Adversarial Machine Learning

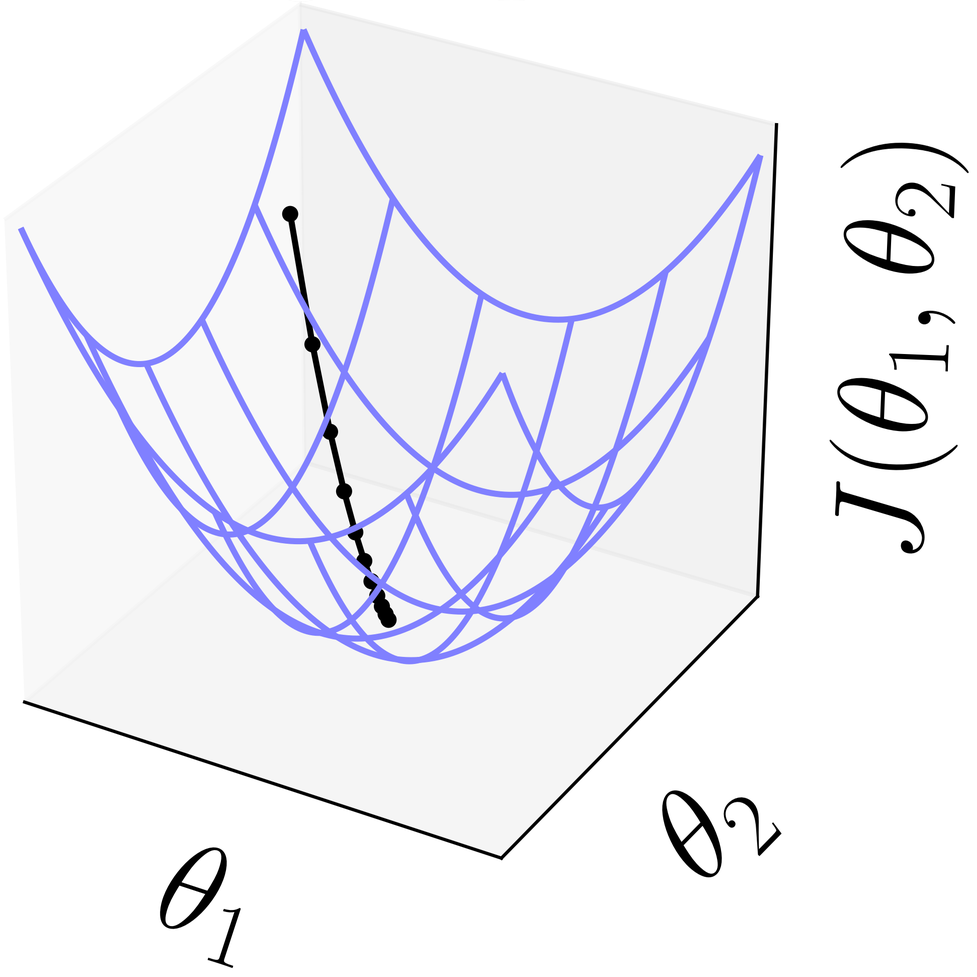
Ian Goodfellow, Senior Staff Research Scientist

AAAI

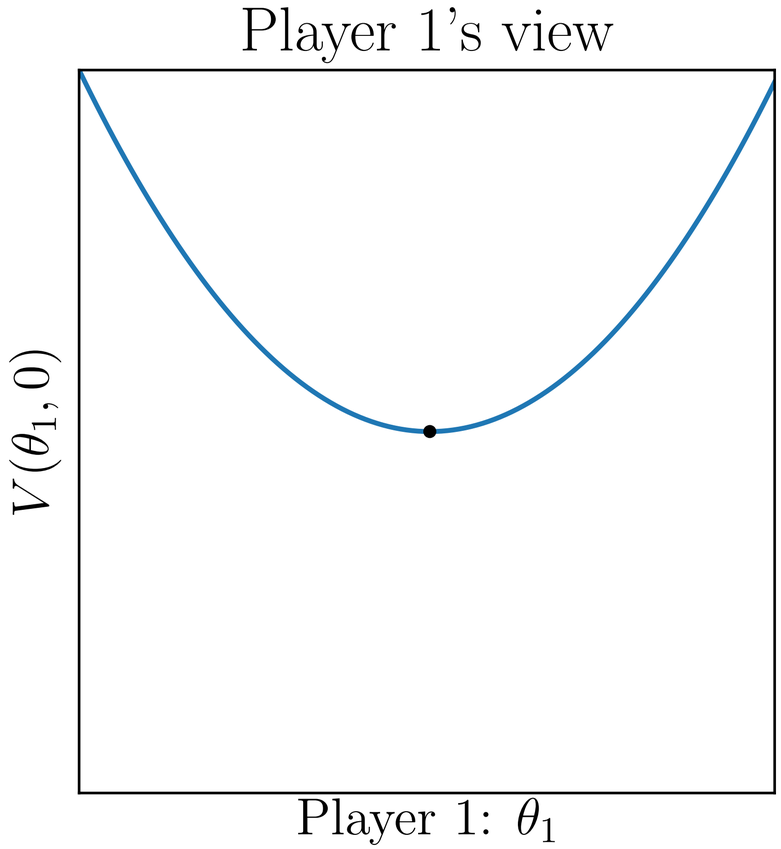
2019-01-30

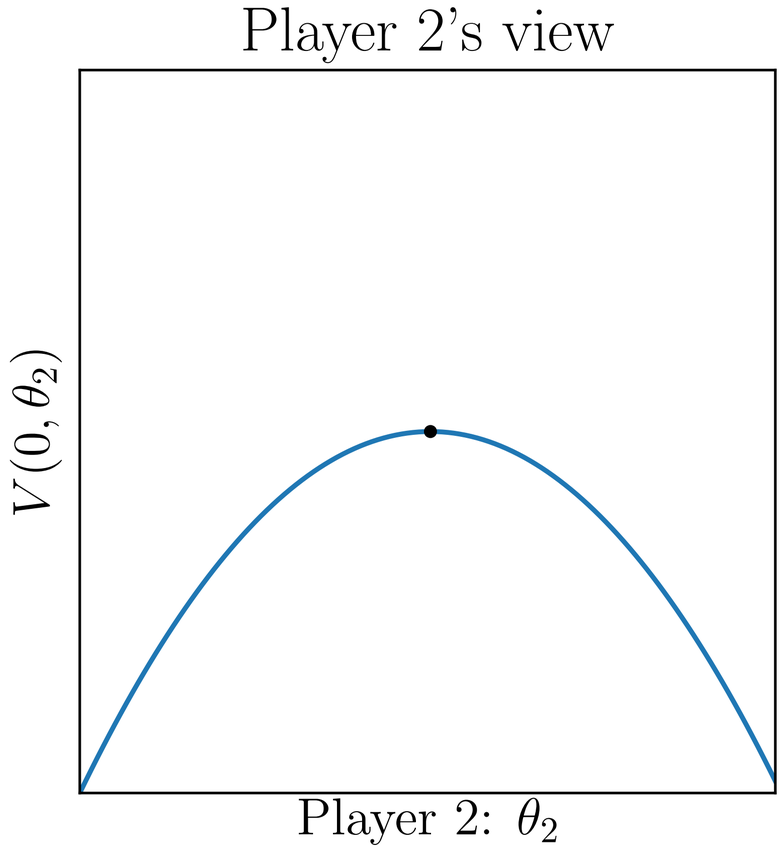
Google AI

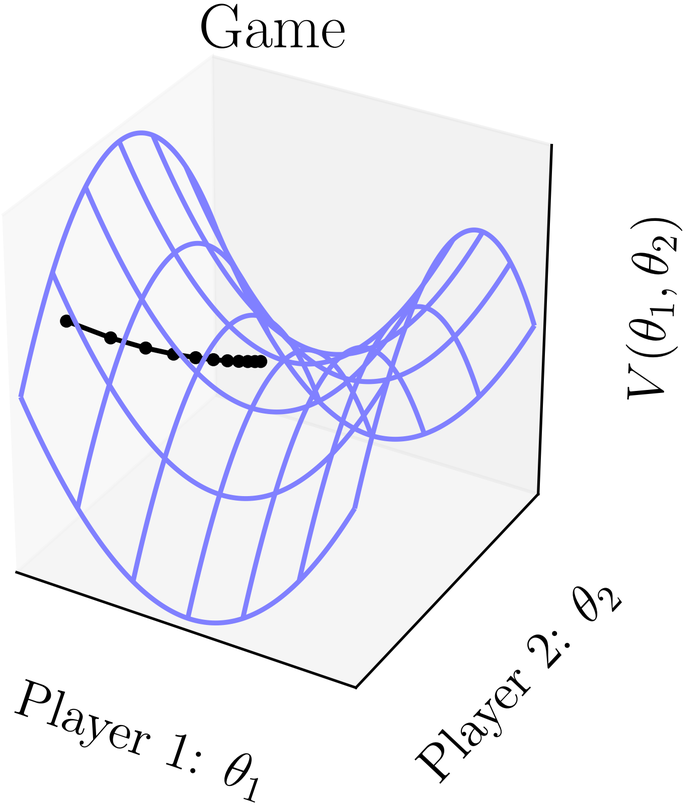
Most Traditional Machine Learning: Optimization



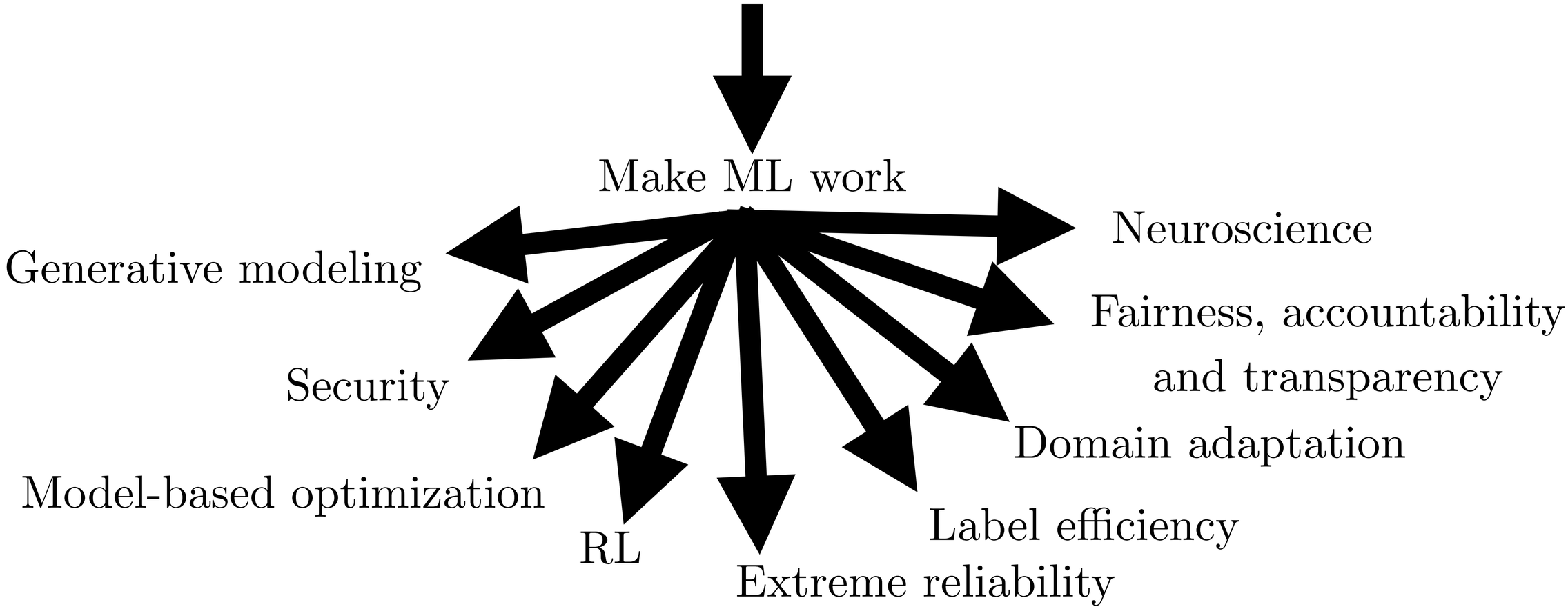
Adversarial Machine Learning: Game Theory



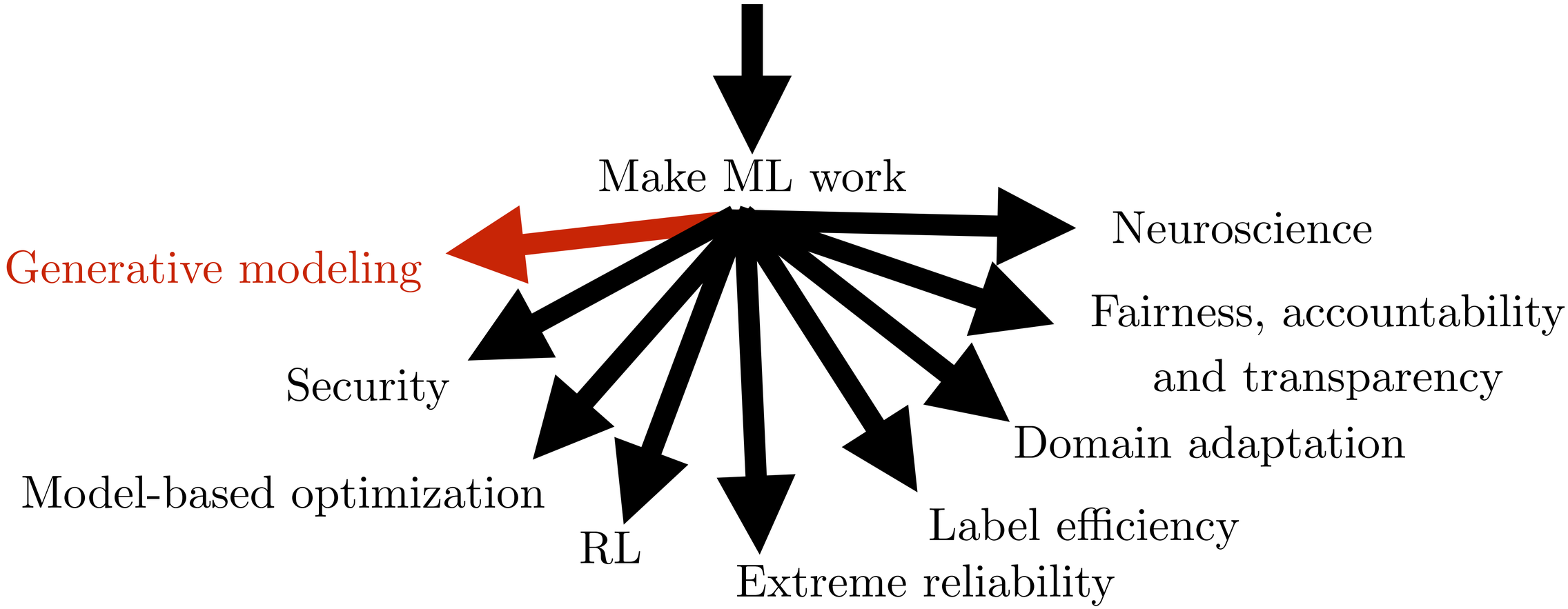




A Cambrian Explosion of Machine Learning Research Topics



A Cambrian Explosion of Machine Learning Research Topics

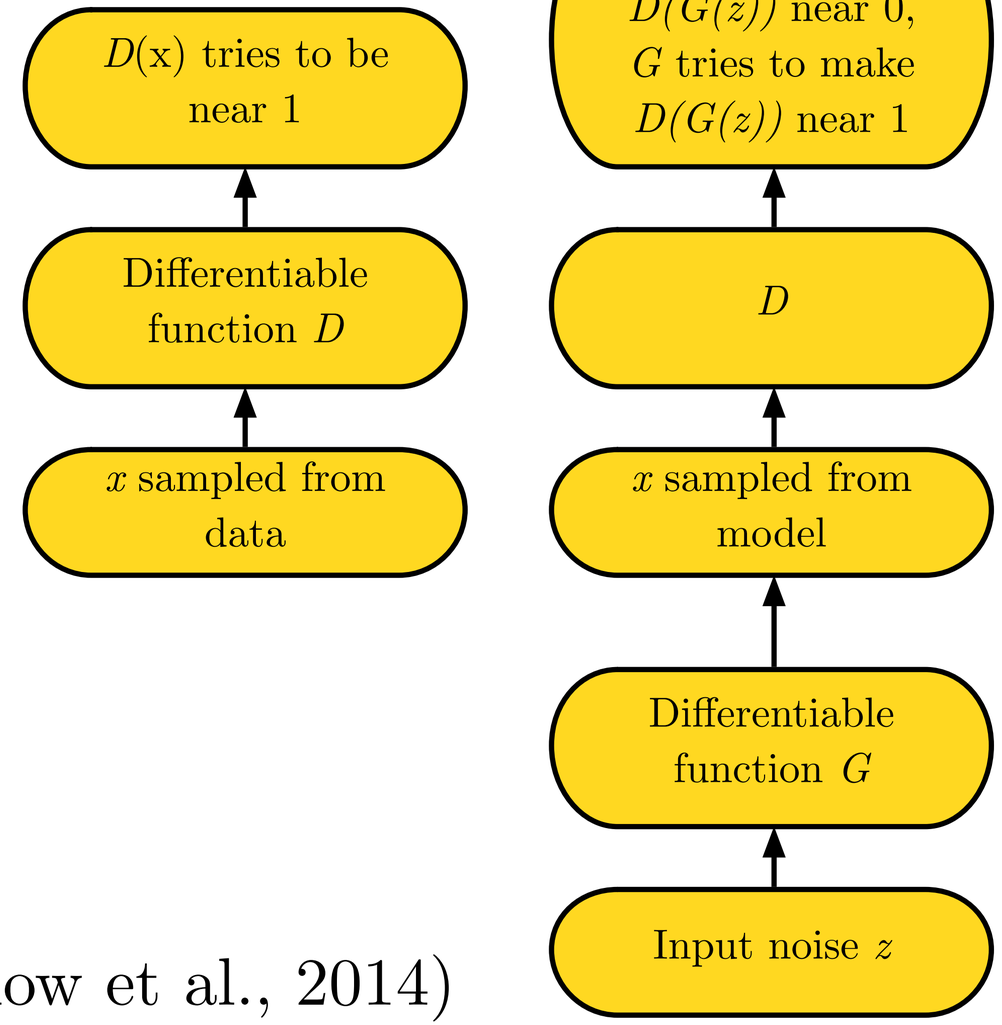


Generative Modeling: Sample Generation

Training Data
(CelebA)

Sample Generator
(Karras et al, 2017)

Adversarial Nets Framework

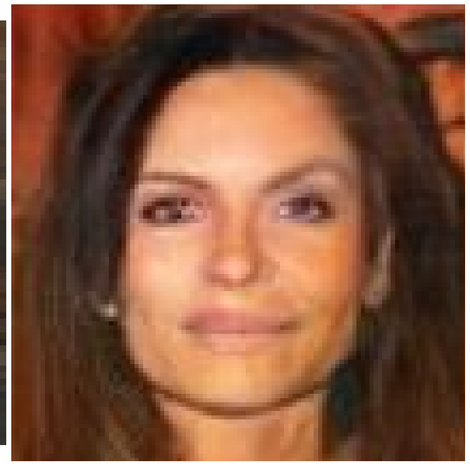


(Goodfellow et al., 2014)

4.5 years of progress on faces

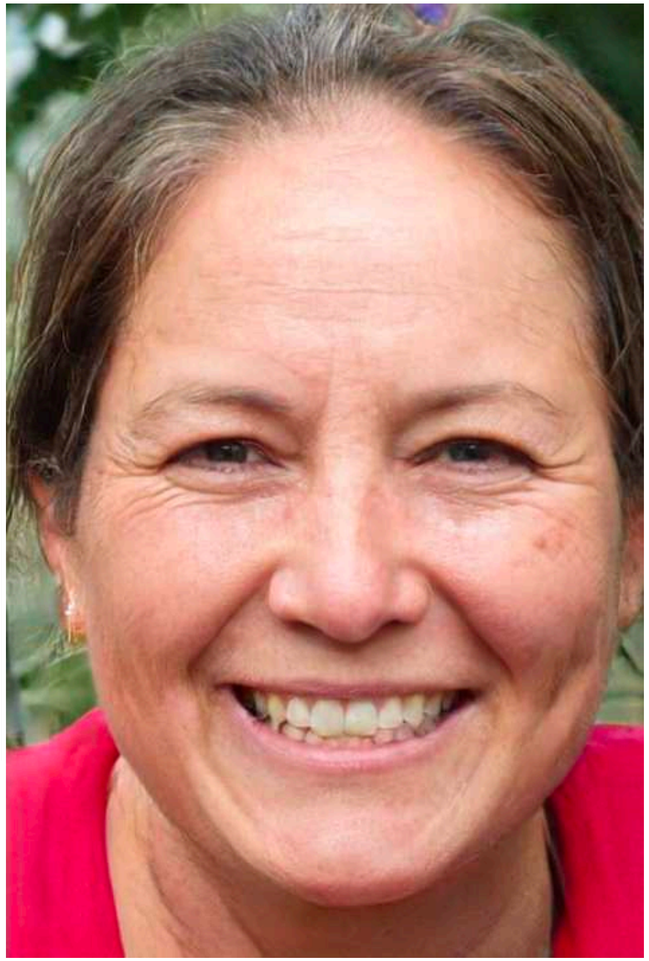
2014

2015



2016

2017



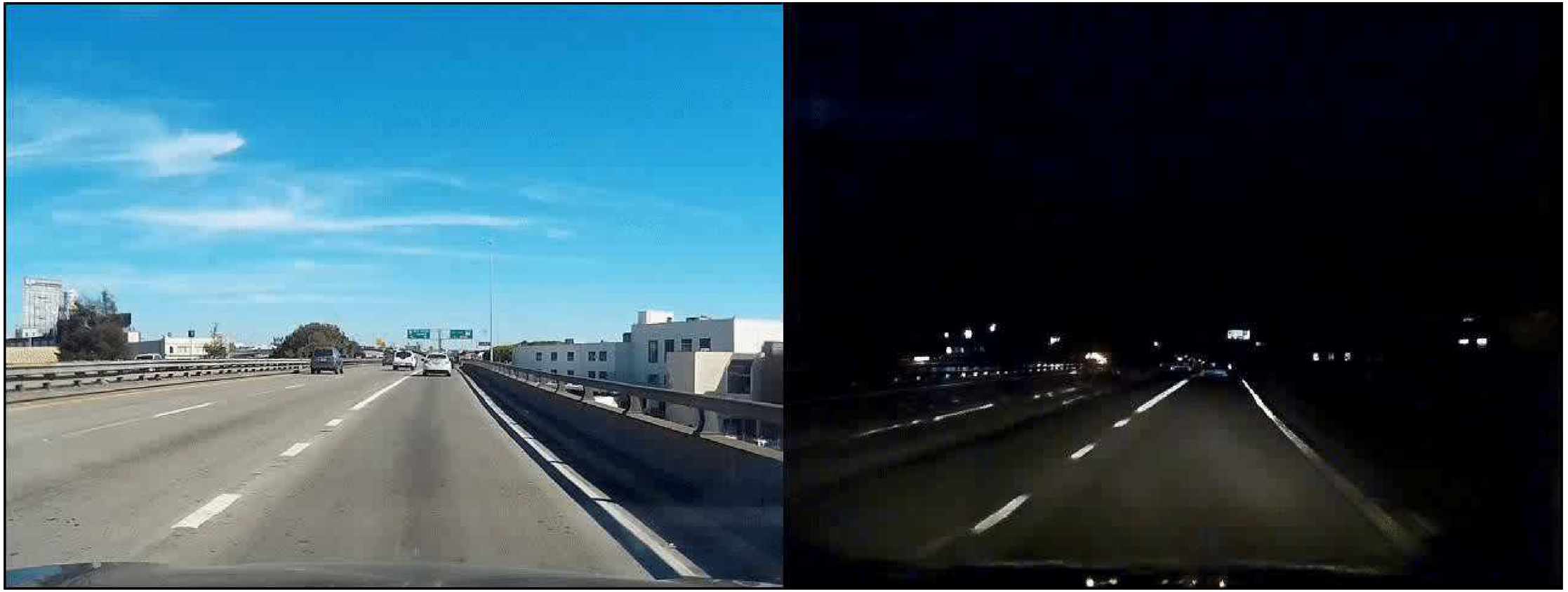
2018

2 Years of Progress on ImageNet

(Odena 2018)

Unsupervised Image-to-Image Translation

Day to night

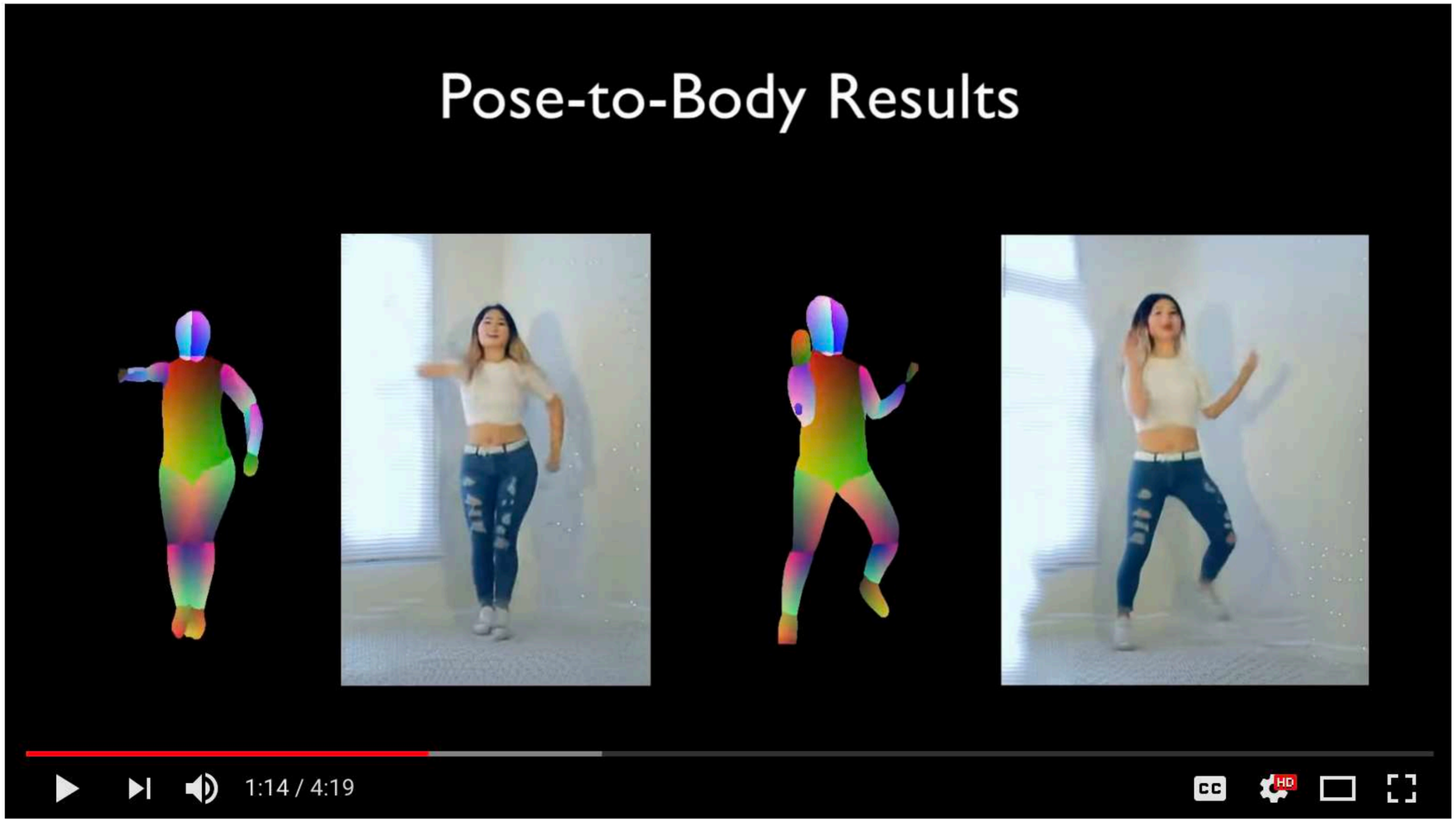


(Liu et al., 2017)

CycleGAN

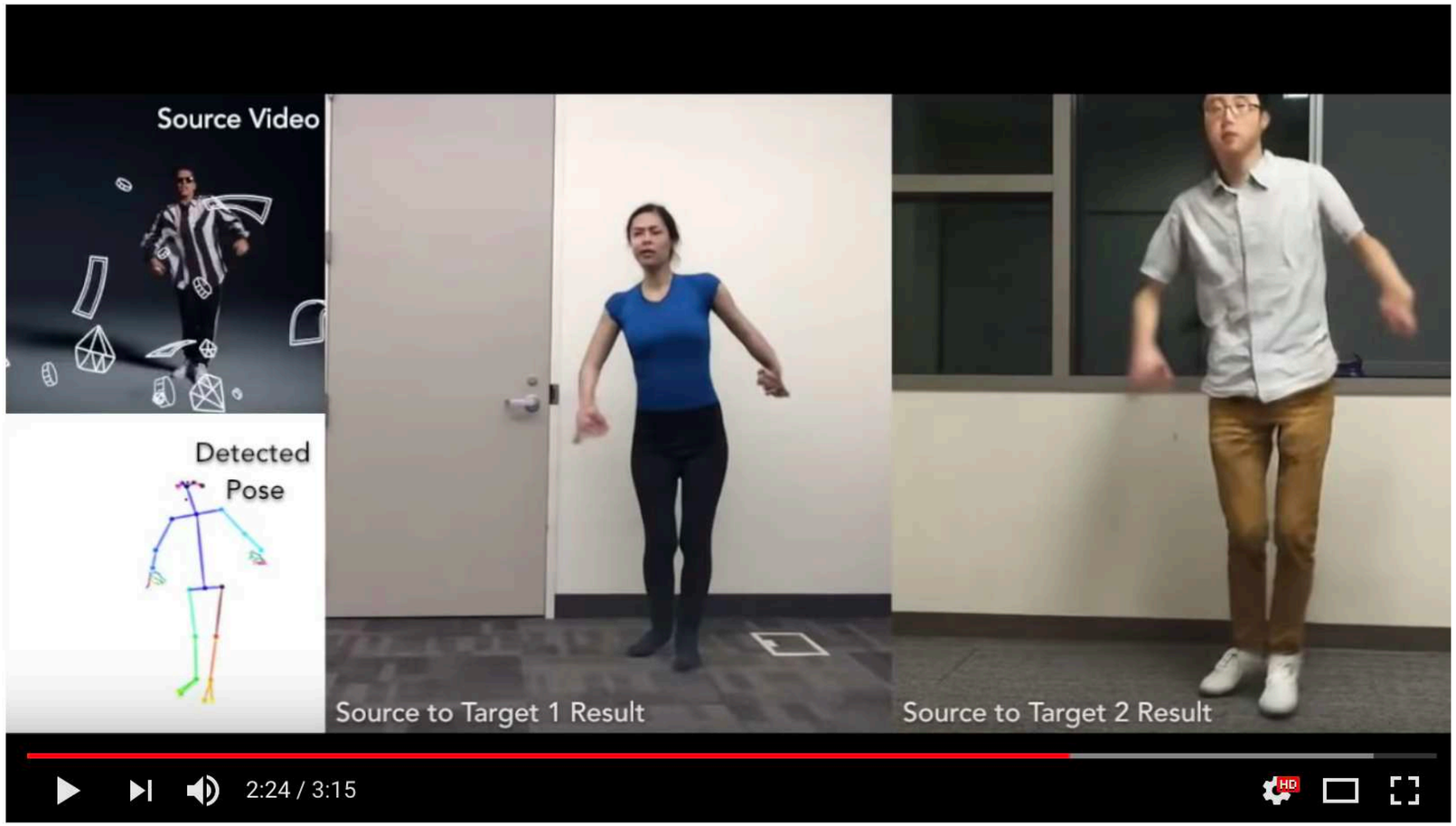
(Zhu et al., 2017)

Video-to-Video



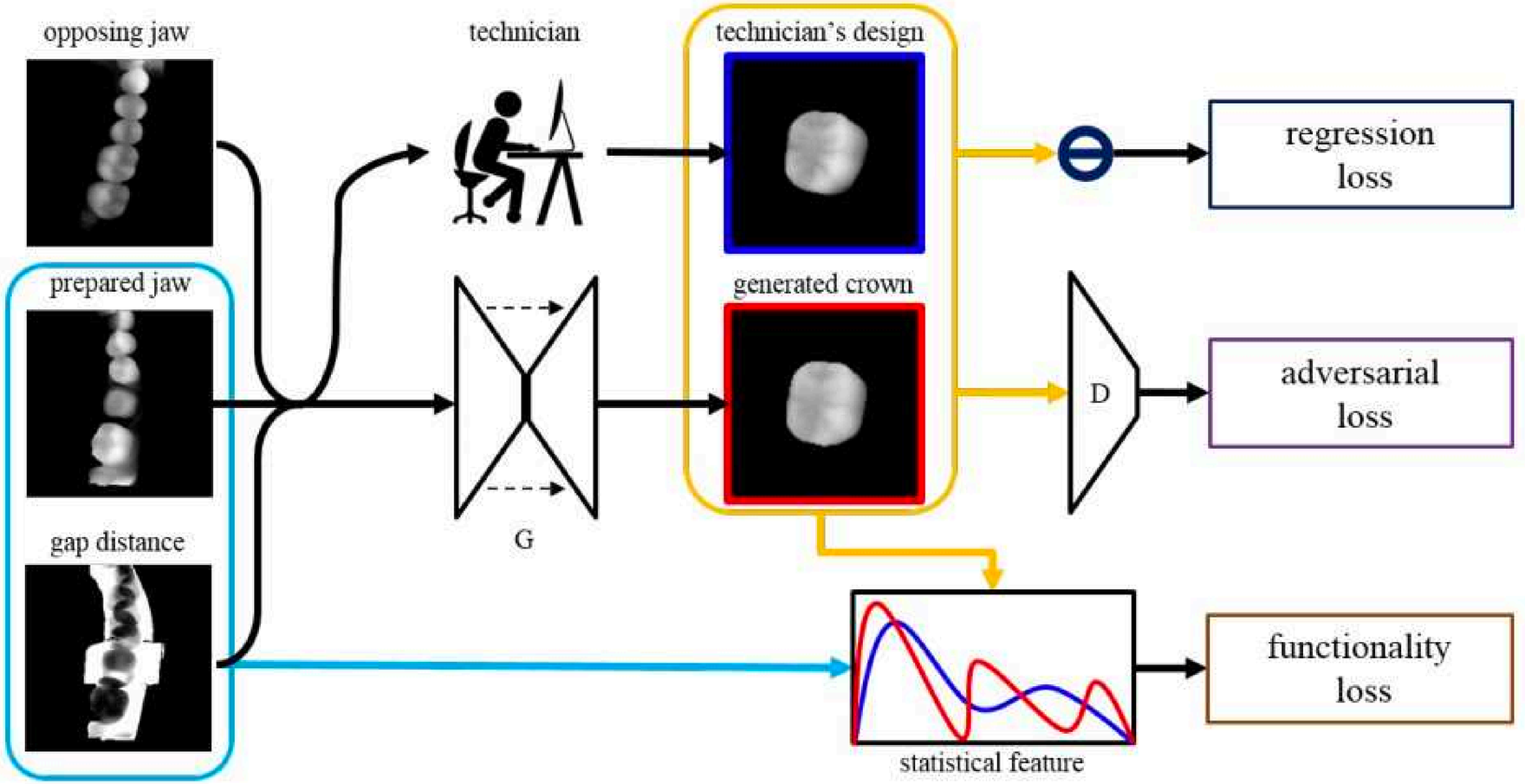
(Wang et al, 2018)

Everybody Dance Now



(Chan et al 2018)

Personalized GANufacturing



(Hwang et al 2018)

Self-Attention



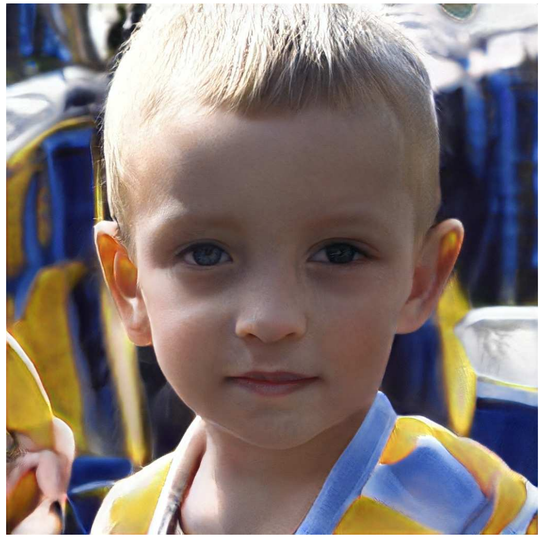

(Zhang et al., 2018)

Use layers from
Wang et al 2018


Recent Advances

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

Starting sample (Fake)

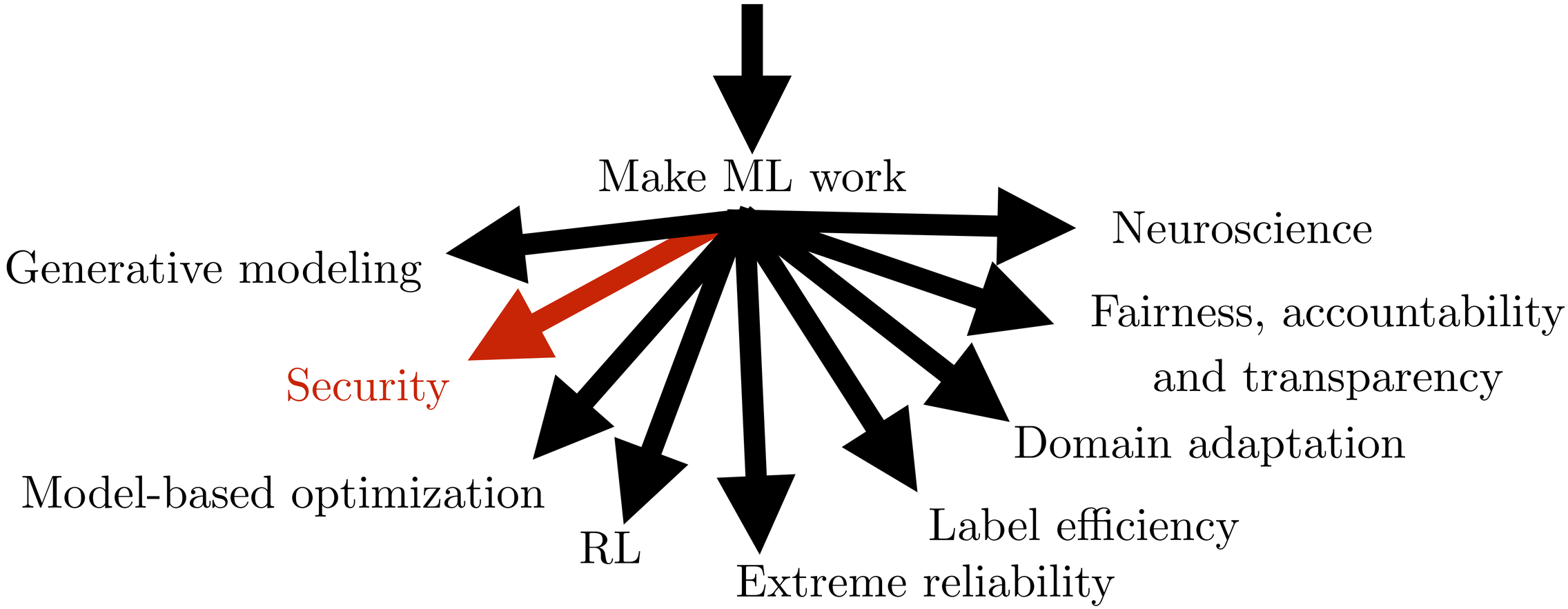


Sample for coarse style (also fake)

Result (still fake)

Style-based generators
(Karras et al, 2018)

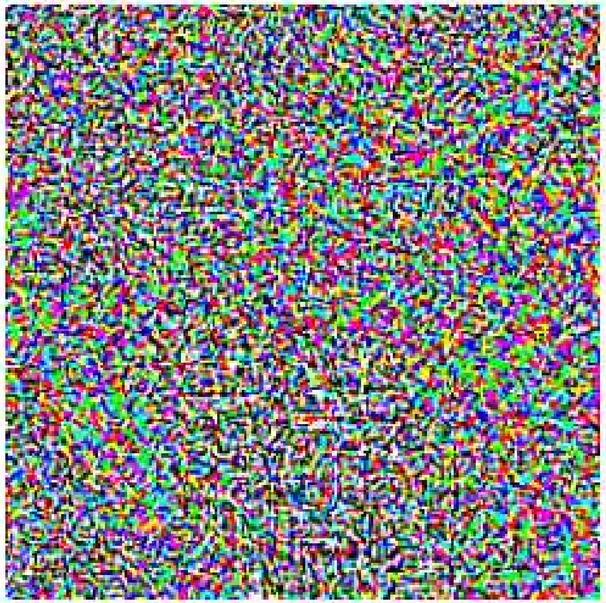
A Cambrian Explosion of Machine Learning Research Topics



Adversarial Examples

58% panda

$+ .007 \times$

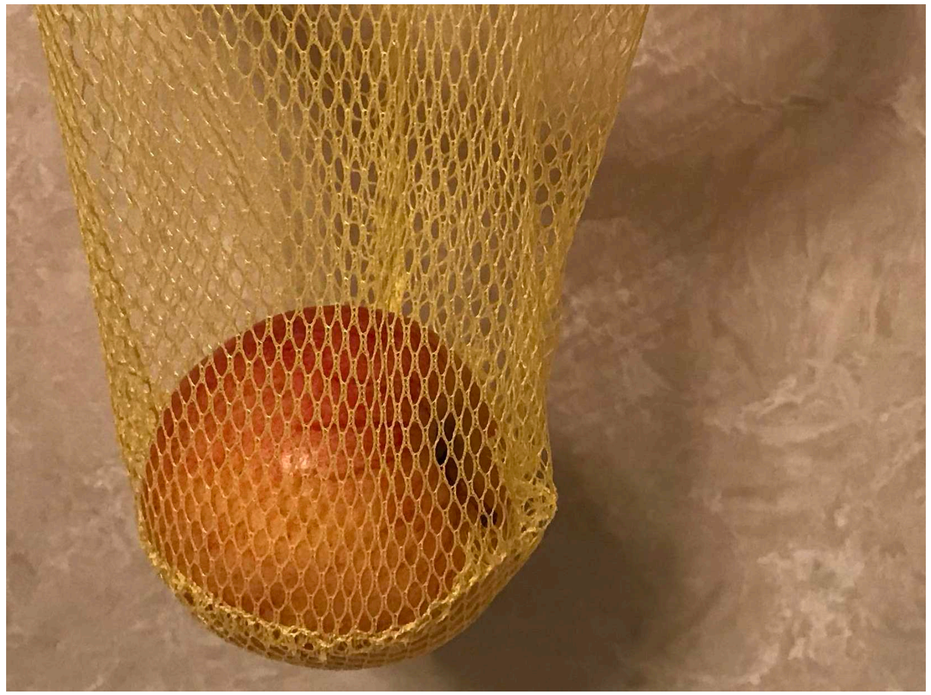


$=$

99% gibbon

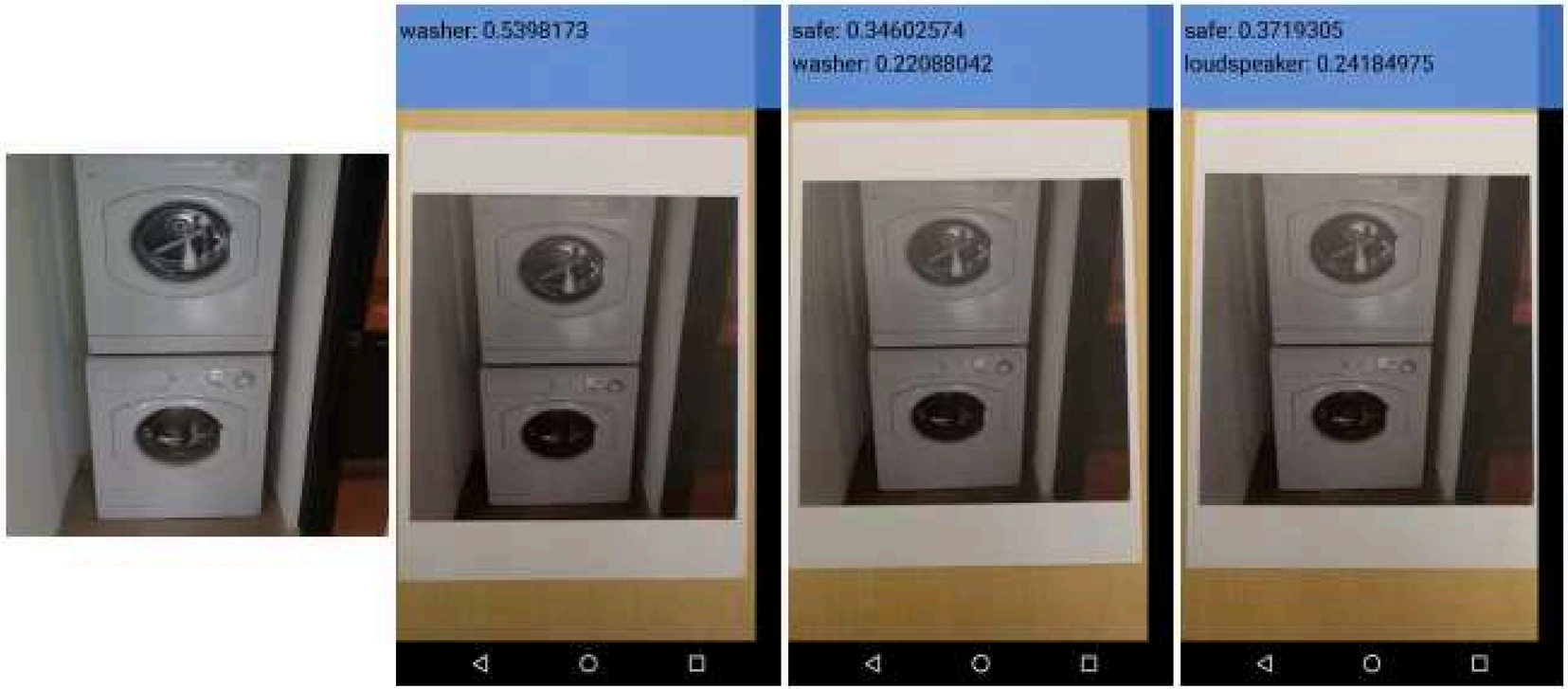
Also Adversarial Examples

(Eykholt et al, 2017)



(Goodfellow 2018)

Adversarial Examples in the Physical World



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

(Kurakin et al, 2016)

Adversarial Training as a Minimax Problem

"Adversarial training can be interpreted as a minimax game,

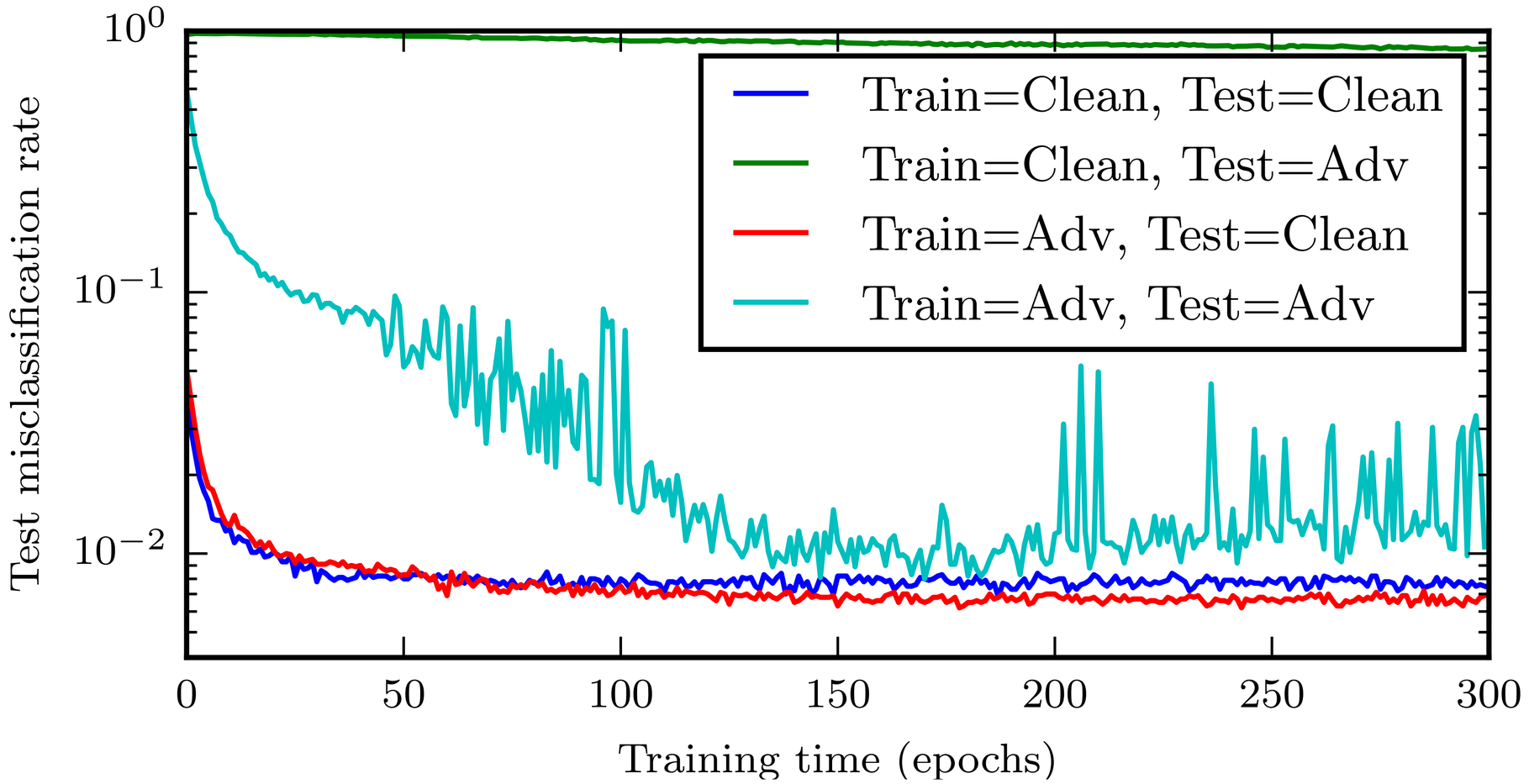
$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x},y} \max_{\boldsymbol{\eta}} \left[J(\boldsymbol{x}, y, \boldsymbol{\theta}) + J(\boldsymbol{x} + \boldsymbol{\eta}, y) \right],$$

with the learning algorithm as the minimizing player and a fixed procedure (such as L-BFGS or the fast gradient sign method) as the maximizing player."

Original implementation: Goodfellow et al 2014

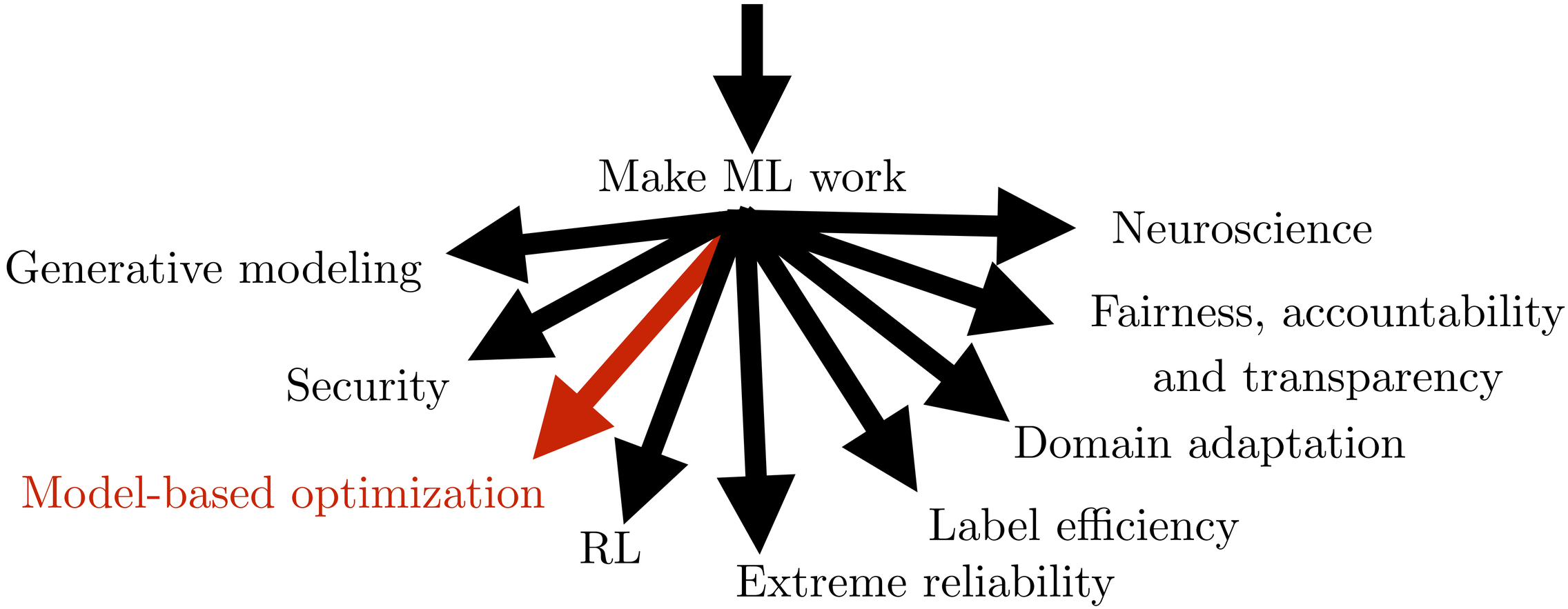
Explicit use of "minimax": Farley and Goodfellow, 2016

Training on Adversarial Examples



(CleverHans tutorial, using method of Goodfellow et al 2014)

A Cambrian Explosion of Machine Learning Research Topics

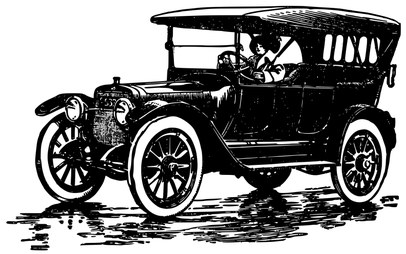


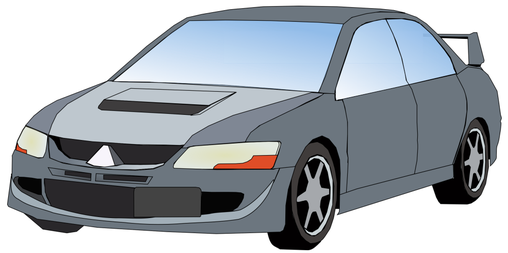
Model-Based Optimization

Make new inventions by finding input that maximizes model's predicted performance

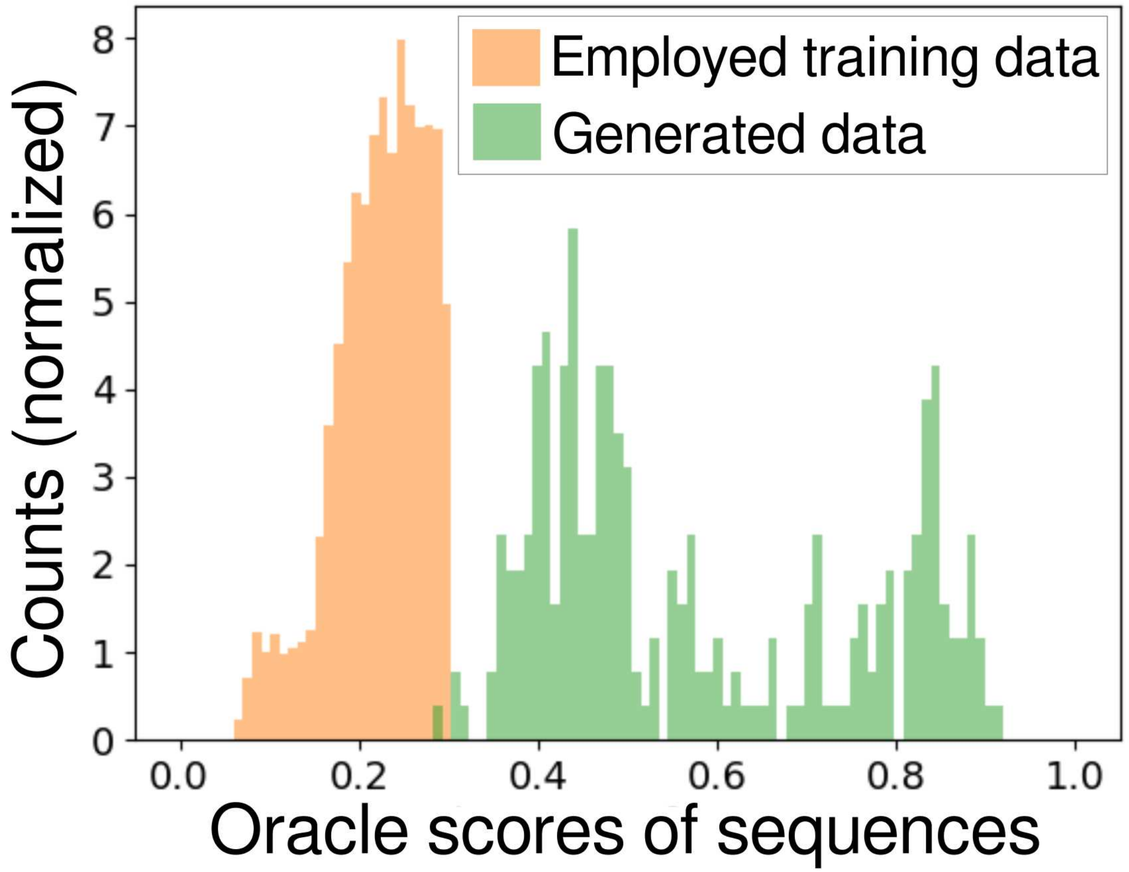
Training data

Extrapolation

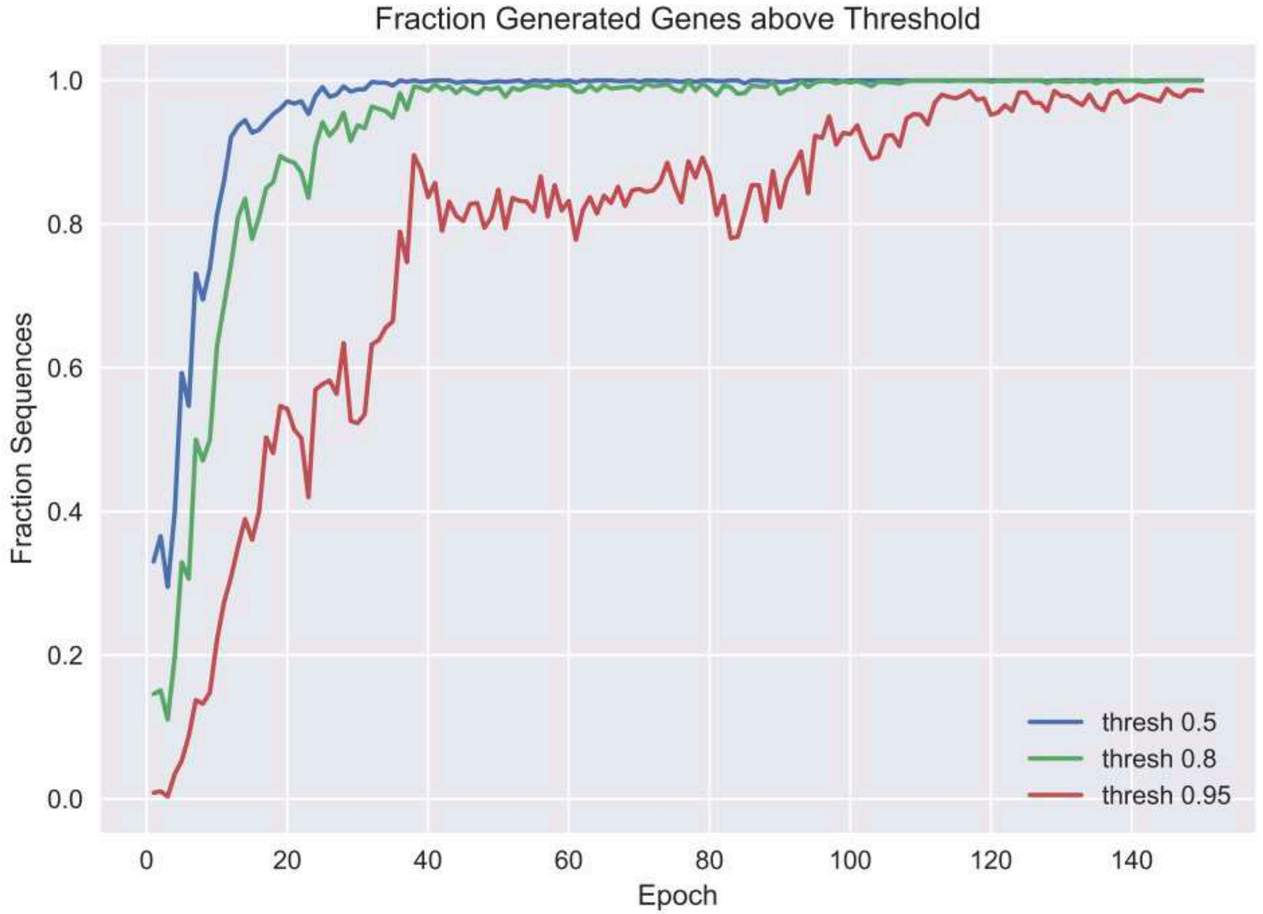




Designing DNA to optimize protein function

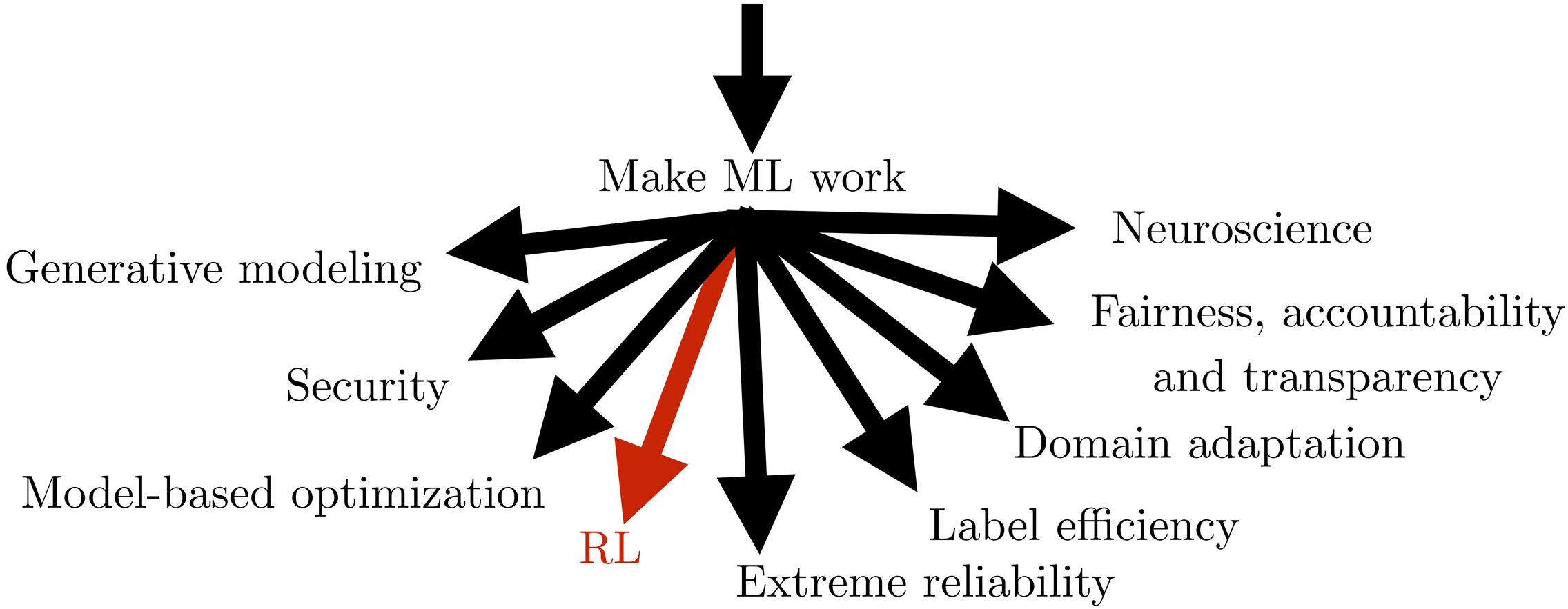


(Killoran et al, 2017)

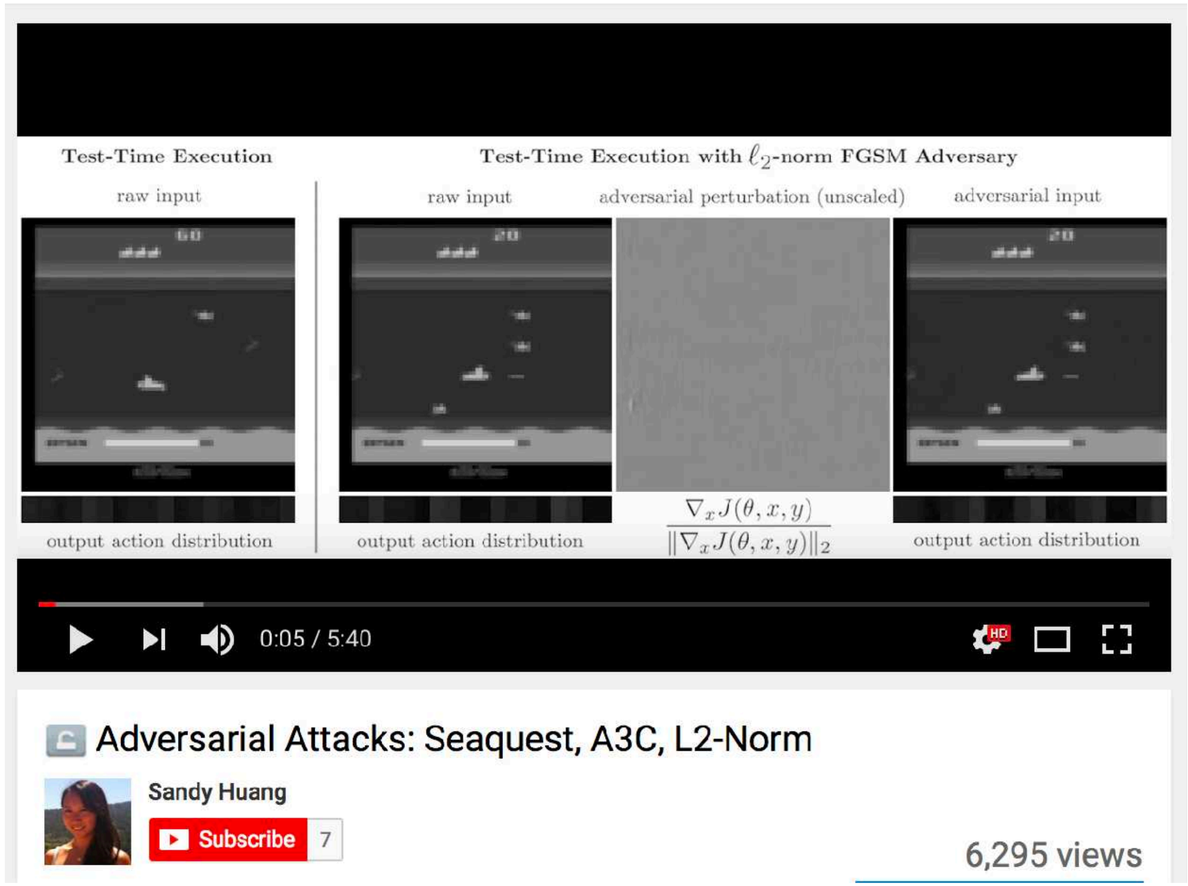


(Gupta and Zou, 2018)

A Cambrian Explosion of Machine Learning Research Topics



Adversarial Examples for RL

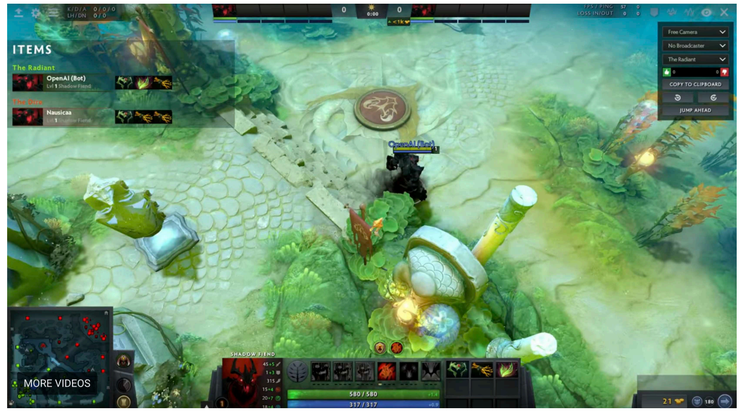


(Huang et al., 2017)

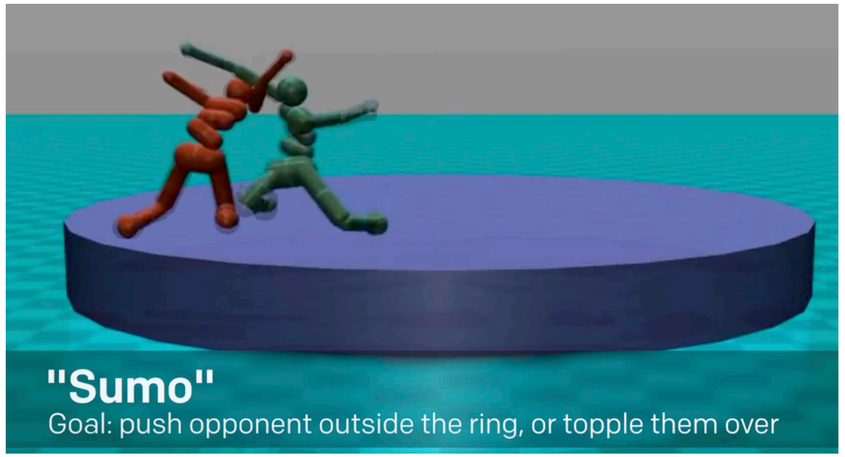
Self-Play

1959: Arthur Samuel's checkers agent

(Silver et al, 2017)



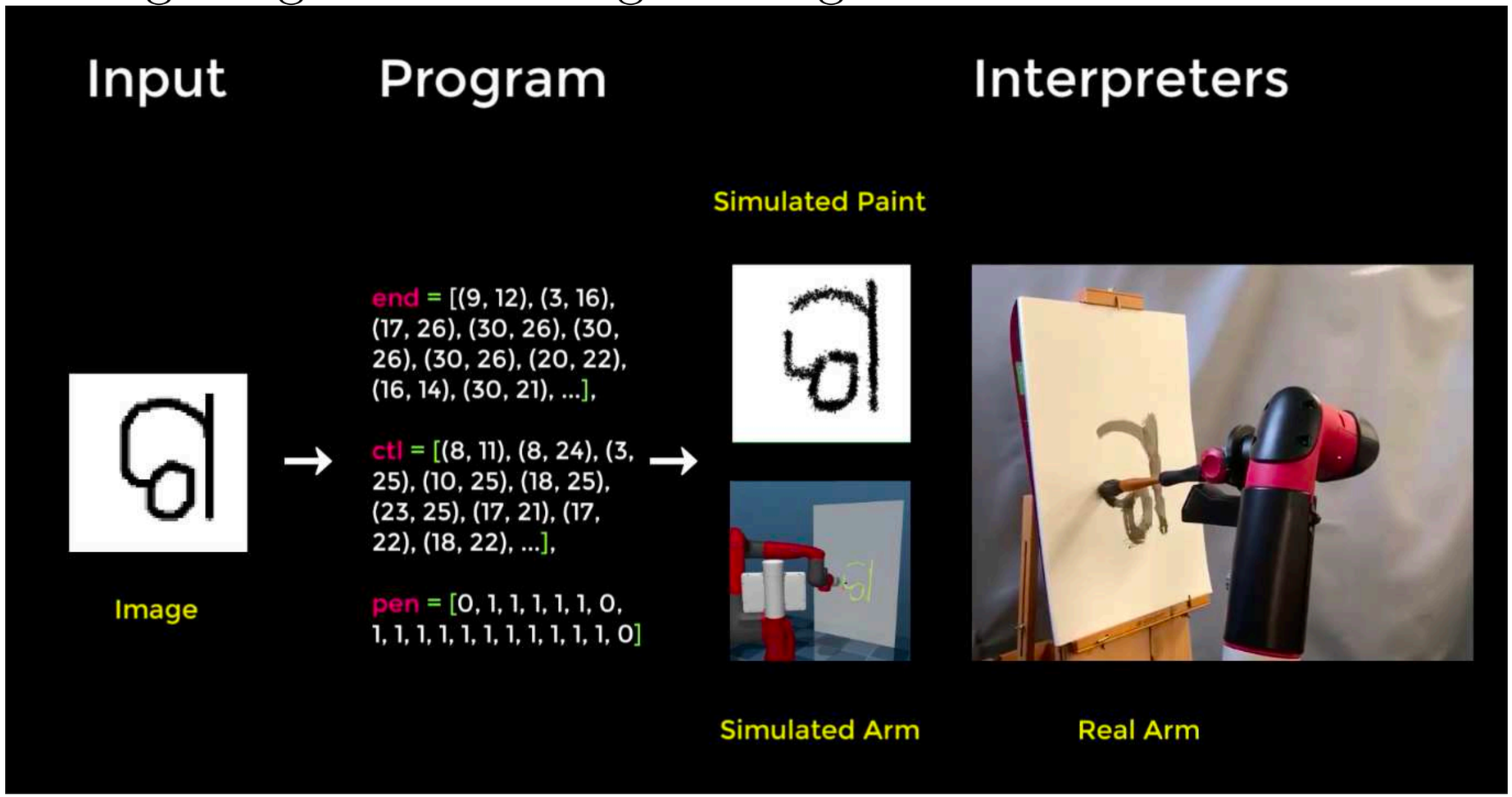
(OpenAI, 2017)



(Bansal et al, 2017)

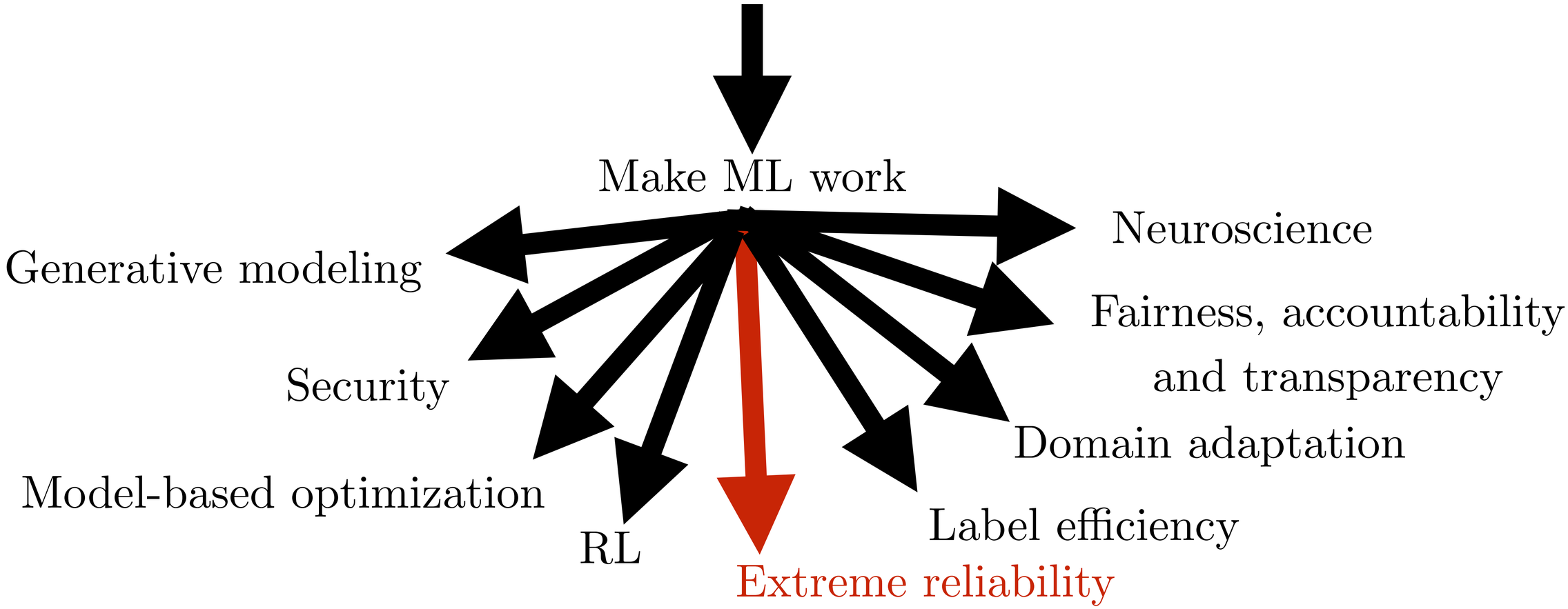
SPIRAL

Synthesizing Programs for Images Using Reinforced Adversarial Learning



(Ganin et al, 2018)

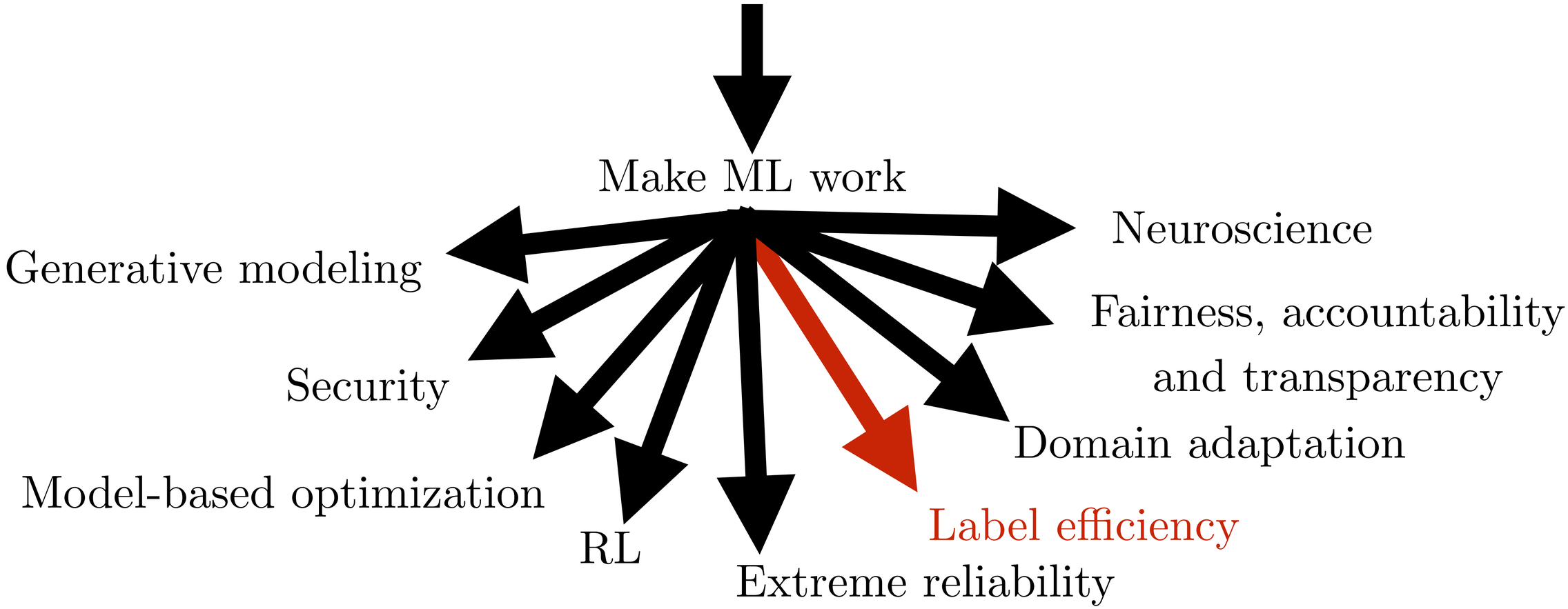
A Cambrian Explosion of Machine Learning Research Topics



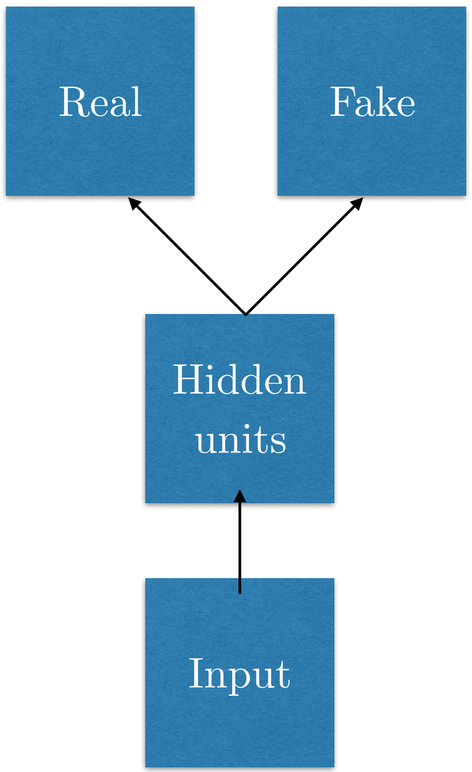
Extreme Reliability

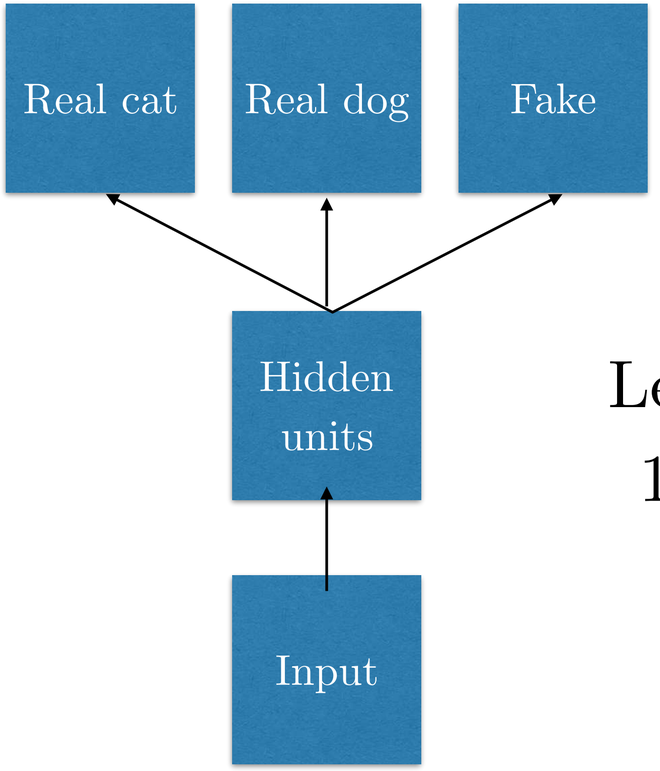
- We want extreme reliability for
  - Autonomous vehicles
  - Air traffic control
  - Surgery robots
  - Medical diagnosis, etc.
- Adversarial machine learning research techniques can help with this
  - Katz et al 2017: verification system, applied to air traffic control

A Cambrian Explosion of Machine Learning Research Topics



Supervised Discriminator for Semi-Supervised Learning



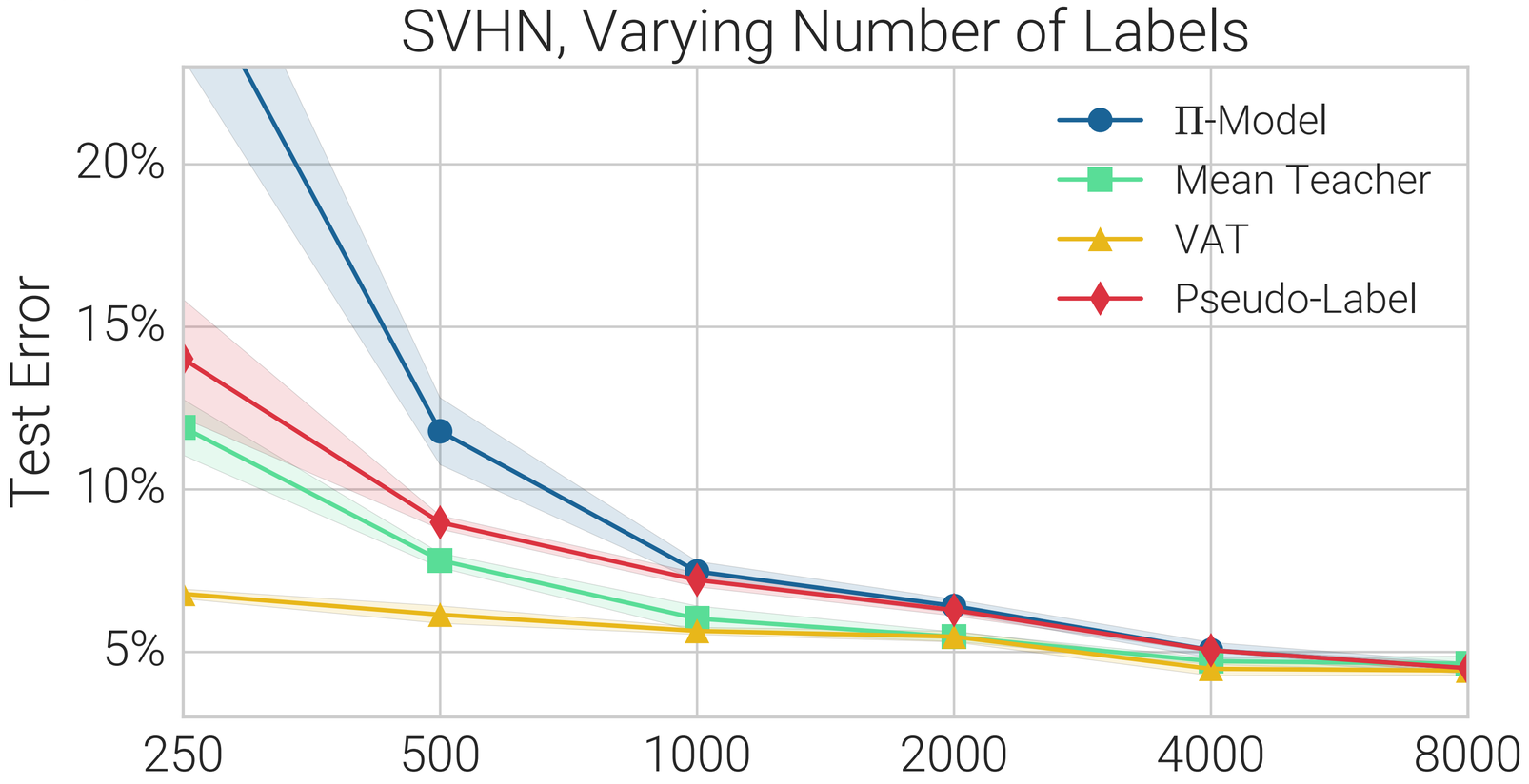


Learn to read with 100 labels rather than 60,000

(Odena 2016, Salimans et al 2016)

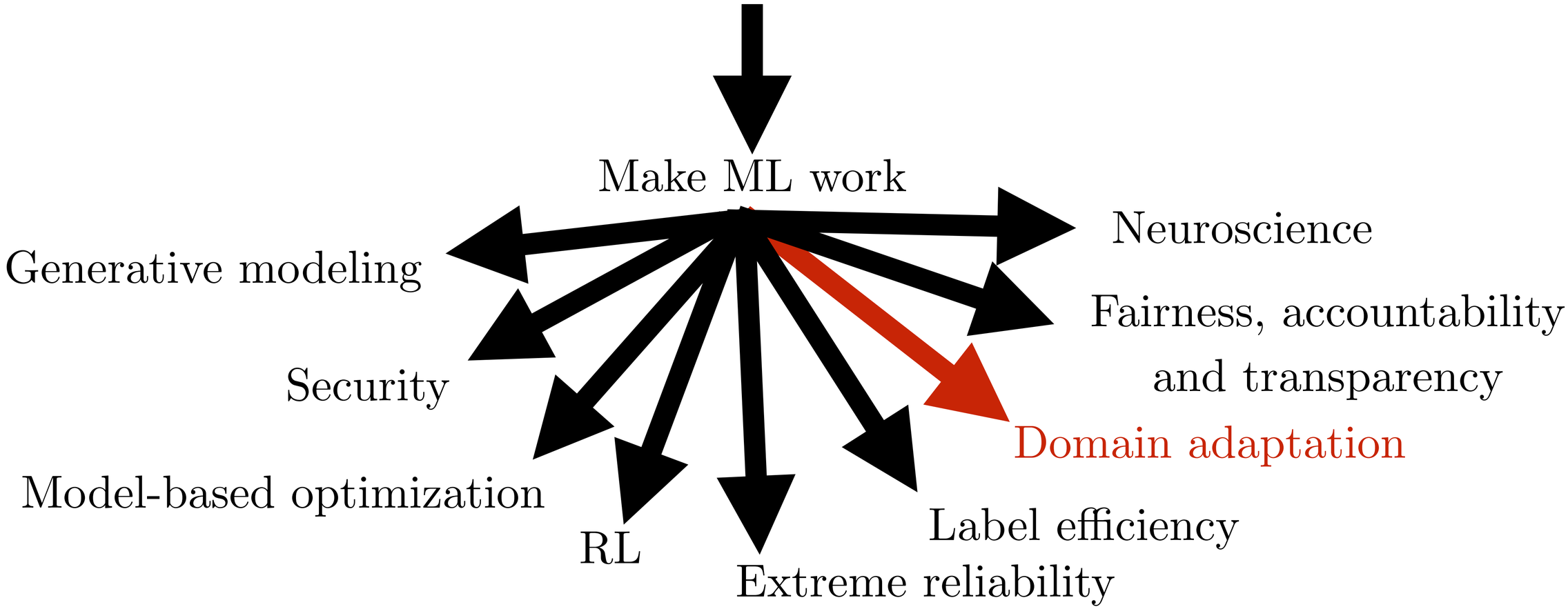
Virtual Adversarial Training

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

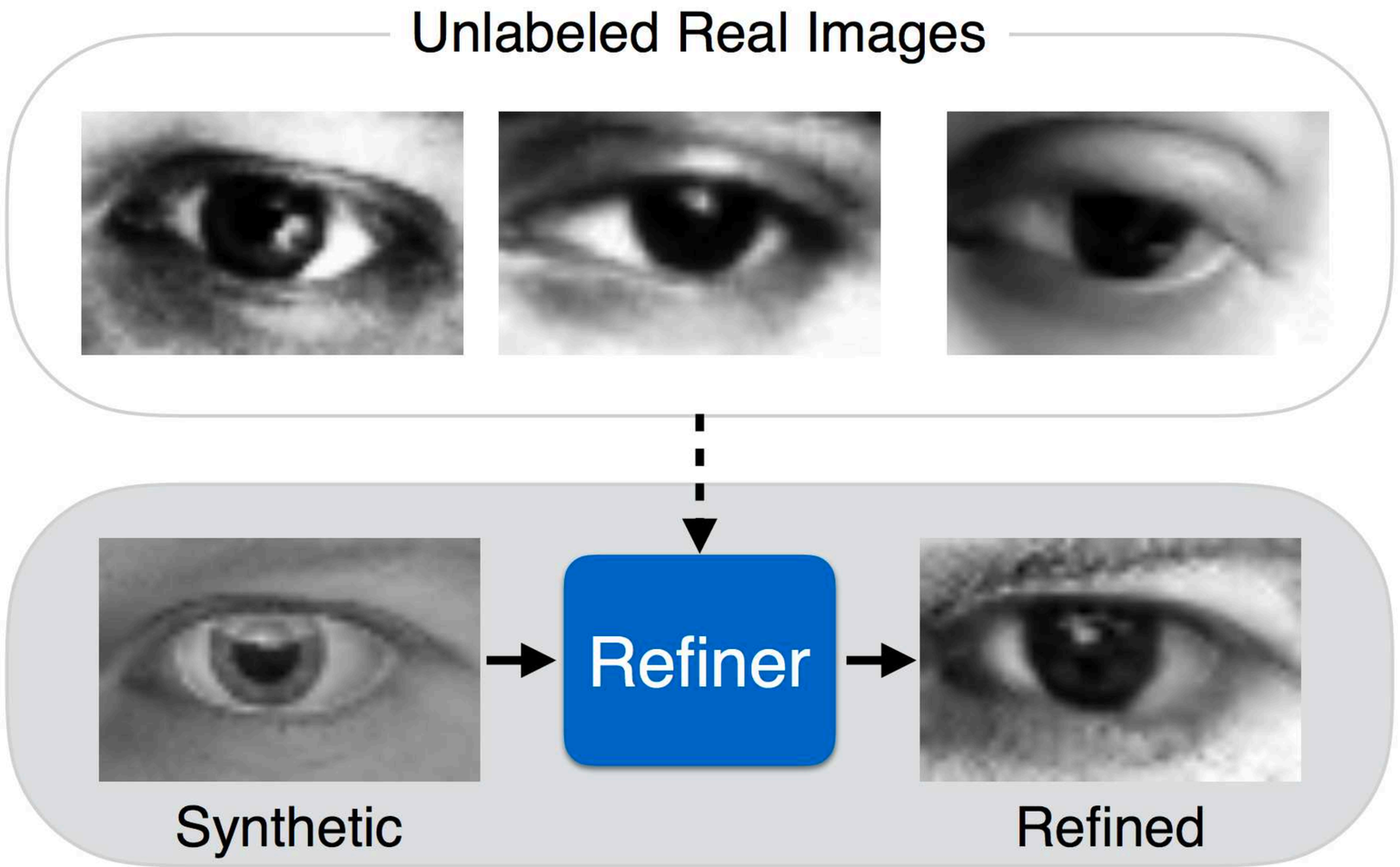


Domain Adaptation

- Domain Adversarial Networks (Ganin et al, 2015)

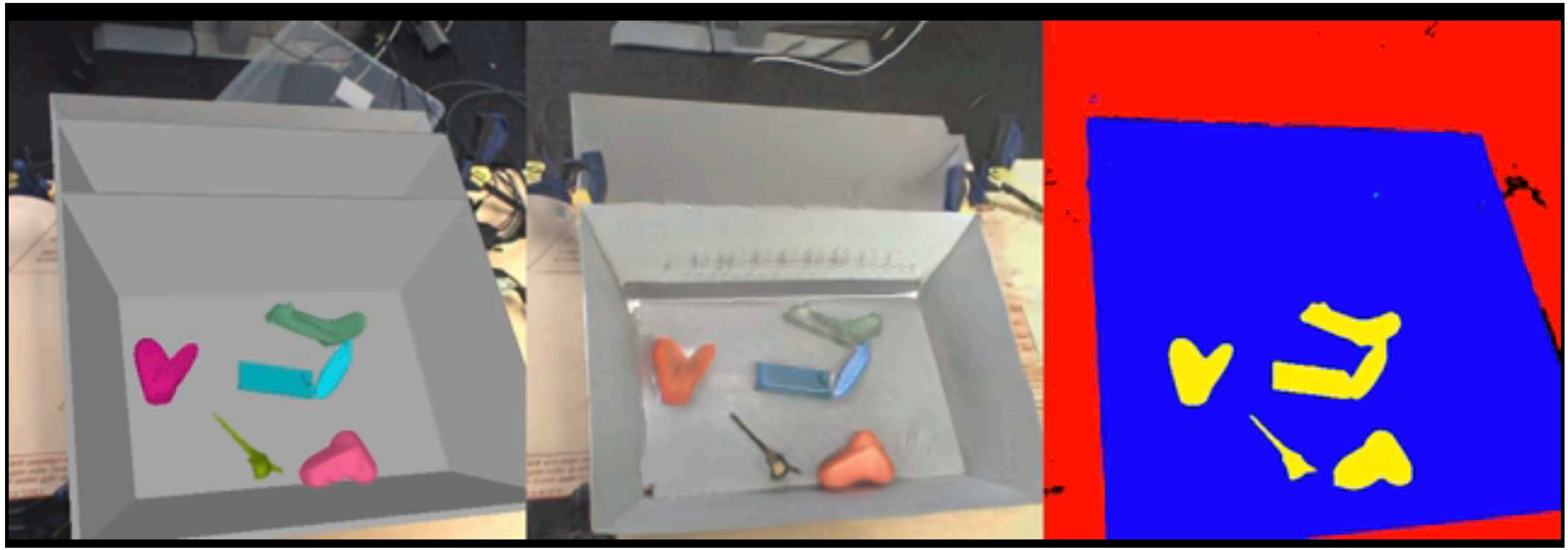
- Professor forcing (Lamb et al, 2016): Domain-Adversarial learning in RNN hidden state

GANs for simulated training data



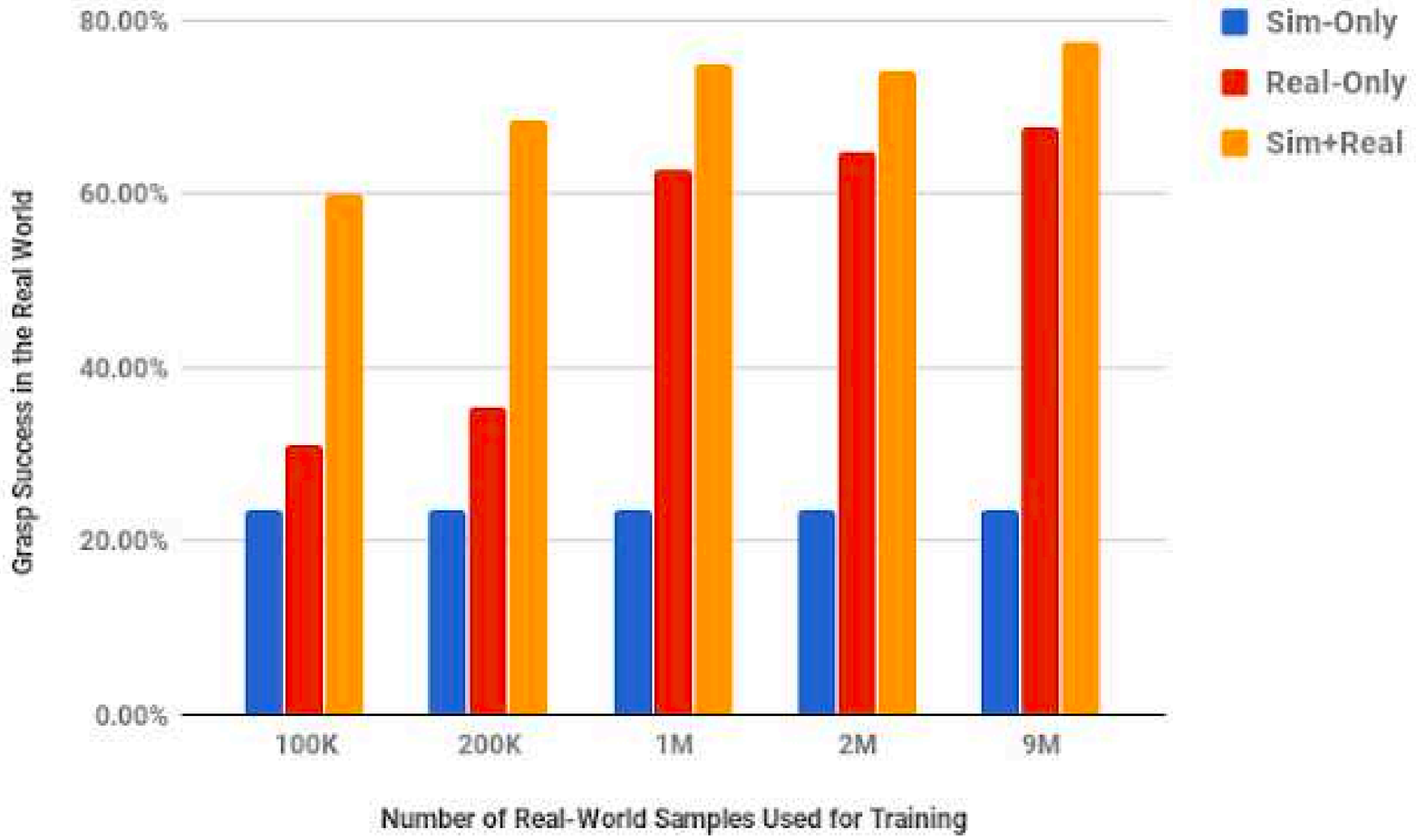
(Shrivastava et al., 2016)

GraspGAN



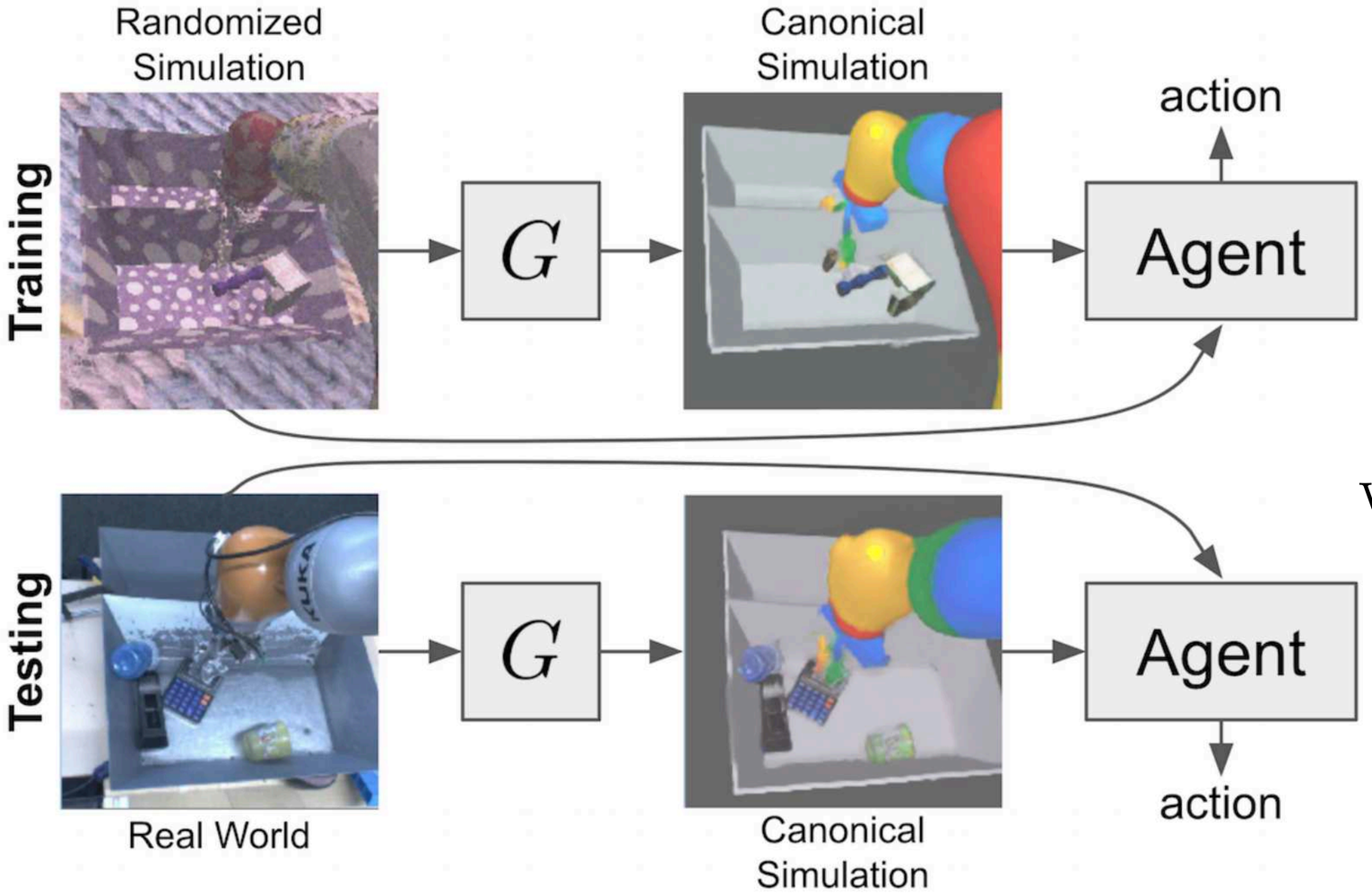
(Bousmalis et al, 2017)

GraspGAN



(Bousmalis et al, 2017)

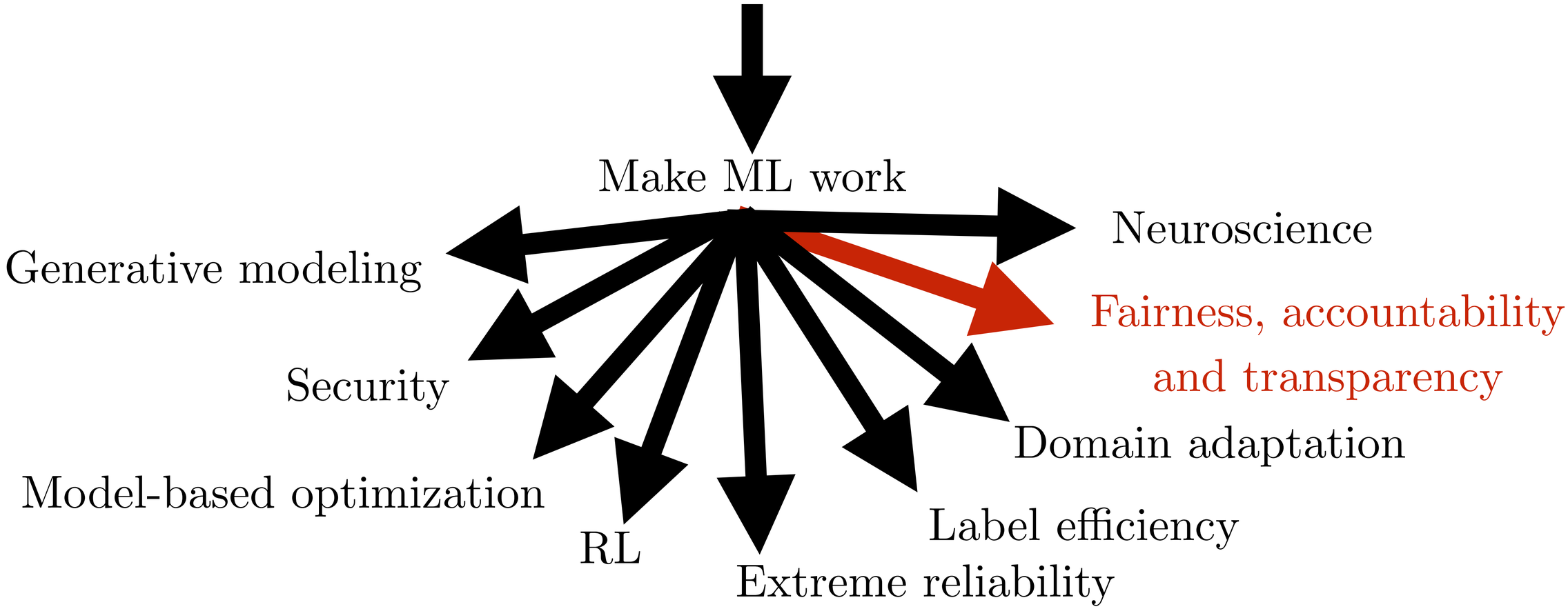
Sim-to-real via sim-to-sim



Learn to grasp without real data!

(James et al, 2018)

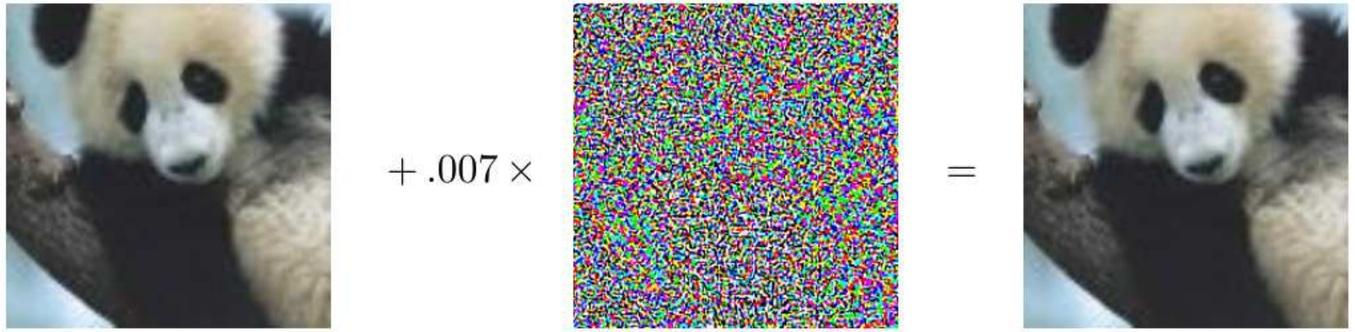
A Cambrian Explosion of Machine Learning Research Topics



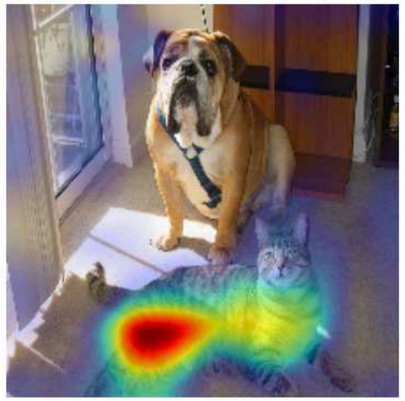
Adversarially Learned Fair Representations

- 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'

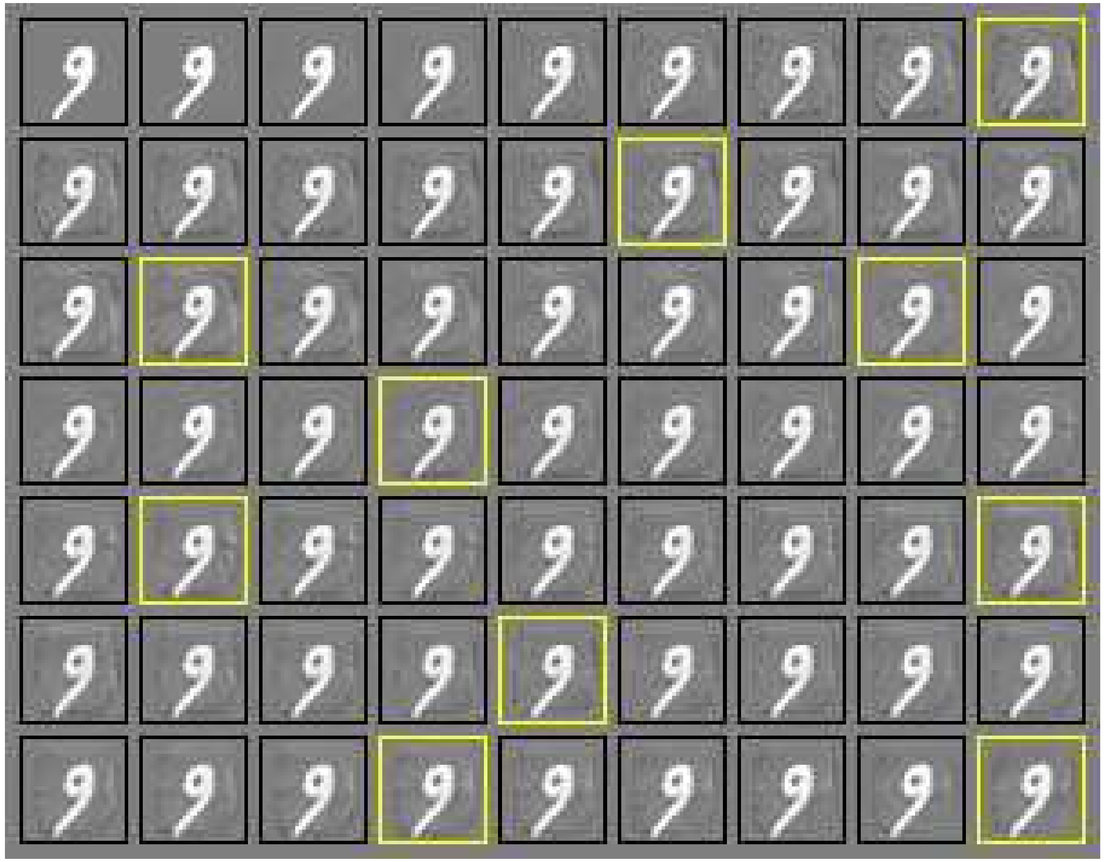
(i) Grad-CAM 'Dog'

(Selvaraju et al, 2016)

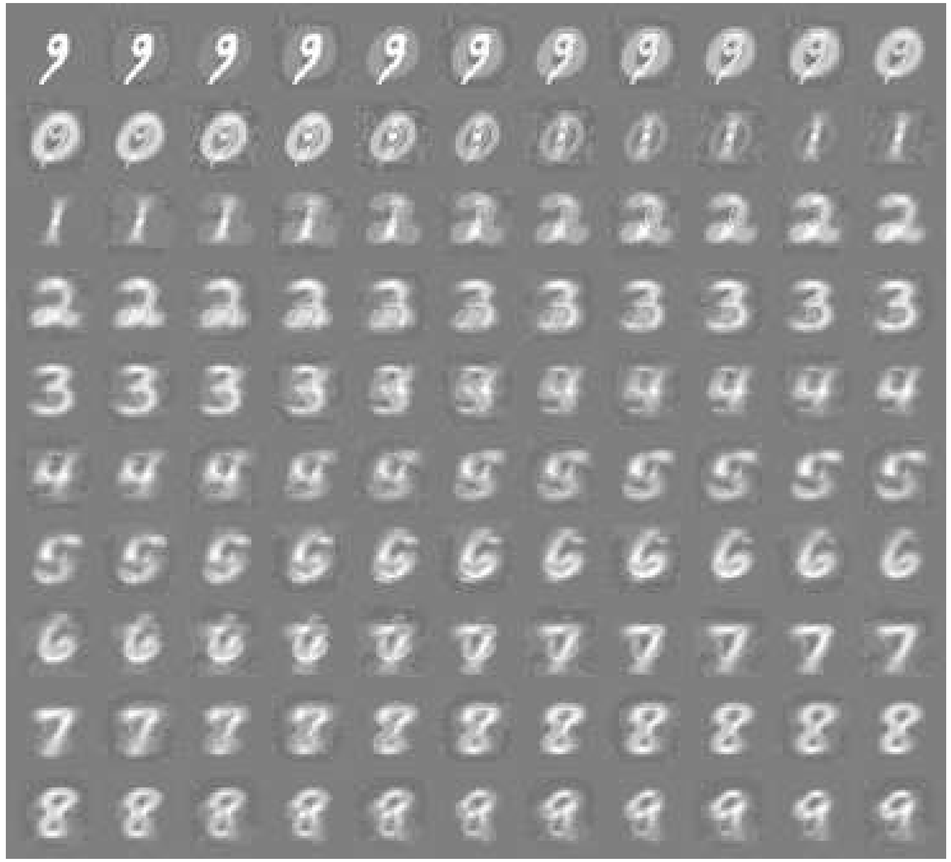
Interpretability literature: our analysis tools show that deep nets work about how you would expect them to.

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

Robust models are more interpretable



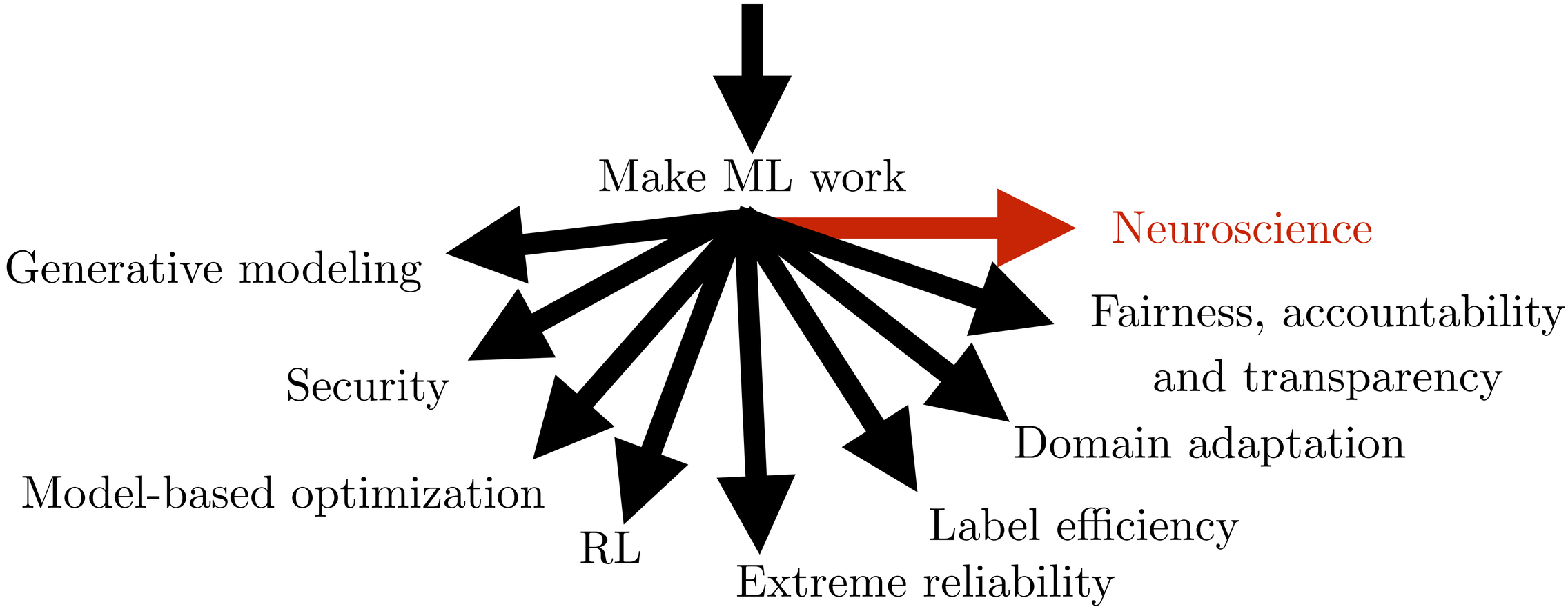
Relatively vulnerable model



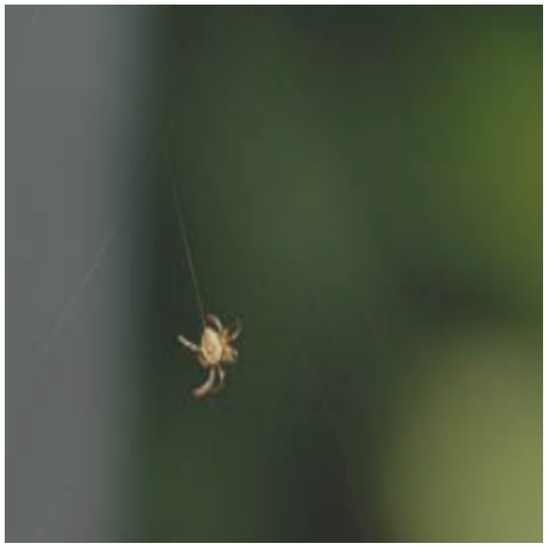
Relatively robust model

(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