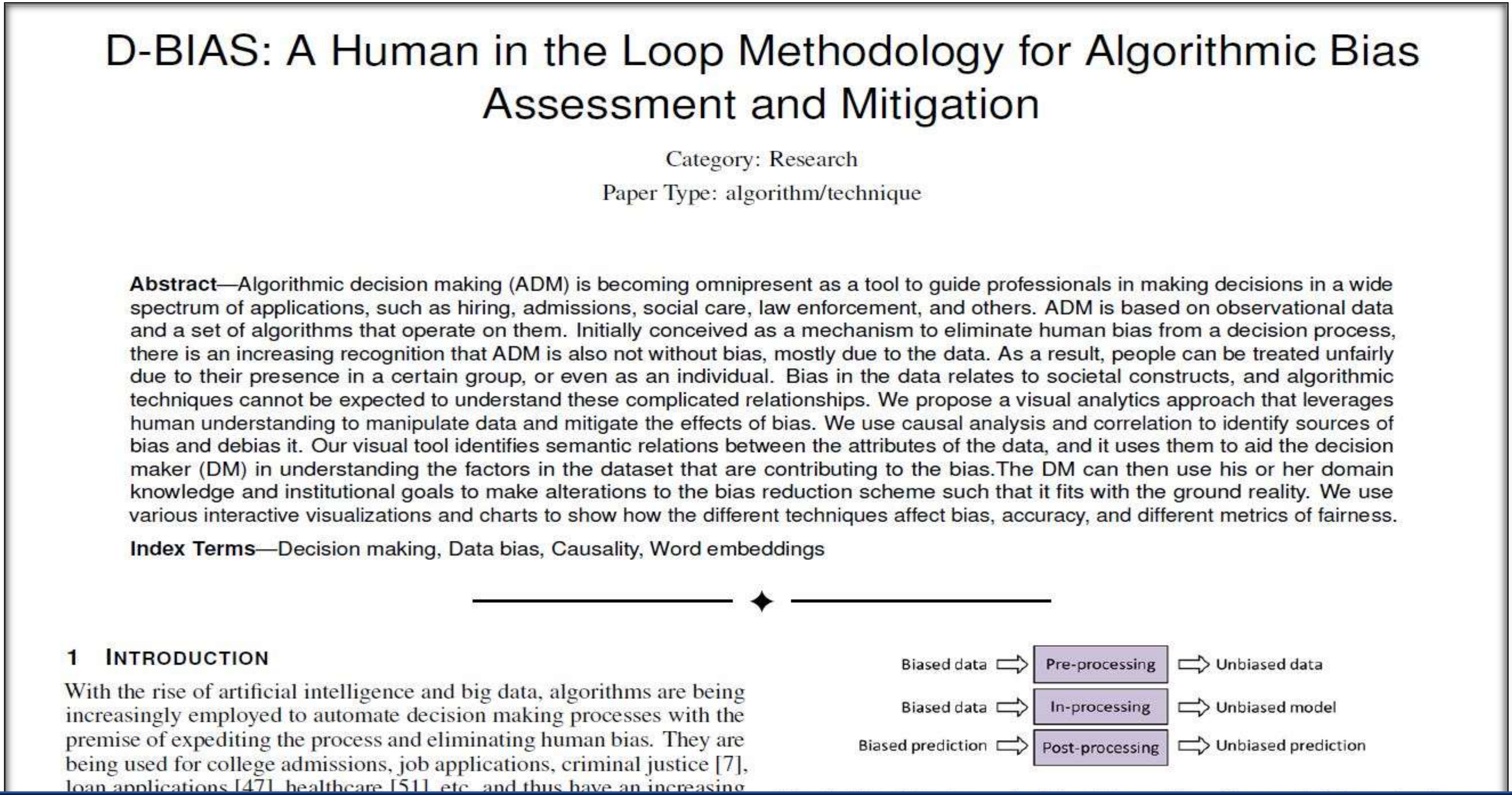
FairText: An Algorithm in the loop writing platform for Inclusive writing

BHAVYA GHAI

PhD Candidate, Computer Science Department

Adviser: Klaus Mueller

Where I left last time …



D-BIAS: A Human in the Loop Methodology for Algorithmic Bias Assessment and Mitigation

Category: Research

Paper Type: algorithm/technique

Abstract—Algorithmic decision making (ADM) is becoming omnipresent as a tool to guide professionals in making decisions in a wide spectrum of applications, such as hiring, admissions, social care, law enforcement, and others. ADM is based on observational data and a set of algorithms that operate on them. Initially conceived as a mechanism to eliminate human bias from a decision process, there is an increasing recognition that ADM is also not without bias, mostly due to the data. As a result, people can be treated unfairly due to their presence in a certain group, or even as an individual. Bias in the data relates to societal constructs, and algorithmic techniques cannot be expected to understand these complicated relationships. We propose a visual analytics approach that leverages human understanding to manipulate data and mitigate the effects of bias. We use causal analysis and correlation to identify sources of bias and debias it. Our visual tool identifies semantic relations between the attributes of the data, and it uses them to aid the decision maker (DM) in understanding the factors in the dataset that are contributing to the bias.The DM can then use his or her domain knowledge and institutional goals to make alterations to the bias reduction scheme such that it fits with the ground reality. We use various interactive visualizations and charts to show how the different techniques affect bias, accuracy, and different metrics of fairness.

Index Terms—Decision making, Data bias, Causality, Word embeddings

1 Introduction

With the rise of artificial intelligence and big data, algorithms are being increasingly employed to automate decision making processes with the premise of expediting the process and eliminating human bias. They are being used for college admissions, job applications, criminal justice [7], loan applications [47], healthcare [51], etc. and thus have an increasing

Biased data ⇨ Pre-processing ⇨ Unbiased data

Biased data ⇨ In-processing ⇨ Unbiased model

Biased prediction ⇨ Post-processing ⇨ Unbiased prediction

Big Thanks to IACS !!!

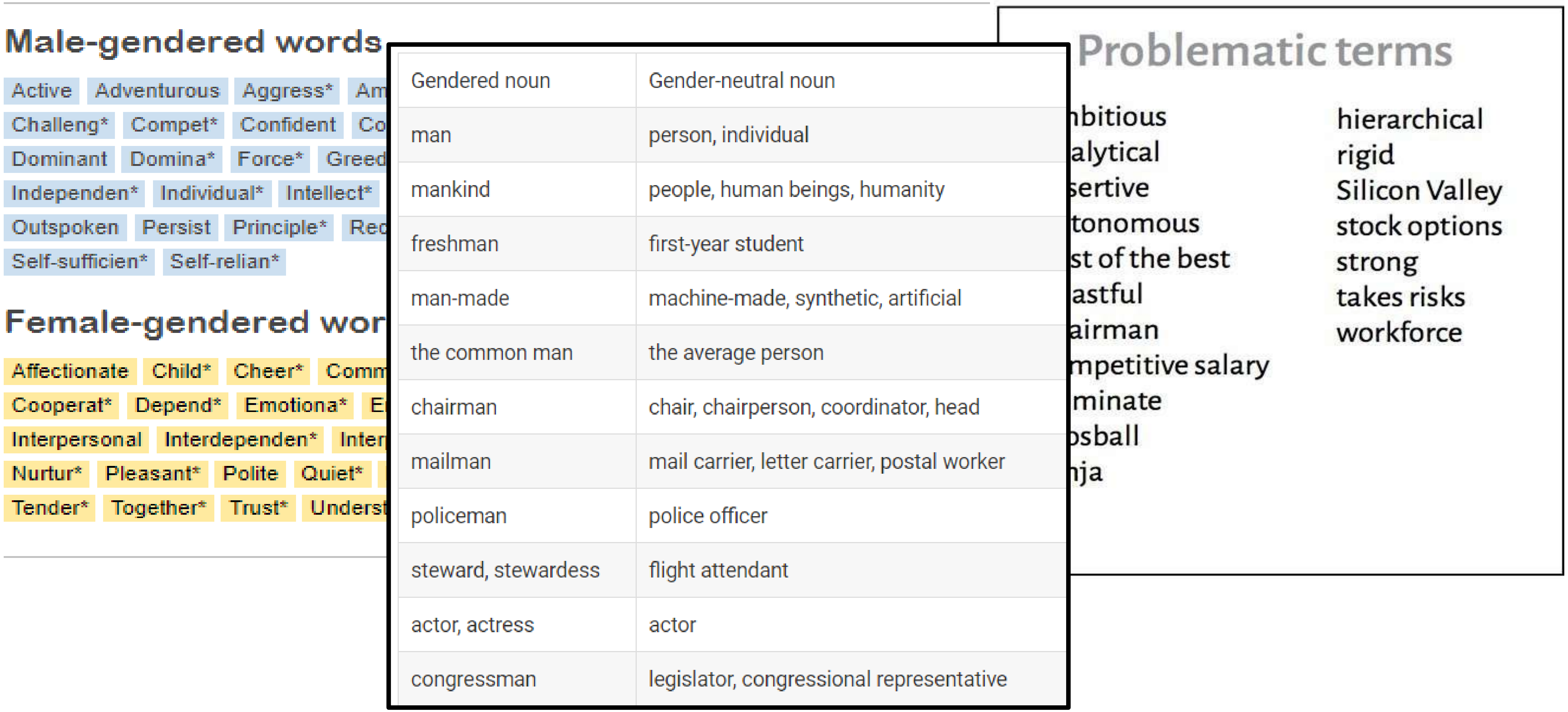
Outline

- Unconscious Bias
- Social Impacts
- Word Embeddings
- Motivation
- Our Approach
- Problem Statement
- Bias Identification
- Ranking Synonym
- Current Status
- Usage Scenarios

Picture a School Teacher …

Everyone holds unconscious beliefs based on past experiences

Unconscious bias in Text



Male-gendered words

Active Adventurous Aggress* Am
Challeng* Compet* Confident Co
Dominant Domina* Force* Greed
Independen* Individual* Intellect*
Outspoken Persist Principle* Rec
Self-sufficien* Self-relian*

Female-gendered wor

Affectionate Child* Cheer* Comm
Cooperat* Depend* Emotiona* E
Interpersonal Interdependen* Inter
Nurtur* Pleasant* Polite Quiet*
Tender* Together* Trust* Underst

Gendered noun	Gender-neutral noun
man	person, individual
mankind	people, human beings, humanity
freshman	first-year student
man-made	machine-made, synthetic, artificial
the common man	the average person
chairman	chair, chairperson, coordinator, head
mailman	mail carrier, letter carrier, postal worker
policeman	police officer
steward, stewardess	flight attendant
actor, actress	actor
congressman	legislator, congressional representative

Problematic terms

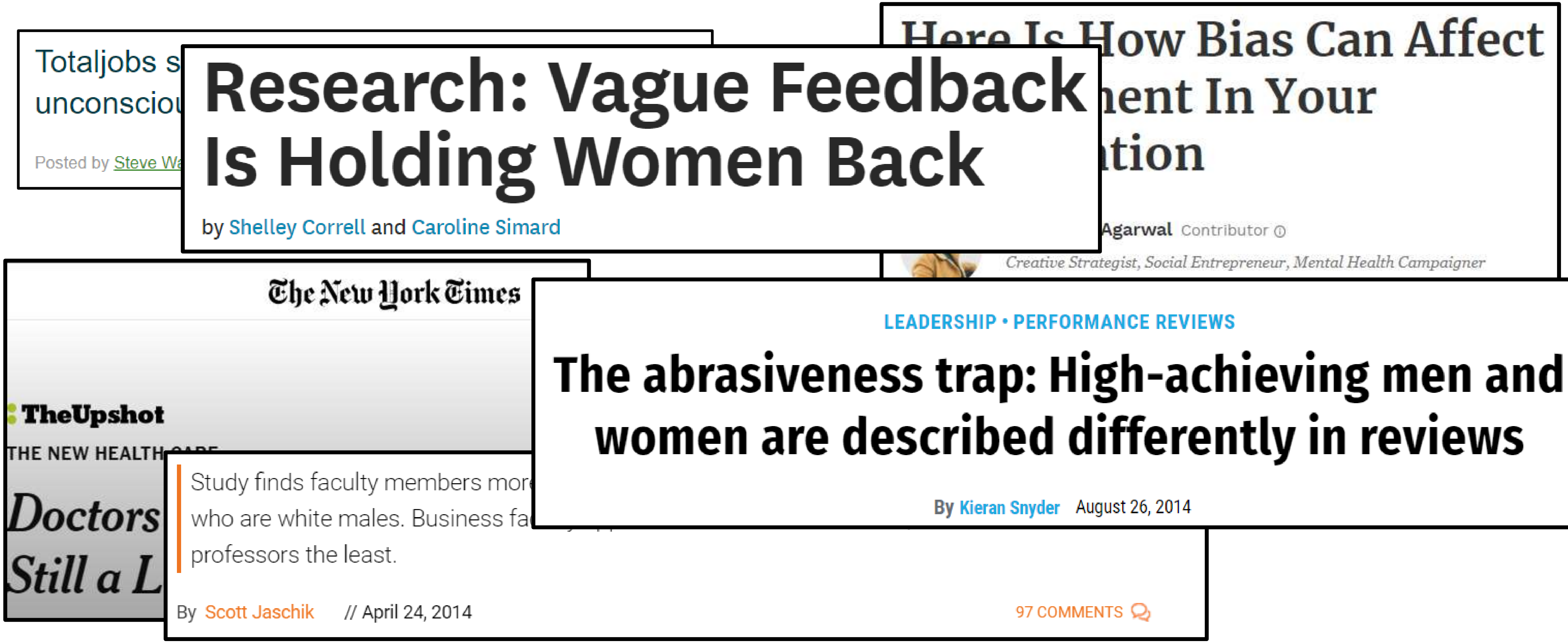
nbitious
alytical
sertive
tonomous
st of the best
astful
airman
mpetitive salary
minate
osball
nja

hierarchical
rigid
Silicon Valley
stock options
strong
takes risks
workforce

Unconscious bias in language isn't always intuitive, but its impact is real

Image: https://textio.ai/watch-your-gender-tone-2728016066ec

Social Impact of Unconscious bias (in text)



Totaljobs s
unconsciou

Posted by Steve Wa

Research: Vague Feedback Is Holding Women Back

by Shelley Correll and Caroline Simard

Here Is How Bias Can Affect
ment In Your
tion

Agarwal Contributor

Creative Strategist, Social Entrepreneur, Mental Health Campaigner

The New York Times

TheUpshot

THE NEW HEALTH CARE

Doctors
Still a L

Study finds faculty members more
who are white males. Business fa
professors the least.

By Scott Jaschik // April 24, 2014

97 COMMENTS

LEADERSHIP • PERFORMANCE REVIEWS

The abrasiveness trap: High-achieving men and women are described differently in reviews

By Kieran Snyder August 26, 2014

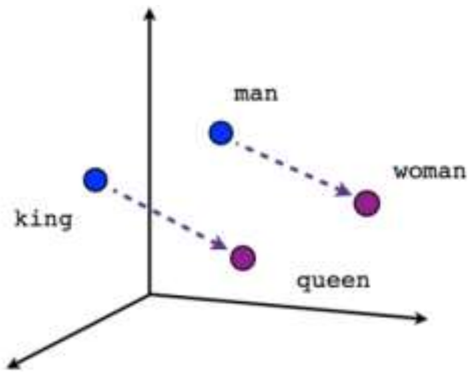
Impact of bias can be felt in all areas including Education, Career, Healthcare, etc.

Image: https://textio.ai/watch-your-gender-tone-2728016066ec

Word Embeddings

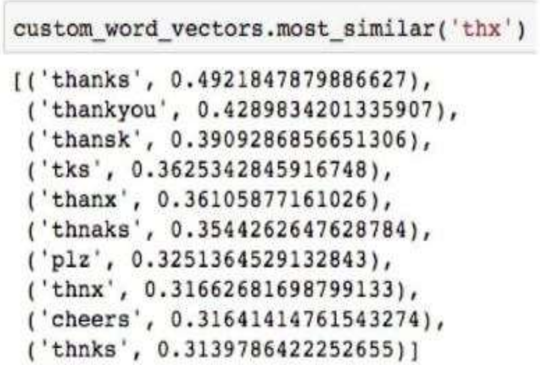
UNSUPERVISED LEARNING

WORD => [0.23, 0.86, 0.19, 0.49,.., .., .., ..]



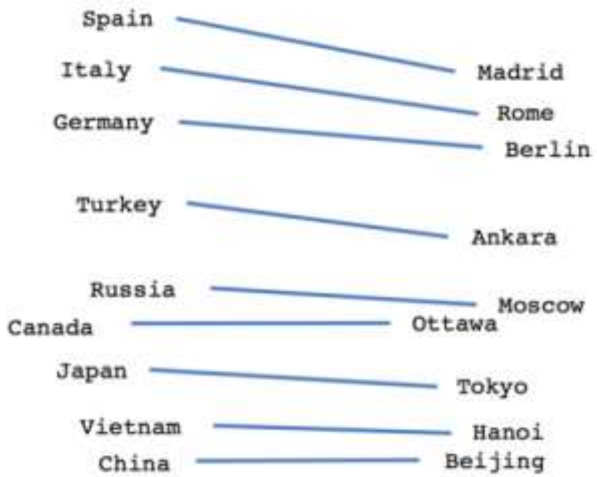
Queen – King + Man = Woman

Popular eg: word2vec, fastText, Glove



```
custom_word_vectors.most_similar('thx')
```

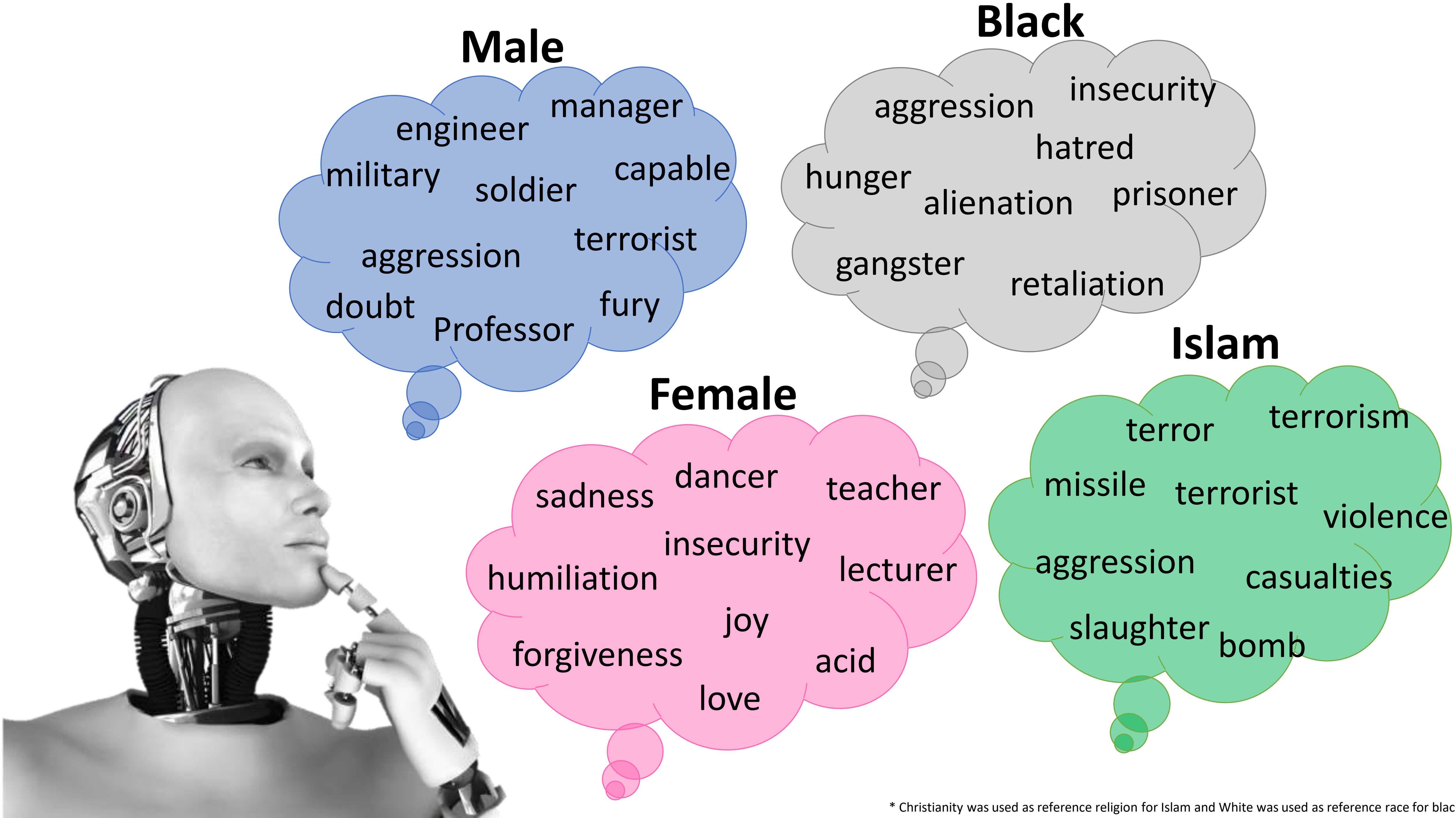
```
[('thanks', 0.4921847879886627),
 ('thankyou', 0.4289834201335907),
 ('thansk', 0.3909286856651306),
 ('tks', 0.3625342845916748),
 ('thanx', 0.36105877161026),
 ('thnaks', 0.3544262647628784),
 ('plz', 0.3251364529132843),
 ('thnx', 0.31662681698799133),
 ('cheers', 0.31641414761543274),
 ('thnks', 0.3139786422252655)]
```



SOME ML MODEL

- Sentiment Analysis
- Machine Translation
- Question Answering
- .
- .
- .
- Named Entity Recognition

Word Embeddings are reflection of society



Male

engineer
manager
military
soldier
capable
terrorist
aggression
doubt
Professor
fury

Black

aggression
insecurity
hatred
hunger
alienation
prisoner
gangster
retaliation

Female

sadness
dancer
teacher
insecurity
humiliation
lecturer
joy
forgiveness
acid
love

Islam

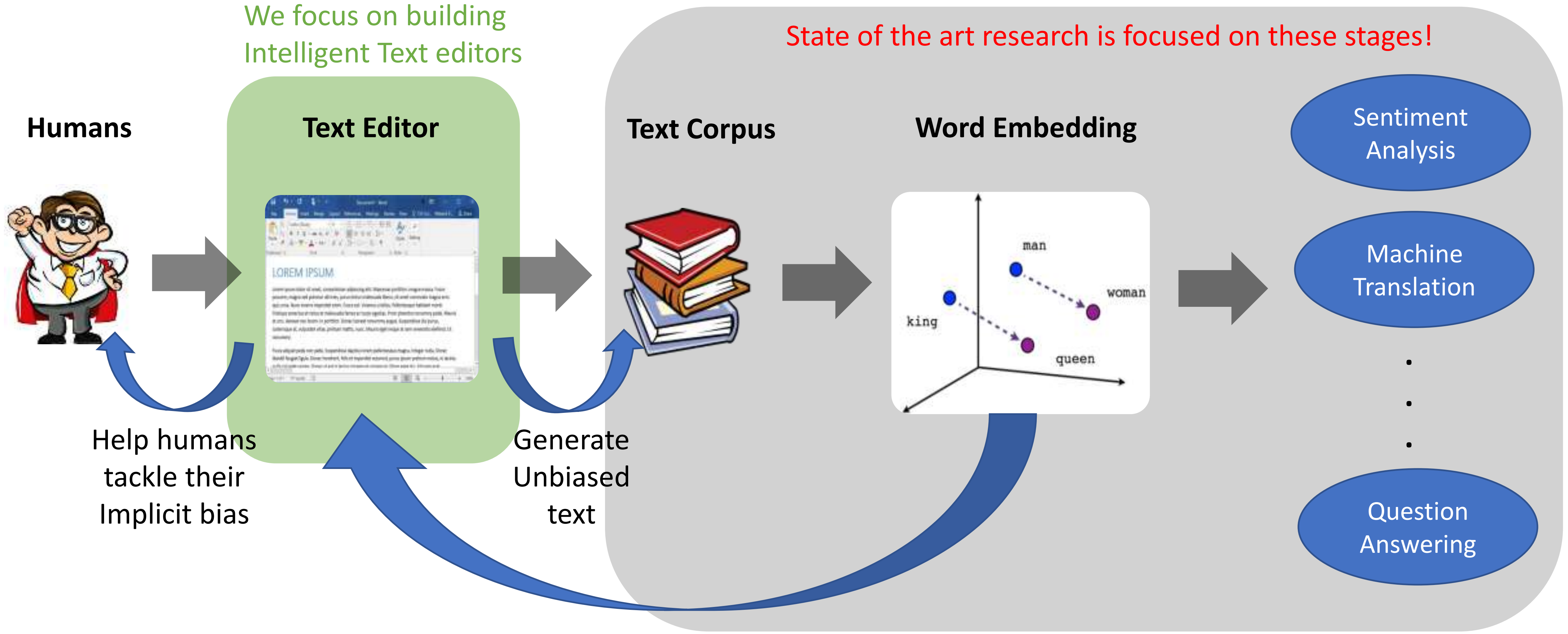
terror
terrorism
missile
terrorist
violence
aggression
casualties
slaughter
bomb

* Christianity was used as reference religion for Islam and White was used as reference race for blacks.

Motivation

- One of the prime ways to tackle Unconscious bias is to make “*the unconscious, conscious*”
- Multiple research papers have established that ML algorithms have captured human like biases against a specific race, gender, etc.
- Specifically, Can we leverage bias encoded in word embeddings for more inclusive writing?

Our Approach



We focus on building Intelligent Text editors

State of the art research is focused on these stages!

Humans

Text Editor

Text Corpus

Word Embedding

Sentiment Analysis

Machine Translation

.
.
.

Question Answering

Help humans tackle their Implicit bias

Generate Unbiased text

We need AI powered Text editors to tackle human bias at its core

Problem Statement

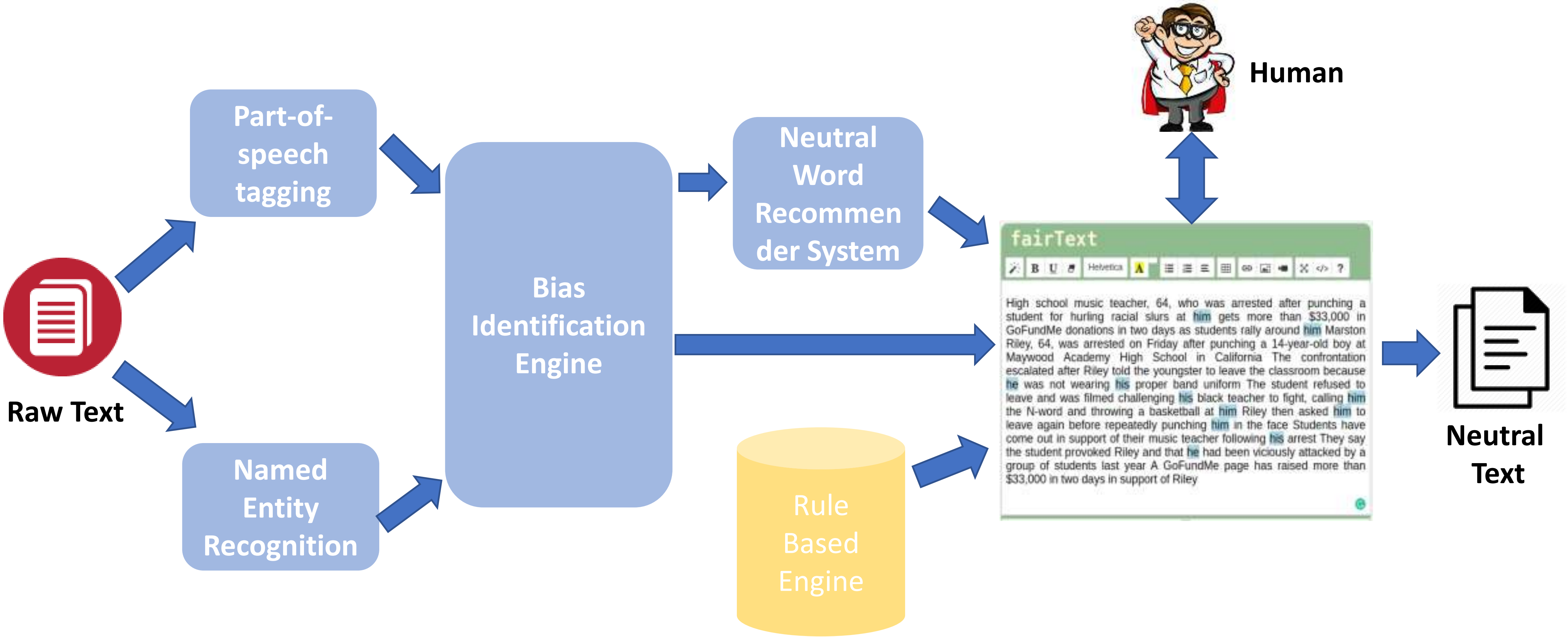
Given some text as input:-

→ Can we identify which words are more likely to incite unconscious bias in the minds of the readers?

→ Can we suggest/recommend alternate words which have similar meaning but doesn’t incite bias?

Return unbiased version of original text with the meaning preserved

Proposed Architecture



Raw Text

Part-of-speech tagging

Named Entity Recognition

Bias Identification Engine

Neutral Word Recommender System

Rule Based Engine

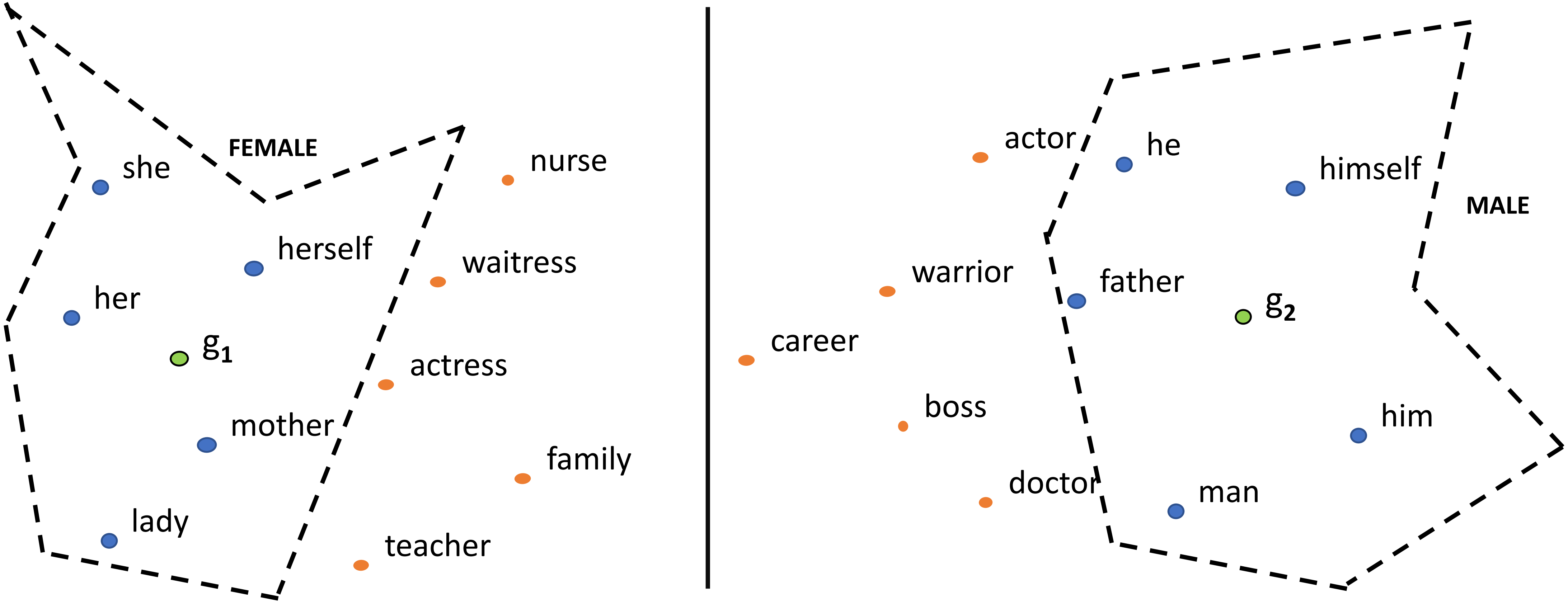
Human

High school music teacher, 64, who was arrested after punching a student for hurling racial slurs at him gets more than $33,000 in GoFundMe donations in two days as students rally around him Marston Riley, 64, was arrested on Friday after punching a 14-year-old boy at Maywood Academy High School in California The confrontation escalated after Riley told the youngster to leave the classroom because he was not wearing his proper band uniform The student refused to leave and was filmed challenging his black teacher to fight, calling him the N-word and throwing a basketball at him Riley then asked him to leave again before repeatedly punching him in the face Students have come out in support of their music teacher following his arrest They say the student provoked Riley and that he had been viciously attacked by a group of students last year A GoFundMe page has raised more than $33,000 in two days in support of Riley

Neutral Text

Human abilities are augmented with knowledge from NLP & Psychology research

Bias Identification



Bias_score(word) = distance(word, g1) - distance(word, g2)

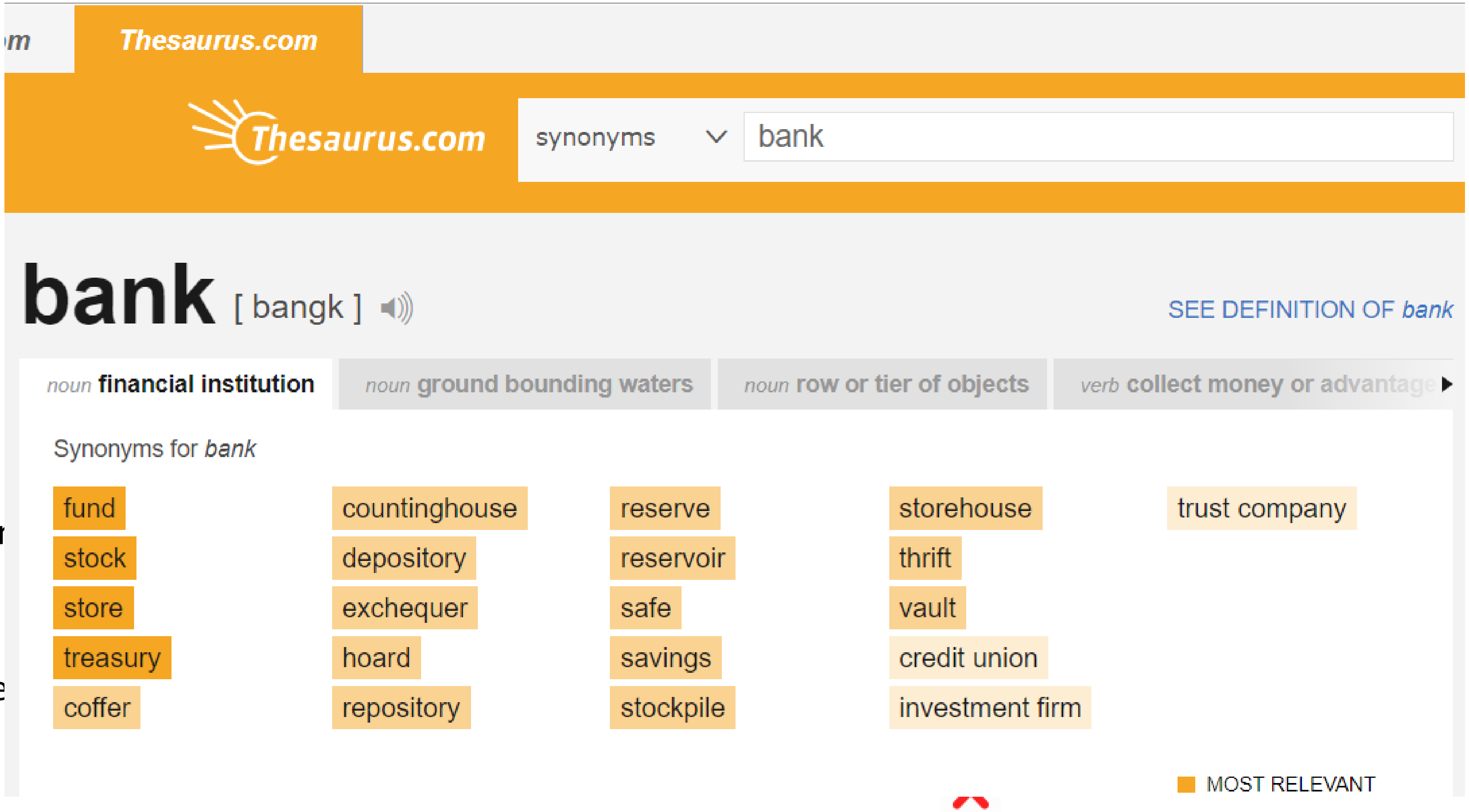
Ideally, neutral words should be equidistant from either cluster

"Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* (2018)

Synonym Retrieval

Thesaurus

- ✓ Good quality synony
- ✗ Limited vocabulary
- ✗ Lacks similarity score



m | Thesaurus.com

Thesaurus.com

synonyms | bank

bank [bangk]

SEE DEFINITION OF *bank*

noun **financial institution** | *noun* ground bounding waters | *noun* row or tier of objects | *verb* collect money or advantage

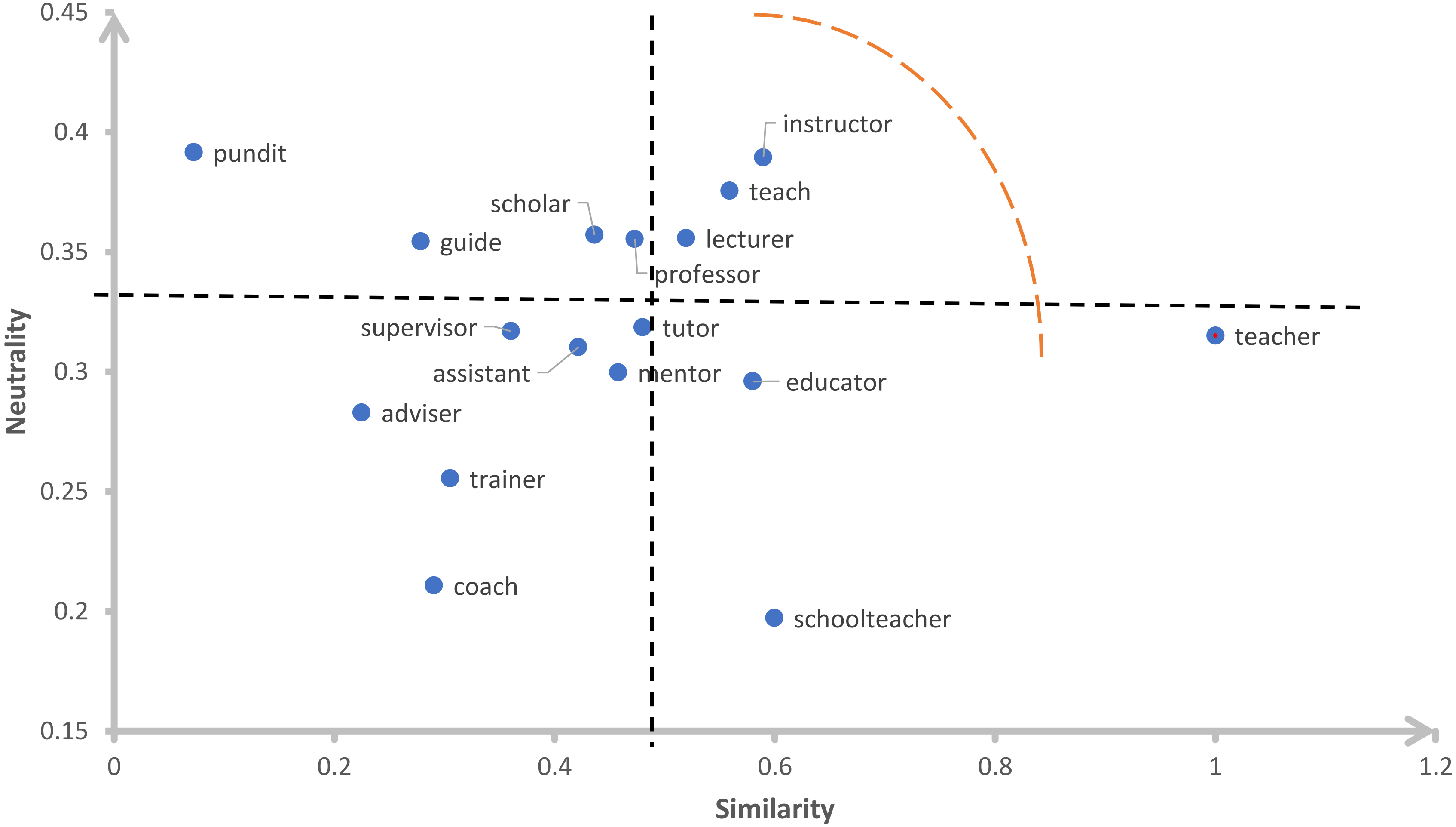
Synonyms for *bank*

fund
stock
store
treasury
coffer
countinghouse
depository
exchequer
hoard
repository
reserve
reservoir
safe
savings
stockpile
storehouse
thrift
vault
credit union
investment firm
trust company

🟧 MOST RELEVANT

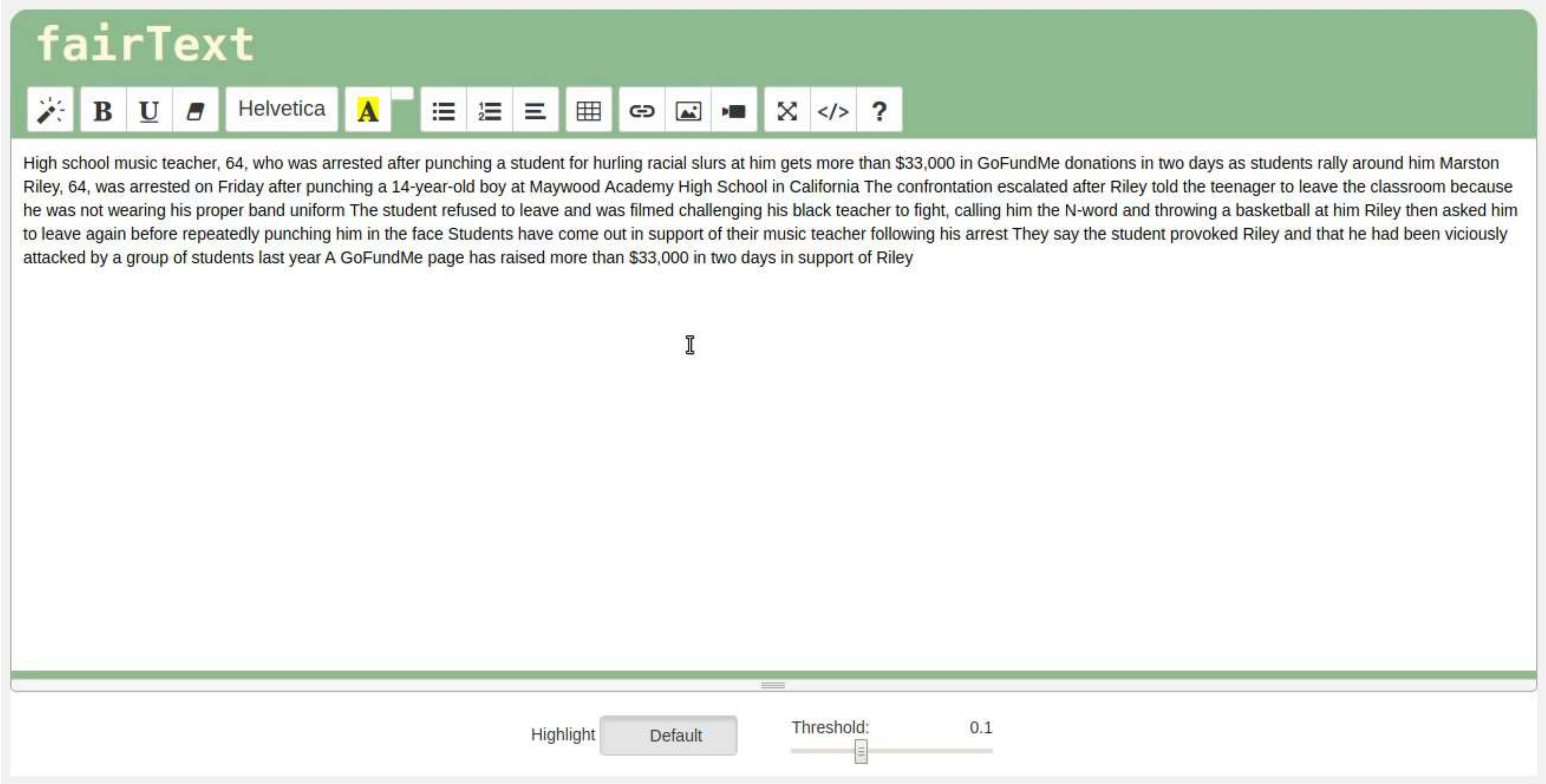
Contextualized word embeddings is the way to go!

Ranking of Synonyms



Finding the right word is a bi-objective optimization problem

Current State



High school music teacher, 64, who was arrested after punching a student for hurling racial slurs at him gets more than $33,000 in GoFundMe donations in two days as students rally around him Marston Riley, 64, was arrested on Friday after punching a 14-year-old boy at Maywood Academy High School in California The confrontation escalated after Riley told the teenager to leave the classroom because he was not wearing his proper band uniform The student refused to leave and was filmed challenging his black teacher to fight, calling him the N-word and throwing a basketball at him Riley then asked him to leave again before repeatedly punching him in the face Students have come out in support of their music teacher following his arrest They say the student provoked Riley and that he had been viciously attacked by a group of students last year A GoFundMe page has raised more than $33,000 in two days in support of Riley

Highlight: Default

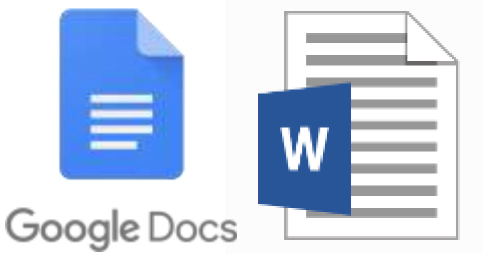
Threshold: 0.1

Basic framework with text highlighting is implemented

Usage Scenarios

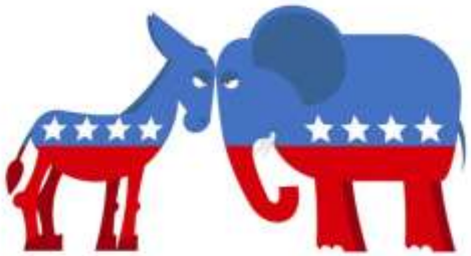
Job Postings

News Articles



Integrate into Mainstream tools

Letter of Recommendation



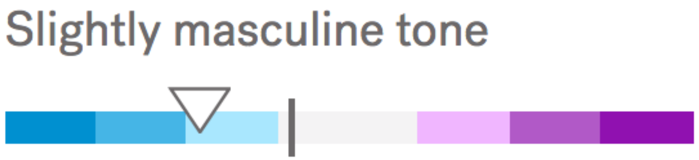
Political Speeches

Natural Language Generation

We haven't made any assumptions on the domain, so possibilities are endless!

Plan of Research

- Improve Bias Identification using POS tagging & Named Entity Recognition
- Improve word recommendations using Contextualized word embedding
- Add overall gender tone meter



- Evaluate using User Study
  - How often do user concur with word highlighting?
  - Do users adopt the suggestion or figure out a new way?
  - Does user demographics play a role?
- Extend to other kinds of bias like political, racial, etc.
- Detect if phrases/sentences are biased.

  E.g.- "Don't be such a drama queen."

It's an uncharted territory & there's a lot to explore!

Conclusion

- Unconscious Bias is ubiquitous & has serious real-world consequences
- Identifying & mitigating bias in language is a tricky Interdisciplinary problem
- We propose a novel approach by leveraging the bias encoded in word embeddings
- Our approach works in real time & can have direct real-world impact as a product
- In future, we will plan to extend to different kinds of biases and detect biases in sentences

Every word matters so choose your words carefully!

iACS
INSTITUTE FOR ADVANCED
COMPUTATIONAL SCIENCE

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

Thank You …

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

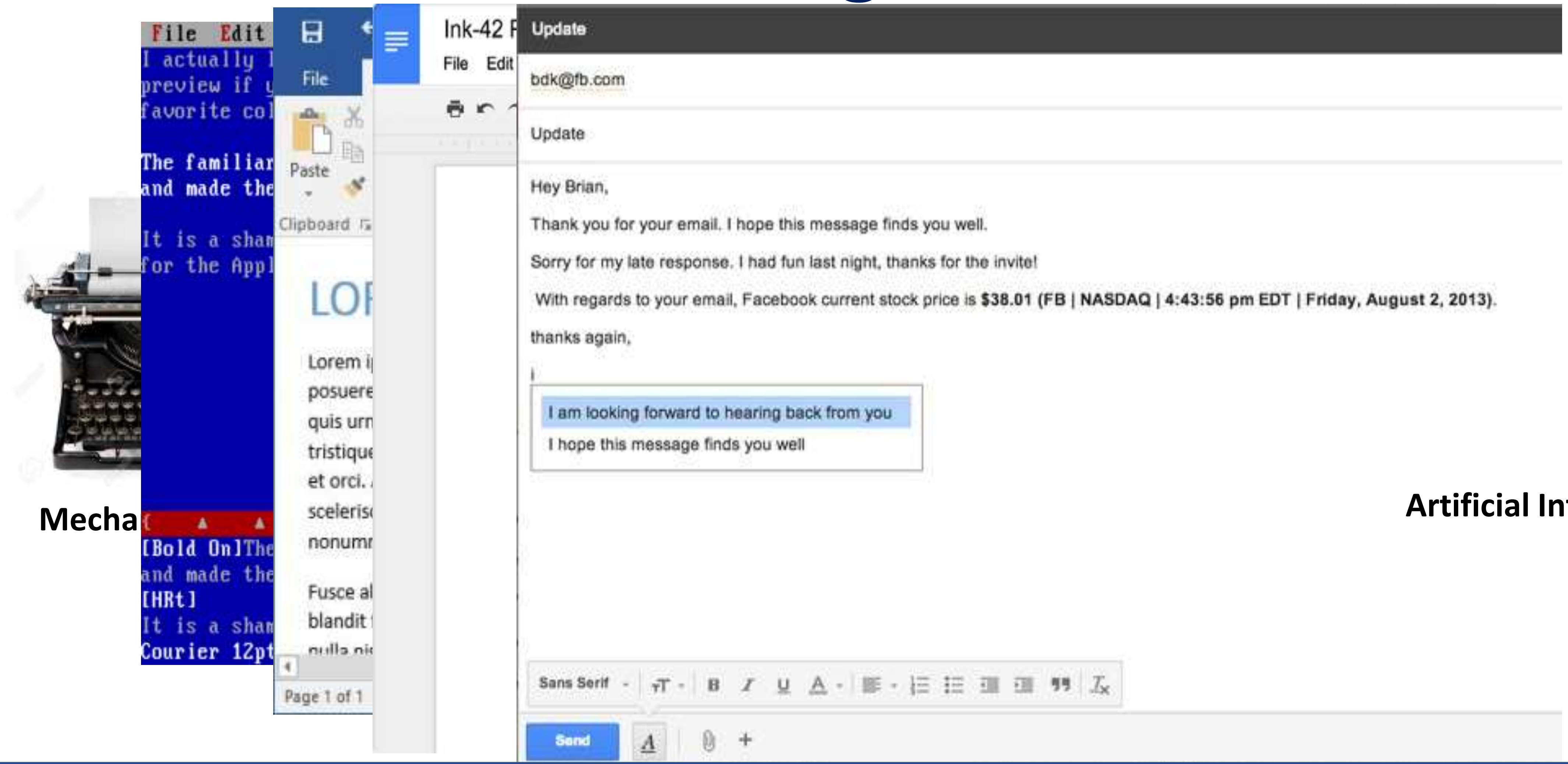
iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Why our approach?

Scalable
Works in real time
Makes unconscious, conscious

Finding the right word is a bi-objective optimization problem

Writing Platforms



bdk@fb.com

Update

Hey Brian,

Thank you for your email. I hope this message finds you well.

Sorry for my late response. I had fun last night, thanks for the invite!

With regards to your email, Facebook current stock price is **$38.01 (FB | NASDAQ | 4:43:56 pm EDT | Friday, August 2, 2013).**

thanks again,

I am looking forward to hearing back from you

I hope this message finds you well

Mecha

Artificial Intelligence Era

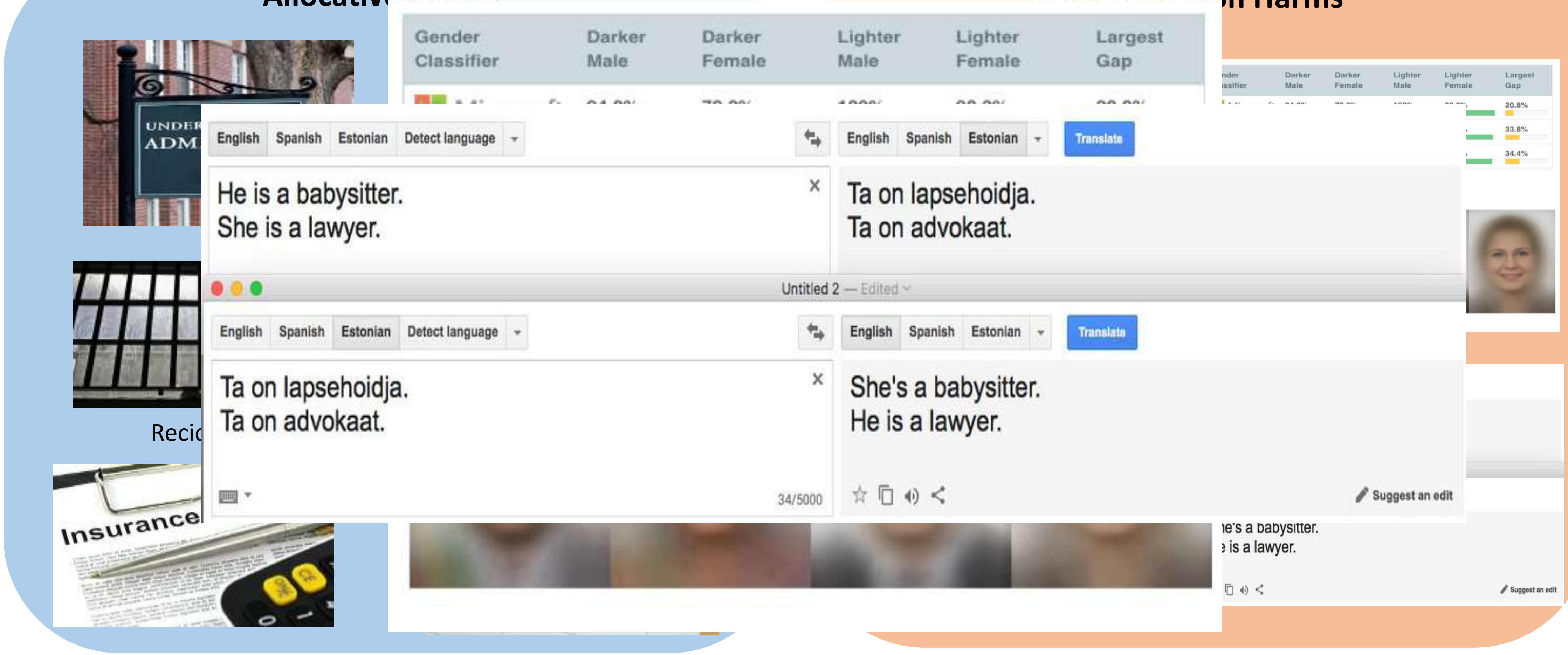
AI powered text editors is the way forward!

Source: https://textio.ai/the-dawn-of-the-augmented-writing-era-a0bb4f4376ff

How Algorithmic Bias is impacting Society?

Allocative Harms

Representation Harms



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

English → Estonian

He is a babysitter.
She is a lawyer.

Ta on lapsehoidja.
Ta on advokaat.

Estonian → English

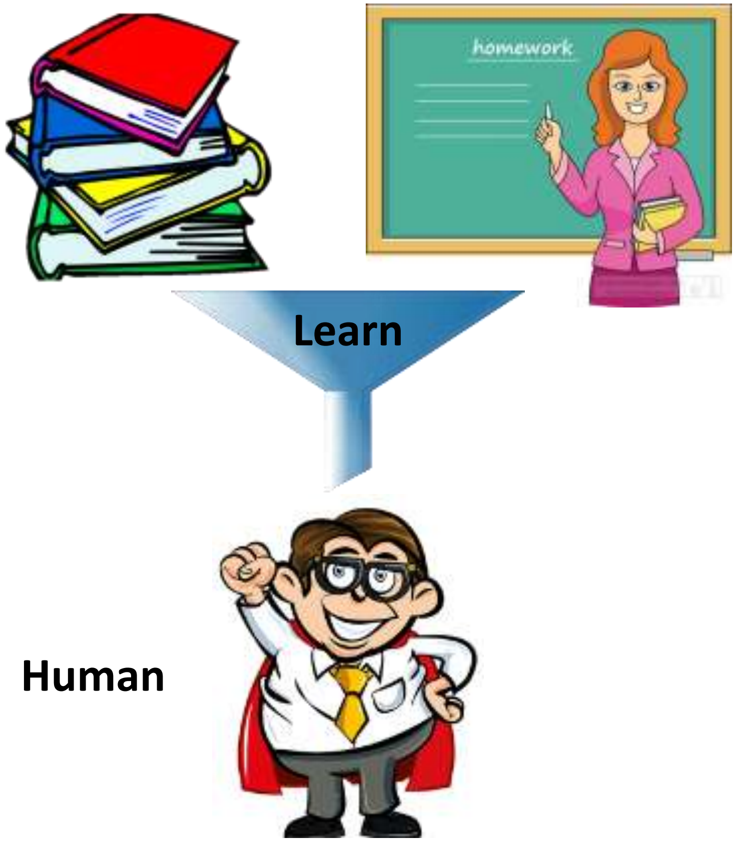
Ta on lapsehoidja.
Ta on advokaat.

She's a babysitter.
He is a lawyer.

Algorithms are trying to replicate the bias encoded in data

iACS
INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Sources of Bias

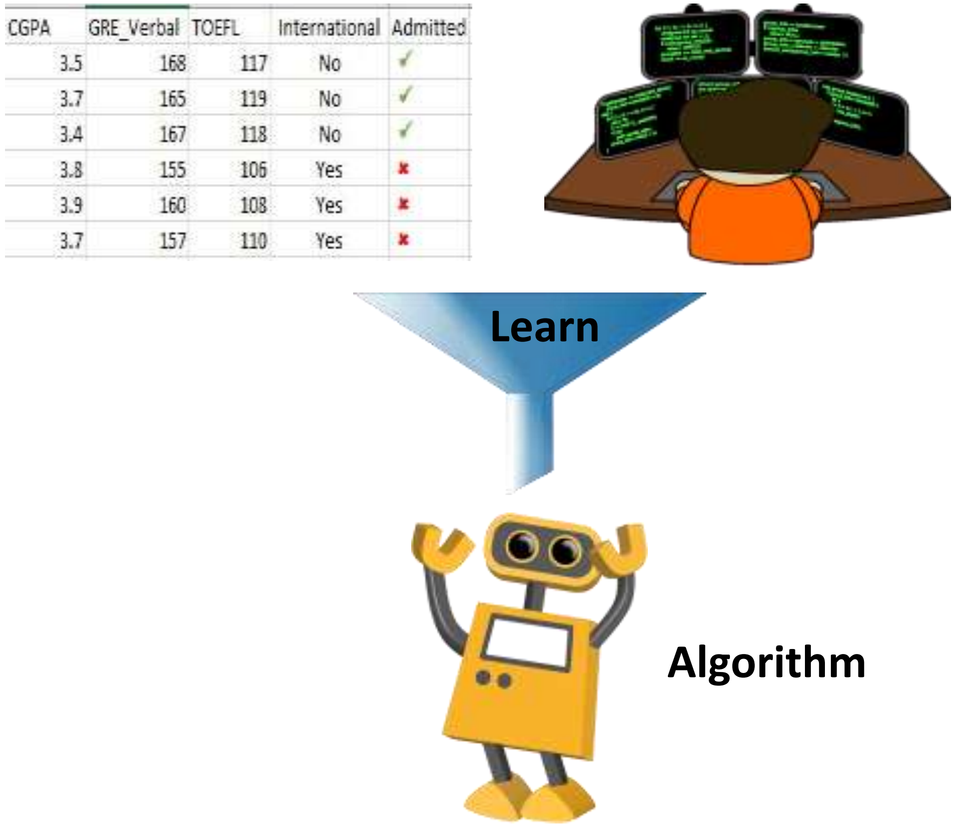


Books

Teacher

Learn

Human



Training data

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	✗

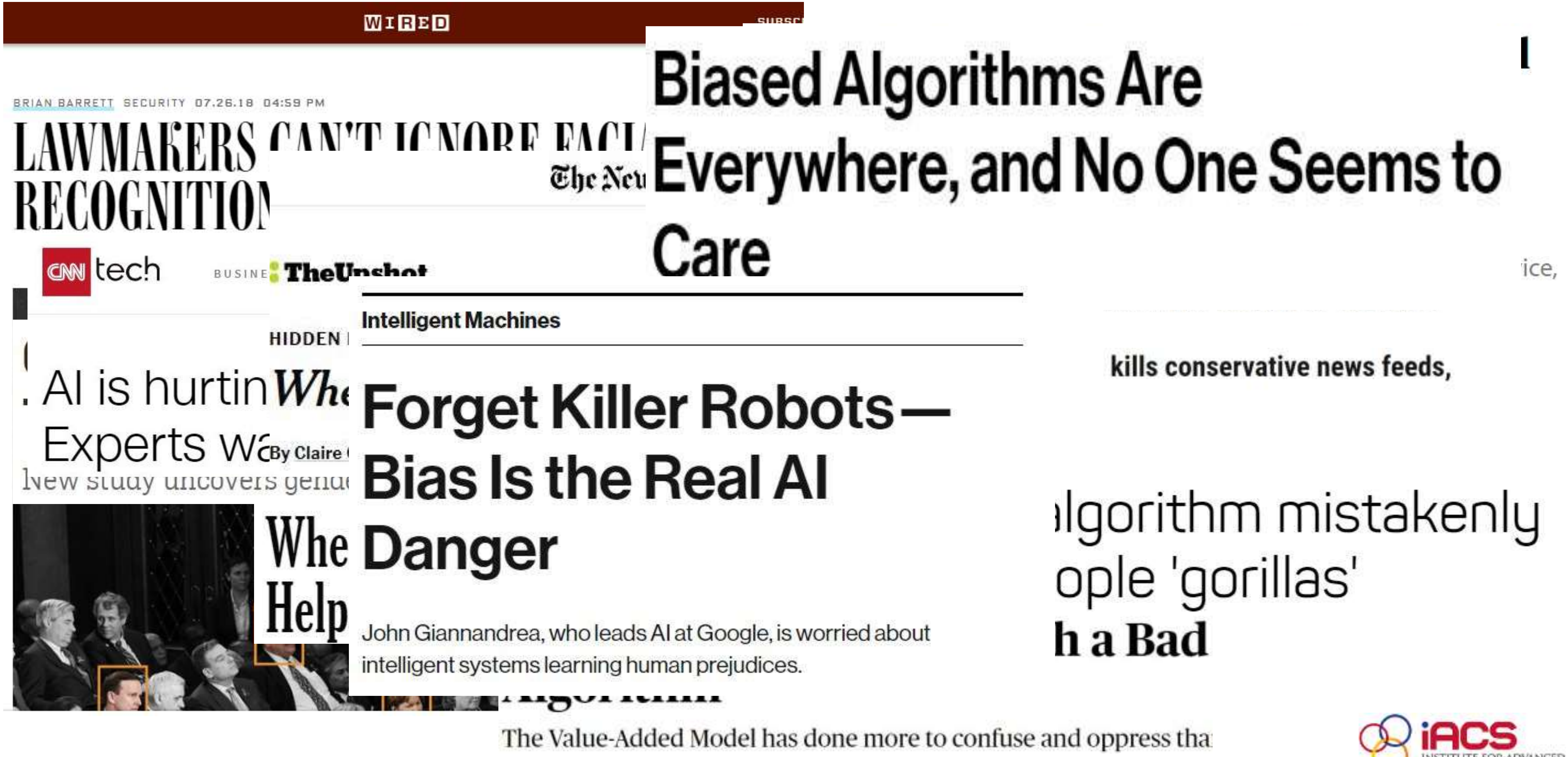
Developers

Learn

Algorithm

If training data or code developer is biased, Algorithms will be biased

In the media …



WIRED

SUBSC

BRIAN BARRETT SECURITY 07.26.18 04:59 PM

LAWMAKERS CAN'T IGNORE FACI RECOGNITIO

The New

Biased Algorithms Are Everywhere, and No One Seems to Care

ice,

CNN tech

BUSINE

TheUpshot

HIDDEN

Intelligent Machines

kills conservative news feeds,

AI is hurtin Experts wa

Wh

By Claire

New study uncovers gende

Forget Killer Robots— Bias Is the Real AI Danger

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

lgorithm mistakenly ople 'gorillas'

h a Bad

Whe Help

Algorithm

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

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Algorithms vs Humans

Human

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

Algorithm

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

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

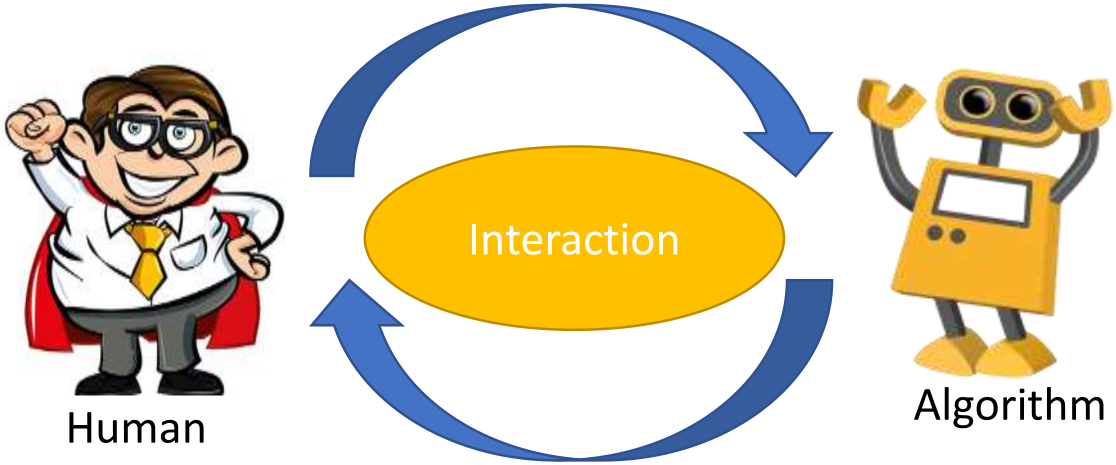
* Algorithms are often used at a much **larger scale** than human decision makers, in many cases, replicating an identical bias at scale (part of the appeal of algorithms is how cheap they are to use)

* 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

Humans and machines have their own pros & cons

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

Our approach – Human Centered AI



- AI Systems should understand humans
- AI help humans understand itself
- Computational creativity

- 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

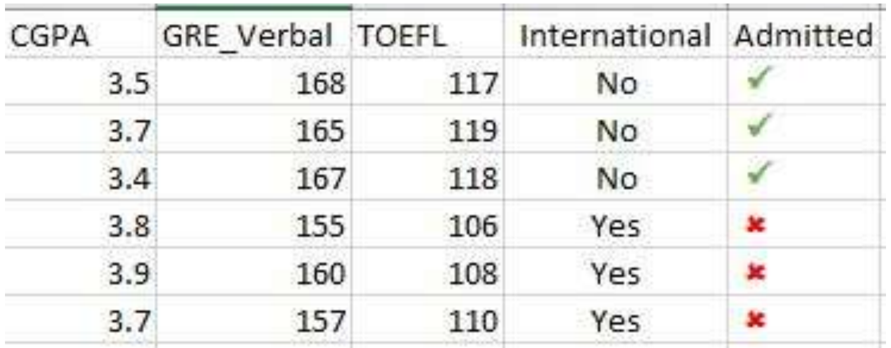
Our approach brings the bests of both worlds!

https://www.nytimes.com/2018/03/07/opinion/artificial-intelligence-human.html

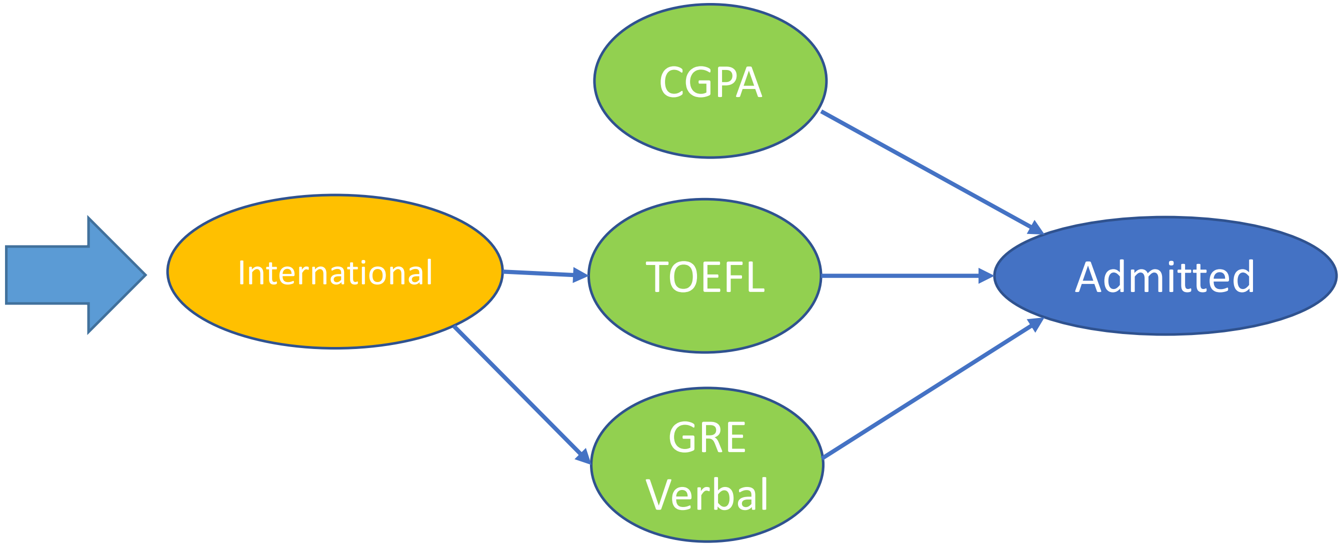
iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

Causal Networks & Debiasing

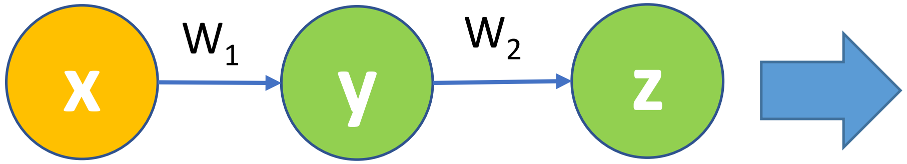
Causal Network



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



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

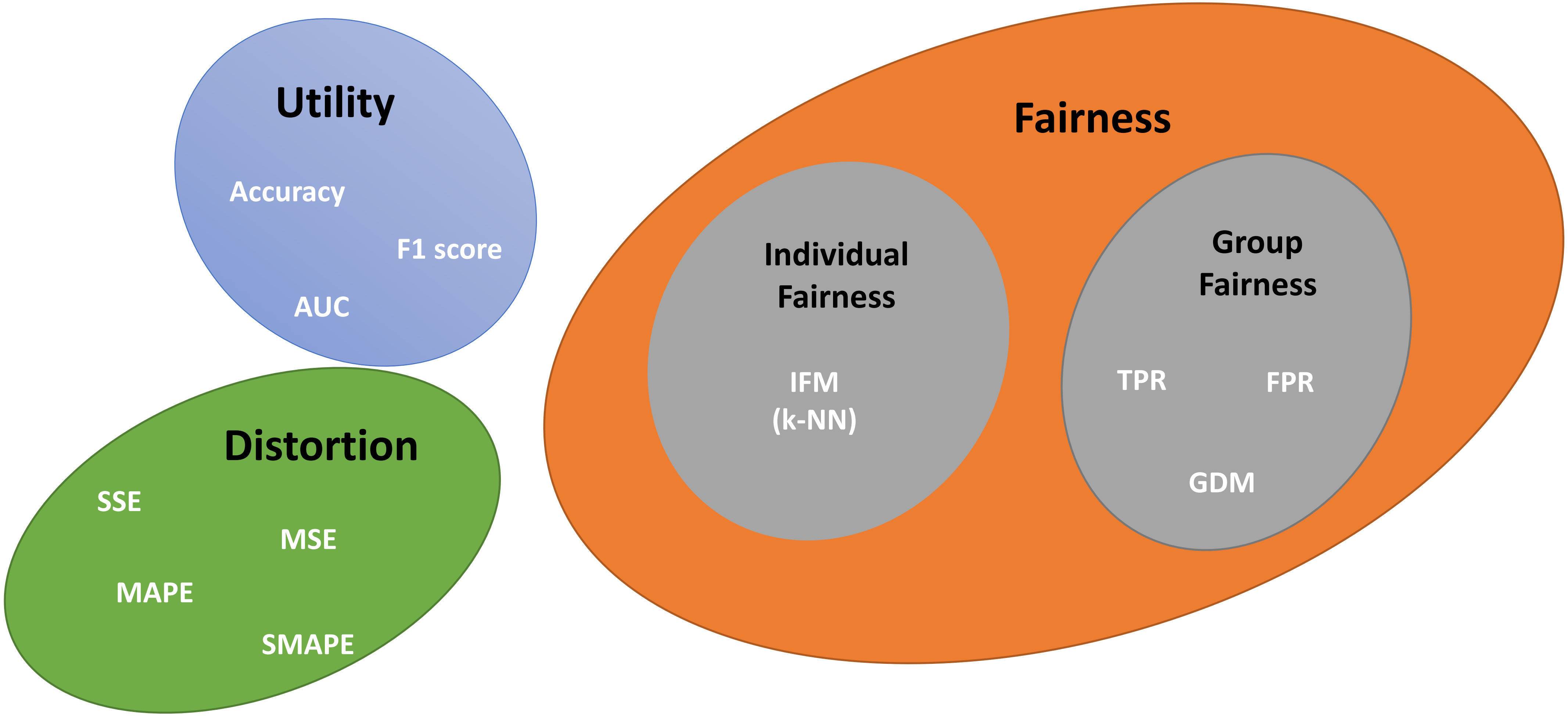
$$z_{new} = z - w_1 w_2 x$$

Partial Debiasing

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

Causal Networks help identify bias

Evaluation Objectives



Utility

Accuracy

F1 score

AUC

Distortion

SSE

MSE

MAPE

SMAPE

Fairness

Individual Fairness

IFM (k-NN)

Group Fairness

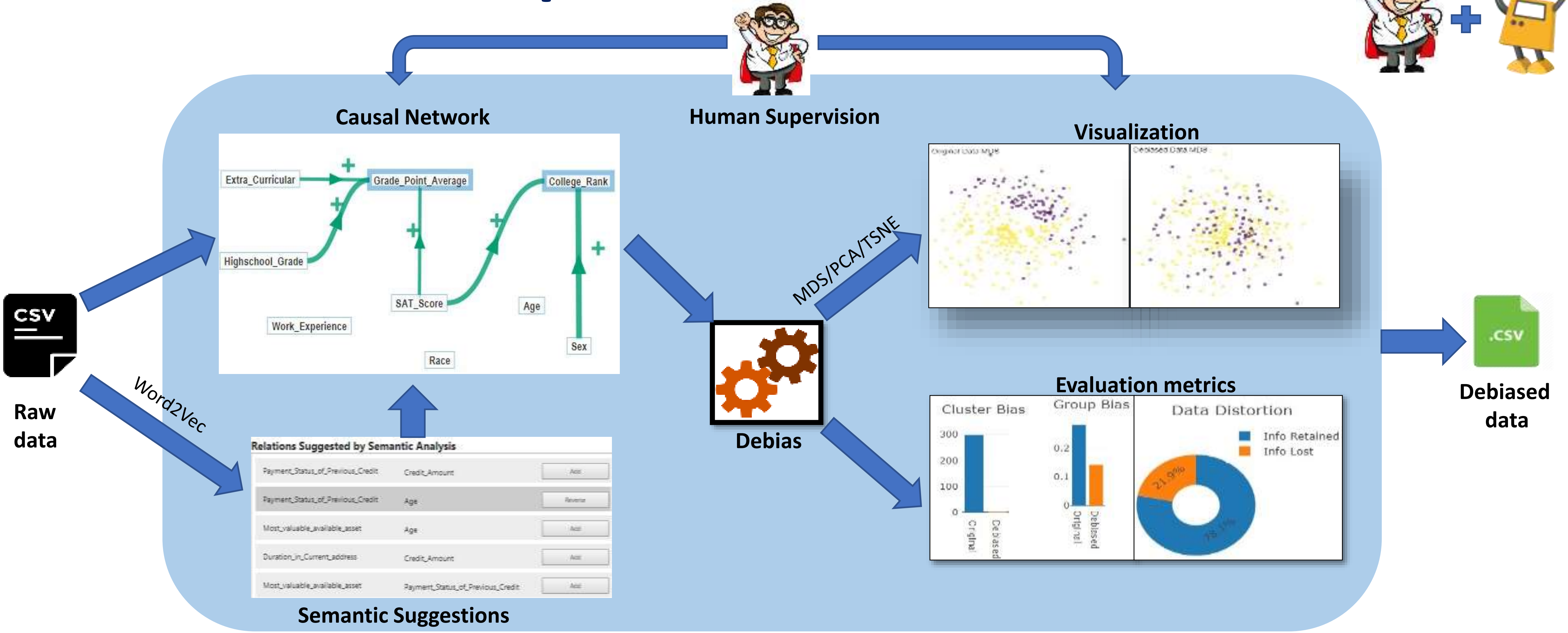
TPR

FPR

GDM

Preserve utility, maximize fairness & minimize distortion

Proposed Architecture



Relations Suggested by Semantic Analysis

Payment_Status_of_Previous_Credit	Credit_Amount	Add
Payment_Status_of_Previous_Credit	Age	Reverse
Most_valuable_available_asset	Age	Add
Duration_in_Current_address	Credit_Amount	Add
Most_valuable_available_asset	Payment_Status_of_Previous_Credit	Add

Humans can infuse domain knowledge by interacting with the causal network

Why our Approach?

Fairness

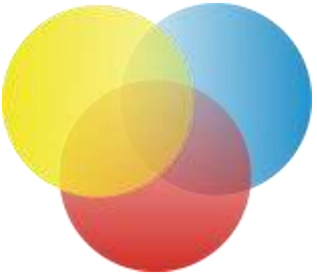
✓ Using multiple fairness definitions

Transparency

✓ Interactive visual interface boosts transparency

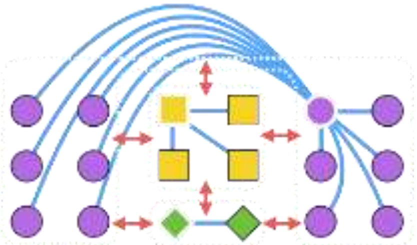
Accountability

✓ 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