

Adversarial Machine Learning

Neural Networks Design And Application

Decompose generative adversarial networks

- Adversarial machine learning
- Generative models

Decompose generative **adversarial** networks

- **Adversarial** machine learning
- Generative models

Decompose **generative** adversarial networks

- Adversarial machine learning
- **Generative** models

Machine learning paradigm



Training data

ML model for bus
recognition



Testing data

Machine learning paradigm



Training data

ML model for
panda recognition

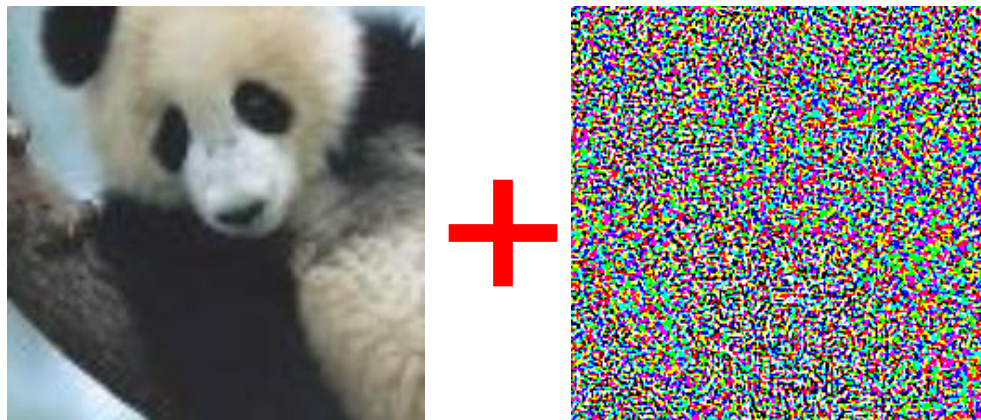


Testing data

Adversarial data



Adversarial data



Adversarial data



Adversarial data



Adversarial data: used to fool the trained model

Adversarial data



Similar to the original one
from human's eyes

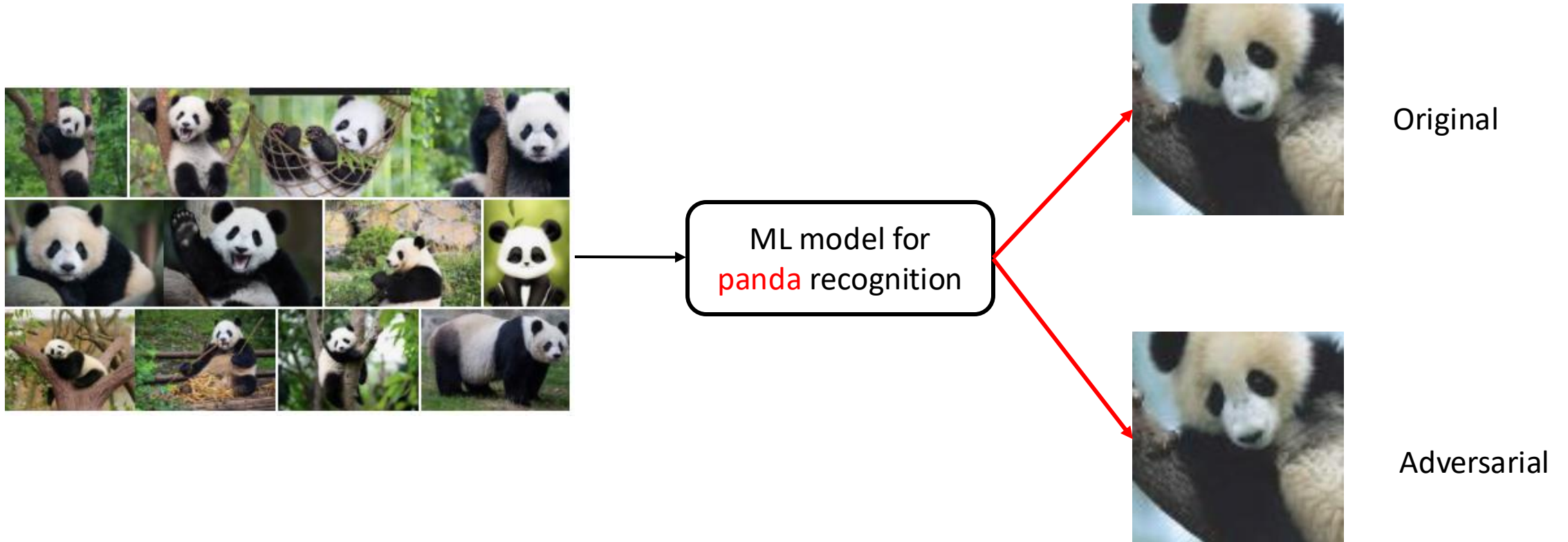
Adversarial data: used to fool the trained model

Machine learning model VS adversarial data



ML model for
panda recognition

Machine learning model VS adversarial data



Machine learning model VS adversarial data

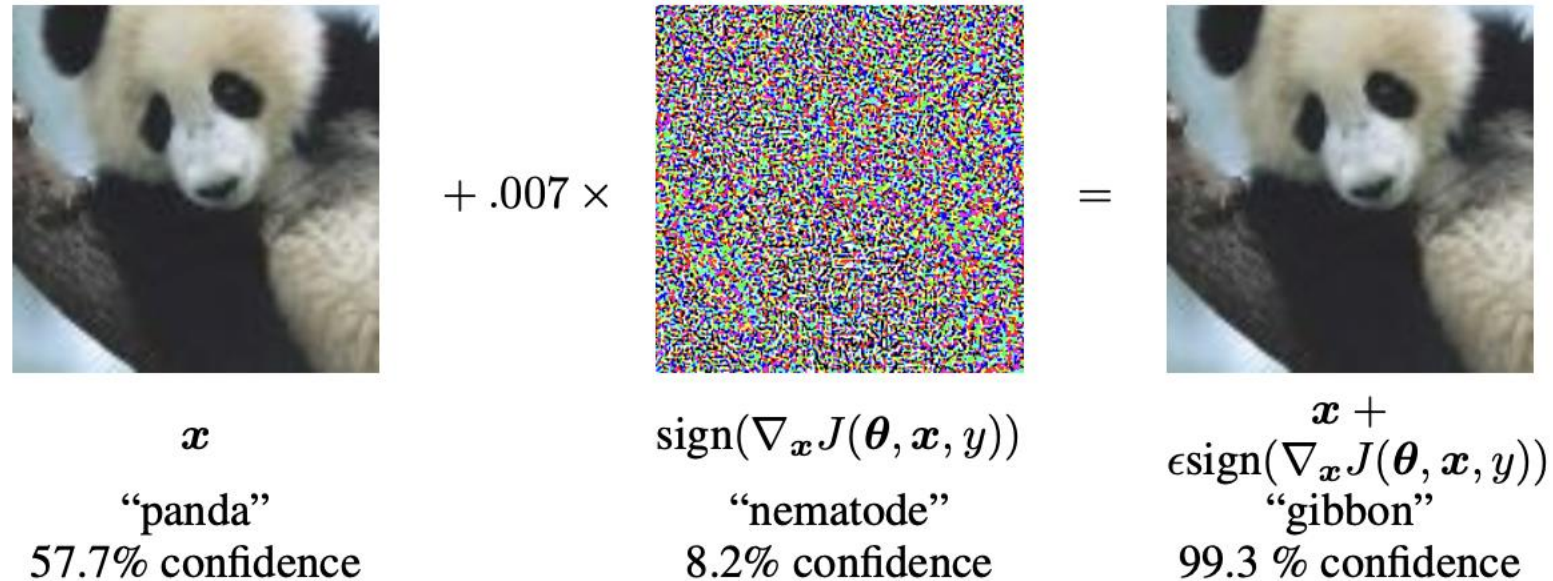


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

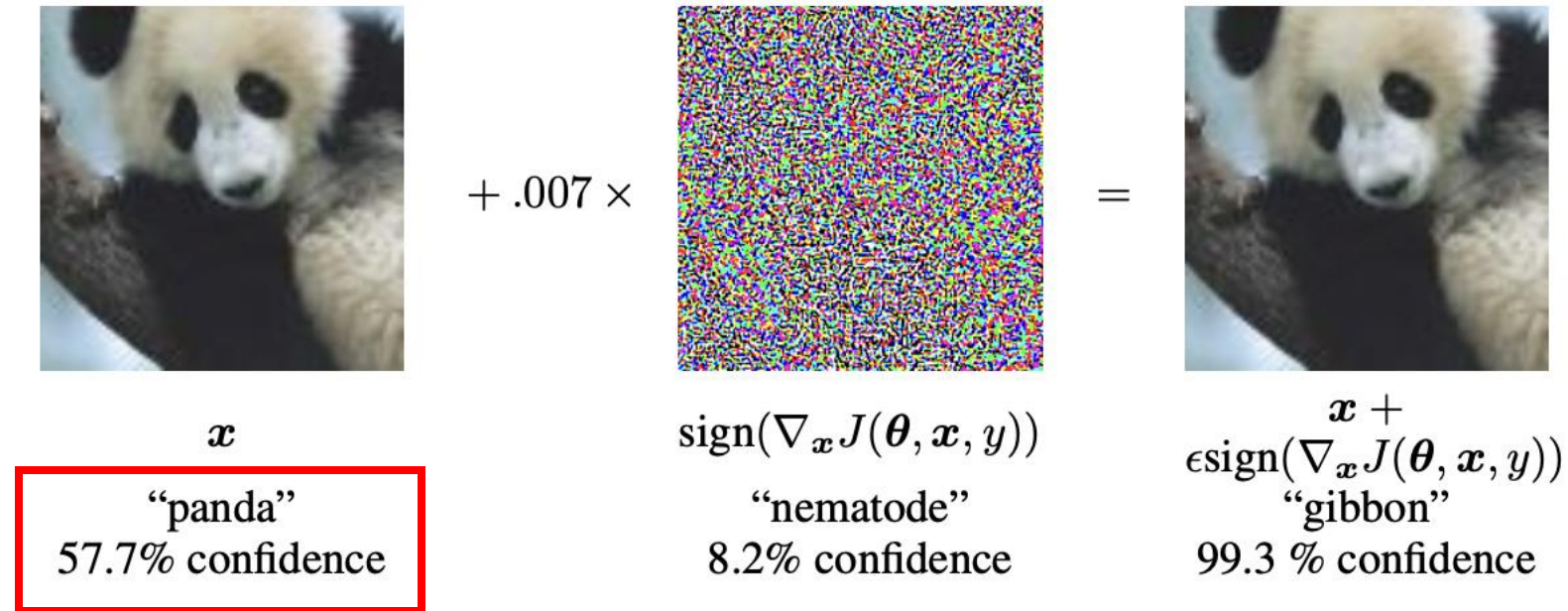


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

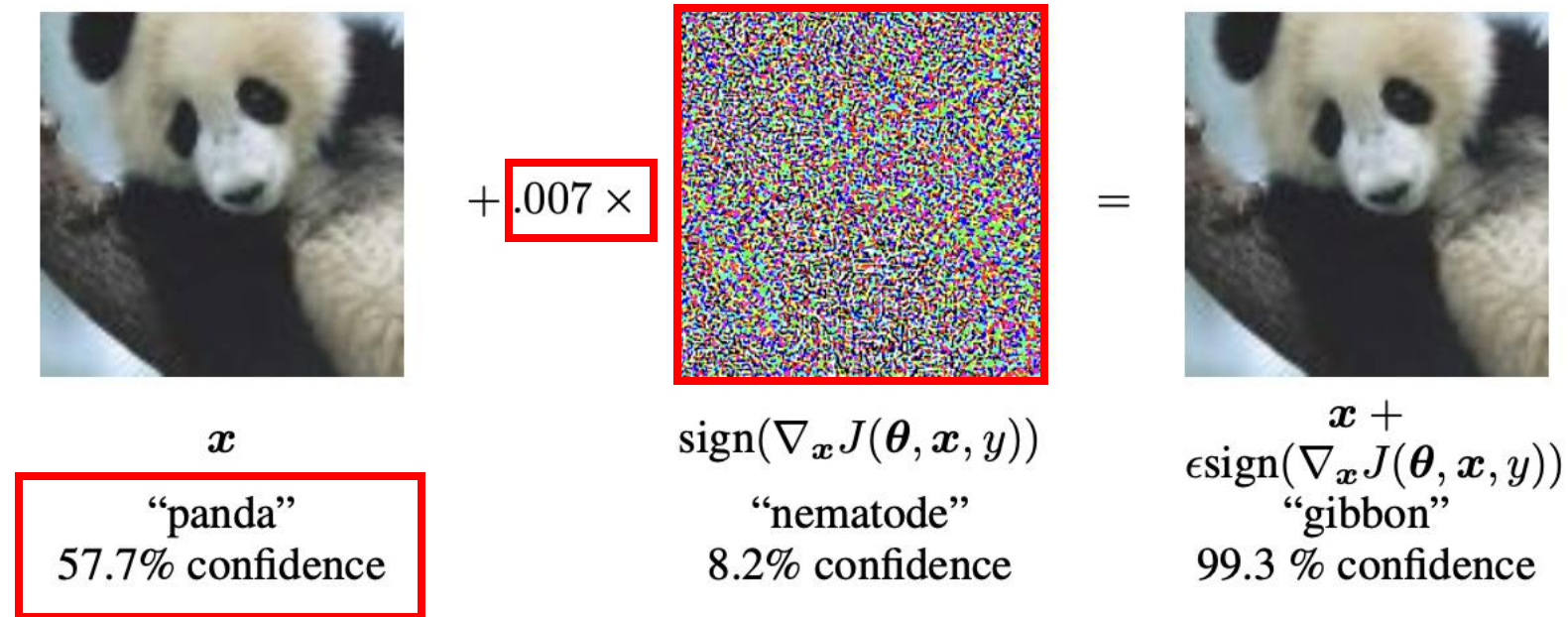


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

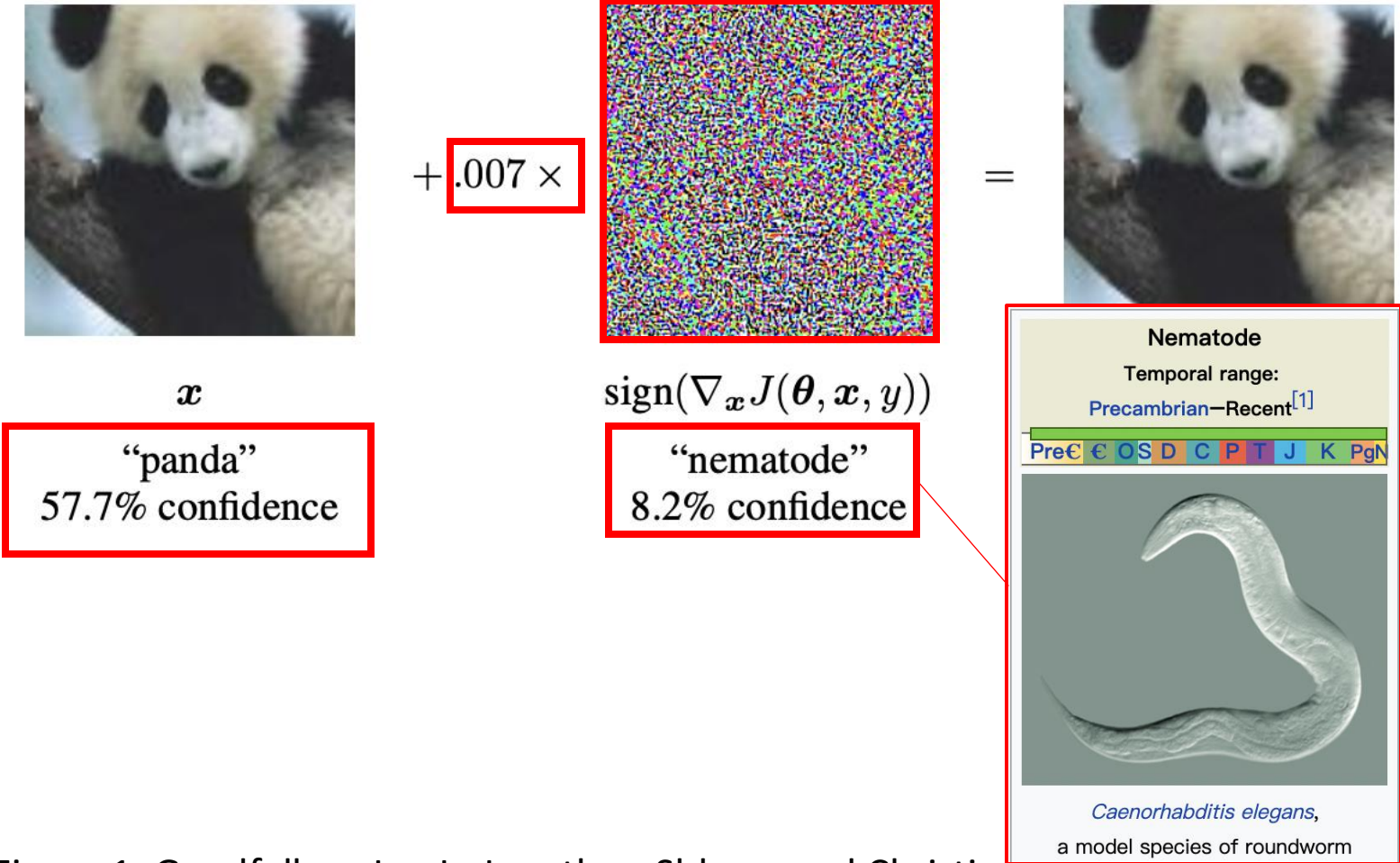


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

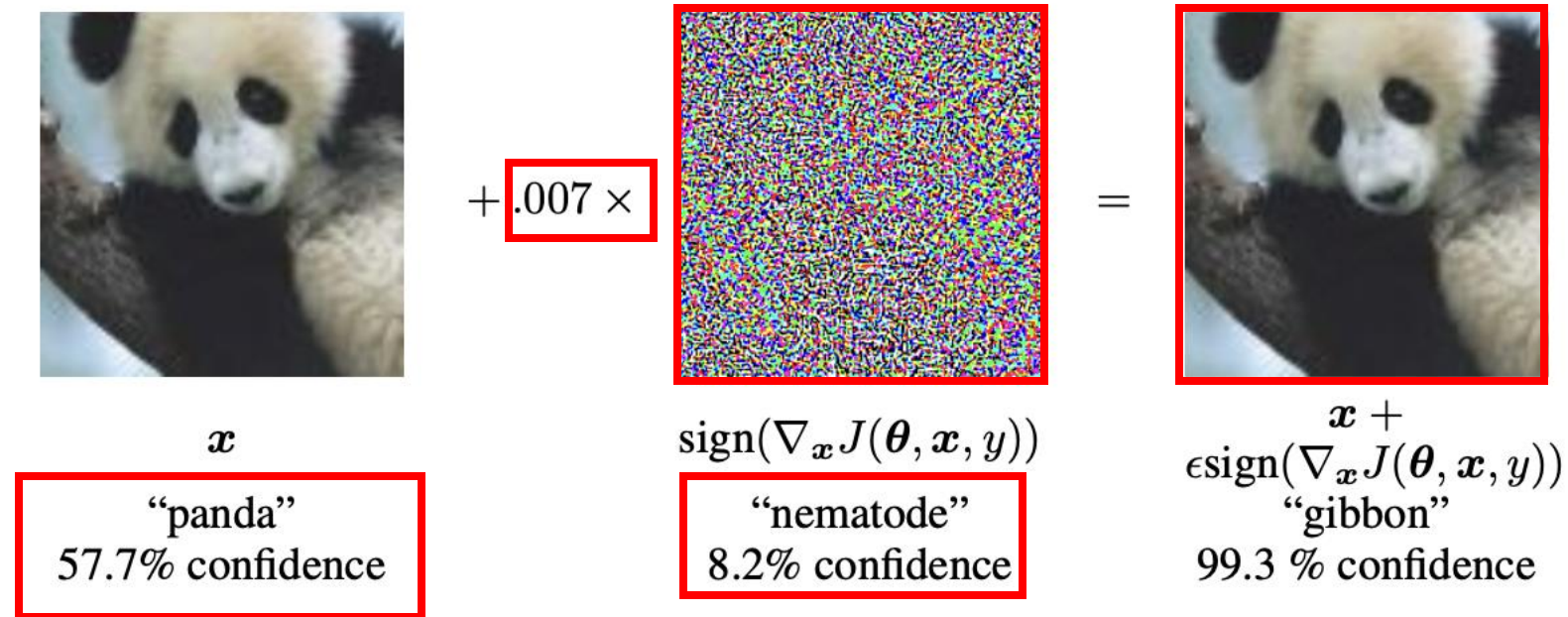


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

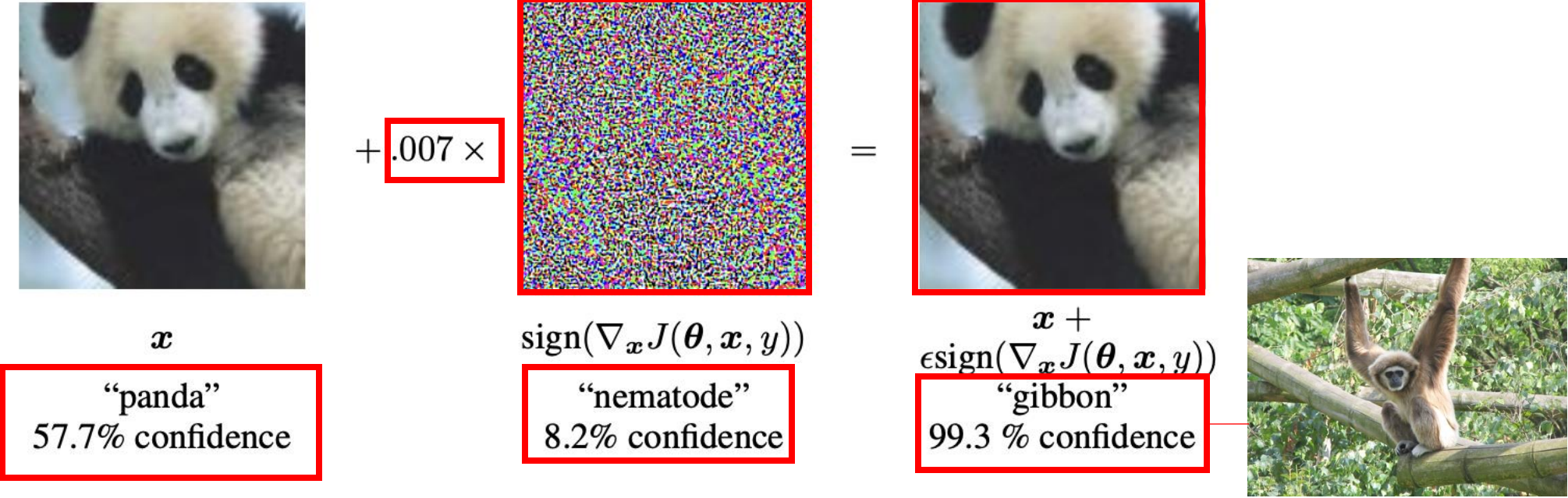
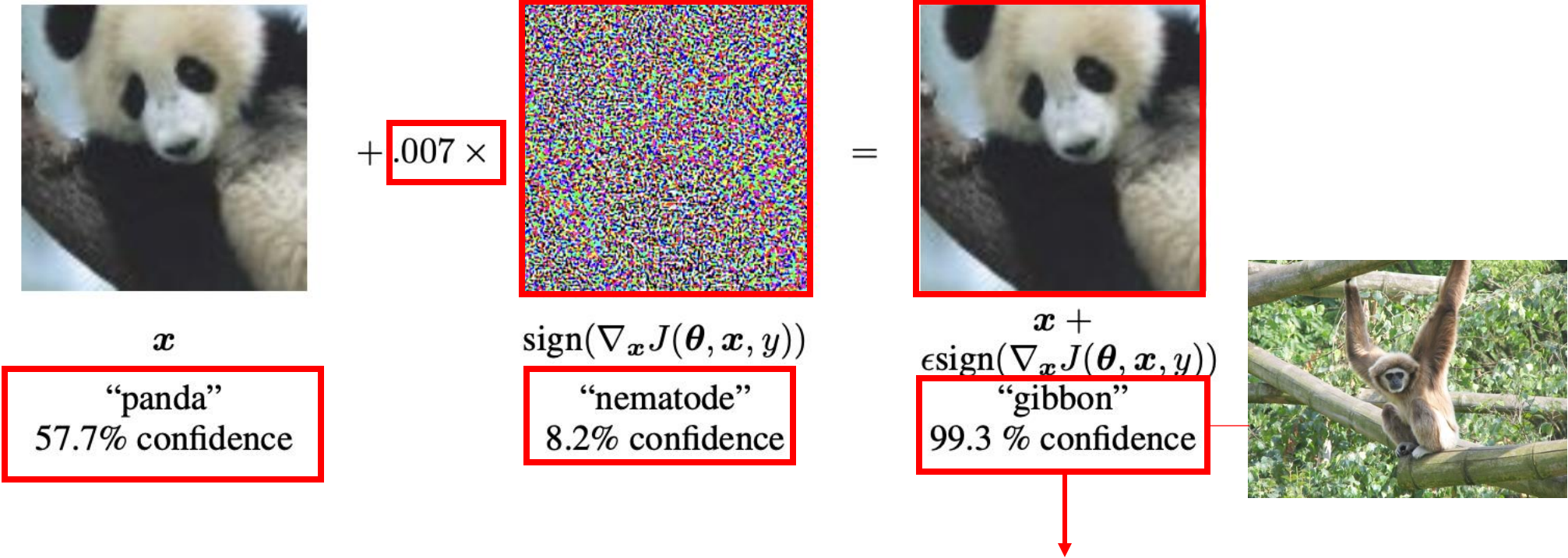


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

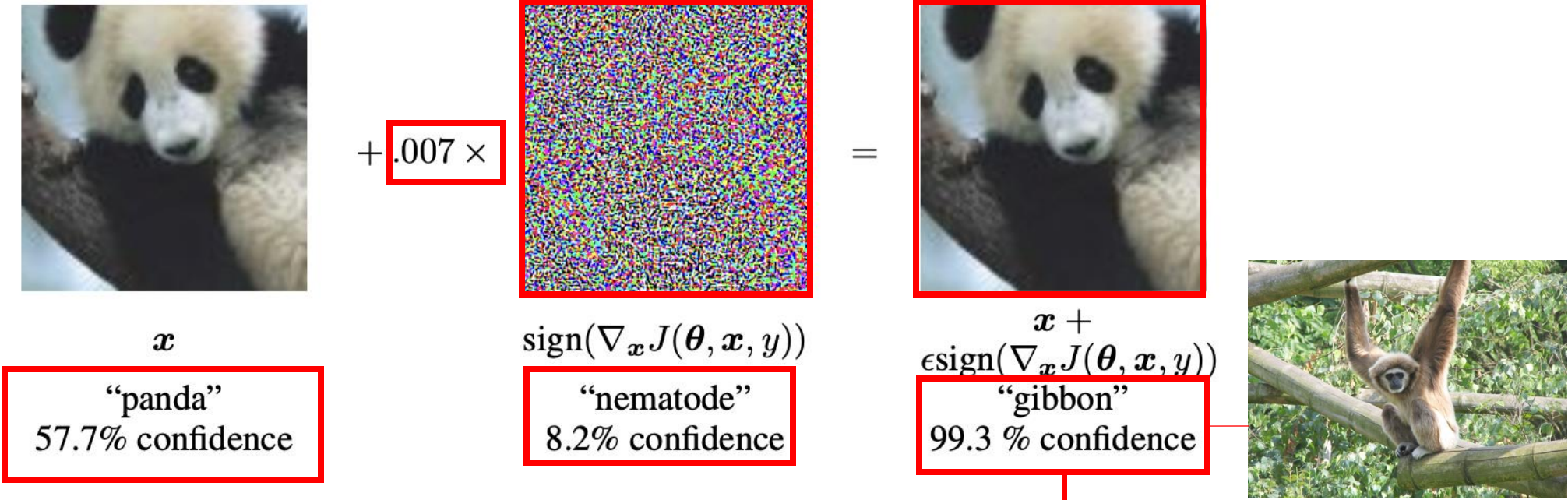
Machine learning model VS adversarial data



Use adversarial data to attack models

Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data



Use adversarial data to attack models
Deep learning models are particularly vulnerable

Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Why care about attacks



Why care about attacks



A real **Stop** sign

Why care about attacks



A real **Stop** sign



A physical perturbation
applied to a **Stop** sign

Why care about attacks



A real **Stop** sign



A physical perturbation applied to a **Stop** sign

Artificial patches

Why care about attacks



A real **Stop** sign



A physical perturbation applied to a **Stop** sign

Prediction result?

Artificial patches

Why care about attacks



A real **Stop** sign



A physical perturbation applied to a **Stop** sign

Prediction result?

Artificial patches

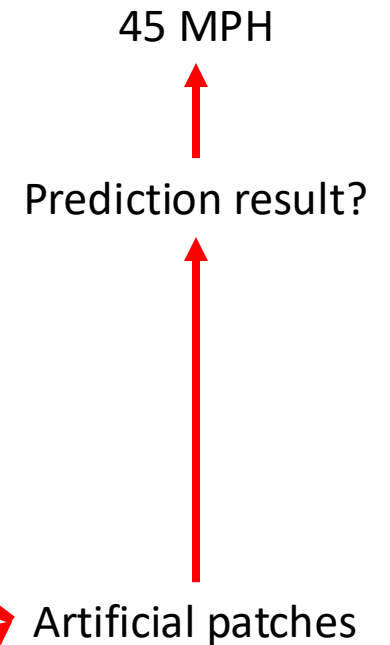
Why care about attacks



A real **Stop** sign



A physical perturbation applied to a **Stop** sign



Why care about attacks

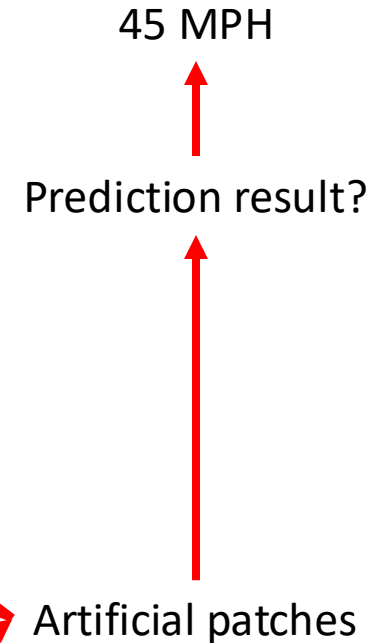
What if a driver recognize **STOP** as **45 MPH**?



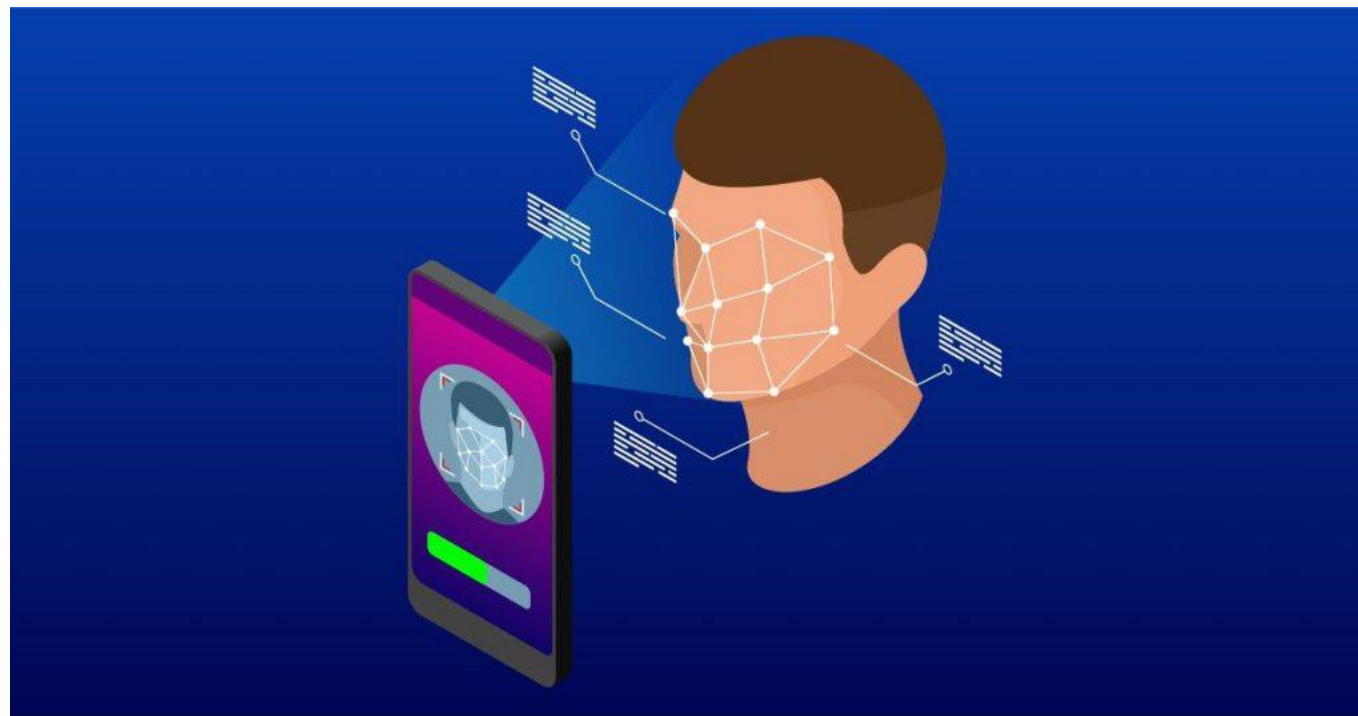
A real **Stop** sign



A physical perturbation applied to a **Stop** sign

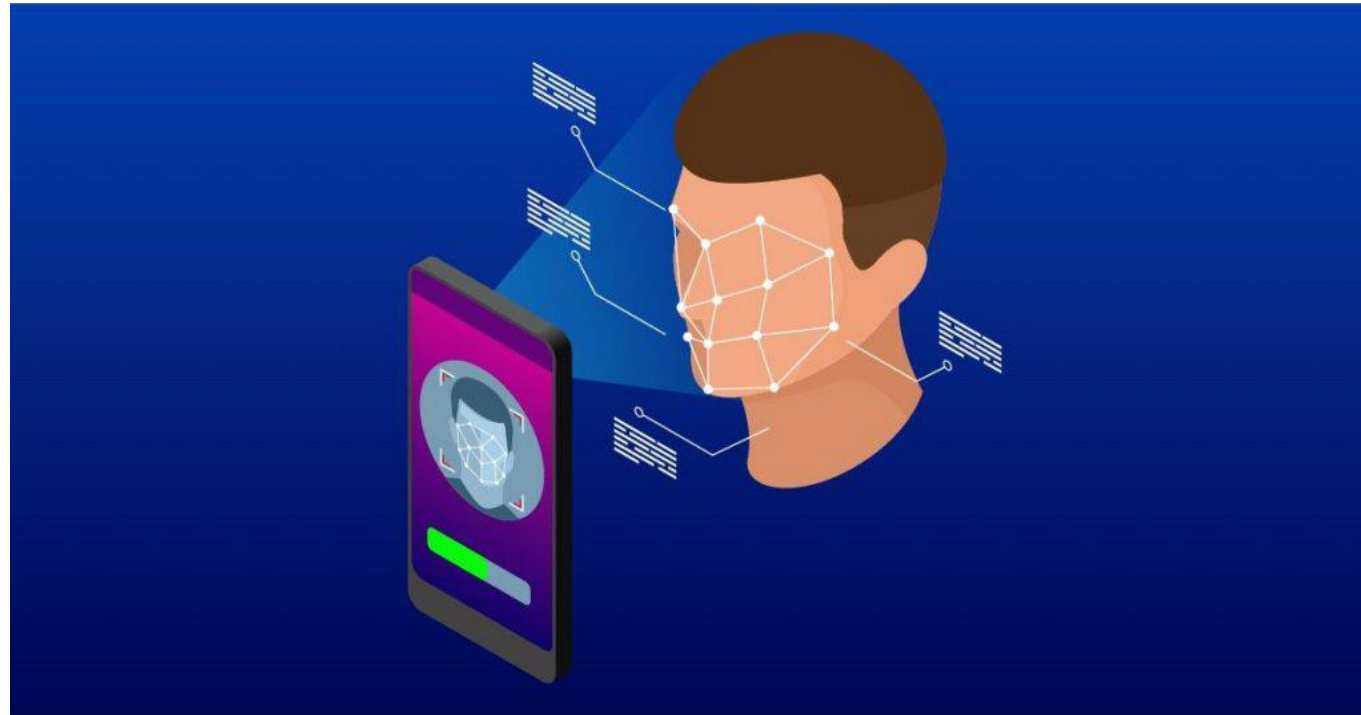


Why care about attacks/adversarial noise?



Q: can we use a simple photo to unlock face recognition system?

Why care about attacks/adversarial noise?



Q: can we use a simple photo to unlock face recognition system?

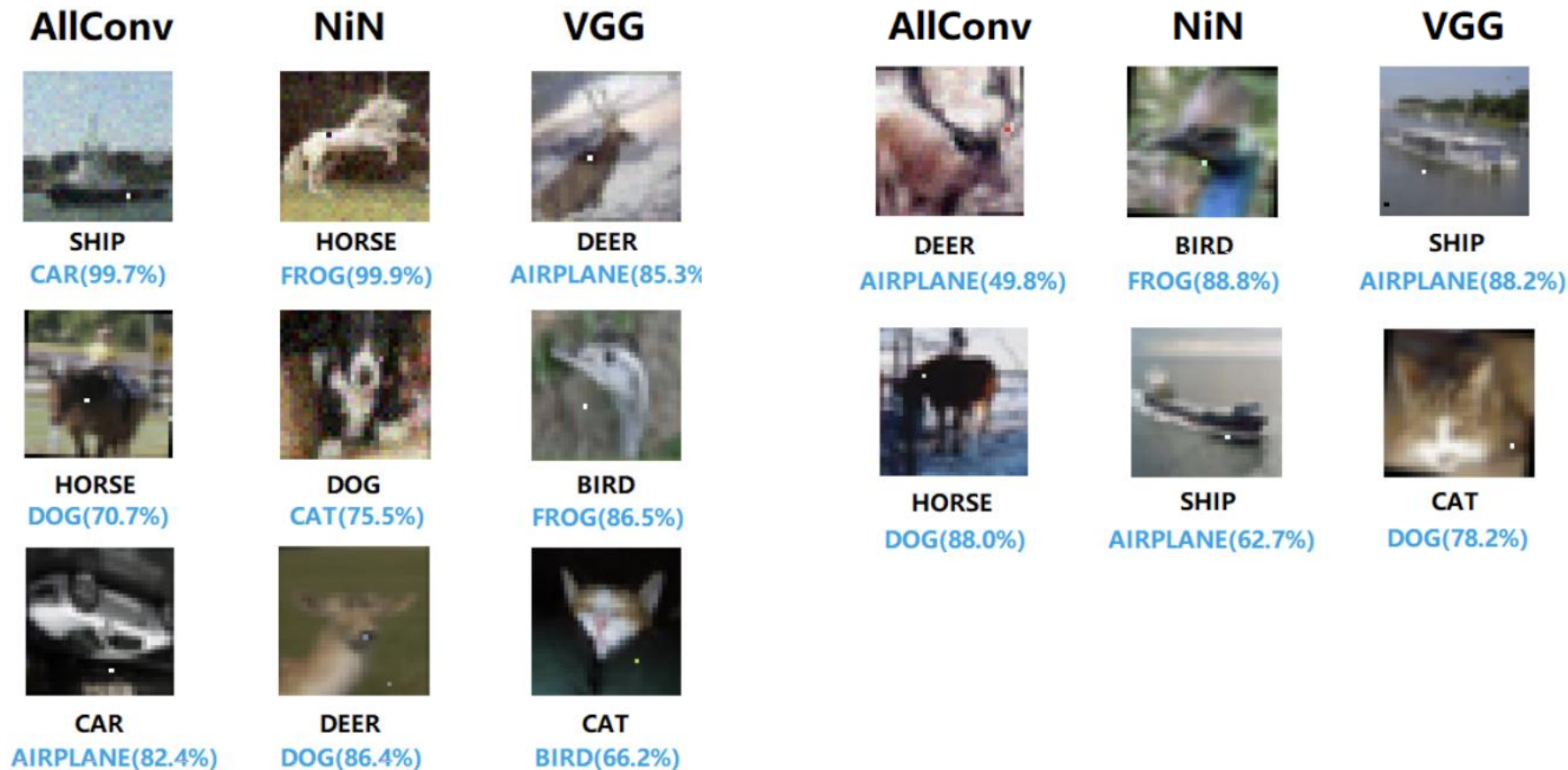
↓
(your smart phone)

Why care about attacks/adversarial noise?

Q: Can we fool deep models with only one pixel modified?











Why care about attacks/adversarial noise?

Q: Can we fool deep models with only **one** pixel modified?



Why care about attacks/adversarial noise?














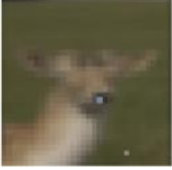
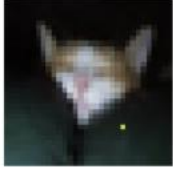
Q: Can we fool deep models with only **one** pixel modified?

AllConv	NiN	VGG	AllConv	NiN	VGG
 SHIP CAR(99.7%)	 HORSE FROG(99.9%)	 DEER AIRPLANE(85.3%)	 DEER AIRPLANE(49.8%)	 BIRD FROG(88.8%)	 SHIP AIRPLANE(88.2%)
 HORSE DOG(70.7%)	 DOG CAT(75.5%)	 BIRD FROG(86.5%)	 HORSE DOG(88.0%)	 SHIP AIRPLANE(62.7%)	 CAT DOG(78.2%)
 CAR AIRPLANE(82.4%)	 DEER DOG(86.4%)	 CAT BIRD(66.2%)			

Why care about attacks/adversarial noise?

Q: Can we fool deep models with only **one** pixel modified?

True label → Predicted label

AllConv	NiN	VGG	AllConv	NiN	VGG
 SHIP CAR(99.7%)	 HORSE FROG(99.9%)	 DEER AIRPLANE(85.3%)	 DEER AIRPLANE(49.8%)	 BIRD FROG(88.8%)	 SHIP AIRPLANE(88.2%)
 HORSE DOG(70.7%)	 DOG CAT(75.5%)	 BIRD FROG(86.5%)	 HORSE DOG(88.0%)	 SHIP AIRPLANE(62.7%)	 CAT DOG(78.2%)
 CAR AIRPLANE(82.4%)	 DEER DOG(86.4%)	 CAT BIRD(66.2%)			

Why care about attacks/adversarial noise?

Q: Can we fool deep models with only **one** pixel modified?

True label →
Predicted label →

Model	True Label	Predicted Label (Confidence)
AllConv	SHIP	CAR(99.7%)
NiN	HORSE	FROG(99.9%)
VGG	DEER	AIRPLANE(85.3%)
AllConv	HORSE	DOG(70.7%)
NiN	DOG	CAT(75.5%)
VGG	BIRD	FROG(86.5%)
AllConv	CAR	AIRPLANE(82.4%)
NiN	DEER	DOG(86.4%)
VGG	CAT	BIRD(66.2%)

Model	True Label	Predicted Label (Confidence)
AllConv	DEER	AIRPLANE(49.8%)
NiN	BIRD	FROG(88.8%)
VGG	SHIP	AIRPLANE(88.2%)
AllConv	HORSE	DOG(88.0%)
NiN	SHIP	AIRPLANE(62.7%)
VGG	CAT	DOG(78.2%)

Why care about attacks/adversarial noise?

Q: Can we fool deep models with only **one** pixel modified?

True label →
Predicted label →

Model	True label	Predicted label
AllConv	SHIP	CAR(99.7%)
NiN	HORSE	FROG(99.9%)
VGG	DEER	AIRPLANE(85.3%)
AllConv	DEER	AIRPLANE(49.8%)
NiN	BIRD	FROG(88.8%)
VGG	SHIP	AIRPLANE(88.2%)
AllConv	HORSE	DOG(88.0%)
NiN	SHIP	AIRPLANE(62.7%)
VGG	CAT	DOG(78.2%)
AllConv	HORSE	DOG(70.7%)
NiN	DOG	CAT(75.5%)
VGG	BIRD	FROG(86.5%)
AllConv	CAR	AIRPLANE(82.4%)
NiN	DEER	DOG(86.4%)
VGG	CAT	BIRD(66.2%)

Why care about attacks/adversarial noise?

Q: Can we fool deep models with only **one** pixel modified?

True label →
Predicted label →

Model	True Label	Predicted Label (Confidence)
AllConv	SHIP	CAR(99.7%)
NiN	HORSE	FROG(99.9%)
VGG	DEER	AIRPLANE(85.3%)
AllConv	DEER	AIRPLANE(49.8%)
NiN	BIRD	FROG(88.8%)
VGG	SHIP	AIRPLANE(88.2%)
AllConv	HORSE	DOG(70.7%)
NiN	DOG	CAT(75.5%)
VGG	BIRD	FROG(86.5%)
AllConv	HORSE	DOG(88.0%)
NiN	SHIP	AIRPLANE(62.7%)
VGG	CAT	DOG(78.2%)
AllConv	CAR	AIRPLANE(82.4%)
NiN	DEER	DOG(86.4%)
VGG	CAT	BIRD(66.2%)

All with high confidence (>49.8%)

Why care about attacks/adversarial noise?

Q: Can we fool deep models with only one pixel modified?

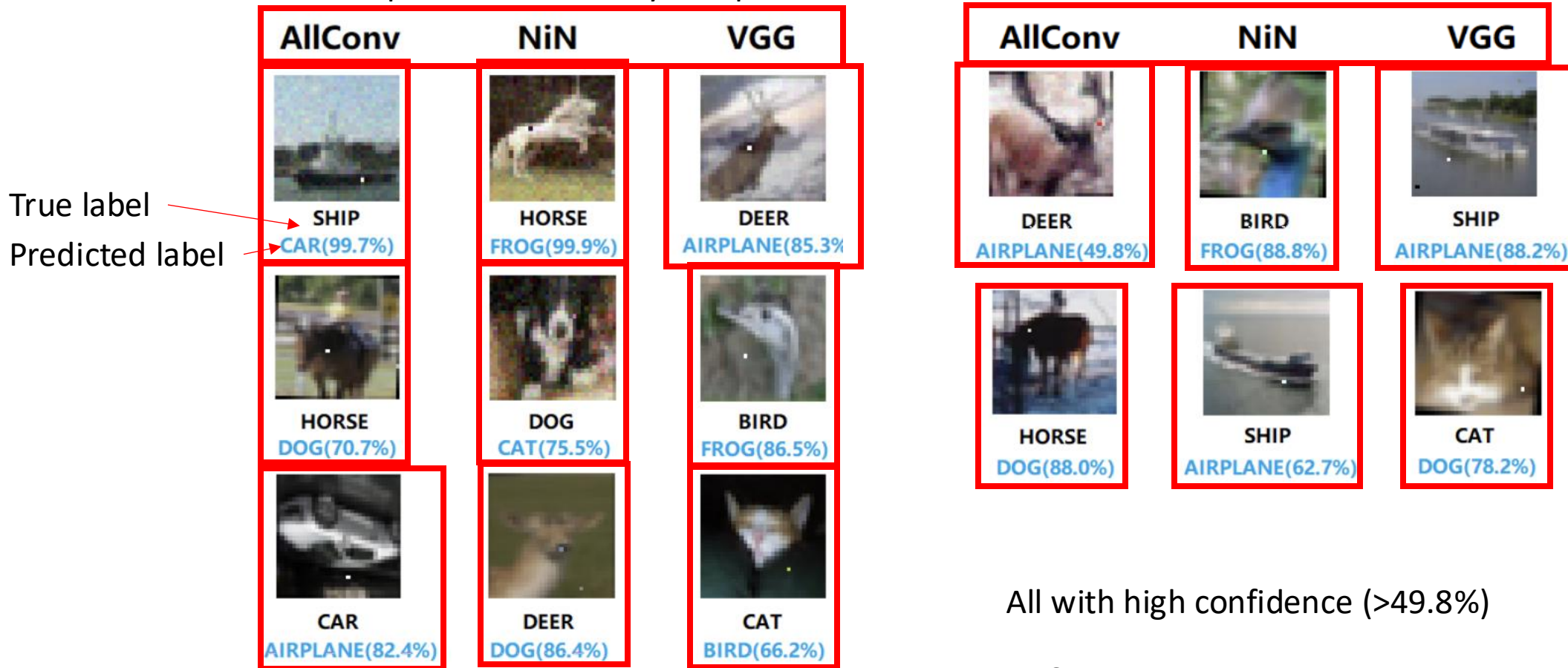


Figure 1, Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation* 23, no. 5 (2019): 828-841.

Robustness of machine learning models



ML model for
panda recognition



Original image:



Adversarial image:

Robustness of machine learning models



ML model for
panda recognition

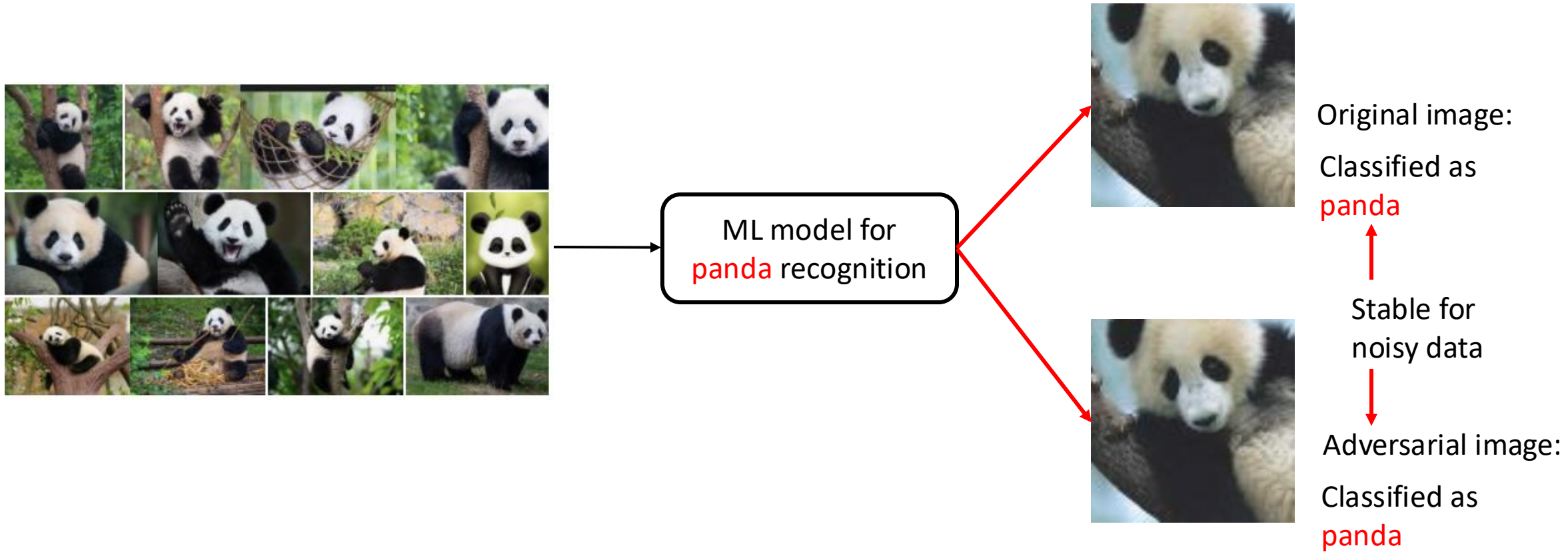


Original image:
Classified as
panda



Adversarial image:
Classified as
panda

Robustness of machine learning models



Learning with adversarial data

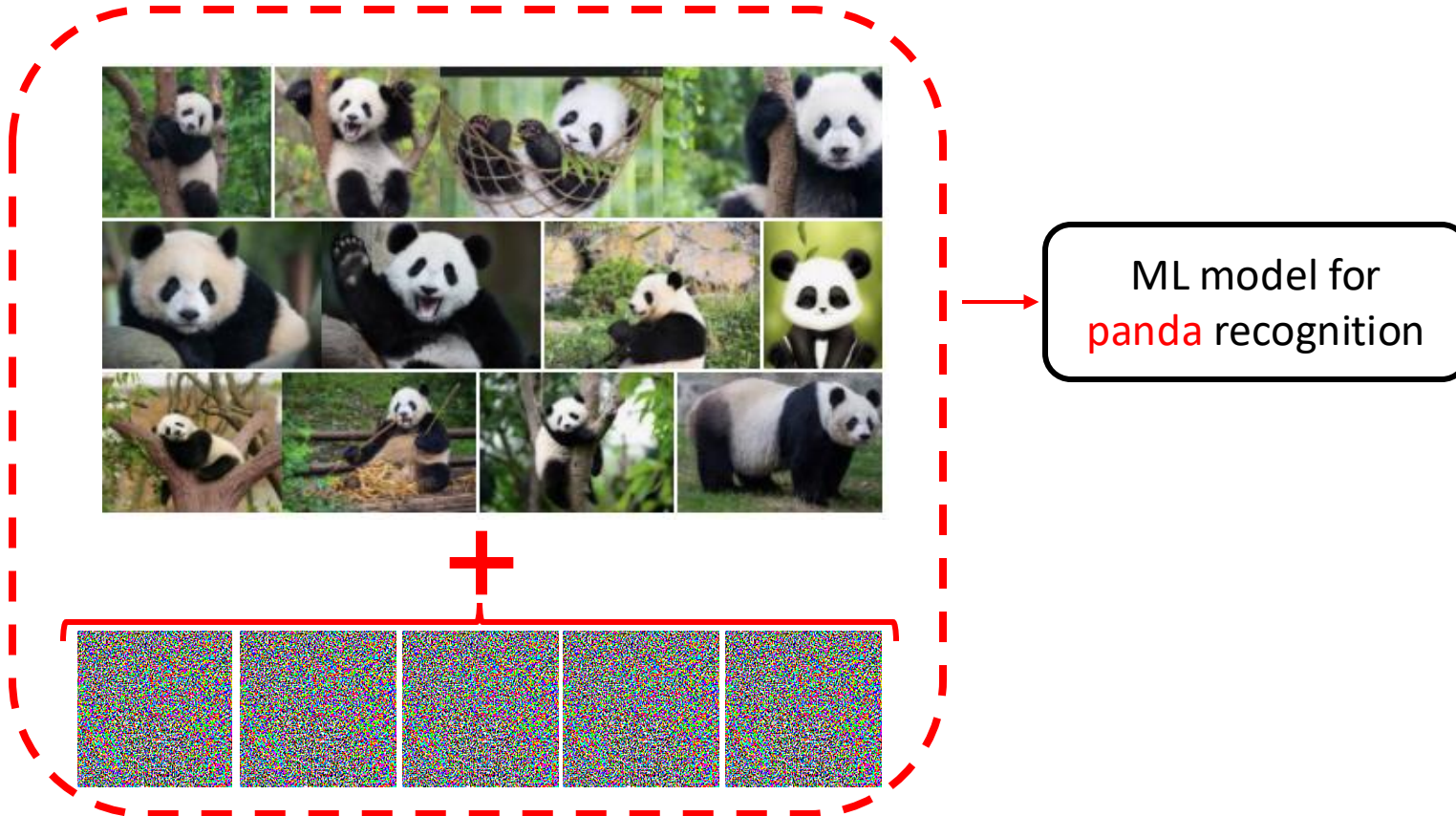
Q: can we build a new training set that includes adversarial data?



→ ML model for
panda recognition

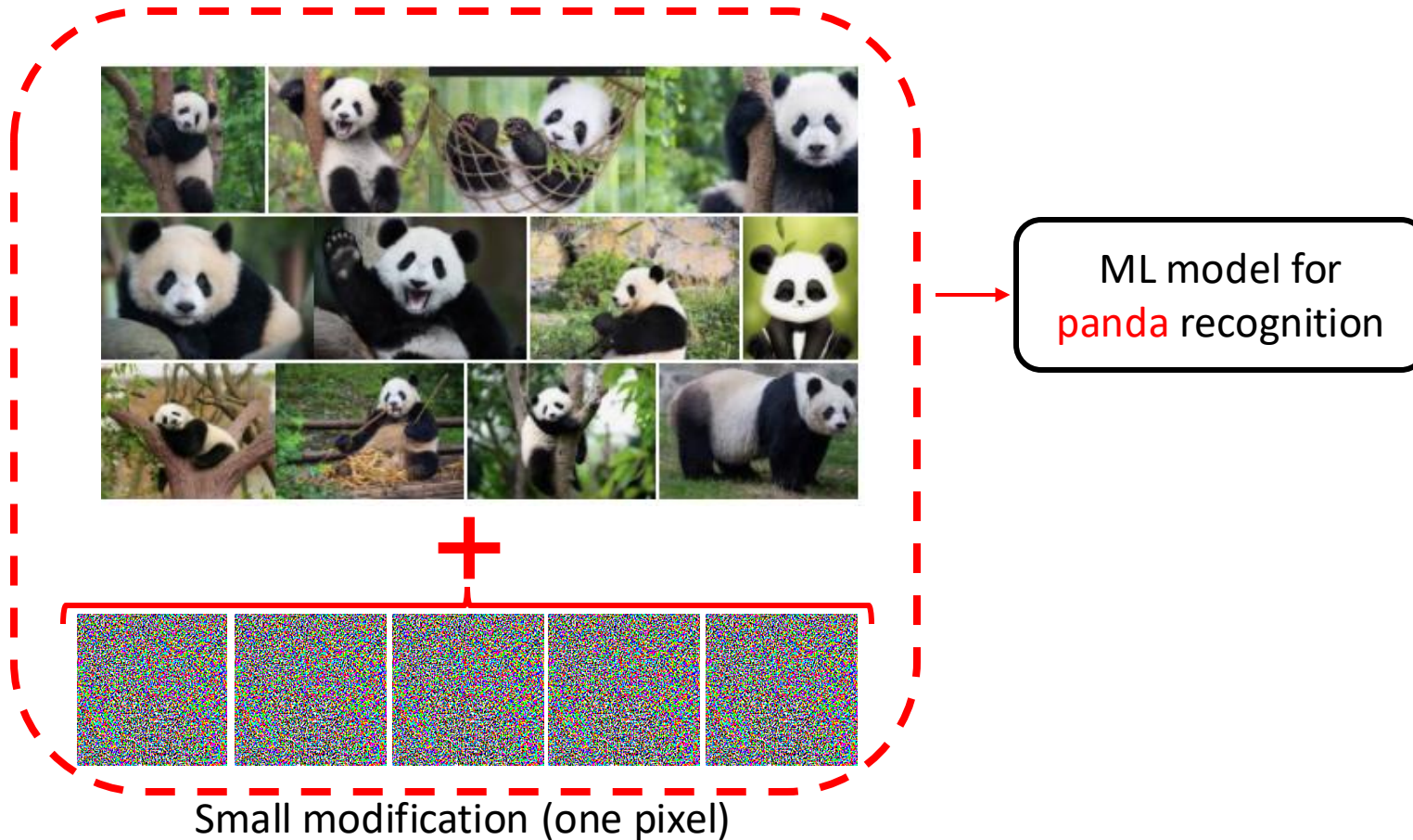
Learning with adversarial data

Q: can we build a new training set that includes adversarial data?



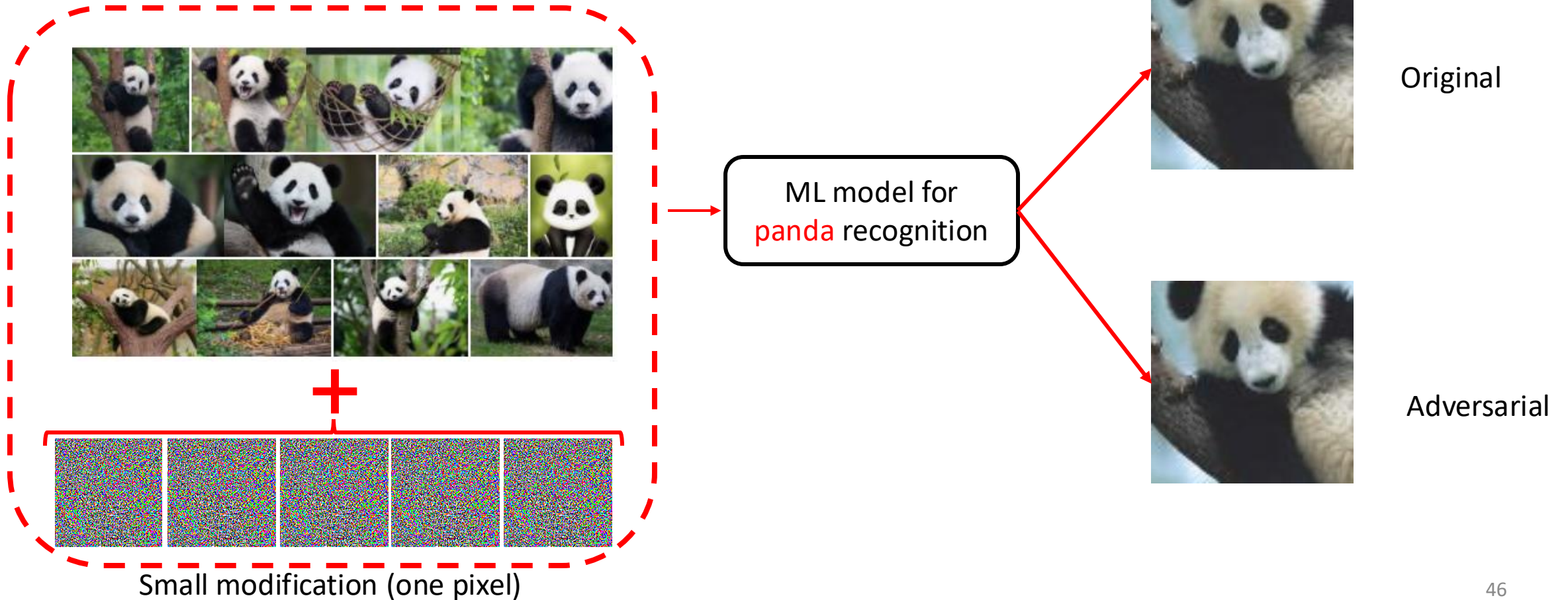
Learning with adversarial data

Q: can we build a new training set that includes adversarial data?



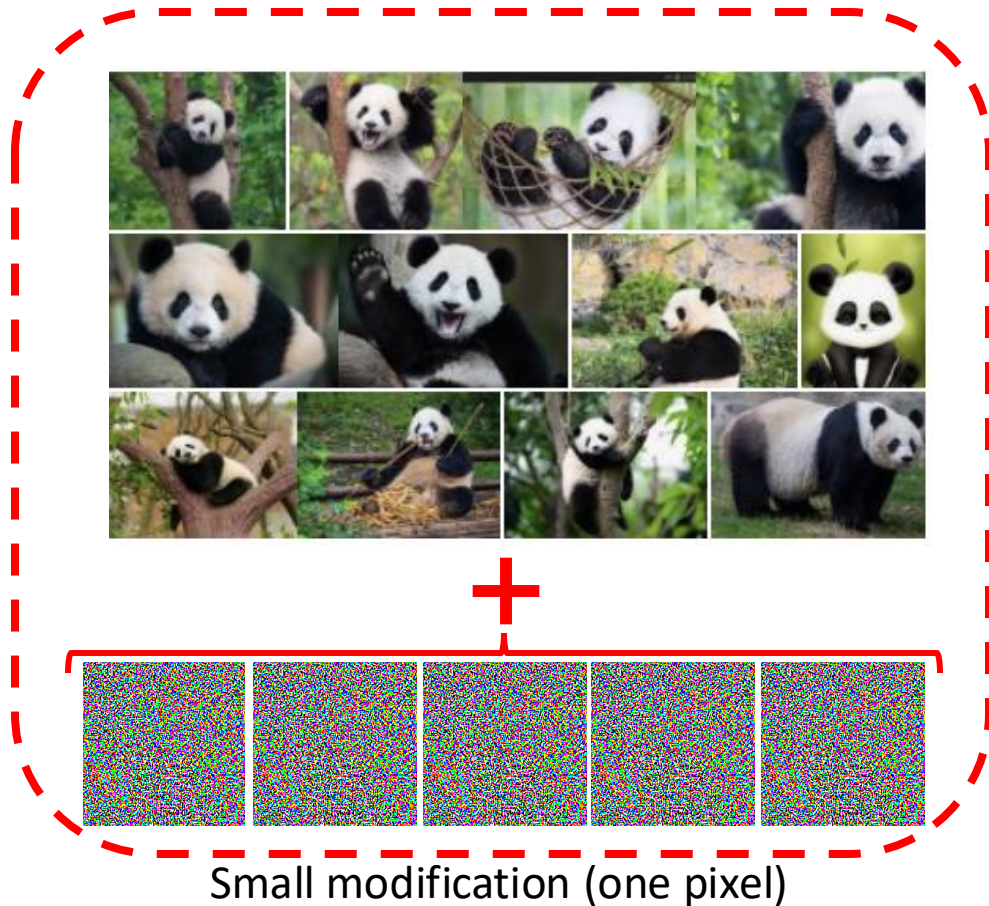
Learning with adversarial data

Q: can we build a new training set that includes adversarial data?

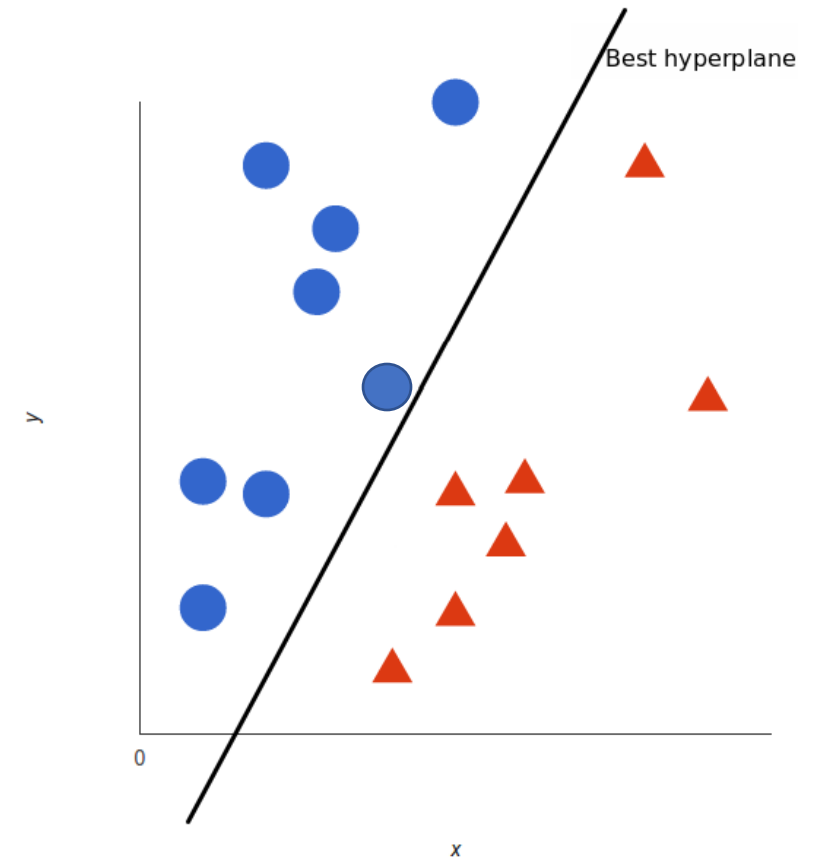


Learning with adversarial data

Q: can we build a new training set that includes adversarial data?

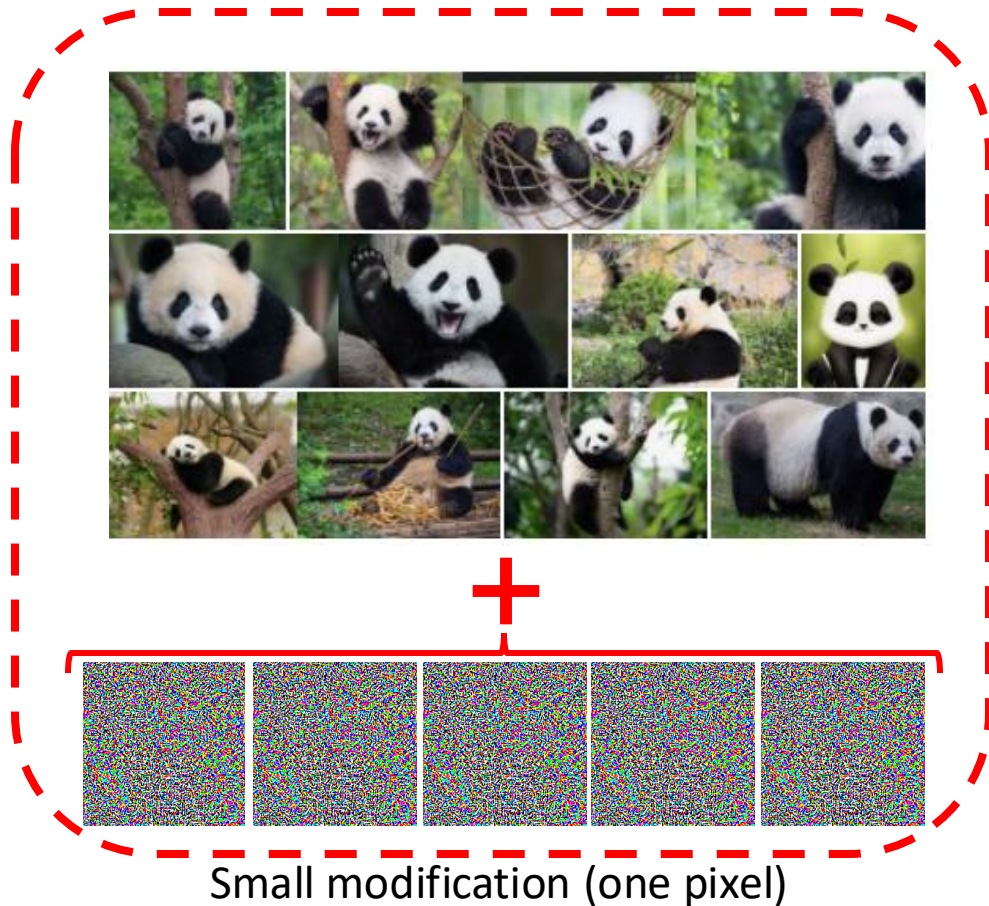


ML model for
panda recognition

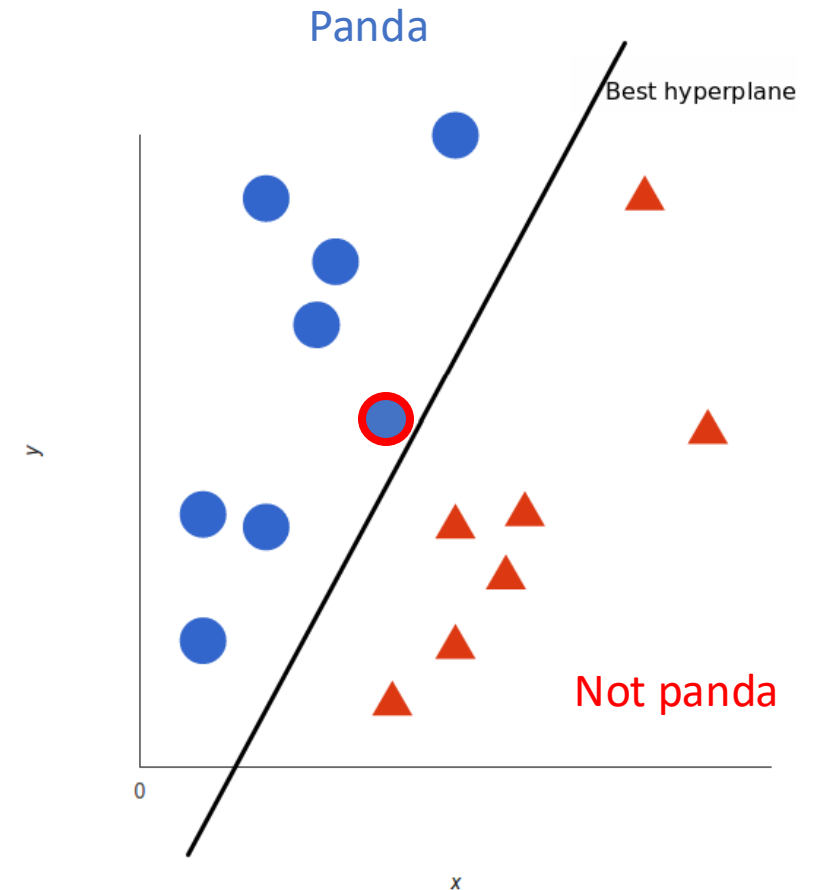


Learning with adversarial data

Q: can we build a new training set that includes adversarial data?

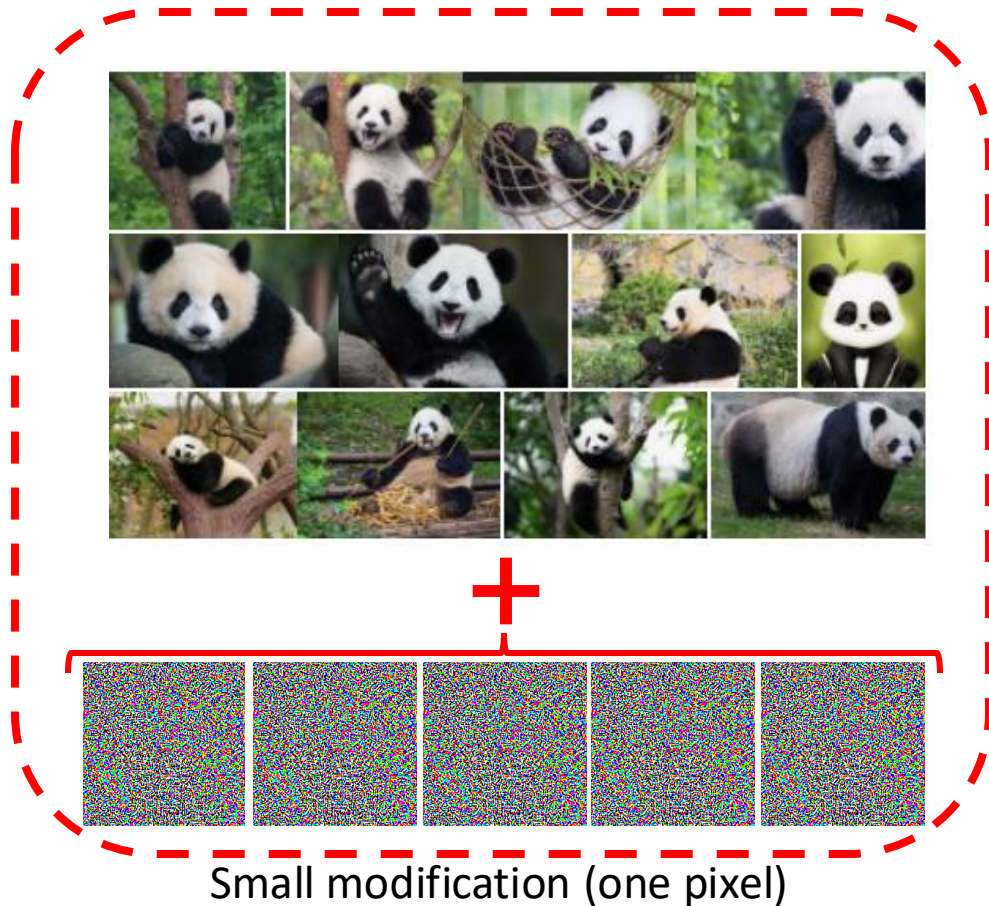


ML model for
panda recognition

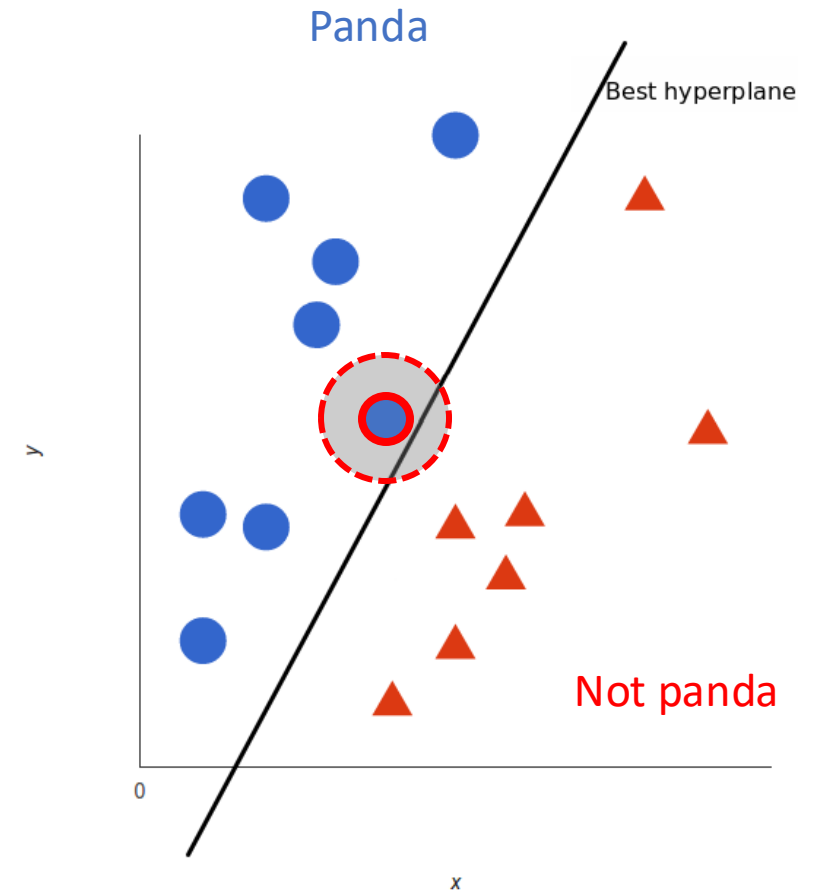


Learning with adversarial data

Q: can we build a new training set that includes adversarial data?

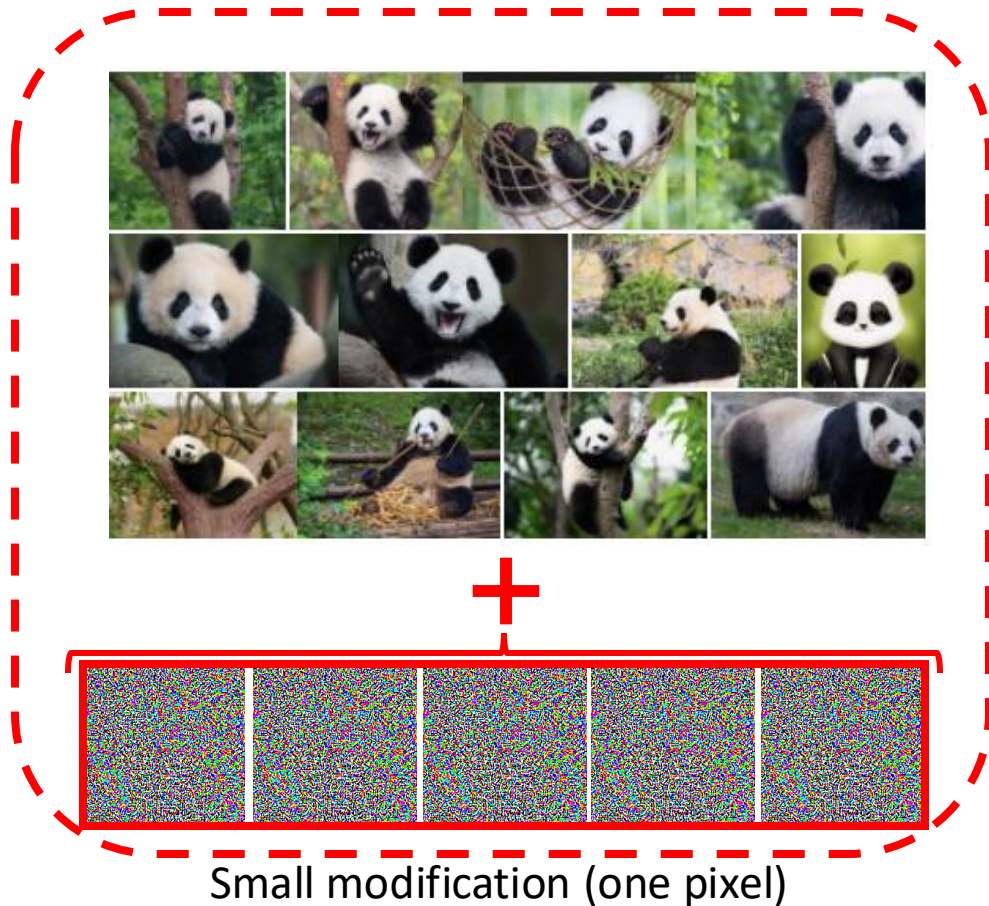


ML model for
panda recognition



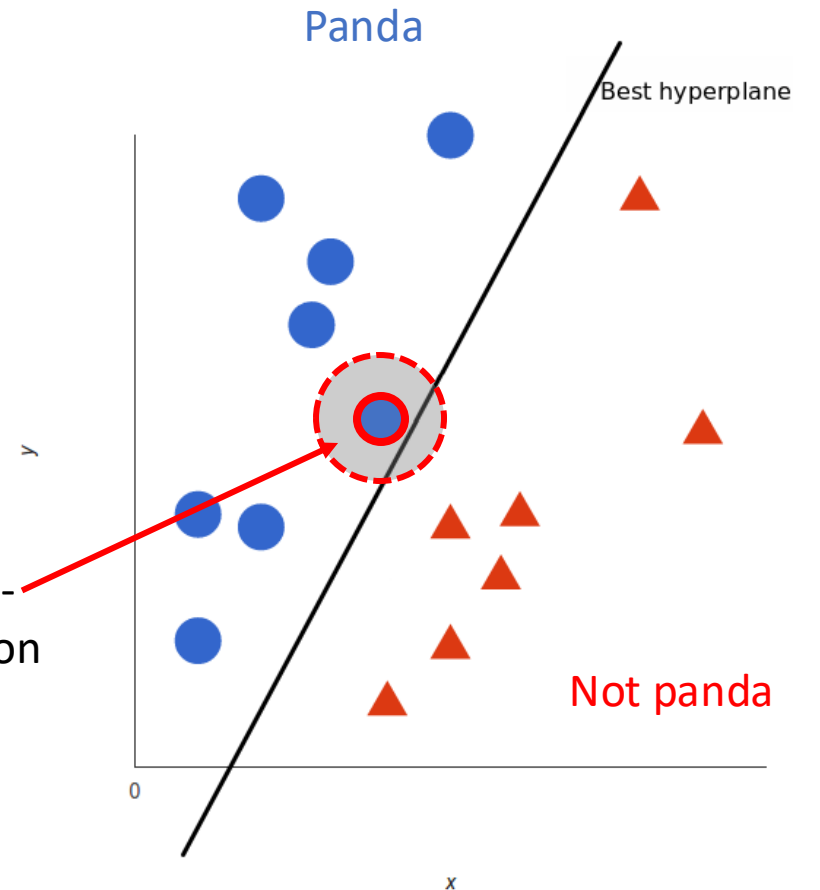
Learning with adversarial data

Q: can we build a new training set that includes adversarial data?



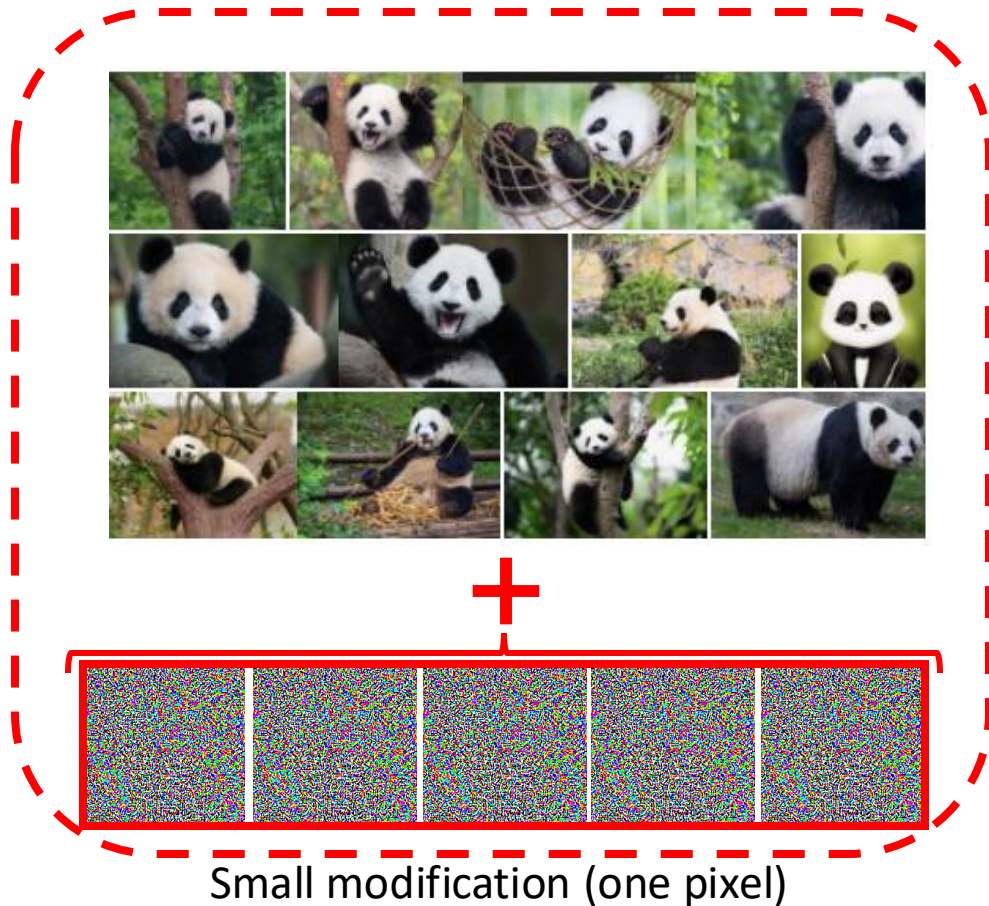
ML model for panda recognition

All possible one-pixel modification



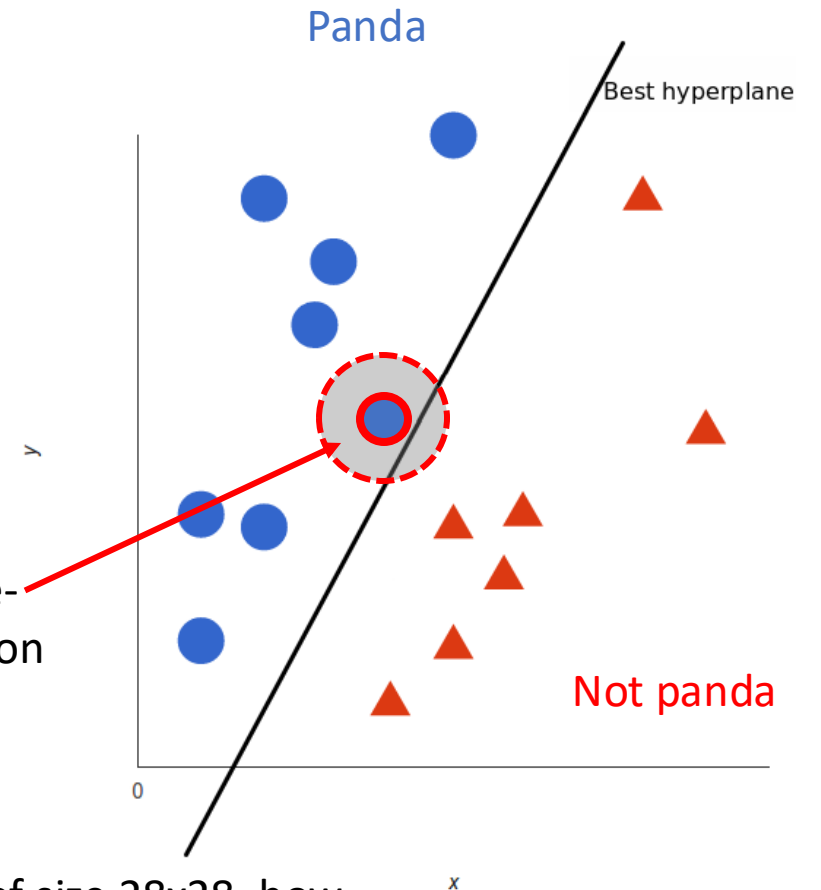
Learning with adversarial data

Q: can we build a new training set that includes adversarial data?



ML model for panda recognition

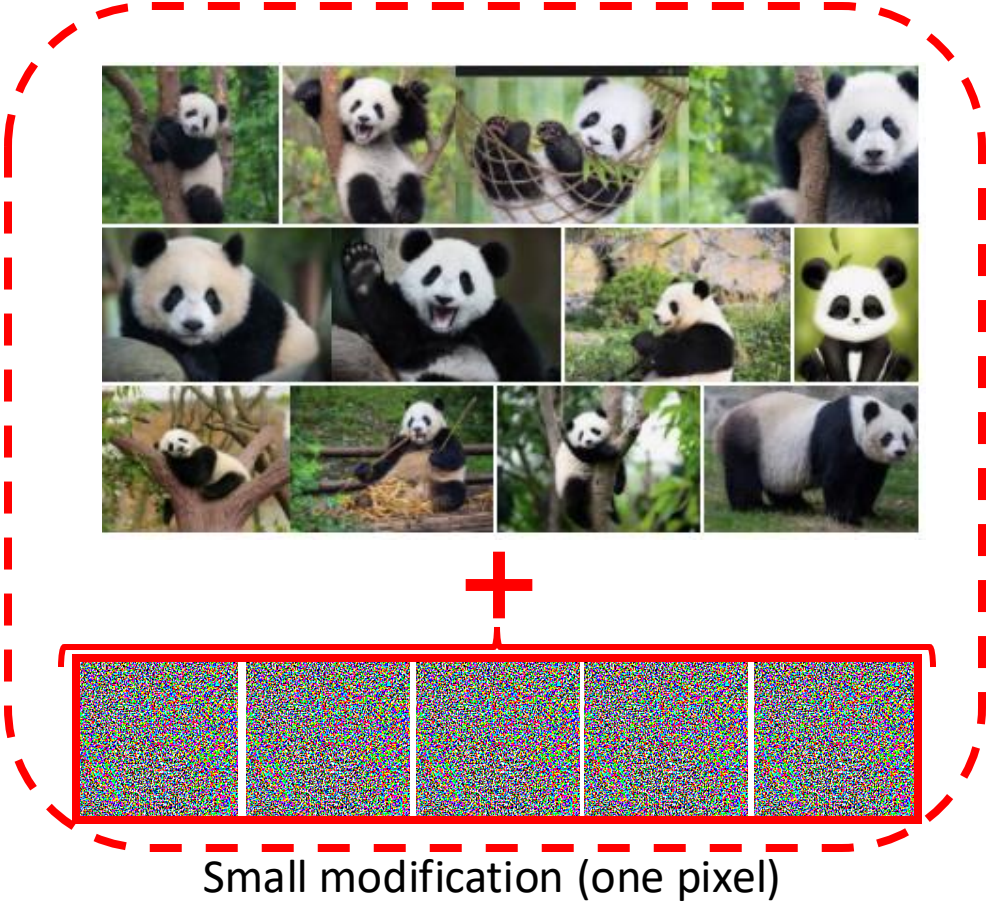
All possible one-pixel modification



Q: for a grey scale images of size 28x28, how many possible one-pixel changes can we have?

Learning with adversarial data

Q: can we build a new training set that includes adversarial data?

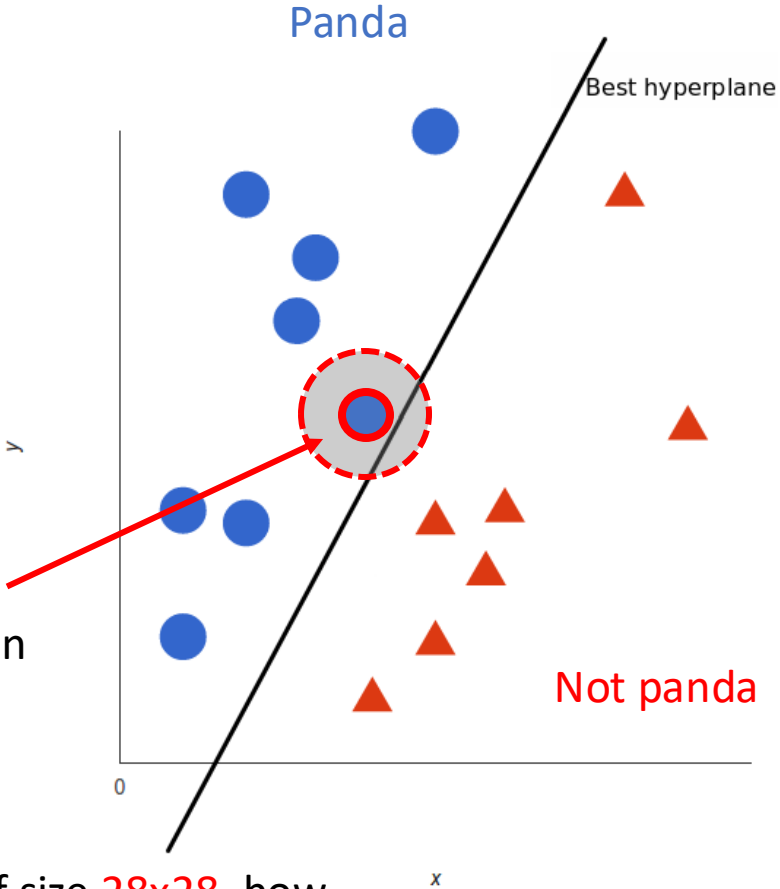


ML model for panda recognition

All possible one-pixel modification

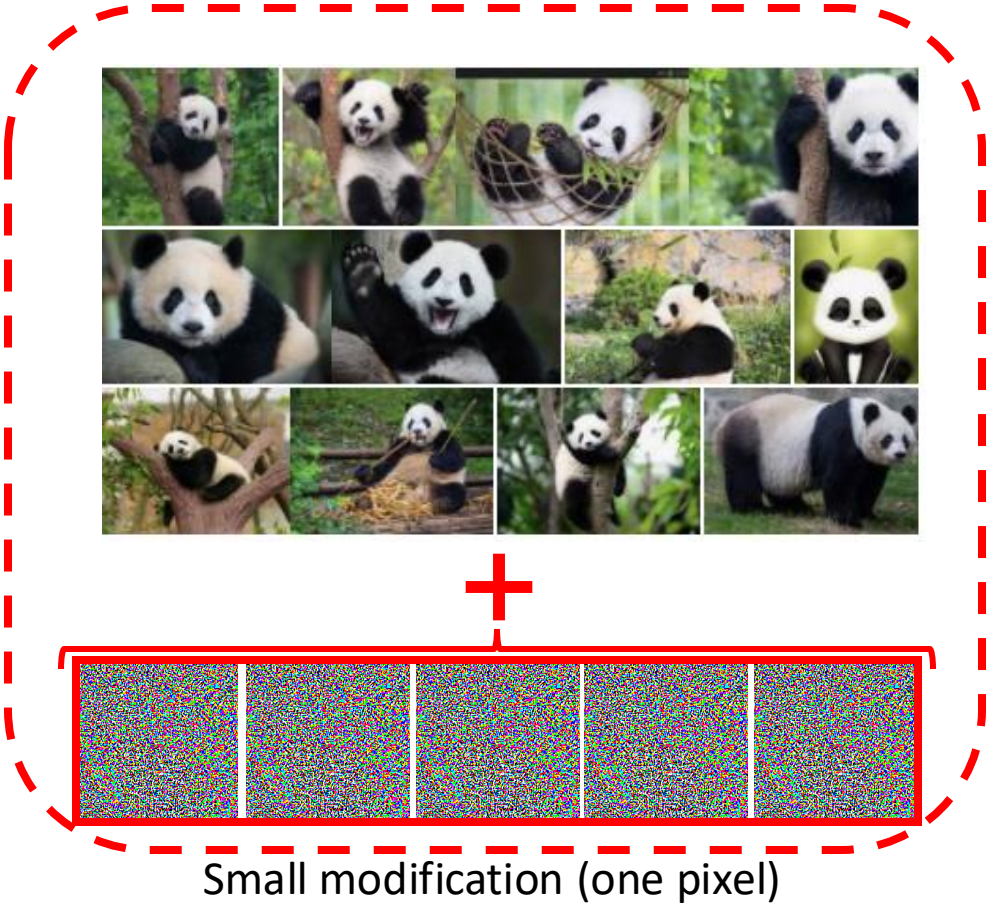
0~255

Q: for a grey scale images of size 28x28, how many possible one-pixel changes can we have?



Learning with adversarial data

Q: can we build a new training set that includes adversarial data?



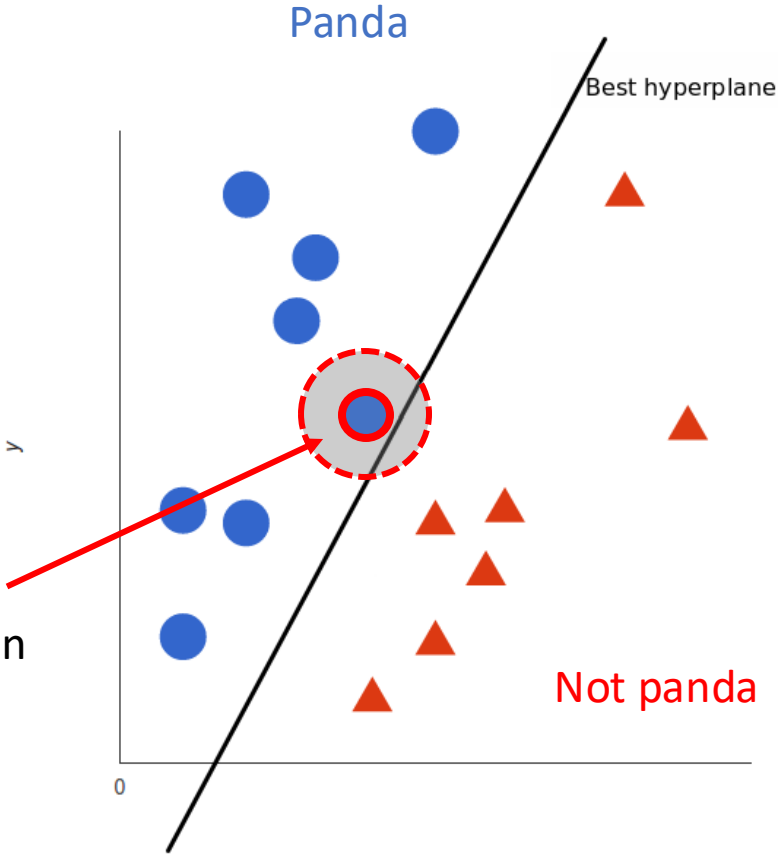
ML model for panda recognition

All possible one-pixel modification

0~255

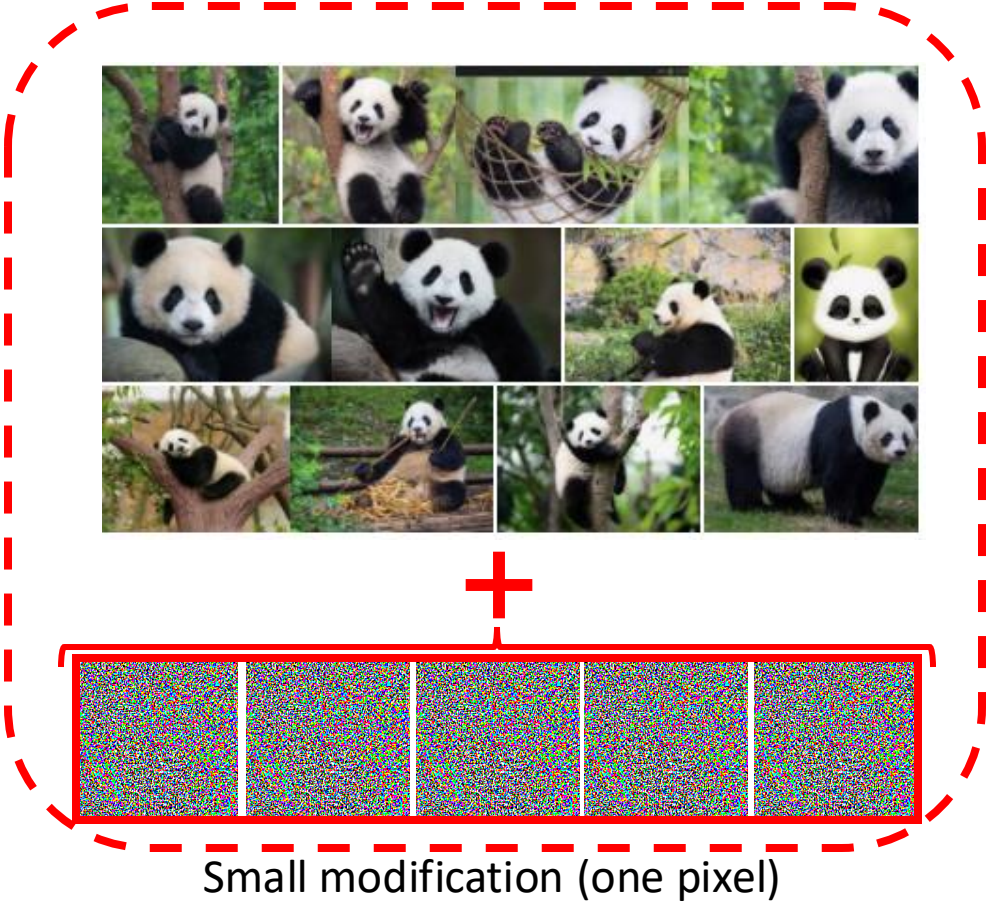
Q: for a grey scale images of size 28x28, how many possible one-pixel changes can we have?

$$255 \times 28 \times 28 = 199920$$



Learning with adversarial data

Q: can we build a new training set that includes adversarial data?



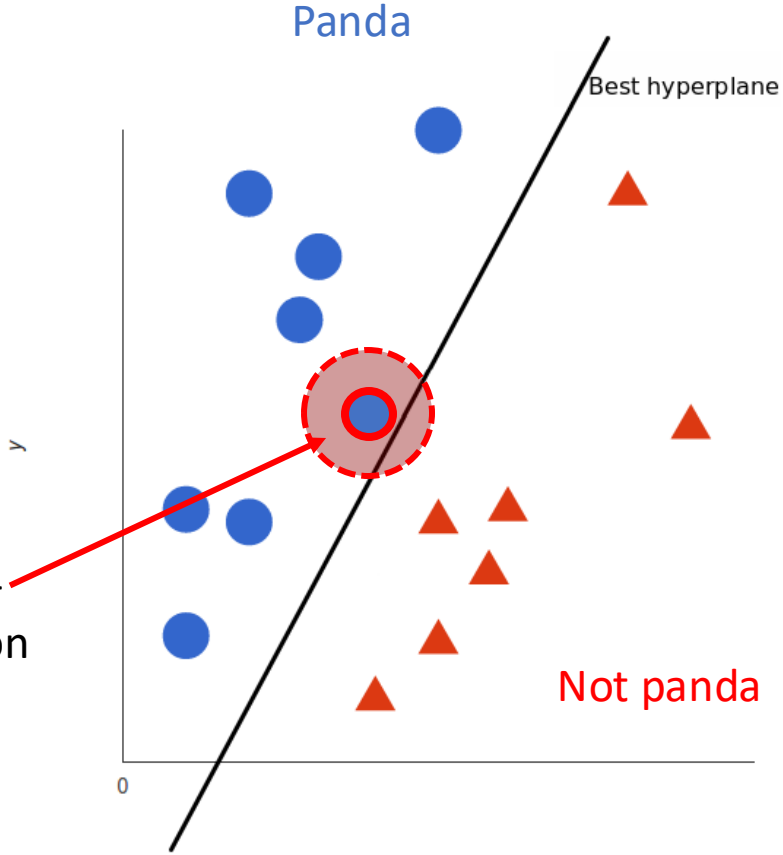
ML model for panda recognition

All possible one-pixel modification

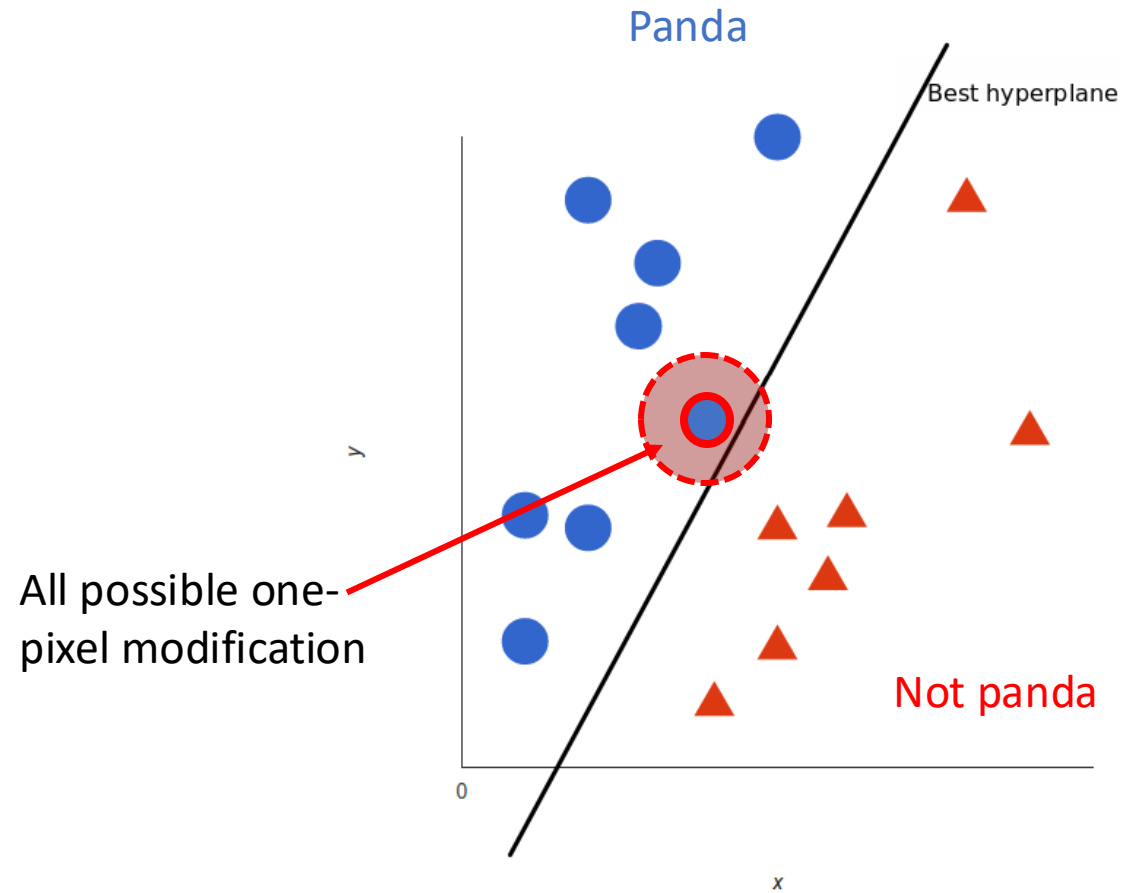
0~255

Q: for a grey scale images of size 28x28, how many possible one-pixel changes can we have?

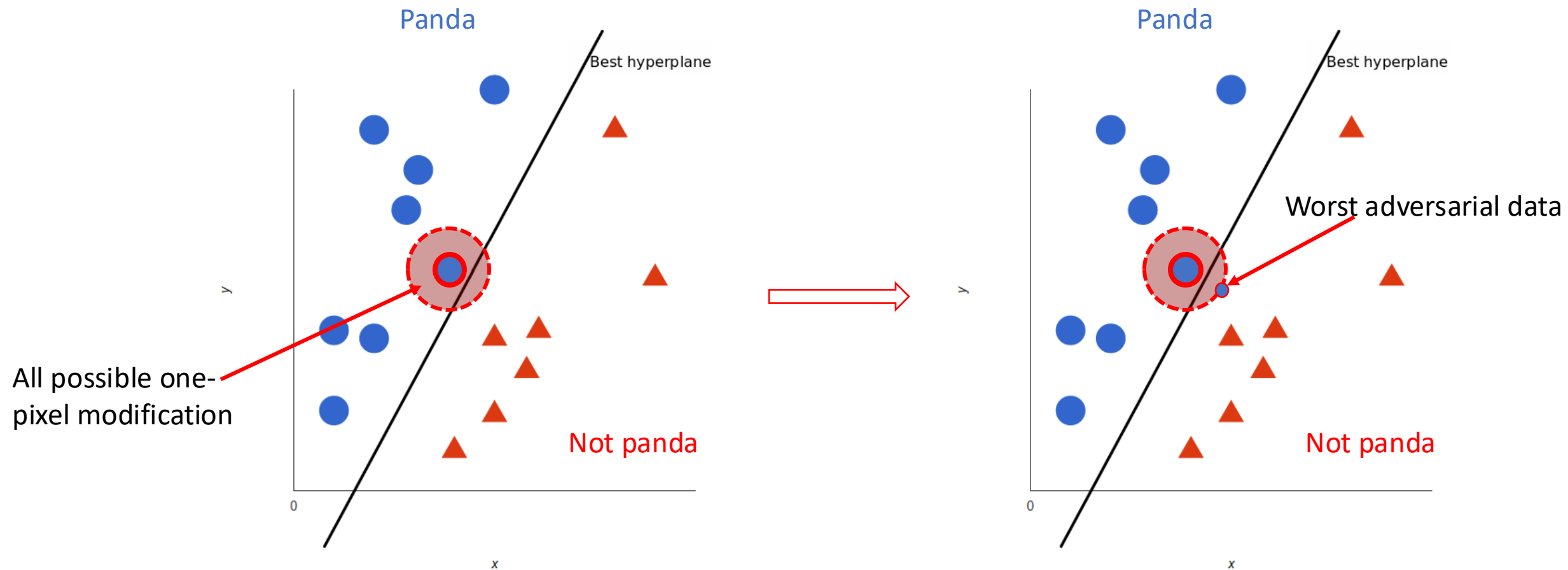
$$255 \times 28 \times 28 = 199920$$



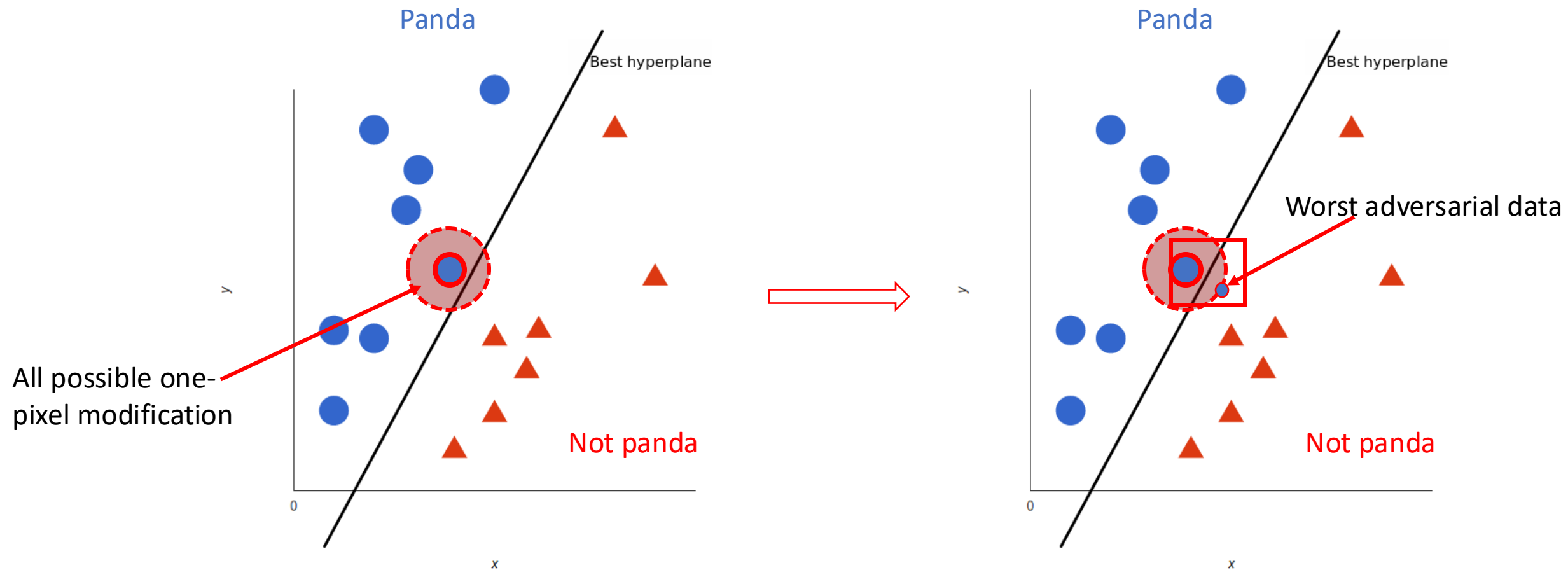
Worst case minimization



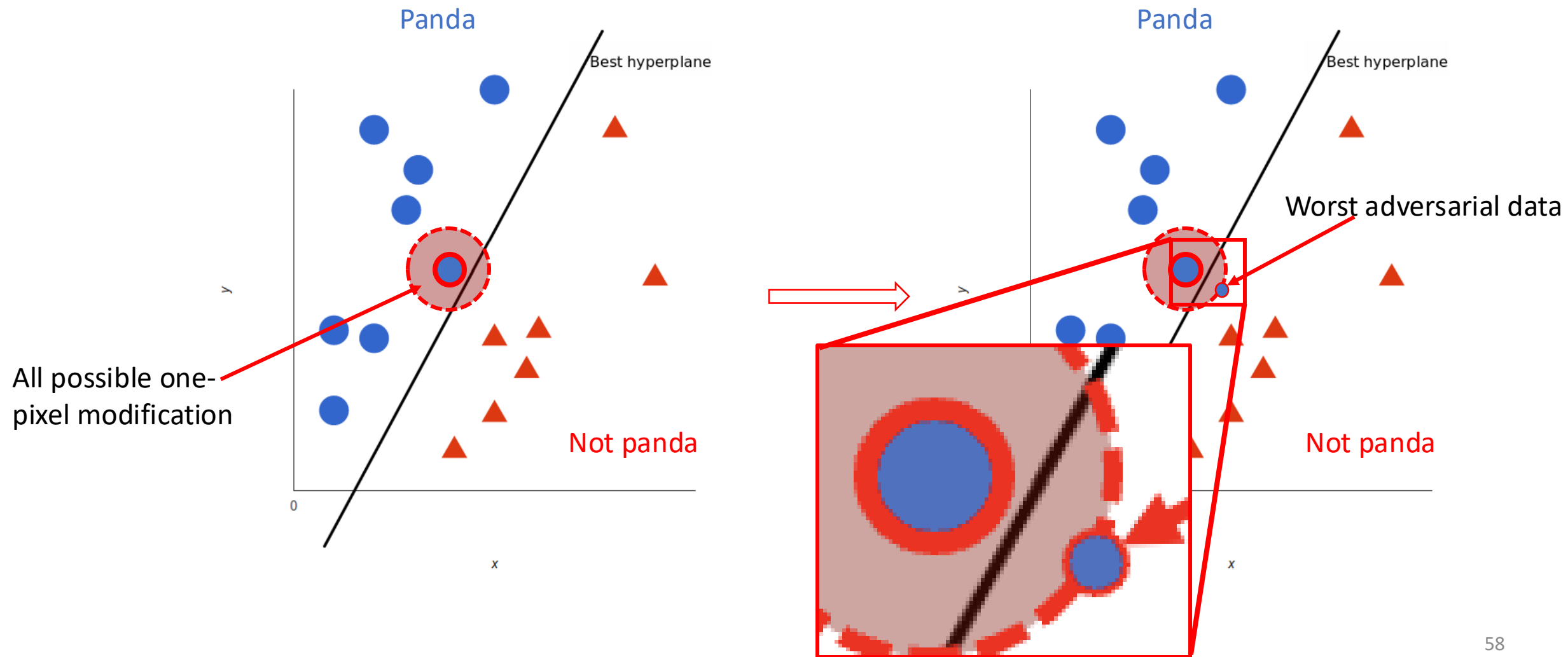
Worst case minimization



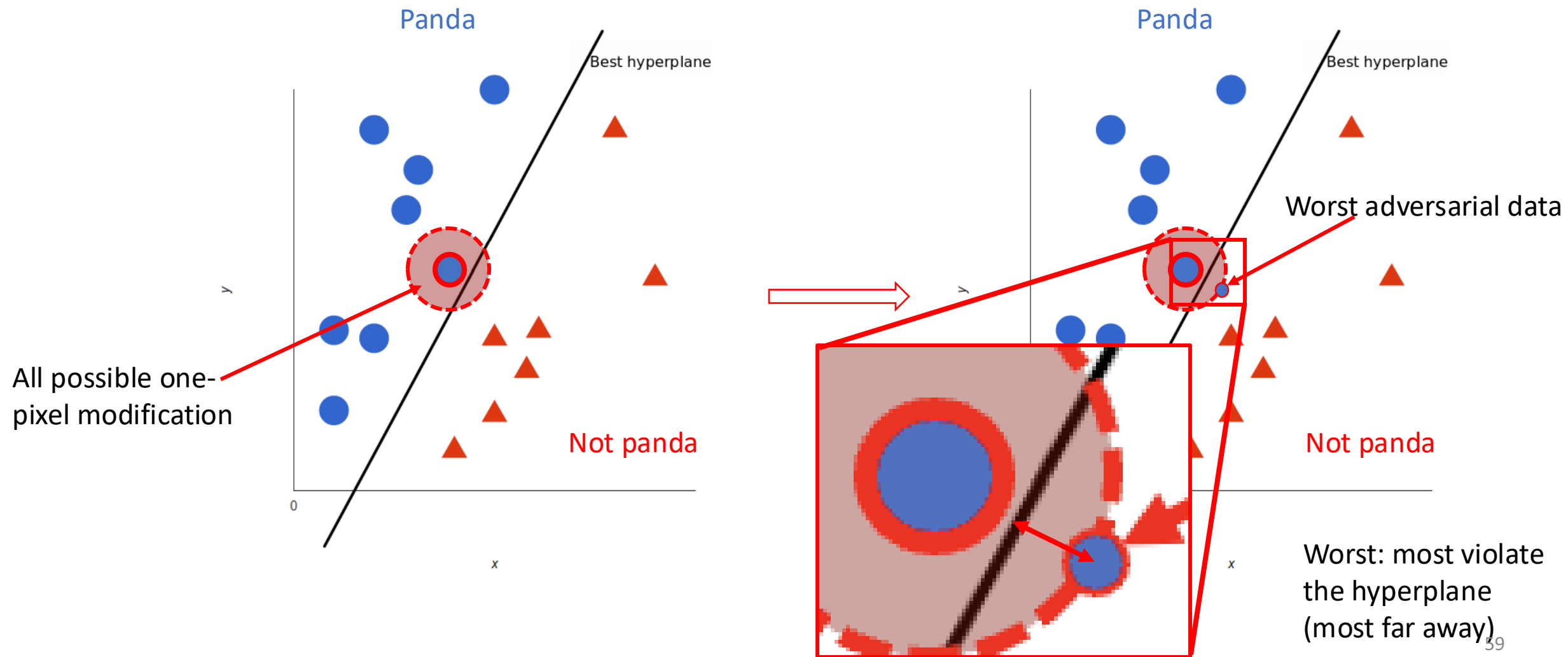
Worst case minimization



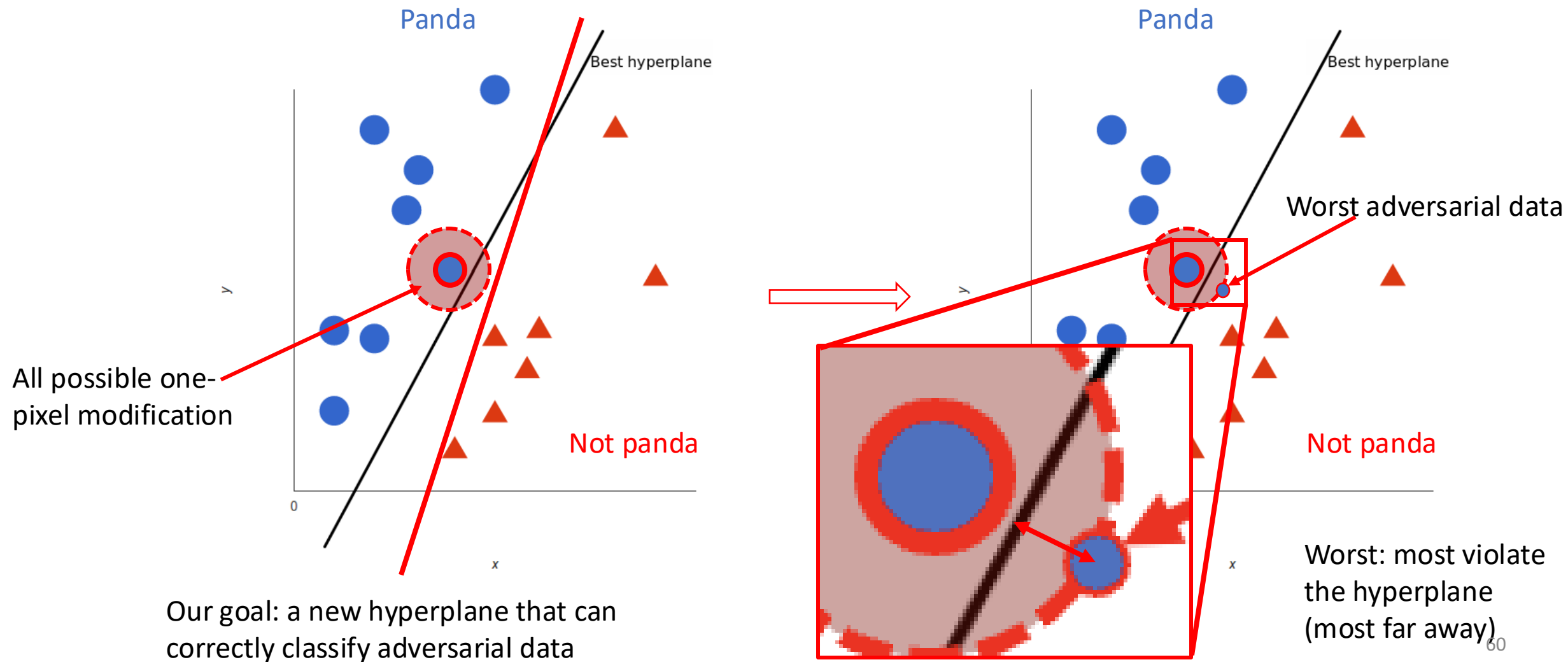
Worst case minimization



Worst case minimization



Worst case minimization



Our goal: a new hyperplane that can correctly classify adversarial data