VECTOR INSTITUTE

Bias in AI:
Week #4 & #5: Mitigating Bias in AI (AI Fairness)

Instructor:
Sayyed Nezhadi

Winter 2023

Why Should We Care?

- Regulations (e.g., anti-discrimination laws)
- Societal Impact:
  - “Weapons of Math Destruction” (award-winning book)
  - AI can accelerate and scale human bias
  - The network effect
- Business Impact:
  - Satisfying the demand from the customers
  - Increasing people's trust in our systems
  - Profit trade offs of fairness (choosing the right fairness is important)

Regulated Domains (USA)

- Credit (Equal credit opportunity act)
- Education (Civil rights act of 1964; Education amendments of 1972)
- Employment (Civil rights act of 1964)
- Housing (Fair housing act)
- Public accommodation (Civil rights act of 1964)
- Complex web of laws that regulates the governments

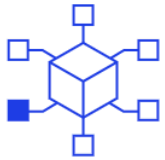
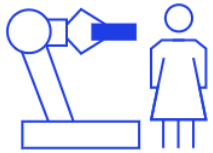
Protected Classes (Canadian Human Rights Act)

Discrimination is an action or a decision that treats a person or a group badly for the following reasons:

- Race
- National or ethnic origin
- Colour
- Religion
- Age
- Sex
- Sexual orientation
- Gender identity or expression
- Marital status
- Family status
- Disability
- Genetic characteristics
- A conviction for which a pardon has been granted or a record suspended

Action Plan

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider

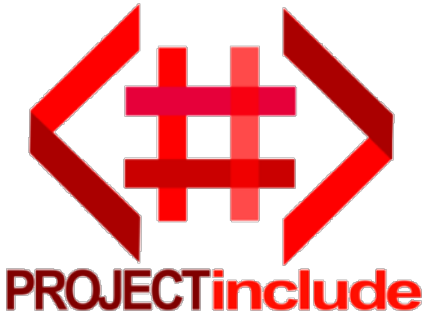
1	2	3	4	5	6
					
Be aware of contexts in which AI can help correct for bias and those in which there is high risk for AI to exacerbate bias	Establish processes and practices to test for and mitigate bias in AI systems	Engage in fact-based conversations about potential biases in human decisions	Fully explore how humans and machines can best work together	Invest more in bias research, make more data available for research (while respecting privacy), and adopt a multidisciplinary approach	Invest more in diversifying the AI field itself

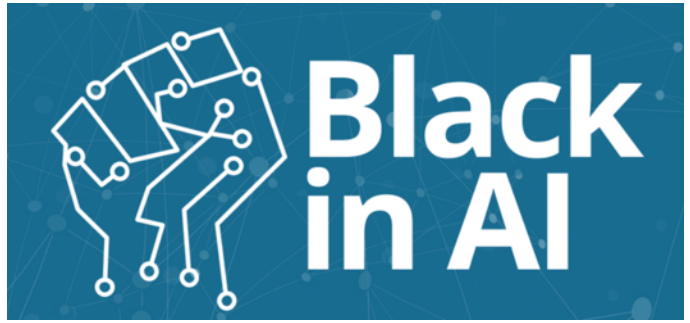
McKinsey & Company

Team's Diversity

- To combat bias in AI, companies need more diverse AI talent.
- AI and data science are not just about number crunching. It requires people with different perspectives who can ask the right questions of the data and design more inclusive systems.

Inclusion & Social Impact





Muslims in ML

Trans-Inclusive AI

Challenges: Fairness Definition

- It is a complex notion debated among philosophers and the public.
- There are many definitions of **fairness**.
- Many of the fairness criterions are mutually exclusive (Impossibility Theorem)

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Source: Fairness and Machine Learning by Solon Barocas, Moritz Hardt, Arvind Narayanan

Challenges: Governance

- Responsibility for Algorithmic Impact Assessment
- Responsibility to Identify beneficiaries of the model/application.
- Accountability in considering and addressing the risk of unfair outcomes (needs to come from the top).
- Fairness, transparency, and explainability are only meaningful when operationalized into the enterprise risk management processes.
- How does fairness align with the business strategy?
- What processes could be modified to improve the outputs?
- What controls need to be in place to track performance and pinpoint problems?

Canada's Algorithmic Impact Assessment (AIA)

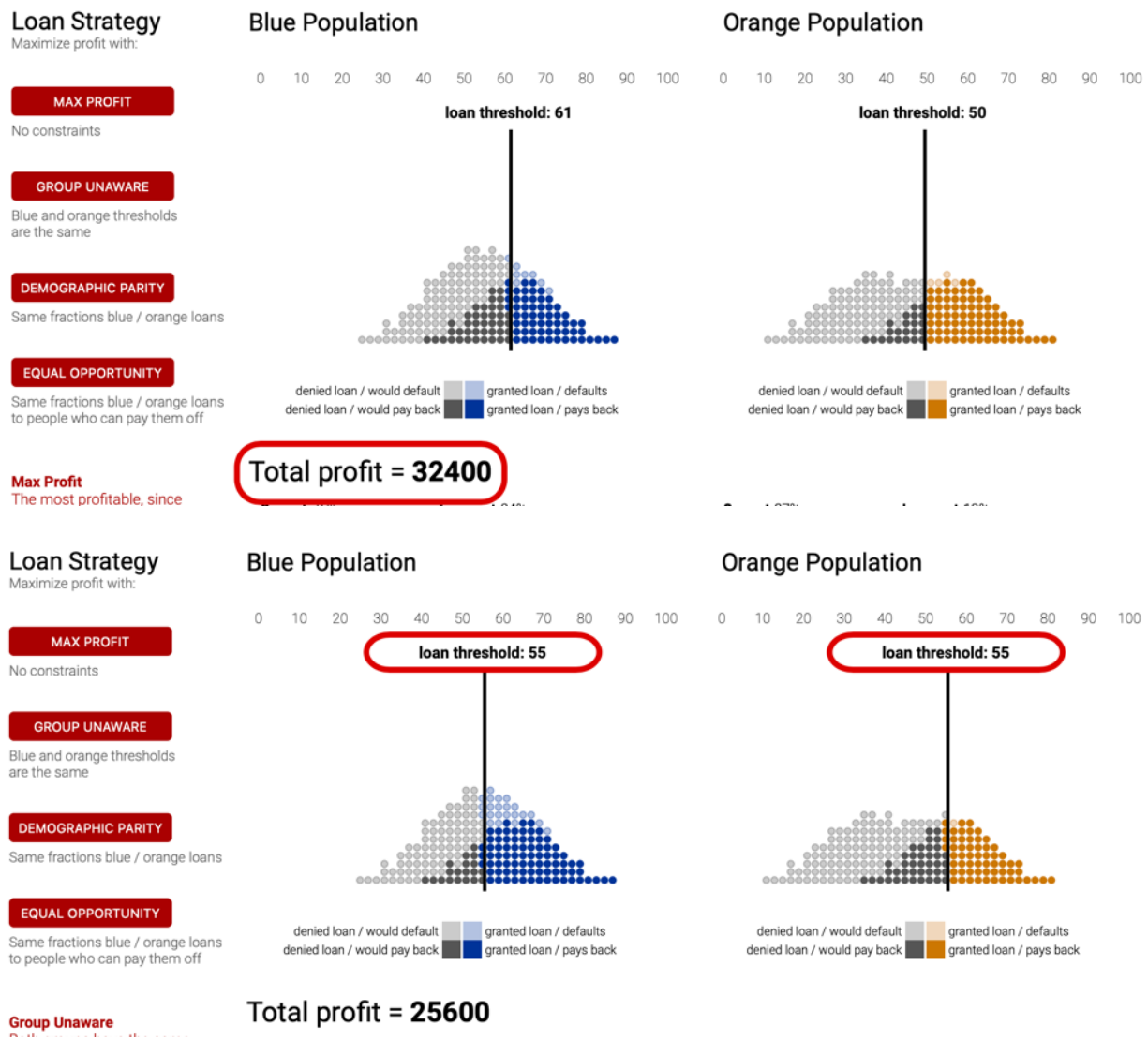
The AIA will be looking to test five areas of interest:

- The impact on individuals, businesses, and communities ("socioeconomic impact")
- The impact on government operations
- The complexity of the system
- Data management practices
- Procedural fairness considerations

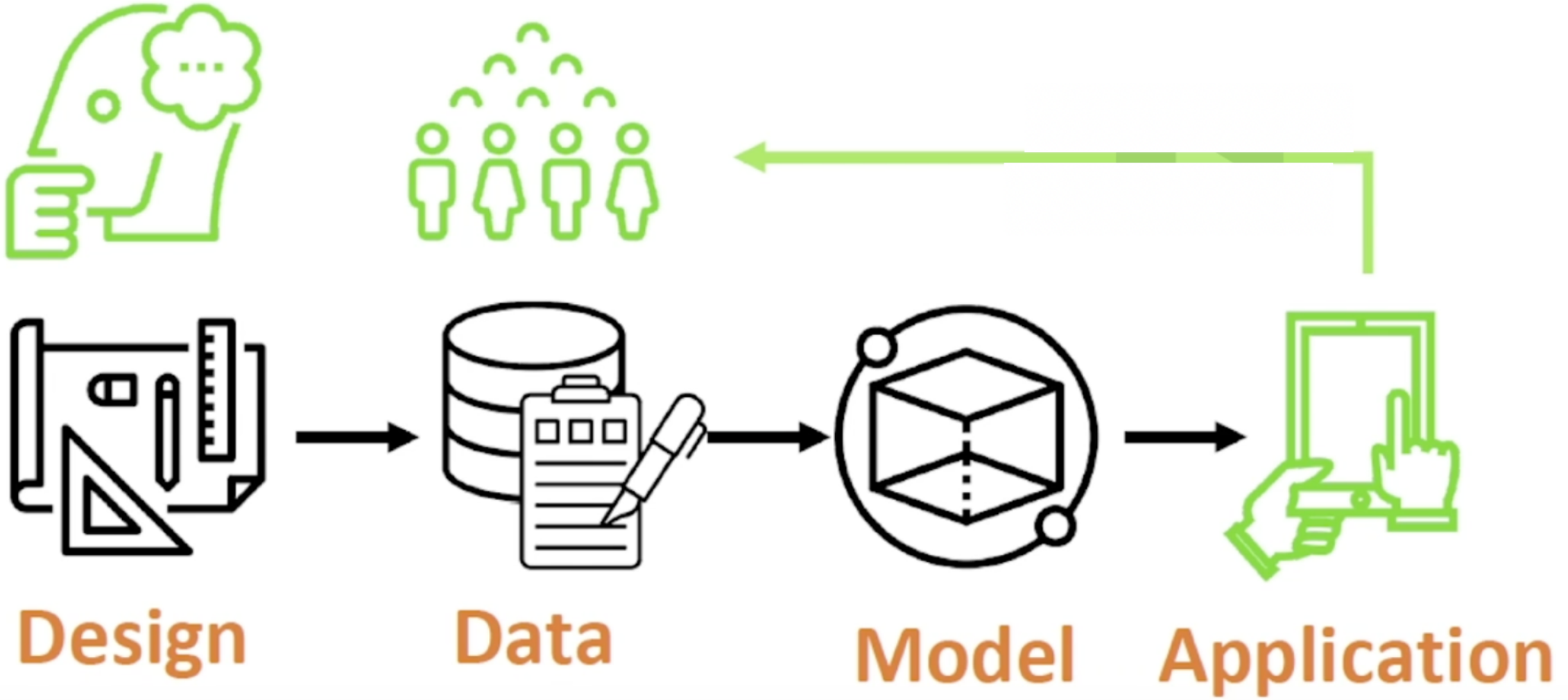
Credit: Noel Corriveau

Challenges: Profit Trade-off

- There is a tension between profits and fairness.
- Firms are not naturally incentivized to ensure fair machine learning algorithms.
- A hypothetical example: one fairness definitions results in profits 21% lower than the theoretical maximum (https://research.google.com/bigpicture/attacking-discrimination-in-ml/)

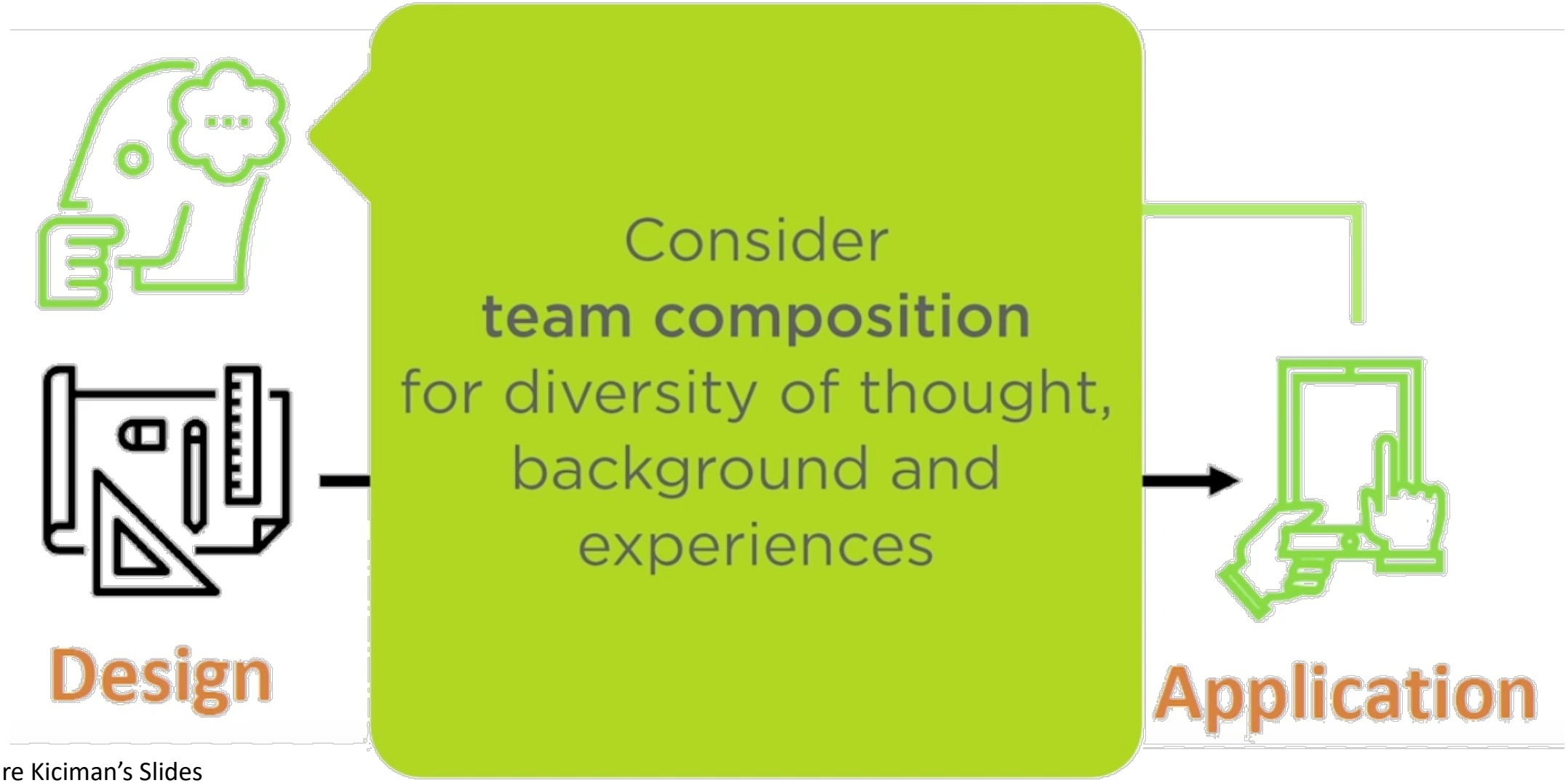


End-to-End Bias Avoidance/Mitigation



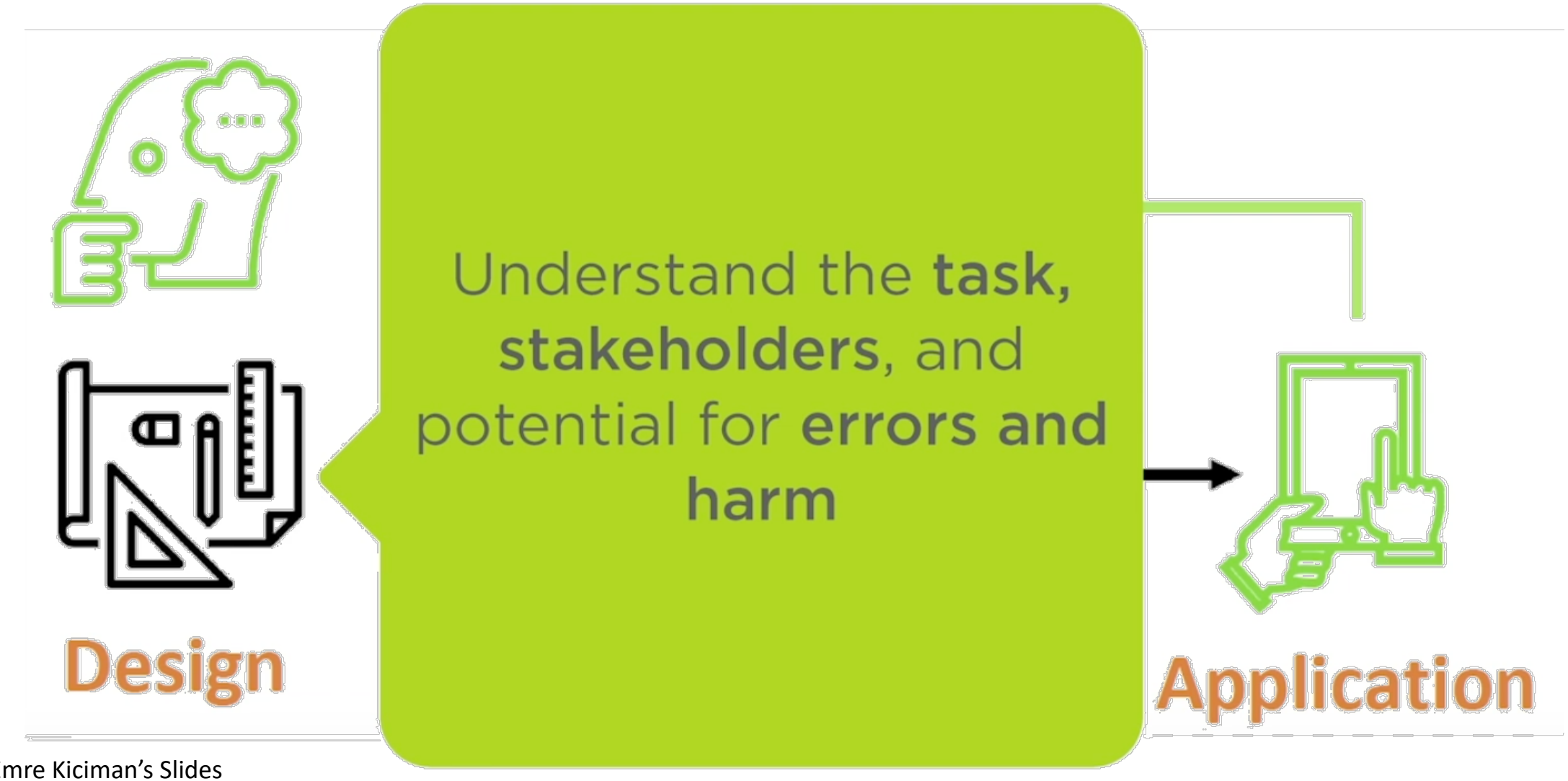
Credit: Emre Kiciman's Slides

End-to-End Bias Avoidance/Mitigation



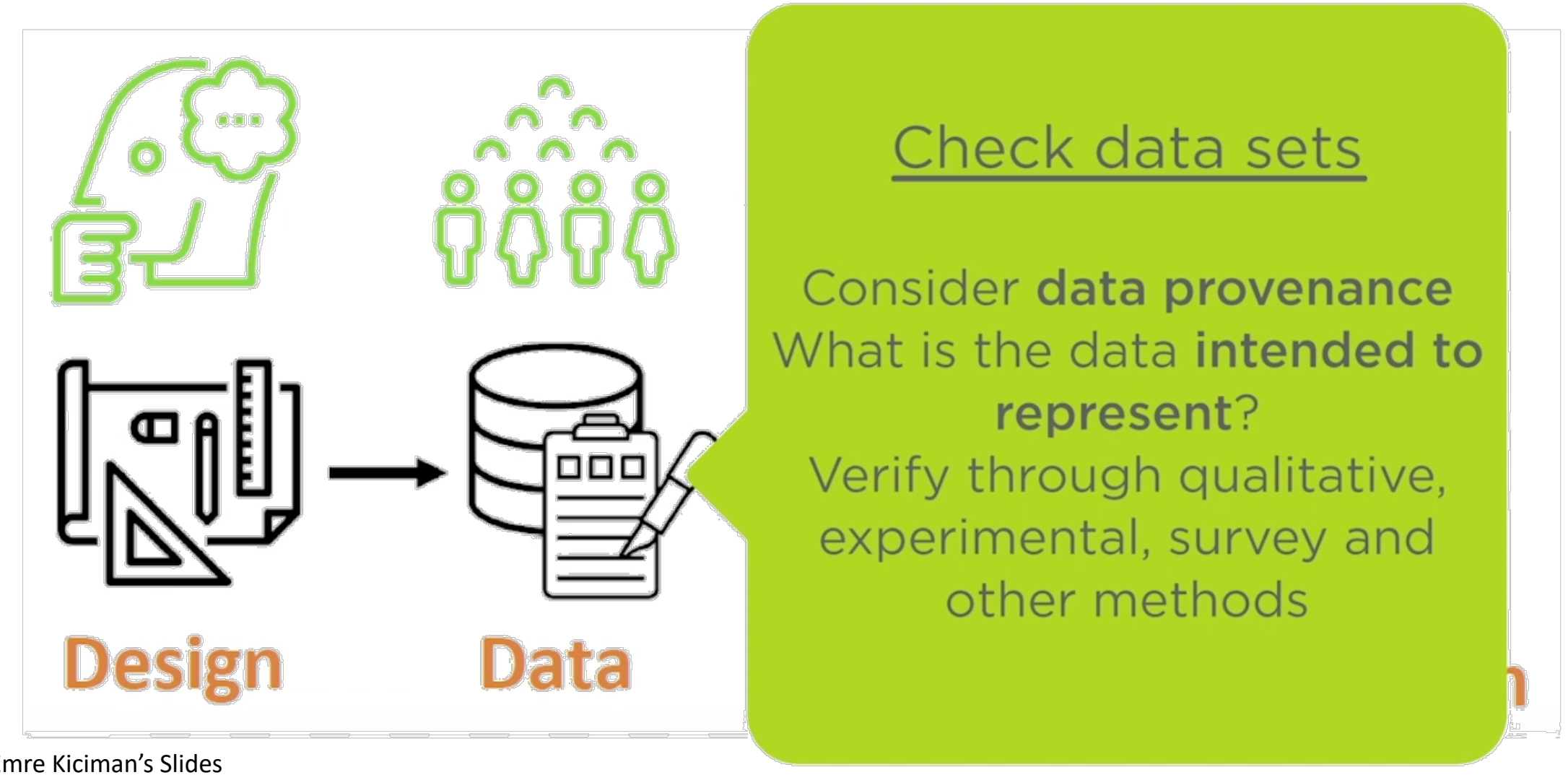
Credit: Emre Kiciman's Slides

End-to-End Bias Avoidance/Mitigation



Credit: Emre Kiciman's Slides

End-to-End Bias Avoidance/Mitigation



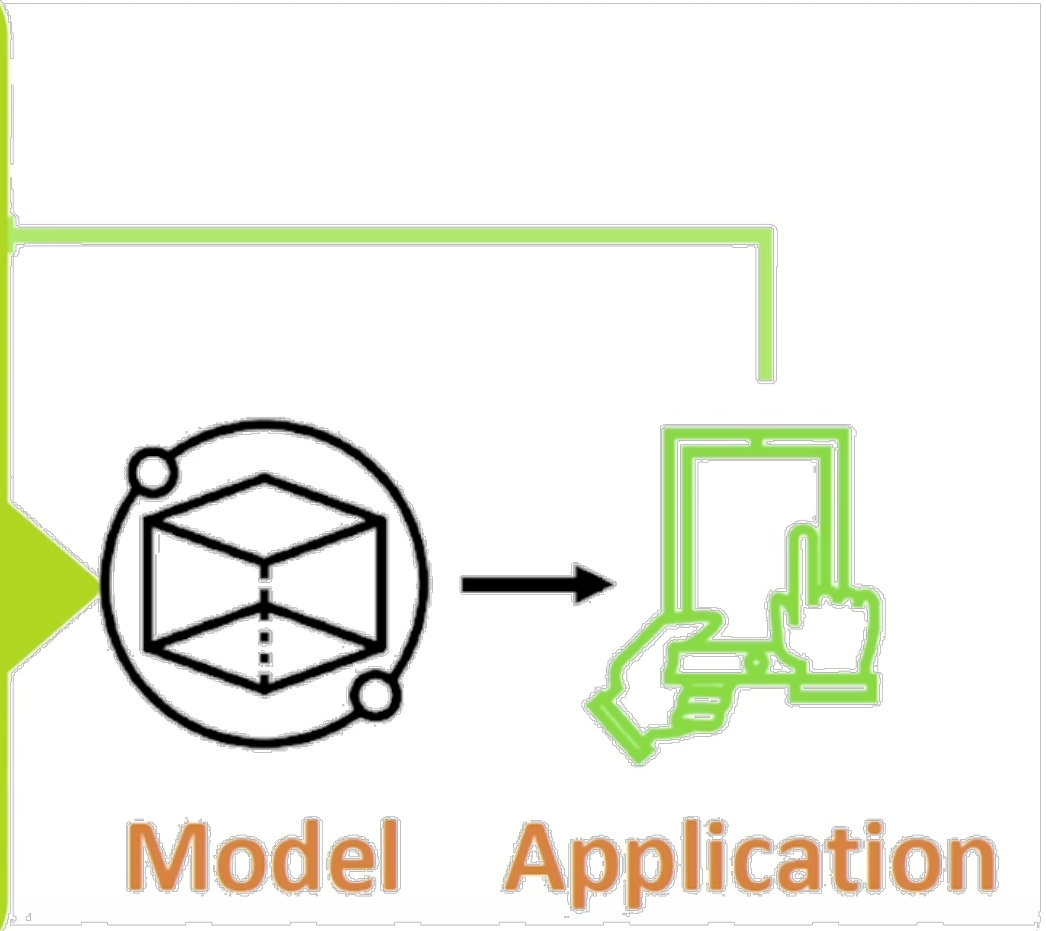
Credit: Emre Kiciman's Slides

End-to-End Bias Avoidance/Mitigation

Check models
and validate results

Why is the model making decision?
What **mechanisms** would
explain results?
Is supporting evidence consistent?

Twyman's law: The more unusual
the result, more likely it's an error



Credit: Emre Kiciman's Slides

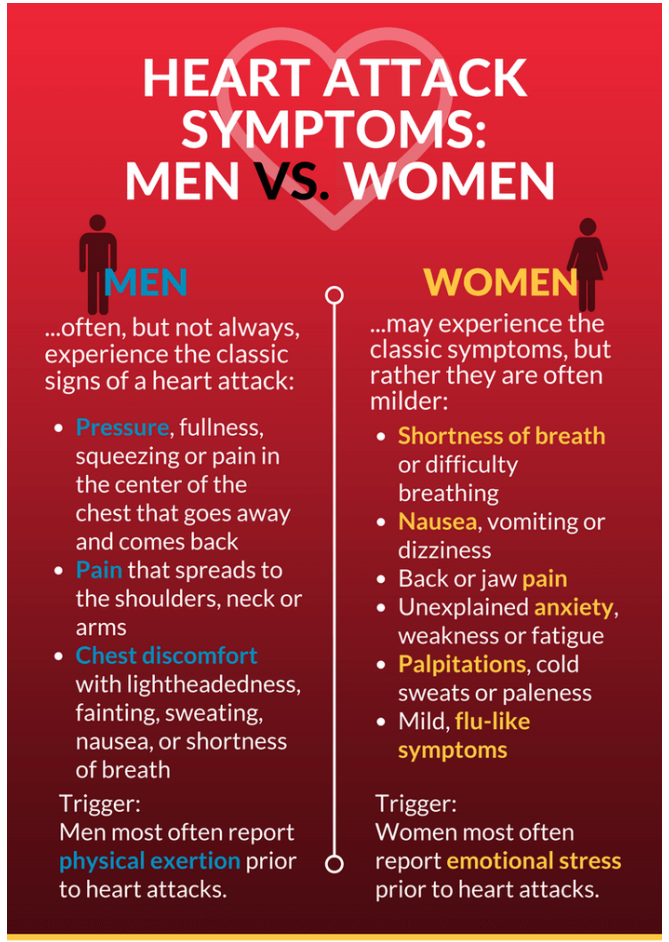
End-to-End Bias Avoidance/Mitigation

Credit: Emre Kiciman's Slides

Bias Should be Mitigated based on the Task

Gender in loan application can be discriminatory

Gender in health application can be necessary

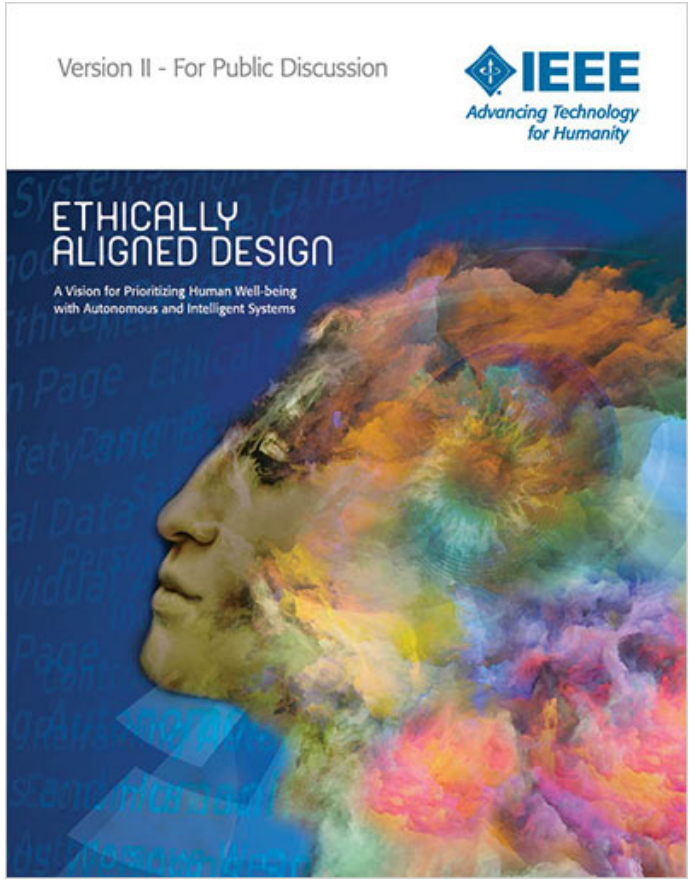


Ethically Aligned Design (IEEE)

II. General Principles

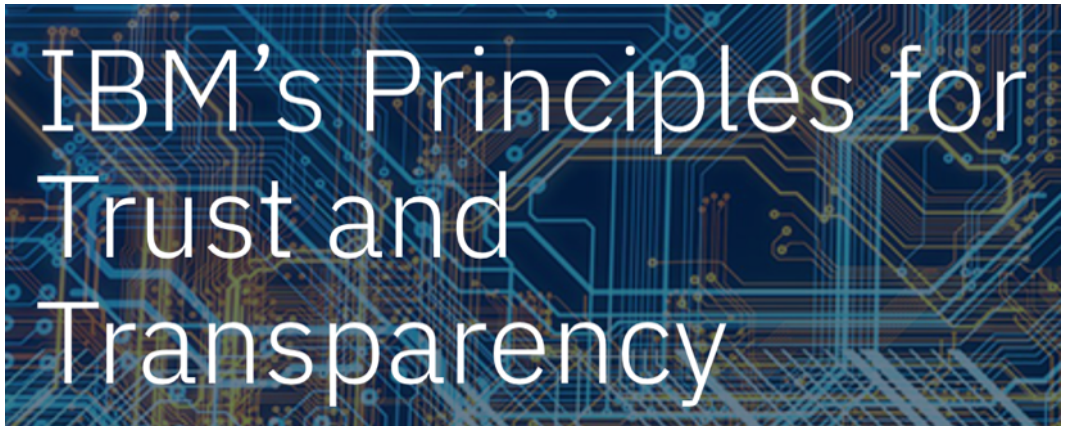
The ethical and values-based design, development, and implementation of autonomous and intelligent systems should be guided by the following General Principles:

1. **Human Rights**
   A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being**
   A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency**
   A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness**
   A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency**
   The basis of a particular A/IS decision should always be discoverable.
6. **Accountability**
   A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse**
   A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence**
   A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.



Corporate Design Standards

Artificial Intelligence at Google:
Our Principles



Guidelines for human-AI interaction design

Published February 1, 2019 (Microsoft)

Documenting Datasets

- One way is to minimize the harm of data bias is to communicate the bias issues with datasets through proper documentation.
- Some attempts are done by the research community to define a standard:
  - Datasheets for Datasets (T. Gebru et al. 2018): describes its operating characteristics, test results, recommended uses, and other information.
  - Data Statement for NLP (E. Bender et al. 2018): a design solution and professional practice for natural language processing technologists, in both research and development.
  - The Montreal Data Licenses (M. Benjamin et al. 2019): a taxonomy for the licensing of data in the fields of artificial intelligence and machine learning. It includes a new family of data license language and a web-based tool to generate license language.

Data Privacy

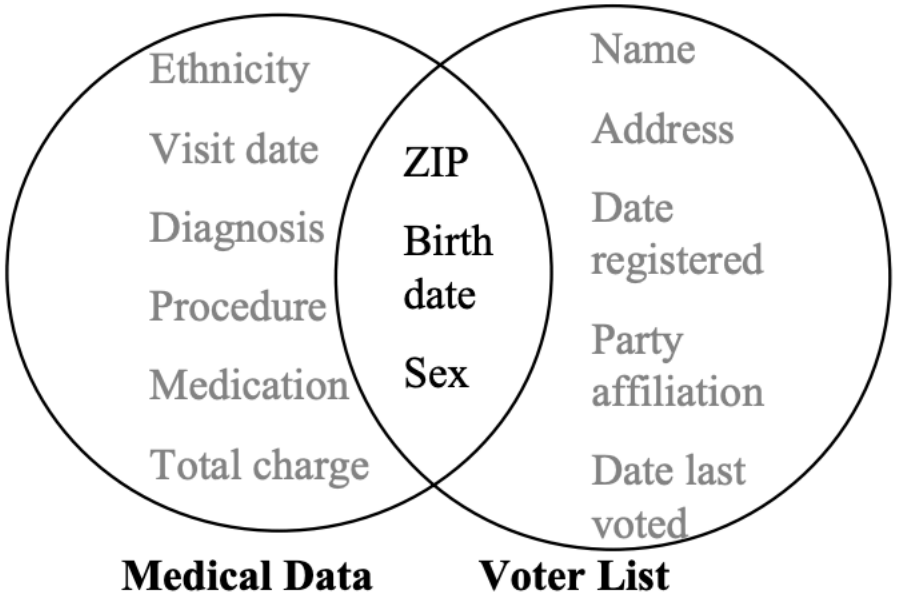
- PIPEDA: personal information includes any factual or subjective information, recorded or not, about an identifiable individual. This includes information in any form, such as:
  - age, name, ID numbers, income, ethnic origin, or blood type;
  - opinions, evaluations, comments, social status, or disciplinary actions; and
  - employee files, credit records, loan records, medical records, existence of a dispute between a consumer and a merchant, intentions (for example, to acquire goods or services, or change jobs).
- GDPR: any piece of information that relates to an identifiable person.

Office of the Privacy Commissioner of Canada

Anonymous Data is Not Always Private

- Anonymous data may not be private because of:
  - High correlation between personal data and other data
  - Can be linked to other datasets

Image credit: Nicholas Vermeys

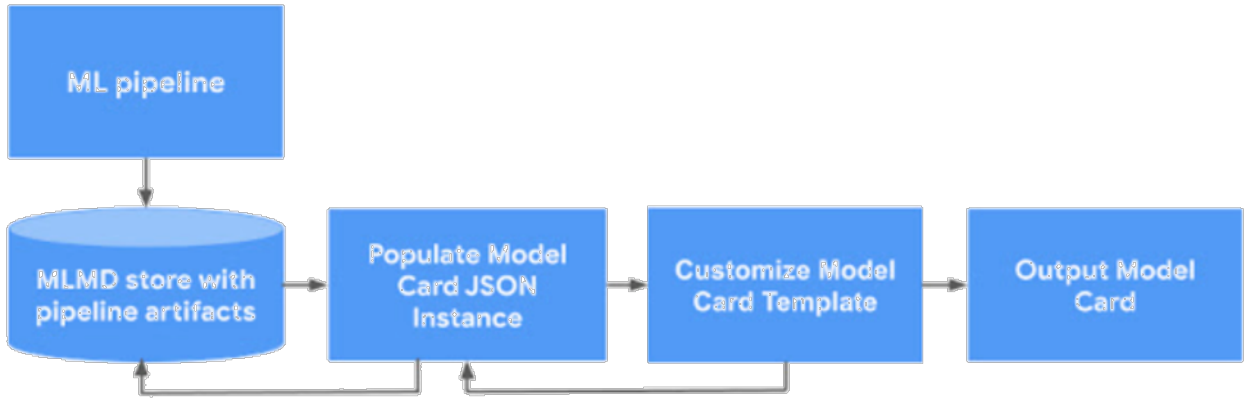


Linking to re-identify data

Source: L. Sweeney 2002

Model Transparency

- Machine learning (ML) model transparency is important across a wide variety of domains that impact peoples' lives, from healthcare to personal finance to employment. Developers can use the information to decide whether a model is appropriate for their use case.
- Google Model Cards: provides a structured framework for reporting on ML model provenance, usage, and ethics-informed evaluation and give a detailed overview of a model's suggested uses and limitations that can benefit developers, regulators, and downstream users alike.



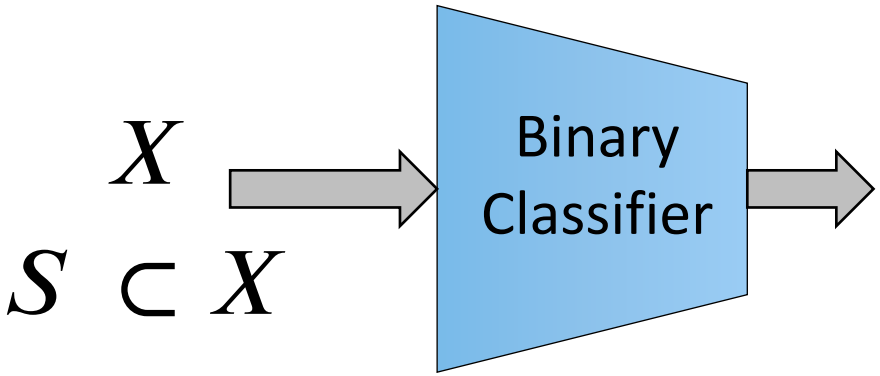
Explainable AI

- The next step to Model Transparency is Explainability.
- We learned that for an AI system the results are not everything (the concept of "black box" in machine learning doesn't work). We need to understand how the model work and if it is fair.
- Explainable AI (XAI) refers to methods and techniques that human can understand how an AI system has come to a decision. It allows us to:
  - Design better models that can generalize better
  - Provide explanation when regulations require that
- XAI is an active area of research and a wide range of methods are suggested (including using AI explain another AI system)

Explainability for Fair Machine Learning

- The concept of "black-box fairness" is not enough to create "trust" in our AI systems.
- In addition, maximum fairness is not achievable without understanding how our models work. We can use model explainability to gain insights into model fairness.
- However, H. Dimanov et al. (2020) show that existing XAI methods are poorly suited for understanding fairness.
- T. Begley et al. (2020) introduced a new approach to explaining fairness in machine learning, based on a value paradigm. The fairness values are able to attribute unfairness in a model's predictions to individual features.
- This is an active and very recent area of research

Fairness Criteria



Demographic Parity (Independence)

- Also referred to as Statistical Parity, Group Fairness, and Disparate impact.
- The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group (e.g., female)

$$P\left(\hat{Y} = 1 \,\middle|\, S = a\right) = P\left(\hat{Y} = 1 \,\middle|\, S = b\right)$$

Example: $\boldsymbol{a}$ = Male, $\boldsymbol{b}$ = Female, c = …

Demographic Parity (Independence)

- Also referred to as Statistical Parity, Group Fairness, and Disparate impact.
- The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group (e.g., female)

$$P(\hat{Y} = 1 \mid S = a) = P(\hat{Y} = 1 \mid S = b)$$

Example: $\boldsymbol{a}$ = Male, $\boldsymbol{b}$ = Female, c = …

Do you see any issue with this criterion?

Limitations of Independence

- It permits the model to select unqualified members of a group to satisfy the criterion (This creates biased perception) .
- Imagine a company that filters the job applicants in group “a" diligently at some rate. But it hires the candidates In group “b” carelessly to match the same rate as group “a”.
- Another problem is when we have much more training data in group “a” than group “b” leading to lower error rates of a learned classifier within that group. This means the classifier does a better job of selection in one group.
- Example: if 95% of training data is from group “a” (with the accuracy of 90% for that group), even randomly selecting from group “b” results in overall accuracy of 85.5%.

Equal Opportunity

- To fix the limitation of independence criterion, we can change it give equal opportunity to all qualified candidates from all groups. That meaning equal True Positive rate across the groups.

$$P\left(\hat{Y}=1 \mid Y=1, S=a\right)=P\left(\hat{Y}=1 \mid Y=1, S=b\right)$$

- Reminder:

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Equalized Odds (Separation)

- A stronger criterion is to equalize both True Positive and False Positive rates across the groups.

$$P\left(\hat{Y}=1 \mid Y=1, S=a\right)=P\left(\hat{Y}=1 \mid Y=1, S=b\right)$$

$$P\left(\hat{Y}=1 \mid Y=0, S=a\right)=P\left(\hat{Y}=1 \mid Y=0, S=b\right)$$

Conditional Statistical Parity

- Sometimes you may want to give equal opportunity based on a different set of legitimate factors(L) than the target labels. That is called Conditional Statistical Parity.

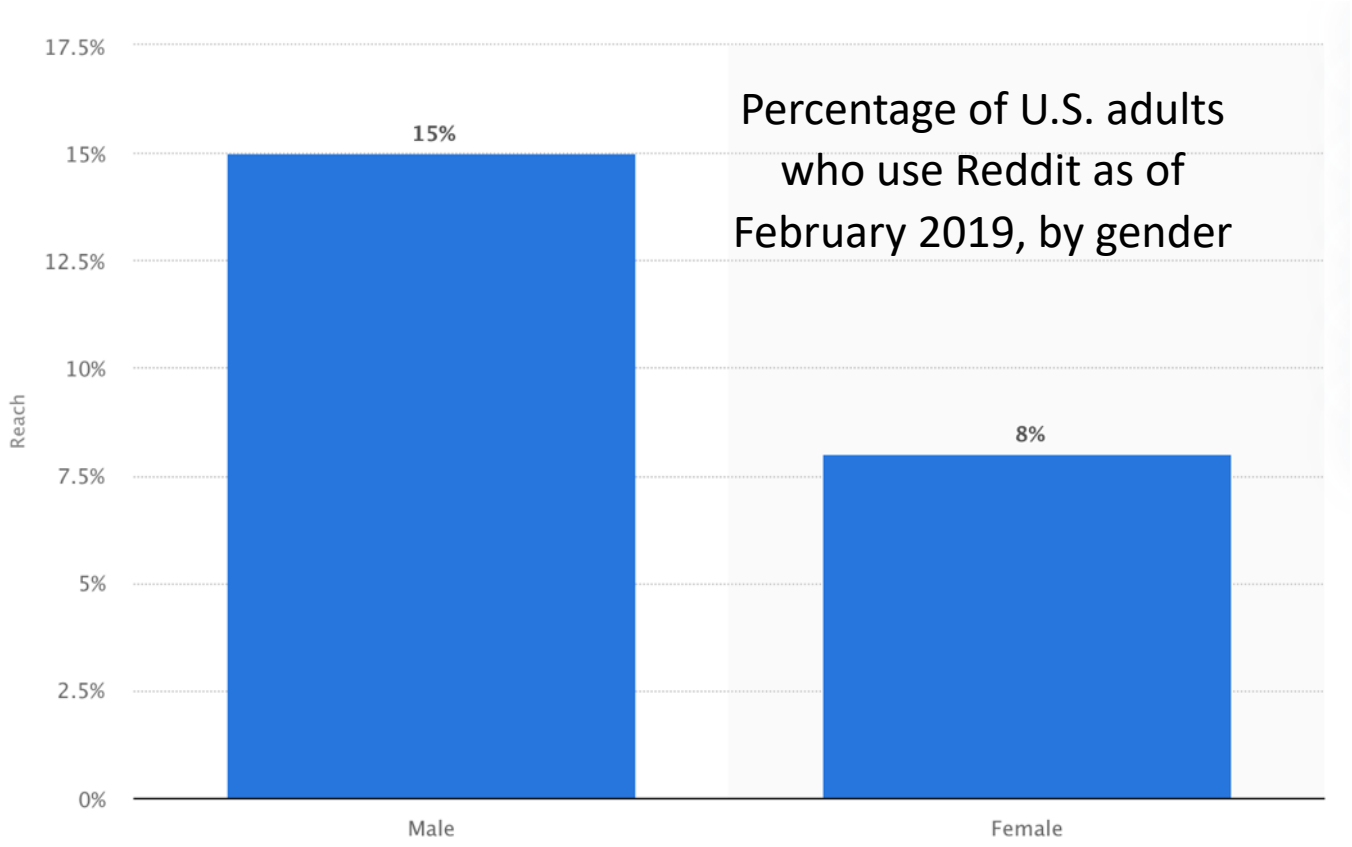
$$P\left(\hat{Y}=1 \;\middle|\; L=1, S=a\right) = P\left(\hat{Y}=1 \;\middle|\; L=1, S=b\right)$$

- For example, among defendants who have the same number of prior convictions, black and white defendants are detained at equal rates (A. Davis et al., 2017)

Likelihood vs. Real Outcome

- Important Note: Fairness criteria talk about equal likelihood (probability) not equal number of selected members from all groups.

- Example: Even if we create an ad that targets all Reddit users equally target, the expected number of outcomes among the men will be almost twice of women (assuming all of them are qualified)

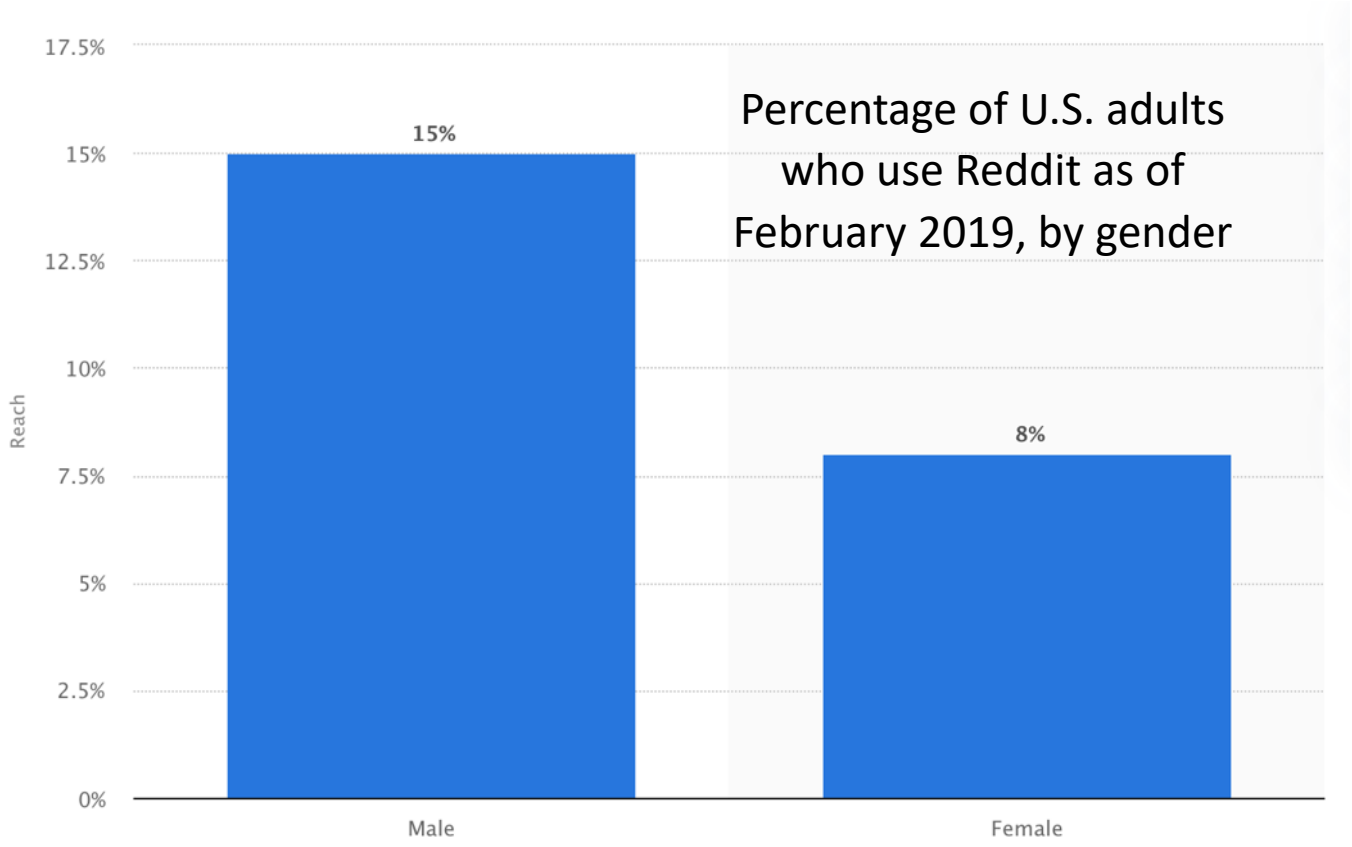


Likelihood vs. Real Outcome

- Important Note: Fairness criteria talk about equal likelihood (probability) not equal number of selected members from all groups.

- Example: Even if we create an ad that targets all Reddit users equally target, the expected number of outcomes among the men will be almost twice of women (assuming all of them are qualified)

Is that fair?

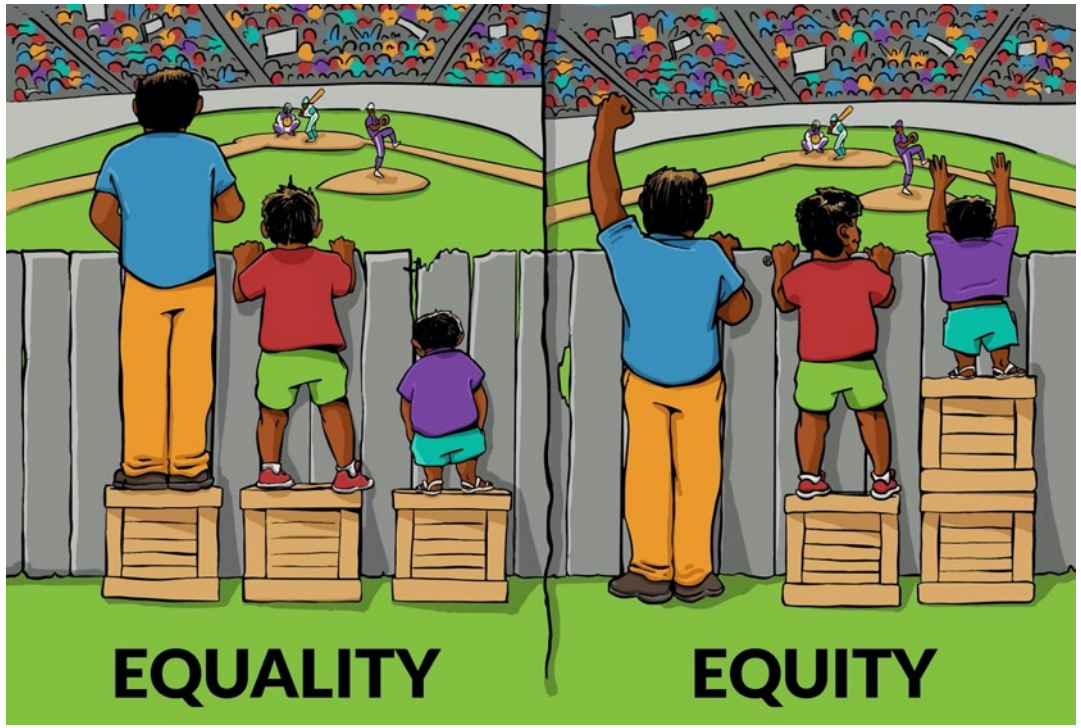


Equality vs. Equity

- Most of research is focused on Equality not Equity. Equity-based fairness remains an important future direction.

- Equality: giving the same opportunities and attention to each group or individual.
- Equity: empowering each individual or group, by giving the resources, to succeed.

Class Discussion



Source: Google Image

Predictive Parity (Sufficiency)

- If the sensitive features S are statistically independent to the target value Y given the prediction $\hat{Y}$.

$$P(Y = 1 \mid \hat{Y} = 1, S = a) = P(Y = 1 \mid \hat{Y} = 1, S = b)$$

$$P(Y = 1 \mid \hat{Y} = 0, S = a) = P(Y = 1 \mid \hat{Y} = 0, S = b)$$

Impossibility Theorem

The Impossibility Theorem states that no more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time for a well calibrated classifier and a sensitive attribute capable of introducing machine bias.

Individual Fairness

- So far, we discussed only fairness criteria toward protected groups. But group fairness doesn't guarantee individual fairness.
- Individual Fairness Through Awareness: giving similar predictions to similar people.
  - Need to define similarity (inverse of distance)
  - Sensitive features can't be used for similarity calculation
- Individual Fairness Through Unawareness: removing all sensitive features from the decision-making process (e.g., training data)

Individual Fairness

- So far, we discussed only fairness criteria toward protected groups. But group fairness doesn't guarantee individual fairness.
- Individual Fairness Through Awareness: giving similar predictions to similar people.
  - Need to define similarity (inverse of distance)
  - Sensitive features can't be used for similarity calculation
- Individual Fairness Through Unawareness: removing all sensitive features from the decision-making process (e.g., training data)

Do you think Unawareness is a good idea?

Limitations of Unawareness

1. The algorithm implicitly learn to predict the sensitive feature from other features (e.g., race from zip code).

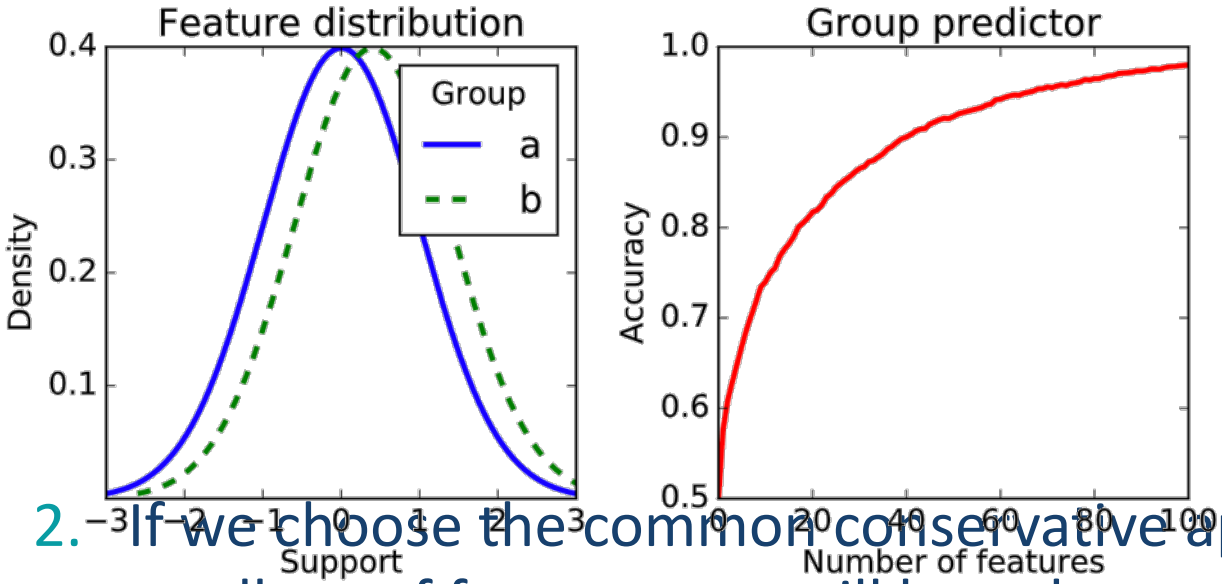
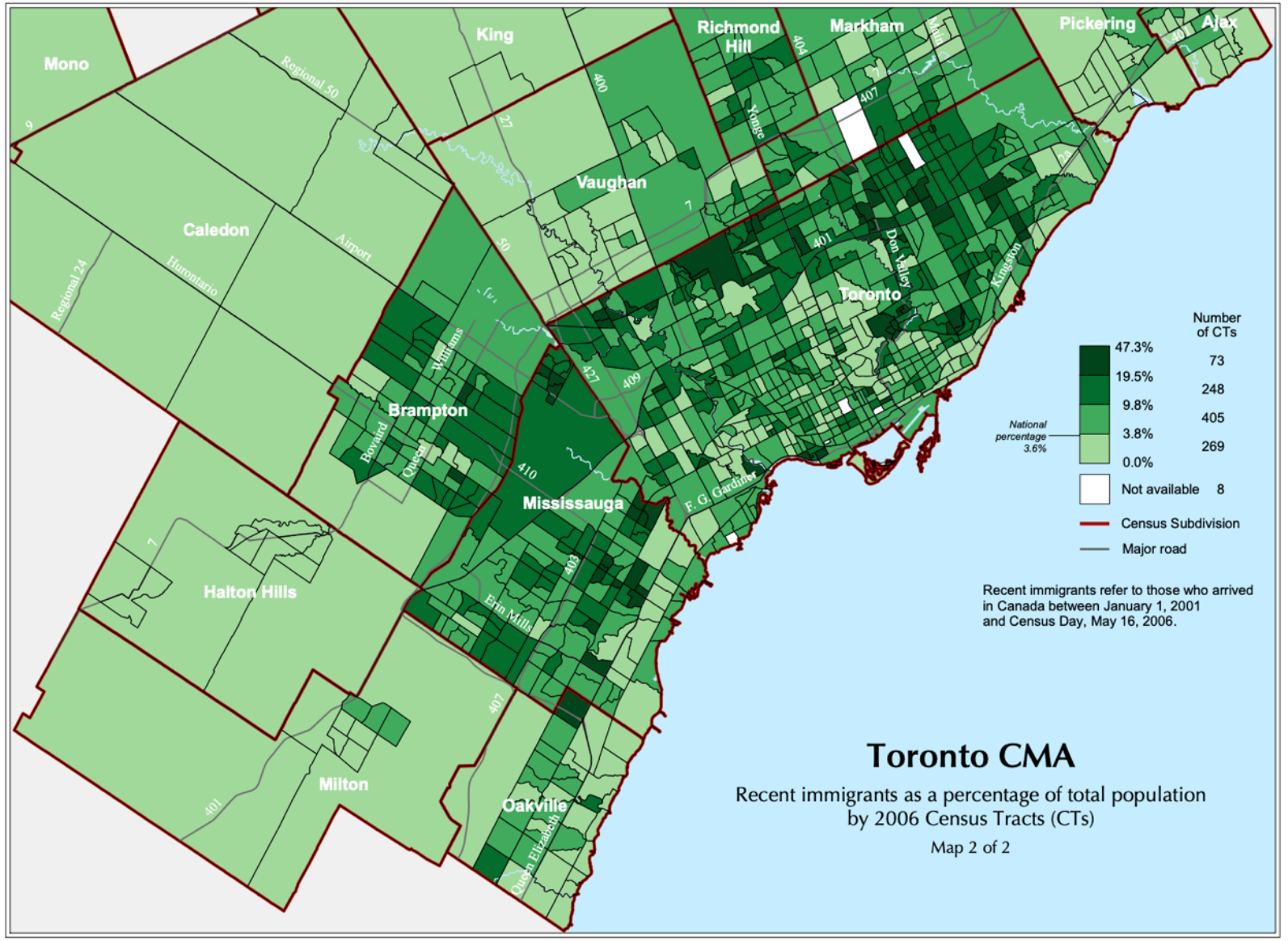


Figure 4: On the left, we see the distribution of a single feature that differs only very slightly between the two groups. In both groups the feature follows a normal distribution. Only the means are slightly different in each group. Multiple features like this can be used to build a high accuracy group membership classifier. On the right, we see how the accuracy grows as more and more features become available.

Source: fairmlbook.org

2. If we choose the common conservative approa small set of features, we will hurt the model accuracy (we will discuss this more in the fair representation learning section)

Example: Toronto's Recent Immigrants Distribution (2006)



Counterfactual Fairness

- Another individual fairness.
- It means if we only change the sensitive feature of an individual (counterfactual world) the likelihood of the outcome stays the same.

- Problem: all the descendants of the sensitive attribute should be avoided as features for classification.

Tools to Evaluate and Visualize Fairness

- Fairness-indicators: Tensorflow's Fairness Evaluation and Visualization Toolkit (Google)
- AI Fairness 360 (IBM)
- Fairlearn: Fairness in machine learning mitigation algorithms (Microsoft)
- The LinkedIn Fairness Toolkit (LiFT)
- Algofairness
- FairSight: Visual Analytics for Fairness in Decision Making
- Aequitas: Bias and Fairness Audit Toolkit
- CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models
- ML-fairness-gym: Google's implementation based on OpenAI's Gym
- scikit-fairness
- Mitigating Gender Bias In Captioning System

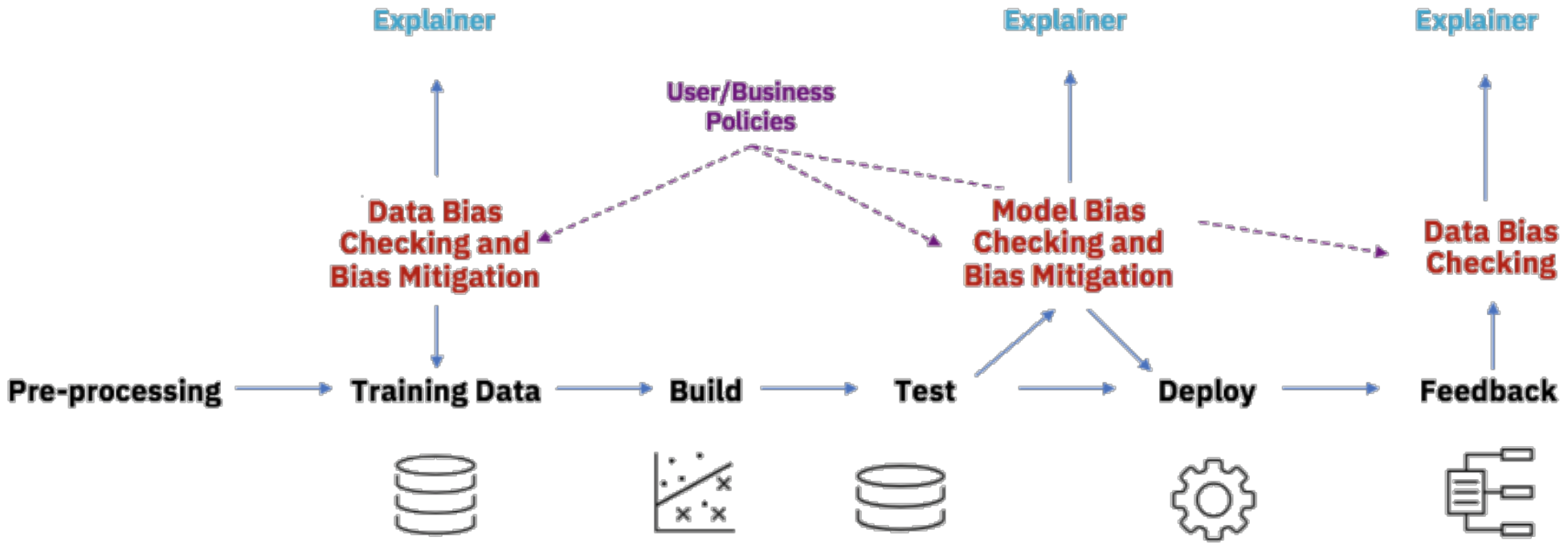
Source: www.linkedin.com/pulse/overview-some-available-fairness-frameworks-packages-murat-durmus/

Example: TensorFlow's Fairness-indicators

- Fairness Indicators is designed to support teams in evaluating and improving models for fairness concerns in partnership with the broader TensorFlow toolkit.
- It includes the ability to:
  - Evaluate the distribution of datasets
  - Evaluate model performance, sliced across defined groups of users
  - Feel confident about your results with confidence intervals and evals at multiple thresholds
  - Dive deep into individual slices to explore root causes and opportunities for improvement

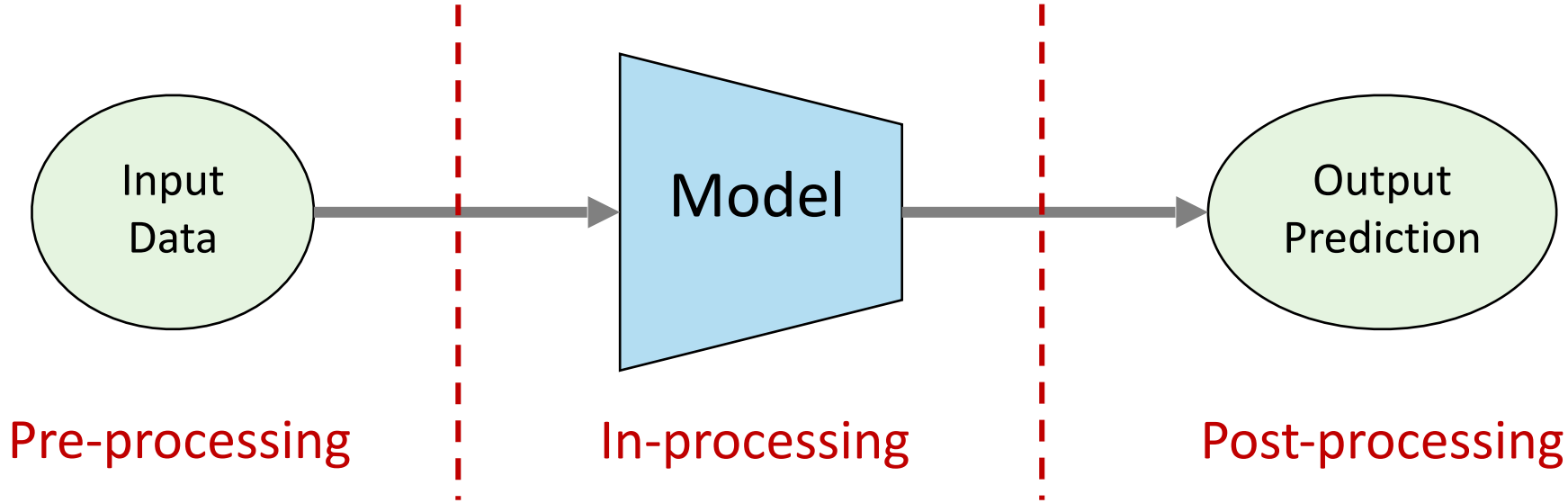
Example: IBM Fairness 360

- A comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias.



General Methods to Satisfy Fairness

- Pre-processing: transform data to remove underlying bias.
- In-processing: modify the state-of-the-art learning algorithms in order to remove bias during the model training process.
- Post-processing: reassign the labels assigned by the black-box model to eliminate bias.



Comparison of the Methods

Method	When to Use	Pros	Cons
Pre-Processing	When you have access and control on data	• Unbiased data can be used for any other tasks too • No need to modify the model • Sometimes it eliminates the need for sensitive features at test time	• Most of fairness criteria, that need predicted labels, can't be used • Eliminating some features can hurt the model accuracy
In-Processing	When you don't have access and control on data but have control on the algorithm	• better fairness vs. accuracy trade off • All fairness criteria can be used • Sometimes it eliminates the need for sensitive features at test time	• Task specific methods • Changing the model can be complex and sometime not possible
Post-processing	When you have a black-box model (no access to training data and the model)	• no need to modify and re-train the main model • All fairness criteria can be used • Can be used with any model	• losing the flexibility in choosing fairness criteria • Requires sensitive features at test time

Balancing Data (pre-processing method)

- Adding new data:
  - An expensive option (labeled data is expensive).
  - Privacy considerations
  - Be careful of exploitation (e.g., CloudWalk Technology and Zimbabwe Government)
- Removing sensitive features or balancing data by removing data:
  - Removing data hurts model accuracy
  - Correlation between other features and the sensitive features

Fair Representation Learning (pre-processing method)

- We already discussed why removing sensitive features is not a a good idea and adding data is also not always feasible.
- New Idea: to learn a representation (Z) of the data (X) that preserves as much information as possible but removes any information about the sensitive attribute (S). (separate the responsibilities of the trusted society and untrusted vendor)

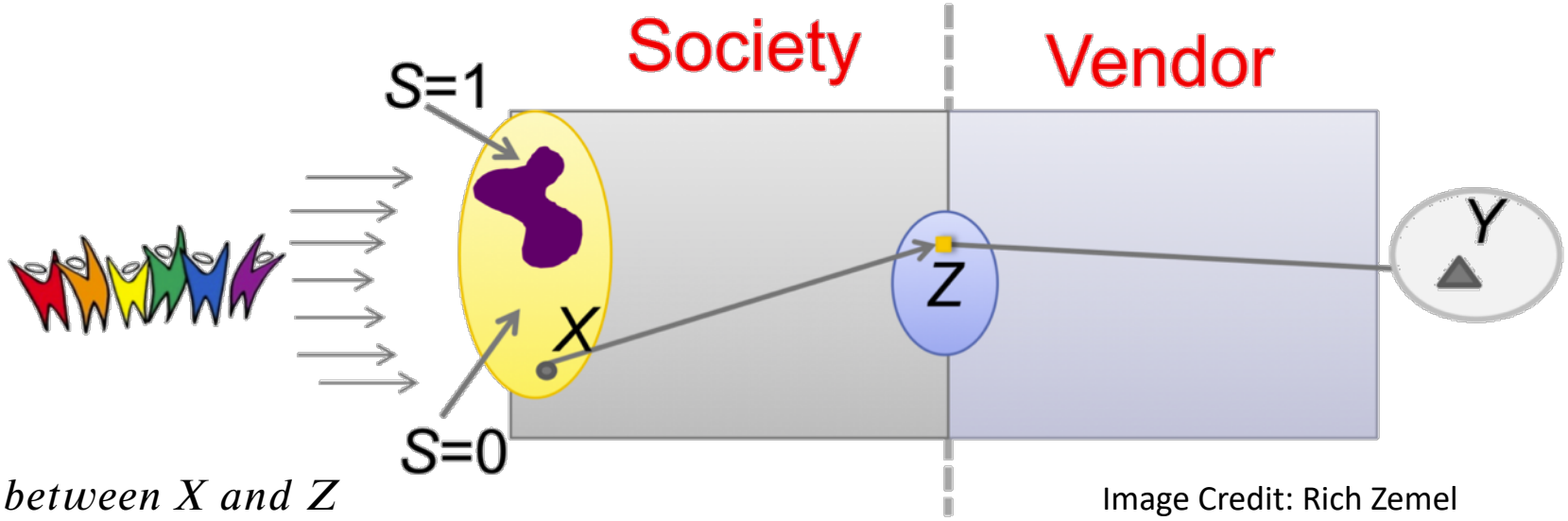


Image Credit: Rich Zemel

$* I(X, Z) = Mutual\ information\ between\ X\ and\ Z$

Fair Representation Learning (pre-processing method)

- We already discussed why removing sensitive features is not a a good idea and adding data is also not always feasible.
- New Idea: to learn a representation (Z) of the data (X) that preserves as much information as possible but removes any information about the sensitive attribute (S). (separate the responsibilities of the trusted society and untrusted vendor)

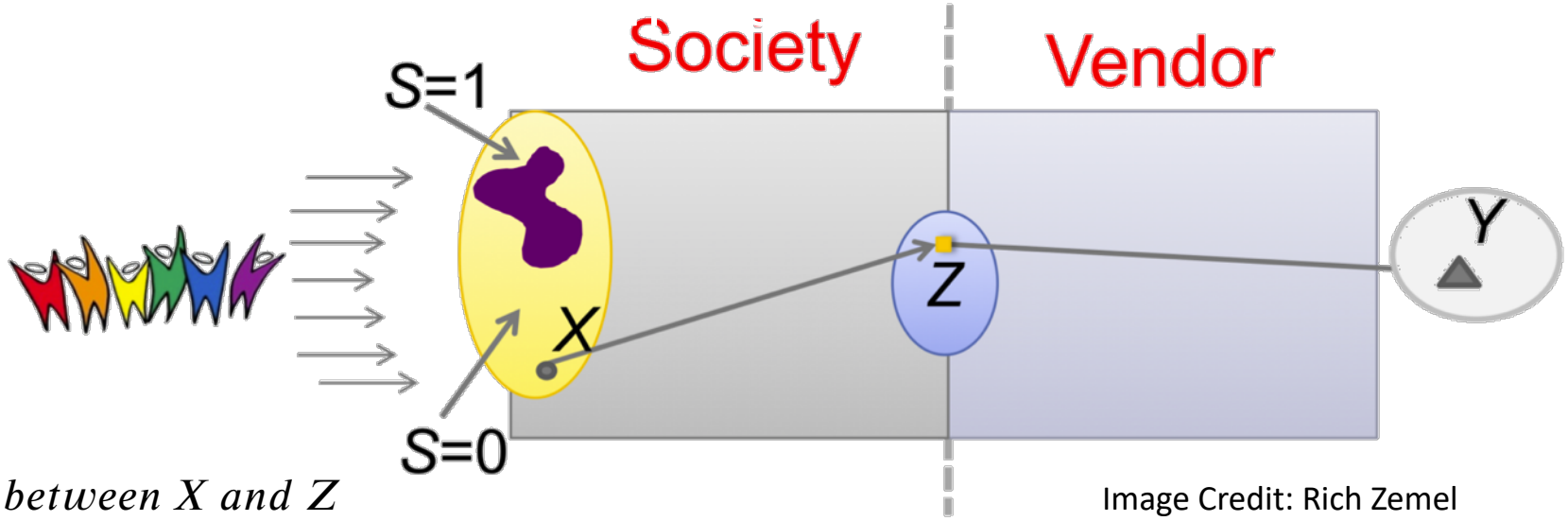


Image Credit: Rich Zemel

$* I(X, Z) = Mutual\ information\ between\ X\ and\ Z$

Fair Model (In-processing method)

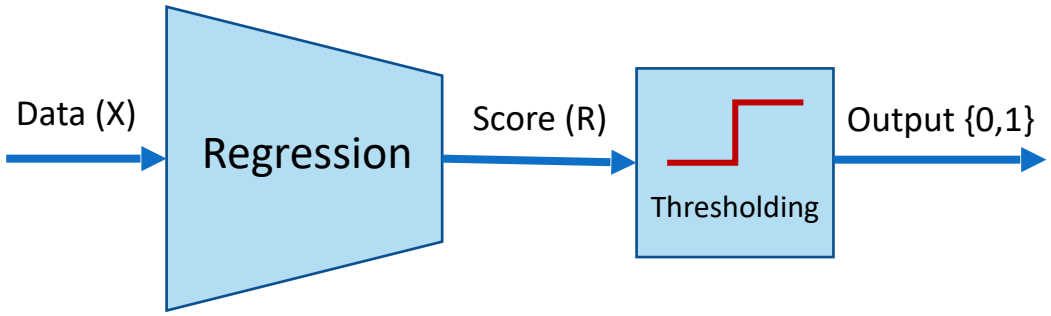
- One simple Idea is to add a constraint term to the objective function to satisfy one of the fairness criteria. This approach is very similar to regularization that we use to mitigate overfitting.
- For example, we can train a classifier with objective function of $L(\theta)$ with the constraint to satisfy equal opportunity criterion:

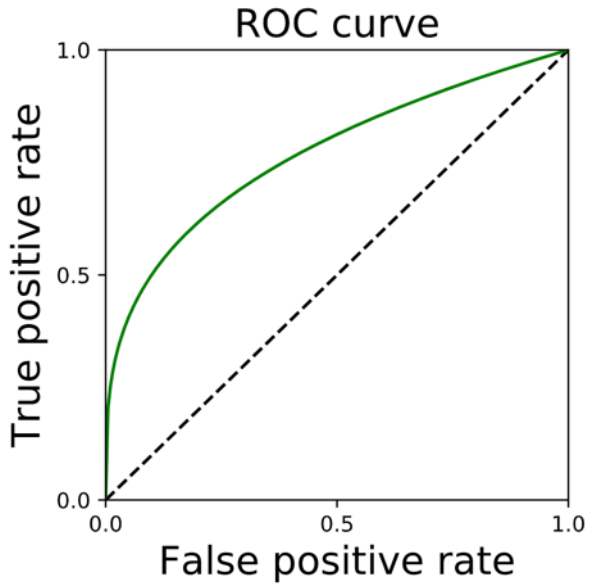
$$Maximize\ L(\theta)\ subject\ to:$$

$$P\left(\hat{Y}=1 \;\middle|\; Y=1, S=0\right) = P\left(\hat{Y}=1 \;\middle|\; Y=1, S=1\right)$$

Post-processing Method

- The main idea is to change the classification thresholds to maximize fairness. No re-training is needed.
- <u>Reminder:</u> a classifier usually uses a regression model to predict a score (or probability) and a threshold to classify. We use a ROC diagram to evaluate the accuracy of a classifier.

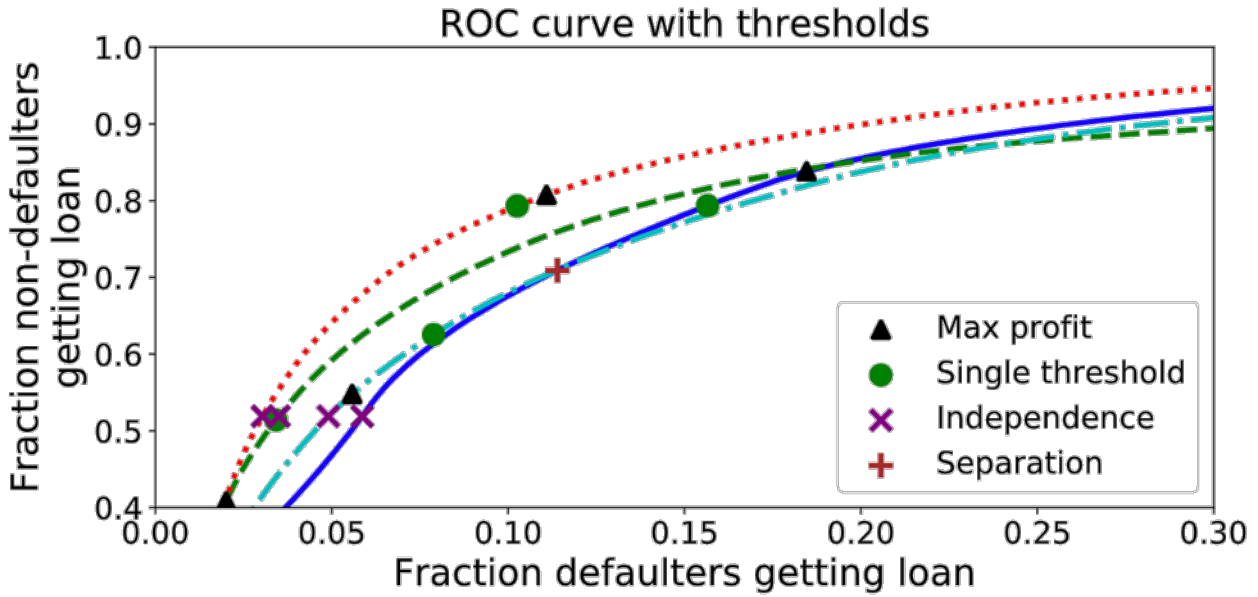




Source: fairmlbook.org/

Post-processing Method - continued

- For example, we use of the following strategies to find the suitable threshold (sensitive feature = ethnicity):
  - Maximum profit: Pick possibly group-dependent score thresholds in a way that maximizes profit.
  - Single threshold: Pick a single uniform score threshold for all groups in a way that maximizes profit.
  - Separation: Achieve an equal true/false positive rate in all groups. Subject to this constraint, maximize profit.
  - Independence: Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.



Source: fairmlbook.org/

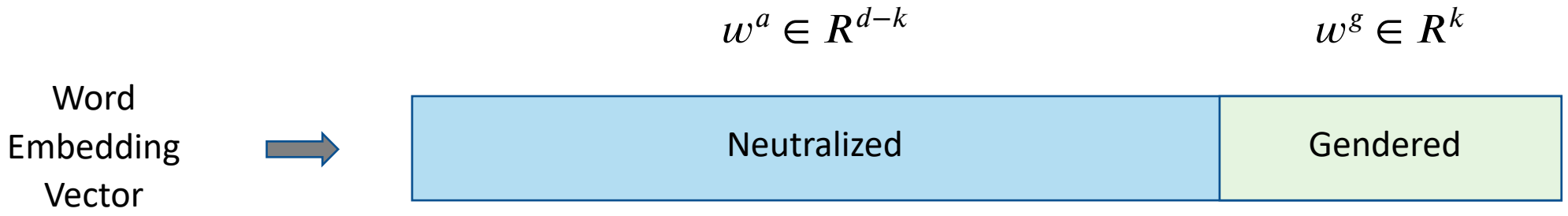
Fair NLP
(Some Examples)

Debiasing Word Embedding

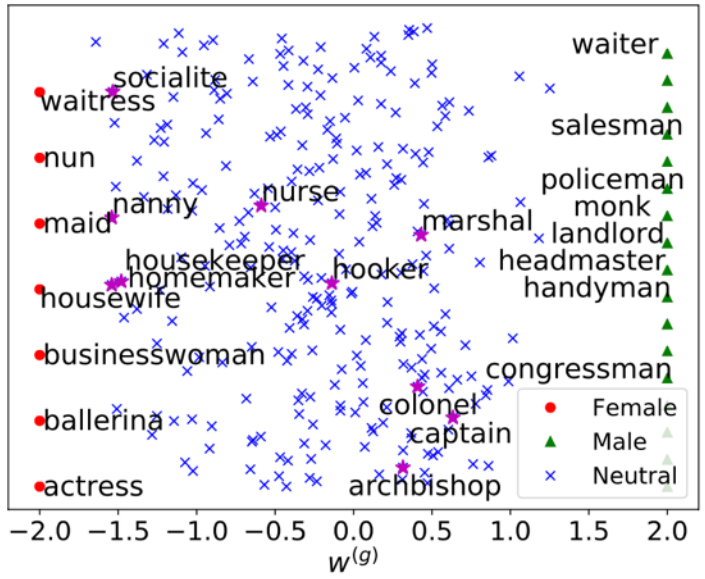
- T. Bolukbasi et al. (2016) suggest a method to reduce gender biases in the word embedding while preserving the useful properties of the embedding:
  1. Reduce bias:
     a) Ensure that gender neutral words such as nurse are equidistant between gender pairs such as he and she.
     b) Reduce gender associations that pervade the embedding even among gender neutral words.
  2. Maintain embedding utility:
     a) Maintain meaningful non-gender-related associations between gender neutral words, including associations within stereotypical categories of words such as fashion-related words or words associated with football.
     b) Correctly maintain definitional gender associations such as between man and father

Debiasing Word Embedding - continued

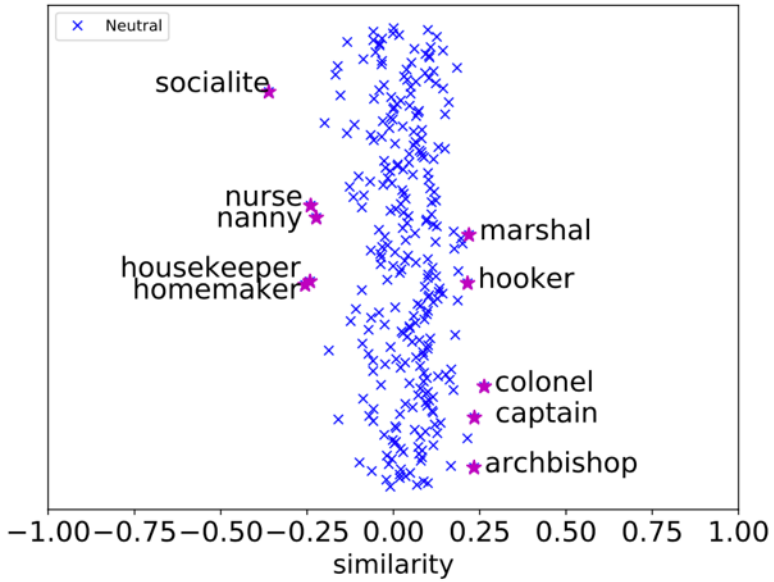
- Zhao et al. (2018) suggest an embedding model (based on GloVe) that has two separate parts of neutralized and gendered (can be used for other protected classes too).
- The idea is to reserve the gender feature into gendered dimensions. Therefore, the information encoded in neutralized dimensions is independent of gender influence.



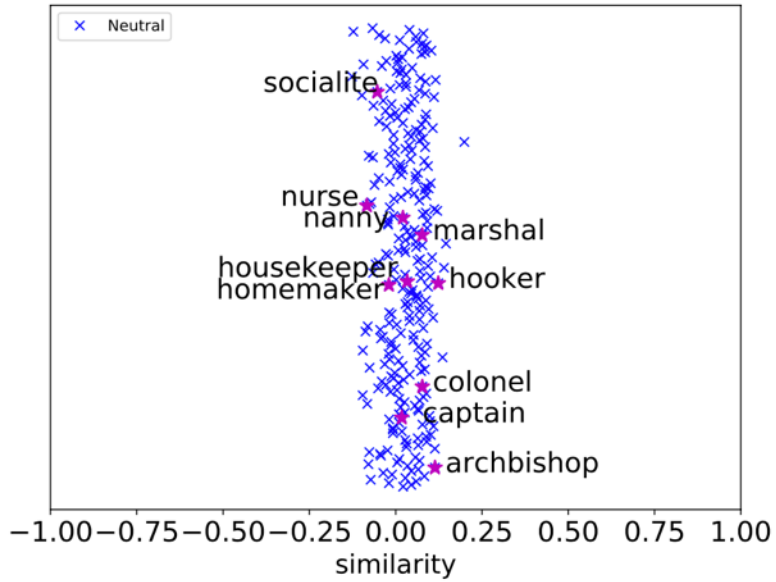
Debiasing Word Embedding - continued



(a) $w^{(g)}$ dimension for all the professions



(b) Gender-neutral profession words projected to gender direction in GloVe

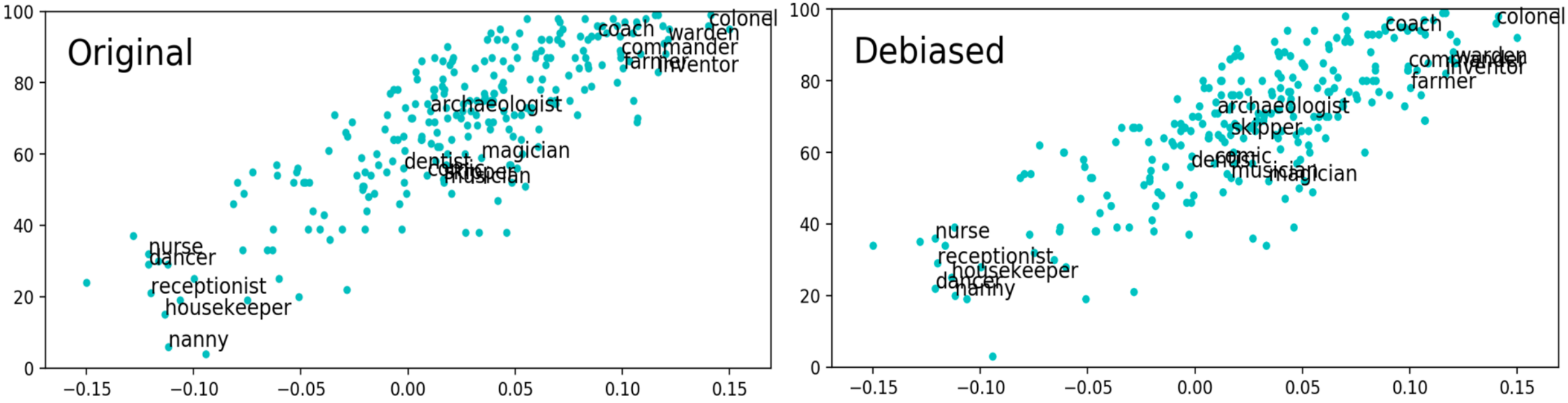


(c) Gender-neutral profession words projected to gender direction in GN-GloVe

Zhao et al. (2018)

Debiasing is Hard (but awareness is better than blindness)

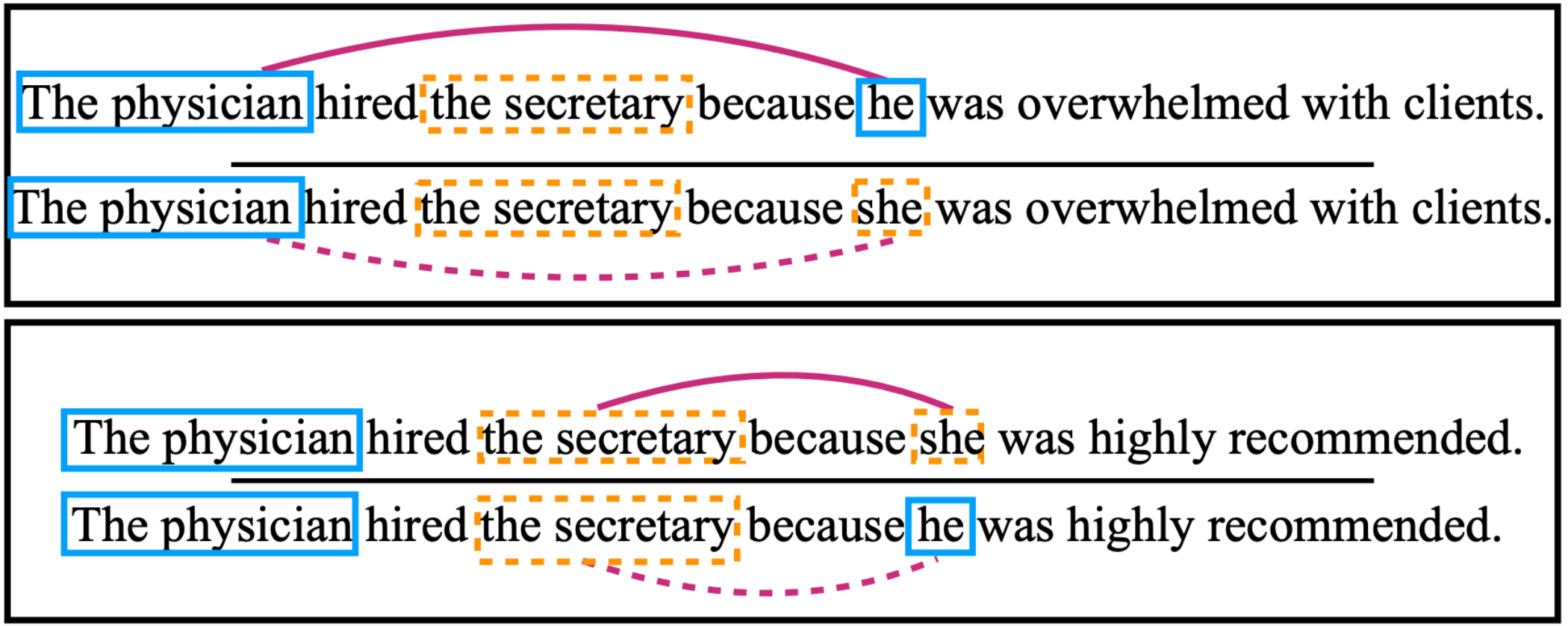
- Gonen et al. (2019) argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it.



The number of male neighbors for each profession as a function of its original bias

Bias in Coreference Resolution (WinoBias)

Zhao et al. (2018)



Pro-stereotypical scenarios (solid purple lines) and anti-stereotypical scenarios (dashed purple lines)

Debiasing Coreference Resolution

- Zhao et al. (2018) introduce a data-augmentation technique that removes bias in the existing state-of-the-art coreferencing methods, in combination with using word2vec debiasing techniques.
- Data Augmentation:
  - Swapping gender words.
  - Anonymizing the named entities

Brian talked to his secretary.

[NE] talked to her secretary.

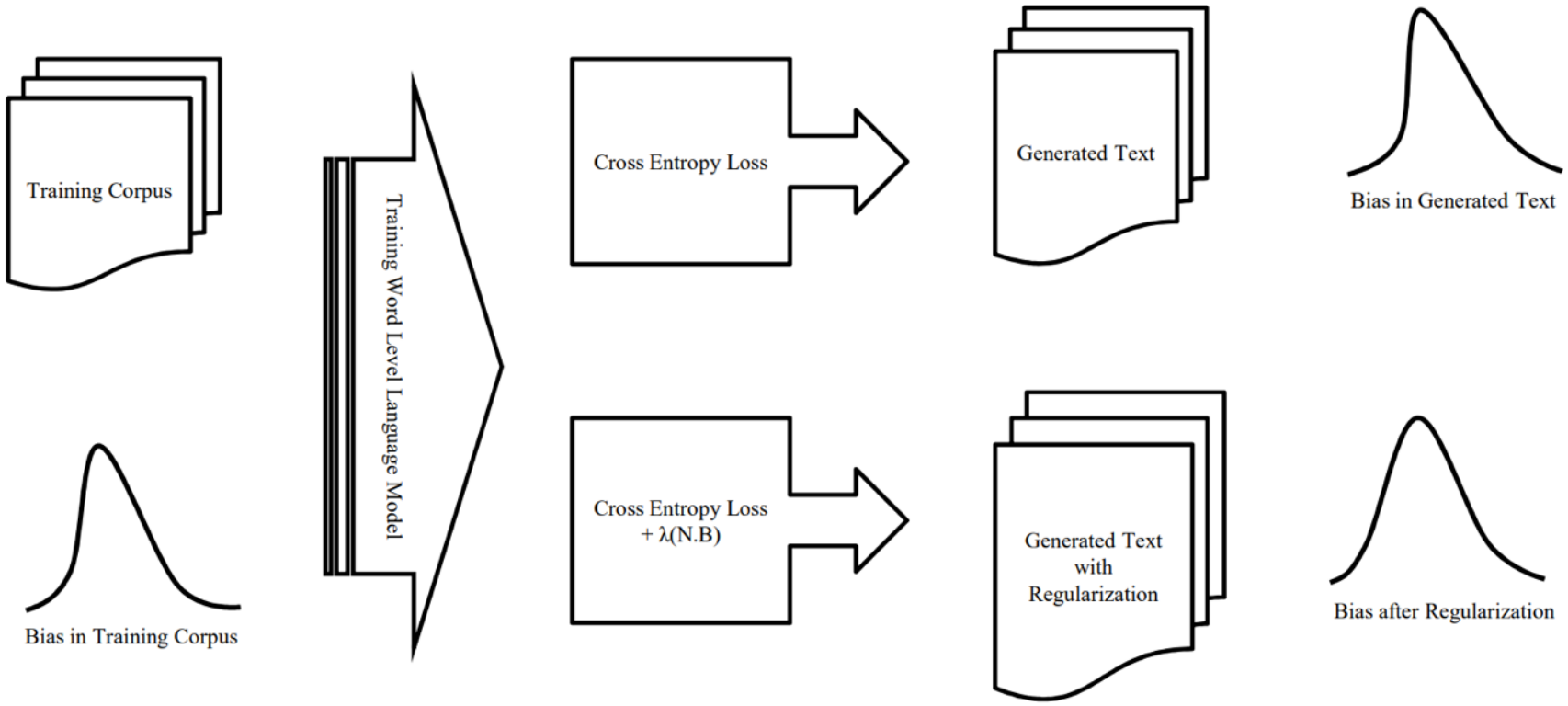
Bias in Language Generation

- Sheng et al. (2019)
- Examples of text continuations generated from OpenAI's medium-size GPT-2 model, given different prompts

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Reducing Gender Bias in Word-Level Language Models

- Bordia et al. (2019) propose a regularization loss term for the language model that minimizes the projection of encoder-trained embeddings onto an embedding subspace that encodes gender.



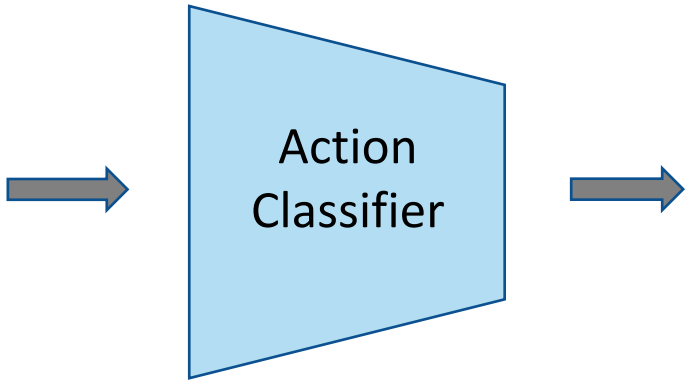
Removing bias in NLP is still an open problem

Fair Computer Vision
(Some Examples Beyond Balancing Datasets)

Bias in Image Classification

- The goal is to classify the action if it is cooking (binary classification)
- Out training dataset is biased, and our model will classify based on gender not the action.
- What should be done without changing training data?

Source: Google Images

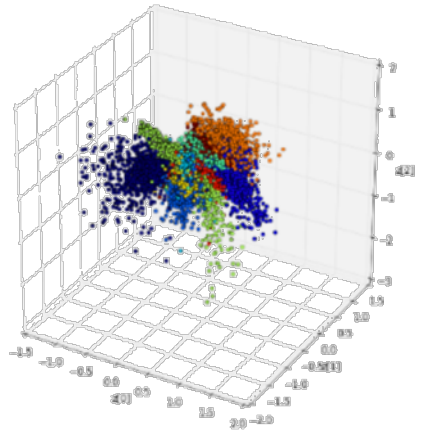


Protected Feature = Gender

Option 1: Fair Representation Learning

Source: Google Images

Nutral Representation (Z)



Protected Feature = Gender

Action Classifier

Cooking? (Yes/No)

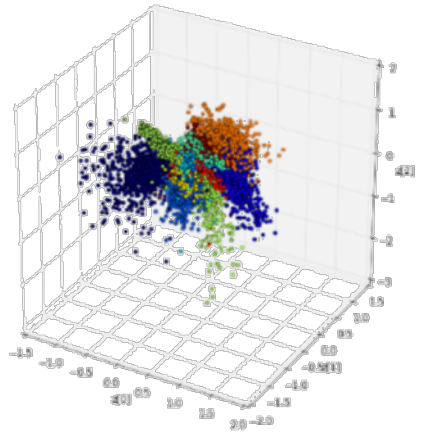
Learning Fair Representations , ICML 2013
Zemel, Wu, Swersky, Pitassi, Dwork.

Option 2: Adversarial Feature Learning

Source: Google Images

Protected Feature = Gender

Nutral Representation (Z)



Maximize Performance

Action Classifier

Cooking? (Yes/No)

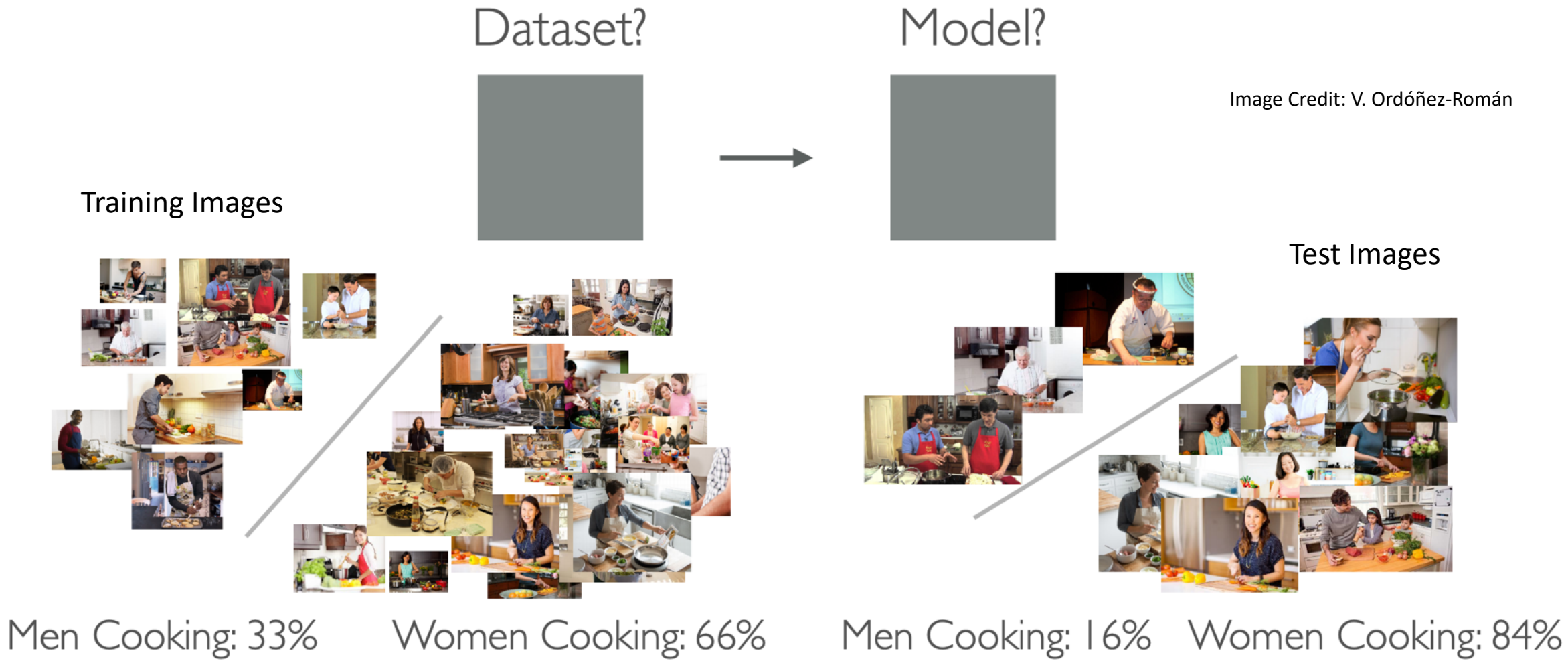
Gender Discriminator

Gender? (M/F)

Minimize Performance

Controllable Invariance through Adversarial Feature Learning, NeurIPS 2017, Xie, Dai, Du, Hovy, Neubig.

Models can Amplify Biases in Data

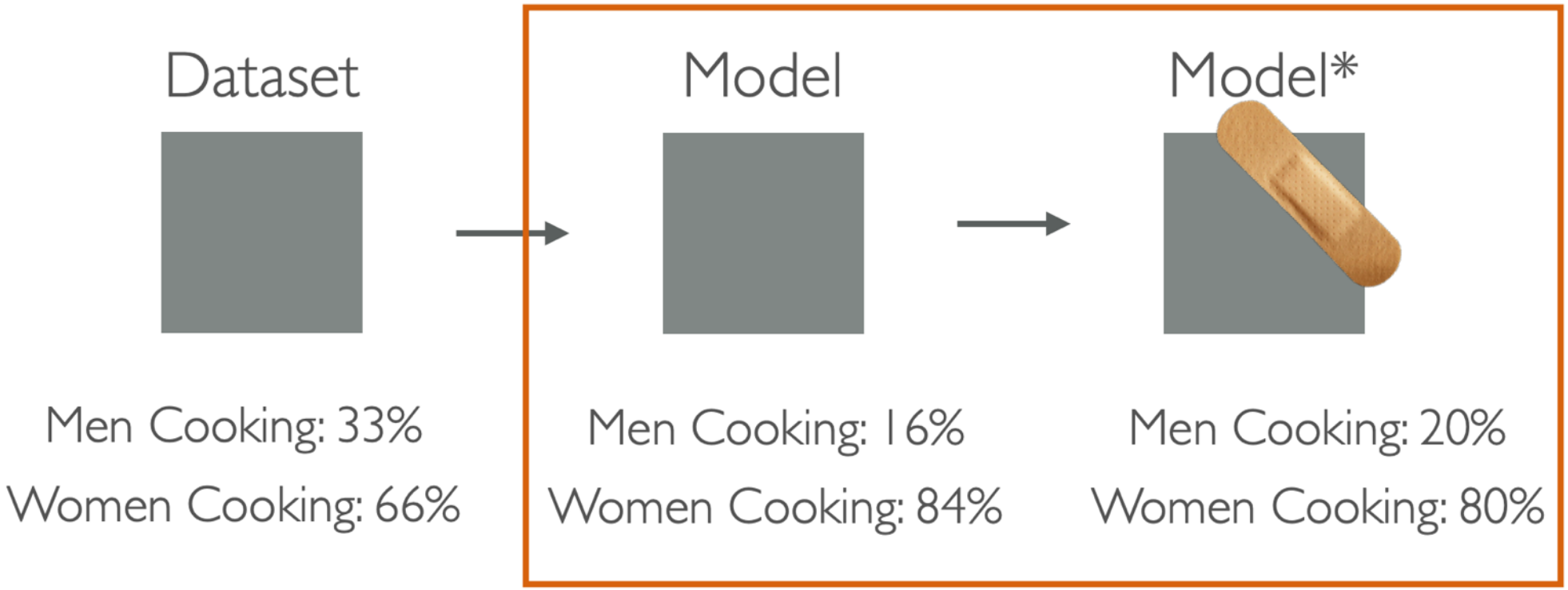


Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. EMNLP 2017

Models can Amplify Biases in Data - continued

Reducing Bias Amplification (RBA): Optimize for accuracy but also to match data distribution

Image Credit: V. Ordóñez-Román

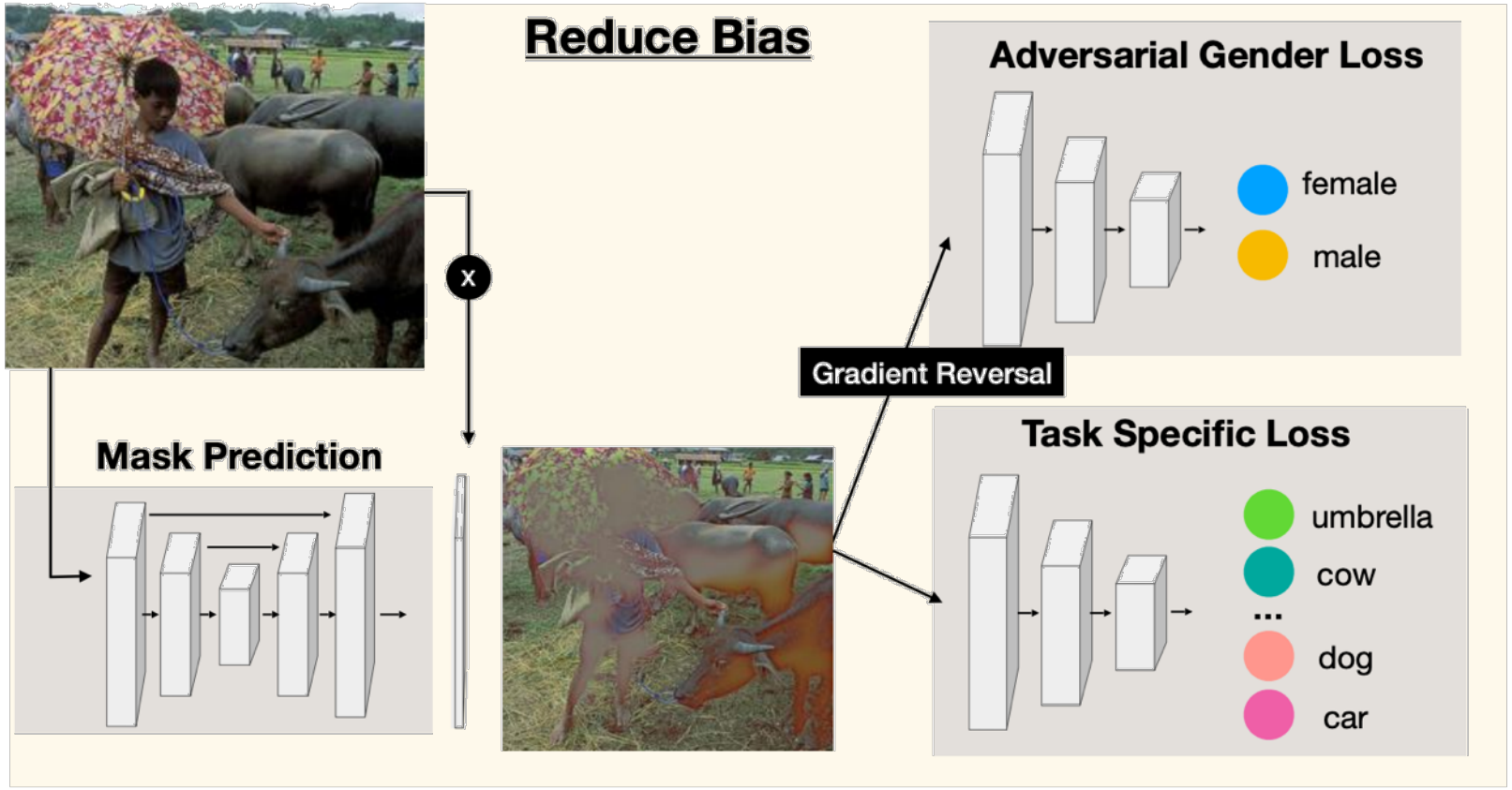


Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. EMNLP 2017

Balanced Datasets Are Not Enough

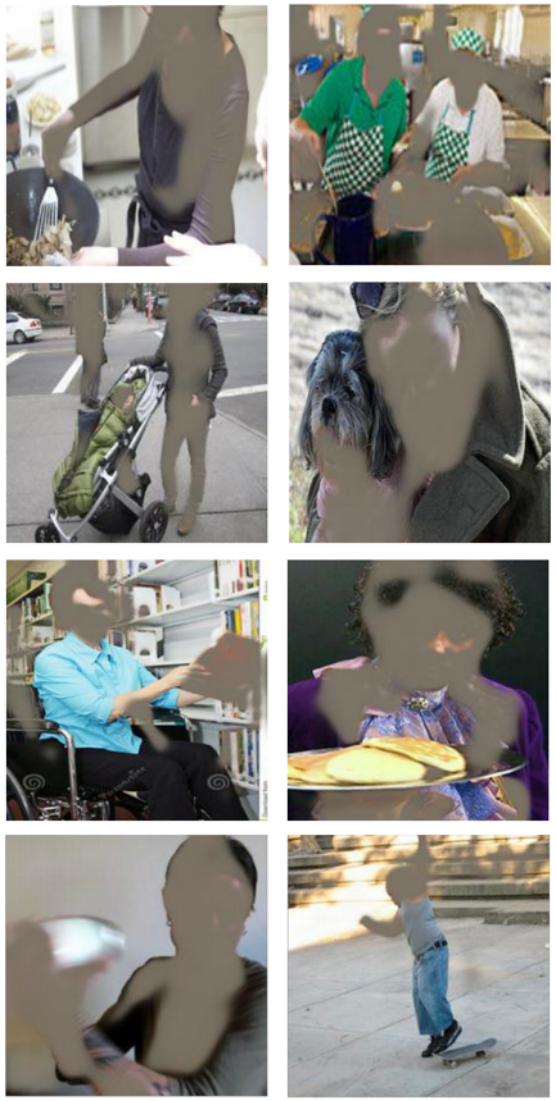
- Even when datasets are balanced such that each label co-occurs equally with each gender, learned models amplify the association between labels and gender, as much as if data had not been balanced!

- Mitigation: Adversarial Removal of Sensitive Features



Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations.
Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez.
ICCV 2019

Balanced Datasets Are Not Enough - continued



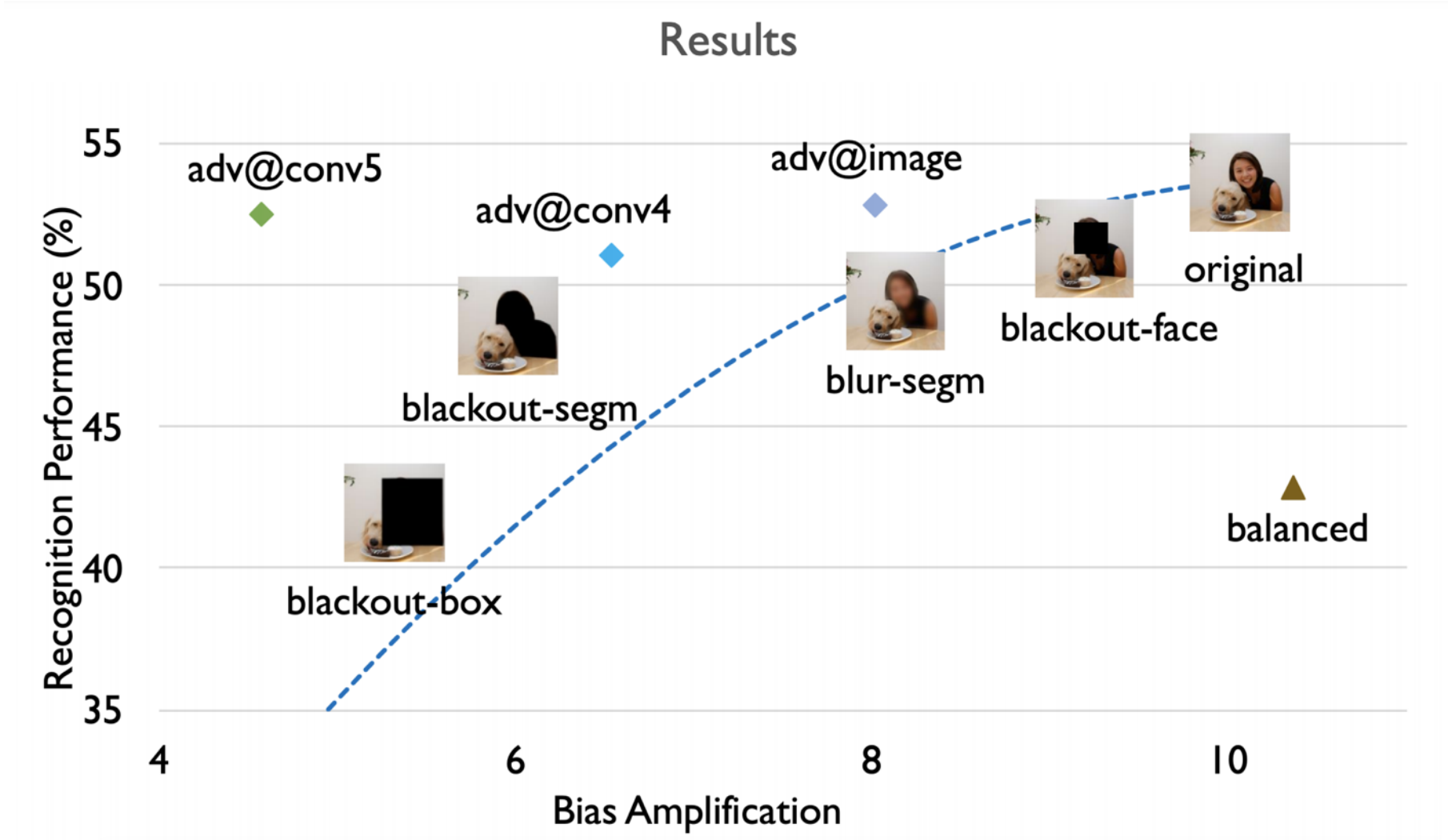
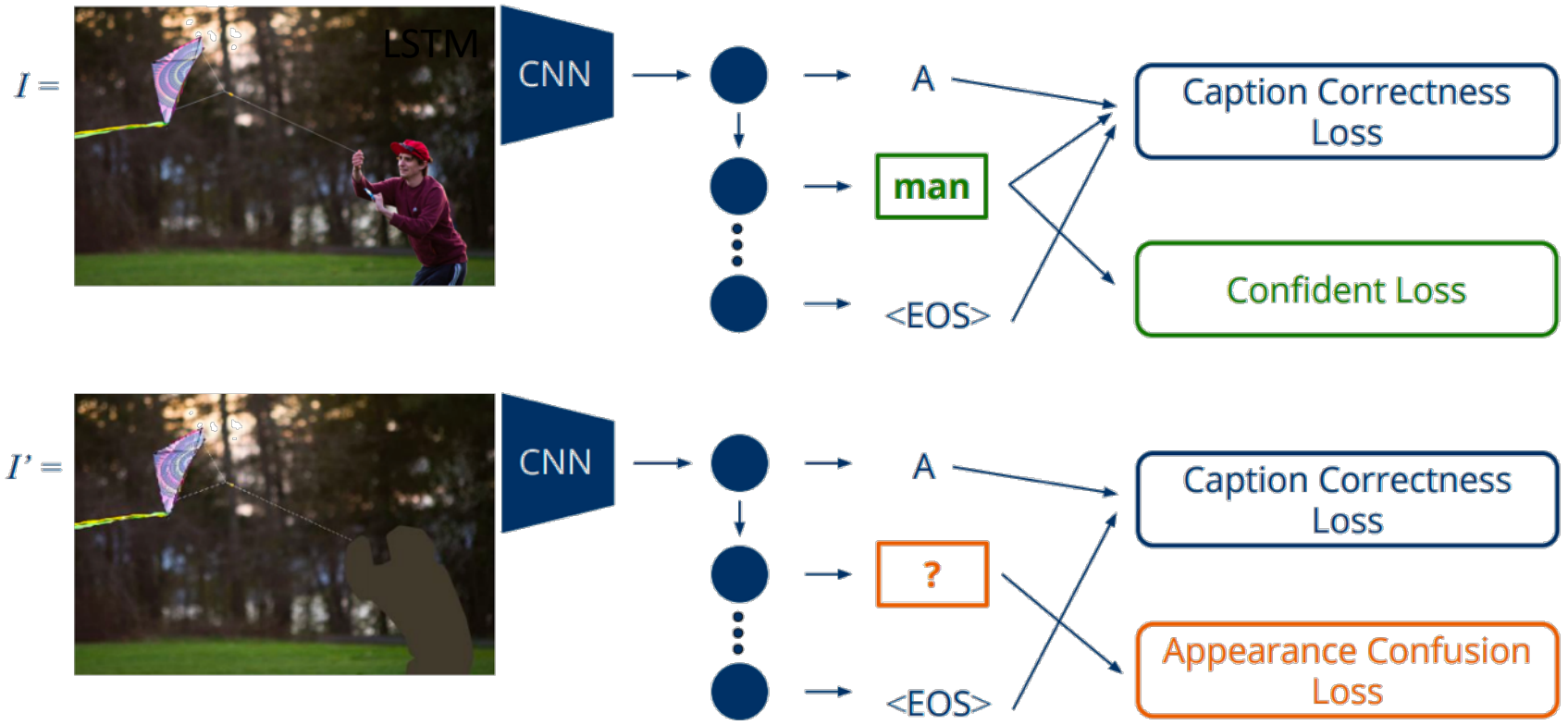


Image Credit: V. Ordóñez-Román

Bias in Image Captioning

- A new Equalizer model that encourages equal gender probability when gender evidence is occluded in a scene and confident predictions when gender evidence is present.



Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach

Major References

- Fairness and machine learning, Limitations and Opportunities (eBook)
  Solon Barocas, Moritz Hardt, Arvind Narayanan
- A Survey on Bias and Fairness in Machine Learning (2019)
  NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA,
  KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI
- A Tutorial on Fairness in Machine Learning (2018)
  Ziyuan Zhong
- A primer on AI fairness (2019)
  Kenn So
- Bias and Discrimination in AI
  UMontrealX and IVADO
- Tutorial: Bias and Fairness in Natural Language Processing, EMNLP 2019
  Kai-Wei Chang, Vicente Ordonez, Margaret Mitchell, Vinodkumar Prabhakaran