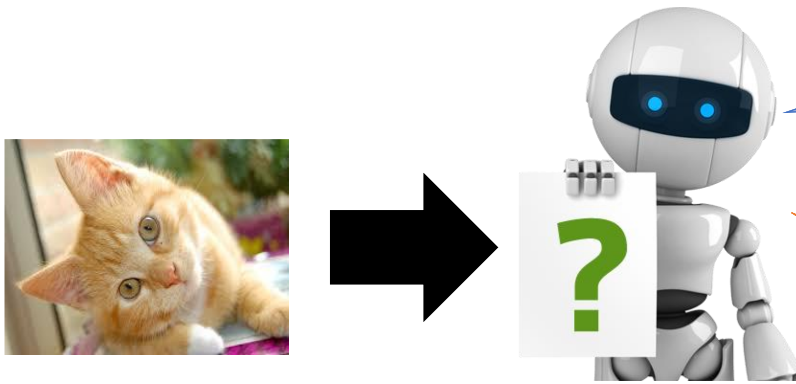
EXPLAINABLE MACHINE LEARNING

Hung-yi Lee 李宏毅

Explainable/Interpretable ML



This is a "cat" .

Because ...

Classifier

Local Explanation

Why do you think *this image* is a cat?

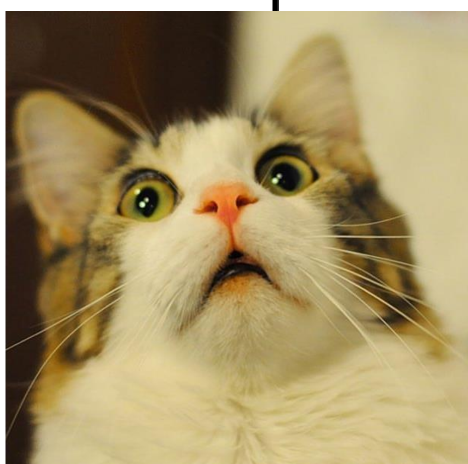
Global Explanation

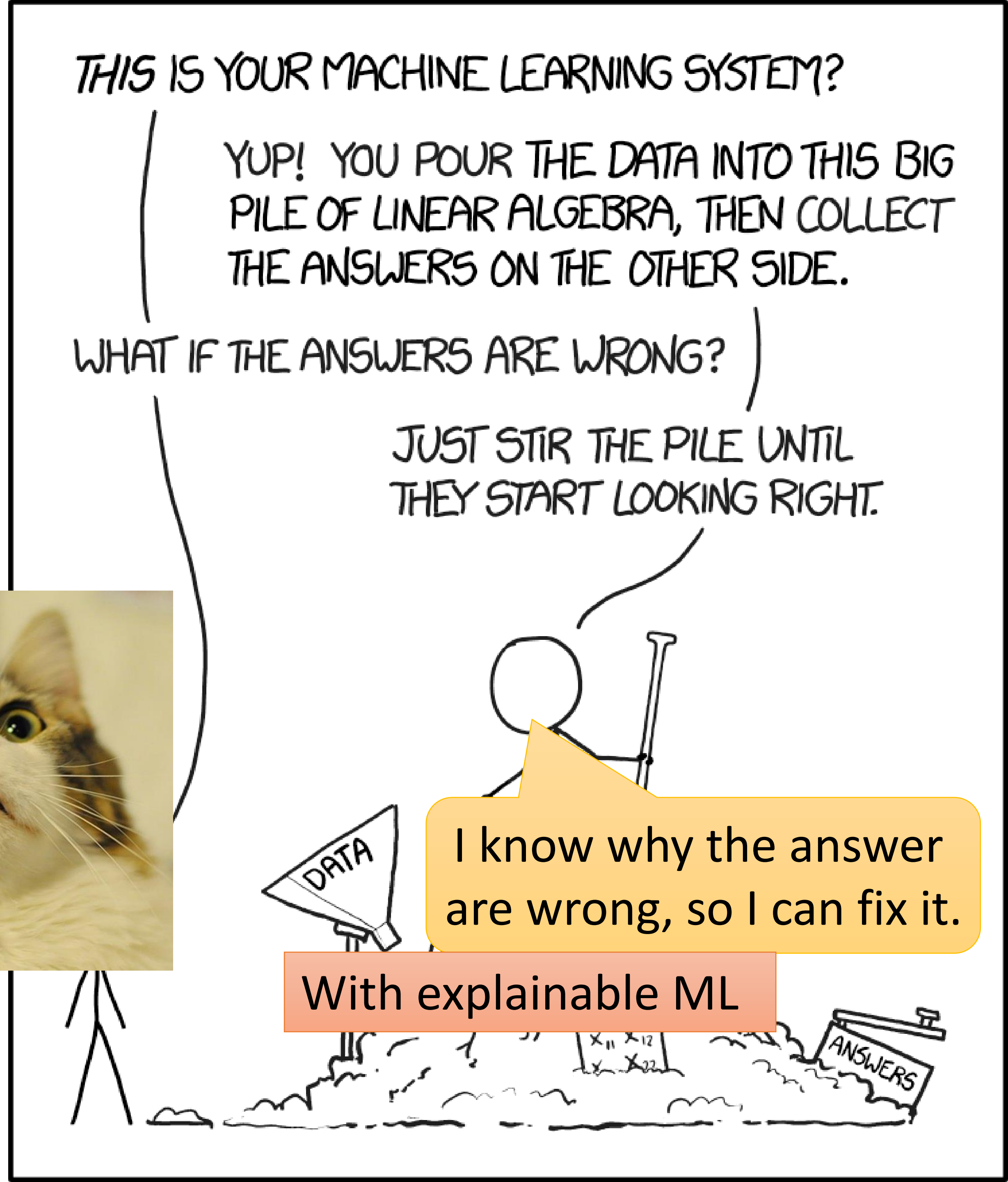
What do you think a "cat" looks like?

Why we need Explainable ML?

- 用機器來協助判斷履歷
  - 具體能力？還是性別？
- 用機器來協助判斷犯人是否可以假釋
  - 具體事證？還是膚色？
- 金融相關的決策常常依法需要提供理由
  - 為什麼拒絕了這個人的貸款？
- 模型診斷：到底機器學到了甚麼
  - 不能只看正確率嗎？想想神馬漢斯的故事

We can improve ML model based on explanation.





https://www.explainxkcd.com/wiki/index.php/1838:_Machine_Learning

My Point of View

- Goal of ML Explanation $\neq$ you completely know how the ML model work
  - Human brain is also a Black Box!
  - People don't trust network because it is Black Box, but you trust the decision of human!
- Goal of ML Explanation is (my point of view)

Make people (your customers, your boss, yourself) comfortable.

讓人覺得爽

Personalized explanation in the future

Interpretable v.s. Powerful

- Some models are intrinsically interpretable.
  - For example, linear model (from weights, you know the importance of features)
  - But ……. not very powerful.
- Deep network is difficult to interpretable.
  - Deep network is a black box.

    Because deep network is a black box, we don’t use it.

    削足適履 ☹

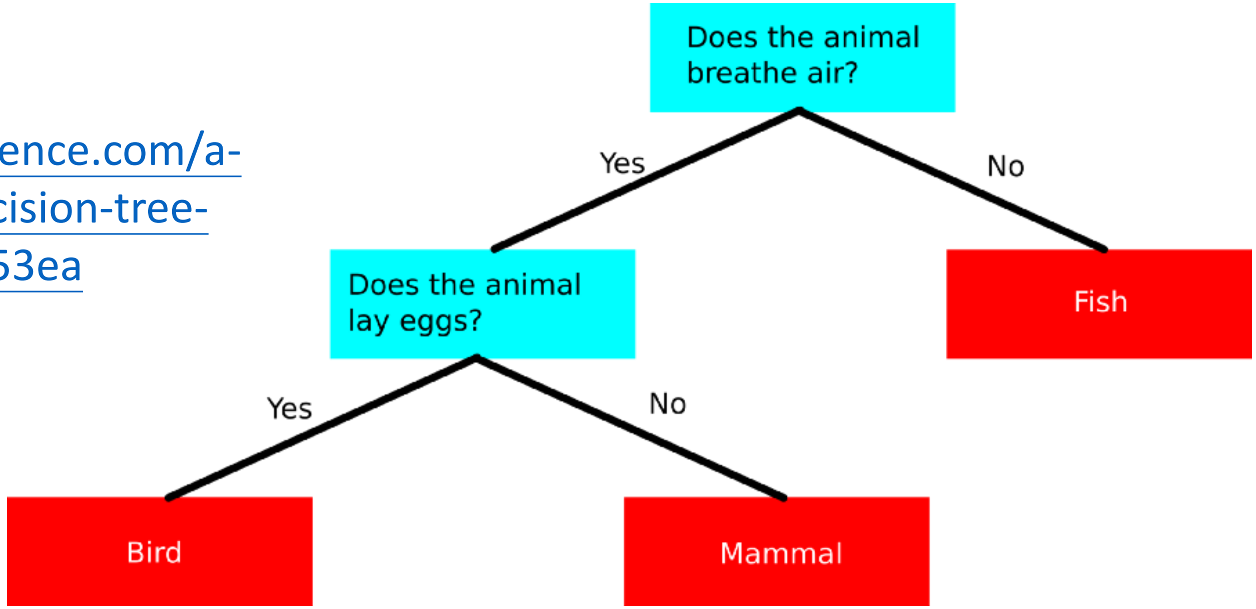
  - But it is more powerful than linear model …

    Let’s make deep network interpretable.

Interpretable v.s. Powerful

- Are there some models interpretable and powerful at the same time?
- How about decision tree?

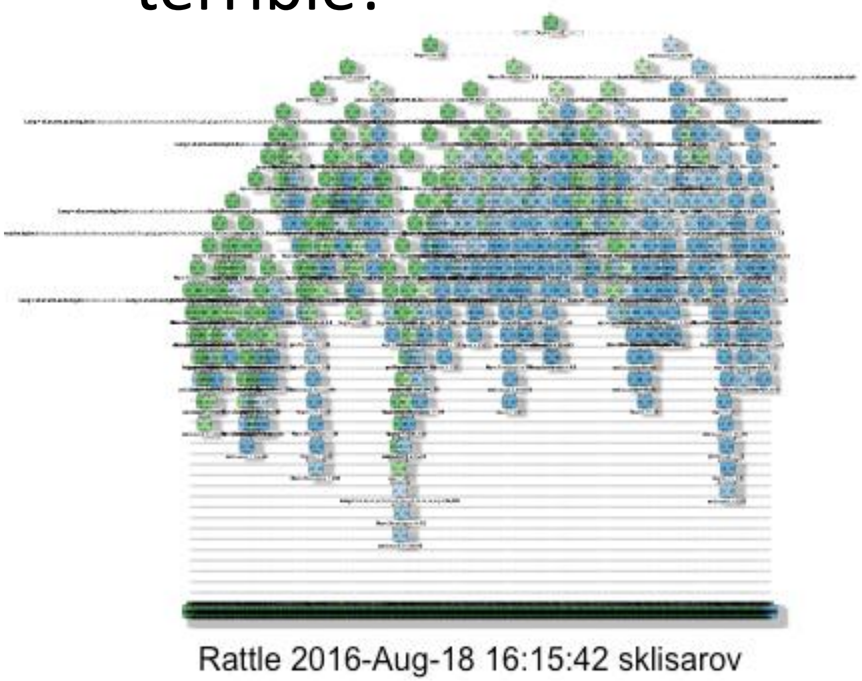
[Source of image: https://towardsdatascience.com/a-beginners-guide-to-decision-tree-classification-6d3209353ea](https://towardsdatascience.com/a-beginners-guide-to-decision-tree-classification-6d3209353ea)




只要用 Decision Tree 就好了
今天這堂課就是在浪費時間 ...

Interpretable v.s. Powerful

- A tree can still be terrible!



https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret

- We use a forest!

Local Explanation: Explain the Decision

Questions: Why do you think this image is a cat?

Source of image: www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Basic Idea

Image: pixel, segment, etc.
Text: a word

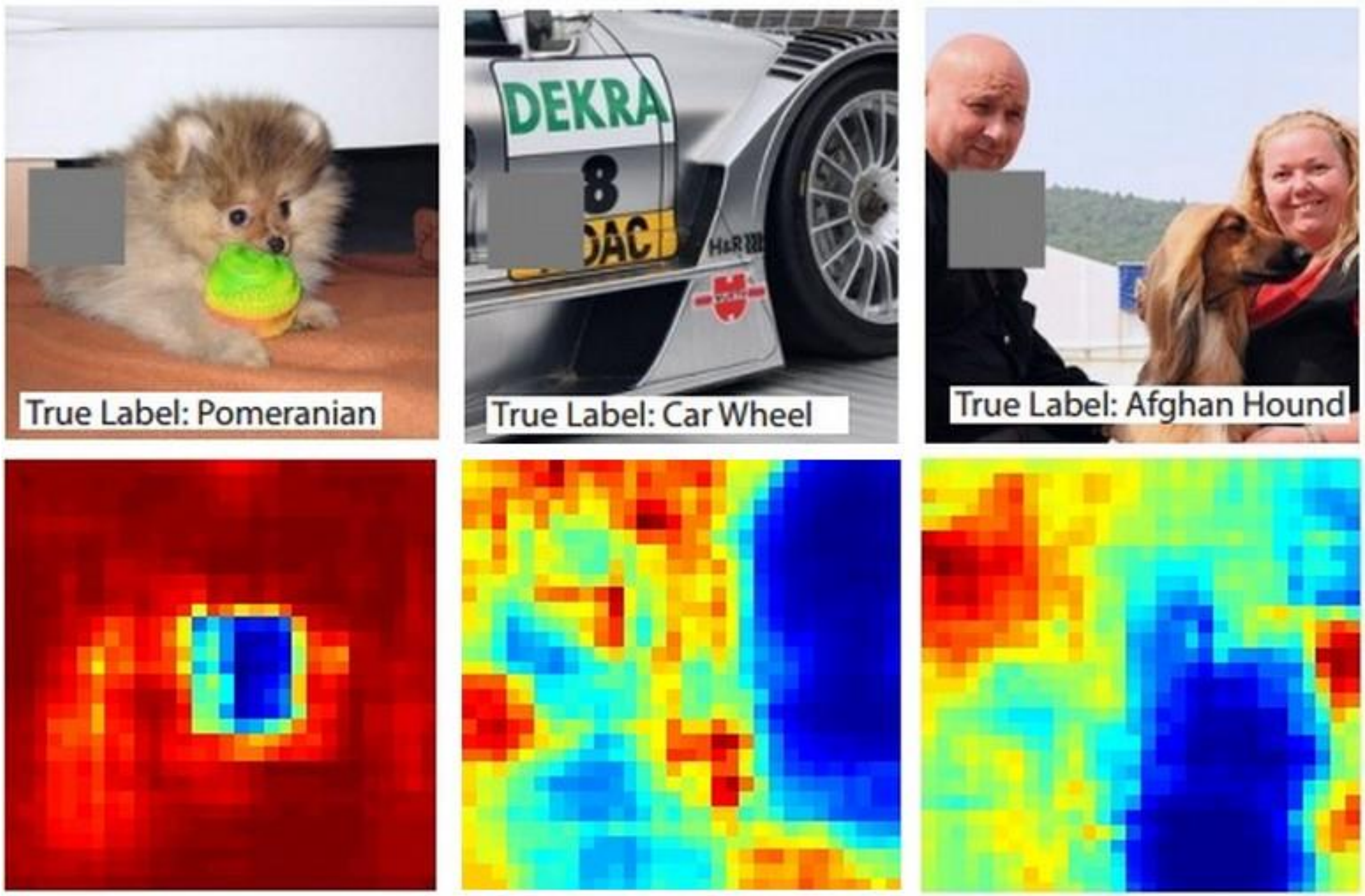
Object x ➡ Components: $\{x_1, \cdots, x_n, \cdots, x_N\}$

We want to know the importance of each components for making the decision.

Idea: Removing or modifying the values of the components, observing the change of decision.

Large decision change ➡ Important component

The size of the gray box can be crucial ……



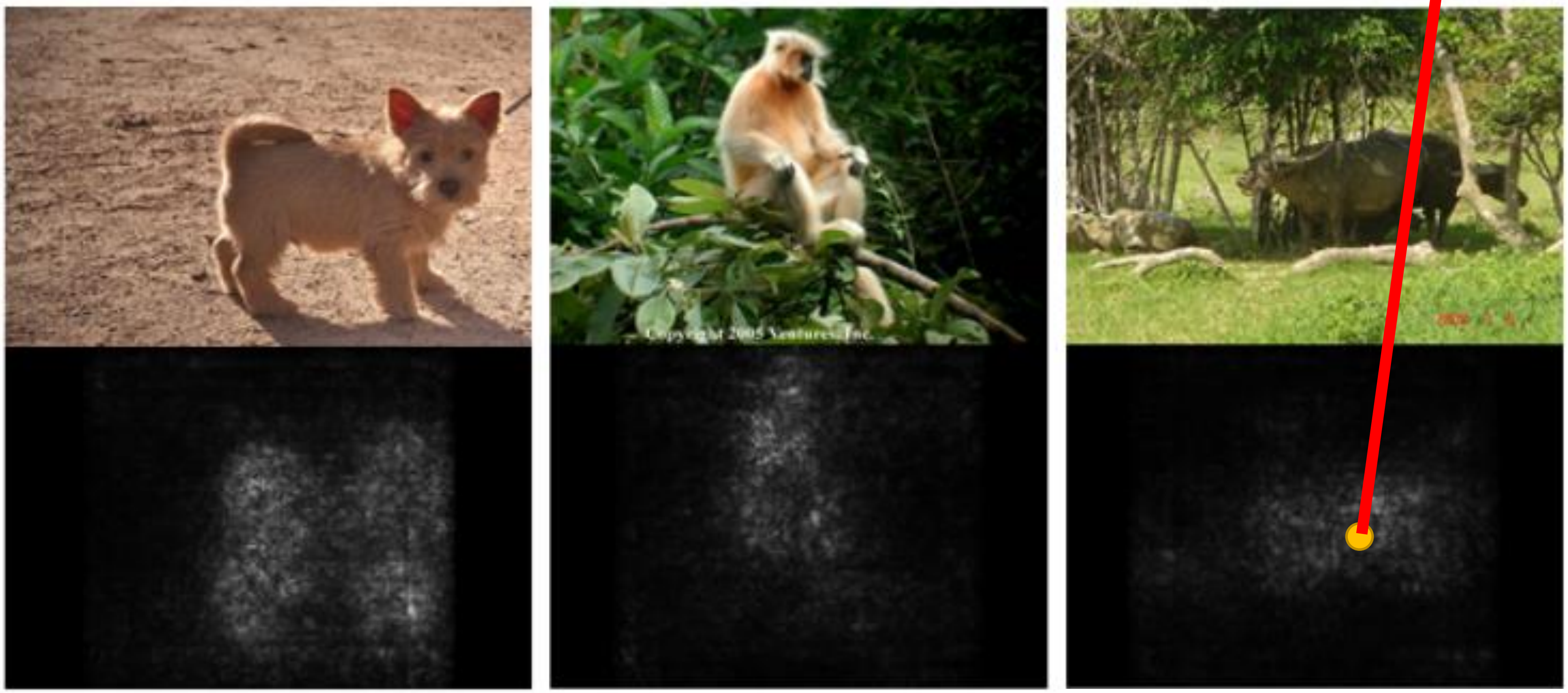
Reference: Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014* (pp. 818-833)

$$\{x_1, \cdots, x_n, \cdots, x_N\} \longrightarrow \{x_1, \cdots, x_n + \Delta x, \cdots, x_N\}$$

$$y_k \longrightarrow y_k + \Delta y$$

y_k: the prob of the predicted class of the model

$$\left|\frac{\Delta y}{\Delta x}\right| \longrightarrow \left|\frac{\partial y_k}{\partial x_n}\right|$$



Saliency Map

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR, 2014

Limitation of Gradient based Approaches

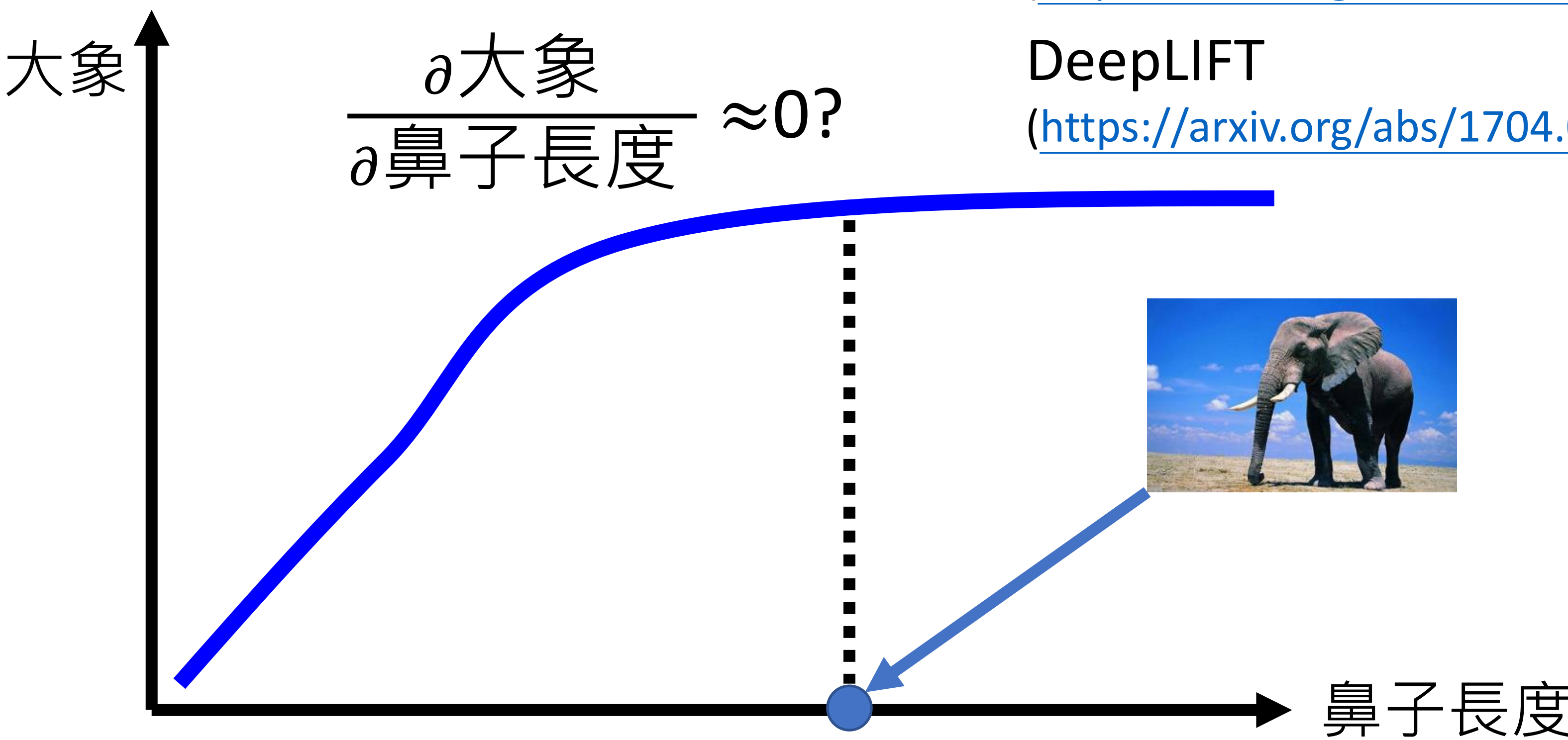
- Gradient Saturation

To deal with this problem:

Integrated gradient
(https://arxiv.org/abs/1611.02639)

DeepLIFT
(https://arxiv.org/abs/1704.02685)

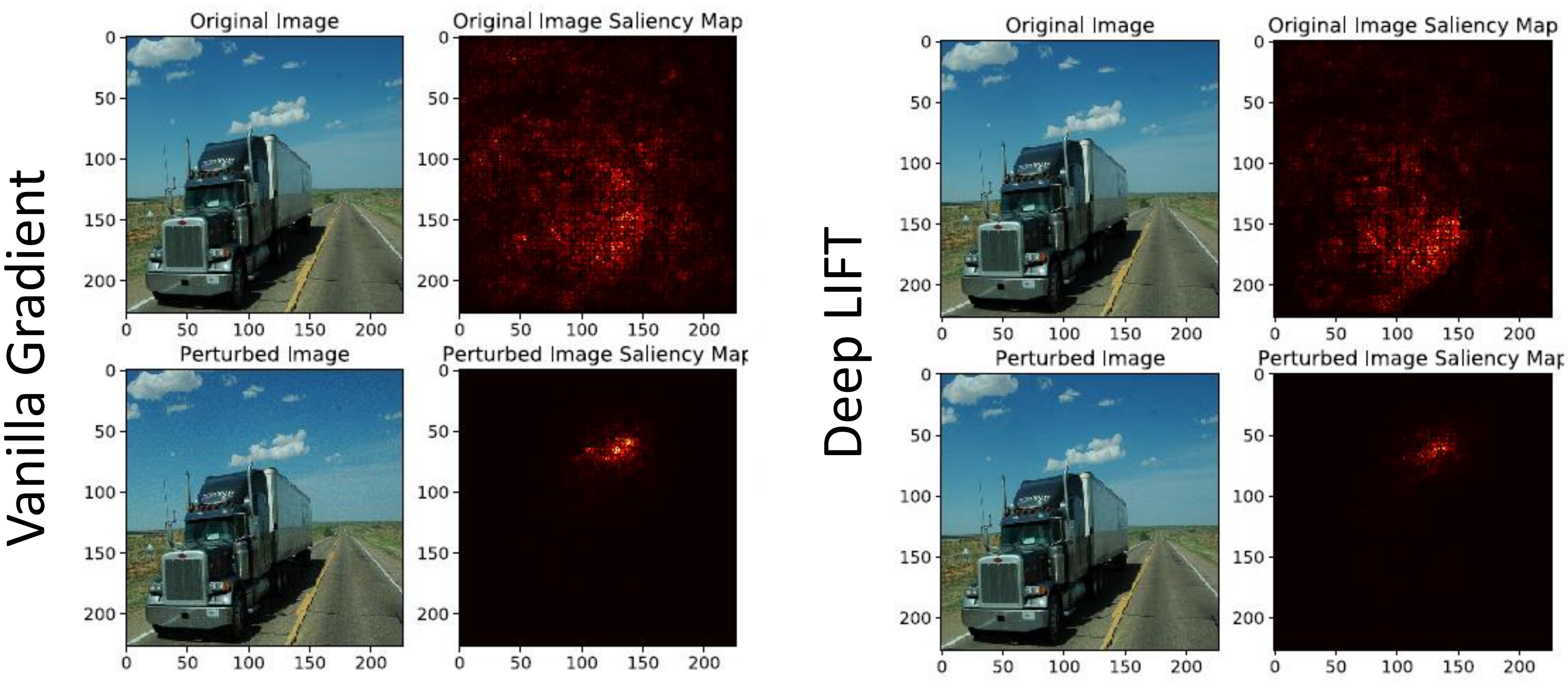
$$\frac{\partial \text{大象}}{\partial \text{鼻子長度}} \approx 0?$$



https://arxiv.org/abs/1710.10547

Attack Interpretation?!

- It is also possible to attack interpretation…



The noise is small, and do not change the classification results.

Case Study: Pokémon v.s. Digimon

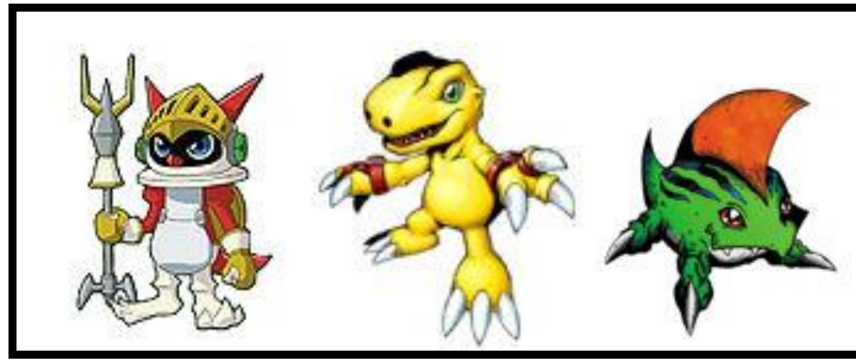
https://medium.com/@tyreeostevenson/teaching-a-computer-to-classify-anime-8c77bc89b881

Task

Pokémon images: https://www.Kaggle.com/kvpratama/pokemon-images-dataset/data

Digimon images: https://github.com/DeathReaper0965/Digimon-Generator-GAN

Pokémon



Digimon

Testing Images:

Experimental Results

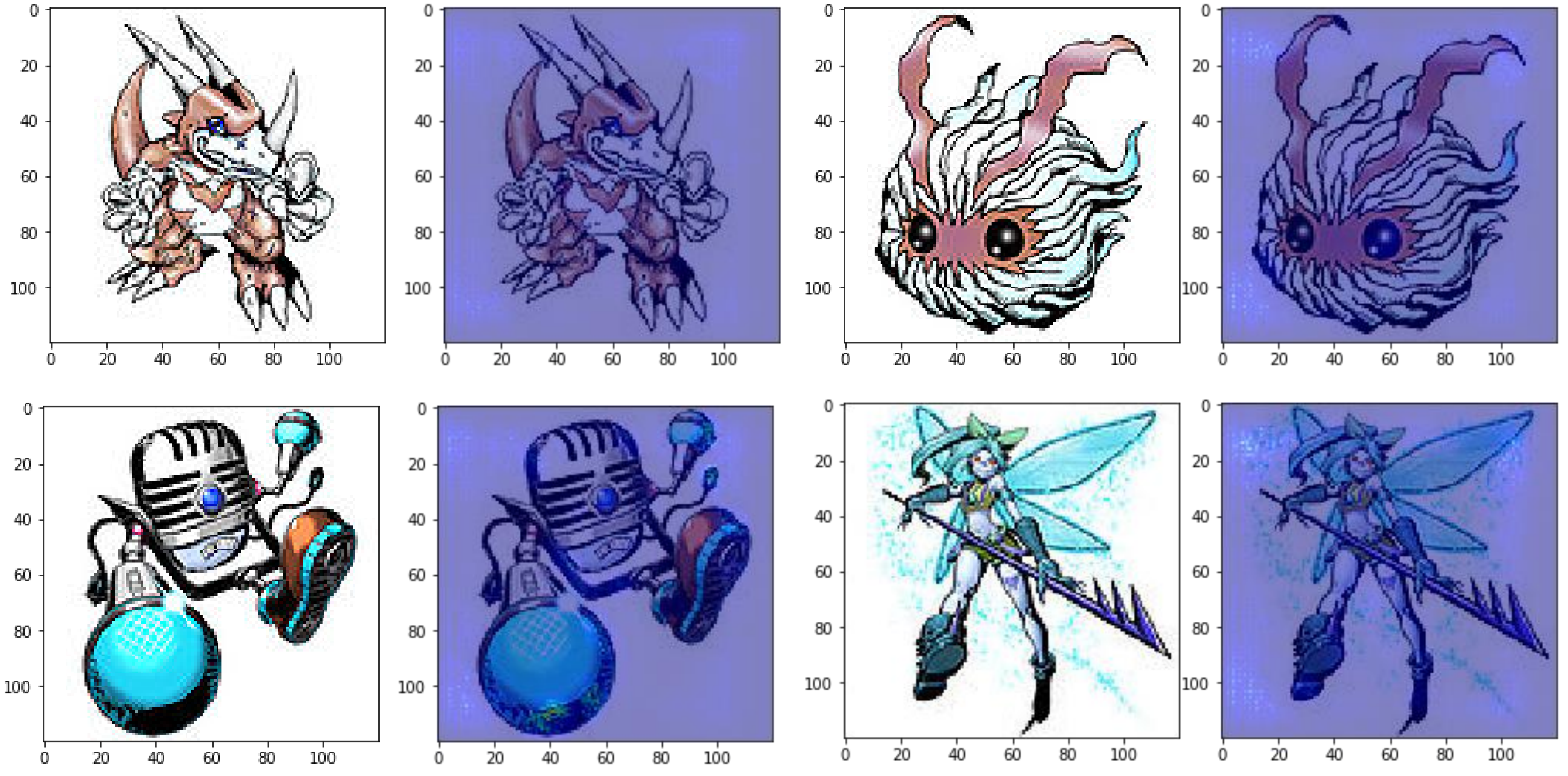
```
model = Sequential()
model.add(Conv2D(32, (3, 3), padding='same', input_shape=(120,120,3)))
model.add(Activation('relu'))
model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(64, (3, 3), padding='same'))
model.add(Activation('relu'))
model.add(Conv2D(64, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(256, (3, 3), padding='same'))
model.add(Activation('relu'))
model.add(Conv2D(256, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Flatten())
model.add(Dense(1024))
model.add(Activation('relu'))
model.add(Dense(2))
model.add(Activation('softmax'))
```

Training Accuracy: 98.9%

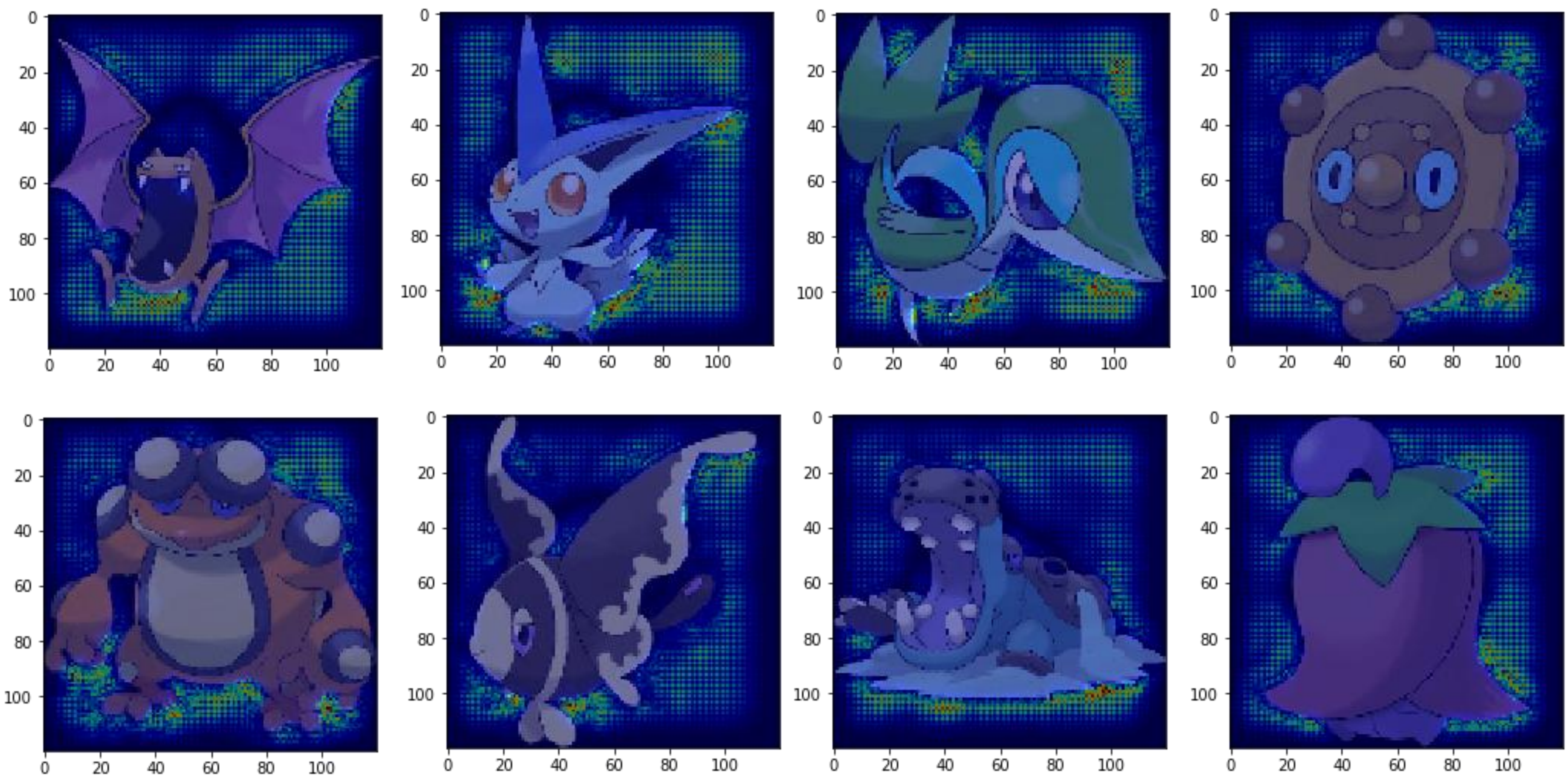
Testing Accuracy: 98.4%

太神啦!!!!!!

Saliency Map



Saliency Map



What Happened?

- All the images of Pokémon are PNG, while most images of Digimon are JPEG.

PNG 檔透明背景

讀檔後背景是黑的!

Machine discriminate Pokémon and Digimon based on Background color.

This shows that explainable ML is very critical.

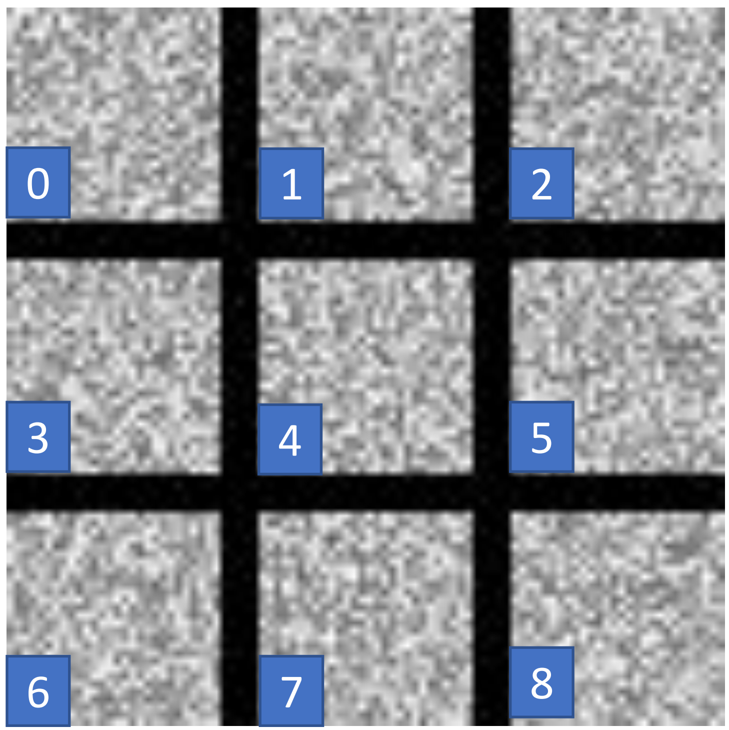
GLOBAL EXPLANATION: EXPLAIN THE WHOLE MODEL

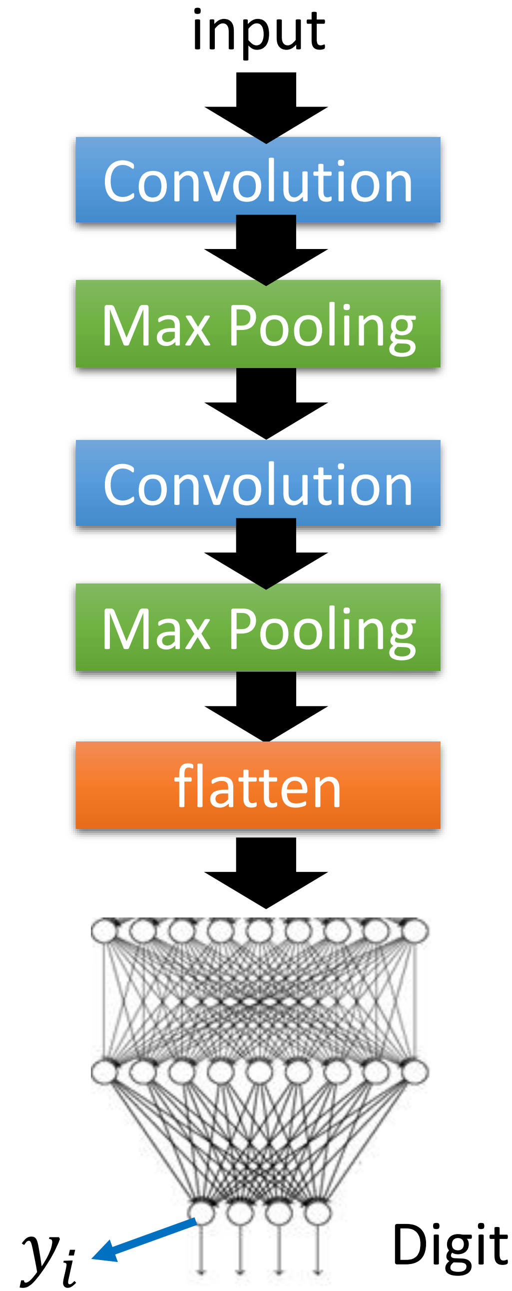
Question: What do you think a “cat” looks like?

Activation Maximization (review)

$$x^* = arg \max_x y_i$$

Can we see digits?

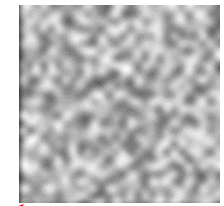




Deep Neural Networks are Easily Fooled

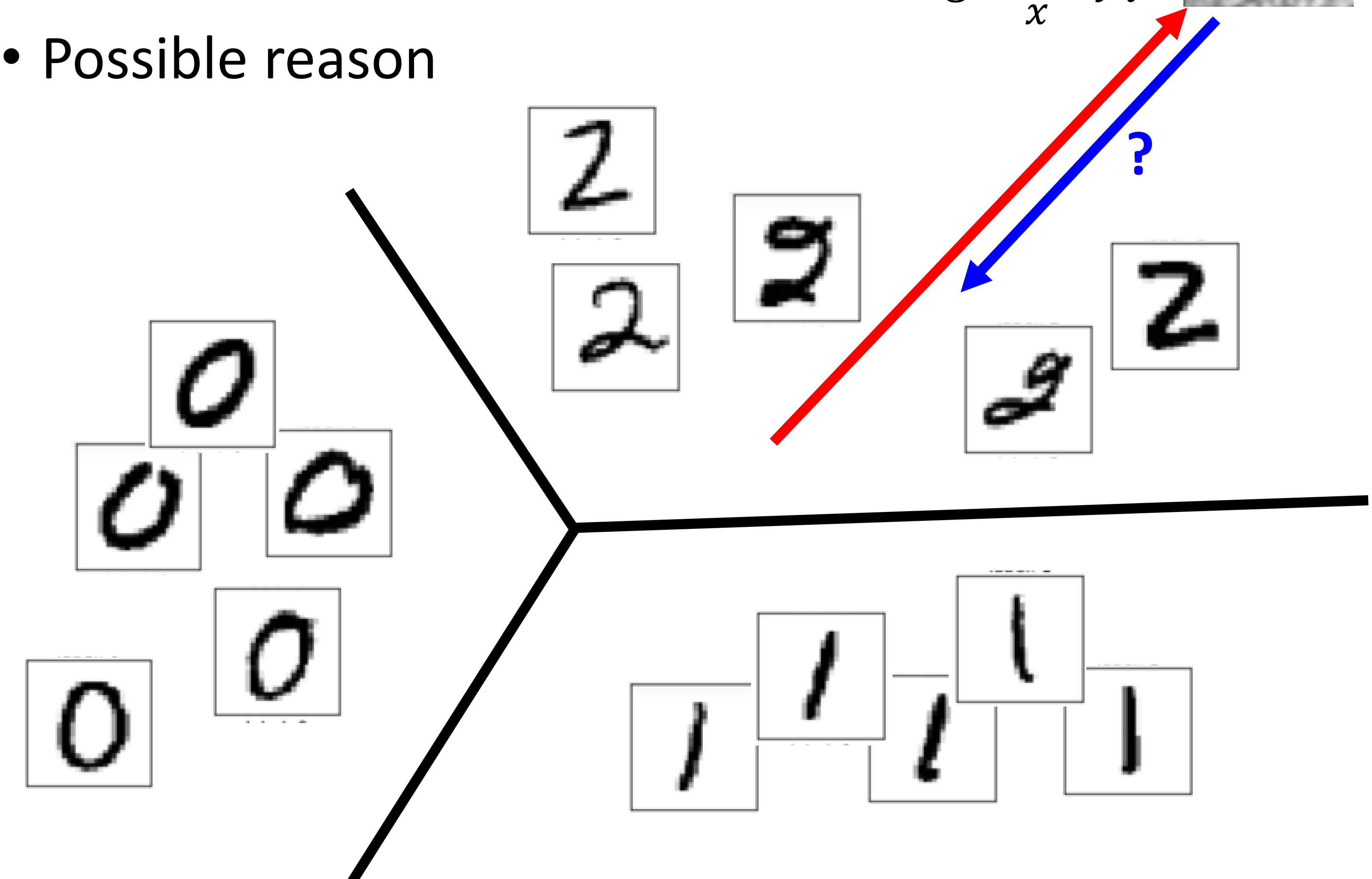
https://www.youtube.com/watch?v=M2IebCN9Ht4

Activation Maximization



$$x^* = arg\max_x y_i$$

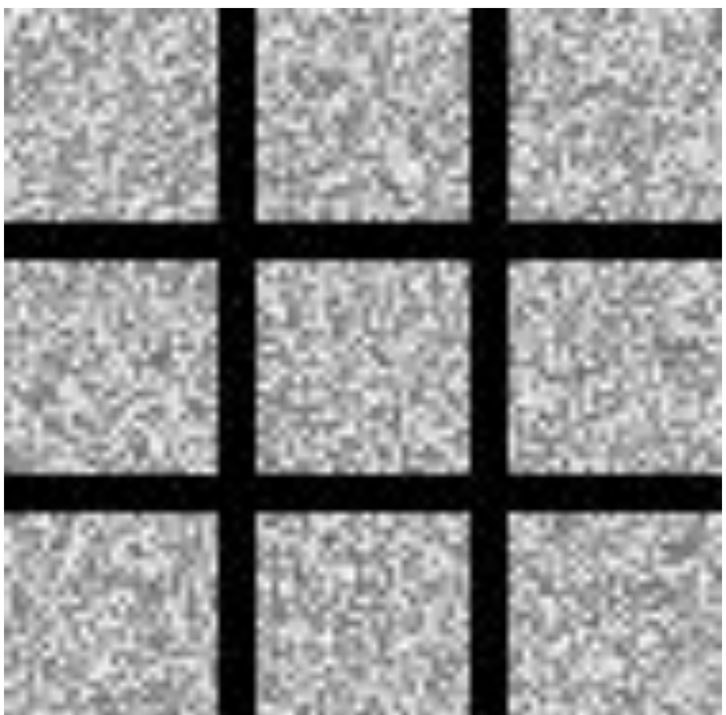
- Possible reason



Activation Maximization (review)

Find the image that maximizes class probability

$$x^* = arg\max_x y_i$$

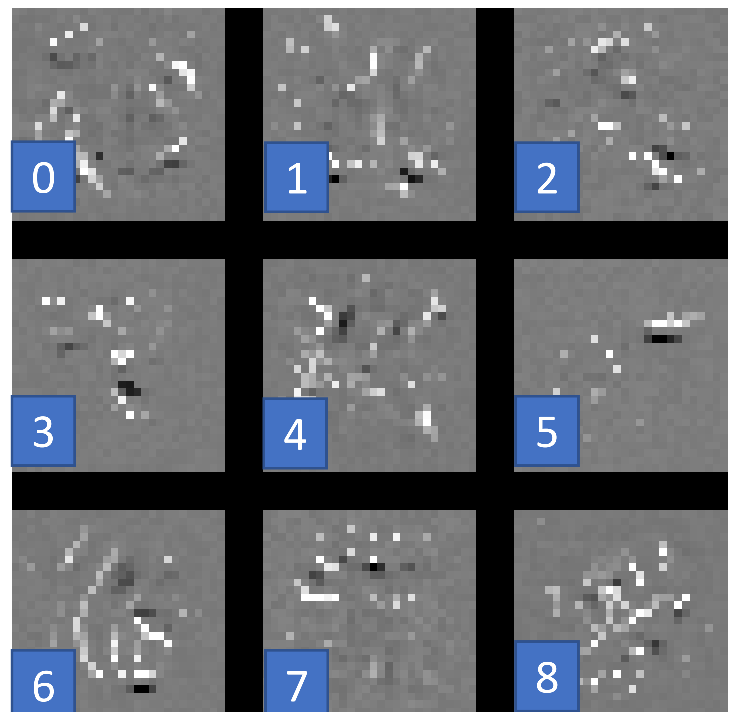


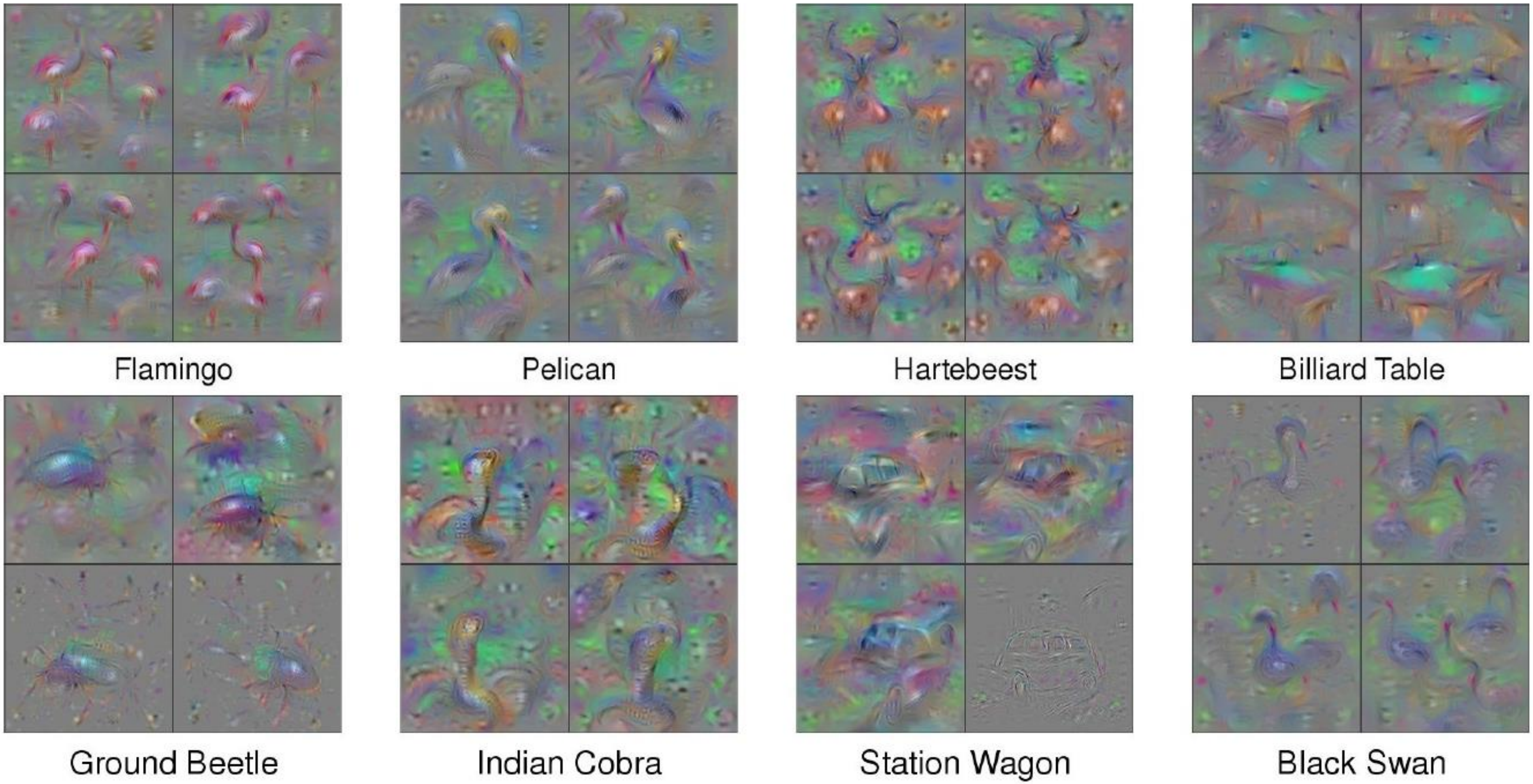
The image also looks like a digit.

$$x^* = arg\max_x y_i \quad + \underline{R(x)}$$

$$R(x) = -\sum_{i,j} |x_{ij}|$$

How likely x is a digit





With several regularization terms, and hyperparameter tuning …..

https://arxiv.org/abs/1506.06579

(Simplified Version)

Constraint from Generator

- Training a generator (by GAN, VAE, etc.)

Training Examples

low-dim vector z → Image Generator G → Image x

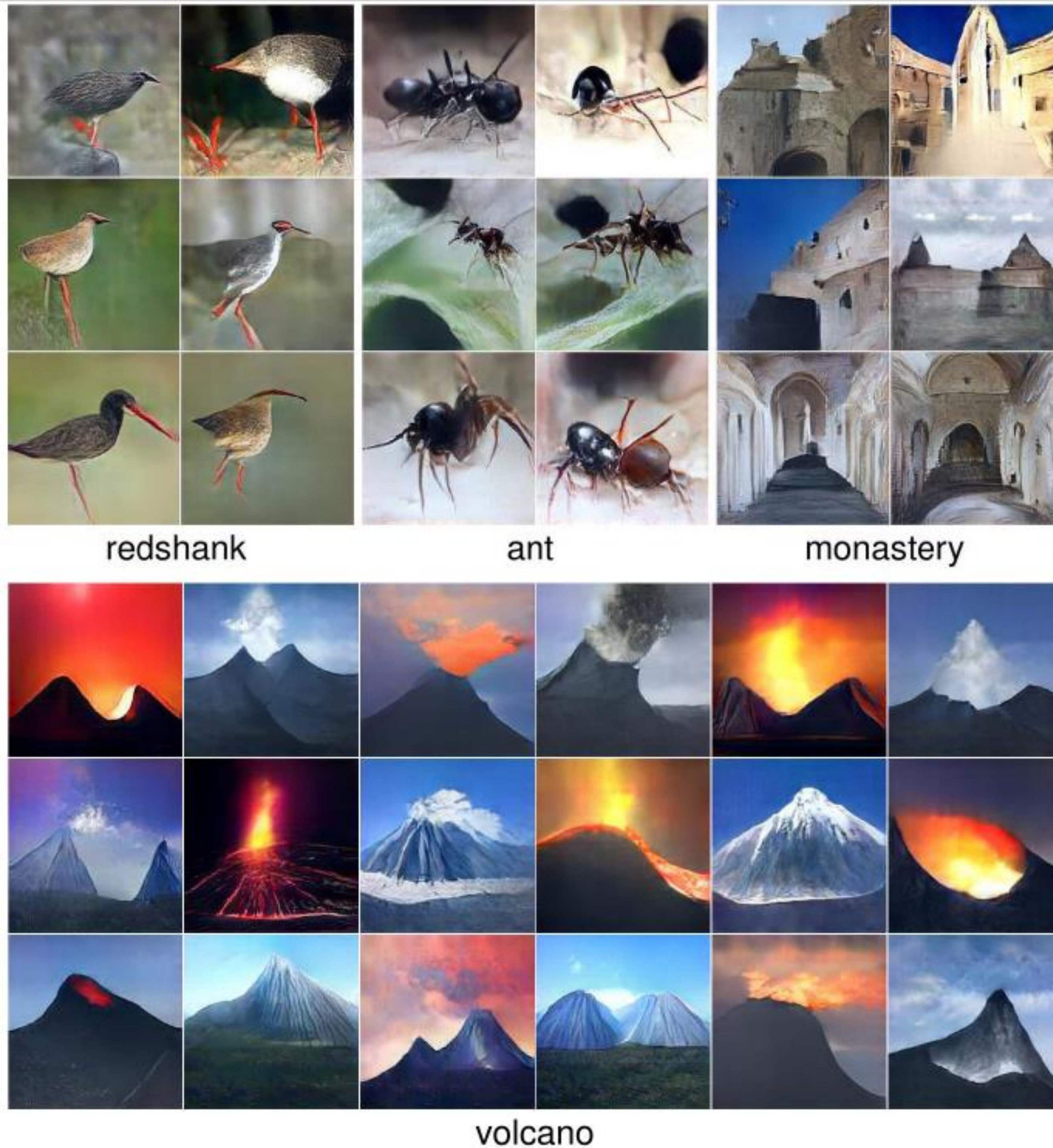
$x = G(z)$

$$x^* = arg\max_x y_i \quad \Rightarrow \quad z^* = arg\max_z y_i$$

Show image:

$$x^* = G(z^*)$$

z → Image Generator → Image x → Image Classifier → y



https://arxiv.org/abs/1612.00005

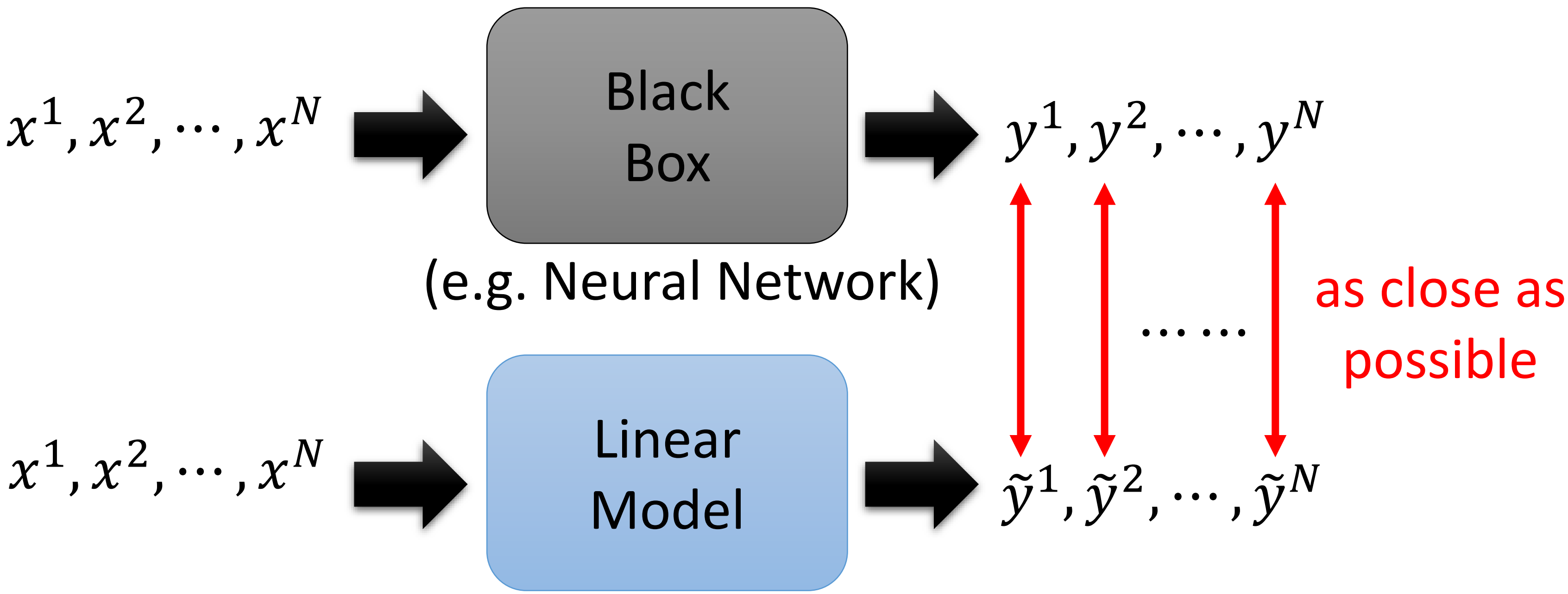
USING A MODEL TO EXPLAIN ANOTHER

Some models are easier to Interpret.

Using interpretable model to mimic uninterpretable models.

Using a model to explain another

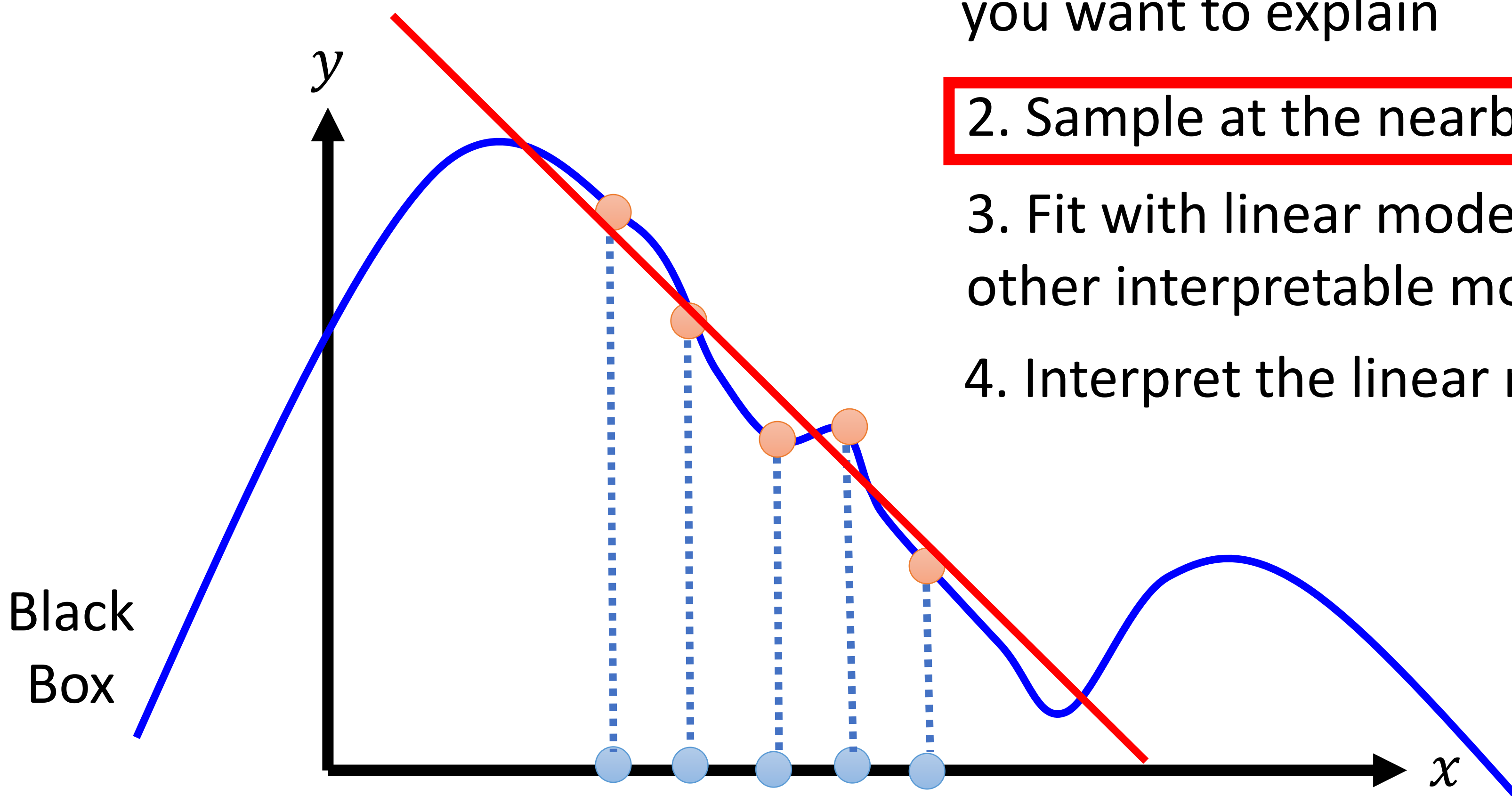
- Using an interpretable model to mimic the behavior of an uninterpretable model.



Problem: Linear model cannot mimic neural network …

However, it can mimic a local region.

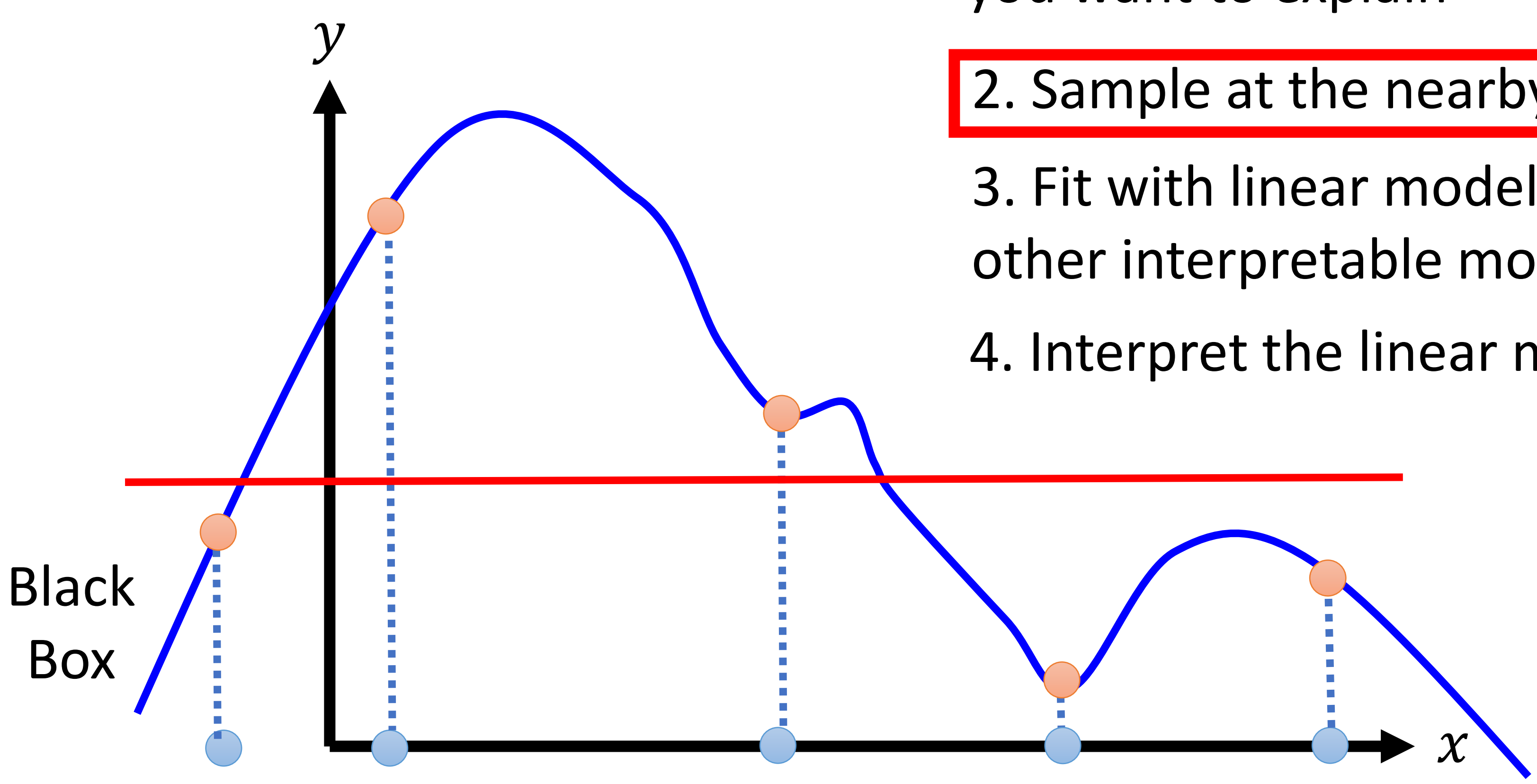
Local Interpretable Model-Agnostic Explanations (LIME)



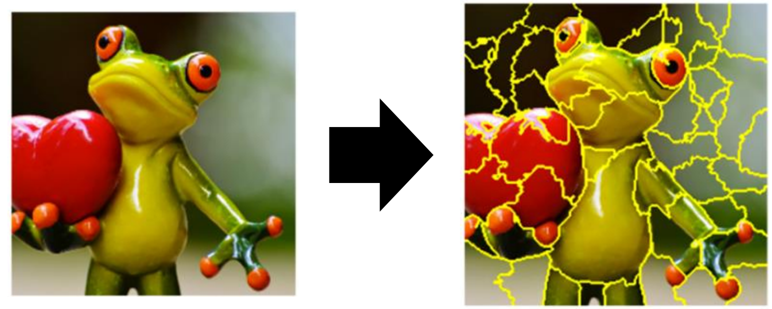
1. Given a data point you want to explain
2. Sample at the nearby
3. Fit with linear model (or other interpretable models)
4. Interpret the linear model

Local Interpretable Model-Agnostic Explanations (LIME)

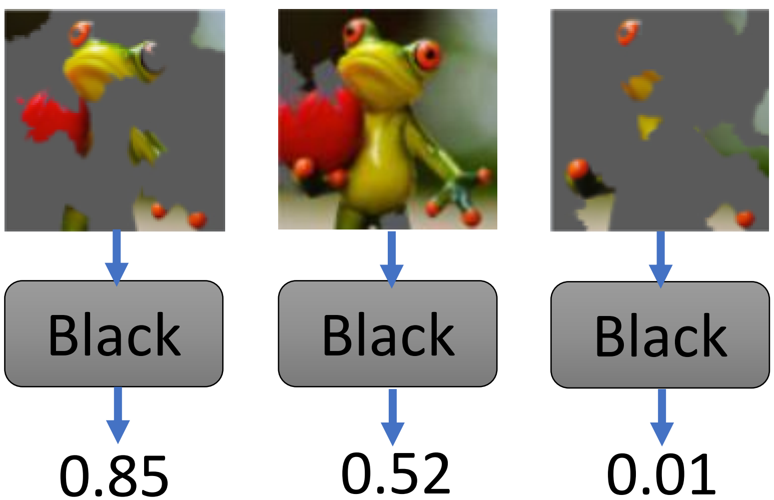
1. Given a data point you want to explain
2. Sample at the nearby
3. Fit with linear model (or other interpretable models)
4. Interpret the linear model



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).

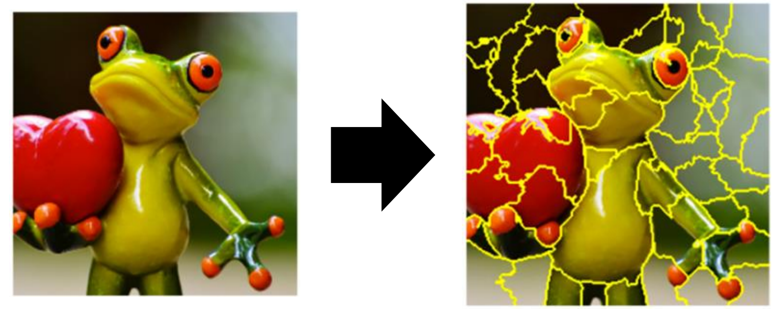


Randomly delete some segments.

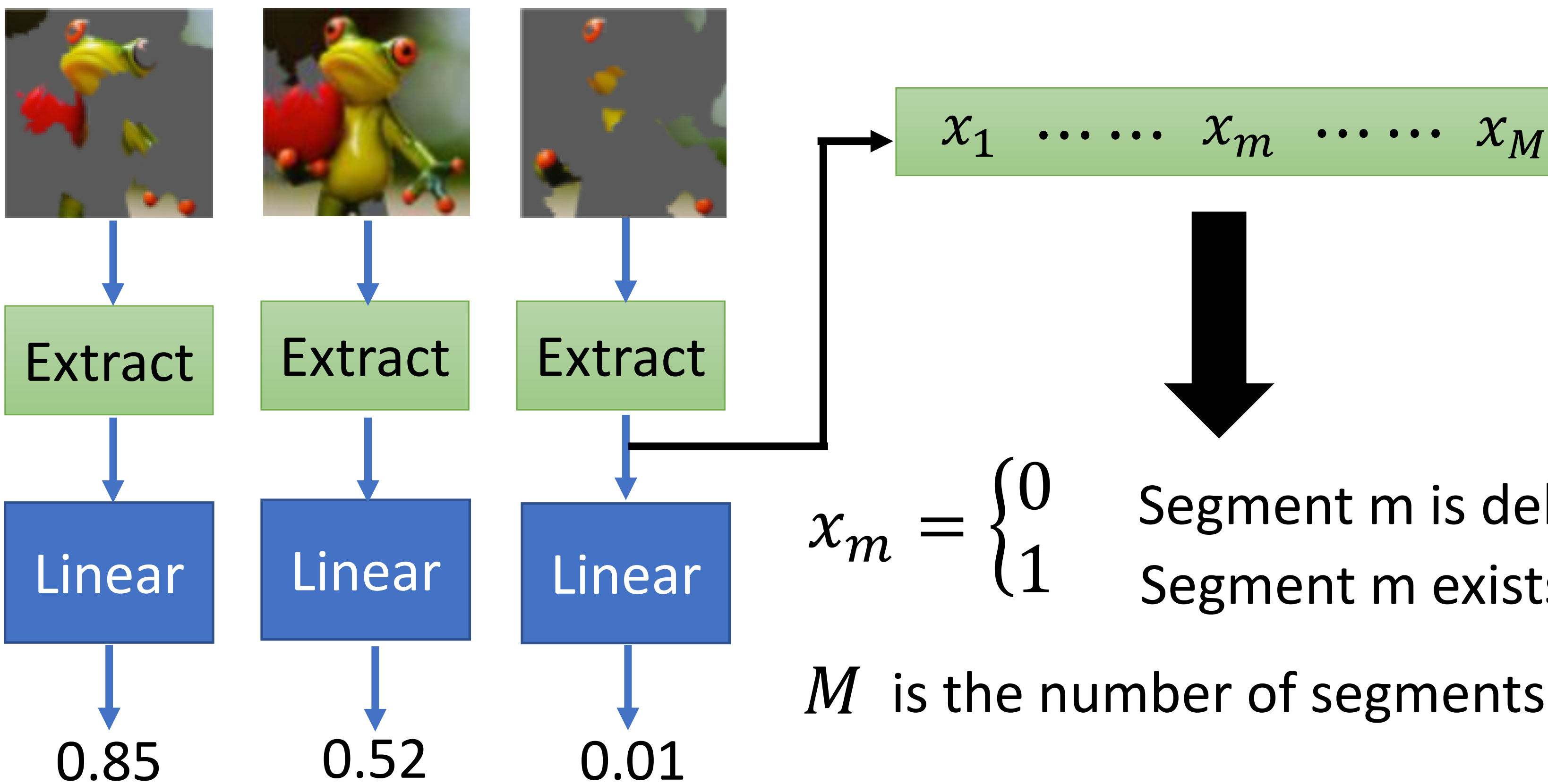
Compute the probability of "frog" by black box

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

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.

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.

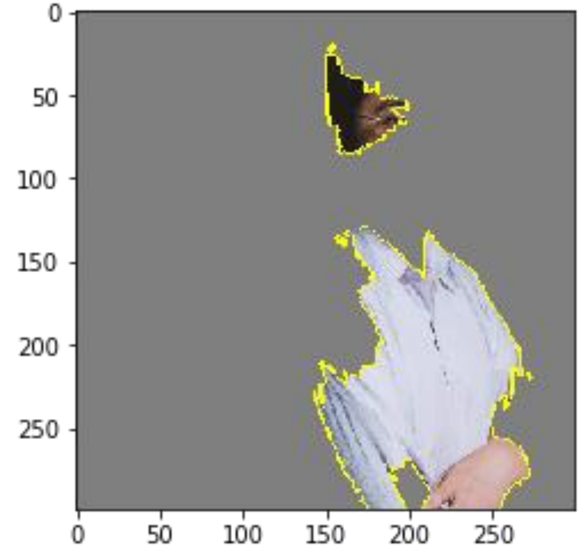
If $w_m \approx 0$ ➡ segment m is not related to “frog”

If w_m is positive
➡ segment m indicates the image is “frog”

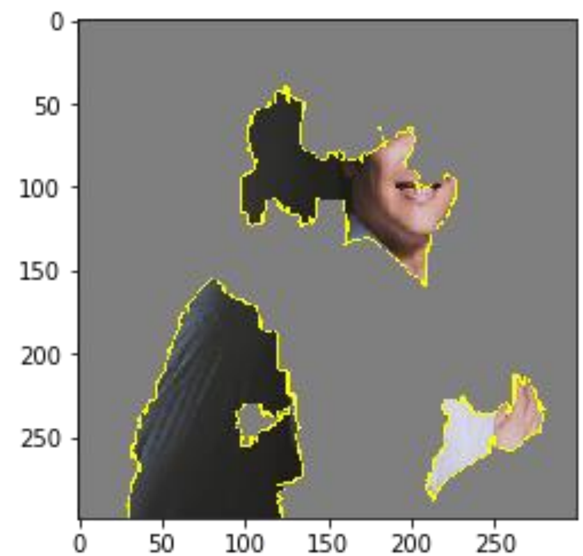
If w_m is negative
➡ segment m indicates the image is not “frog”

LIME - Example

和服：0.25
實驗袍：0.05



實驗袍



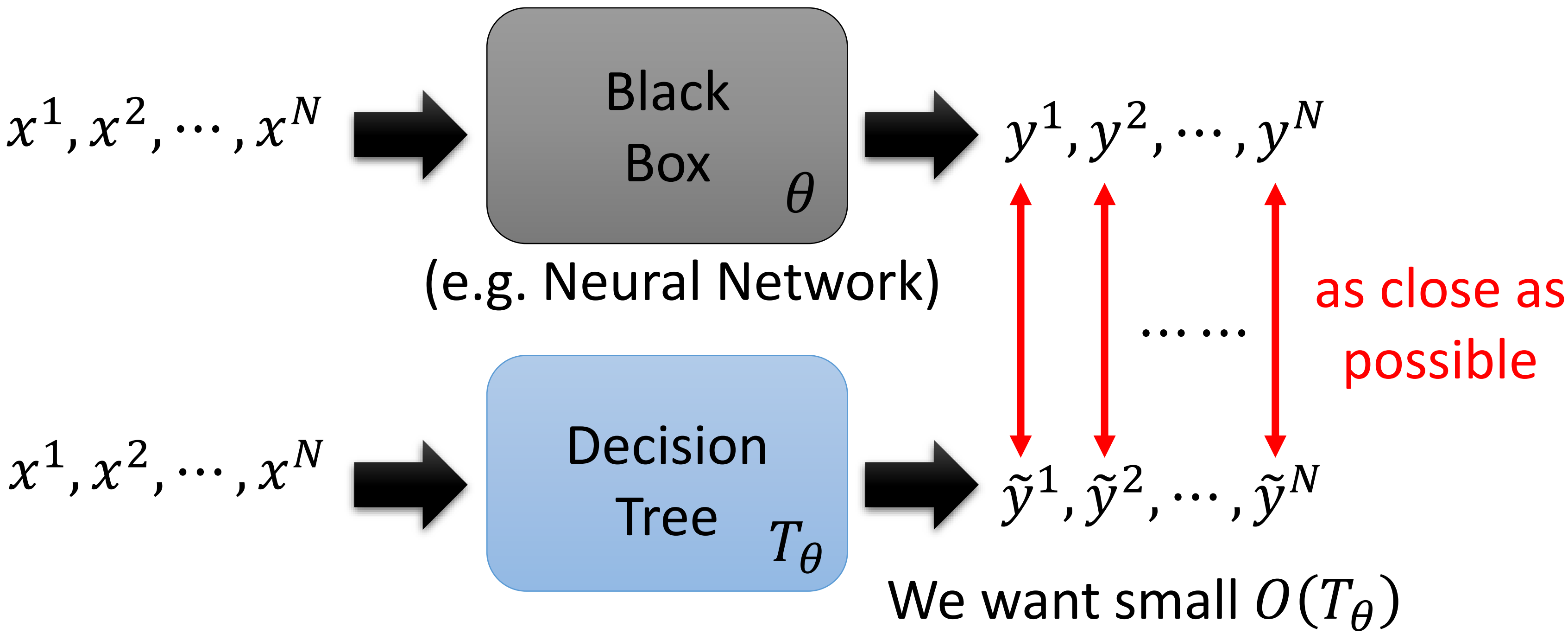
和服

Decision Tree

$O(T_\theta)$: how complex T_θ is

e.g. average depth of T_θ

- Using an interpretable model to mimic the behavior of an uninterpretable model.



Problem: We don't want the tree to be too large.

Decision Tree – Tree regularization

https://arxiv.org/pdf/1711.06178.pdf

- Train a network that is easy to be interpreted by decision tree.

T_θ : tree mimicking network with parameters θ

$O(T_\theta)$: how complex T_θ is

$$\theta^* = arg\min_\theta \underbrace{L(\theta)}_{\text{Original loss function for training network}} + \lambda \underbrace{O(T_\theta)}_{\text{Preference for network parameters}}$$

Original loss function for training network

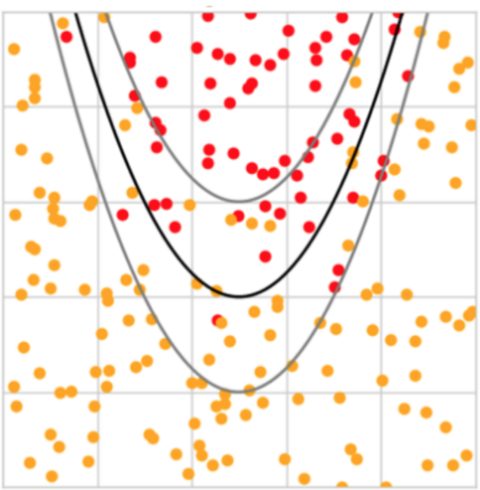
Preference for network parameters

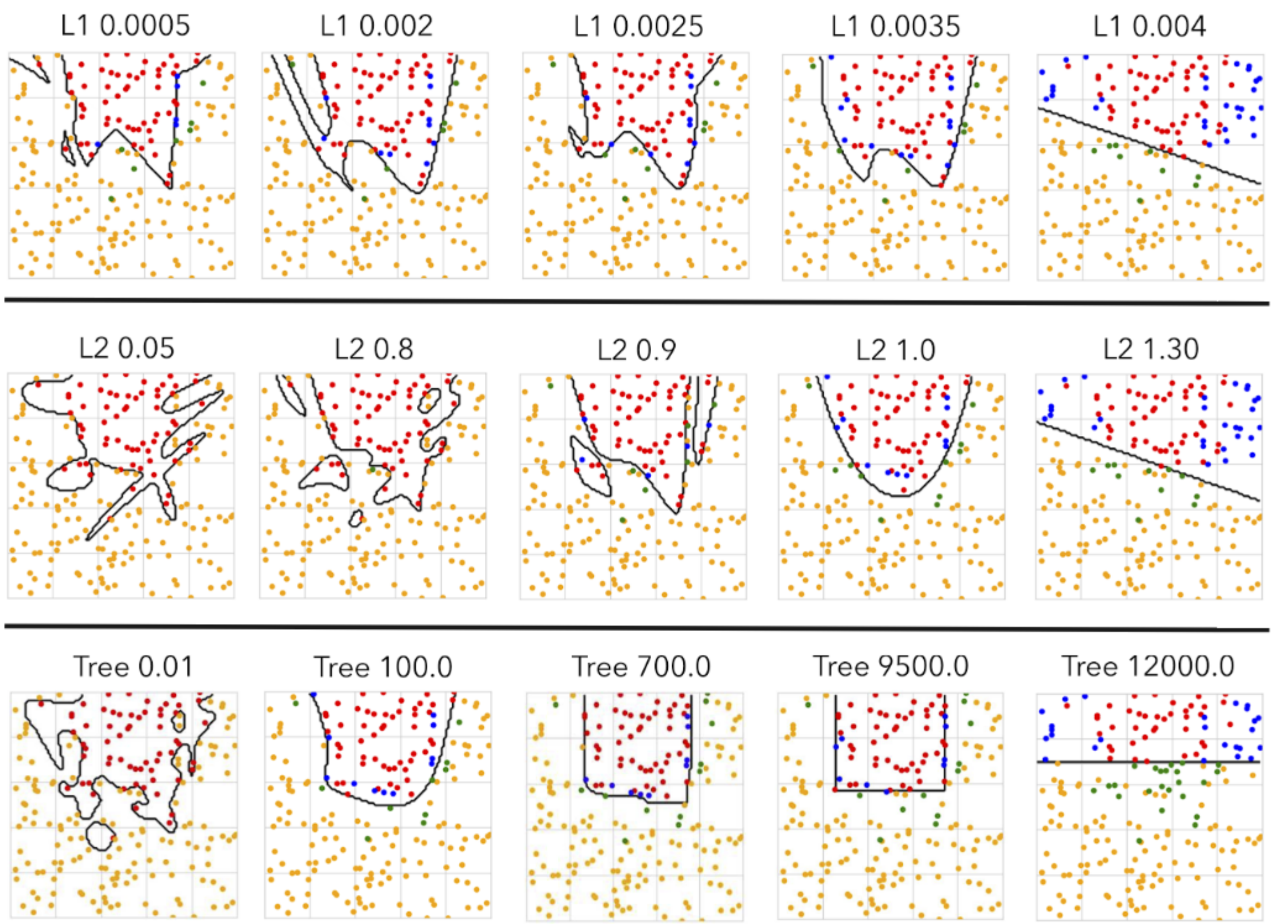
➡ Tree Regularization

Is the objective function with tree regularization differentiable? No! Check the reference for solution.

Decision Tree
– Experimental Results

Dataset

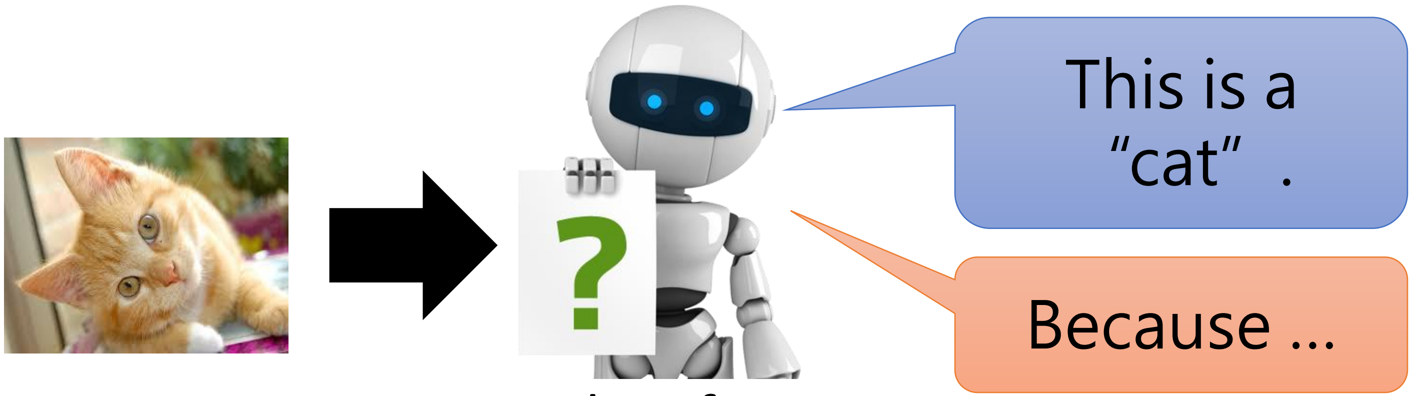




Red: Positive

Yellow: Negative

Concluding Remarks



Classifier

Local Explanation

Why do you think *this image* is a cat?

Global Explanation

What do you think a "cat" look like?

Using an interpretable model to explain an uninterpretable model