Example explanations extracted from their respective papers

1. **RxREN**

Table 2 Properties of six real datasets

Dataset	Total examples	Training examples	Testing examples	No. of Attributes	No. of Continuous Attributes	No. of Classes
iris	150	75	75	4	4	3
wbc	699	349	350	9	9	2
pid	768	390	378	8	8	2
hepatitis	155	81	74	19	6	2
creditg	1000	450	550	20	7	2
iono	351	165	166	34	34	2

Table 7 Extracted rules of 6 real datasets

Dataset	Rules	Ruleset size

iris	if ((sw>=2.0 and sw<=3.4) and (pl <=5.0) and (pw>=1.0 and pw <=1.7)) then
	class=”Iris-versicolor”
	else
	if (pl<=6.9 and pw>=1.4) then
	class=”Iris-virginica”
	else
	class=”Iris-setosa”
		3

wbc	if (Cell_Size_Uniformity <=4 and Bare_Nuclei<=5 and Normal_Nucleoli<=8) then
	class=”benign”
	else
	class=”malignant”
		2

pid	if (Plasma_glucose_concentration<=139) then
	class=“tested _negative”
	else
	class=“tested_positive”
		2

hepatitis	if (ALK_PHOSPHATE <=230) and (SGOT>=14 and SGOT<=420) then
	class =”LIVE”
	else
	class=”DIE”
		2

creditg	if (credit_history = ’existing paid’ and credit_amount <=12204) then
	class=“good”;
	else
	class=“bad”;
		2

iono	if (a01=1 and a03>=0.19 and a05>0 and a08> −1 and a07>= −0.23)
	class=“g”;
	else
	class=“b”;
		2

2. REFNE

Table 1
Data sets used in experiments

Data set	Abbr.	Size	Class	Number of attributes		
				Total	Categorical	Continuous
balance scale	*balance*	625	3	4	0	4
congressional voting records	*voting*	232	2	16	16	0
hepatitis	*hepatitis*	80	2	19	13	6
iris plant	*iris*	150	3	4	0	4
statlog australian credit approval	*credit-a*	690	2	15	9	6
statlog german credit	*credit-g*	1000	2	20	13	7

Table 5
A rule set extracted via REFNE in voting task

No.	Rule
1	physician-fee-freeze → democrat
2	¬education-spending → republican
3	¬handicapped-infants ∧ adoption-of-the-budget-resolution → republican
4	¬handicapped-infants ∧ aid-to-nicaraguan-contras → republican
5	¬water-project-cost-sharing ∧ adoption-of-the-budget-resolution → republican
6	water-project-cost-sharing ∧ mx-missile → republican
7	¬handicapped-infants ∧ ¬ superfund-right-to-sue → democrat
8	¬handicapped-infants ∧ ¬mx-missile → republican
9	¬water-project-cost-sharing ∧ religious-groups-in-schools → republican
10	water-project-cost-sharing ∧ ¬anti-satellite-test-ban → republican
11	water-project-cost-sharing ∧ ¬adoption-of-the-budget-resolution → democrat
12	water-project-cost-sharing ∧ ¬crime → democrat
13	handicapped-infants ∧ synfuels-corporation-cutback → republican
14	¬crime → democrat
15	handicapped-infants ∧ immigration → republican
16	¬religious-groups-in-schools ∧ export-administration-act-south-africa → democrat
17	anti-satellite-test-ban → republican
18	handicapped-infants → democrat

3. Using sampling and queries

Absent

4. **Using genetic algorithms**

- *Problem 1*: Object is a robot if (head shape = body shape) or (jacket color = red).

  Of the 432 possible examples, 124 randomly selected ones are used for the training set. There is no noise, i.e., no misclassifications.

- *Problem 2*: Object is a robot if *exactly* two of the six attributes are in their *first* value.

  Of the 432 possible examples, 169 randomly selected ones are used for the training set. There is again no noise in the output.

- *Problem 3*: Object is a robot if (jacket color = green and holding a sword) or (jacket color is not blue and body shape is not octagon).

  Of the 432 possible examples, 122 randomly selected ones are used for the training set and there is 5 % noise in the output, i.e., 6 examples are deliberately misclassified.

Table 3.1 Six different attributes and possible values of the attributes of robots

Attribute	**Values**			
	1	2	3	4
Head Shape	round	square	octagon	-
Body Shape	round	square	octagon	-
Is Smiling	yes	no	-	-
Holding	sword	balloon	flag	-
Jacket Color	red	yellow	green	blue
Has Tie	yes	no	-	-

5. **Rule generation**

were generated by R1 at zero noise.

```
R1:If
       petal-length is ≤ 2.7 ,
   Then
       it is setosa.
R2:If
       petal-length is >2.7  ≤ 5.0  and
       petal-width is >0.7   ≤ 1.6 ,
   Then
       it is versicolor.
R3:If
       petal-length is >5.0 ,
   Then
       it is viginica.
R4:If
       petal-width is >1.6 ,
   Then
       it is viginica.
R5:If
       septal-width is >3.1  and
       petal-length is >2.7  ≤ 5.0 ,
   Then
       it is versicolor.
```

6. Extracting refined rules

Table 8. Promoter rules extracted by SUBSET.

```
Promoter :- Minus35, Minus10.

Minus35  :- Minus35b, Minus35d.
Minus35  :- Minus35a, Minus35b.
Minus35  :- Minus35a, Minus35d.

Minus10  :- @-14 '-ATA----'.
Minus10  :- @-14 '-T--A-A-'.
Minus10  :- @-14 '--A-T-T-'.
Minus10  :- @-14 '-TA-----'.
Minus10  :- @-14 '-T------'.
Minus10  :- @-14 '---A---T'.
Minus10  :- @-14 '--T----T'.
Minus10  :- @-14 '--TA----'.
Minus10  :- @-14 'T-T-A---'.
Minus10  :- @-14 '-AT--T--'.
Minus10  :- @-14 '--TAA---'.
Minus10  :- @-14 '--TA-T--'.

Minus35  :- @-37 '-T-T--A'.
Minus35  :- @-37 '-TT---A'.
Minus35  :- @-37 '---G-CA'.
Minus35  :- @-37 '--T--C-'.
Minus35  :- @-37 '-T---CA'.

Minus35a :- @-37 'CT--A--'.
Minus35a :- @-37 'C--G-C-'.
Minus35a :- @-37 'C-T--C-'.
Minus35a :- @-37 'CT---C-'.
Minus35a :- @-37 'C---AC-'.

Minus35b :- @-37 '---GA--'.
Minus35b :- @-37 '--T--CA'.
Minus35b :- @-37 '-T---C-'.
Minus35b :- @-37 '---G-CA'.
Minus35b :- @-37 '----ACA'.

Minus35d :- @-37 '-TT--C-'.
Minus35d :- @-37 '-T-G-C-'.
Minus35d :- @-37 '--T-AC-'.
Minus35d :- @-37 '---GAC-'.
Minus35d :- @-37 '-T--AC-'
```

This abbreviated set of rules extracted by SUBSET has a test set accuracy of 75%. Sets whose statistics are reported previously in this section contain about 300 rules.

Table 7. Promoter rules extracted by MOFN.

```
Promoter :- Minus10, Minus35.

Minus10 :- 10 <  3.8 * nt(@-14 '-TA-A-T-') +
                3.0 * nt(@-14 '-G--C---') +
               -3.0 * nt(@-14 '-A--T---').

Minus10 :-           2 of @-14 '--CA---T' and
                 not 1 of @-14 '--RB---S').

Minus10 :- 10 <  3.0 * nt(@-14 '-TAT--T-') +
                1.8 * nt(@-14 '----GA--').

Minus10 :-  1 <  3.5 * nt(@-14 'TAWAAY--') +
               -1.7 * nt(@-14 '-T--TG--') +
               -2.2 * nt(@-14 'CSSK-A--').

Minus35 :- 10 <  4.0 * nt(@-37 '-TTGAT--') +
                1.5 * nt(@-37 '---TCC--') +
               -1.5 * nt(@-37 '-RGAGG--').

Minus35 :- 10 <  5.1 * nt(@-37 '-T-G--A-') +
                3.1 * nt(@-37 '-GT-----') +
               -1.9 * nt(@-37 '-CGW----') +
               -3.1 * nt(@-37 '-A----C-').

Minus35 :-           3 of @-37 'C-TGAC--'.

Minus35 :-               @-37 '-TTG-CA-'.
```

See table 5 for meanings of letters other than A, G, T, and C. "`nt()`" returns the number of named antecedents that match the given sequence. So, `nt(@-14 '---C--G--')` would return 1 when matched against the sequence `@-14 'AAACAAAAA'`.

Table 9. The domain and correct theories for the MONKS problems.

Feature Name		Values
head-shape	$\in$	{round, square, octagon}
body-shape	$\in$	{round, square, octagon}
smiling	$\in$	{yes, no}
holding	$\in$	{sword, balloon, flag}
jacket-color	$\in$	{red, yellow, green, blue}
has-tie	$\in$	{yes, no}

The features and values for the MONKS problems.

```
if (head-shape round)   and (body-shape round)   then MONK.
if (head-shape square)  and (body-shape square)  then MONK.
if (head-shape octagon) and (body-shape octagon) then MONK.
if (jacket-color red)                            then MONK.
```

The correct theory for the first MONKS problem.

```
if exactly two of {(head-shape   round) (body-shape round)
                   (smiling      yes)   (holding    sword)
                   (jacket-color red)   (has-tie    yes)}
then MONK.
```

The correct theory for the second MONKS problem.

```
if two of {(head-shape   round) (body-shape round)
           (smiling      yes)   (holding    sword)
           (jacket-color red)   (has-tie    yes)}
   AND not three of {(head-shape   round) (body-shape round)
                     (smiling      yes)   (holding    sword)
                     (jacket-color red)   (has-tie    yes)}
then MONK.
```

Reformulation of the second MONKS problem.

7. Extracting Rules from distributed DNN

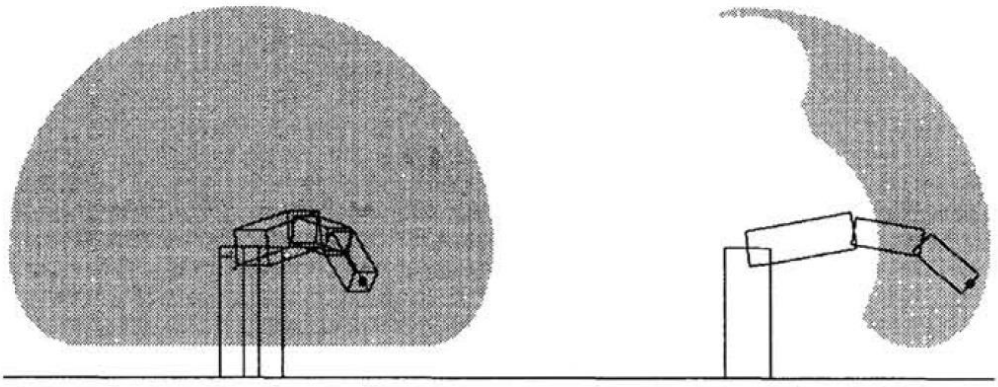


Figure 3: **A single rule, extracted from the network.** (a) Front view. (b) Two-dimensional side view. The grey area indicates safe positions for the tip of the manipulator.

8. FERNN

Rule0: If (body_shape $\neq$ octagon) and (jacket_color $\neq$ blue) then monk.

Rule2: If (holding = sword) and (jacket_color = green) then monk.

9. Symbolic Interpretation

Table 3: Rules Extracted from network *"Iris-Cont"* by Full-RE technique.

Rule No.	Rule Body	Iris Class	Certainty Factor(cf)	Soundness Measure	Completeness Measure	False-Alarm Measure
1	If $I_3 \leq 2.1$	Setosa	0.99	50/50	50/50	0/150
2	If $I_3 \leq 5.1$ and $I_4 \leq 1.7$	Versicolor	0.97	49/50	49/50	3/150
3	If $I_3 \geq 4.8$	Virginica	0.98	47/50	47/50	1/150
Overall Performance					146/150	4/150
%					97.33%	2.67%

Table 7: Rules extracted from network *"Cancer-Cont"* by Full-RE technique.

Rule No.	Rule Body	B-Cancer Class	Certainty Factor	Soundness Measure	Completeness Measure	False-Alarm Measure
1	If $X_1 < 8$ and $X_3 < 3$	Benign	0.96	394/444	394/444	5/683
2	If $X_2 \geq 2$ and $X_7 \geq 3$	Malignant	0.83	227/239	223/239	18/683
3	If $X_1 < 8$ and $X_7 < 3$	Benign	0.75	300/444	27/444	1/683
4	If $X_1 \geq 8$	Malignant	0.89	123/239	9/239	1/683
5	If $X_1 < 8$ and $X_2 < 2$	Benign	0.79	369/444	4/444	1/683
Total For Benign Rules					425/444	7/683
%					95.72%	1.02%
Total For Malignant Rules					232/239	19/683
%					97.07%	2.78%
Overall Performance					657/683	26/683
%					96.19%	3.81%

10. Extracting Rules from Trained NNs (trepan)

Absent

11. Extracting comprehensible models

Absent

12. ICU Outcome Prediction

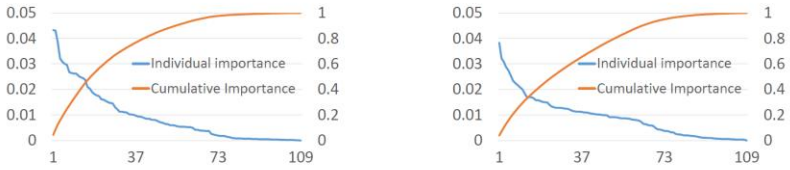


Figure 4: Individual (with left y-axis) and cumulative (with right y-axis) feature importance for MOR (top) and VFD (bottom) tasks. x-axis: sorted features.

Figure 5: Feature importance for static features and temporal features on each day for two tasks.

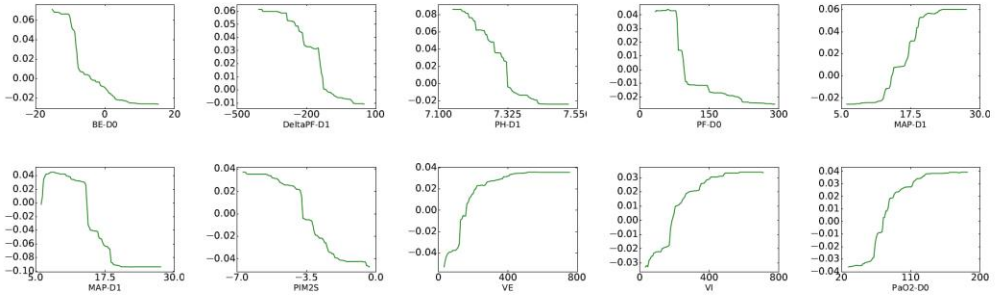


Figure 6: One-way partial dependence plots of the top features from GBTmimic for MOR (top) and VFD (bottom) tasks. x-axis: variable value; y-axis: dependence value.

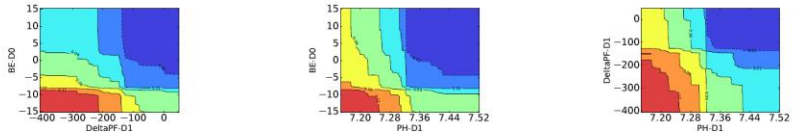


Figure 7: Pairwise partial dependence plots of the top features from GBTmimic for MOR (top) and VFD (bottom) tasks. Red: positive dependence; Blue: negative dependence.

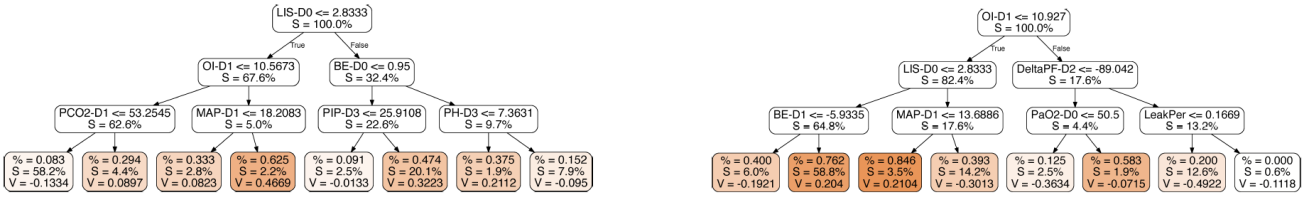
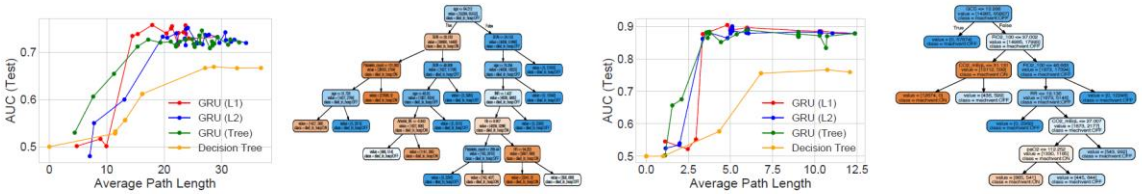


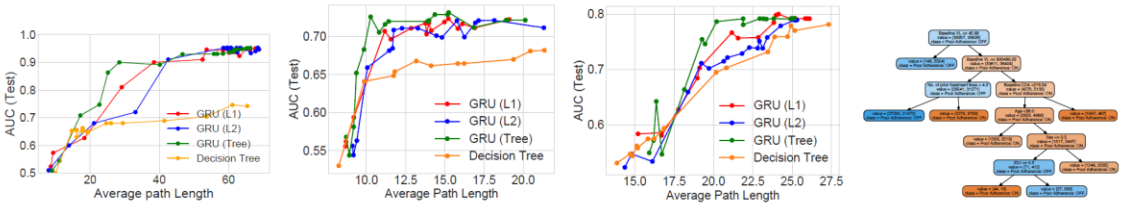
Figure 8: Sample decision trees from best GBTmimic models for MOR (top) and VFD (bottom) tasks. % and leaf color: class distribution for samples belong to that node; S: # of samples to that node; V: prediction value of that node.

13. Tree Regularization



(a) In-Hospital Mortality (b) In-Hospital Mortality (c) Mechanical Ventilation (d) Mechanical Ventilation

Figure 4: *Sepsis task:* Study of different regularization techniques for GRU model with 100 states, trained to jointly predict 5 binary outcomes. Panels (a) and (c) show AUC vs. average path length for 2 of the 5 outcomes (remainder in the supplement); in both cases, tree-regularization provides higher accuracy in the target regime of low-complexity decision trees. Panels (b) and (d) show the associated decision trees for $\lambda = 2\,000$; these were found by clinically interpretable by an ICU clinician (see main text).



(a) TIMIT Stop Phonemes (b) HIV: CD4$^+$ $\leq$ 200 cells/ml (c) HIV Therapy Adherence (d) HIV Therapy Adherence

Figure 5: *TIMIT and HIV tasks:* Study of different regularization techniques for GRU model with 75 states. Panels (a)-(c) are tradeoff curves showing how AUC predictive power and decision-tree complexity evolve with increasing regularization strength under L1, L2 or tree regularization on both TIMIT and HIV tasks. The GRU is trained to jointly predict 15 binary outcomes for HIV, of which 2 are shown here in Panels (b) - (c). The GRU's decision tree proxy for HIV Adherence is shown in (d).

14. Distilling a Neural Network

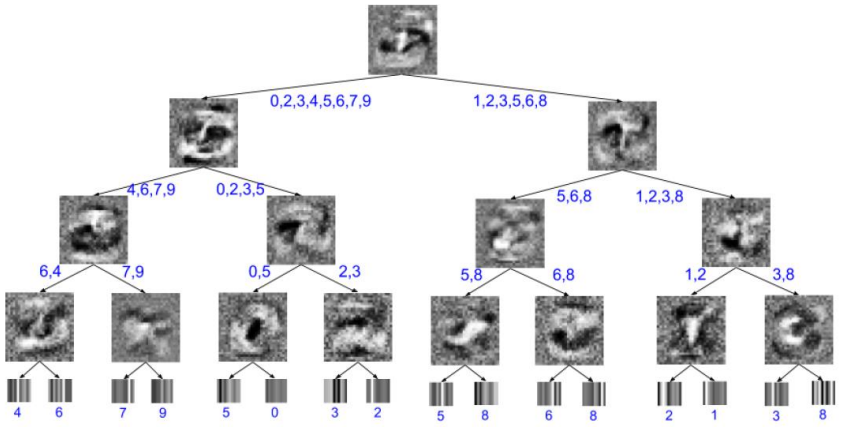


Fig. 2: This is a visualization of a soft decision tree of depth 4 trained on MNIST. The images at the inner nodes are the learned filters, and the images at the leaves are visualizations of the learned probability distribution over classes. The final most likely classification at each leaf, as well as the likely classifications at each edge are annotated. If we take for example the right most internal node, we can see that at that level in the tree the potential classifications are only 3 or 8, thus the learned filter is simply learning to distinguish between those two digits. The result is a filter that looks for the presence of two areas that would join the ends of the 3 to make an 8.

15. Extracting decision trees

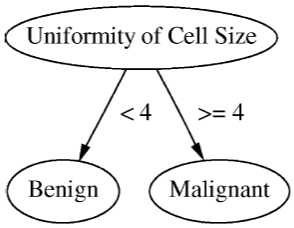


Fig. 4. Decision tree for cancer data using the proposed method (three nodes. two leaves, 94.74% accuracy).

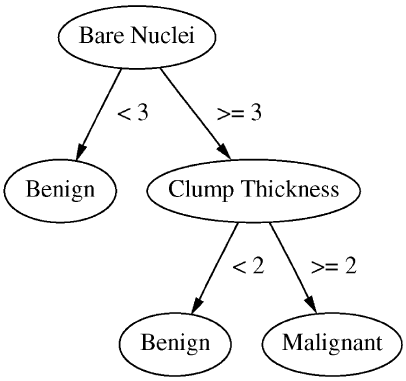


Fig. 5. Decision tree for cancer data using the proposed method (five nodes, three leaves, 93.27% accuracy).

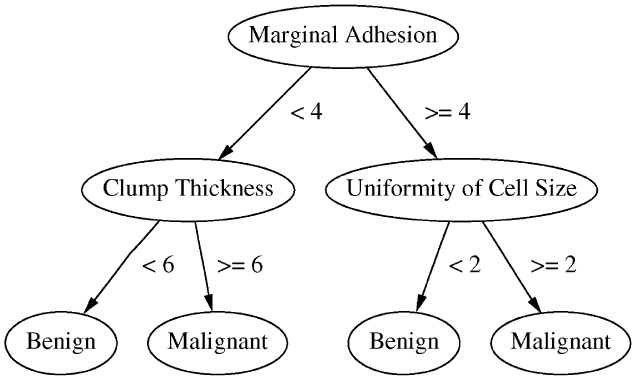


Fig. 6. Decision tree for cancer data using the proposed method (seven nodes, four leaves, 93.57% accuracy).

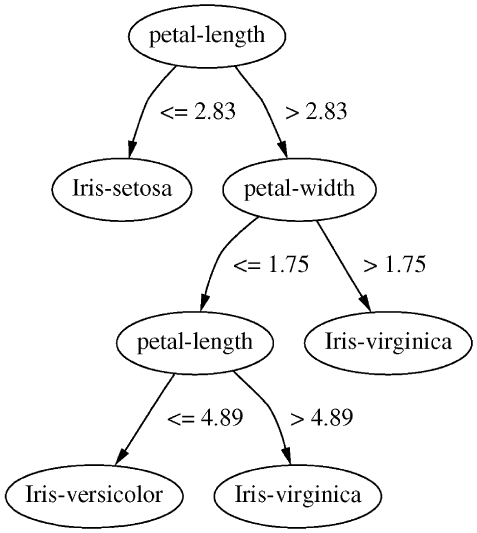


Fig. 3. Decision tree for Iris data using the proposed method (seven nodes, four leaves, 98% accuracy).

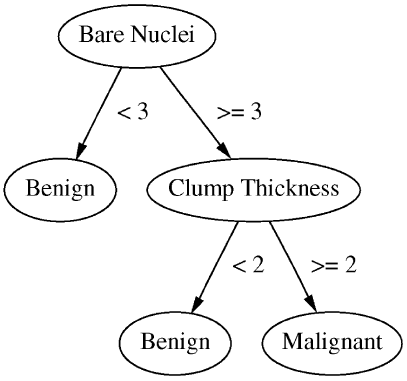


Fig. 5. Decision tree for cancer data using the proposed method (five nodes, three leaves, 93.27% accuracy).

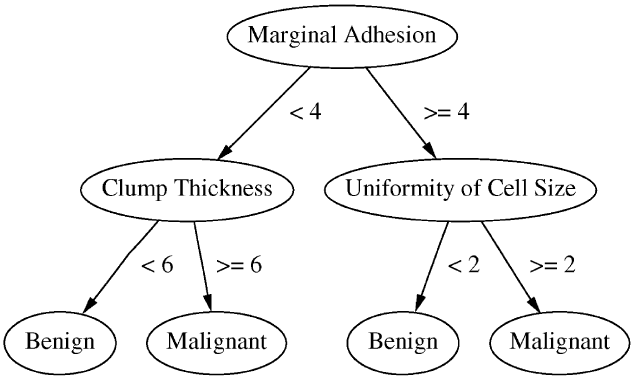


Fig. 6. Decision tree for cancer data using the proposed method (seven nodes, four leaves, 93.57% accuracy).

16. Feature-space partitioning

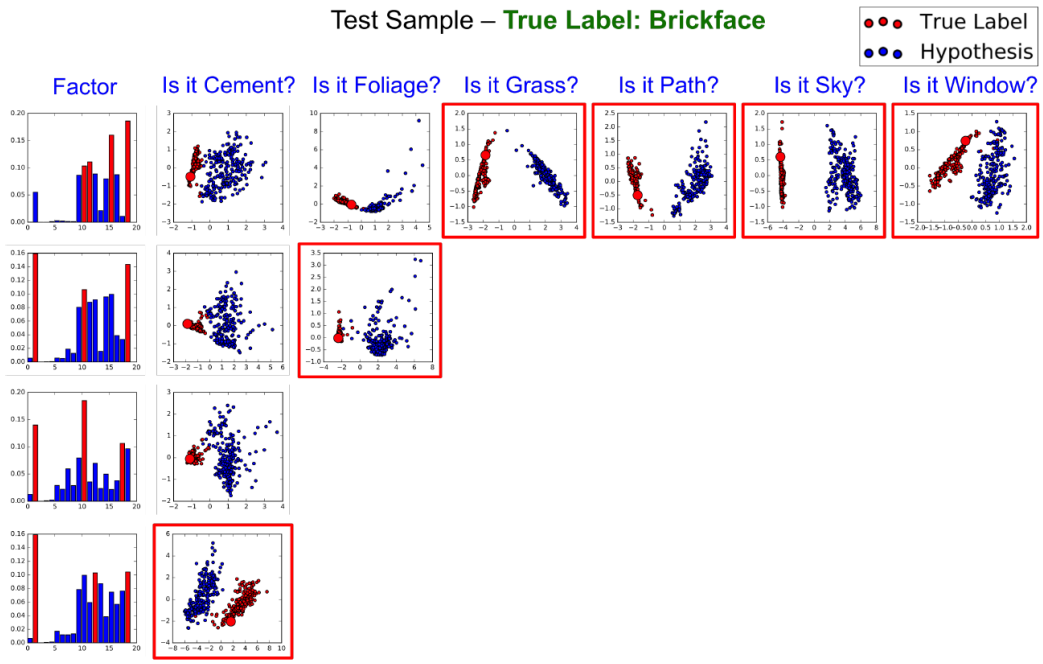


Figure 2: *Treeview* visualization for a sample which is correctly classified by the neural network.

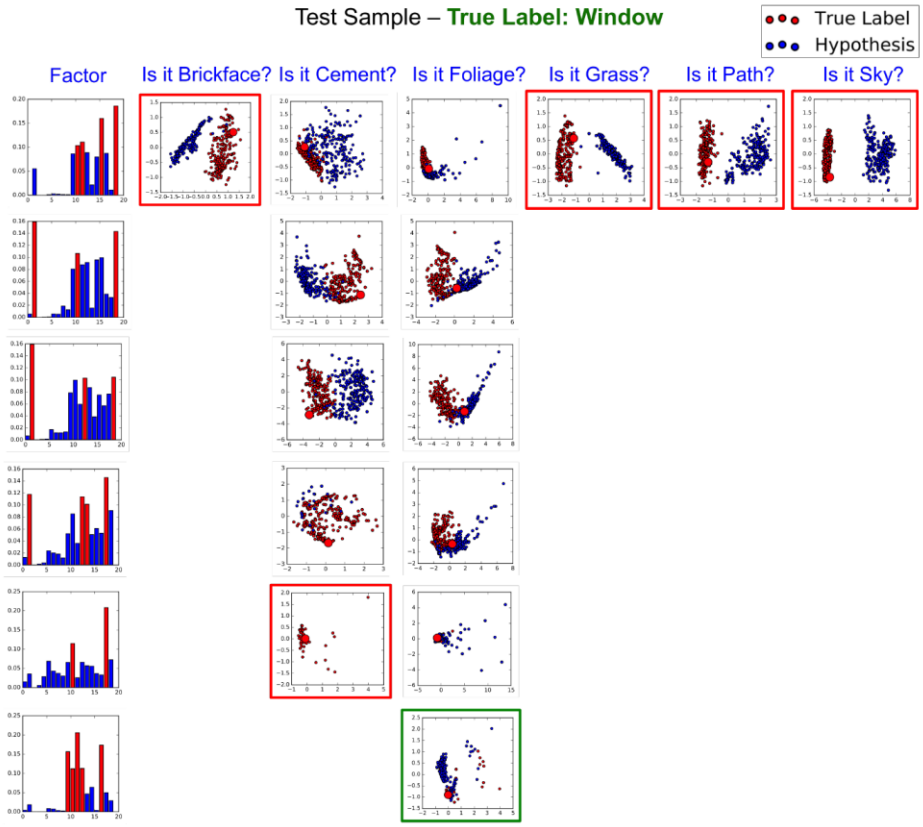


Figure 3: *Treeview* visualization for a sample which is wrongly classified by the neural network.

17. Deepred–rule extraction

Absent

18. ANN-DT

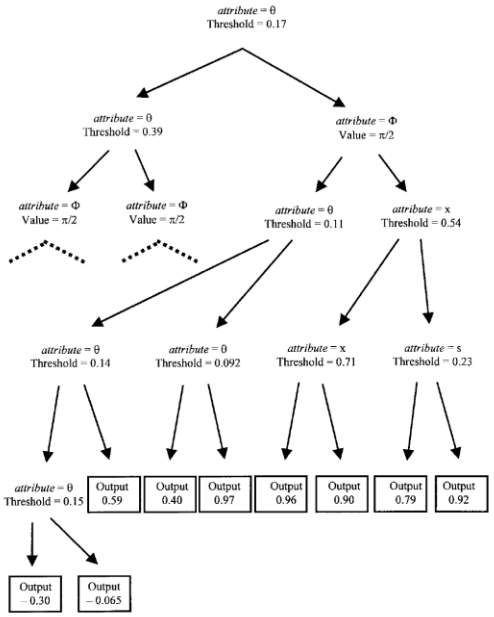


Fig. 4. The decision tree extracted by the ANN-DT(s) algorithm from the trained neural network for case study 1 with $c = 0.3$ using 1000 points to sample the neural network. If attribute a < Threshold the right subtree applies, else the left subtree is valid.

19. Decision tree induction

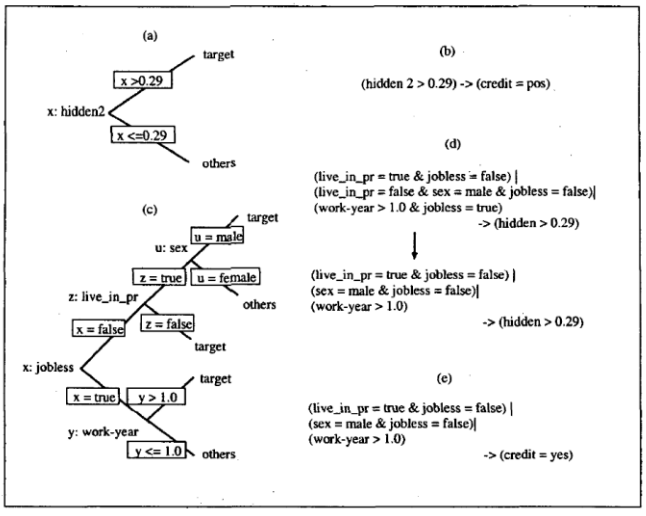


Figure 3: Results on The Credit Data. The hidden-output tree (a), the intermediate-rules (b), the input-hidden tree (c), the input-rules before and after simplification (d), and the total rules (e).

20. Distilling the Knowledge

Absent

21. Deep Taylor decomposition

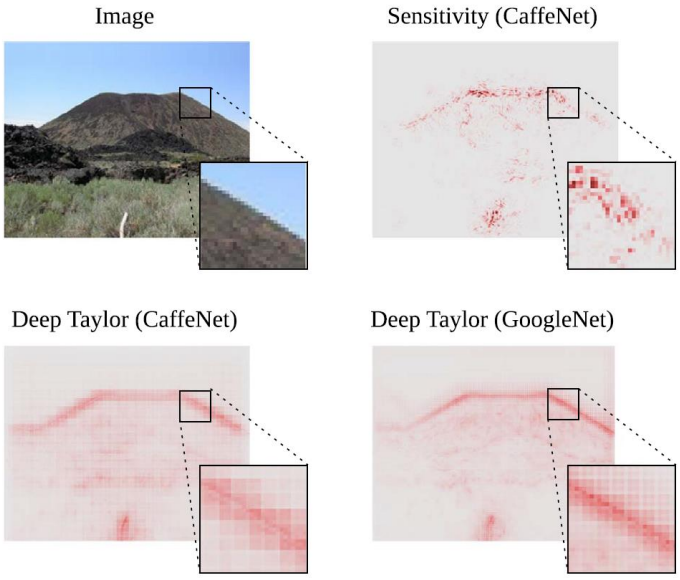


Fig. 8. Image with ILSVRC class "volcano", displayed next to its associated heatmaps and a zoom on a region of interest.

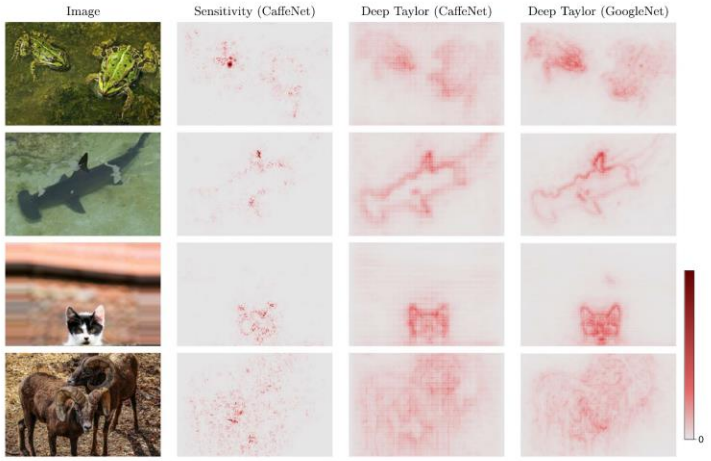


Fig. 7. Images of different ILSVRC classes ("frog", "shark", "cat", and "sheep") given as input to a deep network, and displayed next to the corresponding heatmaps. Heatmap scores are summed over all color channels of the image.

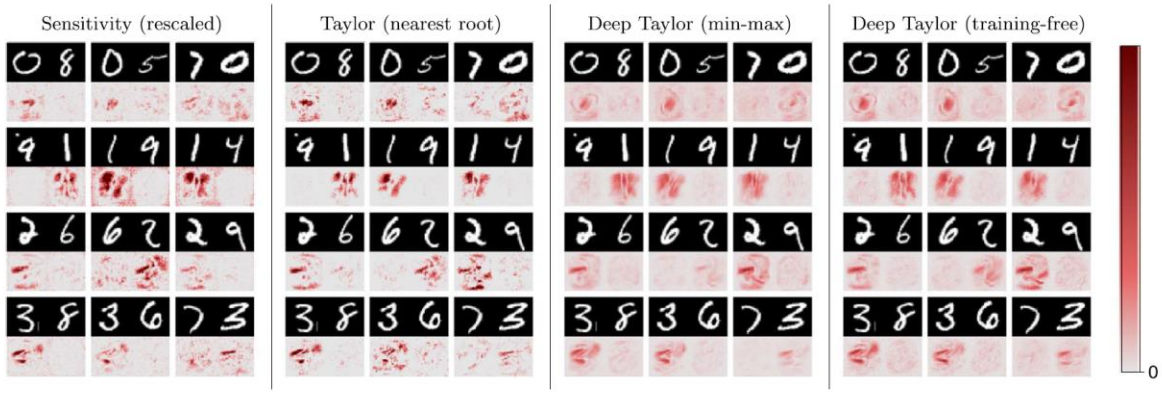


Fig. 5. Comparison of heatmaps produced by various decompositions and relevance models. Each input image is presented with its associated heatmap.

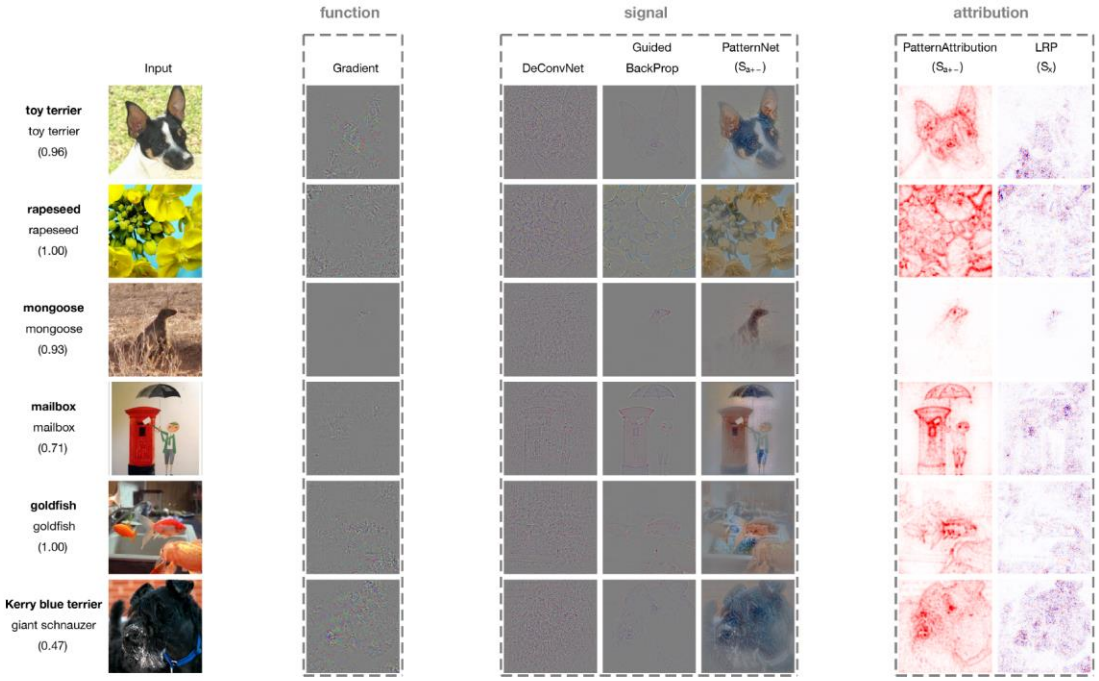
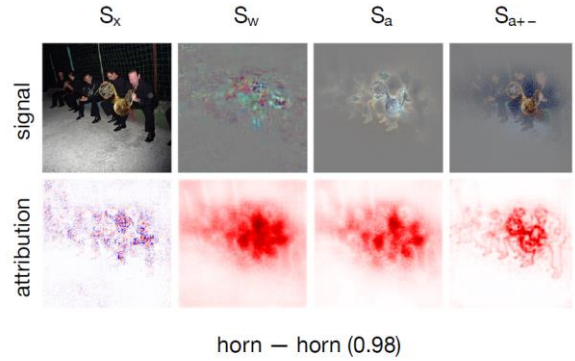


Figure 5: Visualization of random images from ImageNet (validation set). In the leftmost shows column the ground truth, the predicted label and the classifier's confidence. Methods should only be compared within their group. PatternNet, Guided Backprop, DeConvNet and the Gradient (saliency map) are back-projections to input space with the original color channels. They are normalized using $x_{norm} = \frac{x}{2\,\mathrm{max}|x|} + \frac{1}{2}$ to maximize contrast. LRP and PatternAttribution are heat maps showing pixel-wise contributions. Best viewed in electronic format (zoomed in). The supplementary contains more samples.



(b) Top: signal. Bottom: attribution. For the trivial estimator S_x the original input is the signal. This is not informative w.r.t. how the network operates.

23. Propagating Activation Differences

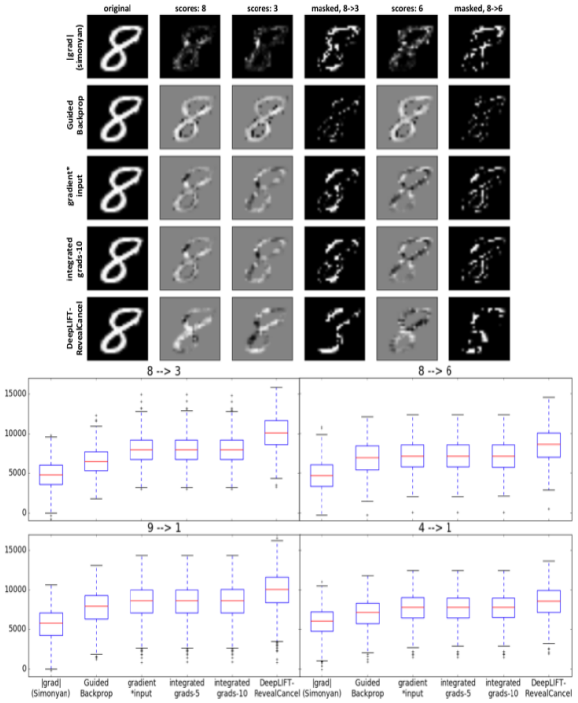


Figure 4. **DeepLIFT with the RevealCancel rule better identifies pixels to convert one digit to another.** Top: result of masking pixels ranked as most important for the original class (8) relative to the target class (3 or 6). Importance scores for class 8, 3 and 6 are also shown. The selected image had the highest change in log-odds scores for the 8→6 conversion using gradient*input or integrated gradients to rank pixels. Bottom: boxplots of increase in log-odds scores of target vs. original class after the mask is applied, for 1K images belonging to the original class in the testing set. "Integrated gradients-n" refers to numerically integrating the gradients over n evenly-spaced intervals using the midpoint rule.

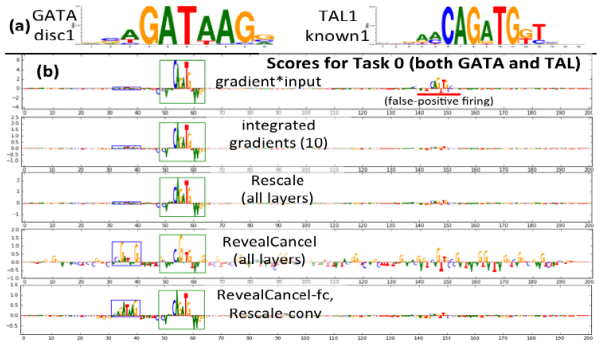


Figure 6. **RevealCancel highlights both TAL1 and GATA1 motifs for Task 0**. (a) PWM representations of the GATA1 motif and TAL1 motif used in the simulation (b) Scores for example sequence containing both TAL1 and GATA1 motifs. Letter height reflects the score. Blue box is location of embedded GATA1 motif, green box is location of embedded TAL1 motif. Red underline is chance occurrence of weak match to TAL1 (CAGTTG instead of CAGATG). Both TAL1 and GATA1 motifs should be highlighted for Task 0. RevealCancel on only the fully-connected layer reduces noise compared to RevealCancel on all layers.

24. Explain neural network Classification

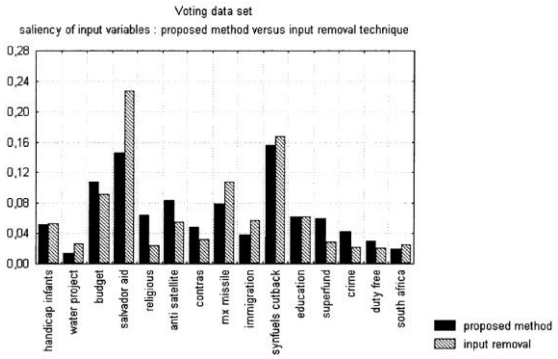


Fig. 4. The saliency according to input removal technique and the saliency according to the proposed method. Each value corresponds to the mean value on 10 trainings.

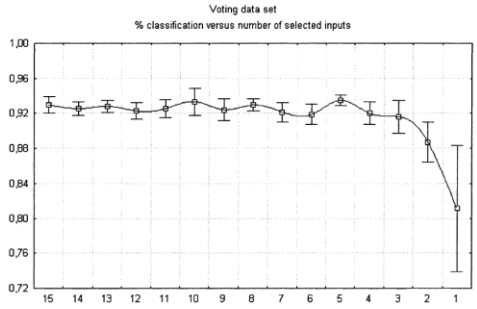


Fig. 5. Results for the variable selection tasks for the voting dataset. Each value corresponds to the mean value on 10 trainings. The error bar shows the 95% confidence interval.

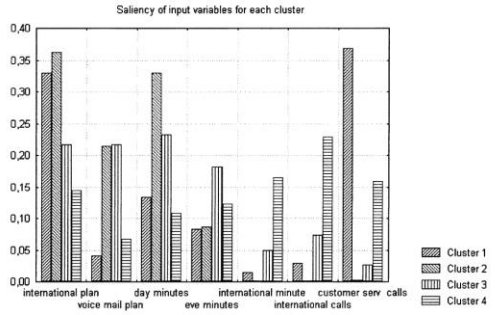


Fig. 8. Saliency of input variables for each cluster. The saliency measurements are normalised by cluster for the comparison.

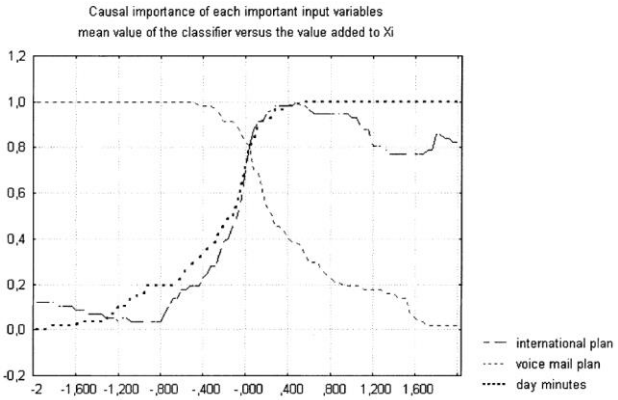


Fig. 9. Causal importance of the three input variables which have the largest saliency for the cluster 3.

25. Axiomatic Attribution

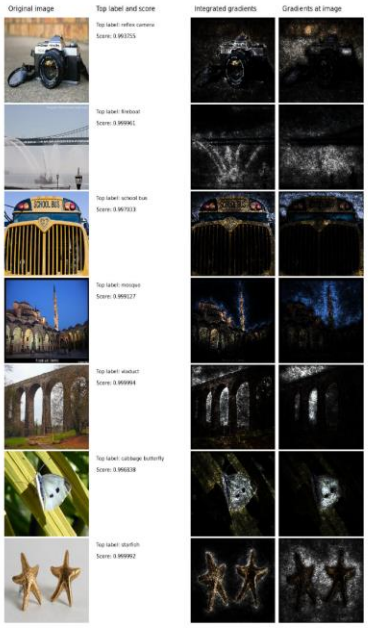


Figure 2. **Comparing integrated gradients with gradients at the image.** Left-to-right: original input image, label and softmax score for the highest scoring class, visualization of integrated gradients, visualization of gradients*image. Notice that the visualizations obtained from integrated gradients are better at reflecting distinctive features of the image.

Figure 3. **Attribution for Diabetic Retinopathy grade prediction from a retinal fundus image.** The original image is show on the left, and the attributions (overlayed on the original image in gray scaee) is shown on the right. On the original image we annotate lesions visible to a human, and confirm that the attributions indeed point to them.

how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Figure 4. **Attributions from question classification model.** Term color indicates attribution strength—Red is positive, Blue is negative, and Gray is neutral (zero). The predicted class is specified in square brackets.

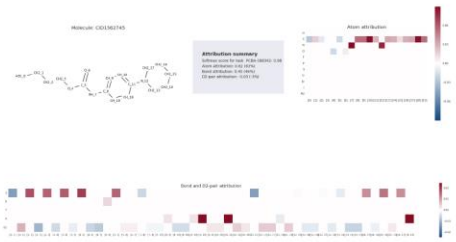


Figure 6. **Attribution for a molecule under the W2N2 network (Kearnes et al., 2016).** The molecules is active on task PCBA-58432.

Figure 5. **Attributions from a language translation model.** Input in English: "good morning ladies and gentlemen". Output in German: "Guten Morgen Damen und Herren". Both input and output are tokenized into word pieces, where a word piece prefixed by underscore indicates that it should be the prefix of a word.

26. Iterative Debugging

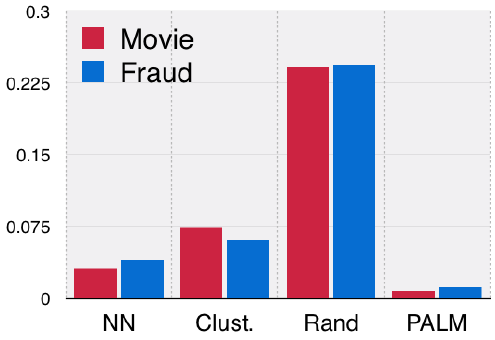


Figure 2: PALM isolates relevant training data more effectively than baselines.

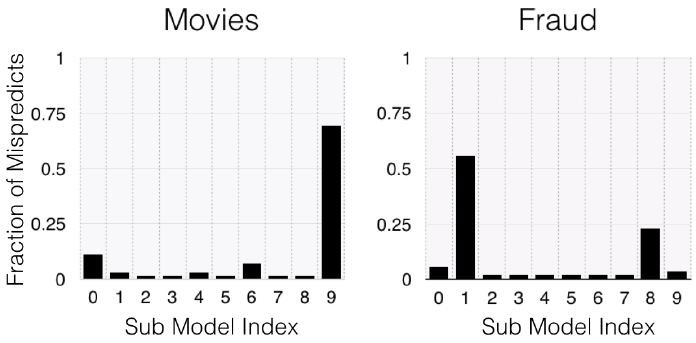


Figure 3: On two datasets, we show how mispredictions concentrate around specific submodels. This means there are specific regions of the feature-space most associated with mispredictions.

27. Local Explanations

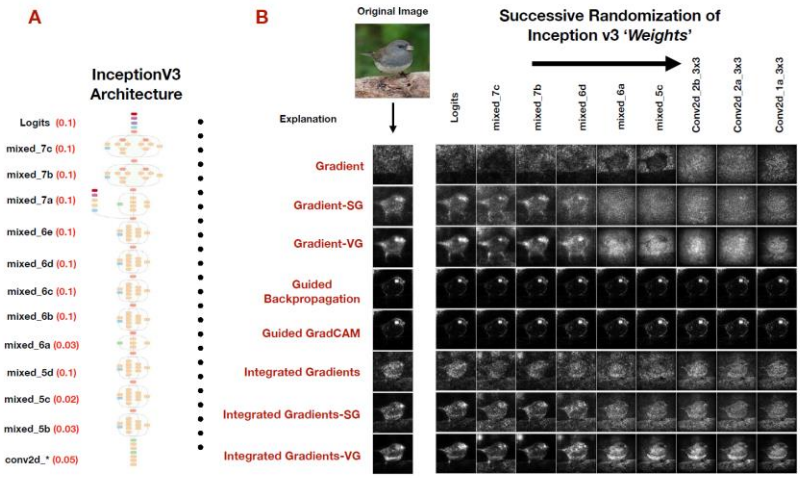


Figure 1: **Change in explanations for various methods as each successive inception block is randomized, starting from the logits layer. A**: Inception v3 architecture along with the names of the different blocks. The number in the parenthesis is the top-1 accuracy of the Inception model on a test set of 1000 images after randomization up that block. Initial top-1 accuracy for this class of images was 97 percent. Conv2d* refers collectively to the last 5 convolutional layers. **B-Left**: Shows the original explanations for the Junco bird in the first column as well as the label for each explanation type shown. **B-Right:** Shows successive explanations as each block is randomized. We show images for 9 blocks of randomization. Coordinate (Gradient, mixed7b) shows the gradient explanation for the network in which the top layers starting from Logits up to mixed7b have been reinitialized. The last column corresponds to a network where all the weights have been completely reinitialized. See Appendix for more examples.

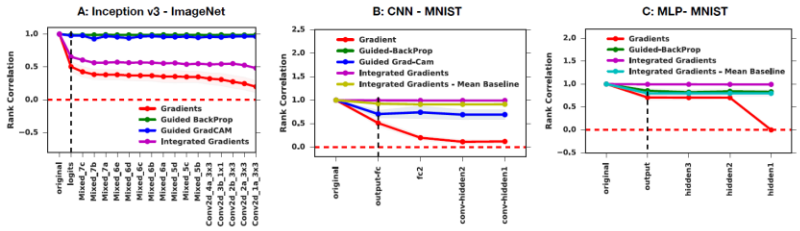


Figure 2: **Successive reinitialization starting from top layers for (A) Inception v3 on ImageNet, (B) CNN on MNIST, and (C) MLP on MNIST.** In all plots, y axis is the rank correlation between original explanation and the randomized explanation derived for randomization up to that layer or block (inception), while the x axis corresponds to the layers/blocks of the DNN starting from the output layer. The black dashed line indicates where successive randomization of the network begins, which is at the first layer. A: Rank correlation plot for Inception v3 trained on ImageNet. B: Rank correlation explanation similarity plot for a 3 hidden-layer CNN on MNIST. C: Rank correlation plot for a 3-hidden layer feed forward network on MNIST.

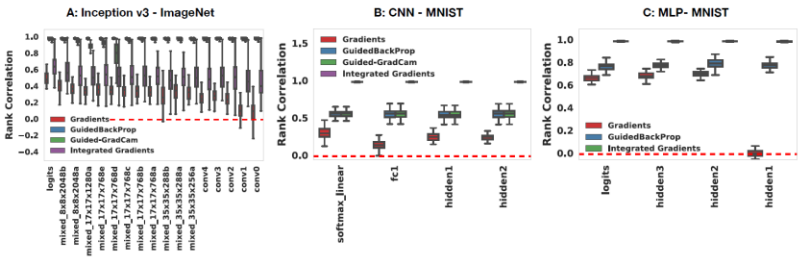


Figure 3: **Independent reinitialization of each layer (A) Inception v3 on ImageNet, (B) CNN on MNIST, and (C) MLP on MNIST.** In all plots, y axis is the rank correlation between original explanation and the randomized explanation derived for independent randomization of that layer or block (inception), while the x axis corresponds to the layers/blocks of the DNN starting from the output layer. The red dashed line indicates zero rank correlation. A: Rank correlation plot for InceptionV3 trained on ImageNet. B: Rank correlation explanation similarity plot for a 3 hidden-layer CNN on MNIST. C: Rank correlation plot for a 3-hidden layer feed forward network on MNIST.

28. Deep k-Nearest Neighbors

Fig. 5: **Debugging ResNet model biases**—This illustrates how the DkNN algorithm helps to understand a bias identified by Stock and Cissé [105] in the ResNet model for ImageNet. The image at the bottom of each column is the test input presented to the DkNN. Each test input is cropped slightly differently to include (left) or exclude (right) the football. Images shown at the top are nearest neighbors in the predicted class according to the representation output by the last hidden layer. This comparison suggests that the "basketball" prediction may have been a consequence of the ball being in the picture. Also note how the white apparel color and general arm positions of players often match the test image of Barack Obama.

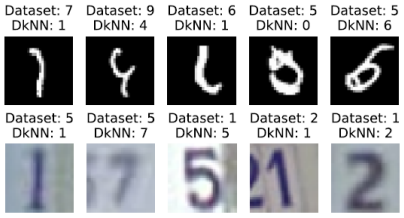
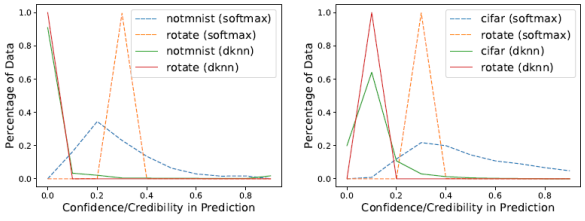
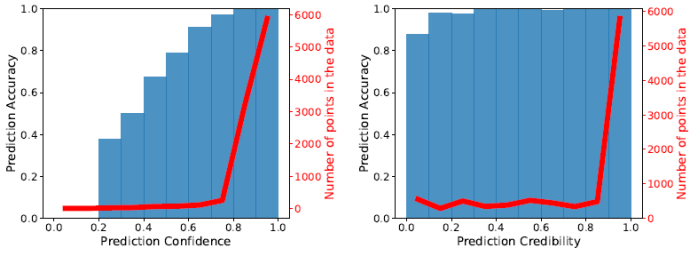


Fig. 3: **Mislabeled inputs from the MNIST (top) and SVHN (bottom) test sets**: we found these points by searching for inputs that are classified with strong credibility by the DkNN in a class that is different than the label found in the dataset.



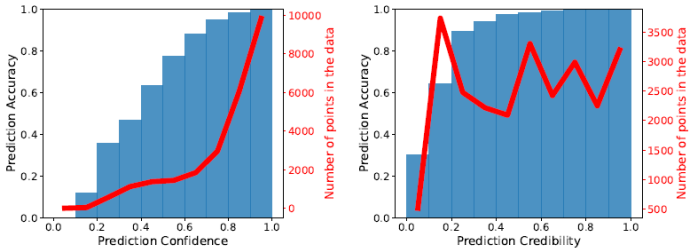
(a) MNIST (b) SVHN

Fig. 4: **DkNN credibility vs. softmax confidence on out-of-distribution test data**: the lower credibility of DkNN predictions (solid lines) compared to the softmax confidence (dotted lines) is desirable here because test inputs are not part of the distribution on which the model was trained—they are from another dataset or created by rotating inputs.



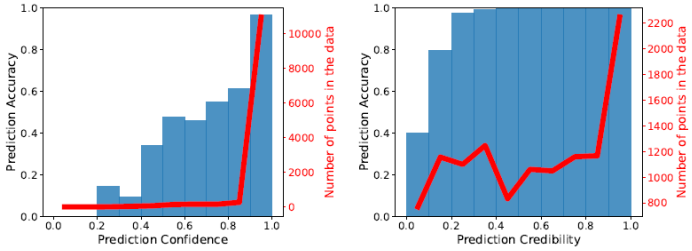
(a) Softmax - MNIST test set

(b) DkNN - MNIST test set



(c) Softmax - SVHN test set

(d) DkNN - SVHN test set



(e) Softmax - GTSRB test set

(f) DkNN - GTSRB test set

Fig. 2: **Reliability diagrams of DNN softmax confidence (left) and DkNN credibility (right) on test data**—bars (left axis) indicate the mean accuracy of predictions binned by credibility; the red line (right axis) illustrates data density across bins. The softmax outputs high confidence on most of the data while DkNN credibility spreads across the value range.

29. Concept Activation Vectors

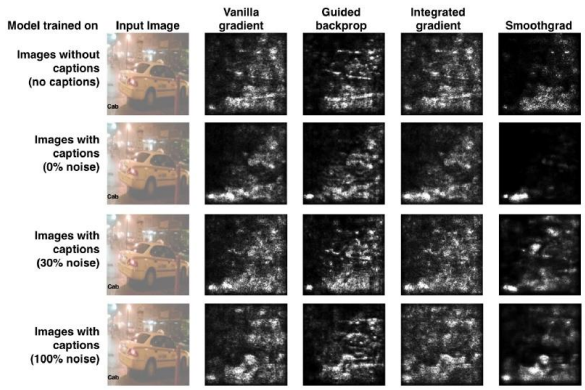


Figure 8. Saliency map results with approximated ground truth: Models trained on datasets with different noise parameter p (rows) and different saliency map methods (columns) are presented. The approximated ground truth is that the network is paying a lot more attention to the image than the caption in all cases, which is not clear from saliency maps.

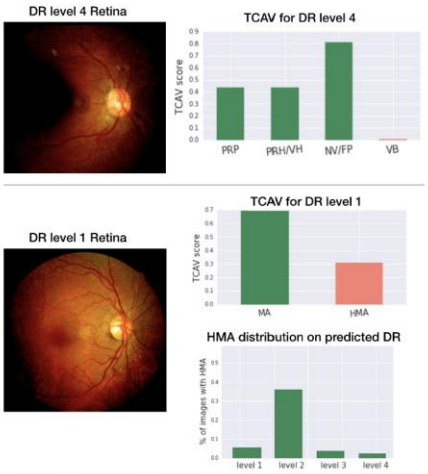


Figure 10. Top: A DR level 4 image and TCAV results. TCAVQ is high for features relevant for this level (green), and low for an irrelevant concept (red). Middle: DR level 1 (mild) TCAV results. The model often incorrectly predicts level 1 as level 2, a model error that could be made more interpretable using TCAV: TCAVQs on concepts typically related to level 1 (green, MA) are high in addition to level 2-related concepts (red, HMA). Bottom: the HMA feature appears more frequently in DR level 2 than DR level 1.

30. Semantic Information

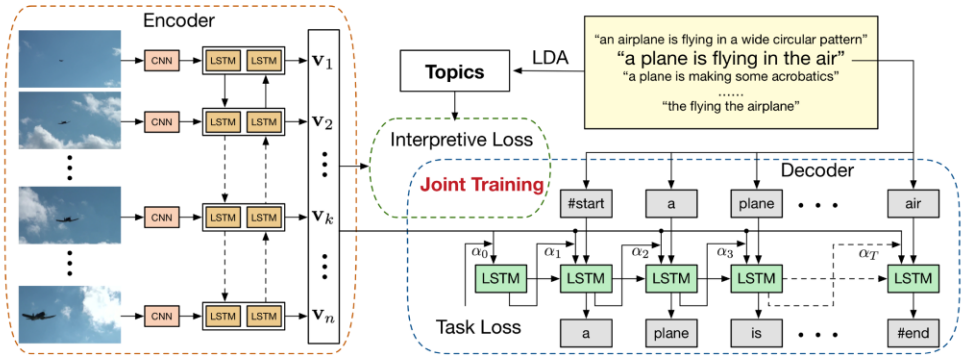


Figure 2. The attentive encoder-decoder framework for the video captioning task, which can automatically learn interpretable features. We stack a CNN model and a bi-directional LSTM model as encoder to extract video features $\{\mathbf{v}_1, ...\mathbf{v}_n\}$, and then feed them to an LSTM decoder to generate descriptions. The attention mechanism is used to let the decoder focus on a weighted sum of temporal features with weight α_t. We extract latent topics from human labeled descriptions as semantic information and introduce an interpretive loss to guide the learning towards interpretable features, which is optimized jointly with the negative log-likelihood of training descriptions.

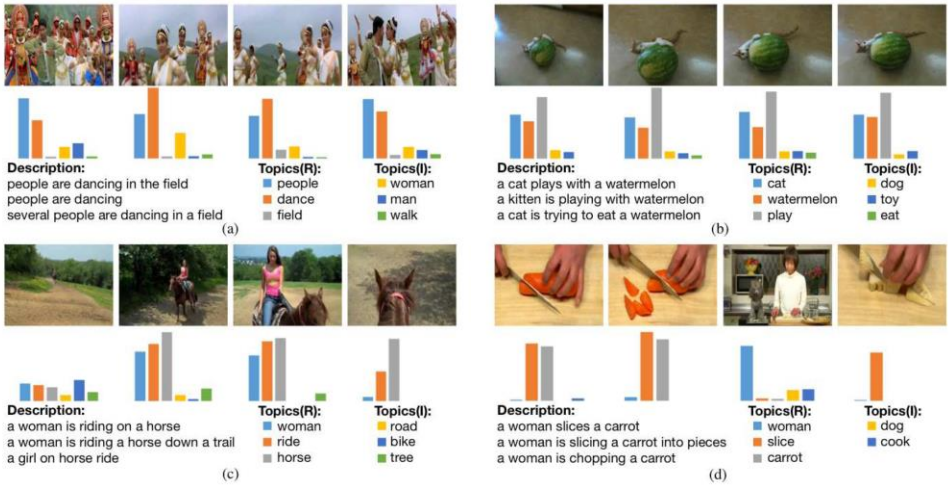


Figure 4. Neuron activations with respect to relevant and irrelevant topics in sampled videos. Topics(R) are relevant topics extracted from video descriptions. Topics(I) are irrelevant topics which are unimportant or easily confusing topics. We plot the activations of one neuron related to each topic through time.

31. Rationalizing Neural Predictions

what is the easiest way to install all the media codec available for ubuntu ? i am having issues with multiple applications prompting me to install codecs before they can play my files . how do i install media codecs ?

---

what should i do when i see <unk> report this <unk> ? an unresolvable problem occurred while initializing the package information . please report this bug against the 'update-manager ' package and include the following error message : e : encountered a section with no package : header e : problem with mergelist <unk> e : the package lists or status file could not be parsed or opened .

---

please any one give the solution for this whenever i try to convert the rpm file to deb file i always get this problem error : <unk> : not an rpm package (or package manifest) error executing `` lang=c rpm -qp -- queryformat % { name } <unk> ' '' : at <unk> line 489 thanks converting rpm file to debian fle

---

how do i mount a hibernated partition with windows 8 in ubuntu ? i ca n't mount my other partition with windows 8 , i have ubuntu 12.10 amd64 : error mounting /dev/sda1 at <unk> : command-line `mount -t `` ntfs '' -o `` uhelper=udisks2 , nodev , nosuid , uid=1000 , gid=1000 , dmask=0077 , fmask=0177 '' `` /dev/sda1 '' `` <unk> '' ' exited with non-zero exit status 14 : windows is hibernated , refused to mount . failed to mount '/dev/sda1 ' : operation not permitted the ntfs partition is hibernated . please resume and shutdown windows properly , or mount the volume read-only with the 'ro ' mount option

Figure 7: Examples of extracted rationales of questions in the AskUbuntu domain.

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

---

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

---

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

---

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate … … m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Figure 3: Examples of extracted rationales indicating the sentiments of various aspects. The extracted texts for appearance, smell and palate are shown in red, blue and green color respectively. The last example is shortened for space.

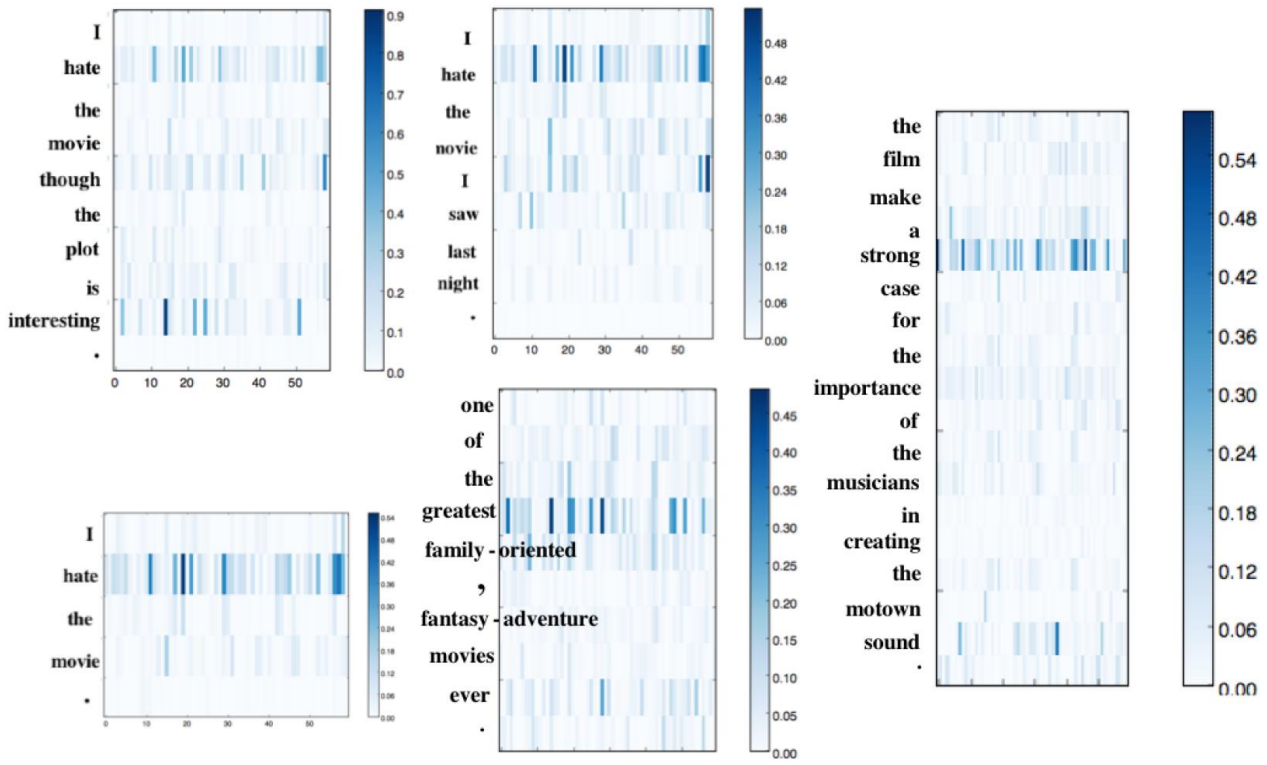


Figure 8: Variance visualization.

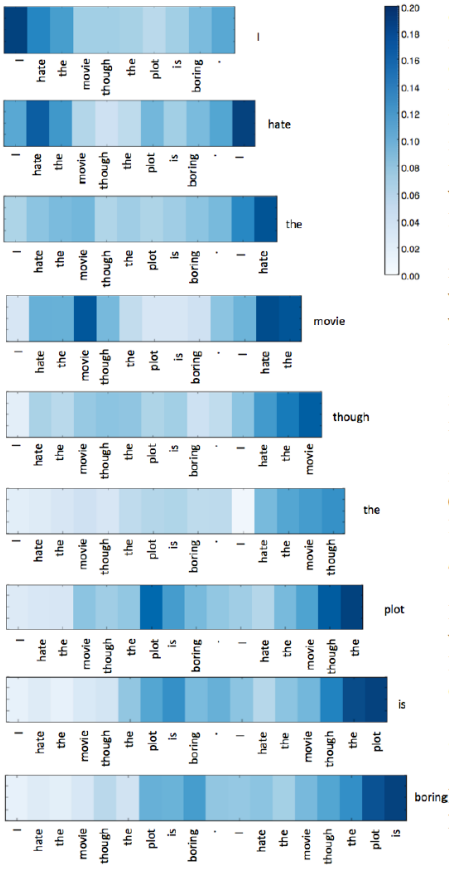


Figure 9: Saliency heatmap for SEQ2SEQ auto-encoder in terms of predicting correspondent token at each time step.

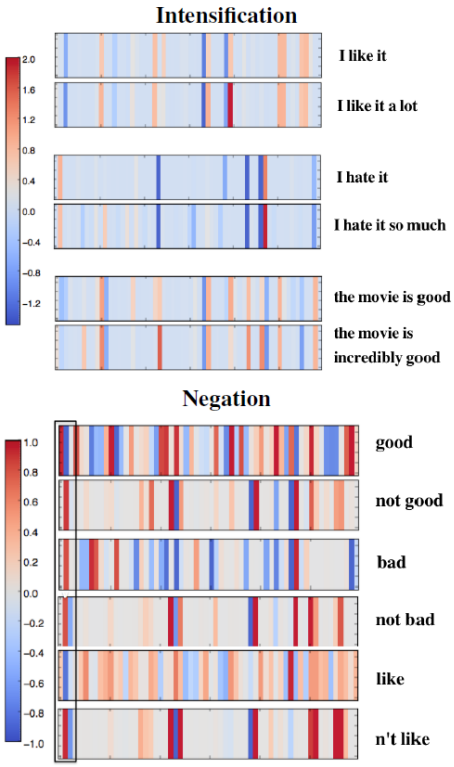


Figure 1: Visualizing intensification and negation. Each vertical bar shows the value of one dimension in the final sentence/phrase representation after compositions. Embeddings for phrases or sentences are attained by composing word representations from the pretrained model.

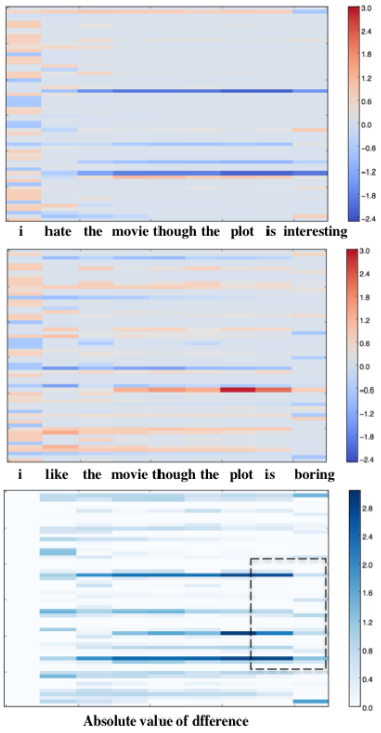


Figure 3: Representations over time from LSTMs. Each column corresponds to outputs from LSTM at each time-step (representations obtained after combining current word embedding with previous build embeddings). Each grid from the column corresponds to each dimension of current time-step representation. The last rows correspond to absolute differences for each time step between two sequences.

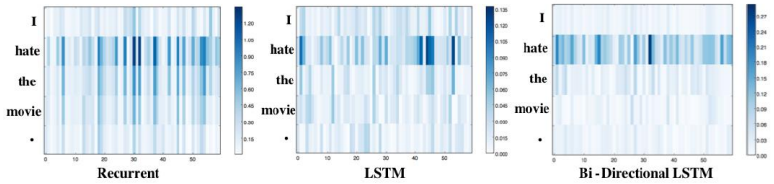


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.

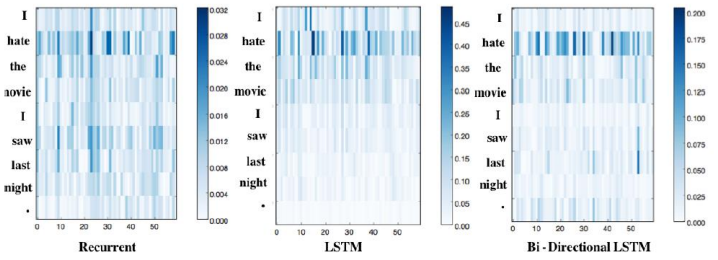


Figure 6: Saliency heatmap for "I hate the movie I saw last night ." .

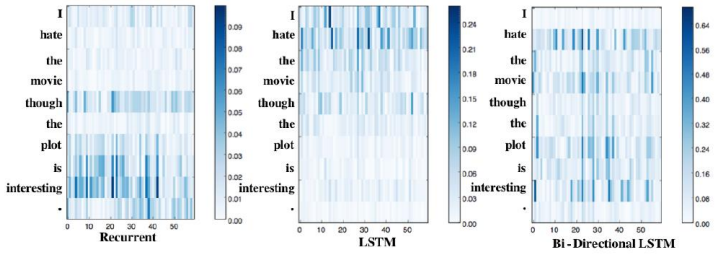


Figure 7: Saliency heatmap for "I hate the movie though the plot is interesting ." .

33. Simulated learning

Absent

34. Why should I trust you?

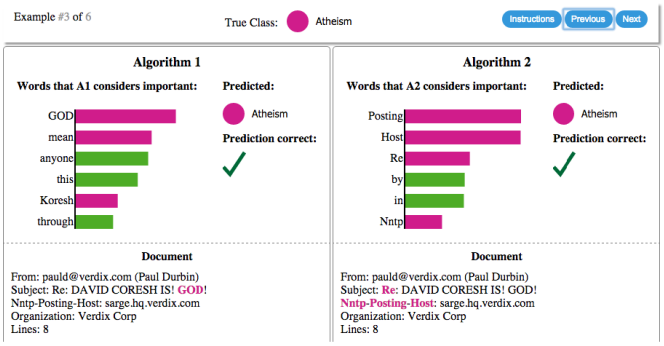
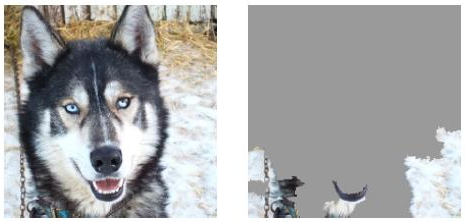


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

(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" ($p = 0.21$)



(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.

35. Boolean rule extraction

Table 4. Set 1 and Set 2 rules for malignancy in descending order of PPV

Rule Number	Conjunctive Rule Statements	For Individual Rule			Cumulative Value in Set		
		Sensitivity	Specificity	PPV	Sensitivity	Specificity	PPV
Set 1							
1	69 <= MaxLes <= 410 Pain = 0 44 <= MaxSolid <= 230 ColScore =4	0.18	1	1	0.18	1	1
2	54 <= Age <= 94 Pain = 0 Ascites = 1 WallRegularity = 1 Shadows = 0 ColScore =2 or 4	0.18	0.99	0.97	0.3	0.99	0.98
3	108 <= MaxLes <= 410 Pain = 0 42 <= MaxSolid <= 230 WallRegularity = 1	0.22	0.99	0.97	0.39	0.99	0.97
4	HormTherapy = 0 46 <= Age <= 94 PapFlow = 1 15 <= MaxSolid <= 230 ColScore = 3 4	0.19	0.99	0.95	0.49	0.99	0.96
5	31 <= Age <= 94 77 <= MaxLes <= 410 Pain = 0 30 <= MaxSolid <= 230 WallRegularity = 1 Shadows = 0 ColScore = 2 or 4	0.32	0.99	0.94	0.54	0.99	0.95
Set 2							
1	WallRegularity = 1 ColScore = 4 51.6<= Age <= 93.5	0.22	1	1	0.22	1	1
2	HormTherapy = 0 WallRegularity = 1 ColScore = 4 108<=MaxLes <=403	0.14	1	1	0.28	1	1
3	PapNr = 5 PapFlow = 1 ColScore = 3 4 59<=MaxLes <= 401 19.9<=MaxSolid<=227	0.18	1	1	0.39	1	1
4	WallRegularity = 1 Locul5 = 1 49.72<= Age <= 92.25	0.20	0.99	0.96	0.49	0.99	0.98
5	HormTherapy = 0 PapNr = 3 4 5 ColScore = 3 4 28.7<=MaxSolid<=224.2	0.17	0.99	0.97	0.52	0.99	0.97

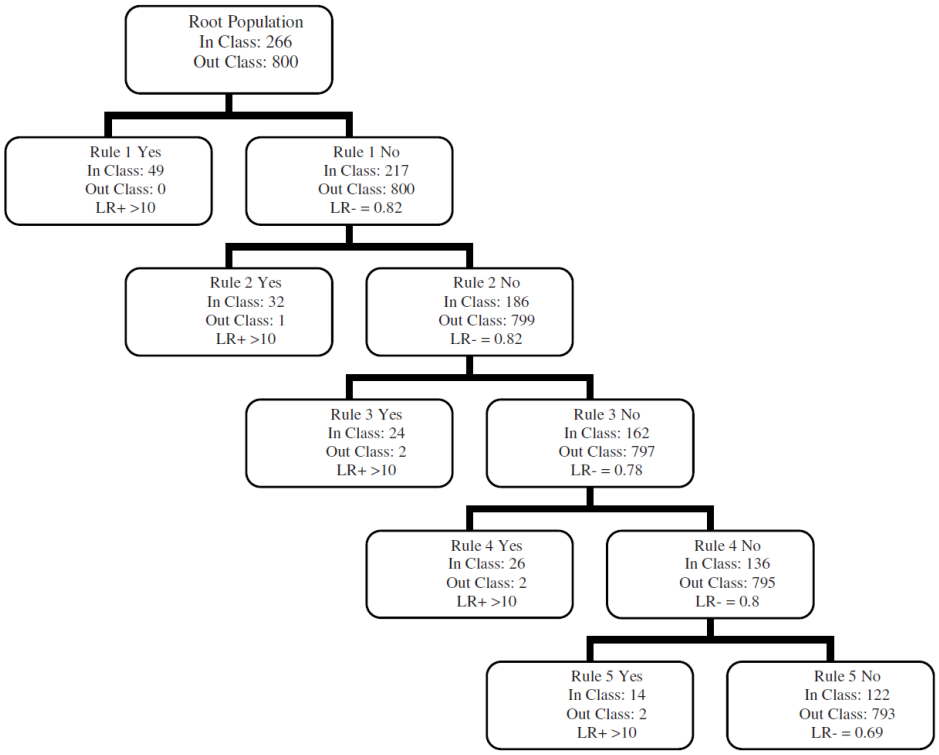


Fig. 3. Hierarchical Rule Tree for Set 1: The Positive and Negative Likelihood Ratios for each rule is shown

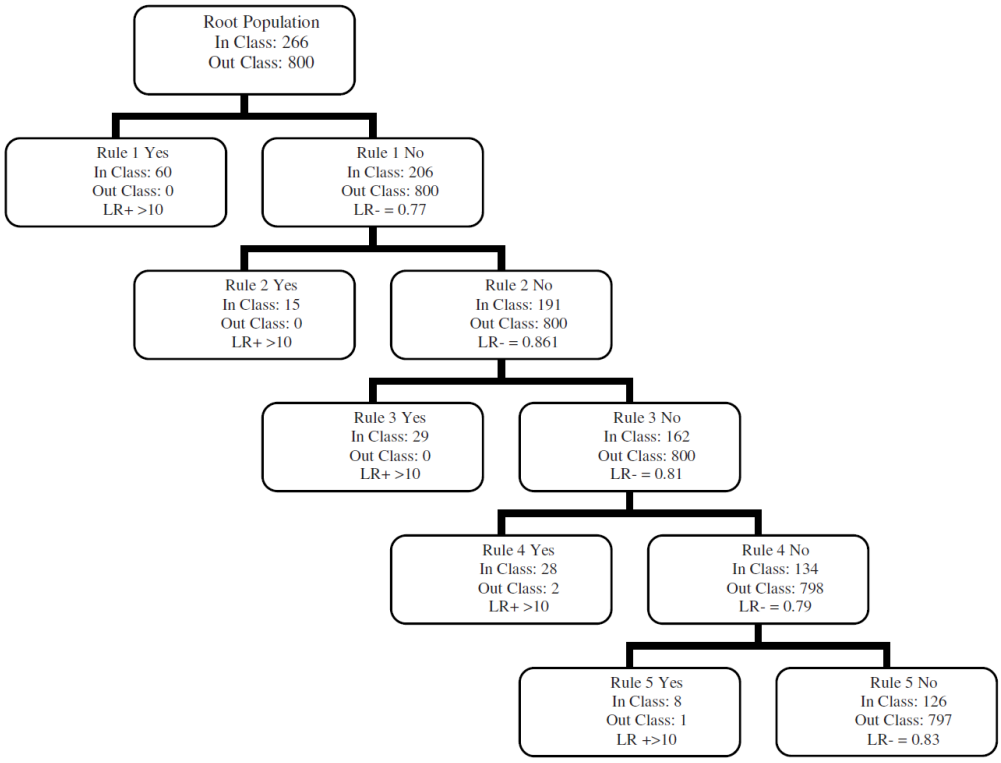


Fig. 4. Hierarchical Rule Tree for Set 2: The Positive and Negative Likelihood Ratios for each rule is shown

36. Transparent model distillation

Absent

37. Genetic programming

Fig. 2 and Fig. 3. show predictions from the ANN and G-REX, plotted against the target values.

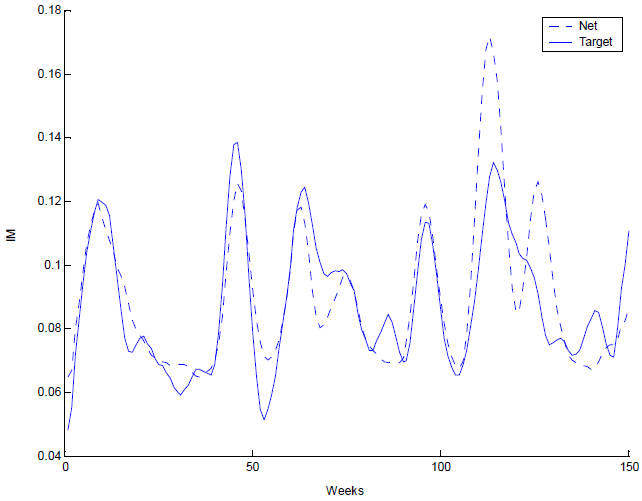


Fig. 2: ANN prediction for Ford IM. Training and test set.

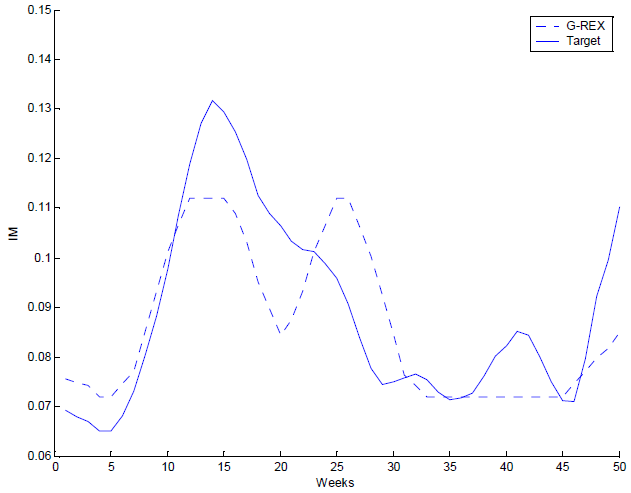


Fig. 3: G-REX prediction for Ford IM. Test set only.

38. Comprehensibility

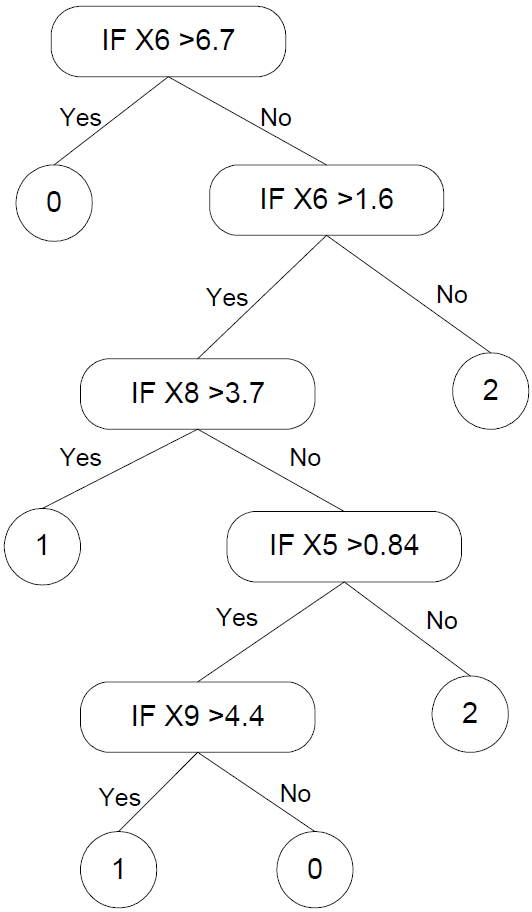


Fig. 4: Rule for WAV extracted by G-REX.

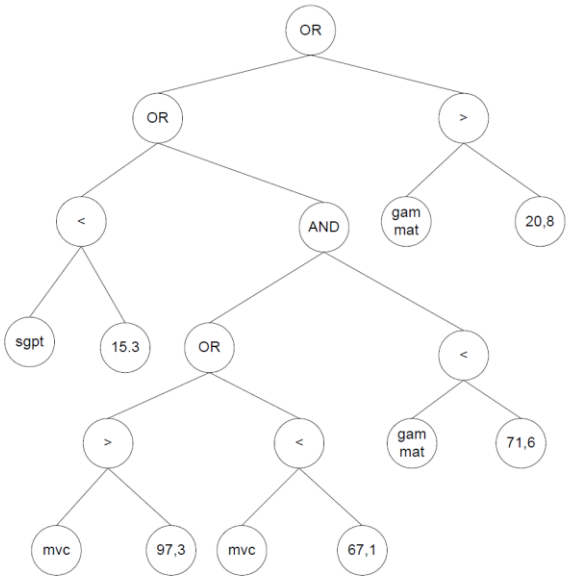


Fig. 1: Rule for BLD extracted by G-REX.

39. G-REX

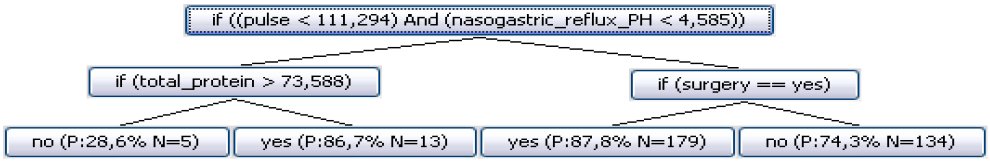


Figure 1 – Tree with complex conditions

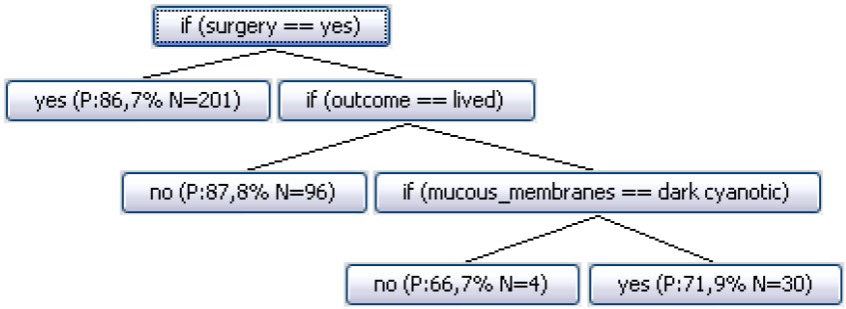


Figure 2 - Basic decision list model

40. Explorable Approximations

If Age <50 and Male =Yes:

If Past-Depression =Yes and Insomnia =No and Melancholy =No, then Healthy

If Past-Depression =Yes and Insomnia =Yes and Melancholy =Yes and Tiredness =Yes, then Depression

If Age ≥ 50 and Male =No:

If Family-Depression =Yes and Insomnia =No and Melancholy =Yes and Tiredness =Yes, then Depression

If Family-Depression =No and Insomnia =No and Melancholy =No and Tiredness =No, then Healthy

Default:

If Past-Depression =Yes and Tiredness =No and Exercise =No and Insomnia =Yes, then Depression

If Past-Depression =No and Weight-Gain =Yes and Tiredness =Yes and Melancholy =Yes, then Depression

If Family-Depression =Yes and Insomnia =Yes and Melancholy =Yes and Tiredness =Yes, then Depression

Figure 1: Explanations generated by our approach on depression dataset when approximating a deep neural network

41. Music Content Analysis

Absent

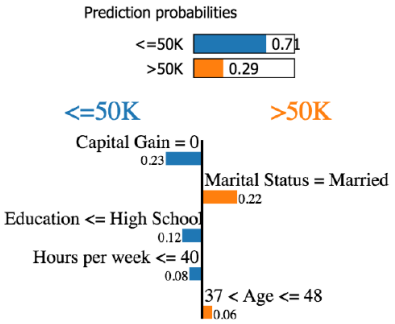
42. Two-level boolean rule learning

Absent

43. Identifying Prediction Invariance

Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	$\leq$ High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	≤ 40
Country	United States

(a) Instance



(b) Linear LIME explanation

IF Education $\leq$ High School
THEN PREDICT Salary $\leq 50K$

(c) aLIME explanation (*anchor*)

Figure 1: Explaining a prediction from the UCI adult dataset. The task is to predict if a person's salary is higher than 50,000 dollars (>50k) or not (≤50K).

IF Education $\leq$ High School
THEN PREDICT Salary $\leq 50K$

IF Marital status = Never married
THEN PREDICT Salary $\leq 50K$

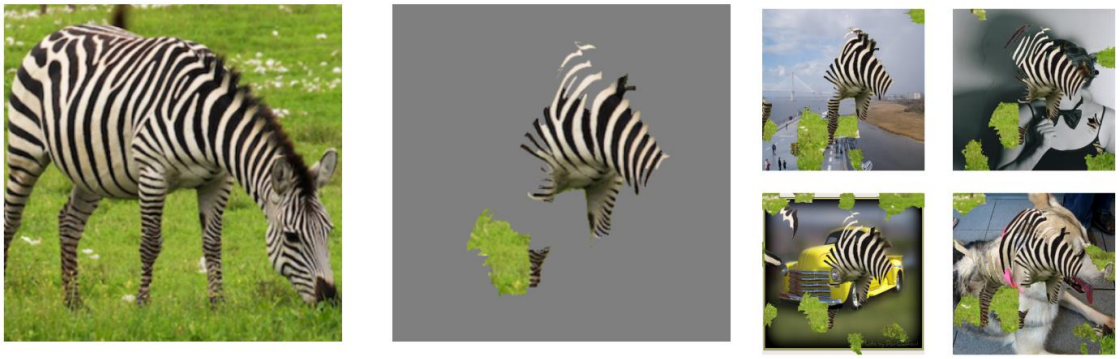
(a) Adult

IF Inpatient visits = 0
THEN PREDICT Never

IF Inpatient visits ≥ 2 **AND** Emergency visits ≥ 1
AND Outpatient visits ≥ 1
THEN PREDICT > 30 days

(b) Hospital readmission

Figure 4: Top-2 anchors chosen with Submodular Pick for both datasets



(a) Original image (b) Anchor for "Zebra" (c) Images with $P(\text{zebra}) > 90\%$

Figure 5: **Image classification**: explaining a prediction from Inception, and examples from $\mathcal{D}(z|c, x)$

(a) Original Image

What is the mustache made of?	banana
What is the ground made of ?	banana
What is the bed made of ?	banana
What is this mustache ?	banana
What is the man made of?	banana
What is the picture of ?	banana

(b)

How **many** bananas are in the picture?	2
How **many** are in the picture?	2
many animals the picture ?	2
How **many** people are in the picture ?	2
How **many** zebras are in the picture ?	2
How **many** planes are on the picture ?	2

(c)

Figure 6: **Visual QA:** explaining predictions from a CNN-LSTM model by looking at the question text (image is fixed), and examples from $\mathcal{D}(z|c, x)$

44. Inductive learning

Absent

45. Knowledge discovery via multiple models

Absent

46. Single Tree Approximation

Absent

47. Using oracle guides

```
if (Body_mass_index > 29.132)
 |T: if (plasma_glucose < 127.40)
 |    |T: [Negative] {56/12}
 |    |F: [Positive] {29/21}
 |F: [Negative] {63/11}
```

Figure 3: Sample tree evolved on diabetes dataset

48. Extracting comprehensible models

Absent

49. Model extraction

Absent

50. ICU Outcome Prediction

Duplicate

51. Discovering additive structure

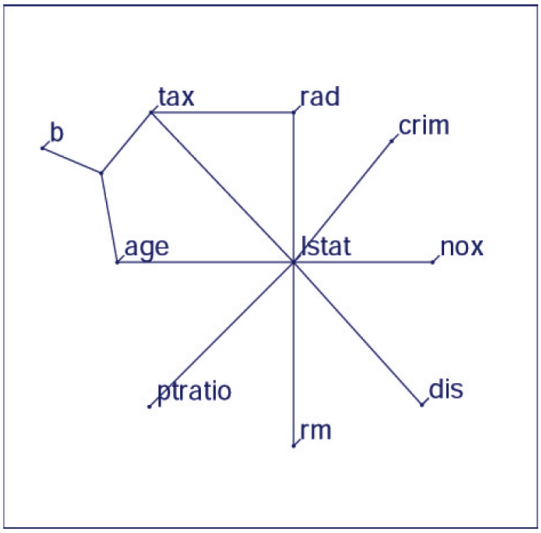
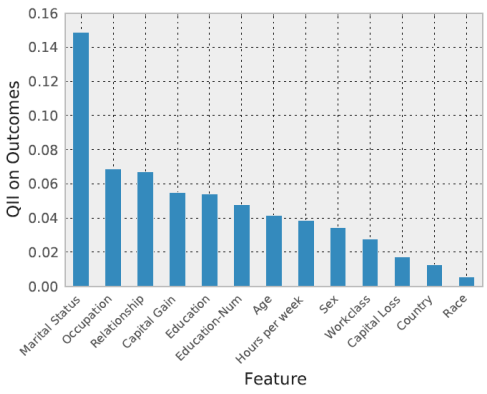
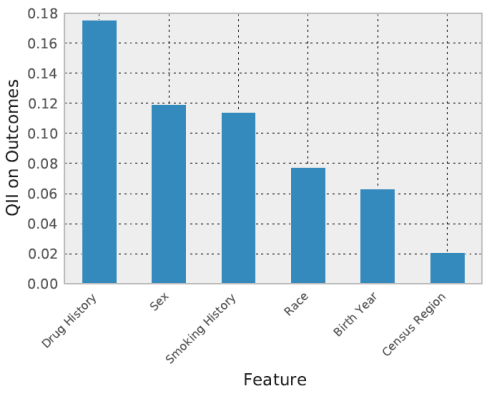


Figure 4: Variable Interaction Network for a Neural Network trained on the Boston Housing Data, with cutoff $\epsilon = 0.7$.

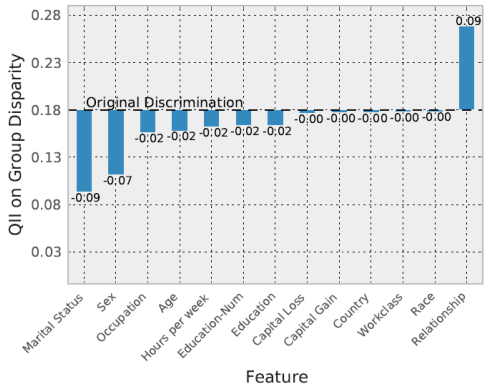
52. Quantitative input influence



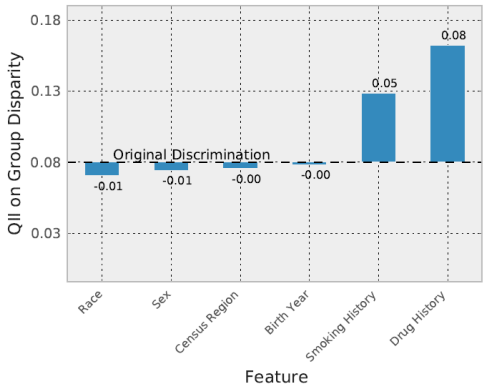
(a) QII of inputs on Outcomes for the `adult` dataset



(b) QII of inputs on Outcomes for the `arrests` dataset

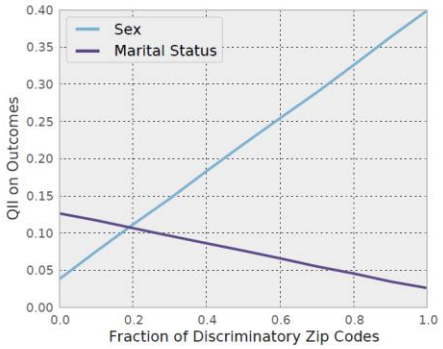


(c) QII of Inputs on Group Disparity by Sex in the `adult` dataset

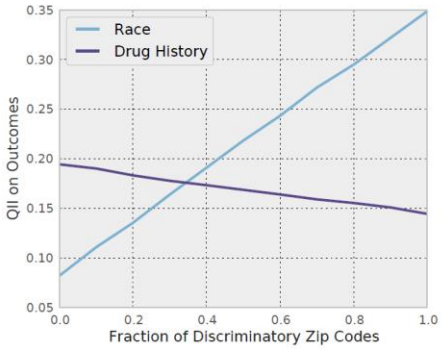


(d) Influence on Group Disparity by Race in the `arrests` dataset

Fig. 2: QII measures for the `adult` and `arrests` datasets



(a) Change in QII of inputs as discrimination by Zip Code increases in the `adult` dataset

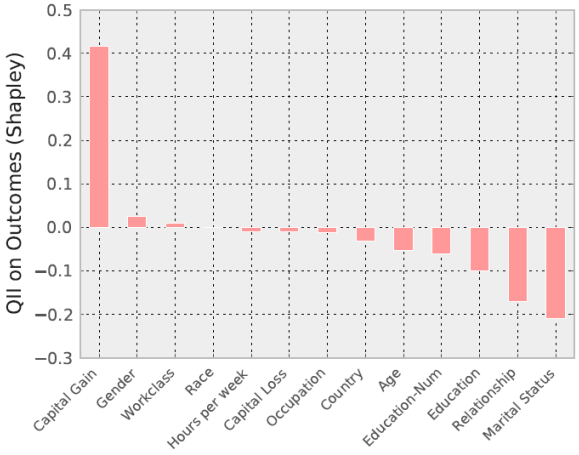


(b) Change in QII of inputs as discrimination by Zip Code increases in the `arrests` dataset

Fig. 3: The effect of discrimination on QII.

Age	23
Workclass	Private
Education	11th
Education-Num	7
Marital Status	Never-married
Occupation	Craft-repair
Relationship	Own-child
Race	Asian-Pac-Islander
Gender	Male
Capital Gain	14344
Capital Loss	0
Hours per week	40
Country	Vietnam

(a) Mr. X's profile

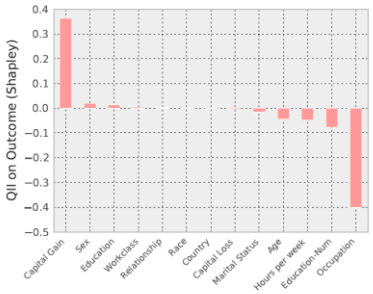


(b) Transparency report for Mr. X's negative classification

Fig. 4: Mr. X

Age	27
Workclass	Private
Education	Preschool
Education-Num	1
Marital Status	Married-civ-spouse
Occupation	Farming-fishing
Relationship	Other-relative
Race	White
Gender	Male
Capital Gain	41310
Capital Loss	0
Hours per week	24
Country	Mexico

(a) Mr. Y's profile

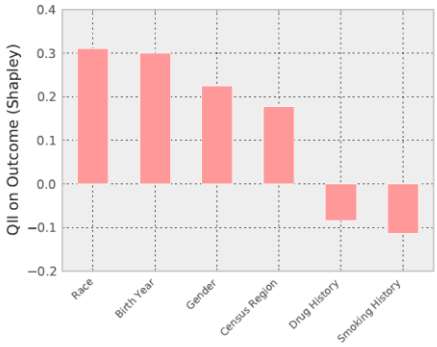


(b) Transparency report for Mr. Y's negative classification

Fig. 5: Mr. Y.

Birth Year	1984
Drug History	None
Smoking History	None
Census Region	West
Race	Black
Gender	Male

(a) Mr. Z's profile



(b) Transparency report for Mr. Z's positive classification

Fig. 6: Mr. Z.

53. Indirect Influence

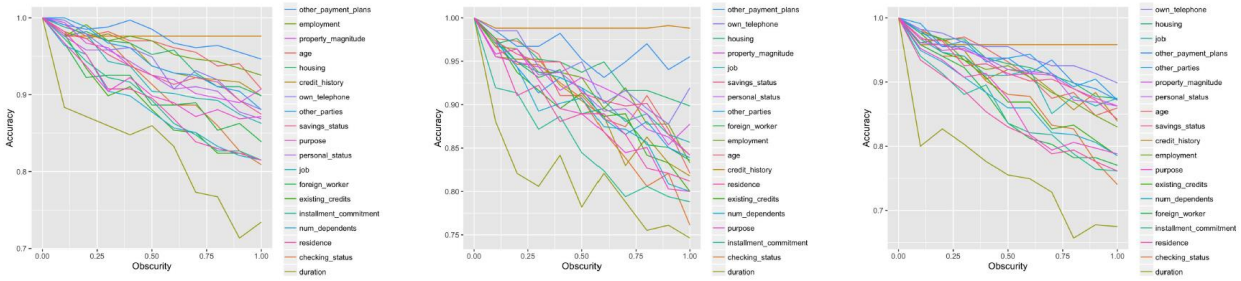


Fig. 2. Obscurity vs. consistency plots for the German Credit data. First column: decision tree model. Second column: SVM. Third column: FNN.

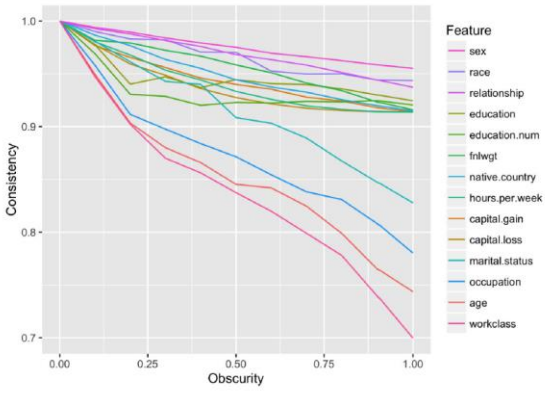


Fig. 3. Obscurity vs. consistency for Adult Income data modeled by an FNN.

Fig. 4. Model of a model decision tree probability rankings vs. GFA consistency scores shown with a linear regression and 95% confidence interval. The outlying features contain proxy information in the data set.

To consider the cases where these rankings don't match, we look at Adult Income under the C4.5 decision tree model. As shown in Figure 4, linear regression confirms that most features have similar scores, but there are a few important outliers: `marital-status`, `education`, `race`, `age`, and `capital-gain`. We hypothesize that the information in these features can be reconstructed from the remaining attributes, and so they are scored differently under an indirect influence audit. We explore this hypothesis next.

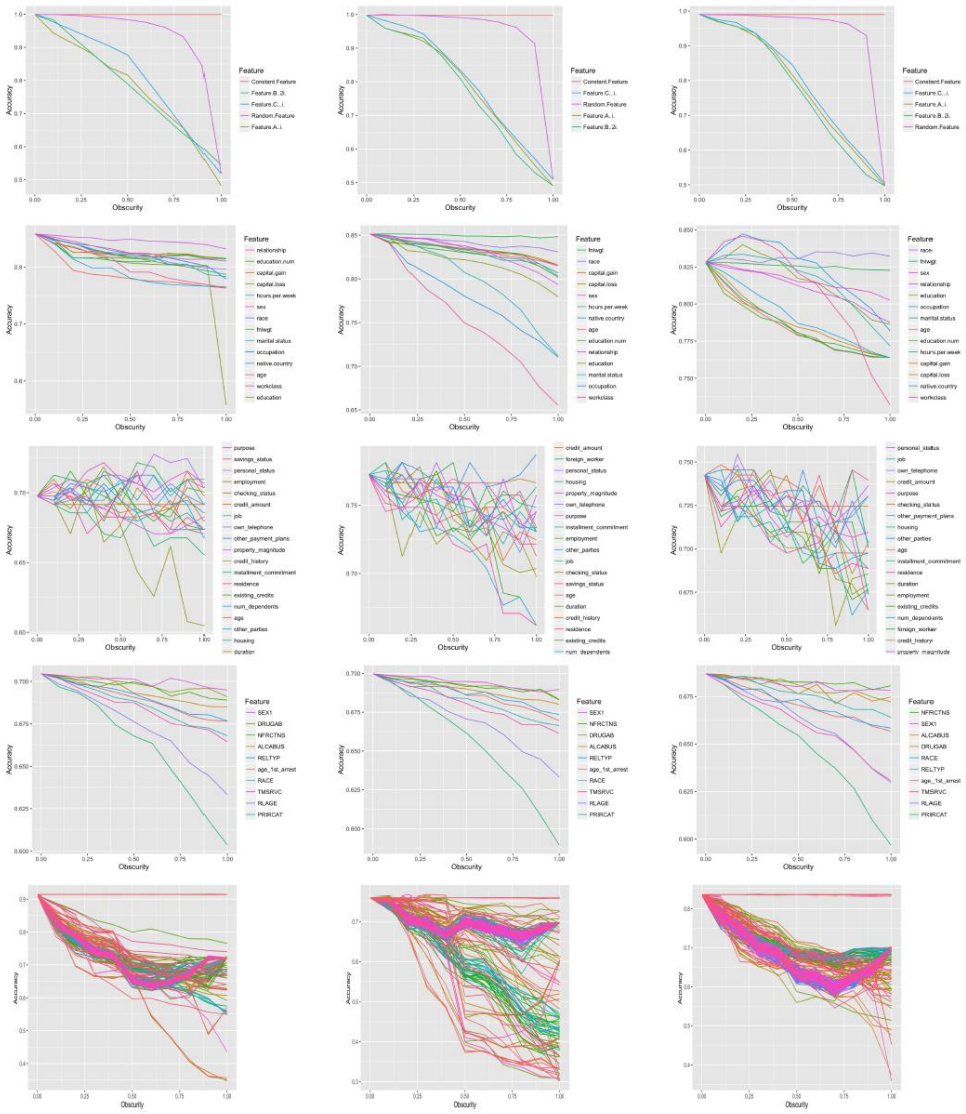


Fig. 1. Obscurity vs. accuracy plots for each model and each data set considered. First column: C4.5 decision trees. Second column: SVMs. Third column: FNNs. First row: Synthetic data. Second row: Adult income data set. Third row: German credit data. Fourth row: Recidivism data. Final row: Dark Reaction data, shown without a feature legend due to the large number of features.

54. Influence Functions

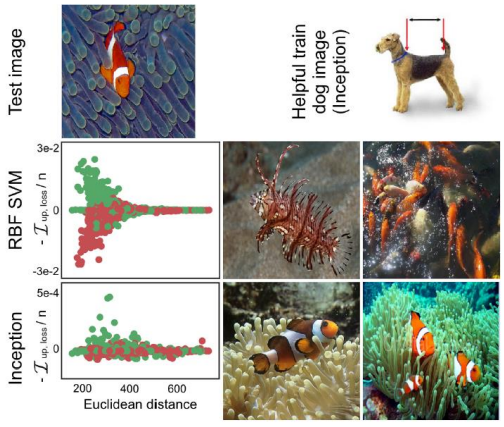


Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in the training set that helped the Inception model correctly classify the test image as a fish.

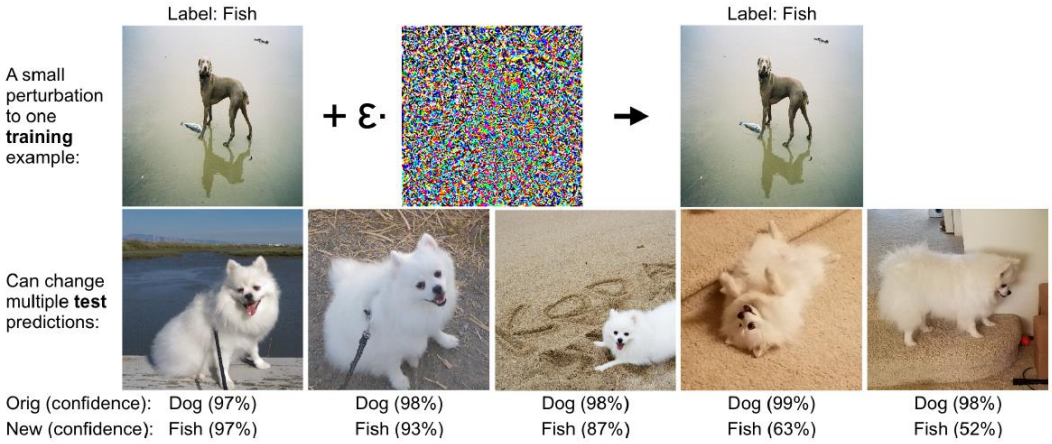


Figure 5. **Training-set attacks.** We targeted a set of 30 test images featuring the first author's dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, we created a visually-imperceptible change to the particular training image (shown on top) that flipped predictions on 16 test images.

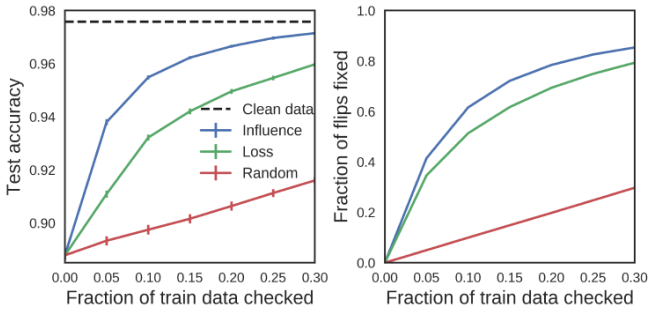


Figure 6. **Fixing mislabeled examples.** Plots of how test accuracy (left) and the fraction of flipped data detected (right) change with the fraction of train data checked, using different algorithms for picking points to check. Error bars show the std. dev. across 40 repeats of this experiment, with a different subset of labels flipped in each; error bars on the right are too small to be seen. These results are on the Enron1 spam dataset (Metsis et al., 2006), with 4,147 training and 1,035 test examples; we trained logistic regression on a bag-of-words representation of the emails.

55. Sensitivity Analysis

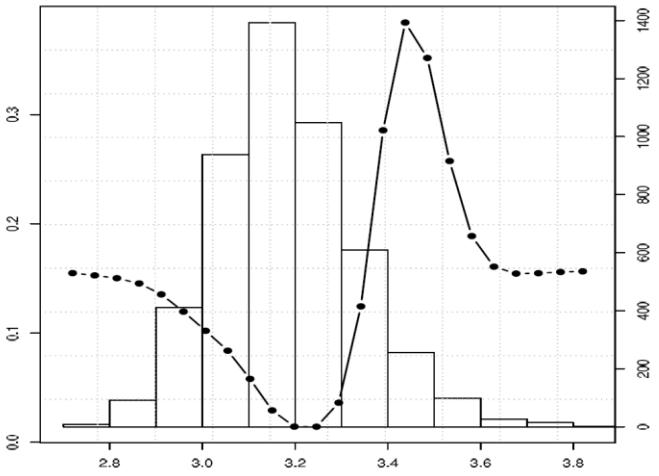


Fig. 5. VEC curve and histogram for the pH input (x-axis) and the respective high quality wine probability (left of y-axis) and frequency (right of y-axis)

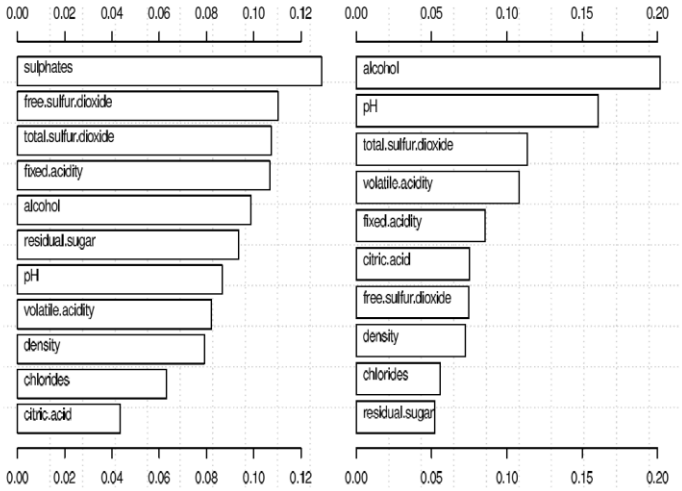


Fig. 3. Bar plots with the 1-D input importances (left) and 2-D interactions with the sulphates input (right) for the wwq-reg dataset

56. Sensitivity Analysis and Visualization

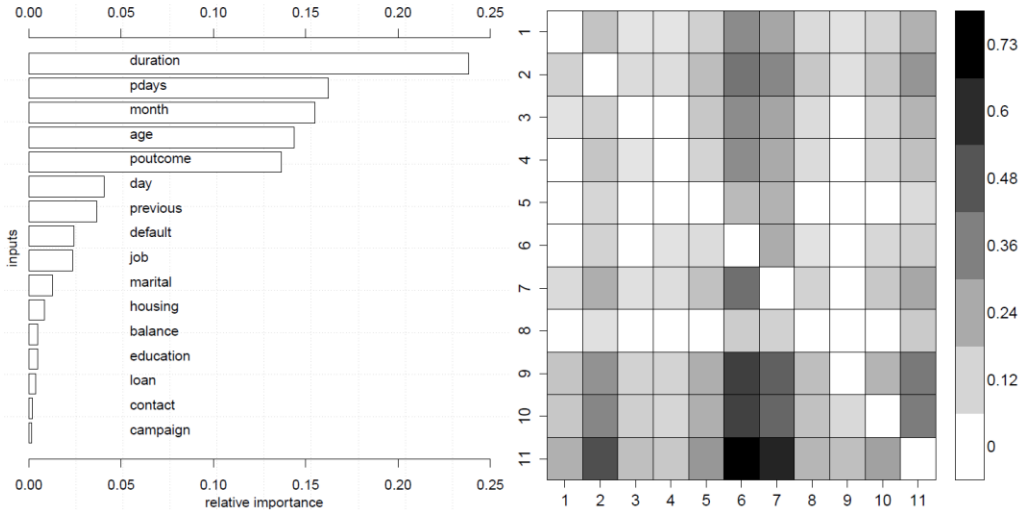


Figure 6: Bar plot with the 1D input importances for the bank data (left) and color matrix with 2D input pair sensitivity for the wwq dataset.

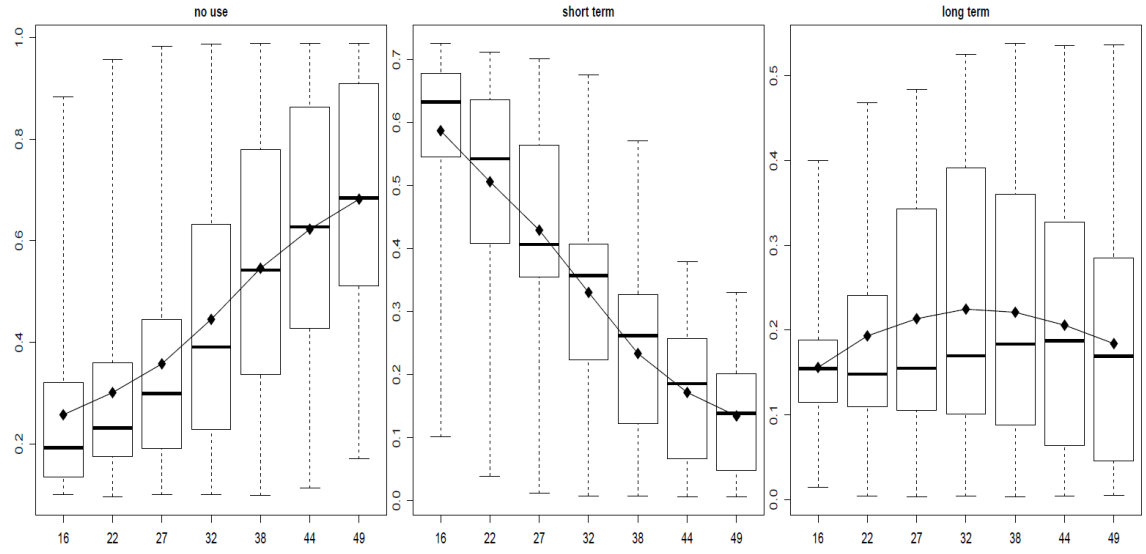


Figure 8: VEC curves with box plots for the wife's age influence (x-axis) on cmc task (y-axis) and classes "no use", "short term" and "long term".

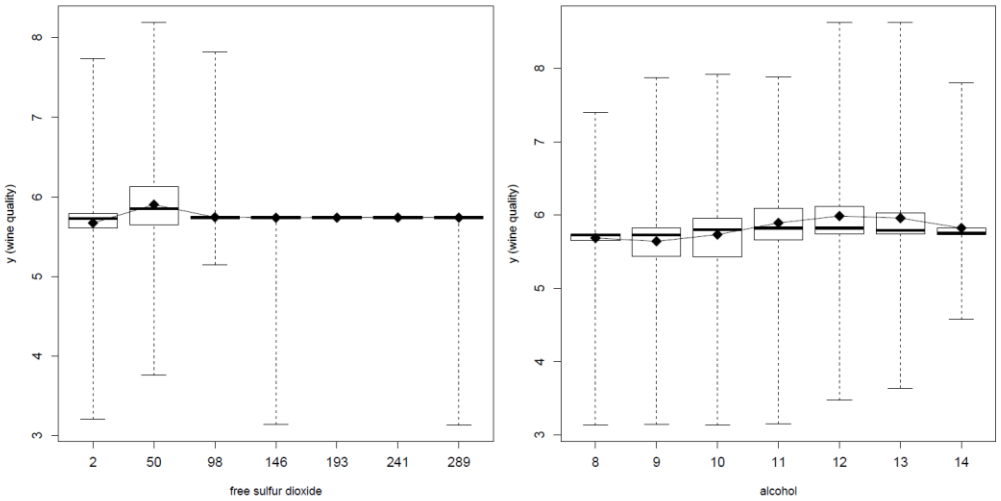


Figure 9: VEC curves with box plots for free sulfur dioxide (left) and alcohol (right).

57. A Unified Approach

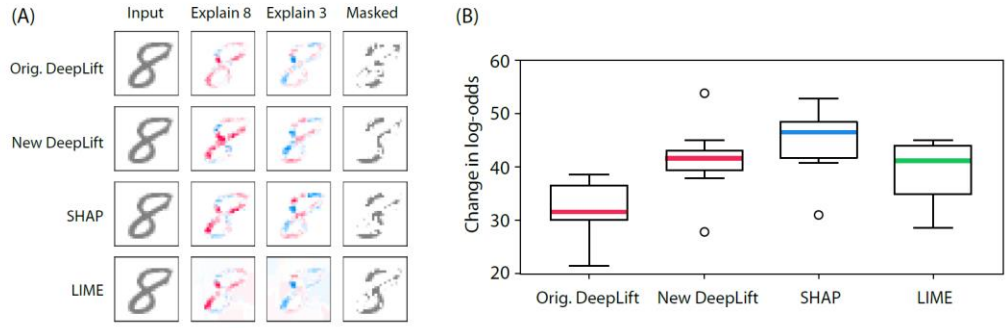


Figure 5: Explaining the output of a convolutional network trained on the MNIST digit dataset. Orig. DeepLIFT has no explicit Shapley approximations, while New DeepLIFT seeks to better approximate Shapley values. (A) Red areas increase the probability of that class, and blue areas decrease the probability. Masked removes pixels in order to go from 8 to 3. (B) The change in log odds when masking over 20 random images supports the use of better estimates of SHAP values.

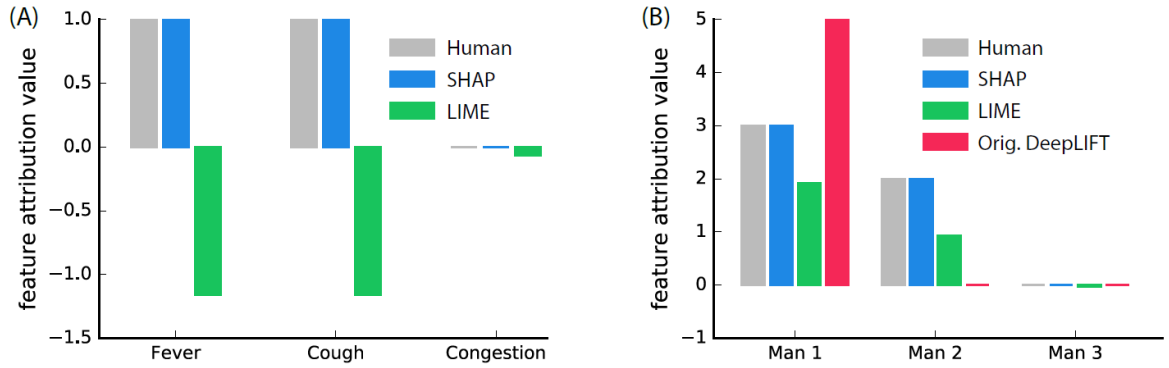
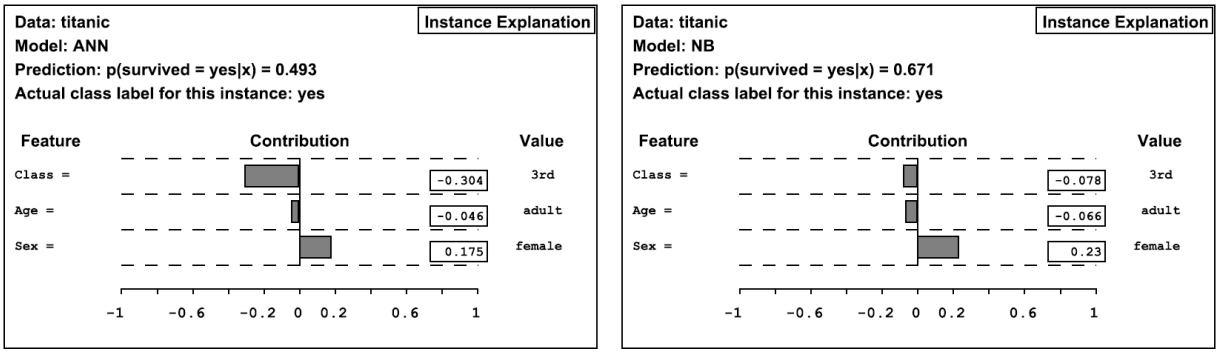


Figure 4: Human feature impact estimates are shown as the most common explanation given among 30 (A) and 52 (B) random individuals, respectively. (A) Feature attributions for a model output value (sickness score) of 2. The model output is 2 when fever and cough are both present, 5 when only one of fever or cough is present, and 0 otherwise. (B) Attributions of profit among three men, given according to the maximum number of questions any man got right. The first man got 5 questions right, the second 4 questions, and the third got none right, so the profit is $5.

58. Using Game Theory



(a) *ANN model* (b) *NB model*

Figure 7: Two explanations for the Titanic instance from the introduction. The left hand side explanation is for the ANN model. The right hand side explanation is for the NB model.

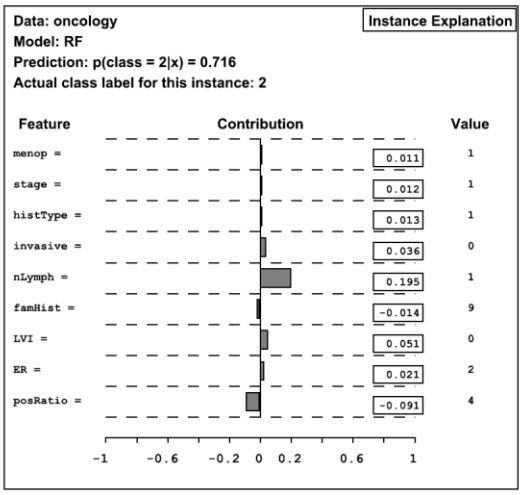
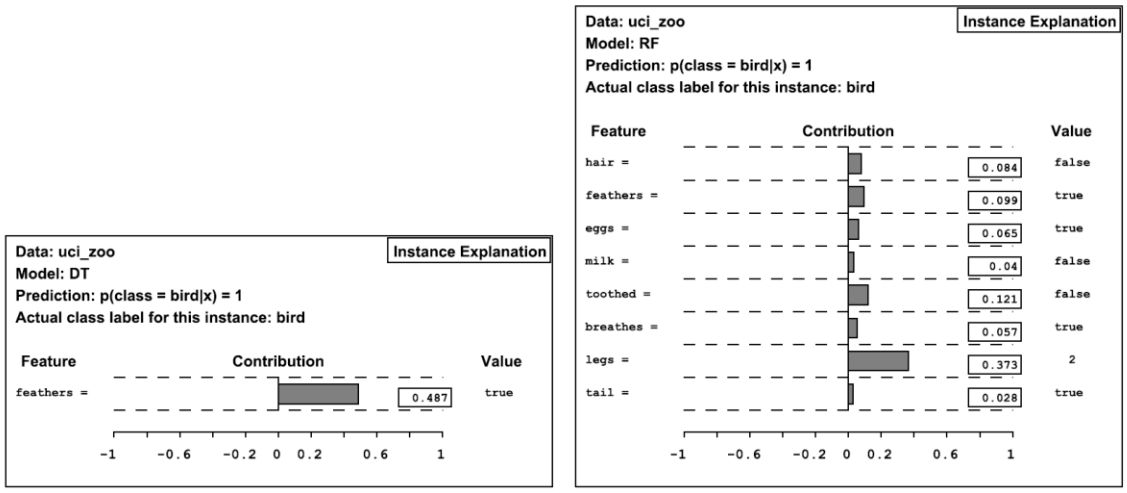


Figure 8: An explanation or the RF model's prediction for a patient from the Oncology data set.



(a) *DT model* (b) *RF model*

Figure 5: Explanations for an instance from the Zoo data set. The DT model uses a single feature, while several feature values influence the RT model. Feature values with contributions ≤ 0.01 have been removed for clarity.

59. Meaningful Perturbation

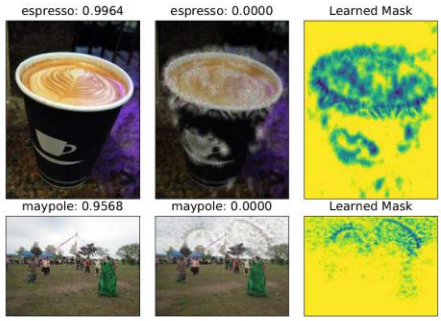


Figure 5. From left to right: an image correctly classified with large confidence by GoogLeNet [17]; a perturbed image that is not recognized correctly anymore; the deletion mask learned with artifacts. Top: A mask learned by minimizing the top five predicted classes by jointly applying the constant, random noise, and blur perturbations. Note that the mask learns to add highly structured swirls along the rim of the cup ($\gamma = 1, \lambda_1 = 10^{-5}, \lambda_2 = 10^{-3}, \beta = 3$). Bottom: A minimizing-top5 mask learned by applying a constant perturbation. Notice that the mask learns to introduce sharp, unnatural artifacts in the sky instead of deleting the pole ($\gamma = 0.1, \lambda_1 = 10^{-4}, \lambda_2 = 10^{-2}, \beta = 3$).

Figure 1. An example of a mask learned (right) by blurring an image (middle) to suppress the softmax probability of its target class (left: original image; softmax scores above images).

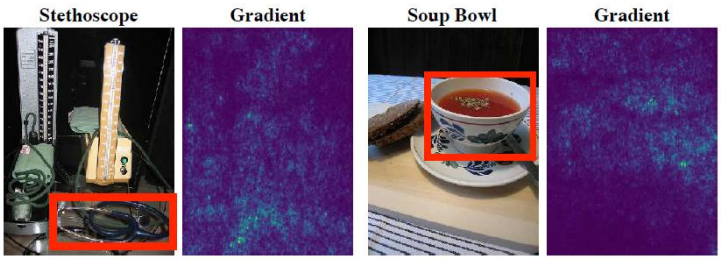


Figure 3. Gradient saliency maps of [15]. A red bounding box highlight the object which is meant to be recognized in the image. Note the strong response in apparently non-relevant image regions.

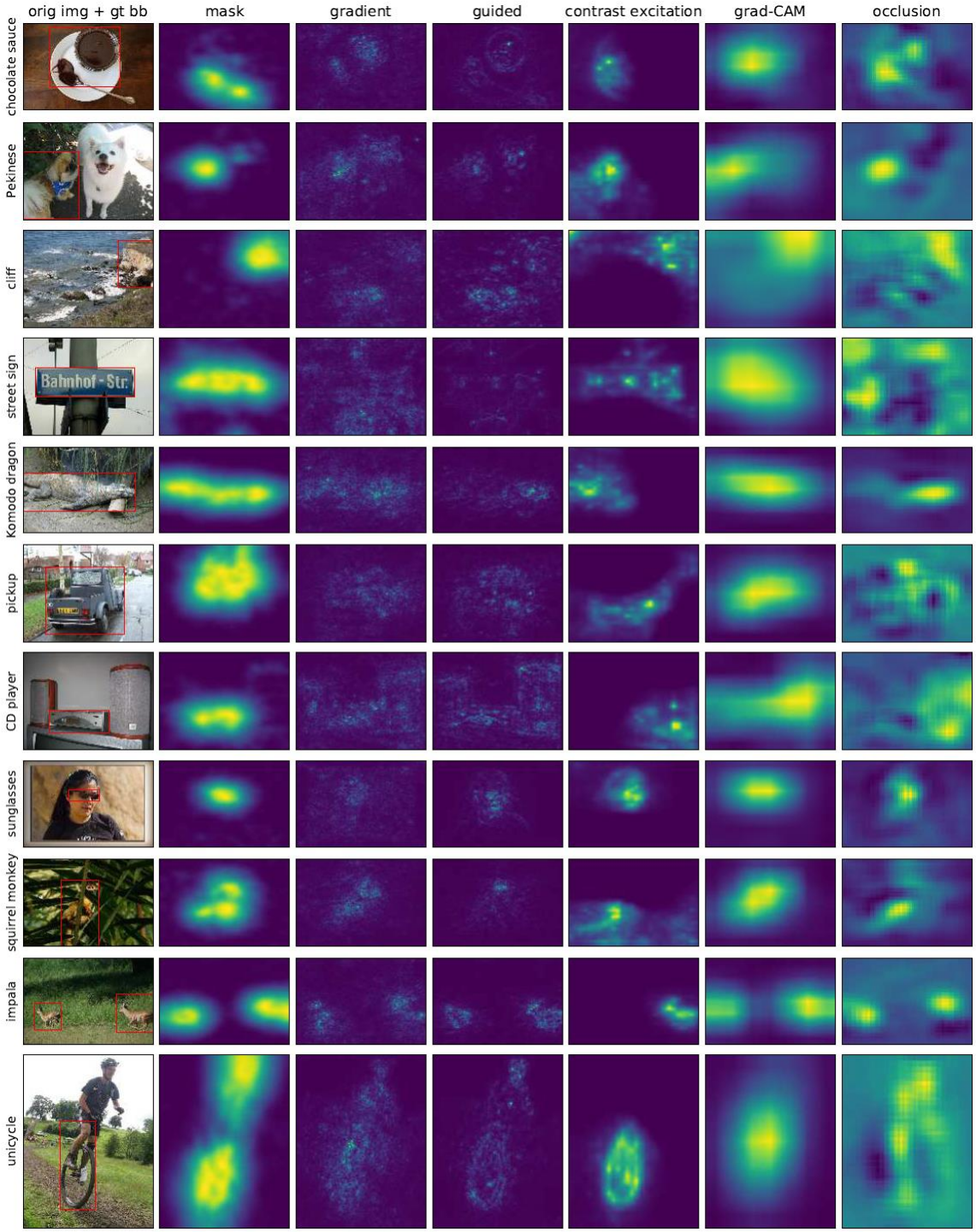
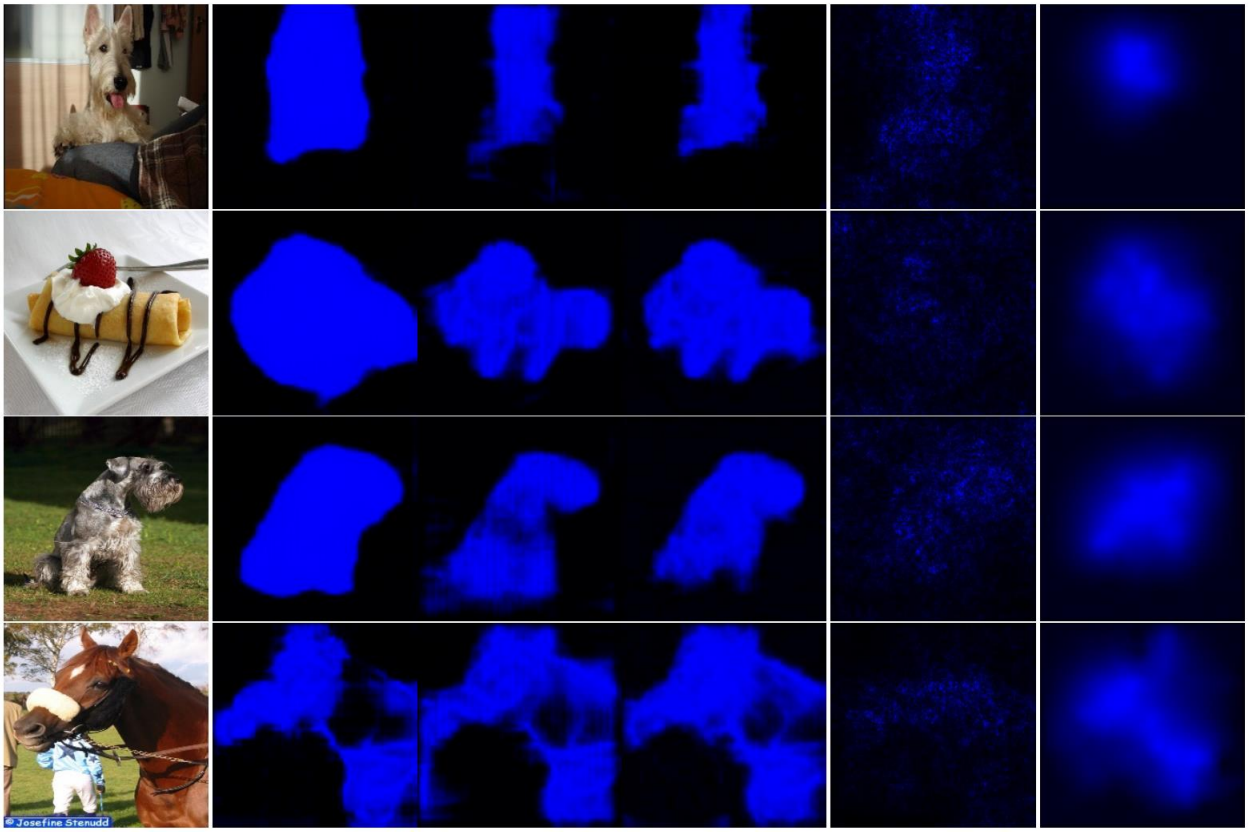


Figure 2. Comparison with other saliency methods. From left to right: original image with ground truth bounding box, learned mask subtracted from 1 (our method), gradient-based saliency [15], guided backprop [16, 8], contrastive excitation backprop [20], Grad-CAM [14], and occlusion [19].

60. Real Time Image Saliency



(a) Input Image (b) Model & AlexNet (c) Model & GoogleNet (d) Model & ResNet-50 (e) Grad [11] (f) Mask [2]

Figure 5: Saliency maps generated by different methods for the ground truth class. The ground truth classes, starting from the first row are: Scottish terrier, chocolate syrup, standard schnauzer and sorrel. Columns b, c, d show the masks generated by *our* masking models, each trained on a different black box classifier (from left to right: AlexNet, GoogleNet, ResNet-50). Last two columns e, f show saliency maps for GoogleNet generated respectively by gradient [11] and the recently introduced iterative mask optimisation approach [2].

Figure 6: Saliency maps generated by our model for images from CIFAR-10 validation set.

61. Attribute Interactions in Datasets

Table 3: Groupings for UCI datasets. Columns as follows: number of attributes in the dataset (N), size of the grouping (k), size of the largest (N_1) and second-largest (N_2) groups, baseline accuracy for the classifier trained with unshuffled data (a_0) and the CI. p_{OG} is the p-value of Test 2 in Ojala & Garriga (2010) ($p \geq 0.05$ highlighted).

Dataset	N	k	N_1	N_2	a_0	CI	p_{OG}
				SVM			
balance-scale	4	3	2	1	0.891	[0.821, 0.897]	0.03
credit-a	15	12	4	1	0.871	[0.847, 0.871]	0.04
diabetes	8	8	1	1	0.714	[0.688, 0.740]	0.59
kr-vs-kp	36	33	4	1	0.917	[0.922, 0.924]	0.00
mushroom	21	15	7	1	0.995	[0.991, 0.995]	0.00
segment	18	3	16	1	0.948	[0.936, 0.948]	0.00
soybean	35	35	1	1	0.844	[0.820, 0.850]	0.26
vehicle	18	3	15	2	0.767	[0.719, 0.781]	0.00
vote	16	8	9	1	0.931	[0.897, 0.931]	0.00
vowel	13	3	11	1	0.806	[0.760, 0.806]	0.00
				random forest			
balance-scale	4	3	2	1	0.821	[0.731, 0.833]	0.02
credit-a	15	15	1	1	0.877	[0.847, 0.883]	0.19
diabetes	8	8	1	1	0.703	[0.698, 0.740]	0.89
kr-vs-kp	36	16	21	1	0.982	[0.972, 0.982]	0.00
mushroom	21	14	8	1	1.000	[0.996, 1.000]	0.00
segment	18	4	15	1	0.986	[0.979, 0.986]	0.00
soybean	35	24	12	1	0.964	[0.946, 0.964]	0.00
vehicle	18	3	13	4	0.752	[0.710, 0.757]	0.00
vote	16	10	7	1	0.948	[0.897, 0.948]	0.00
vowel	13	3	11	1	0.917	[0.901, 0.917]	0.00

Table 2: The `synthetic` dataset. The cardinality of the grouping is k and CI is the confidence interval for accuracy. Original accuracy using unshuffled data (a_0) and the final grouping ($\mathcal{S}$, highlighted row) shown above and below the table, respectively. An asterisk ($*$) denotes that the factorisation is valid.

(a) SVM

$a_0 = 0.908$

k	CI		a3	a4	a2	a1
2	[0.900, 0.920]	*	(A)	(B	B	B)
3	[0.896, 0.920]	*	(A)	(B)	(C	C)
4	[0.696, 0.784]		(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1, 2\}, \{3\}, \{4\}\}$

(b) Random forest

$a_0 = 0.904$

k	CI		a3	a4	a1	a2
2	[0.896, 0.928]	*	(A)	(B	B	B)
3	[0.896, 0.928]	*	(A)	(B)	(C	C)
4	[0.668, 0.756]		(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1, 2\}, \{3\}, \{4\}\}$

(c) Naïve Bayes

$a_0 = 0.760$

k	CI		a1	a2	a3	a4
2	[0.760, 0.760]	*	(A)	(B	B	B)
3	[0.760, 0.760]	*	(A)	(B)	(C	C)
4	[0.760, 0.760]	*	(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

62. Randomization

Table 5 Grouping of the `diabetes` dataset

`diabetes`		original accuracy	final accuracy	Bayes fidelity	final fidelity	plas	mass	preg	age	pedi	insu	pres	skin	plas	mass	preg	age	pedi	insu	pres	skin
size:	768																				
classes:	2																				
attributes:	8																				
major class:	0.65																				
Classifier																					
DecisionStump		0.66	0.66	1.00	1.00	·	·	·	·	·	·	·	·	2	6	1	8	7	5	3	4
SMO radial		0.76	0.70	0.88	0.88	A	·	A	·	A	A	·	·	1	4	2	6	3	5	7	8
IBk		0.69	0.59	0.67	0.69	A	·	A	A	A	·	·	·	2	7	1	3	4	8	5	6
AdaBoostM1		0.75	0.75	0.91	1.00	A	A	·	A	·	·	·	·	1	2	4	3	8	7	5	6
J48		0.75	0.72	0.86	0.95	A	A	·	A	·	·	·	·	1	3	8	2	6	7	4	5
Bagging		0.74	0.70	0.85	0.90	A	A	·	○	·	A	·	·	1	2	7	3	4	8	6	5
LogitBoost		0.76	0.74	0.89	0.95	A	A	·	○	·	A	A	·	1	2	7	3	6	5	4	8
Logistic		0.77	0.74	0.88	0.93	A	A	A	·	·	·	A	·	1	2	3	7	6	5	4	8
SMO		0.77	0.75	0.92	0.98	A	A	A	·	A	·	A	·	1	3	2	8	4	6	5	7
LMT		0.78	0.76	0.91	0.97	A	A	A	·	A	A	·	·	1	2	3	8	4	6	5	7
naiveBayes		0.78	0.75	0.90	0.93	A	A	A	A	·	○	A	·	1	3	2	4	8	5	6	7
randomForest		0.76	0.74	0.87	0.95	A	A	A	A	A	·	·	·	1	2	5	4	3	6	8	7
JRip		0.71	0.69	0.83	0.92	A	A	B	·	B	B	·	B	1	3	4	8	2	5	6	7
PART		0.78	0.75	0.88	0.95	A	○	·	A	·	·	·	·	1	3	5	2	8	7	4	6
OneR		0.73	0.73	1.00	1.00	○	·	·	·	·	·	·	·	1	6	2	8	7	5	3	4

The *original accuracy* shows the accuracy of the classifier without any randomization, *final accuracy* is the accuracy after randomizing using the grouping. *Bayes fidelity* is the fidelity when all attributes are in singleton groups. The final fidelity is calculated using the grouping of attributes discovered. Groups are denoted by roman letters. Singleton groups are marked with ○ and pruned-out singletons with ·. The rank order of importance of individual attributes is shown to the right of the groupings with numbers

63. Visualizing the effects of predictor variables

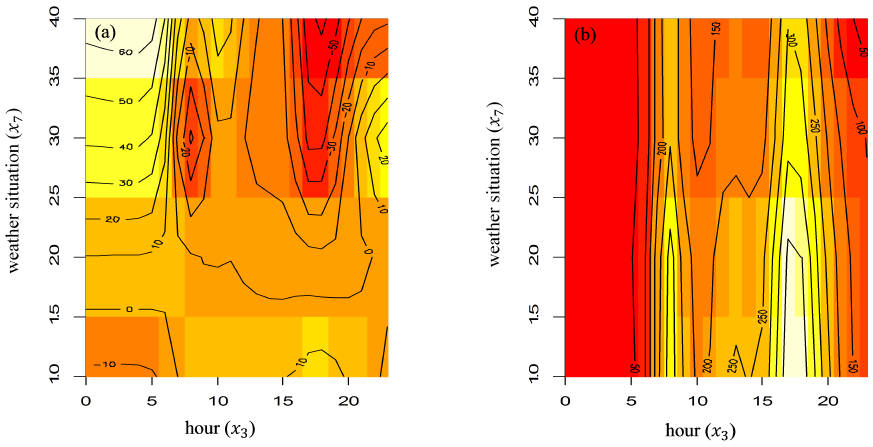


Fig. 10. ALE second-order interaction plots for the predictors hour (X_3) and weather situation (X_7) without (left panel) and with (right panel) the main effects of X_3 and X_7 included. The left panel plots $\hat{f}_{\{3,7\},ALE}(x_3, x_7)$, and the right panel plots $\mathbb{E}[f(\mathbf{X})] + \hat{f}_{3,ALE}(x_3) + \hat{f}_{7,ALE}(x_7) + \hat{f}_{\{3,7\},ALE}(x_3, x_7)$. The numbers on the contours are the function values.

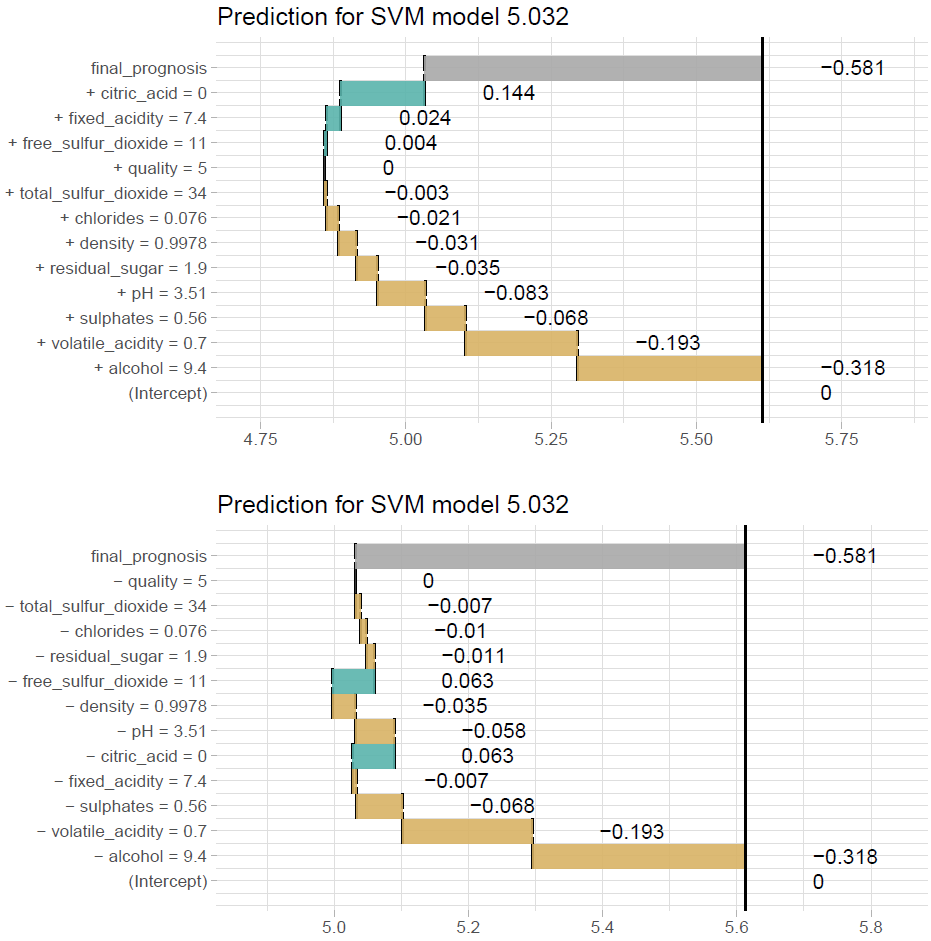


Figure 6: ag-break feature attributions for SVM model calculated for the 5th wine. The upper plot presents feature attributions for the step-up strategy, while the bottom plot presents results for the step-down strategy. Attributions are very similar even if the ordering is different. Vertical black line shows the average prediction for the SVM model. The 5th wine gets final prediction of 5.032 which is below the average for this model by 0.581 point.

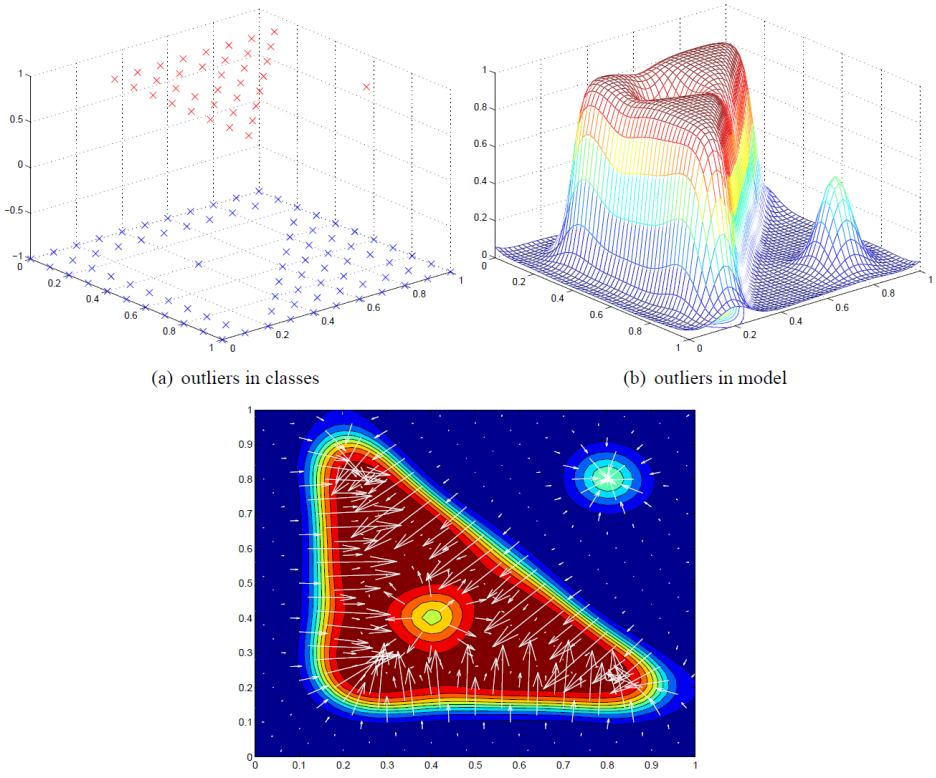
```
library("breakDown")
explain_5 <- broken(wine_svm_model, new_observation = nobs,
                    data = wine,
                    baseline = "intercept",
                    direction = "up")
explain_5
##                                contribution
## baseline                          5.613
## + alcohol = 9.4                  −0.318
## + volatile_acidity = 0.7         −0.193
## + sulphates = 0.56               −0.068
## + pH = 3.51                      −0.083
## + residual_sugar = 1.9           −0.035
## + density = 0.9978               −0.031
## + chlorides = 0.076              −0.021
## + total_sulfur_dioxide = 34      −0.003
## + quality = 5                     0.000
## + free_sulfur_dioxide = 11        0.004
## + fixed_acidity = 7.4             0.024
## + citric_acid = 0                 0.144
## final_prognosis                   5.032

plot(explain_5)
```

Figure 6 shows variable contributions for step-up and step-down strategy. Variable ordering is different but the contributions are consistent across both strategies.

65. High-Dimensional Sparse Data

Absent

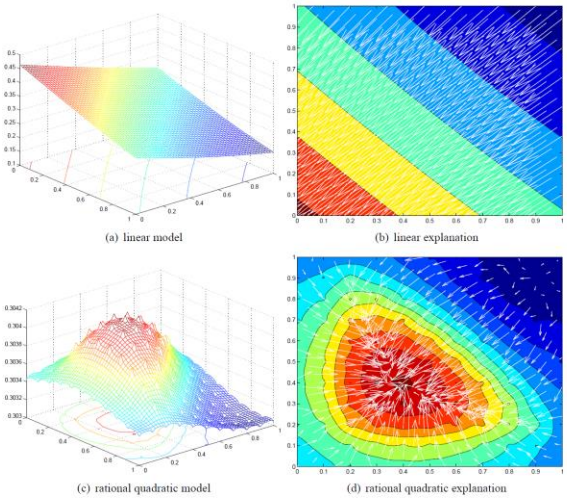
66. Explain Individual Classification



(a) outliers in classes (b) outliers in model

(c) outlier explanation

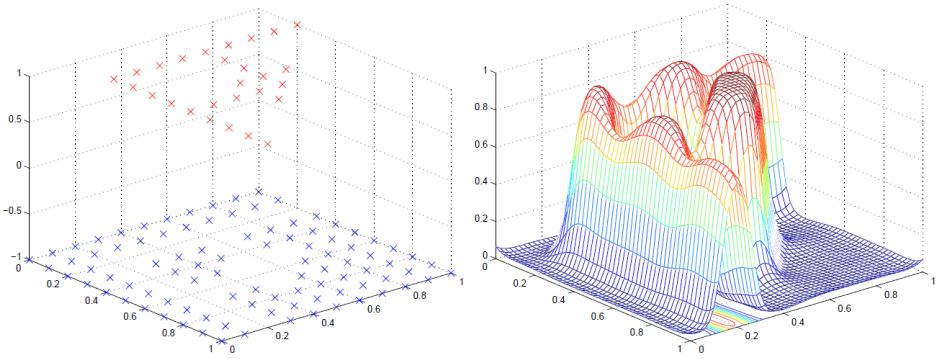
Figure 12: The effect of outliers to the local gradient explanations



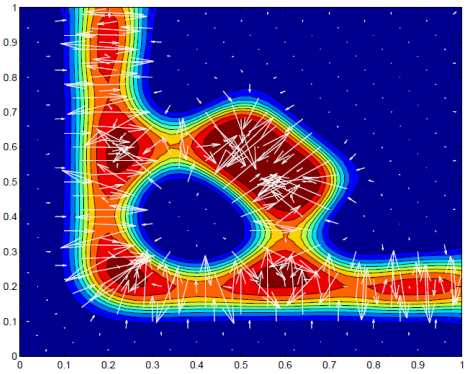
(a) linear model (b) linear explanation

(c) rational quadratic model (d) rational quadratic explanation

Figure 11: The effect of different kernel functions to the local gradient explanations

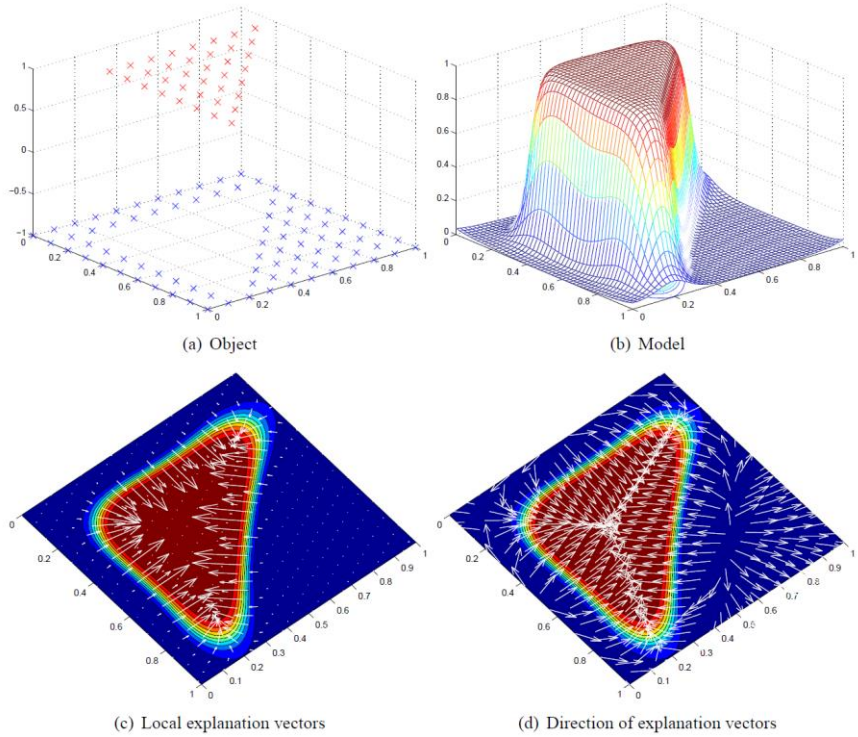


(a) locally non-linear object (b) locally non-linear model



(c) locally non-linear explanation

Figure 13: The effect of local non-linearity to the local gradient explanations



(a) Object (b) Model

(c) Local explanation vectors (d) Direction of explanation vectors

Figure 1: Explaining simple object classification with Gaussian Processes

Figure 4: USPS digits (training set): "twos" (left) and "eights" (right) with correct classification. For each digit from left to right: (i) explanation vector (with black being negative, white being positive), (ii) the original digit, (iii-end) artificial digits along the explanation vector towards the other class.

Figure 5: USPS digits (test set bottom part): "twos" (left) and "eights" (right) with correct classification. For each digit from left to right: (i) explanation vector (with black being negative, white being positive), (ii) the original digit, (iii-end) artificial digits along the explanation vector towards the other class.

67. Diagnosing Bias

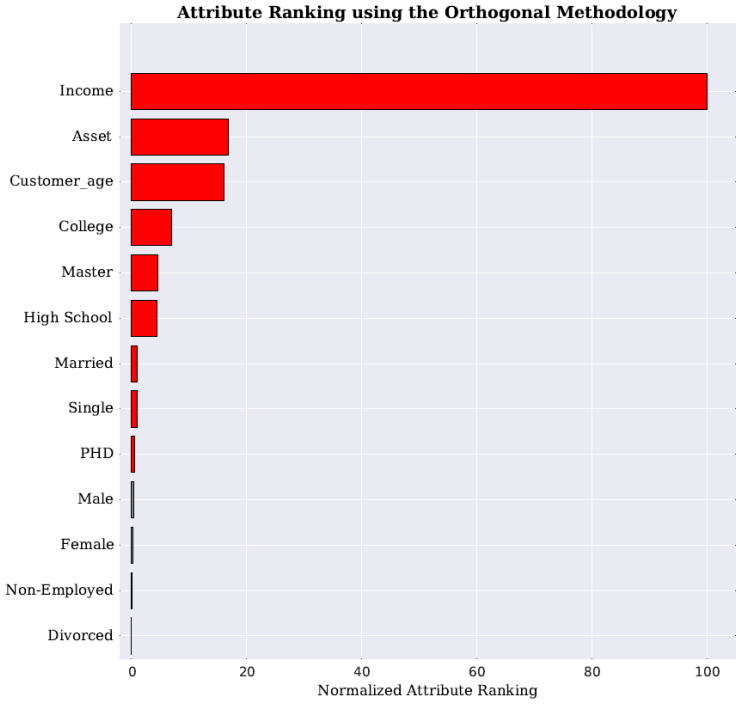


Figure 1. Figure shows the ranking derived from our iterative orthogonal projection algorithm. The rankings have been normalized so that the most significant variable is scaled to 100 and the others relative to the most significant one.

68. Local Rule-Based Explanations (LORE)

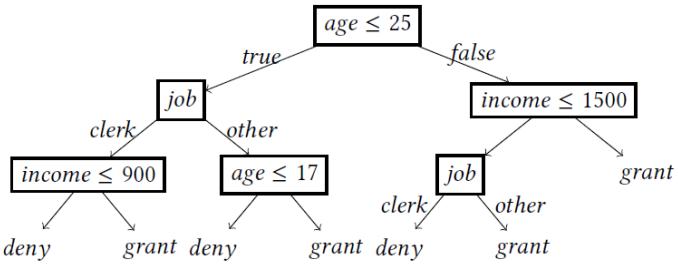


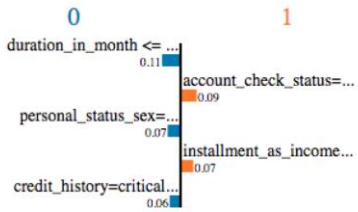
Figure 4: Example decision tree.

- LORE

$r =$ *({credit_amount > 836, housing = own, other_debtors = none, credit_history = critical account} → decision = 0)*

$\Phi =$ *{ ({credit_amount ≤ 836, housing = own, other_debtors = none, credit_history = critical account} → decision = 1), ({credit_amount > 836, housing = own, other_debtors = none, credit_history = all paid back} → decision = 1) }*

- LIME



- Anchor

$a =$ *({credit_history = critical account, duration_in_month ∈ [0, 18.00]} → decision = 0)*

Figure 9: Explanations of LORE, LIME and Anchor.

Dataset	*Method*	*hit*	*fidelity*	*l-fidelity*	*tree depth*
adult	lore	**.912 ± .29**	**.959 ± .17**	**.892 ± .29**	**4.16 ± 0.21**
	global	.901 ± .28	.750 ± .00	.873 ± .27	12.00 ± 0.00
compas	lore	**.942 ± .23**	**.992 ± .03**	**.937 ± .23**	**4.72 ± 2.15**
	global	.902 ± .29	.935 ± .00	.857 ± .29	12.00 ± 0.00
german	lore	**.925 ± .26**	**.988 ± .07**	**.920 ± .26**	**4.95 ± 2.54**
	global	.880 ± .32	.571 ± .00	.824 ± .31	6.00 ± 0.00

Table 2: Local vs global approach.

69. Visual Inspection

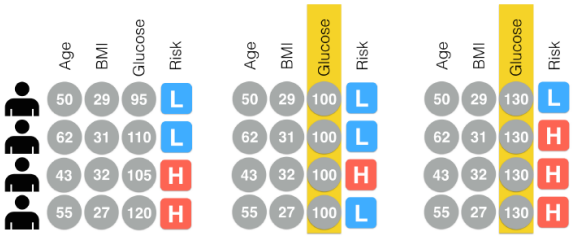


Figure 1. An illustration of how partial dependence is computed for the feature "Glucose". On the left are four patients' original feature values. In the middle, the "Glucose" values are all changed to 100, and the corresponding predictions change. Similarly, on the right, the "Glucose" values are all changed to 130, and again the risks are different. This demonstrates the impact of the "Glucose" feature on risk prediction.

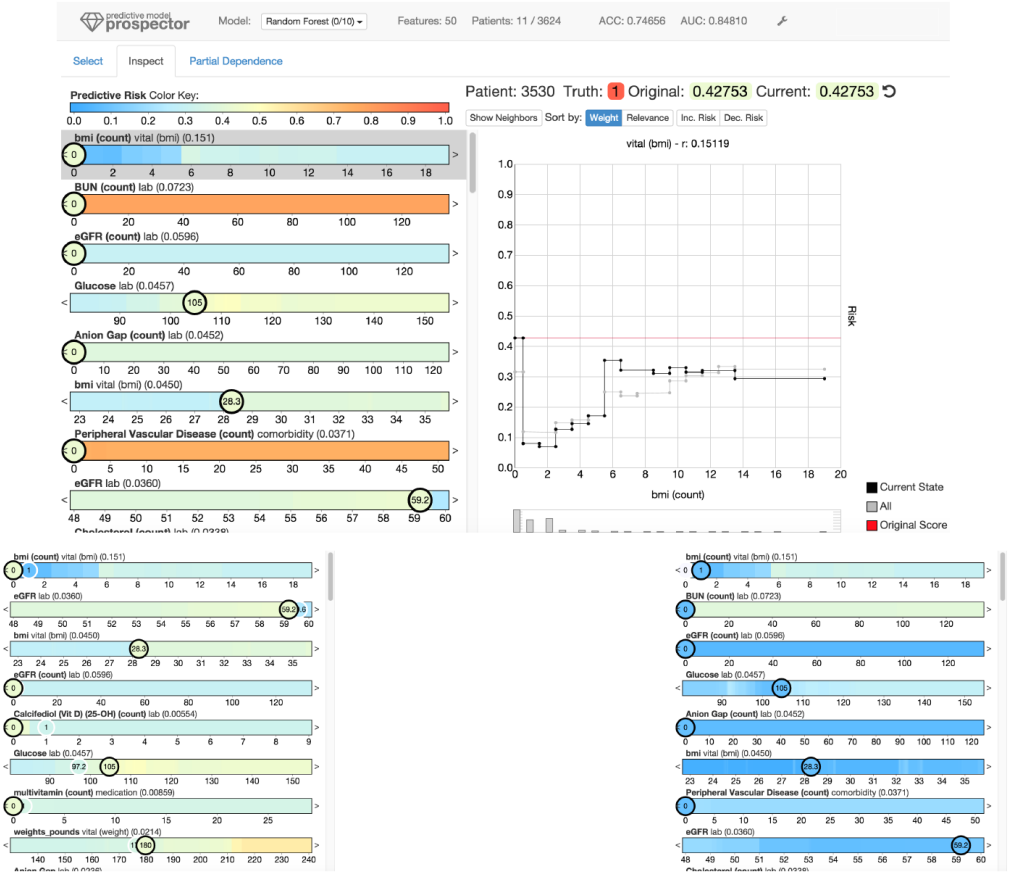


Figure 9. The user interface of *Prospector* is shown at the top. The bottom left shows suggestions on what changes (white circles) would decrease the predicted risk the most. The bottom right shows how the color plots change due to changing a value (namely changing the bmi value from 0 to 1). Fully white circles show the original value of the given patient.

Patient: 3530 Truth: 1 Original: 0.42753

Decreasing Risk:

Feature	Current	Suggested Change
bmi (count) vital (bmi)	0	1 (0.08021)
eGFR lab	59.18887	59.59549 (0.23110)
bmi vital (bmi)	28.27873	28.23937 (0.27954)
eGFR (count) lab	0	1 (0.28705)
Calcifediol (Vit D) (25-O…	0	1 (0.31857)

Increasing Risk:

Feature	Current	Suggested Change
BUN (count) lab	0	1 (0.77246)
Peripheral Vascular Dis…	0	1 (0.68666)
Uric Acid (count) lab	0	1 (0.64202)
Calcium lab	9.37486	9.38749 (0.59175)
Carbon Dioxide lab	26.56109	27.35469 (0.59025)

Figure 8. The summary of one patient. The header line indicates the patient id, the ground truth, and the predicted risk. For both decreasing and increasing the predicted risk the top 5 most impactful features are shown. Each feature shows its current value and the suggested change with the highest impact along with how the predicted risk would change.

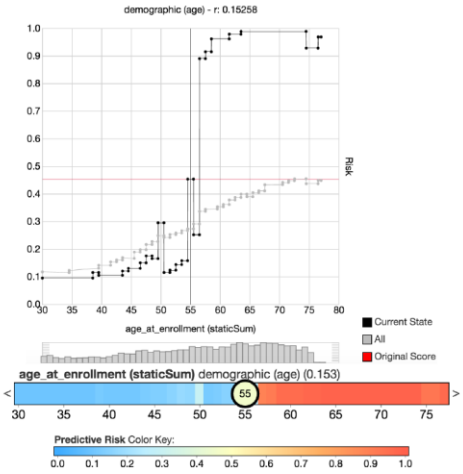


Figure 5. The same feature shown as line plot (top) and partial dependence bar (middle). Color indicates the predicted risk for the outcome. The color mapping is shown at the bottom.

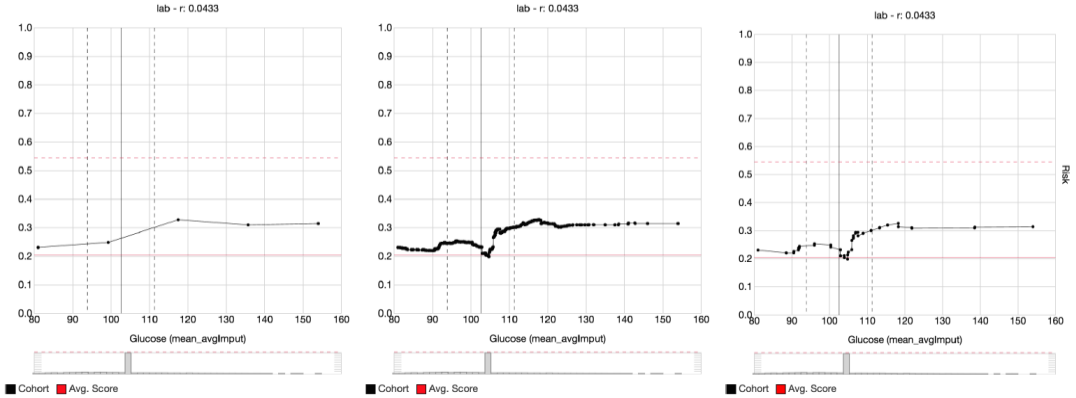


Figure 3. Different sampling strategies for partial dependence plots. The leftmost plot uses a naïve sampling which misses the dip in the predicted risk for a Glucose value around 105. Using the thresholds of the trees in the random forest the middle plot shows all details of the displayed model. The rightmost plot simplifies this by detecting co-linear points and summarizing them into lines improving readability. The dip in the predicted risk is due to imputation of missing values to the mean of the observed values. This increase in local noise shifts the prediction towards the overall average predicted risk (the horizontal red line). Most patients have never had their Glucose value measured.

70. Explaining for individual instances

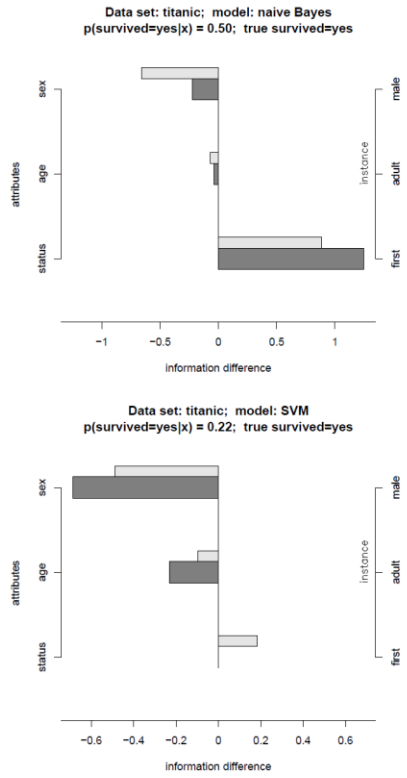


Figure 2: Explanation of NB and SVM models for one of the first class, adult, male passengers in Titanic data set. Explanations for particular instance are depicted with dark bars. Average positive and negative explanations for given attributes' values are presented with light shaded half-height bars above them.

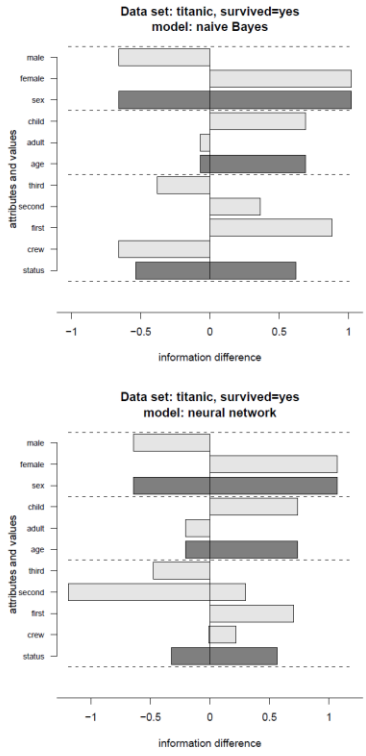


Figure 3: Model explanation of the NB (left-hand side) and ANN (right-hand side) on the Titanic data set. Light bars are average explanation for attributes' values and dark bars are averages (separately for positive and negative scores) over all values of each attribute.

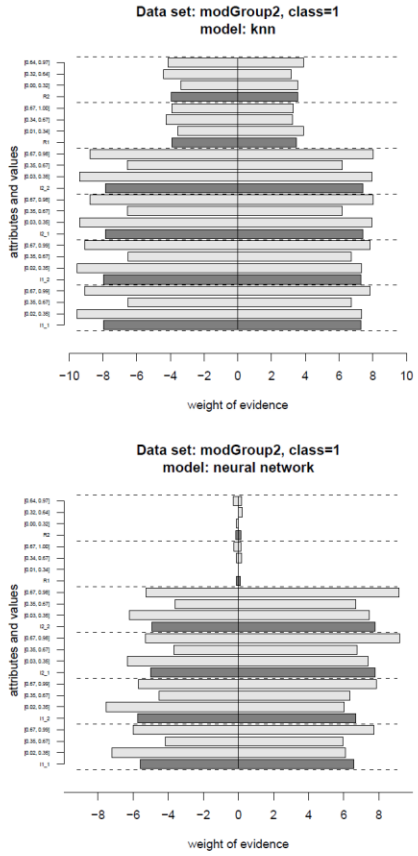


Figure 8: Model explanations for kNN (top) and ANN (bottom) on the groups data set with duplicated important attributes. As attributes are numerical, explanations are averaged and presented for intervals.

71. Anchors

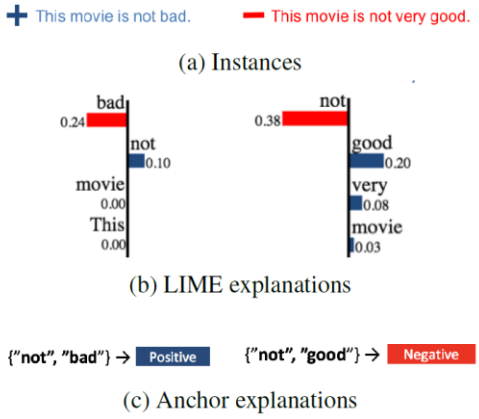


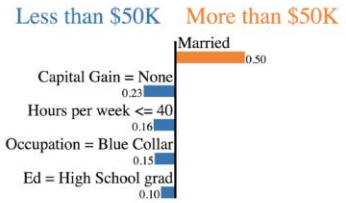
Figure 1: Sentiment predictions, LSTM

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K
recdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score $\leq$ 649	Bad Loan
	649 $\leq$ FICO score $\leq$ 699 and \$5,400 $\leq$ loan amount $\leq$ \$10,000	Good Loan

Table 3: Generated anchors for Tabular datasets

28 < Age ≤ 37
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week ≤ 40.00
Country = United-States
P(**Salary** $>$ \$50**K**) = 0.57

(a) Instance and prediction



(b) LIME explanation

IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 < Age ≤ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > \$50K

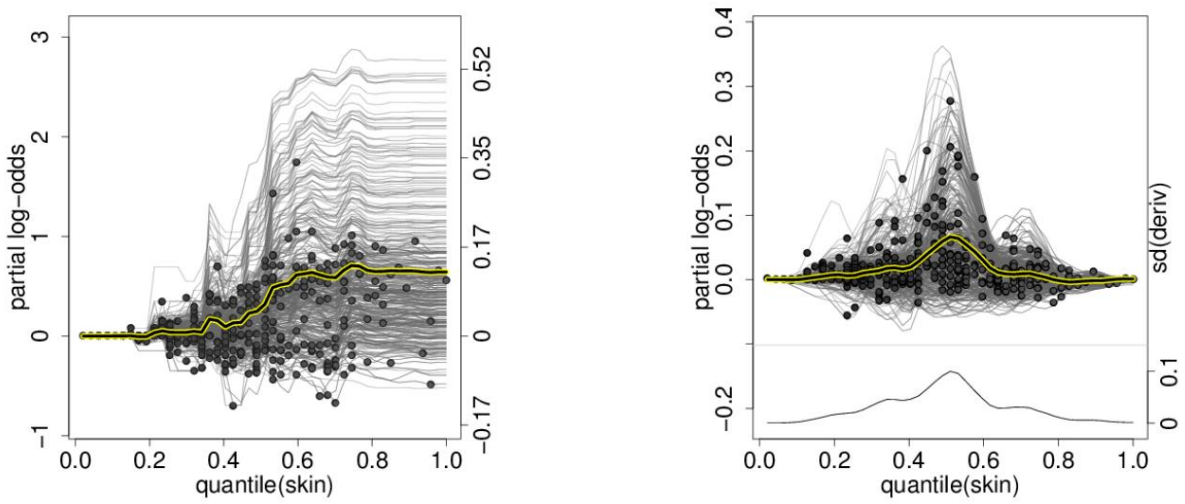
(c) An *anchor* explanation

Figure 5: Explaining a prediction near the decision boundary in the UCI adult dataset.

English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar
This is the **problem** we must address	Este é o problema que temos que enfrentar
This is what we must address	É isso que temos de enfrentar

Table 2: Anchors (in bold) of a machine translation system for the Portuguese word for "This" (in pink).

72. Visualizing statistical learning



(a) c-ICE (b) d-ICE

Figure 12: ICE plots of a `RF` model for estimated centered logit of the probability of contracting diabetes versus `skin` colored by subject `age`.

73. Visualizing the Feature Importance

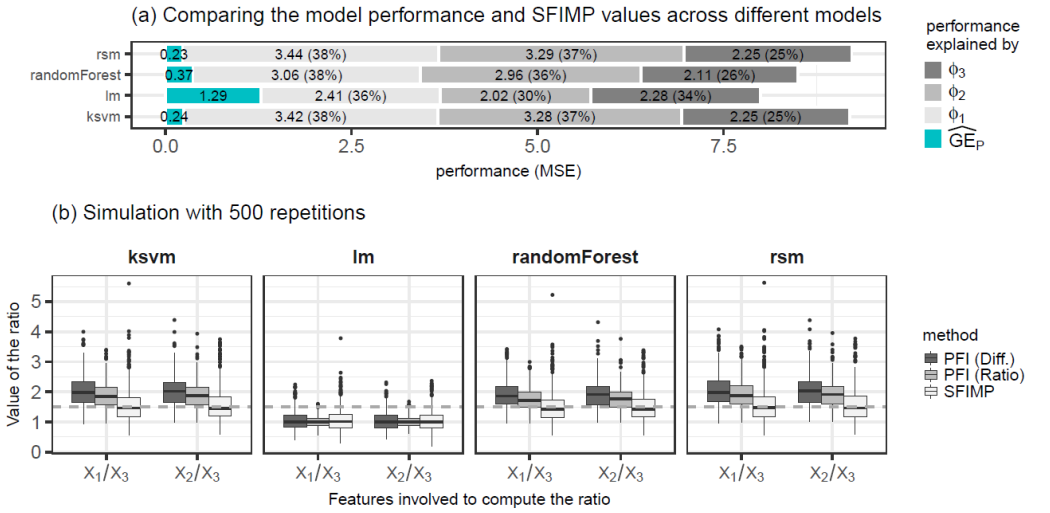
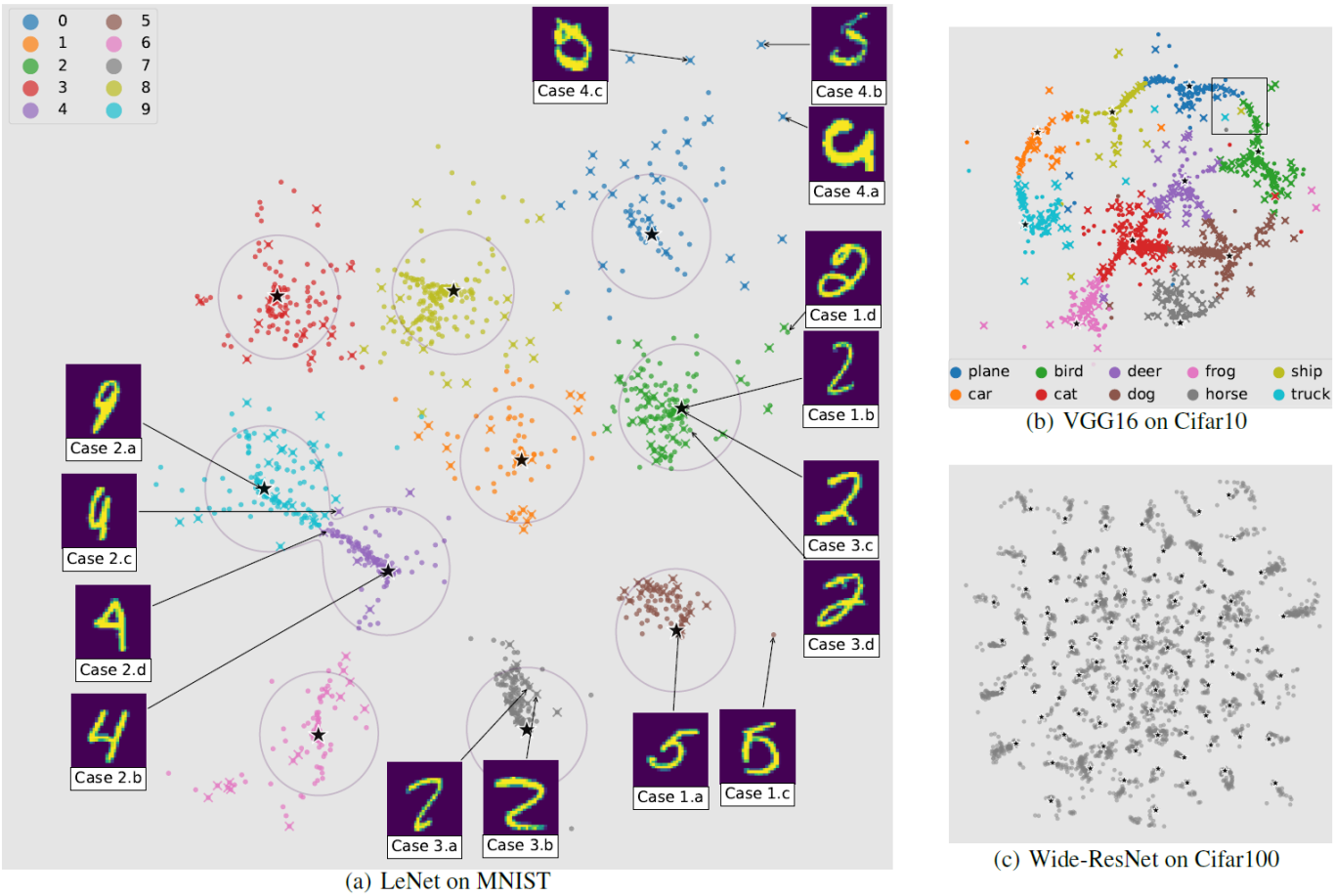


Fig. 5. Panel (a) shows the results of a single run, consisting of sampling test data and computing the importance on the previously fitted models. The first numbers on the left refer to the model performance (MSE) using all features. The other numbers are the SFIMP values which sum up to the total explainable performance $v_{GE}(P)$ from Eq. (10). The percentages refer to the proportion of explained importance. Panel (b) shows the results of 500 repetitions of the experiment. The plots display the distribution of ratios of the importance values for X_1 and X_2 with respect to X_3 computed by SFIMP, by the difference-based PFI, and by the ratio-based PFI.

74. Visual Distillation of Dark Knowledge



(a) LeNet on MNIST

(b) VGG16 on Cifar10

(c) Wide-ResNet on Cifar100

Figure 5. Scatter plots generated by DarkSight for LeNet (MNIST), VGG16 (Cifar10) and Wide-ResNet (Cifar100). For (a) and (b), points are colored by the teacher's predictions. Rounded points means they are correctly classified by the teacher and crossings means they are wrongly classified by the teacher. For (c), we show the monochrome scatter plot simply because there are too many clusters which are hard to assign colors to. Stars in all plots are μs of the Student's t-distributions. In (a), the contour is where $P_S(y_i;\theta)$ equals to 0.001.

75. Using natural frequencies

Absent

76. Trepan

Absent

77. DecText

Absent

78. Pairwise Interactions

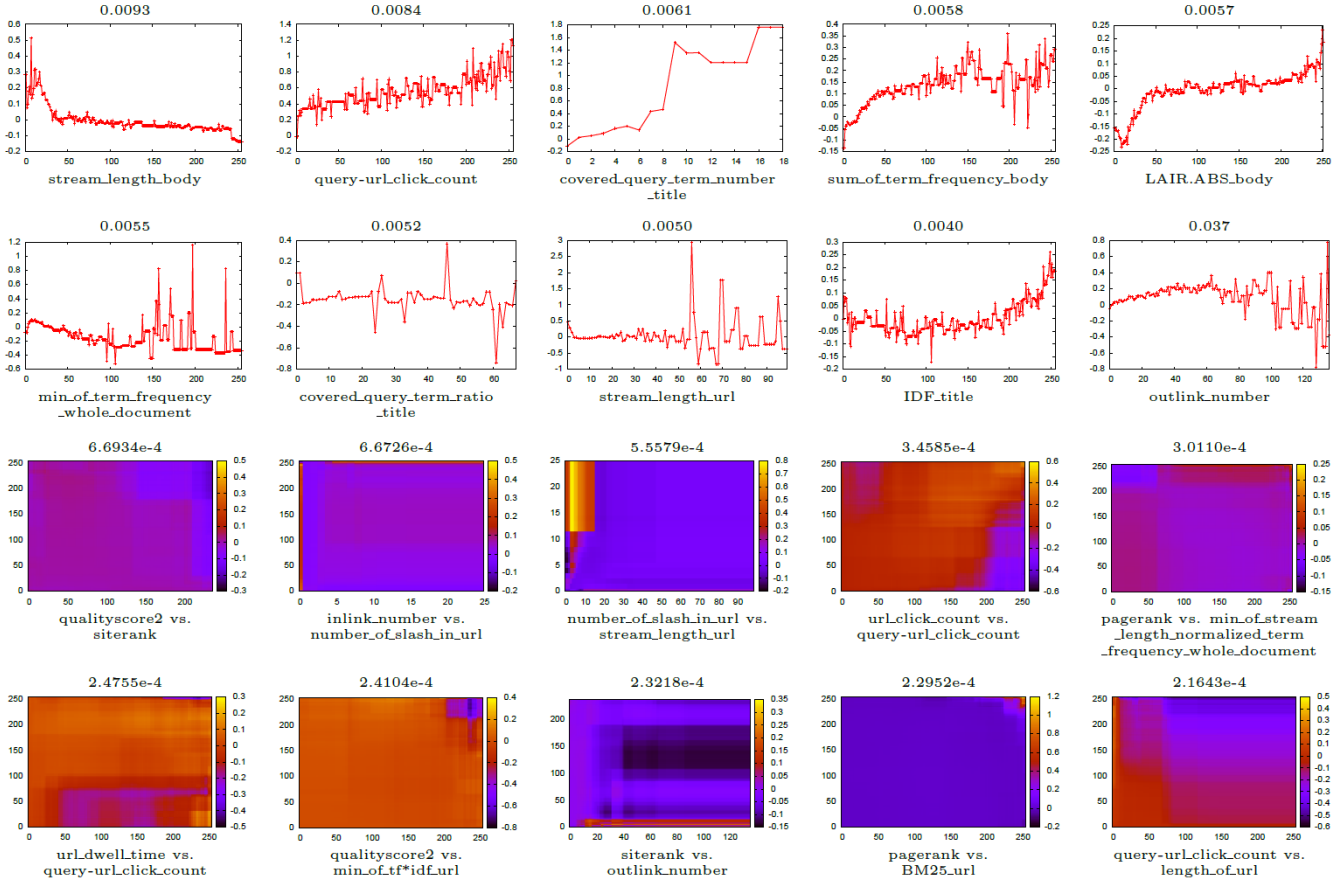


Figure 9: Shapes of features and pairwise interactions for the "MSLR10k" dataset with weights. Top two rows show top 10 strongest features. Next two rows show top 10 strongest interactions.

79. Firm

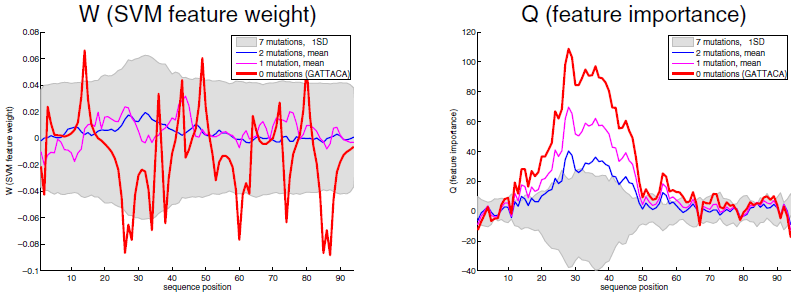


Fig. 3. Feature importance analyses based on (left) the SVM feature weighting $\mathbf{w}$ and (right) FIRM. The shaded area shows the ± 1 SD range of the importance of completely irrelevant features (length 7 sequences that disagree to `GATTACA` at every position). The red lines indicate the positional importances of the exact motif `GATTACA`; the magenta and blue lines represent average importances of all length 7 sequences with edit distances 1 and 2, respectively, to `GATTACA`. While the feature weighting approach cannot distinguish the decisive motiv from random sequences, FIRM identifies it confidently.

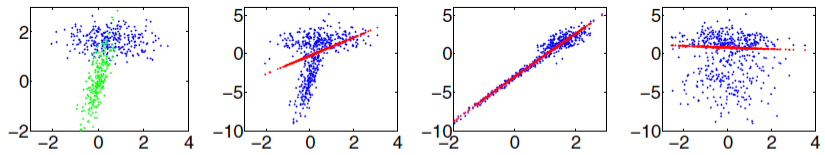


Fig. 2. Binary classification performed on continuous data that consists of two 3d Gaussians constituting the two classes (with x_3 being pure noise). From left to right a) Of the raw data set x_1, x_2 are displayed. b) Score of the linear discrimination function $s(\mathbf{x}_i)$ (blue) and conditional expected score $q_1((\mathbf{x}_i)_1)$ (red) for the first dimension of $\mathbf{x}$. c) $s(\mathbf{x}_i)$ and $q_2((\mathbf{x}_i)_2)$ for varying x_2. As the variance of q is highest here, this is the discriminating dimension (closely resembling the truth). d) $s(\mathbf{x}_i)$ and $q_3((\mathbf{x}_i)_3)$ for varying x_3. Note that x_3 is the noise dimension and does not contain discriminating information (as can be seen from the small slope of q_3).

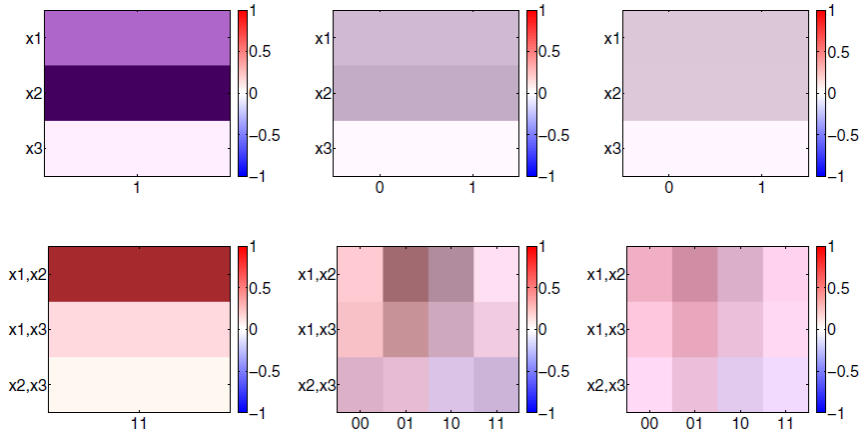


Fig. 1. FIRMs and SVM-w for the Boolean formula $x_1 \vee (\neg x_1 \wedge \neg x_2)$. The figures display heat maps of the scores, blue denotes negative label, red positive label, white is neutral. The upper row of heat maps shows the scores assigned to a single variable, the lower row shows the scores assigned to pairs of variables. The first column shows the SVM-w assigning a weight to the monomials x_1, x_2, x_3 and x_1x_2, x_1x_3, x_2x_3 respectively. The second column shows FIRMs obtained from the trained SVM classifier. The third column shows FIRMs obtained from the true labeling.

80. MFI

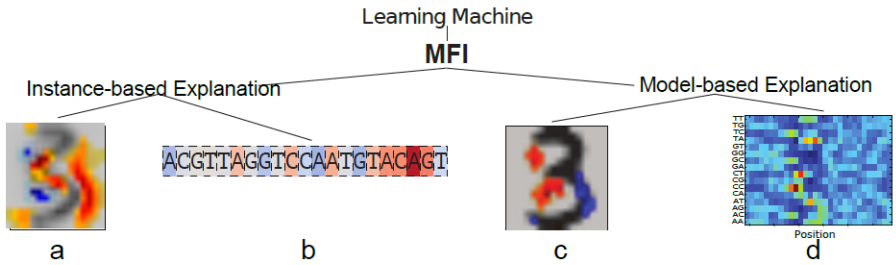
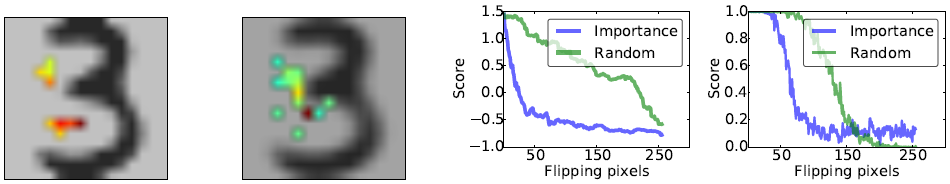


Figure 1: **MFI Examples** We consider two possible flavors of feature importance: (left) instance-based importance measures (e.g. *Why is this specific example of '3' classified as '3' using my trained RBF-SVM classifier?*); (right) model-based importance measure (e.g. *Which regions are generally important for the classifier decision?*).

In the following, we will explain the "explanation mode" of above definition in terms of model-based and instance-based proceed exemplary for both, sequence and image data.



(a) SVM (b) CNN (c) SVM (d) CNN

Figure 3: **Results are shown for the USPS data set using kernel MFI for SVM (a) and CNN (b)**, where the most important pixels found by kernel MFI are embedded in the mean picture of digit `three`. Figure (c) and (d) show the classifier performance loss when successively blurring the pixel regarding their relevance found by kernel MFI compared to a random pixel blurring.

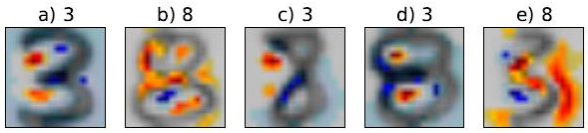


Figure 4: **Instance-based explanation of the SVM decision for five USPS test data images**. The highlighted pixels are informative for the individual SVM decisions (plotted at the image top) – only the first two images were correctly classified.

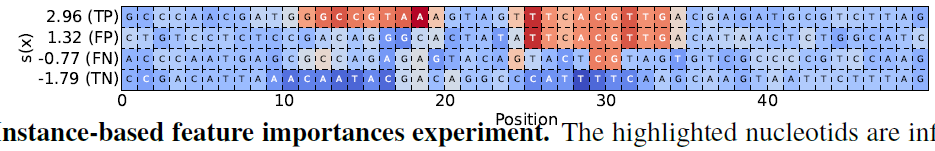


Figure 5: **Instance-based feature importances experiment.** The highlighted nucleotids are informative for the SVM decision for four test sequences that have been correctly (TN and TP) and incorrectly (FP and FN) classified.

81. MES

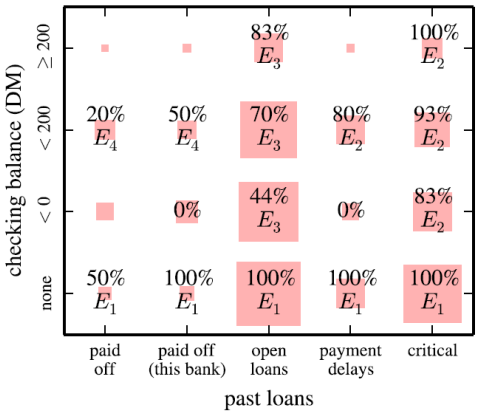
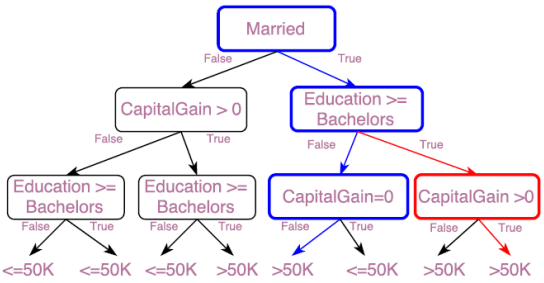


Figure 3. MES applied to German credit data with LR classifier f. The shaded boxes represent the marginal distribution on the two variables (past loans and checking balance). The area is proportional to the frequency in the training data. The percentages show how often test points with those values result in a classification of 1 by f. We show the *most common* explanation for data points in each box. The explanations within a box vary as there are another 18 features not plotted. The explanations are: E_1 individual has no checking account; E_2 past payment delays or worse; E_3 individual already has loans out; E_4 loan duration less than 22 months. It is unclear why shorter loans are more likely to be predicted as risky by the model. However, E_4 is only used 1% of the time and for individuals who are otherwise low risk.

Figure 2. Example of MES explaining a correct prediction of Powell by the (nonlinear) SVM classifier. This example used extended MES (Algo. 3 followed by Algos. 1 and 2) to learn the optimal linear explanation. We subtract out the explanation face (**right**) from the (mean removed) original (**left**) to make the image on the **far left**. In these images: gray $= 0$, white > 0, and black < 0. The product image (**far right**) is the Hadamard product of the original face and the explanation face. Here, the explanation is that the product image has net white balance $1.6\% > 0.5\%$, with a score of $S = 0.865$. We have added the red annotations as cues to the reader on the important areas. **Technical details:** The above images are created as follows: Let $\mathbf{x}$ be the mean removed input face (left) reshaped as a vector. This is transformed by PCA to get $\mathbf{x}_{\text{PCA}} := \mathbf{C}\mathbf{x}$, where $\mathbf{C}$ is the principal component matrix. The explanation is: $\mathbf{w}^\top \mathbf{x}_{\text{PCA}} > a$. Thus we set the right image to be $\mathbf{x}_E := \mathbf{C}^\top \mathbf{w}$. We then set the far right image to be $\mathbf{x}_H := \mathbf{x}_E \odot \mathbf{x}$. Then the explanation becomes: $\mathbf{x}_E \cdot \mathbf{x} = \sum \mathbf{x}_H > a$. We set the corrected image to be $\mathbf{x}_F := \mathbf{x} - \alpha \mathbf{x}_E / ||\mathbf{x}_E||^2$. When applying the explanation to the corrected image we get: $\mathbf{x}_E \cdot \mathbf{x}_F = \mathbf{x}_E \cdot \mathbf{x} - \alpha$. Thus, by setting $\alpha > \sum \mathbf{x}_H - a$, the explanation is false: $E(\mathbf{x}_F) = 0$. Here, $\alpha = 2$.

82. Programs as Black-Box



(a) Decision tree (with the path highlighted)

Random Forests (true implies > 50K):

```
(if HoursPerWeek<=40:
   CapitalGain>0
else: True) and Married
```

Decision Tree (true implies > 50K):

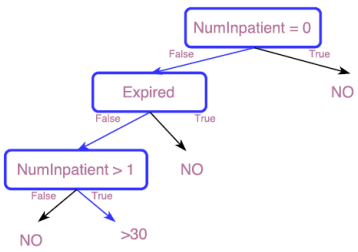
```
CapitalGain>0 and Married
```

Linear model (true implies > 50K):

```
if CapitalGain>0: Married
else : False
```

(b) Generated program explanations

Figure 3: **Adult dataset:** In (a), we show the learned tree, with the path for the instance in blue, and in red, we show that `Education` doesn't really matter for this instance. (b) shows the explanations for three classifiers (they got the prediction right), in particular showing that the explanation for the decision tree gets the more compact form.



(a) Decision tree (with the path highlighted)

Random Forests (true implies < 30):

```
if Diag:Other and not Tolbutamide:
    Discharged:Home
else: Diag:Other
```

Decision Tree (true implies > 30):

```
NumInpatient > 1.00
```

Linear model (true implies NO):

```
not Tolazamide
```

(b) Generated program explanations

Figure 4: **Hospital Readmission data:** (a) shows the learned tree, with the path for the instance in blue. Again, (b) shows the explanations for three classifiers (only the tree had the correct prediction), with the compact explanation for tree almost correct, except that it assumes the patient is alive.

83. A randomization approach

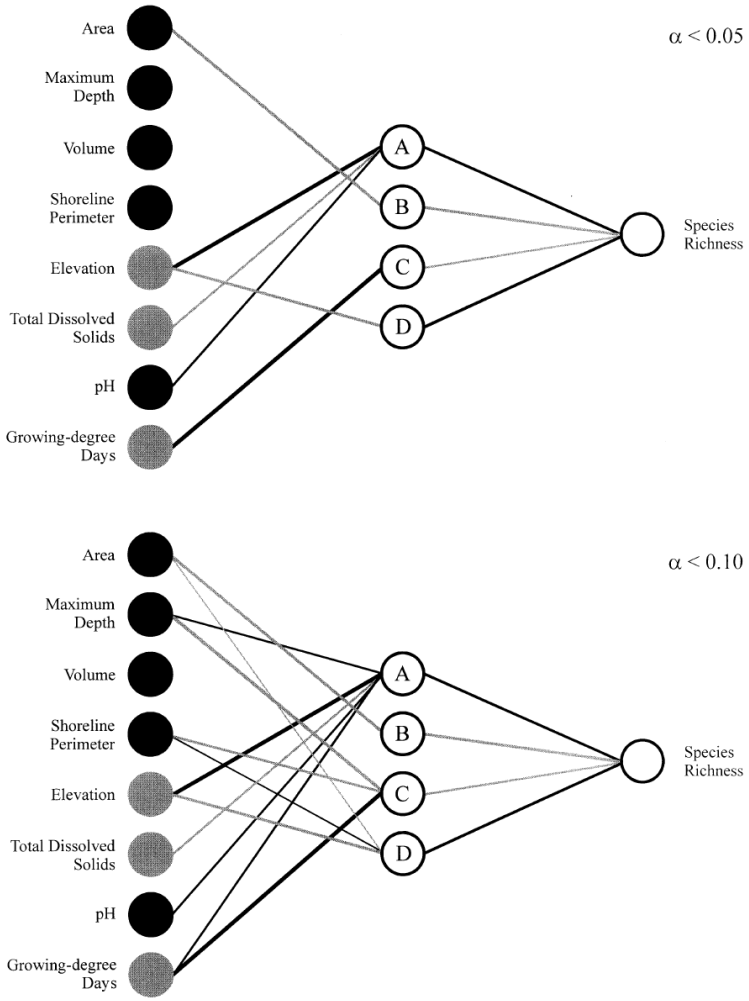


Fig. 5. NID after non-significant input-hidden-output connection weights are eliminated using the randomization test (i.e. connection weights statistically different from zero based on $\alpha = 0.05$ and $\alpha = 0.10$). The thickness of the lines joining neurons is proportional to the magnitude of the connection weight, and the shade of the line indicates the direction of the interaction between neurons: black connections are positive (excitator) and gray connections are negative (inhibitor). Black input neurons indicate habitat variables that have an overall positive influence on species richness, and gray input neurons indicate an overall negative influence on species richness (based on overall connection weights).

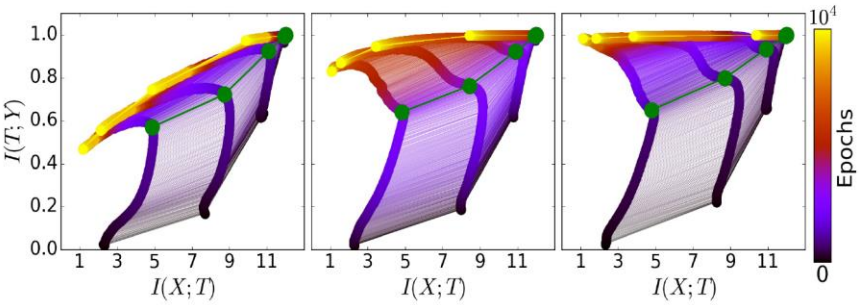


Figure 3: The evolution of the layers with the training epochs in the information plane, for different training samples. On the left - 5% of the data, middle - 45% of the data, and right - 85% of the data. The colors indicate the number of training epochs with Stochastic Gradient Descent from 0 to 10000. The network architecture was fully connected layers, with widths: input=12-10-8-6-4-2-1=output. The examples were generated by the spherical symmetric rule described in the text. The green paths correspond to the SGD drift-diffusion phase transition - grey line on Figure 4

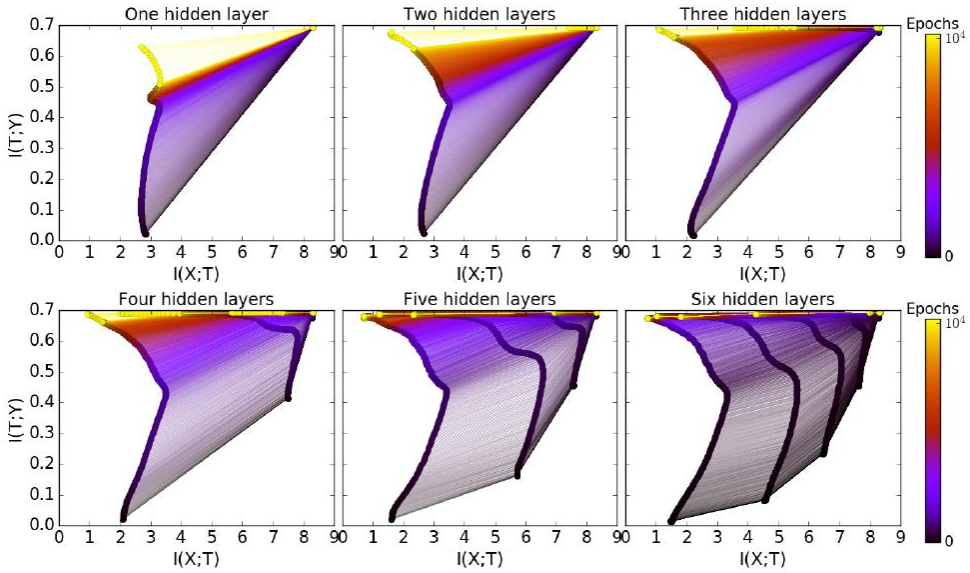


Figure 5: **The layers information paths during the SGD optimization for different architectures.** Each panel is the *information plane* for a network with a different number of hidden layers. The width of the hidden layers start with 12, and each additional layer has 2 fewer neurons. The final layer with 2 neurons is shown in all panels. The line colors correspond to the number of training epochs.

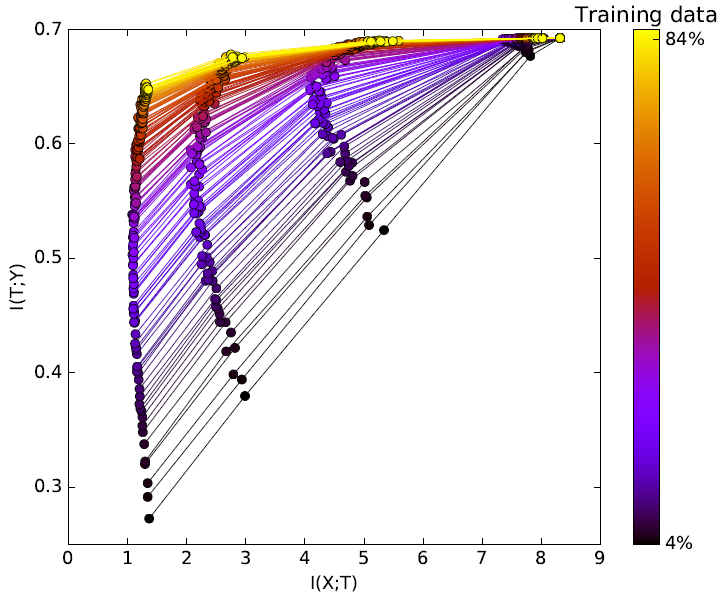


Figure 7: **The effect of the training data size on the layers in the *information plane*.** Each line (color) represents a converged network with a different training sample size. Along each line there are 6 points for the different layers, each averaged over 50 random training samples and randomized initial weights.

85. Generate Reviews

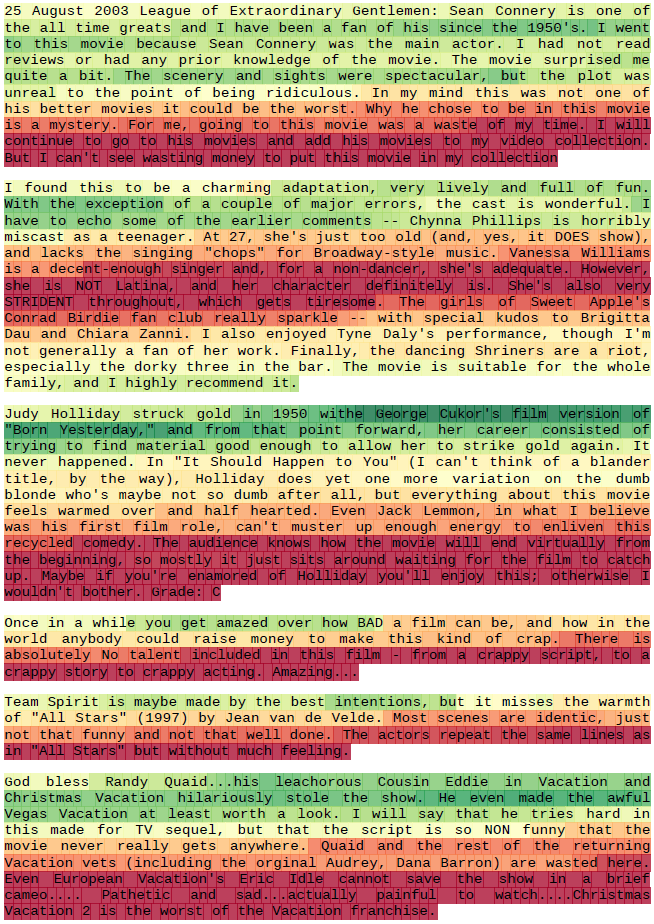


Figure 4. Visualizing the value of the sentiment cell as it processes six randomly selected high contrast IMDB reviews. Red indicates negative sentiment while green indicates positive sentiment. Best seen in color.