

Logistics: Project

- Project presentation
 - Dec 15/17
 - See Piazza on the presentation order
 - Prepare a 12~13 minutes presentation. Might impose a hard cut-off in 13 minutes if time is an issue.
- Project reports
 - Due: Dec 20
 - Up to 6 pages (with additional pages containing only references/citations)

Reminder on The Grades

- Homework assignments / reviews / presentations are lightly graded
- Condition on you complete all other requirements satisfactorily, your final grades are determined by your final project. Your final grades will be
 - A+: Your project is close to be published in top venues
 - A: I'll be proud to use your project as model projects in the future
 - A-: Overall good, maybe with minor issues
 - B+ or lower: There are obvious flaws in the project (e.g., poorly motivated problems, etc)
- Your final grades will be decreased from the above for missing reviews / homework issues / non-participation using the grading scheme (following the standard mapping)

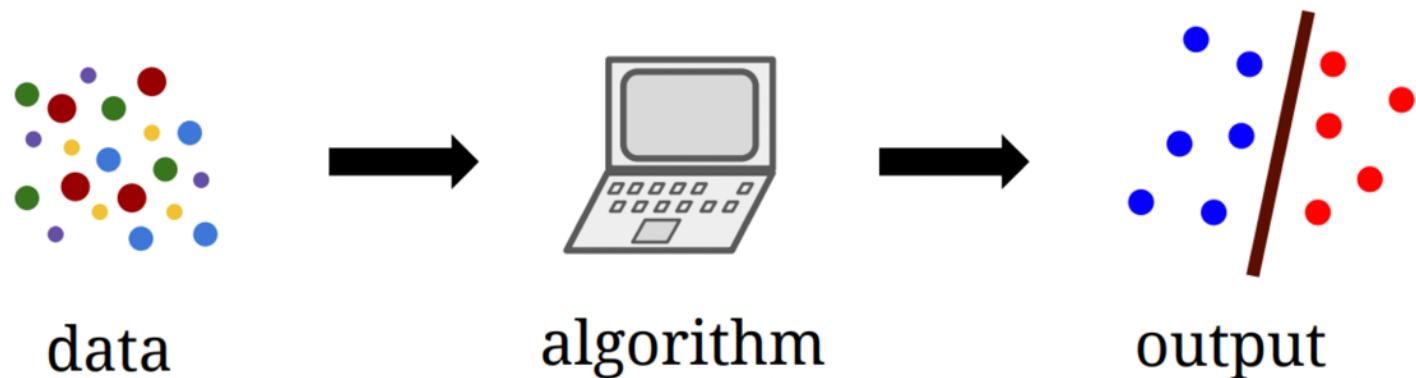
Lecture 22

Strategic Classification

Instructor: Chien-Ju (CJ) Ho

Machine Learning

- Standard setup of (supervised) machine learning



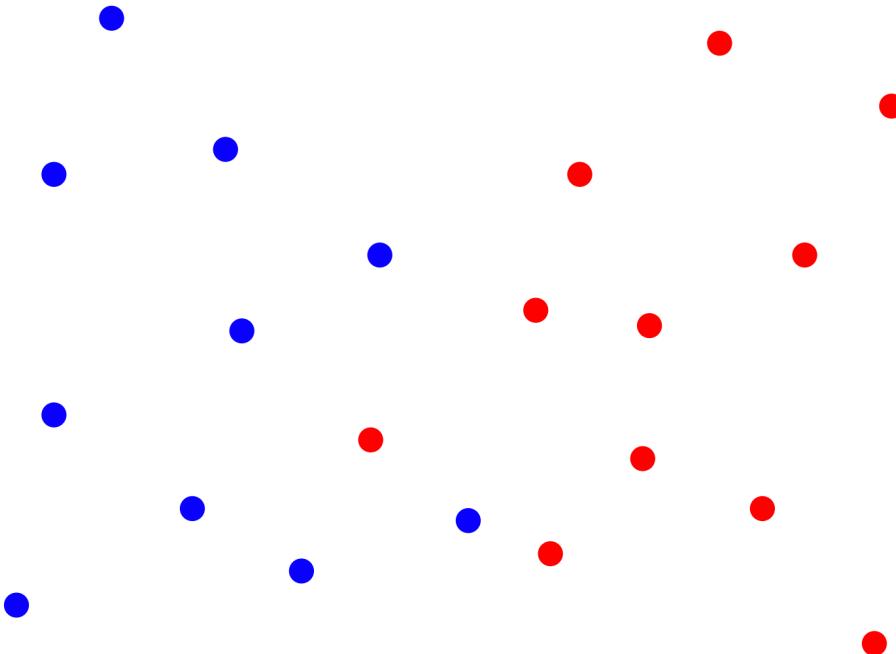
- Finding patterns from the given training datasets
- Use the pattern to make predictions on new testing data

- Fundamental assumption:
 - Training and testing data points are i.i.d. drawn from the same distribution

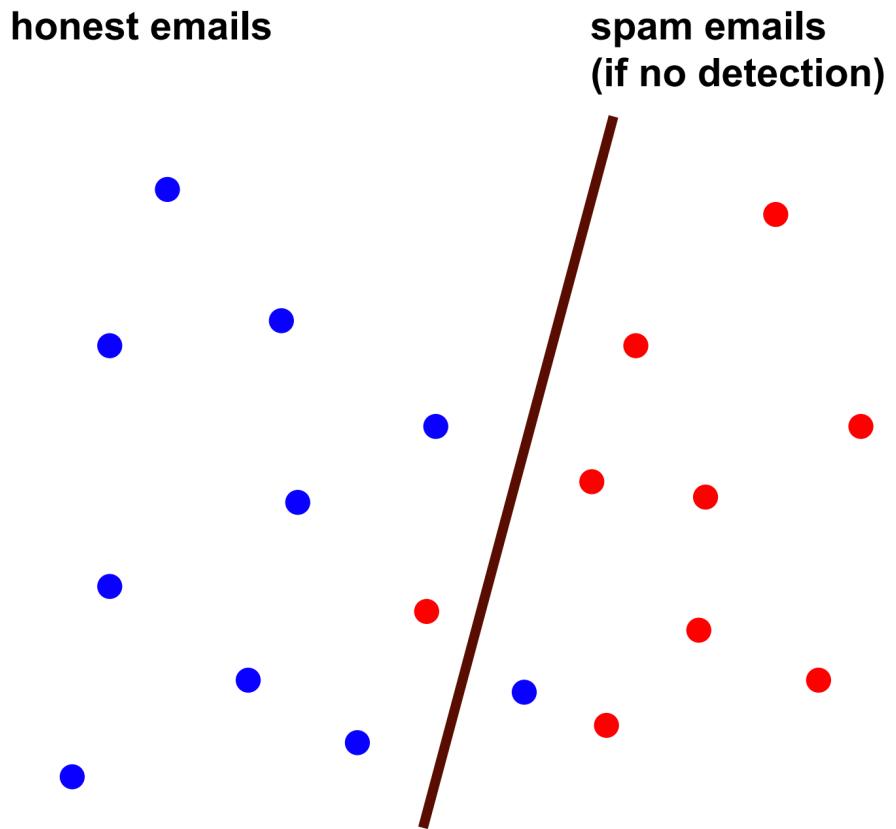
Example: Spam Filter

honest emails

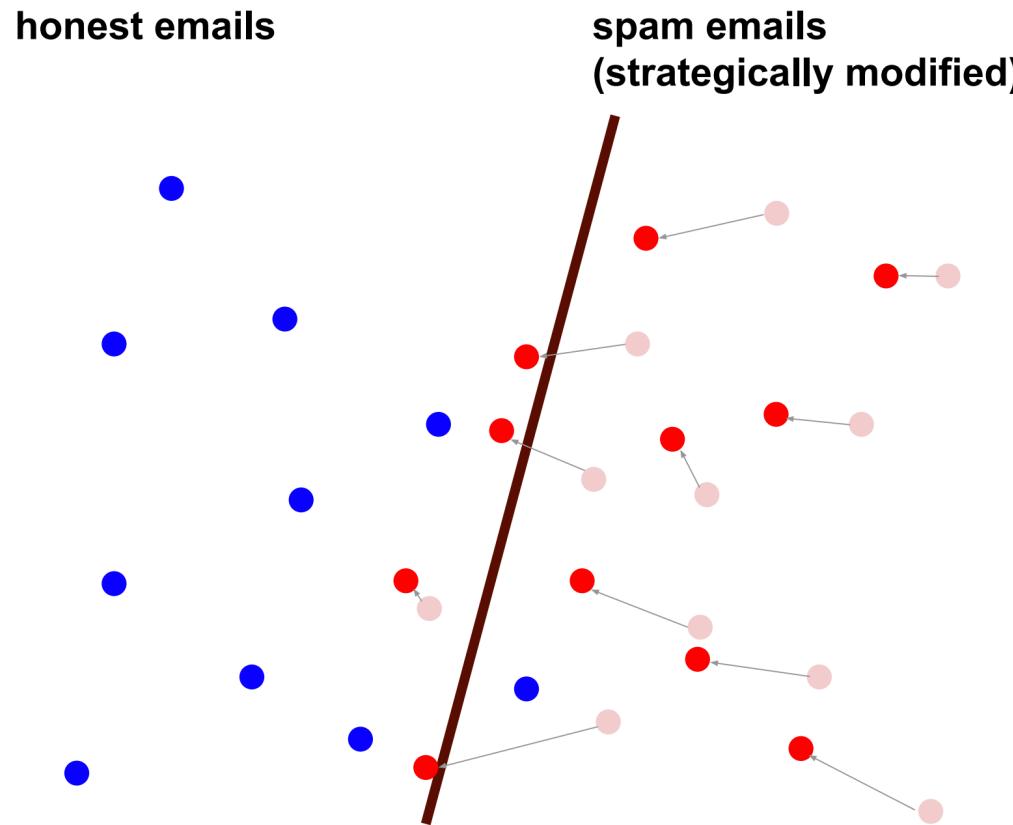
**spam emails
(if no detection)**



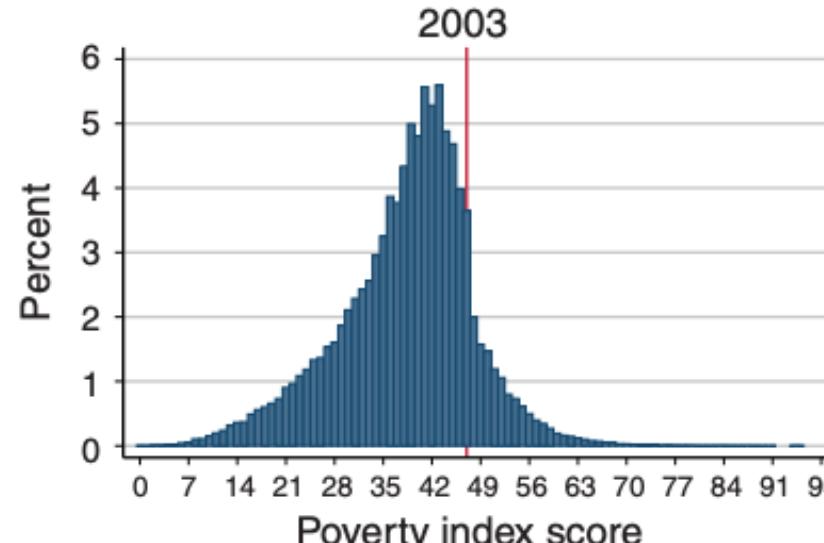
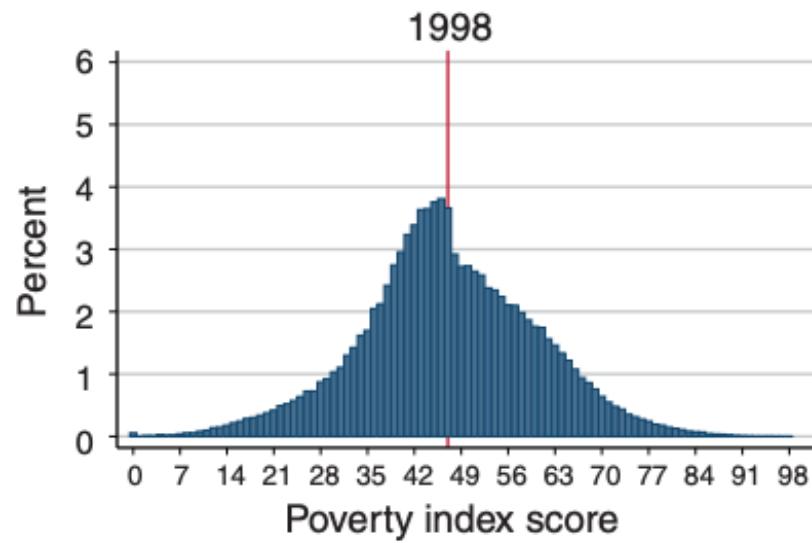
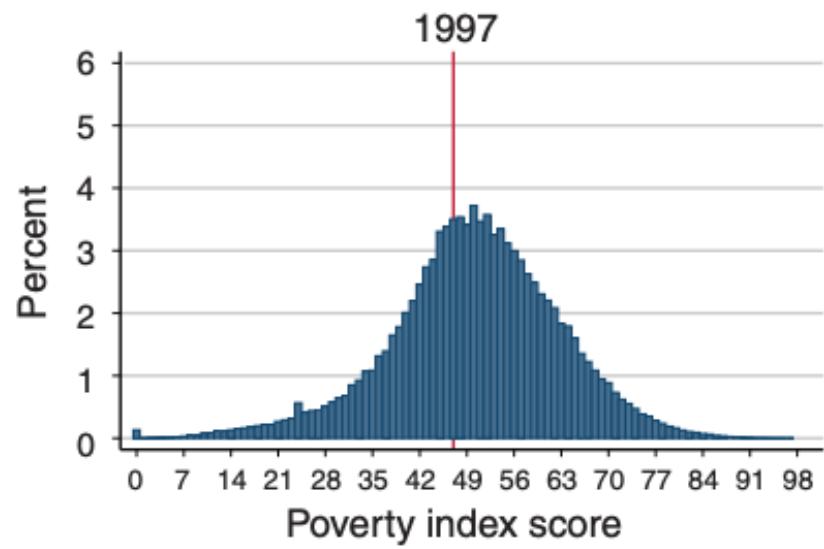
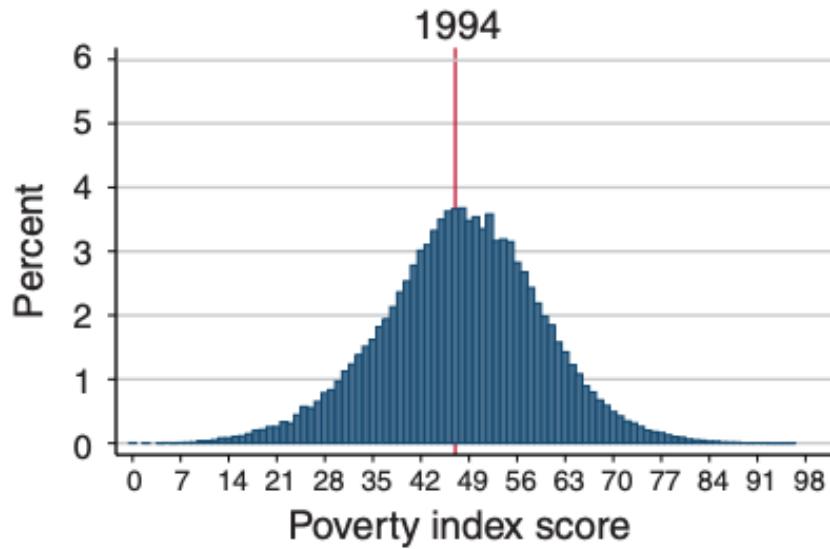
Example: Spam Filter



Example: Spam Filter



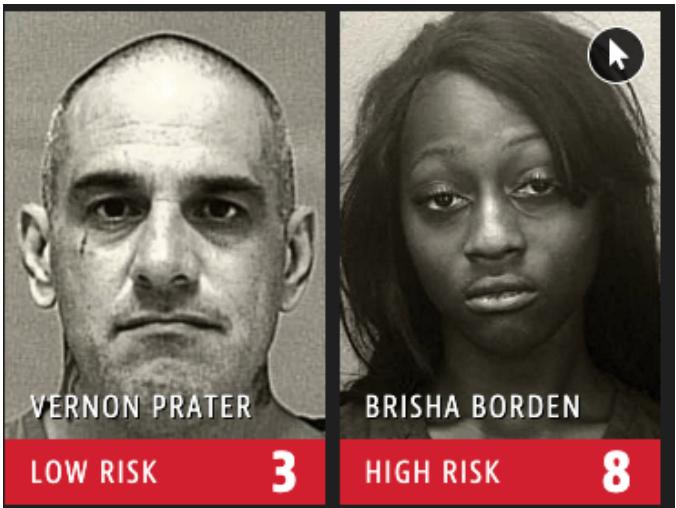
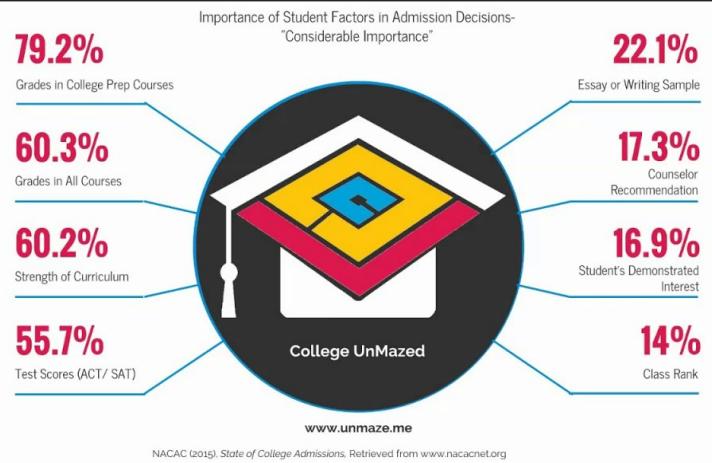
Social Program Eligibility [Camacho and Conover, 2012]



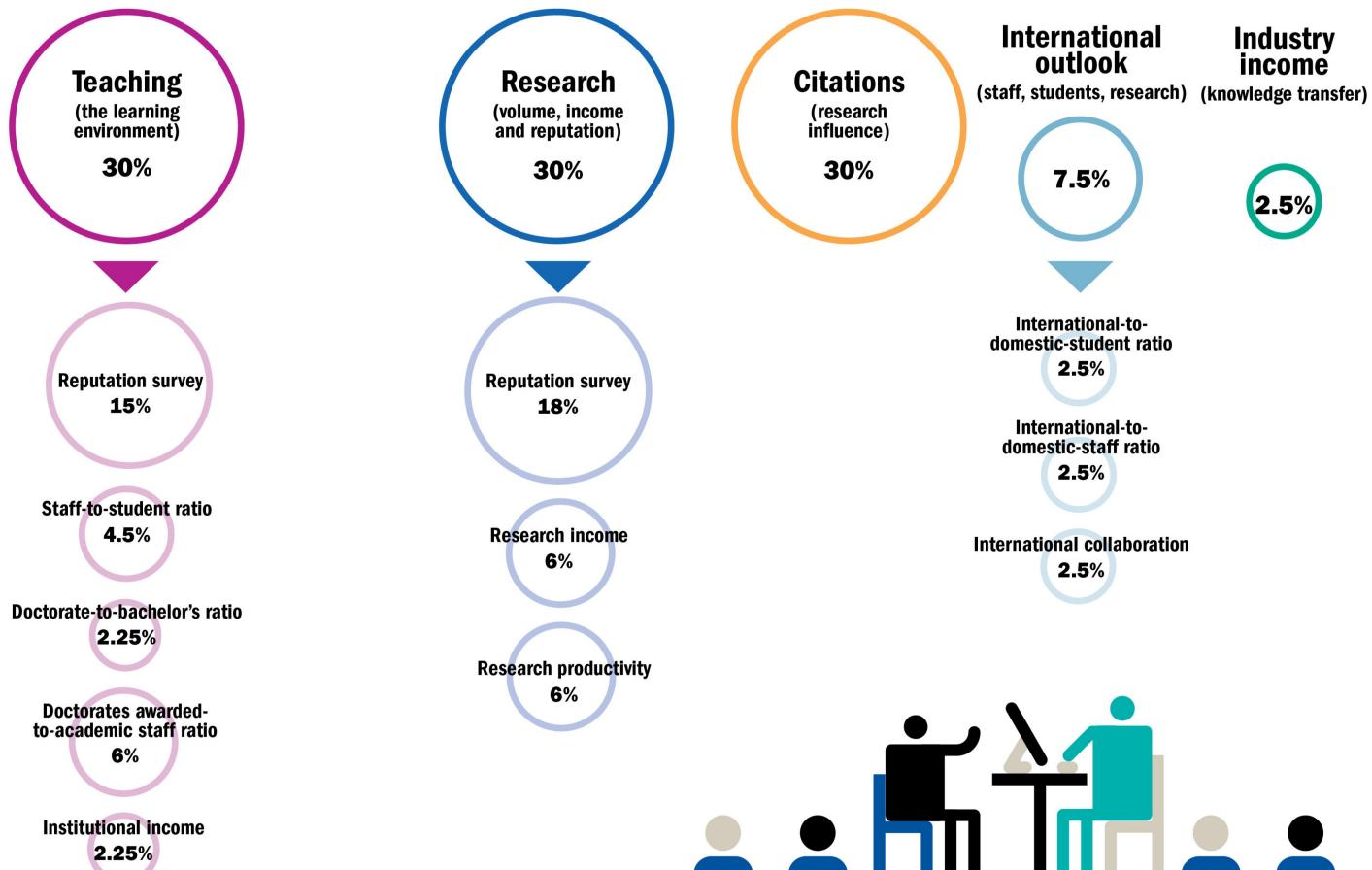
Goodhart's law:

“If a measure becomes the public’s goal,
it is no longer a good measure.”

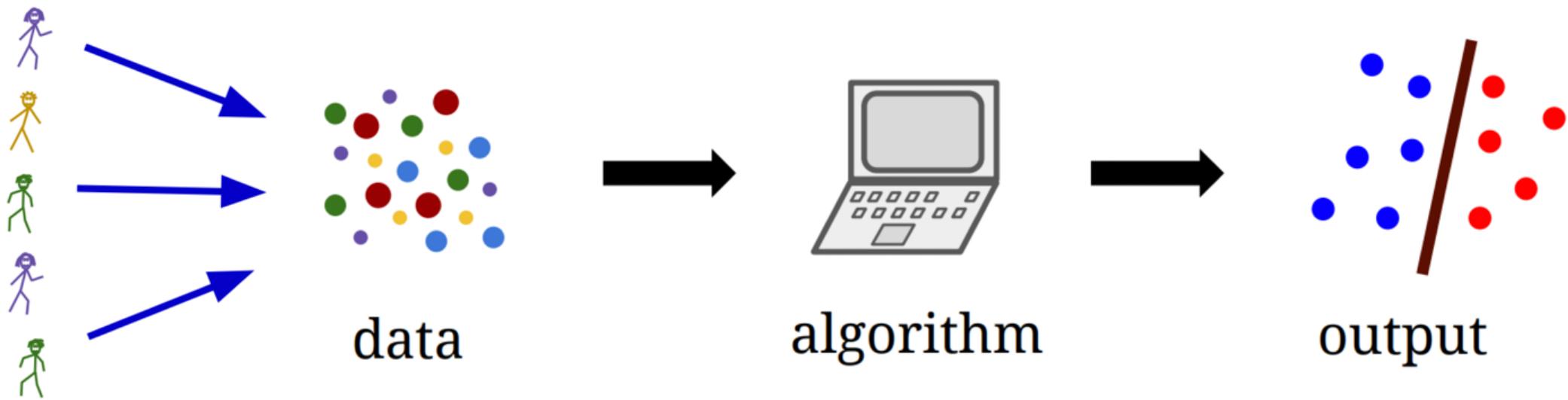
COLLEGE ADMISSIONS



Methodology



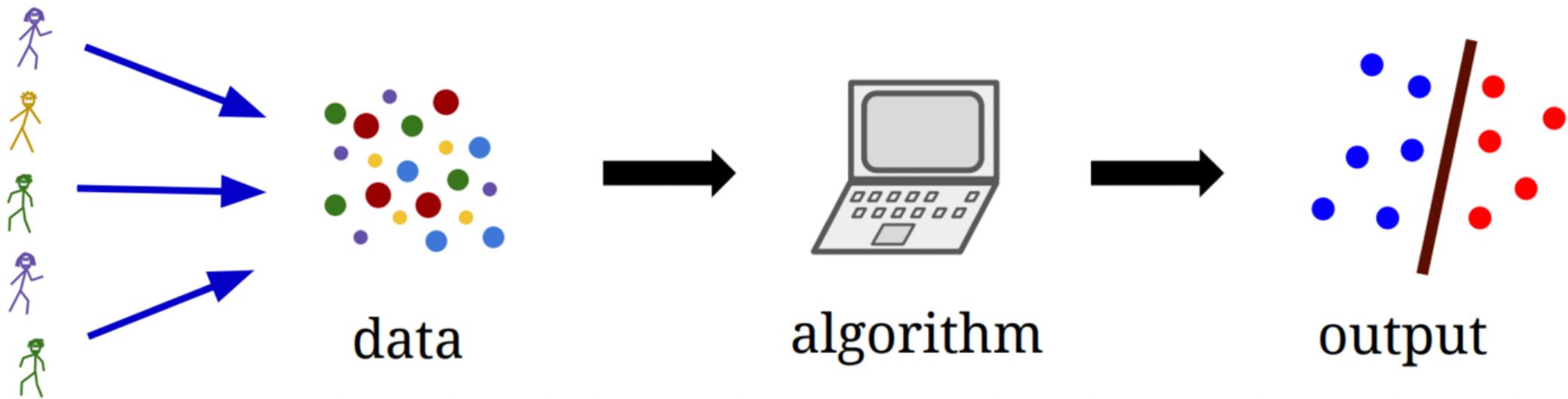
Strategic Classification



Warm-Up Discussion

- What are other examples of strategic classifications in the real world? What are the potential consequences of not considering the issues of strategic manipulations?
- On a high-level, what do you think can be done to help mitigate or address the issues of strategic manipulations?
 - As an example, Google keep **secret** of their ranking algorithms in the search results. Then companies that focus on SEO (search engine optimization) try to find out how the algorithms works. Is secrecy a good approach? What other options do we have?

Strategic Classification



How to take this interaction between ML algorithms and data-holders into account?

Game theoretical modeling

Game Theoretical Modeling

- Key elements:
 - **Players, actions, payoffs**
- **Players:** Jury (e.g., university) and Contestants (student applicants)
- **Actions:**
 - First, Jury decides on the machine learning model (binary classification)
 - Then, Contestant decides how to alter their features based on the model
- **Payoffs**
 - Jury wants to maximize the probability of correct predictions
 - Contestants want to be selected (being predicted as 1)



Choose a classifier
 $f: X \rightarrow \{0,1\}$



Represented by initial features
 $\vec{x} = (\text{SAT score}, \text{GPA}, \text{etc})$

True label $y = h(\vec{x}) \in \{0,1\}$

Choose manipulation (new \vec{x}')
 $\text{cost}(\text{initial } \vec{x}, \text{new } \vec{x}')$

Student distribution D

Given f , student with \vec{x} chooses to manipulate her feature to \vec{x}' to maximize
 $f(\vec{x}') - \text{cost}(\vec{x}, \vec{x}')$

The university chooses the classifier f with the goal to maximize

$$\Pr_{\vec{x} \sim D} [f(\vec{x}') = h(\vec{x})]$$

Stackelberg Game

- Similar to the “contract design” problem
- The “classifier” is a contract



University

Choose a classifier

$$f: X \rightarrow \{0,1\}$$



Represented by initial features
 $\vec{x} = (\text{SAT score}, \text{GPA}, \text{etc})$

True label $y = h(\vec{x}) \in \{0,1\}$

Choose manipulation (**new \vec{x}'**)
cost(initial \vec{x} , new \vec{x}')

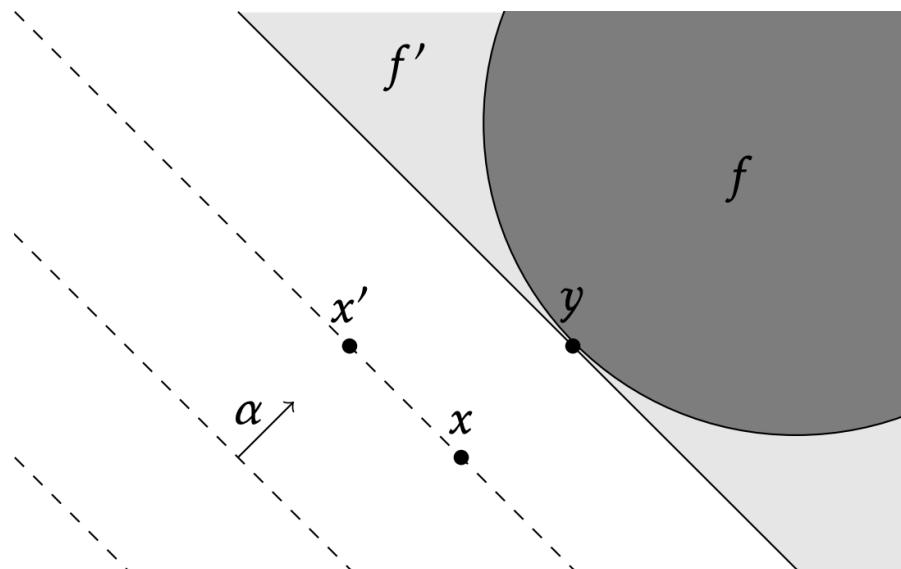
Student distribution D

Research question of the required reading:

Design the algorithm for computing the optimal f when taking strategic manipulations into account.

Main Results

- In general cases, finding the classification rule with near-optimal performance is NP-hard.
- In special cases, there exist efficient algorithms
 - E.g., when $\text{cost}(\vec{x}, \vec{x}') = \vec{\alpha} \cdot (\vec{x}' - \vec{x})$



More Aspects of Strategic Classification

1. Evaluation vs. Incentive
2. Social costs

Types of Manipulations

- Say that your insurance company wants you to stay healthy (so they can pay less) and decide reward you if you hit the step target on your phone.



Shake Wiggle Device Mobile Phone Holder
Automatic Swing Motion for Mobile Phone Run
Step Count Program

[Visit the zjchao Store](#)

4.5 ★★★★☆ 14 ratings

Price: \$33.61 & FREE Shipping. [Details](#) & [FREE Returns](#)

Coupon Save an extra 11% when you apply this coupon. [Details](#)

New (2) from \$33.61 + FREE Shipping

Report incorrect product information.



Save on the items you want with Alexa
Alexa now gives you deal alerts for items on your Wish List. [Learn more about this feature >](#)

Gaming vs. Improvement

- **Gaming**
 - Alter the decision by manipulating proxy features without changing the underlying label
 - Unjustifiable or pointless effort
 - What's considered in the required reading
- **Improvement**
 - Change the decision by manipulation that changes the underlying label
 - Maybe this is the positive effect that we want to utilize

The Purposes of Decision Rules

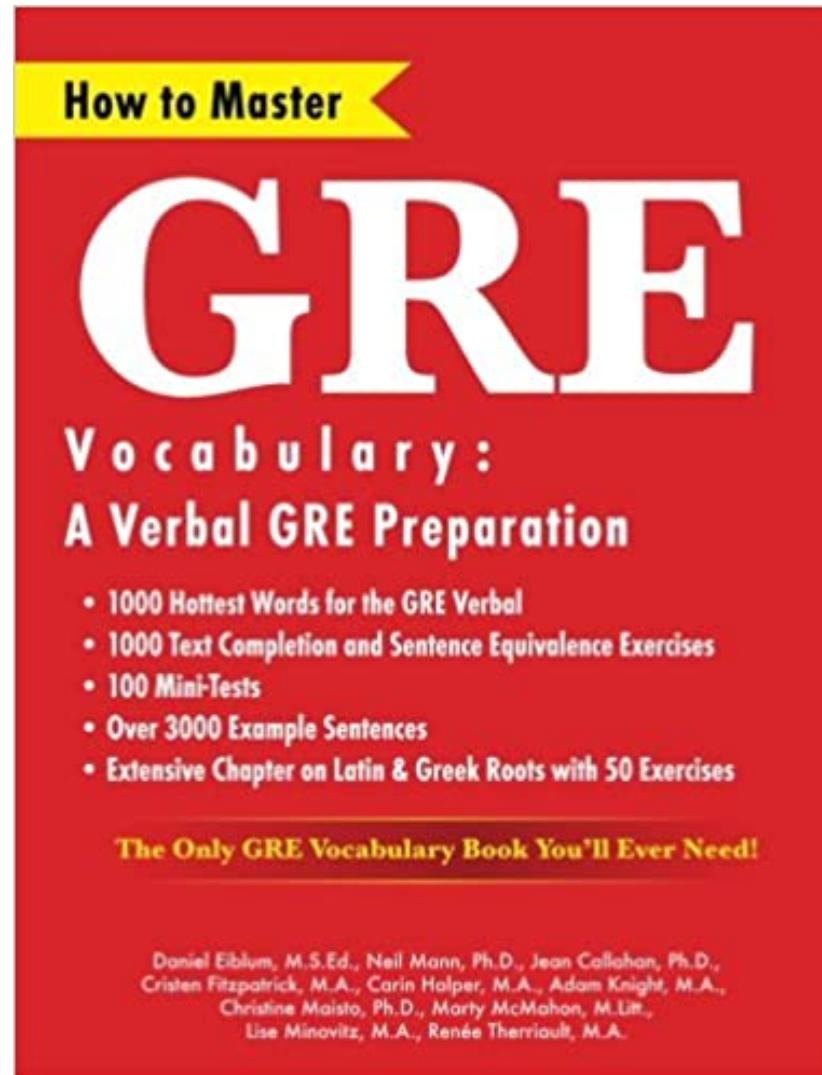
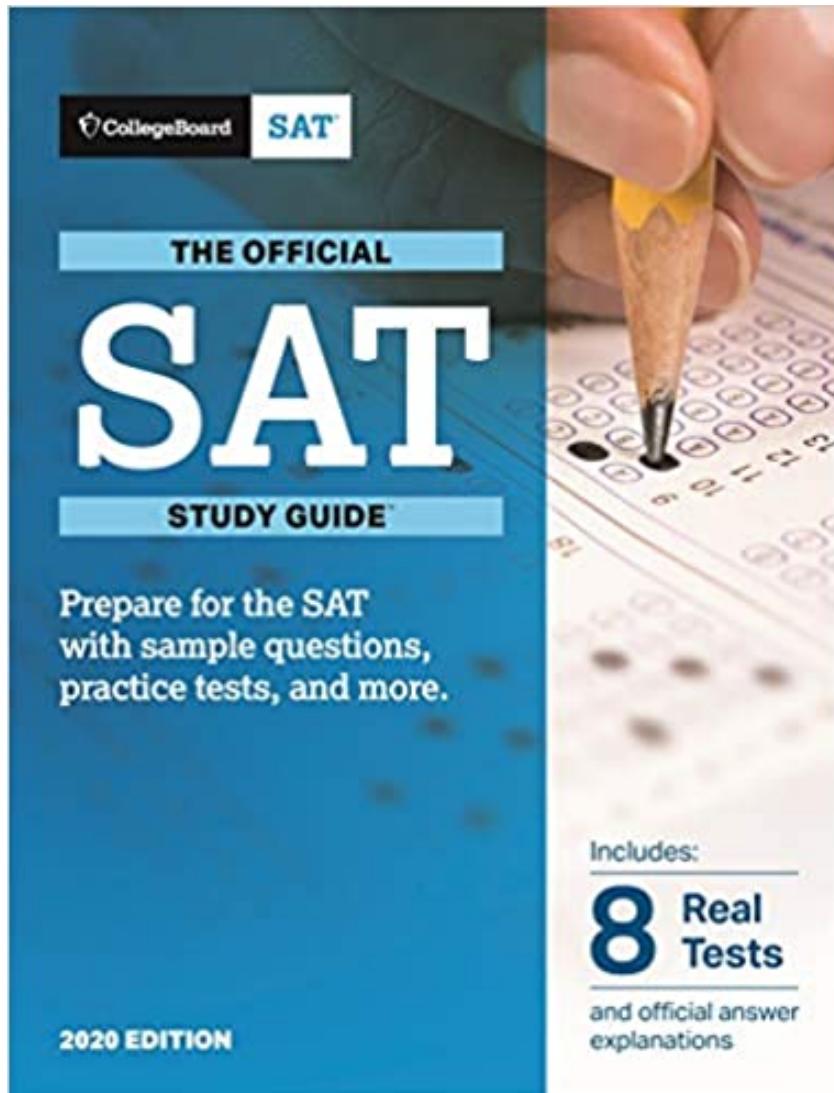
- To evaluate the candidates
 - Assume there is a true unobservable quality that we care about
 - e.g., whether the student will succeed or not
 - Unobservable quality leads to observable but manipulatable features
 - Aims to classify based on the true quality with gaming behavior
- To incentivize the candidates
 - Assume the true unobservable quality can be improved
 - Incentivize candidates to perform desired improvements

Discussion

- Think about examples that we can use **classification as incentive** to motivate desired actions.
- How can we design the incentives?
 - Informally, what behavior do you want to encourage, and how to make the classification rule achieve that?
 - Formally, how the classification rule should look like? How to find the “optimal” rule?
- Example: how the course grades are split up would impact what you do for the course. What’s the “best” way to design the grades? (What should be the desired behavior?)

How do classifiers induce agents to invest effort strategically?

Kleinberg and Raghavan. EC 2019.

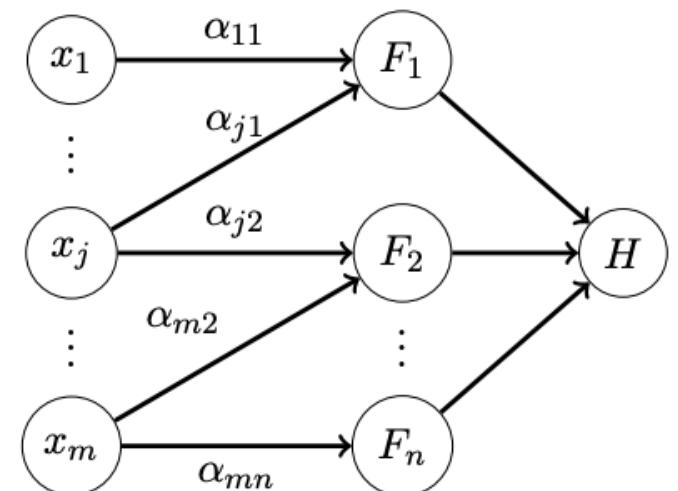


- “Test preparation has been the focus of intense argument for many years, and all sorts of different terms have been used to describe both good and bad forms. . . I think it’s best to. . . distinguish between seven different types of test preparation: **Working more effectively**; **Teaching more**; **Working harder**; **Reallocation**; **Alignment**; **Coaching**; **Cheating**. The first three are what proponents of high-stakes testing want to see”
- - Daniel Koretz, Measuring Up.

How to induce the desired behavior?

Graphical Representation of the Model

- Consider a single student
 - x_1, \dots, x_m : the effort students spent on actions
 - E.g., studying, working in a reading group, copying answers online
 - F_1, \dots, F_n : the set of observable features after actions
 - E.g., homework grades, exam grades
 - $\alpha_{i,j}$: the contribution to feature F_i from action x_j
 - $F_i = f(\sum_{j=1}^m \alpha_{j,i} x_j)$
- The teacher designs the final grades H based on F
 - $H = \sum_{i=1}^n \beta_i F_i$



Example: Course Grades

cheating

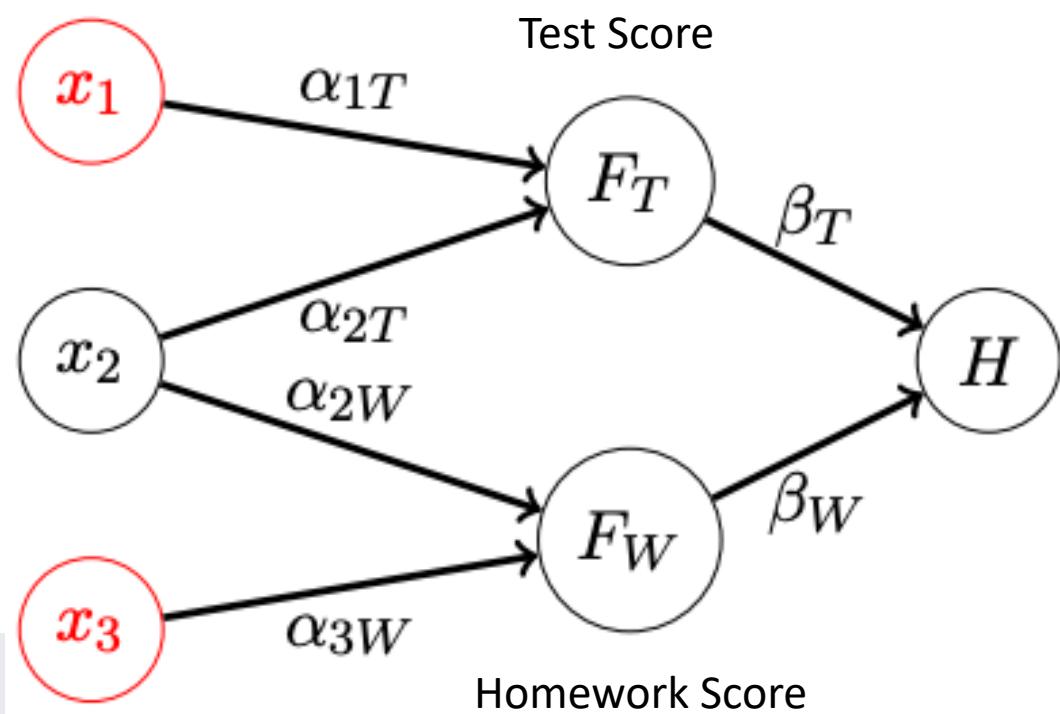


studying



copying
homework

SOLUTION MANUAL
www.SOLUTIONMANUAL.INFO



Example: Course Grades

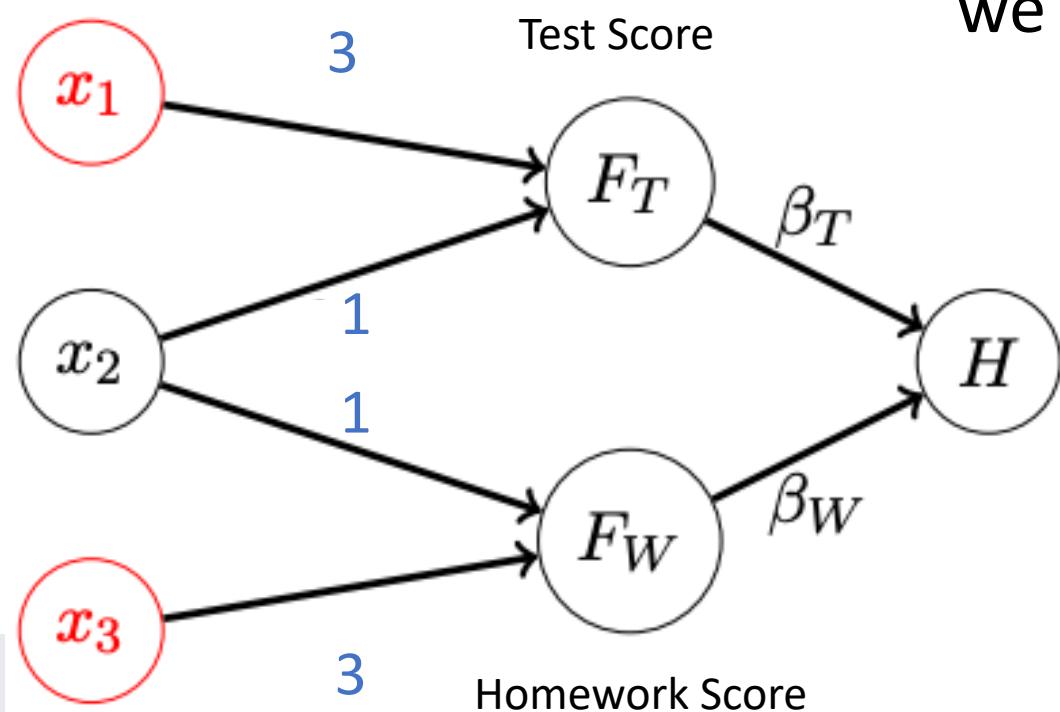
cheating



studying



copying
homework



When the desired effort is **substitutable**, we can't incentivize that effort.

Example: Course Grades

cheating

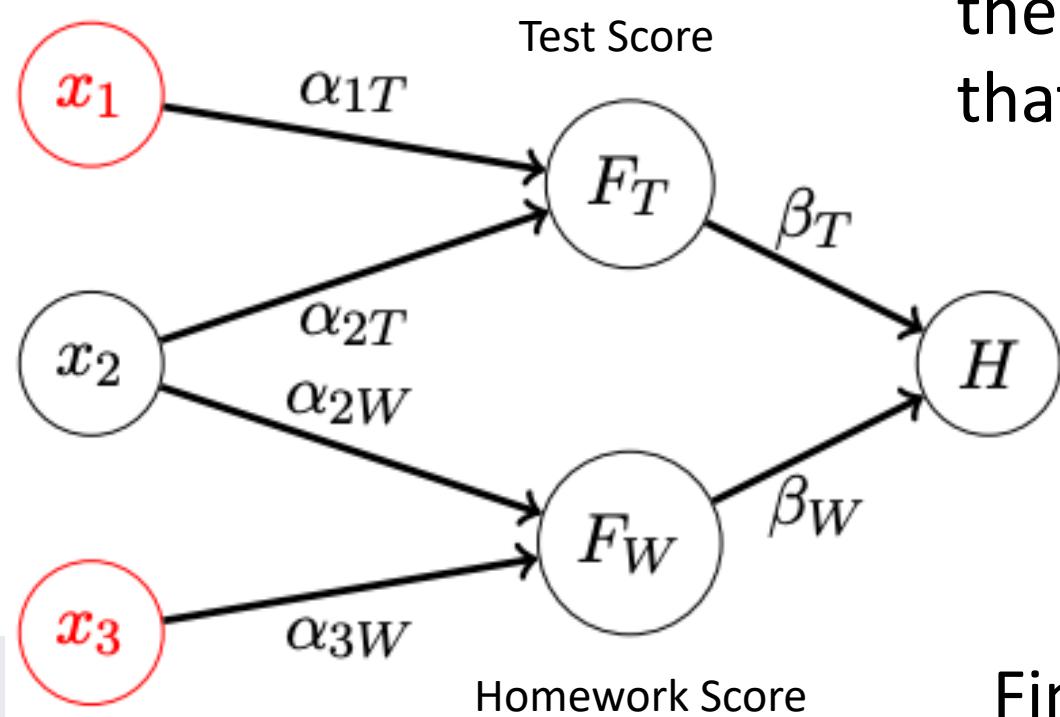


studying



copying
homework

SOLUTION MANUAL
www.solutionmanual.info



When the effort is not **substitutable**, there is a **linear mechanism** incentivizing that effort.

Finding that mechanism is NP-hard.

The Purposes of Decision Rules

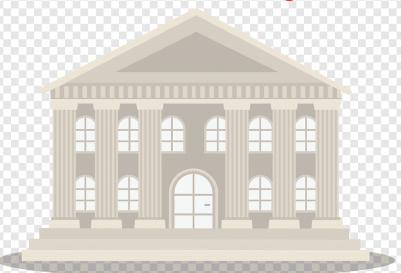
- To evaluate the candidates
 - Assume there is a true unobservable quality that we care about
 - Whether the student will succeed or not
 - Unobservable quality leads to observable but manipulatable features
 - Aims to classify based on the true quality with gaming behavior
- To incentivize the candidates
 - Assume the true unobservable quality can be improved
 - Incentivize candidates to perform desired improvements

More Aspects of Strategic Classification

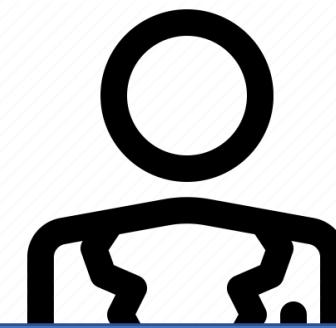
1. Evaluation vs. Incentive
2. Social costs

There Are Two Parties in Strategic Classification

University



Choose a classifier
 $f: X \rightarrow \{0,1\}$



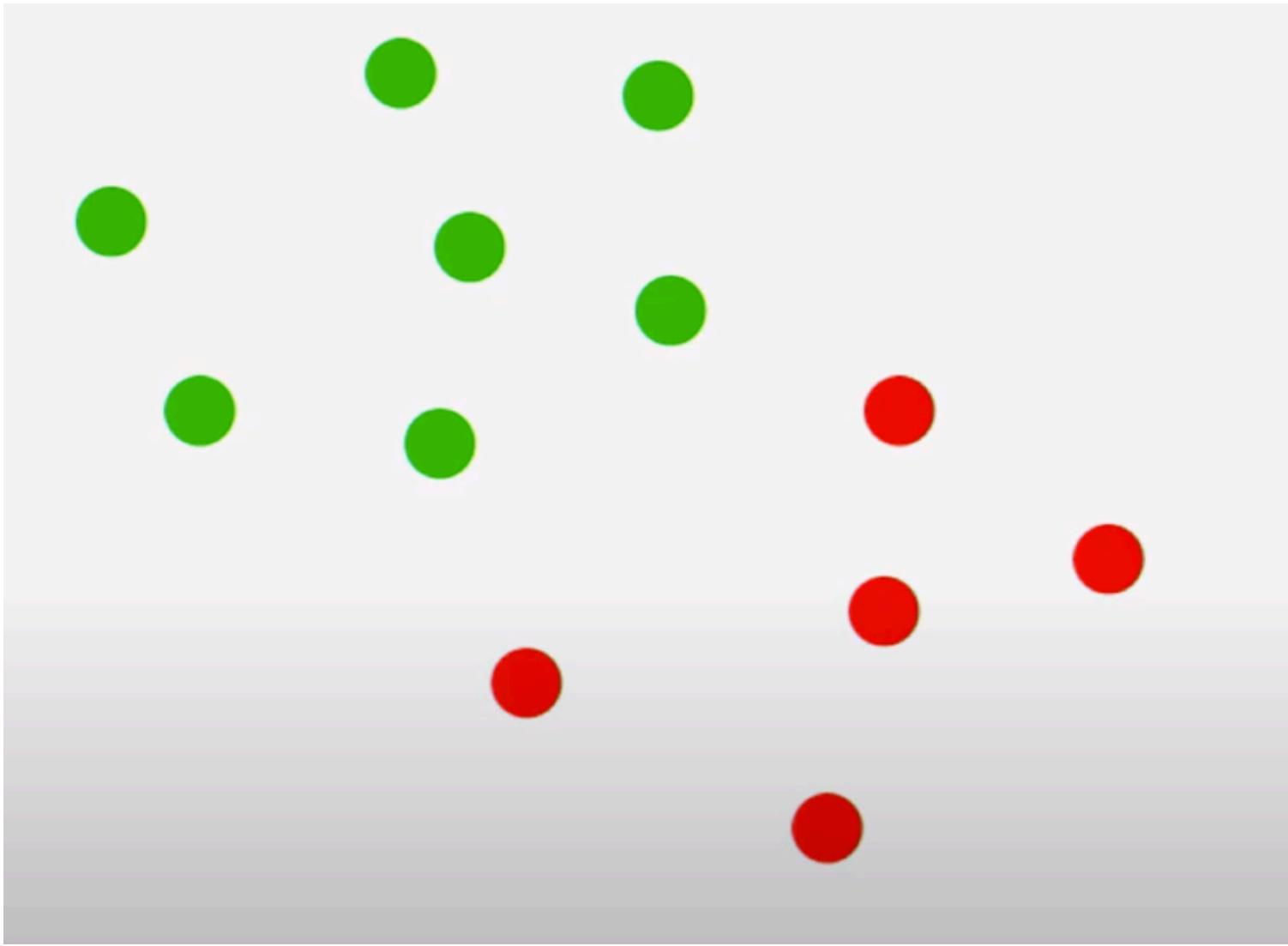
Represented by initial features
 $\vec{x} = (\text{SAT score, grades, etc})$

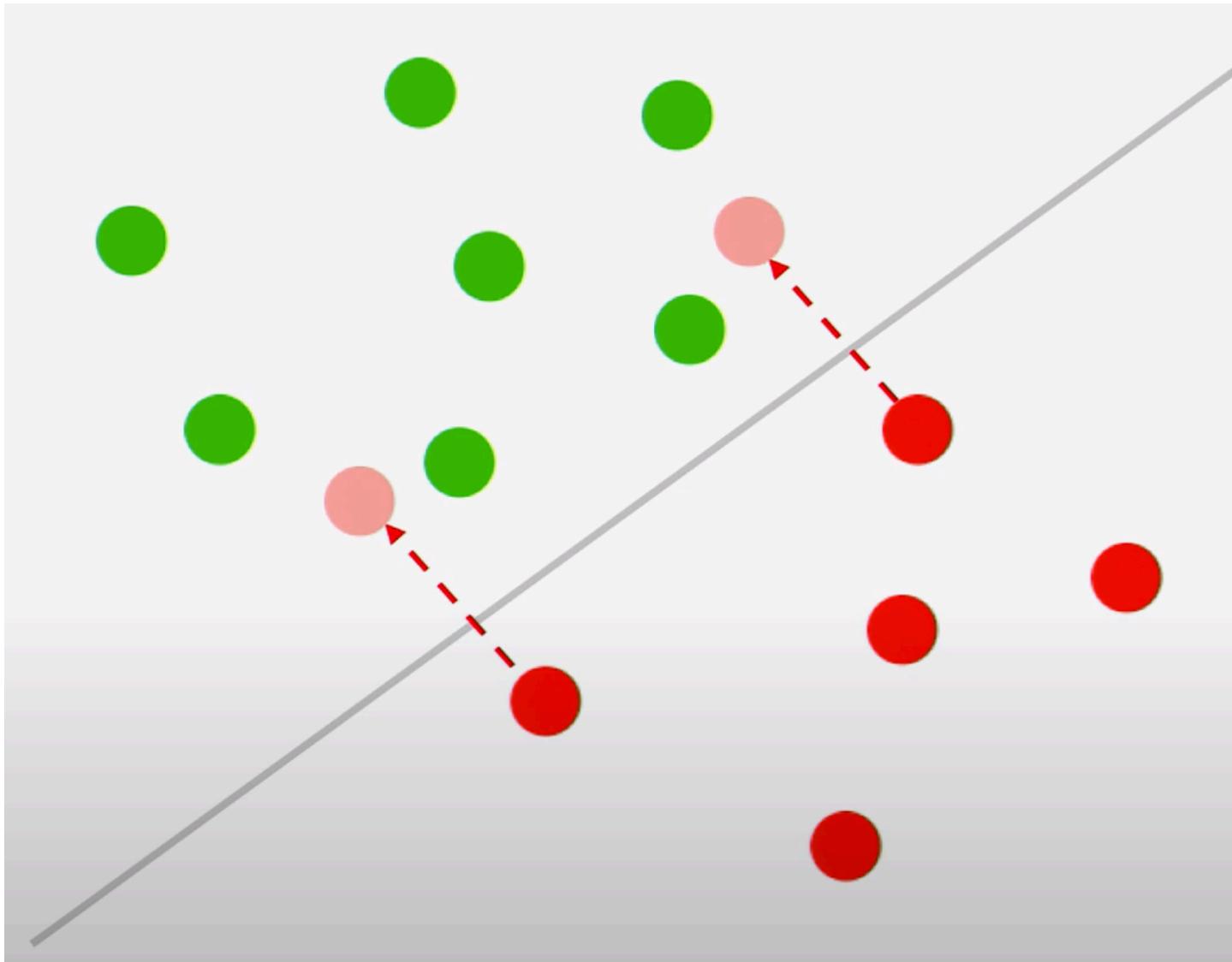
True label $y = h(\vec{x}) \in \{0,1\}$

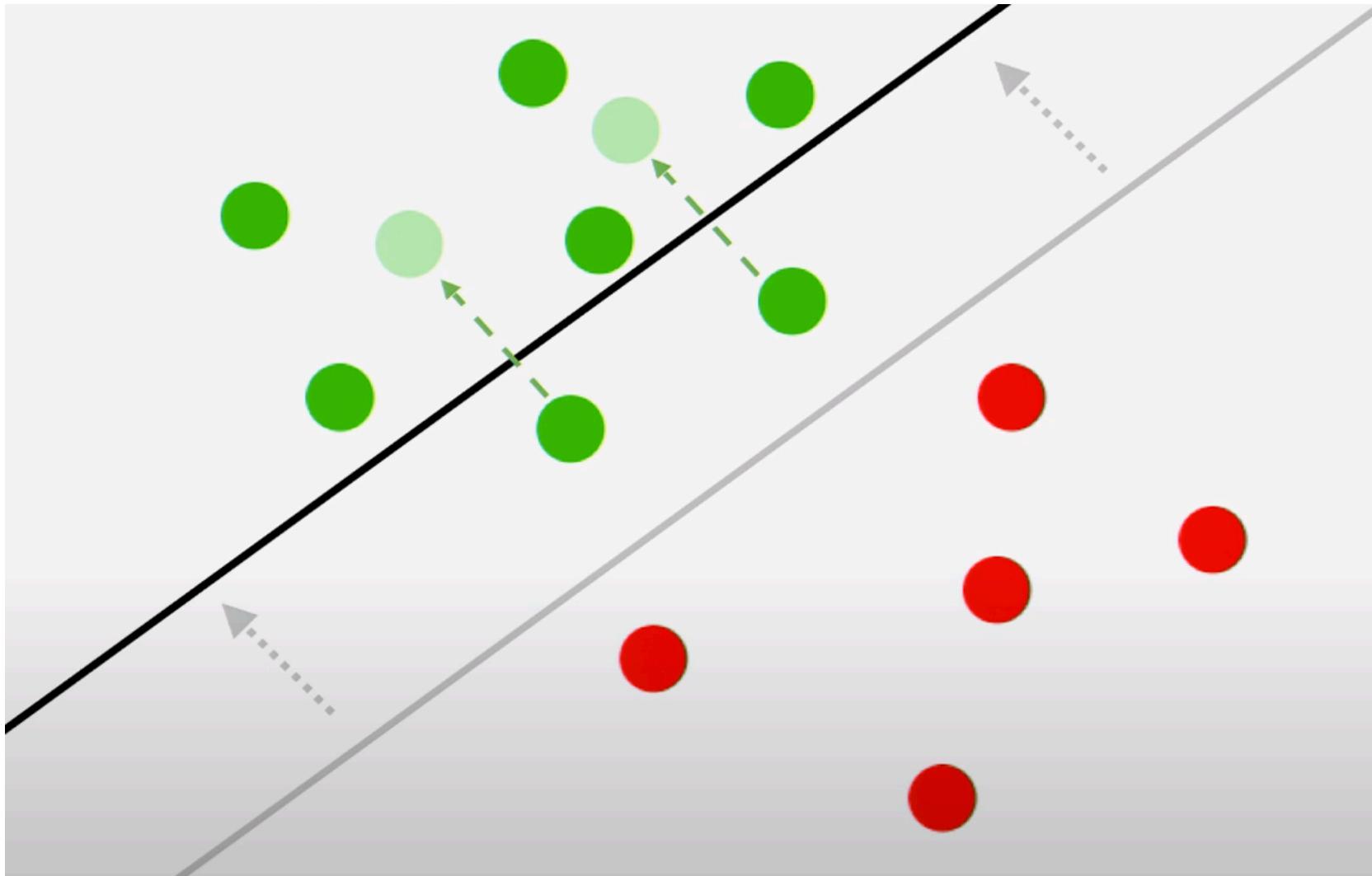
Choose manipulation (new \vec{x}')
 $\vec{x}, \text{new } \vec{x}'$

distribution D

Is it reasonable to only optimize
the benefit of the institution?

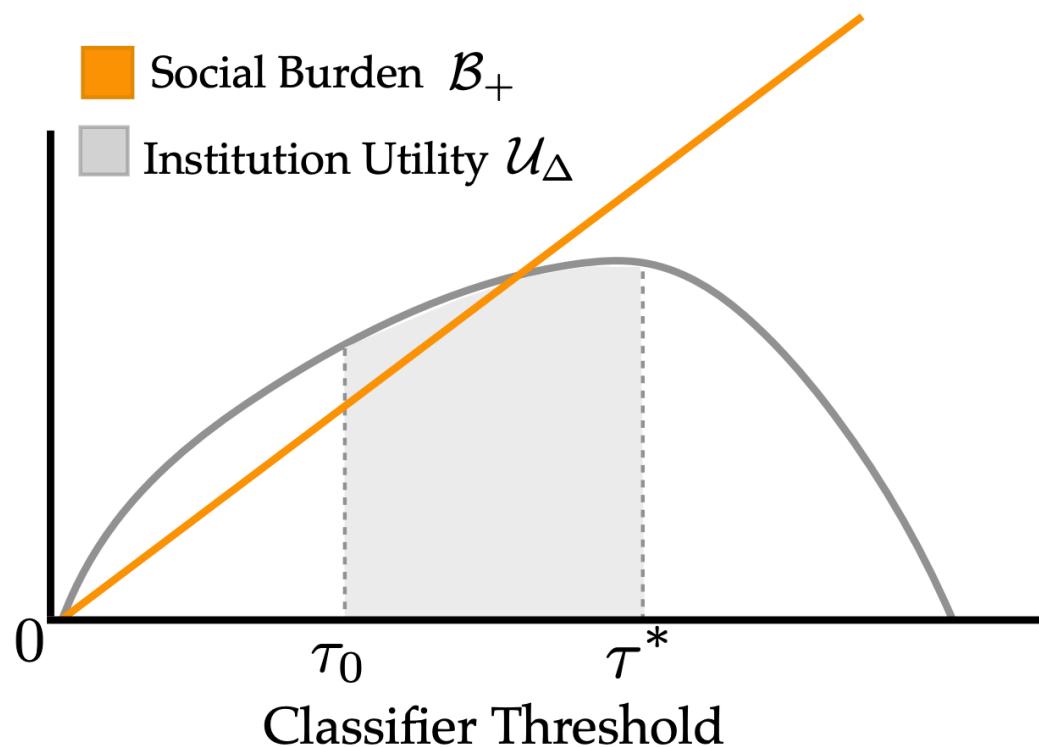






Institution Utility vs. Social Burden [Milli et al. 2018]

- Institution utility: the utility for deploying the decision rule
- Social burden: the amount of effort contestants need to put in



Read our COVID-19 research and news.

SHARE



Science

Contents ▾

News ▾

Careers ▾

Journals ▾

Read our COVID-19 research and news.



GREs don't predict grad school success. What does?

By Beryl Lieff Benderly | Jun. 7, 2017, 8:30 AM

EDUCATION

The Problem With the GRE

The exam “is a proxy for asking ‘Are you rich?’ ‘Are you white?’ ‘Are you male?’”

VICTORIA CLAYTON MARCH 1, 2016

EDUCATION AND OUTREACH | NEWS

US graduate entry exams not a predictor of PhD success, says study

28 Jan 2019



ETS Home > GRE Home > Research > Validity Evidence: Predicting Success in Graduate Education

Validity Evidence: Predicting Success in Graduate Education

Reports in this section support the validity arguments for the interpretation of scores from the GRE® General Test and Subject Tests.

GRE Research

➤ Validity Evidence: Predicting Success in Graduate Education

Validity Evidence: Constructs and Content

Fairness and Accessibility

Psychometric Issues

Developing New Measures

Other Research

Key Reports

- [The Validity of GRE General Test Scores for Predicting Academic Performance at U.S. Law Schools](#)
By D. M. Klieger, B. Bridgeman, R. J. Tannenbaum, F. A. Cline, M. Olivera-Aguilar (2018)
ETS Research Report No. RR-18-26
- [The Validity of Scores from the GRE revised General Test for Forecasting Performance in Business Schools: Phase One](#)
By J. W. Young, D. Klieger, J. Bochenek, C. Li, and F. Cline (2014)
GRE Board Report No.14-01
- [The Role of Noncognitive Constructs and Other Background Variables in Graduate Education](#)
By P. C. Kyllonen, A. M. Walters, and J. C. Kaufman (2011)
GRE Board Report No. 00-11
- [Understanding What the Numbers Mean: A Straightforward Approach to GRE Predictive Validity](#)
By B. Bridgeman, N. Burton, and F. Cline (2008)
GRE Board Report No. 04-03
- [Predicting Long-Term Success in Graduate School: A Collaborative Validity Study](#)
By N. W. Burton and M. Wang (2005)
GRE Board Report No. 99-14R

Discussion

- People in advantage groups might be able to pay smaller costs to manipulate the features. What are the example applications that this could happen (e.g., SAT scores)?
- What do you think are the bad consequences with this imbalanced manipulations? What are potential ways we can deal with them?

The Disparate Effects of Strategic Manipulation

Hu, Immorlica, and Vaughan. FAT* 19.



University

Choose a classifier
 $f:X \rightarrow \{0,1\}$

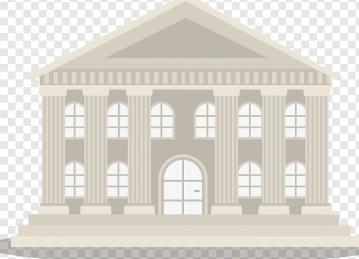


Represented by initial features
 $\vec{x} = (\text{SAT score, grades, etc})$

True label $y = h(\vec{x}) \in \{0,1\}$

Choose manipulation (**new \vec{x}**)
cost(initial \vec{x} , new \vec{x})

Student distribution D



University

Choose a classifier
 $f:X \rightarrow \{0,1\}$



Represented by initial features
 $\vec{x} = (\text{SAT score, grades, etc})$

True label $y = h(\vec{x}) \in \{0,1\}$

Choose manipulation (new \vec{x})
cost(initial \vec{x} , new \vec{x})

Student distribution D

What if this cost is different across groups?

Main Results

- Group differences
 - Group A: Advantage group
 - Group B: Disadvantage group
- Result 1: Reinforcing inequalities
 - Under mild conditions, the equilibrium classifier more likely to **mistakenly exclude people from group B** and **mistakenly admit people from group A**
- Result 2: Subsidy interventions?
 - There exists cases that both groups are worse-off when a subsidy is offered compared to no subsidy at all.

Summary: Strategic Classification

- Mitigate the effect of gaming
 - How do we ensure we still have a good classifier even if we know that people will game the system
- Using classification to incentivize improvements
 - Classification as incentives
- Social costs
 - The tradeoffs between the ML utility and the costs incurred on the individuals