University of Colorado **Boulder**

# CSCI 5622: Machine Learning

## Lecture 9

# Explainable AI

- Basics on explainability and interpretability
- Visualizing CNN activations
- Prediction difference analysis
- Local Interpretable Model-agnostic Explanations (LIME)

# Explainable AI

- Basics on explainability and interpretability
- Visualizing CNN activations
- Prediction difference analysis
- Local Interpretable Model-agnostic Explanations (LIME)

# Basics on explainability and interpretability

## Deep neural networks are easily fooled! Chihuahua or muffin?

# Basics on explainability and interpretability

## Chihuahua or muffin? Image captioning

test13.png

| | |
|---|---|
| **msft_captions** | a brown and white teddy bear (69) |
| **msft_tags** | |

test11.png

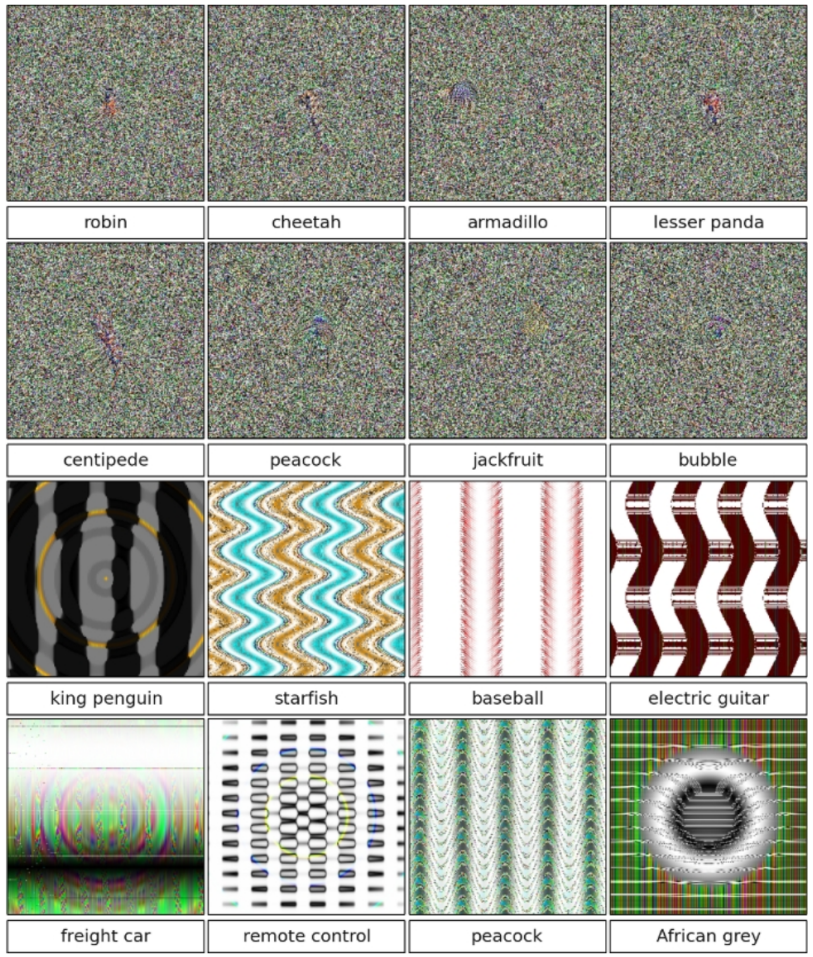
| | |
|---|---|
| **msft_captions** | a close up of a stuffed animal (55) |
| **msft_tags** | indoor (94) , food (87) , bread (74) , dessert (30) |

test1.png

| | |
|---|---|
| **msft_captions** | a close up of a stuffed animal (65) |
| **msft_tags** | indoor (88) , bread (33) |

# Basics on explainability and interpretability

## Deep neural networks are easily fooled!

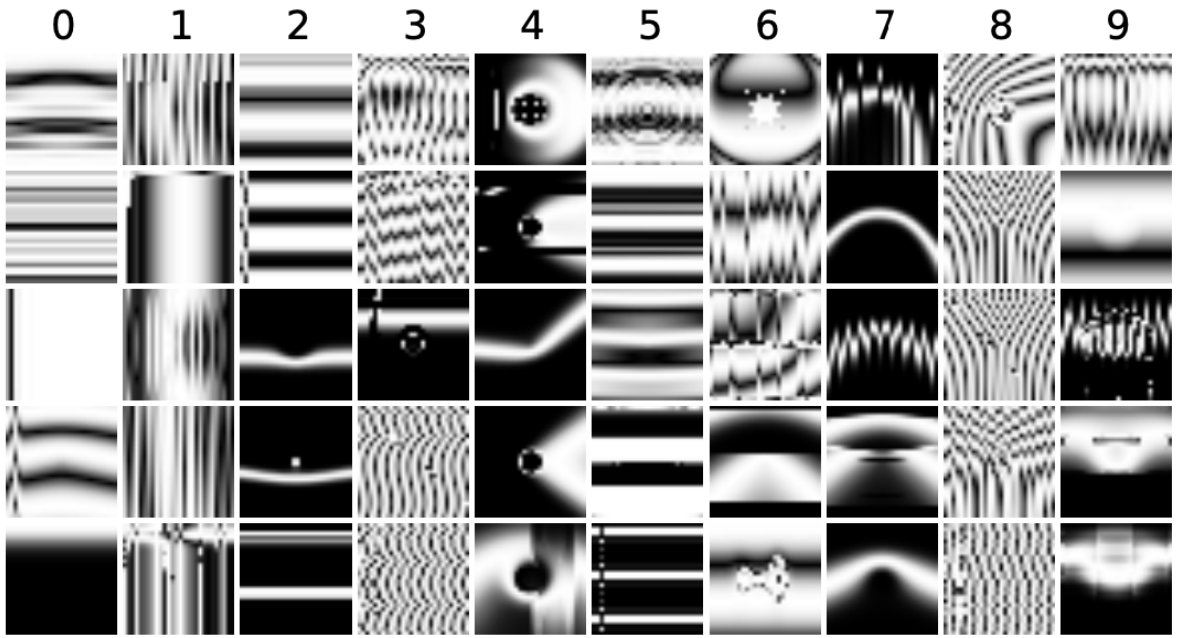



Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

# Basics on explainability and interpretability

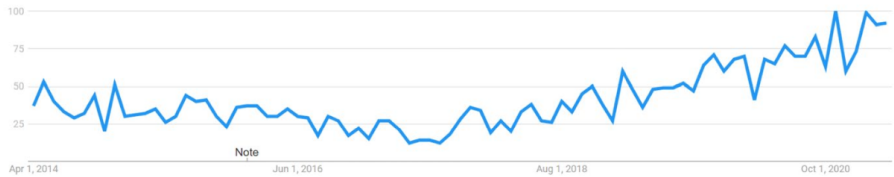
## Deep neural networks are easily fooled!

Images that MNIST DNNs believe with 99.99% confidence are digits 0-9

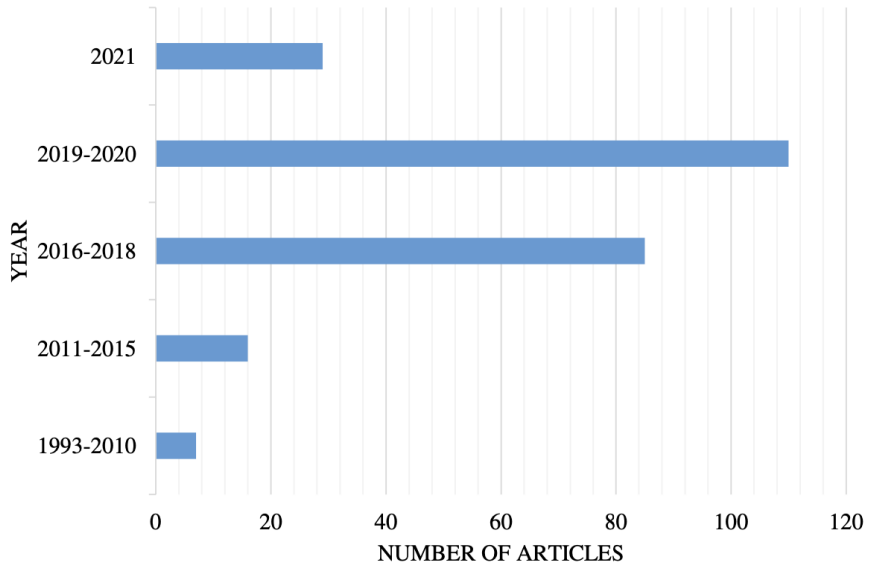



Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

# Basics on explainability and interpretability



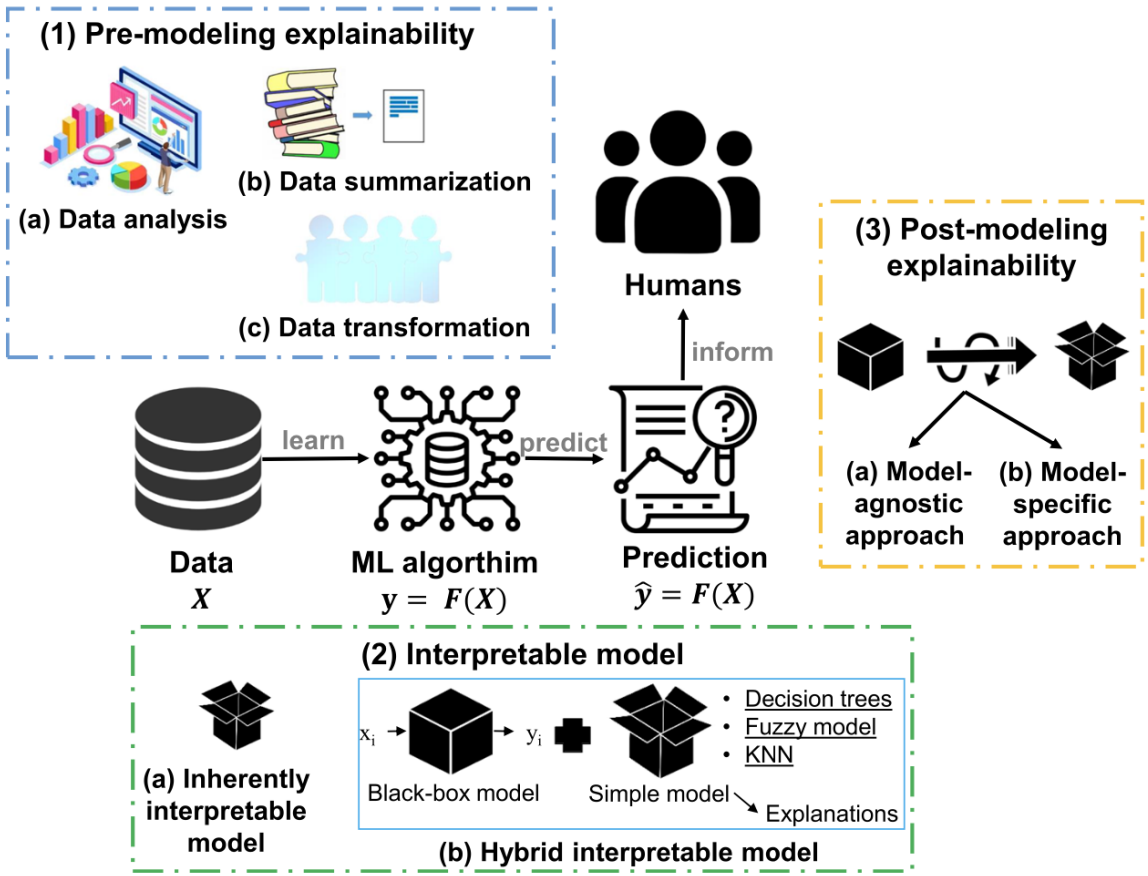

**(1) Google trends result for the Explainable AI**




**(2) Distribution of published scientific articles over time**

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66.

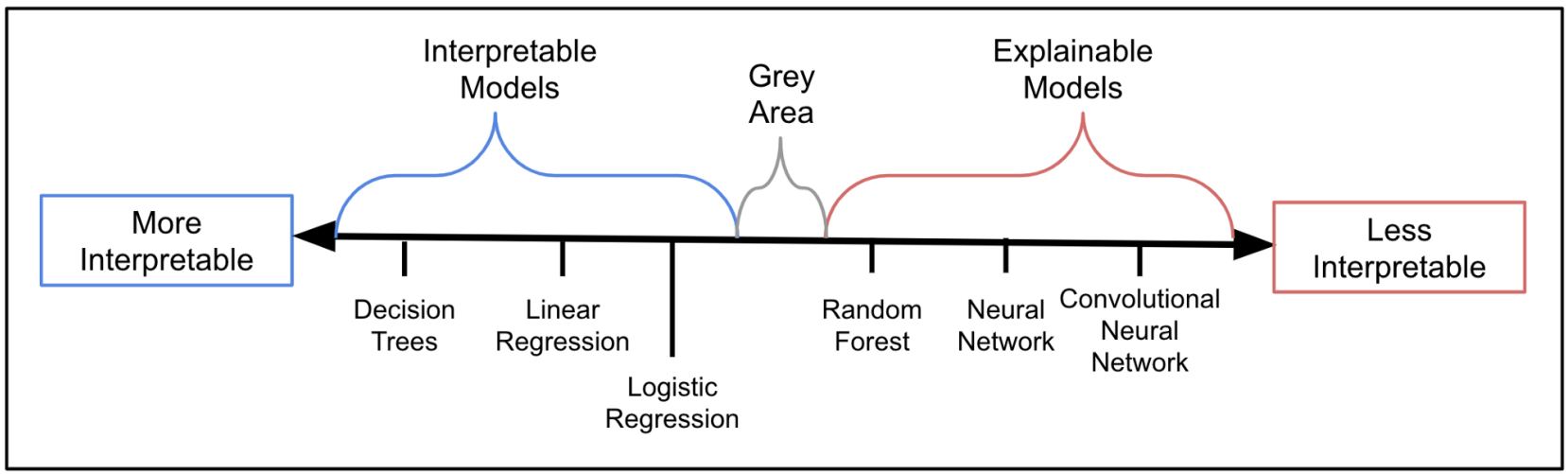
# Basics on explainability and interpretability




Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66.

# Basics on explainability and interpretability

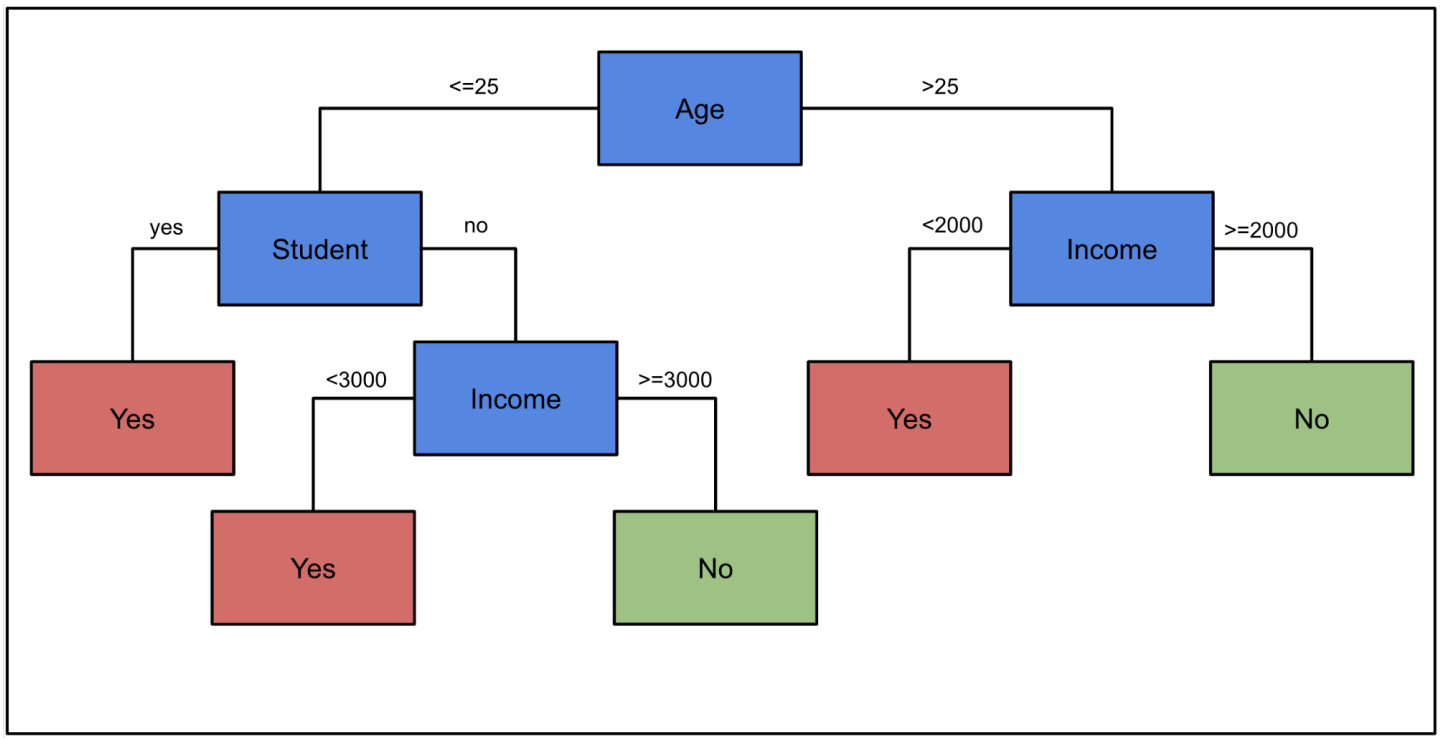
## Spectrum of interpretability



# Basics on explainability and interpretability

## Interpretable machine learning model

- Ability of the model to be understood on its own
- By looking at the model parameters or model summary, we can understand exactly why it made a certain decision
- Examples: linear regression, decision tree

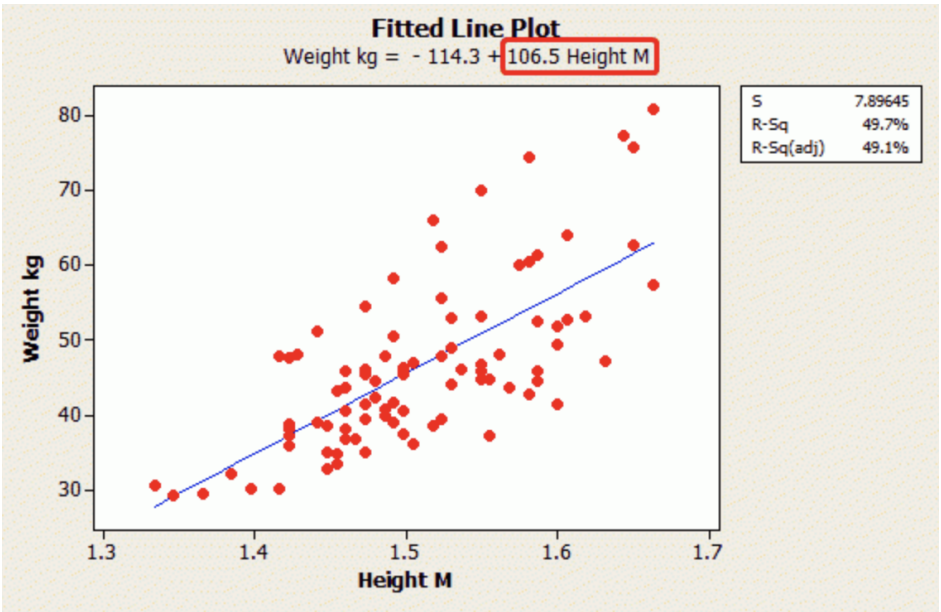



Decision tree predicting whether someone would default (Yes) or not default (No) on a car loan

# Basics on explainability and interpretability

## Interpretable machine learning model

- Interpretability of linear regression coefficients
- If the height increases by 1 meter, the average weight increases by 106.5 kilograms




Linear regression model predicting weight from the height.

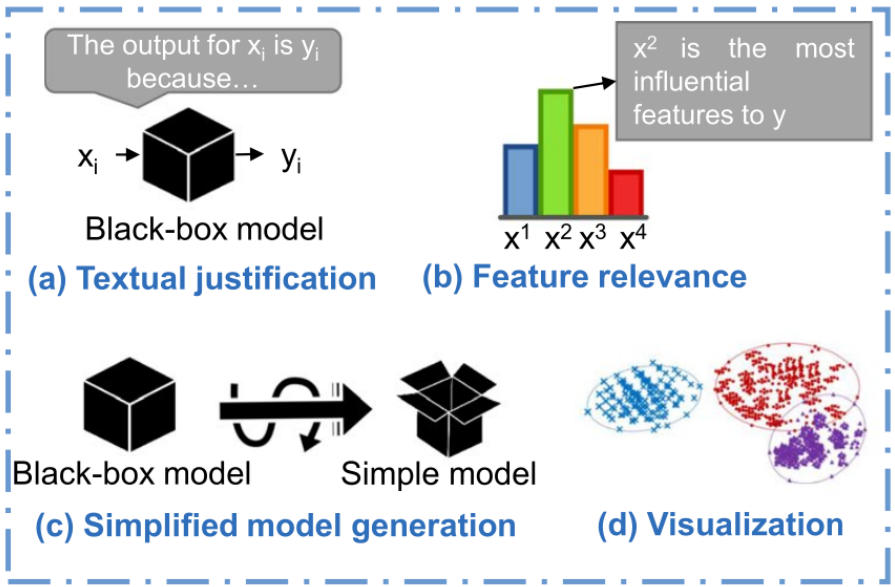
# Basics on explainability and interpretability

## Explainable machine learning model

- A model that needs an additional method for a human to understand how it works
- It is not possible for humans to understand how complicated ("black box") models work
- Examples: neural networks, convolutional neural networks, random forests

# Basics on explainability and interpretability

## Model-agnostic explainable AI methods



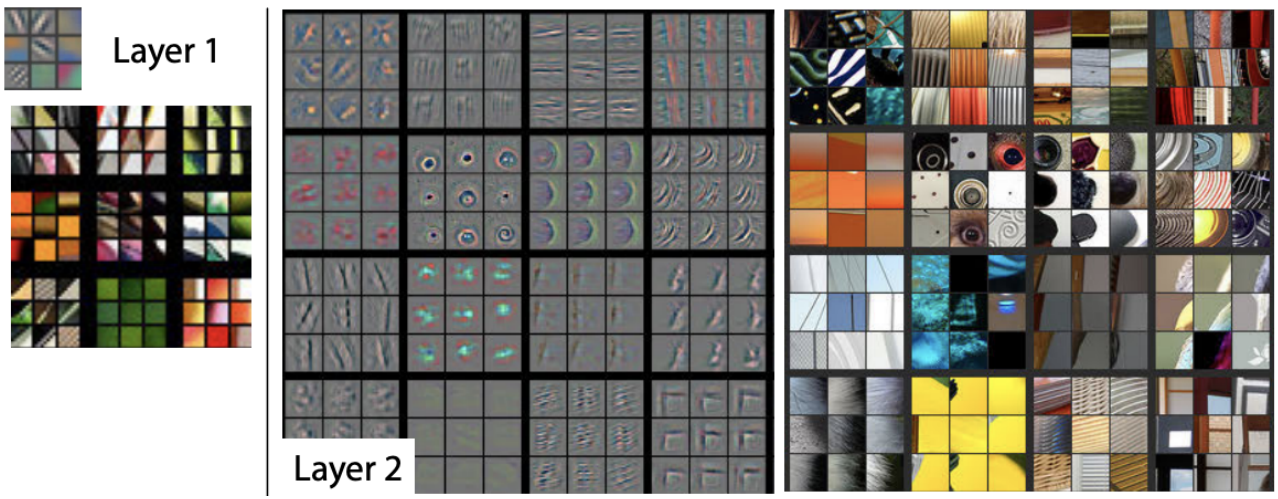

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66.

# Explainable AI

- Basics on explainability and interpretability
- Visualizing CNN activations
- Prediction difference analysis
- Local Interpretable Model-agnostic Explanations (LIME)

# Visualizing CNN activations

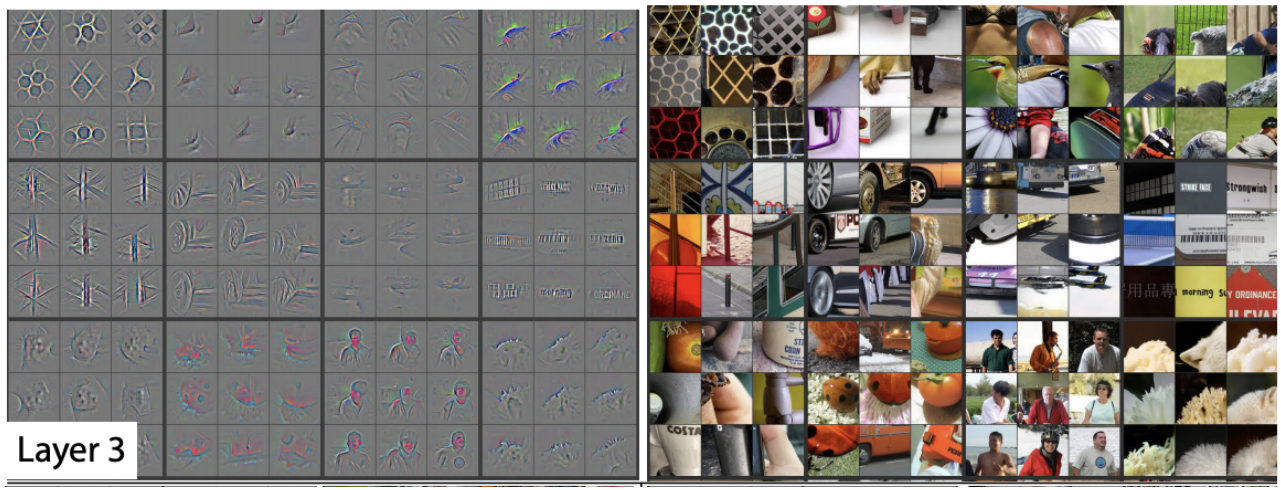
- Top 9 activations once training is complete
- Higher variance in the first layers of the network




Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

# Visualizing CNN activations

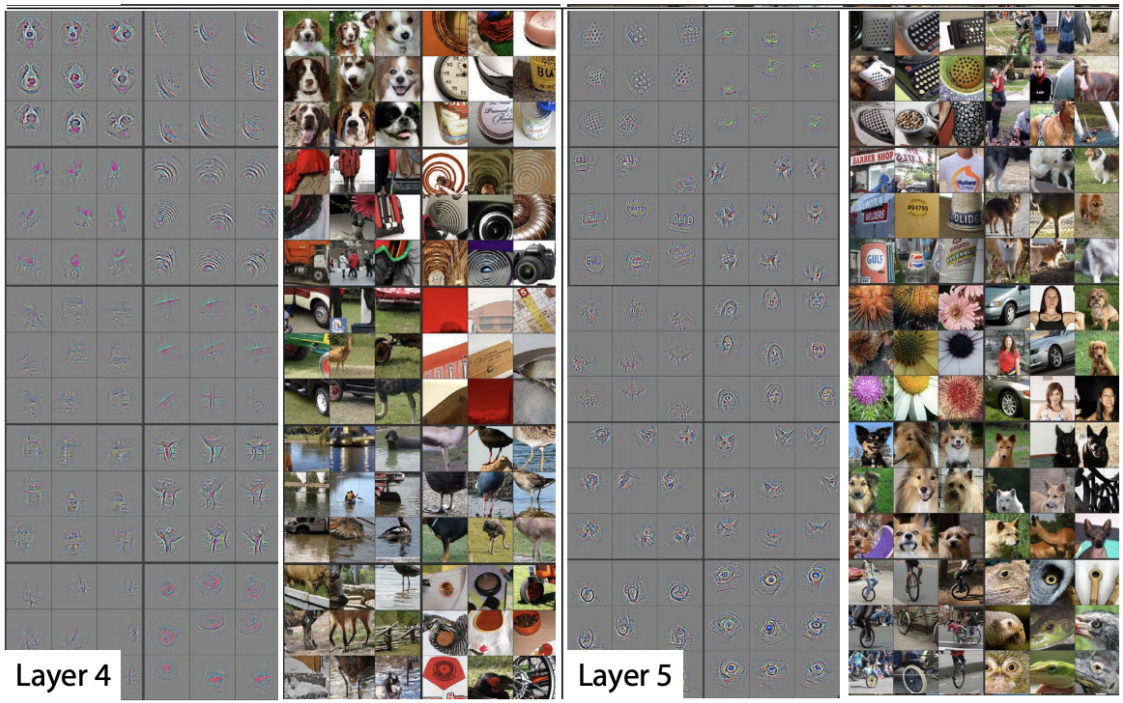
- Top 9 activations once training is complete




Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

# Visualizing CNN activations

- Top 9 activations once training is complete
- Less variance toward the last layers
- Exaggeration of discriminative parts of the image (e.g. eyes, nose)




Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

# Explainable AI

- Basics on explainability and interpretability
- Visualizing CNN activations
- Prediction difference analysis
- Local Interpretable Model-agnostic Explanations (LIME)

# Prediction difference analysis

- For a given prediction, the method assigns a relevance value to each input feature with respect to a class $c$
- The relevance of a feature $x_i$ is estimated by measuring how the prediction changes if the feature is unknown, i.e., $p(c|\mathbf{x}) - p(c|\mathbf{x} \setminus x_i)$
- Substituting the missing component with components from its neighborhood

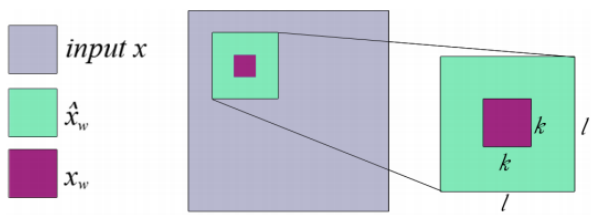
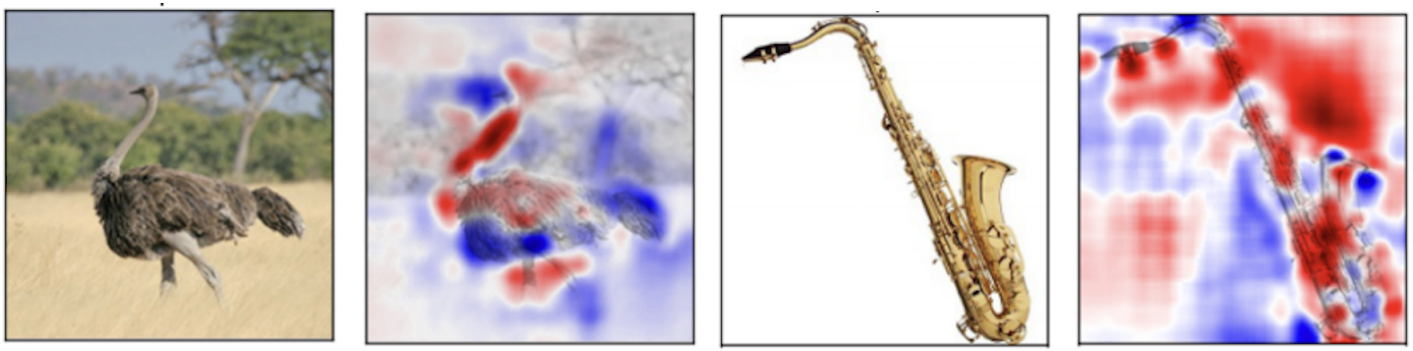



Figure 2: Simple illustration of the sampling procedure in algorithm 1. Given the input image $x$, we select every possible patch $x_w$ (in a sliding window fashion) of size $k \times k$ and place a larger patch $\hat{x}_w$ of size $l \times l$ around it. We can then conditionally sample $x_w$ by conditioning on the surrounding patch $\hat{x}_w$.

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595.

# Prediction difference analysis

Visualization of the effects of marginal versus conditional sampling using the GoogLeNet classifier. The classifier makes correct predictions (ostrich and saxophone), and we show the evidence for (red) and against (blue) this decision at the output layer.



Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595.

# Prediction difference analysis

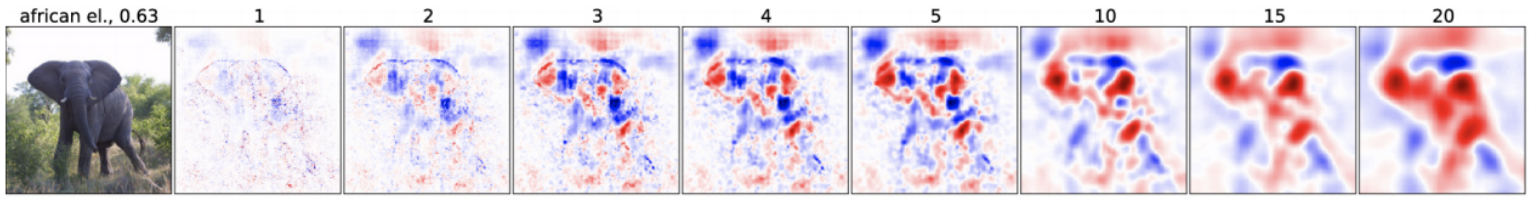



Figure 4: **Visualization of how different window sizes influence the visualization result.** We used the conditional sampling method and the AlexNet classifier with $l = k + 4$ and varying $k$. We can see that even when removing single pixels ($k = 1$), this has a noticeable effect on the classifier and more important pixels get a higher score. By increasing the window size we can get a more easily interpretable, smooth result until the image gets blurry for very large window sizes.

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595.

# Explainable AI

- Basics on explainability and interpretability
- Visualizing CNN activations
- Prediction difference analysis
- Local Interpretable Model-agnostic Explanations (LIME)

# Local Interpretable Model-agnostic Explanations (LIME)

- Trusting a prediction: whether a user trusts an individual prediction sufficiently to take some action based on it
  - e.g., using a machine learning model for medical diagnosis or terrorism detection, predictions cannot be acted upon on blind faith, as the consequences may be highly consequential
  - LIME: explains the predictions of any classifier or regressor in a faithful way
- Trusting a model: whether the user trusts a model to behave in reasonable ways if deployed
  - SP-LIME: selects a set of representative instances with explanations to address the "trusting the model" problem

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). `https://github.com/marcotcr/lime`

# Local Interpretable Model-agnostic Explanations (LIME)

## Interpretable data representations

- Text: A binary vector indicating the presence or absence of a word contributing to the model decision
- Image: A binary vector indicating the "presence" or "absence" of a contiguous patch of similar pixels (a super-pixel)

# Local Interpretable Model-agnostic Explanations (LIME)

## Explaining a prediction

- Presenting artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction

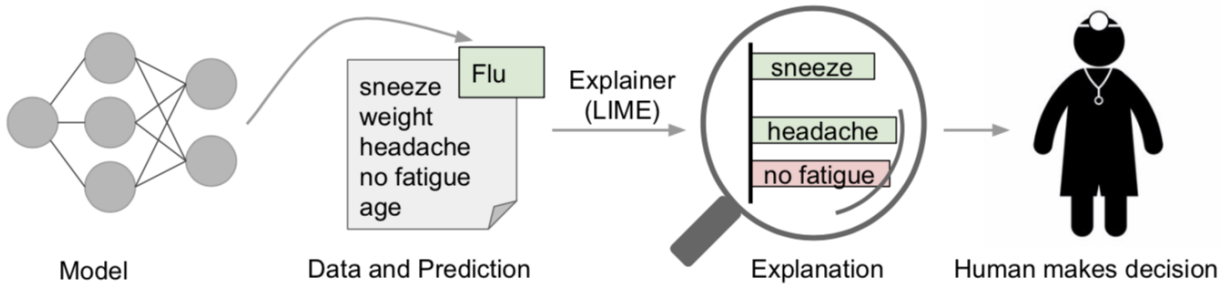



**Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.**

# Local Interpretable Model-agnostic Explanations (LIME)

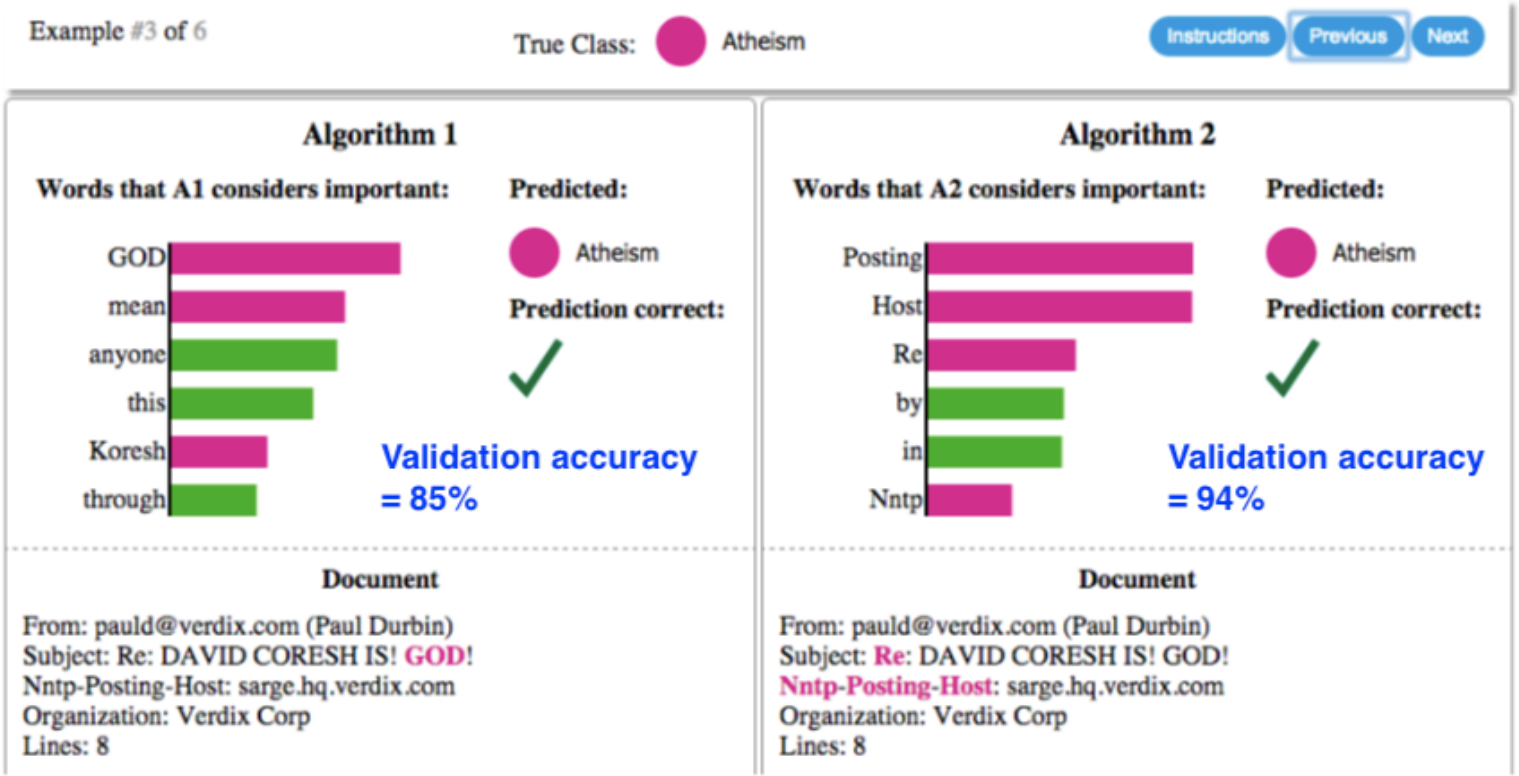
## Explaining a prediction

- Presenting artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction

# Local Interpretable Model-agnostic Explanations (LIME)

## Which model would you trust?

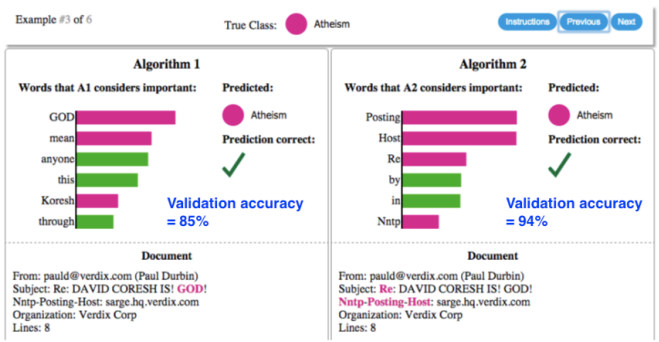
A classifier was trained to determine if a document is about "Christianity" or "Atheism"



# Local Interpretable Model-agnostic Explanations (LIME)

## Explaining a prediction

- The right algorithm has higher accuracy on the validation set, but its explanations actually do not make much sense
- Frequent mismatch between performance metrics (e.g. accuracy) and metrics related to user engagement and retention
- Interpretability depends on the audience: ML practitioners may interpret small networks, but the general audience may be more comfortable with a few weighted features as an explanation



# Local Interpretable Model-agnostic Explanations (LIME)

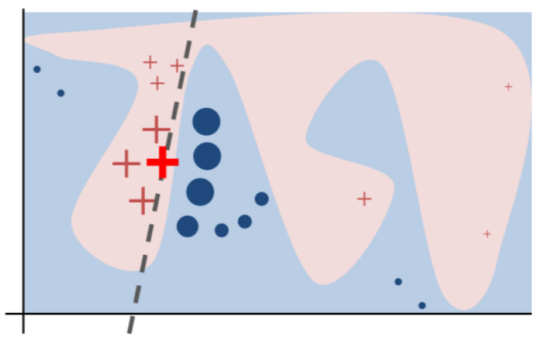
## Desired characteristics of explainability

- Local fidelity: An explanation has to correspond to how the model behaves at least in the vicinity of the instance being predicted
- Model agnostic: An explainer should be able to explain any model (i.e. treat the original model as a black box)
- Global perspective: We select a few explanations to present to the user, such that they are representative of the model

# Local Interpretable Model-agnostic Explanations (LIME)

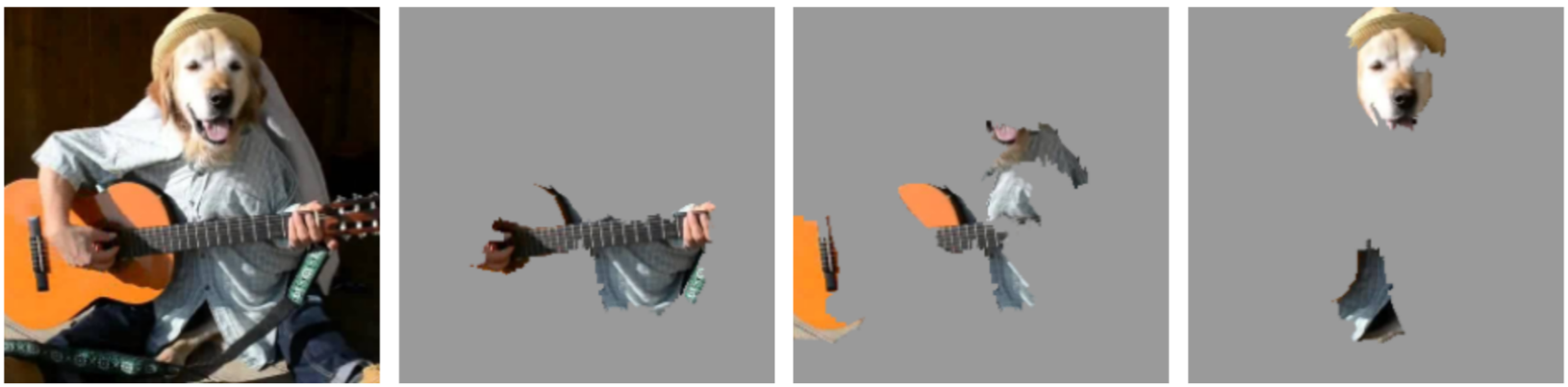
## Explaining a prediction

- Complex decision function $f$ represented by the red/blue boundaries
- Explanation model $g$ represented with the black line
  - **Approximates** the decision boundary **locally**
  - The linear model ensures interpretability through the weight of the features and by imposing a **sparsity** constraint (i.e., the line is learned such that most weights are zero, except the weights corresponding to the most important features)



# Local Interpretable Model-agnostic Explanations (LIME)

## Explaining a prediction



(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" ($p = 0.21$)**

# Local Interpretable Model-agnostic Explanations (LIME)

## Explaining a model

- SP-LIME choses a diverse, representative, and non-redundant set of samples to show and explain to the user
- Explanation matrix: local importance of the interpretable components for each instance
- Computing global importance for each component/feature
  - Features that explain many different instances have higher importance scores
- **Submodular optimization** to select the components and samples

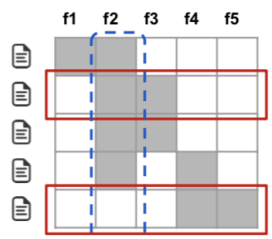



Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

# Local Interpretable Model-agnostic Explanations (LIME)

## User evaluation: Can users select the best model?

- Amazon Turkers were asked to choose the classifier they would deploy "in the wild" for classifying a text between "atheism" and "christanity". The two classifiers are:
  - SVM trained on the original dataset (84% accuracy)
  - SVM trained on clean data with a subset of features (88% acc)
- Greedy feature selection (remove features that contribute the most to the predicted class until the prediction changes) vs LIME
- Submodular pick (see previous slide) vs random pick (randomly picking features to show to the user)

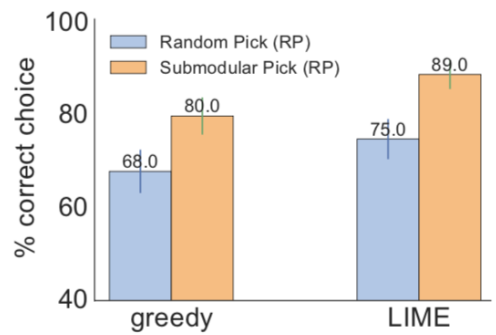
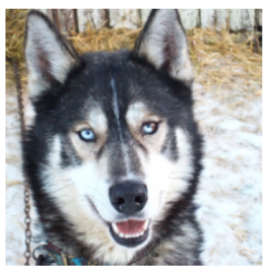



**Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.**

# Local Interpretable Model-agnostic Explanations (LIME)

## User evaluation: Do explanations lead to insights?

- Classifying between Wolves and Huskies using synthetic dataset that induced undesirable biases
  - All images of wolves with snow in the background
  - All images of huskies without snow in the background
- Amazon Turkers were presented with the images and predictions first without and then with explanations, and were asked: (1) Whether they trust the algorithm; and (2) How the algorithm reaches decision



(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.**

# In-class activity

## Detecting deceptive speech

Tutul, A. A., Chaspari, T., Levitan, S. I., & Hirschberg, J. (2023, September). Human-AI Collaboration for the Detection of Deceptive Speech. In 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-4). IEEE.

# In-class activity

## Detecting deceptive speech - Follow-up!