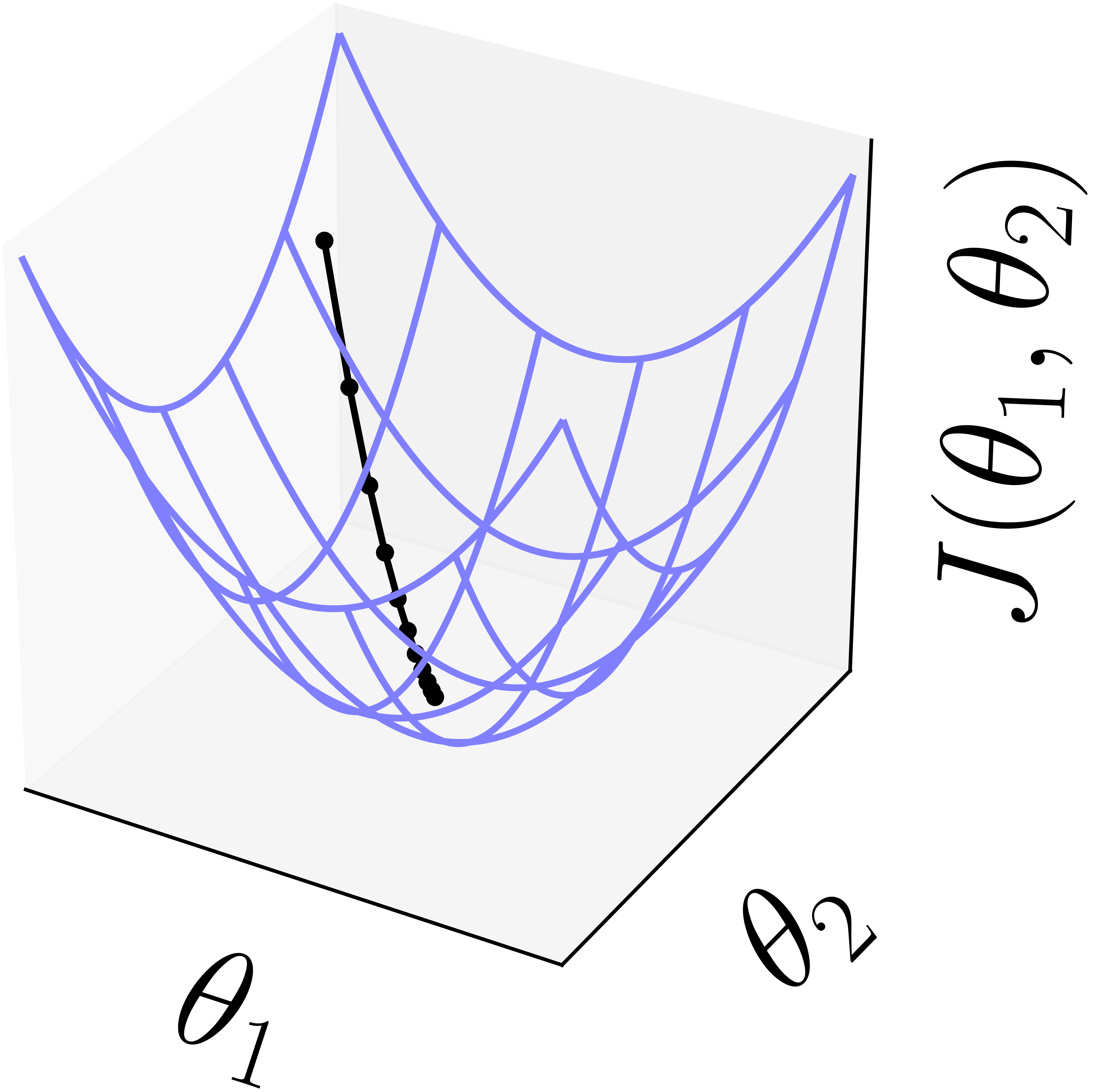
Adversarial Machine Learning

Ian Goodfellow

ICLR

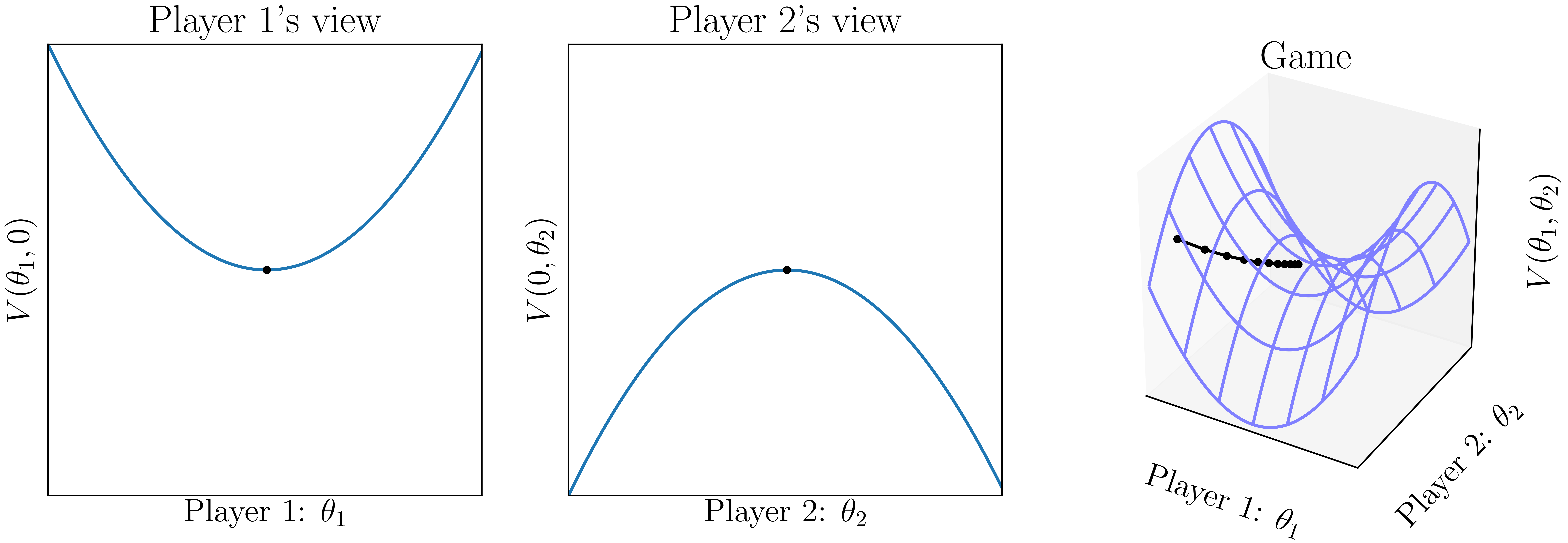
2019-05-07

Most Traditional Machine Learning: Optimization



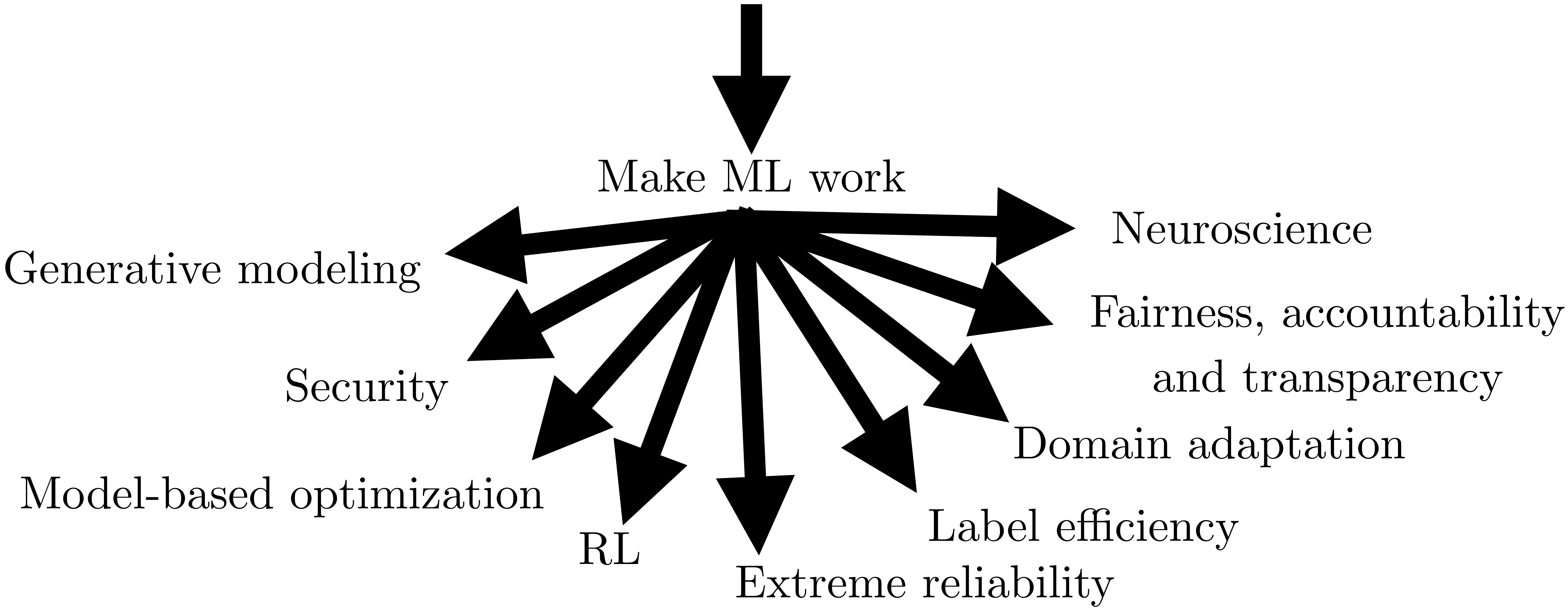
(Goodfellow 2019)

Adversarial Machine Learning: Game Theory

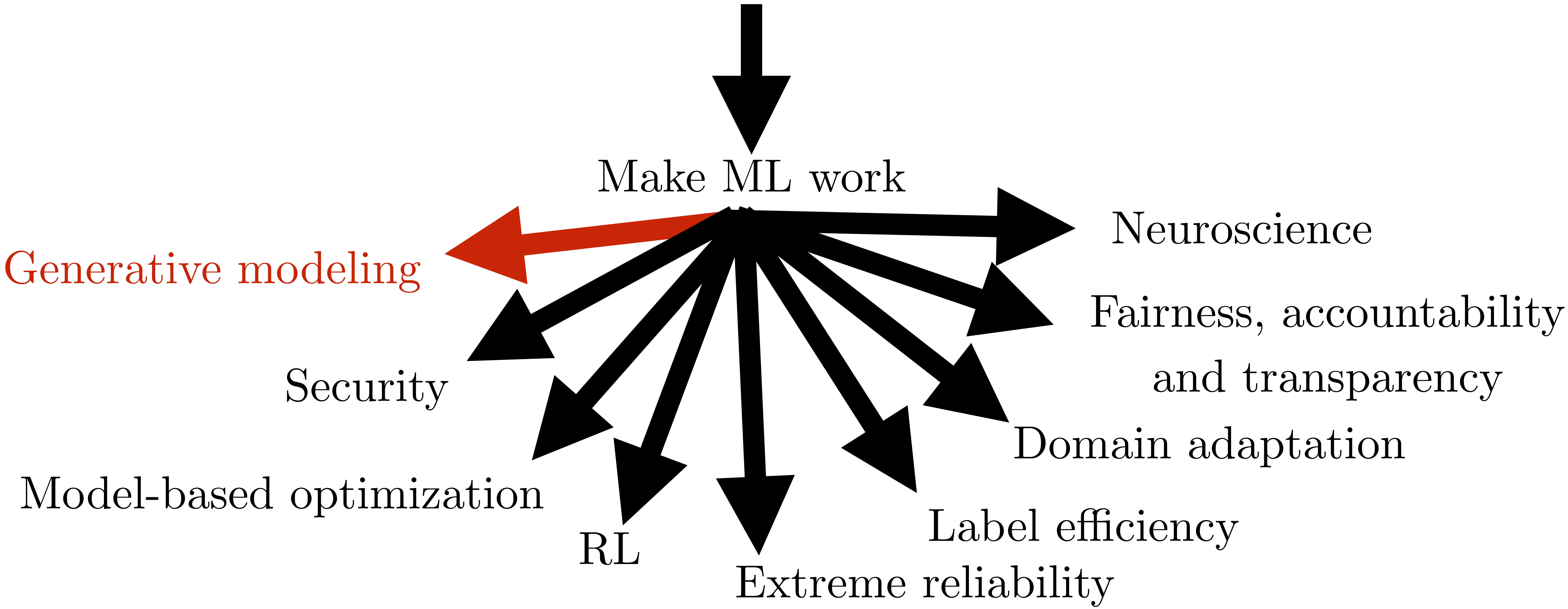


(Goodfellow 2019)

A Cambrian Explosion of Machine Learning Research Topics



A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2019)

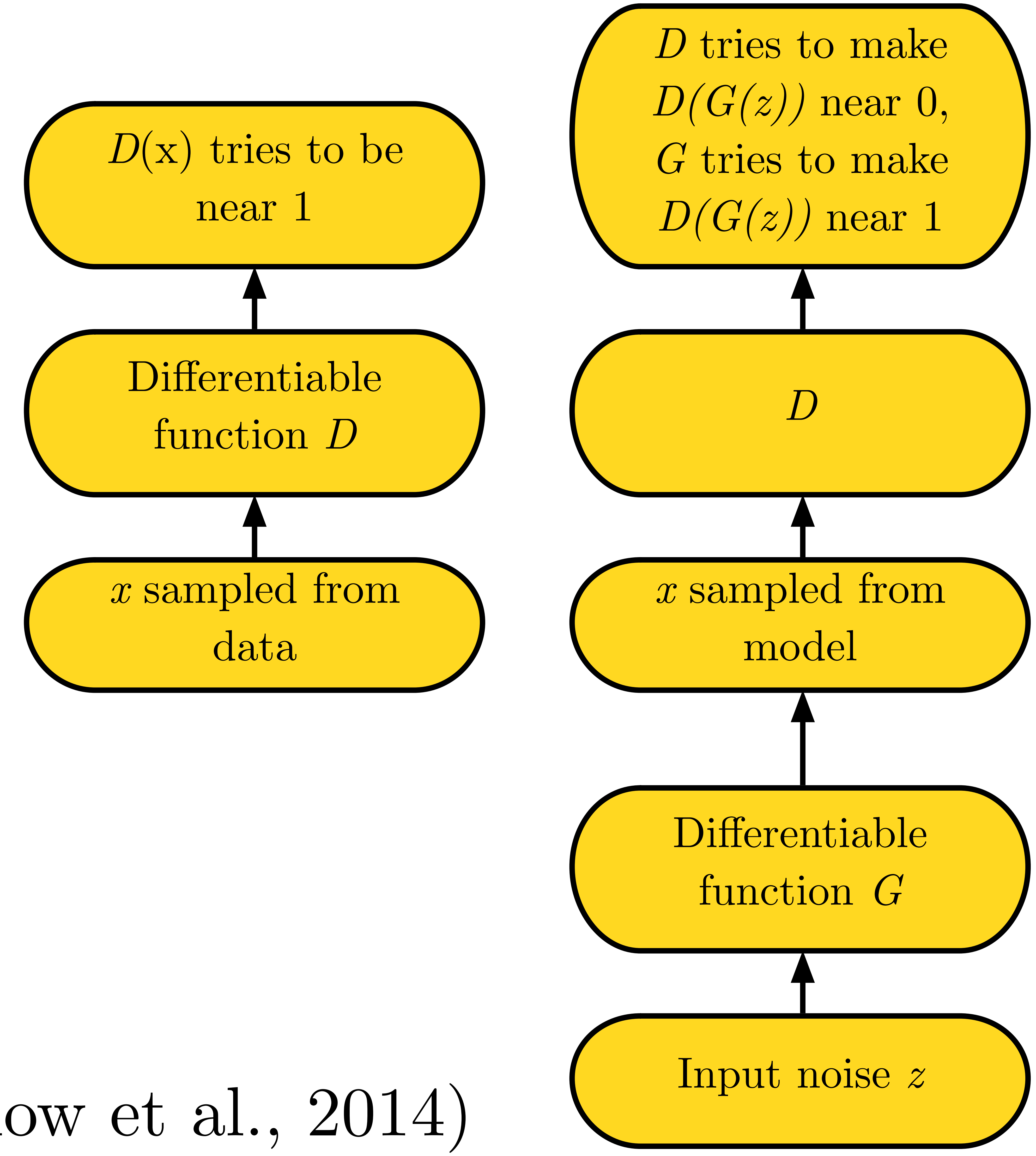
Generative Modeling: Sample Generation

Training Data
(CelebA)

Sample Generator
(Karras et al, 2017)

(Goodfellow 2019)

Generative Adversarial Networks



(Goodfellow et al., 2014)

4.5 years of progress on faces

2014

2015

2016

2017

2018

(Goodfellow 2019)

2 Years of Progress on ImageNet

Odena et al
2016

Miyato et al
2017

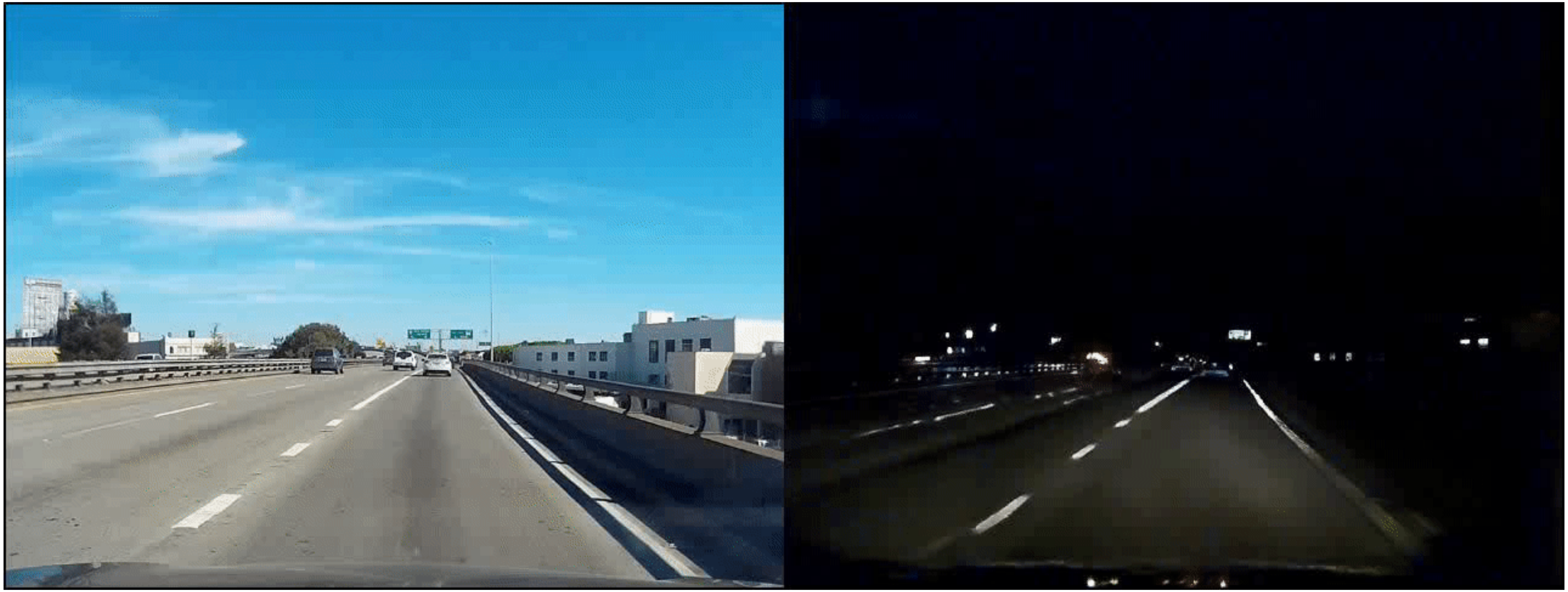
Zhang et al
2018

Brock et al
2018

(Odena 2018)

Unsupervised Image-to-Image Translation

Day to night

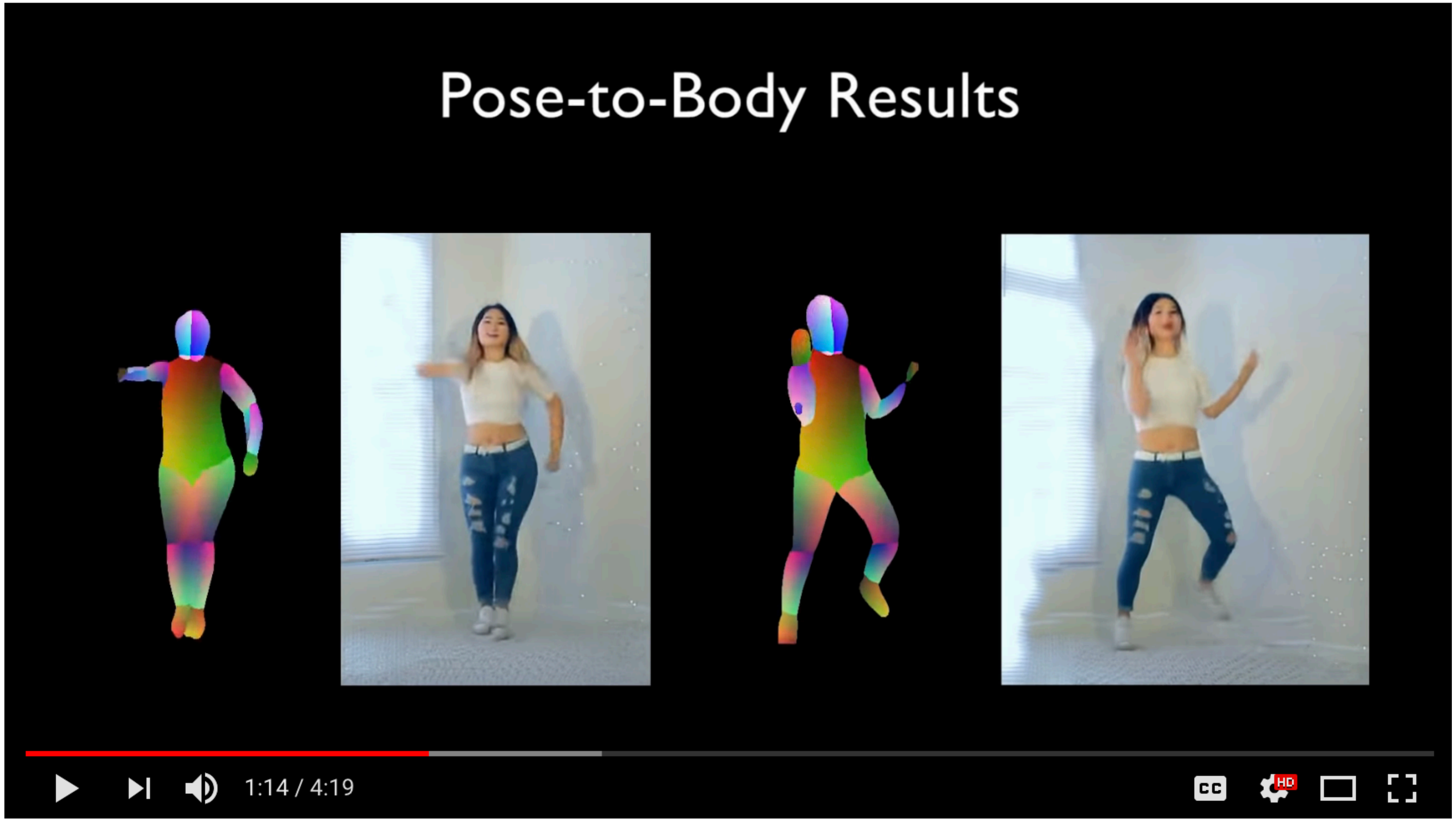


(Liu et al., 2017)

CycleGAN

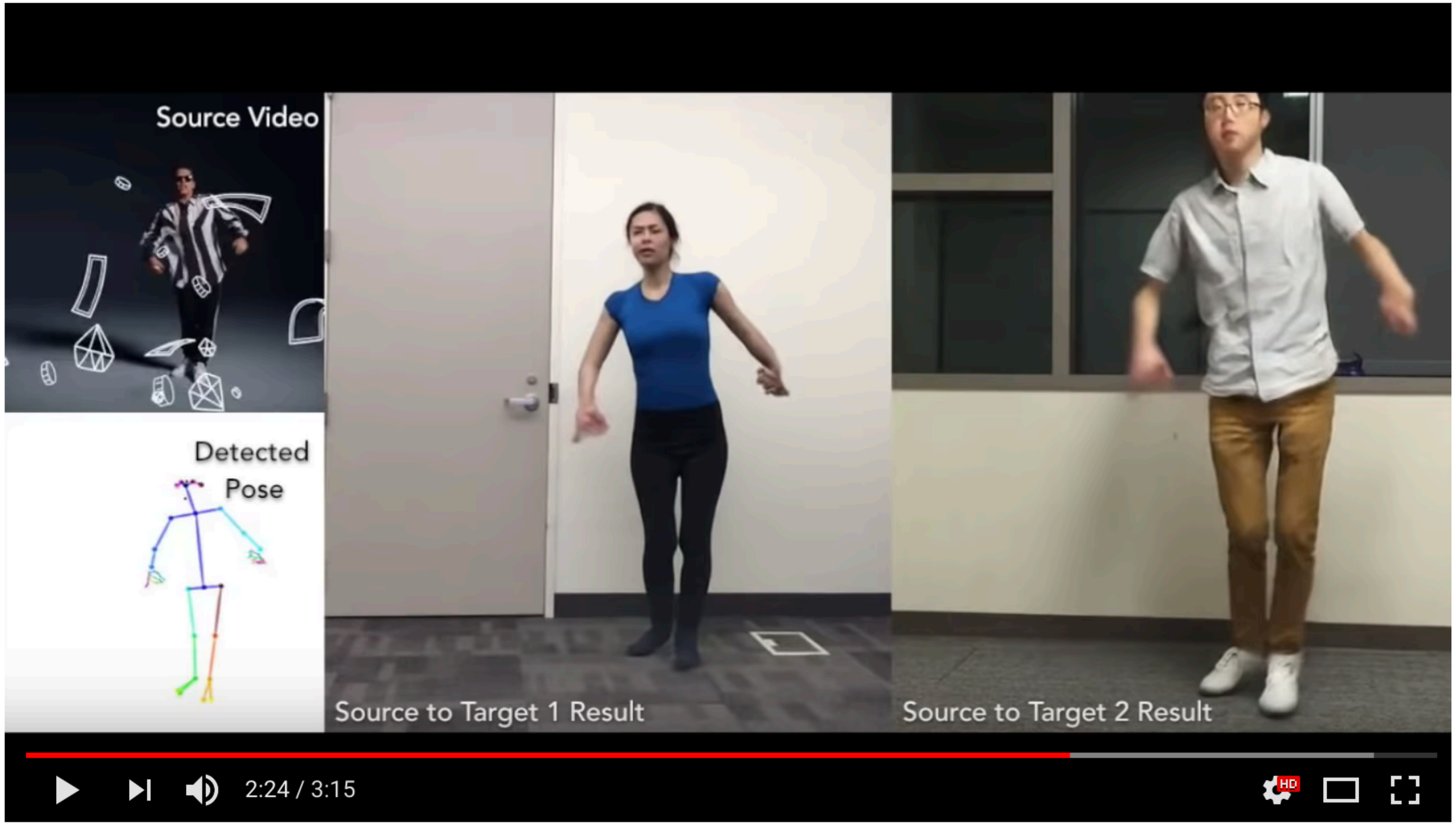
(Zhu et al., 2017)

Video-to-Video



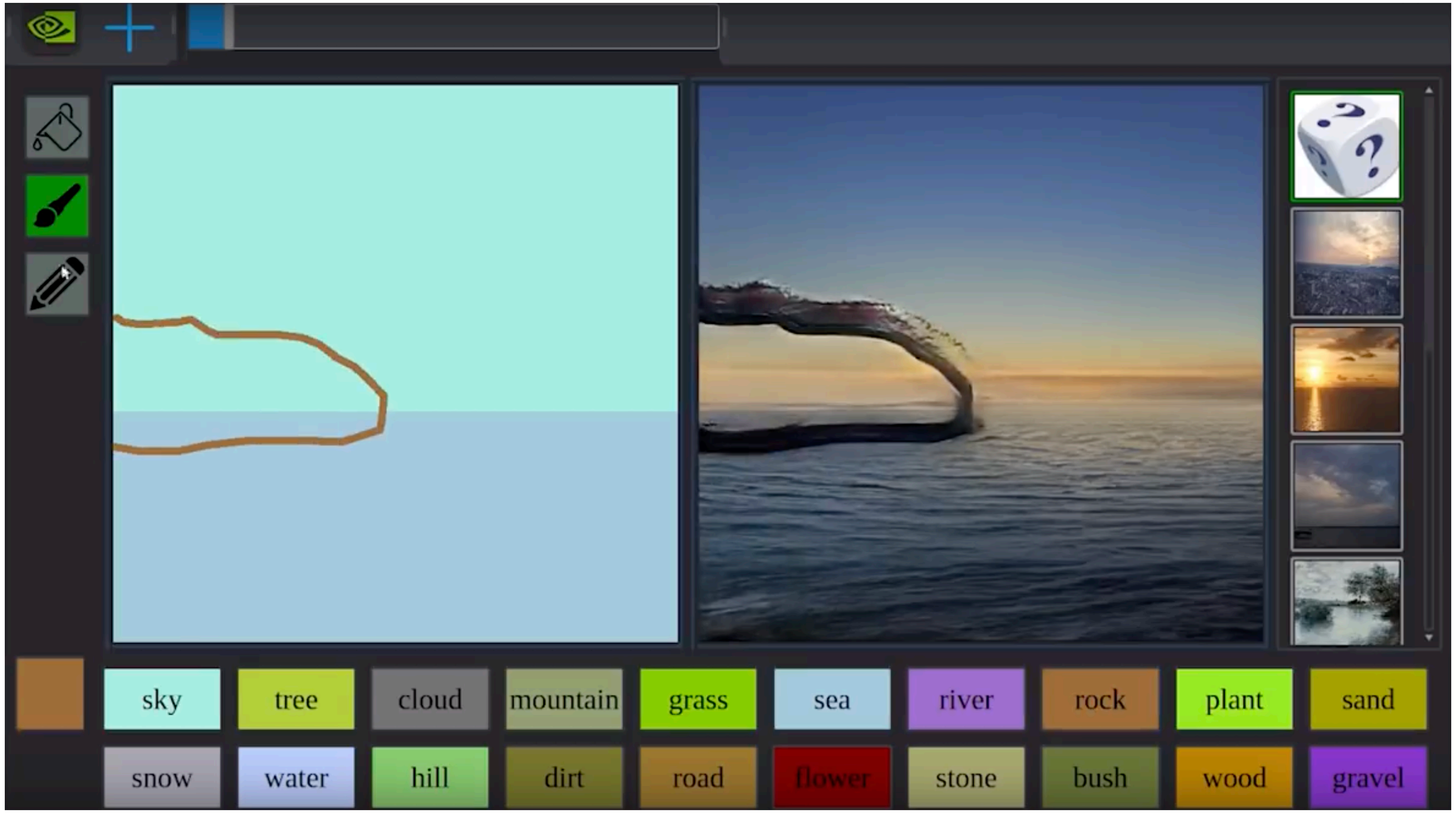
(Wang et al, 2018)

Everybody Dance Now



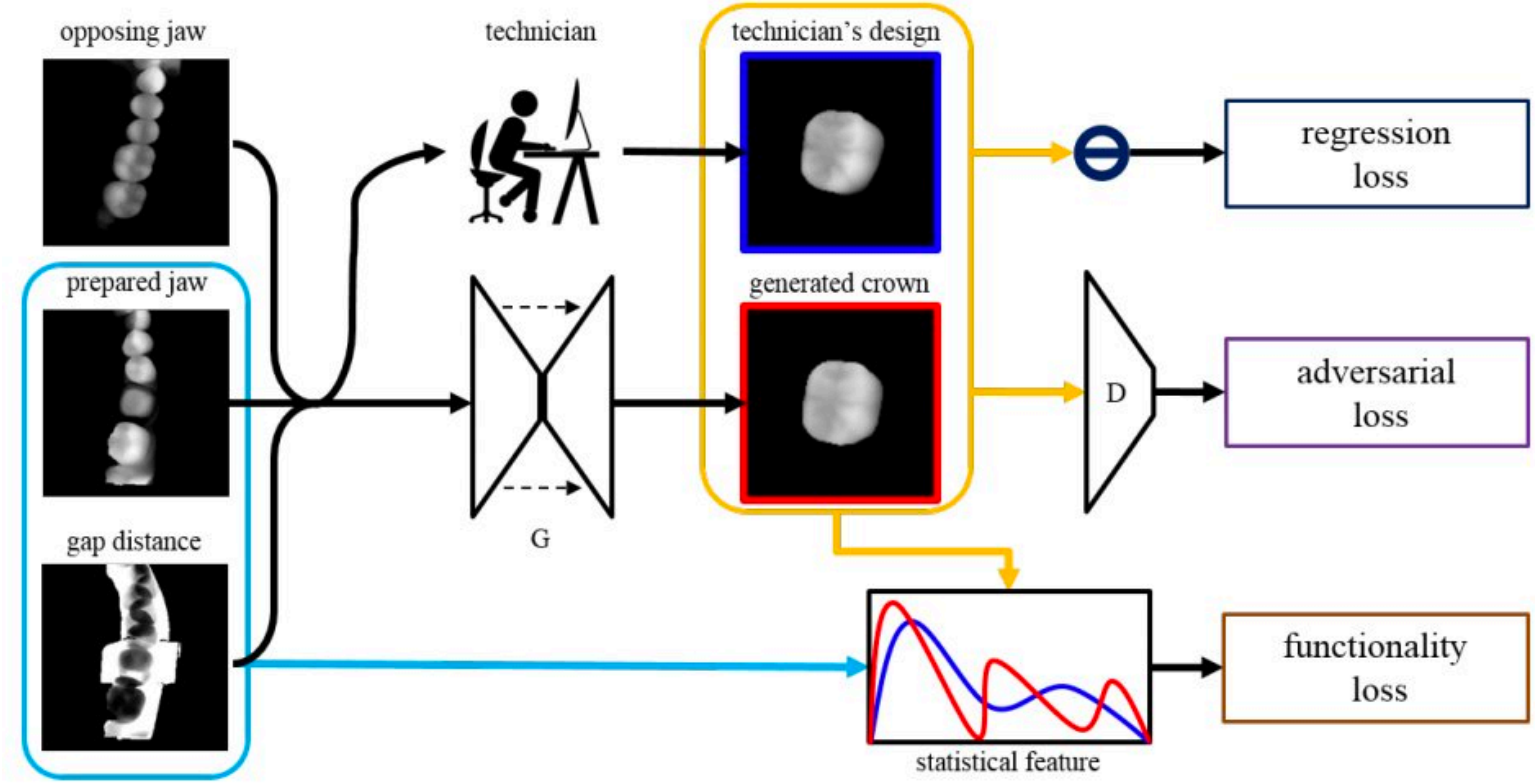
(Chan et al 2018)

GauGAN



(Park et al 2019)

Personalized GANufacturing



(Hwang et al 2018)

Recent Advances

(Brock et al, 2018)
BigGAN
Large scale TPU
implementation

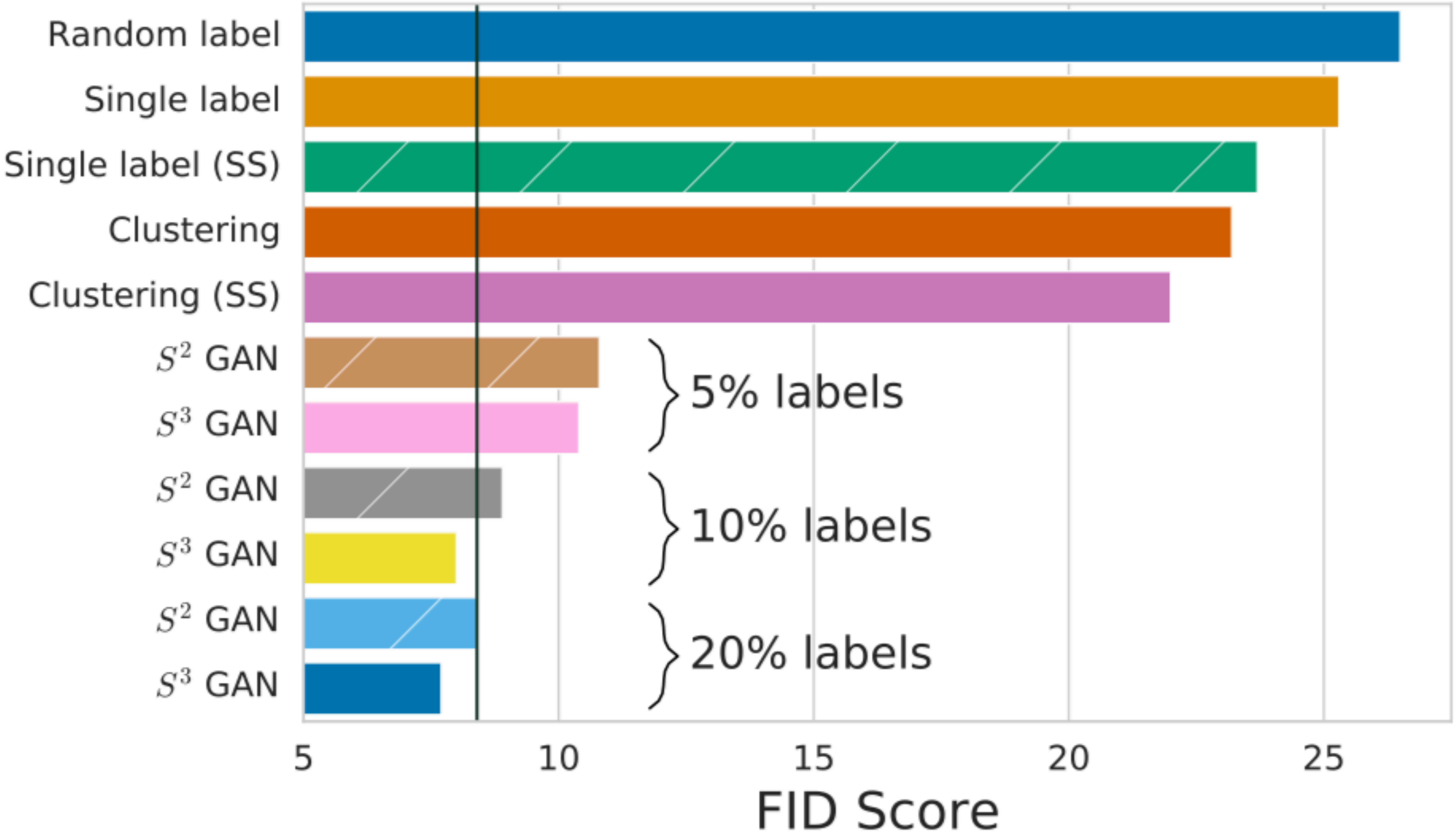
Starting sample
(Fake)

Sample for coarse
style (also fake)

Result
(still fake)

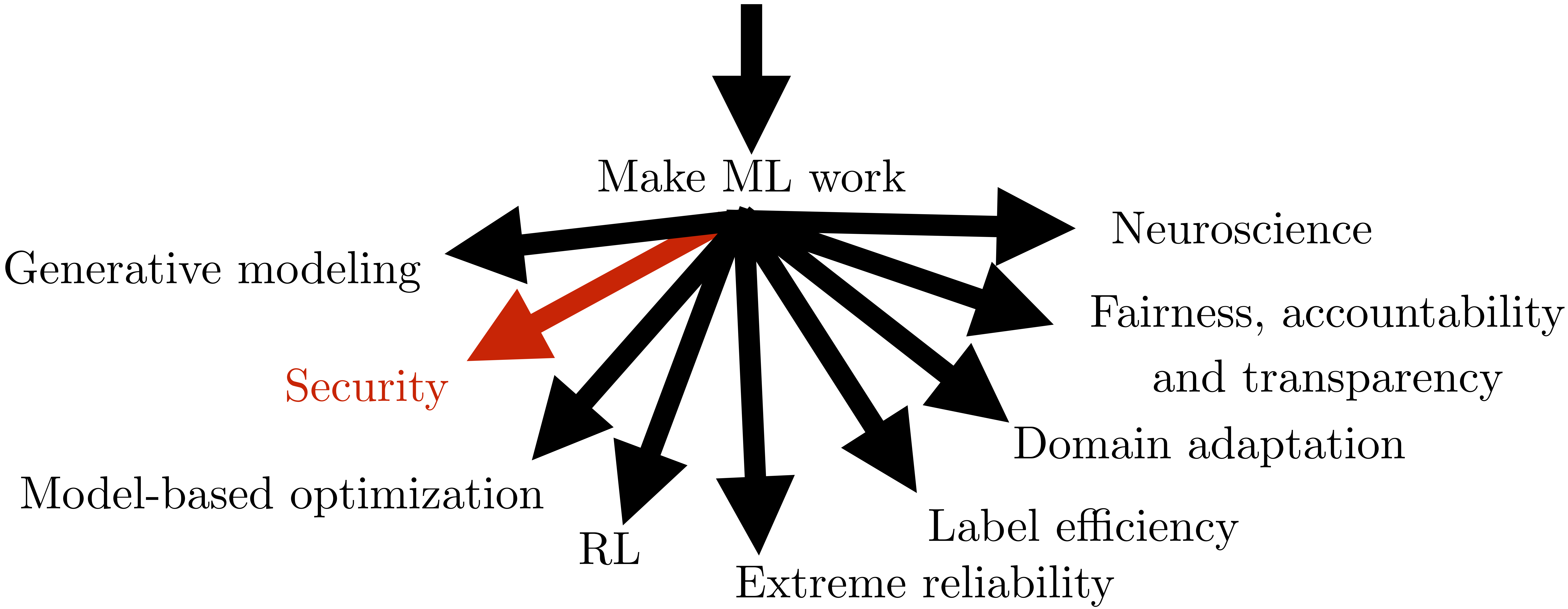
Style-based generators
(Karras et al, 2018)

Reducing Supervision Needed for "Unsupervised" Learning



(Lucic+Tschannen+Ritter et al 2019)

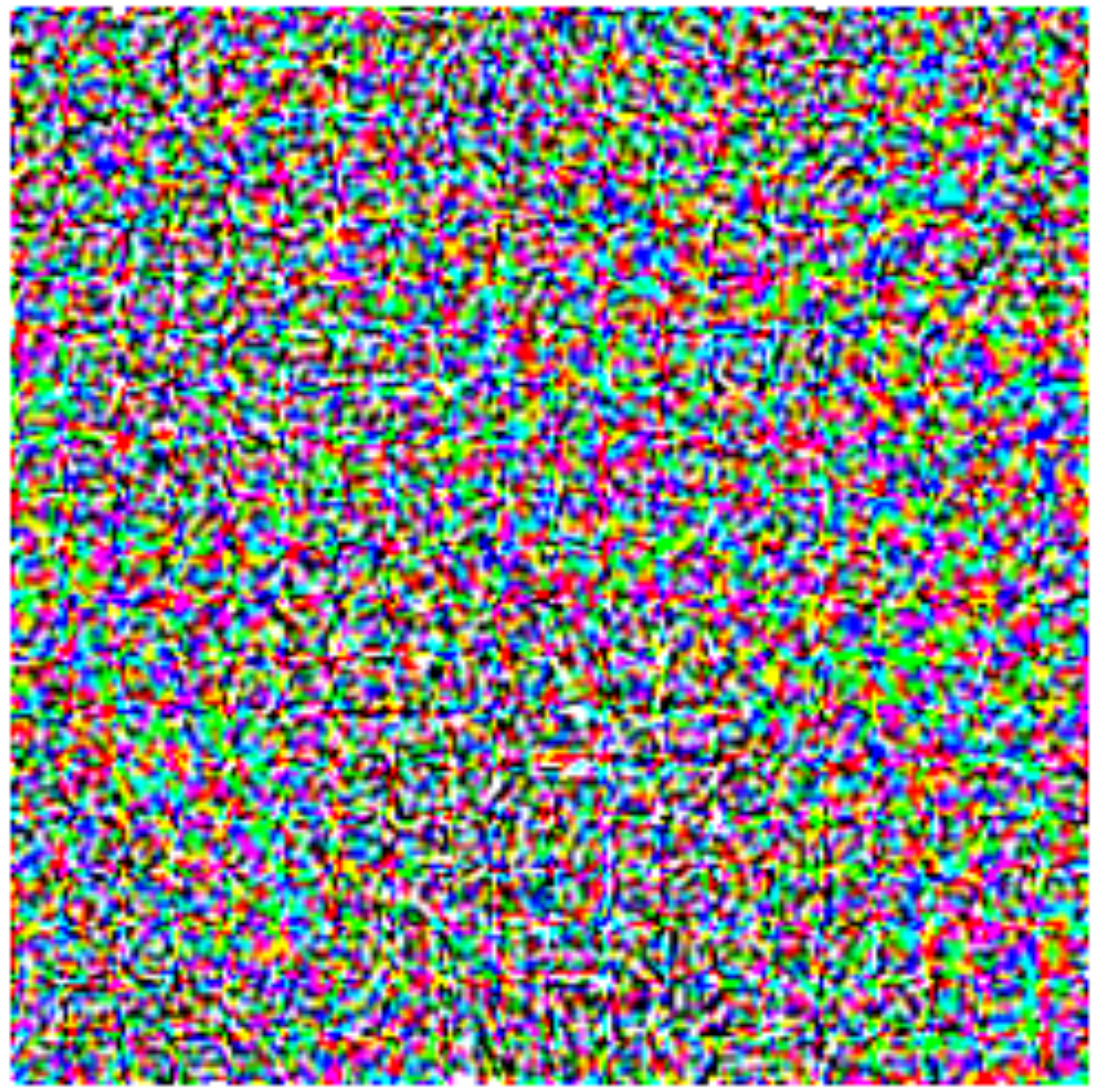
A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2019)

Adversarial Examples

$+ .007 \times$



$=$

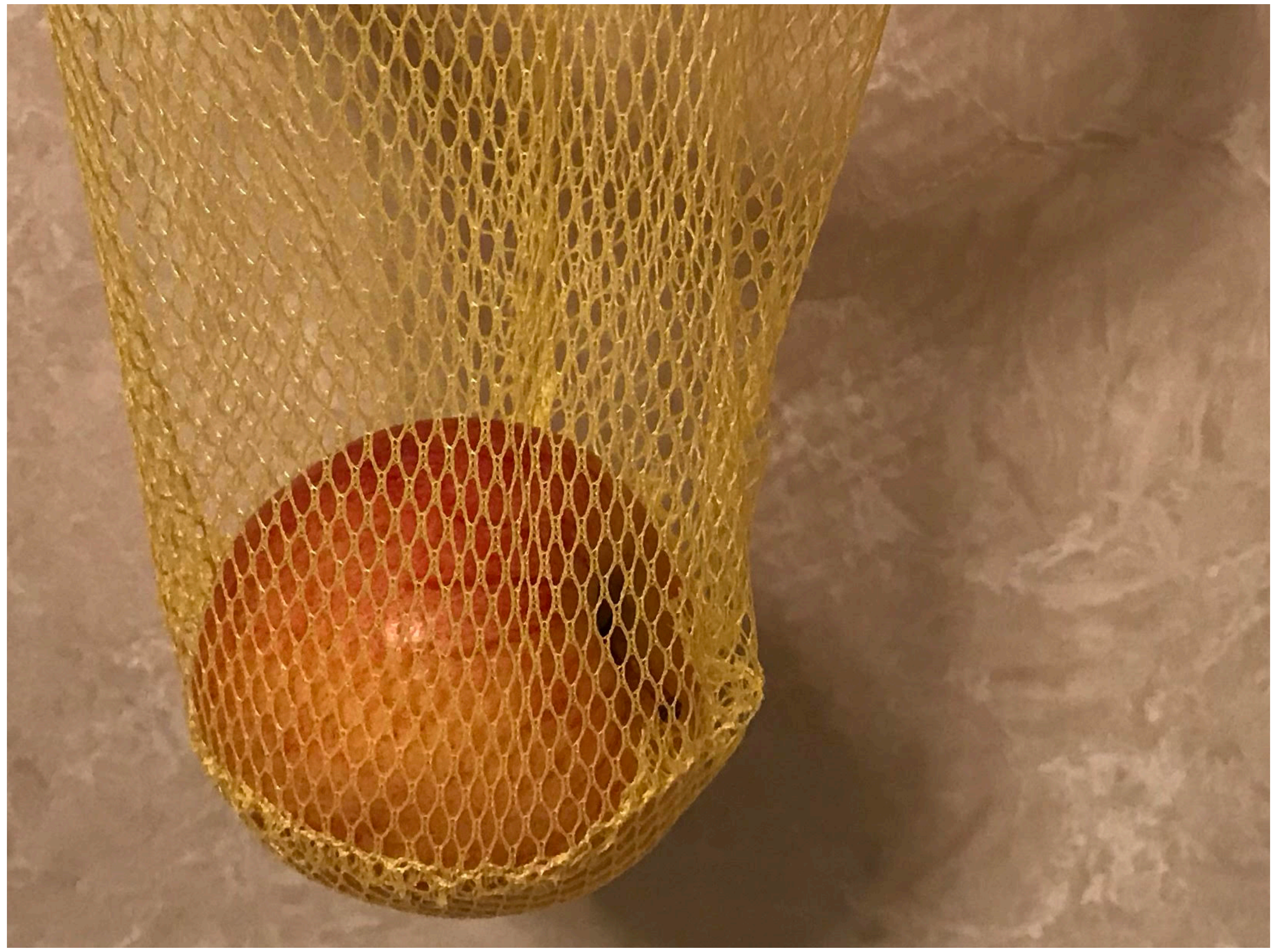
58% panda

99% gibbon

(Goodfellow et al, 2014)

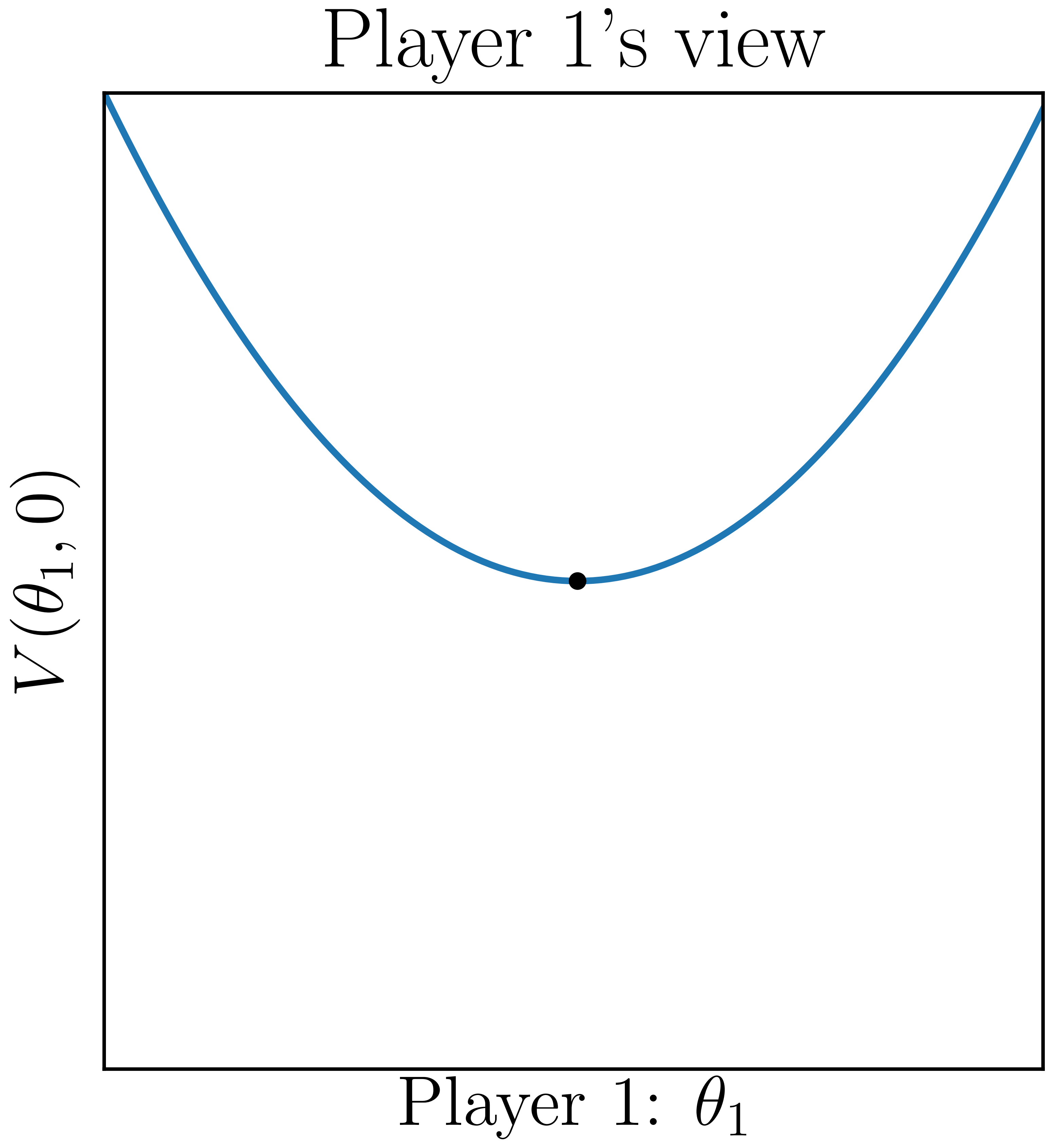
Also Adversarial Examples

(Eykholt et al, 2017)

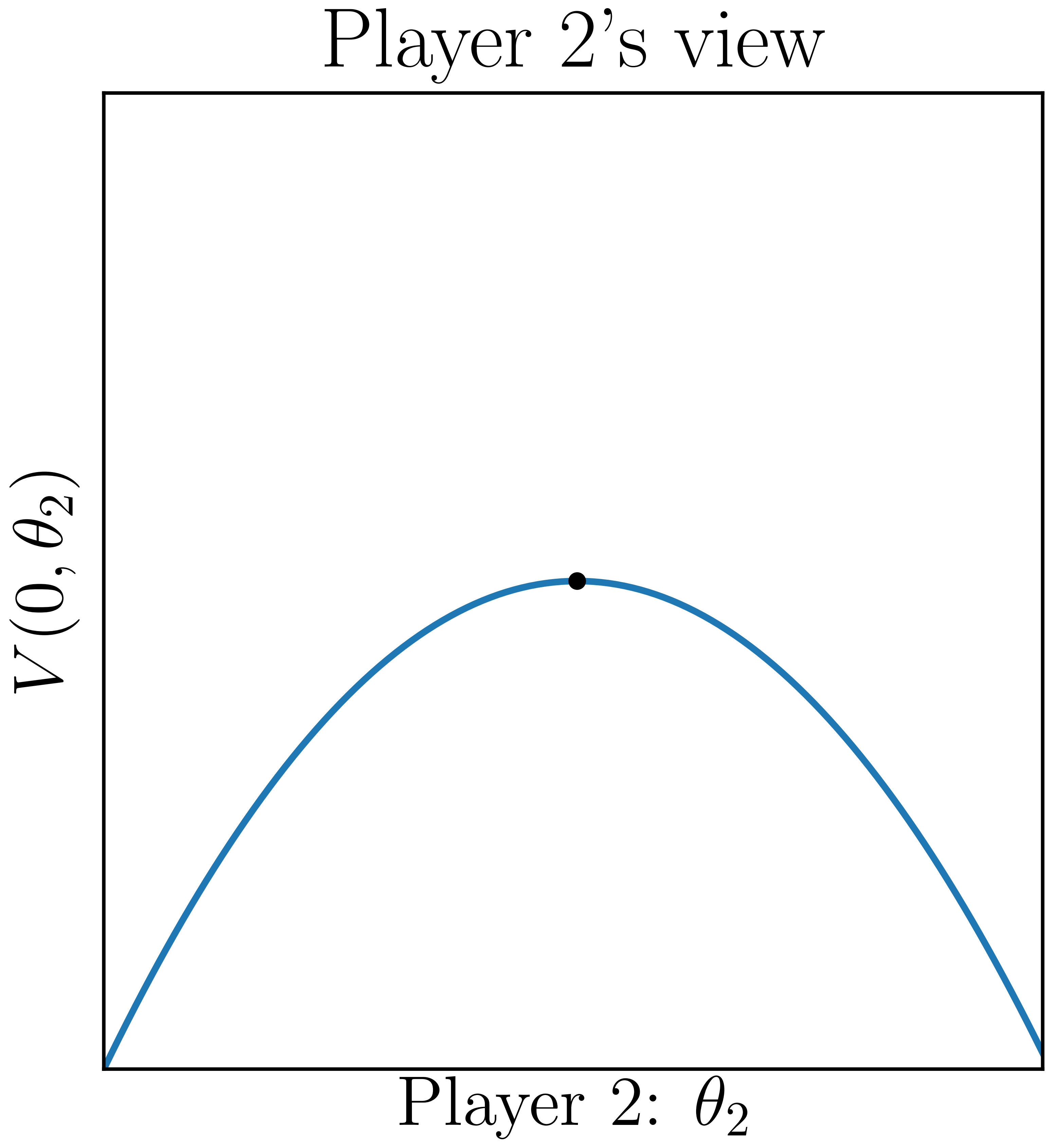


(Goodfellow 2018)

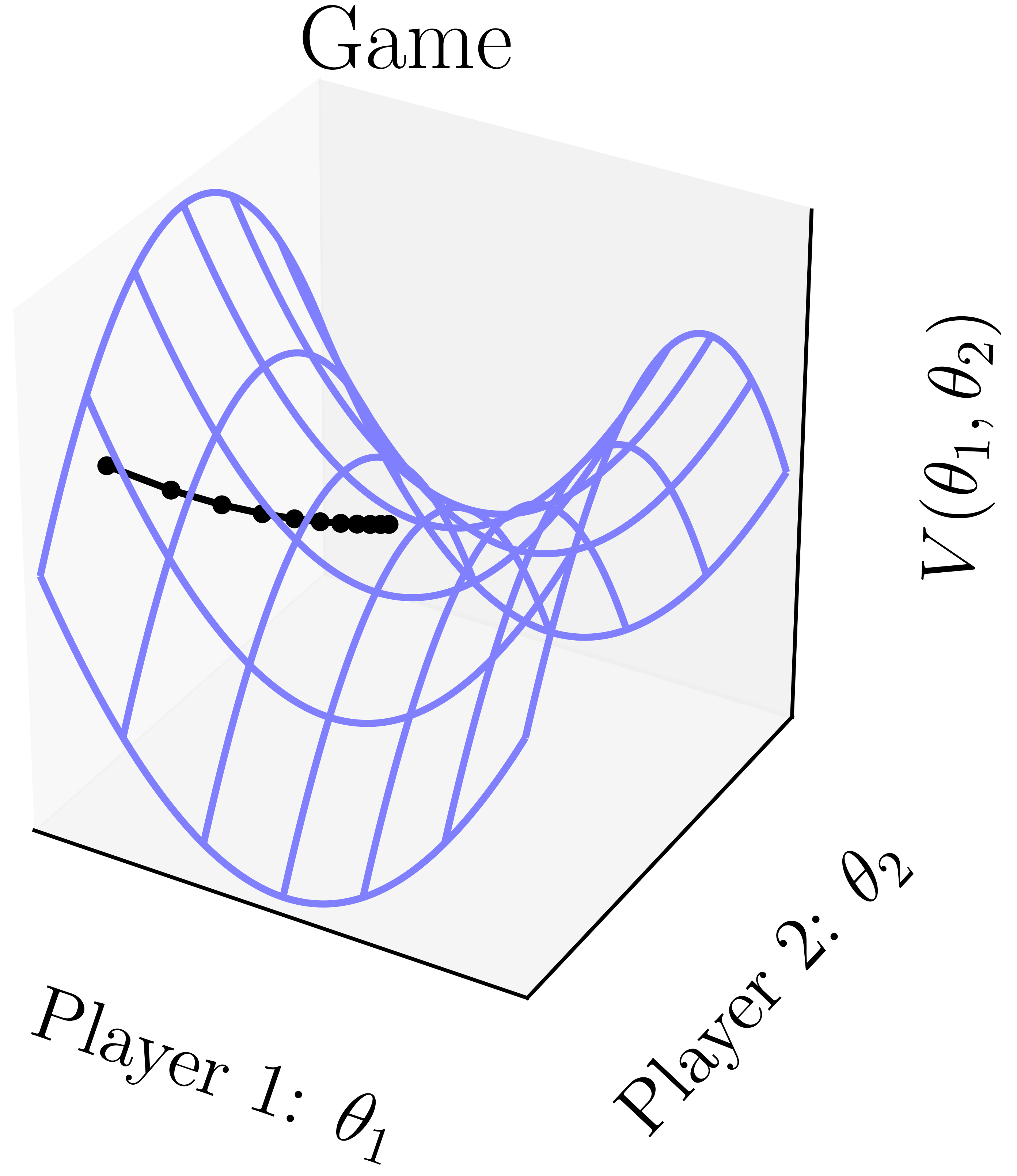
Adversarial Training



ML model learns parameters

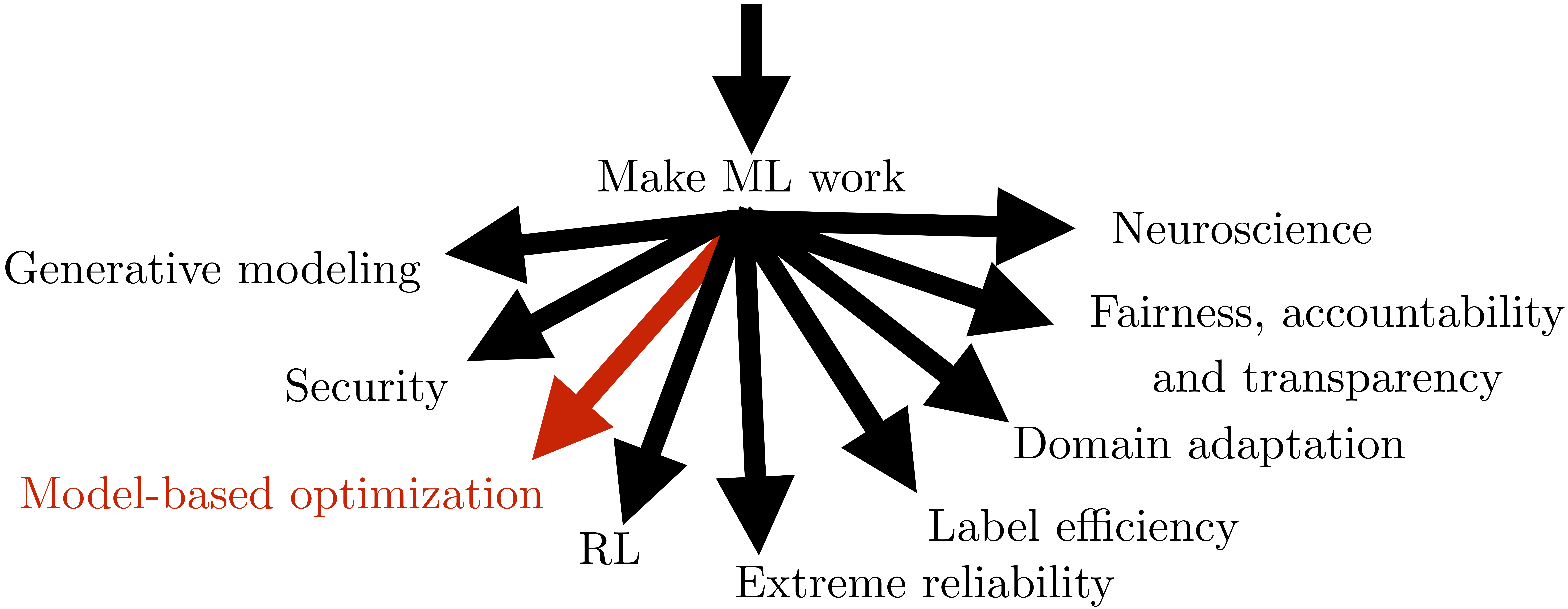


Attacker chooses input point



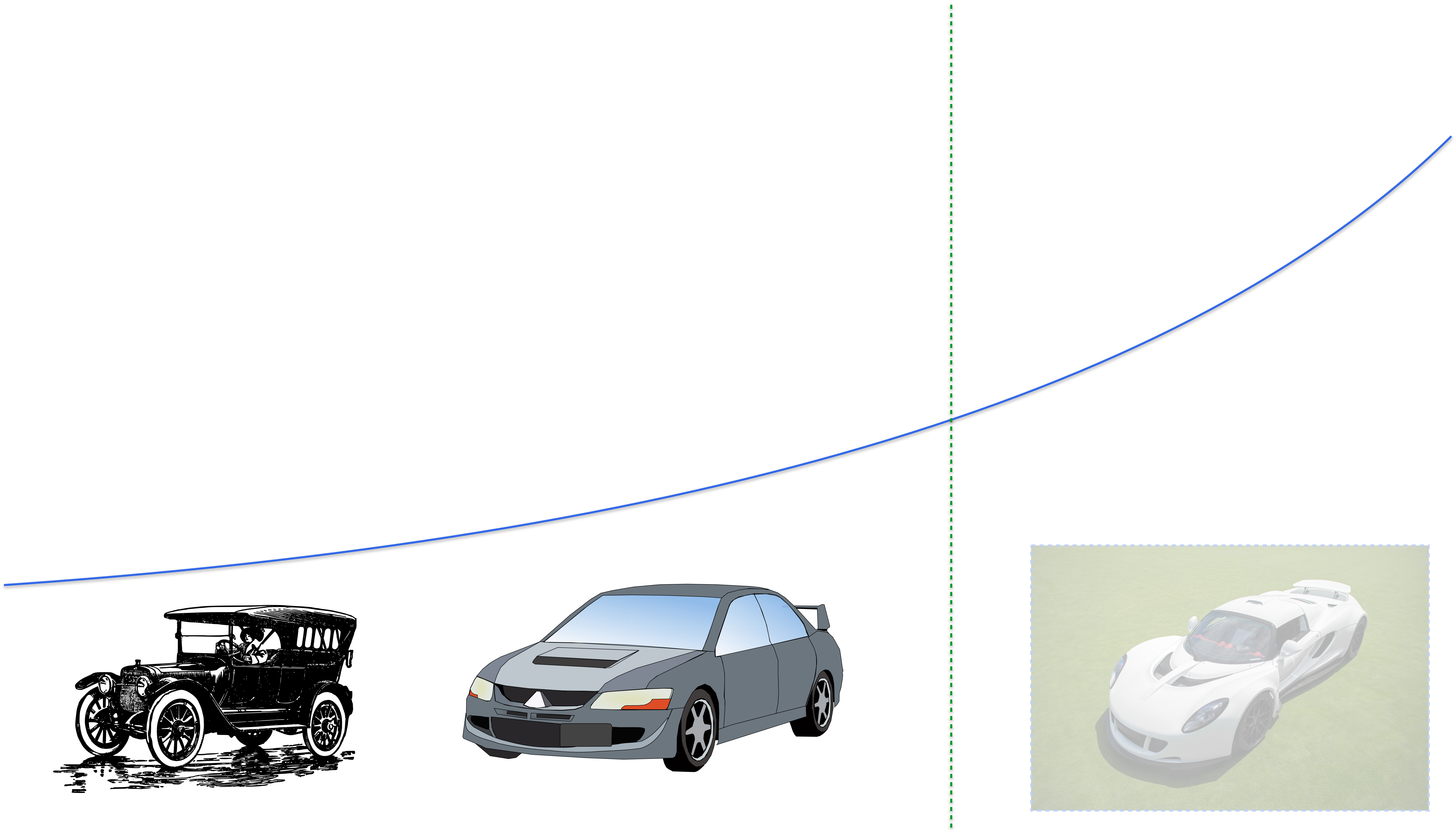
(Goodfellow et al 2014, Farley and Goodfellow 2016)

A Cambrian Explosion of Machine Learning Research Topics



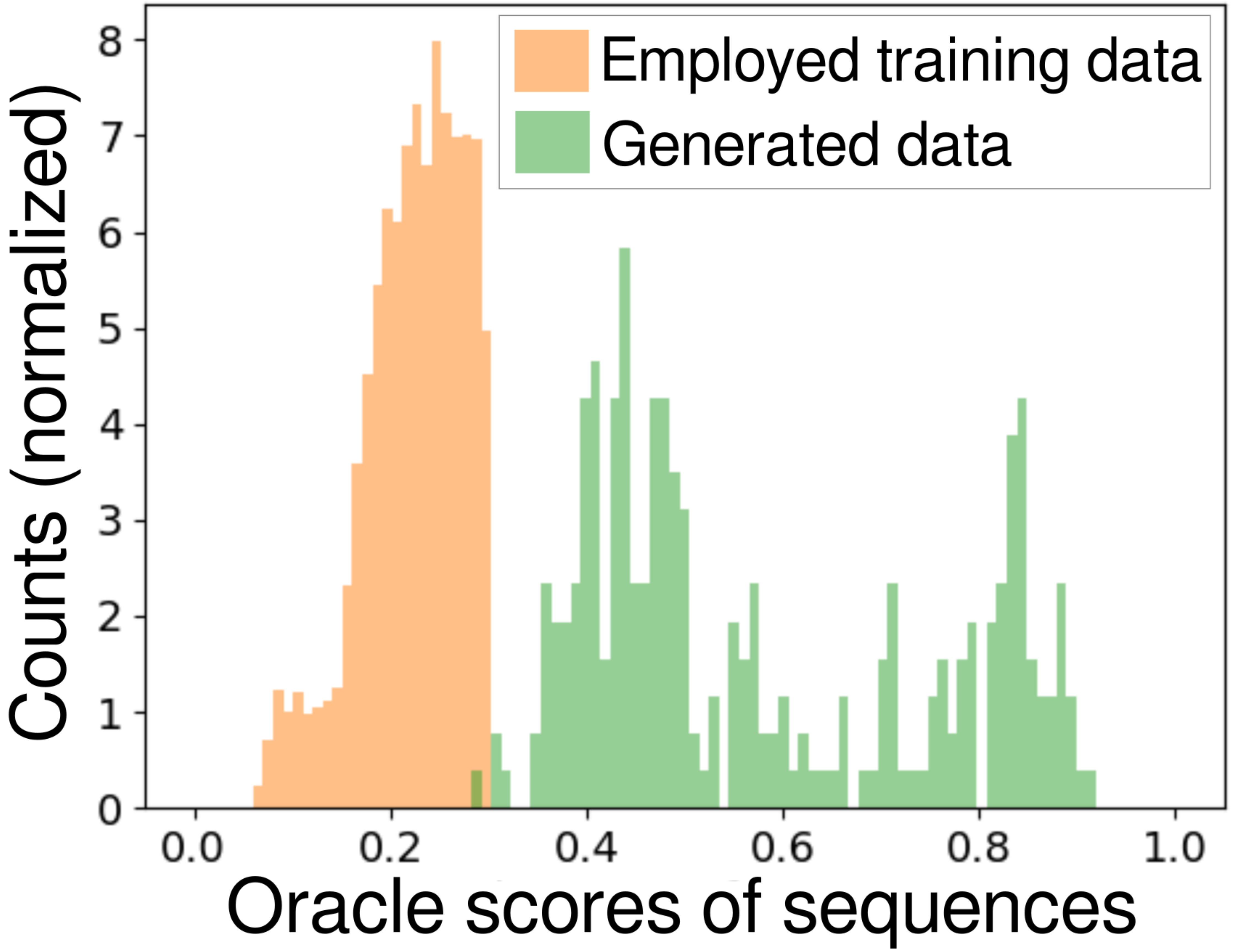
(Goodfellow 2019)

Model-Based Optimization



(Goodfellow 2019)

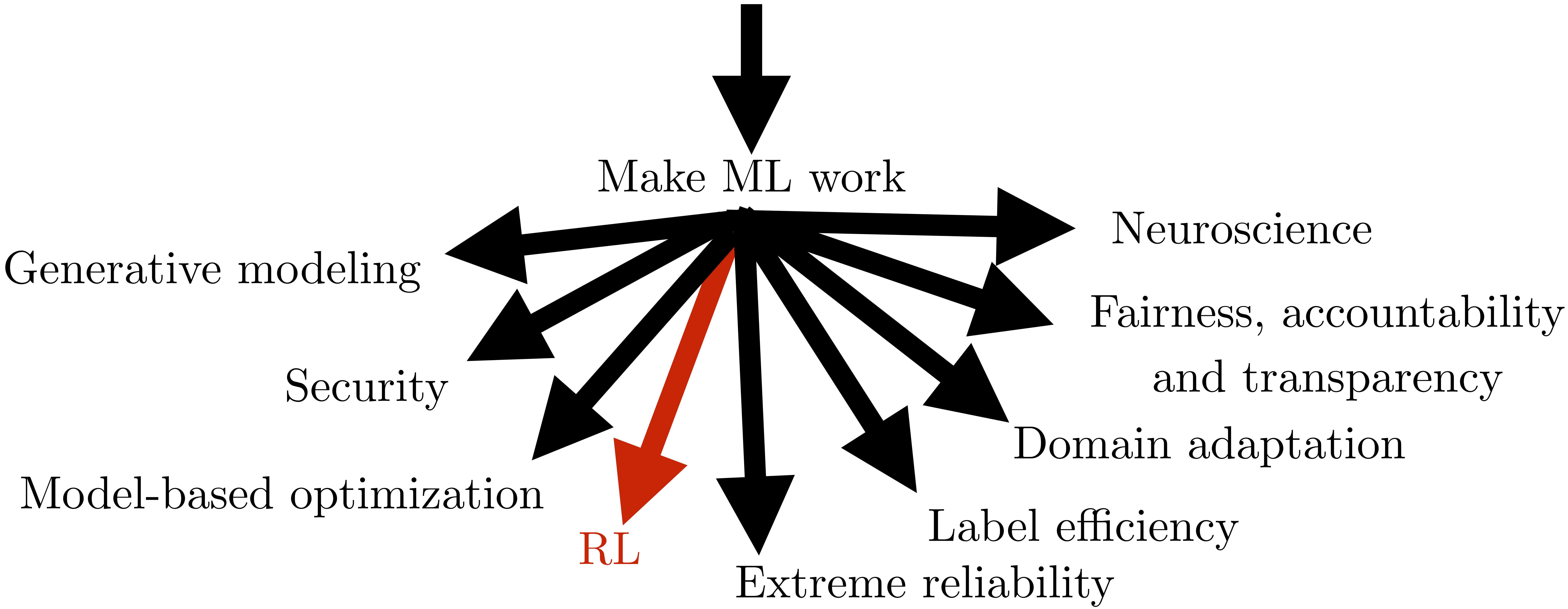
Designing DNA to optimize protein function



(Killoran et al, 2017)

(Gupta and Zou, 2018)

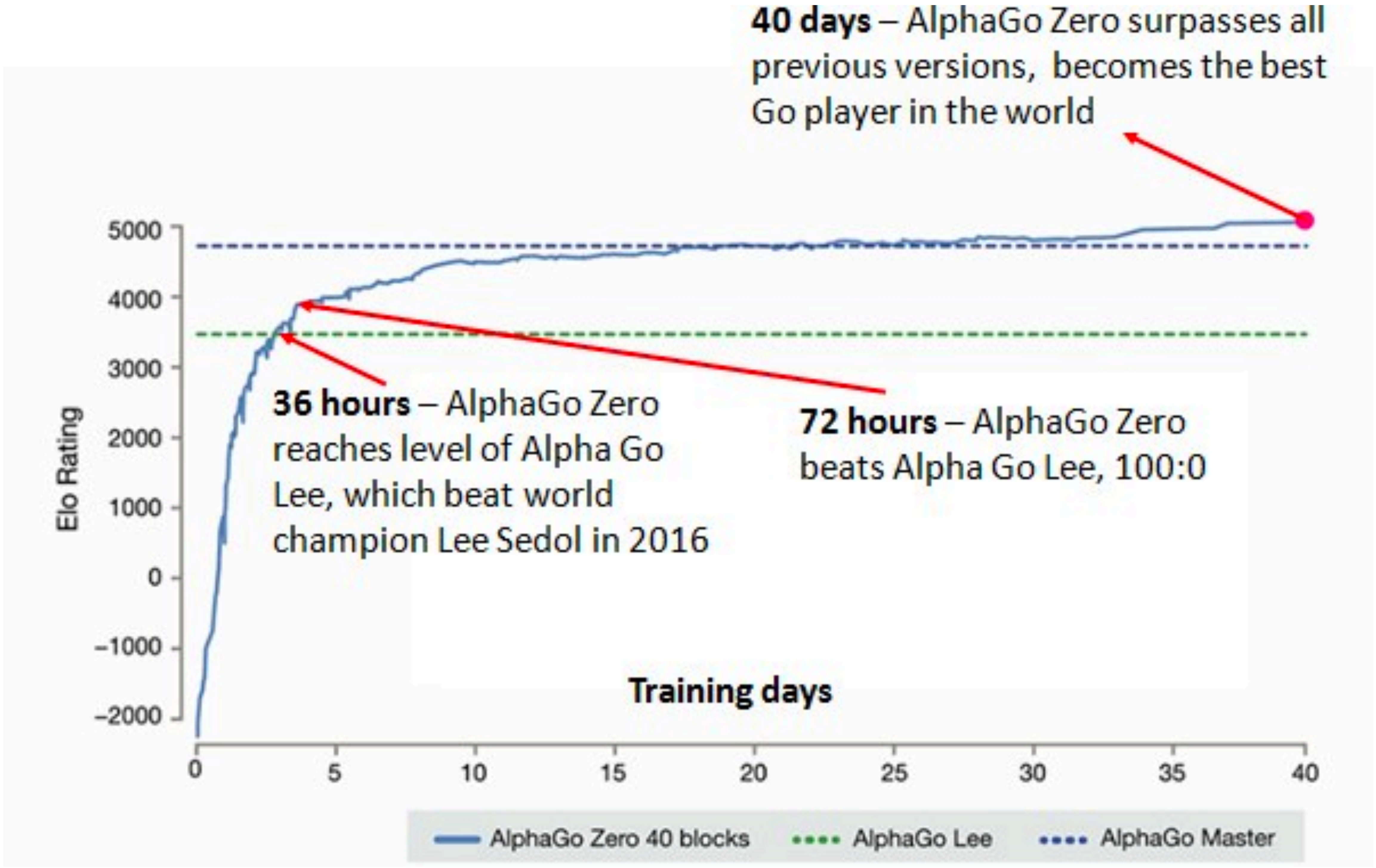
A Cambrian Explosion of Machine Learning Research Topics



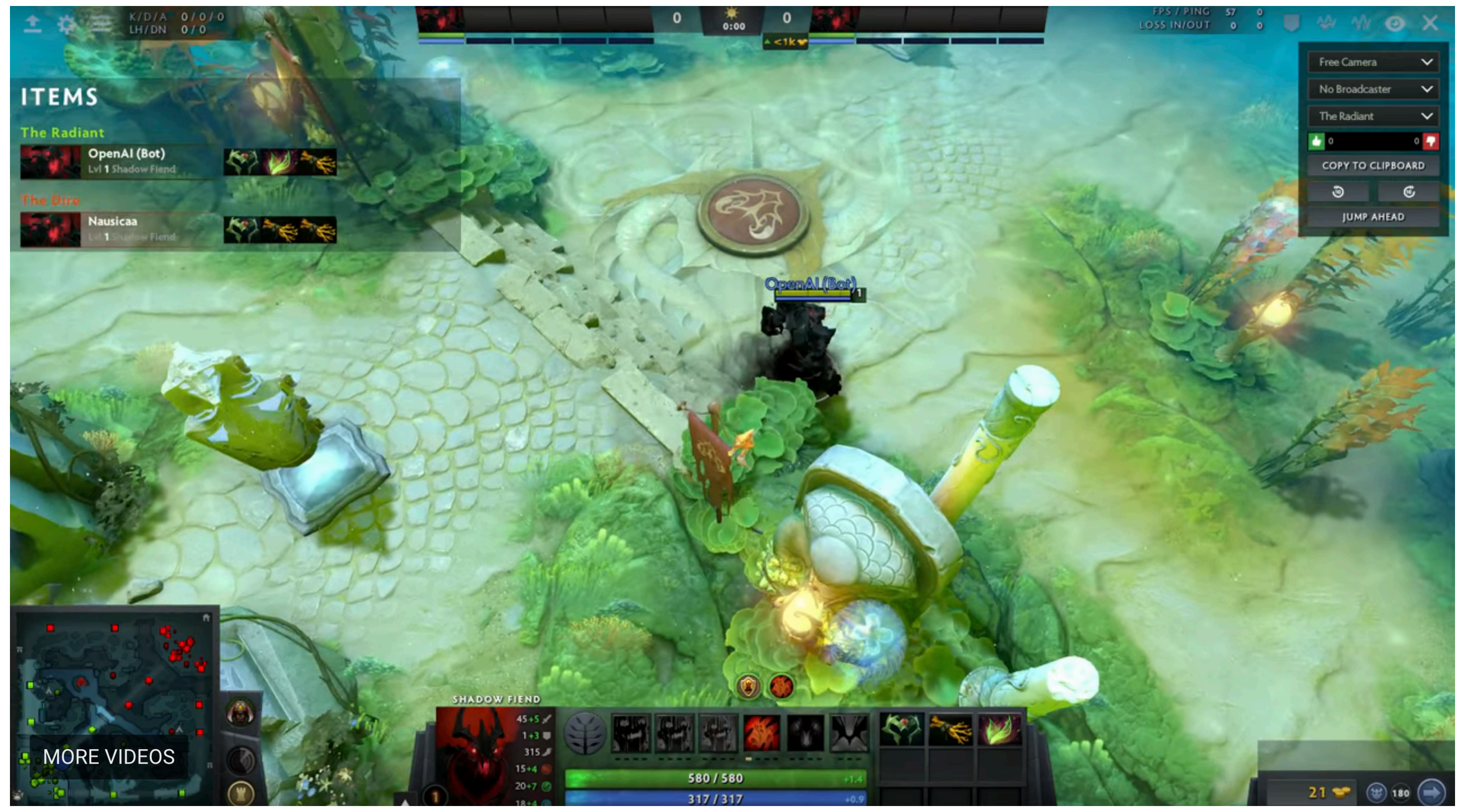
(Goodfellow 2019)

Self-Play

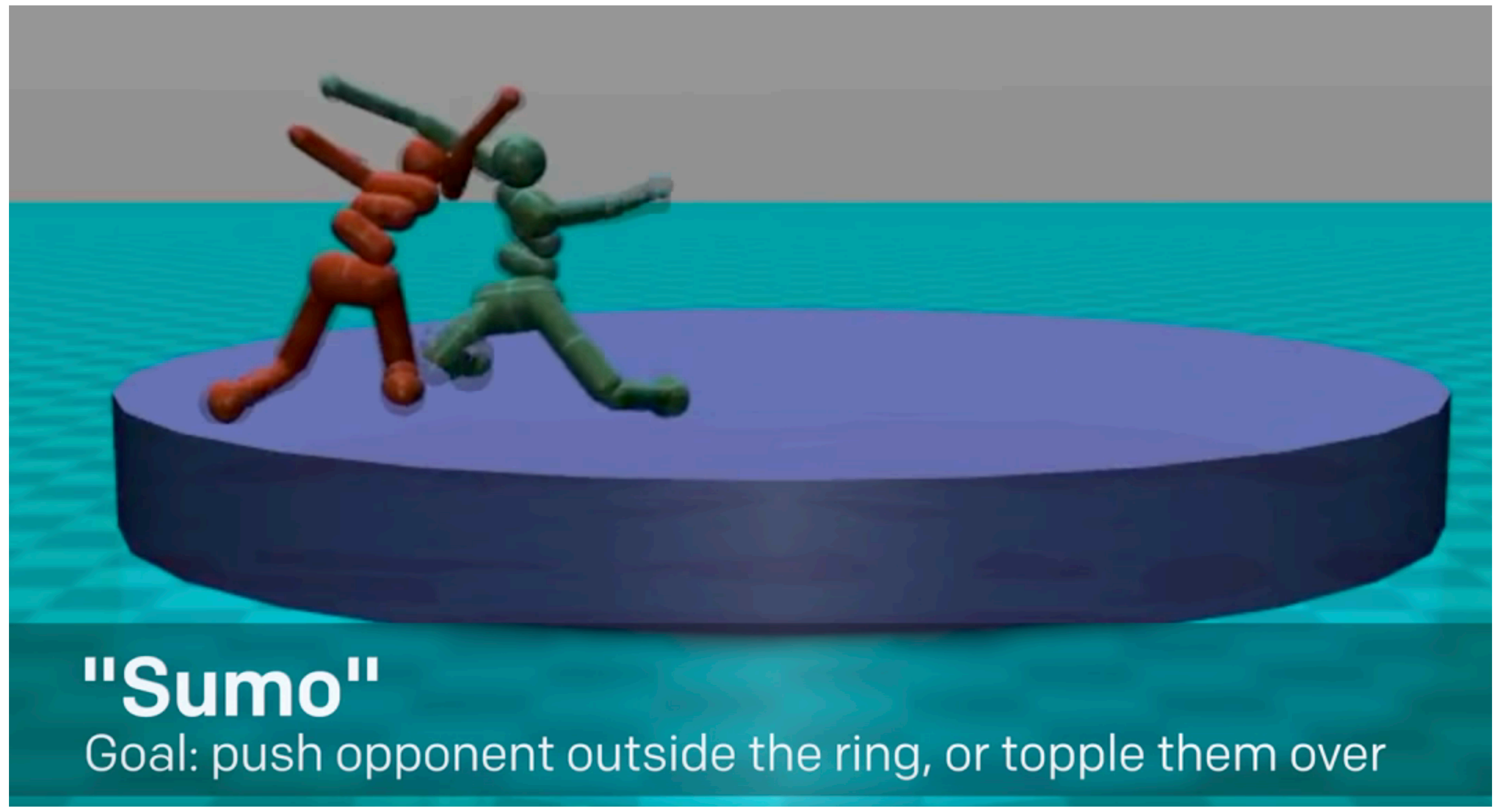
1959: Arthur Samuel's checkers agent



(Silver et al, 2017)

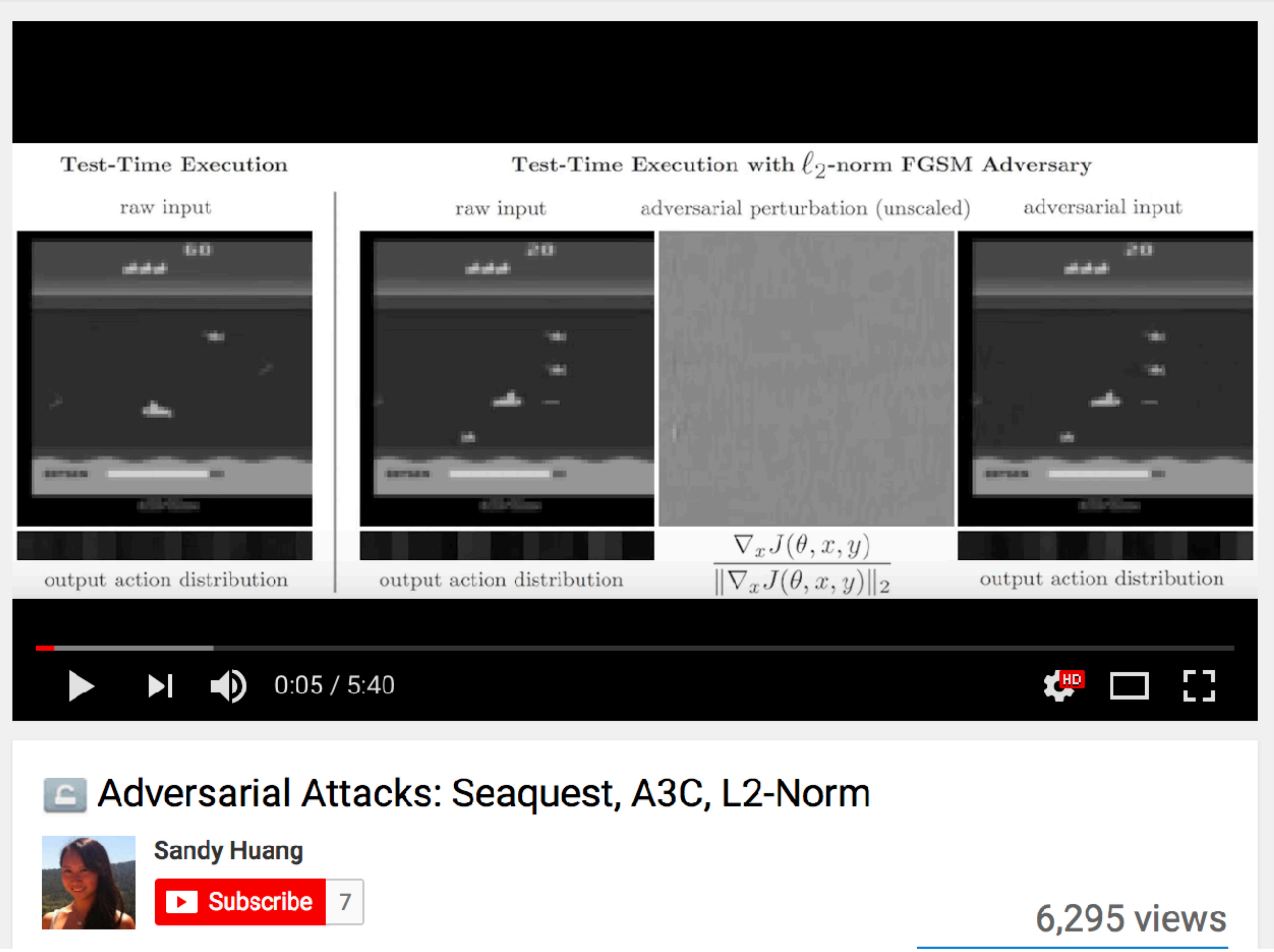


(OpenAI, 2017)



(Bansal et al, 2017)

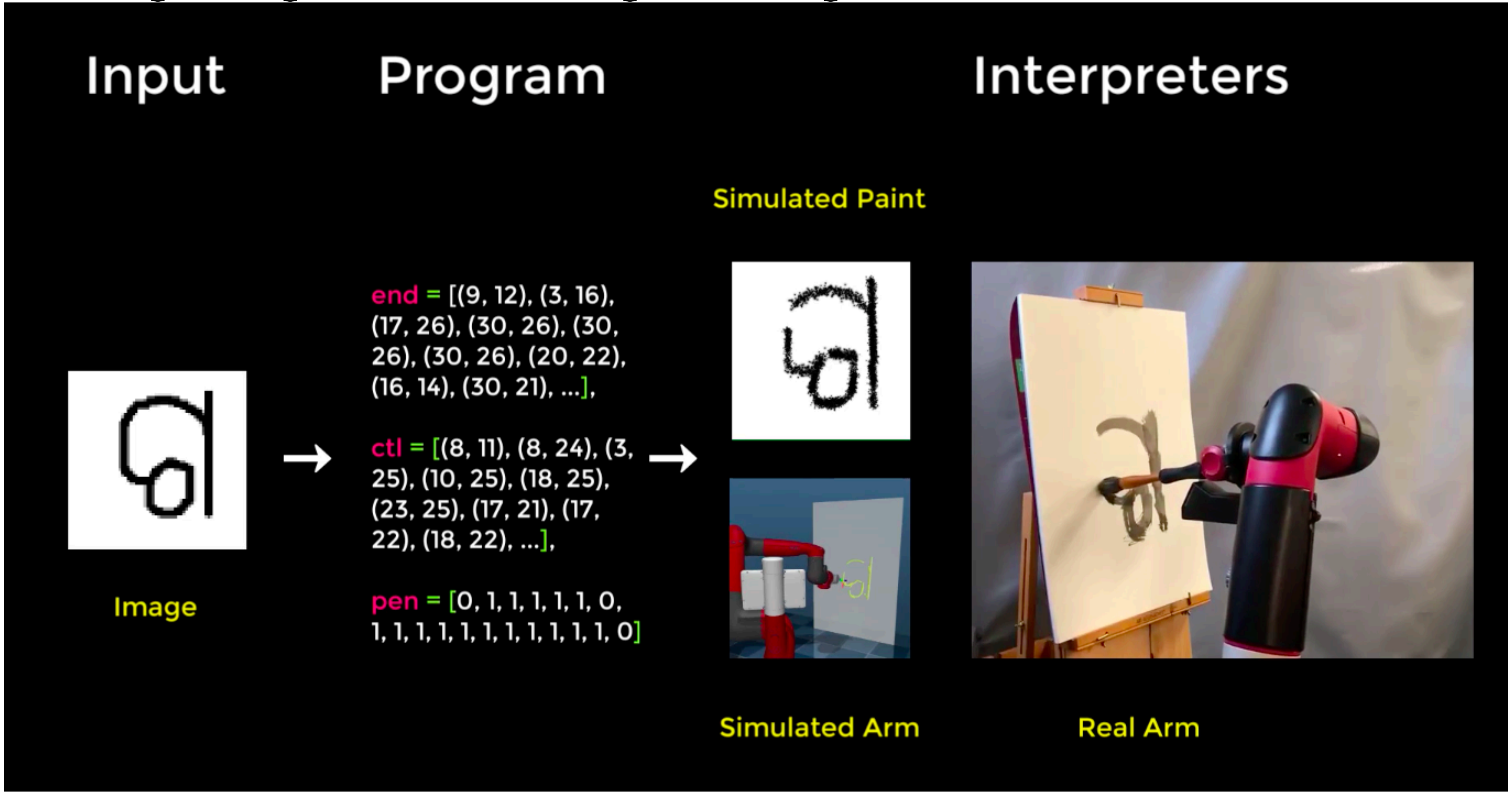
Adversarial Examples for RL



(Huang et al., 2017)

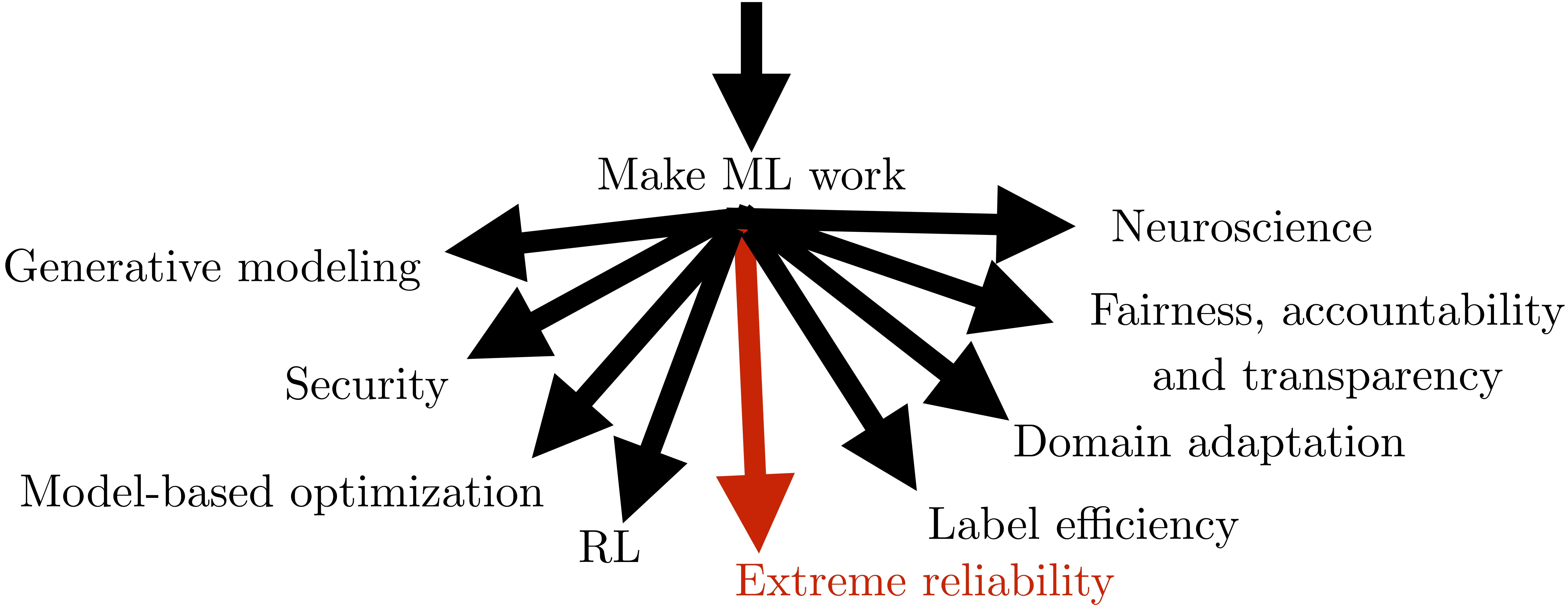
SPIRAL

Synthesizing Programs for Images Using Reinforced Adversarial Learning



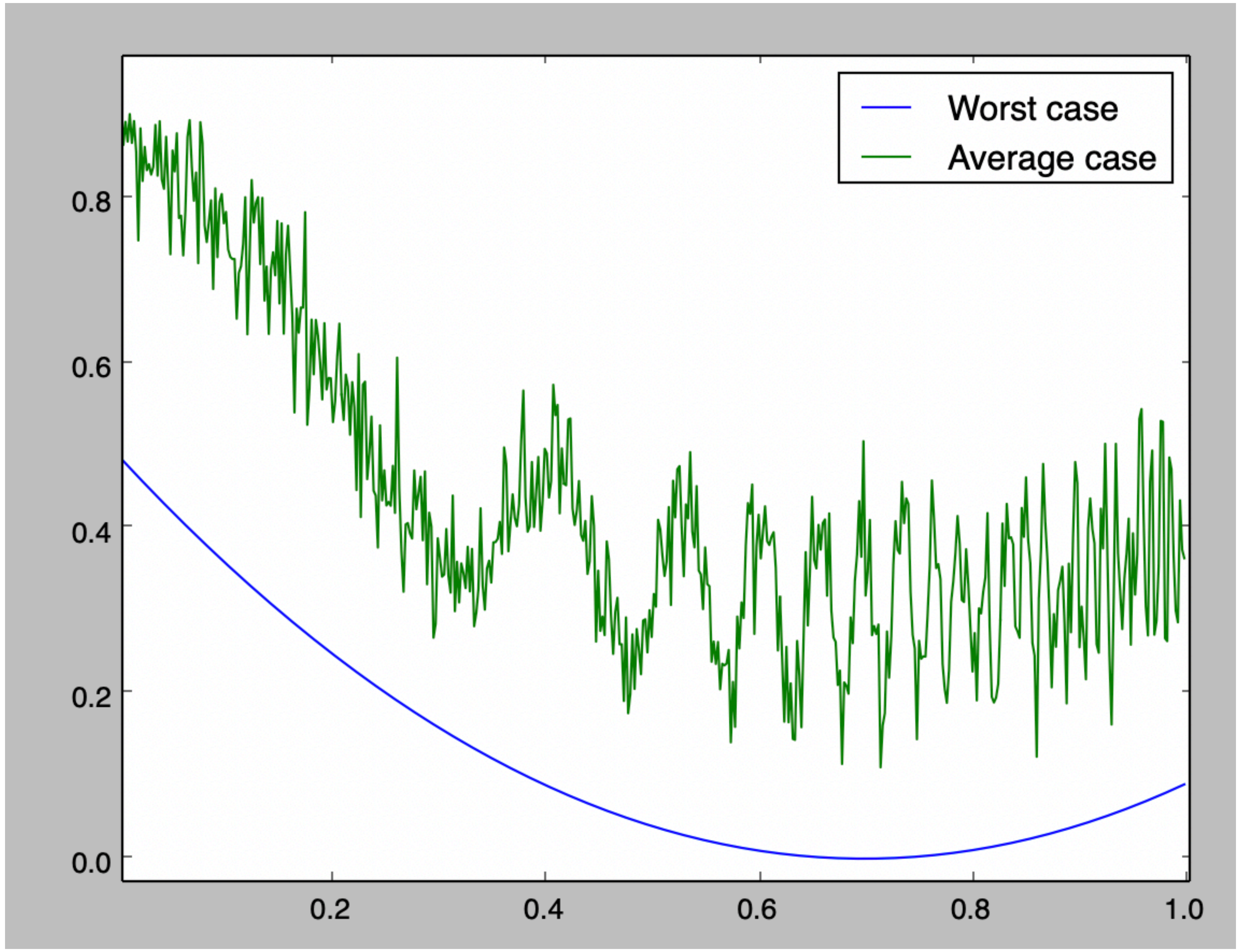
(Ganin et al, 2018)

A Cambrian Explosion of Machine Learning Research Topics



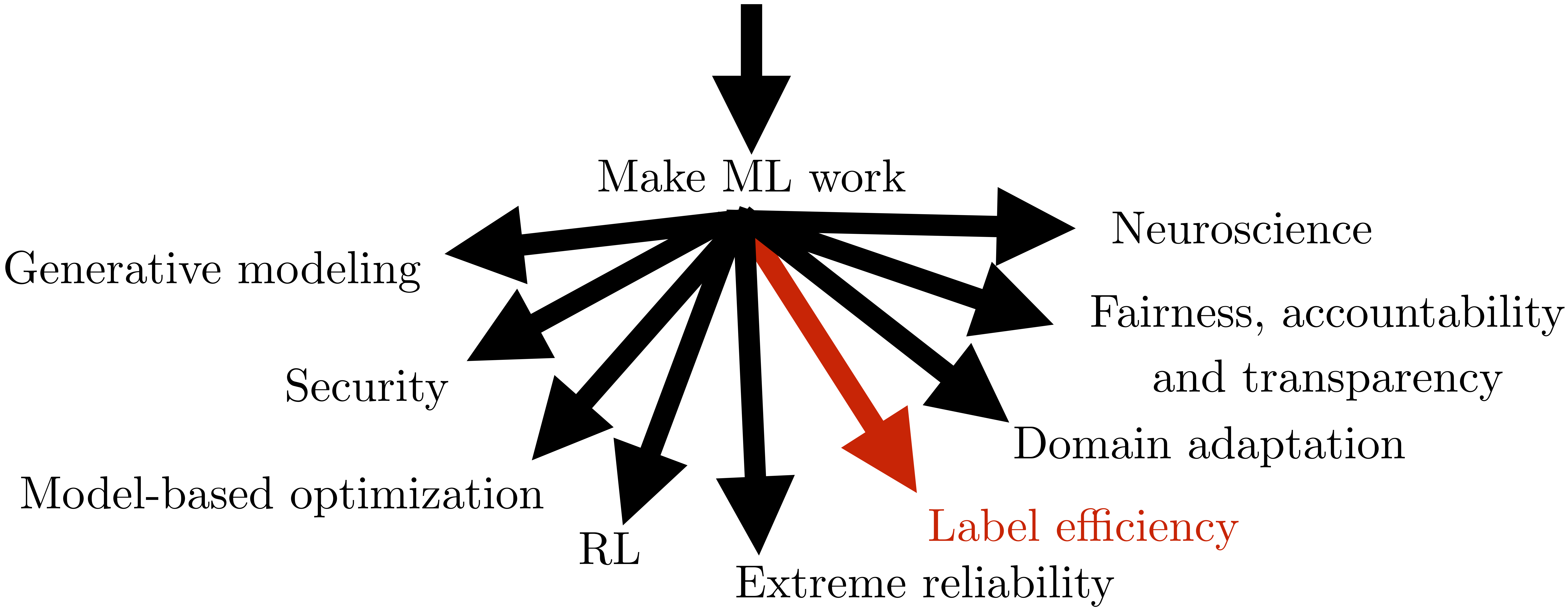
Extreme Reliability

Robustness and verification techniques
essential for air traffic control, surgery robots, etc.

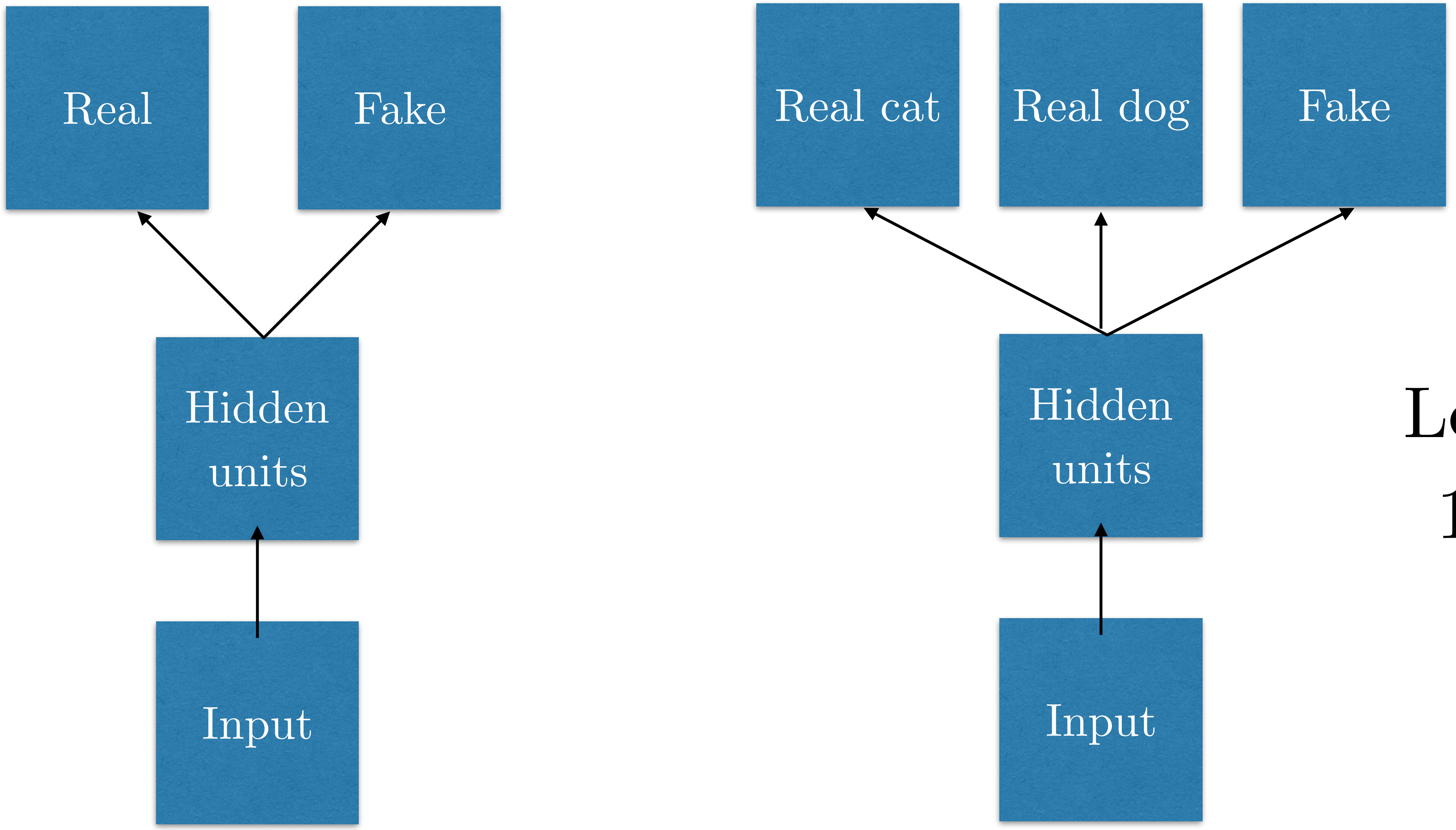


(Goodfellow 2019)

A Cambrian Explosion of Machine Learning Research Topics



Supervised Discriminator for Semi-Supervised Learning

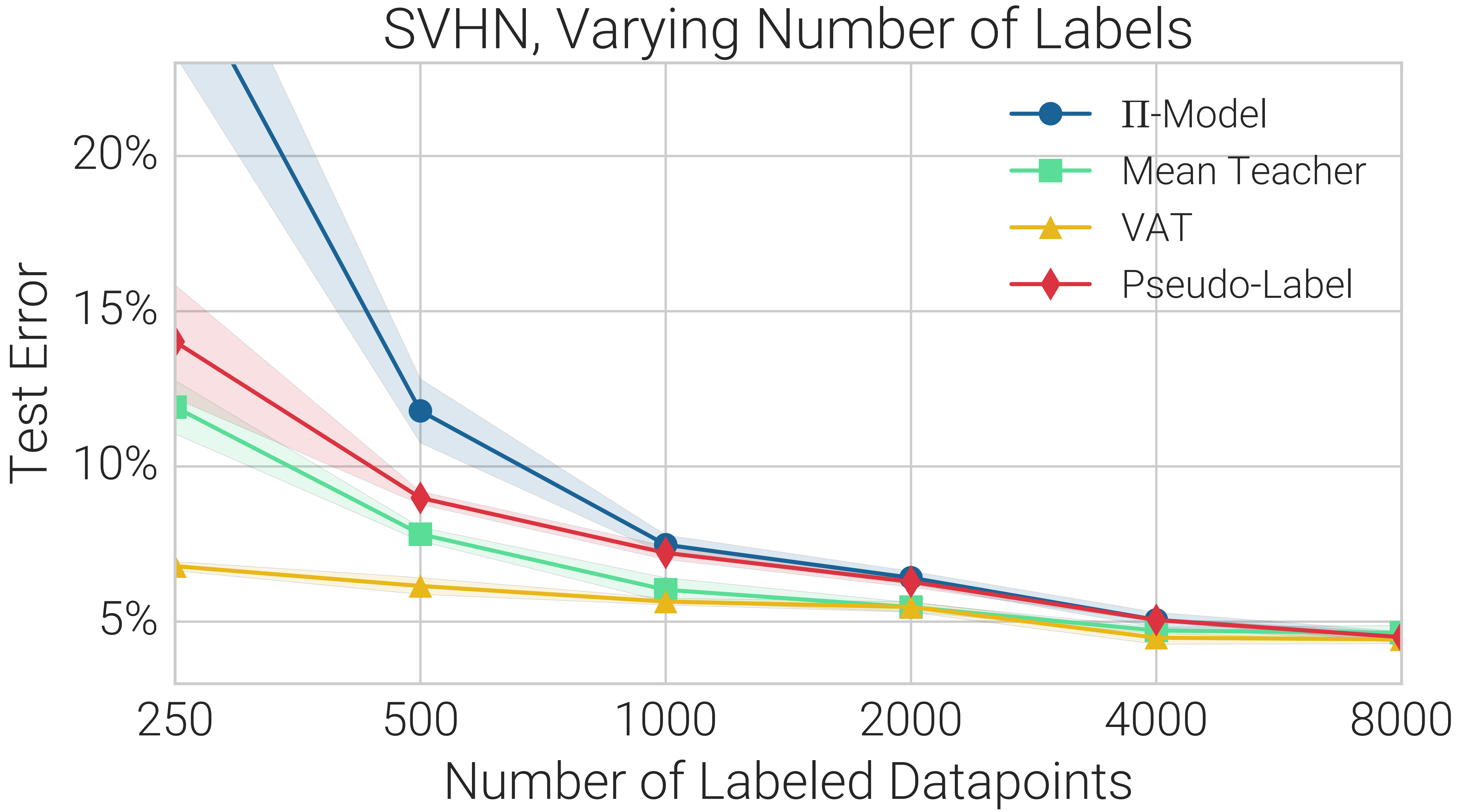


Learn to read with 100 labels rather than 60,000

(Odena 2016, Salimans et al 2016)

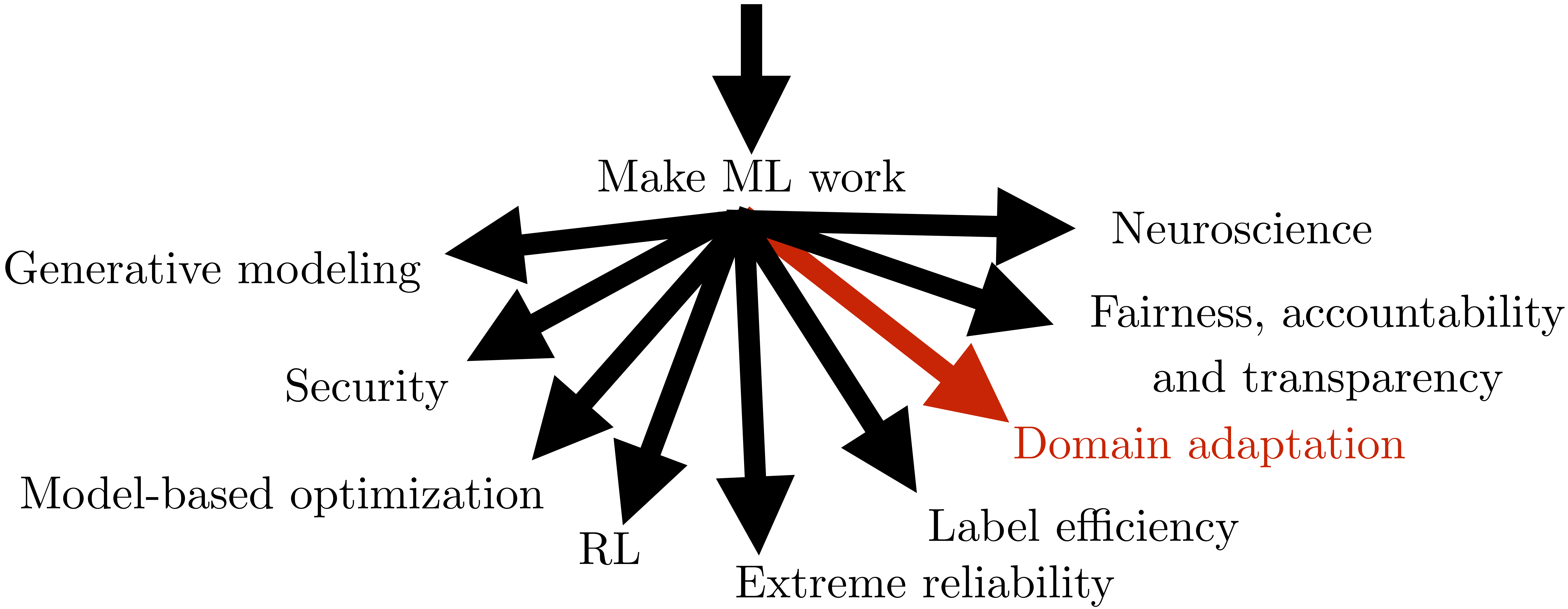
Virtual Adversarial Training

Miyato et al 2015: regularize for robustness to adversarial perturbations of *unlabeled* data



(Oliver+Odena+Raffel et al, 2018)

A Cambrian Explosion of Machine Learning Research Topics



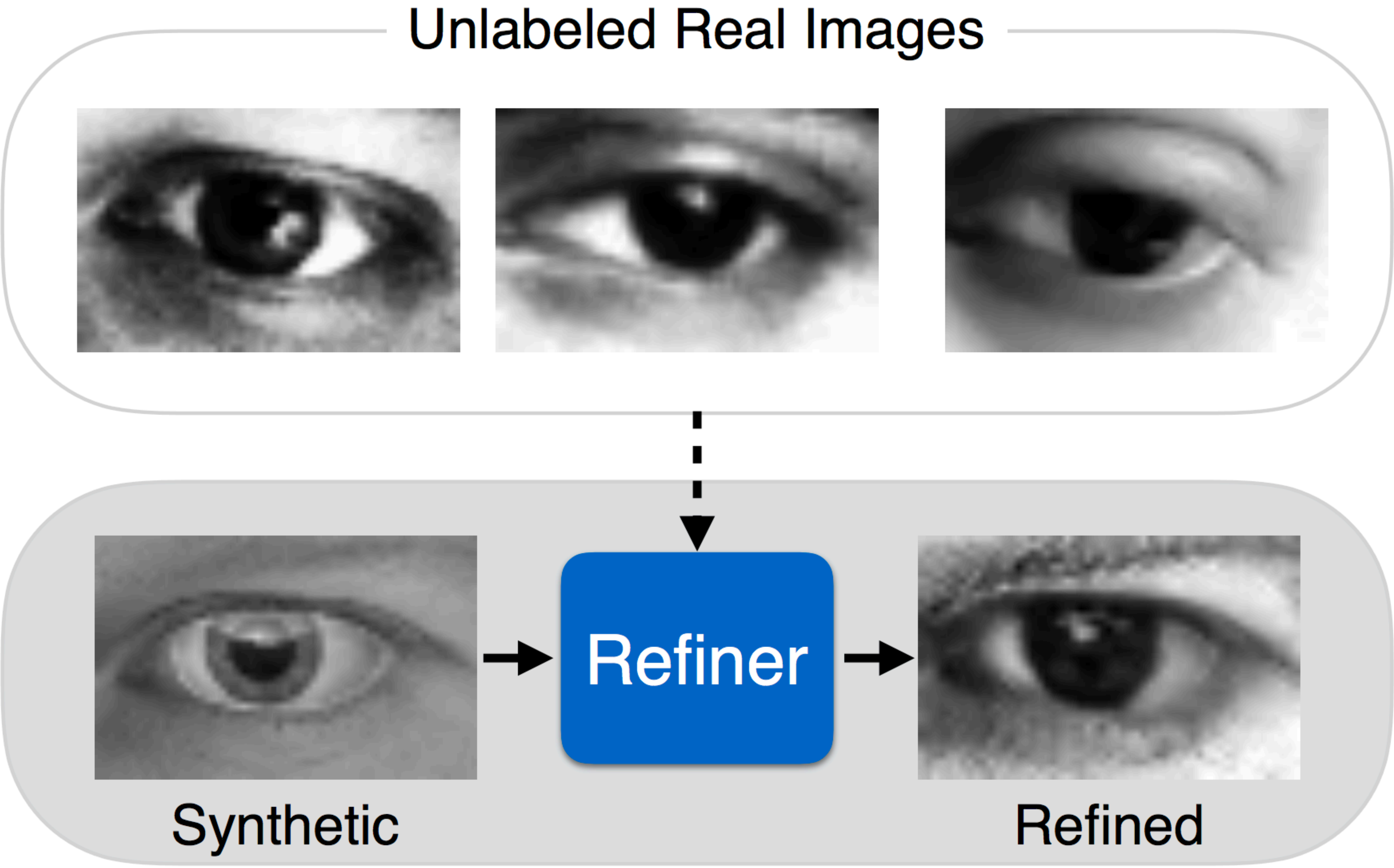
(Goodfellow 2019)

Domain Adaptation

- Domain Adversarial Networks (Ganin et al, 2015)

- Professor forcing (Lamb et al, 2016): Domain-Adversarial learning in RNN hidden state

GANs for simulated training data

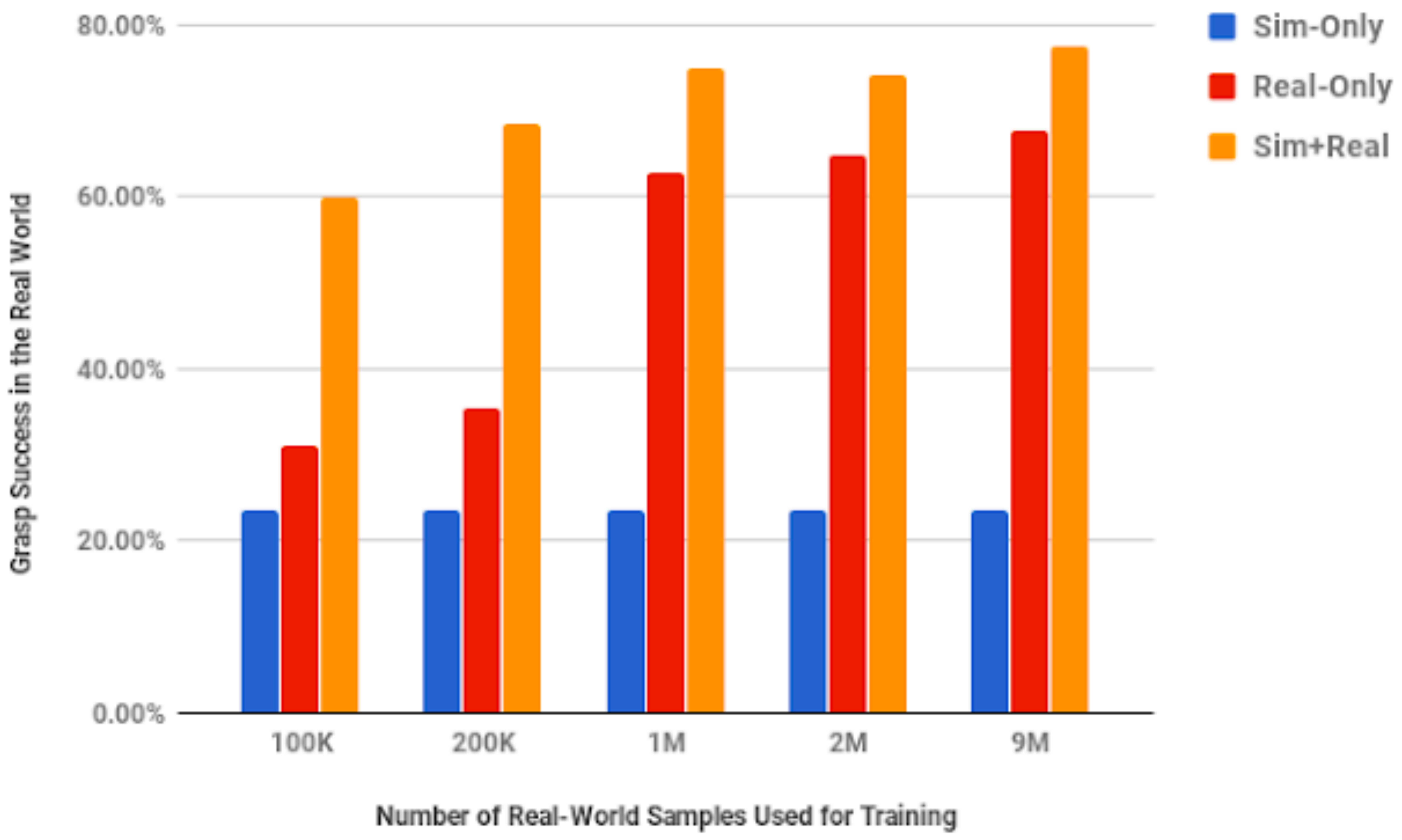


(Shrivastava et al., 2016)

GraspGAN

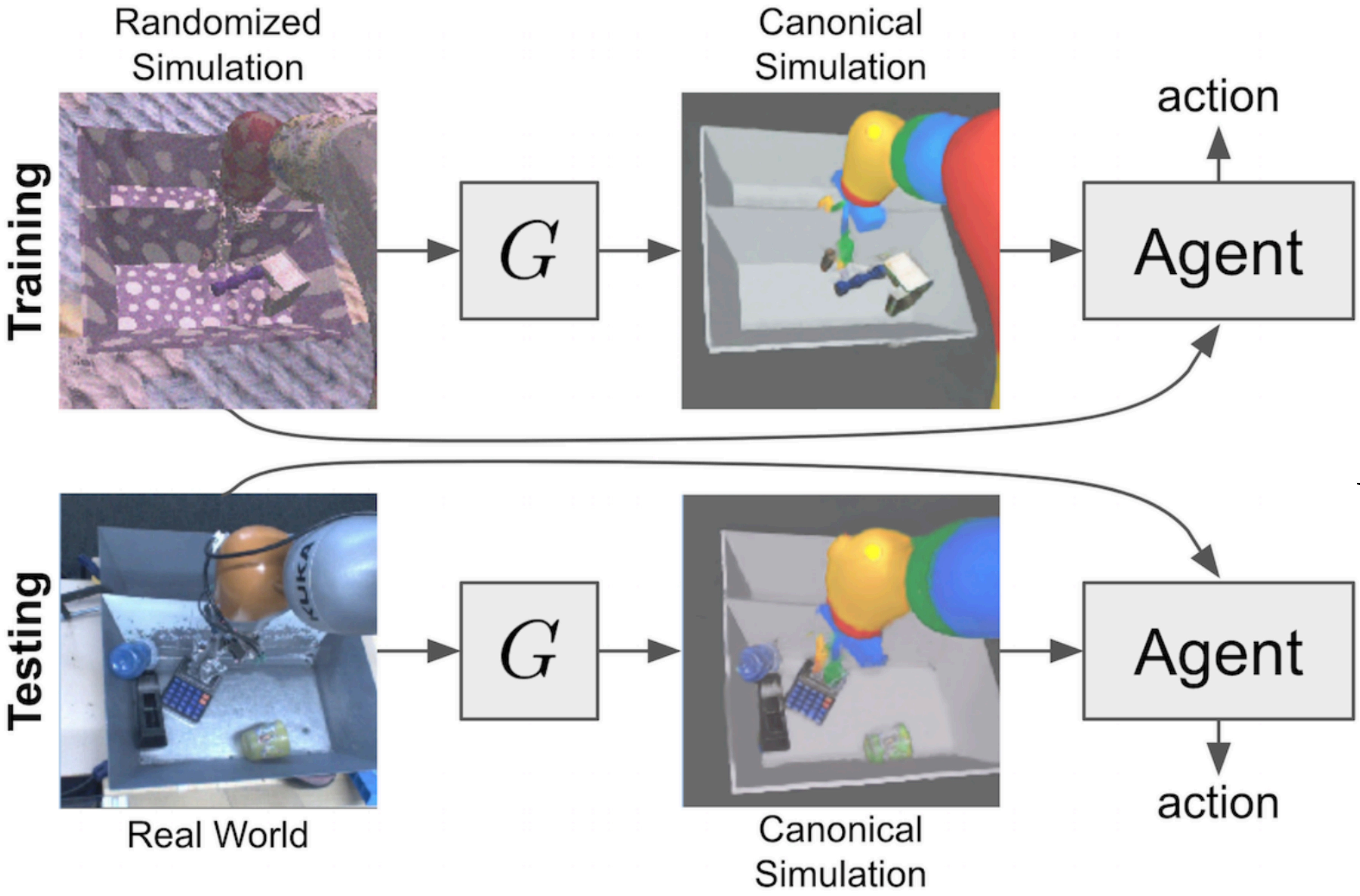
(Bousmalis et al, 2017)

GraspGAN



(Bousmalis et al, 2017)

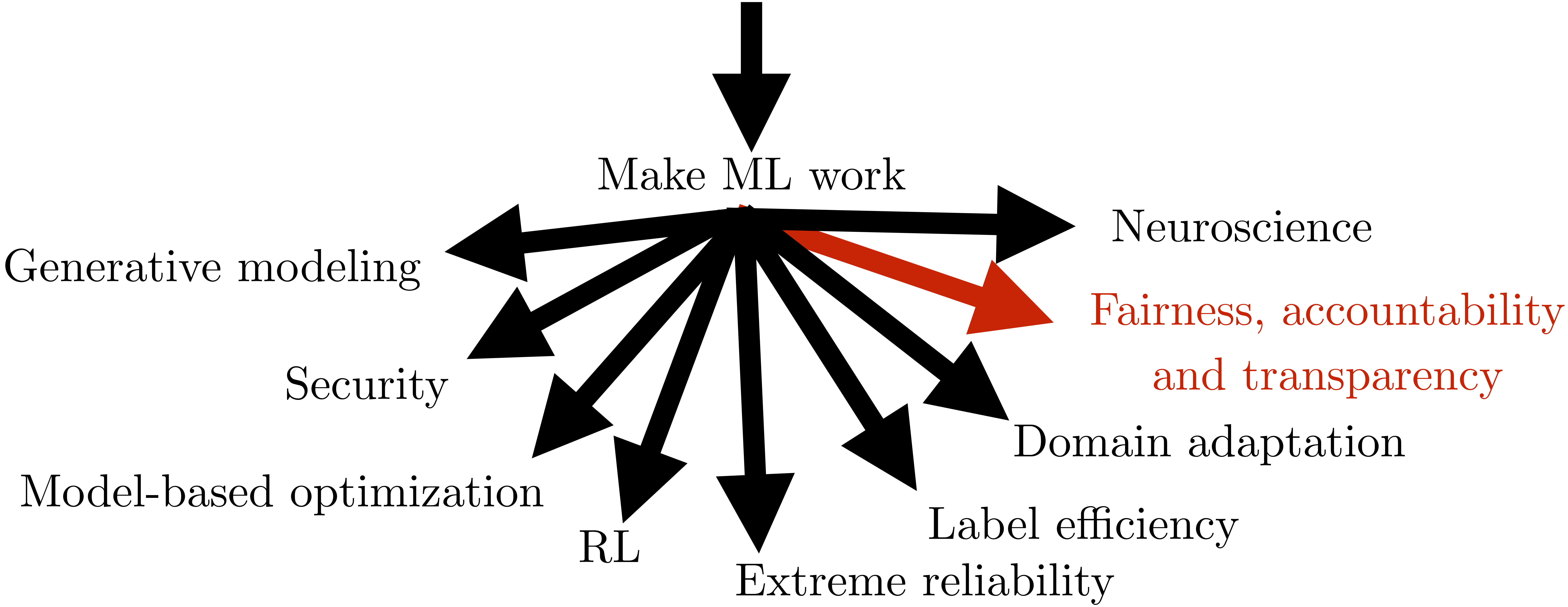
Sim-to-real via sim-to-sim



Learn to grasp without real data!

(James et al, 2018)

A Cambrian Explosion of Machine Learning Research Topics

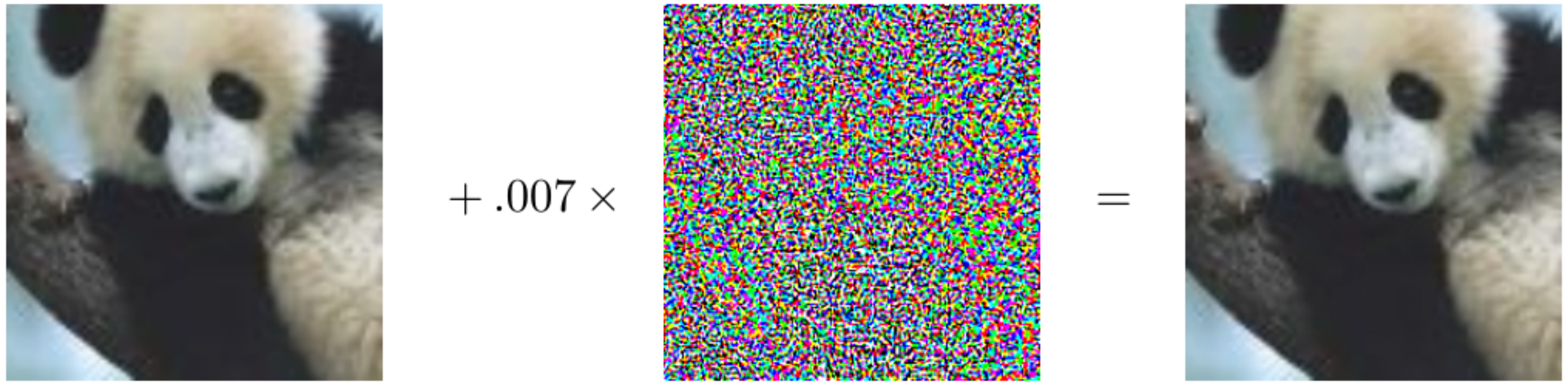


(Goodfellow 2019)

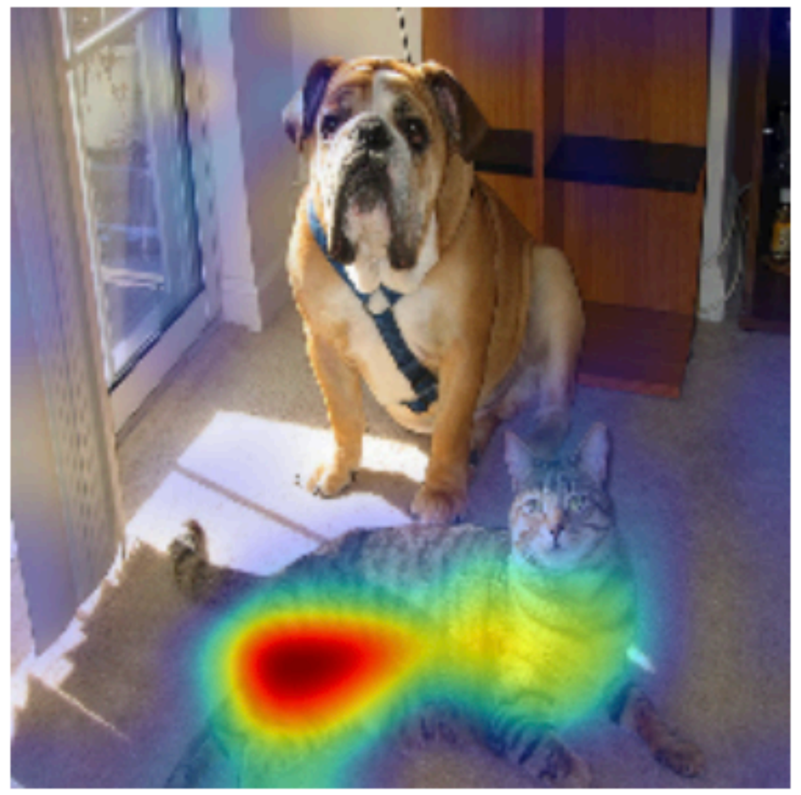
Adversarially Learned Fair Representations

- Edwards and Storkey 2015
- Learn representations that are useful for classification
- An adversary tries to recover a sensitive variable S from the representation. Primary learner tries to make S impossible to recover
- Final decision does not depend on S

How do machine learning models work?



(Goodfellow et al, 2014)



(c) Grad-CAM 'Cat'

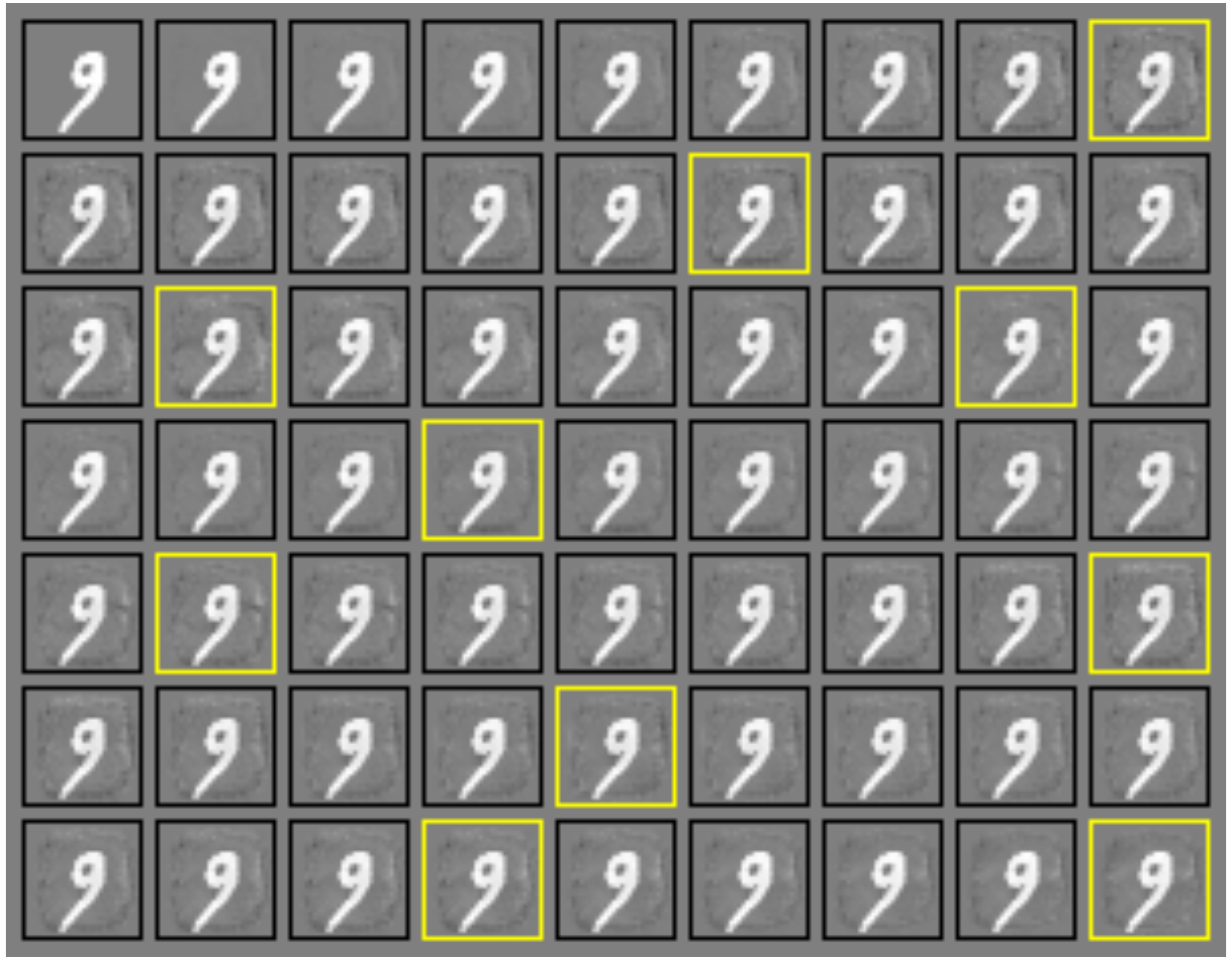
(i) Grad-CAM 'Dog'

(Selvaraju et al, 2016)

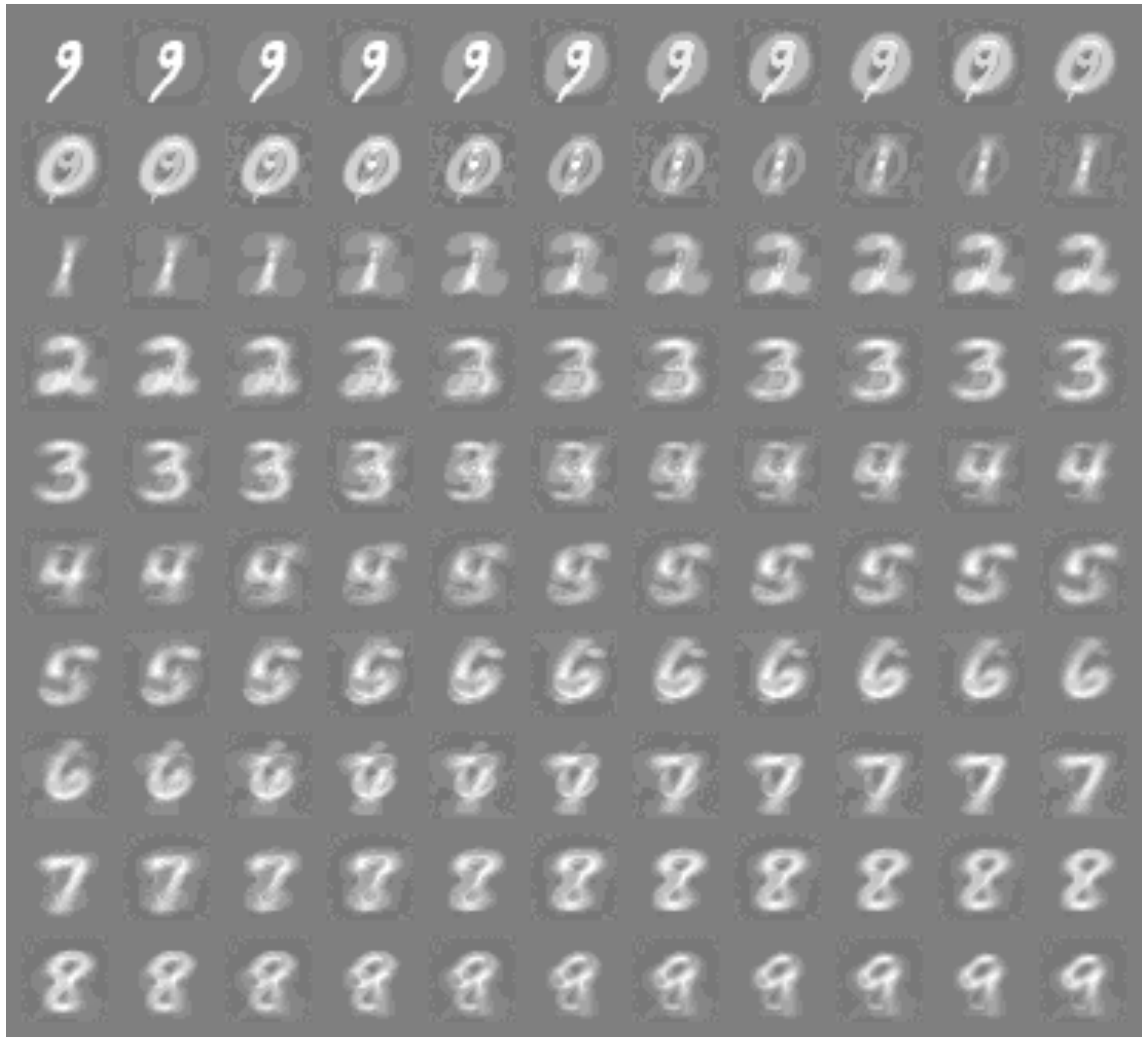
Interpretability literature: our analysis tools show that deep nets work about how you would expect them to.

Adversarial ML literature: ML models are very easy to fool and even linear models work in counter-intuitive ways.

Robust models are more interpretable



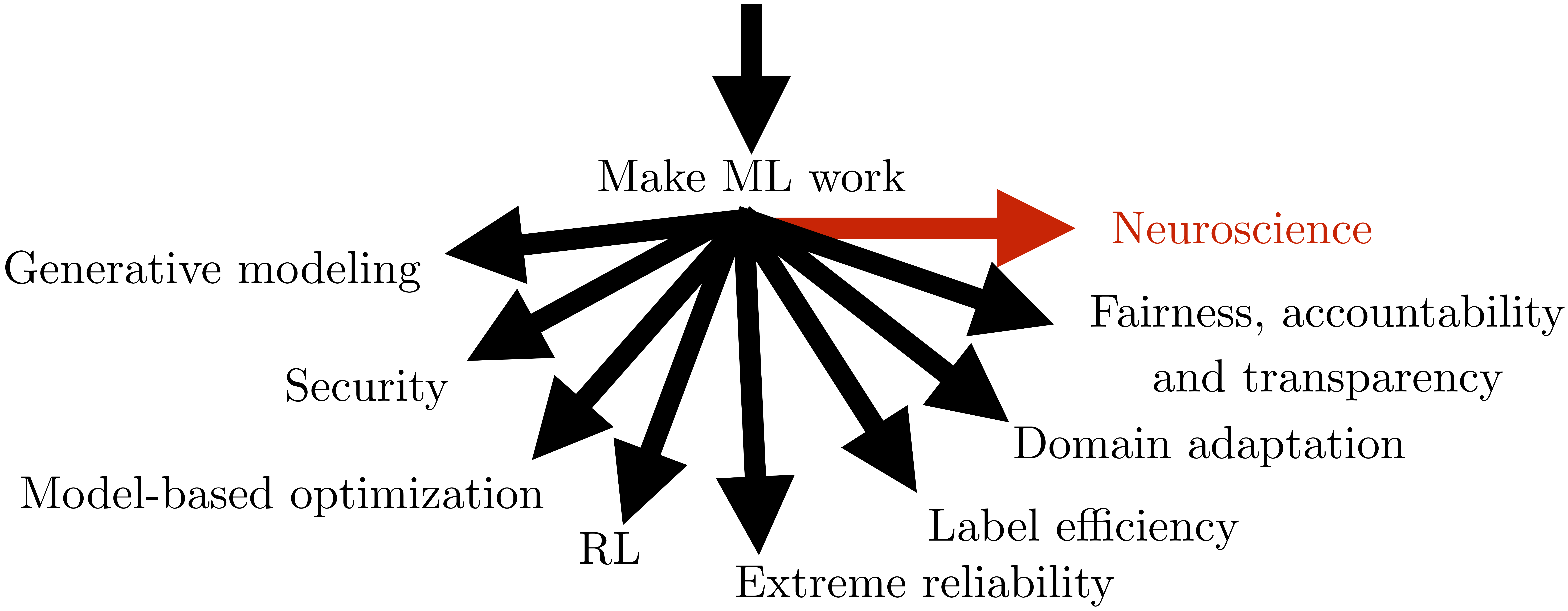
Relatively vulnerable model



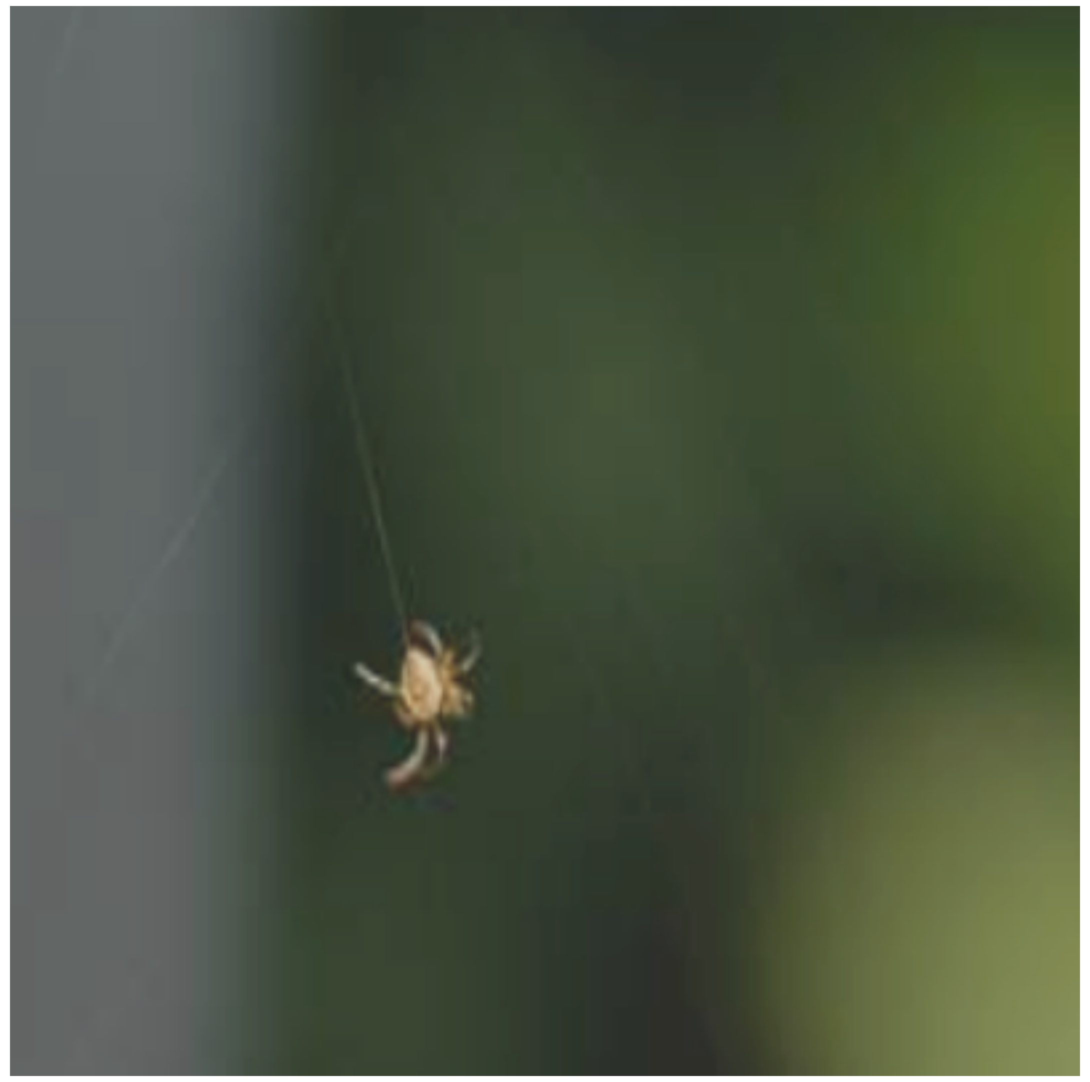
Relatively robust model

(Goodfellow 2015)

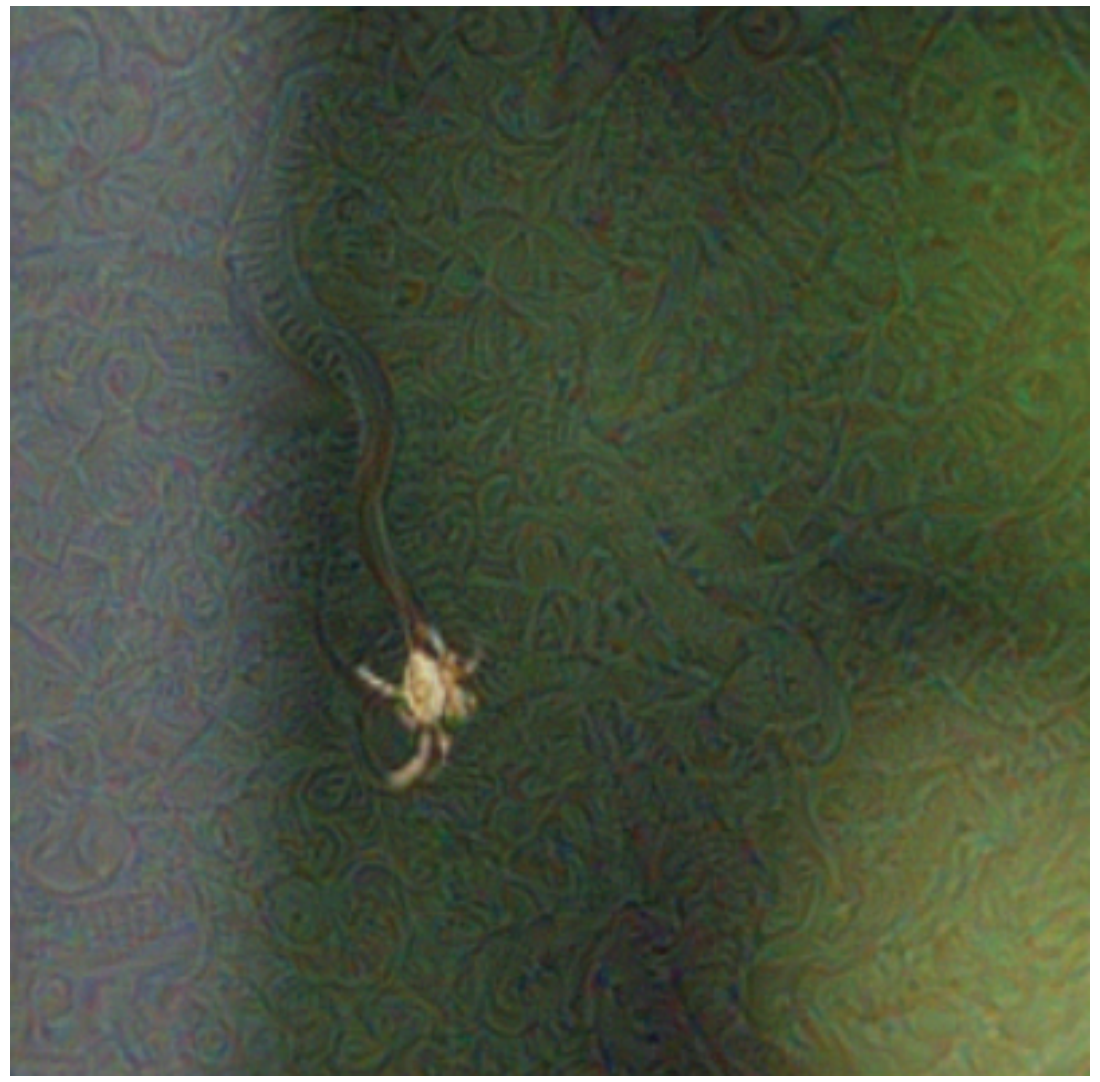
A Cambrian Explosion of Machine Learning Research Topics



Adversarial examples that affect both computer and time-limited human vision



25% snake



67% snake

Elsayed et al 2018

Questions