CSC 2515 Lecture 12: Algorithmic Fairness

Roger Grosse

University of Toronto

Final Exam

- Tuesday, Dec. 17, from 3–6pm, in the Banting Institute, room 131.
- Covers up through Lecture 11 (i.e. last week), with slightly more emphasis on the second half of the course.
- Similar in format and difficulty to midterm
- You are only responsible for material covered in lecture, but topics additionally covered in tutorials and homeworks will receive more emphasis.
- Closed book, no aids permitted.
- Practice exams will be posted.
- **Use blue or black ink.**
- **Please bring photo ID.**

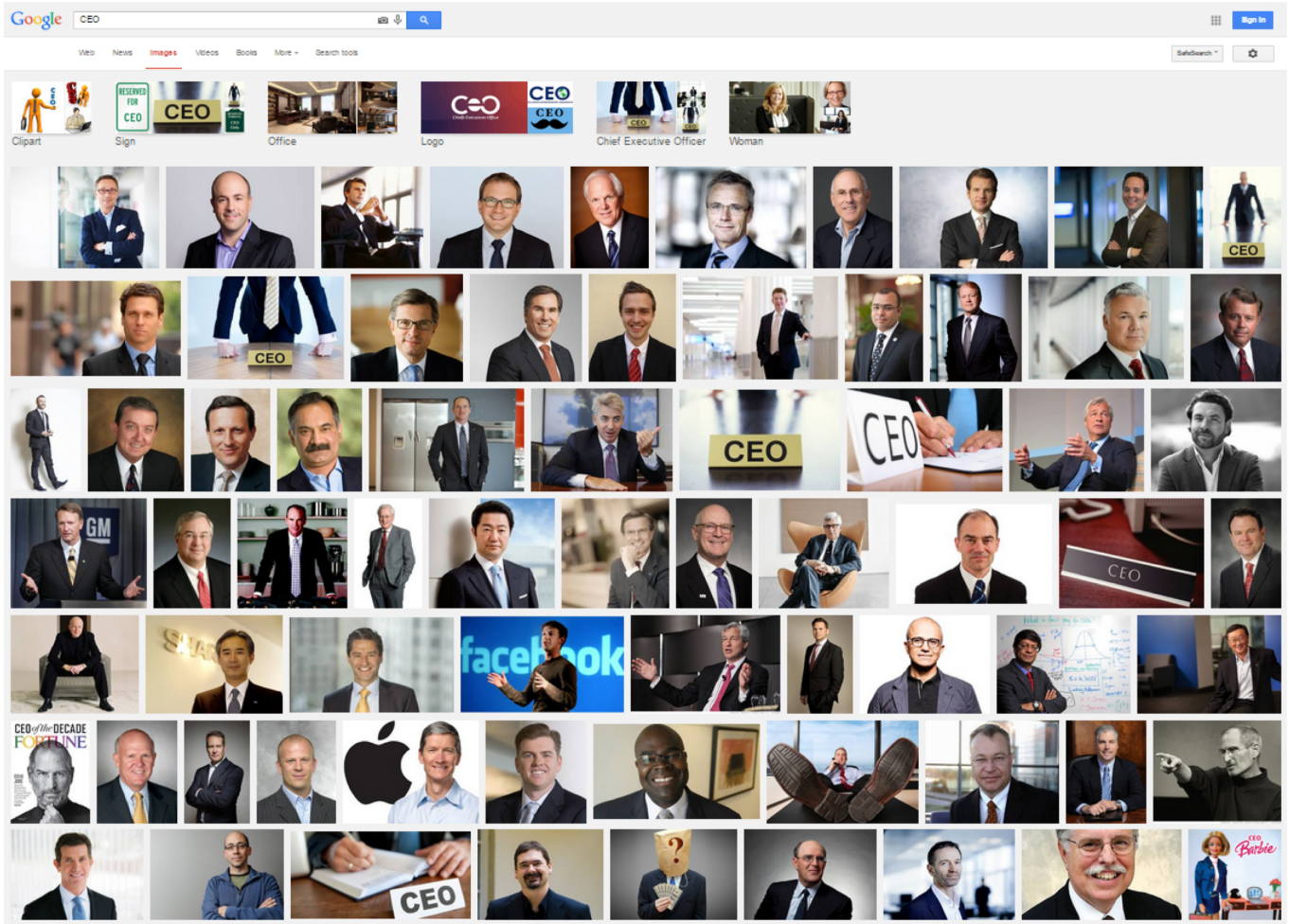
WHY WAS I NOT SHOWN THIS AD?

Credit: Richard Zemel

FAIRNESS IN AUTOMATED DECISIONS

Credit: Richard Zemel

SUBTLER BIAS



Credit: Richard Zemel

Overview: Fairness

- This lecture: algorithmic fairness
- Goal: identify and mitigate bias in ML-based decision making, in all aspects of the pipeline
- Sources of bias/discrimination
  - Data
    - Imbalanced/impoverished data
    - Labeled data imbalance (more data on white recidivism outcomes)
    - Labeled data incorrect / noisy (historical bias)
  - Model
    - ML prediction error imbalanced
    - Compound injustices (Hellman)

Credit: Richard Zemel

Overview: Fairness

- Notation
  - X: input to classifier
  - S: sensitive feature (age, gender, race, etc.)
  - Z: latent representation
  - Y: prediction
  - T: true label
- We use capital letters to emphasize that these are random variables.

Fairness Criteria

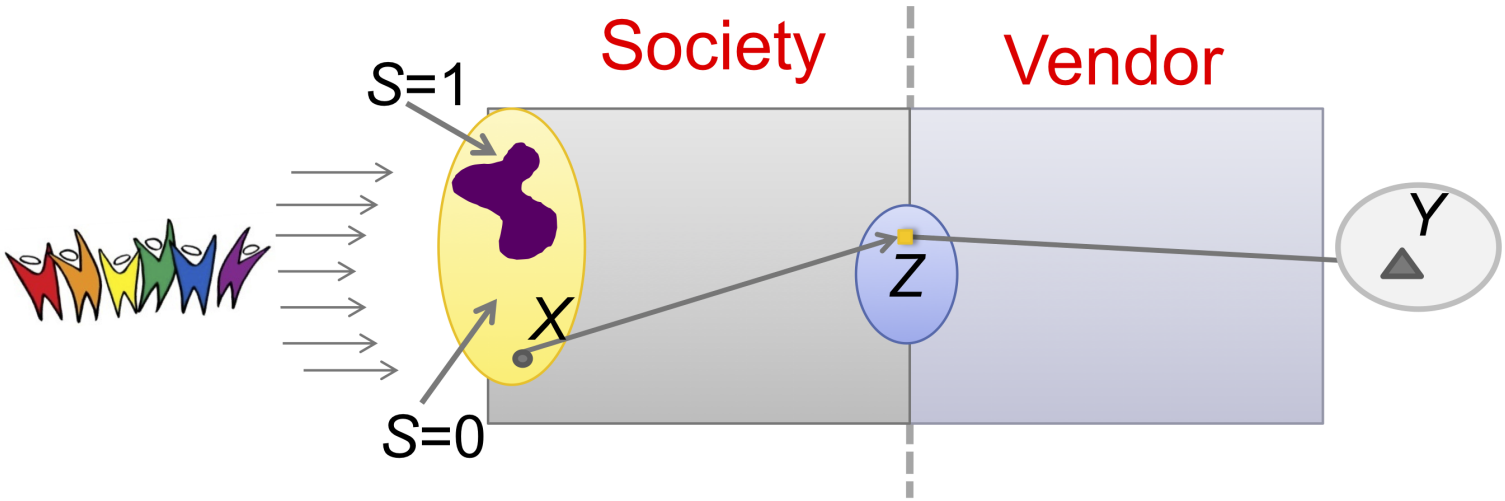
- Most common way to define fair classification is to require some invariance with respect to the sensitive attribute
  - Demographic parity: $Y \perp\!\!\!\perp S$
  - Equalized odds: $Y \perp\!\!\!\perp S \mid T$
  - Equal opportunity: $Y \perp\!\!\!\perp S \mid T = t$, for some t
  - Equal (weak) calibration: $T \perp\!\!\!\perp S \mid Y$
  - Equal (strong) calibration: $T \perp\!\!\!\perp S \mid Y$ and $Y = \Pr(T = 1)$
  - Fair subgroup accuracy: $\mathbb{1}[T = Y] \perp\!\!\!\perp S$
- $\perp\!\!\!\perp$ denotes stochastic independence
- Many of these definitions are incompatible!

Credit: Richard Zemel

Learning Fair Representations

Learning Fair Representations

- Idea: separate the responsibilities of the (trusted) society and (untrusted) vendor



- Goal: find a representation Z that removes any information about the sensitive attribute
- Then the vendor can do whatever they want!

Image Credit: Richard Zemel

Learning Fair Representations

- A naïve attempt: simply don't use the sensitive feature.
  - Problem: the algorithm implicitly learn to predict the sensitive feature from other features (e.g. race from zip code)
- Another idea: limit the algorithm to a small set of features you're pretty sure are safe and task-relevant
  - This is the conservative approach, and commonly used for both human and machine decision making
  - But removing features hurts the classification accuracy. Maybe we can make more accurate decisions if we include more features and somehow enforce fairness algorithmically?
- Can we learn fair representations, which can make accurate classifications without implicitly using the sensitive attribute?

Learning Fair Representations

Desiderata for the representation:

Retain information about X	$\Rightarrow$	high mutual information between X and Z
Obfuscate S	$\Rightarrow$	low mutual information between S and Z
Allow high classification accuracy	$\Rightarrow$	high mutual information between T and Z

Learning Fair Representations

First approach: Zemel et al., 2013, "Learning fair representations"

- Let Z be a discrete representation (like K-means)
- Determine Z stochastically based on distance to a prototype for the cluster (like the cluster center in K-means)

$$\Pr(Z = k \mid \mathbf{x}) \propto \exp(-d(\mathbf{x}, \mathbf{v}_k)),$$

  where d is some distance function (e.g. Euclidean distance)
- Use the Bayes classifier $y = \Pr(T = 1 \mid Z)$
- Need to fit the prototypes $\mathbf{v}_k$

Learning Fair Representations

- Retain information about X: penalize reconstruction error

$$\mathcal{L}_{\text{reconst}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \tilde{\mathbf{x}}^{(i)}\|^2$$

- Predict accurately: cross-entropy loss

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{i=1}^{N} -t^{(i)} \log y^{(i)} - (1 - t^{(i)}) \log(1 - y^{(i)})$$

- Obfuscate S:

$$\mathcal{L}_{\text{discrim}} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{1}{N_0} \sum_{i:s^{(i)}=0} \Pr(Z = k \mid \mathbf{x}^{(i)}) - \frac{1}{N_1} \sum_{i:s^{(i)}=1} \Pr(Z = k \mid \mathbf{x}^{(i)}) \right|,$$

where we assume for simplicity $S \in \{0, 1\}$ and N_0 is the count for $s = 0$.

Learning Fair Representations

- Obfuscate S:

$$\mathcal{L}_{\text{discrim}} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{1}{N_0}\sum_{i:s^{(i)}=0}\Pr(Z=k\,|\,\mathbf{x}^{(i)}) - \frac{1}{N_1}\sum_{i:s^{(i)}=1}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\right|,$$

- Is this about individual-level or group-level fairness?
- If discrimination loss is 0, we satisfy demographic parity

$$\begin{aligned}
\Pr(Y=1\,|\,s^{(i)}=1) &= \frac{1}{N_1}\sum_{i:s^{(i)}=1}\sum_{k=1}^{K}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\Pr(Y=1\,|\,Z=k) \\
&= \sum_{k=1}^{K}\left[\frac{1}{N_1}\sum_{i:s^{(i)}=1}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\right]\Pr(Y=1\,|\,Z=k) \\
&= \sum_{k=1}^{K}\left[\frac{1}{N_0}\sum_{i:s^{(i)}=0}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\right]\Pr(Y=1\,|\,Z=k) \\
&= \Pr(Y=1\,|\,s^{(i)}=0)
\end{aligned}$$

Learning Fair Representations

Datasets

1. **German Credit**

   Task: classify individual as good or bad credit risk

   Sensitive feature: Age

2. **Adult Income**

   Size: 45,222 instances, 14 attributes

   Task: predict whether or not annual income > 50K

   Sensitive feature: Gender

3. **Heritage Health**

   Size: 147,473 instances, 139 attributes

   Task: predict whether patient spends any nights in hospital

   Sensitive feature: Age

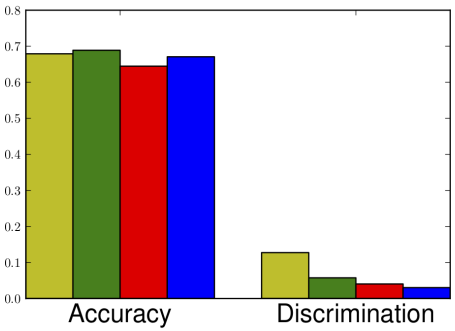
Learning Fair Representations

Metrics

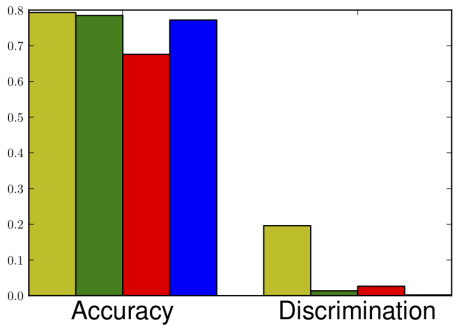
- Classification accuracy
- Discrimination

$$\left| \frac{\sum_{i:s^{(i)}=1}^{N} y^{(i)}}{N_1} - \frac{\sum_{i:s^{(i)}=0}^{N} y^{(i)}}{N_0} \right|$$

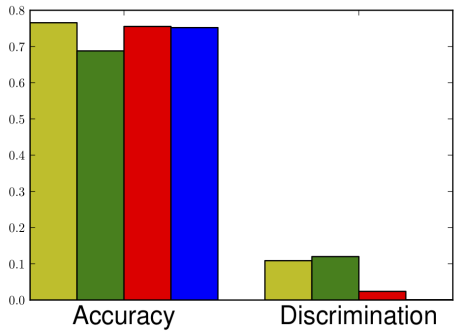
German



Adult



Health



Yellow = unrestricted; Blue = theirs

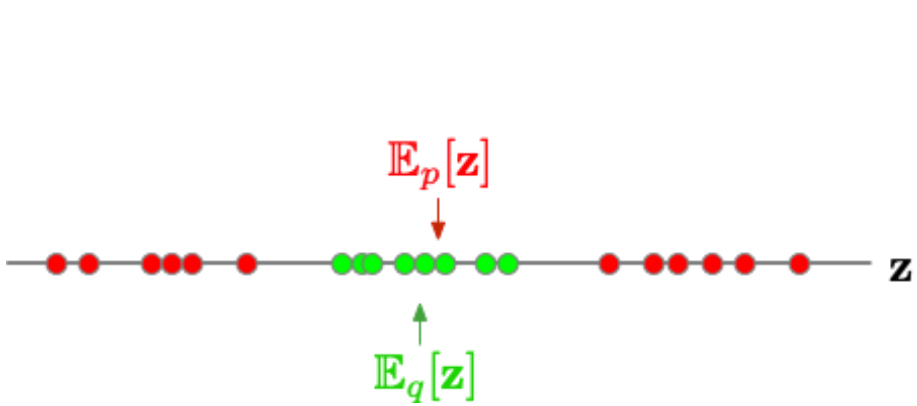
Fair VAE

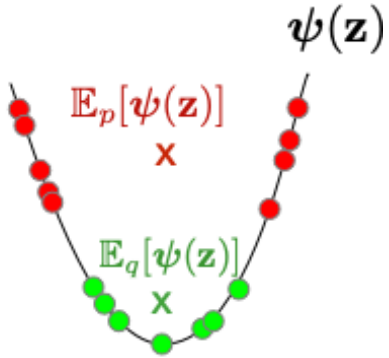
- Discrete Z based on prototypes is very limiting. Can we learn a more flexible representation?
- Louizos et al., 2015, "The variational fair autoencoder"
- The variational autoencoder (VAE) is a kind of autoencoder that represents a probabilistic model, and can be trained with a variational objective similar to the one we used for E-M.
  - For this lecture, just think of it as an autoencoder.
  - How can we learn an autoencoder such that the code vector $\mathbf{z}$ loses information about s?

Fair VAE: Maximum Mean Discrepancy

- Our previous non-discrimination criterion only makes sense for discrete Z.
- New criterion: ensure that $p(Z \mid s)$ is indistinguishable for different values of s.
- Maximum mean discrepancy (MMD) is a quantitative measure of distance between two distributions. Pick a feature map ψ.

$$\mathrm{MMD}(p; q) = \left\| \mathbb{E}_{\mathbf{z} \sim p}[\psi(\mathbf{z})] - \mathbb{E}_{\mathbf{z} \sim q}[\psi(\mathbf{z})] \right\|^2$$

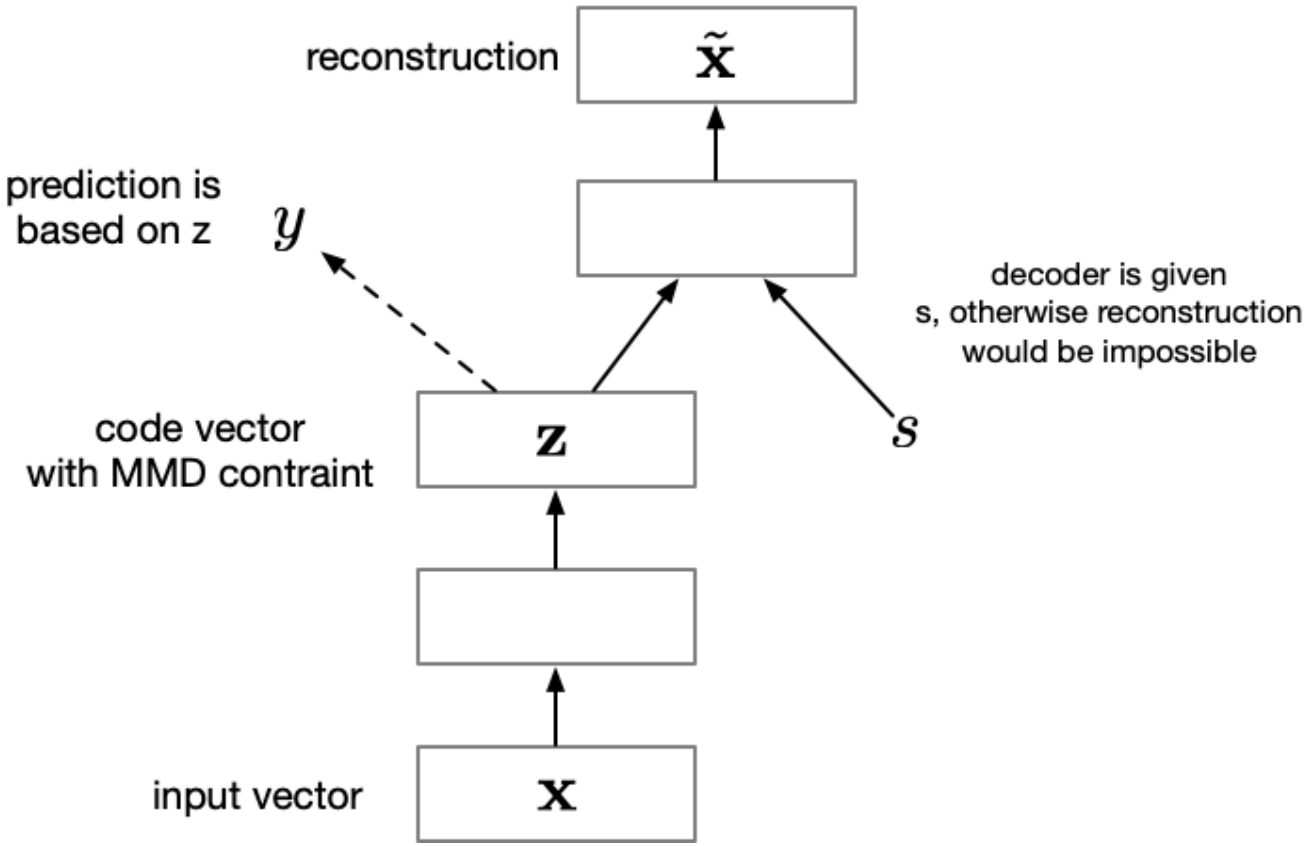




- If ψ is sufficiently expressive, then the MMD is only 0 if the distributions match. (Making this precise requires the idea of *kernels*.)

Fair VAE

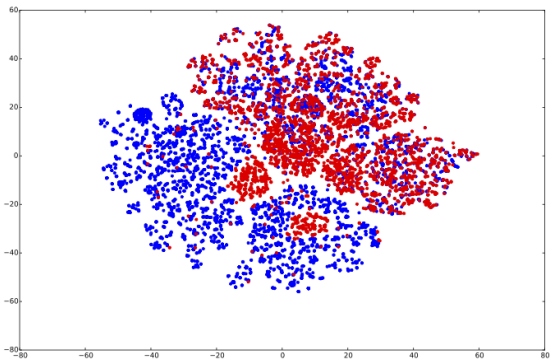
Train a VAE, with the constraint that the MMD between $p(\mathbf{z} \mid s = 0)$ and $p(\mathbf{z} \mid s = 1)$ is small.



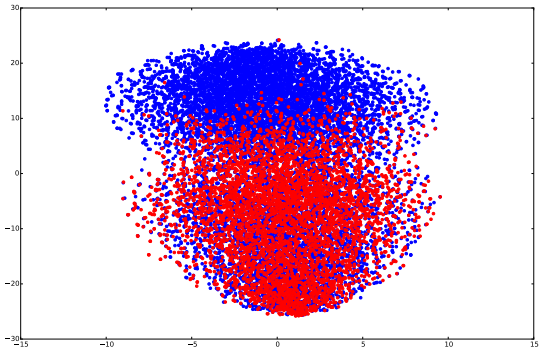
Fair VAE: tSNE embeddings

- tSNE is an unsupervised learning algorithm for visualizing high-dimensional datasets. It tries to embed points in low dimensions in a way that preserves distances as accurately as possible.
- Here are tSNE embeddings of different distributions, color-coded by the sensitive feature:

Original inputs



VAE latent space



Fair VAE latent space

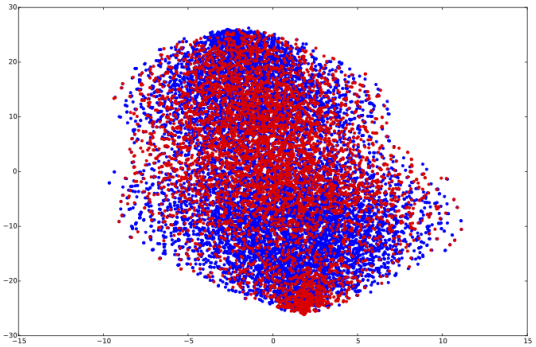


Figure Credit: Louizos et al., 2015

Individual Fairness

Individual Fairness

- The work on fair representations was geared towards group fairness
- Another notion of fairness is individual level: ensuring that similar individuals are treated similarly by the algorithm
  - This depends heavily on the notion of "similar".
- One way to define similarity is in terms of the "true label" T (e.g. whether this individual is in fact likely to repay their loan)
  - Can you think of a problem with this definition?
  - The label may itself be biased
    - if based on human judgments
    - if, e.g., societal biases make it harder for one group to pay off their loans
  - We'll ignore this issue in our analysis. But keep in mind that you'd need to carefully consider the assumptions when applying one of these methods!

Equal Opportunity

- Now we'll turn to Hardt et al., 2016, "Equality of opportunity in supervised learning".
- Assume we make a binary prediction by computing a real-valued score $R = f(X, S)$, and then thresholding this score to obtain the prediction Y.
- As before, assume $S \in \{0, 1\}$.
- Motivating example: predict whether an individual is likely to repay their loan
- Two notions of individual fairness:
  - Equalized odds: equal false positive and false negative rates

$$\Pr(Y = 1 \mid S = 0, T = t) = \Pr(Y = 1 \mid S = 1, T = t) \quad \text{for } t \in \{0, 1\}$$

  - Equal opportunity: equal false negative rates

$$\Pr(Y = 1 \mid S = 0, T = 1) = \Pr(Y = 1 \mid S = 1, T = 1)$$

Equal Opportunity

- Consider derived predictors, which are a function of the real-valued score R and the sensitive feature S.
  - I.e., we don't need to check the original input X. This simplifies the analysis.
- Define a loss function $\mathcal{L}(Y, T)$. Since Y and T are binary, there are 4 values to specify.
- They show that:
  - Without a constraint, the optimal predictor is obtained from thresholding R.
  - With an equal opportunity constraints, the optimal predictor is obtained by thresholding R, but with a different treshold for different values of S.
  - Satisfying equalized odds is overconstrained, and may require randomizing Y.

Equal Opportunity

- Case study: FICO scores
- Aim to predict whether an individual has less than an 18% rate of default (which is the treshold for profitability)

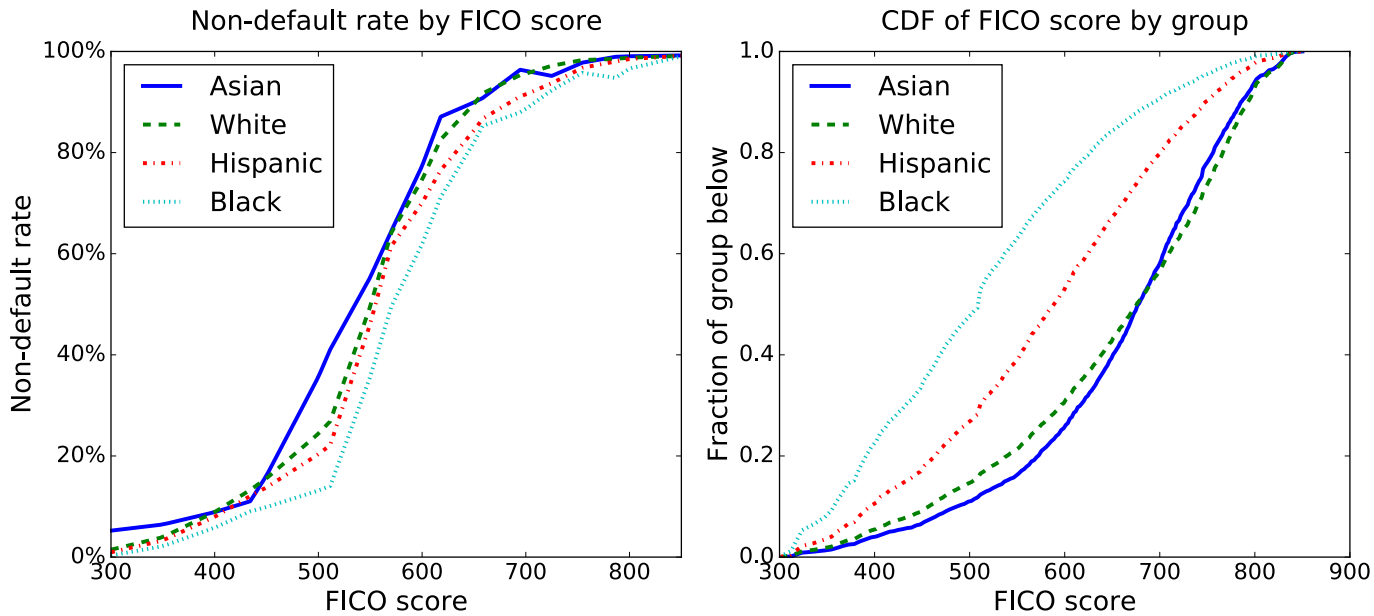


Figure: Hardt et al., 2016

Equal Opportunity

- The "race-blind" solution applies the same threshold for all the groups.
- Problem: non-defaulting black applicants are much less likely to be approved than non-defaulting white applicants.
  - Fraction of non-defaulting applicants in each group = fraction of area under curve which is shaded

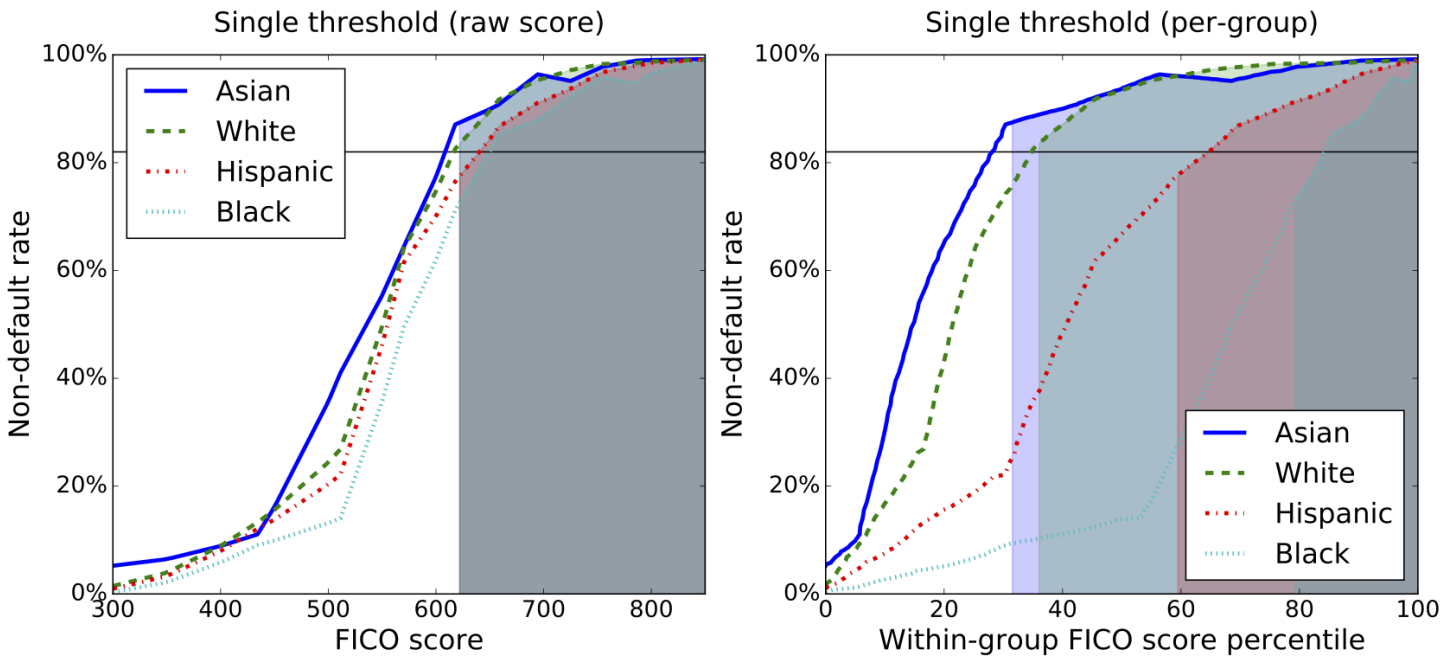


Figure: Hardt et al., 2016

Equal Opportunity

- Can obtain equal opportunity, equalized odds, demographic parity by setting group-specific thresholds (except equalized odds requires randomizing).

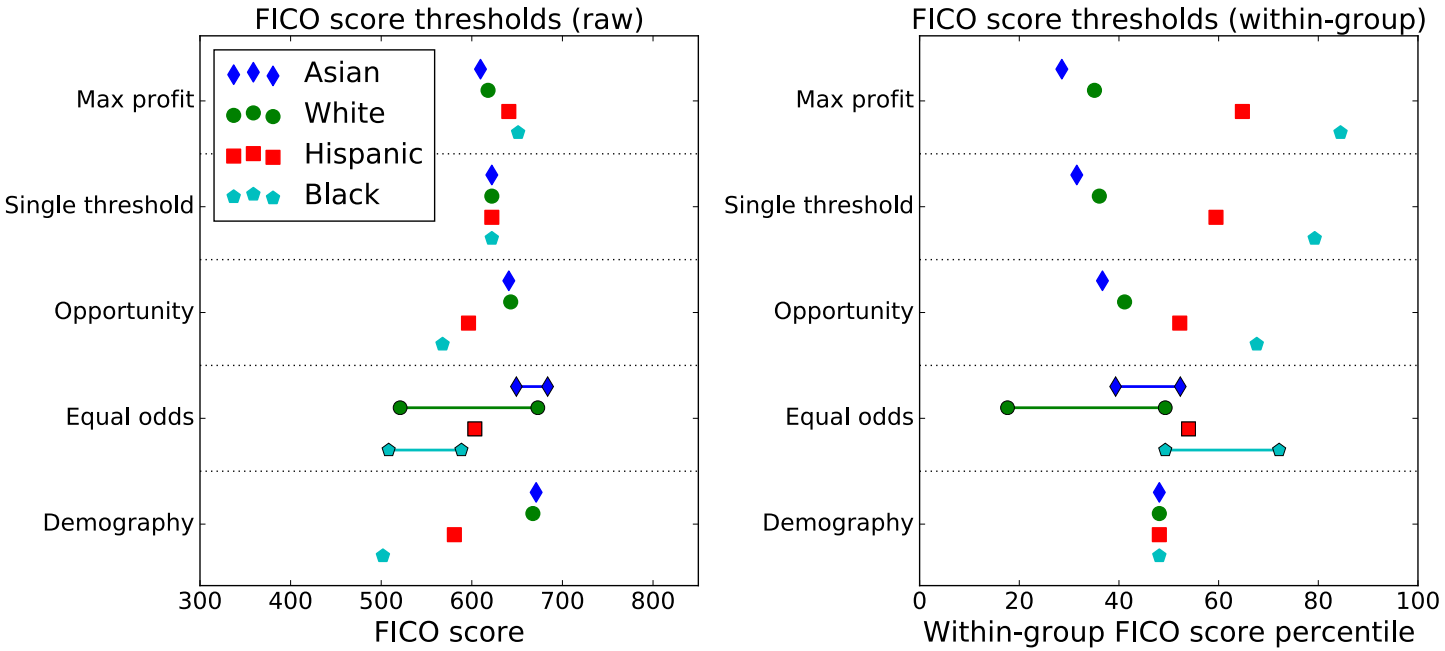


Figure: Hardt et al., 2016

Equal Opportunity

- Different notions of fairness often come into conflict. E.g., demographic parity conflicts with equal opportunity (left).
- Some notions of fairness are harder to achieve than others, in terms of lost profit (right).
- Choosing the right criterion requires careful consideration of the causal relationships between the variables.

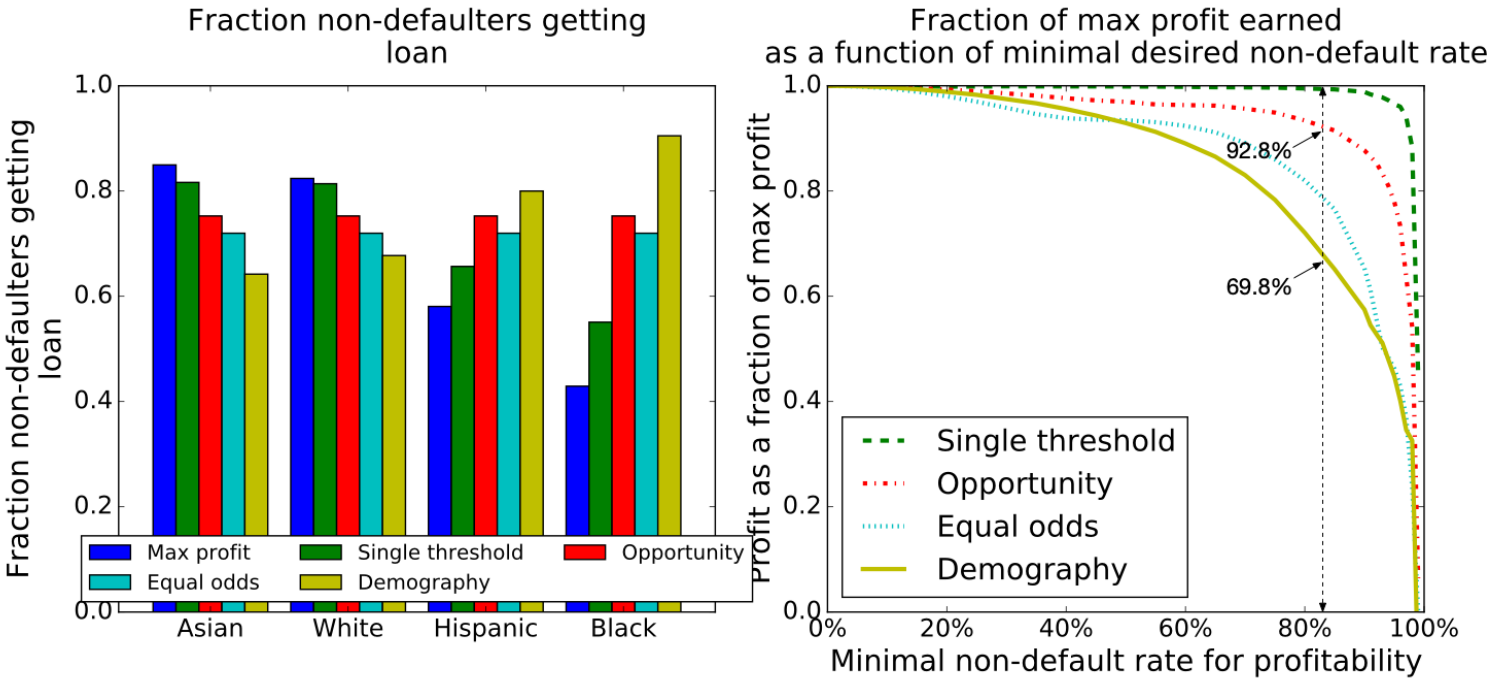


Figure: Hardt et al., 2016

Summary

- Fairness is a challenging issue to address
  - Not something you can just measure on a validation set
  - Philosophers and lawyers have been trying to define it for thousands of years
  - Different notions are incompatible. Need to carefully consider the particular problem.
    - individual vs. group
- Explosion of interest in ML over the last few years
- New conference on Fairness, Accountability, and Transparency (FAT*)
- New textbook: `https://fairmlbook.org/`

Closing Thoughts and Next Steps

Recap

What this course focused on:

- Supervised learning: regression, classification
  - Choose model, loss function, optimizer
  - Parametric vs. nonparametric
  - Generative vs. discriminative
  - Iterative optimization vs. closed-form solutions
- Unsupervised learning: dimensionality reduction and clustering
- Reinforcement learning: value iteration

This lecture: what we left out, and teasers for other courses

CSC2516 Teaser: Neural Nets

- This course covered some fundamental ideas, most of which are more than 10 years old.
- Big shift of the past decade: neural nets and deep learning
  - 2010: neural nets significantly improved speech recognition accuracy (after 20 years of stagnation)
  - 2012–2015: neural nets reduced error rates for object recognition by a factor of 6
  - 2016: a program called AlphaGo defeated the human Go champion
  - 2016: neural nets bridged half the gap between machine and human translation
  - 2015–2018: neural nets learned to produce convincing high-resolution images
  - 2017–2019: attention-based architectures (e.g. Transformers)

CSC2516 Teaser: Automatic Differentiation

- In this course, you derived update rules by hand
- Backprop is totally mechanical. Now we have automatic differentiation tools that compute gradients for you.
- In CSC2516, you learn how an autodiff package can be implemented
  - Lets you do fancy things like differentiate through the whole training procedure to compute the gradient of validation loss with respect to the hyperparameters.
- With TensorFlow, PyTorch, etc., we can build much more complex neural net architectures that we could previously.

CSC2516 Teaser: Beyond Scalar/Discrete Targets

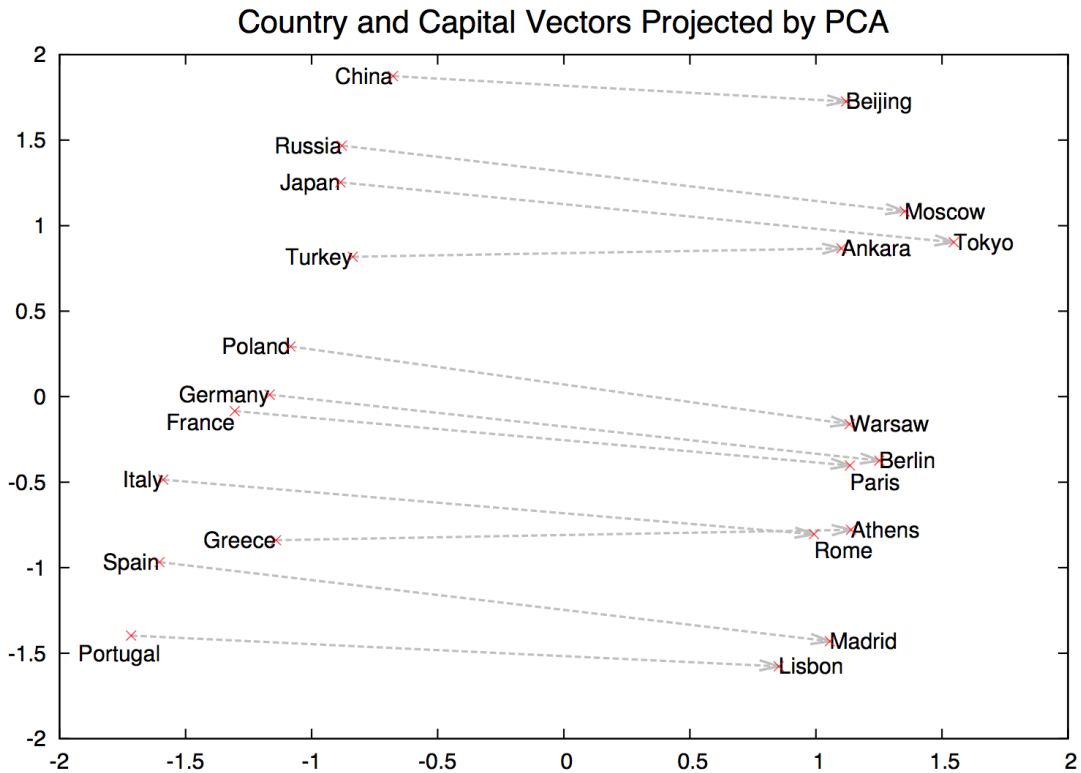
- This course focused on regression and classification, i.e. scalar-valued or discrete outputs
- That only covers a small fraction of use cases. Often, we want to output something more structured:
  - text (e.g. image question answering, machine translation)
  - dense labels of images (e.g. semantic segmentation)
  - graphs (e.g. molecule design)
- This used to be known as structured prediction, but now it's so routine we don't need a name for it.

CSC2516 Teaser: Representation Learning

- We talked about neural nets as learning feature maps you can use for regression/classification
- More generally, want to learn a representation of the data such that mathematical operations on the representation are semantically meaningful
- Classic (decades-old) example: representing words as vectors
  - Measure semantic similarity using the dot product between word vectors (or dissimilarity using Euclidean distance)
  - Represent a web page with the average of its word vectors

CSC2516 Teaser: Representation Learning

- Here's a linear projection of word representations for cities and capitals into 2 dimensions (part of a representation learned using word2vec)
- The mapping $\text{city} \rightarrow \text{capital}$ corresponds roughly to a single direction in the vector space:



Mikolov et al., 2013, "Efficient estimation of word representations in vector space"

CSC2516 Teaser: Representation Learning

- In other words, $\text{vec}(\text{Paris}) - \text{vec}(\text{France}) \approx \text{vec}(\text{London}) - \text{vec}(\text{England})$
- This means we can analogies by doing arithmetic on word vectors:
  - e.g. "Paris is to France as London is to _____"
  - Find the word whose vector is closest to $\text{vec}(\text{France}) - \text{vec}(\text{Paris}) + \text{vec}(\text{London})$
- Example analogies:

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Mikolov et al., 2013, "Efficient estimation of word representations in vector space"

CSC2516 Teaser: Representation Learning

One of the big goals is to learn *disentangled* representations, where individual dimensions tell you something meaningful

(a) Baldness (-6, 6)

(b) Face width (0, 6)

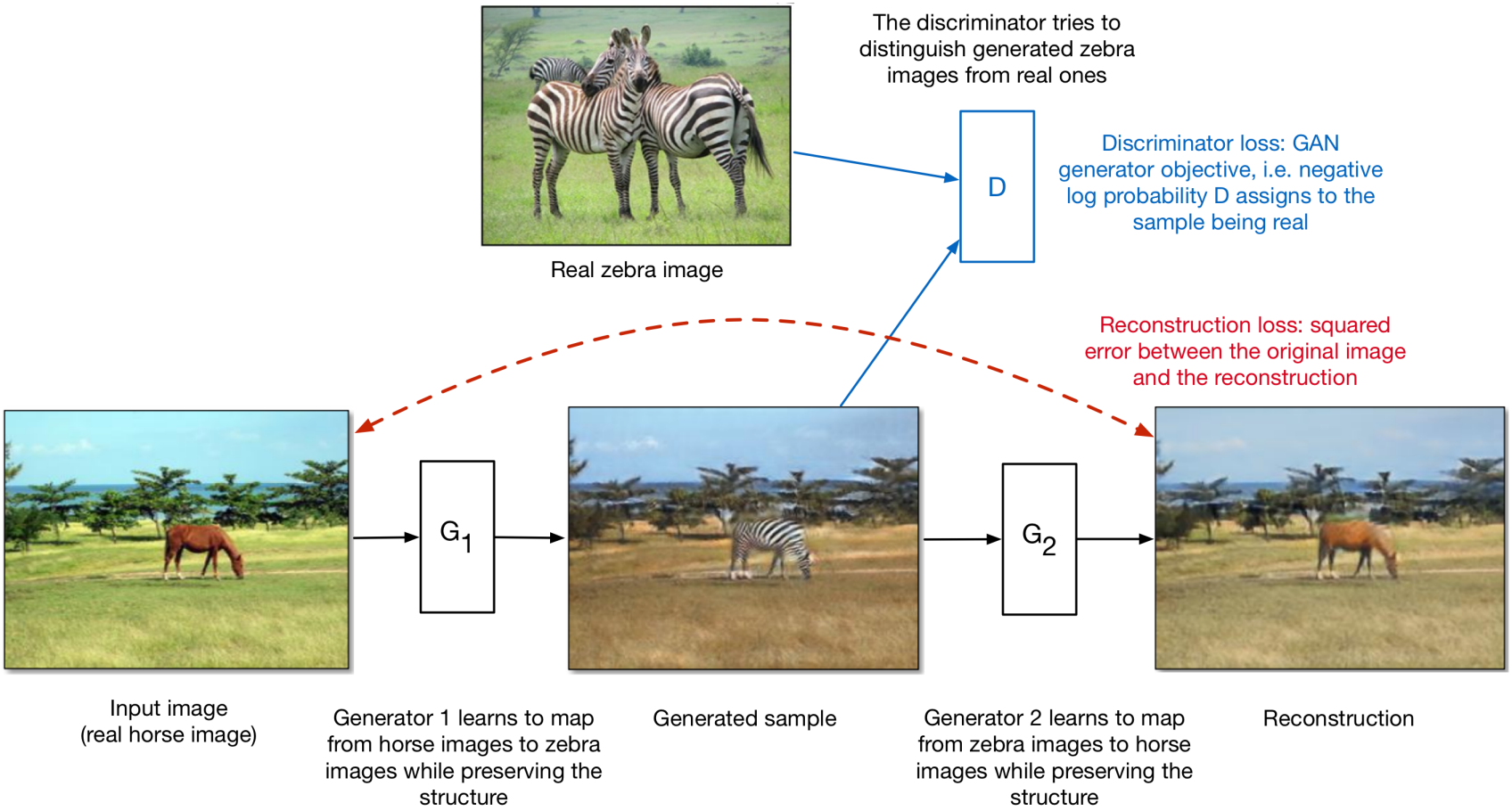
(c) Gender (-6, 6)

(d) Mustache (-6, 0)

Chen et al., 2018, "Isolating sources of disentanglement in variational autoencoders"

CSC2516 Teaser: Image-to-Image Translation

Due to convenient autodiff frameworks, we can combine multiple neural nets together into fancy architectures. Here's the CycleGAN.

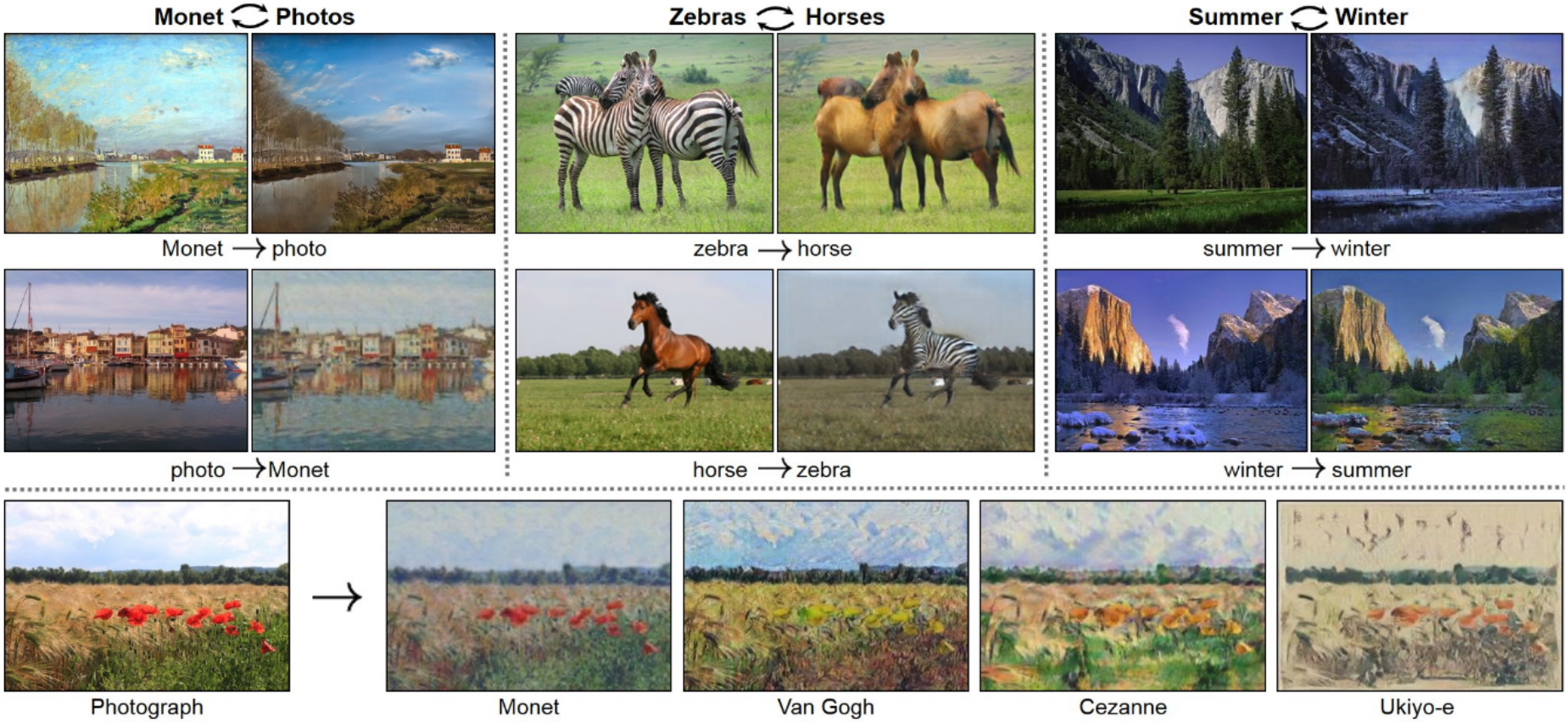


Total loss = discriminator loss + reconstruction loss

Zhu et al., 2017, "Unpaired image-to-image translation using cycle-consistent adversarial networks"

CSC2516 Teaser: Image-to-Image Translation

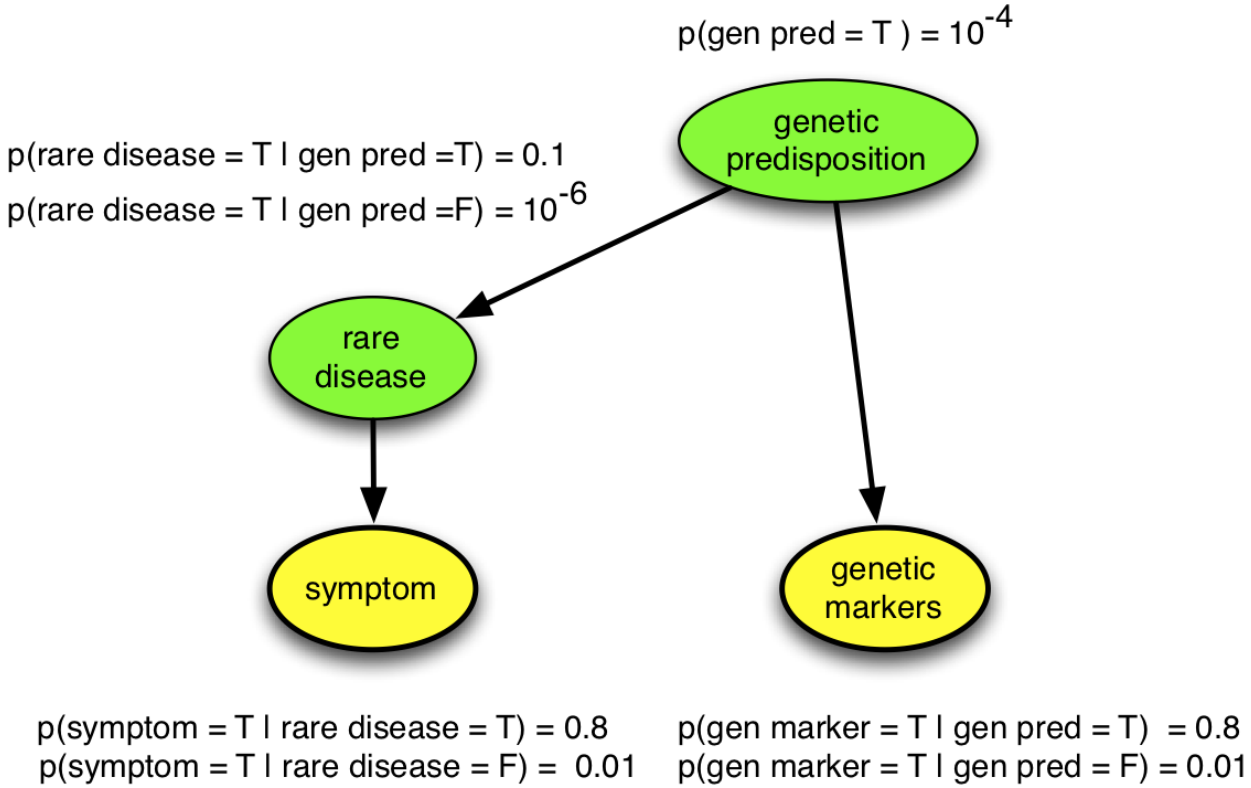
Style transfer problem: change the style of an image while preserving the content.



Data: Two unrelated collections of images, one for each style

CSC2506 Teaser: Probabilistic Graphical Models

- In this course, we just scratched the surface of probabilistic models.
- Probabilistic graphical models (PGMs) let you encode complex probabilistic relationships between lots of variables.



Ghahramani, 2015, "Probabilistic ML and artificial intelligence"

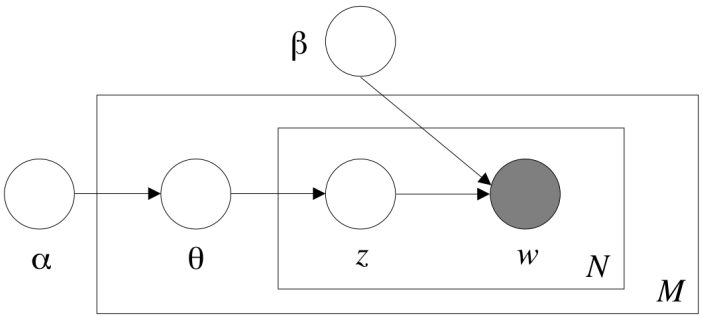
CSC2506 Teaser: PGM Inference

- We derived inference methods by inspection for some easy special cases (e.g. GDA, naïve Bayes)
- In CSC2506, you'll learn much more general and powerful inference techniques that expand the range of models you can build
  - Exact inference using dynamic programming, for certain types of graph structures (e.g. chains)
  - Markov chain Monte Carlo
    - forms the basis of a powerful probabilistic modeling tool called Stan
  - Variational inference: try to approximate a complex, intractable, high-dimensional distribution using a tractable one
    - Try to minimze the KL divergence
    - Based on the same math from our EM lecture

CSC2506 Teaser: Beyond Clustering

- We've seen unsupervised learning algorithms based on two ways of organizing your data
  - low-dimensional spaces (dimensionality reduction)
  - discrete categories (clustering)
- Other ways to organize/model data
  - hierarchies
  - dynamical systems
  - sets of attributes
  - topic models (each document is a mixture of topics)
- Motifs can be combined in all sorts of different ways

Latent Dirichlet Allocation (LDA)



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Blei et al., 2003, "Latent Dirichlet Allocation"

CSC2506 Teaser: Beyond Clustering

Automatic mouse tracking

- When biologists do behavioral genetics researchers on mice, it's very time consuming for a person to sit and label everything a mouse does
- The Datta lab at Harvard built a system for automatically tracking mouse behaviors
- Goal: show the researchers a summary of how much time different mice spend on various behaviors, so they can determine the effects of the genetic manipulations
- One of the major challenges is that we don't know the right "vocabulary" for describing the behaviors — clustering the observations into meaningful groups is an unsupervised learning task

Switching linear dynamical system model

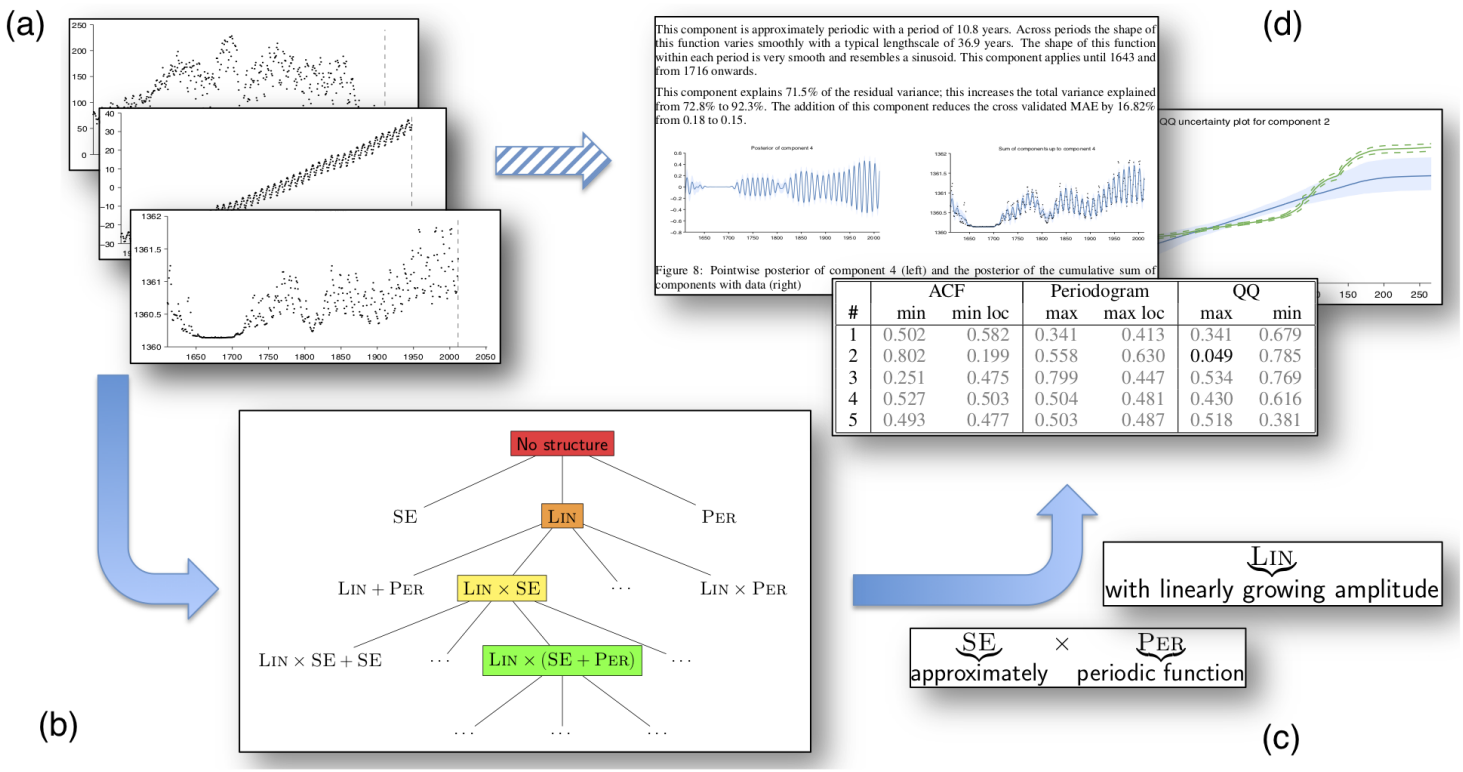
- Mouse's movements are modeled as a dynamical system
- System parameters depend on what behavior the mouse is currently engaging in
- Mice transition stochastically between behaviors according to some distribution

Videos

- `https://www.cell.com/neuron/fulltext/S0896-6273(15)01037-5`
- `https://www.youtube.com/watch?v=btr1poCYIzw`

CSC2506 Teaser: Automatic Statistician

Automatic search over Gaussian process kernel structures



#	ACF min	ACF min loc	Periodogram max	Periodogram max loc	QQ max	QQ min
1	0.502	0.582	0.341	0.413	0.341	0.679
2	0.802	0.199	0.558	0.630	0.049	0.785
3	0.251	0.475	0.799	0.447	0.534	0.769
4	0.527	0.503	0.504	0.481	0.430	0.616
5	0.493	0.477	0.503	0.487	0.518	0.381

Duvenaud et al., 2013, "Structure discovery in nonparametric regression through compositional kernel search"
Image: Ghahramani, 2015, "Probabilistic ML and artificial intelligence"

Resources

Continuing with machine learning

- Courses
  - csc2516, "Neural Networks and Deep Learning"
  - csc2506, "Probabilistic Learning and Reasoning"
  - csc2547, "Topics in Statistical Learning Theory"
  - Various topics courses (varies from year to year)
- Videos from top ML conferences (NIPS/NeurIPS, ICML, ICLR, UAI)
  - Tutorials and keynote talks are aimed at people with your level of background (know the basics, but not experts in a subfield)
- Try to reproduce results from papers
  - If they've released code, you can use that as a guide if you get stuck
- Lots of excellent free resources avaiable online!