Explaining algorithms and automation: A guide for lawyers

Dr. Reuben Binns, Dept. Computer Science, University of Oxford

reuben.binns@cs.ox.ac.uk

Overview

Digitisation, automation, decision-making

Rule-based vs statistical systems

Bias, error, discrimination

Automation bias

Unequal and fettered discretion

UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

WHAT ARE HUMAN RIGHTS? | DONATE

HOME | ABOUT US | ISSUES | HUMAN RIGHTS BY COUNTRY | WHERE WE WORK | HUMAN RIGHTS BODIES | NEWS AND EVENTS | PUBLICATIONS RESOURCE

nglish > News and Events > **DisplayNews**

World stumbling zombie-like into a digital welfare dystopia, warns UN human rights expert

NEW YORK (17 October 2019) – A UN human rights expert has expressed concerns about the emergence of the "digital welfare state", saying that all too often the real motives behind such programs are to slash welfare spending, set up intrusive government surveillance systems and generate profits for private corporate interests.

"As humankind moves, perhaps inexorably, towards the digital welfare future it needs to alter course significantly and rapidly to avoid stumbling zombie-like into a digital welfare dystopia," the Special Rapporteur on extreme poverty and human rights, Philip Alston, says in a *report* to be presented to the General Assembly on Friday.

https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25156&LangID=E

'Digital welfare state': big tech allowed to target and surveil the poor, UN is warned

https://www.theguardian.com/technology/2019/oct/16/digital-welfare-state-big-tech-allowed-to-target-and-surveil-the-poor-un-warns

Digitisation, automation, decision-making

Digitisation of paper forms (e.g. tax returns online)

Automation of processes (e.g. automatically recurring payments)

Computer-supported /automated **decision-making (ADM)**, e.g.**:**

- Determining eligibility for benefit
- Risk scoring based on statistical models
- Fraud detection

Rules-based systems

e.g.

IF “years_in_residence” > 5:
THEN:
“settled_status_eligibility” = TRUE

THE EPISTEMOLOGY OF A RULEBASED EXPERT SYSTEM

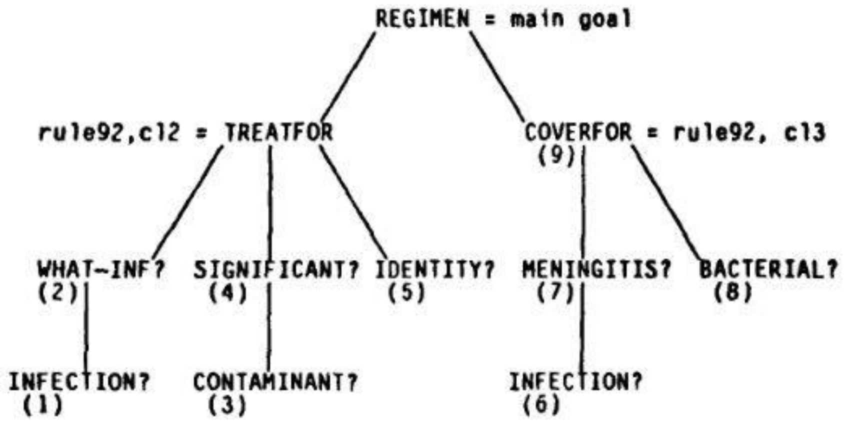


Figure 14. Portion of MYCIN's inference structure
(Numbers give the order in which non-place-holder goals are achieved by the depth-first interpreter.)

Clancey, William J. "The epistemology of a rule-based expert system—a framework for explanation." *Artificial intelligence* 20.3 (1983): 215-251.

Statistical systems

Aim to *classify*, *predict*, or *score*

How similar is this benefits application to previously fraudulent ones?

How likely is this person to re-offend (based on statistics from previous cases?)

How risky is the person behind this visa application?

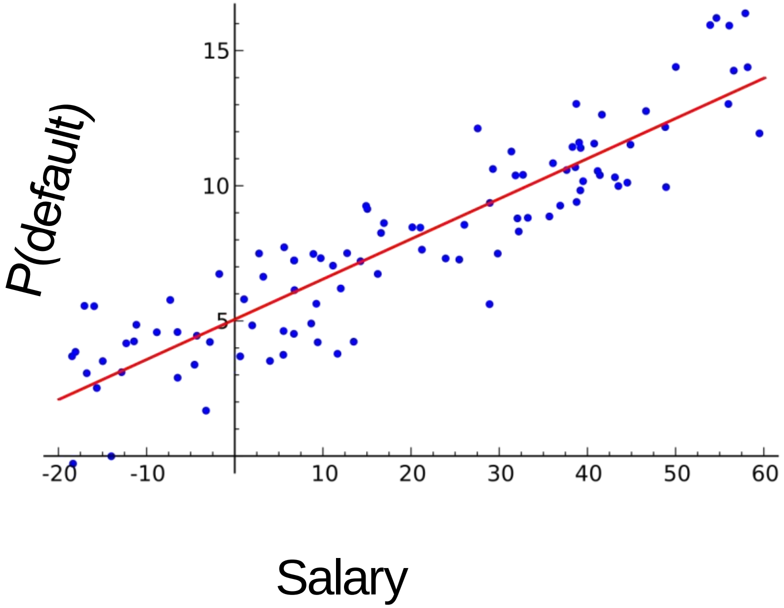
Default

Salary

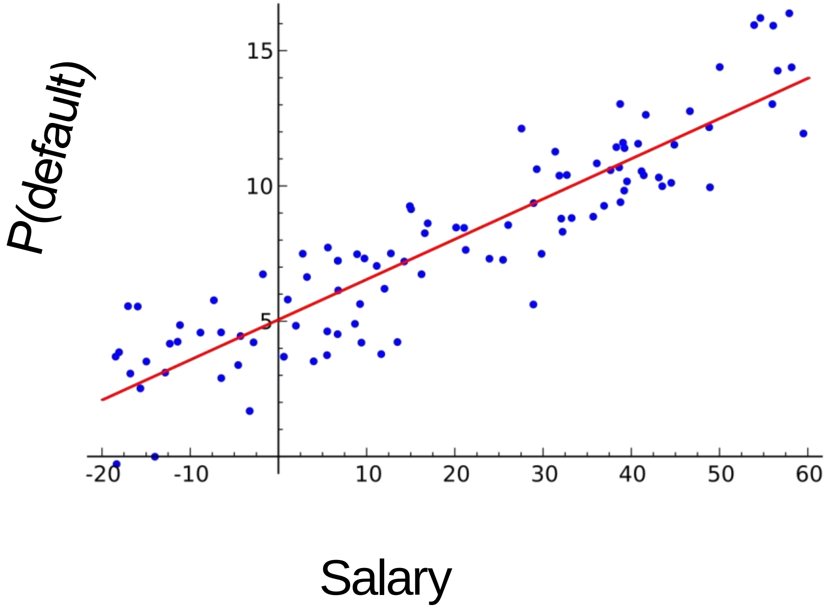
Observation Number	Temperature (x_i)	Yield (y_i)
1	50	122
2	53	118
3	54	128
4	55	121
5	56	125
6	59	136
7	62	144
8	65	142
9	67	149
10	71	161
11	72	167
12	74	168
13	75	162
14	76	171
15	79	175
16	80	182
17	82	180
18	85	183
19	87	188
20	90	200
21	93	194
22	94	206
23	95	207
24	97	210
25	100	219

Data

Learning Algorithm



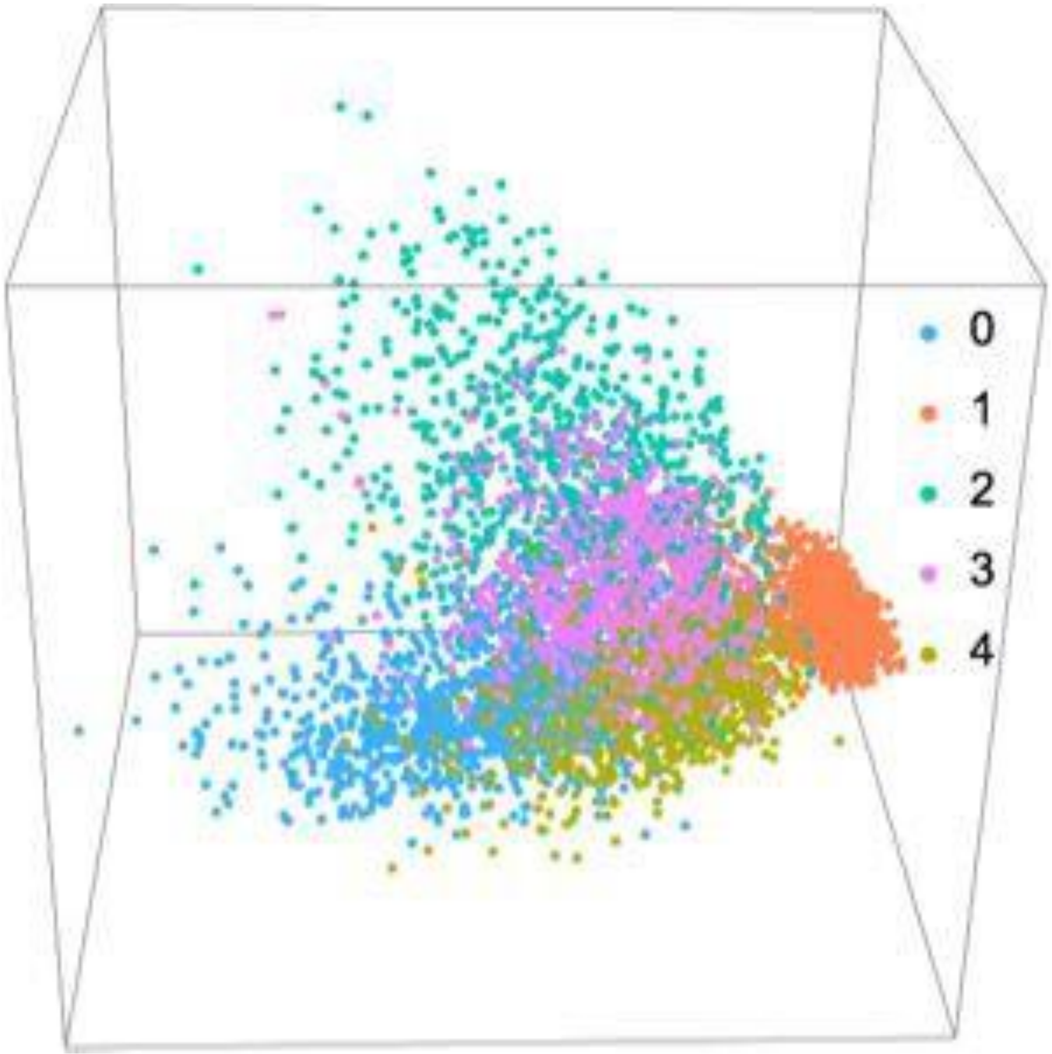
Model



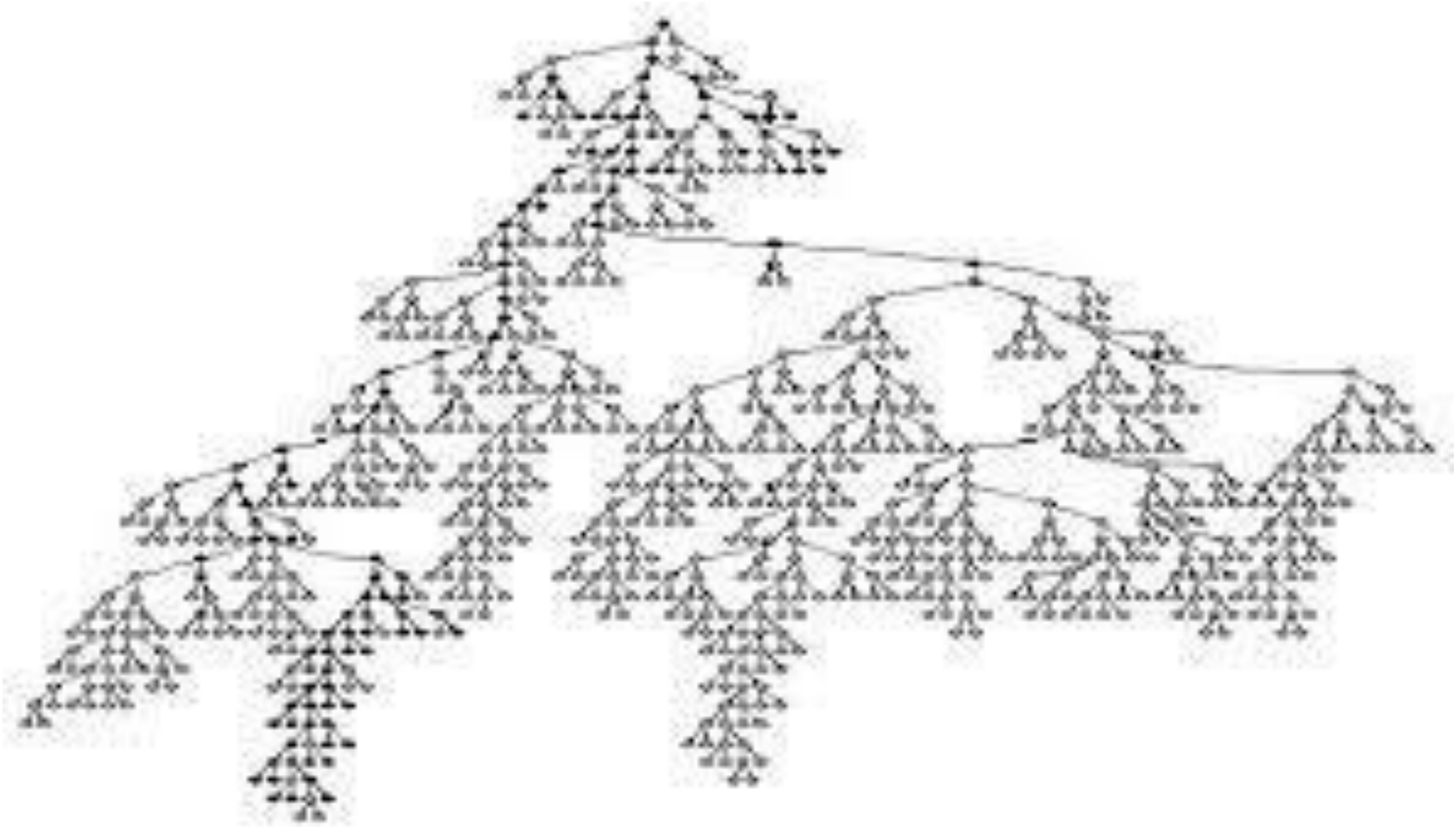
->

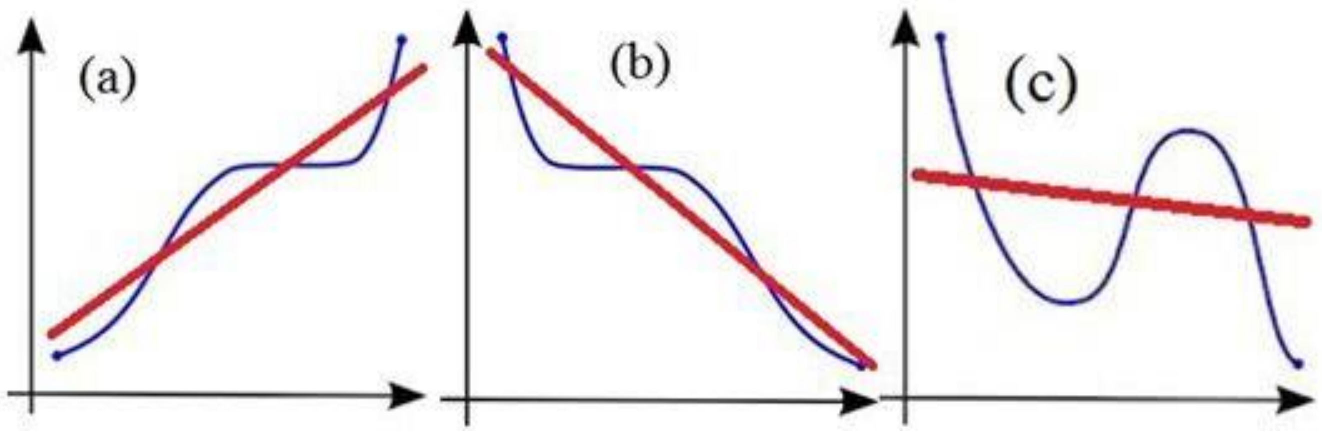
If P(default) > threshold, *then* deny credit

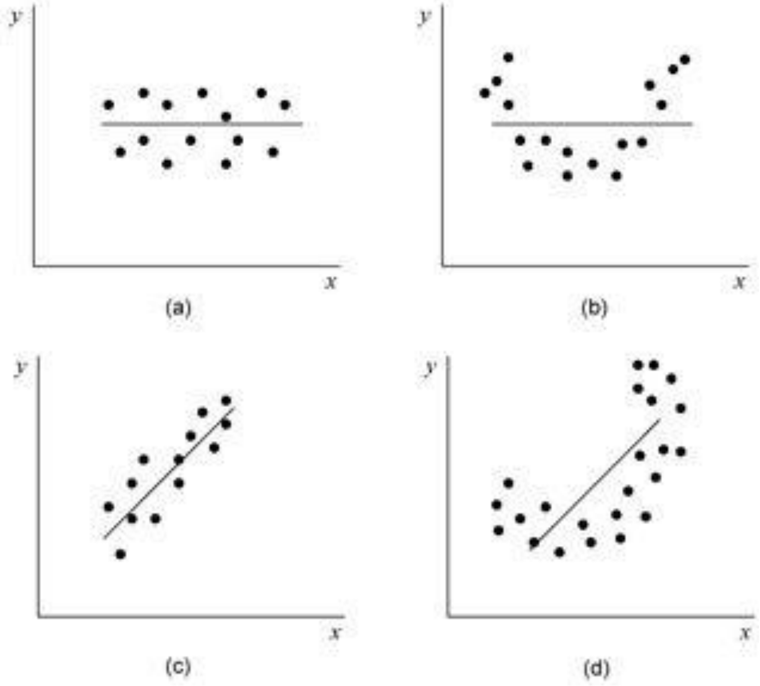
High-dimensionality



Complexity



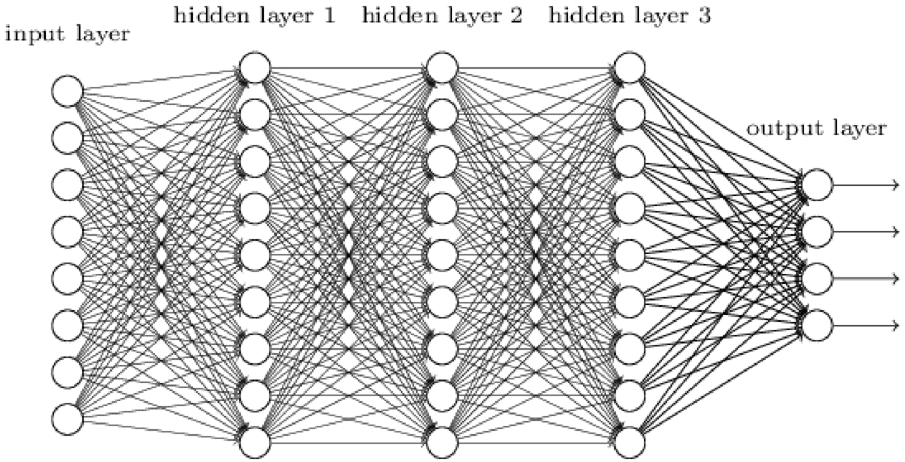




- Non-linear
- Non-monotonic

‘Deep learning’

features: { 1,1 = black, 1,2 = brown, 1,3 = grey …}



hidden layers: {?}

…labrador?

Bias, error, discrimination in statistical models

- False positives vs false negatives
- Fitting to the majority population
- Reflecting (and compounding) structural discrimination

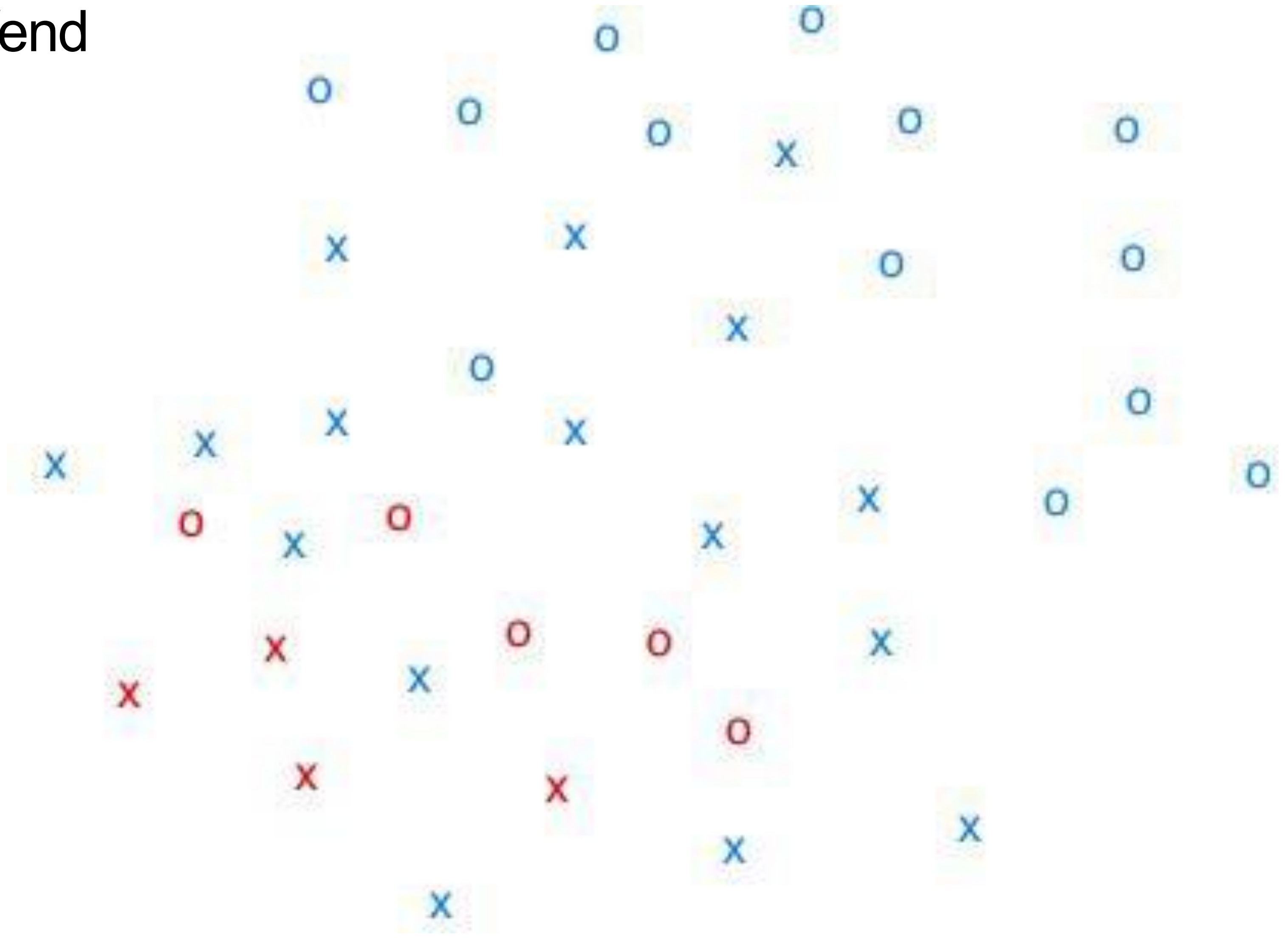
Bias, error, discrimination in statistical models

- **False positives vs false negatives**
- Fitting to the majority population
- Reflecting (and compounding) structural discrimination

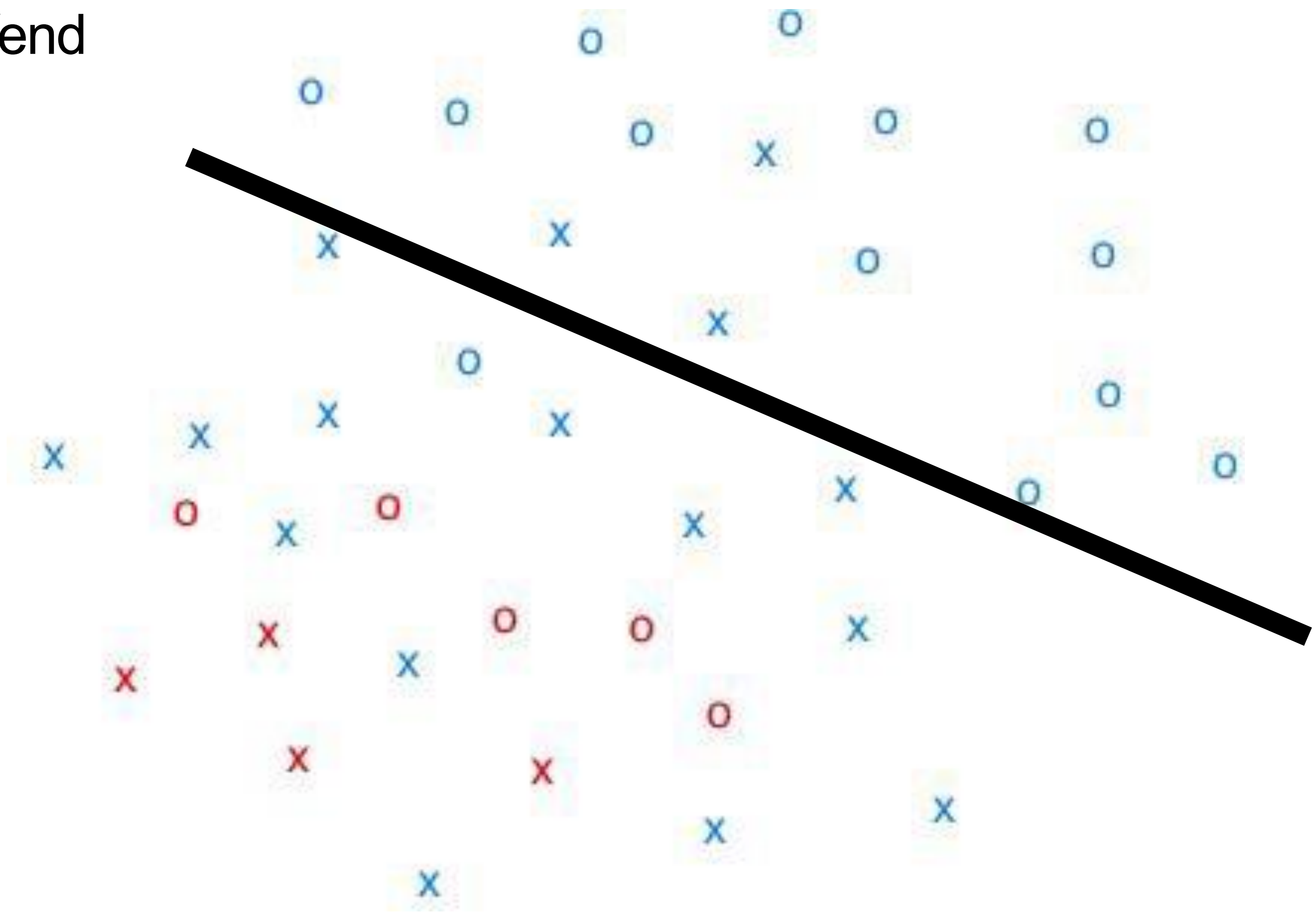
False Positive:
the boy cried wolf… but no wolf

False Negative:
The villagers thought ‘no wolf’
… but wolf!

x = reoffend
o = not reoffend

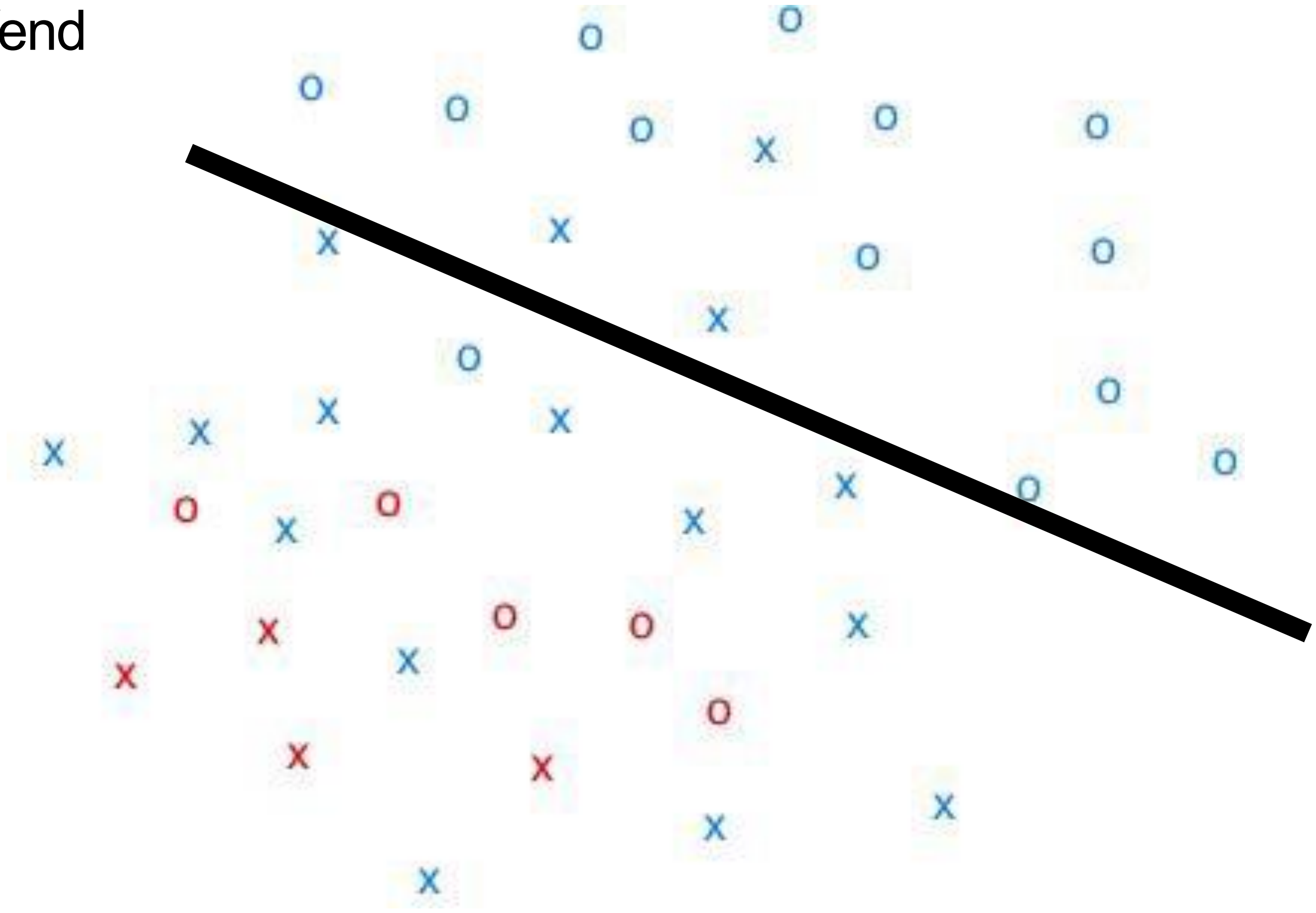


x = reoffend
o = not reoffend



x = reoffend
o = not reoffend

False positive rate = 3/14 = 21%



False negative rate = 6/24 = 25%

Better that ten guilty persons escape than that one innocent suffer

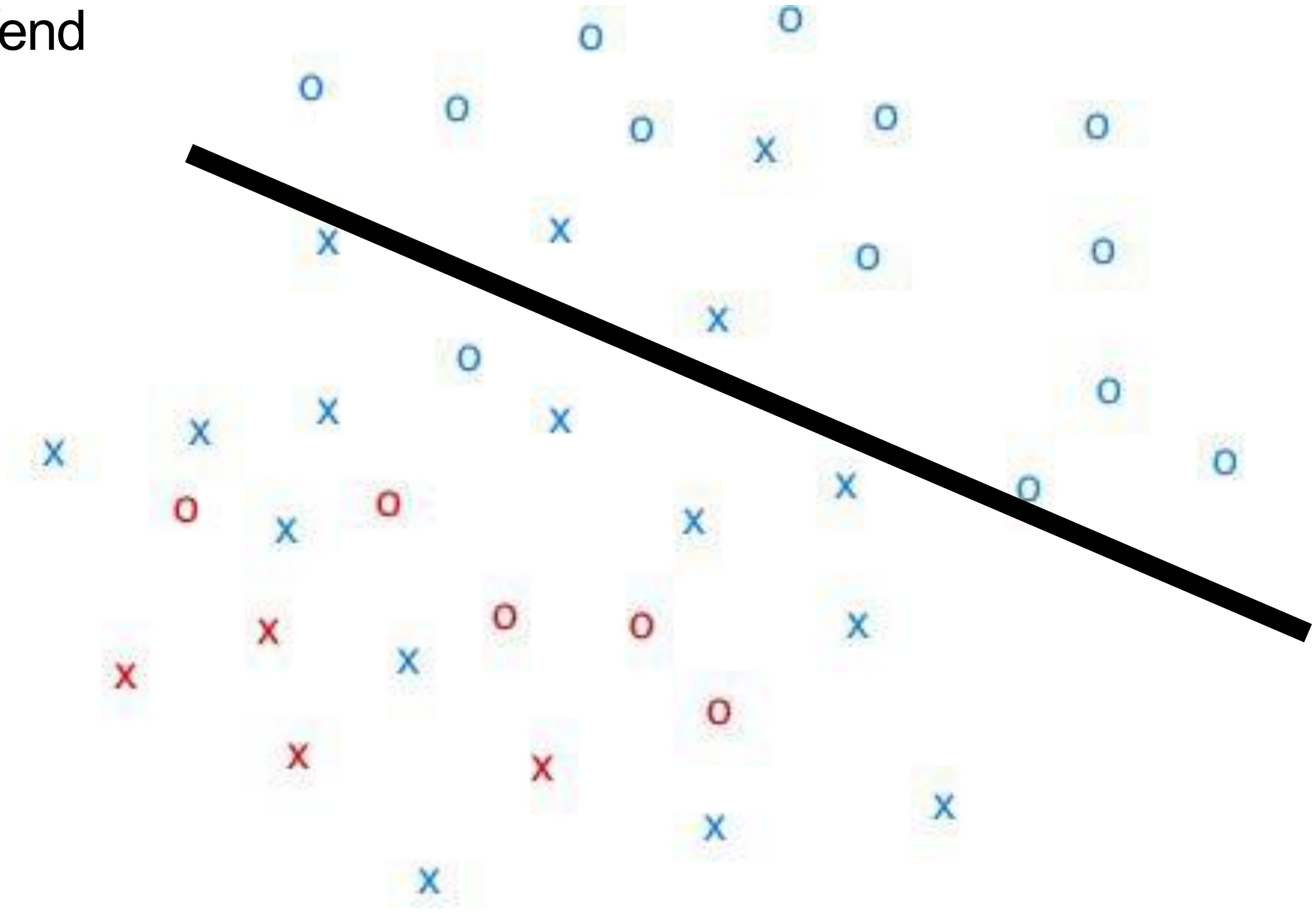
— *Sir William Blackstone (1765)*

Sir William Blackstone by Paul Wayland Bartlett - Washington, D.C.

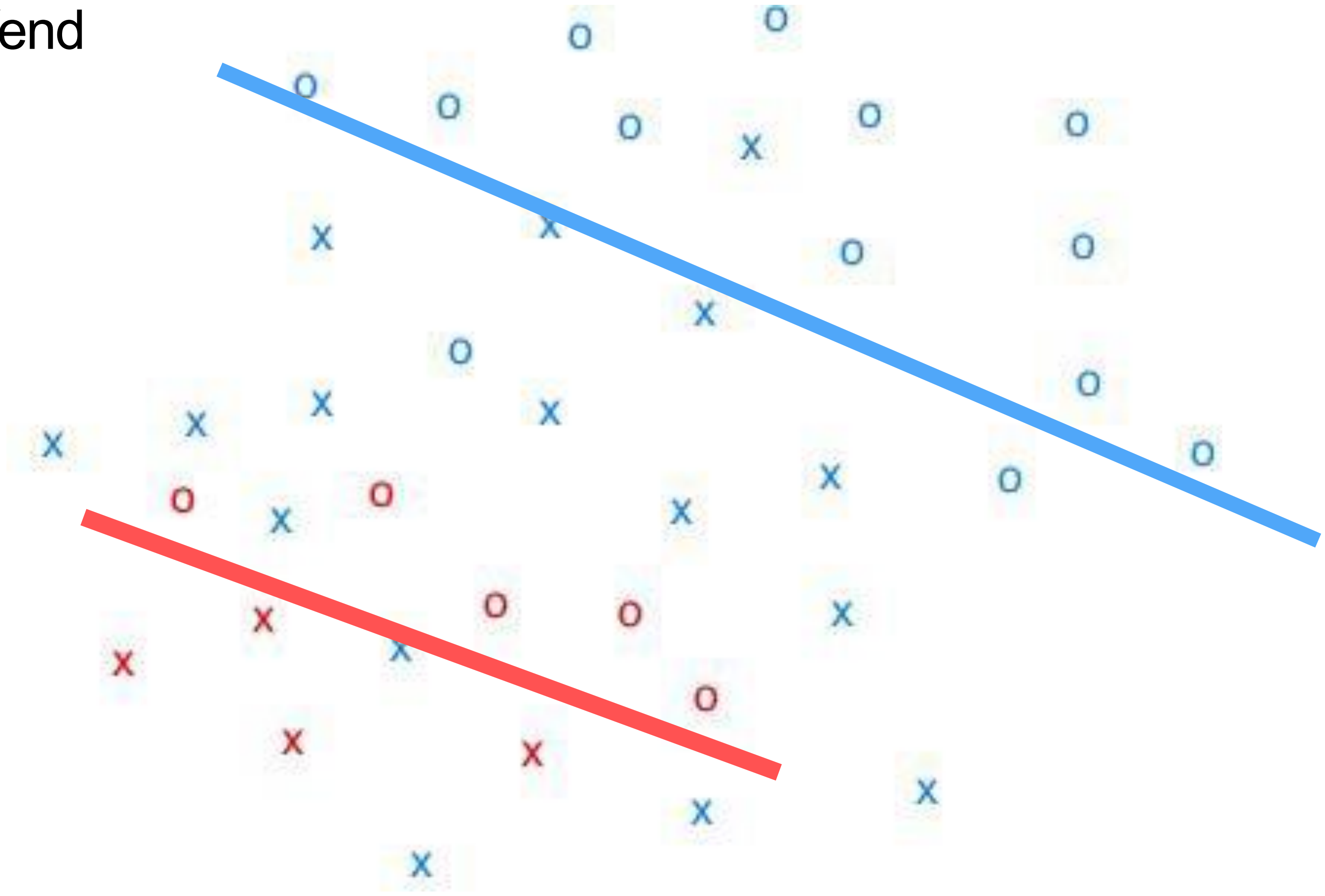
Bias, error, discrimination in statistical models

- False positives vs false negatives
- **Fitting to the majority population**
- Reflecting (and compounding) structural discrimination

x = reoffend
o = not reoffend



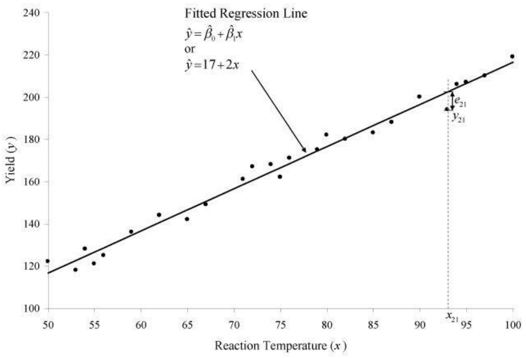
x = reoffend
o = not reoffend



Bias, error, discrimination in statistical models

- False positives vs false negatives
- Fitting to the majority population
- **Reflecting (and compounding) structural discrimination**

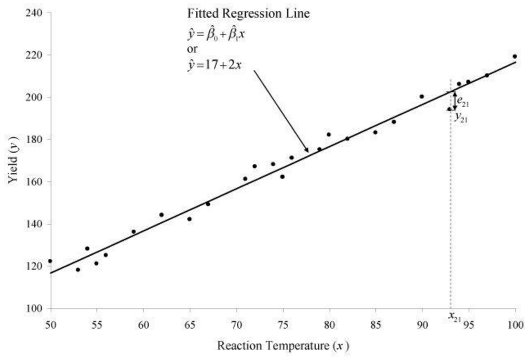
prior offence



Model

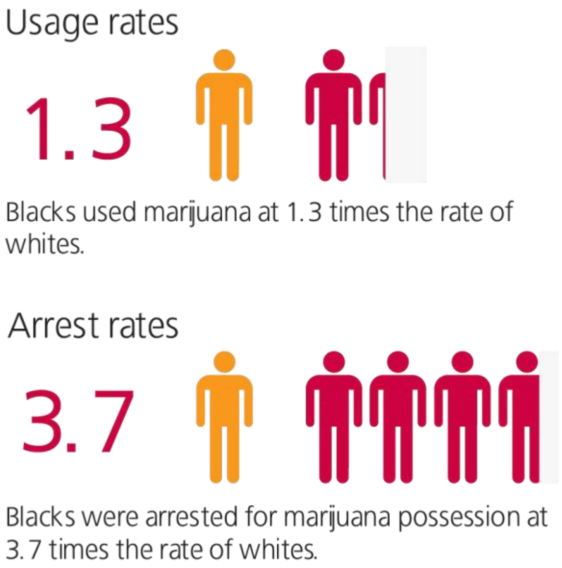
re-offend

prior arrests

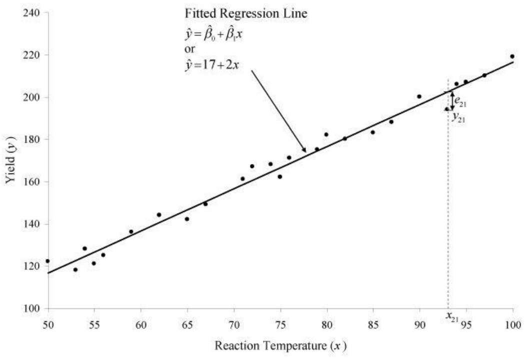


Model

re-arrest



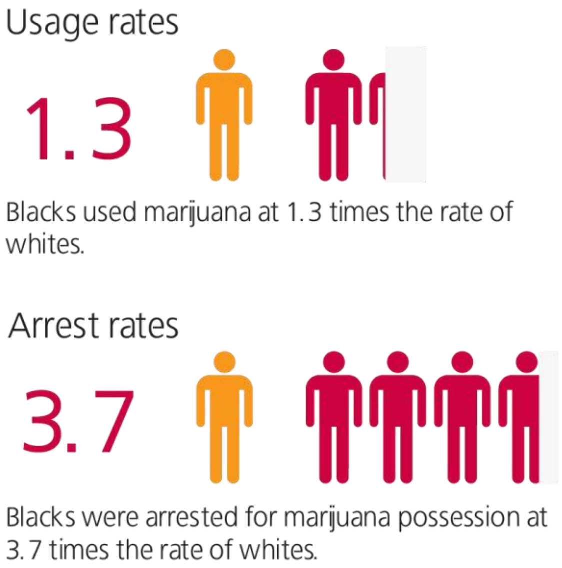
prior arrests



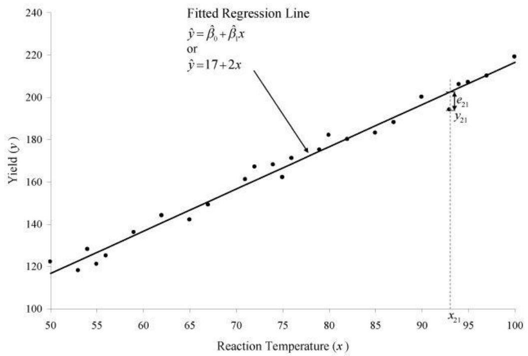
Model

re-arrest

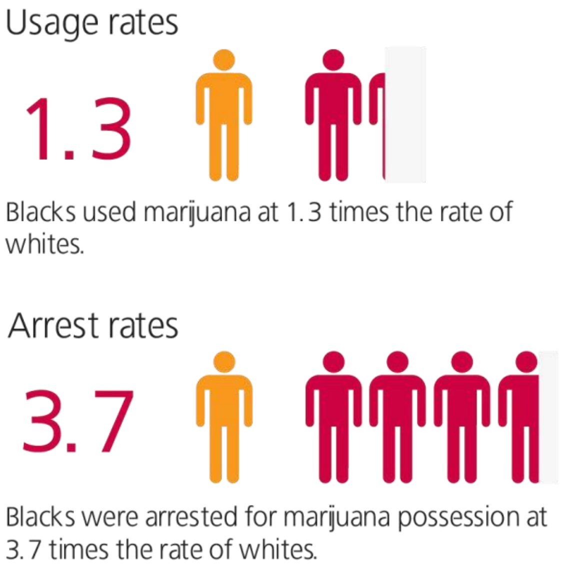
Report: The War on Marijuana in Black and White, ACLU
https://www.aclu.org/report/report-war-marijuana-black-and-white?redirect=criminal-law-reform/war-marijuana-black-and-white



prior arrests



Model



re-arrest

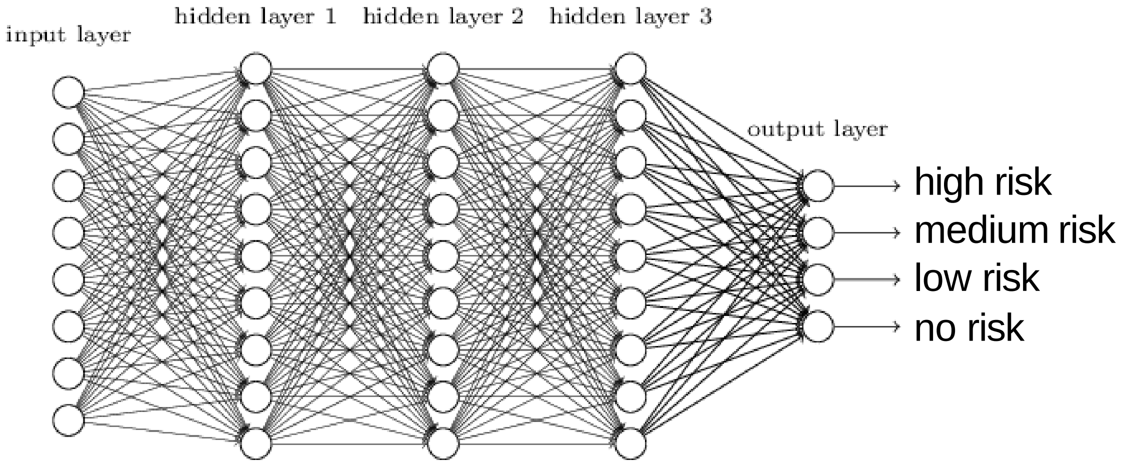
Report: The War on Marijuana in Black and White, ACLU
https://www.aclu.org/report/report-war-marijuana-black-and-white?redirect=criminal-law-reform/war-marijuana-black-and-white

"black people are now nine times more likely to be stopped and searched for drugs despite using drugs at a lower rate than white people"

https://www.release.org.uk/publications/ColourOfInjustice

Applicant	Monthly Income	Age	Default?
A	$1800	34	No
B	$600	21	Yes
C	$350	84	No
D	$1100	46	No
E	$2100	39	Yes
...	...	...	...

features: { qualifications, postcode, place of birth, occupation, behavioural data, …}



latent features: {?}

{ ~gender?, ~ethnicity? }

Parity of errors between protected classes

protected groups receive equal proportion of errors

Model performance on male applicants

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Model performance on female applicants

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Parity of calibration between protected classes

Calibration: of those given a particular risk score S, S% should result in the predicted outcome.

Calibration should be equal between protected groups

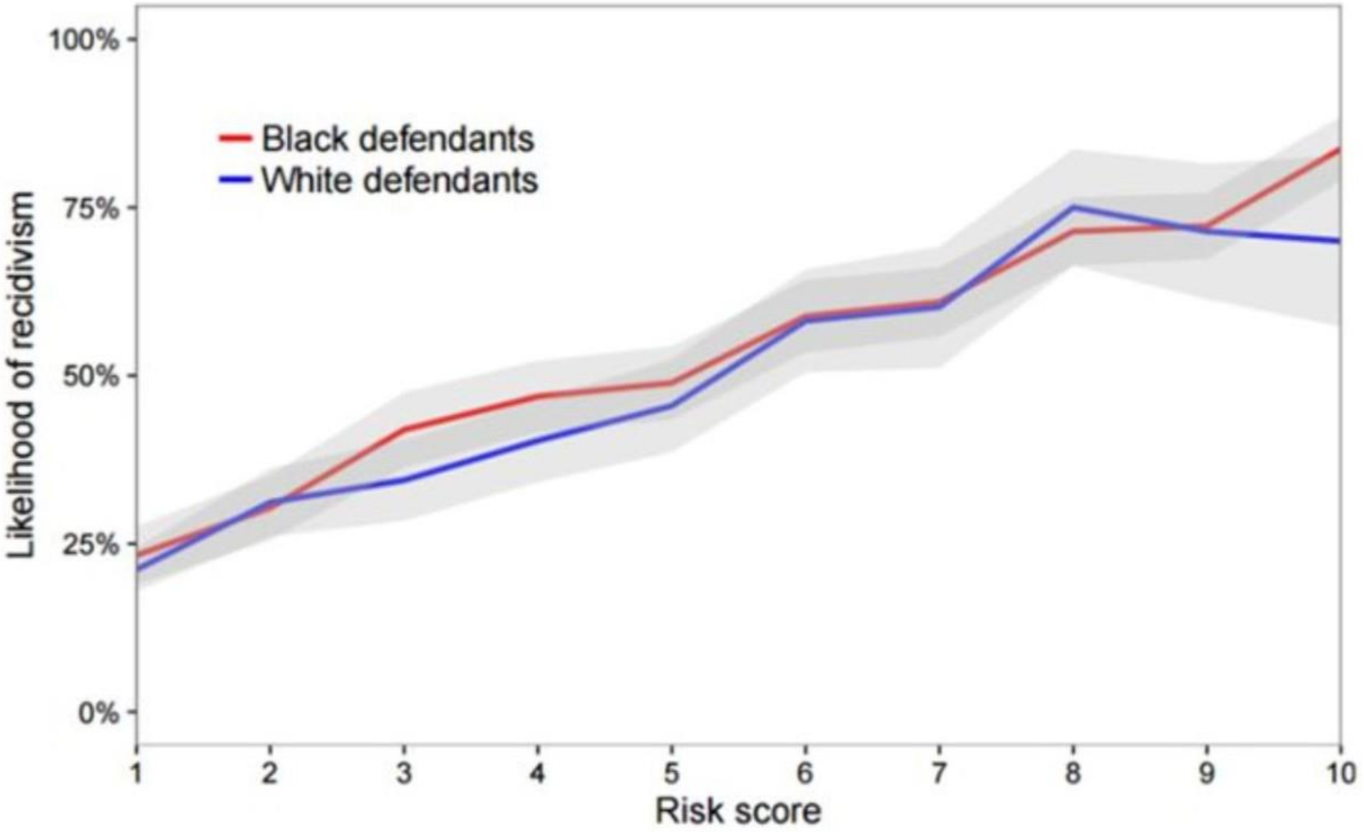


Image from "Defining and Designing Fair Algorithms"
Sam Corbett-Davies and Sharad Goel Stanford University. EC18 Fairness tutorial

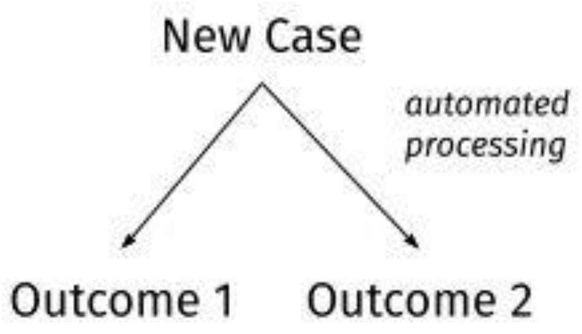
Roles for automated decision-making

Decision **support** vs **full automation**

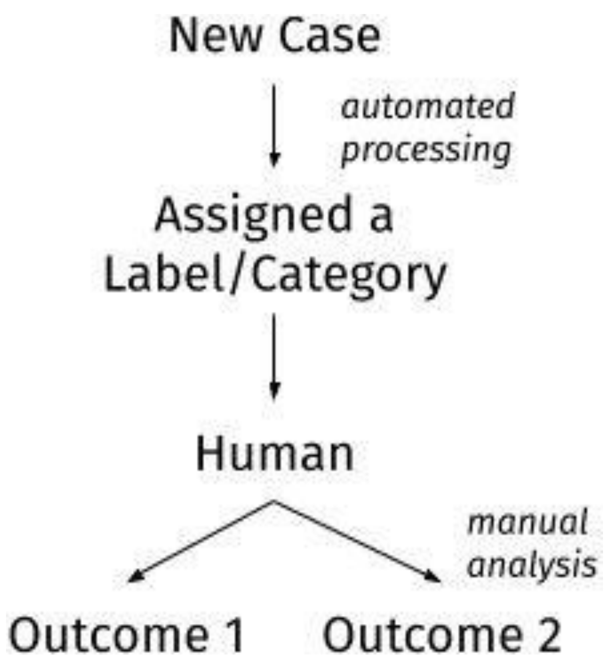
- **Decision support**: providing additional information, informed by statistical or rules-based systems, to aid a human decision-maker in their decision.
  - E.g. a risk score presented to a parole officer to inform their assessment of an offender
- **Fully automated**: the system takes a decision and action in relation to a person or group without human input.
  - E.g. a visa application is automatically assessed and approved

NB: implications for data protection (GDPR Article 22 'solely automated' decisions)

Fully automated



Decision support



Automation bias

Human decision-makers may either systematically:

Under-rely on computer outputs, ignoring good information

Over-rely on computer outputs, ignoring their own judgement and supplemental information from other sources

Daniel Schwen / Wikimedia Ccmmons. Boeing 787 cockpit at the Museum of Flight near Seattle

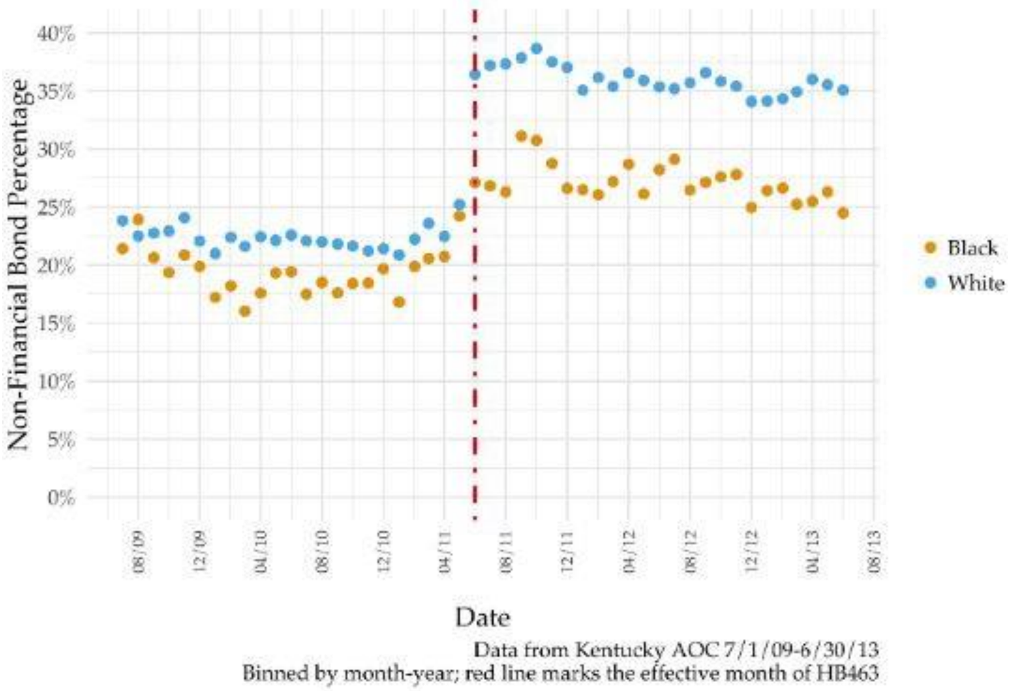
Unequal application of discretion

Under-reliance and over-reliance might be applied unequally between different groups.

Even if the algorithm is not biased, the way that human decision-makers use it may interact with existing prejudice / bias

See introduction of COMPAS in US (Albright (2019), Cowgill (2019))

Figure 11: Bond Outcomes Before and After HB463 by Race

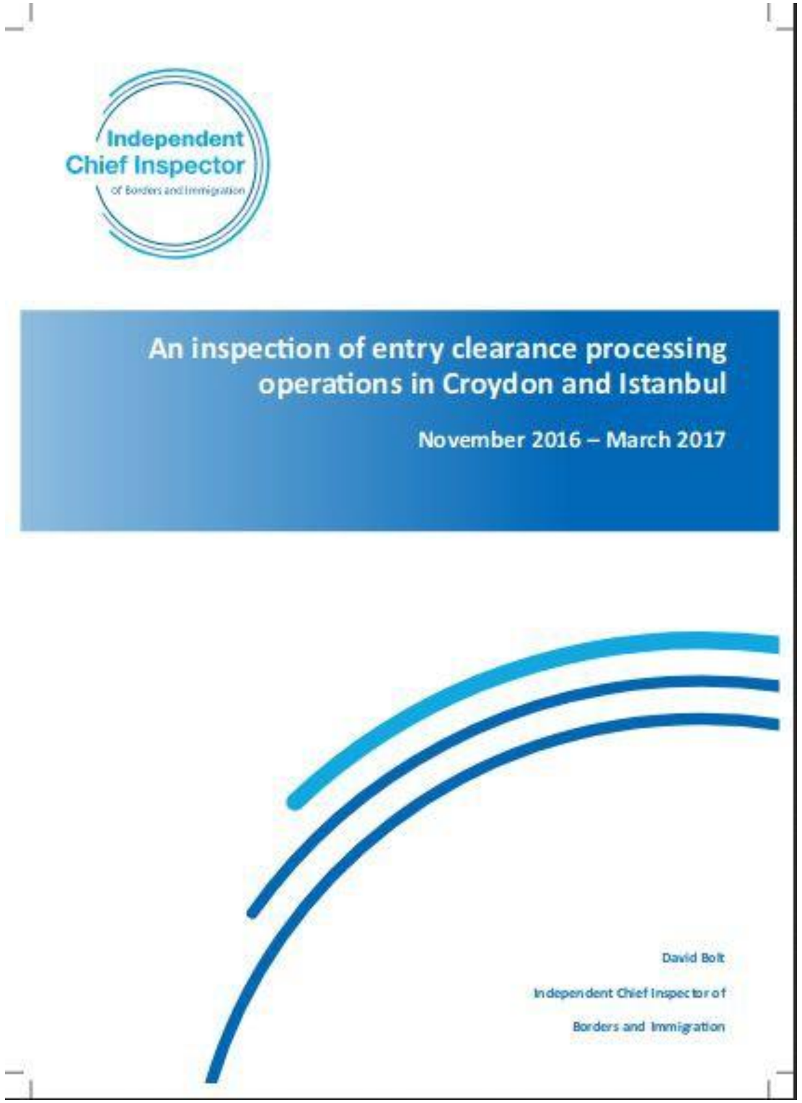


Alex Albright. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions.The John M. Olin Centerfor Law, Economics, and Business Fellows' Discussion Paper Series85 (2019).

Unequal application of discretion

An initial ADM stage may determine *which* human decision makers make the assessment

Even if no decision is taken without a human, the algorithmic step determines the type and quality of human judgement

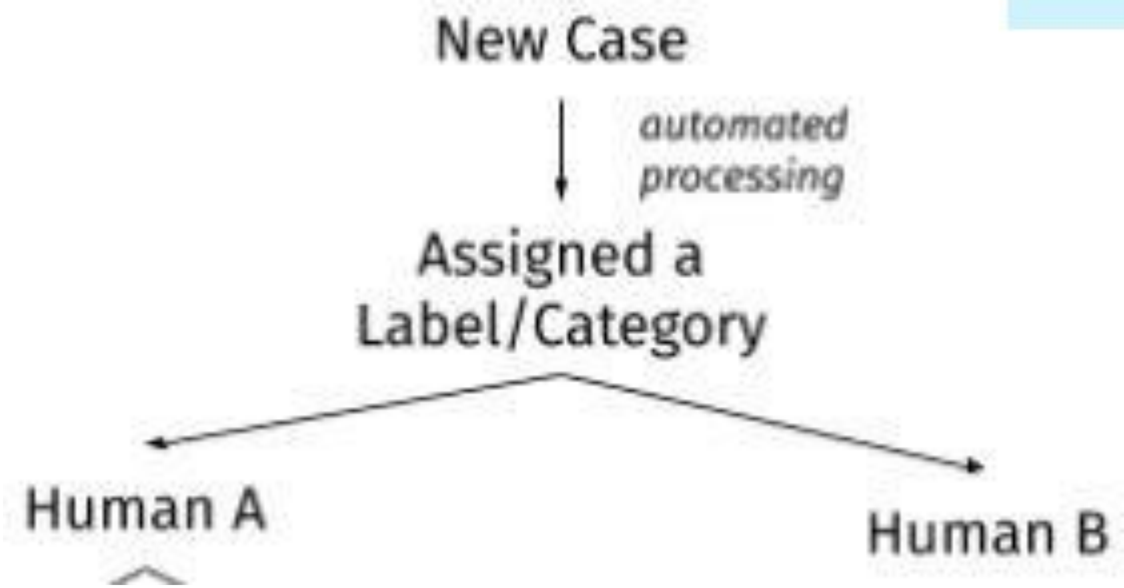


Independent Chief Inspector of Borders and Immigration, 'An inspection of entry clearance processing operations in Croydon and Istanbul' https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/631520/An-inspection-of-entry-clearance-processing-operations-in-Croydon-and-Istanbul1.pdf

Unequal application of discretion

Figure 4: Decisions and streaming tool ratings for Croydon visit visa applications 1 January to 28 February 2107

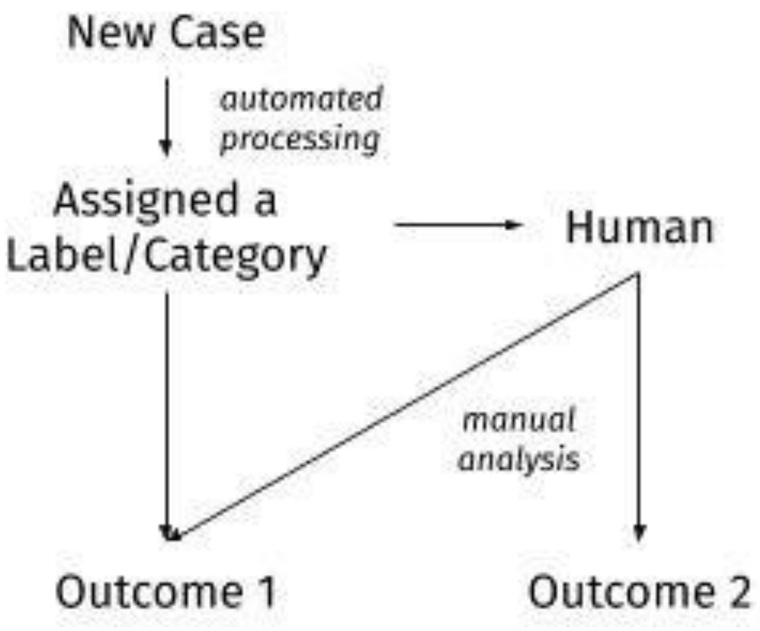
Streaming rating	Applications	Percentage issued	Percentage refused
Green	13,560	96.36%	3.64%
Amber	3,662	81.08%	18.92%
Red	6,421	48.59%	51.41%

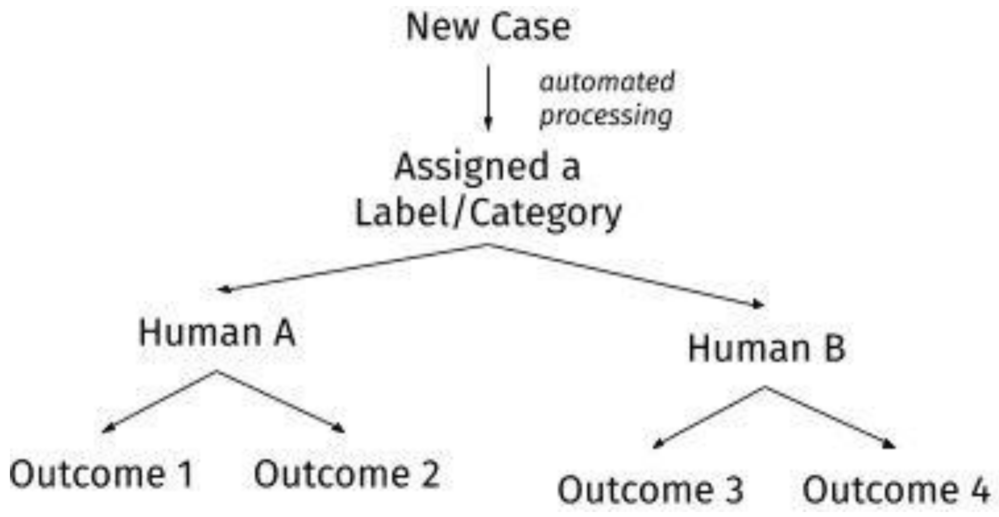


Upstream automation may fetter downstream discretion

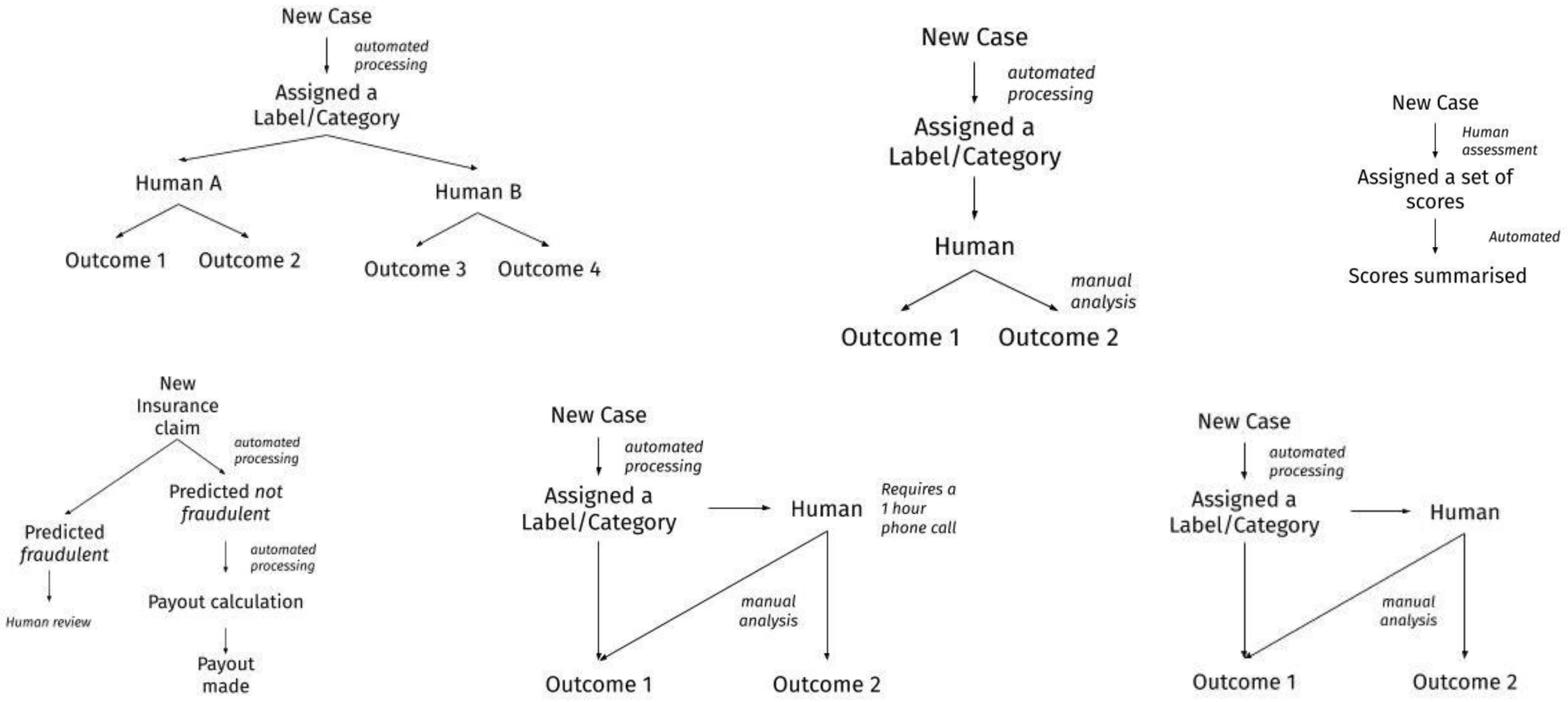
Figure 5: Daily benchmarks for deciding visit applications

	Streaming Rating			
Location	**Super Green**	**Green**	**Amber**	**Red**
Croydon	N/A	75	35	25
Istanbul	100	70	35	30





Where is the decision? Who /what made it?



Thanks!

Reuben Binns

reuben.binns@cs.ox.ac.uk

Twitter: @RDBinns