Booz | Allen | Hamilton

GRAD: GRADIENT REVERSAL AGAINST DISCRIMINATION

A Fair Neural Network Learning Approach

Jared Sylvester, PhD

3 October 2018 | DSAA 2018 | Turin, Italy

WHY DO WE CARE?

OBLIGATORY TROUBLESOME AI HEADLINES

nature

COMMENT · 18 JULY 2018

AI can be sexist and racist — it's time to make it fair

Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.

James Zou & Londa Schiebinger

THE VERGE

TECH · SCIENCE · CULTURE · MORE

REPORT · SCIENCE · TECH · 67

The invention of AI 'gaydar' could be the start of something much worse

Researchers claim they can spot gay people from a photo, but critics say we're revisiting pseudoscience

By James Vincent | @jjvincent | Sep 21, 2017, 1:24pm EDT

Illustrations by Alex Castro

SHARE

The Economist

Subscribe

The Economist explains

The Economist explains

Why Uber's self-driving car killed a pedestrian

It was the first fatal accident of its kind

The Economist explains ›

May 29th 2018 | by T.S.

WHY DO WE CARE ABOUT AI ETHICS?

- It’s the right thing to do.

re and

al issu

g ethic

st for A

AI mea /techn

ing soc

  - Ethical concerns about AI are widespread even if you don’t share them.
  - If you want more AI in the world, you’ll need to assuage those fears in others.
  - This is true even if your domain doesn’t have obvious ethical implications.

WHY DO WE CARE ABOUT AI ETHICS?

- It's the right thing to do.
  - AI is affecting more and more of our lives.
  - Life is full of ethical issues.
  - ∴ AI is confronting ethical issues.
- Appeal to self-interest for AI practitioners:
  - Producing more AI means overcoming practical/technical problems.
  - But also overcoming social/PR problems.
  - Ethical concerns about AI are widespread even if you don't share them.
  - If you want more AI in the world, you'll need to assuage those fears in others.
  - This is true even if your domain doesn't have obvious ethical implications.

MORE OBLIGATORY TROUBLESOME AI HEADLINES

Forbes

Sep 24, 2018, 06:00am

Artificial Intelligence Can Reinforce Bias, Cloud Giants Announce Tools For AI Fairness

Paul Teich Contributor

Enterprise & Cloud

I write about new technologies and usage models transforming business.

TWEET THIS

Ultimately, the best answer to addressing bias in trained machine learning models will be to build transparent models.

Unfairly trained Artificial Intelligence (AI) systems can reinforce bias, therefore AI systems must be trained fairly. Experts say AI fairness is a dataset issue for each specific machine learning model. AI fairness is a newly recognized challenge.

THE GLOBE AND MAIL

SUBSCRIBE LOG IN

LEADERSHIP LAB

Is artificial intelligence sexist?

JODIE WALLIS
SPECIAL TO THE GLOBE AND MAIL
PUBLISHED SEPTEMBER 27, 2018
UPDATED 2 HOURS AGO

Managing director of AI at Accenture in Canada, host of The AI Effect podcast with Amber Mac, which launches Season 2 on Oct. 23

Artificial intelligence (AI) is bringing amazing changes to the workplace, and it's raising a perplexing question: Are those robots sexist?

While it may sound strange that AI could be gender-biased, there's evidence that it's happening when organizations aren't taking the right steps.

In the age of #MeToo and the drive to achieve gender parity in the workplace, it's critical to understand how and why this occurs and to continue to take steps to address the imbalance. At Accenture, a global professional services company, we have set a goal to have a gender-balanced work force by 2025. There is no shortage of examples that demonstrate how a diverse

FORTUNE

SUBSCRIBE

AI

How to Fight the Growing Scourge of Algorithmic Bias in AI

Researcher Joy Buolamwini has started the Algorithmic Justice League to combat algorithmic bias in AI and machine learning apps. Courtesy of Affectiva and Steve Nisotel Photography

By AARON PRESSMAN September 14, 2018

Joy Buolamwini was a graduate student at MIT a few years ago when she was working on an art and science project called the Aspire Mirror. The set up was

ASSESSING FAIRNESS

MEASURES OF FAIRNESS

GOAL: ACCURATE DECISIONS THAT ARE INVARIANT TO PROTECTED ATTRIBUTES

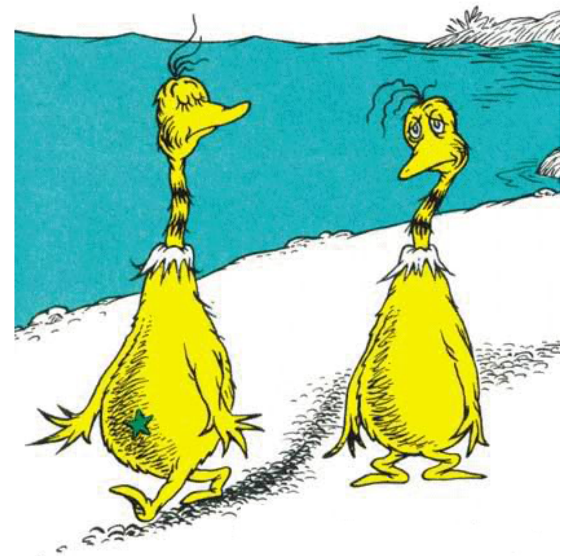
- e.g., Predict credit-worthiness, recidivism, job performance, etc. but do not consider race, gender, nationality, etc. in our decision.
- "Fairness through unawareness" is insufficient
  - Even if the protected attribute is completely removed from the dataset, other features may be highly correlated with it and function as proxies.

ID	Age	Name	Fav. Musician	Fav. Food	Vehicle
1	*REMOVED*	Ethel	Frank Sinatra	Tuna Casserole	Buick LaCrosse
2	*REMOVED*	Hermione	deadmau5	Quinoa	None (Uber)

- There are many ways to measure whether you've succeeded

MEASURES OF FAIRNESS

(WARNING: THIS IS A MINEFIELD)



- Discrimination
  - A.k.a. “group fairness” or “statistical parity”
  - Difference between average predicted scores for each protected attribute-value.
    (Average output for Star-Bellied Sneetches and average output for Smooth-Bellied Sneeches should be the same.)
- Problems include:
  - Allows discrimination within sub-populations
  - Can’t account for different base rates across groups

$$\text{Discrimination} = \frac{2}{k}\sum_{i=1}^{k}\left|\frac{\sum_{x_j \in T}\hat{y}_j}{|T|} - \frac{\sum_{x_j \in T_i}\hat{y}_j}{|T_i|}\right|$$

MEASURES OF FAIRNESS

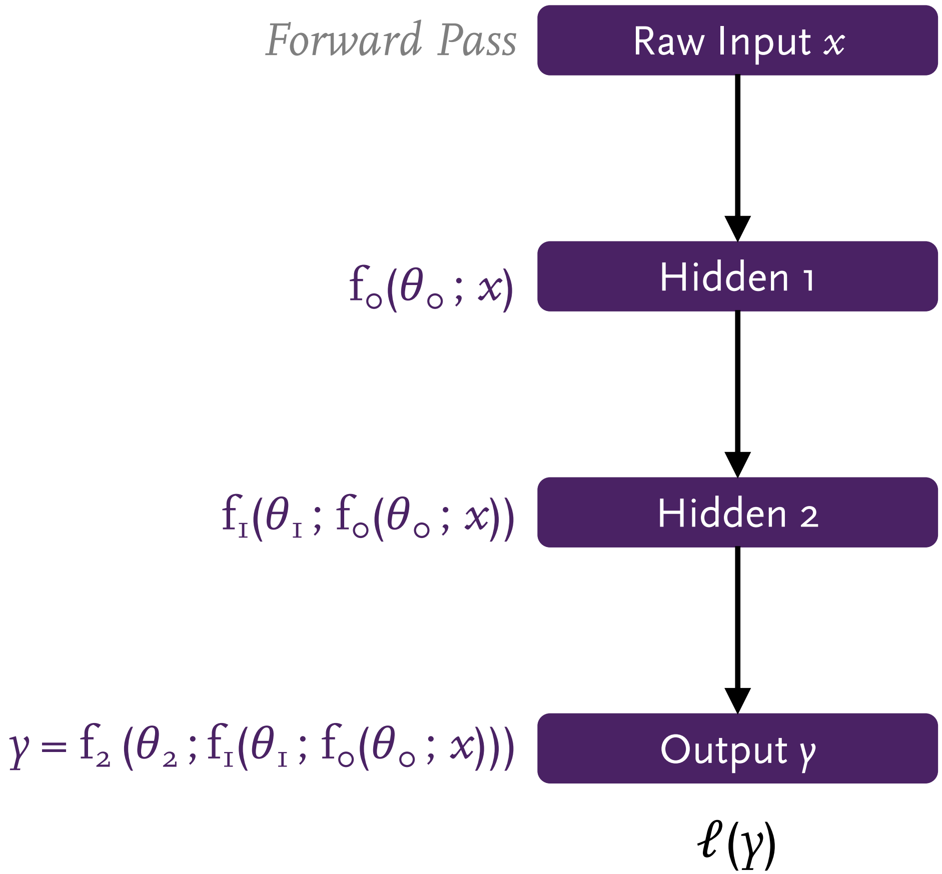
- Consistency
  - A.k.a. "individual fairness"
  - Similar samples should receive similar outputs.
    (A Star-Bellied Sneetch with PhD, five years of experience & 93% score on qualifying test should get the same output as a Smooth-Bellied Sneetch with PhD, five years of experience & 93% score on qualifying test.)
- Problems:
  - May still result in "headline figures" that seem quite unfair.
  - What does "similar samples" mean?

$$\text{Consistency} = 1 - \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - \frac{1}{k}\sum_{j \in k\cdot\mathrm{NN}(x_i)} \hat{y}_j\right|$$

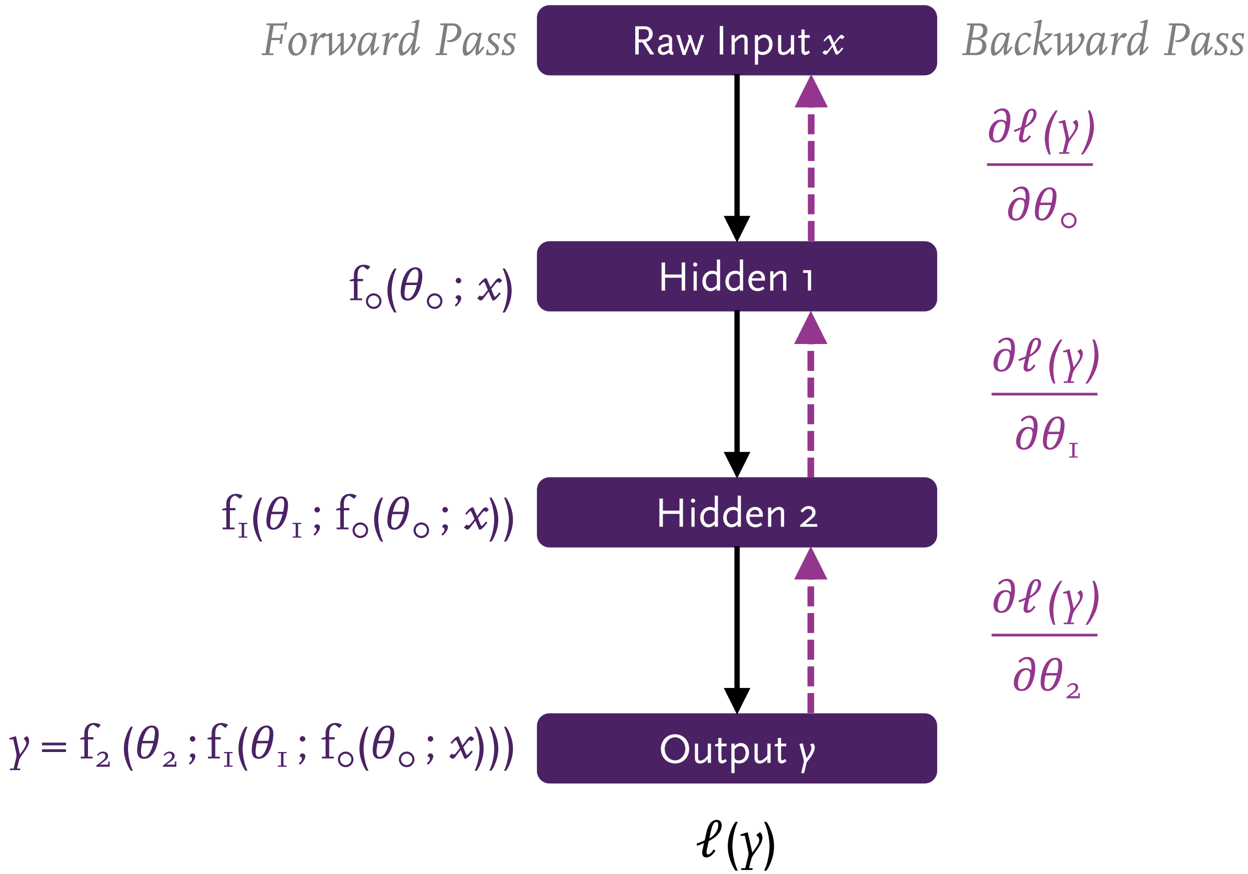
- Accuracy
  - We still want something usable
- Delta: Accuracy – Discrimination
  - Way to balance performance & fairness (though quite crude)

GRAD:
GRADIENT REVERSAL AGAINST DISCRIMINATION

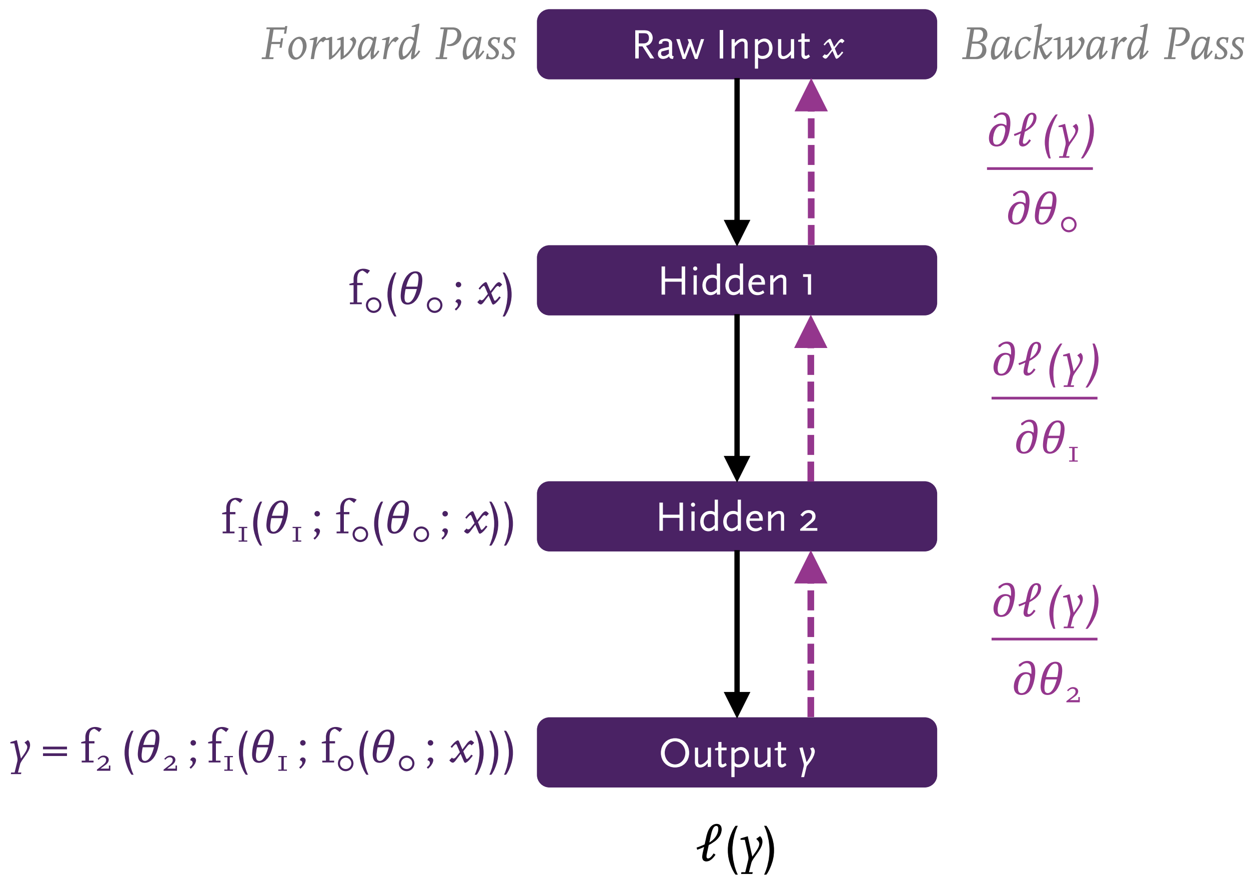
NEURAL NETS & GRADIENT DESCENT



NEURAL NETS & GRADIENT DESCENT

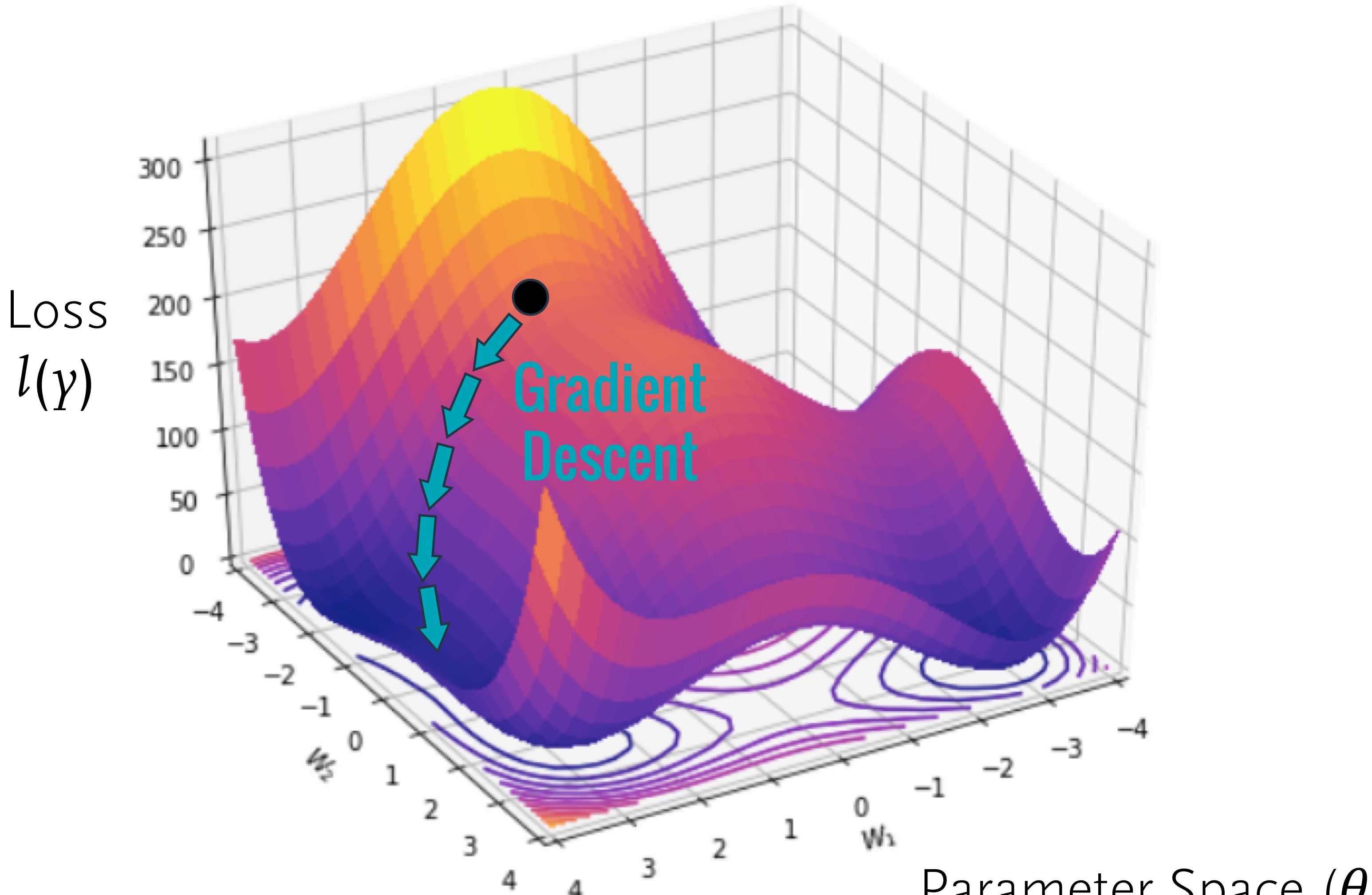


NEURAL NETS & GRADIENT DESCENT



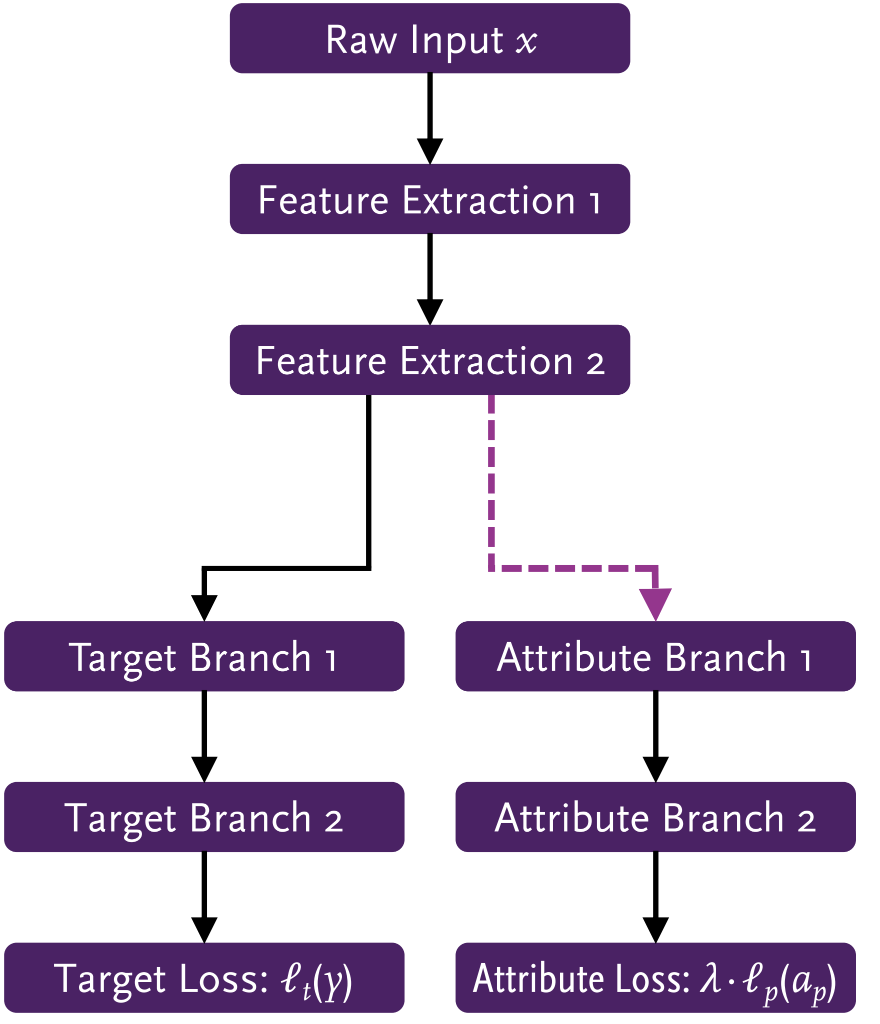
Guess Check Adjust

NEURAL NETS & GRADIENT DESCENT



GRAD

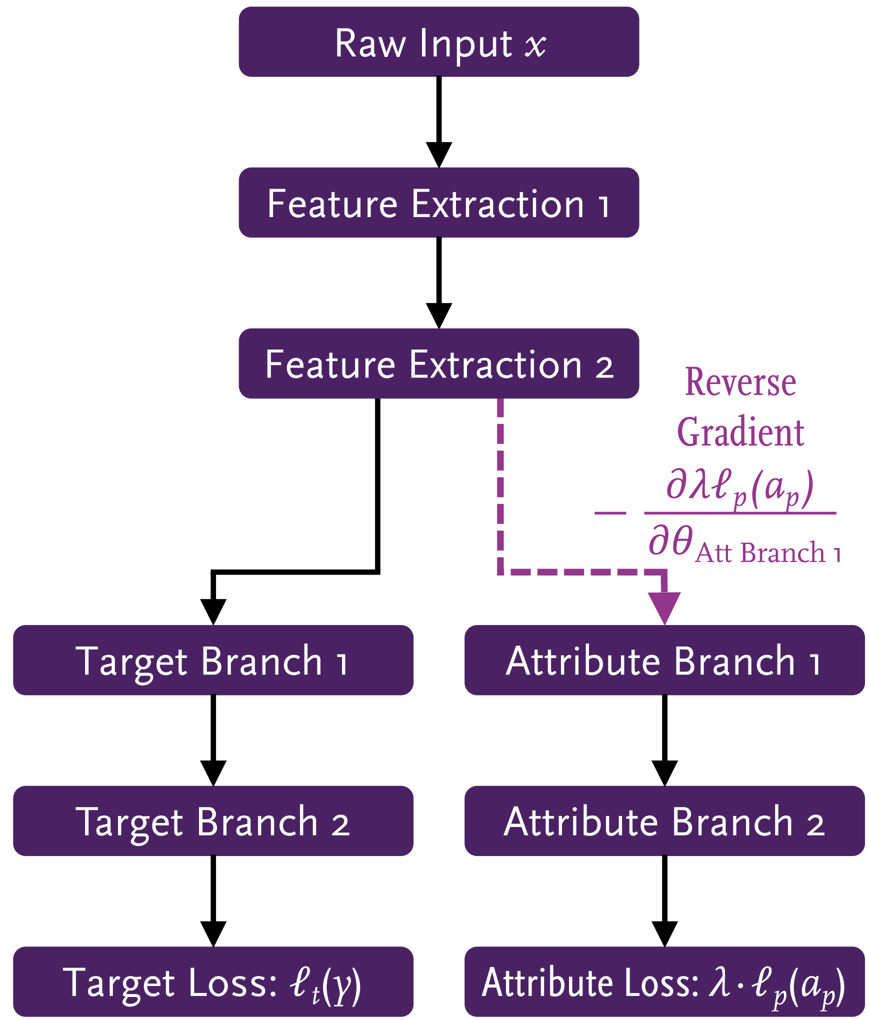
- Neural network architecture with two sets of outputs
  - Inspired by domain adaptation
- Several feature extraction layers (the "trunk"), followed by split into two "branches":
  - "*Target branch*" learns to predict target y
  - "*Attribute branch*" learns to predict protected attribute a_p
- Architecture agnostic:
  - Target branch can be either an autoencoder (GRAD–Auto) or a classifier/regressor (GRAD–Pred)



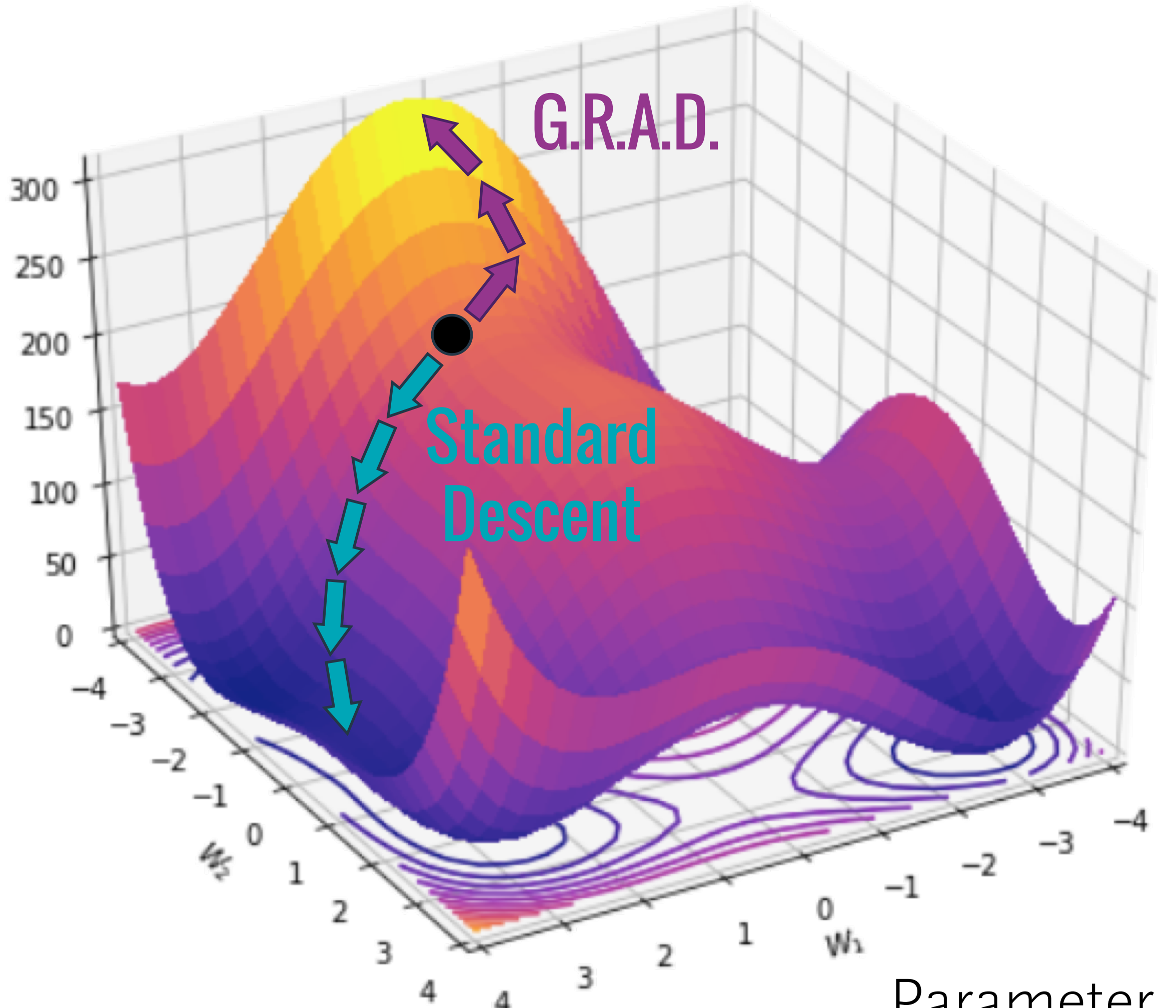
GRAD

$$\ell(y, a_p) = \ell_t(y) + \lambda \cdot l_p(a_p)$$

- Losses for both $\ell_t(y)$ and $\ell_p(a_p)$ are calculated and gradients are used for weight updates as normal…
- *Except: once propagated down the attribute branch, gradients are reversed (i.e. multiplied by -1) before being applied to the trunk.*
- Effect:
  - Network can still accurately predict target
  - Network moves away from optima in predictions of a_p
  - *Enforces ignorance of protected attribute*



GRADIENT ~~DESCENT~~ ASCENT



Loss
$l_p(a_p)$

Parameter Space

ARCHITECTURE FLEXIBILITY

GRAD–PRED

- Target branch outputs discrete class or regression value directly.
  - E.g. output creditworthiness.
- Allows greater task-specificity.

$$\ell^{\text{pred}}(\cdot) = \log\left(1 + \exp\left(-y \cdot h_{\text{target}}\right)\right)$$

GRAD–AUTO

- Target branch attempts to output representation of input (x) w/o sensitive feature ($\tilde{x}$).
  - $a_p \notin \tilde{x}$
- New representation is less biased version of input.
  - Train other classifiers on output. (e.g. Logistic Regression; same approach as LFR & VFAE)
  - Distribute data to others.
- Allows maximum flexibility.

$$\ell^{\text{auto}}(\cdot) = \|h_{\text{target}} - \widetilde{x}\|_2^2$$

METHODS

- Data sets (Zemel et al., 2013 / Edwards & Storkey, 2016)
  - German Credit
  - Adult Income
  - Heritage Health
  - Diabetes
- Comparison techniques:
  - Baseline neural nets
    (same architecture as GRAD-Pred & GRAD-Auto, but no Attribute Branch)
  - LRF: Fair Logistic Regression (Kamishima et al., 2011)
  - NBF: Fair Naive Bayes (Kamiran & Calders, 2009)
  - FF: Fair Random Forests (Raff, Sylvester & Mills, 2018)
  - LFR: Learning Fair Representations (Zemel et al., 2013)
  - VFAE: Variation Fair Auto-Encoders (Louizos, 2016)
  - ALFR: Adversarial Learned Fair Representations (Edwards & Storkey, 2016)

RESULTS

Results on the Heritage Health dataset. Best results in bold, second best in italics.

NN = standard neural nets; NBF = fair Naïve Bayes; FF=Fair Forests; LR = Logistic Regression; LRF = fair Logistic Regression; LFR = Learned Fair Representations; VFAE = Variational Fair Autoencoders.

Algorithm	Acc	Delta	Discr	Cons
NN-Auto	**0.8506**	0.7939	0.0567	0.9730
GRAD-Auto	0.8491	**0.8491**	**0.0000**	**1.0000**
NN-Pred	0.8440	0.7511	0.0929	0.9453
GRAD-Pred	*0.8493*	0.8486	*0.0007*	*0.9999*
NBF	0.6878	0.5678	0.1200	0.5893
FF	0.8474	0.8474	**0.0000**	**1.0000**
LR	0.7547	0.6482	0.1064	0.7233
LRF	0.7212	0.7038	0.0174	0.6223
LFR	0.7365	0.7365	**0.0000**	**1.0000**
VFAE	0.8490	*0.8490*	**0.0000**	—

RESULTS

- GRAD is typically best or 2nd best in each metric
- Competitive with prior methods in all metrics
- Both GRAD–Auto & GRAD–Pred reliably produce very high Consistency scores
  - One of the two is always the best in Consistency
- Capable of achieving Discrimination=0.00 & Consistency=1.00

Results on the Heritage Health dataset. Best results in bold, second best in italics.

NN = standard neural nets; NBF = fair Naïve Bayes; FF=Fair Forests; LR = Logistic Regression; LRF = fair Logistic Regression; LFR = Learned Fair Representations; VFAE = Variational Fair Autoencoders.

Algorithm	Acc	Delta	Discr	Cons
NN-Auto	**0.8506**	0.7939	0.0567	0.9730
GRAD-Auto	0.8491	**0.8491**	**0.0000**	**1.0000**
NN-Pred	0.8440	0.7511	0.0929	0.9453
GRAD-Pred	*0.8493*	0.8486	*0.0007*	*0.9999*
NBF	0.6878	0.5678	0.1200	0.5893
FF	0.8474	0.8474	**0.0000**	**1.0000**
LR	0.7547	0.6482	0.1064	0.7233
LRF	0.7212	0.7038	0.0174	0.6223
LFR	0.7365	0.7365	**0.0000**	**1.0000**
VFAE	0.8490	*0.8490*	**0.0000**	—

RESULTS: MULTIPLE ATTRIBUTES

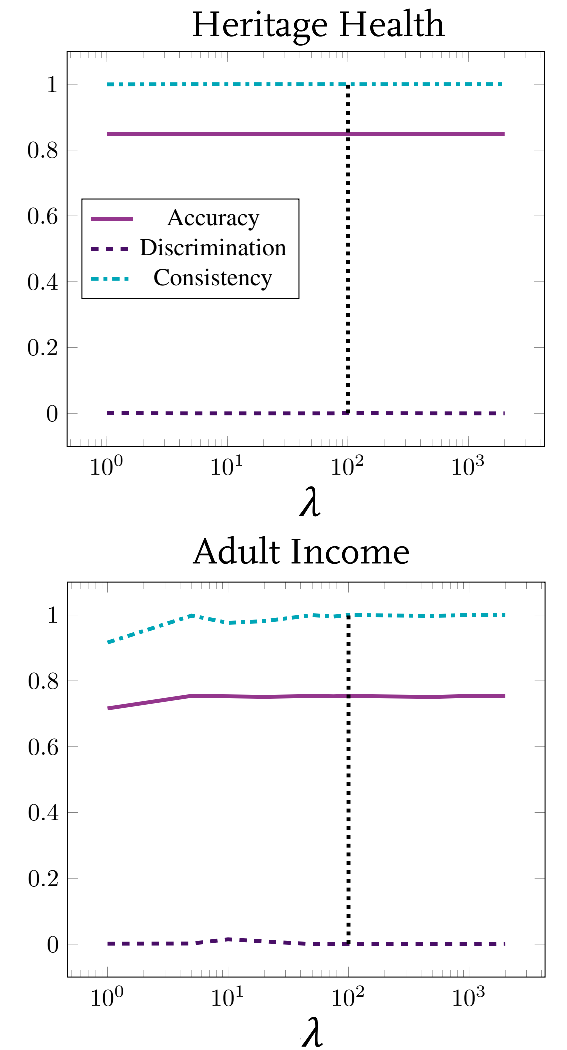
- ∃ 11 federal protected classes in the US
- What if >1 occurs in the dataset?
  - Not a hypothetical question
- No prior work has rigorously examined protecting multiple attributes.
- Protecting one attribute causes decreased fairness w.r.t. the other:

Need to explicitly protect both at once.

			Discrimination		
Algorithms	Acc	Delta	Race	Gender	Cons
NN-Auto	0.5735	0.5392	0.0412	0.0275	0.6411
GRAD-Auto	0.5765	0.5723	*0.0055*	**0.0030**	0.6288
NN-Pred	**0.6286**	*0.5848*	0.0418	0.0458	*0.6464*
GRAD-Pred	*0.5980*	**0.5949**	**0.0028**	*0.0034*	**0.7180**
GRAD-Auto-R	0.5851	0.5749	0.0003	0.0201	0.6404
GRAD-Auto-G	0.5640	0.5143	0.0981	0.0013	0.6093
GRAD-Pred-R	0.5844	0.5478	0.0020	0.0713	0.7538
GRAD-Pred-G	0.5941	0.5526	0.0785	0.0045	0.6849

RESULTS: ROBUSTNESS TO LAMBDA

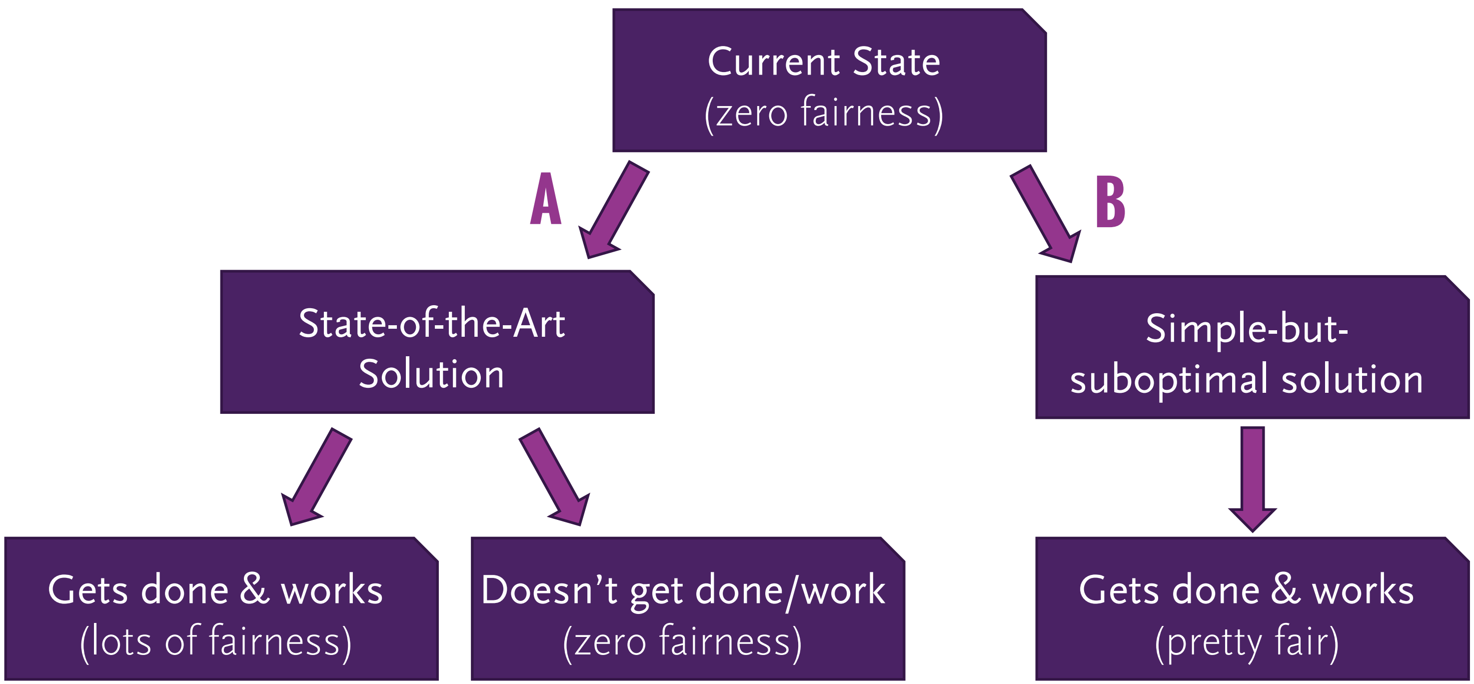
- λ controls trade-off between goals of predicting y and not predicting a_p
- Didn't do any hyper-parameter search for λ
  - Used λ=100 for all experiments
  - Keeps things simple for practitioners
- Any values in [20,1000] would have been acceptable



WHY SIMPLICITY MATTERS

- AI is hard.
- AI practitioners have many competing demands.
- If Fair AI solutions are too difficult in practice, they won't get built.
- Shipping is a feature:

A perfect solution that isn't/can't be implemented will never make the world more fair.



GRAD CONCLUSIONS

- Simple to implement
- Requires only one (insensitive) hyper-parameter
- Applicable to any architecture:
  - Autoencoders
  - Direct predictive networks
  - Allows trade-off between generality and specificity
- Competitive with other approaches
- The first neural network shown to protect multiple attributes concurrently

THANK YOU

Booz | Allen | Hamilton

For more information, please contact us:

Jared Sylvester

@jsylvest

sylvester_jared@bah.com

Edward Raff

@EdwardRaffML

Raff_edward@bah.com

Raff & Sylvester. *"Gradient Reversal Against Discrimination."* Fairness, Accountability & Transparency in Machine Learning (FAT/ML). 2018.

arxiv.org/abs/1807.00392

Raff, Sylvester & Mills. *"Fair Forests: Regularized Tree Induction to Minimize Model Bias."* AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES). 2018.

arxiv.org/abs/1712.08197

Sylvester & Raff. *"What about applied fairness?"* ICML: The Debates. 2018.

arxiv.org/abs/1806.05250

BOOZALLEN.COM/MACHINEINTELLIGENCE

CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER