Machine Learning & Ethics

William Seymour | Human Centred Computing
william.seymour@cs.ox.ac.uk

Overview

- ~~Introduction to Machine Learning (13:30-14:10, Access Grid)~~
- ~~Practical: Building a Classifier (14:20-15:00, Comlab)~~
- Ethics in Machine Learning (15:10-15:50, Access Grid)
  - Ethical challenges in computer science
  - Academic work on ethics in ML
  - What does it mean to be fair?
- Practical: Operationalising Ethics (16:00:16:30, Comlab)
  - Explaining image classifiers
  - Revisiting pred pol classifiers from earlier

Examples?

Woman Follows GPS, Drives Car Into Canada's Georgian Bay

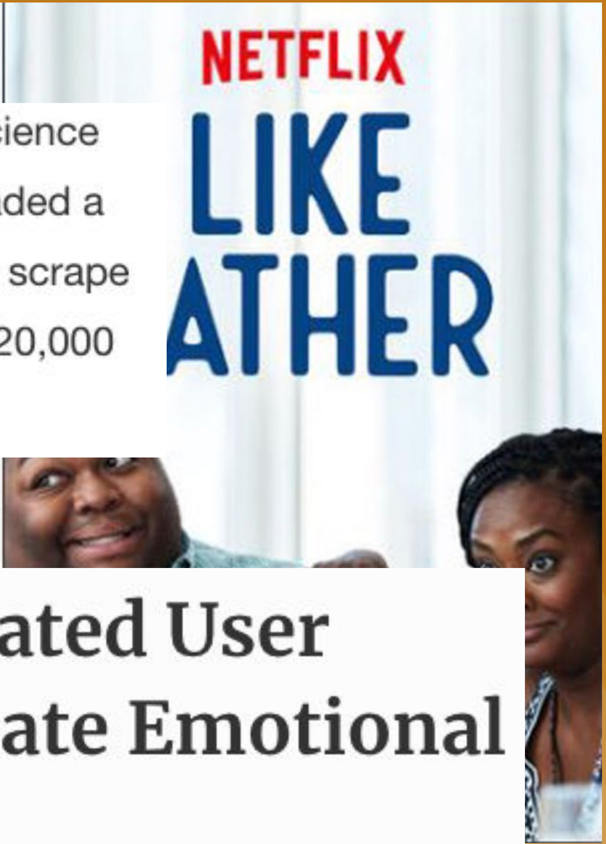
The 23-year-old Canadian woman took a wrong turn onto a boat ramp to the bay.

Top Stories

1 in 4 people near Congo's Ebola outbreak believe virus isn't real, new study says

Pope Francis travels to Morocco,

A user of **Kaggle,** a platform for machine learning and data science competitions which was recently acquired by Google, has uploaded a facial data set he says was created by exploiting Tinder's API to scrape 40,000 profile photos from Bay Area users of the dating app — 20,000 apiece from profiles of each gender.



200
Ok

men rating women	BLACK men rating...	3%	-3%	3%	-3%
	LATINO men rating...	7%	-22%	6%	9%
	WHITE men rating...				
		ASIAN m			
women rating men	ASIAN women rating...	1			
	BLACK women rating...	-1			
	LATINA women rating...	-16%	-4%	11%	10%
	WHITE women rating...	-12%	-6%	1%	17%

Facebook Manipulated User News Feeds To Create Emotional Responses

What can we do?

- Codes of conduct / ethical guidance
- Derive working definitions of what we consider “fair”
- Data protection laws (e.g. GDPR)
- FAT/ML: Fairness, Accountability, and Transparency

What can we do?

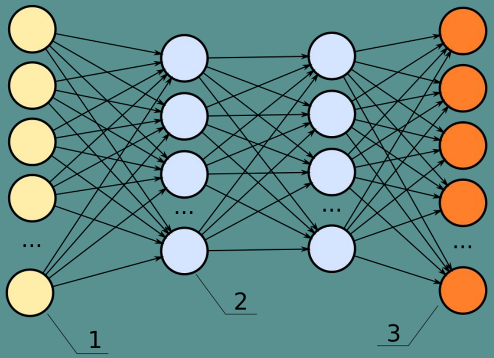
- Codes of conduct / ethical guidance
- Derive working definitions of what we consider “fair”
- Data protection laws (e.g. GDPR)
- **FAT/ML: Fairness, Accountability, and Transparency**

The Ethics of Algorithms

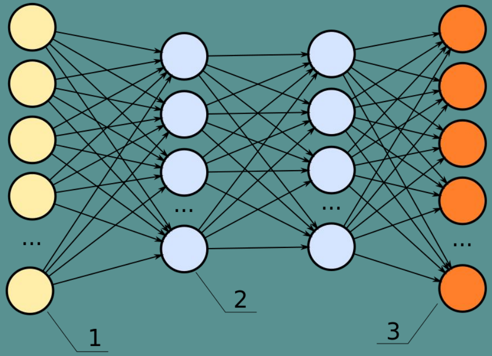
1. Inconclusive Evidence
2. Inscrutable Evidence
3. Misguided Evidence
4. Unfair Outcomes
5. Transformative Effects
6. Traceability

The Ethics of Algorithms

1. Inconclusive Evidence
2. **Inscrutable Evidence**
3. Misguided Evidence
4. Unfair Outcomes
5. Transformative Effects
6. Traceability

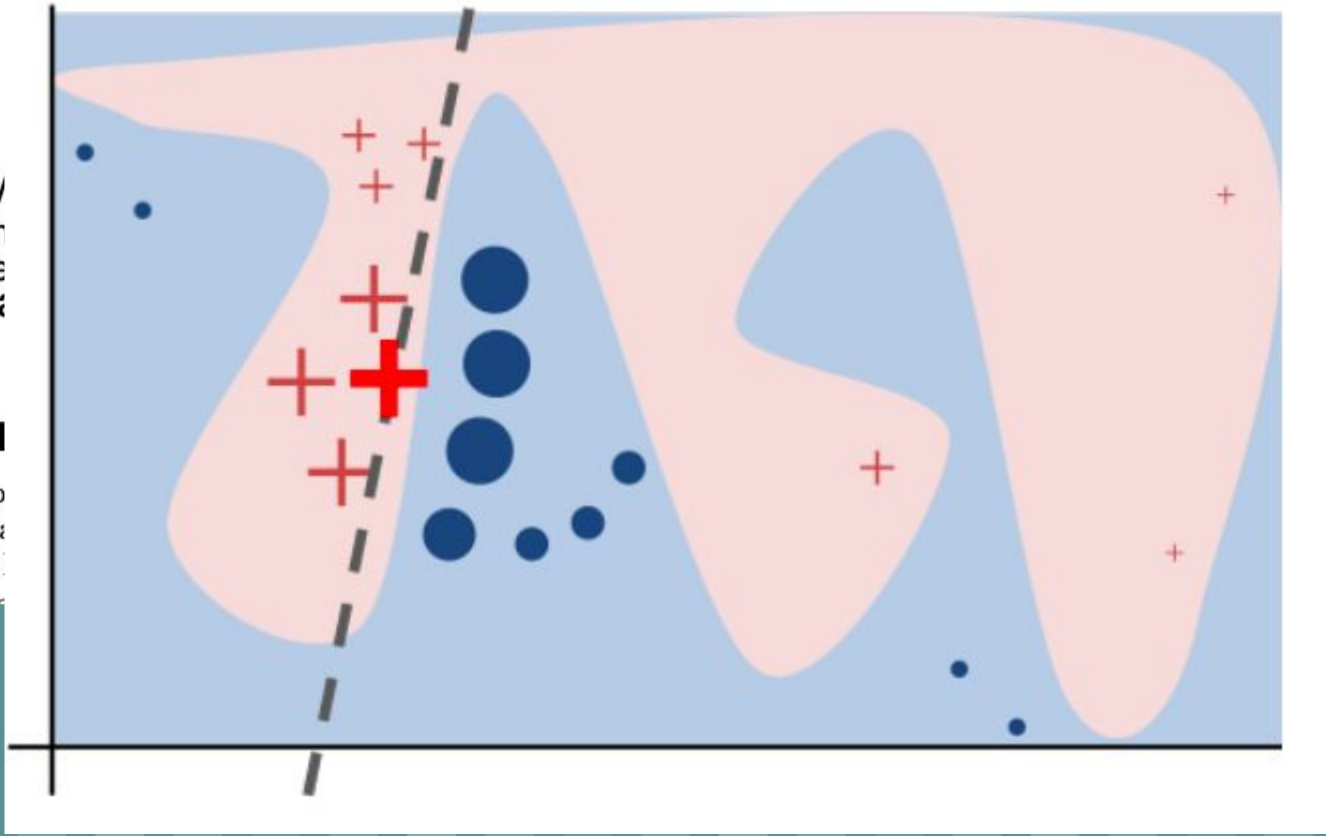


It’s a rabbit



You can't get a loan

“Why Should I Trust You?”



ABSTRAC

Despite widesp
main mostly bl
predictions is,

’s behaviour, as
s is an important
making. When
[6] or terrorism

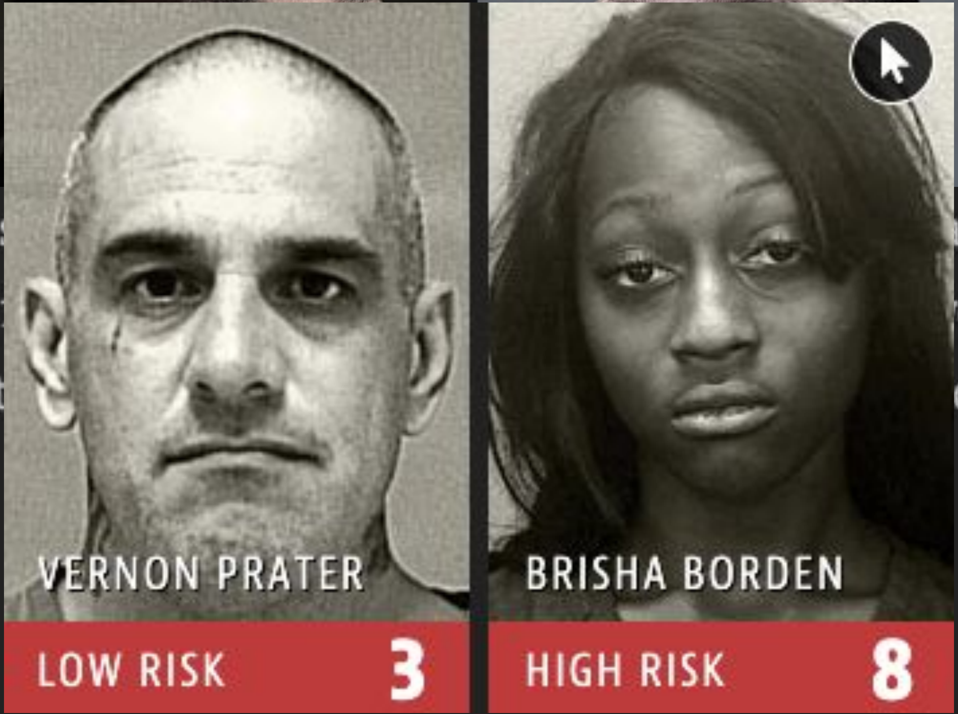
The Ethics of Algorithms

1. Inconclusive Evidence
2. Inscrutable Evidence
3. Misguided Evidence
4. **Unfair Outcomes**
5. Transformative Effects
6. Traceability

H1: Maximum Profit

About that data set we used earlier…

We obtained the risk … arrested in Broward County, Florida, in 20… were charged with new crimes over the next t… creators of the algorithm.



ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

An aside: the personal impact of classification

- Everyone that COMPAS classified was a real person
- Classification scores were used to distribute services within prisons
- In some states, COMPAS scores are used during sentencing
- How many years of life disappeared due to this algorithm?

H1: Maximum Profit

Group	Count
Combined	6172
African American	3175
Caucasian	2103
Hispanic	509
Asian	31
Native American	11
Other	343

Group	Count	Accuracy (%)
Combined	6172	64.4
African American	3175	64.6
Caucasian	2103	64.9
Hispanic	509	58.7
Asian	31	74.2
Native American	11	81.8
Other	343	65.9

H2: Demographic Parity

Group	Count	Accuracy (%)
Combined	6172	64.4
African American	3175	64.6
Caucasian	2103	64.9
Hispanic	509	58.7
Asian	31	74.2
Native American	11	81.8
Other	343	65.9

Group	Count	Accuracy (%)	P(Recid) (%)
Combined	6172	64.4	48.4
African American	3175	64.6	55.8
Caucasian	2103	64.9	41.6
Hispanic	509	58.7	38.7
Asian	31	74.2	32.3
Native American	11	81.8	54.6
Other	343	65.9	37.9

H3：Equal Accuracy

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035

FP rate: 32.35

FN rate: 37.40

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035
FP rate: 32.35		
FN rate: 37.40		

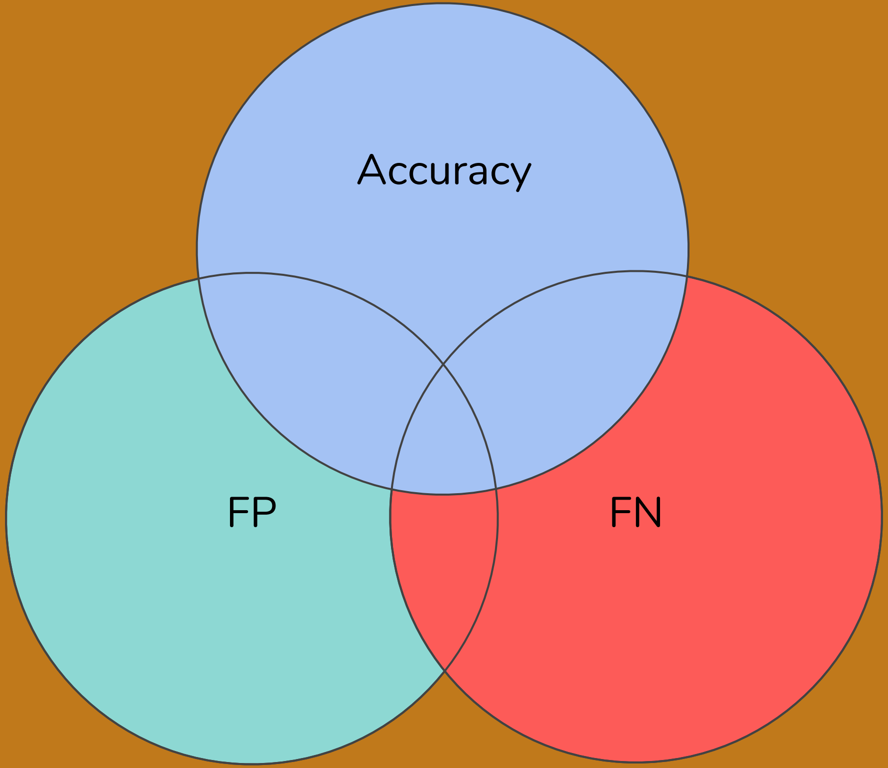
Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		

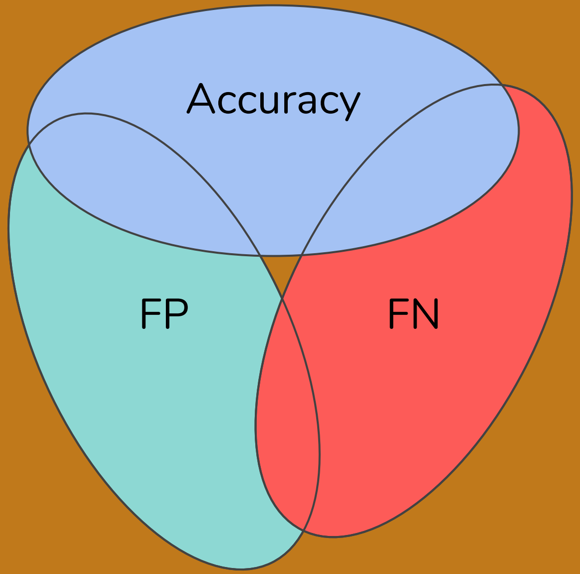
White Defendants	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		

Who is Wronged?

- True negative: low risk criminal given more lenient sentence
- True positive: high risk criminal given harsher sentence
- False negative: high risk criminal given more lenient sentence
- False positive: low risk criminal given harsher sentence

H4: Equal Opportunity


Accuracy
FP
FN


Accuracy
FP
FN

Big Data
Volume 5 Number 2, 2017
© Mary Ann Liebert, Inc.
DOI: 10.1089/big.2016.0047

Group	P(Recid) (%)
Combined	48.4
African American	55.8
Caucasian	41.6
Hispanic	38.7
Asian	32.3
Native American	54.6
Other	37.9

Can we just ignore {race, age, gender}?

- Just excluding protected class data seems like an obvious option
- Other features can act as **proxies** for protected classes
- Might cause us to unwittingly perpetuate systemic discrimination

Practical #2

- Explaining image predictions using LIME
- Exploring unfairness in your classifiers from earlier
    - Do you think your classifier is fair?
    - How does it treat people of different groups?
    - Does it adhere to any of the fairness definitions we just covered?