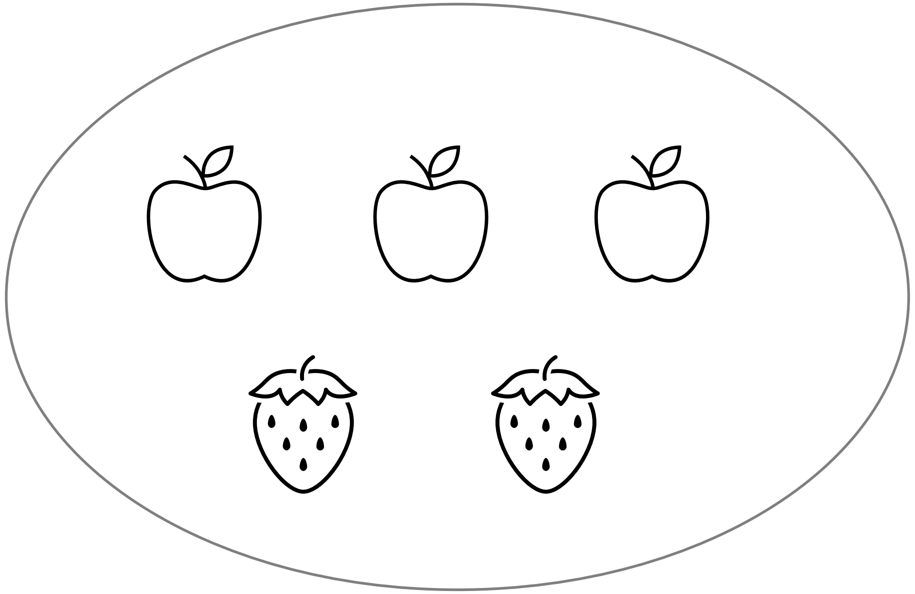
CS 4501/6501 Interpretable Machine Learning

Introduction to Interpretability

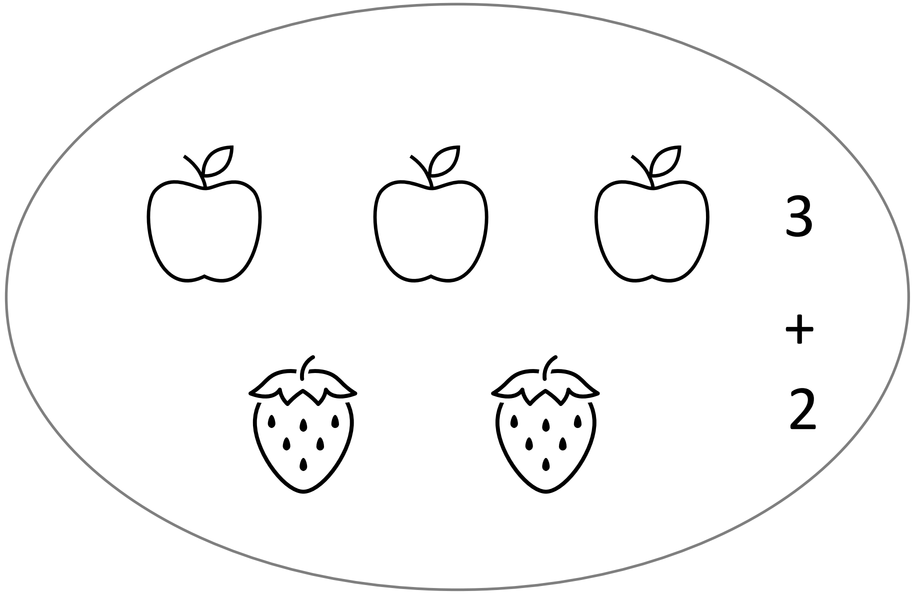
Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

Machine Learning



How many fruits?

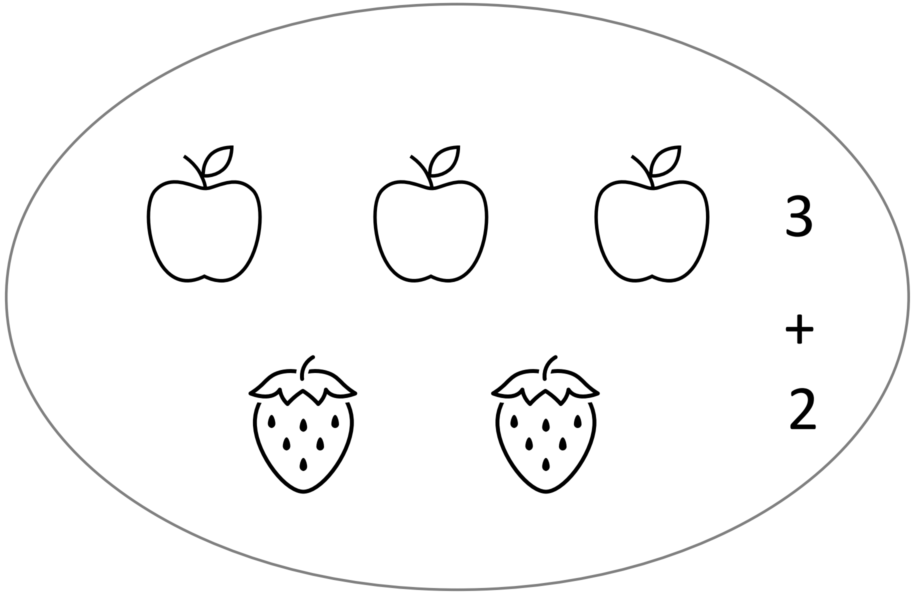
Machine Learning



How many fruits?

5

Machine Learning



How many fruits?

5

Features

x_1: the number of apples

x_2: the number of strawberries

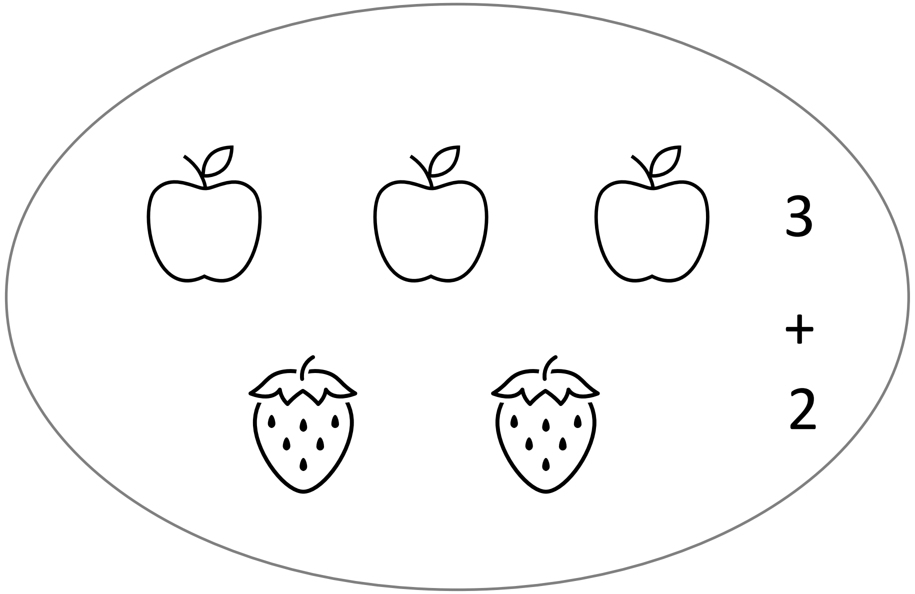
Rule

$x_1 + x_2$

Output

y: the total number of fruits

Machine Learning



How many fruits?

5

What is the contribution of each feature?

Features

x_1: the number of apples

x_2: the number of strawberries

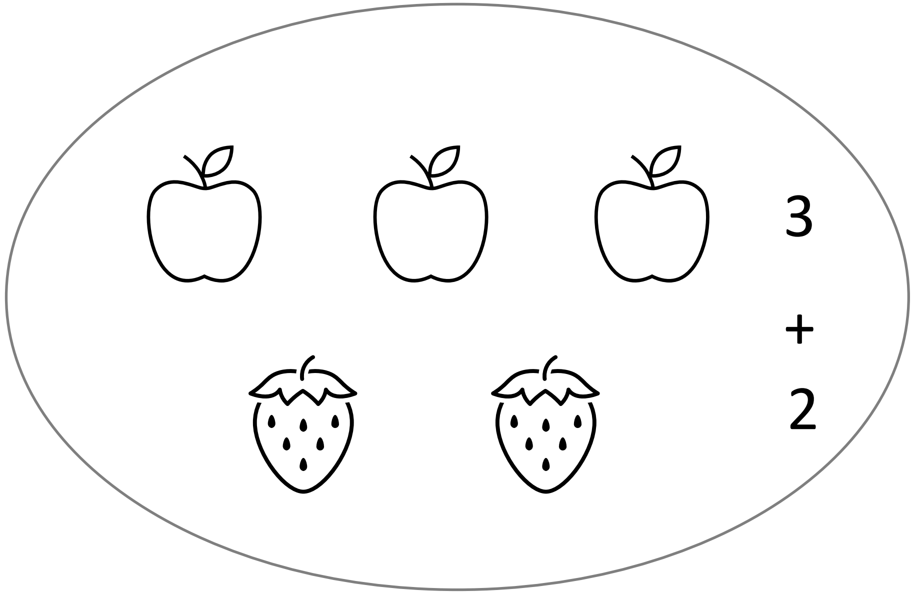
Rule

$x_1 + x_2$

Output

y: the total number of fruits

Machine Learning



How many fruits?

5

The contributions of apple and strawberry are 3 and 2 respectively

Features

x_1: the number of apples

x_2: the number of strawberries

Rule

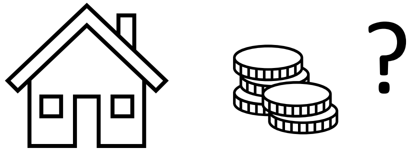
$x_1 + x_2$

Output

y: the total number of fruits

Machine Learning

A more complex problem: predict the value of a house



Features

x_1: house size

x_2: location

x_3: floor type

Rule

$$0.6x_1 + 0.3x_2 + 0.1x_3$$

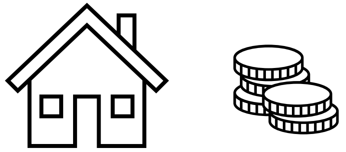
house size, location, and floor type account for 60%, 30%, 10% respectively

Output

y: house value

Machine Learning

A more complex problem: predict the value of a house



?

Features

x_1: house size

x_2: location

x_3: floor type

Rule

$$0.6x_1 + 0.3x_2 + 0.1x_3$$

house size, location, and floor type account for 60%, 30%, 10% respectively

Output

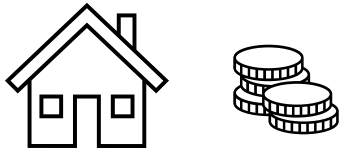
y: house value

$$x_1 = 100,$$
$$x_2 = 300,$$
$$x_3 = 200$$

$$y = 170$$

Machine Learning

A more complex problem: predict the value of a house



?

Features	**Rule**	**Output**
x_1: house size x_2: location x_3: floor type	$0.6x_1 + 0.3x_2 + 0.1x_3$ $\longrightarrow$ house size, location, and floor type account for 60%, 30%, 10% respectively	y: house value

$$x_1 = 100,\quad x_2 = 300,\quad x_3 = 200$$

$$y = 170$$

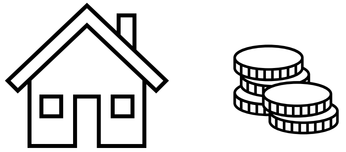
Contributions:

$x_1: 100 \times 0.6 = 60$

$x_2: 300 \times 0.3 = 90$

$x_3: 200 \times 0.1 = 20$

Machine Learning

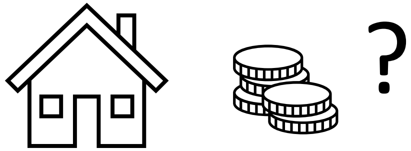
A more complex problem: predict the value of a house  ?

Features	**Rule**	**Output**
x_1: house size	$\xrightarrow{\quad ? \quad}$	y: house value
x_2: location		
x_3: floor type		

What if we do not know the rule?

Machine Learning

Learn a rule from past house sales



Features

$\{\boldsymbol{x} = (x_1, x_2, x_3)\}$

Machine Learning

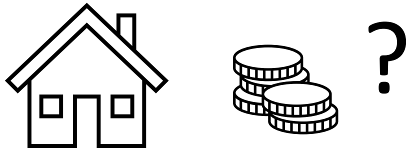
$f_{\boldsymbol{w}}(\cdot)$

Output

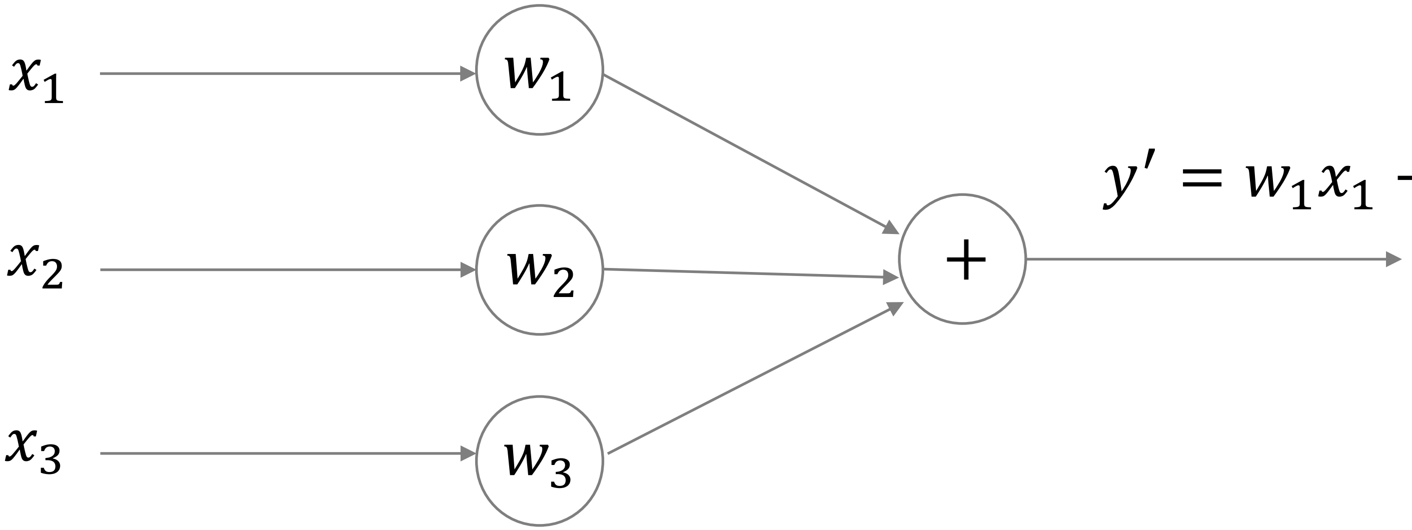
$\{y\}$

Machine Learning

Learn a rule from past house sales



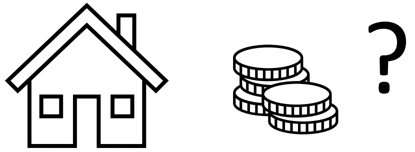
Features	**Machine Learning**	**Output**
$\{\boldsymbol{x} = (x_1, x_2, x_3)\}$	$f_{\boldsymbol{w}}(\cdot)$	$\{y\}$



$$y' = w_1 x_1 + w_2 x_2 + w_3 x_3$$

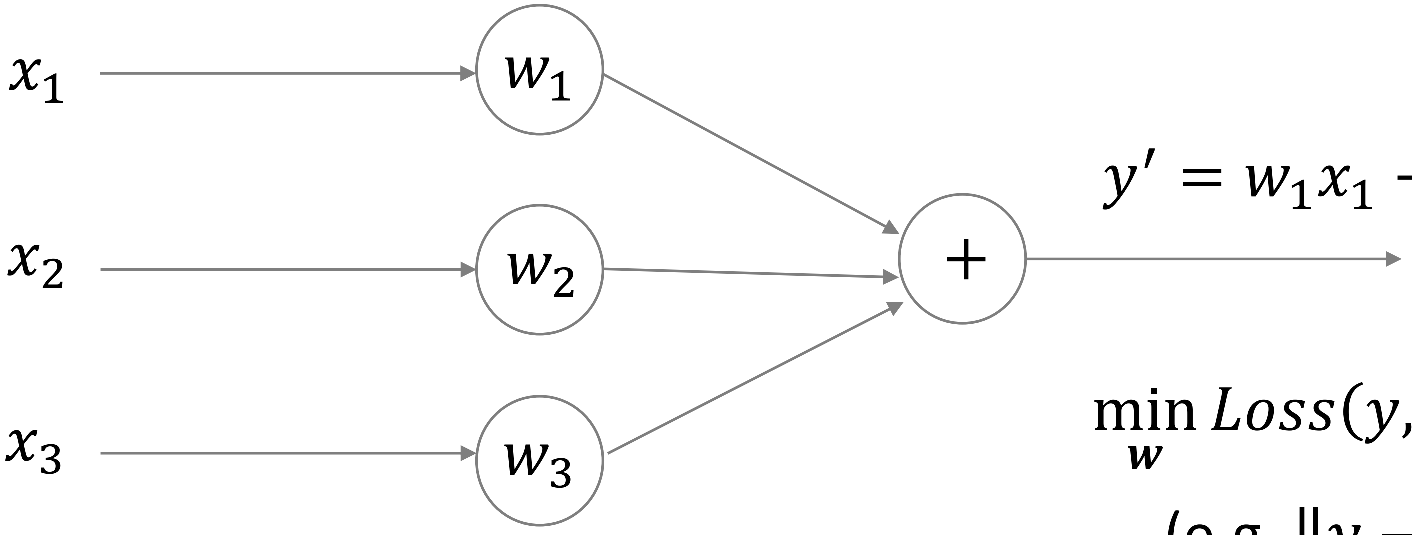
Machine Learning

Learn a rule from past house sales



Features | **Machine Learning** | **Output**

$\{\boldsymbol{x} = (x_1, x_2, x_3)\}$ | $f_{\boldsymbol{w}}(\cdot)$ | $\{y\}$



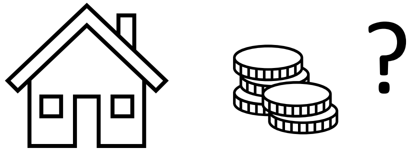
$$y' = w_1x_1 + w_2x_2 + w_3x_3$$

$$\min_{\boldsymbol{w}} Loss(y, y')$$

(e.g. $\|y - y'\|_2$)

Machine Learning

Learn a rule from past house sales



Features

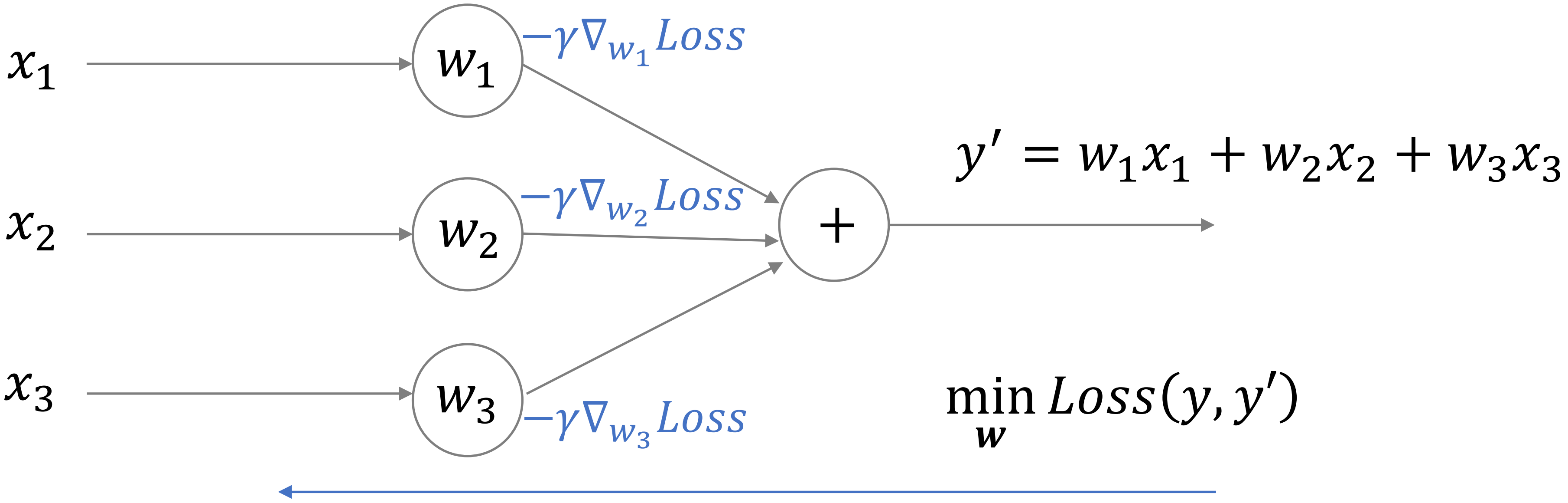
$\{\boldsymbol{x} = (x_1, x_2, x_3)\}$

Machine Learning

$f_{\boldsymbol{w}}(\cdot)$

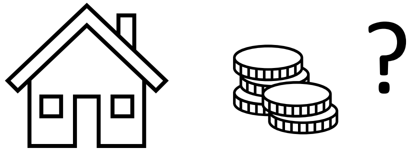
Output

$\{y\}$

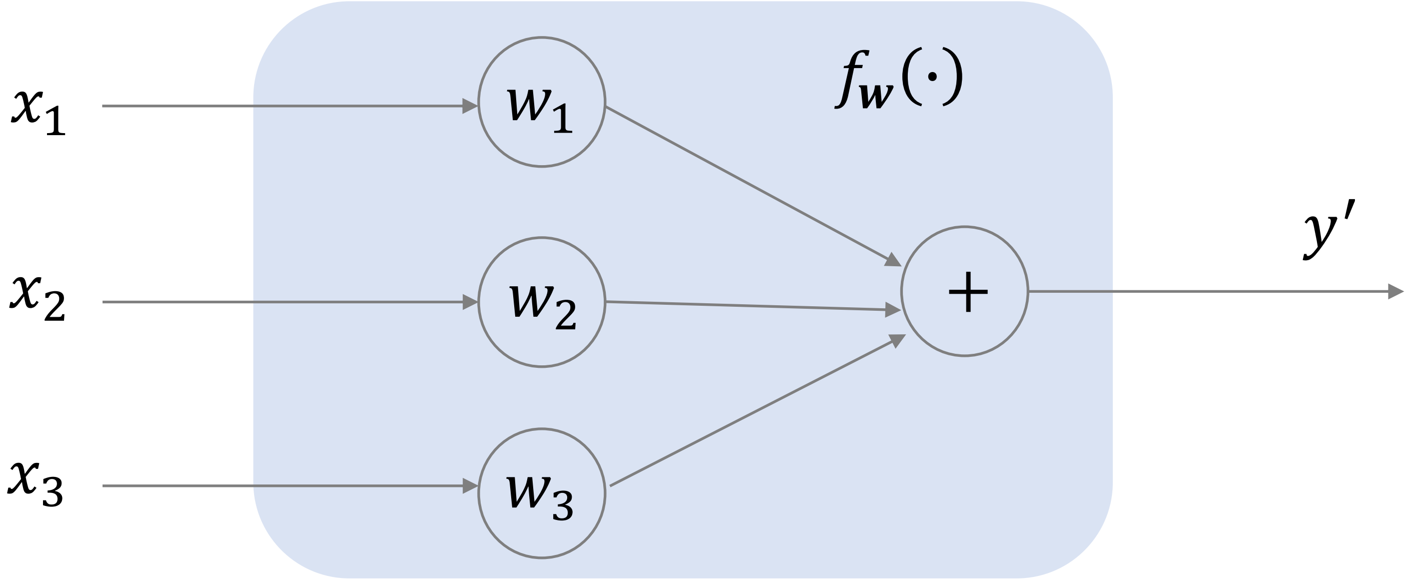


Machine Learning

Predict the house value via the learned machine learning model

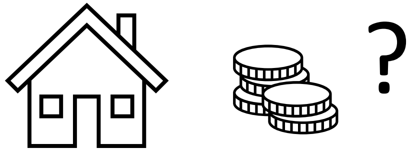


Machine Learning model

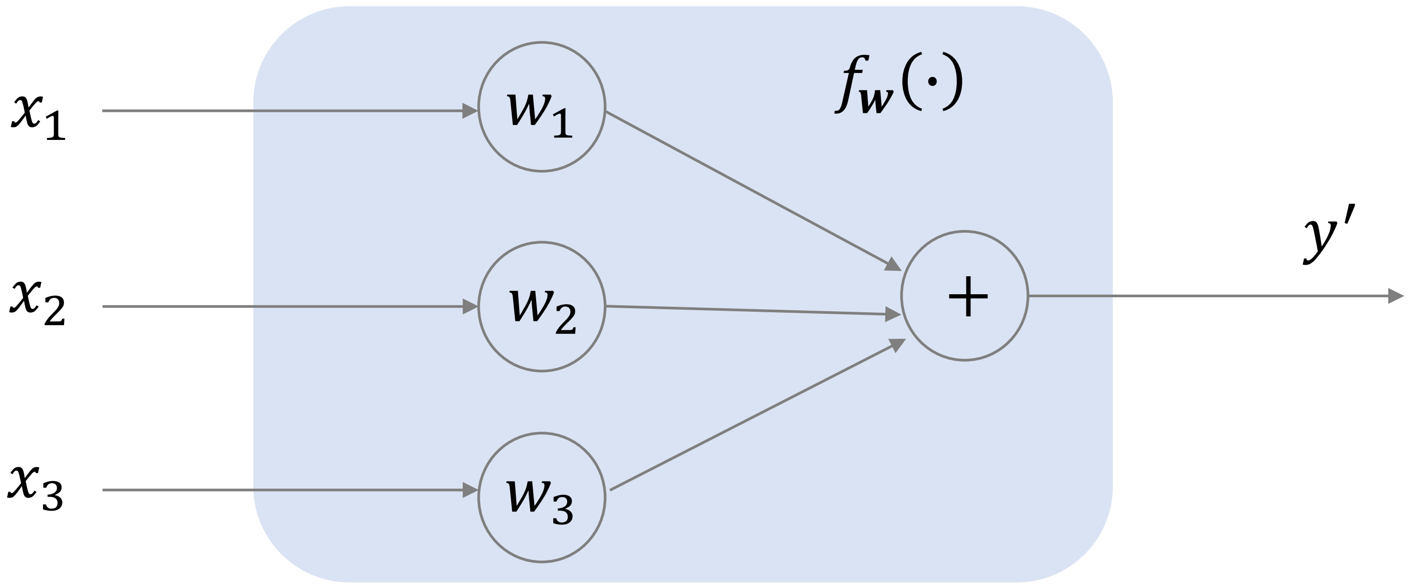


Machine Learning

Predict the house value via the learned machine learning model



Machine Learning model → Interpretable



Contributions:

$x_1 : w_1 x_1$

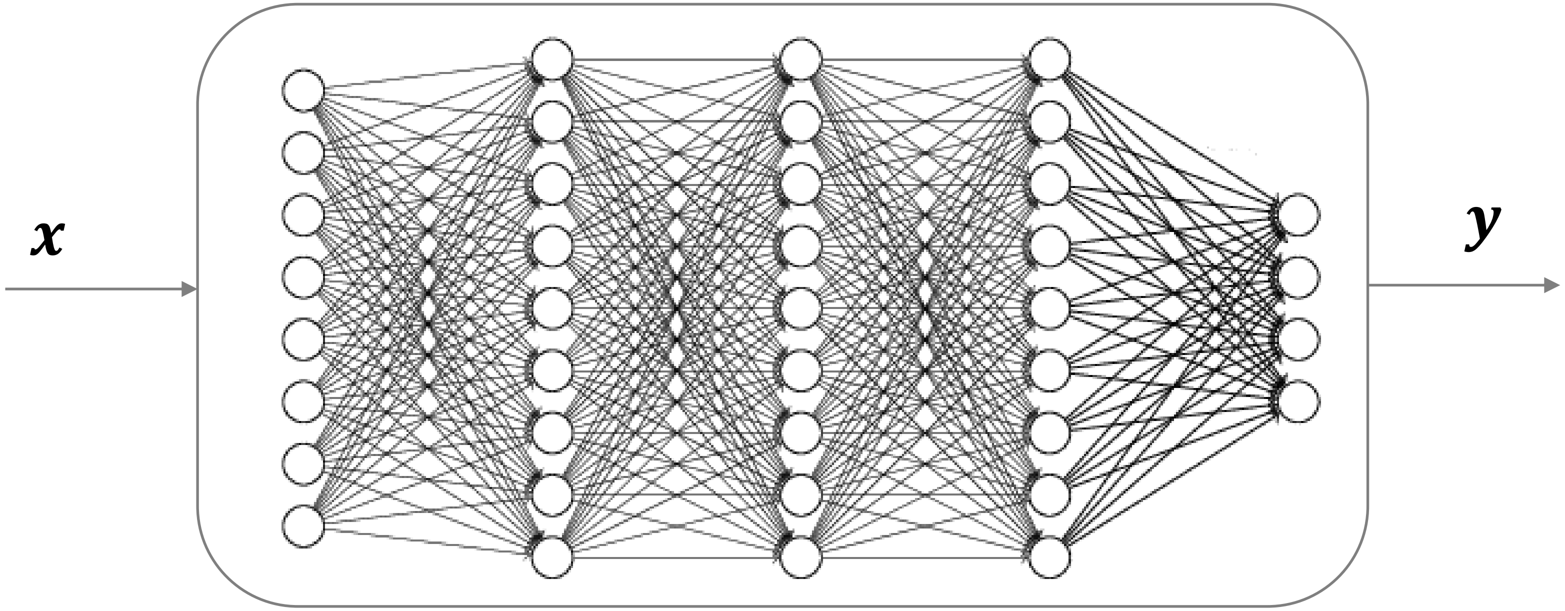
$x_2 : w_2 x_2$

$x_3 : w_3 x_3$

Machine Learning

In reality, features and relationships can be more complex

Machine Learning model



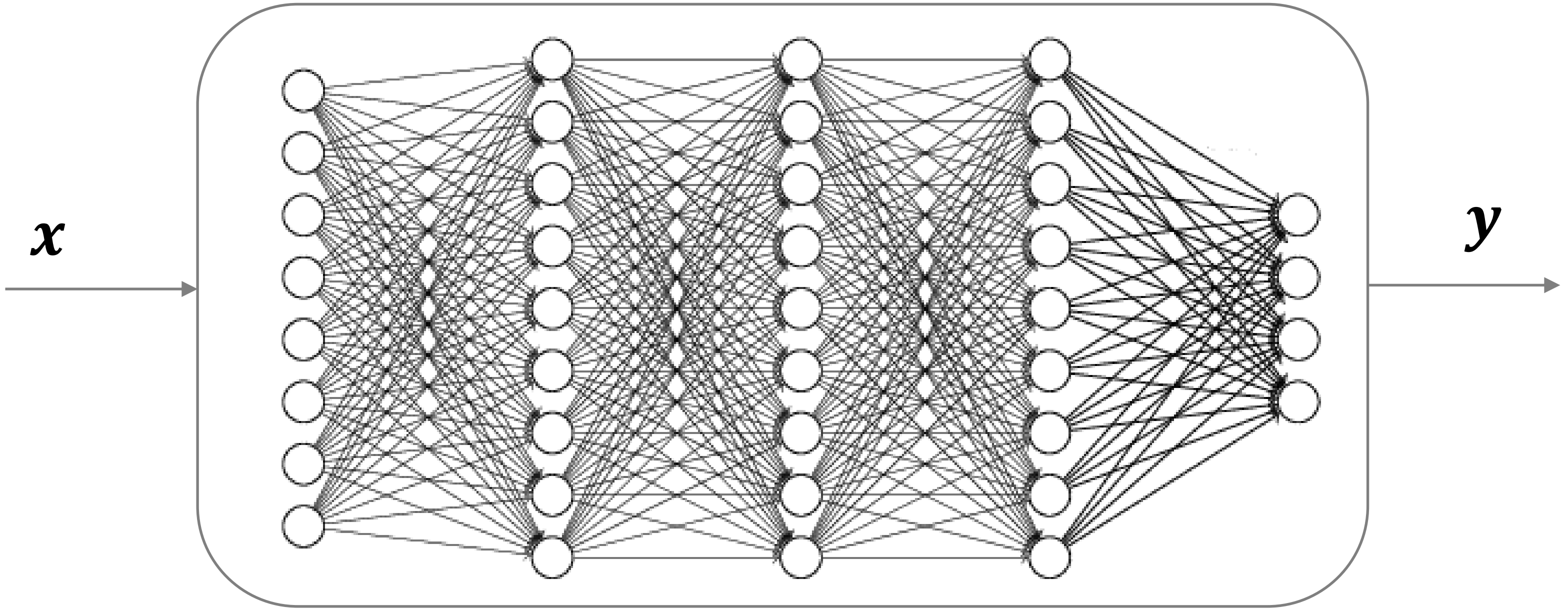
(nonlinear and complex transformations)

Machine Learning

In reality, features and relationships can be more complex

Machine Learning model

Uninterpretable



It is difficult to understand the model's inner working and trace the contributions of input features

(nonlinear and complex transformations)

Machine Learning

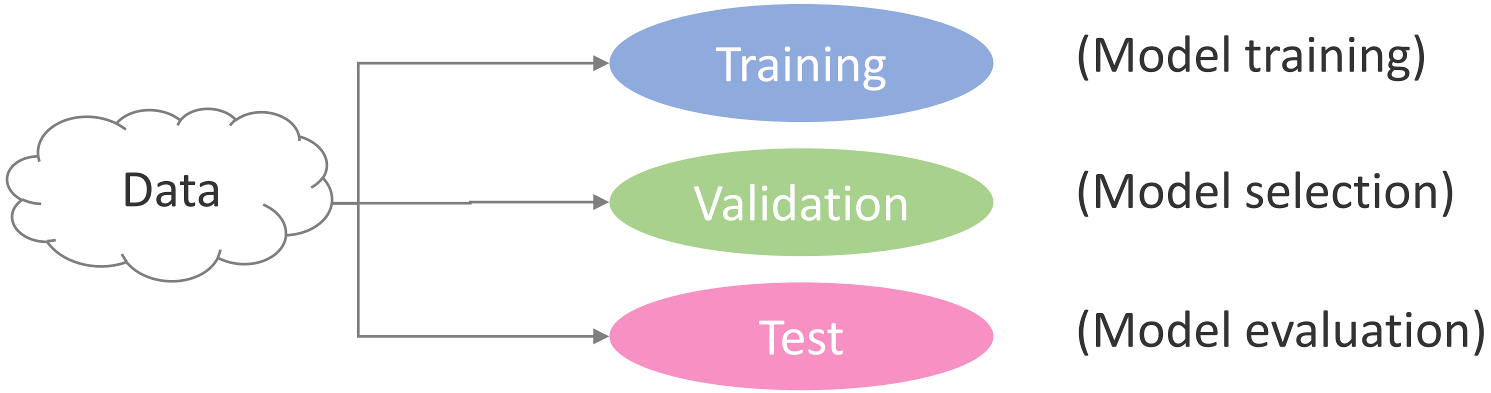
Machine learning is a set of methods that computers use to make and improve predictions or behaviors based on data

Machine Learning

Machine learning is a set of methods that computers use to make and improve predictions or behaviors based on data

1 Collecting data

$$\{(\boldsymbol{x}, \boldsymbol{y})\}$$

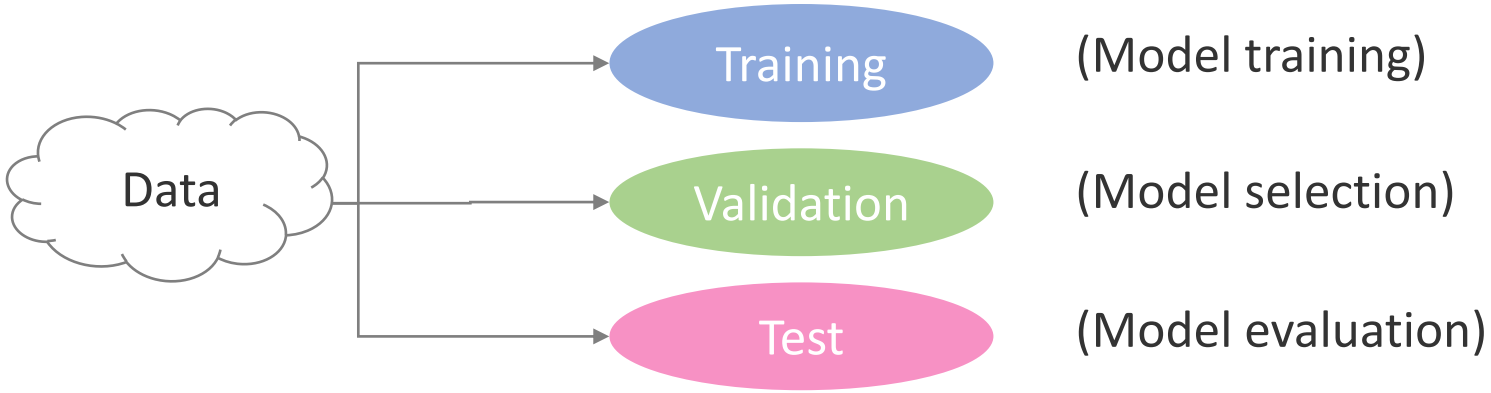


Machine Learning

Machine learning is a set of methods that computers use to make and improve predictions or behaviors based on data

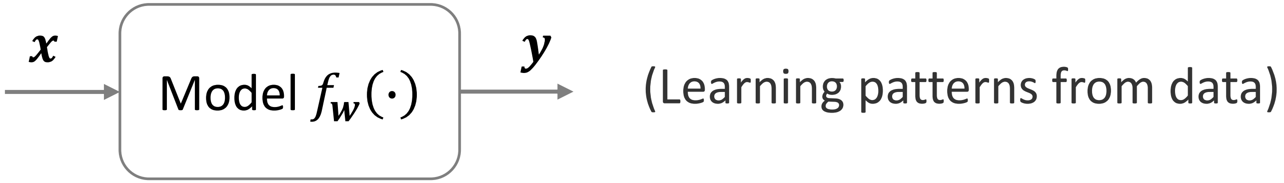
1. Collecting data

   $\{(\boldsymbol{x}, \boldsymbol{y})\}$



2. Training a machine learning model

   $f_w(\cdot)$

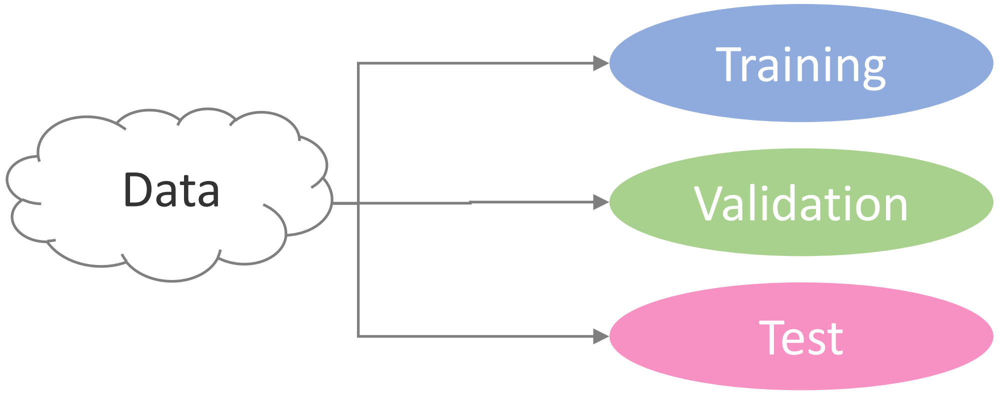


Machine Learning

Machine learning is a set of methods that computers use to make and improve predictions or behaviors based on data

1. Collecting data

$$\{(\boldsymbol{x}, \boldsymbol{y})\}$$



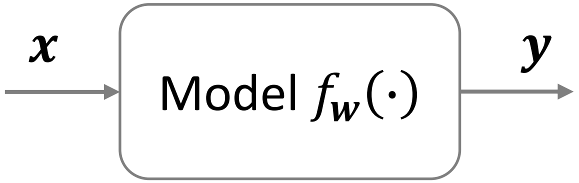
(Model training)

(Model selection)

(Model evaluation)

2. Training a machine learning model

$$f_w(\cdot)$$



(Learning patterns from data)

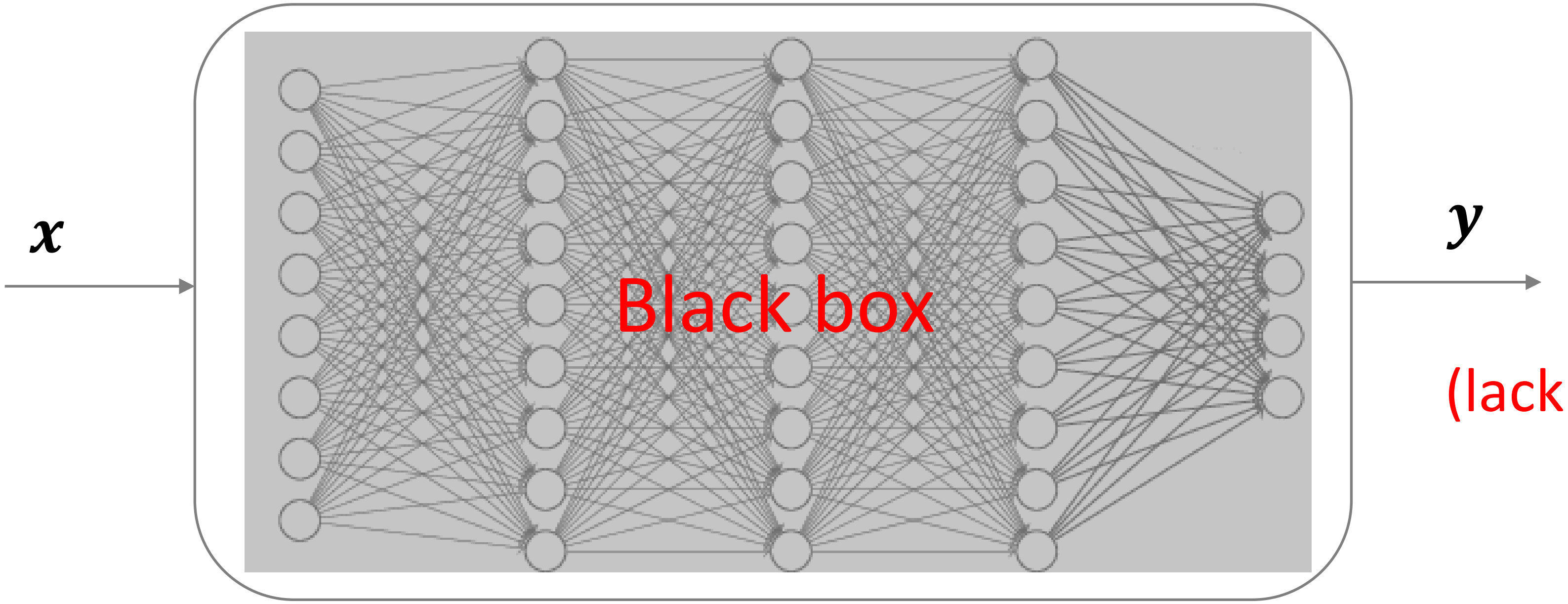
3. Testing the model

$$\boldsymbol{y}' = f_w(\boldsymbol{x})$$

Interpretability

When data and tasks are complex, machine learning models are becoming bigger and sophisticated

Machine Learning model



Interpretability

- What is interpretability?

- Why interpretability is important?

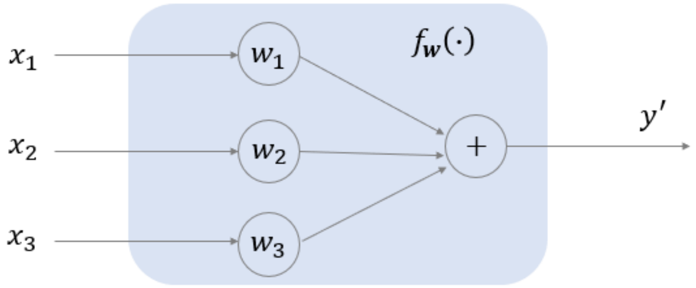
Interpretability

There is no standard or mathematical definition of interpretability

- Interpretability is the degree to which a human can understand the cause of a decision [Miller, 2019]
- Interpretability is the degree to which a human can consistently predict the model's result [Kim et al., 2016]

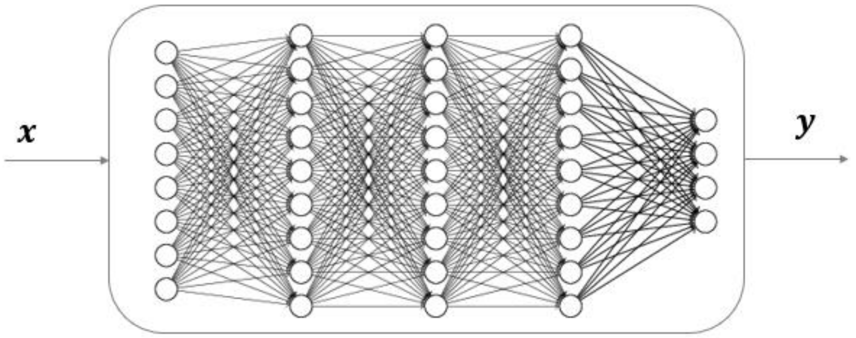
Interpretability

A simple model is usually more interpretable than a complex neural network model



- Three parameters (w_1, w_2, w_3)
- $y' = w_1x_1 + w_2x_2 + w_3x_3$
- Contributions:

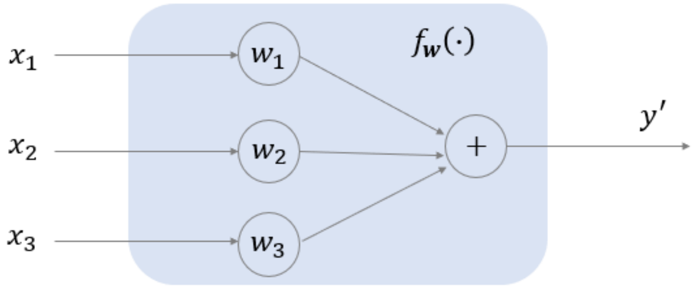
$$x_1 : w_1x_1$$
$$x_2 : w_2x_2$$
$$x_3 : w_3x_3$$

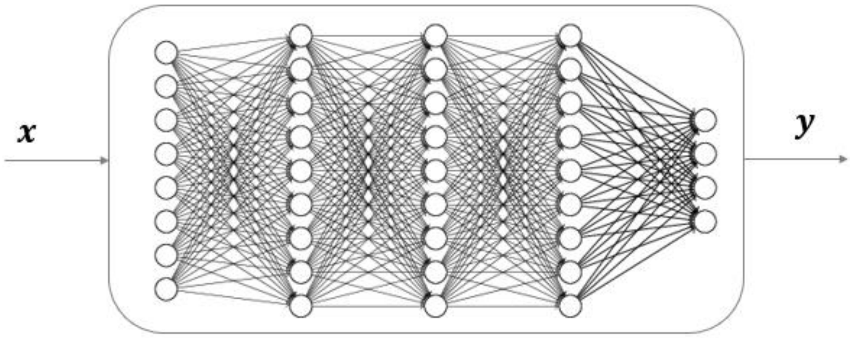


- Millions of parameters
- $\boldsymbol{y}' = f_{\boldsymbol{w}}(\boldsymbol{x})$ (complex transformations)
- Model decision-making and feature attributions are unclear

Interpretability

There is a trade-off between model performance and interpretability



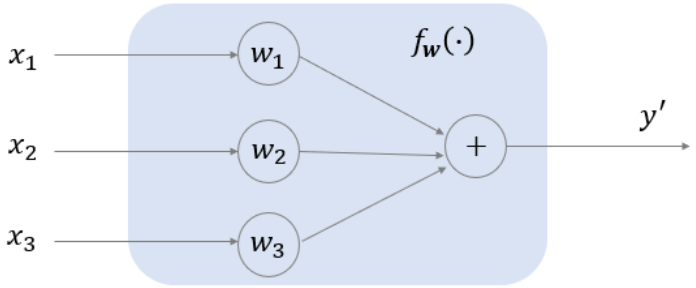


Bad performance
Good interpretability

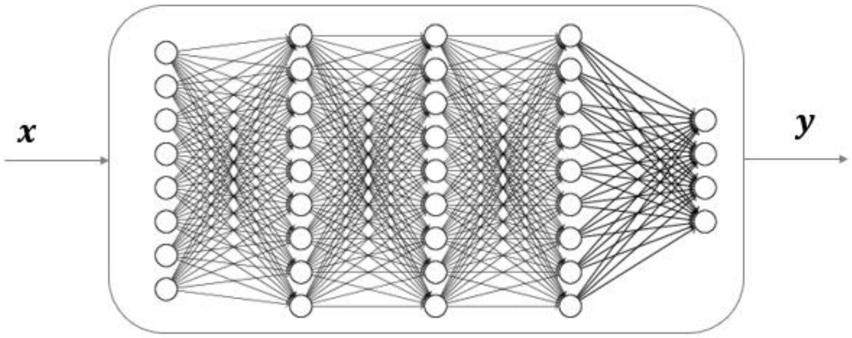
Good performance
Bad interpretability

Interpretability

Depending on the specific task…

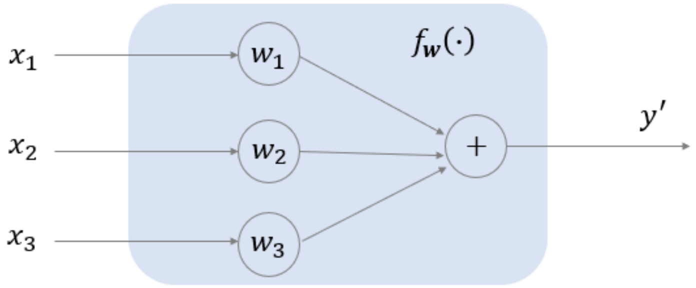


Simple task

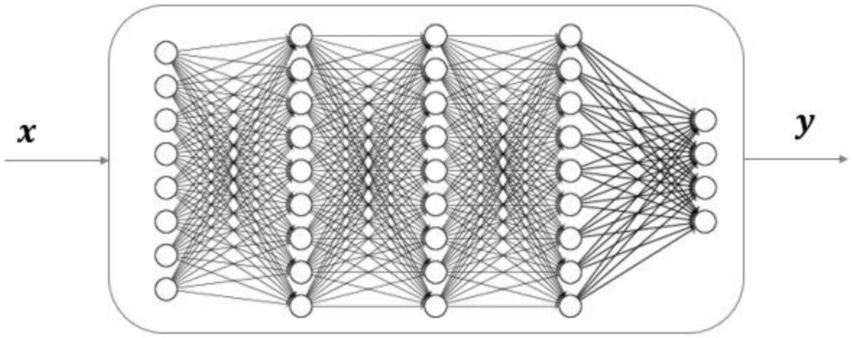


Interpretability

Depending on the specific task…



Simple task

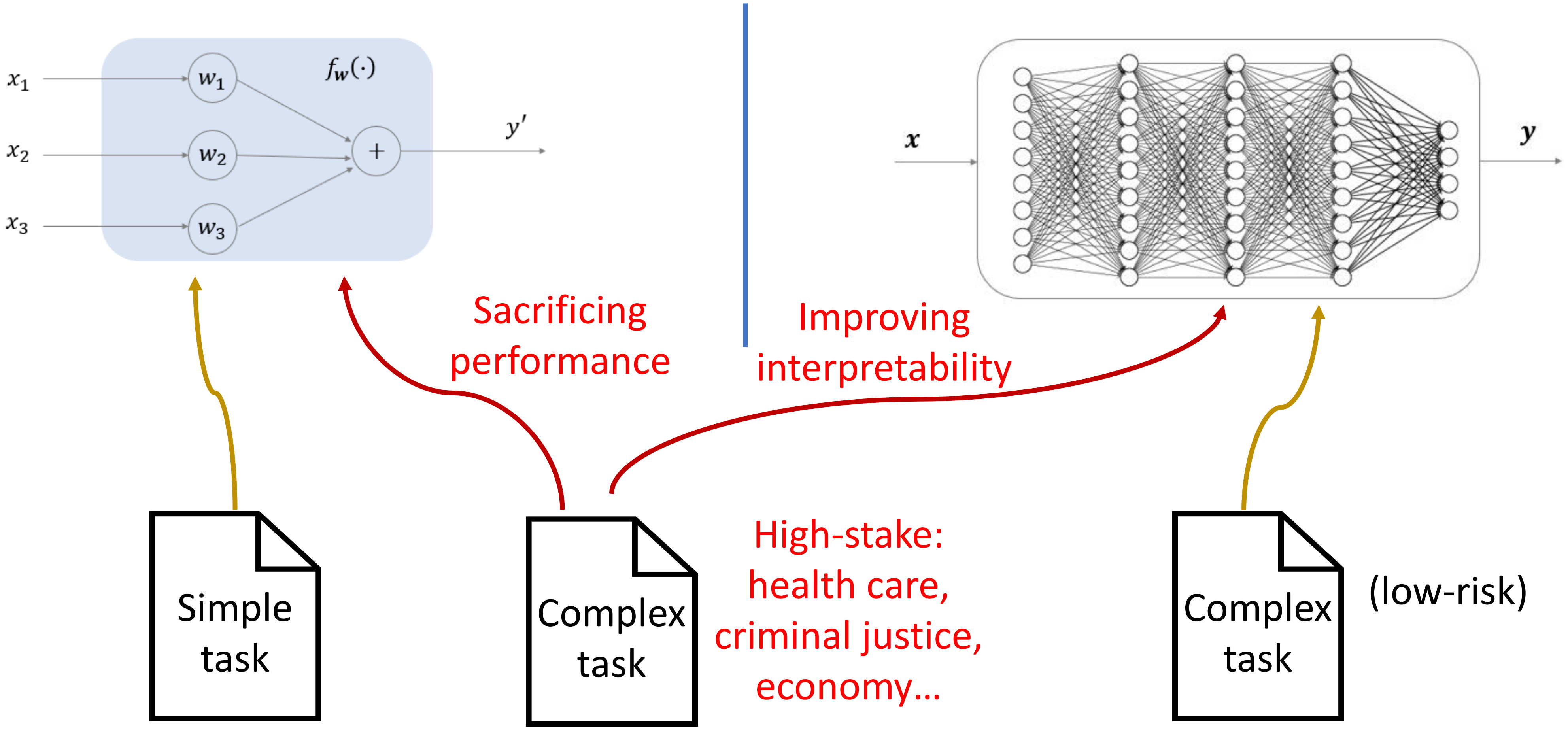


Complex task

(low-risk)

Interpretability

Depending on the specific task…



Interpretability

Building a machine learning model

- Performance (**What** the prediction is?)
- Interpretability (**Why** it came to the prediction?)

Desiderata of Interpretability

- Trust
- Causality
- Transferability
- Informativeness
- Fair and Ethical Decision Making

Desiderata of Interpretability

- Trust

  - What is trust?
  - Is it simply confidence that a model will perform well?

Desiderata of Interpretability

- Trust

  - What is trust?
  - Is it simply confidence that a model will perform well?
  - Trust can be defined subjectively

For example:

- ❑ People may trust an ML model if they are comfortable with relinquishing control to it

Desiderata of Interpretability

- Trust

- What is trust?
- Is it simply confidence that a model will perform well?
- Trust can be defined subjectively

For example:

- ❑ People may trust an ML model if they are comfortable with relinquishing control to it
- ❑ People may not only care about *how often* a model is right, but also *for which examples* it is right
    - If the model tends to make mistakes on only those kinds of inputs where humans also make mistakes ✅
    - If a model tends to make mistakes for inputs that humans classify accurately ❌

Desiderata of Interpretability

- Causality
    - Machine learning models are optimized to make associations
    - They are expected to infer properties of the natural world (e.g., smoking and lung cancer)
    - The associations learned by models may not reflect causal relationships
    - Interpreting ML models can help provide clues about the causal relationships between associated variables

Desiderata of Interpretability

- Transferability
  - Training and test data are randomly sampled from the same distribution
  - A model’s generalization error (transferability) is judged by the gap between its performance on training and test data
  - Humans exhibit a far richer capacity to generalize, transferring learned skills to unfamiliar situations

Desiderata of Interpretability

- Transferability

  - Training and test data are randomly sampled from the same distribution
  - A model's generalization error (transferability) is judged by the gap between its performance on training and test data
  - Humans exhibit a far richer capacity to generalize, transferring learned skills to unfamiliar situations
  - Interpretability provides insights on model's transferability

For example:

❑ A model trained to predict probability of death from pneumonia assigns *less risk* to patients if they also had asthma

Reason: The patients with asthma received more aggressive treatment

asthma $\xleftrightarrow{\text{Spurious correlation}}$ less risk of death from pneumonia

Desiderata of Interpretability

- Informativeness

  - A model conveys information via its outputs
  - Interpretability can provide additional information to human users

For example:

- [ ] A diagnosis model might provide intuition to a human decision maker by pointing to similar cases in support of a diagnostic decision

(skin cancer)

Desiderata of Interpretability

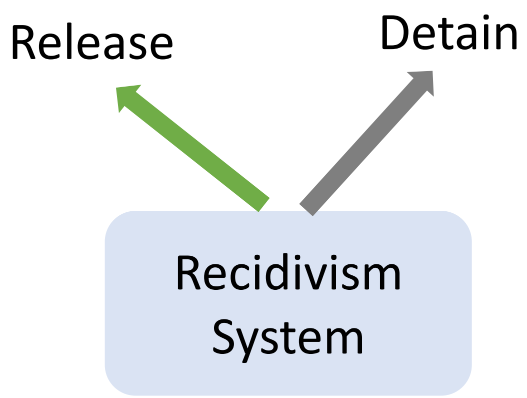
- Fair and Ethical Decision Making

Politicians, journalists, and researchers have expressed concern that interpretations must be produced for assessing whether decisions produced automatically by algorithms conform to ethical standards [Lipton, 2018]

Desiderata of Interpretability

- Fair and Ethical Decision Making

Politicians, journalists, and researchers have expressed concern that interpretations must be produced for assessing whether decisions produced automatically by algorithms conform to ethical standards [Lipton, 2018]



- Predictions do not discriminate on race?
- Accuracy or AUC (area under the curve) offer little assurance that ML-based decisions will behave acceptably
- Demands for fairness often lead to demands for interpretable models

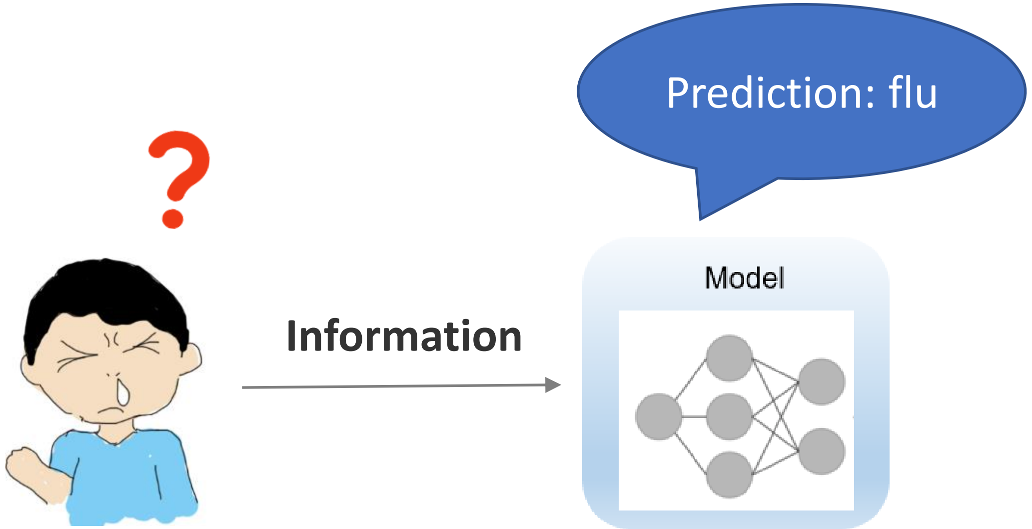
Interpretability

- What is interpretability?

- Why interpretability is important?

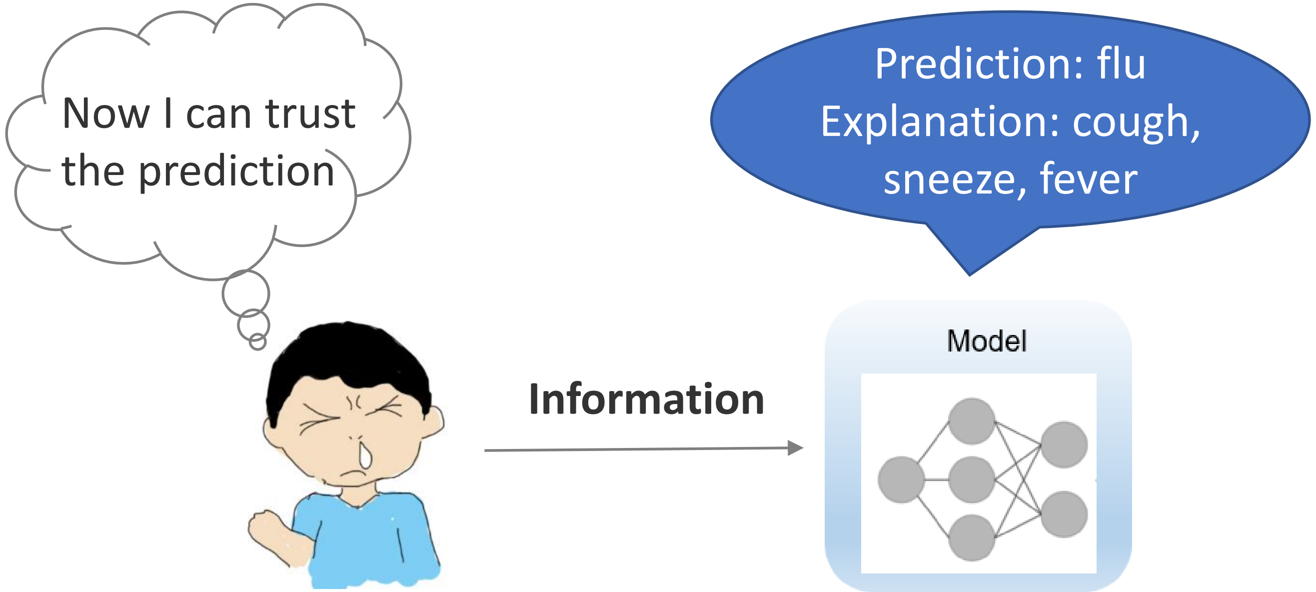
Why interpretability is important?

The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior



Why interpretability is important?

The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior



Why interpretability is important?

Interpretability reveals the knowledge captured by the model

A recommendation system trained on a large dataset

- It is impossible for human to understand the data
- It is hard to decide whether the model prediction is trustworthy

Why interpretability is important?

Interpretability reveals the knowledge captured by the model

You bought some paint

Recommendation: brush and ladder

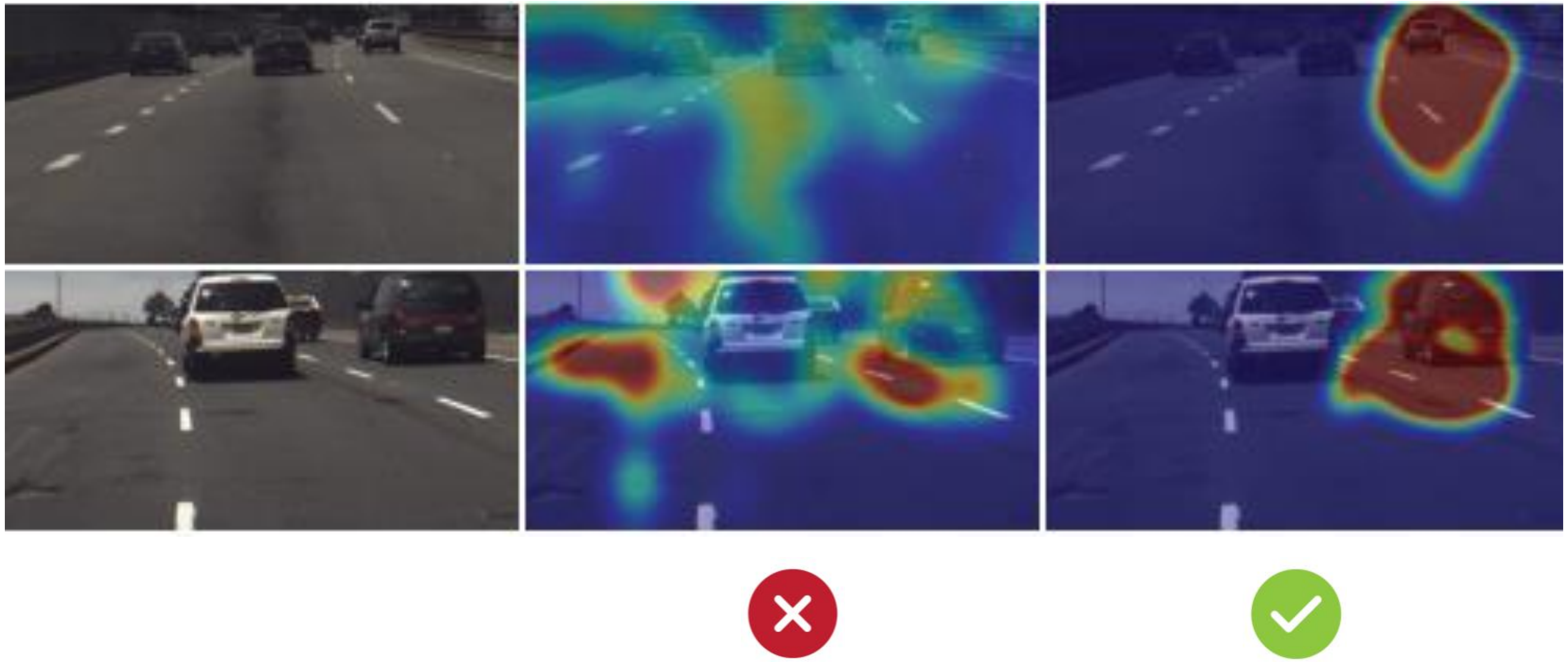
Interpretation: paint, brush and ladder are frequently bought together

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the trustworthiness of model predictions

Object recognition
Interpretation: highlighted pixels



[Kim et al., 2017]

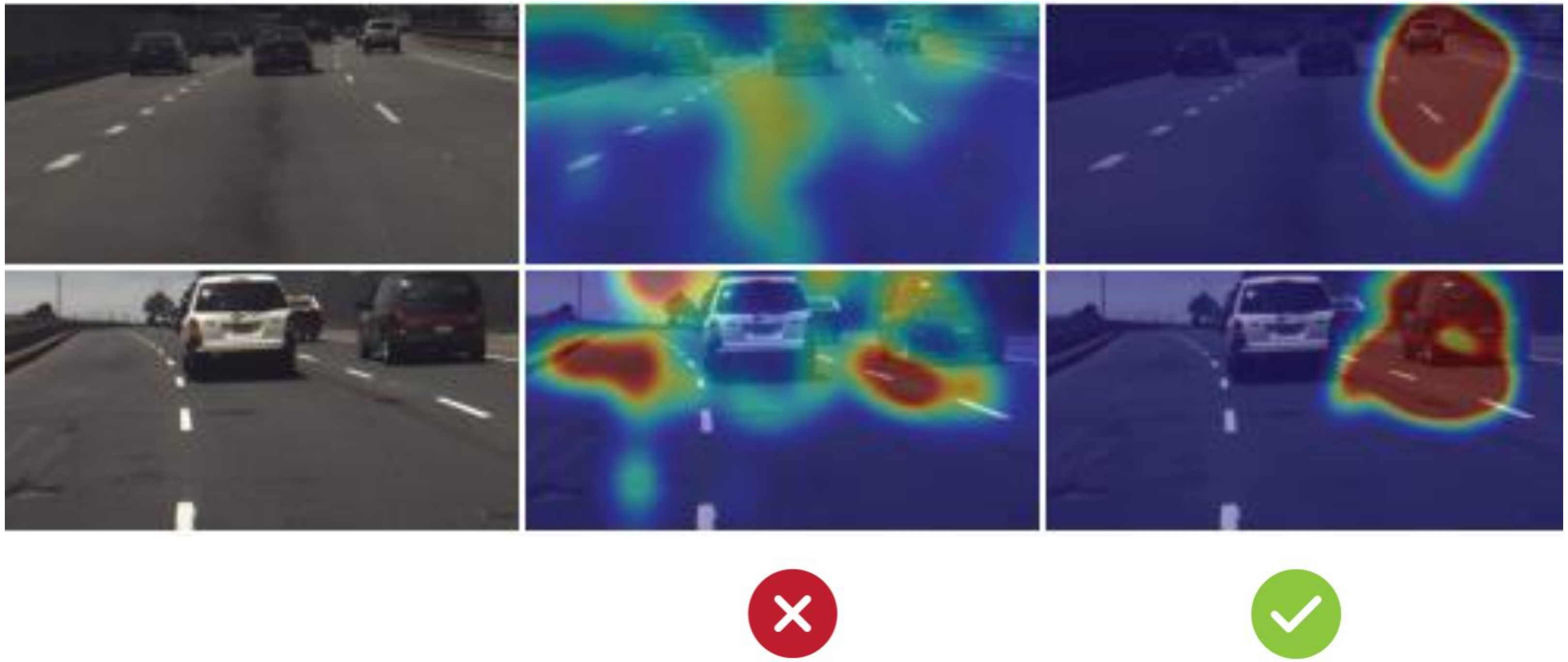
Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the trustworthiness of model predictions

Object recognition
Interpretation: highlighted pixels

Interpretations tell people whether the model makes correct predictions based on right reasons



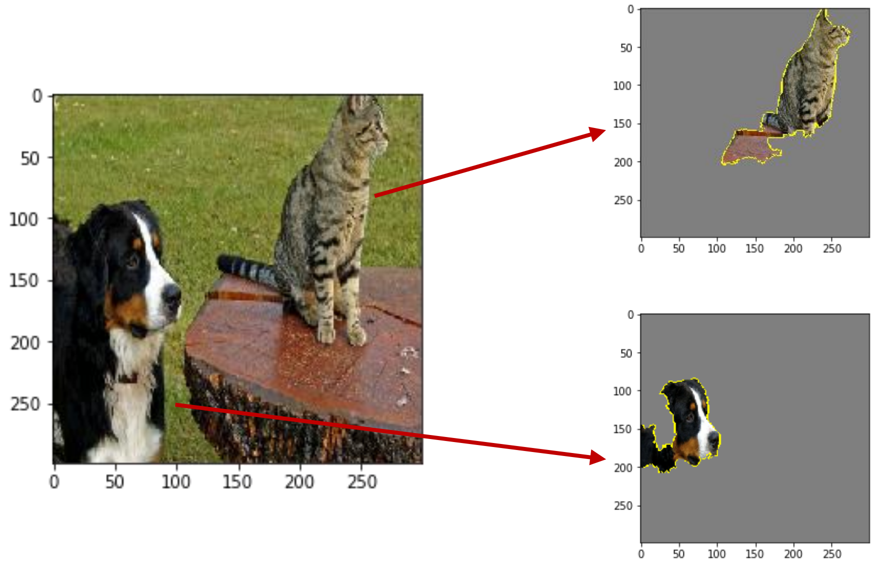
[Kim et al., 2017]

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the trustworthiness of model predictions

Object recognition
Interpretation: highlighted pixels



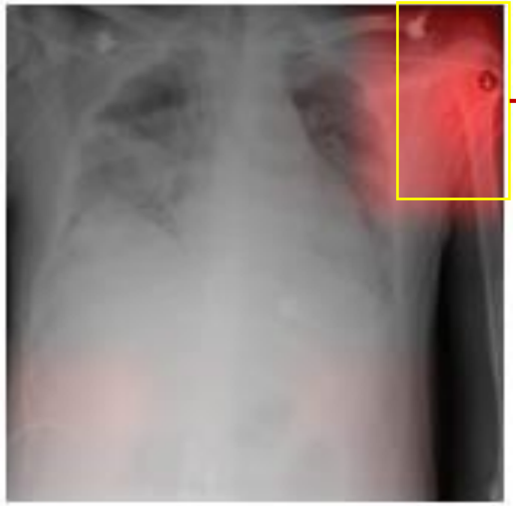
[Ribeiro et al., 2021]

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the trustworthiness of model predictions

Diagnose pneumonia
Interpretation: highlighted pixels



The model prediction is based on the hospital logo, not lung

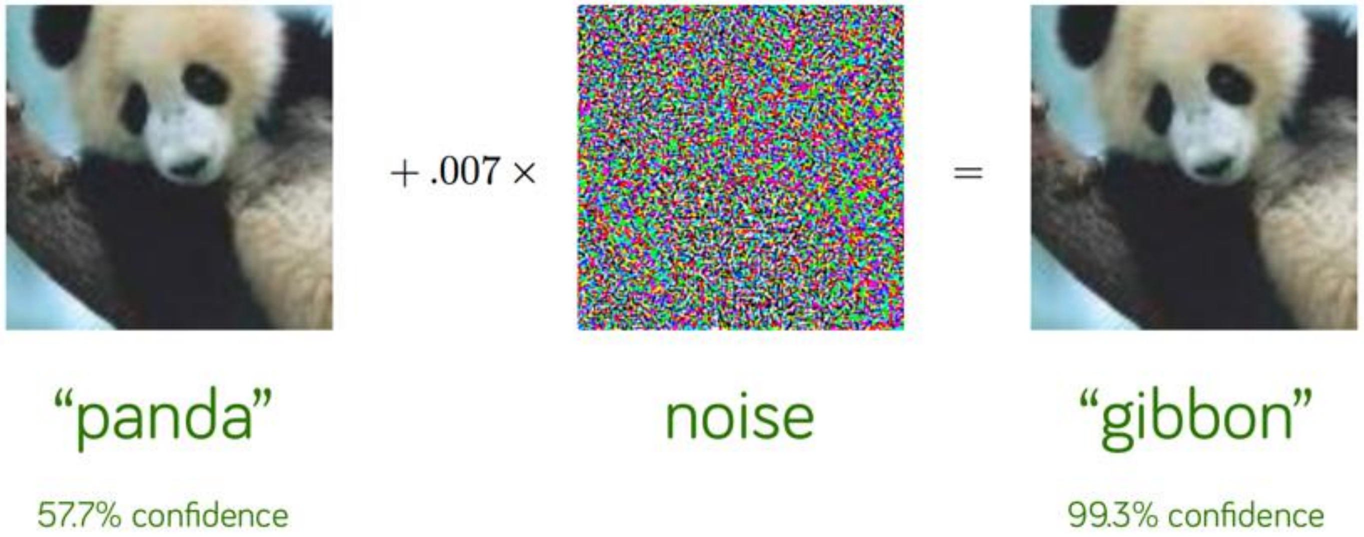
[Geirhos et al., 2021]

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the reliability of model predictions

Neural network models are vulnerable to adversarial attacks



[Goodfellow et al., 2015]

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the reliability of model predictions

Neural network models are vulnerable to adversarial attacks

Original prediction: **Entailment** **Confidence:** 99%

Premise: A runner wearing purple strives for the finish line

Hypothesis: A **runner** wants to head for the finish line

Adversarial prediction: **Contradiction** **Confidence:** 98%

Premise: A runner wearing purple strives for the finish line

Hypothesis: A **racer** wants to head for the finish line

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the reliability of model predictions

Interpretations for debugging

Prediction:
monkey

Prediction:
fish

Interpretation

[Boopathy et al., 2020]

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the reliability of model predictions

Interpretations for debugging

Original text	**Prediction**
an exceedingly clever piece of cinema	Positive
Adversarial text	
an shockingly proficient piece of cinema	Negative

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the reliability of model predictions

Interpretations for debugging

Original text	Prediction
an exceedingly clever piece of cinema	Positive

Adversarial text	
an shockingly proficient piece of cinema	Negative

Interpretation Pos Neg

an	exceedingly	clever	piece	of	cinema
an	shockingly	proficient	piece	of	cinema

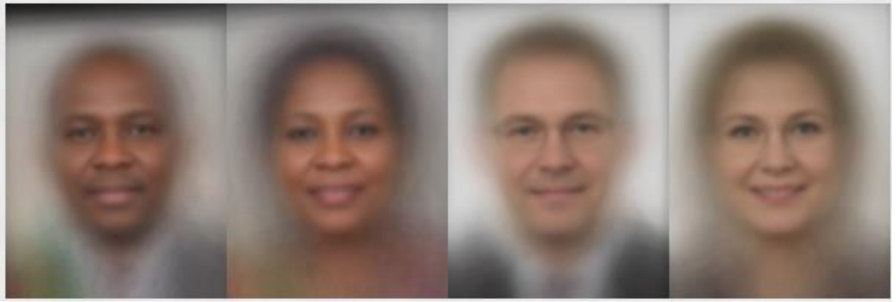
Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the fairness of model predictions

Machine learning models are making biased decisions

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Higher error rate on darker female

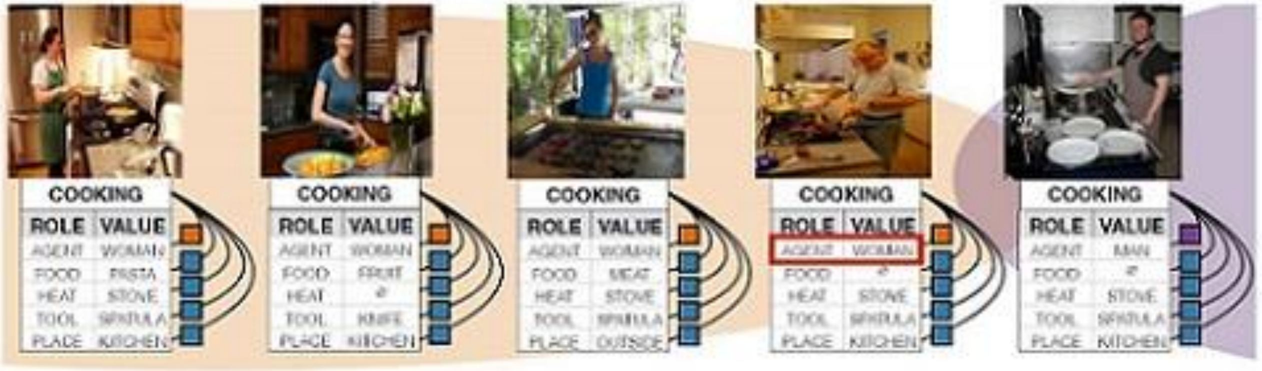
Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the fairness of model predictions

Machine learning can amplify bias

Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints



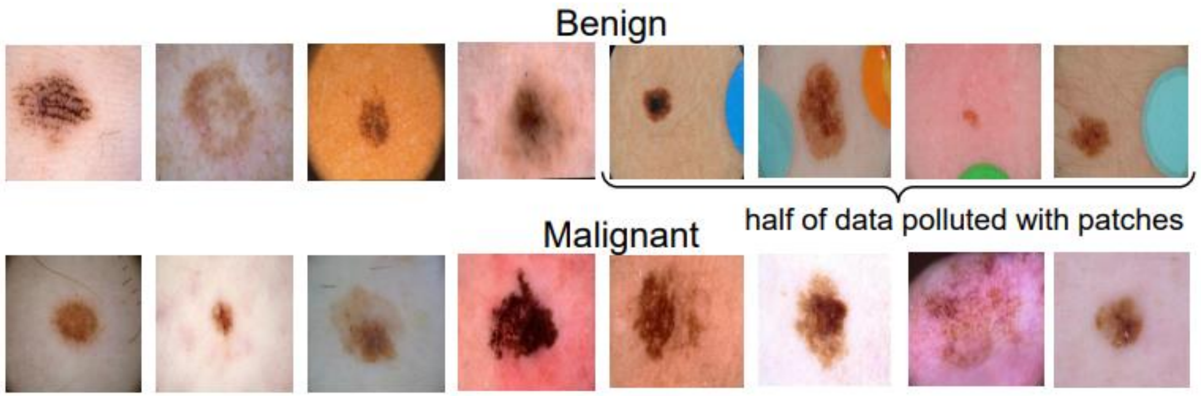
- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the fairness of model predictions

Detecting and mitigating bias via interpretations



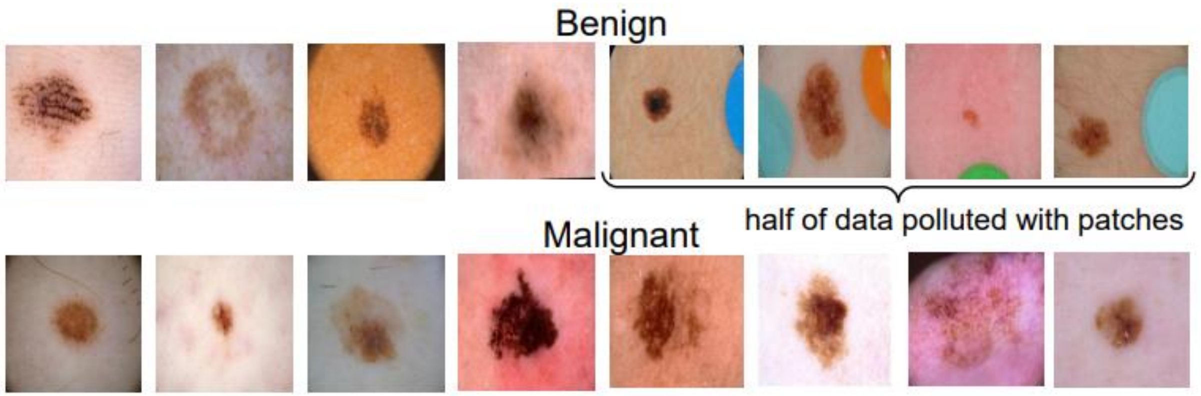
[Rieger et al., 2020]

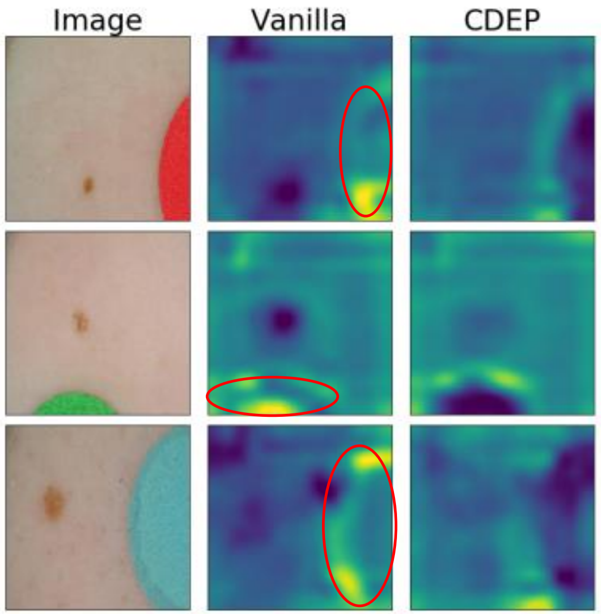
Why interpretability is important?

Interpretability for trustworthy AI

- Increasing the fairness of model predictions

Detecting and mitigating bias via interpretations





[Rieger et al., 2020]

Summary

- To solve complex problems, machine learning models are becoming bigger and sophisticated (uninterpretable)
- Model interpretability is an important criterion beyond performance
- Improving model interpretability
    - Increasing social acceptance
    - Building trustworthy AI (trustworthiness, reliability, fairness)
    - Debugging and developing

Evaluation

- Faithfulness to model

  How accurately an interpretation reflects the true reasoning process of the model

- Plausibility to humans

  How convincing the interpretation is to humans

[Jacovi, 2020]

Evaluation

- Faithfulness to model

  How accurately an interpretation reflects the true reasoning process of the model

- Plausibility to humans

  How convincing the interpretation is to humans

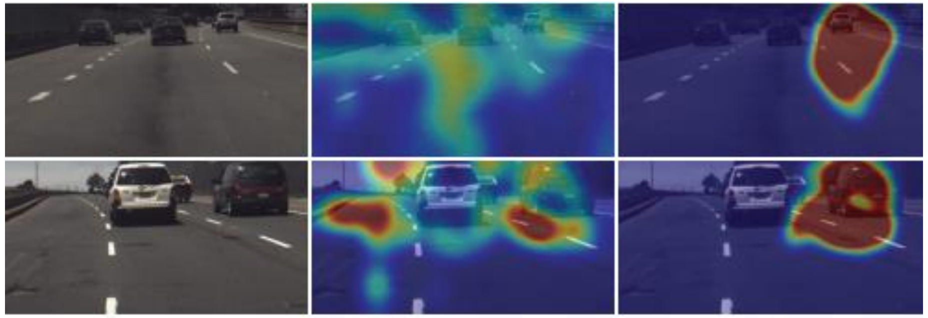
- Generally, we cannot satisfy both criteria because of the gap between model reasoning and human understanding
- Faithfulness is the primary criterion

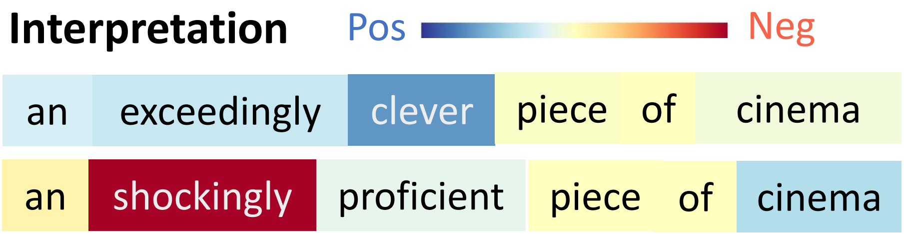
[Jacovi, 2020]

Research topics

- Post-hoc explanations (Week 4-6)

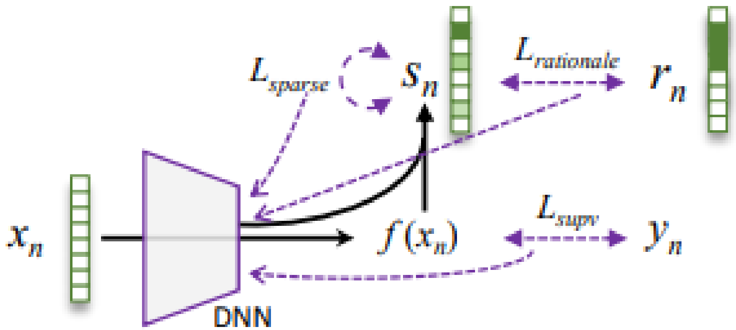
  - In the inference stage
  - Explaining well-trained models’ predictions
  - Inferring model decision making (perturbation, gradients, attention, interaction)





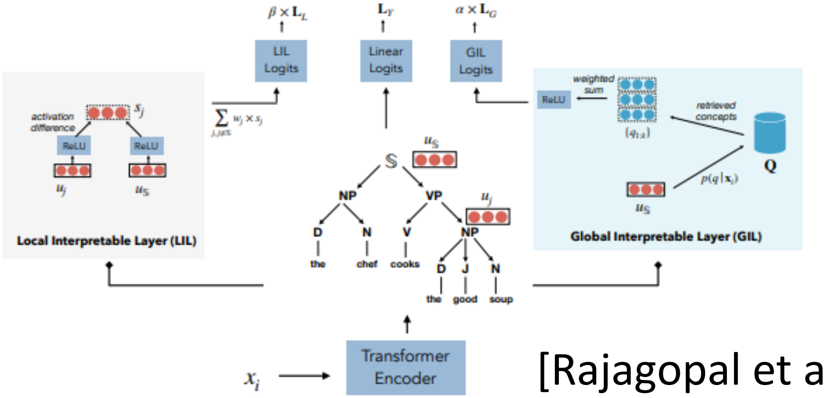
Research topics

- Improving neural network intrinsic interpretability (Week 7)
  - In the training stage
  - Do not change model architecture
  - Manipulating model prediction behavior (to be more interpretable)



[Du et al., 2019]

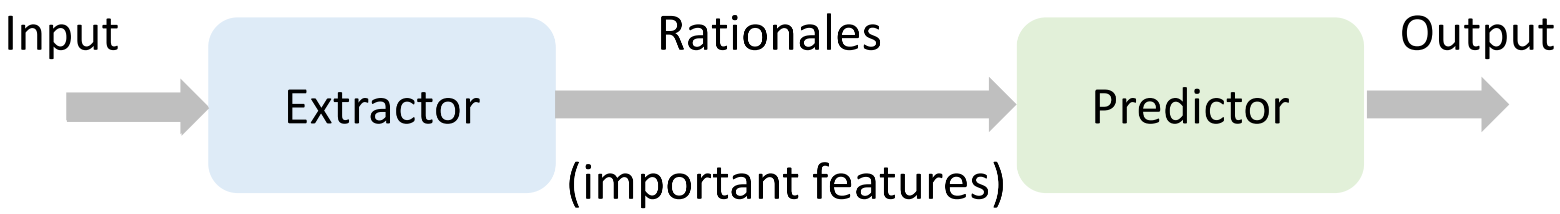
- Building interpretable neural network models (Week 9)
  - Model engineering and expert knowledge
  - Designing self-interpretable models



[Rajagopal et al., 2021]

Research topics

- Rationalized neural networks (Week 10)

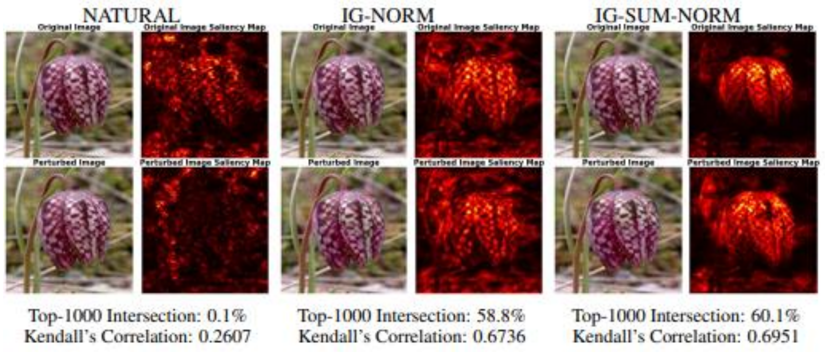


- Interpretation and human understanding (Week 11)

  - Interpretation can help human understanding?
  - Interpretation may fool human decision?
  - How humans and models interact via interpretations?

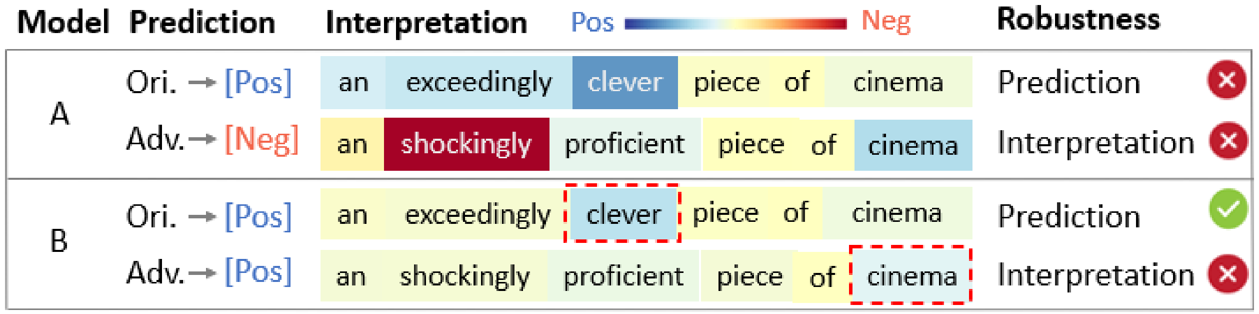
Research topics

- Robust interpretations (Week 12)
  - Robustness of interpretations to input perturbations
  - Robustness of interpretations to model manipulations
  - Risks of interpretation vulnerability



[Chen et al., 2019]

- Connections with model performance, robustness, fairness (Week 13)



[Chen et al., 2022]

Reference

- Christoph Molnar, Interpretable Machine Learning, 2021
- Murdoch, W. James, et al. "Definitions, methods, and applications in interpretable machine learning." *Proceedings of the National Academy of Sciences* 116.44 (2019): 22071-22080.
- 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).
- Kim, Jinkyu, and John Canny. "Interpretable learning for self-driving cars by visualizing causal attention." *Proceedings of the IEEE international conference on computer vision*. 2017.
- Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673.
- 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.
- Boopathy, Akhilan, et al. "Proper network interpretability helps adversarial robustness in classification." *International Conference on Machine Learning*. PMLR, 2020.
- Rieger, Laura, et al. "Interpretations are useful: penalizing explanations to align neural networks with prior knowledge." *International Conference on Machine Learning*. PMLR, 2020.

Reference

- Du, Mengnan, et al. "Learning credible deep neural networks with rationale regularization." *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019.
- Rajagopal, Dheeraj, et al. "SelfExplain: A Self-Explaining Architecture for Neural Text Classifiers." *arXiv preprint arXiv:2103.12279* (2021).
- Jain, Sarthak, et al. "Learning to faithfully rationalize by construction." *arXiv preprint arXiv:2005.00115* (2020).
- Hase, Peter, and Mohit Bansal. "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?." *arXiv preprint arXiv:2005.01831* (2020).
- Chen, Jiefeng, et al. "Robust attribution regularization." *arXiv preprint arXiv:1905.09957* (2019).
- Chen, Hanjie, and Ji, Yangfeng. "Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation." The 36th AAAI Conference on Artificial Intelligence (2022).
- Lipton, Zachary C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue* 16.3 (2018): 31-57.
- Jacovi, Alon, and Yoav Goldberg. "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?." *arXiv preprint arXiv:2004.03685* (2020).