FAIROD: Fairness-aware Outlier Detection

Shubhranshu Shekhar

Neil Shah

Leman Akoglu

https://tinyurl.com/fairOD

Longer version: https://arxiv.org/pdf/2012.03063.pdf

Fourth AAAI /ACM Conference on Artificial Intelligence, Ethics, and Society

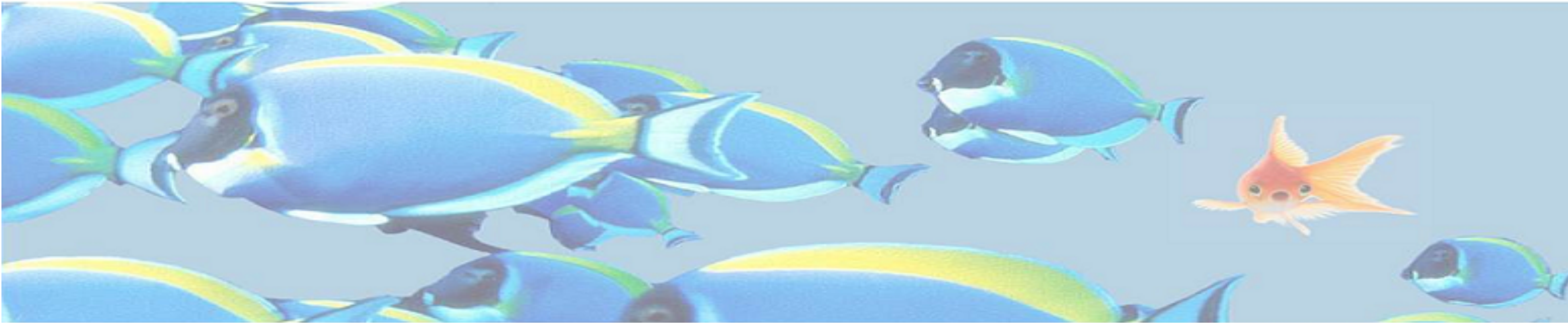
Carnegie Mellon University Heinz College

Snap Inc.

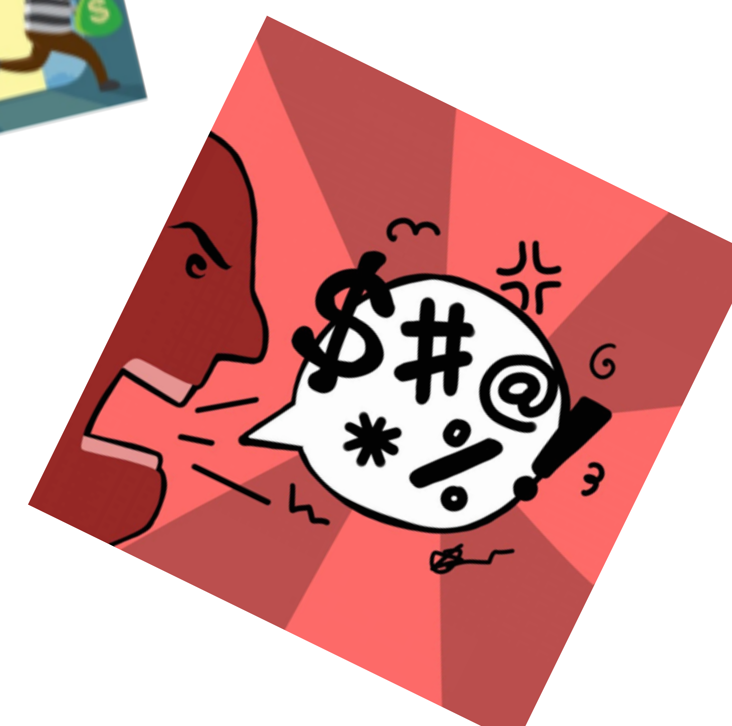
What is an outlier?

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]



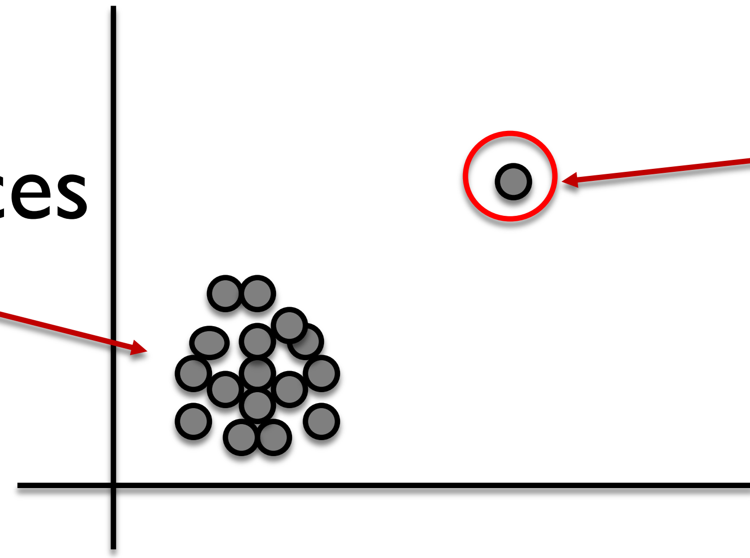
Outlier Detection: Use-cases



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,

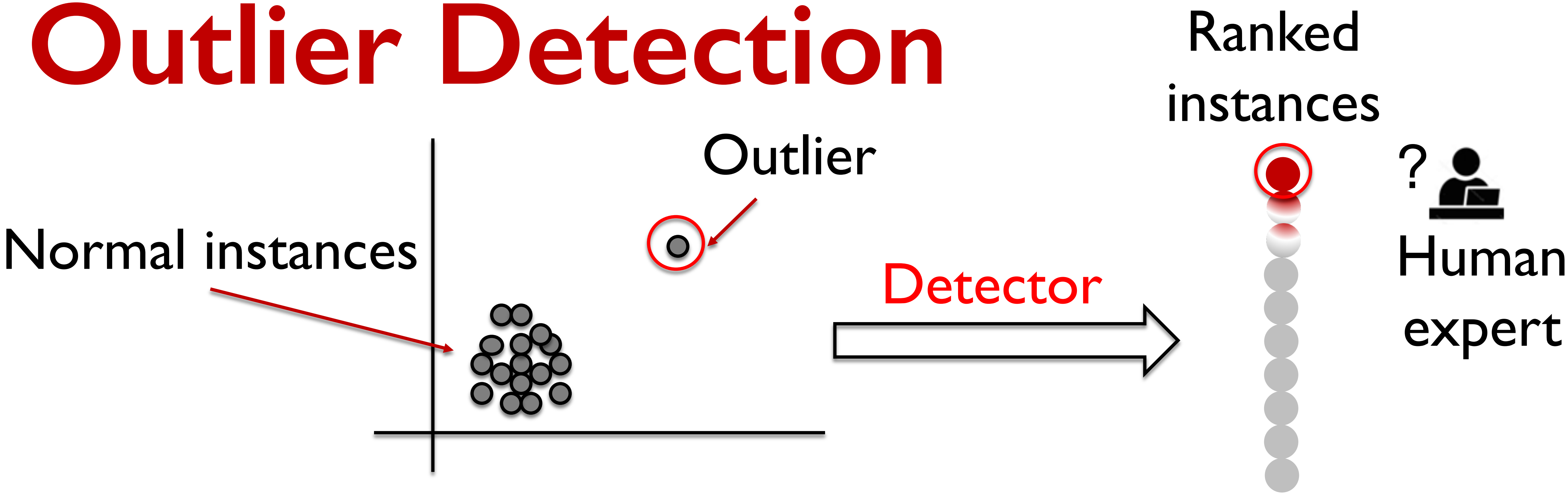
Outlier Detection

Normal instances



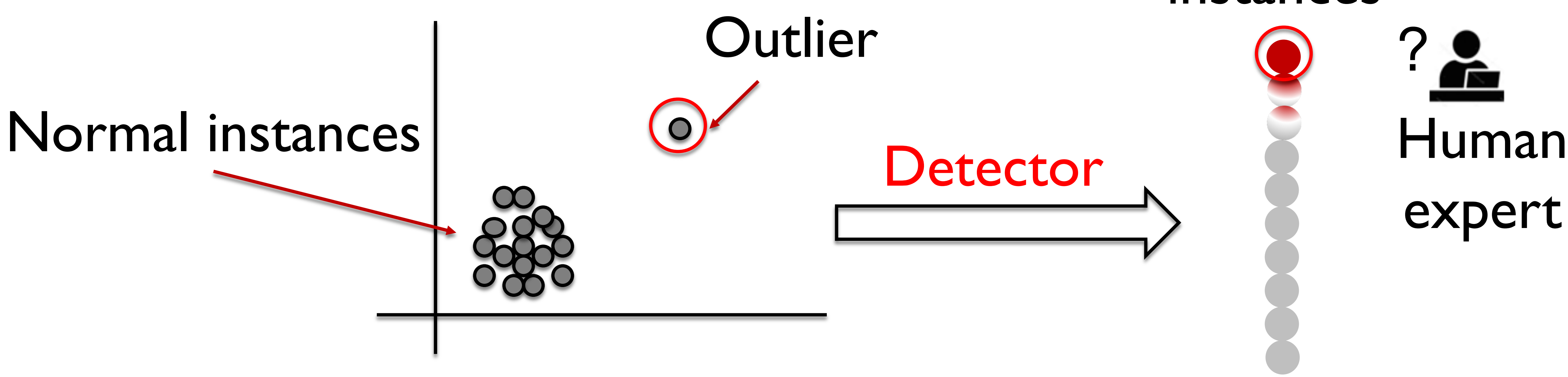
Inconsistent with normal observations

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

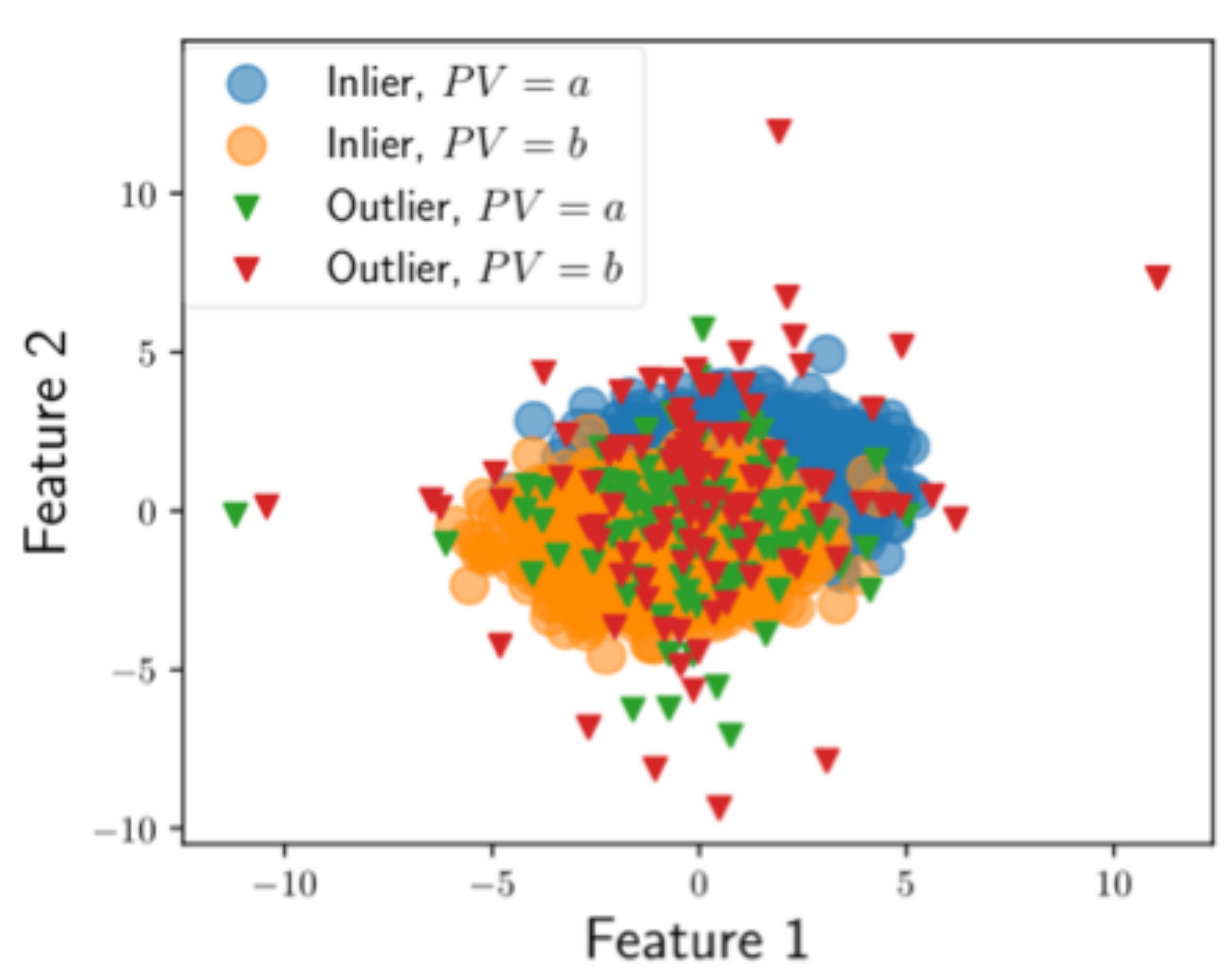
Assumes outlierness reflects true *riskiness*.

- designed to spot/flag rare, 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

Roadmap

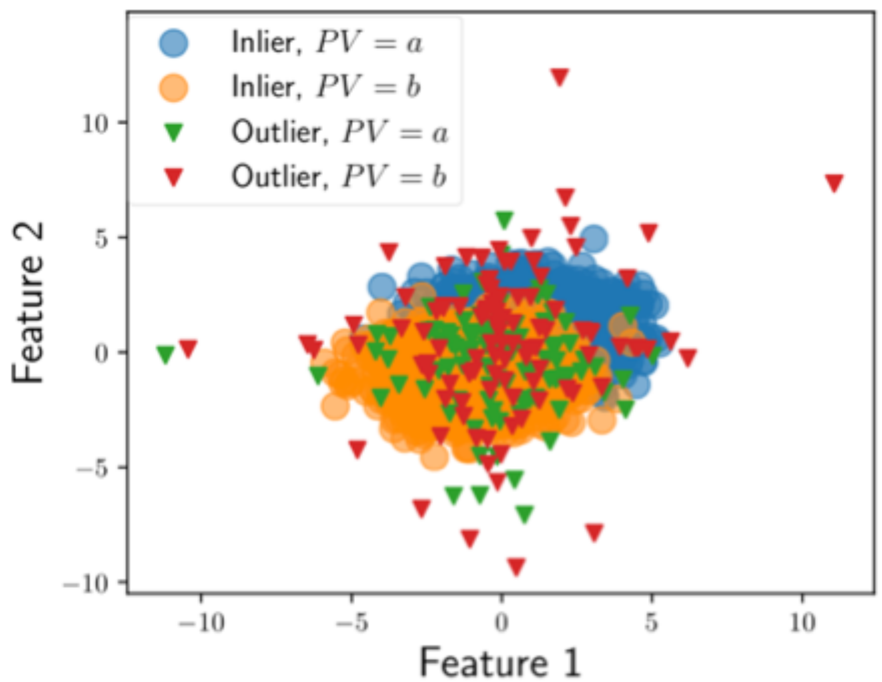
- Introduction
- Problem: Fairness in OD
- Desiderata
- Fairness-aware OD
- Evaluation

Bias in Outlier Detection

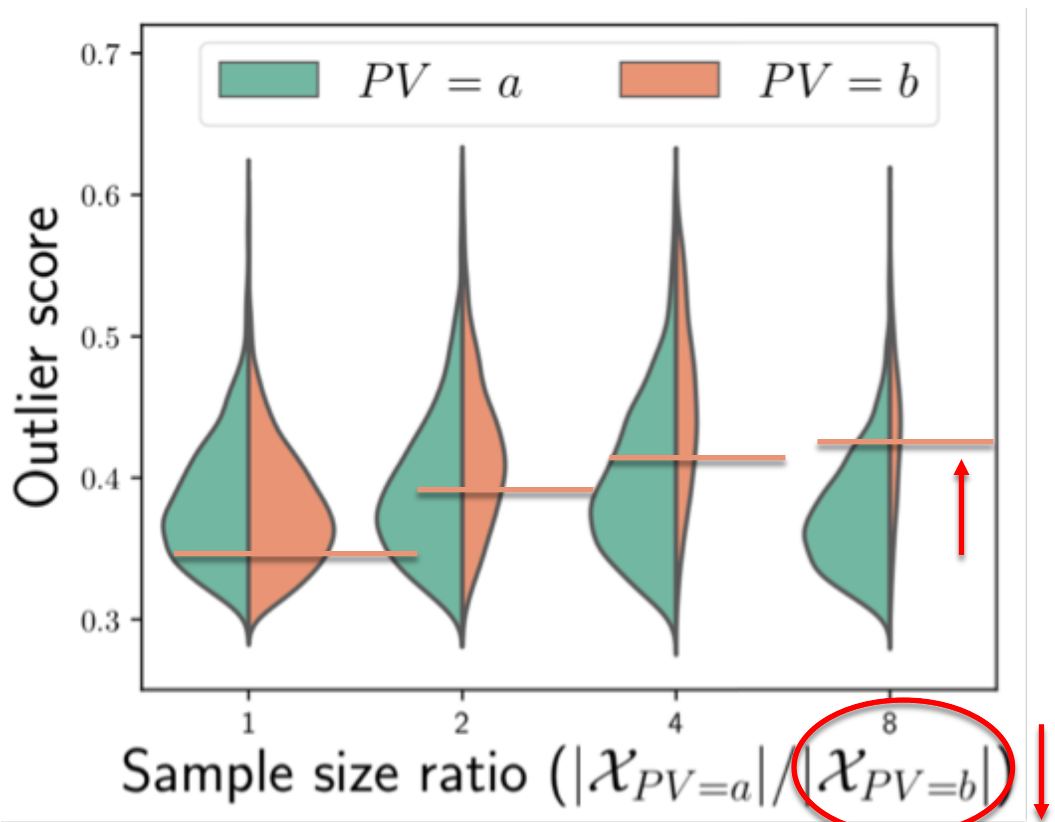


- Simulated dataset
  - equal sized groups
  - groups induced by
    $PV = a$ and $PV = b$

Bias in Outlier Detection

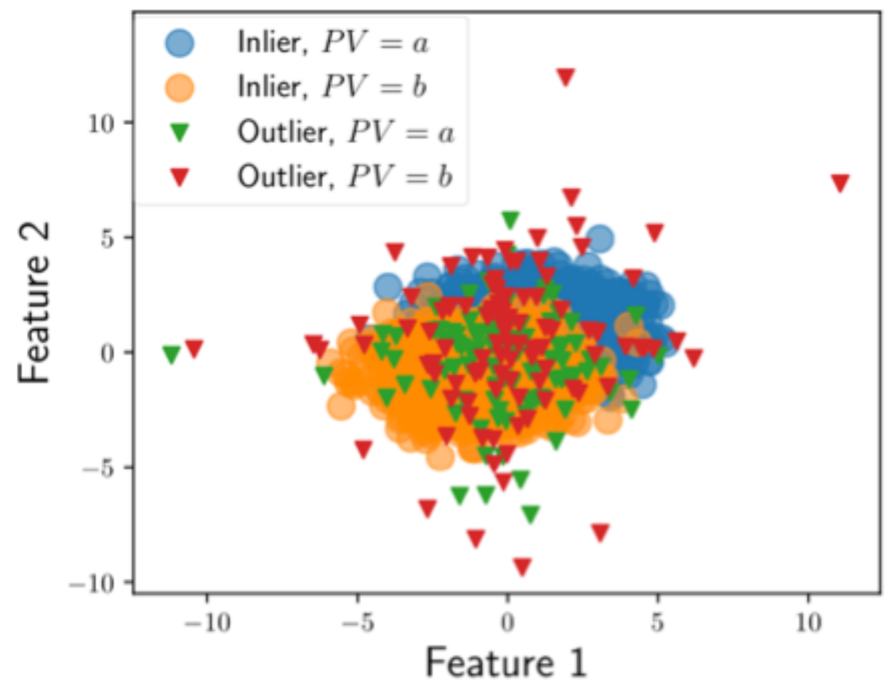


- Simulated dataset
  - equal sized groups
  - groups induced by PV = a and PV = b



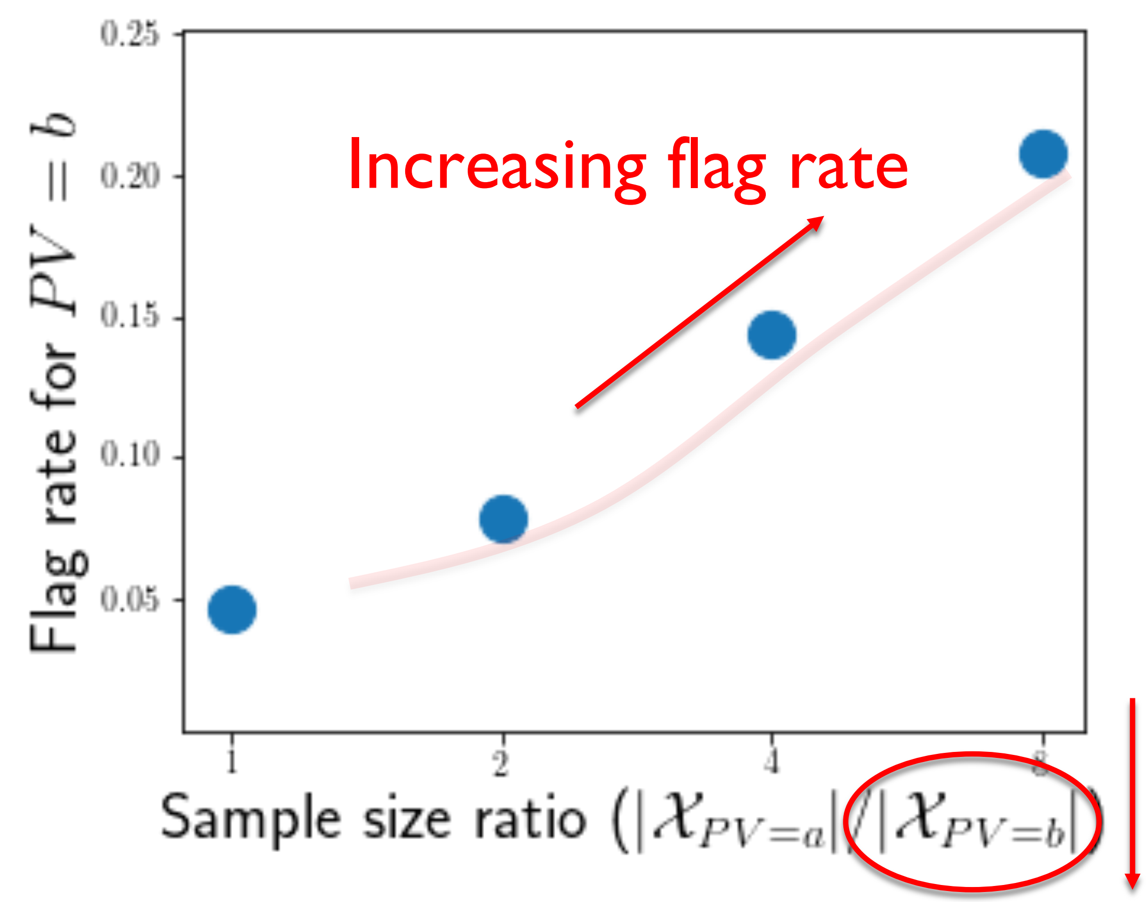
Higher outlier scores as sample size of $PV = b$ is decreased

Bias in Outlier Detection



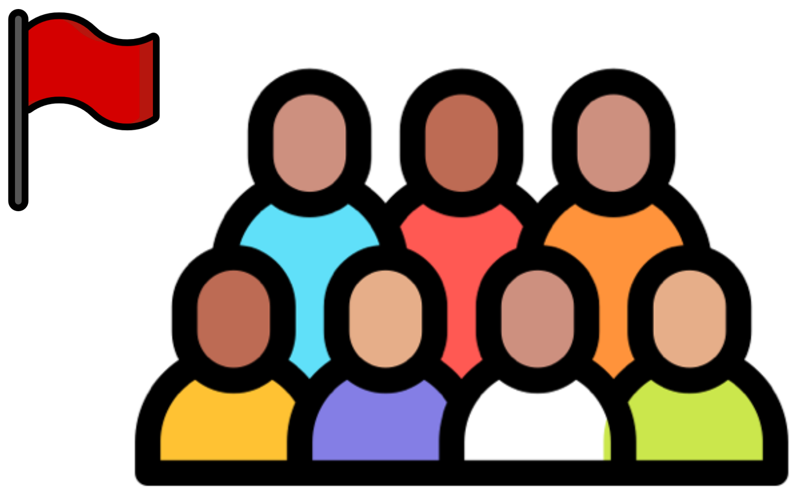
- Simulated dataset
  - equal sized groups
  - groups induced by
    PV = a and PV = b

Corresponding flag rate for $PV = b$ increases



Bias in Outlier Detection

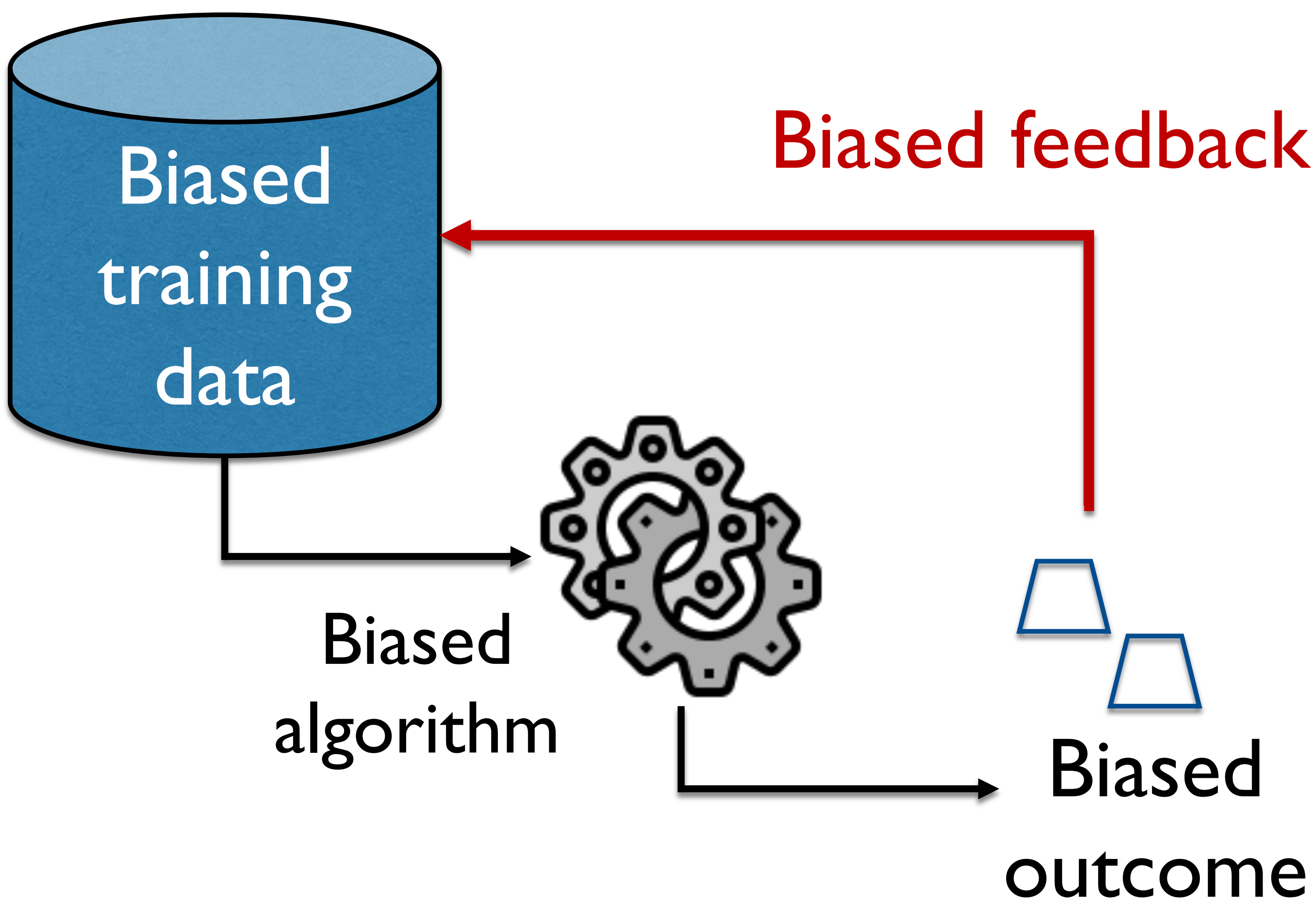
- Societal minorities may be statistical minorities
  - defined by protected variable (PV) :
    race/ ethnicity/gender/age etc.



$\neq$ 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

- <u>Given</u>:
  - Observations $\mathcal{X} = \{X_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$
    - $PV_i = a$ identifies majority group

What constitutes a “<u>fair</u>” outcome in OD?

- <u>Build</u> 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 $\mathcal{Y}$

Literature on Fairness in OD

- Algorithmic fairness – mostly for supervised ML
  - Unsupervised OD adds challenge
  - Numerous notions of fairness and associated incompatibility results
- Possible approach: pre-processing
  - ➢ re-purpose (unsupervised) fair representation learning
  1. PV-obfuscated/masked new embeddings
  2. Re-weighted/adjusted data distributions
  - Issue: an isolated/detached step to OD task at hand

Literature on Fairness in OD

- Algorithmic fairness – mostly for supervised ML
  - Unsupervised OD adds challenge
  - Numerous notions of fairness and associated incompatibility results
- Countably-few work on fairness for OD
  1. A Framework for Determining the Fairness of Outlier Detection. [Ravi & Davidson, ECAI 2020]
     - ❖ Quantify/measure (detect) the (un)fairness of OD model outcomes post hoc (i.e. proceeding detection)
  2. Fair Outlier Detection. [P & Abraham, WISE 2020]
  3. Towards Fair Deep Anomaly Detection. [Zhang & Davidson, FAccT 2021]
  4. Deep Clustering based Fair Outlier Detection. [Song+, KDD 2021]
  5. Fairness-aware Outlier Ensemble. [Liu+, 2021 - unpublished]

Roadmap

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

Proposed Desiderata

D1. Detection effectiveness

detection performance

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity

D5. Base rate preservation

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

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

- ensures OD-decisions are "blindfolded" to PV

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity – decision avoids use of PV

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

- ensures OD-decisions are "blindfolded" to PV
- (!) may allow discriminatory OD results for minority:
  - due to several other features that (partially-)redundantly encode the PV (e.g. zipcode & race).
  - OD will use the PV indirectly, through <u>proxy</u> features.

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

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

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

- a.k.a. demographic parity, or group fairness

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

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

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

$\Longrightarrow$ fraction of minority (majority) members in flagged set is the same as fraction of minority (majority) in overall population.

$$fr_a = fr_b \text{ (SP)} \iff P(PV = a|O = 1) = P(PV = a) \text{ and}$$
$$P(PV = b|O = 1) = P(PV = b) \ .$$

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

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

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

$$\Longrightarrow \quad P(PV = a|O = 1) = P(PV = a) \ \text{ and}$$

$$P(PV = b|O = 1) = P(PV = b)\ .$$

- Derives from "luck egalitarianism" : [Carl Knight, 2009]
  counteract the distributive effects of "brute luck"
  – by redistributing equality to those who suffer through no fault of their own choosing of race, gender, etc.

Proposed Desiderata

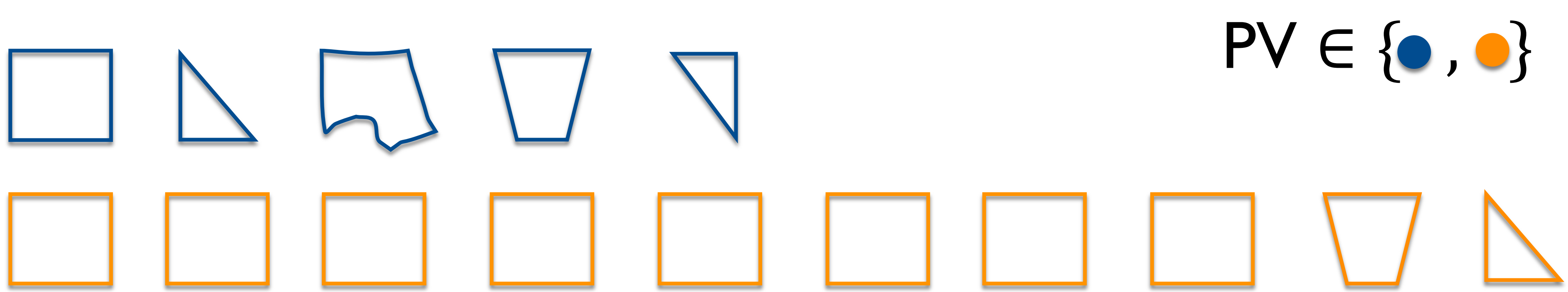
D1. Detection effectiveness

D2. Treatment parity

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

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

- ➢ permits *"laziness"*; may disadvantage some groups <u>despite</u> SP [Barocas et al.'2017]



Proposed Desiderata

D1. Detection effectiveness

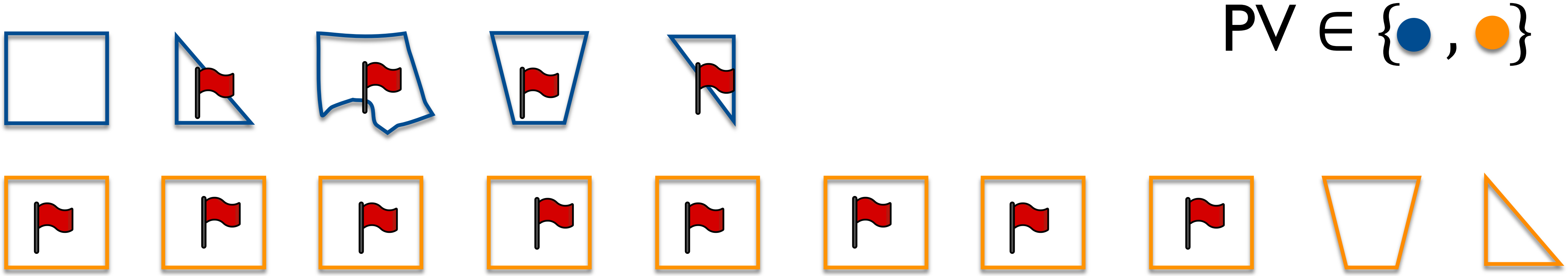
D2. Treatment parity

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

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

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

PV ∈ {●, ●}



Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity – decision faithful to ground-truth

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

- penalizes *"laziness"*
- equivalent to the so-called Equality of Opportunity*
- same true positive rate (TPR) for all groups

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

D4. Group fidelity – decision faithful to ground-truth

$$P(O{=}1 \mid Y{=}1, PV{=}a) \; = \; P(O{=}1 \mid Y{=}1, PV{=}b)$$

- ➢ 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

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

- approx.: enforce group-level rank preservation
- fidelity to within-group ranking from the $BASE$ model
  - $\pi_{PV=v}^{BASE} = \pi_{PV=v}; \quad \forall v \in \{a, b\}$
  - 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 the population

$$P(Y = 1 | O = 1, PV = v) = \underbrace{P(Y = 1 | PV = v)}_{\substack{\textbf{Base rate/Prevalence} \\ \text{for } PV = v}}, \ \forall v \in \{a, b\}$$

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 the population

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

➢ Incompatibility: 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 the population

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

- ➢ relaxation: preservation of the ratio of base rates
  - Leads to overestimation of true group-level base rates *(Claim 2)*
- ➢ still, D5 cannot be enforced: relies on ground-truth

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

Proposed Desiderata

D1. Detection effectiveness

D2. Treatment parity

D3. Statistical parity (SP)

✓ Enforceable

Fair OD model follows the proposed desiderata D1 - D4.

D4. Group fidelity

✓ Enforceable via proposed proxy

D5. Base rate preservation

✗ Can't be enforced

Literature on Fairness in OD

- Countably-few work on fair OD
  1. Fair Outlier Detection. [P and Abraham, WISE 2020]
     - ➢ Seminal paper
     - ➢ disparate treatment (i.e. uses PV) at decision time (may be unlawful for some settings!)
     - ➢ prioritizes statistical parity (SP); may permit "laziness"
     - ➢ not end-to-end but rather heuristic
  2. Towards Fair Deep Anomaly Detection. [Zhang & Davidson, FAccT 2021]
     - ➢ focus on SP
     - ➢ one-class objective & adversarial training for PV prediction

Literature on Fairness in OD

- Countably-few work on fairness for OD
  3. Deep Clustering based Fair Outlier Detection. [Song+, KDD 2021]
     - Again, sole focus on SP

  4. Fairness-aware Outlier Ensemble. [Liu+, 2021; not publ.]
     - assumes the outlier scores "obtained from the base outlier ensemble method is an optimal result" (why do anything if this is true!)
     - notions of *group* fairness : focus on SP only & *individual* fairness : similarity "based on original feature values excluding sensitive features" (proxy variables!)

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]*

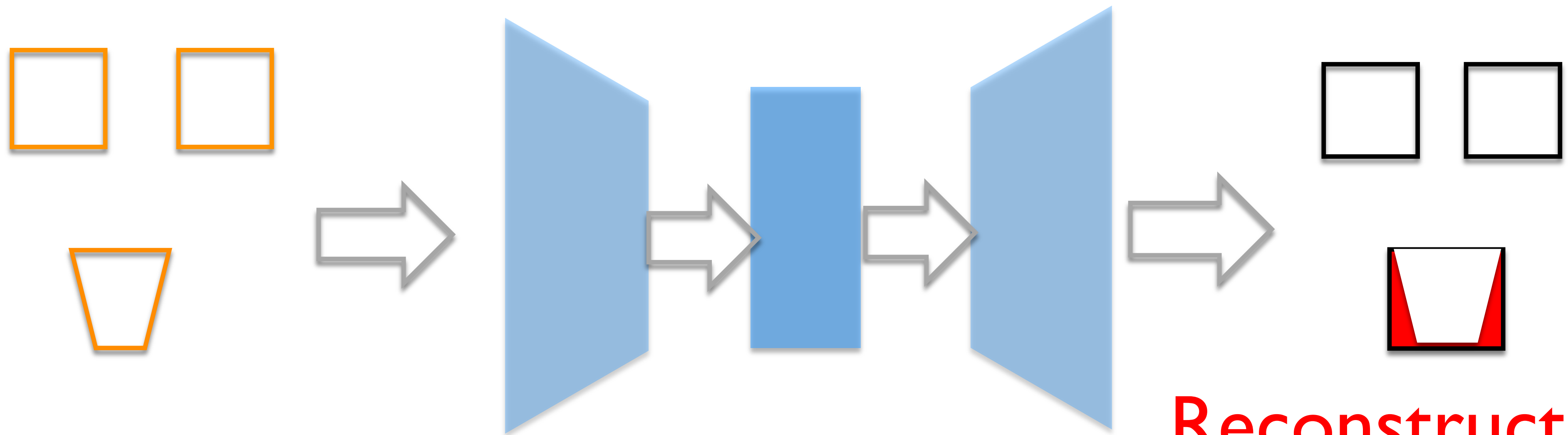
iii. $P(O{=}1|PV{=}a) = P(O{=}1|PV{=}b)$ *[D3]*

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

BASE is fairness-agnostic detector

FairOD

- Instantiates deep-autoencoder as BASE detector



- Minimizes the regularized loss:

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

FairOD

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

$$\mathcal{L}_{\text{BASE}} = \sum_{i=1}^{N} \|X_i - G(X_i)\|_2^2$$

$$\mathcal{L}_{SP} = \left| \frac{\left(\sum_{i=1}^{N} s(X_i) - \mu_s\right) \left(\sum_{i=1}^{N} PV_i - \mu_{PV}\right)}{\sigma_s \; \sigma_{PV}} \right|$$

$$\mathcal{L}_{GF} = \sum_{v \in \{a,b\}} \left(1 - \sum_{X_i \in \mathcal{X}_{PV=v}} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\log_2 \left(1 + \sum_{X_k \in \mathcal{X}_{PV=v}} \text{sigm}(s(X_k) - s(X_i))\right) \cdot IDCG_{PV=v}} \right)$$

See paper for details : https://arxiv.org/pdf/2012.03063.pdf

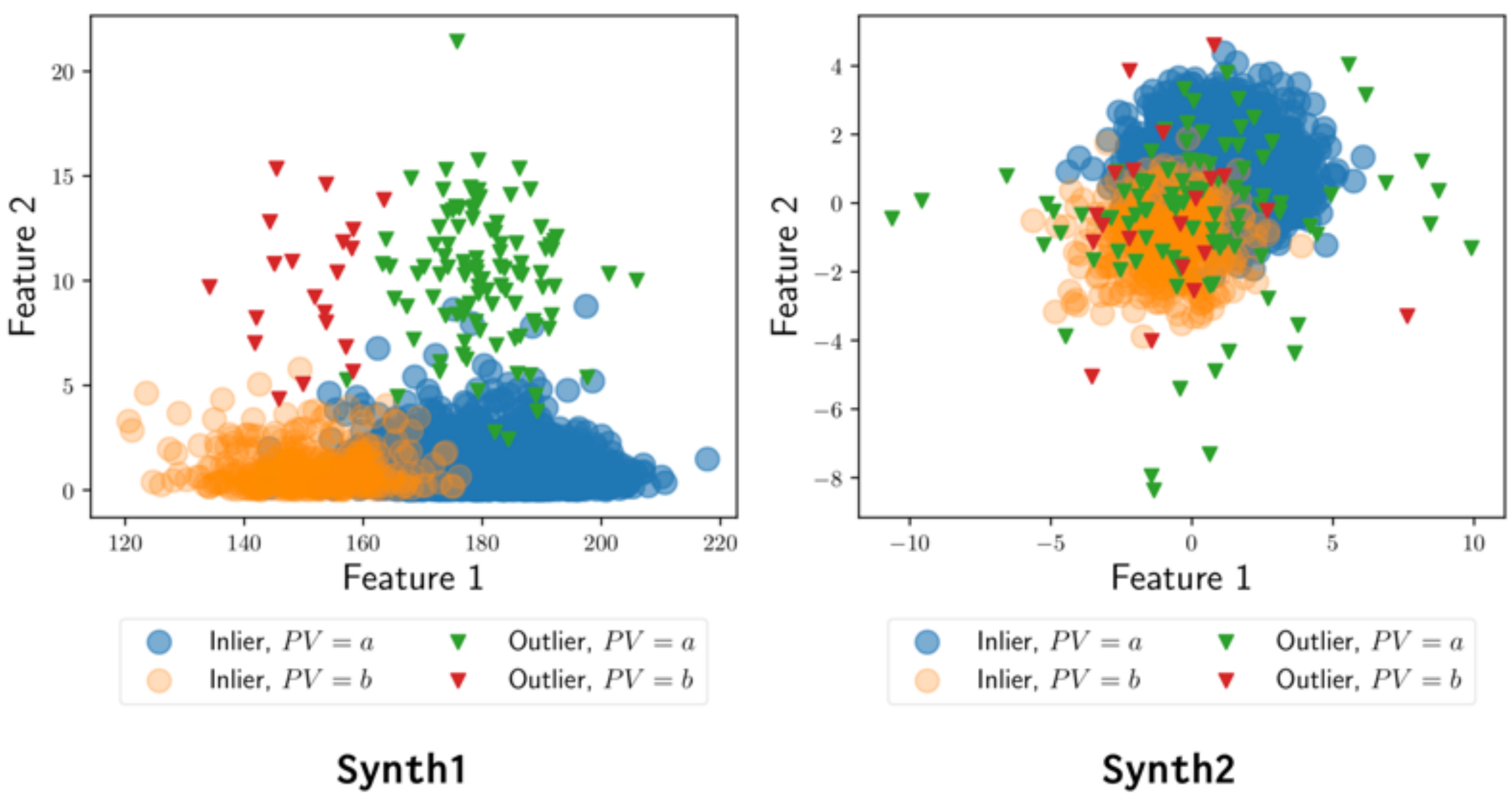
Roadmap

- 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 de-correlate 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 $r = \frac{P\ (O=1|PV=a)}{P\ (O=1|PV=\mathrm{b})}$

[D3]

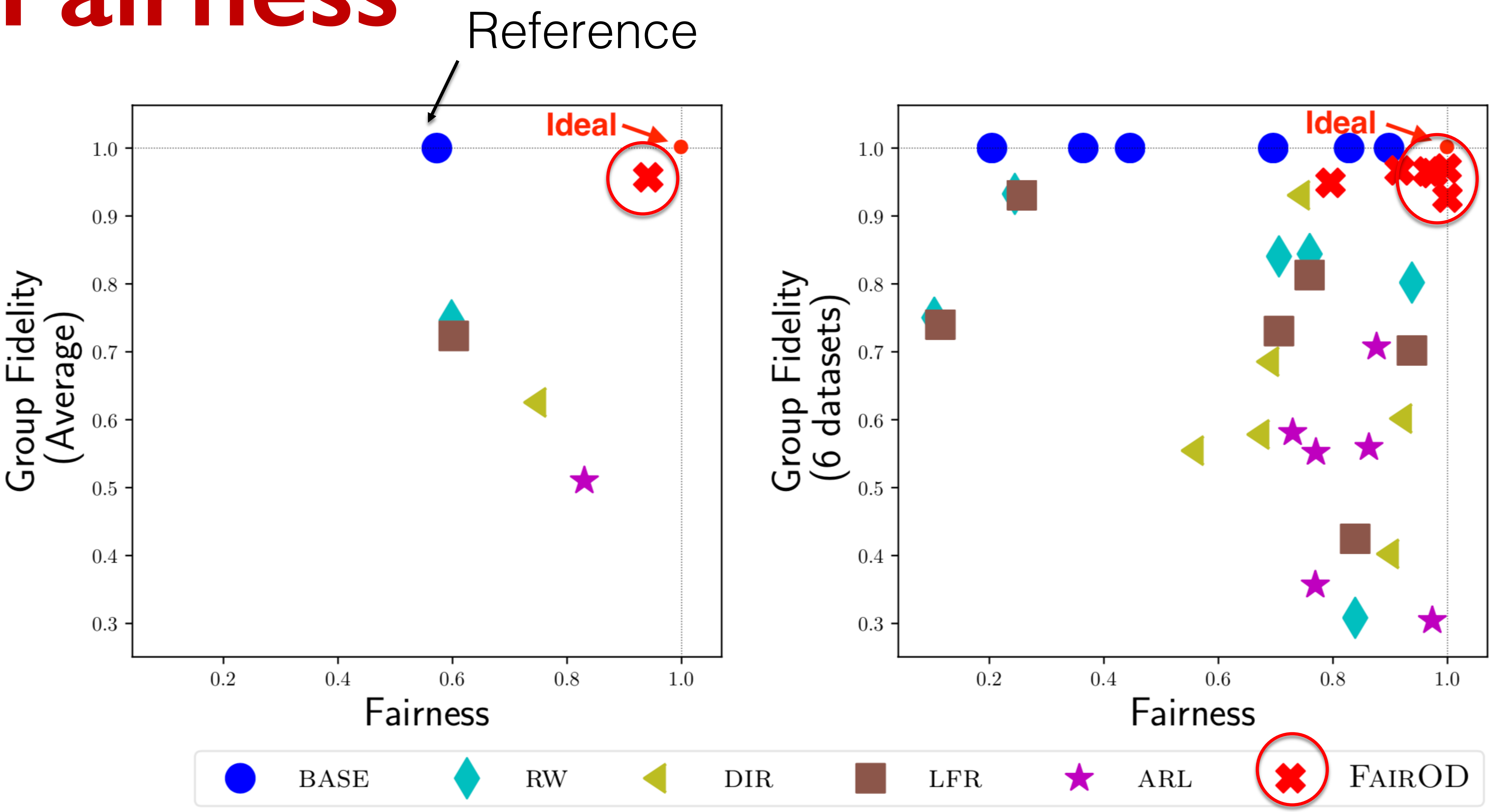
- Group Fidelity = $HM(NDCG_{PV=a}, NDCG_{PV=b})$

[D4]

- 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

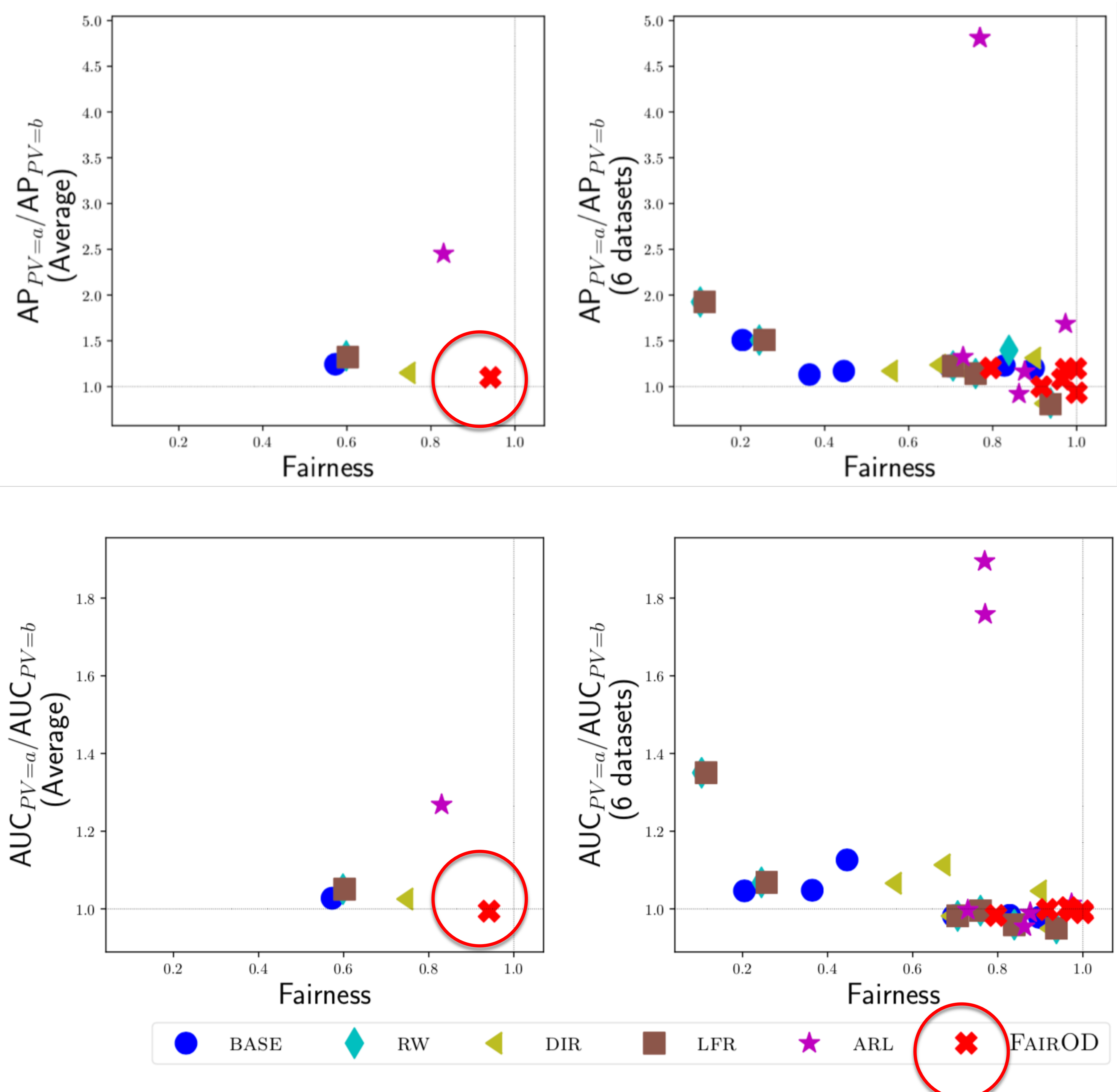
Fairness



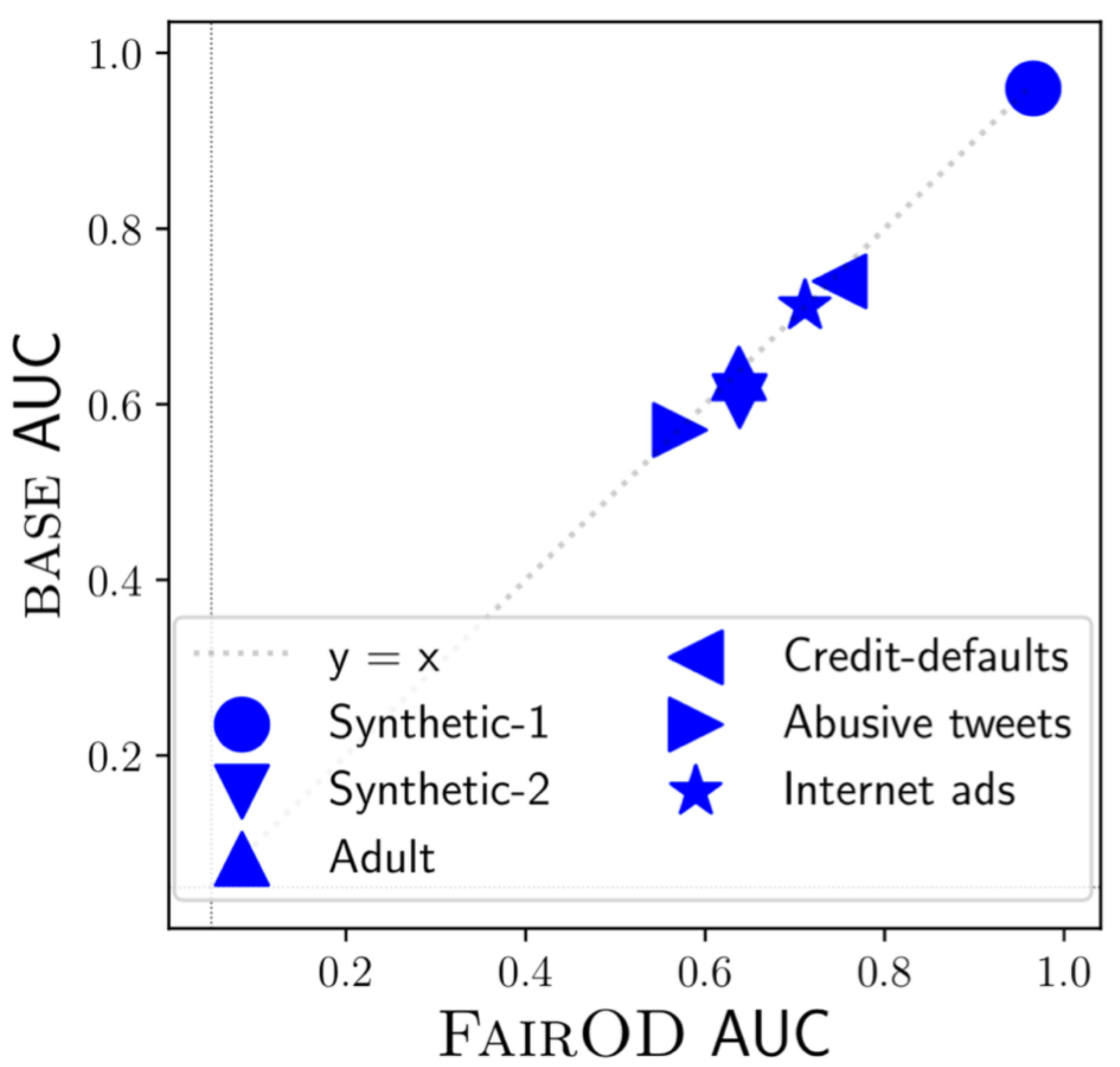
Group Fidelity vs Fairness

Fairness

Label-aware parity measures vs Fairness

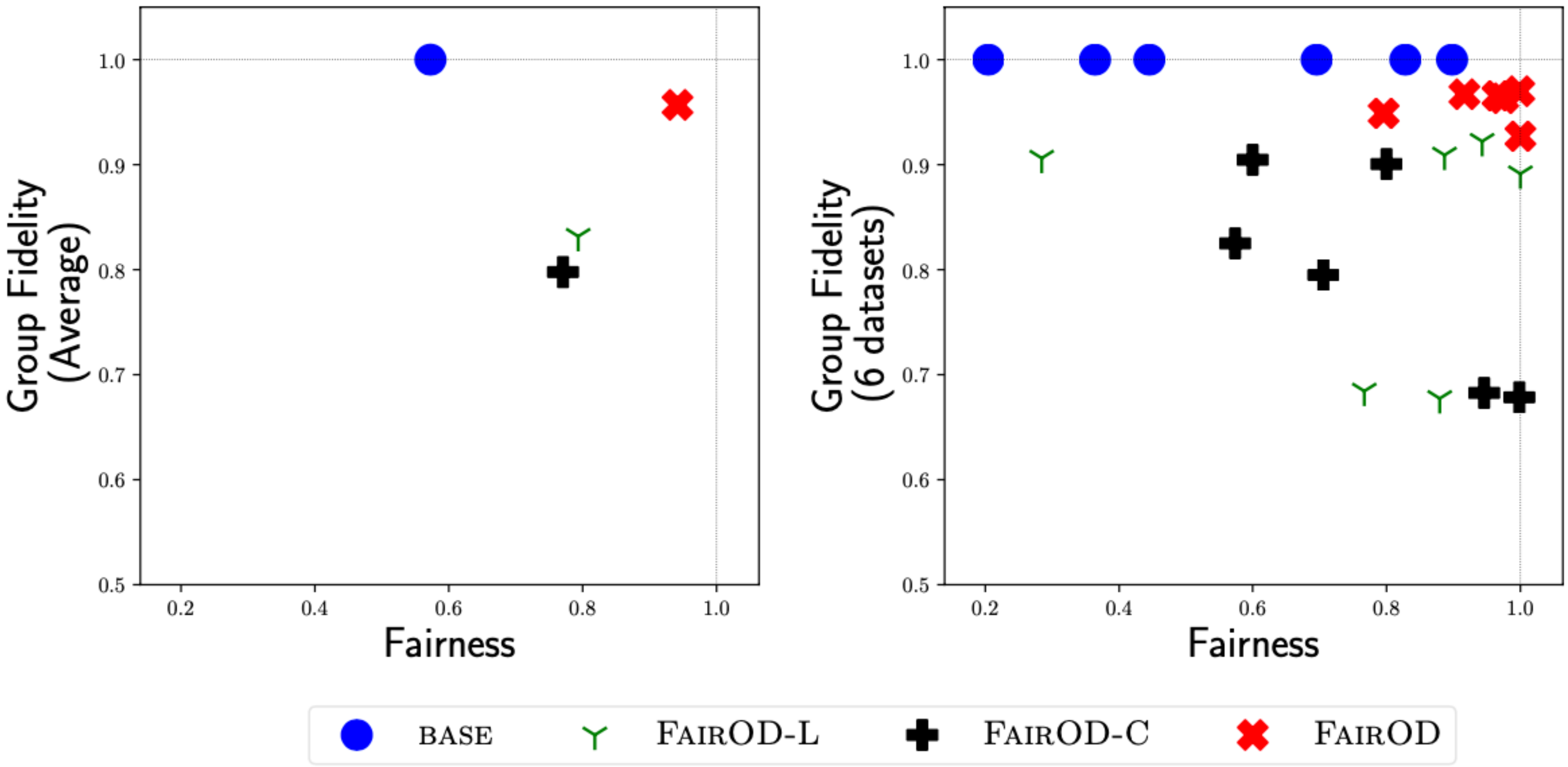


Fairness-accuracy trade-off



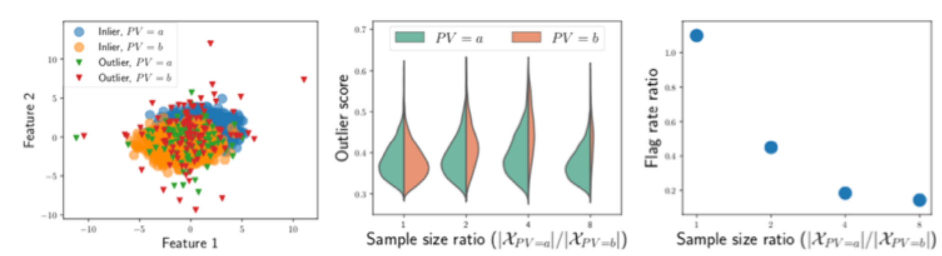
Ablation study

- FairOD-**L** : only SP-based regularization (permits "**L**aziness")
- FairOD-**C** : **C**orrelation-based group fidelity regularization



Conclusion

- ✓ Guiding desiderata for, and concrete formalization of the fair OD problem

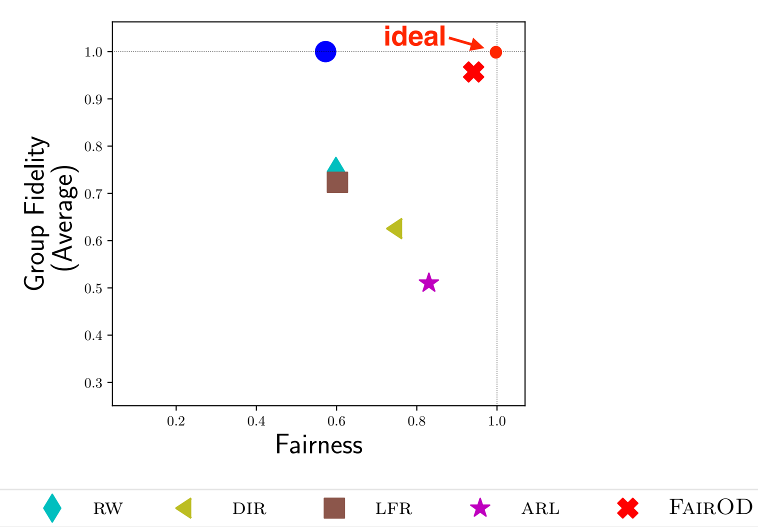


- ✓ Introduced well-motivated fairness criteria

- ✓ Proposed FairOD

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

  - End-to-end detector w/ prescribed criteria
  - Accurate detection that achieves fairness goals



Code, paper, and slides

https://tinyurl.com/fairOD

Thanks!

NSF

Carnegie Mellon University Heinz College

Snap Inc.

Artificial Intelligence, Ethics, and Society

Carnegie Mellon University DATA Lab

Dimitris Berberidis