MAE 688: Machine Learning for Mechanical Engineers

Lecture 8- Transfer Learning and Explainable AI

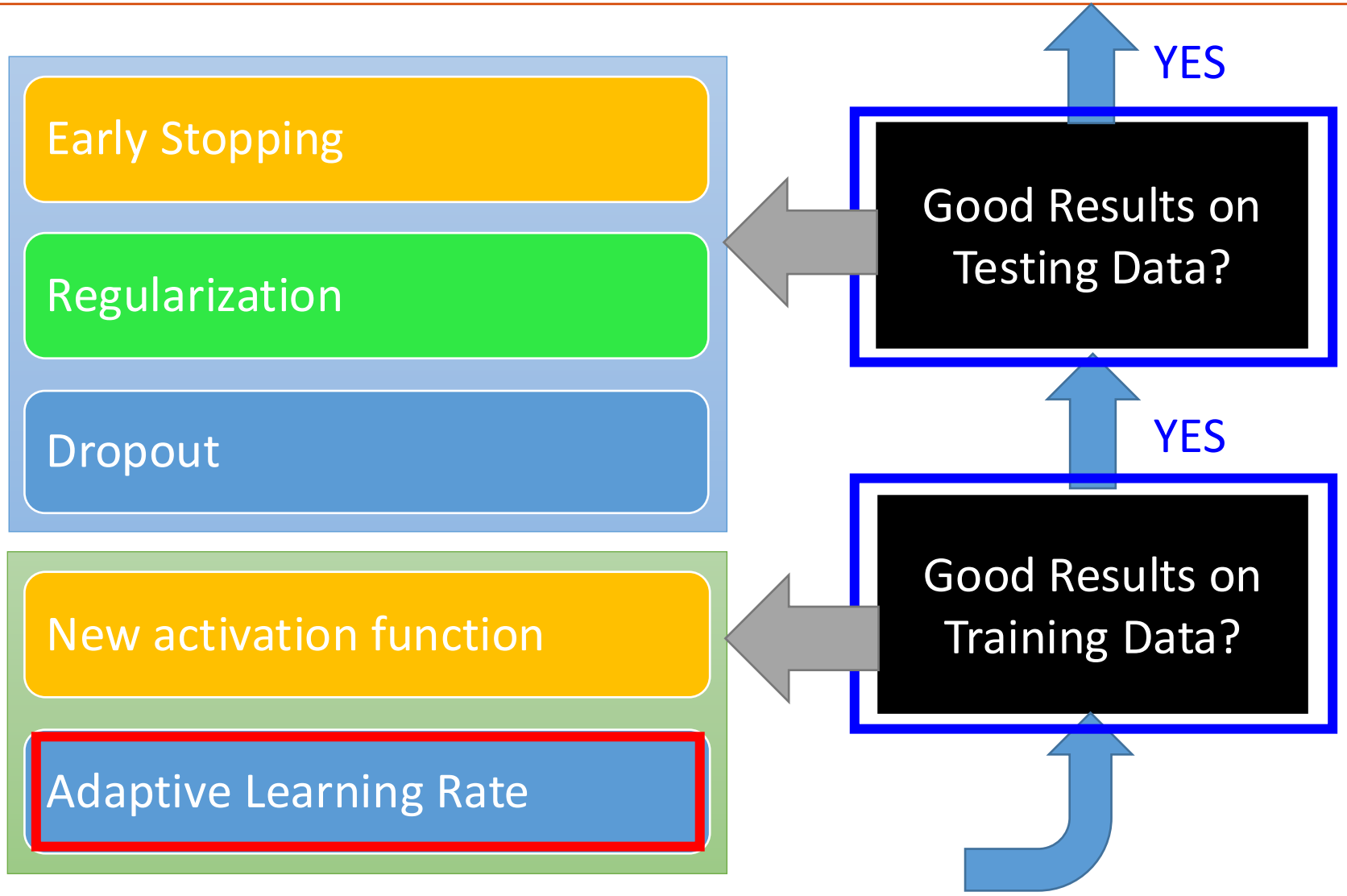
Dr. Bing Dong

Director, Built Environment Science and Technology (BEST) Lab
Professor
Mechanical and Aerospace Engineering

Syracuse University

Mach 28, 2023

Review NN Training Tips:



What is Transfer Learning

"You need a lot of a data if you want to train/use DNN"

What is Transfer Learning

What is Transfer Learning

Dog/Cat
Classifier

cat

dog

Data ***not directly related to*** the task considered

elephant

tiger

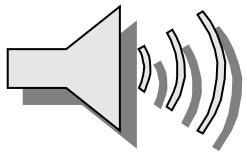
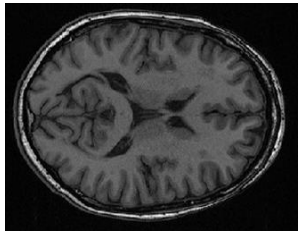
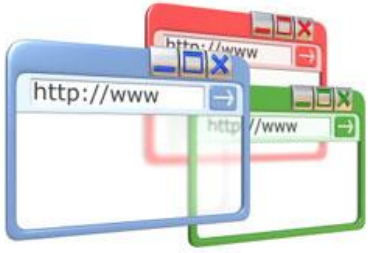
Similar domain, different tasks

dog

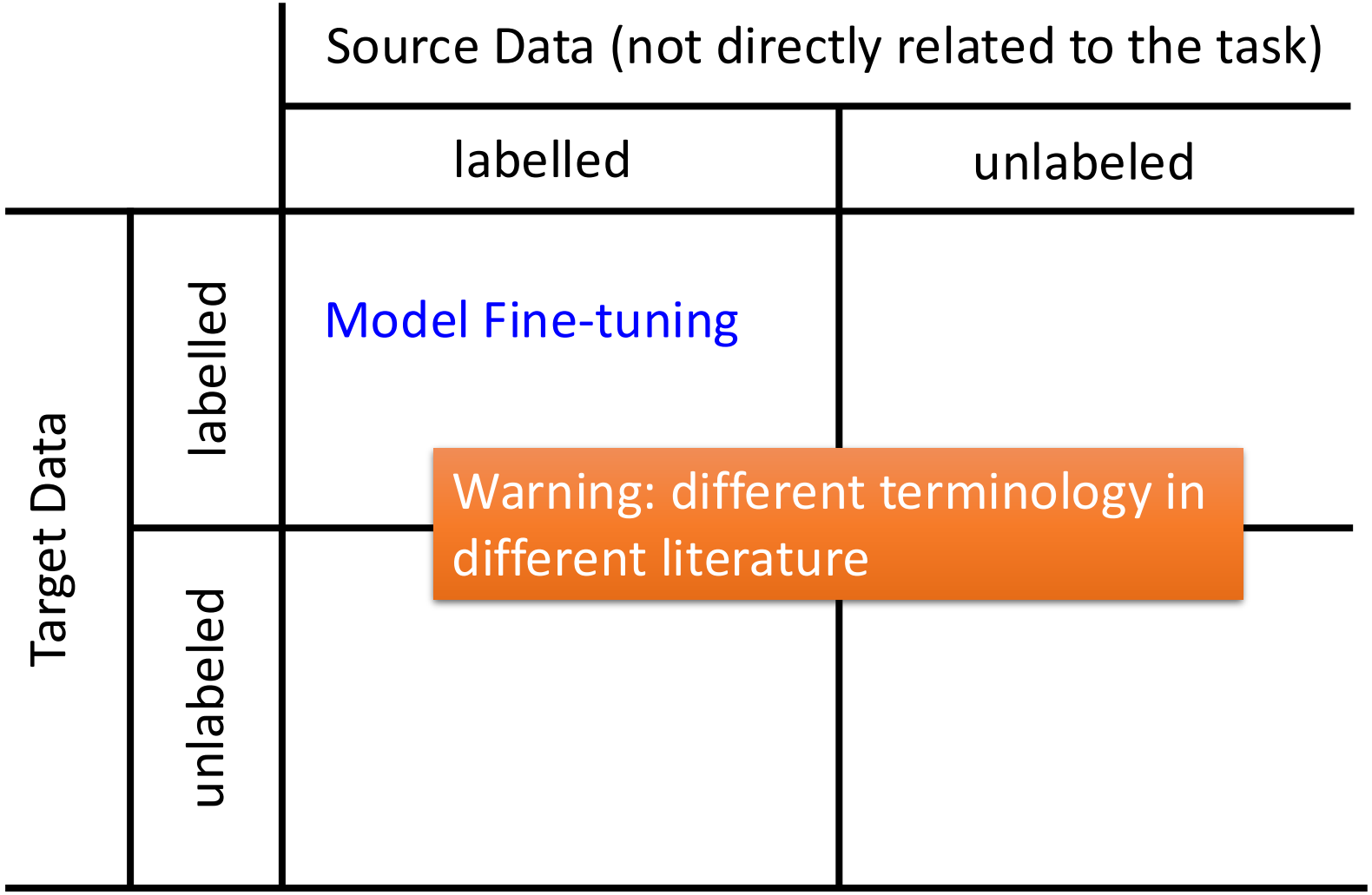
cat

Different domains, same task

Application of Transfer Learning

	Task Considered	Data not directly related
Speech Recognition	 Chinese	YouTube English Japanese ……
Image Recognition	 Medical Images	
Text Analysis	 Specific domain	 Webpages

What is Transfer Learning

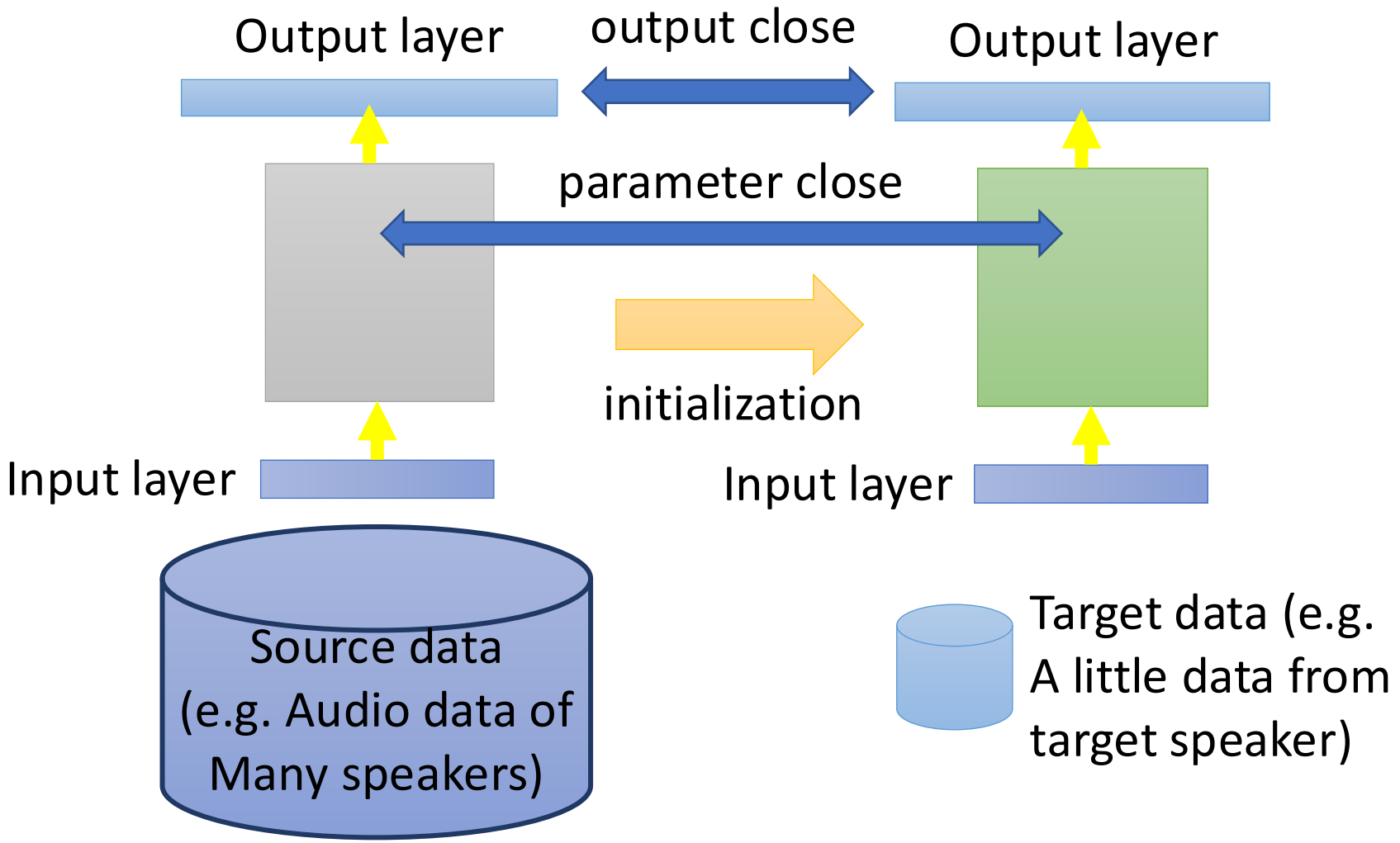


Model Fine-tuning

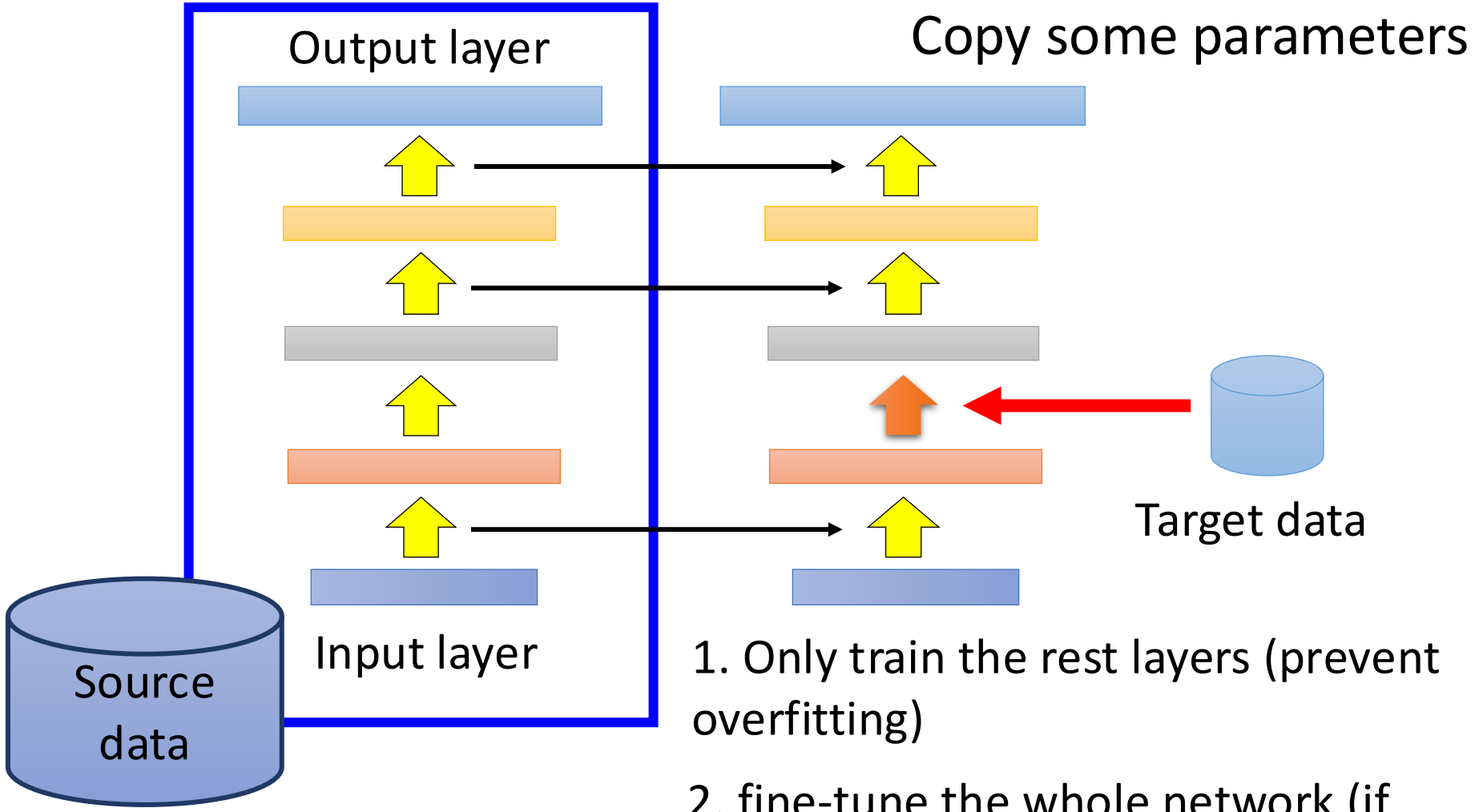
- Task description
  - Source data: (x^s, y^s) ⬅ A large amount
  - Target data: (x^t, y^t) ⬅ Very little
- Example: (supervised) speaker adaption
  - Source data: audio data and transcriptions from many speakers
  - Target data: audio data and its transcriptions of specific user
- Idea: training a model by source data, then fine-tune the model by target data
  - Challenge: only limited target data, so be careful about overfitting

One-shot learning: only a few examples in target domain

Conservative Training



Layer Transfer

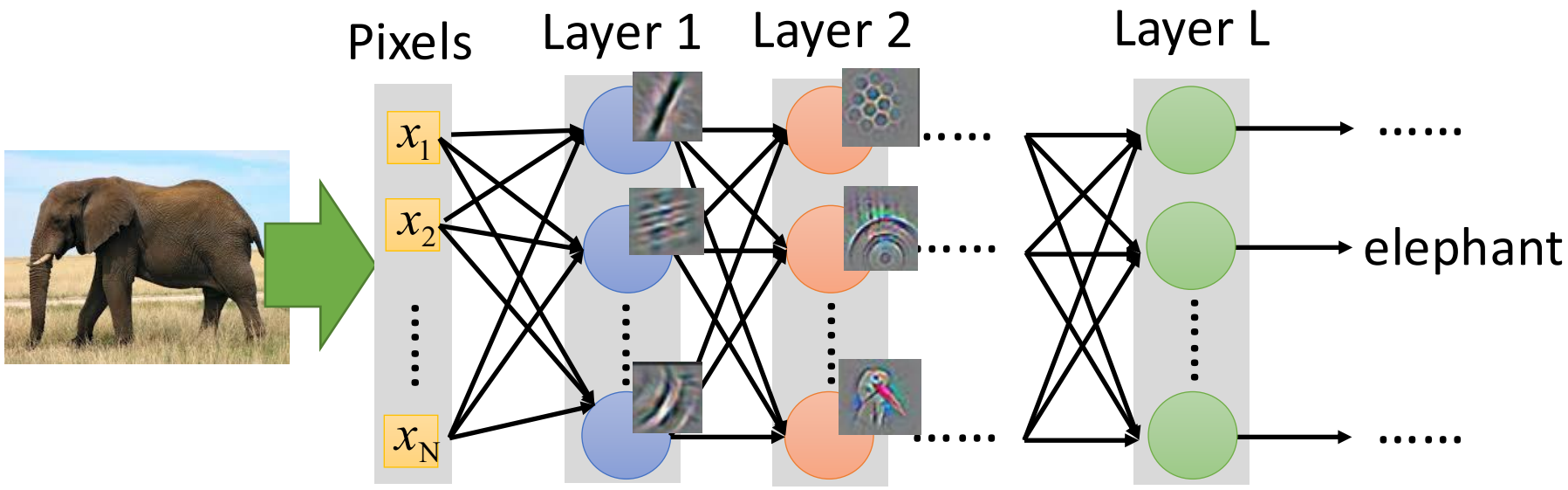


1. Only train the rest layers (prevent overfitting)

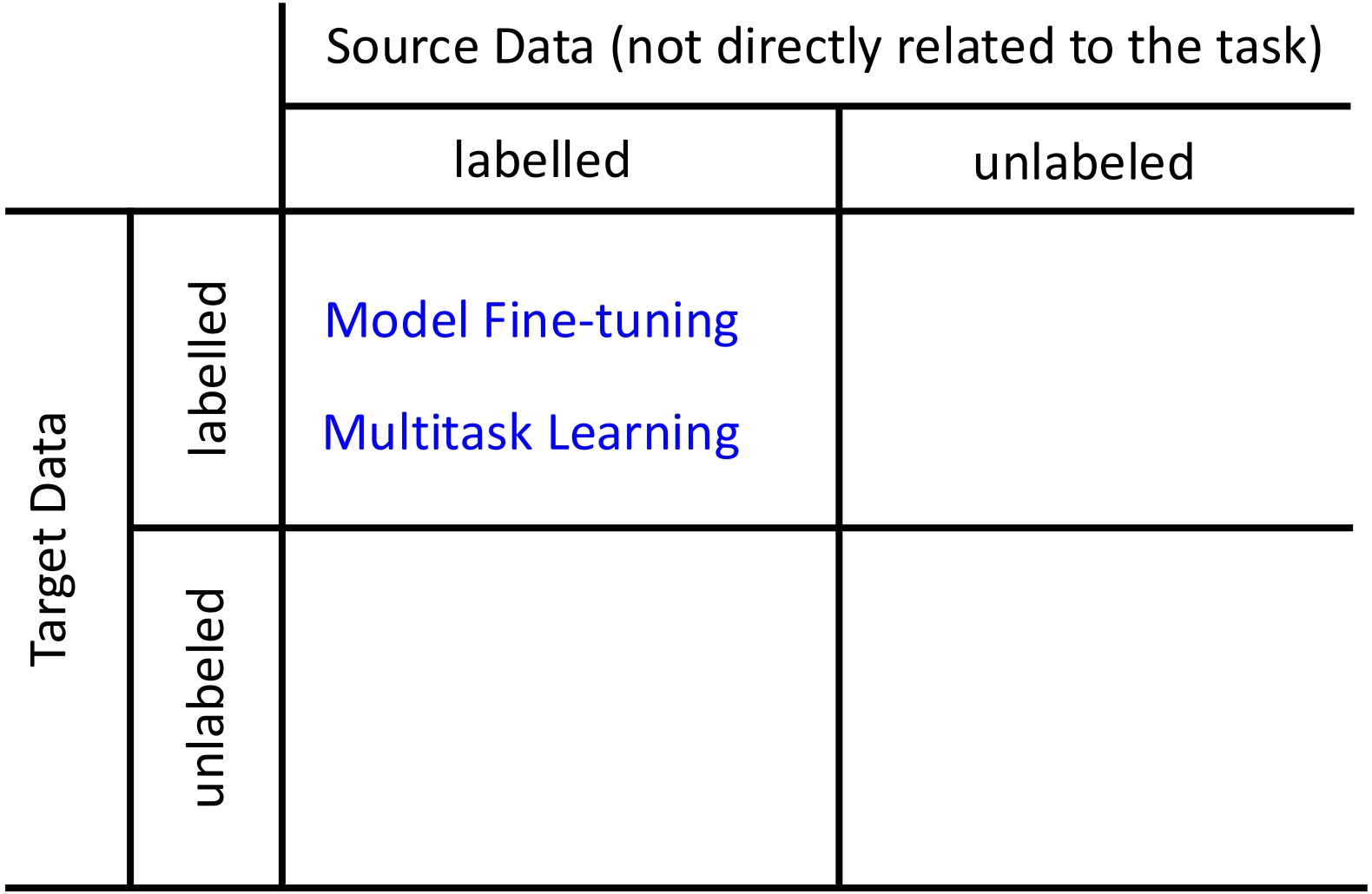
2. fine-tune the whole network (if there is sufficient data)

Layer Transfer

- Which layer can be transferred (copied)?
  - Speech: usually copy the last few layers
  - Image: usually copy the first few layers

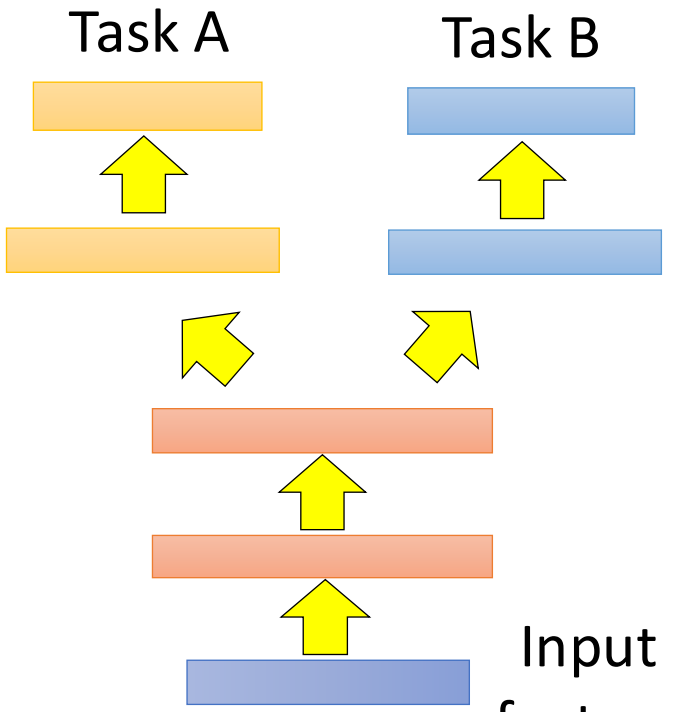


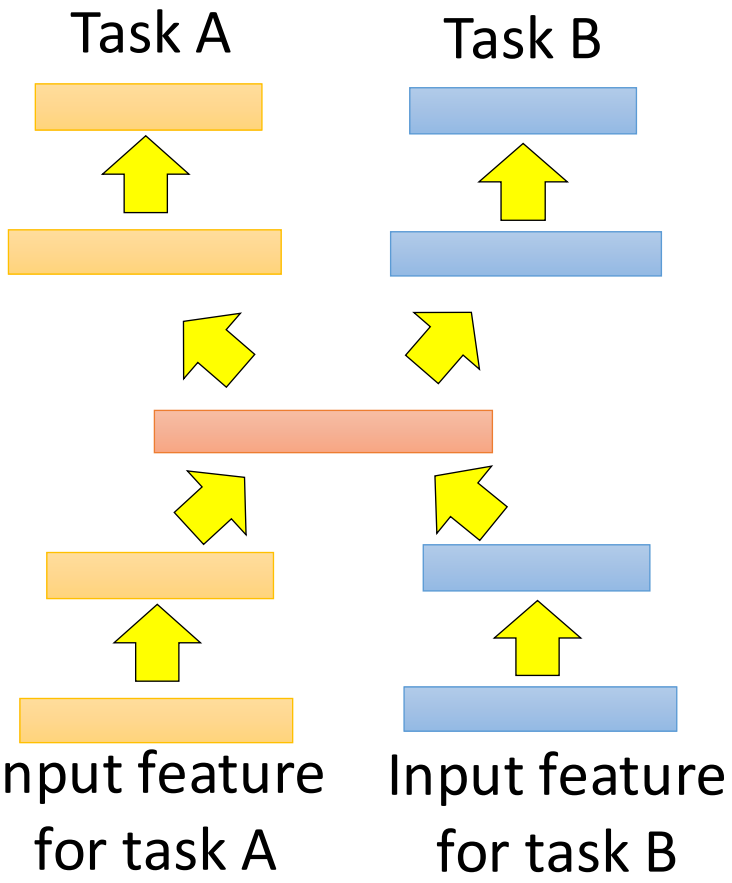
What is Transfer Learning



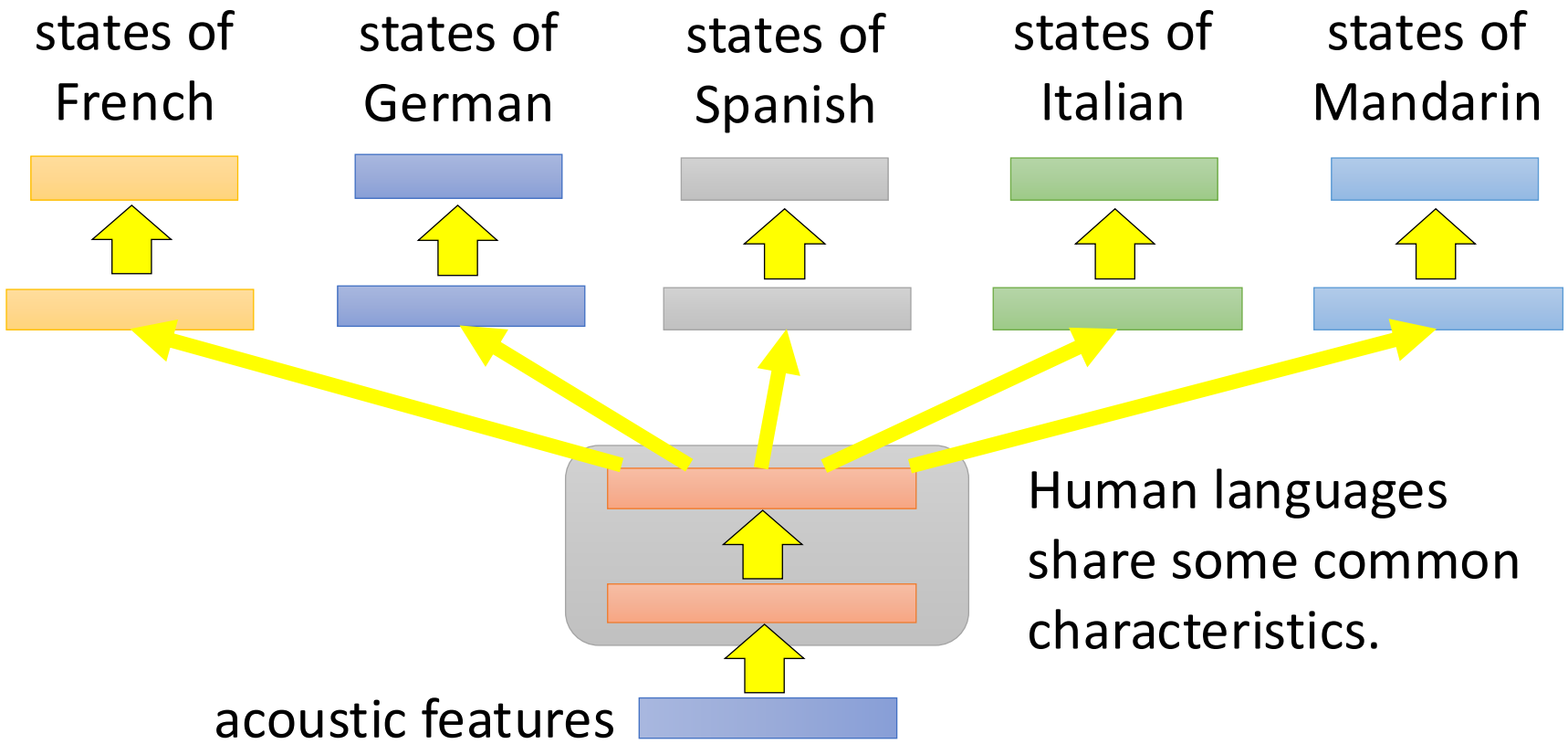
Multitask Learning

- The multi-layer structure makes NN suitable for multitask learning



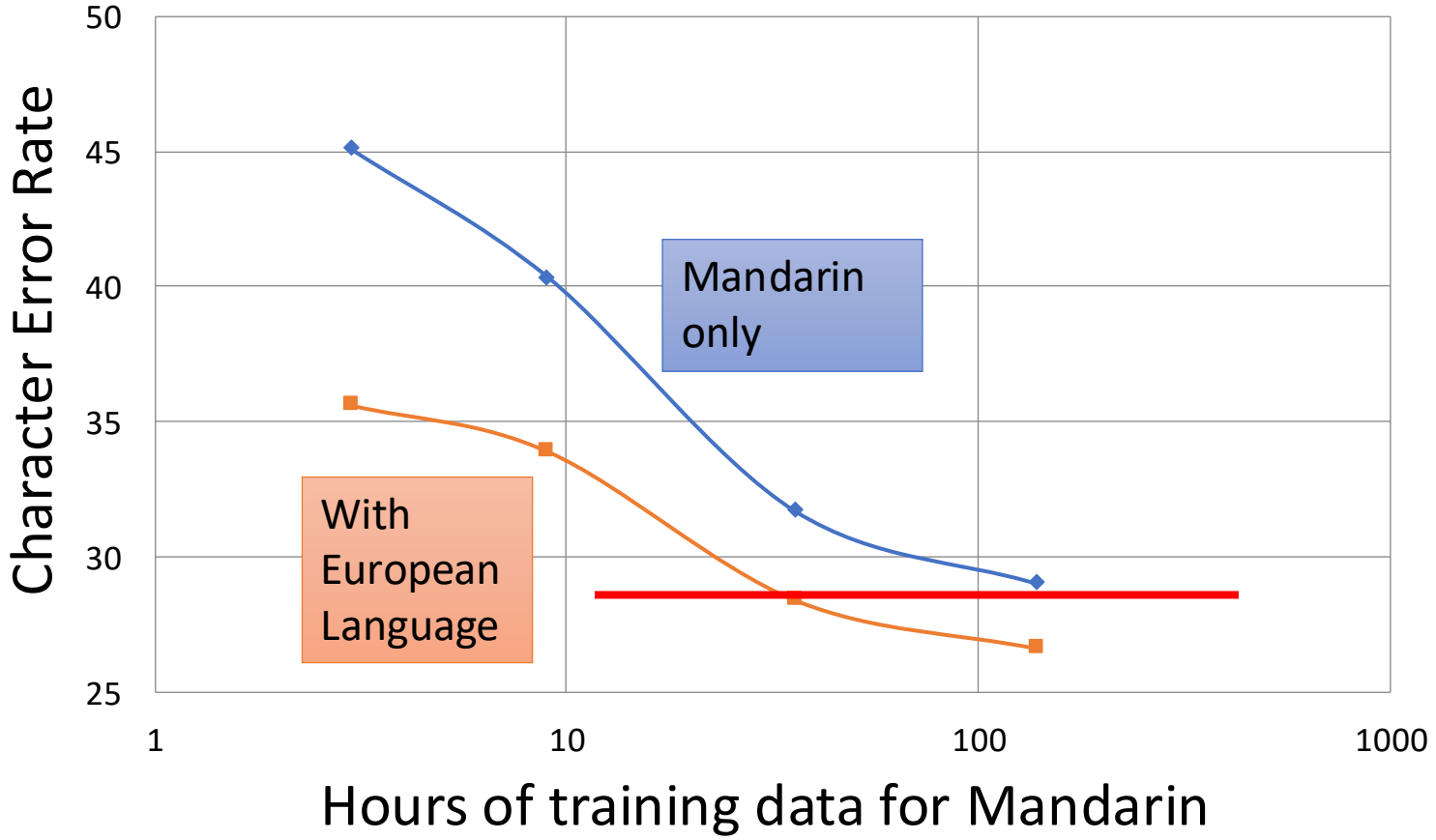


Multitask Learning
- Multilingual Speech Recognition



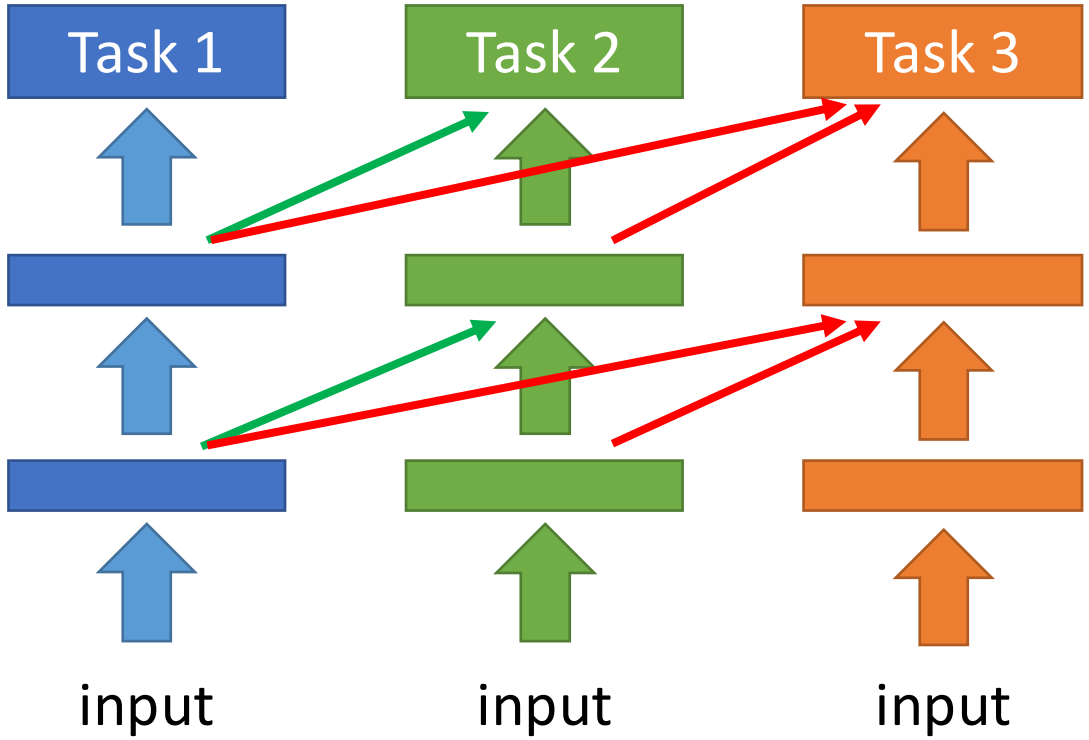
Similar idea in translation: Daxiang Dong, Hua Wu, Wei He, Dianhai Yu and Haifeng Wang, "Multi-task learning for multiple language translation.“, ACL 2015

Multitask Learning - Multilingual



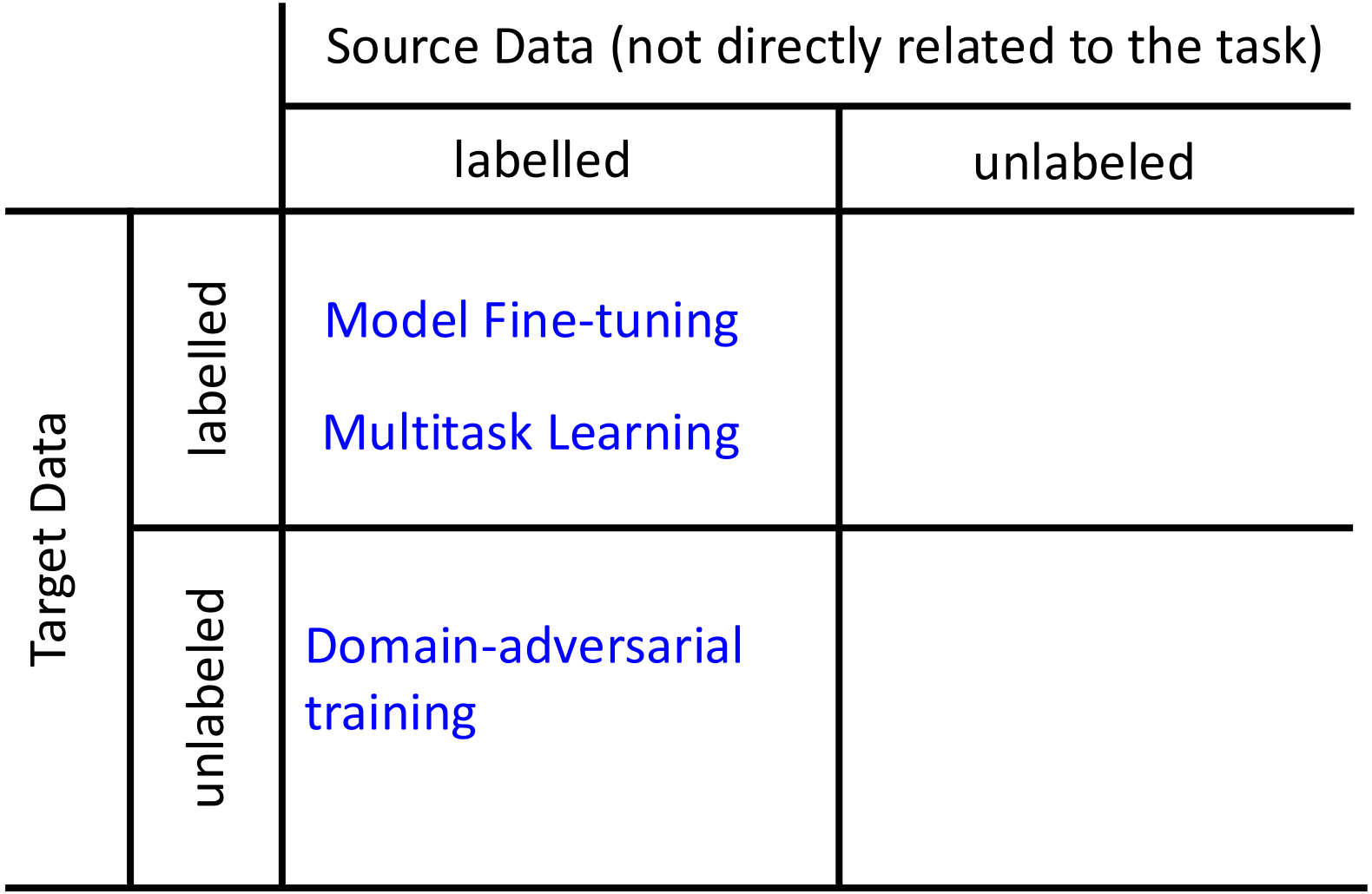
Huang, Jui-Ting, et al. "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers." *ICASSP, 2013*

Progressive Neural Networks



Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, Raia Hadsell, "Progressive Neural Networks", arXiv preprint 2016

What is Transfer Learning



Task Description

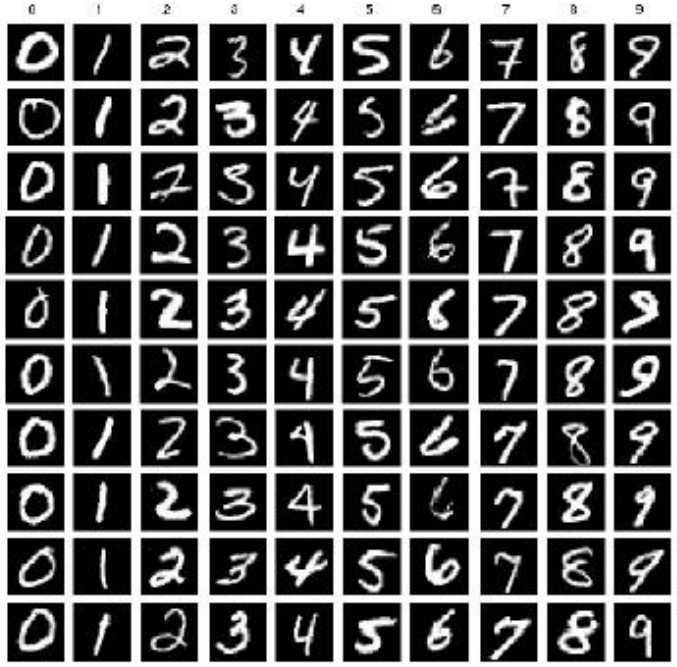
- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

} Same task, mismatch

Domain Adaption

★ Often times, the data of the actual task we want to tackle is largely different from the training data we use.

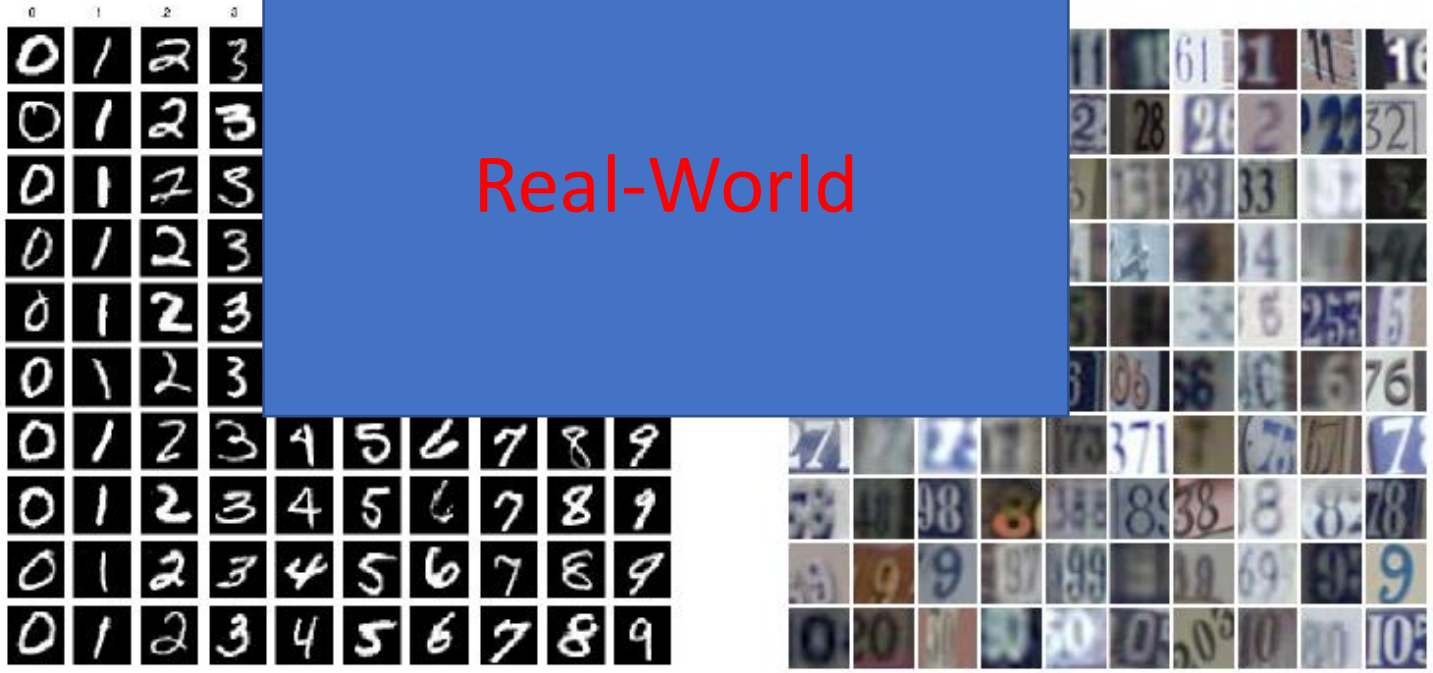
★ We need to think of ways to alleviate this problem.



Domain Adaption

★ Often times, the data of the actual task we want to tackle is largely different from the training data we use.

★ We need to think of ways to alleviate this problem.



Definition: Domain Adaption

Domain D

Definition: Domain Adaption

Task T

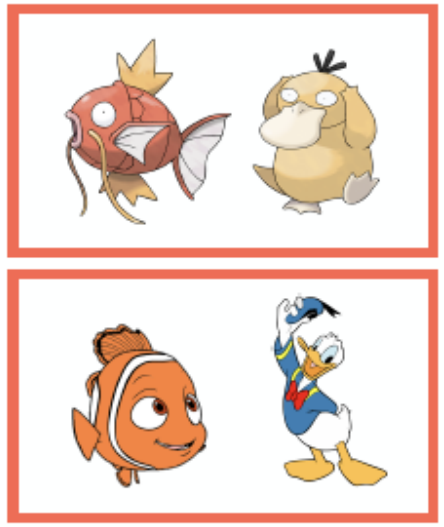
Classification "Biker"

Segmentation

Bounding box detection

Definition: Domain Adaption

D^s

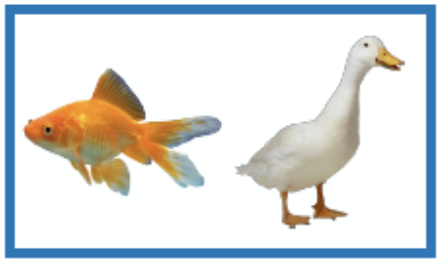


Source

Fully labelled

$D^s \neq D^t$

D^t



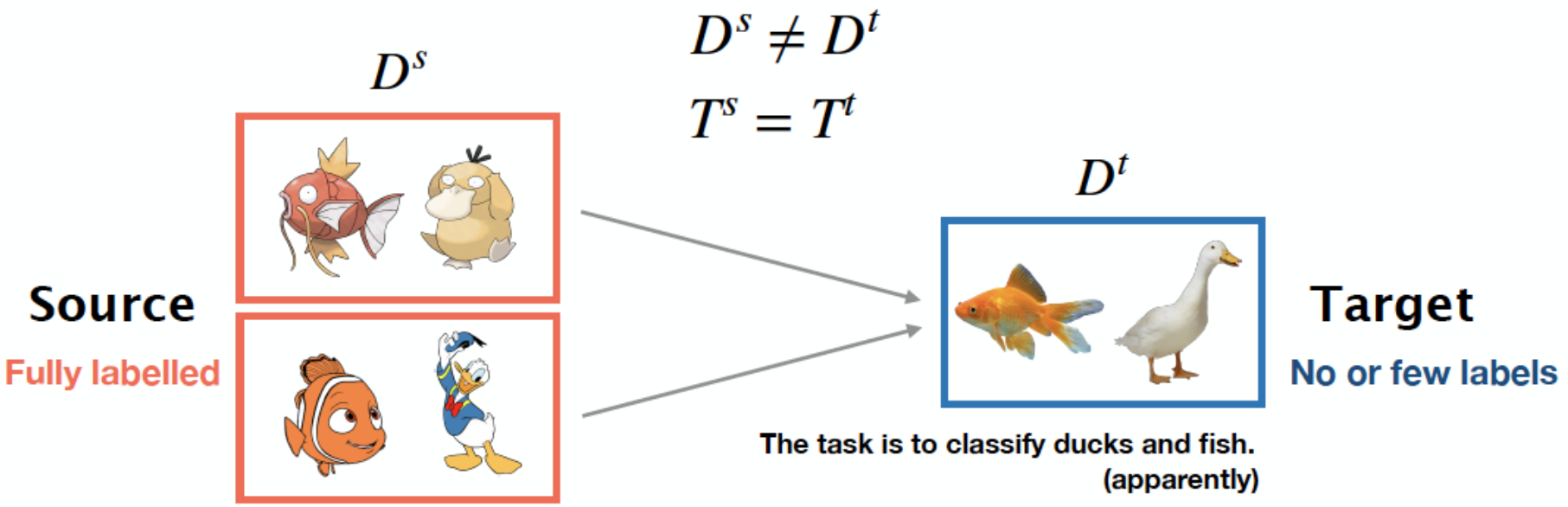
Target

No or few labels

Testing

Definition: Domain Adaption

Tasks are the same



Why: Domain Adaption

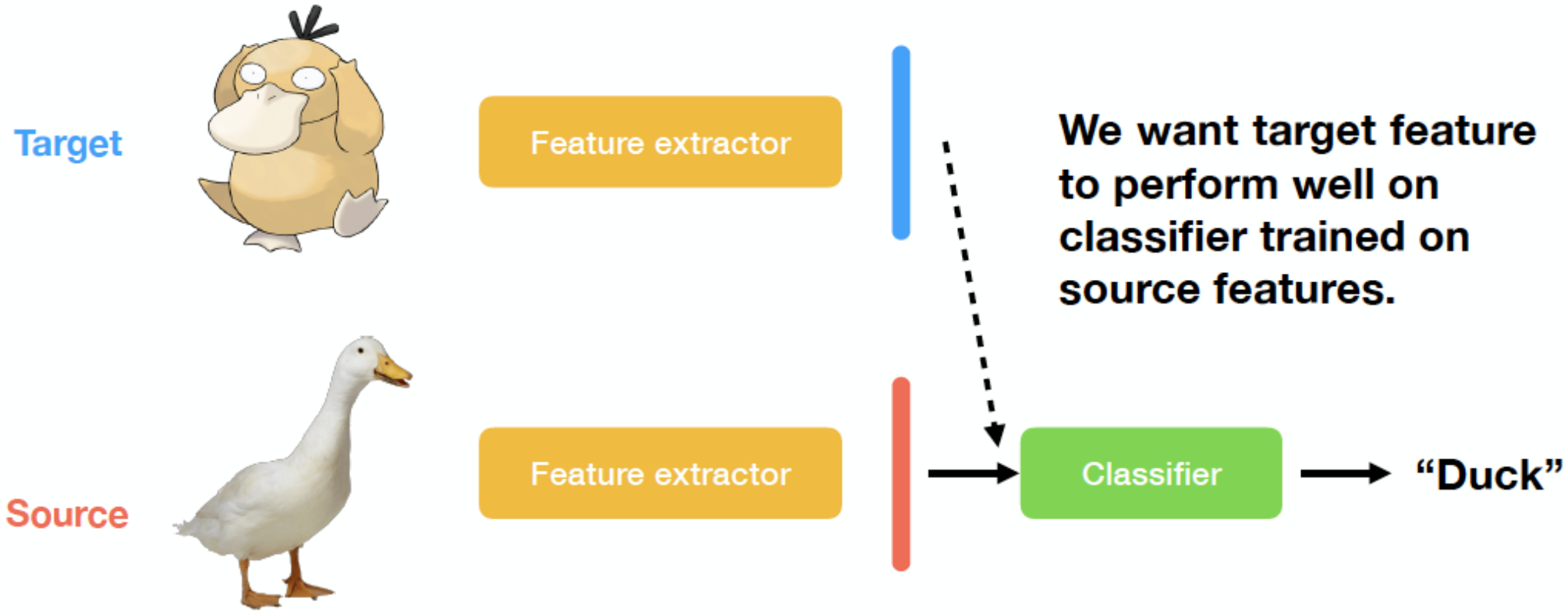
When it comes to autonomous driving ...

We are able to easily amass unlimited amounts of street view data from Grand Theft Auto.

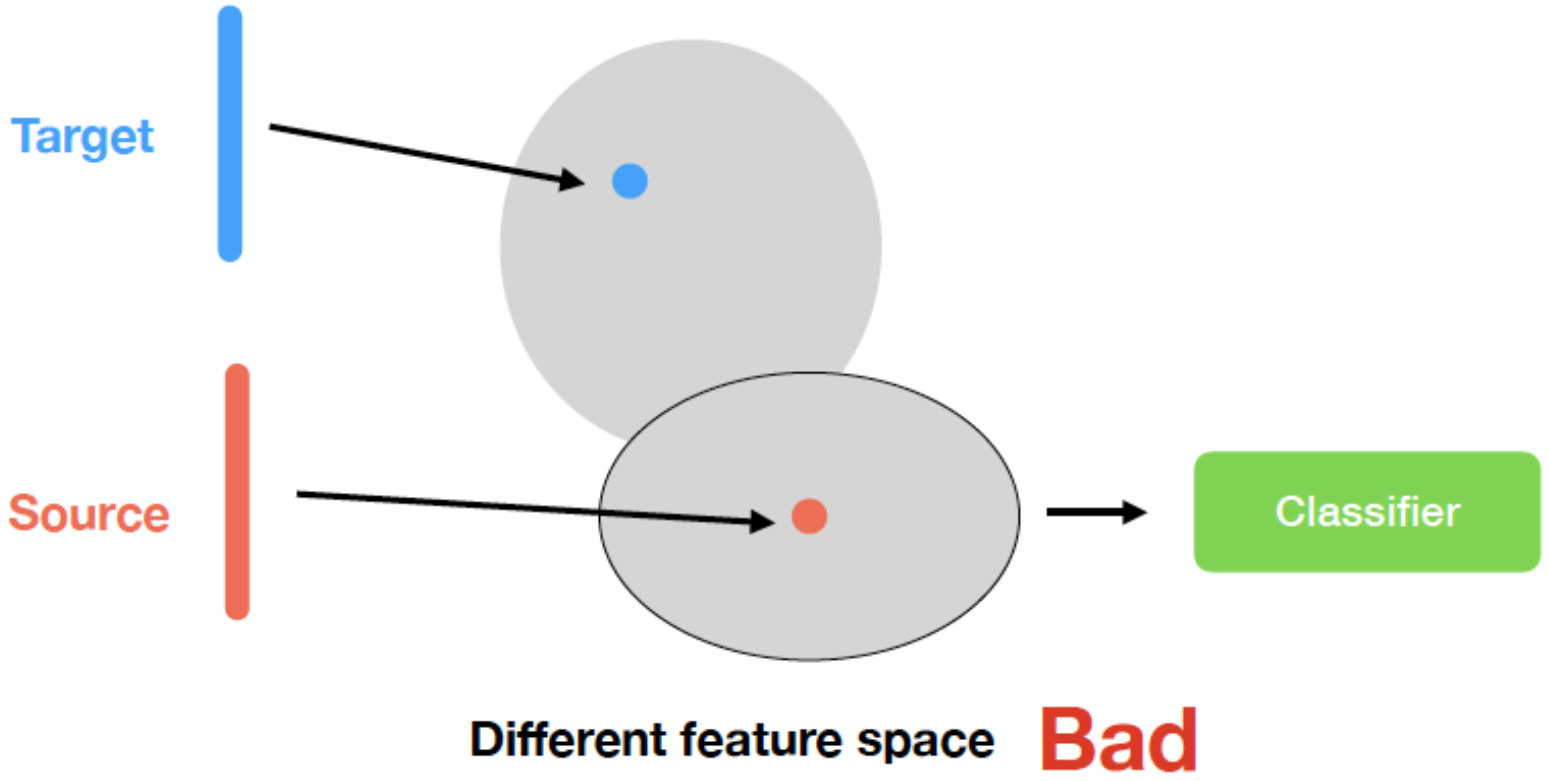
Not the case for arbitrary real-life scenarios.

Why: Domain Adaption

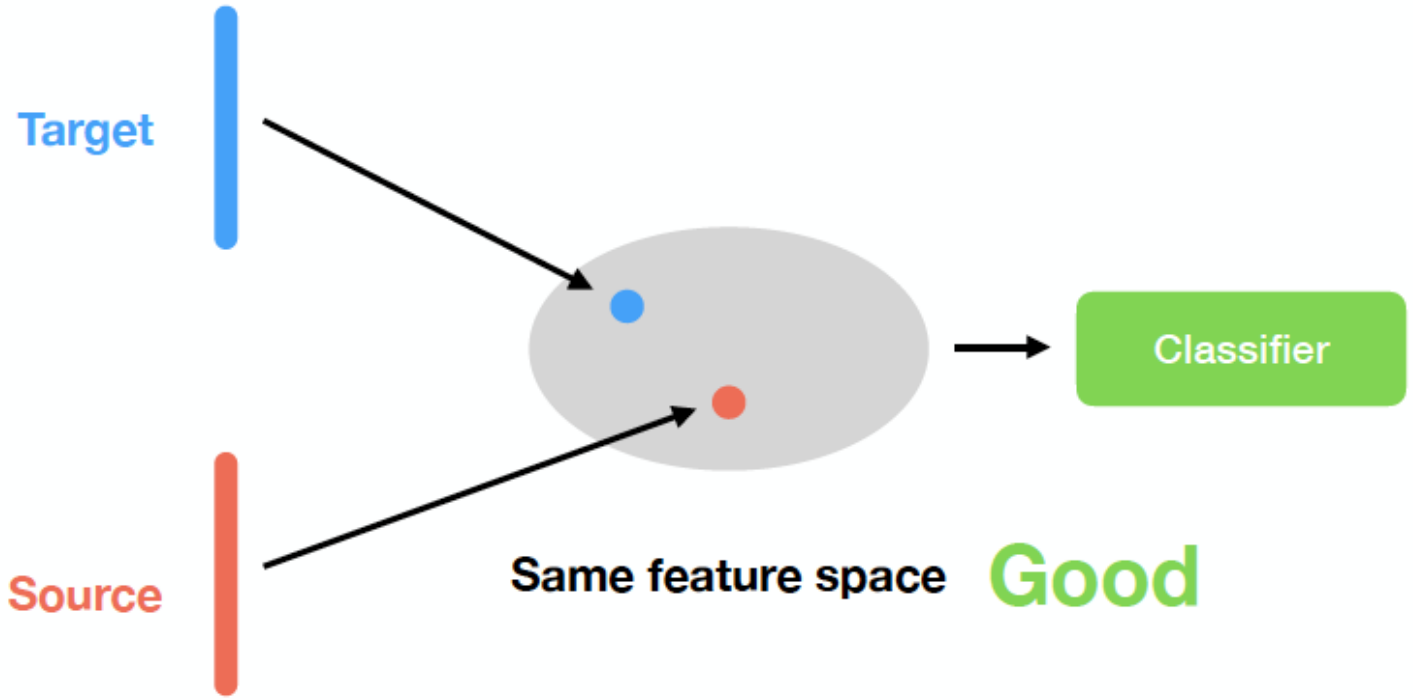
Feature: Domain Adaption



Feature: Domain Adaption



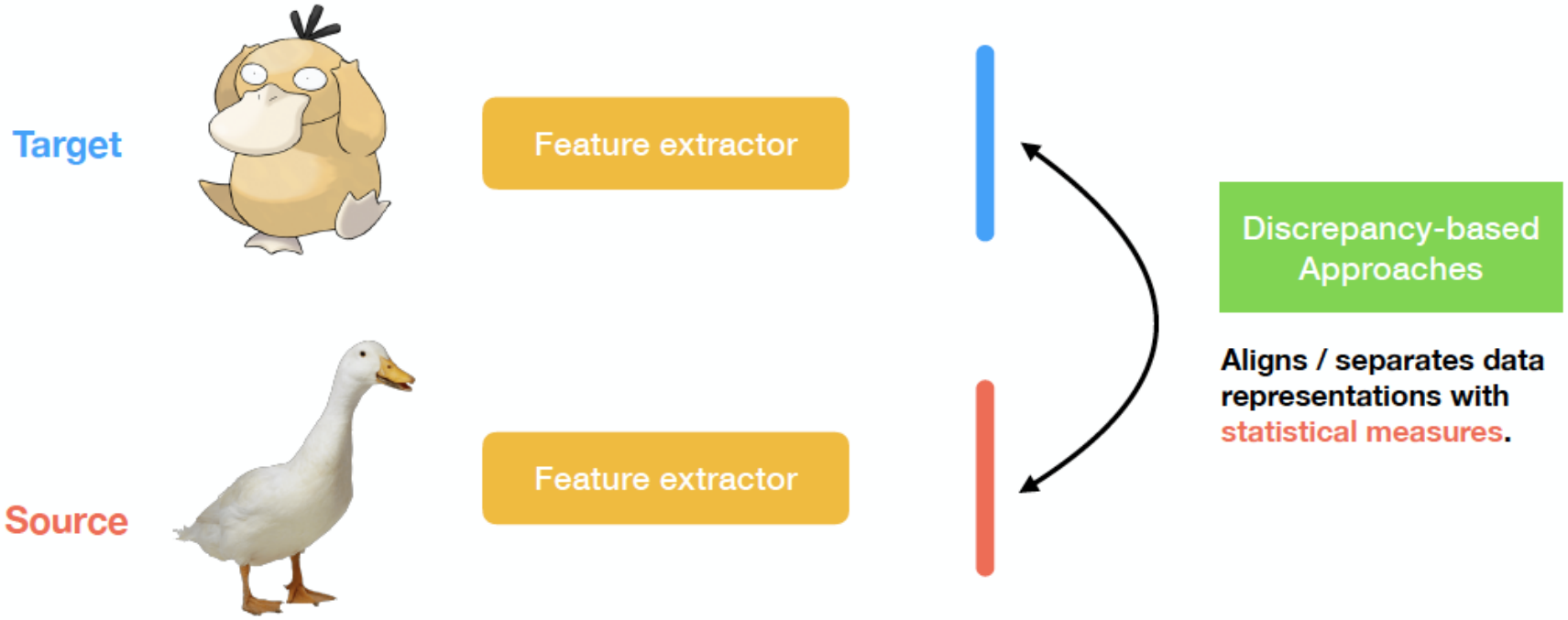
Feature: Domain Adaption



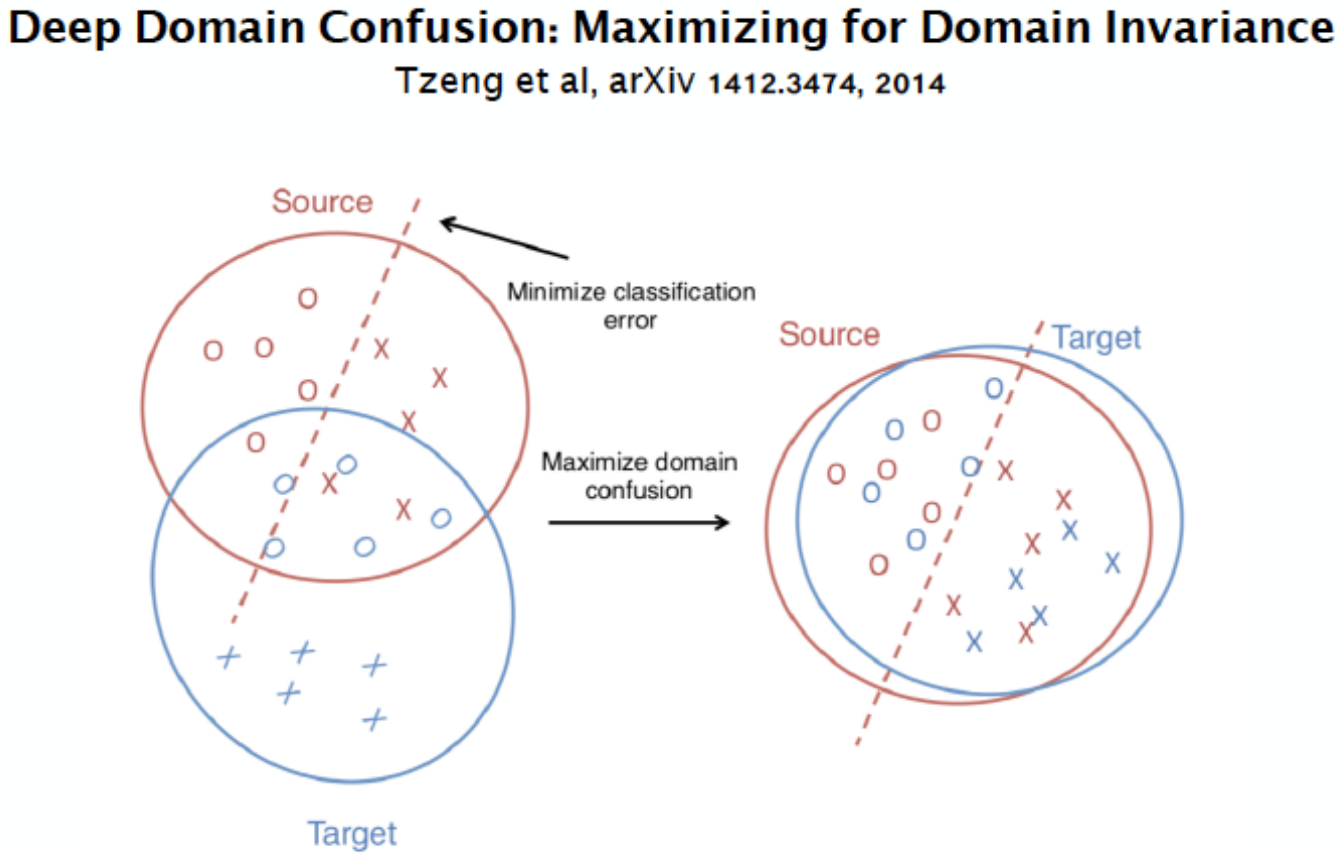
Methods for Domain Adaption

1. **Discrepancy-based methods**
2. **Adversarial-based methods**
3. **Reconstruction-based methods**

Discrepancy-based

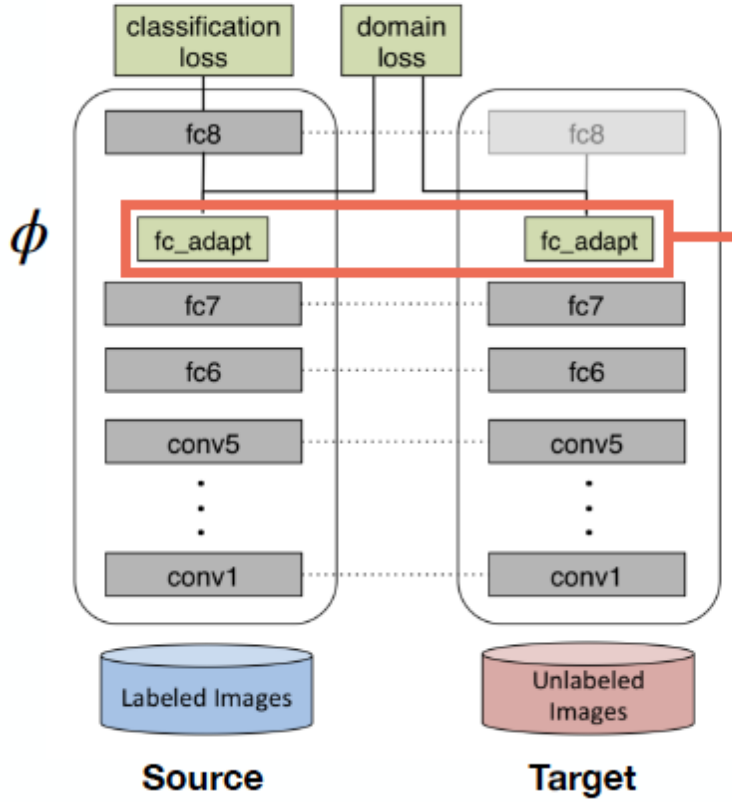


Deep Domain Confusion



Deep Domain Confusion

https://en.wikipedia.org/wiki/Kernel_embedding_of_distributions
https://en.wikipedia.org/wiki/Kernel_method



Maximum Mean Discrepancy

Measuring distance between distributions

$$\mathrm{MMD}(X_S, X_T) = \left\| \frac{1}{|X_S|} \sum_{x_s \in X_S} \phi(x_s) - \frac{1}{|X_T|} \sum_{x_t \in X_T} \phi(x_t) \right\|$$

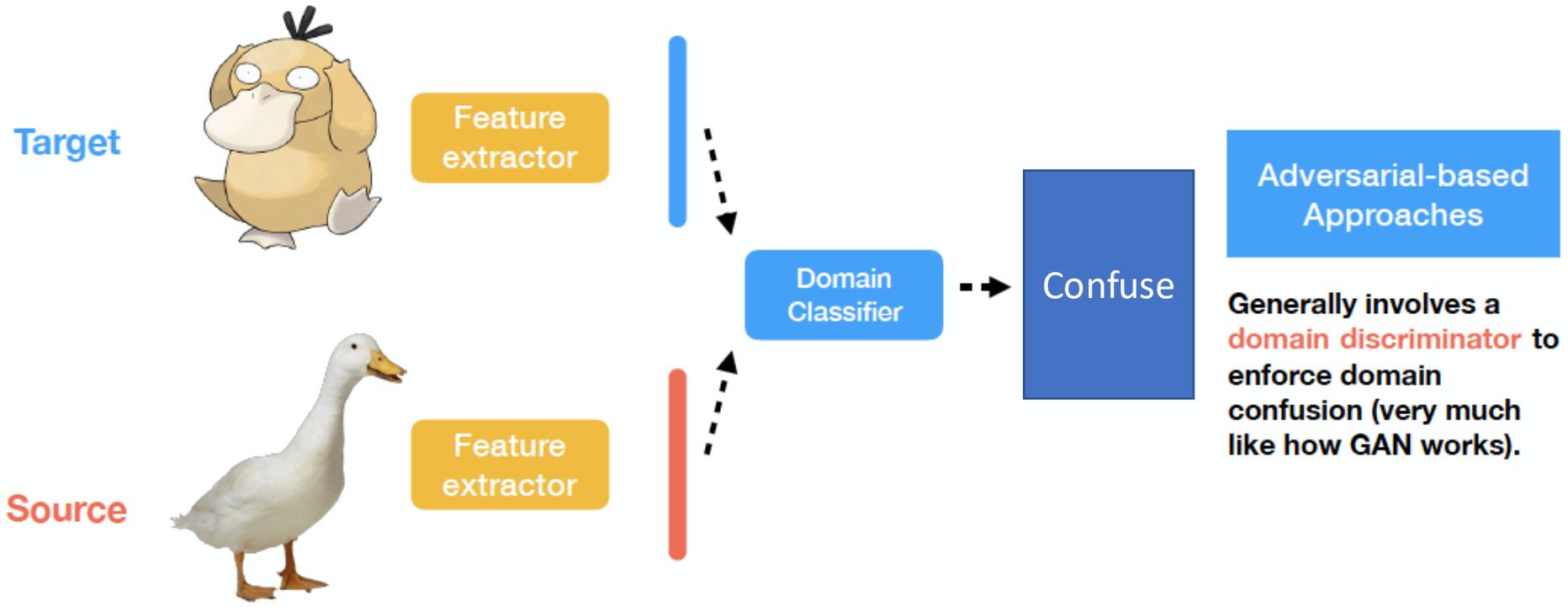
Objective Function

Classification + MMD

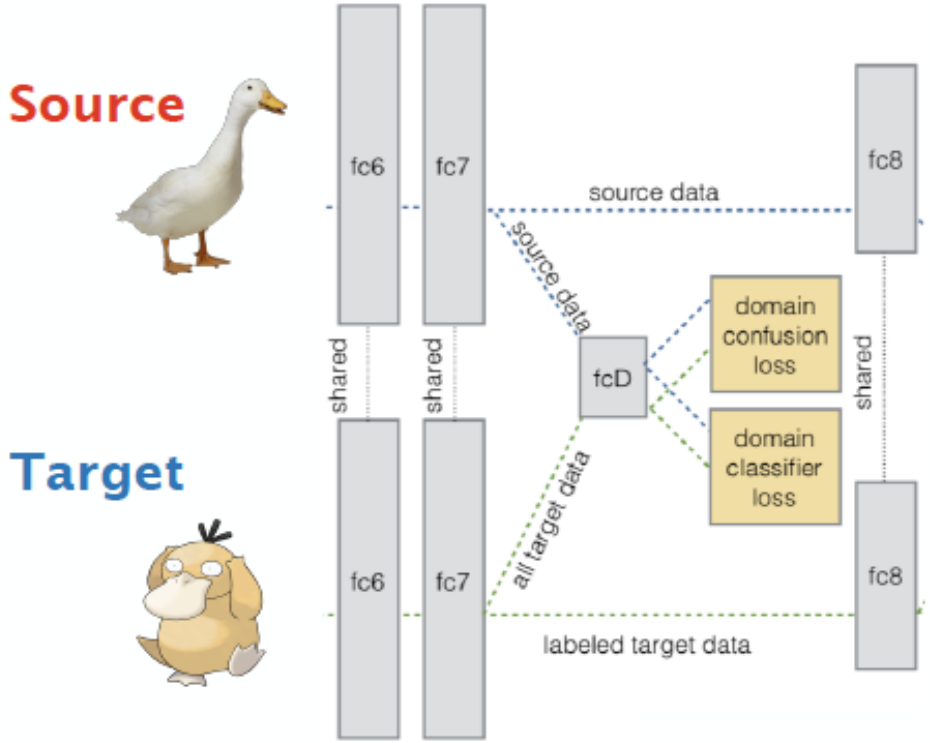
$$\mathcal{L} = \mathcal{L}_C(X_L, y) + \lambda \mathrm{MMD}^2(X_S, X_T)$$

How do you choose ϕ?

Adversarial based



Domain-adversarial training



Iteration t

Learn a good domain classifier

fix feature extractor

$$L_{cls} \quad -\sum \mathbb{1}[y_D = d] \log q_d$$

Iteration t + 1

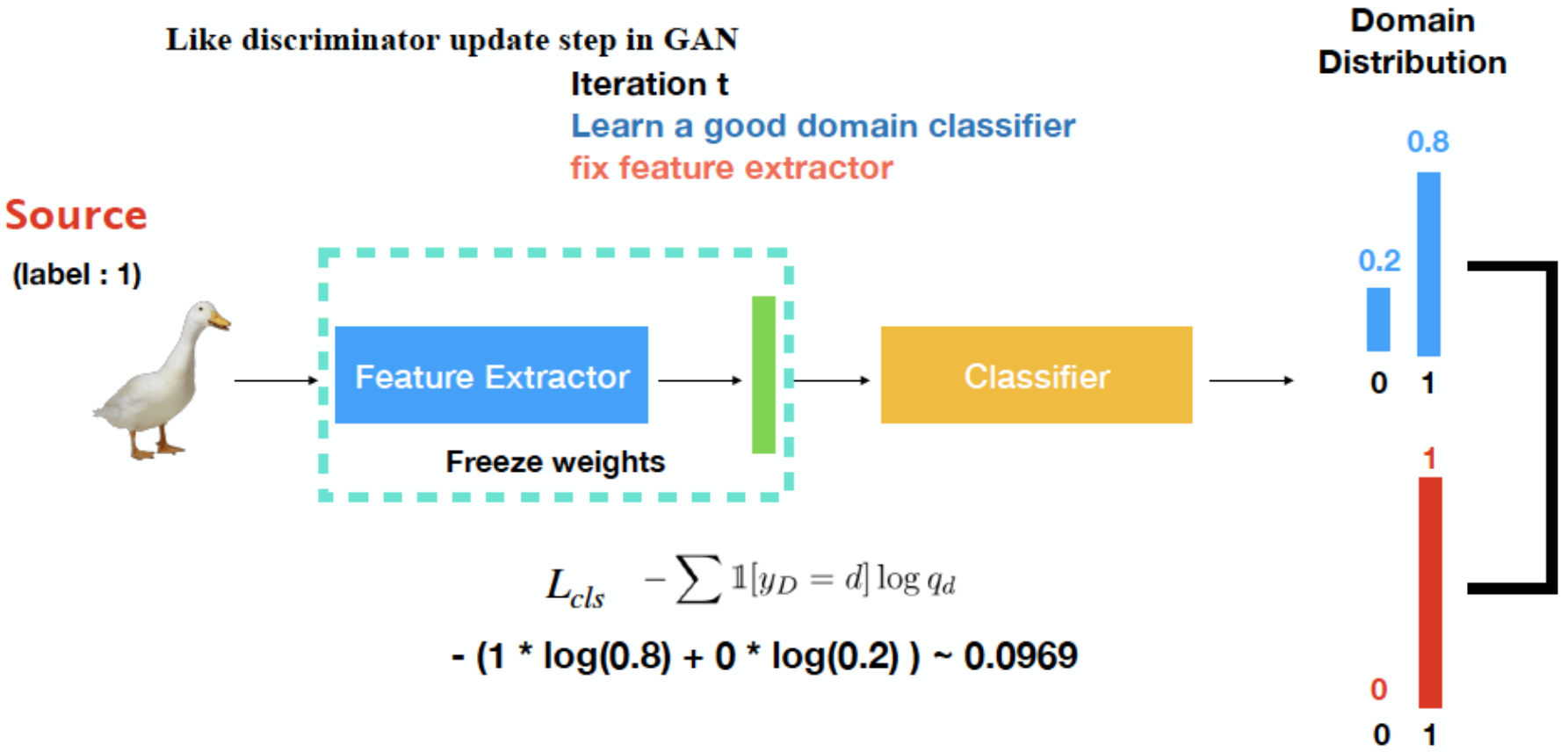
Learn a representation that confuses the classifier (uniform distribution)

fix domain classifier

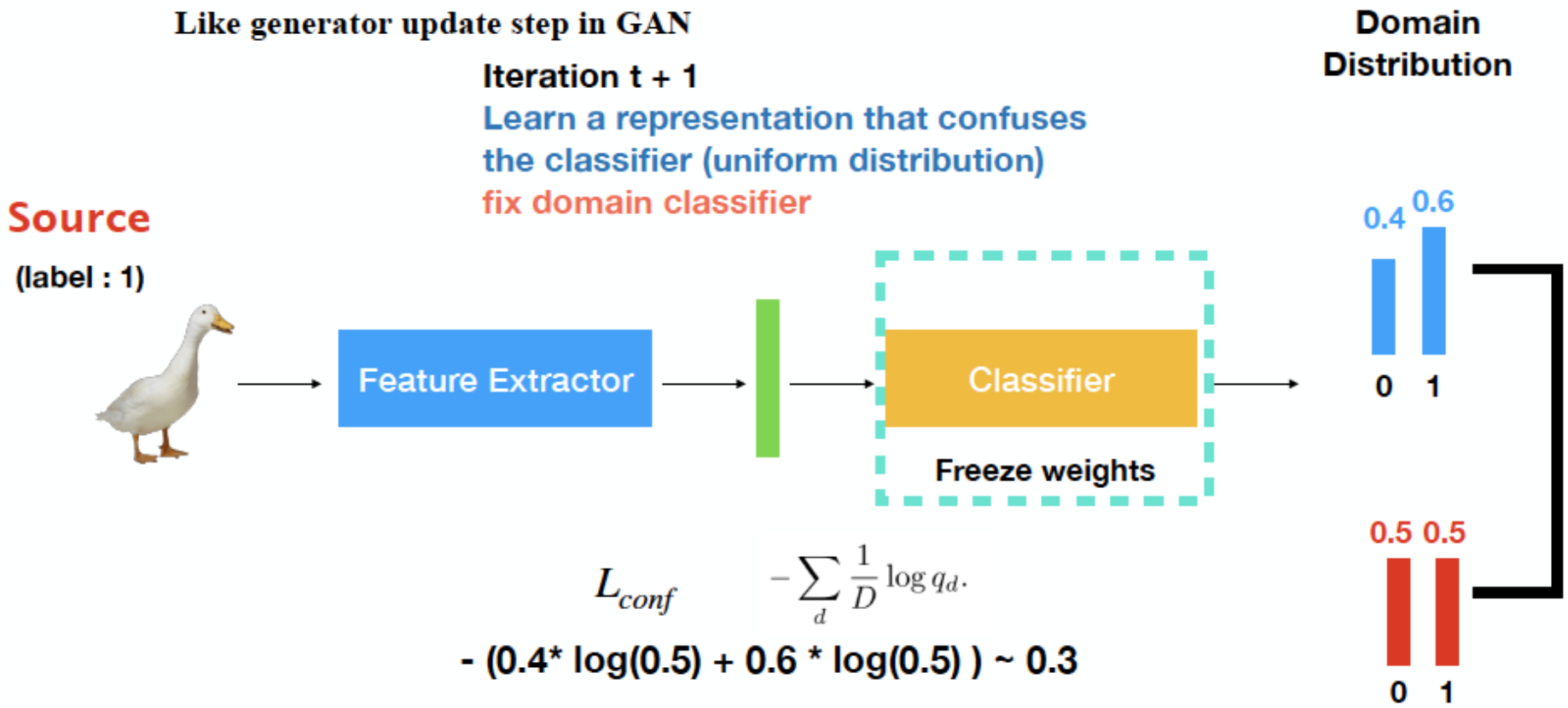
$$L_{conf} \quad -\sum_d \frac{1}{D} \log q_d.$$

Both Cross Entropy !

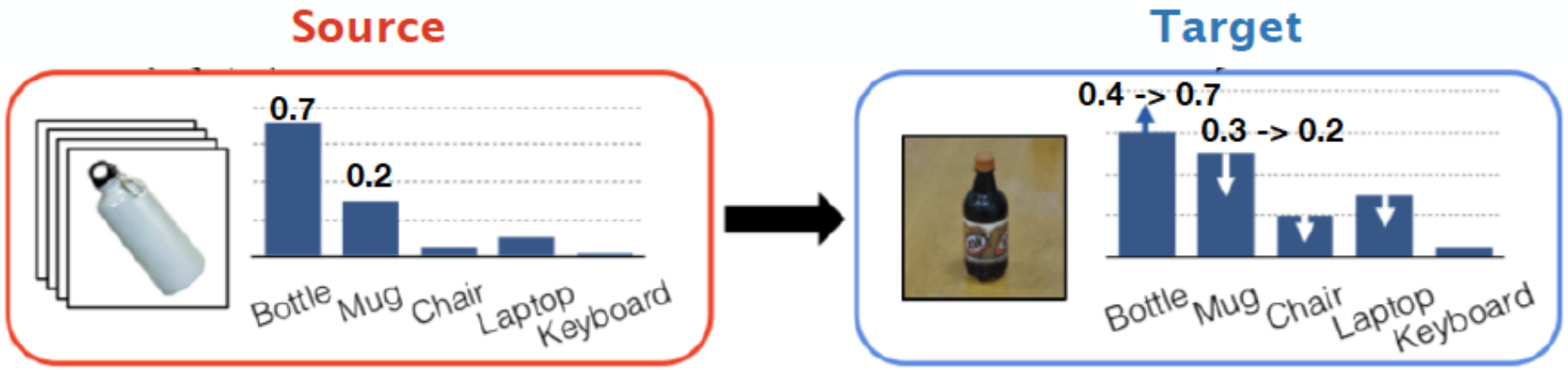
Domain-adversarial training



Domain-adversarial training



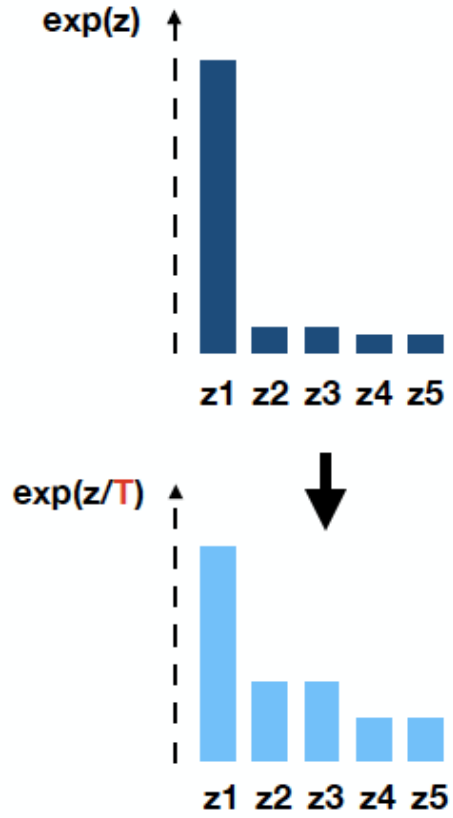
Domain-adversarial training-Label Correlation



Idea: "bottle" should be similar to "mug" but less similar to "laptop"

Class distributions of source and target should be similar

Domain-adversarial training-Label Correlation



Harder label

Soft label

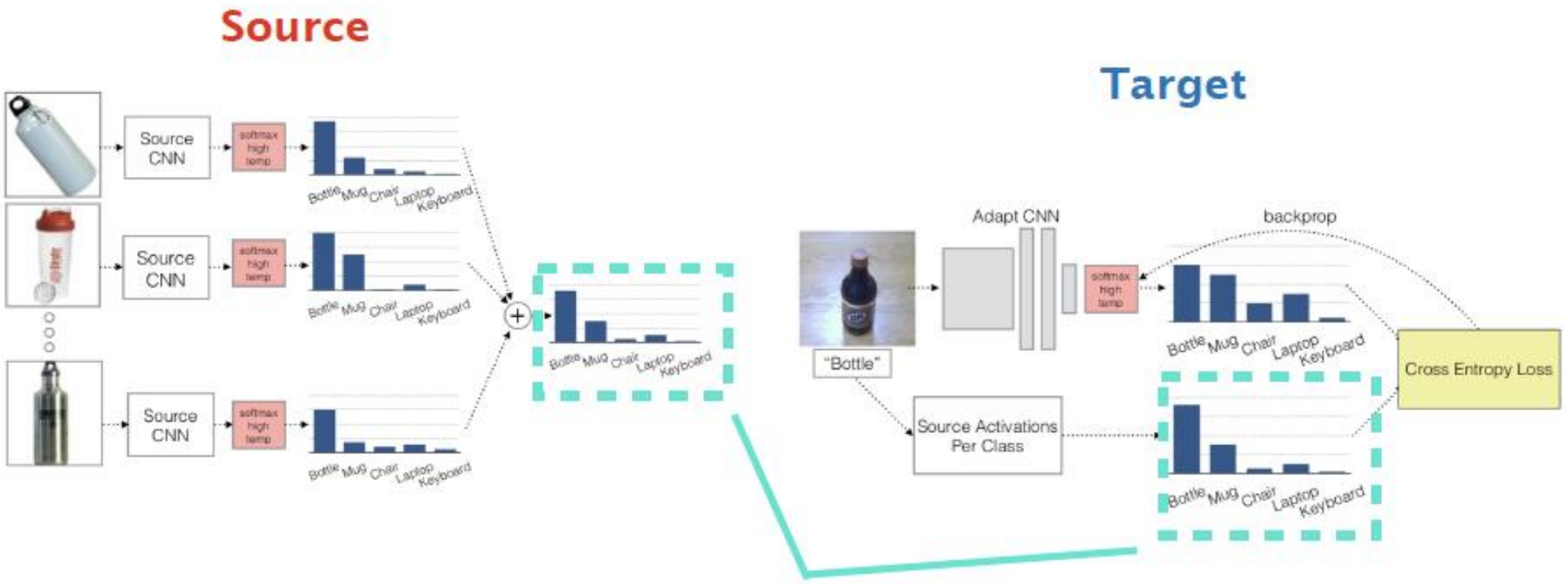
Softmax with Temperature

Enables a smoother class probability distribution to address inter-class correlations

$$\frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

"Scale down probabilities"

Domain-adversarial training-Label Correlation

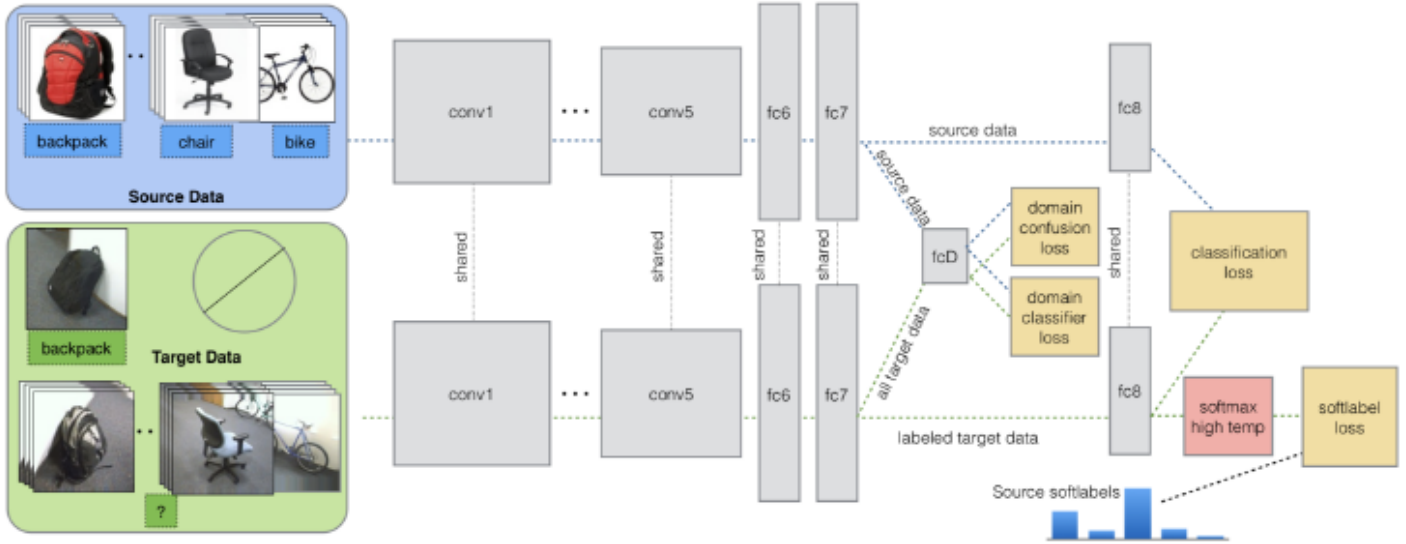


Domain-adversarial training-Final

Simultaneous Deep Transfer Across Domains and Tasks

Tzeng et al, ICCV 2015

$$L = L_{cls} + L_{conf} + L_{soft}$$



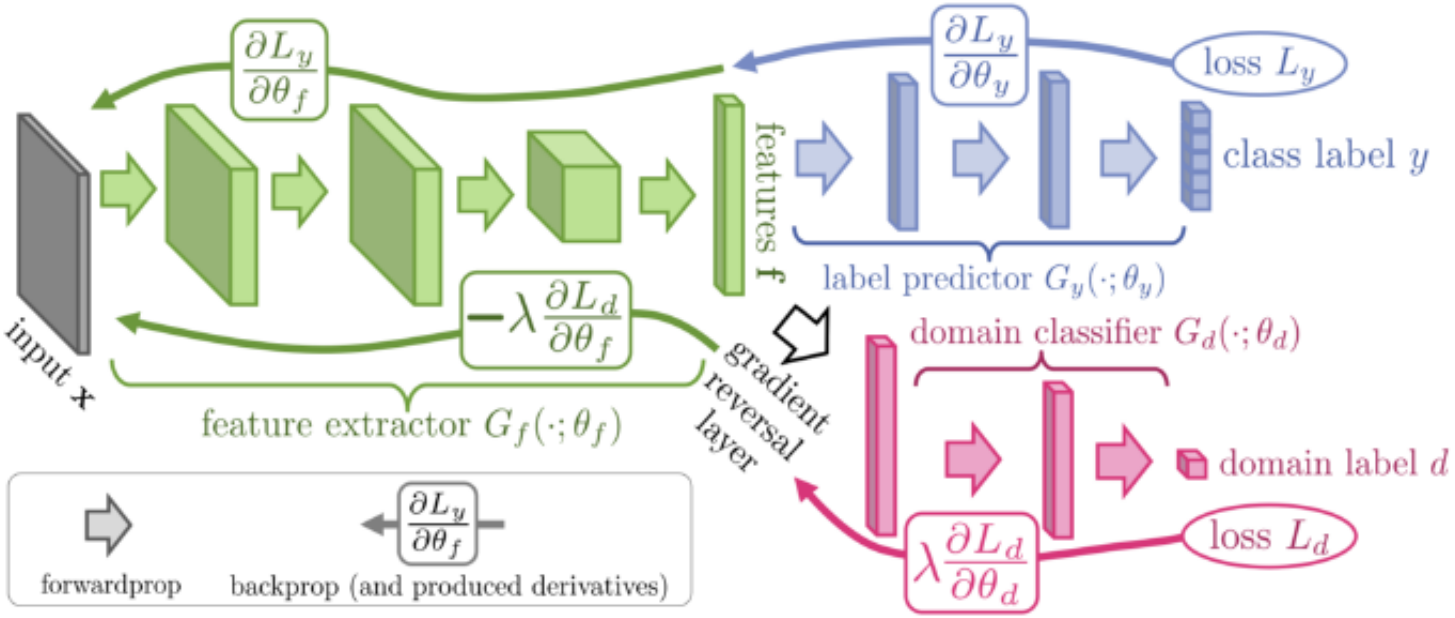
Domain-adversarial training

DANN

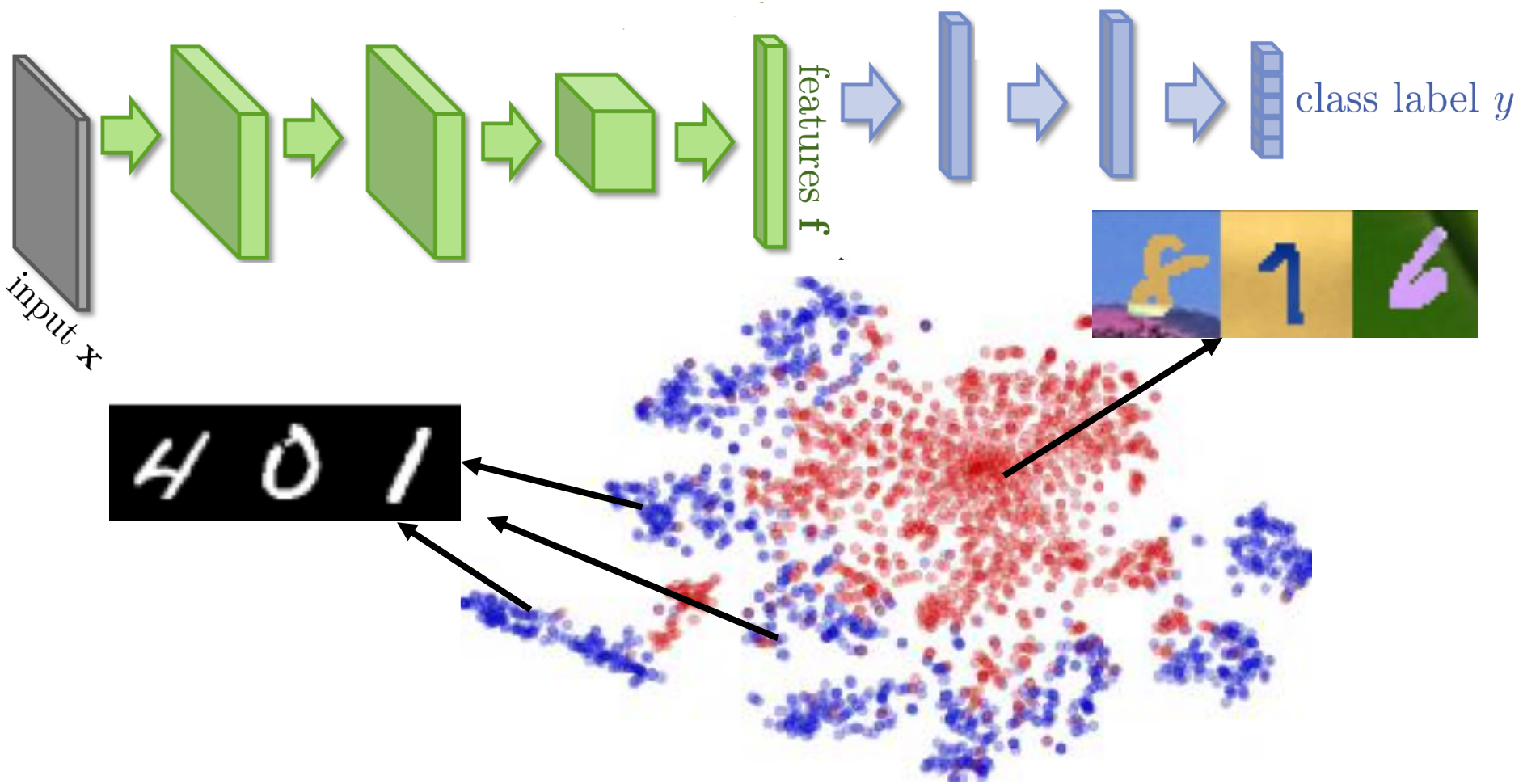
Domain Adversarial Training of Neural Networks

Ganin et al, NIPS 2016

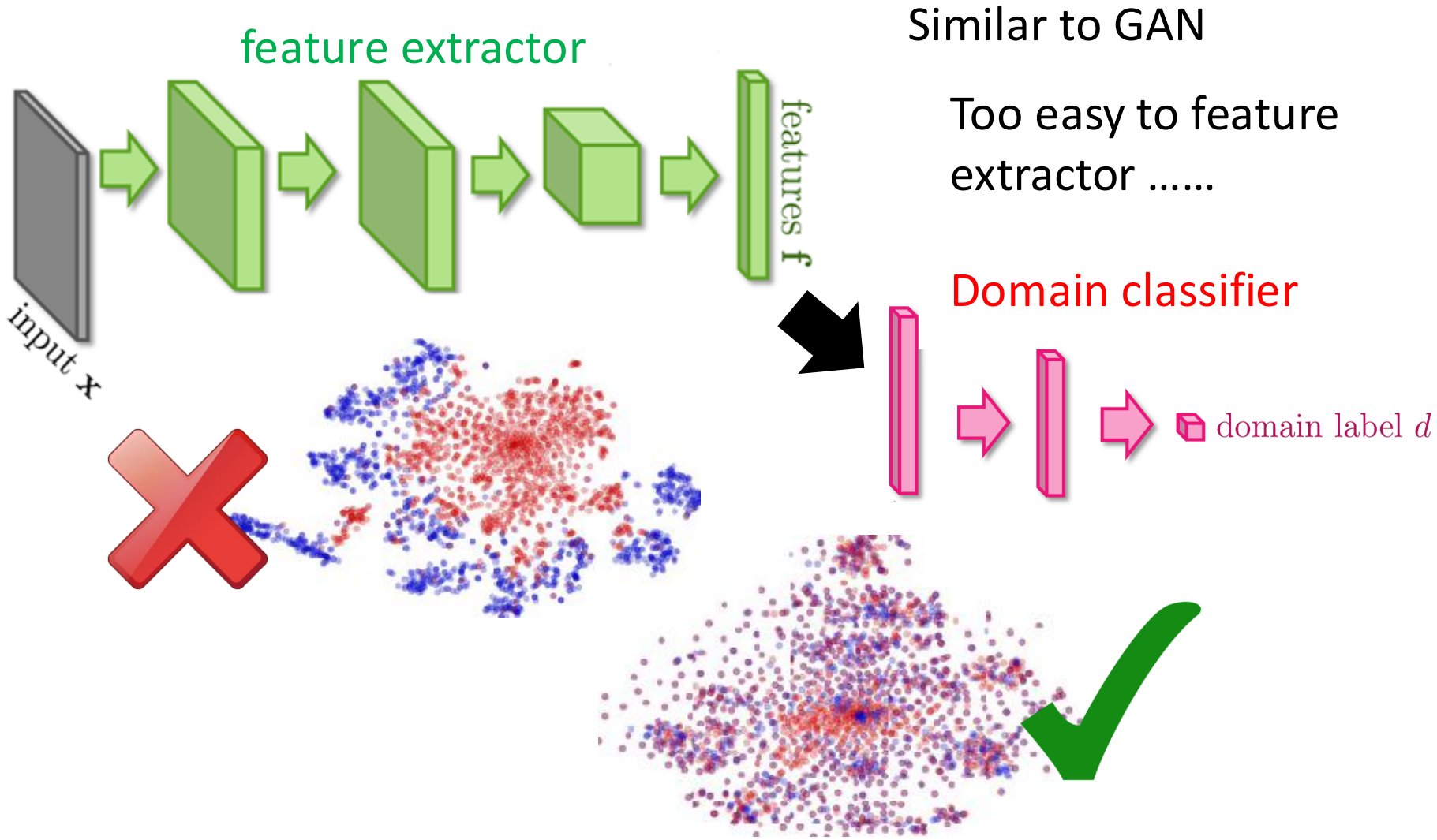
- **Spotlight :** Domain Classifier + Gradient Reversal



Domain-adversarial training



Domain-adversarial training



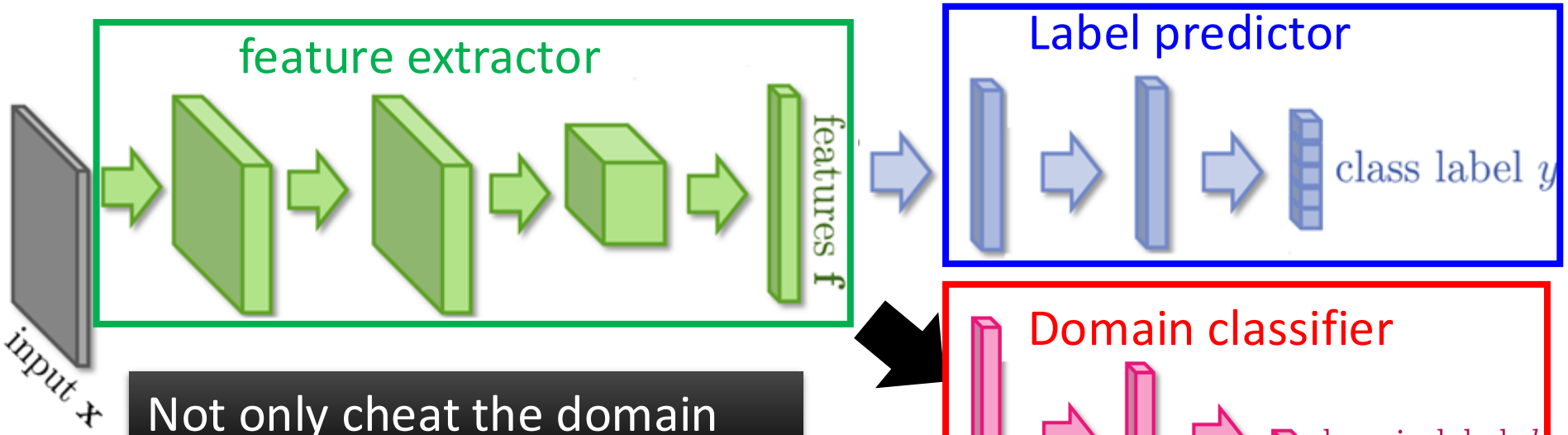
Similar to GAN

Too easy to feature extractor ……

Domain-adversarial training

Maximize label classification accuracy + minimize domain classification accuracy

Maximize label classification accuracy

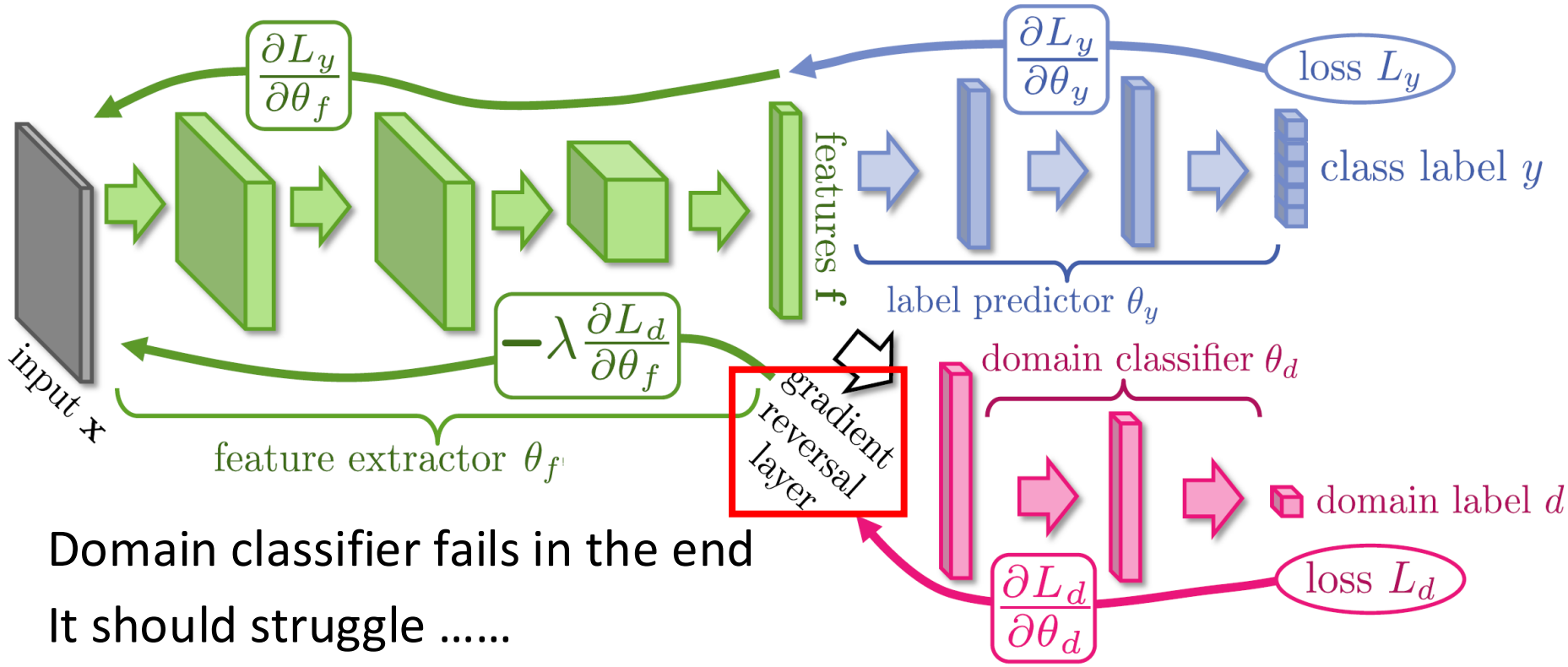


Not only cheat the domain classifier, but satisfying label classifier at the same time

Maximize domain classification accuracy

This is a big network, but different parts have different goals.

Domain-adversarial training



Domain classifier fails in the end

It should struggle ……

Yaroslav Ganin, Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, ICML, 2015

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Domain-Adversarial Training of Neural Networks, JMLR, 2016

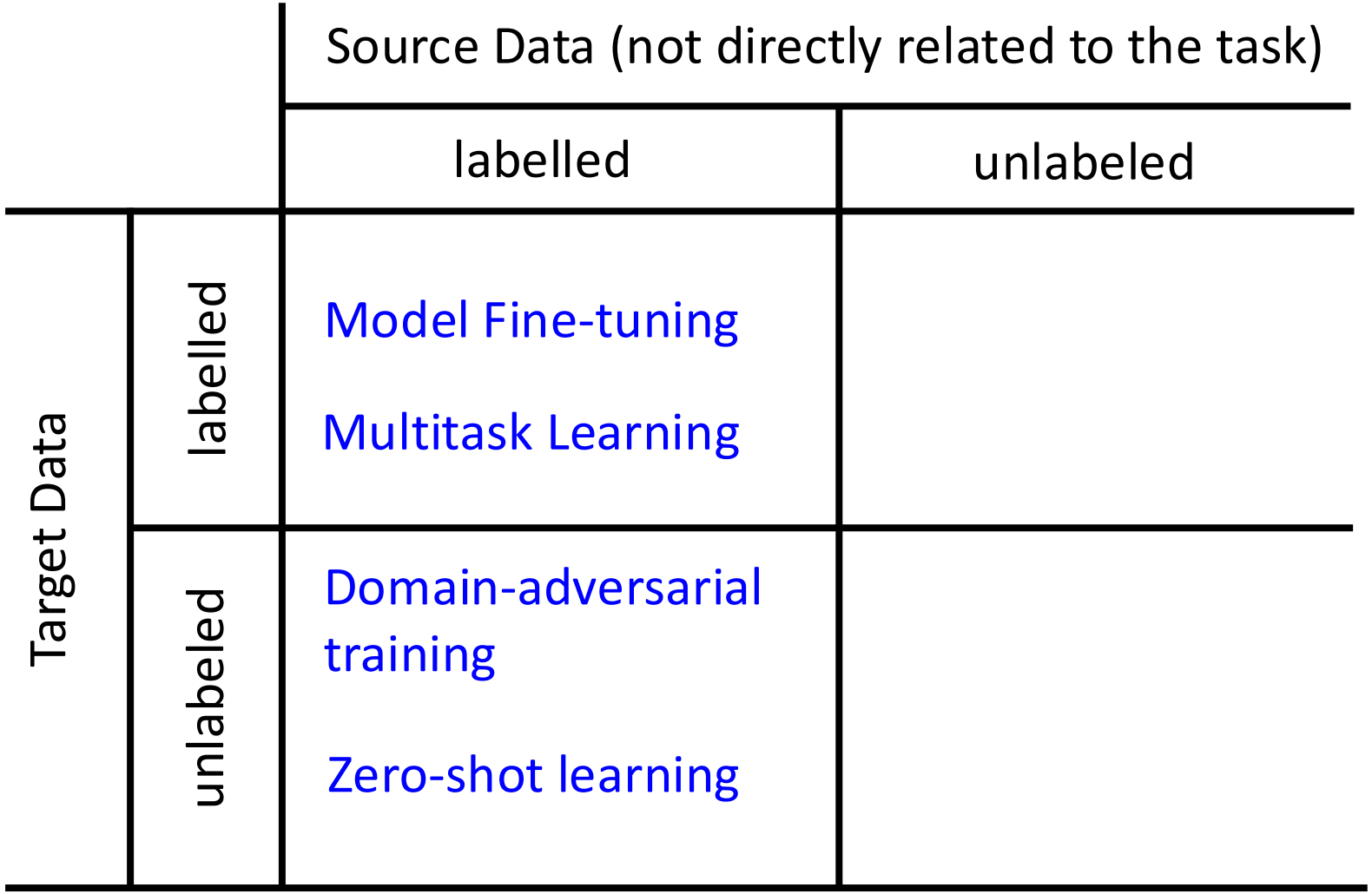
Domain-adversarial training

METHOD	SOURCE / TARGET	MNIST / MNIST-M	SYN NUMBERS / SVHN	SVHN / MNIST	SYN SIGNS / GTSRB
SOURCE ONLY		.5749	.8665	.5919	.7400
SA (FERNANDO ET AL., 2013)		.6078 (7.9%)	.8672 (1.3%)	.6157 (5.9%)	.7635 (9.1%)
PROPOSED APPROACH		**.8149** (57.9%)	**.9048** (66.1%)	**.7107** (29.3%)	**.8866** (56.7%)
TRAIN ON TARGET		.9891	.9244	.9951	.9987

Yaroslav Ganin, Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, ICML, 2015

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Domain-Adversarial Training of Neural Networks, JMLR, 2016

What is Transfer Learning



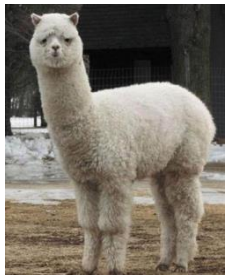
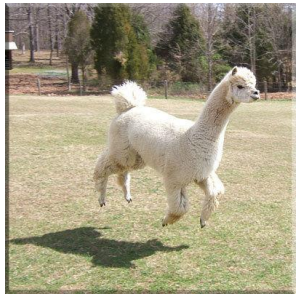
Zero-shot Learning

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

Different tasks

x^s:   ……

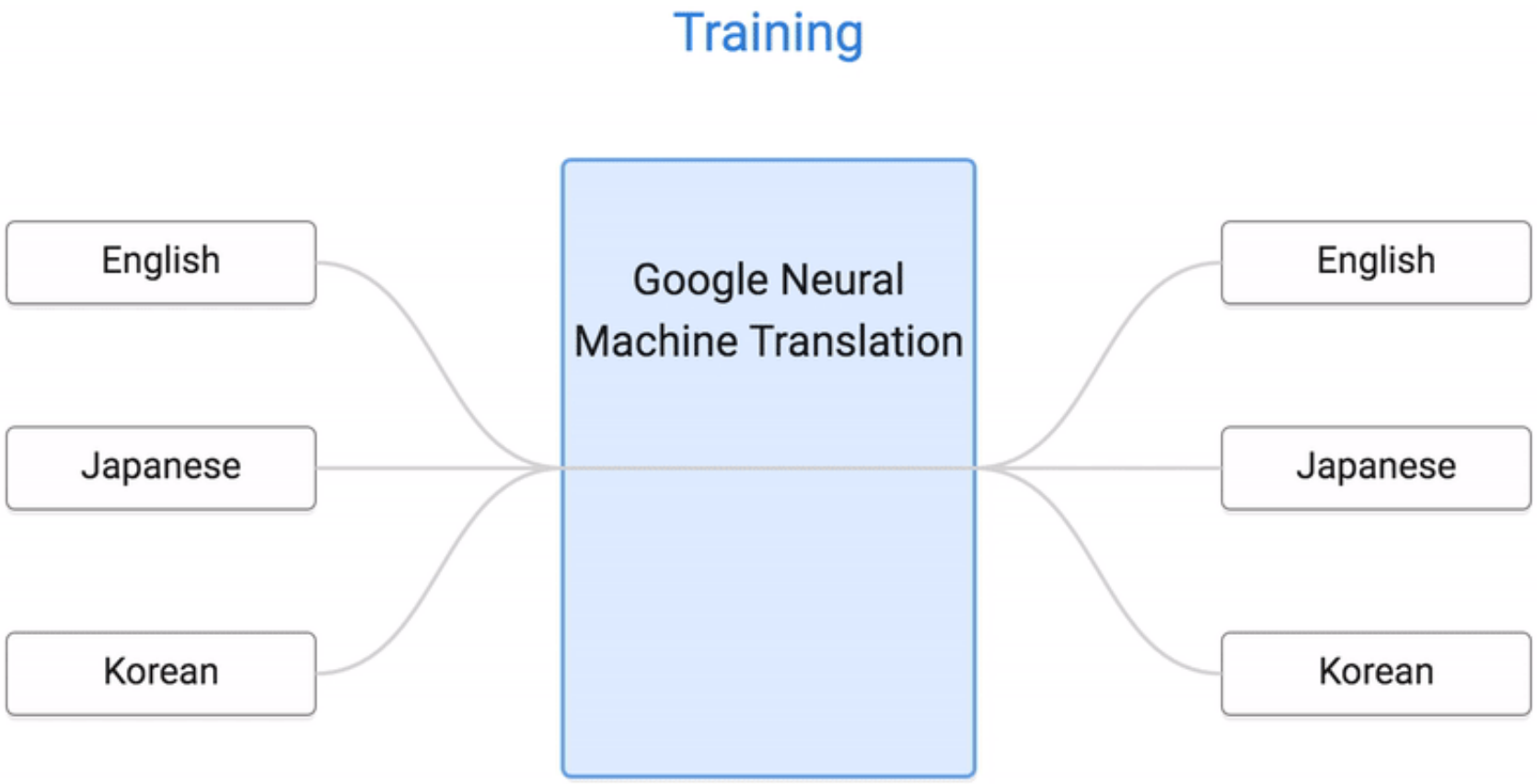
y^s: cat dog ……

x^t :  

In speech recognition, we can not have all possible words in the source (training) data.

How we solve this problem in speech recognition?

Example of Zero-Shot Learning



Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, arXiv preprint 2016

What is Transfer Learning

Target Data	Source Data (not directly related to the task): labelled	Source Data (not directly related to the task): unlabeled
labelled	Model Fine-tuning Multitask Learning	Self-taught learning Rajat Raina , Alexis Battle , Honglak Lee , Benjamin Packer , Andrew Y. Ng, Self-taught learning: transfer learning from unlabeled data, ICML, 2007
unlabeled	Domain-adversarial training Zero-shot learning	Self-taught Clustering Wenyuan Dai, Qiang Yang, Gui-Rong Xue, Yong Yu, "Self-taught clustering", ICML 2008

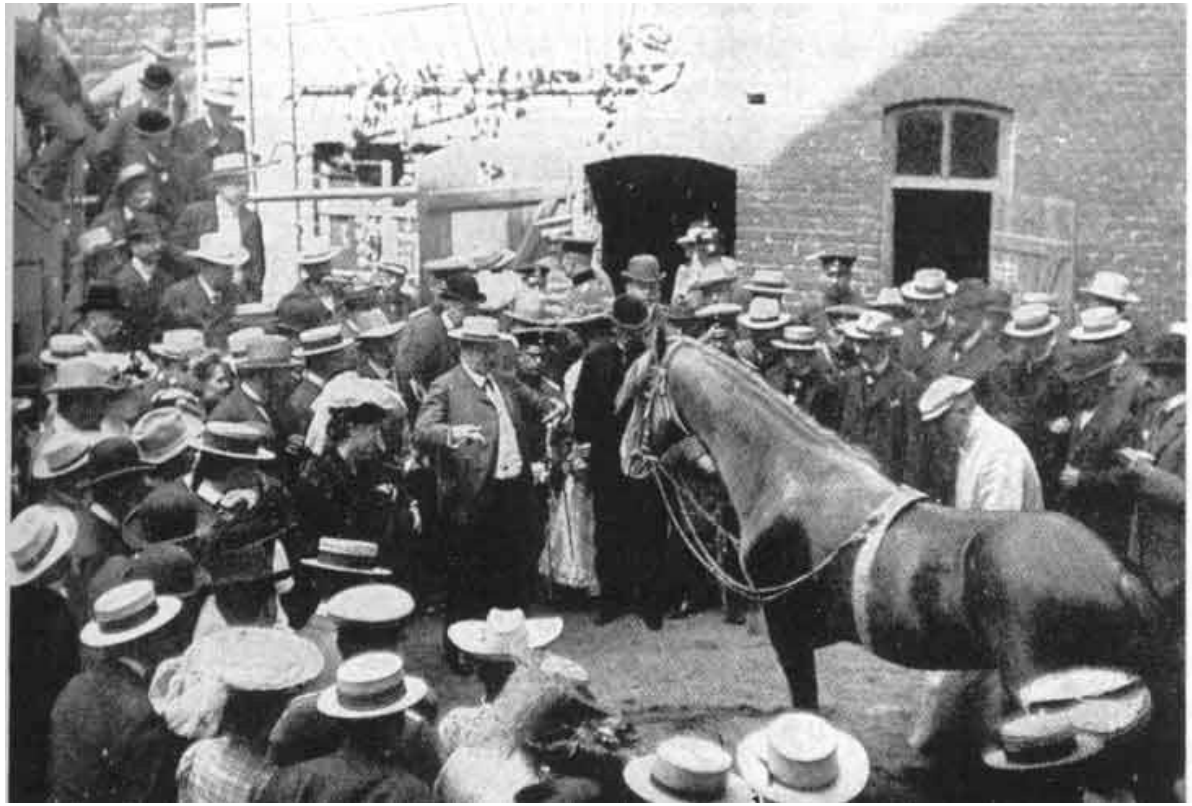
Different from semi-supervised learning

Explainable Machine Learning

Why we need Explainable ML?

- Correct answers ≠ Intelligent

Clever Hans performing in 1904



Why we need Explainable ML?

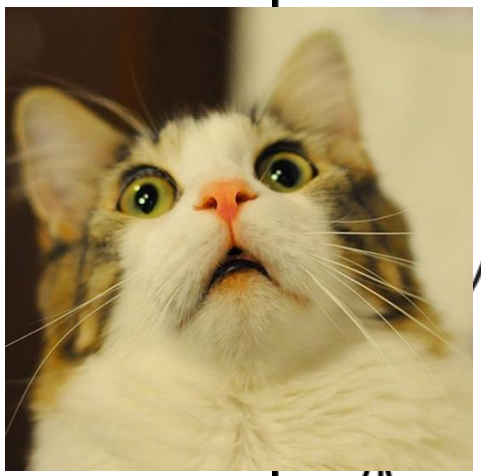
Loan issuers are required by law to explain their models.

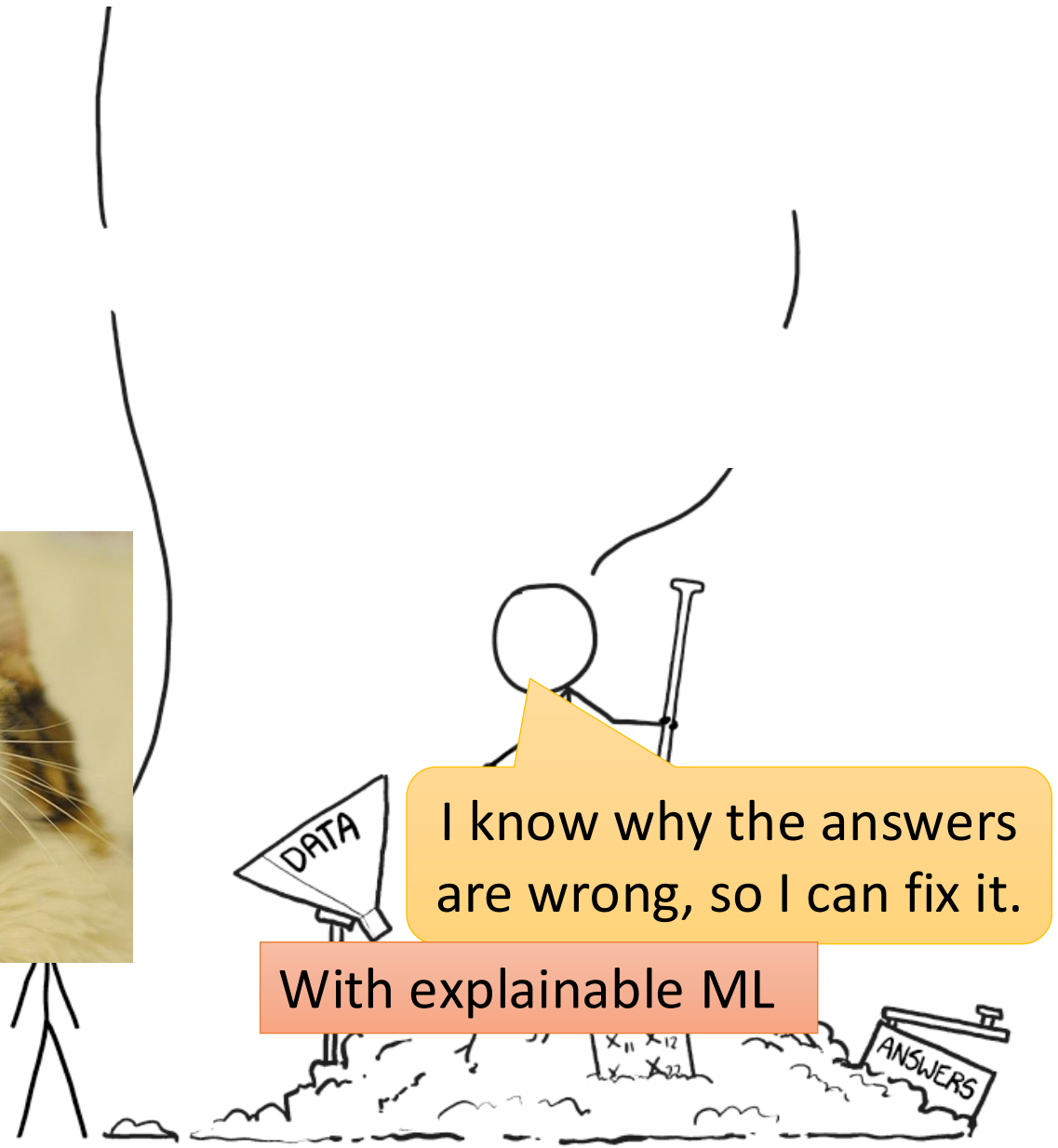
Medical diagnosis model is responsible for human life. Can it be a black box?

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.





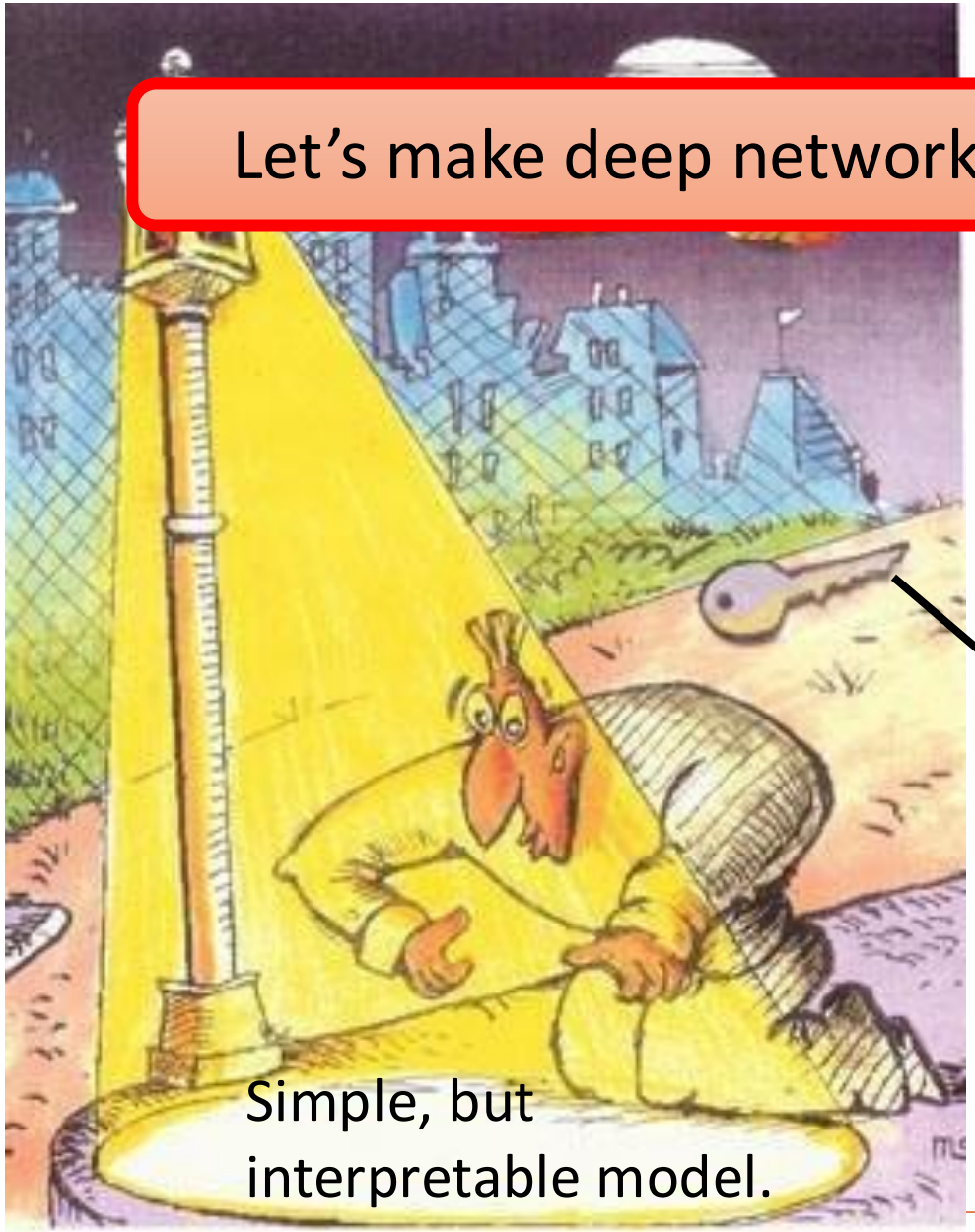
https://www.explainxkcd.com/wiki/index.php/1838:_Machine_Learning

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."



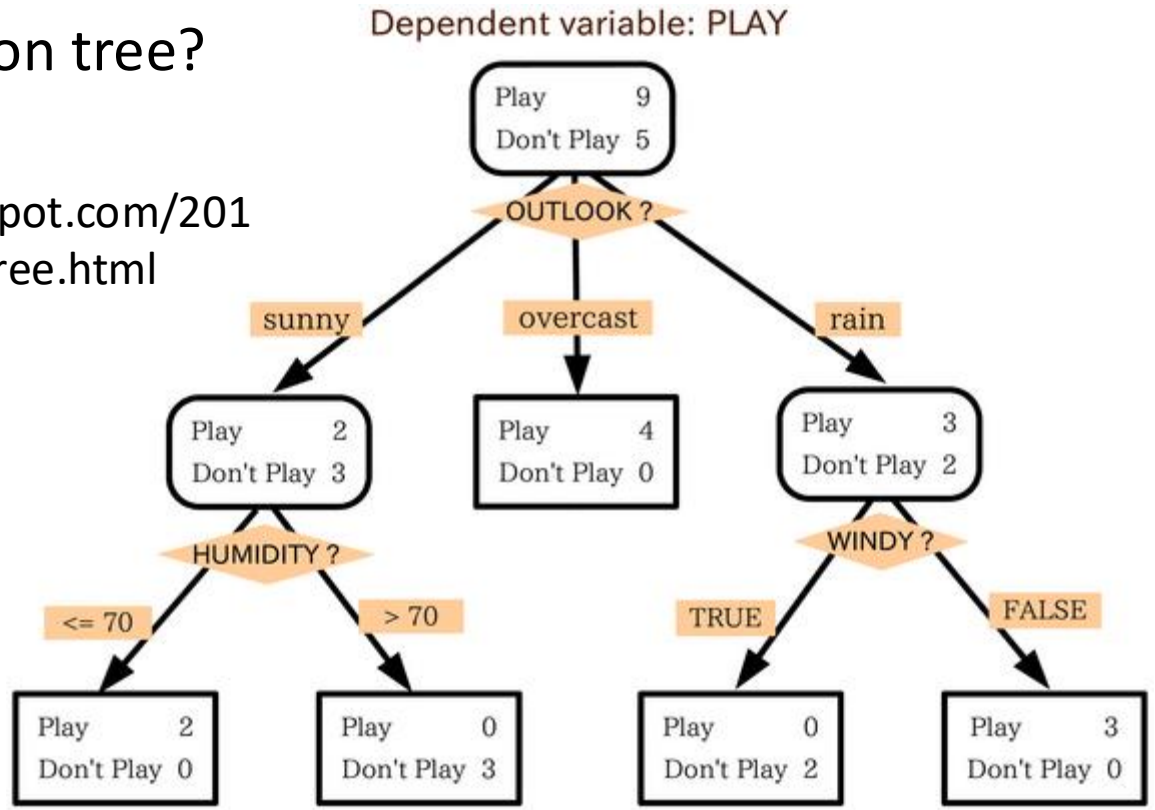
Powerful model

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?

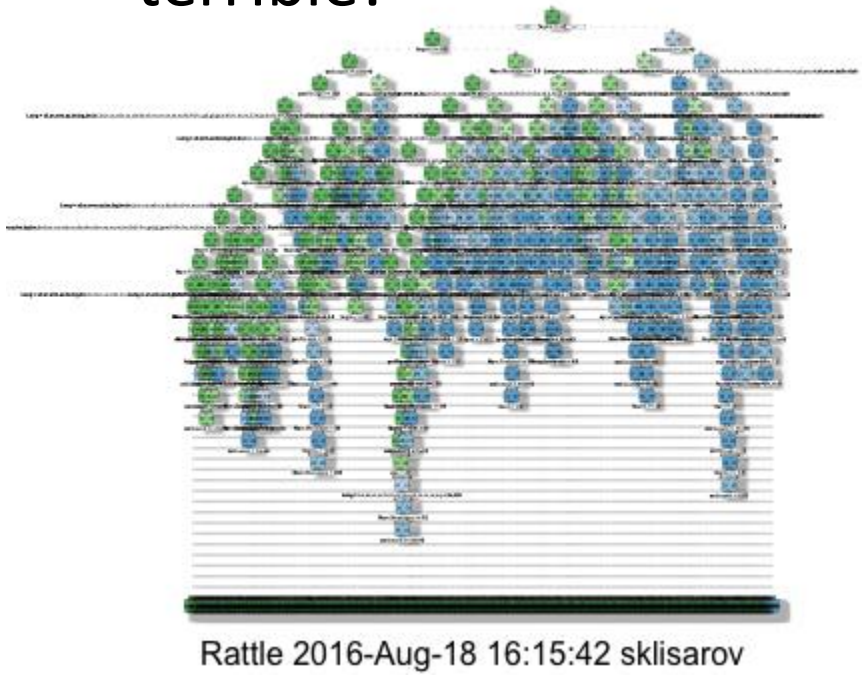
Source of image:
https://mropengate.blogspot.com/2015/06/ai-ch13-2-decision-tree.html



Decision tree is all you need!?

Interpretable v.s. Powerful

- A tree can still be terrible!



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

- We use a forest!

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!

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

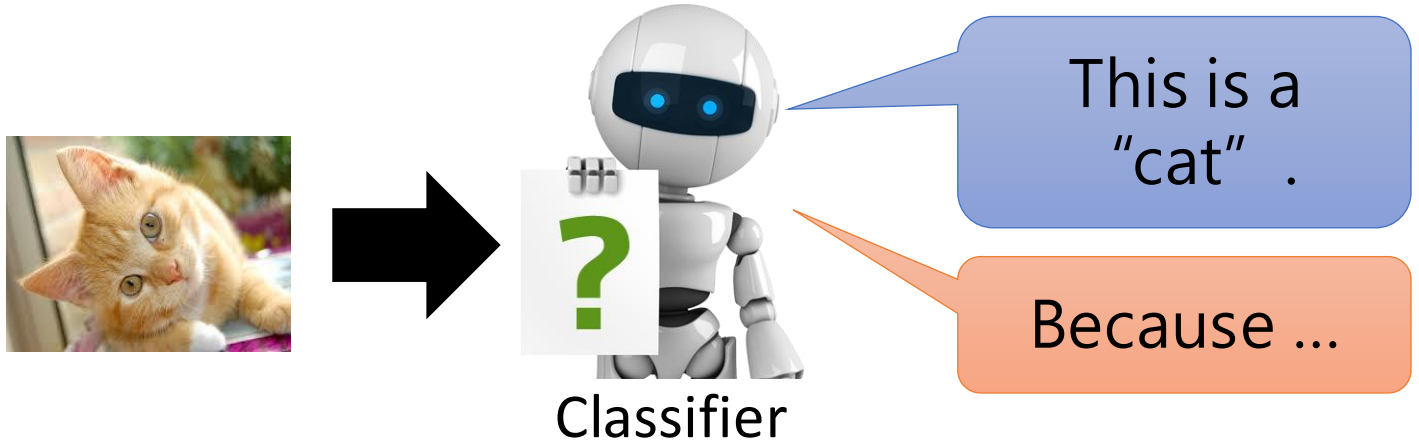
"Excuse me, I have 5 pages. May I use the Xerox machine, **because I have to make copies**?"

93% accept

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

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

Explainable ML



Local Explanation

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

Global Explanation

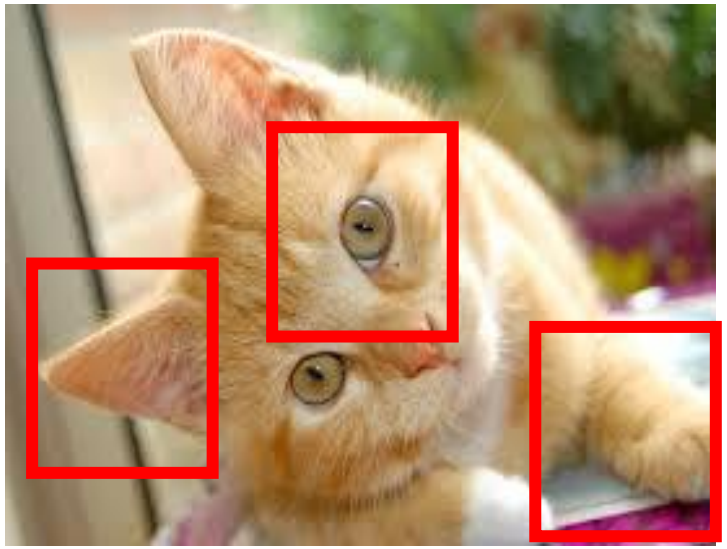
What does a "cat" look like?

(not referred to a specific image)

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?

Object x ⟶

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

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

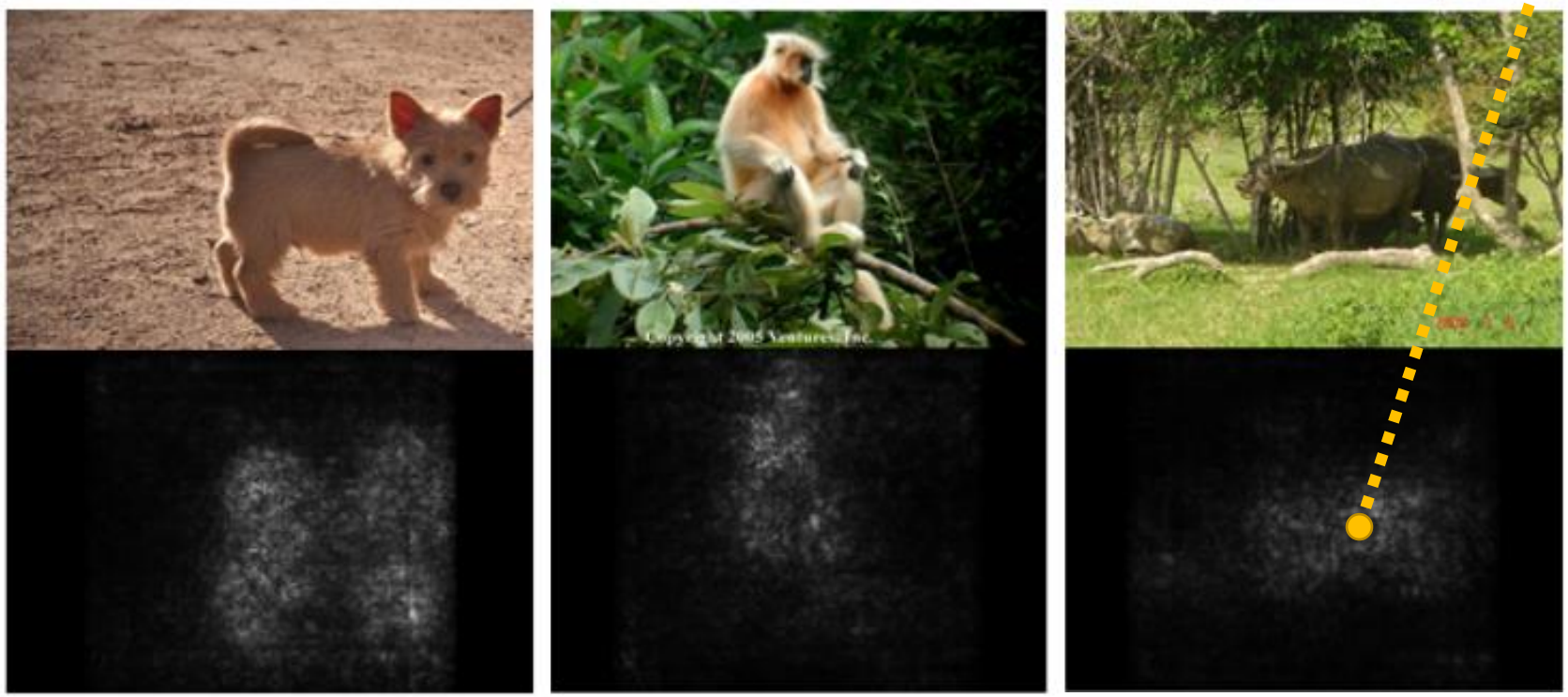
$\{x_1, \cdots, x_n, \cdots, x_N\} \rightarrow \{x_1, \cdots, x_n + \Delta x, \cdots, x_N\}$

pixels

$e \rightarrow e + \Delta e$

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

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



Saliency Map

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

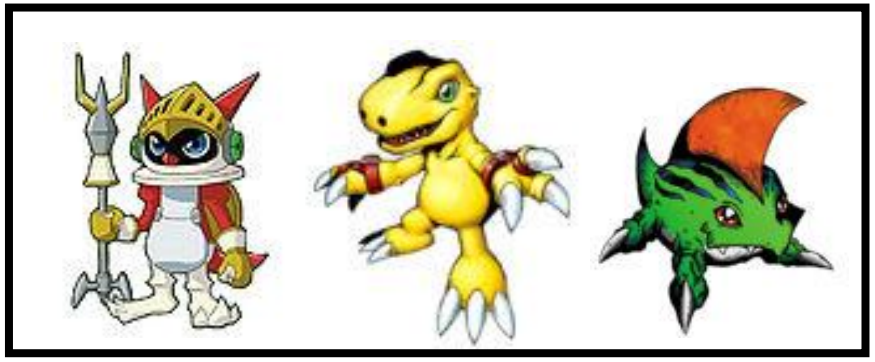
https://medium.com/@tyreeostevenson/teaching-a-computer-to-classify-anime-8c77bc89b881

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

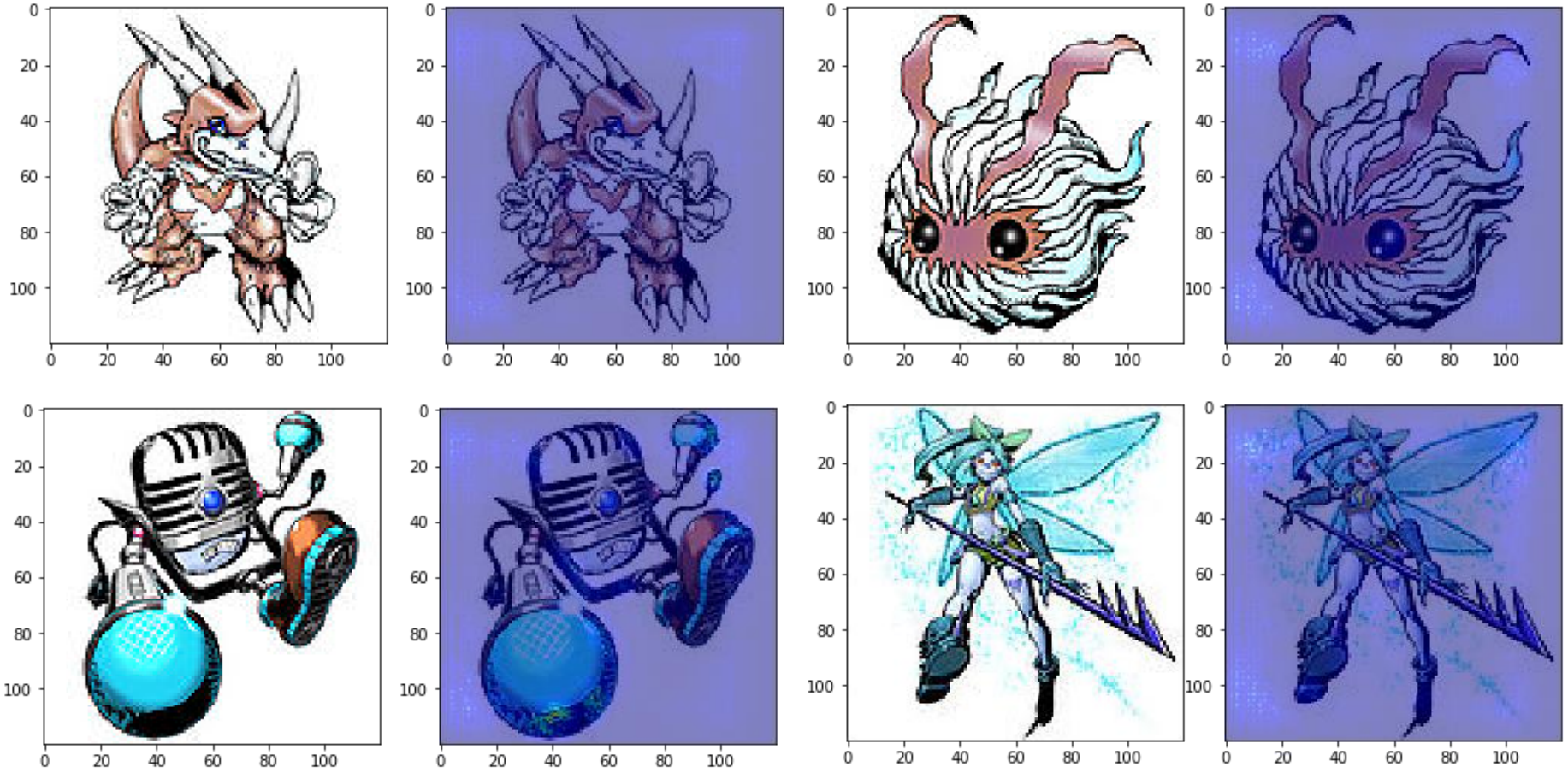
```
model = Sequential()
model.add(Conv2D(32, (3, 3), padding='same', input_shape=(120,120,3)))
model.add(Activation('relu'))
model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

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%

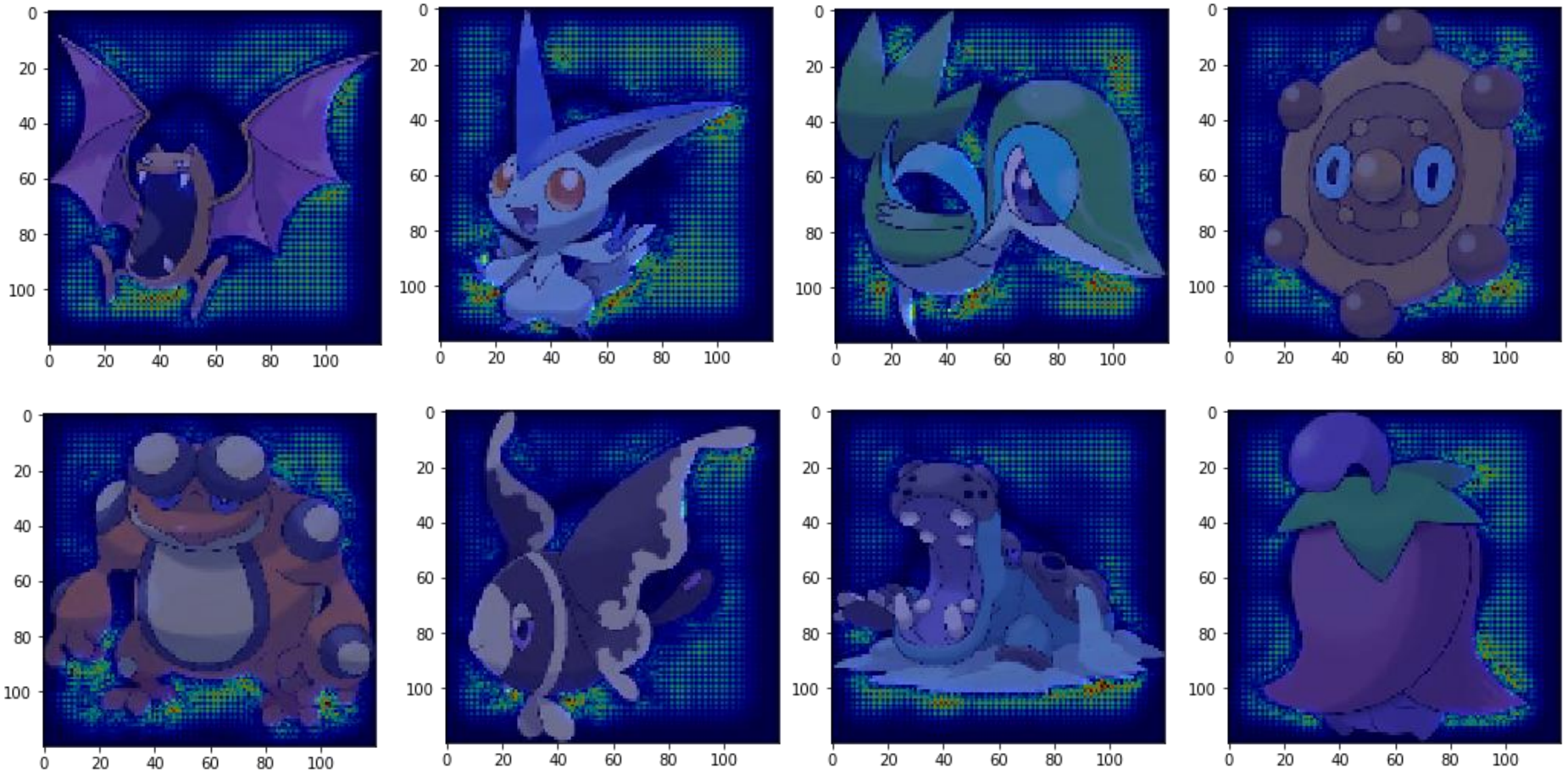
Testing Accuracy: 98.4%

Amazing!!!!!!

Saliency Map



Saliency Map



What Happened?

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

loading the files

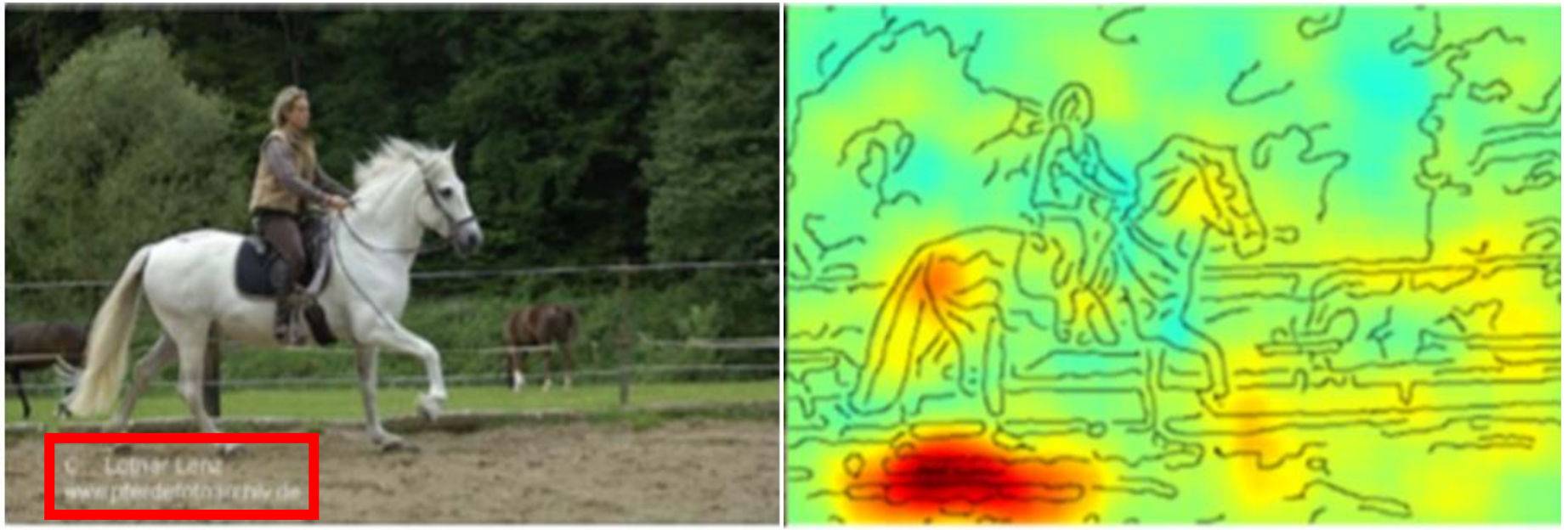
png files have transparent background

transparent background becomes black

Machine discriminates Pokémon and Digimon based on the background colors.

More Examples …

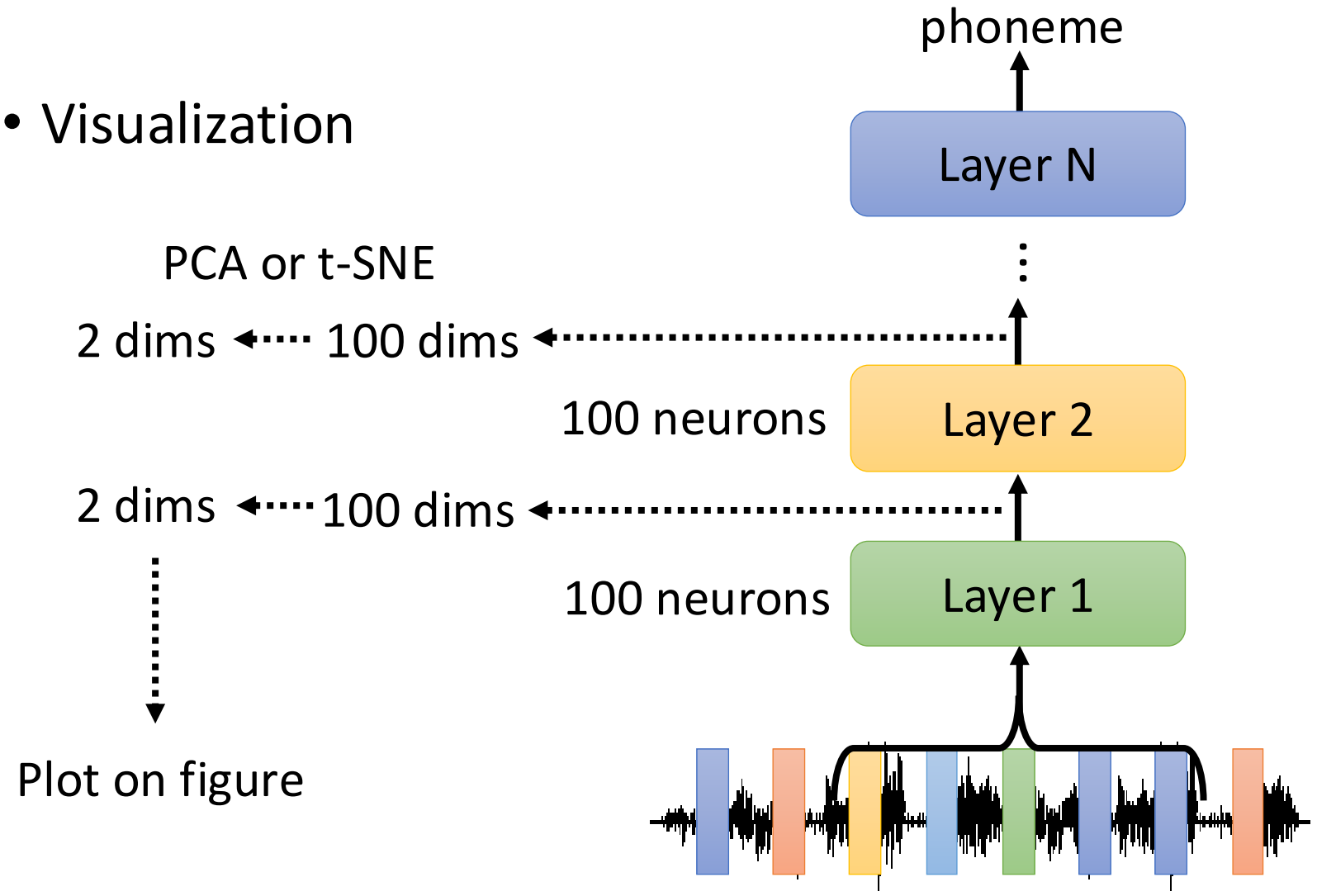
- PASCAL VOC 2007 data set



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

How a network processes the input data?

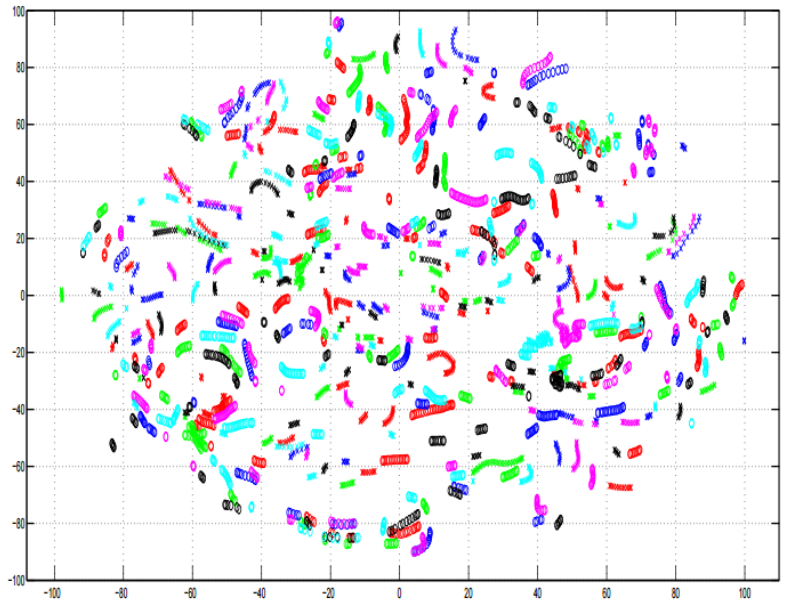
- Visualization



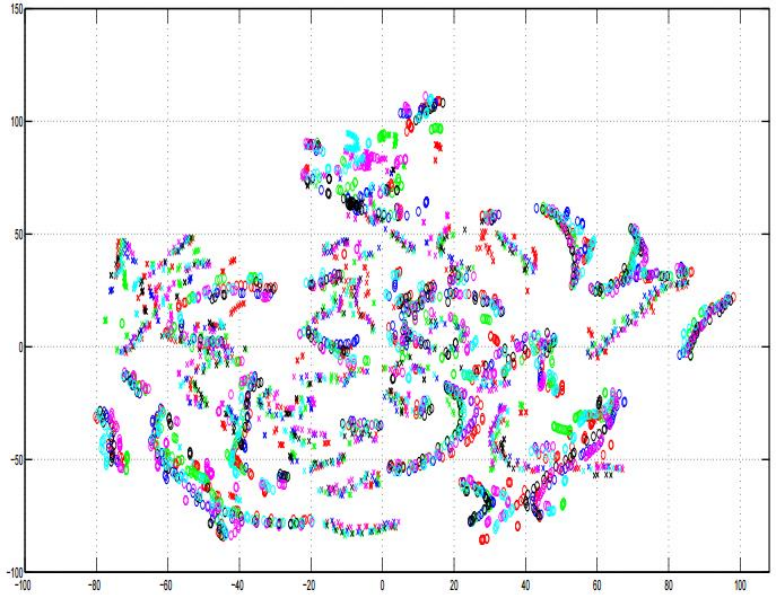
How a network processes the input data?

- Visualization
  Colors: speakers

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



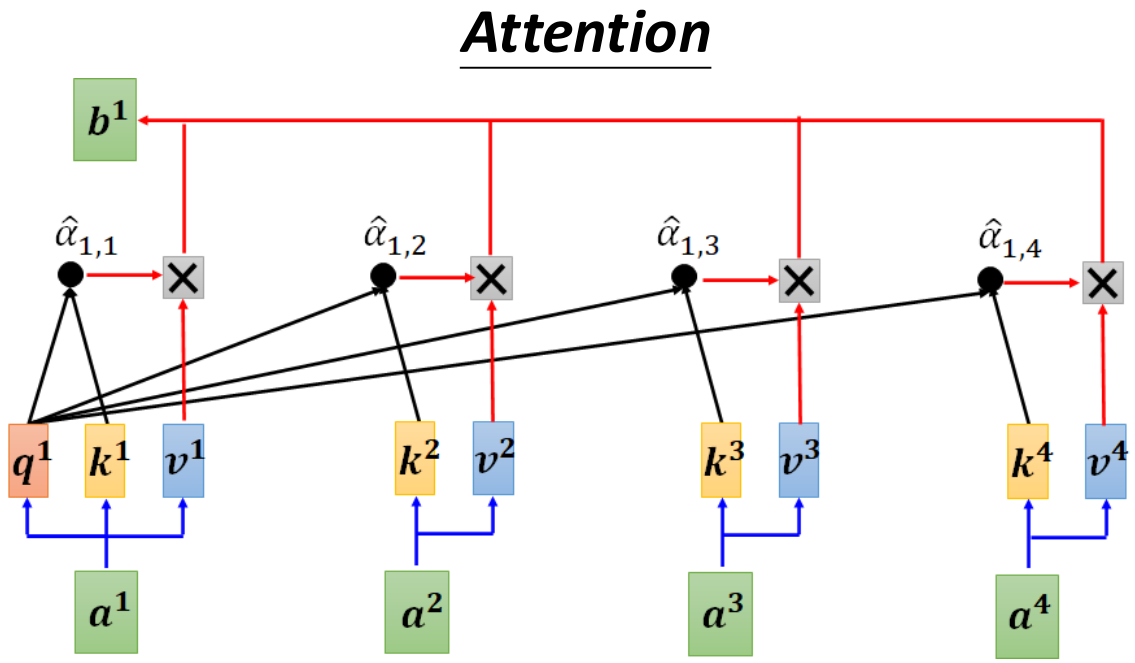
Input Acoustic Feature (MFCC)



8-th Hidden Layer

How a network processes the input data?

- Visualization

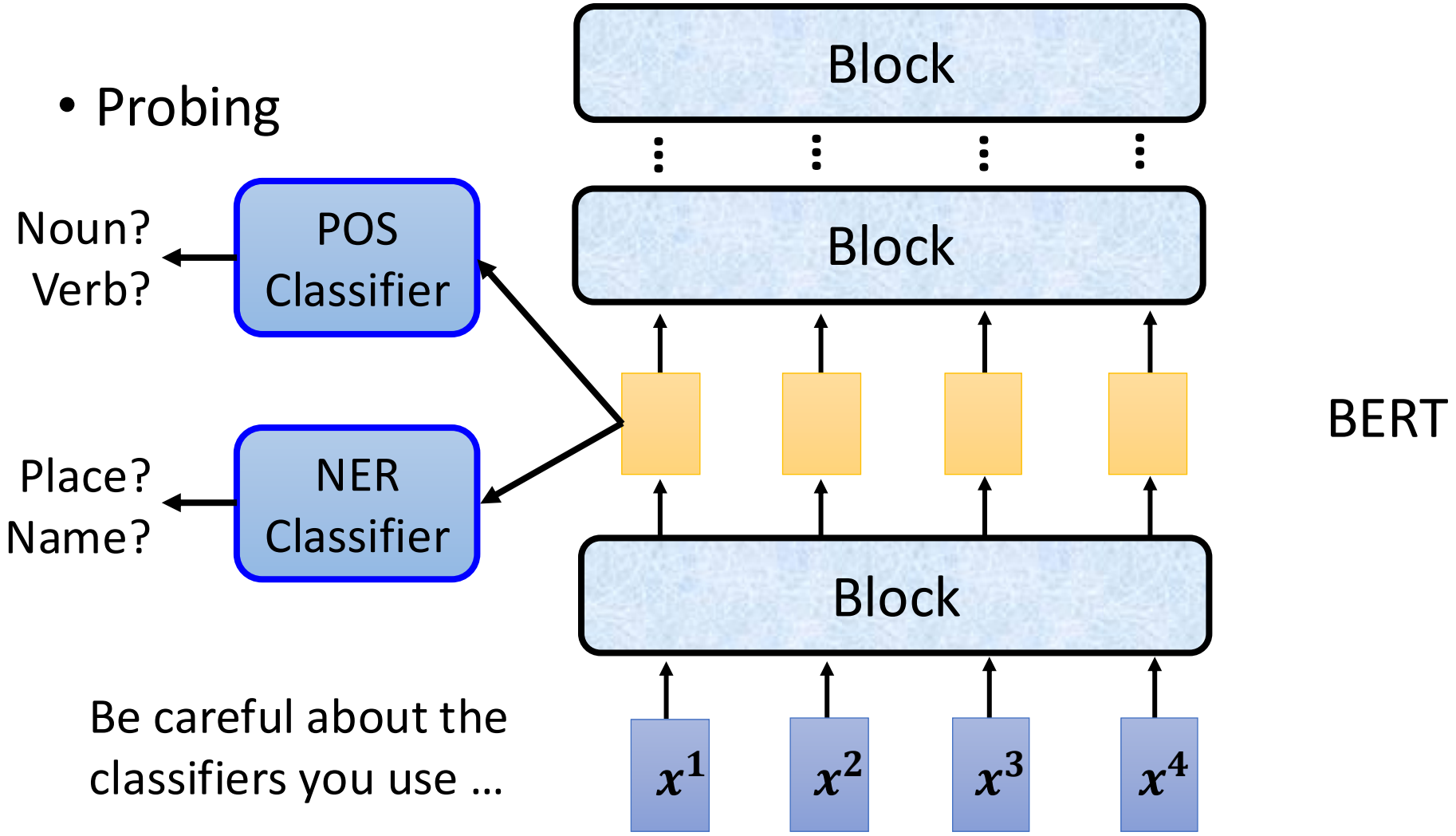


Attention is not Explanation
https://arxiv.org/abs/1902.10186 (02. 2019)

Attention is not not Explanation
https://arxiv.org/abs/1908.04626 (05. 2019)

How a network processes the input data?

- Probing



BERT

Be careful about the classifiers you use …

Global Explanation: Explain the whole Model

Question: What does a “cat” look like?

What does a filter detect?

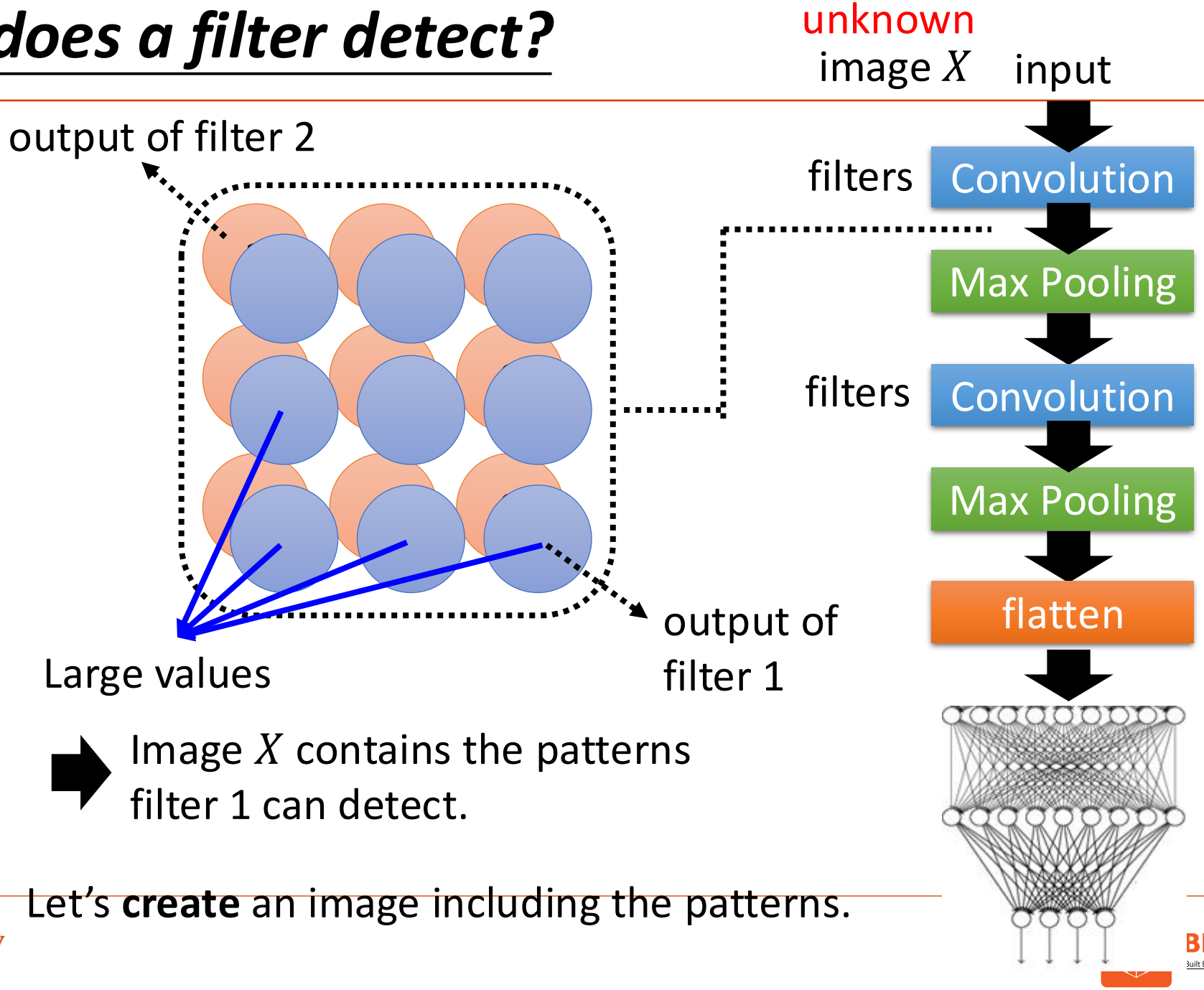


Image X contains the patterns filter 1 can detect.

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

What does a filter detect?

unknown image X

input

filters

Convolution

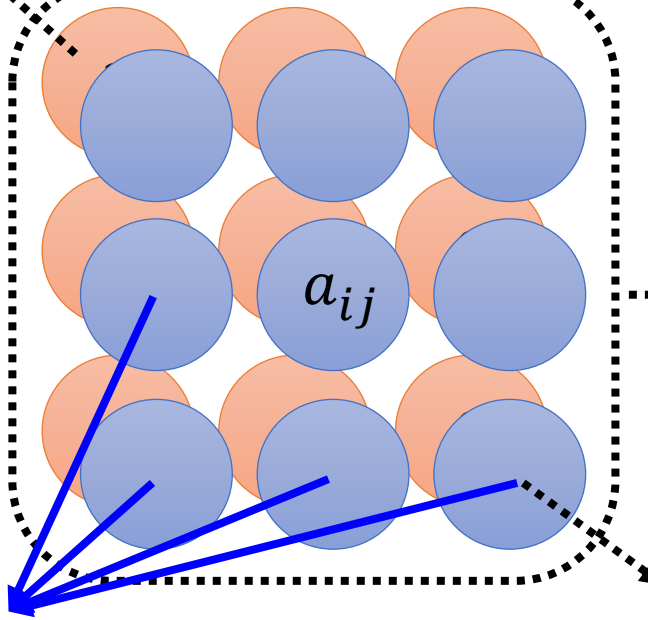
Max Pooling

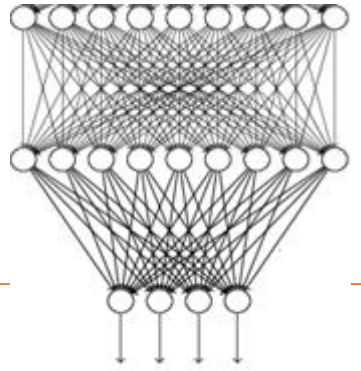
filters

Convolution

Max Pooling

flatten





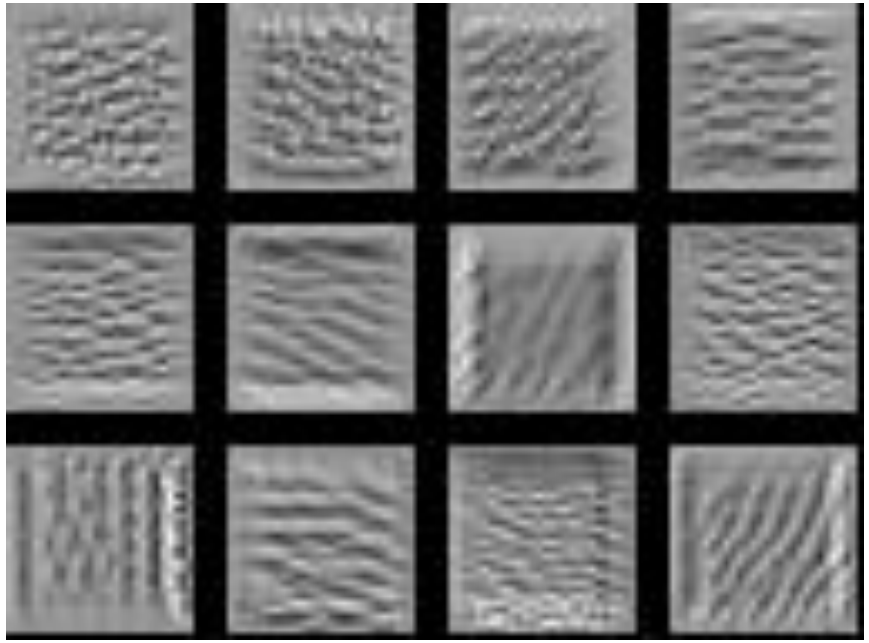
$$X^* = arg\max_X \sum_i \sum_j a_{ij}$$ (gradient ascent)

The image contains the patterns filter 1 can detect.

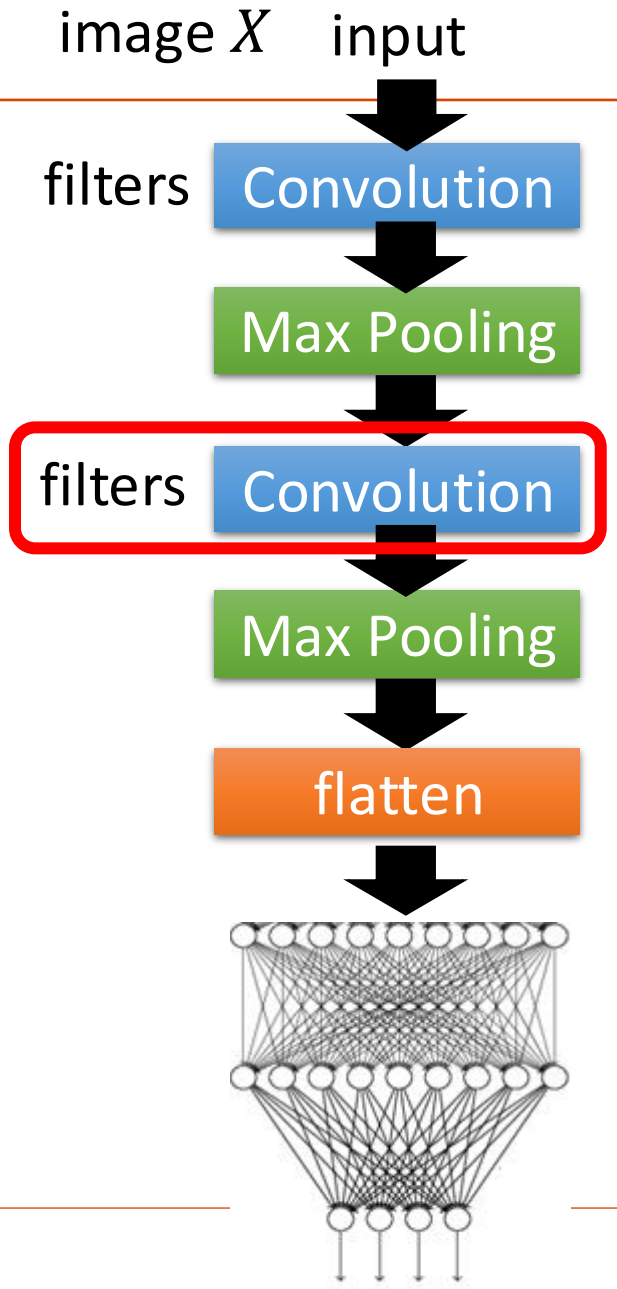
What does a filter detect?

E.g., Digit classifier

X^* for each filter



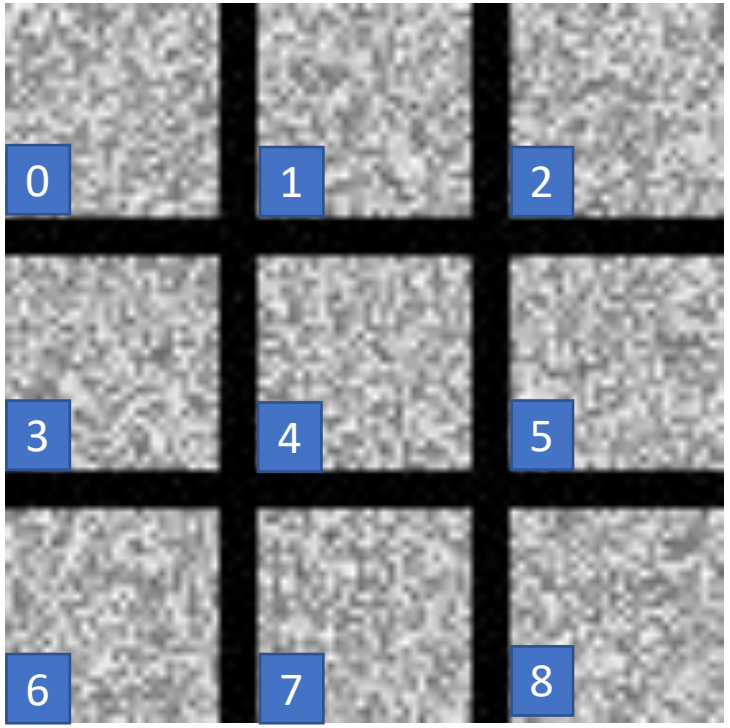
Use NMIST handwrite database

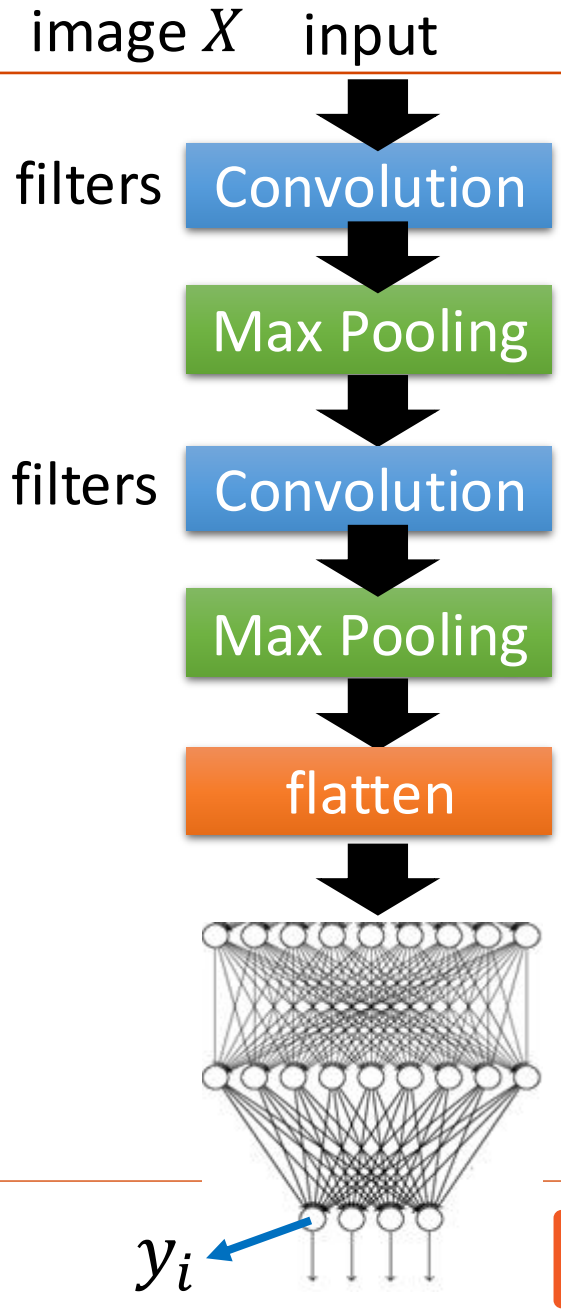


What does a digit look like for CNN?

E.g., Digit classifier

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



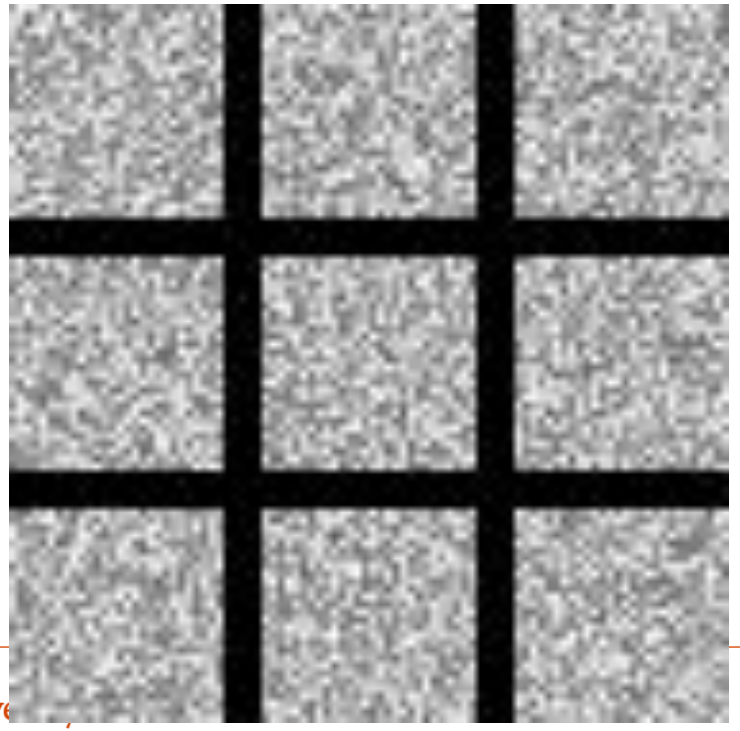


Surprise? Consider adversarial attack!

What does a digit look like for CNN?

Find the image that maximizes class probability

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

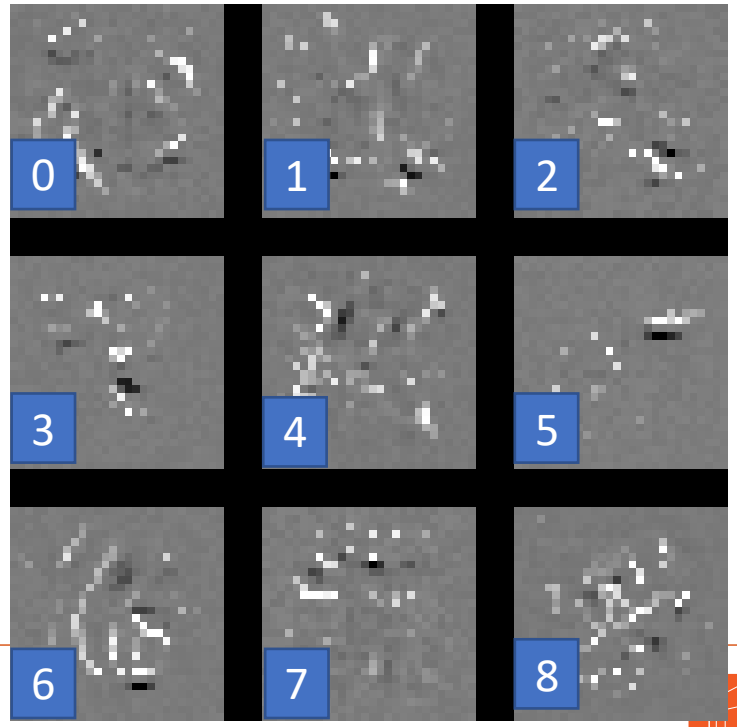


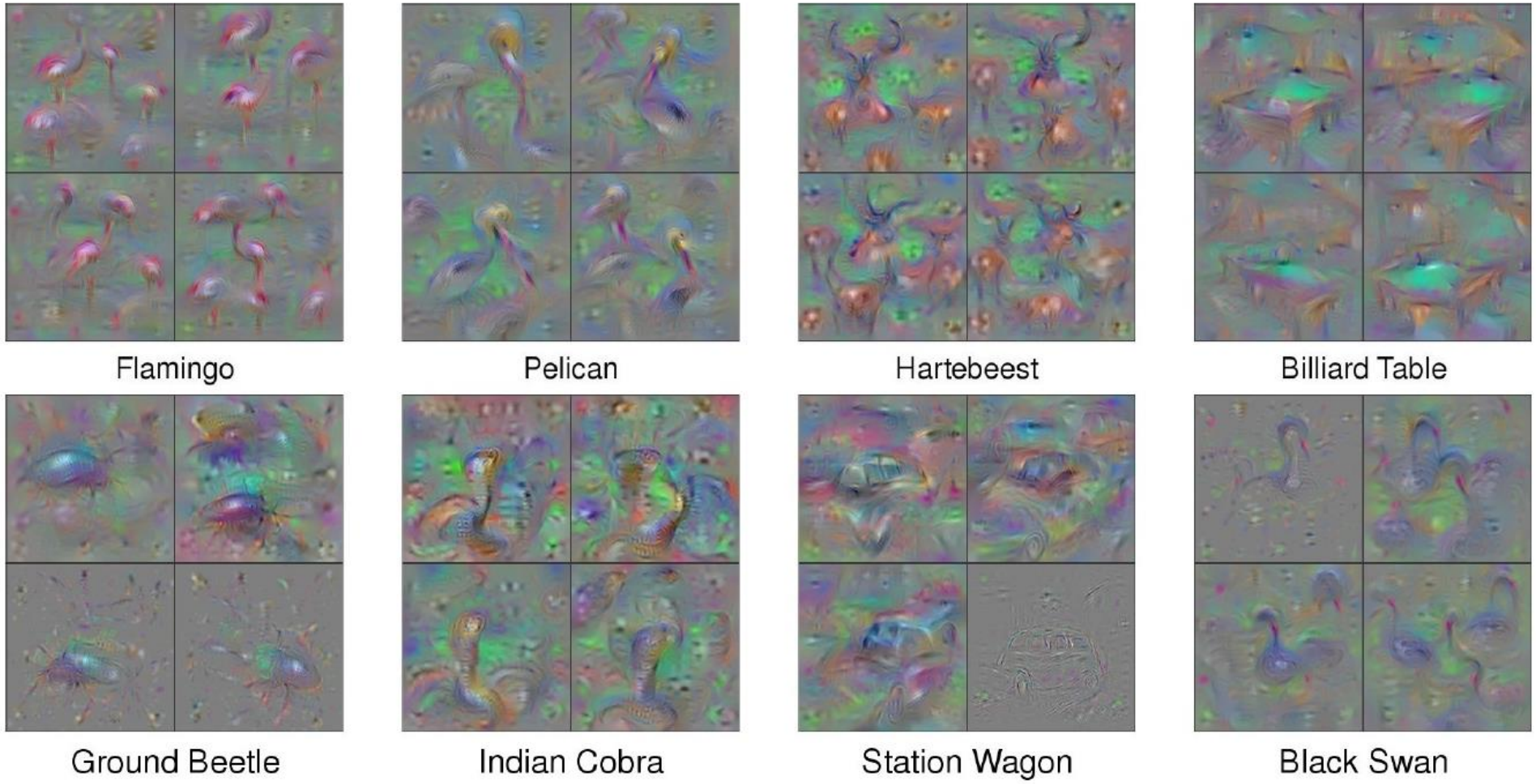
The image should looks like a digit.

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

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

How likely X is a digit





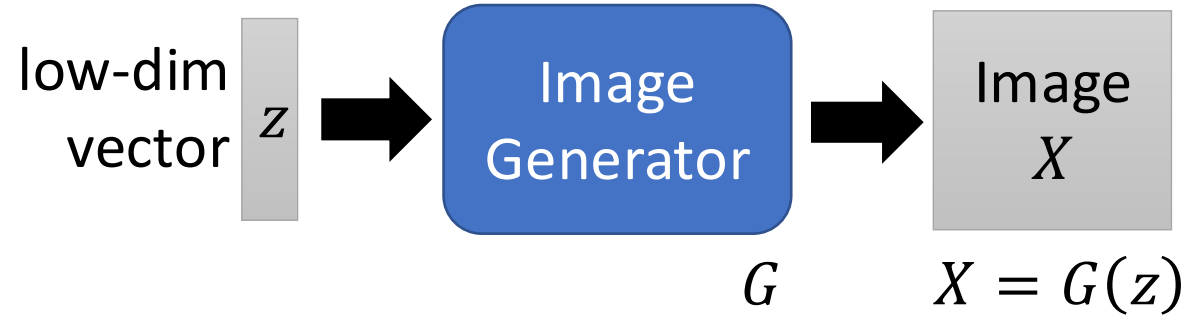
With several regularization terms, and hyperparameter tuning …..

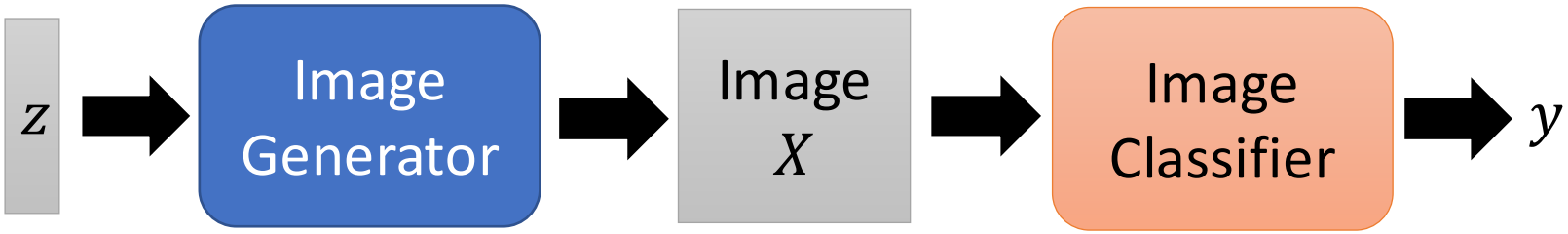
https://arxiv.org/abs/1506.06579 (2015)

Constraint from Generator

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

Training Examples





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

Show image:

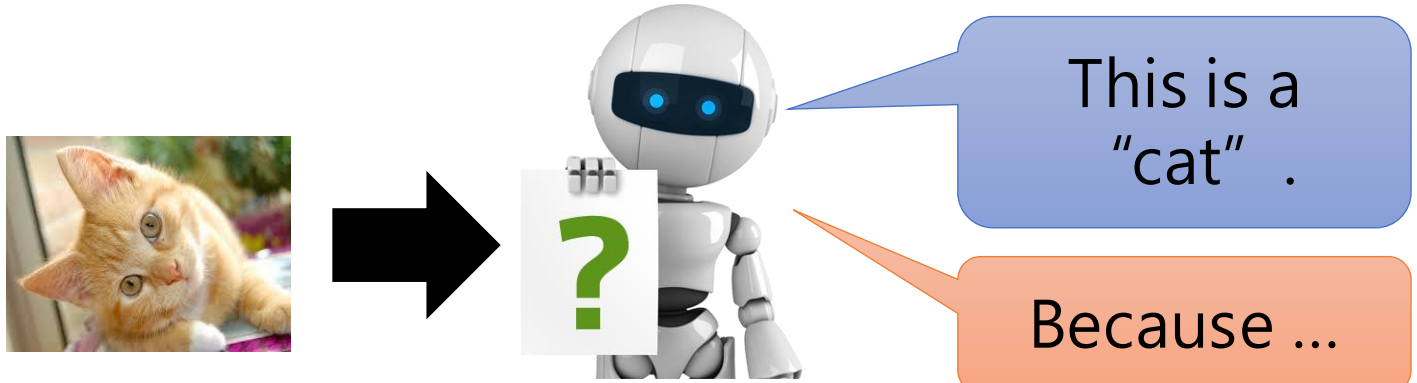
$$X^* = G(z^*)$$

Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

（2017）

https://arxiv.org/abs/1612.00005

Conclusion 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)

Any Questions?

bidong@syr.edu