APPLIED MECHANISM DESIGN FOR SOCIAL GOOD

JOHN P DICKERSON & MARINA KNITTEL

Lecture #22 – 04/12/2022

CMSC498T
Mondays & Wednesdays
2:00pm – 3:15pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

ANNOUNCEMENTS

Due tonight: Fair Allocation Quiz

Due on Monday, 4/25: Project Checkup

- Kind of like the project proposal, but regarding the current state of things
- Will be graded in a similar manner
- Remember that proposal comments are up!

WHAT IS MACHINE LEARNING?

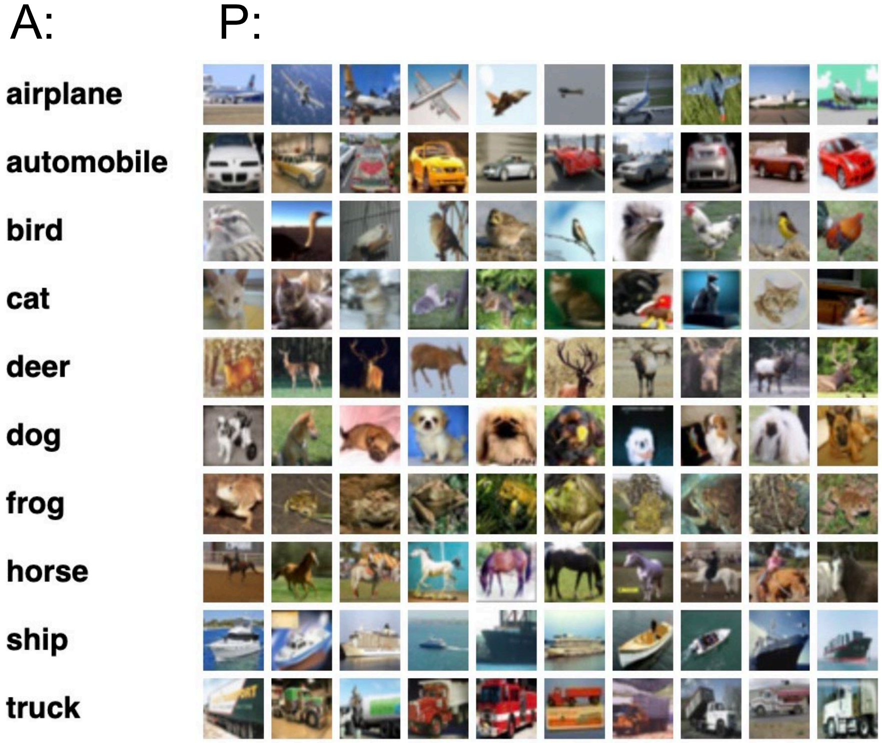
"The study of computer algorithms that can improve automatically through experience and by the use of data."

- Wikipedia

Let P be data. Let A be a set of labels.

Find a mapping M : P → A in an attempt to most accurately identify the labels.

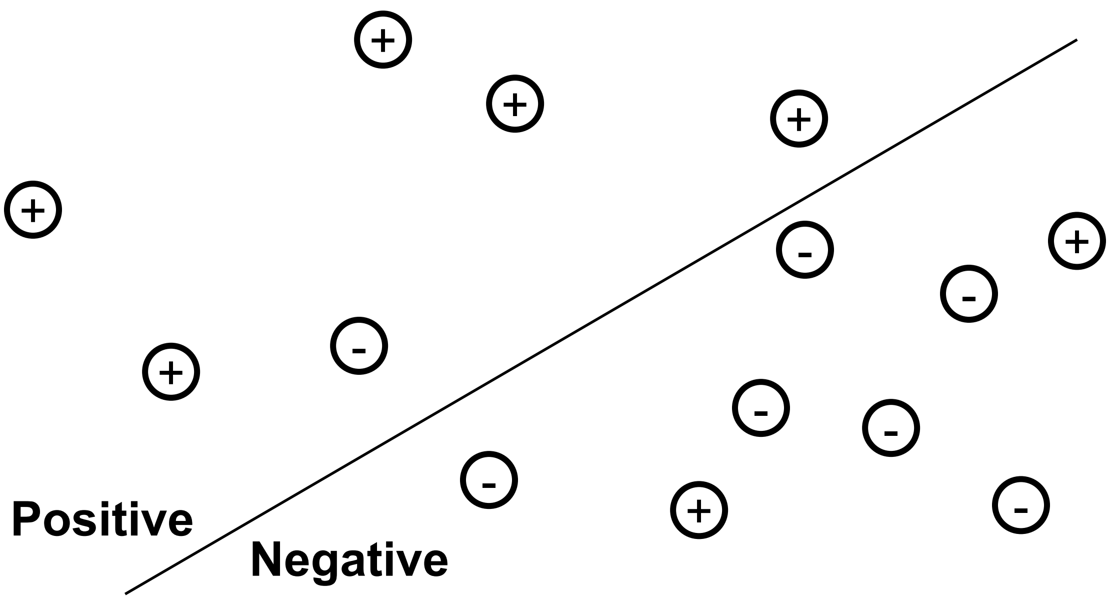
We want to estimate the label as best as possible under *constraints*.



CONSTRAINT: LINEAR SEPARATOR

A constraint is any restriction on the solution map M.

Example: M must be a linear separator.



False Positives:
1

False Negatives:
2

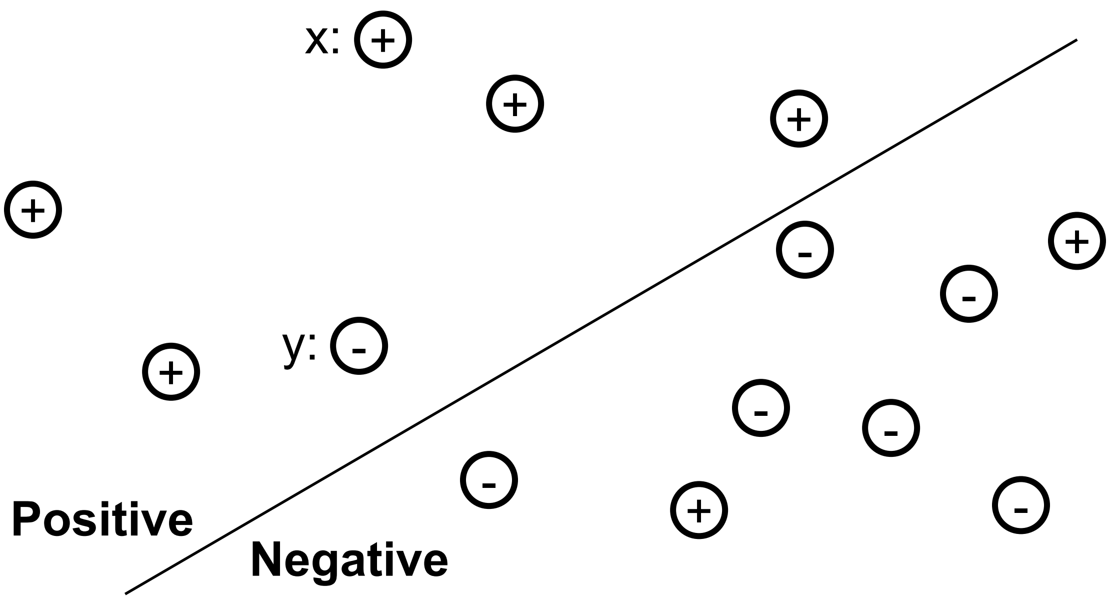
This may be the best linear separator

LOSS FUNCTION AND LINEAR PROGRAMS

Loss function: A function L : P x A → R which tells you "how far away you are from a solution.

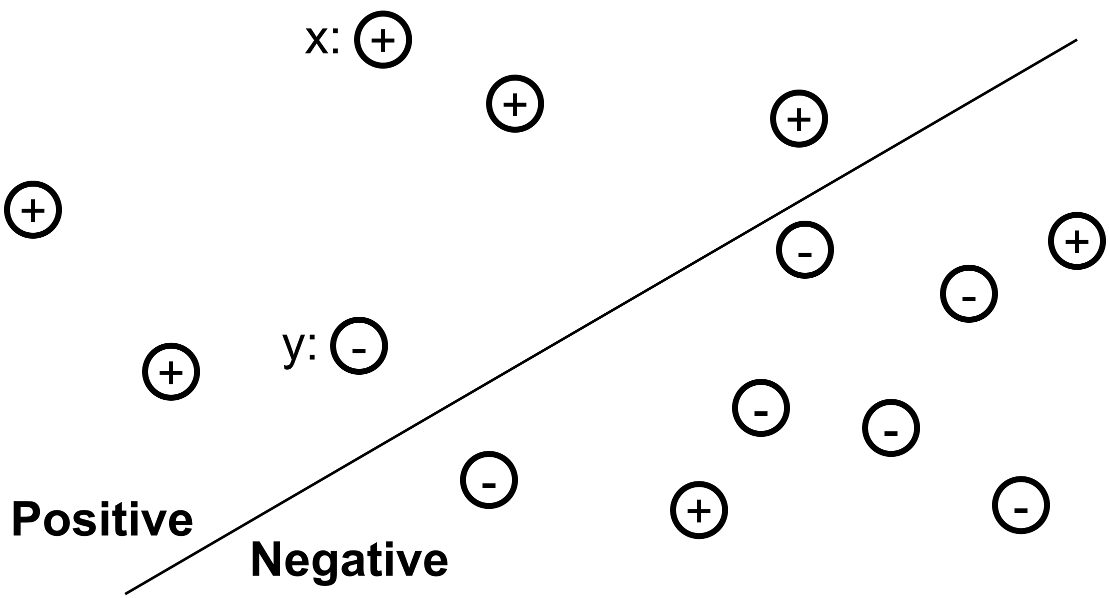
Say L(p, a) is 0 if x's label is a, 1 otherwise (0-1 loss).

L(x, +) =



LOSS FUNCTION AND LINEAR PROGRAMS

Loss function: A function L : P x A → R which tells you “how far away you are from a solution.



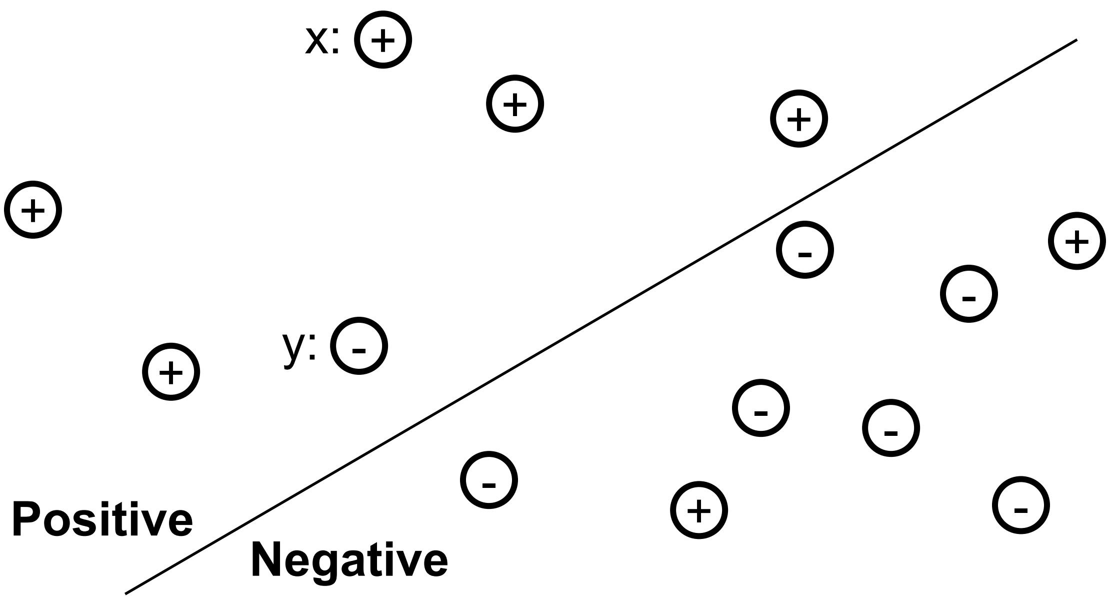
Say L(p, a) is 0 if x’s label is a, 1 otherwise (0-1 loss).

L(x, +) = 0

L(y, +) =

LOSS FUNCTION AND LINEAR PROGRAMS

Loss function: A function L : P x A → R which tells you "how far away you are from a solution.



Say L(p, a) is 0 if x's label is a, 1 otherwise (0-1 loss).

L(x, +) = 0

L(y, +) = 1

Thus y is a bad label, x is a good label.

Goal: minimize total loss.

IF IT AIN'T BROKE, DON'T FIX IT

Unfortunately, it is broken 

IF IT AIN'T BROKE, DON'T FIX IT

Unfortunately, it is broken  in lots of ways. 

But let's focus on a few.

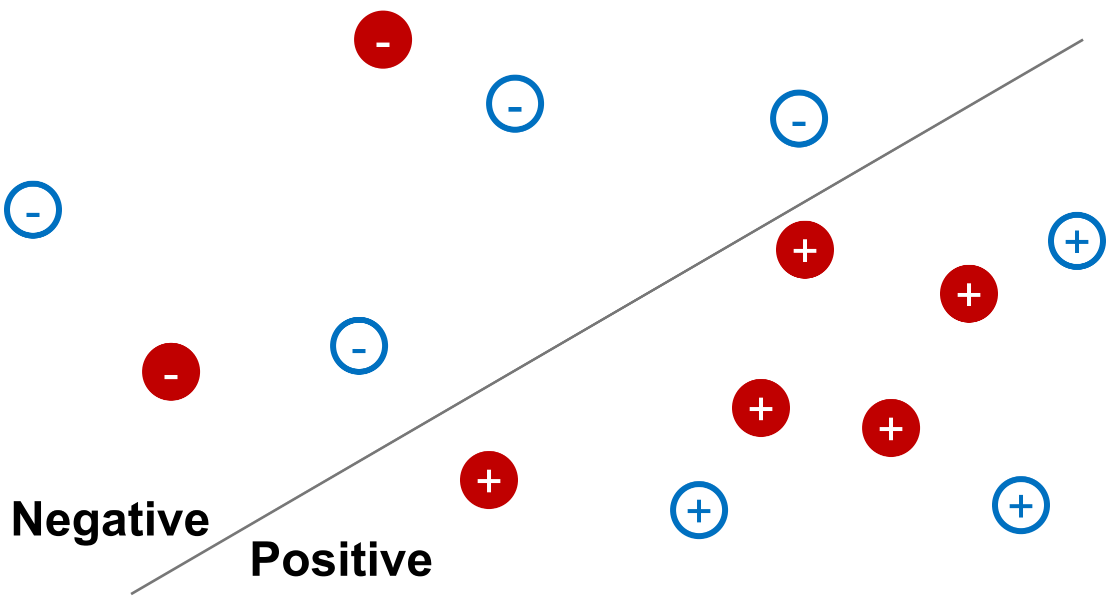
Many machine learning projects that we rely on today discriminate against real people.

- Credit card advertisements
- Google Ad selection
- Google name advertisements
- Recidivism risk
- Others: hiring decisions, school admission, etc.

GROUP FAIRNESS

"Group fairness" or "statistical parity": demographics in the positive group and negative group are the same as the whole distribution.

Say our demographics are blue and red points.



50% of the points are red, 50% are blue.

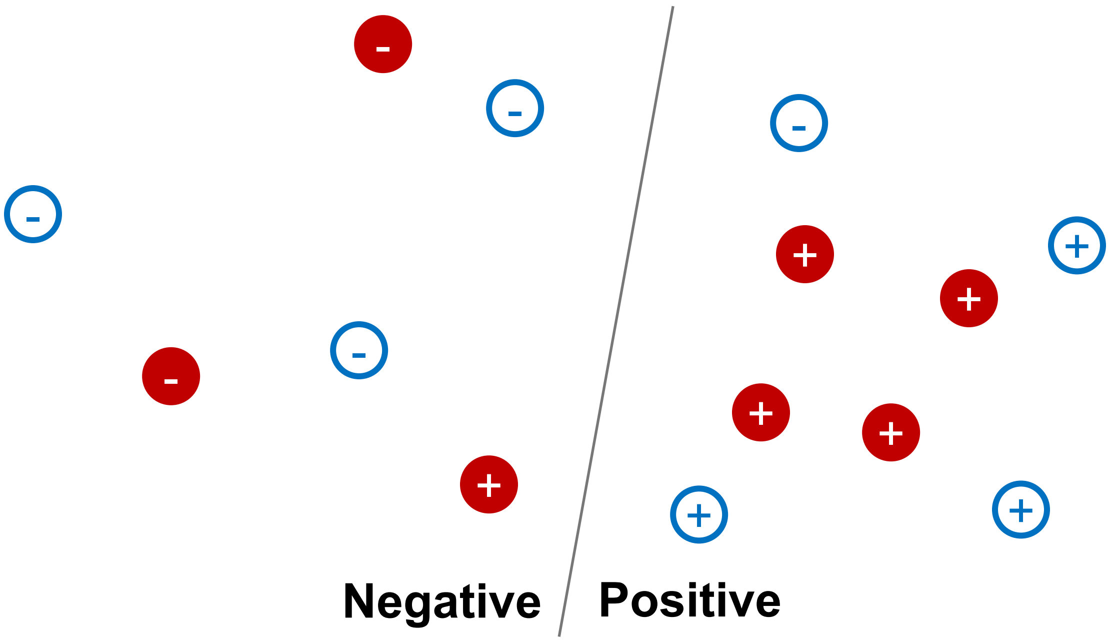
To be fair, 50% of the positive points must be red, 50% must be blue. Same with negative points.

This is not fair.

GROUP FAIRNESS

“Group fairness” or “statistical parity”: demographics in the positive group and negative group are the same as the whole distribution.

Say our demographics are blue and red points.



50% of the points are red, 50% are blue.

To be fair, 50% of the positive points must be red, 50% must be blue. Same with negative points.

This is fair.

EXAMPLE: LOAN DECISIONS

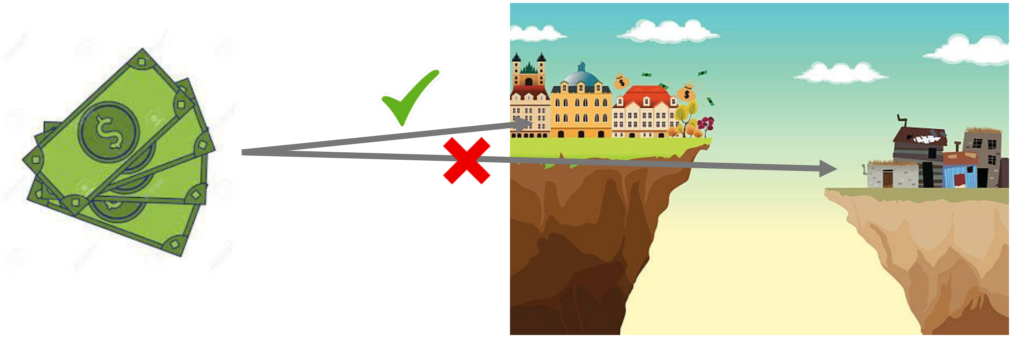
http://research.google.com/bigpicture/attacking-discrimination-in-ml/

DISCRIMINATORY PRACTICES

We are running a classification task on a point set P with labels A. Say S is a protected class (i.e., a racial minority).

Discriminatory practices:

- **Blatant discrimination:** membership in S is explicitly used to give a worse outcome.
- **Redundant encoding:** blatant discrimination but you use a “proxy” metric.
  - **Redlining:** discriminating against neighborhoods because occupants are mostly minorities and/or low-income



DISCRIMINATORY PRACTICES

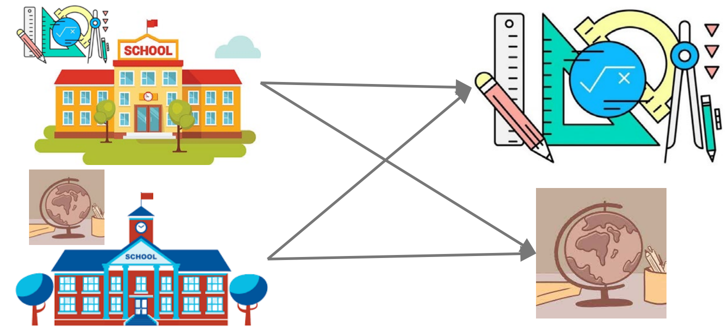
We are running a classification task on a point set P with labels A. Say S is a protected class (i.e., a racial minority).

Discriminatory practices:

- **Disproportionate discrimination:** discriminating against groups because occupants are *disproportionately* minorities and/or low-income.
- **Self-fulfilling prophecy:** deliberately choosing a specific subset of S to discriminate against S.
- **Reverse tokenism:** excusing discrimination against S by citing a "good" member of S^c who is denied service.

LIMITATIONS OF GROUP FAIRNESS

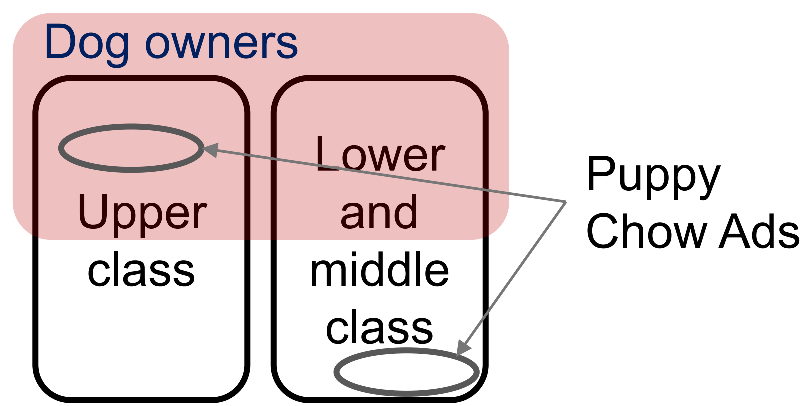
Reduced utility: statistical parity can yield low-utility solutions.



Self-fulfilling prophecy: you can select sub-optimal example and use that as a basis to discriminate.

Kent IQ: 145	Mark IQ: 115	Zak IQ: 95
Kate IQ: 145	Eve IQ: 115	Rose IQ: 95

Subset targeting: you can target irrelevant individuals in S, thereby catering more to S^c.

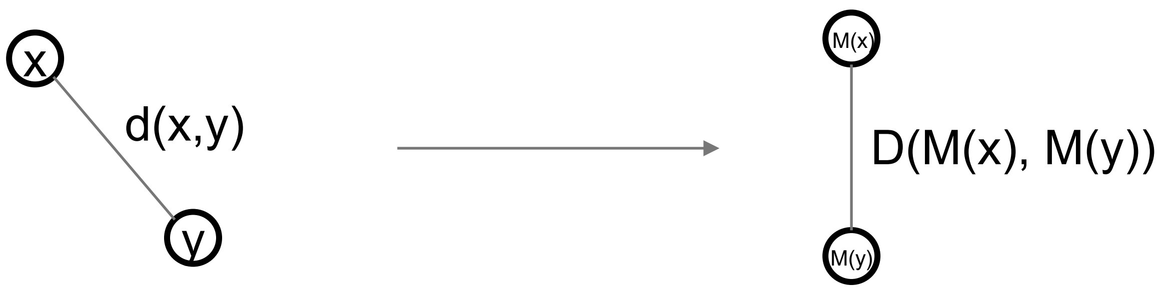


INDIVIDUAL FAIRNESS

"Lipschitz fairness" or **"individual fairness"**: The closer two points are together, the closer their labels should be.

Recall our classifier is M, and M(x) is the label of a point x. Let d,D be a distance function. M is Lipschitz if for any x,y in P:

$$\mathbf{d(x,y) \leq D(M(x), M(y))}$$



Point space, distance function: d

Label space, distance function: D

WHY WE LIKE INDIVIDUAL FAIRNESS

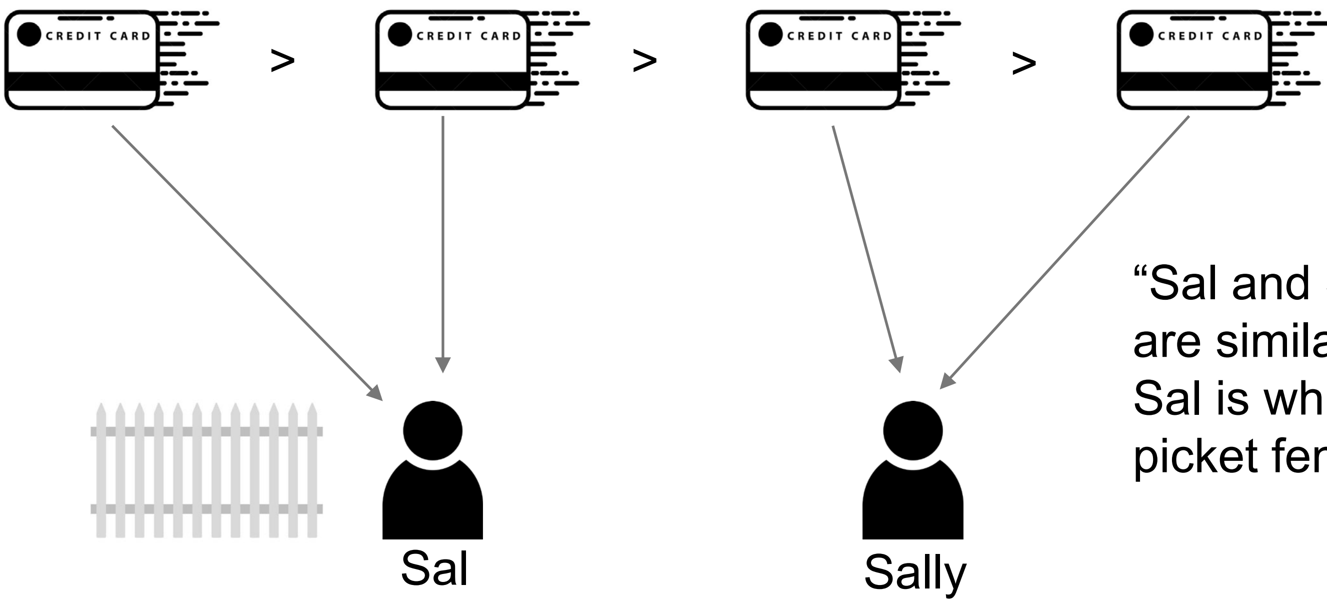
Theorem: In certain circumstances (i.e., certain distance measures), individual fairness implies group fairness.

- In other cases, you can force group fairness while retaining some individual fairness.

Property: Individual fairness is a generalization of *differential privacy.*

Property: Individual fairness prevents reverse tokenism, the self-fulfilling prophecy, and redundant encodings.

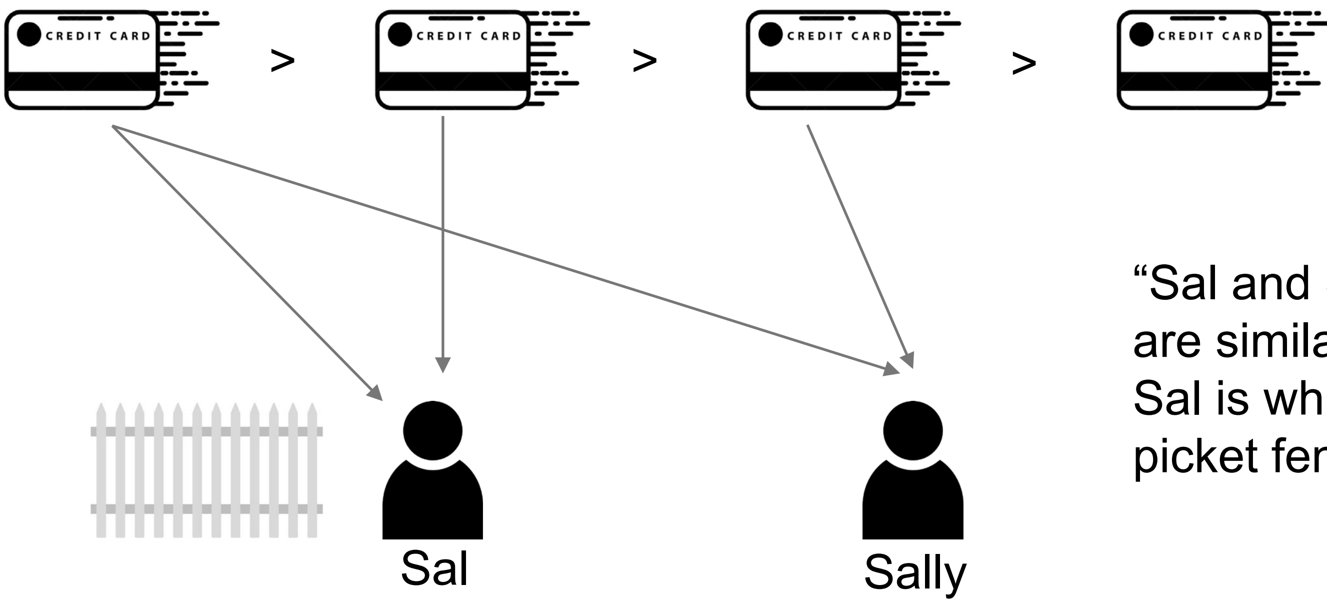
EXAMPLE: AD NETWORK



"Sal and Sally are similar, but Sal is white picket fencer"

Individual fairness guarantees that Sally and Sal are expected to have similar classifications. Prevents reverse tokenism and self-fulfilling prophecy (have them guess!)

EXAMPLE: AD NETWORK



"Sal and Sally are similar, but Sal is white picket fencer"

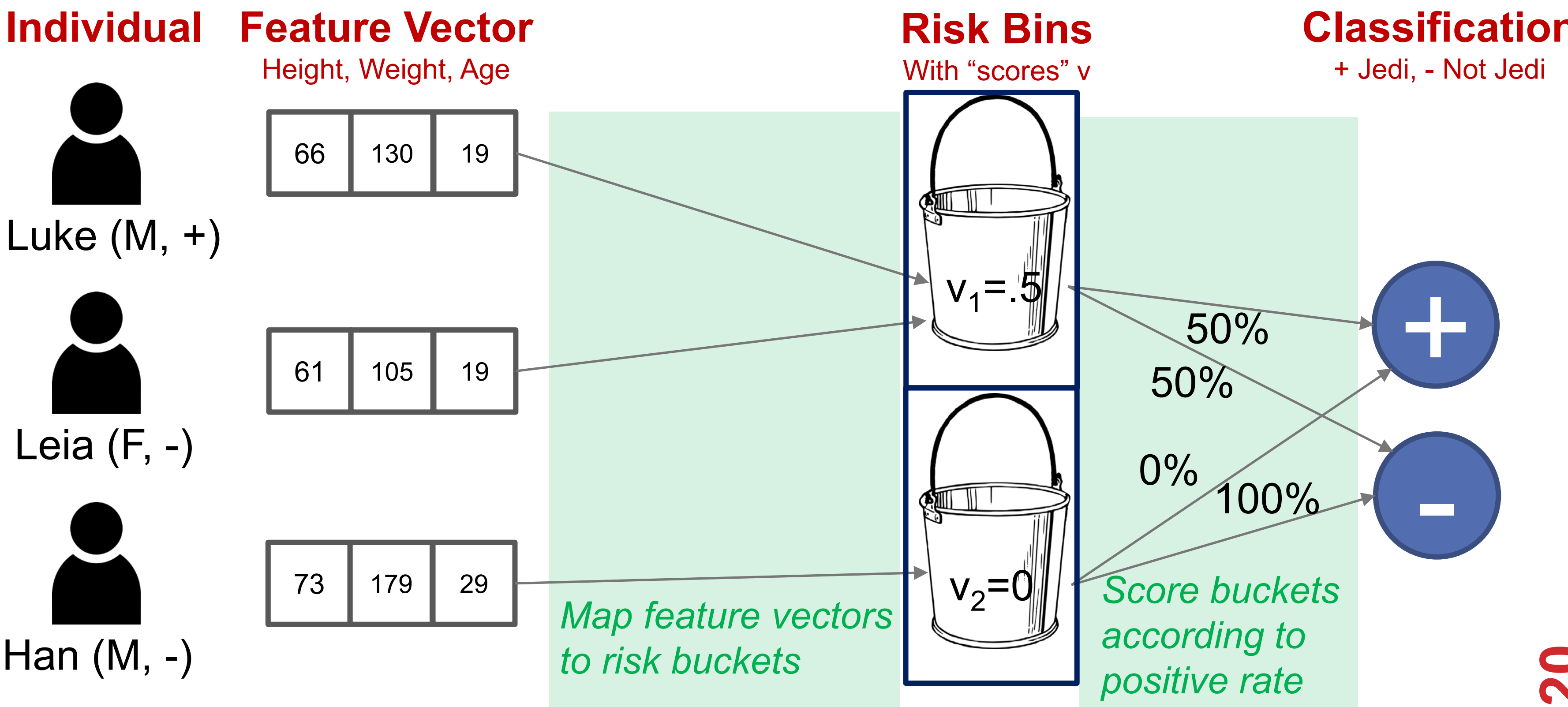
Individual fairness guarantees that Sally and Sal are expected to have similar classifications. Prevents reverse tokenism and self-fulfilling prophecy (have them guess!)

RISK ASSIGNMENTS

Warning: For simplicity purposes, this uses binary gender labels, which may not reflect all possible groups in the data. The issue of datasets using binary gender labels is common and a current topic of interest in fair data collection.

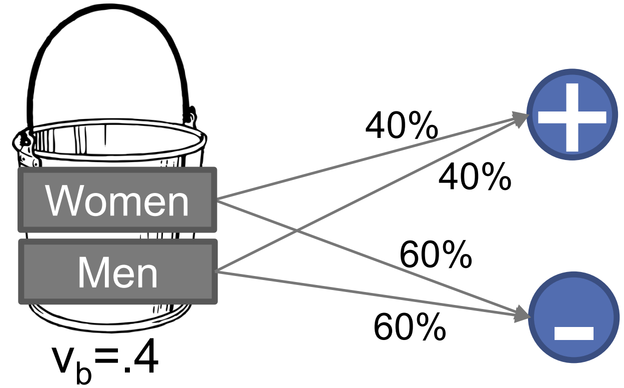
We can also quantify fairness through *risk assignments*.

Task: output a probability someone is a jedi. Protect for gender.

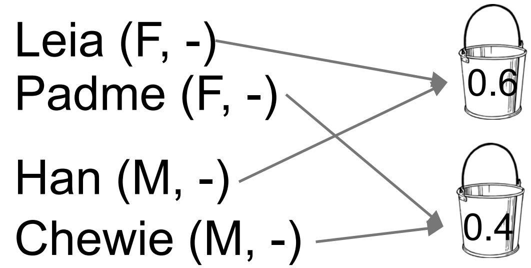


FAIRNESS BY RISK ASSIGNMENT

Calibration within groups: in any bin, men and women have the same chance of jedi classification.

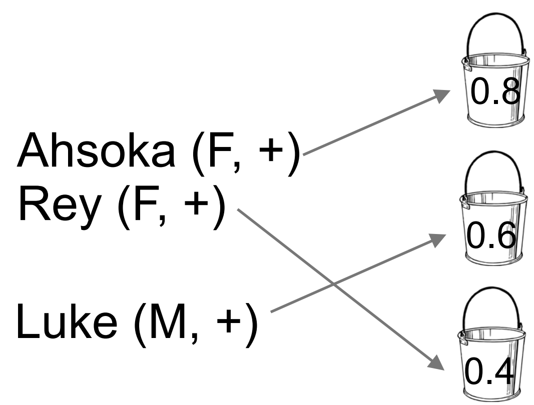


Negative class balance: the average scores of non-jedi men and women are equal



Average Scores
Women: 0.5
Men: 0.5

Positive class balance: the average score of jedi men and women are equal.



Average Scores
Women: 0.6
Men: 0.6

A PERPLEXING CASE: RECIDIVISM PREDICTION

COMPAS risk tool: an intelligent system used by the criminal justice system to assign an estimated chance of convicted criminals to commit reoffenses.

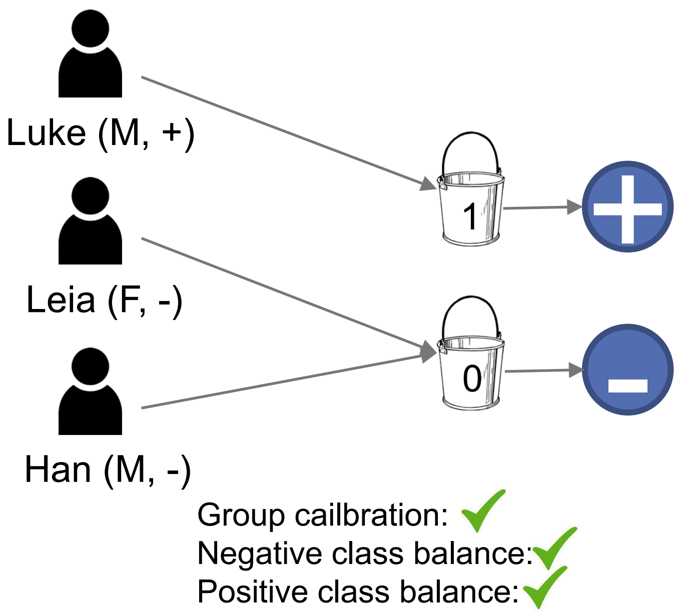

Angwin et al.: claimed that COMPAS discriminated against race because it failed to achieve both *negative class balance* and *positive class balance.*

Counter: claimed COMPAS does not discriminate because it achieves *calibration within groups*.

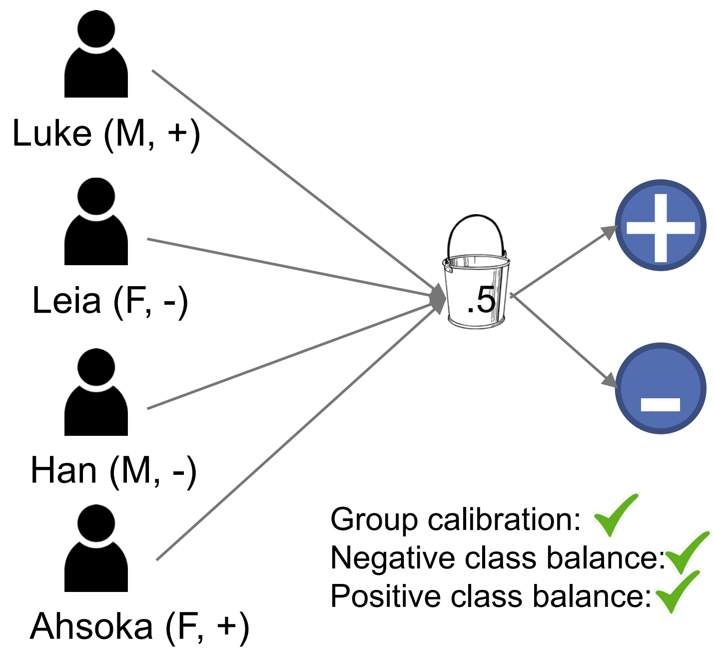
Does that mean it's okay or bad?

HOW MANY CONDITIONS CAN WE GET SIMULTANEOUSLY?

Perfect prediction: We are given who is a jedi.

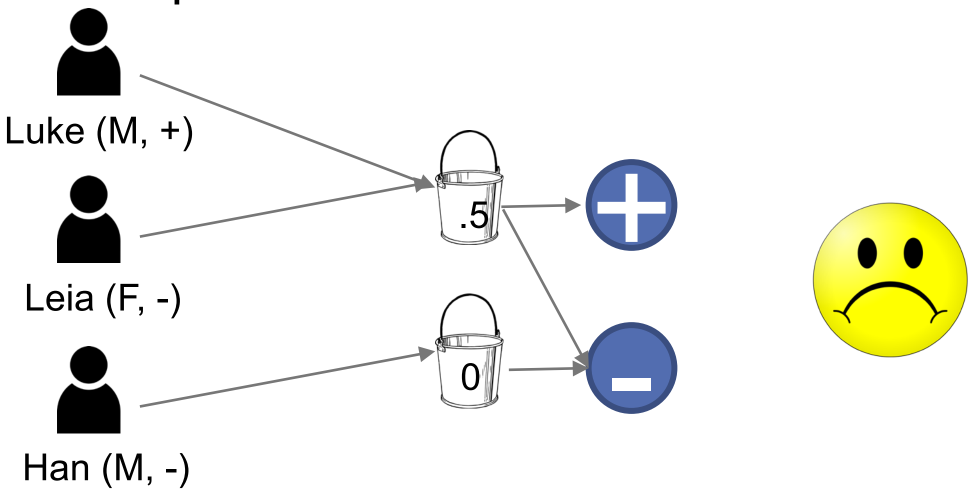


Equal base rates: Men and women are equally likely to be jedis.



BEYOND SPECIAL CASES

Theorem: group calibration, negative class balance, and positive class balance can be achieved all together if and only if there is perfect prediction or equal base rates.



COMPAS: Never could have achieved all 3! But maybe could have done 2.

ETHICAL GUIDE TO FAIR MACHINE LEARNING

Keep in mind: algorithmic fairness inherently interacts with vulnerable and marginalized communities.

Big question: How do we ensure that we serve and give back to these communities

- Do not exploit fair algorithms

Other big question: How do we avoid harming these communities?

Some groups have codes for this, including **the AAAI code of ethics**.

FAIR ALGORITHMS CAN CAUSE HARM!

Consider: We know employers discriminate against applicant's criminal history. Since the criminal justice system exhibits racial discrimination, this issue can propagate to hiring.

- Solve this by banning employers from asking about criminal history?
- No: we know that employers then use race as a proxy for unknown criminal history. This *increases* racial discrimination.

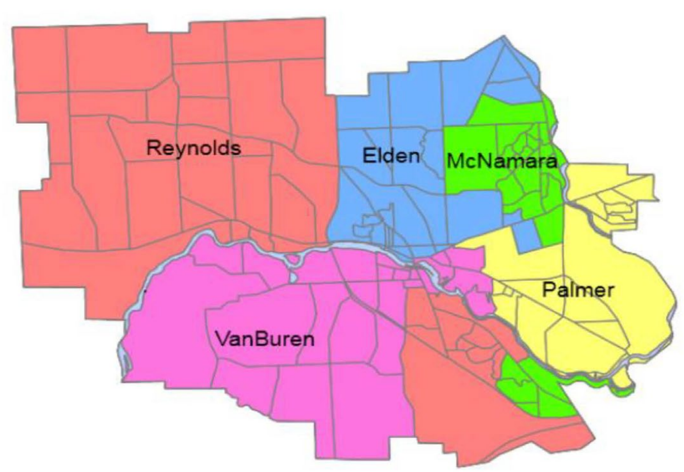
"Imposing a fairness constraint can make the disadvantaged group worse off if the fairness constraint and utilities of the population mismatch."

EXAMPLE: SCHOOL DISTRICTING

In the US, the population is divided ("clustered") into geographic districts. People in the same district use the same school system.

- Funding and resources are not distributed equitably
- Districts are segregated

Baldwinsville Central School District Elementary Boundary Map



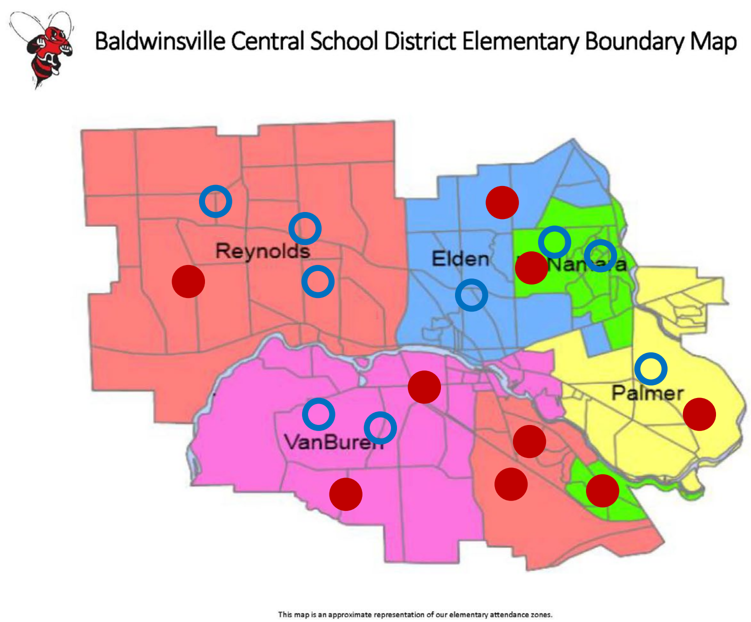
This map is an approximate representation of our elementary attendance zones.

EXAMPLE: SCHOOL DISTRICTING

What if we impose fairness on this clustering problem?

- Ensure the clustering is group fair

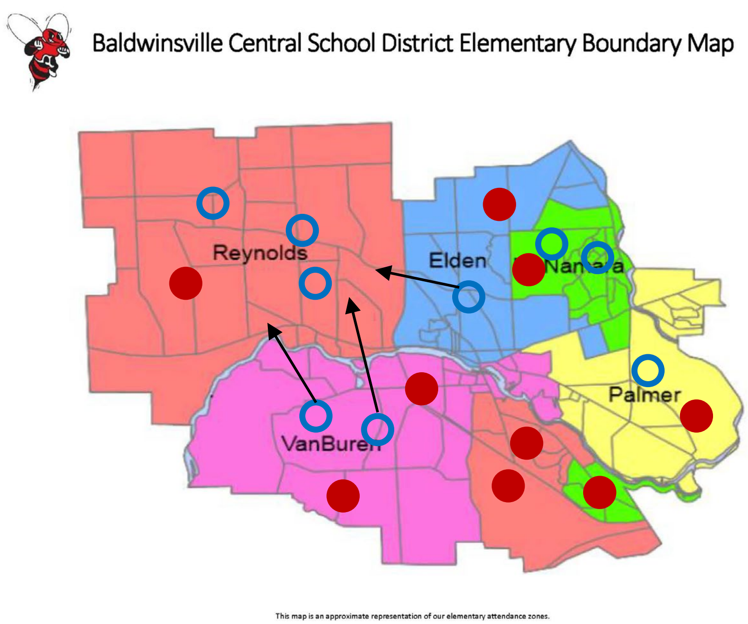
Consider: Who are we serving, and how does this impact them?



EXAMPLE: SCHOOL DISTRICTING

Important considerations:

- Logistics and cost
- People move for schools!



WHEN YOU ARE TRYING TO APPLY FAIRNESS…

Applications and context matter

- Define and model fairness for specific social problems
- General abstractions are useful but often over-sold
- Use caution mapping ideas from fair classification to other fair problems (i.e., fair clustering…)

Fairness interventions do not act in a vacuum

- Broader context and upstream/downstream effects are important
- Different bad inputs require different fair algorithms
- How the algorithm's output is used must also be considered

WHEN YOU ARE TRYING TO APPLY FAIRNESS…

Interdisciplinary research is the best way to use fairness well

- Know your limitations as a researcher, programmer, etc.
- Know relevant work in related areas
- Understand what compromises are most acceptable when ideals can't be achieved
- Establish what is/isn't allowed in practice (i.e., code of ethics)

Real people are involved!

- Who is this for?
- Who are we being fair to?
- What do they want?
- How do they want fairness defined?