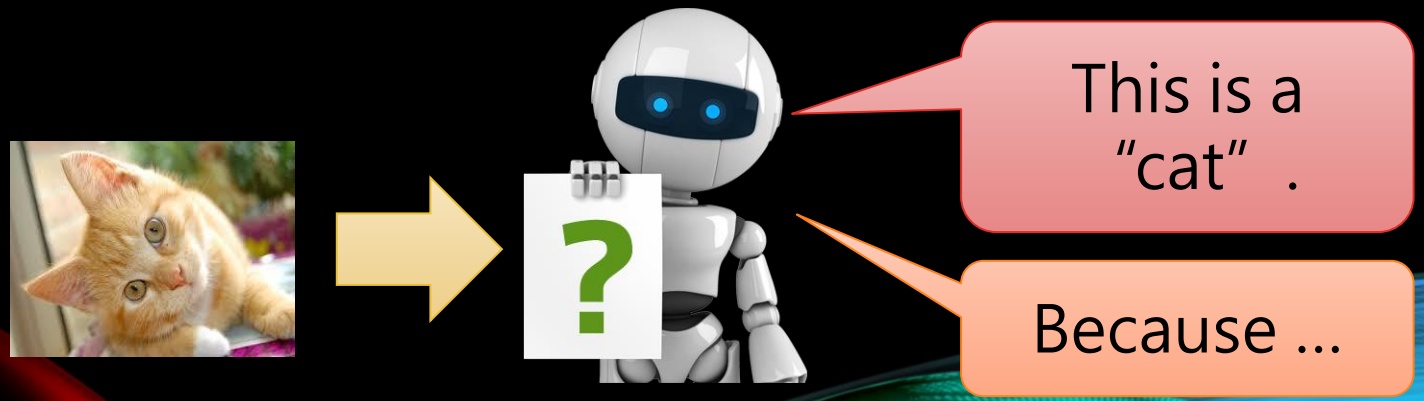
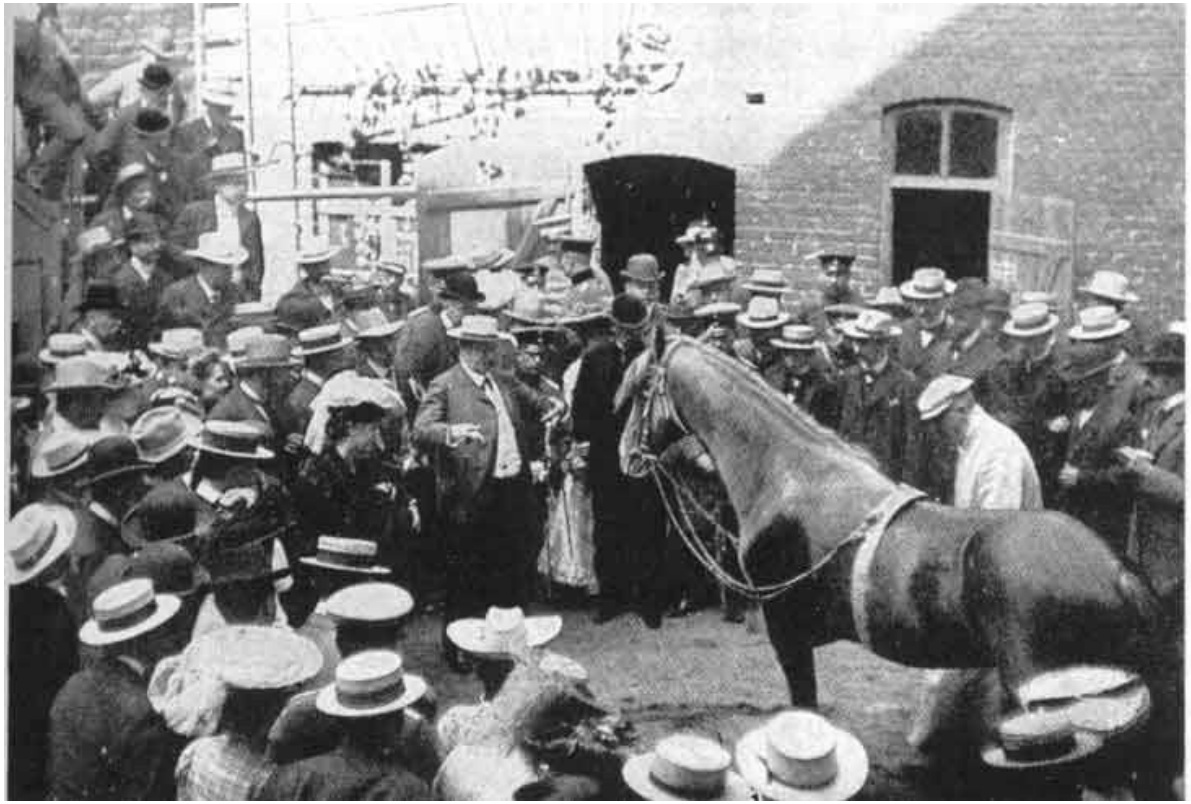
EXPLAINABLE MACHINE LEARNING

Hung-yi Lee 李宏毅



Why we need Explainable ML?

- Correct answers ≠ Intelligent

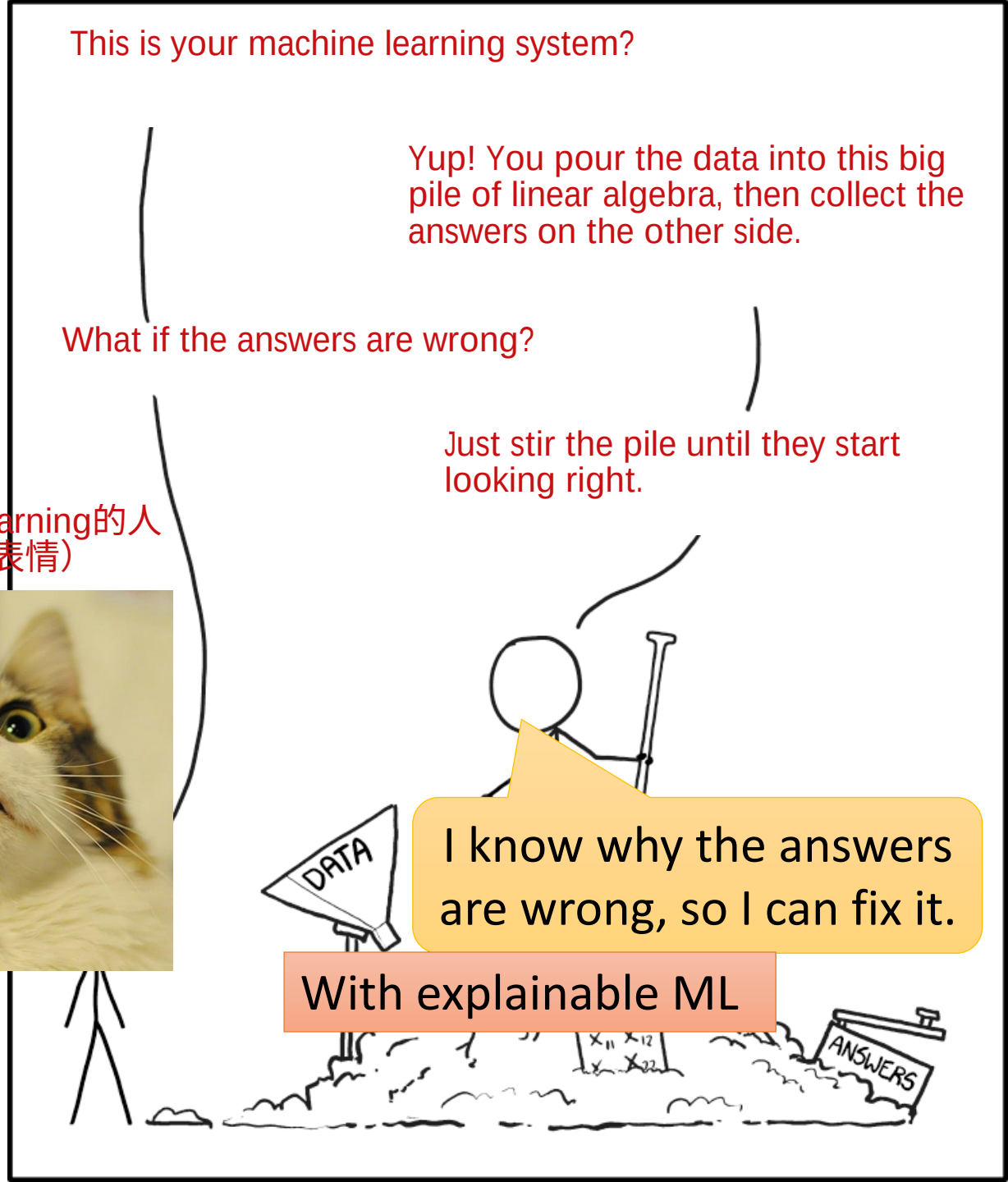


Why we need Explainable ML?

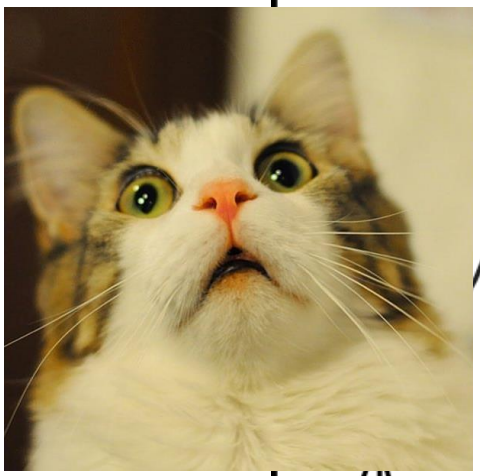
我们需要决策背后的理由

- Loan issuers are required by law to explain their models.
- Medical diagnosis model is responsible for human life. Can it be a black box?
- If a model is used at the court, we must make sure the model behaves in a nondiscriminatory manner.
- If a self-driving car suddenly acts abnormally, we need to explain why.

We can improve ML model based on explanation.



（没做过deep learning的人知道真相之后的表情）



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

可解释性 强大

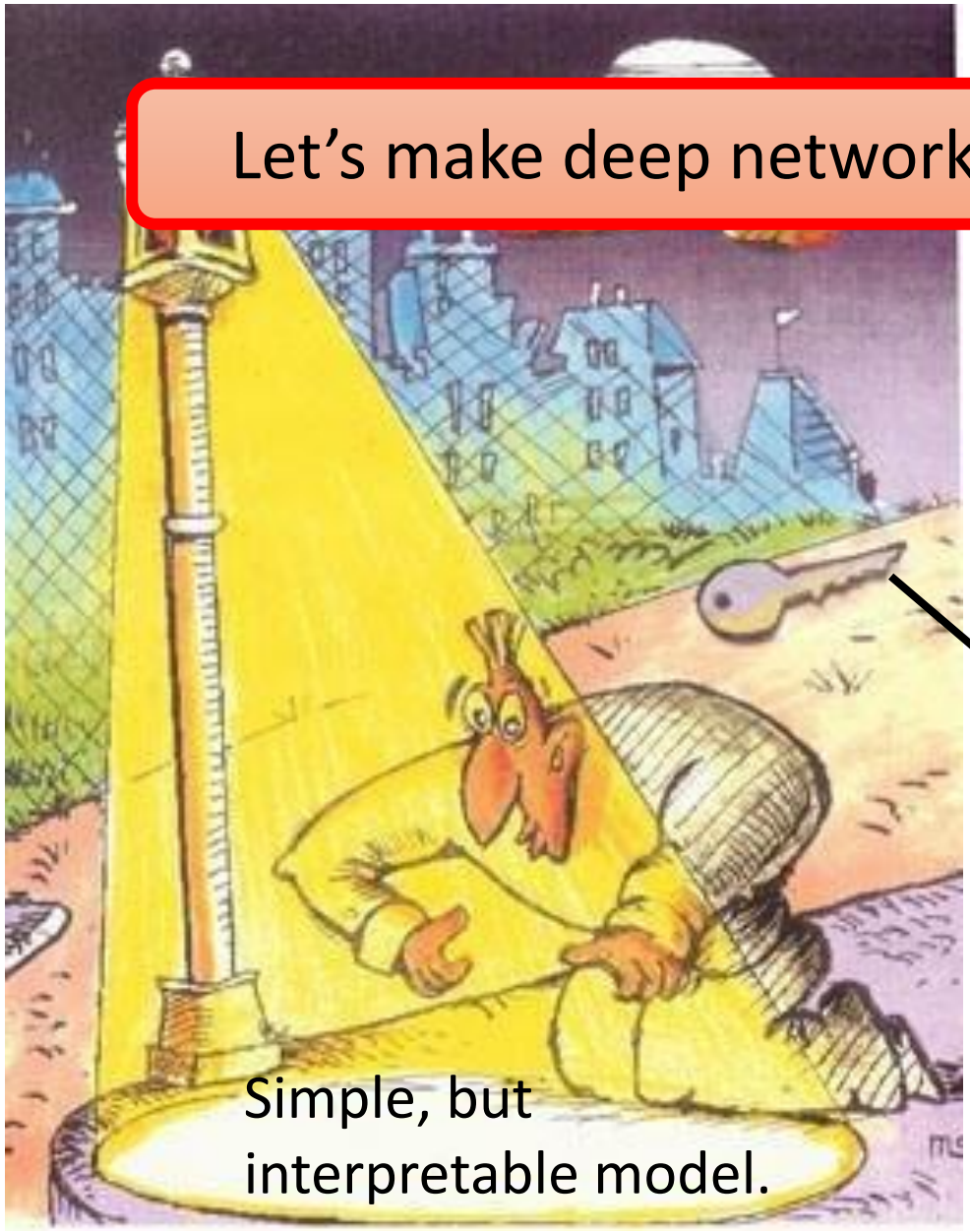
Interpretable v.s. Powerful

- 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 networks are black boxes … but powerful than a linear model.

We don't want to use a more powerful model because it is a black box.

This is "cut the feet to fit the shoes." (削足適履)

改变路灯的范围



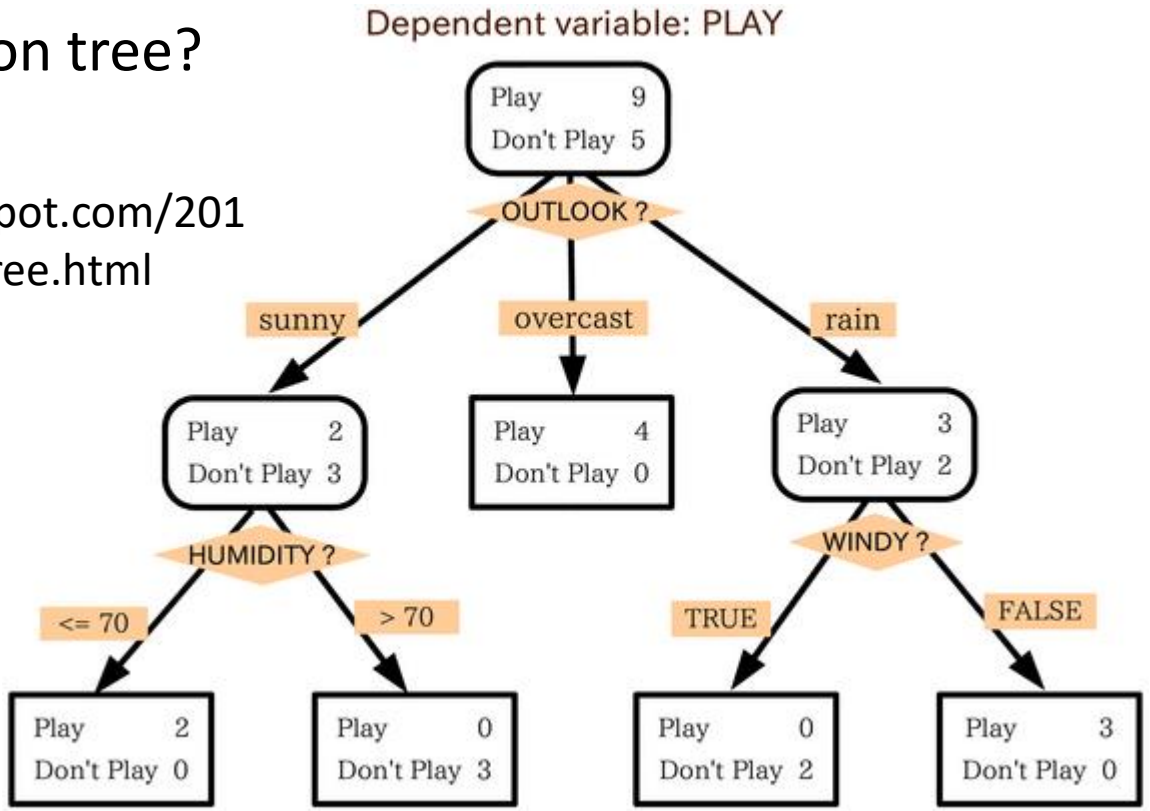
explainable：本来是黑箱
interpretable：本来不是黑箱，可以被解读的

Source of image: https://kknews.cc/news/pnynzgp.html

Interpretable v.s. Powerful

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

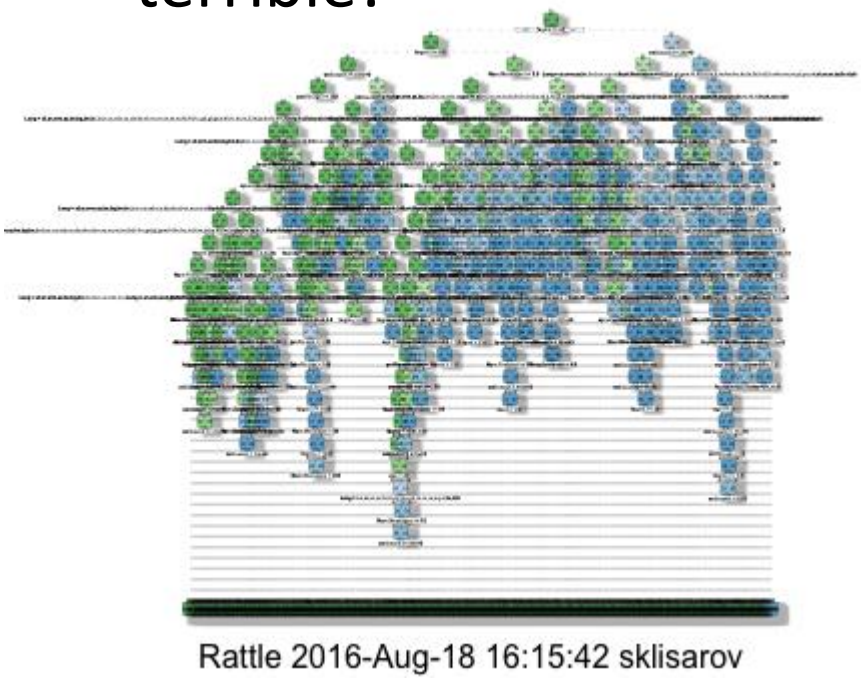
Source of image: https://mropengate.blogspot.com/2015/06/ai-ch13-2-decision-tree.html




Decision tree is all you need!?

Interpretable v.s. Powerful

- A tree can still be terrible! 决策树也可以是特别复杂的



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

- We use a forest! 随机森林

Goal of Explainable ML

- Completely know how an ML model works?
  - We do not completely know how brains work!
  - But we trust the decision of humans!

人脑也是黑盒子

理由很重要

The Copy Machine Study (Ellen Langer, Harvard University)

"Excuse me, I have 5 pages. May I use the Xerox machine?"

60% accept

"Excuse me, I have 5 pages. May I use the Xerox machine, **because I'm in a rush**?"

94% accept

"Excuse me, I have 5 pages. May I use the Xerox machine, **because I have to make copies**?"

随便说个理由，别人也会接受

93% accept

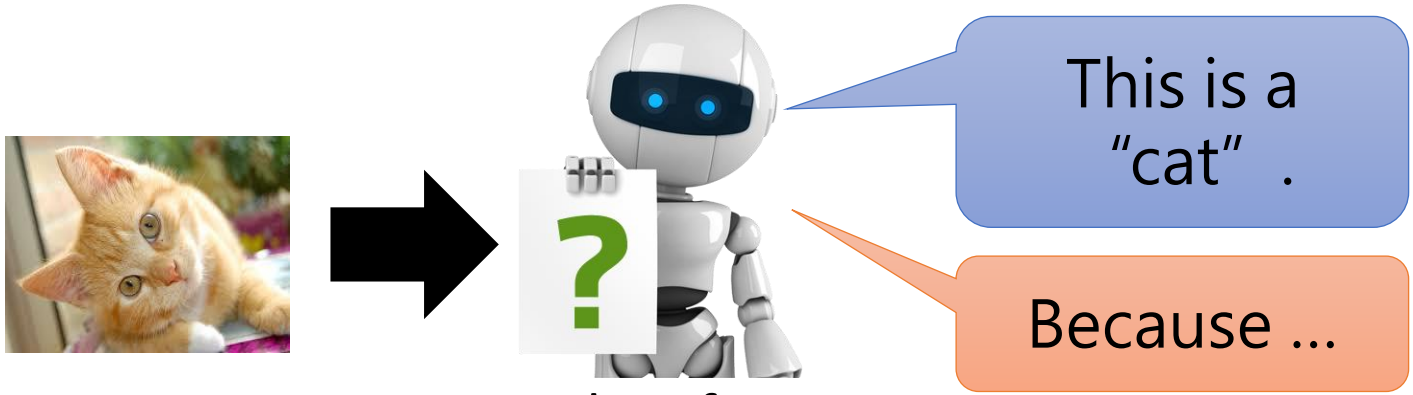
https://jamesclear.com/wp-content/uploads/2015/03/copy-machine-study-ellen-langer.pdf

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

(my two cents)

有两类

Explainable ML



1 ***Local Explanation*** 对特定的图片进行分析

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

2 ***Global Explanation***

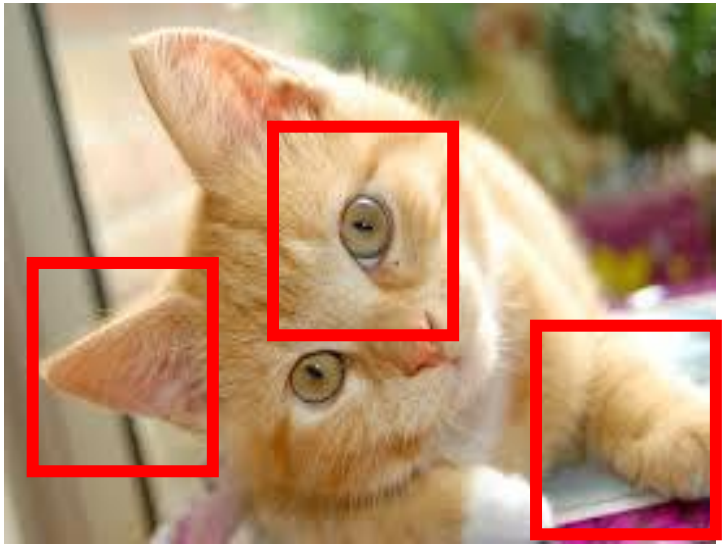
What does a "cat" look like?

(not referred to a specific image)

Local Explanation: Explain the Decision

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

Which component is critical?



Which component is critical for making decision?

模型的输入x，可以是图片，文字等

Object x ⟶ Image, text, etc.

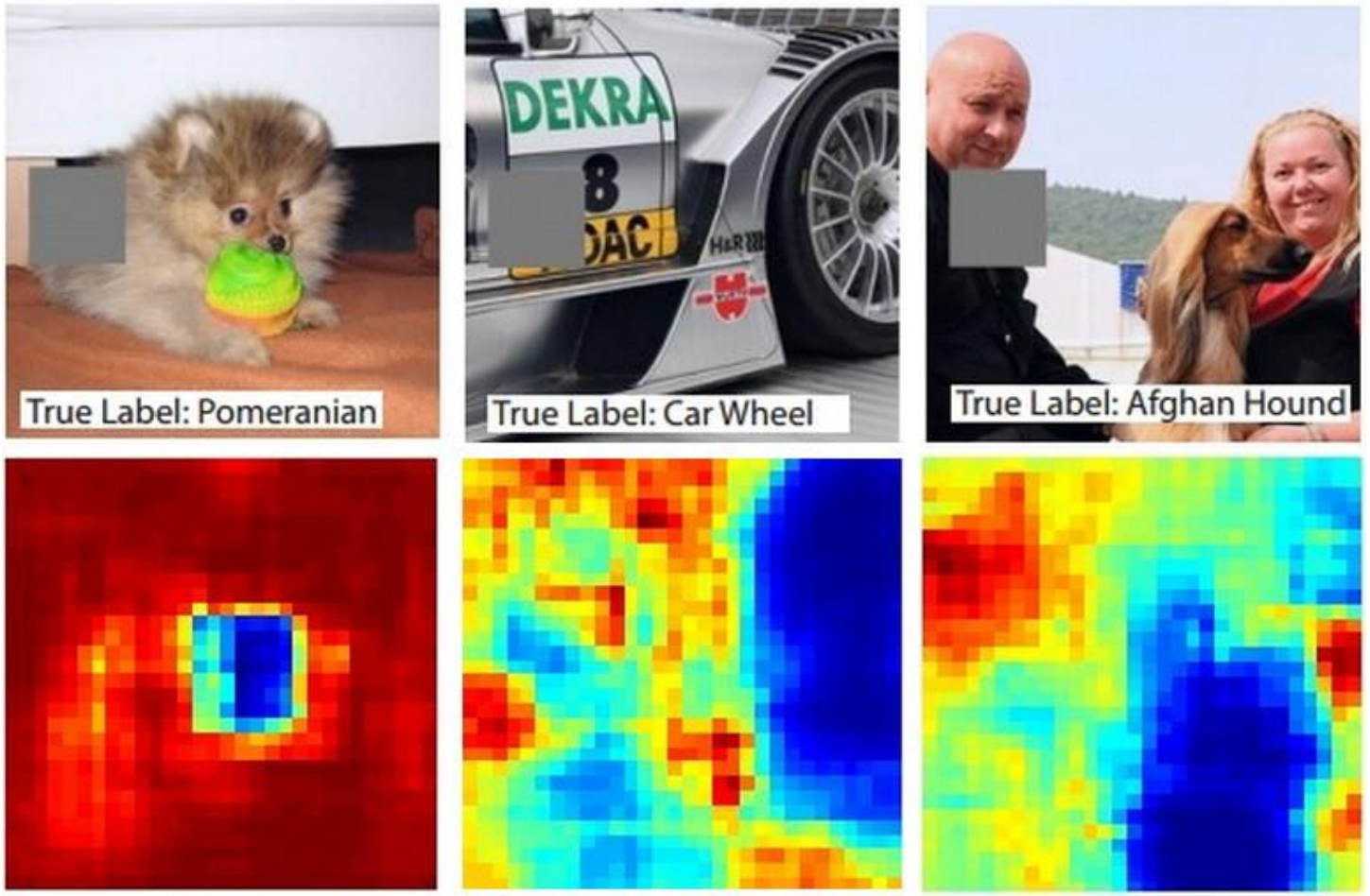
Components:

$$\{x_1, \cdots, x_n, \cdots, x_N\}$$

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

- Removing or modifying the components
- Large decision change

➡ Important component



放灰色方块

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

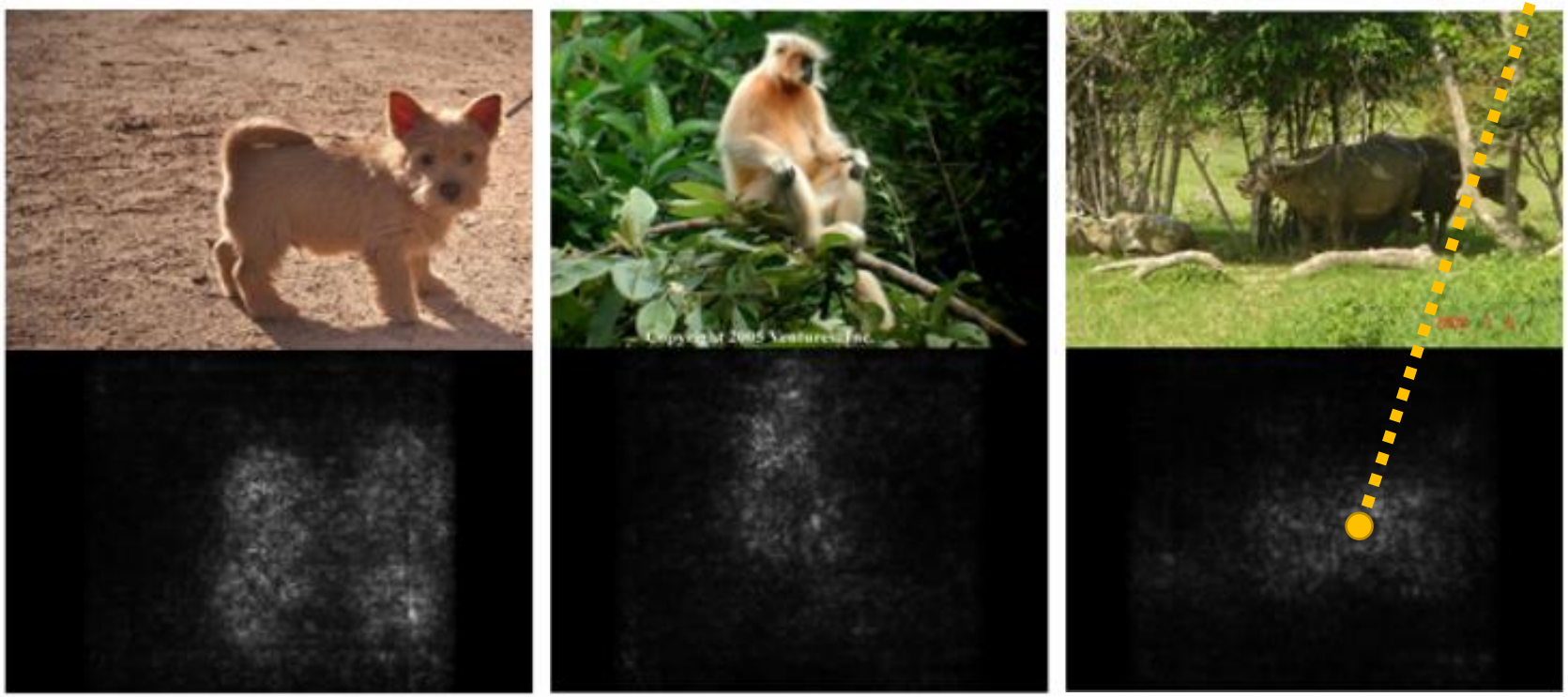
$\{x_1, \cdots, x_n, \cdots, x_N\}$ ➡ $\{x_1, \cdots, x_n + \Delta x, \cdots, x_N\}$ （对输入做小变化）

pixels

e ➡ $e + \Delta e$ （loss的差距）

$\left|\frac{\Delta e}{\Delta x}\right|$ ➡ $\left|\frac{\partial e}{\partial x_n}\right|$

loss of an example (the difference between model output and ground truth)



Saliency Map

越白色，值越大，这块越重要

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

Case Study: Pokémon v.s. Digimon

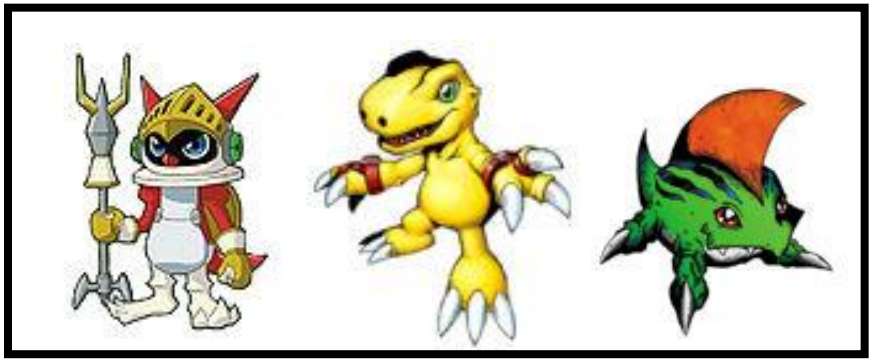
https://medium.com/@tyreeostevenson/teaching-a-computer-to-classify-anime-8c77bc89b881

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'))
```

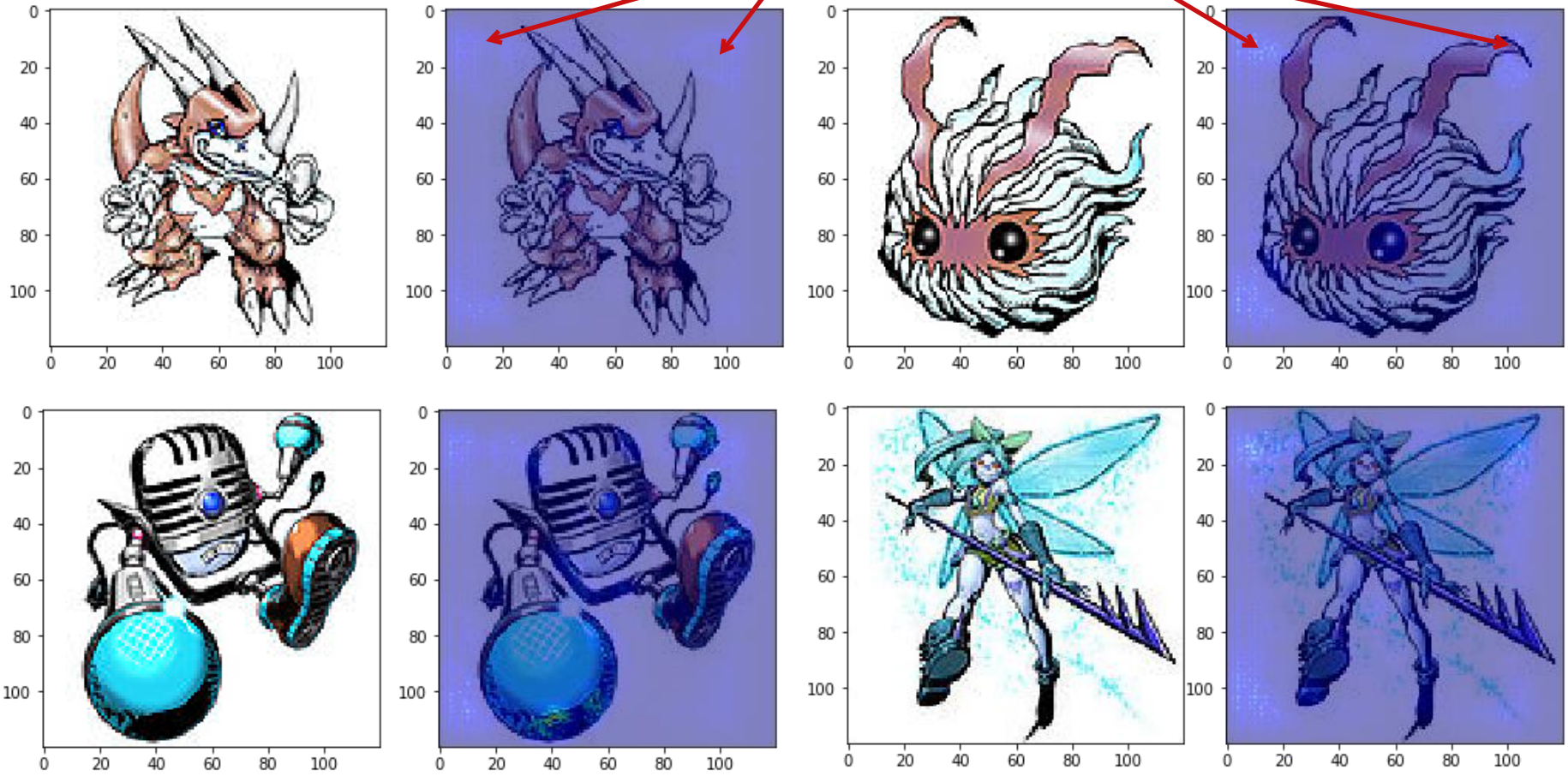
Training Accuracy: 98.9%

Testing Accuracy: 98.4%

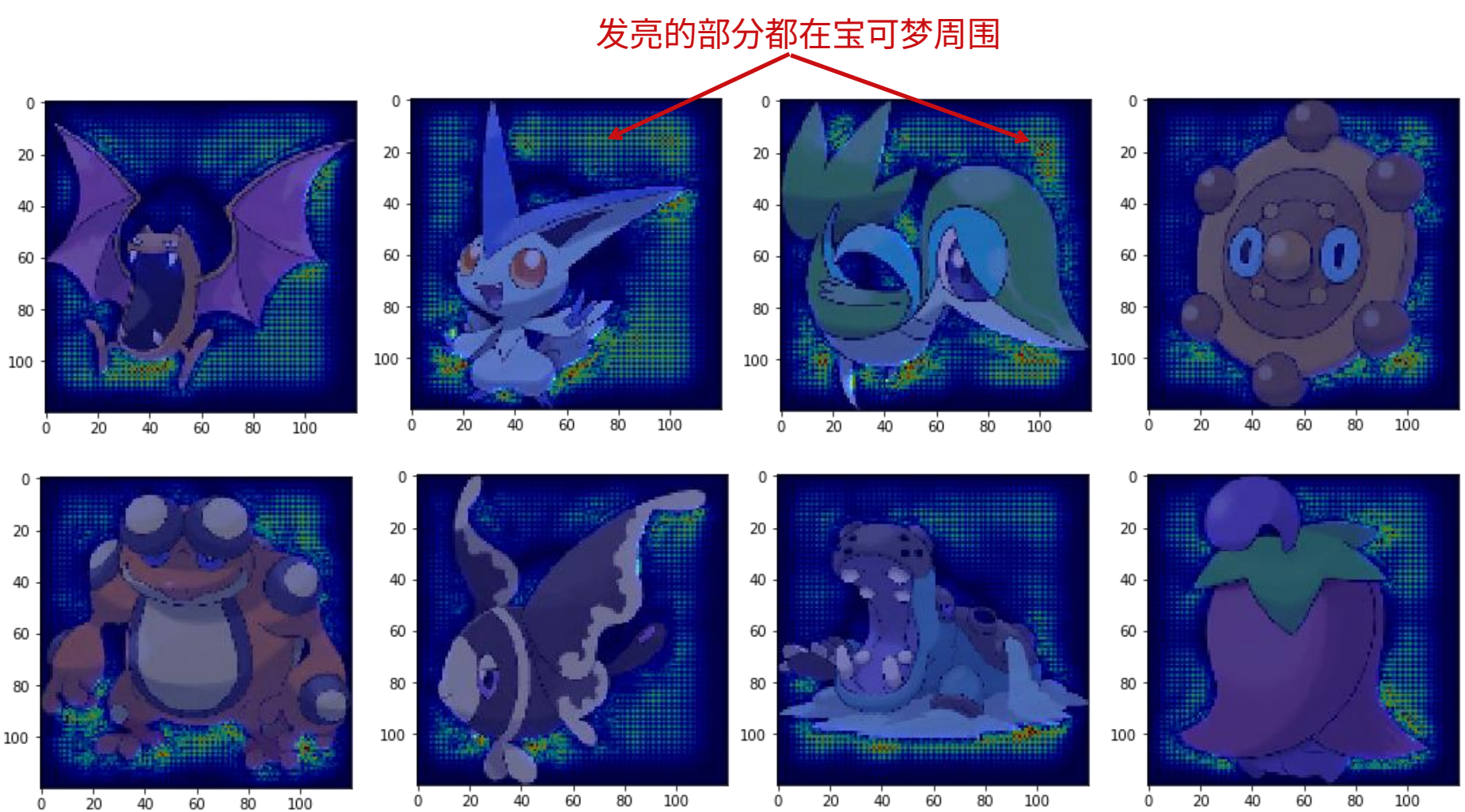
Amazing!!!!!!

Saliency Map

发亮的部分都在数码宝贝周围



Saliency Map



What Happened?

模型会根据数码宝贝和宝可梦的图片的背景颜色来判断，导致模型的准确率很高

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

loading the files

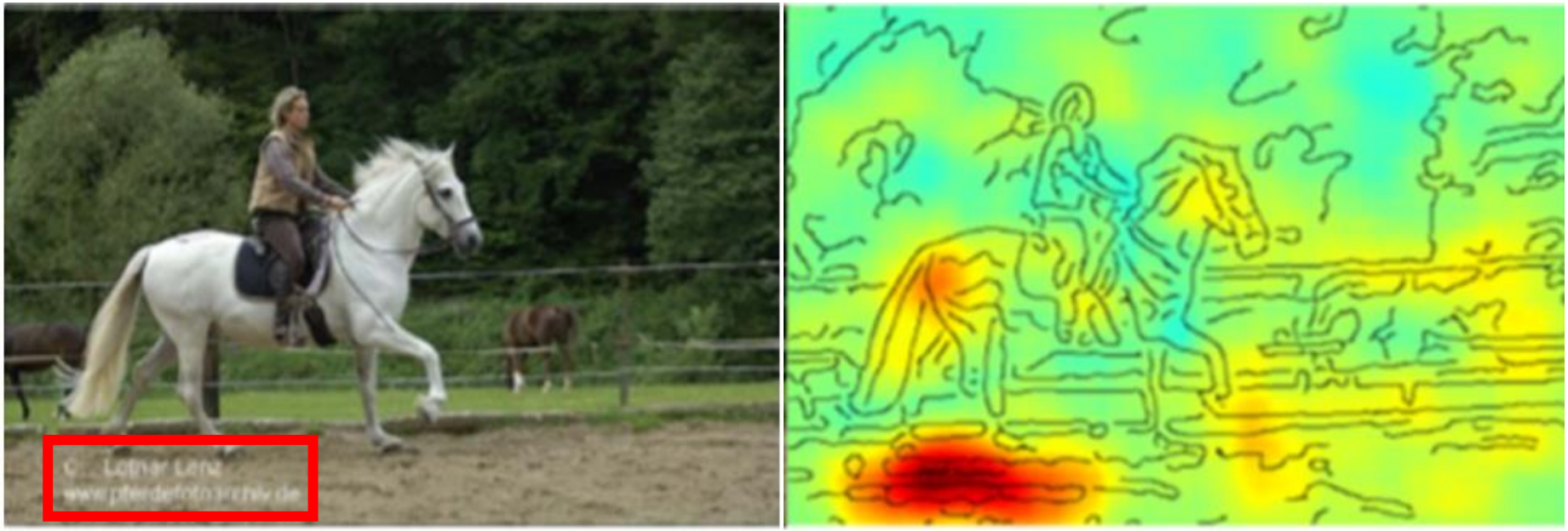
png files have transparent background

transparent background becomes black

Machine discriminates Pokémon and Digimon based on the background colors.

More Examples …

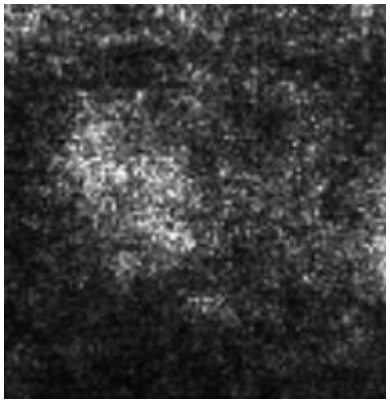
- PASCAL VOC 2007 data set 模型看重的是图片里的文字



This slide is from: GCPR 2017 Tutorial — W. Samek & K.-R. Müller

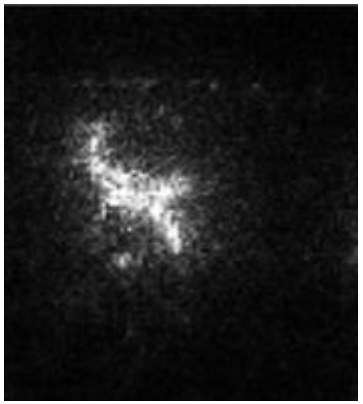
Limitation: Noisy Gradient

Gazelle
(瞪羚)



Typical

普通的saliency map的结果



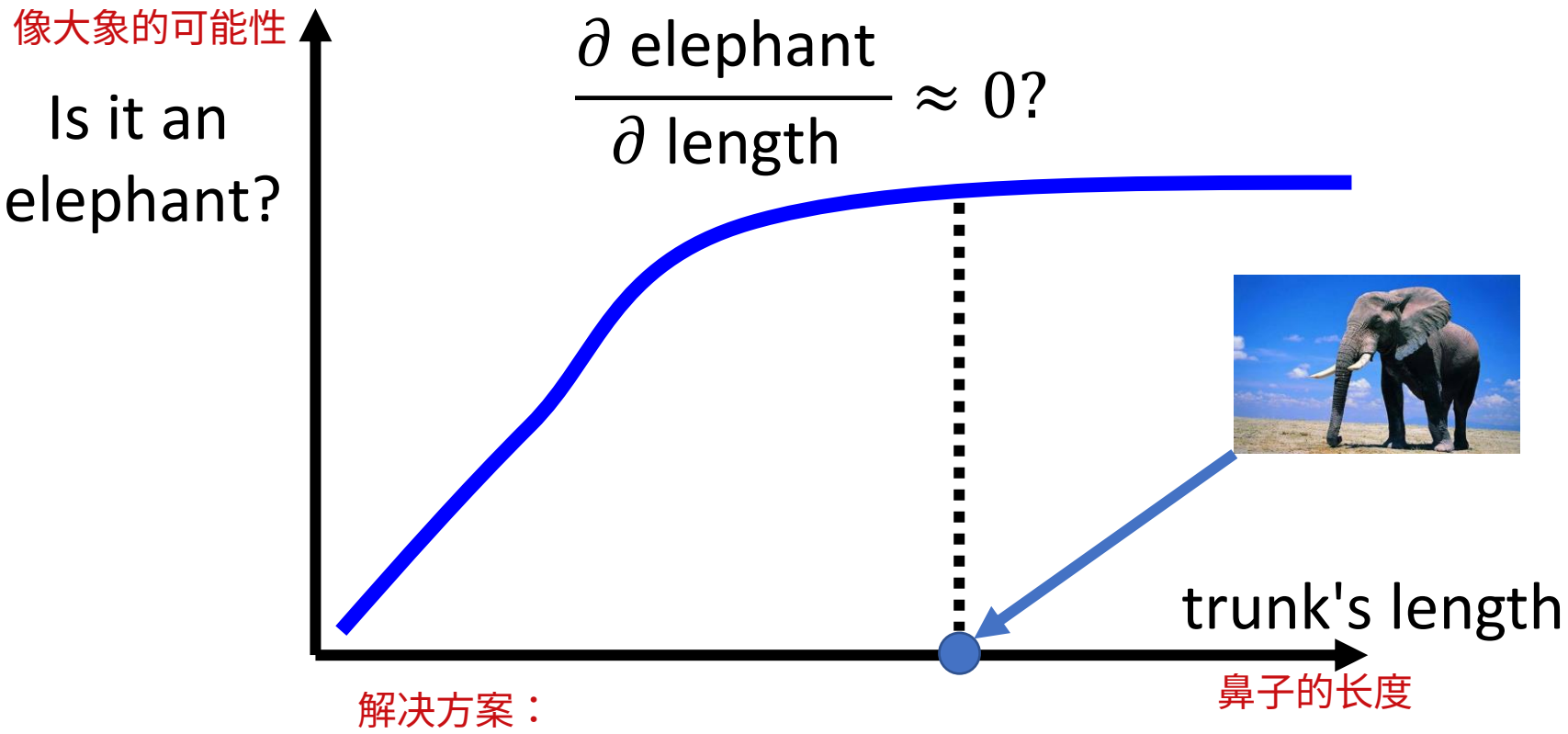
SmoothGrad

SmoothGrad: Randomly add noises[1] to the input image, get[2] saliency maps of the noisy images, and average them[3].

https://arxiv.org/abs/1706.03825

Limitation: Gradient Saturation

Gradient cannot always reflect importance

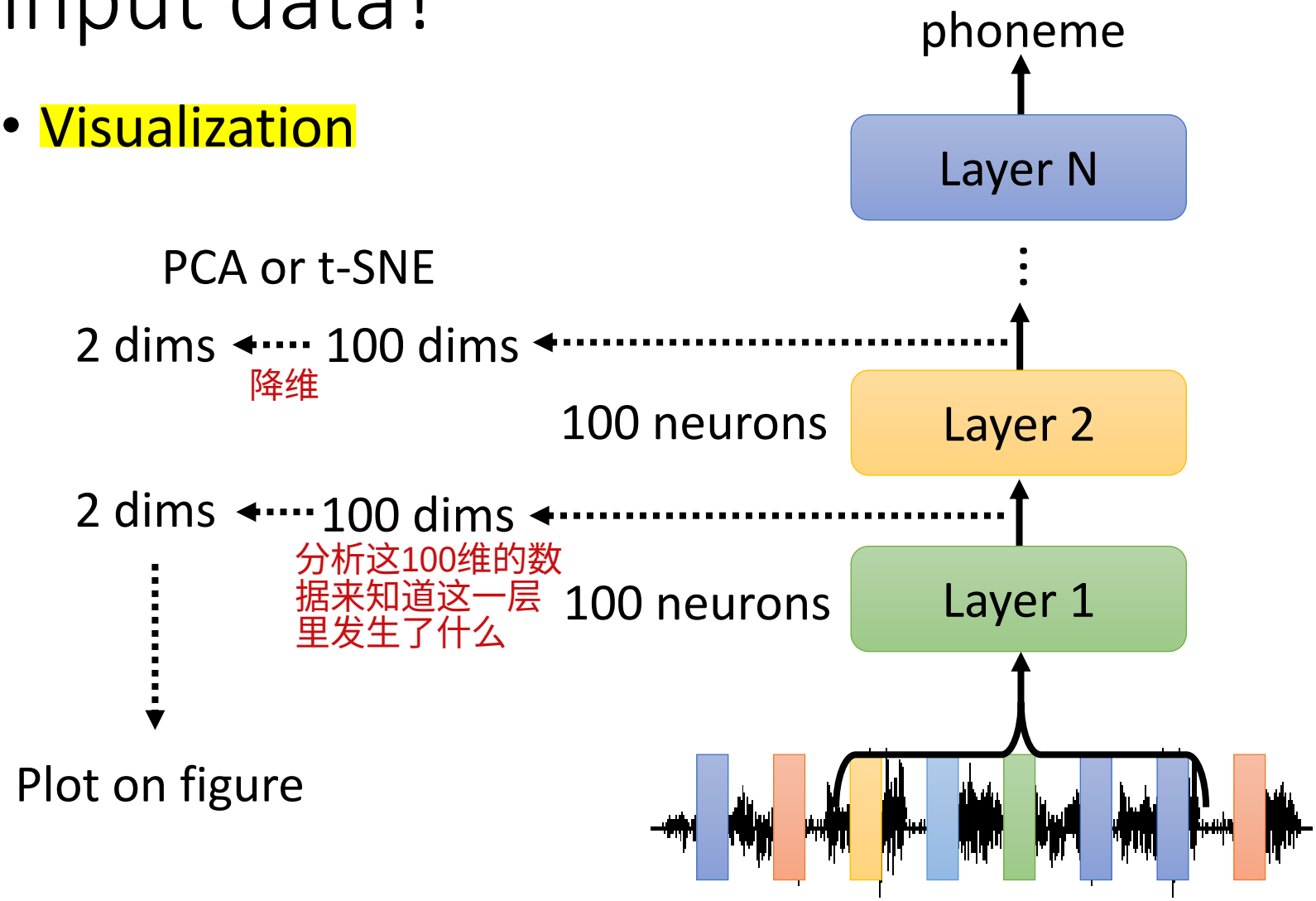


Alternative: Integrated gradient (IG)

https://arxiv.org/abs/1611.02639

How a network processes the input data?

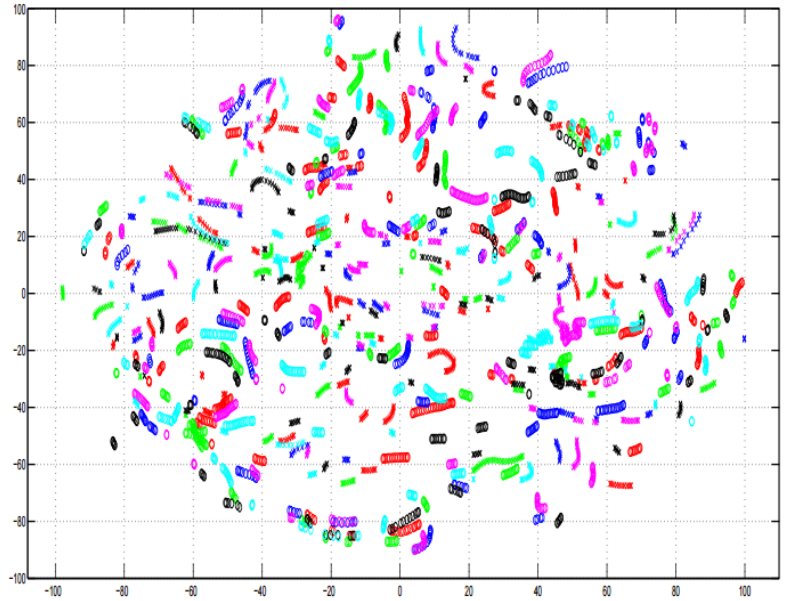
1 • Visualization



How a network processes the input data?

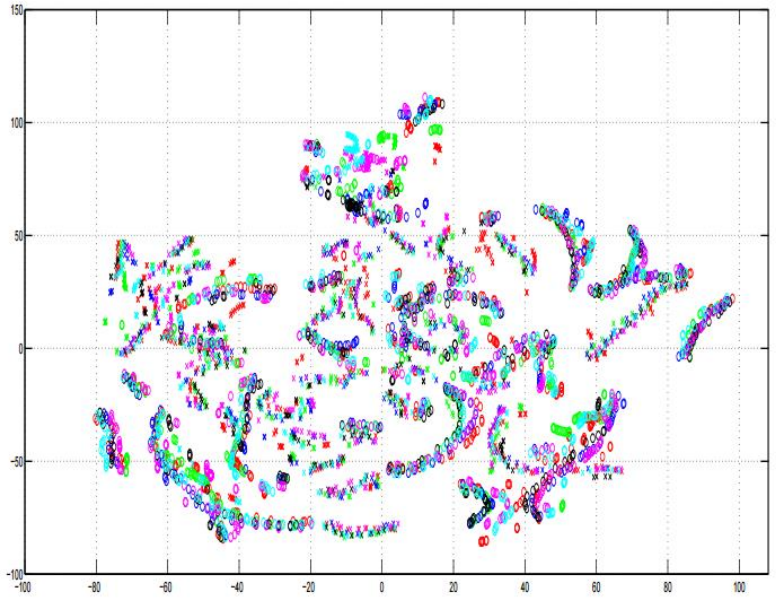
- Visualization
  Colors: speakers

A. Mohamed, G. Hinton, and G. Penn, "Understanding how Deep Belief Networks Perform Acoustic Modelling," in ICASSP, 2012.



Input Acoustic Feature (MFCC)

每个点代表一小段声音讯号
每个颜色代表了某个speaker

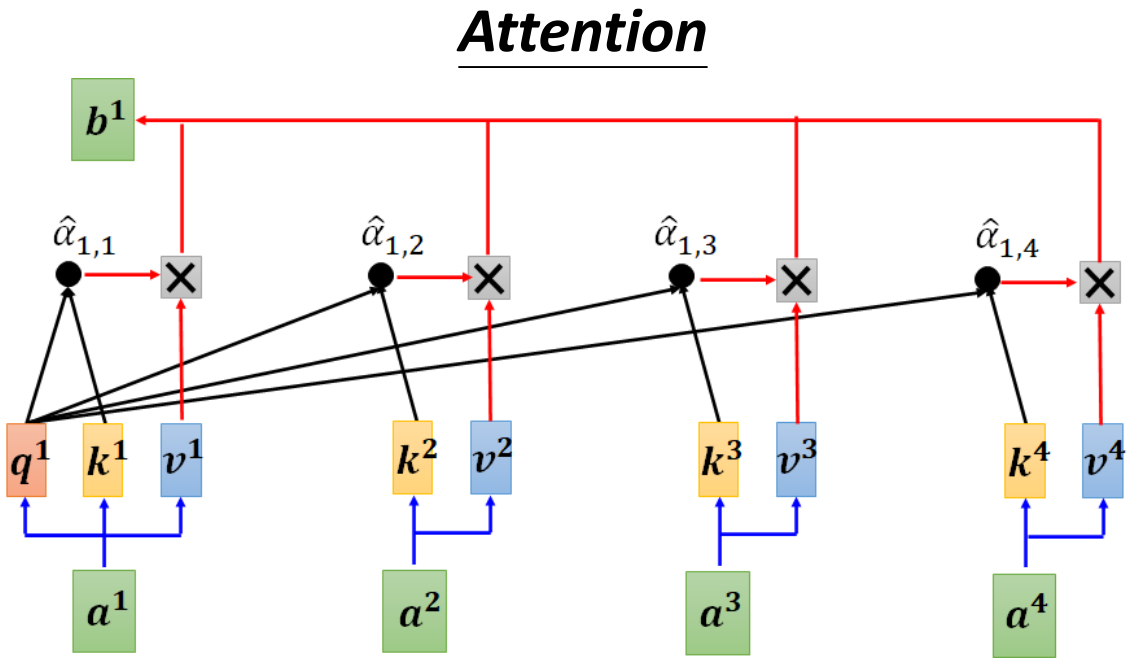


8-th Hidden Layer

每一条代表不同的speaker
说的同样的内容

How a network processes the input data?

- Visualization

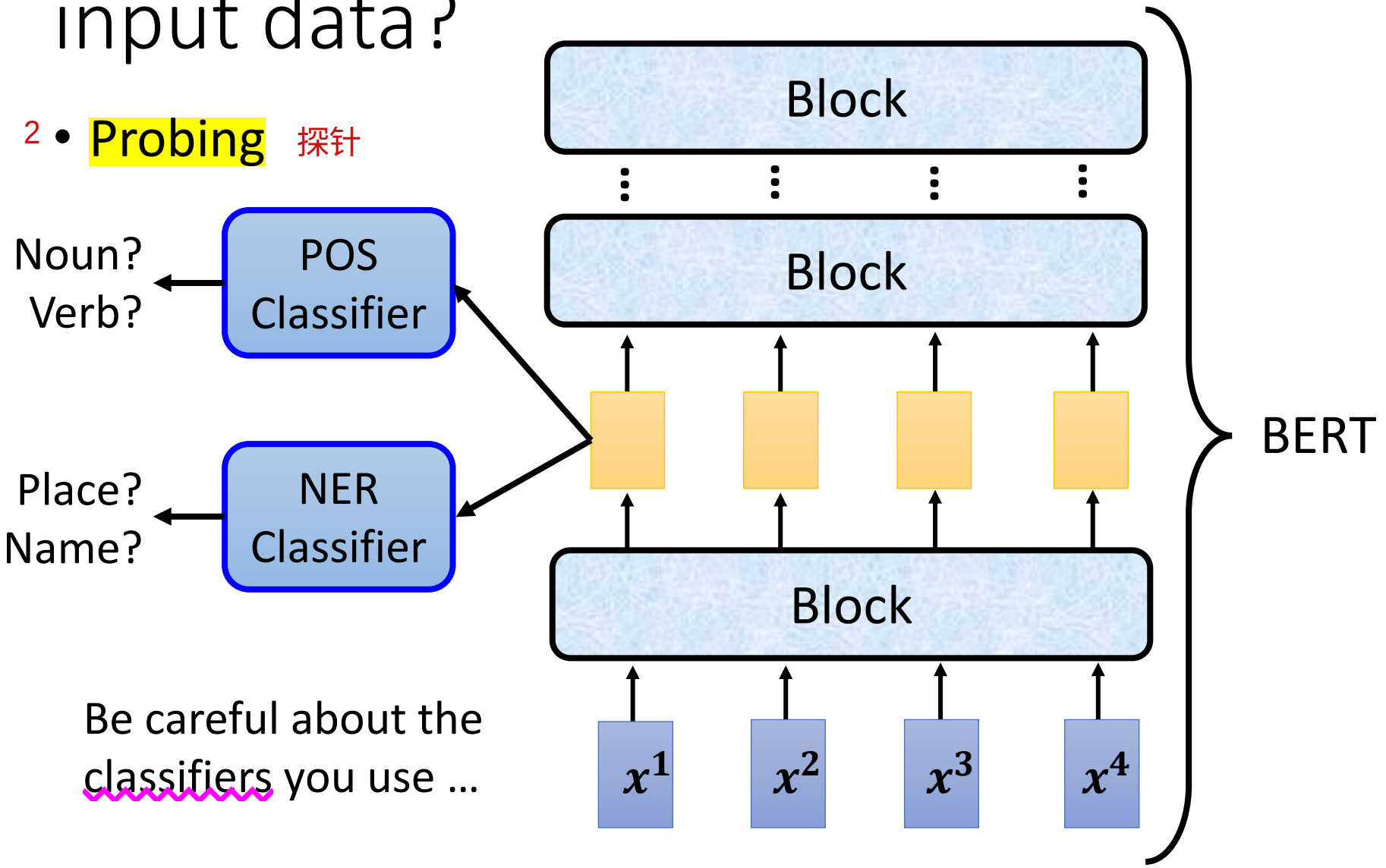


Attention is not Explanation
https://arxiv.org/abs/1902.10186

Attention is not not Explanation
https://arxiv.org/abs/1908.04626

How a network processes the input data?

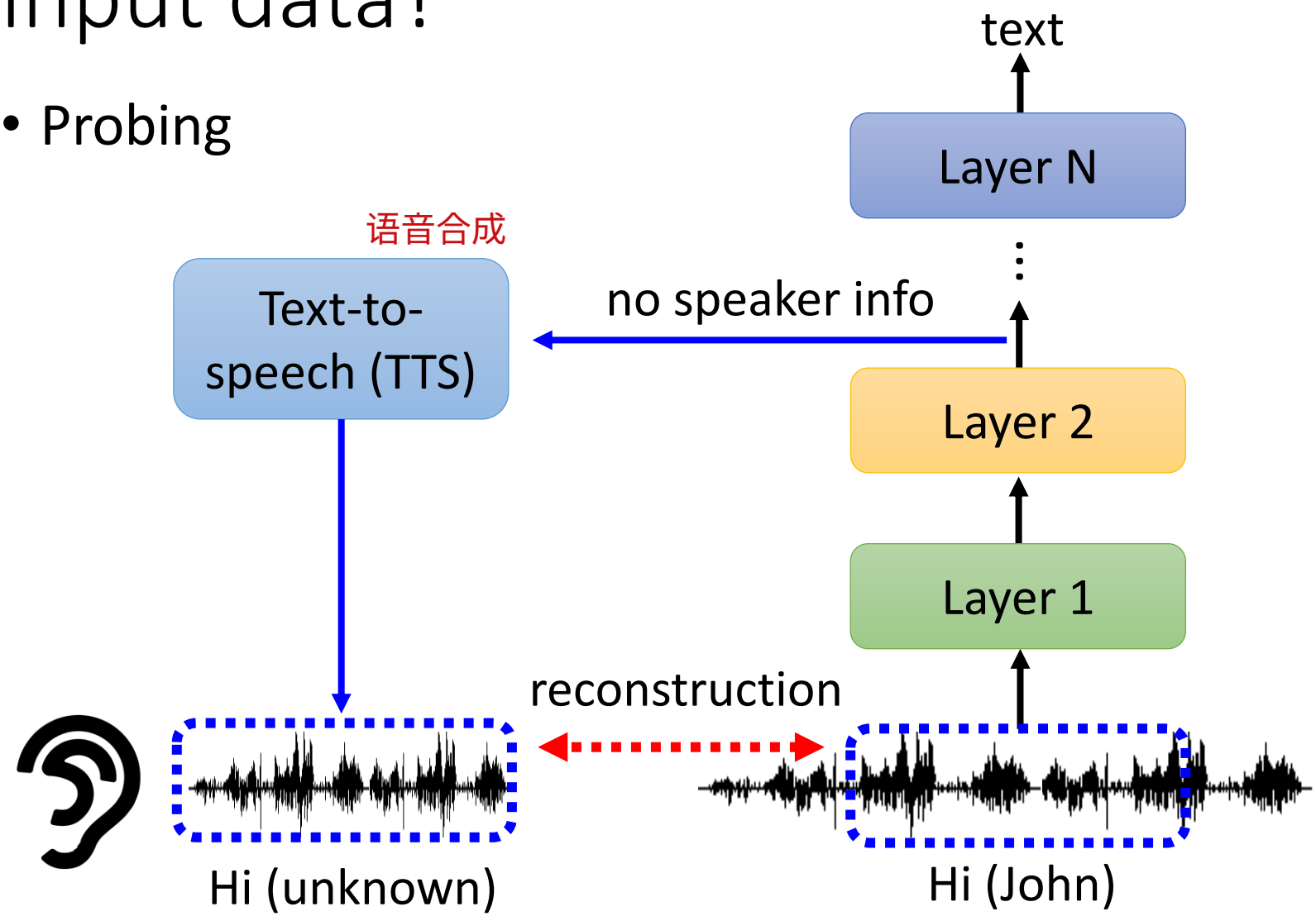
- Probing 探针



Be careful about the classifiers you use …

How a network processes the input data?

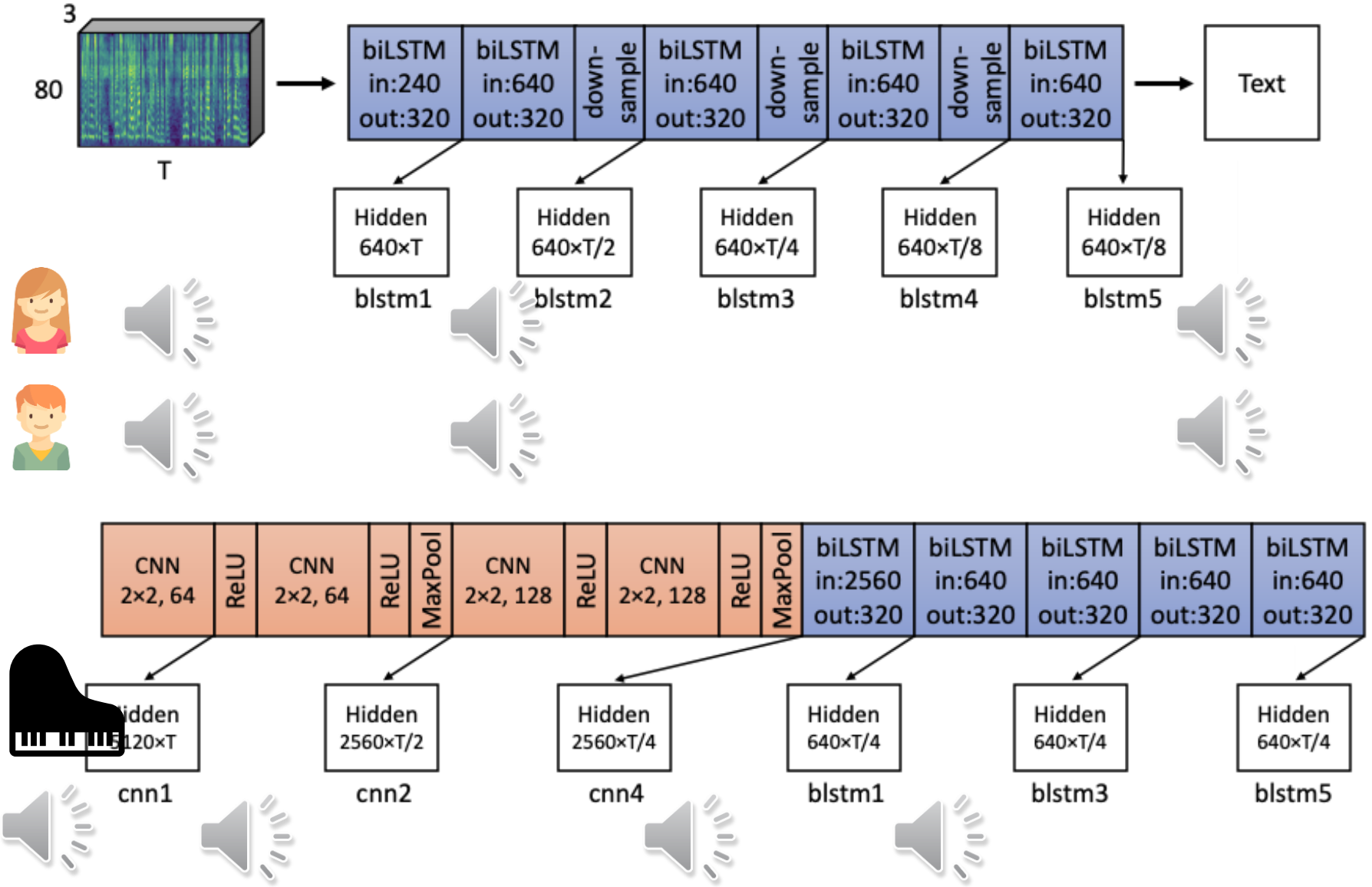
- Probing



What does a network layer hear? Analyzing hidden representations of end-to-end ASR through speech synthesis

https://arxiv.org/abs/1911.01102

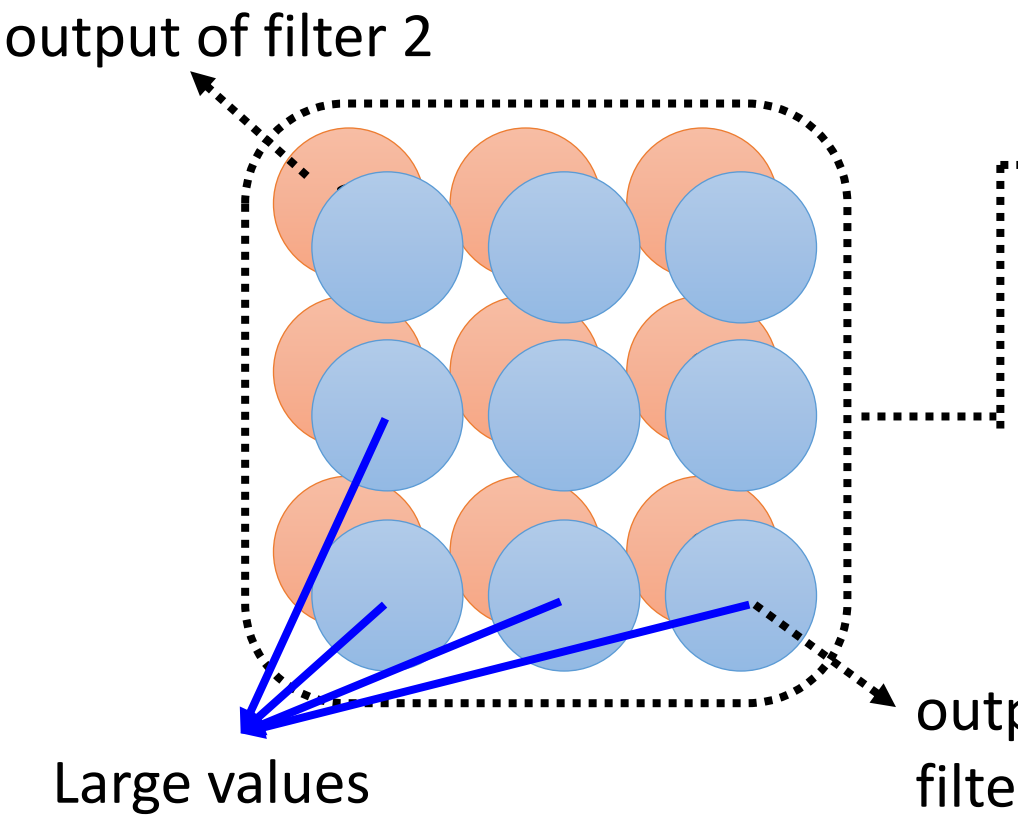
https://youtu.be/6gtn7H-pWr8

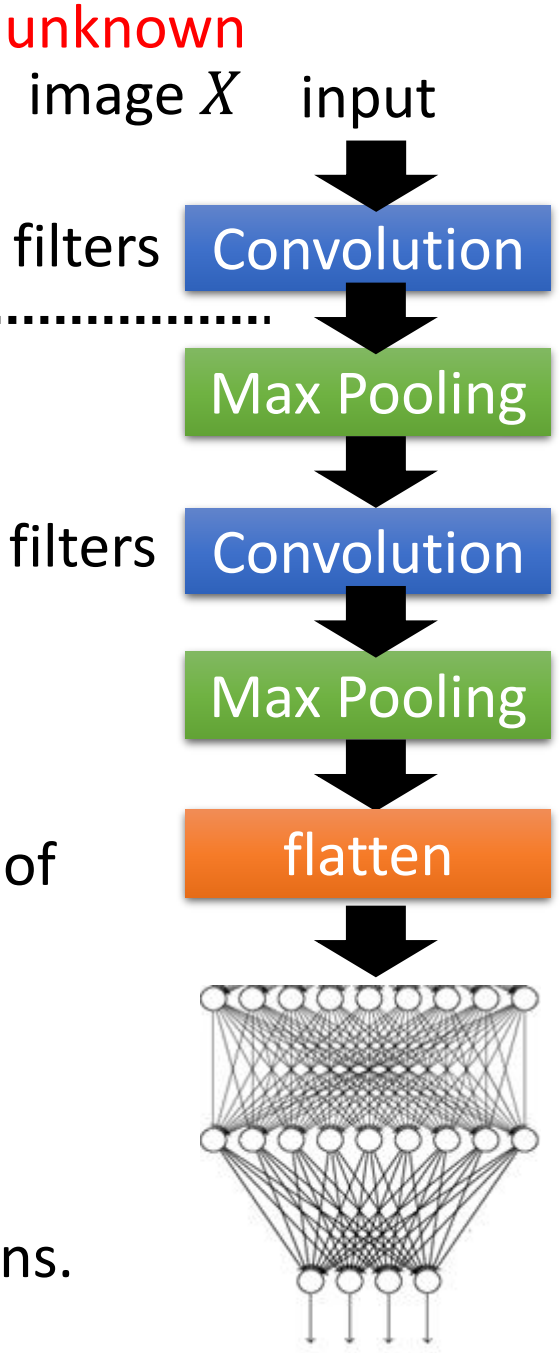


GLOBAL EXPLANATION: EXPLAIN THE WHOLE MODEL

Question: What does a “cat” look like?

What does a filter detect?

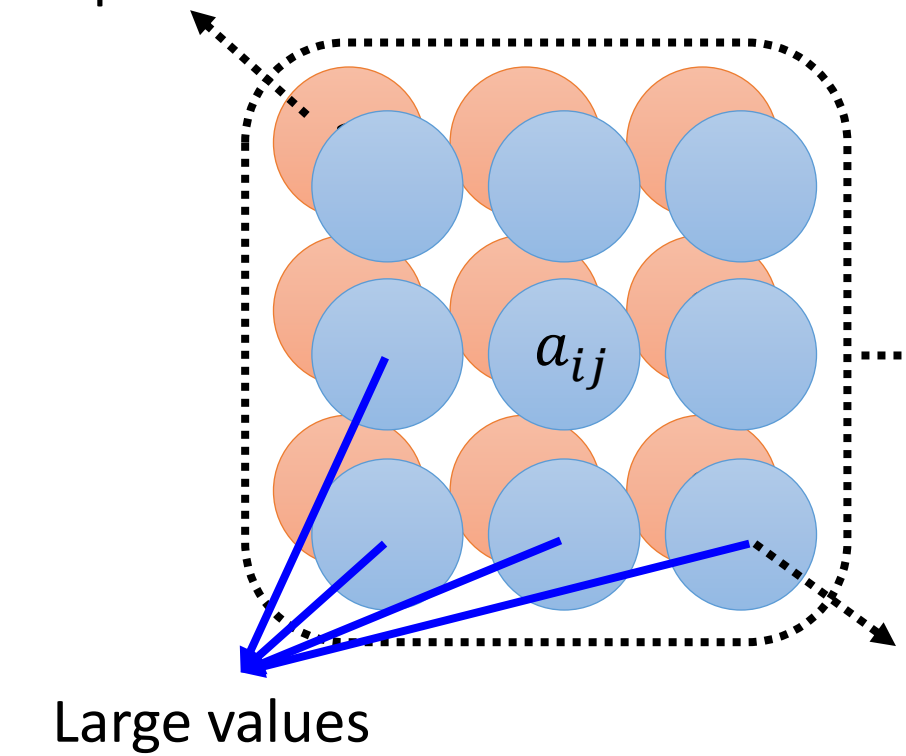


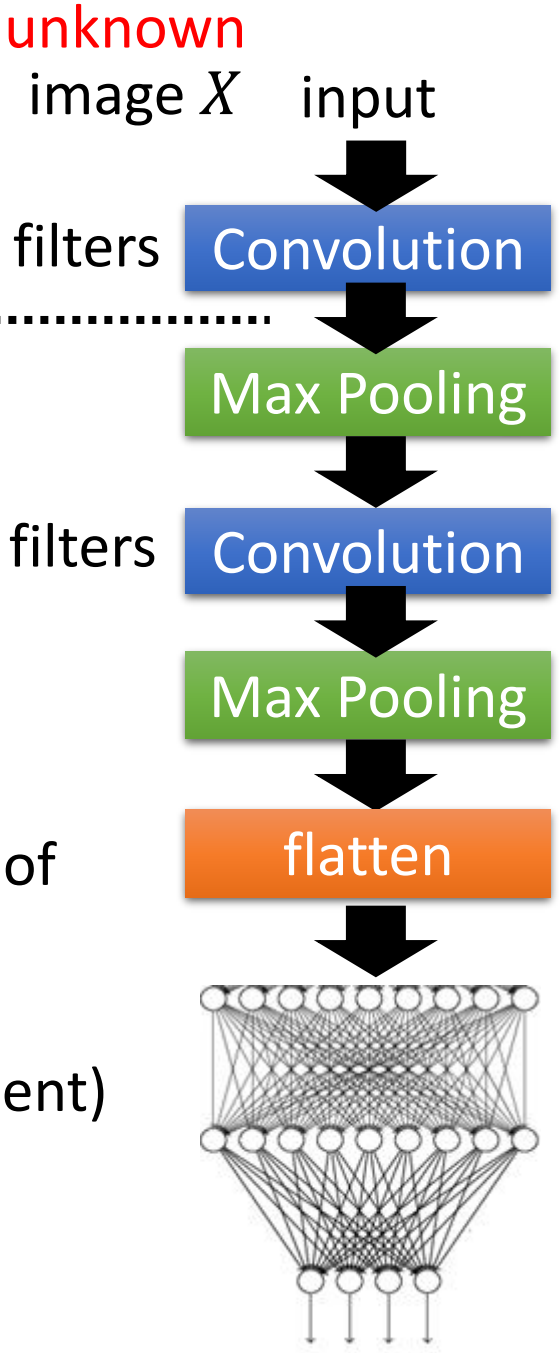


➡ Image X contains the patterns filter 1 can detect.

Let's **create** an image including the patterns.

What does a filter detect?





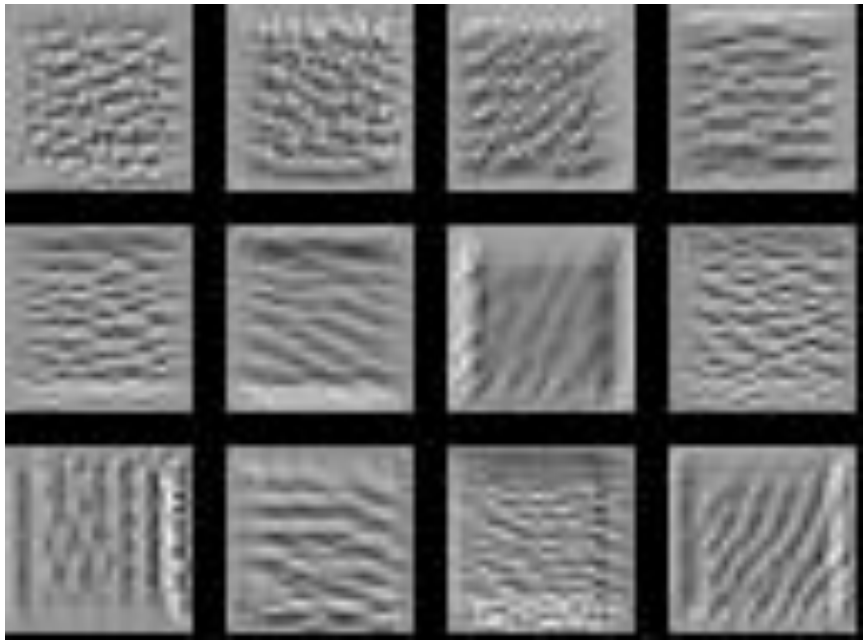
$$X^* = arg\max_X \sum_i \sum_j a_{ij}$$ (gradient ascent)

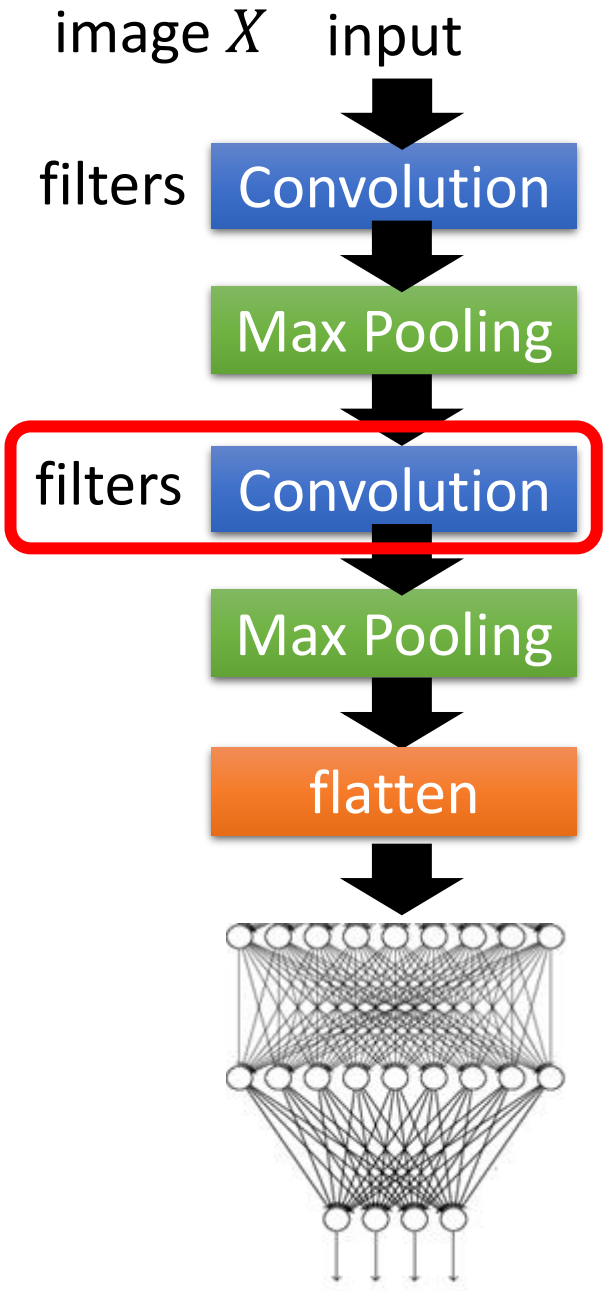
The image contains the patterns filter 1 can detect.

What does a filter detect?

E.g., Digit classifier

X^* for each filter

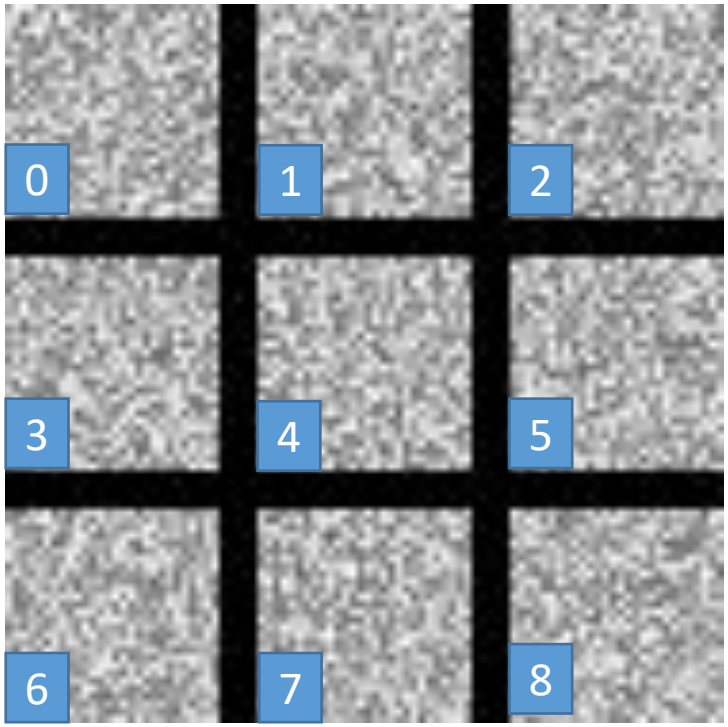




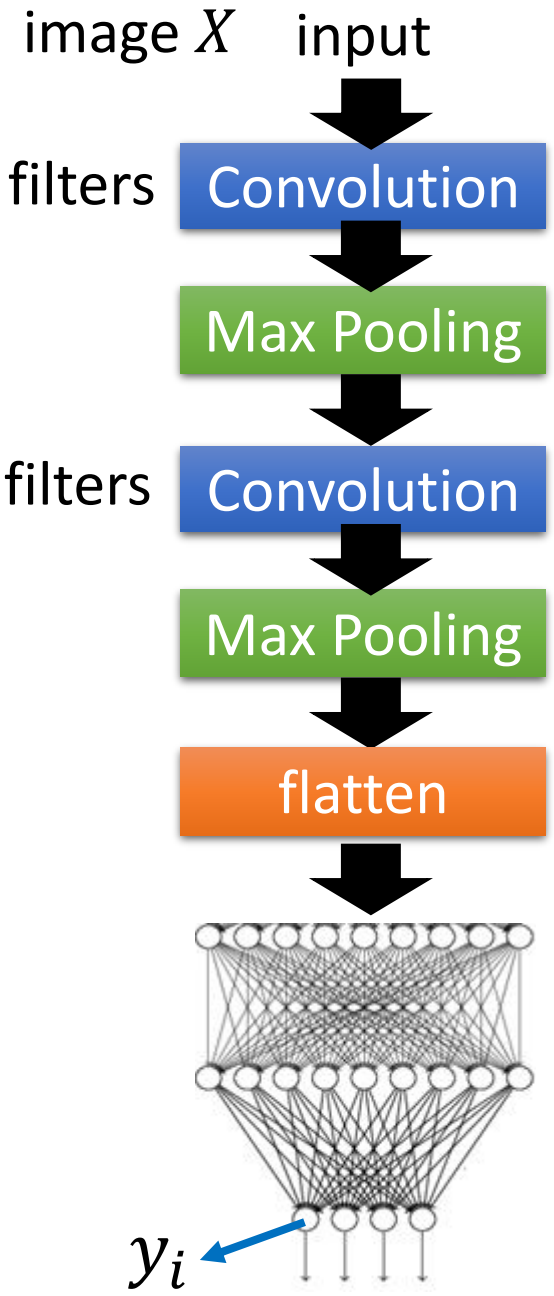
What does a digit look like for CNN?

E.g., Digit classifier

$$X^* = arg\max_X y_i$$ Can we see digits?



Surprise? Consider adversarial attack!



What does a digit look like for CNN?

Find the image that maximizes class probability

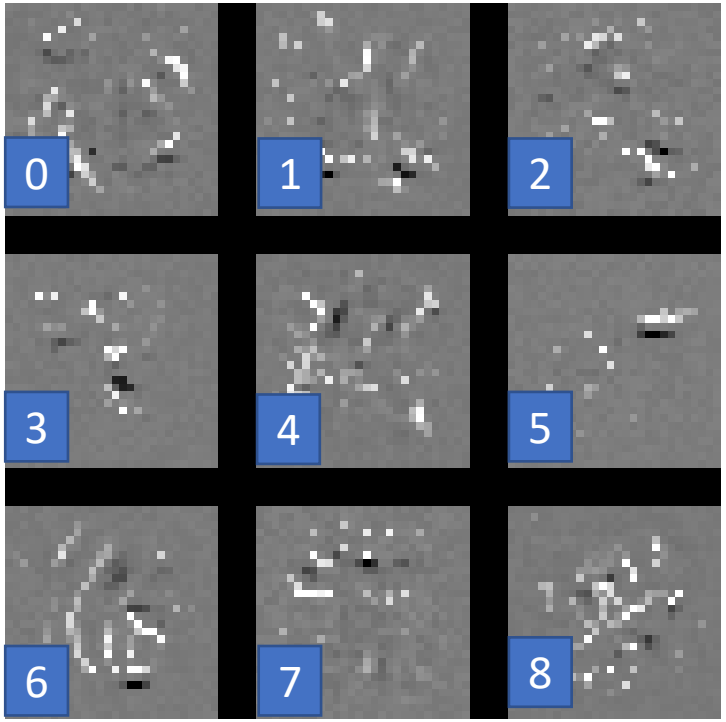
$$X^* = arg\max_X y_i$$

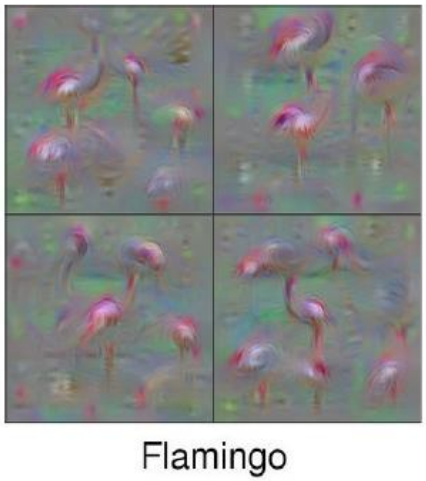
The image should looks like a digit.

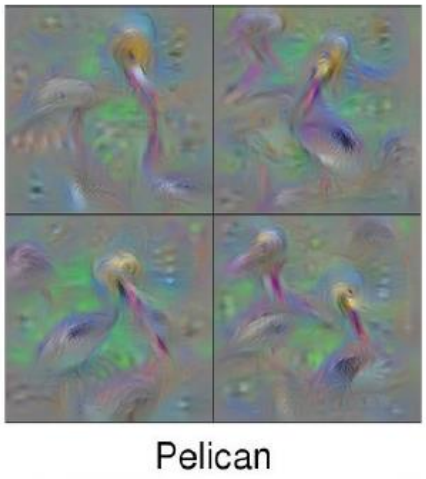
$$X^* = arg\max_X y_i + R(X)$$

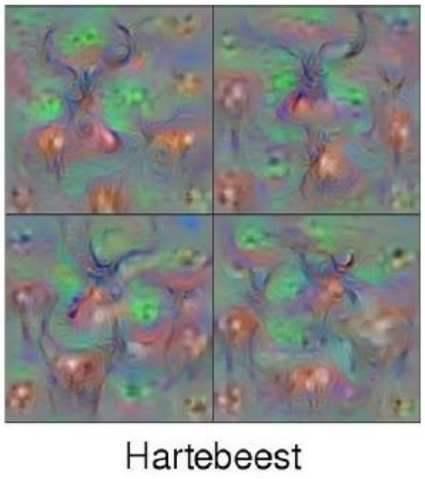
$$R(X) = -\sum_{i,j} |X_{ij}|$$

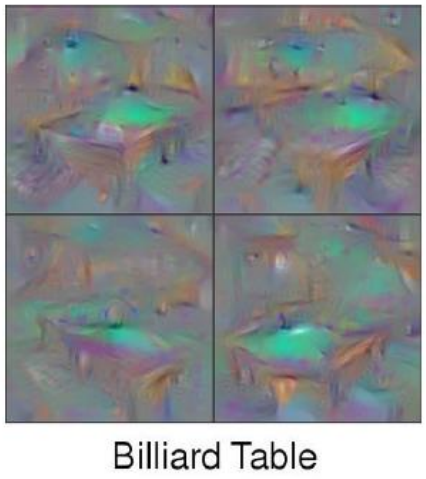
How likely X is a digit

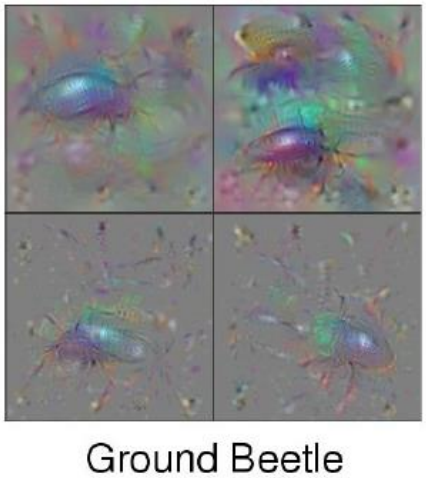


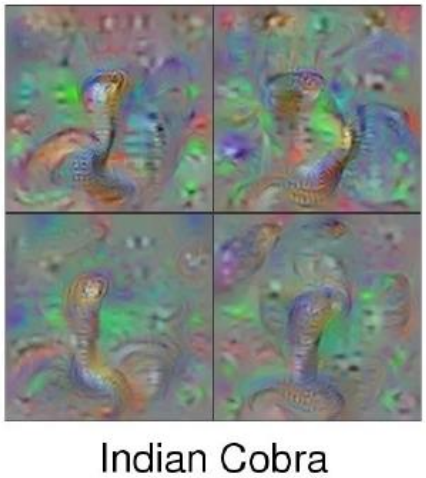

Flamingo

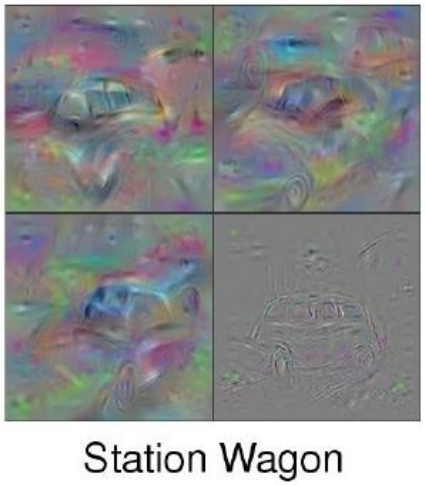

Pelican

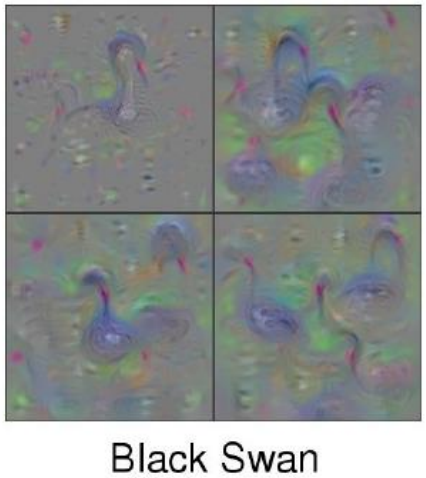

Hartebeest


Billiard Table


Ground Beetle


Indian Cobra


Station Wagon


Black Swan

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 $X = G(z)$

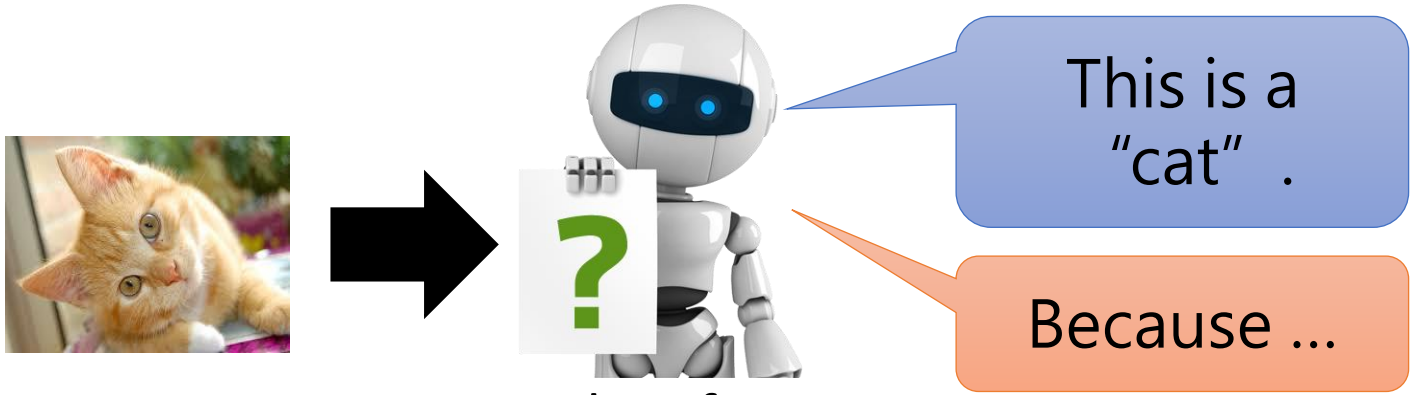
z → Image Generator → Image X → Image Classifier → y

$$X^* = arg\max_X y_i \Rightarrow z^* = arg\max_z y_i$$

Show image: $X^* = G(z^*)$

https://arxiv.org/abs/1612.00005

Concluding Remarks



Local Explanation

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

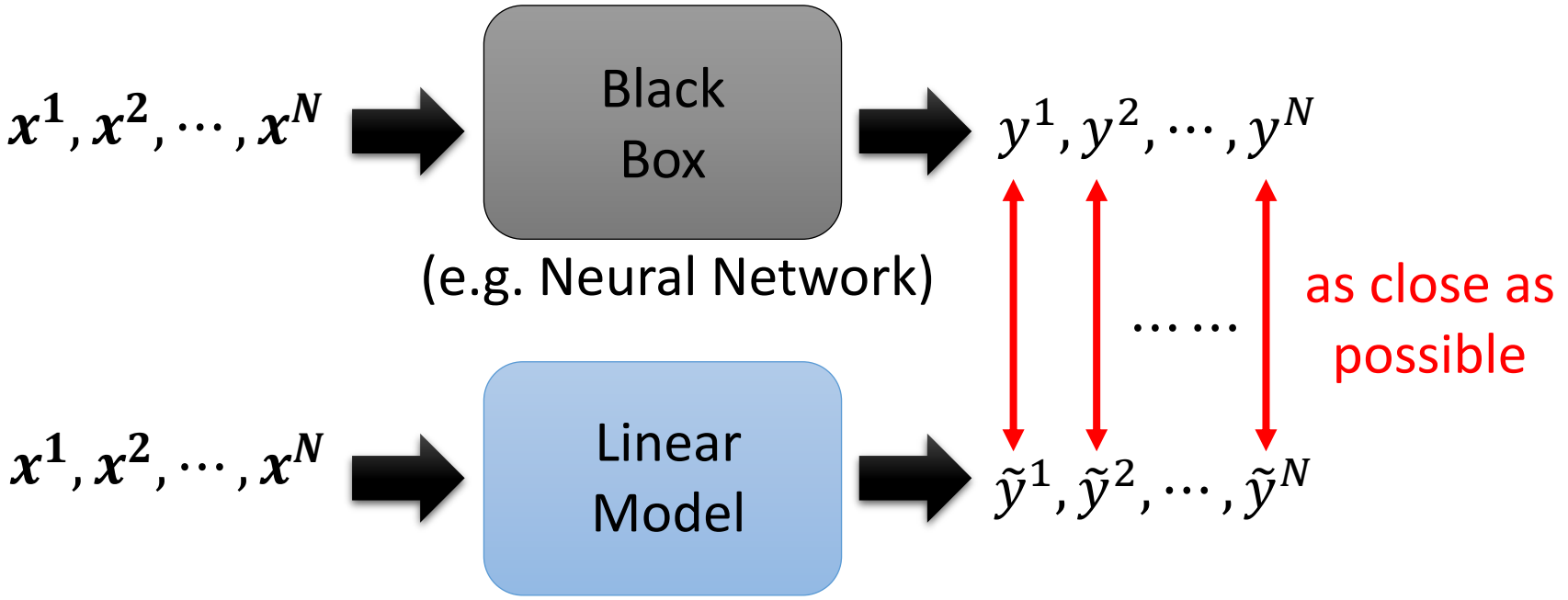
Global Explanation

What does a "cat" look like?

(not referred to a specific image)

Outlook

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



Local Interpretable Model-Agnostic Explanations (LIME)

https://youtu.be/K1mWgthGS-A
https://youtu.be/OjqIVSwly4k