A geometric perspective on the robustness of deep networks

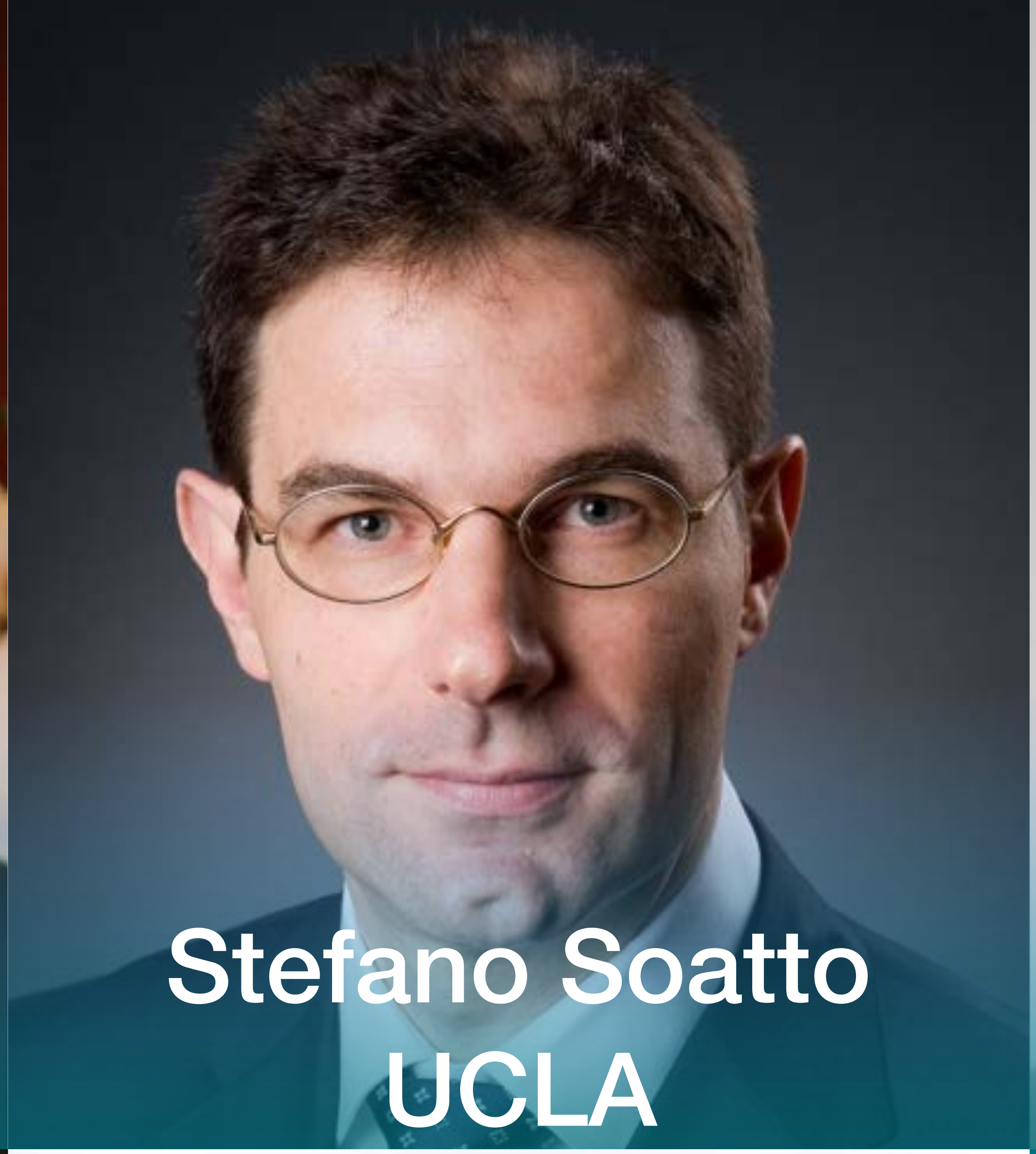
Seyed-Mohsen Moosavi-Dezfooli

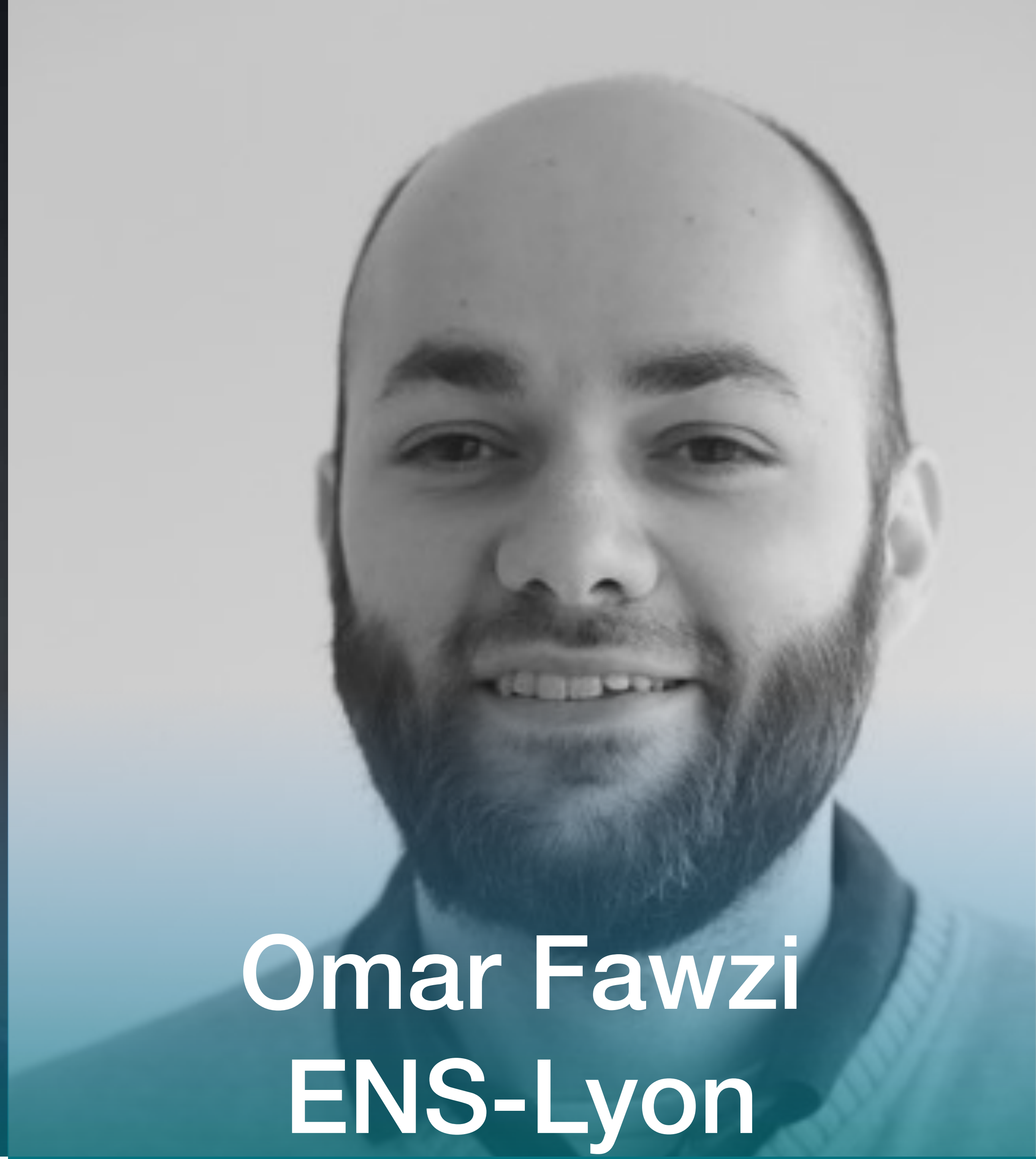
Amirkabir Artificial Intelligence Summer Summit
July 2019

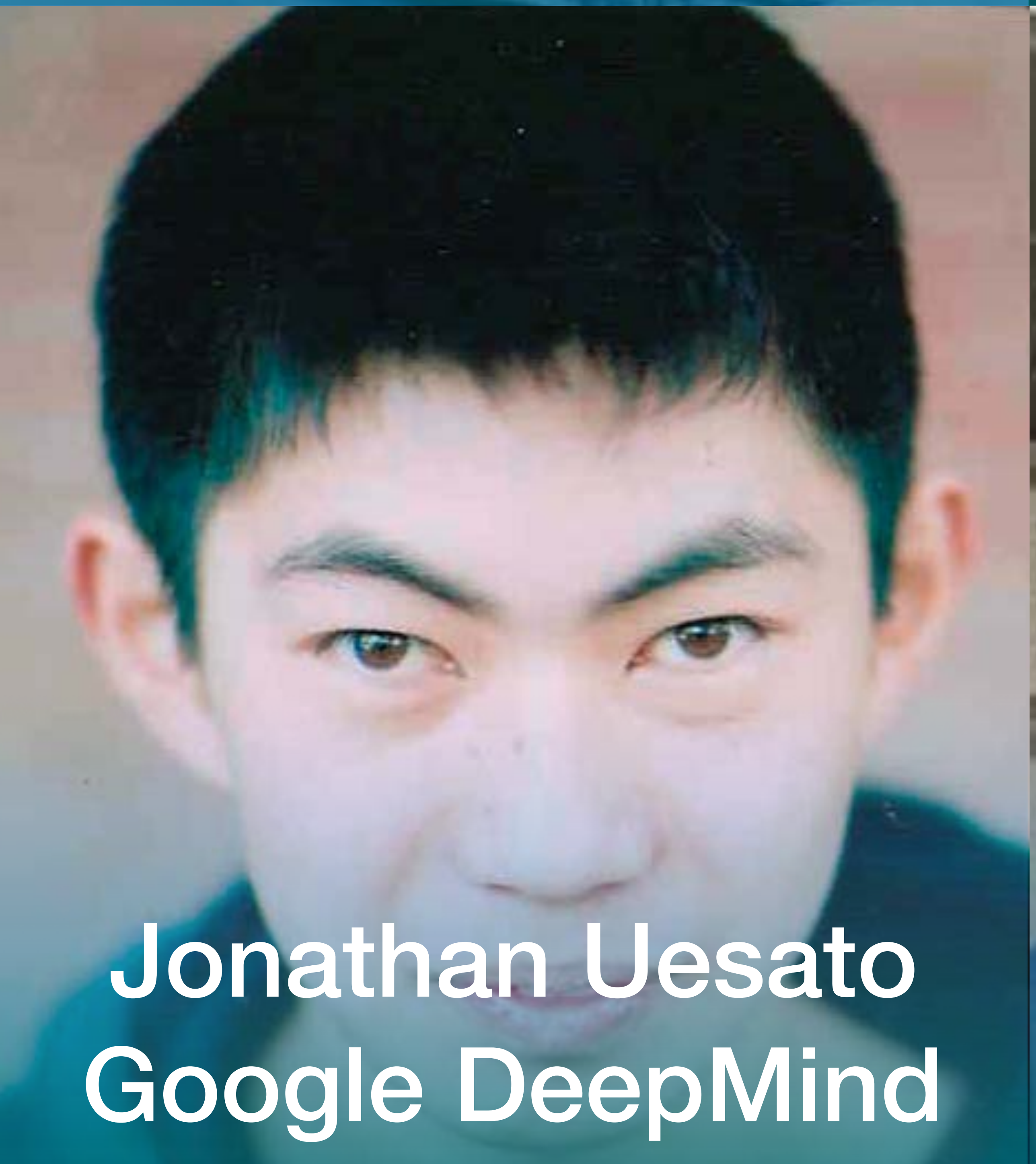
Laboratoire
de traitement
des signaux 4

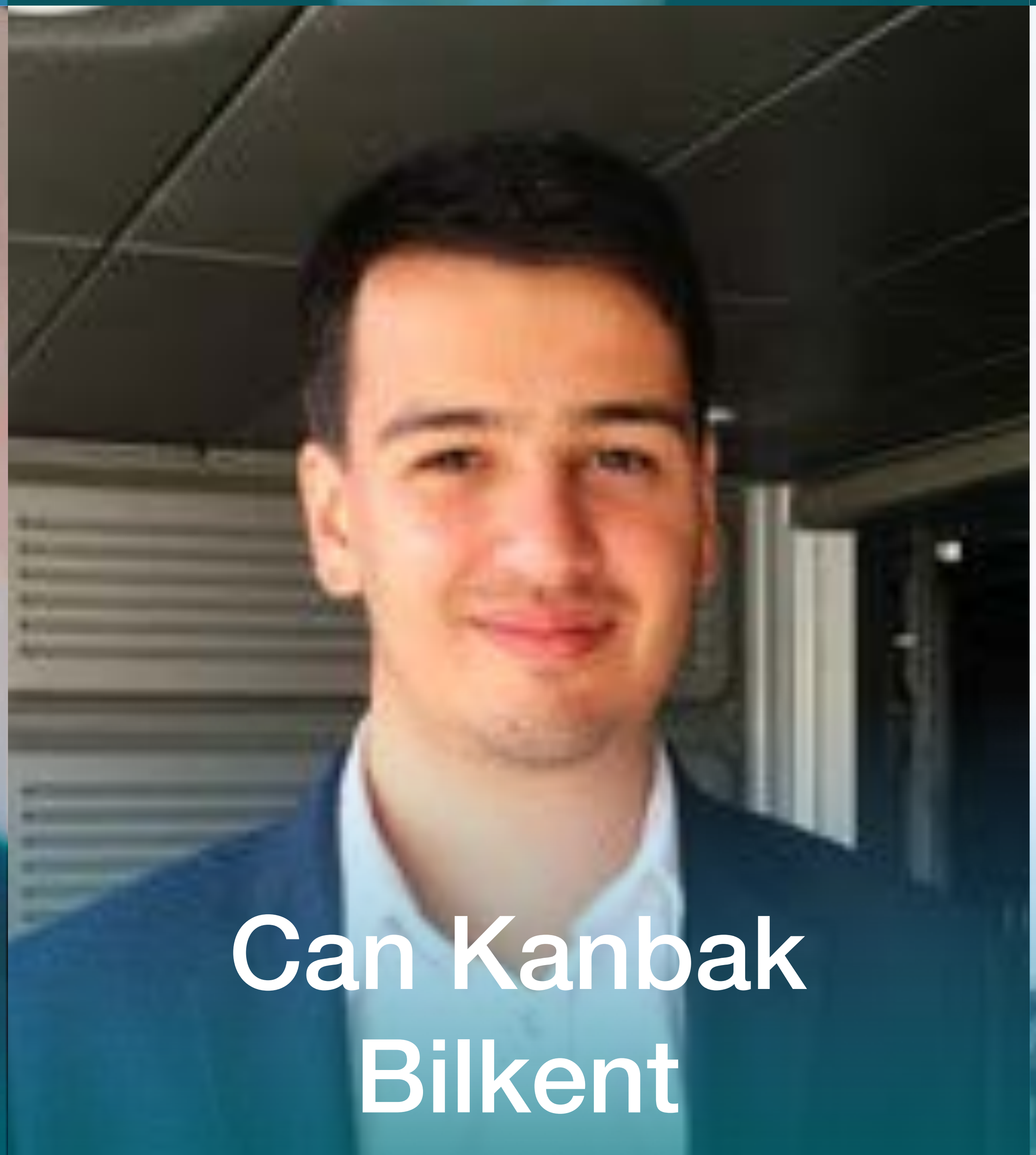
Tehran
Polytechnic
Iran

EPFL Lausanne, Switzerland



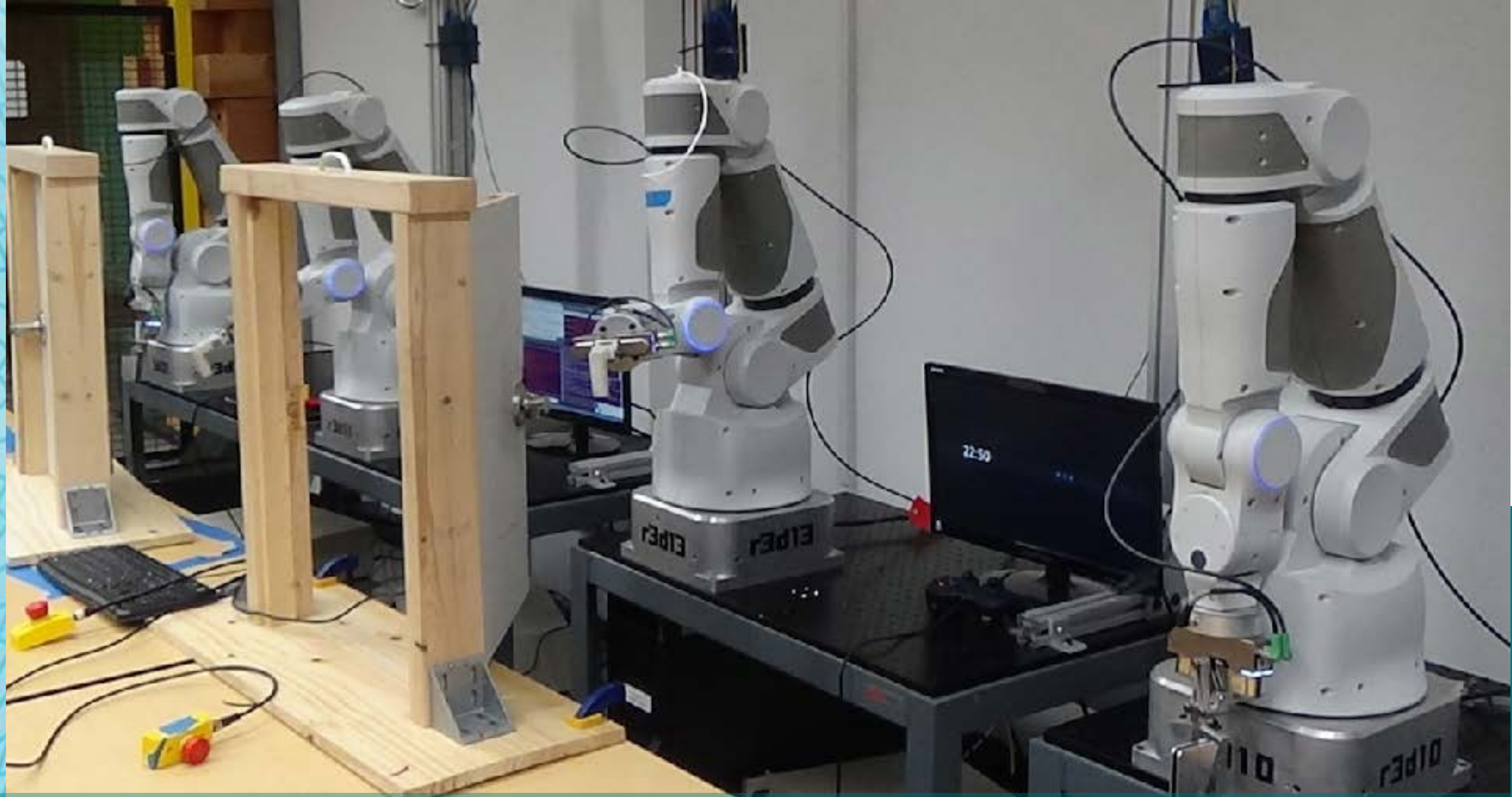






Collaborators

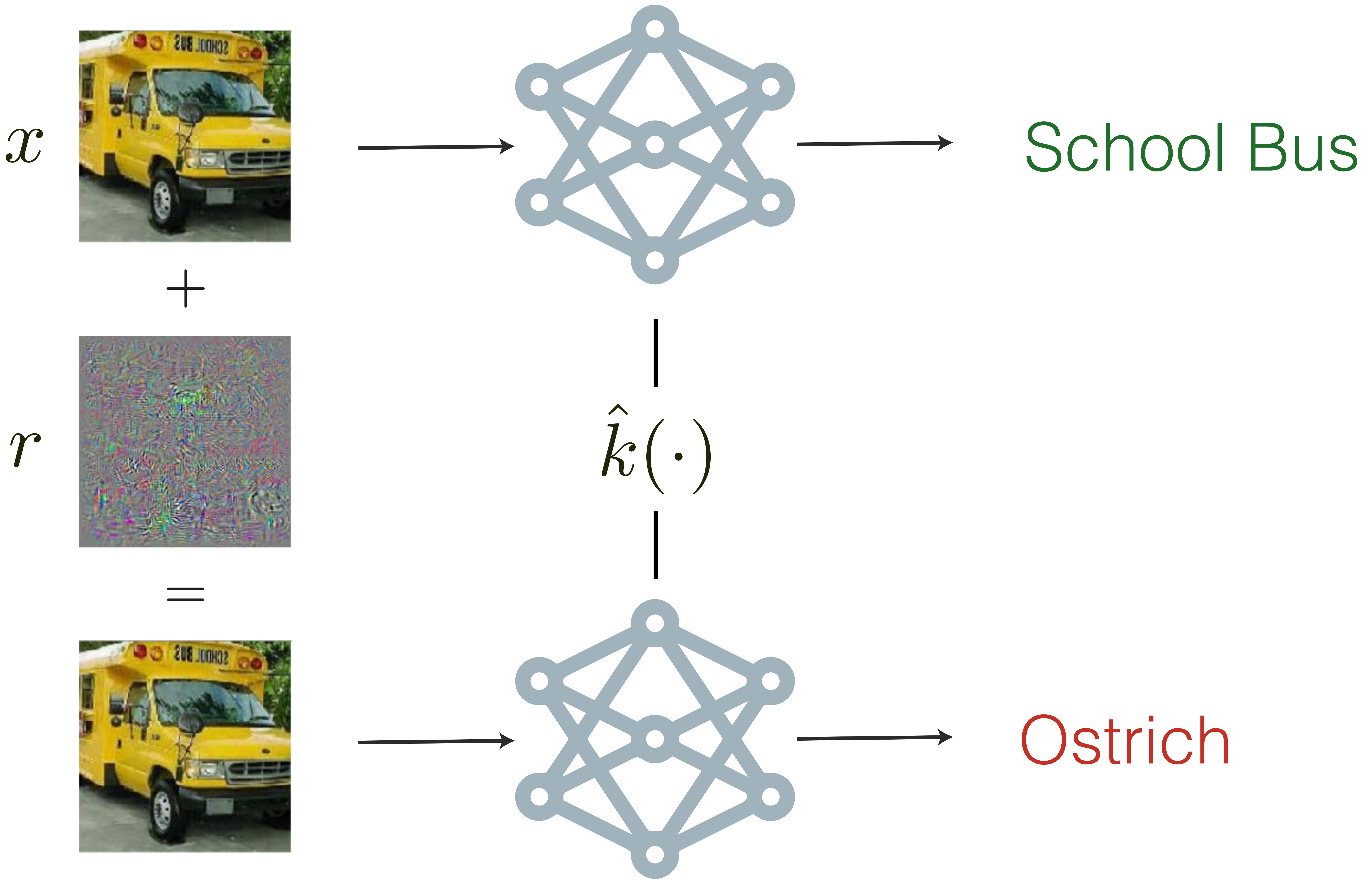
Rachel Jones—Biomedical Computation Review



Google Research

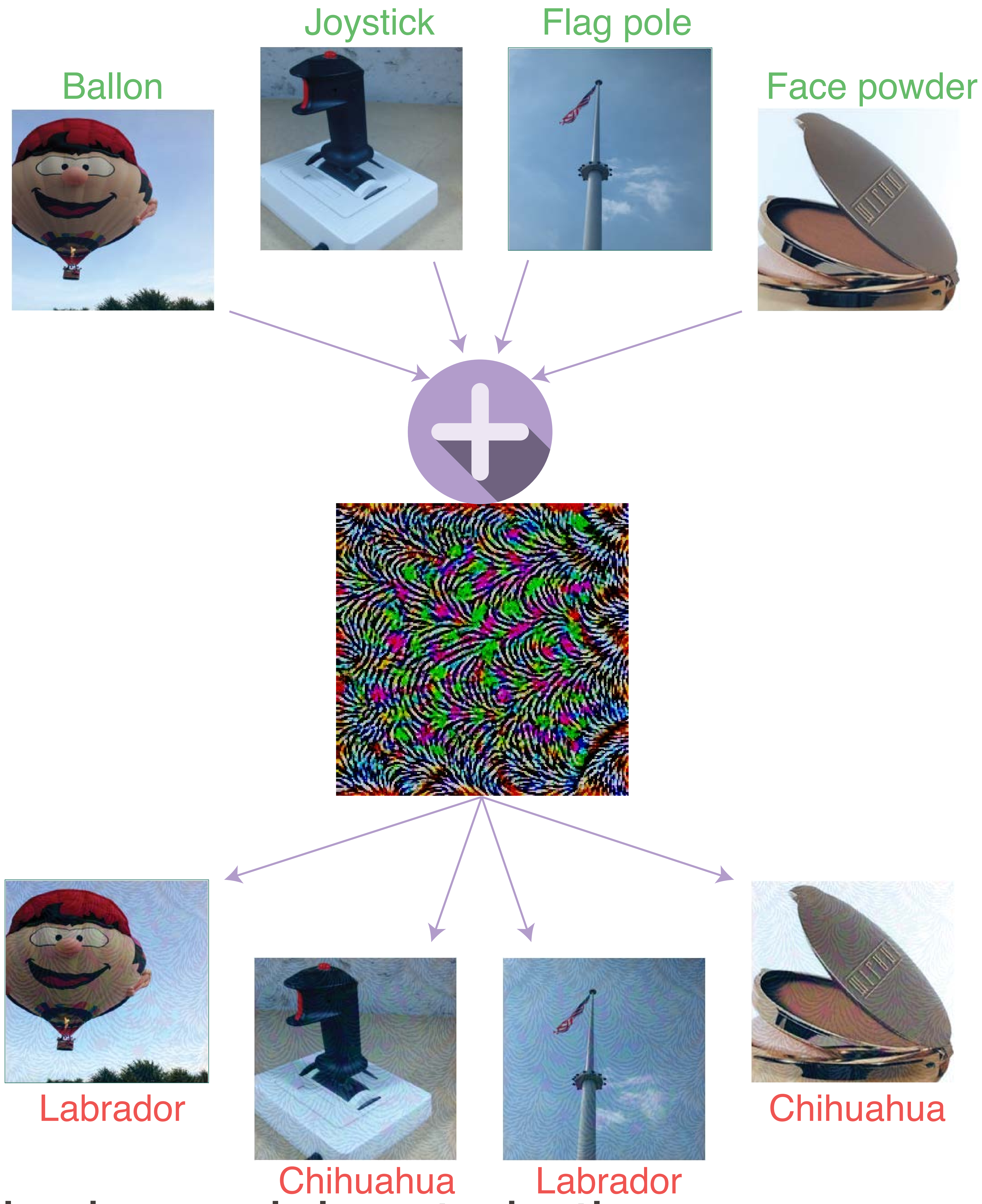
David Paul Morris—Bloomberg/Getty Images

Are we ready?



- Intriguing properties of neural networks, Szegedy et al., *ICLR 2014.*

Adversarial perturbations



- Universal adversarial perturbations, Moosavi et al., *CVPR 2017.*

Universal (adversarial) perturbations

“Geometry is not true, it is advantageous.”

Henri Poincaré

Adversarial perturbations

How large is the “space” of adversarial examples?

Universal perturbations

What causes the vulnerability of deep networks to universal perturbations?

Adversarial training

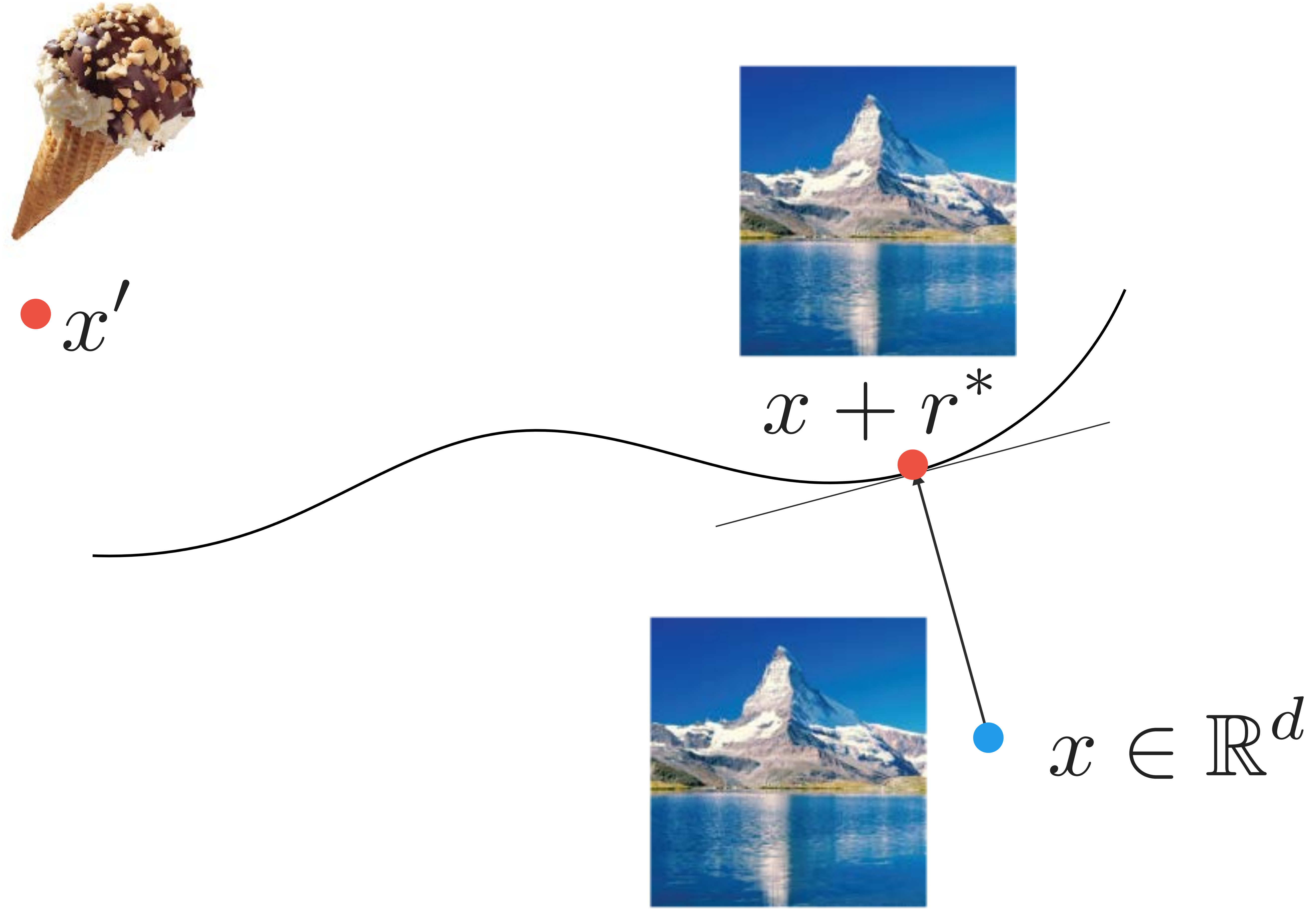
What geometric features contribute to a better robustness properties?

Geometry of …

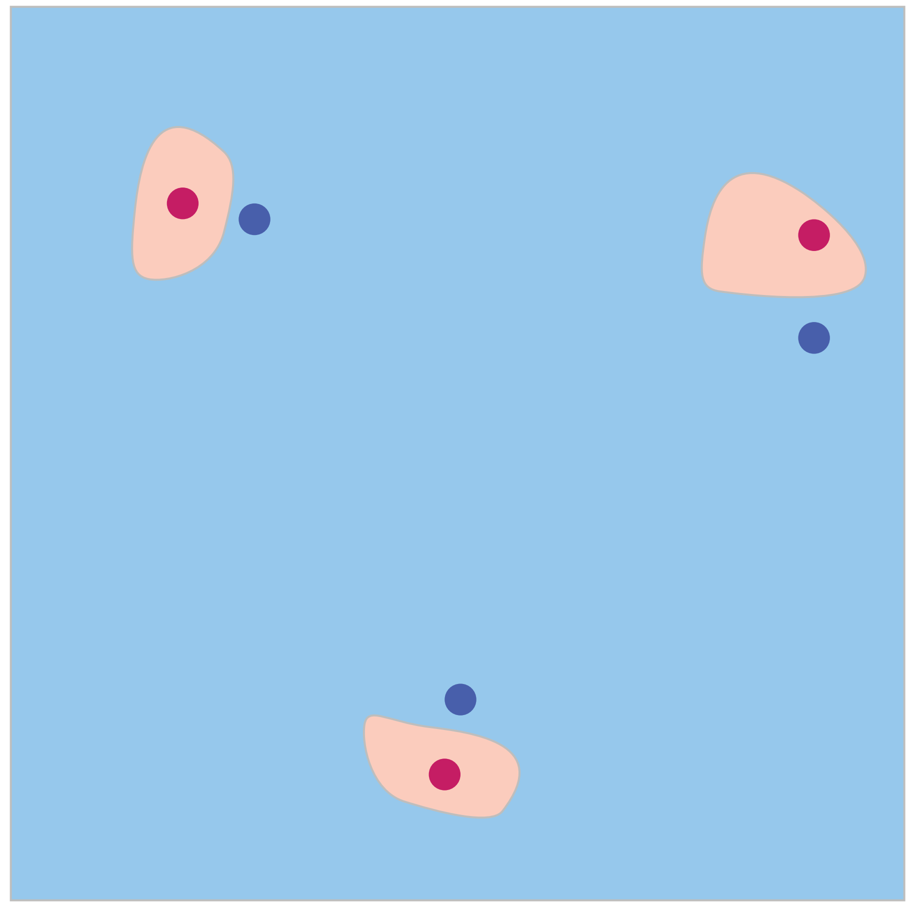
Geometry of adversarial perturbations

Geometric interpretation of adversarial perturbations

$$r^* = \underset{r}{\operatorname{argmin}} \|r\|_2 \text{ s.t. } \hat{k}(x+r) \neq \hat{k}(x)$$

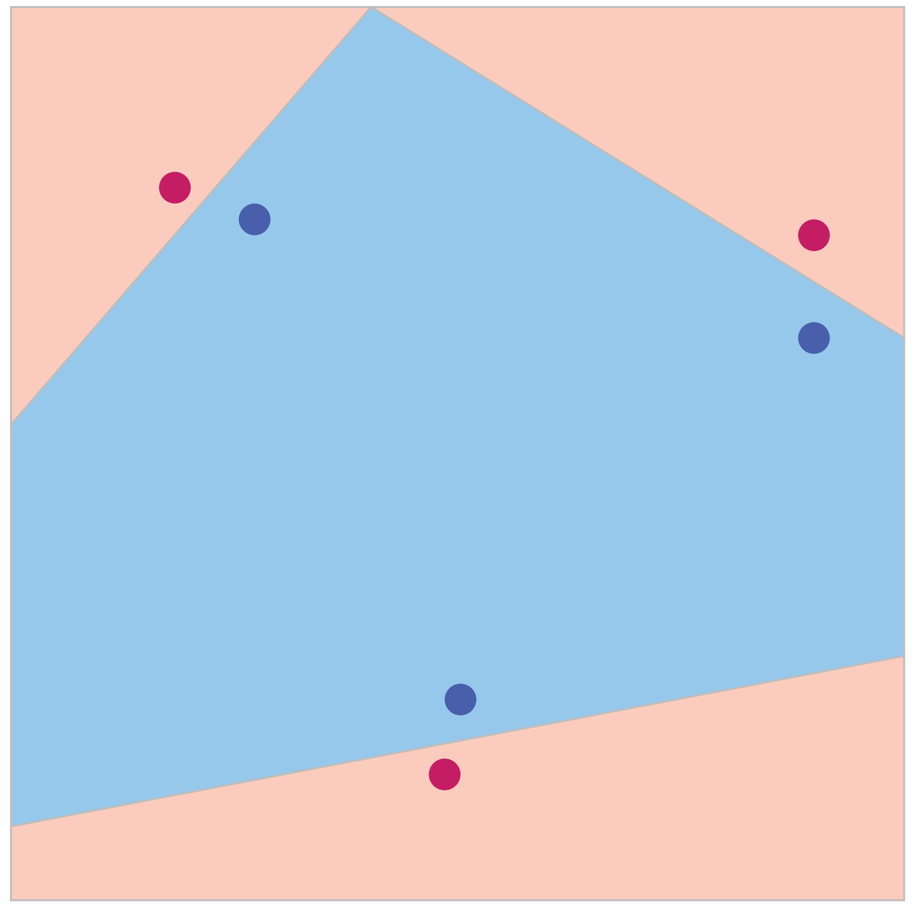


Adversarial examples are “blind spots”.



- Intriguing properties of neural networks, Szegedy et al., *ICLR 2014*.

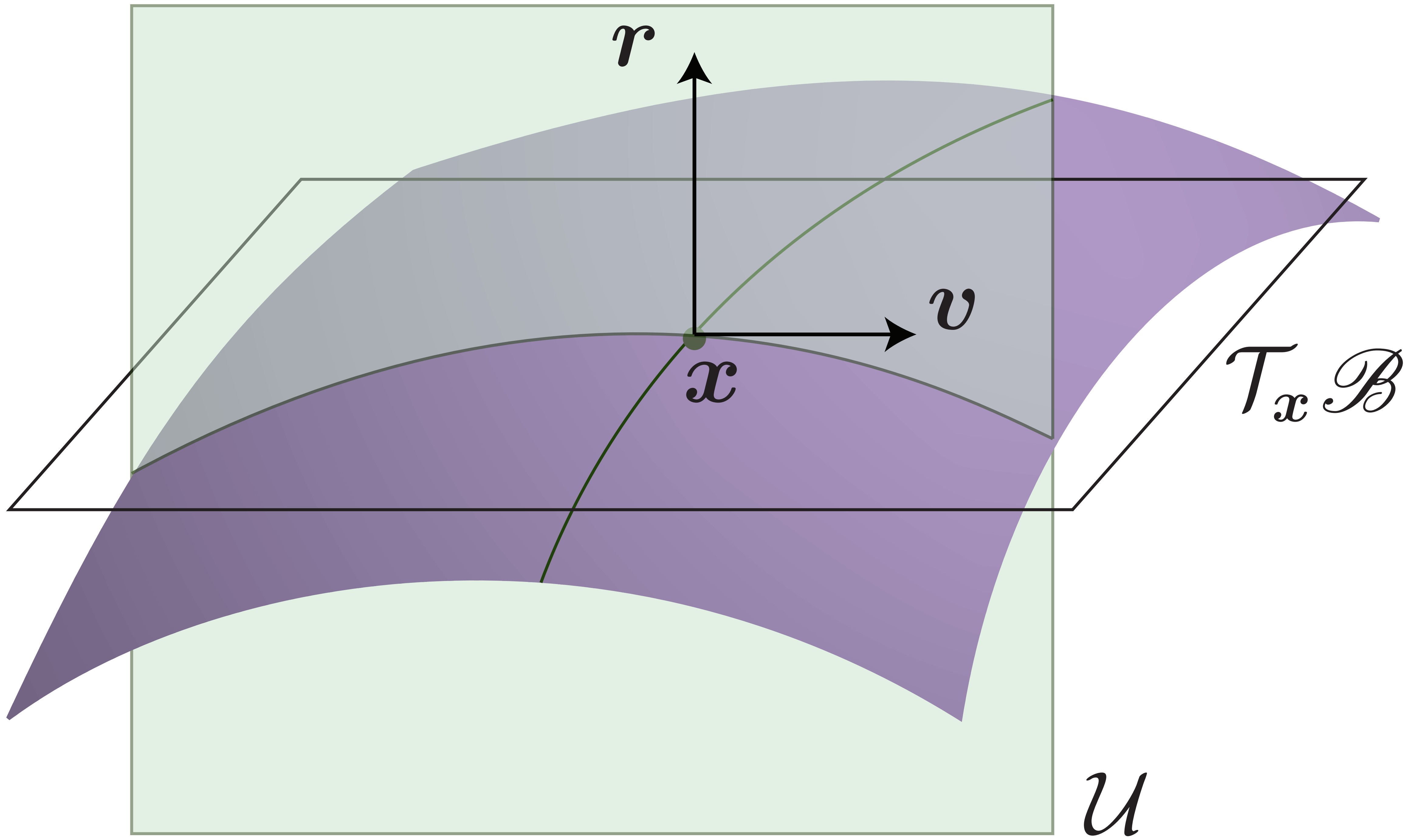
Deep classifiers are “too linear”.



- Explaining and harnessing adversarial examples, Goodfellow et al., *ICLR 2015*.

Two hypotheses

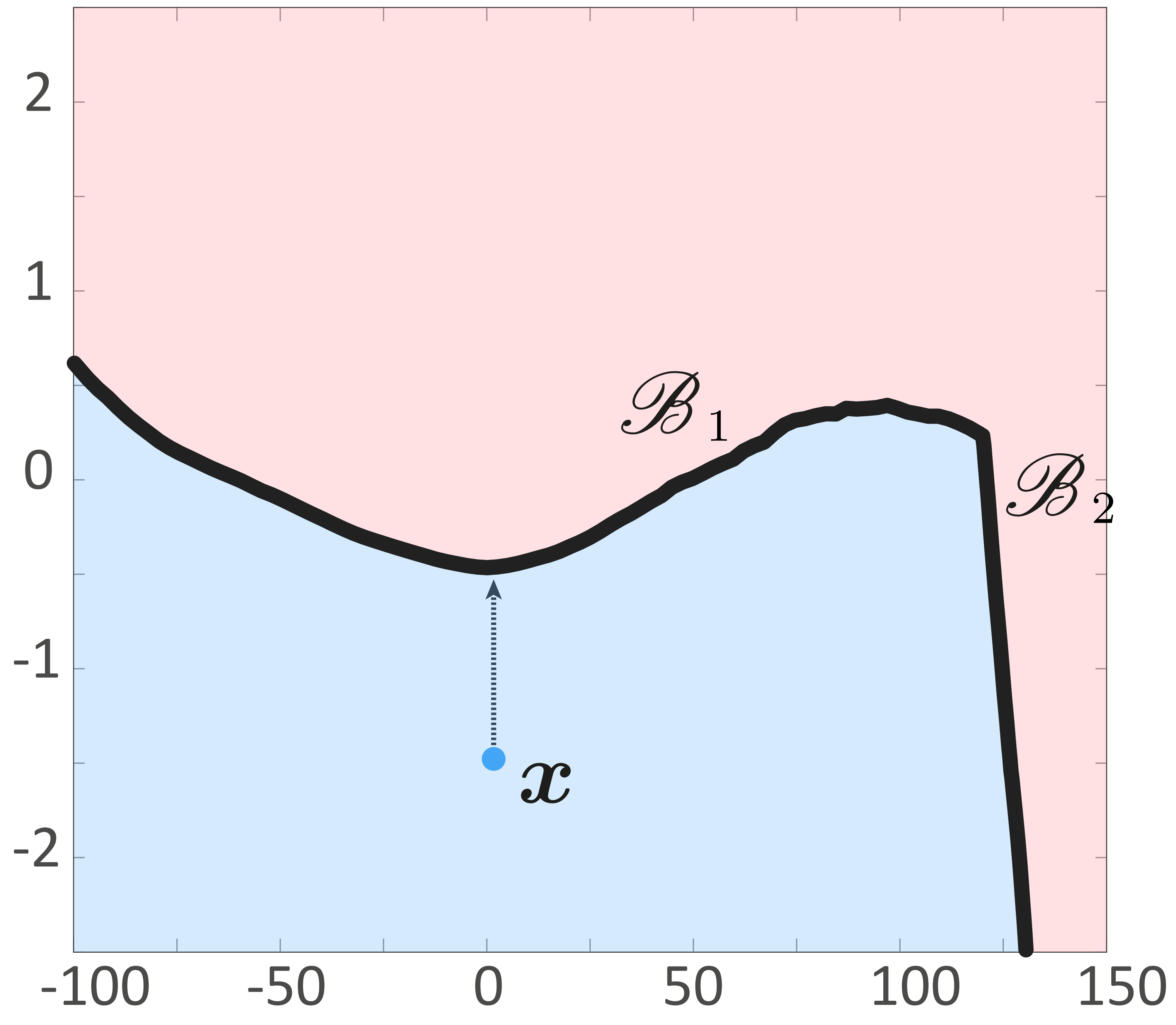
Normal cross-sections of decision boundary



- Robustness of classifiers: from adversarial to random noise, Fawzi, **Moosavi**, Frossard, *NIPS 2016*.

Curvature of decision boundary of deep nets

Decision boundary of CNNs is almost flat along *random* directions.



- Robustness of classifiers: from adversarial to random noise, Fawzi, **Moosavi**, Frossard, *NIPS 2016*.

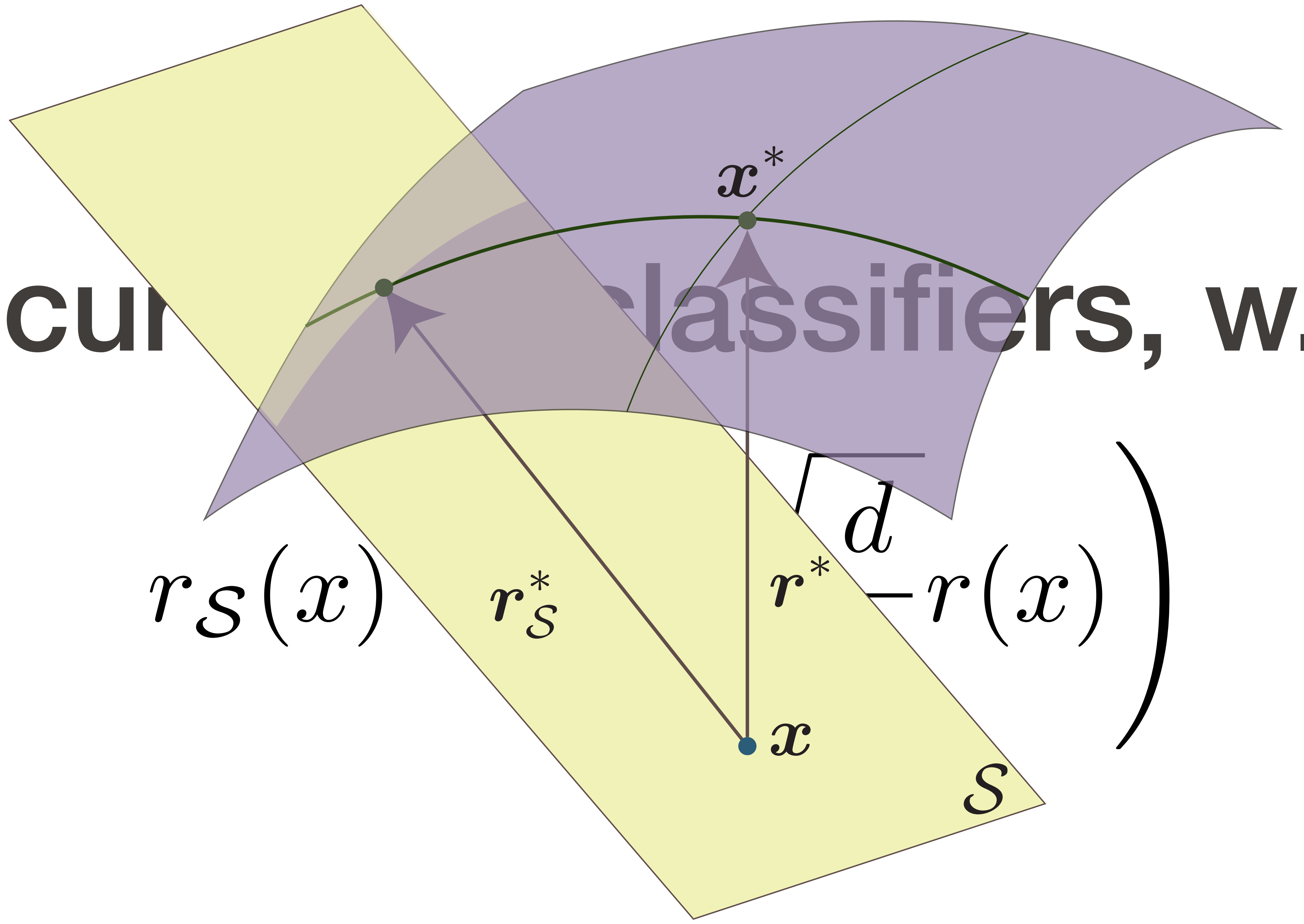
Space of adversarial perturbations

Adversarial perturbations constrained to a random subspace of dimension m.

$$r_{\mathcal{S}}(x) = \arg\min_{r \in \mathcal{S}} \|r\| \text{ s.t. } \hat{k}(x+r) \neq \hat{k}(x)$$

For low cur… classifiers, w.h.p., we have

$$r_{\mathcal{S}}(x) \;\ldots\; \sqrt{\frac{d}{\ldots}}\, r(x) \Big)$$



Structured additive perturbations

The “space” of adversarial examples is quite vast.

Flowerpot

Pineapple

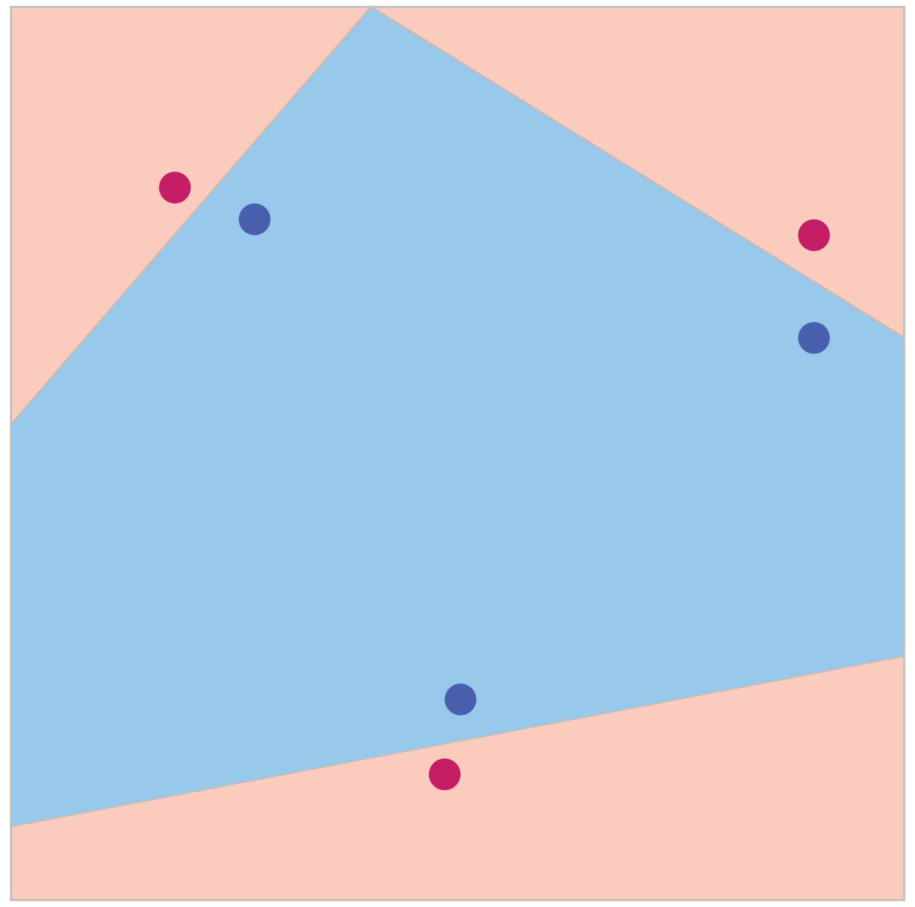
- Robustness of classifiers: from adversarial to random noise, Fawzi, **Moosavi**, Frossard, *NIPS 2016*.

Summary

Geometry of adversarial examples

Decision boundary is “locally” almost flat.

Datapoints lie close to the decision boundary.

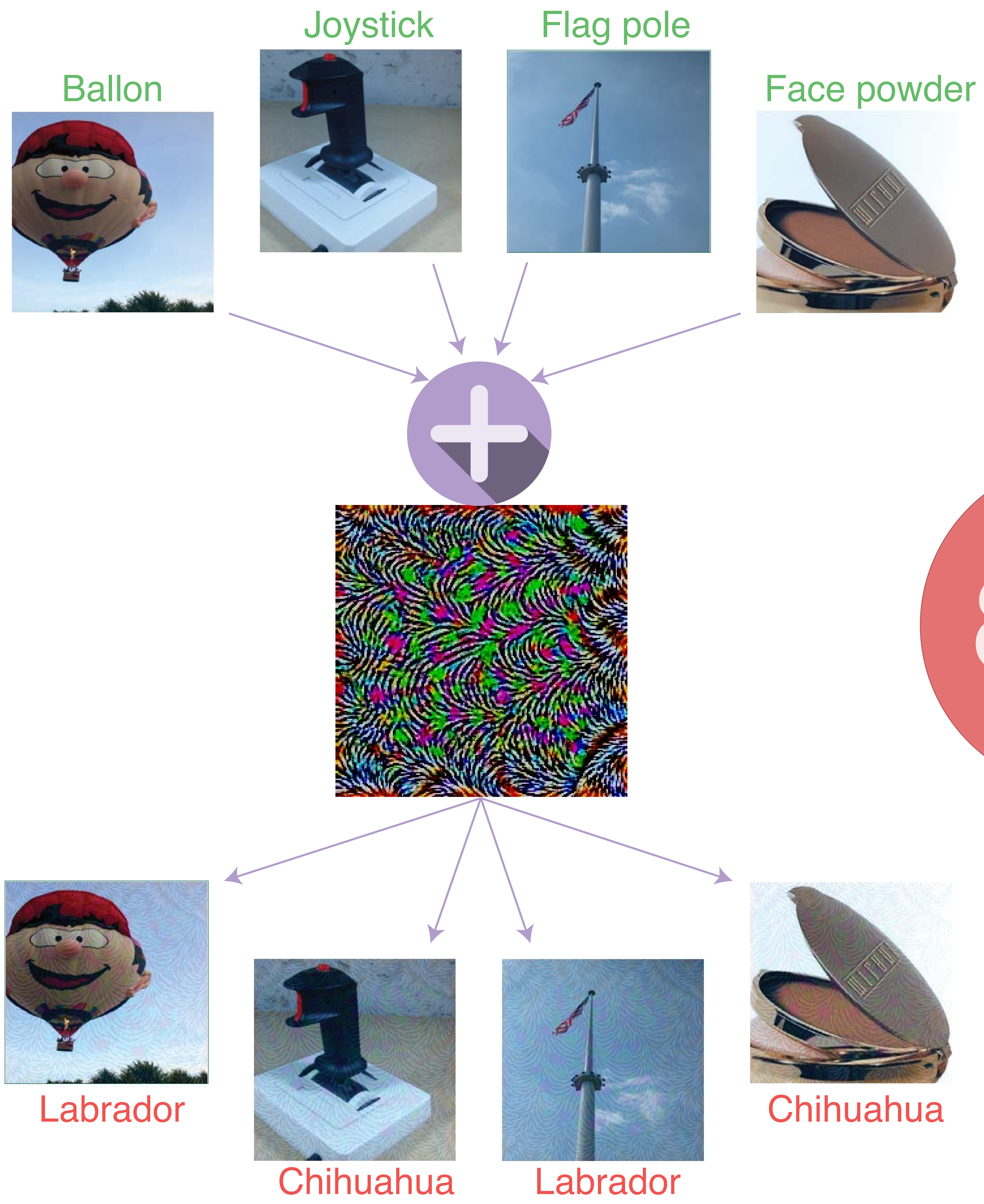


Flatness can be used to

construct diverse set of perturbations.

design efficient attacks.

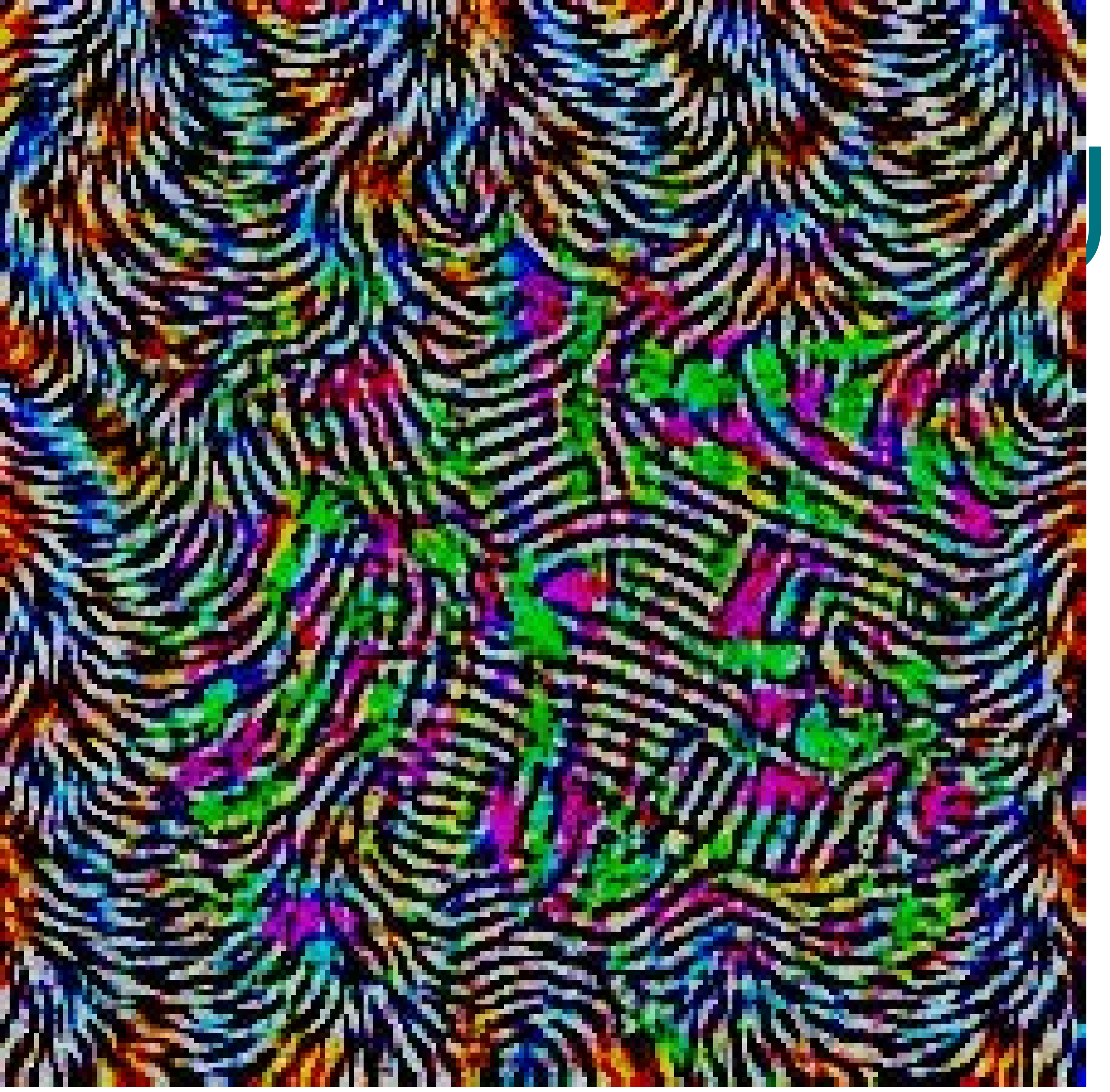
Geometry of universal perturbations



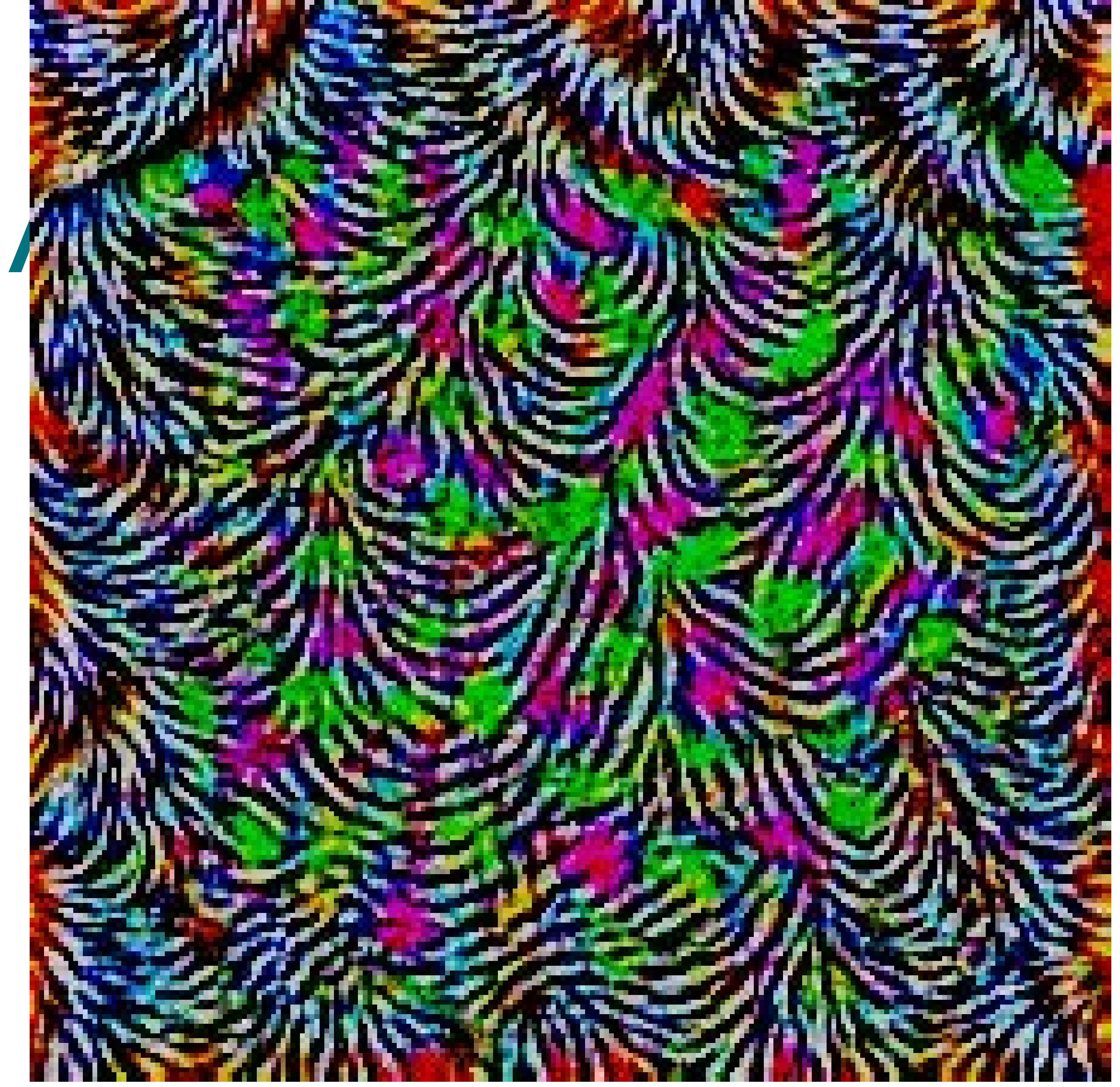
85%

- Universal adversarial perturbations, Moosavi et al., *CVPR 2017.*

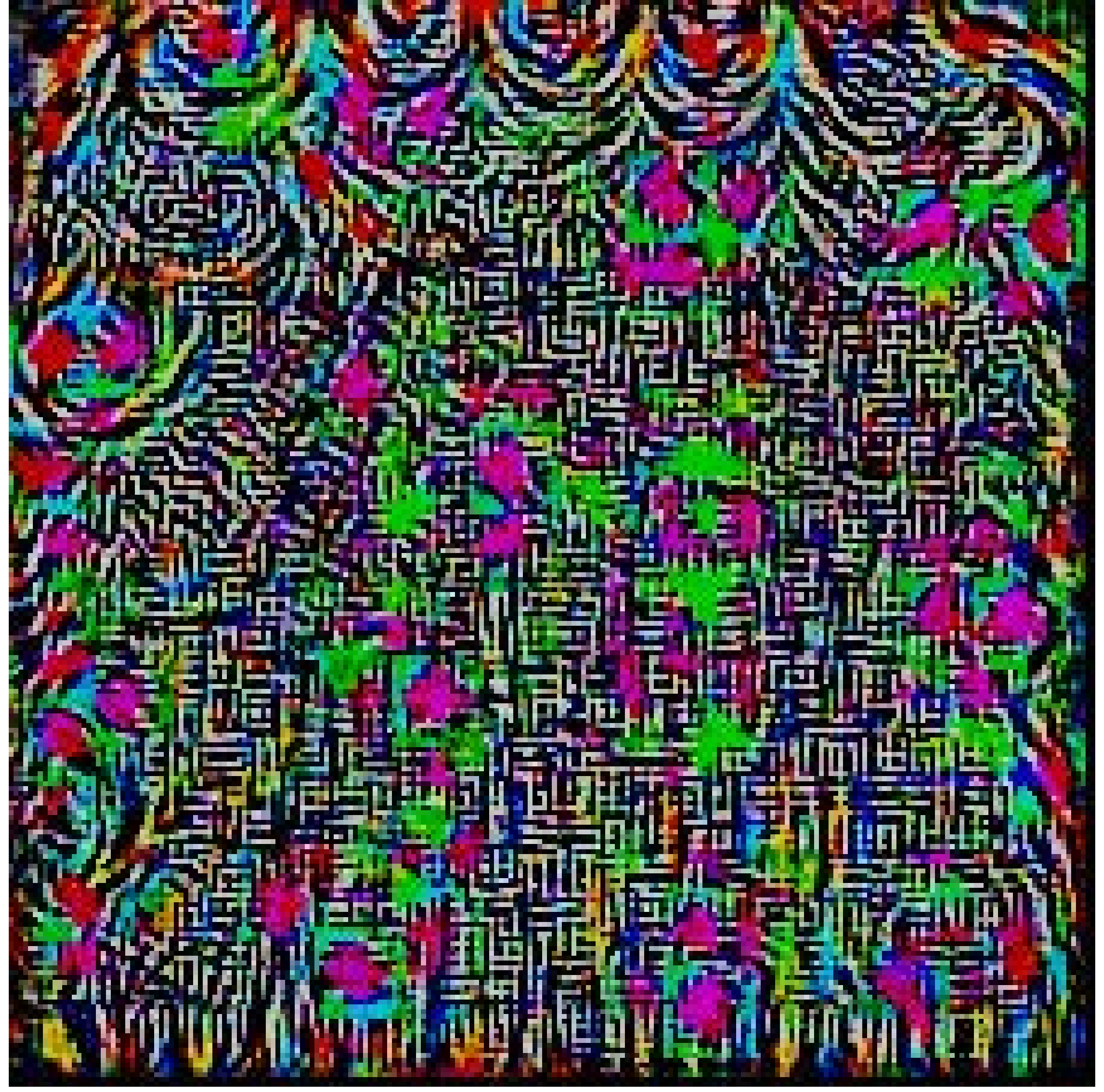
Universal adversarial perturbations (UAP)



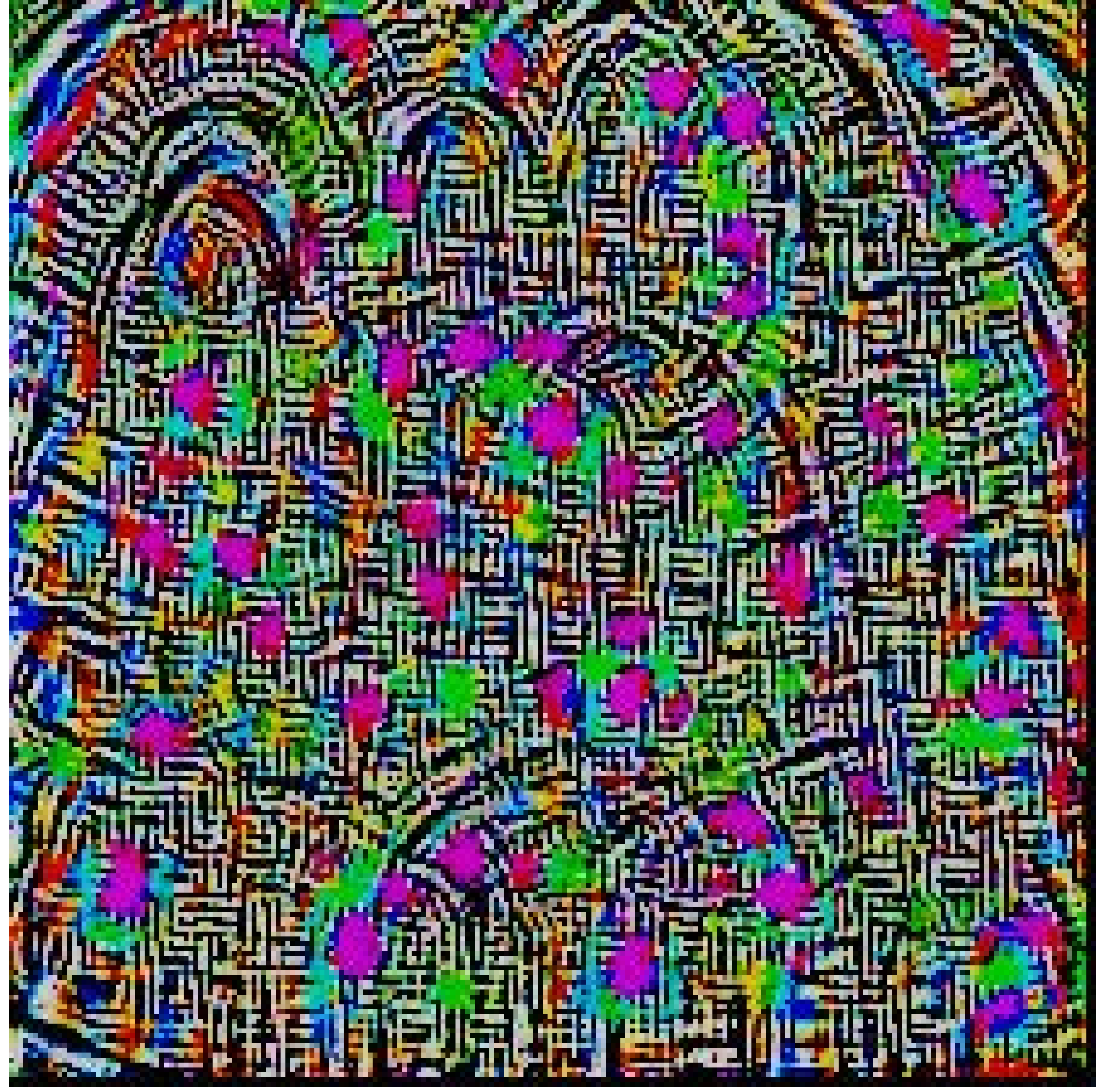
VGG-19



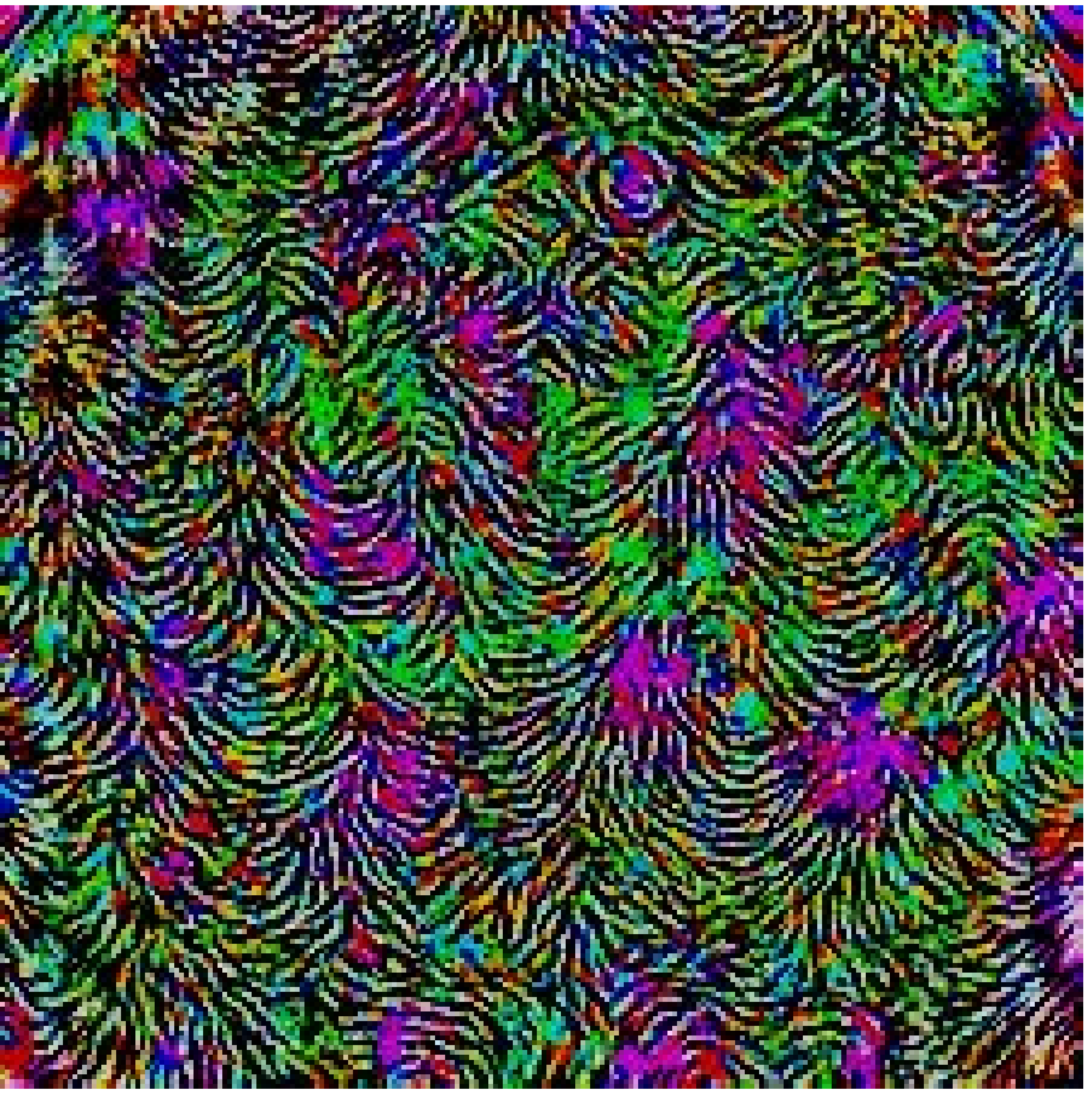
VGG-16



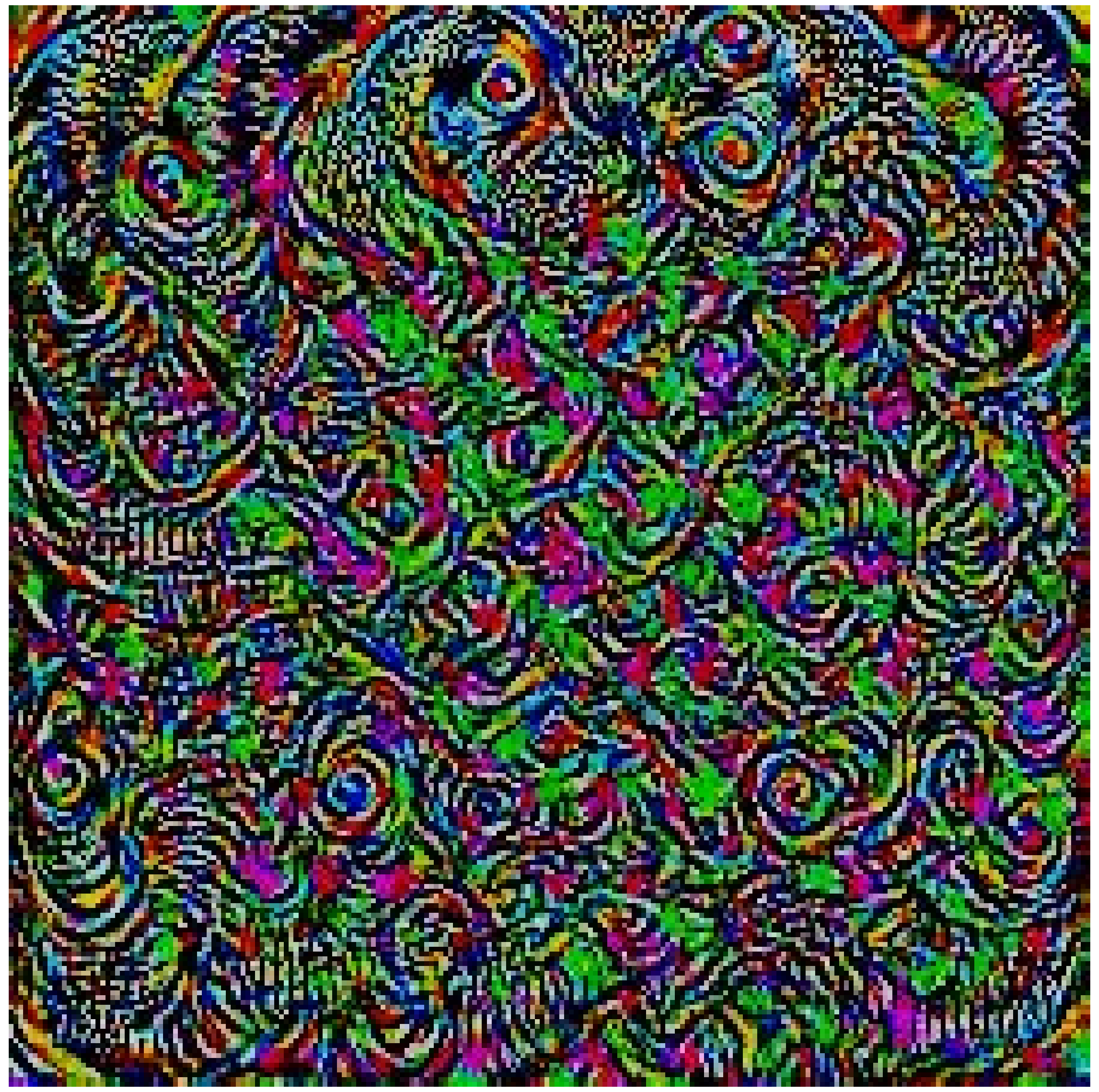
VGG-F



CaffeNet



ResNet-152



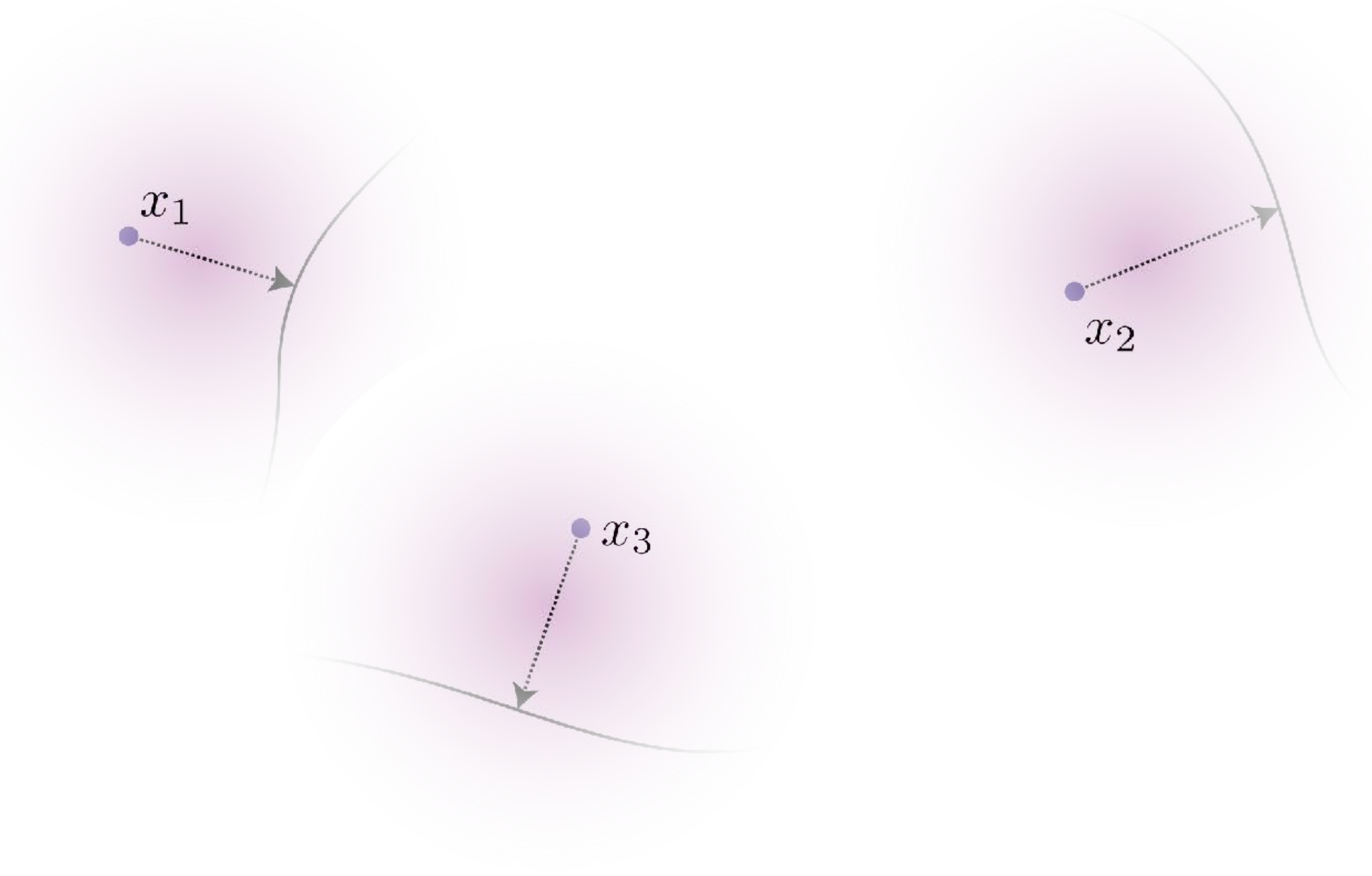
GoogLeNet

Diversity of perturbations

Flat model

Curved model

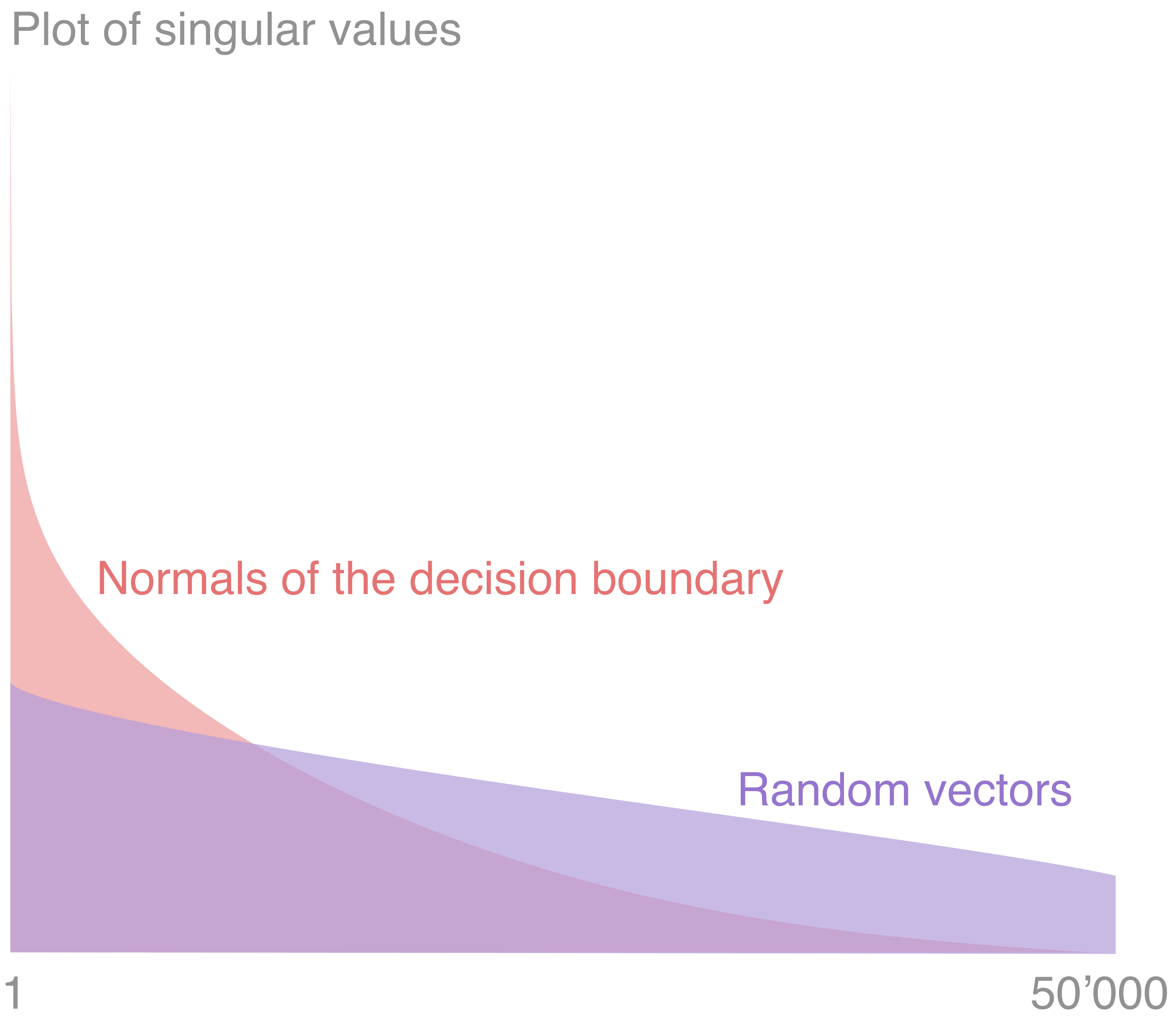
Why do universal perturbations exist?



- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Flat model

Normals to the decision boundary are "globally" correlated.



- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Flat model (cont'd)

The flat model only partially explains the universality.

Random

13%

38%

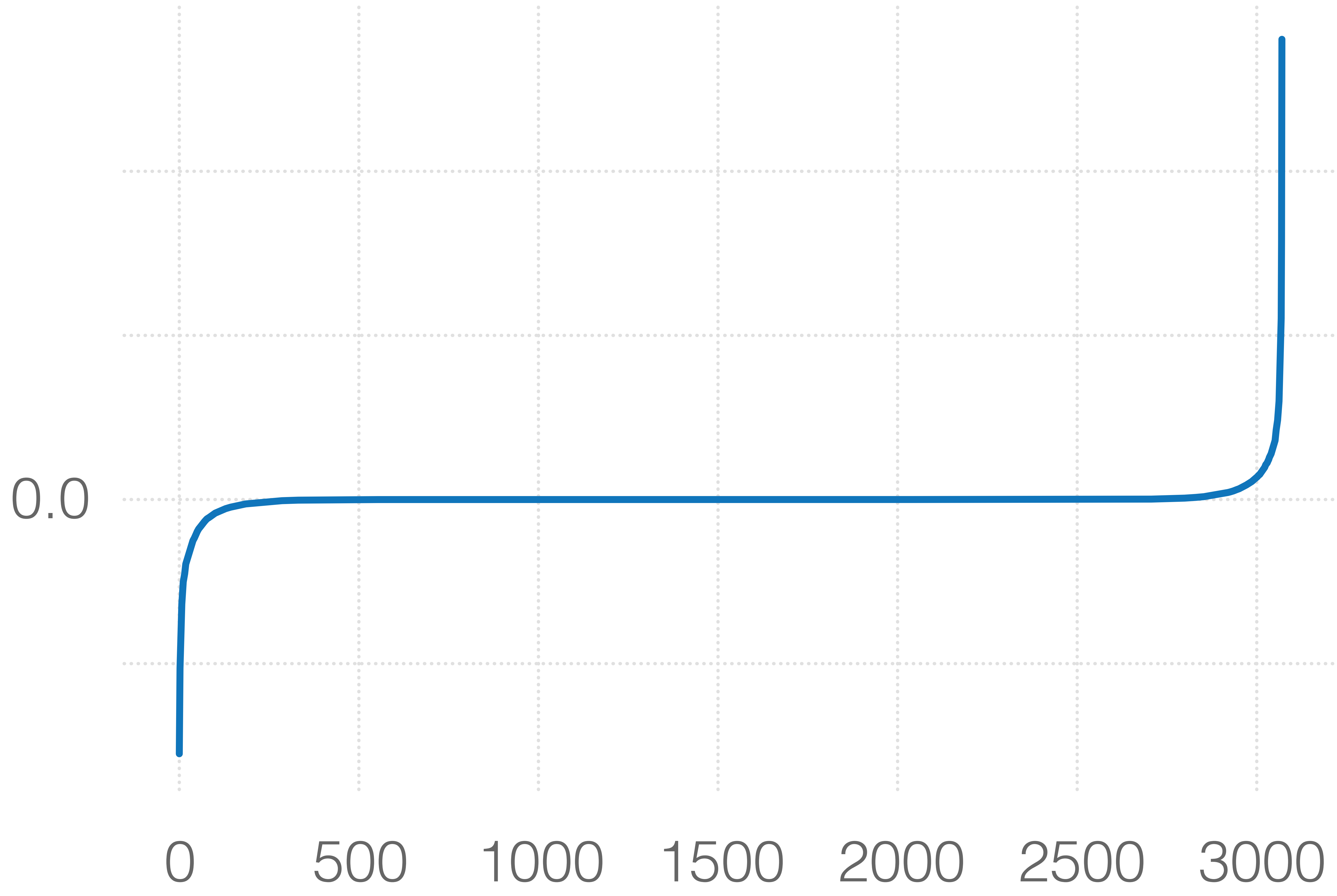
UAP
(greedy algorithm)

85%

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Flat model (cont'd)

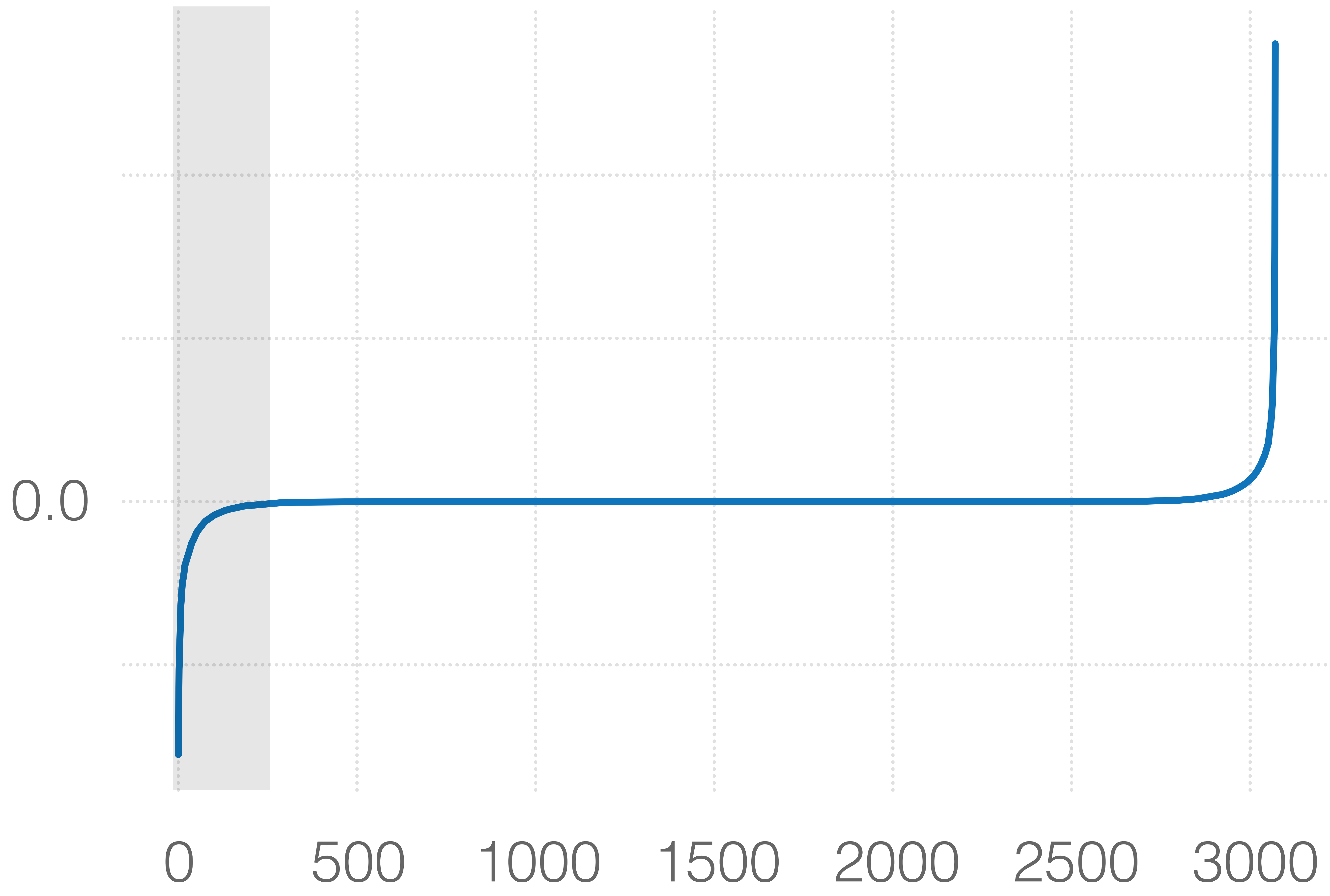
The principal curvatures of the decision boundary:

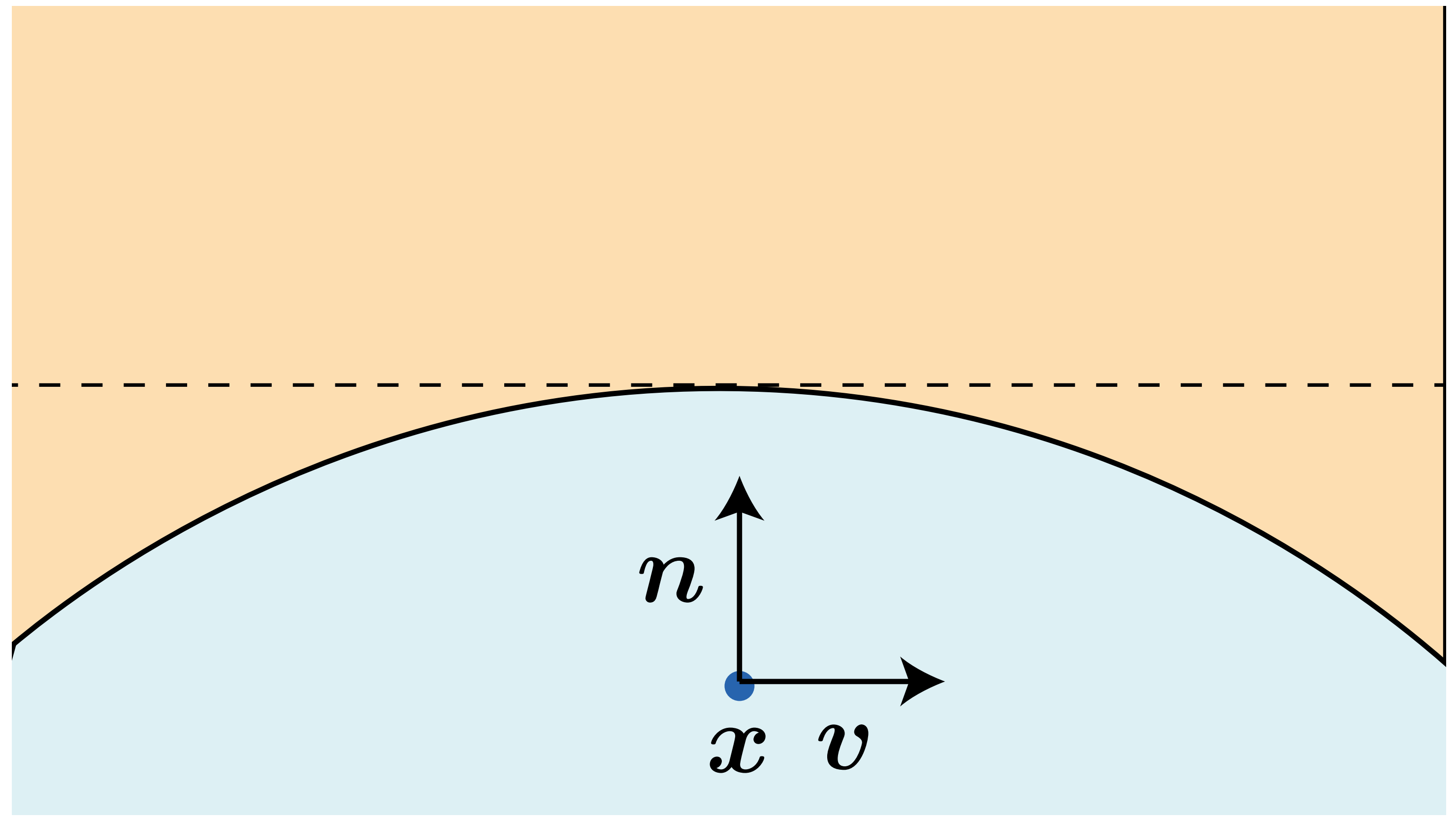


- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Curved model

The principal curvatures of the decision boundary:

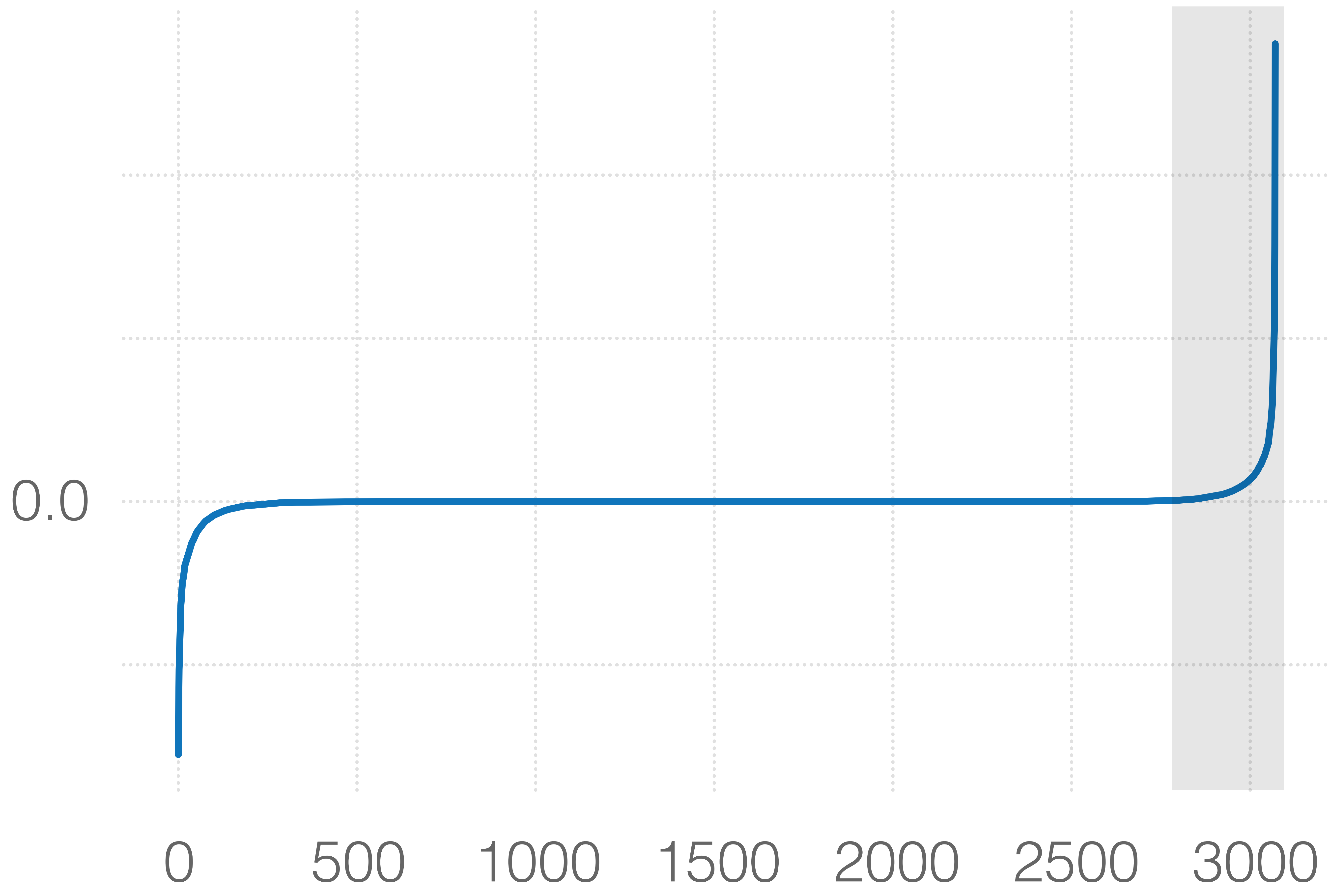


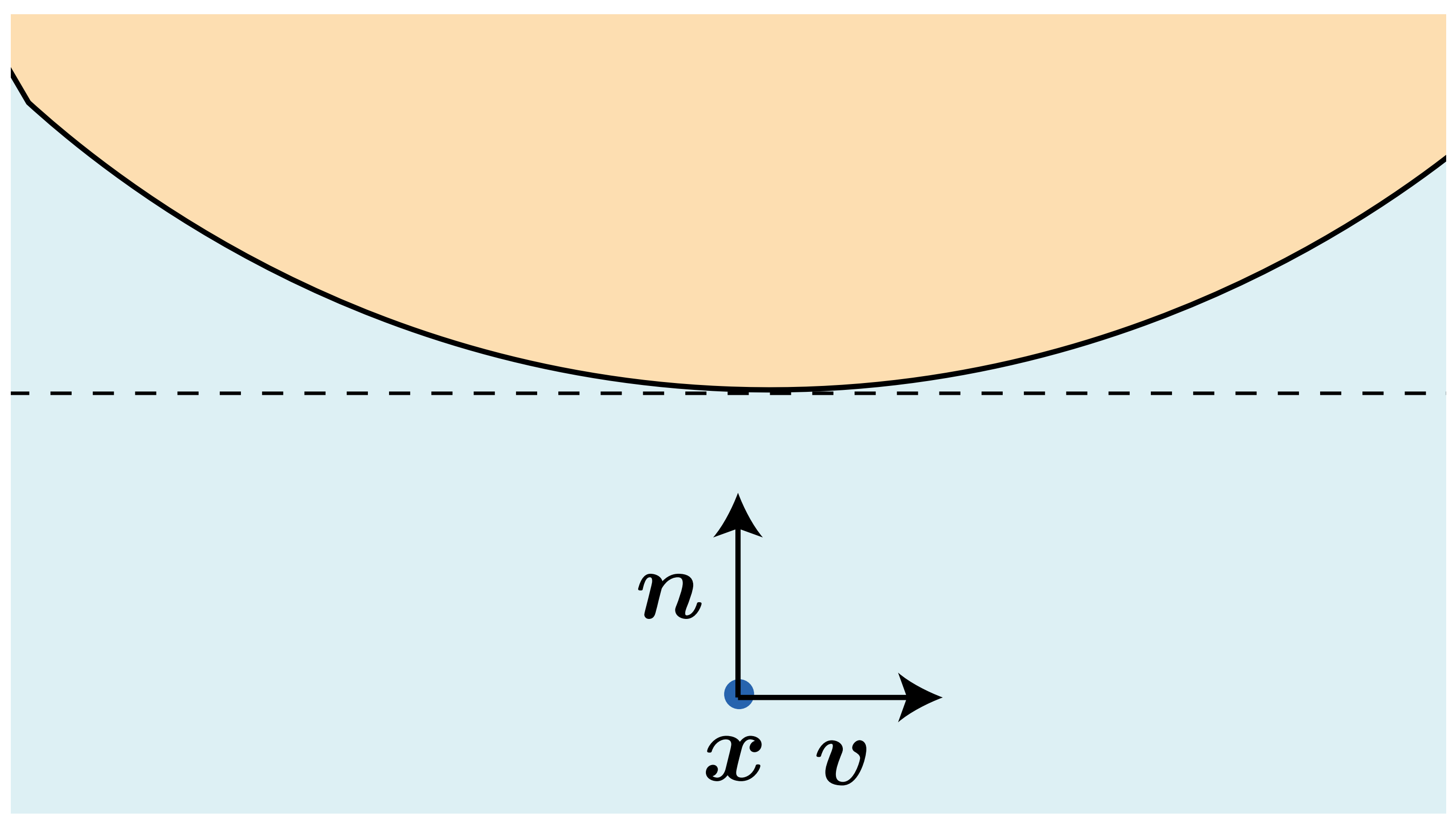


Curved model (cont'd)

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

The principal curvatures of the decision boundary:

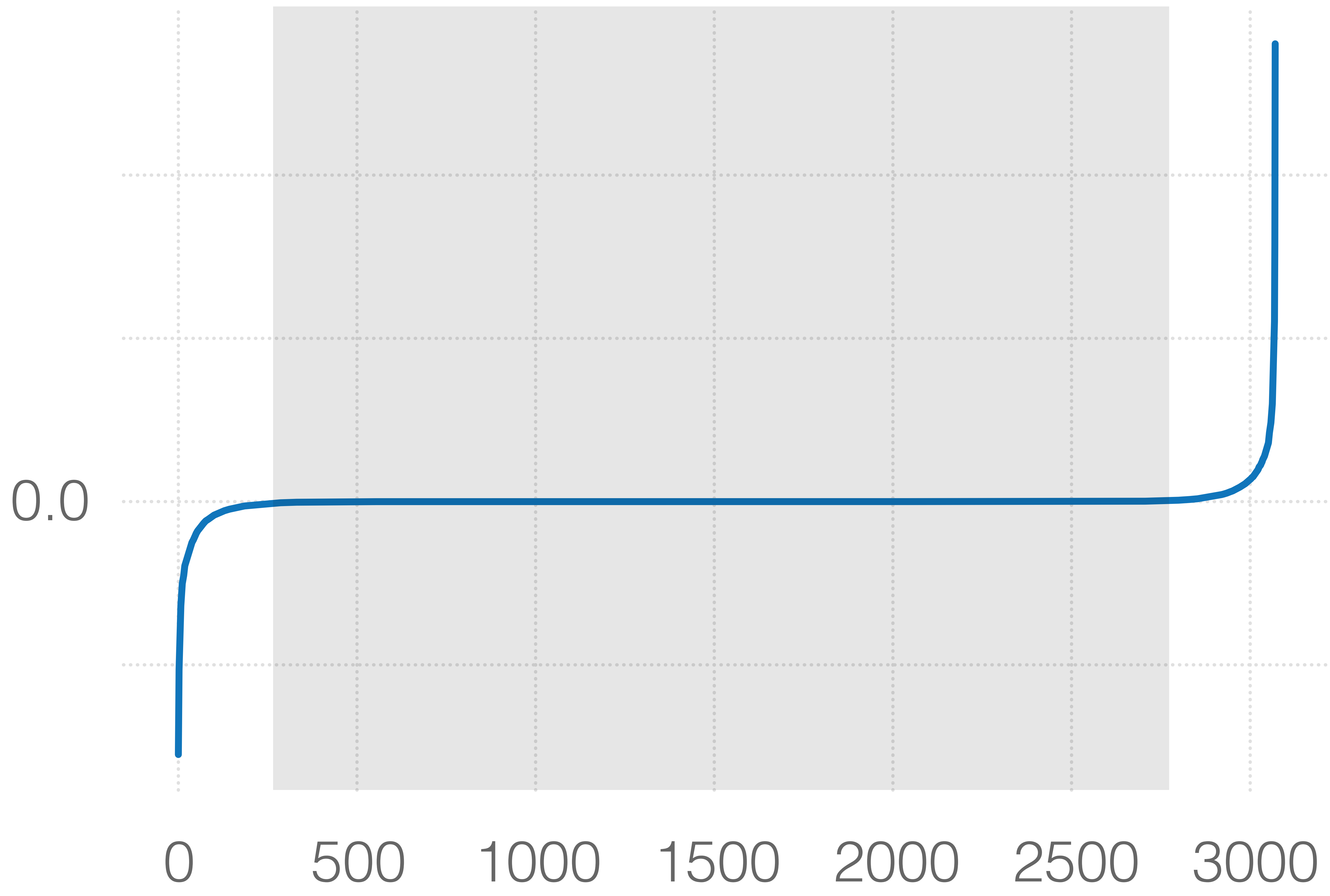


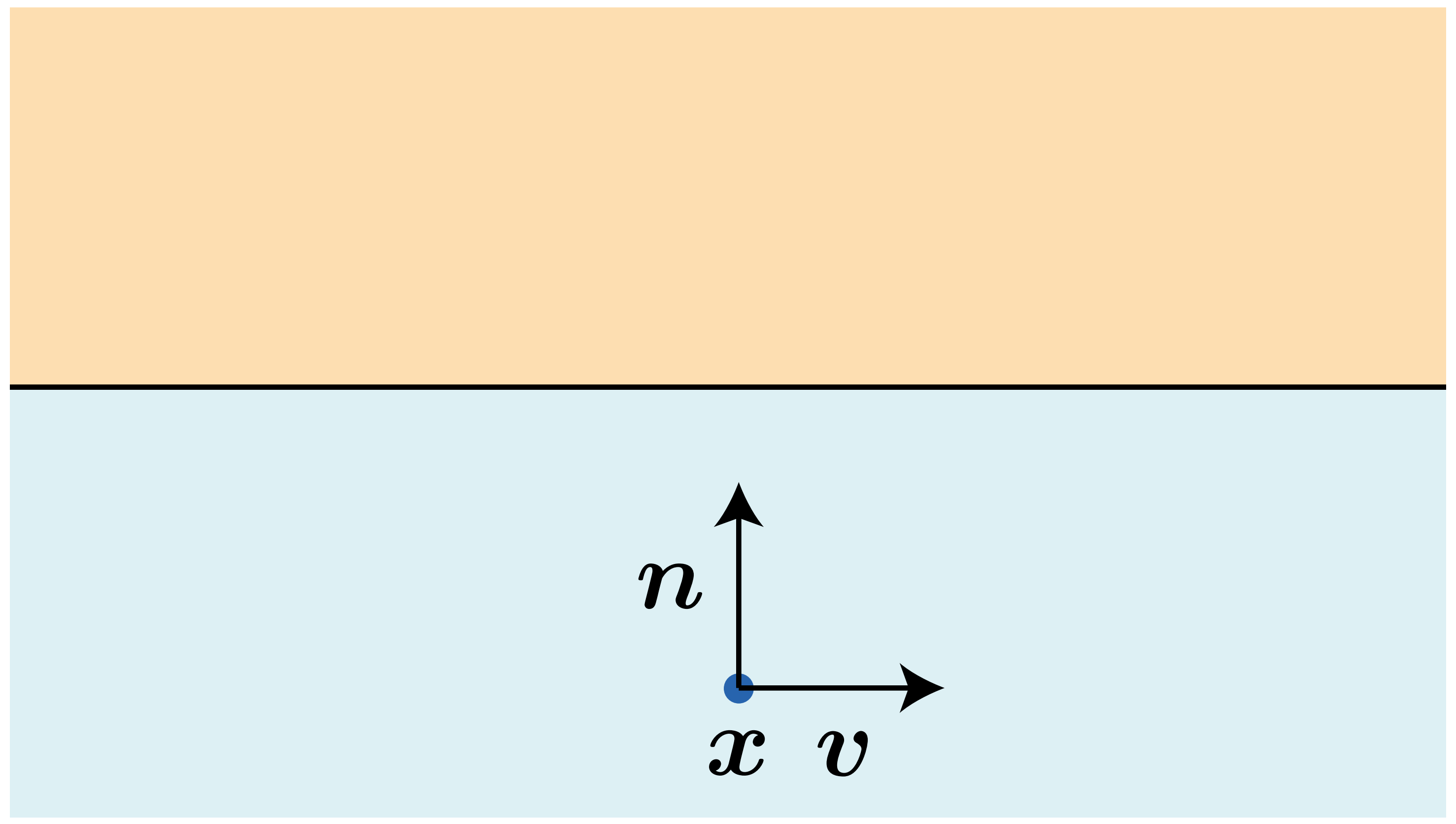


- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Curved model (cont'd)

The principal curvatures of the decision boundary:





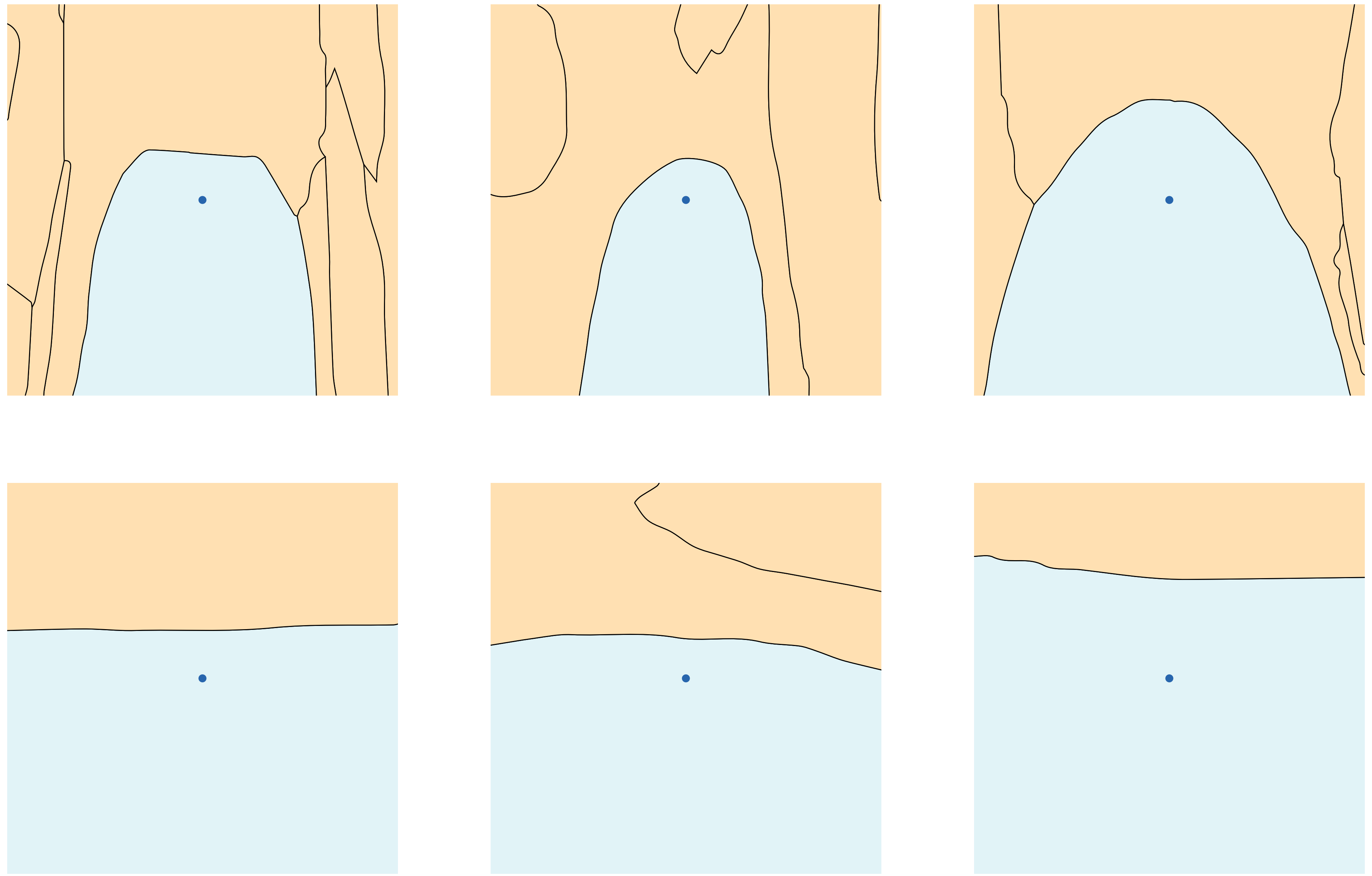
- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Curved model (cont'd)

Normal sections of the decision boundary (for different datapoints) along a single direction:

UAP direction

Random direction



- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Curved directions are shared

The curved model better explains the existence of universal perturbations.

Random	Flat model	Curved model	UAP
13%	38%	67%	85%

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*

Curved directions are shared (cont'd)

Universality of perturbations

Shared curved directions explain this vulnerability.

A possible solution

Regularizing the geometry to combat against universal perturbations.

Why are deep nets curved?

- With friends like these, who needs adversaries?, Jetley et al., *NeurIPS 2018.*

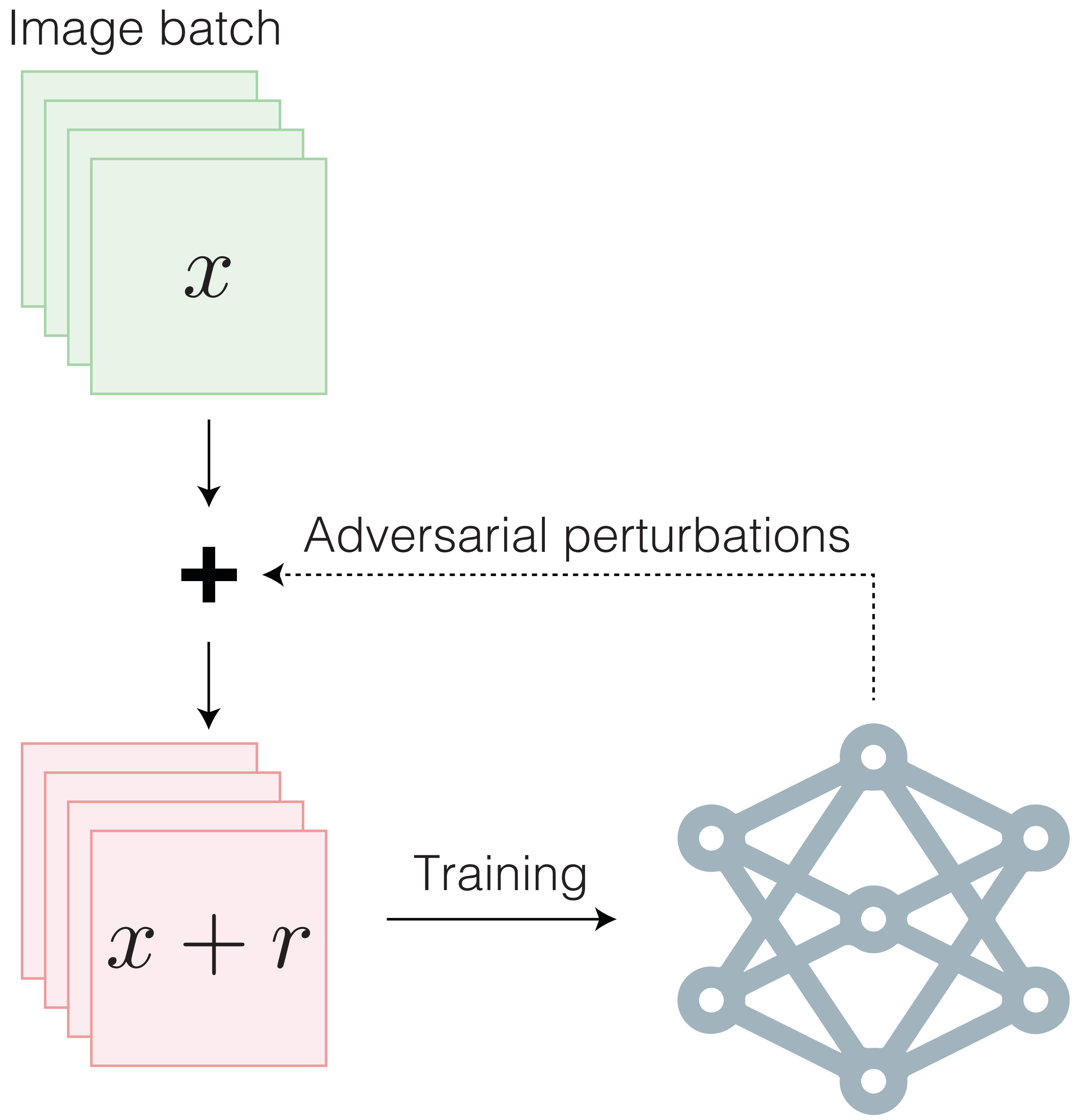
Summary

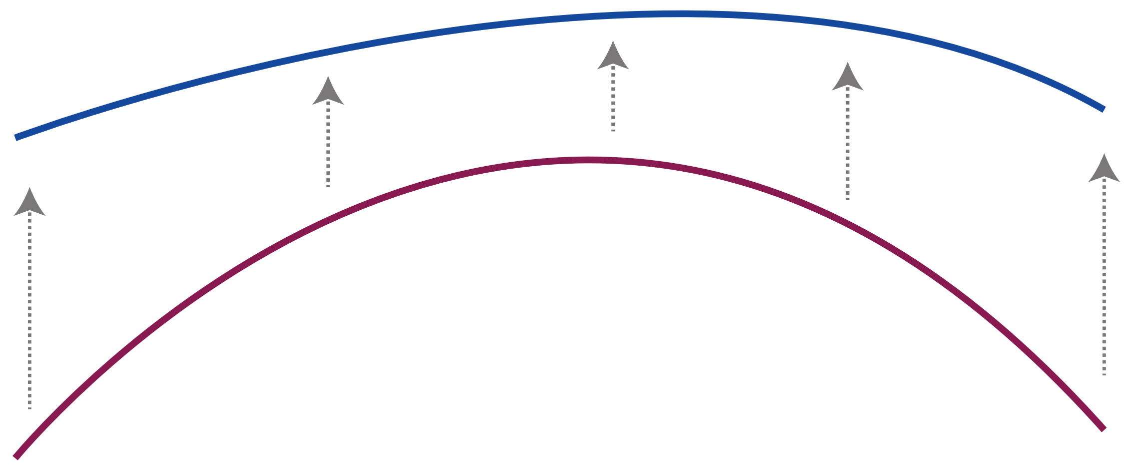
Geometry of adversarial training

In a nutshell

Adversarial training

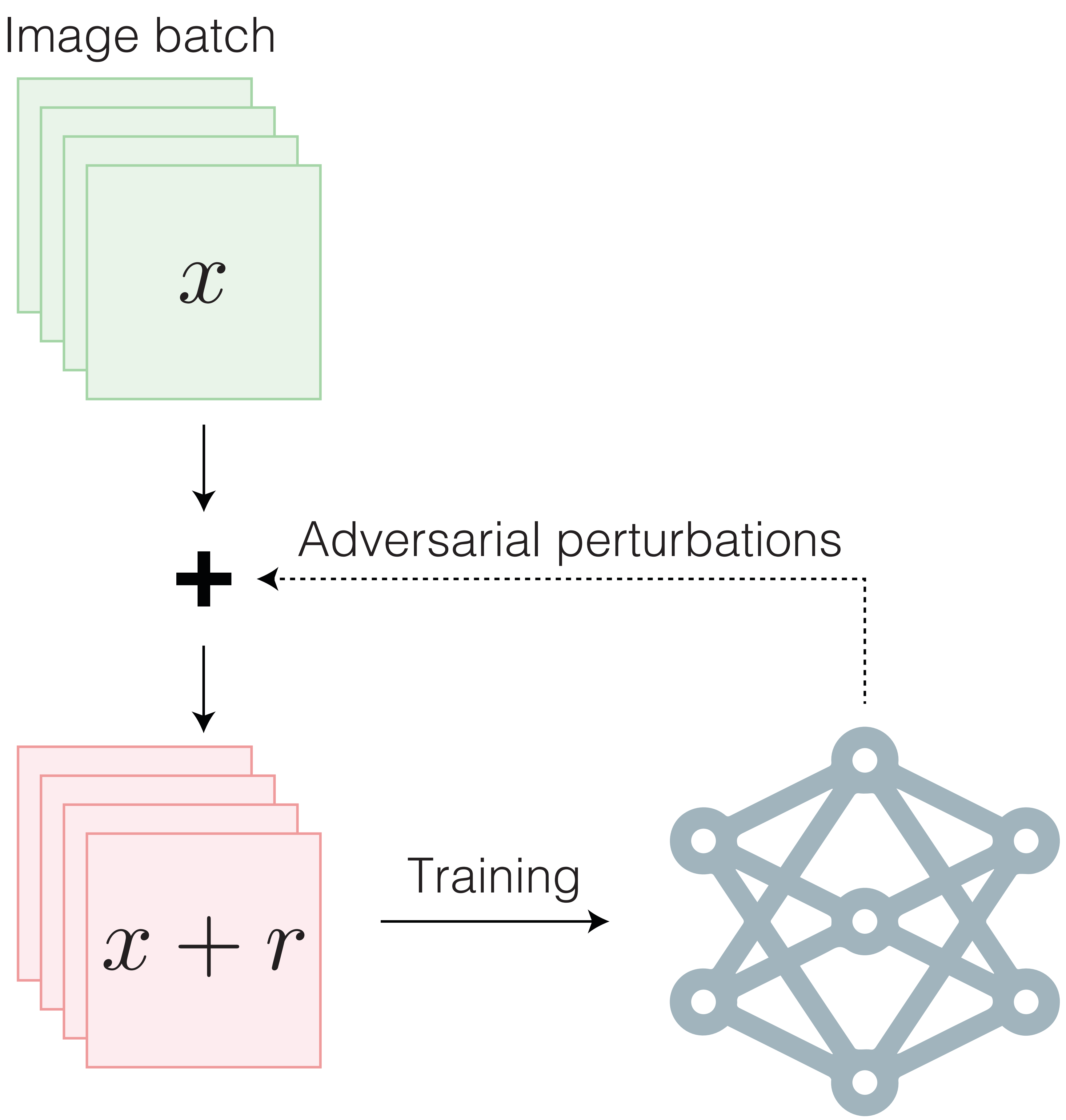
Curvature regularization





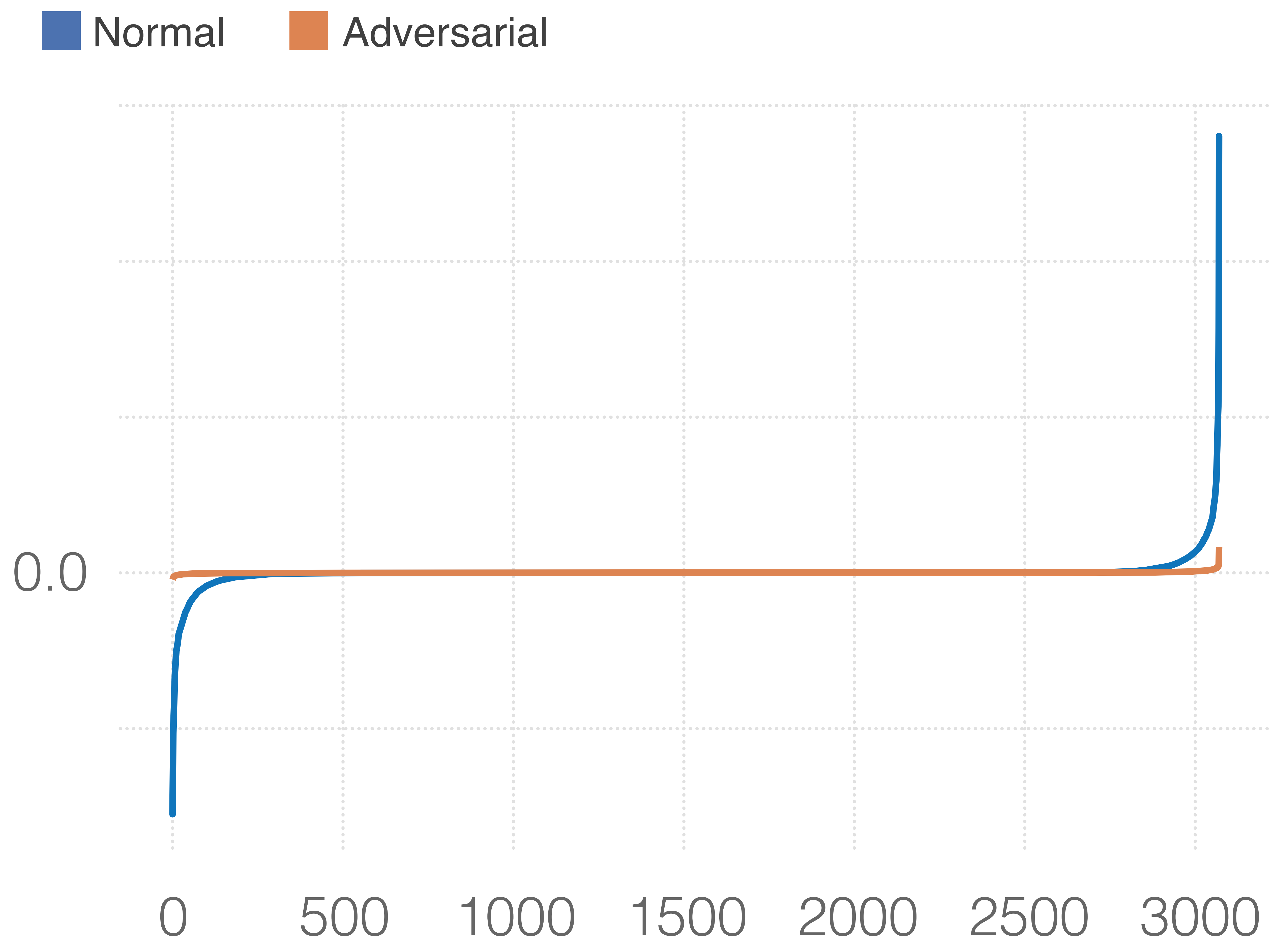
Adversarial training

One of the most effective methods to improve adversarial robustness…



- Obfuscated gradients give a false sense of security, Athalye et al., *ICML 2018*. (Best paper)

Curvature profiles of normally and adversarially trained networks:



- Robustness via curvature regularisation, and vice versa, Moosavi et al., *CVPR 2019*.

Geometry of adversarial training

Curvature Regularization (CURE)

	Normal training	CURE	Adversarial training
Clean	94.9%	81.2%	79.4%
PGD with $\|r^*\|_\infty = 8$	0.0%	36.3%	43.7%

- Robustness via curvature regularisation, and vice versa, Moosavi et al., *CVPR 2019*.

AT vs CURE

AT	CURE
Implicit regularization	Explicit regularization
Time consuming	3x to 5x faster
SOTA robustness	On par with SOTA

- Robustness via curvature regularisation, and vice versa, Moosavi et al., *CVPR 2019*.

Summary

Inherently more robust classifiers

Curvature regularization can significantly improve the robustness properties.

Counter-intuitive observation

Due to a more linear nature, an adversarially trained net is “easier” to fool.

A better trade-off?

- Adversarial Robustness through Local Linearization, Qin et al., *arXiv.*

Future challenges

Architectures

Batch-norm, dropout, depth, width, etc.

Data

of modes, convexity, distinguishability, etc.

Training

Batch size, solver, learning rate, etc.

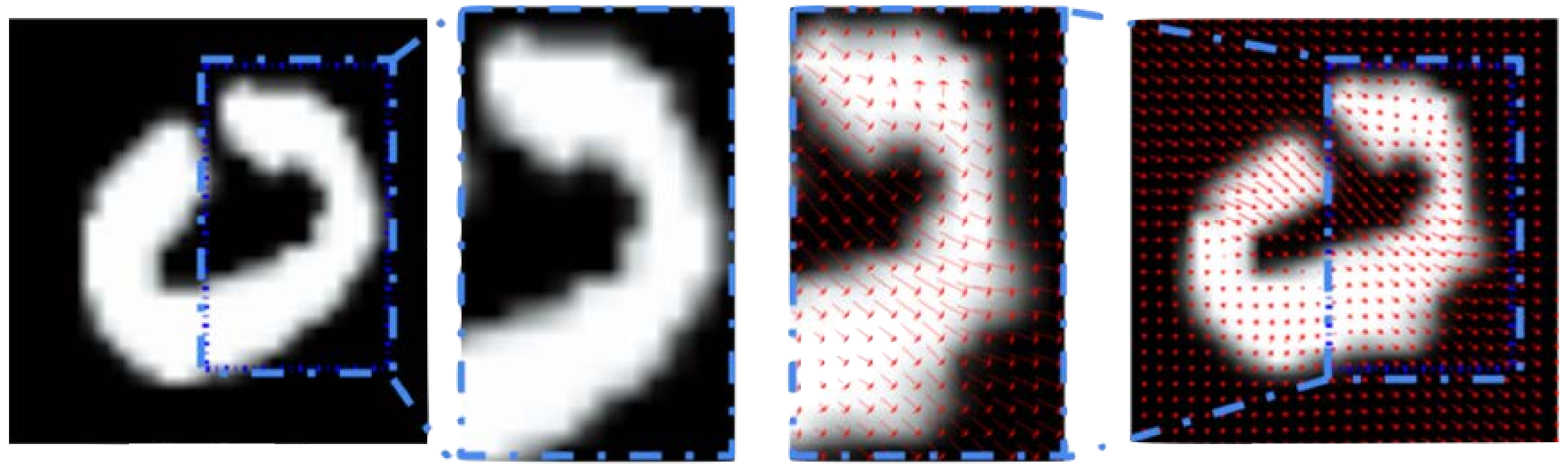
Disentangling different factors

Beyond additive perturbations

- **Geometric robustness of deep networks**, Canbak, **Moosavi**, Frossard, *CVPR 2018*.

Bear Fox

- **Spatially transformed adversarial examples**, Xiao et al., *ICLR 2018*.



“0” “2”

Original image	Standard training	Adversarial training

- Robustness may be at odds with accuracy, Tsipras et al., *NeurIPS 2018.*

“Interpretability” and robustness

ETHZ Zürich, Switzerland

Google Zürich

Interested in my research?

smoosavi.me

moosavi.sm@gmail.com