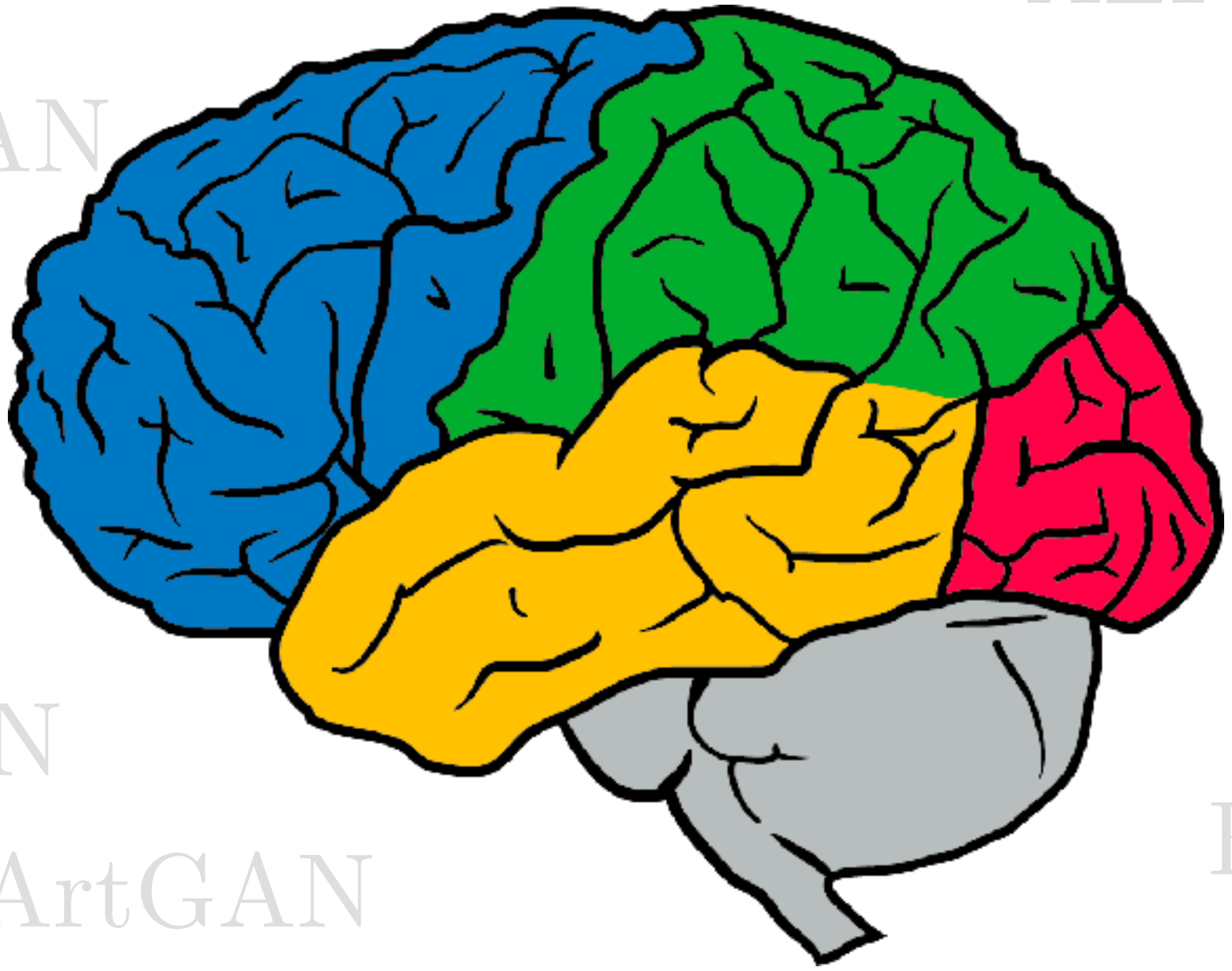
Adversarial Machine Learning

Ian Goodfellow, Staff Research Scientist, Google Brain

ACM Webinar

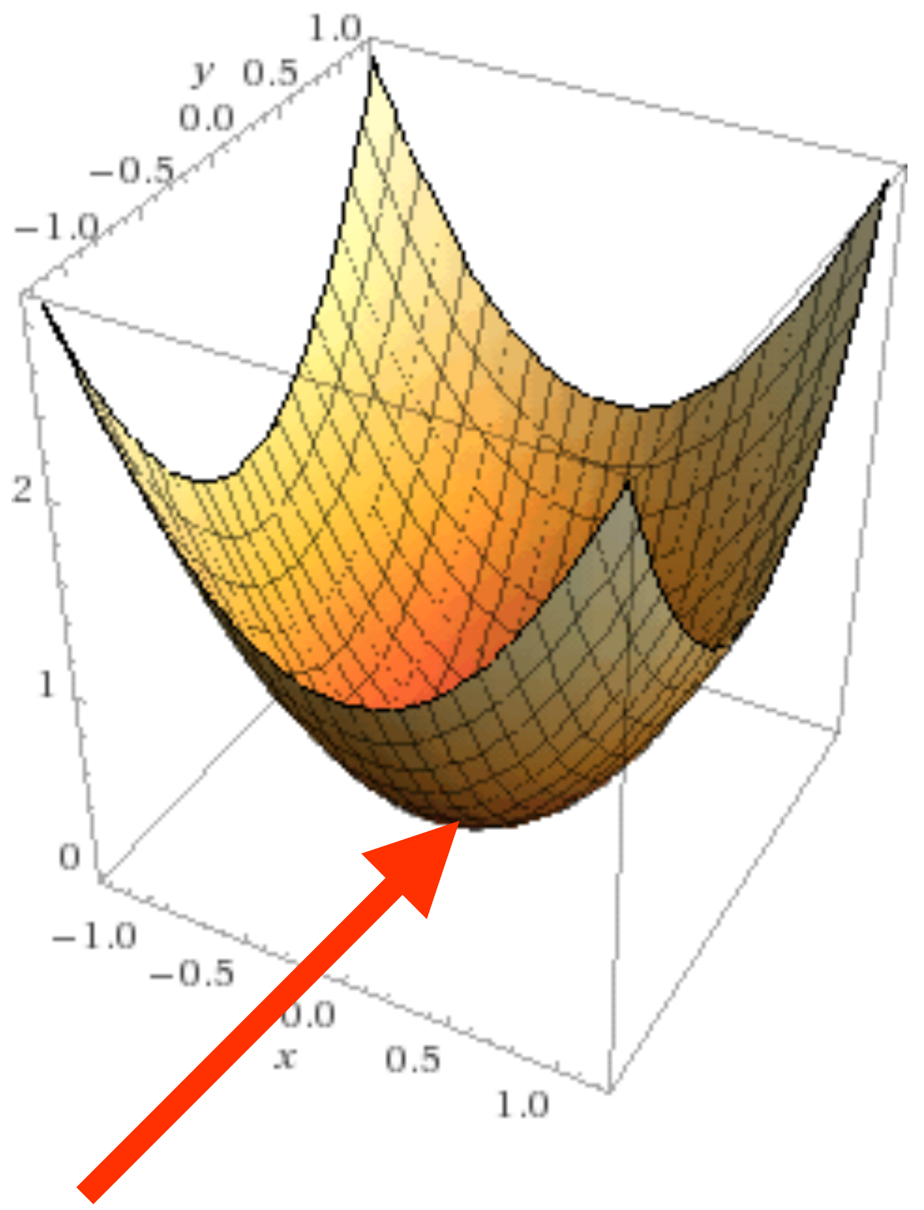
2018-07-24



MedGAN Progressive GAN CoGAN LR-GAN CGAN IcGAN
BIM LS-GAN AffGAN LAPGAN DiscoGAN MPM-GAN AdaGAN
LSGAN InfoGAN ATN FGSM iGAN IAN
McGAN MIX+GAN
FF-GAN BPDA
MGAN
C-VAE-GAN C-RNN-GAN SAGAN ALP DR-GAN BS-GAN
MAGAN 3D-GAN CCGAN AC-GAN DCGAN
Adversarial Training BiGAN
GAWWN CycleGAN
Bayesian GAN Gradient Masking
AnoGAN
EBGAN SN-GAN DTN
Context-RNN-GAN MAD-GAN
ALI
PGD f-GAN ArtGAN BEGAN MalGAN AL-CGAN

Adversarial Machine Learning

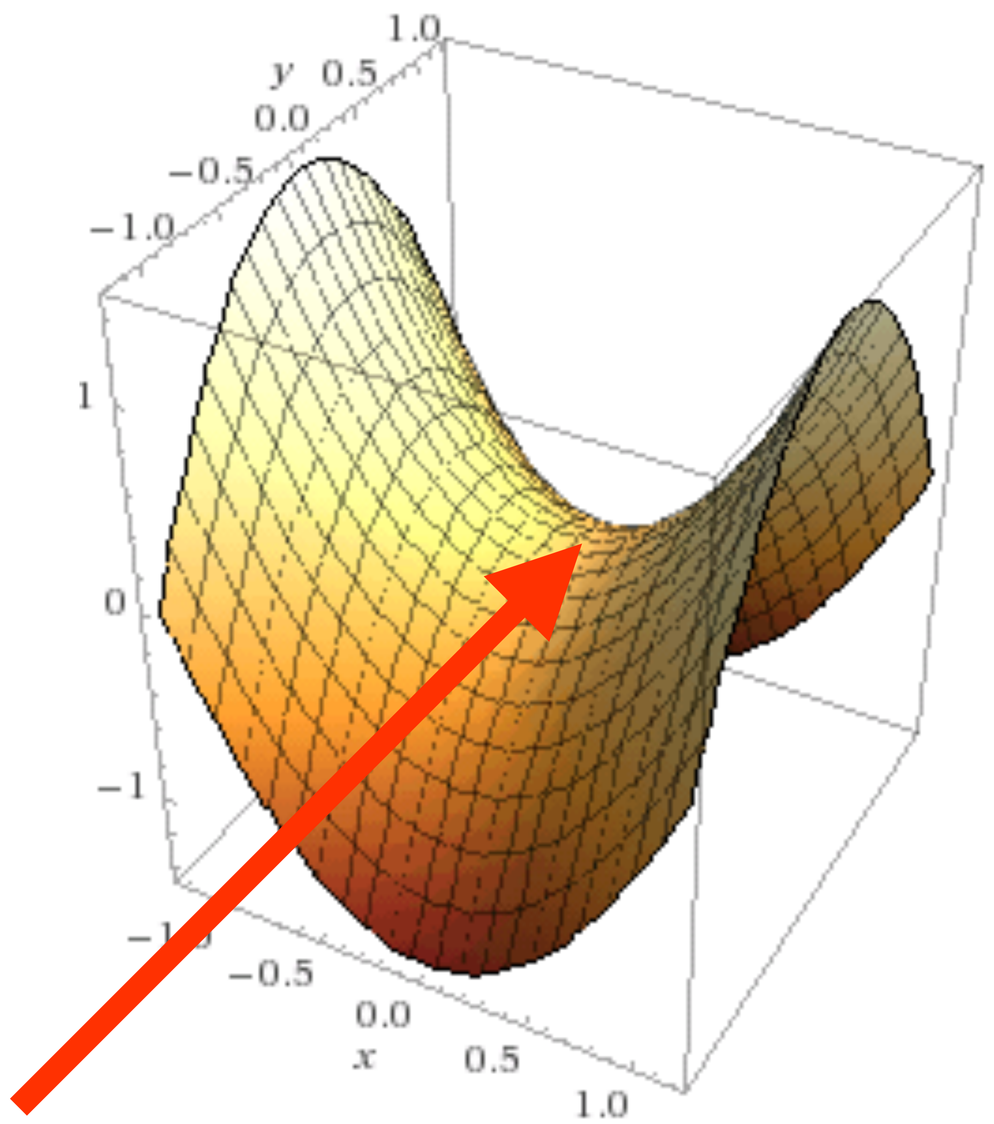
Traditional ML:
optimization



Minimum

One player,
one cost

Adversarial ML:
game theory

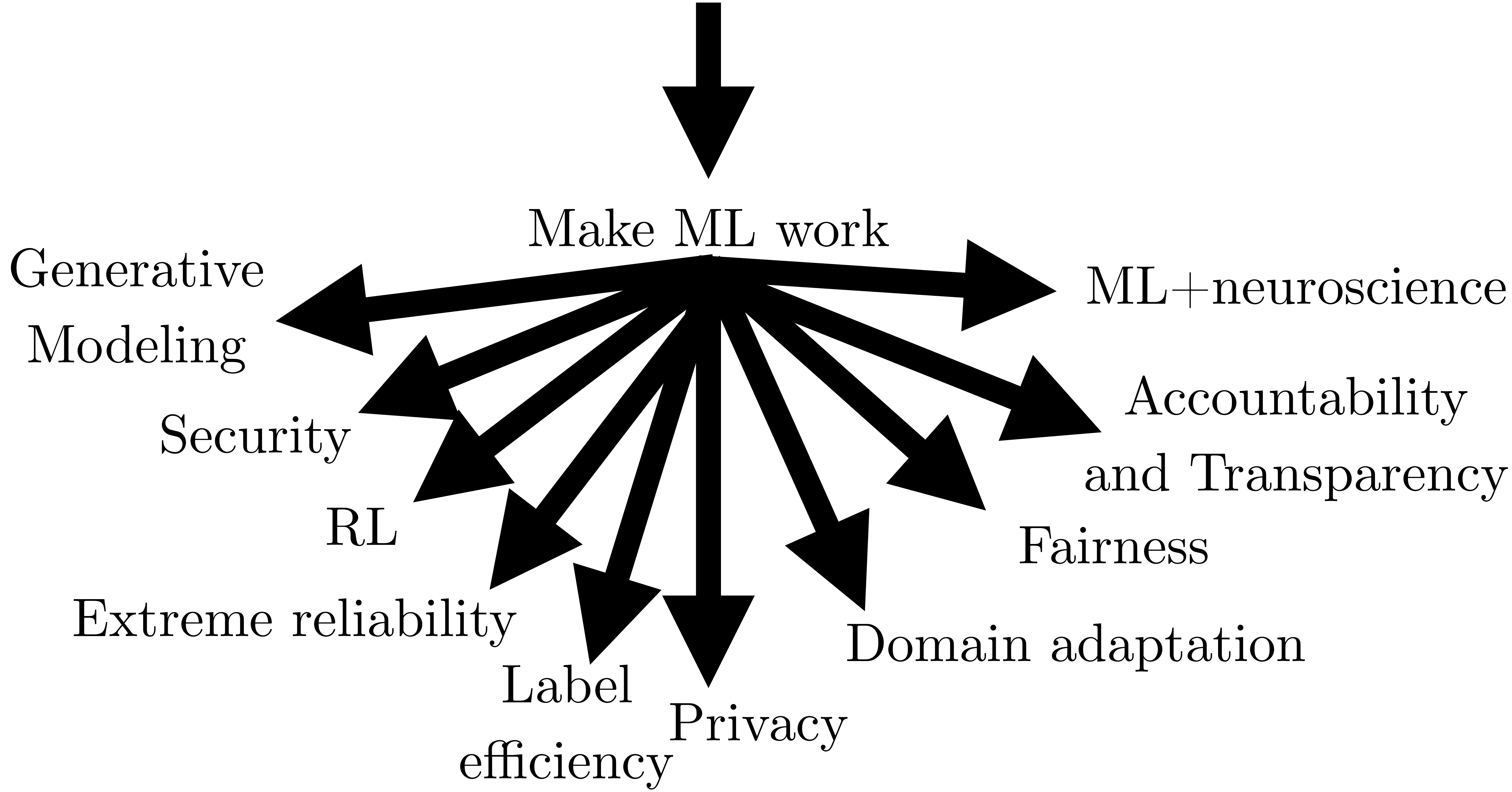


Equilibrium

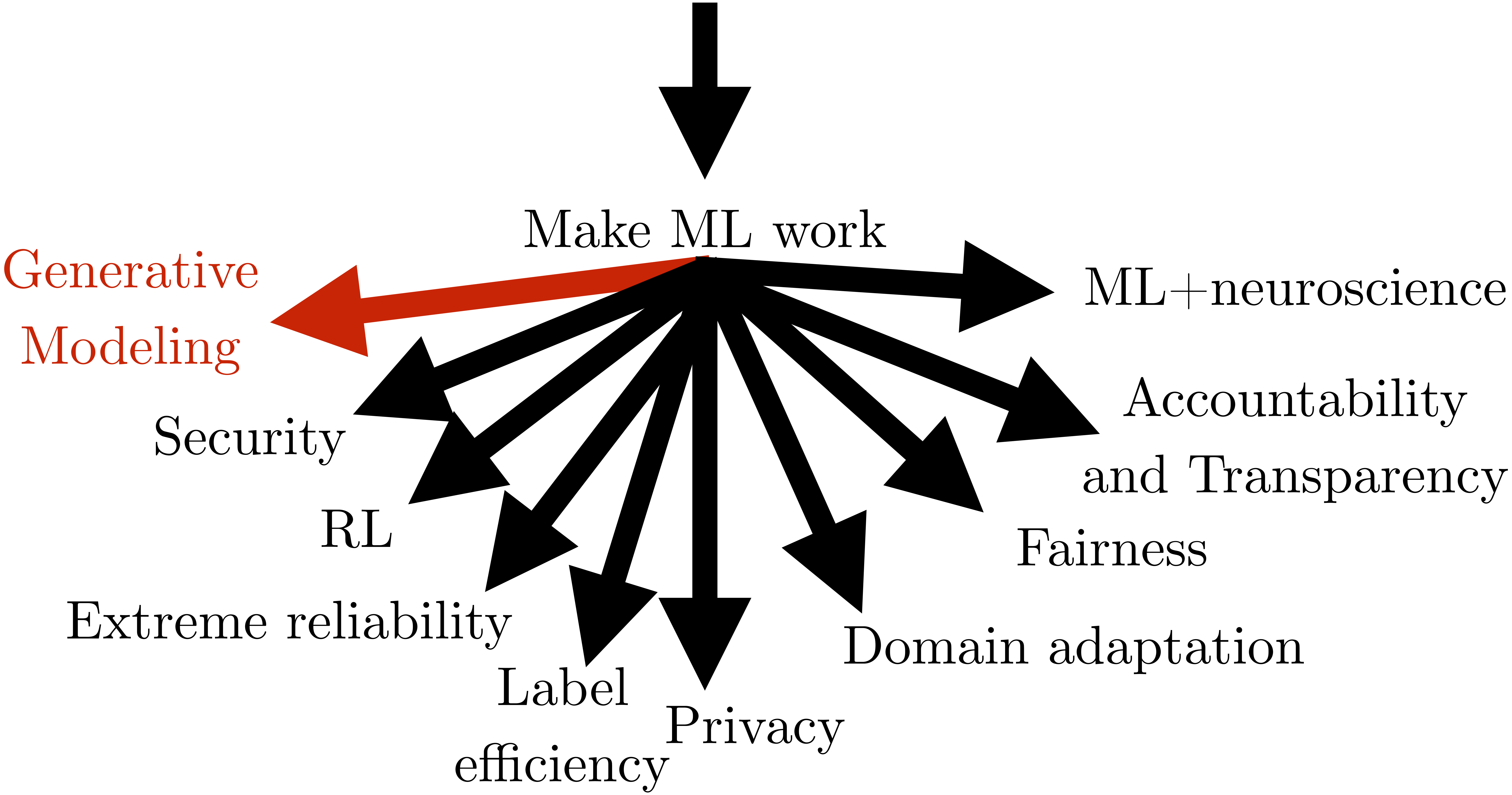
More than one player,
more than one cost

(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

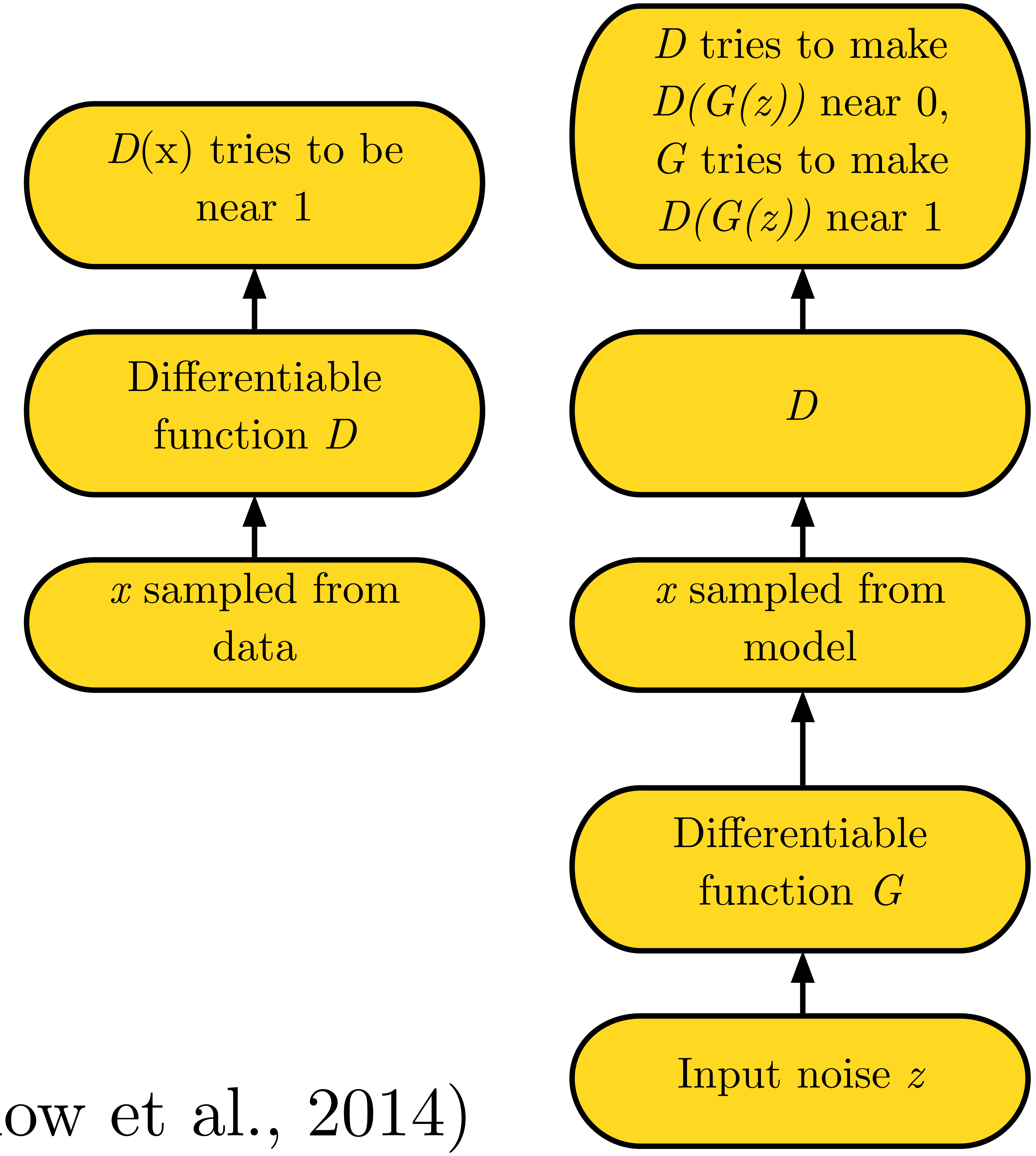
Generative Modeling: Sample Generation

Training Data
(CelebA)

Sample Generator
(Karras et al, 2017)

(Goodfellow 2018)

Adversarial Nets Framework



(Goodfellow et al., 2014)

(Goodfellow 2018)

3.5 Years of Progress on Faces

2014 2015 2016 2017

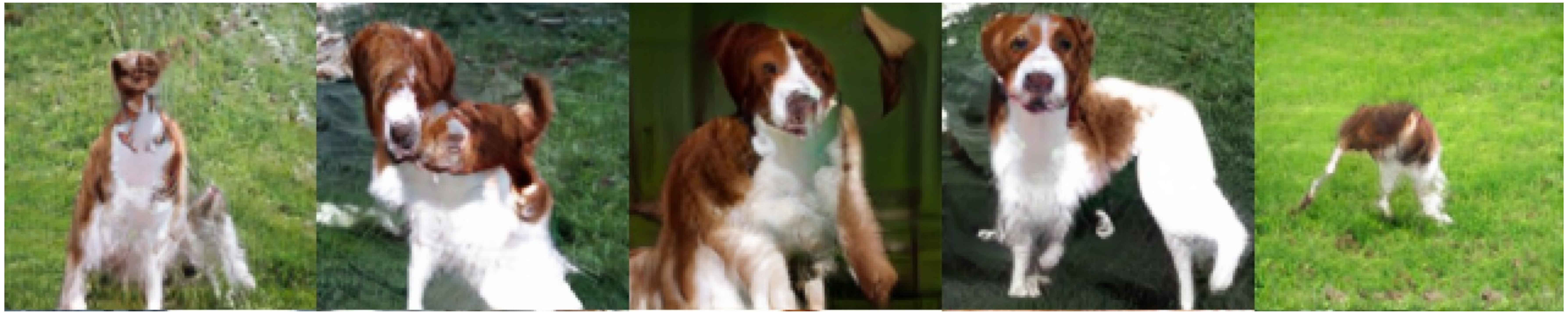
(Brundage et al, 2018)

(Goodfellow 2018)

<2 Years of Progress on ImageNet

Odena et al
2016

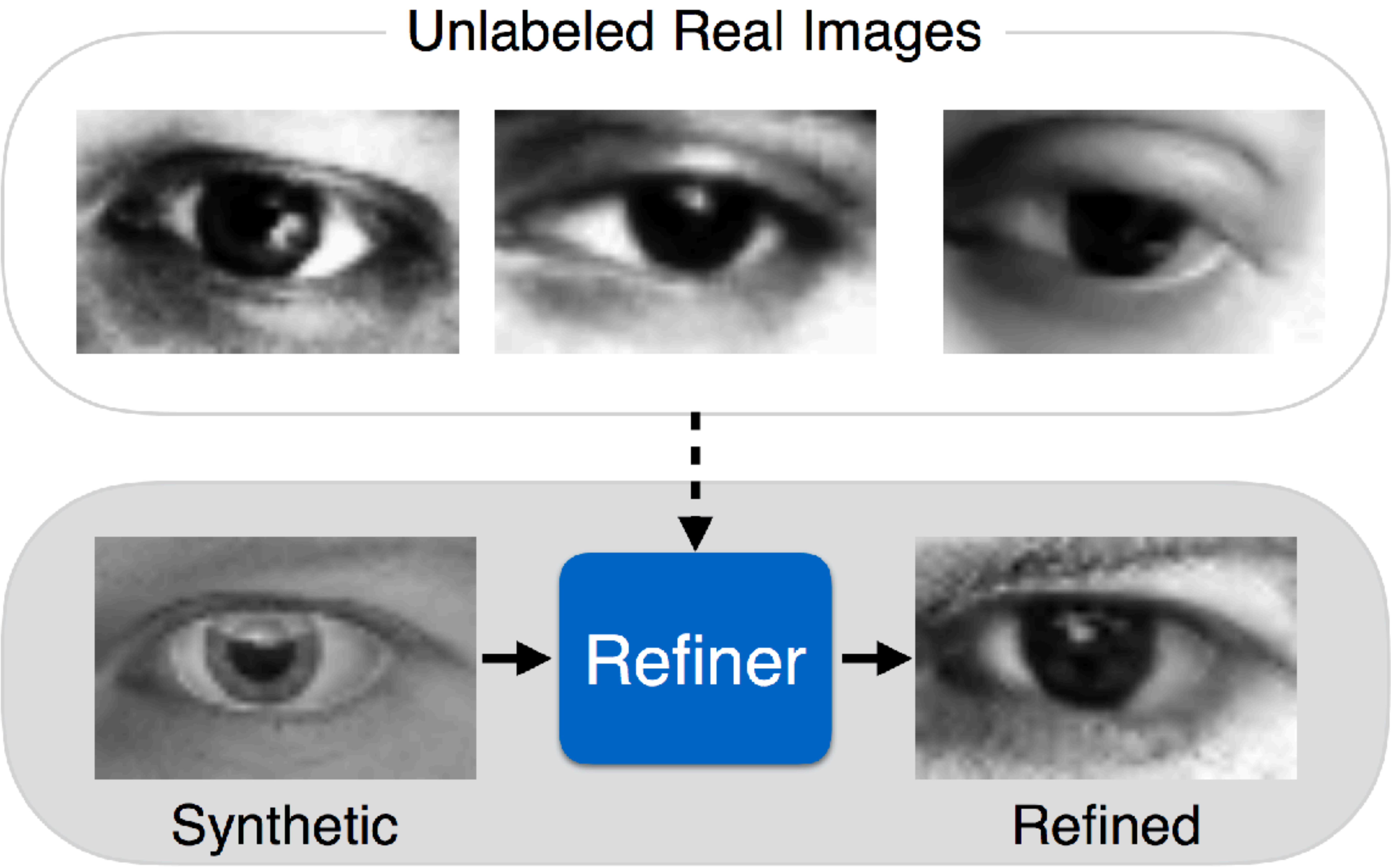
Miyato et al
2017



Zhang et al
2018

(Goodfellow 2018)

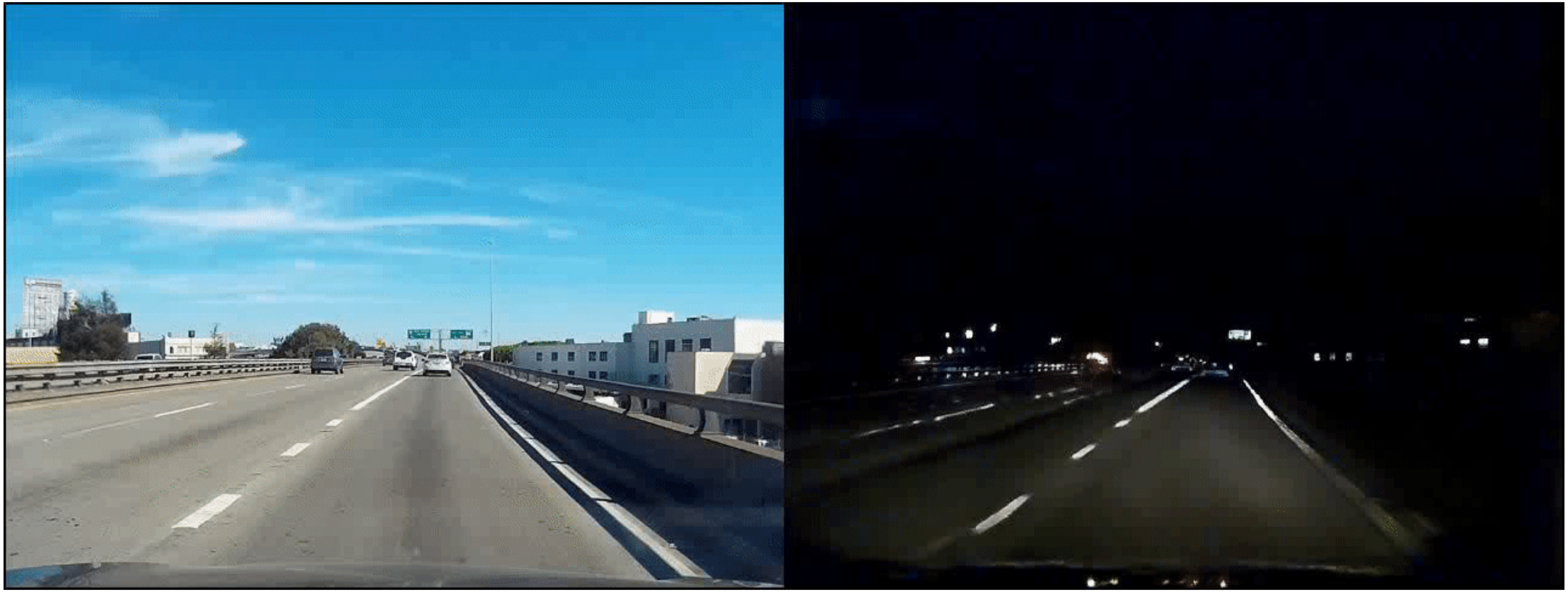
GANs for simulated training data



(Shrivastava et al., 2016)

Unsupervised Image-to-Image Translation

Day to night

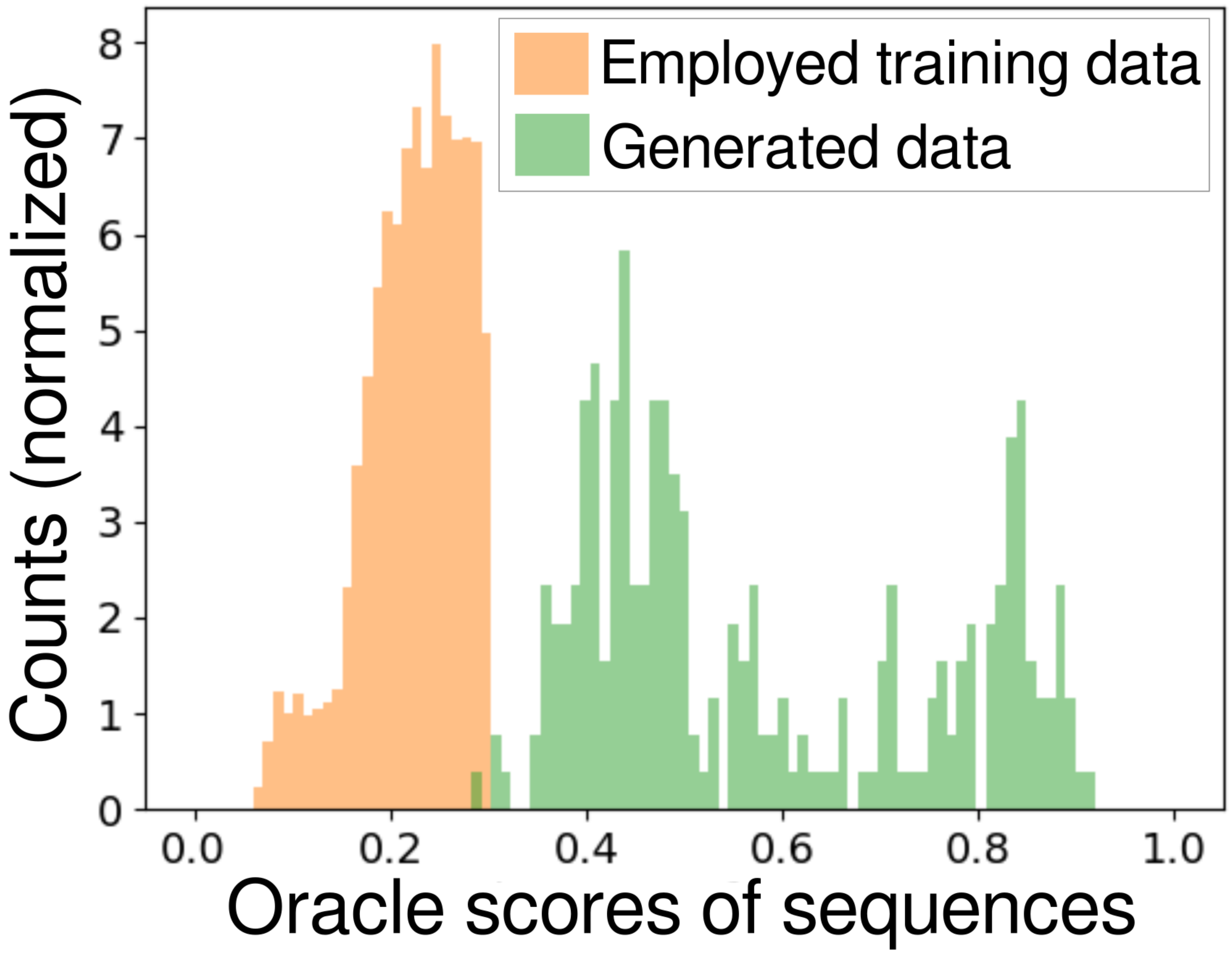


(Liu et al., 2017)

CycleGAN

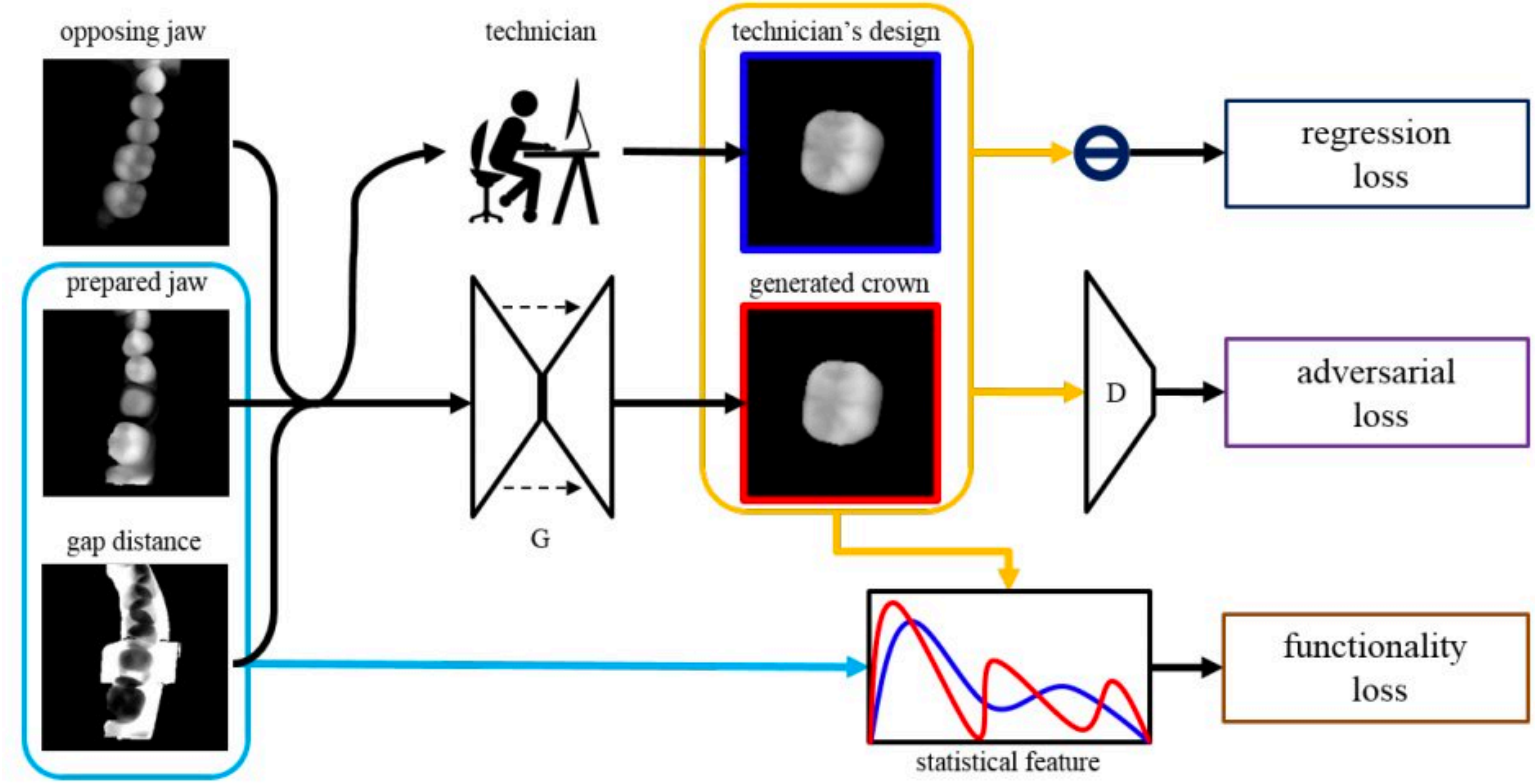
(Zhu et al., 2017)

Designing DNA to optimize protein binding



(Killoran et al, 2017)

Personalized GANufacturing



(Hwang et al 2018)

Self-Attention GAN

State of the art FID on ImageNet: 1000 categories, 128x128 pixels

Goldfish

Redshank

Broccoli

Tiger Cat

Geyser

Indigo Bunting

Stone Wall

Saint Bernard

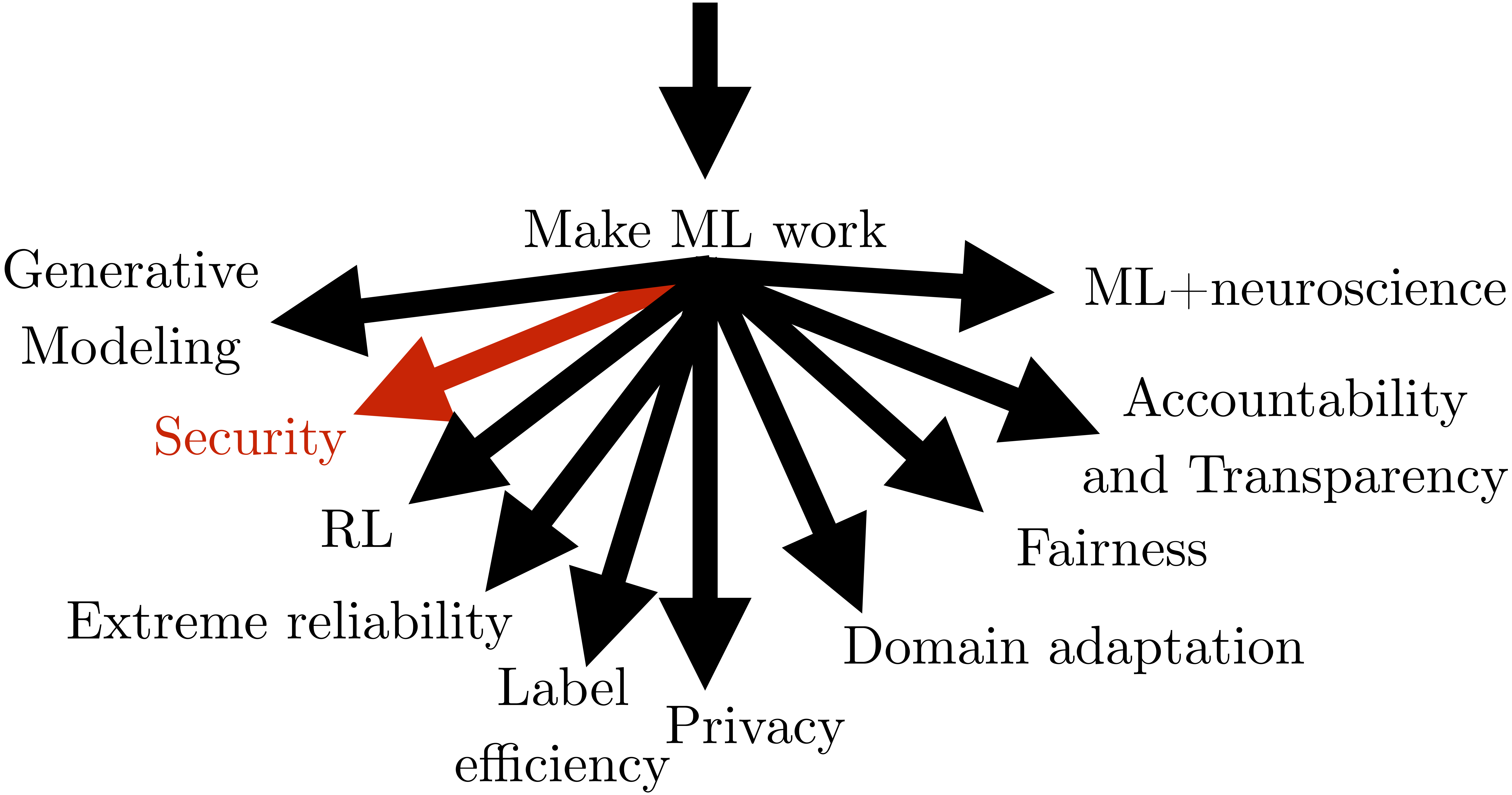
(Zhang et al., 2018)

Self-Attention

Use layers from
Wang et al 2018

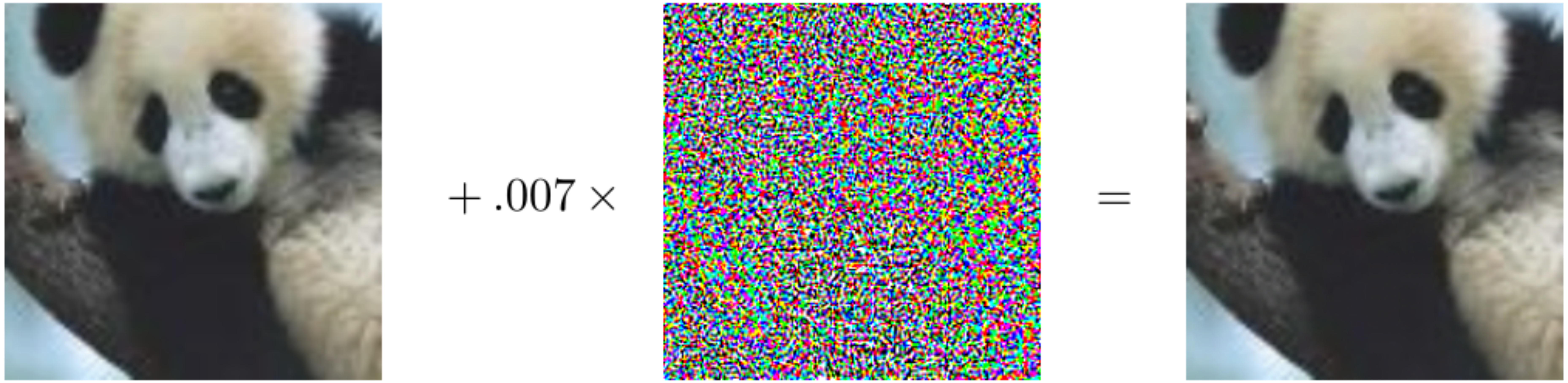
(Goodfellow 2018)

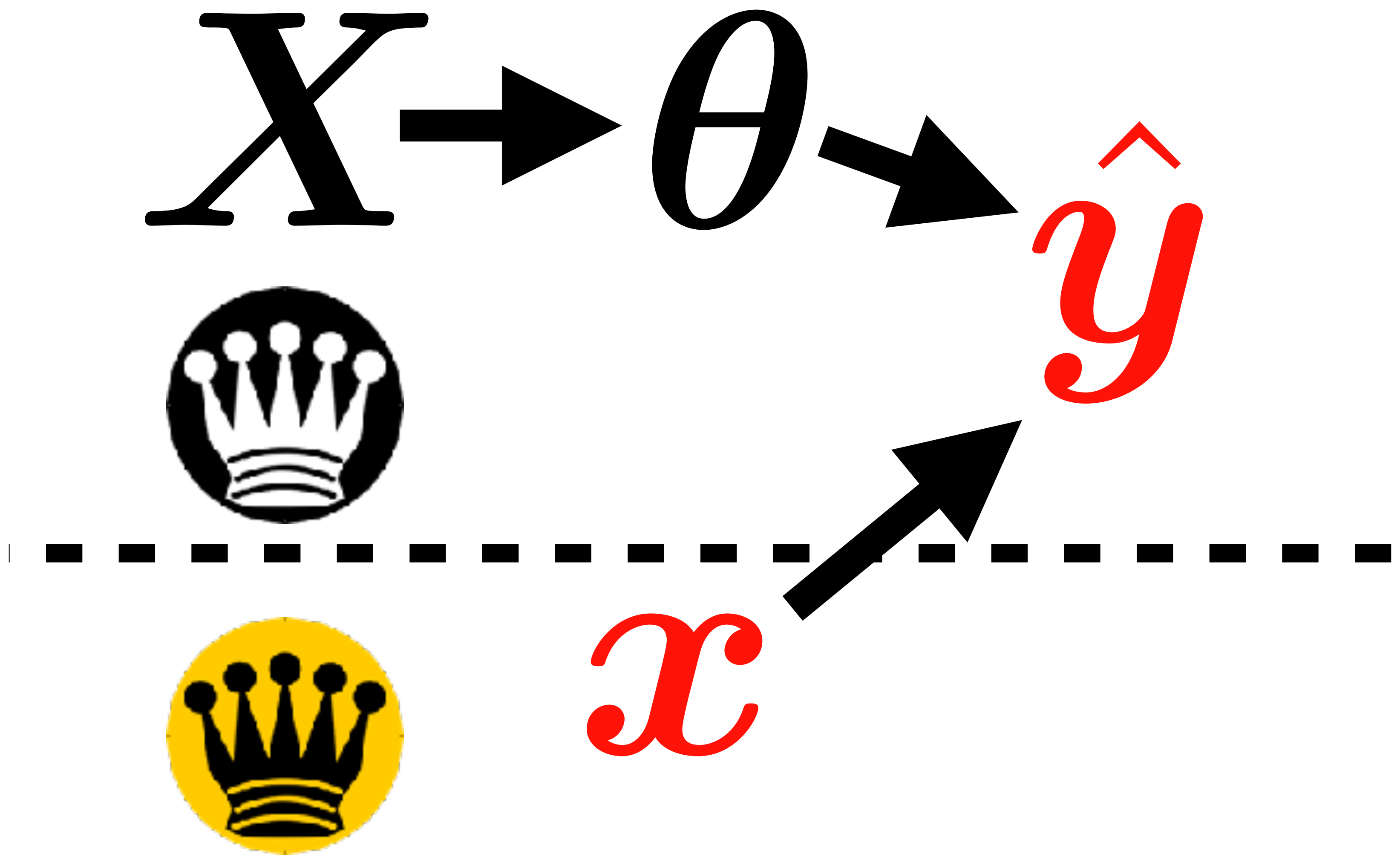
A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

Adversarial Examples

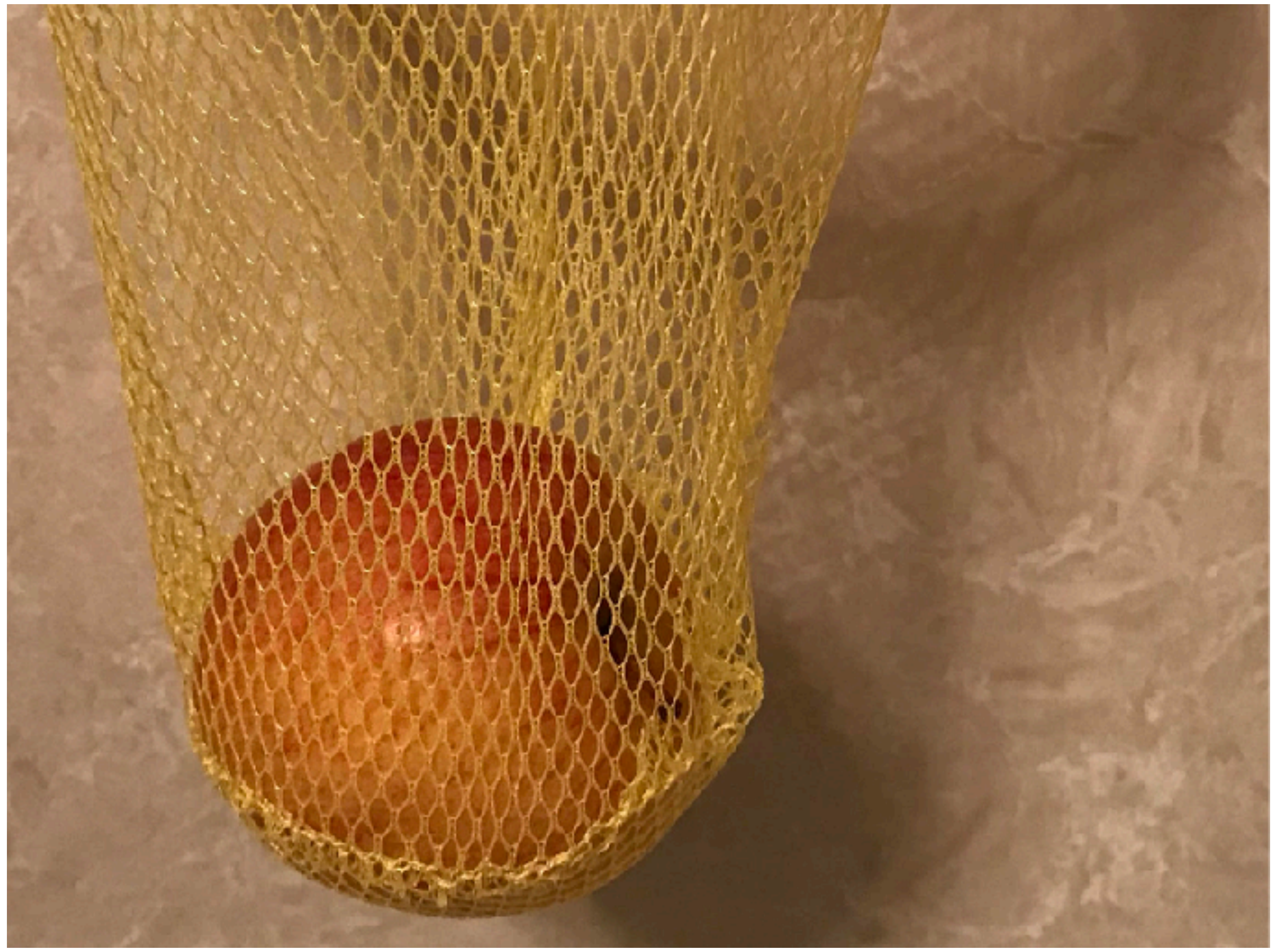




(Goodfellow 2018)

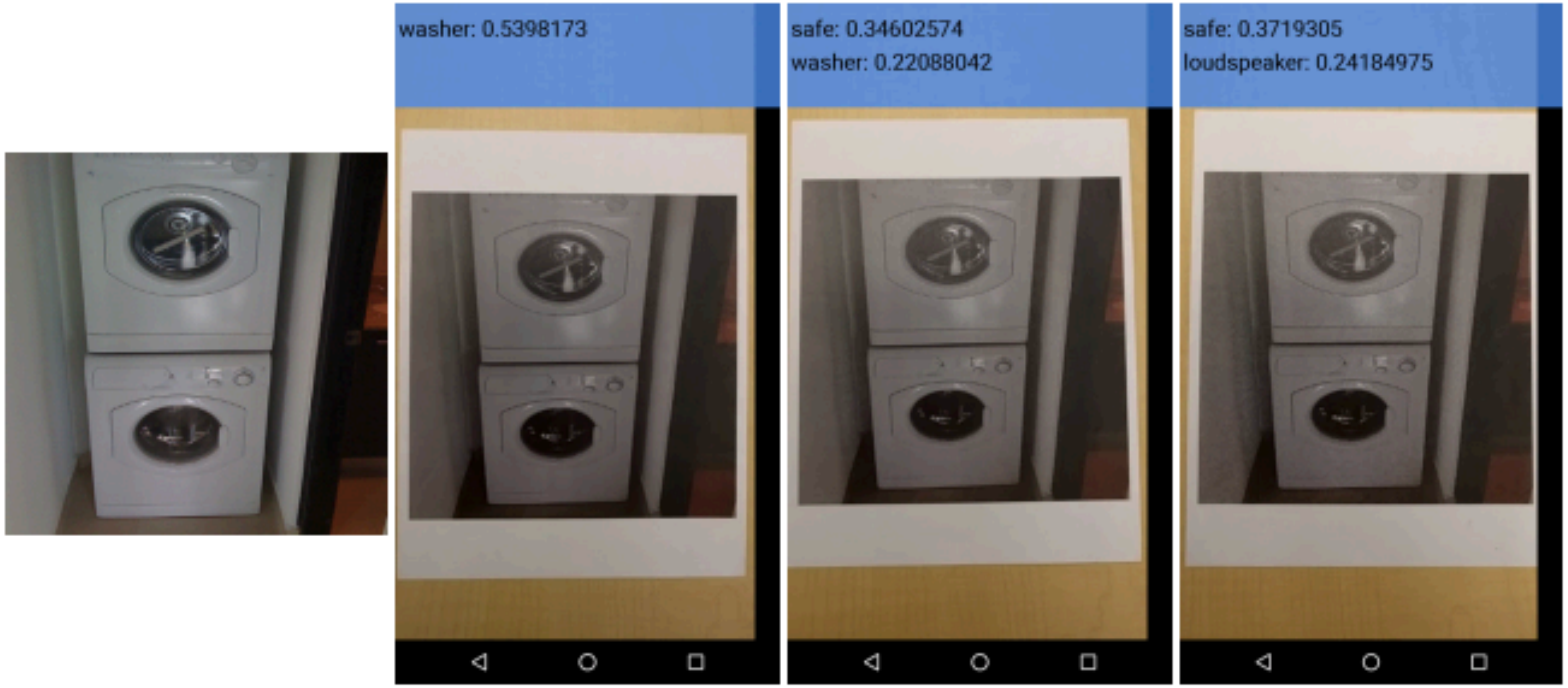
Also Adversarial Examples

(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,

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

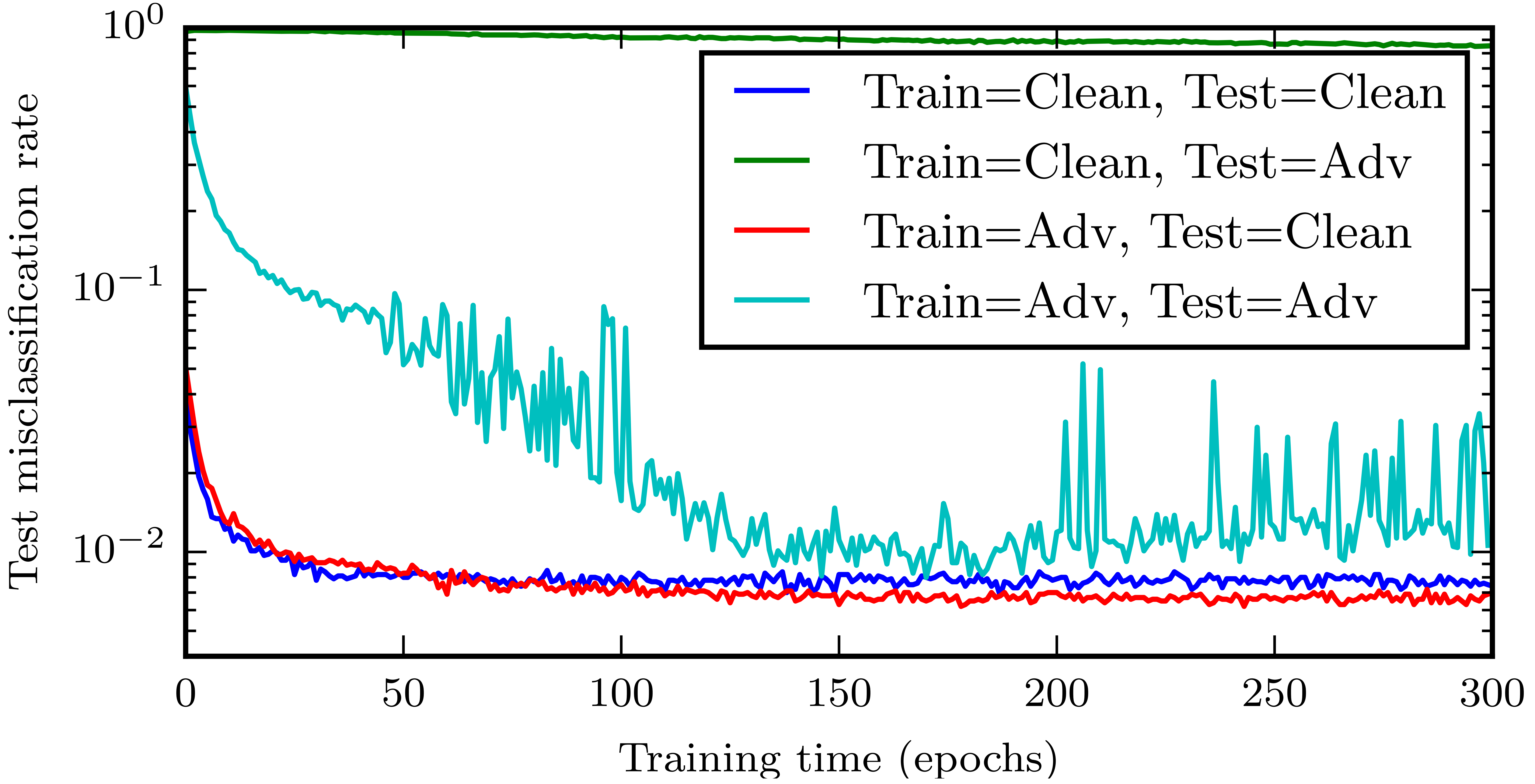
with the learning algorithm as the minimizing player and a fixed procedure (such as L-BFGS or the fast gradient sign method) as the maximizing player."

Original implementation: Goodfellow et al 2014

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

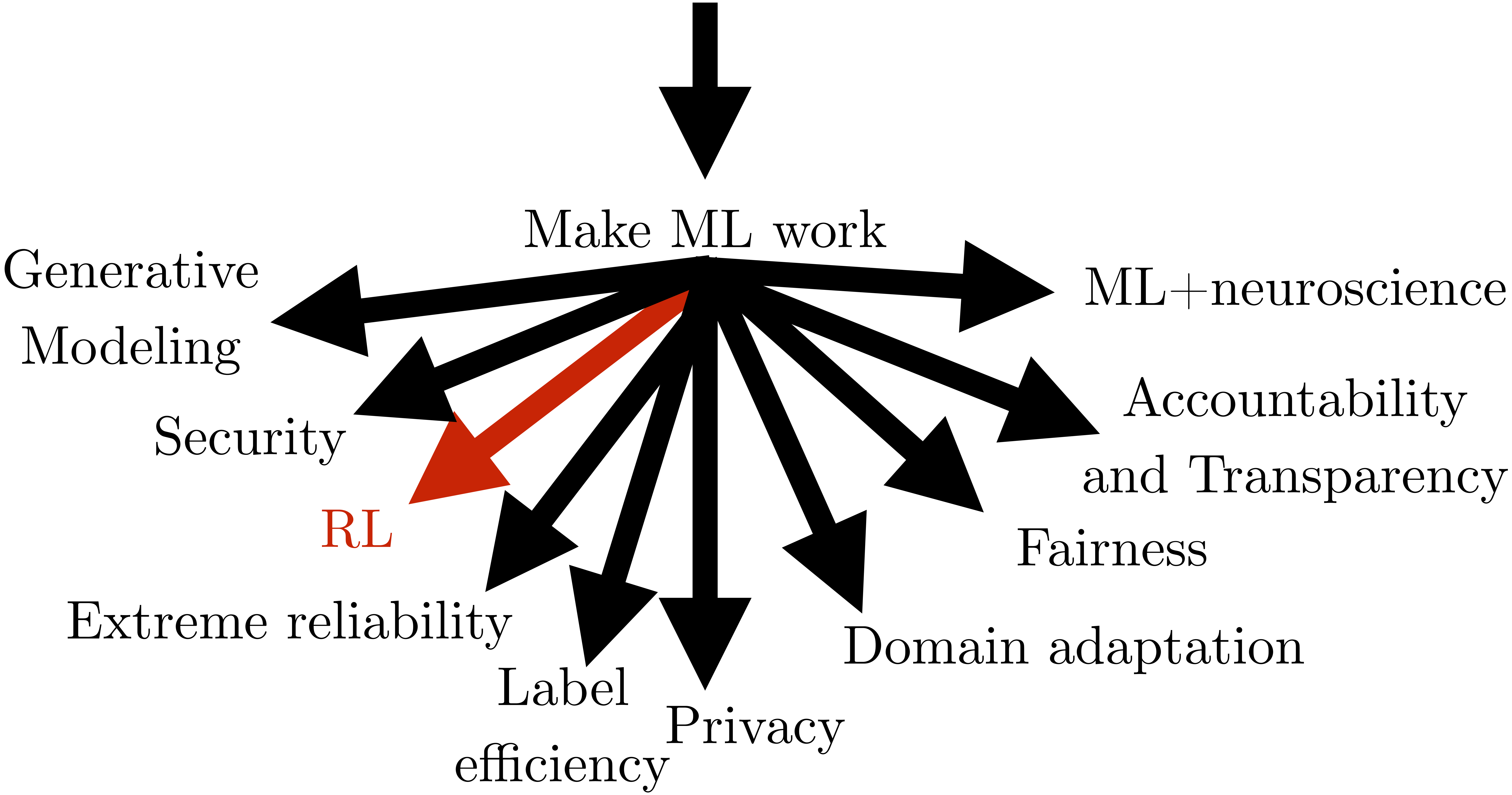
(Goodfellow 2018)

Training on Adversarial Examples

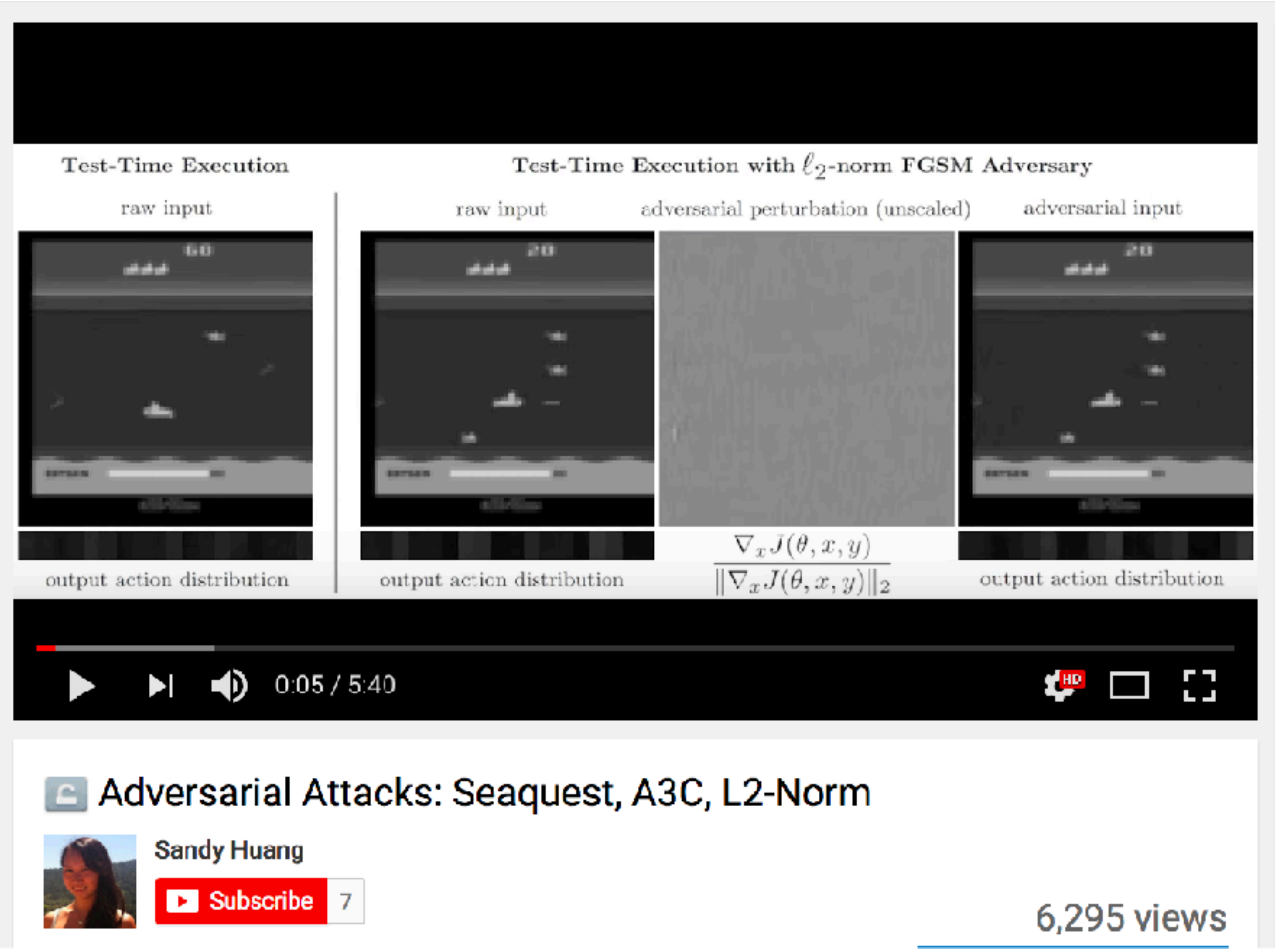


(CleverHans tutorial, using method of Goodfellow et al 2014)

A Cambrian Explosion of Machine Learning Research Topics



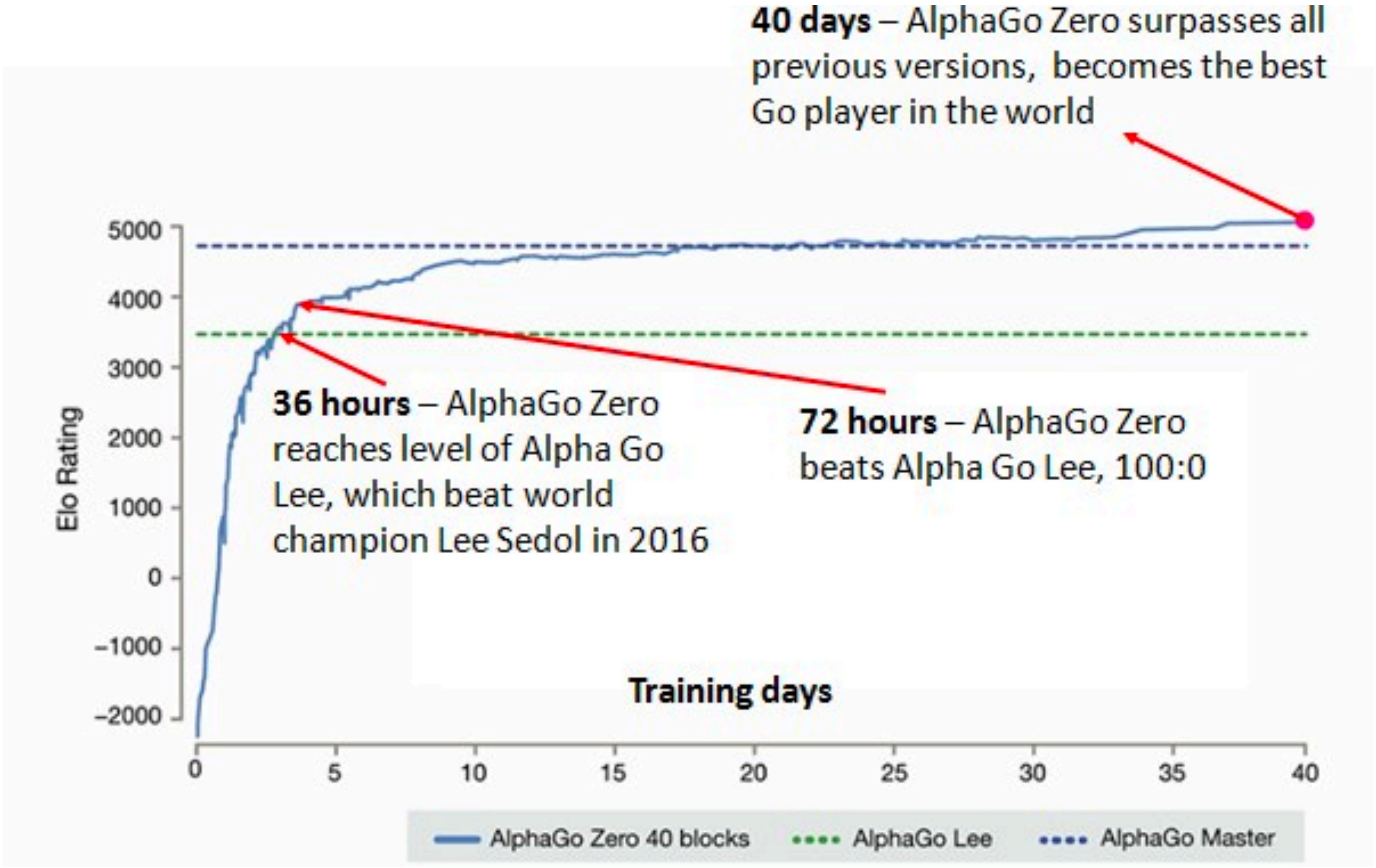
Adversarial Examples for RL



(Huang et al., 2017)

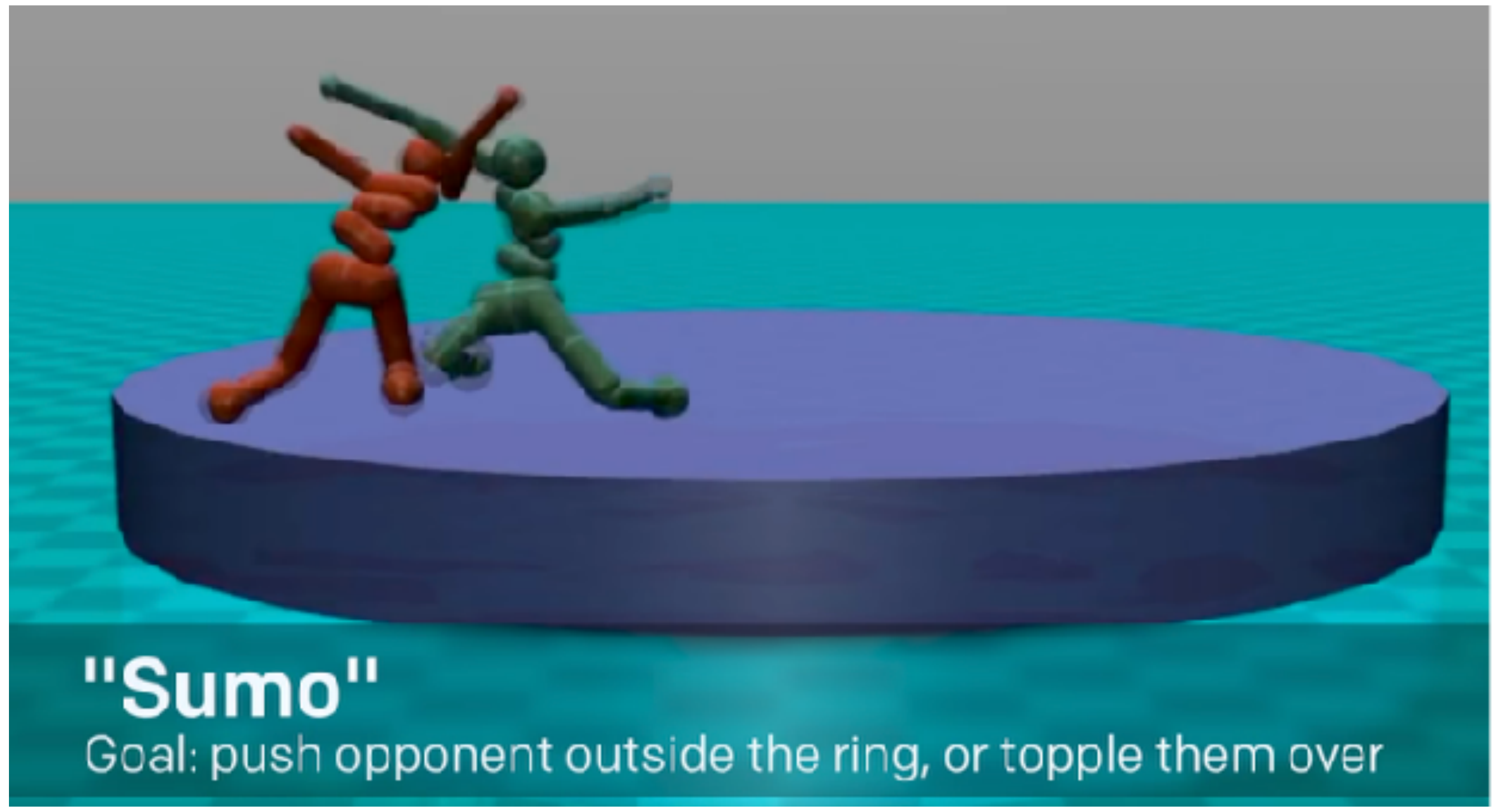
Self-Play

1959: Arthur Samuel's checkers agent



(Silver et al, 2017)

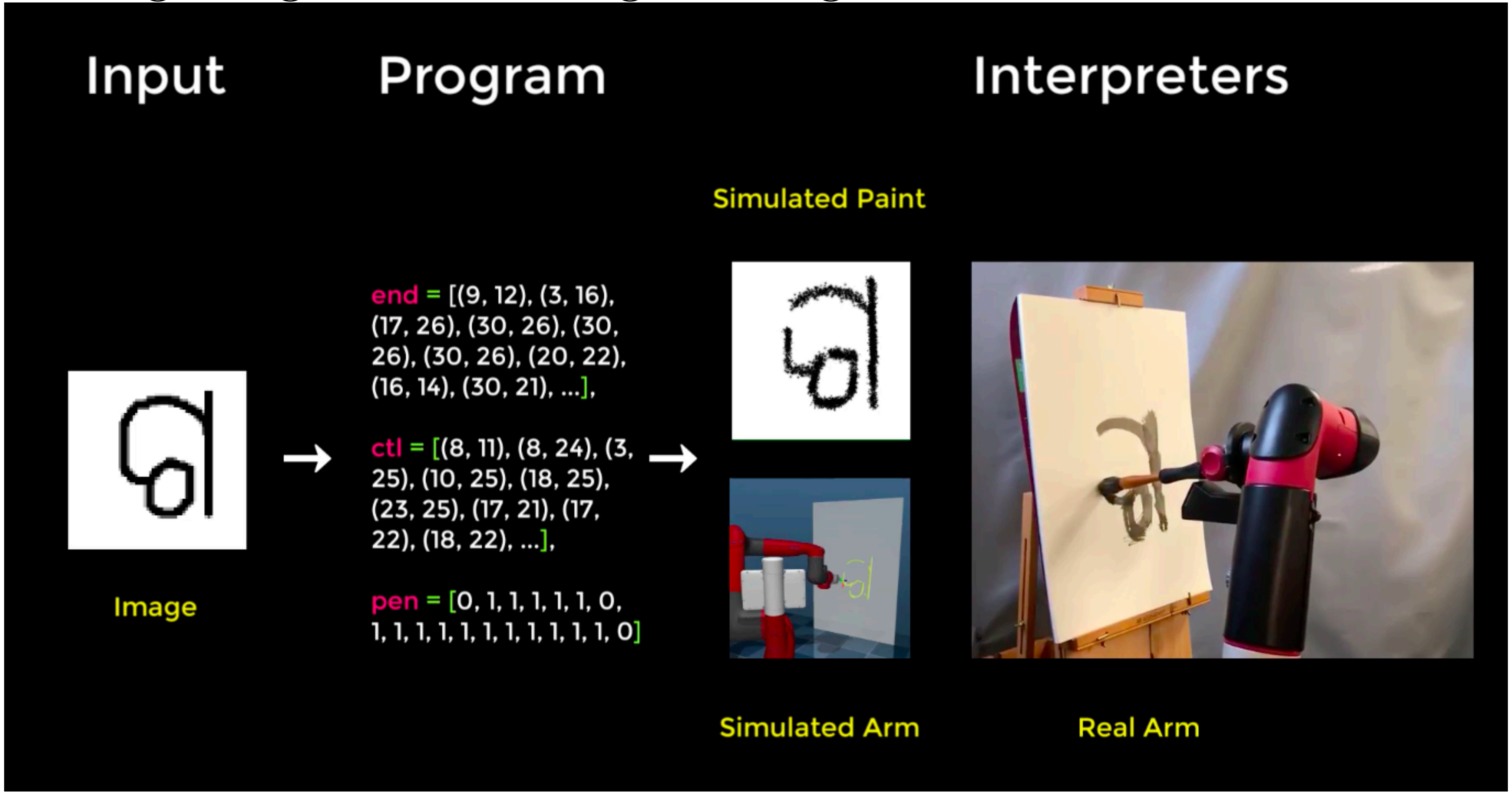
(OpenAI, 2017)



(Bansal et al, 2017)

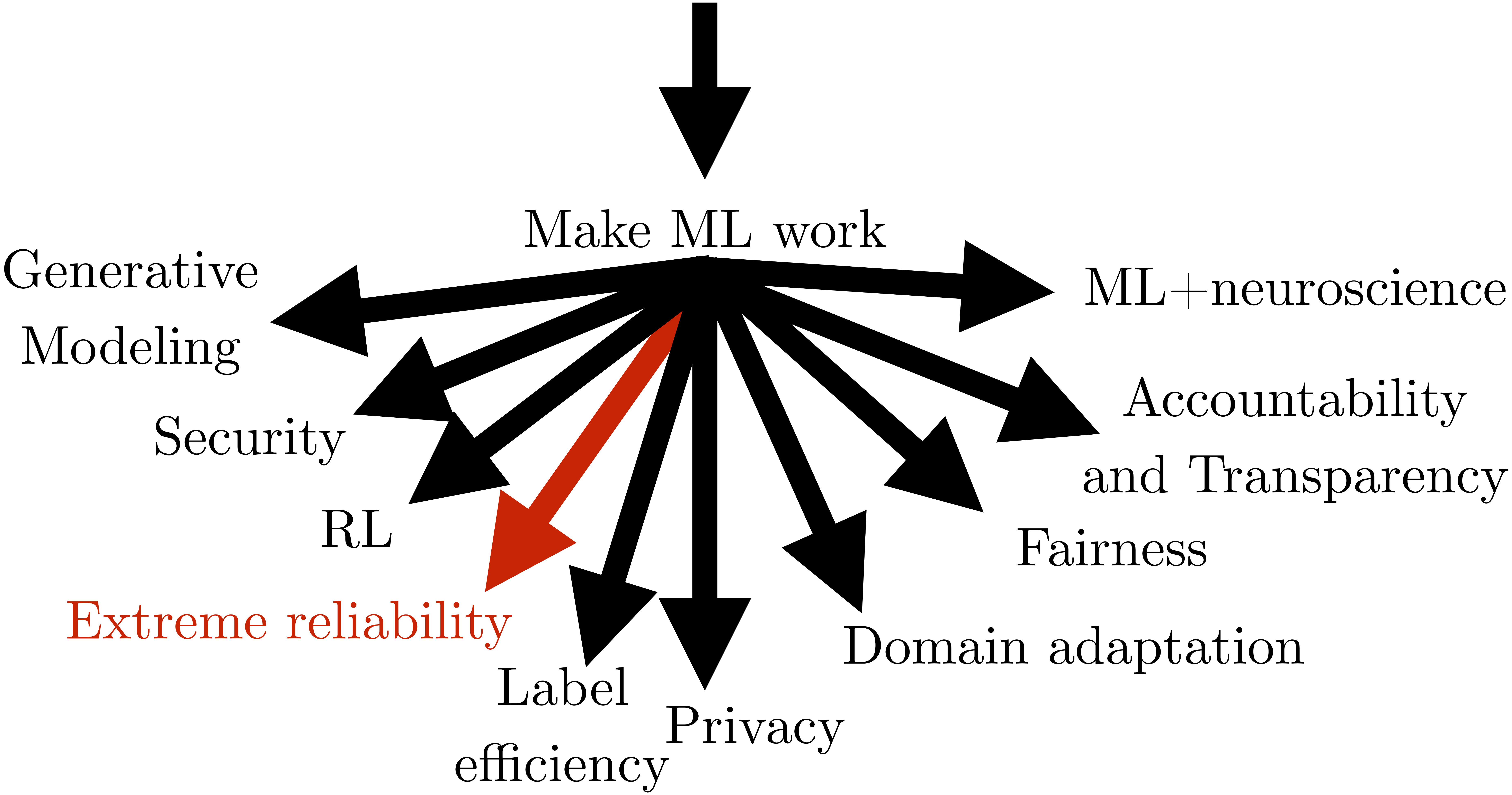
SPIRAL

Synthesizing Programs for Images Using Reinforced Adversarial Learning



(Ganin et al, 2018)

A Cambrian Explosion of Machine Learning Research Topics

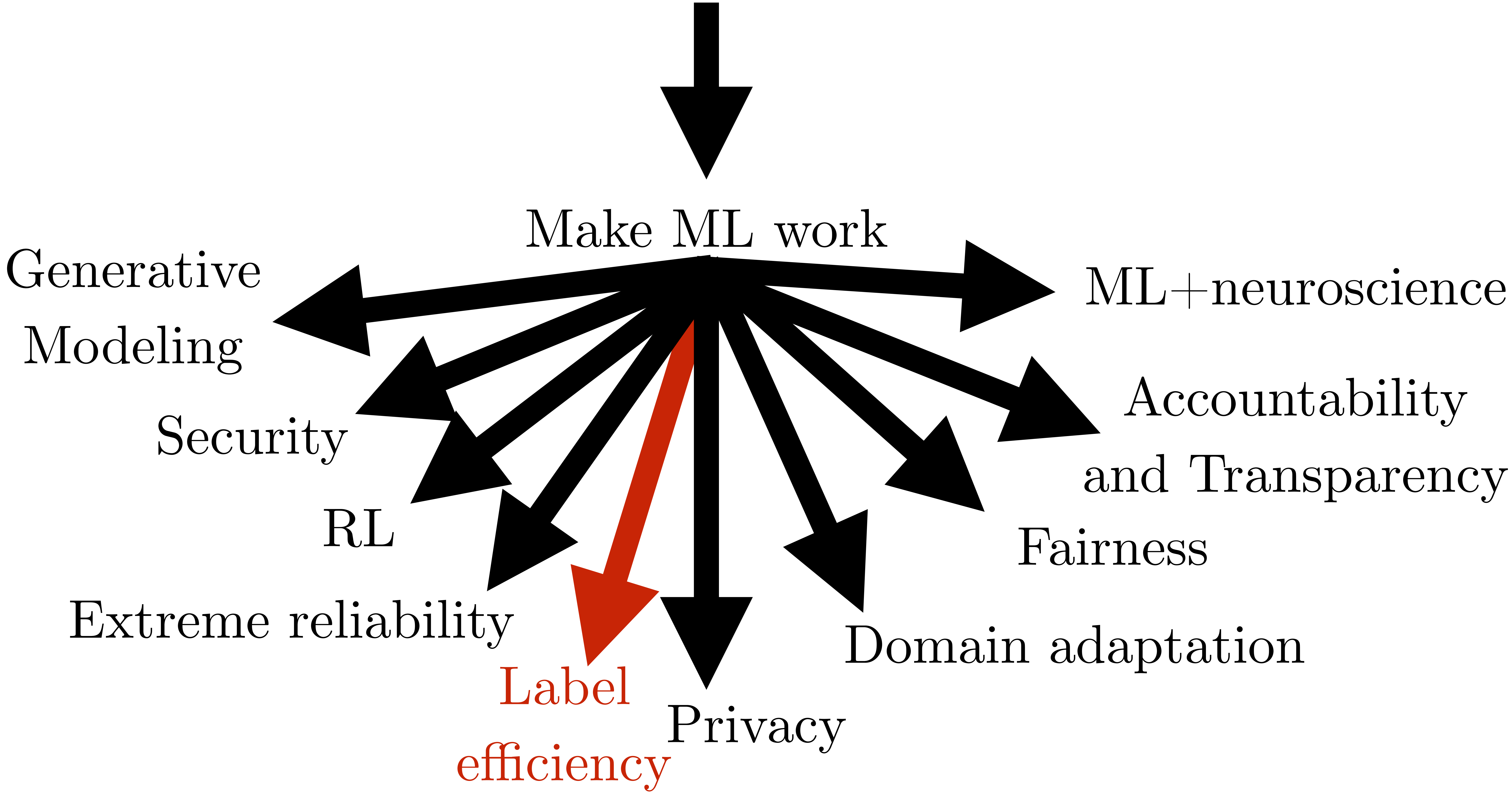


(Goodfellow 2018)

Extreme Reliability

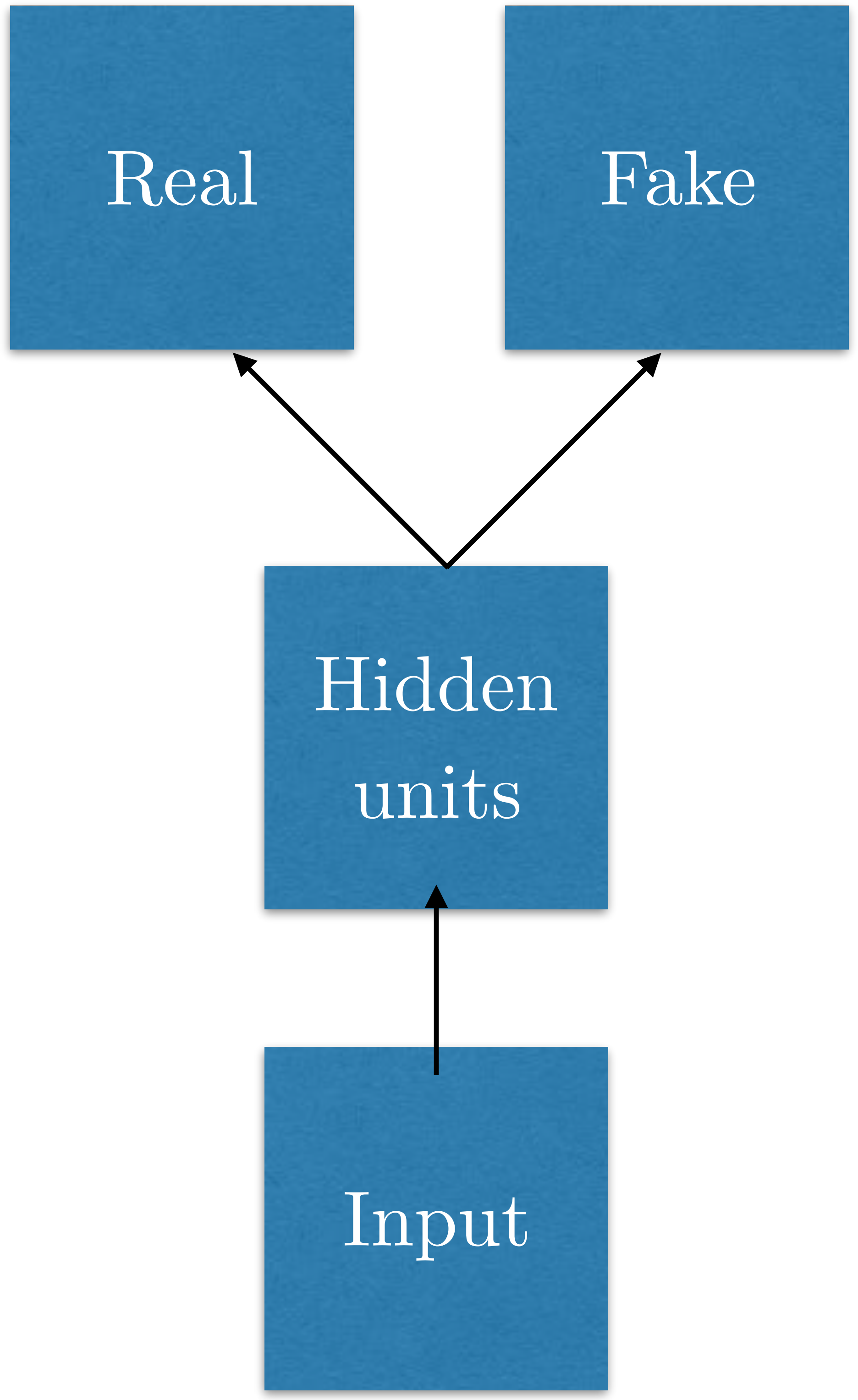
- We want extreme reliability for
  - Autonomous vehicles
  - Air traffic control
  - Surgery robots
  - Medical diagnosis, etc.
- Adversarial machine learning research techniques can help with this
  - Katz et al 2017: verification system, applied to air traffic control

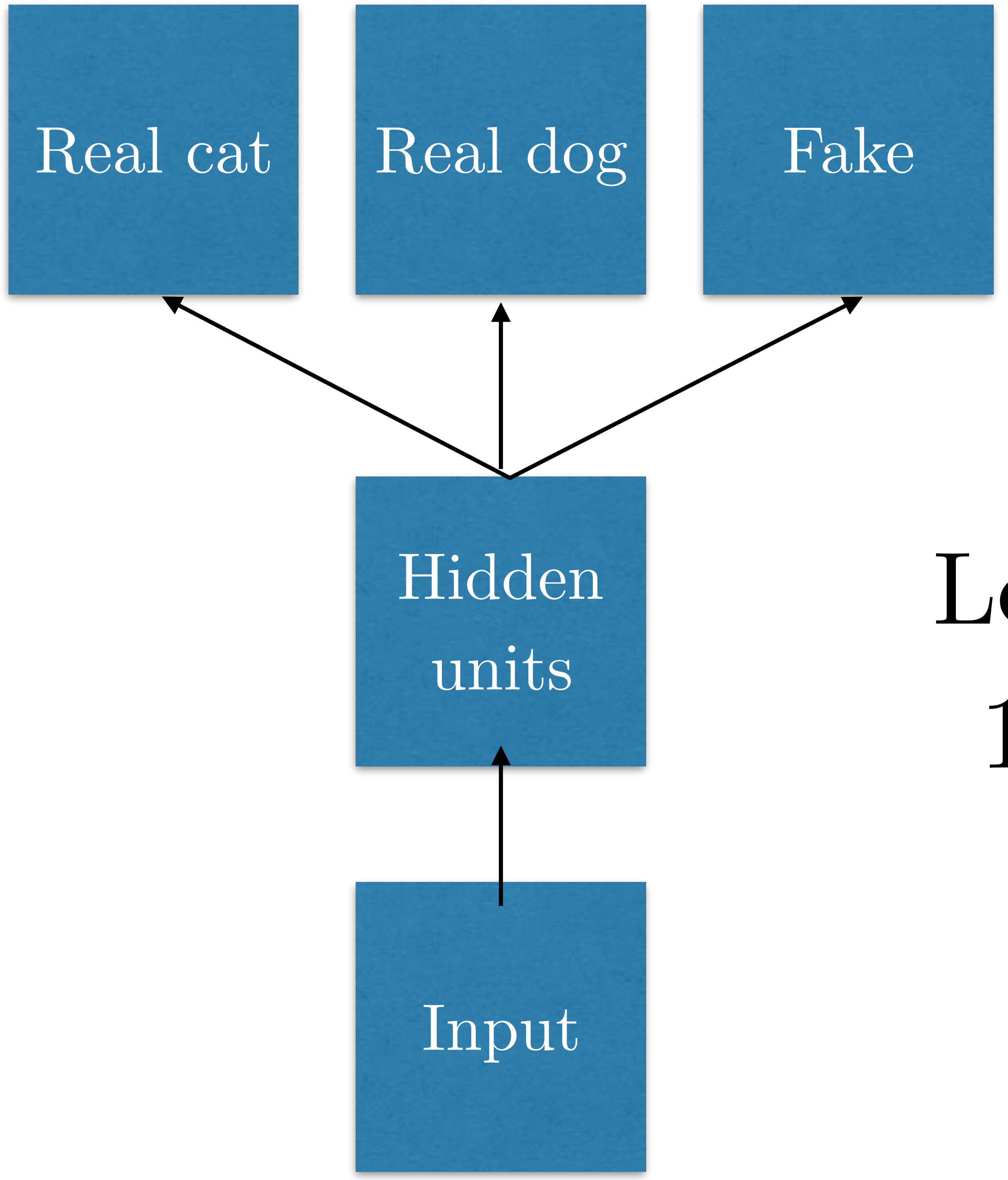
A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

Supervised Discriminator for Semi-Supervised Learning



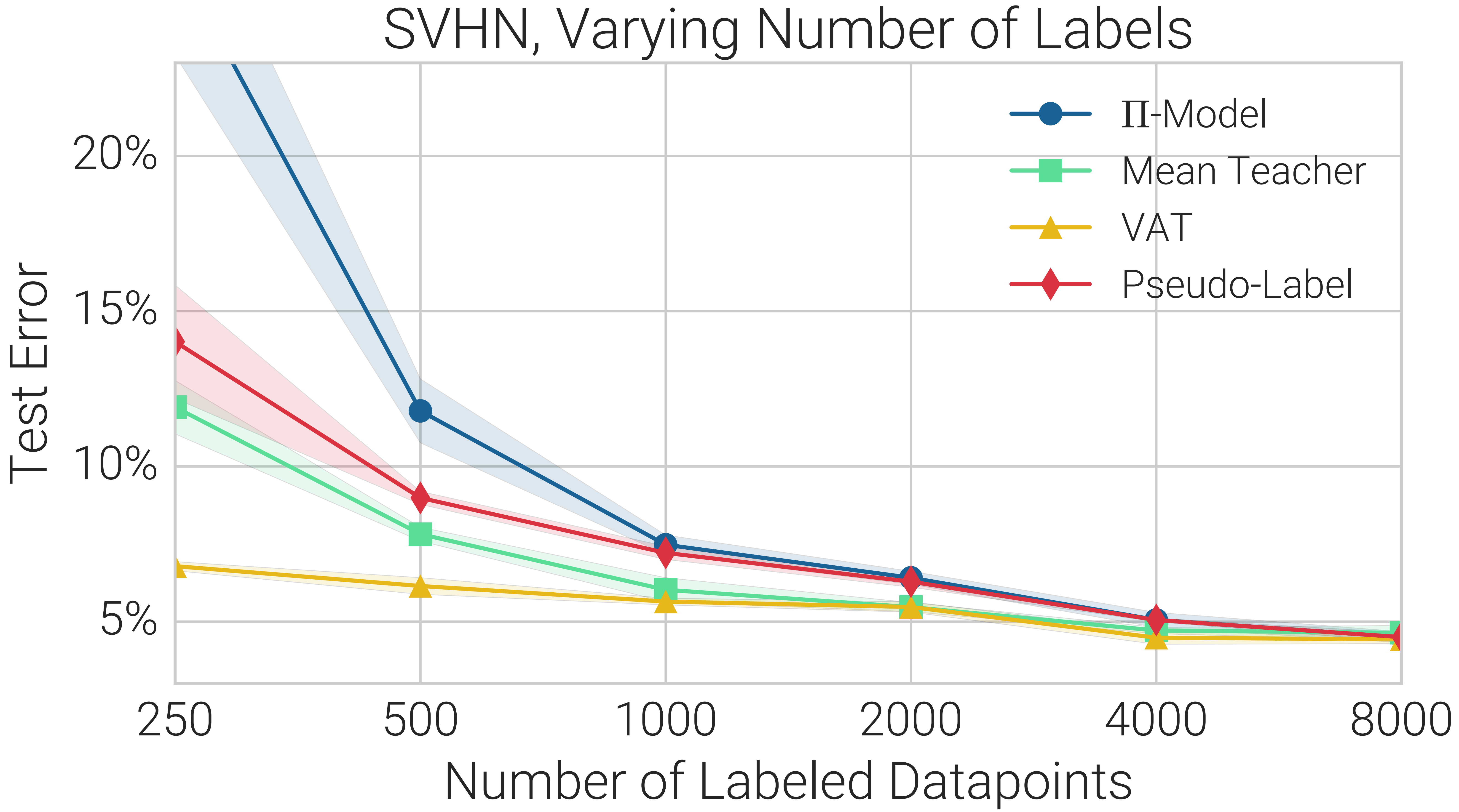


Learn to read with 100 labels rather than 60,000

(Odena 2016, Salimans et al 2016)

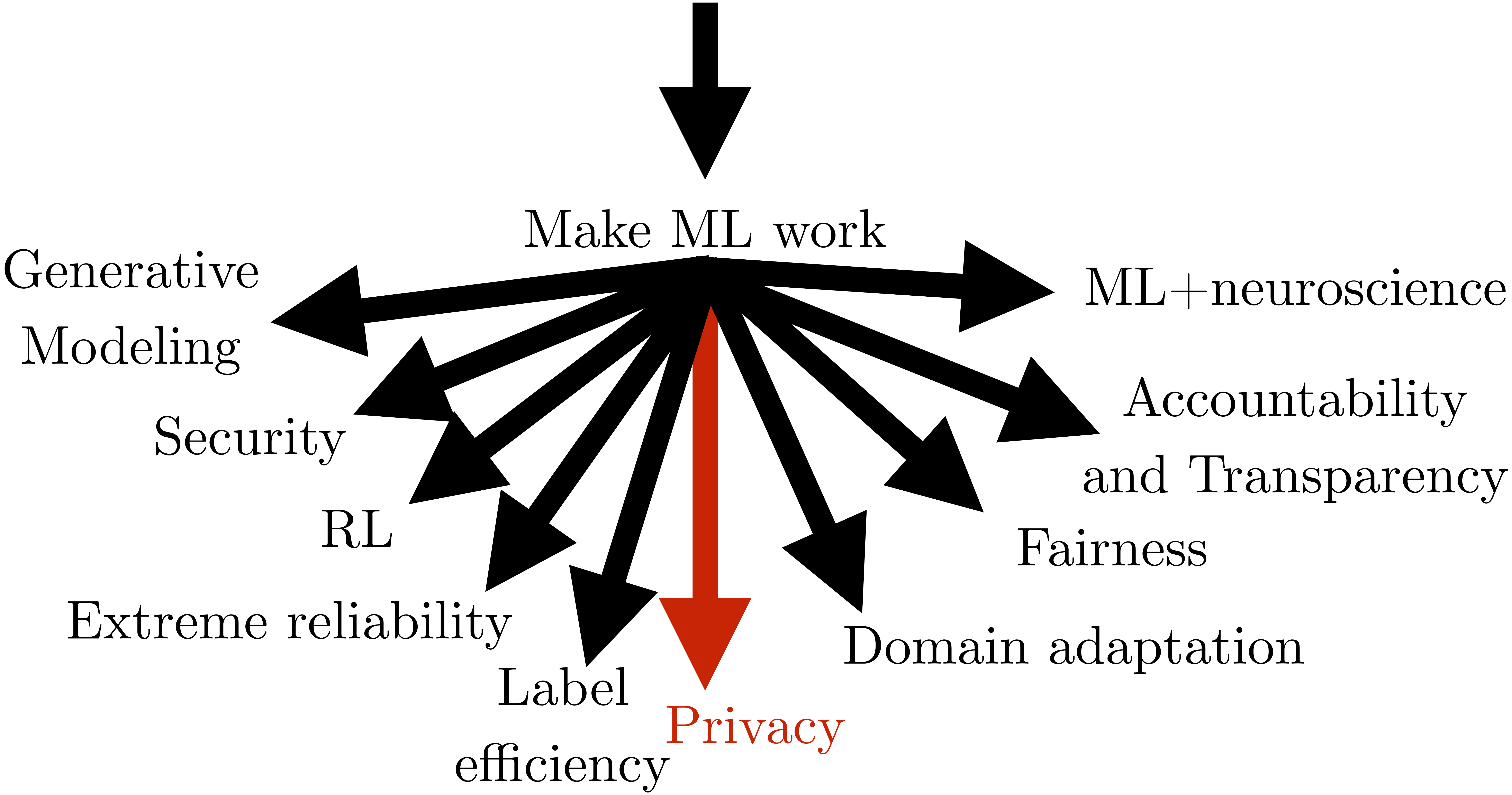
Virtual Adversarial Training

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



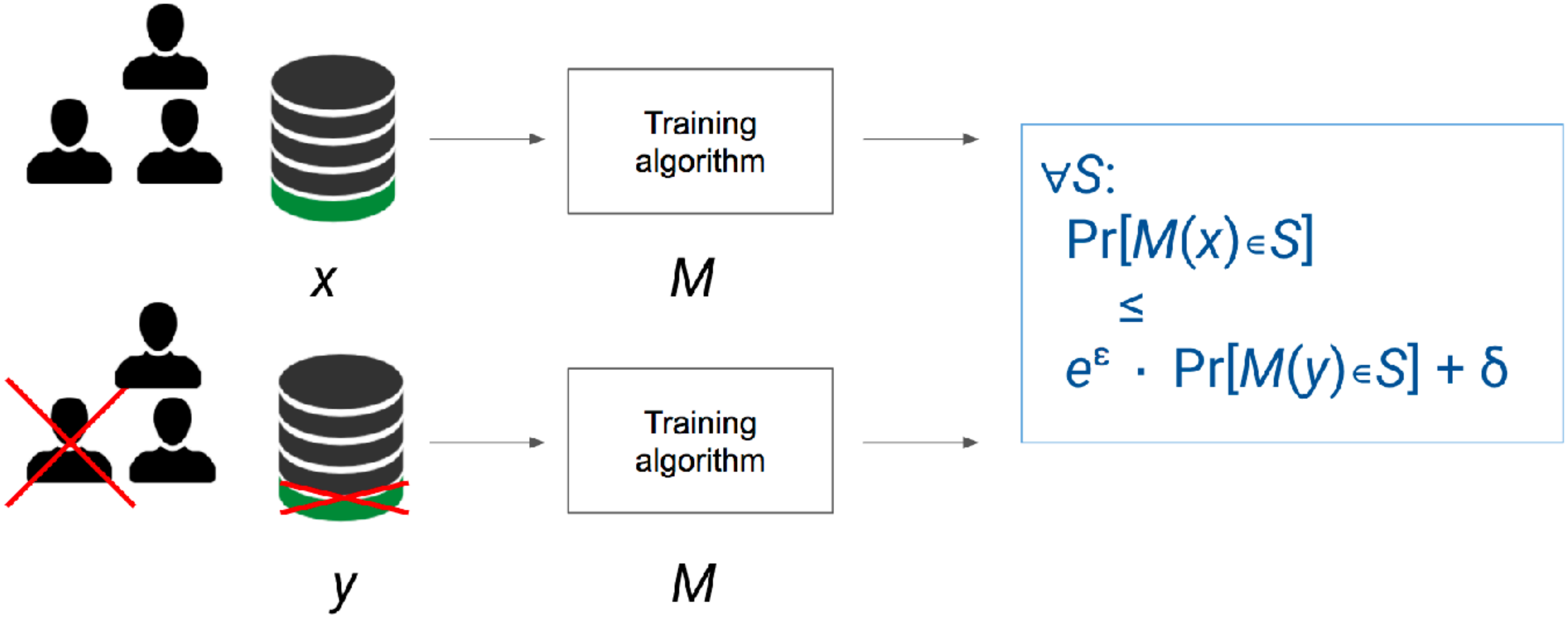
(Goodfellow 2018)

Privacy of training data

$$\boldsymbol{X} \rightarrow \;\vdots\; \boldsymbol{\theta} \rightarrow \hat{\boldsymbol{X}}$$

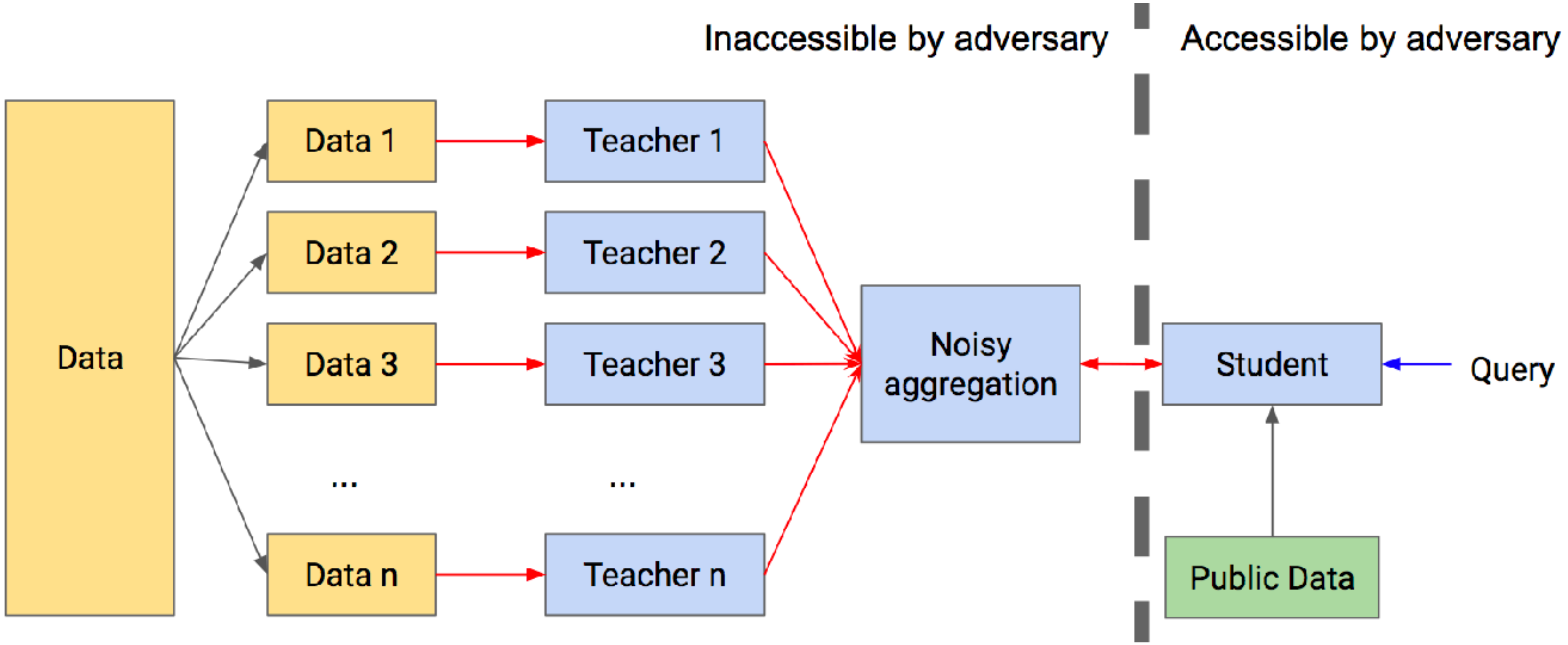
(Goodfellow 2018)

Defining (ε, δ)-Differential Privacy



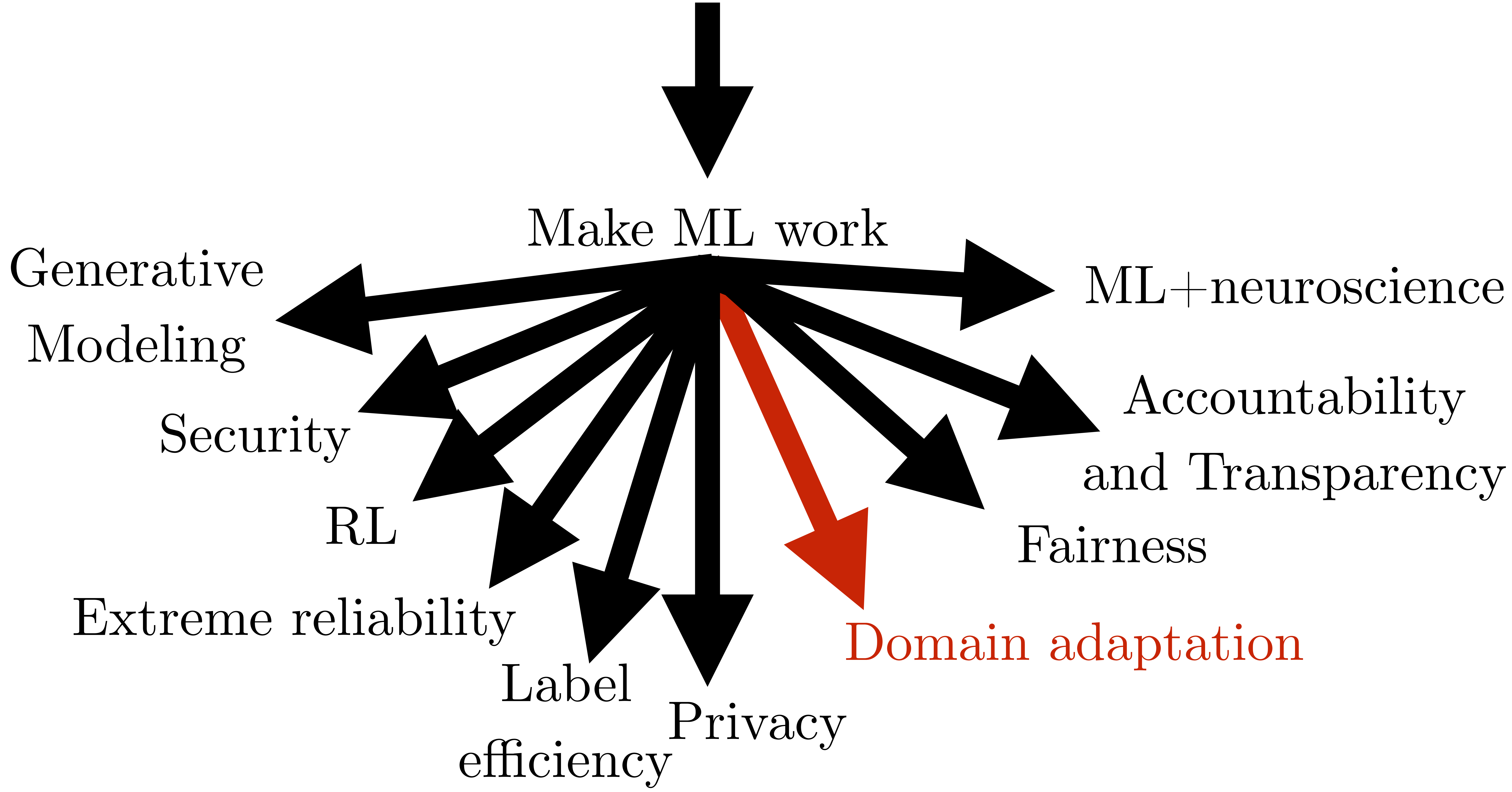
(Abadi 2017)

Private Aggregation of Teacher Ensembles



(Papernot et al 2016)

A Cambrian Explosion of Machine Learning Research Topics

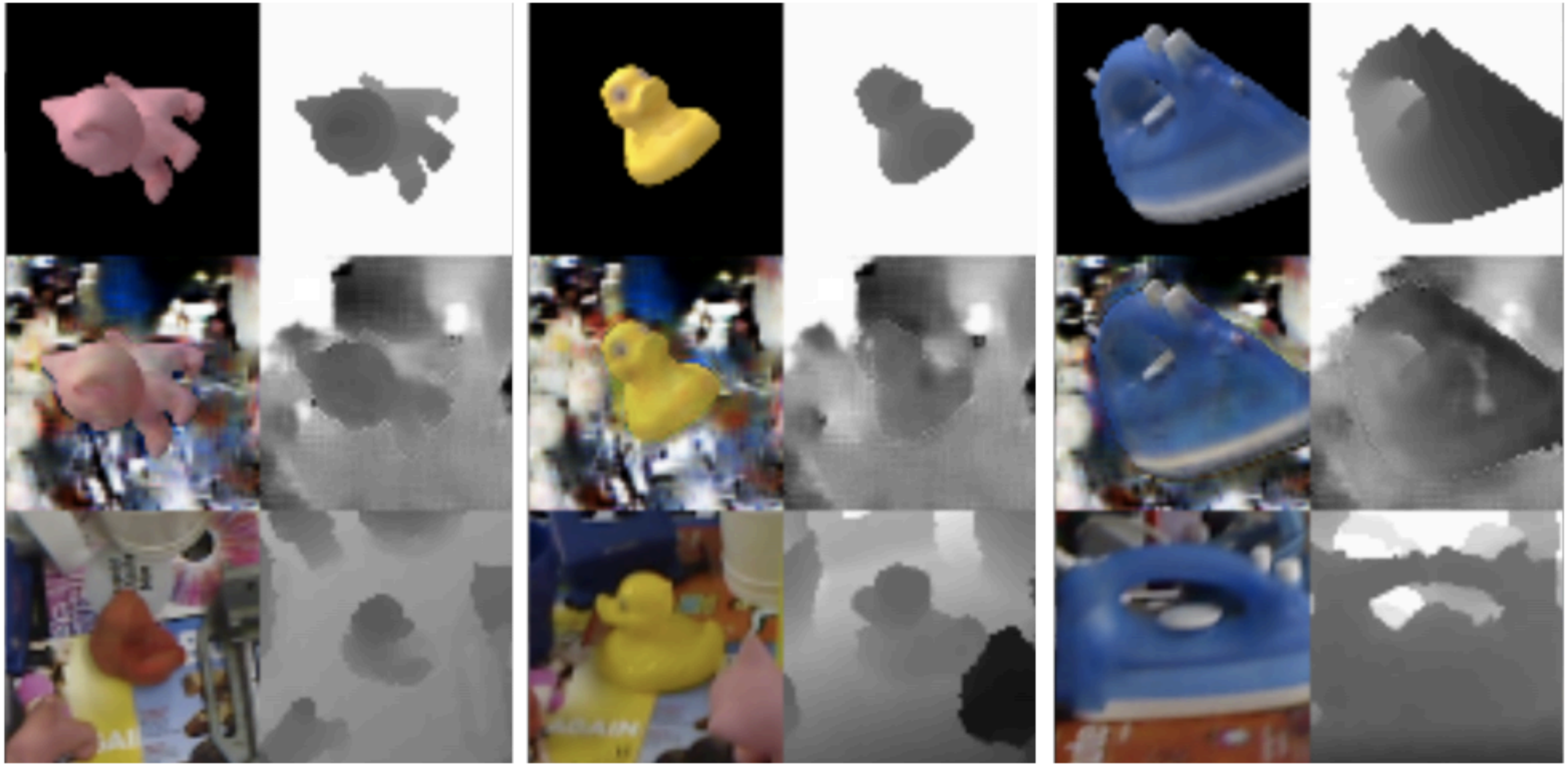


Domain Adaptation

- Domain Adversarial Networks (Ganin et al, 2015)

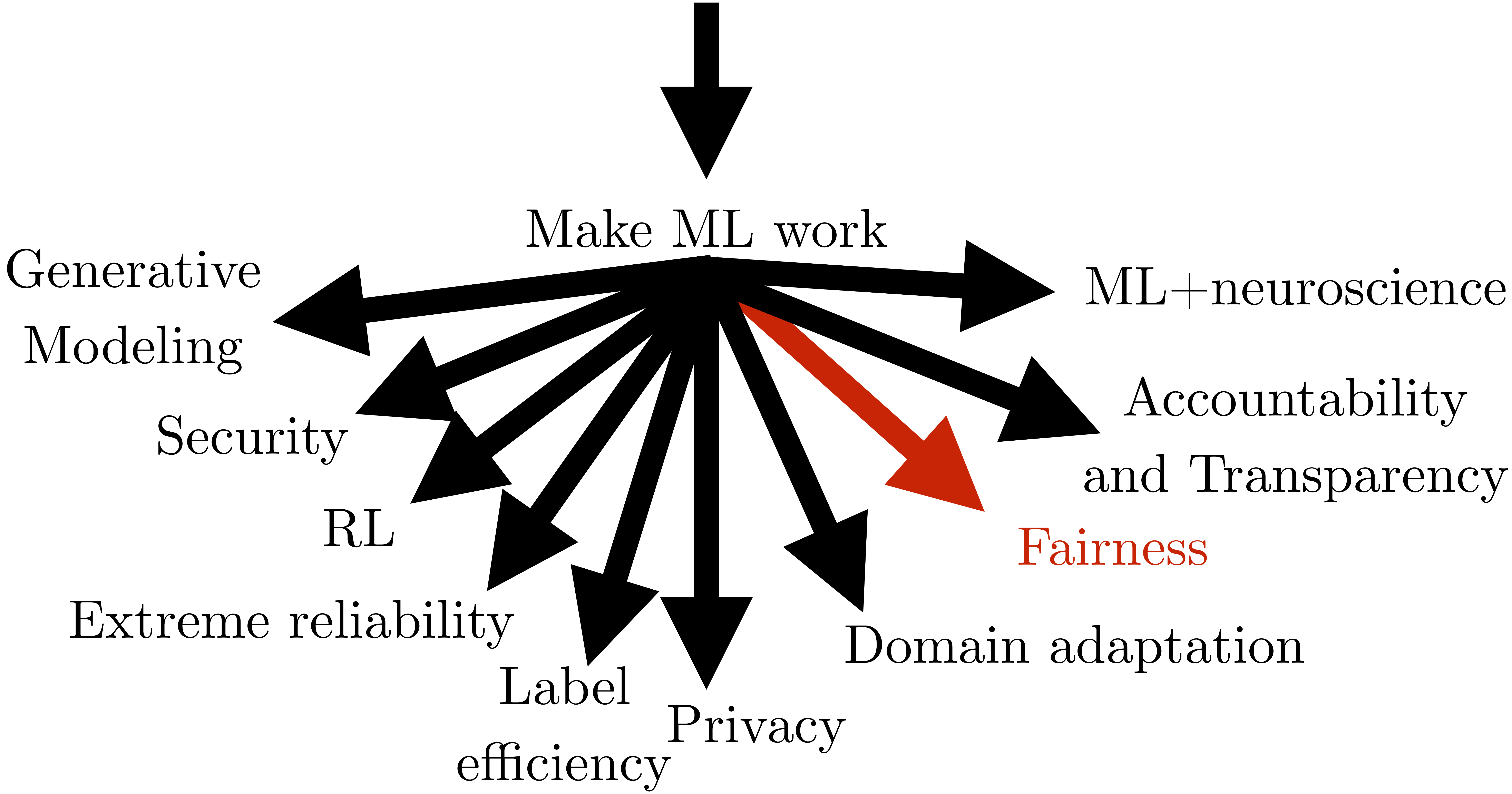
- Professor forcing (Lamb et al, 2016): Domain-Adversarial learning in RNN hidden state

GANs for domain adaptation



(Bousmalis et al., 2016)

A Cambrian Explosion of Machine Learning Research Topics

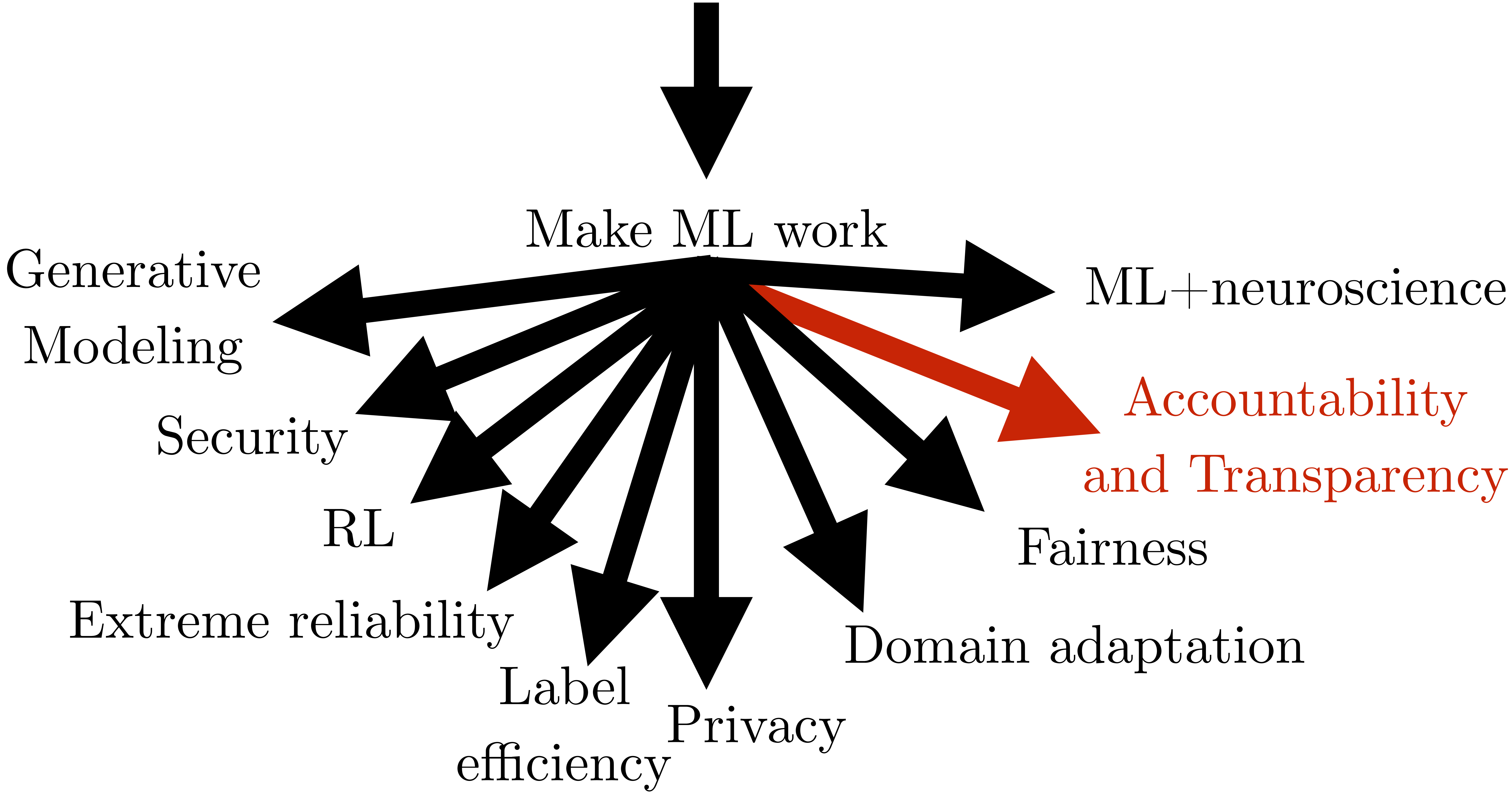


(Goodfellow 2018)

Adversarially Learned Fair Representations

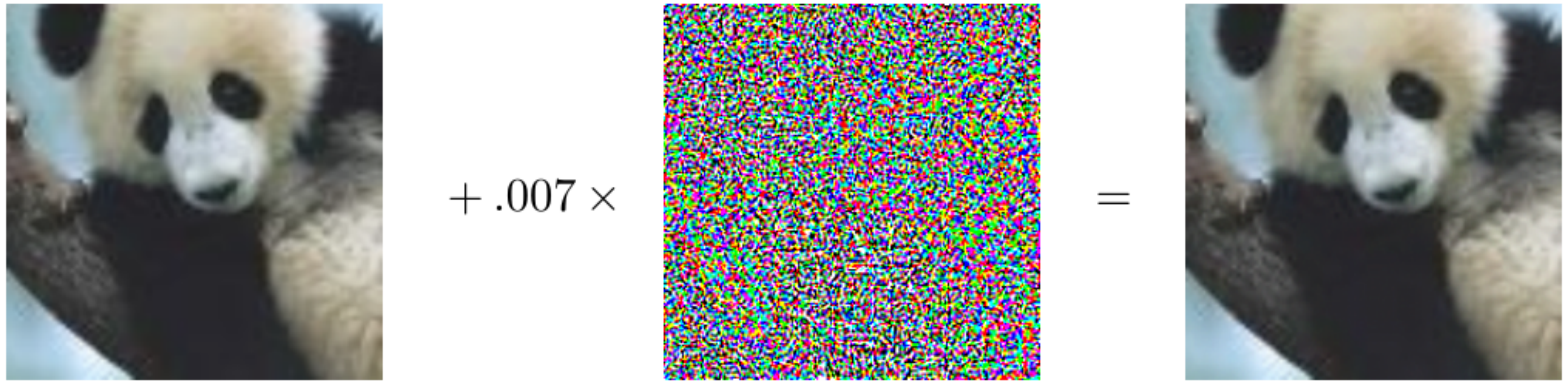
- Edwards and Storkey 2015
- Learn representations that are useful for classification
- An adversary tries to recover a sensitive variable S from the representation. Primary learner tries to make S impossible to recover
- Final decision does not depend on S

A Cambrian Explosion of Machine Learning Research Topics

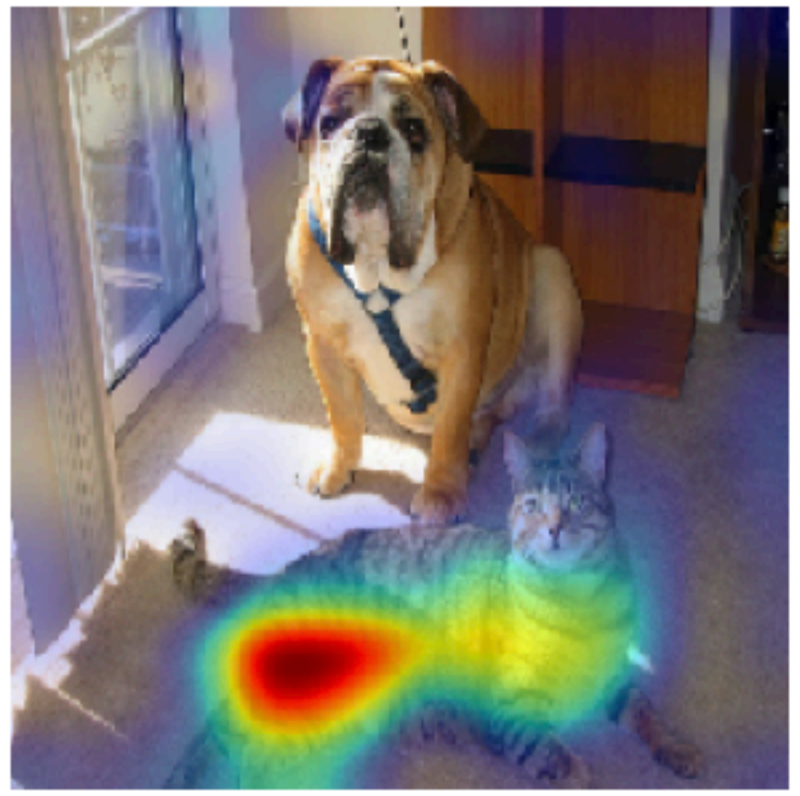


(Goodfellow 2018)

How do machine learning models work?



(Goodfellow et al, 2014)



(c) Grad-CAM ‘Cat’

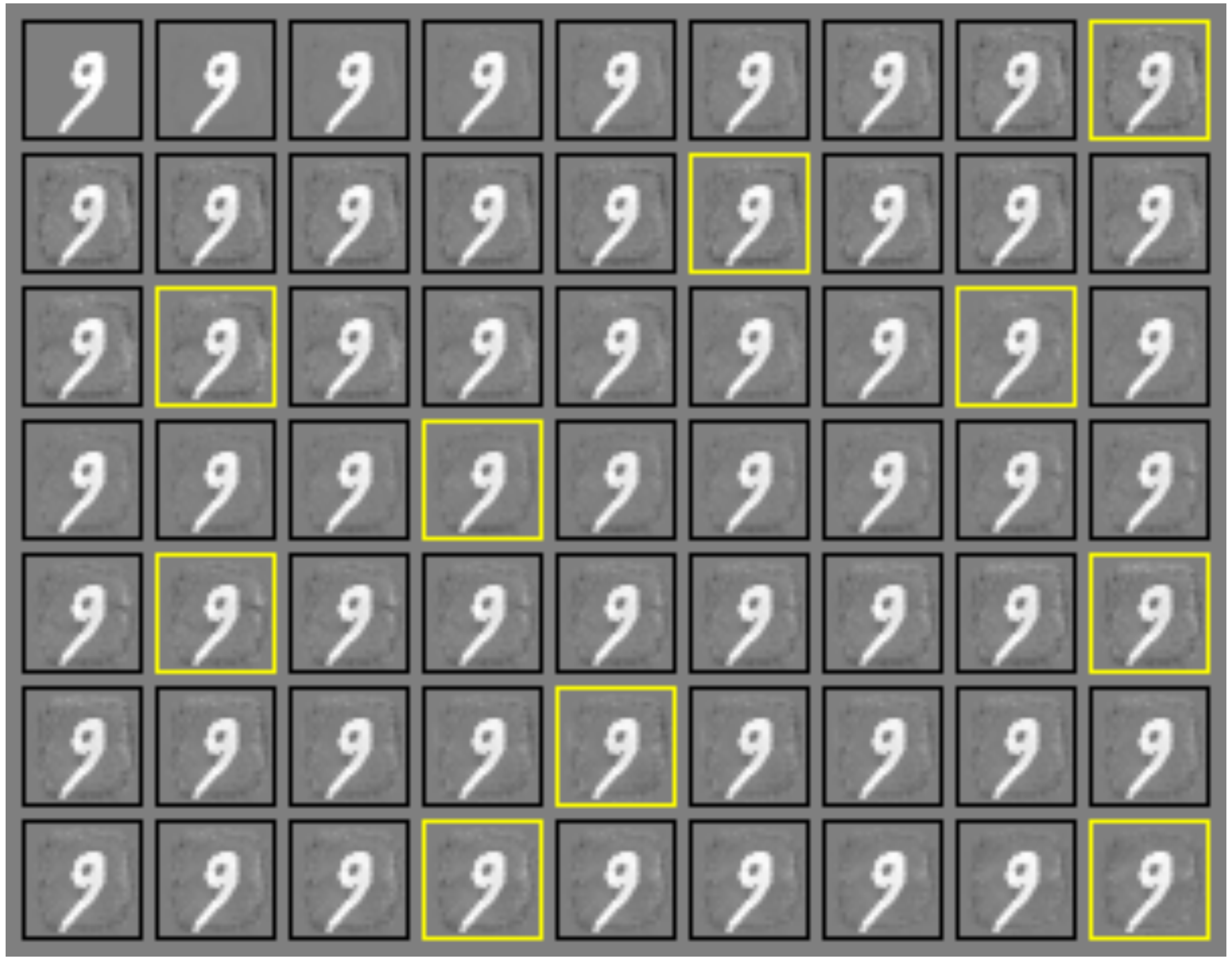
(i) Grad-CAM ‘Dog’

(Selvaraju et al, 2016)

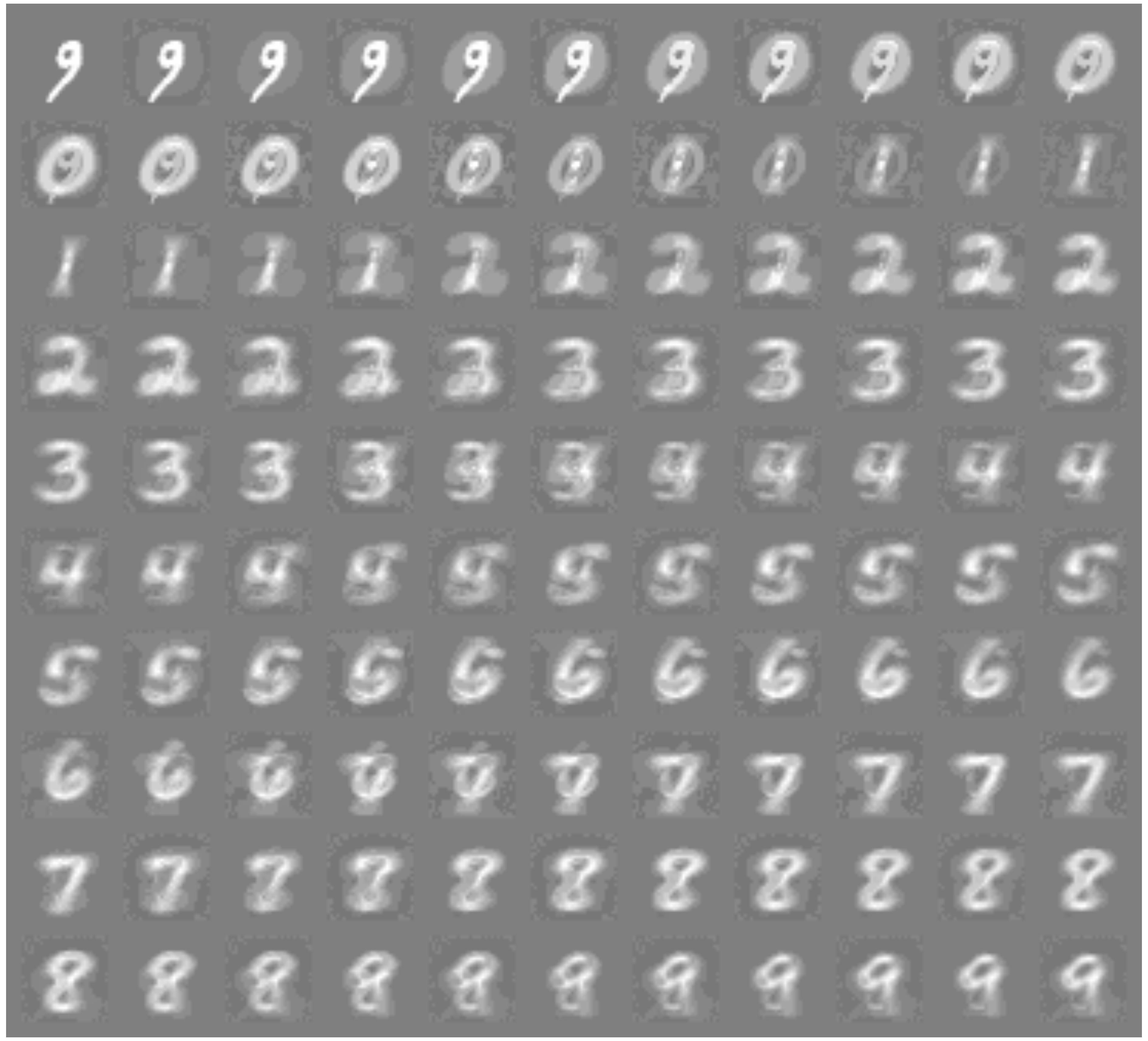
Interpretability literature: our analysis tools show that deep nets work about how you would expect them to.

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

Robust models are more interpretable



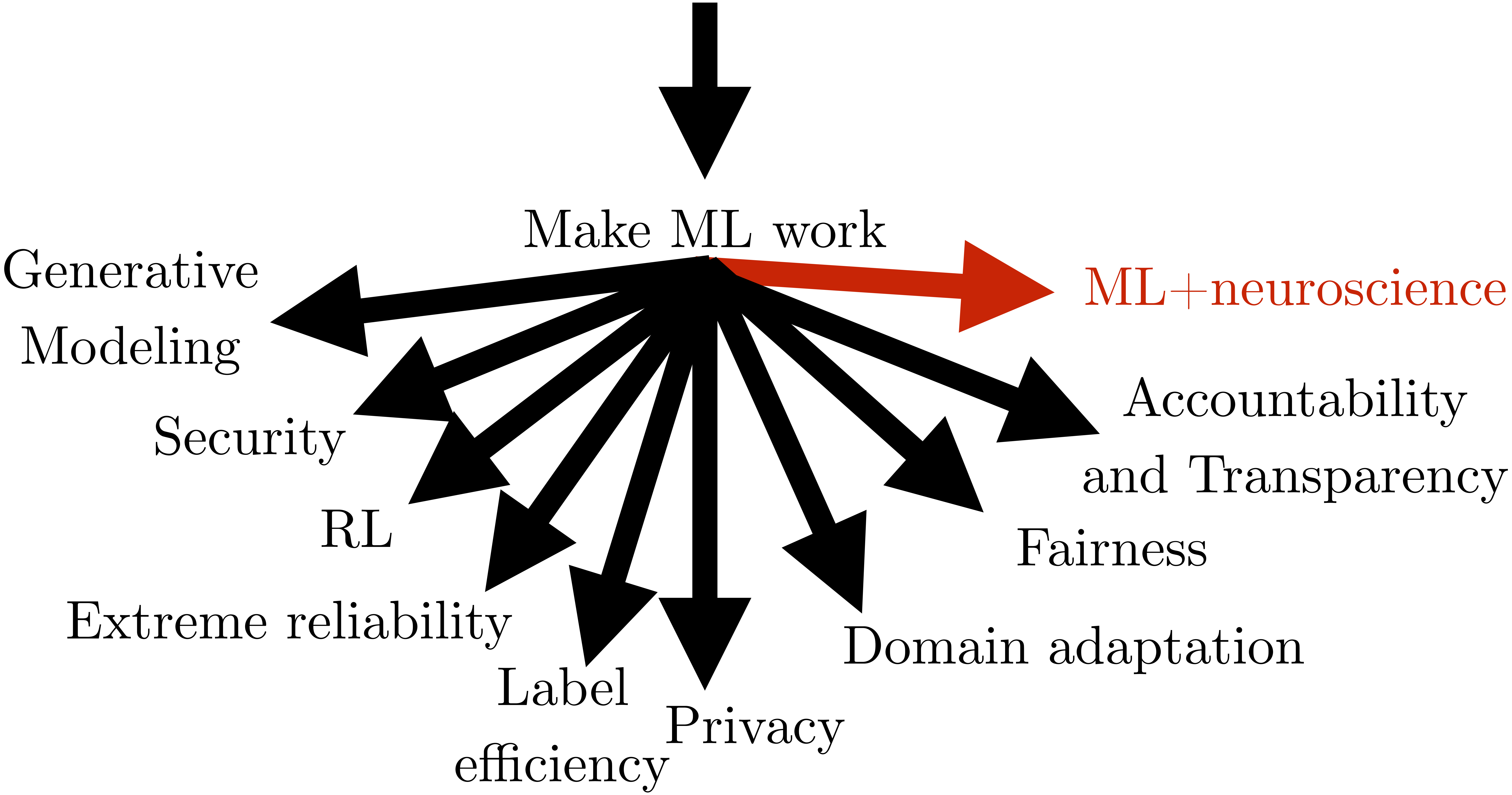
Relatively vulnerable model



Relatively robust model

(Goodfellow 2015)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

Adversarial Examples that Fool both Human and Computer Vision

Gamaleldin et al 2018

Questions