*DS-GA 3001.009: Responsible Data Science*

# Algorithmic Fairness

Prof. Julia Stoyanovich
Center for Data Science
Computer Science and Engineering at Tandon

@stoyanoj

http://stoyanovich.org/
https://dataresponsibly.github.io/

# Slack

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Sign In

LEGISLATIVE RESEARCH CENTER

Council Home | Legislation | Calendar | City Council | Committees

RSS | Alerts

Details | Reports

| | | | |
|---|---|---|---|
| File #: | Int 1696-2017 Version: A | Name: | Automated decision systems used by agencies. |
| Type: | Introduction | Status: | Enacted |
| | | Committee: | Committee on Technology |
| On agenda: | 8/24/2017 | | |
| Enactment date: | 1/11/2018 | Law number: | 2018/049 |
| Title: | A Local Law in relation to automated decision systems used by agencies | | |
| Sponsors: | James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca | | |
| Council Member Sponsors: | 9 | | |
| Summary: | This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems. | | |
| Indexes: | Oversight | | |
| Attachments: | 1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017 | | |

# NYC ADS transparency law

1/11/2018

**Int. No. 1696-A**: A Local Law in relation to automated decision systems used by agencies

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Sign In

LEGISLATIVE RESEARCH CENTER

Council Home | Legislation | Calendar | City Council | Committees

RSS | Alerts

Details | Reports

| | | | |
|---|---|---|---|
| File #: | Int 1696-2017 Version: A | Name: | Automated decision systems used by agencies. |
| Type: | Introduction | Status: | Enacted |
| | | Committee: | Committee on Technology |
| On agenda: | 8/24/2017 | | |
| Enactment date: | 1/11/2018 | Law number: | 2018/049 |
| Title: | A Local Law in relation to automated decision systems used by agencies | | |
| Sponsors: | James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca | | |
| Council Member Sponsors: | 9 | | |
| Summary: | This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems. | | |
| Indexes: | Oversight | | |
| Attachments: | 1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017 | | |

# The original draft

8/16/2017

Int. No. 1696

By Council Member Vacca

A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons

Be it enacted by the Council as follows:

Section 1. Section 23-502 of the administrative code of the city of New York is amended to add a new subdivision g to read as follows:

g. Each agency that uses, for the purposes of targeting services to persons, imposing penalties upon persons or policing, an algorithm or any other method of automated processing system of data shall:

1. Publish on such agency's website, the source code of such system; and

2. Permit a user to (i) submit data into such system for self-testing and (ii) receive the results of having such data processed by such system.

§ 2. This local law takes effect 120 days after it becomes law.

MAJ
LS# 10948
8/16/17 2:13 PM

this is **NOT** what was adopted

# Summary of Int. No. 1696-A

Form an automated decision systems (**ADS**) task force that surveys current use of algorithms and data in City agencies and develops procedures for:

- requesting and receiving an **explanation** of an algorithmic decision affecting an individual (3(b))
- interrogating ADS for **bias and discrimination** against members of legally-protected groups (3(c) and 3(d))
- allowing the **public** to **assess** how ADS function and are used (3(e)), and archiving ADS together with the data they use (3(f))

we've come a long way from the original draft!

# Get engaged!

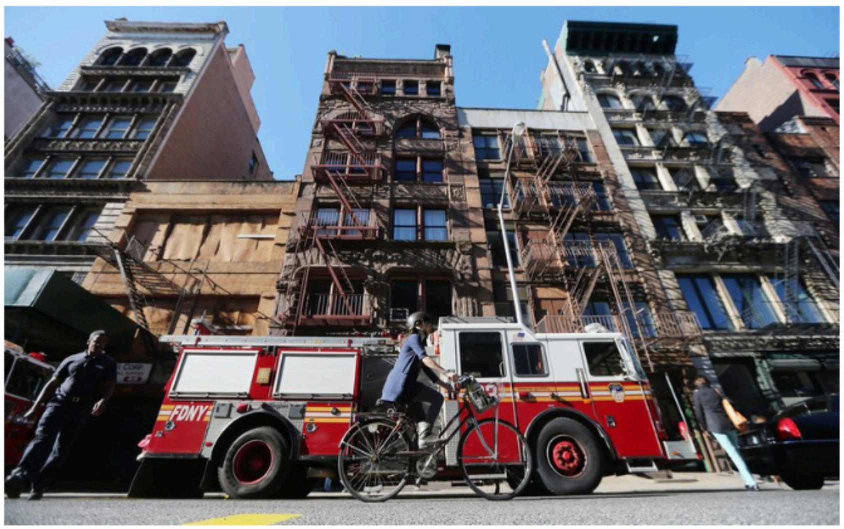
10/16/2017

THE NEW YORKER

By Julia Powles December 20, 2017

ELEMENTS

## NEW YORK CITY'S BOLD, FLAWED ATTEMPT TO MAKE ALGORITHMS ACCOUNTABLE




*Automated systems guide the allocation of everything from firehouses to food stamps. So why don't we know more about them?*

Photograph by Mario Tama / Getty

# The ADS Task Force

Visit **alpha.nyc.gov** to help us test out new ideas for NYC's website.

The Official Website of the City of New York

NYC

简体中文 Translate Text Size

NYC Resources | NYC311 | Office of the Mayor | Events | Connect | Jobs | Search

Mayor | First Lady | News | Officials

SHARE

Email

Print

## Mayor de Blasio Announces First-In-Nation Task Force To Examine Automated Decision Systems Used By The City

**May 16, 2018**

**NEW YORK—** Today, Mayor de Blasio announced the creation of the Automated Decision Systems Task Force which will explore how New York City uses algorithms. The task force, the first of its kind in the U.S., will work to develop a process for reviewing "automated decision systems," commonly known as algorithms, through the lens of equity, fairness and accountability.

"As data and technology become more central to the work of city government, the algorithms we use to aid decision making must be aligned with our goals and values," said **Mayor de Blasio**. "The establishment of the Automated Decision Systems Task Force is an important first step towards greater transparency and equity in our use of technology."

# February 12, 2019

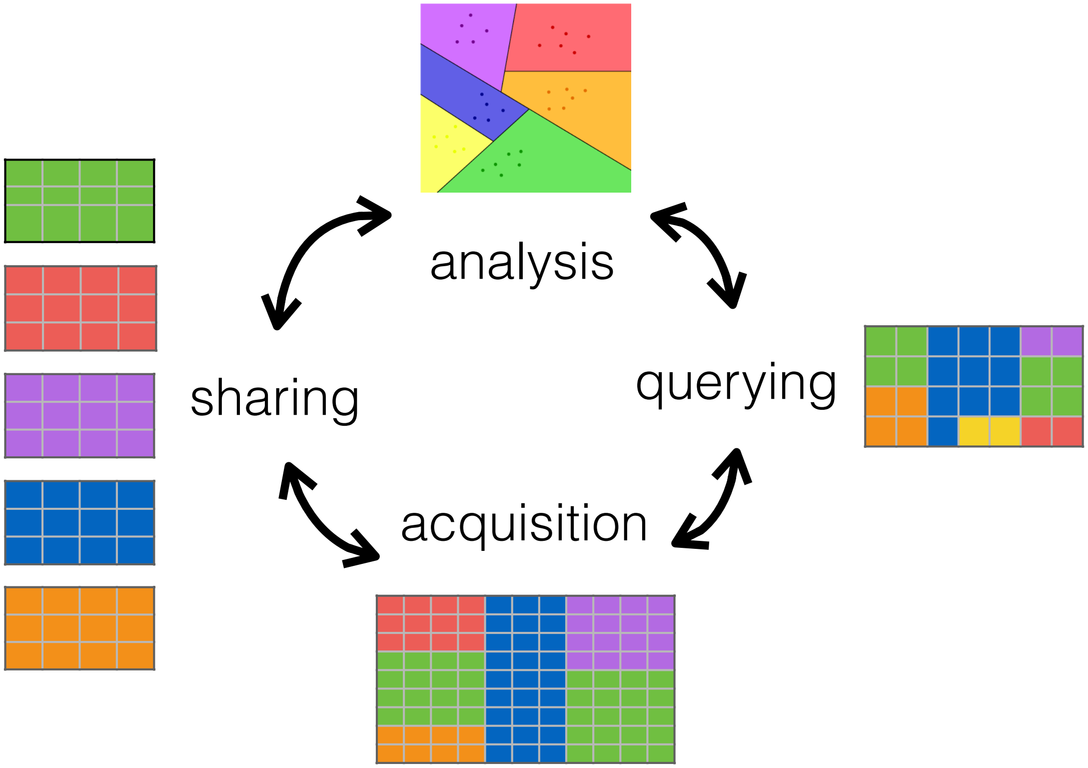
THE NEW YORK CITY COUNCIL

Corey Johnson, Speaker

Please be advised about the changes to the next Technology Committee hearing. The hearing will be held jointly with the **Commission on Public Information and Communication (COPIC)** on T**uesday, February 12, 2019 at 1 pm in the 14th Floor Committee Room, 250 Broadway, New York, NY 10007**.

**The Committees will take testimony on the role of COPIC with respect to improving government transparency, improving the public's access to government information, protecting personal information privacy, and facilitating data sharing between city agencies.** You are hereby invited to attend this meeting and testify therein. Please feel free to bring with you such members of your staff you deem appropriate to the subject matter.

# The big picture



# Urban homelessness

## *Mayor de Blasio Scrambles to Curb Homelessness After Years of Not Keeping Pace*

By J. DAVID GOODMAN and NIKITA STEWART JAN. 13, 2017

The New York Times

https://www.nytimes.com/2017/01/13/nyregion/mayor-de-blasio-scrambles-to-curb-homelessness-after-years-of-not-keeping-pace.html

Volunteers during the homeless census in February 2015. In a decision made by M
York City stopped opening shelters for much of that year. Stephanie Keith for The New

Ms. Glen emphasized that the construction of new housing takes several years, a long-term solution whose effect on homelessness could not yet be evaluated.

# Urban homelessness

## *Homeless Young People of New York, Overlooked and Underserved*

By NIKITA STEWART FEB. 5, 2016

The New York Times

https://www.nytimes.com/2016/02/06/nyregion/young-and-homeless-in-new-york-overlooked-and-underserved.html

Abdul, 23, at Safe Horizon in Harlem, has been homeless since 2010. Jake Naugh

Last year, the total number of sheltered and unsheltered homeless people in the city was 75,323, which included 1,706 people between ages 18 and 24. The actual number of young people is significantly higher, according to the service providers, who said the census mostly captured young people who received social services. The census takers were not allowed to enter private businesses, including many of the late-night spots where young people often create an ad hoc shelter by pretending to be customers.

# ADS example: urban homelessness

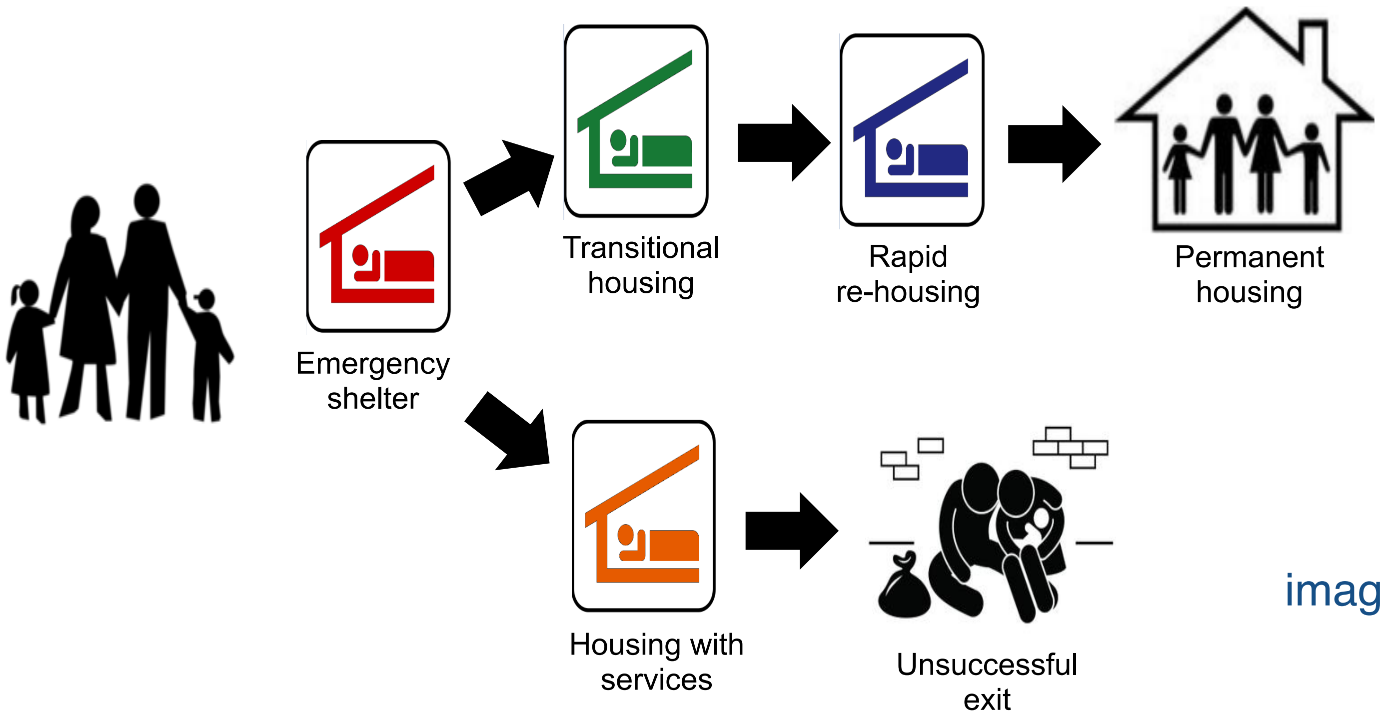



image by Bill Howe

- **Services:** rapid rehousing, transitional housing, emergency shelter, permanent supportive housing
- **Support mechanisms:** substance abuse treatment, mental health treatment, protection for victims of domestic violence

# ADS example: urban homelessness

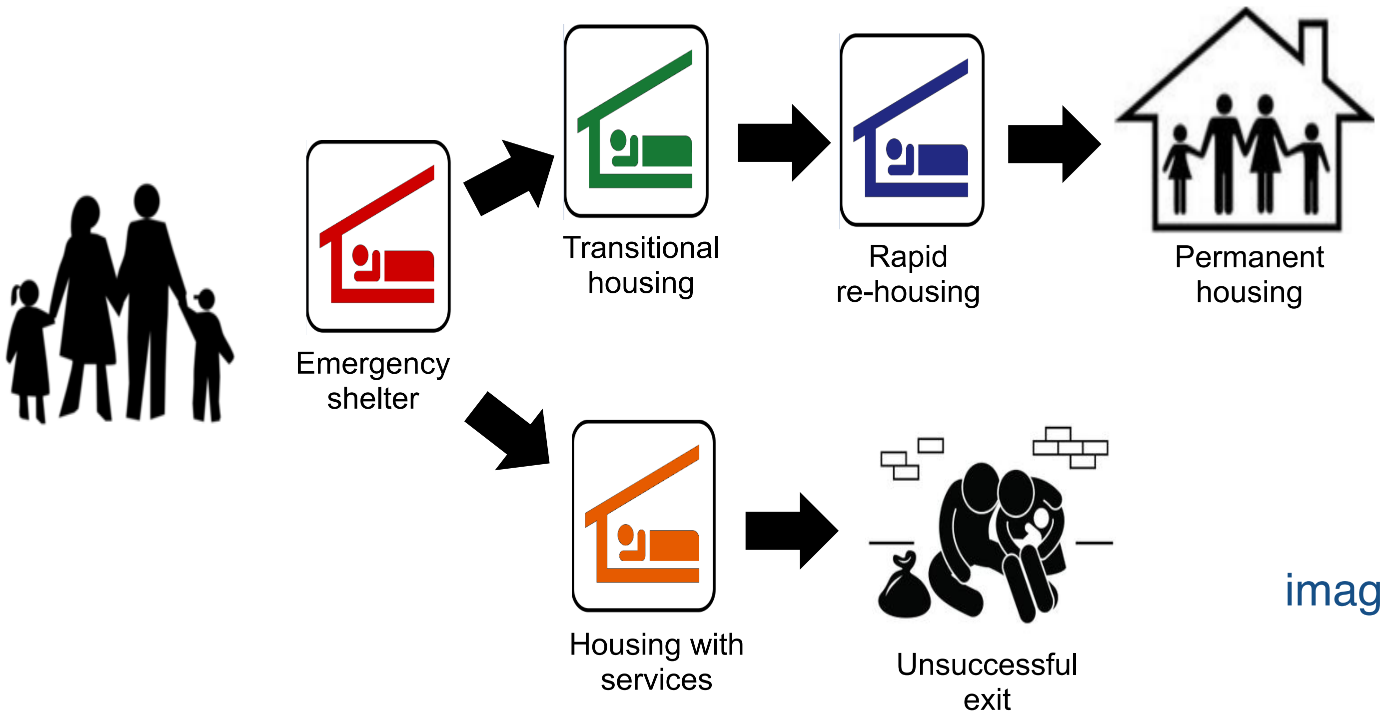



image by Bill Howe

- **Allocate** interventions: services and support mechanisms
- **Recommend** pathways through the system
- **Evaluate** effectiveness of interventions, pathways, over-all system

# ADS example: urban homelessness

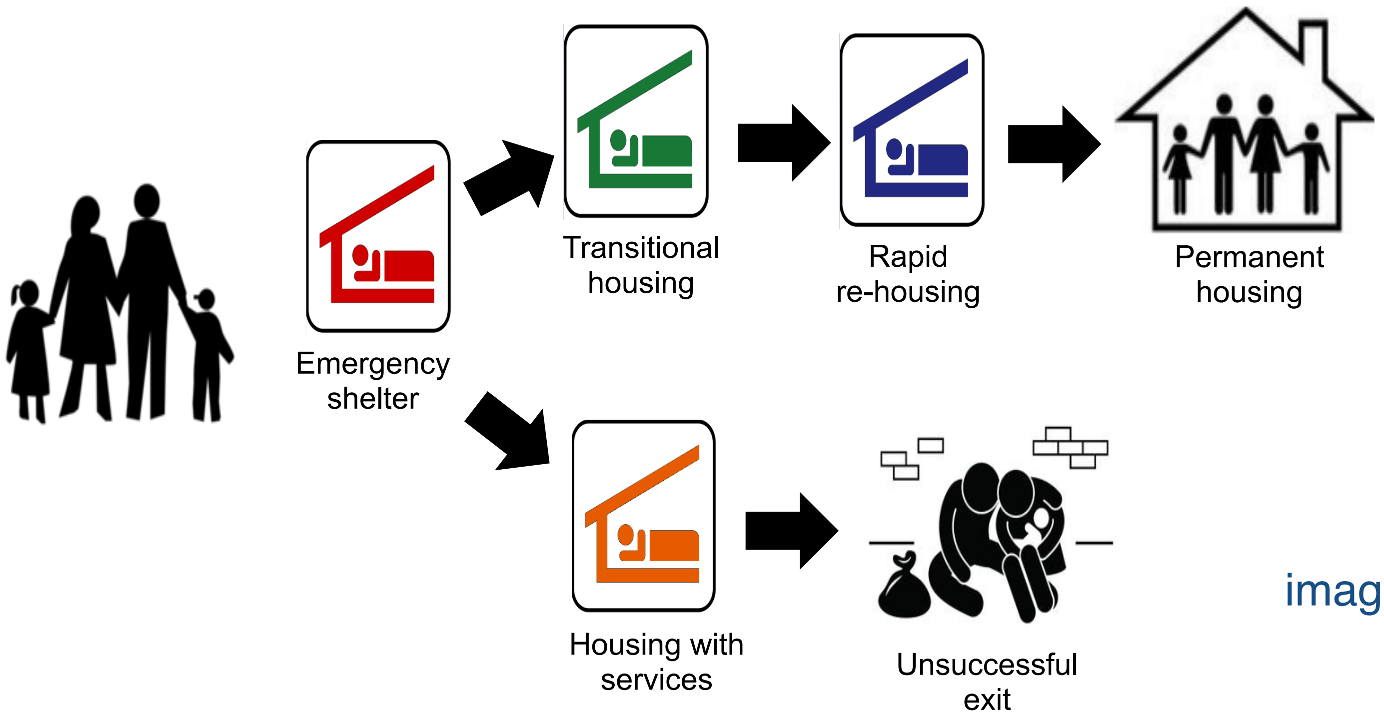
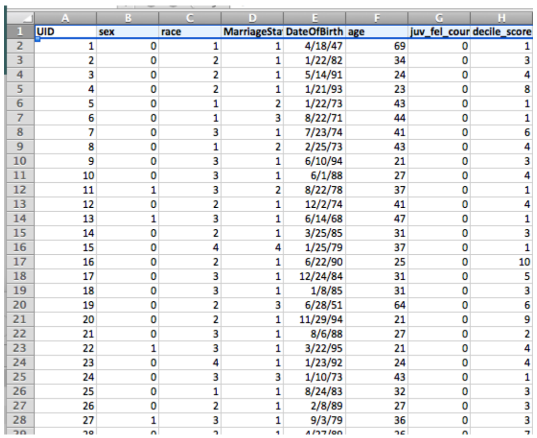


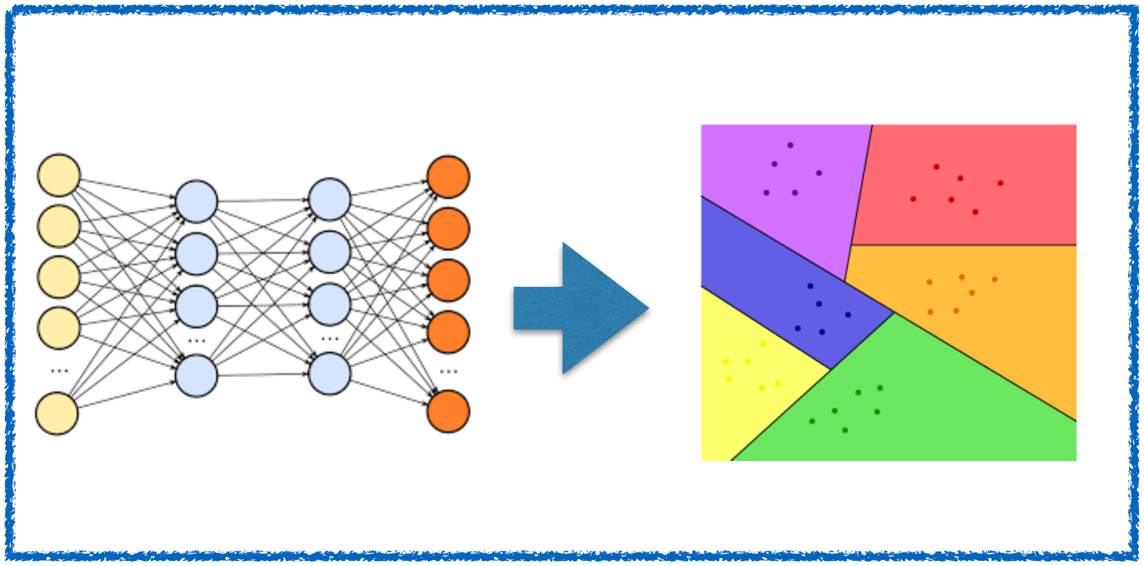

image by Bill Howe

- Be **transparent** and **accountable**
- Achieve **equitable** resource distribution
- Be cognizant of the **rights** and **preferences** of individuals

# Responsible data science



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | UID | sex | race | MarriageSta | DateOfBirth | age | juv_fel_cour | decile_score |
| 2 | 1 | 0 | 1 | 1 | 4/18/47 | 69 | 0 | 1 |
| 3 | 2 | 0 | 2 | 1 | 1/22/82 | 34 | 0 | 3 |
| 4 | 3 | 0 | 2 | 1 | 5/14/91 | 24 | 0 | 4 |
| 5 | 4 | 0 | 2 | 1 | 1/21/93 | 23 | 0 | 8 |
| 6 | 5 | 0 | 1 | 2 | 1/22/73 | 43 | 0 | 1 |
| 7 | 6 | 0 | 1 | 3 | 8/22/71 | 44 | 0 | 1 |
| 8 | 7 | 0 | 3 | 1 | 7/23/74 | 41 | 0 | 6 |
| 9 | 8 | 0 | 1 | 2 | 2/25/73 | 43 | 0 | 4 |
| 10 | 9 | 0 | 3 | 1 | 6/10/94 | 21 | 0 | 3 |
| 11 | 10 | 0 | 3 | 1 | 6/1/88 | 27 | 0 | 4 |
| 12 | 11 | 1 | 3 | 2 | 8/22/78 | 37 | 0 | 1 |
| 13 | 12 | 0 | 2 | 1 | 12/2/74 | 41 | 0 | 4 |
| 14 | 13 | 1 | 3 | 1 | 6/14/68 | 47 | 0 | 1 |
| 15 | 14 | 0 | 2 | 1 | 3/25/85 | 31 | 0 | 3 |
| 16 | 15 | 0 | 4 | 4 | 1/25/79 | 37 | 0 | 1 |
| 17 | 16 | 0 | 2 | 1 | 6/22/90 | 25 | 0 | 10 |
| 18 | 17 | 0 | 3 | 1 | 12/24/84 | 31 | 0 | 5 |
| 19 | 18 | 0 | 3 | 1 | 1/8/85 | 31 | 0 | 3 |
| 20 | 19 | 0 | 2 | 3 | 6/28/51 | 64 | 0 | 6 |
| 21 | 20 | 0 | 2 | 1 | 11/29/94 | 21 | 0 | 9 |
| 22 | 21 | 0 | 3 | 1 | 8/6/88 | 27 | 0 | 2 |
| 23 | 22 | 1 | 3 | 1 | 3/22/95 | 21 | 0 | 4 |
| 24 | 23 | 0 | 4 | 1 | 1/23/92 | 24 | 0 | 4 |
| 25 | 24 | 0 | 3 | 3 | 1/10/73 | 43 | 0 | 1 |
| 26 | 25 | 0 | 1 | 1 | 8/24/83 | 32 | 0 | 3 |
| 27 | 26 | 0 | 2 | 1 | 2/8/89 | 27 | 0 | 3 |
| 28 | 27 | 1 | 3 | 1 | 9/3/79 | 36 | 0 | 3 |



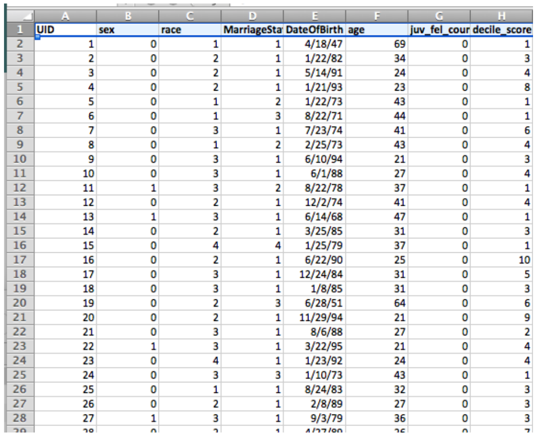
done?

but where does the data come from?

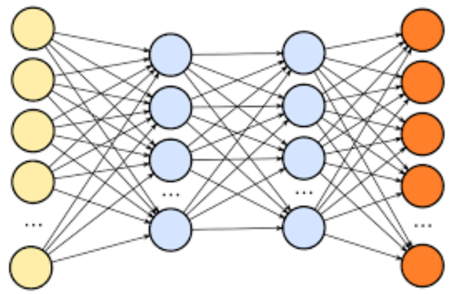
# How did we get the data?

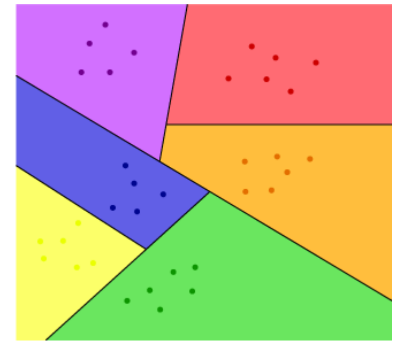
- A multitude of datasets gathered from local communities, data is **weakly structured**: inconsistencies, missing values, hidden and apparent bias
- Some data was **anonymized**, other data was **not shared** in fear of violating regulations or the trust of participants
- Shared data was **triaged**, **aligned**, **integrated** (ETL + SQL)
- Integrated data was then **filtered** (SQL) and **prioritized** (sorted/ranked), and only then passed as input to the learning module

# Mitigating urban homelessness



| UID | sex | race | MarriageSta | DateOfBirth | age | juv_fel_cour | decile_score |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 4/18/47 | 69 | 0 | 1 |
| 2 | 0 | 2 | 1 | 1/22/82 | 34 | 0 | 3 |
| 3 | 0 | 2 | 1 | 5/14/91 | 24 | 0 | 4 |
| 4 | 0 | 2 | 1 | 1/21/93 | 23 | 0 | 8 |
| 5 | 0 | 1 | 2 | 1/22/73 | 43 | 0 | 1 |
| 6 | 0 | 1 | 3 | 8/22/71 | 44 | 0 | 1 |
| 7 | 0 | 3 | 1 | 7/23/74 | 41 | 0 | 6 |
| 8 | 0 | 1 | 2 | 2/25/73 | 43 | 0 | 4 |
| 9 | 0 | 3 | 1 | 6/10/94 | 21 | 0 | 3 |
| 10 | 0 | 3 | 1 | 6/1/88 | 27 | 0 | 4 |
| 11 | 1 | 3 | 2 | 8/22/78 | 37 | 0 | 1 |
| 12 | 0 | 2 | 1 | 12/2/74 | 41 | 0 | 4 |
| 13 | 1 | 3 | 1 | 6/14/68 | 47 | 0 | 1 |
| 14 | 0 | 2 | 1 | 3/25/85 | 31 | 0 | 3 |
| 15 | 0 | 4 | 4 | 1/25/79 | 37 | 0 | 1 |
| 16 | 0 | 2 | 1 | 6/22/90 | 25 | 0 | 10 |
| 17 | 0 | 3 | 1 | 12/24/84 | 31 | 0 | 5 |
| 18 | 0 | 3 | 1 | 1/8/85 | 31 | 0 | 3 |
| 19 | 0 | 2 | 3 | 6/28/51 | 64 | 0 | 6 |
| 20 | 0 | 2 | 1 | 11/29/94 | 21 | 0 | 9 |
| 21 | 0 | 3 | 1 | 8/6/88 | 27 | 0 | 2 |
| 22 | 1 | 3 | 1 | 3/22/95 | 21 | 0 | 4 |
| 23 | 0 | 4 | 1 | 1/23/92 | 24 | 0 | 4 |
| 24 | 0 | 3 | 3 | 1/10/73 | 43 | 0 | 1 |
| 25 | 0 | 1 | 1 | 8/24/83 | 32 | 0 | 3 |
| 26 | 0 | 2 | 1 | 2/8/89 | 27 | 0 | 3 |
| 27 | 1 | 3 | 1 | 9/3/79 | 36 | 0 | 3 |





finding: women are underrepresented in the favorable outcome groups (group fairness)

**fix the model!**

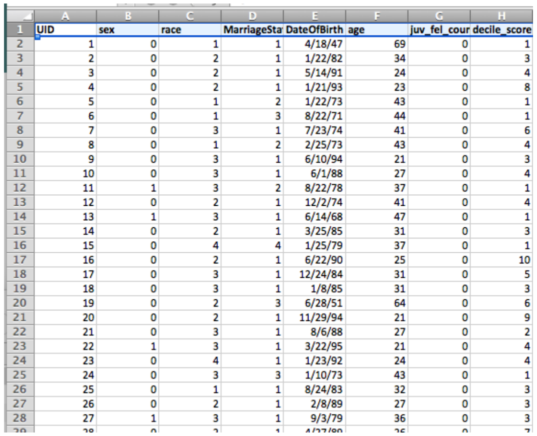
**of course, but maybe… the input was generated with:**

select * from R
where status = ‘unsheltered’
and length > 2 1 month

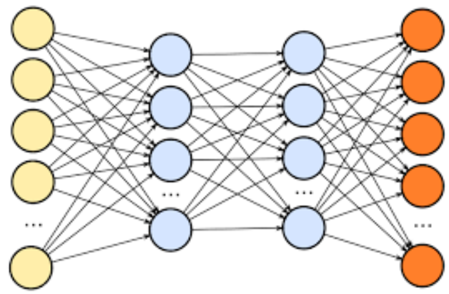
10% female

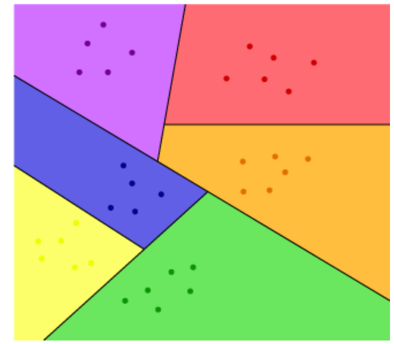
40% female

# Mitigating urban homelessness



| UID | sex | race | MarriageSta | DateOfBirth | age | juv_fel_cour | decile_score |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 4/18/47 | 69 | 0 | 1 |
| 2 | 0 | 2 | 1 | 1/22/82 | 34 | 0 | 3 |
| 3 | 0 | 2 | 1 | 5/14/91 | 24 | 0 | 4 |
| 4 | 0 | 2 | 1 | 1/21/93 | 23 | 0 | 8 |
| 5 | 0 | 1 | 2 | 1/22/73 | 43 | 0 | 1 |
| 6 | 0 | 1 | 3 | 8/22/71 | 44 | 0 | 1 |
| 7 | 0 | 3 | 1 | 7/23/74 | 41 | 0 | 6 |
| 8 | 0 | 1 | 2 | 2/25/73 | 43 | 0 | 4 |
| 9 | 0 | 3 | 1 | 6/10/94 | 21 | 0 | 3 |
| 10 | 0 | 3 | 1 | 6/1/88 | 27 | 0 | 4 |
| 11 | 1 | 3 | 2 | 8/22/78 | 37 | 0 | 1 |
| 12 | 0 | 2 | 1 | 12/2/74 | 41 | 0 | 4 |
| 13 | 1 | 3 | 1 | 6/14/68 | 47 | 0 | 1 |
| 14 | 0 | 2 | 1 | 3/25/85 | 31 | 0 | 3 |
| 15 | 0 | 4 | 4 | 1/25/79 | 37 | 0 | 1 |
| 16 | 0 | 2 | 1 | 6/22/90 | 25 | 0 | 10 |
| 17 | 0 | 3 | 1 | 12/24/84 | 31 | 0 | 5 |
| 18 | 0 | 3 | 1 | 1/8/85 | 31 | 0 | 3 |
| 19 | 0 | 2 | 3 | 6/28/51 | 64 | 0 | 6 |
| 20 | 0 | 2 | 1 | 11/29/94 | 21 | 0 | 9 |
| 21 | 0 | 3 | 1 | 8/6/88 | 27 | 0 | 2 |
| 22 | 1 | 3 | 1 | 3/22/95 | 21 | 0 | 4 |
| 23 | 0 | 4 | 1 | 1/23/92 | 24 | 0 | 4 |
| 24 | 0 | 3 | 3 | 1/10/73 | 43 | 0 | 1 |
| 25 | 0 | 1 | 1 | 8/24/83 | 32 | 0 | 3 |
| 26 | 0 | 2 | 1 | 2/8/89 | 27 | 0 | 3 |
| 27 | 1 | 3 | 1 | 9/3/79 | 36 | 0 | 3 |





finding: young people are recommended pathways of lower effectiveness (high error rate)

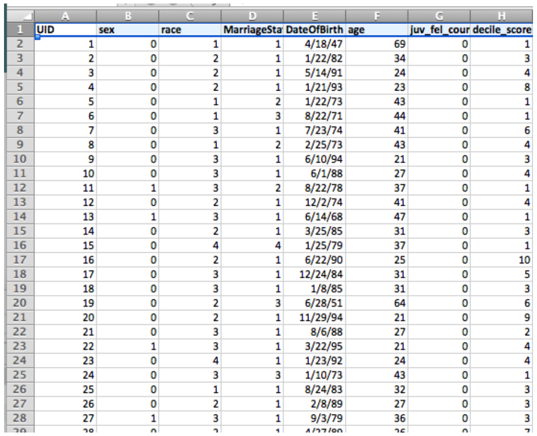
**fix the model!**

**of course, but maybe…**

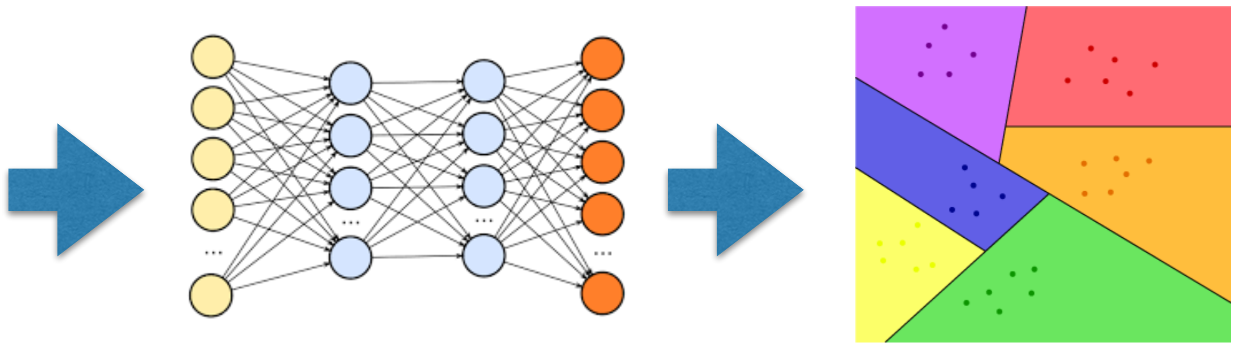
**mental health info was missing for this population**

go back to the data acquisition step, look for additional datasets

# Mitigating urban homelessness



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | UID | sex | race | MarriageSta | DateOfBirth | age | juv_fel_cour | decile_score |
| 2 | 1 | 0 | 1 | 1 | 4/18/47 | 69 | 0 | 1 |
| 3 | 2 | 0 | 2 | 1 | 1/22/82 | 34 | 0 | 3 |
| 4 | 3 | 0 | 2 | 1 | 5/14/91 | 24 | 0 | 4 |
| 5 | 4 | 0 | 2 | 1 | 1/21/93 | 23 | 0 | 8 |
| 6 | 5 | 0 | 1 | 2 | 1/22/73 | 43 | 0 | 1 |
| 7 | 6 | 0 | 1 | 3 | 8/22/71 | 44 | 0 | 1 |
| 8 | 7 | 0 | 3 | 1 | 7/23/74 | 41 | 0 | 6 |
| 9 | 8 | 0 | 1 | 2 | 2/25/73 | 43 | 0 | 4 |
| 10 | 9 | 0 | 3 | 1 | 6/10/94 | 21 | 0 | 3 |
| 11 | 10 | 0 | 3 | 1 | 6/1/88 | 27 | 0 | 4 |
| 12 | 11 | 1 | 3 | 2 | 8/22/78 | 37 | 0 | 1 |
| 13 | 12 | 0 | 2 | 1 | 12/2/74 | 41 | 0 | 4 |
| 14 | 13 | 1 | 3 | 1 | 6/14/68 | 47 | 0 | 1 |
| 15 | 14 | 0 | 2 | 1 | 3/25/85 | 31 | 0 | 3 |
| 16 | 15 | 0 | 4 | 4 | 1/25/79 | 37 | 0 | 1 |
| 17 | 16 | 0 | 2 | 1 | 6/22/90 | 25 | 0 | 10 |
| 18 | 17 | 0 | 3 | 1 | 12/24/84 | 31 | 0 | 5 |
| 19 | 18 | 0 | 3 | 1 | 1/8/85 | 31 | 0 | 3 |
| 20 | 19 | 0 | 2 | 3 | 6/28/51 | 64 | 0 | 6 |
| 21 | 20 | 0 | 2 | 1 | 11/29/94 | 21 | 0 | 9 |
| 22 | 21 | 0 | 3 | 1 | 8/6/88 | 27 | 0 | 2 |
| 23 | 22 | 1 | 3 | 1 | 3/22/95 | 21 | 0 | 4 |
| 24 | 23 | 0 | 4 | 1 | 1/23/92 | 24 | 0 | 4 |
| 25 | 24 | 0 | 3 | 3 | 1/10/73 | 43 | 0 | 1 |
| 26 | 25 | 0 | 1 | 1 | 8/24/83 | 32 | 0 | 3 |
| 27 | 26 | 0 | 2 | 1 | 2/8/89 | 27 | 0 | 3 |
| 28 | 27 | 1 | 3 | 1 | 9/3/79 | 36 | 0 | 3 |



**finding**: minors are underrepresented in the input, compared to their actual proportion in the population (insufficient data)

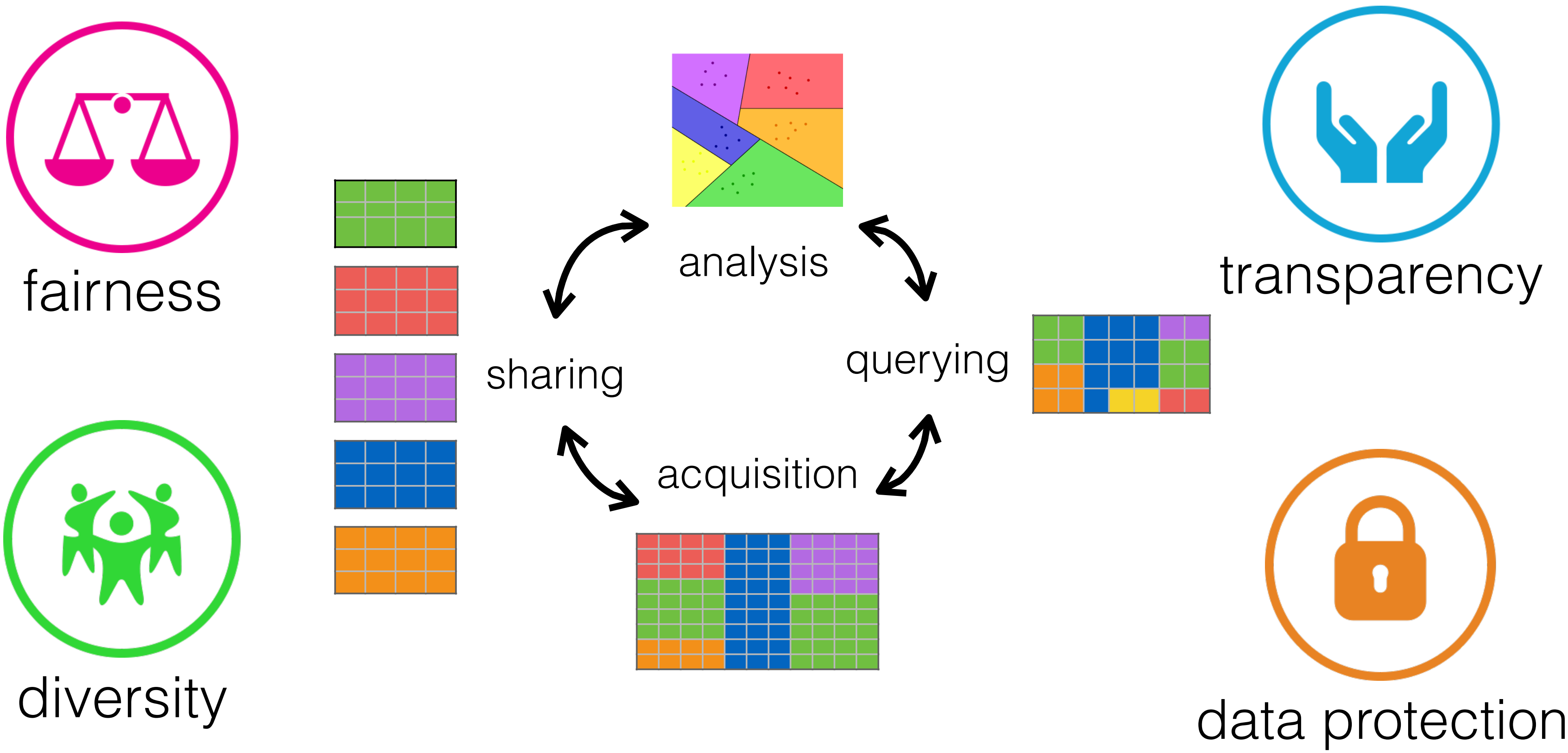
unlikely to help!

fix the model??

minors data was not shared

go back to the data sharing step, help data providers share their data while adhering to laws and upholding the trust of the participants

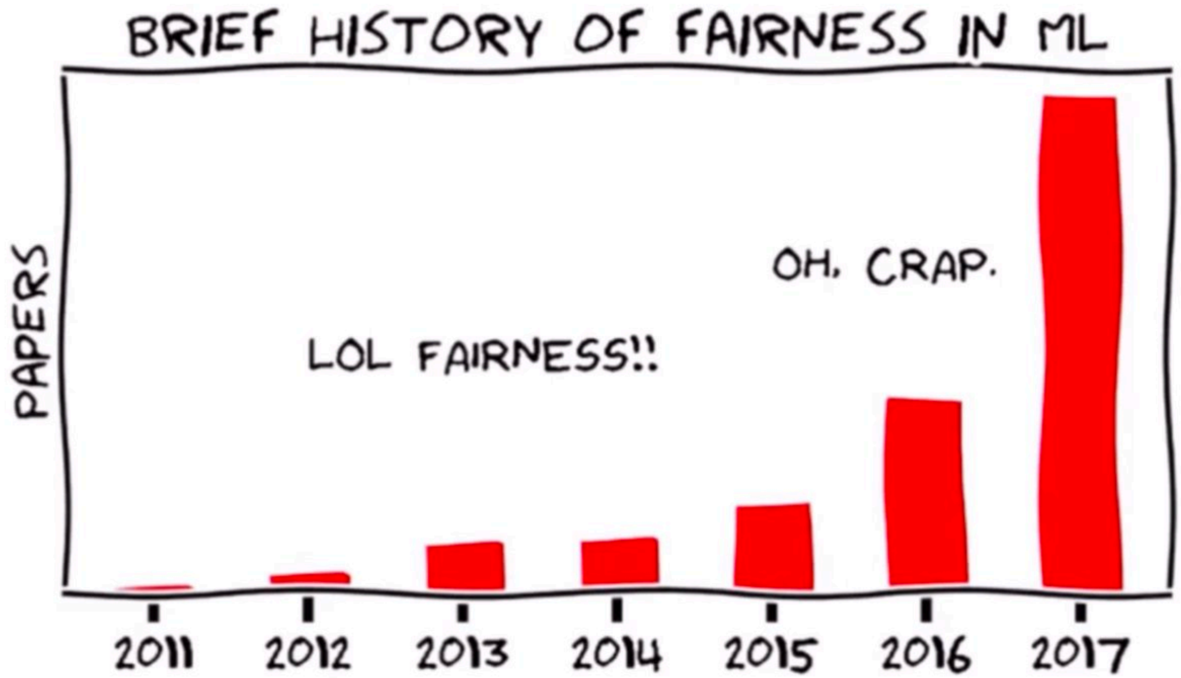
# The data science lifecycle




**responsible data science** requires a holistic view of the data lifecycle

# Fairness

# Fairness in ML




by Moritz Hardt

# Fairness is lack of "bias"

- What are the tasks we are interested in?
  - for now, let's say: predictive analytics
- What do we mean by **bias**?
  - **statistical bias**: a model is biased if it doesn't summarize the data correctly
  - **societal bias**: a dataset or a model is biased if it does not represent the world "correctly", e.g., data is not representative, there is measurement error, or the **world is "incorrect"**

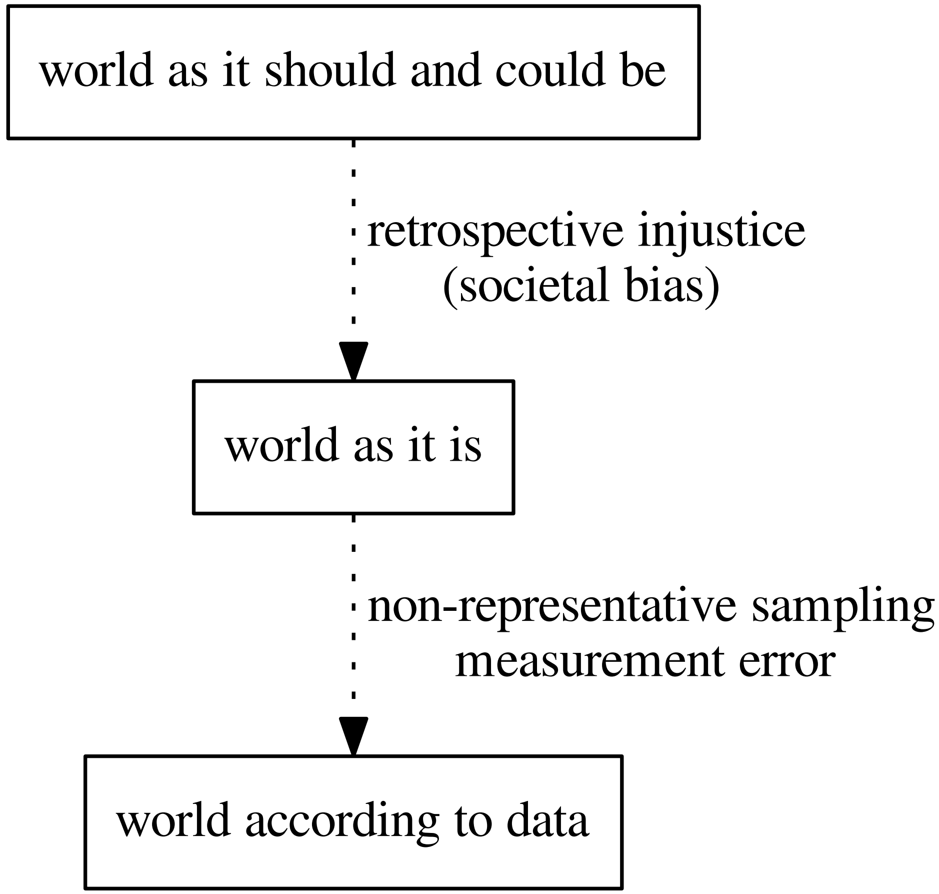
**the world as it is or as it should be?**

# More on statistical bias

- Is statistical **bias** sufficient?
  - A common view: "The model summarizes the data correctly. If the data is biased - it's not the algorithm's fault"
- But:
  - statistical bias says nothing about error distribution
  - data biases are inevitable - training data is not identical between groups - we must account for them
- **Reframing**: focus on designing systems that support human values.

**Sometimes we may decide to introduce statistical bias to correct for societal bias!**

# “Biased data”




from “Prediction-Based Decisions and Fairness” by Mitchell, Potash and Barocas, 2018

**when data is about people, bias can lead to discrimination**

# The evils of discrimination

**Disparate treatment** is the illegal practice of treating an entity, such as a creditor or employee, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

**Disparate impact** is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

# Regulated domains

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

**Credit** - Equal Credit Opportunity Act

**Education** - Civil Rights Act of 1964

**Employment** - Civil Rights Act of 1964

**Housing** - Fair Housing Act

# The 80% rule

[M/ Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian; *KDD 2015*]

DEFINITION 1.1 (DISPARATE IMPACT ("80% RULE")). *Given data set* $D = (X, Y, C)$, *with* protected attribute $X$ *(e.g., race, sex, religion, etc.), remaining attributes* $Y$, *and binary class to be predicted* $C$ *(e.g., "will hire"), we will say that* $D$ *has* disparate impact *if*

$$\frac{\Pr(C = YES \mid X = 0)}{\Pr(C = YES \mid X = 1)} \leq \tau = 0.8$$

*for positive outcome class* $YES$ *and majority protected attribute* 1 *where* $\Pr(C = c \mid X = x)$ *denotes the conditional probability (evaluated over* $D$*) that the class outcome is* $c \in C$ *given protected attribute* $x \in X$.[1]

[1]Note that under this definition disparate impact is determined based on the given data set and decision outcomes.

# Disparate impact vs. the 80% rule

- Advocated by the US Equal Employment Opportunity Commission (EEOC).
- Violating the 80% rule is not automatically illegal: Business necessity arguments can be made to excuse disparate impact
- To have disparate impact impact: violation of the rule has to be shown as **unjustified** or **avoidable**

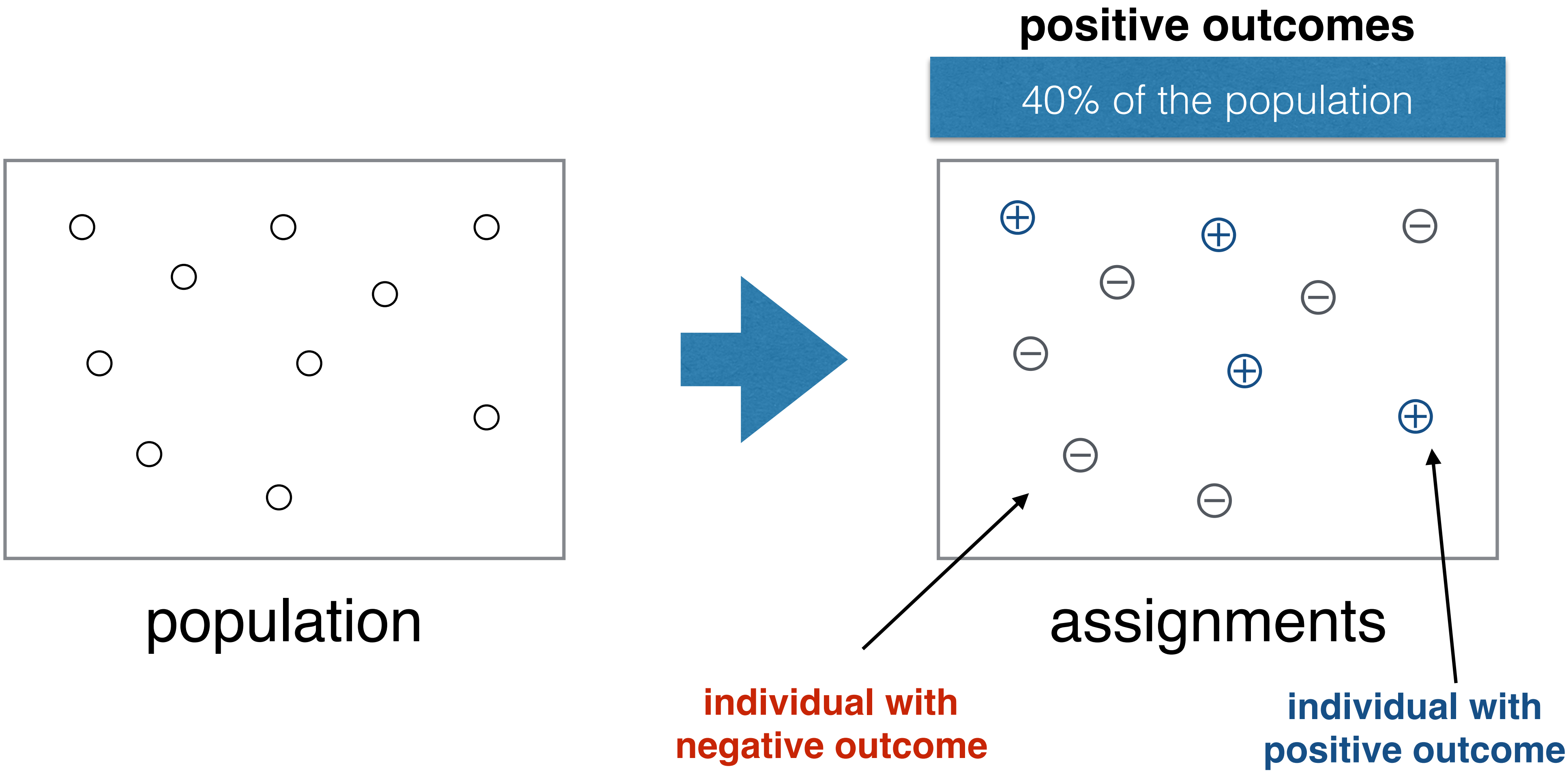
# Vendors and outcomes

Consider a **vendor** assigning positive or negative **outcomes** to individuals.

| Positive Outcomes | Negative Outcomes |
|---|---|
| offered employment | denied employment |
| accepted to school | rejected from school |
| offered a loan | denied a loan |
| offered a discount | not offered a discount |

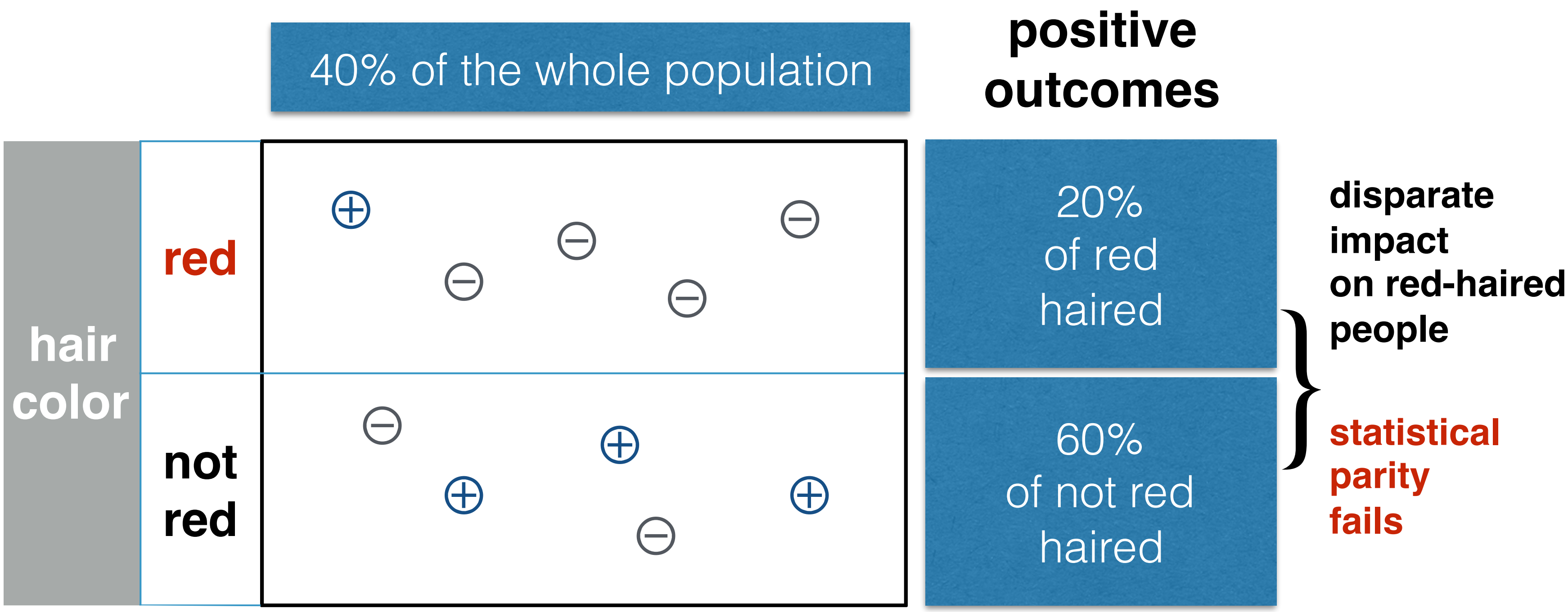
# Assigning outcomes to populations

**Fairness** is concerned with how outcomes are assigned to a population



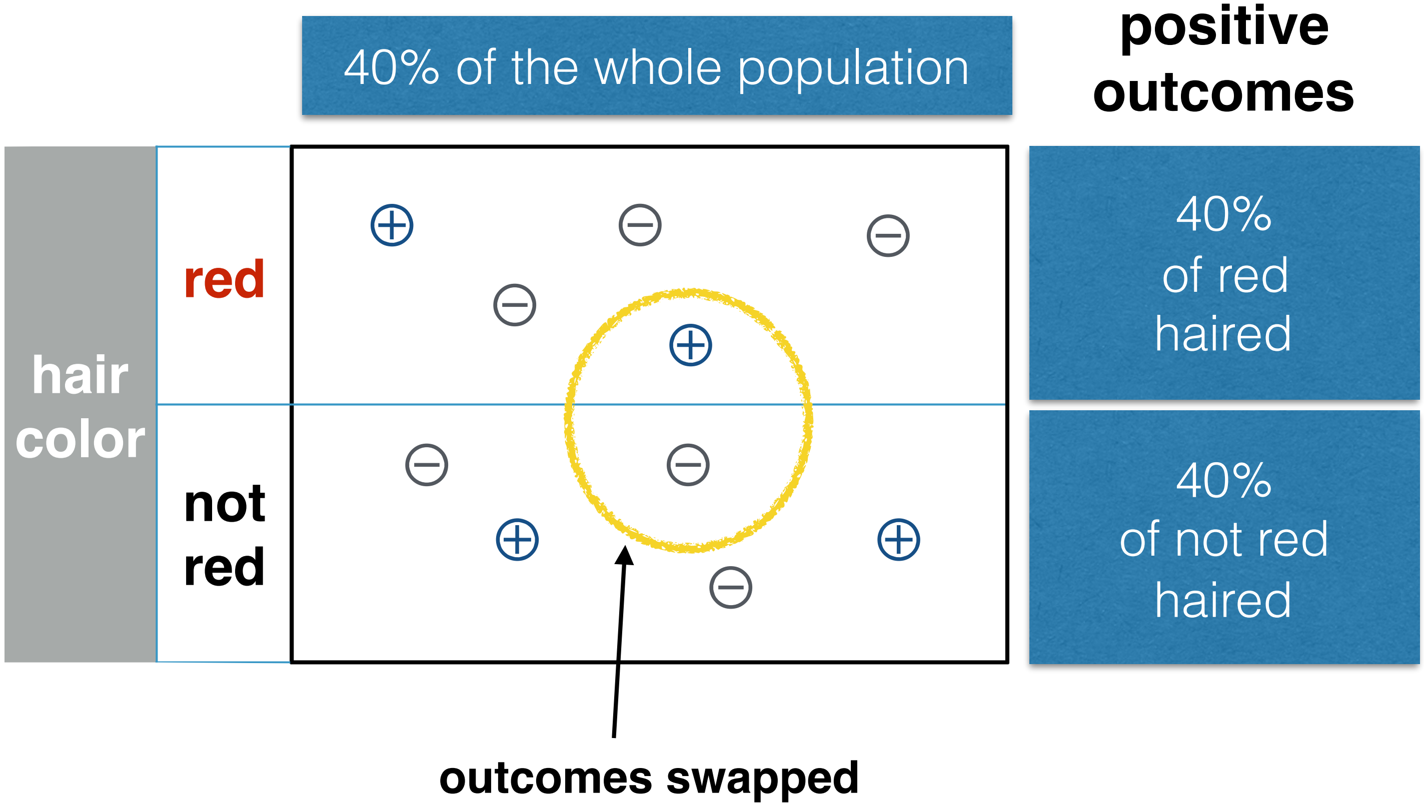
# Sub-populations may be treated differently

**Sub-population**: those with red hair
(under the same assignment of outcomes)



# Statistical parity

**Statistical parity** (a popular **group fairness** measure)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



# Redundant encoding

Now consider the assignments under both
**hair color** (protected) and **hair length** (innocuous)

| | | hair length | | |
|---|---|---|---|---|
| | | long | not long | positive outcomes |
| hair color | red | ⊕ | ⊖ ⊖ ⊖ ⊖ | 20% of red haired |
| | not red | ⊕ ⊕ ⊕ | ⊖ ⊖ | 60% of not red haired |

**Deniability**
The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

# Blinding is not an excuse

Removing **hair color** from the vendor's assignment process does not prevent discrimination!

| | | hair length | |
|---|---|---|---|
| | | **long** | **not long** |
| **hair color** | **red** | ⊕ | ⊖ ⊖ ⊖ ⊖ |
| | **not red** | ⊕ ⊕ ⊕ | ⊖ ⊖ |

**positive outcomes**

20% of red haired

60% of not red haired

**Assessing disparate impact**

Discrimination is assessed by the <u>effect</u> on the protected sub-population, not by the input or by the process that lead to the effect.

# Redundant encoding

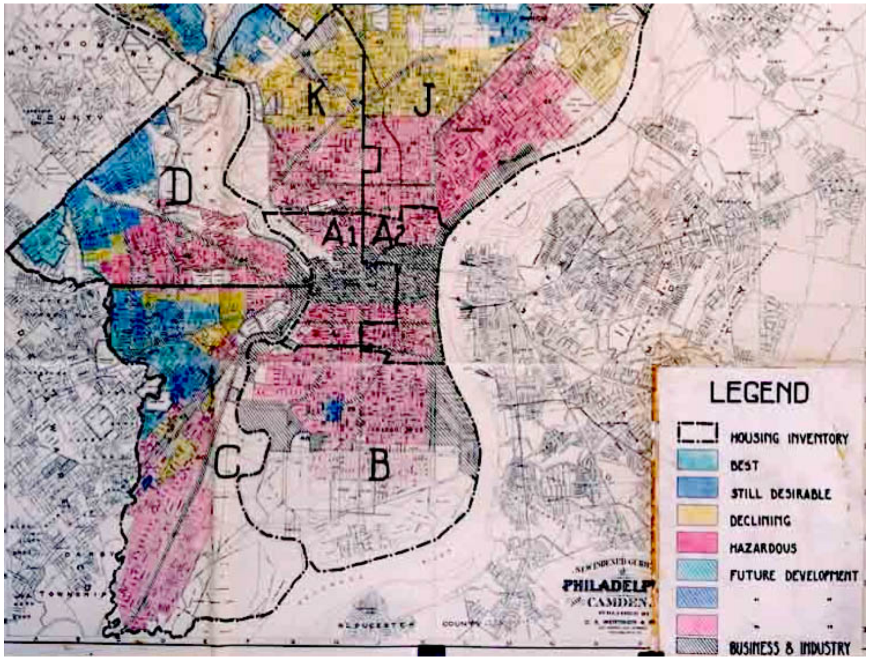
Let's replace hair color with **race** (protected),
hair length with **zip code** (innocuous)

| | | zip code | | positive outcomes |
|---|---|---|---|---|
| | | **10025** | **10027** | |
| **race** | **black** | ⊕ | ⊖ ⊖ ⊖ ⊖ | 20% of black |
| | **white** | ⊕ ⊕ ⊕ | ⊖ ⊖ | 60% of white |

# Redlining

**Redlining** is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Philadelphia, 1936



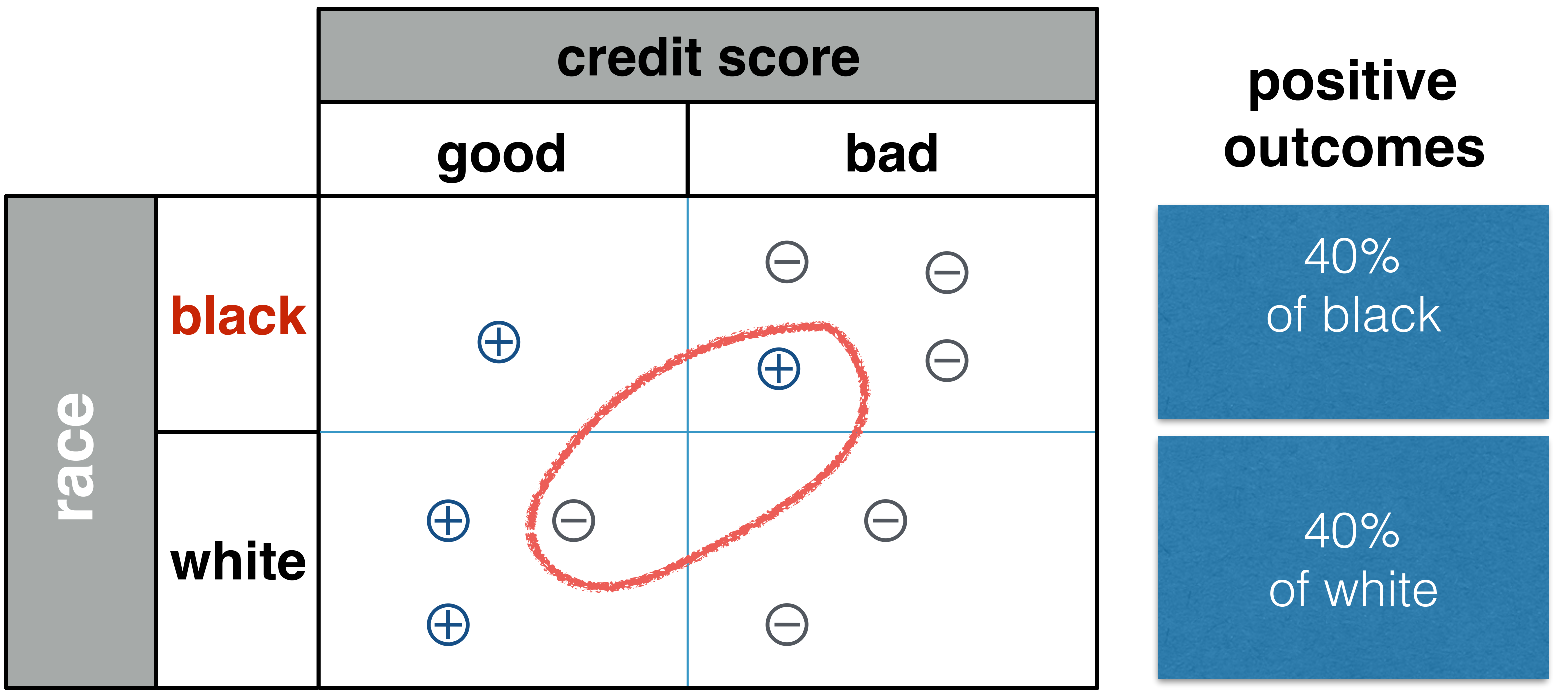

wikipedia

Households and businesses in the red zones could not get mortgages or business loans.

# Imposing statistical parity

May be contrary to the goals of the vendor

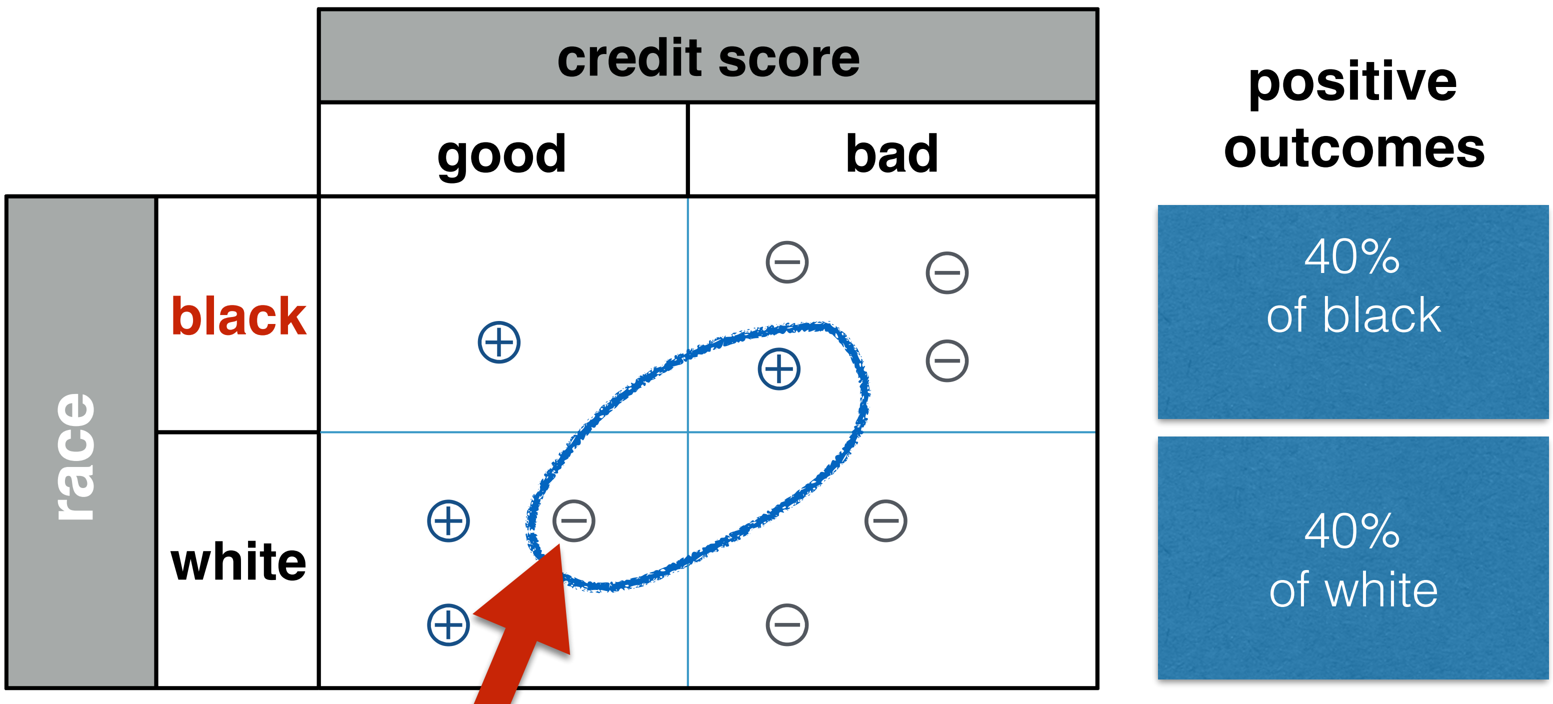
**positive outcome: offered a loan**




Impossible to predict loan payback accurately.
Use past information, which may itself be biased.

# Is statistical parity sufficient?

**Statistical parity** (a popular **group fairness** measure)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population




**Individual fairness**
any two individuals who are similar w.r.t. a particular task should receive similar outcomes

# Justifying exclusion

**Self-fulfilling prophecy**
deliberately choosing the “wrong” (lesser qualified) members of the protected group to build bad track record

| race | credit score: good | credit score: bad | positive outcomes |
|---|---|---|---|
| black | ⊕ | ⊖ ⊕ ⊖ ⊖ | 40% of black |
| white | ⊕ ⊖ ⊕ | ⊖ ⊖ | 40% of white |

# Effect on sub-populations

**Simpson's paradox**
disparate impact at the full population level disappears or reverses when looking at sub-populations!

| gender | | grad school admissions: admitted | grad school admissions: denied | positive outcomes |
|---|---|---|---|---|
| | F | 1512 | 2809 | 35% of women |
| | M | 3715 | 4727 | 44% of men |

UC Berkeley 1973: it appears men were admitted at higher rate.

# Effect on sub-populations

**Simpson's paradox**
disparate impact at the full population level disappears or reverses when looking at sub-populations!

| Department | Men | | Women | | favored group |
|---|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted | |
| **A** | 825 | 62% | 108 | **82%** | women |
| **B** | 560 | 63% | 25 | **68%** | women |
| **C** | 325 | **37%** | 593 | 34% | men |
| **D** | 417 | 33% | 375 | **35%** | women |
| **E** | 191 | **28%** | 393 | 24% | men |
| **F** | 373 | 6% | 341 | **7%** | women |

**whole population**

35% of women

44% of men

UC Berkeley 1973: women applied to more competitive departments, with low rates of admission among qualified applicants.

# A word of caution: Observational data

**Correlation is not causation!**

Cannot claim a causal relationship based on observational data alone. Need a story.

# Discrimination-aware data analysis

- **Detecting discrimination**
  - mining for discriminatory patterns in (input) data
  - verifying data-driven applications
- **Preventing discrimination**
  - data pre-processing
  - model post-processing
  - model regularization
  - data post-processing

[Ruggieri *et al.*; 2010]
[Luong *et al.*; 2011]
[Pedresci *et al.*; 2012]
[Romei *et al.*; 2012]
[Hajian & Domingo-Ferrer; 2013]
[Mancuhan & Clifton; 2014]
[Kamiran & Calders; 2009]
[Kamishima *et al.*; 2011]
[Mancuhan & Clifton; 2014]
[Feldman *et al.*; 2015]
[Dwork *et al.*; 2012]
[Zemel *et al.*; 2013]

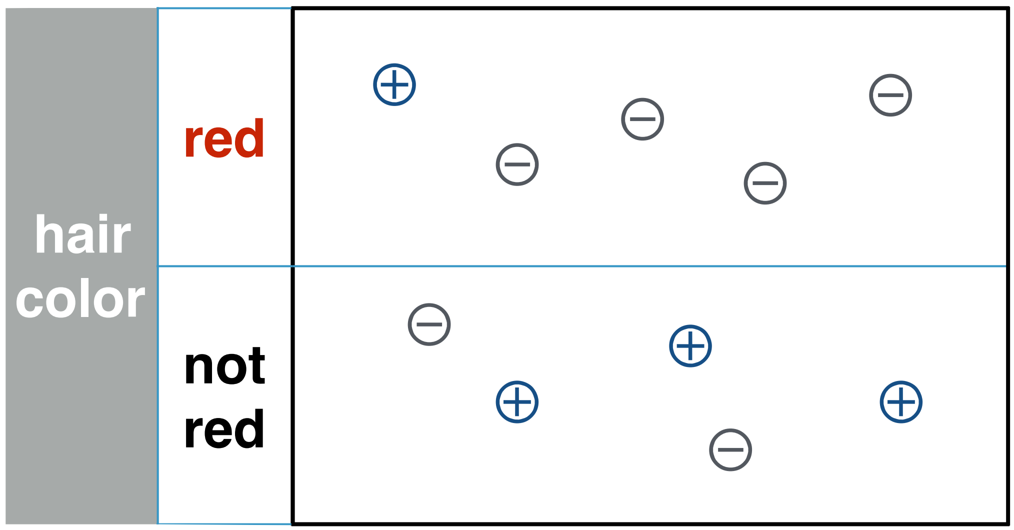
many more….

**both rely on discrimination criteria**

# Quantifying discrimination

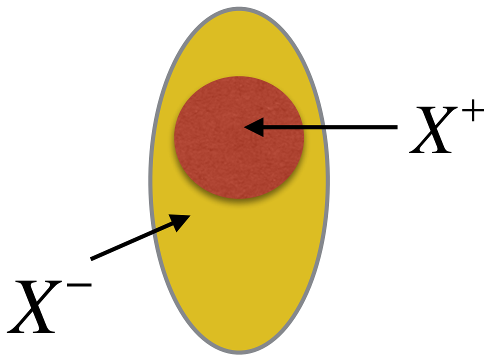
40% of the whole population

**positive outcomes** $Y=1$




20% of red hair $Y=1 \mid X^{+}$

60% of not red hair $Y=1 \mid X^{-}$




discrete (binary) protected feature *S*

$X^{+}$ are members of $X$ with $S=1$
$X^{-}$ are members of $X$ with $S=0$

# Discrimination criteria

[I. Zliobaite, Data Mining & Knowledge Discovery (2017)]

- **Statistical tests** check how likely the difference between groups is due to chance - is there discrimination?
- **Absolute measures** express the absolute difference between groups, quantifying the magnitude of discrimination
- **Conditional measures** express how much of the difference between groups cannot be explained by other attributes, while also quantifying the magnitude of discrimination
- **Structural measures** how wide-spread is discrimination? Measures the number of individuals impacted by direct discrimination.

# Discrimination measures

[I. Zliobaite, Data Mining & Knowledge Discovery (2017)]

**a proliferation of task-specific measures**

Table III. Summary of absolute measures. Checkmark (✓) indicates that it is directly applicable in a given machine learning setting. Tilde (∼) indicates that a straightforward extension exists (for instance, measuring pairwise).

| | Protected variable | | | Target variable | | |
|---|---|---|---|---|---|---|
| Measure | Binary | Categoric | Numeric | Binary | Ordinal | Numeric |
| Mean difference | ✓ | ∼ | | ✓ | | ✓ |
| Normalized difference | ✓ | ∼ | | ✓ | | |
| Area under curve | ✓ | ∼ | | ✓ | ✓ | ✓ |
| Impact ratio | ✓ | ∼ | | ✓ | | |
| Elift ratio | ✓ | ∼ | | ✓ | | |
| Odds ratio | ✓ | ∼ | | ✓ | | |
| Mutual information | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Balanced residuals | ✓ | ∼ | | ∼ | ✓ | ✓ |
| Correlation | ✓ | | ✓ | ✓ | | ✓ |

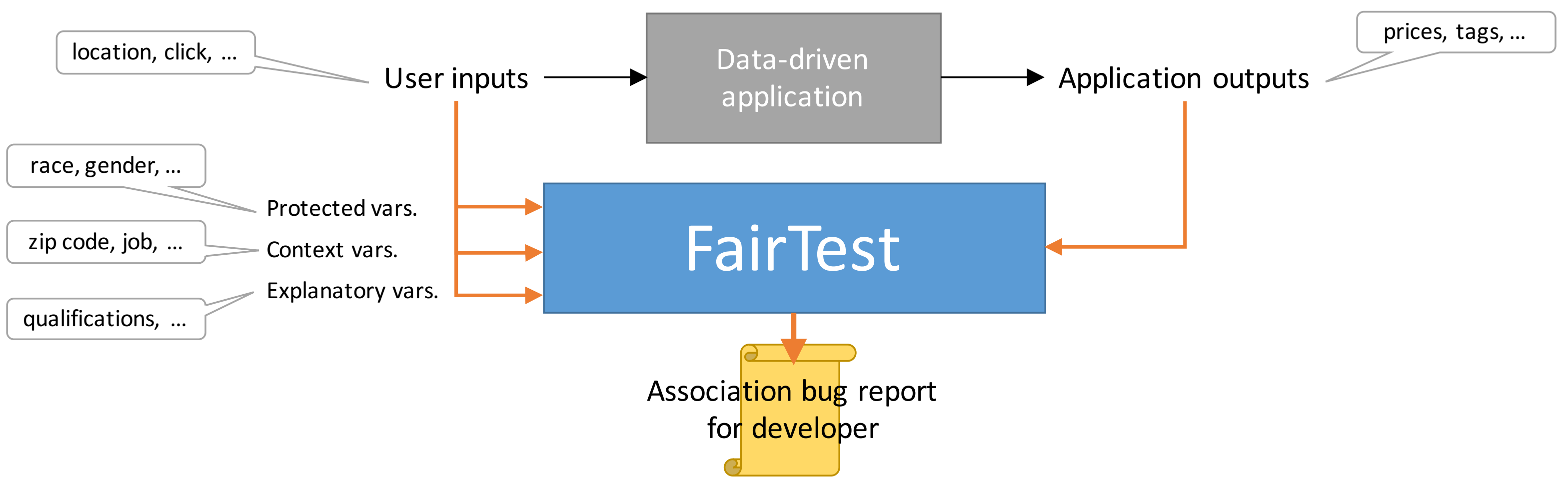
used for statistical parity: $\dfrac{\text{\% of + for protected class}}{\text{\% of + for population}}$

# FairTest: identifying discrimination

[F. Tramèr *et al.*, arXiv:1510.02377 (2015)]

A test suite for data analysis applications

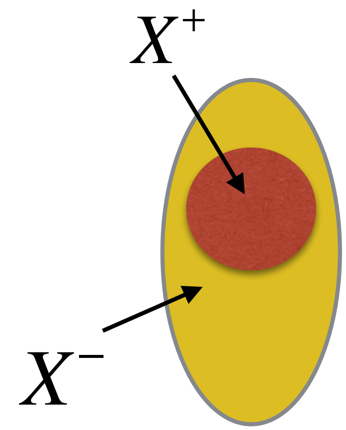
- Tests for **unintentional discrimination** according to several representative discrimination measures.
- Automates search for **context-specific associations** between protected variables and application outputs
- Report findings, ranked by association **strength** and affected **population size**




http://www.cs.columbia.edu/~djhsu/papers/fairtest-privacycon.pdf

# FairTest: discrimination measures

[F. Tramèr *et al.*, arXiv:1510.02377 (2015)]




**Binary ratio / difference** compares probabilities of a single output for two groups

$$\Pr(Y=1 \mid X^{+}) - \Pr(Y=1 \mid X^{-})$$

Easy to extend to non-binary outputs, not easy to overcome binary protected class membership

$$\frac{\Pr(Y=1 \mid X^{+})}{\Pr(Y=1 \mid X^{-})} - 1$$

---

**Mutual information** measures statistical dependence between outcomes and protected group membership

Works for non-binary outputs, class membership, can be normalized; bad for continuous values, does not incorporate of order among values

$$\sum \Pr(y,s) \ln \frac{\Pr(y,s)}{\Pr(y)\Pr(s)}$$

---

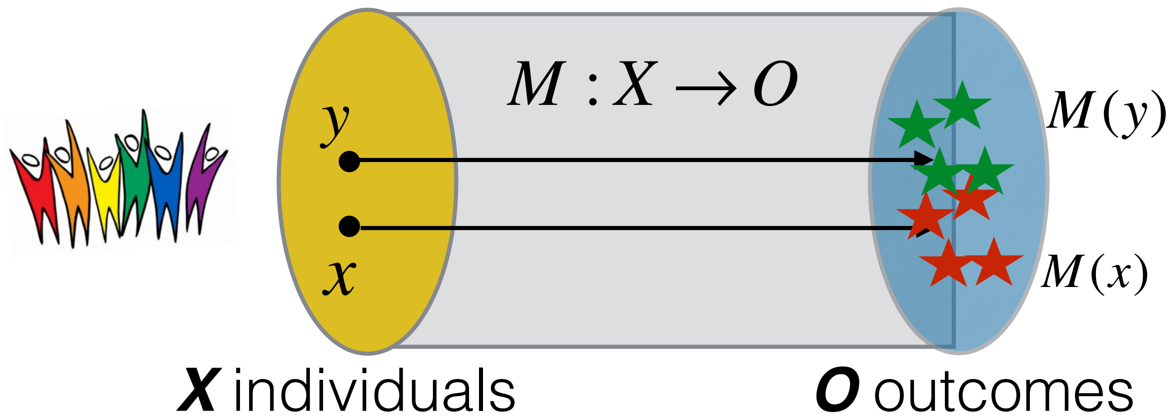
**Pearson's correlation** measures strength of linear relationship between outcomes and protected group membership

Works well for ordinal and continuous values, may detect non-linear correlations, is easy to interpret; finding a 0 correlation does not imply that S and Y are independent

# Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

**Fairness:** Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



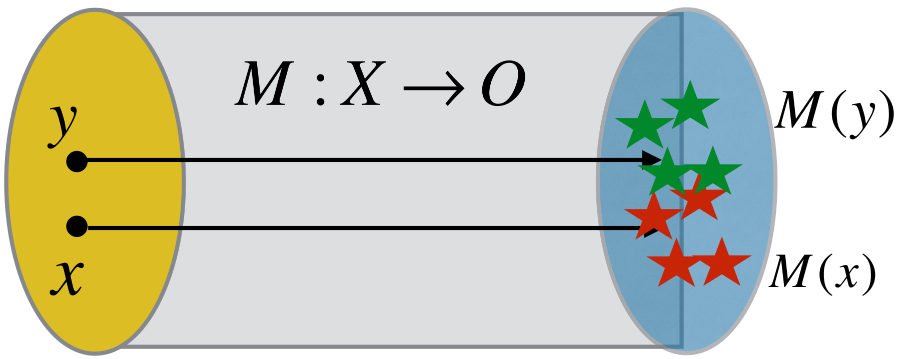

A task-specific similarity metric is given $d(x,y)$

$M : X \to O$ is a **randomized mapping**: an individual is mapped to a distribution over outcomes

# Fairness through a Lipschitz mapping

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Individuals who are **similar** for the purpose of classification task should be **treated similarly**.




***X*** individuals ***O*** intermediate mapping

A task-specific similarity metric is given $d(x,y)$

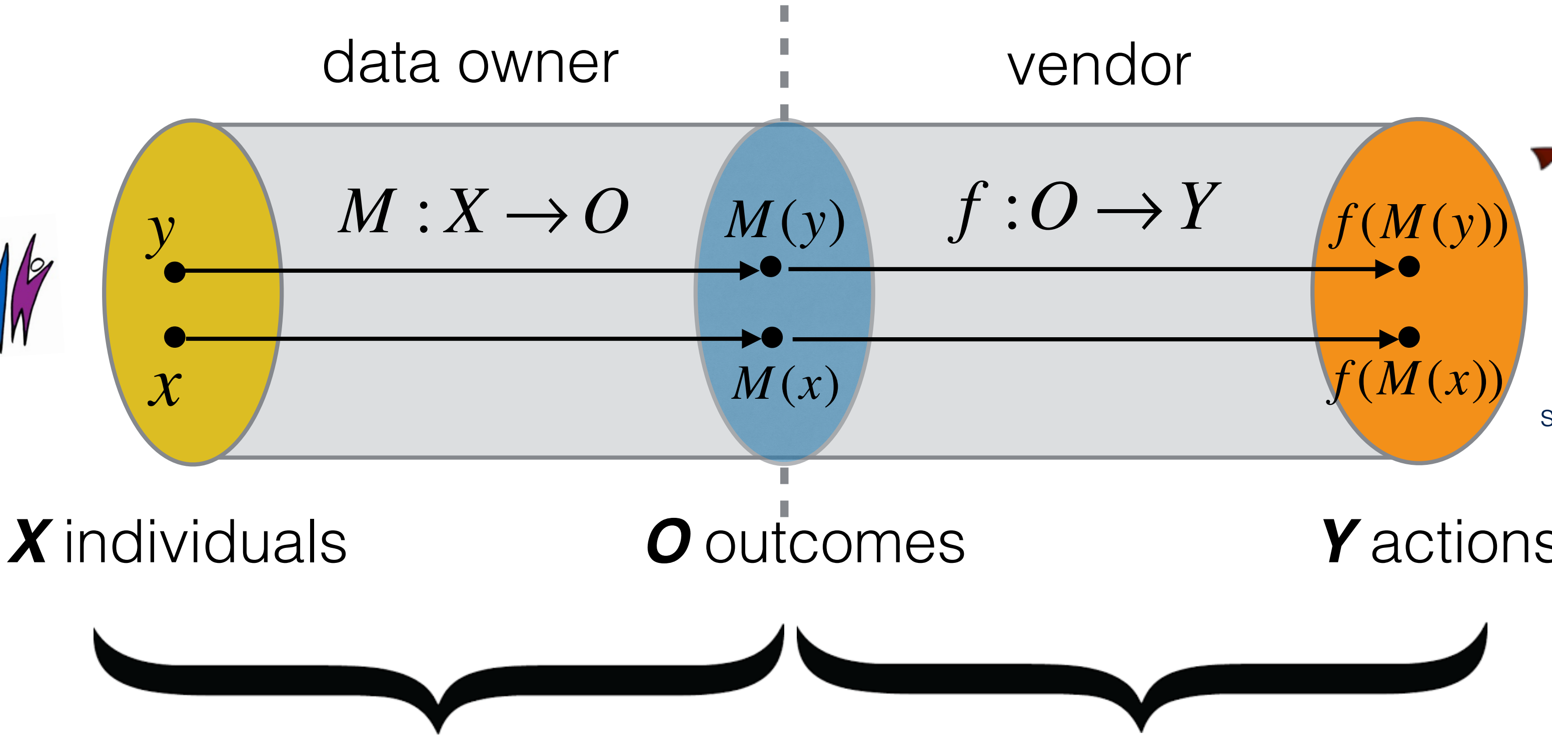
***M*** is a Lipschitz mapping if $\forall x,y \in X \quad \|M(x),M(y)\| \leq d(x,y)$

**close individuals map to close distributions**

**there always exists a Lipschitz mapping - which?**

# Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

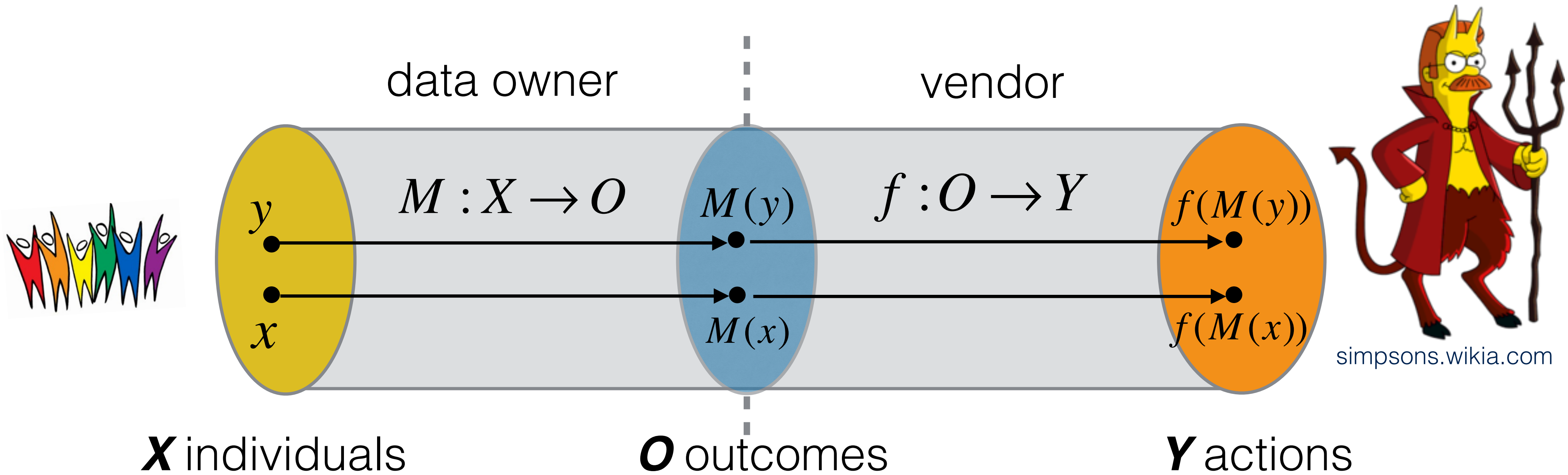



simpsons.wikia.com

**fairness enforced at this step** **vendor cannot introduce bias**

# Objective of a data owner

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

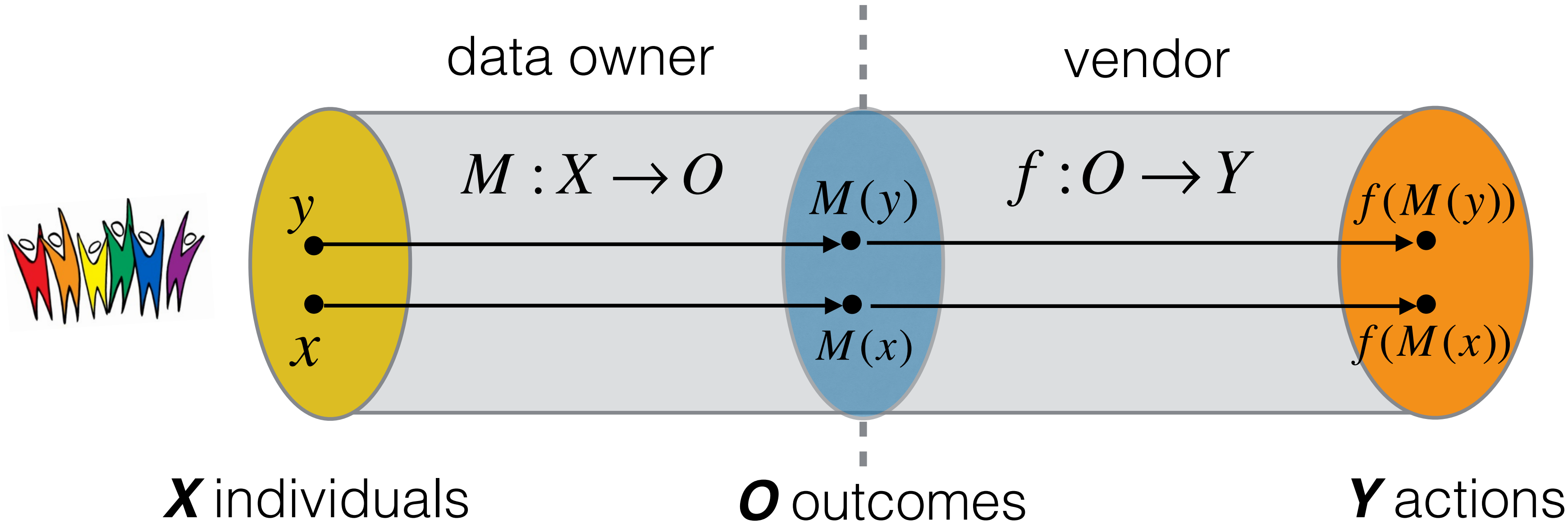



Find a mapping from individuals to distributions over outcomes that minimizes expected loss, **subject to the Lipschitz condition**. Optimization problem: minimize an arbitrary loss function.

# What about the vendor?

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Vendors can efficiently maximize expected utility, subject to the Lipschitz condition




Computed with a linear program of size $poly(|X|,|Y|)$

**the same mapping can be used by multiple vendors**

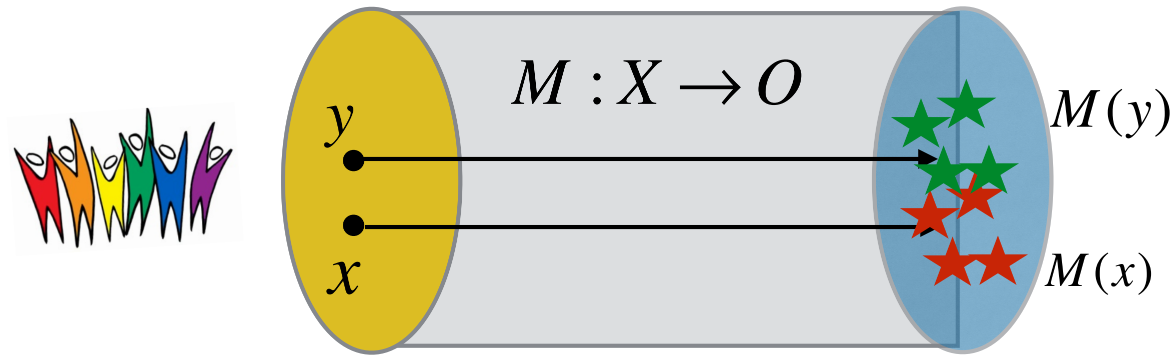
# Some philosophical background

[C. Calsamiglia; *PhD thesis 2005*]

"**Equality of opportunity** defines an important welfare criterion in political philosophy and policy analysis. **Philosophers** define equality of opportunity as the requirement that an individual's well being be independent of his or her irrelevant characteristics. **The difference among philosophers is mainly about which characteristics should be considered irrelevant.** **Policymakers**, however, are often called upon to address more specific questions: How should admissions policies be designed so as to provide equal opportunities for college? Or how should tax schemes be designed so as to equalize opportunities for income? These are called local distributive justice problems, because each policymaker is in charge of achieving equality of opportunity to a specific issue."

# Connection to privacy

Fairness through awareness generalizes differential privacy




***X*** databases ***O*** *sanitized* output

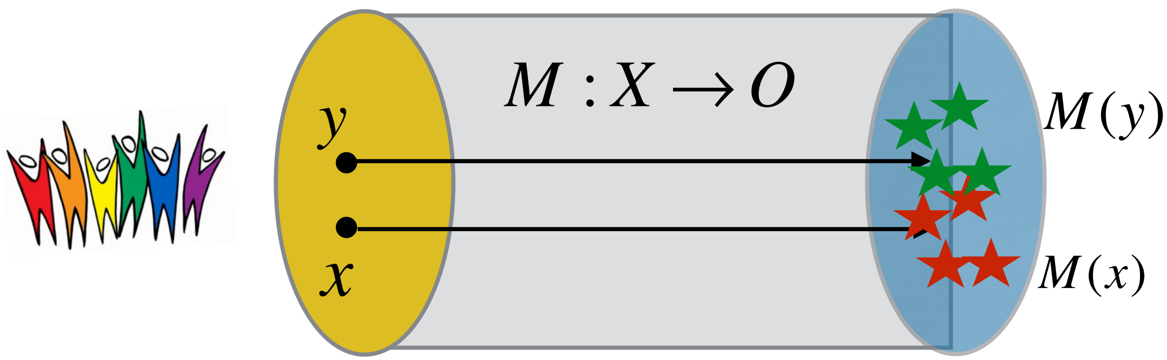
**close databases map to close output distributions**

Databases that differ in one record.

# Connection to privacy

## Does the fairness mapping provide privacy?

Similar individuals (according to $d(x,y)$) are hard to distinguish in the intermediate mapping. This provides a form of protection similar to anonymity-based privacy.




***X*** individuals ***O*** intermediate mapping

It depends on the metric *d* and on whether individual similarity is based on sensitive properties.

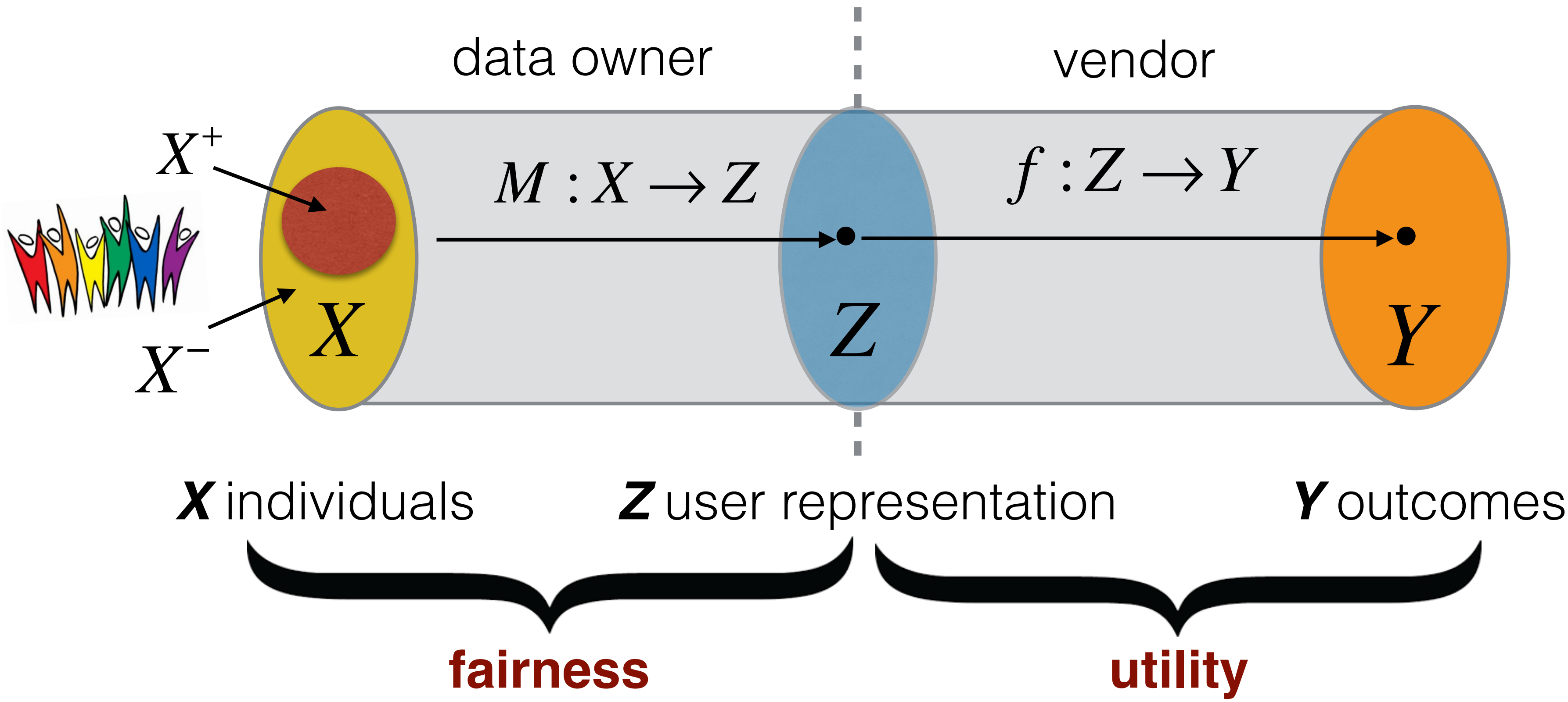
# Fairness through awareness: summary

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

- An early work in this space, proposes a principled data pre-processing approach
- Stated as an **individual fairness** condition but also sometimes leads to **group fairness**
- Relies on an externally-supplied task-specific similarity metric - magic!
- Is not formulated as a learning problem, does not generalize to unseen data

# Learning fair representations

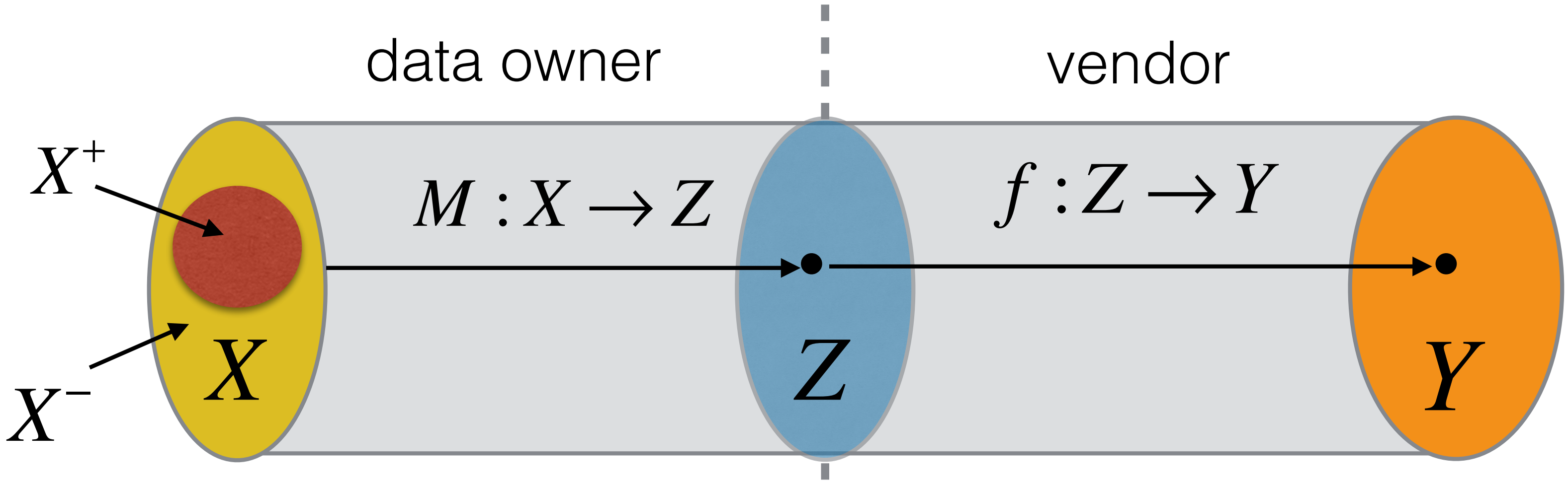
[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]




- **Idea**: remove reliance on a “fair” similarity measure, instead **learn** representations of individuals, distances

# Fairness and utility

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]




Learn a **randomized mapping** M(X) to a set of K prototypes Z

M(X) should lose information about membership in S $P(Z \mid S=0) = P(Z \mid S=1)$

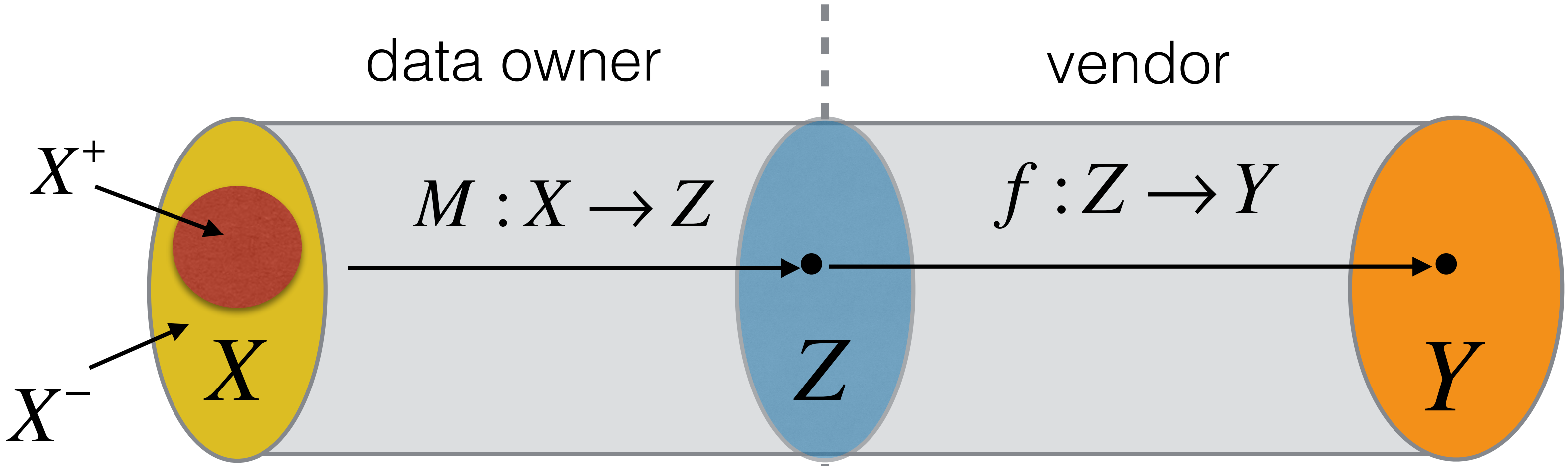
M(X) should preserve other information so that vendor can maximize utility

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

**group fairness** **individual fairness** **utility**

# The objective function

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]




$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

**group fairness** **individual fairness** **utility**

$$P_k^+ = P(Z = k \mid x \in X^+)$$

$$P_k^- = P(Z = k \mid x \in X^-)$$

$$L_z = \sum_k \left| P_k^+ - P_k^- \right| \qquad L_x = \sum_n (x_n - \widehat{x}_n)^2$$

$$L_y = \sum_n -y_n \log \widehat{y}_n - (1 - y_n) \log(1 - \widehat{y}_n)$$

# Learning fair representations: summary

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]

- A principled learning framework in the data pre-processing / classifier regularization category
- **Evaluation** of accuracy, discrimination (group fairness) and consistency (individual fairness), promising results on real datasets
- Not clear how to set $K$, so as to trade off accuracy / fairness
- The mapping is **task-specific**

# Ricci v. DeStefano (2009)

## *Supreme Court Finds Bias Against White Firefighters*

By **ADAM LIPTAK** JUNE 29, 2009

The New York Times

| Case opinions | |
|---|---|
| **Majority** | Kennedy, joined by Roberts, Scalia, Thomas, Alito |
| **Concurrence** | Scalia |
| **Concurrence** | Alito, joined by Scalia, Thomas |
| **Dissent** | Ginsburg, joined by Stevens, Souter, Breyer |
| **Laws applied** | |
| Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq. | |

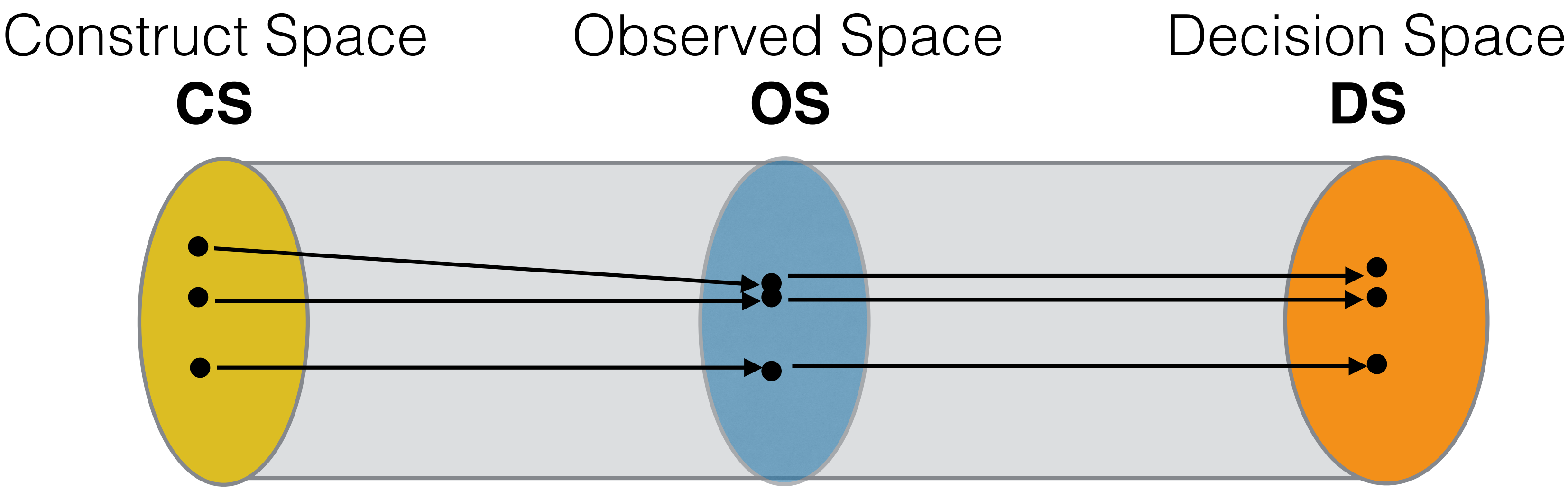
Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

# On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

**Goal**: tease out the difference between *beliefs* and *mechanisms* that logically follow from those beliefs.

**Main insight**: To study algorithmic fairness is to study the interactions between different spaces that make up the decision pipeline for a task



# Examples of features and outcomes

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

| Construct Space | Observed Space | Decision Space |
|---|---|---|
| intelligence | SAT score | performance in college |
| grit | high-school GPA | |
| propensity to commit crime | family history | recidivism |
| risk-averseness | age | |

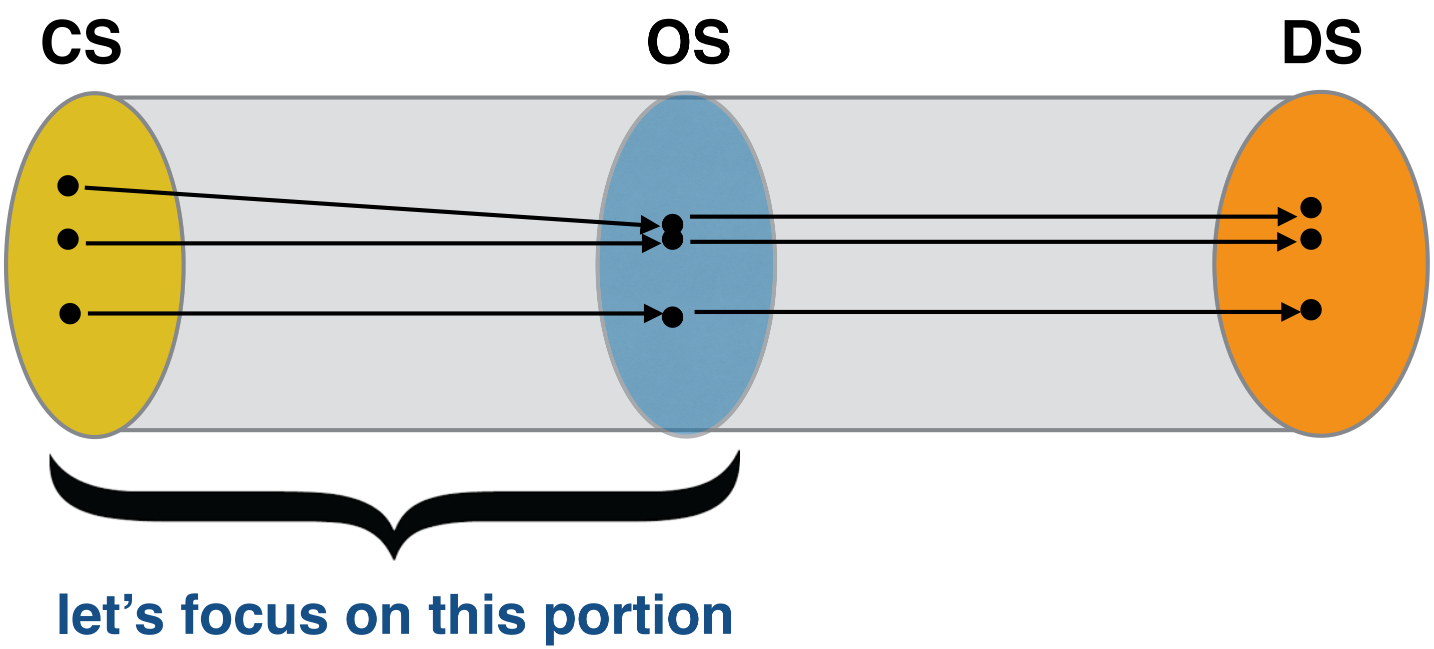
**define fairness through properties of mappings between CS, OS and DS**

# Fairness through mappings

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

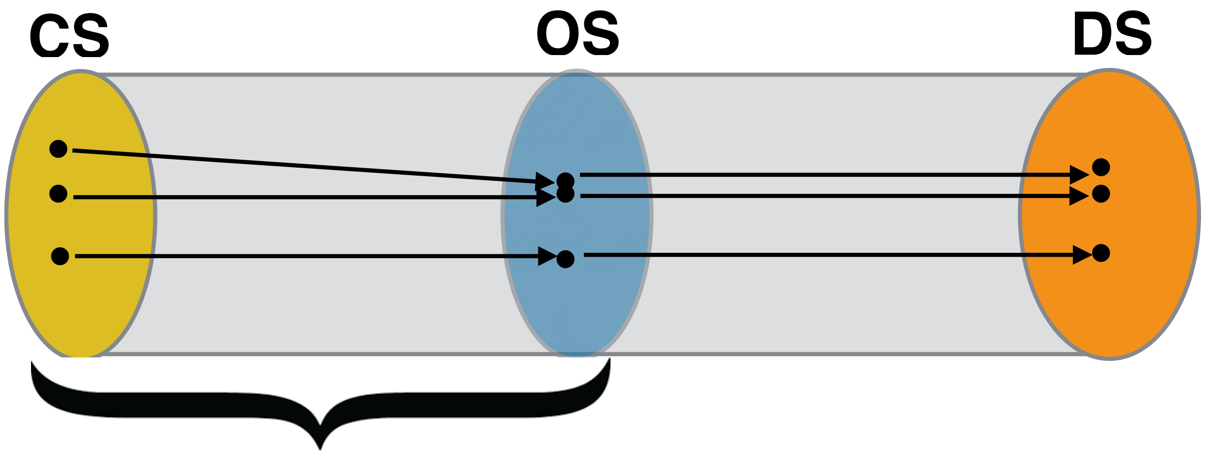
**Fairness**: a mapping from **CS** to **DS** is **(ε, ε')**-fair if two objects that are no further than **ε** in **CS** map to objects that are no further than **ε'** in **DS**.

$$f : CS \rightarrow DS \qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$



# A world view: What you see is what you get

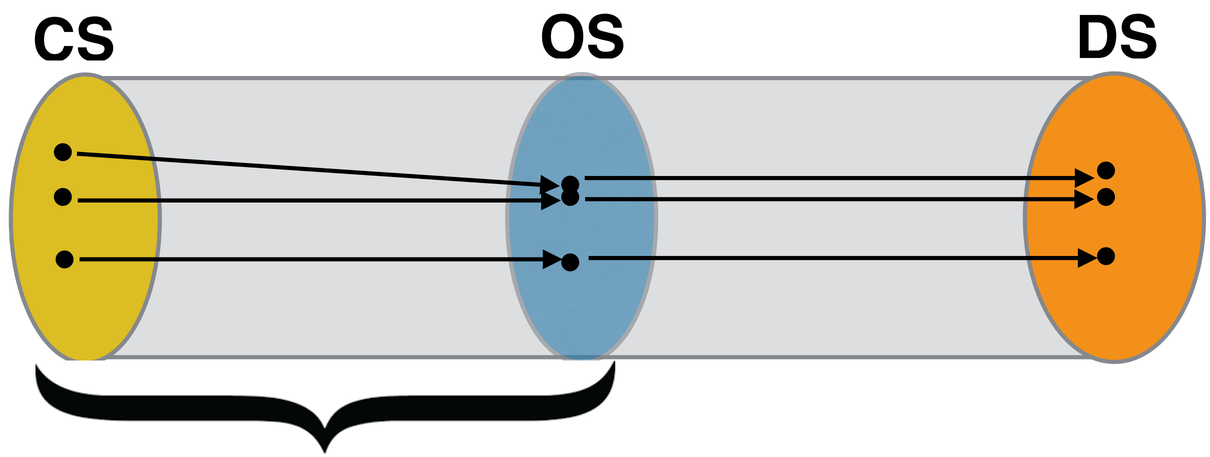
[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]




What you see is what you get (**WYSIWYG**): there exists a mapping from **CS** to **OS** that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

# A world view: Structural bias

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]




We are all equal (**WAE**): the mapping from **CS** to **OS** introduces **structural bias** - there is a distortion that aligns with the group structure of **CS**. **This is the group fairness world view.**

**Structural bias examples**: SAT verbal questions function differently in the African-American and in the Caucasian subgroups in the US. Other examples?

# A word of caution: Observational data

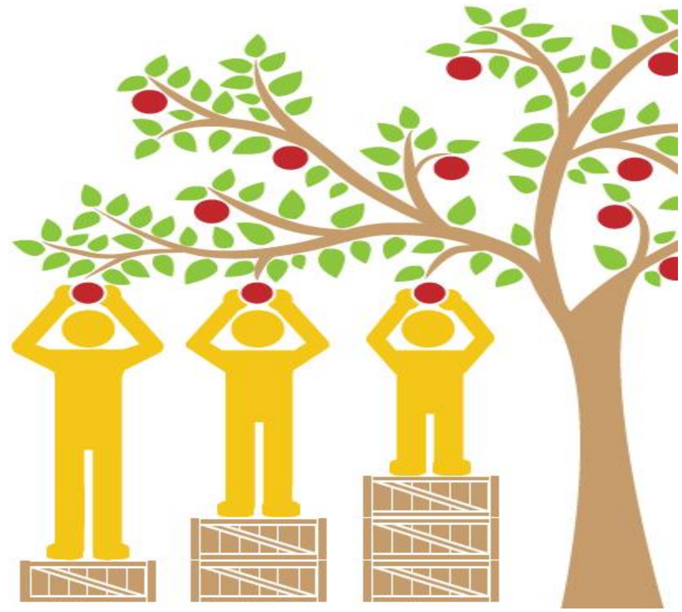
We cannot tell, based on observational data alone, whether the world is **WYSIWYG** or **WAE**

**Other examples where observational data is insufficient?**

# Two notions of fairness

individual fairness

group fairness

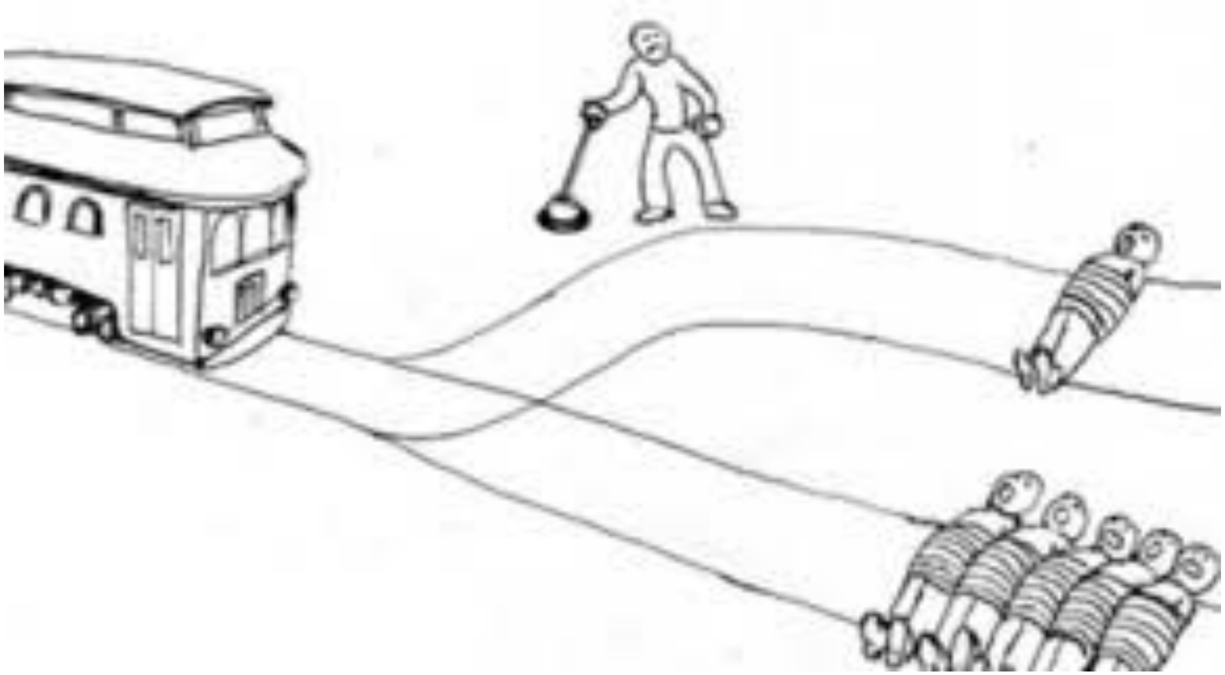


equality

equity

**two intrinsically different world views**

# Fairness definitions as “trolley problems”



# Racial bias in criminal sentencing

## Machine Bias

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

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*

May 23, 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

### Prediction Fails Differently for Black Defendants

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

***Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.*** *(Source: ProPublica analysis of data from Broward County, Fla.)*

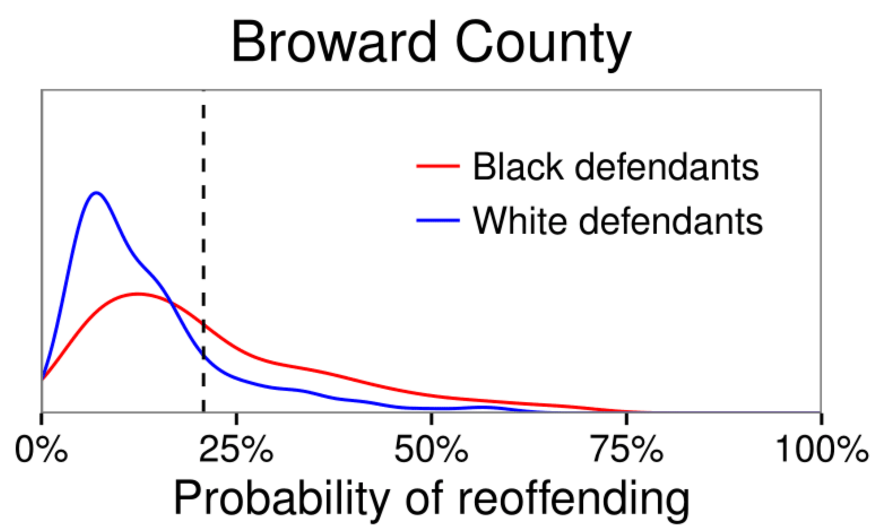
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# COMPAS as a predictive instrument

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

**Predictive parity** (also called **calibration**)
an instrument identifies a set of instances as having probability *x* of constituting positive instances, then approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is **well-calibrated**: in the window around 40%, the fraction of defendants who were re-arrested is ~40%, both over-all and per group.




[plot from Corbett-Davies et al.; *KDD 2017*]

# Group fairness impossibility result

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups

Recidivism rates in the ProPublica dataset are higher for the black group than for the white group

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

**What is recidivism?**: Northpointe [*the maker of COMPAS*] defined recidivism as "**a finger-printable arrest** involving a charge and a filing for any uniform crime reporting (UCR) code."

# Fairness for whom?

based on a slide by Arvind Narayanan

**Decision-maker**: of those I've labeled high-risk, how many will recidivate?

**Defendant**: how likely am I to be incorrectly classified high-risk?

**Society**: (think positive interventions) is the selected set demographically balanced?

| | labeled low-risk | labeled high-risk |
|---|---|---|
| did not recidivate | TN | **FP** |
| recidivated | **FN** | TP |

**different metrics matter to different stakeholders**

https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story

# Impossibility theorem

based on a slide by Arvind Narayanan

| Metric | Equalized under |
|---|---|
| Selection probability | Demographic parity |
| Pos. predictive value | Predictive parity |
| Neg. predictive value | |
| False positive rate | Error rate balance |
| False negative rate | Error rate balance |
| Accuracy | Accuracy equity |

Chouldechova paper

All these metrics can be expressed in terms of FP, FN, TP, TN

If these metrics are equal for 2 groups, some trivial algebra shows that the prevalence (in the COMPAS example, of recidivism, as measured by re-arrest) is also the same for 2 groups

**Nothing special about these metrics, can pick any 3!**

# Ways to evaluate binary classifiers

based on a slide by Arvind Narayanan

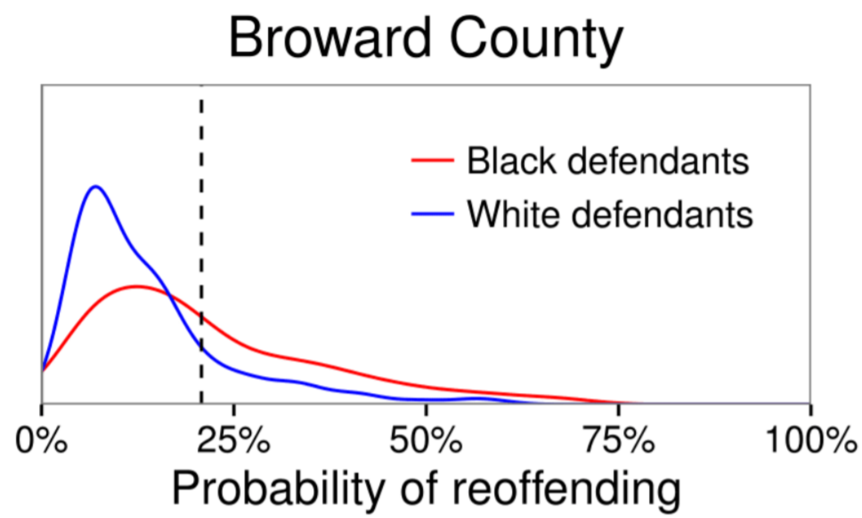
| | | | | | | |
|---|---|---|---|---|---|---|
| | | **True condition** | | | | |
| | Total population | Condition positive | Condition negative | Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$ | Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$ | |
| **Predicted condition** | Predicted condition positive | **True positive**, Power | **False positive**, Type I error | Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$ | False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$ | |
| | Predicted condition negative | **False negative**, Type II error | **True negative** | False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$ | Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$ | |
| | | True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$ | False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$ | Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$ | Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ | $F_1$ score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$ |
| | | False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$ | True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$ | Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$ | | |

**364 impossibility theorems :)**

# Individual fairness

based slides by Arvind Narayanan

**Individual fairness**: assuming scores are calibrated, we cannot pick a single threshold for 2 groups that equalizes both the False Positives Rate and the False Negatives Rate



# What's the right answer?

**There is no single answer!**

**Need transparency and public debate**

- Consider harms and benefits to different stakeholders
- Being transparent about which fairness criteria we use, how we trade them off
- Recall "Learning Fair Representations": a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

**group fairness** **individual fairness** **utility**

**apples + oranges + fairness = ?**

## AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ Get Code ↗

## Not sure what to do first? Start here!

### Read More

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

→

### Try a Web Demo

Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

→

### Watch a Video

Watch a video to learn more about AI Fairness 360.

→

### Read a paper

Read a paper describing how we designed AI Fairness 360.

→

### Use Tutorials

Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.

→