Statistical Collusion by Collectives on Learning Platforms

Etienne Gauthier, Francis Bach, Michael I. Jordan
(INRIA, Ecole Normale Supérieure)

Numerous examples of collectives emerging to strategically influence platforms

BUSINESS INSIDER

Uber

Uber drivers are reportedly colluding to trigger 'surge' prices because they say the company is not paying them enough

By Isobel Asher Hamilton

- Uber drivers deactivate the app to create a supply shortage and drive up prices

Numerous examples of collectives emerging to strategically influence platforms

BUSINESS INSIDER

Uber drivers are reportedly colluding to trigger 'surge' prices because they say the company is not paying them enough

By Isobel Asher Hamilton

- Uber drivers deactivate the app to create a supply shortage and drive up prices

How merchants use Facebook to flood Amazon with fake reviews

April 23, 2018 More than 7 years ago

- Amazon users coordinate to post fake reviews, manipulating ratings and search rankings

Numerous examples of collectives emerging to strategically influence platforms

BUSINESS INSIDER

Uber drivers are reportedly colluding to trigger 'surge' prices because they say the company is not paying them enough

By Isobel Asher Hamilton

- Uber drivers deactivate the app to create a supply shortage and drive up prices

How merchants use Facebook to flood Amazon with fake reviews

April 23, 2018 More than 7 years ago

- Amazon users coordinate to post fake reviews, manipulating ratings and search rankings

Numerous examples of collectives emerging to strategically influence platforms

The Geotagging Counterpublic: The Case of Facebook Remote Check-Ins to Standing Rock

- Facebook users relocalized themselves to Standing Rock to disrupt surveillance and blur police tracking

BUSINESS INSIDER

Uber drivers are reportedly colluding to trigger 'surge' prices because they say the company is not paying them enough

By Isobel Asher Hamilton

- Uber drivers deactivate the app to create a supply shortage and drive up prices

How merchants use Facebook to flood Amazon with fake reviews

April 23, 2018 More than 7 years ago

- Amazon users coordinate to post fake reviews, manipulating ratings and search rankings

Numerous examples of collectives emerging to strategically influence platforms

The Geotagging Counterpublic:
The Case of Facebook Remote Check-Ins to Standing Rock

- Facebook users relocalized themselves to Standing Rock to disrupt surveillance and blur police tracking

Home

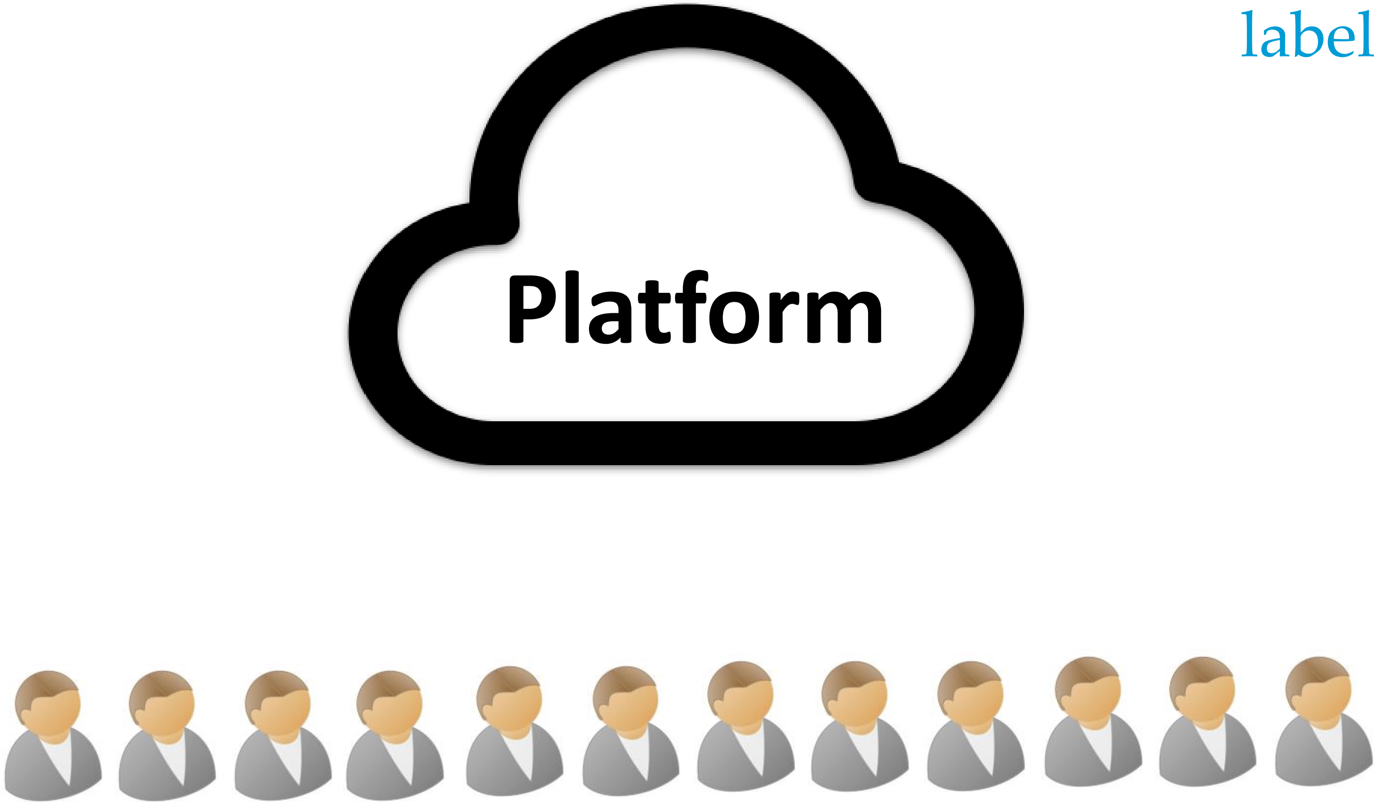
How Neighborhoods Are Fighting Off Traffic That Waze Sends Their Way

When Waze or Google Maps turns your sleepy street into a veritable highway, you don't just have to sit there and take it.

- Waze users falsely report accidents to keep traffic out of their neighborhoods

Model

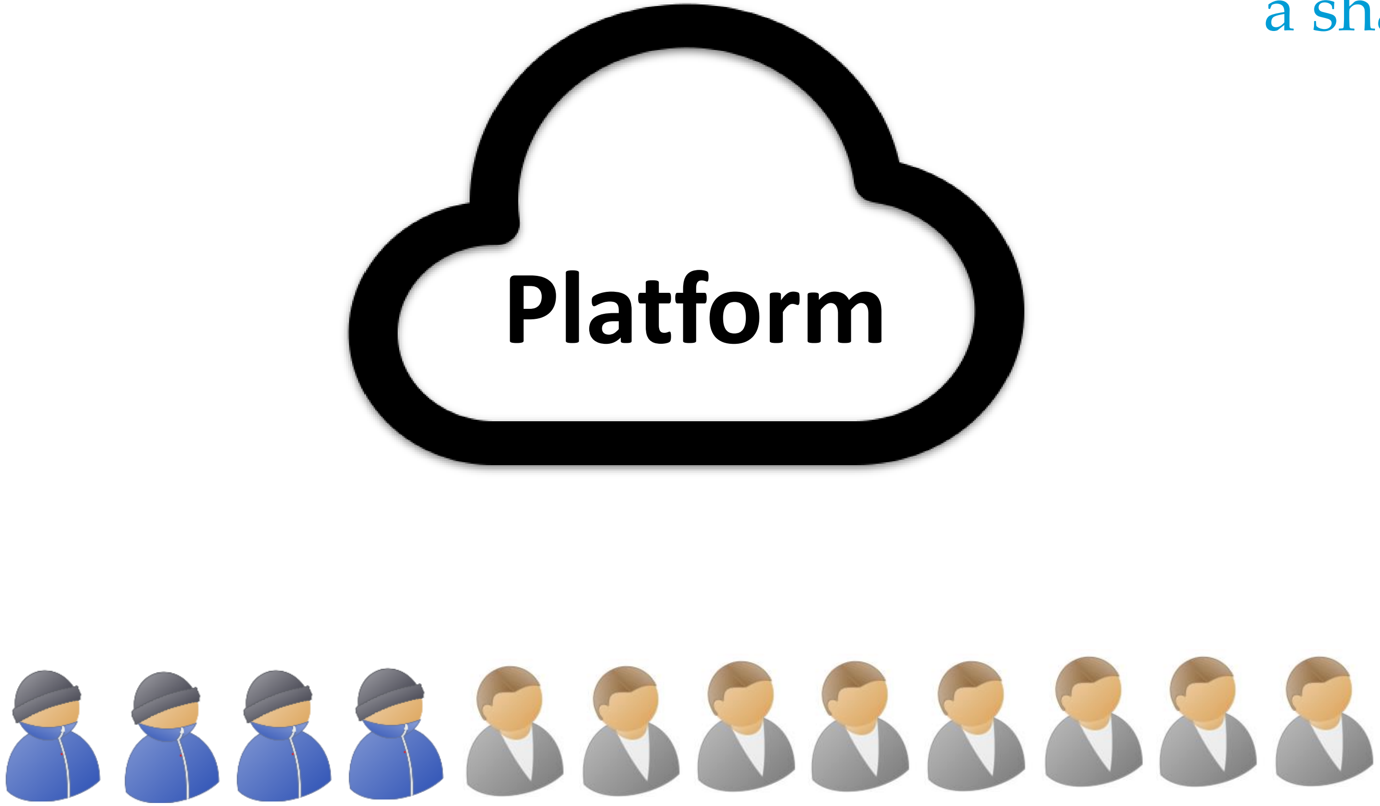
Initially, each user is drawn from the same probability distribution $\mathcal{D}$ over feature-label pairs X×Y



N consumers $\overset{\text{i.i.d.}}{\sim} \mathcal{D}$

Model

A collective forms to influence the platform's behavior toward a shared goal



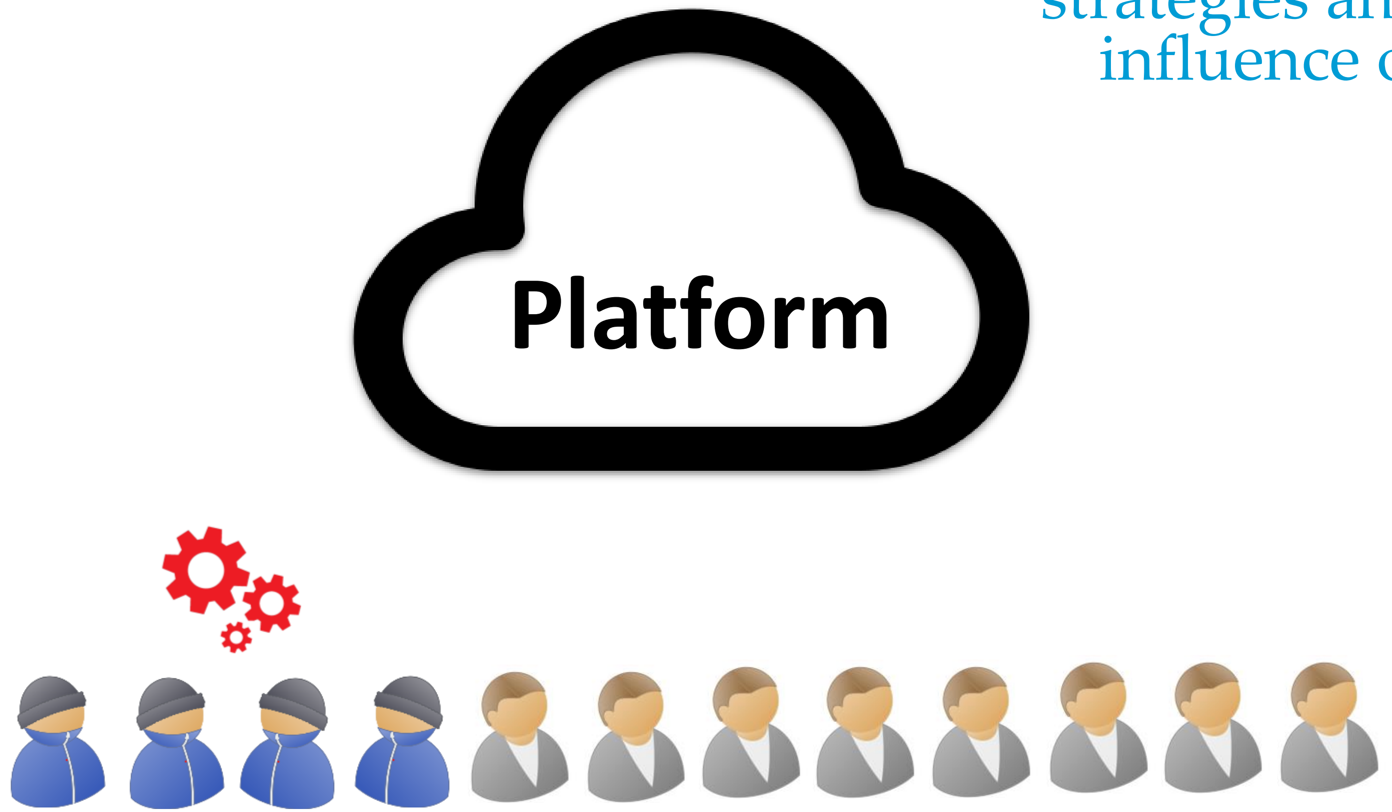
N consumers $\overset{\text{i.i.d.}}{\sim} \mathcal{D}$

Collective
(size n)

Rest of the population
(size $N - n$)

Model

Collective members share their data to identify effective strategies and anticipate their influence on the platform



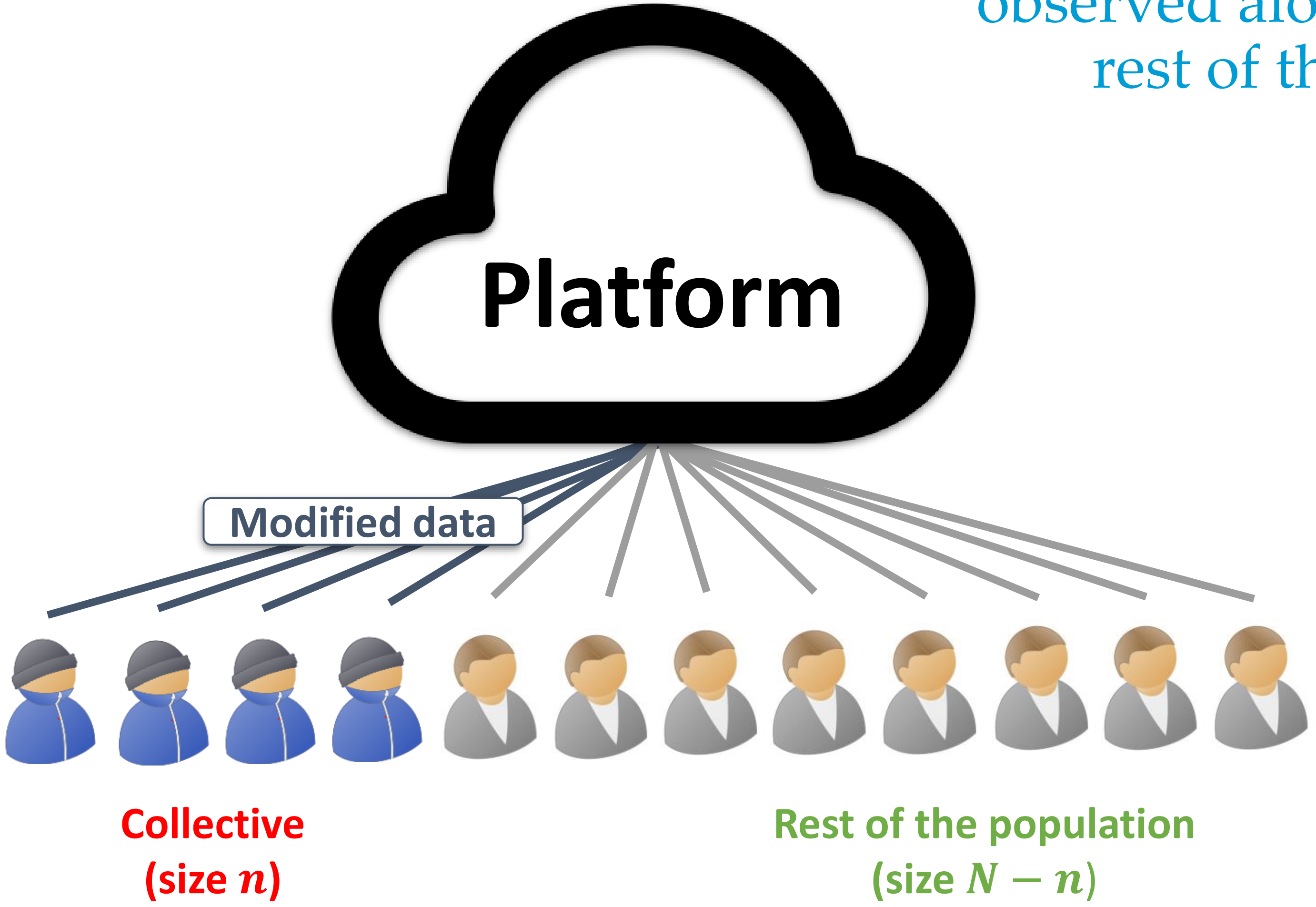
N consumers $\overset{\text{i.i.d.}}{\sim} \mathcal{D}$

Collective
(size n)

Rest of the population
(size $N-n$)

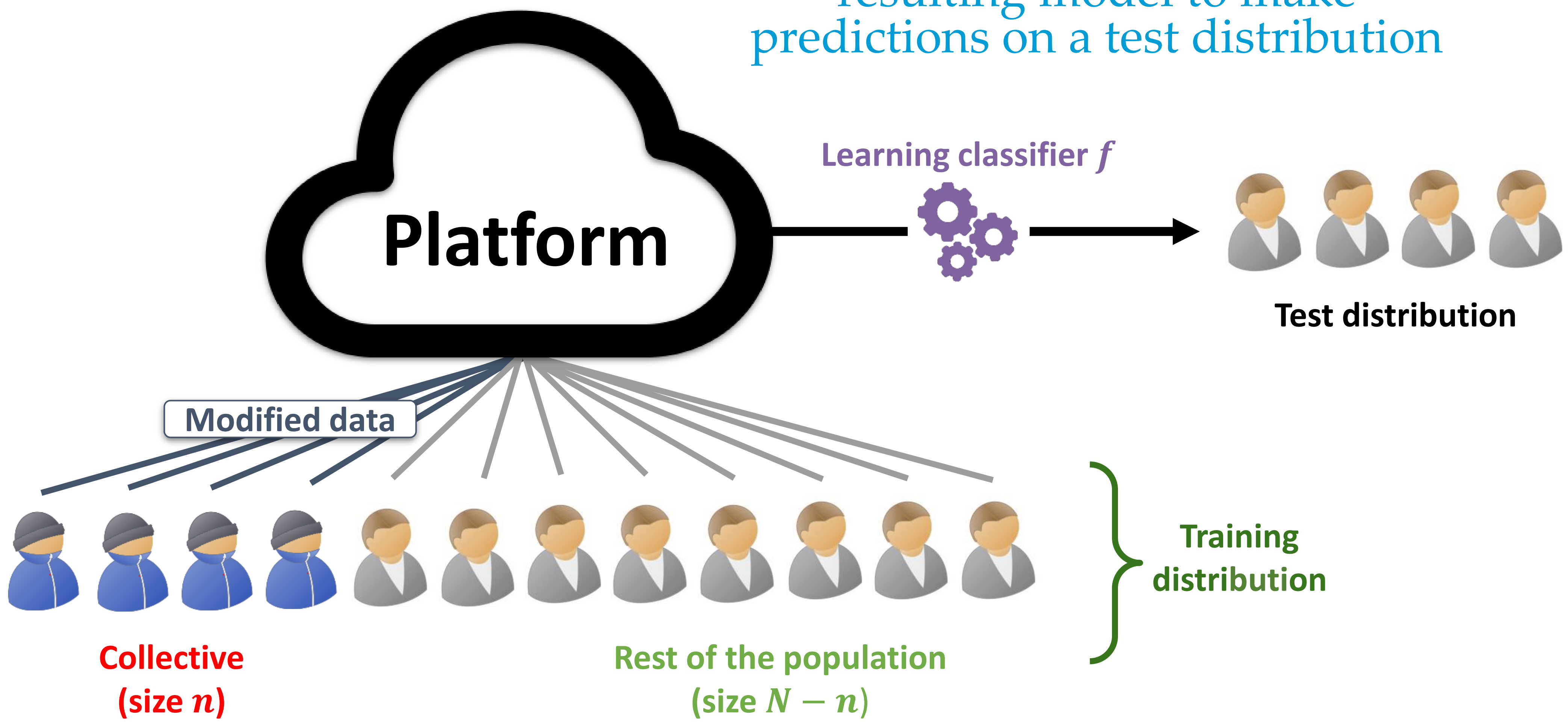
Model

Collective members modify their data, which is then observed alongside that of the rest of the population



Model

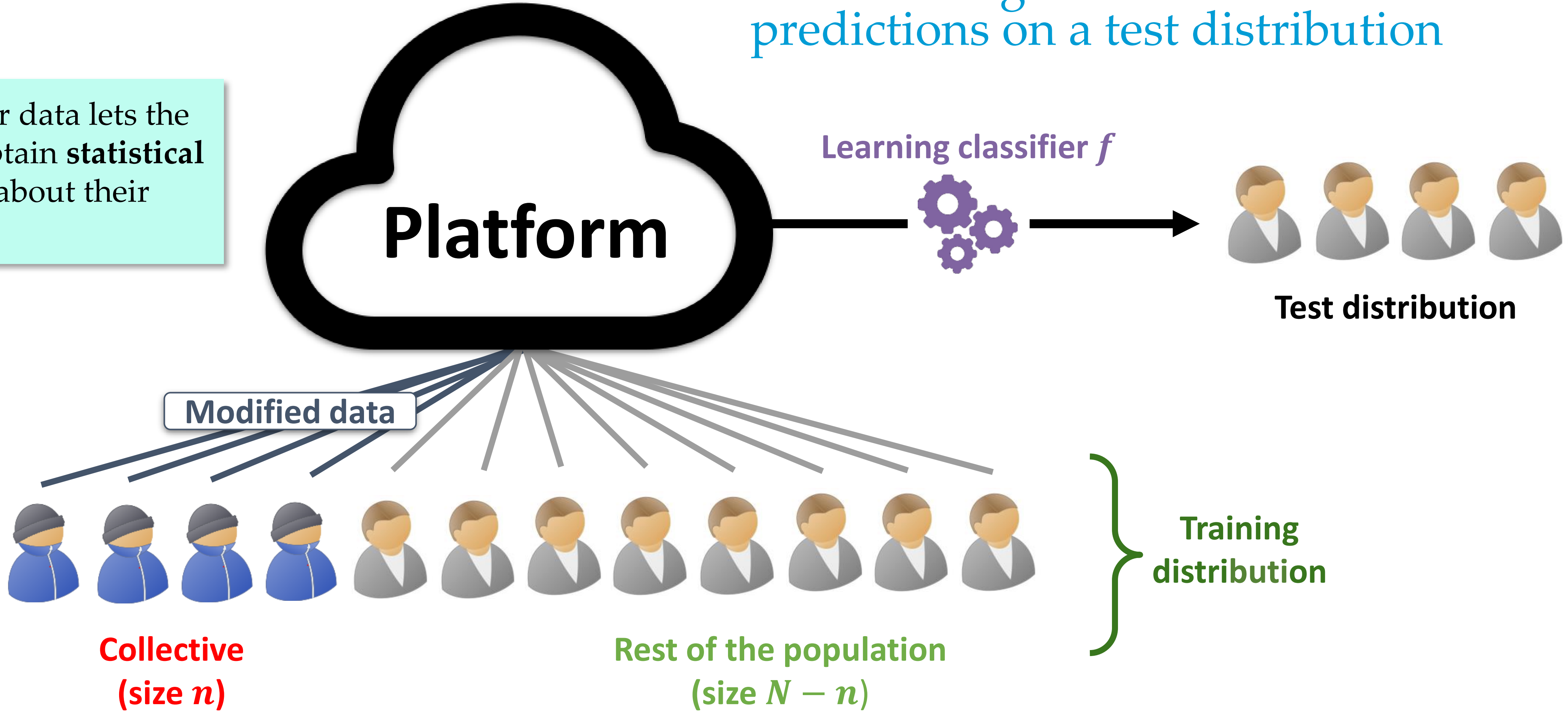
The platform learns from the training data and uses the resulting model to make predictions on a test distribution



Model

The platform learns from the training data and uses the resulting model to make predictions on a test distribution

- Pooling their data lets the collective obtain **statistical guarantees** about their impact



Platform's behavior

$$f(x) = \underset{y}{\operatorname{argmax}}\ \mathcal{P}(y|x)$$

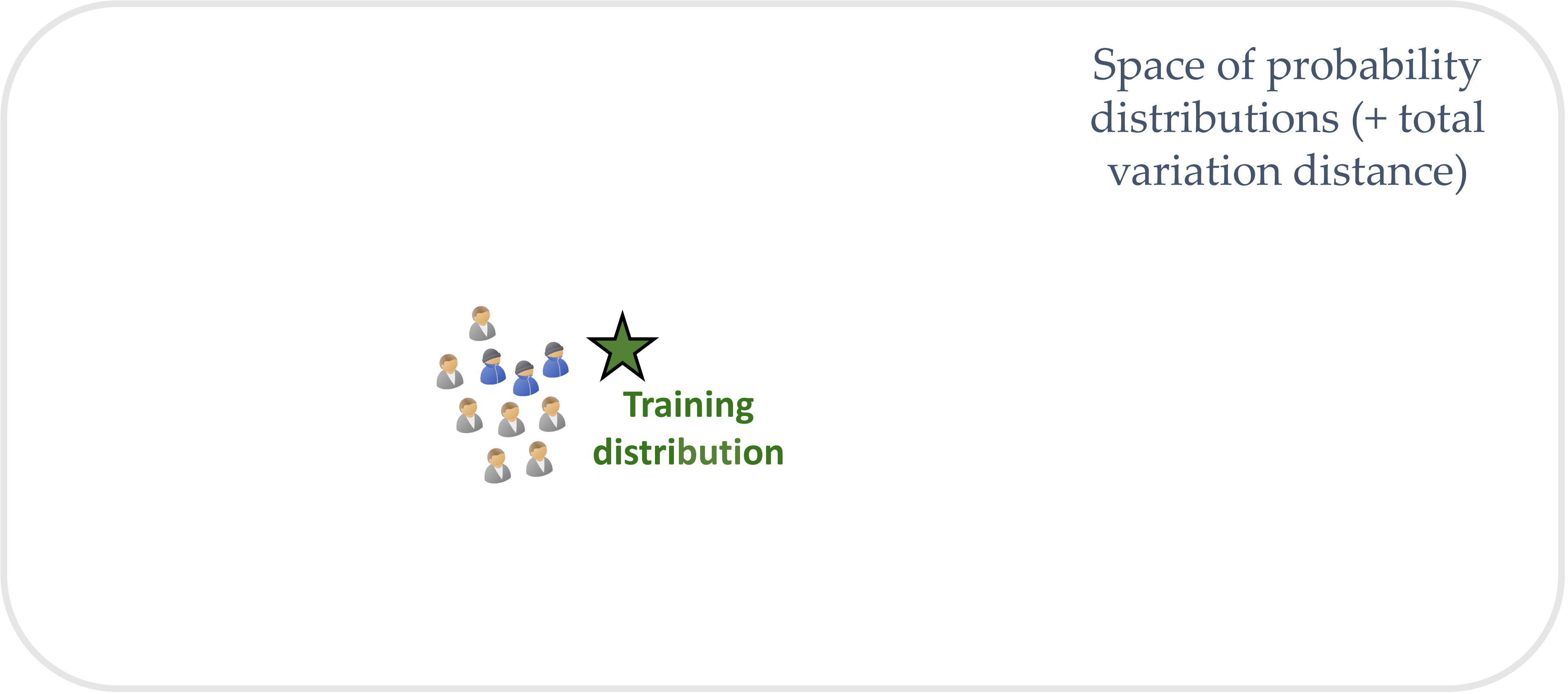
[Hardt et al., 2023]

Space of probability distributions (+ total variation distance)

Platform's behavior

$$f(x) = \underset{y}{\operatorname{argmax}}\ \mathcal{P}(y|x)$$

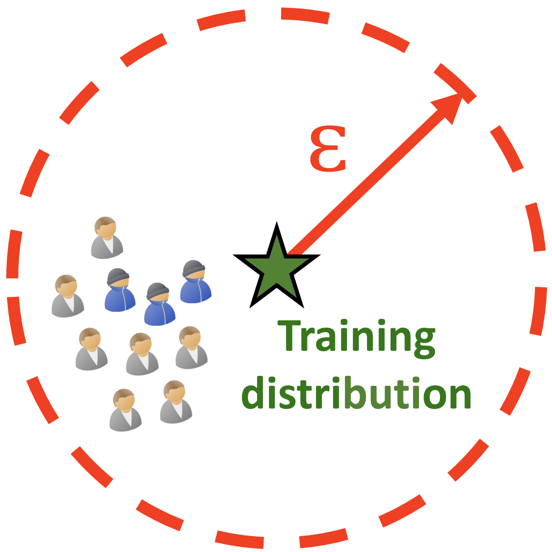
[Hardt et al., 2023]



Platform's behavior

$$f(x) = \operatorname*{argmax}_{y} \mathcal{P}(y|x)$$

[Hardt et al., 2023]

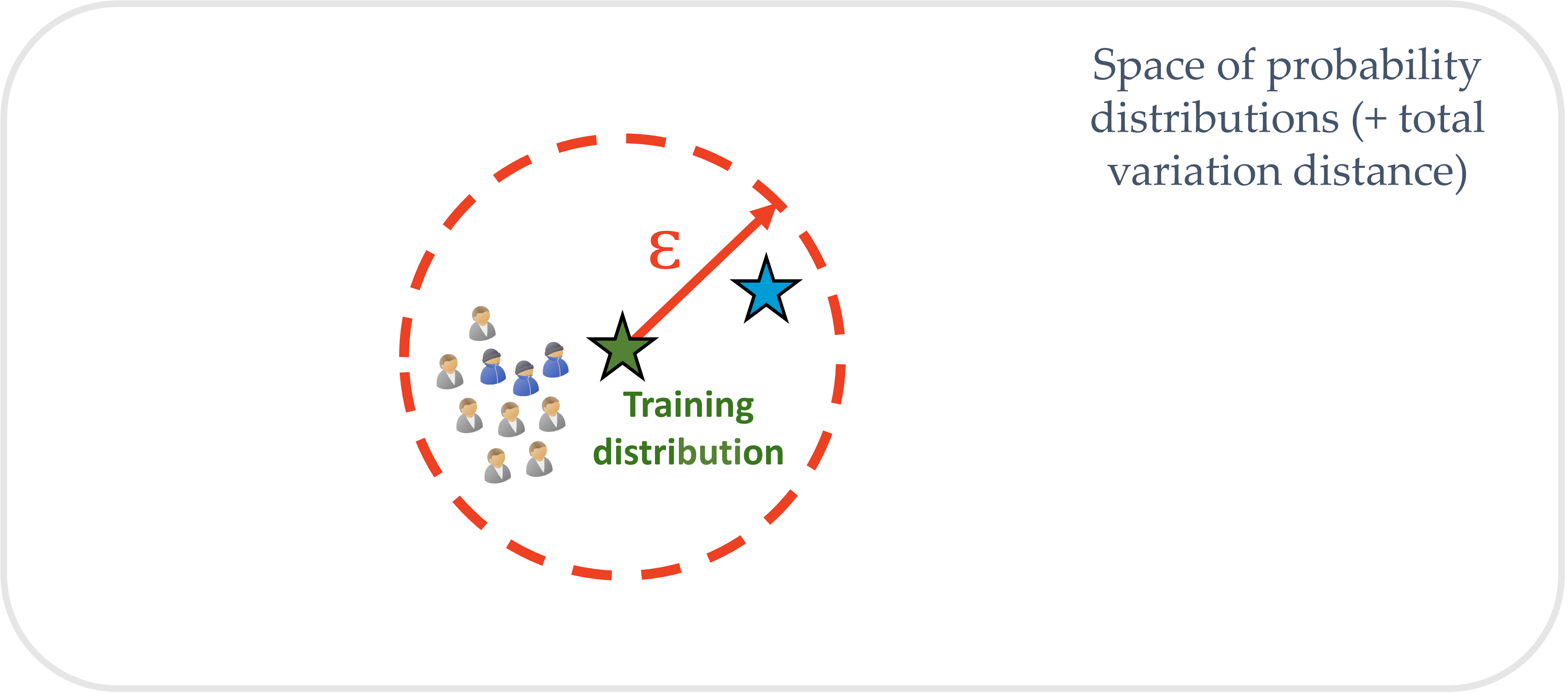


Space of probability distributions (+ total variation distance)

Platform's behavior

$$f(x) = \underset{y}{\operatorname{argmax}}\ \mathcal{P}(y|x)$$

[Hardt et al., 2023]



Collective's objective

Maximize a measure of success $S(n)$

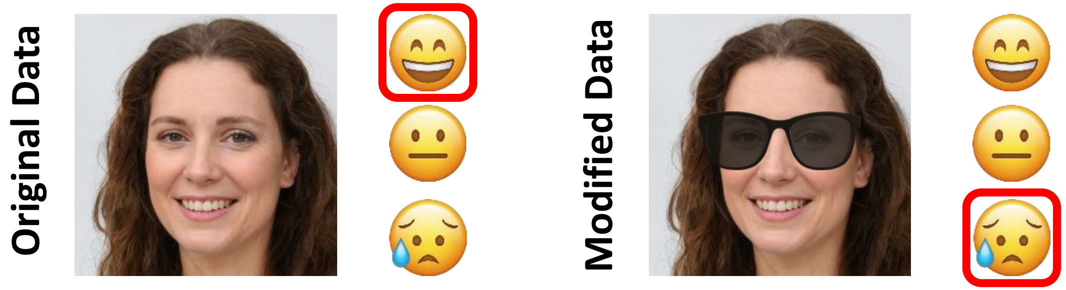
Collective's objective

Maximize a measure of success $S(n)$

- **Signal planting:**
  - Given $g\colon X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$

Collective's objective

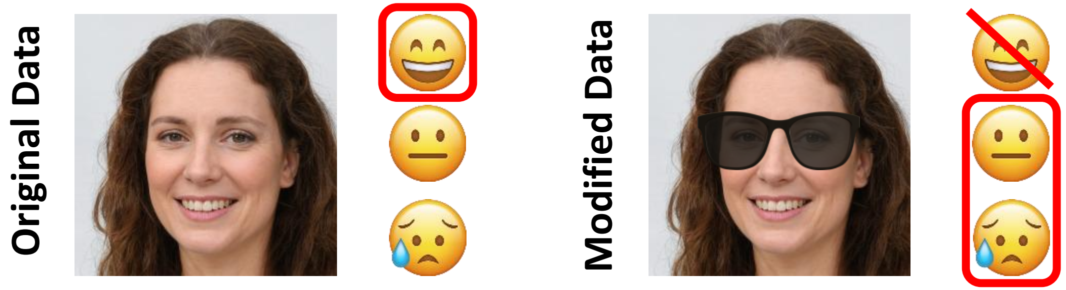
Maximize a measure of success $S(n)$



- **Signal planting:**
  - Given $g\colon X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$

Collective's objective

Maximize a measure of success $S(n)$

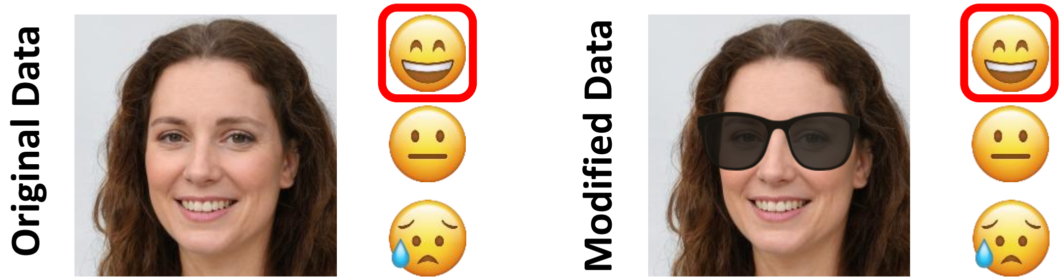


- **Signal planting:**
  - Given $g: X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$

- **Signal unplanting:**
  - Given $g: X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) \neq y^*\big)$

Collective’s objective

Maximize a measure of success $S(n)$



- **Signal planting:**
  - Given $g\colon X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$

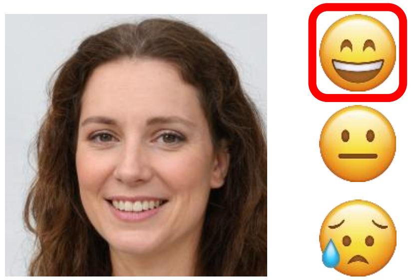
- **Signal unplanting:**
  - Given $g\colon X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) \neq y^*\big)$

- **Signal erasing:**
  - Given $g\colon X \to X$, $\mathbb{P}\big(f\big(g(x)\big) = f(x)\big)$

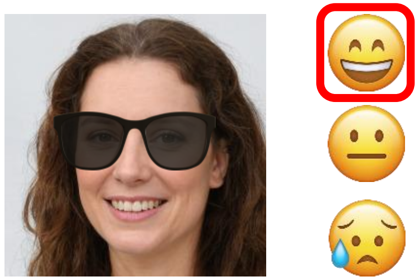
Collective's objective

Maximize a measure of success $S(n)$

Original Data



Modified Data



- **Signal planting:**
  - Given $g\colon X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$

- **Signal unplanting:**
  - Given $g\colon X \to X$ and $y^* \in Y$, $S(n) = \mathbb{P}\big(f\big(g(x)\big) \neq y^*\big)$

- **Signal erasing:**
  - Given $g\colon X \to X$, $\mathbb{P}\big(f\big(g(x)\big) = f(x)\big)$

> **Results:** for each objective, we analyze **strategies** that the collective can set and we derive **strategy-dependent high-probability lower bounds** on $S(n)$

Example: signal planting

$$S(n) = \mathbb{P}\left(f\left(g(x)\right) = y^*\right)$$

Example: signal planting

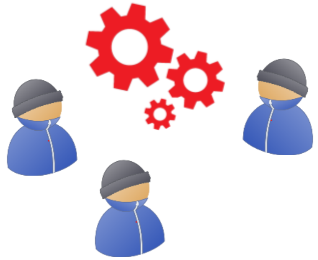
$$S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$$

- **A natural strategy:**

Example: signal planting

$$S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$$

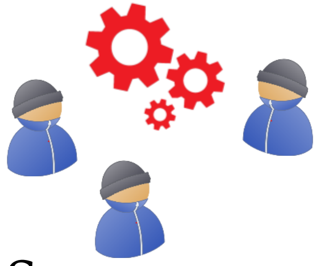
- **A natural strategy:**
  - The collective can change its data as follows: $(x, y) \rightarrow (g(x), y^*)$



Example: signal planting

$$S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$$

- **A natural strategy:**
  - The collective can change its data as follows: $(x, y) \rightarrow (g(x), y^*)$



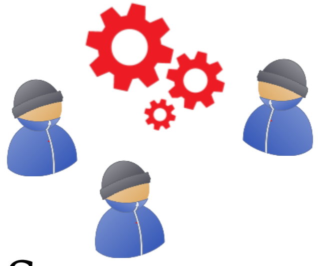
- **Results:** with high probability, up to $1/\sqrt{n}$ estimation terms,

$$S(n) \geq \widehat{\mathbb{P}}_{\mathit{Collective\ data}}[\text{Prevalence} - \text{Counteracting Influence} - \text{Robustness} > 0]$$

Example: signal planting

$$S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$$

- **A natural strategy:**
  - The collective can change its data as follows: $(x, y) \rightarrow (g(x), y^*)$



- **Results:** with high probability, up to $1/\sqrt{n}$ estimation terms,

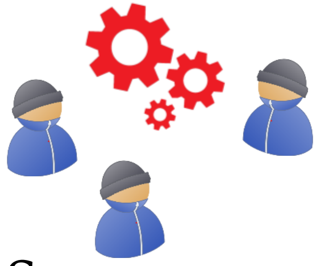
$$S(n) \geq \hat{\mathbb{P}}_{Collective\ data}[\text{Prevalence} - \text{Counteracting Influence} - \text{Robustness} > 0]$$

Indicates the prevalence of the modified feature in the modified dataset ($\sim n/N$)

Example: signal planting

$$S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$$

- **A natural strategy:**
  - The collective can change its data as follows: $(x, y) \rightarrow (g(x), y^*)$



- **Results:** with high probability, up to $1/\sqrt{n}$ estimation terms,

$$S(n) \geq \underset{\textit{Collective data}}{\widehat{\mathbb{P}}}[\textcolor{red}{\text{Prevalence}} - \textcolor{green}{\text{Counteracting Influence}} - \text{Robustness} > 0]$$

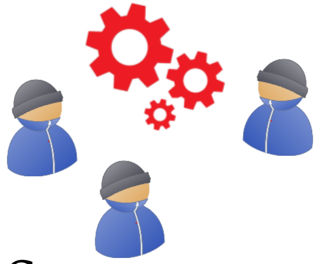
Indicates the prevalence of the modified feature in the modified dataset ($\sim n/N$)

Captures how non-collective individuals hinder the collective's impact ($\sim 1 - n/N$)

Example: signal planting

$$S(n) = \mathbb{P}\big(f\big(g(x)\big) = y^*\big)$$

- **A natural strategy:**
  - The collective can change its data as follows: $(x, y) \to (g(x), y^*)$



- **Results:** with high probability, up to $1/\sqrt{n}$ estimation terms,

$$S(n) \geq \underset{Collective\ data}{\hat{\mathbb{P}}}\left[\text{Prevalence} - \text{Counteracting Influence} - \text{Robustness} > 0\right]$$

Indicates the prevalence of the modified feature in the modified dataset ($\sim n/N$)

Captures how non-collective individuals hinder the collective's impact ($\sim 1 - n/N$)

Platform robustness ($\nearrow \varepsilon$)

Experiments

Synthetic Dataset: Car Evaluation

- **Features:** 18 attributes including *Model Type*, *Fuel Type*, *Country of Manufacture*, etc.
- **Labels (4 classes): Excellent, Good, Average, Poor**

Experiments

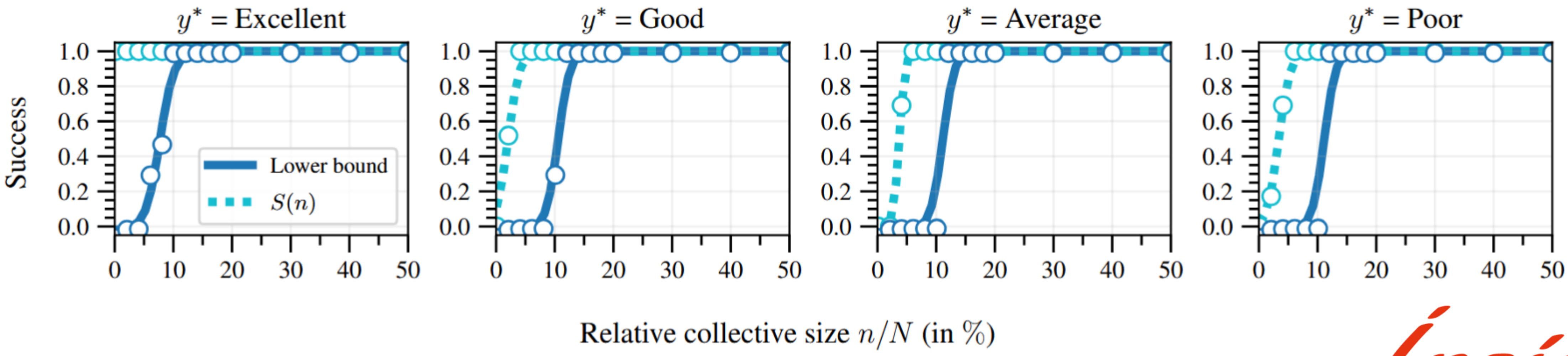
Synthetic Dataset: Car Evaluation

- **Features:** 18 attributes including *Model Type, Fuel Type, Country of Manufacture*, etc.
- **Labels (4 classes): Excellent, Good, Average, Poor**
- **Experiment Setup:** Apply transformation $g: X \to X$ targeting specific types of vehicles
- **Collective Influence:** A "lobby" group advocating *for* or *against* these specific vehicles

Experiments

Synthetic Dataset: Car Evaluation

- **Features:** 18 attributes including *Model Type, Fuel Type, Country of Manufacture*, etc.
- **Labels (4 classes): Excellent, Good, Average, Poor**
- **Experiment Setup:** Apply transformation $g: X \rightarrow X$ targeting specific types of vehicles
- **Collective Influence:** A "lobby" group advocating *for* or *against* these specific vehicles



Beyond This Talk: What's in the Paper

- **General Framework:** formal modelization
- **Different Objectives:** signal planting, unplanting, and erasing
- **More Strategies:** feature-label vs feature-only, adaptive vs static
- **Theory:** explicit lower bounds, algorithmic implementations
- **Parameters Influence:** how impact varies with collective size n and number of consumers N
  - platforms interacting with large user bases are more exposed to collectives altering their data

Conclusion

- By **sharing their data**, collectives can **infer** and put into practice impactful strategies
- Our approach enables collectives to **anticipate their potential impact** on learning platforms
- Opens new directions for understanding **multi-agent influence** on learning platforms

Thank you! Questions?

Paper

Poster:

East Exhibition Hall

A-B #E-700

4:30 to 7:00 PM