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

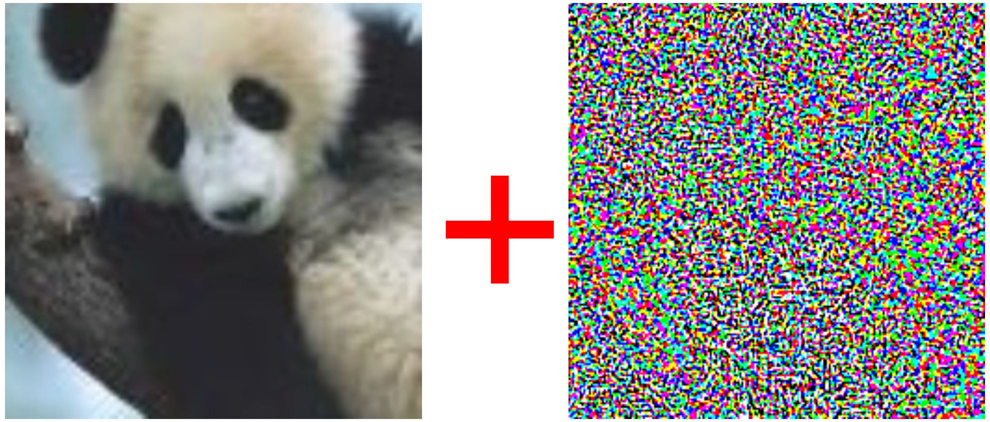
ML model for panda recognition

Testing data

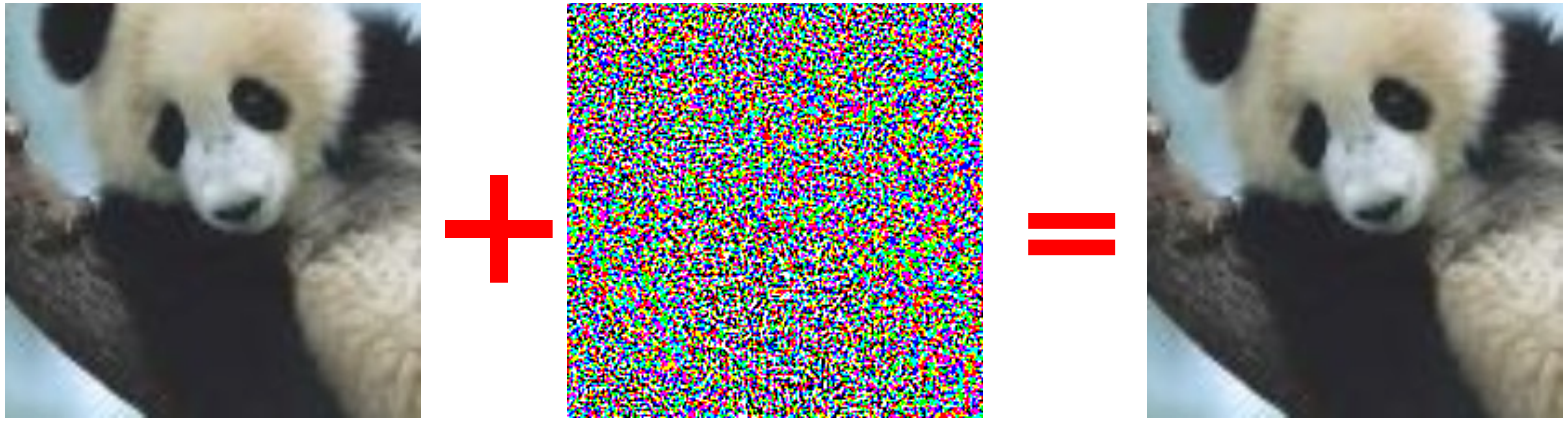
Training data

Adversarial data

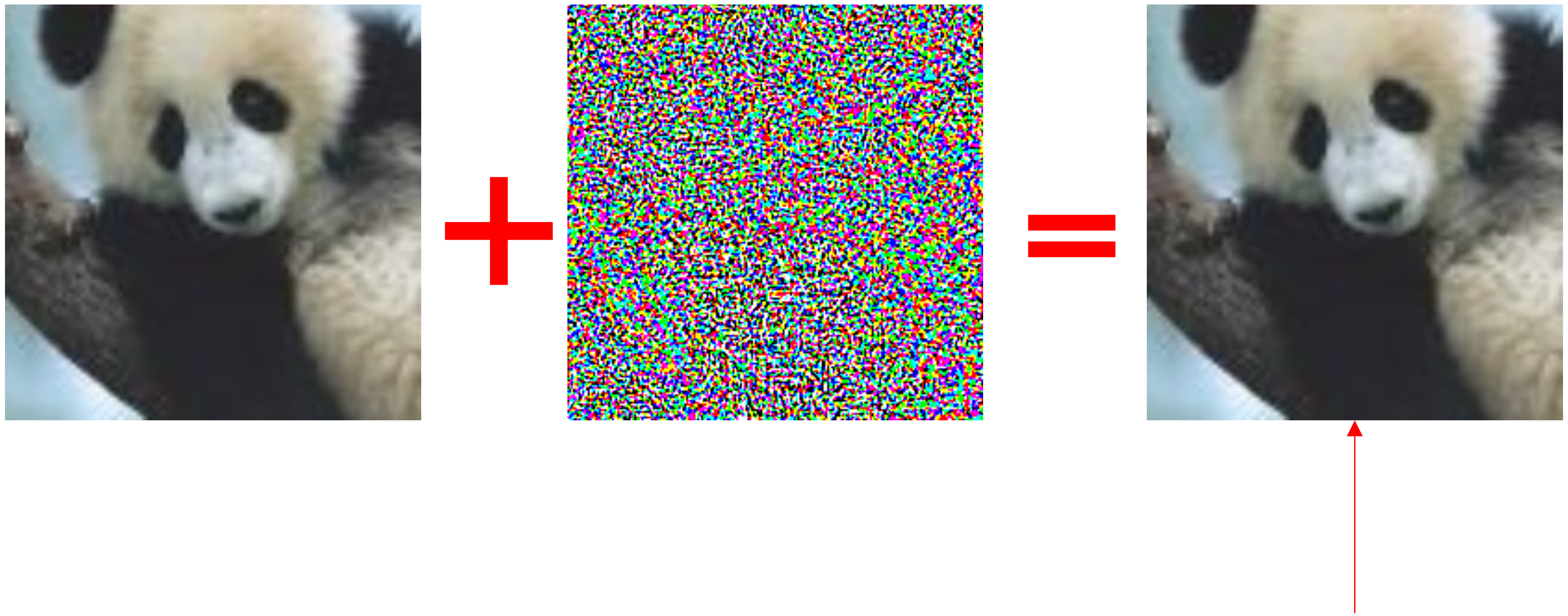
Adversarial data



Adversarial data

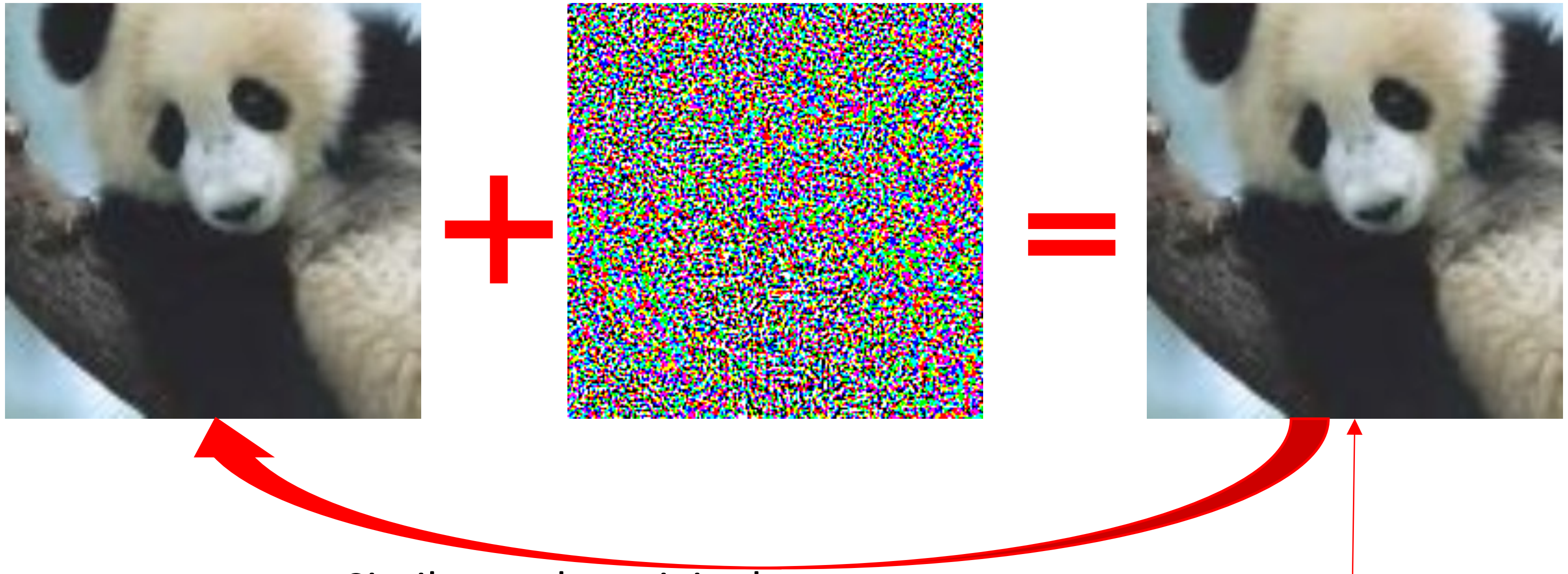


Adversarial data



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

ML model for panda recognition

Original

Adversarial

Machine learning model VS adversarial data

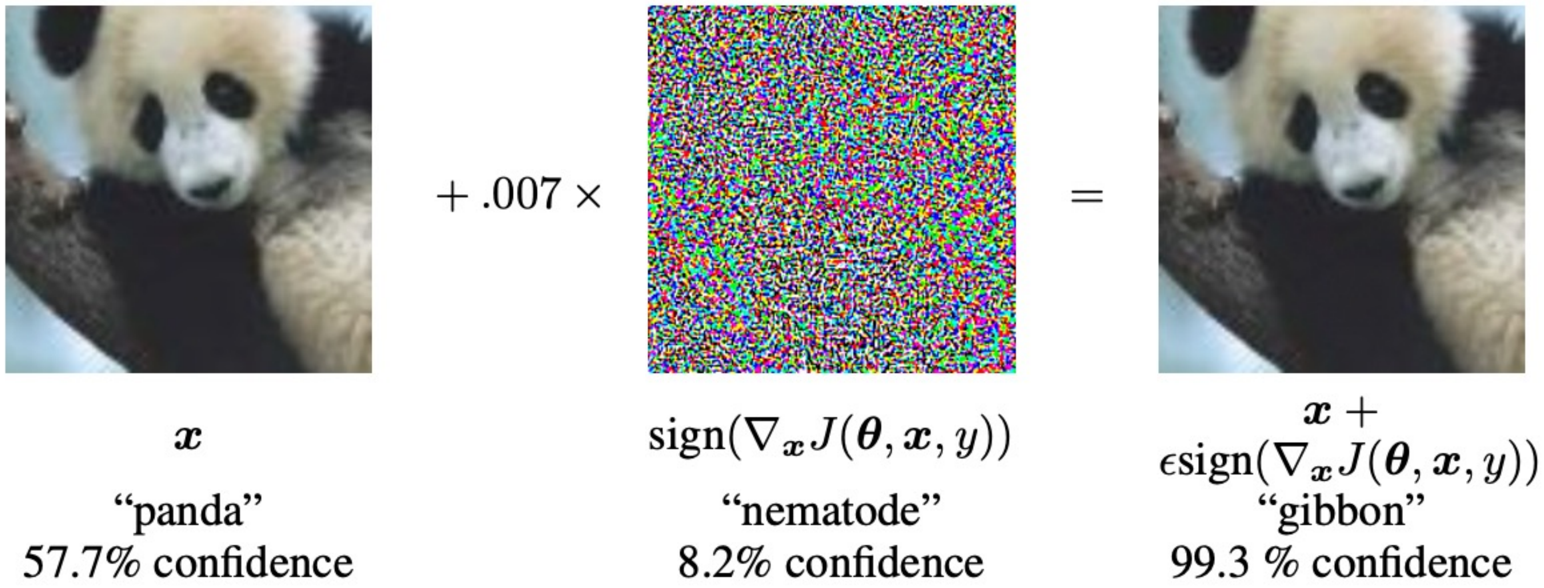


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

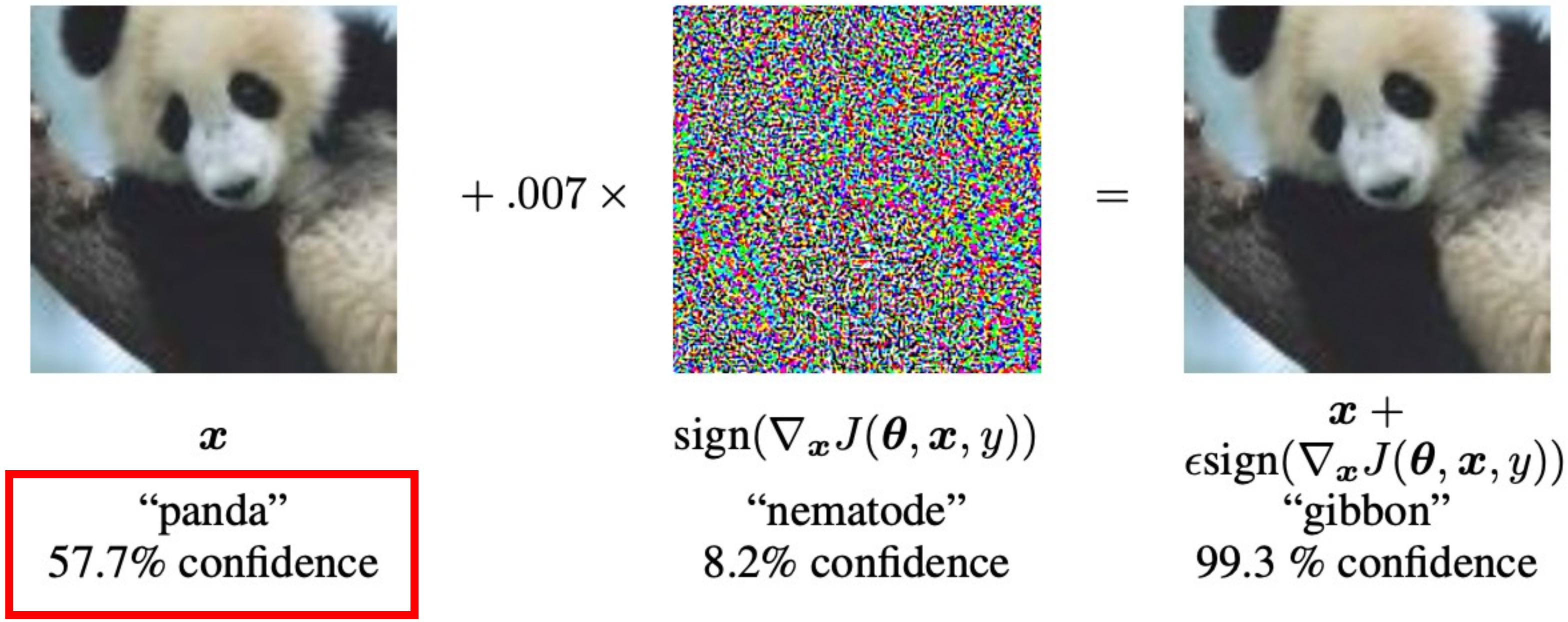


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

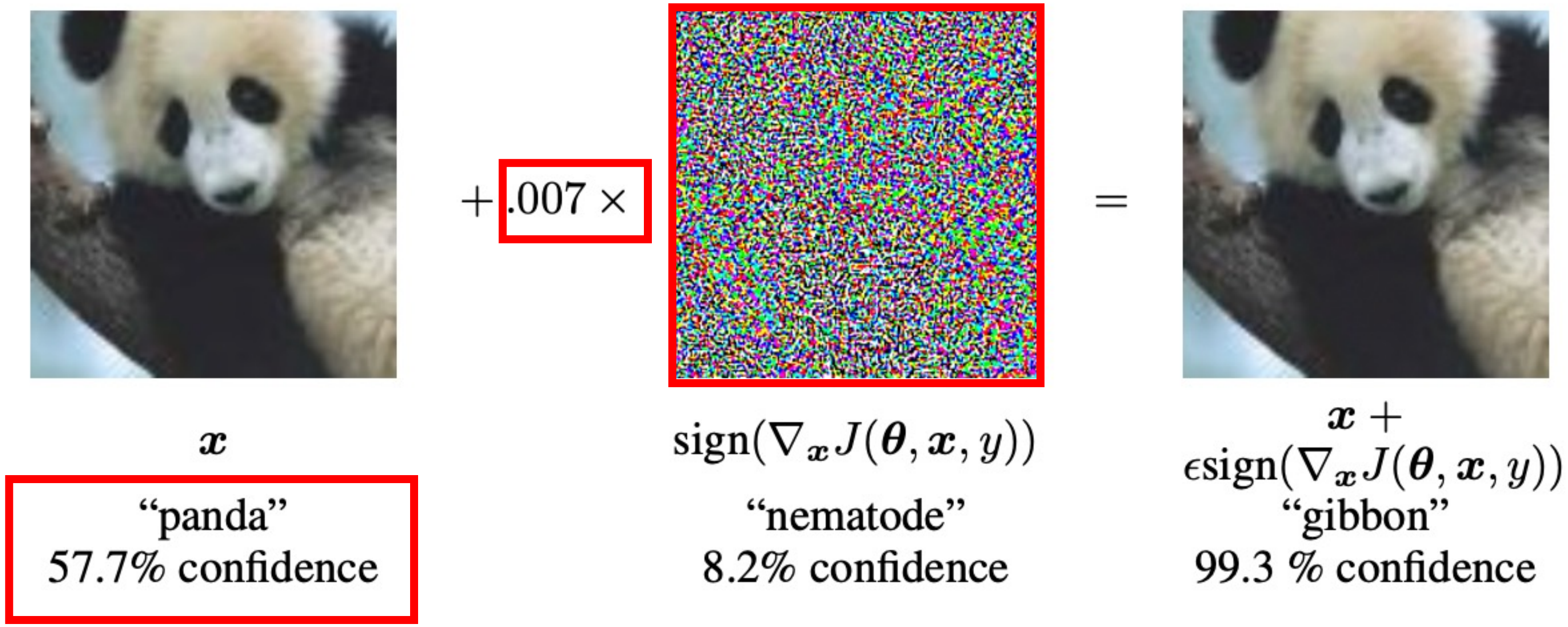


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

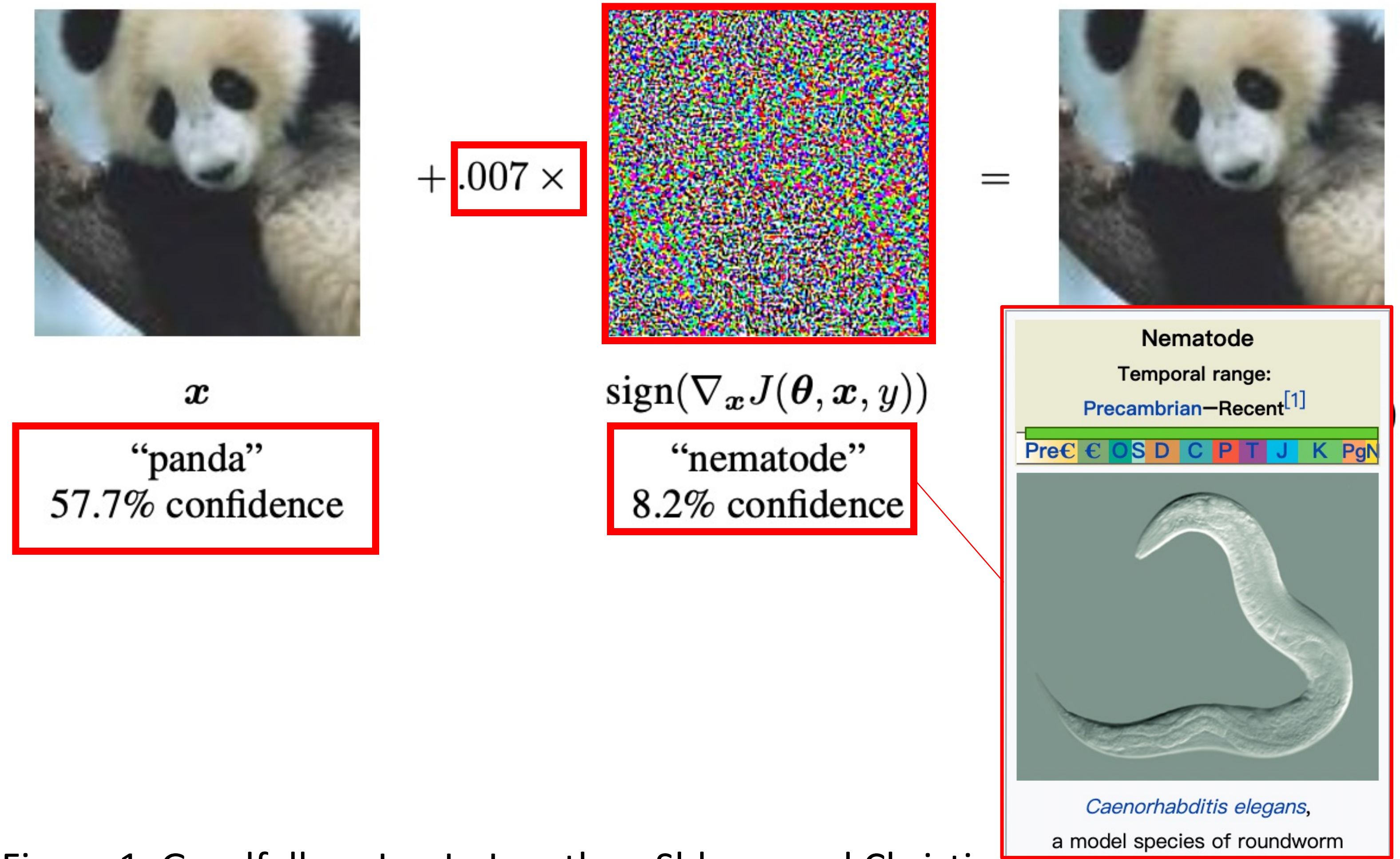


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

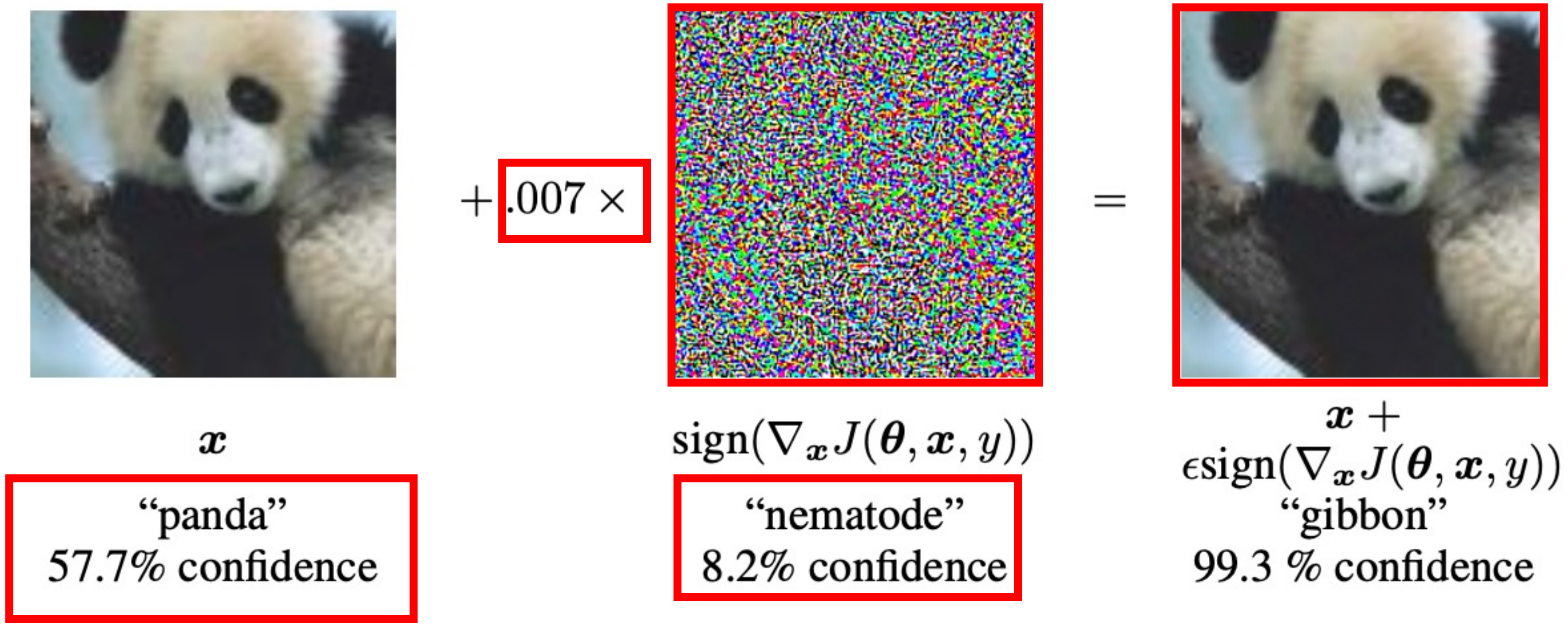


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Machine learning model VS adversarial data

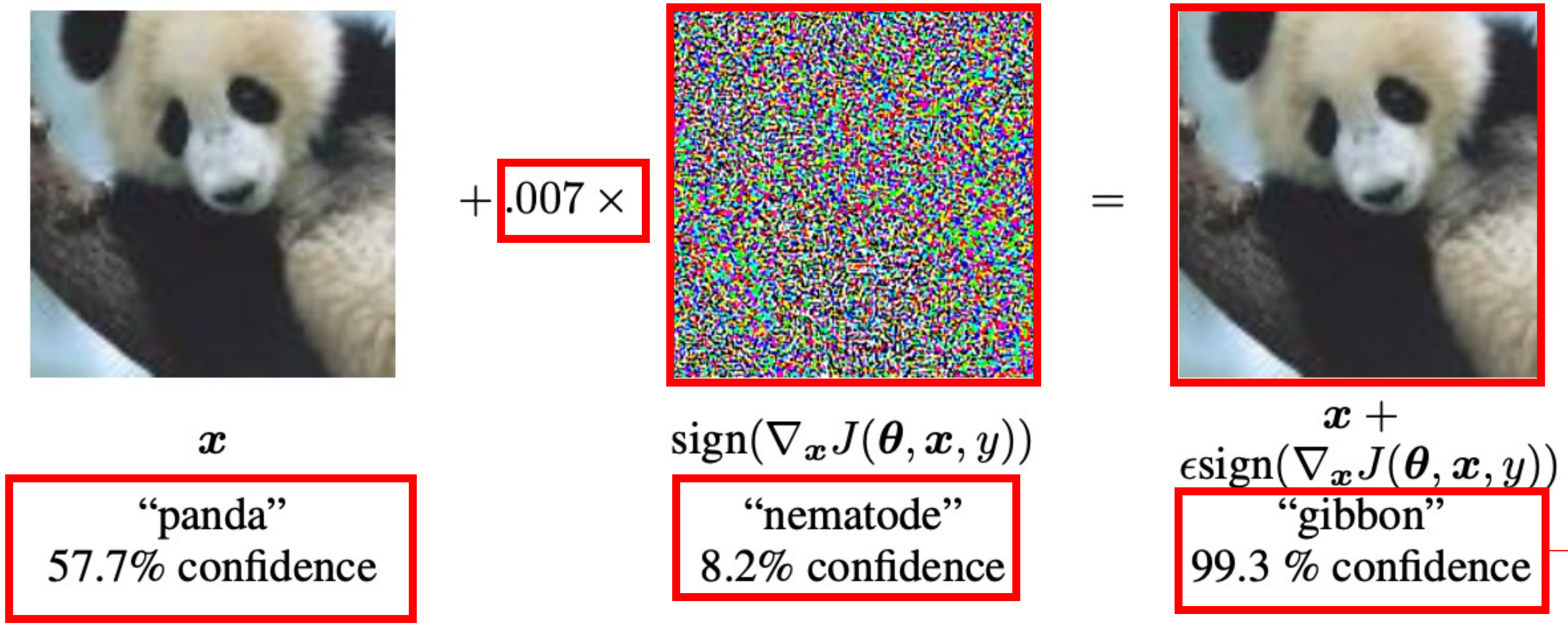
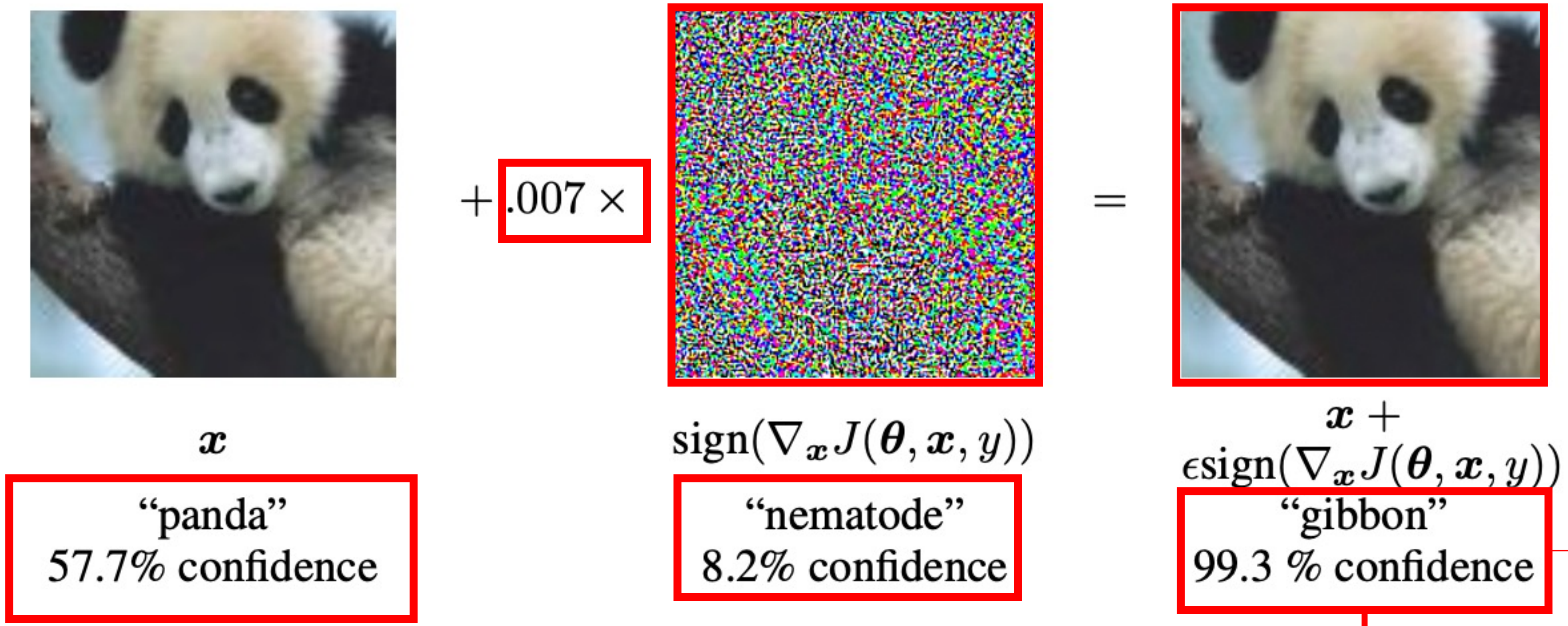


Figure 1, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

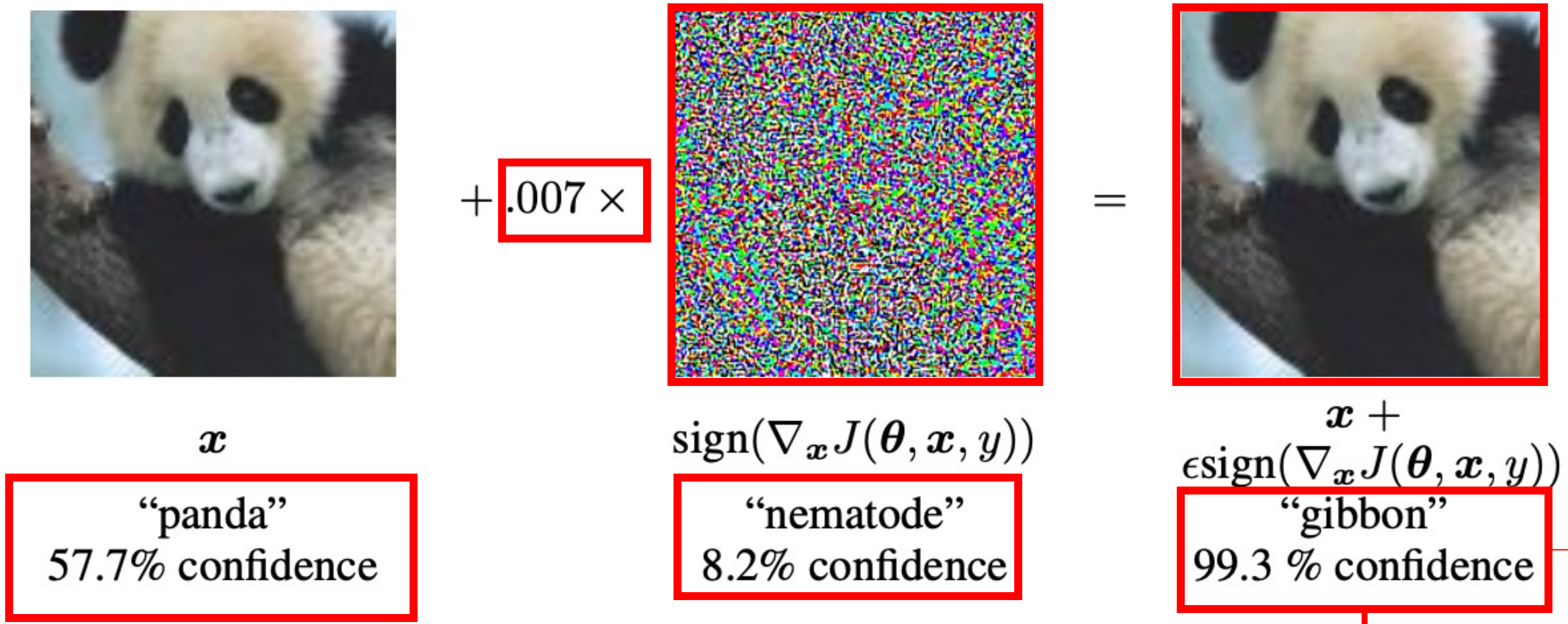
Machine learning model VS adversarial data



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

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625-1634. 2018.

Why care about attacks

A real Stop sign

A physical perturbation applied to a Stop sign

Artificial patches

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625-1634. 2018.

Why care about attacks

A real Stop sign

A physical perturbation applied to a Stop sign

Prediction result?

Artificial patches

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625-1634. 2018.

Why care about attacks

A real Stop sign

A physical perturbation applied to a Stop sign

Prediction result?

Artificial patches

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625-1634. 2018.

Why care about attacks

45 MPH

Prediction result?

A real Stop sign

A physical perturbation applied to a Stop sign

Artificial patches

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625-1634. 2018.

Why care about attacks

What if a driver recognize STOP as 45 MPH?

45 MPH

Prediction result?

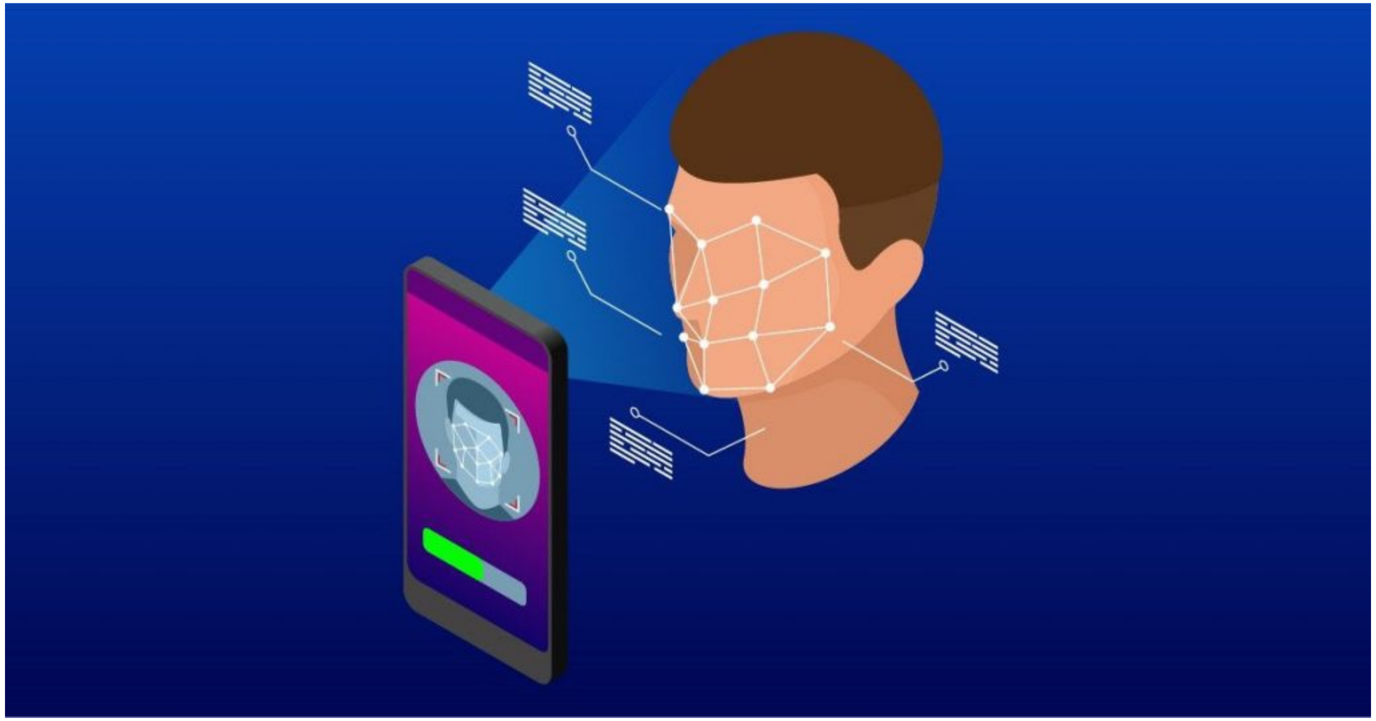
A real Stop sign

Artificial patches

A physical perturbation applied to a Stop sign

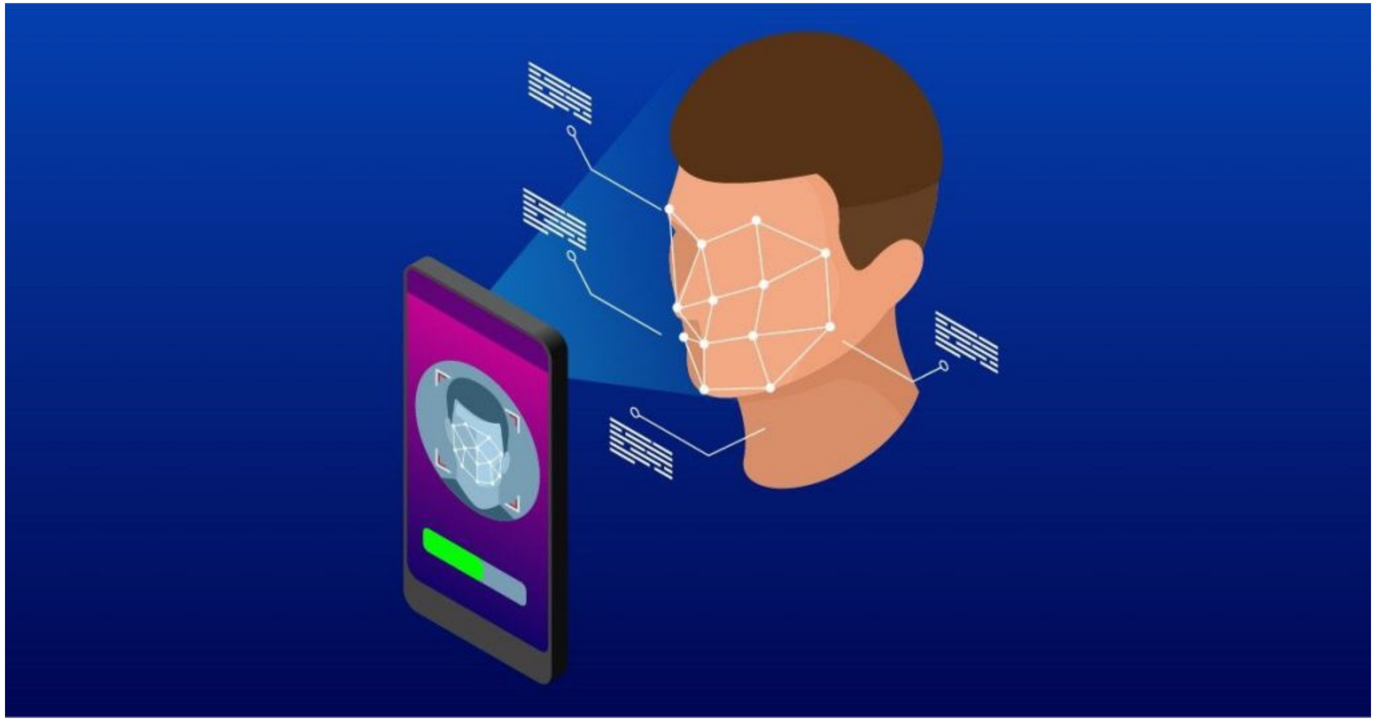
Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625-1634. 2018.

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?

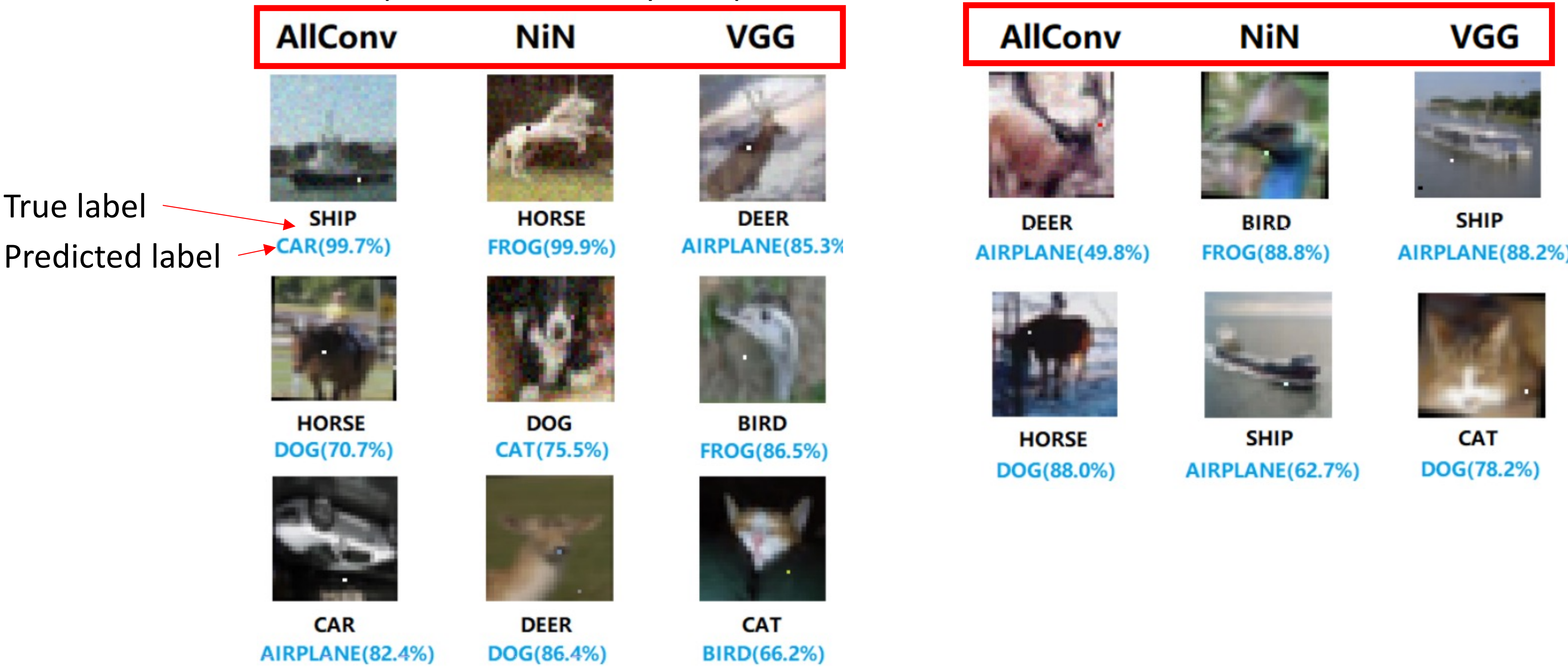
Q: Can we fool deep models with only one pixel modified?

Why care about attacks/adversarial noise?

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

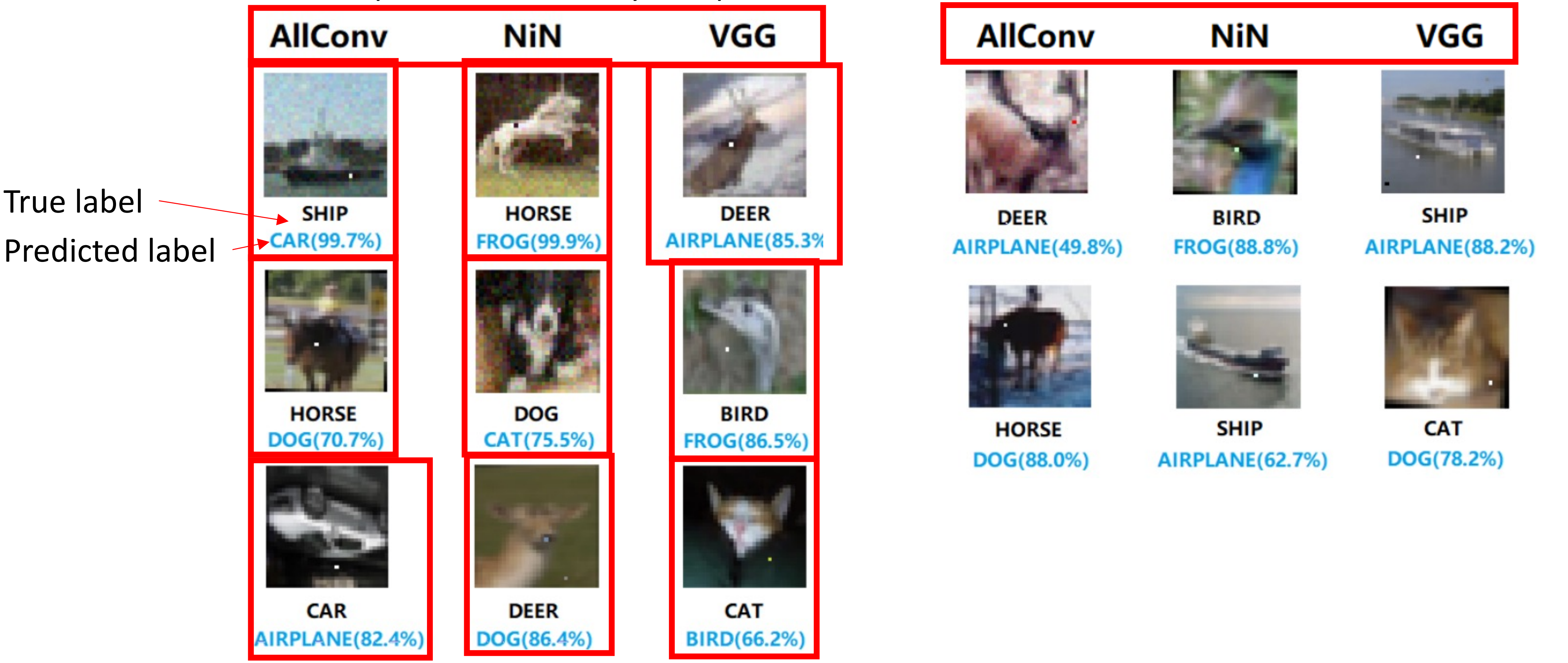
Why care about attacks/adversarial noise?

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



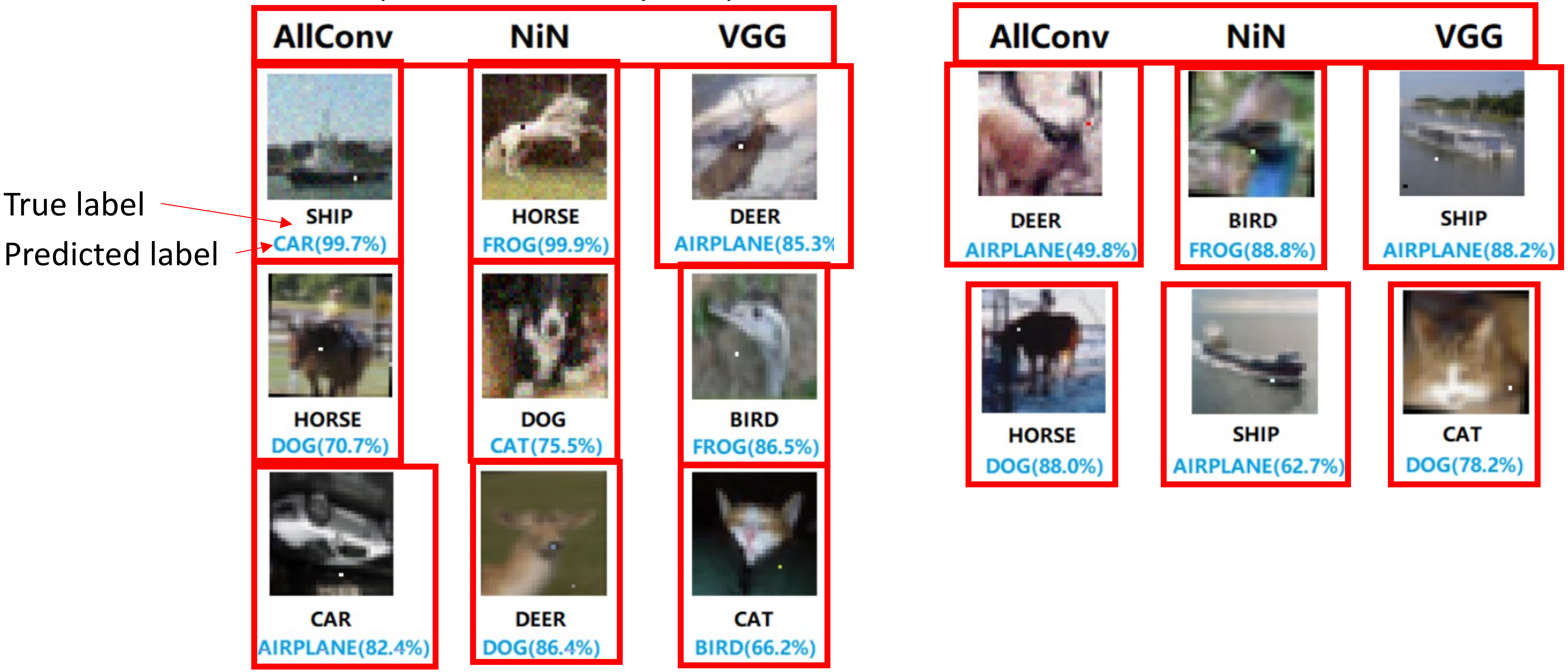
Why care about attacks/adversarial noise?

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



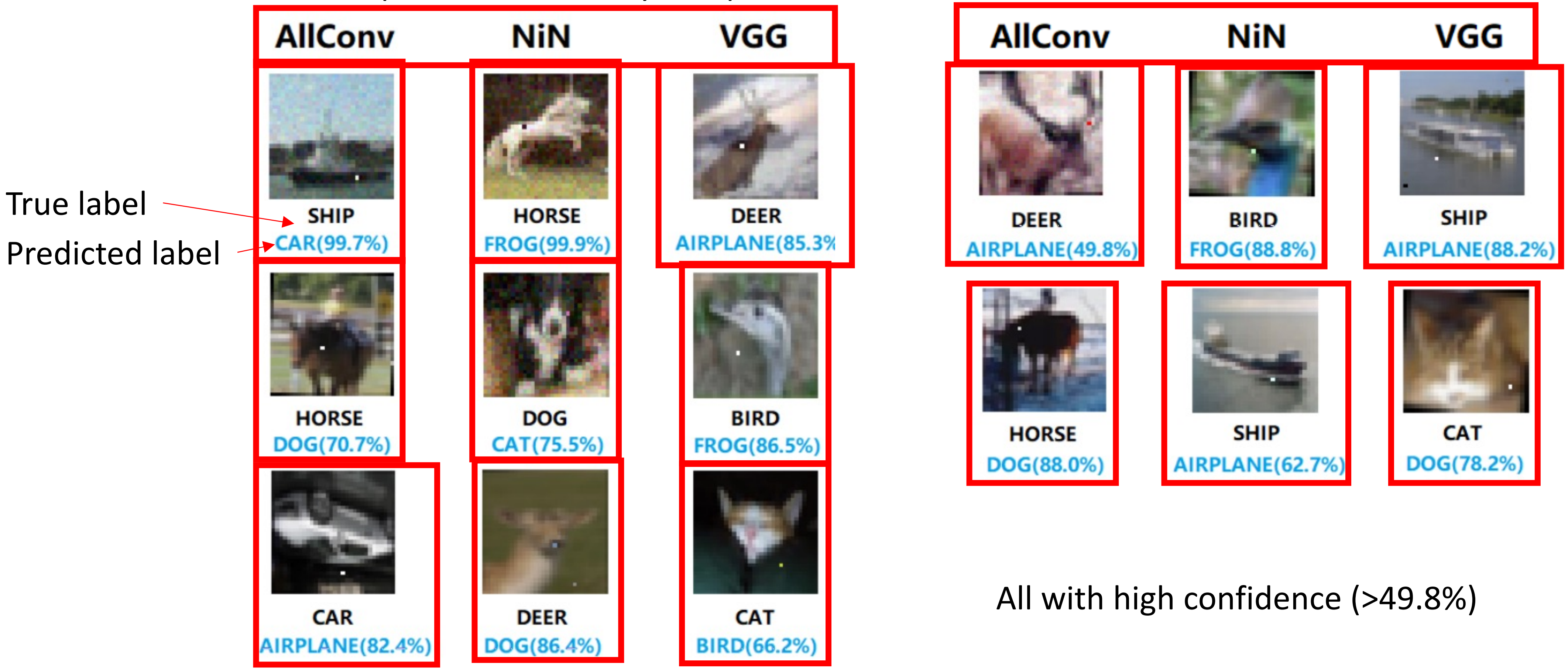
Why care about attacks/adversarial noise?

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



Why care about attacks/adversarial noise?

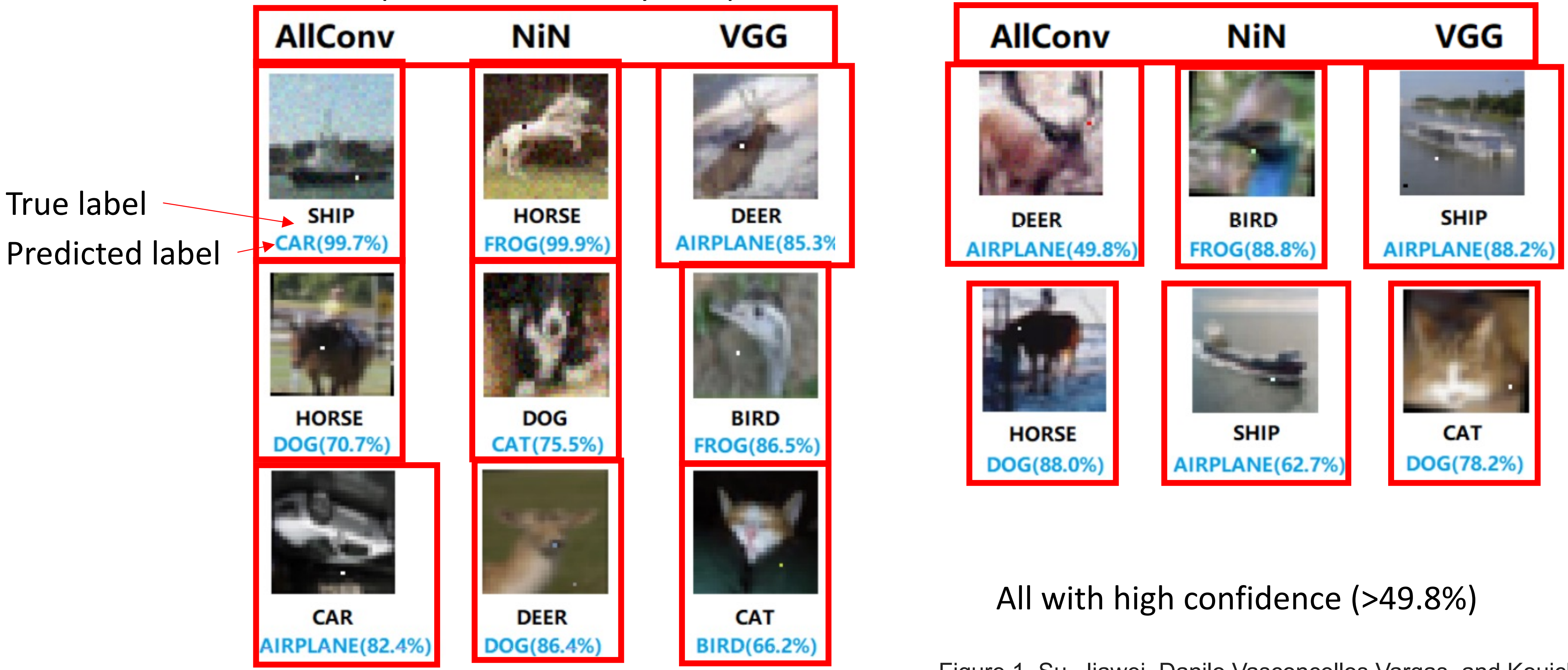
Q: Can we fool deep models with only one pixel modified?



All with high confidence (>49.8%)

Why care about attacks/adversarial noise?

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



All with high confidence (>49.8%)

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

ML model for panda recognition

Original image:

Classified as panda

Stable for noisy data

Adversarial image:

Classified as panda

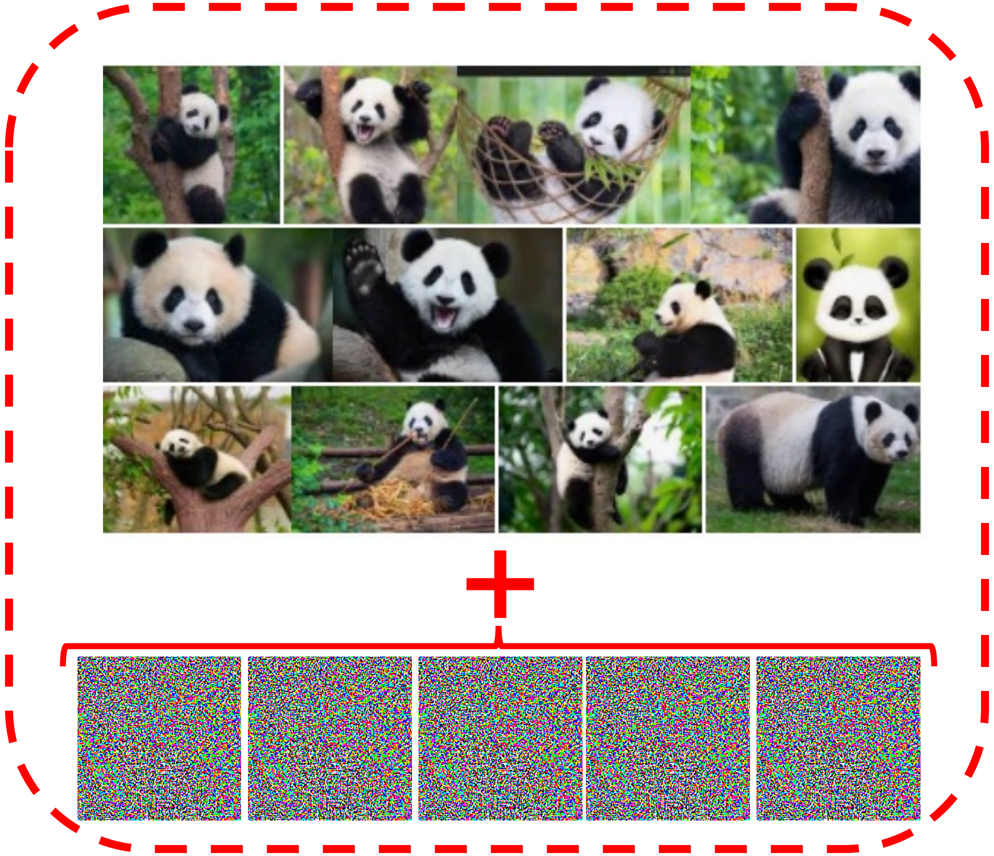
Learning with adversarial data

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

ML model for
panda recognition

Learning with adversarial data

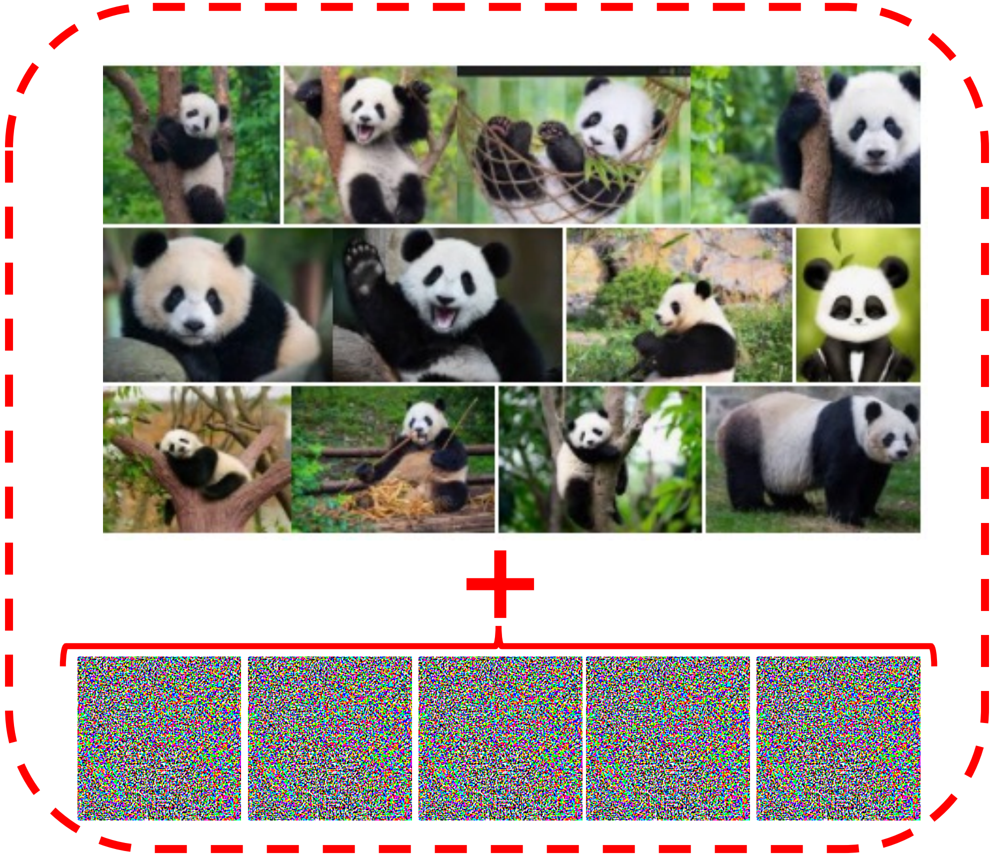
Q: can we build a new training set that includes adversarial data?



ML model for
panda recognition

Learning with adversarial data

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



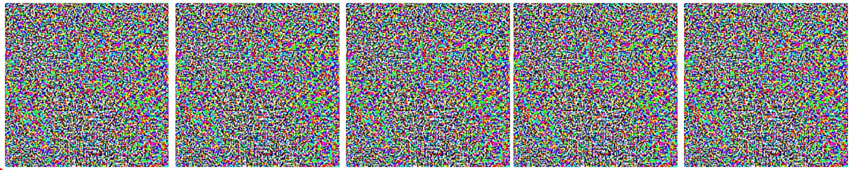
ML model for
panda recognition

Small modification (one pixel)

Learning with adversarial data

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

+



Small modification (one pixel)

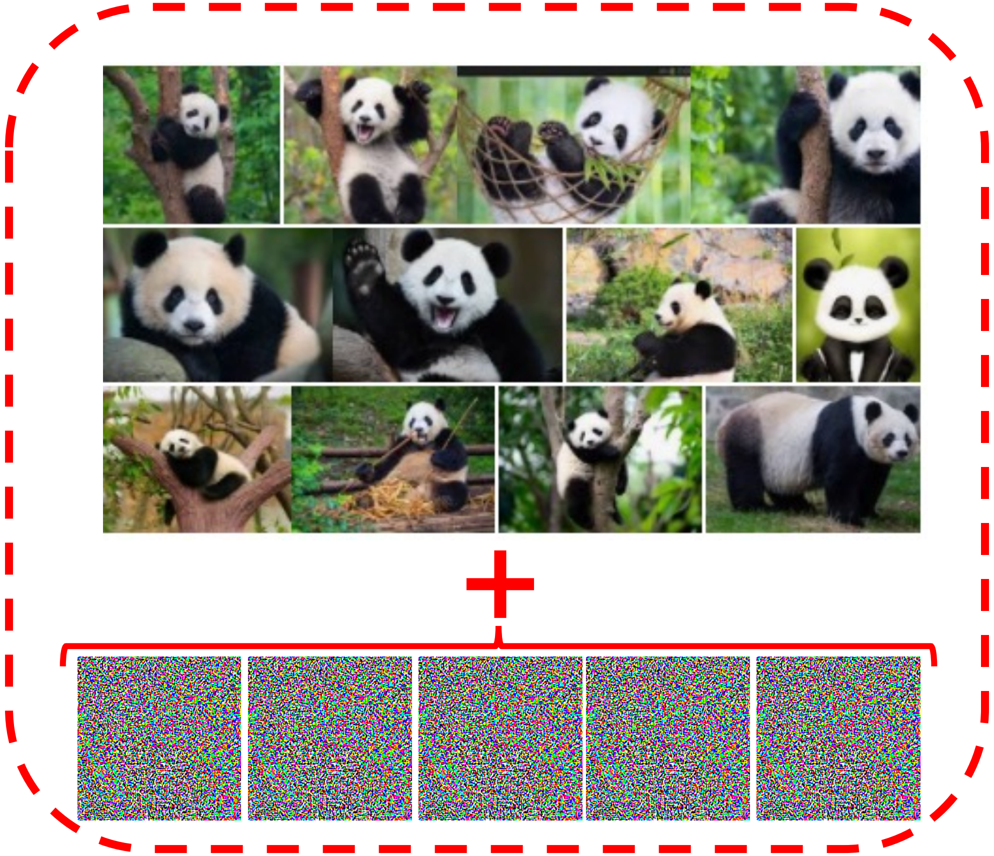
ML model for panda recognition

Original

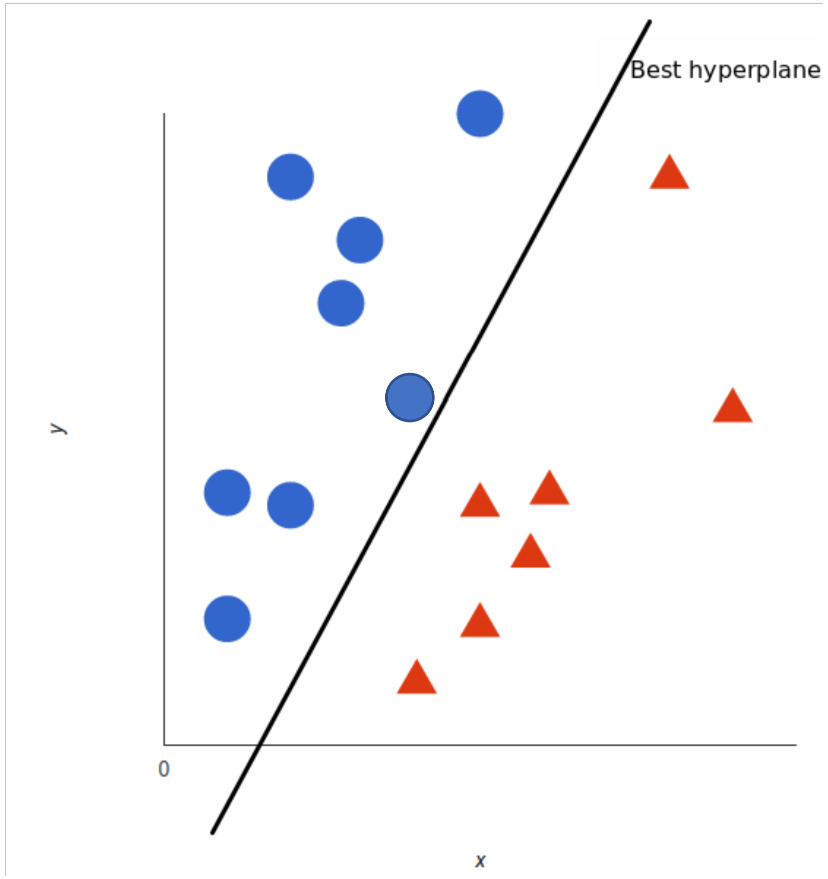
Adversarial

Learning with adversarial data

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

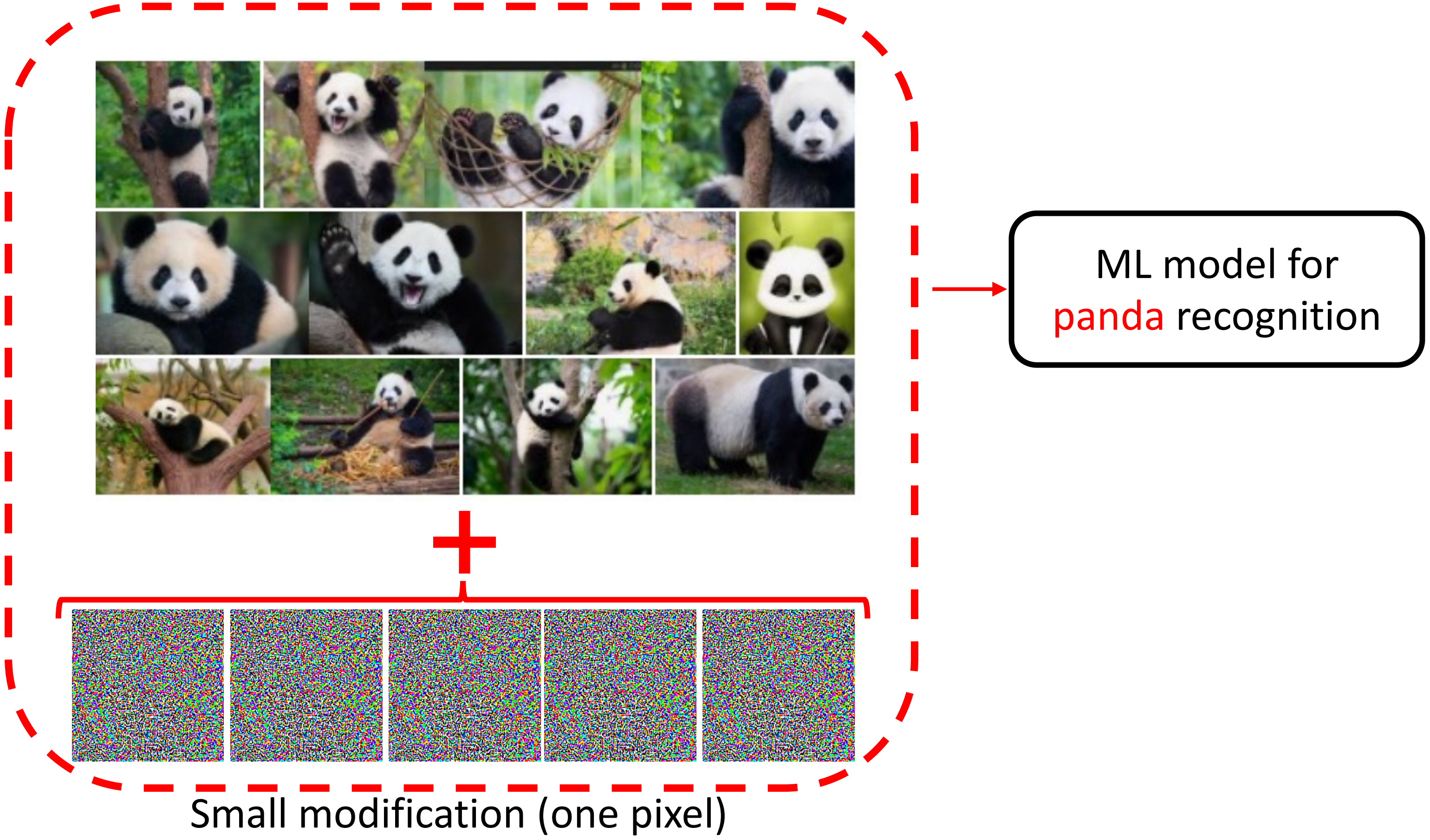


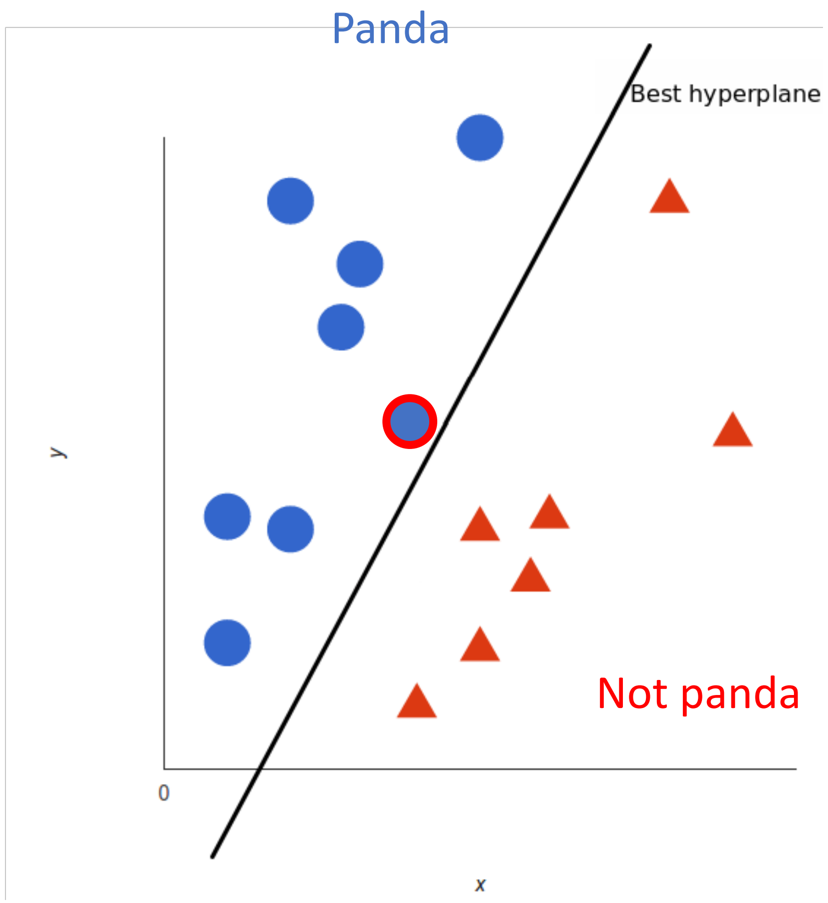
ML model for
panda recognition



Learning with adversarial data

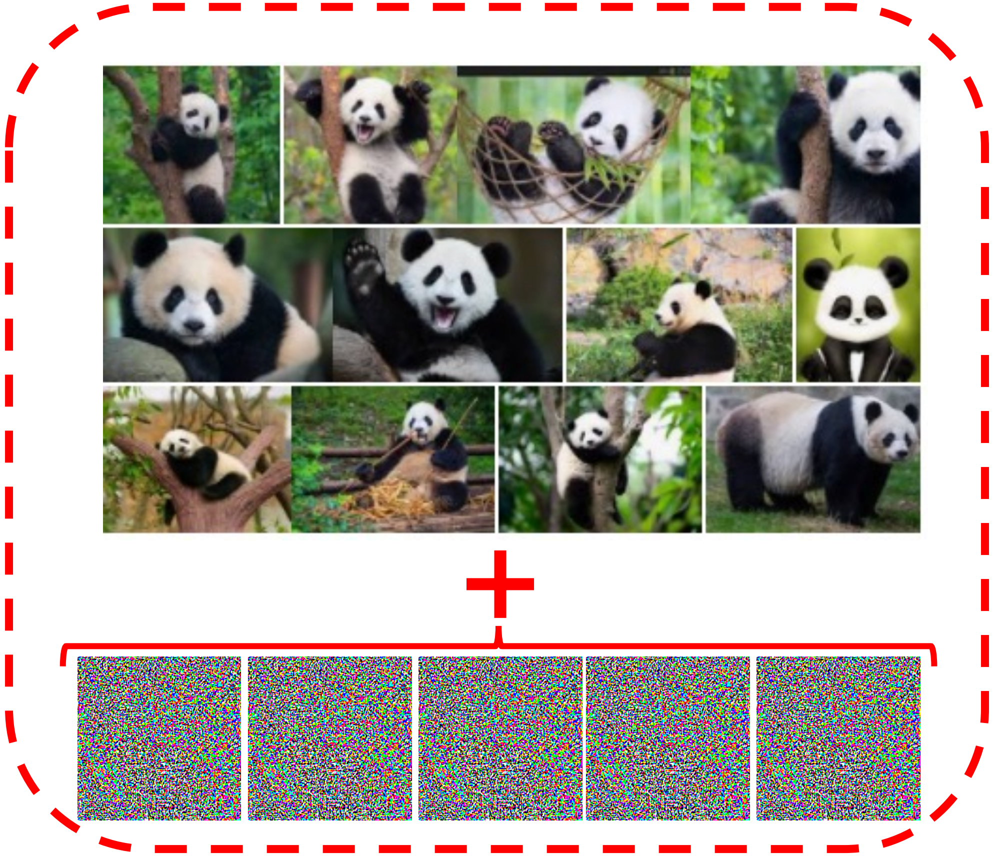
Q: can we build a new training set that includes adversarial data?





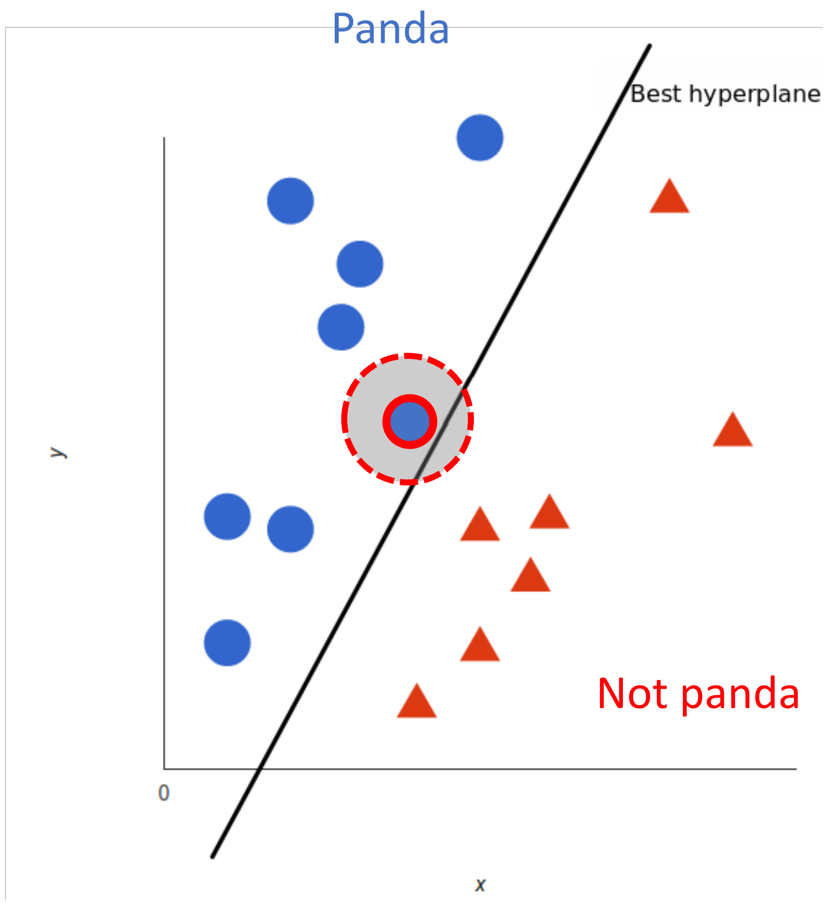
Learning with adversarial data

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



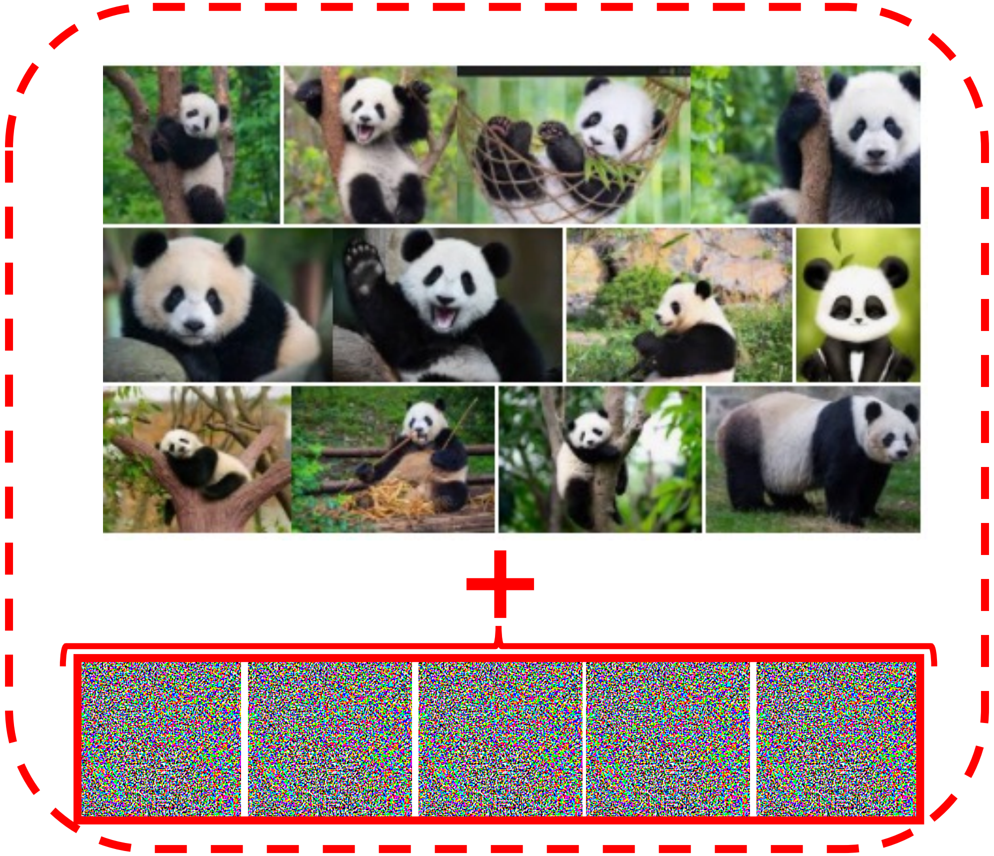
Small modification (one pixel)

ML model for panda recognition



Learning with adversarial data

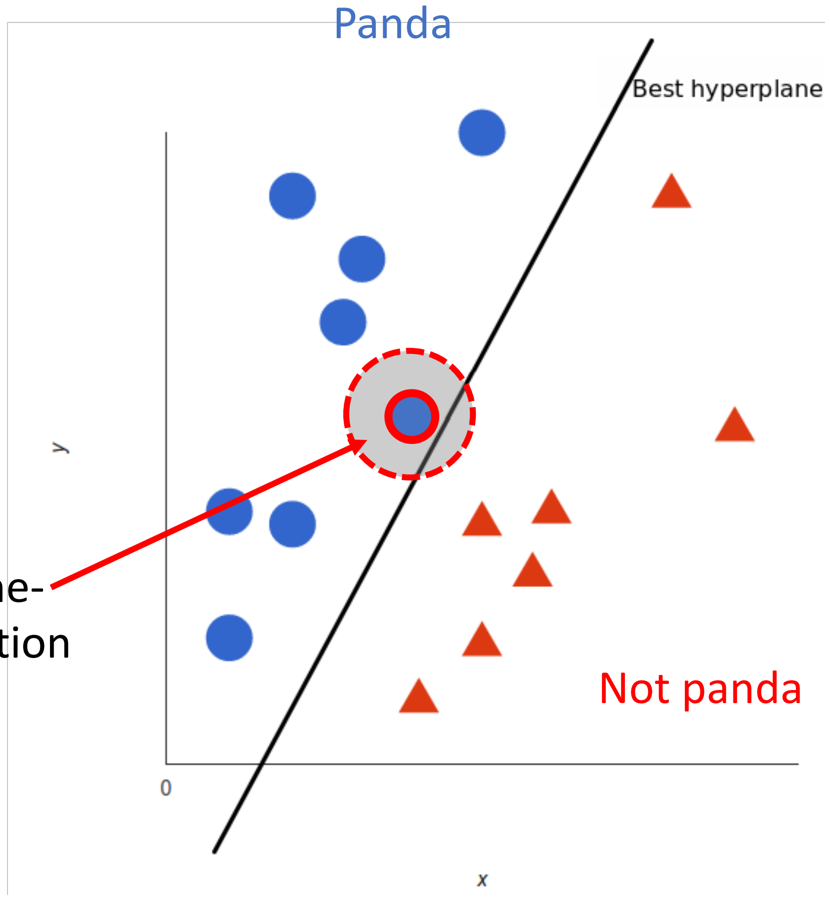
Q: can we build a new training set that includes adversarial data?



Small modification (one pixel)

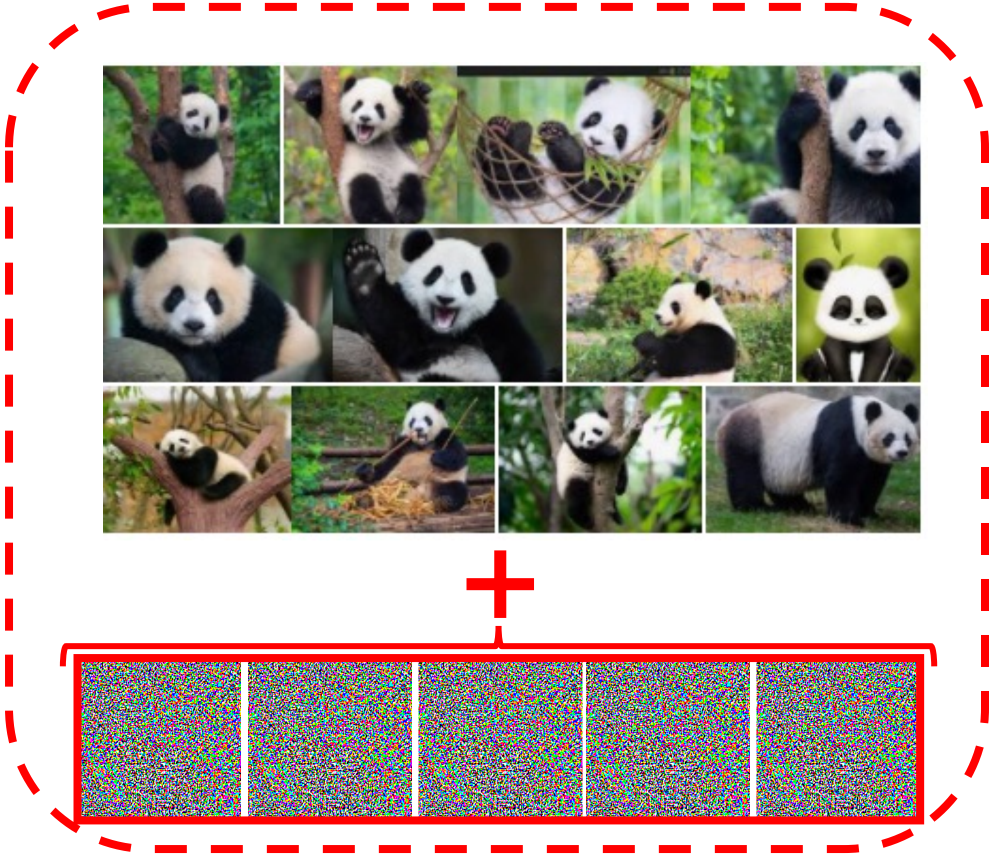
ML model for panda recognition

All possible one-pixel modification



Learning with adversarial data

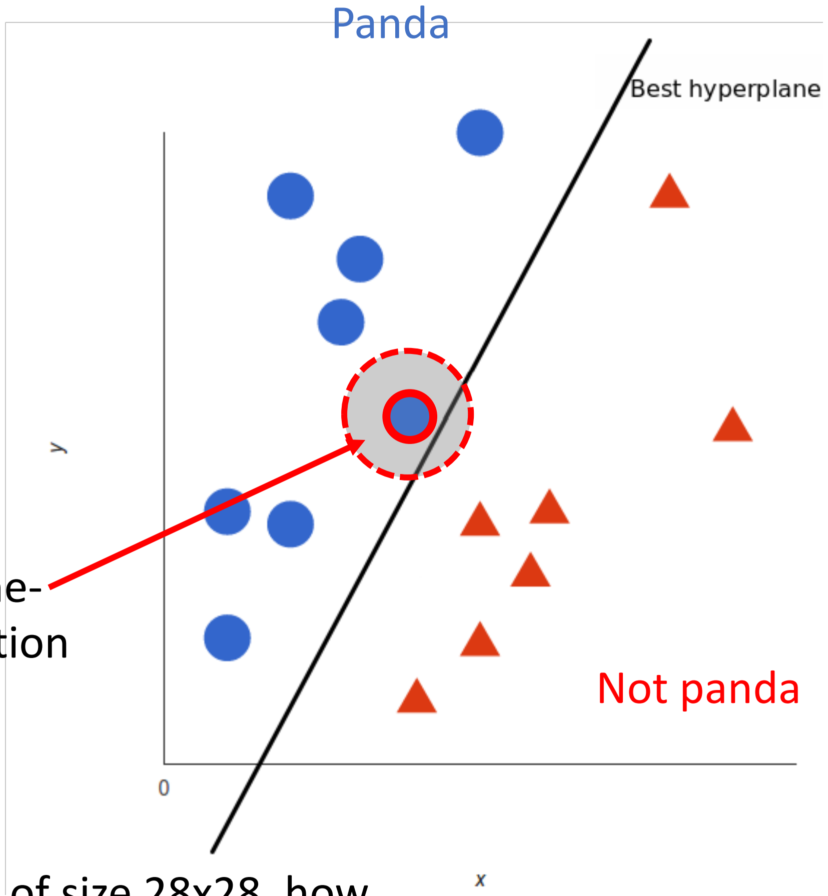
Q: can we build a new training set that includes adversarial data?



Small modification (one pixel)

ML model for panda recognition

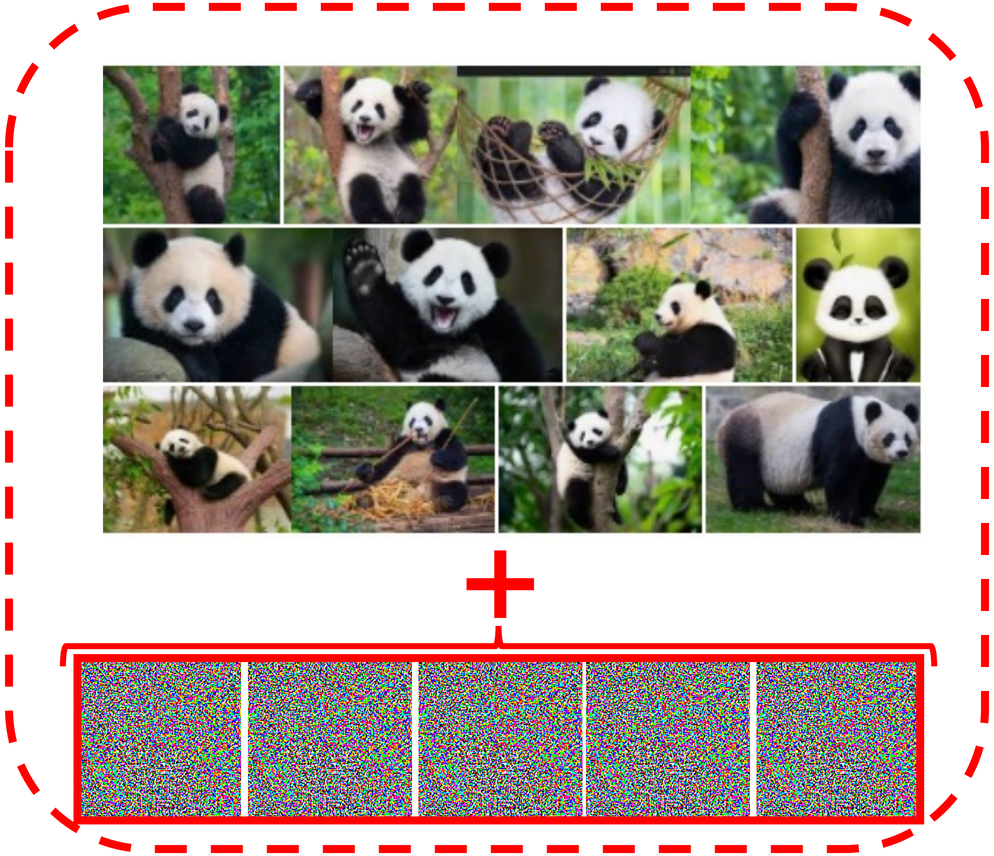
All possible one-pixel modification



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

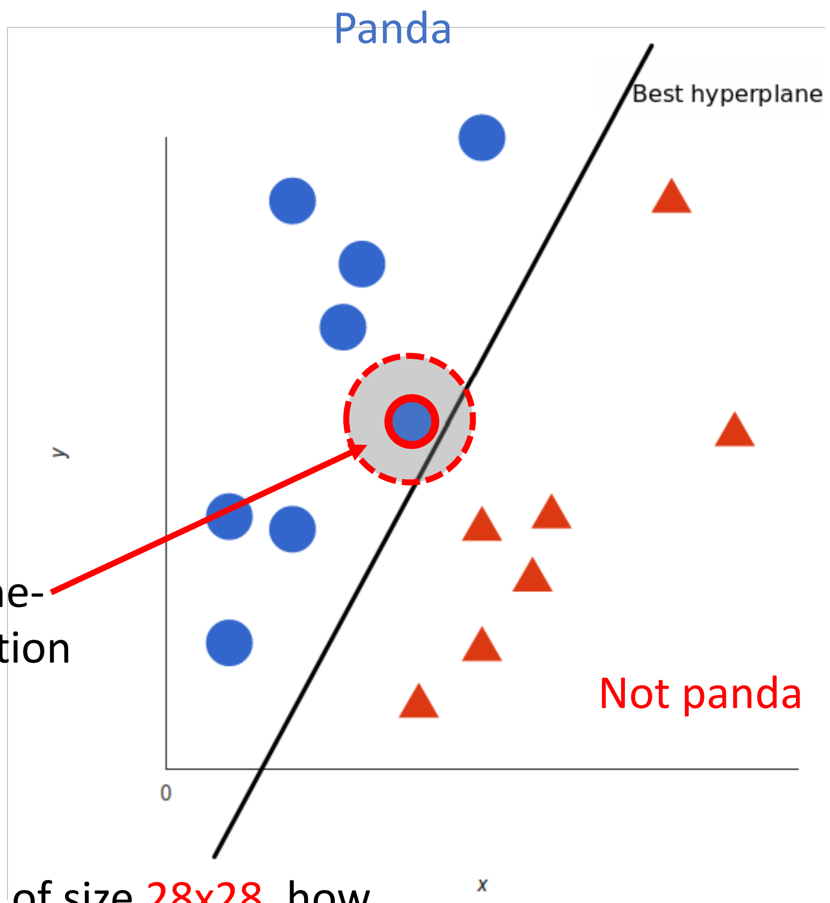
Learning with adversarial data

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



Small modification (one pixel)

ML model for panda recognition



All possible one-pixel modification

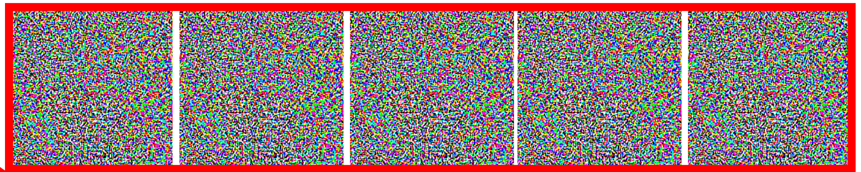
0~255

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?

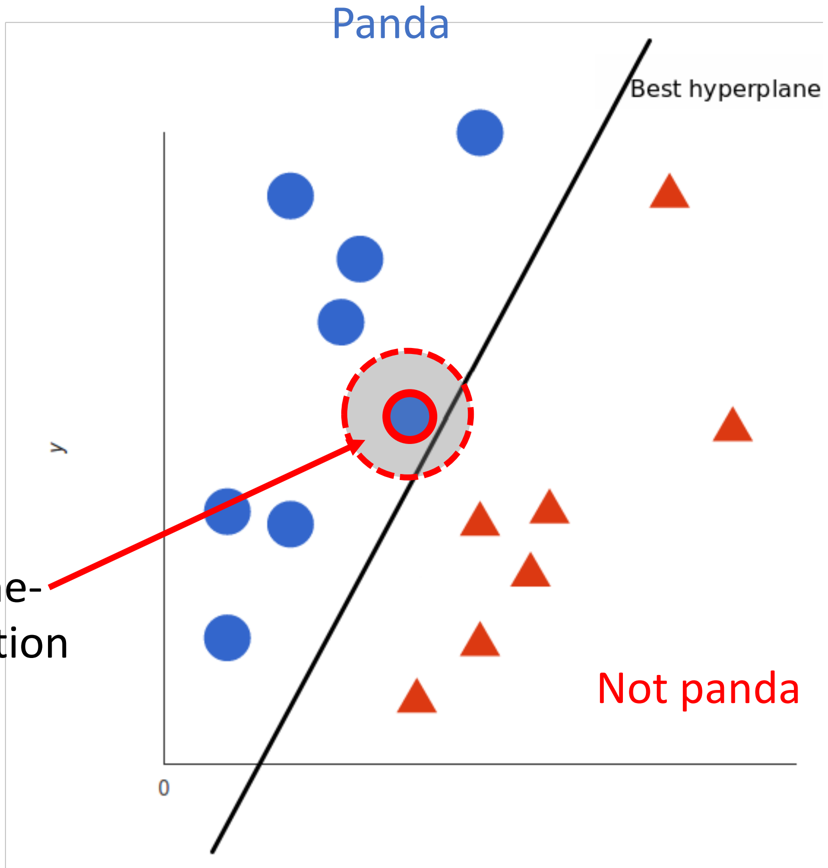
+



Small modification (one pixel)

ML model for panda recognition

All possible one-pixel modification



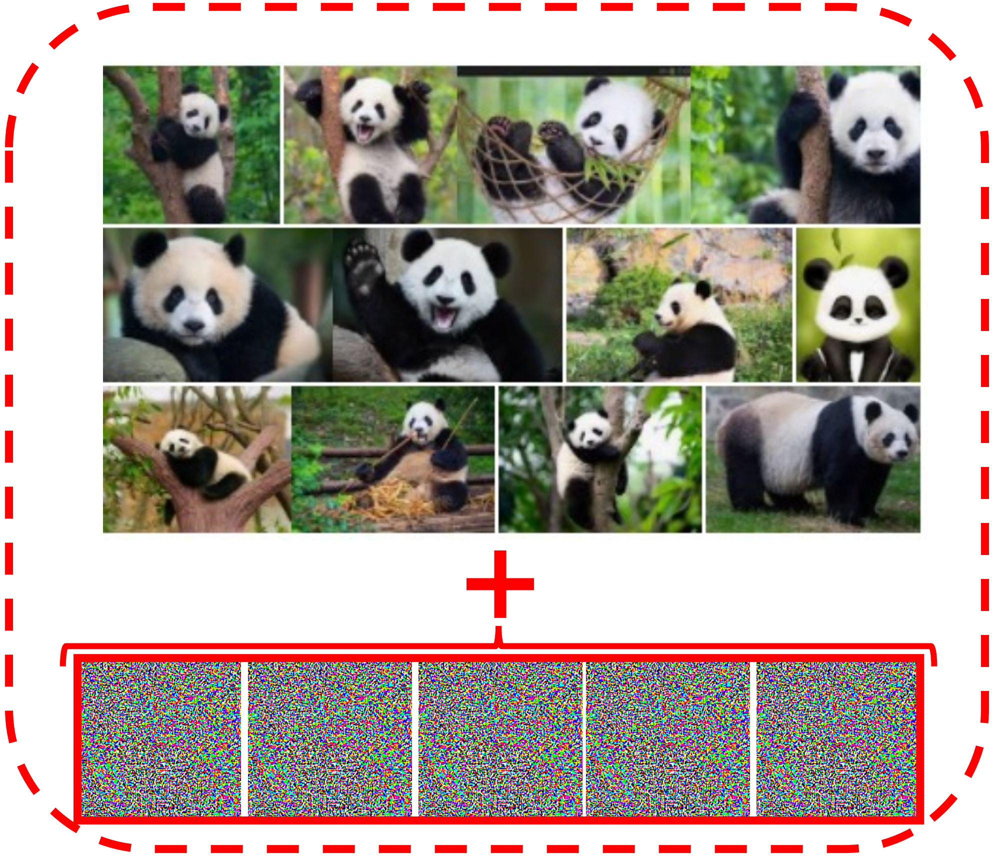
0~255

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

255 x 28 x 28
= 199920

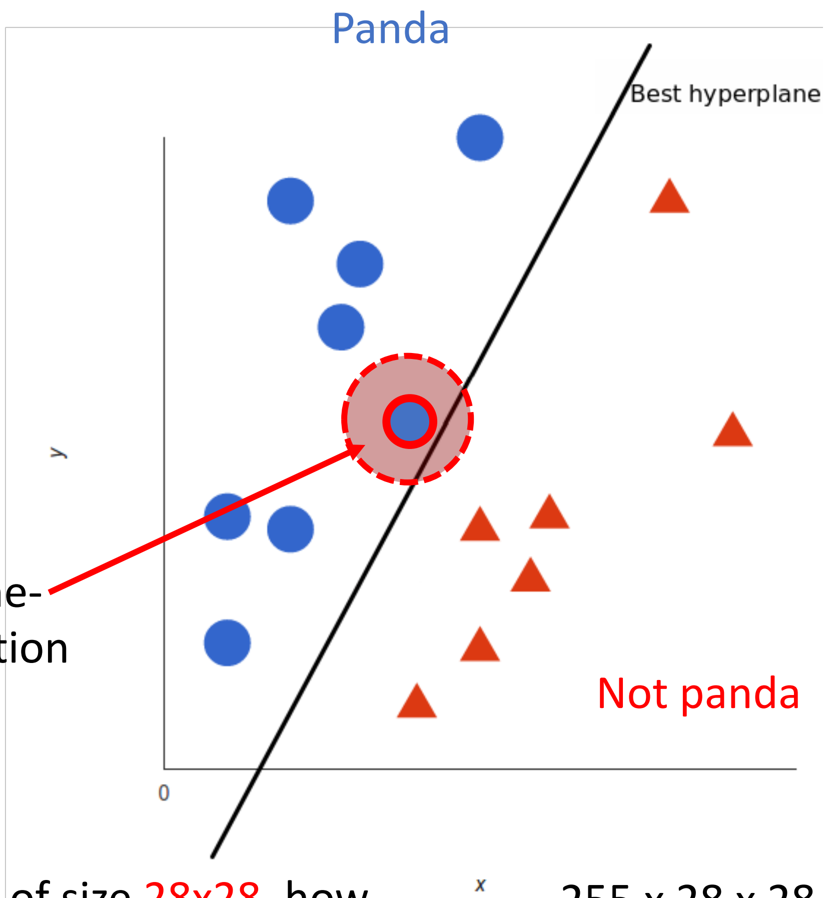
Learning with adversarial data

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



Small modification (one pixel)

ML model for
panda recognition



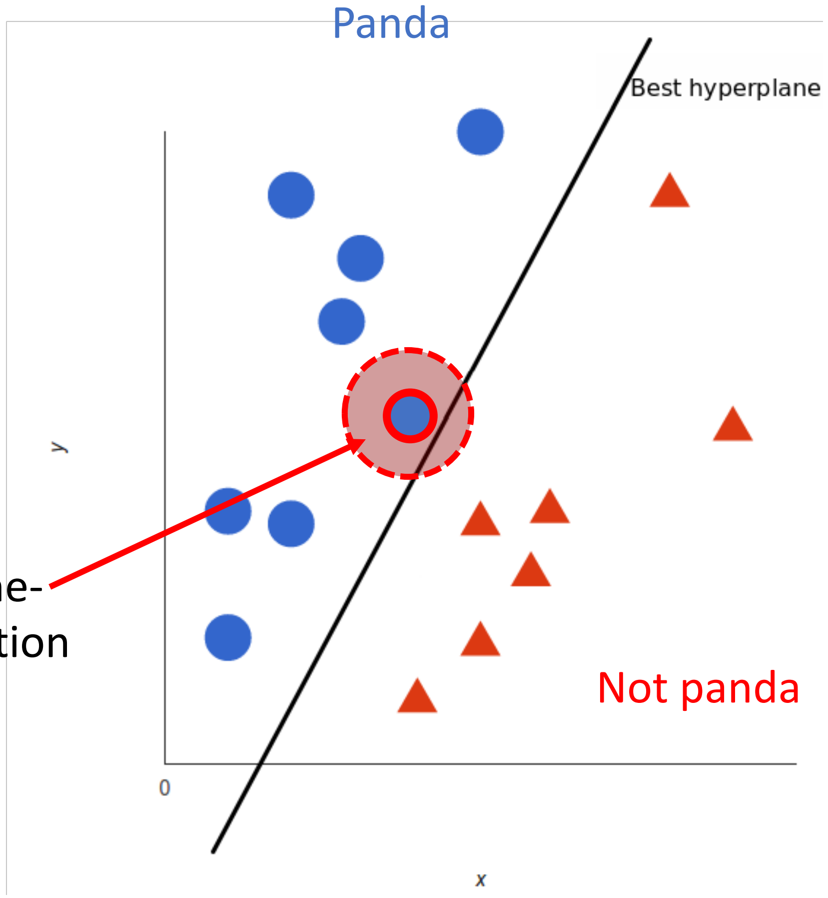
All possible one-pixel modification

0~255

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

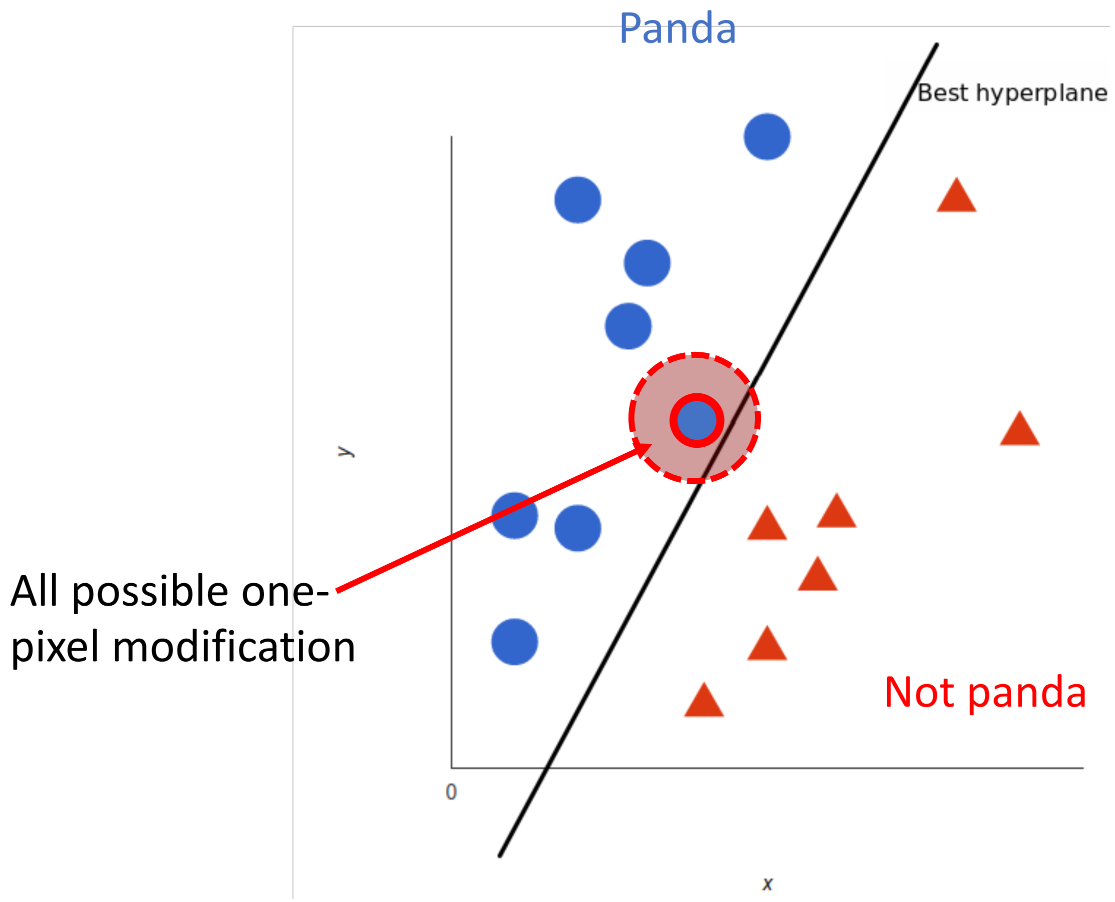
255 x 28 x 28
= 199920

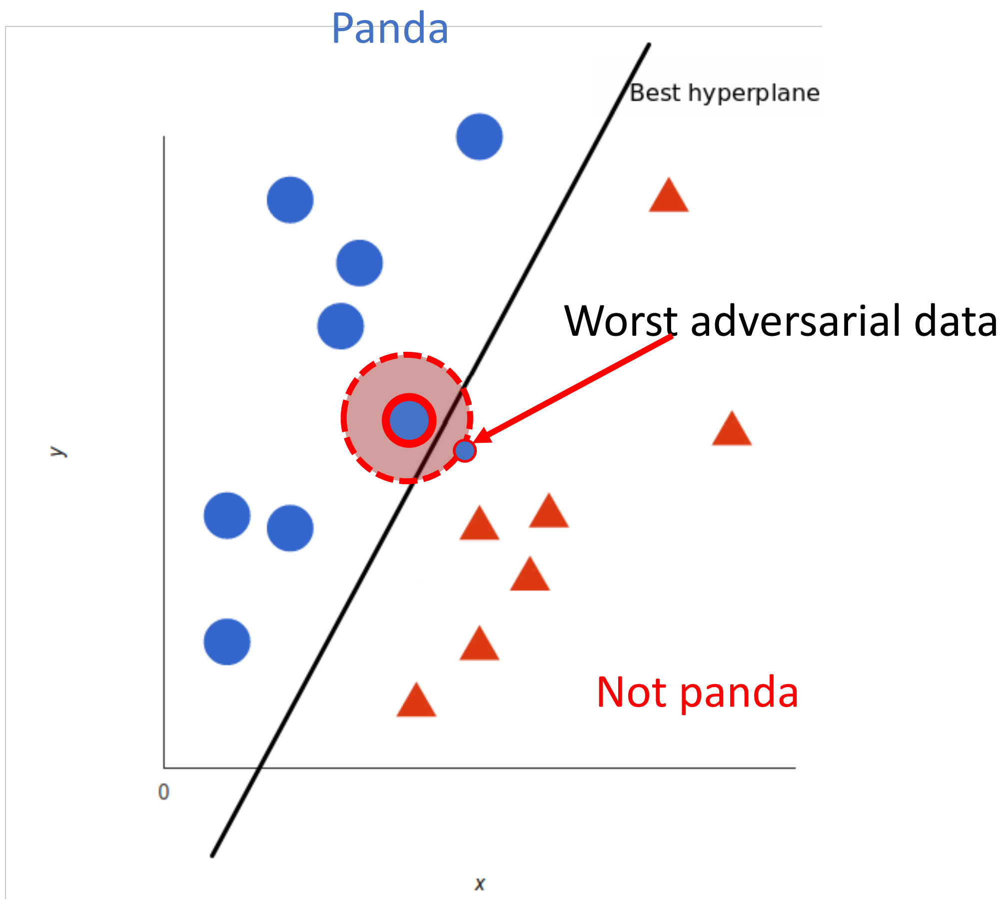
Worst case minimization



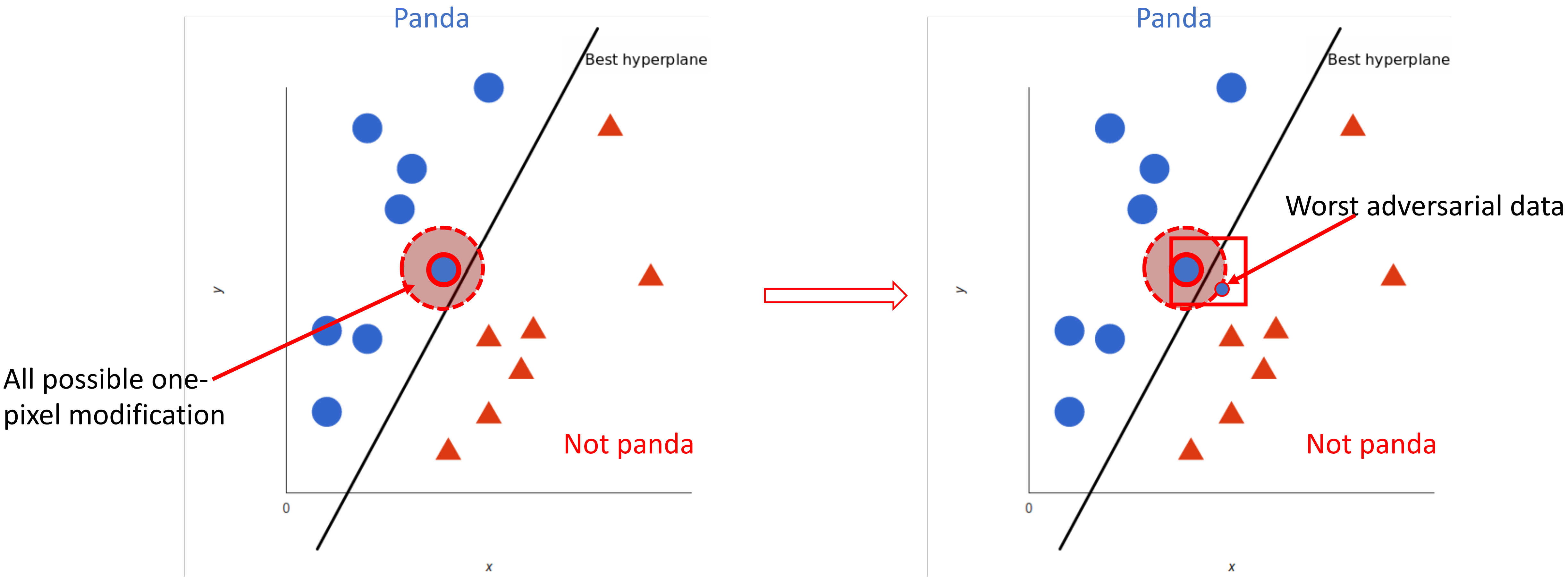
All possible one-pixel modification

Worst case minimization

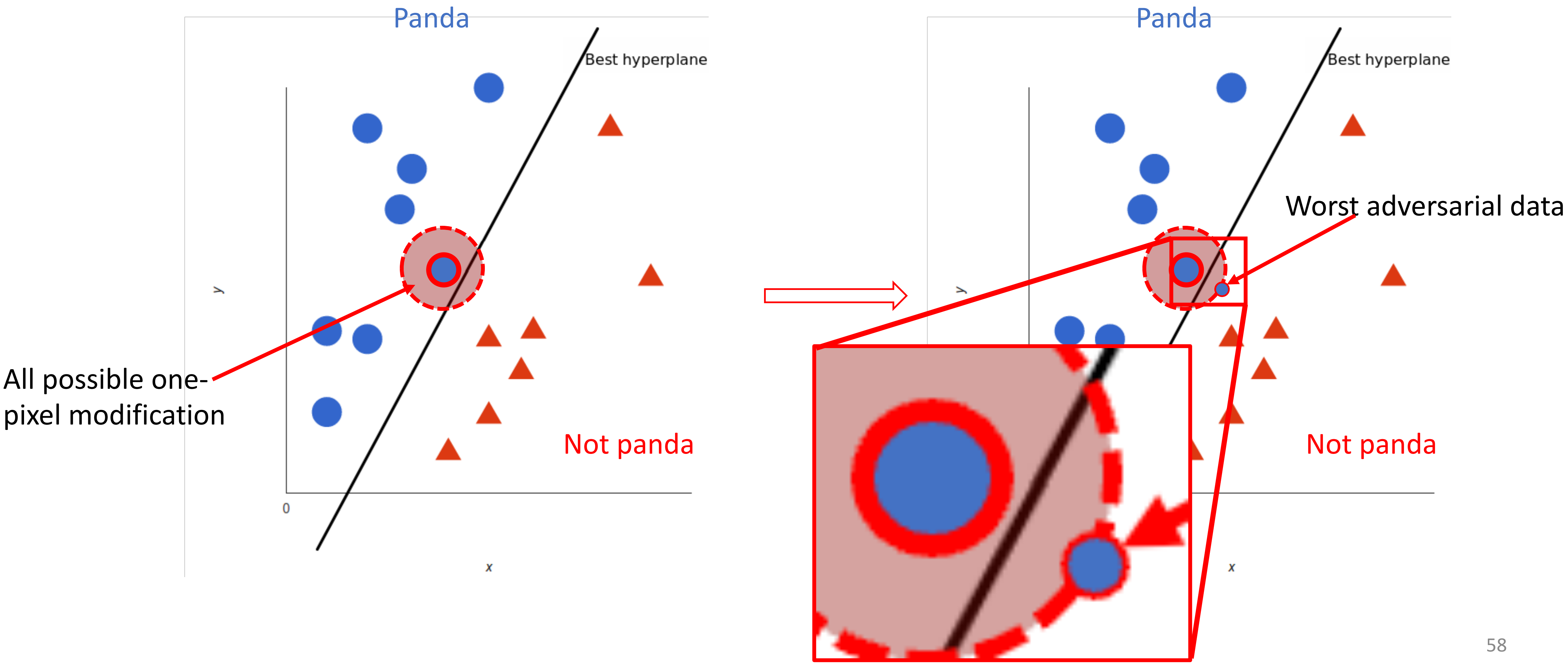




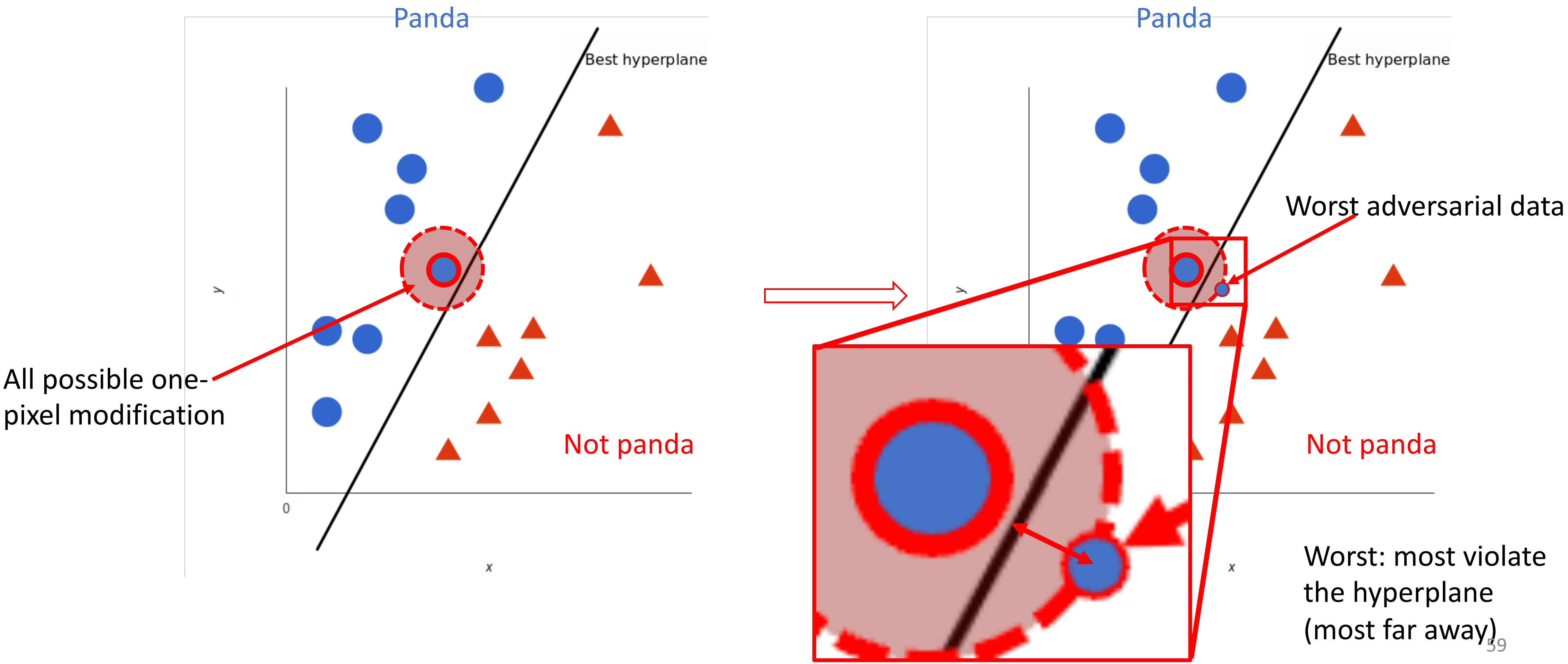
Worst case minimization



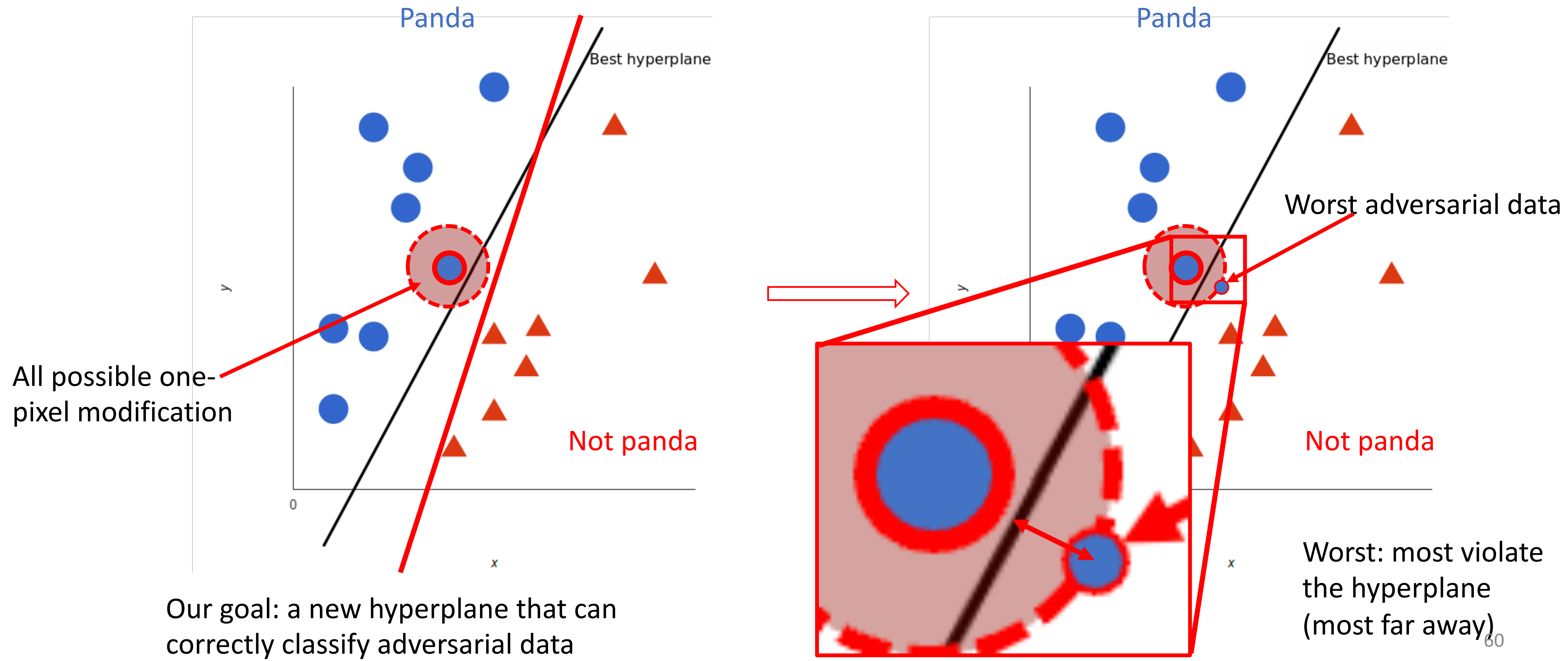
Worst case minimization



Worst case minimization



Worst case minimization



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

Worst: most violate the hyperplane (most far away)