

RESPONSIBLE AI

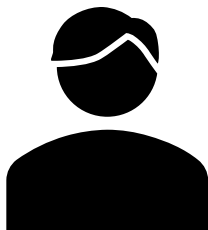
Week 2: A multi-stakeholder perspective on ethical AI



TODAY'S OBJECTIVES

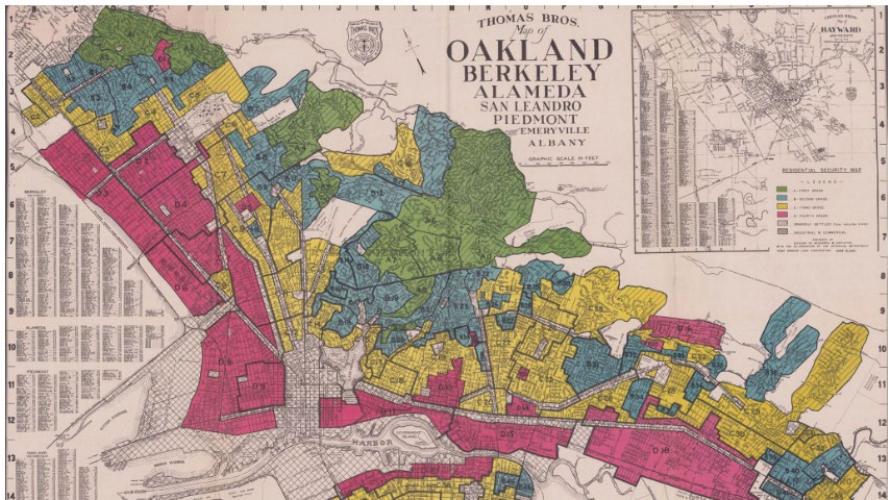
- What are the main doctrines and frameworks of AI fairness?
- What are some limitations to “ethical AI”?
- How does responsible AI affect different fields?
- How do ethical considerations affect different stakeholders?

DIFFERING INTENTS, SIMILAR OUTCOMES



A loan provider from the mid-1900s

“I do not want to give loans to non-white applicants, so I deliberately do not offer services to minority neighborhoods.”

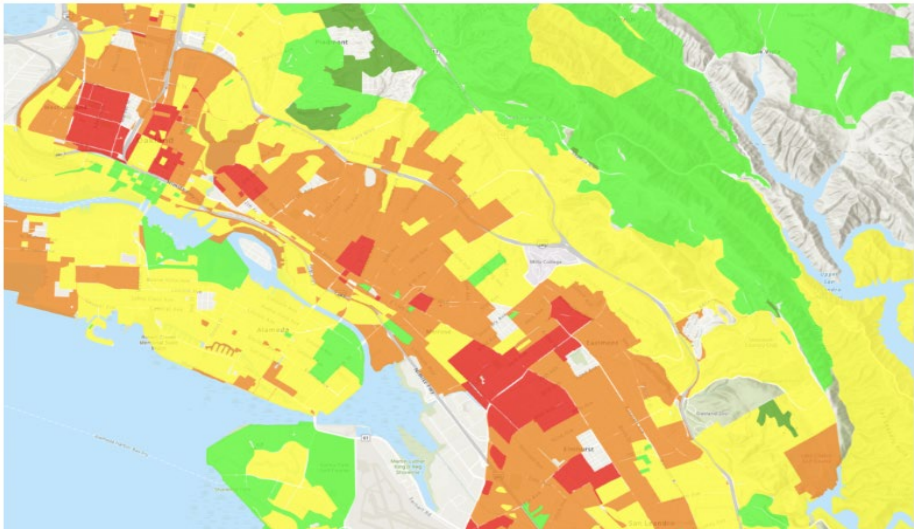


A historical redlining map of the City of Oakland and neighboring towns



An ISP service manager in 2020

“I don’t want to discriminate against anyone, so we choose where to offer high-speed service by looking at expected returns – and never at race.”



A heat map of the City of Oakland's Digital Divide in 2020



Sources:
<https://greenlining.org/publications/online-resources/2020/on-the-wrong-side-of-the-digital-divide/>
<https://itif.org/publications/2022/04/13/broadband-myths-do-isps-engage-digital-redlining/>
<https://www.calcities.org/news/post/2021/10/20/oakland-is-closing-the-digital-divide-through-oakwifi-and-education>

FAIRNESS DOCTRINES: OUTCOMES V. PROCESSES

From Justice¹

Distributive



Procedural

Cumulative Weighted High School GPA	Merit Scholarships	
3.90+	President's Scholarship	\$21,000
3.75–3.89	Trustees' Scholarship	\$19,000
3.35–3.74	Deans' Scholarship	\$17,000
3.00–3.34	Fellows' Scholarship	\$15,000
under 3.00	BW Yellow Jacket Success Grant	\$12,000

outcome-
focused

process-
focused

From Harms²

Allocative

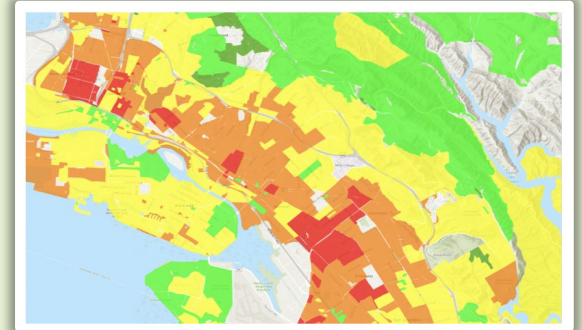


Representational

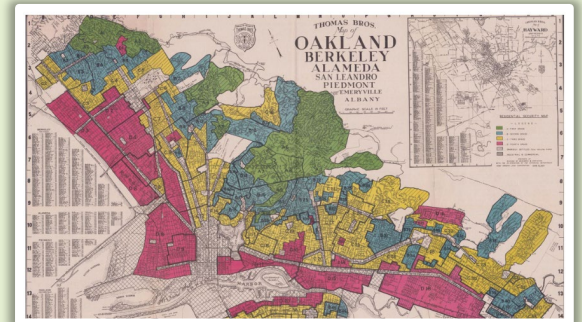


From Law³

Disparate Impact



Disparate Treatment



1: A Theory of Justice (Rawls): <https://www.jstor.org/stable/j.ctvjf9z6v>

2: Kate Crawford's NIPS 2017 Keynote, "The Trouble With Bias": https://www.youtube.com/watch?v=fMym_BKWQzk; A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle: <https://arxiv.org/pdf/1901.10002.pdf>

3: Big Data's Disparate Impact (Barocas and Selbst): <https://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf>; Bloomberg News: <https://www.bloomberg.com/graphics/2016-amazon-same-day/>

INTERDISCIPLINARY DEFINITIONS OF FAIRNESS

The general concept of fairness is familiar to most but assumes different pursuit of end goals in different disciplines



Law

Fairness includes protecting individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories.

Fair Housing Act, Equal Credit Opportunity Act, Americans with Disabilities Acts, Title VII of the Civil Rights Act and the Age Discrimination in Employment Act



Social science

“often considers fairness in light of social relationships, power dynamics, institutions and markets.”

Social contract theory



Quantitative fields

(i.e., math, computer science, statistics, economics): questions of fairness are seen as mathematical problems. Fairness tends to match to some sort of criteria, such as equal or equitable allocation, representation, or error rates, for a particular task or problem.

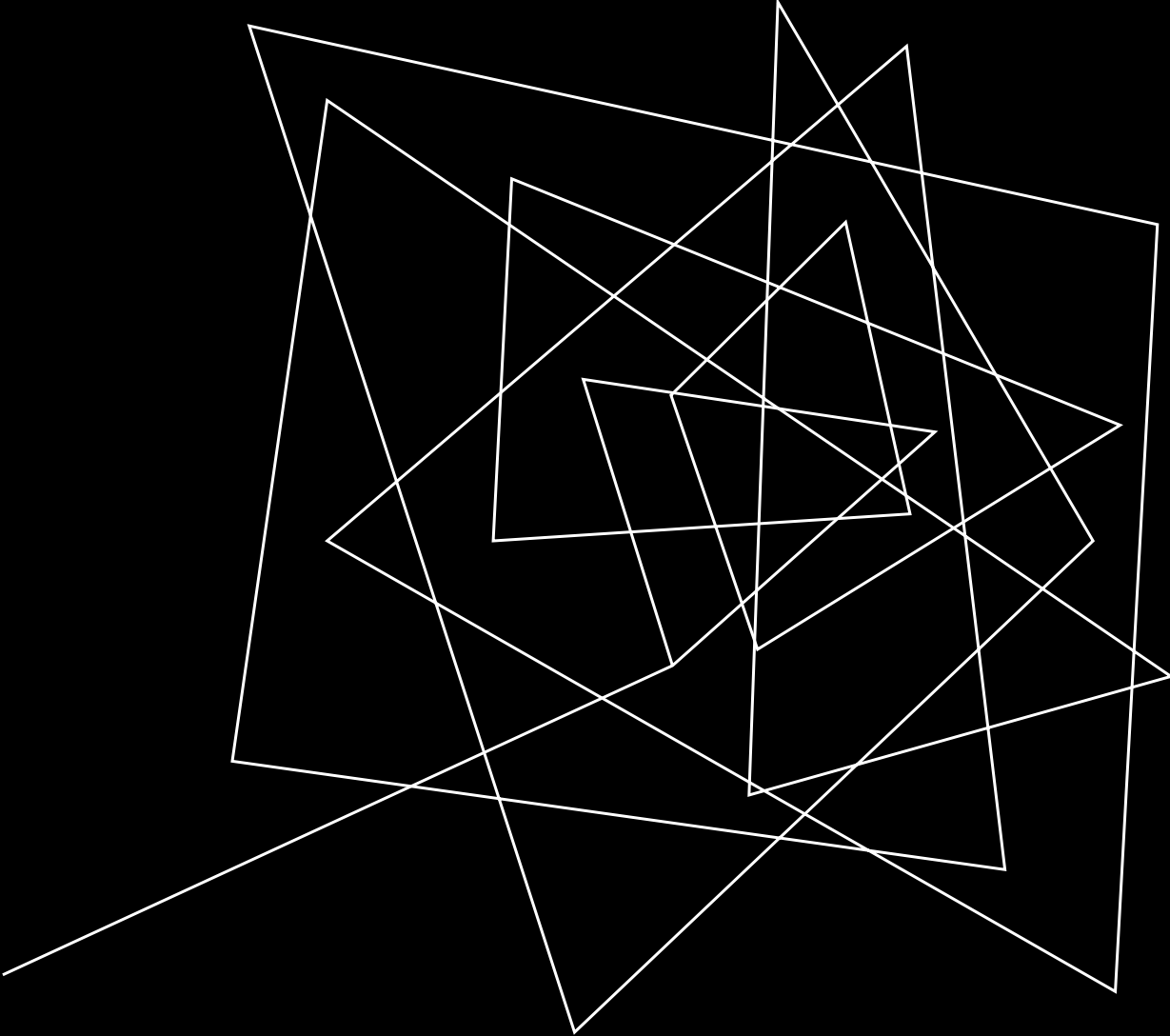
Design of market, Voting mechanisms, statistical and machine learning models.



Philosophy

Ideas of fairness “rest on a sense that what is fair is also what is morally right.” Political philosophy connects fairness to notions of justice and equity.

Virtue ethics, deontological ethics



READING PRESENTATION #1:

*Inherent Trade-Offs in the
Fair Determination of Risk
Scores*

THE IMPOSSIBLE QUEST...

False Positive & False Negative Rates

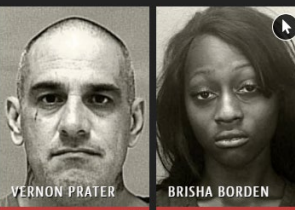
“Do the algorithm’s mistakes affect Black and white defendants in the same way?”

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.)

Two Petty Theft Arrests

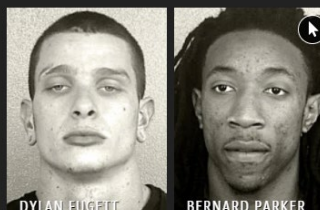


VERNON PRATER LOW RISK 3

BRISHA BORDEN HIGH RISK 8

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

Two Drug Possession Arrests



DYLAN FUGETT LOW RISK 3

BERNARD PARKER HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Two Shoplifting Arrests



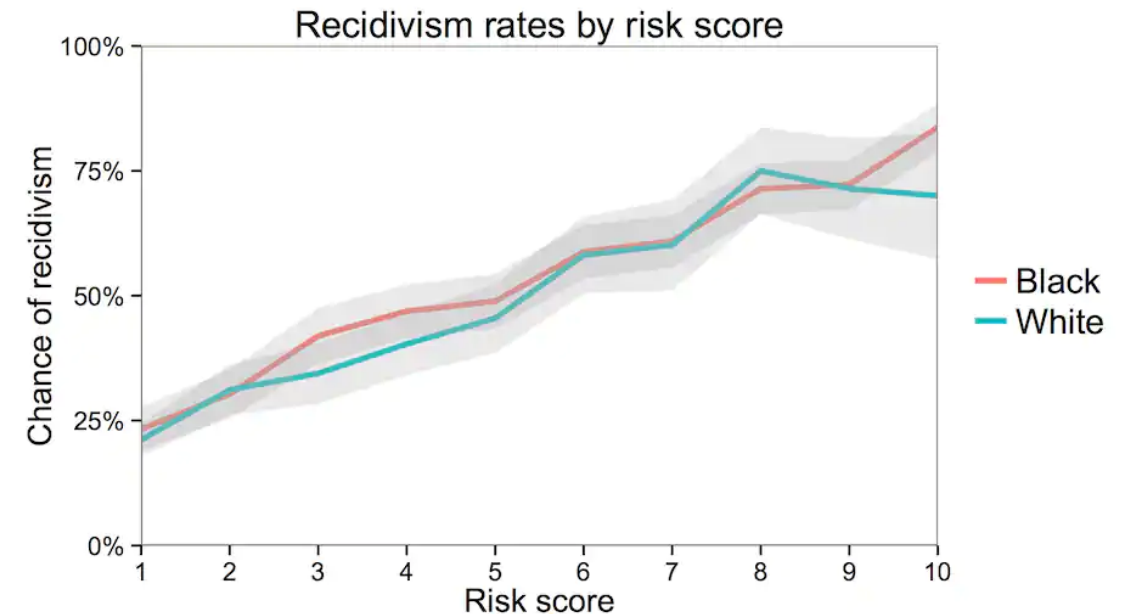
JAMES RIVELLI LOW RISK 3

ROBERT CANNON MEDIUM RISK 6

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted \$1,000 worth of tools from a Home Depot.

Calibration

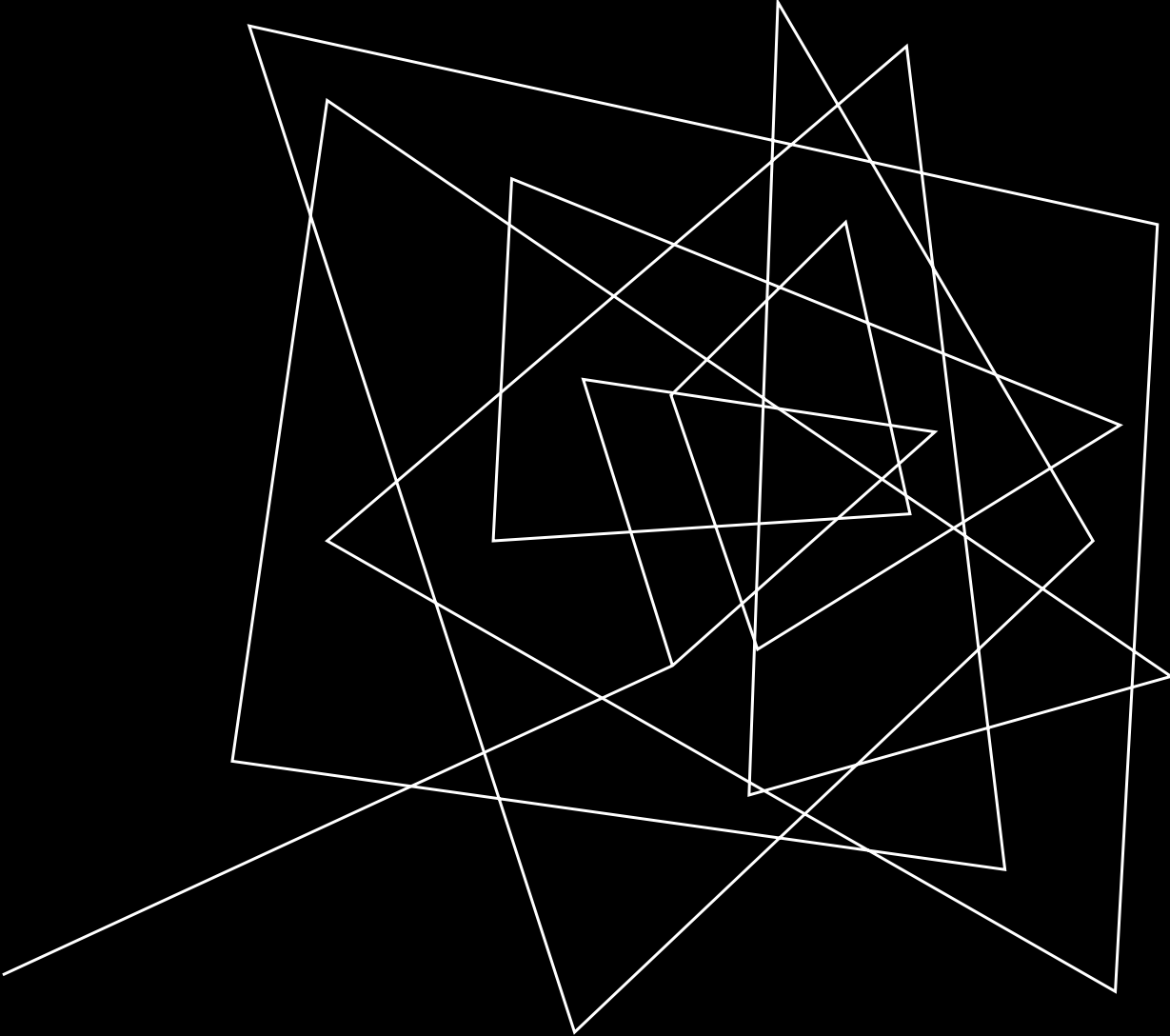
“Can scores be interpreted the same way for both Black and white defendants?”



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

<https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?variant=116ae929826d1fd3>

http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf



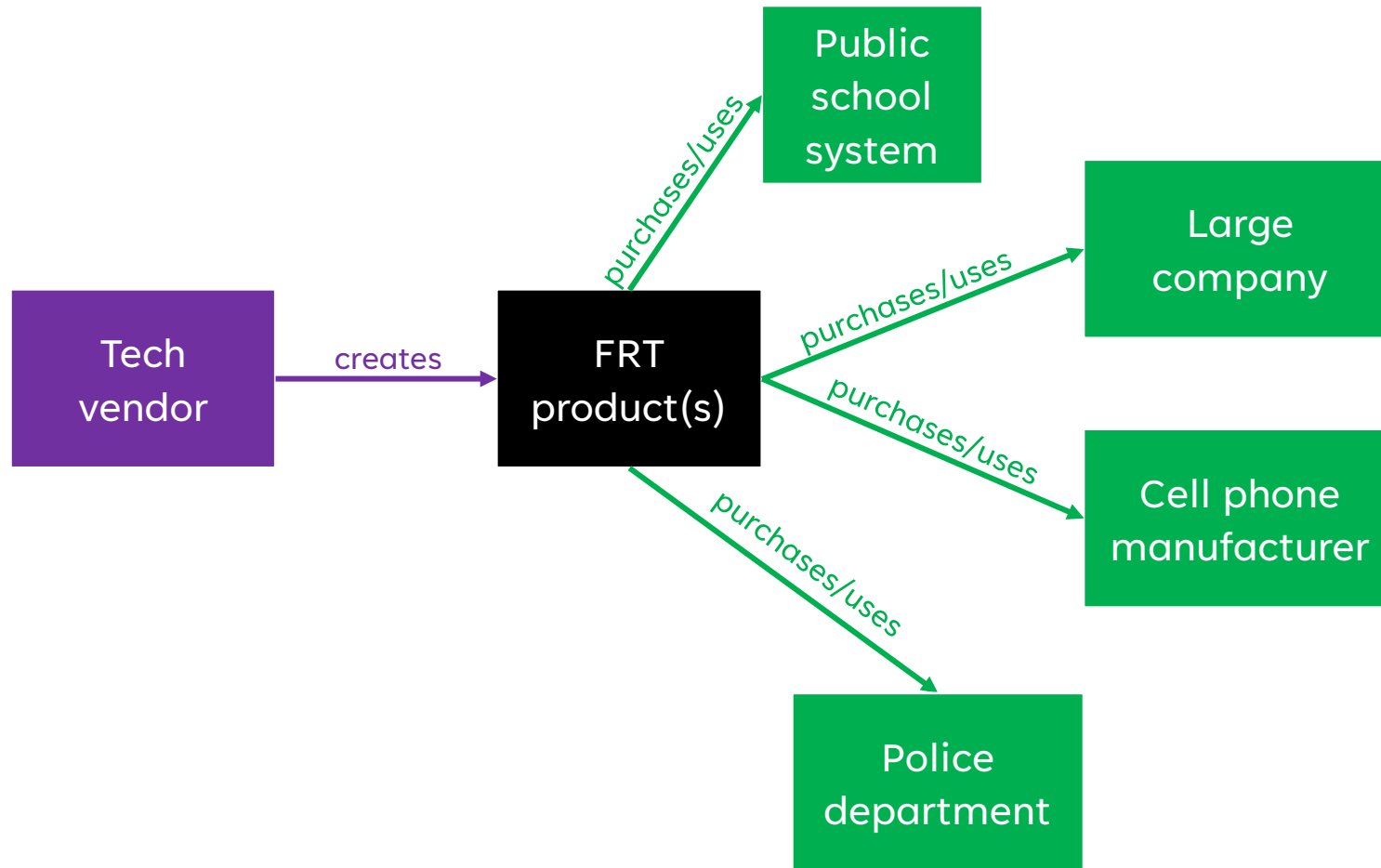
READING PRESENTATION #2:

*Moving Beyond Audits (AI
Now Report, pg. 34-42)*

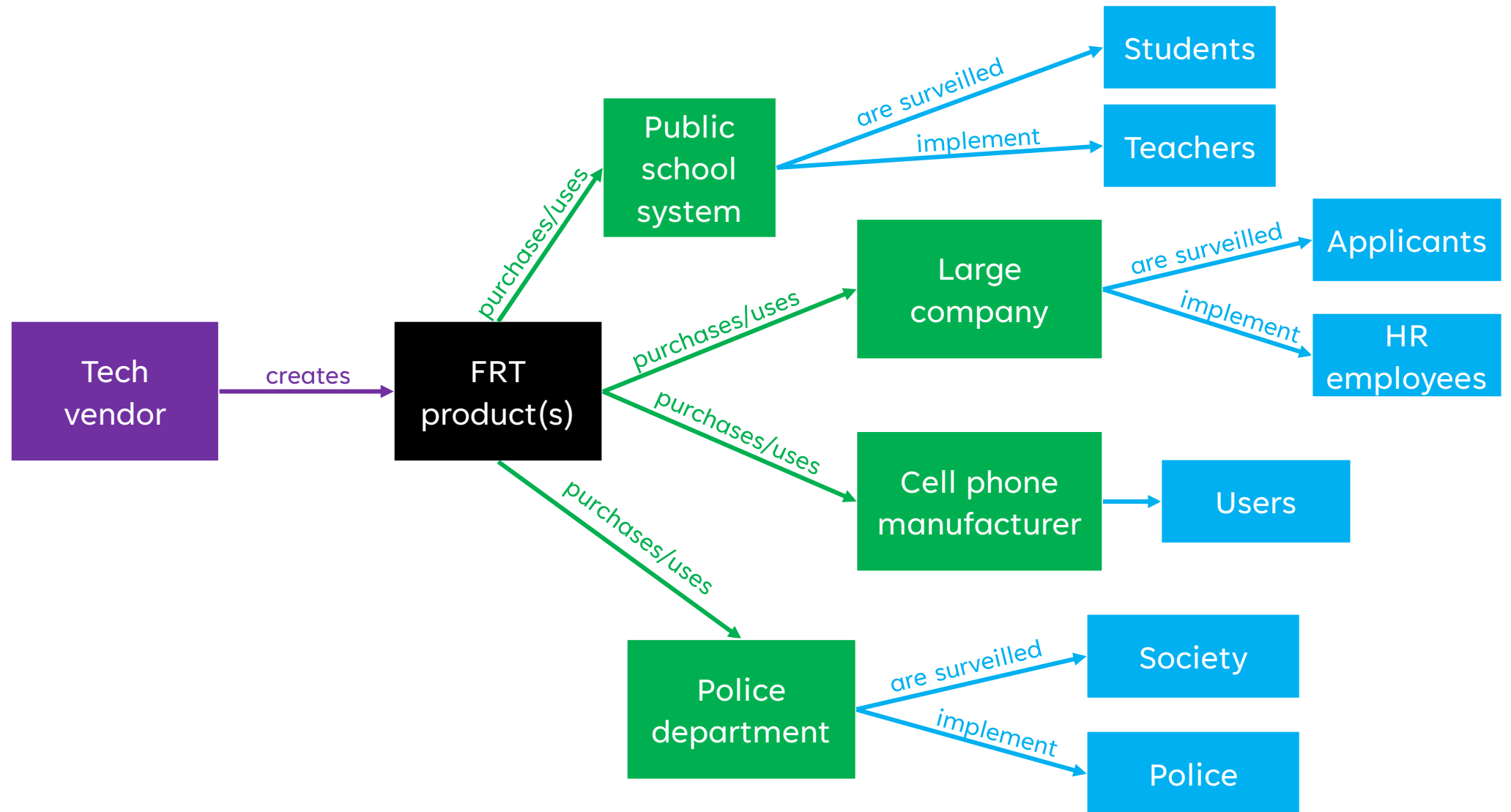
EXAMPLE: FACIAL RECOGNITION TECHNOLOGY (FRT)

FRT
product(s)

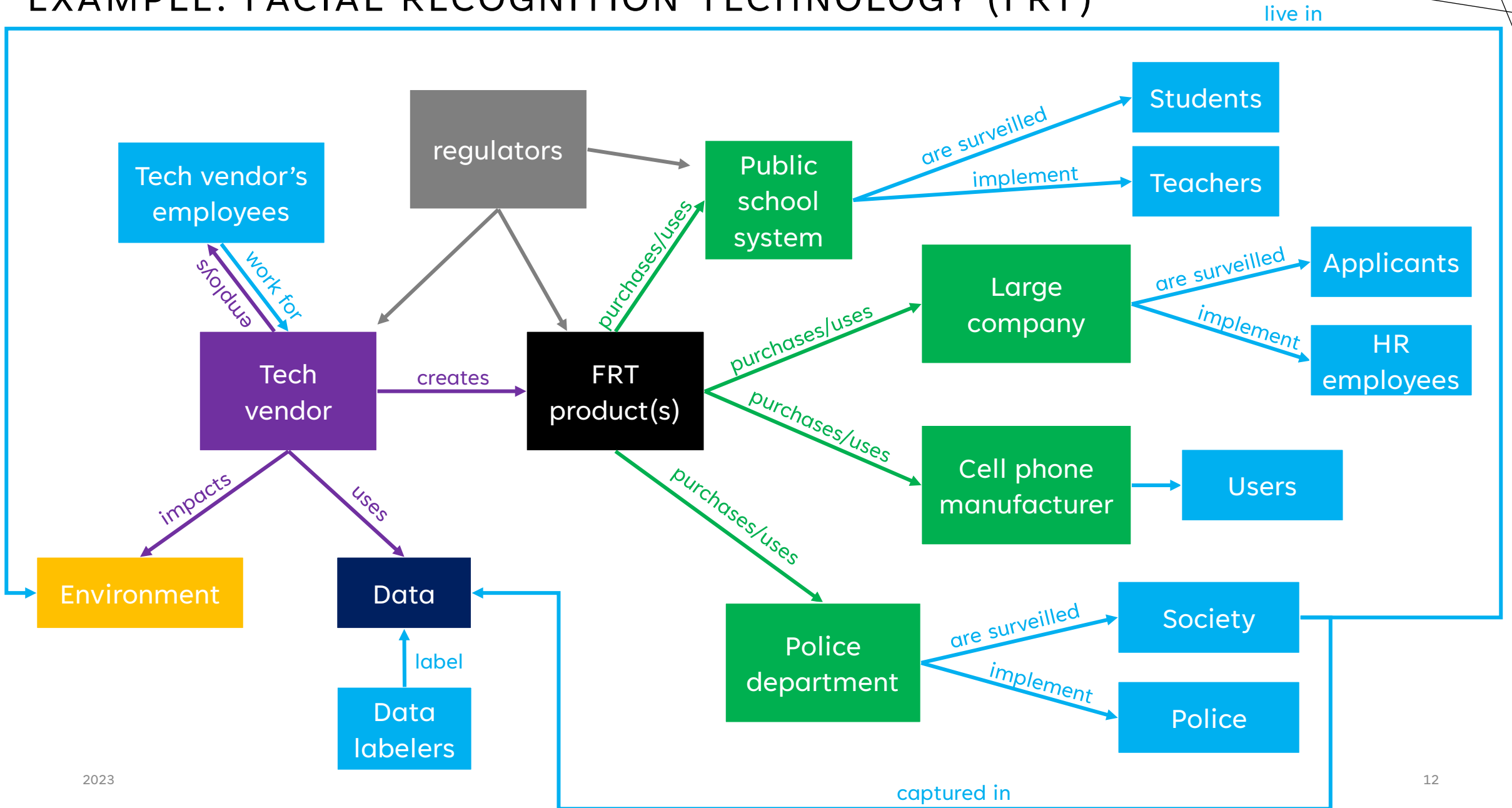
EXAMPLE: FACIAL RECOGNITION TECHNOLOGY (FRT)



EXAMPLE: FACIAL RECOGNITION TECHNOLOGY (FRT)



EXAMPLE: FACIAL RECOGNITION TECHNOLOGY (FRT)





DISCUSSION

Consider a home loan provider that takes a variety of demographic and financial information from applicants, puts the features into a model, and uses it to decide whether an applicant is eligible for a loan.

Who all are the likely **stakeholders** in this example? What is the **impact** of a correct* decision on each type of stakeholder? An incorrect decision?

In breakout rooms, take **five minutes to whiteboard a stakeholder map** and discuss the potential impact of true/false positives/negatives to different stakeholders.

Make sure to nominate someone to (1) screenshot your stakeholder map before the breakout rooms close and (2) share it with the room when we reconvene!

**Bonus: how are you defining “correct”? How do you think the model defines it?*

FOR NEXT WEEK

- Complete next week's **readings**
 - If you sign up to present on *Datasheets for Datasets*, come prepared to present next week and submit your presentation on Gradescope by 10AM PT on Friday, October 20
- Submit your response to **Writeup #2** to Gradescope by 10AM PT on Friday, October 20