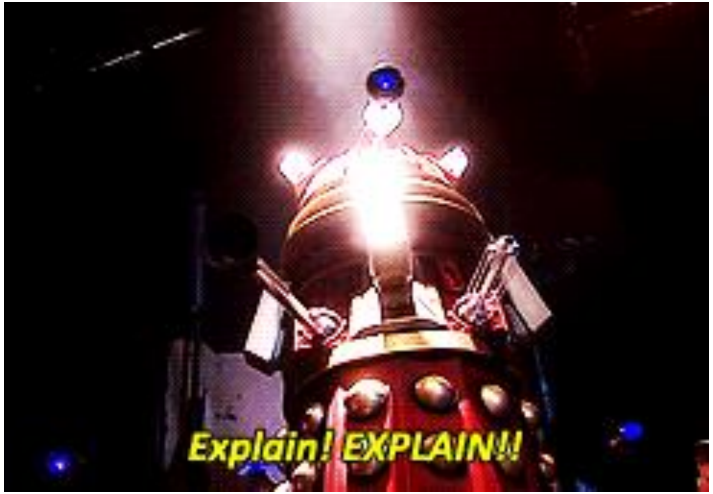
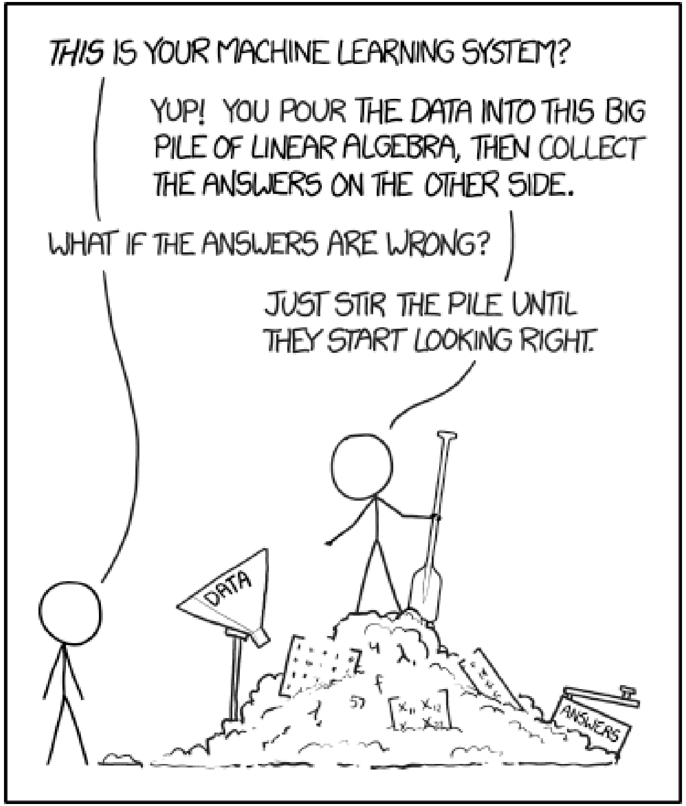
Interpretability in Machine Learning



Why Interpret ?

The current state of machine learning

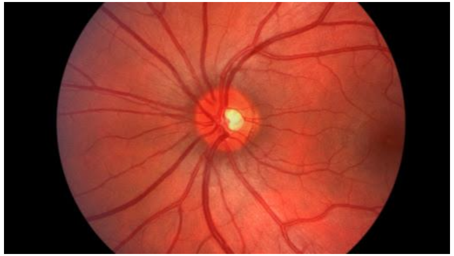


And its uses ...

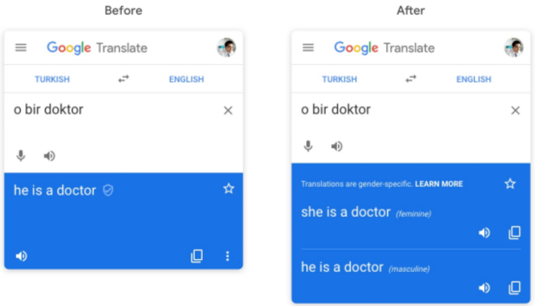
https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long


NYPost

MIT Technology Review


DeepMind


DeepMind



So are we in the golden age of AI ?

Safety and well being

Tesla hit parked police car 'while using Autopilot'

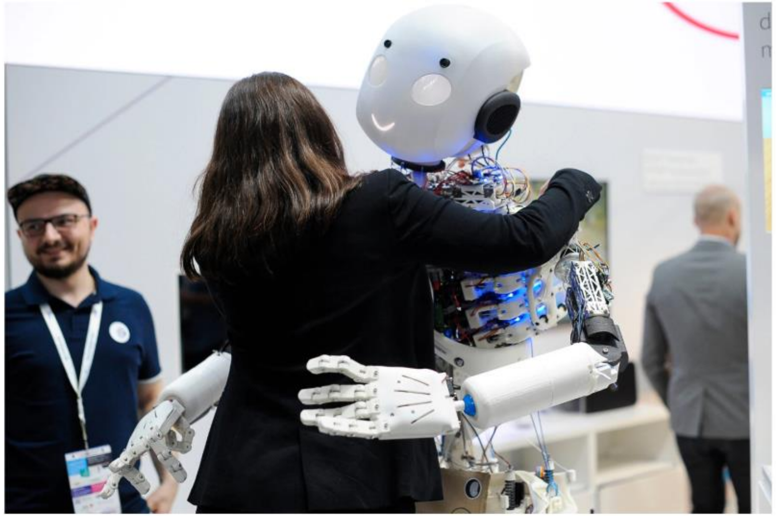
30 May 2018

Share

LAGUNA BEACH POLICE DEPARTMENT

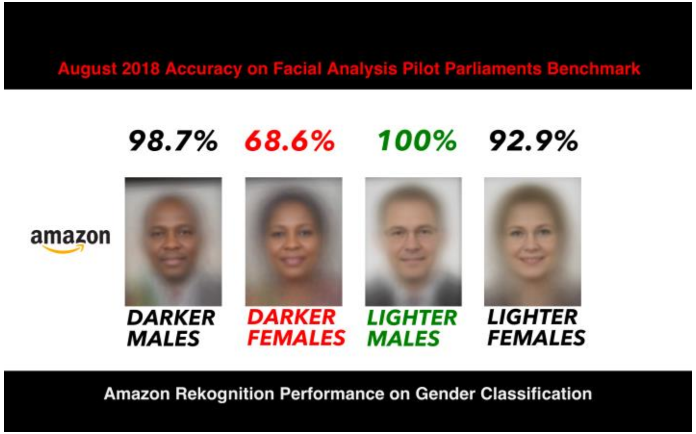
A number of Tesla vehicles have been involved in crashes.

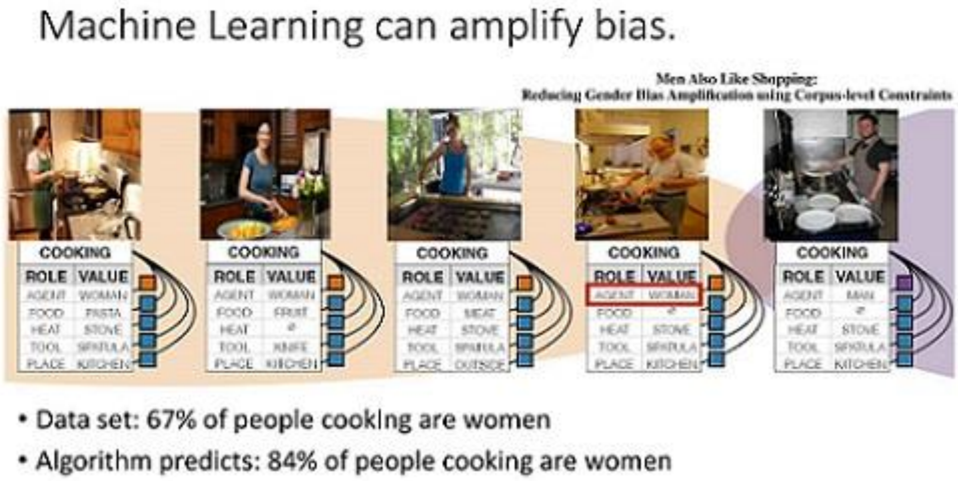
Warnings of a Dark Side to A.I. in Health Care



Scientists worry that with just tiny tweaks to data, neural networks can be fooled into committing "adversarial attacks" that mislead rather than help. Joan Cros/NurPhoto, via Getty Images

Bias in algorithms

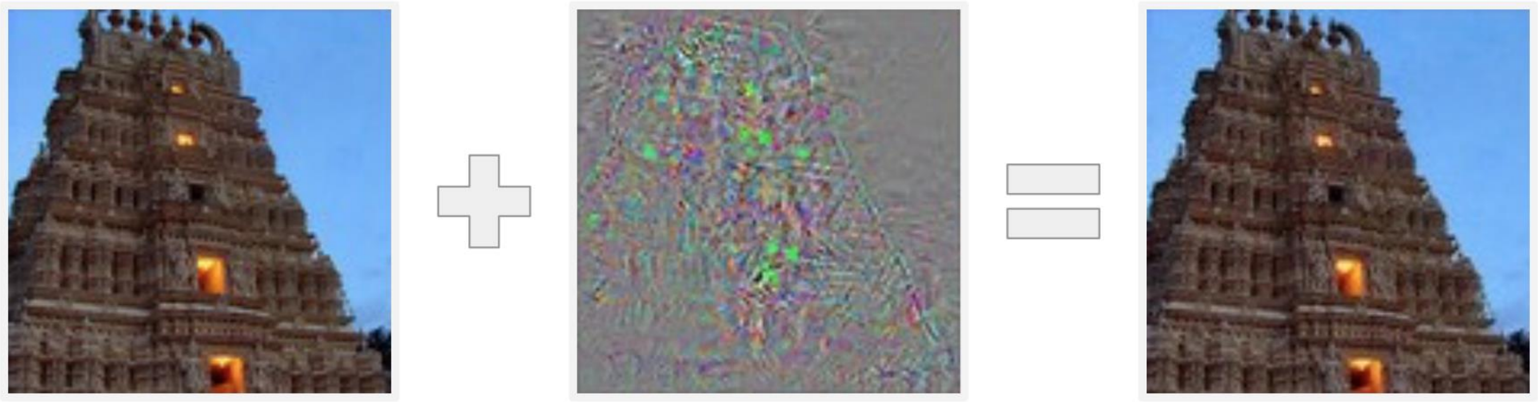




https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

https://www.infoq.com/presentations/unconscious-bias-machine-learning/

Adversarial Examples



Original image Temple (97%)

Perturbations

Adversarial example Ostrich (98%)

Legal Issues - GDPR

And more ...

- Interactive feedback - can model learn from human actions in online setting ? (Can you tell a model to not repeat a specific mistake ?)

- Recourse – Can a model tell us what actions we can take to change its output ? (For example, what can you do to improve your credit score?)

In general, it seems like there are few fundamental problems –

- We don’t trust the models
- We don’t know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model makes similar mistakes as humans ?
- How to change model when things go wrong ?

Interpretability is one way we try to deal with these problems

What is interpretability ?

There is no standard definition –

Most agree it is something different from performance.

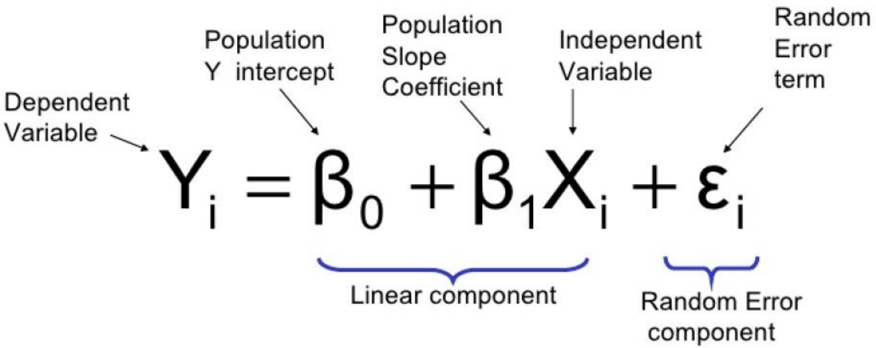
Ability to explain or to present a model in understandable terms to humans (Doshi-Velez 2017)

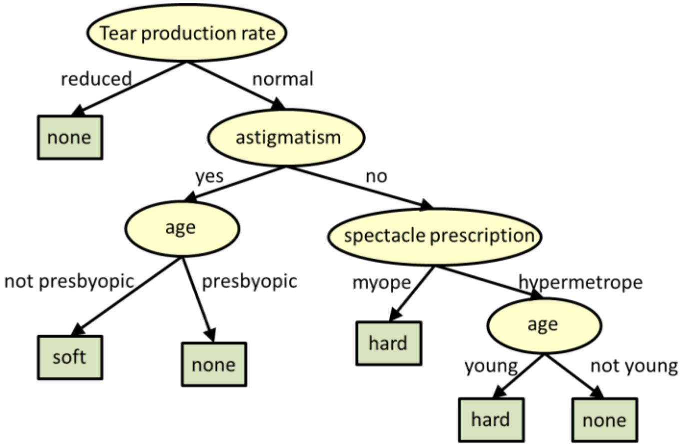
Cynical view – It is what makes you feel good about the model.

It really depends on target audience.

What does interpretation looks like ?

- In pre-deep learning models, some models are considered “interpretable”





What does interpretation look like ?

- Heatmap Visualization

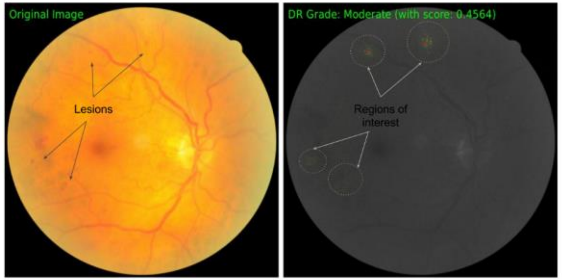


Figure 3. **Attribution for Diabetic Retinopathy grade prediction from a retinal fundus image.** The original image is show on the left, and the attributions (overlayed on the original image in gray scaee) is shown on the right. On the original image we annotate lesions visible to a human, and confirm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

[Sundarajan 2017]

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

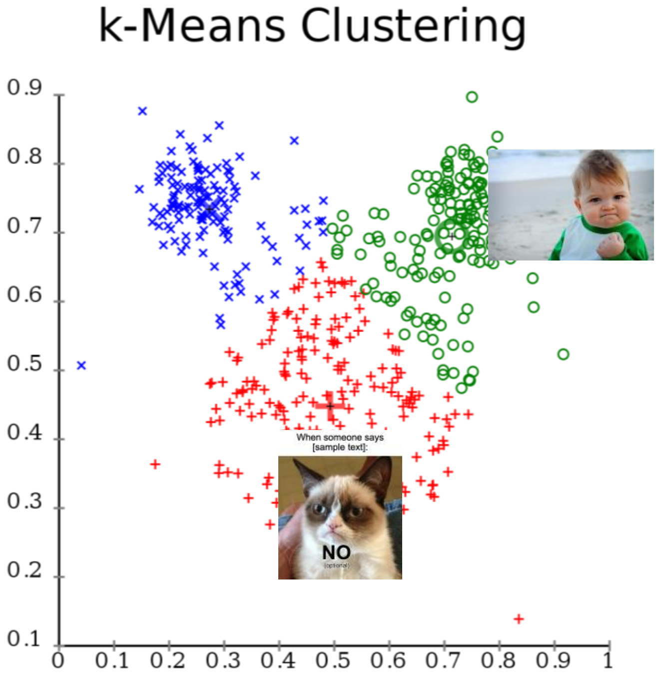
Table 2: Gate activations for each aspect in a PICC abstract. Note that because gates are calculated a the final convolution layer, activations are not i exact 1-1 correspondence with words.

[Jain 2018]

What does interpretation looks like ?

- Give prototypical examples

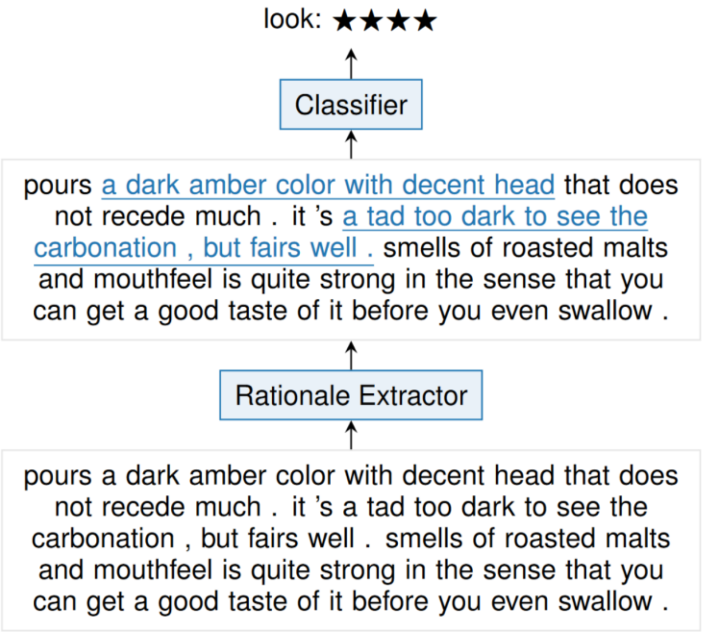
[Kim 2016]



By Chire - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11765684

What does interpretation look like ?

- Bake it into the model



[Bastings et al 2019]

What does interpretation looks like ?

- Provide explanation as text

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Table 1: Examples from our CoS-E dataset.

[Rajani et al 2019]

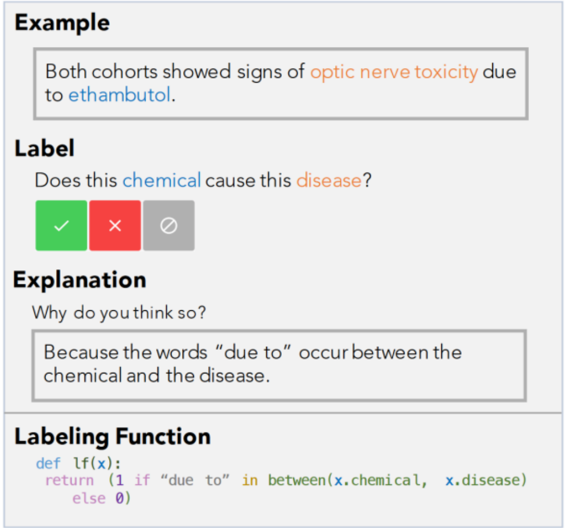


Figure 1: In BabbleLabble, the user provides a natural language explanation for each labeling decision. These explanations are parsed into labeling functions that convert unlabeled data into a large labeled dataset for training a classifier.

[Hancock et al 2018]

Some properties of Interpretations

- **Faithfulness** - how to provide explanations that accurately represent the true reasoning behind the model's final decision.

- **Plausibility** – Is the explanation correct or something we can believe is true, given our current knowledge of the problem ?

- **Understandable** – Can I put it in terms that end user without in-depth knowledge of the system can understand ?

- **Stability** – Does similar instances have similar interpretations ?

Evaluating Interpretability [Doshi-Velez 2017]

- Application level evaluation – Put the model in practice and have the end users interact with explanations to see if they are useful .
- Human evaluation – Set up a Mechanical Turk task and ask non-experts to judge the explanations
- Functional evaluation – Design metrics that directly test properties of your explanation.

How to “interpret” ? Some definitions

Global vs Local

- **Do we explain individual prediction ?**

Example –

Heatmaps
Rationales

- **Do we explain entire model ?**

Example –

Prototypes
Linear Regression
Decision Trees

Inherent vs Post-hoc

- **Is the explainability built into the model ?**

Example –

Rationales
Linear Regression
Decision Trees
Natural Language Explanations

- **Is the model black-box and we use external method to try to understand it ?**

Example –

Heatmaps (Some forms)
Prototypes

Model based vs Model Agnostic

- **Can it explain only few classes of models ?**

Example –

Rationales
LR / Decision Trees
Attention
Gradients (Differentiable Models only)

- **Can it explain any model ?**

Example –

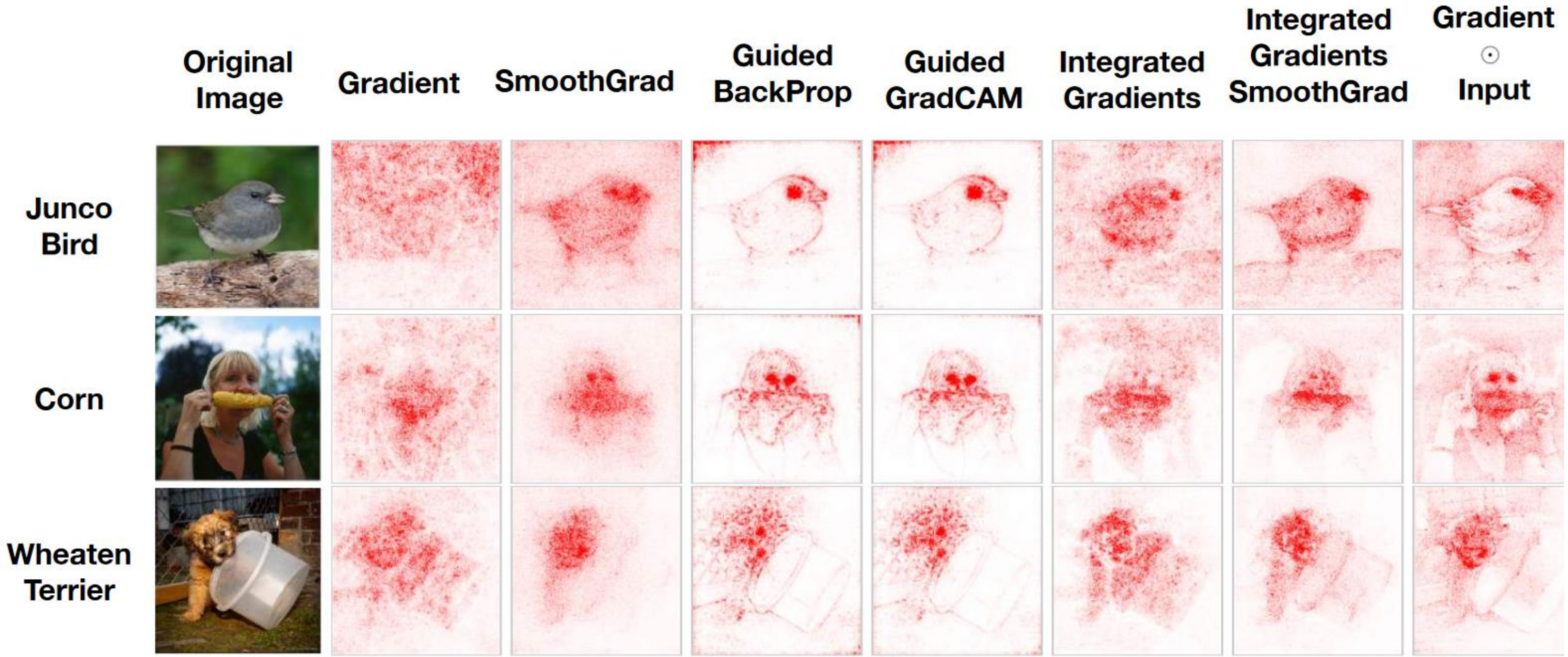
LIME – Locally Interpretable Model Agnostic Explanations

SHAP – Shapley Values

Some
Locally Interpretable,
Post-hoc
methods

Saliency Based Methods

- Heatmap based visualization
- Need differentiable model in most cases
- Normally involve gradient



[Adebayo et al 2018]

Saliency Example - Gradients

$$f(x): R^d \to R$$

$$E(f)(x) = \frac{df(x)}{dx}$$

How do we take gradient with respect to words ?

Take gradient with respect to embedding of the word .

Saliency Example – Leave-one-out

$$f(x): R^d \to R$$

$$E(f)(x)_i = f(x) - f(x\backslash \mathrm{i})$$

How to remove ?

1. Zero out pixels in image
2. Remove word from the text
3. Replace the value with population mean in tabular data

Problems with Saliency Maps

- Only capture first order information
- Strange things can happen to heatmaps in second order.

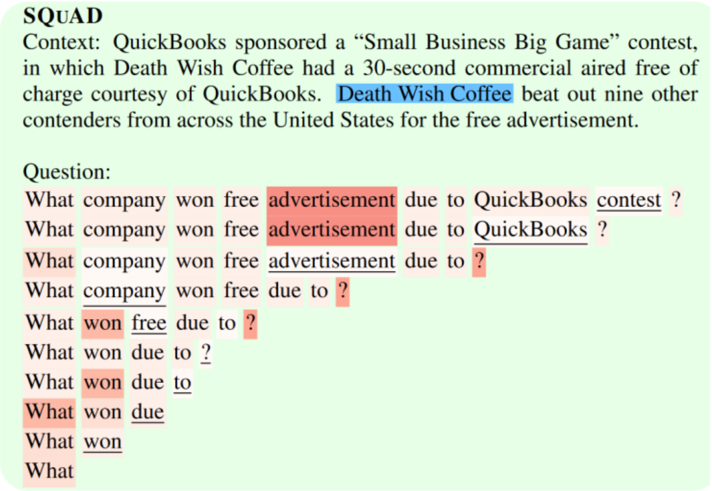
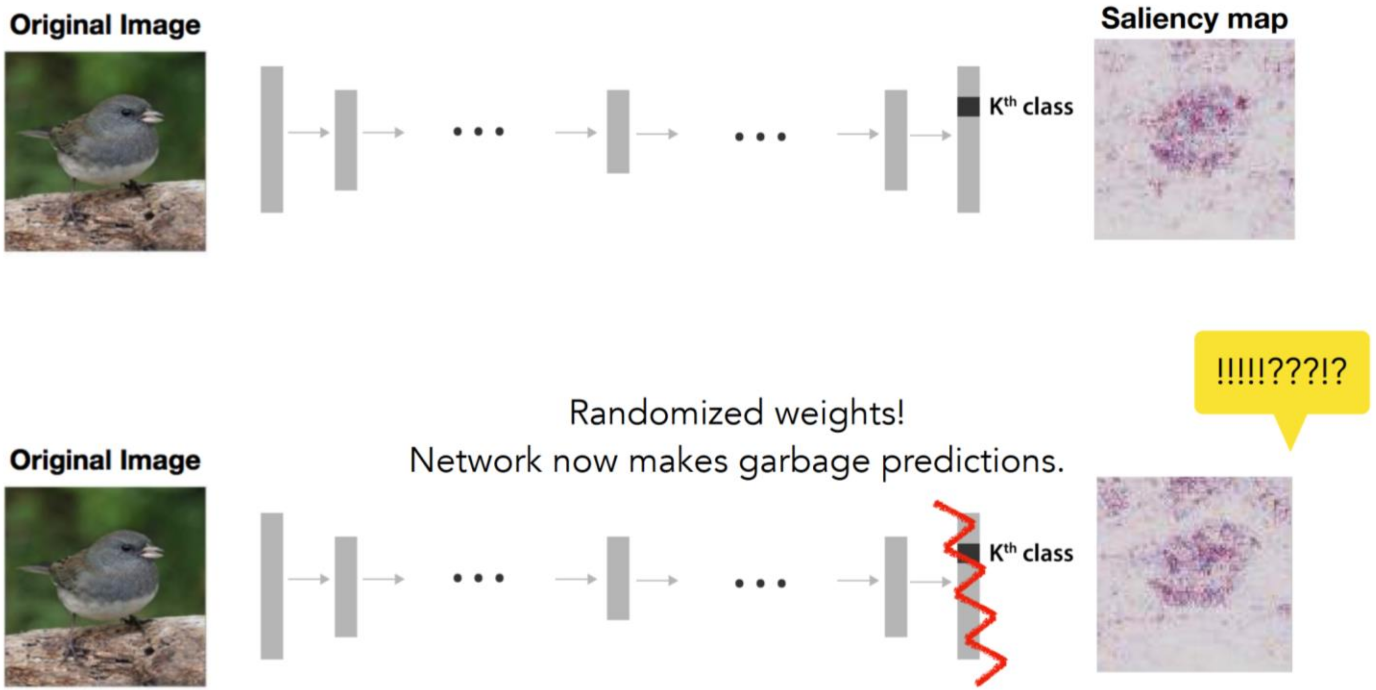


Figure 6: Heatmap generated with leave-one-out shifts drastically despite only removing the least important word (underlined) at each step. For instance, “advertisement”, is the most important word in step two but becomes the least important in step three.

[Feng et al 2018]

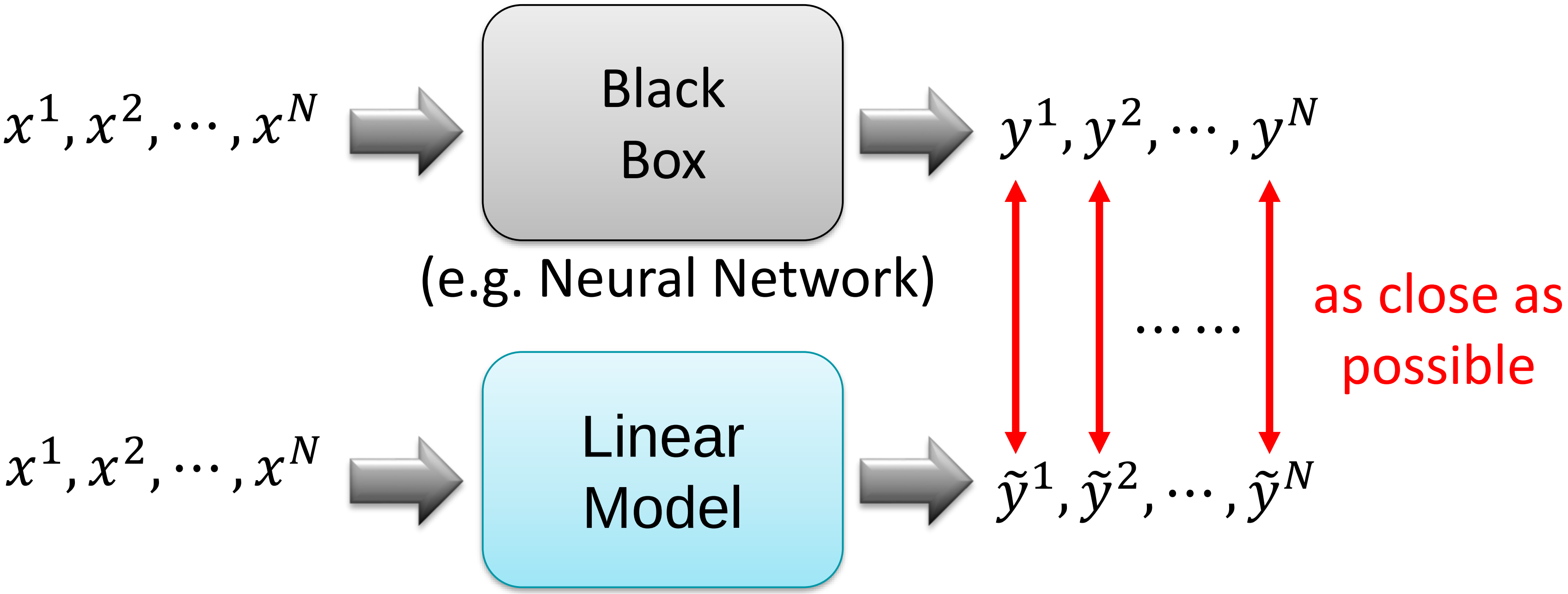
Sanity check:
When prediction changes, do explanations change?



(Slide Credit – Julius Adebayo)

(Image Credit – Hung-yi Lee)

LIME – locally interpretable model agnostic



Can't do it globally of course, but locally ? Main Idea behind LIME

Intuition behind LIME

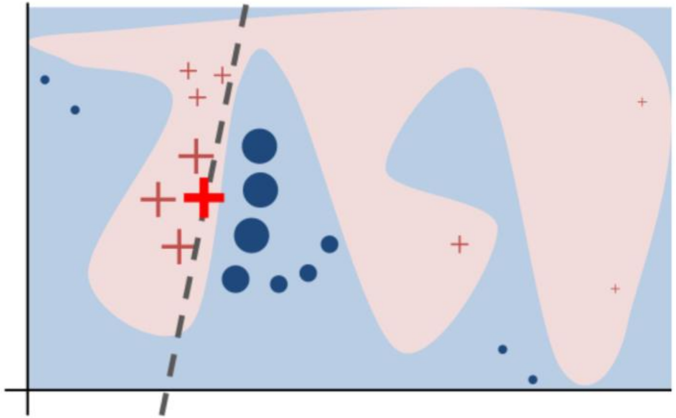
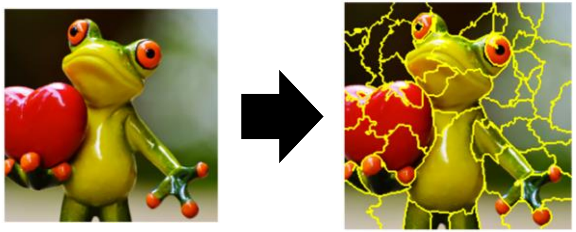


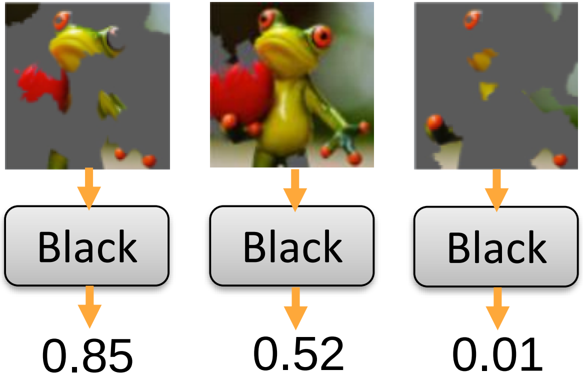
Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

[Ribeiro et al 2016]

LIME – Image



- 1. Given a data point you want to explain
- 2. Sample at the nearby - Each image is represented as a set of superpixels (segments).



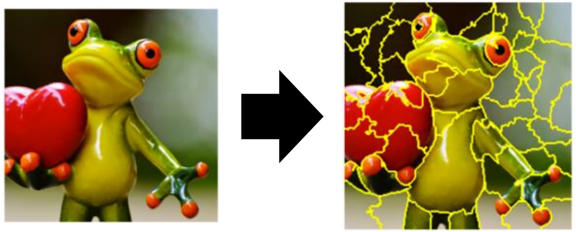
Randomly delete some segments.

Compute the probability of “frog” by black box

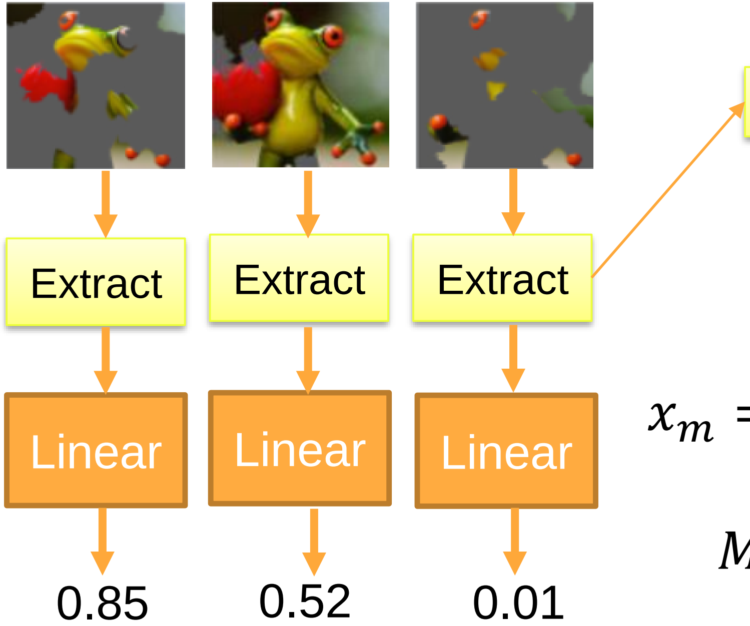
Ref: https://medium.com/@kstseng/lime-local-interpretable-model-agnostic-explanation%E6%8A%80%E8%A1%93%E4%BB%8B%E7%B4%B9-a67b6c34c3f8

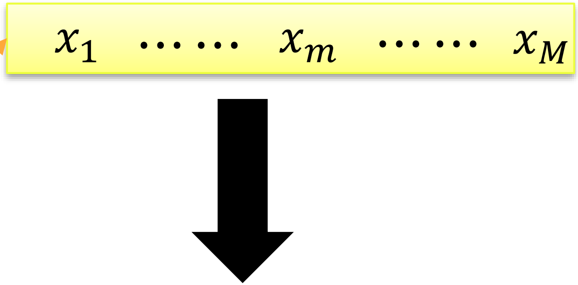
(Slide Credit – Hung-yi Lee)

LIME – Image



- 3. Fit with linear (or interpretable) model



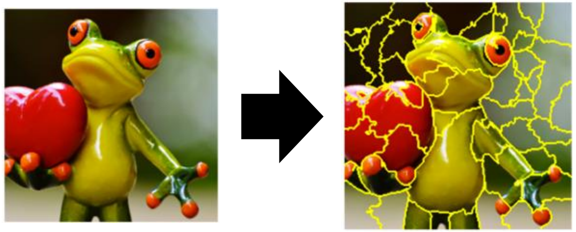


$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

(Slide Credit – Hung-yi Lee)

LIME – Image



- 4. Interpret the model you learned

Extract

Linear

0.85

$$y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M$$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

If $w_m \approx 0$ → segment m is not related to "frog"

If w_m is positive → segment m indicates the image is "frog"

If w_m is negative → segment m indicates the image is not "frog"

(Slide Credit – Hung-yi Lee)

The Math behind LIME

Algorithm 1 Sparse Linear Explanations using LIME

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

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for

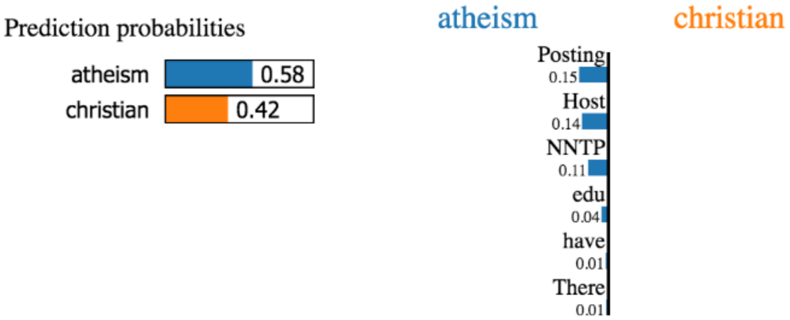
Match interpretable model to black box

Control complexity of the model

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

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

Example from NLP



Text with highlighted words

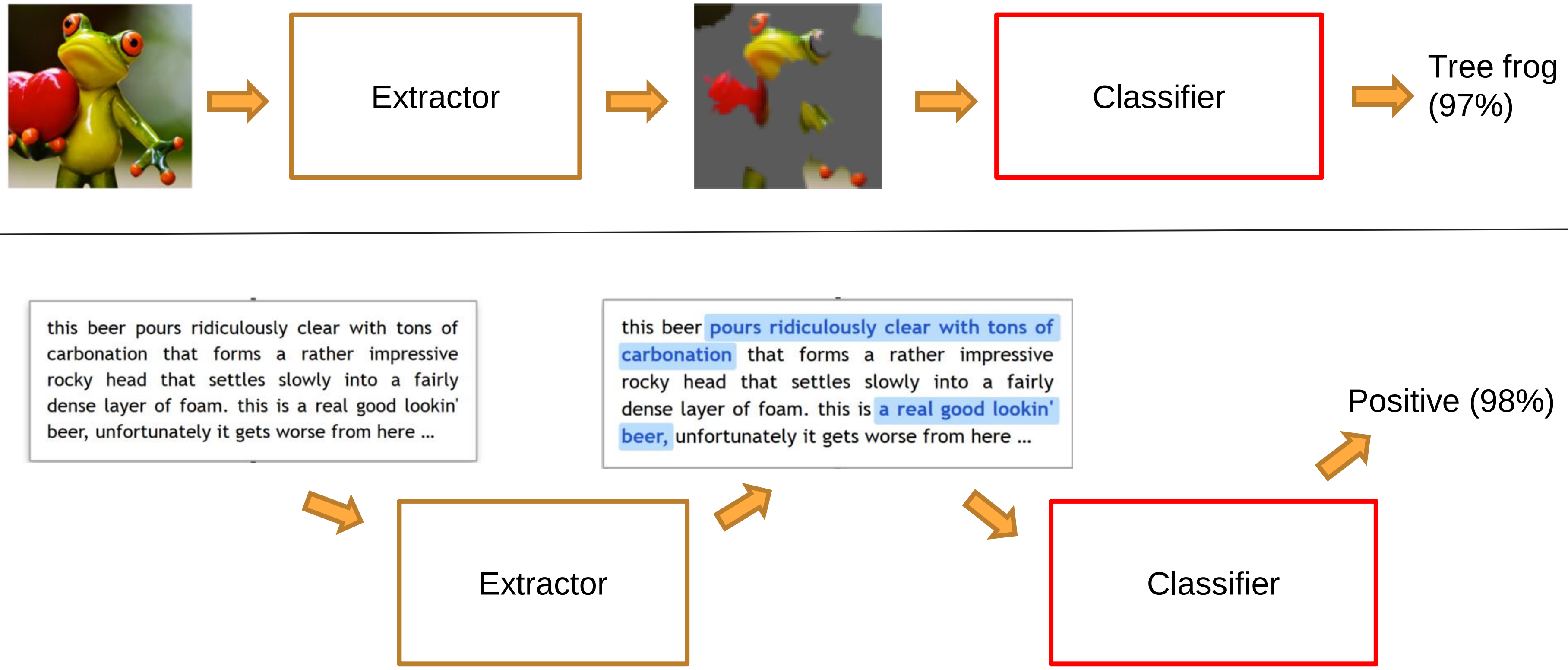
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

Rationalization Models

General Idea



Massachusetts
Institute of
Technology

CSAIL

Rationalizing Neural Predictions

Tao Lei

Regina Barzilay   Tommi Jaakkola

EMNLP 2016

(Slides Credit – Tao Lei)

Model Architecture

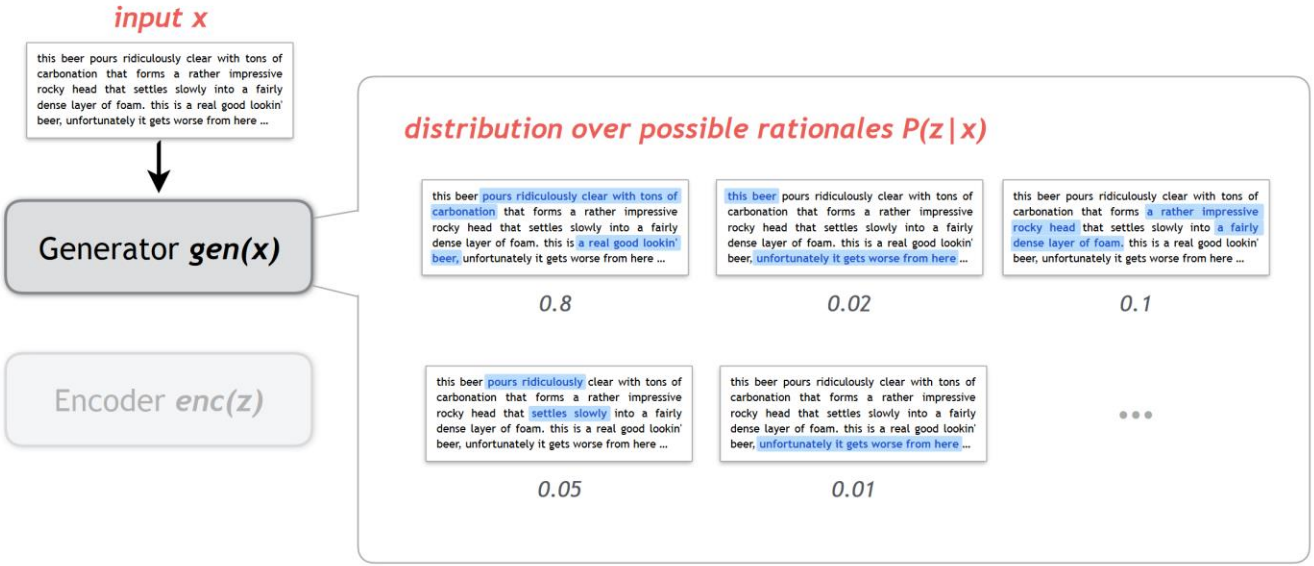
Generator ***gen(x)***

Encoder ***enc(z)***

two modular components ***gen()*** and ***enc()***

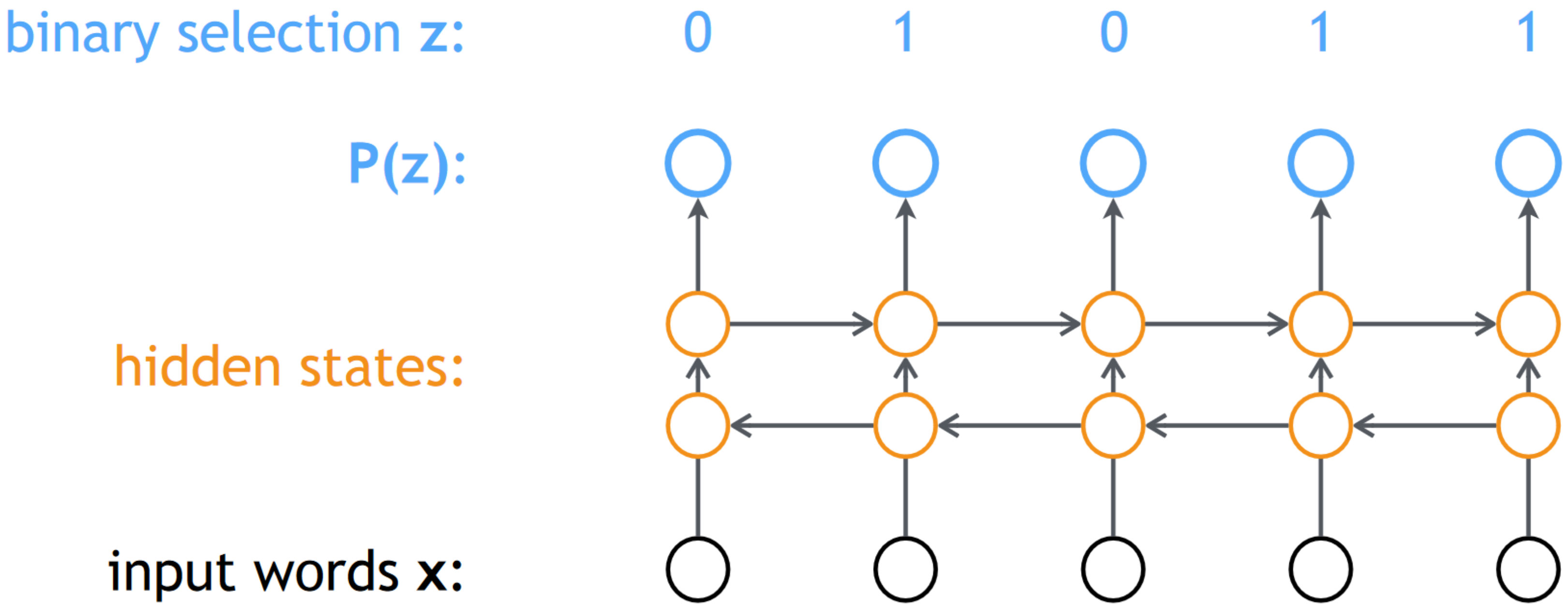
(Slides Credit – Tao Lei)

Model Architecture



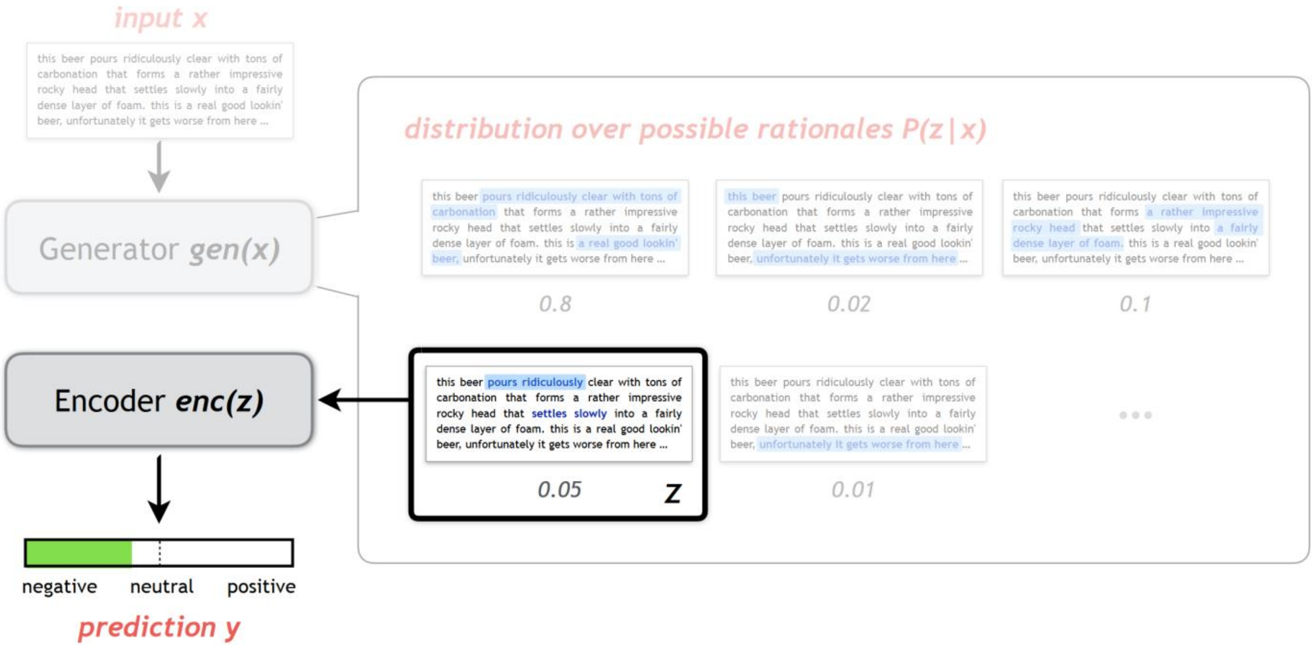
generator specifies the distribution of rationales

(Slides Credit – Tao Lei)



(Slides Credit – Tao Lei)

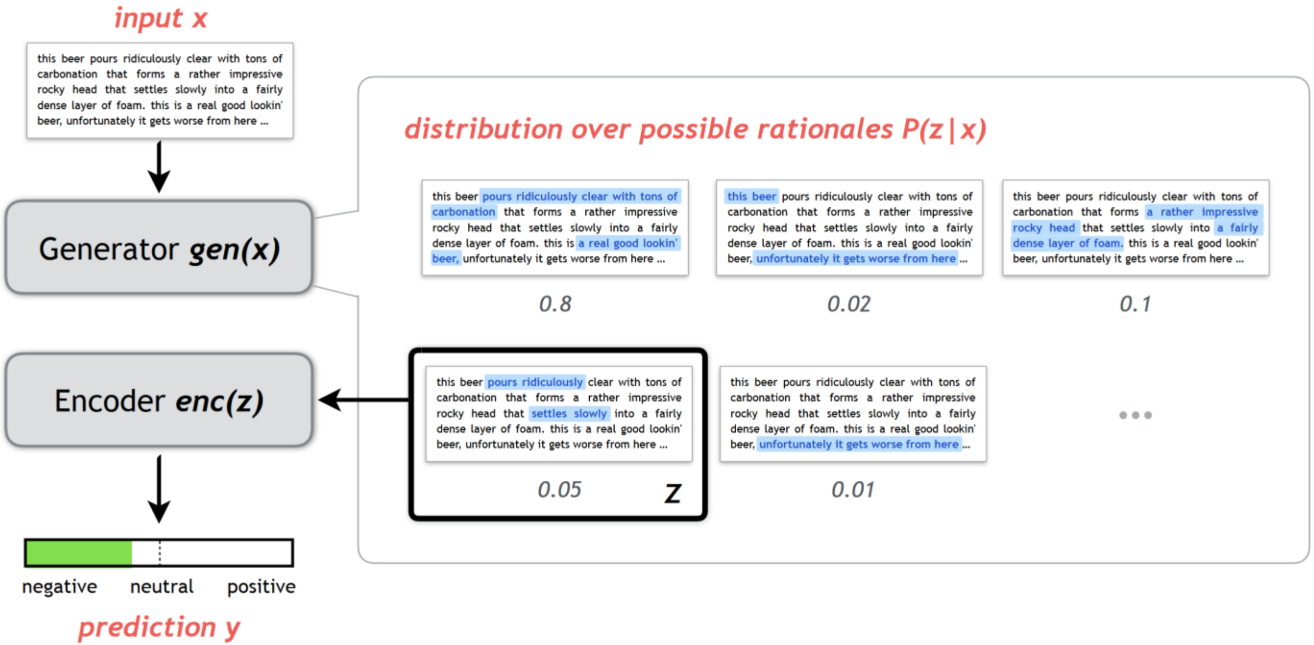
Model Architecture



encoder makes prediction given rationale

(Slides Credit – Tao Lei)

Model Architecture



two components optimized jointly

(Slides Credit – Tao Lei)

Training Objective

$$\mathrm{cost}(\mathbf{z},\mathbf{y}) = \boxed{\mathrm{loss}(\mathbf{z},\mathbf{y})} + \boxed{\lambda_1 |\mathbf{z}|_1} + \boxed{\lambda_2 \sum_i |\mathbf{z}_i - \mathbf{z}_{i-1}|}$$

sufficiency
correct prediction

sparsity
rationale is short

coherency
continuous selection

- receive this training signal after z is produced

Minimizing expected cost:

$$\min_\theta \sum_{(\mathbf{x},\mathbf{y})\in D} \mathbb{E}_{\mathbf{z}\sim\mathrm{gen}(\mathbf{x})}\left[\mathrm{cost}(\mathbf{z},\mathbf{y})\right]$$

- intractable because summation over z is exponential

(Slides Credit – Tao Lei)

Learning Method

- Possible to sample the gradient, e.g.:

$$\mathbb{E}_{\mathbf{z}\sim \text{gen}(\mathbf{x})}\left[\text{cost}(\mathbf{z},\mathbf{y})\ \frac{\partial \log P(\mathbf{z}|\mathbf{x})}{\partial \theta_g}\right]$$

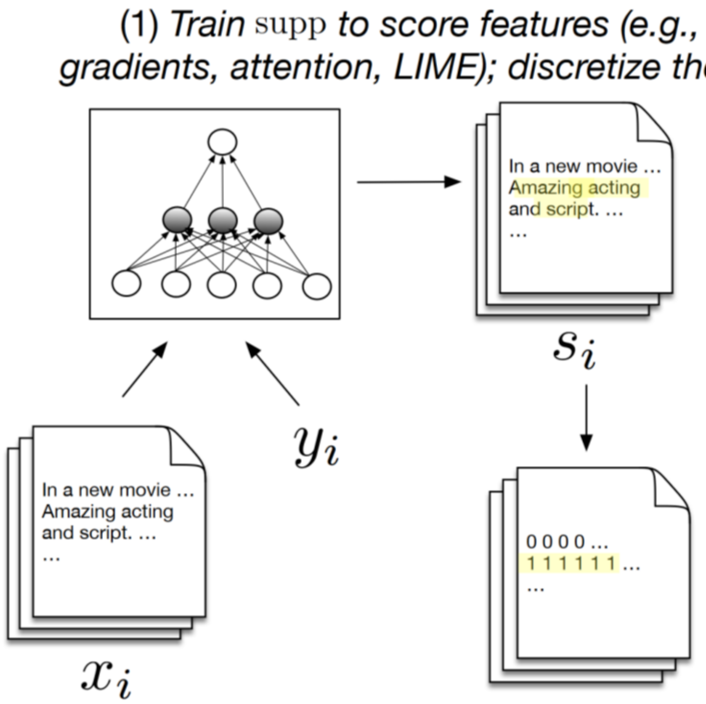
$$\approx \ \frac{1}{N}\sum_{i=1}^{N}\text{cost}(\mathbf{z}_i,\mathbf{y}_i)\frac{\partial \log P(\mathbf{z}_i|\mathbf{x}_i)}{\partial \theta_g}$$

where z_i are sampled rationales

- Stochastic gradient decent on sampled gradients

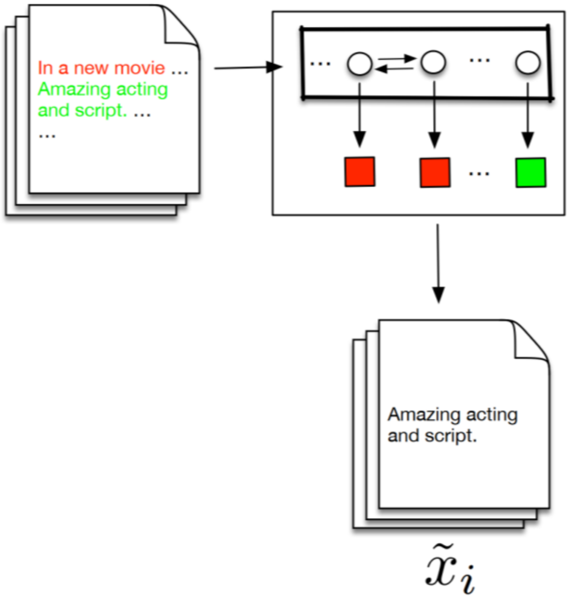
(Slides Credit – Tao Lei)

FRESH Model – Faithful Rationale Extraction using Saliency Thresholding



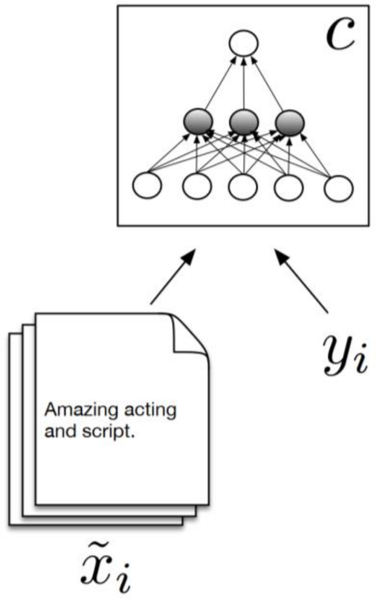
FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(2) *Train* ext *to extract snippets; use to create* $\tilde{x}_i$



FRESH Model – Faithful Rationale Extraction using Saliency Thresholding

(3) *Train* pred *on* $(\tilde{x}_i, y_i)$



Some Results – Functional Evaluation

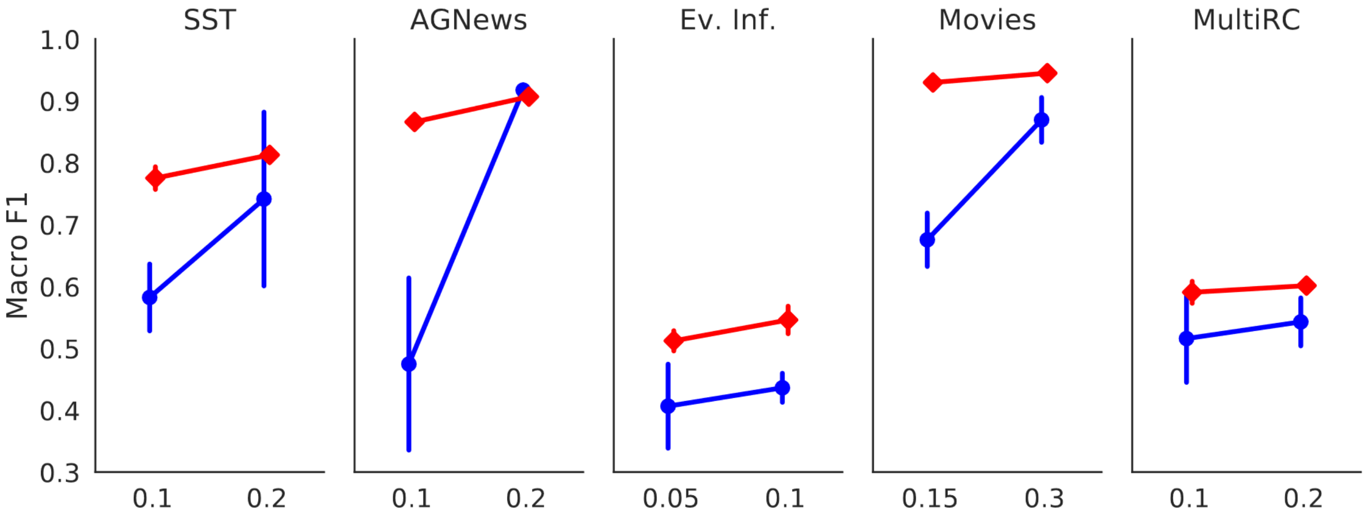


Figure 2: Results for Lei *et al.* (●) and FRESH (♦) evaluated across five datasets at two different desired raionale lengths (as % of document length). Vertical bars depict standard deviations observed over five random seeds.

Some Results – Human Evaluation

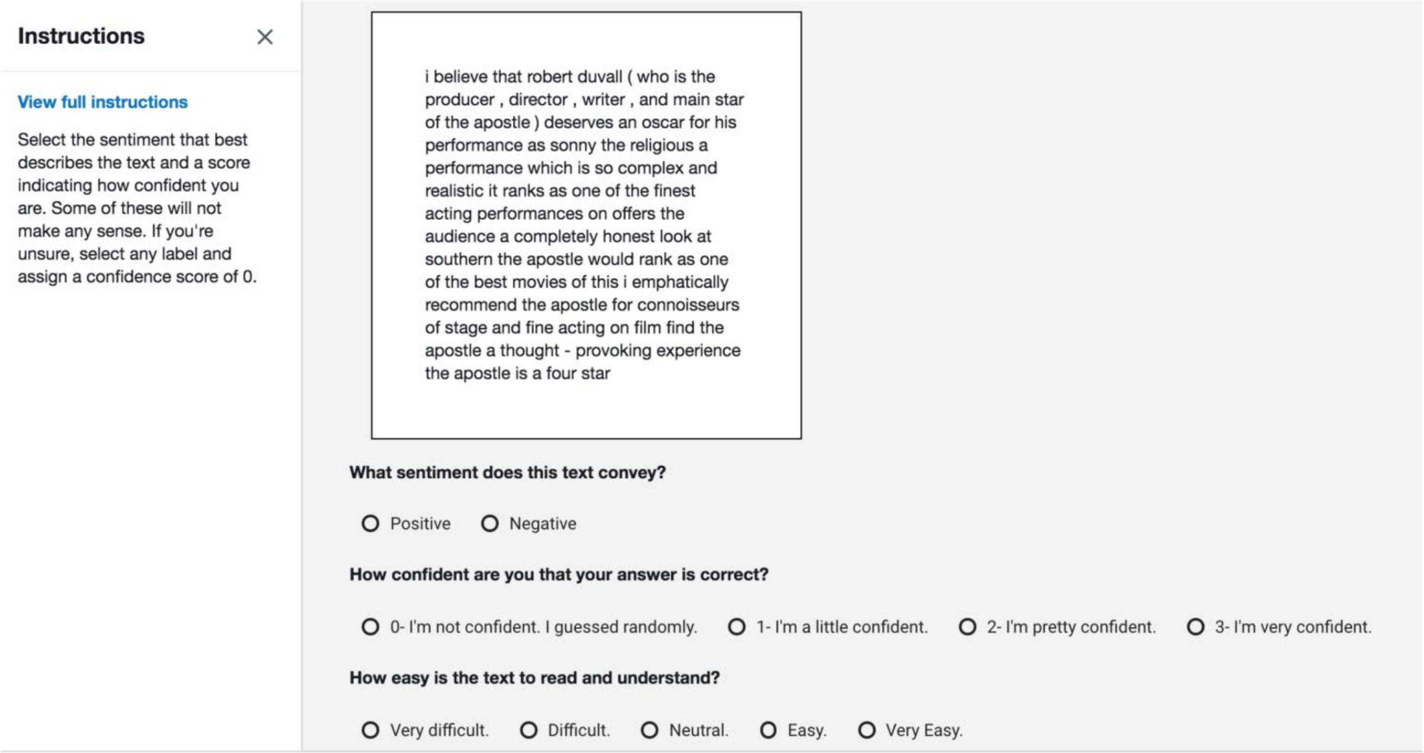


Figure 8: Amazon Mechanical Turk layout for Movies tasks.

Some Results – Human Evaluation

Rationale Source	Human Acc.	Confidence (1–4)	Readability (1–5)
Human	.99	3.44 ±0.53	3.82 ±0.56
Random			
Contiguous	.84	3.18 ±0.55	3.80 ±0.57
Non-Contiguous	.65	2.09 ±0.51	2.07 ±0.69
Lei et al. 2016			
Contiguous	.88	3.39 ±0.48	4.17 ±0.59
Non-Contiguous	.84	2.97 ±0.72	2.90 ±0.88
Our Best			
Contiguous	.92	3.31 ±0.48	3.88 ±0.57
Non-Contiguous	.87	3.23 ±0.47	3.63 ±0.59

Important Points to take away

- Interpretability – no consistent definition
- When designing new system, ask your stakeholders what they want out of it .
- See if you can use inherently interpretable model .
- If not, what method can you use to interpret the black box ?
- Ask – does this method make sense ? Question Assumptions !!!
- Stress Test and Evaluate !