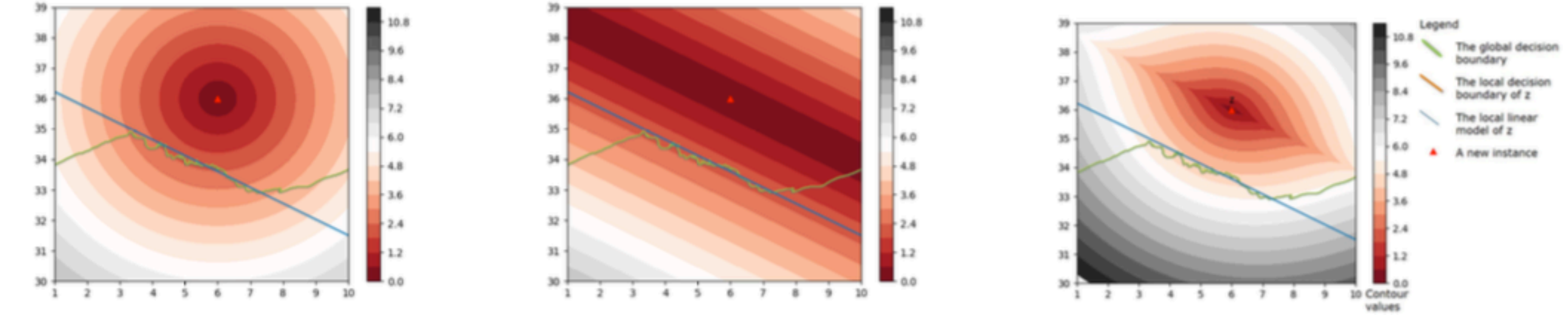
Adhikari, A., Tax, D. M. J., Satta, R., & Fath, M. (2018, December 21). Example and Feature importance-based Explanations for Black-box Machine Learning Models. arXiv.

LEAFAGE - Local Example and Feature importance-based model AGnostic Explanations

- Experts often reason by analogy from previous cases
  - In law, this is formally enshrined as *precedent*
  - In medicine, we see it in the behavior of experts
- Case-based reasoning: retrieve, adapt, learn
- Contrastive justification
  - Not “why did you choose *x*?”,
  - but “why did you choose *x* rather than *y*?”
- Assume that a black-box model $f : \mathcal{X} \rightarrow \mathcal{Y}$ solves a classification problem where
  - $\mathcal{X} = \mathbb{R}^d \quad \text{and} \quad \mathcal{Y} = \{c_1, c_2\}$
  - training set $X = [x_1, \ldots, x_n]$ and $y_{\text{true}} = [y_1, \ldots, y_n]$, $y_{\text{predicted}} = \{f(x) \mid x_i \in X\}$
- To explain $f(z) = c_z$, use
  - $\text{allies} = \{x \in X \mid f(x) = c_z\}$, $\text{enemies} = \{x \in X \mid f(x) \neq c_z\}$

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.

- Choose a subset of training examples in the neighborhood of *z*
- Build a linear model from that subset
- Compute importance of each feature in that model
- Define a similarity measure based on features weighted by their importance
  - $g(\boldsymbol{x}) = \boldsymbol{w}_z \boldsymbol{x} + c$ defines the decision boundary
  - $b(t) = \sqrt{d} \cdot ||\boldsymbol{w}_z^T \boldsymbol{t} - \boldsymbol{w}_z^T \boldsymbol{z}|| + ||\boldsymbol{t} - \boldsymbol{z}||$ is the distance function, $\boldsymbol{w}_z = (w_{z1}, \ldots, w_{zd})^T$
- Explanation gives
  - Most important features
  - Most similar examples that give the same answer
- *(details in paper)*

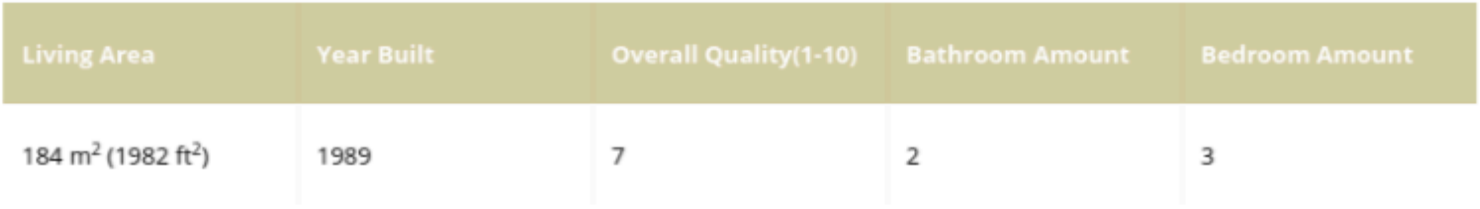


(a) Contour-lines of the first term of the similarity measure.

(b) Contour-lines of the second term of the similarity measure.

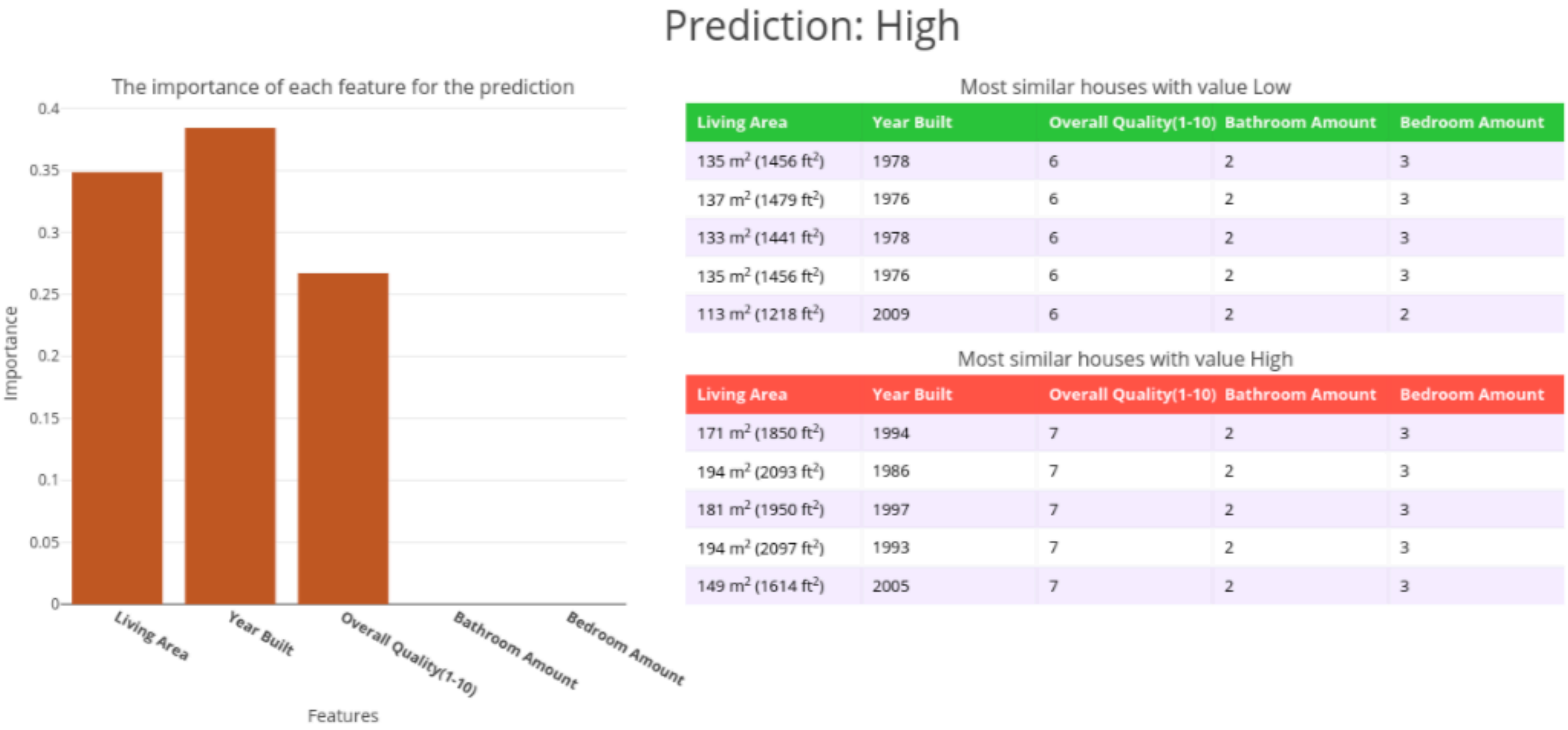
(c) Contour-lines of the black-box similarity measure.

Figure 5: An example to illustrate the black-box similarity measure.



Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
184 m^2 (1982 ft^2)	1989	7	2	3

Figure 7: Example of a house that is predicted as value low by the machine learning model.



Prediction: High

Most similar houses with value Low

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
135 m^2 (1456 ft^2)	1978	6	2	3
137 m^2 (1479 ft^2)	1976	6	2	3
133 m^2 (1441 ft^2)	1978	6	2	3
135 m^2 (1456 ft^2)	1976	6	2	3
113 m^2 (1218 ft^2)	2009	6	2	2

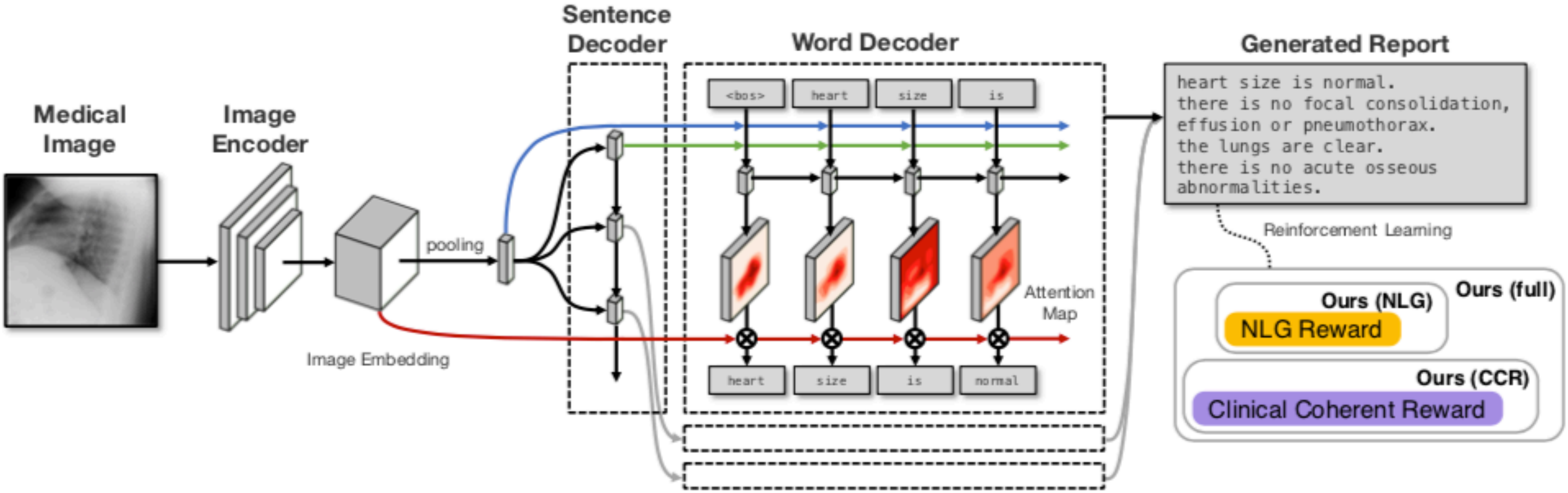
Most similar houses with value High

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
171 m^2 (1850 ft^2)	1994	7	2	3
194 m^2 (2093 ft^2)	1986	7	2	3
181 m^2 (1950 ft^2)	1997	7	2	3
194 m^2 (2097 ft^2)	1993	7	2	3
149 m^2 (1614 ft^2)	2005	7	2	3

See user study in paper

Figure 8: Example of a LEAFAGE explanation.

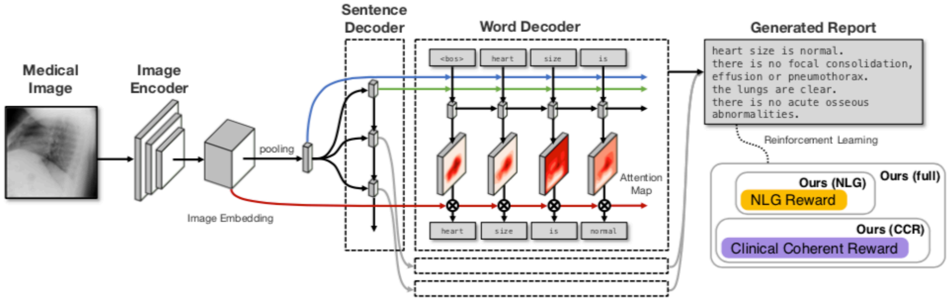
Can Attention Models in Deep Learning Serve as Explanations?



```
heart size is normal.
there is no focal consolidation,
effusion or pneumothorax.
the lungs are clear.
there is no acute osseous
abnormalities.
```

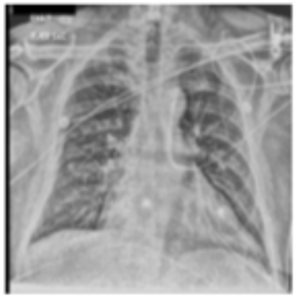
Figure 2: **The model for our proposed *Clinically Coherent Reward*.** Images are first encoded into image embedding maps, and a sentence decoder takes the pooled embedding to recurrently generate topics for sentences. The word decoder then generates the sequence from the topic with attention on the 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.



```
heart size is normal.
there is no focal consolidation,
effusion or pneumothorax.
the lungs are clear.
there is no acute osseous
abnormalities.
```

- 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 $\boldsymbol{\tau}_i = \mathrm{ReLU}(\boldsymbol{W}_{\tau}^{T}\boldsymbol{h}_i + \boldsymbol{b}_{\tau}),\ \ u_i = \sigma(\boldsymbol{w}_u^{T}\boldsymbol{h}_i + b_u)$
- 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

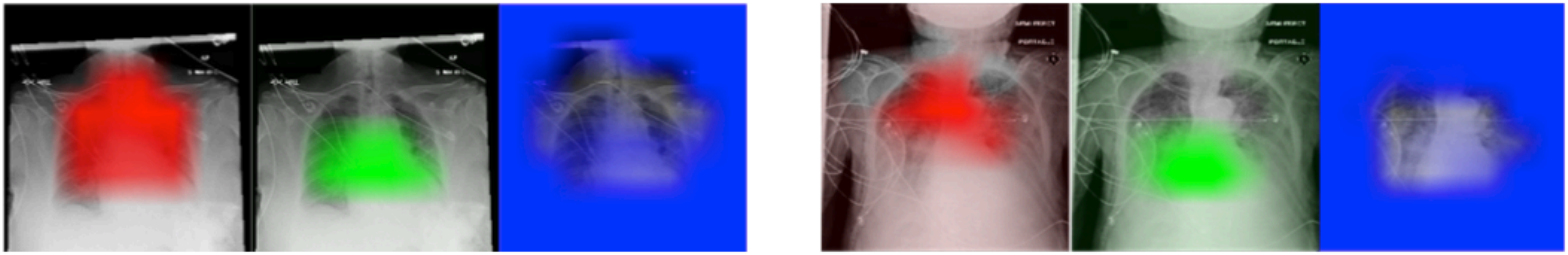


Ground Truth

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.

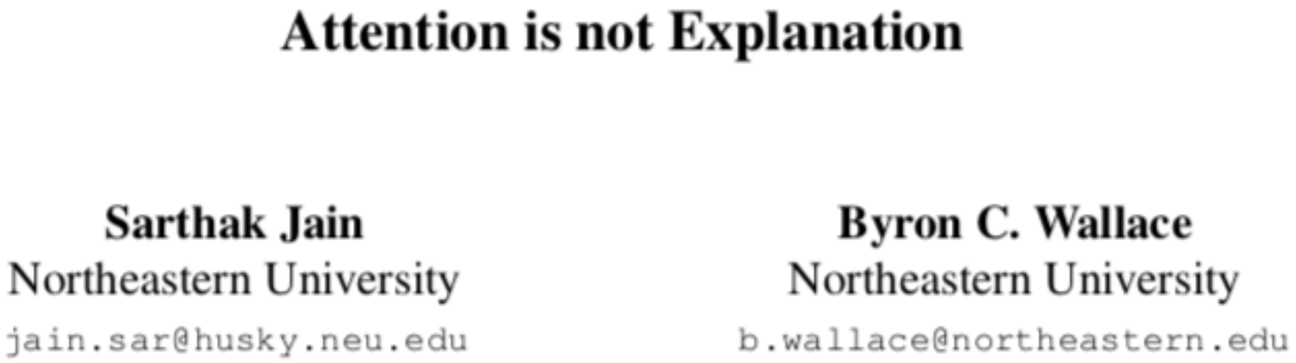
(a)

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.

But



Attention is not Explanation

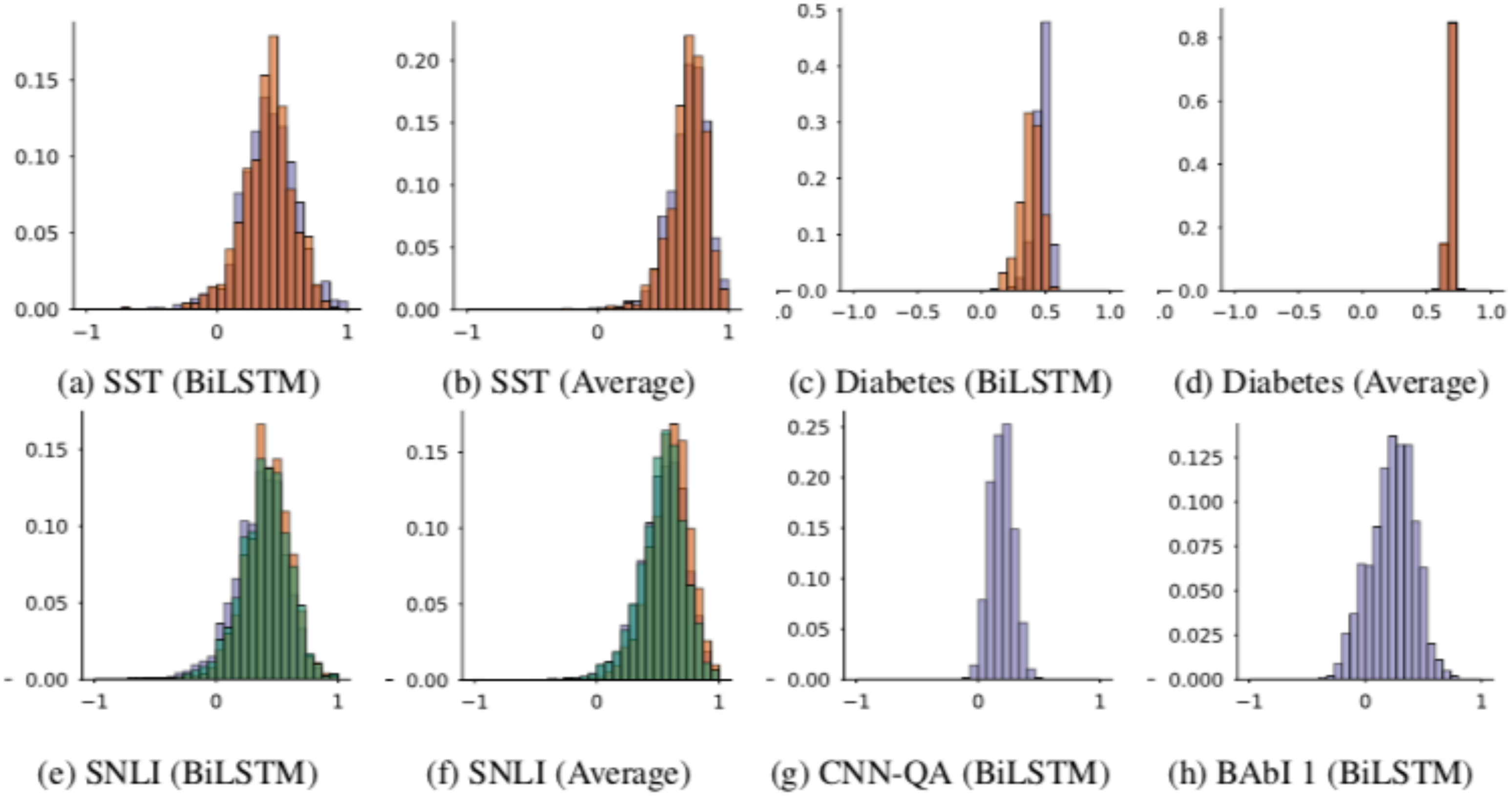
Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

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.



(a) SST (BiLSTM)

(b) SST (Average)

(c) Diabetes (BiLSTM)

(d) Diabetes (Average)

(e) SNLI (BiLSTM)

(f) SNLI (Average)

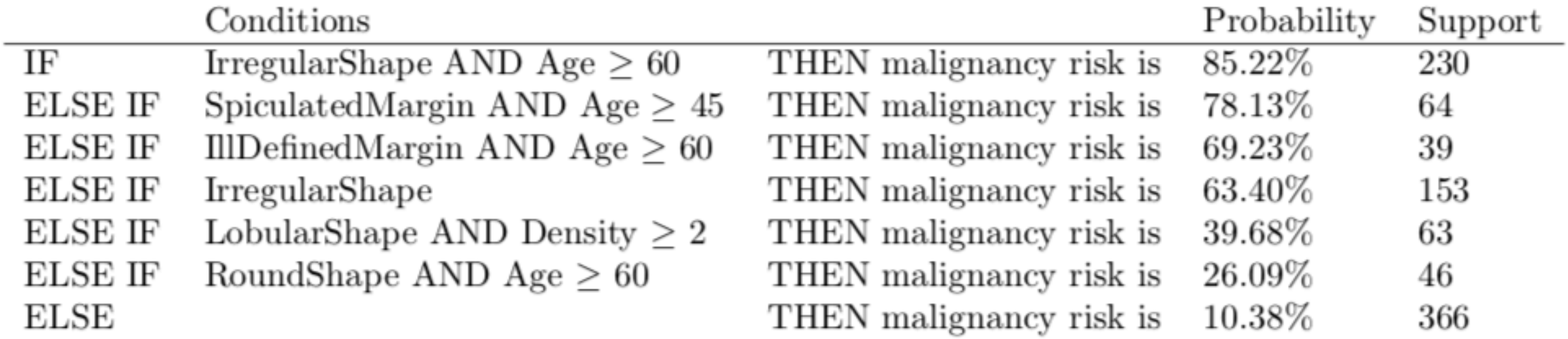
(g) CNN-QA (BiLSTM)

(h) BAbI 1 (BiLSTM)

Figure 2: Histogram of **Kendall** τ between attention and gradients. Encoder variants are denoted parenthetically; colors indicate predicted classes. Exhaustive results are available for perusal online.

Building Simple Models
Falling Rule Lists

- Willing to sacrifice (some) performance for simplicity of model
- Falling Rule List is a form of Decision List, a one-sided Decision Tree
  - the order of rules determines which example should be classified by each rule
  - the estimated probability of success decreases monotonically down the list
- Rank rules to form a predictive model
- Stratify patients into decreasing risk sets



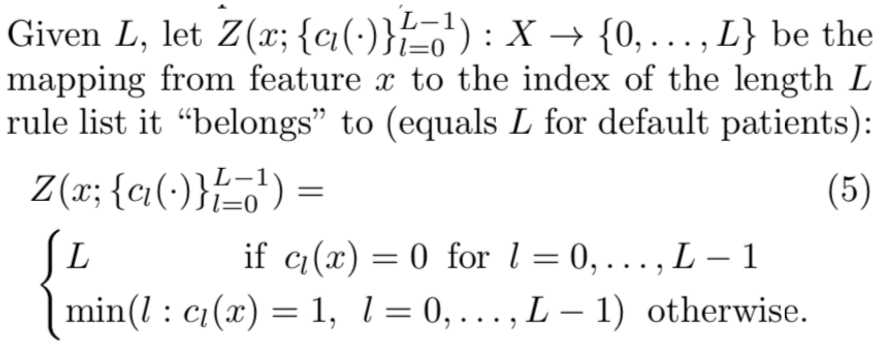
	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk 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. (2015). Falling Rule Lists. Aistats.

Learning Falling Rule Lists

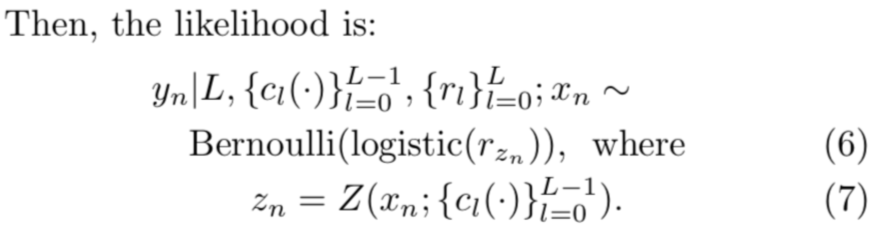
- Data: $D = \{(x_n, y_n)\}_{n=1,\ldots,N}, \quad x_n \in X, \quad y_n \in \{0,1\}$
- Bayesian approach:
  - Hyperparameters H
  - Falling Rule List parameters $\theta \text{ with prior } p_\theta(\cdot; H)$
  - Likelihood $p_{\boldsymbol{Y}}(\{y_n\} \mid \theta; \{x_n\})$
  - Size of rule list $L \in \mathbb{Z}^+$
  - Space of possible IF clauses (Boolean functions on X) $B_X(\cdot)$
  - Clauses $c_l(\cdot) \in B_X(\cdot) \ni c_l = 1$ iff *x* satisfies a set of conditions, for $l = 1, \ldots, L-1$
  - Risk scores $r_l \in \mathbb{R}, \text{ for } l = 0, \ldots, L \ni r_{l+1} \leq r_l$
    - These will be scaled by logistic function to yield a probability



Given L, let $Z(x;\{c_l(\cdot)\}_{l=0}^{L-1}) : X \to \{0,\dots,L\}$ be the mapping from feature x to the index of the length L rule list it "belongs" to (equals L for default patients):

$$Z(x;\{c_l(\cdot)\}_{l=0}^{L-1}) = \tag{5}$$

$$\begin{cases} L & \text{if } c_l(x) = 0 \text{ for } l = 0,\dots,L-1 \\ \min(l : c_l(x) = 1,\ l = 0,\dots,L-1) & \text{otherwise.} \end{cases}$$



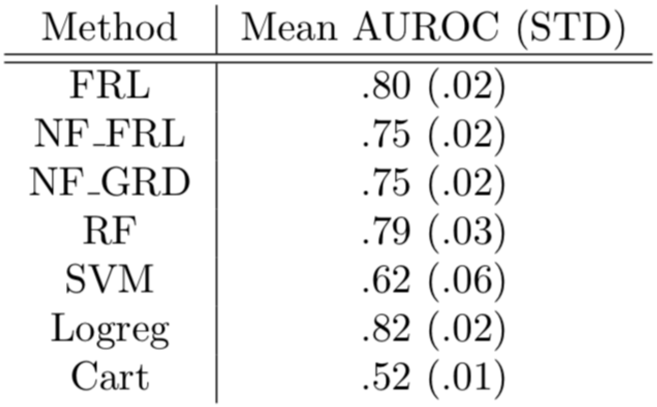
Then, the likelihood is:

$$y_n | L, \{c_l(\cdot)\}_{l=0}^{L-1}, \{r_l\}_{l=0}^{L}; x_n \sim$$
$$\text{Bernoulli}(\text{logistic}(r_{z_n})), \quad \text{where} \tag{6}$$
$$z_n = Z(x_n; \{c_l(\cdot)\}_{l=0}^{L-1}). \tag{7}$$

- Lots of details (see the paper)
  - use a "frequent itemset mining" algorithm to find clauses with enough support
  - choose r_l to be log of products of real numbers
  - L is drawn from a Poisson distribution
  - use can express preference over lengths of clauses
  - MAP decision list is computed by simulated annealing: {swap, replace, add, delete} a clause
  - Gibbs sampling to estimate posteriors

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)

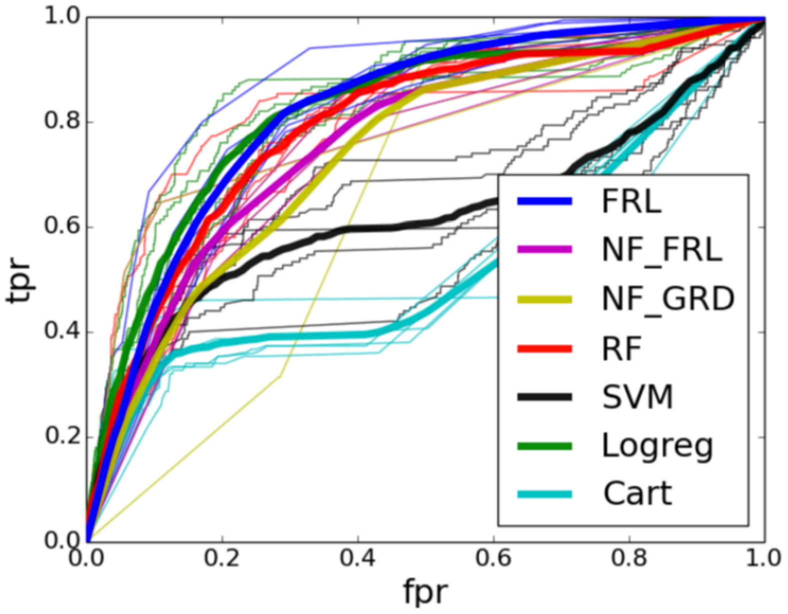
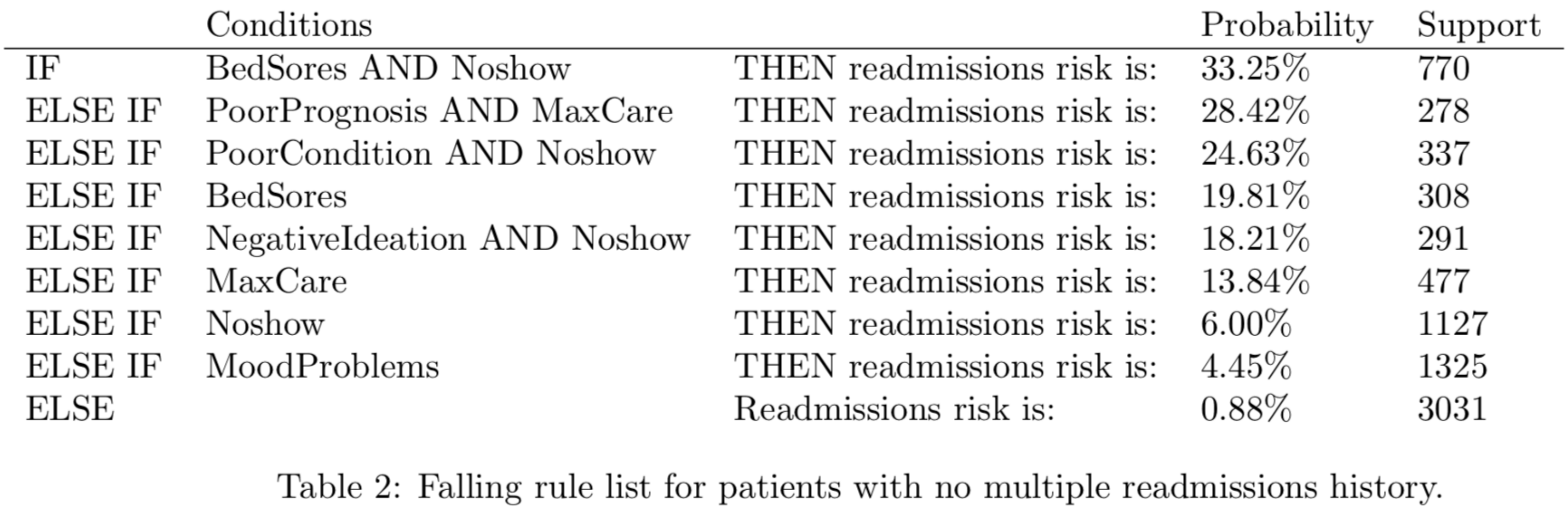


Figure 2: ROC curves for readmissions prediction.

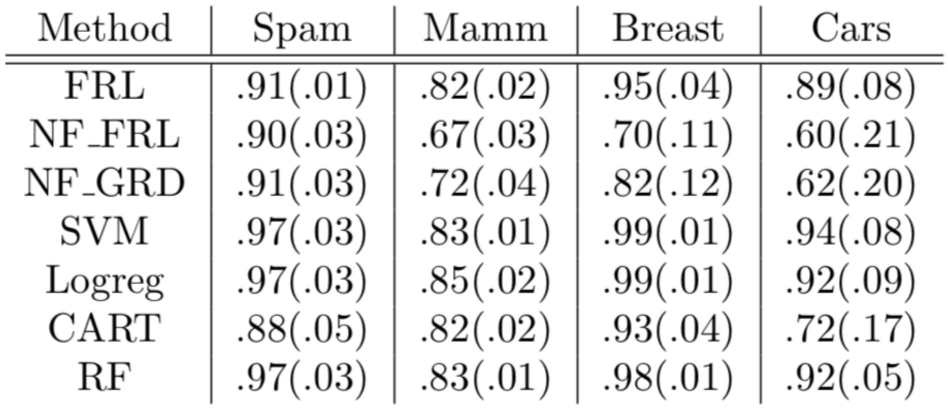
Readmission Rule List



	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.

Test on Various UCI Data Sets



Method	Spam	Mamm	Breast	Cars
FRL	.91(.01)	.82(.02)	.95(.04)	.89(.08)
NF_FRL	.90(.03)	.67(.03)	.70(.11)	.60(.21)
NF_GRD	.91(.03)	.72(.04)	.82(.12)	.62(.20)
SVM	.97(.03)	.83(.01)	.99(.01)	.94(.08)
Logreg	.97(.03)	.85(.02)	.99(.01)	.92(.09)
CART	.88(.05)	.82(.02)	.93(.04)	.72(.17)
RF	.97(.03)	.83(.01)	.98(.01)	.92(.05)

Table 4: AUROC value comparisons over datasets

Self-Explaining AI

- What counts as an explanation?
  - understandable and relevant to user
  - mechanistic, but at a level relevant to user (not “nuts and bolts”)
- Requirements on AI model
  - a measure of mutual information between the explanation and the decision
  - an uncertainty on both the explanation and decision
  - a “warning system” which warns the user when the decision falls outside the domain of applicability of the system
    - e.g., warn if test data fall outside convex hull of the latent layer’s training data

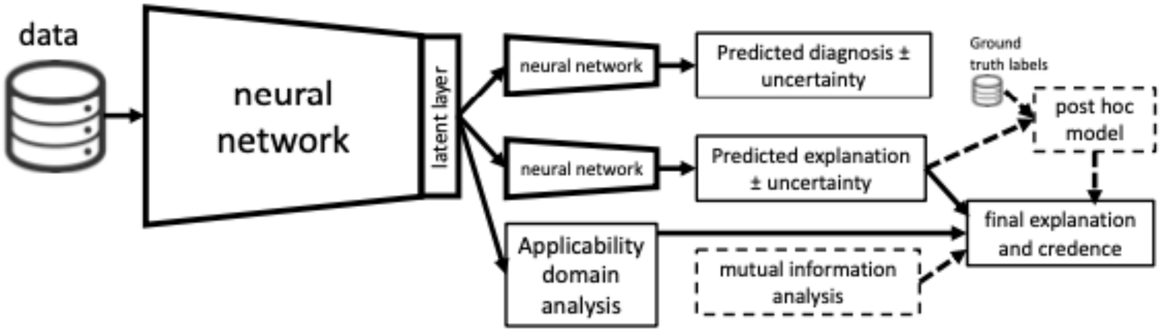

data
neural network
latent layer
neural network
Predicted diagnosis ± uncertainty
neural network
Predicted explanation ± uncertainty
Applicability domain analysis
mutual information analysis
Ground truth labels
post hoc model
final explanation and credence

Fig. 1. Sketch of a simple self-explaining AI system. Optional components are shown with dashed lines.

Elton, D. C. (2020, February 12). Self-explaining AI as an alternative to interpretable AI. Iclr 2020.

Extrapolation vs. Interpolation

- Recent evidence is that ANN models *interpolate* among training data points rather than learning general principles about the domain
  - Improving performance of models with vast numbers ($> 10^9$) of parameters

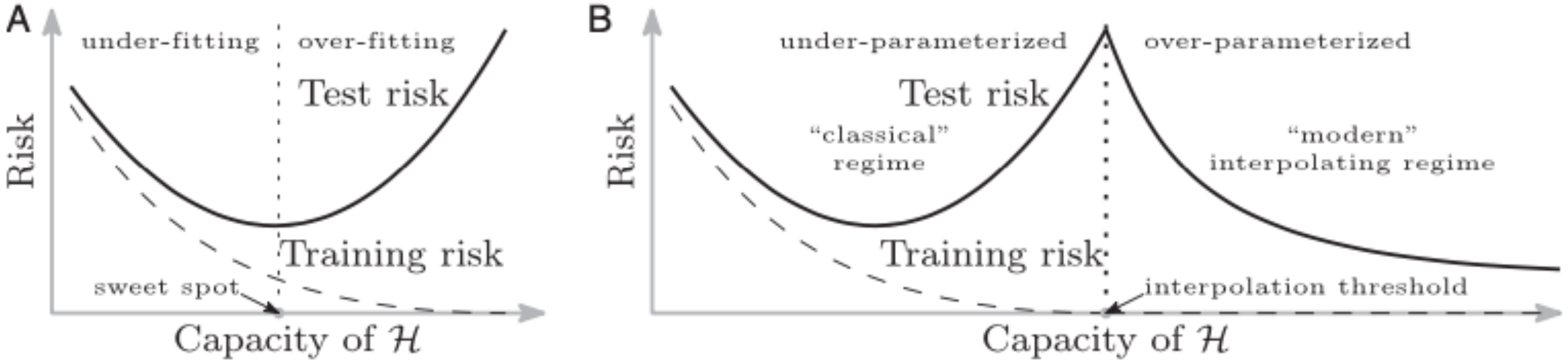
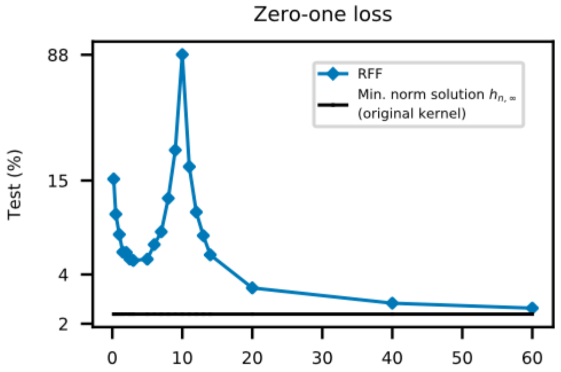


Fig. 1. Curves for training risk (dashed line) and test risk (solid line). (*A*) The classical U-shaped risk curve arising from the bias–variance trade-off. (*B*) The double-descent risk curve, which incorporates the U-shaped risk curve (i.e., the “classical” regime) together with the observed behavior from using high-capacity function classes (i.e., the “modern” interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

- Simplicity (smoothness) of the model class seems important
- Ex: Random Fourier Feature model on MNIST
  - x-axis is number of features



Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proceedings of the National Academy of Sciences of the United States of America, 116(32), 15849–15854. http://doi.org/10.1073/pnas.1903070116

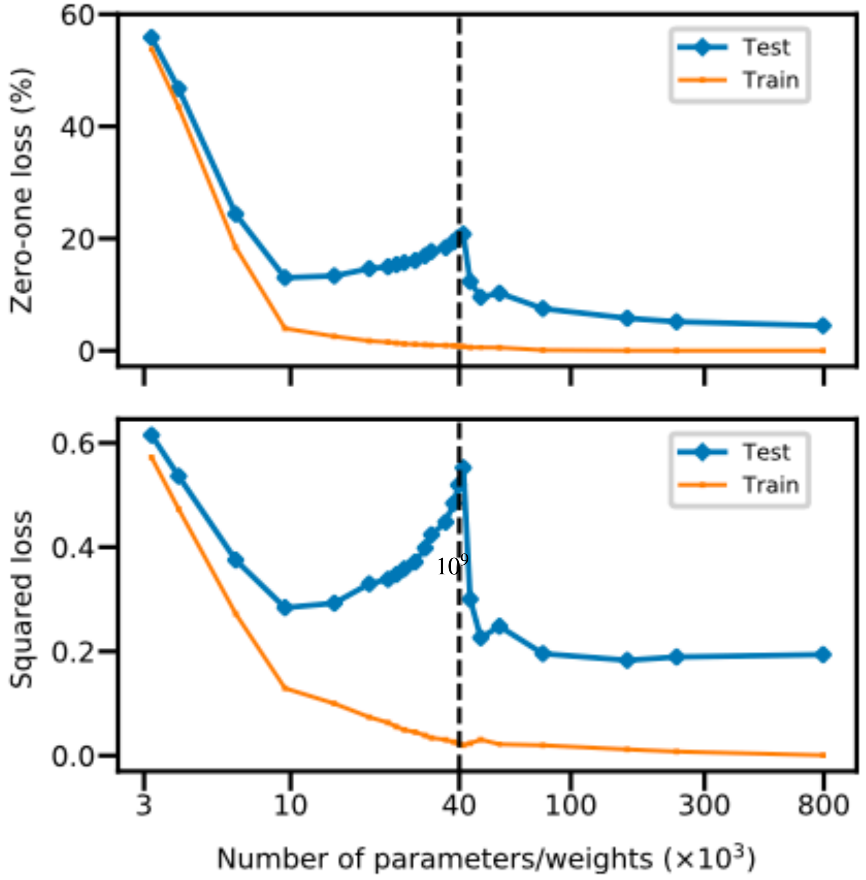


Fig. 3. Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of H hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d + 1) \cdot H + (H + 1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.

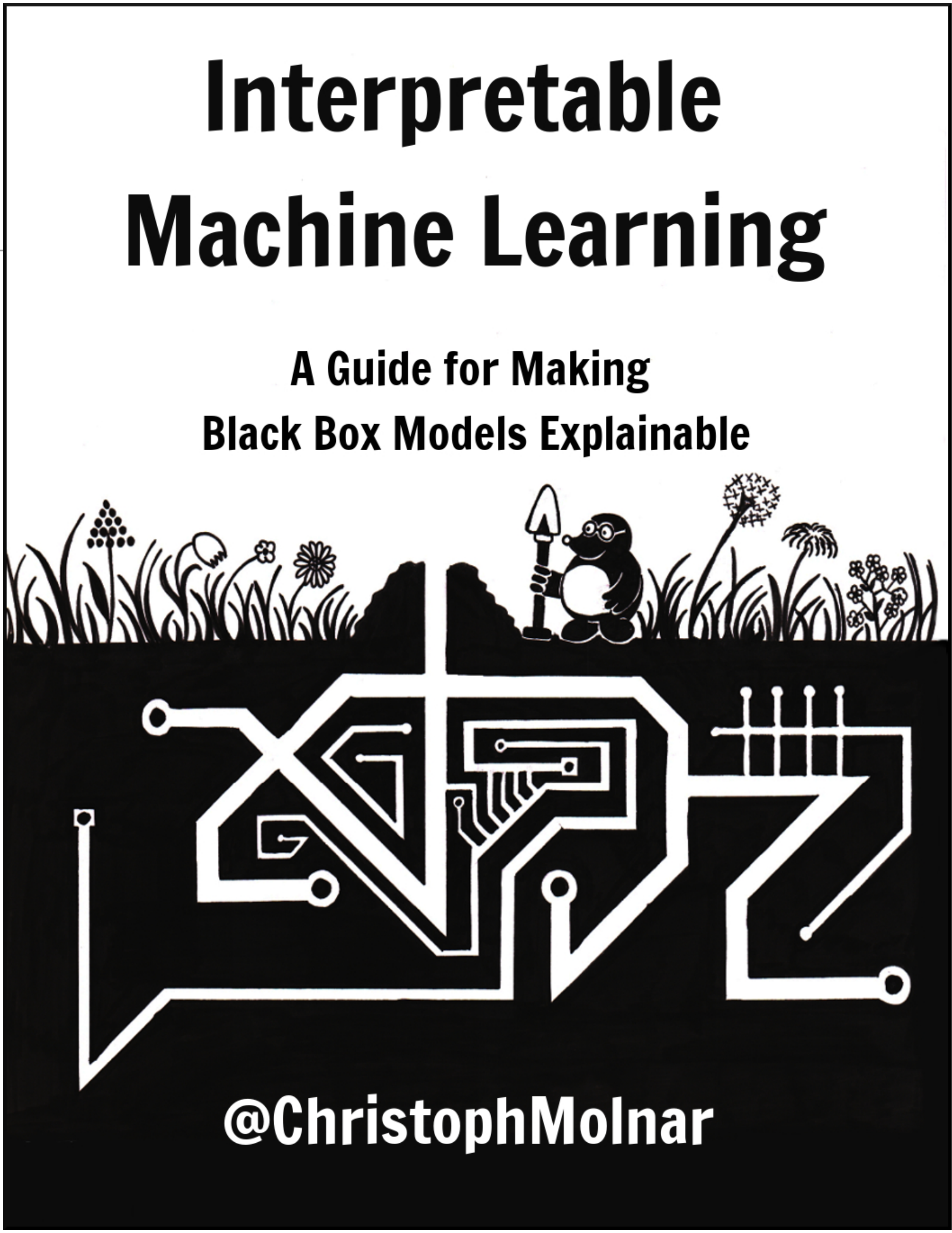
Achieving Interpretability *for Humans*

- **Why**: Incompleteness in problem formalization
  - Scientific understanding
  - Safety
  - Ethics
  - Indirect objectives
  - Competing objectives
- **How**: Methods
  - Application-grounded; in the context of its end-task
    - Compare to value of human-generated explanation to help other people
  - Human-grounded; simplified tasks
    - Choose better explanation; predict model outcome based on inputs and explanation; counterfactual (what input must change to change output)
  - Functionally-grounded; formal definition of interpretability
    - Posit certain classes of models to be interpretable; e.g., decision lists

Doshi-Velez, F., & Kim, B. (2017, February 27). Towards A Rigorous Science of Interpretable Machine Learning. Iclr 2020.

Useful source

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



Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

@ChristophMolnar