On Computing Counterfactuals for Causal Fairness

Master's Thesis

Ayan Majumdar

Supervisors

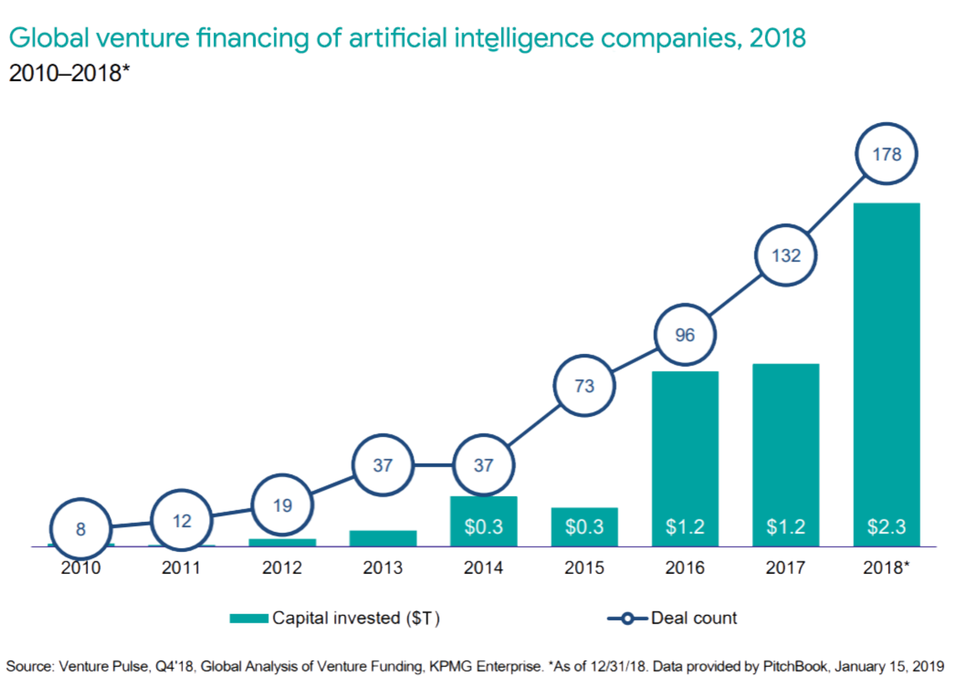
Krishna Gummadi, Isabel Valera

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

UNIVERSITÄT
DES
SAARLANDES

ML and our society

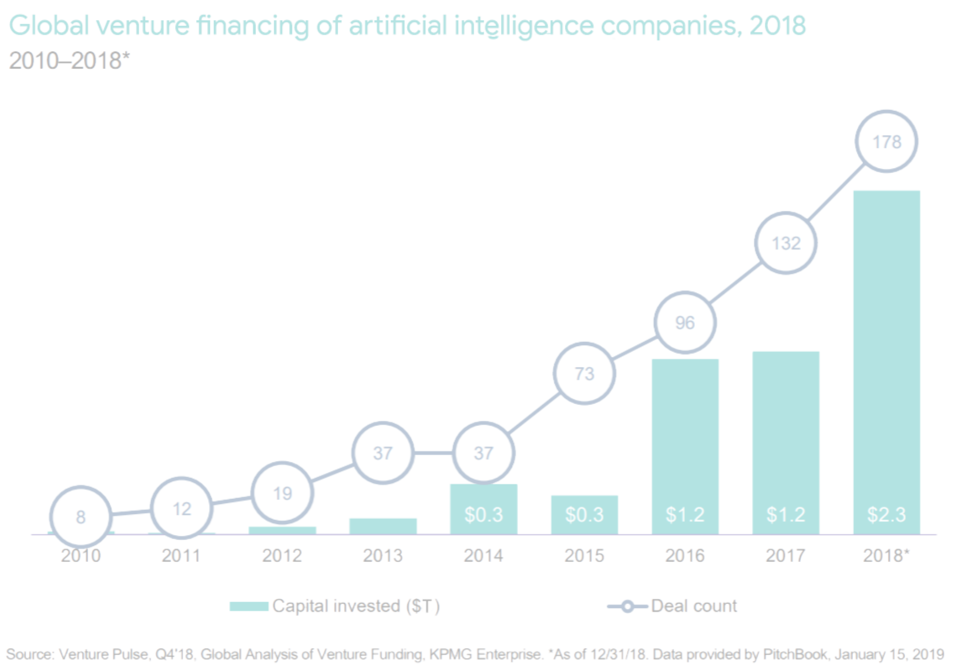
Data-driven ML algorithms **heavily** deployed in today's tech industry



Increased industry financing for AI and ML

ML and our society

Data-driven ML algorithms
make **critical** predictions!



Increased industry financing for AI and ML

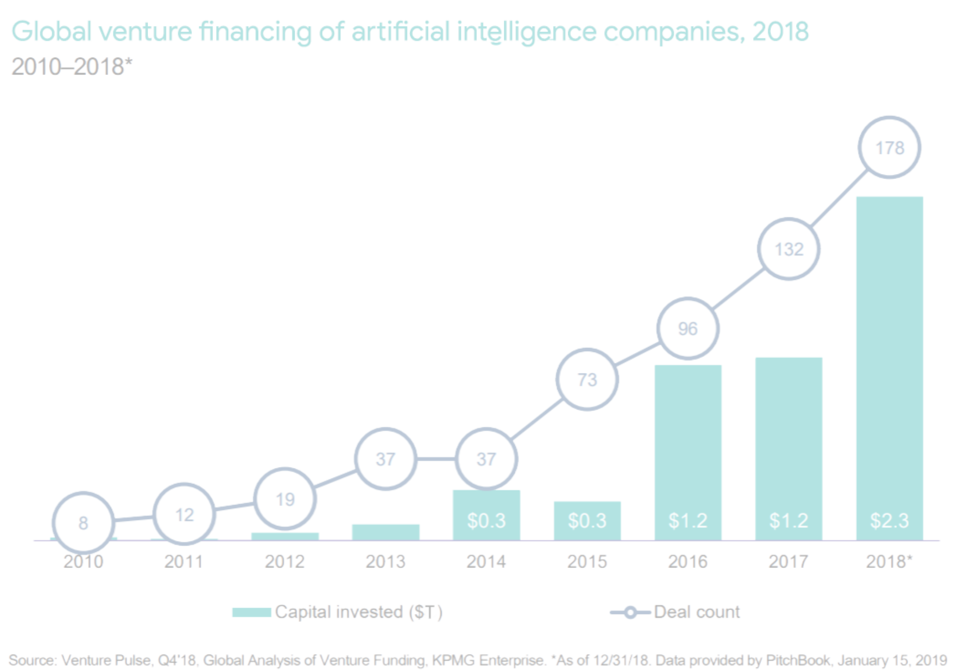
ML and our society

Data-driven ML algorithms make **critical** predictions!

NORTHPOINTE

PredPol
Predict Crime in Real Time
PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

ShotSpotter®



Increased industry financing for AI and ML

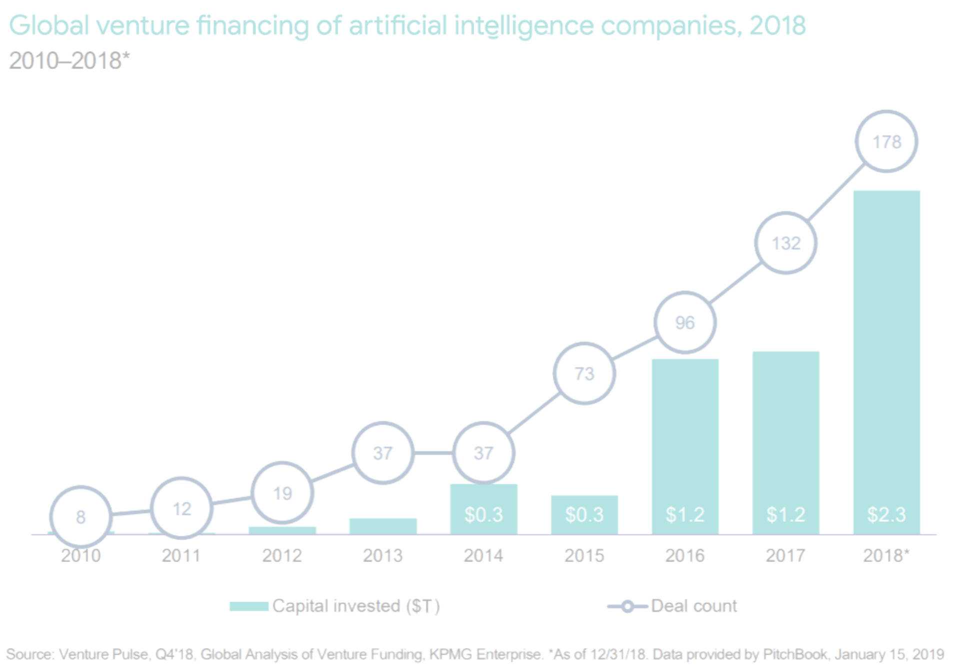
ML and our society

Data-driven ML algorithms make **critical** predictions!

NORTHPOINTE

PredPol
Predict Crime in Real Time
PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

ShotSpotter®



Card

EQUIFAX

ZEST AI

Increased industry financing for AI and ML

ML and our society

Data-driven ML algorithms make **critical** predictions!

NORTHPOINTE

PredPol Predict Crime in Real Time

PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

ShotSpotter®

Card

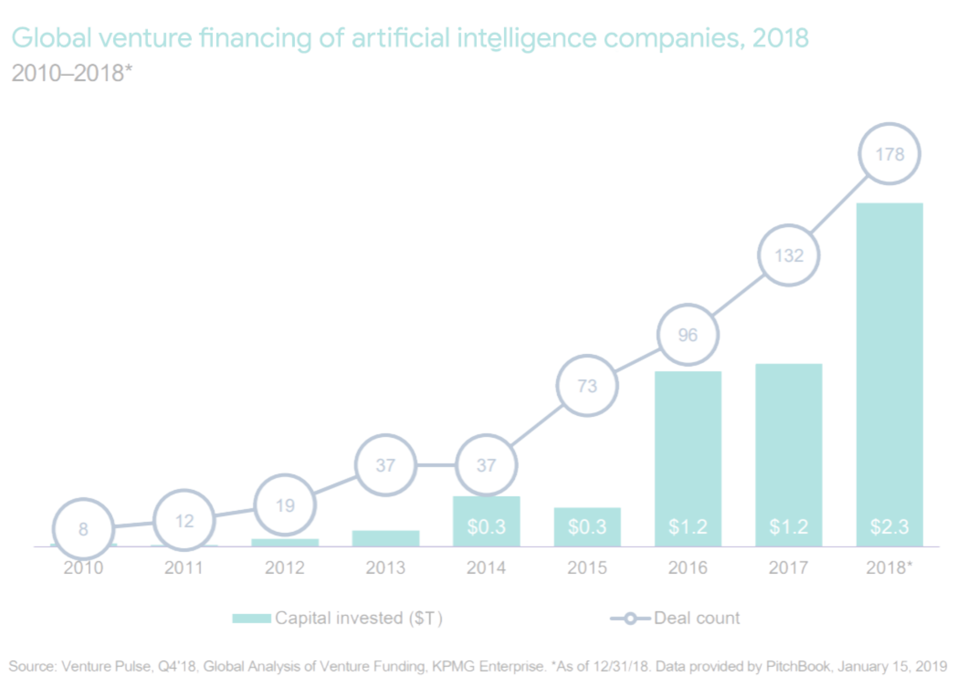
EQUIFAX

ZEST AI

HireVue

amazon

entelo



Increased industry financing for AI and ML

Fairness in ML systems

Studies have shown potential **bias**!

Business

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the higher credit score

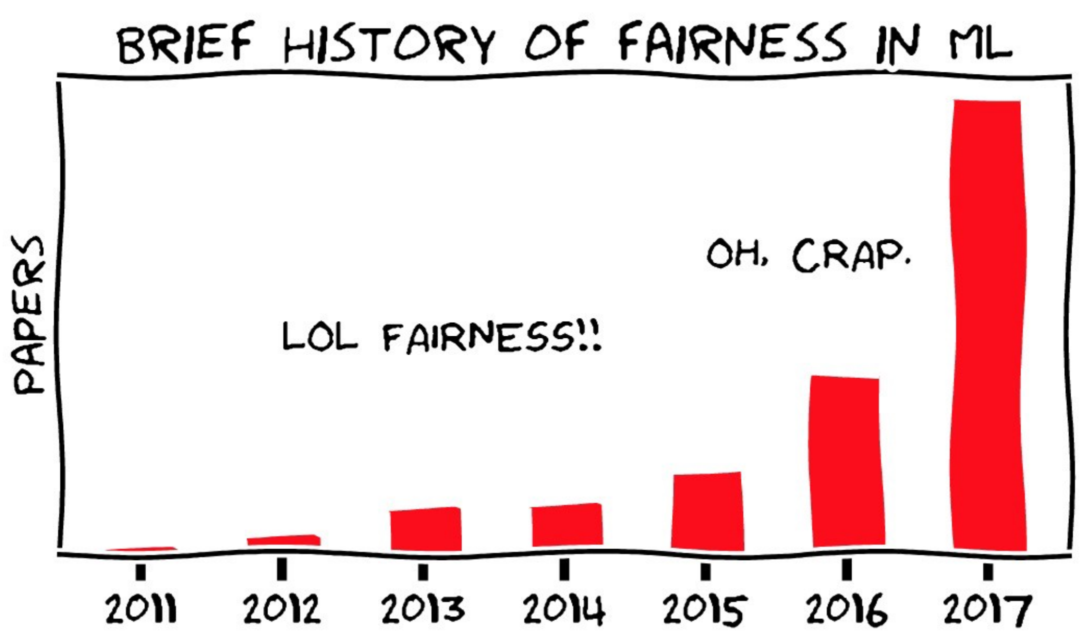
Why Amazon's Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program
OCTOBER 12, 2018 | 1:00 PM

TAGS: Women's Rights in the Workplace, Women's Rights, Privacy & Technology

Fairness in ML systems

Led to **extensive** research in the domain…



Credits: Moritz Hardt, CS 294-Fairness in Machine Learning

Business

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the higher credit score

Why Amazon's Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program

OCTOBER 12, 2018 | 1:00 PM

TAGS: Women's Rights in the Workplace, Women's Rights, Privacy & Technology

Existing notions of fairness

Many definitions

- Group: *Fairness through unawareness, demographic parity, …*
- Individual: *individual fairness*

Existing notions of fairness

Many definitions

- Group: *Fairness through unawareness, demographic parity, …*
- Individual: *individual fairness*

However…

Existing notions of fairness

Many definitions

- Group: *Fairness through unawareness, demographic parity, …*
- Individual: *individual fairness*

However…

- What is the **cause** of bias?
- How to **eliminate** bias?
- Which individuals get **similar treatment**?

Existing notions of fairness

Many definitions

- Group: *Fairness through unawareness, demographic parity, …*
- Individual: *individual fairness*

However…

- What is the **cause** of bias?
- How to **eliminate** bias?
- Which individuals get **similar treatment**?

Not clear!

Use causation in fairness!

Use causation in fairness!

Is the law school admission process fair?

Jacob is a **black male** law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was **rejected**.

Use causation in fairness!

Is the law school admission process fair?

Jacob is a **black male** law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was **rejected**.

- Had Jacob been **white** instead, would he had been **accepted**?
  - *counterfactual*

Use causation in fairness!

Is the law school admission process fair?

Jacob is a **black male** law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was **rejected**.

- Had Jacob been **white** instead, would he had been **accepted**?
  — *counterfactual*
- Did Jacob's race **cause** him to get negative outcome?
  — *counterfactual fairness* (*Kusner et al. 2017*)[1]

[1]*Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Use causation in fairness!

Is the law school admission process fair?

Jacob is a **black male** law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was **rejected**.

- Had Jacob been **white** instead, would he had been **accepted**?
  — *counterfactual*
- Did Jacob's race **cause** him to get negative outcome?
  — *counterfactual fairness* (*Kusner et al. 2017*)[1]

Such questions of fairness need **counterfactual data**

[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Use causation in fairness!

Is the law school admission process fair?

Jacob is a **black male** law school applicant. He scored 55

Need to **know** data generating process…

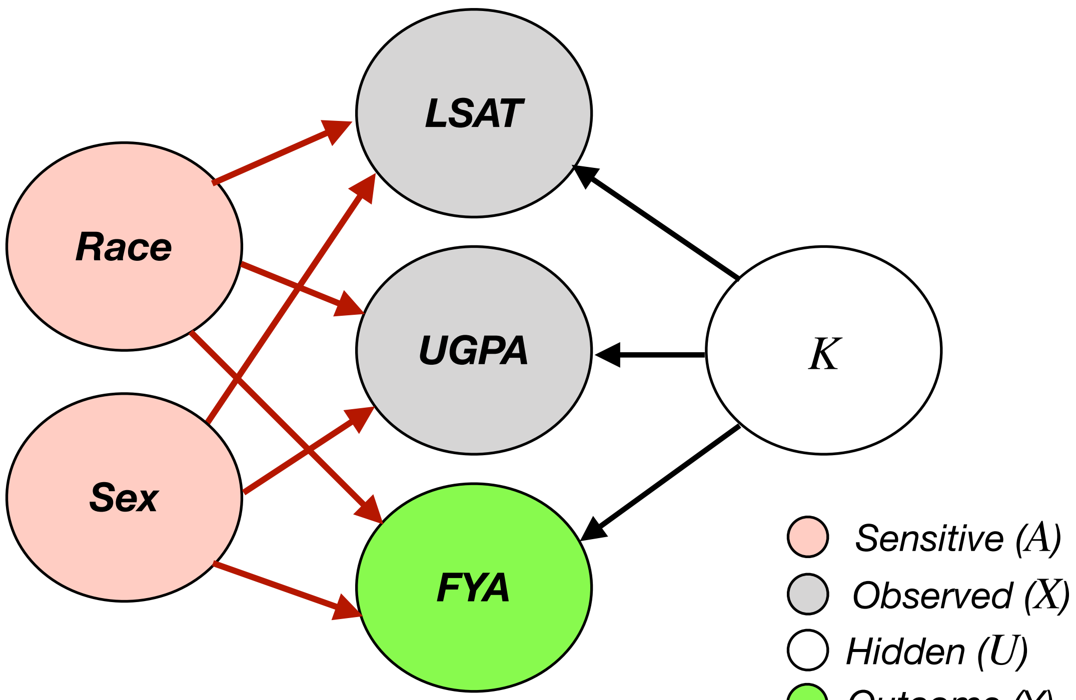
Causal models!

Such questions of fairness need **counterfactual data**

[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Causal models

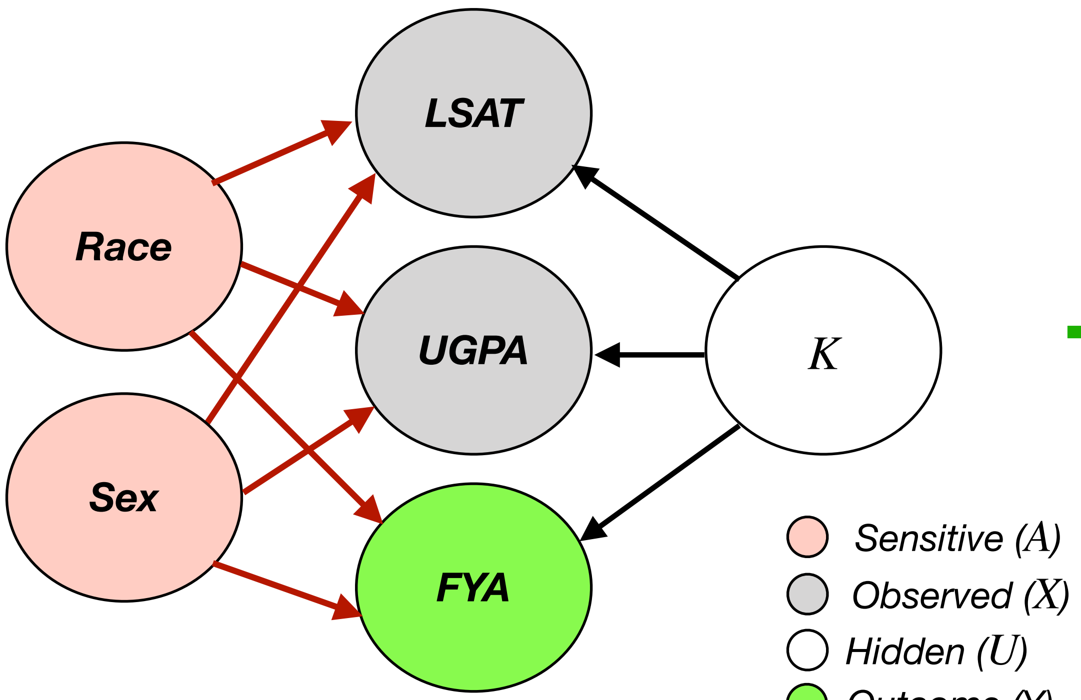
Causal graph



Relations between the features

Causal models

Causal graph



Relations between the features

+

Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

$$\textbf{UGPA} := \mathcal{N}(b_G + w_G^R R + w_G^S S + w_G^K K, \sigma_G)$$

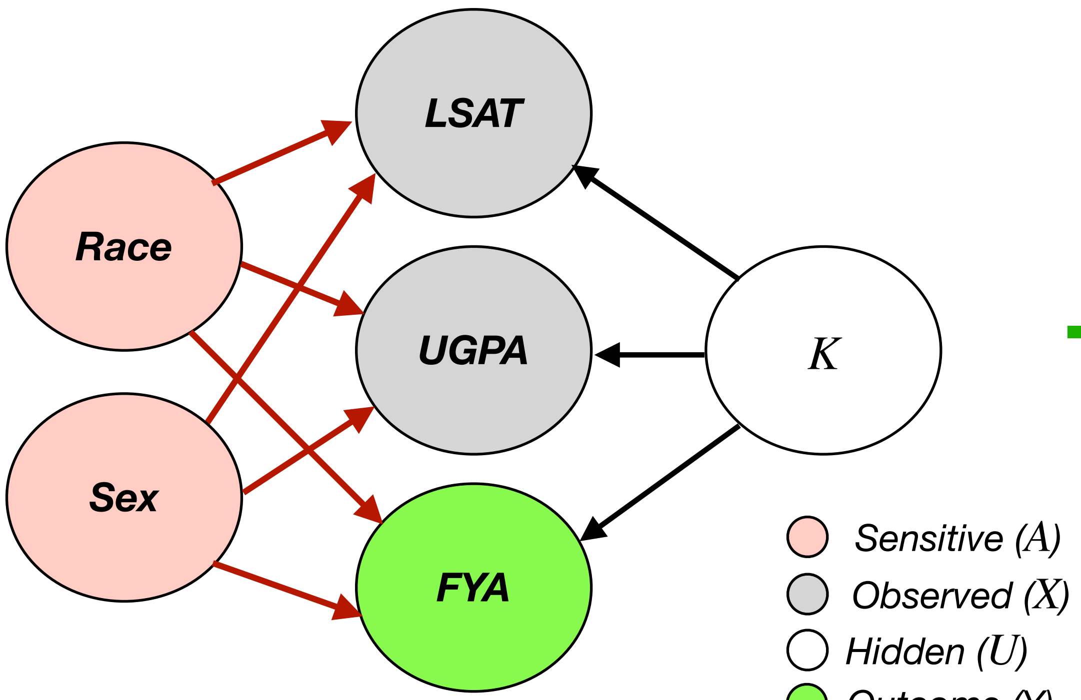
$$\textbf{FYA} := \mathcal{N}(w_F^R R + w_F^S S + w_F^K K, 1)$$

$$K \sim \mathcal{N}(0,1)$$

Quantification of the relations

Causal models

Causal graph



Relations between the features

+

Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

$$\textbf{UGPA} := \mathcal{N}(b_G + w_G^R R + w_G^S S + w_G^K K, \sigma_G)$$

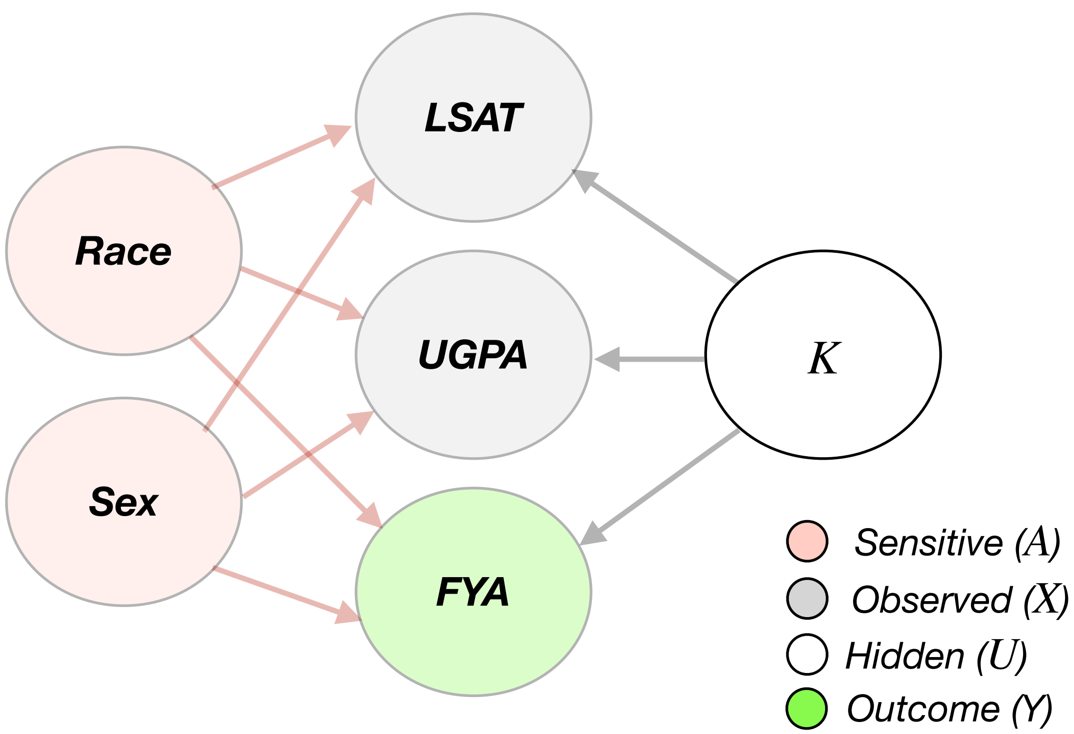
$$\textbf{FYA} := \mathcal{N}(w_F^R R + w_F^S S + w_F^K K, 1)$$

$$K \sim \mathcal{N}(0,1)$$

Quantification of the relations

Strict assumptions allow **counterfactual** generation

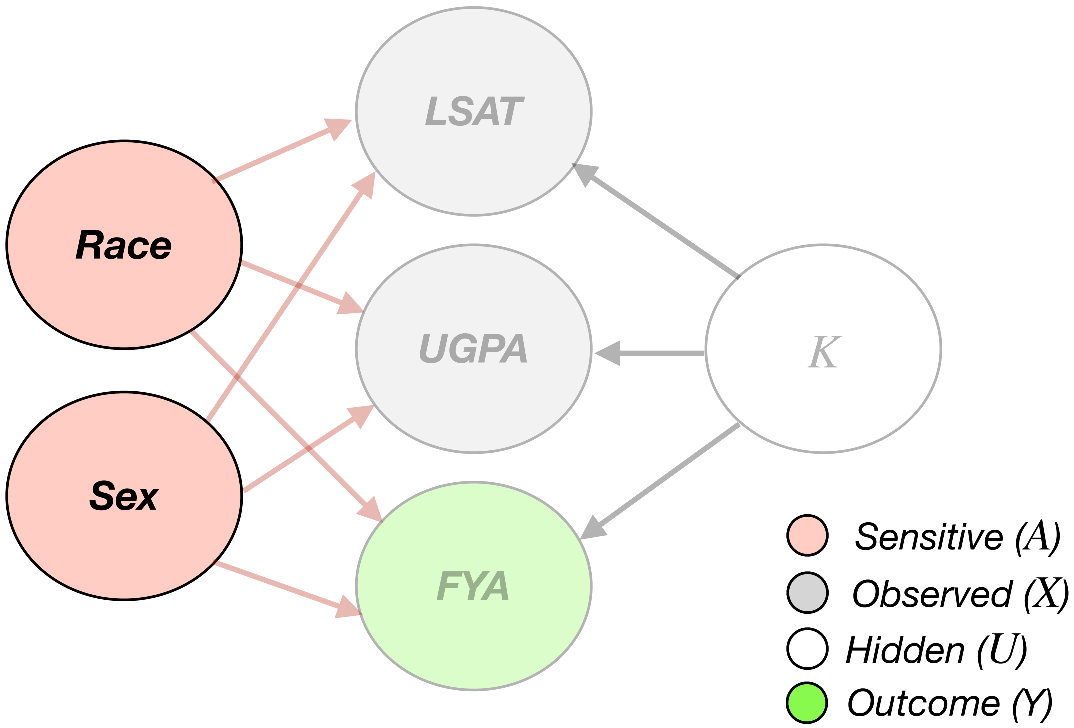
Counterfactuals from causal model[1]



1. *Abduction*: Given $X, A = a$ **estimate** U

[1] Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.

Counterfactuals from causal model[1]

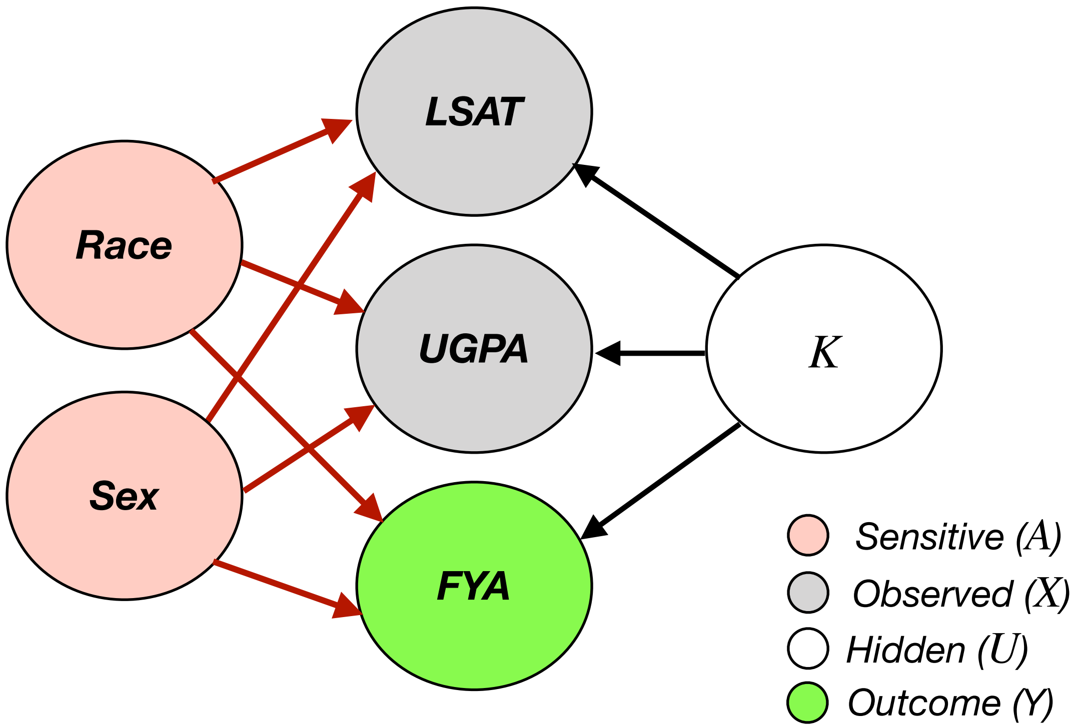


1. *Abduction*: Given $X, A = a$ **estimate** ϵ

2. *Action*: **Intervene** on A by setting it to a'

[1] Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.

Counterfactuals from causal model[1]



1. *Abduction*: Given $X, A = a$ **estimate** ϵ

2. *Action*: **Intervene** on A by setting it to a'

3. *Prediction*: **Counterfactual** X^c using U under intervention $do(A = a')$

[1] *Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

Achieving fairness with counterfactuals

- Prediction $\hat{Y}$ (for any individual) should not change while[1]:

[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Achieving fairness with counterfactuals

- Prediction $\hat{Y}$ (for any individual) should not change while[1]:
  - **Intervene** on sensitive feature A
  - Keep everything not dependent on A **constant**

[1]Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.

Achieving fairness with counterfactuals

- Prediction $\hat{Y}$ (for any individual) should not change while[1]:
  - **Intervene** on sensitive feature A
  - Keep everything not dependent on A **constant**

$$P\left(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a\right) = P\left(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a\right)$$

[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Achieving fairness with counterfactuals

- Prediction $\hat{Y}$ (for any individual) should not change while[1]:
  - **Intervene** on sensitive feature A
  - Keep everything not dependent on A **constant**

$$P\left(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a\right) = P\left(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a\right)$$

- Feature A should not **cause** $\hat{Y}$ in any **individual** instance!

[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Achieving fairness with counterfactuals

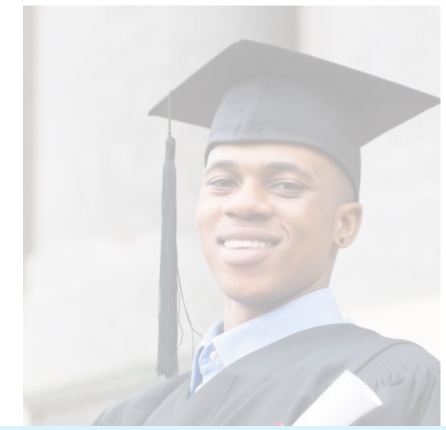
- Prediction $\hat{Y}$ (for any individual) should not change while[1]:

Complete causal knowledge is **infeasible** in practice!

- Feature A should not **cause** $\hat{Y}$ in any **individual** instance!

[1] Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.

Achieving fairness with counterfactuals



- Prediction $\hat{Y}$ (for any individual) should not change while[1]:

Complete causal knowledge is **infeasible** in practice!

Wrong assumptions → high **errors**!

- Feature A should not **cause** $\hat{Y}$ in any **individual** instance!

[1]*Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Research Question

Can we generate **counterfactuals** for **counterfactual fairness** without complete **causal** knowledge?

Research Question

Can we generate **counterfactuals** for **counterfactual fairness** without complete **causal** knowledge?

1. Use generated **counterfactuals** to audit trained predictive models?

Research Question

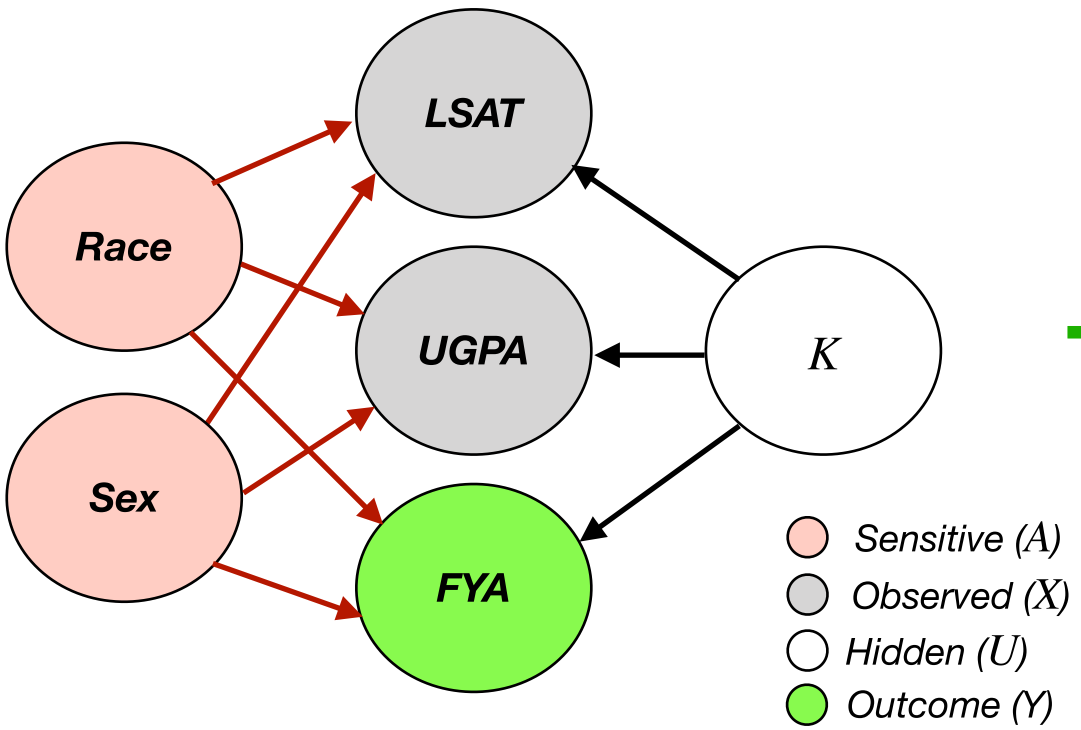
Can we generate **counterfactuals** for **counterfactual fairness** without complete **causal** knowledge?

1. Use generated **counterfactuals** to audit trained predictive models?

2. Build a predictive model that is **counterfactually fair**?

Recap: Causal counterfactuals

Causal graph



Relations between the features

+

Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

$$\textbf{UGPA} := \mathcal{N}(b_G + w_G^R R + w_G^S S + w_G^K K, \sigma_G)$$

$$\textbf{FYA} := \mathcal{N}(w_F^R R + w_F^S S + w_F^K K, 1)$$

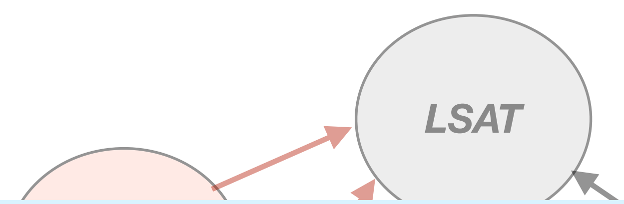
$$K \sim \mathcal{N}(0,1)$$

Quantification of the relations

Generate counterfactuals by Pearl's 3 steps

Recap: Causal counterfactuals

Causal graph



Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

How to **generate** **counterfactuals** in the **absence** of complete causal knowledge?

Relations between the features

Quantification of the relations

Generate counterfactuals by Pearl's 3 steps

Fairness scenarios have implicit structures

Fairness scenarios have implicit structures

1. Sensitive features **intrinsic** factors for individuals

   $\longrightarrow$ A *root nodes in causal graph*

Fairness scenarios have implicit structures

1. Sensitive features **intrinsic** factors for individuals

   $\longrightarrow$ A *root nodes in causal graph*

2. Sensitive features affect some **observed** features

   $\longrightarrow$ *Causal links from* A *to some* X

Fairness scenarios have implicit structures

1. Sensitive features **intrinsic** factors for individuals

   $\longrightarrow$ A *root nodes in causal graph*

2. Sensitive features affect some **observed** features

   $\longrightarrow$ *Causal links from* A *to some* X

3. Hidden factors **independent** of sensitive features

   $\longrightarrow$ ϵ *independent root nodes in causal graph*

Fairness scenarios have implicit structures

1. Sensitive features **intrinsic** factors for individuals

   $\longrightarrow$ A *root nodes in causal graph*

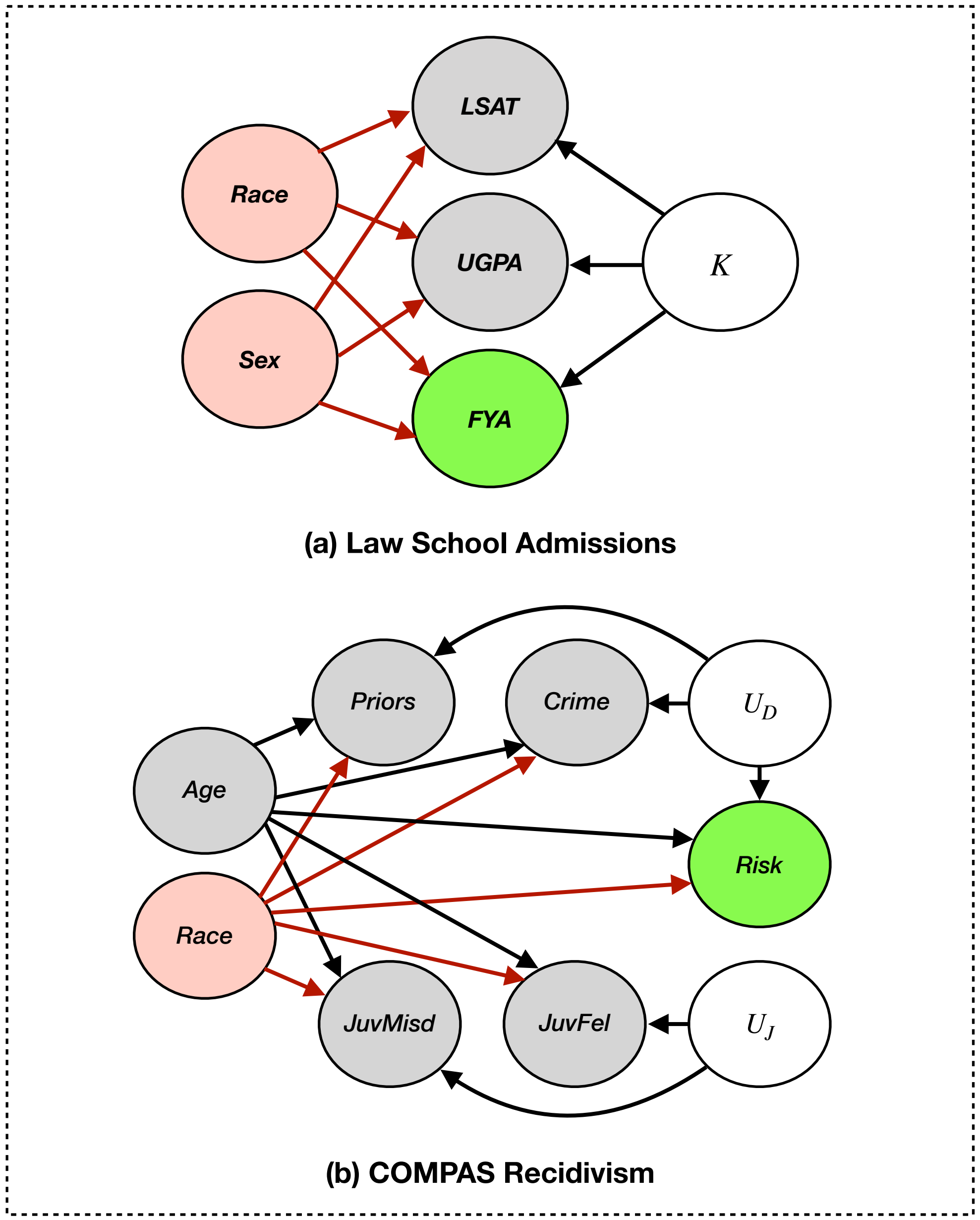
Can work with simpler assumptions!

3. Hidden factors **independent** of sensitive features

   $\longrightarrow$ ϵ *independent root nodes in causal graph*

Assumptions

✓ **Fairness** scenarios allow using **simpler** causal assumptions!



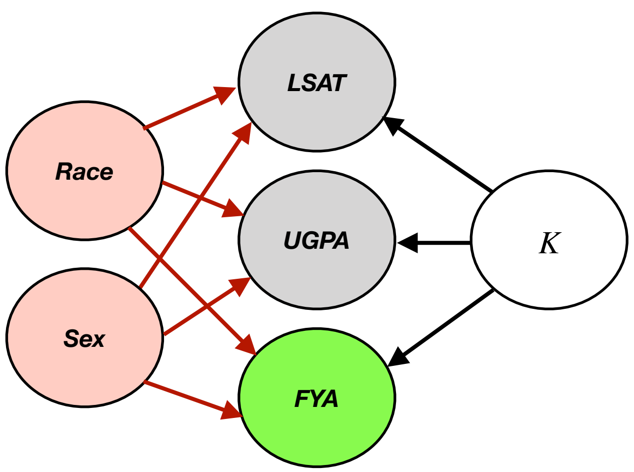
Example fairness causal graphs[1,2]

[1] *Matt J Kusner et al. "Counterfactual Fairness". In NIPS 2017*

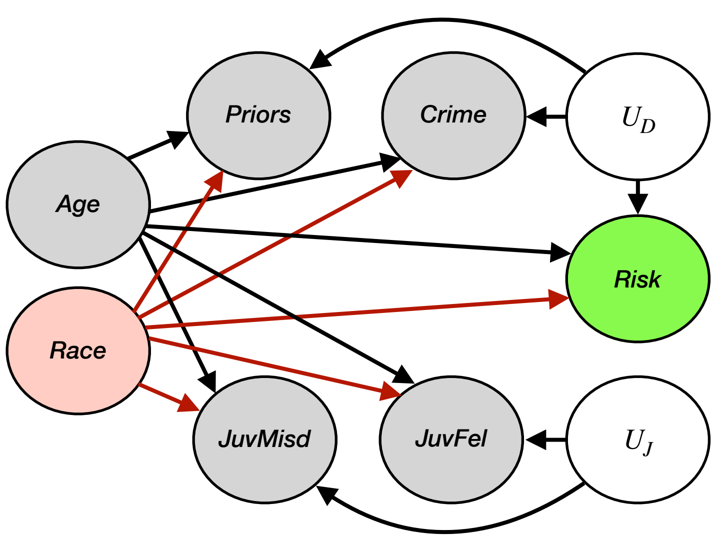
[2] *Chris Russell et al. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness". In NIPS 2017*

Assumptions

✓ **Fairness** scenarios allow using **simpler** causal assumptions!



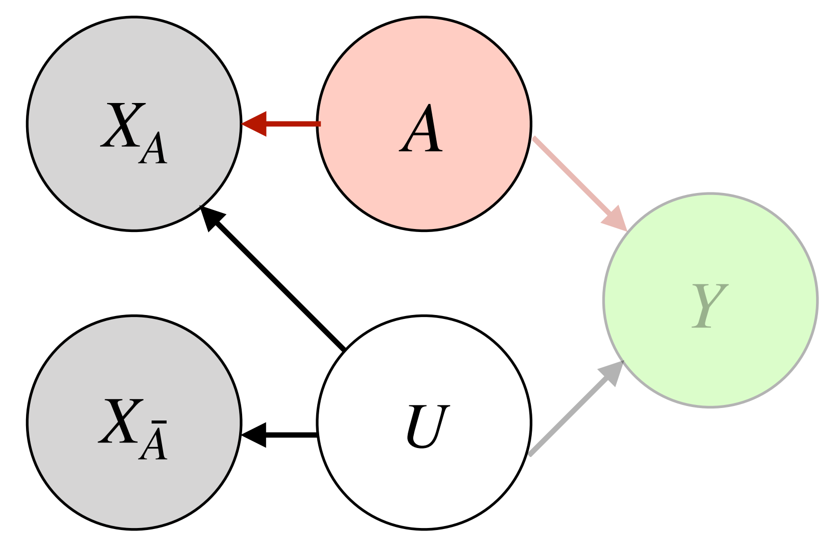
(a) Law School Admissions



(b) COMPAS Recidivism

Example fairness causal graphs[1,2]

Simplify



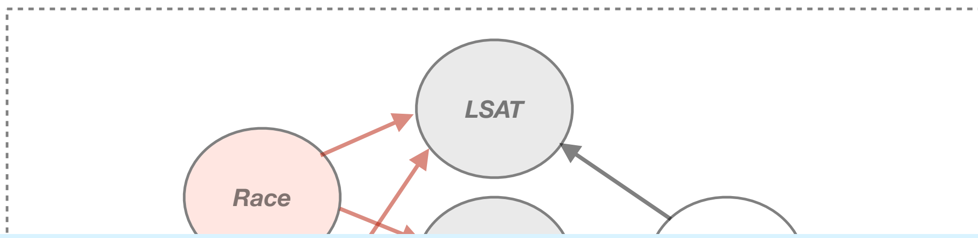
Assumed causal graph

[1] *Matt J Kusner et al. "Counterfactual Fairness". In NIPS 2017*

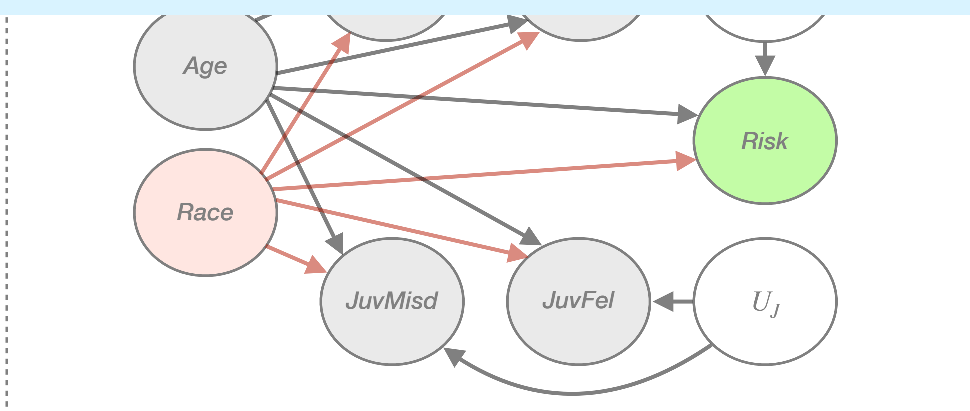
[2] *Chris Russell et al. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness". In NIPS 2017*

Assumptions

✓ **Fairness** scenarios allow using **simpler causal** assumptions!

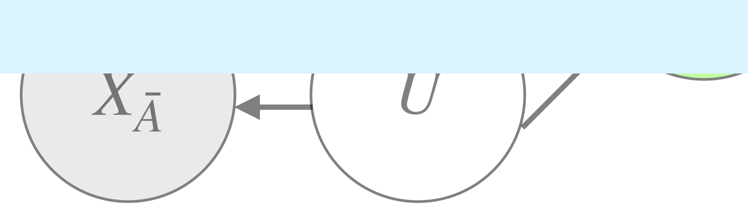


How to **model** data **generating** process?



(b) COMPAS Recidivism

Example fairness causal graphs[1,2]



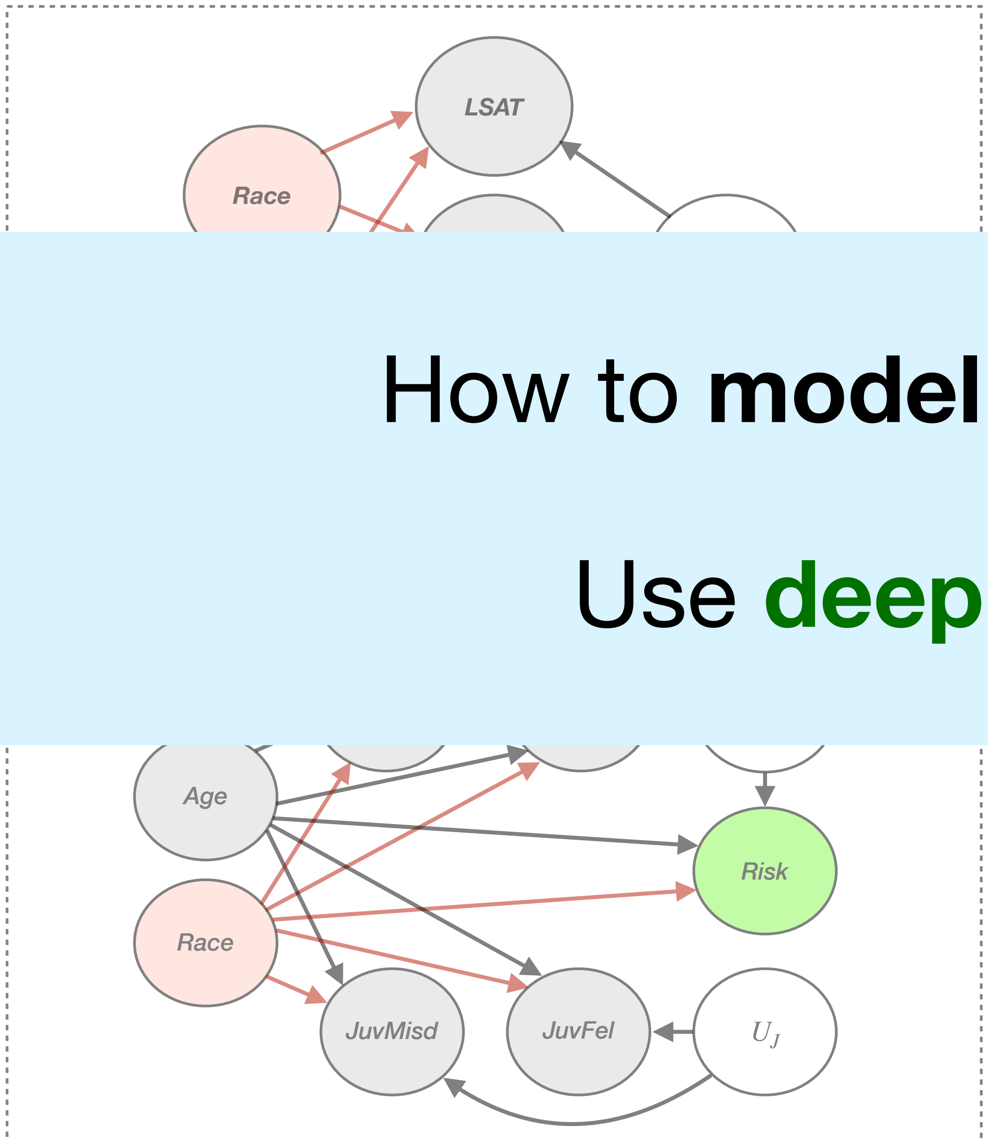
Assumed causal graph

[1] *Matt J Kusner et al. "Counterfactual Fairness". In NIPS 2017*

[2] *Chris Russell et al. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness". In NIPS 2017*

Assumptions

✓ **Fairness** scenarios allow using **simpler** causal assumptions!

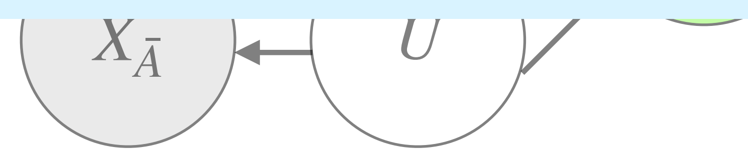


(b) COMPAS Recidivism

Example fairness causal graphs[1,2]

How to **model** data **generating** process?

Use **deep generative** modeling!



Assumed causal graph

[1]Matt J Kusner et al. "Counterfactual Fairness". In NIPS 2017

[2]Chris Russell et al. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness". In NIPS 2017

Modeling

1. **Intervene** on A for counterfactual fairness

   $\rightarrow$ *learn generative process* ***conditioned*** *on* A

Modeling

1. **Intervene** on A for counterfactual fairness

   → *learn generative process **conditioned** on* A

2. Estimate total effect of causal hidden factors in data

   → *use **latent-space** generative modeling*

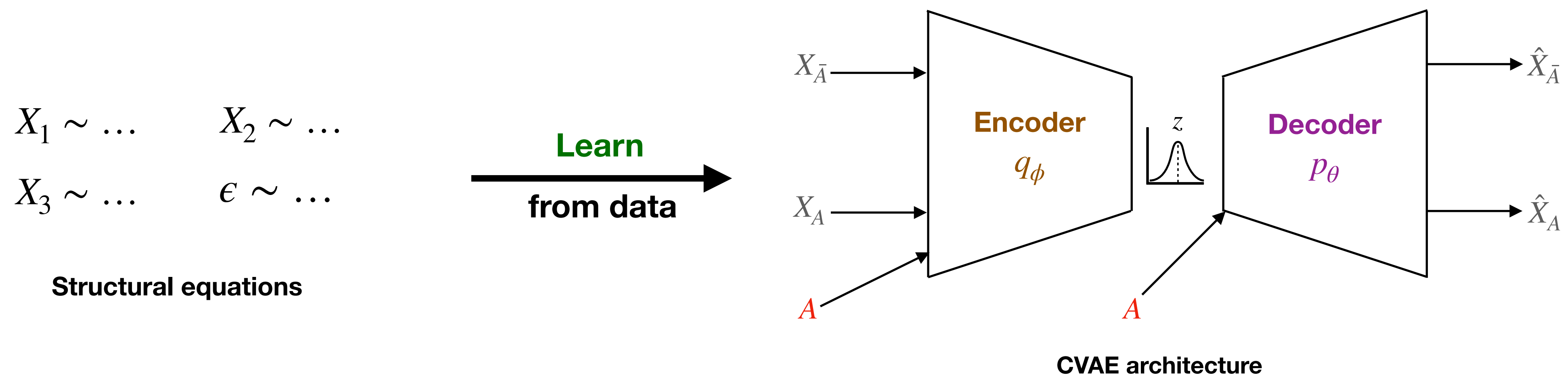
Modeling

1. **Intervene** on A for counterfactual fairness

   → *learn generative process **conditioned** on* A

2. Estimate total effect of causal hidden factors in data

   → *use **latent-space** generative modeling*



Use **C**onditional **V**ariational **A**uto**E**ncoders!

CVAE: Brief Overview

Two deep neural networks:

- **Encoder**: Given data, learn latent z conditioned on A
  - z should match a **prior** distribution *(Gaussian)*

CVAE: Brief Overview

Two deep neural networks:

- **Encoder**: Given data, learn latent z conditioned on A
  - z should match a **prior** distribution *(Gaussian)*
- **Decoder**: Generate realistic data given A and z
  - Faithful **reconstruction** of input data

CVAE: Brief Overview

Two deep neural networks:

- **Encoder**: Given data, learn latent z conditioned on A
  - z should match a **prior** distribution *(Gaussian)*
- **Decoder**: Generate realistic data given A and z
  - Faithful **reconstruction** of input data

Train both models **end-to-end** to maximize data-likelihood

CVAE: Brief Overview

Two deep neural networks:

- **Encoder**: Given data, learn latent z conditioned on A
  - z should match a **prior** distribution *(Gaussian)*
- **Decoder**: Generate realistic data given A and z
  - Faithful **reconstruction** of input data

Train both models **end-to-end** to maximize data-likelihood

$$\log p_\theta(X|A) \geq \underbrace{\mathbb{E}_{q_\phi(z|X,A)}[\log p_\theta(X|z,A)]}_{\textbf{Decoder}} - \underbrace{\mathbb{D}_{KL}[q_\phi(z|X,A)\,||\,p(z)]}_{\textbf{Encoder}}$$

Evidence **L**ower **BO**und (**ELBO**) Loss

CVAE Counterfactuals

Causal

CVAE

CVAE Counterfactuals

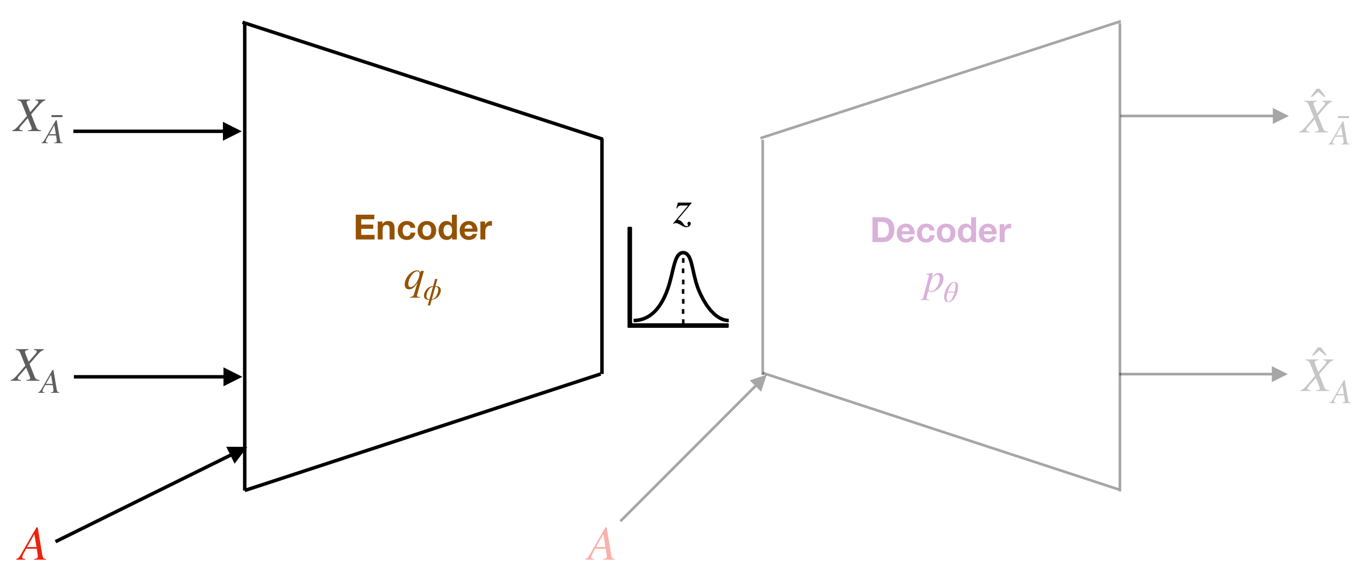
Causal

CVAE

1. Abduction

Infer U from $X, A = a$

Infer z from $X, A = a$



CVAE architecture

CVAE Counterfactuals

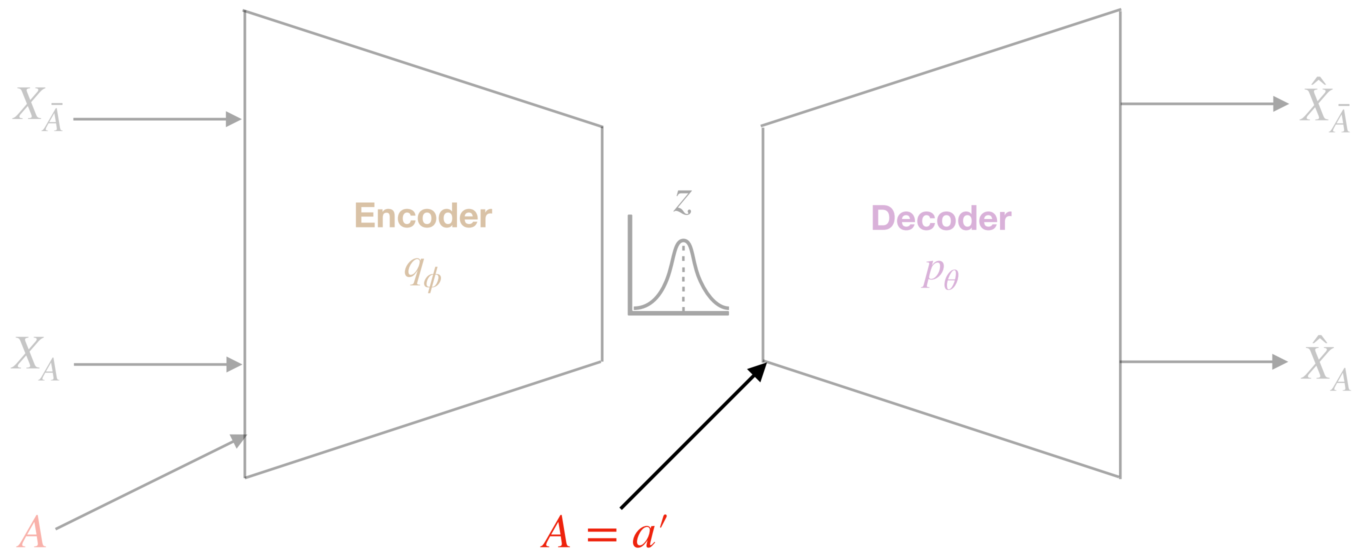
Causal

CVAE

Set $A = a'$ in the graph

2. Action

Set $A = a'$ at decoder



CVAE architecture

CVAE Counterfactuals

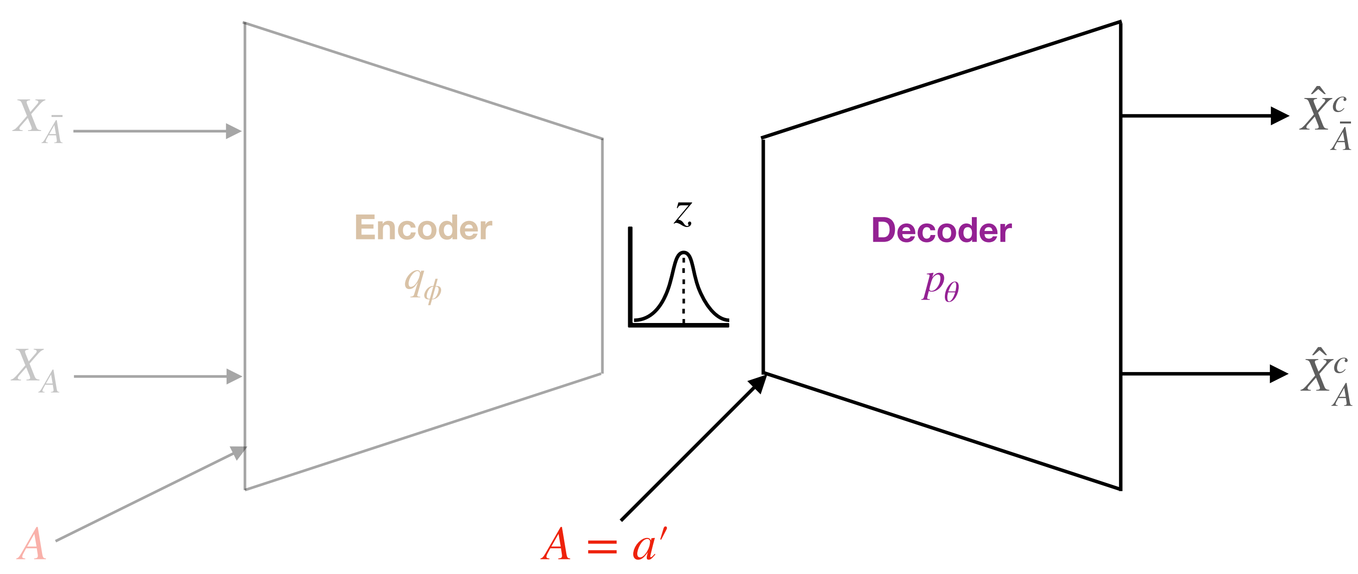
Causal

$\boxed{\text{Get } X^c \text{ from } U \text{ and } A = a'}$

3. Prediction

CVAE

$\boxed{\text{Get } \hat{X}^c \text{ from } z \text{ and } A = a'}$



CVAE architecture

Results

Can we practically operationalize counterfactual fairness?

Baseline Methods

Counterfactual Fairness[1]

- Ideal causal knowledge to generate counterfactuals
- *Use MCMC* for estimation with causal models
- **Flexible**, need **strict** causal assumptions!

FlipTest[2]

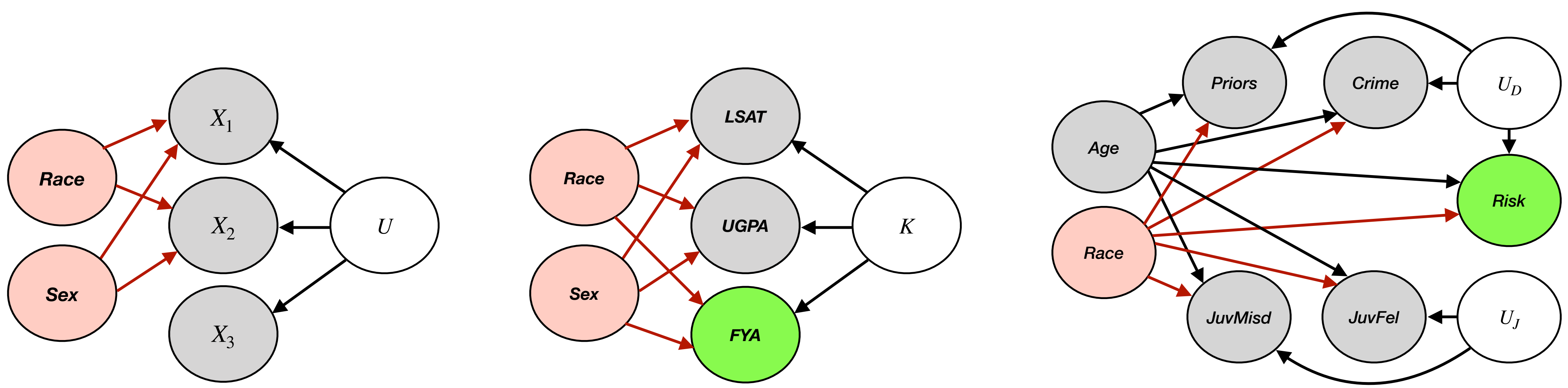
- Approximate counterfactuals via **optimal transport**
- *Use GAN* with **no** latent factor modeling
- **Inflexible**, fewer assumptions but **not clear**!

[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*
[2] *Emily Black et al. "FlipTest: fairness testing via optimal transport": 2020 Conference on Fairness, Accountability and Transparency*

Experimental Setup

Datasets

- Synthetic
  - Various functional models
- Semi-synthetic
  - Law School Admissions
  - COMPAS Recidivism risk



(a) Synthetic (b) Law School Admissions (c) COMPAS Recidivism

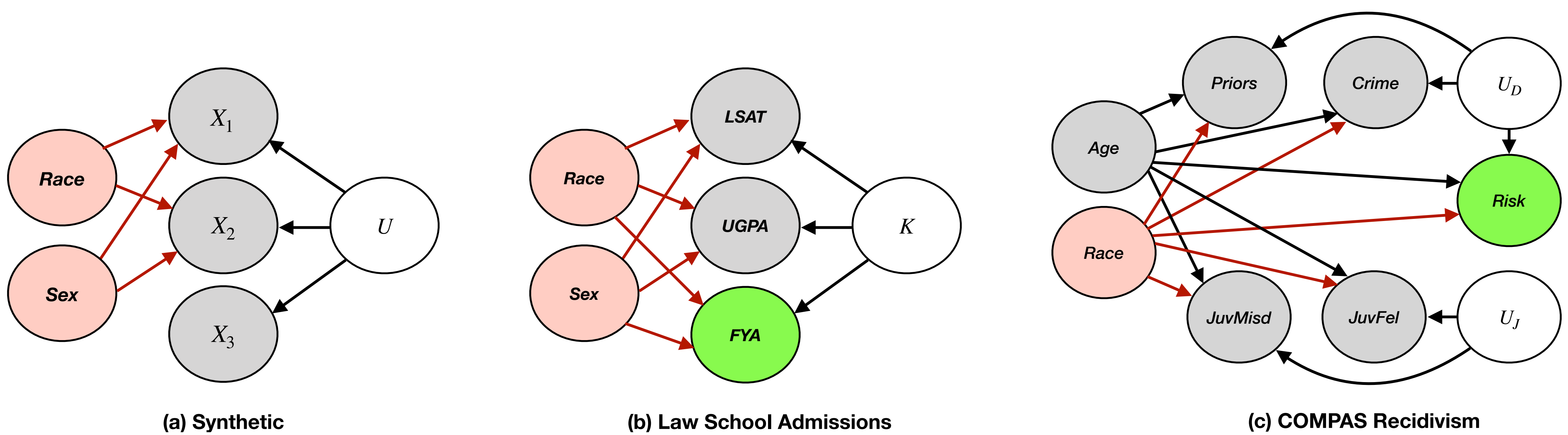
Experimental Setup

Datasets

- Synthetic
  - Various functional models
- Semi-synthetic
  - Law School Admissions
  - COMPAS Recidivism risk

Models

- Causal MCMC
  - Varying **causal** assumptions (*ideal, linear*)
- FlipTest GAN
  - Needs training **more** models!
- CVAE (*ours*)



Approximating counterfactuals

- **Goal:** **Faithful** counterfactuals for fairness using **reduced** assumptions
- **Metric:** **Mean absolute error** b/w approx. & ground-truth counterfactuals

$$\textbf{Err} = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \hat{X}_i^c\right|$$

Approximating counterfactuals

- **Goal: Faithful** counterfactuals for fairness using **reduced** assumptions
- **Metric: Mean absolute error** b/w approx. & ground-truth counterfactuals

$$\textbf{Err} = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \hat{X}_i^c\right|$$

Dataset	MCMC-ideal	MCMC-linear	FlipTest	CVAE
Synthetic *(Non-linear)*	0.0035 +/- 0.0005	0.035 +/- 0.012	0.033 +/- 0.007	0.008 +/- 0.002
Synthetic *(Non-additive)*	0.022 +/- 0.002	0.023 +/- 0.005	0.042 +/- 0.004	0.021 +/- 0.001
Law School	0.27 +/- 0.001	0.32 +/- 0.02	0.3 +/- 0.02	0.25 +/- 0.011
COMPAS	0.035 +/- 0.018	0.17 +/- 0.03	0.12 +/- 0.016	0.06 +/- 0.012

Counterfactual generation quality **(Race: Black to White)**

Approximating counterfactuals

- **Goal:** **Faithful** counterfactuals for fairness using **reduced** assumptions
- **Metric:** **Mean absolute error** b/w approx. & ground-truth counterfactuals

CVAE can generate faithful counterfactuals!

(Fewer assumptions)

Synthetic *(Non-linear)*	0.0035 +/- 0.0005	0.035 +/- 0.012	0.033 +/- 0.007	0.008 +/- 0.002
Synthetic *(Non-additive)*	0.022 +/- 0.002	0.023 +/- 0.005	0.042 +/- 0.004	0.021 +/- 0.001
Law School	0.27 +/- 0.001	0.32 +/- 0.02	0.3 +/- 0.02	0.25 +/- 0.011
COMPAS	0.035 +/- 0.018	0.17 +/- 0.03	0.12 +/- 0.016	0.06 +/- 0.012

Counterfactual generation quality **(Race: Black to White)**

Can we use generated **counterfactuals** for auditing?

Auditing setup

- **Trained** regression model (*COMPAS*)
  - Predict output score (*recidivism risk*)
  - Audit w.r.t. **race** (*Black*→*White*)

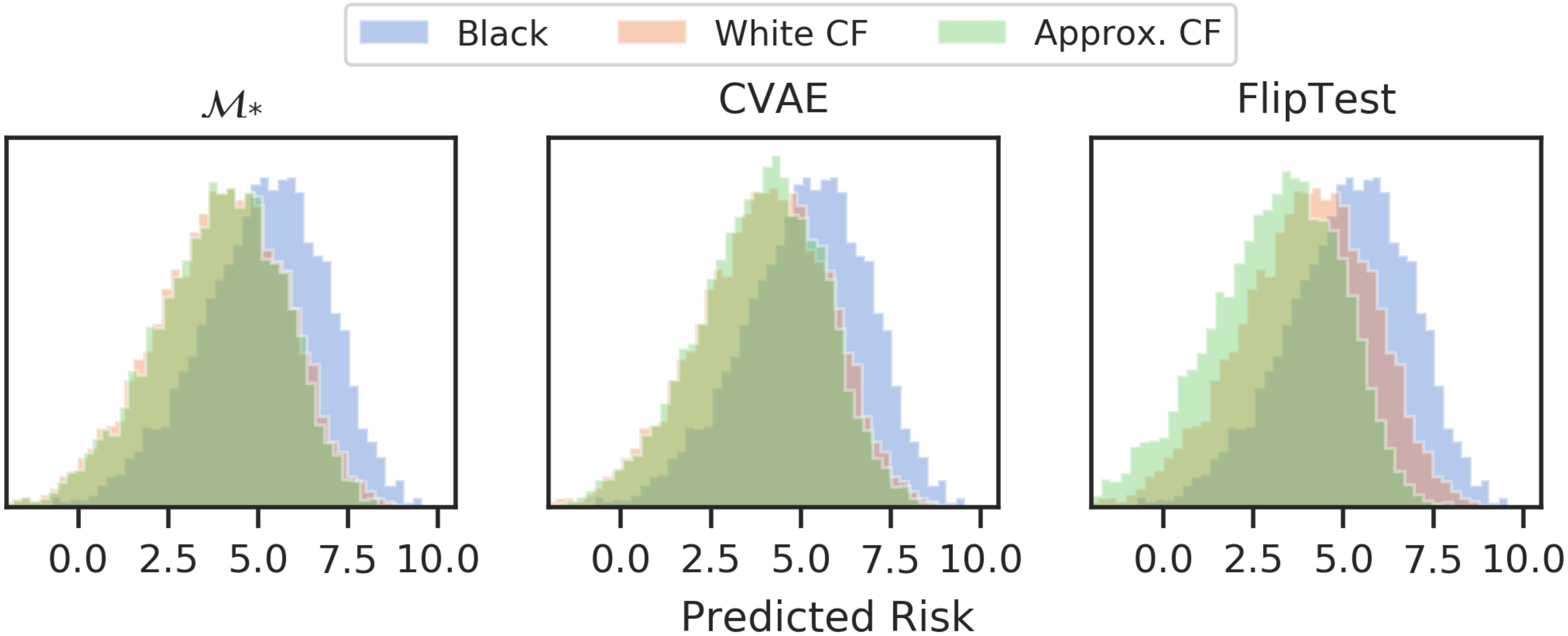
Auditing setup

- **Trained** regression model (*COMPAS*)
  - Predict output score (*recidivism risk*)
  - Audit w.r.t. **race** (*Black*→*White*)
- Audit counterfactual fairness:
  - ***Black** inmate was predicted to have risk of 9.*
  - *If inmate was **white** instead, would the **predicted risk change**?*

Auditing setup

- **Trained** regression model (*COMPAS*)
  - Predict output score (*recidivism risk*)
  - Audit w.r.t. **race** (*Black*→*White*)
- Audit counterfactual fairness:
  - ***Black*** *inmate was predicted to have risk of 9.*
  - *If inmate was **white** instead, would the **predicted risk change**?*
- Approximated counterfactuals to audit model
  - *How well can we match the **true causal** auditing?*

Audit counterfactual fairness

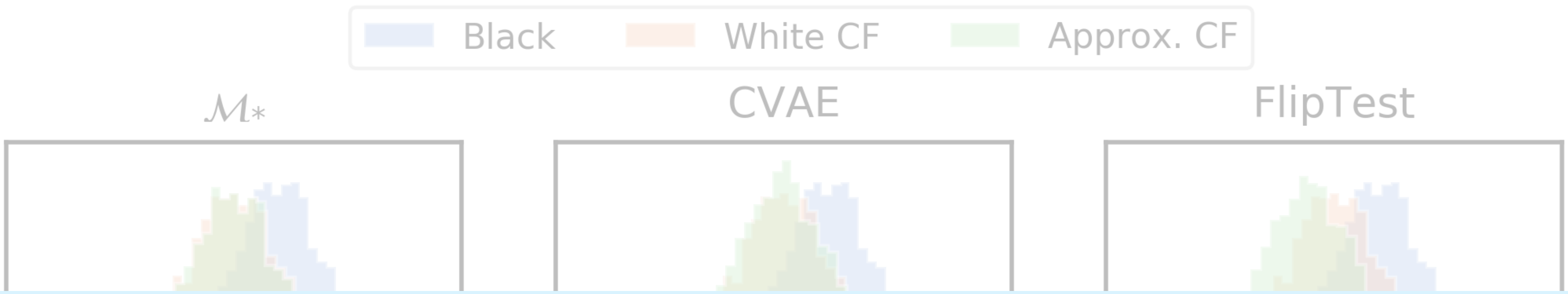


Black → White : : Predicted risk reduces!

Model biased negatively towards blacks!

FlipTest inaccurate, mismatch in auditing!

Audit counterfactual fairness



CVAE auditing $\simeq$ True causal auditing

(Fewer assumptions)

Black $\rightarrow$ White : : Predicted risk reduces!

Model biased negatively towards blacks!

FlipTest inaccurate, mismatch in auditing!

Can we train a **fair** predictive system using our model?

Fair predictor setup

★ **Goal:** Train a fair predictive model (*Law School*)

Compare following models:

Fair predictor setup

★ **Goal:** Train a fair predictive model (*Law School*)

Compare following models:

- **Full**: Use all data features (incl. A)
- **Unaware**: Use all features except A

Fair predictor setup

★ **Goal:** Train a fair predictive model (*Law School*)

Compare following models:

- **Full**: Use all data features (incl. A)
- **Unaware**: Use all features except A
- **Fair-U**: Train on ideal MCMC hidden U

Fair predictor setup

★ **Goal:** Train a fair predictive model (*Law School*)

Compare following models:

- **Full**: Use all data features (incl. A)
- **Unaware**: Use all features except A
- **Fair-U**: Train on ideal MCMC hidden U
- **Fair-z**: Train on CVAE latent z

Fair predictor setup

★ **Goal:** Train a fair predictive model (*Law School*)

Compare following models:

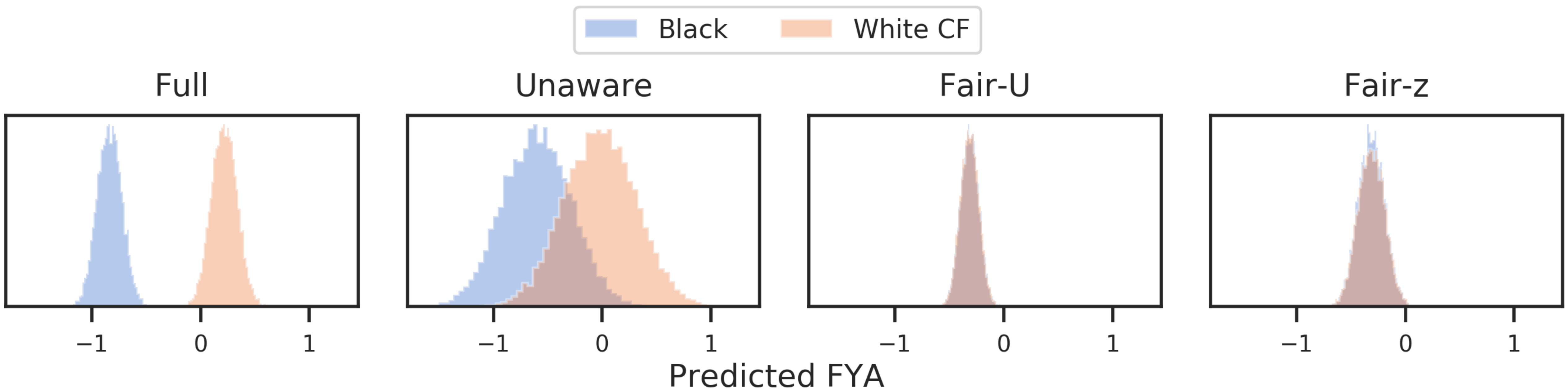
- **Full**: Use all data features (incl. A)
- **Unaware**: Use all features except A
- **Fair-U**: Train on ideal MCMC hidden U
- **Fair-z**: Train on CVAE latent z

Metrics:

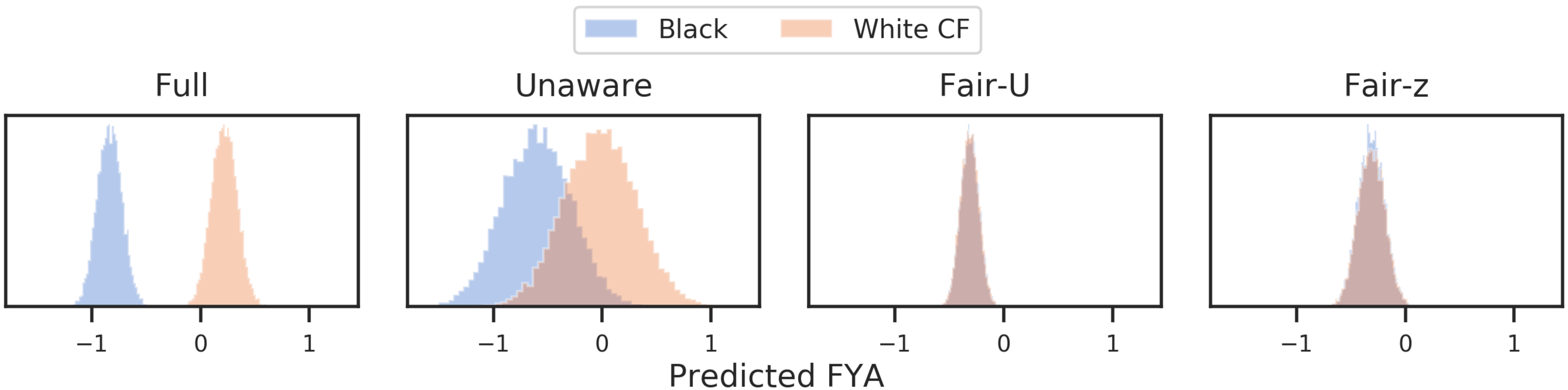
- **Accuracy**: Root mean squared error (*RMSE*)
- **Unfairness**: Absolute difference in outcome to counterfactual

Use data and its **causal counterfactual** for testing

Training fair predictor

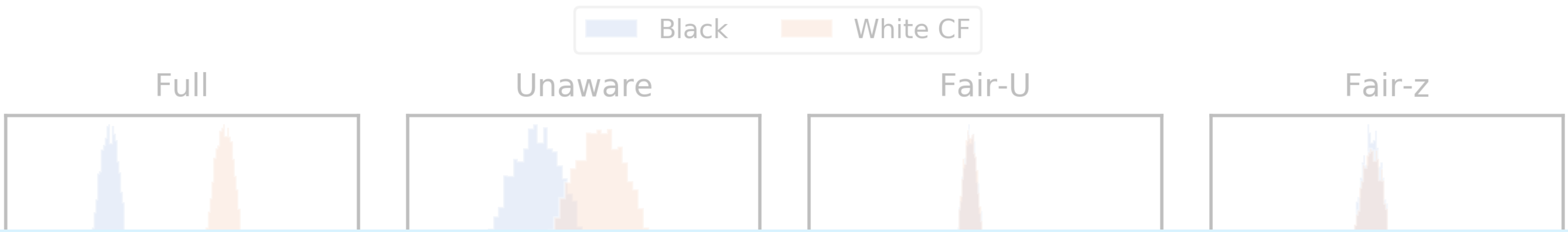


Training fair predictor



Model	Pred. Error (*RMSE*)	Unfairness (*Abs. Diff.*)
Full	**1 (very accurate)**	**1.05 (highly biased)**
Unaware	**1.04 (accurate)**	**0.58 (less biased)**
Fair-U	**1.12 (less accurate)**	**0.01 (fair)**
Fair-z	**1.12 (less accurate)**	**0.01 (fair)**

Training fair predictor



CVAE can be used for fair predictions!

(Fewer assumptions)

Full	1 (very accurate)	1.05 (highly biased)
Unaware	1.04 (accurate)	0.58 (less biased)
Fair-U	1.12 (less accurate)	0.01 (fair)
Fair-z	1.12 (less accurate)	0.01 (fair)

Conclusion

- Causal analysis **useful** for fairness: counterfactual fairness
  - Requires strict assumptions → impractical!
- CVAE **generates** counterfactuals under reduced causal assumptions
  - Possible for scenarios of counterfactual fairness!
- Approximate counterfactuals allow for **reliable** auditing
- CVAE **latent** factors help train **fair** prediction model

Discussion

- Incorporate more assumptions in our approach for other causal fairness definitions
- Analyze scenarios where our assumptions fail/do not hold
- Rethink practical deployment, legal and societal factors
- Study human experts' rating of counterfactual mappings

Thank you!