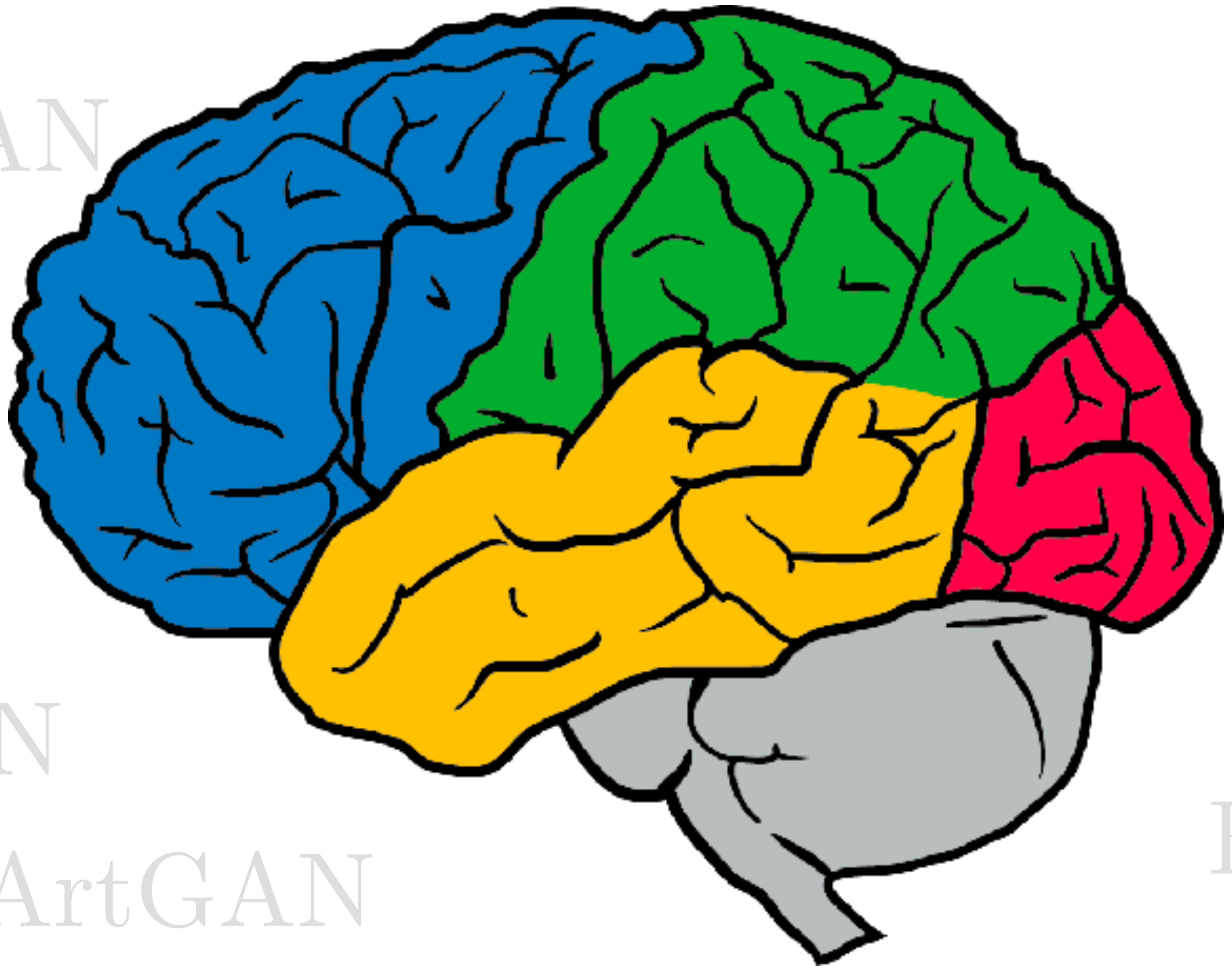
Adversarial Machine Learning

Ian Goodfellow, Staff Research Scientist, Google Brain

South Park Commons

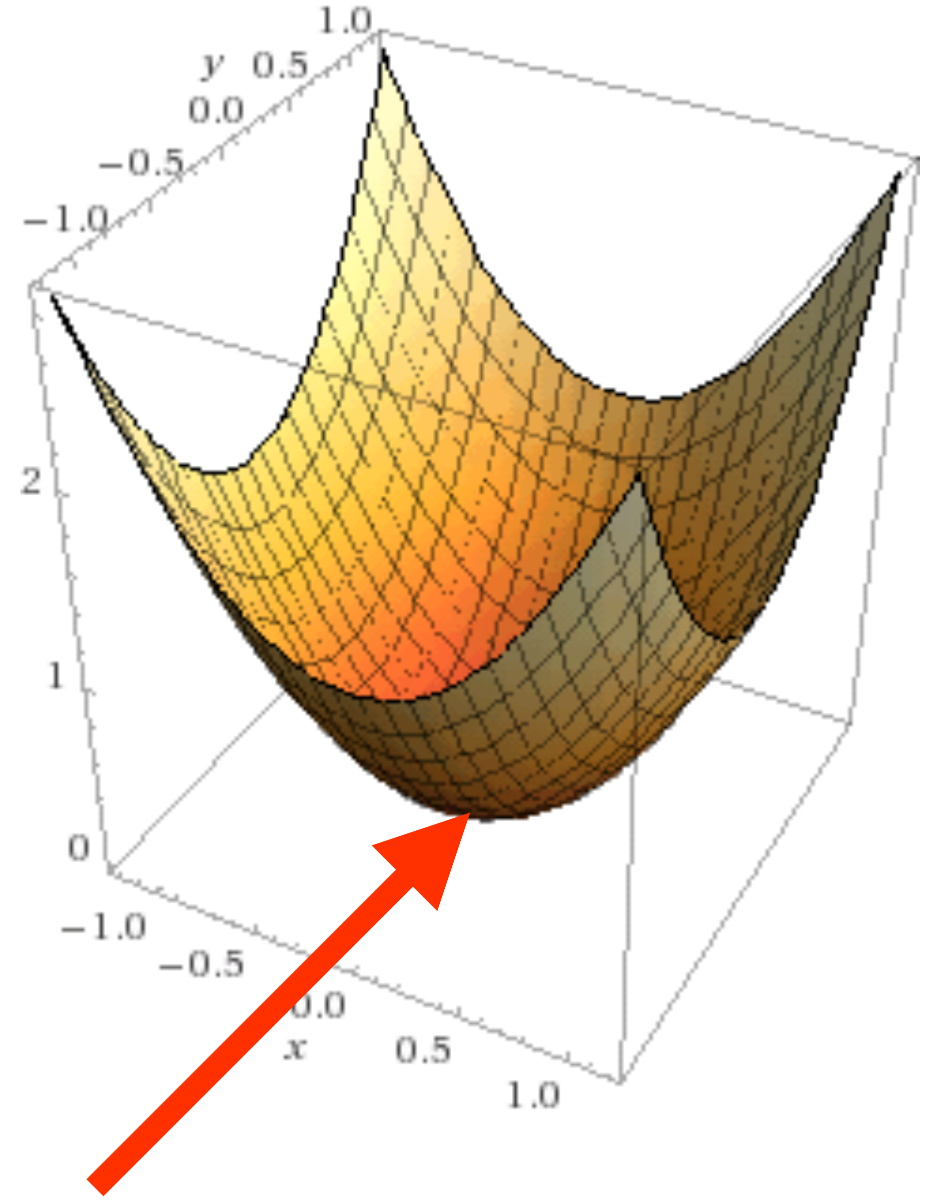
San Francisco, 2018-05-24



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 DR-GAN BS-GAN C-VAE-GAN C-RNN-GAN DCGAN CCGAN AC-GAN MAGAN 3D-GAN BiGAN Adversarial Training CycleGAN GAWWN Gradient Masking Bayesian GAN AnoGAN SN-GAN EBGAN DTN Context-RNN-GAN MAD-GAN ALI BEGAN AL-CGAN PGD f-GAN ArtGAN MalGAN

Adversarial Machine Learning

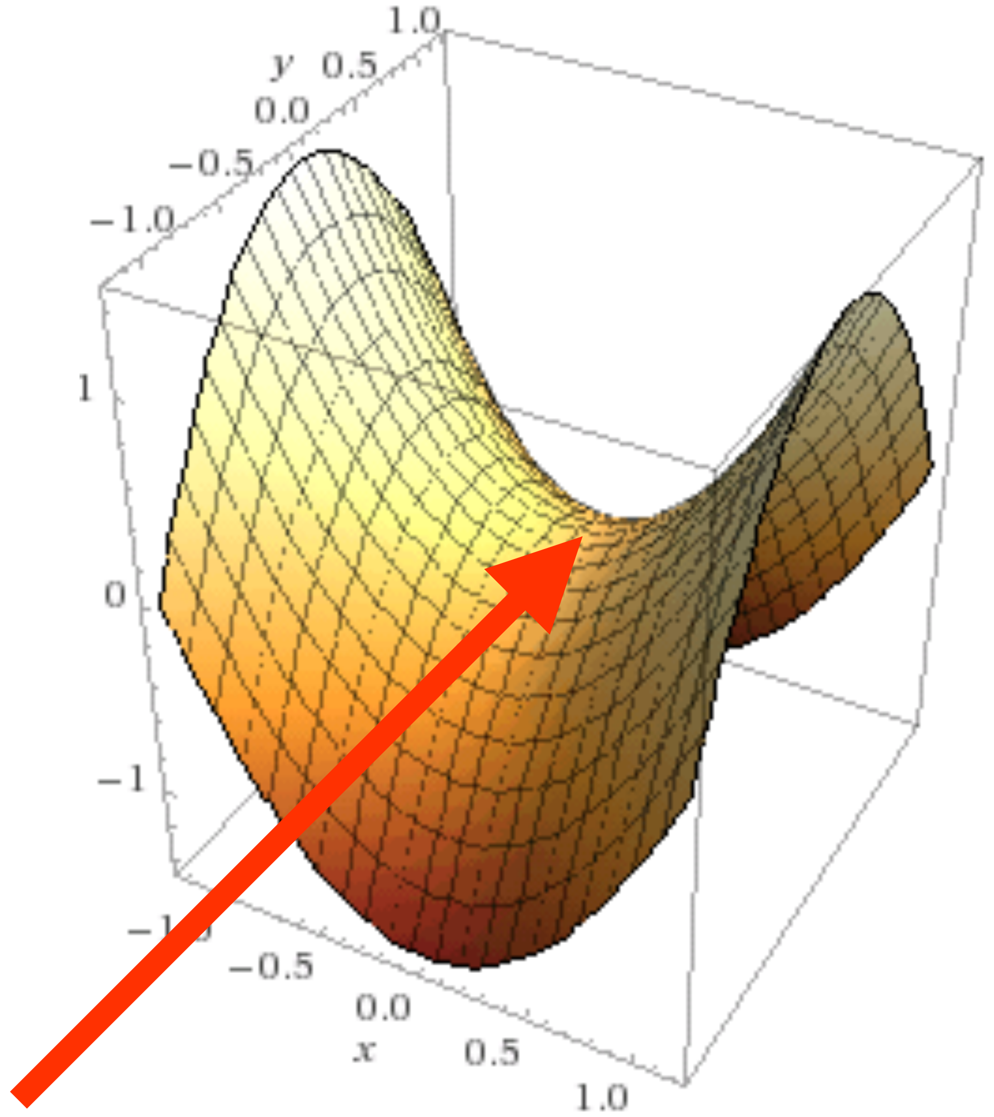
Traditional ML:
optimization



Minimum

One player,
one cost

Adversarial ML:
game theory

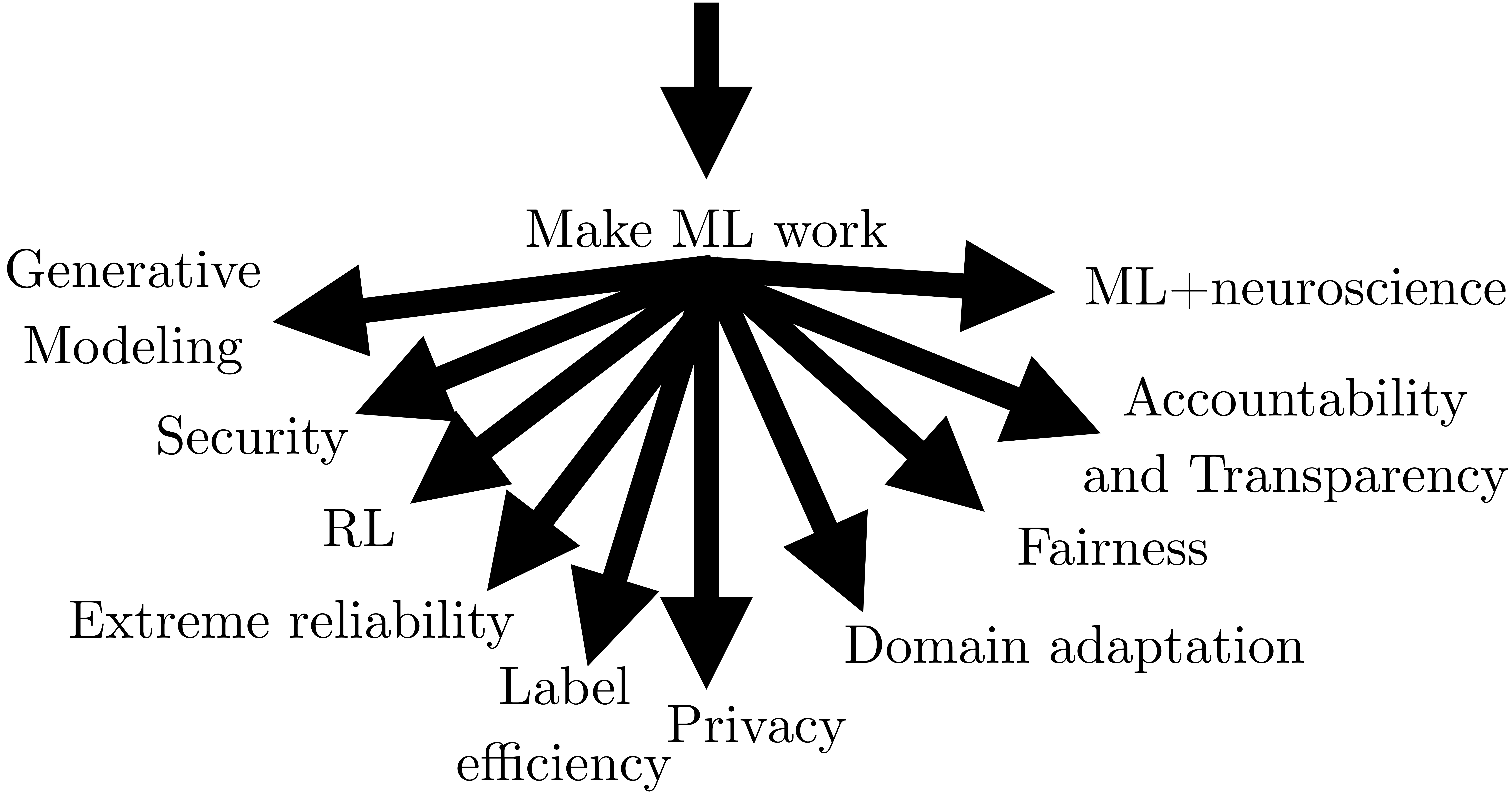


Equilibrium

More than one player,
more than one cost

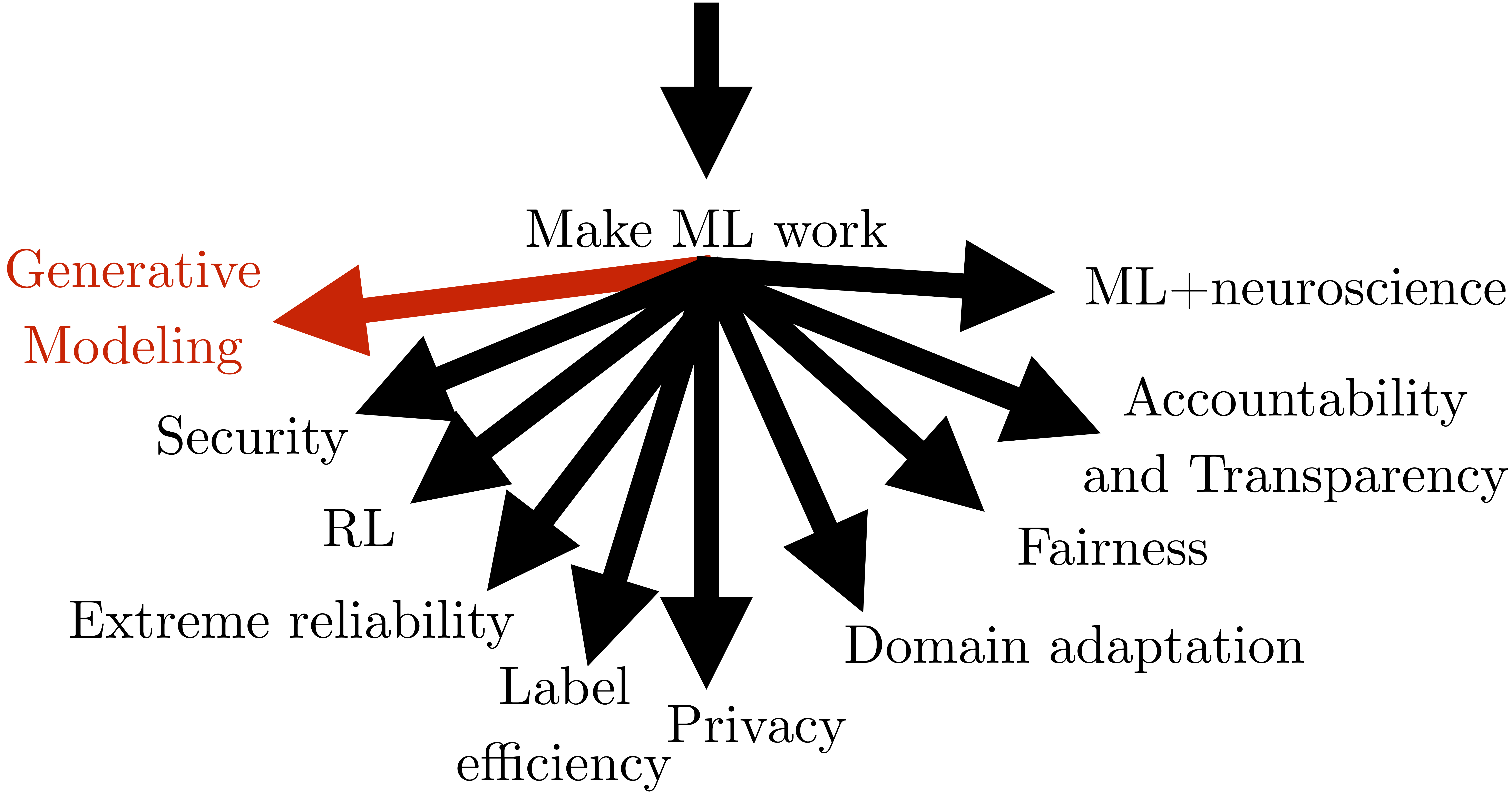
(Goodfellow 2017)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2017)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2017)

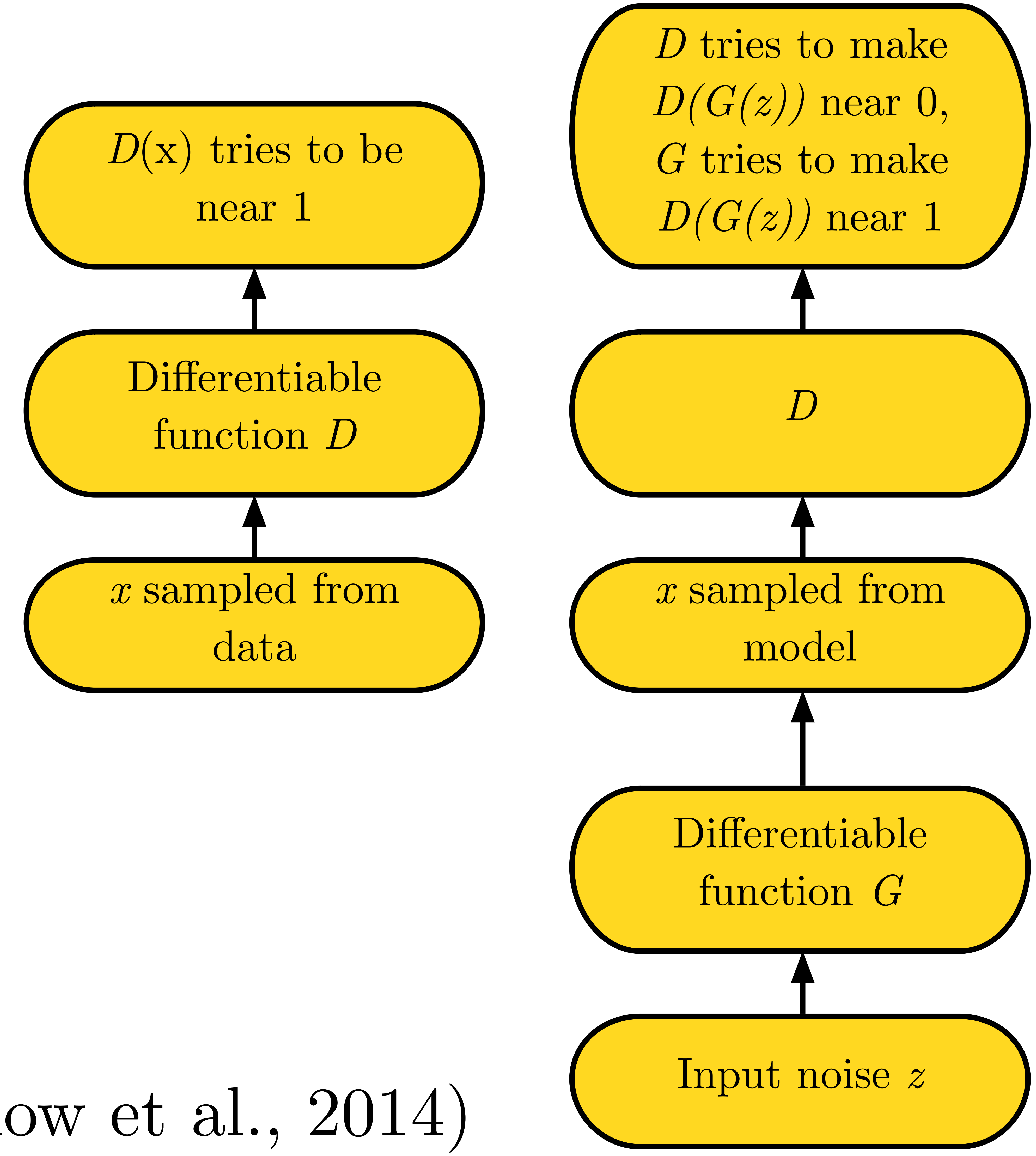
Generative Modeling: Sample Generation

Training Data
(CelebA)

Sample Generator
(Karras et al, 2017)

(Goodfellow 2018)

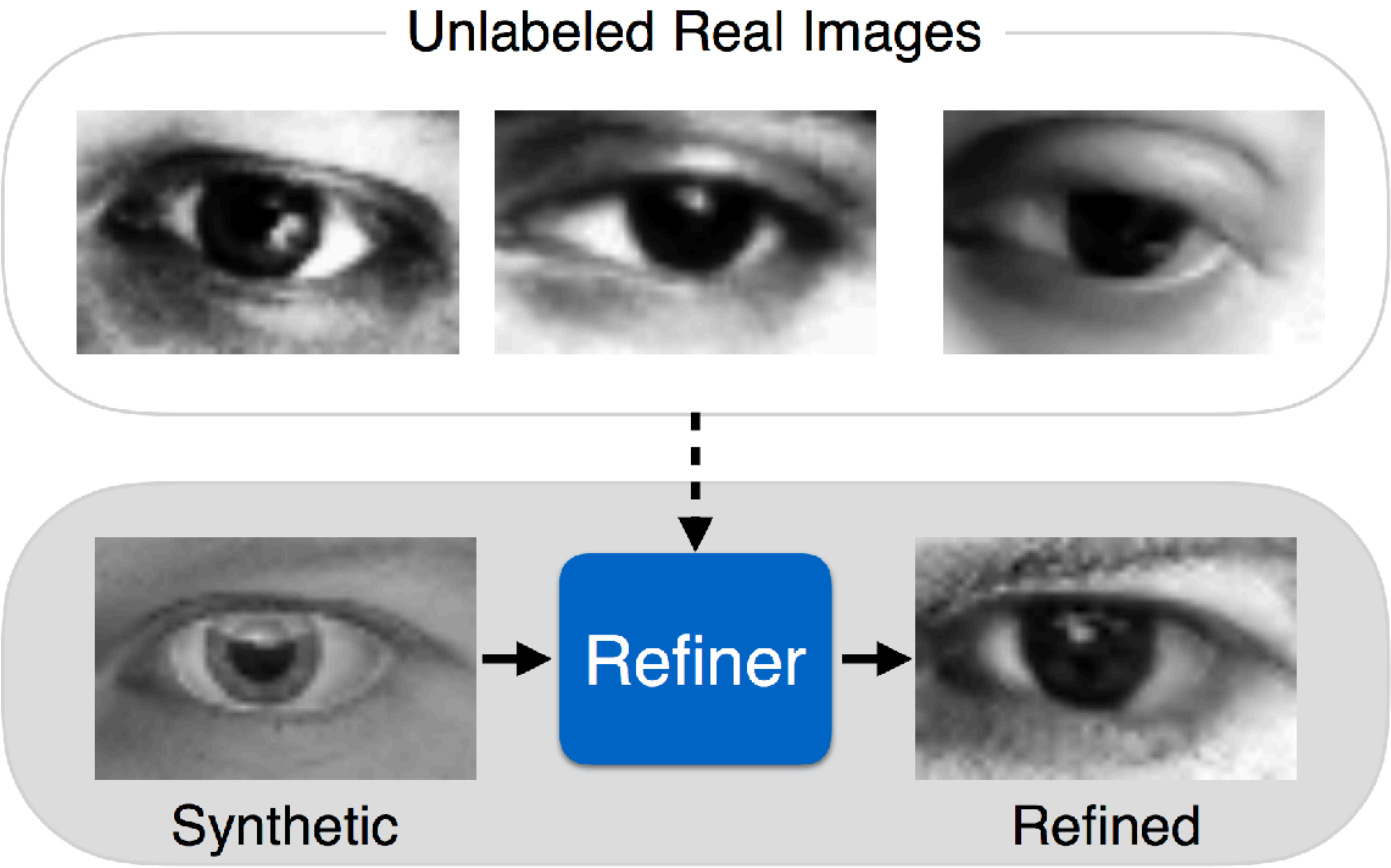
Adversarial Nets Framework



(Goodfellow et al., 2014)

(Goodfellow 2018)

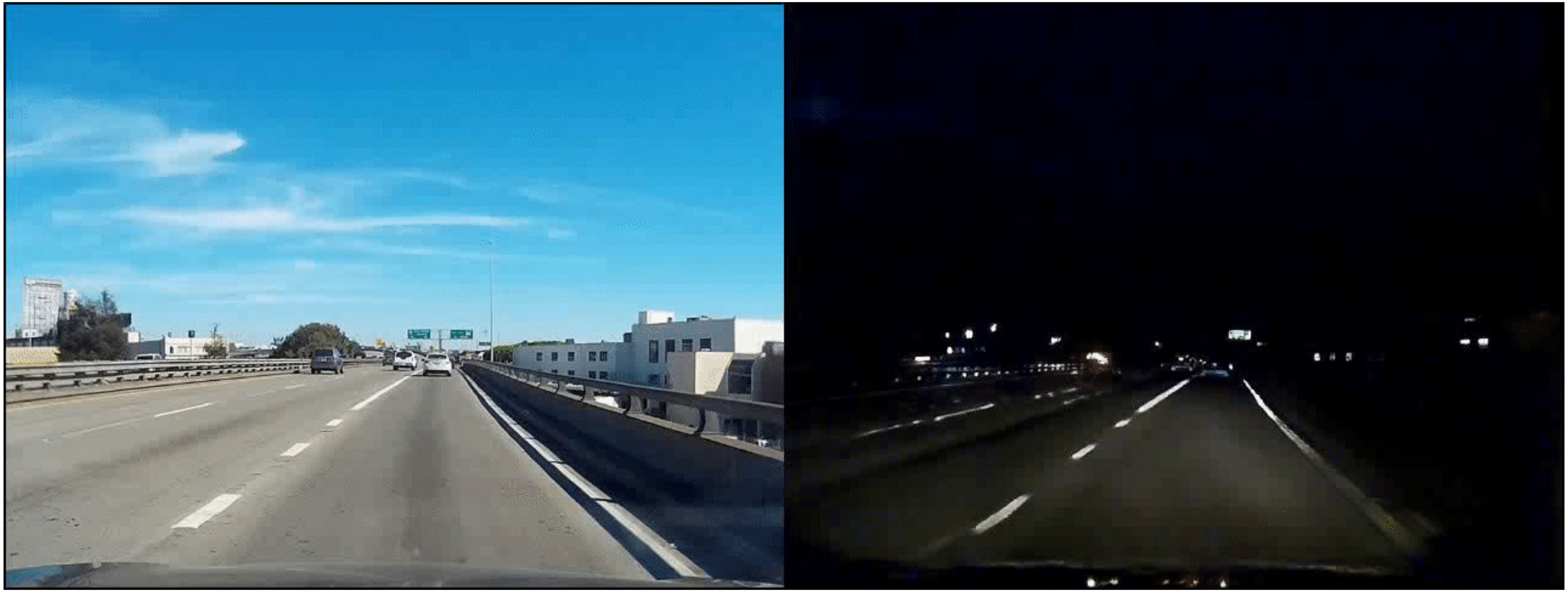
GANs for simulated training data



(Shrivastava et al., 2016)

Unsupervised Image-to-Image Translation

Day to night

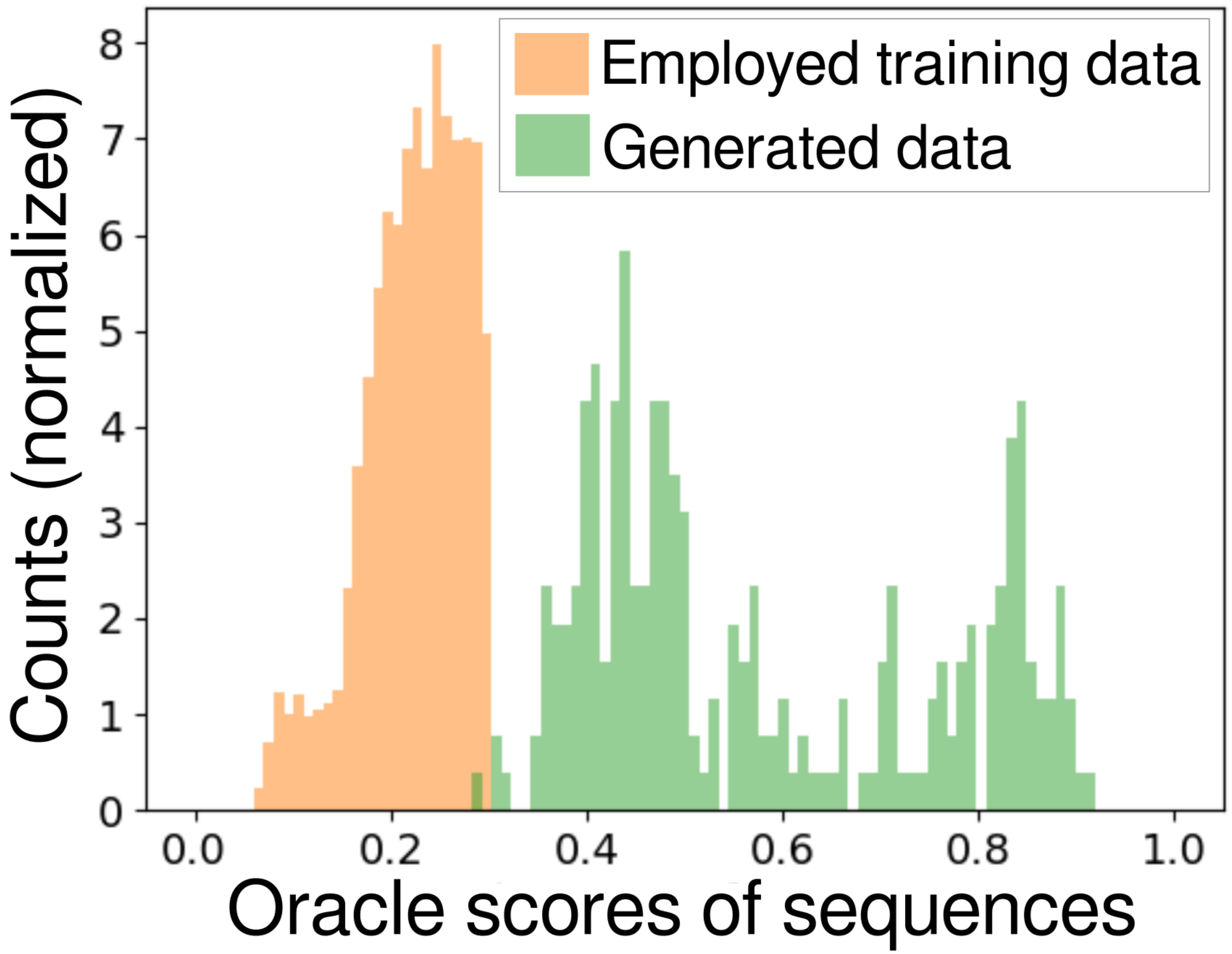


(Liu et al., 2017)

CycleGAN

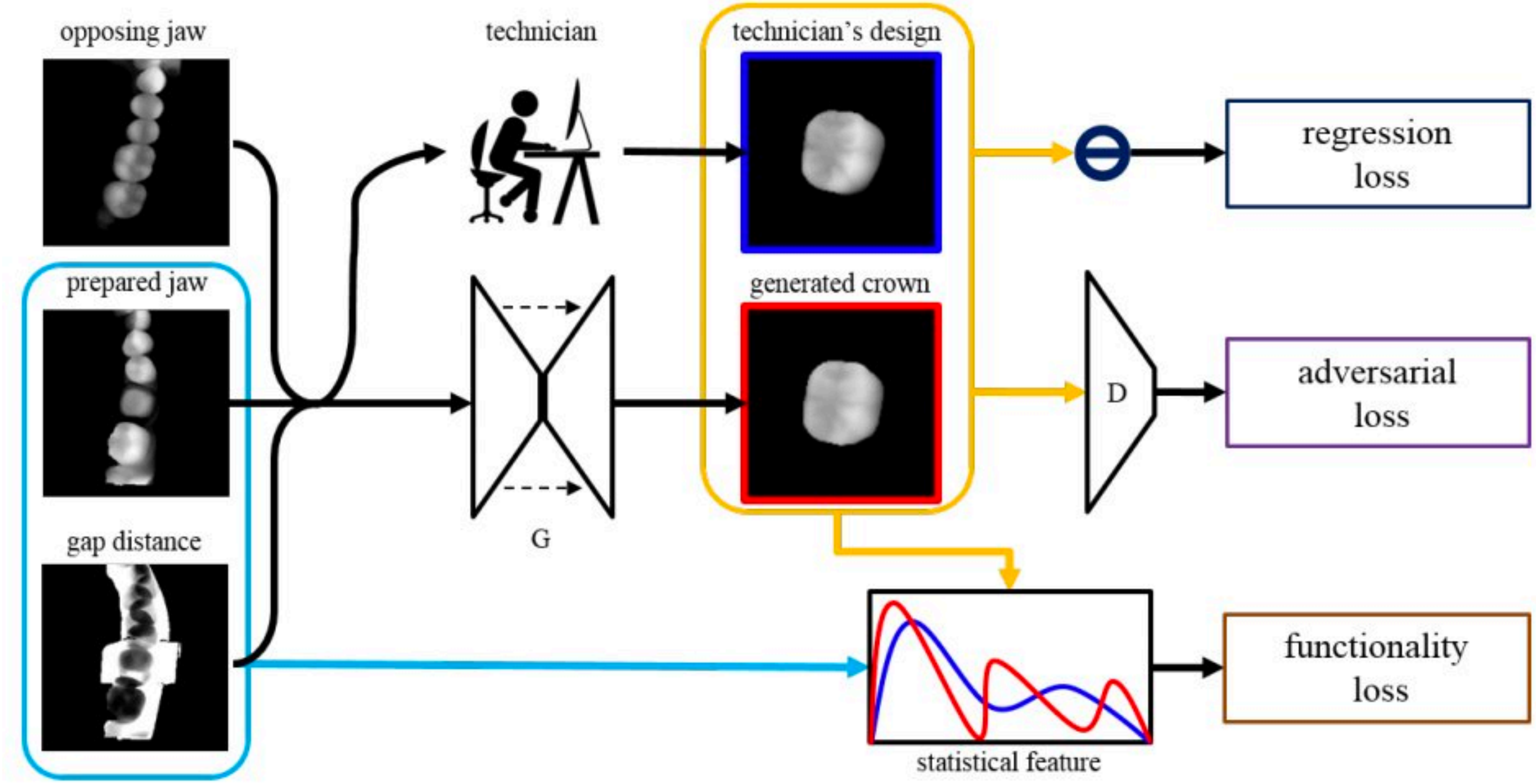
(Zhu et al., 2017)

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

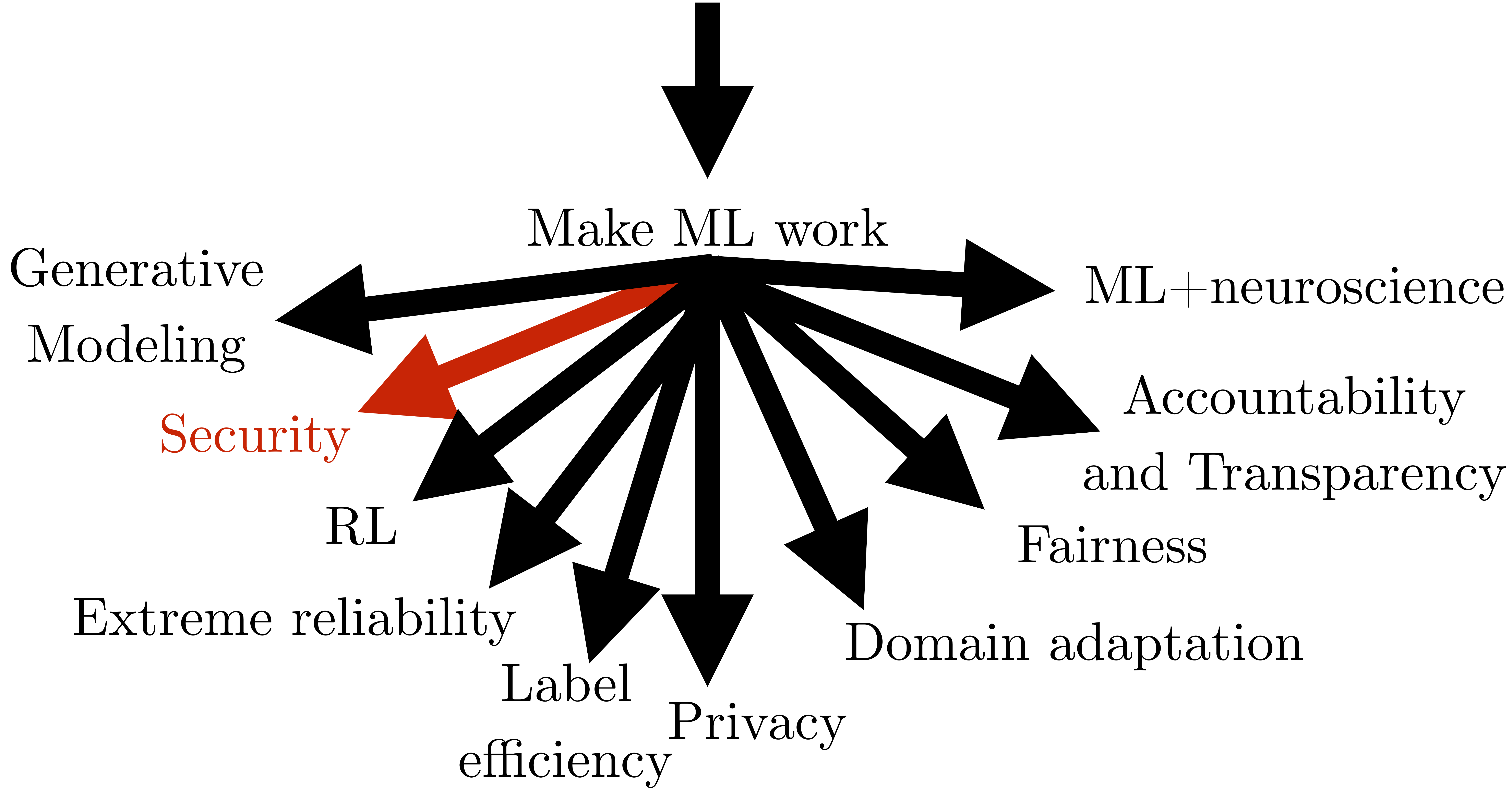
Indigo Bunting

Stone Wall

Saint Bernard

(Zhang et al, 2018)

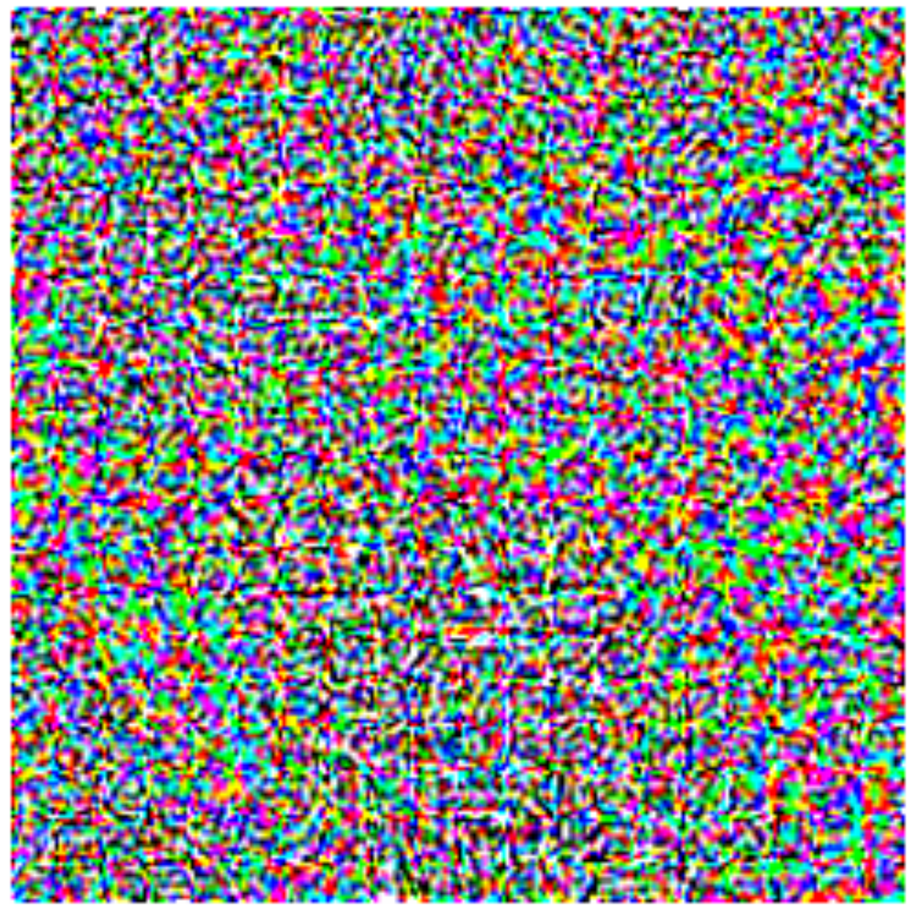
A Cambrian Explosion of Machine Learning Research Topics



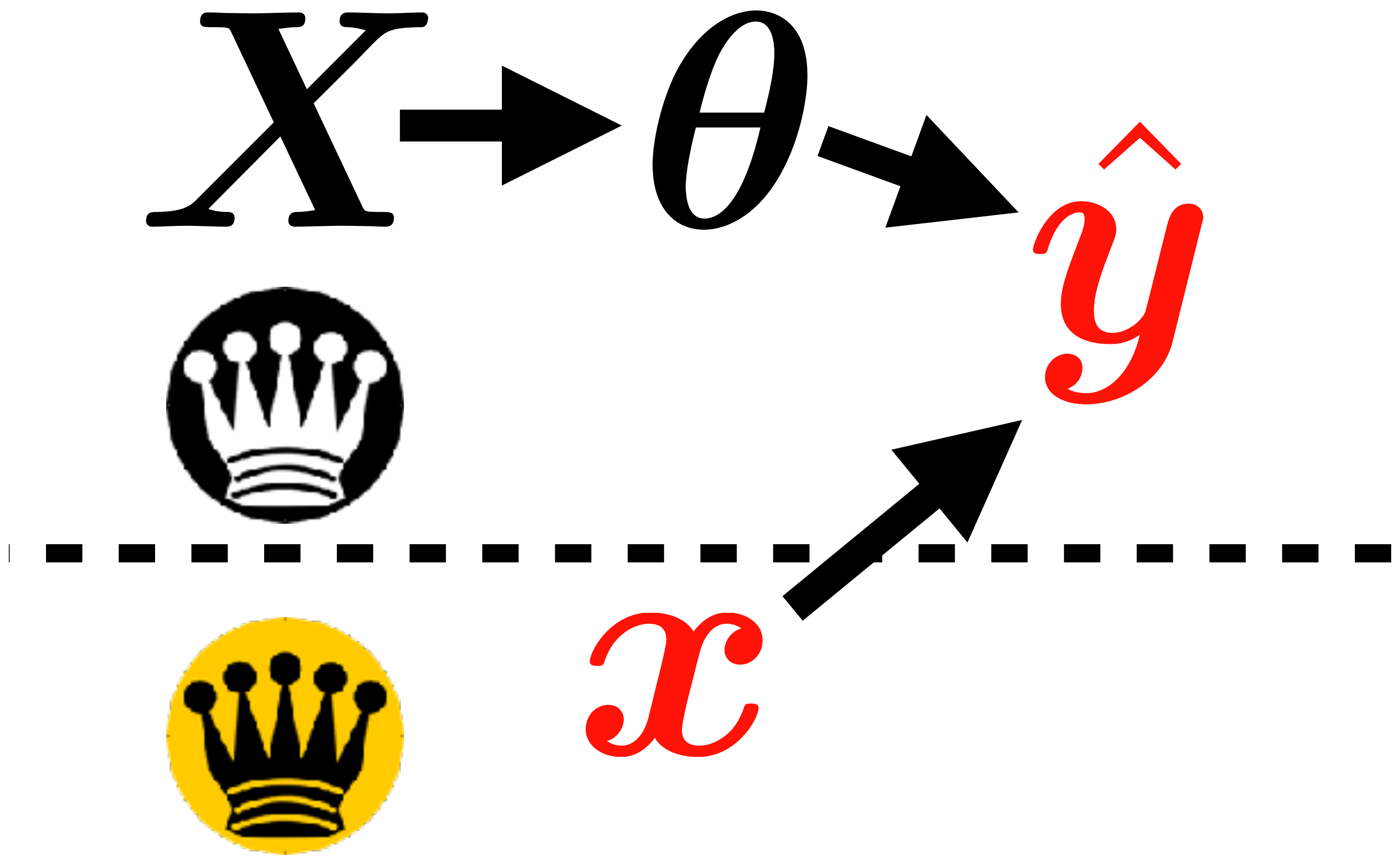
(Goodfellow 2017)

Adversarial Examples

$+ .007 \times$

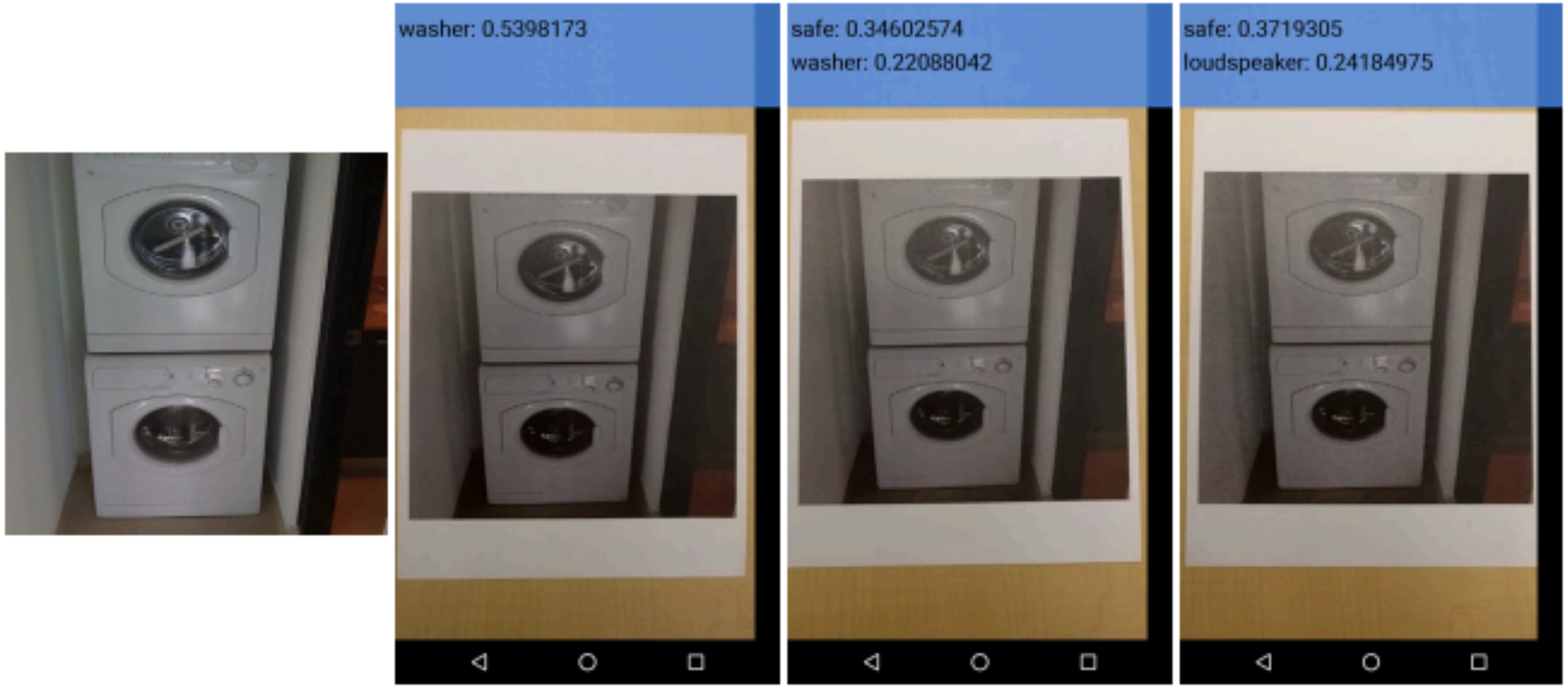


$=$



(Goodfellow 2017)

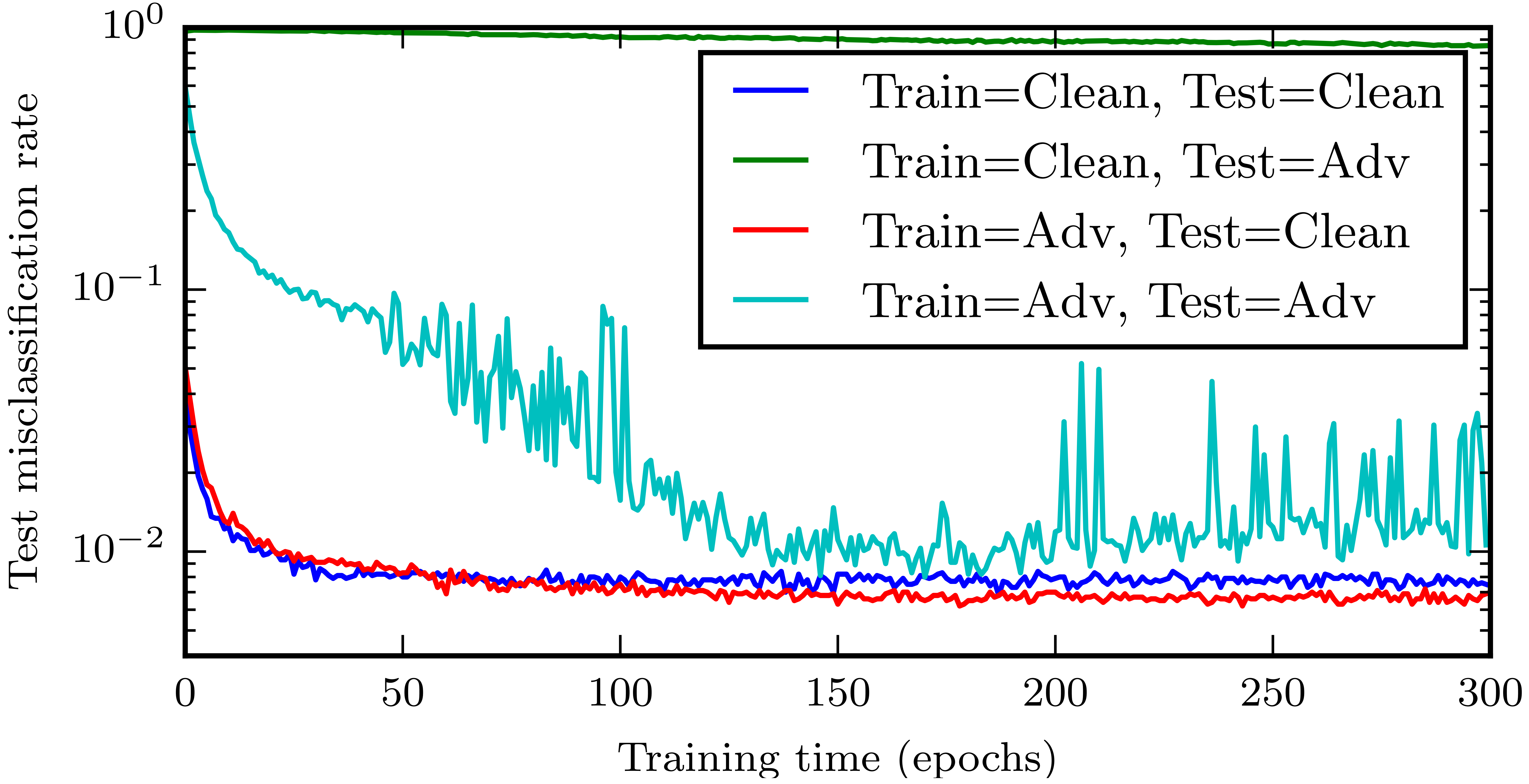
Adversarial Examples in the Physical World



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

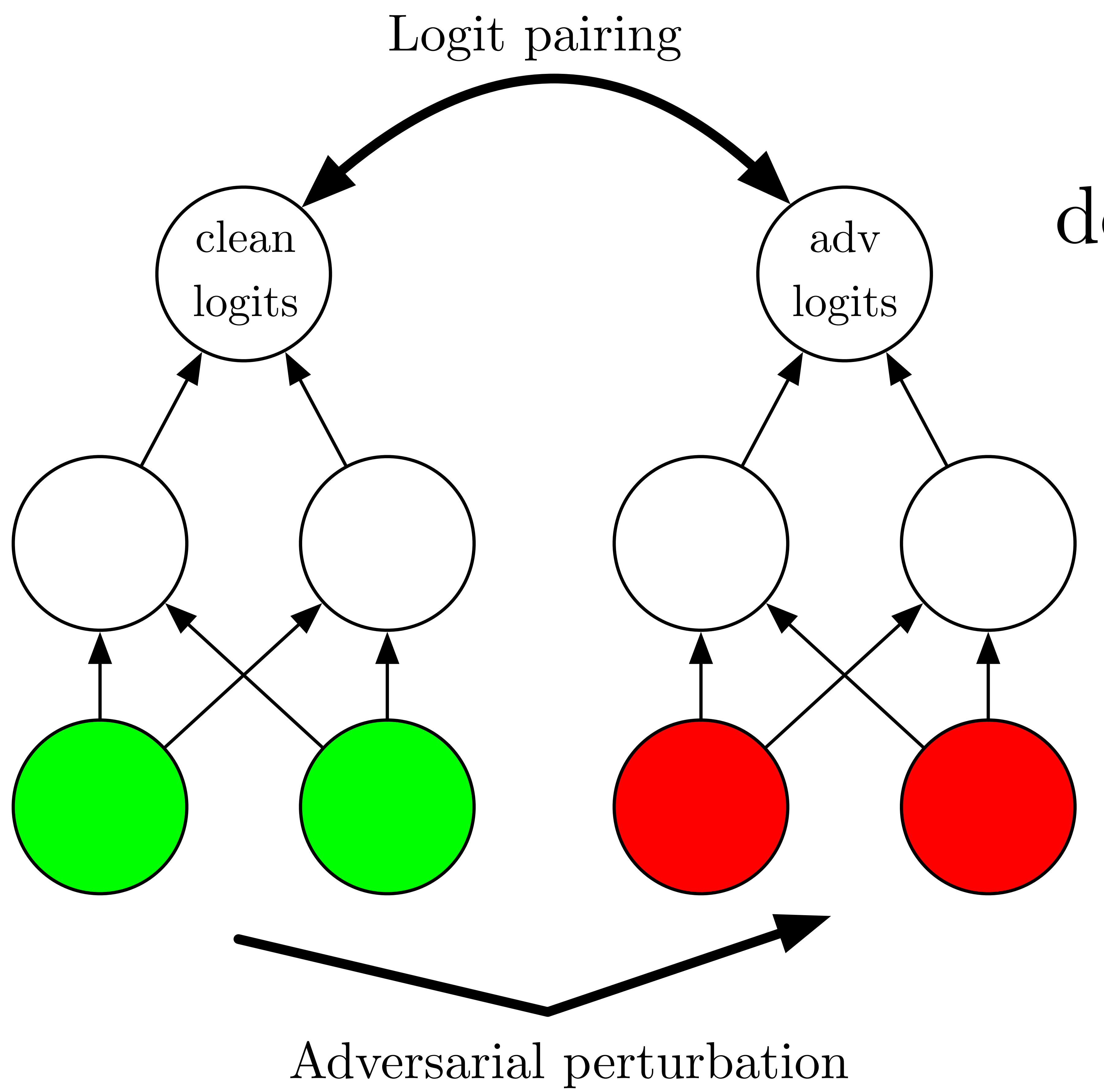
(Kurakin et al, 2016)

Training on Adversarial Examples



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

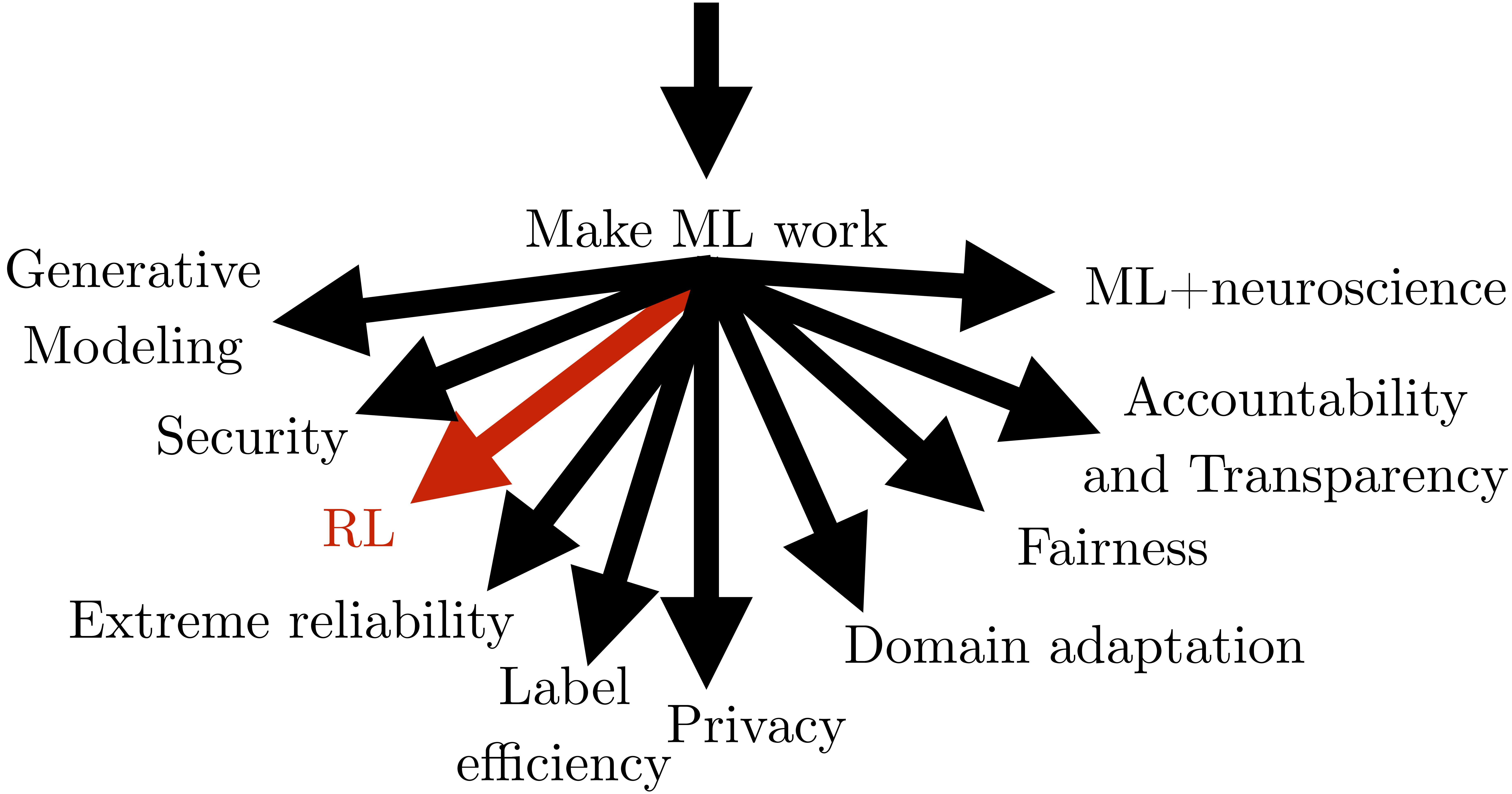
Adversarial Logit Pairing



State of the art defense on ImageNet

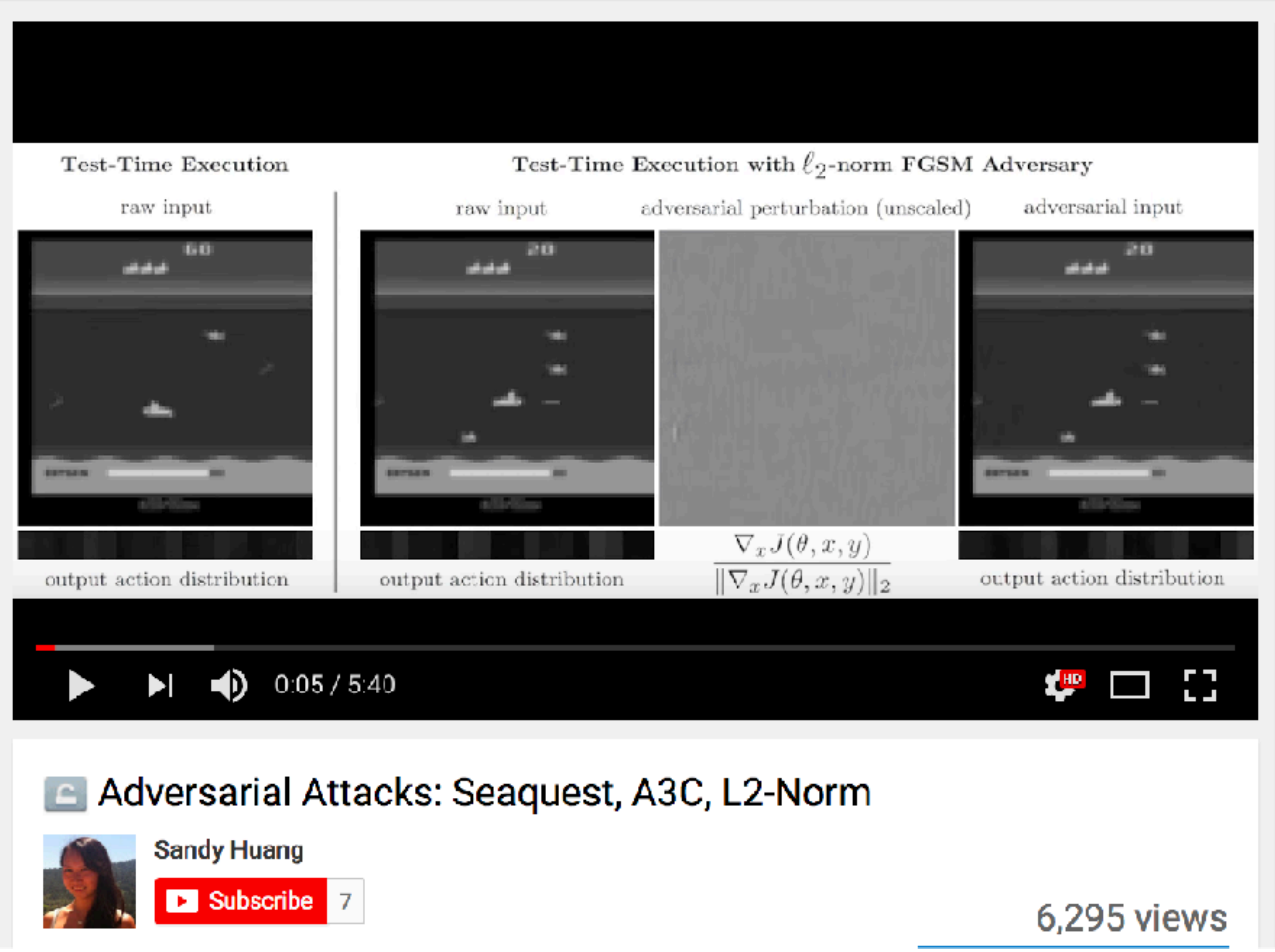
(Kannan et al, 2018)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2017)

Adversarial Examples for RL



(Huang et al., 2017)

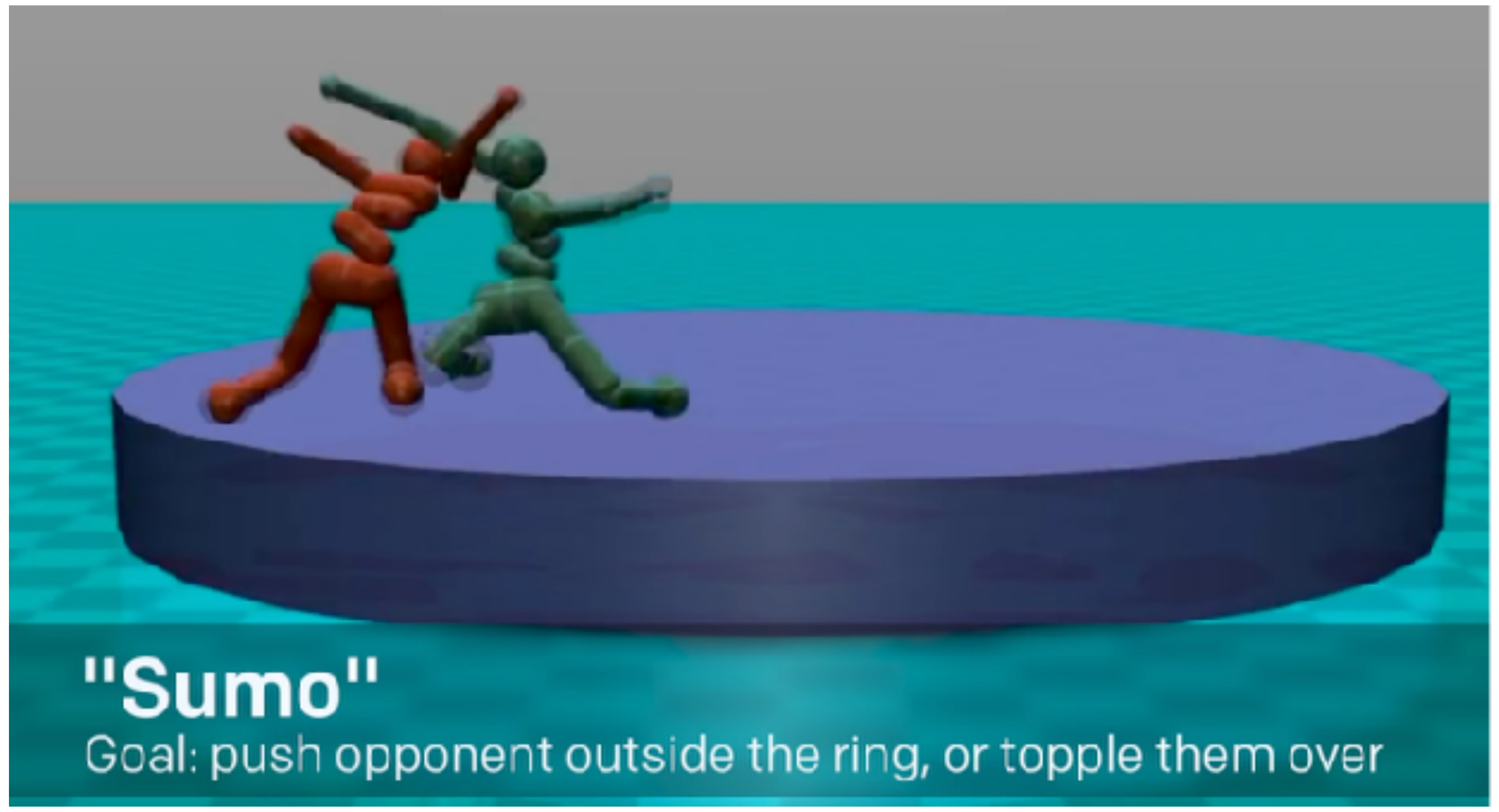
(Goodfellow 2017)

Self-Play

1959: Arthur Samuel's checkers agent

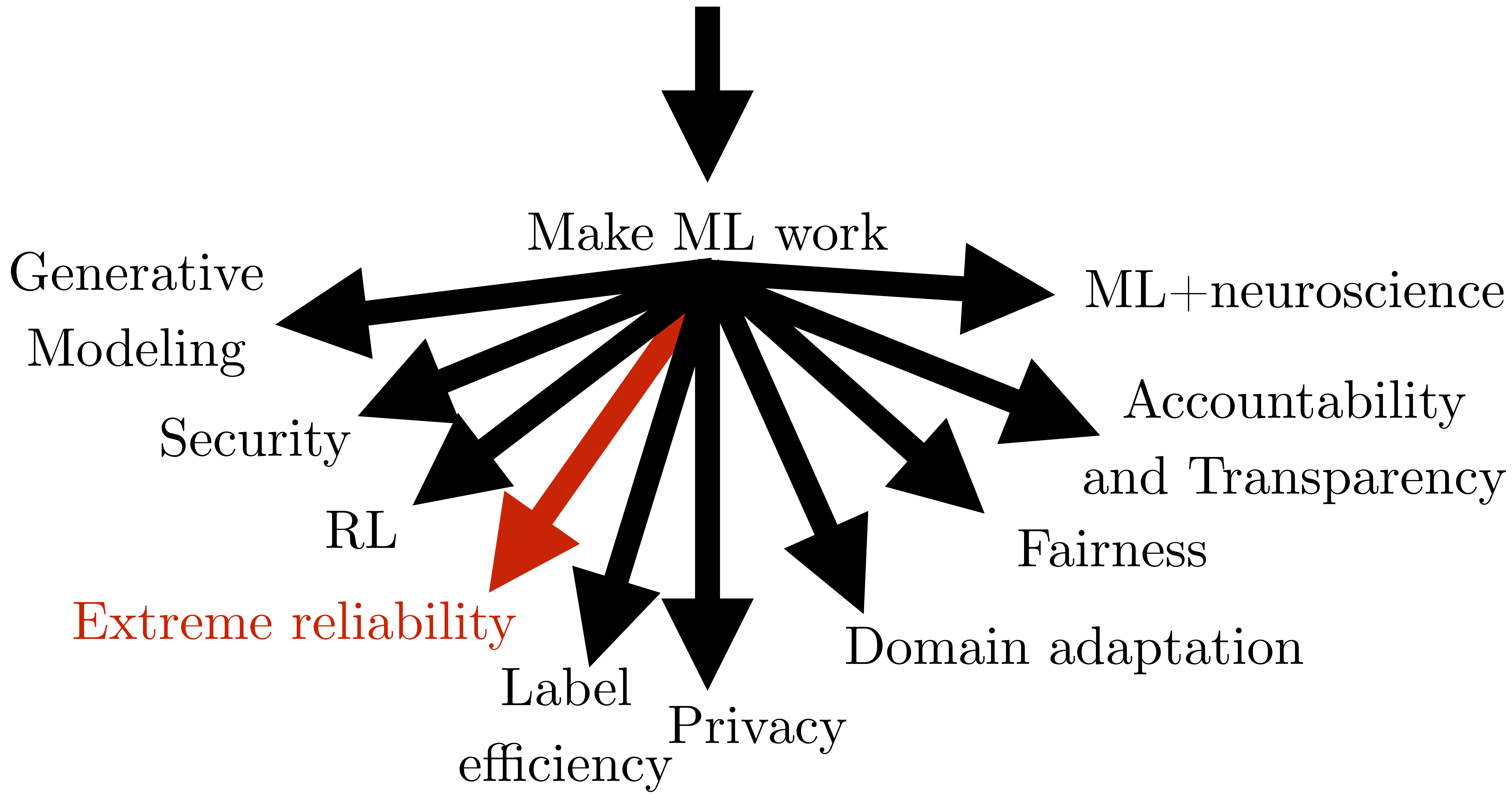
(Silver et al, 2017)

(OpenAI, 2017)



(Bansal et al, 2017)

A Cambrian Explosion of Machine Learning Research Topics

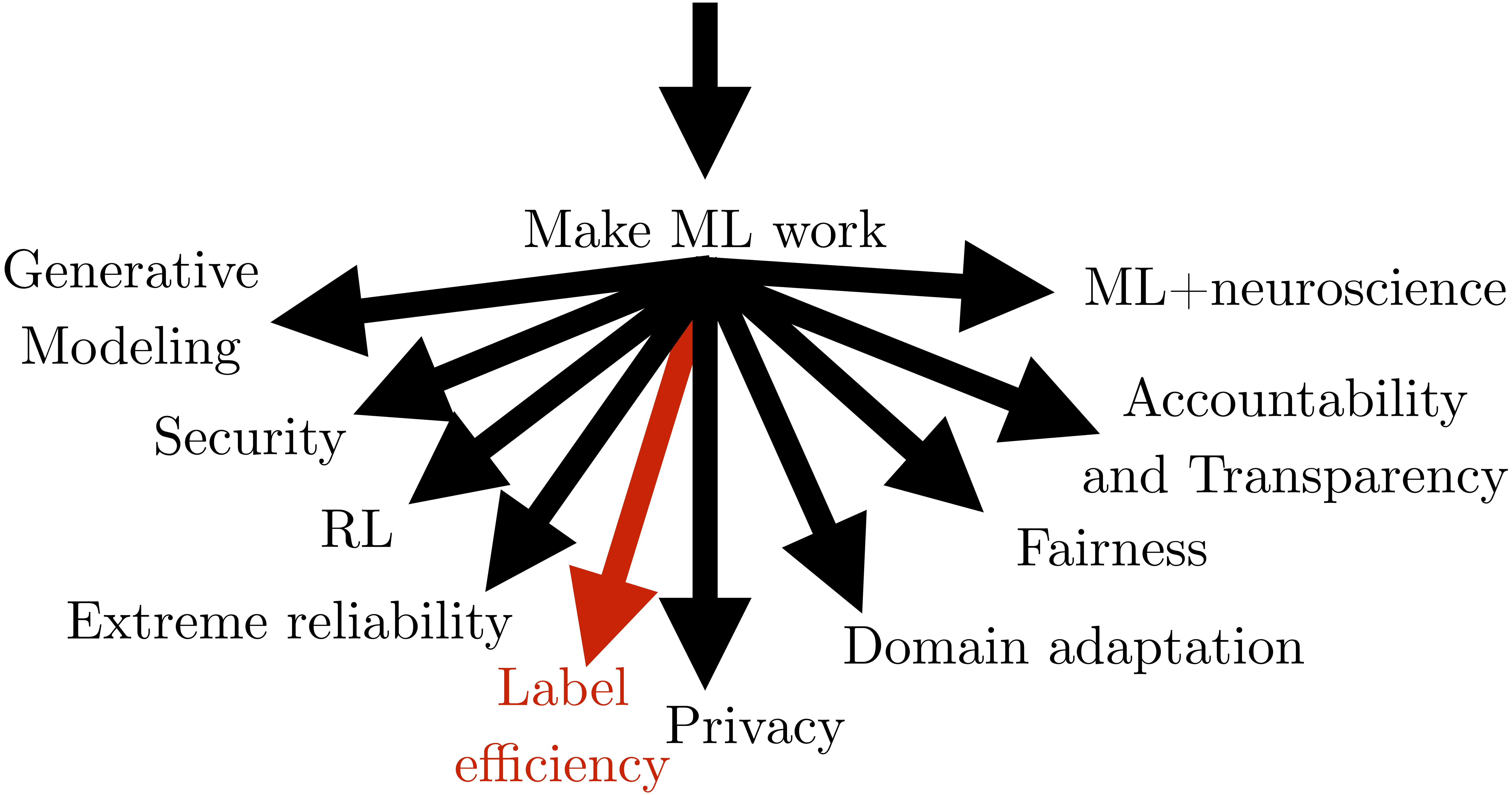


(Goodfellow 2017)

Extreme Reliability

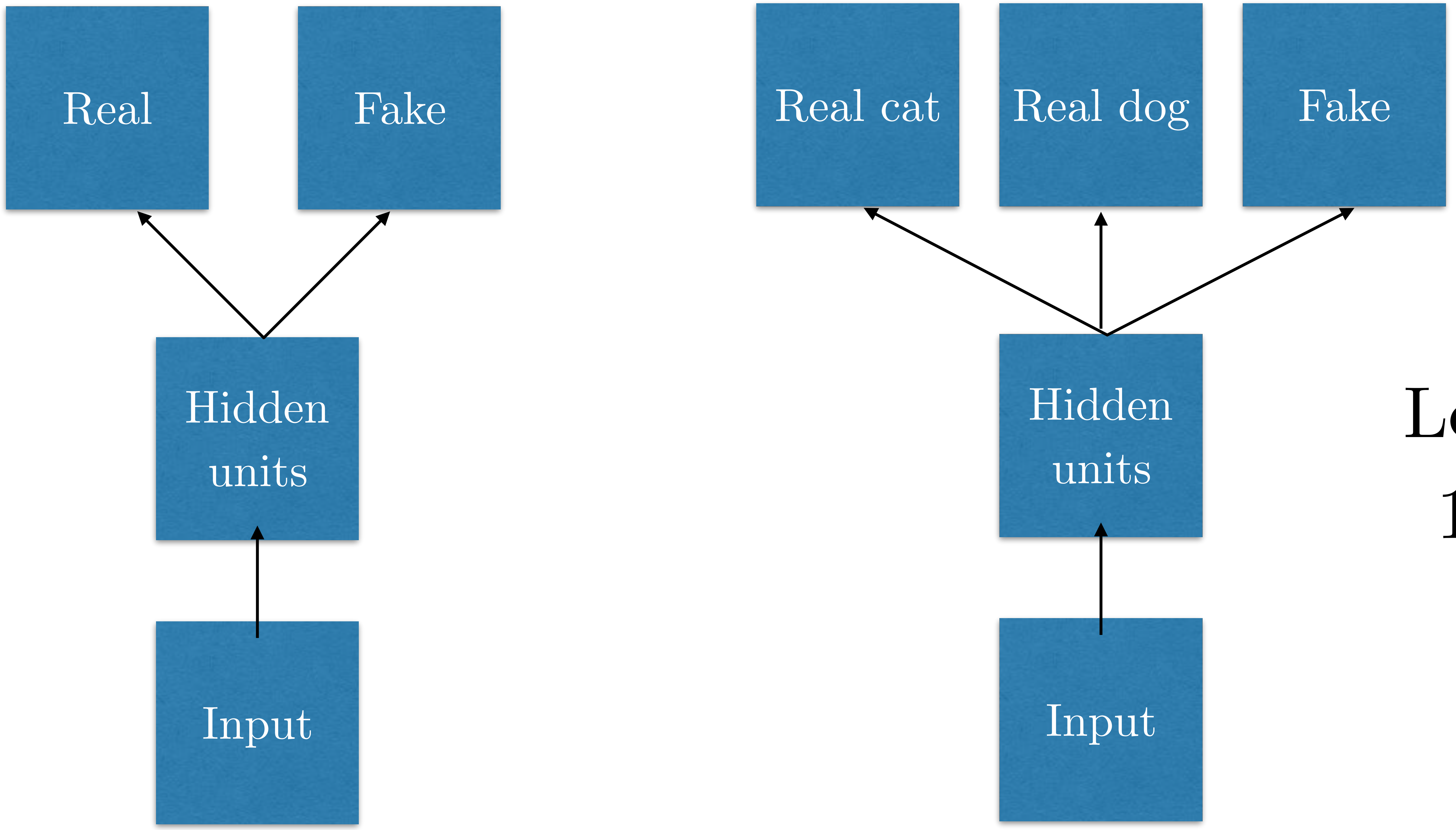
- We want extreme reliability for
  - Autonomous vehicles
  - Air traffic control
  - Surgery robots
  - Medical diagnosis, etc.
- Adversarial machine learning research techniques can help with this
  - Katz et al 2017: verification system, applied to air traffic control

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2017)

Supervised Discriminator for Semi-Supervised Learning

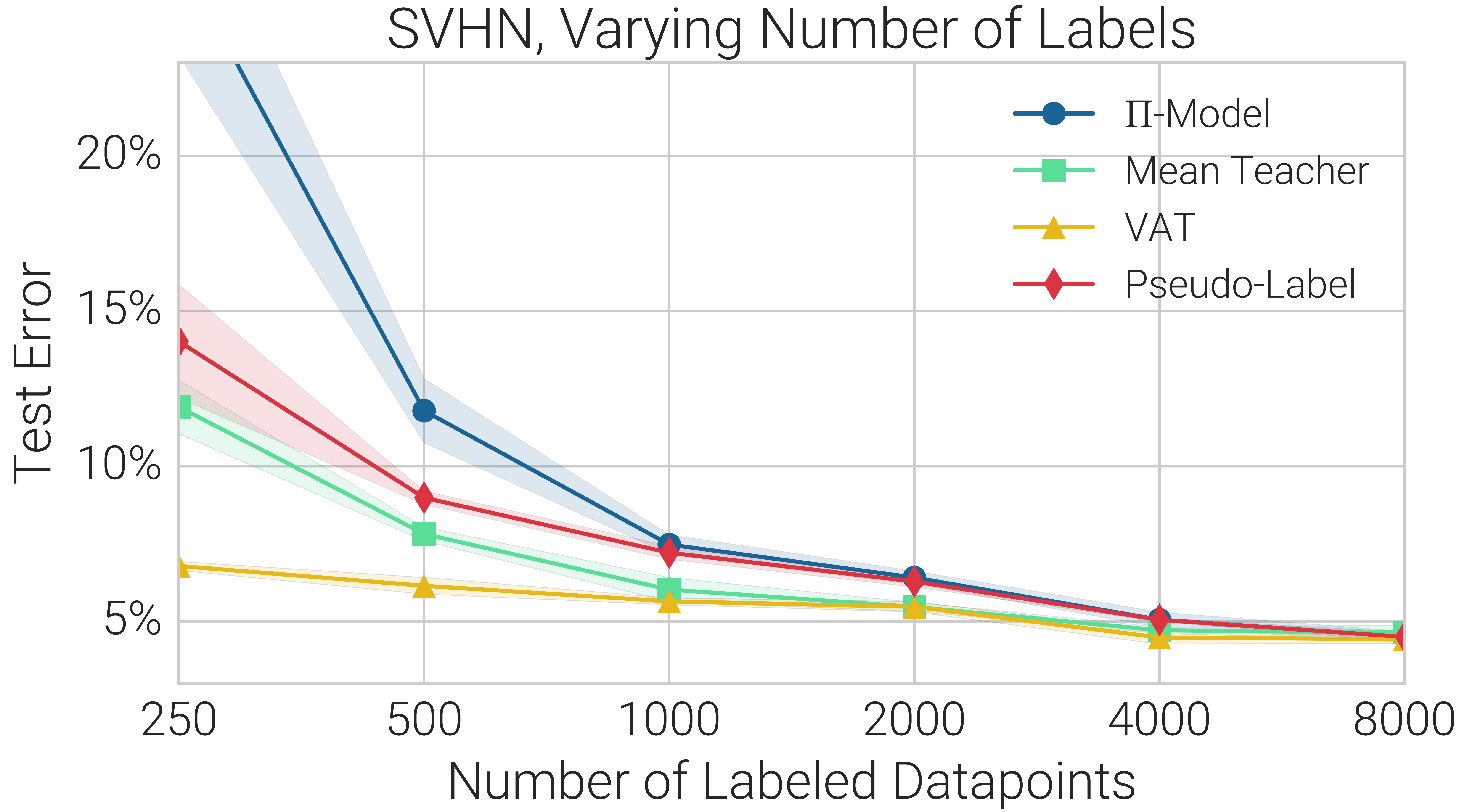


Learn to read with 100 labels rather than 60,000

(Odena 2016, Salimans et al 2016)

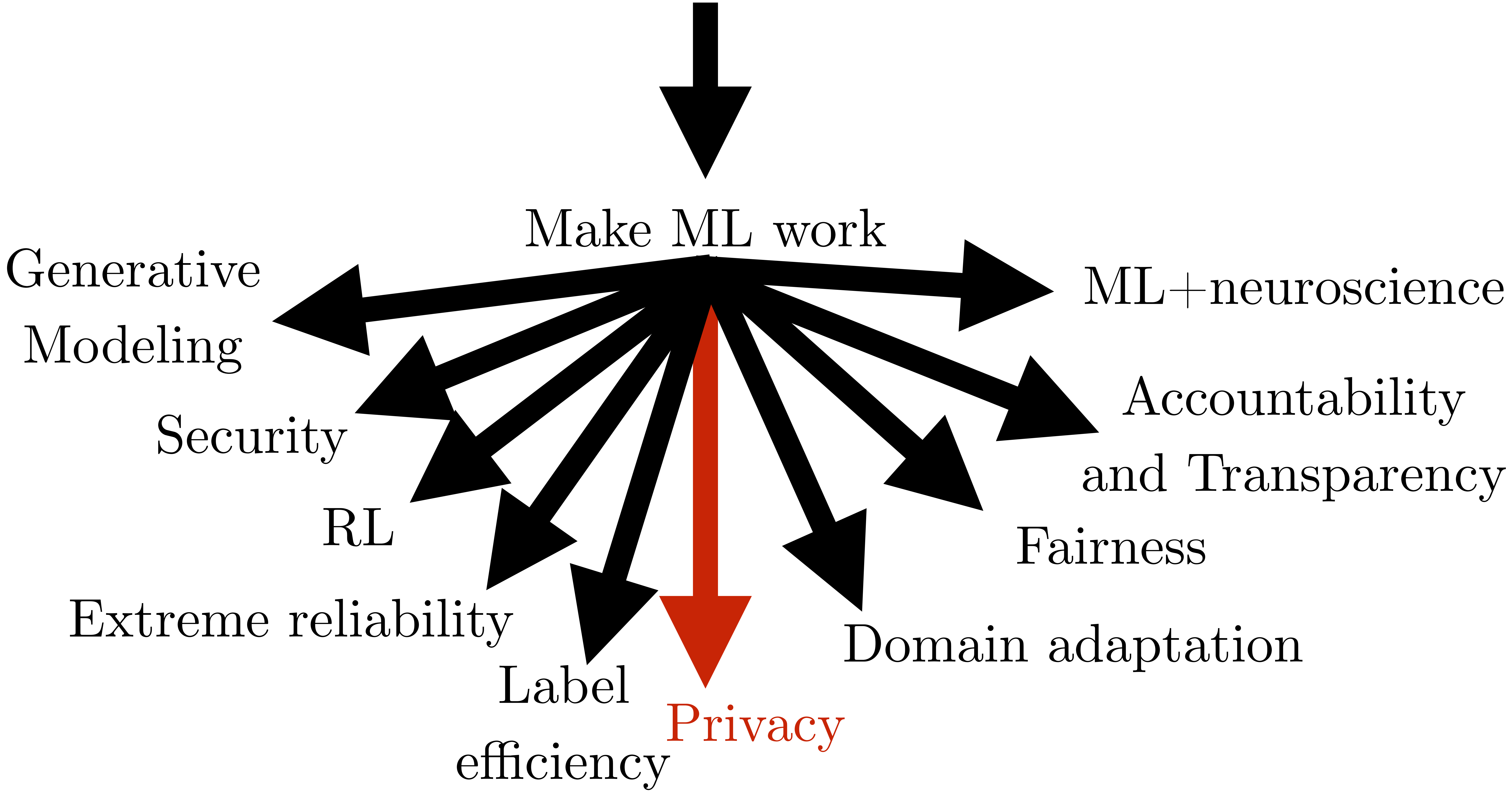
Virtual Adversarial Training

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



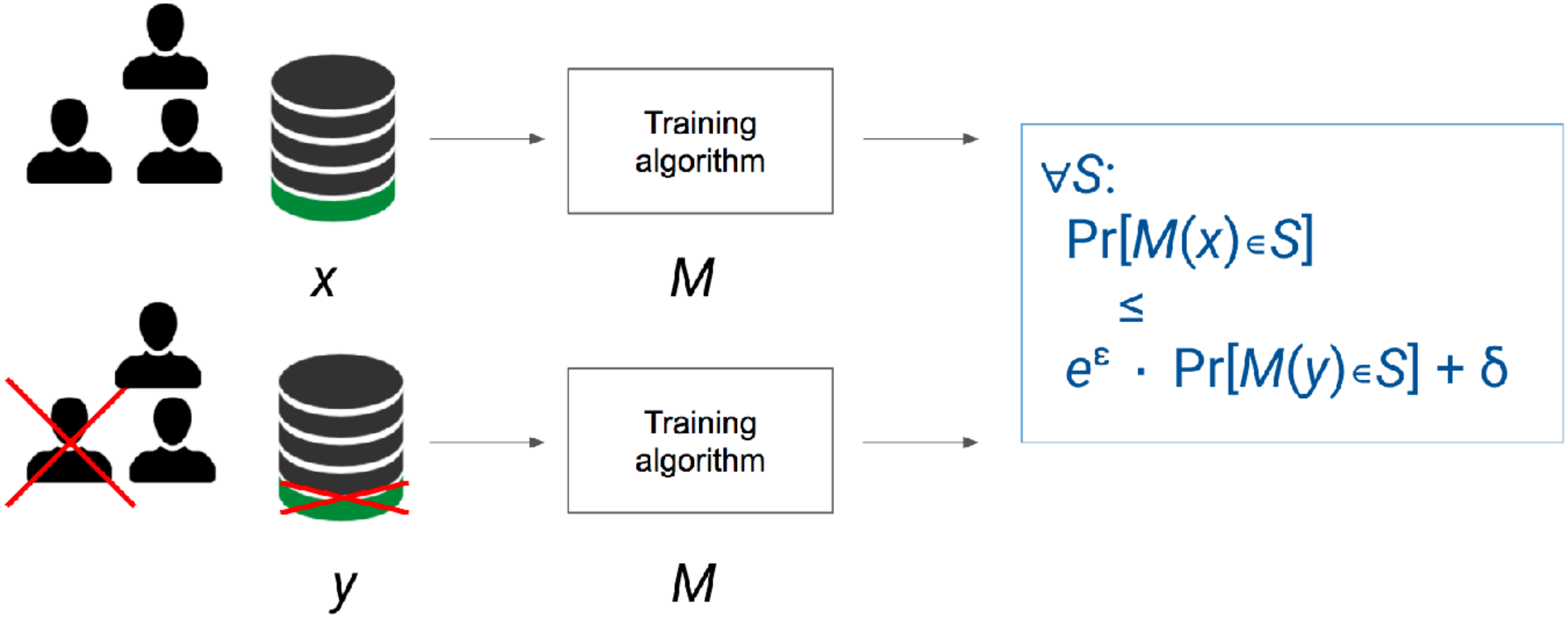
(Goodfellow 2017)

Privacy of training data

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

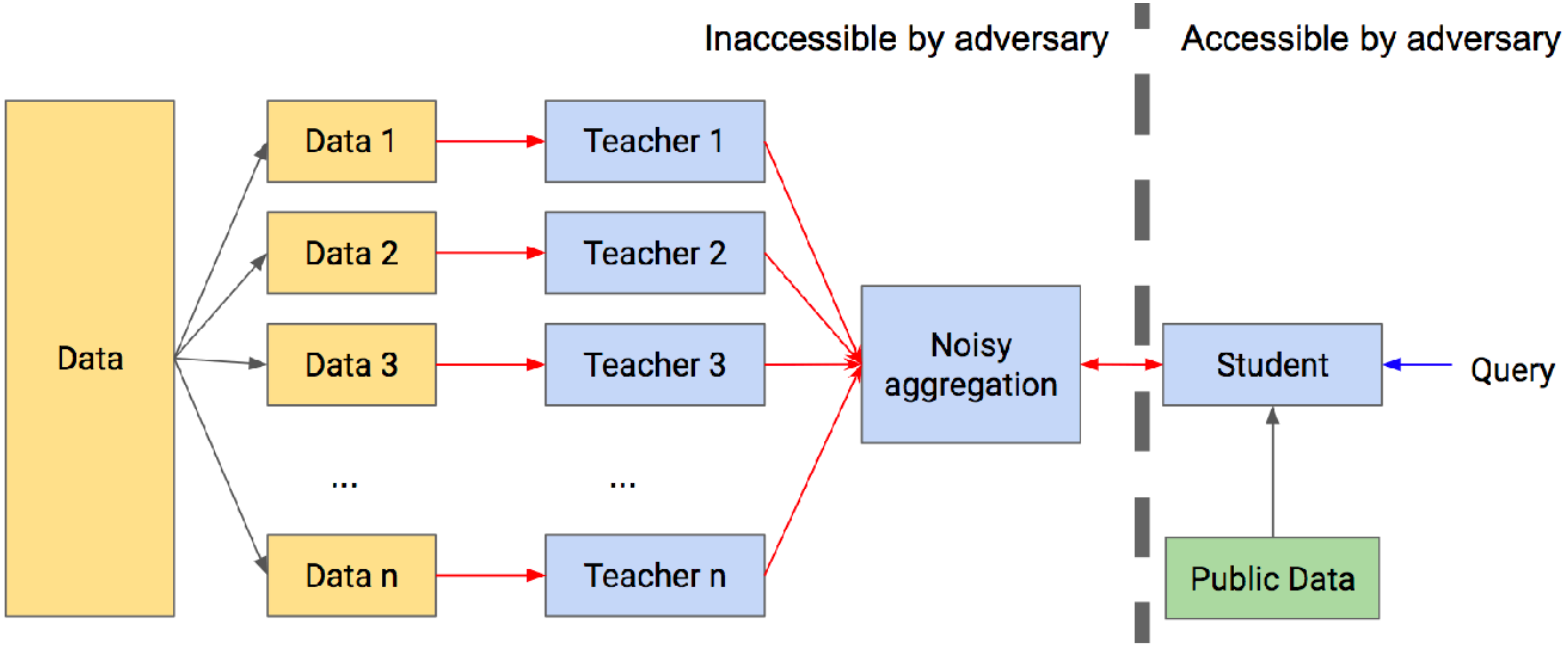
(Goodfellow 2018)

Defining (ε, δ)-Differential Privacy



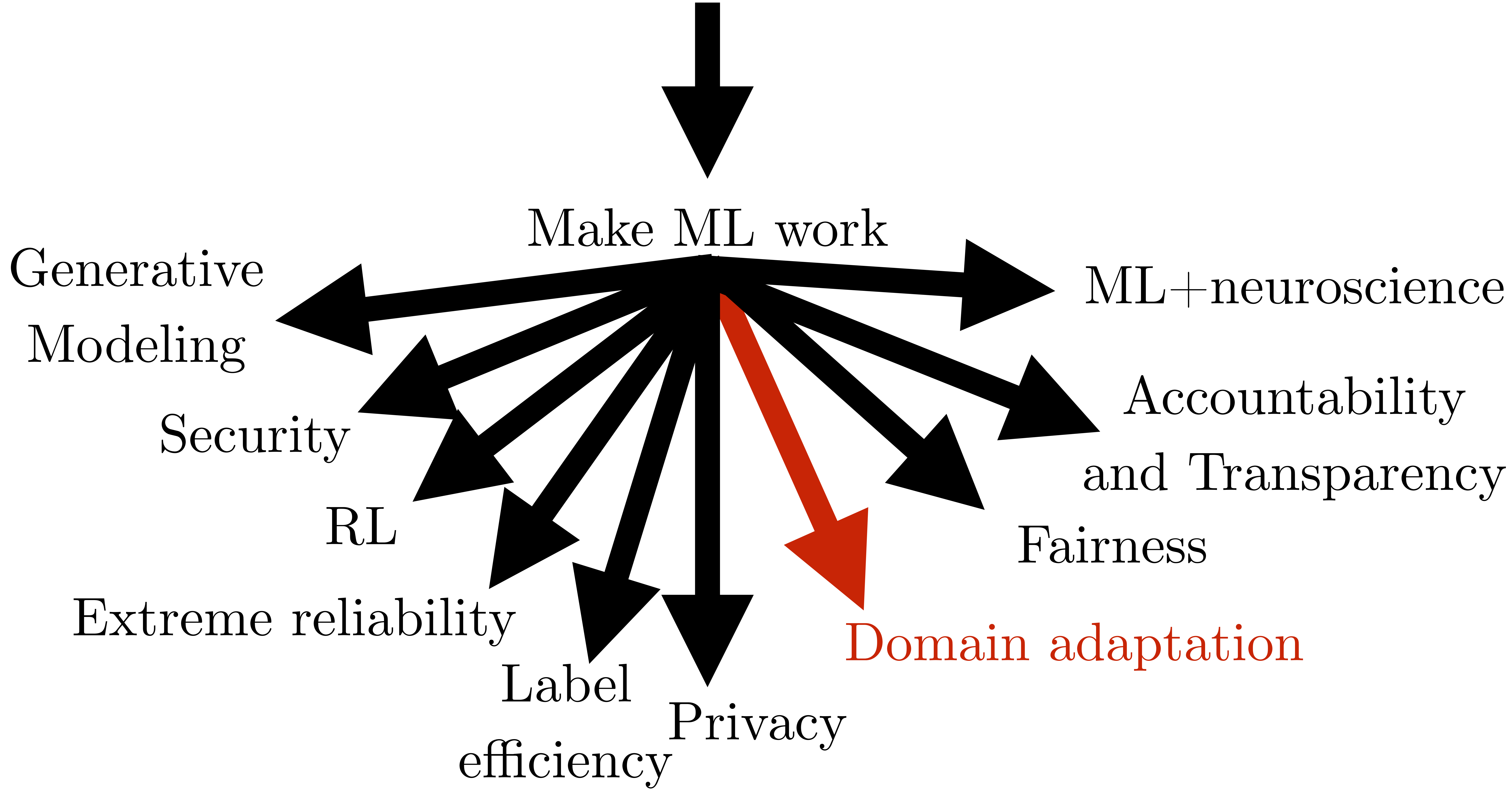
(Abadi 2017)

Private Aggregation of Teacher Ensembles



(Papernot et al 2016)

A Cambrian Explosion of Machine Learning Research Topics



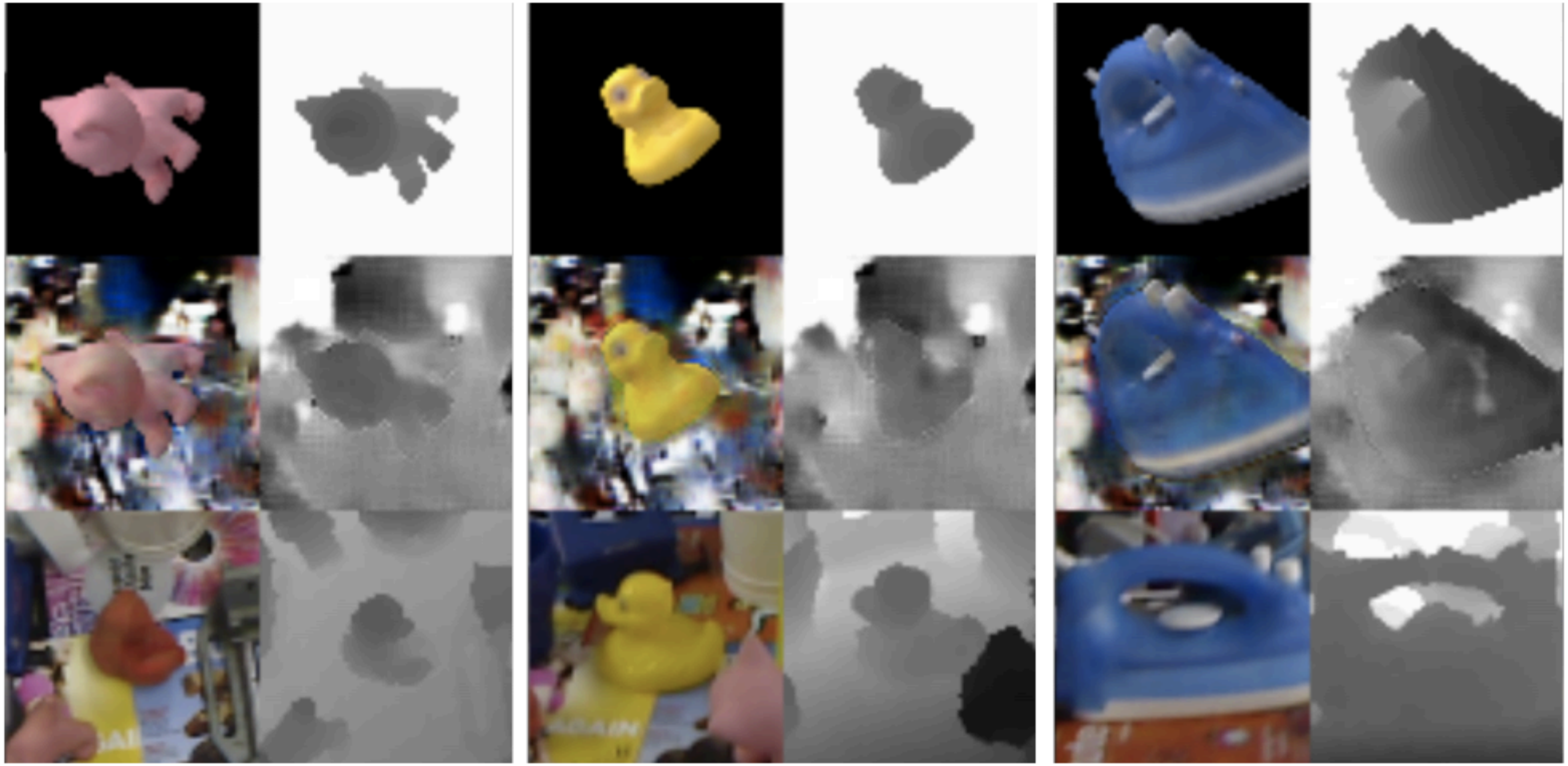
(Goodfellow 2017)

Domain Adaptation

- Domain Adversarial Networks (Ganin et al, 2015)

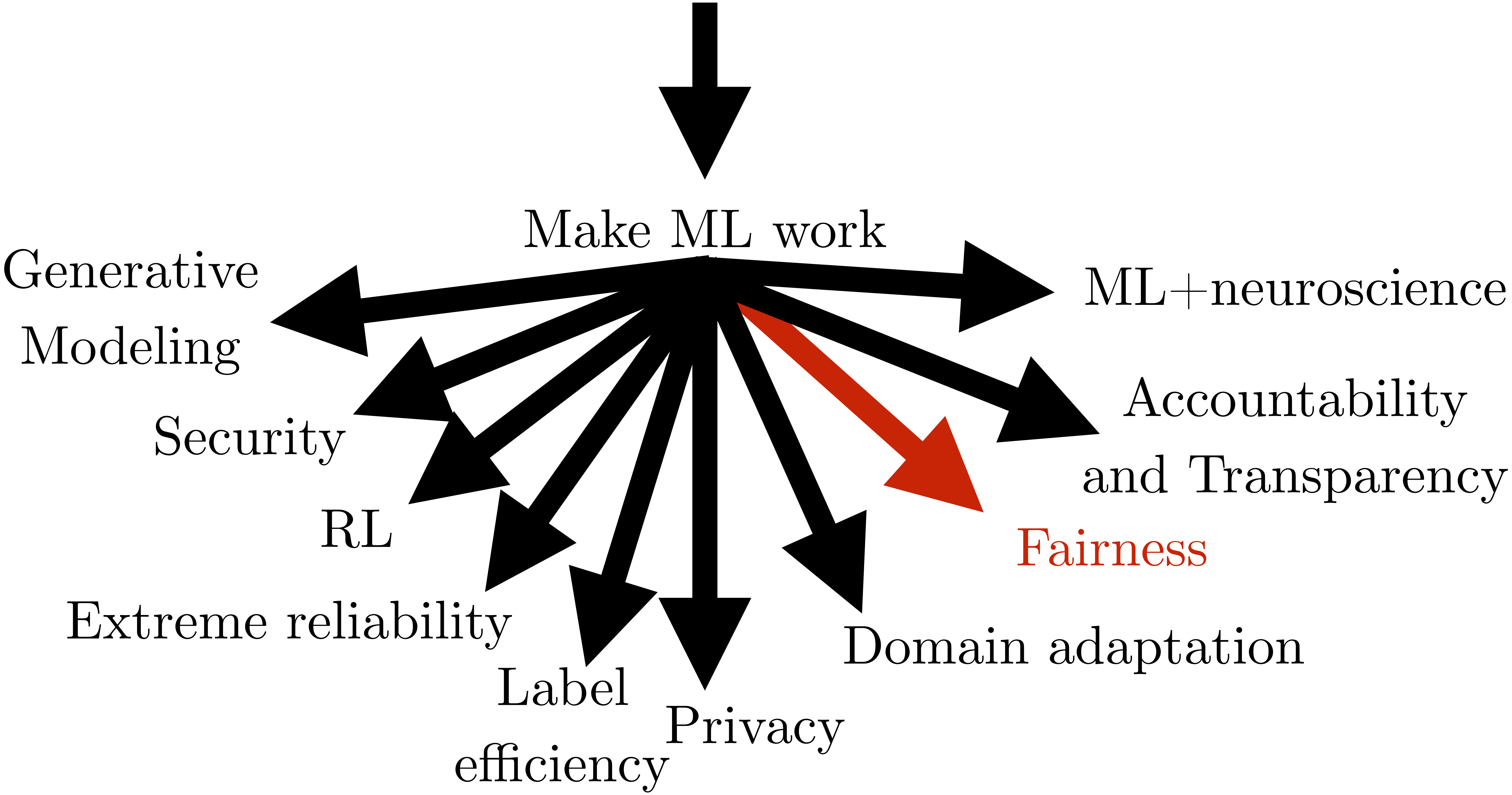
- Professor forcing (Lamb et al, 2016): Domain-Adversarial learning in RNN hidden state

GANs for domain adaptation



(Bousmalis et al., 2016)

A Cambrian Explosion of Machine Learning Research Topics

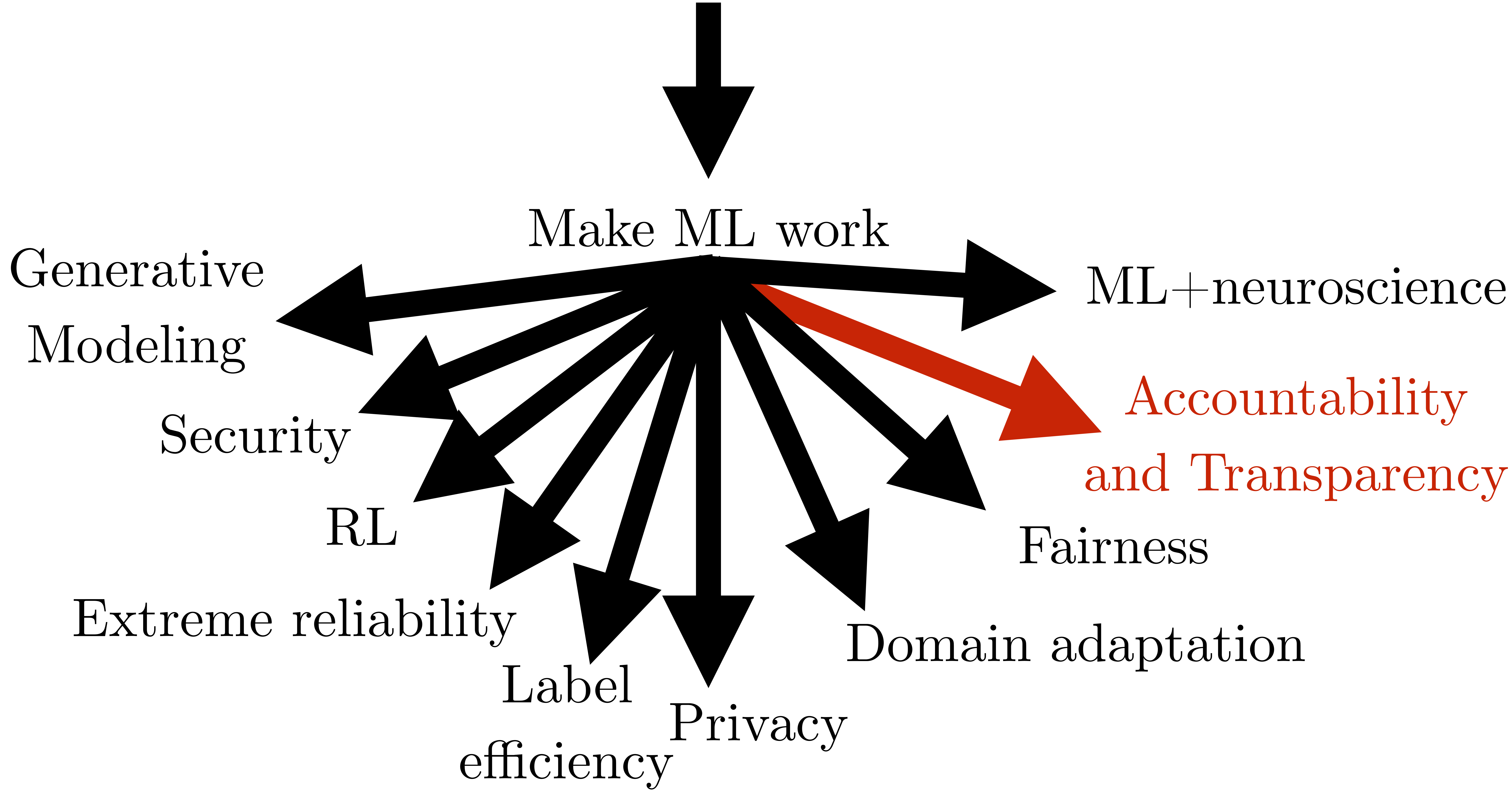


(Goodfellow 2017)

Adversarially Learned Fair Representations

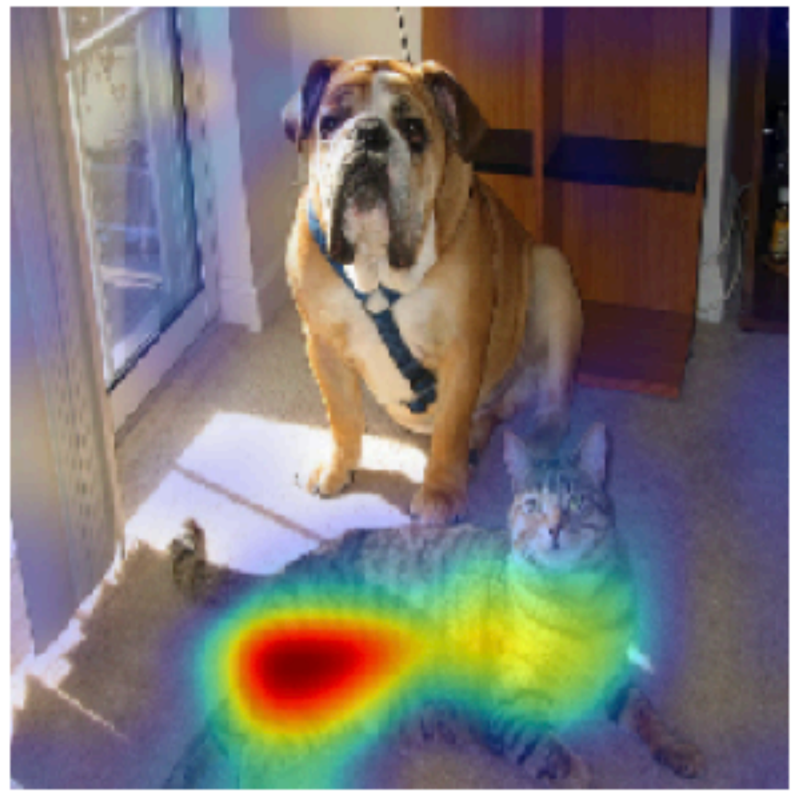
- Edwards and Storkey 2015
- Learn representations that are useful for classification
- An adversary tries to recover a sensitive variable S from the representation. Primary learner tries to make S impossible to recover
- Final decision does not depend on S

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2017)

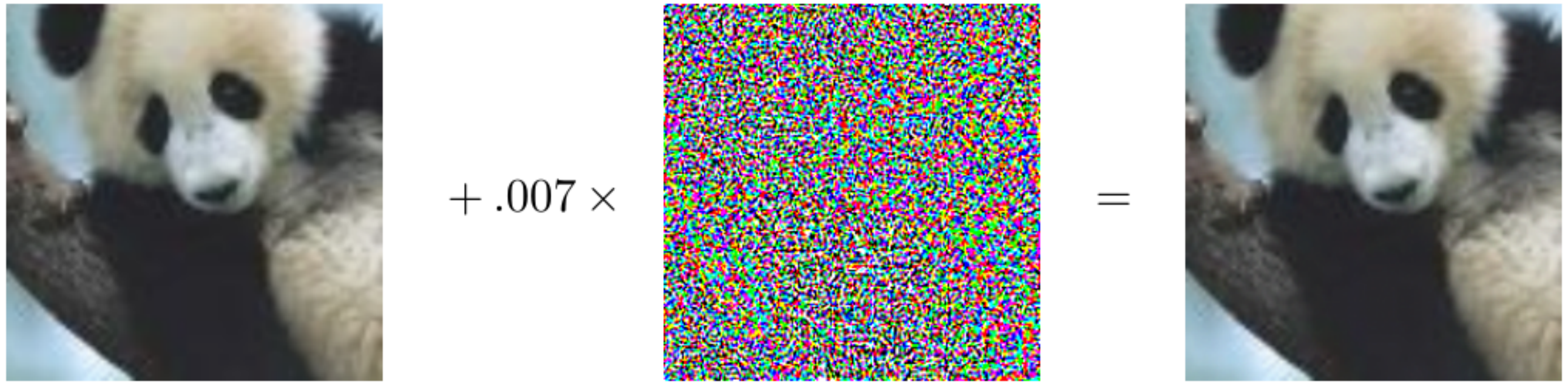
How do machine learning models work?



(c) Grad-CAM 'Cat'

(i) Grad-CAM 'Dog'

(Selvaraju et al, 2016)

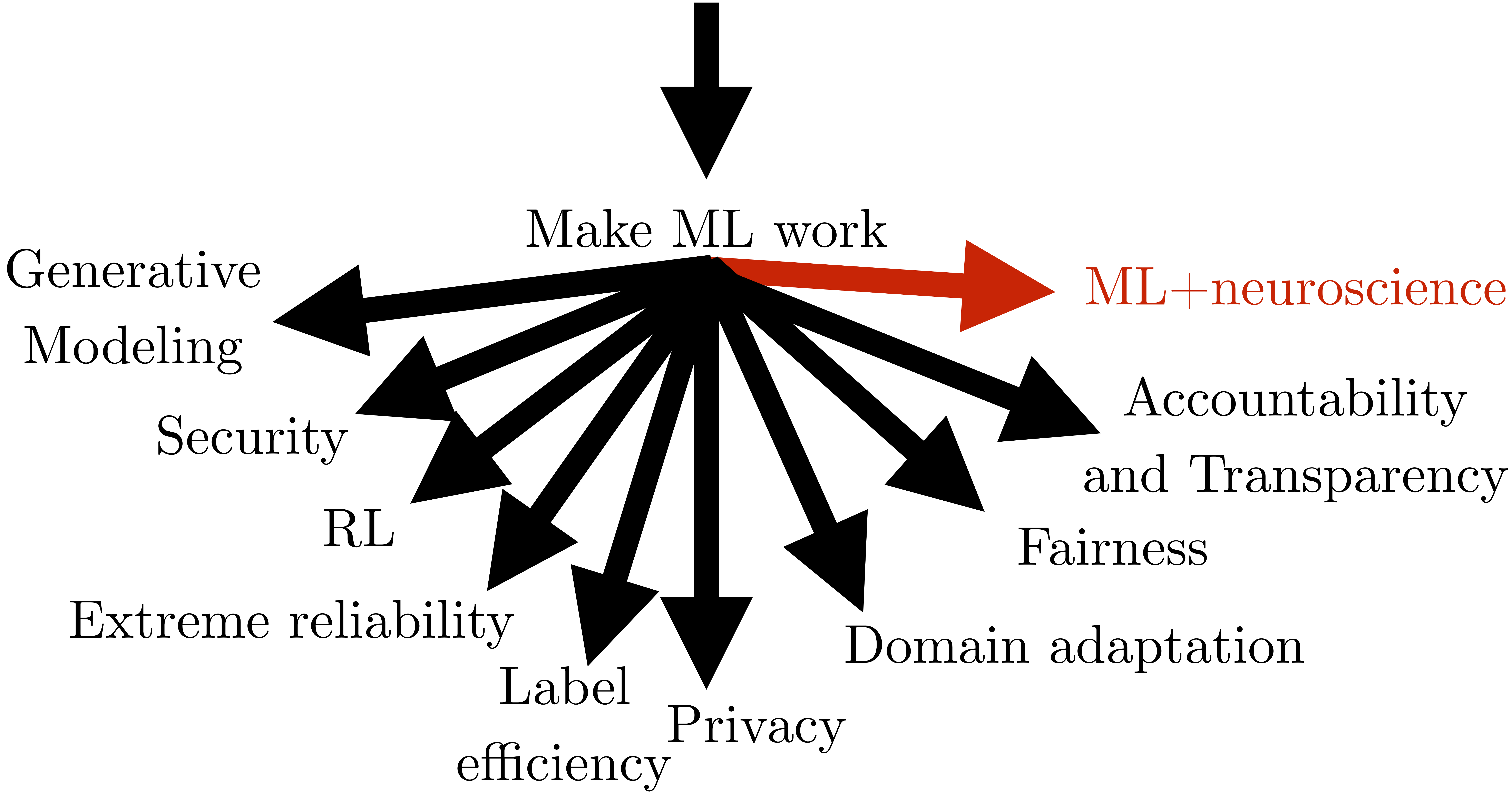


(Goodfellow et al, 2014)

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.

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2017)

Adversarial Examples that Fool both Human and Computer Vision

Gamaleldin et al 2018

Questions

(Goodfellow 2018)