FairOD: Fairness-aware Outlier Detection

Shubhranshu Shekhar

Neil Shah

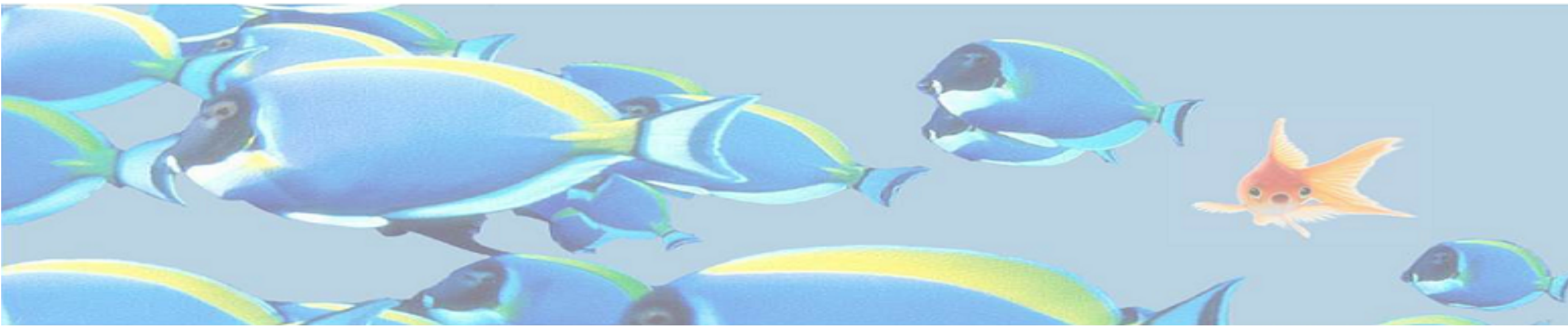
Leman Akoglu

Fourth AAAI /ACM Conference on

Artificial Intelligence, Ethics, and Society

AIES'21

What is an outlier?

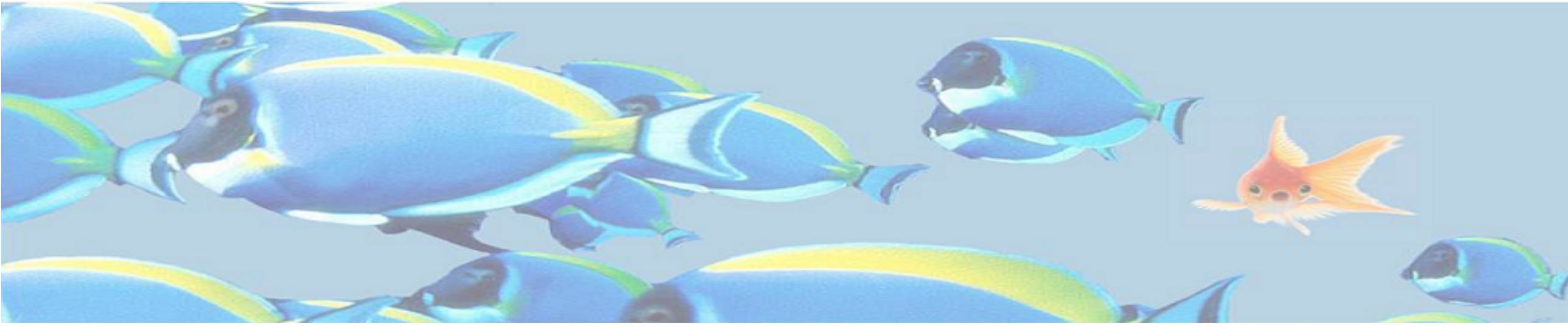


Slide Courtesy:http://www.andrew.cmu.edu/user/lakoglu/index.html

What is an outlier?

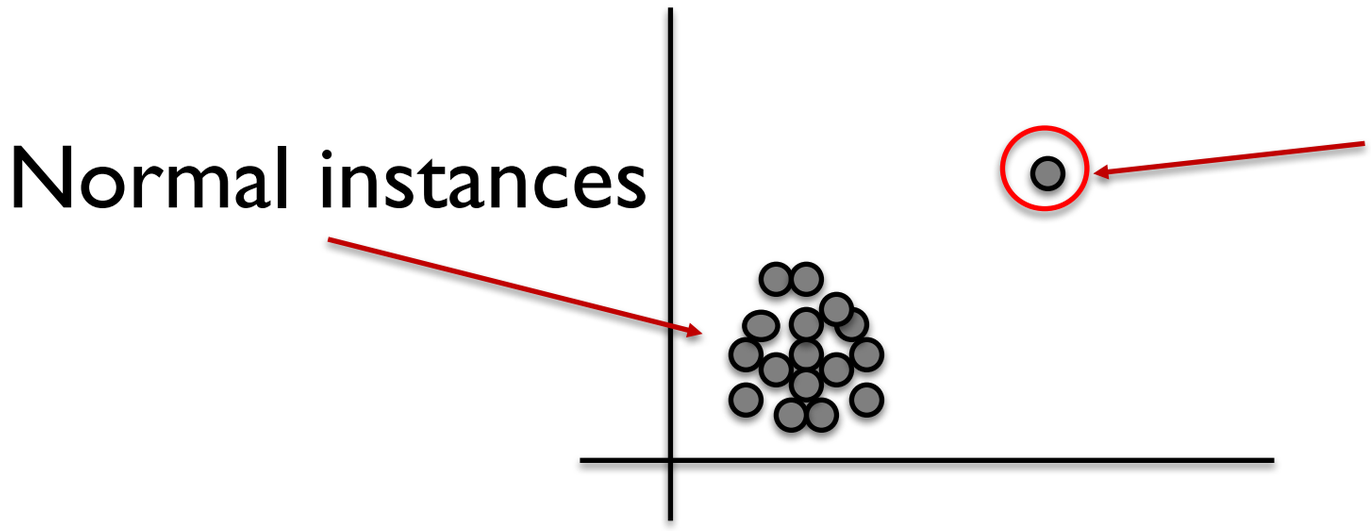
Observations that…

- "…are **inconsistent** with the remainder…" [Barnett&Lewis'94]
- "… deviate so much … as to arouse suspicions … they were generated by a **different mechanism**" [Hawkins '80]
- "… **deviate markedly** from other members of sample in which it occurs" [Grubbs '69]



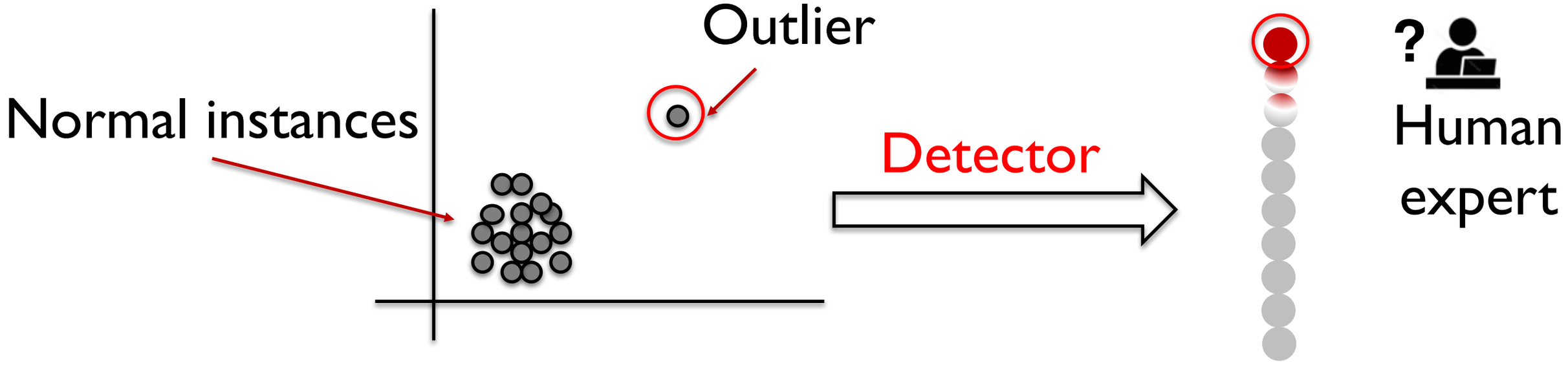
Slide Courtesy:http://www.andrew.cmu.edu/user/lakoglu/index.html

Outlier Detection



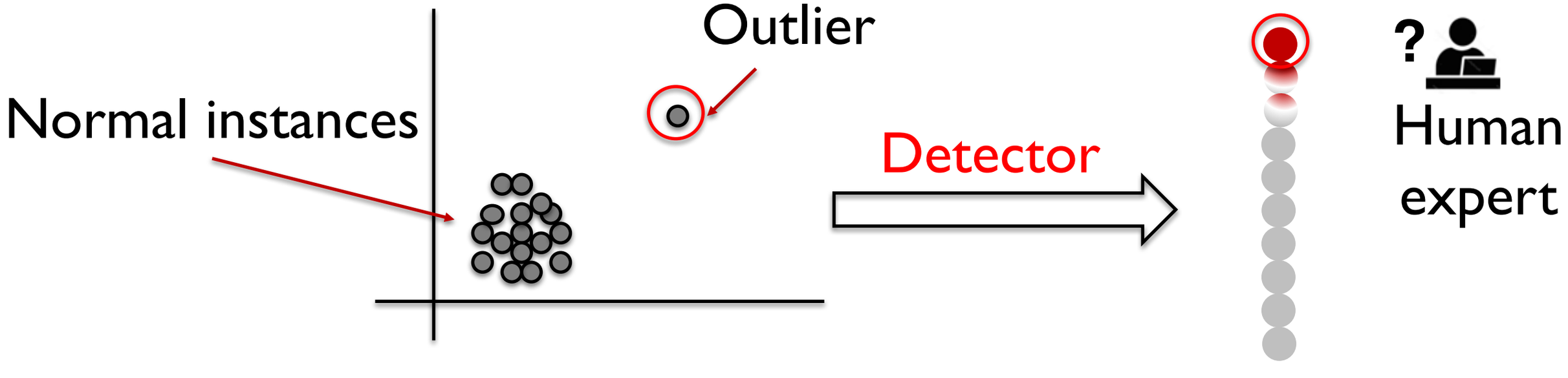
Inconsistent with normal observations

Outlier Detection



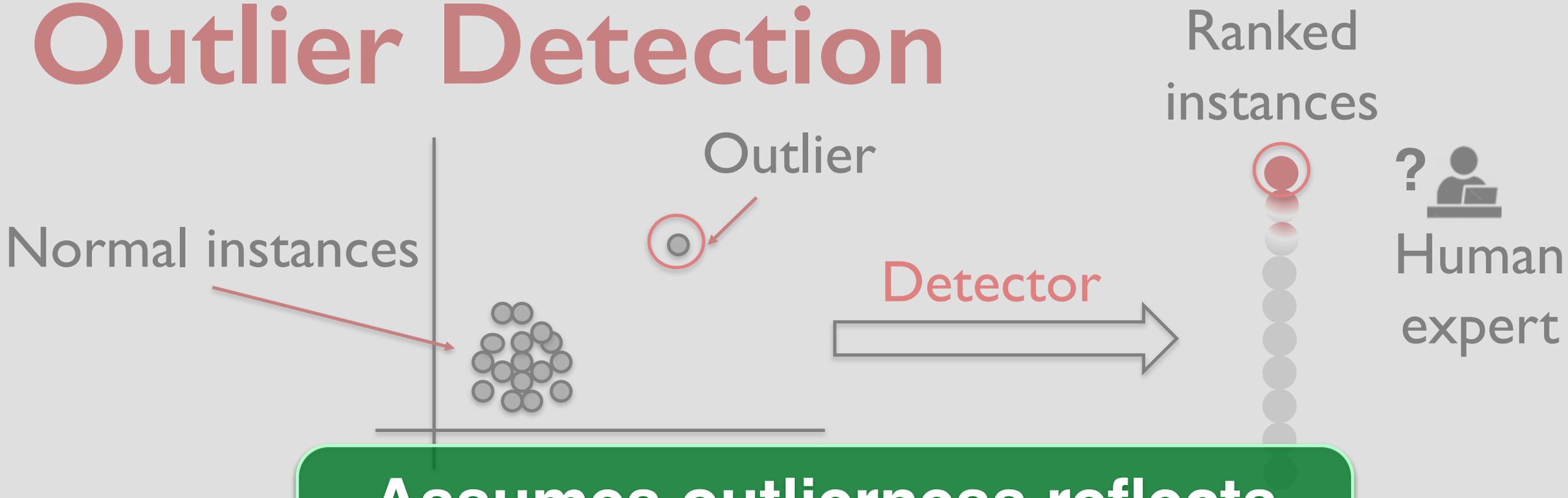
- designed to spot/flag rare, minority samples
  - e.g. suspicious activity, abnormal heart rate etc.

Outlier Detection



- designed to spot/flag rare, minority samples
  - e.g. suspicious activity, abnormal heart rate etc.
- facilitates auditing ("*policing*") by human experts
  - e.g. stop-and-frisk in automated surveillance flagged instances
  - human labeled data for downstream learning tasks

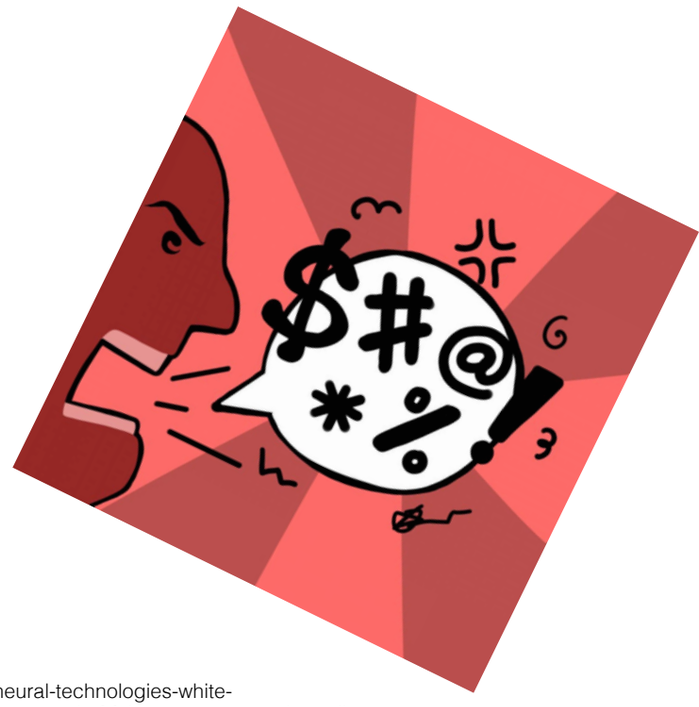
Outlier Detection



Assumes outlierness reflects true *riskiness.*

- designed to spot/flag rare, anomalous samples
  - e.g. suspicious activity, abnormal heart rate etc.
- facilitates auditing ("*policing*") by human experts
  - e.g. stop-and-frisk in automated surveillance flagged instances
  - human labeled data for downstream learning tasks

Outlier Detection: Use-cases

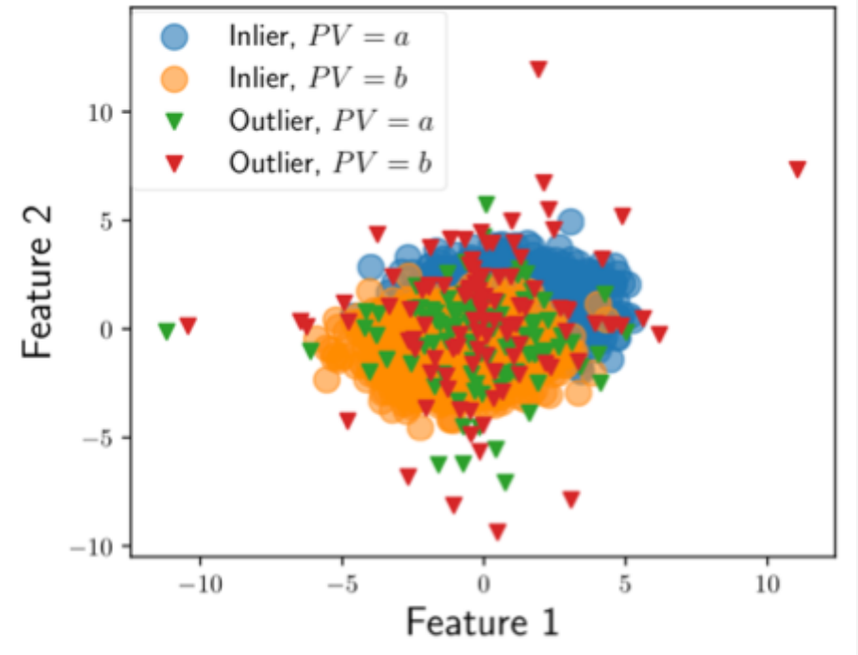


Sources: https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6, https://www.google.com/url?q=https://www.the-digital-insurer.com/insurance-fraud-digital-age-neural-technologies-white-paper/&sa=D&source=hangouts&ust=1620381203046000&usg=AFQjCNGpeSoWM0xriR0YhGq3vXzrhdisLg,, https://www.google.com/url?q=https://www.internetmatters.org/hub/news-blogs/stopping-the-spread-of-fake-news-on-popular-online-platforms/&sa=D&source=hangouts&ust=1620381203046000&usg=AFQjCNHTmHYACxrqcOX0A-vTMcTpM3_Fxw , https://www.investopedia.com,, https://traderdefenseadvisory.com/,, https://www.google.com/url?q=https://blog.volkovlaw.com/2015/01/healthcare-fraud-aggressive-enforcement-strategies/&sa=D&source=hangouts&ust=1620386116751000&usg=AFQjCNGw2wgs6uMWfIB8D2L6qXeJWPnibg,

Roadmap

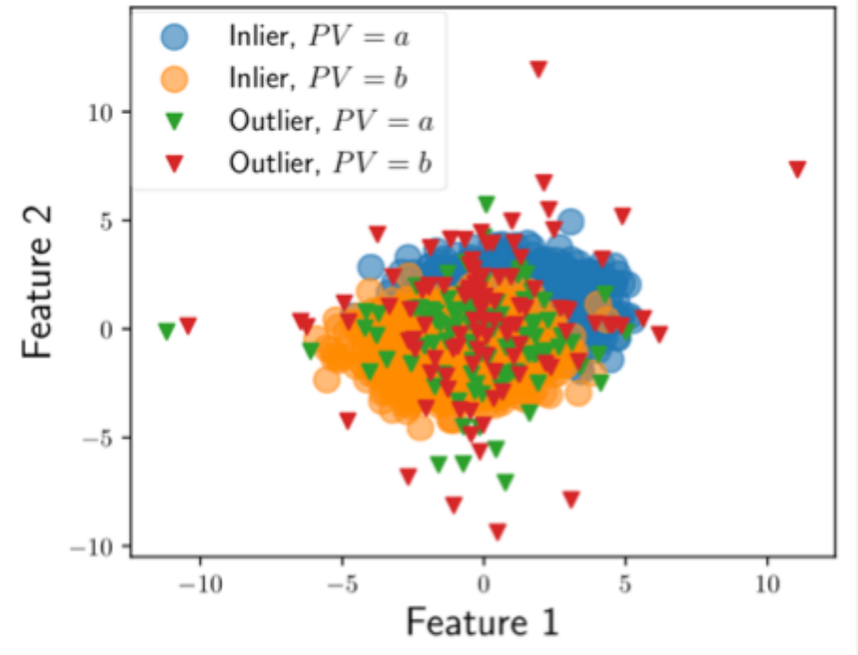
- Introduction
- Problem: Fairness in OD
- Desiderata
- Fairness aware OD
- Evaluation

Bias in Outlier Detection

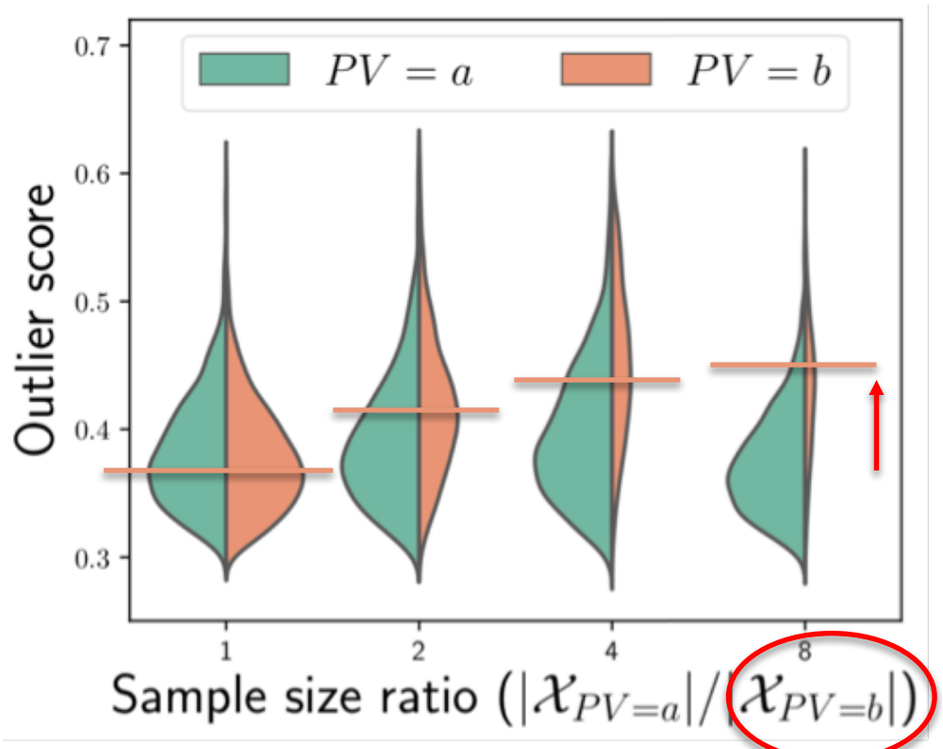


- Simulated dataset
  - equal sized groups
  - groups induced by $\mathrm{PV} = \mathrm{a}$ and $\mathrm{PV} = \mathrm{b}$

Bias in Outlier Detection

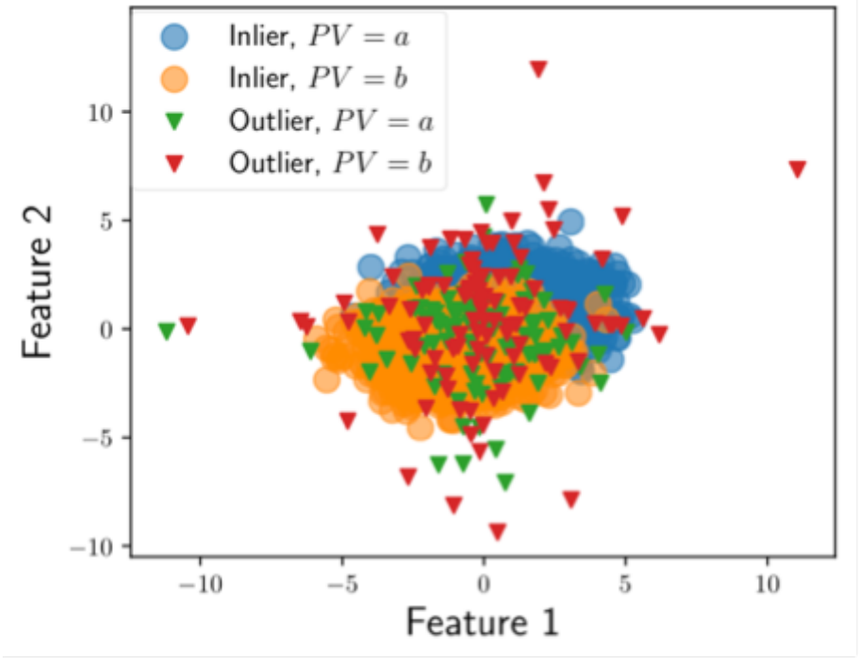


- Simulated dataset
  - equal sized groups
  - groups induced by $\mathrm{PV} = \mathrm{a}$ and $\mathrm{PV} = \mathrm{b}$



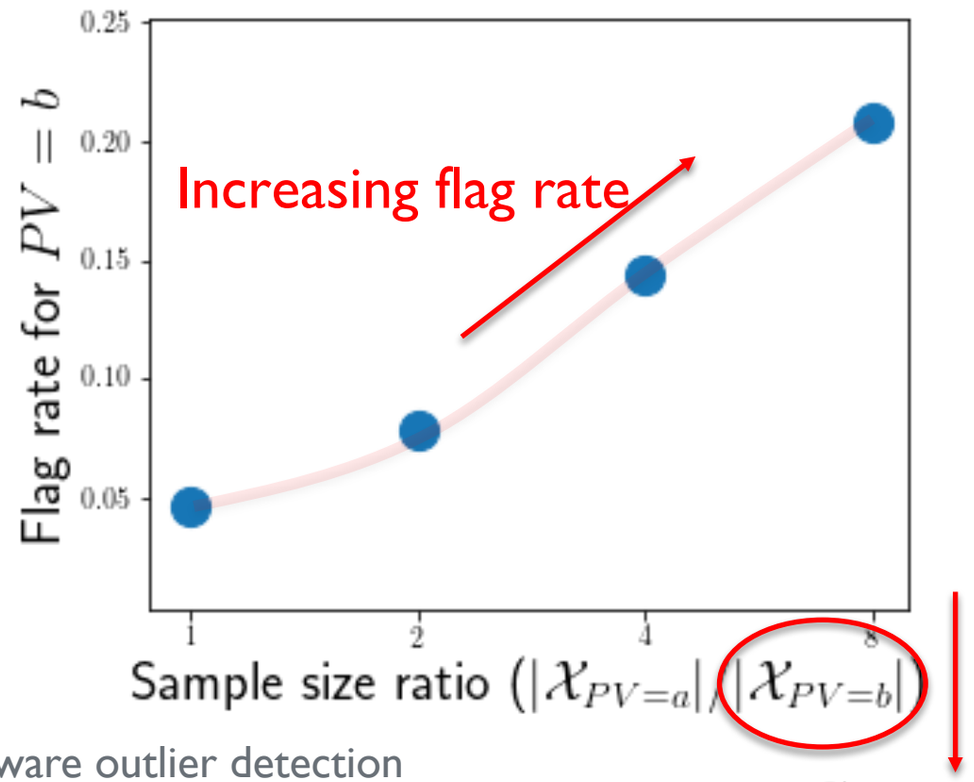
Higher outlier scores as sample size of $PV = b$ is decreased

Bias in Outlier Detection



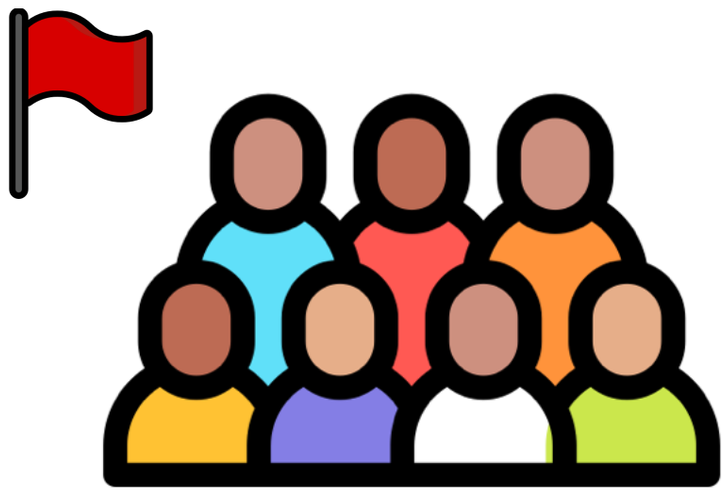
- Simulated dataset
  - equal sized groups
  - groups induced by $\mathrm{PV} = \mathrm{a}$ and $\mathrm{PV} = \mathrm{b}$

Corresponding flag rate for $PV = b$ increases



Bias in Outlier Detection

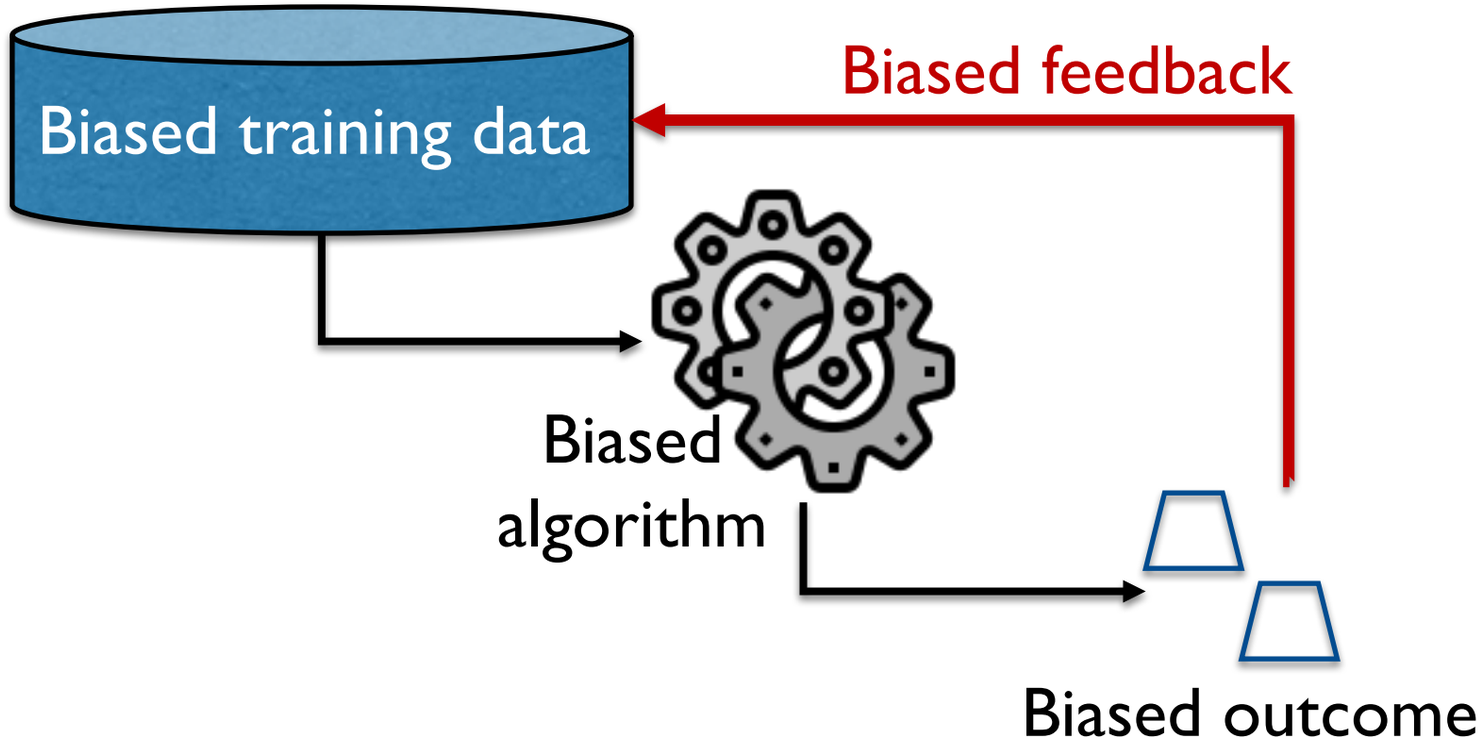
- Societal minorities may be statistical minorities
  - defined by protected variable (PV) race/ ethnicity/gender/age etc.



≠ riskiness

Bias in Outlier Detection

- Disparate Impact
  - unjust flagging leads to over-policing
  - feedback loop results in further skewness



Fair Outlier Detection

- Given:
  - Observations $\mathcal{X} = \{X_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$
  - $\mathcal{PV} = \{PV_i\}_{i=1}^{N},\ PV_i \in \{a, b\}$
    - $PV_i = a$ identifies majority group

- Build a **detector** that estimates outlier scores $\mathcal{S}$ and assigns outlier labels $\mathcal{O}$ s.t.

i. assigned labels and scores are "fair" w.r.t. the PV

ii. higher scores correspond to higher riskiness encoded by the underlying (unobserved) true labels $\mathcal{Y}$

Fair Outlier Detection

- Given:
  - Observations $\mathcal{X} = \{X_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$
  - $\mathcal{PV} = \{PV_i\}_{i=1}^{N},\ PV_i \in \{a, b\}$
  - $PV_i = a$ identifies majority group

What constitutes a "fair" outcome in OD?

- Build a detector that estimates outlier scores $\mathcal{S}$ and assigns outlier labels $\mathcal{O}$ s.t.

i. assigned labels and scores are "fair" w.r.t. the PV

ii. higher scores correspond to higher riskiness encoded by the underlying (unobserved) true labels $\mathcal{Y}$

Roadmap

- Introduction
- Problem: Fairness in OD

- Desiderata
- Fairness aware OD
- Evaluation

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity

D5. Base rate preservation

detection performance

fairness related

Proposed Desiderata

D1. Detection effectiveness - accurate at detection

$$P\ (Y = 1 \mid O = 1) > P\ (Y = 1)$$

- related to detection performance

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity – decision avoids use of PV

$$P(O{=}1|X) = P(O{=}1|X,PV{=}v), \forall\, v$$

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity – decision avoids use of PV

$$P(O{=}1|X) = P(O{=}1|X,PV{=}v), \forall\, v$$

- may allow discriminatory OD results for minority

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP) – decision independent of PV

$$P(O{=}1|PV{=}a) = P(O{=}1|PV{=}b)$$

Proposed Desiderata

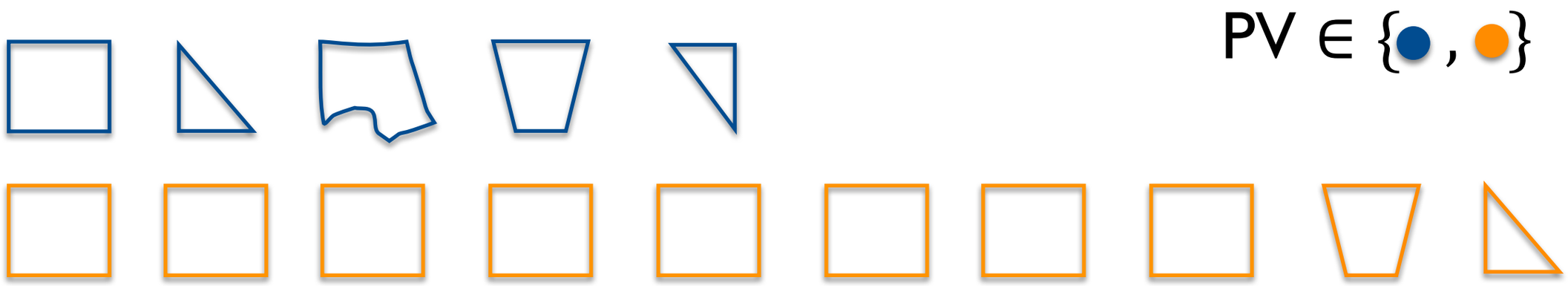
D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP) – decision independent of PV

$$P(O{=}1|PV{=}a) = P(O{=}1|PV{=}b)$$

➢ permits *"laziness"* [Barocas et al.'2017]



Proposed Desiderata

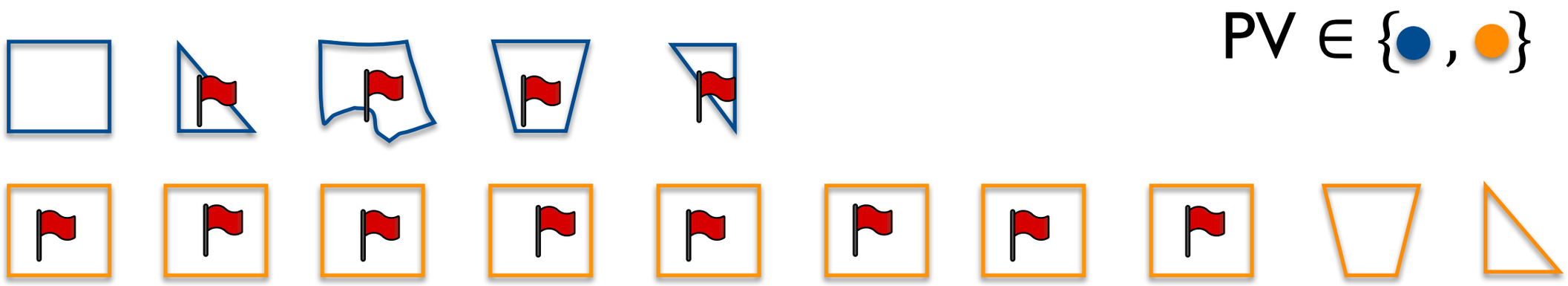
D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP) – decision independent of PV

$$P(O{=}1|PV{=}a) = P(O{=}1|PV{=}b)$$

➢ permits *"laziness"* [Barocas et al.'2017]



Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity – decision faithful to ground-truth

$$P(O=1|Y=1,PV=a) = P(O=1|Y=1,PV=b)$$

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity – decision faithful to ground-truth

$$P(O=1|Y=1,PV=a) = P(O=1|Y=1,PV=b)$$

- requires access to the ground-truth
  - unavailable for unsupervised OD task

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity – decision faithful to ground-truth

$$P(O=1|Y=1,PV=a) = P(O=1|Y=1,PV=b)$$

- requires access to the ground-truth
  - unavailable for unsupervised OD task
- D3 (SP) and D4 are incompatible

[Barocas et al.'2017]

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity – decision faithful to ground-truth

$$P(O{=}1|Y{=}1,PV{=}a) = P(O{=}1|Y{=}1,PV{=}b)$$

- proxy enforces group-level rank preservation
- fidelity to within-group ranking from the base model i.e.
  - $\pi_{PV=v}^{BASE} = \pi_{PV=v}; \forall\, v \in \{a, b\}$
  - π denotes ranking

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity

D5. Base rate preservation – equal base rate in flagged instances and population

$$P(Y = 1|O = 1, PV = v) = P(Y = 1|PV = v) \ , \forall v \in \{a, b\}$$

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity

D5. Base rate preservation – equal base rate in flagged instances and population

$$P(Y = 1|O = 1, PV = v) = P(Y = 1|PV = v) \text{ , } \forall v \in \{a, b\}$$

- given OD satisfies D1 and D3, it cannot also satisfy D5

See Claim 1 in the paper

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity

D5. Base rate preservation – equal base rate in flagged instances and population

$$P(Y = 1|O = 1, PV = v) = P(Y = 1|PV = v) \;, \forall v \in \{a, b\}$$

- relaxed notion: equal base rate proportions
- still D5 cannot be satisfied

See Claim 2 in the paper

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

Fair OD model follows the proposed desiderata.

D4. Group fidelity

D5. Base rate preservation

detection performance

fairness related

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

✓ Enforceable

D4. Group fidelity

✓ Enforceable via proposed proxy

D5. Base rate preservation

✗ Can't be enforced

Roadmap

- Introduction
- Problem: Fairness in OD

- Desiderata
- Fairness aware OD
- Evaluation

Fairness-aware Outlier detection

- Given:
  - Observations $\mathcal{X} = \{X_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$
  - $\mathcal{PV} = \{PV_i\}_{i=1}^{N},\ PV_i \in \{a, b\}$
    - $PV_i = a$ identifies majority group

- Build a **detector** that estimates outlier scores $\mathcal{S}$ and assigns outlier labels $\mathcal{O}$ to achieve

i. $P(Y = 1 \mid O = 1) > P(Y = 1)$ *[D1]*

ii. $P(O=1|X) = P(O=1|X,PV=v), \forall\, v$ *[D2]*

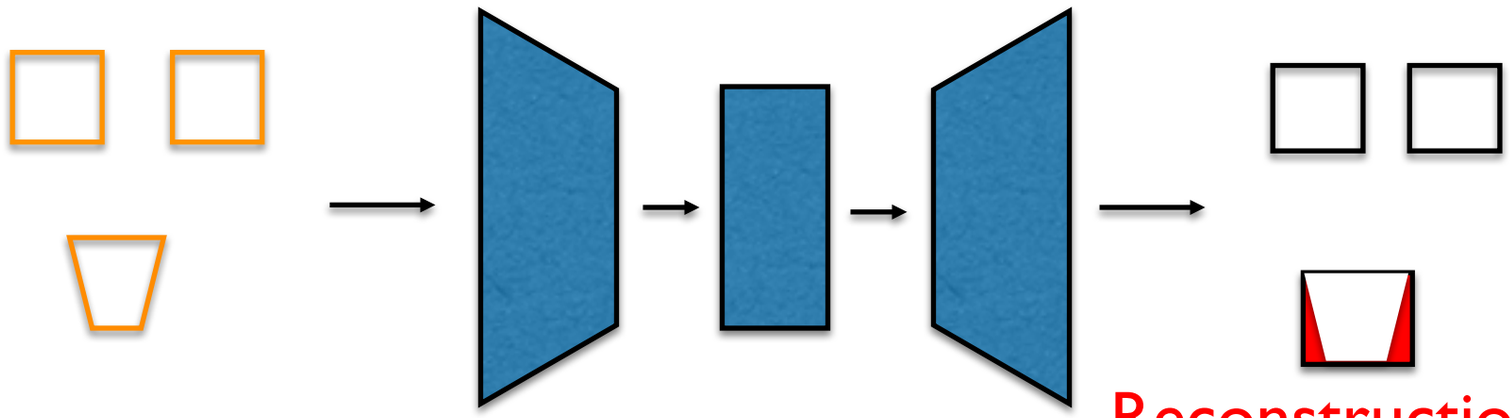
iii. $P(O=1|PV=a) = P(O=1|PV=b)$ *[D3]*

iv. $\pi^{\text{BASE}}_{\text{PV=v}} = \pi_{\text{PV=v}}; \forall\, v$, *[D4]*

BASE is fairness-agnostic detector

FAIROD

- Instantiates deep-autoencoder as BASE detector



Reconstruction error as outlier score

- Minimizes the regularized loss

$$\mathcal{L} = \alpha \underbrace{\mathcal{L}_{\text{BASE}}}_{\text{Reconstruction}} + (1-\alpha) \underbrace{\mathcal{L}_{SP}}_{\text{Statistical Parity}} + \gamma \underbrace{\mathcal{L}_{GF}}_{\text{Group Fidelity}}$$

Refer the paper for details on equations

Roadmap

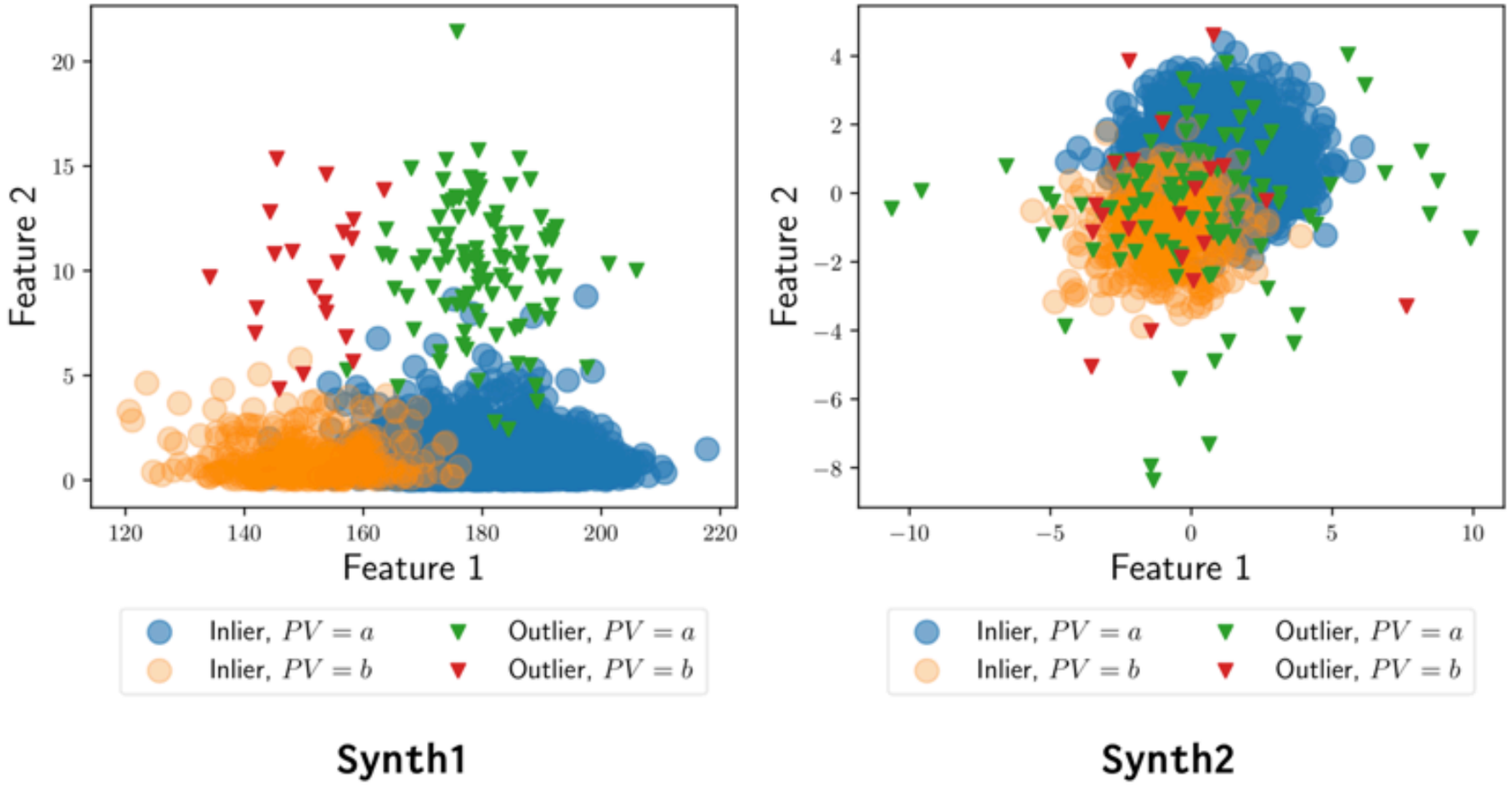
- Introduction
- Problem: Fairness in OD

- Desiderata
- Fairness aware OD
- Evaluation

Datasets

Dataset	N	d	PV	PV = b	$\lvert\mathcal{X}_{\text{PV=a}}\rvert/\lvert\mathcal{X}_{\text{PV=b}}\rvert$	% outliers	Labels
Adult	25262	11	gender	*female*	4	5	{income $\leq 50K$, income $> 50K$}
Credit	24593	1549	age	*age* ≤ 25	4	5	{paid, delinquent}
Tweets	3982	10000	racial dialect	*African-American*	4	5	{normal, abusive}
Ads	1682	1558	simulated	1	4	5	{non-ad, ad}
Synth1	2400	2	simulated	1	4	5	{0, 1}
Synth2	2400	2	simulated	1	4	5	{0, 1}

Synthetic datasets



Baselines

- BASE – fairness-agnostic deep anomaly detector

Preprocessing based methods

- RW – reweights instances [Kamiran et al.'2012]
- DIR – edits features to decorrelate PV [Feldman et al.'2015]
- LFR – latent representation obfuscating PV information [Zemel et al.'2013]
- ARL – latent representation via adversarial training [Beutel et al.'2017]

Evaluation Measures

- Fairness = $\min\left(r, \frac{1}{r}\right)$, where $r = \frac{P\ (O{=}\mathsf{l}\,|PV\ {=}a)}{P\ (O{=}\mathsf{l}\,|PV\ {=}\,\mathsf{b})}$

[D3]

Evaluation Measures

- Fairness = $\min\left(r, \frac{1}{r}\right)$, where $r = \frac{P\ (O{=}1 | PV = a)}{P\ (O{=}1 | PV = b)}$
- Group Fidelity = $HM(NDCG_{PV=a}, NDCG_{PV=b})$

[D4]

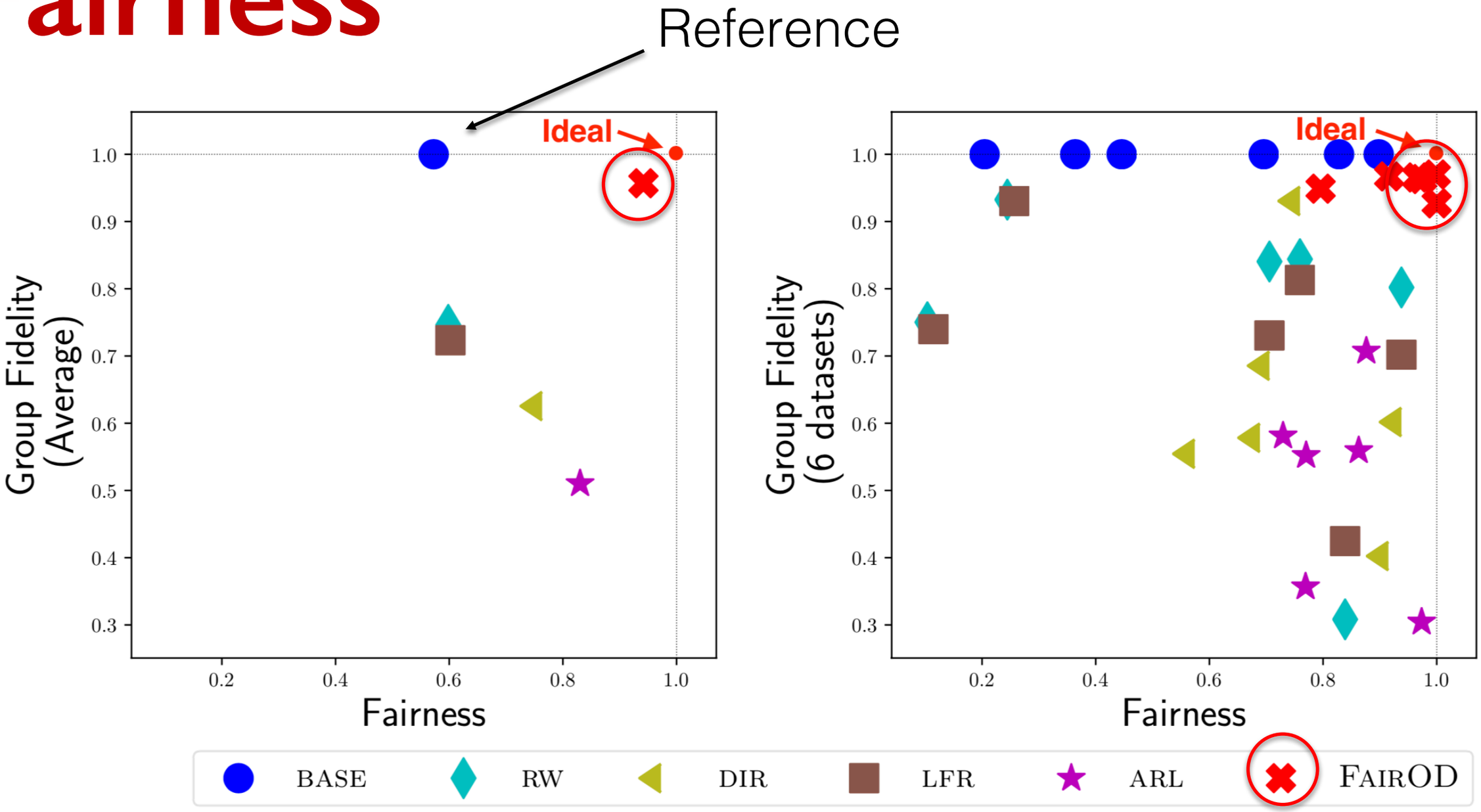
Evaluation Measures

- Fairness = $\min\left(r, \frac{1}{r}\right)$, where $r = \frac{P\ (O{=}1 | PV = a)}{P\ (O{=}1 | PV = b)}$
- Group Fidelity = $HM(NDCG_{PV=a}, NDCG_{PV=b})$

- AUC-ratio $= \frac{AUC_{PV=a}}{AUC_{PV=b}}$
- AP-ratio $= \frac{AP_{PV=a}}{AP_{PV=b}}$

} label aware parity measures used when ground truth labels are available

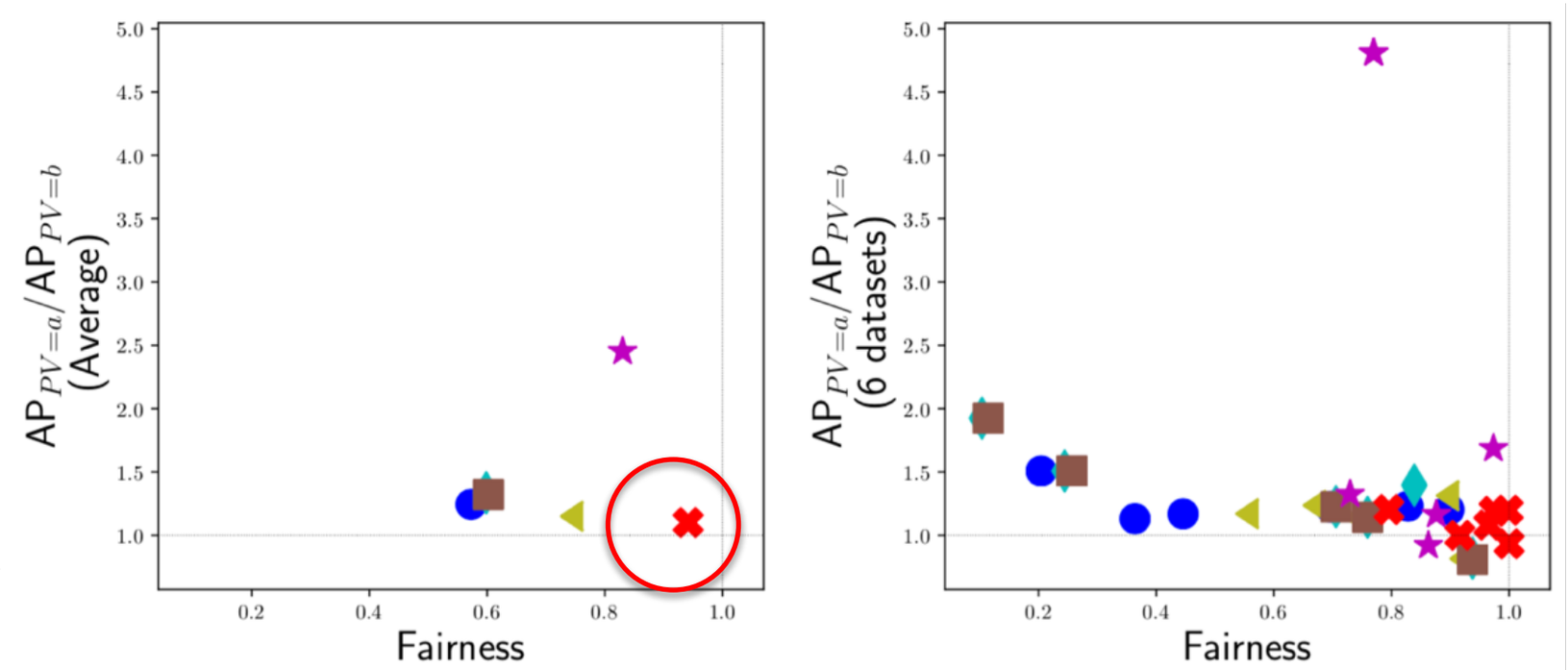
Fairness



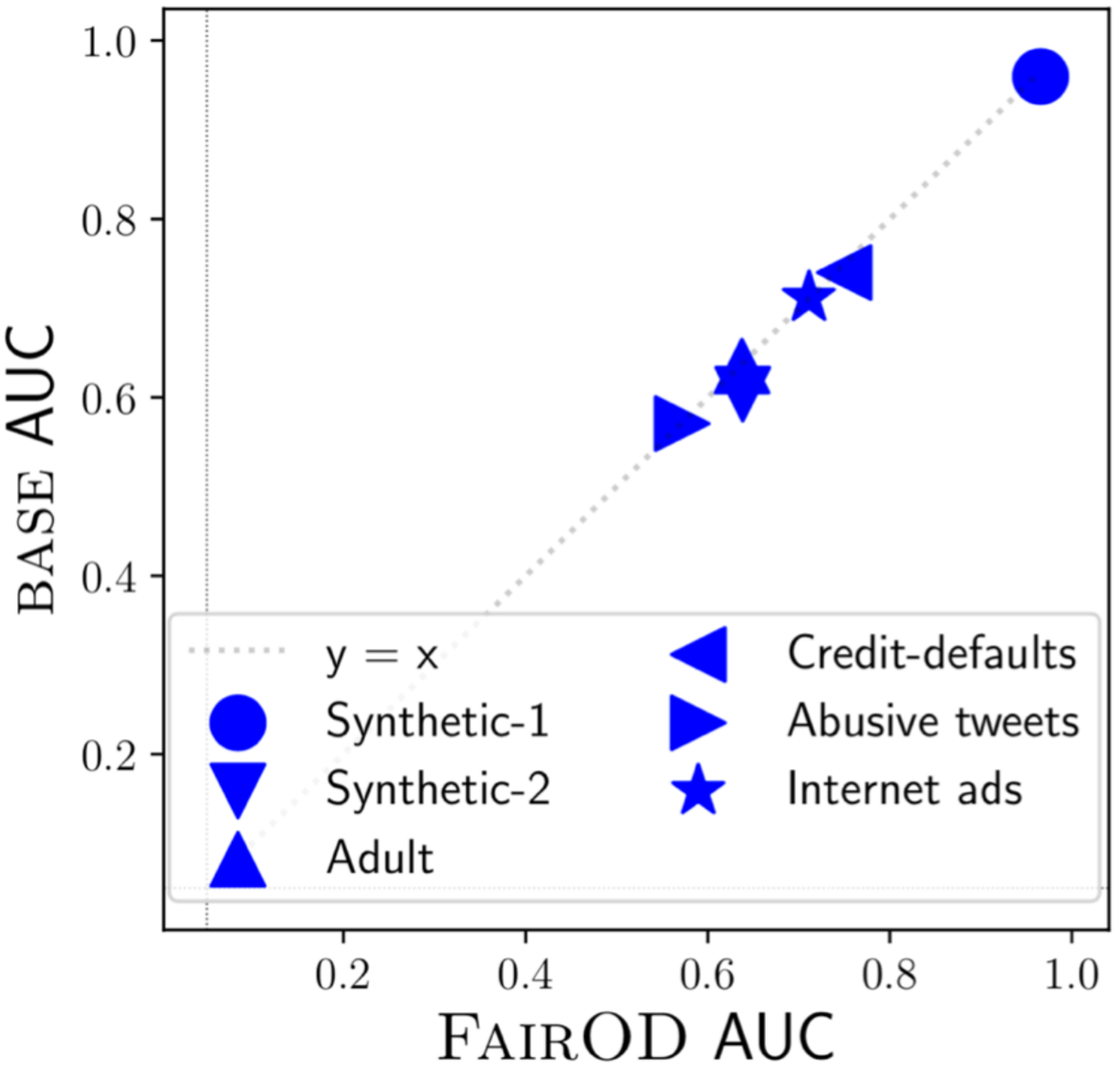
Group Fidelity vs Fairness

Fairness

Label aware parity measures vs Fairness

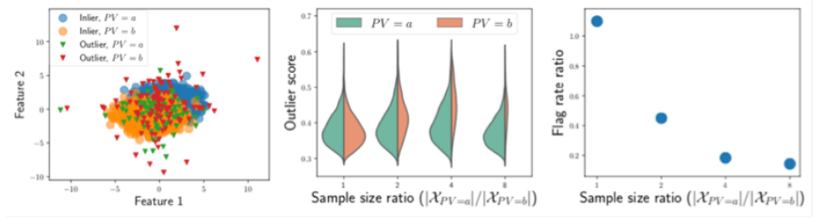


Fairness-accuracy trade-off



Conclusion

- Identified guiding desiderata for, and concrete formalization of the fair OD problem



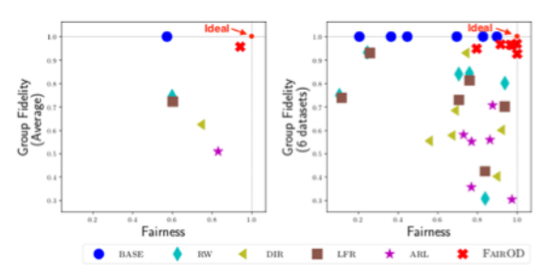
- Introduced well-motivated fairness criteria

$$\mathcal{L} = \alpha \underbrace{\mathcal{L}_{\text{BASE}}}_{\text{Reconstruction}} + (1-\alpha) \underbrace{\mathcal{L}_{SP}}_{\text{Statistical Parity}} + \gamma \underbrace{\mathcal{L}_{GF}}_{\text{Group Fidelity}}$$

- We proposed FAIROD

  End-to-end detector that incorporates prescribed criteria ✓

  Experiments demonstrate effectiveness in achieving fairness goals ✓ and accurate detection ✓



Thanks!

Carnegie Mellon University
Heinz College

Snap Inc.

Artificial Intelligence, Ethics, and Society

Carnegie Mellon University
DATA Lab

Dimitris Berberidis

Code, paper, and slides

https://shubhranshu-shekhar.github.io/#publications