On The Utility of Conditional Generation Based Mutual Information For Characterizing Adversarial Subspaces

Chia–Yi Hsu*, Pei–Hsuan Lu*, Pin–Yu Chen†, and Chia–Mu Yu*

*National Chung Hsing University, Taiwan

†IBM Research, USA

IBM Research

國立中興大學
National Chung Hsing University

Applications of Neural Networks

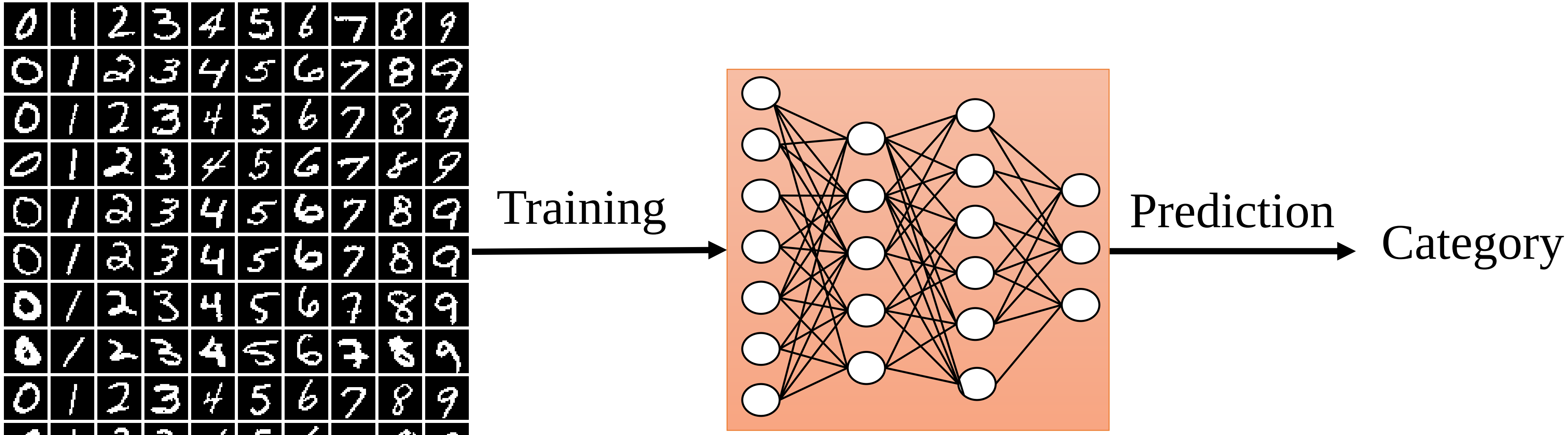
- Game-playing

[https://www.youtube.com/watch?v=Ipi40cb_RsI]

- Autonomous Vehicle

[https://www.youtube.com/watch?v=0rc4RqYLtEU]

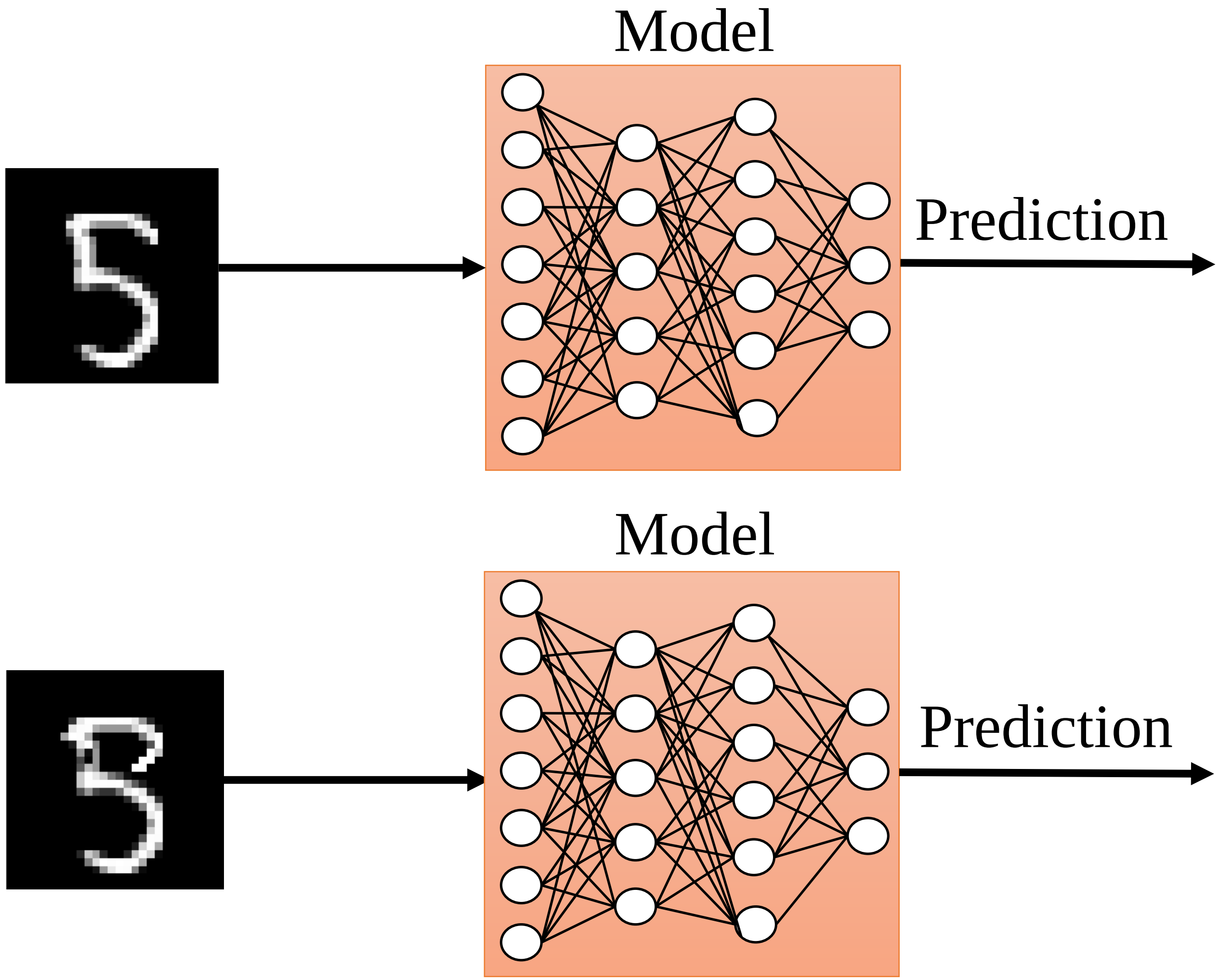
Neural networks as classifiers



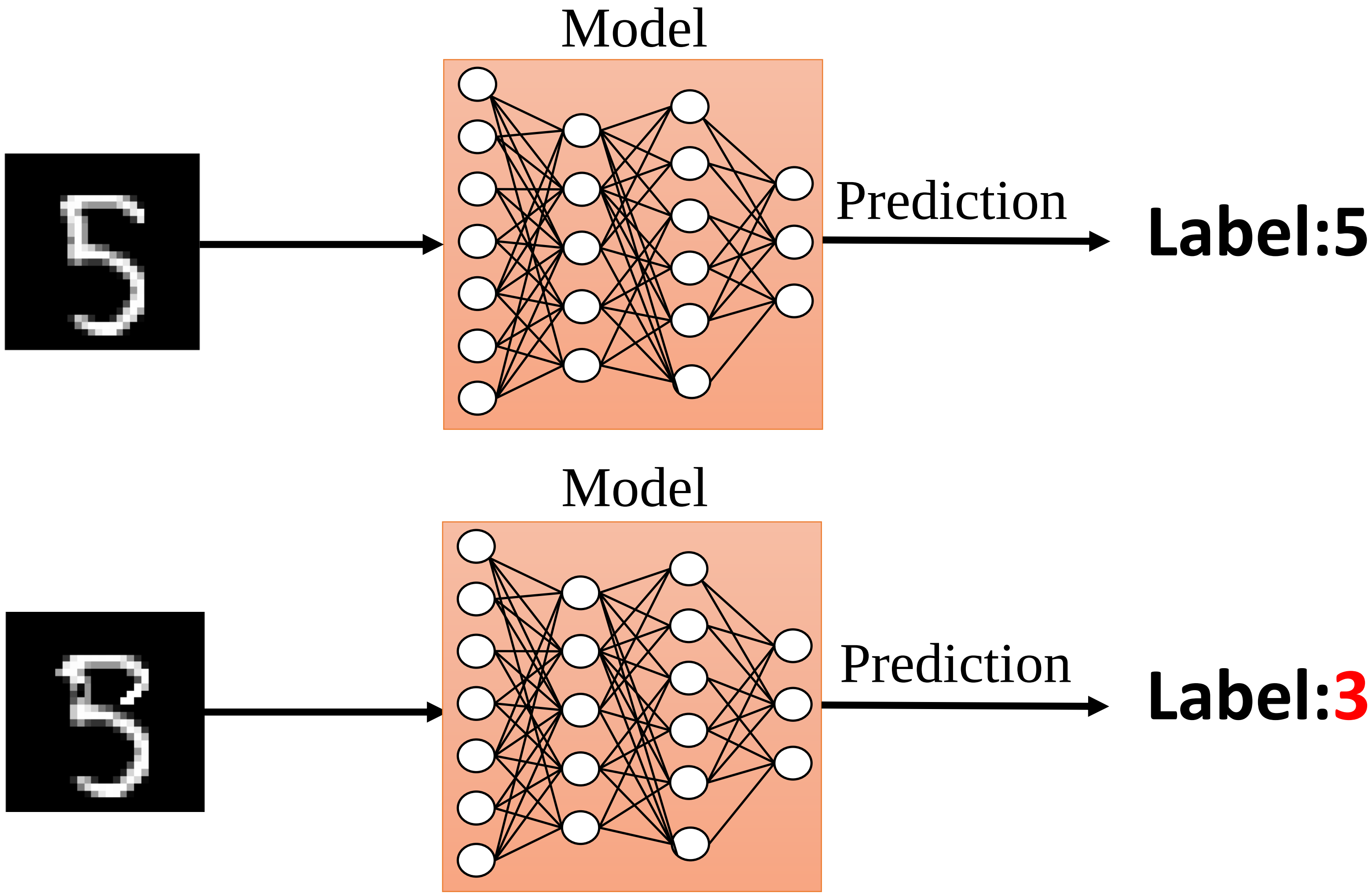
Classification Model

Handwritten Digits
from the MNIST dataset

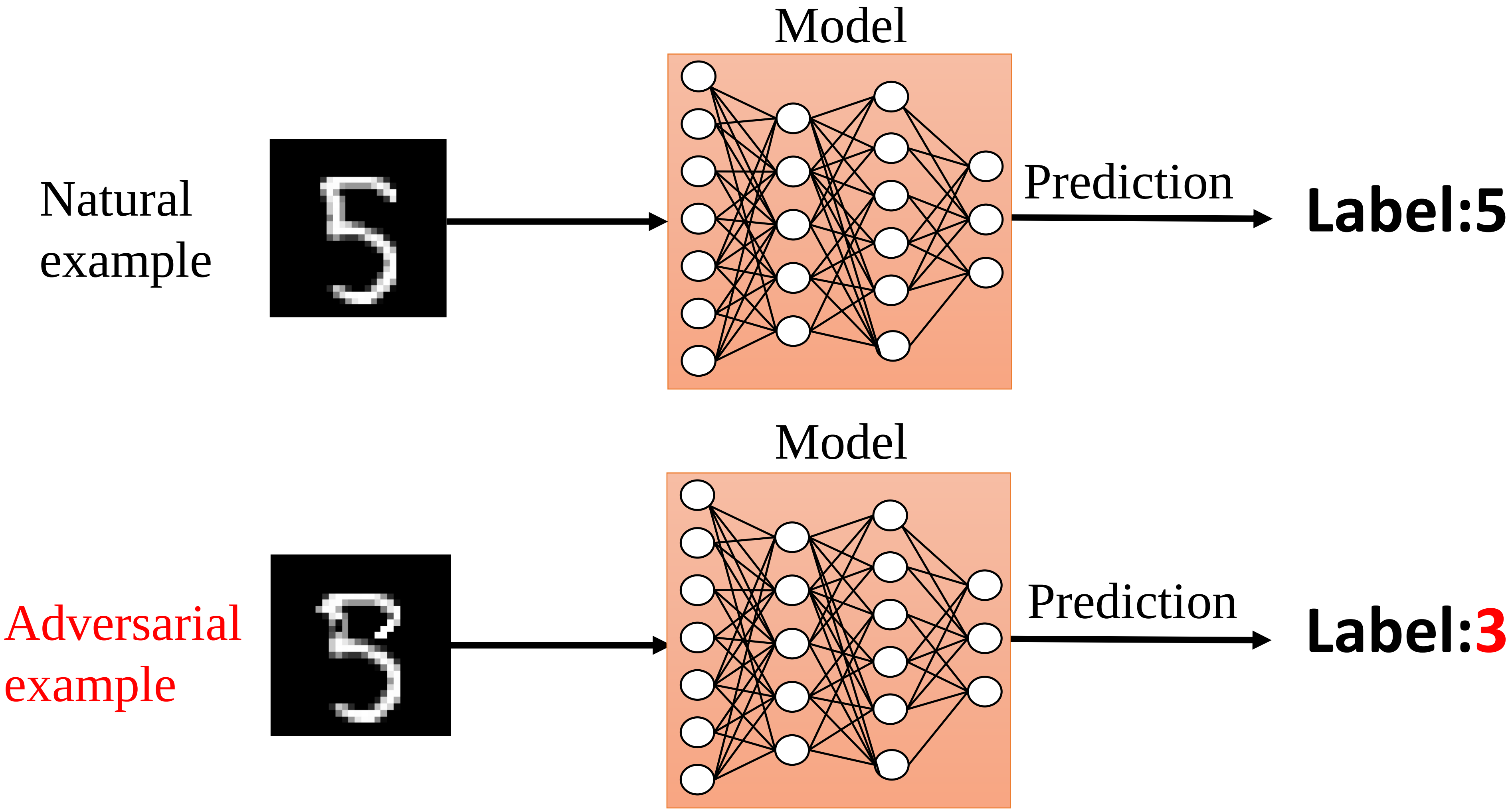
Adversarial Examples



Adversarial Examples

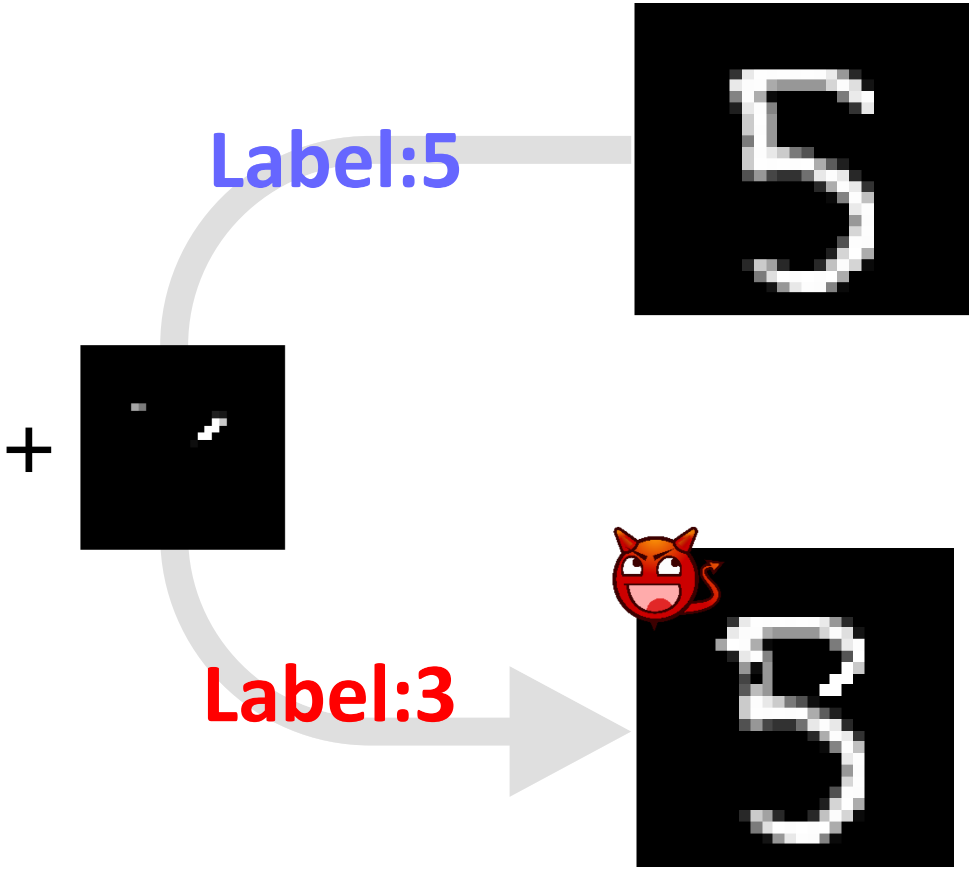


Adversarial Examples



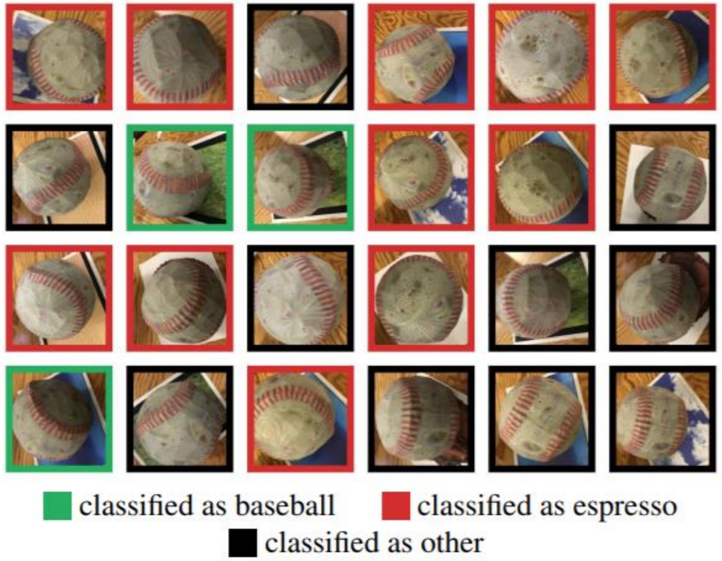
Adversarial Examples

- Look like natural examples
- Cause misclassification



Real word adversarial example

Physical Adversarial Sticker Perturbations for YOLO [Eykholt, 2018]



Adversarial baseball [Athalye, 2017]

Carlini and Wanger's attack (C&W attack) [Carlini, 2017]

- Optimization-based method with carefully designed attack loss (L_2)
- Minimize$_x$ $\|x - x_0\|_2 + c \cdot f(x, t_0)$ s.t. $x_0 + \delta \in [0,1]^p$

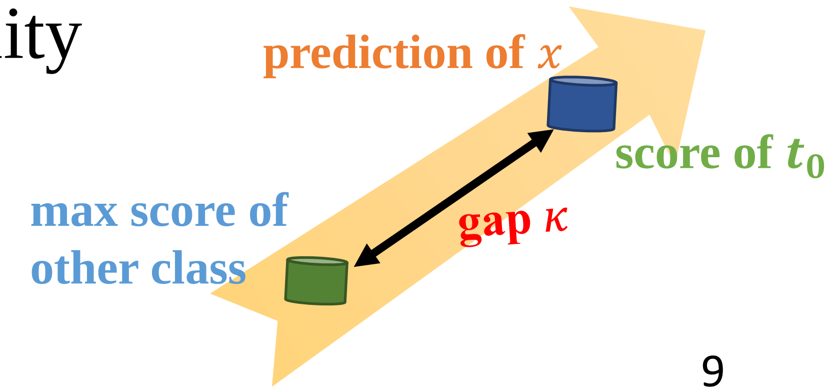
confidence

most probable class prediction other than t_0

- $f(x, t_0) = max\{-\kappa, [Z(x)]_{t_0} - max_{i \neq t_0}[Z(x)]_i\}$ (untargeted attack)

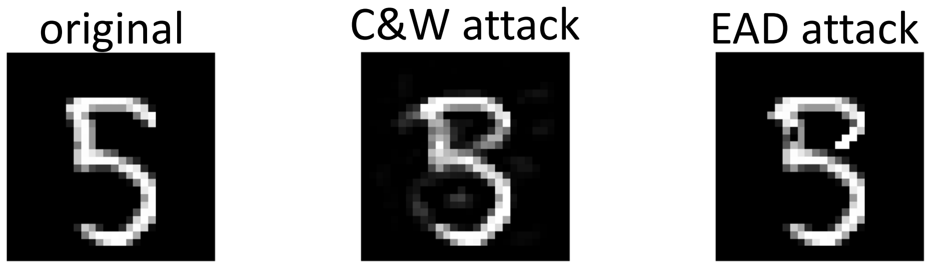
ground truth label's probability

- $\kappa \geq 0$: confidence parameter for transferability



EAD: Elastic-net Attacks to DNNs [Chen, 2018]

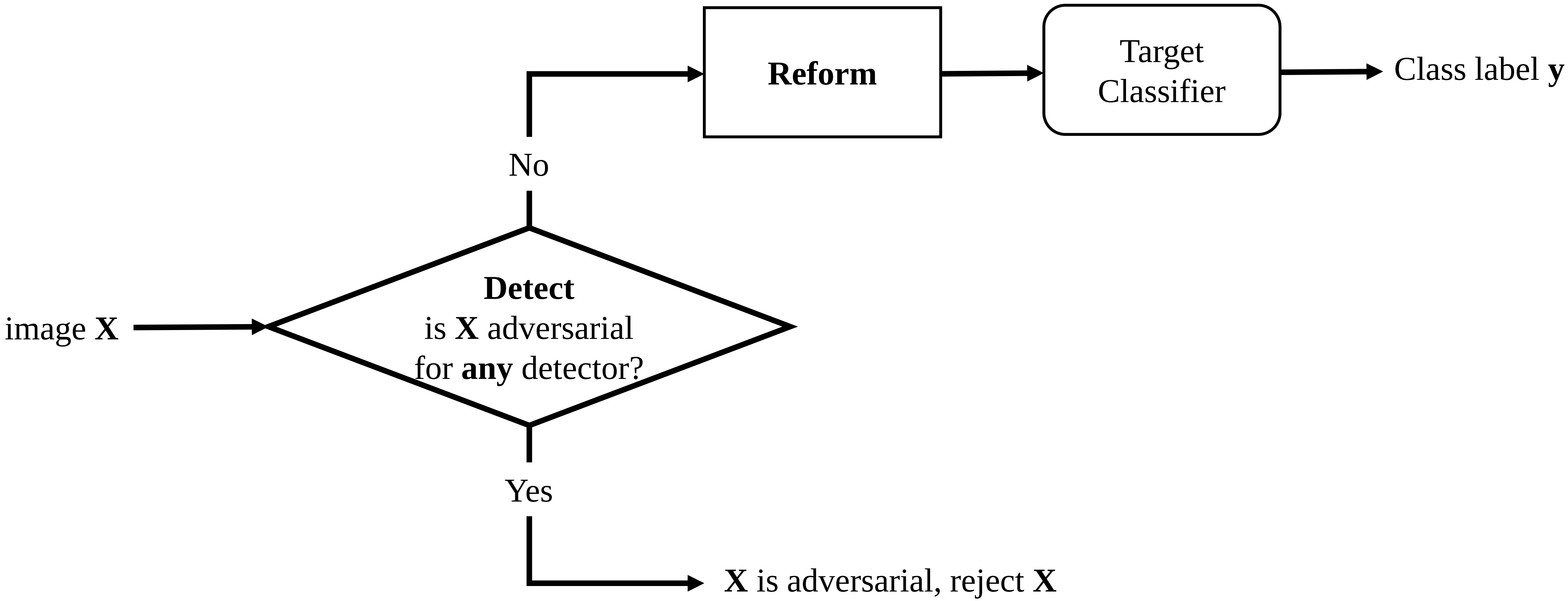
- Recall C&W attack: $\text{Minimize}_x \ \|x - x_0\|_2 + c \cdot f(x, t_0)$ s.t. $x_0 + \delta \in [0,1]^p$
- EAD: $\text{Minimize}_x \ \boldsymbol{\beta\|x - x_0\|_1} + \|x - x_0\|_2 + c \cdot f(x, t_0)$ s.t. $x_0 + \delta \in [0,1]^p$
- The advantages of EAD attack(L_1 regularizer):
  - $\|x - x_0\|_1 = \|\delta\|_1$ is a convex regularizer that encourages sparsity and hence transferability in the adversarial perturbation δ
  - Craft adversarial images while denoising unnecessary noises towards more effective attacks



Defense of adversarial example

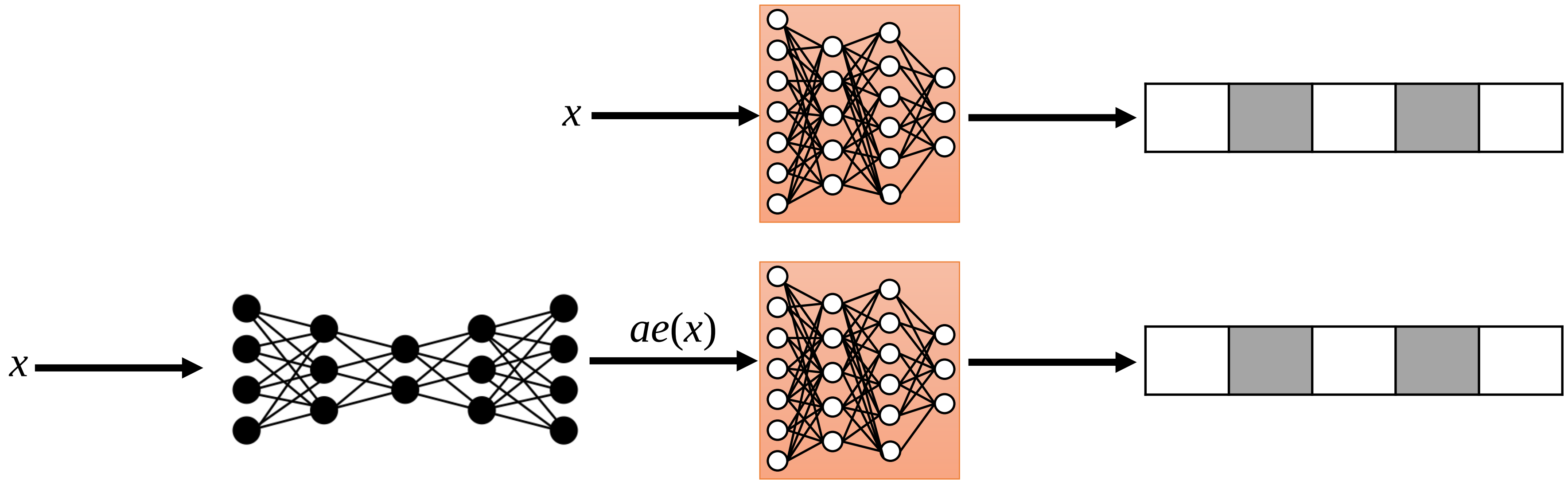
- Detection approach
  - Separate natural examples and adversarial examples
- Manifold-based approach
  - Correcting adversarial examples by projection to data manifold
- Gradient masking
- Adversarial training
  - Iteratively retrain a DNN while augmenting adversarial examples

MagNet


Reform
Target Classifier
Class label **y**
No
Detect
is **X** adversarial
for **any** detector?
image **X**
Yes
X is adversarial, reject **X**

MagNet

- Detector
  - Based on reconstruction error: $\|x - \mathrm{ae}(x)\|_2$ < threshold, MagNet accepts input
  - Based on probability divergence: $D_{KL}(P||Q)$ < threshold, MagNet accepts input



Mutual Information Detector (MID)

$$I(X,Y) = H(Y) - H(Y|X)$$

$$H(Y|X) = -\sum_{x} p(x) \left(\sum_{y} p(y|x)\ log\ \ p(y|x) \right)$$

Conditional generation

$$H(Y) = -\sum_{y} \log p(y)\, p(y)$$

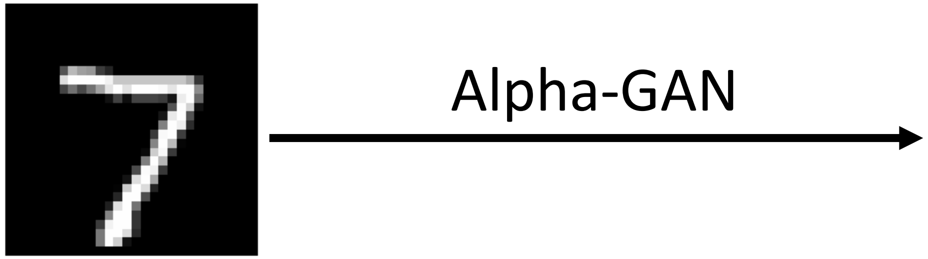
Classifier

$X = f(x)$

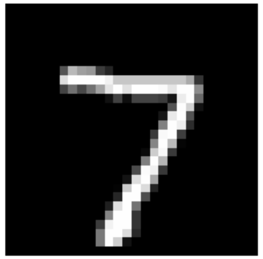
Auto-encoder

$Y = f(ae(x))$

Conditional generation



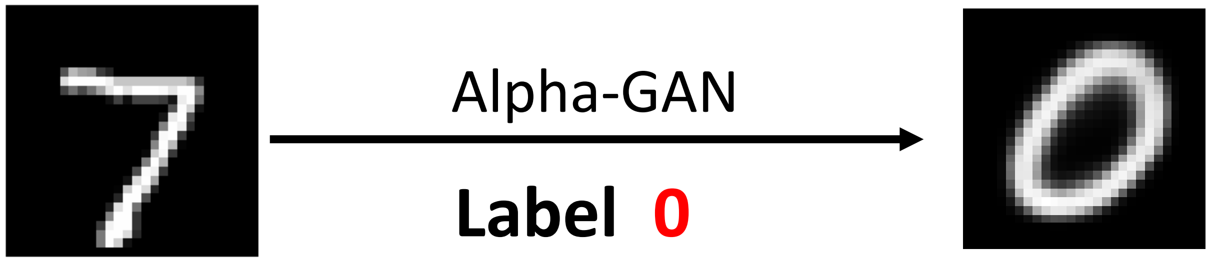
Conditional generation



Alpha-GAN

Label 0

Conditional generation



Jaccard distance

$$d(X,Y) = H(Y) + H(X) - 2I(X,Y)$$

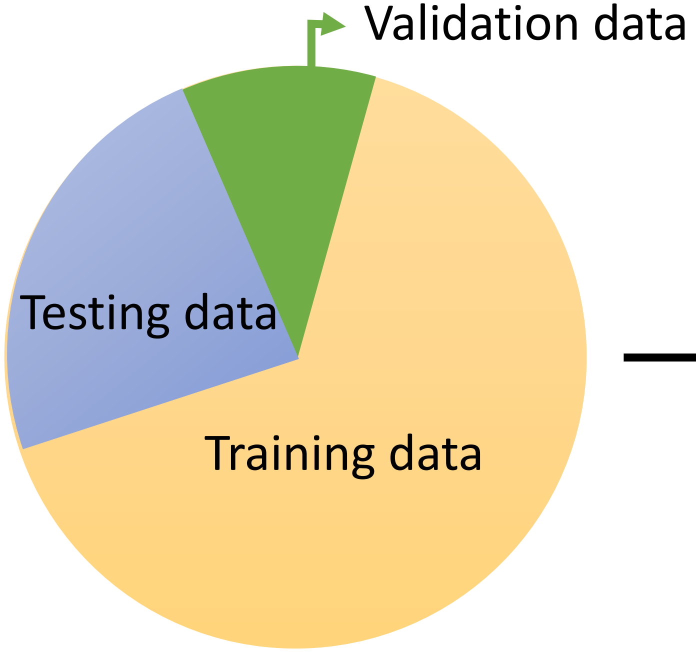
$$\text{Jaccard distance} = 1 - \frac{d(X,Y)}{H(X,Y)}$$

$$H(X,Y) = H(Y) + H(X) - I(X,Y)$$

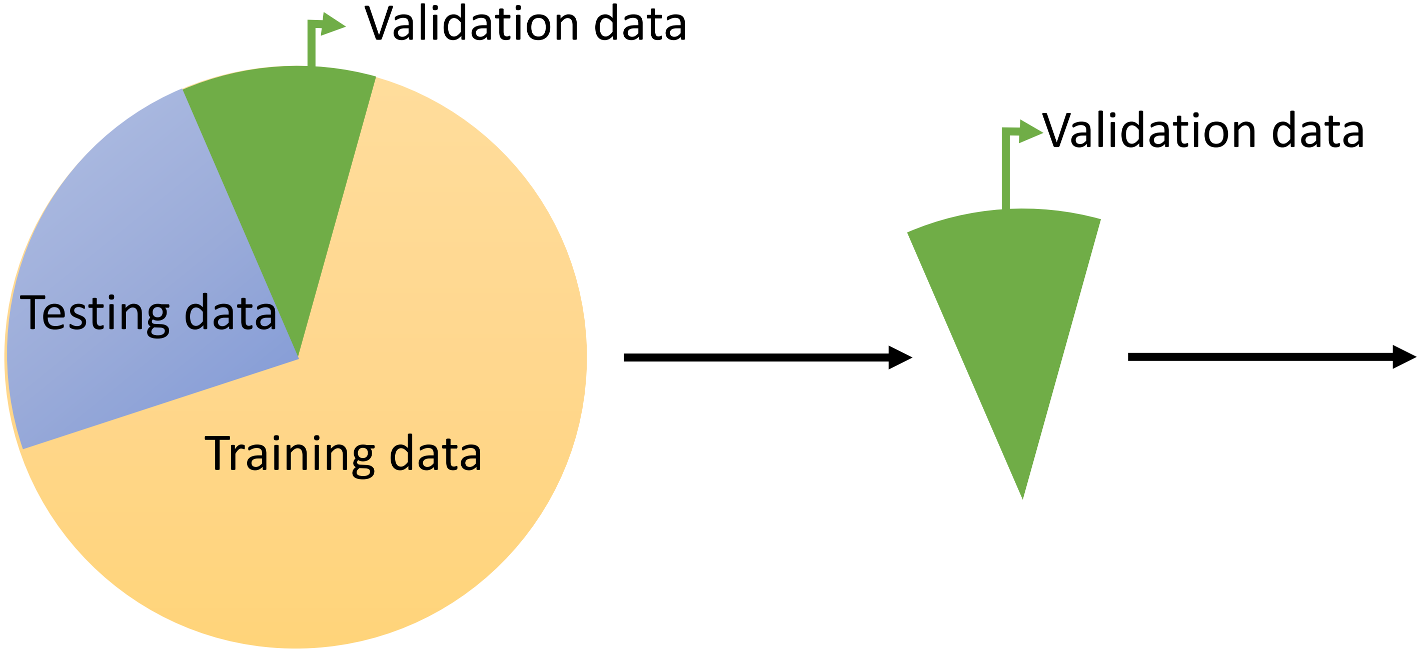
where, $X = f(x),\ Y = f(ae(x))$ *and* $H(Y) = -\sum_{y} \log p(y)\, p(y)$

classifier Auto-encoder

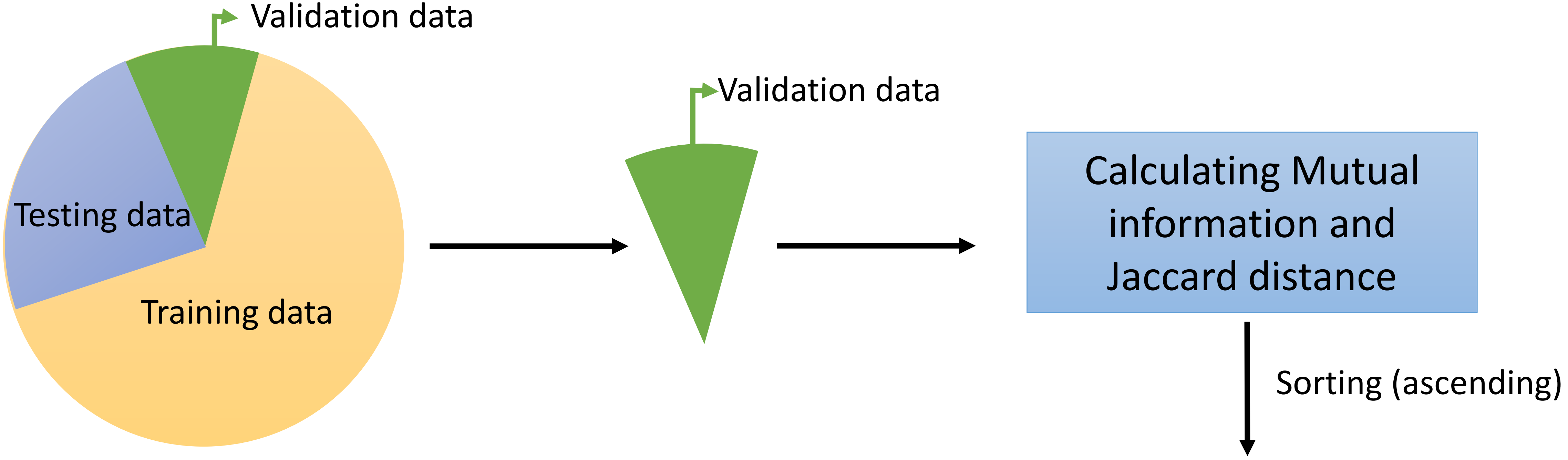
Getting threshold



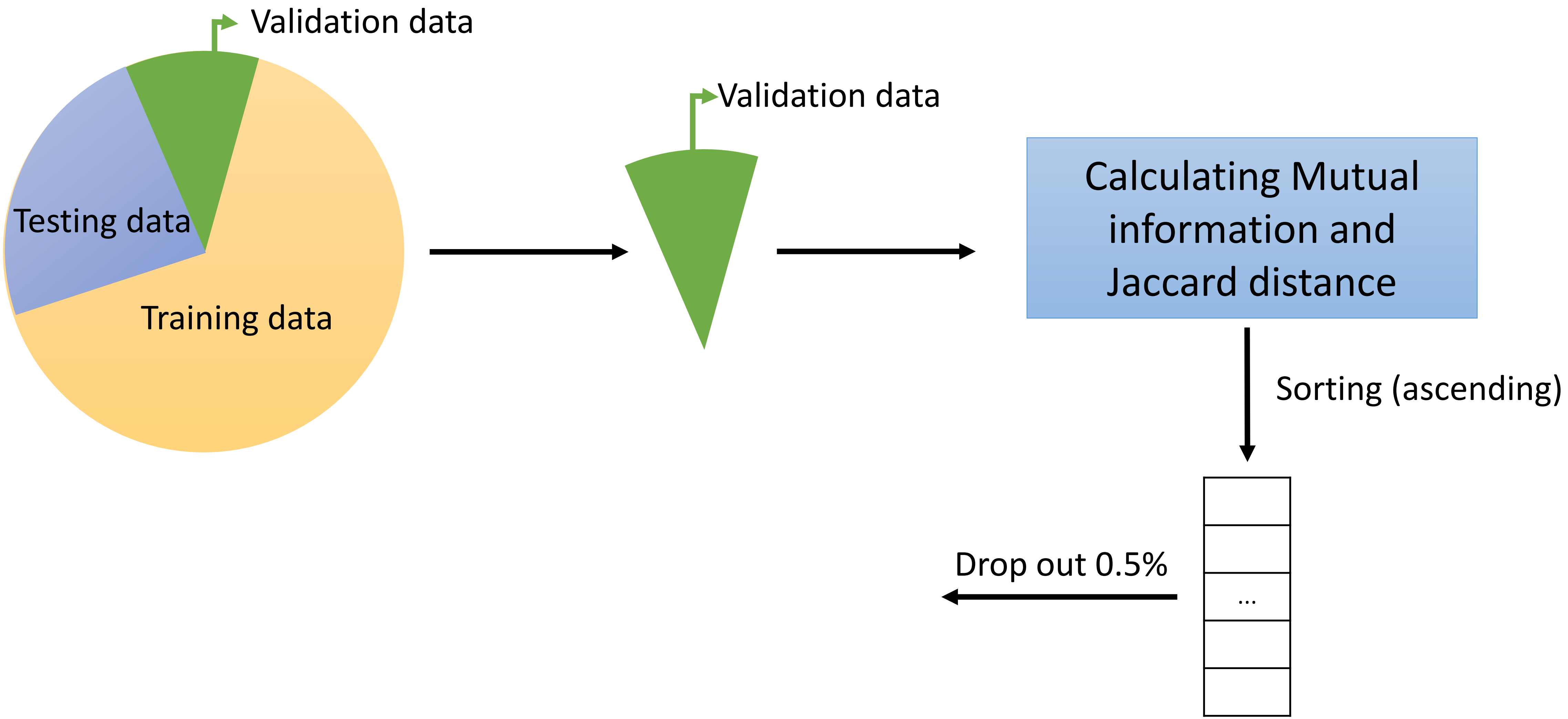
Getting threshold



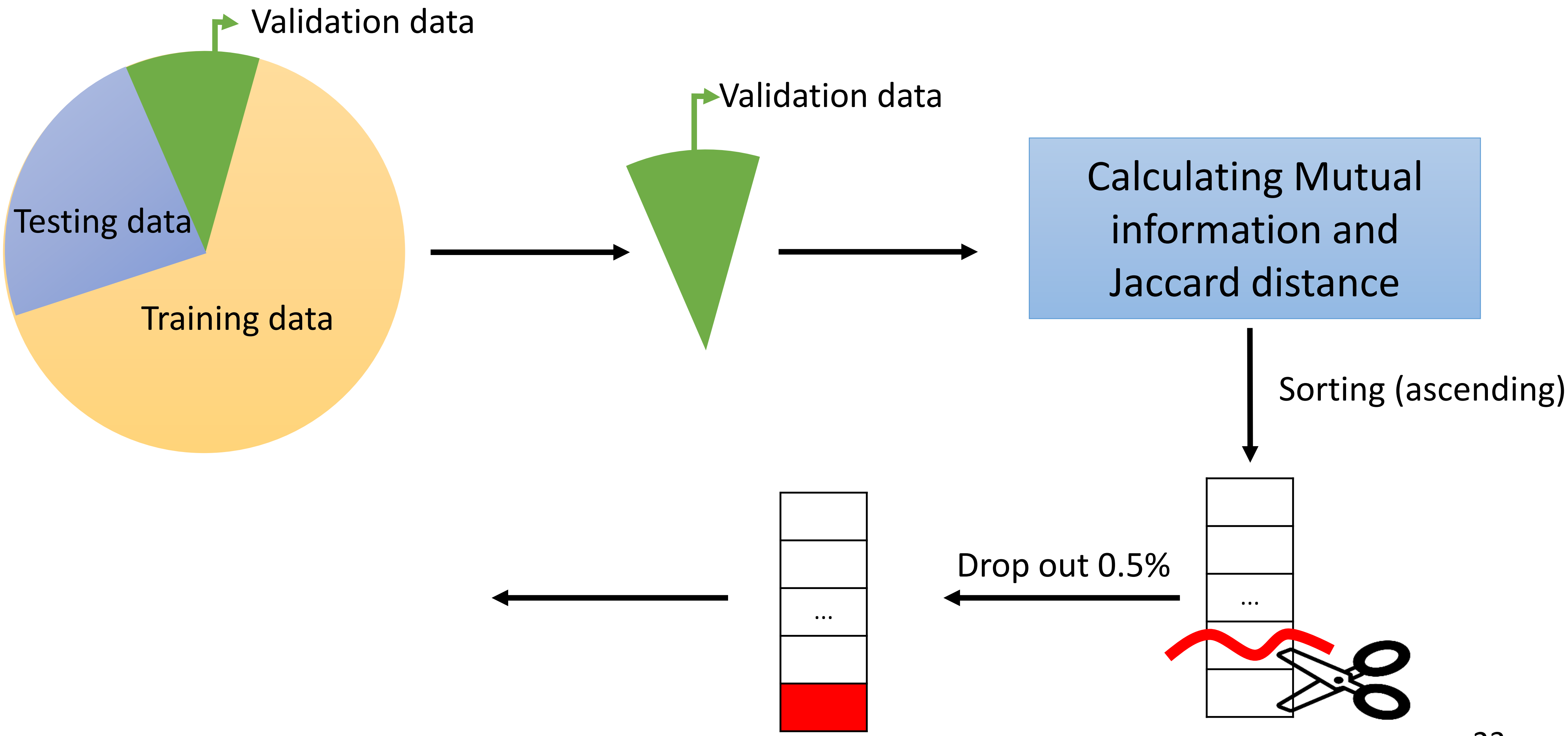
Getting threshold



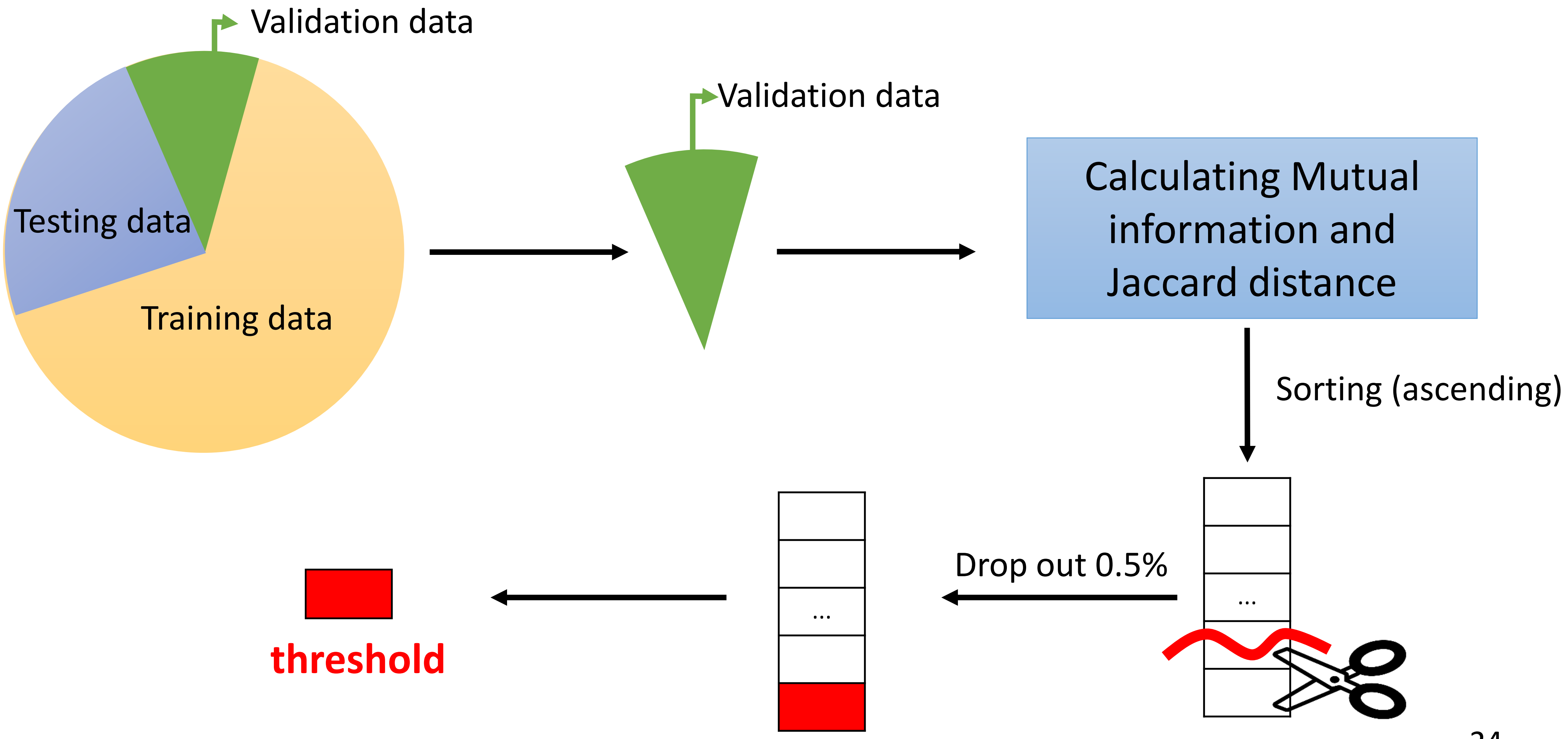
Getting threshold



Getting threshold



Getting threshold



Experimental results—MNIST

MNIST						
	MagNet			MID		
Attack method	C&W attack (L_2 version)	EAD attack (L_1 rule, $\beta = 10^{-1}$)	EAD attack (EN rule, $\beta = 10^{-1}$)	C&W attack (L_2 version)	EAD attack (L_1 rule, $\beta = 10^{-1}$)	EAD attack (EN rule, $\beta = 10^{-1}$)
κ						
0	98.7	78.8	78.1	98.7	78.8	78.1
5	94.6	33.5	26.6	95.8	39.4	37.4
10	91.5	17.9	11.7	97.8	46.9	44
15	90	16.2	9.7	98.0	47.4	41.8
20	91.4	19.6	12.1	98.2	45.1	36.8
25	93.9	26.1	16.8	98.4	44.3	35.6
30	96.2	34.5	22.5	98.5	44.3	32.9
35	97.7	41.1	28.6	99.0	47.3	35.4
40	98.5	47.8	33.1	98.9	52.0	37.9

Experimental results—CIFAR10

CIFAR-10						
	MagNet			MID		
Attack method	C&W attack (L_2 version)	EAD attack (L_1 rule, $\beta = 10^{-1}$)	EAD attack (EN rule, $\beta = 10^{-1}$)	C&W attack (L_2 version)	EAD attack (L_1 rule, $\beta = 10^{-1}$)	EAD attack (EN rule, $\beta = 10^{-1}$)
κ						
0	80.1	70.5	70.7	80.1	70.3	70.6
10	50.3	26.2	26.4	50.9	28.2	28.5
20	48.0	26.8	26.8	51.4	29.7	29.7
30	62.9	37.1	38.4	63.8	38.0	39.6
40	72.3	48.4	45.3	72.8	49.2	46.0
50	81.4	61.0	60.0	81.7	61.5	60.3
60	89.6	73.8	71.7	89.6	74.1	71.7
70	94.6	84.6	81.5	94.6	84.6	81.5
80	97.3	90.6	90.4	97.3	90.6	90.4

Conclusion

- Mutual information is a promising approach to characterize adversarial subspaces
- We will continue to improve the quality of image generated by auto-encoder to strengthen the effectiveness of mutual information detector