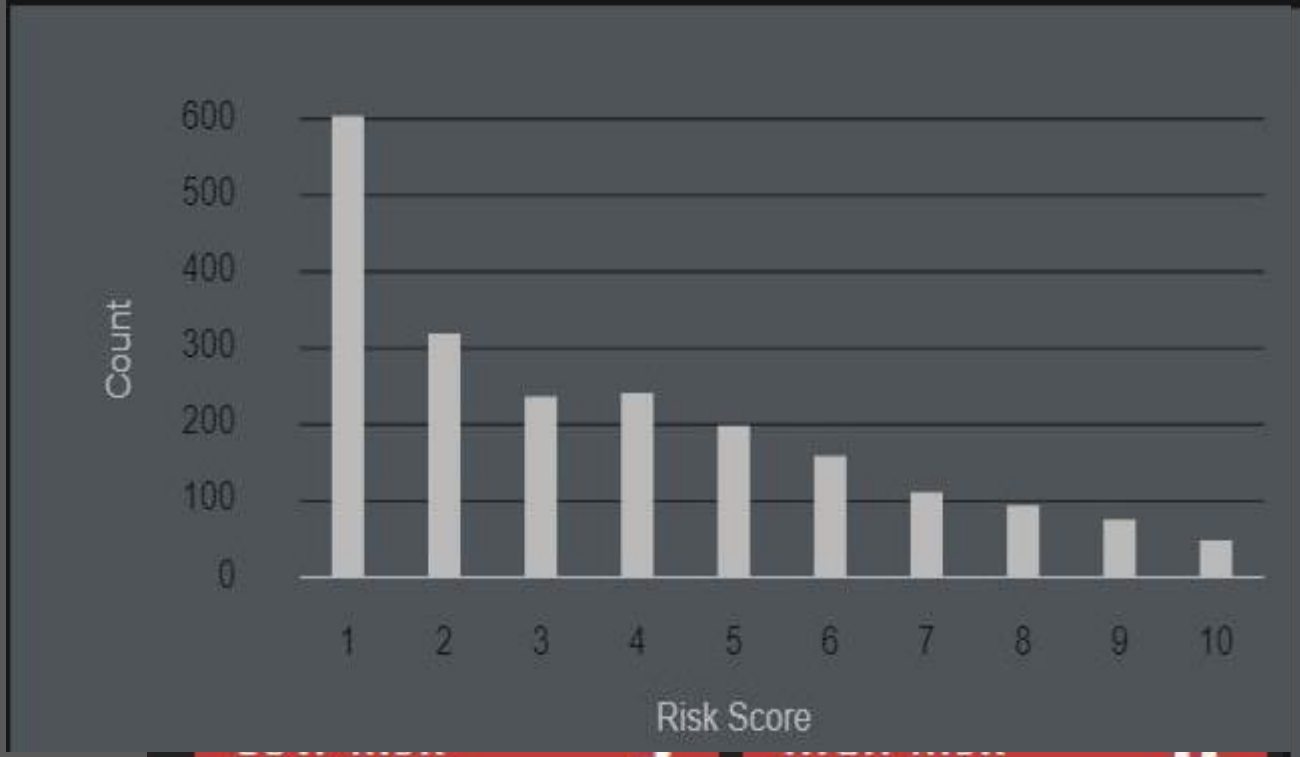
Towards Fairness, Accountability & Transparency in Algorithmic Decision Making

BHAVYA GHAI
PhD Student, Computer Science Department
Adviser: Klaus Mueller
STRIDE Adviser: Liliana Davalos

Machine Bias

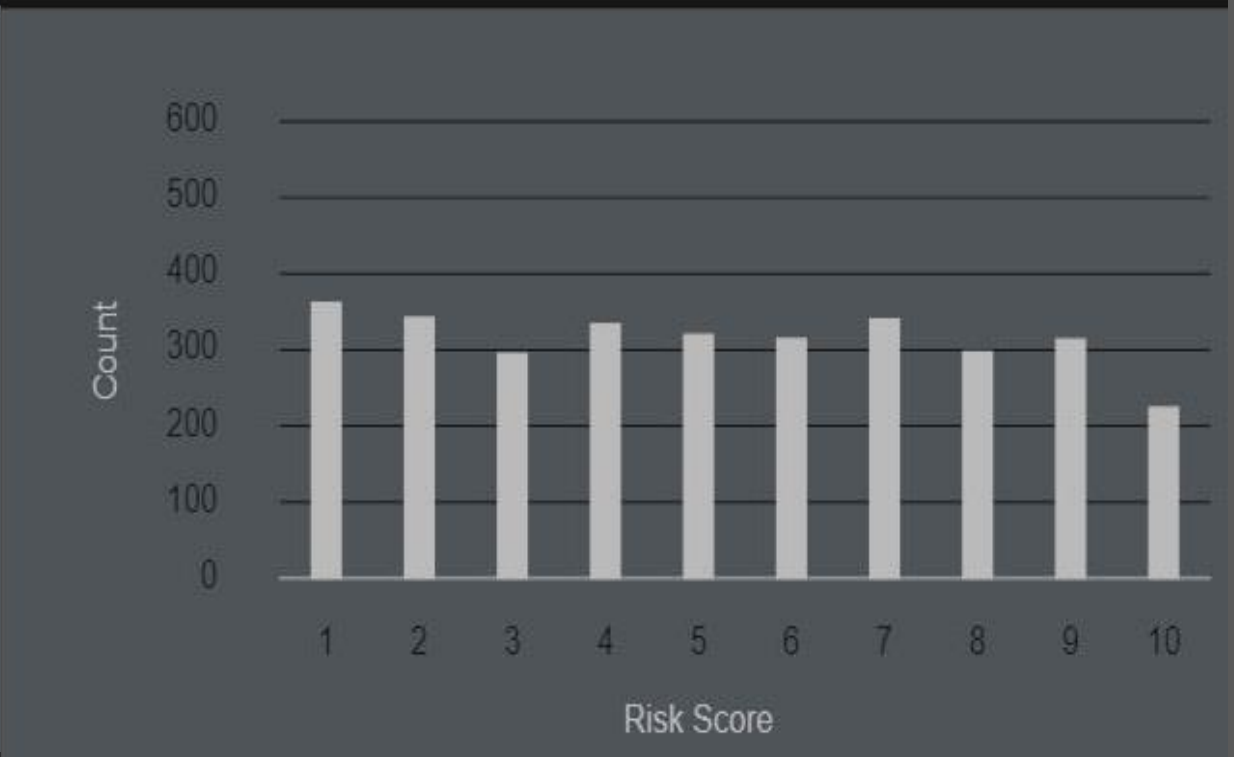
There's software used across the country to predict future criminals. And it's biased against blacks.

White Defendants' Risk Scores



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Black Defendants' Risk Scores



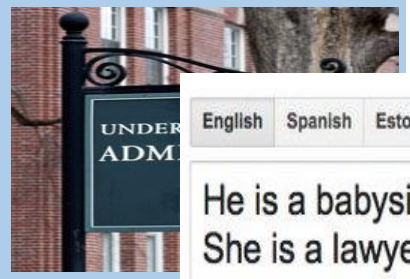
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

How Algorithmic Bias is impacting Society?

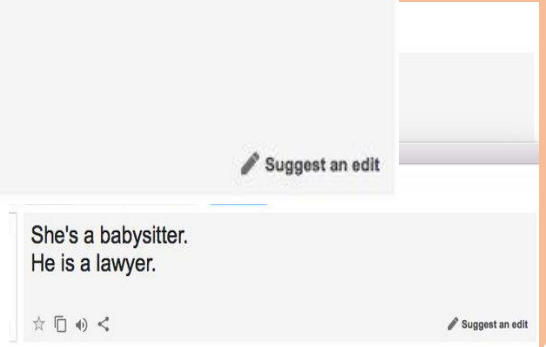
Allocat

tion Harms



Recid

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap



Algorithms are trying to replicate the bias encoded in data

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

In the media …

WIRED

BRIAN BARRETT SECURITY 07.26.18 04:59 PM

LAWMAKERS CAN'T IGNORE FACI RECOGNITION

The New

CNN tech

BUSINE TheUnshot

HIDDEN

. AI is hurtin Whe

Experts wa By Claire

New study uncovers genda

Whe Help

Biased Algorithms Are Everywhere, and No One Seems to Care

Intelligent Machines

Forget Killer Robots—Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.

kills conservative news feeds,

lgorithm mistakenly

ople 'gorillas'

With a Bad

Algorithm

The Value-Added Model has done more to confuse and oppress tha

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Existing work

Data Model Interpretation

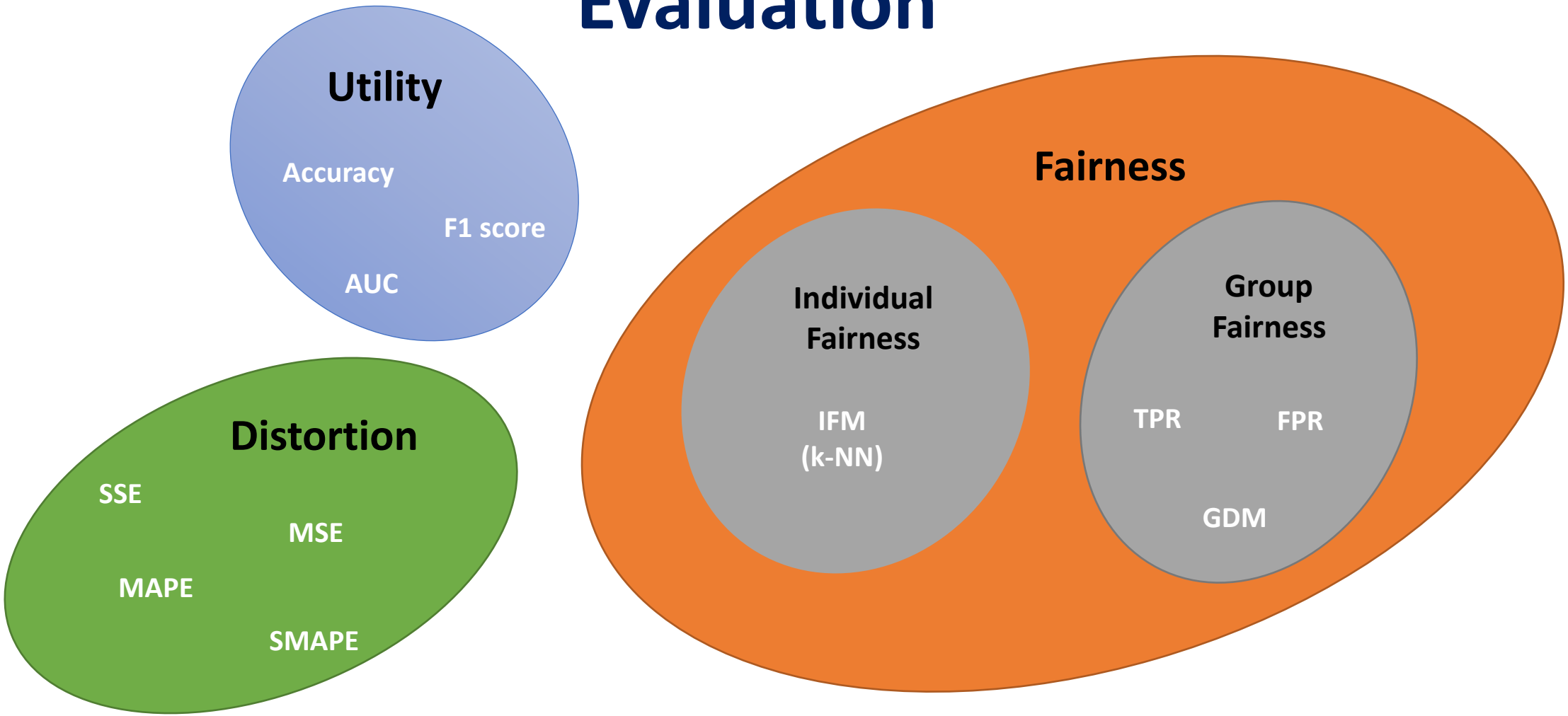
- Data Stage
  - Fairness through unawareness
  - Sampling/Re-weighting
  - Modifying output variable
  - Non-interpretable transformations
- Model Phase
  - Add constraints to loss function
  - Regularization

CGPA	GRE_Verbal	TOEFL	International	Admitted
3.5	168	117	No	✓
3.7	165	119	No	✓
3.4	167	118	No	✓
3.8	155	106	Yes	✗
3.9	160	108	Yes	✗
3.7	157	110	Yes	✗

Synthetic Admissions data

Dealing with Bias at the Data stage provides most flexibility

Evaluation



Preserve utility, maximize fairness & minimize distortion

Gaps in Literature

Fairness

Transparency

Accountability

Domain Knowledge

Trust

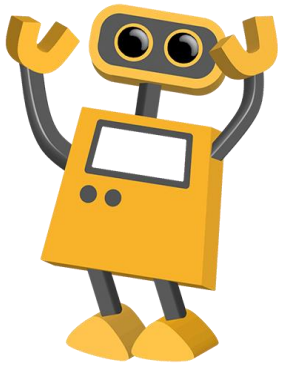
We can’t rely on existing Techniques to take life changing decisions

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Our approach – Human Centered AI

Human

- ✓ Domain Expertise
- ✓ Interpretable
- ✓ Storytelling
- ✗ Expensive
- ✗ Biased
- ✗ Slow



Algorithm

- ✓ Fast
- ✓ Economical
- ✓ Unbiased
- ✗ Opaque
- ✗ Non-culpable
- ✗ No domain Knowledge

- Propose an interactive visual interface to identify and tackle bias
- Understand underlying structures in data using interpretable model like causal inference
- Infuse domain knowledge into the system by modifying causal network
- Evaluate debiased data using Utility, Distortion, Individual fairness & group fairness

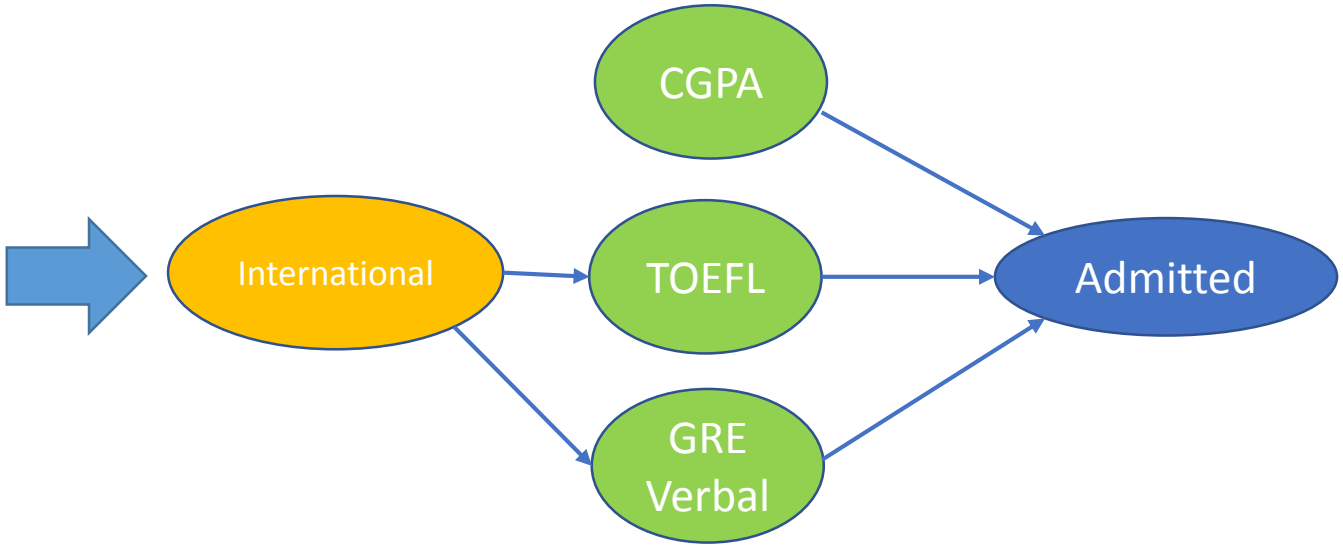
Our approach brings the bests of both worlds!

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

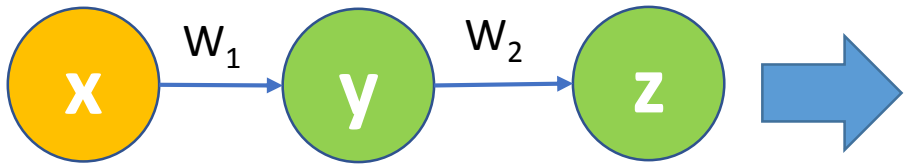
Computational Components

Causal Network

CGPA	GRE_Verbal	TOEFL	International	Admitted
3.5	168	117	No	✓
3.7	165	119	No	✓
3.4	167	118	No	✓
3.8	155	106	Yes	✗
3.9	160	108	Yes	✗
3.7	157	110	Yes	✗



Debiasing



$$y_{new} = y - w_1 x$$

$$z_{new} = z - w_1 w_2 x$$

Causal Networks are interpretable and enable data-driven Storytelling

Computational Components cont.

Dimensionality Reduction

Duration_	Payment_	Purpose	Credit_An	Value_Sav	Length_of	Instalmen	Sex_and_I	Guarantor	Duration_
18	4	2	1049	1	2	4	2	1	4
9	4	0	2799	1	3	2	3	1	2
12	2	9	841	2	4	2	2	1	4
12	4	0	2122	1	3	3	3	1	2
12	4	0	2171	1	3	4	3	1	4
10	4	0	2241	1	2	1	3	1	3
8	4	0	3398	1	4	1	3	1	4
6	4	0	1361	1	2	2	3	1	4
18	4	3	1098	1	1	4	2	1	4
24	2	3	3758	3	1	1	2	1	4
11	4	0	3905	1	3	2	3	1	2
30	4	1	6187	2	4	1	4	1	4
6	4	3	1957	1	4	1	2	1	4
48	3	10	7582	2	1	2	3	1	4
18	2	3	1936	5	4	2	4	1	4
6	2	3	2647	3	3	2	3	1	3
11	4	0	3939	1	3	1	3	1	2

MDS/PCA/TSNE

Evaluation Metrics

Distortion

Symmetric mean absolute percentage error (SMAPE)

Utility

Mean accuracy of an ensemble of ML models

Individual Bias

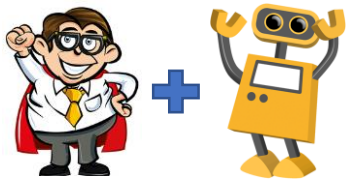
Mean number of neighbors with same label (k-NN)

Group Bias

GDM = $|FPR_{max} - FPR_{min}| + |FNR_{max} - FNR_{min}|$

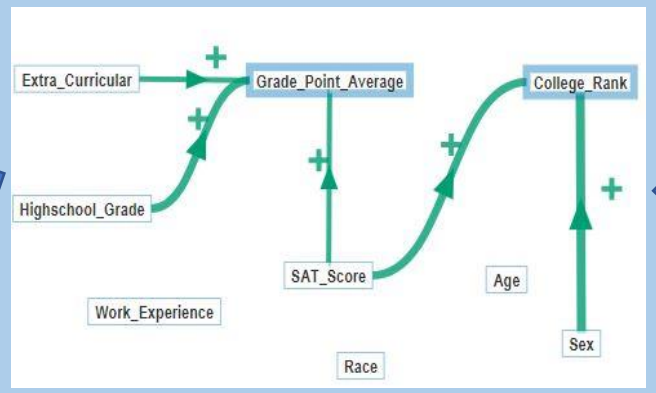
Visual inspection along with evaluation metrics infuses more trust

Proposed Architecture

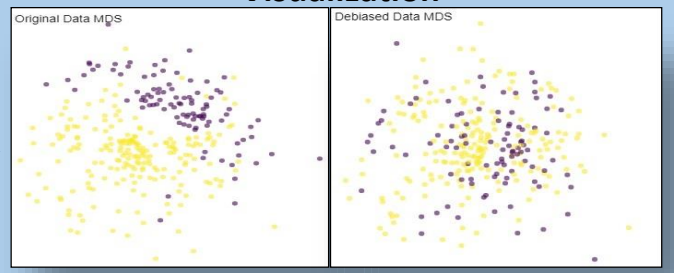


Human Supervision

Causal Network



Visualization

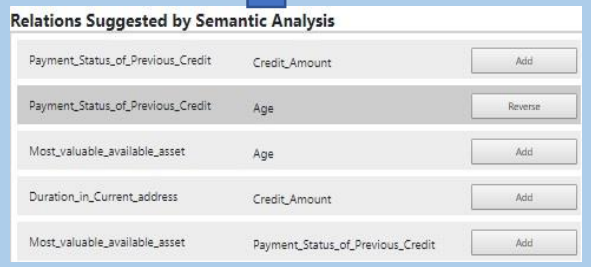


MDS/PCA/TSNE

CSV

Raw data

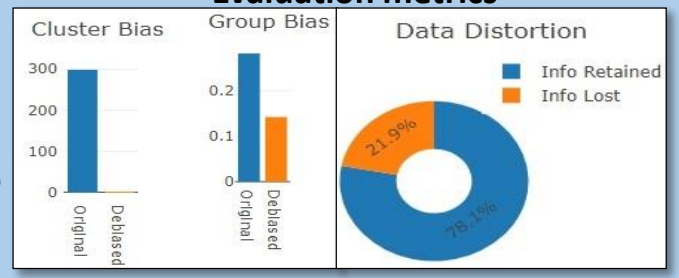
Word2Vec



Semantic Suggestions

Debias

Evaluation metrics



.csv

Debiased data

Humans can infuse domain knowledge by interacting with the causal network

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Our Contribution

Fairness

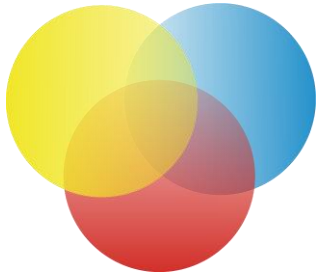
✓ Using multiple fairness definitions

Transparency

✓ Interactive visual interface boosts transparency

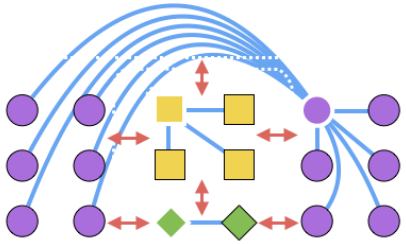
Accountability

✓ Human in-charge can be held accountable



Multidisciplinary

✓ Human expert infuses domain knowledge into system



Data-driven Storytelling

✓ Investigate policies by traversing causal network

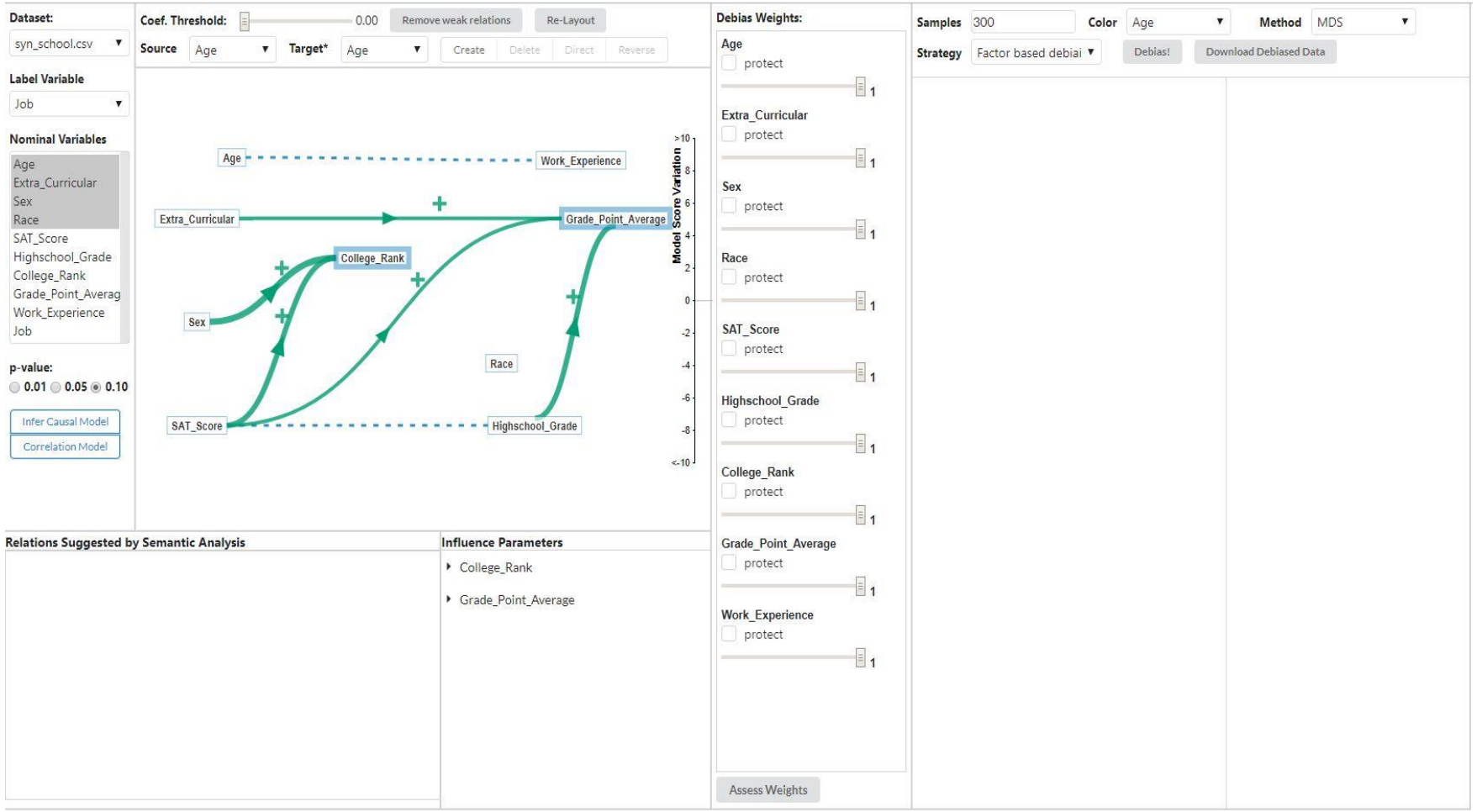
Trust

✓ Human brings more trust into the system

Introducing Human in the loop is the way forward!

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Current state

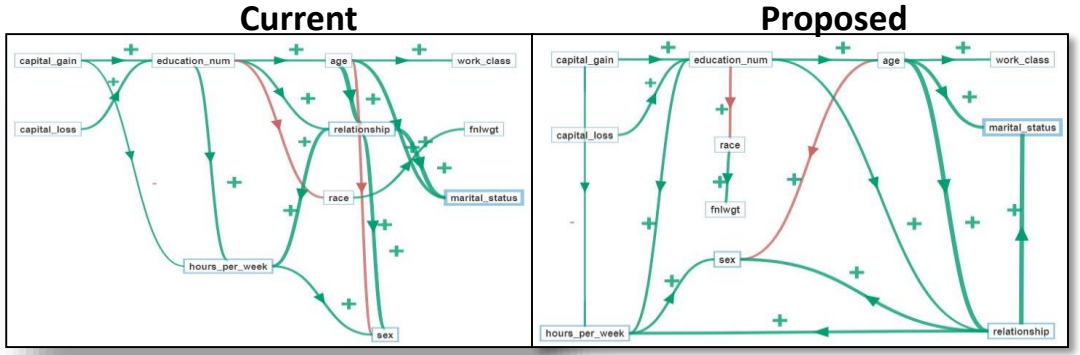


Basic framework along with causal network is implemented

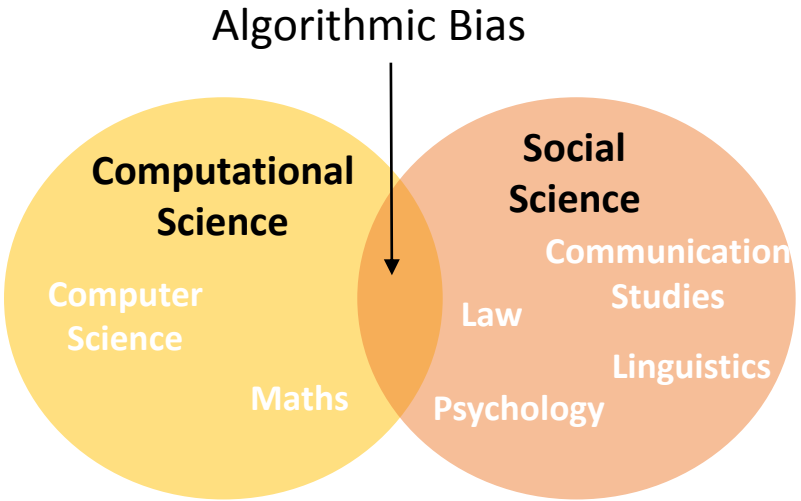
iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Future Work

- Work on different components of the visual interface
- Improve graph layout algorithm to reduce number of intersections



- Improve semantic suggestions by combining with correlation
- Select optimal hyperparameters to calculate utility
- Test our framework on broad set of use cases.
  (IACS collaboration can be very useful here)
- If we get an extension, We will tackle Representation bias & stereotypes



IACS collaboration can give this project new wings!!!

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Conclusion

- Algorithmic Bias is the real AI danger which can have broad social implications
- Existing black box models can’t be used for life changing decisions
- Proposed a novel human centric approach which brings best of both worlds
- Our approach enables humans to monitor, intervene and override if required
- In future, we will test our framework on different use cases & tackle representation bias

Don’t trust algorithms blindly. They can only be as neutral as the training data & the people developing them.

Image: https://www.dreamstime.com/royalty-free-stock-images-finish-line-image29185929

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Thank You …

Image: https://depositphotos.com/99431064/stock-photo-man-hand-writing-any-questions.html

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

References

Biased algorithms are everywhere & no one seems to care

AI programs exhibit racial and gender biases, research reveals

When Algorithms Discriminate

AI is hurting people of color and the poor. Experts want to fix that

How to Fix Silicon Valley's Sexist Algorithms

Houston teachers sue over controversial teacher evaluation method

Algorithms vs Humans

* Algorithms are often implemented **without any appeals method** in place (due to the misconception that algorithms are objective, accurate, and won't make mistakes)

* Algorithms are often used at a much **larger scale** than human decision makers, in many cases, replicating an identical bias at scale (part of the appeal of algorithms is how cheap they are to use)

* Users of algorithms **may not understand probabilities or confidence intervals** (even if these are provided), and may not feel comfortable overriding the algorithm in practice (even if this is technically an option)

* Instead of just focusing on the least-terrible existing option, it is more valuable to ask how we can create **better, less biased decision-making tools** by leveraging the strengths of humans and machines working together

http://www.fast.ai/2018/08/07/hbr-bias-algorithms/

Long term solution

Who code matters?

-- Have diverse teams to cover each others blind spots

How we code matters?

-- Don’t just optimize for accuracy, factor in fairness

Why we code matters?

-- End objective shouldn’t just be profits. Unlock greater equality if social change a priority

Problem Statement

How can we make Algorithmic Decision Making more fair, transparent &

Agenda

- Algorithmic Bias
- Motivation
- Existing Work
- Our Approach
- Demo
- Future Work

Algorithmic Bias

Human

- ✓ Domain Expertise
- ✓ Interpretable
- ✗ Expensive
- ✗ Biased
- ✗ Slow

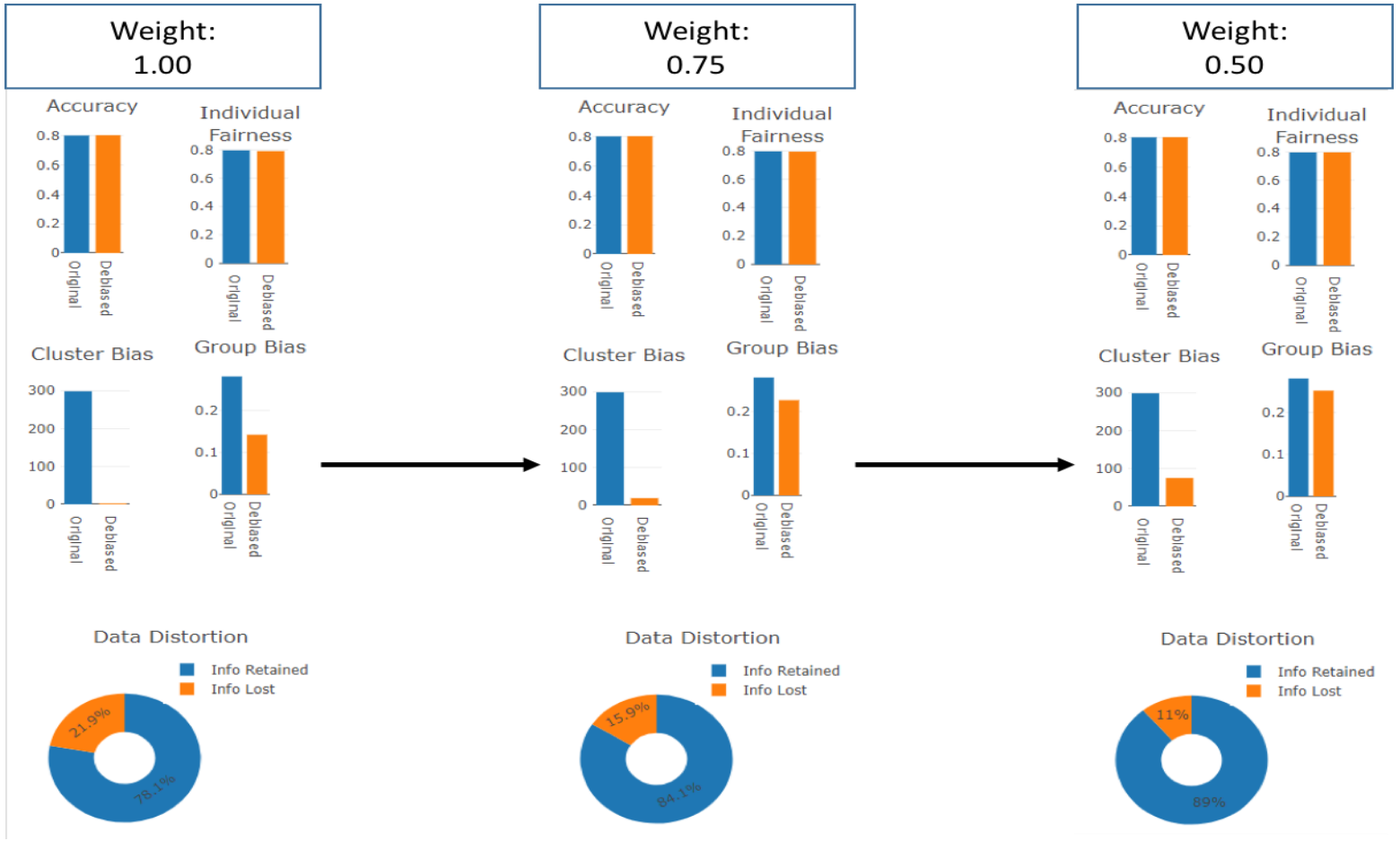
Algorithm

- ✓ Fast
- ✓ Economical
- ✓ Unbiased
- ✗ Biased
- ✗ Opaque
- ✗ Non-culpable

- Algorithms are not intrinsically biased but we are.
- **Type of Bias**: Gender, Race, Age, Personality, etc.
- **Sources of Bias**: Training data, Developers

"Algorithms are opinions expressed in code" – Cathy O'Neil

Partial Debiasing

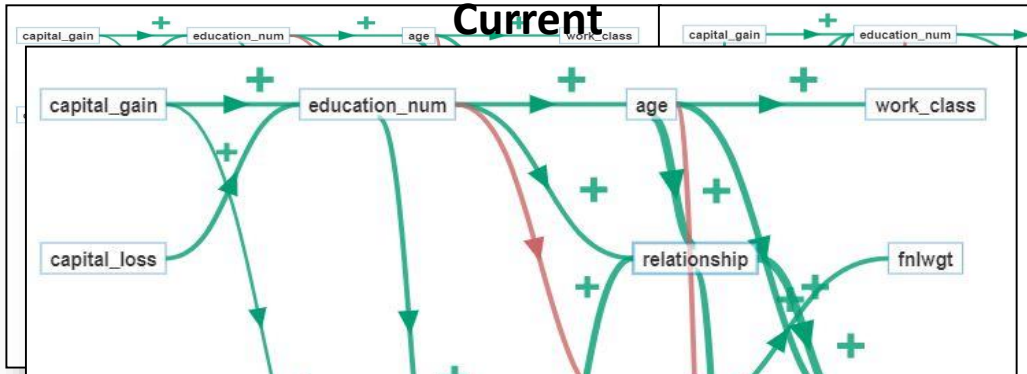


More Fairness causes more data distortion

Future Work

- Improve graph layout algorithm to reduce number of intersections

Current



Proposed

- S
- T

- If we get an extension, We will tackle Representation bias & stereotypes

IACS collaboration can give this project new wings!!!

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE