Title: Adversarial Machine Learning

Date: Nov 28, 2018 02:00 PM

URL: http://pirsa.org/18110086

Abstract: Most machine learning algorithms involve optimizing a single set of parameters to decrease a single cost function. In adversarial machine learning, two or more "players" each adapt their own parameters to decrease their own cost, in competition with the other players. In some adversarial machine learning algorithms, the algorithm designer contrives this competition between two machine learning models in order to produce a beneficial side effect. For example, the generative adversarial networks framework involves a contrived conflict between a generator network and a discriminator network that results in the generator learning to produce realistic data samples. In other contexts, adversarial machine learning models a real conflict, for example, between spam detectors and spammers. In general, moving machine learning from optimization and a single cost to game theory and multiple costs has led to new insights in many application areas.

Adversarial Machine Learning

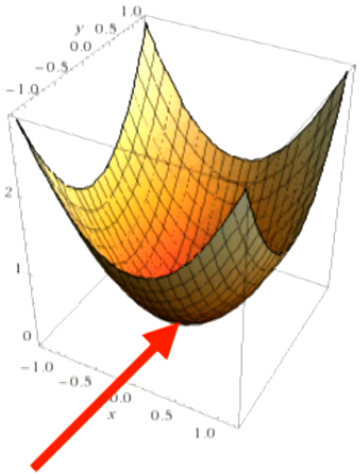
Ian Goodfellow, Staff Research Scientist, Google Brain

ACM Webinar

2018-07-24

Adversarial Machine Learning

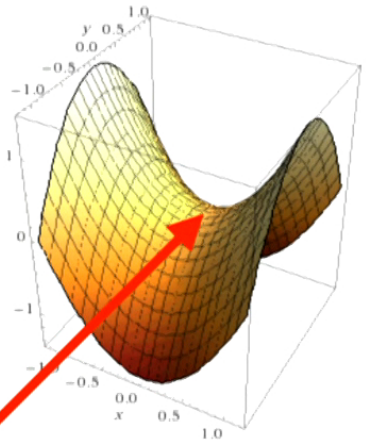
Traditional ML:
optimization



Minimum

One player,
one cost

Adversarial ML:
game theory

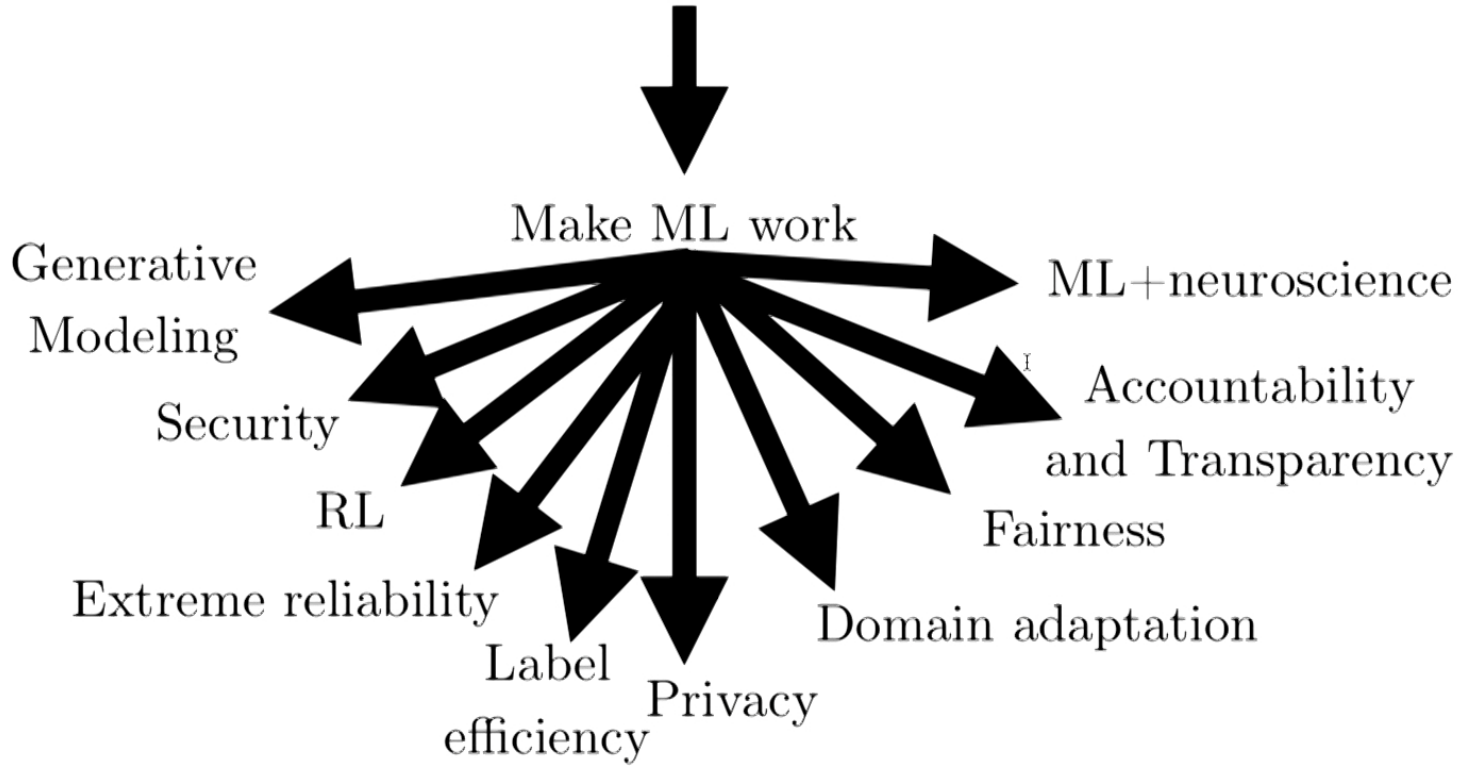


Equilibrium

More than one player,
more than one cost

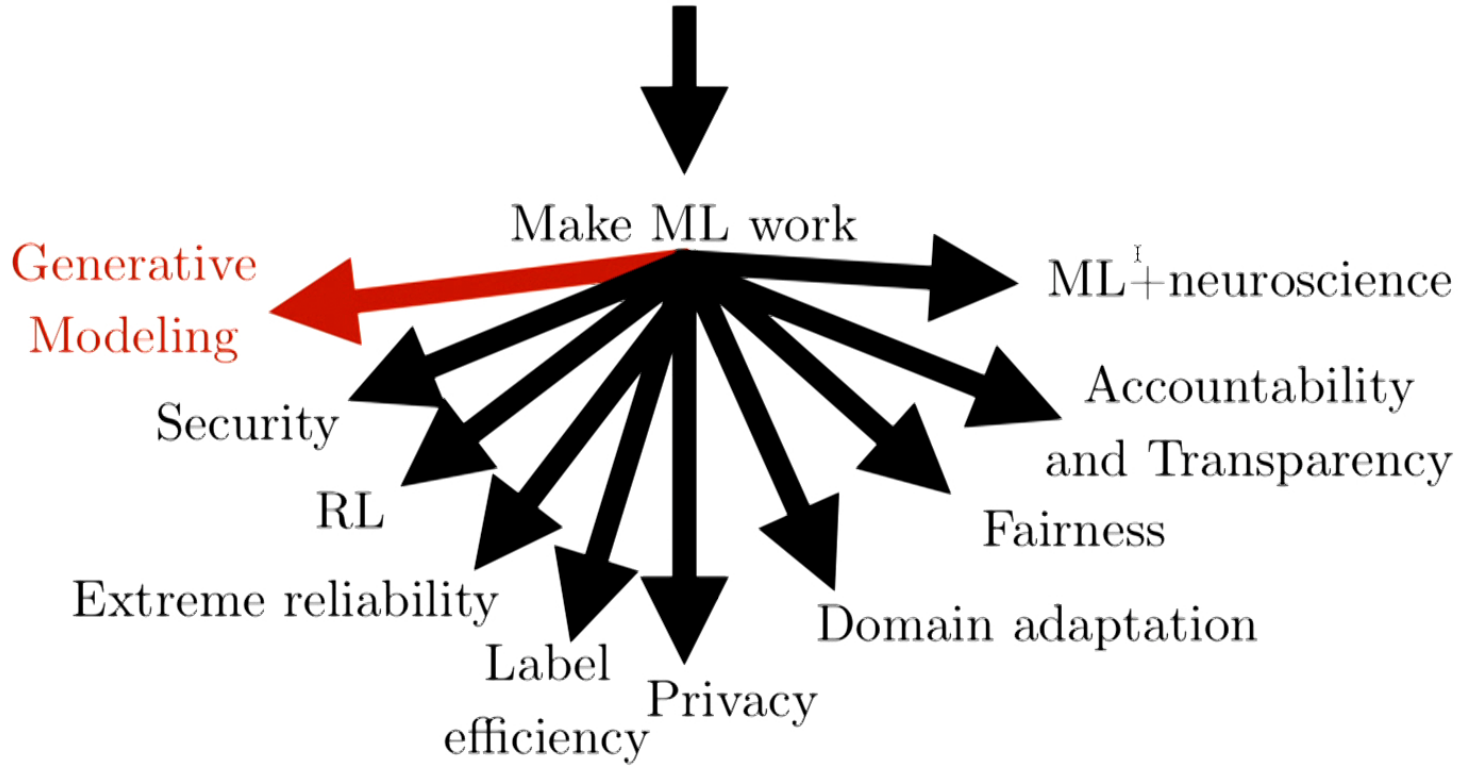
(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

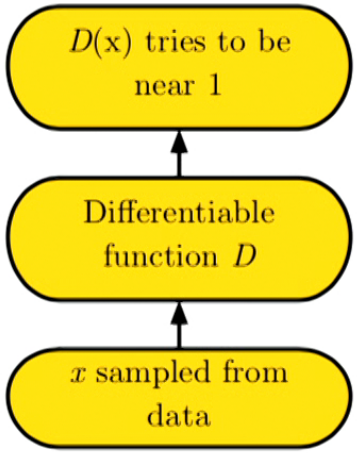
Generative Modeling: Sample Generation

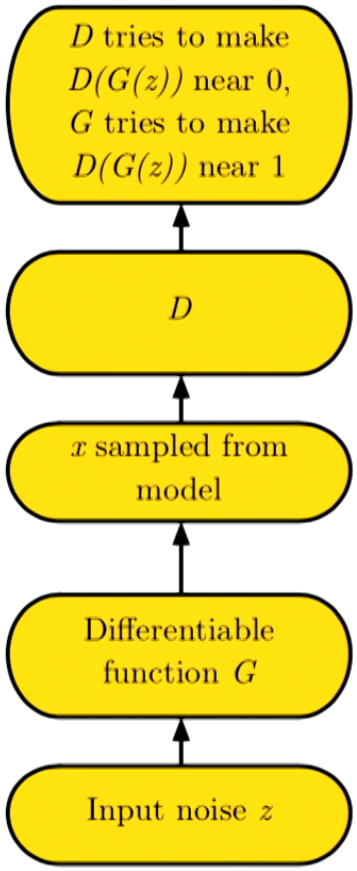
Training Data
(CelebA)

Sample Generator
(Karras et al, 2017)

(Goodfellow 2018)

Adversarial Nets Framework

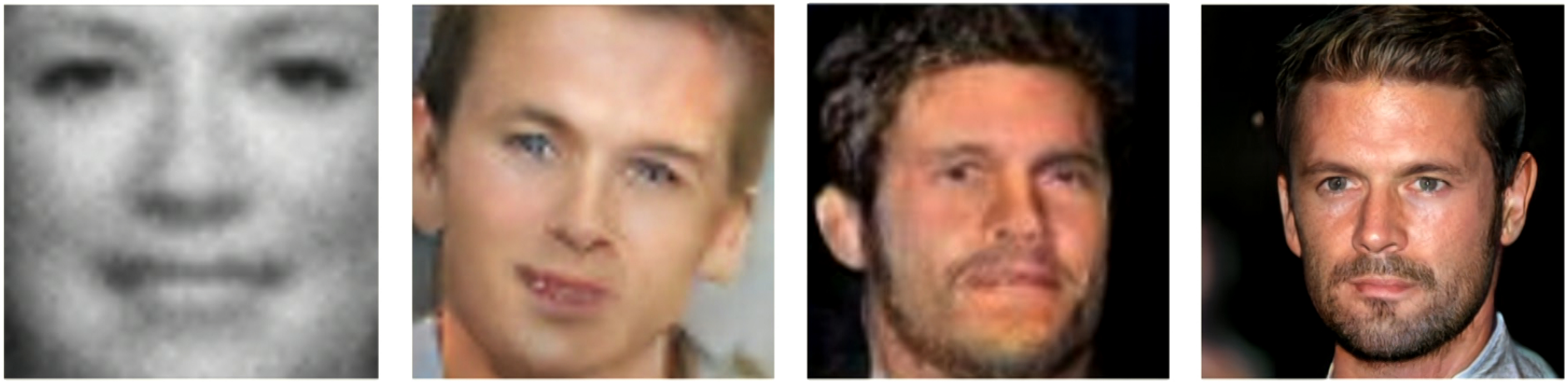




(Goodfellow et al., 2014)

(Goodfellow 2018)

3.5 Years of Progress on Faces



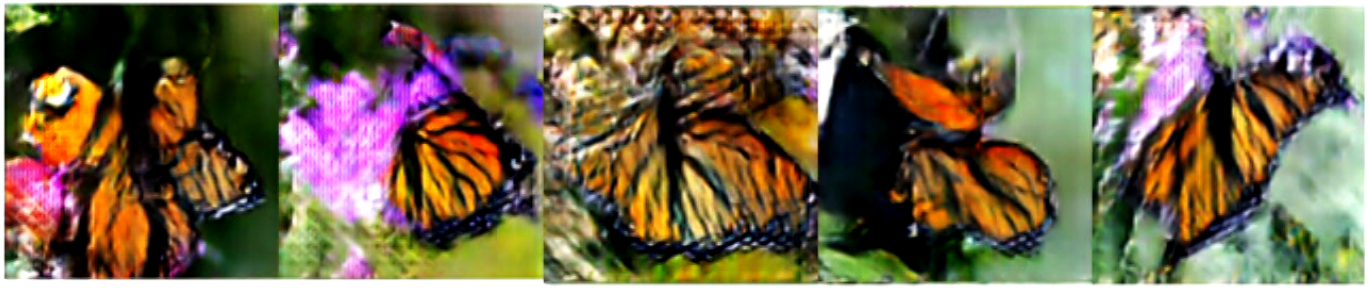
2014 2015 2016 2017

(Brundage et al, 2018)

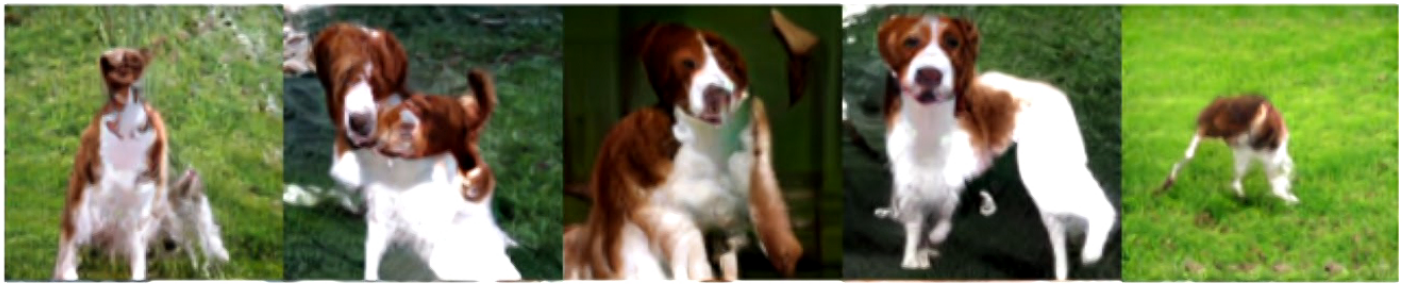
(Goodfellow 2018)

<2 Years of Progress on ImageNet

Odena et al 2016



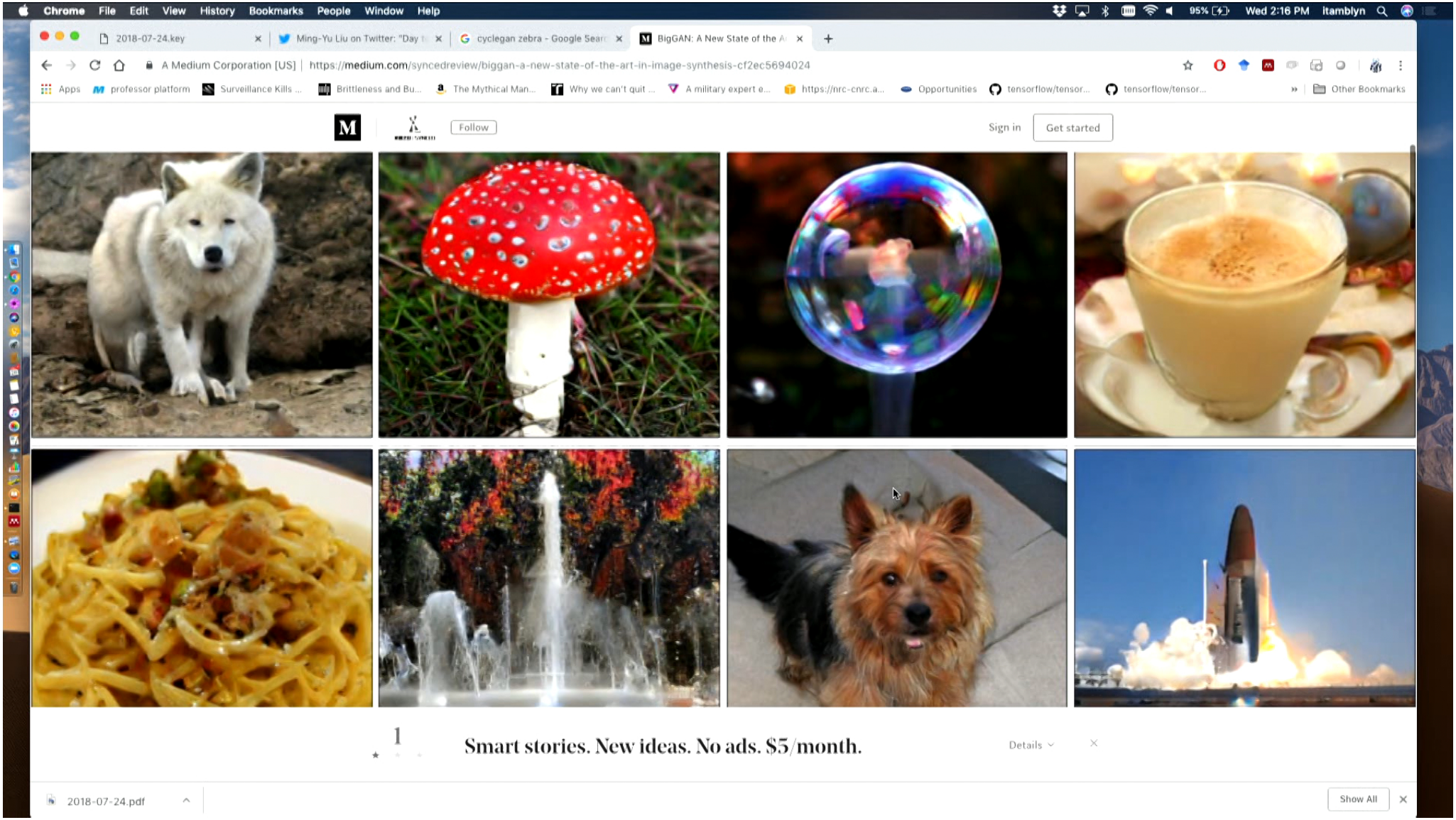
Miyato et al 2017



Zhang et al 2018

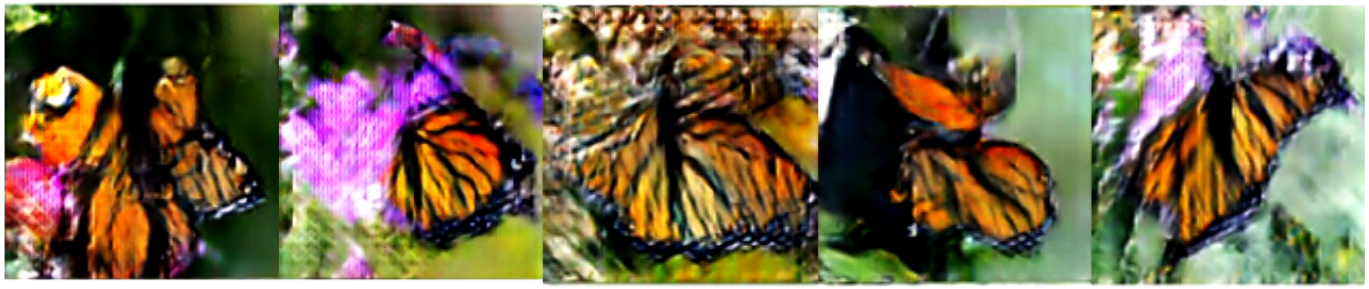
(Goodfellow 2018)

Figure 1: Class-conditional samples generated by our model.

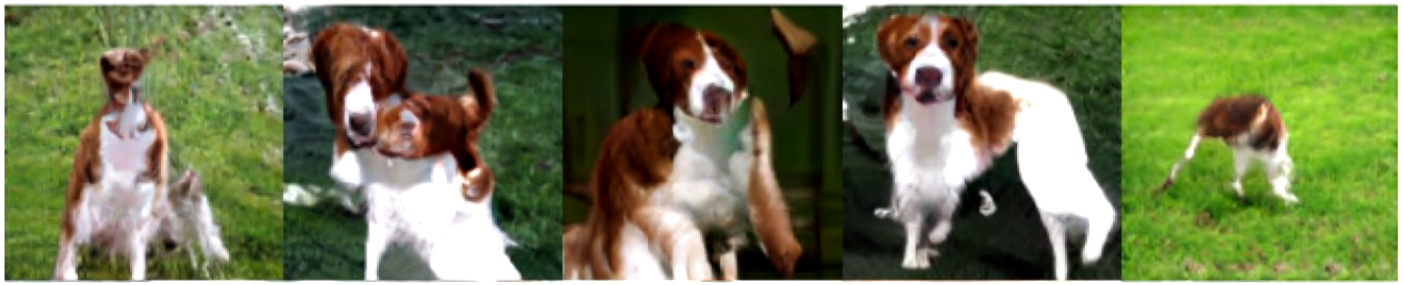

Chrome File Edit View History Bookmarks People Window Help
95% Wed 2:16 PM itamblyn
2018-07-24.key
Ming-Yu Liu on Twitter: "Day
cyclegan zebra - Google Sear
BigGAN: A New State of the A
A Medium Corporation [US] | https://medium.com/syncedreview/biggan-a-new-state-of-the-art-in-image-synthesis-cf2ec5694024
Apps
professor platform
Surveillance Kills ...
Brittleness and Bu...
The Mythical Man...
Why we can't quit ...
A military expert e...
https://nrc-cnrc.a...
Opportunities
tensorflow/tensor...
tensorflow/tensor...
Other Bookmarks
Follow
Sign in
Get started
1
Smart stories. New ideas. No ads. $5/month.
Details
2018-07-24.pdf
Show All

<2 Years of Progress on ImageNet

Odena et al 2016



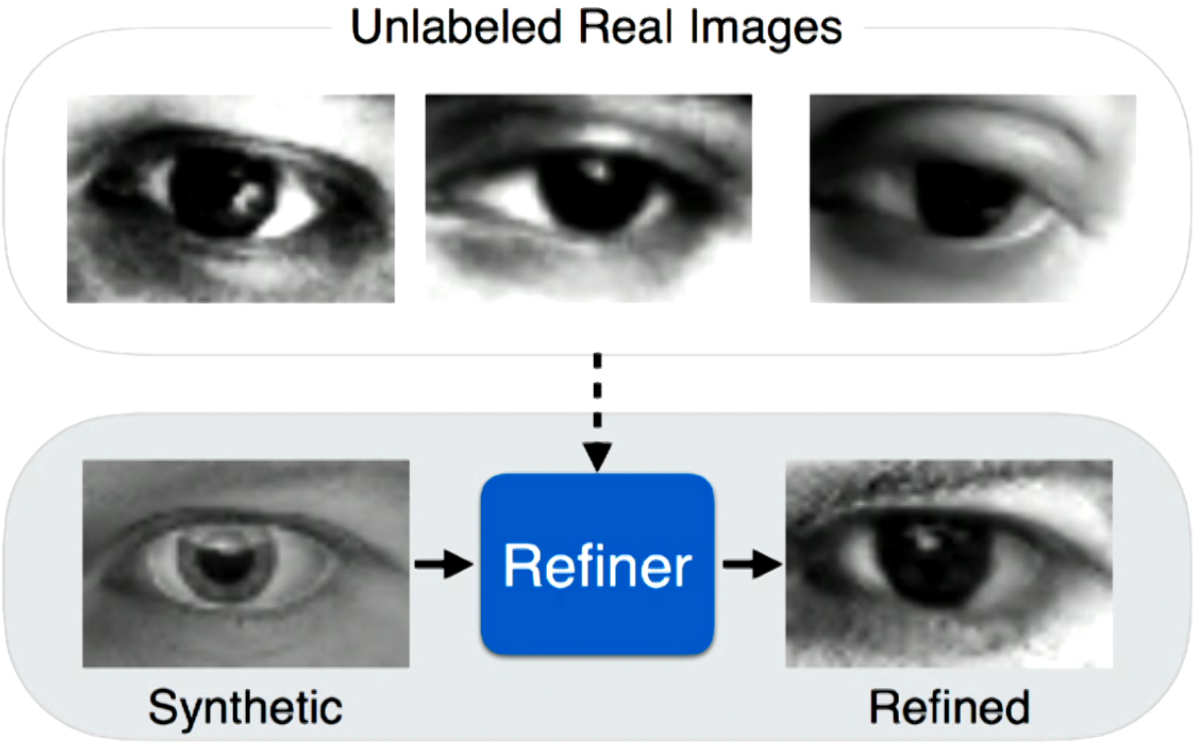
Miyato et al 2017



Zhang et al 2018

(Goodfellow 2018)

GANs for simulated training data



(Shrivastava et al., 2016)

(Goodfellow 2018)

Unsupervised Image-to-Image Translation

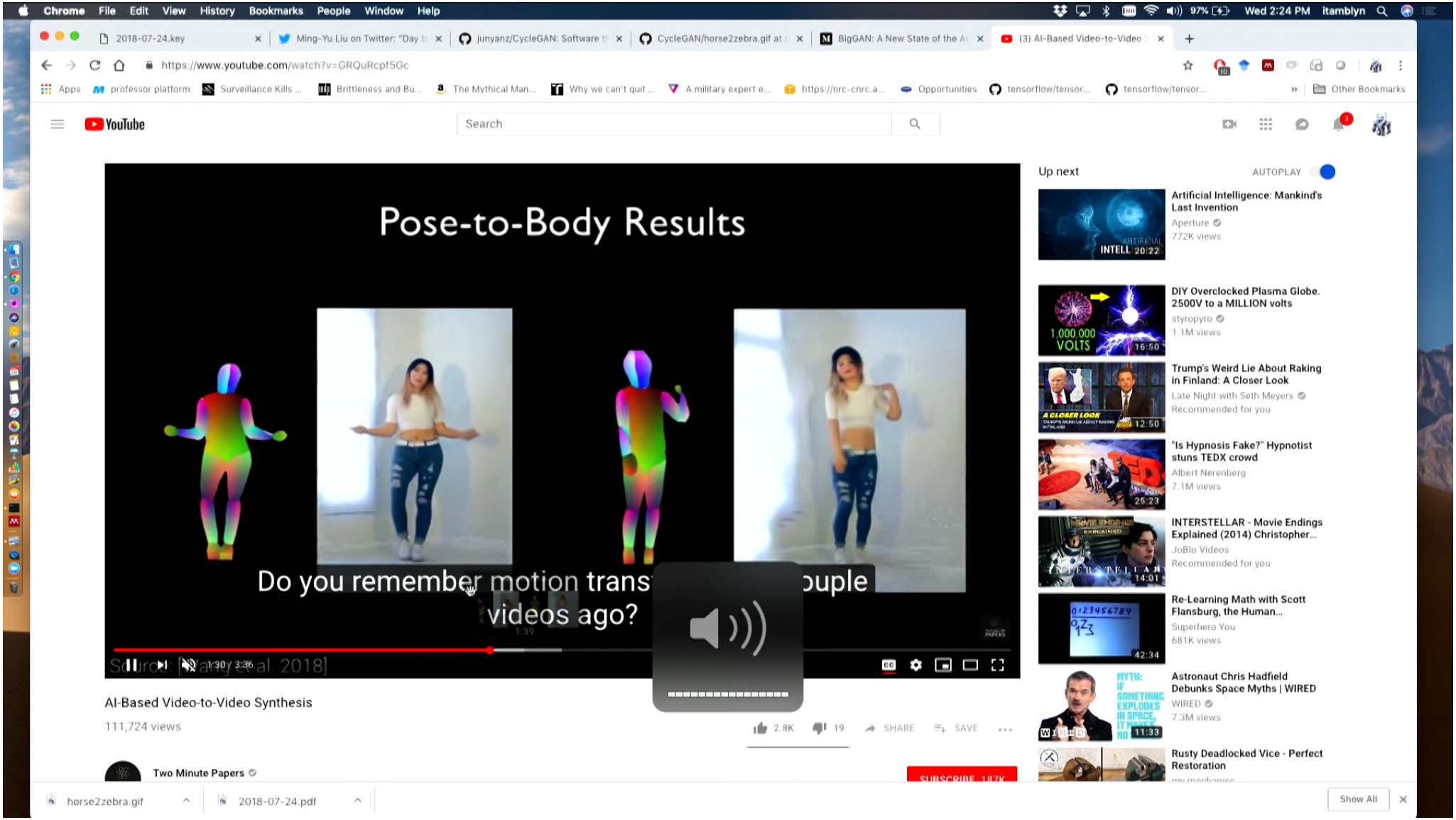
Day to night

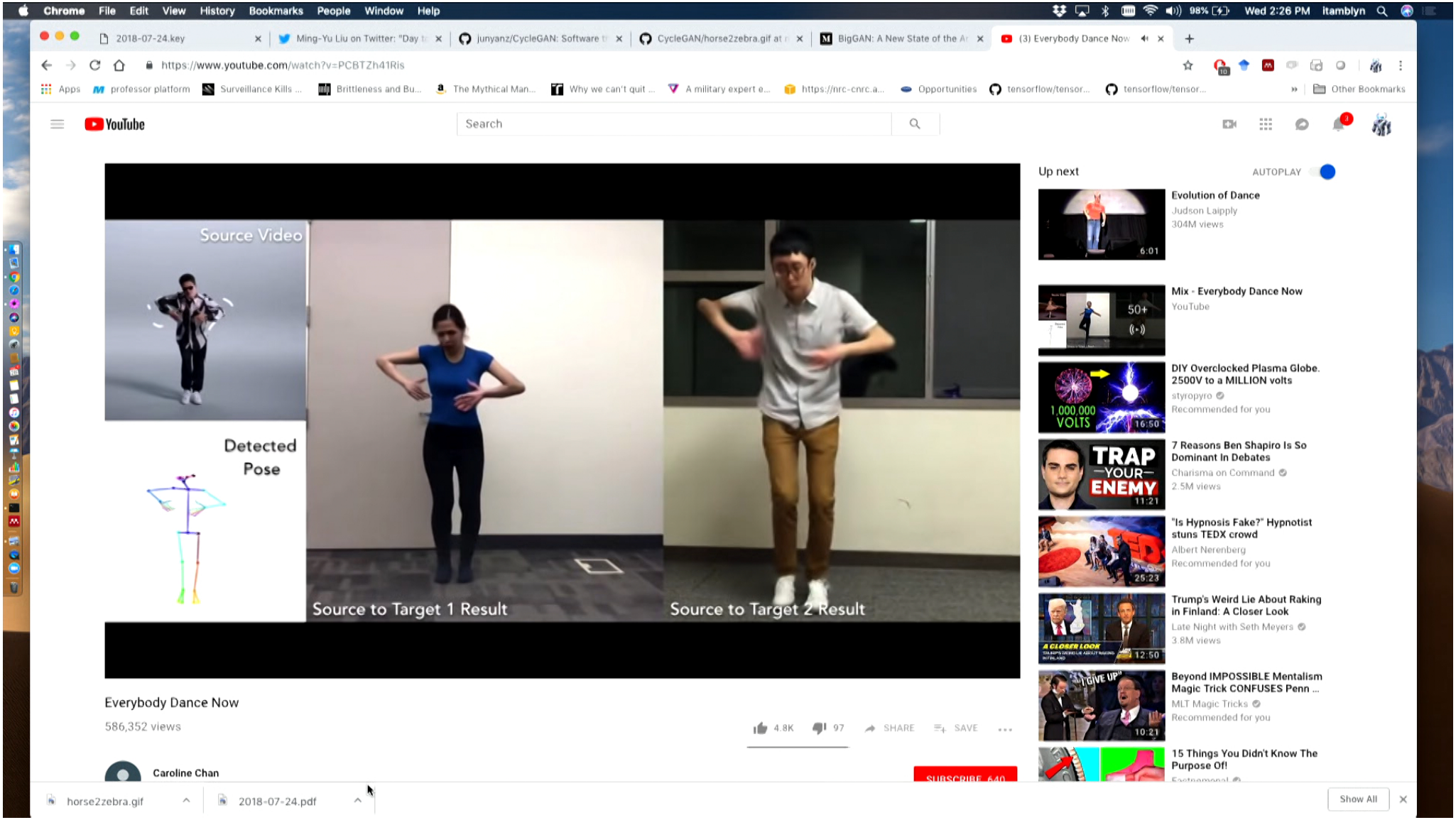
(Liu et al., 2017)

(Goodfellow 2018)

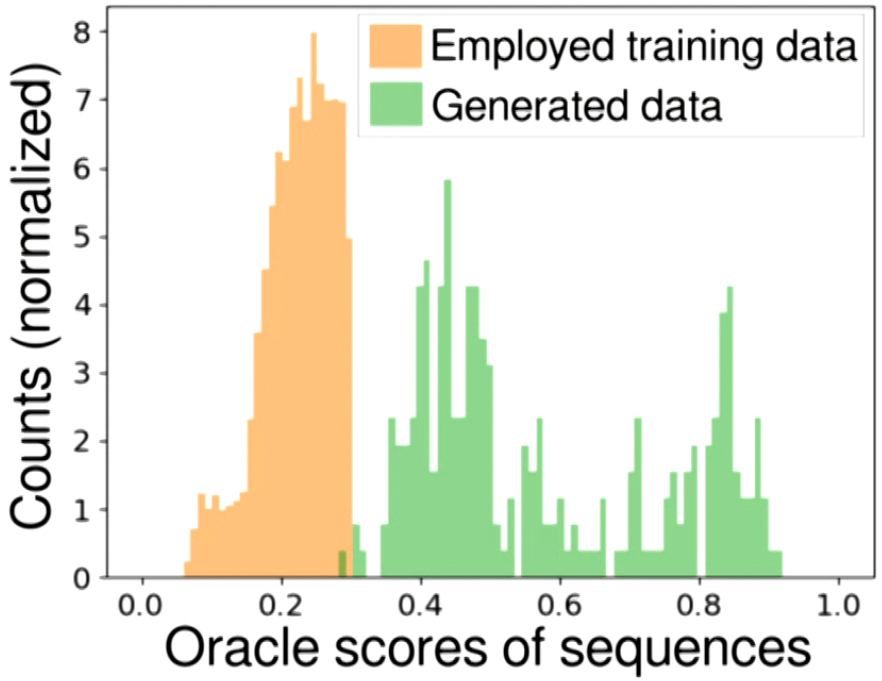
CycleGAN

(Zhu et al., 2017)


Pose-to-Body Results
Do you remember motion trans ouple
videos ago?
AI-Based Video-to-Video Synthesis
111,724 views
Two Minute Papers
Up next
AUTOPLAY
Artificial Intelligence: Mankind's Last Invention
Aperture
772K views
DIY Overclocked Plasma Globe. 2500V to a MILLION volts
styropyro
1.1M views
Trump's Weird Lie About Raking in Finland: A Closer Look
Late Night with Seth Meyers
Recommended for you
"Is Hypnosis Fake?" Hypnotist stuns TEDX crowd
Albert Nerenberg
7.1M views
INTERSTELLAR - Movie Endings Explained (2014) Christopher...
JoBlo Videos
Recommended for you
Re-Learning Math with Scott Flansburg, the Human...
Superhero You
681K views
Astronaut Chris Hadfield Debunks Space Myths | WIRED
WIRED
7.3M views
Rusty Deadlocked Vice - Perfect Restoration
horse2zebra.gif
2018-07-24.pdf
Show All

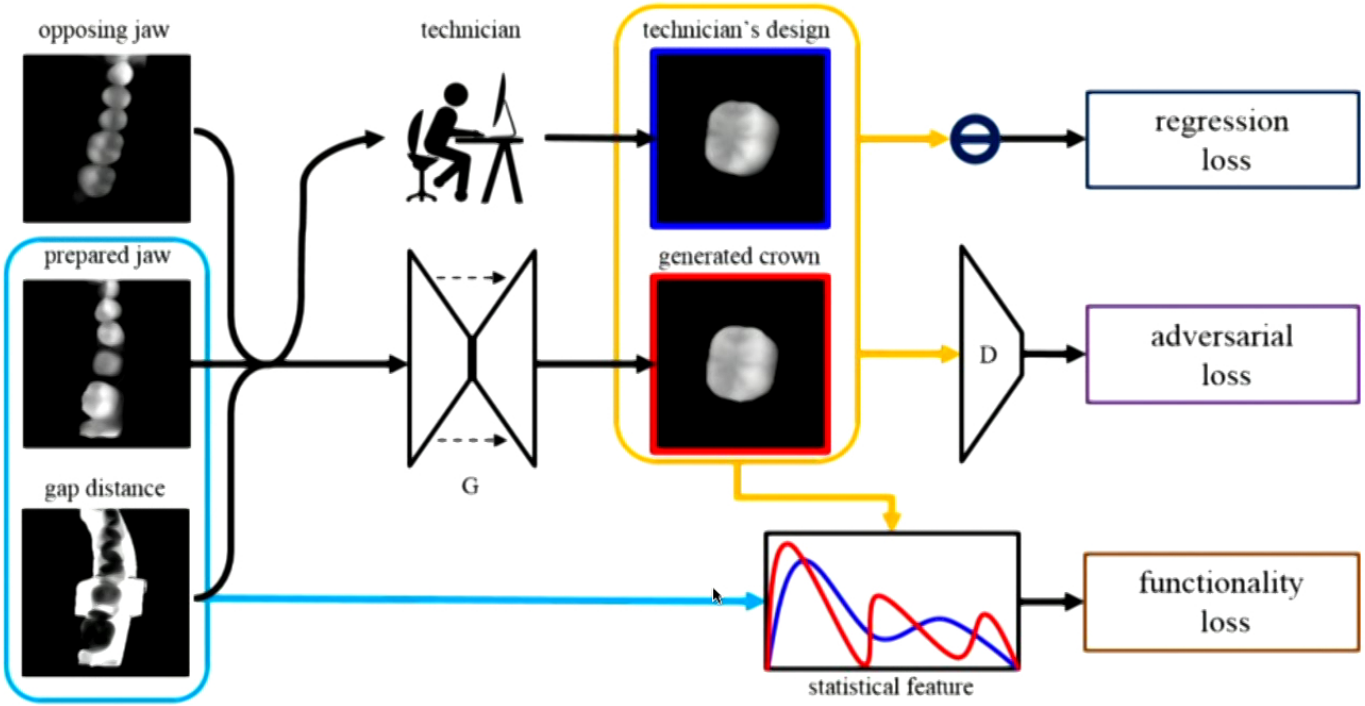

Source Video
Detected Pose
Source to Target 1 Result
Source to Target 2 Result
Everybody Dance Now
586,352 views
Up next
AUTOPLAY
Evolution of Dance
Judson Laipply
304M views
Mix - Everybody Dance Now
YouTube
DIY Overclocked Plasma Globe. 2500V to a MILLION volts
styropyro
Recommended for you
7 Reasons Ben Shapiro Is So Dominant In Debates
Charisma on Command
2.5M views
"Is Hypnosis Fake?" Hypnotist stuns TEDX crowd
Albert Nerenberg
Recommended for you
Trump's Weird Lie About Raking in Finland: A Closer Look
Late Night with Seth Meyers
3.8M views
Beyond IMPOSSIBLE Mentalism Magic Trick CONFUSES Penn ...
MLT Magic Tricks
Recommended for you
15 Things You Didn't Know The Purpose Of!
horse2zebra.gif
2018-07-24.pdf
Show All

Designing DNA to optimize protein binding



(Killoran et al, 2017)

Personalized GANufacturing

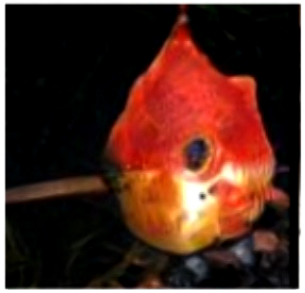


(Hwang et al 2018)

(Goodfellow 2018)

Self-Attention GAN

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



Goldfish

Redshank

Broccoli

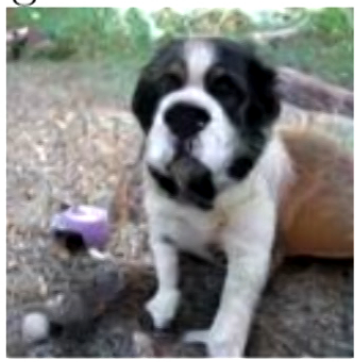
Tiger Cat

Geyser

Indigo Bunting

Stone Wall

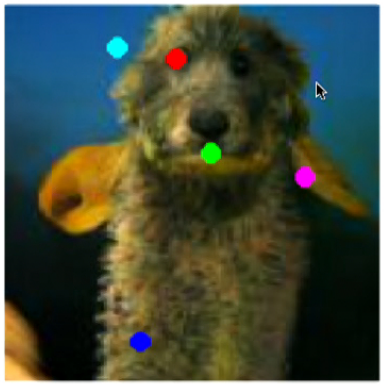
(Zhang et al., 2018)



Saint Bernard

(Goodfellow 2018)

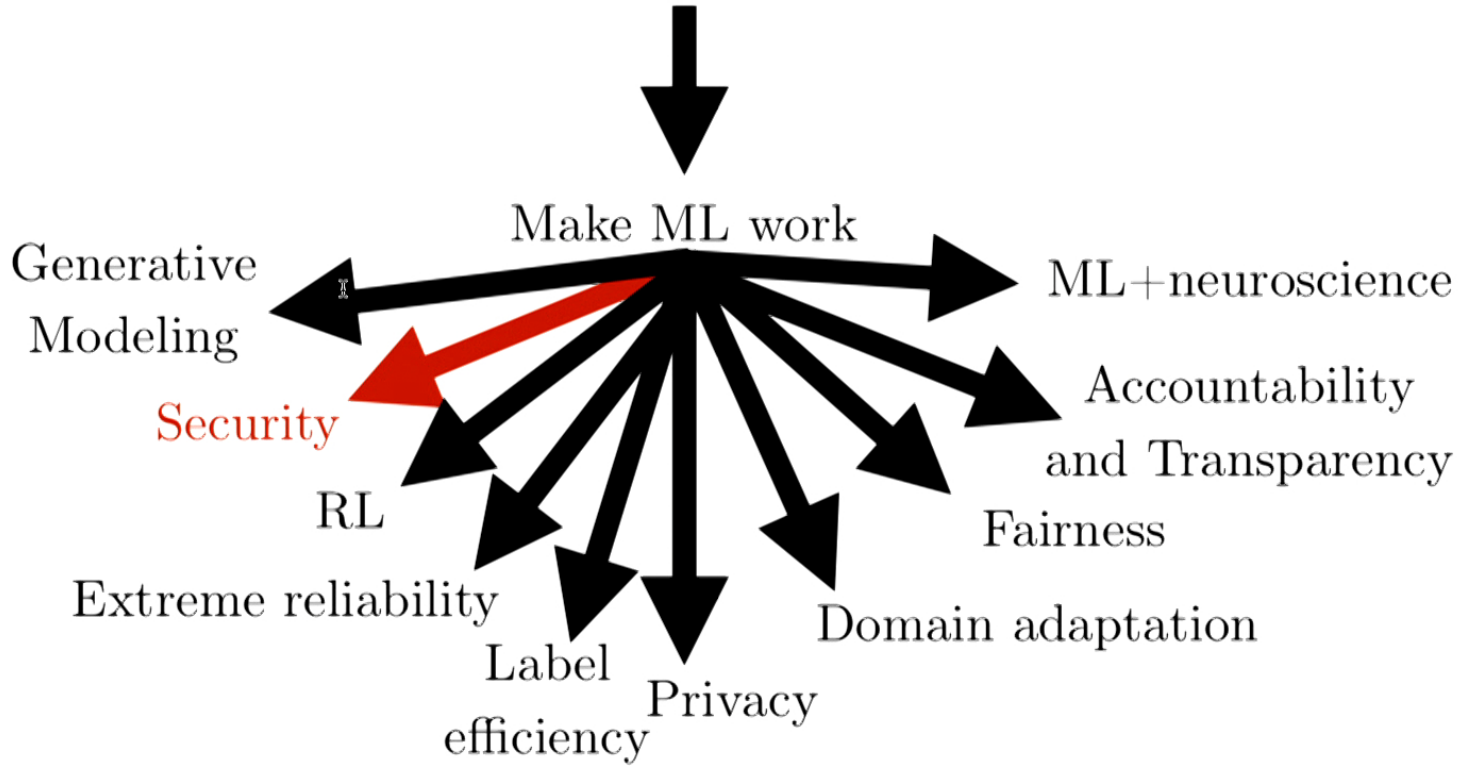
Self-Attention




Use layers from
Wang et al 2018

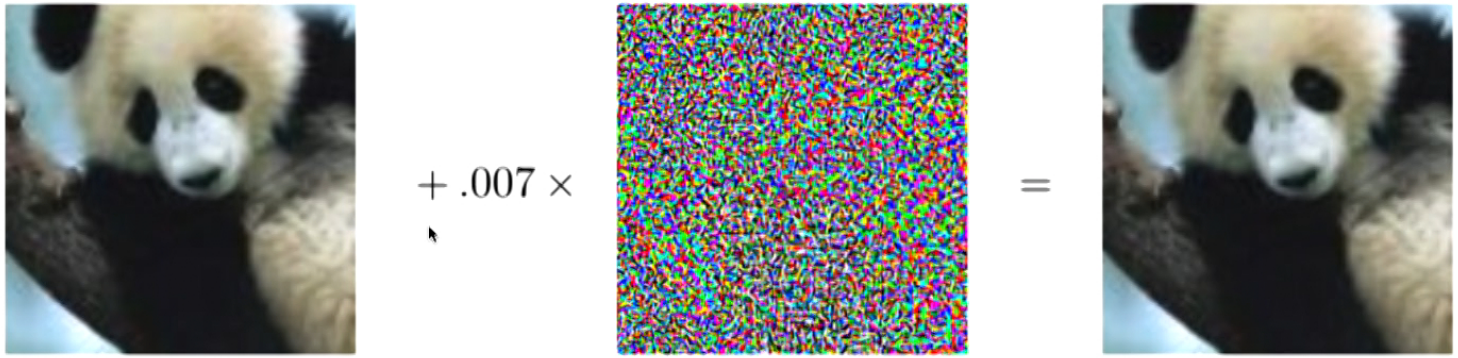

(Goodfellow 2018)

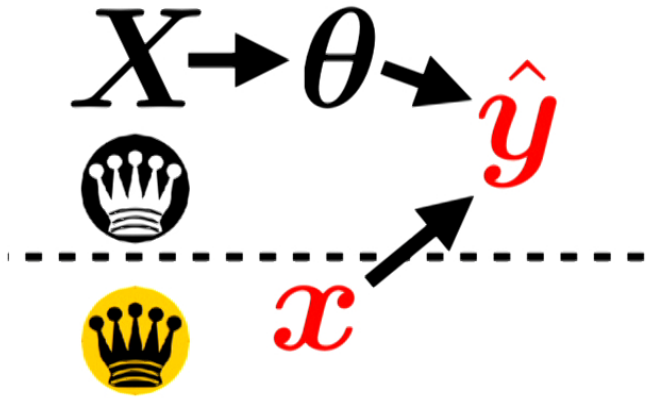
A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

Adversarial Examples

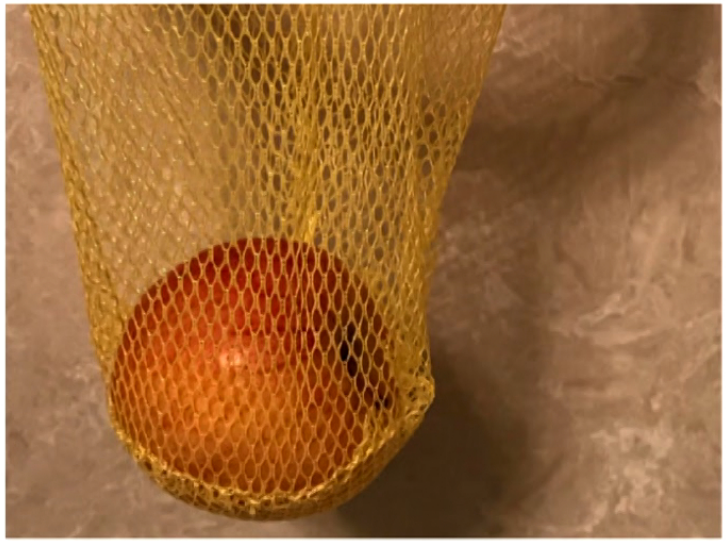




(Goodfellow 2018)

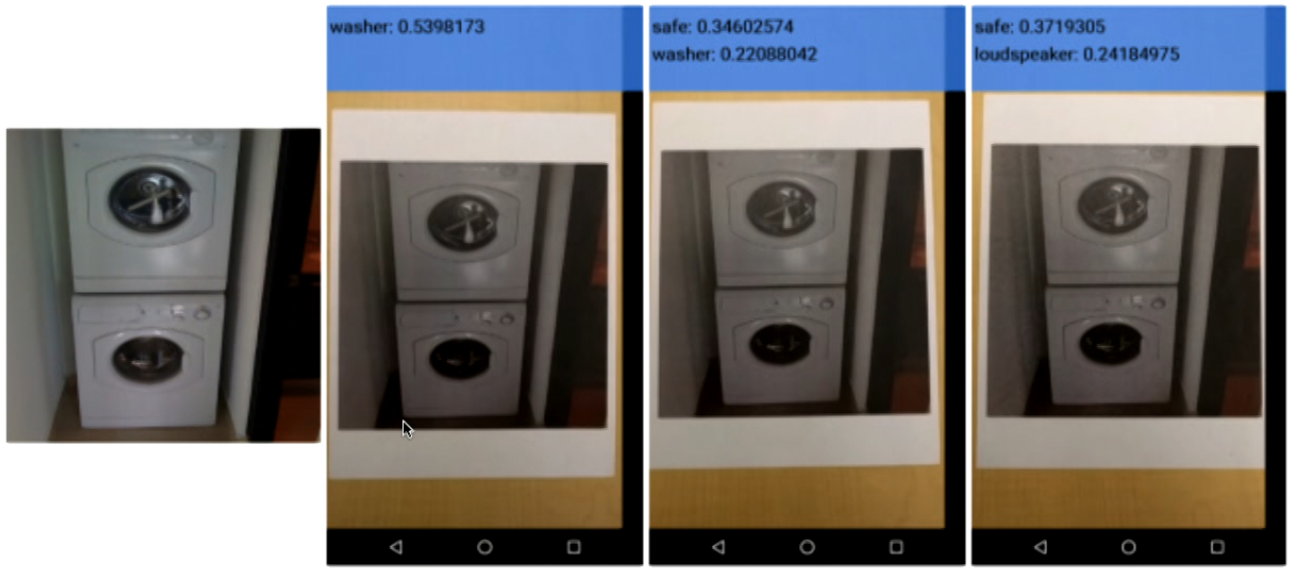
Also Adversarial Examples

(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,

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

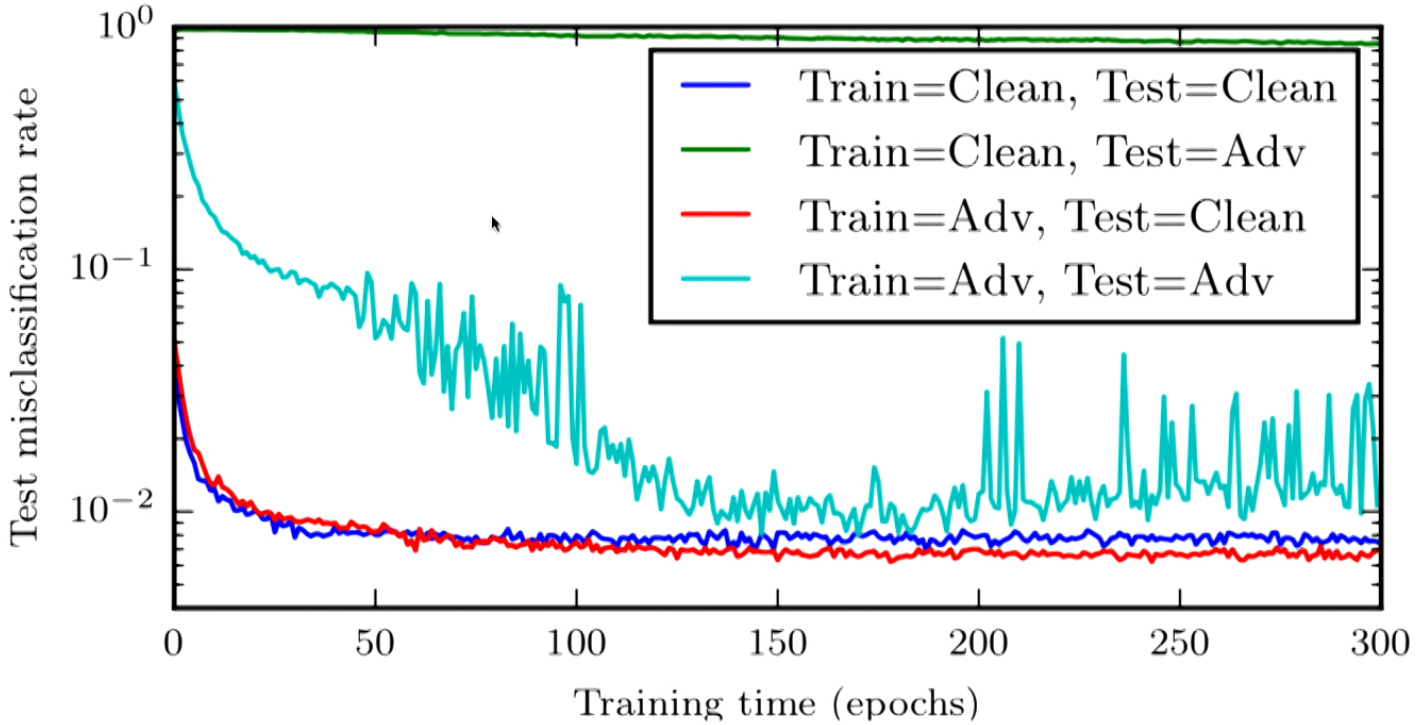
with the learning algorithm as the minimizing player and a fixed procedure (such as L-BFGS or the fast gradient sign method) as the maximizing player."

Original implementation: Goodfellow et al 2014

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

(Goodfellow 2018)

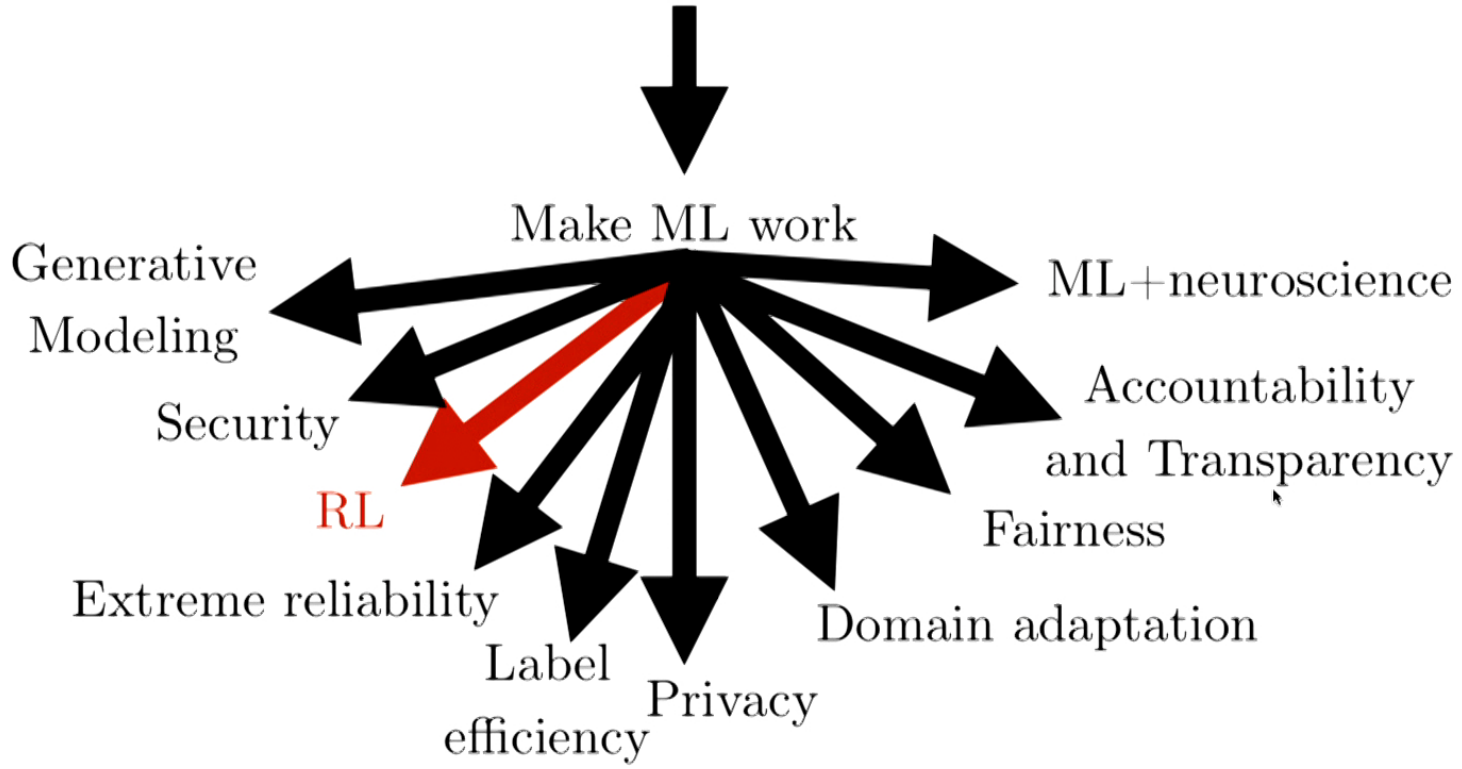
Training on Adversarial Examples



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

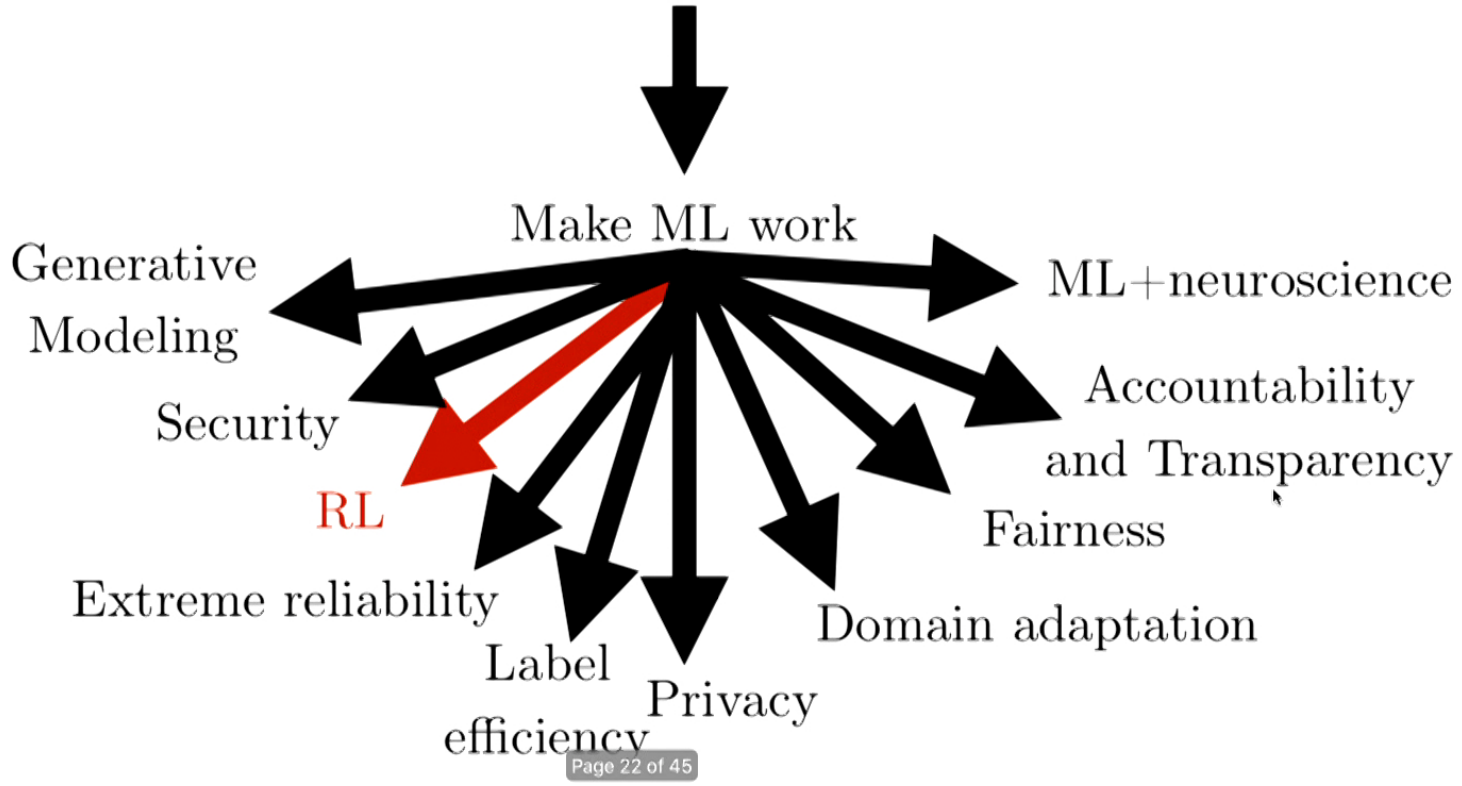
(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



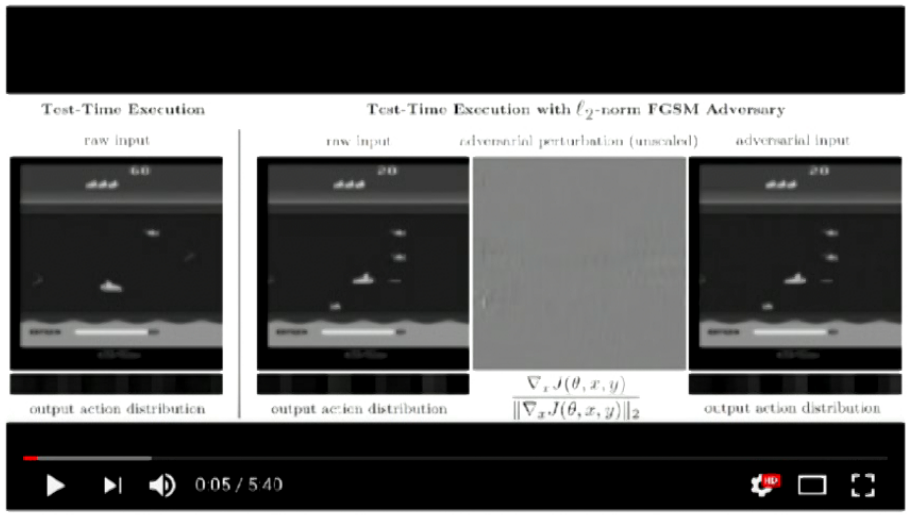
(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

Adversarial Examples for RL

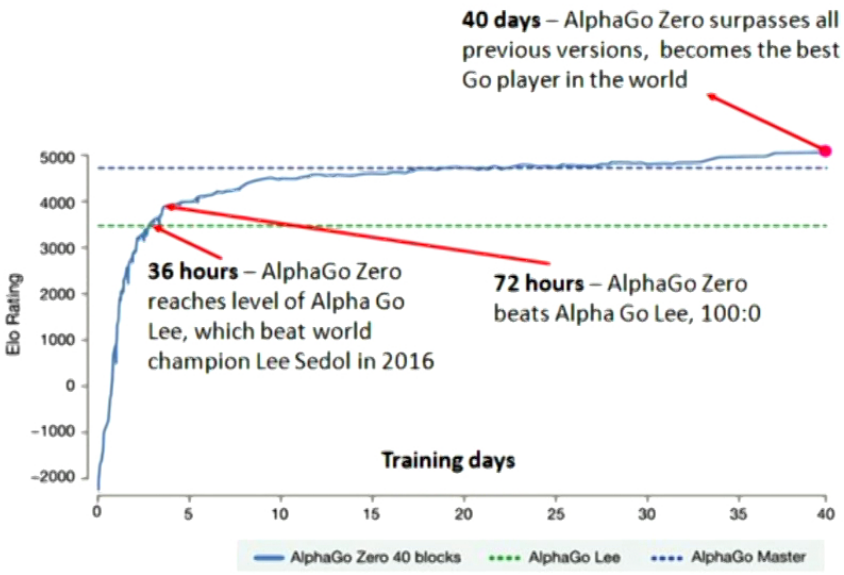


(Huang et al., 2017)

(Goodfellow 2018)

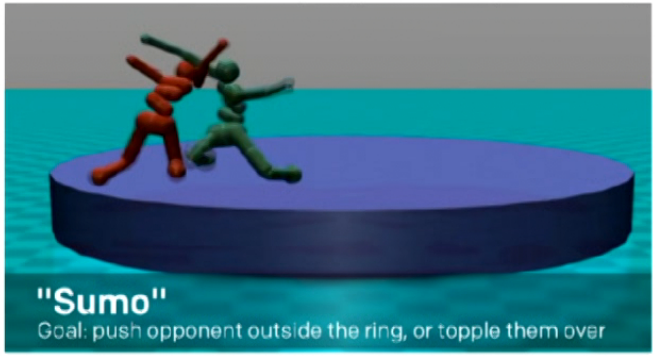
Self-Play

1959: Arthur Samuel's checkers agent



(Silver et al, 2017)

(OpenAI, 2017)

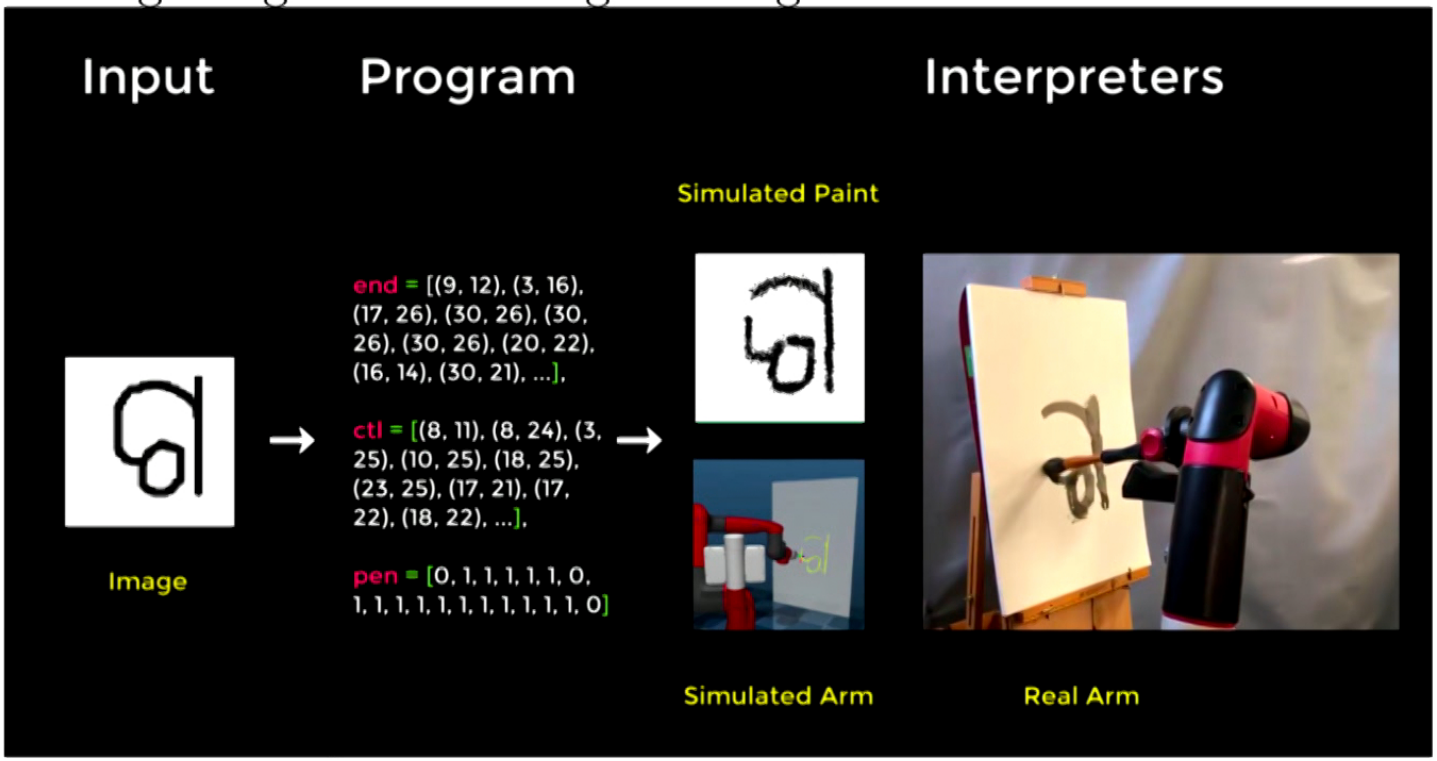


(Bansal et al, 2017)

(Goodfellow 2018)

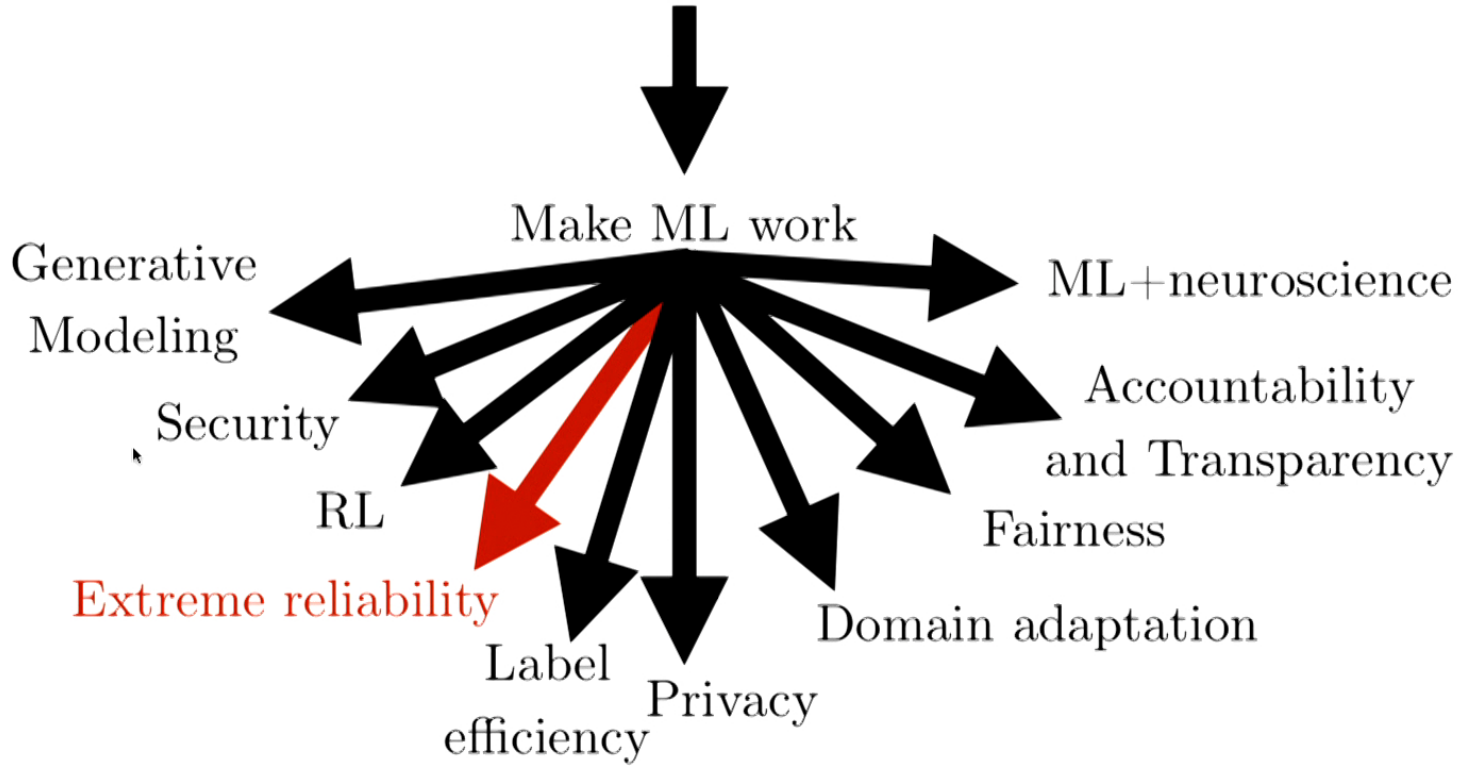
SPIRAL

Synthesizing Programs for Images Using Reinforced Adversarial Learning



(Ganin et al, 2018)

A Cambrian Explosion of Machine Learning Research Topics



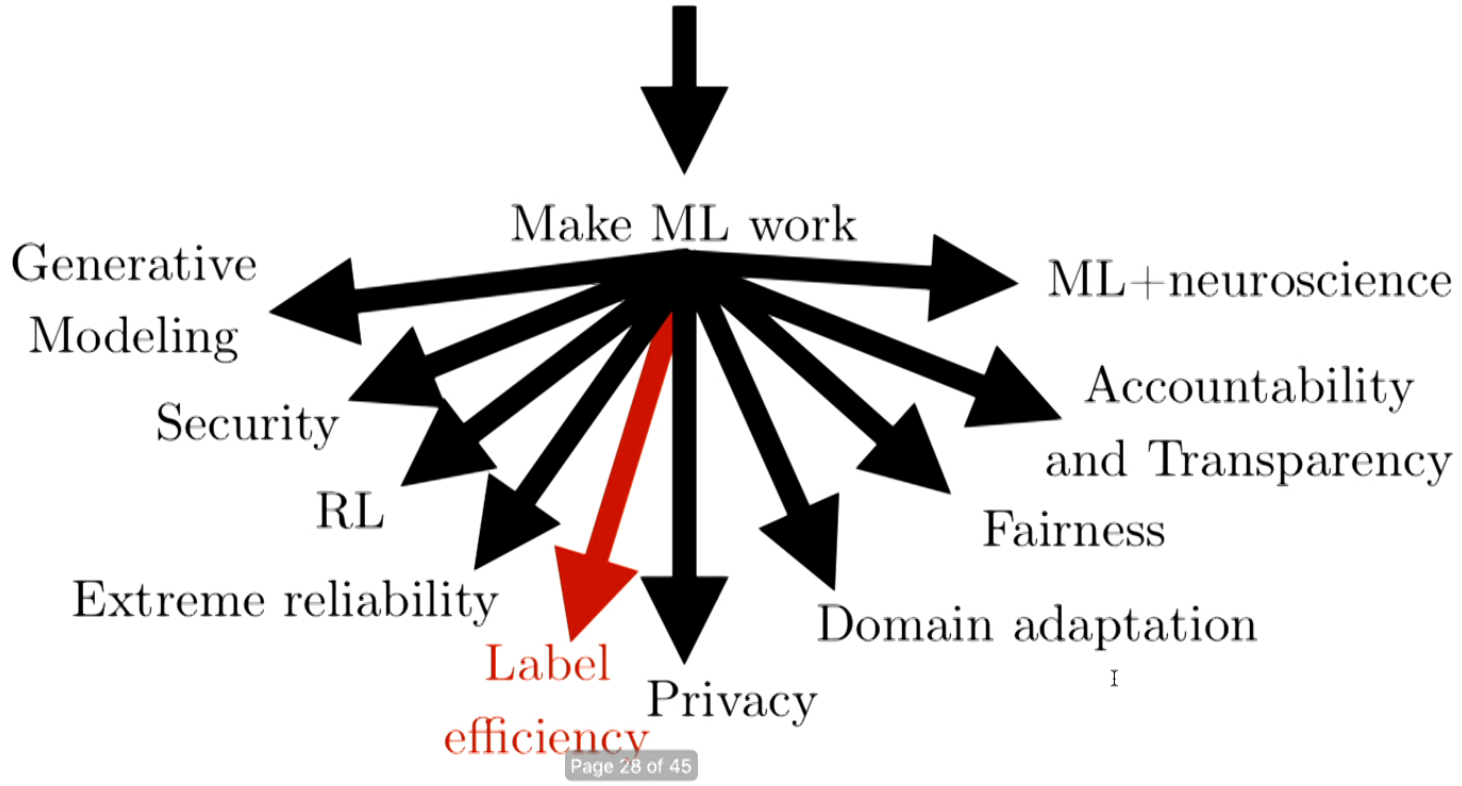
(Goodfellow 2018)

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

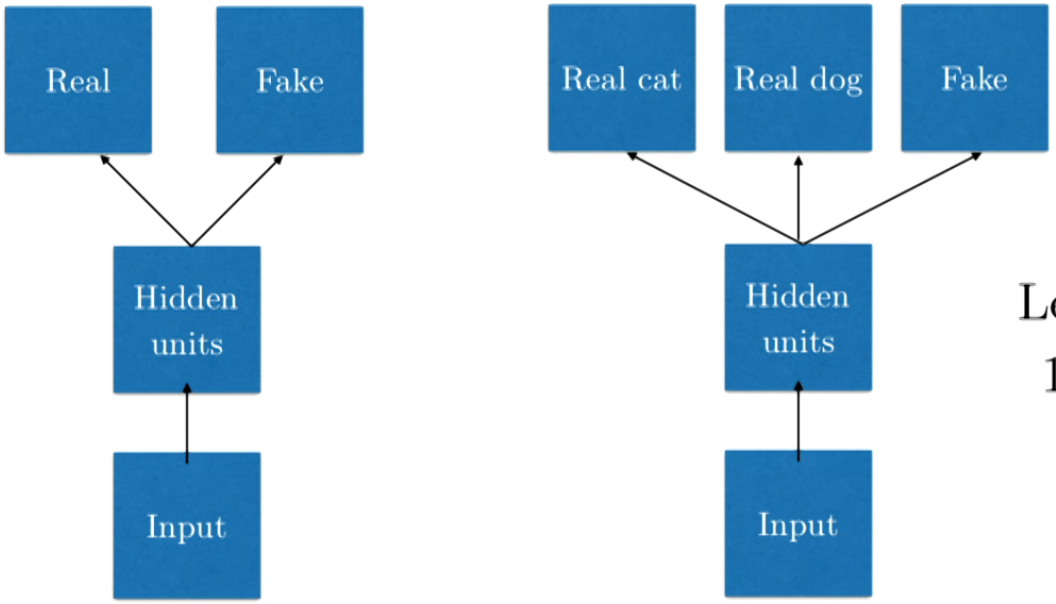
(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

Supervised Discriminator for Semi-Supervised Learning



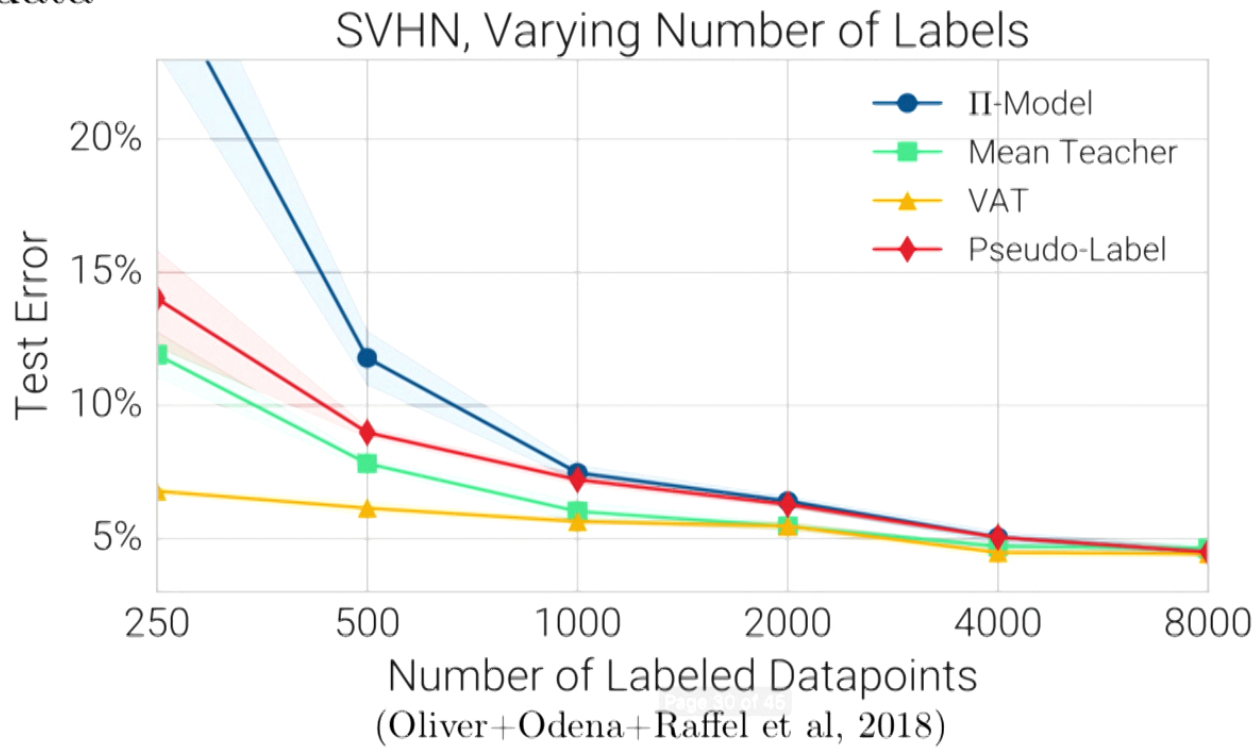
Learn to read with 100 labels rather than 60,000

(Odena 2016, Salimans et al 2016)

(Goodfellow 2018)

Virtual Adversarial Training

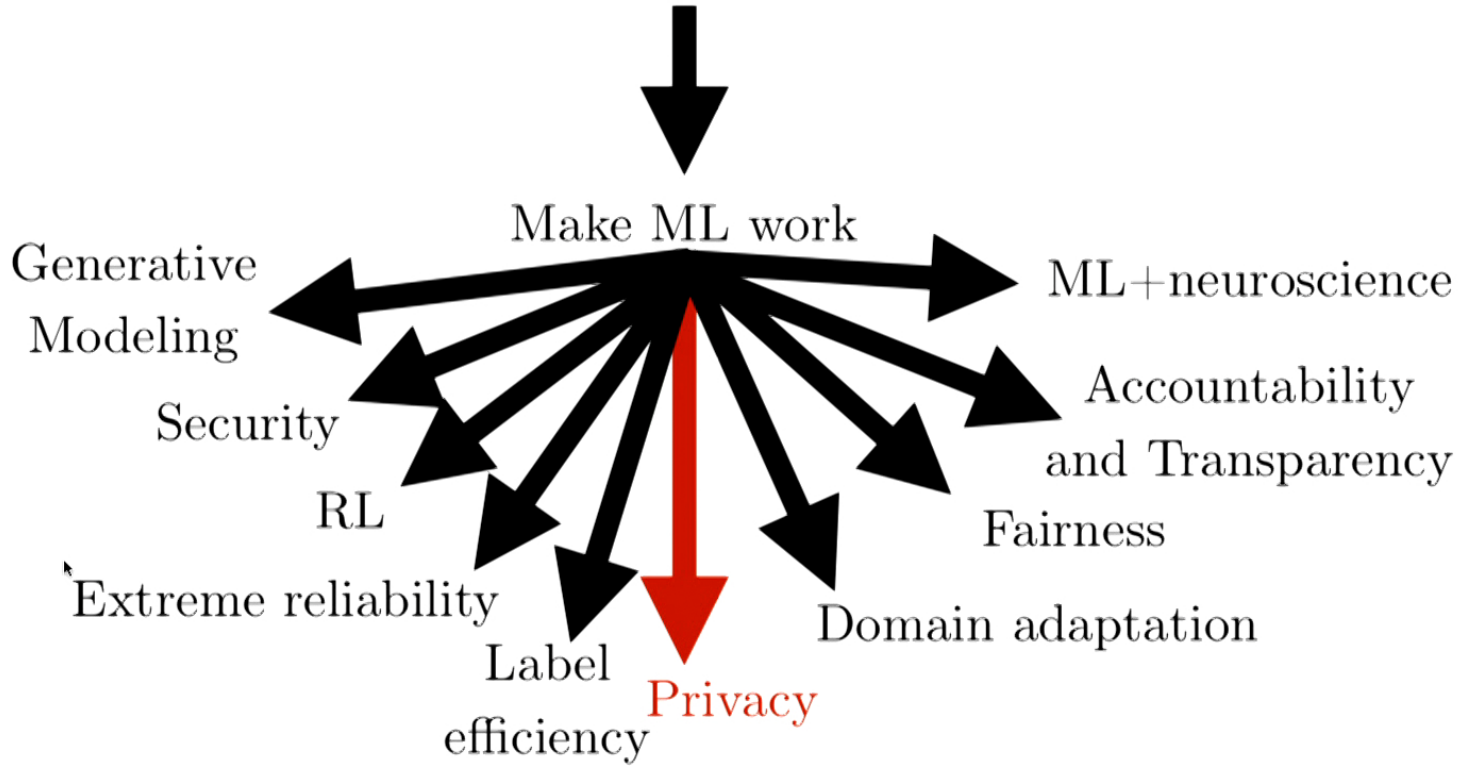
Miyato et al 2015: regularize for robustness to adversarial perturbations of *unlabeled* data



(Oliver+Odena+Raffel et al, 2018)

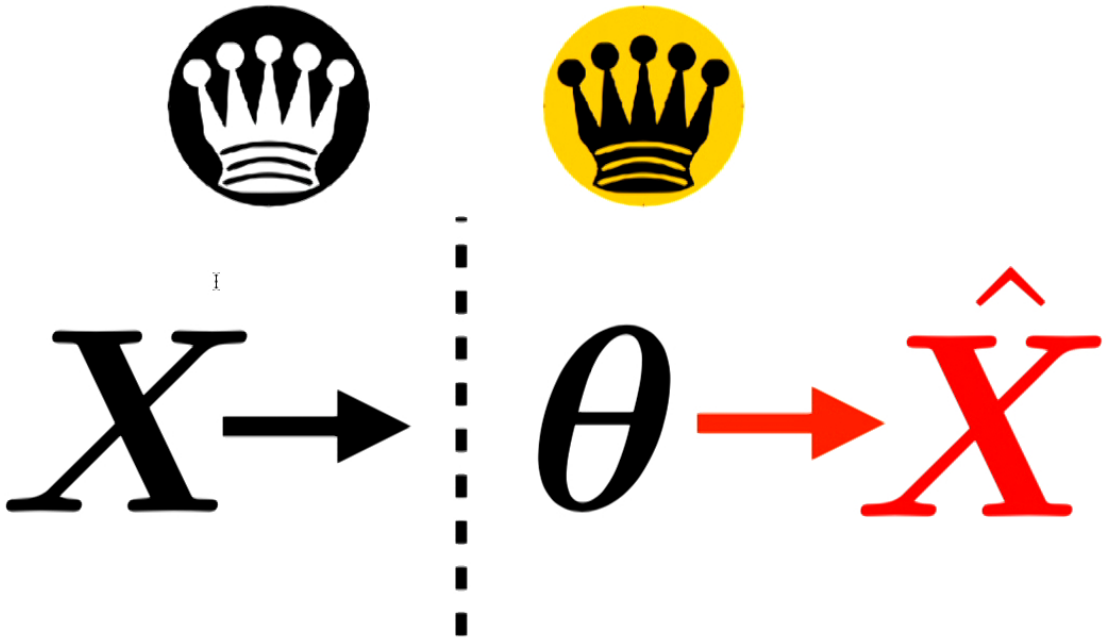
(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics



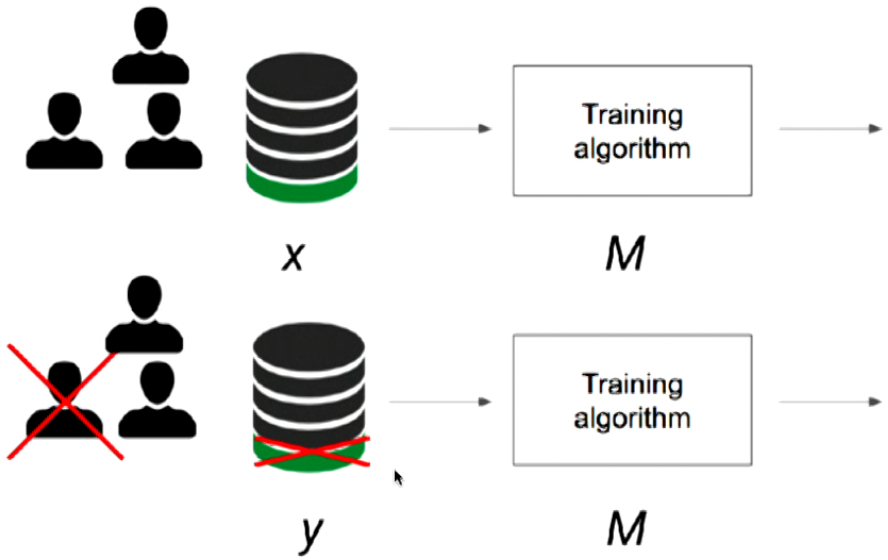
(Goodfellow 2018)

Privacy of training data



(Goodfellow 2018)

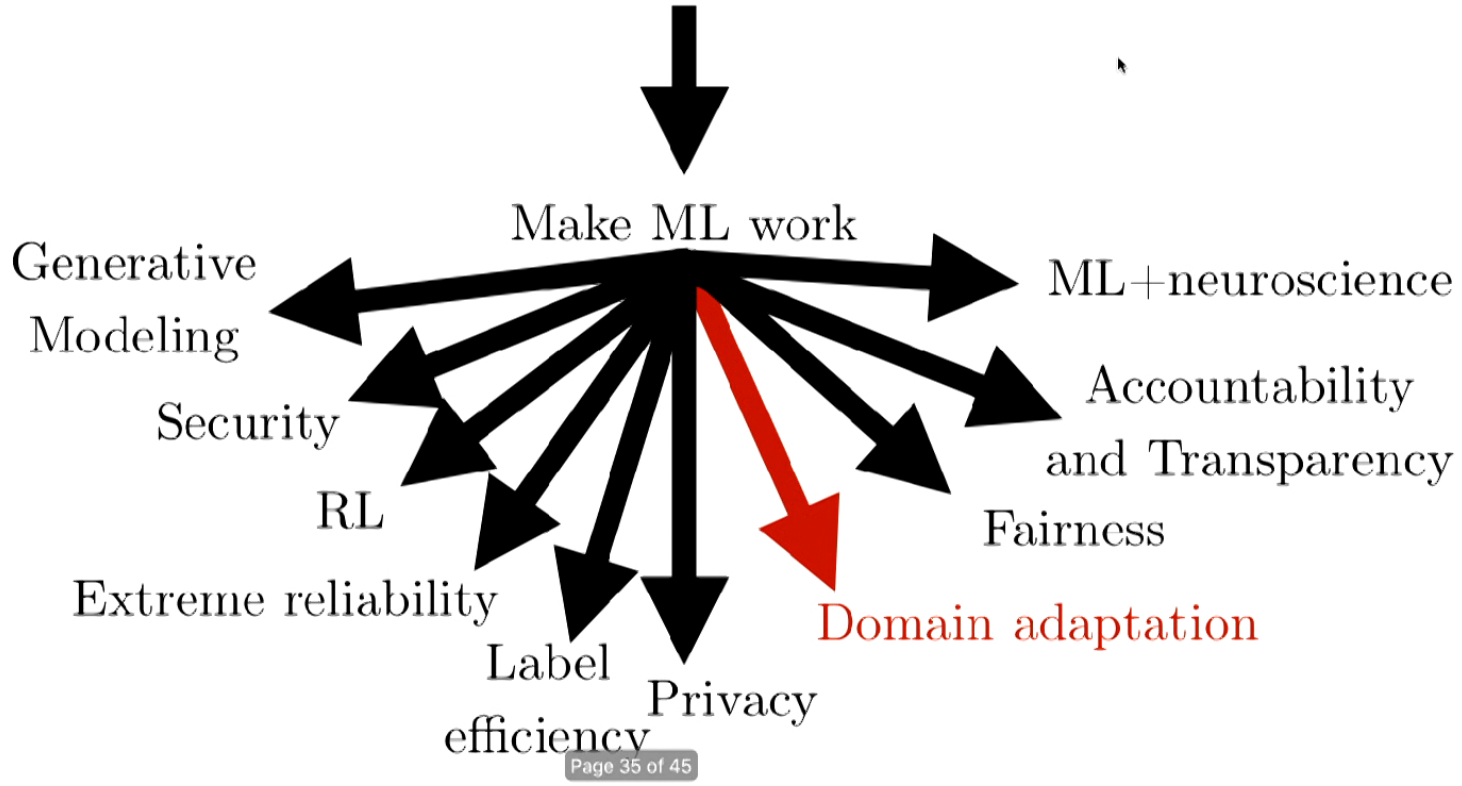
Defining (ε, δ)-Differential Privacy



$$\forall S: \Pr[M(x) \in S] \leq e^{\varepsilon} \cdot \Pr[M(y) \in S] + \delta$$

(Abadi 2017)

A Cambrian Explosion of Machine Learning Research Topics



(Goodfellow 2018)

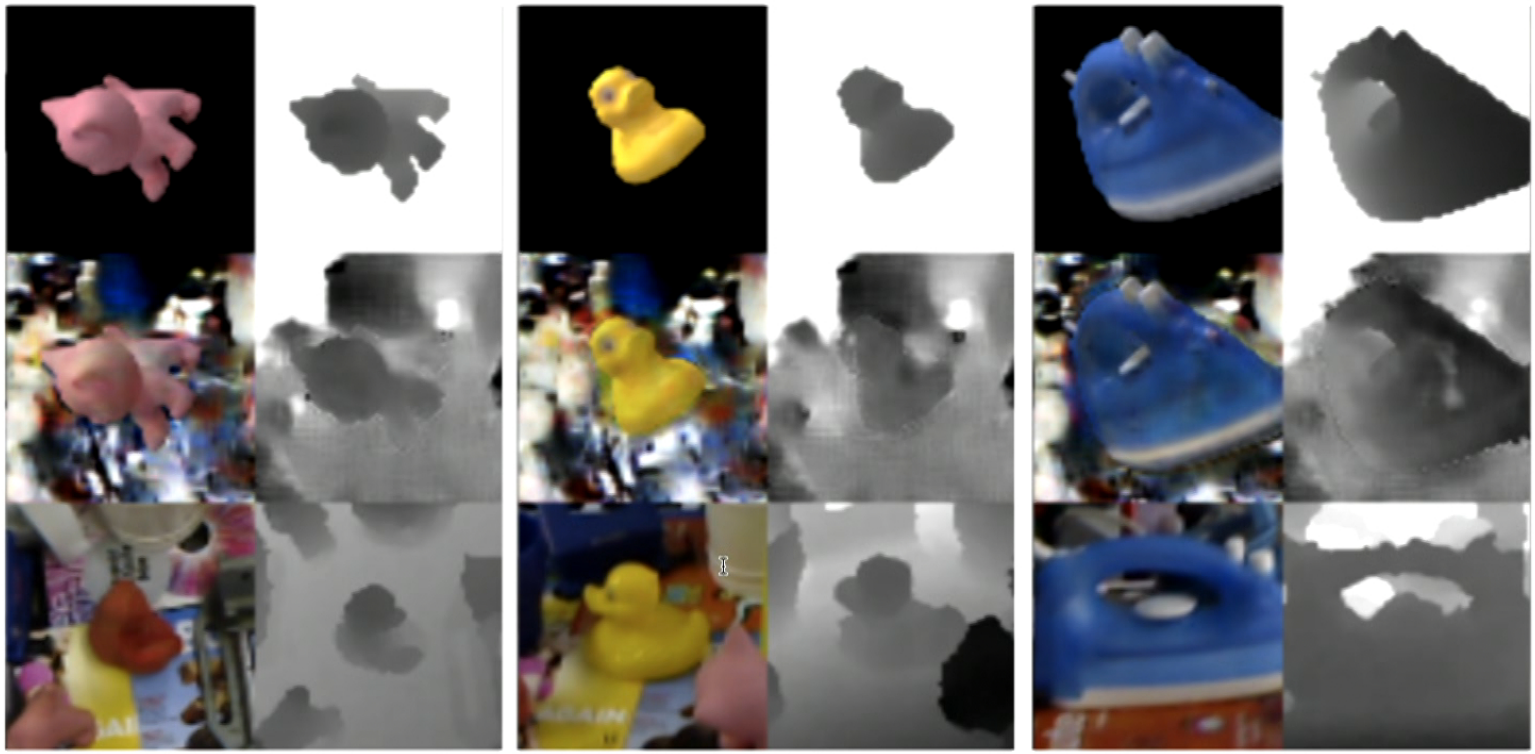
Domain Adaptation

- Domain Adversarial Networks (Ganin et al, 2015)

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

(Goodfellow 2018)

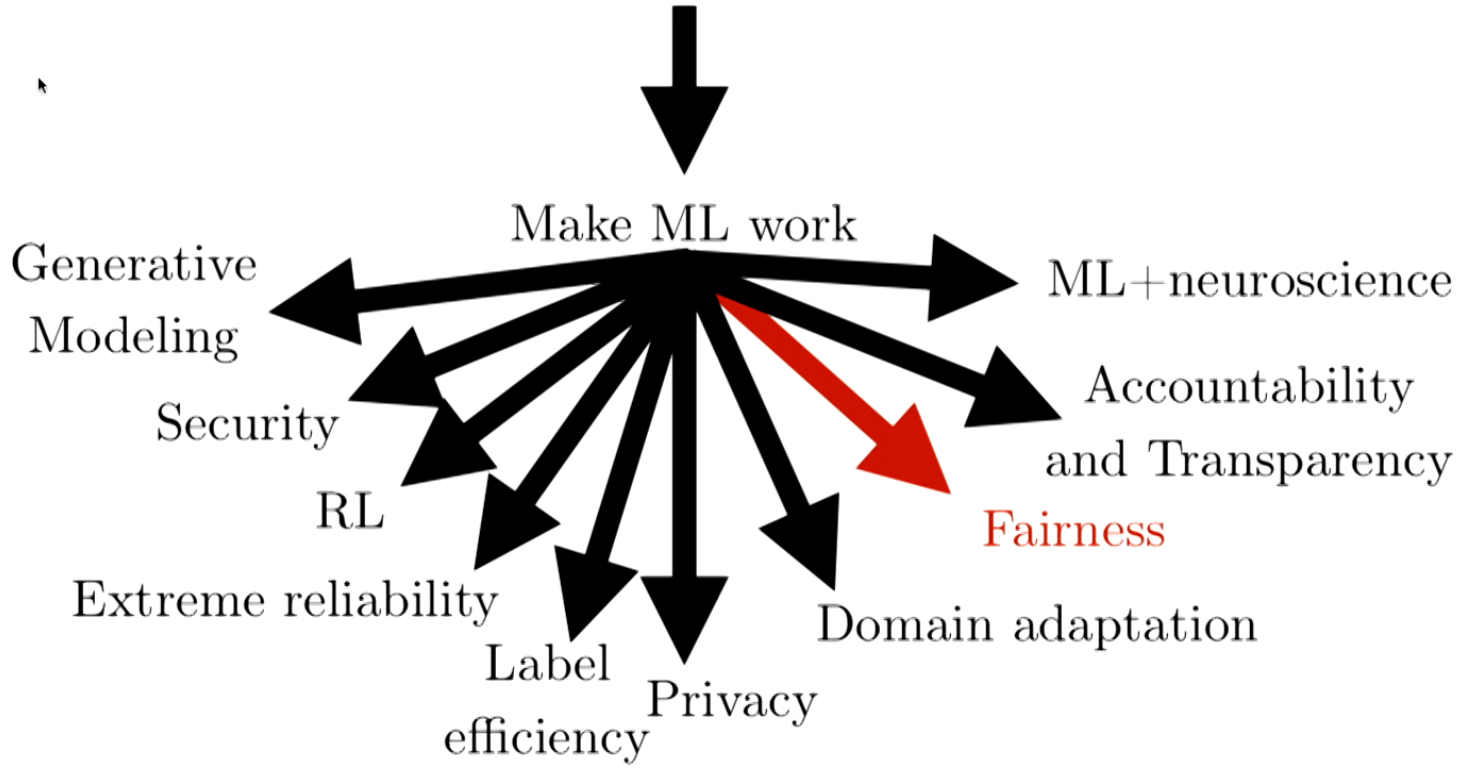
GANs for domain adaptation



(Bousmalis et al., 2016)

(Raffel, 2017)

A Cambrian Explosion of Machine Learning Research Topics



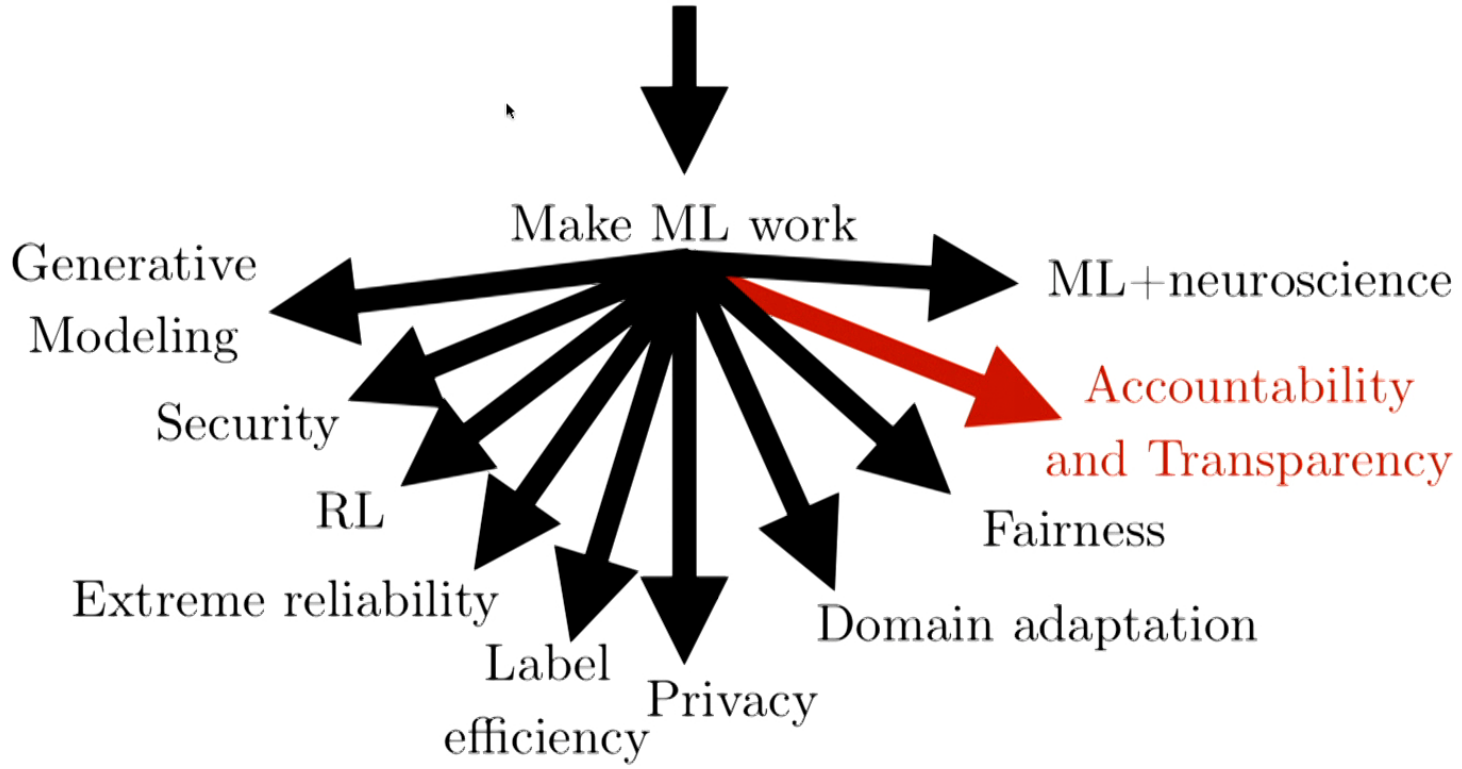
(Goodfellow 2018)

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

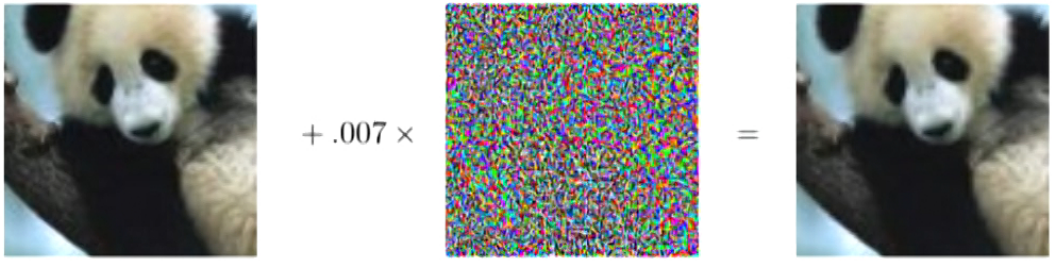
(Goodfellow 2018)

A Cambrian Explosion of Machine Learning Research Topics

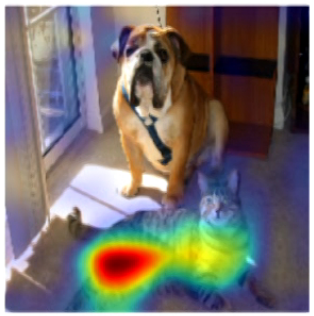


(Goodfellow 2018)

How do machine learning models work?



(Goodfellow et al, 2014)



(c) Grad-CAM 'Cat'

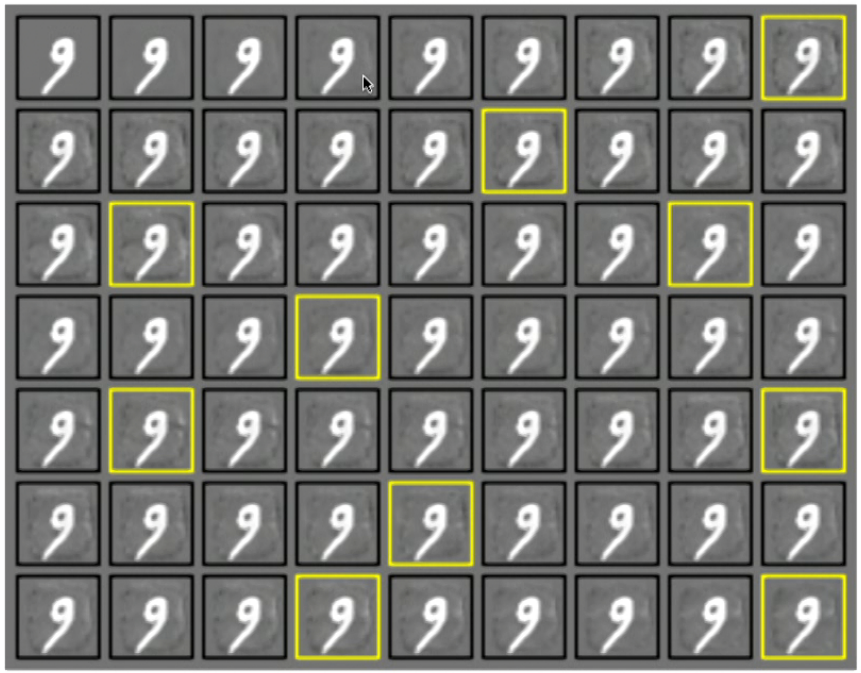
(i) Grad-CAM 'Dog'

(Selvaraju et al, 2016)

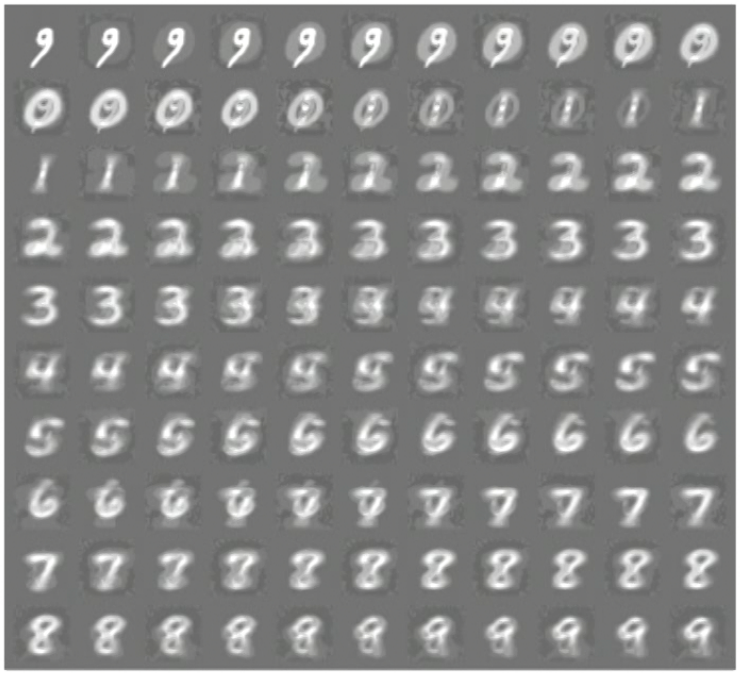
Interpretability literature: our analysis tools show that deep nets work about how you would expect them to.

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

Robust models are more interpretable



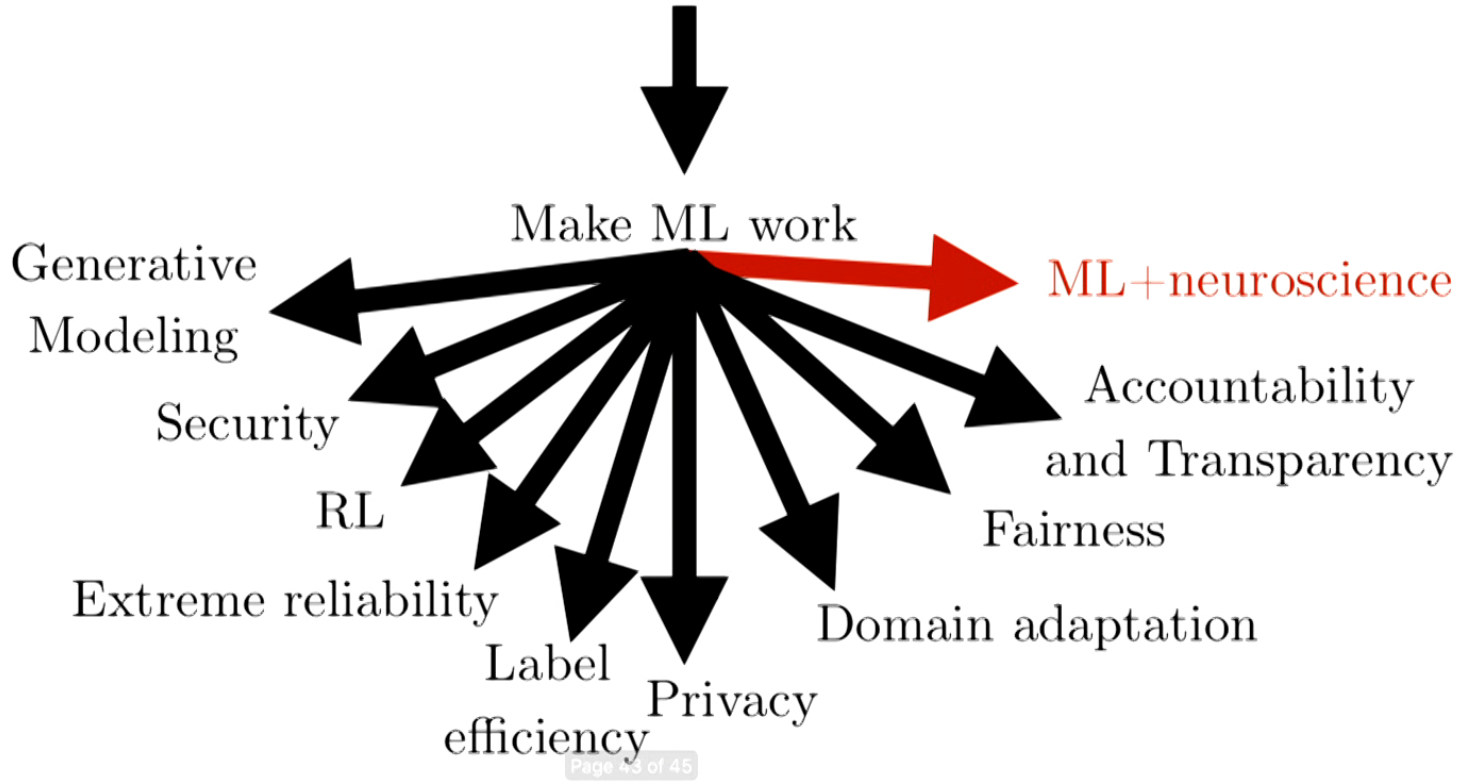
Relatively vulnerable model



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

(Goodfellow 2018)