UNIVERSITY OF GEORGIA

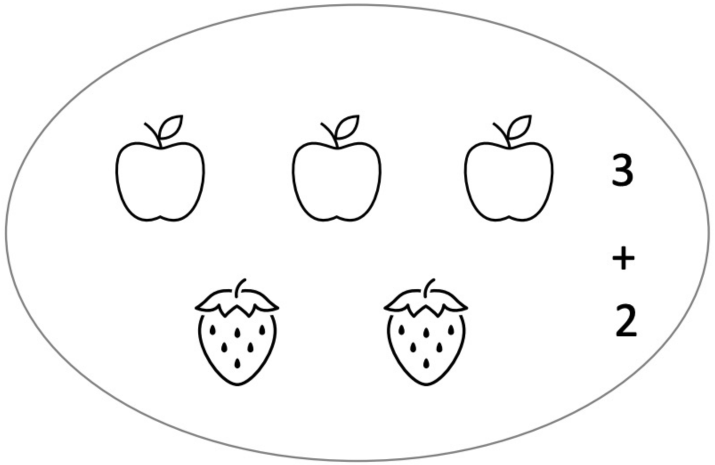
CSCI 4360/6360: Data Science II

Machine Learning Interpretation – Preliminaries

Ninghao Liu
Assistant Professor
School of Computing
University of Georgia

Many contents are adopted from “Interpretable Machine Learning” by Hanjie Chen, Yangfeng Ji

Machine Learning



How many fruits?

5

Features

x_1: the number of apples

x_2: the number of strawberries

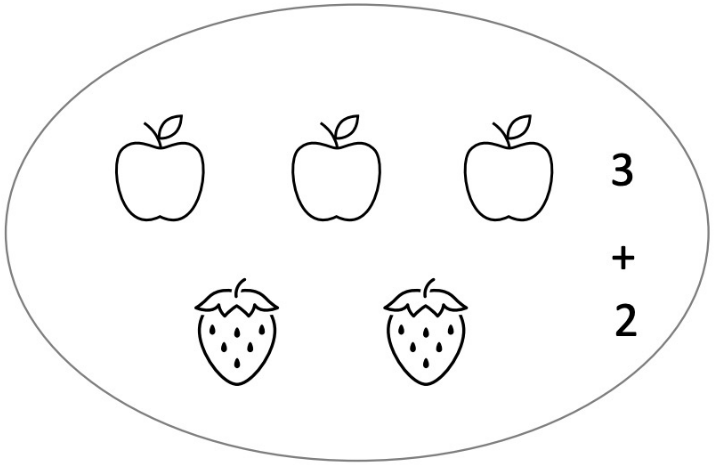
Rule

$x_1 + x_2$

Output

y: the total number of fruits

Machine Learning



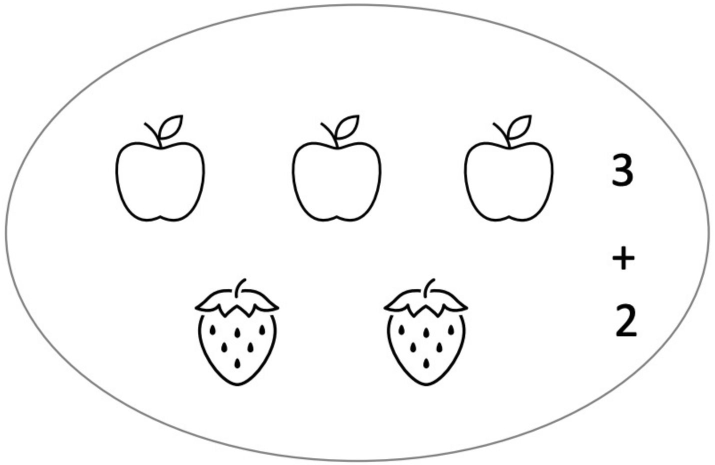
How many fruits?

5

What is the contribution of each feature?

Features	**Rule**	**Output**
x_1: the number of apples	$x_1 + x_2$ →	y: the total number of fruits
x_2: the number of strawberries		

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

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	$\xrightarrow{0.6x_1 + 0.3x_2 + 0.1x_3}$ house size, location, and floor type account for 60%, 30%, 10% respectively	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	$\xrightarrow{0.6x_1 + 0.3x_2 + 0.1x_3}$ house size, location, and floor type account for 60%, 30%, 10% respectively	y: house value
$x_1 = 100,$ $x_2 = 300,$ $x_3 = 200$		$y = 170$

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	$\xrightarrow{?}$	y: house value
x_2: location		
x_3: floor type		

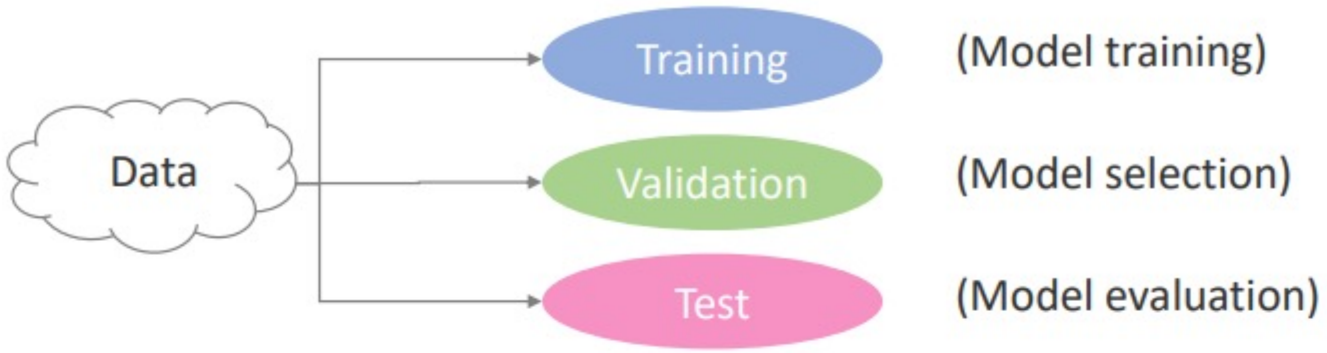
What if we do not know the rule?

Machine Learning

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

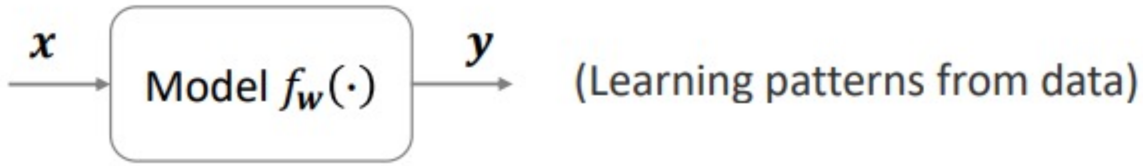
1. Collecting data

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



2. Training a machine learning model

$$f_w(\cdot)$$



3. Testing the model

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

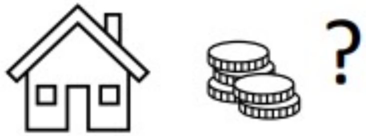
Machine Learning

Learn a rule from past house sales

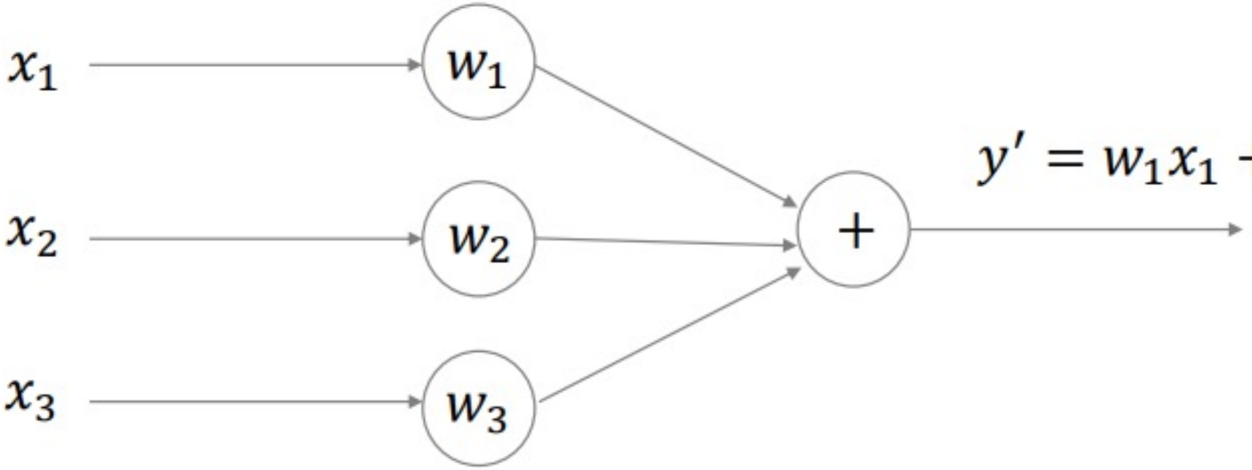
Features	**Machine Learning**	**Output**
$\{\boldsymbol{x} = (x_1, x_2, x_3)\}$	$f_{\boldsymbol{w}}(\cdot)$	$\{y\}$

Machine Learning

Learn a rule from past house sales



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

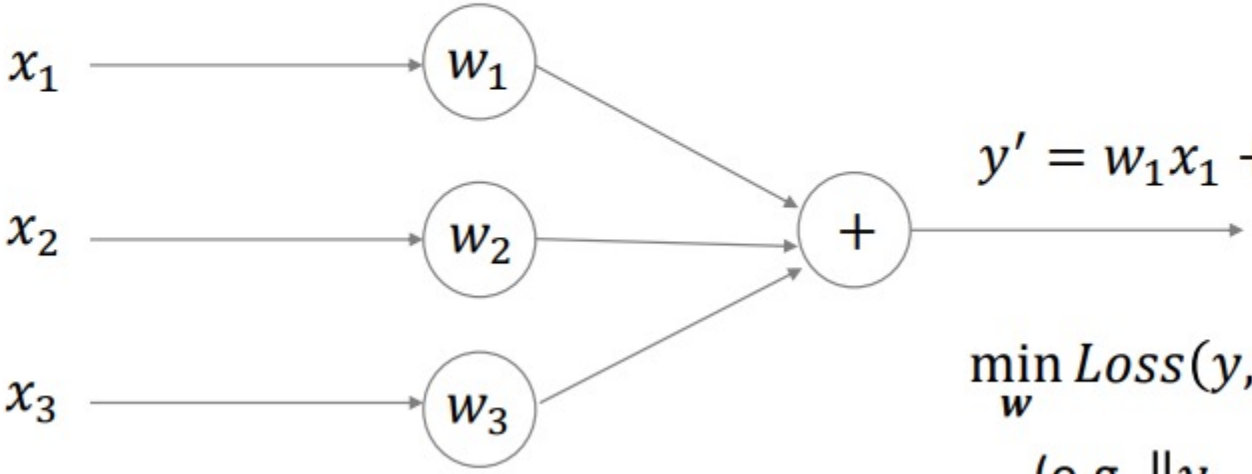


$$y' = w_1 x_1 + w_2 x_2 + w_3 x_3$$

Machine Learning

Learn a rule from past house sales

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



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

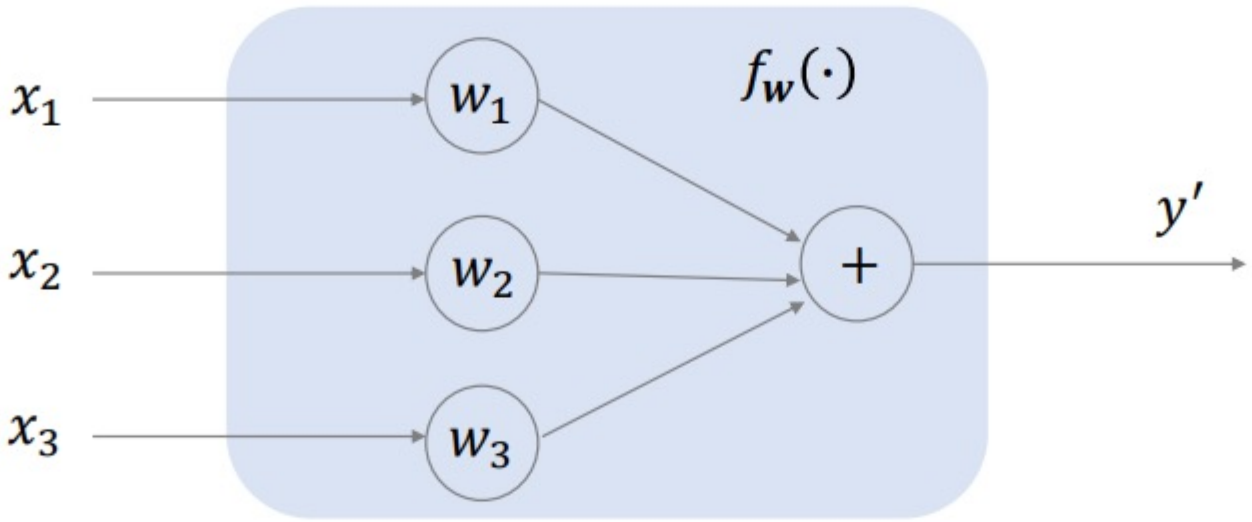
$$\min_{\boldsymbol{w}} Loss(y, y')$$

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

Machine Learning

Predict the house value via the learned machine learning model

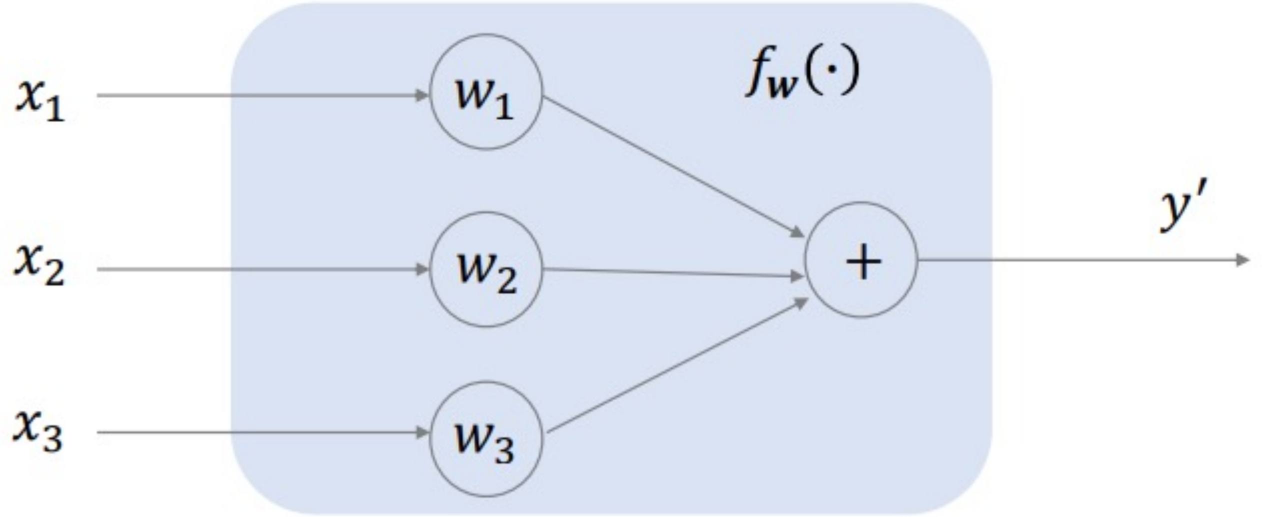
Machine Learning model



Machine Learning

Predict the house value via the learned machine learning model

Machine Learning model → Interpretable

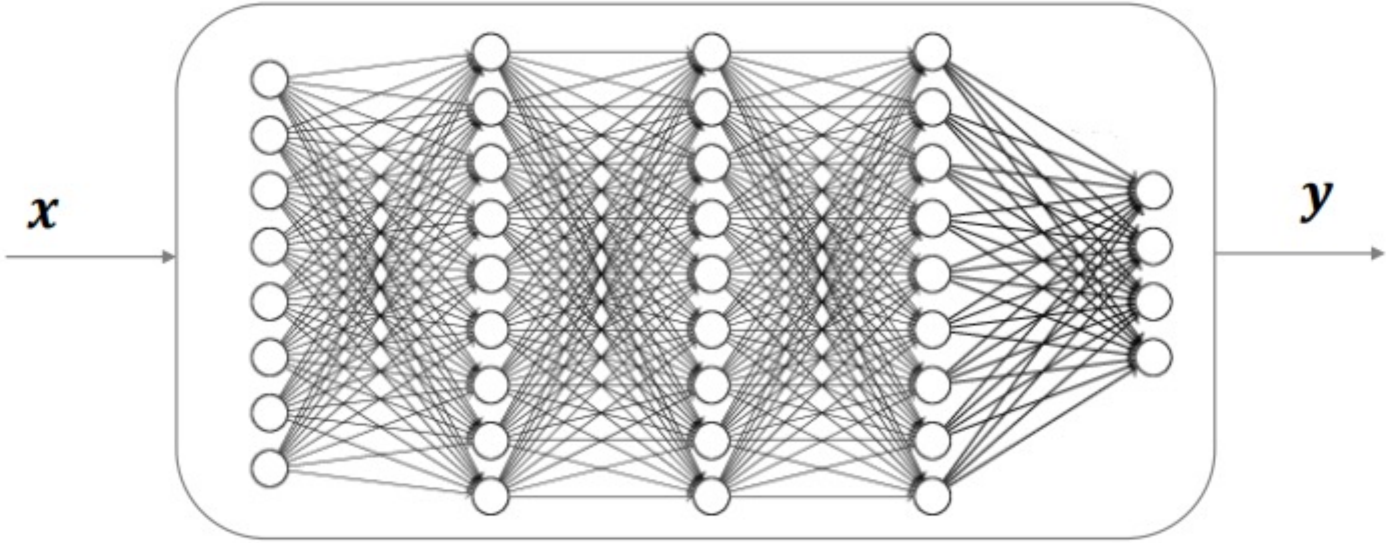


Contributions:
$x_1 : w_1 x_1$
$x_2 : w_2 x_2$
$x_3 : w_3 x_3$

Machine Learning - Issue

In reality, features and relationships can be more complex

Machine Learning model → Uninterpretable

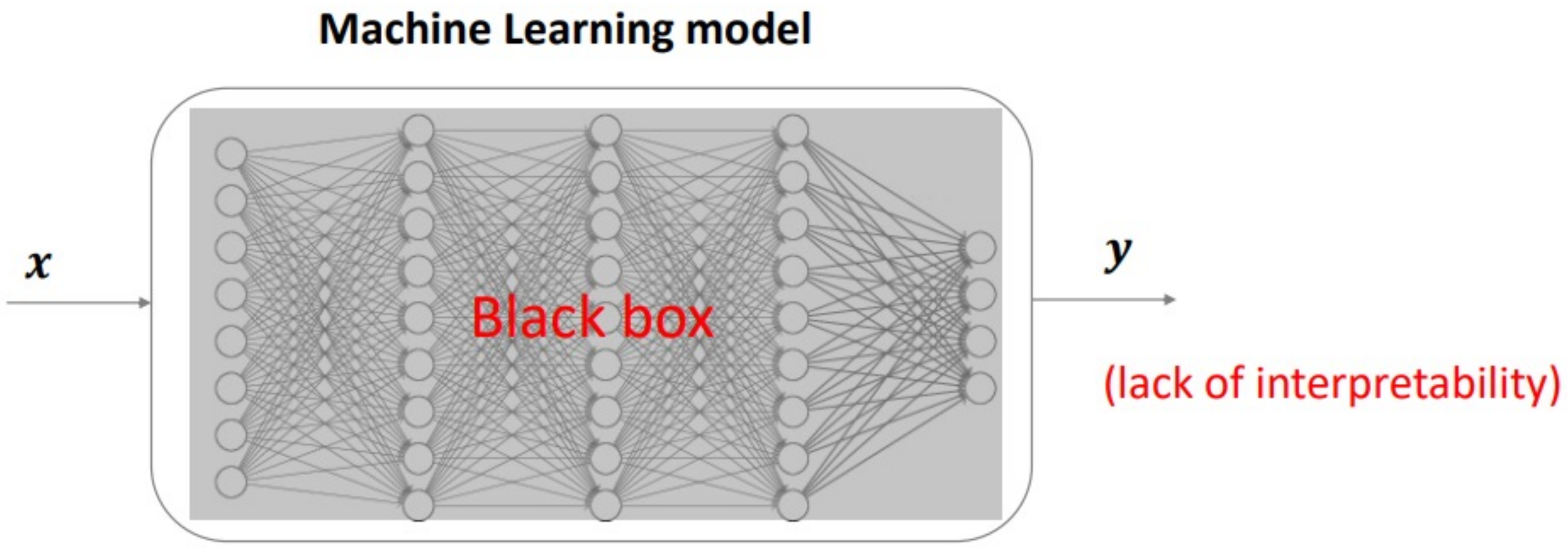


(nonlinear and complex transformations)

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

Machine Learning - Issue

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



Interpretability

- What is interpretability?
- Why is interpretability important?

Interpretability

- **What is interpretability?**
- Why is interpretability 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]

At this time, it is good for us to define interpretation as estimating the contribution of each feature to the final prediction.

Interpretability

- Trust
- Informativeness
- Causality

Trust

- Trust

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

Trust

- 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

Trust

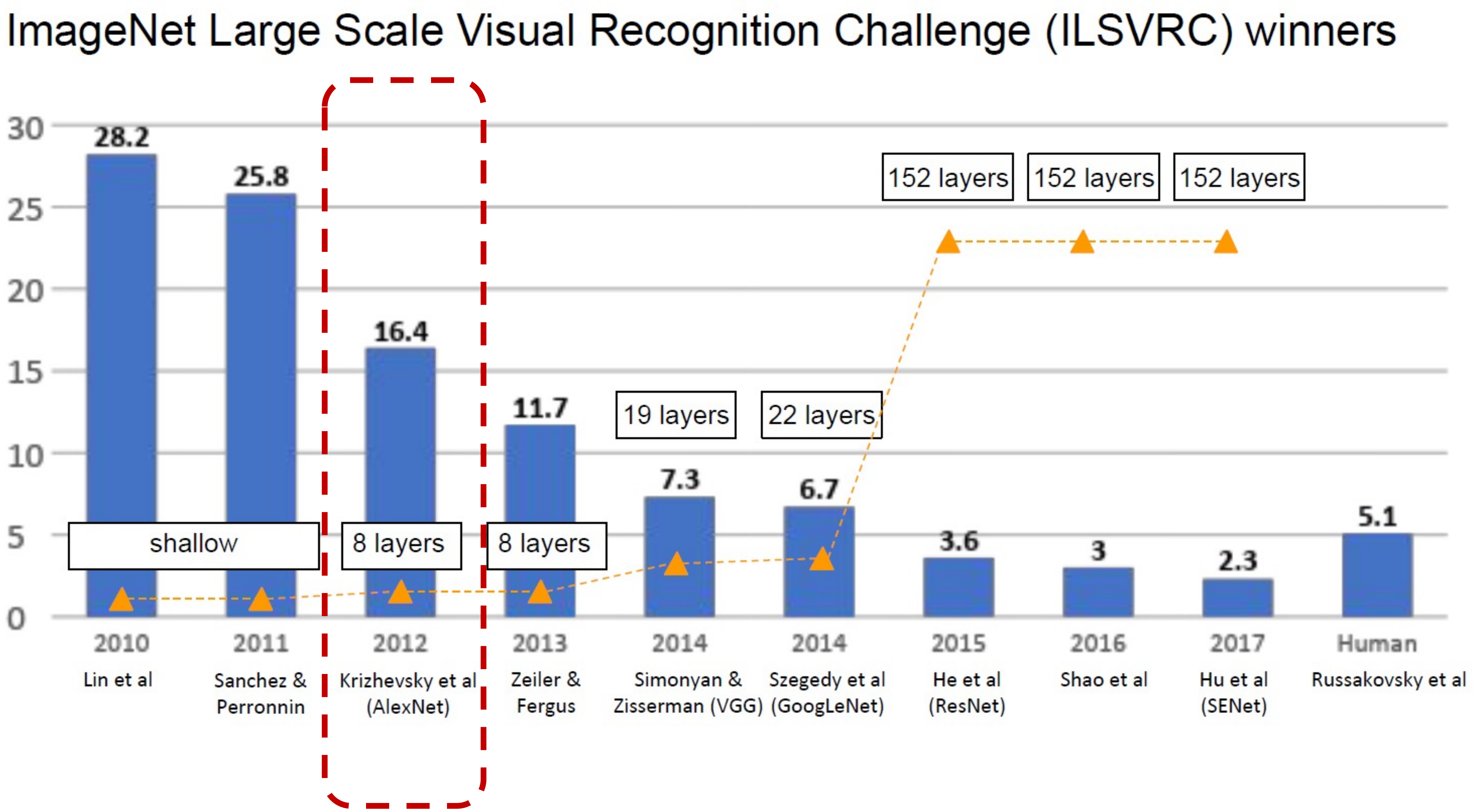
- Trust

- 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 ✗

Trust


ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners
30
25
20
15
10
5
0
28.2
25.8
16.4
11.7
7.3
6.7
3.6
3
2.3
5.1
shallow
8 layers
8 layers
19 layers
22 layers
152 layers
152 layers
152 layers
2010
Lin et al
2011
Sanchez & Perronnin
2012
Krizhevsky et al (AlexNet)
2013
Zeiler & Fergus
2014
Simonyan & Zisserman (VGG)
2014
Szegedy et al (GoogLeNet)
2015
He et al (ResNet)
2016
Shao et al
2017
Hu et al (SENet)
Human
Russakovsky et al

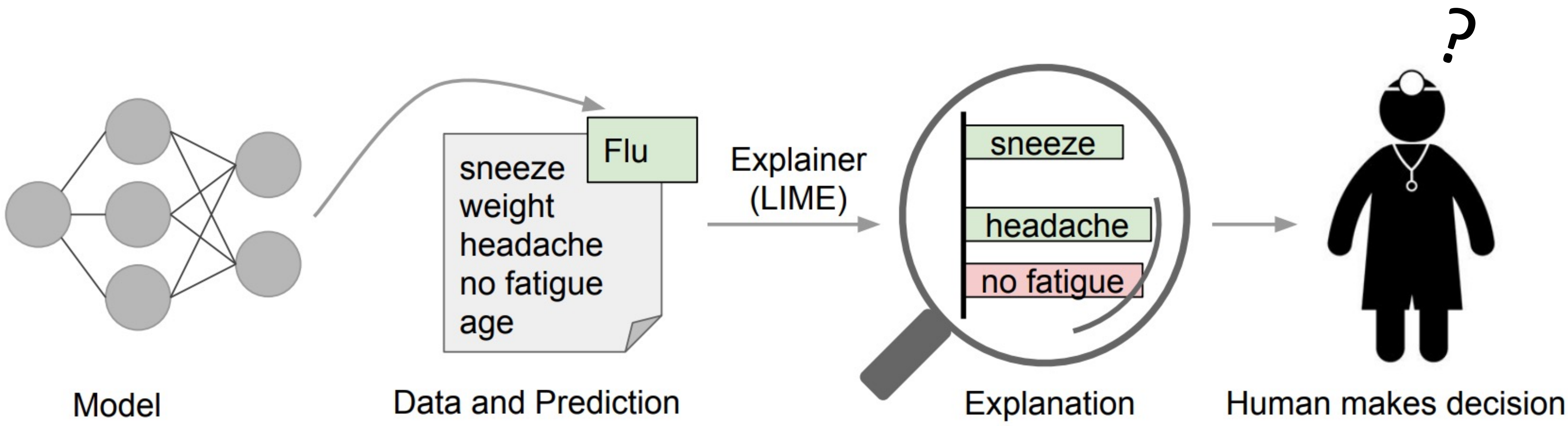
Trust

Learning data mining involves understanding the concepts, techniques, and tools used to extract valuable insights and patterns from large datasets. Here's a step-by-step guide to help you get started:

1. **Understand the Basics:**
   Begin by grasping the fundamental concepts of data mining, such as data preprocessing, data transformation, pattern recognition, and statistical analysis. A solid foundation will make advanced topics easier to comprehend.
2. **Learn about Data and Databases:**
   Understand the different types of data (structured, unstructured, semi-structured) and various database systems (relational, NoSQL) where data is stored. This knowledge is crucial for data extraction and manipulation.
3. **Master Programming Languages:**
   Learn programming languages commonly used in data mining, such as Python or R. These languages offer libraries and tools for data manipulation, analysis, and visualization.
4. **Study Statistics and Probability:**
   Data mining heavily relies on statistical techniques to identify patterns. Familiarize yourself with concepts like probability distributions, regression, clustering, and hypothesis testing.
5. **Explore Machine Learning:**

Trust

- Different users expect different explanation.



Informativeness

- Informativeness

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

  For example:

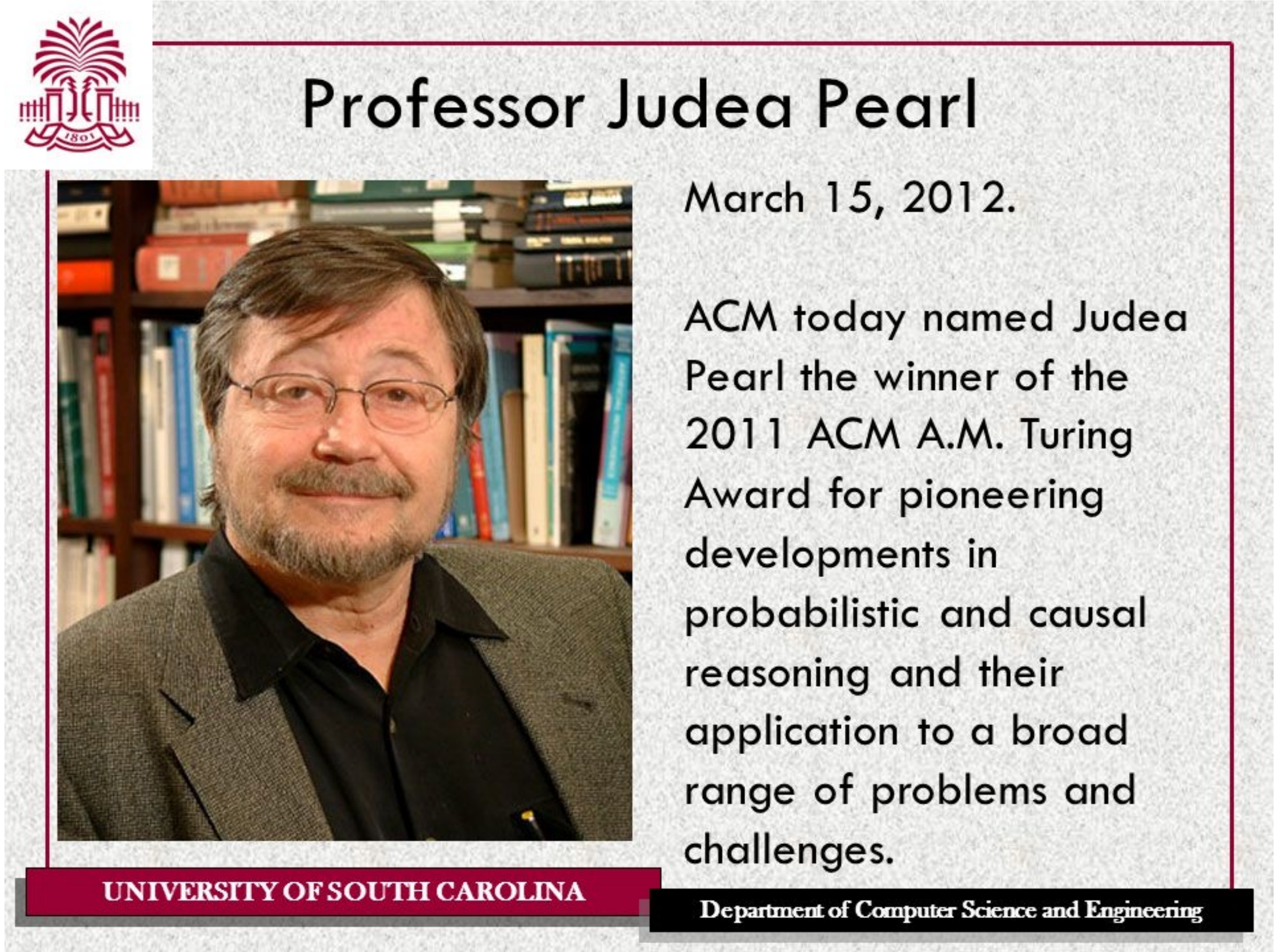
  - [ ] A diagnosis model might provide intuition to a human decision maker by pointing to similar cases in support of a diagnostic decision

(skin cancer)

Causality

- 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

Causality

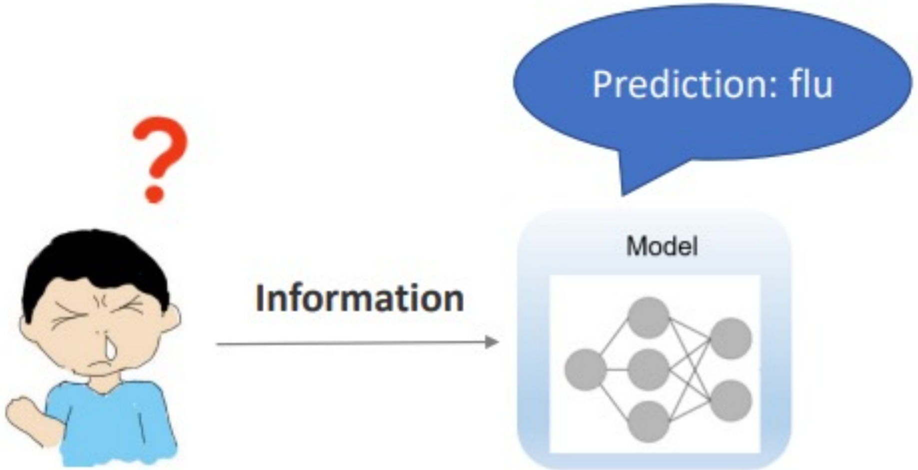

Professor Judea Pearl
March 15, 2012.
ACM today named Judea Pearl the winner of the 2011 ACM A.M. Turing Award for pioneering developments in probabilistic and causal reasoning and their application to a broad range of problems and challenges.
UNIVERSITY OF SOUTH CAROLINA
Department of Computer Science and Engineering

Interpretability

- What is interpretability?
- **Why is interpretability important?**

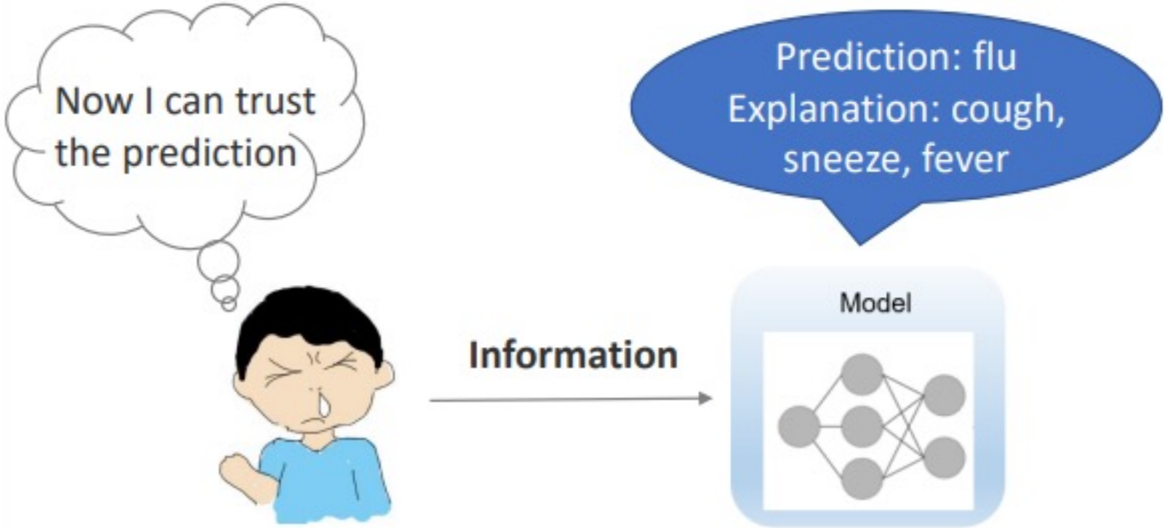
Interpretability - Why

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



Interpretability - Why

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



Interpretability - Why

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

Interpretability - Why

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

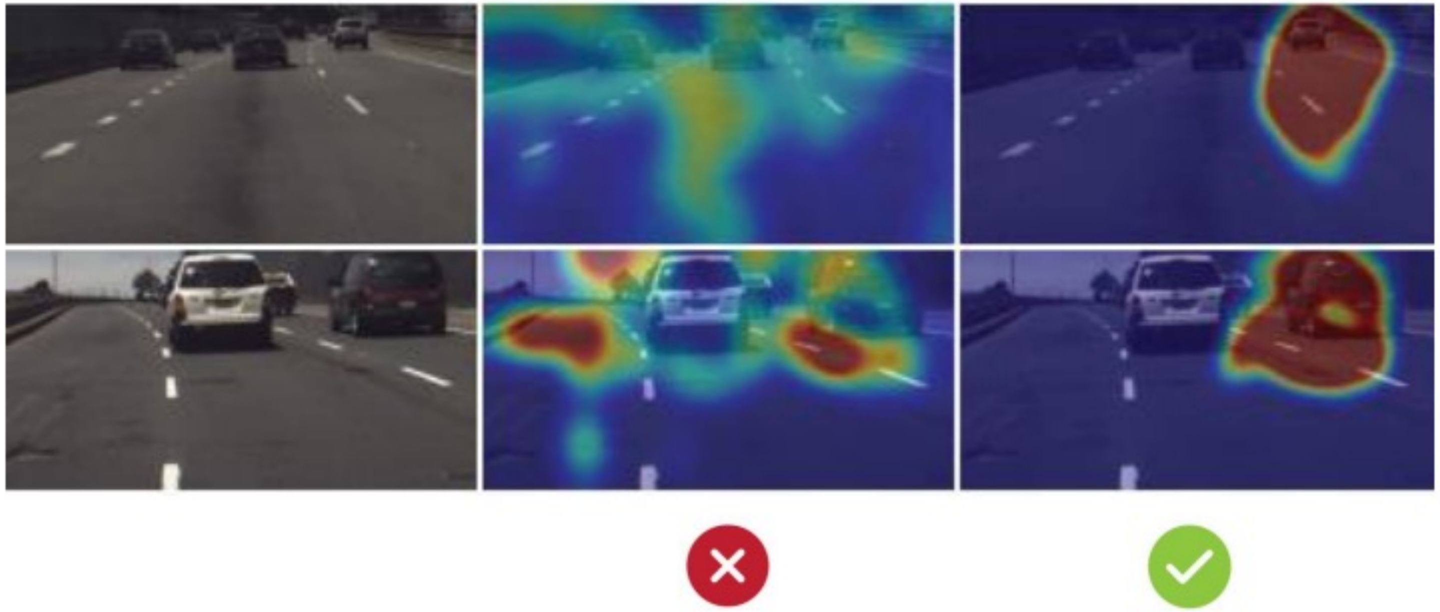
Interpretability - Why

Interpretability for trustworthy AI

- Increasing the trustworthiness of model predictions

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

Object recognition
Interpretation: highlighted pixels



[Kim et al., 2017]

Interpretability - Why

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

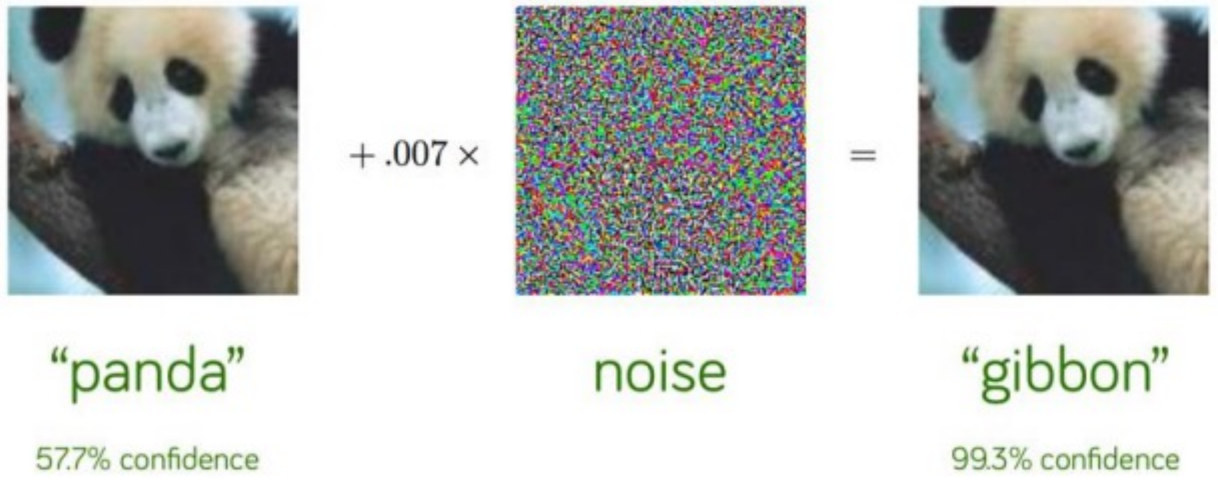
[Geirhos et al., 2021]

Interpretability - Why

Interpretability for trustworthy AI

- Increasing the reliability of model predictions

Neural network models are vulnerable to adversarial attacks



[Goodfellow et al., 2015]

Interpretability - 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