STOC 2022: 54th Annual ACM Symposium
on Theory of Computing
June 20-24, 2022 in Rome, Italy

SIGACT

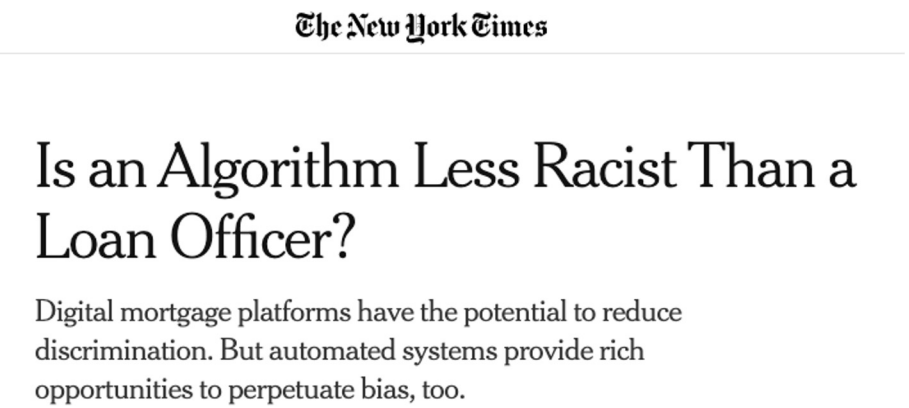
Do we incentivize honest effort or gaming in incentive-aware learning?

Chara Podimata, Harvard

STOC22 Workshop on “Optimizing the Effort of Others”, 06/20/2022

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.

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.

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.

- 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**.

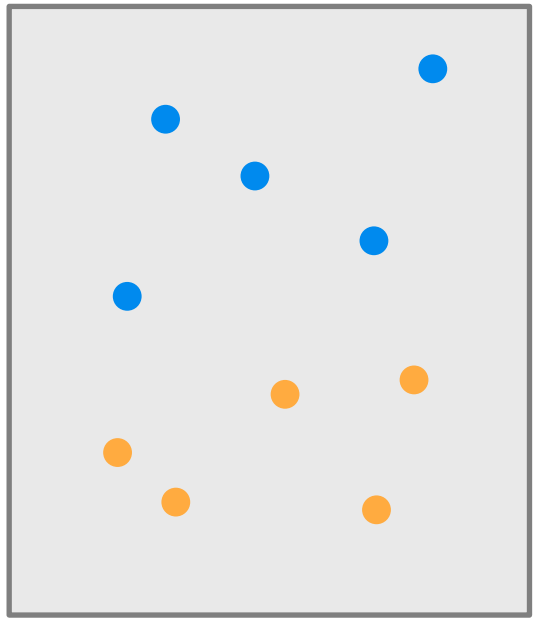
Training Data

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**.

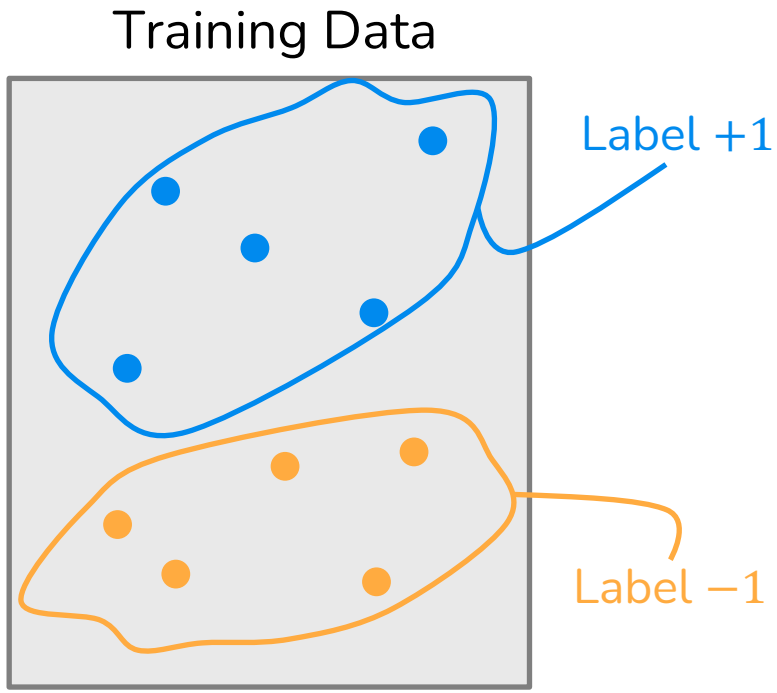
Training Data

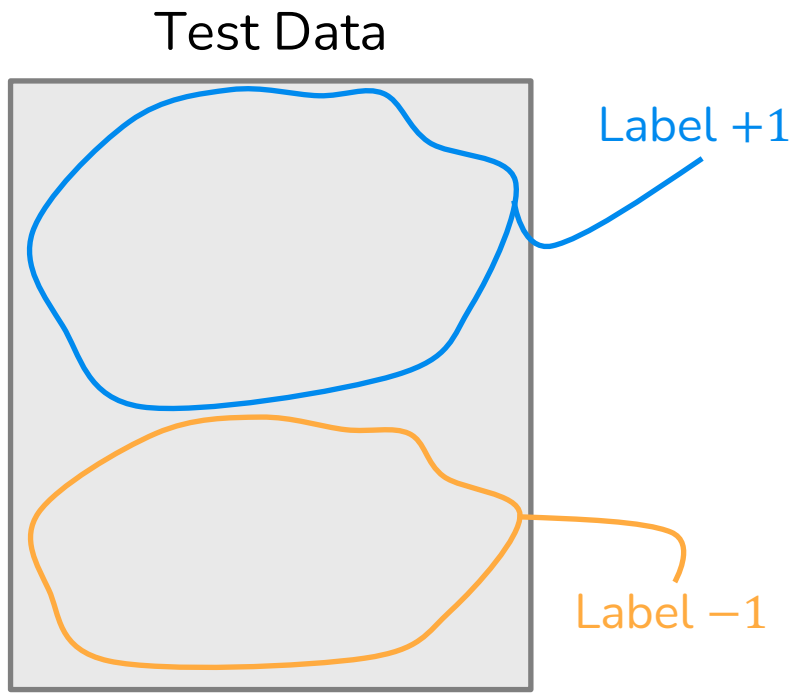


Test Data

What Can Go Wrong?

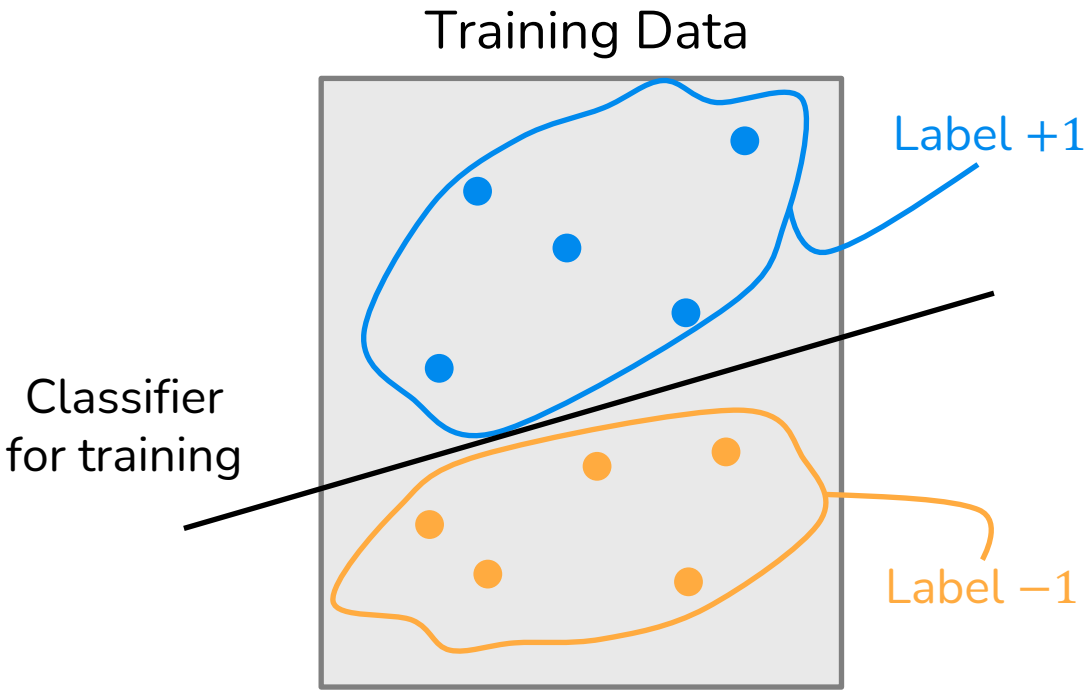
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.

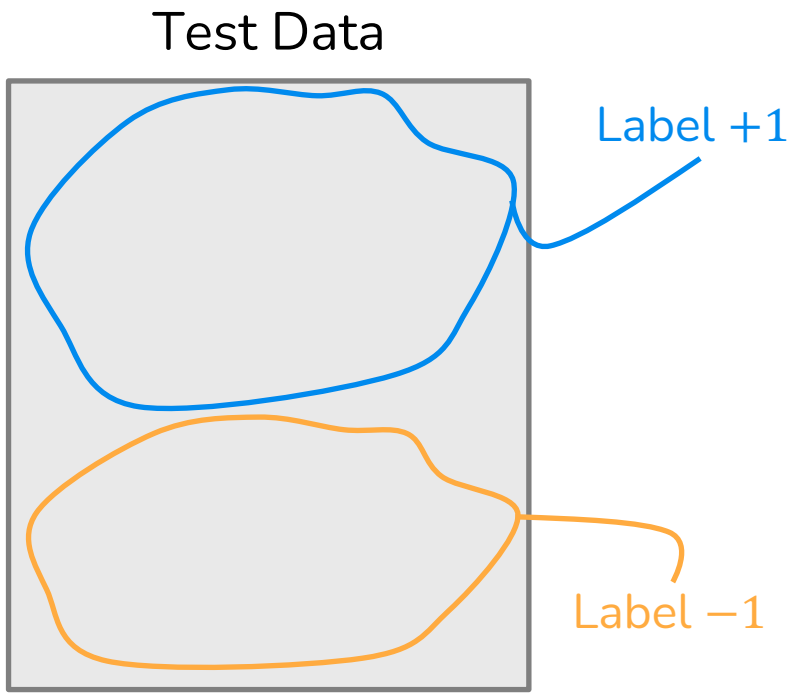




What Can Go Wrong?

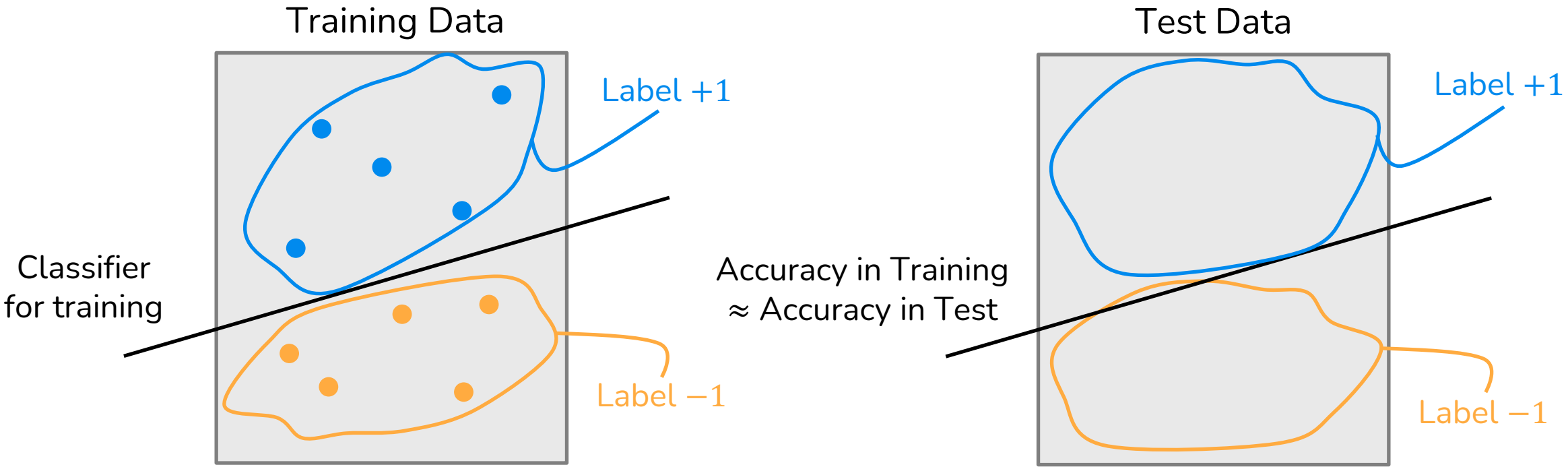
Policy makers' and mechanism designers' goal in using ML for decision-making: **learn from human data to create better decisions**.





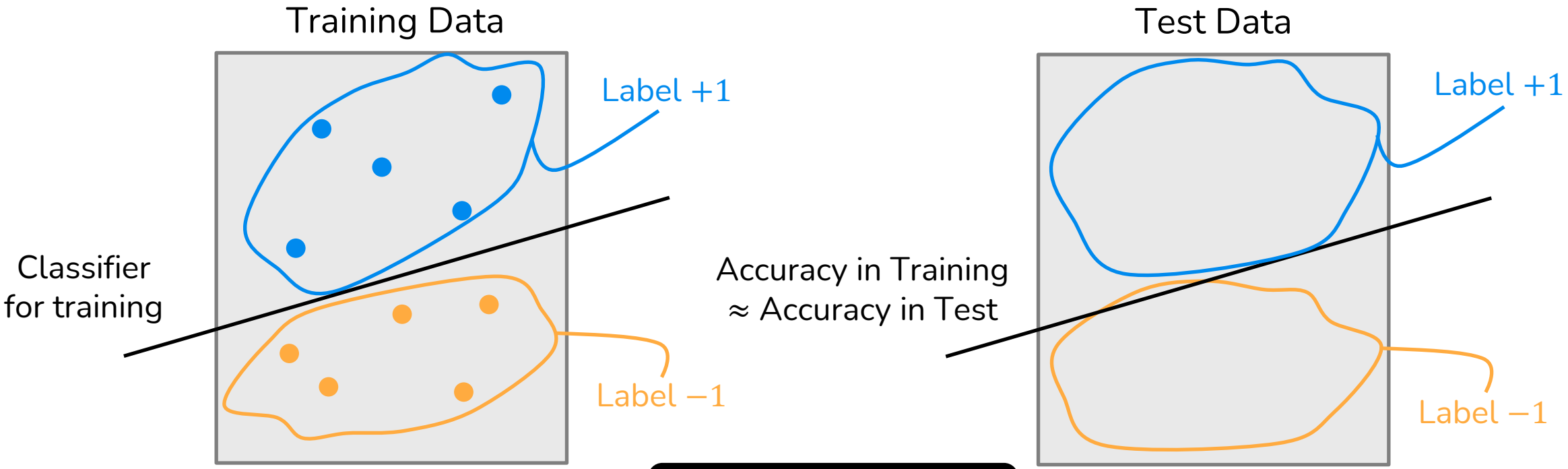
What Can Go Wrong?

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**.

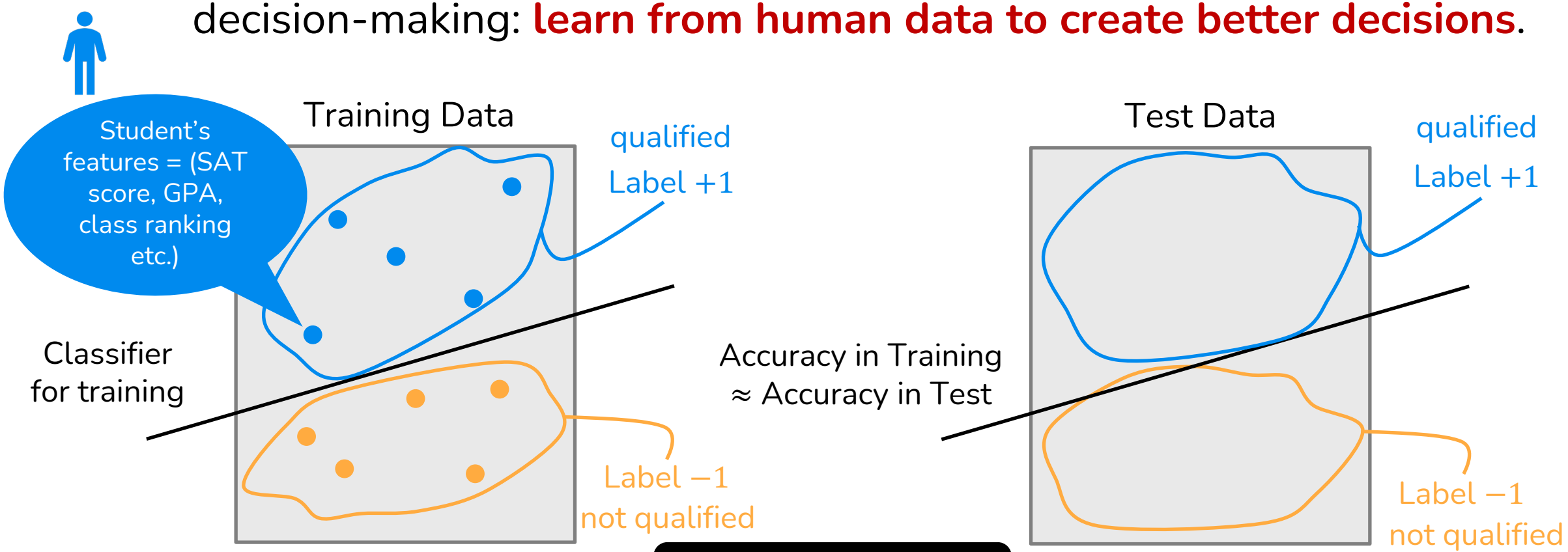


Root of Problem

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**.

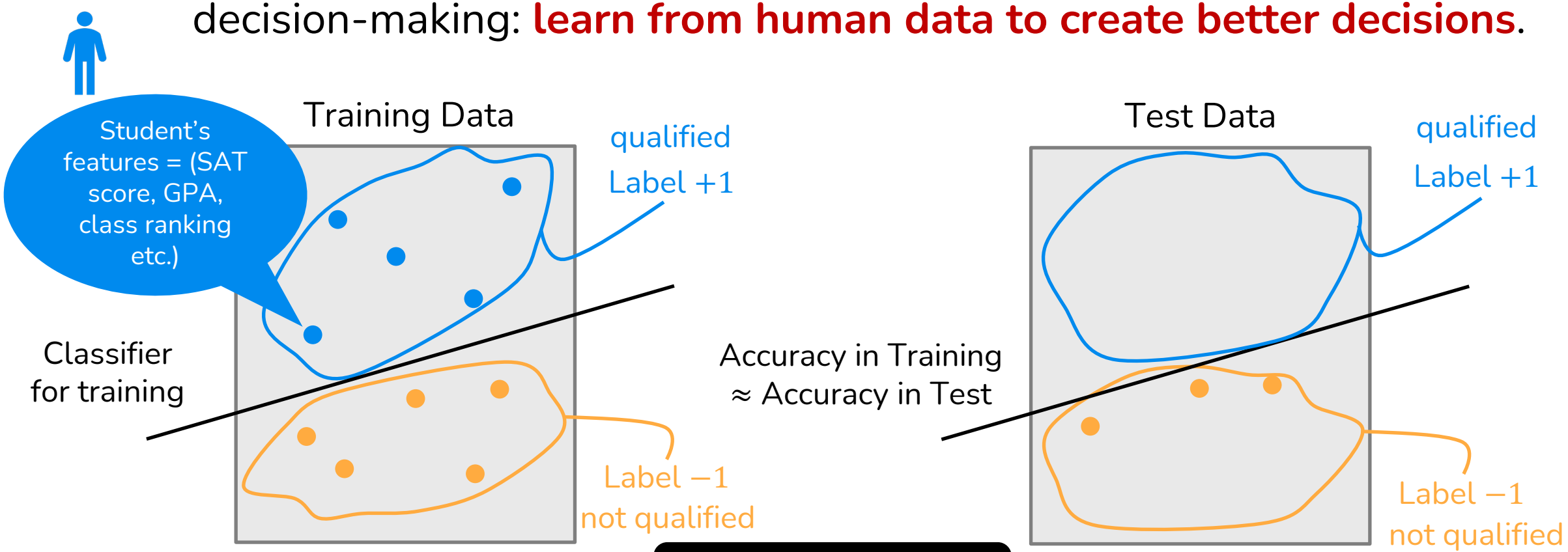


Root of Problem

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**.

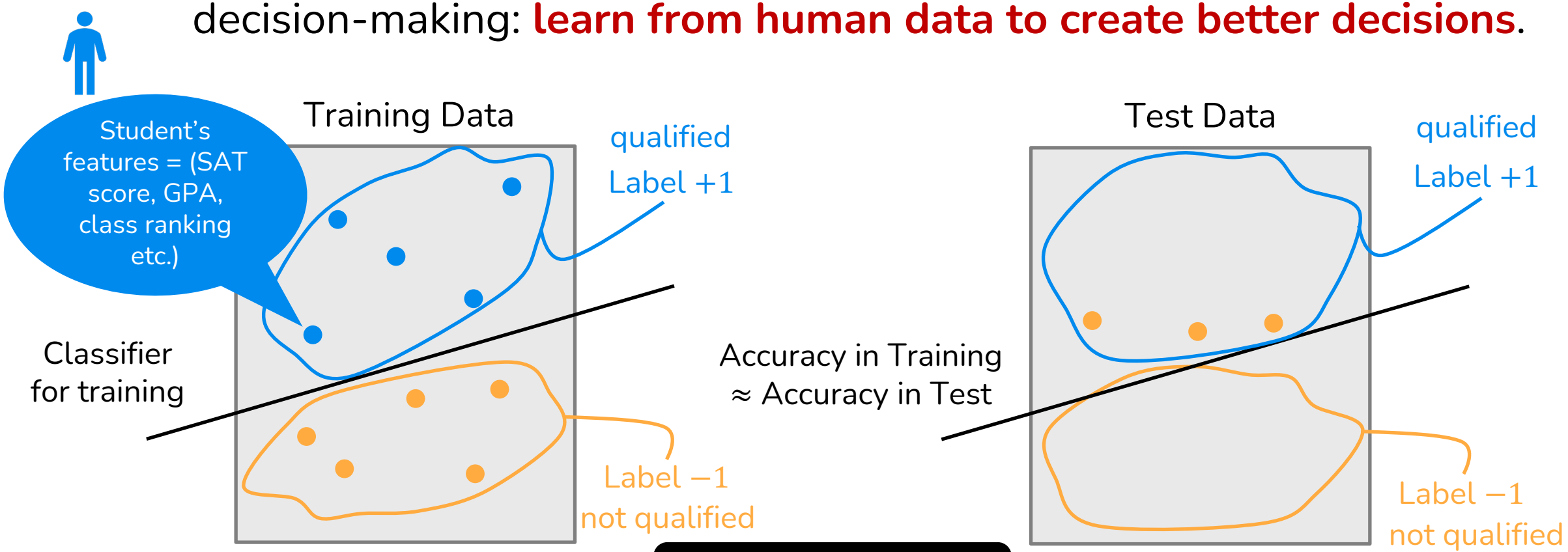


Root of Problem

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**.

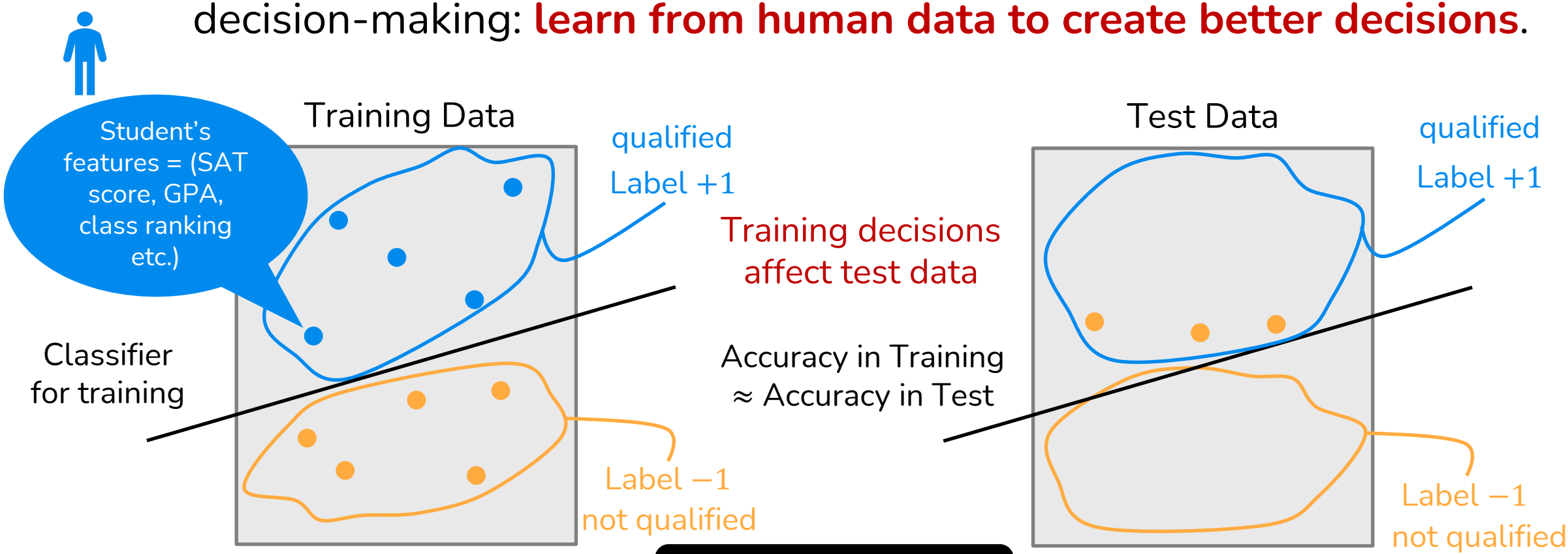


Root of Problem

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**.

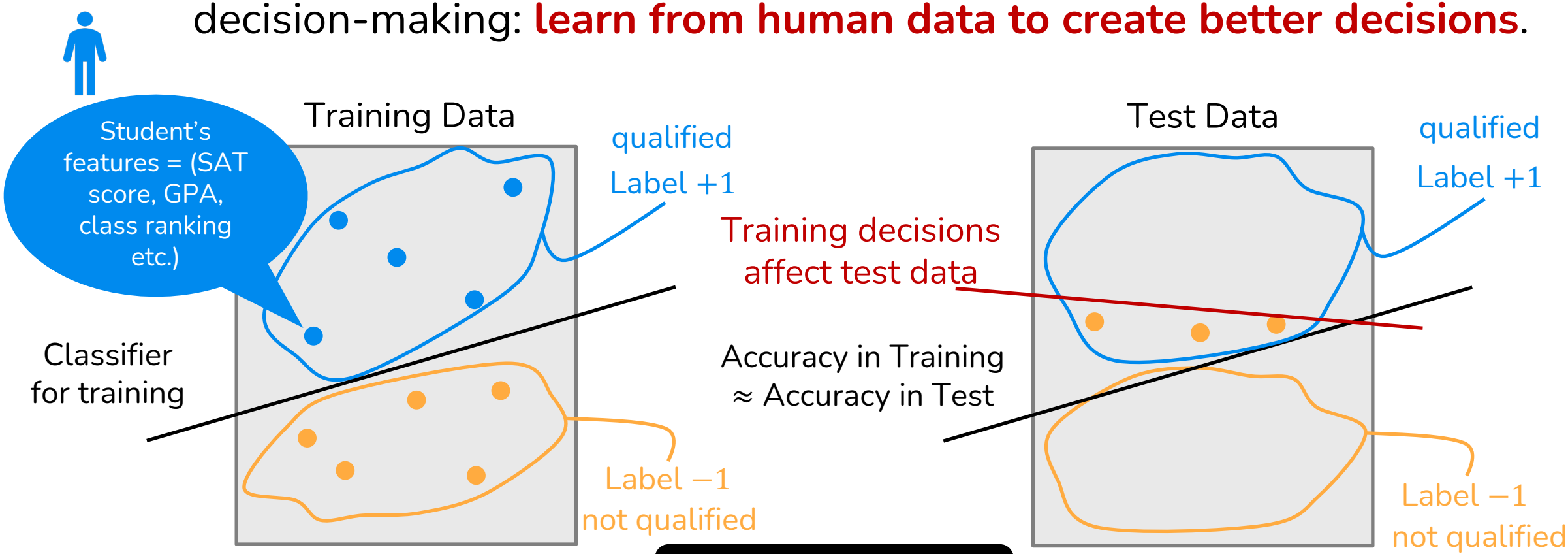


Root of Problem

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**.



Root of Problem

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

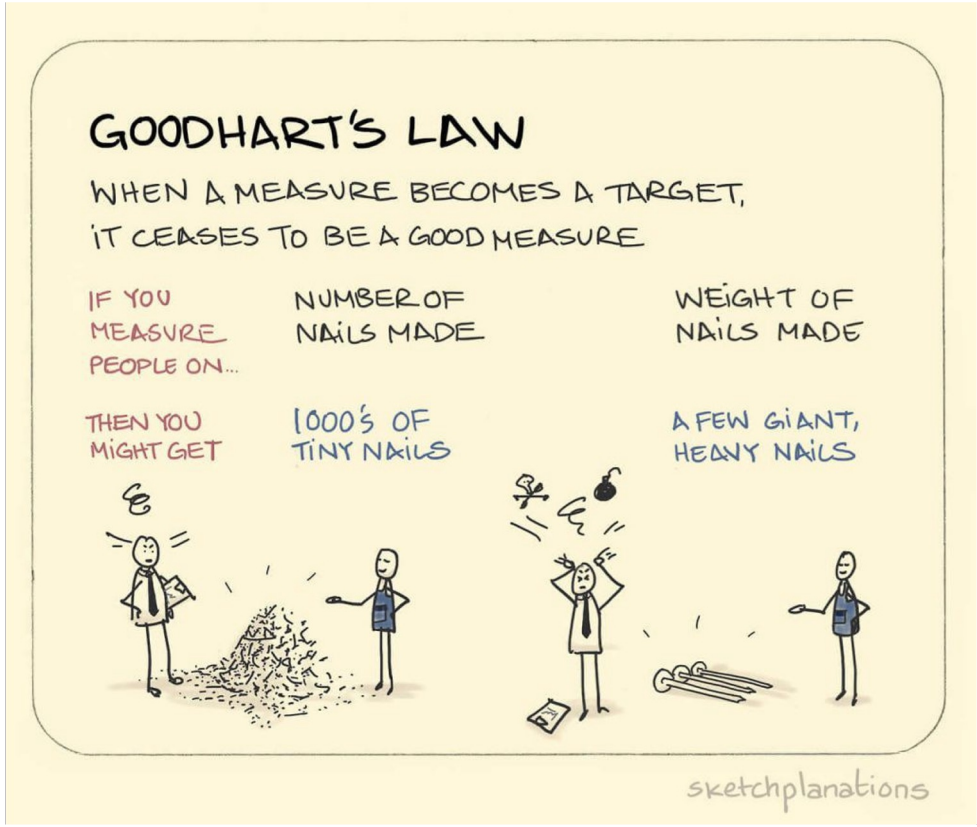
Research Agenda: Incentive-Aware ML

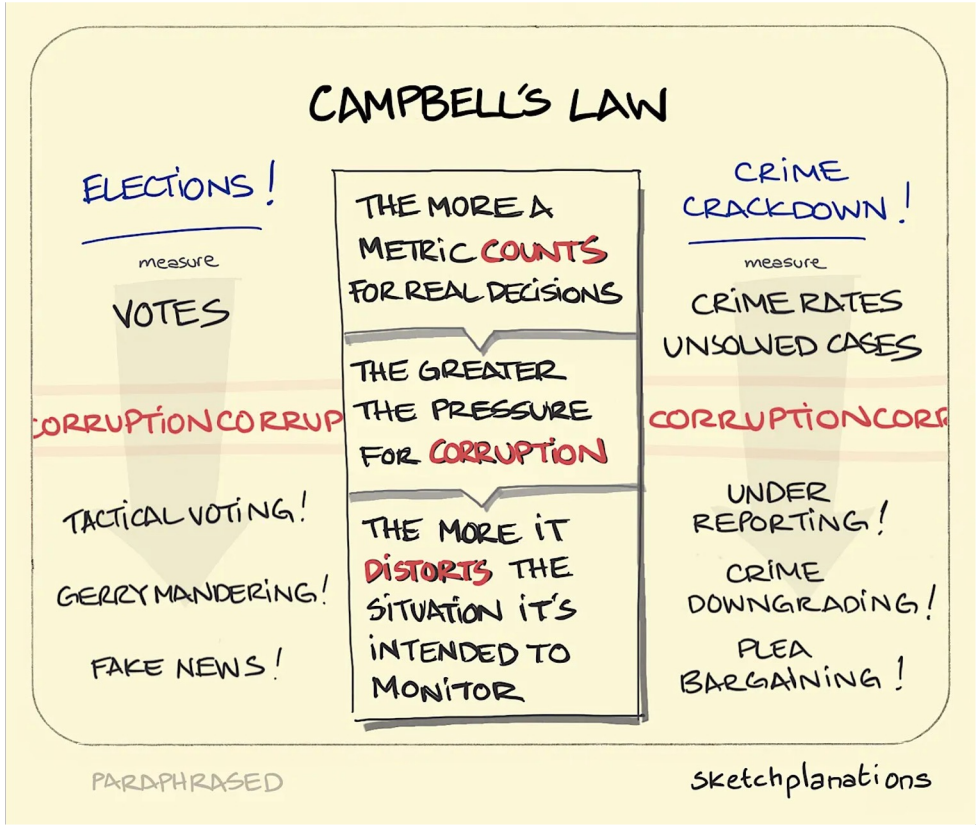
1) study the **effects of strategic behavior** to ML algorithms
2) propose **ways to robustify ML** algorithms
3) propose **ways to incentivize honest effort exertion**

Lots of Recent, Exciting Work

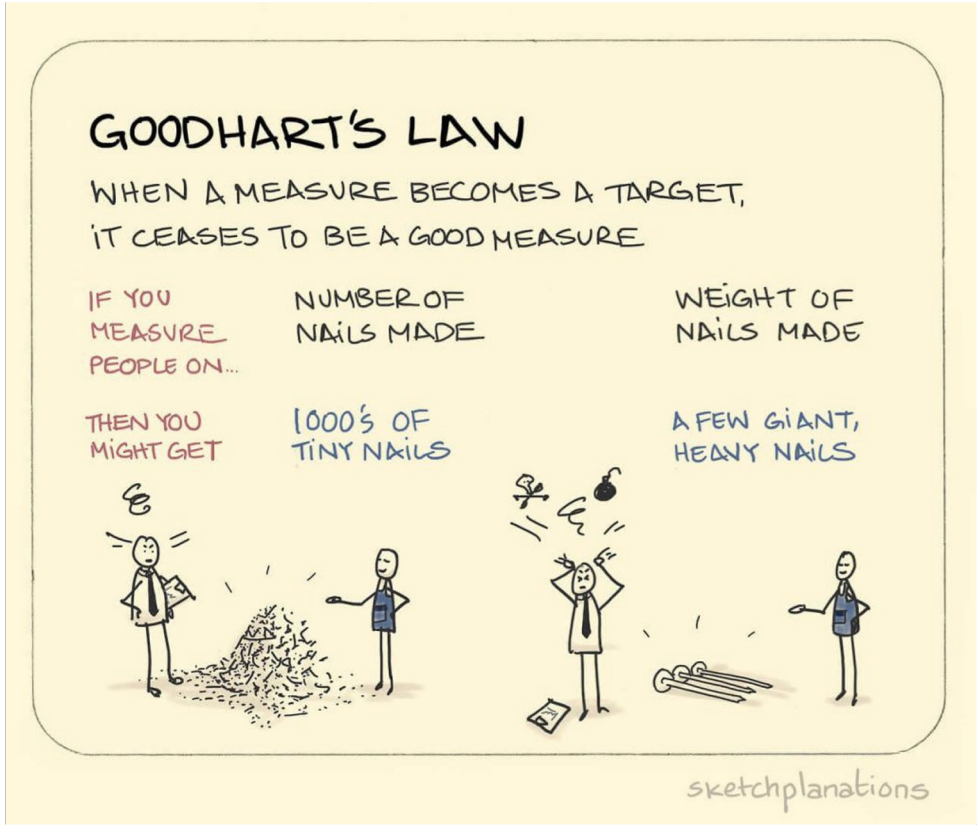
- **Robustness**: [Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**], [Dong, Roth, Schutzman, Waggoner, Wu, **EC18**], [Chen, Liu, **P.**, **NeurIPS20**], [Ahmadi, Beyhaghi, Blum, Naggita, **EC21**], [Sundaraman, Vullikanti, Xu, Yao, **ICML21**], [Ghalme , Nair, Eilat, Talgam-Cohen, Rosenfeld, **ICML21**], [Zrnic, Mazumdar, Sastry, Jordan, **NeurIPS21**], [Jagadeesan, Mendler-Dünner, Hardt, **ICML21**]
- **Fairness:** [Milli, Miller, Dragan, Hardt, **FAT*19**], [Hu, Immorlica, Vaughan, **FAT*19**], [Liu, Wilson, Haghtalab, Kalai, Borgs, Chayes, **FAT*19**], [Braverman, Garg, **FORC20**]
- **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**], [Bechavod, **P.**, Wu, Ziani, **ICML22**]
- **Causality:** [Miller, Milli, Hardt, **FAT*19**], [Shavit, Edelman, Axelrod, **ICML20**], [Bechavod, Ligett, Wu, Ziani, **AISTATS21**]
- **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**].

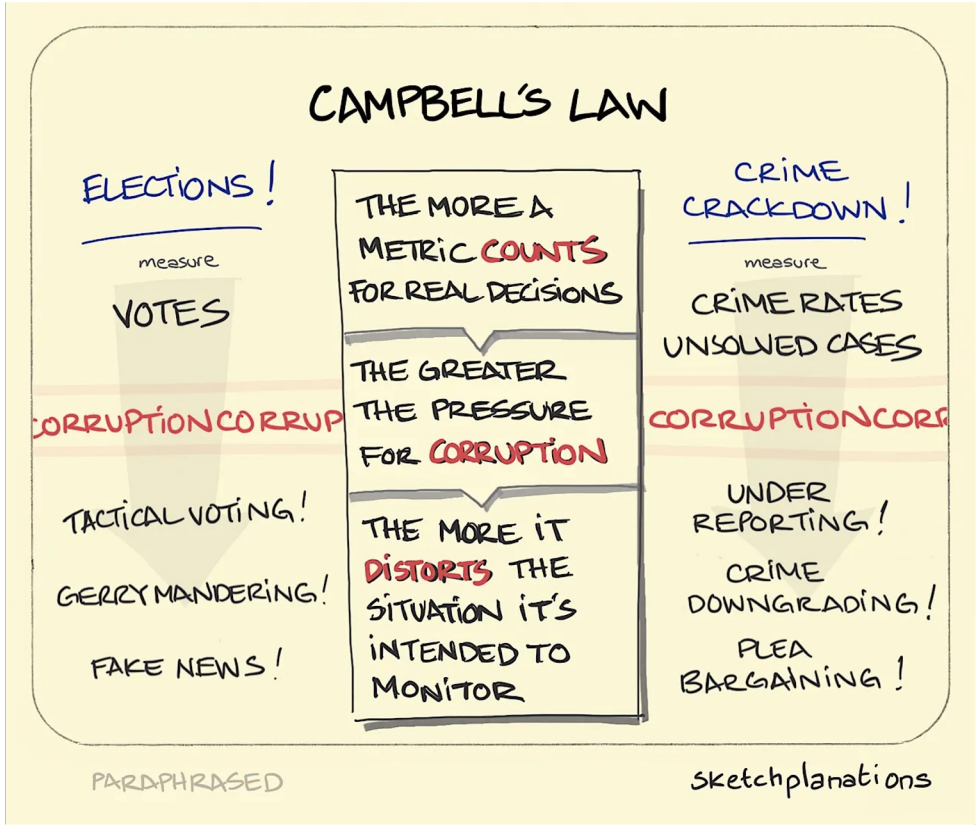
Similar Problem, Different Fields





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

What's New in Today's Research?

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

What's New in Today's Research?

- Too many predictors to scrutinize individually

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

What's New in Today's Research?

- Too many predictors to scrutinize individually
- Black-box models

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

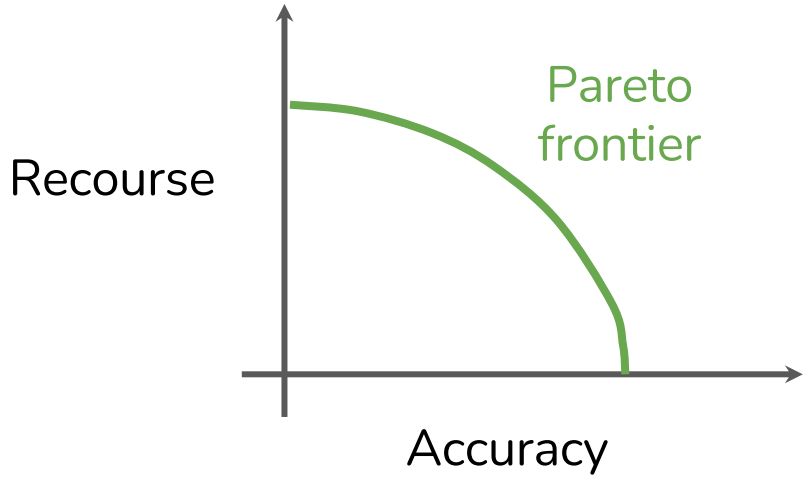
What's New in Today's Research?

- Too many predictors to scrutinize individually
- Black-box models
- Formalizing objectives in high-dimensions

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

What's New in Today's Research?

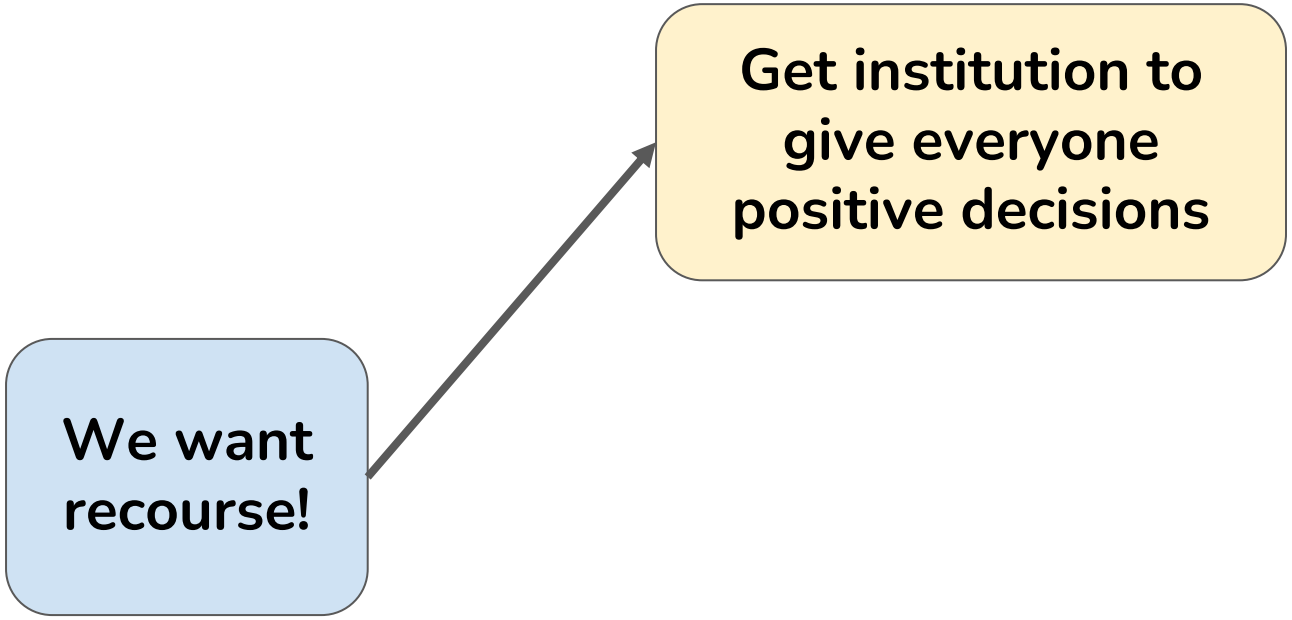
- Too many predictors to scrutinize individually
- Black-box models
- Formalizing objectives in high-dimensions
- Revealing Pareto frontier



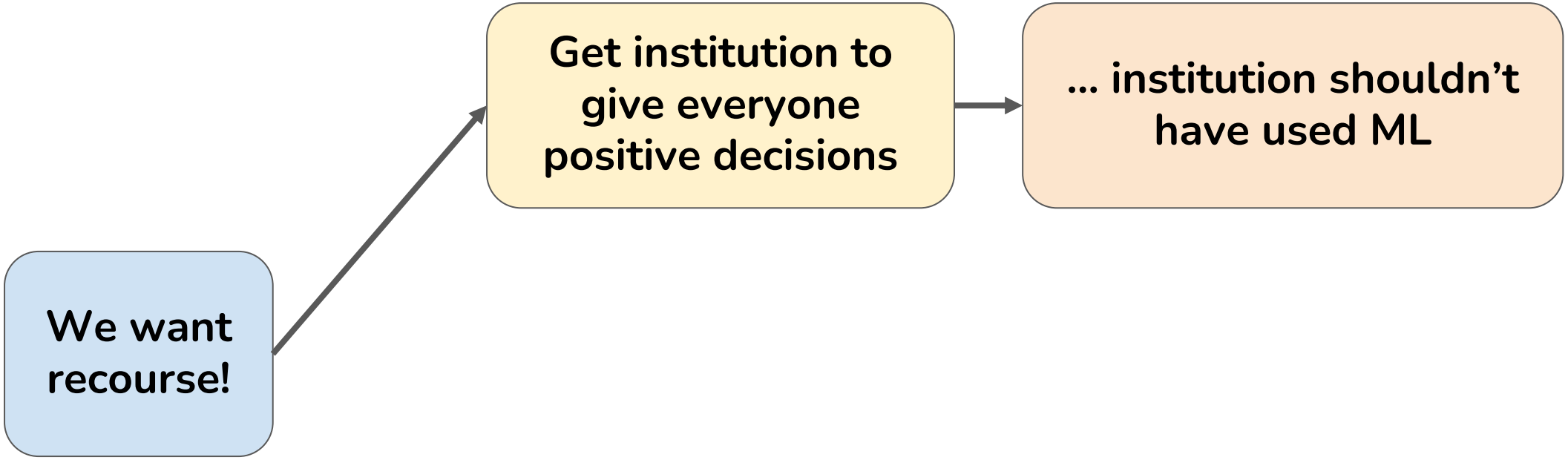
Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

We want recourse!

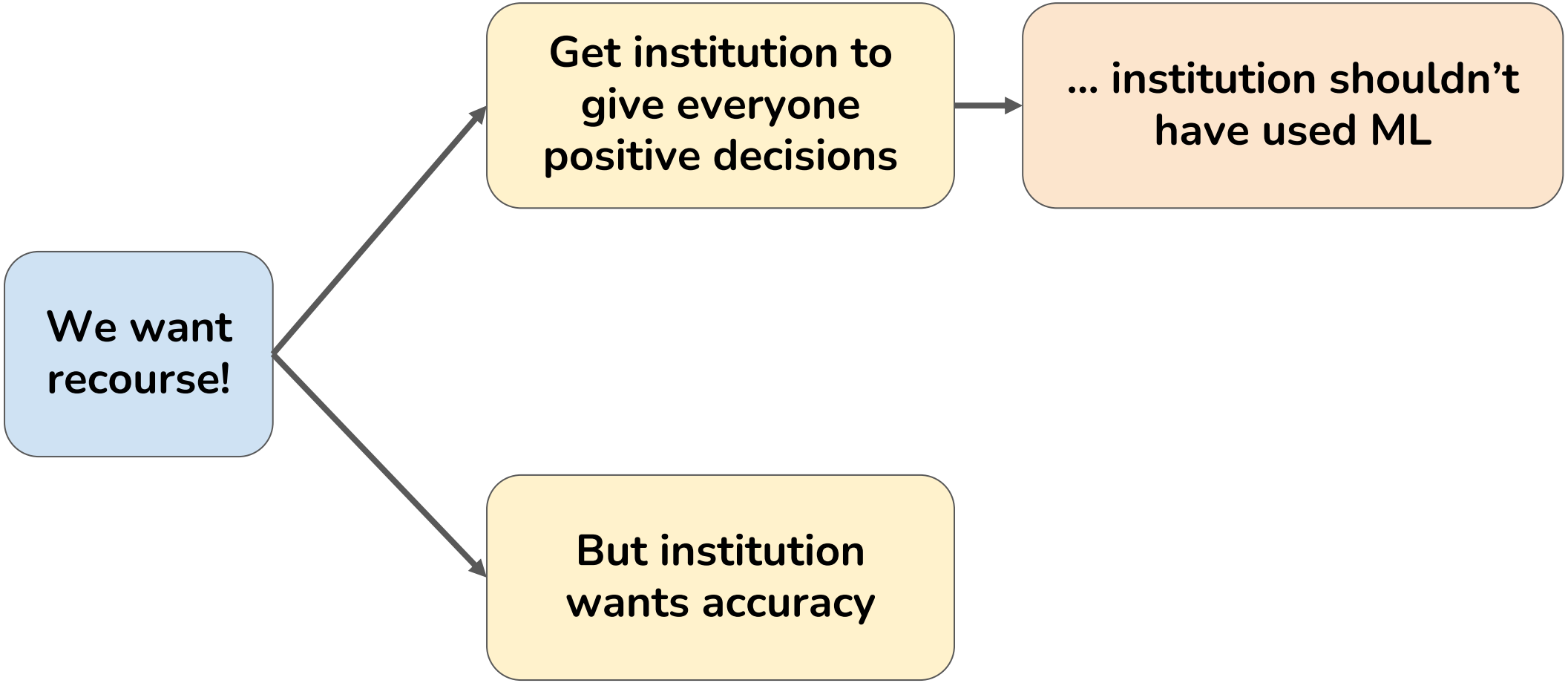
Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit



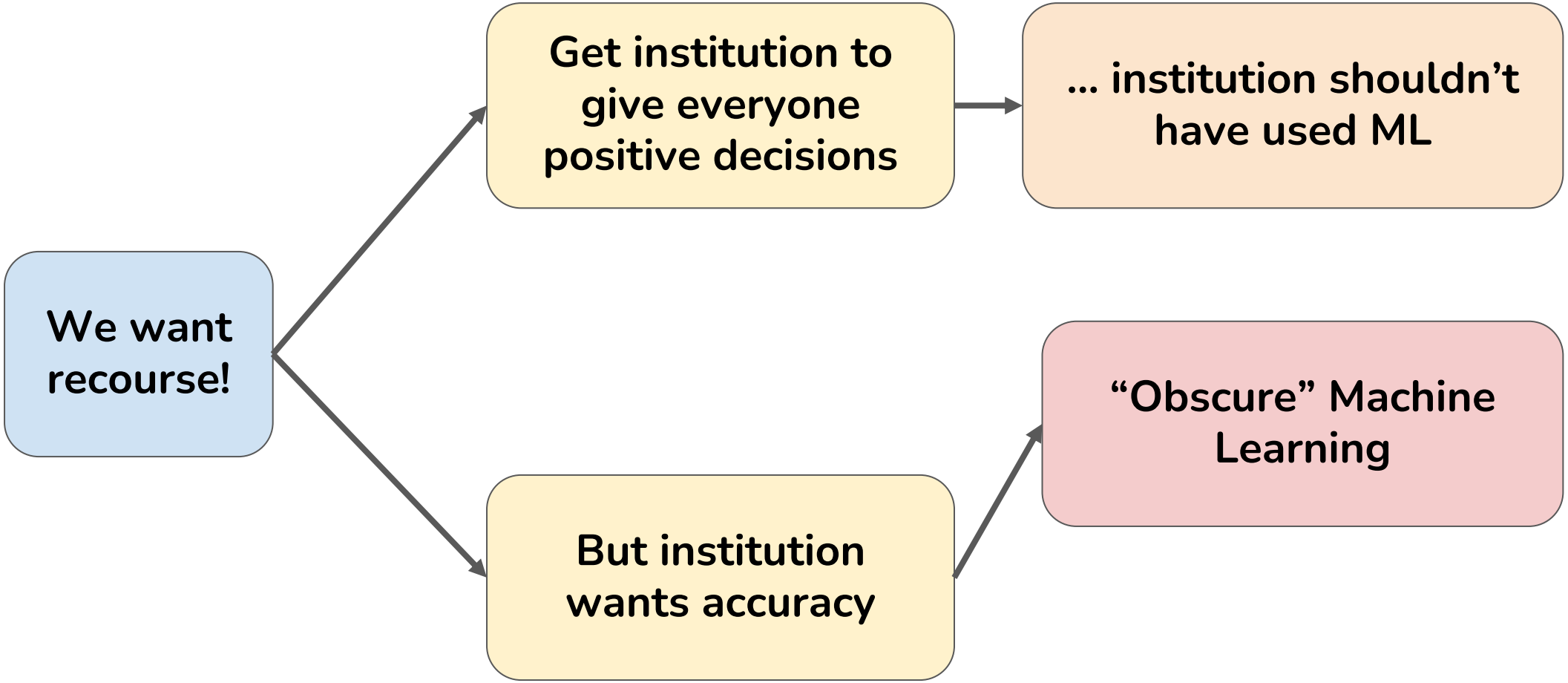
Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit



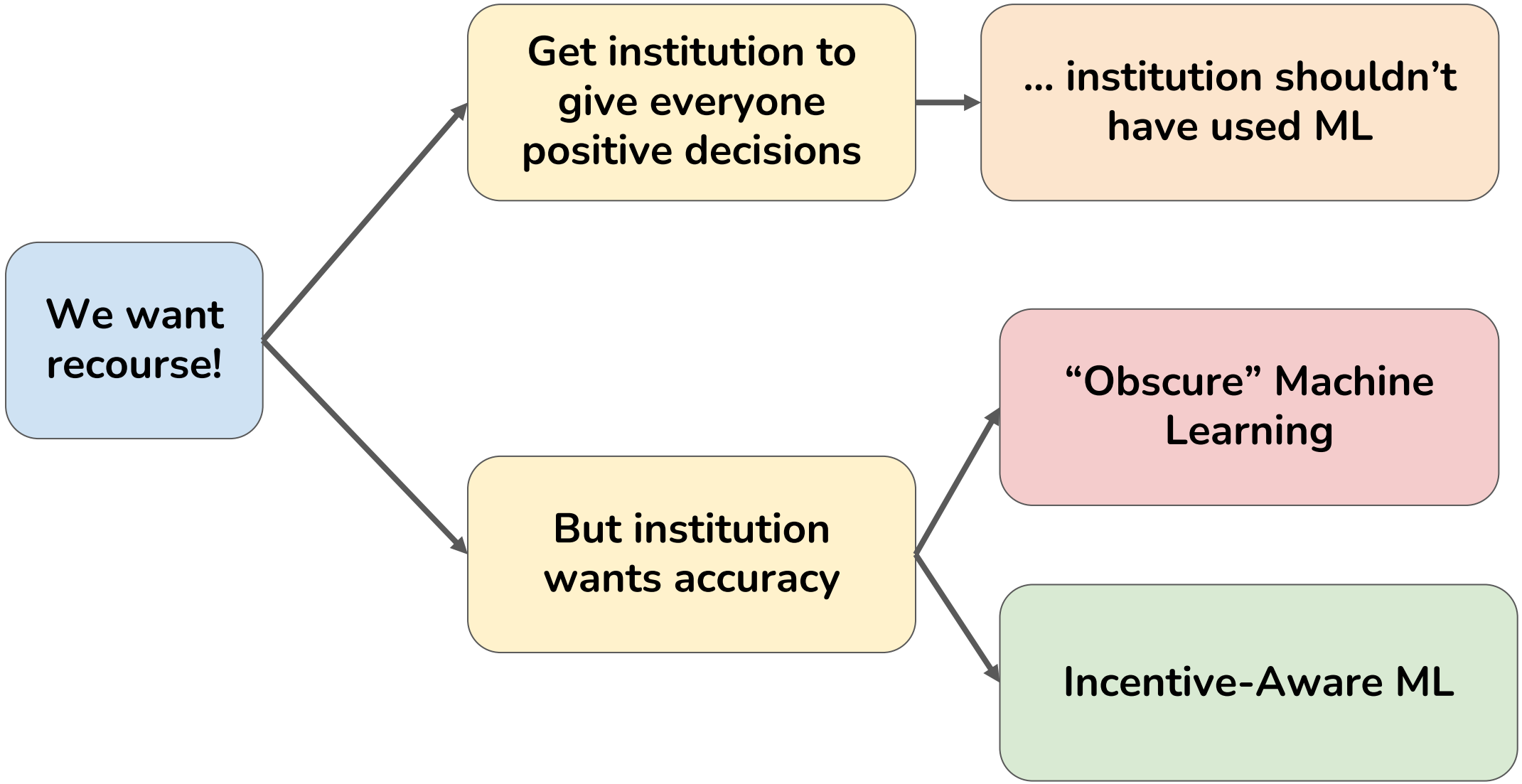
Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit



Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

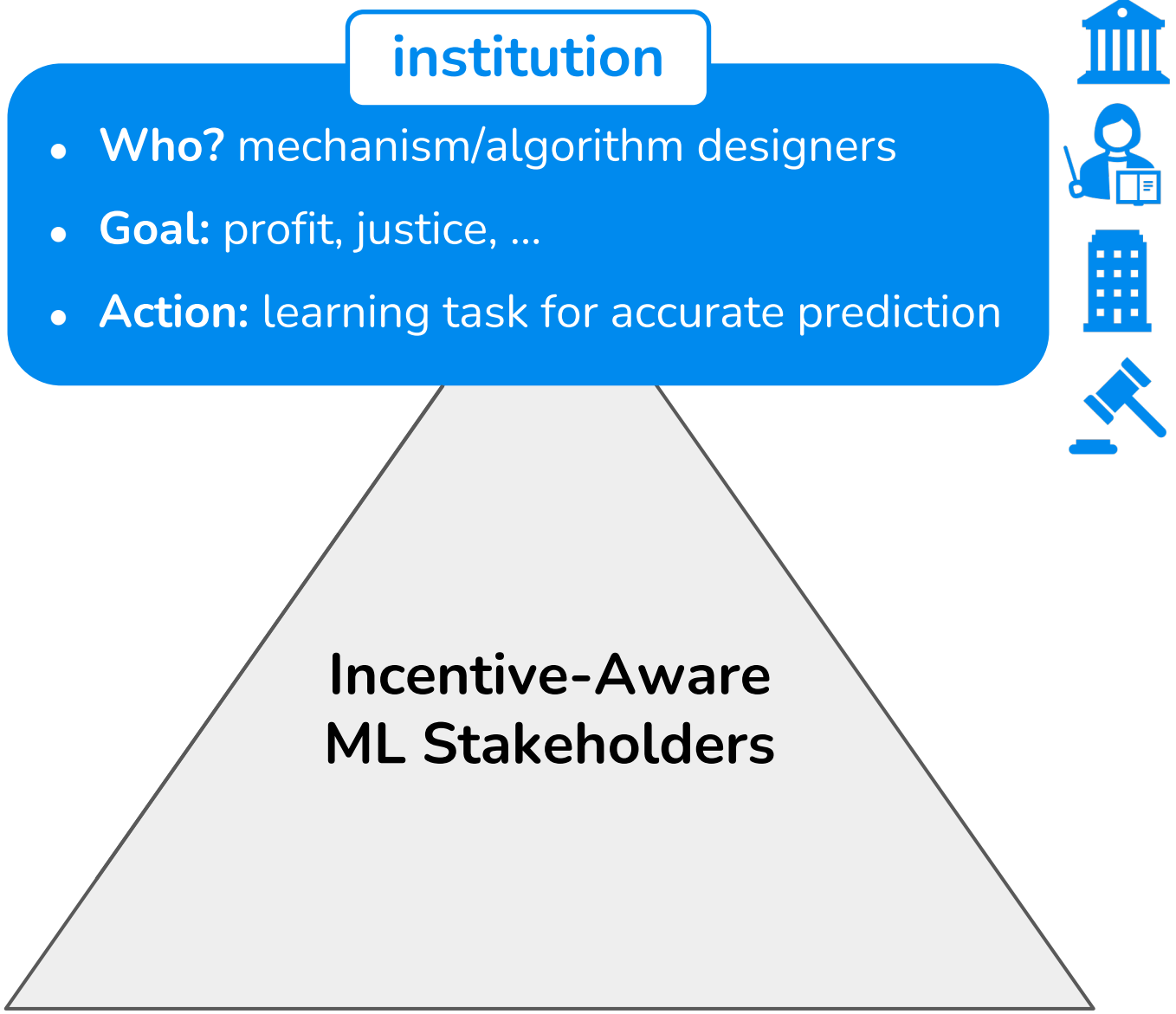


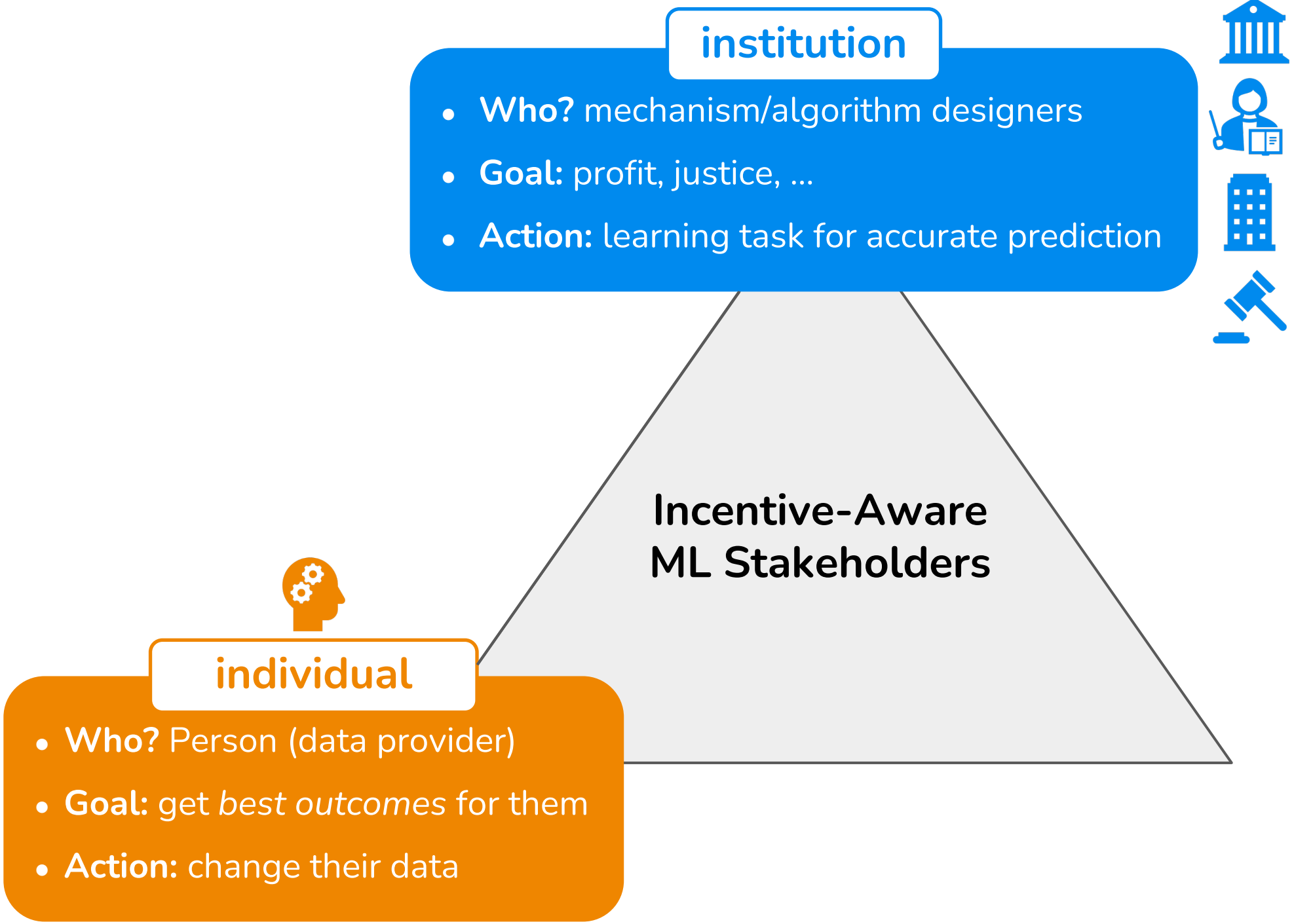
Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

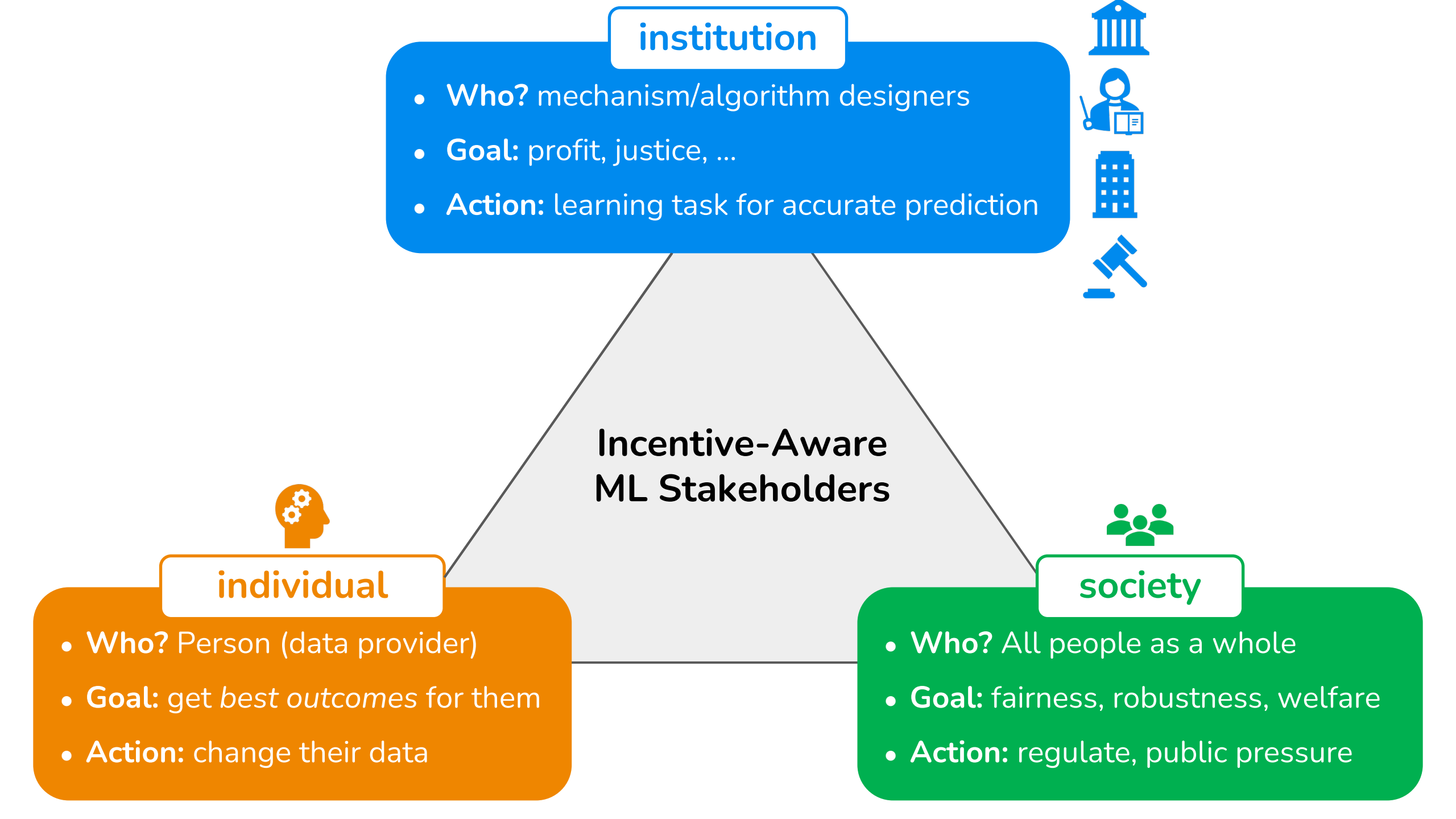


Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit


Incentive-Aware
ML Stakeholders


institution
Who? mechanism/algorithm designers
Goal: profit, justice, ...
Action: learning task for accurate prediction
Incentive-Aware
ML Stakeholders


institution
• Who? mechanism/algorithm designers
• Goal: profit, justice, ...
• Action: learning task for accurate prediction
Incentive-Aware
ML Stakeholders
individual
• Who? Person (data provider)
• Goal: get best outcomes for them
• Action: change their data


institution
Who? mechanism/algorithm designers
Goal: profit, justice, ...
Action: learning task for accurate prediction
Incentive-Aware
ML Stakeholders
individual
Who? Person (data provider)
Goal: get best outcomes for them
Action: change their data
society
Who? All people as a whole
Goal: fairness, robustness, welfare
Action: regulate, public pressure

Tutorial Outline

- ☑ Introduction
- ❑ Robustness
- ❑ Fairness
- ❑ Recourse/Performativity/Causality
- ❑ Future Directions/Open Questions

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

institution

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

1. Nature draws agent's features (e.g., SAT score, class ranking, ...) $x \in \mathcal{X}$ from distribution $\mathcal{D}$.

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

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 $\alpha \in \mathcal{A} : \mathcal{X} \to \{-1, +1\}$.

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

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.

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

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 $\Delta(x)$ $(\neq x)$.

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

Assumption: Learner knows cost function $c(x, y)$.

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

value for passing classifier

manipulation cost

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

Assumption: Learner knows cost function $c(x, y)$.

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

value for passing classifier — manipulation cost

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

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

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

Assumption: Learner knows cost function $c(x, y)$.

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

value for passing classifier — manipulation cost

$c(x, y)$: "separable", i.e.,
$c(x, y) = \max\{0, c_2(y) - c_1(x)\}$

Goal: Compute Stackelberg Equilibrium
$\alpha^\star = \arg\max_{f \in \mathcal{H}} \Pr_{x \sim \mathcal{D}}[h(x) = f(\Delta(x))]$

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

Assumption: Learner knows cost function $c(x, y)$.

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

Goal: Compute Stackelberg Equilibrium
$\alpha^\star = \arg\max_{f \in \mathcal{H}} \Pr_{x \sim \mathcal{D}}[h(x) = f(\Delta(x))]$

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

value for passing classifier — manipulation cost

$c(x, y)$: "separable", i.e.,
$c(x, y) = \max\{0, c_2(y) - c_1(x)\}$

Main Result

Algorithm that learns α^* with polynomial time and sample complexity.

Strategic Classification Offline Model

[Hardt, Megiddo, Papadimitriou, Wooters, **ITCS16**]:
Stackelberg Game origin model for strategic classification

Assumption: Learner knows cost function $c(x, y)$.

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 $h(x)$, where $h \in \mathcal{H}$ is the "ground truth" classifier.
6. Learner gets utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$.

Goal: Compute Stackelberg Equilibrium

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

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

value for passing classifier — manipulation cost

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

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

Main Result

Algorithm that learns α^* with polynomial time and sample complexity.

[Zrnic, Mazumdar, Sastry, Jordan, **NeurIPS21**]:

- order of play is determined by how fast principal-agent adapt to each other
- agent's equilibria may be favorable for both

Strategic Classification Offline Model

Moral Hazard

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

Moral Hazard

1. Agent's inherent effort level ϵ
   - If drawn from distr. → "noise" level in agent's effort

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

2. Learner's classifier $\alpha \in \mathcal{A}$

Moral Hazard

1. Agent's inherent effort level ϵ
   - If drawn from distr. → "noise" level in agent's effort

2. Principal's contract $w(q) = a + \beta q$

outcome ↑ base salary ↑ bonus rate ↑

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

2. Learner's classifier $\alpha \in \mathcal{A}$

3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

Moral Hazard

1. Agent's inherent effort level ϵ
   - If drawn from distr. → "noise" level in agent's effort
2. Principal's contract $w(q) = a + \beta q$

3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

2. Learner's classifier $\alpha \in \mathcal{A}$

3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.

Moral Hazard

1. Agent's inherent effort level ϵ
   - If drawn from distr. → "noise" level in agent's effort

2. Principal's contract $w(q) = a + \beta q$

3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

2. Learner's classifier $\alpha \in \mathcal{A}$

3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.

5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$
(or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level ϵ
- If drawn from distr. → "noise" level in agent's effort

2. Principal's contract $w(q) = a + \beta q$

3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.

5. Principal's utility: $q - w(q)$.

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

2. Learner's classifier $\alpha \in \mathcal{A}$

3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.

5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$
(or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level ϵ
- If drawn from distr. → "noise" level in agent's effort

2. Principal's contract $w(q) = a + \beta q$

3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.

5. Principal's utility: $q - w(q)$.

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$

2. Learner's classifier $\alpha \in \mathcal{A}$

3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.

5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$
(or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level ϵ
- If drawn from distr. → "noise" level in agent's effort

2. Principal's contract $w(q) = a + \beta q$

3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome
$q = \Delta + \epsilon$.

5. Principal's utility: $q - w(q)$.

Optimality/necessity of linearity in ML:
[Kleinberg & Raghavan, **EC19**]

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$
2. Learner's classifier $\alpha \in \mathcal{A}$
3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.
5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$ (or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level
2. Principal's contract $w(q) = a + \beta q$
3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.
5. Principal's utility: $q - w(q)$.

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$
2. Learner's classifier $\alpha \in \mathcal{A}$
3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.
5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$ (or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level
2. Principal's contract $w(q) = a + \beta q$
3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.
5. Principal's utility: $q - w(q)$.

Differences

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$
2. Learner's classifier $\alpha \in \mathcal{A}$
3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.
5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$ (or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level
2. Principal's contract $w(q) = a + \beta q$
3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.
5. Principal's utility: $q - w(q)$.

Differences

1. Action vs outcome observed.

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$
2. Learner's classifier $\alpha \in \mathcal{A}$
3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.
5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$ (or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level
2. Principal's contract $w(q) = a + \beta q$
3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.
5. Principal's utility: $q - w(q)$.

Differences

1. Action vs outcome observed.
2. Single action versus continuum of actions (i.e., feature vector). Maybe **Combinatorial Contracts** can help here?

Strategic Classification Offline Model

1. Agent's original features $x \in \mathcal{X}$
2. Learner's classifier $\alpha \in \mathcal{A}$
3. Agent chooses features $\Delta(x)$ as:

$$\Delta(x) = \arg\max_{y \in \mathcal{X}} \mathbb{E}_x[\alpha(y) - c(x, y)]$$

4. Learner observes $\Delta(x)$.
5. Learner's utility: $\Pr_{x \sim \mathcal{D}}[h(x) = \alpha(\Delta(x))]$ (or some other loss func, see online model)

Moral Hazard

1. Agent's inherent effort level
2. Principal's contract $w(q) = a + \beta q$
3. Agent chooses action Δ as:

$$\Delta = \arg\max_{y \in \mathcal{X}} \mathbb{E}\left[-e^{-r \cdot (w(q) - c(y))}\right]$$

4. Principal observes outcome $q = \Delta + \epsilon$.
5. Principal's utility: $q - w(q)$.

Differences

1. Action vs outcome observed.
2. Single action versus continuum of actions (i.e., feature vector). Maybe **Combinatorial Contracts** can help here?
3. "Outcomes" observed only in a "censored" way in Strategic Classification.

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **NeurIPS20**]
[Ahmadi, Beyhaghi, Blum, Naggita, **EC21**]

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

Strategic Classification Online Model

For round $t \in [T]$:

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **NeurIPS20**]
[Ahmadi, Beyhaghi, Blum, Naggita, **EC21**]

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **NeurIPS20**]
[Ahmadi, Beyhaghi, Blum, Naggita, **EC21**]

For round $t \in [T]$:

1. Nature chooses agent's features (e.g., SAT score, class ranking, ...) $x_t \in \mathcal{X} \subseteq [0,1]^d$.

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **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 $\alpha_t \in \mathcal{A} \subseteq [-1,1]^{d+1}$.

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **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 $y_t \in \{-1, 1\}$.

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
5. Learner observes label y_t.

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
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

individual

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

Strategic Classification Online Model

[Dong, Roth, Schutzman, Waggoner, Wu, **EC18**]
[Chen, Liu., **P.**, **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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
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

$$\mathcal{R}(T) = \sum_{t=1}^{T} \ell\left(\alpha_t, \hat{x}_t(\alpha_t)\right) - \min_{\alpha^\star \in \mathcal{A}} \sum_{t=1}^{T} \ell\left(\alpha^\star, \hat{x}_t(\alpha^\star)\right)$$

Strategic Classification Online Model

[Hardt, Roth, Schuttman, Wagoner, Wu, EC18]

[Chen, Liu., **P.**, **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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs binary classification loss:

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

institution

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

Goal: Minimize Stackelberg Regret

$$\mathcal{R}(T) = \sum_{t=1}^{T} \ell\big(\alpha_t, \hat{x}_t(\alpha_t)\big) - \min_{\alpha^\star \in \mathcal{A}} \sum_{t=1}^{T} \ell\big(\alpha^\star, \hat{x}_t(\alpha^\star)\big)$$

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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs logistic/hinge classification loss:

$\ell(\alpha_t, \hat{x}_t(\alpha_t)) = \log(1 + e^{y_t \cdot \langle \hat{x}_t, \alpha_t \rangle})$ **or** $\ell(\alpha_t, \hat{x}_t(\alpha_t))$ $= \max(0, 1 - y_t \cdot \langle \hat{x}_t, \alpha_t \rangle)$.

institution

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

Goal: Minimize Stackelberg Regret

$$\mathcal{R}(T) = \sum_{t=1}^{T} \ell\big(\alpha_t, \hat{x}_t(\alpha_t)\big) - \min_{\alpha^\star \in \mathcal{A}} \sum_{t=1}^{T} \ell\big(\alpha^\star, \hat{x}_t(\alpha^\star)\big)$$

Strategic Classification Online Model

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

[Chen, Liu, P., 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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs logistic/hinge classification loss:

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

$= \max(0, 1 - y_t \cdot \langle \hat{x}_t, \alpha_t \rangle)$.

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

individual

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

Strategic Classification Online Model

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

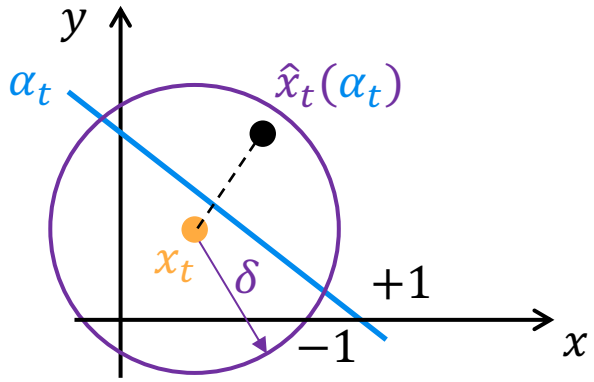
For round $t \in [T]$:

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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs binary classification loss:

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

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

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

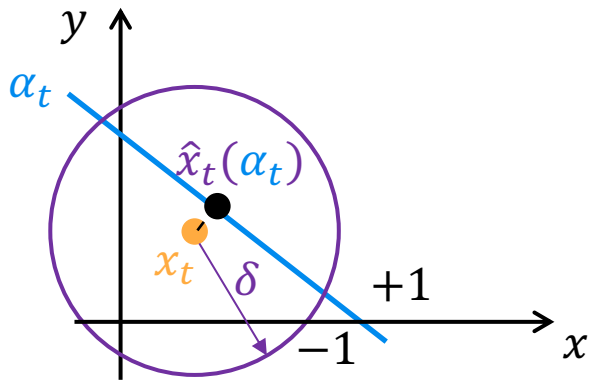
For round $t \in [T]$:

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 $\hat{x}_t(\alpha_t)$ $(\neq x_t)$.
5. Learner observes label y_t.
6. Learner incurs binary classification loss:

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

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 Question

How does **the learner learn** to classify strategic agents with **diminishing regret**?

Main Results

Main Results

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

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

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

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

Main Results

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

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

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

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

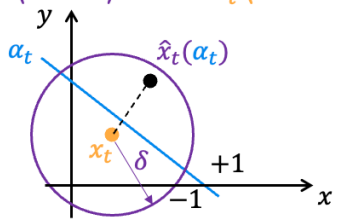
[Chen, Liu, **P.**, **NeurIPS20]**

$$(\textbf{nearly tight})\ Regret(T) = O\left(\sqrt{T \cdot \log^2(T \cdot F(\delta))}\right),$$

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).

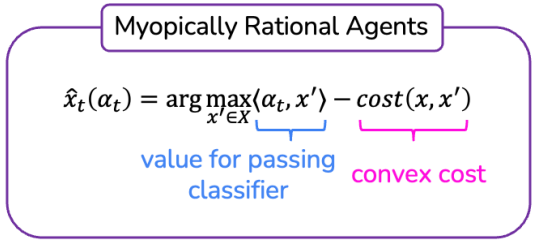


Main Results

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

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

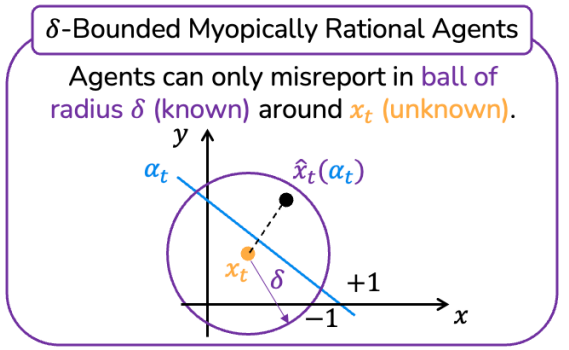
→ bandit convex opt → $Regret = O(\sqrt{d}T^{3/4})$



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

$(\boldsymbol{nearly\ tight})\ Regret(T) = O\left(\sqrt{T \cdot \log^2(T \cdot F(\delta))}\right),$

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

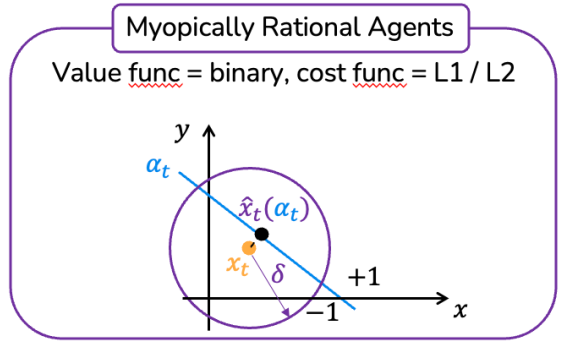


[Ahmadi, Beyhaghi, Blum, Naggita, **EC21]**

If data are linearly separable with margin γ:

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

L1: $Regret(T) = O\left(\frac{d \cdot (1+ManipulationPower)^2}{\gamma^2}\right)$

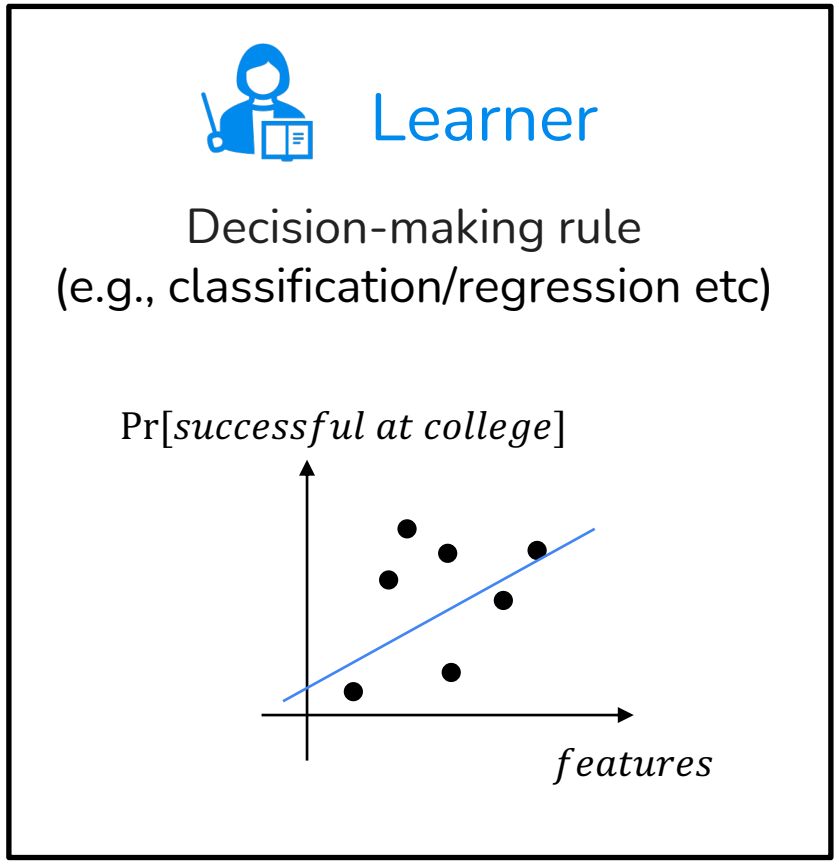


Tutorial Outline

- ☑ Introduction
- ☑ Robustness
- ❑ Fairness
- ❑ Recourse/Performativity/Causality
- ❑ Future Directions/Open Questions

Implicit Assumption so Far: Homogeneous Population

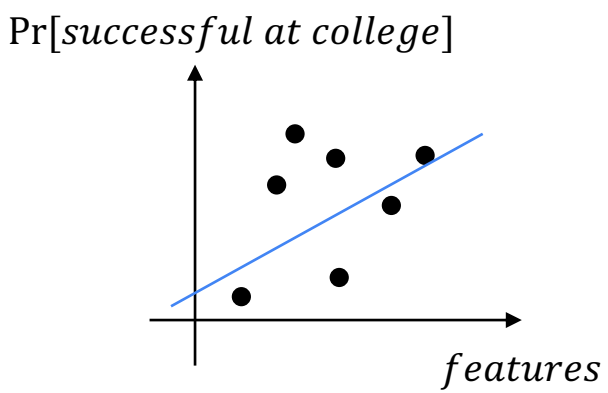
Implicit Assumption so Far: Homogeneous Population



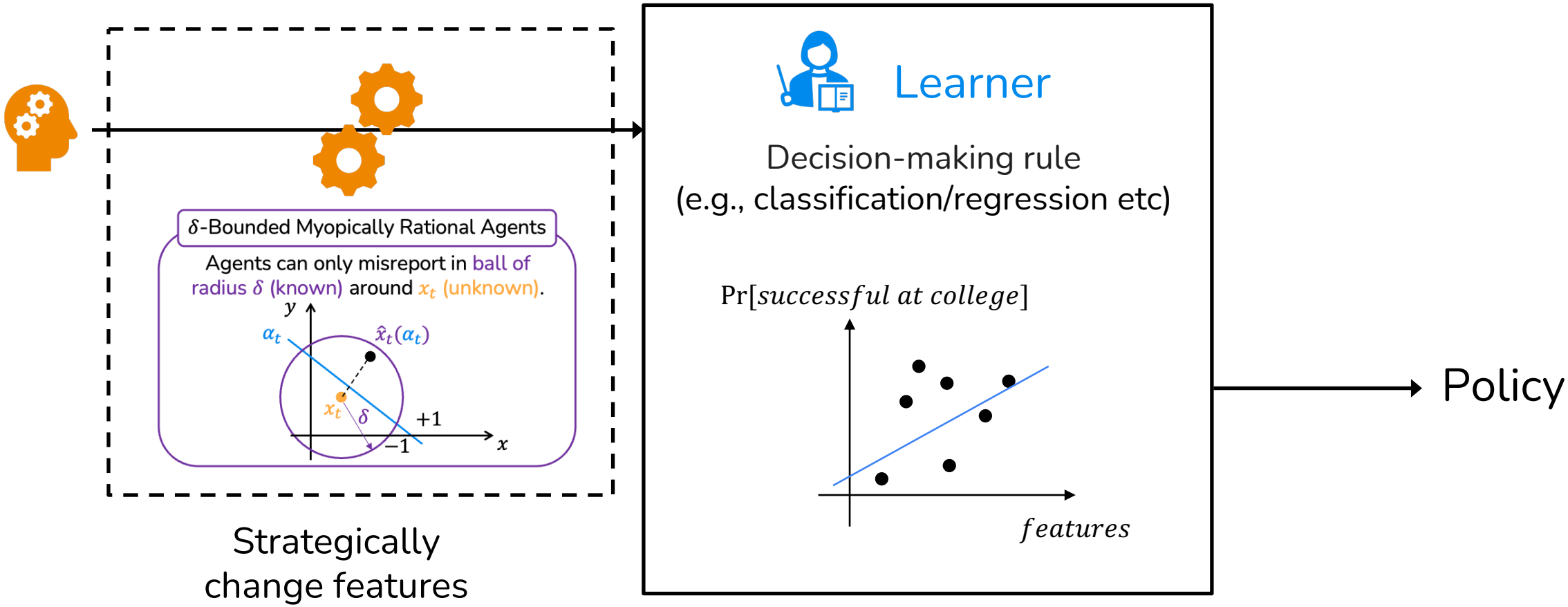
Implicit Assumption so Far: Homogeneous Population

Learner

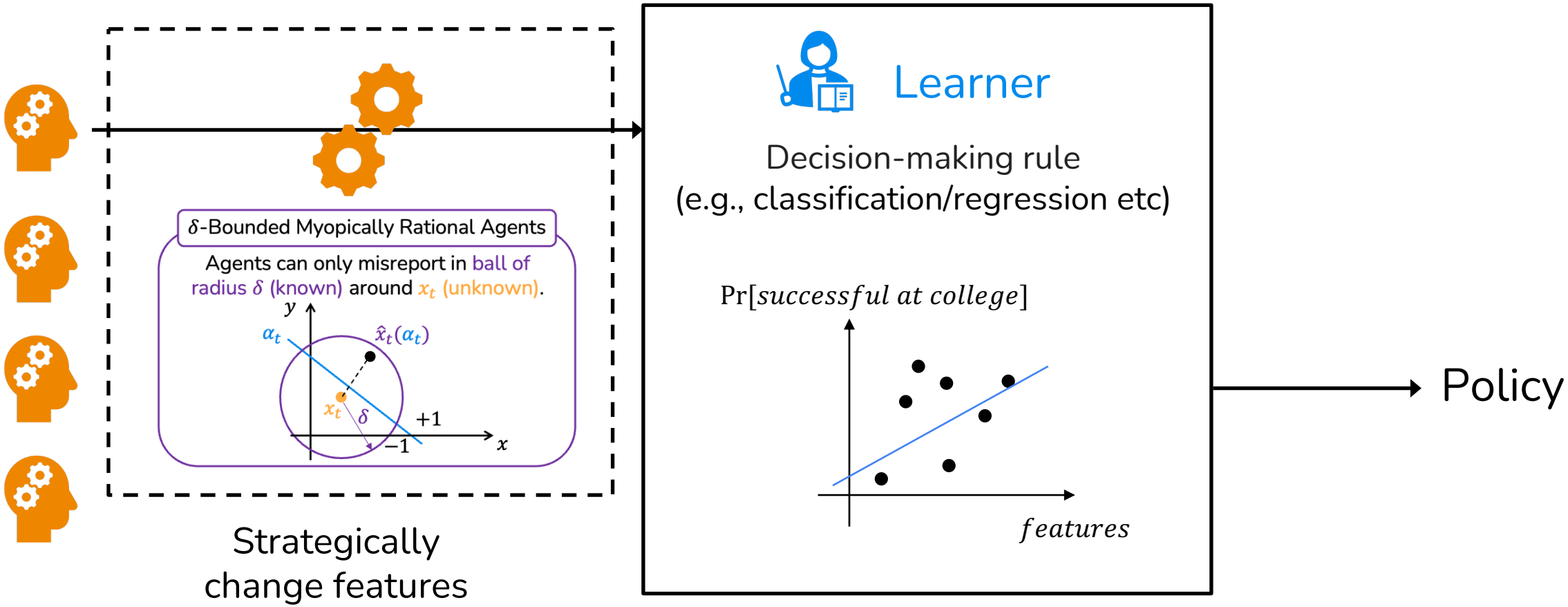
Decision-making rule
(e.g., classification/regression etc)



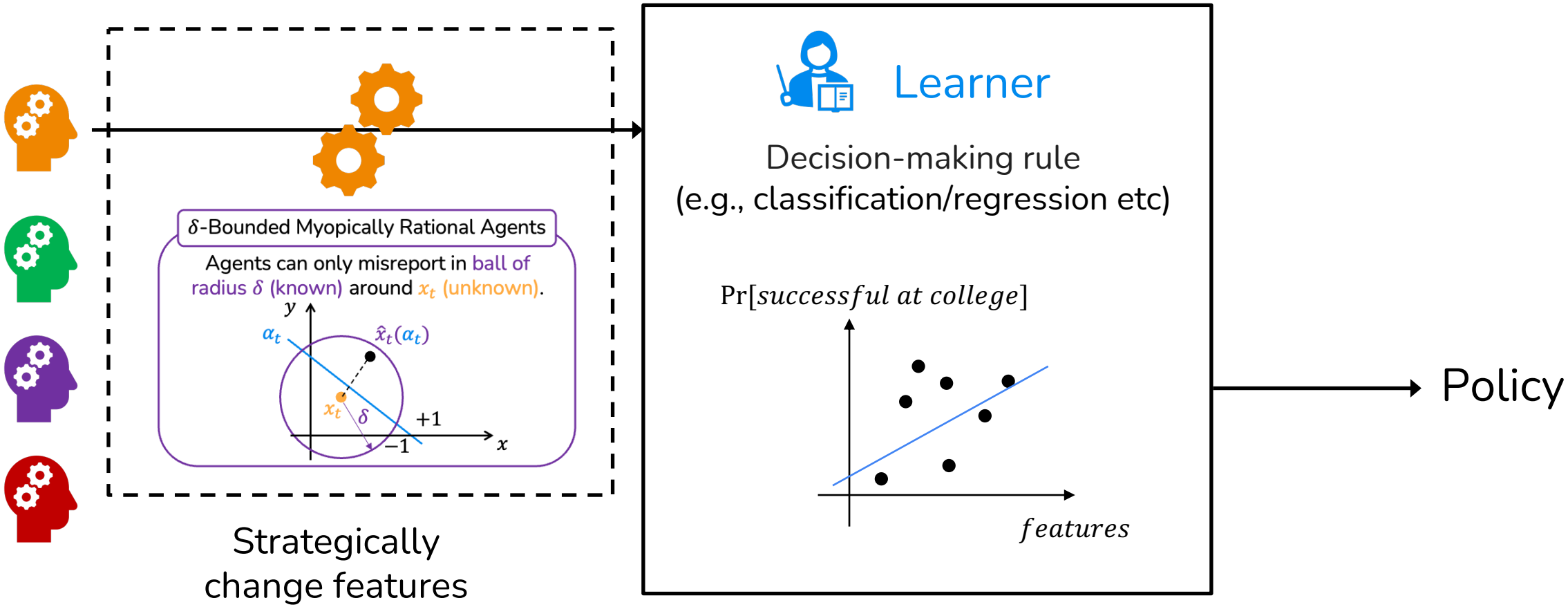
Implicit Assumption so Far: Homogeneous Population



Implicit Assumption so Far: Homogeneous Population



Reality: Highly Heterogeneous!



Reality: Highly Heterogeneous!

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

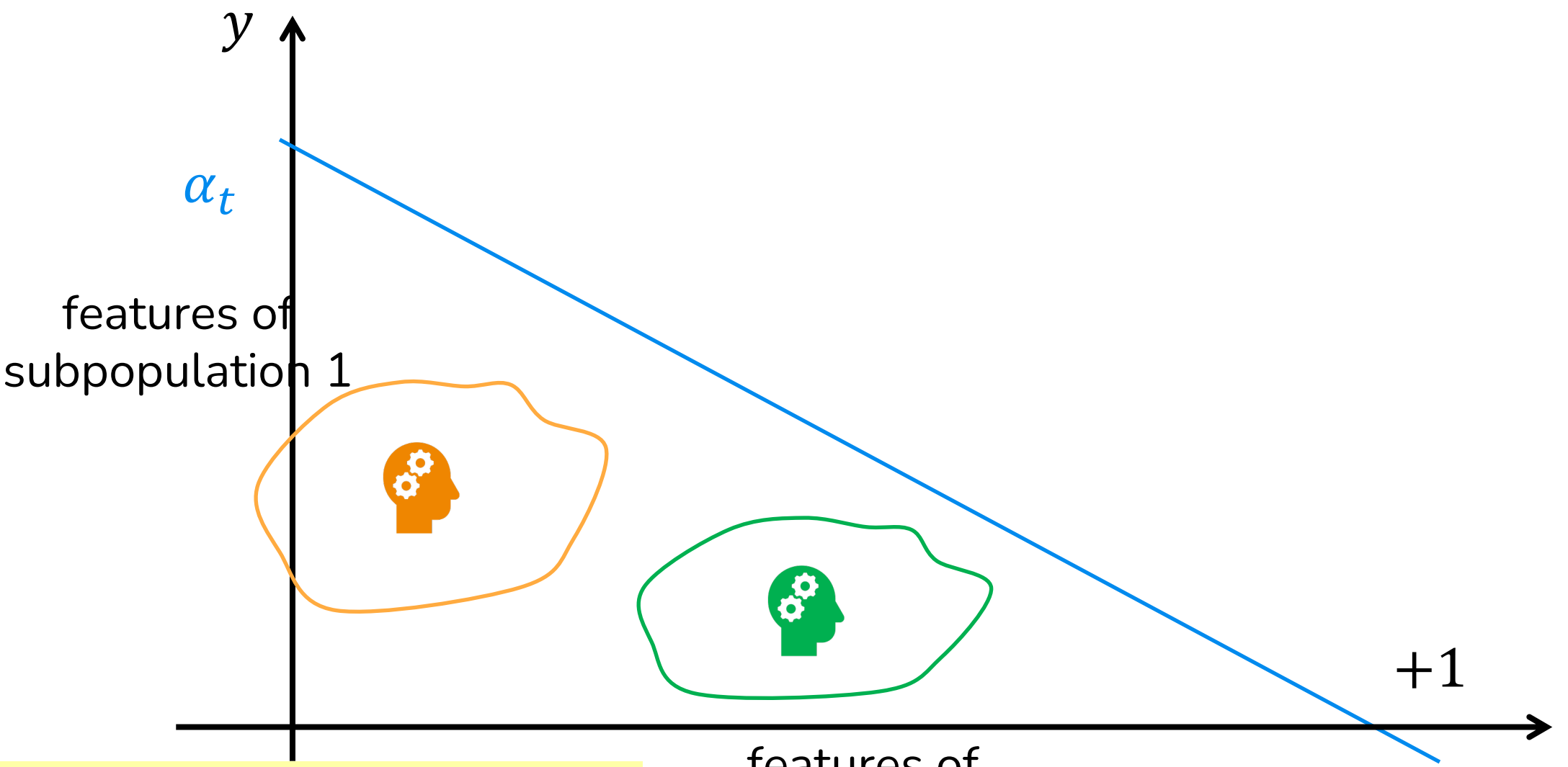
Reality: Highly Heterogeneous!

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

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

Reality: Highly Heterogeneous!

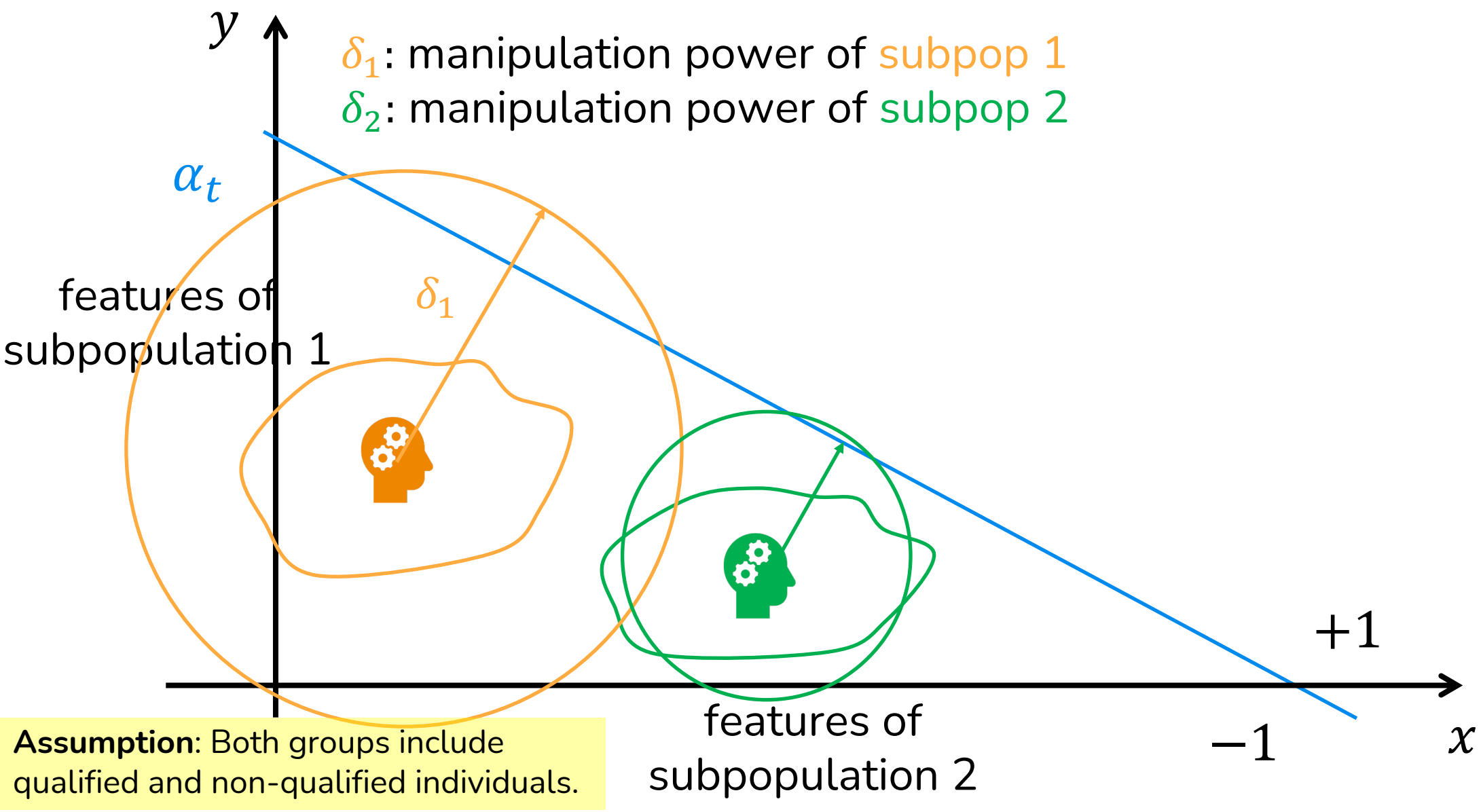
[Hu, Immorlica, Vaughan, **FAT*19**],
[Milli, Miller, Dragan, Hardt, **FAT*19**]



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

Reality: Highly Heterogeneous!

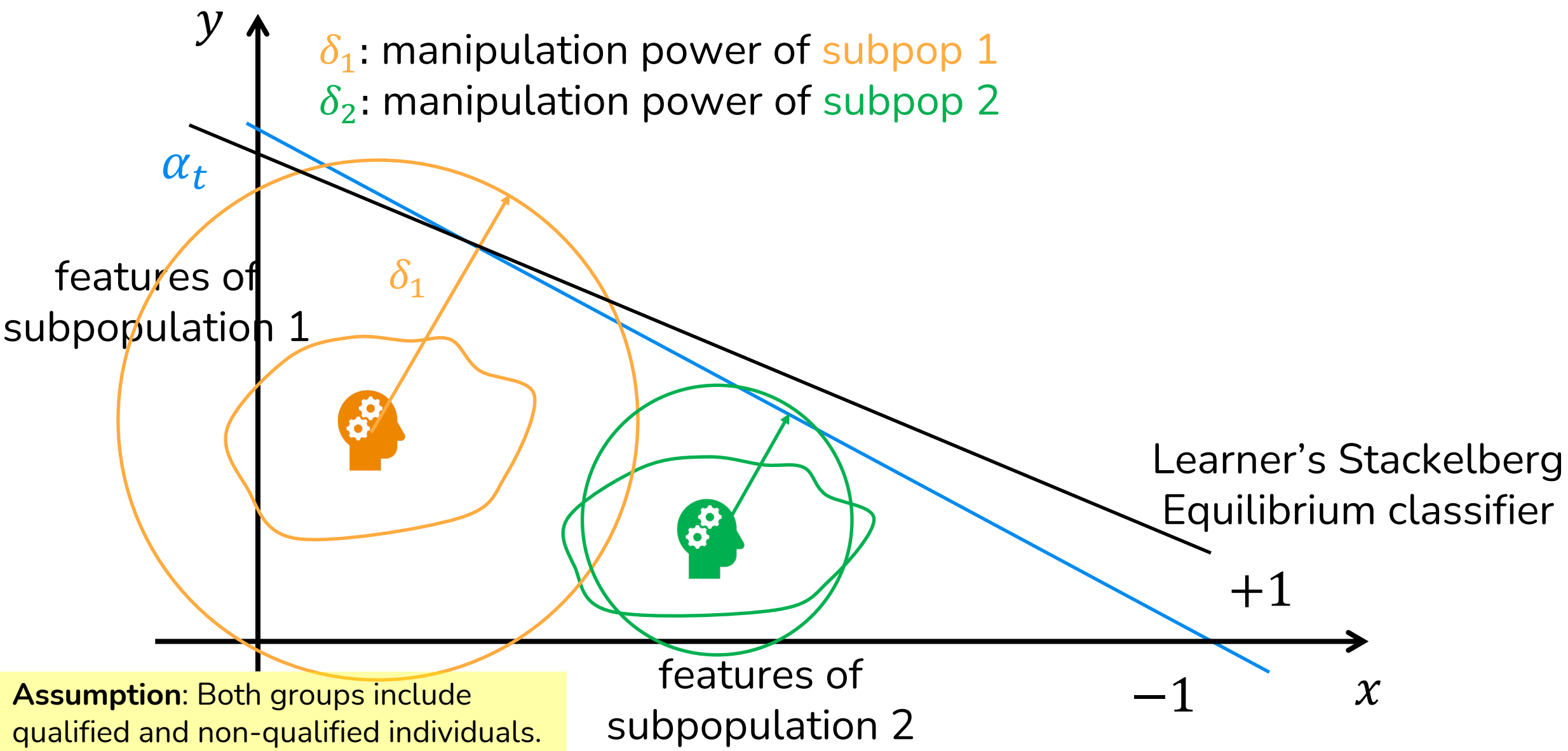
[Hu, Immorlica, Vaughan, **FAT*19**],
[Milli, Miller, Dragan, Hardt, **FAT*19**]



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

Reality: Highly Heterogeneous!

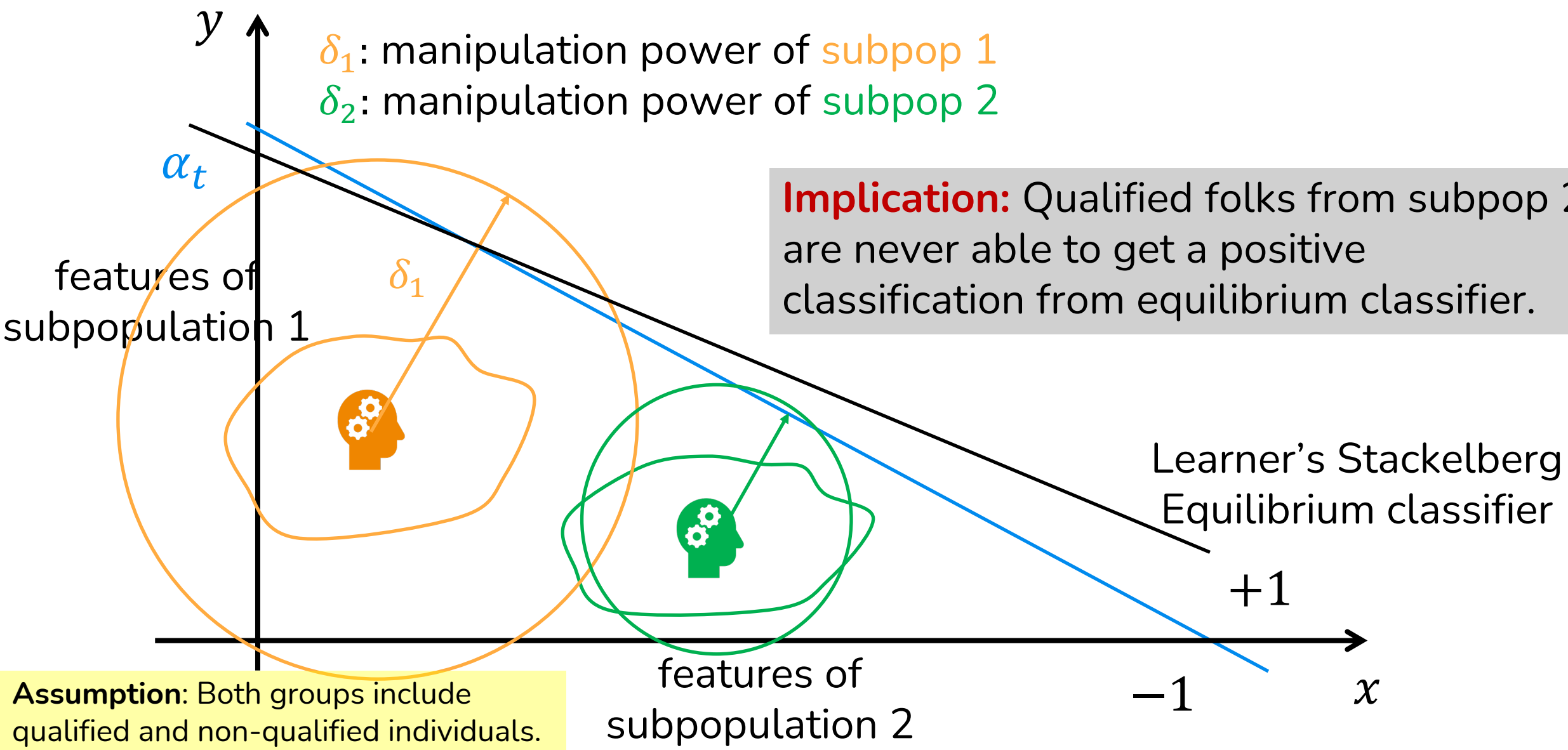
[Hu, Immorlica, Vaughan, **FAT*19**],
[Milli, Miller, Dragan, Hardt, **FAT*19**]



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

Reality: Highly Heterogeneous!

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

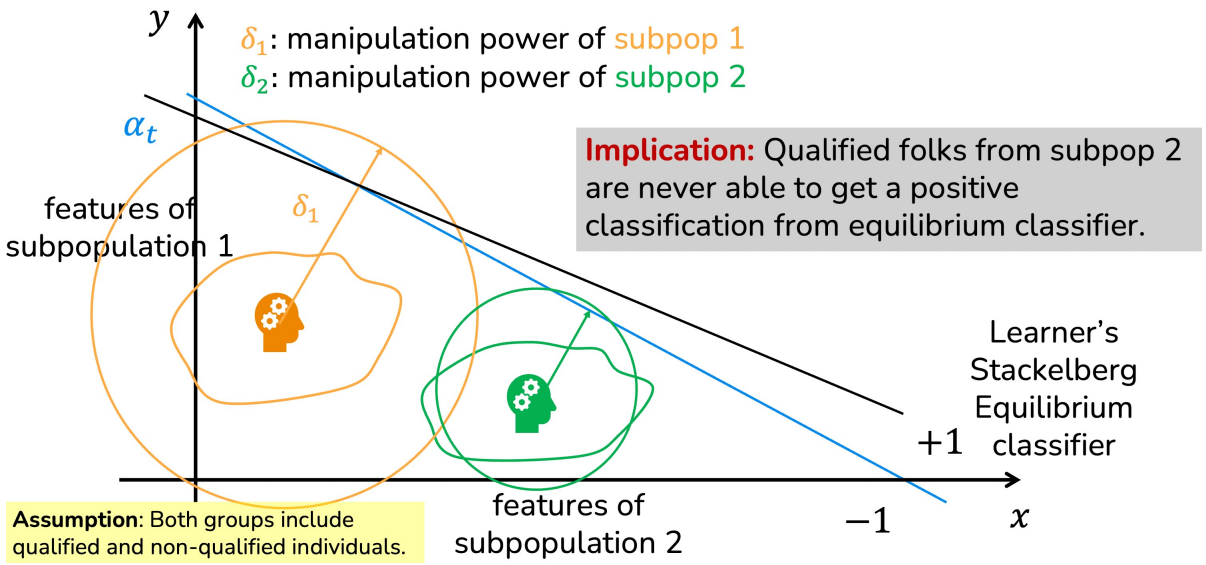


Implication: Qualified folks from subpop 2 are never able to get a positive classification from equilibrium classifier.

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

Reality: Highly Heterogeneous!

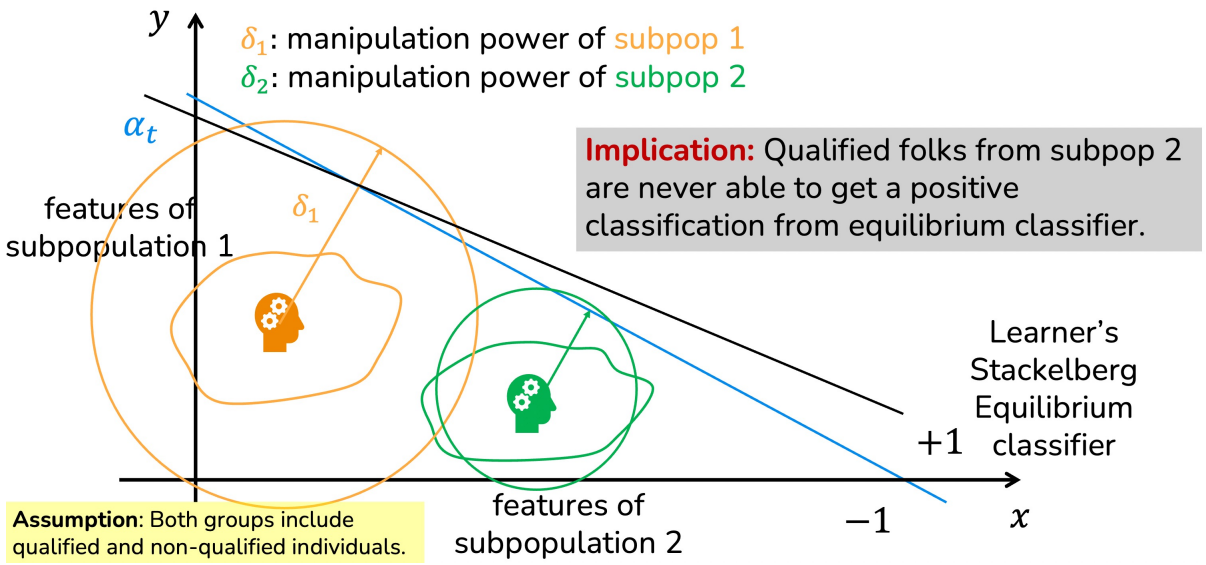
[Hu, Immorlica, Vaughan, **FAT*19**],
[Milli, Miller, Dragan, Hardt, **FAT*19**]



Summary

Reality: Highly Heterogeneous!

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

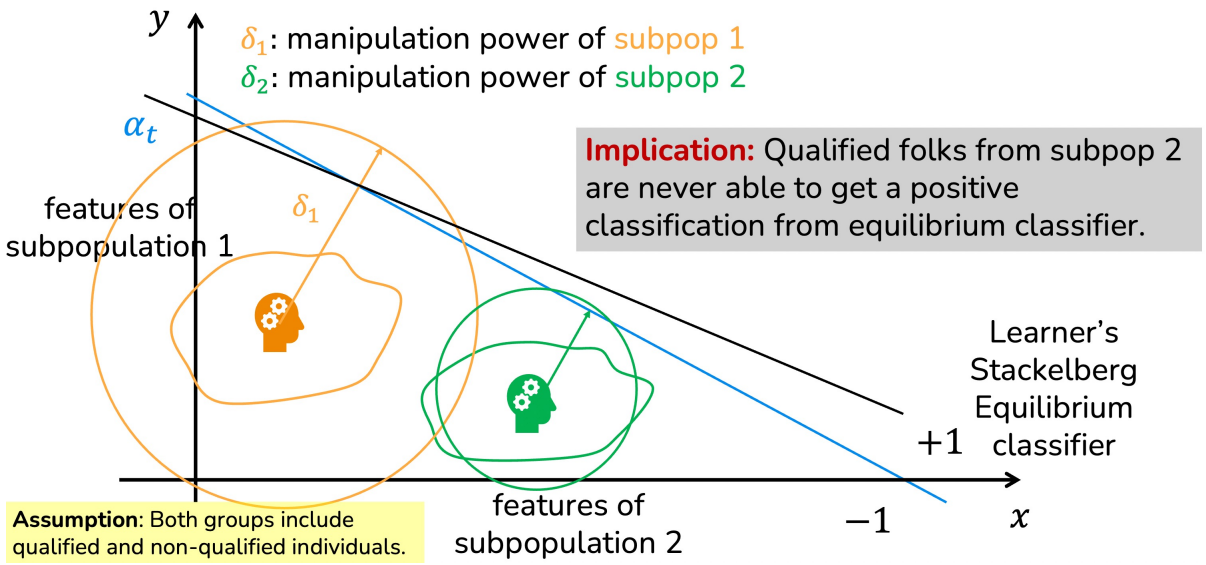


Summary

1) Strategic classification disproportionately affects disadvantaged population.

Reality: Highly Heterogeneous!

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

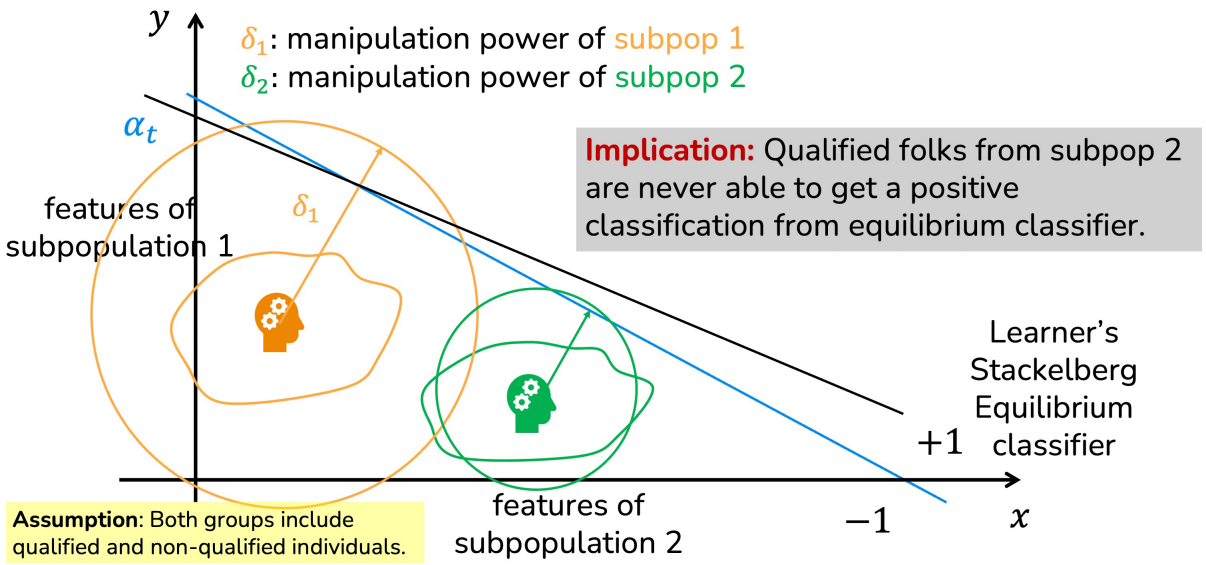


Summary

1) Strategic classification disproportionately affects disadvantaged population.
2) There are cases where subsidies make both subpopulations worse off, while making the learner better off.

Reality: Highly Heterogeneous!

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



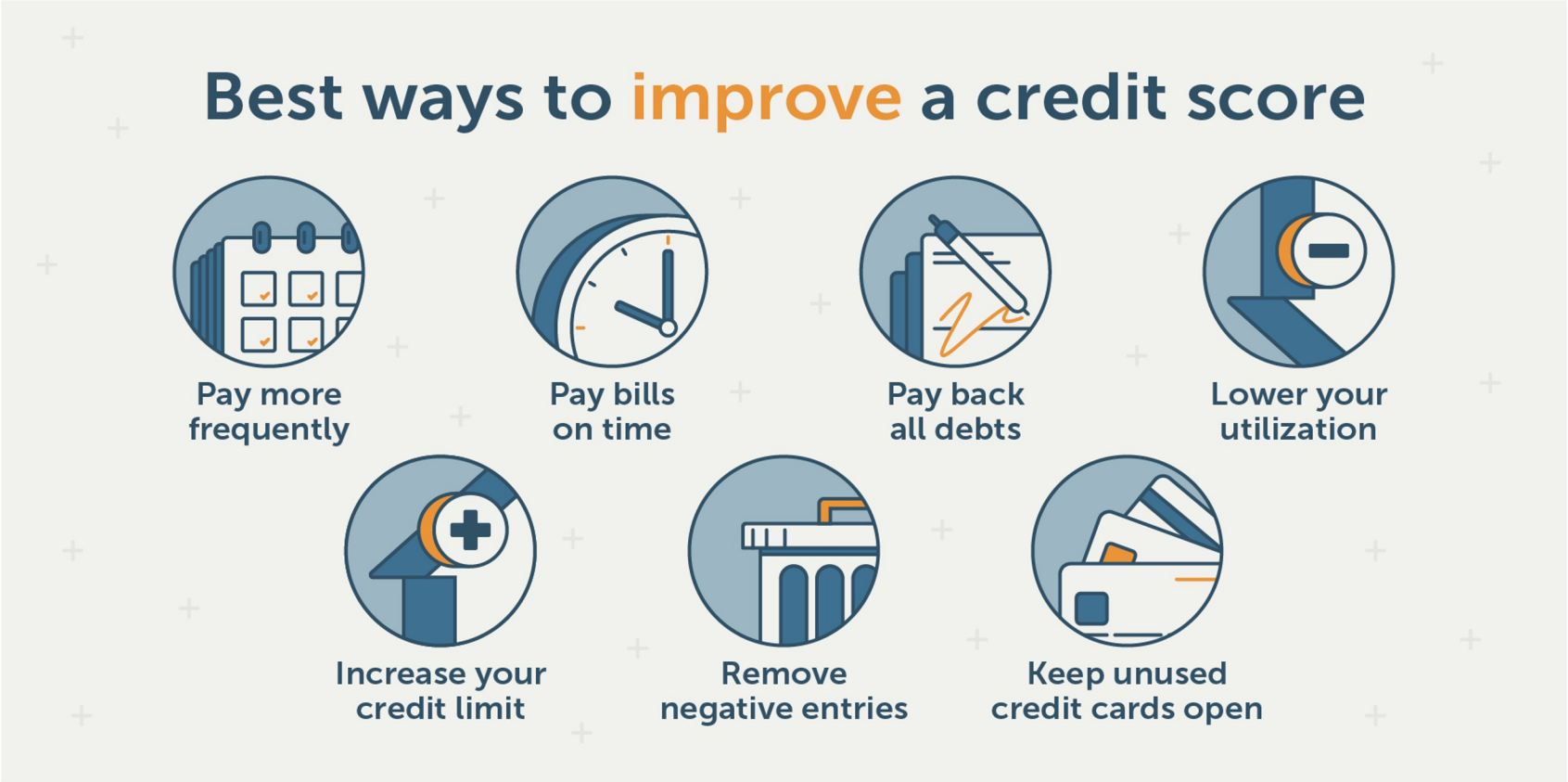
Summary

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.

Tutorial Outline

- ☑ Introduction
- ☑ Robustness
- ☑ Fairness
- ❑ Recourse/Performativity/Causality
- ❑ Future Directions/Open Questions

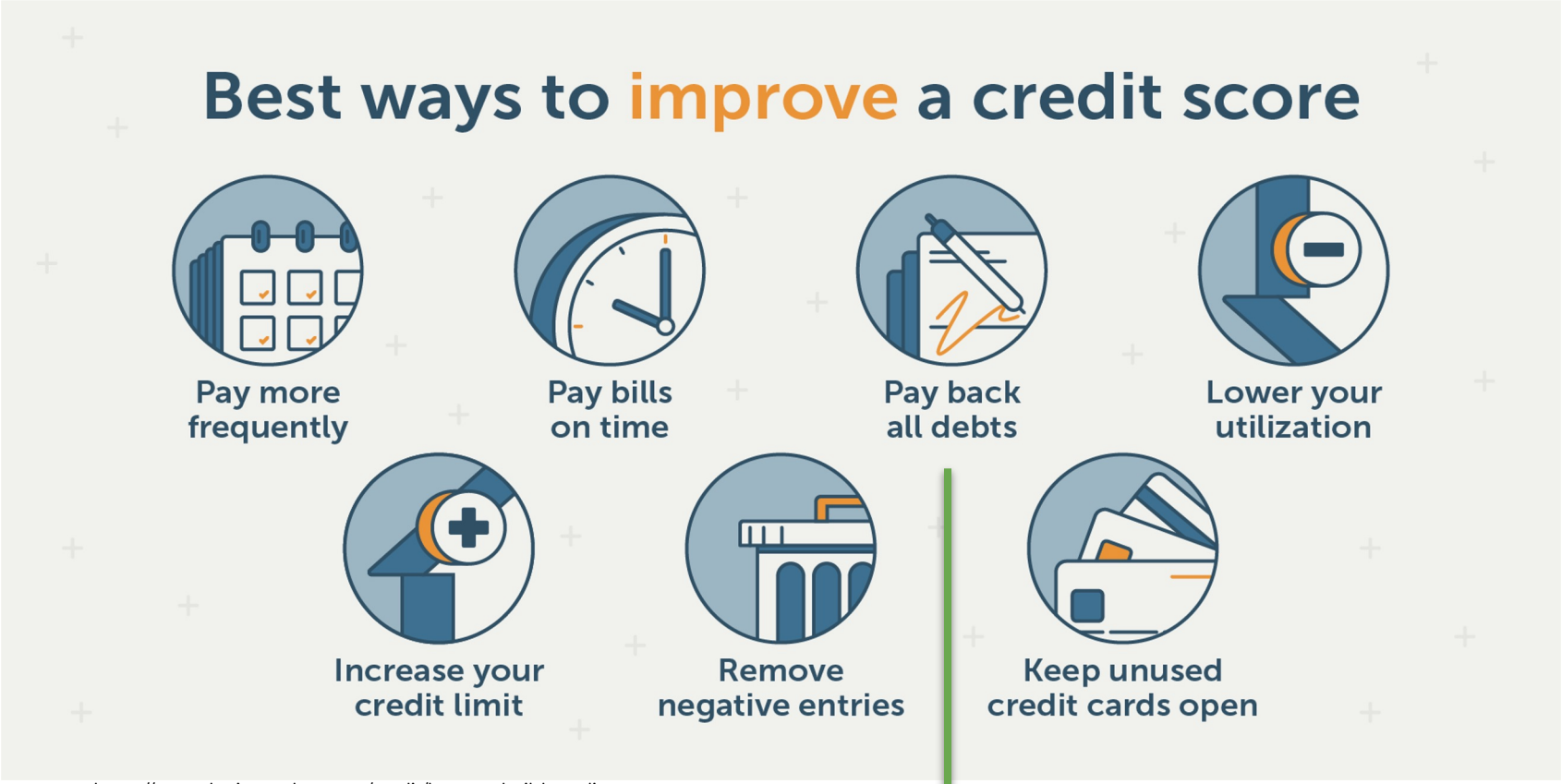
Is It All Just Gaming?



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

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Is It All Just Gaming?



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

ability to pay back future loans

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

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

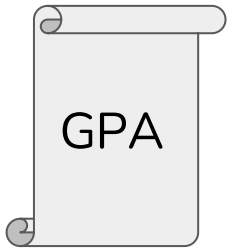
institution

Model

> Prediction

Features

True label



individual

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

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

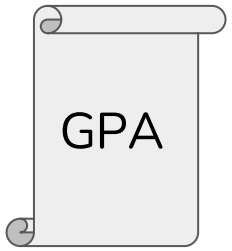
institution

Model

> Prediction

Features

True label



individual

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

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

institution

Model

> Prediction

Features

True label

GPA

individual

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

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

institution

Model

> Prediction

CollegeBoard
Advanced Placement Program

Features

True label

GPA

individual

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

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

institution

Model

> Prediction

CollegeBoard
Advanced Placement
Program

Features

True label

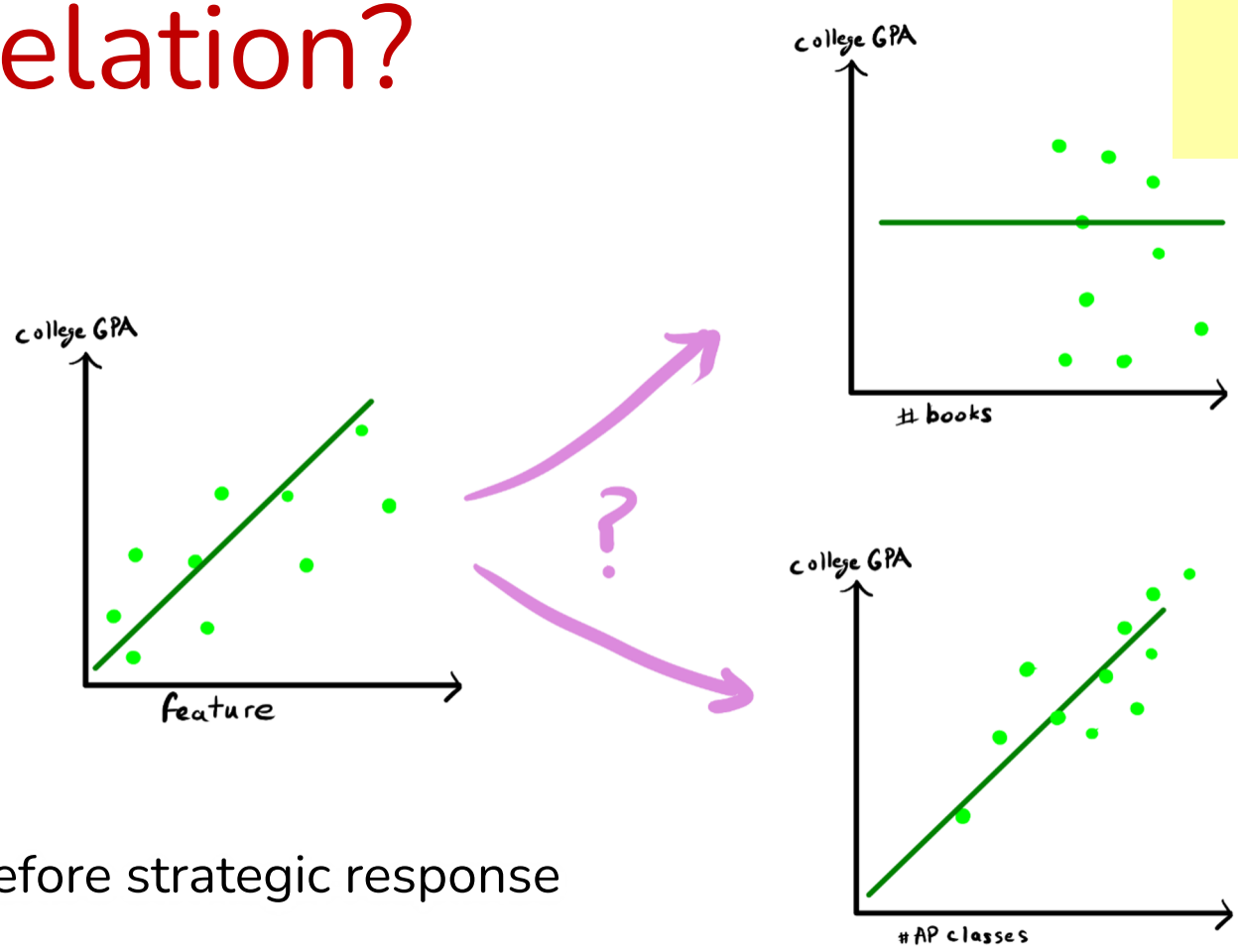
GPA

individual

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Causation or Just Correlation?

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]
[Bechavod, Ligett, Wu, Ziani, **AISTATS20**]



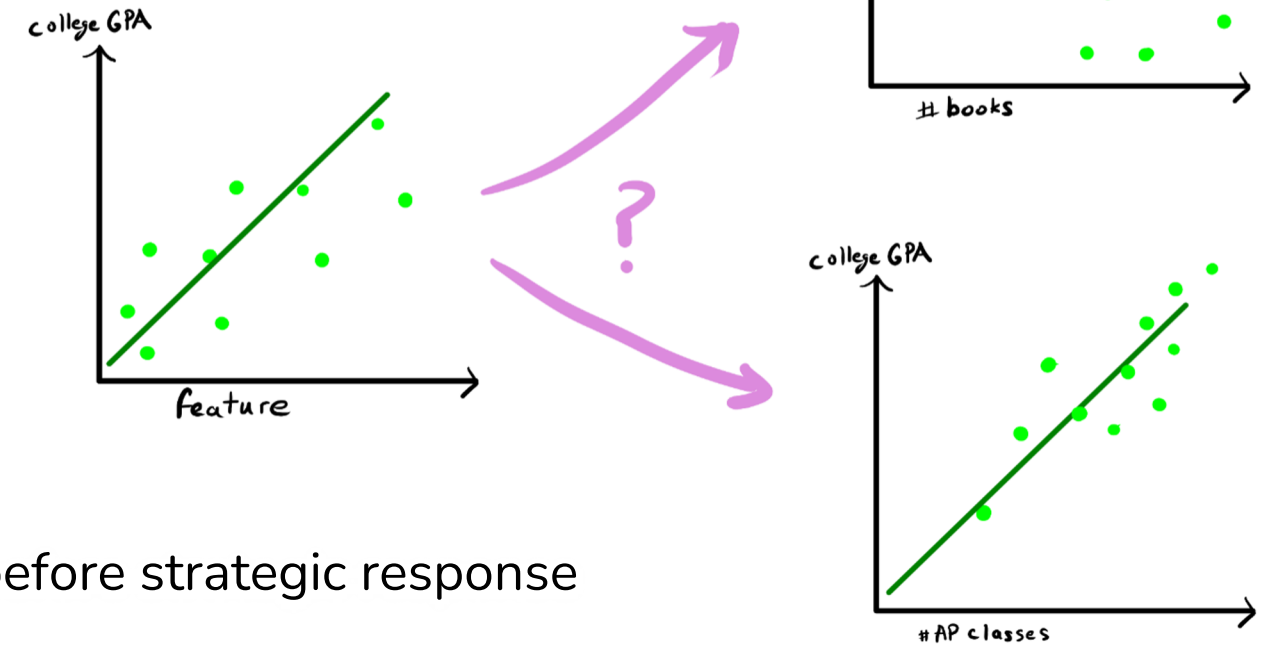
before strategic response

after strategic response

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Causation or Just Correlation?

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]
[Bechavod, Ligett, Wu, Ziani, **AISTATS20**]



before strategic response

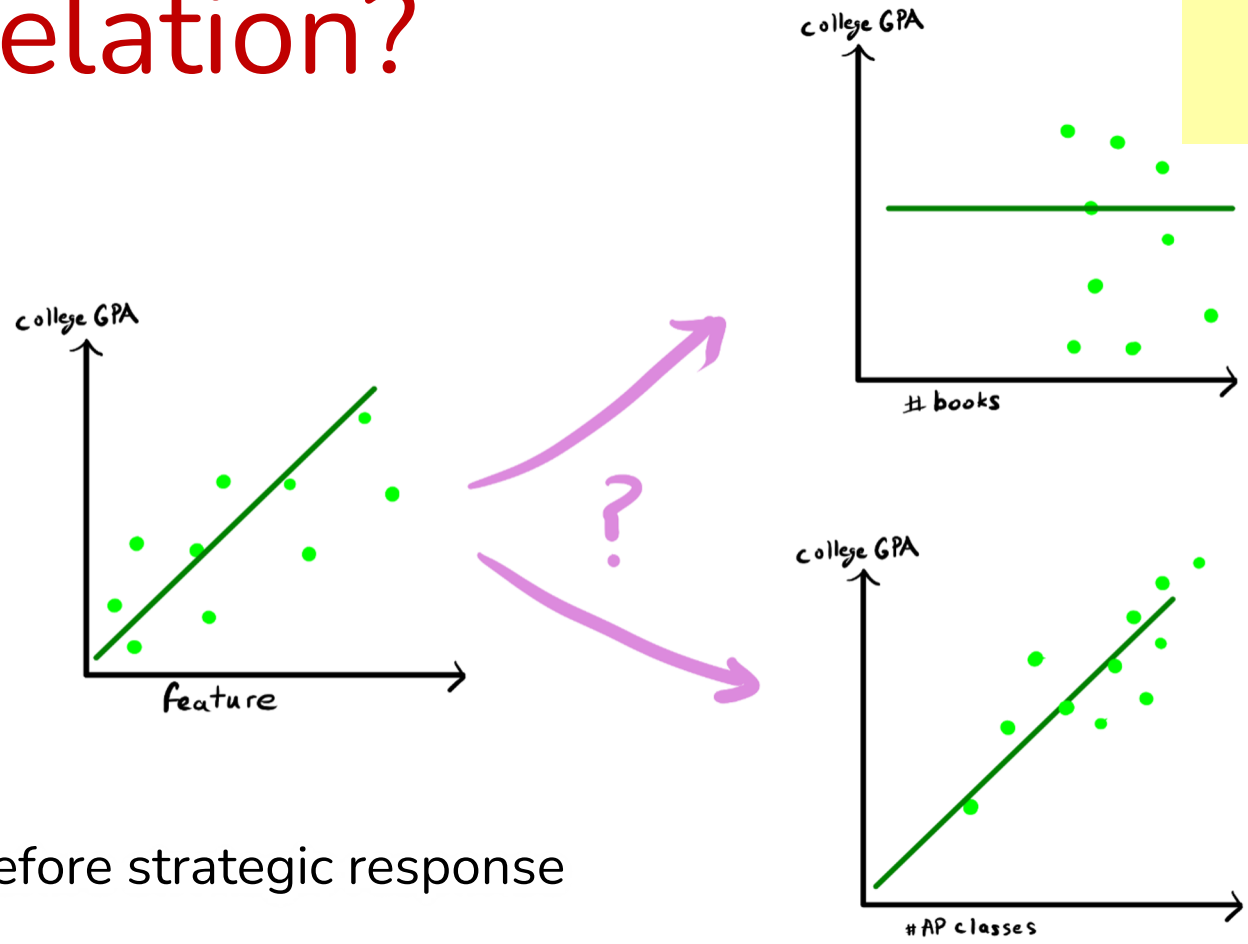
gaming

after strategic response

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Causation or Just Correlation?

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]
[Bechavod, Ligett, Wu, Ziani, **AISTATS20**]



before strategic response

gaming

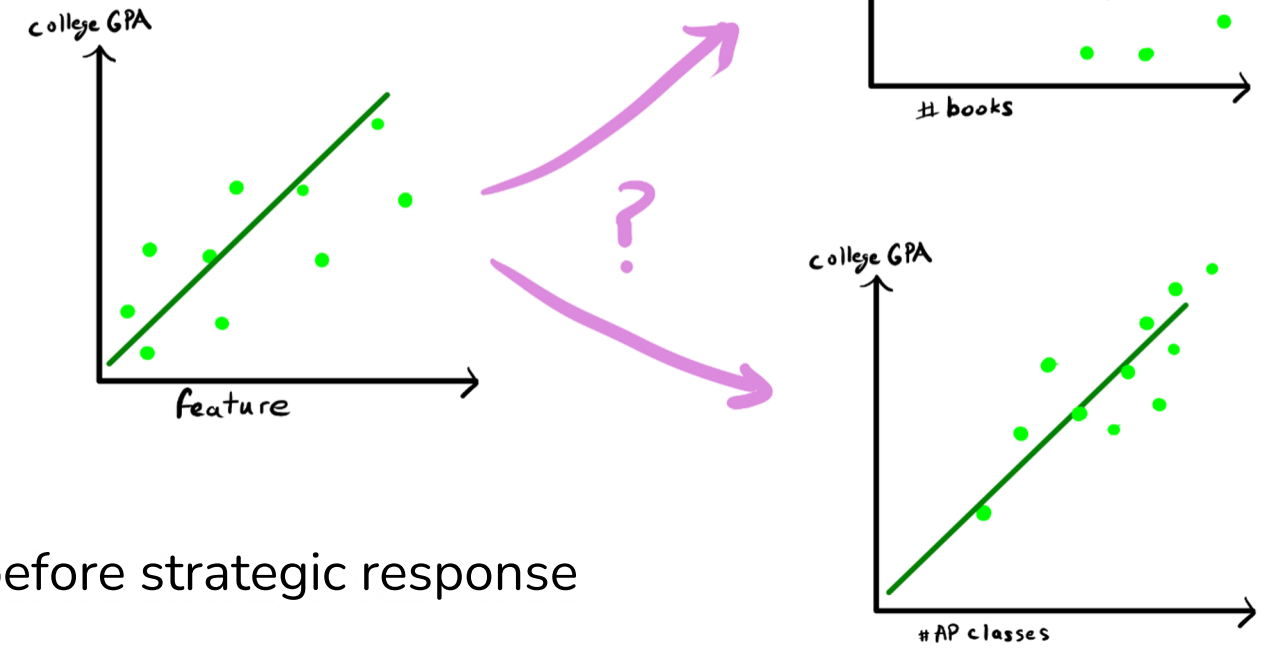
improvement:
no cost to
transparency

after strategic response

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Causation or Just Correlation?

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]
[Bechavod, Ligett, Wu, Ziani, **AISTATS20**]



before strategic response

gaming

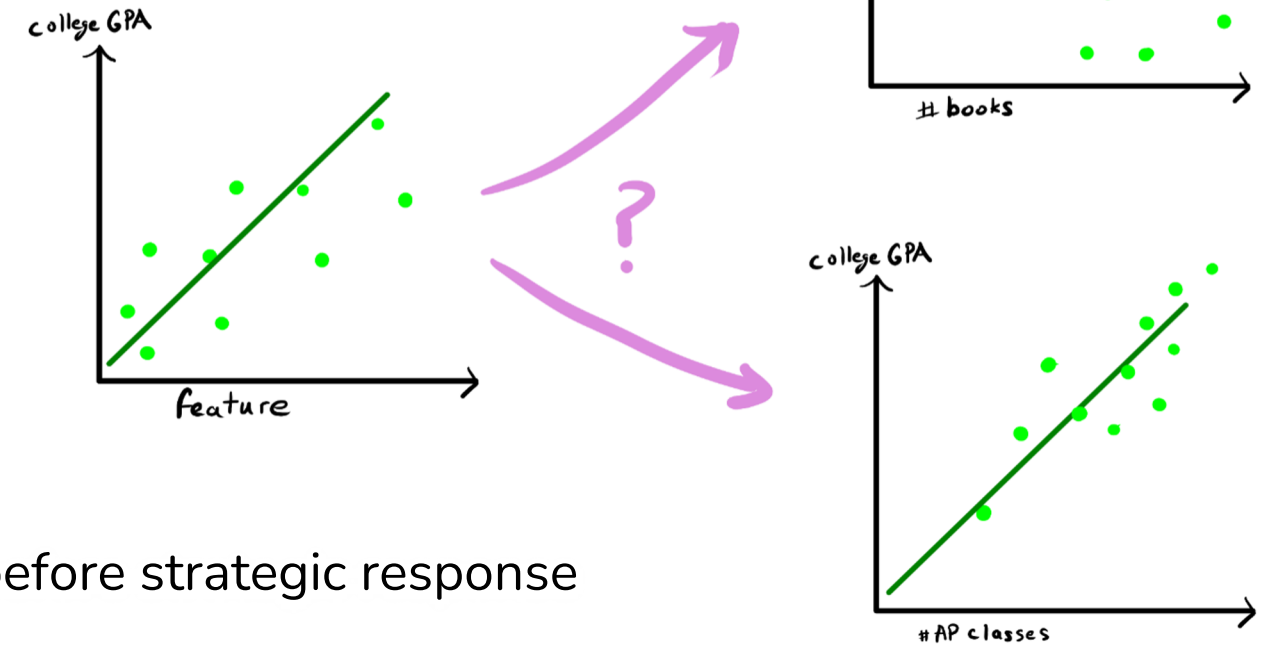
improvement:
no cost to
transparency

after strategic response

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Causation or Just Correlation?

[Kleinberg & Raghavan, **EC19**]
[Miller, Milli, Hardt, **ICML20**]
[Haghtalab, Immorlica, Lucier, Wang, **IJCAI20**]
[Shavit, Edelman, Axelrod, **ICML20**]
[Bechavod, Ligett, Wu, Ziani, **AISTATS20**]



before strategic response

gaming

improvement:
no cost to
transparency

after strategic response

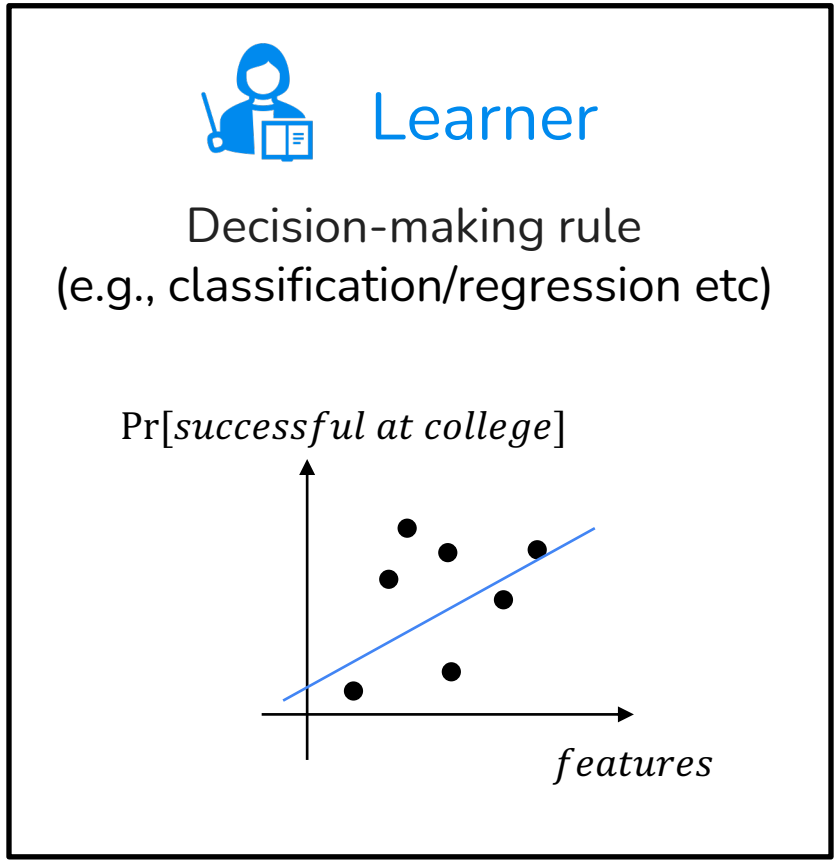
Is the "transparency" assumption realistic?

Slide adapted from FAccT21 tutorial co-taught with Ben Edelman, Yo Shavit

Strategic Learning Revisited

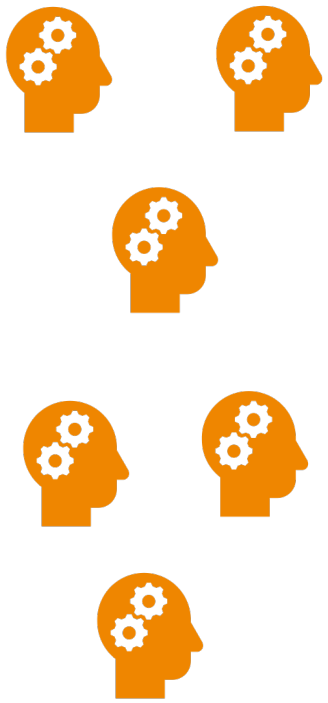
[Bechavod, P., Wu, Ziani, **ICML22]**

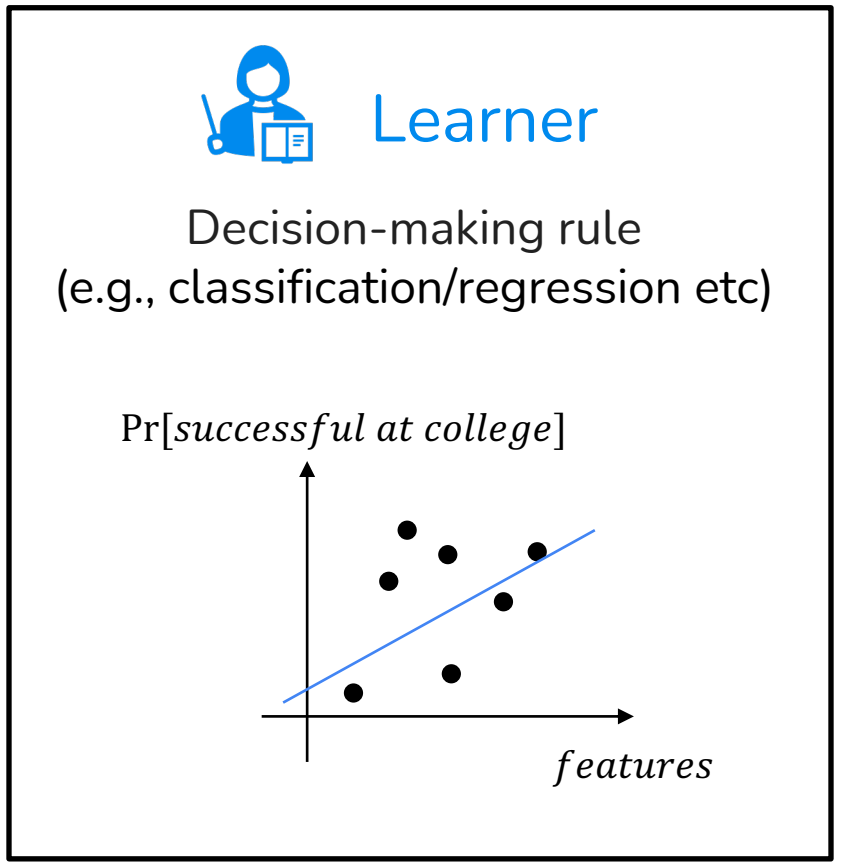
Strategic Learning Revisited



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

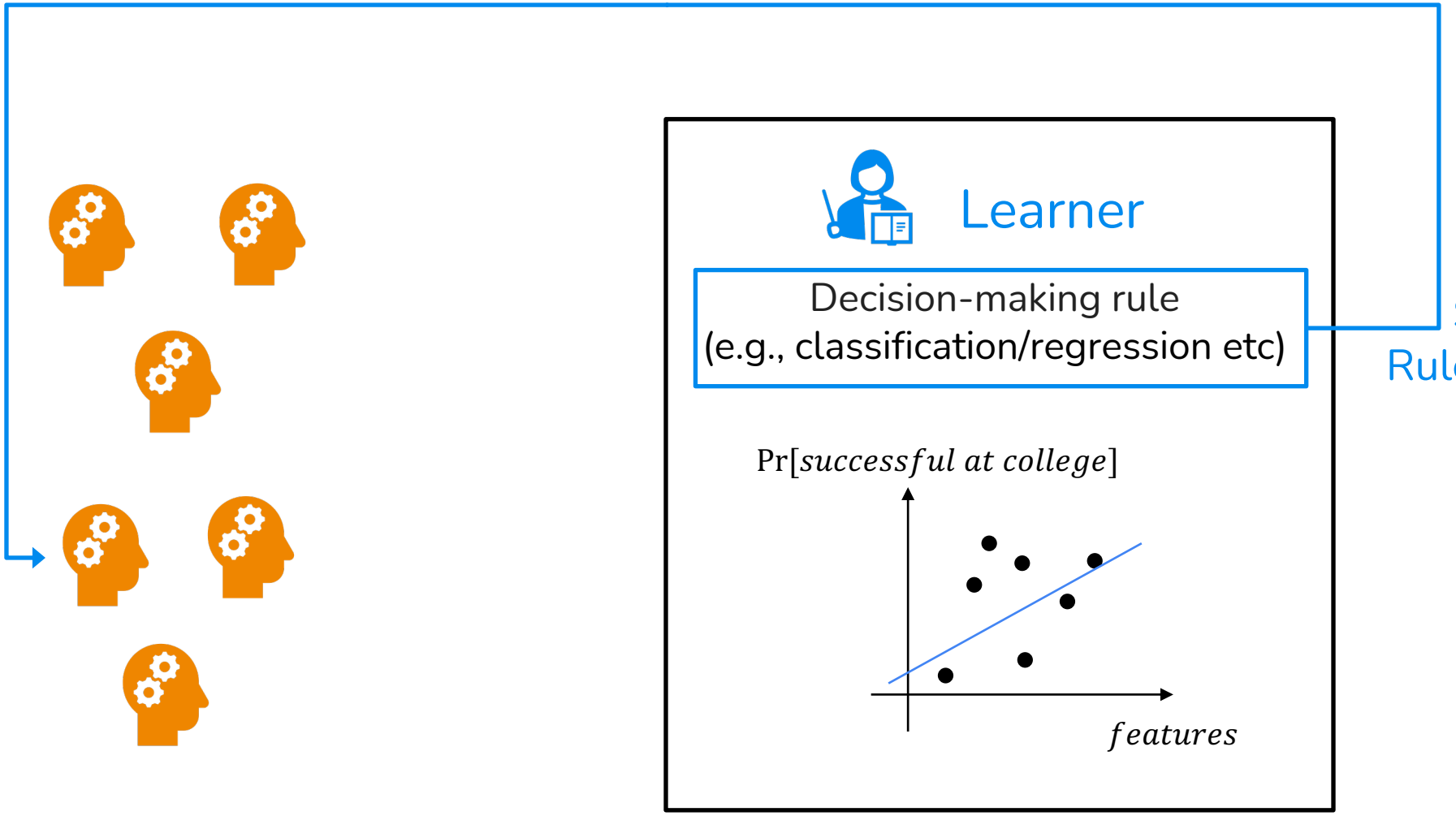
Strategic Learning Revisited





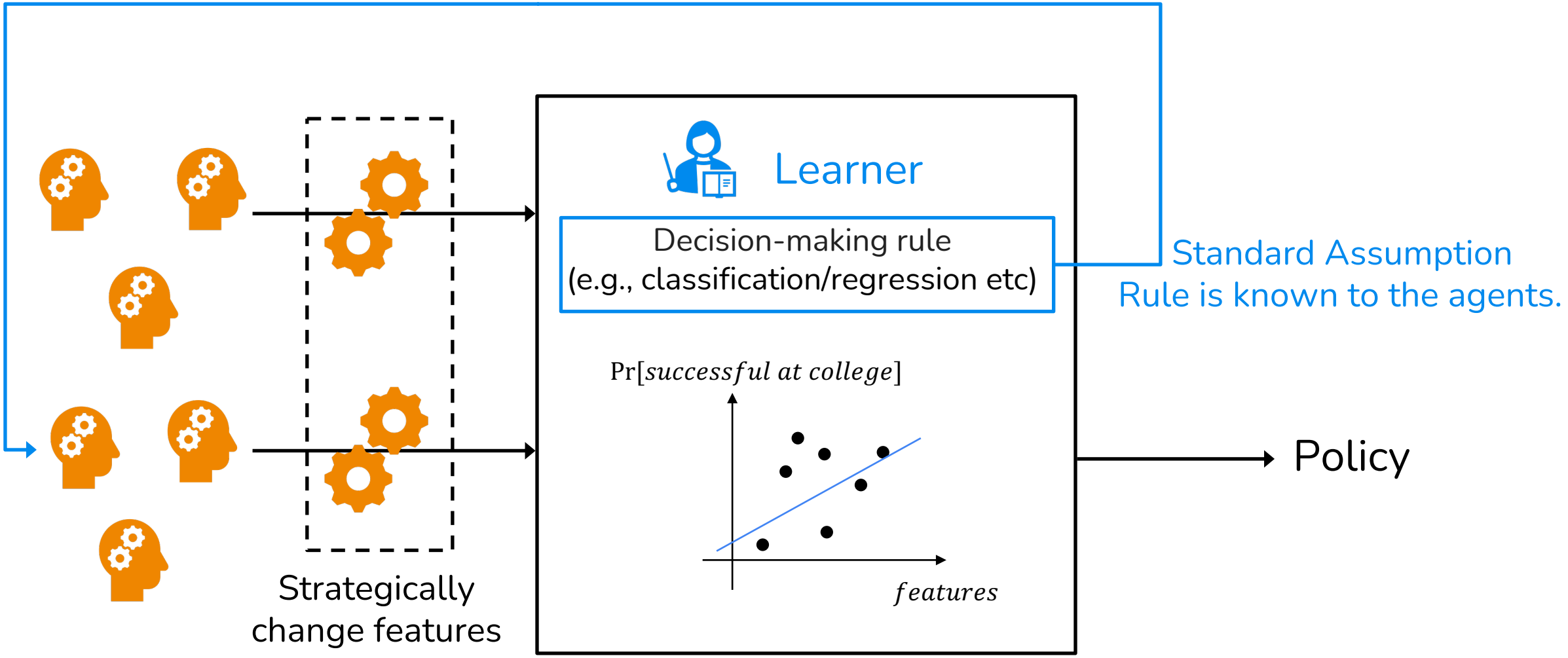
[Bechavod, P., Wu, Ziani, **ICML22]**

Strategic Learning Revisited



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

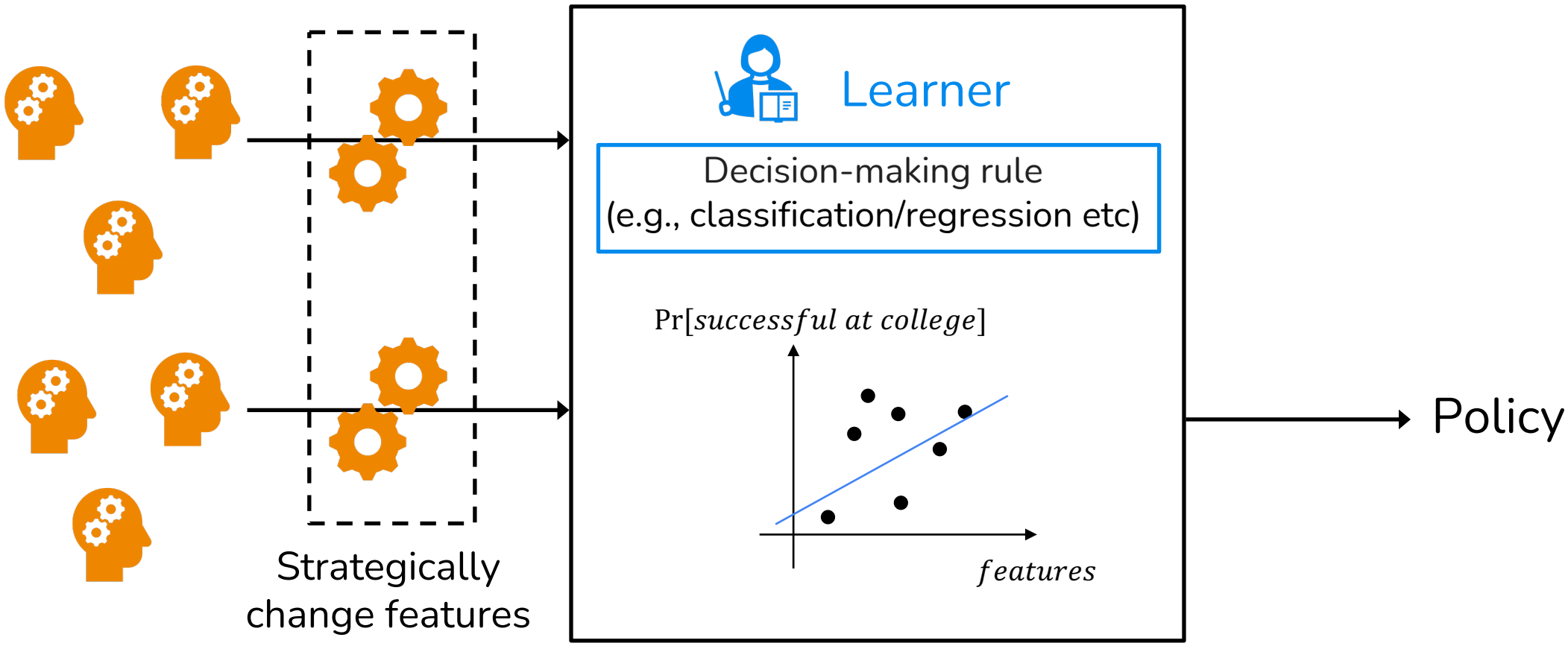
Strategic Learning Revisited



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

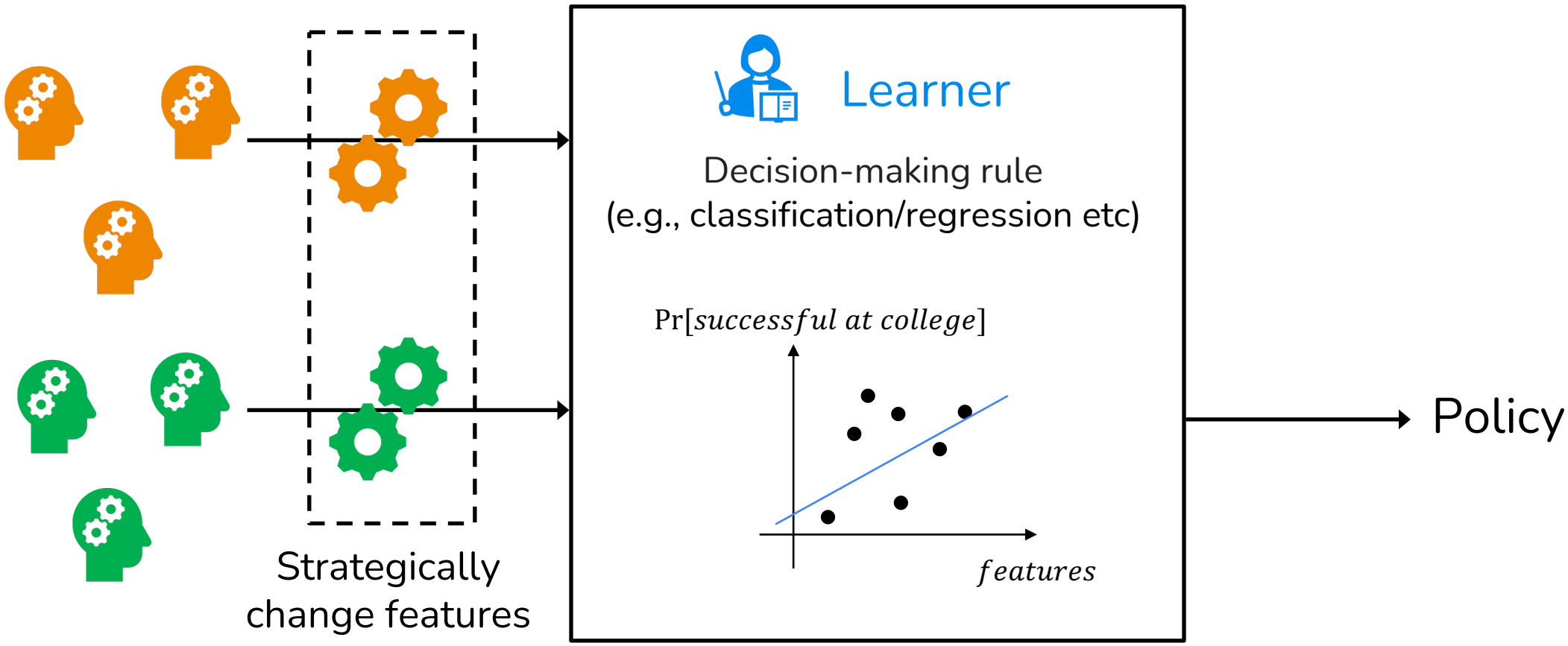
Strategic Learning Revisited

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



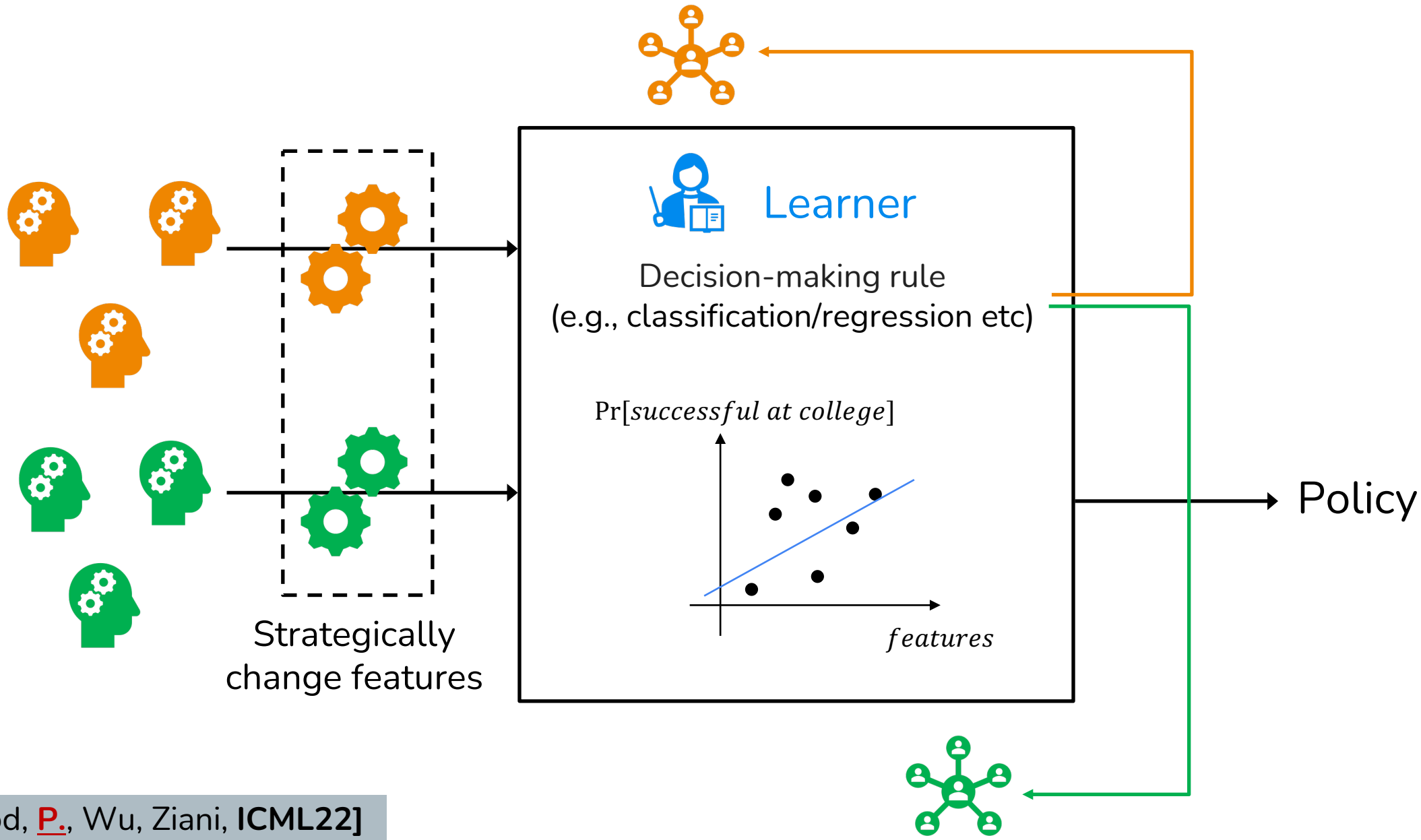
[Bechavod, P., Wu, Ziani, **ICML22]**

Strategic Learning Revisited



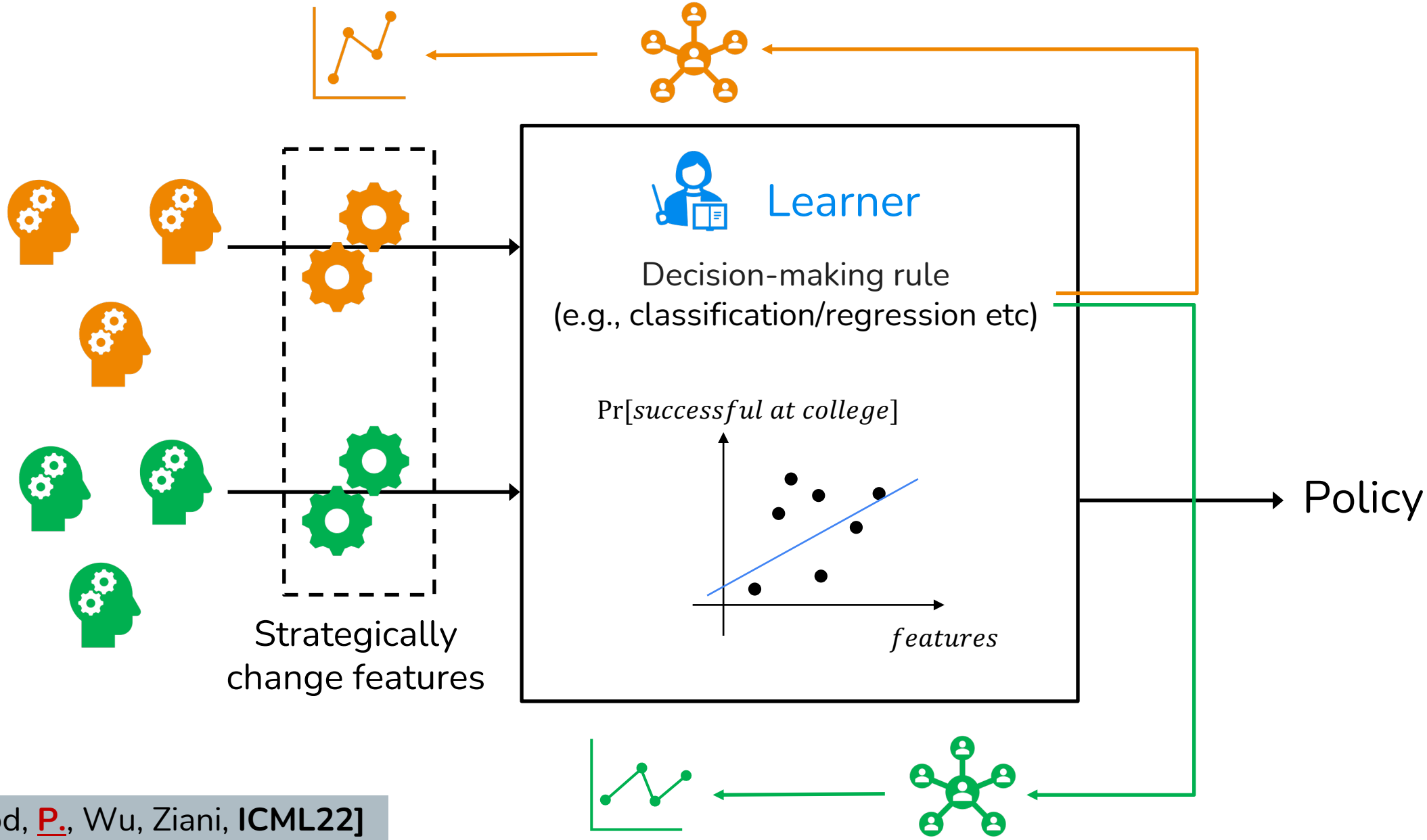
[Bechavod, P., Wu, Ziani, **ICML22]**

Strategic Learning Revisited



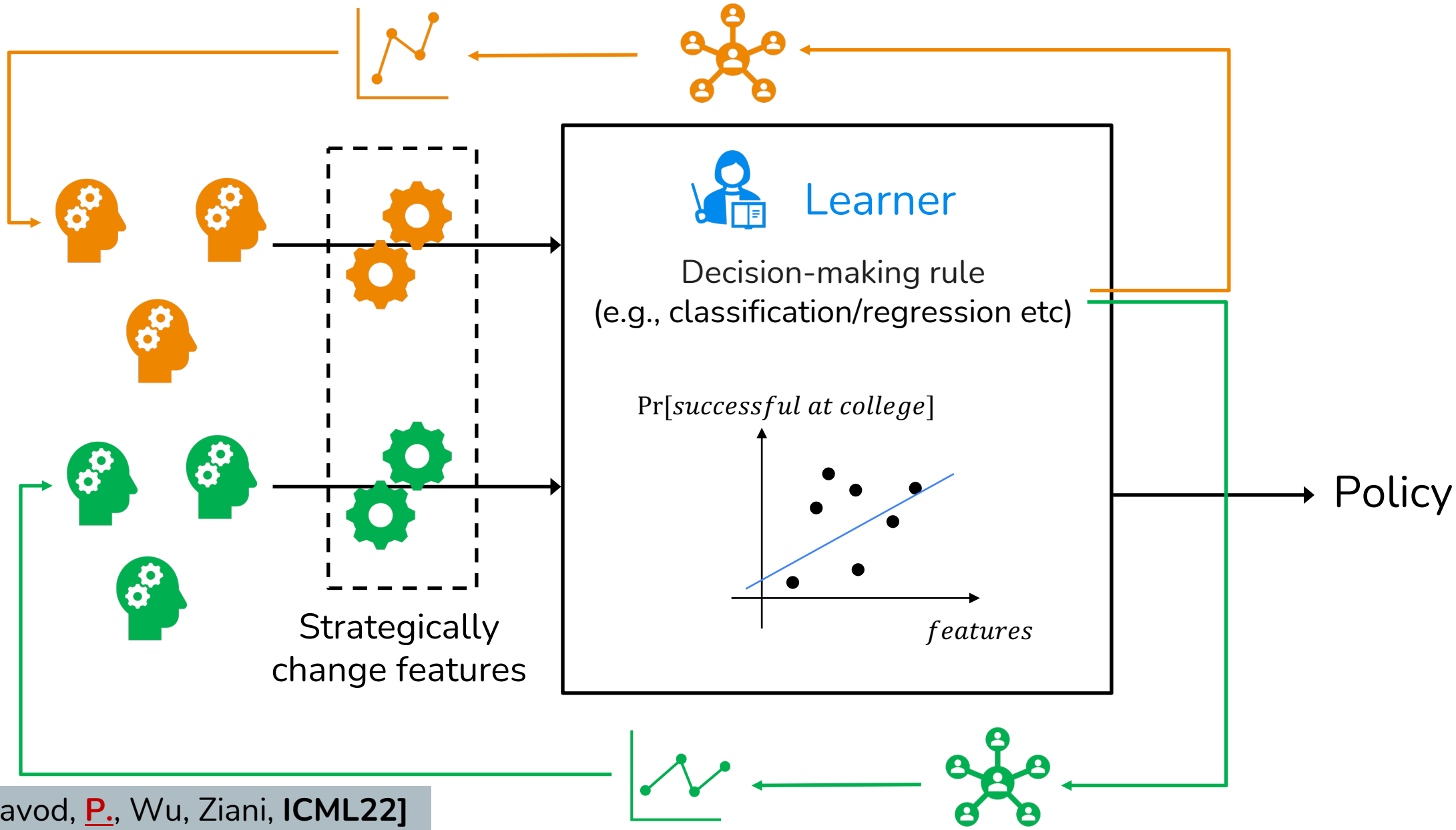
[Bechavod, P., Wu, Ziani, **ICML22]**

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]**

The Adult Dataset

- Publicly available at UCI repository: https://archive.ics.uci.edu/ml/datasets/adult
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): <50K, >= 50K

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

The Adult Dataset

- Publicly available at UCI repository: https://archive.ics.uci.edu/ml/datasets/adult
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): <50K, >= 50K

Our process:

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

The Adult Dataset

- Publicly available at UCI repository: https://archive.ics.uci.edu/ml/datasets/adult
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): <50K, >= 50K

Our process:

- 4 experiments separating subpopulations based on:

Characteristic	Subpopulation 1	Subpopulation 2
Age	<35 yrs old	>=35 yrs old
Country	All others	Western countries
Education	All others	Above high school
Race	All others	White

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

The Adult Dataset

- Publicly available at UCI repository: https://archive.ics.uci.edu/ml/datasets/adult
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): <50K, >= 50K

Our process:

- 4 experiments separating subpopulations based on:

Characteristic	Subpopulation 1	Subpopulation 2
Age	<35 yrs old	>=35 yrs old
Country	All others	Western countries
Education	All others	Above high school
Race	All others	White

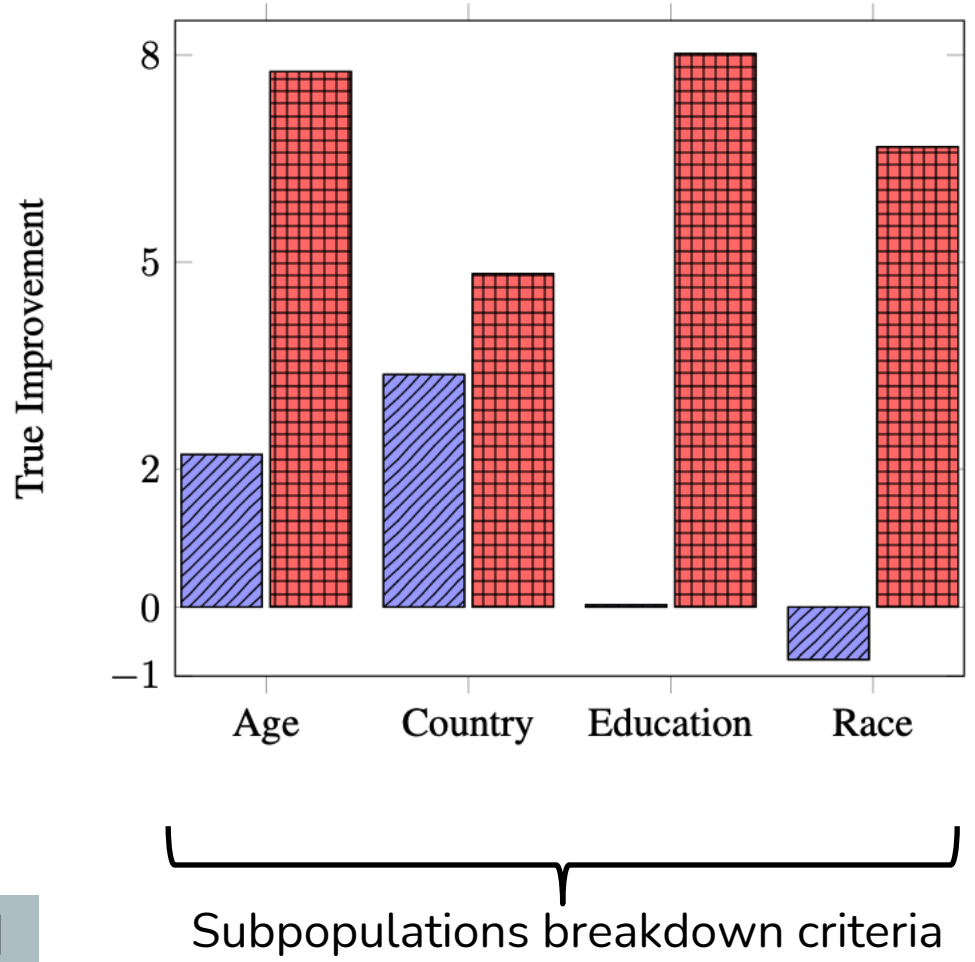
- Predict income improvement (final income – original income) for each subpopulation.

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

Results Snapshot: Adult Dataset

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

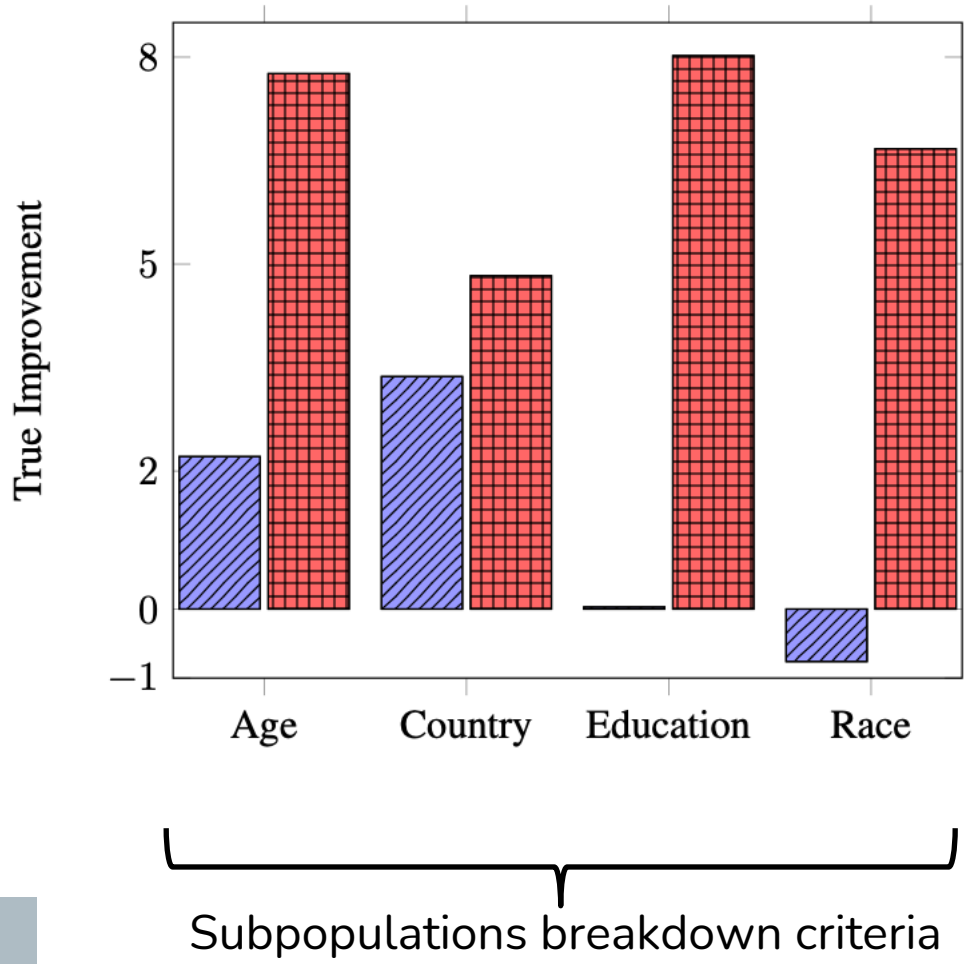
Results Snapshot: Adult Dataset



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

Results Snapshot: Adult Dataset

- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2

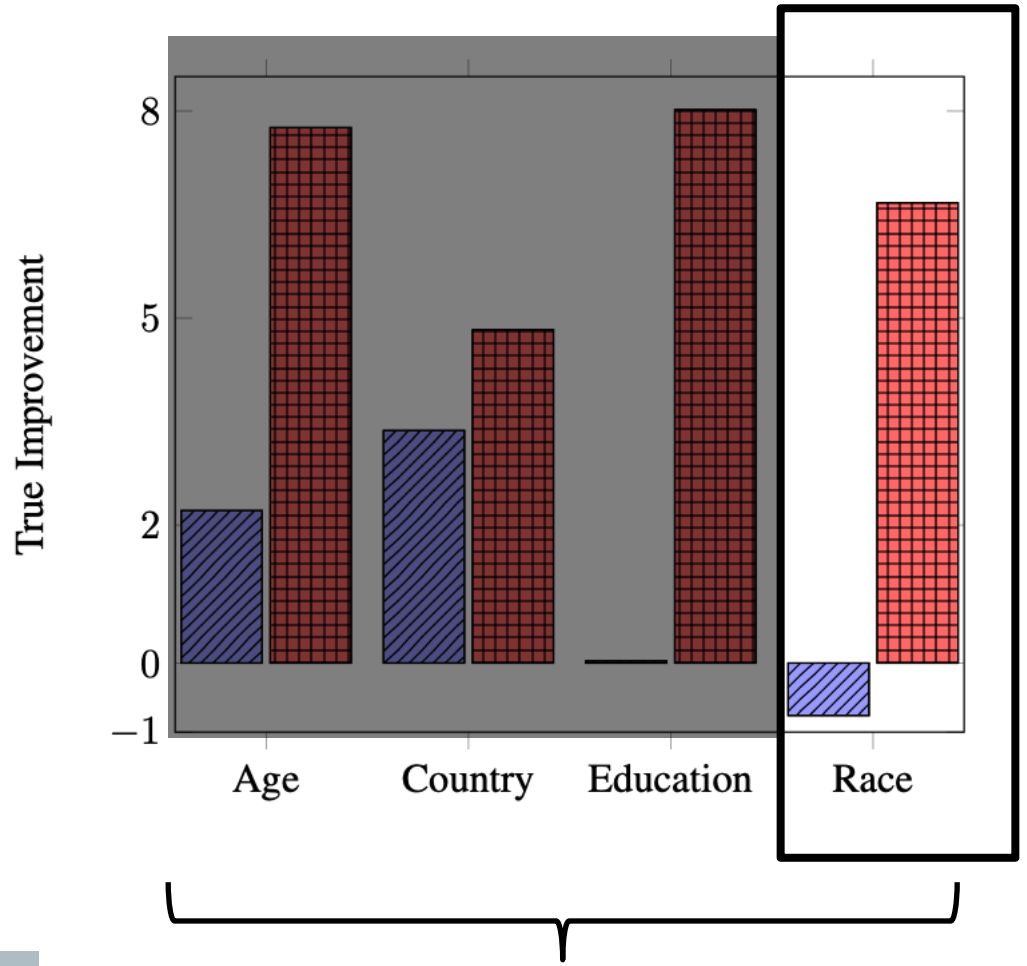


[Bechavod, P., Wu, Ziani, **ICML22]**

Results Snapshot: Adult Dataset

1 One subpopulation may get **worse off.**

- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2



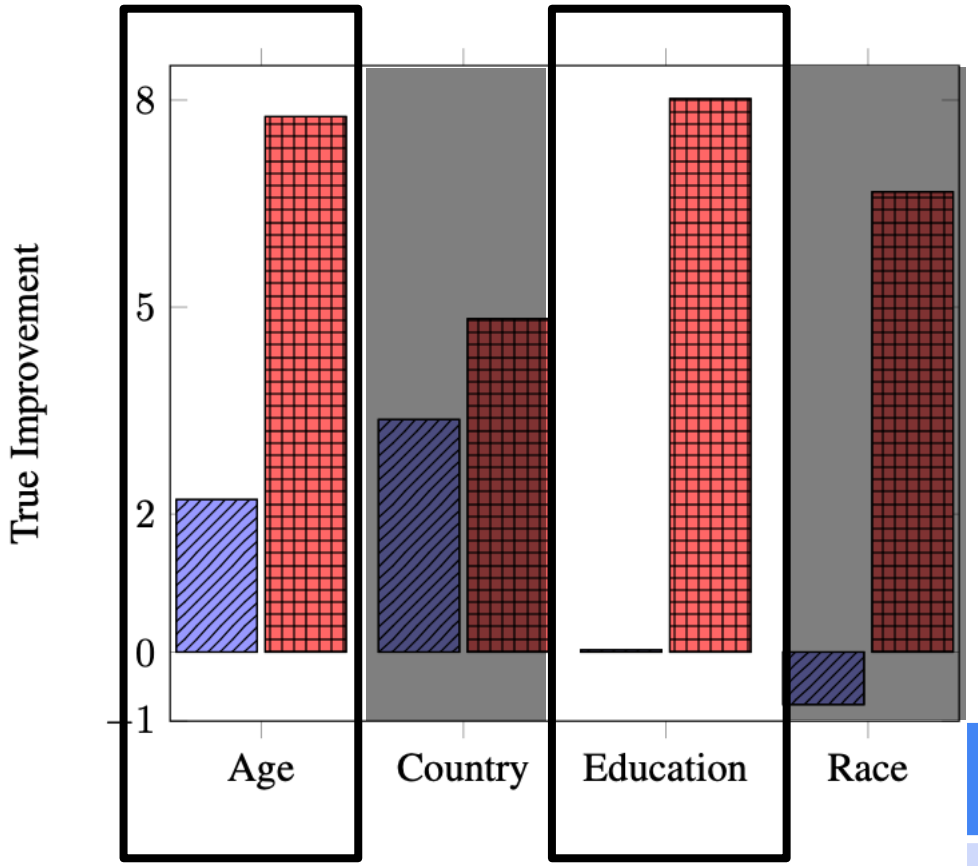
	Subpopulation 1	Subpopulation 2
Race	All others	White

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

Results Snapshot: Adult Dataset

2 Total improvement may be **very unequal** across subpopulations.

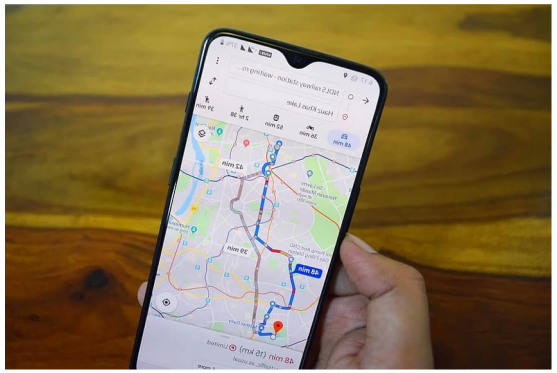
- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2



	Subpopulation 1	Subpopulation 2
Age	<35 yrs old	>=35 yrs old
Education	All others	Above HS

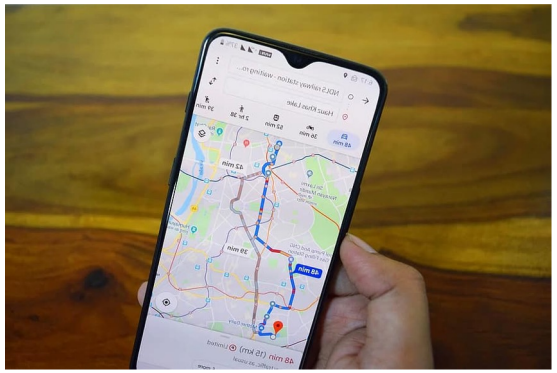
[Bechavod, P., Wu, Ziani, **ICML22]**

Performativity Beyond Just Strategizing



Performative Prediction: When **predictions** influence the **data**
(not just strategic prediction)

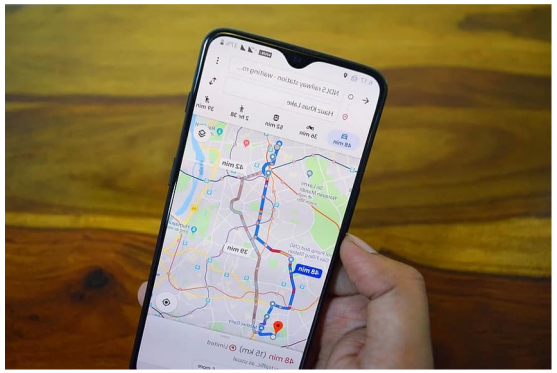
Performativity Beyond Just Strategizing



Performative Prediction: When **predictions** influence the **data**
(not just strategic prediction)

Main Results

Performativity Beyond Just Strategizing

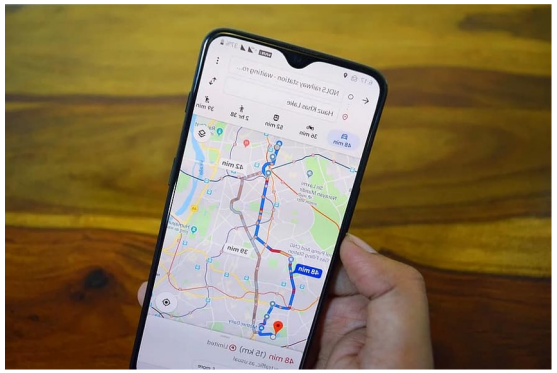


Performative Prediction: When **predictions** influence the **data**
(not just strategic prediction)

Main Results

1) When does repeated retraining lead to **stable** rules? [Perdomo, Zrnic, Mendler-Dünner, Hardt, **ICML20**]

Performativity Beyond Just Strategizing

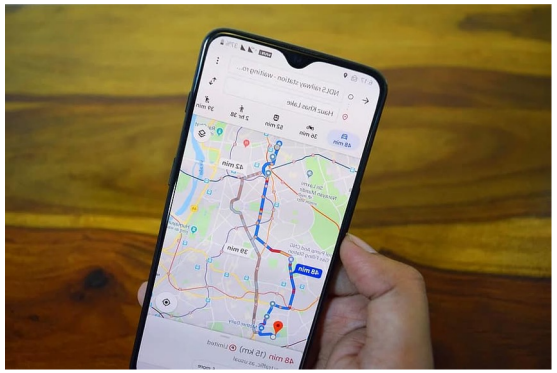


Performative Prediction: When **predictions** influence the **data**
(not just strategic prediction)

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**]

Performativity Beyond Just Strategizing

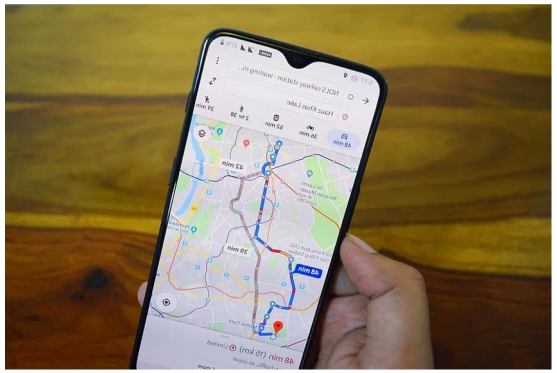


Performative Prediction: When **predictions** influence the **data**
(not just strategic prediction)

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**]

Performativity Beyond Just Strategizing



Performative Prediction: When **predictions** influence the **data**
(not just strategic prediction)

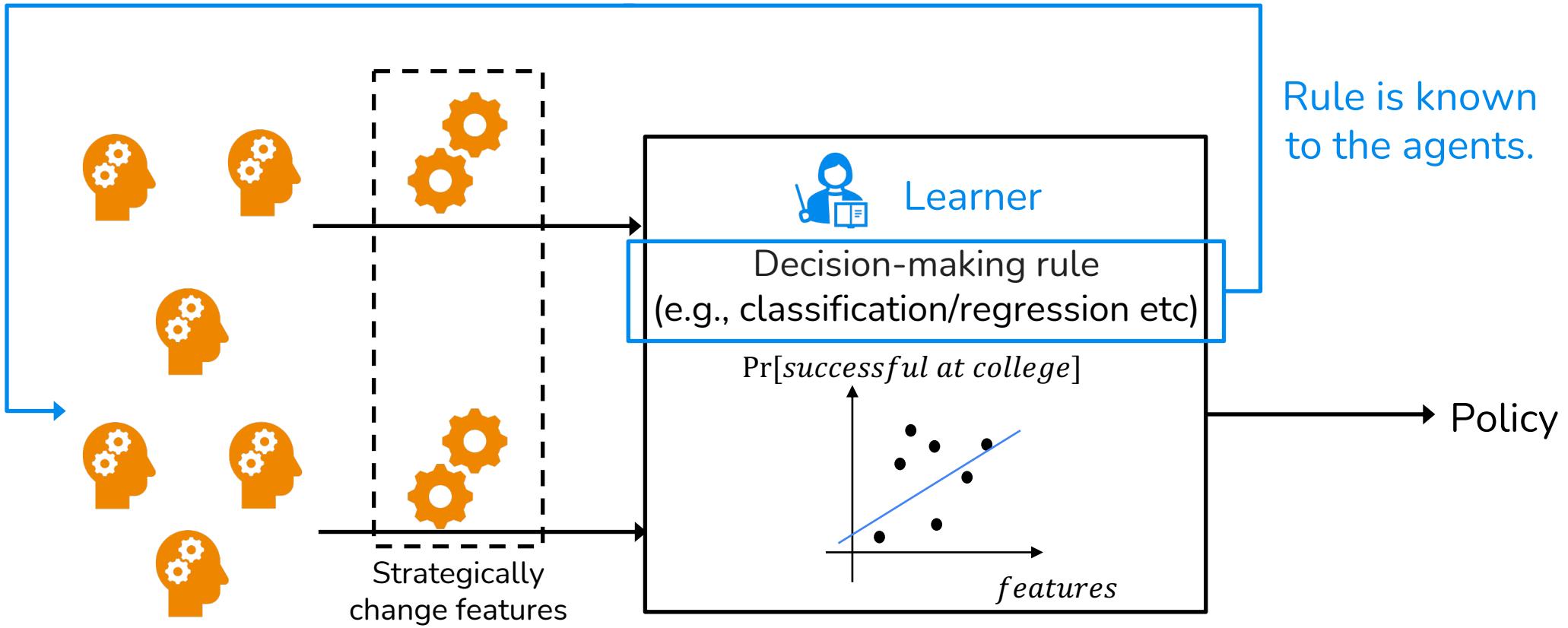
Main Results

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**]

Tutorial Outline

- ☑ Introduction
- ☑ Robustness
- ☑ Fairness
- ☑ Recourse/Performativity/Causality
- ❑ Future Directions/Open Questions

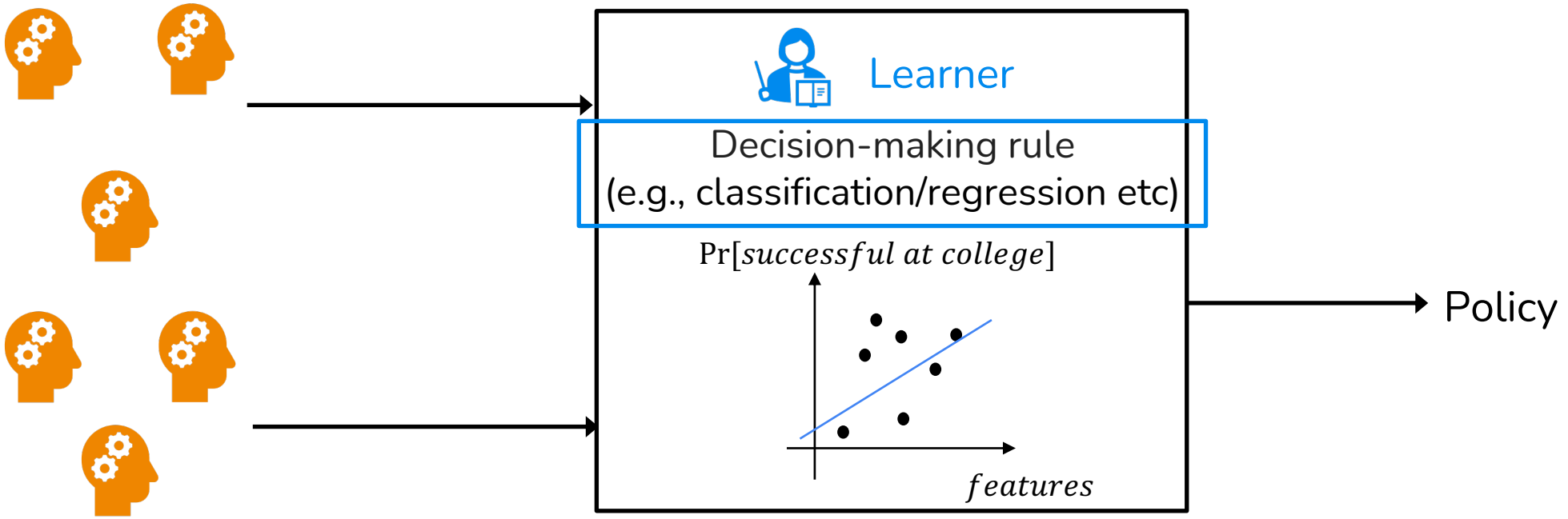
Theme 1: Interpretability and Incentives



Theme 1: Interpretability and Incentives

"Obscure" ML algorithms

+ Stop strategic behavior
- Non-transparent



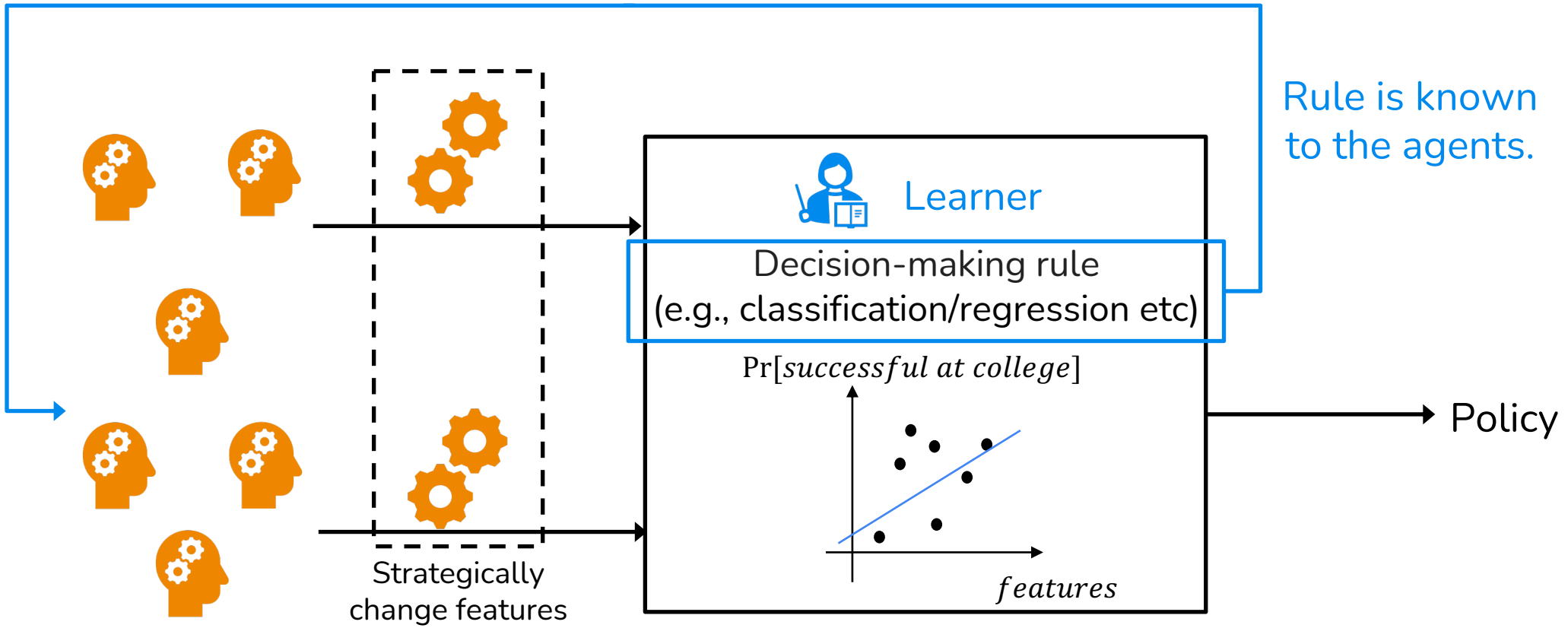
Theme 1: Interpretability and Incentives

"Obscure" ML algorithms

+ Stop strategic behavior
- Non-transparent

Public ML algorithms

+ Incentivize efforts for outcome improvement [KR, **EC19**]
- Prone to strategic behavior

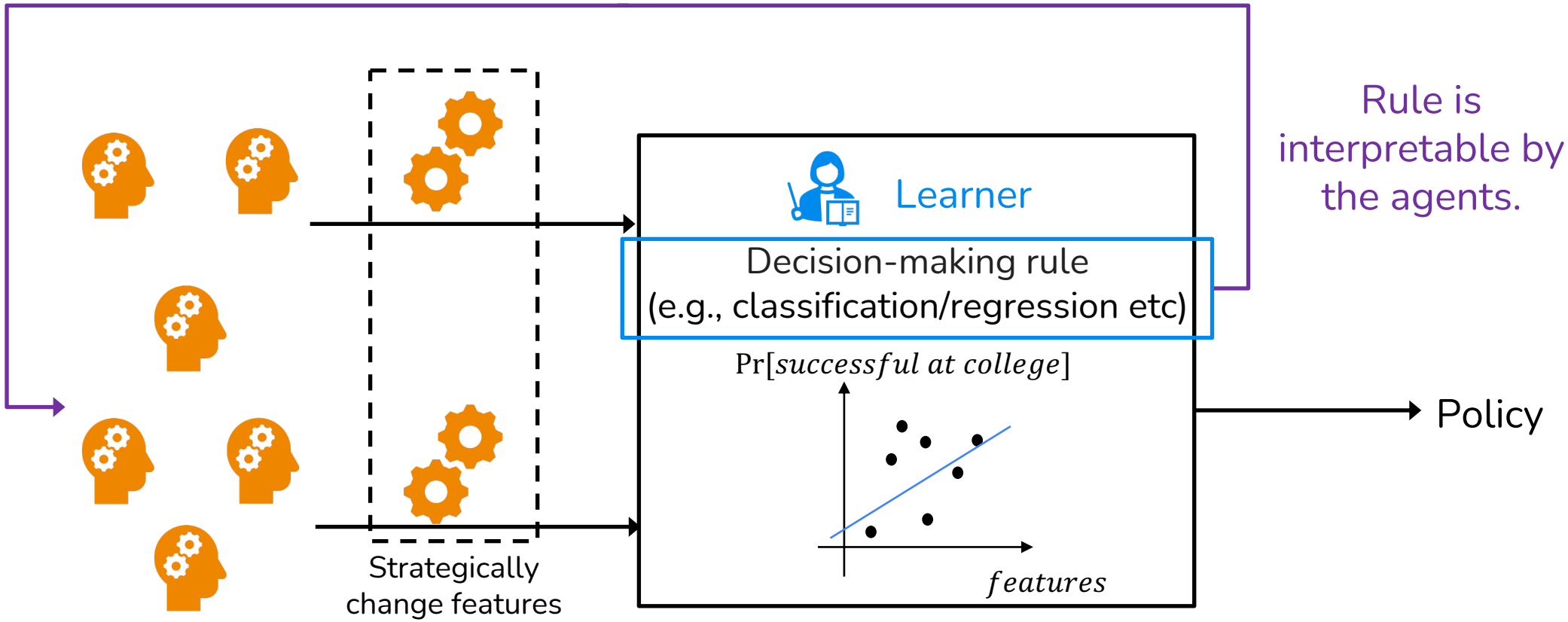


Theme 1: Interpretability and Incentives

"Obscure" ML algorithms
+ Stop strategic behavior
- Non-transparent

Interpretable ML Algorithms

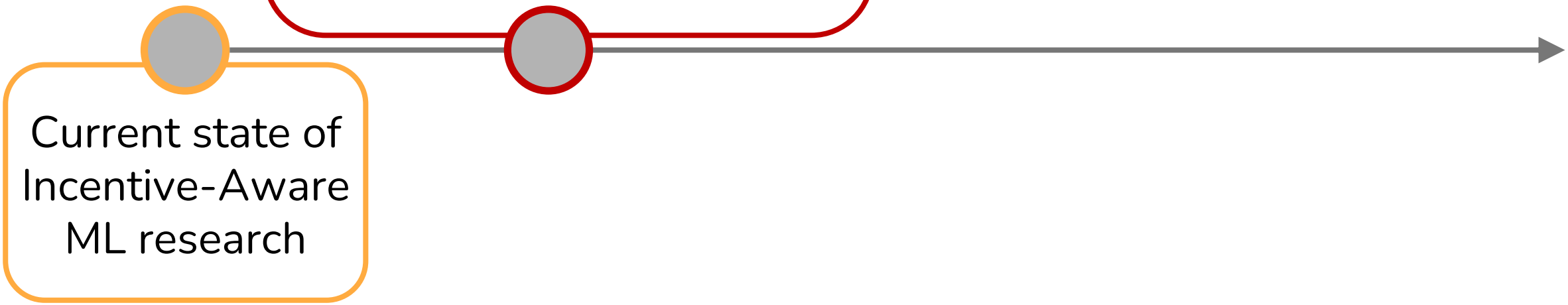
Public ML algorithms
+ Incentivize efforts for outcome improvement [KR, **EC19**]
- Prone to strategic behavior



Theme 1: Interpretability and Incentives

Theme 1: Interpretability and Incentives

- Learner: **Non-linear rules** (e.g., coming from neural nets).
- Agent: **understand** rules fully + **best-respond**



Current state of Incentive-Aware ML research

Theme 1: Interpretability and Incentives

- Learner: **Non-linear rules** (e.g., coming from neural nets).
- Agent: **understand** rules fully + **best-respond**

Current state of Incentive-Aware ML research

Large **case studies** to move from theory to practice and drive policy change.

Theme 1: Interpretability and Incentives

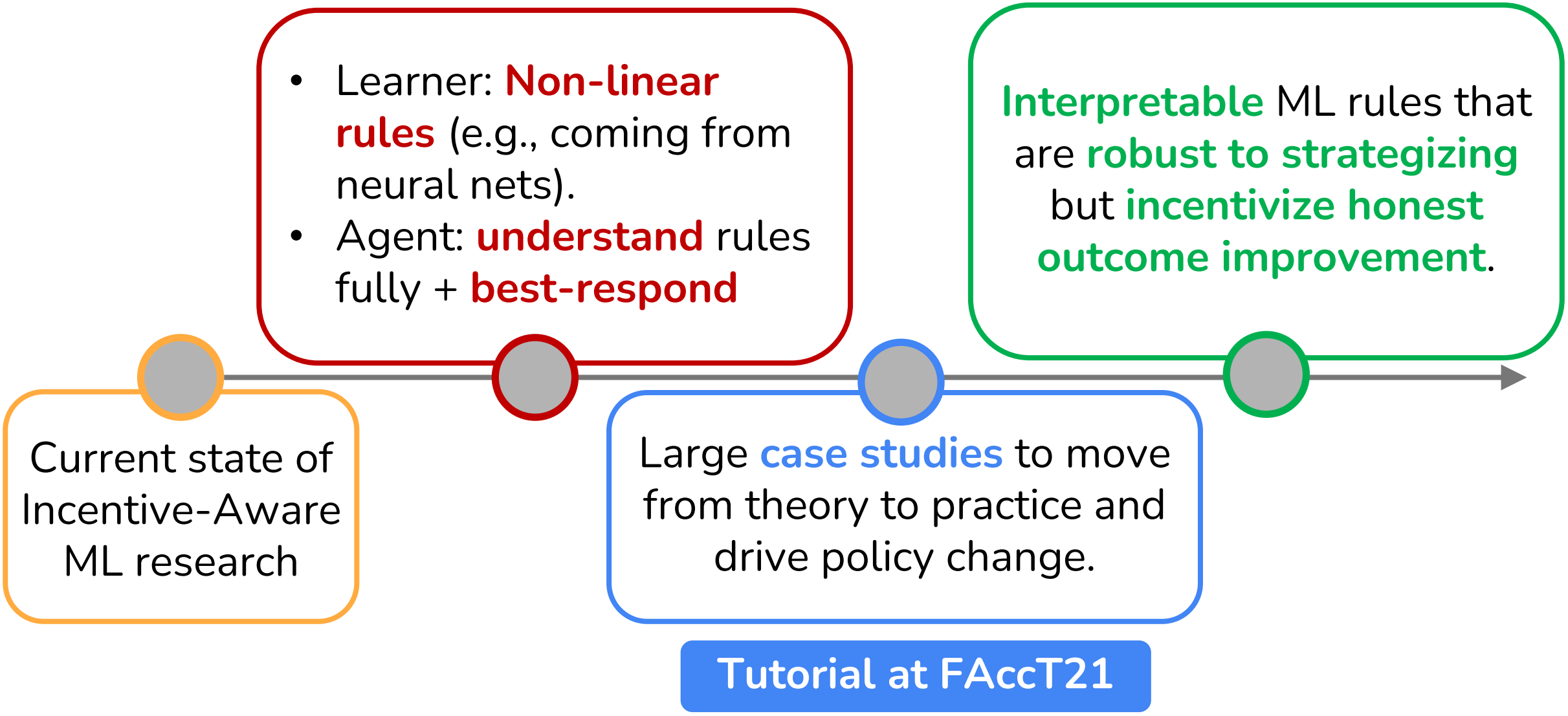
- Learner: **Non-linear rules** (e.g., coming from neural nets).
- Agent: **understand** rules fully + **best-respond**

Current state of Incentive-Aware ML research

Large **case studies** to move from theory to practice and drive policy change.

Tutorial at FAccT21

Theme 1: Interpretability and Incentives



Theme 2: Agent Behavior

Theme 2: Agent Behavior

Agent Behavior Assumptions

Theme 2: Agent Behavior

Extreme 1: Full Structure

Agent Behavior Assumptions

Most of theoretical works in incentive-aware ML:
myopically best-responding

Theme 2: Agent Behavior

Extreme 1: Full Structure

Most of theoretical works in incentive-aware ML: **myopically best-responding**

Agent Behavior Assumptions

Adversarial viewpoint: agents/adversaries want to "destroy" the algorithm.

Extreme 2: No Structure

Theme 2: Agent Behavior

Extreme 1: Full Structure

Agent Behavior Assumptions

Most of theoretical works in incentive-aware ML: **myopically best-responding**

Beyond Myopia and Best-Response

See also [Krishnamurthy, Lykouris, **P.**, Schapire, **STOC21**], [Paes Leme, **P.**, Schneider, **COLT22**]

Adversarial viewpoint: agents/adversaries want to "destroy" the algorithm.

Extreme 2: No Structure

Theme 2: Agent Behavior

Extreme 1: Full Structure

Agent Behavior Assumptions

Most of theoretical works in incentive-aware ML: **myopically best-responding**

Beyond Myopia and Best-Response

See also [Krishnamurthy, Lykouris, **P.**, Schapire, **STOC21**], [Paes Leme, **P.**, Schneider, **COLT22**]

Adversarial viewpoint: agents/adversaries want to "destroy" the algorithm.

Extreme 2: No Structure

Theme 2: Agent Behavior

Extreme 1: Full Structure

Agent Behavior Assumptions

Most of theoretical works in incentive-aware ML: **myopically best-responding**

Beyond Myopia and Best-Response

See also [Krishnamurthy, Lykouris, **P.**, Schapire, **STOC21**], [Paes Leme, **P.**, Schneider, **COLT22**]

Adversarial viewpoint: agents/adversaries want to "destroy" the algorithm.

Extreme 2: No Structure

Working at the population, rather than individual level e.g., [Jagadeesan, Mendler-Dünner, Hardt, **ICML21**]

Thank You!