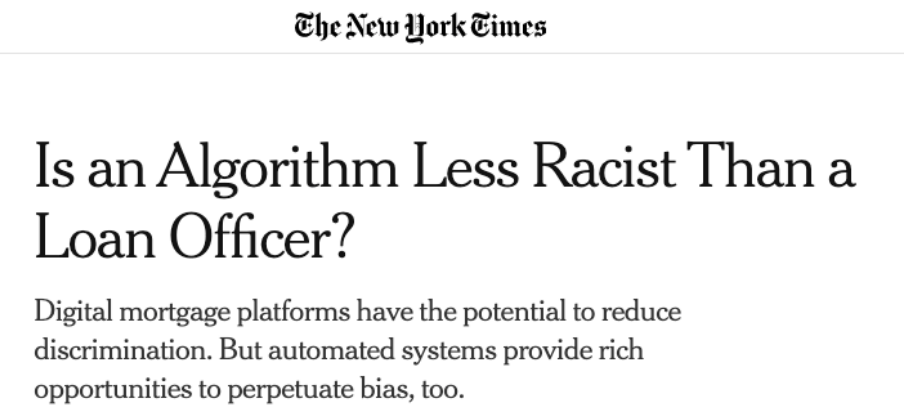
The 3 Perspectives of Incentive-Aware Machine Learning: Robustness, Fairness, Improvement & Causality

Chara Podimata (podimata@mit.edu)
MIT

BEACH day, April 2026

ML algorithms for **decision-making** are almost everywhere nowadays.

ML algorithms for **decision-making** are almost everywhere nowadays.



ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

HireVue Platform Why HireVue Hiring Resources

Your end-to-end hiring platform with video interview software, conversational AI, and assessments.

Build a faster, fairer, friendlier hiring process with HireVue's end-to-end hiring platform. Together, we can improve the way you discover, engage, and hire talent.

ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

HireVue Platform Why HireVue Hiring Resources

Your end-to-end hiring platform with video interview software, conversational AI, and assessments.

Build a faster, fairer, friendlier hiring process with HireVue's end-to-end hiring platform. Together, we can improve the way you discover, engage, and hire talent.

The Washington Post
Democracy Dies in Darkness
Get one yea

Business

Student tracking, secret scores: How college admissions offices rank prospects before they apply

Before many schools even look at an application, they comb through prospective students' personal data, such as web-browsing habits and financial history

ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

→ increase # credit cards
→ increase # bank accounts
→ improve credit history

The Washington Post
Democracy Dies in Darkness

Business

Student tracking, secret scores: How college admissions offices rank prospects before they apply

Before many schools even look at an application, they comb through prospective students' personal data, such as web-browsing habits and financial history

HireVue Platform Why HireVue Hiring Resources

Your end-to-end hiring platform with video interview software, conversational AI, and assessments.

Build a faster, fairer, friendlier hiring process with HireVue's end-to-end hiring platform. Together, we can improve the way you discover, engage, and hire talent.

ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

- increase # credit cards
- increase # bank accounts
- improve credit history

HireVue Platform Why HireVue Hiring Resources

Your end-to-end hiring platform with video interview software, conversational AI, and assessments.

Build a faster, fairer, friendlier hiring process with HireVue's end-to-end hiring platform. Together, we can improve the way you discover, engage, and hire talent.

The Washington Post
Democracy Dies in Darkness
Get one yea

Business

Student tracking, secret scores: How college admissions offices rank prospects before they apply

Before many schools even look at an application, they comb through prospective students' personal data, such as web-browsing habits and financial history

- improve GPA
- retake GRE / pay for classes
- change schools

ML algorithms for **decision-making** are almost everywhere nowadays.

The New York Times

Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.

→ increase # credit cards
→ increase # bank accounts
→ improve credit history

The Washington Post
Democracy Dies in Darkness

Business

Student tracking, secret scores: How college admissions offices rank prospects before they apply

Before many schools even look at an application, they comb through prospective students' personal data, such as web-browsing habits and financial history

→ improve GPA
→ retake GRE / pay for classes
→ change schools

HireVue Platform Why HireVue Hiring Resources

Your end-to-end hiring platform with video interview software, conversational AI, and assessments.

Build a faster, fairer, friendlier hiring process with HireVue's end-to-end hiring platform. Together, we can improve the way you discover, engage, and hire talent.

→ dress a certain way
→ hide piercings / tattoos
→ change way you talk

Problem

If ML algorithms **ignore** this **strategic behavior**,

they risk making **policy decisions** that are

incompatible with the original policy's goal.

What Can Go Wrong?

What Can Go Wrong?

Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.

What Can Go Wrong?

Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.

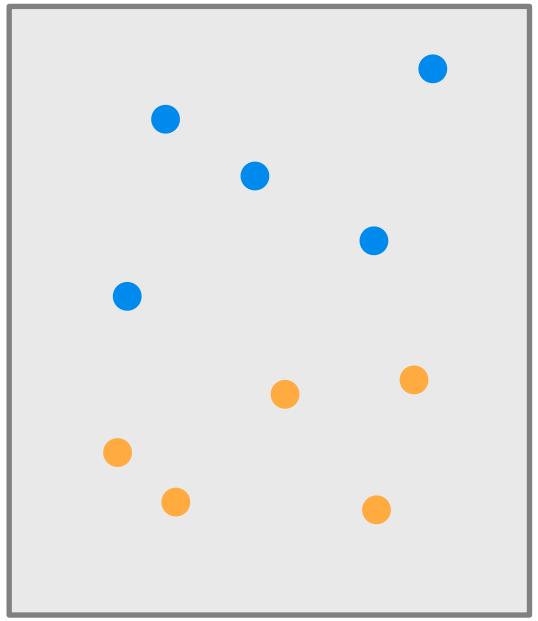
Training Data

Test Data

What Can Go Wrong?

Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.

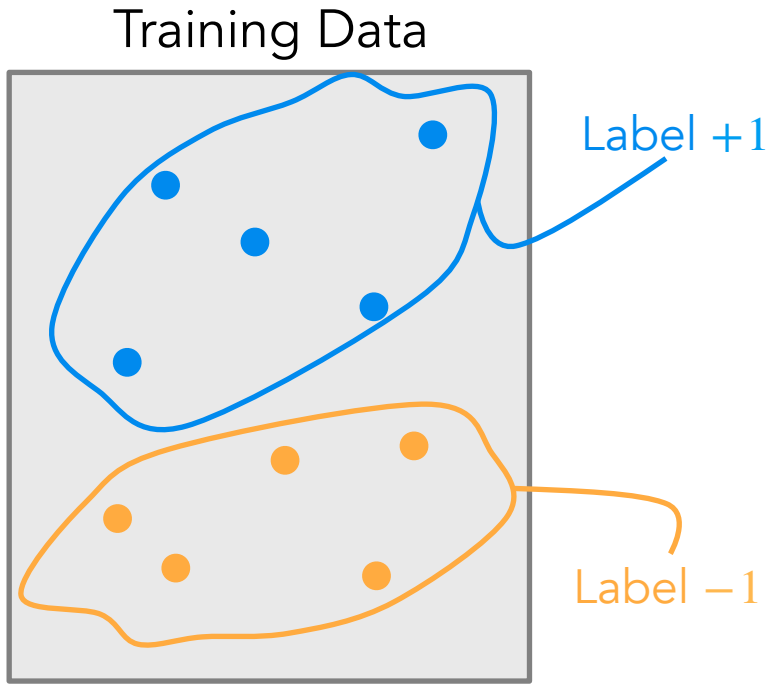
Training Data

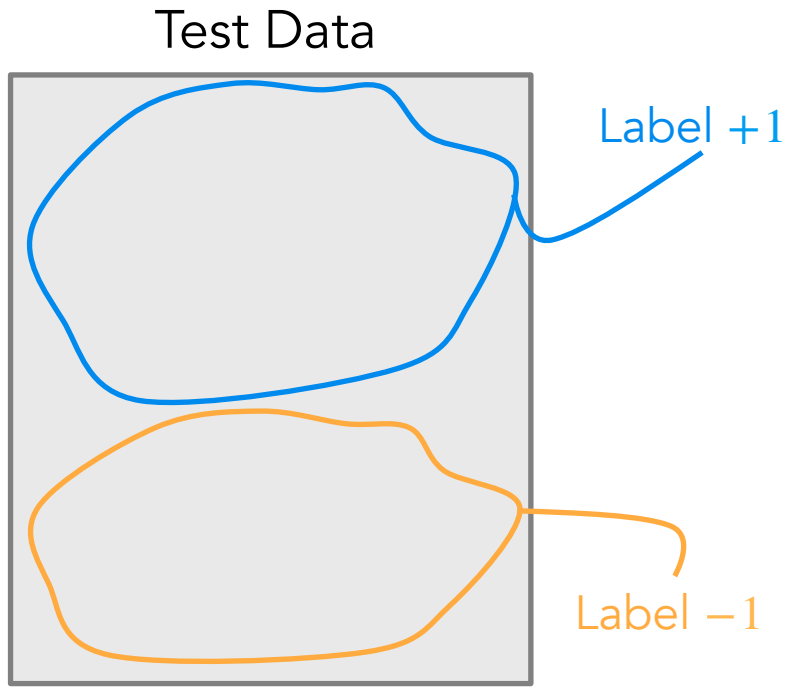


Test Data

What Can Go Wrong?

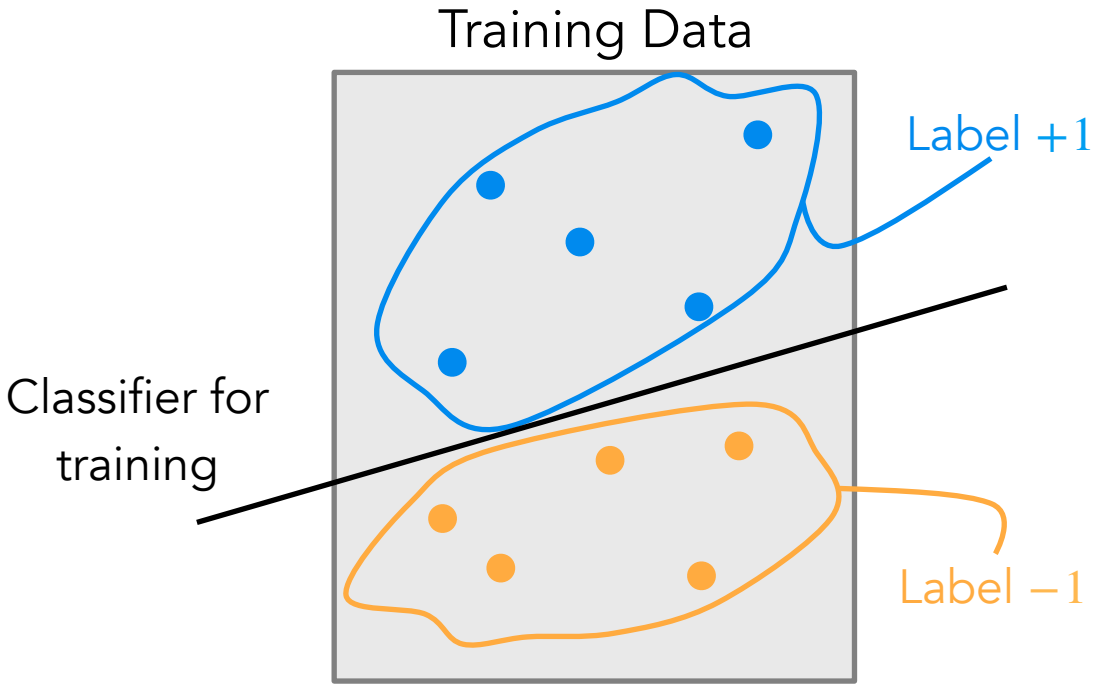
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.

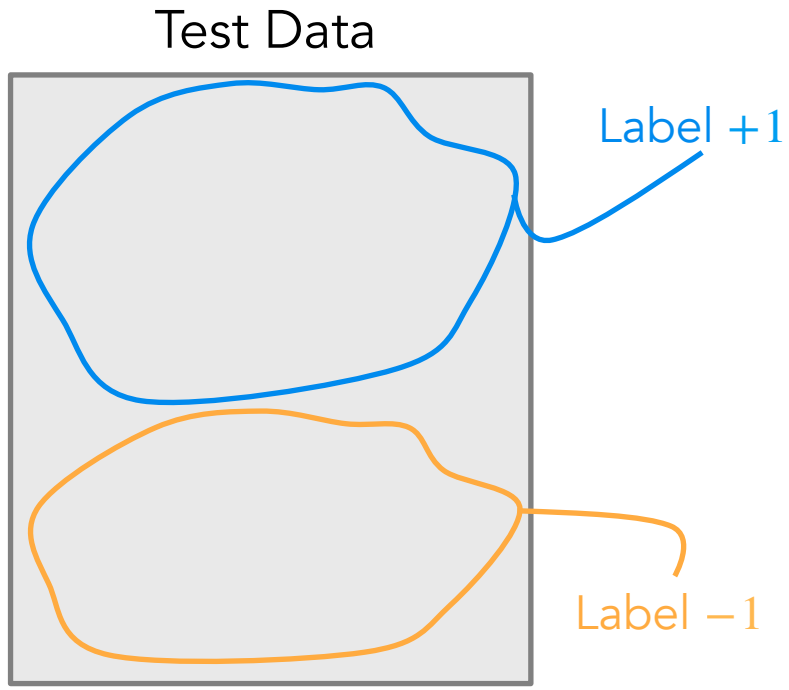




What Can Go Wrong?

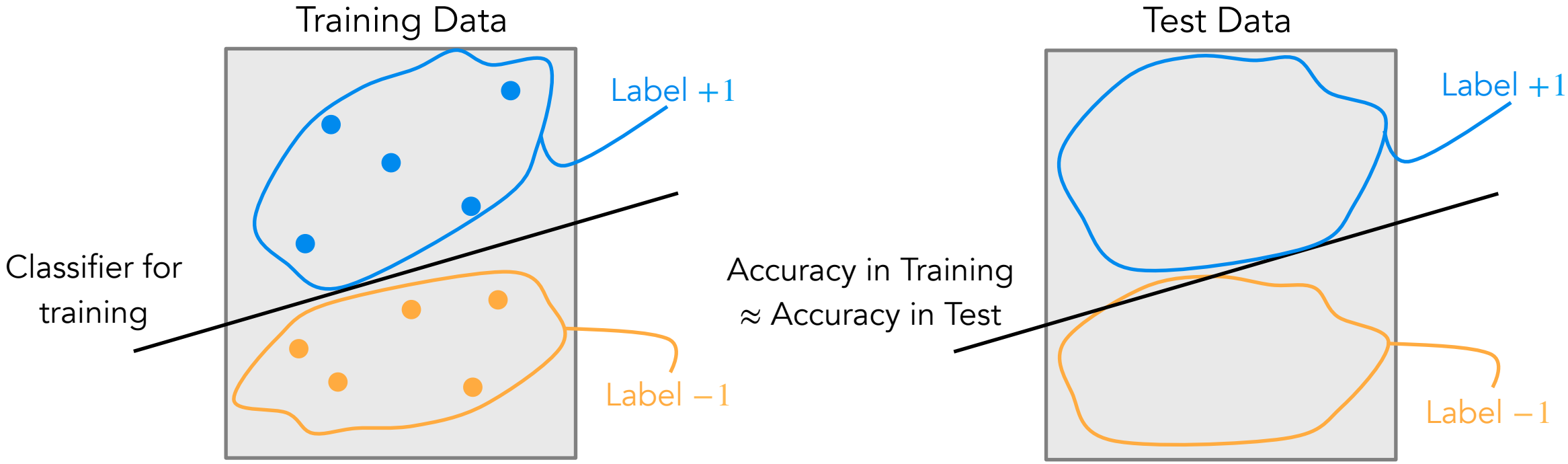
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.





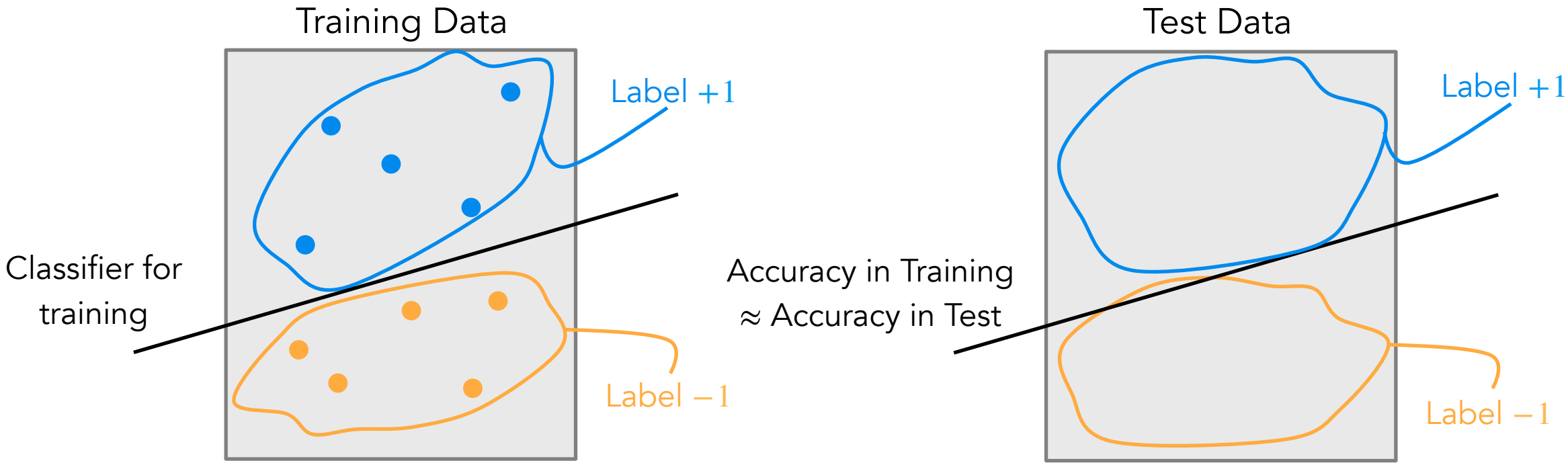
What Can Go Wrong?

Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.



What Can Go Wrong?

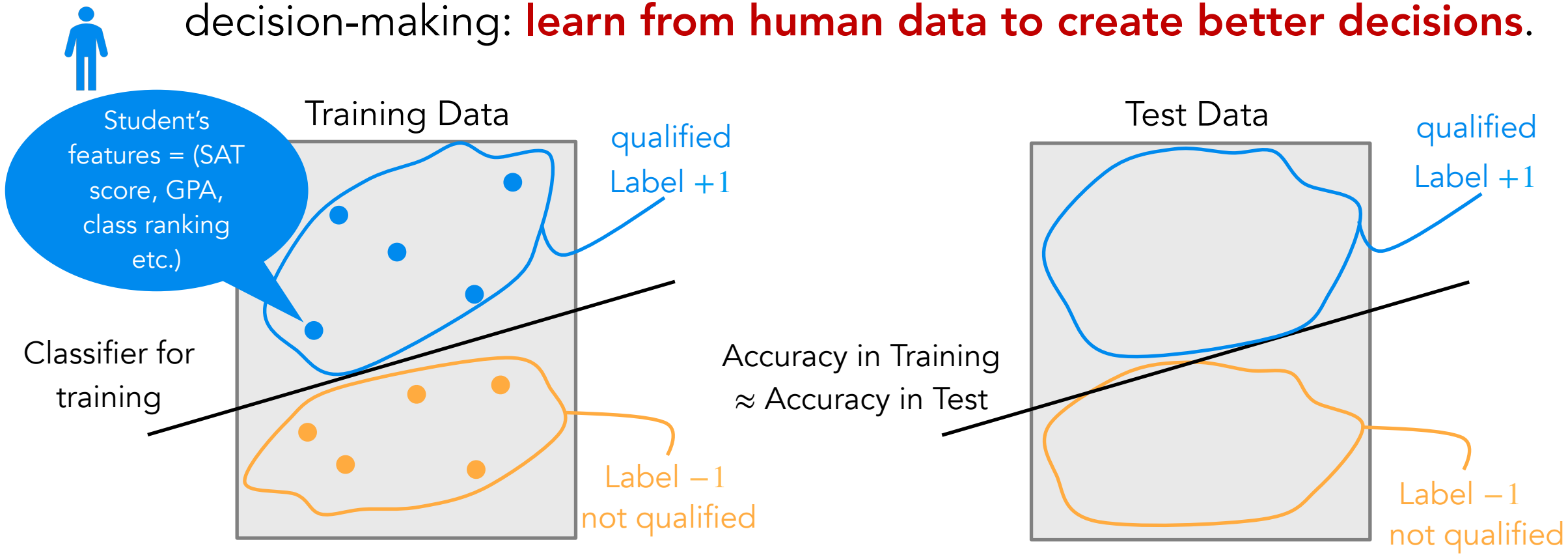
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.



Data **corresponds to individuals who have agency** and want to affect the decisions made on them by the ML algorithms.

What Can Go Wrong?

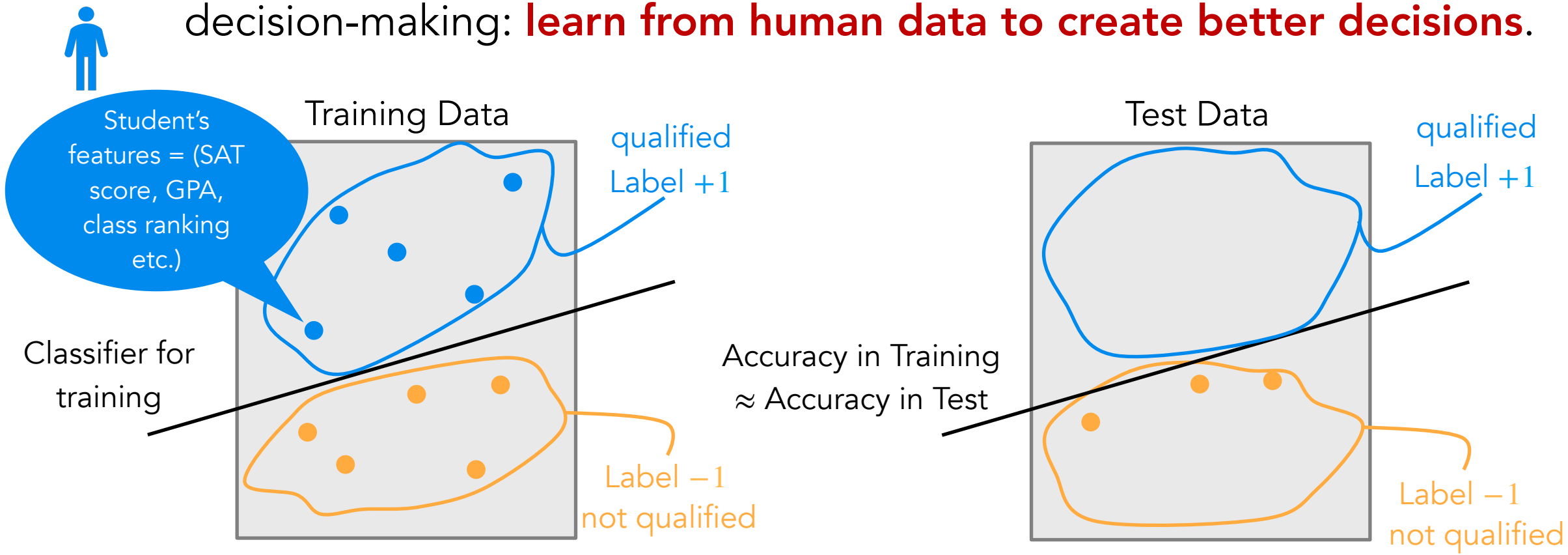
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.



Data **corresponds to individuals who have agency** and want to affect the decisions made on them by the ML algorithms.

What Can Go Wrong?

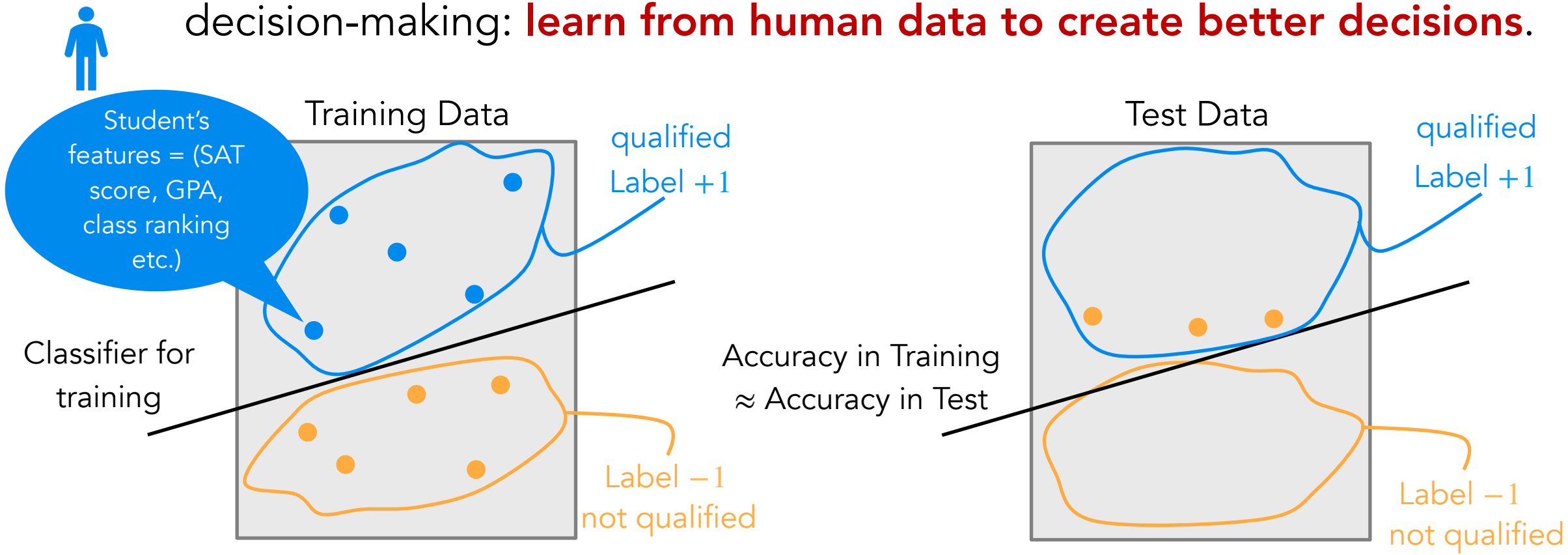
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.



Data **corresponds to individuals who have agency** and want to affect the decisions made on them by the ML algorithms.

What Can Go Wrong?

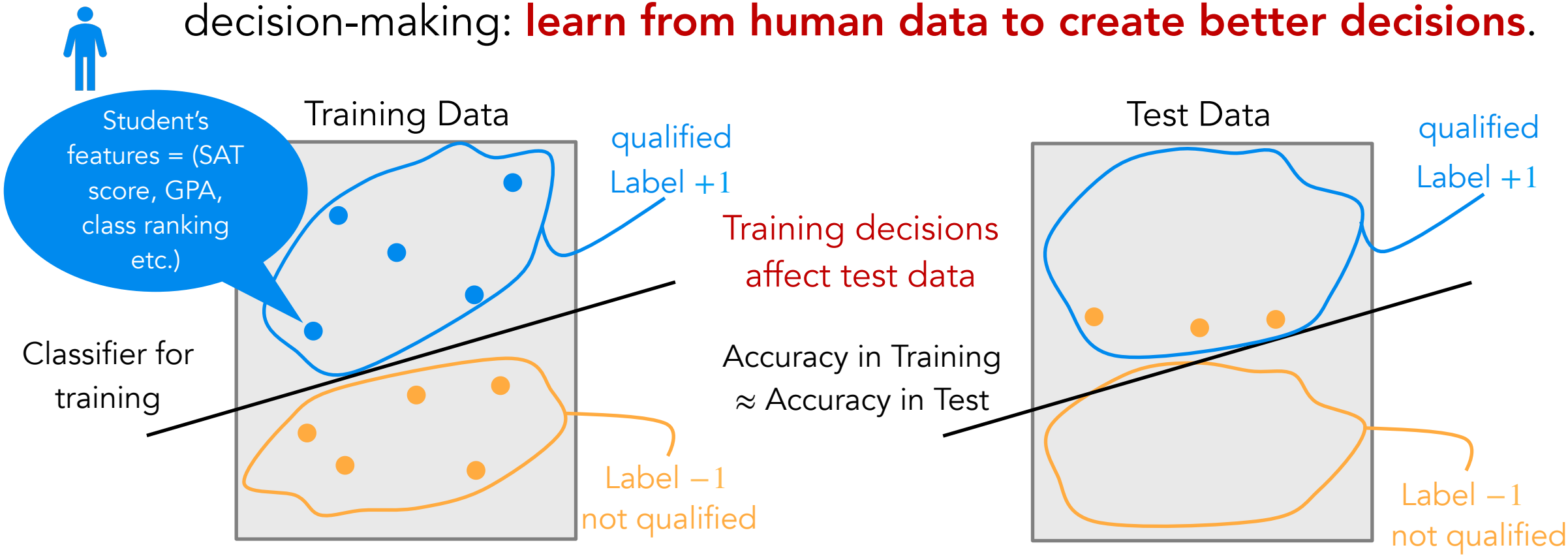
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.



Data **corresponds to individuals who have agency** and want to affect the decisions made on them by the ML algorithms.

What Can Go Wrong?

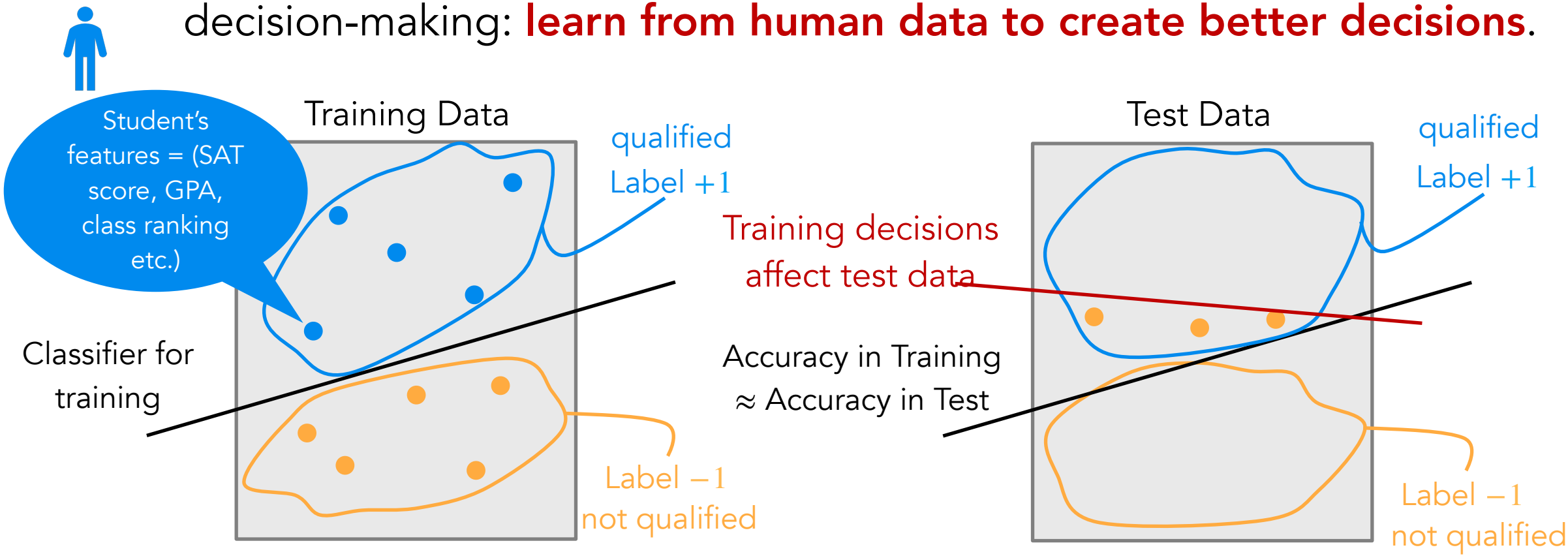
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.



Data **corresponds to individuals who have agency** and want to affect the decisions made on them by the ML algorithms.

What Can Go Wrong?

Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.

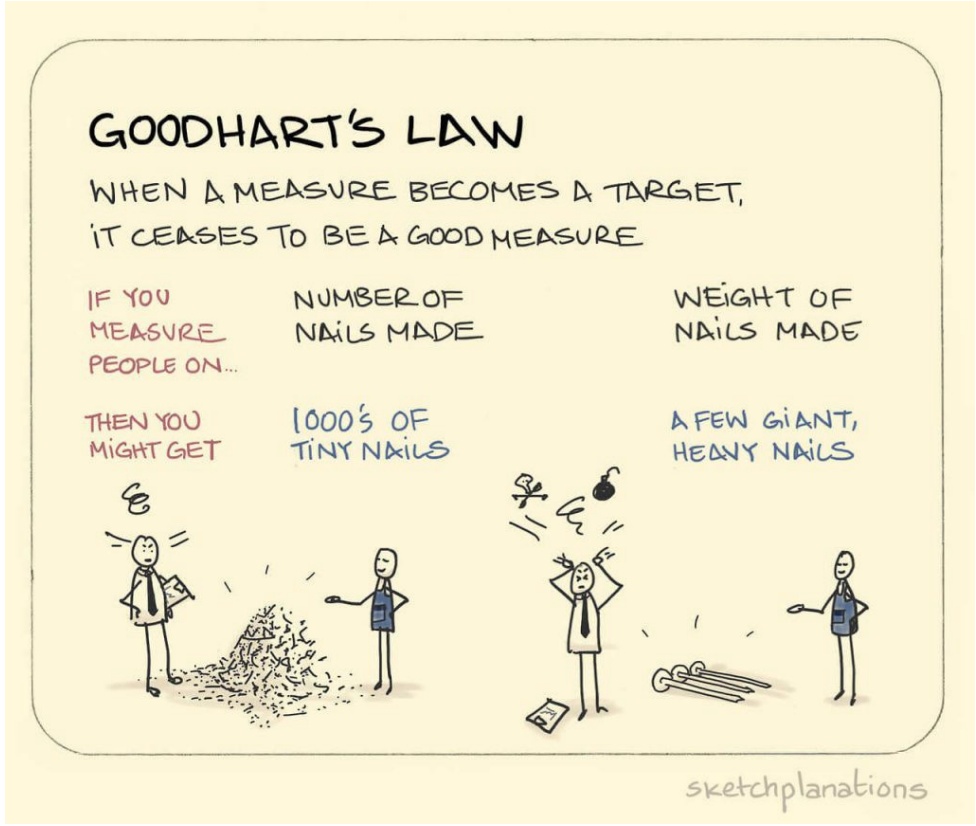


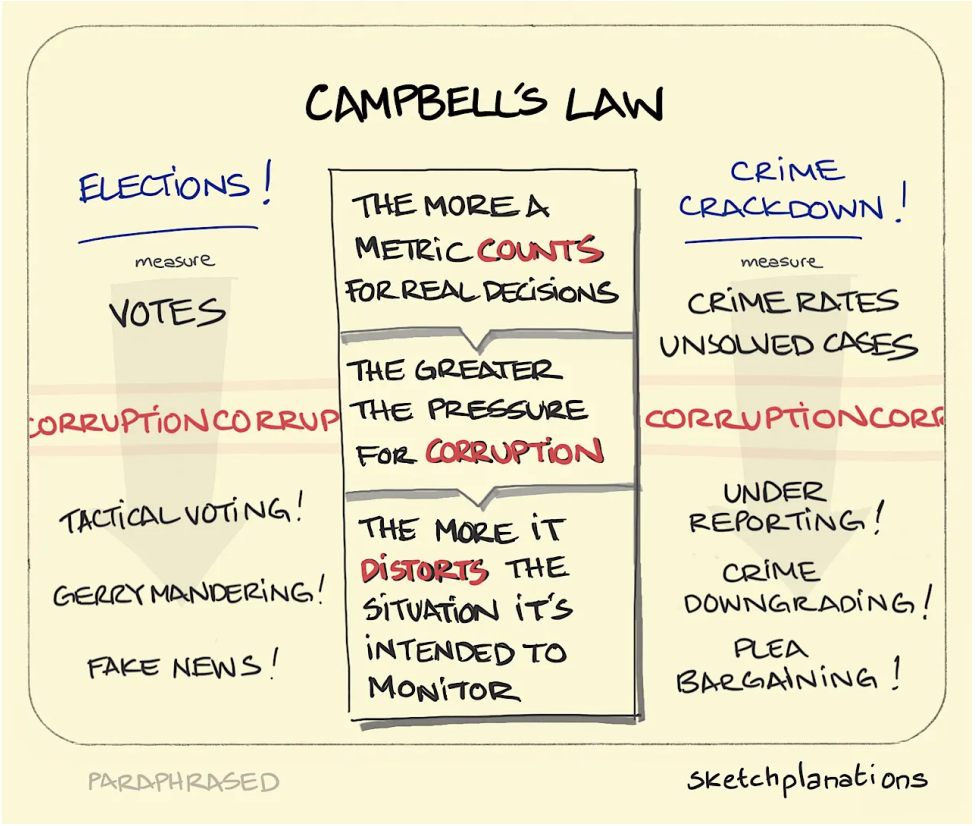
Data **corresponds to individuals who have agency** and want to affect the decisions made on them by the ML algorithms.

Lots of Recent, Exciting Work

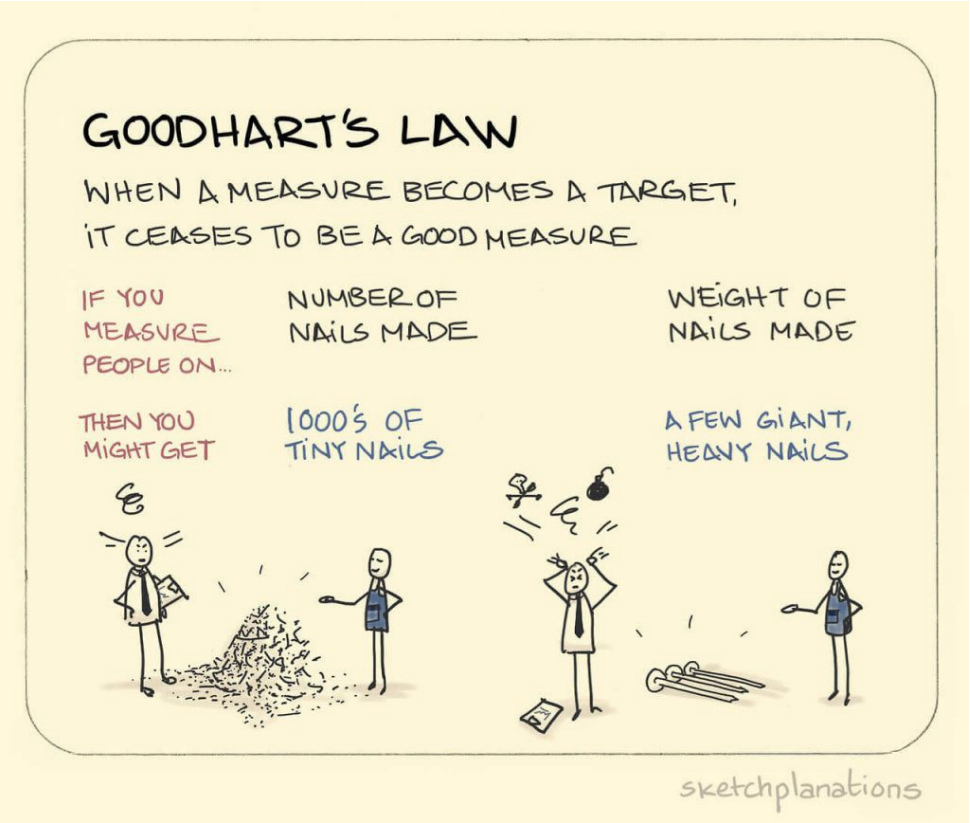
- **Robustness**: [Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**], [Dong, Roth, Schutzman, Waggoner, Wu, **EC18**], [Chen, Liu, **P.**, **NeurIPS20**], [Ahmadi, Beyhaghi, Blum, Naggita, **EC21**], [Sundaraman, Vullikanti, Xu, Yao, **ICML21**], [Ghalme , Nair, Eilat, Talgam-Cohen, Rosenfeld, **ICML21**], [Zrnic, Mazumdar, Sastry, Jordan, **NeurIPS21**], [Jagadeesan, Mendler-Dünner, Hardt, **ICML21**], [Levanon & Rosenfeld, **ICML21**], [Zhang & Conitzer, **ICML21**], [Lechner & Urner, **AAAI22**], [Sundaram, Vullikanti, Xu, Yao, **JMLR23**], [Ahmadi, Blum, Yang, **EC23**], [Shao, Blum, Montasser **NeurIPS23**], [Rosenfeld & Rosenfeld, **ICML24**], [Cohen, Mansour, Moran, Shao, **COLT24**], [Shao, Xie, Yang, **ICML25**]
- **Fairness:** [Milli, Miller, Dragan, Hardt, **FAT*19**], [Hu, Immorlica, Vaughan, **FAT*19**], [Liu, Wilson, Haghtalab, Kalai, Borgs, Chayes, **FAT*19**], [Braverman, Garg, **FORC20**], [Ahmadi, Beyhaghi, Blum, Naggita, **arXiv23**], [Estornell, Das, Liu, Vorobeychik, **FAccT23**]
- **Recourse/Incentivizing Effort:** [Ustun, Spangher, Liu, **FAT*19**], [Kleinberg and Raghavan, **EC19**], [Khajehnejad, Tabibian, Scholkopf, Singla, Gomez-Rodriguez, arXiv19], [Gupta, Nokhiz, Roy, Venkatasubramanian, **arXiv19**], [Chen, Wang, Liu, **arXiv20**], [Tsirtsis, Gomez-Rodriguez, **NeurIPS20**], [Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**], [Harris, Heidari, Wu, **NeurIPS21**], [Bechavod, **P.**, Wu, Ziani, **ICML22**], [Efthymiou, **P.**, Sen, Ziani, **NeurIPS25**], [Efthymiou, Fedorova, **P.**, **arXiv25**]
- **Causality:** [Miller, Milli, Hardt, **FAT*19**], [Shavit, Edelman, Axelrod, **ICML20**], [Bechavod, Ligett, Wu, Ziani, **AISTATS21**], [Horowitz & Rosenfeld, **ICML23**]
- **Performative Prediction:** [Perdomo, Zrnic, Mendler-Dünner, Hardt, **ICML20**], [Mendler-Dünner, Perdomo, Zrnic, Hardt, **NeurIPS20**], [Miller, Perdomo, Zrnic, **ICML21**] [Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner, **ICML22**] **and many (many) more**

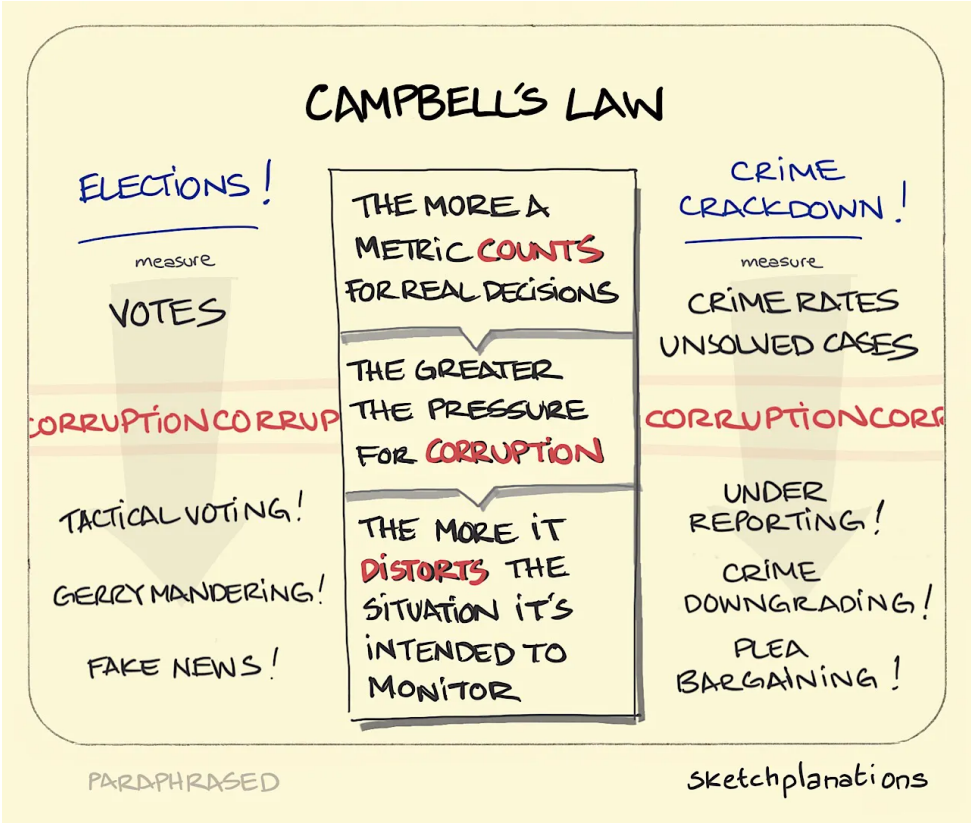
Similar Problem, Different Fields





Similar Problem, Different Fields





- School's admission rule: admit anyone who has more than 100 books in their house.
- Students with (say) 90 and more books can "easily" buy (**but need not read!**) 10 more and get admitted.

→ defeats the purpose of having the # books as a measure of qualifications

The Robustness Perspective

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]

Strategic Classification Offline Model

1. Nature draws agent's features (e.g., SAT score, class ranking, …) $x \in \mathcal{X}$ from distribution $\mathcal{D}$.
2. Learner commits to classifier $\alpha \in \mathcal{A} : \mathcal{X} \to \{-1, +1\}$.
3. Agent observes the classifier α and the x.
4. Agent reports to learner feature vector $\Delta(x)$ ($\neq x$).
5. Learner observes label y
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[y = \alpha(\Delta(x))]$

institution

- **Who?** School/College
- **Goal:** admit most qualified candidates
- **Action:** *linear* classification

individual

- **Who?** Students applying to the school
- **Goal:** be admitted
- **Action:** *strategically* change features

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]

Strategic Classification Offline Model

1. Nature draws agent's features (e.g., SAT score, class ranking, …) $x \in \mathcal{X}$ from distribution $\mathcal{D}$.
2. Learner commits to classifier $\alpha \in \mathcal{A} : \mathcal{X} \to \{-1, +1\}$.
3. Agent observes the classifier α and the x.
4. Agent reports to learner feature vector $\Delta(x)$ ($\neq x$).
5. Learner observes label y
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[y = \alpha(\Delta(x))]$

$$\Delta(x) = \operatorname*{argmax}_{z \in \mathcal{X}} \mathbb{E}_x [\alpha(z) - c(x, z)]$$

value for passing classifier

manipulation cost

$c(x, z)$: "separable", i.e.,

$$c(x, z) = \max\{0, c_2(z) - c_1(x)\}$$

Goal: Compute Stackelberg Equilibrium

$$\alpha^\star = \arg\max_{f \in \mathcal{H}} \Pr_{x \sim \mathcal{D}}[y = f(\Delta(x))]$$

Main Result

Algorithm that learns $\alpha^\star$ with **polynomial time** and **sample complexity**.

Strategic Classification Online Model

For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, …) $x_t \in \mathcal{X} \subseteq [0,1]^d$.
2. Learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.
3. Agent observes the classifier α_t and the datapoint (x_t, y_t), where $y_t \in \{-1, 1\}$.
4. Agent reports to learner feature vector $\hat{x}(a_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs classification loss: $\ell(\alpha_t, \hat{x}_t(\alpha_t))$

institution

- **Who?** School/College
- **Goal:** admit most qualified candidates
- **Action:** *linear* classification

Goal: Minimize Stackelberg Regret

$$\mathcal{R}(T) = \sum_{t\in[T]} \ell(\alpha_t, \hat{x}_t(\alpha_t)) - \min_{\alpha^\star \in \mathcal{A}} \sum_{t\in[T]} \ell(\alpha^\star, \hat{x}_t(\alpha^\star))$$

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]

Strategic Classification Online Model

For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, …) $x_t \in \mathcal{X} \subseteq [0,1]^d$.
2. Learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.
3. Agent observes the classifier α_t and the datapoint (x_t, y_t), where $y_t \in \{-1, 1\}$.
4. Agent reports to learner feature vector $\hat{x}(a_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs hinge / logistic classification loss:

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = \max(0, 1 - y_t \cdot \alpha_t(\hat{x}_t(\alpha_t)) \text{ or}$$

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = \log(1 + e^{-y_t \langle \alpha_t, \hat{x}_t(\alpha_t) \rangle})$$

institution

- **Who?** School/College
- **Goal:** admit most qualified candidates
- **Action:** *linear* classification

Goal: Minimize Stackelberg Regret

$$\mathcal{R}(T) = \sum_{t \in [T]} \ell(\alpha_t, \hat{x}_t(\alpha_t)) - \min_{\alpha^\star \in \mathcal{A}} \sum_{t \in [T]} \ell(\alpha^\star, \hat{x}_t(\alpha^\star))$$

Strategic Classification Online Model

[Chen, P., Liu, **NeurIPS20**]

[Ahmadi, Beyhaghi, Blum, Naggita, **EC21**]

For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, …) $x_t \in \mathcal{X} \subseteq [0,1]^d$.
2. Learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.
3. Agent observes the classifier α_t and the datapoint (x_t, y_t), where $y_t \in \{-1, 1\}$.
4. Agent reports to learner feature vector $\hat{x}(a_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs binary classification loss:

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = 1\{y_t \neq \text{PredictedLabel}(\hat{x}_t(\alpha_t), \alpha_t)\}$$

institution

- **Who?** School/College
- **Goal:** admit most qualified candidates
- **Action:** *linear* classification

Goal: Minimize Stackelberg Regret

$$\mathcal{R}(T) = \sum_{t\in[T]} \ell(\alpha_t, \hat{x}_t(\alpha_t)) - \min_{\alpha^\star \in \mathcal{A}} \sum_{t\in[T]} \ell(\alpha^\star, \hat{x}_t(\alpha^\star))$$

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]

For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, …) $x_t \in \mathcal{X} \subseteq [0,1]^d$.
2. Learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.
3. Agent observes the classifier α_t and the datapoint (x_t, y_t), where $y_t \in \{-1, 1\}$.
4. Agent reports to learner feature vector $\hat{x}(a_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs hinge / logistic classification loss:

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = \max(0, 1 - y_t \cdot \alpha_t(\hat{x}_t(\alpha_t)) \text{ or}$$

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = \log(1 + e^{-y_t \langle \alpha_t, \hat{x}_t(\alpha_t) \rangle})$$

Myopically Rational Agents

$$\hat{x}_t(\alpha_t) = \arg\max_{x'} \langle \alpha_t, x' \rangle - \underbrace{\text{cost}(x, x')}_{\text{convex cost}}$$

individual

- **Who?** Students applying to the school
- **Goal:** be admitted
- **Action:** *strategically* change features

[Chen, P., Liu, NeurIPS20]

Strategic Classification Online Model

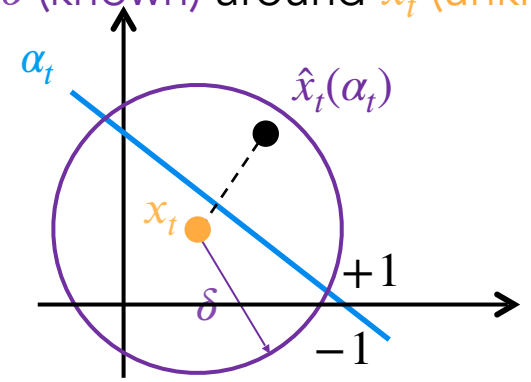
For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, …) $x_t \in \mathcal{X} \subseteq [0,1]^d$.
2. Learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.
3. Agent observes the classifier α_t and the datapoint (x_t, y_t), where $y_t \in \{-1, 1\}$.
4. Agent reports to learner feature vector $\hat{x}(a_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs binary classification loss:

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = 1\{y_t \neq \text{PredictedLabel}(\hat{x}_t(\alpha_t), \alpha_t)\}$$

δ-Bounded Myopically Rational Agents

Agents can only misreport in ball of radius δ (known) around x_t (unknown).



individual

- **Who?** Students applying to the school
- **Goal:** be admitted
- **Action:** *strategically* change features

Strategic Classification Online Model

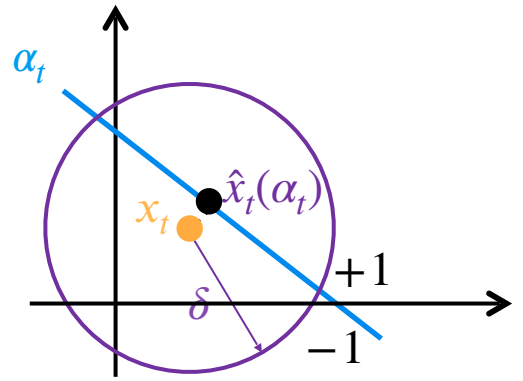
For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, …) $x_t \in \mathcal{X} \subseteq [0,1]^d$.
2. Learner picks classification rule $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.
3. Agent observes the classifier α_t and the datapoint (x_t, y_t), where $y_t \in \{-1, 1\}$.
4. Agent reports to learner feature vector $\hat{x}(a_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs binary classification loss:

$$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = 1\{y_t \neq \text{PredictedLabel}(\hat{x}_t(\alpha_t), \alpha_t)\}$$

Myopically Rational Agents

Value func = binary, cost func = L1 / L2



individual

- **Who?** Students applying to the school
- **Goal:** be admitted
- **Action:** *strategically* change features

Main Results

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]

Value func: linear & cost func: convex + positive homogenous

→ bandit convex opt → $\mathscr{R}(T) = O(\sqrt{d}T^{3/4})$

Myopically Rational Agents

$\hat{x}_t(\alpha_t) = \arg\max_{x' \in X} \langle \alpha_t, x' \rangle - cost(x, x')$

value for passing classifier — convex cost

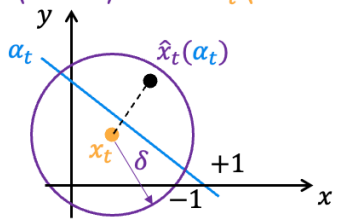
[Chen, Liu, **P.**, **NeurIPS20**]

$$\mathscr{R}(T) = O\left(\sqrt{T \log^2(T \cdot F(\delta)}\right),$$

where $F(\delta)$: function that depends on $(\delta, \{x_t\}_t)$ terms

δ-Bounded Myopically Rational Agents

Agents can only misreport in ball of radius δ (known) around x_t (unknown).



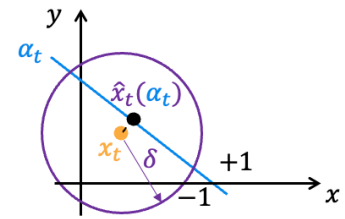
[Ahmadi, Beyhaghi, Blum, Naggita, **EC21**]

If data are linearly separable with margin γ:

L2: $\mathscr{R}(T) = O\left(\frac{(1 + \text{ManipulationPower})^2}{\gamma^2}\right)$

Myopically Rational Agents

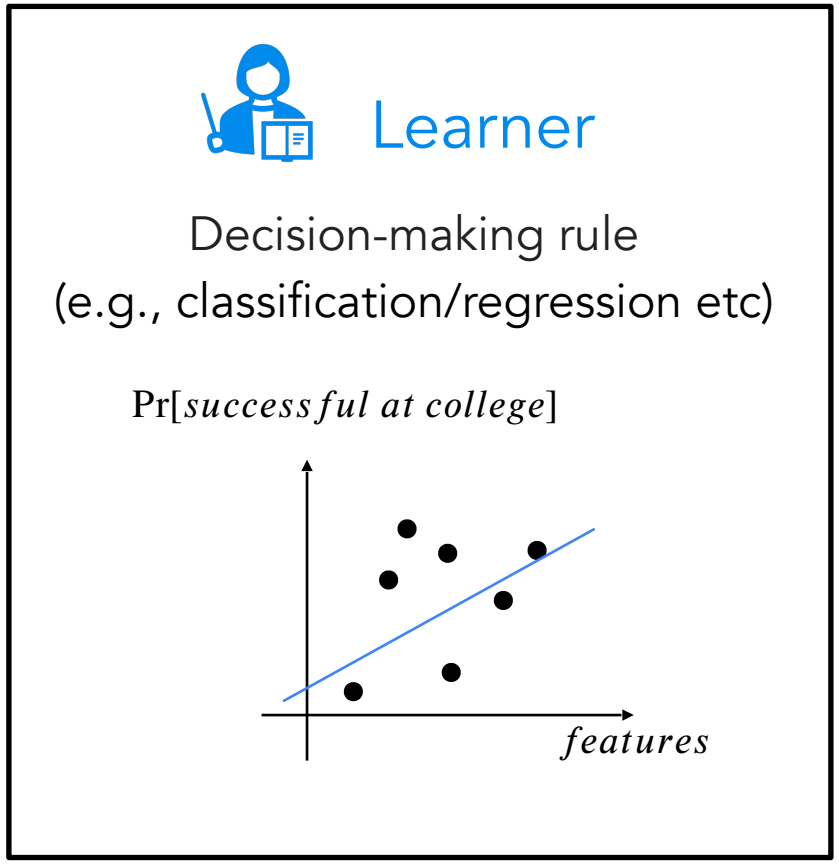
Value func = binary, cost func = L1 / L2



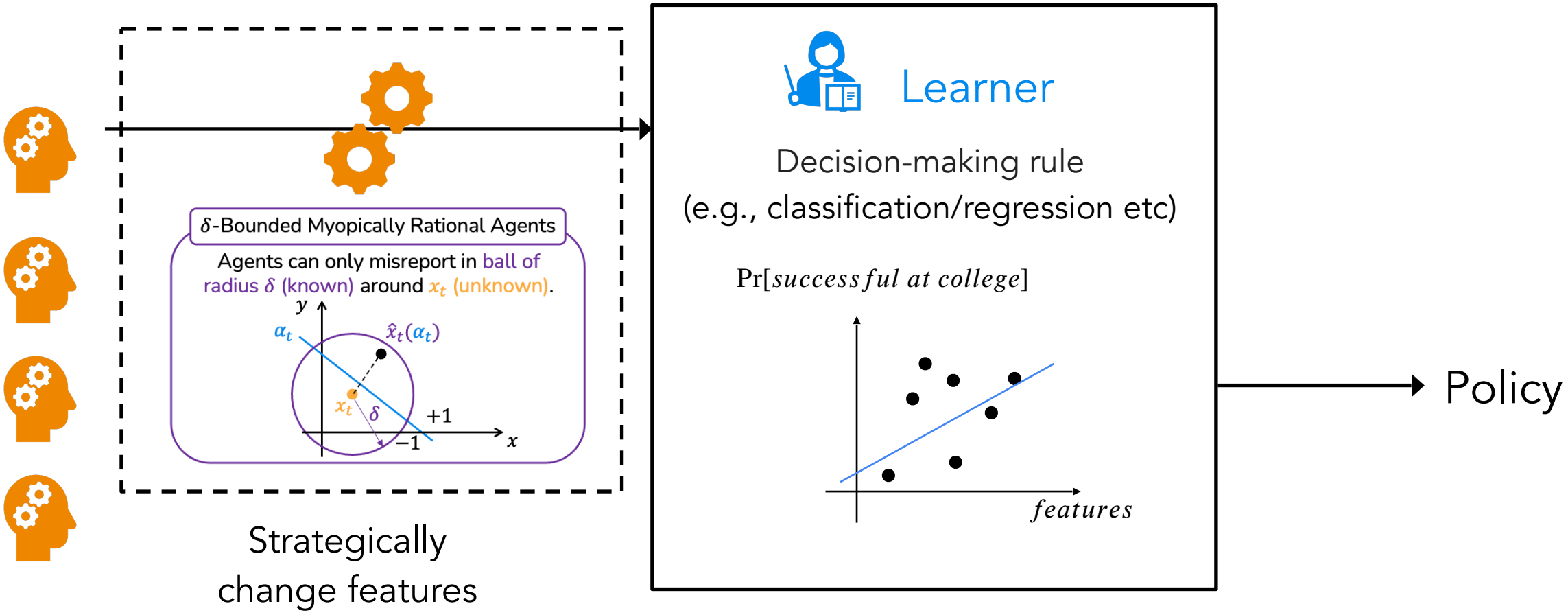
The Fairness Perspective

Implicit Assumption so Far: Homogeneous Population

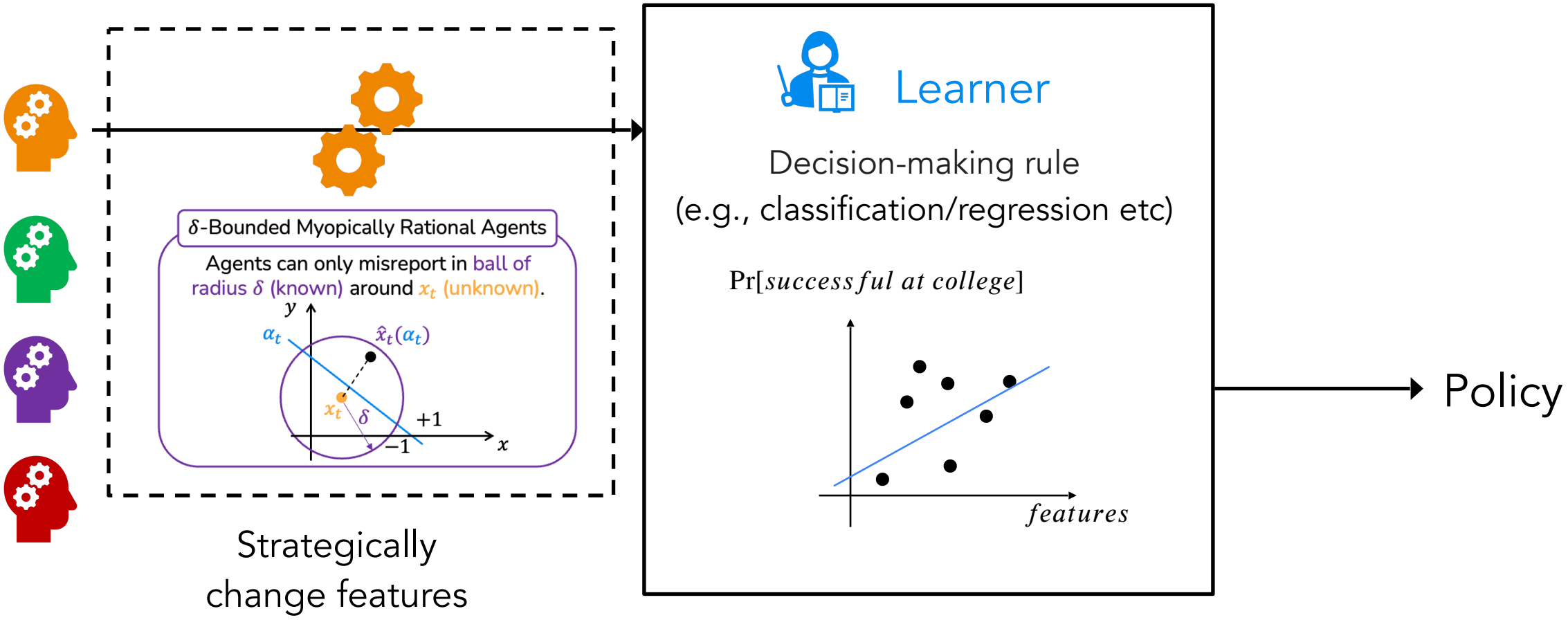
Implicit Assumption so Far: Homogeneous Population



Implicit Assumption so Far: Homogeneous Population



Reality: Highly Heterogeneous!



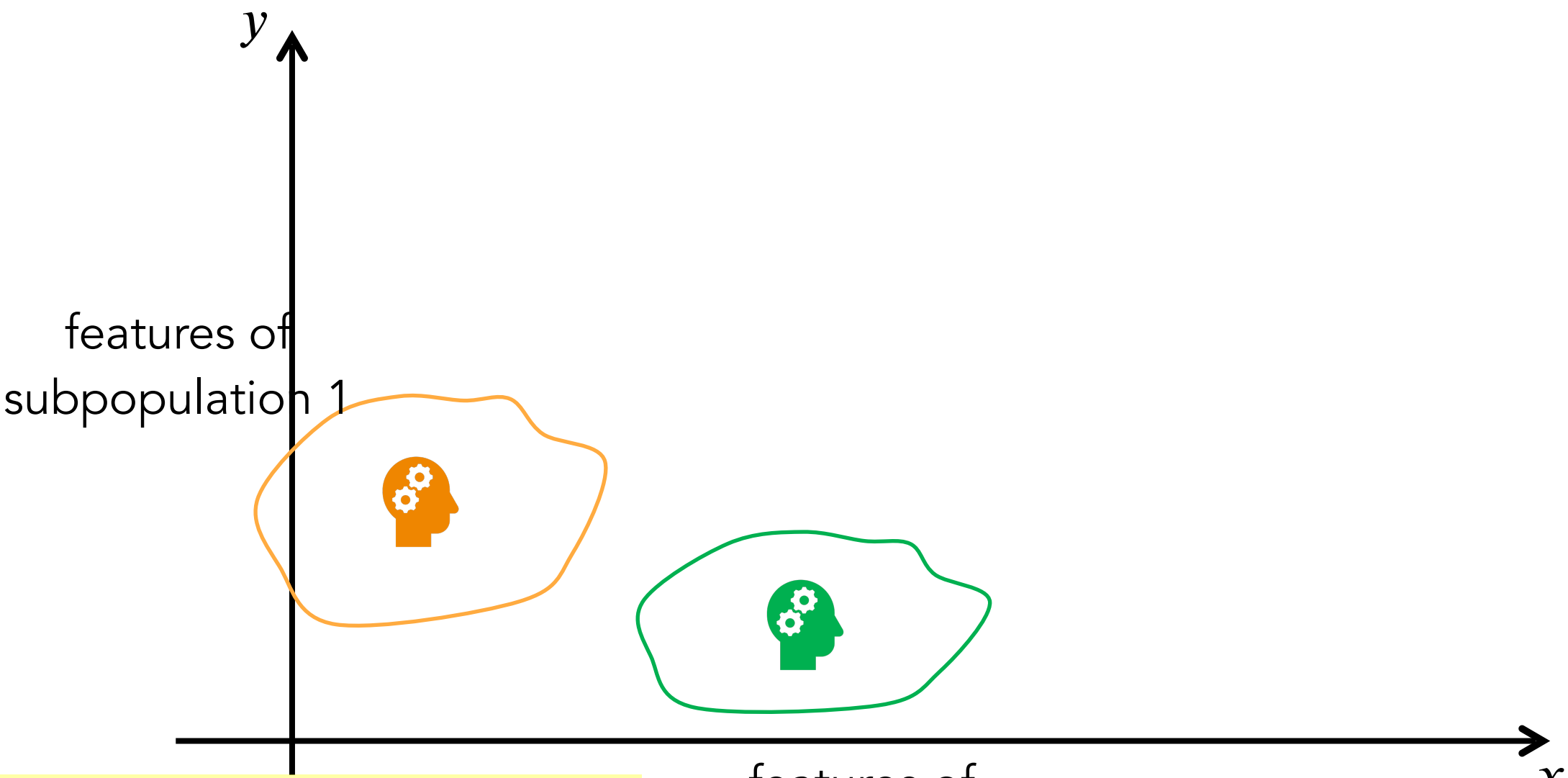
Reality: Highly Heterogeneous!

[Hu, Immorlica, Vaughan, **FAT*19**]

[Milli, Miller, Dragan, Hardt, **FAT*19**]

Reality: Highly Heterogeneous!

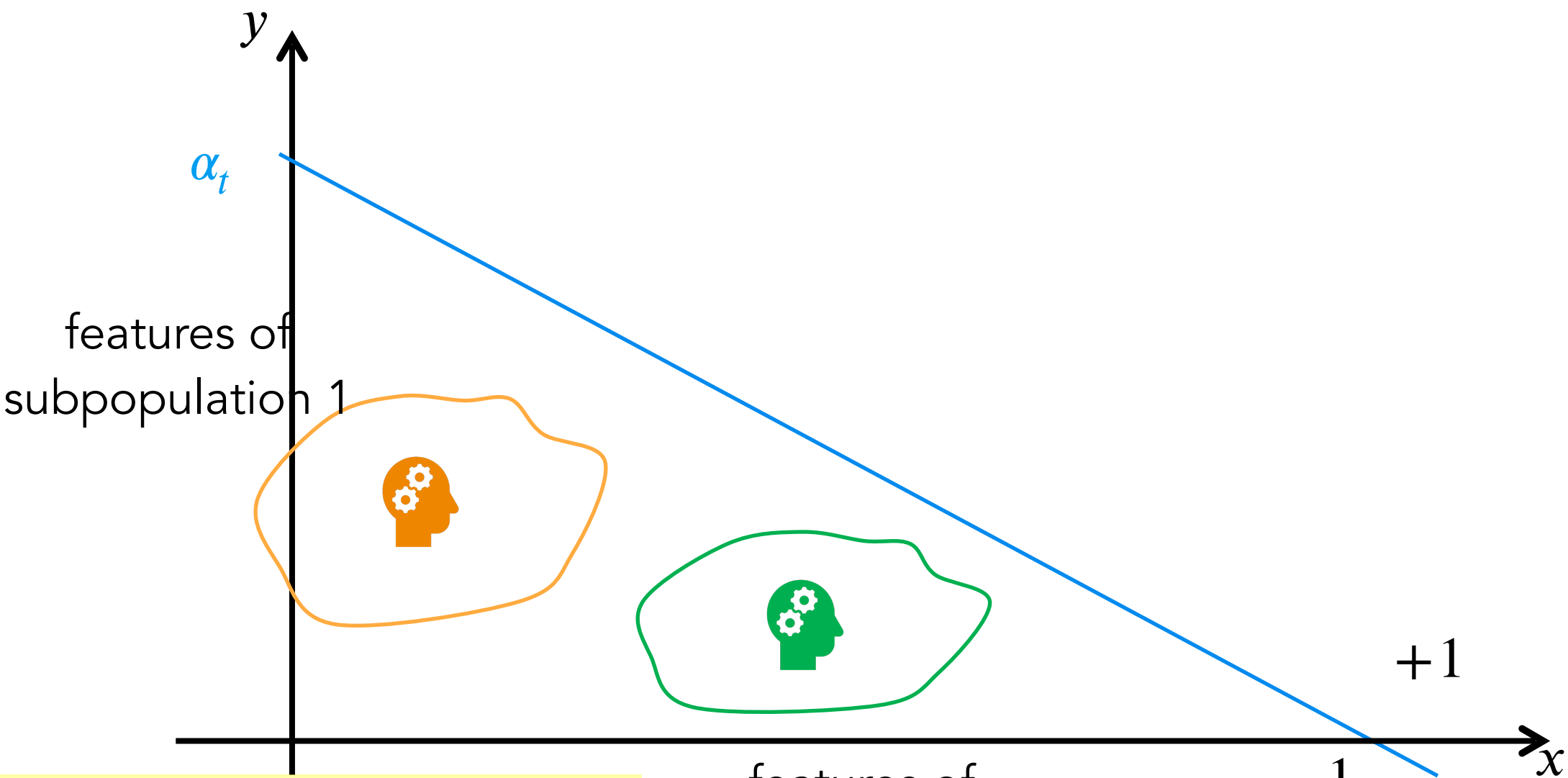
[Hu, Immorlica, Vaughan, **FAT*19**]
[Milli, Miller, Dragan, Hardt, **FAT*19**]



Assumption: Both groups include qualified and non-qualified individuals.

Reality: Highly Heterogeneous!

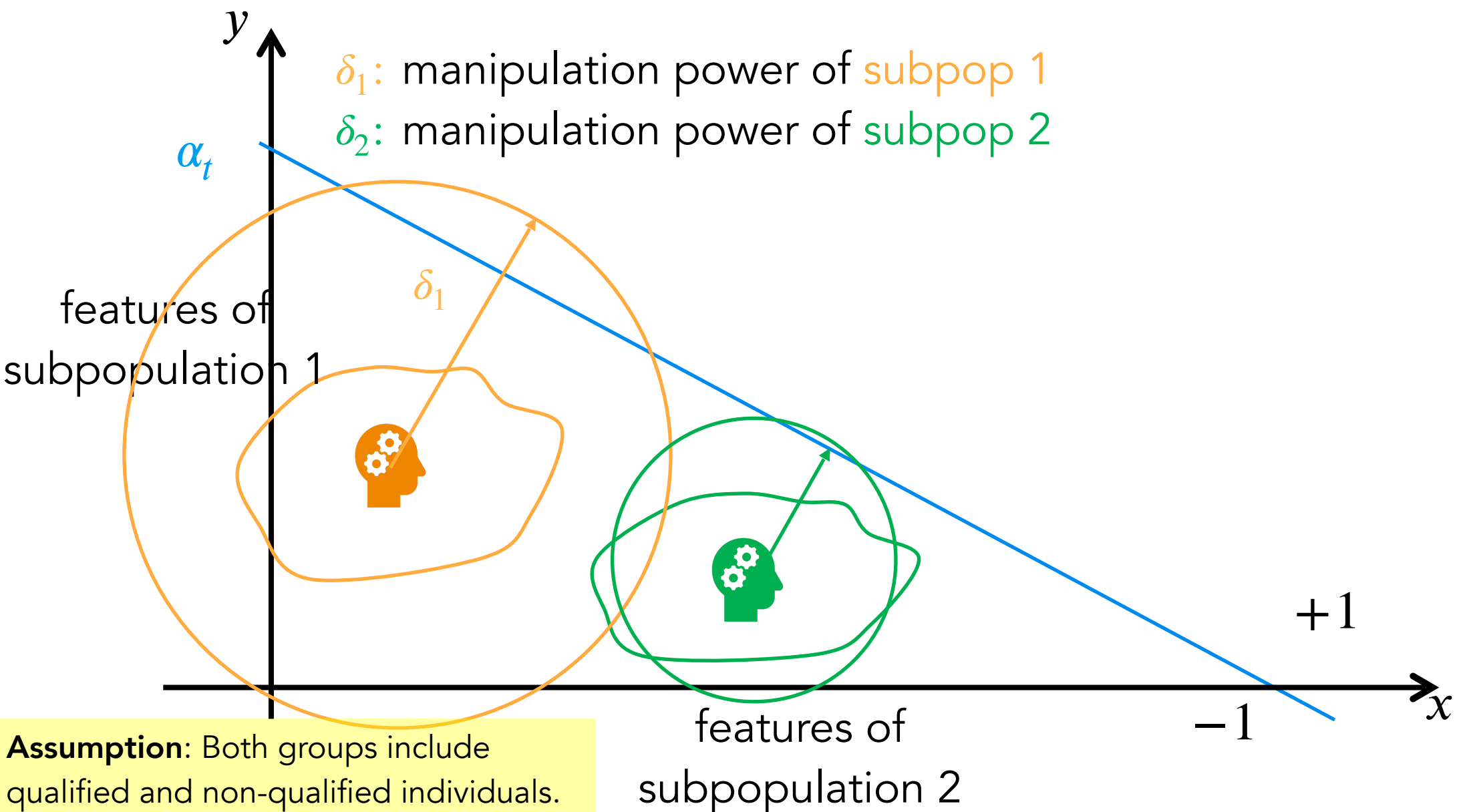
[Hu, Immorlica, Vaughan, **FAT*19**]
[Milli, Miller, Dragan, Hardt, **FAT*19**]



Assumption: Both groups include qualified and non-qualified individuals.

Reality: Highly Heterogeneous!

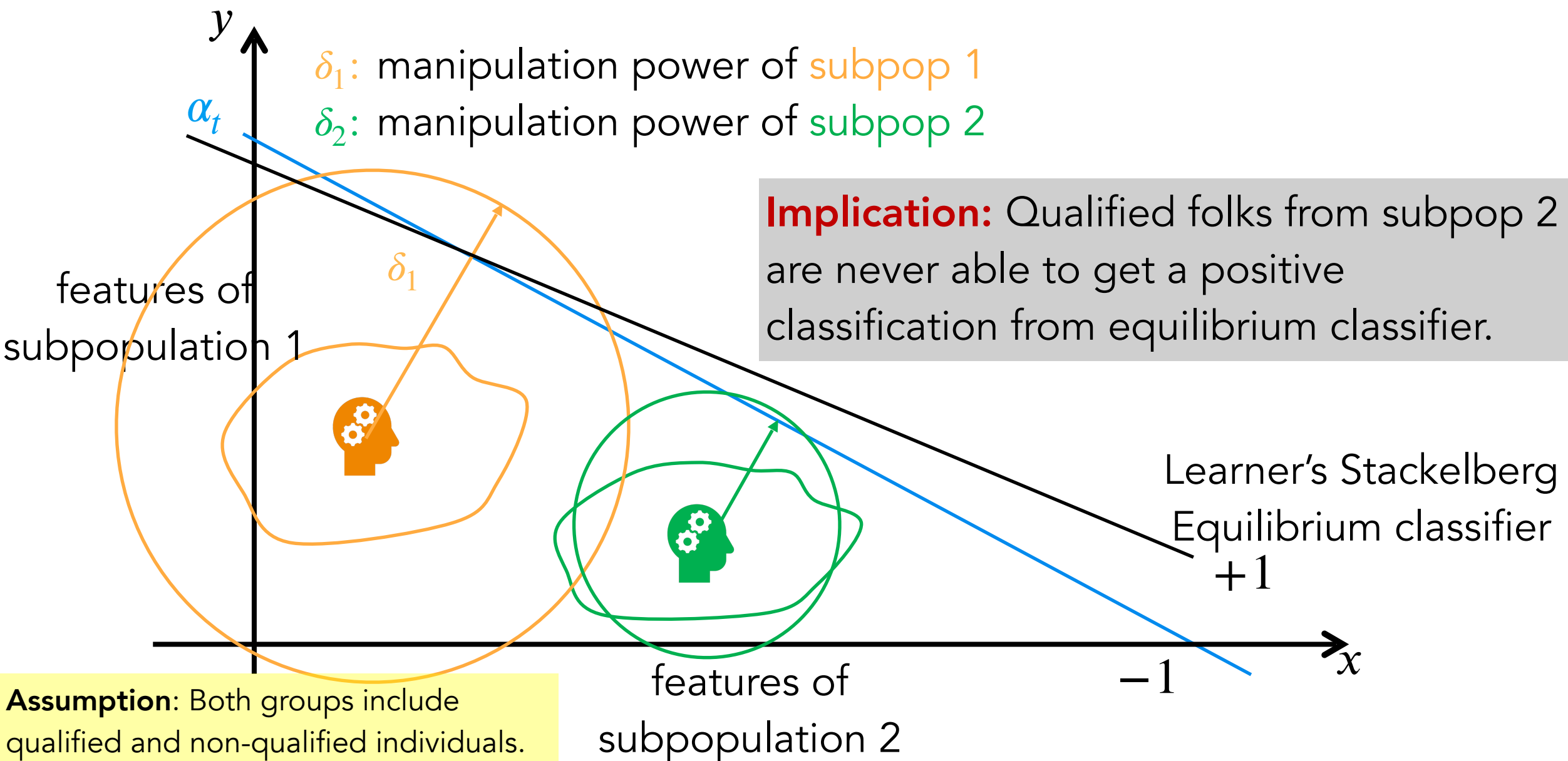
[Hu, Immorlica, Vaughan, **FAT*19**]
[Milli, Miller, Dragan, Hardt, **FAT*19**]



Assumption: Both groups include qualified and non-qualified individuals.

Reality: Highly Heterogeneous!

[Hu, Immorlica, Vaughan, **FAT*19**]
[Milli, Miller, Dragan, Hardt, **FAT*19**]

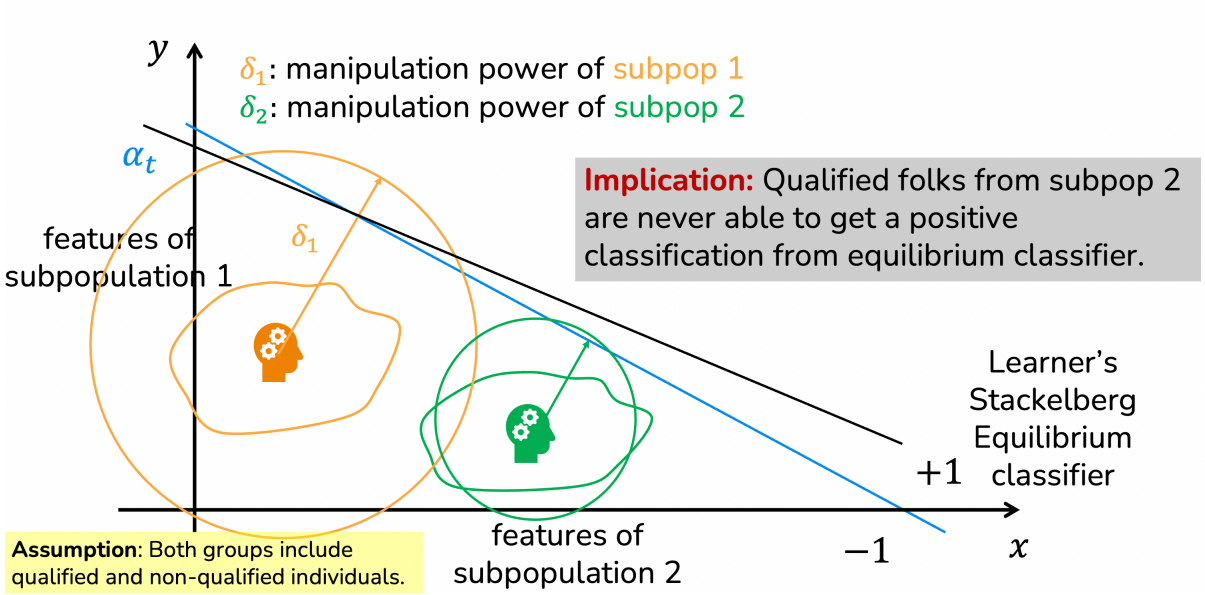


Implication: Qualified folks from subpop 2 are never able to get a positive classification from equilibrium classifier.

Assumption: Both groups include qualified and non-qualified individuals.

Reality: Highly Heterogeneous!

[Hu, Immorlica, Vaughan, **FAT*19**]
[Milli, Miller, Dragan, Hardt, **FAT*19**]



Summary

1) Strategic classification **disproportionately affects** disadvantaged population.

2) There are cases where **subsidies make both subpopulations worse off**, while making the **learner better off**.

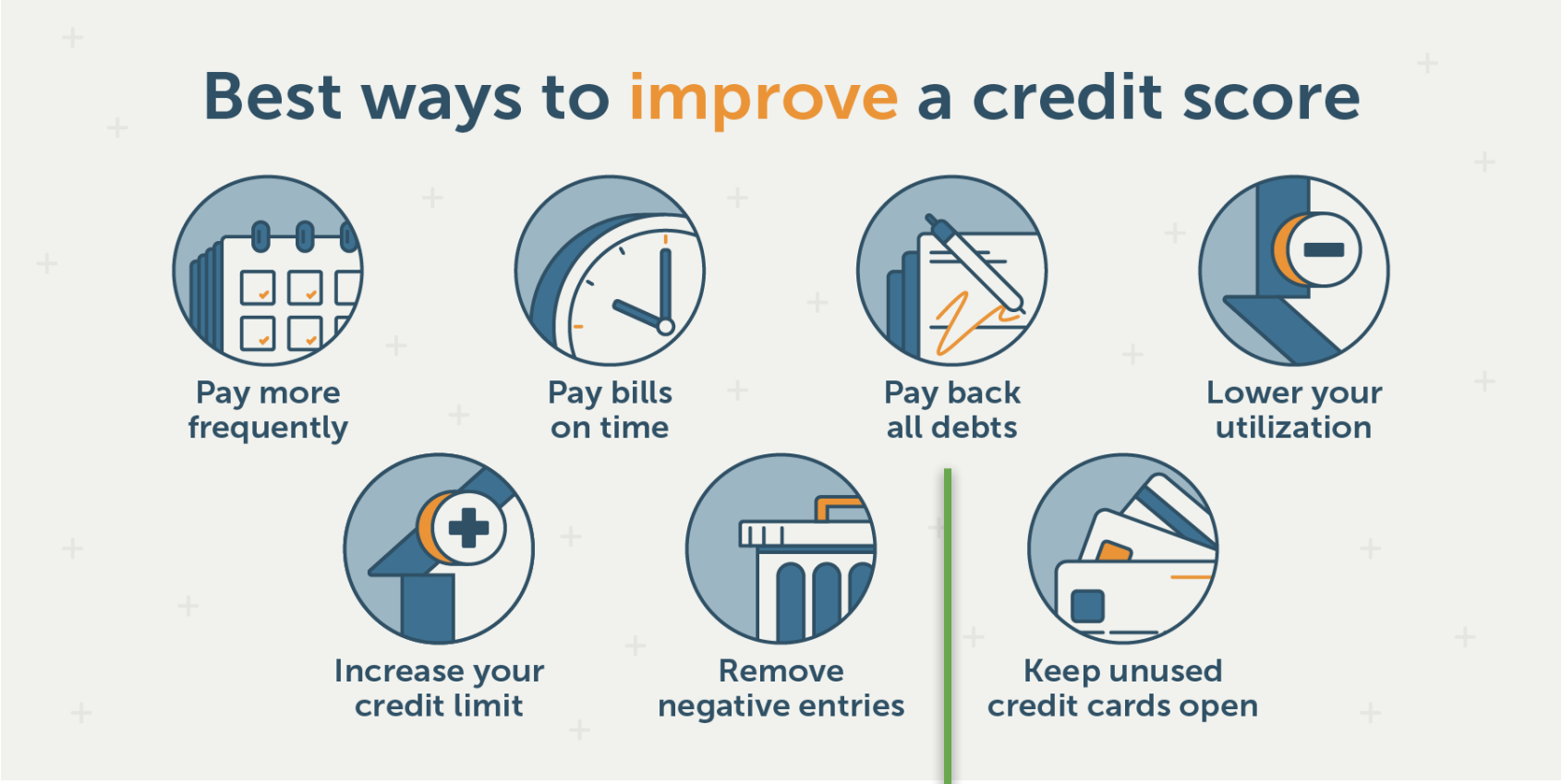
3) Insights hold for cases where classification rule is revealed to agents.

The Causality / Improvement Perspective

Is It All Just Gaming?

source: https://www.lexingtonlaw.com/credit/how-to-build-credit

Is It All Just Gaming?

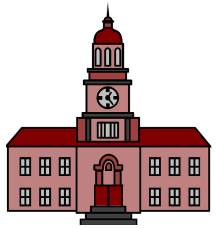


source: https://www.lexingtonlaw.com/credit/how-to-build-credit

ability to pay back future loans

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]

institution



Model

> Prediction

Features

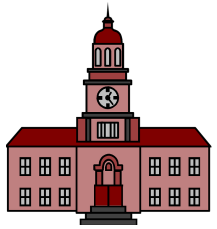
True label

GPA

individual

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]

institution



Model

> Prediction

Features

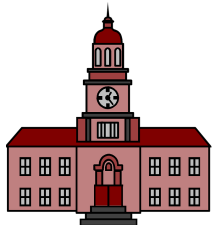
True label

GPA

individual

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]

institution



Model

> Prediction

Features

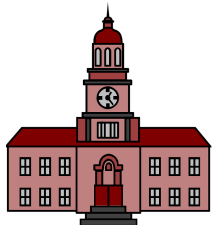
True label

GPA

individual

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]

institution



Model

> Prediction

Features

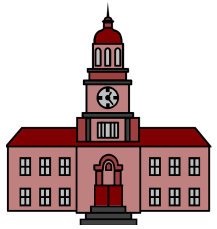
True label

GPA

individual

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]

institution



Model

> Prediction

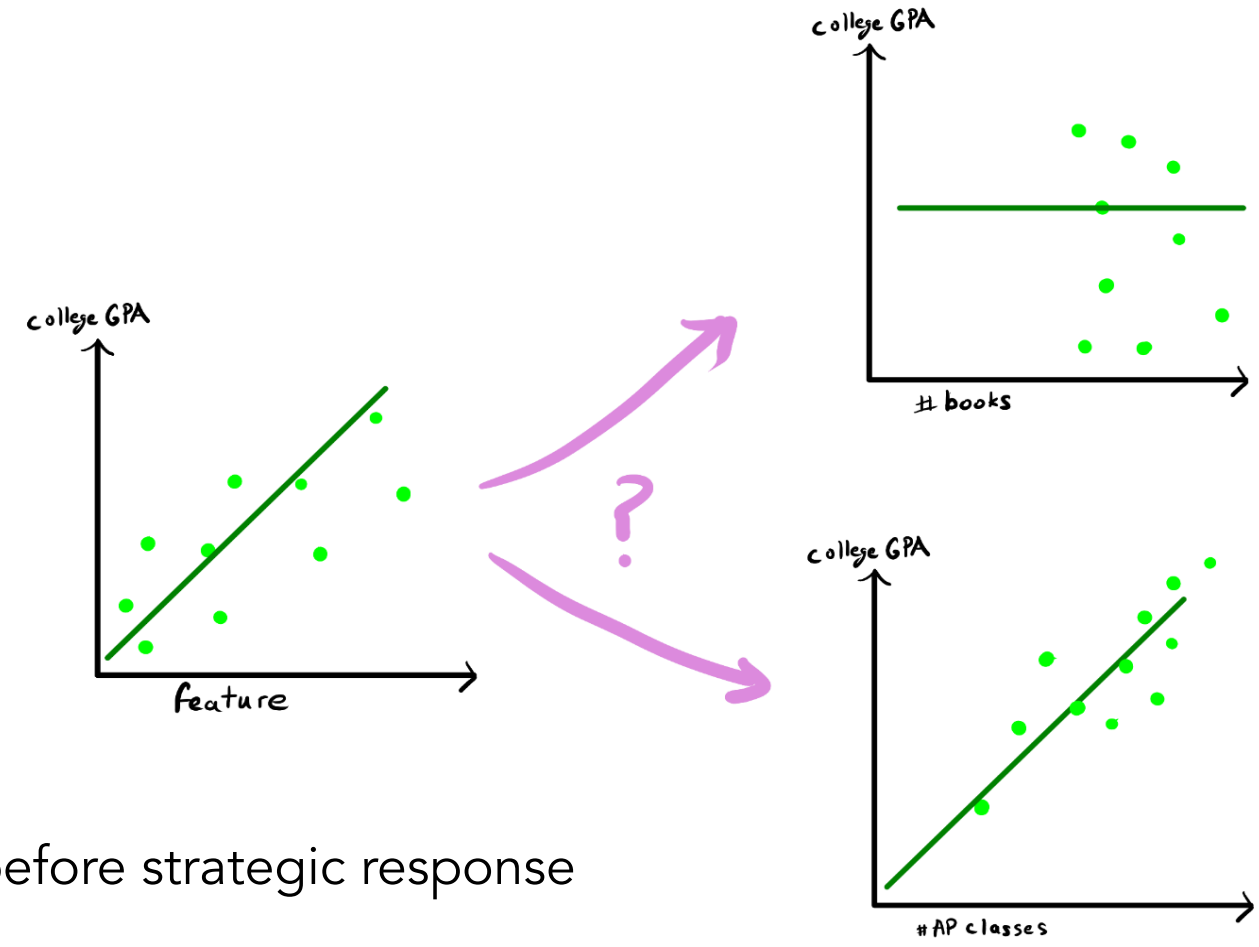
Features

True label

GPA

individual

Causation or Just Correlation?



before strategic response

after strategic response

Gaming

Improvement

Gaming vs Improvement

Incentivizing Desirable Effort Profiles in Strategic Classification: The Role of Causality and Uncertainty

Valia Efthymiou* Chara Podimata† Diptangshu Sen‡ Juba Ziani§

February 11, 2025

Abstract

We study strategic classification in binary decision-making settings where agents can modify their features in order to improve their classification outcomes. Importantly, our work considers the causal structure across different features, acknowledging that effort in a given feature may affect other features. The main goal of our work is to understand *when and how much agent effort is invested towards desirable features*, and how this is influenced by the deployed classifier, the causal structure of the agent's features, their ability to modify them, and the information available to the agent about the classifier and the feature causal graph.

In the complete information case, when agents know the classifier and the causal structure of the problem, we derive conditions ensuring that rational agents focus on features favored by the principal. We show that designing classifiers to induce desirable behavior is generally non-convex, though tractable in special cases. We also extend our analysis to settings where agents have incomplete information about the classifier or the causal graph. While optimal effort selection is again a non-convex problem under general uncertainty, we highlight special cases of partial uncertainty where this selection problem becomes tractable. Our results indicate that uncertainty drives agents to favor features with higher expected importance and lower variance, potentially misaligning with principal preferences. Finally, numerical experiments based on a cardiovascular disease risk study illustrate how to incentivize desirable modifications under uncertainty.

Desirable Effort Fairness and Optimality Trade-offs in Strategic Learning

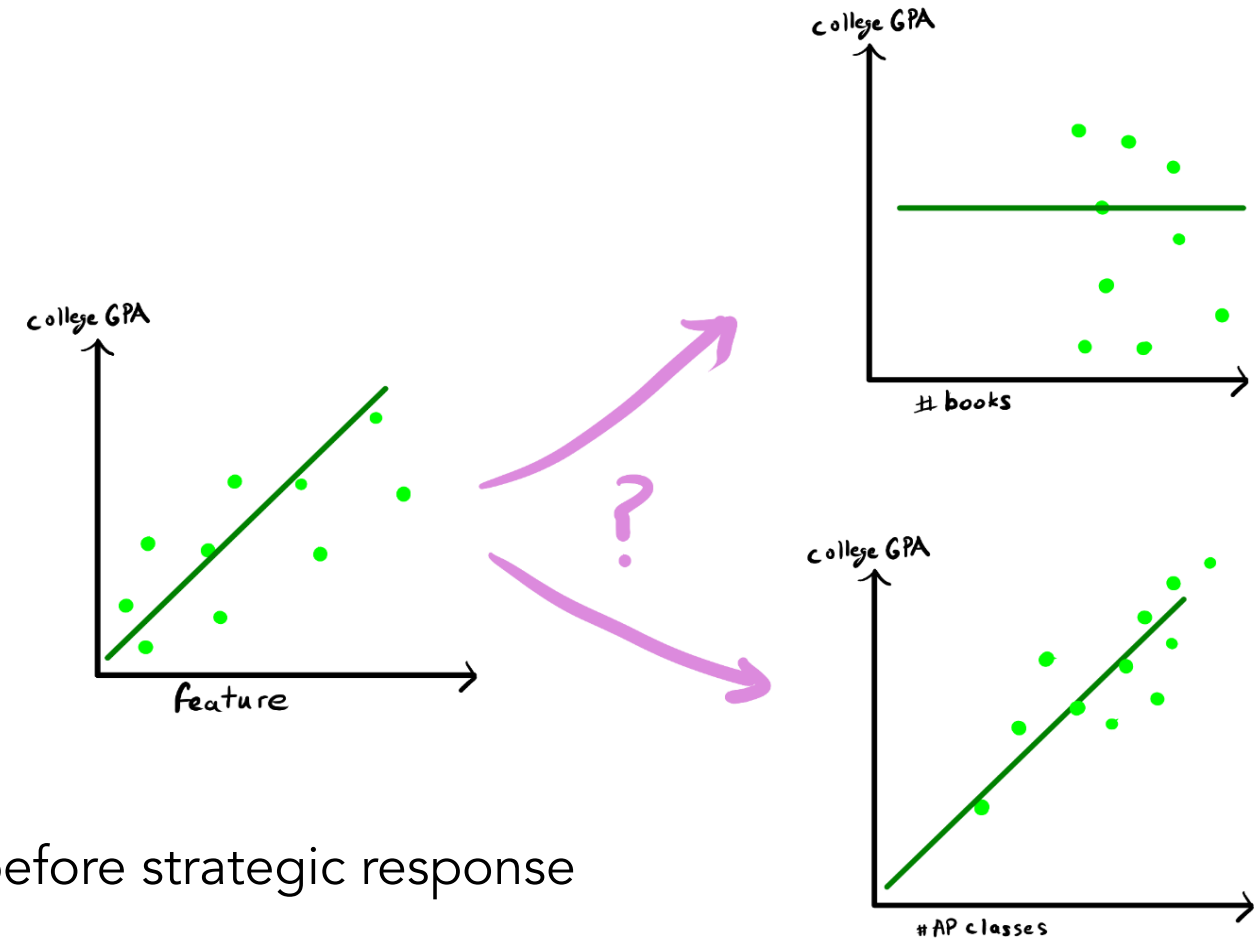
Valia Efthymiou* Ekaterina Fedorova† Chara Podimata‡

October 2025

Abstract

Strategic learning studies how decision rules interact with agents who may strategically change their inputs/features to achieve better outcomes. In standard settings, models assume that the decision-maker's sole scope is to learn a classifier that maximizes an objective (e.g., accuracy) assuming that agents best respond. However, real decision-making systems' goals do not align *exclusively* with producing good predictions. They may consider the external effects of inducing certain incentives, which translates to the change of certain features being more *desirable* for the decision maker. Further, the principal may also need to incentivize desirable feature changes fairly across heterogeneous agents. *How much does this constrained optimization (i.e., maximize the objective, but restrict agents' incentive disparity) cost the principal?* We propose a unified model of principal-agent interaction that captures this trade-off under three additional components: (1) causal dependencies between features, such that changes in one feature affect others; (2) heterogeneous manipulation costs between agents; and (3) peer learning, through which agents infer the principal's rule. We provide theoretical guarantees on the principal's optimality loss constrained to a particular desirability fairness tolerance for multiple broad classes of fairness measures. Finally, through experiments on real datasets, we show the explicit tradeoff between maximizing accuracy and fairness in desirability effort.

Causation or Just Correlation?



before strategic response

after strategic response

Gaming

Improvement

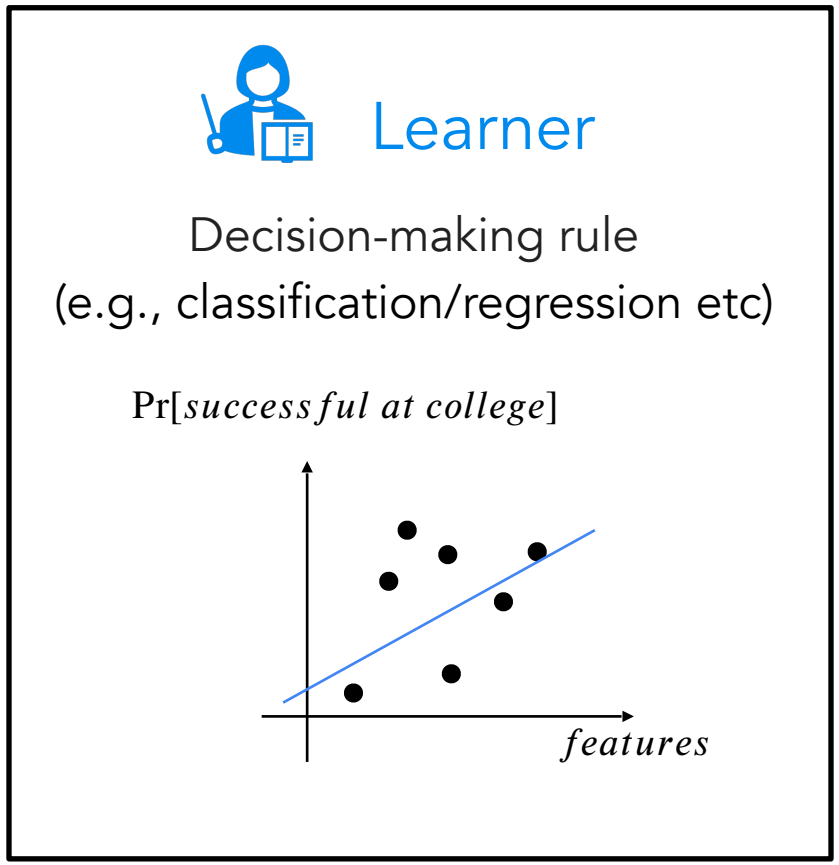
no cost to **transparency**

Is the "transparency" assumption realistic?

Strategic Learning Revisited

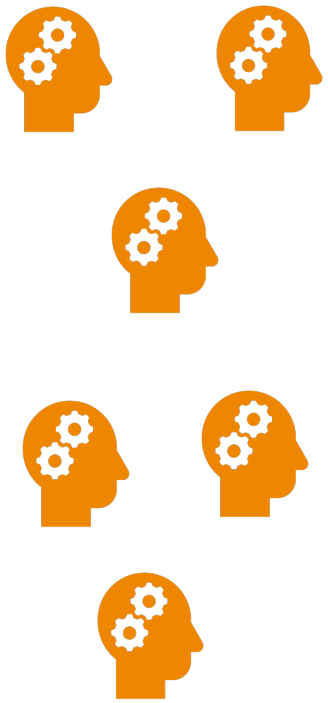
[Bechavod, P., Wu, Ziani, **ICML22**]

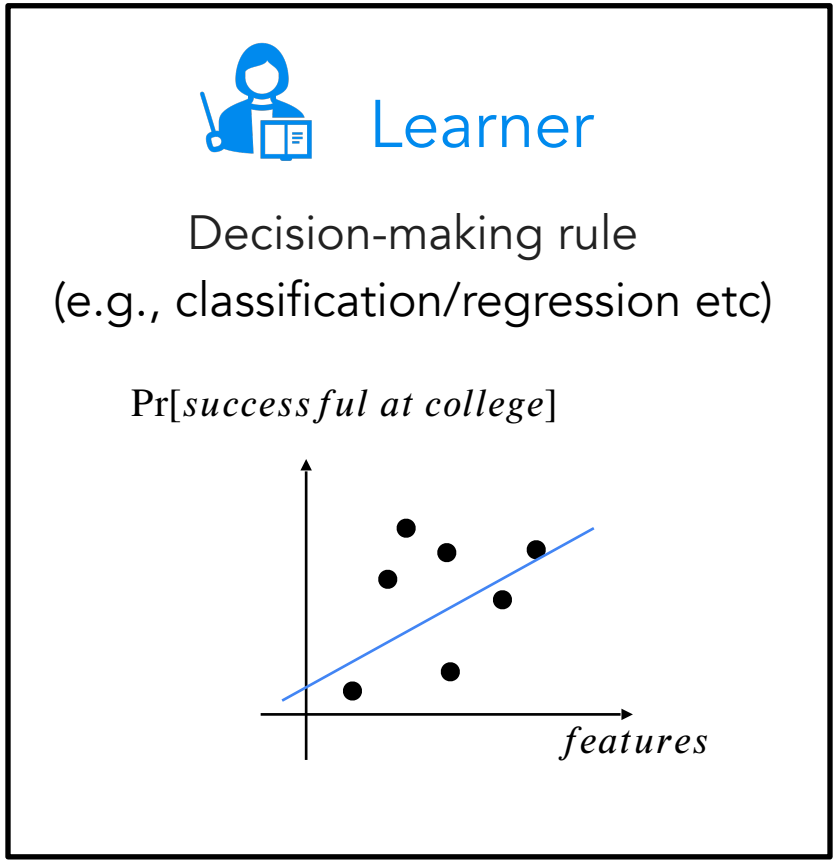
Strategic Learning Revisited



[Bechavod, P., Wu, Ziani, **ICML22**]

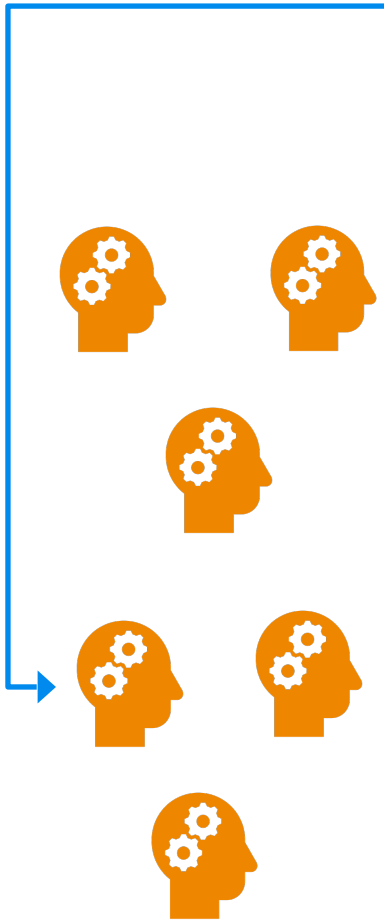
Strategic Learning Revisited

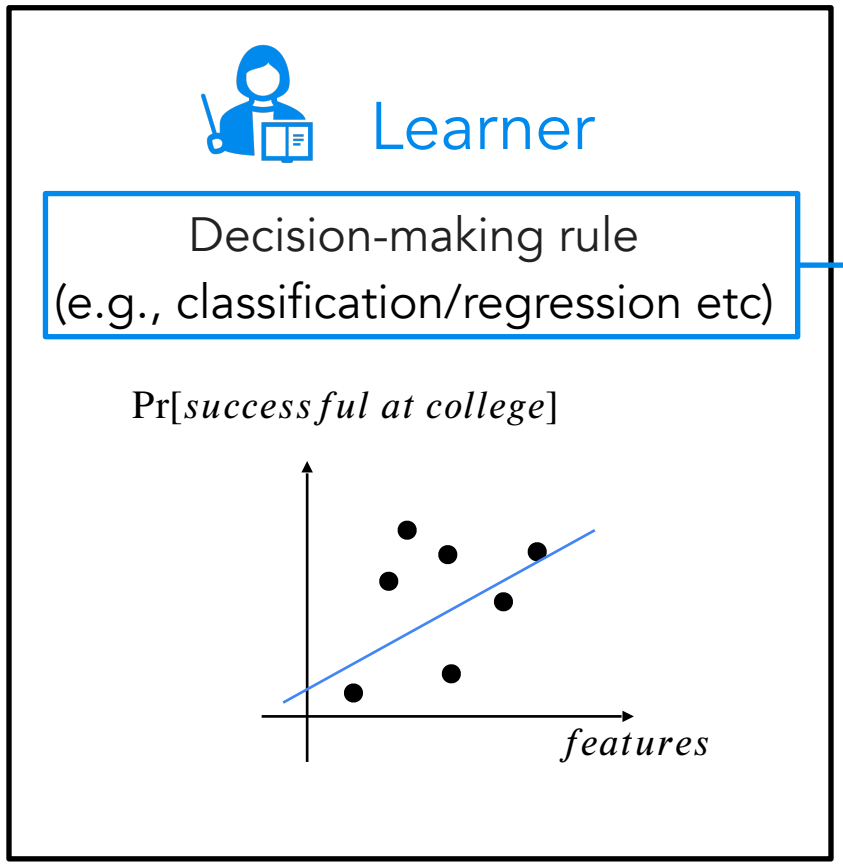




[Bechavod, P., Wu, Ziani, **ICML22**]

Strategic Learning Revisited

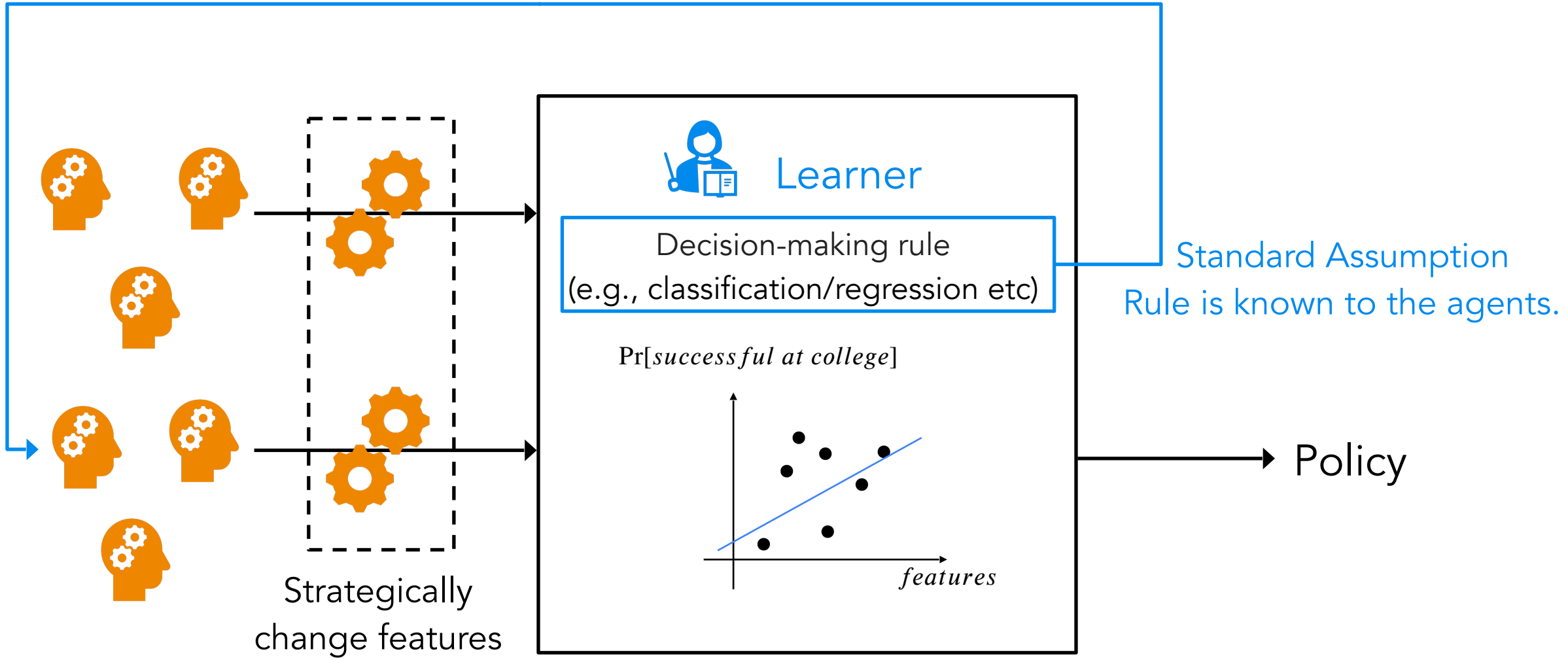




Standard Assumption
Rule is known to the agents.

[Bechavod, P., Wu, Ziani, **ICML22**]

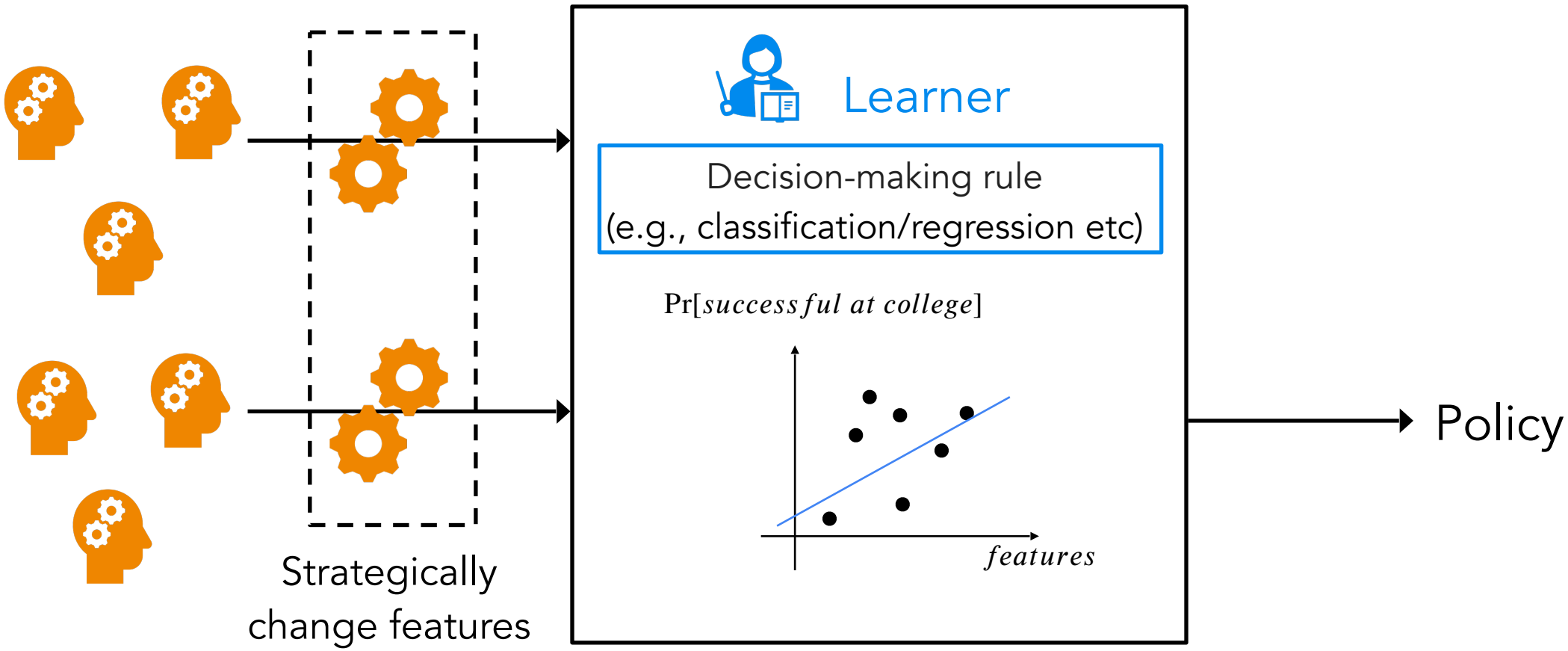
Strategic Learning Revisited



[Bechavod, P., Wu, Ziani, **ICML22**]

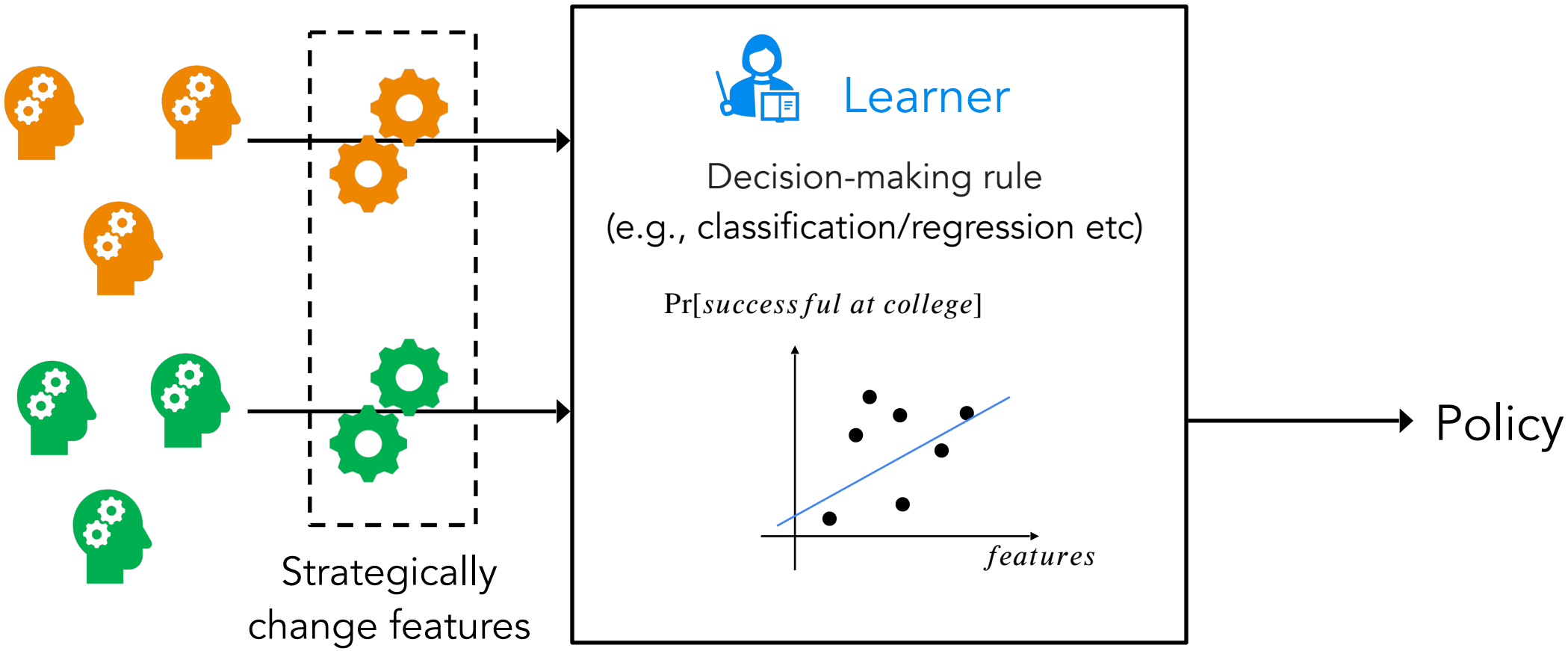
Strategic Learning Revisited

In reality: institutions **rarely reveal** their decision rules (reasons: privacy, proprietary software etc)!



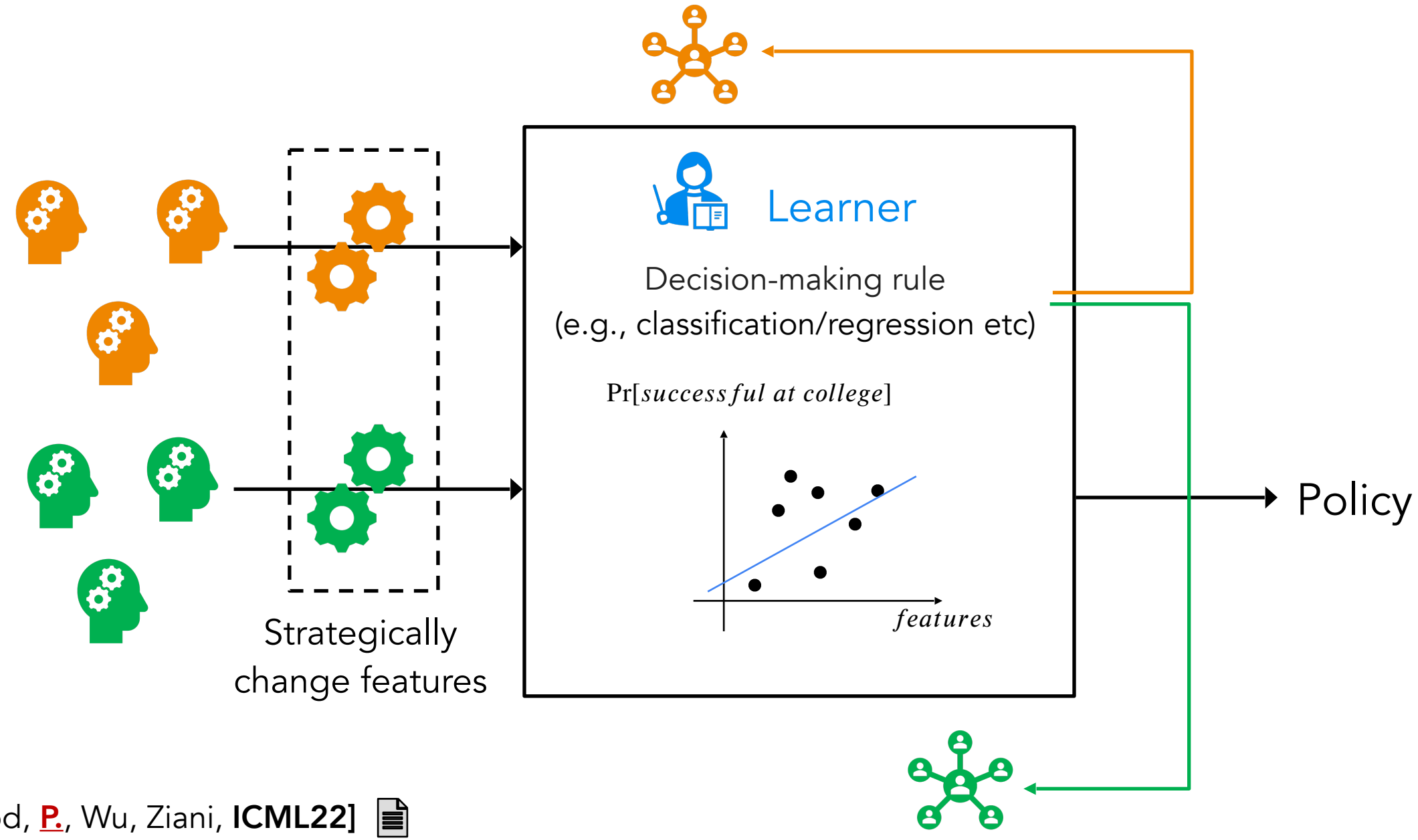
[Bechavod, P., Wu, Ziani, **ICML22**]

Strategic Learning Revisited



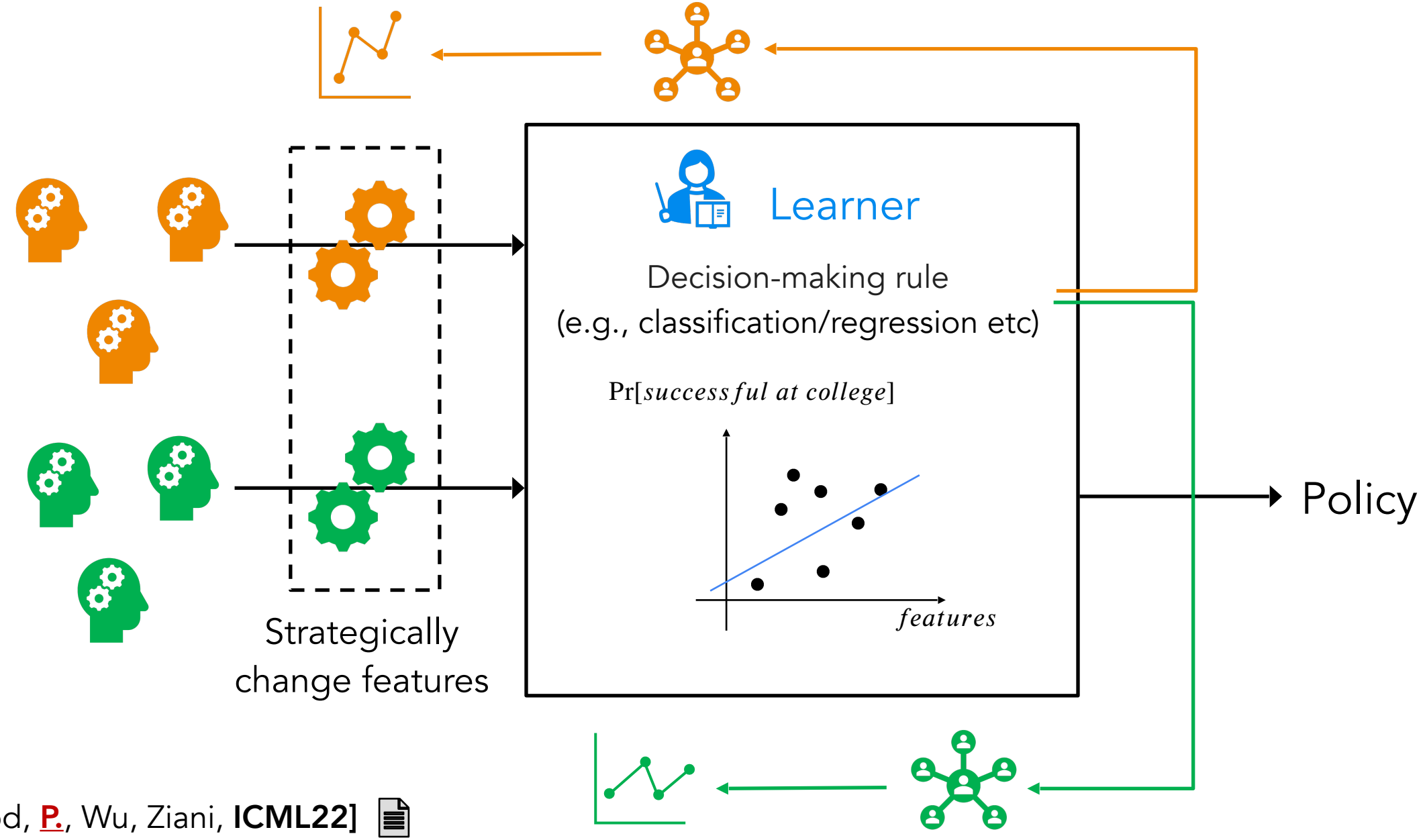
[Bechavod, P., Wu, Ziani, **ICML22**]

Strategic Learning Revisited



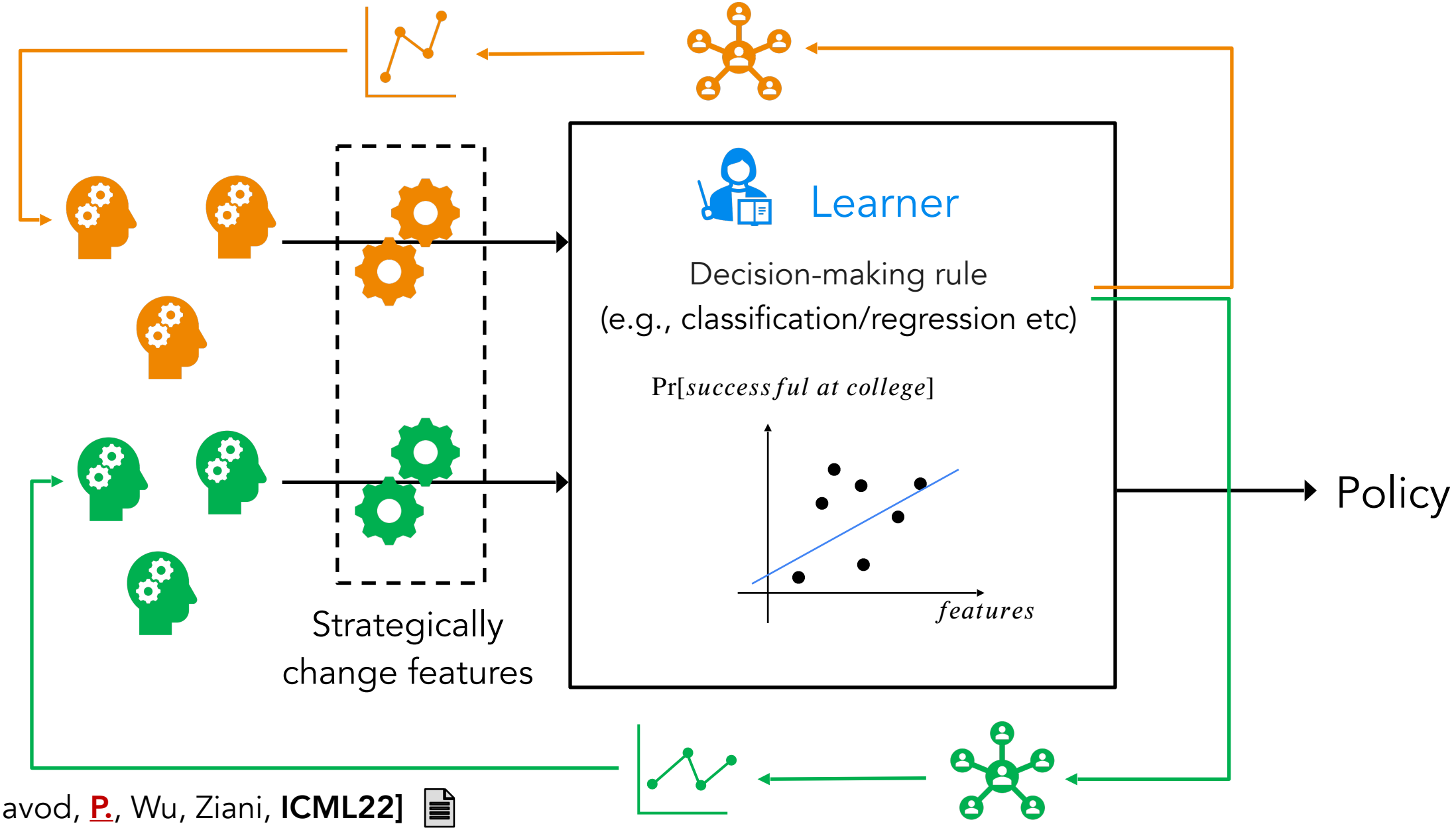
[Bechavod, P., Wu, Ziani, **ICML22**]

Strategic Learning Revisited



[Bechavod, P., Wu, Ziani, **ICML22**]

Strategic Learning Revisited



[Bechavod, P., Wu, Ziani, **ICML22**]

Question

If learner maximizes **social welfare of total population**, how does **information discrepancy** affect the subpopulations' **ability to improve**?

[Bechavod, P., Wu, Ziani, **ICML22**] 

Question

If learner maximizes **social welfare of total population**, how does **information discrepancy** affect the subpopulations' **ability to improve**?

Results

[Bechavod, P., Wu, Ziani, **ICML22**]

Question

If learner maximizes **social welfare of total population**, how does **information discrepancy** affect the subpopulations' **ability to improve**?

Results

1) In general, disadvantaged subpopulation may end up being **strictly worse off** (i.e., NO improvement).

[Bechavod, **P.**, Wu, Ziani, **ICML22]**

Question

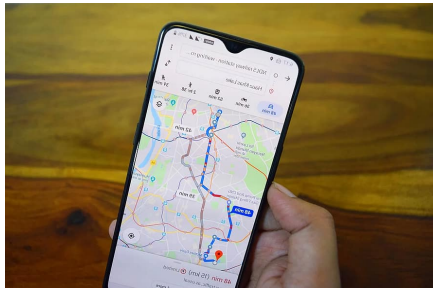
If learner maximizes **social welfare of total population**, how does **information discrepancy** affect the subpopulations' **ability to improve**?

Results

1) In general, disadvantaged subpopulation may end up being **strictly worse off** (i.e., NO improvement).
2) Subpopulation-optimal outcome **is achievable** if information for two subpopulations is independent!

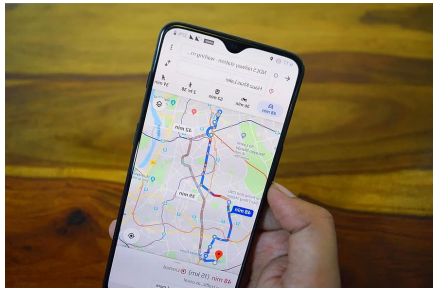
[Bechavod, **P.**, Wu, Ziani, **ICML22]** 

Performative Prediction



Performative Prediction: When **predictions** influence the **data** beyond specific utility functions

Performative Prediction



Performative Prediction: When **predictions** influence the **data** beyond specific utility functions

Main Results

1) When does repeated retraining lead to **stable** rules? [Perdomo, Zrnic, Mendler-Dünner, Hardt, **ICML20**]
2) Stoch optimization techniques to identify **stable** solutions: [Mendler-Dünner, Perdomo, Zrnic, Hardt, **NeurIPS20**]
3) Conditions for having a convex optimization problem when searching for **performatively optimal** rules. [Miller, Perdomo, Zrnic, **NeurIPS20**]
4) Regret minimization techniques that draw inspiration from zooming to learn adaptively better than standard Lipschitz bandits. [Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner, **ICML22**]

Thank You!