Adversarial Domain Adaptation and Adversarial Robustness

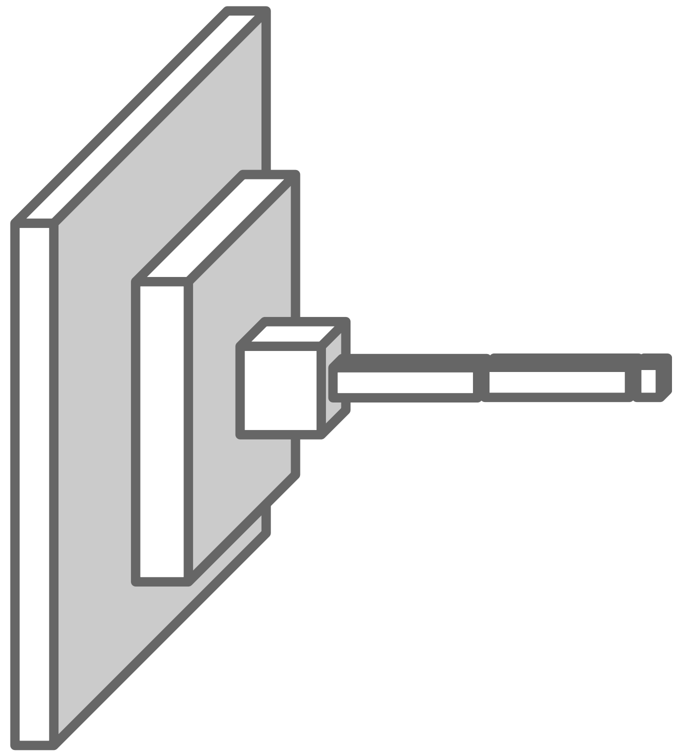
Judy Hoffman

facebook Artificial Intelligence Research

Georgia Institute of Technology

Big data

+

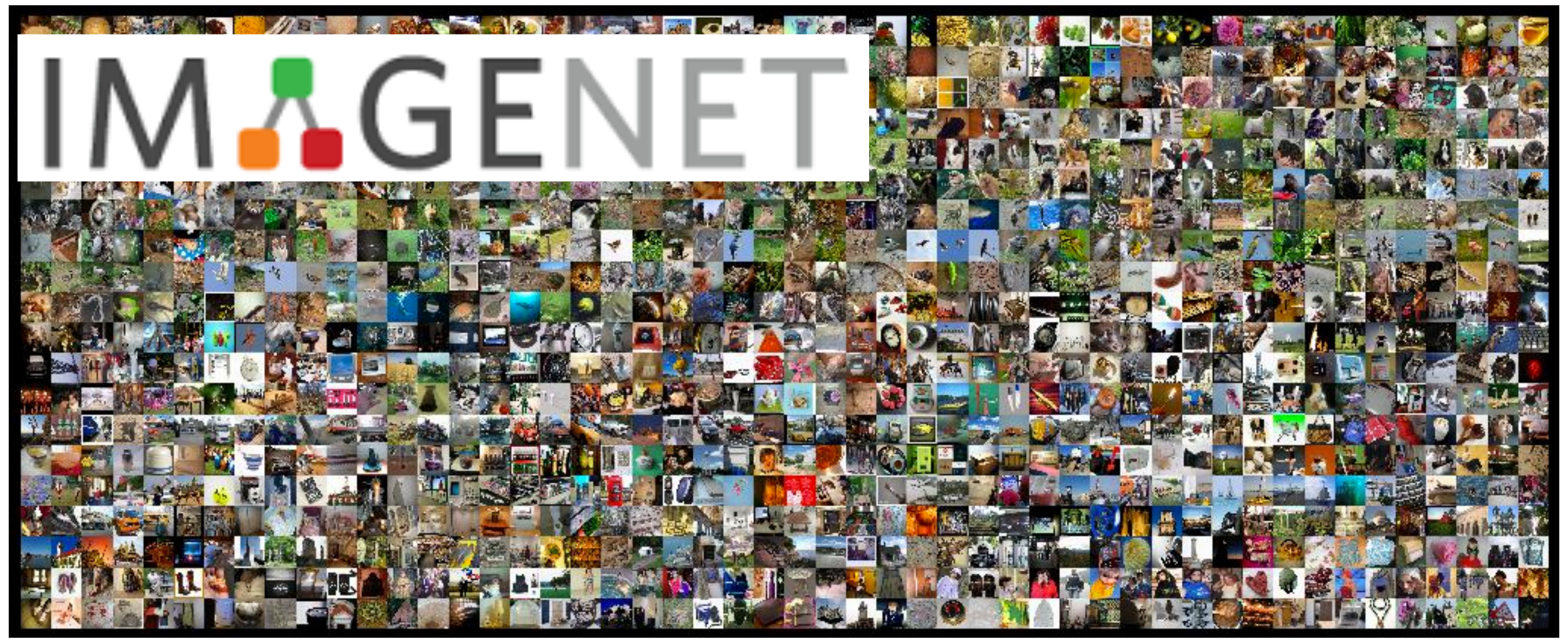


Deep learning

=

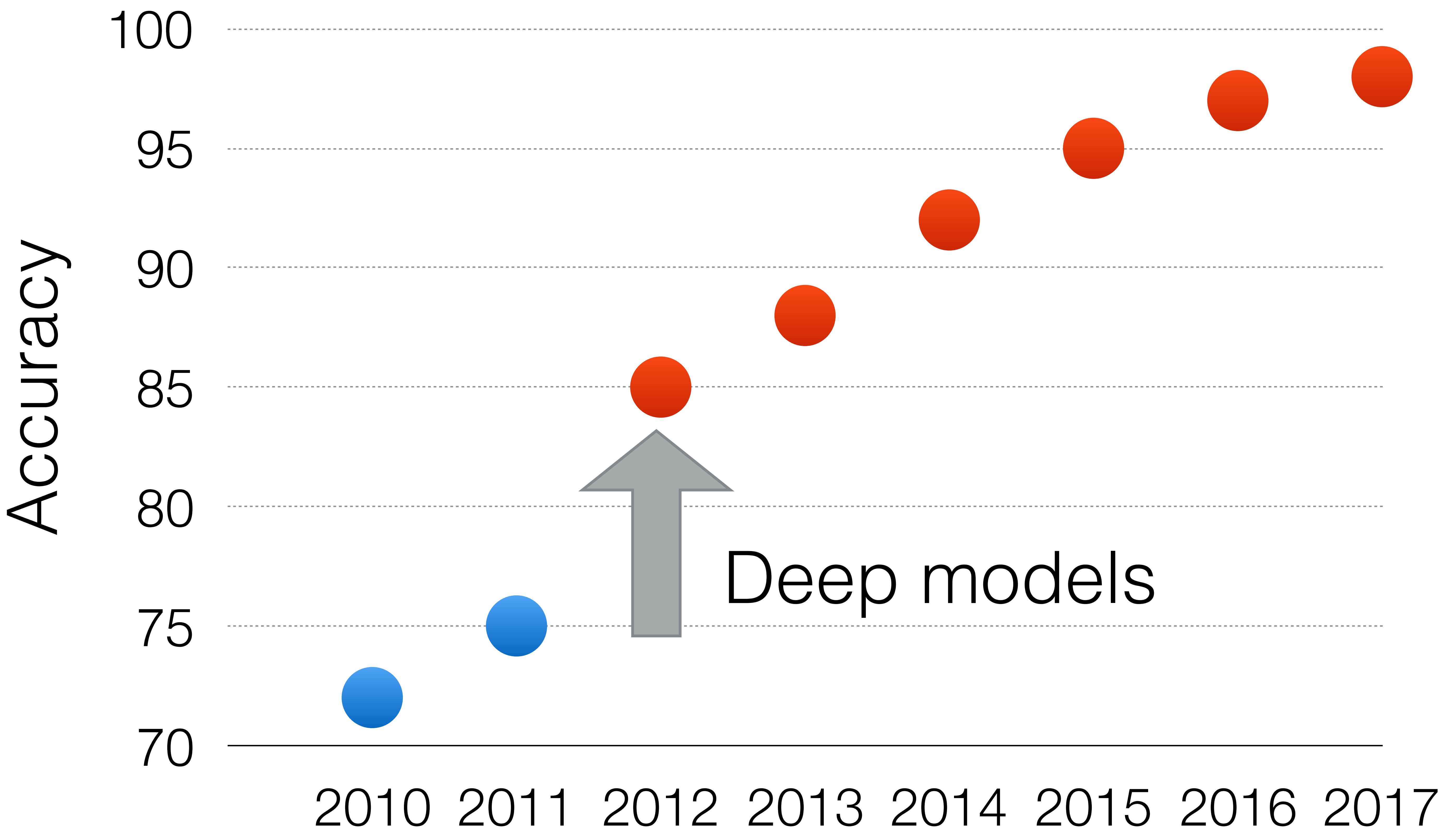
success

Benchmark Performance

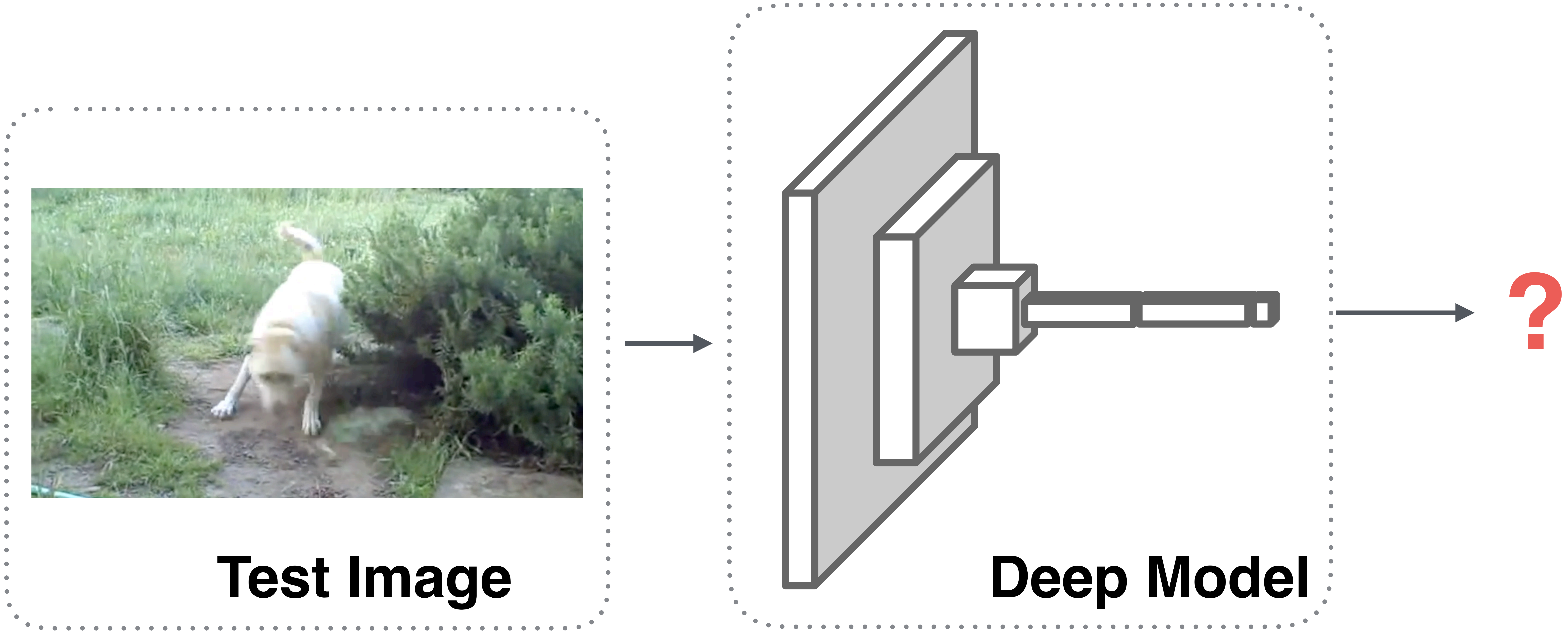


Millions of Images

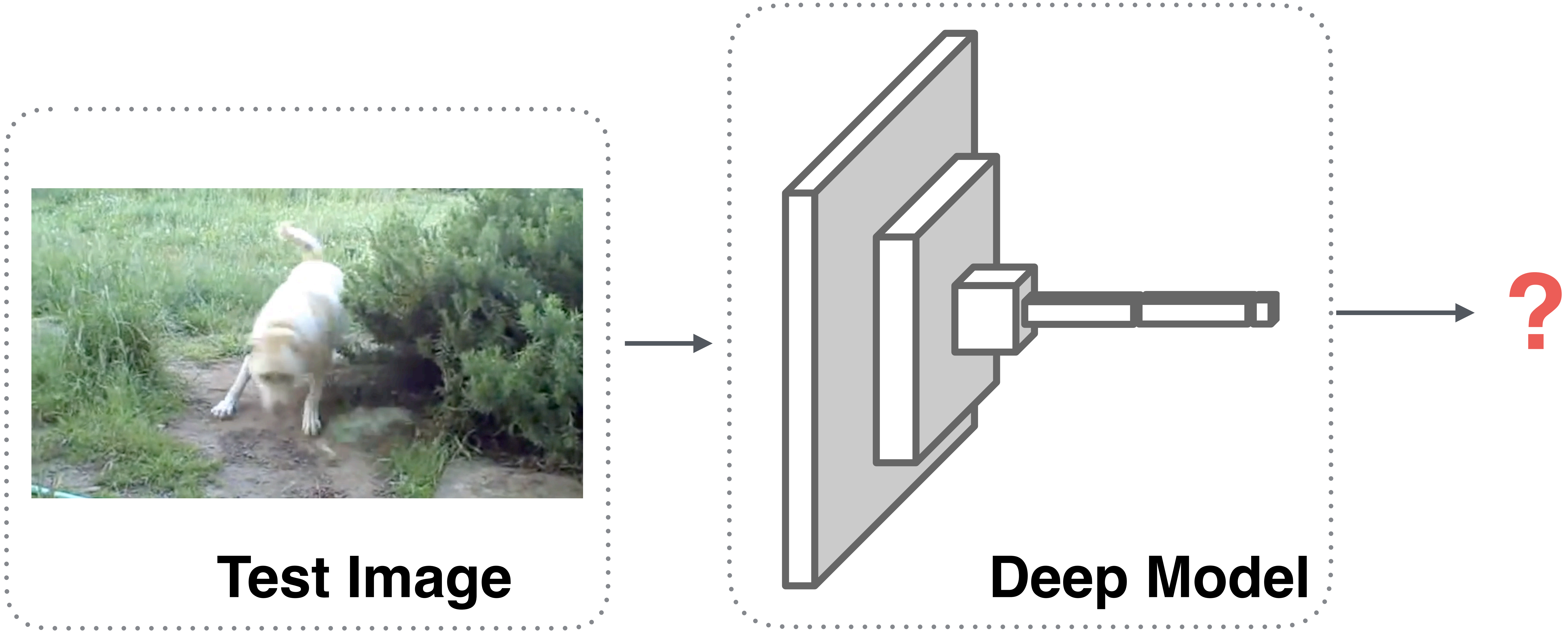
Challenge to recognize 1000 categories



Dataset Bias



Dataset Bias



Dataset Bias

Dataset Bias

IMAGENET

Dataset Bias

Low resolution

Dataset Bias

Low resolution

Motion Blur

Dataset Bias

Low resolution

Motion Blur

Pose Variety

Why not collect new annotations?

Why not collect new annotations?

Car
Road
Sidewalk
Person
Sky
Vegetation
Street Sign
Building

Why not collect new annotations?

Car
Road
Sidewalk
Person
Sky
Vegetation
Street Sign
Building

Expensive
($10-12 per image)

Why not collect new annotations?

Large Potential for Change
Different: Weather, City, Car

Car
Road
Sidewalk
Person
Sky
Vegetation
Street Sign
Building

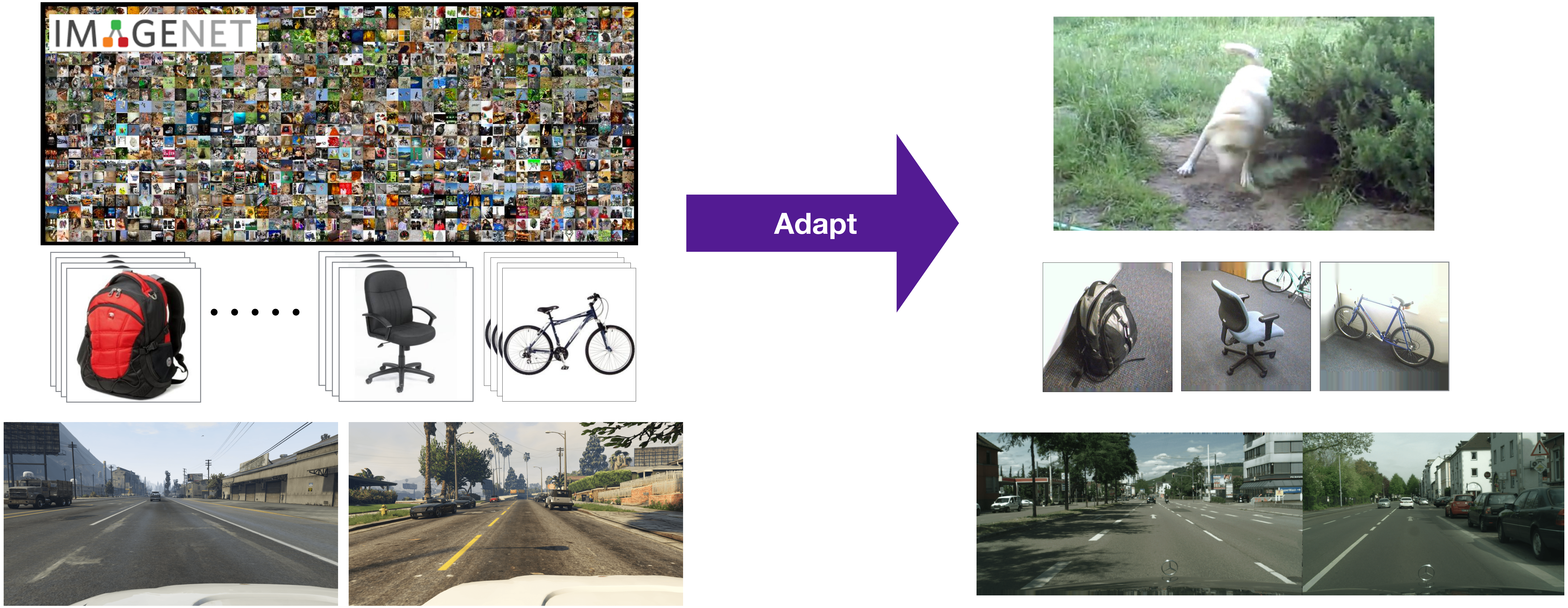
Expensive
($10-12 per image)

Why not collect new annotations?

Proprietary

Private

Domain Adaptation: Train on Source Test on Target



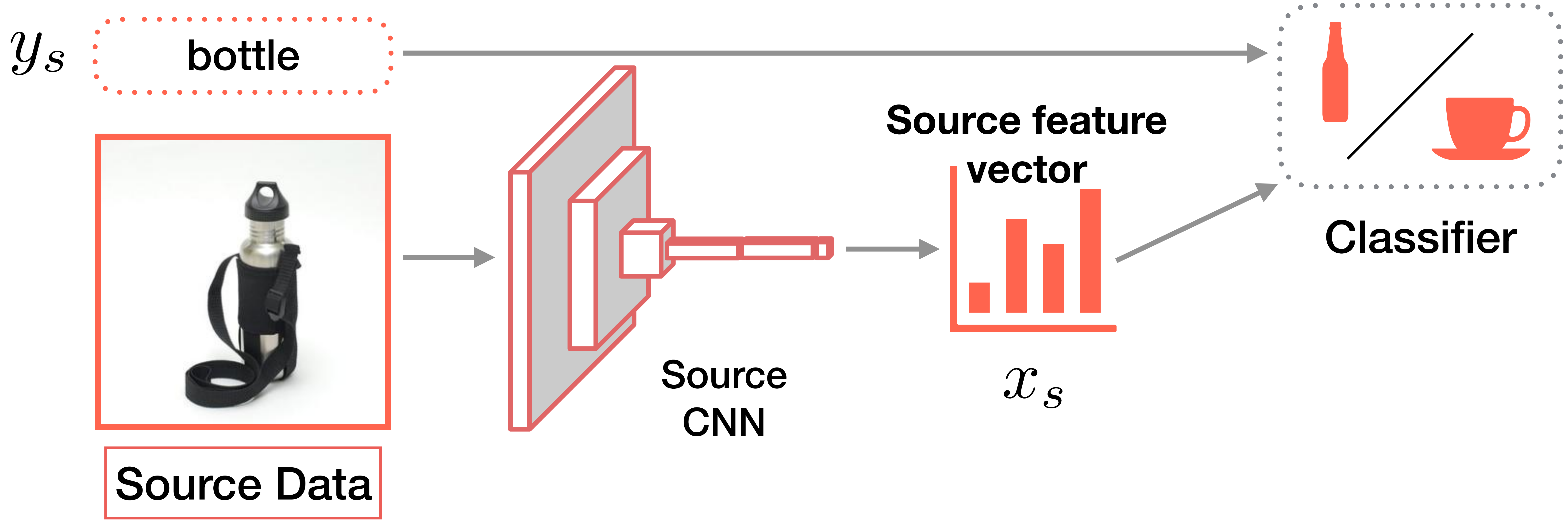
Source Domain $\sim P_S(X_S, Y_S)$

lots of **labeled** data

Target Domain $\sim P_T(X_T, Y_T)$

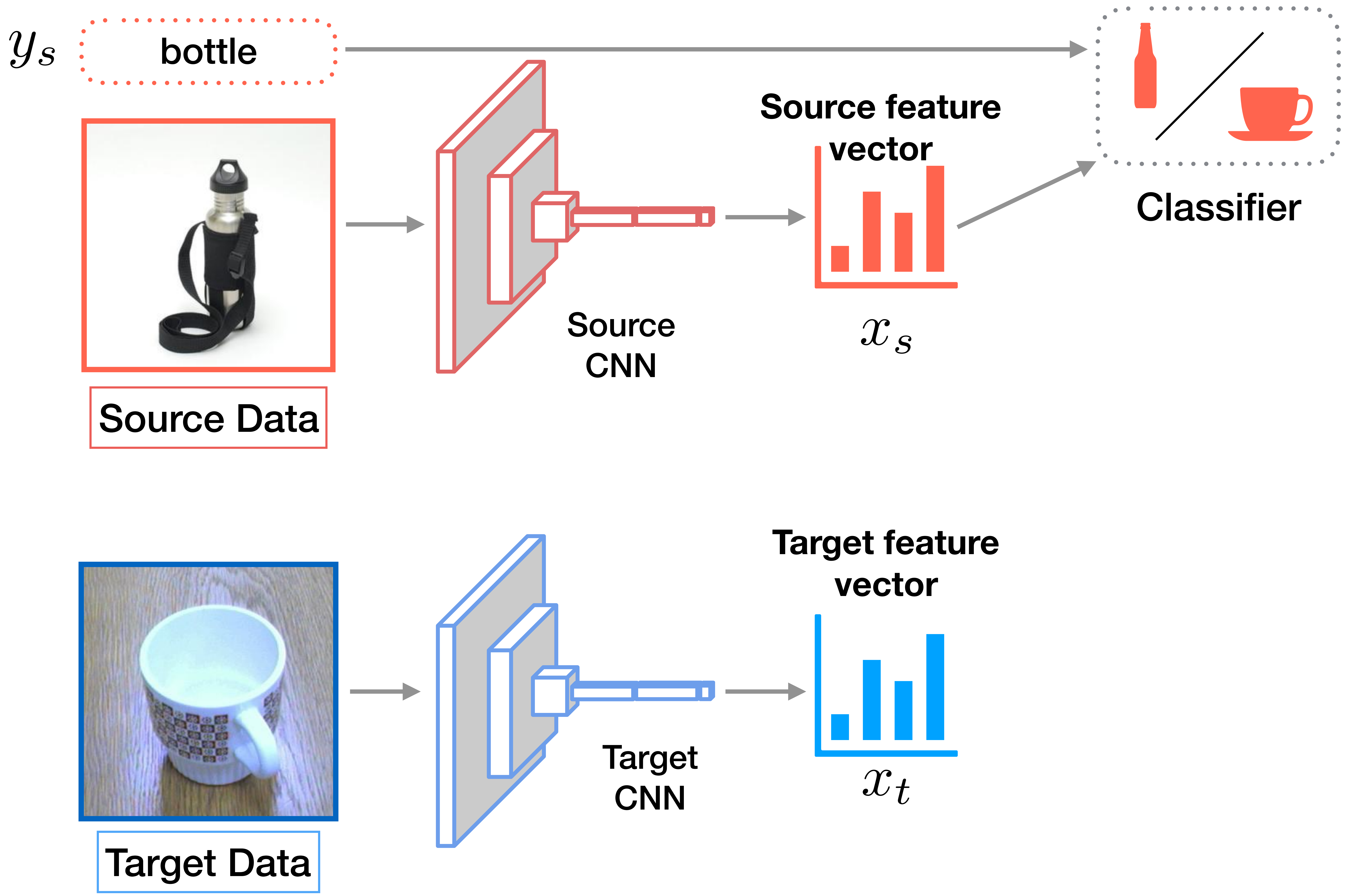
unlabeled or limited labels

Adversarial Domain Adaptation



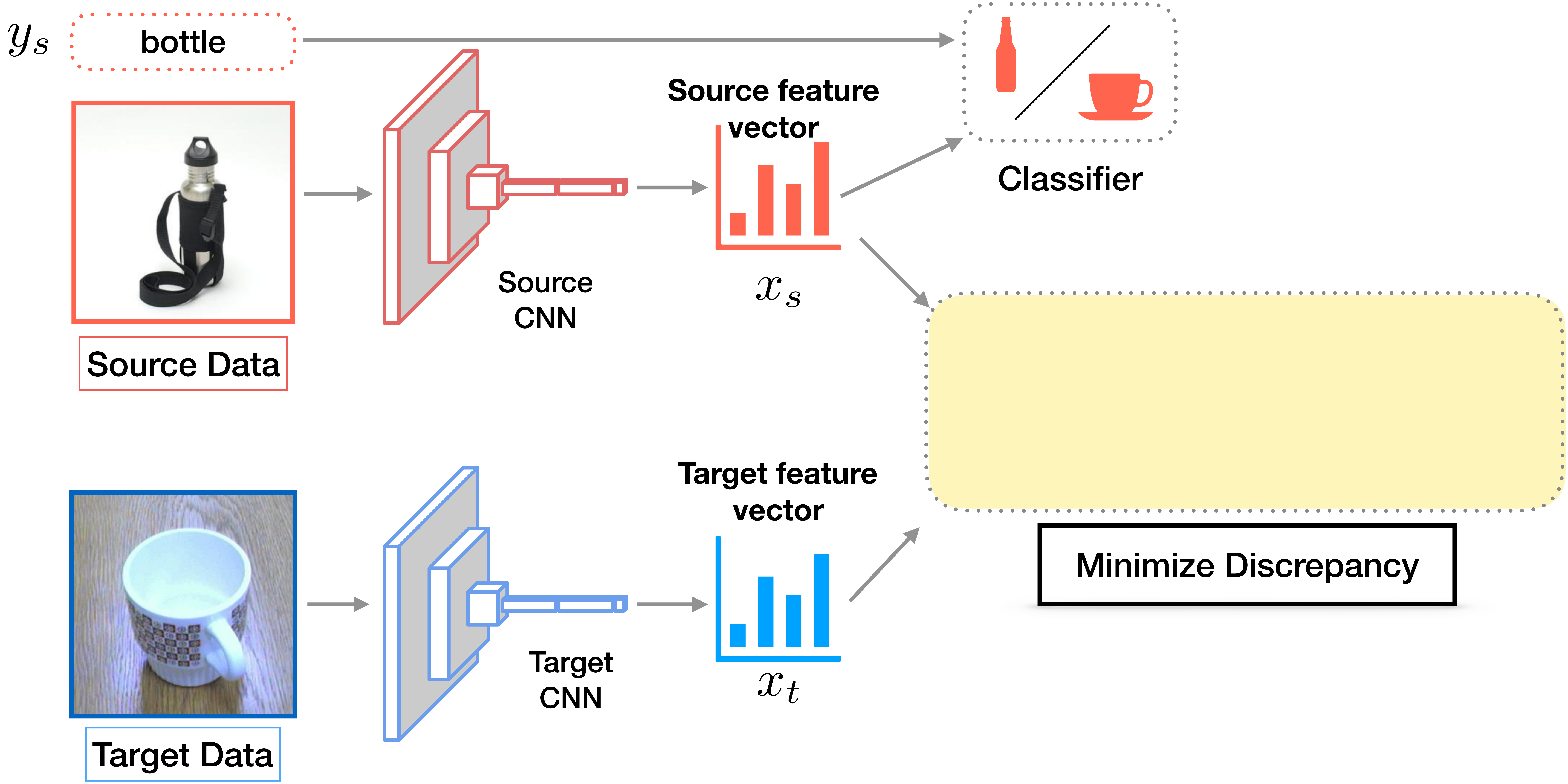
Ganin & Lempinsky, ICML 2015. Tzeng*, Hoffman*, Saenko, Darrell, ICCV 2015. Tzeng, Hoffman, Saenko, Darrell. *CVPR* 2017.

Adversarial Domain Adaptation



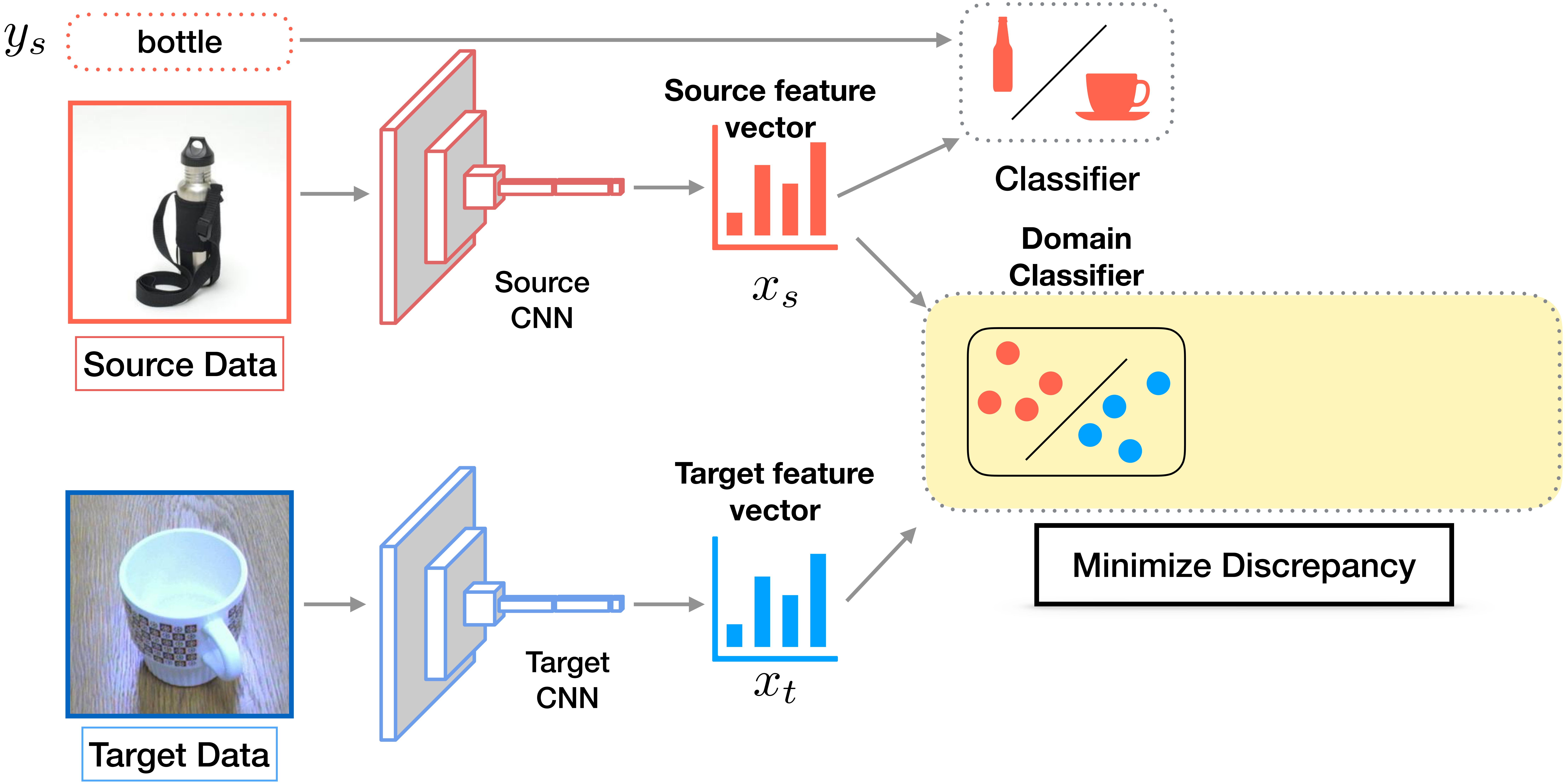
Ganin & Lempinsky, ICML 2015. Tzeng*, Hoffman*, Saenko, Darrell, ICCV 2015. Tzeng, Hoffman, Saenko, Darrell. *CVPR* 2017.

Adversarial Domain Adaptation



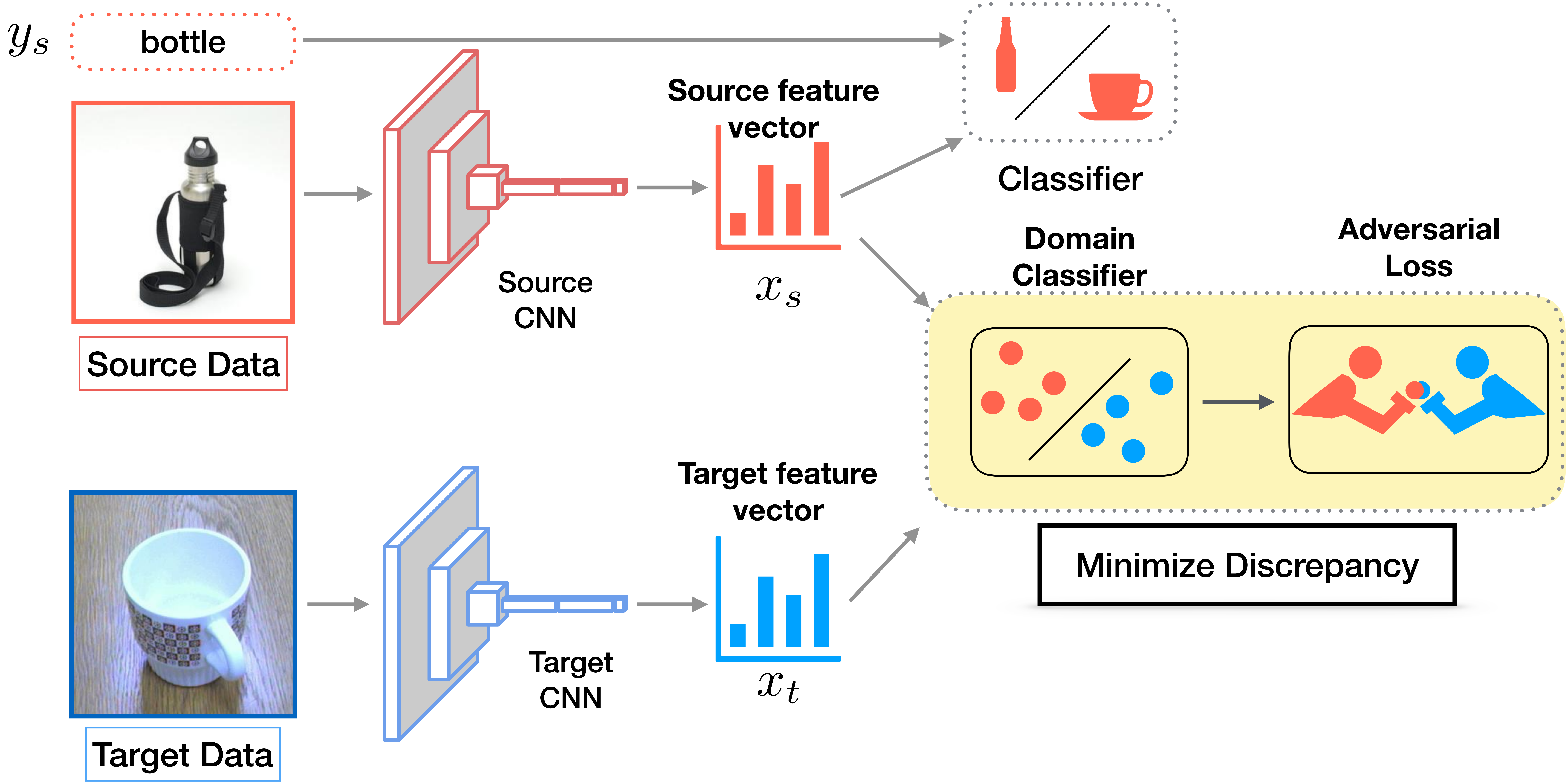
Ganin & Lempinsky, ICML 2015. Tzeng*, Hoffman*, Saenko, Darrell, ICCV 2015. Tzeng, Hoffman, Saenko, Darrell. *CVPR* 2017.

Adversarial Domain Adaptation



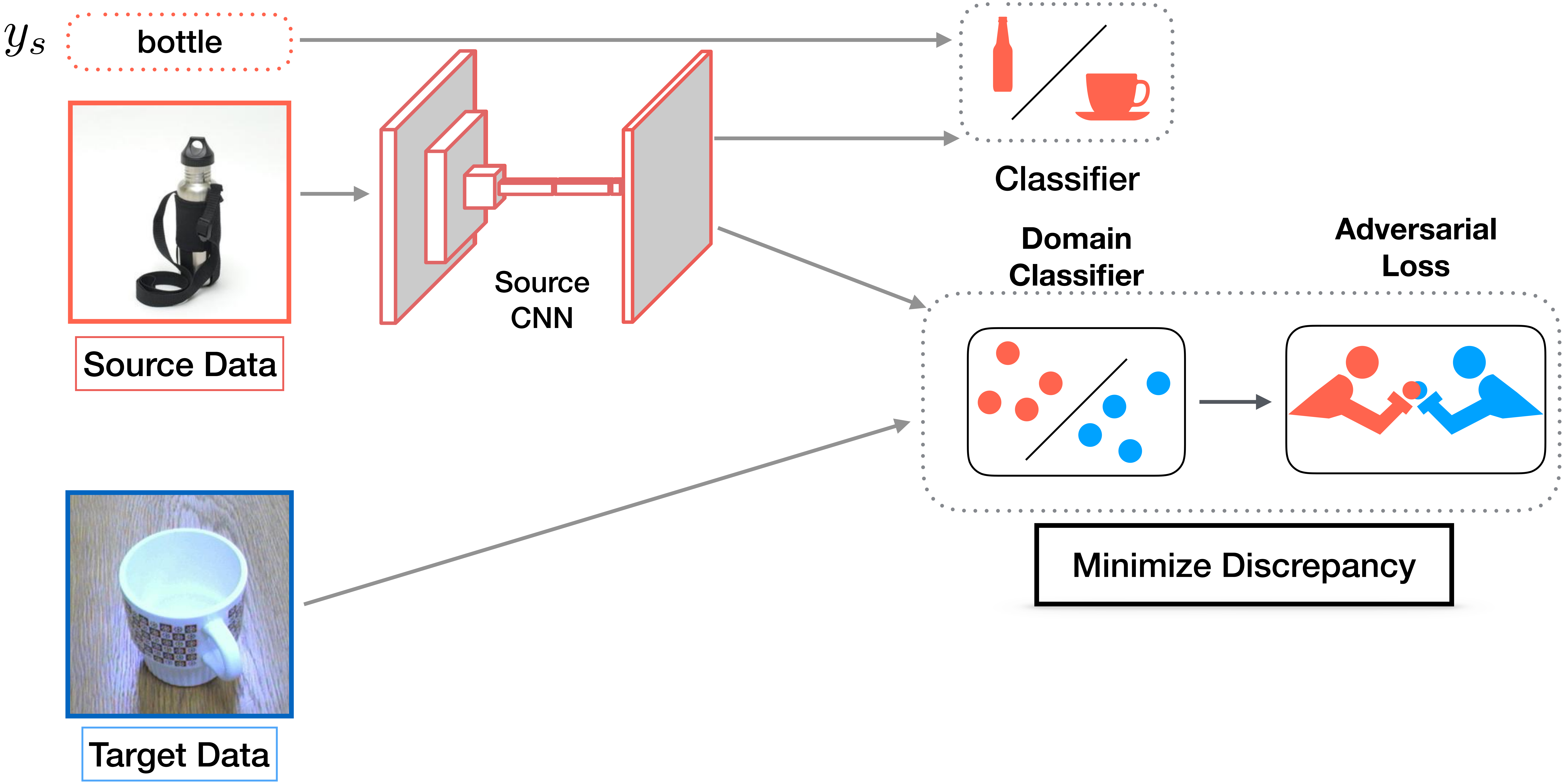
Ganin & Lempinsky, ICML 2015. Tzeng*, Hoffman*, Saenko, Darrell, ICCV 2015. Tzeng, Hoffman, Saenko, Darrell. *CVPR* 2017.

Adversarial Domain Adaptation



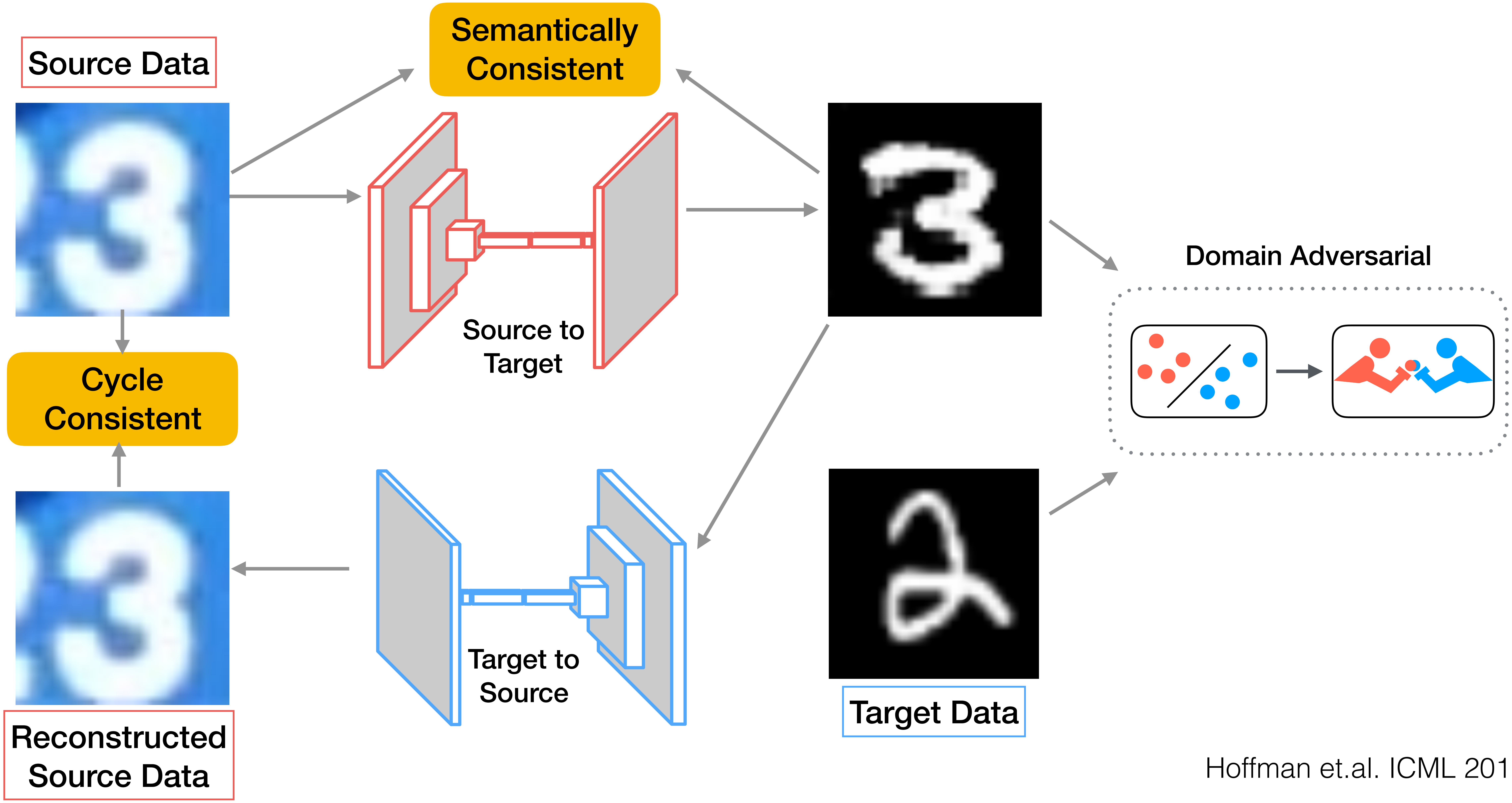
Ganin & Lempinsky, ICML 2015. Tzeng*, Hoffman*, Saenko, Darrell, ICCV 2015. Tzeng, Hoffman, Saenko, Darrell. *CVPR* 2017.

Adversarial Domain Adaptation



Liu 2016. Taigman 2016. Bousmalis 2017. Liu 2017. Kim 2017. Sankaranarayanan 2018. Hoffman 2018.

CyCADA: Cycle Consistent Adversarial DA



Hoffman et.al. ICML 2018

Synthetic to Real Pixel Adaptation

Train

GTA (synthetic)

Test

CityScapes (Germany)

Hoffman et.al. ICML 2018

Synthetic to Real Pixel Adaptation

Hoffman et.al. ICML 2018

Synthetic to Real Pixel Adaptation

Hoffman et.al. ICML 2018

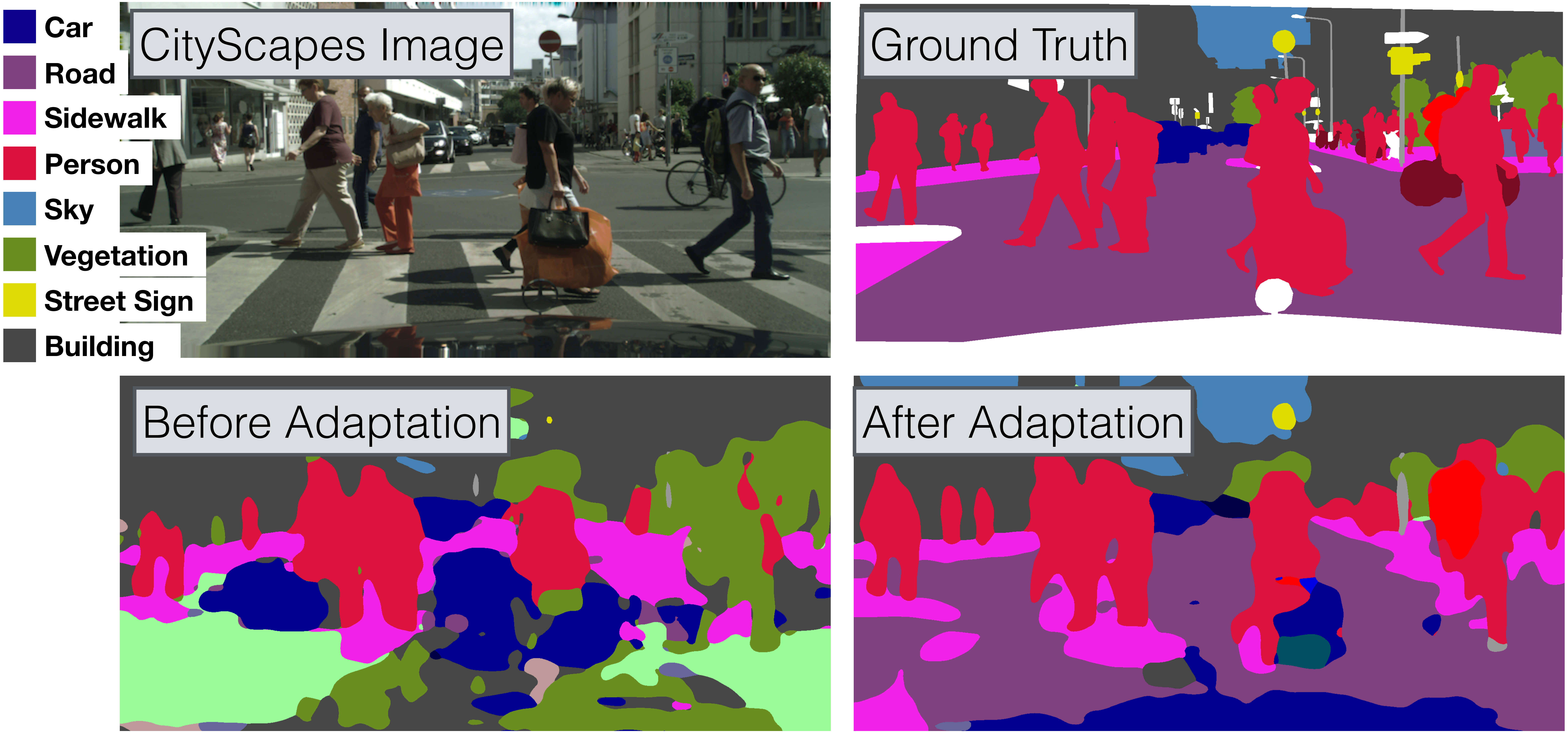
Synthetic to Real Pixel Adaptation

Zhu*, Park*, Isola, Efros. ICCV 2017.

Synthetic to Real Pixel Adaptation

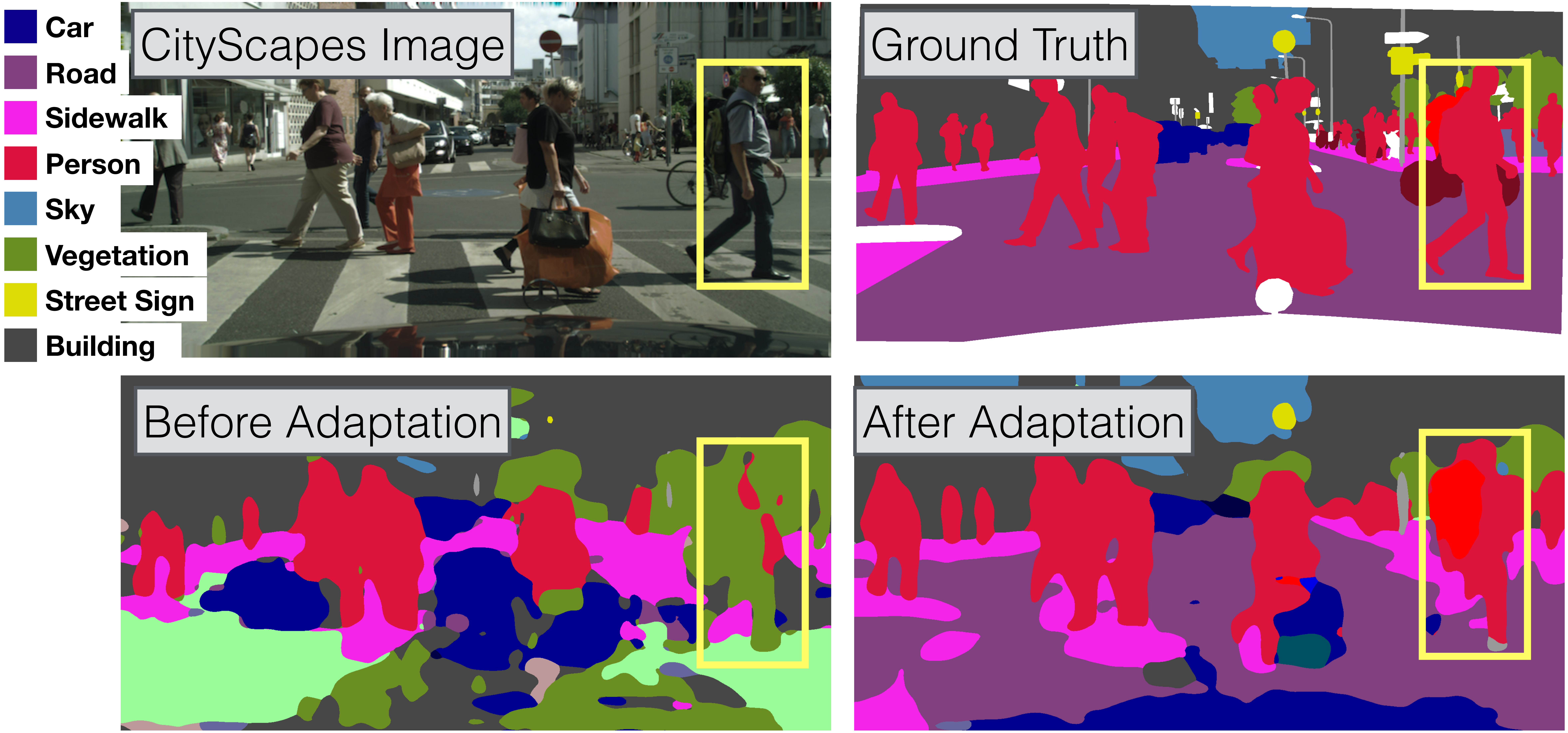
Zhu*, Park*, Isola, Efros. ICCV 2017.

CyCADA Results: CityScapes Evaluation



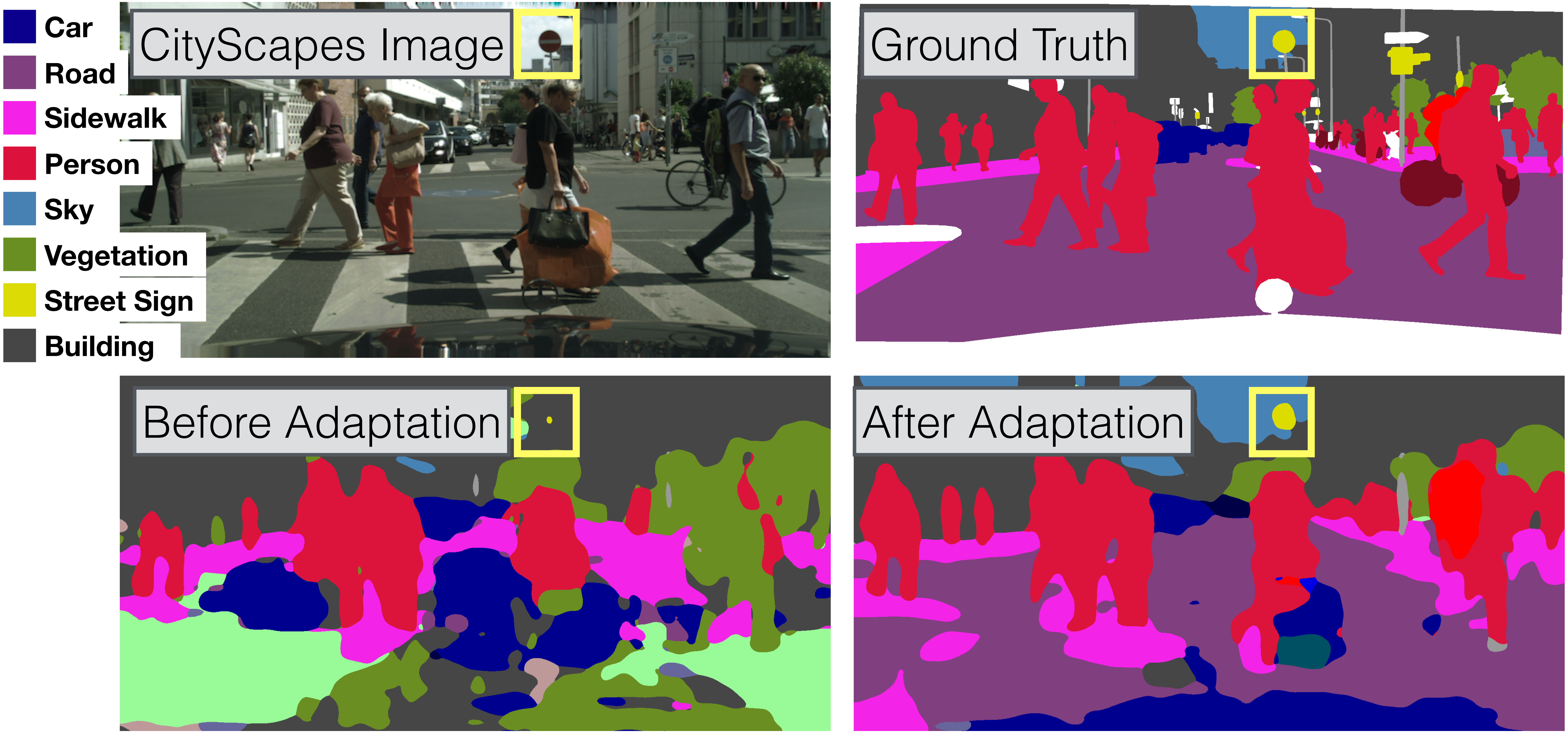
Hoffman et.al. ICML 2018

CyCADA Results: CityScapes Evaluation



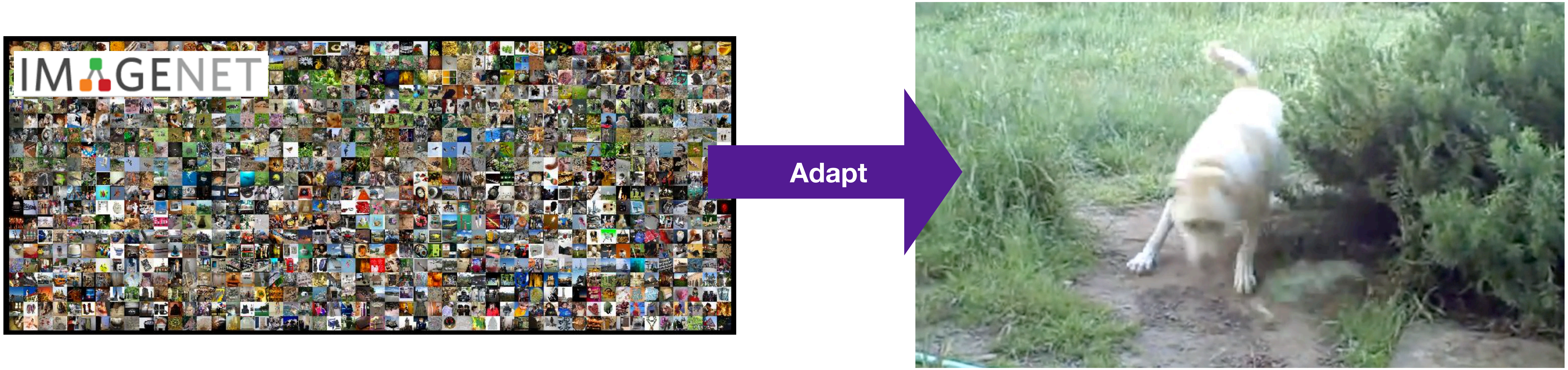
Hoffman et.al. ICML 2018

CyCADA Results: CityScapes Evaluation



Hoffman et.al. ICML 2018

So Far: Adapting to Natural Shifts



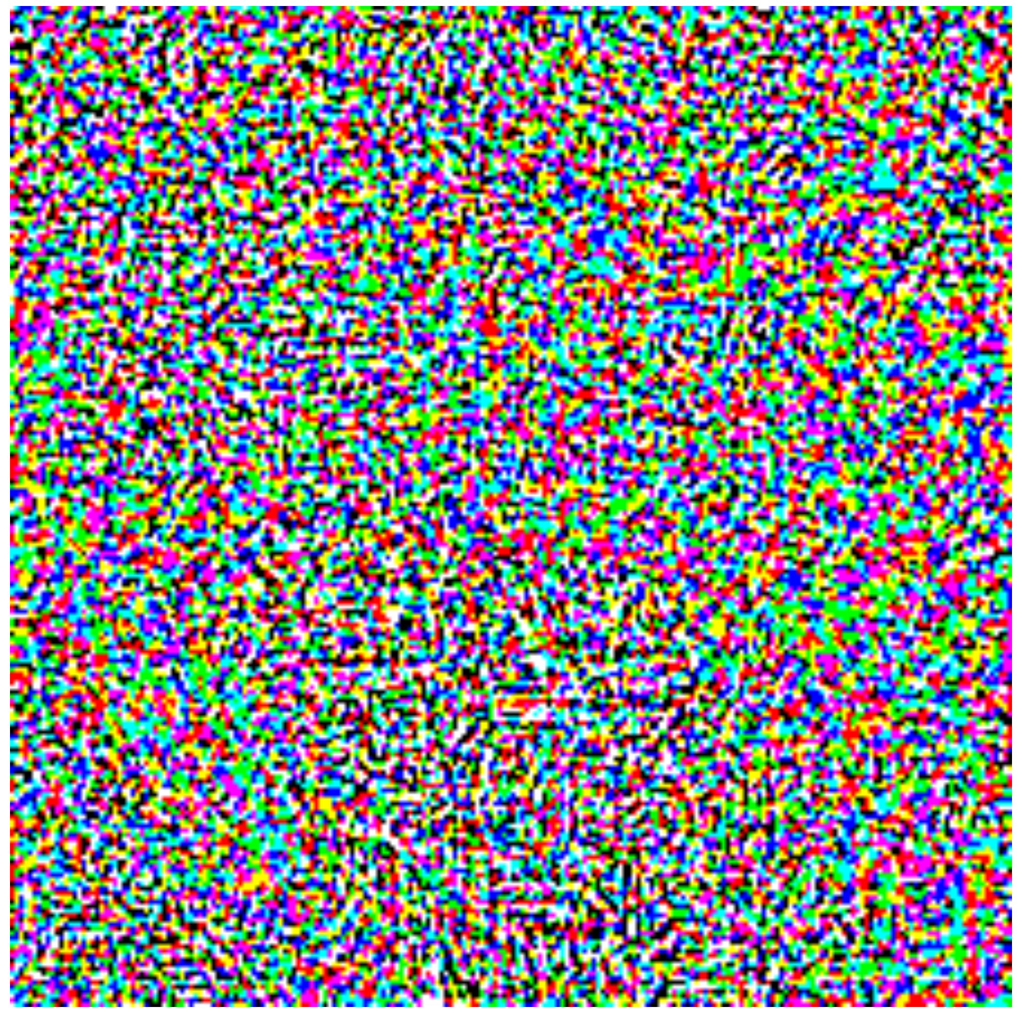
So Far: Adapting to Natural Shifts

Adapt

What about adversarial shifts?

Adversarial Examples

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} +$
$\epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

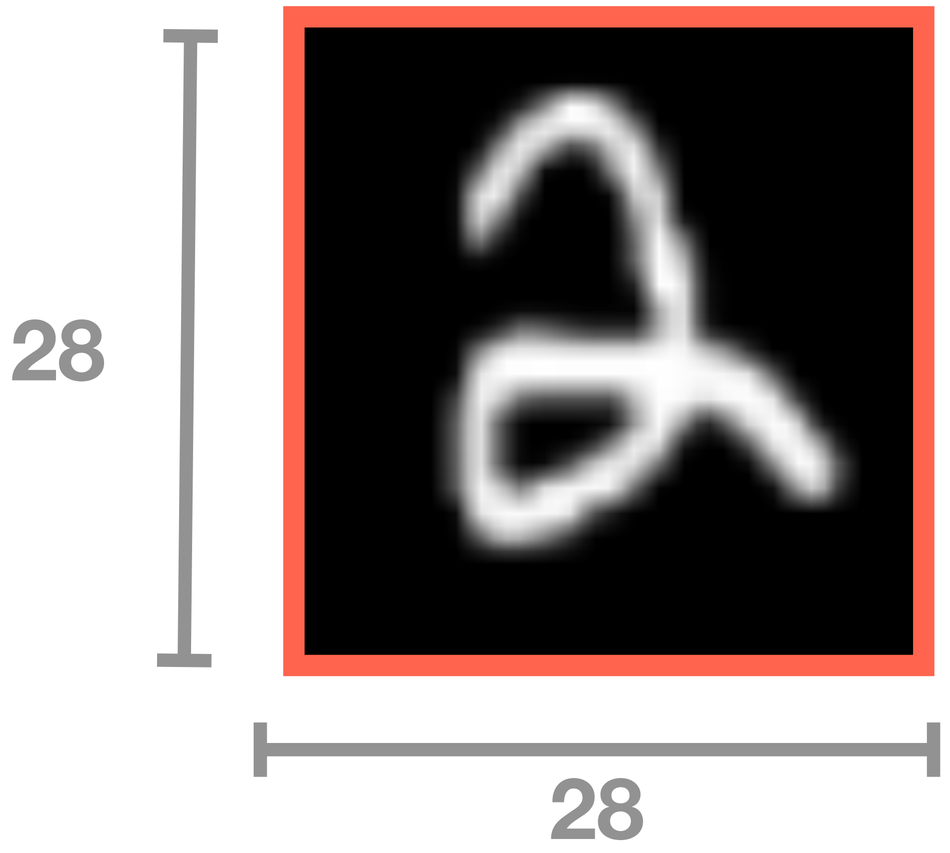
"gibbon"
99.3 % confidence

Goodfellow et al. ICLR 2015.

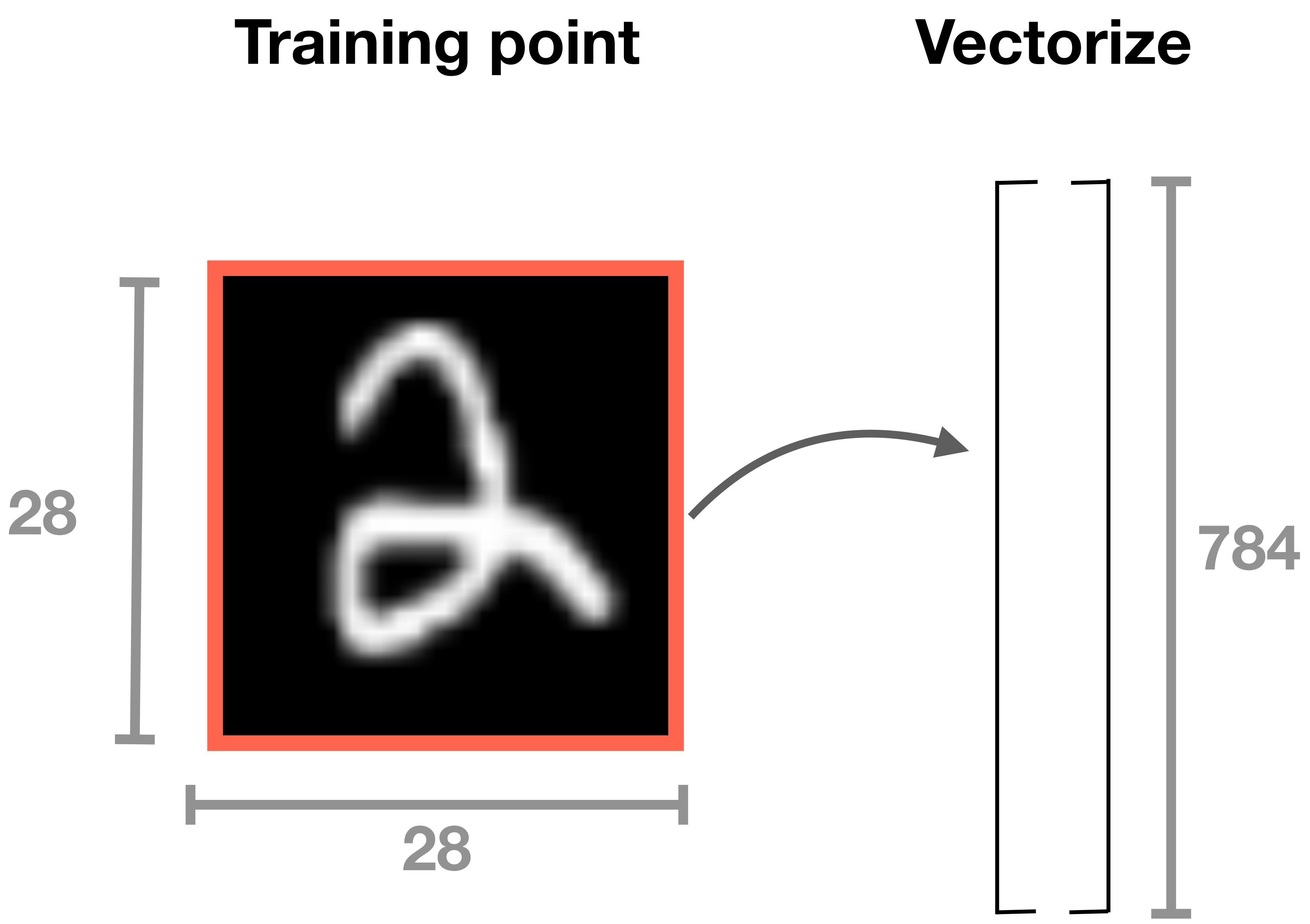
Visualize Perturbation Space

Visualize Perturbation Space

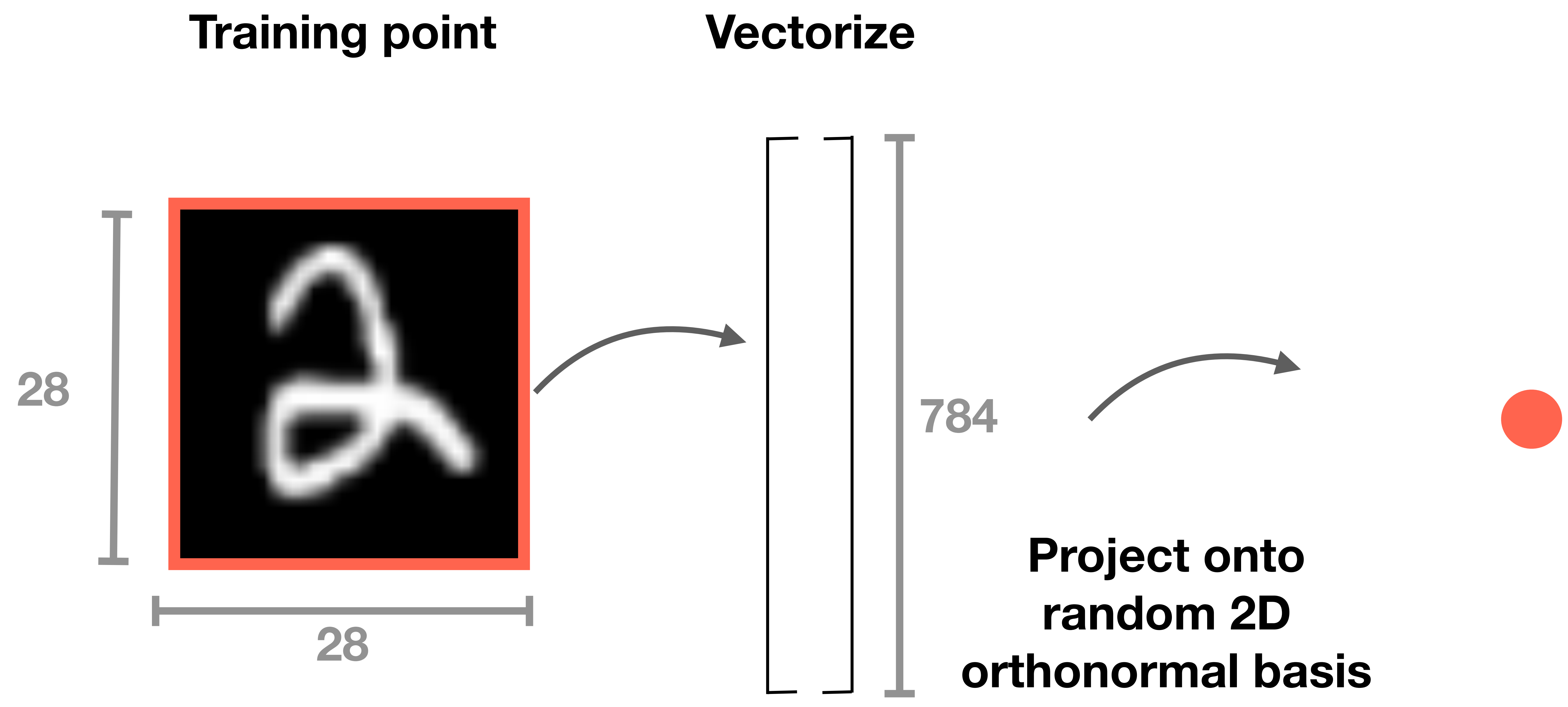
Training point



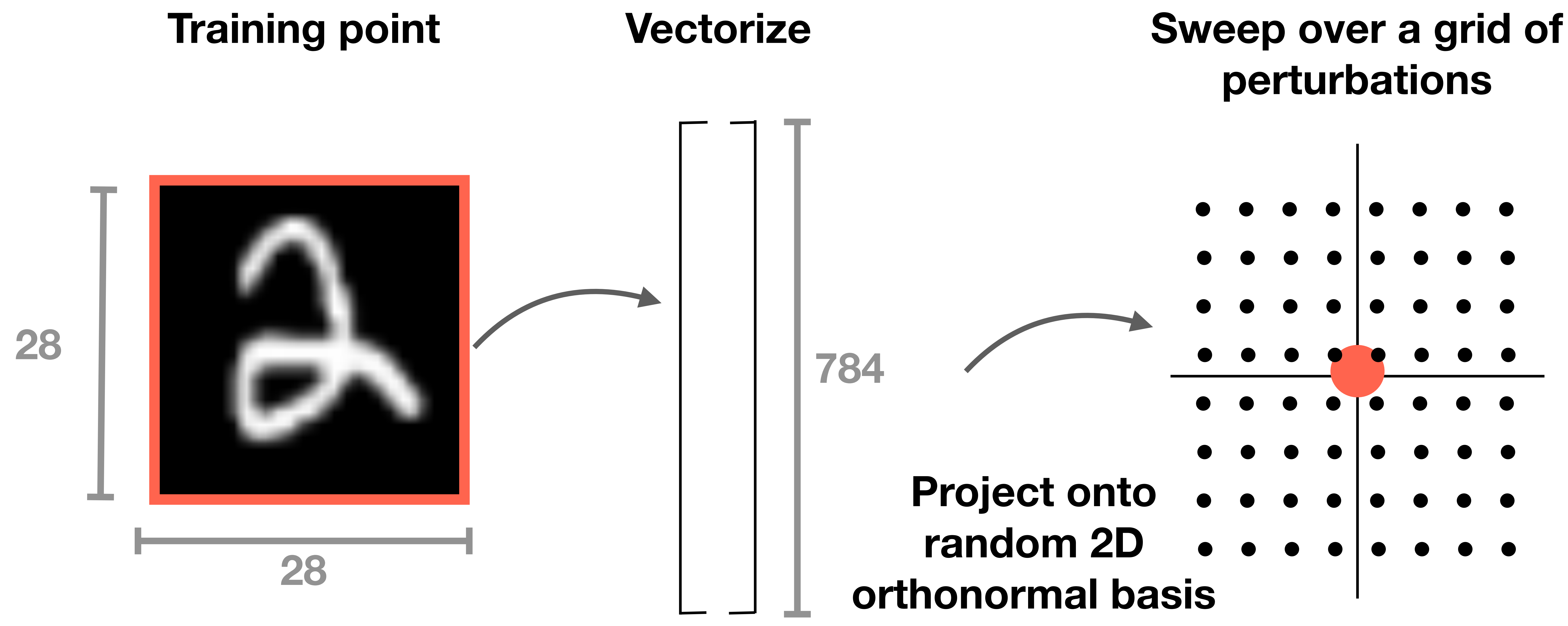
Visualize Perturbation Space



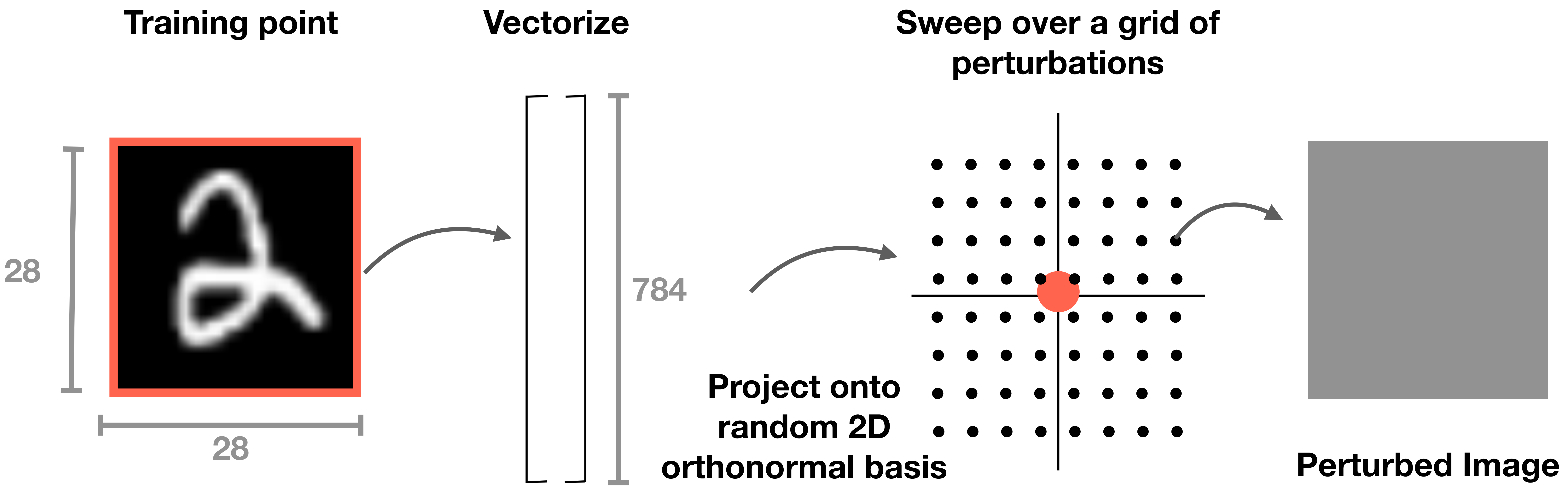
Visualize Perturbation Space



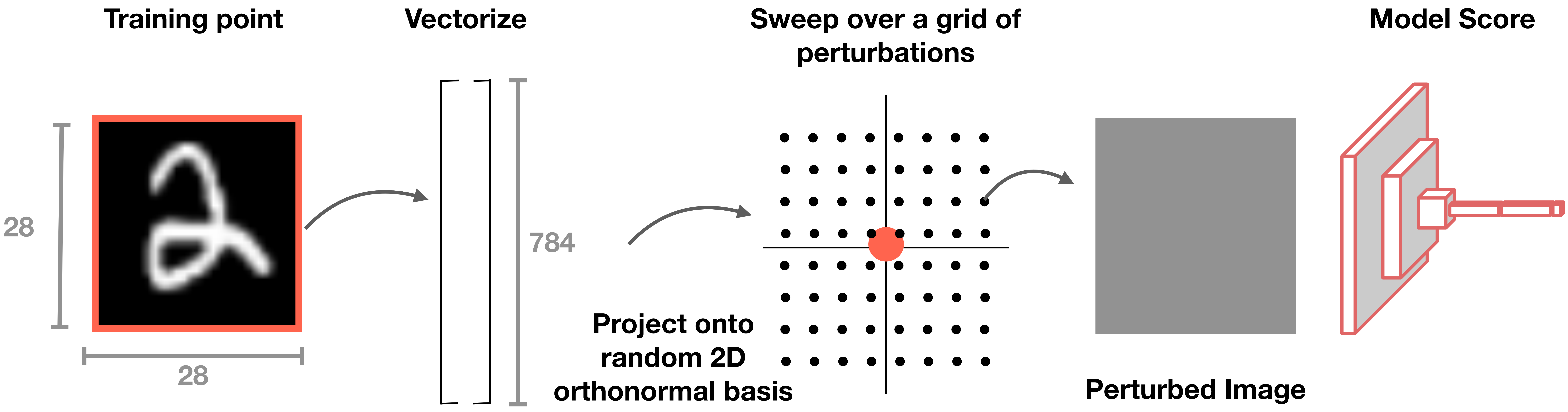
Visualize Perturbation Space



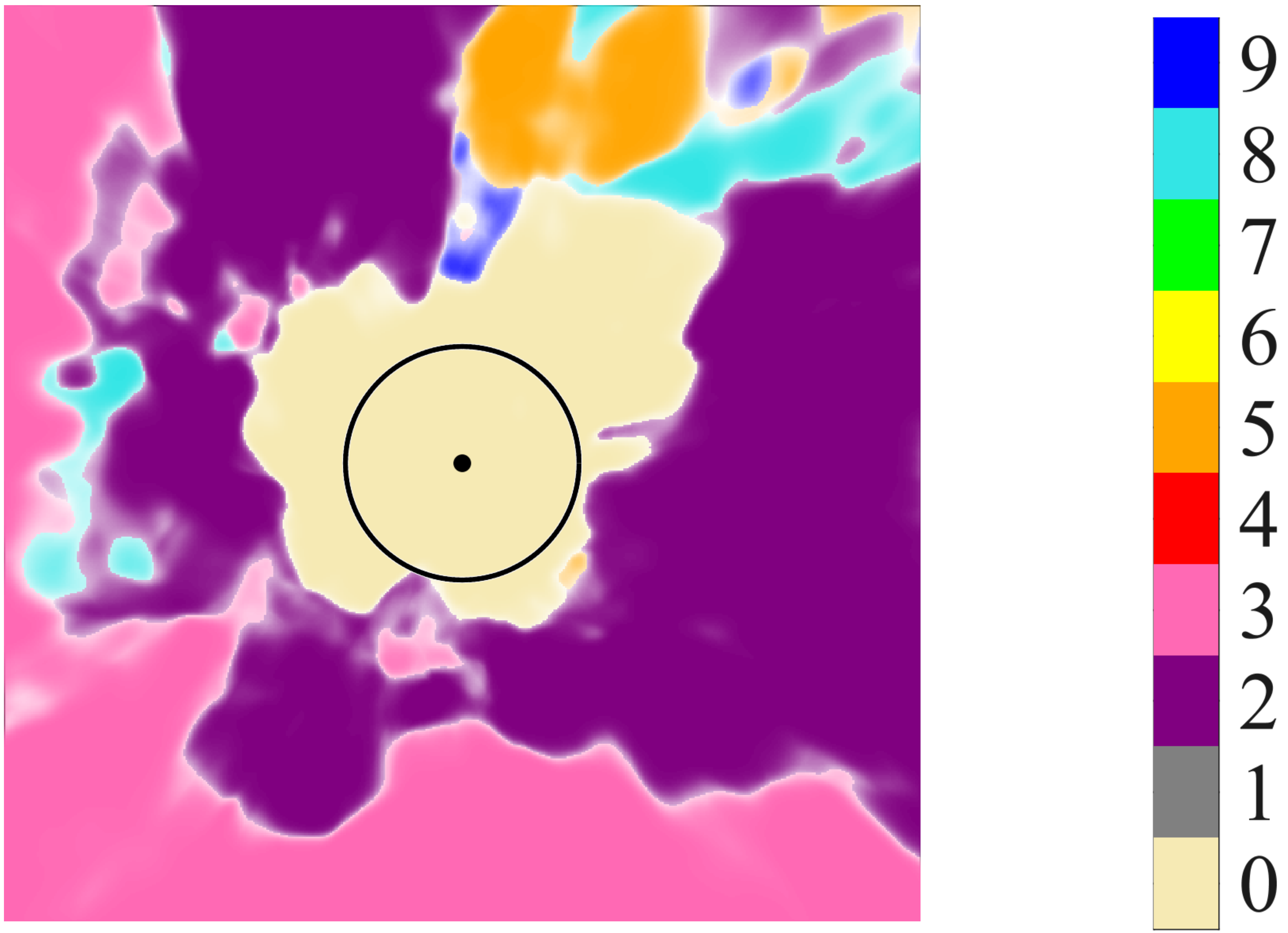
Visualize Perturbation Space



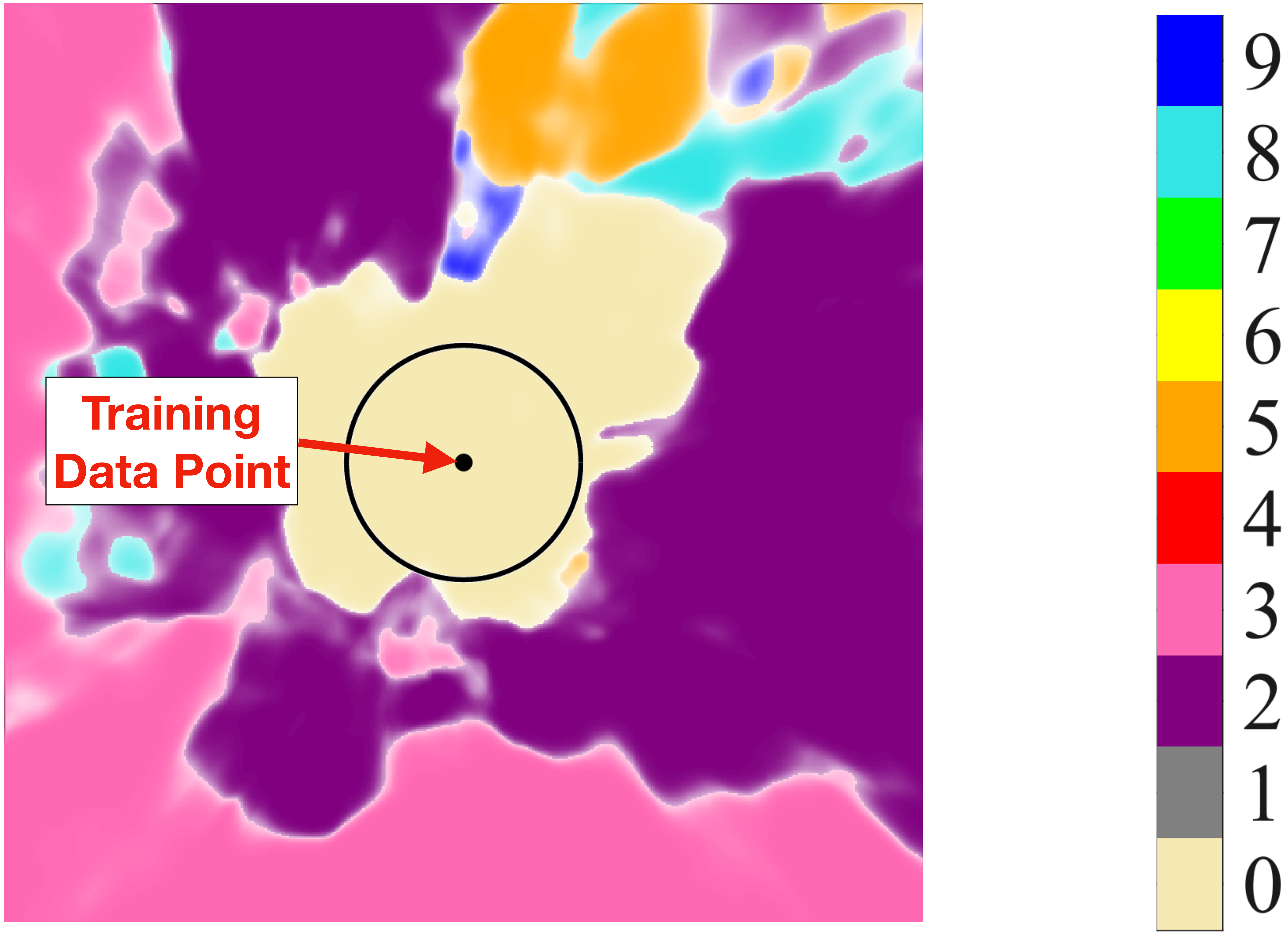
Visualize Perturbation Space



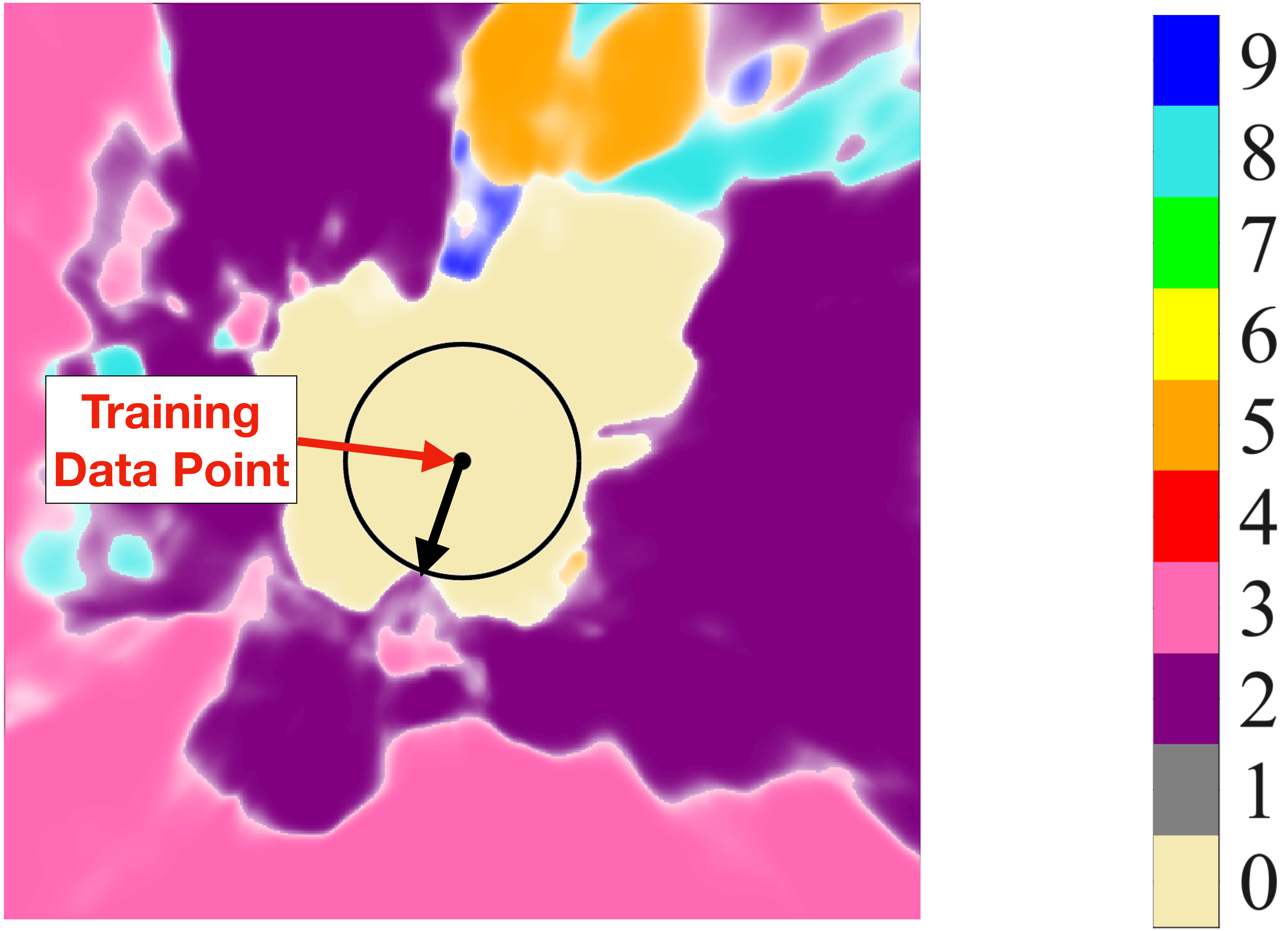
MNIST LeNet Decisions Around Training Point



MNIST LeNet Decisions Around Training Point

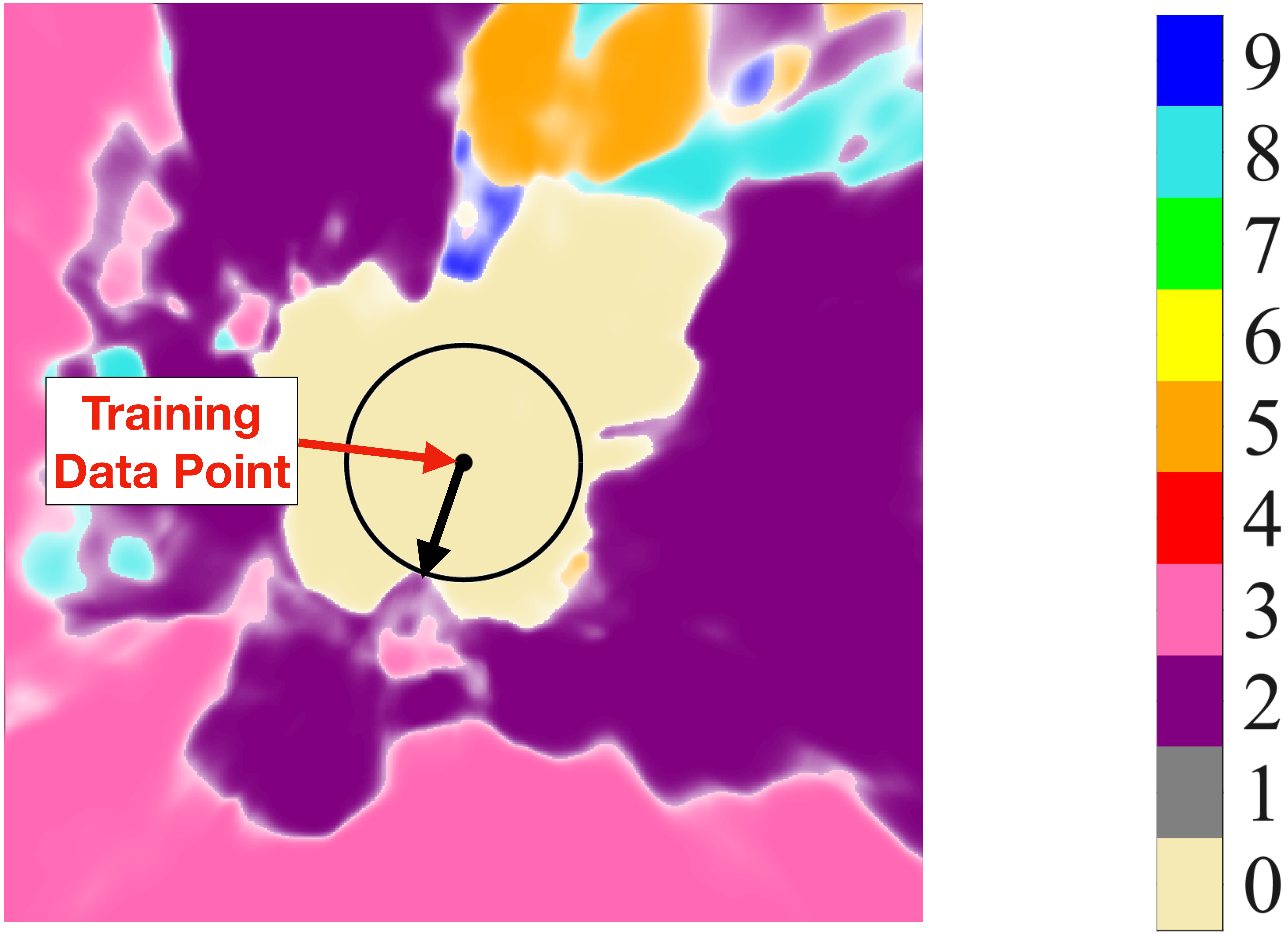


MNIST LeNet Decisions Around Training Point



MNIST LeNet Decisions Around Training Point

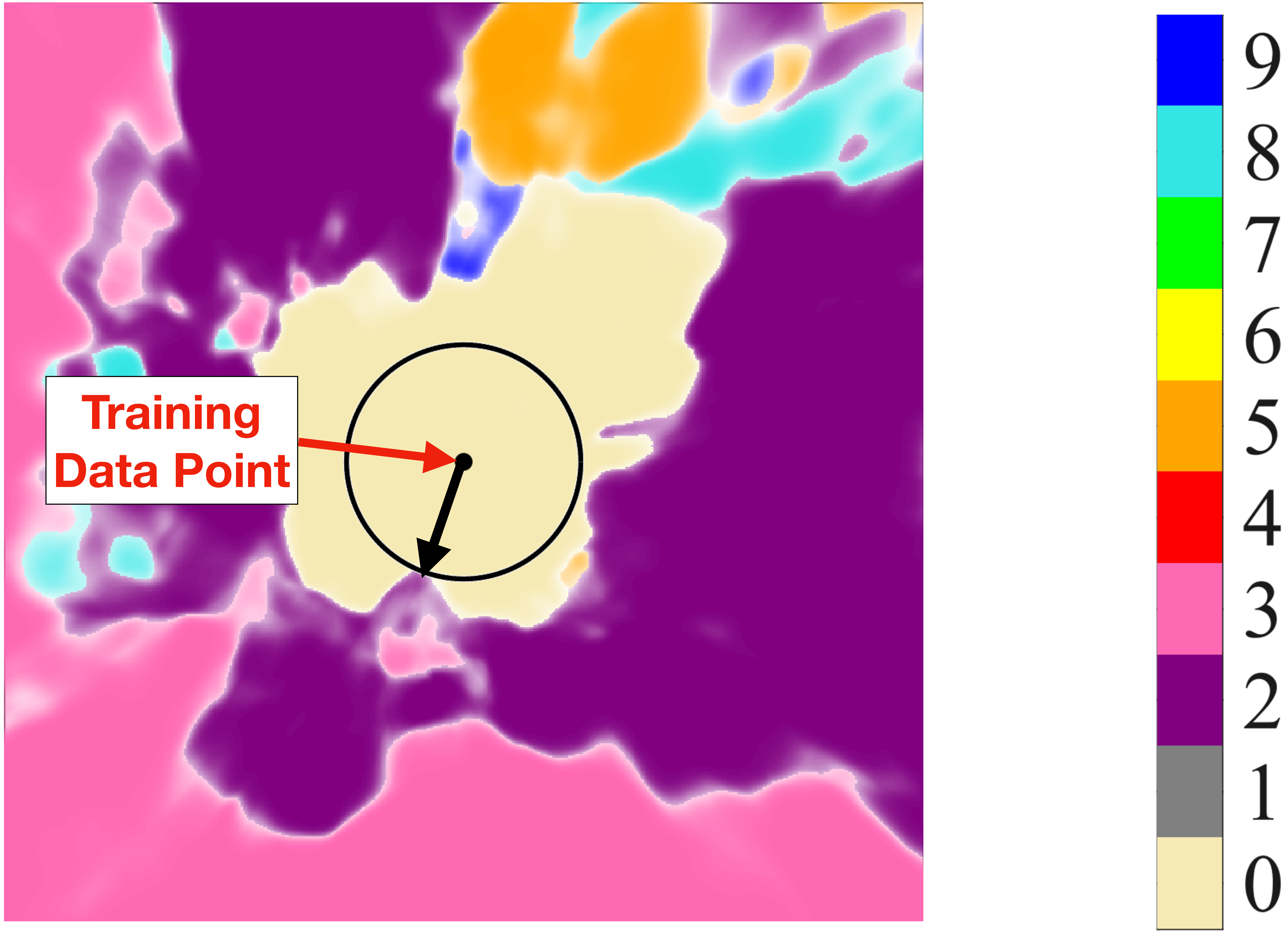
Non-smooth Decision Boundary



MNIST LeNet Decisions Around Training Point

Non-smooth Decision Boundary

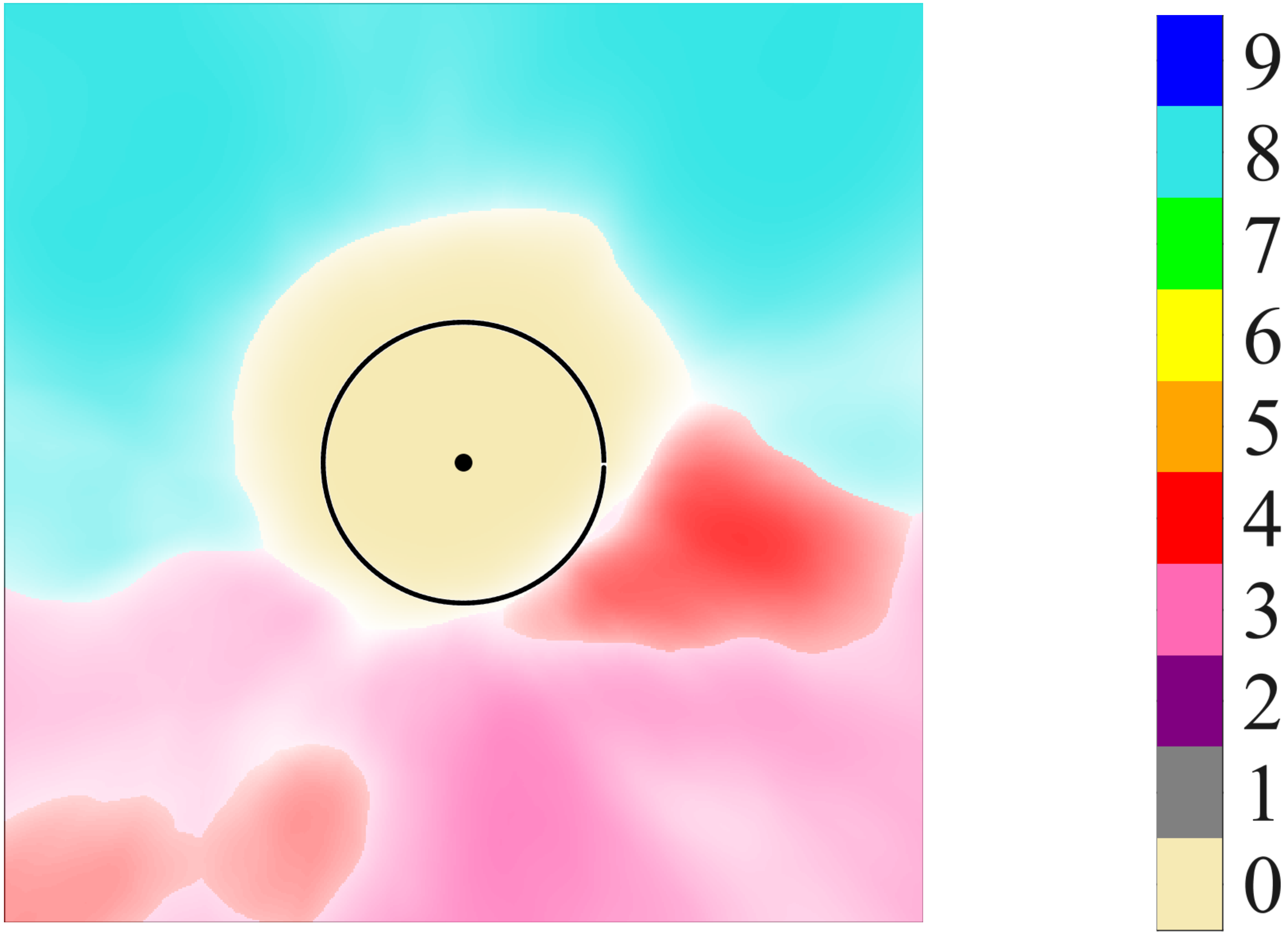
Small perturbations lead to new outputs



MNIST LeNet with L2 Regularization

Smooth Decision Boundary

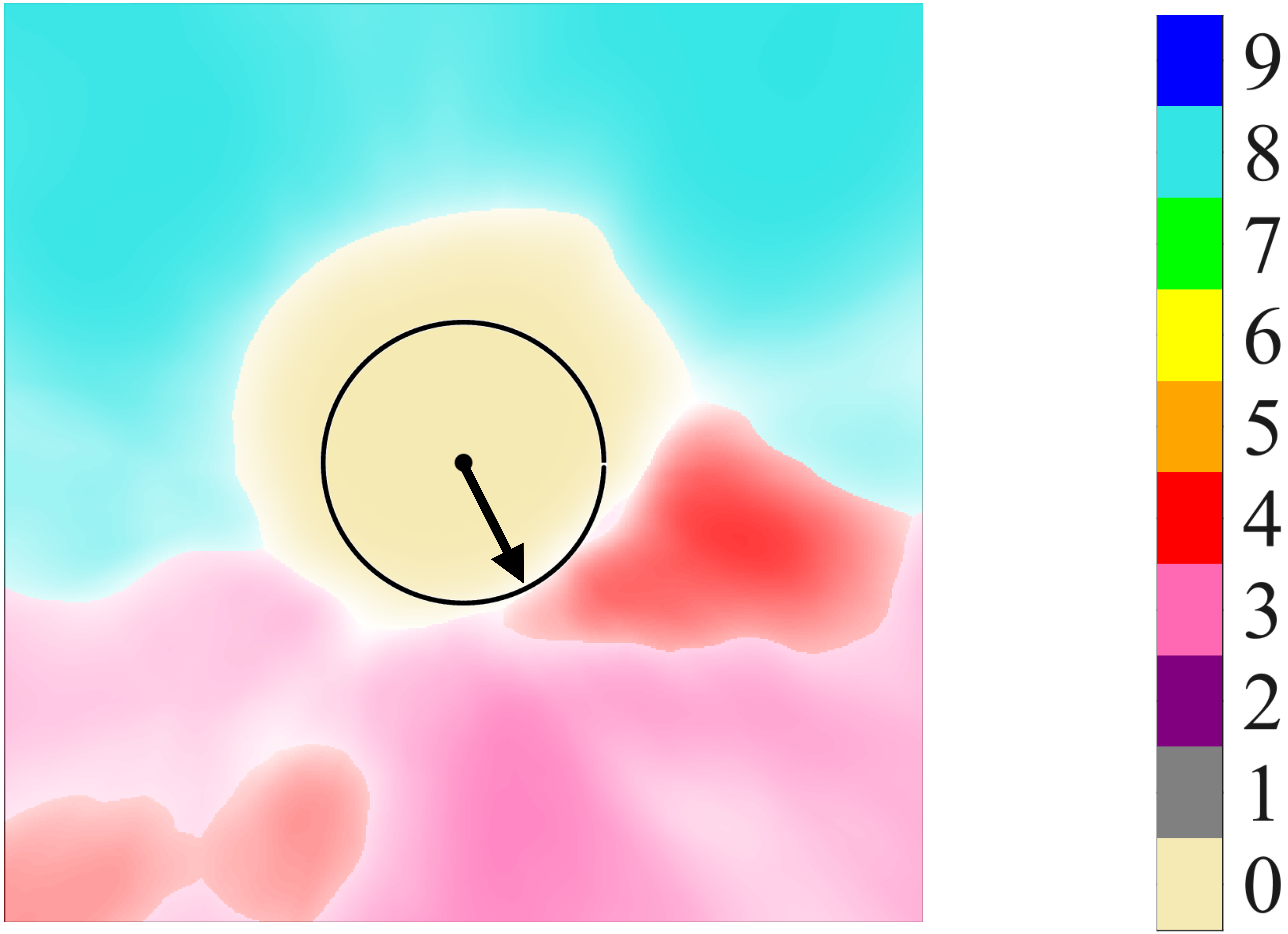
Small perturbations lead to new outputs



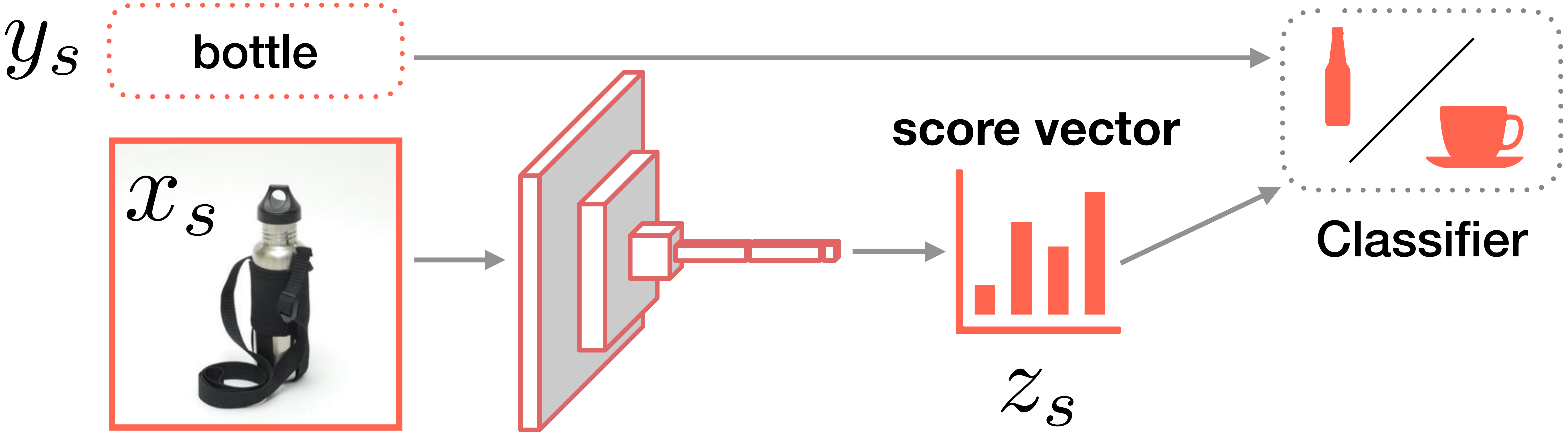
MNIST LeNet with L2 Regularization

Smooth Decision Boundary

Small perturbations lead to new outputs

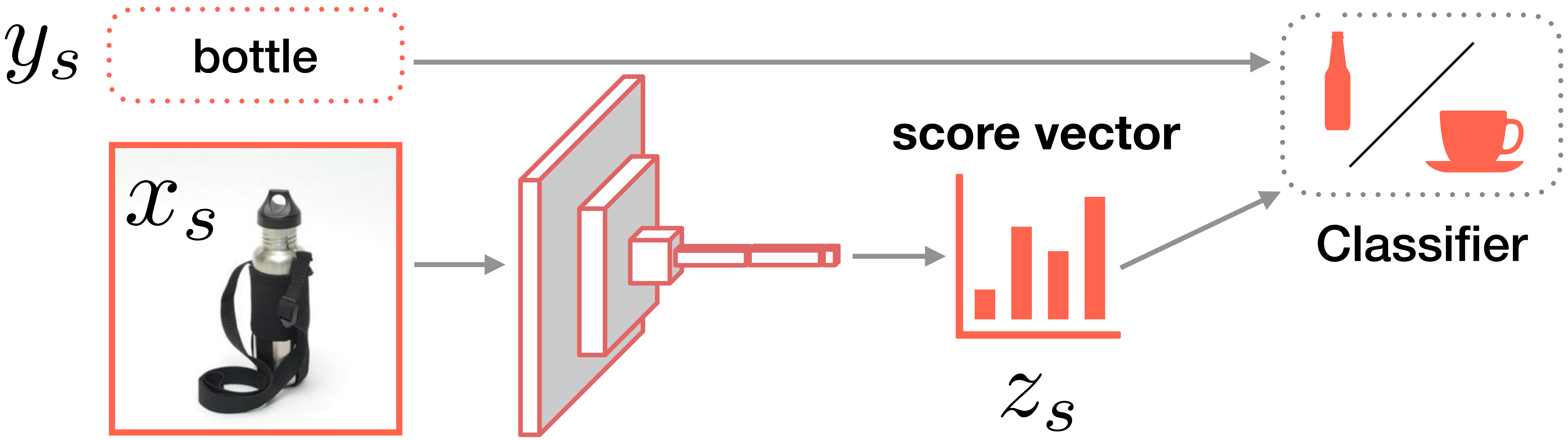


Jacobian Regularization



Hoffman, Roberts, Yaida, In submission, 2019.

Jacobian Regularization

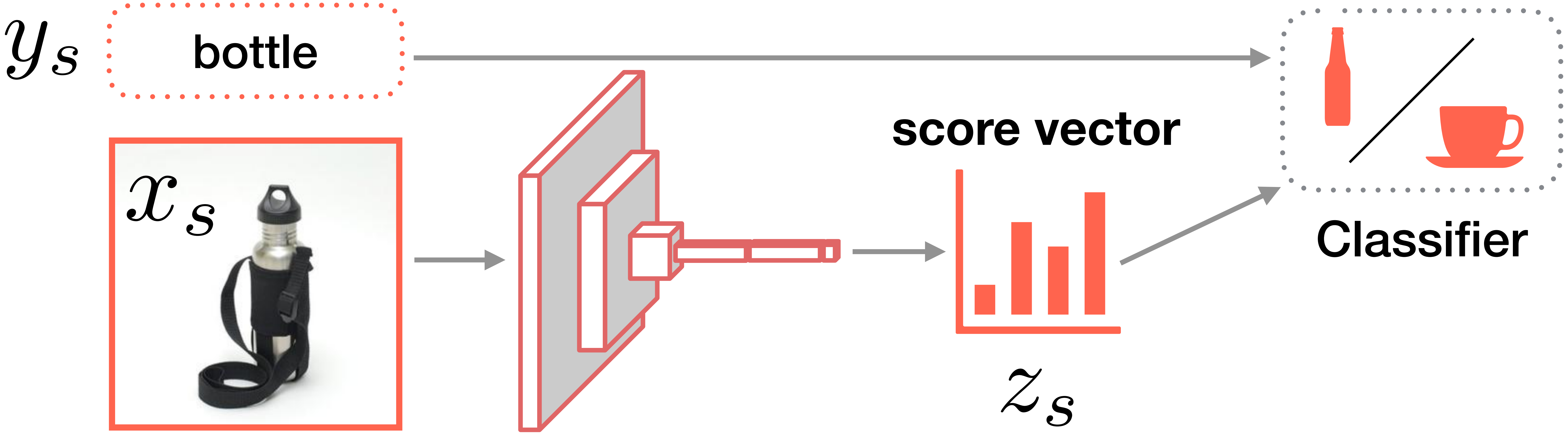


Input-output Jacobian matrix

$$J_{c,i} = \frac{\partial z_c}{\partial x_i}$$

Hoffman, Roberts, Yaida, In submission, 2019.

Jacobian Regularization



Input-output Jacobian matrix

$$J_{c,i} = \frac{\partial z_c}{\partial x_i}$$

Minimize Frobenius Norm

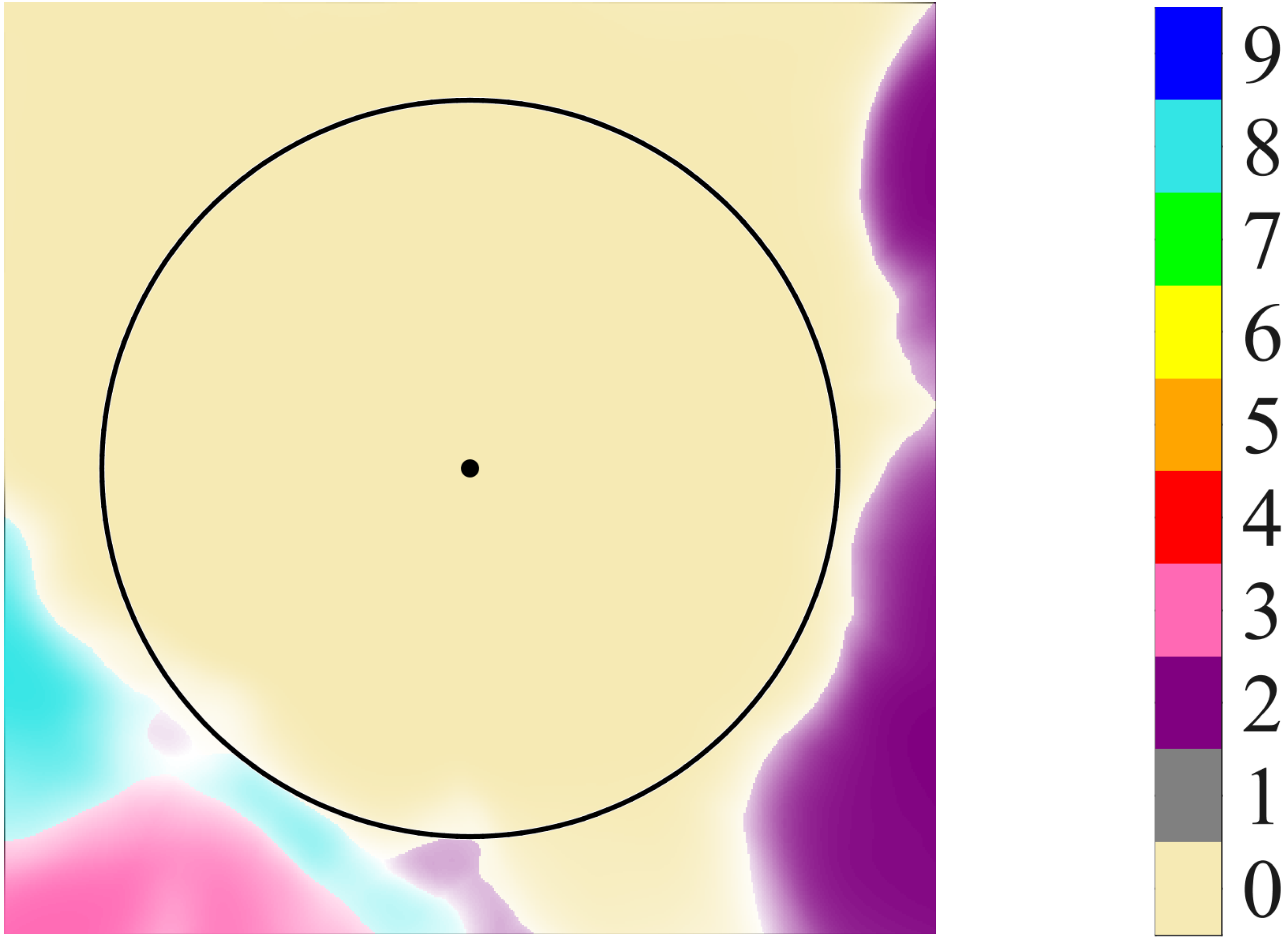
$$||J||_F^2$$

Hoffman, Roberts, Yaida, In submission, 2019.

MNIST LeNet with Jacobian Regularization

Mostly Smooth Decision Boundary

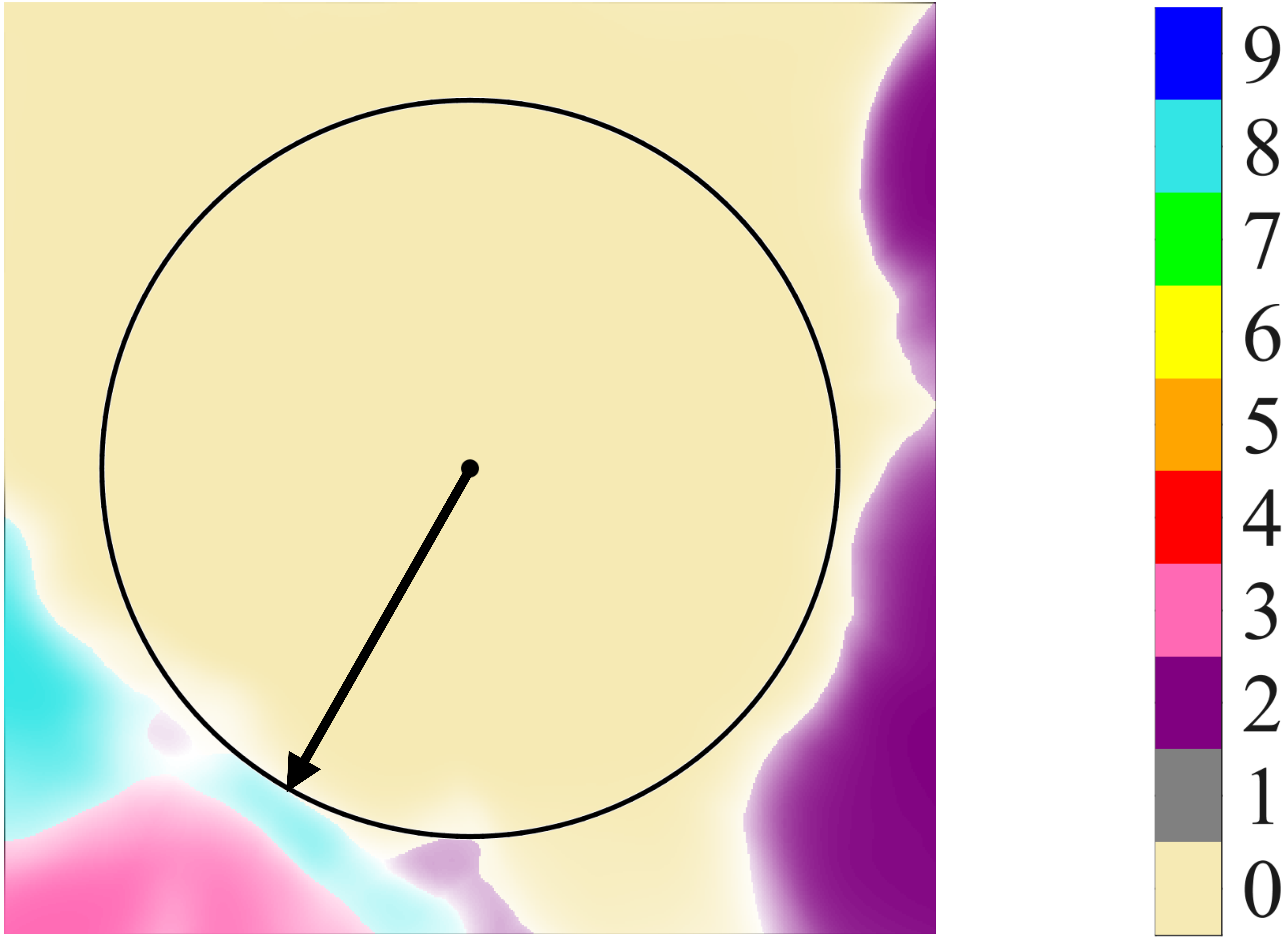
Larger perturbations needed to lead to new outputs



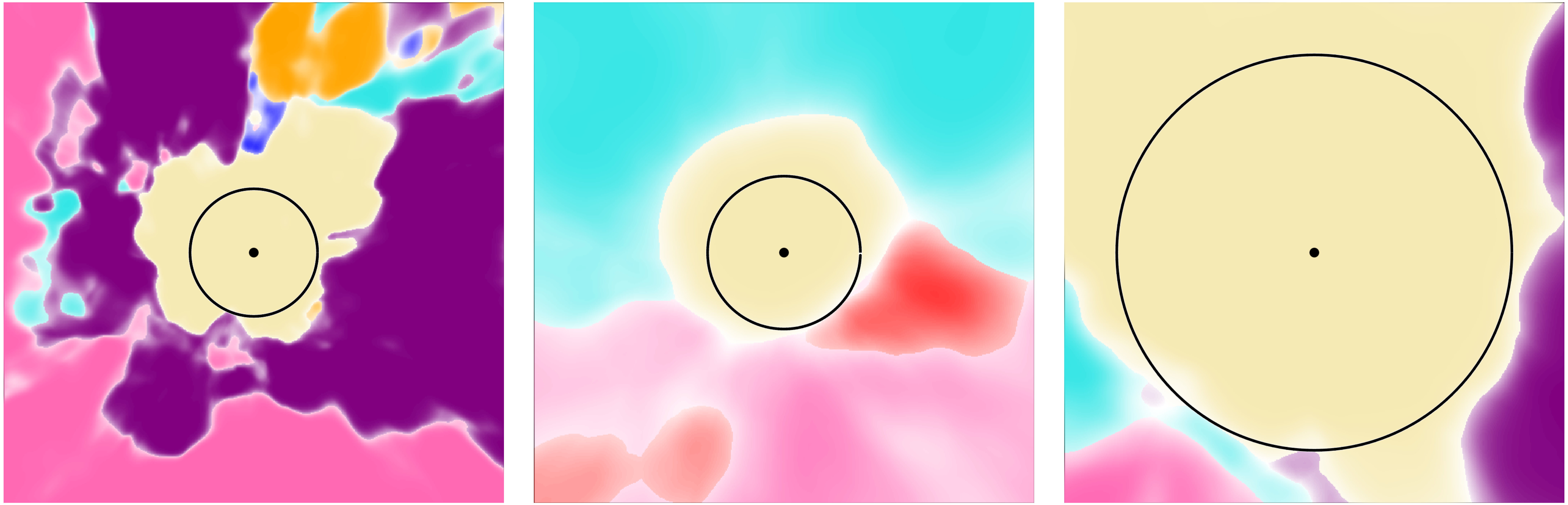
MNIST LeNet with Jacobian Regularization

Mostly Smooth Decision Boundary

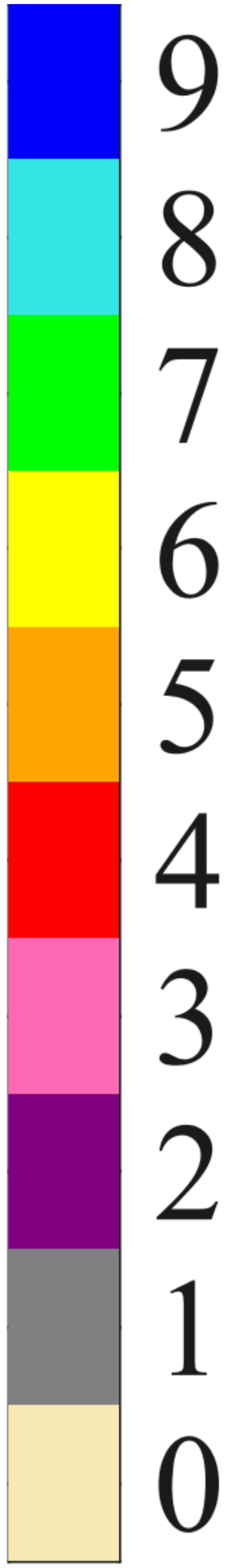
Larger perturbations needed to lead to new outputs



Decision Boundary Comparison

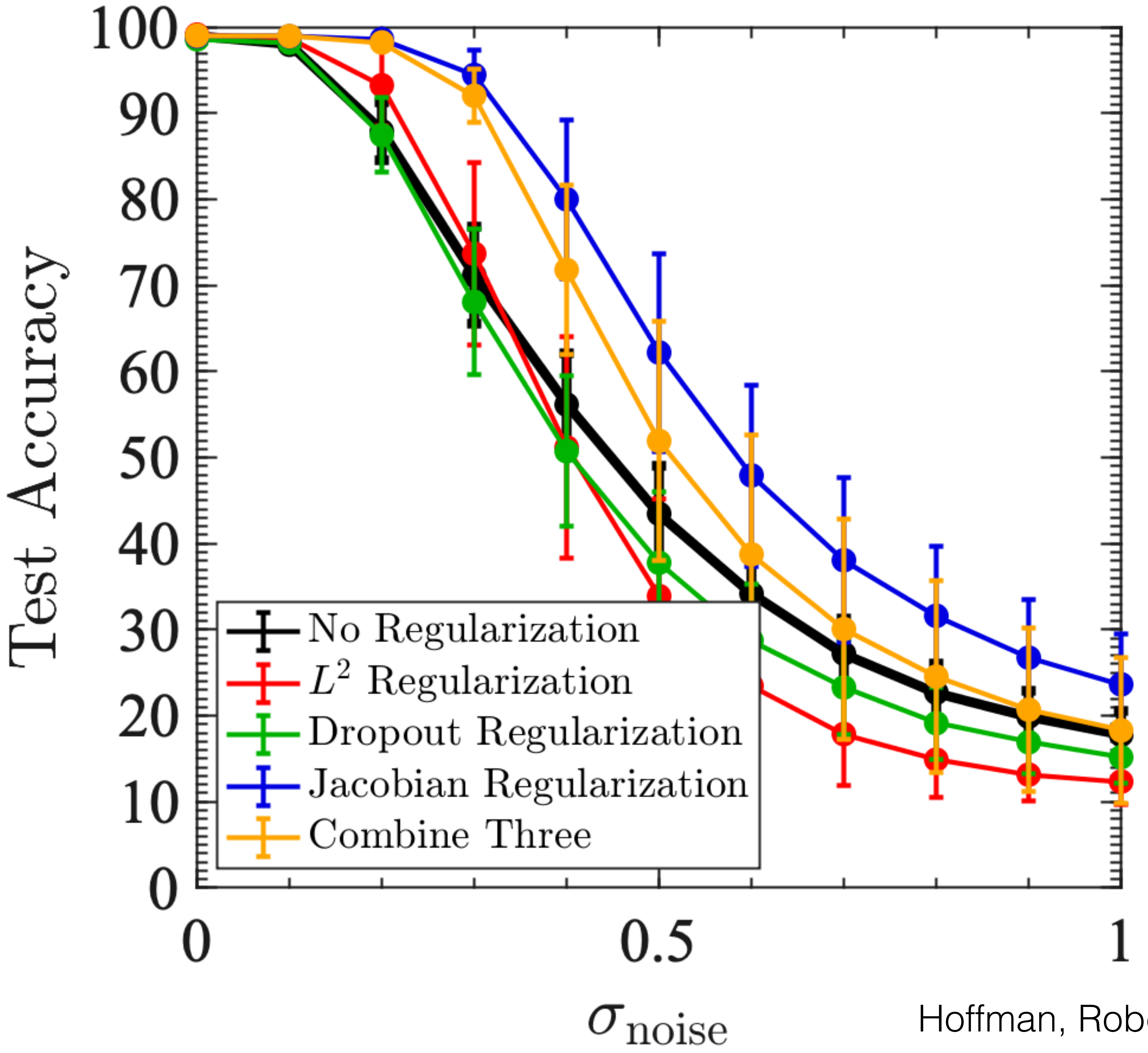


No Regularization | **L2 Regularization** | **Jacobian Regularization**



Hoffman, Roberts, Yaida, In submission, 2019.

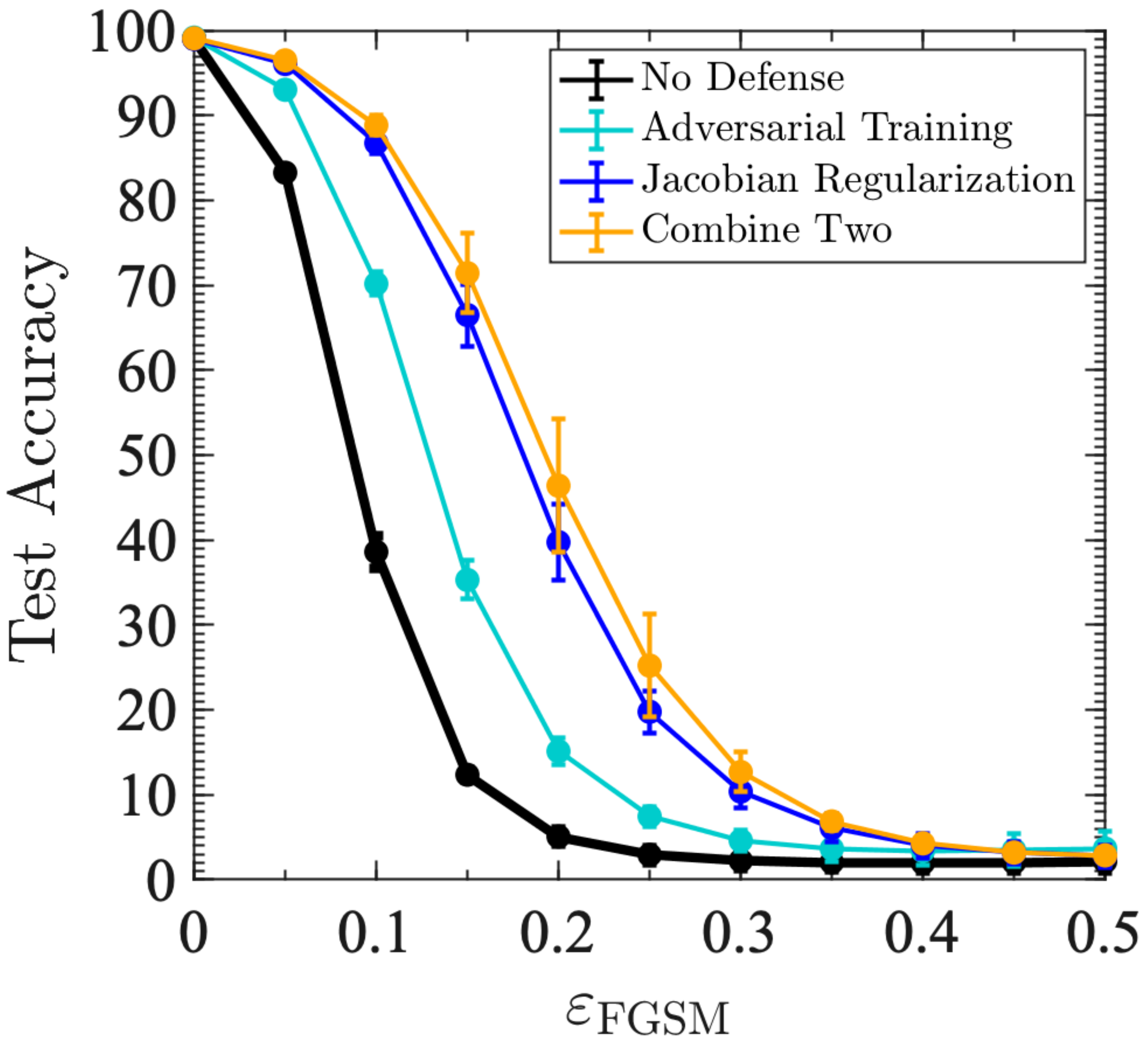
Robustness to Random Perturbations

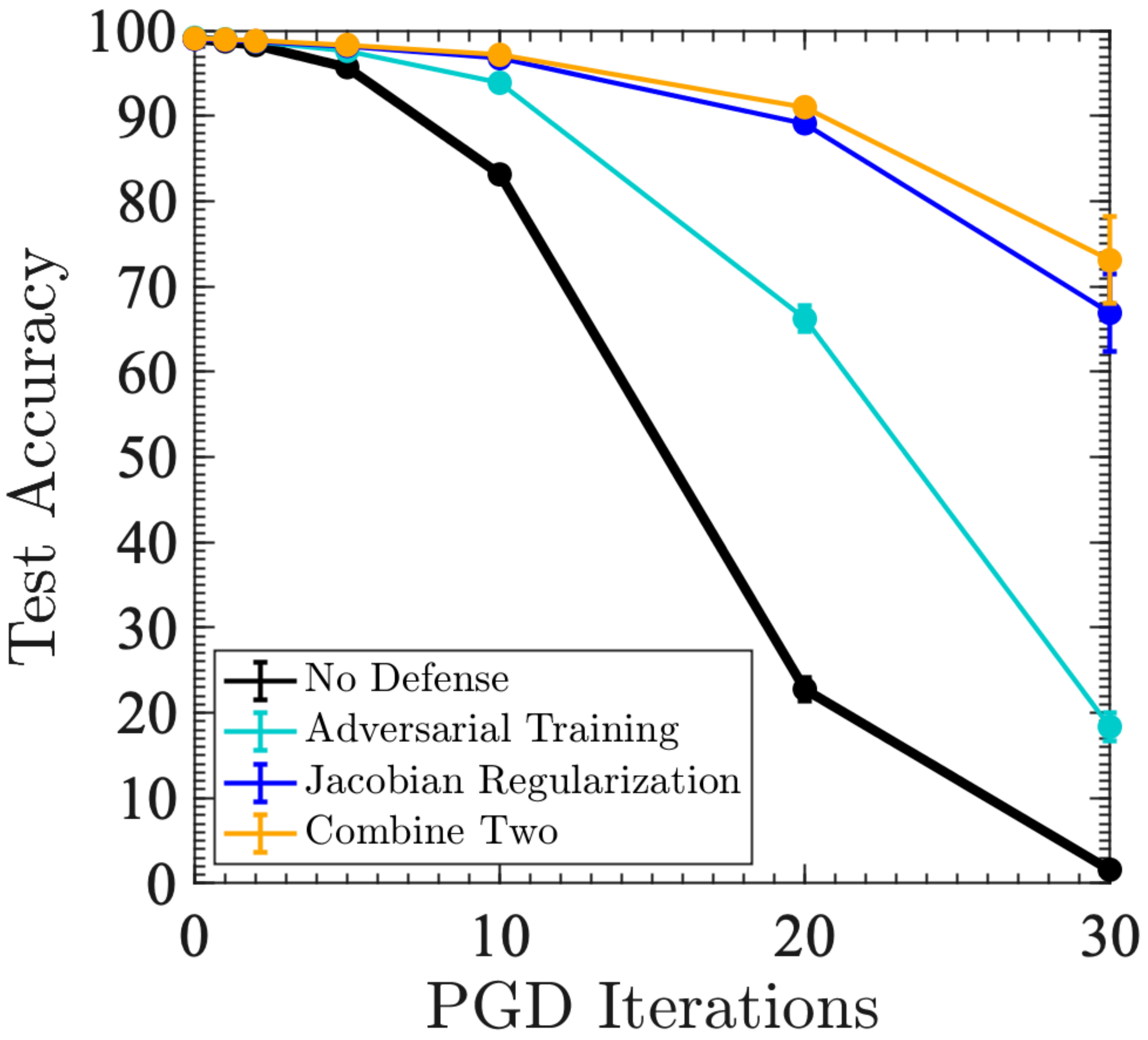


MNIST
LeNet Model

Hoffman, Roberts, Yaida, In submission, 2019.

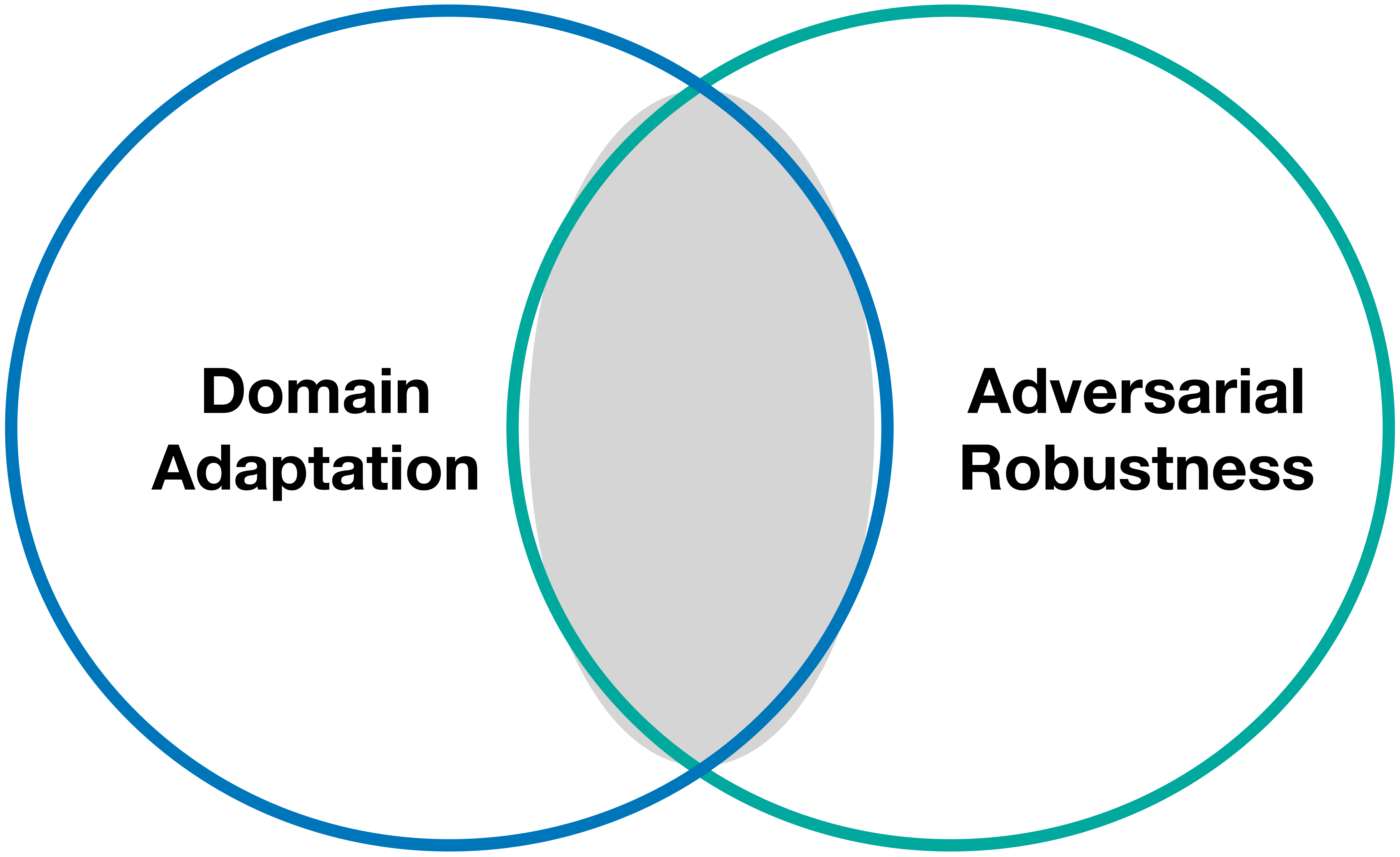
Robustness to Adversarial Perturbations





Hoffman, Roberts, Yaida, In submission, 2019.

Next Steps

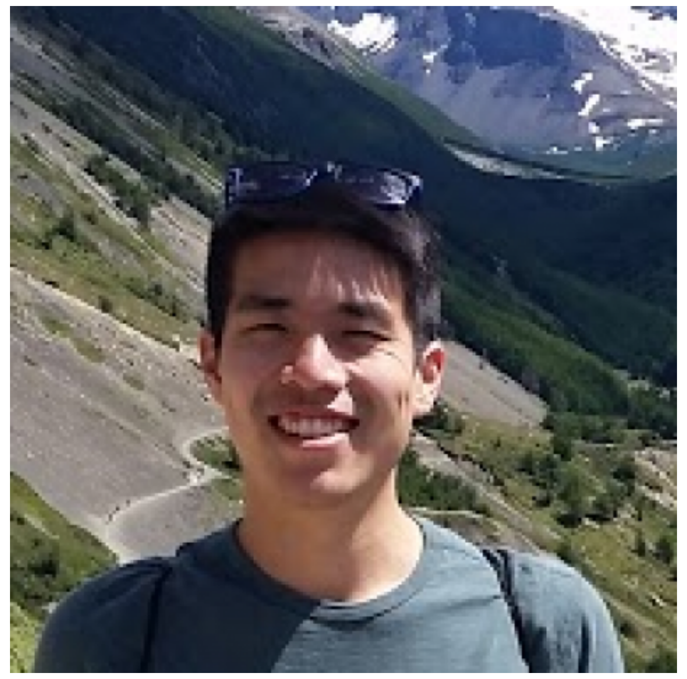


Jacobian regularizer as unsupervised adaptive loss?

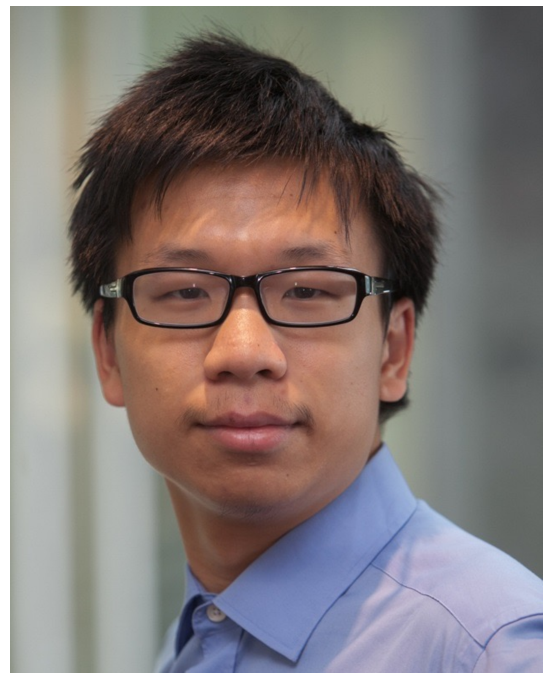
Adaptation to an adversarial domain?

Thank you

Taesung Park
UC Berkeley



Eric Tzeng
UC Berkeley



Jun-Yan Zhu
MIT

Dan Roberts
Diffeo

Phil Isola
MIT

Kate Saenko
Boston University

Trevor Darrell
UC Berkeley

Alyosha Efros
UC Berkeley

Sho Yaida
FAIR

Judy Hoffman
judyhoffman.io