SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

2021-22 Term 1

IS457: Fairness in Socio-technical Systems

Week 10 - Interpretability of algorithmic systems

KWAK Haewoon

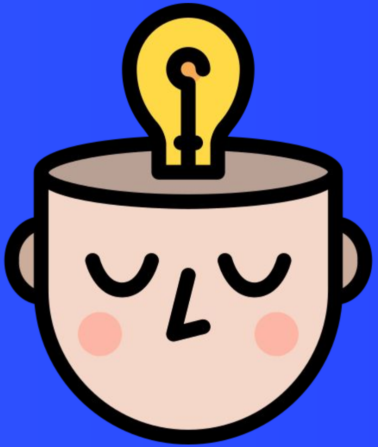


Image from Flaticon.com

Study questions

What is interpretability?

What is the “right to explanation”?

What are the three levels of transparency?

What are post-hoc explanations?

What types of complexity affect the interpretability (simulatability)?

Interpretability

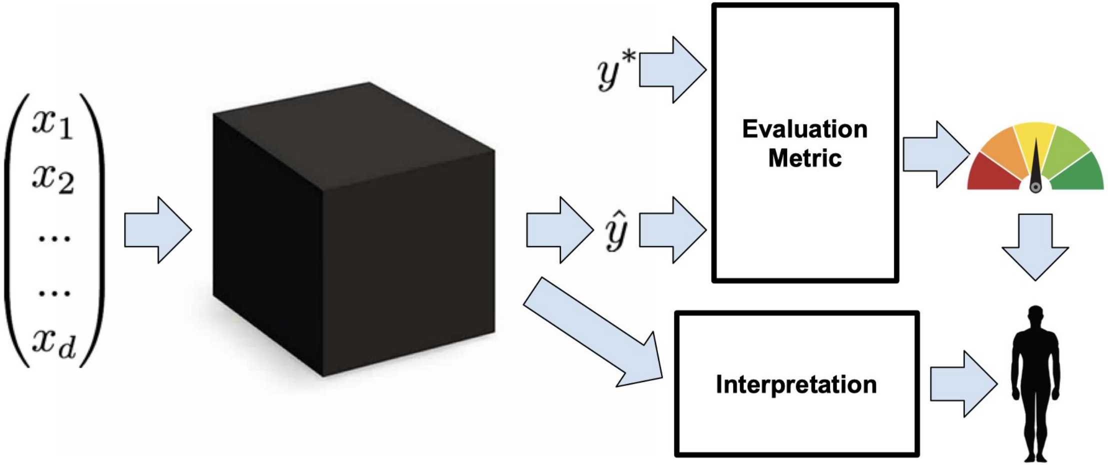
Interpretability is the degree to which a human can understand the cause of a decision.

Interpretability is the degree to which a human can consistently predict the model's result.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.
Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability." Advances in neural information processing systems 29 (2016).

Interpretability as an additional dimension

Predictions and evaluation metrics (e.g., accuracy) do not suffice to characterize the model in terms of interpretation.



Interpretability in linear regression model

Consider a linear regression model to predict # of rented bikes on a particular day.

"An increase of the temperature by 1 degree Celsius increases the predicted number of bicycles by 110.7, when all other features remain fixed."

	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

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

Racial discrimination in the (opaque) algorithm

In Week 3

The visa algorithm discriminated on the basis of nationality.

Applications made by people holding ‘suspect’ nationalities received a higher risk score.

Their applications received intensive scrutiny by Home Office officials, were approached with more scepticism, took longer to determine, and were much more likely to be refused.

HireVue's opaque algorithms

In Week 3

Job candidates don't know their final scores, what they got wrong, and what they could do better because:

- The algorithm is protected as trade secrets
- Even HireVue doesn't always know how the system decides on who gets high scores.

Instead, HireVue has given only vague explanations.

- E.g., for a call center job, "supportive" words might be encouraged.

Interpretability as a solution

As machine learning models penetrate critical areas, such as medicine, the criminal justice system, and employment, the inability of humans to understand these models seems problematic.

Interpretability has been proposed as a remedy.

Interpretability for trust

If users do not trust a machine learning model or a prediction, they will not use it.

Two different definitions of trust in machine learning:

- Trusting a prediction: whether a user trusts an individual prediction sufficiently to take some action based on it
- Trusting a model: whether a user trusts a model to behave in reasonable ways if deployed

Both trusts are directly impacted by the understanding of model's behavior.

Interpretability for better transferability

Users need to be confident that the model will perform well on real-world data.

Currently, models are evaluated using accuracy metrics on testing datasets, but real-world data is often significantly different.

Even worse, a deployment environment might be actively adversarial.

Adversarial examples

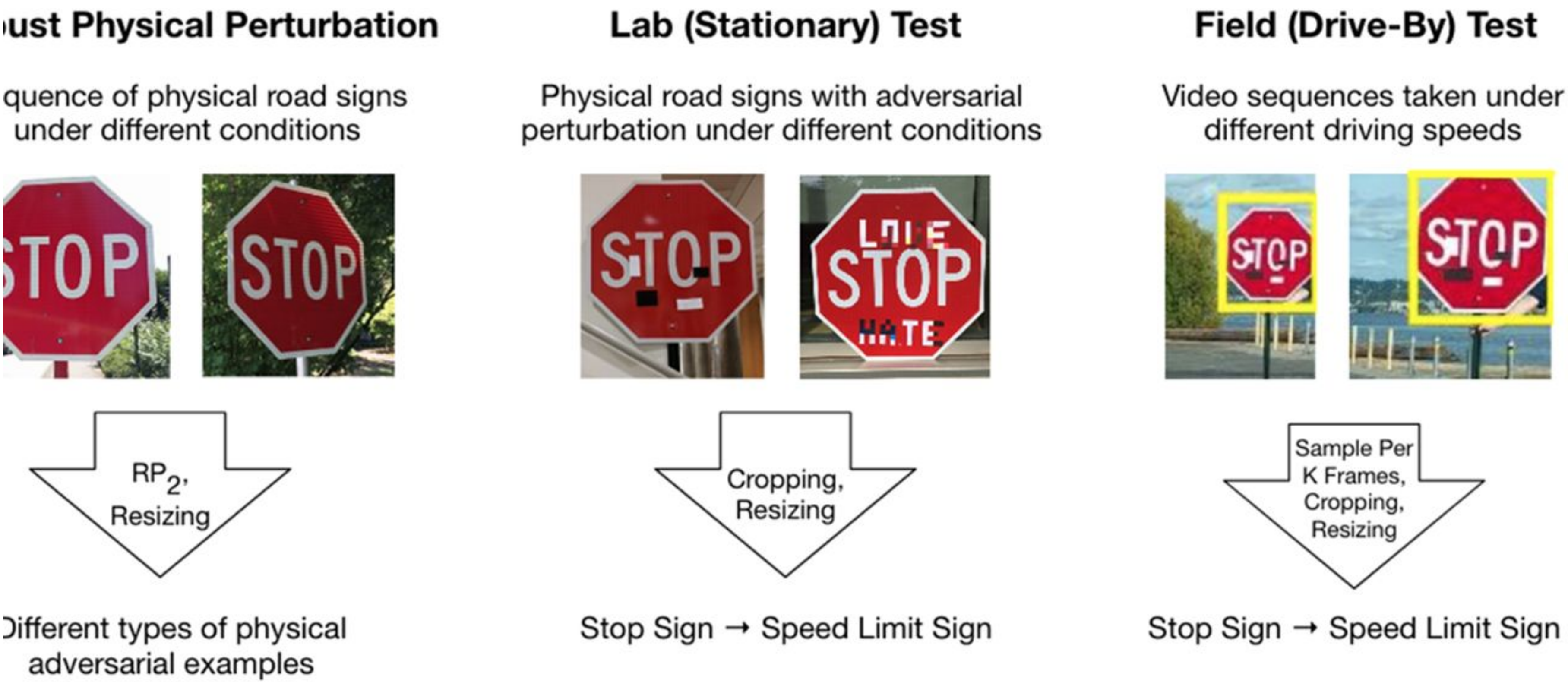


Image from https://deepdrive.berkeley.edu/node/212

Interpretability for fair and ethical outcomes

Algorithmic decision-making more and more influences our social experiences.

We must provide interpretations for assessing whether these “decisions” conform to ethical standards.

Likewise, fairness in these decisions lead to demands for interpretable models.

Right to explanation and three barriers

EU regulations on algorithmic decision-making: “A user can ask for an explanation of an algorithmic decision that significantly affects them.”

There are three barriers:

1. Intentional concealment on the part of corporations or other institutions
2. Gaps in technical literacy, which mean that, for most people, simply having access to underlying code is insufficient
3. A “mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation.”

Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a “right to explanation”." AI magazine 38.3 (2017): 50-57.

GDPR provides some foundations

Article 13: Information to be made available or given to the data subject goes some way

1. Intentional concealment on the part of corporations or other institutions
2. Gaps in technical literacy, which mean that, for most people, simply having access to underlying code is insufficient
3. A "mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation."

Article 12: Communication and modalities for exercising the rights of the data subject attempts to solve the second by requiring that communication with data subjects is in "concise, intelligible and easily accessible form."

Properties for interpretations

Two categories of techniques and model properties to enable interpretations

- Transparency
- Post-hoc explanations

Transparency

Transparency ↔ Blackbox-ness or opacity

Three levels of transparency

- Entire model level (simulatability)
- Individual component level (decomposability)
- Training algorithm level (algorithmic transparency)

Entire model level: Simulatability

We might call a model transparent if a person can contemplate the entire model at once.

"Simple" model: A human should be able to take the input data together with the parameters of the model and go through every calculation required to produce a prediction in <u>reasonable</u> time step.

Individual component level: Decomposability

Each part of the model - input, parameter, and calculation - admits an intuitive explanation.

Inputs themselves are need to be individually interpretable, disqualifying some models with highly engineered or anonymous features.

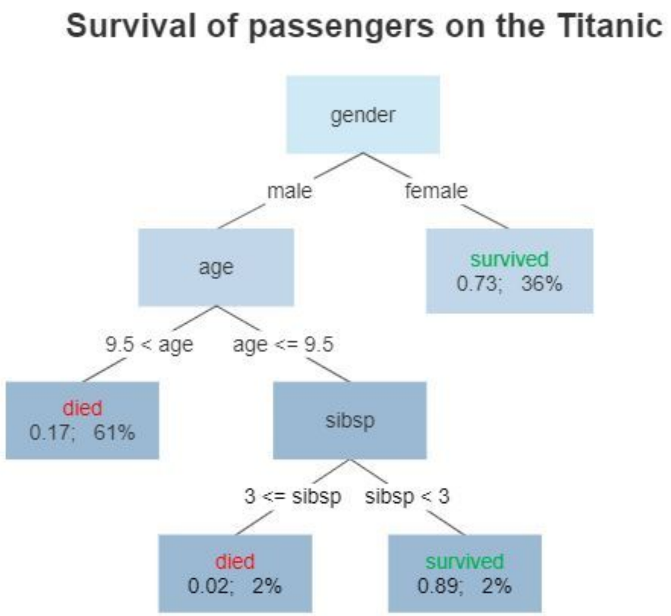


Image from https://en.wikipedia.org/wiki/Decision_tree_learning

Training algorithm level: Algorithmic transparency

Transparency can be provided at the level of the learning algorithm itself.

- Prove that training will converge to a unique solution, even for unseen data.
- Give confidence how the model behaves in the real-world setting

Note: Modern deep learning methods lack this algorithmic transparency (We cannot guarantee that they will work on new problems.)

Linear models vs. deep neural networks

Linear models are more interpretable than DNNs in terms of algorithmic transparency.

However, given high dimensional or heavily engineered features, linear models lose simulatability or decomposability, respectively.

Post-hoc explanations

Post-hoc interpretability presents an approach to extracting information from learned models.

While post-hoc interpretations often do not elucidate precisely how a model works, they confer useful information for practitioners and end users.

Strong advantage: we can interpret opaque models after-the-fact, without sacrificing predictive performance.

Common approaches to post-hoc interpretability:

- Text explanations
- Visualizations of learned representations or models
- Explanations by example

Text explanations

Train one model to generate predictions and a separate model to generate an explanation.

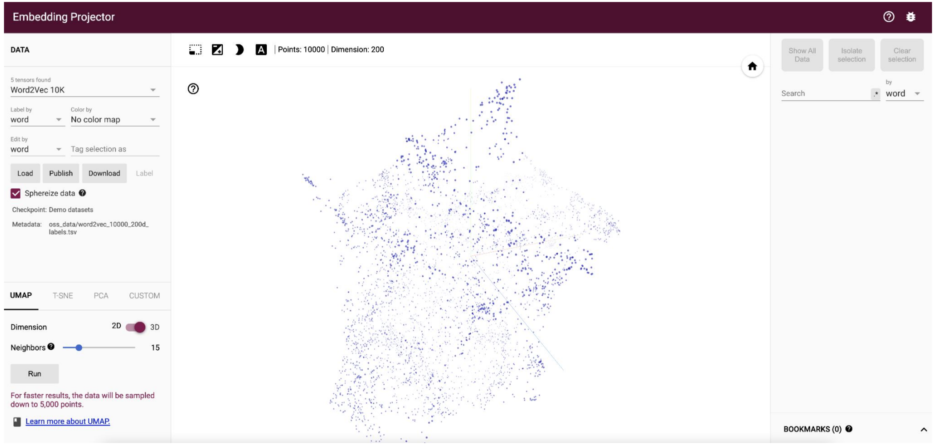
These explanations are trained to maximize the likelihood of previously observed ground truth explanations from human players, and may not faithfully describe the agent's decisions, however plausible they appear.

Visualization

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

Render visualizations with the aim of determining qualitatively what a model has learned.

E.g., t-SNE visualization of learned representations



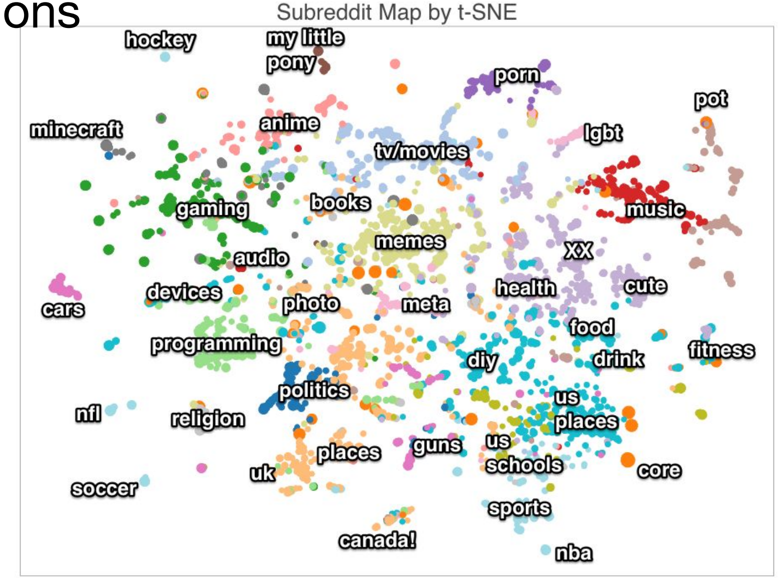


Image from http://nicolas.kruchten.com/content/2014/12/subreddit-map/

Explanations by example

Report other examples that the model considers to be most similar.

Humans also sometimes justify actions by analogy (explanations by example).

- Doctors often refer to case studies to support a planned treatment protocol.

Example of Post-hoc explanations: LIME

Local Interpretable Model-Agnostic Explanations (LIME)

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Can we trust these predictions?

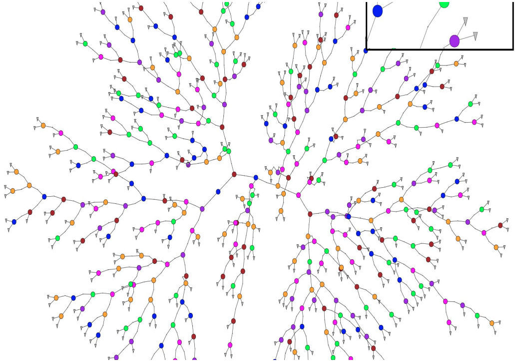
5 out of 6 predictions are correct, and only 1 prediction is incorrect.

Three must-haves for a good explanation

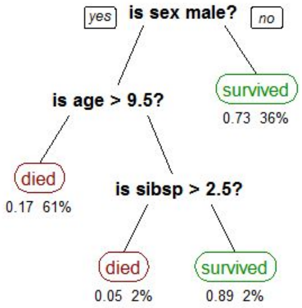
SMU Singapore Management University | School of Computing and Information Systems

Interpretable

- Humans can easily interpret reasoning



Definitely
not interpretable



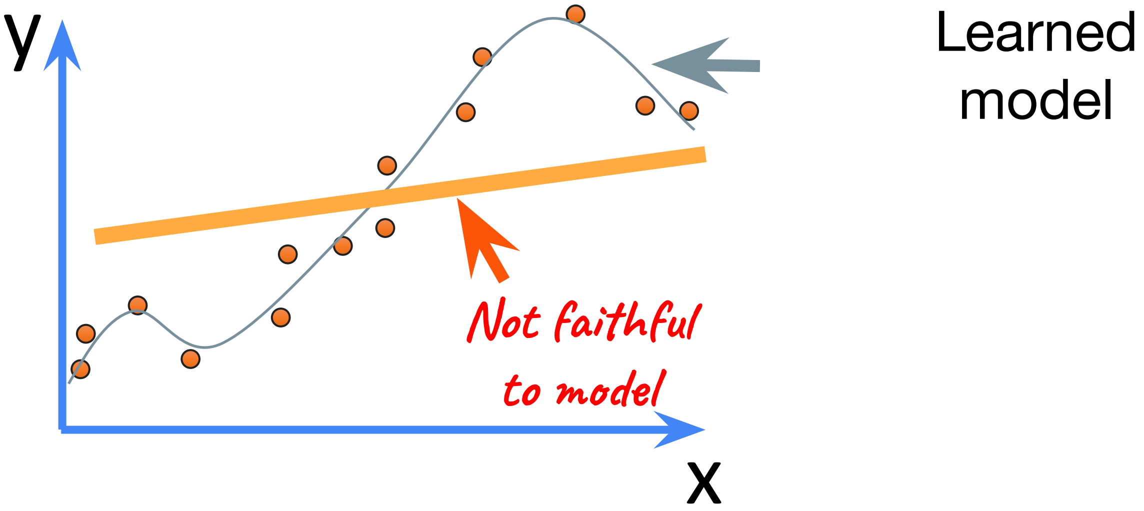
Potentially
interpretable

From authors' slides

Three must-haves for a good explanation

SMU Singapore Management University | School of Computing and Information Systems

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves

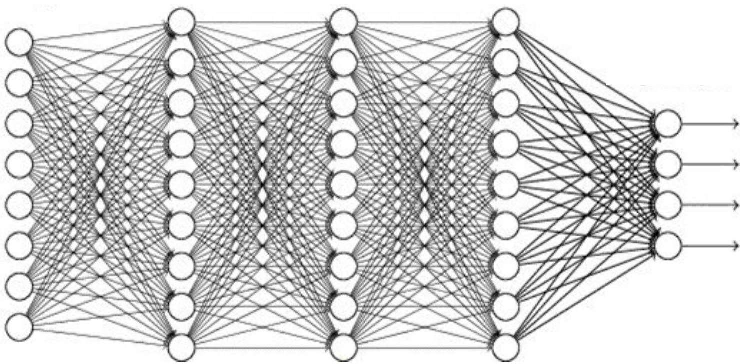


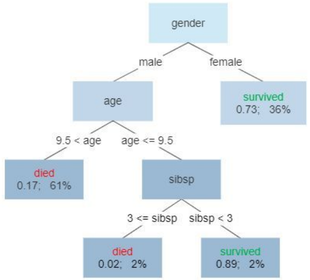
From authors' slides

Three must-haves for a good explanation

SMU SINGAPORE MANAGEMENT UNIVERSITY
School of Computing and Information Systems

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves
Model agnostic	• Can be used for *any* ML model





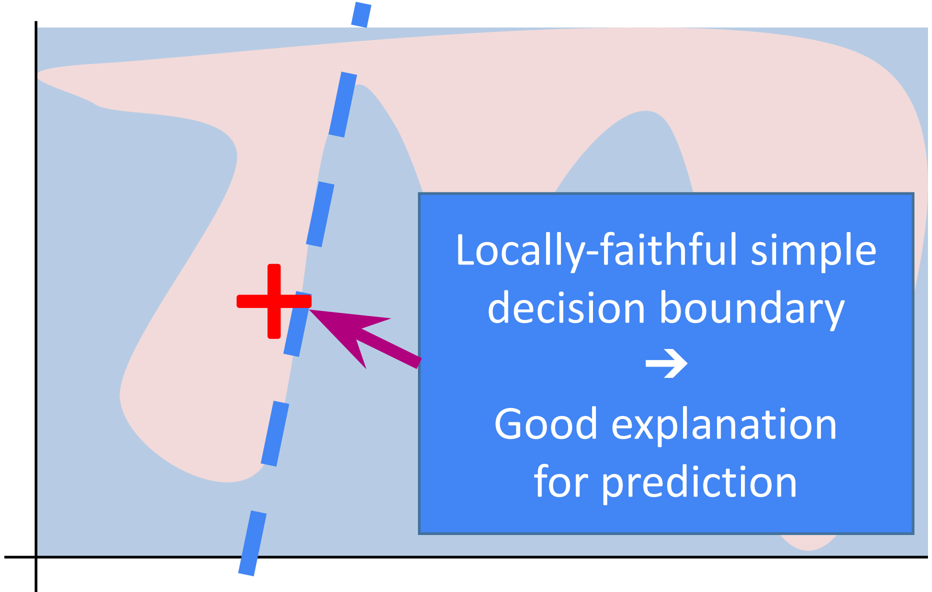
From authors' slides

Key ideas behind LIME

1. Pick a model class interpretable by humans
   - Not globally faithful… ☹️

2. Locally approximate global (blackbox) model
   - Simple model globally bad, but locally good

Line,
shallow decision tree,
sparse features, …

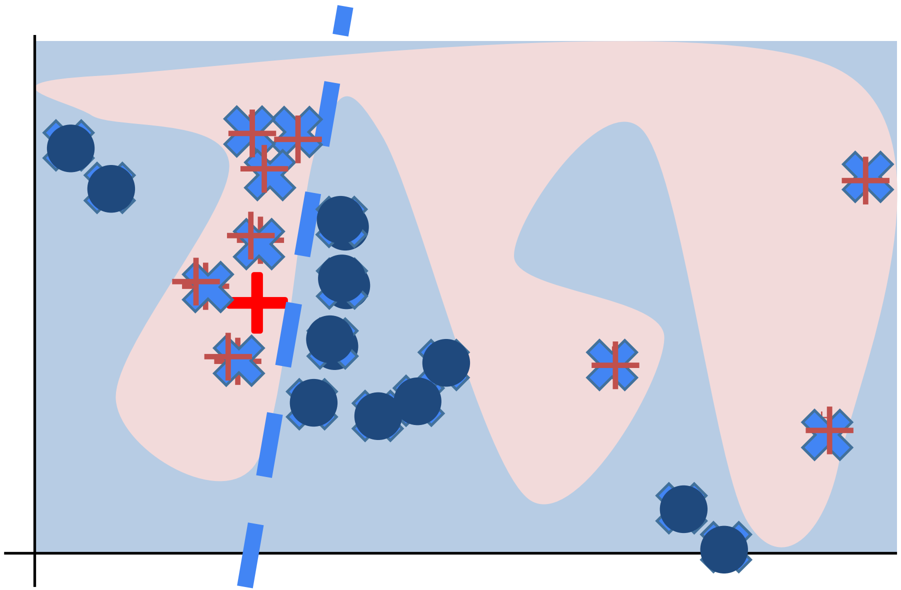


From authors' slides

Using LIME to explain a complex model

SMU Singapore Management University | School of Computing and Information Systems

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

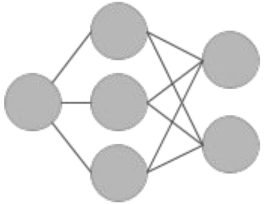


From authors' slides

Interpretable representation: images

School of
Computing and
Information Systems

x (3 color channels / pixel)



Model

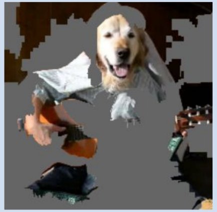
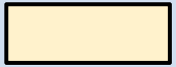
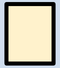
x' (contiguous superpixels)

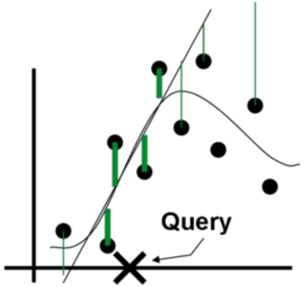
Human

From authors’ slides

Sampling example - images

Original Image
P(labrador) = 0.21

Perturbed Instances	P(Labrador)
	 0.92
	0.001
	 0.34



Locally weighted regression

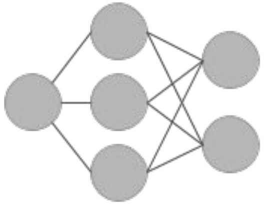
Explanation

From authors' slides

Interpretable representations

x (embeddings)

0.5	0.3	1.3	4.4	1.1	…



Model

x' (words)

This is a horrible movie.

Human-readable text is what to perturb and what to use in the interpretable approximation

From authors' slides

lime/lime_text.py

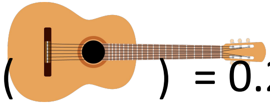

```python
    def __data_labels_distances(self,
                                indexed_string,
                                classifier_fn,
                                num_samples,
                                distance_metric='cosine'):
        """Generates a neighborhood around a prediction.

        Generates neighborhood data by randomly removing words from
        the instance, and predicting with the classifier. Uses cosine distance
        to compute distances between original and perturbed instances.
```

https://github.com/marcotcr/lime/blob/master/lime/lime_text.py

Explaining Google's Inception NN

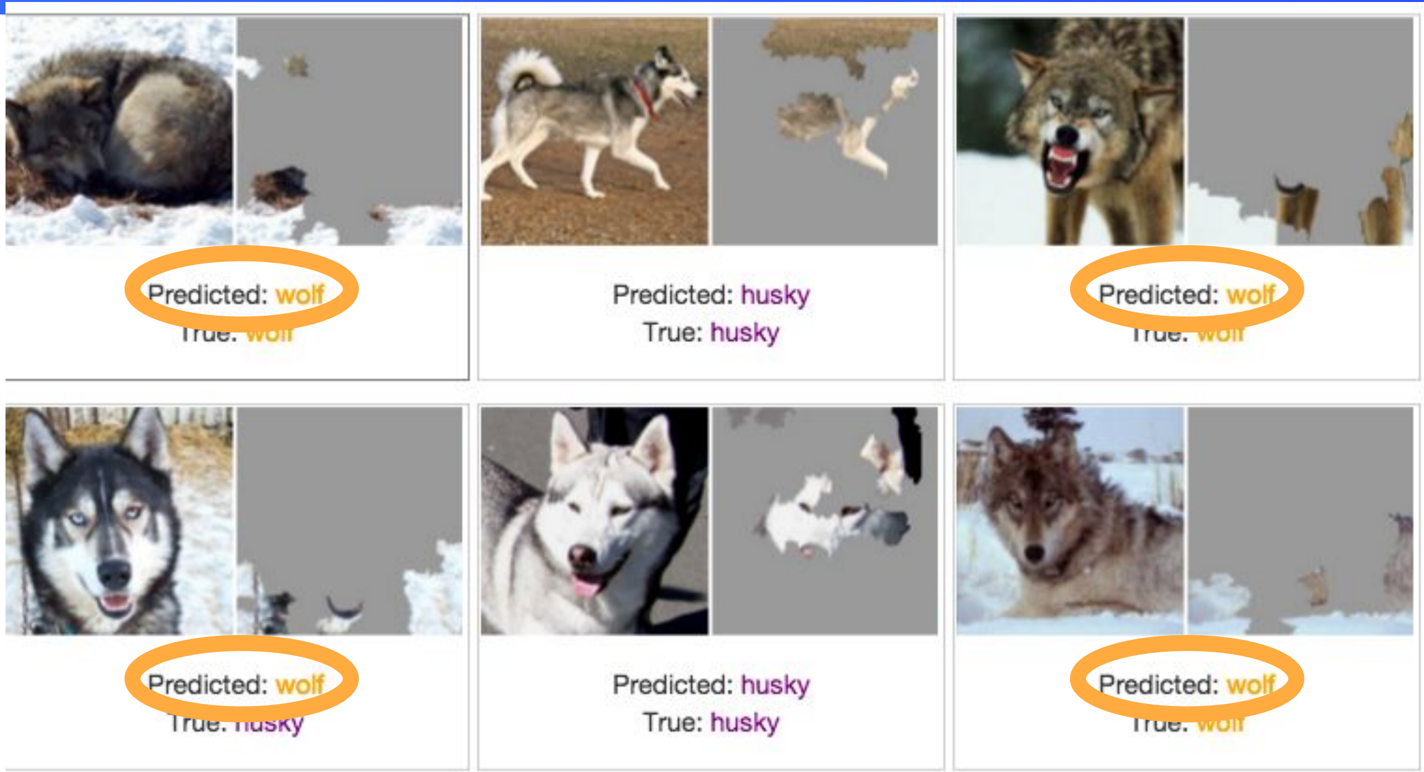
P() = 0.32

P() = 0.24

P() = 0.21

From authors' slides

Can we trust these predictions? :)



It's actually a SNOW detector!

Lab

Visit https://github.com/haewoon/lab-interpretable-machine-learning

Click Open in Colab

Check SHAP https://github.com/slundberg/shap if you are interested in more.

Transparency vs. Post-hoc explanations

Requiring model transparency would create an important change to ML as it is being done today—essentially that we forgo deep learning altogether and whatever benefits it may entail.

Post-hoc explanations for the black-box outputs become widely-used these days but less accurate. Thus, transparent model could be preferred in some cases.

Babic, Boris, et al. "Beware explanations from AI in health care." Science 373.6552 (2021): 284-286.

Counterfactual explanations

Example: “You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan.”

Typical explanation typically refers to an attempt to convey the internal state or logic of an algorithm that leads to a decision. In contrast, counterfactuals describe a dependency on the external facts that led to that decision.

Advantages of counterfactual explanations

Counterfactuals bypass the challenge of explaining the internal workings of complex machine learning systems.

Counterfactuals provide information that is both easily digestible and practically useful for understanding the reasons for a decision, challenging them, and altering future behaviour for a better result.

The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

Human Evaluation of Models Built for Interpretability

Isaac Lage,*[1] Emily Chen,*[1] Jeffrey He,*[1] Menaka Narayanan,*[1]
Been Kim,[2] Samuel J. Gershman,[1] Finale Doshi-Velez[1]
[1]Harvard University, [2]Google
isaaclage@g.harvard.edu, {emily-chen, jdhe, menakanarayanan}@college.harvard.edu
beenkim@google.com, gershman@fas.harvard.edu, finale@seas.harvard.edu

Model complexity and simulatability

School of Computing and Information Systems

The challenge is determining which types of complexity affect human-simulatability and by how much, since this will guide the choice of regularizers for interpretability.

Lage, Isaac, et al. "Human evaluation of models built for interpretability." Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. Vol. 7. No. 1. 2019.

Simulation task - Decision sets

Logic-based models

Each line contains a clause in disjunctive normal form (an or-of-ands) of the inputs (blue words), which, if true, maps to the output (orange words–also in disjunctive normal form).

The alien's preferences:

frowning or raining and **puffy eyes** and **chest pain** → **laxatives or vitamins** and **antibiotics**
sweating and **frowning** and **raining or anxious** → **laxatives and antibiotics** or **stimulants**
hoarse and blurry vision and frowning or **sweating** → **painkillers and antibiotics** or **vitamins**
squinting or **chest pain and raining and sweating** → **antibiotics or tranquilizers** and **painkillers**
puffy eyes and **hoarse** and **blurry vision or anxious** → **vitamins** and **antibiotics or tranquilizers**
hives and squinting and raining or **frowning** → **tranquilizers** or **painkillers and antibiotics**

Observations: hoarse, blurry vision, puffy eyes

Disease Medications:

- **antibiotics:** Aerove, Adenon, Athoxin
- **painkillers:** Poxin, Parola, Pelapin
- **vitamins:** Vipryl, Vyorix, Votasol
- **stimulants:** Silvax, Setoxin, Soderal
- **tranquilizers:** Trasmin, Tydesol, Texopal
- **laxatives:** Lantone, Lezanto, Lexerol

What prescription would you recommend to treat the alien's symptoms?

- [ ] Vitamins
- [ ] Antibiotics
- [ ] Laxatives
- [ ] Tranquilizers
- [ ] Stimulants
- [ ] Painkillers

Submit Answer

Complexity dimension

Model size:

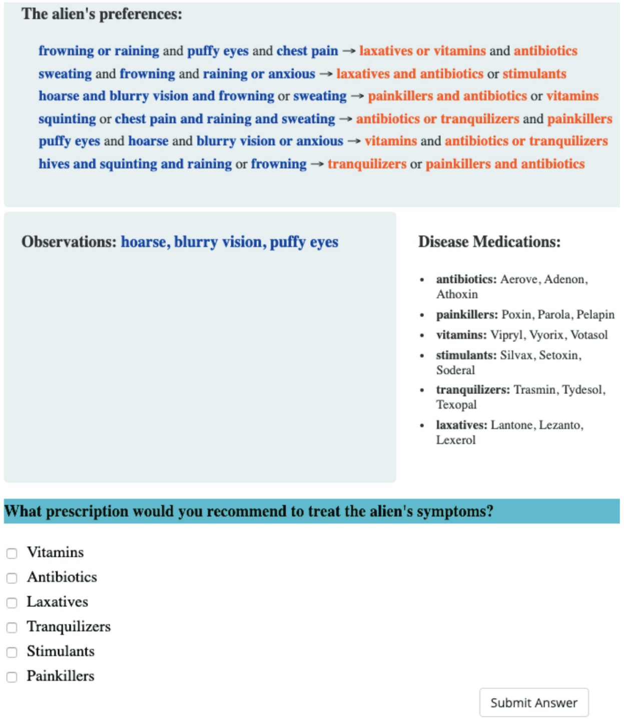
- Total number of lines in the decision sets (2,5,10)
- Number of terms within the output clause (2,5)

Cognitive chunks

- Number of clauses in disjunctive normal (1,3,5)

Repeated terms

- Number of variable repetitions (2,3,4,5)



Evaluation metrics

Response time (seconds)

Subjective difficulty of use (5-point Likert scale, 1-very easy .. 5-very difficult)

Accuracy

Results

Greater complexity results in longer response times, with the most marked effects for cognitive chunks, followed by model size, then number of variable repetitions.

Subjective difficulty of use largely replicates the findings of response time.

The effect of different types of complexity on accuracy was less clear.

Reflection

School of Computing and Information Systems

https://smu.sg/IS457r10