

# CMSC 3540I: The Interplay of Learning and Game Theory (Autumn 2022)

## Learning From Strategic Data Sources

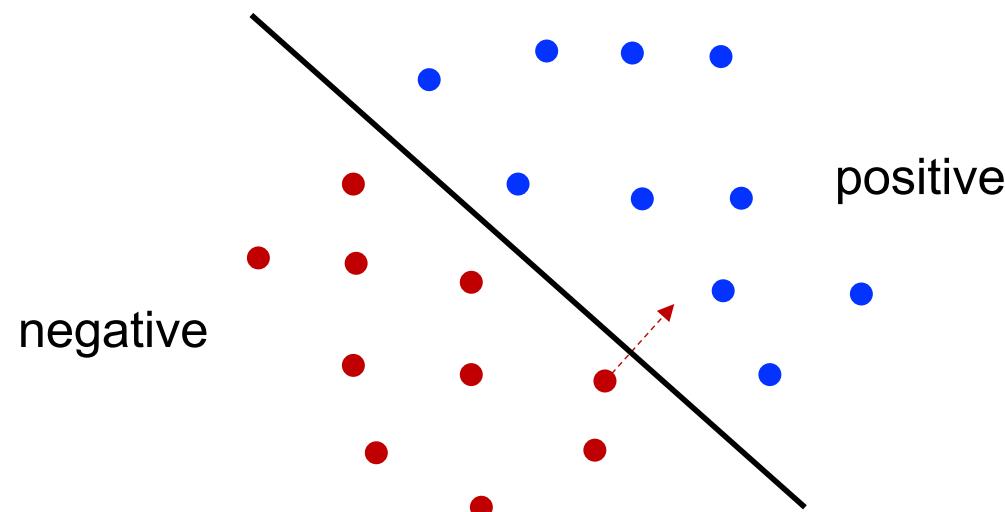
Instructor: Haifeng Xu



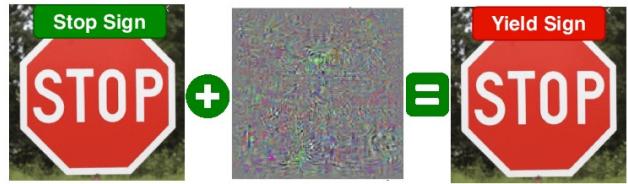
# Outline

- Introduction to Strategic Classification
- Learnability and Computability of Strategic Classifiers
- Beyond Classification

# Classification



Data points' features may be manipulated



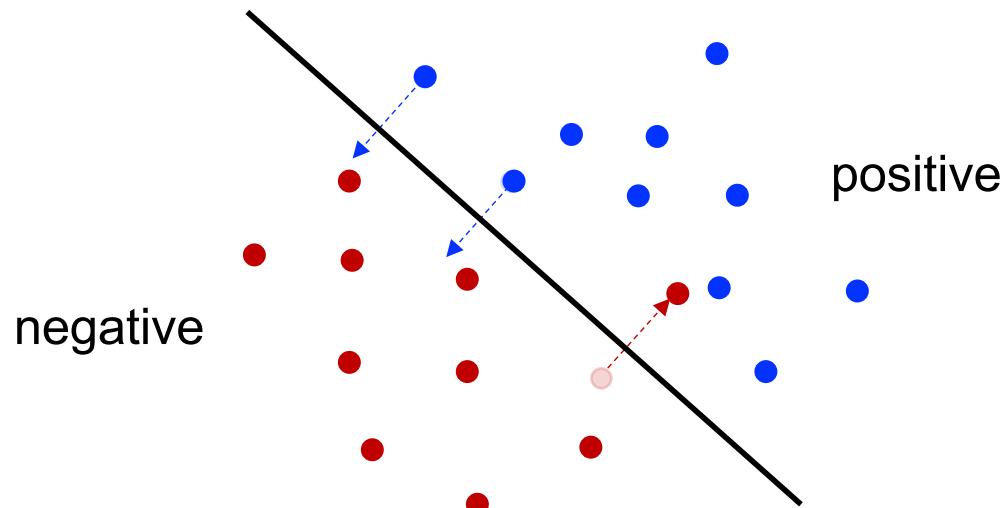
## Adversarial attack

[Goodfellow et al.'15]

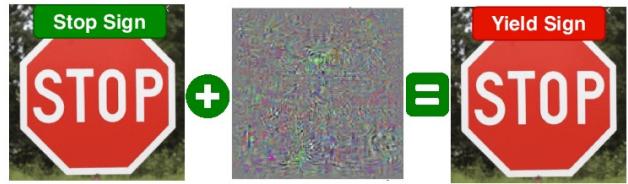
[Eykholt et al.'18]

[Cullina et al.'18]

.....



Data points' features may be manipulated



Adversarial attack



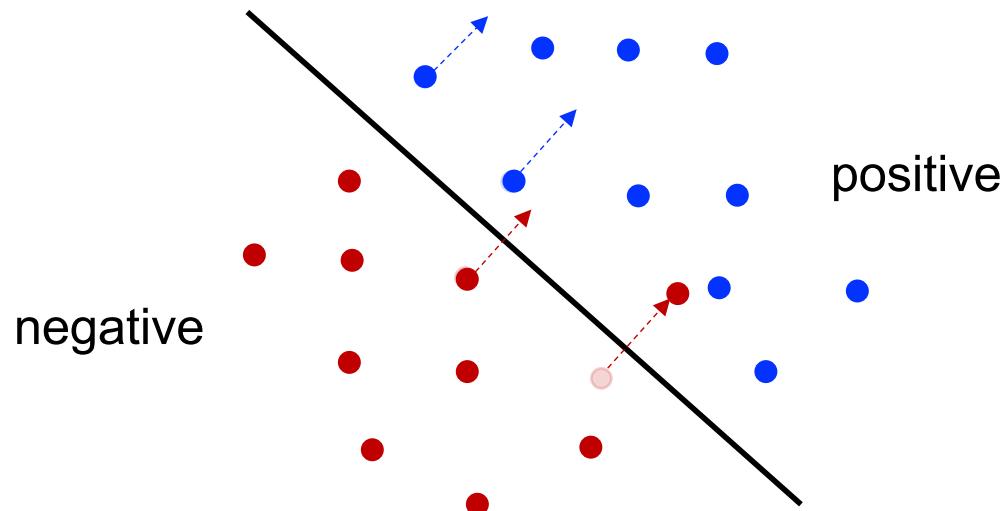
Strategic manipulation

[Hardt et al.'16]

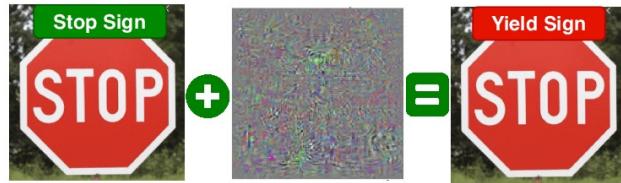
[Hu et al.'19]

[Ghalme et al.'21]

.....



Data points' features may be manipulated

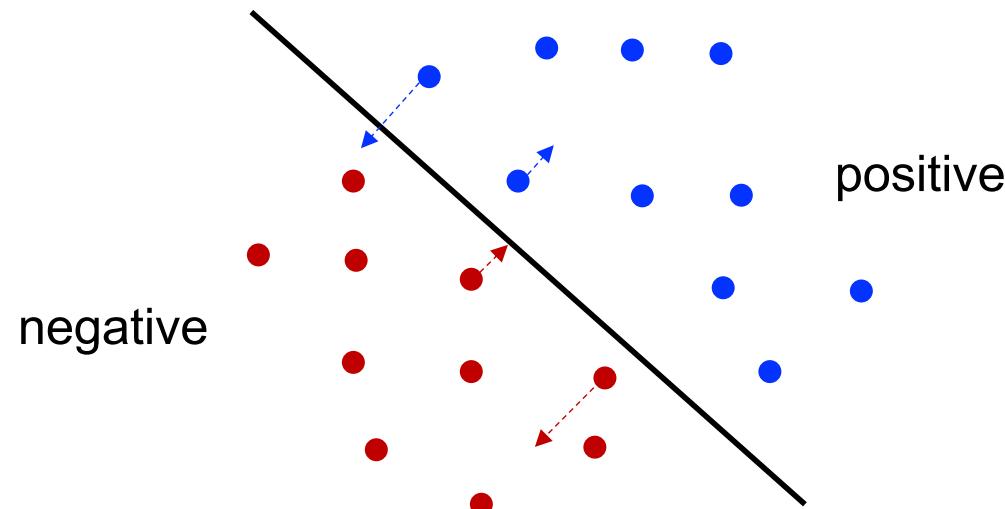


Adversarial attack

[SVXY'21]



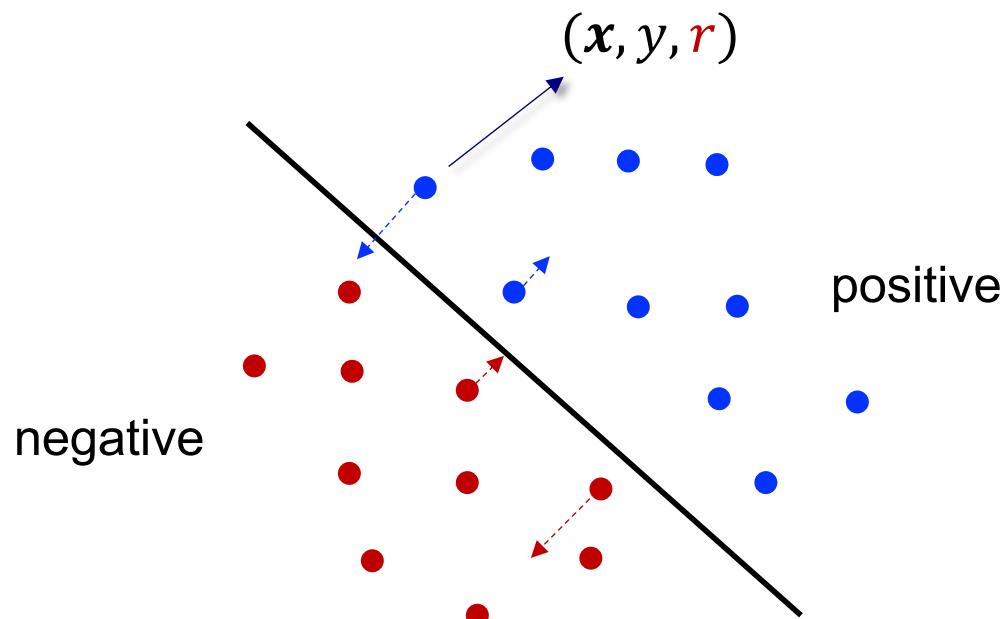
Strategic manipulation



Data points' features may be manipulated

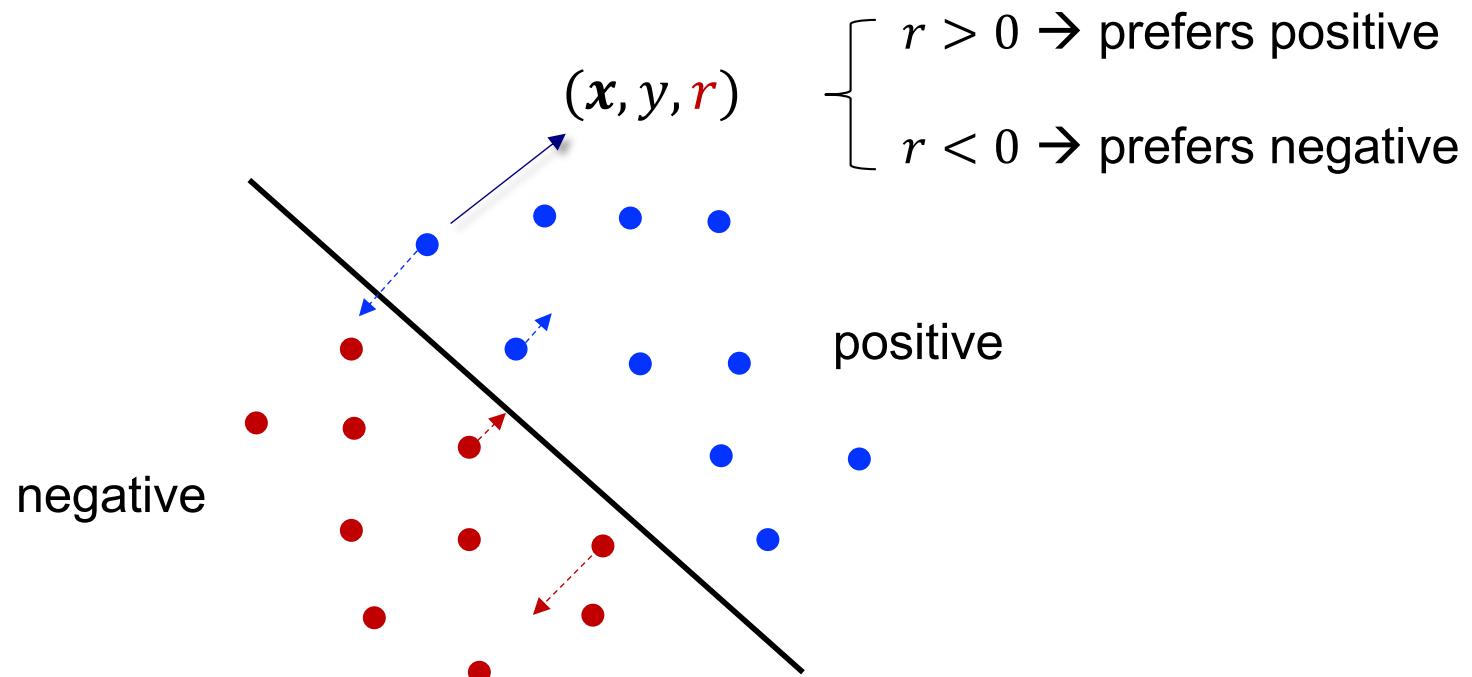
# A Unified Model of Strategic Classification

- Each data point is an individual agent, represented by  $(x, y, r)$ 
  - $r \in \mathbb{R}$  capture the point's incentive of being classified as positive



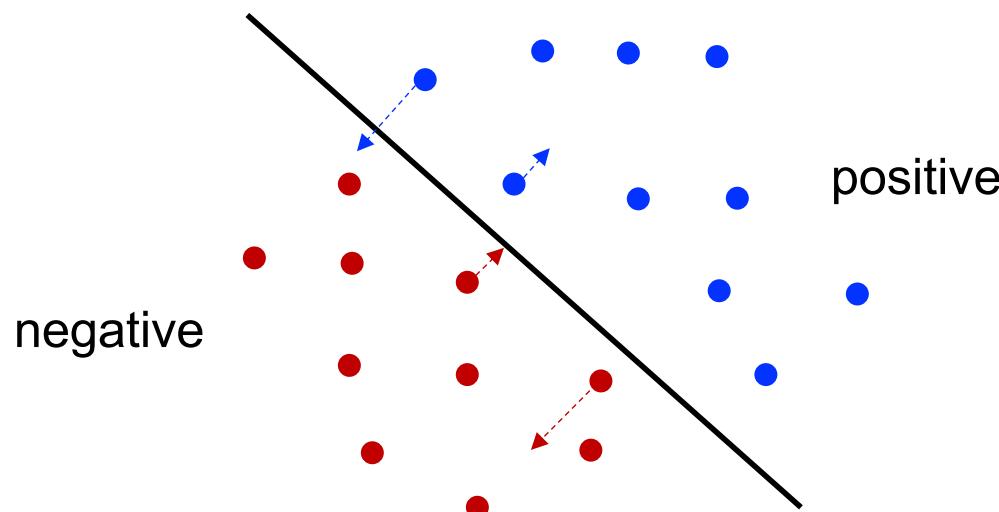
# A Unified Model of Strategic Classification

- Each data point is an individual agent, represented by  $(x, y, r)$ 
  - $r \in \mathbb{R}$  capture the point's incentive of being classified as positive



# A Unified Model of Strategic Classification

- Each data point is an individual agent, represented by  $(x, y, r)$ 
  - $r \in \mathbb{R}$  capture the point's incentive of being classified as positive
- Manipulating feature from  $x$  to  $z$  incurs cost  $c(x - z)$ 
  - $c$  is an arbitrary semi-norm



# A Unified Model of Strategic Classification

- Each data point is an individual agent, represented by  $(x, y, r)$ 
  - $r \in \mathbb{R}$  capture the point's incentive of being classified as positive
- Manipulating feature from  $x$  to  $z$  incurs cost  $c(x - z)$ 
  - $c$  is an arbitrary semi-norm
- Given classifier  $f: X \rightarrow \{0, 1\}$ , data point  $(x, y, r)$  will manipulate its feature to  $z$  that maximizes utility

$$\underbrace{r \cdot \mathbb{I}(f(z) = 1)}_{\text{reward from classification outcome}} - \underbrace{c(x - z)}_{\text{Manipulation cost}}$$

# A Unified Model of Strategic Classification

- Each data point is an individual agent, represented by  $(x, y, r)$ 
  - $r \in \mathbb{R}$  capture the point's incentive of being classified as positive
- Manipulating feature from  $x$  to  $z$  incurs cost  $c(x - z)$ 
  - $c$  is an arbitrary semi-norm
- Given classifier  $f: X \rightarrow \{0, 1\}$ , data point  $(x, y, r)$  will manipulate its feature to

$$z^*(x, r; f) = \arg \max_{z \in X} [r \cdot \mathbb{I}(f(z) = 1) - c(x - z)]$$

This is a game now!

# A Unified Model of Strategic Classification

## The Strategic Classification Problem

**Input:**  $n$  uncontaminated training data  $(\mathbf{x}_1, y_1, r_1), \dots, (\mathbf{x}_n, y_n, r_n) \sim \mathcal{D}$

**Learning goal:** compute a classifier  $f$  that predicts well based only on the manipulated feature  $\mathbf{z}^*(\mathbf{x}, r; f)$

$$\mathbf{z}^*(\mathbf{x}, r; f) = \arg \max_{\mathbf{z} \in X} [r \cdot \mathbb{I}(f(\mathbf{z}) = 1) - c(\mathbf{x} - \mathbf{z})]$$

# A Unified Model of Strategic Classification

## The Strategic Classification Problem

**Input:**  $n$  uncontaminated training data  $(x_1, y_1, r_1), \dots, (x_n, y_n, r_n) \sim \mathcal{D}$

**Learning goal:** compute a classifier  $f$  that predicts well based only on the manipulated feature  $z^*(x, r; f)$

Some notably special cases

- ✓  $r \equiv 0 \rightarrow$  classic classification
- ✓  $r \equiv 1 \rightarrow$  strategic classification (cf. [Hardt et al.'16])
- ✓  $r = -y \rightarrow$  adversarial classification (cf. [Cullina et al.'18])
- ✓  $\text{sgn}(r) = -y \rightarrow$  generalized adversarial classification

Remark: manipulation here does not change true label

# Strategic Classifications are Everywhere

## ➤ University admissions

- Students academic records are selectively revealed
- Heterogeneous preferences: not all students prefer the same school

The screenshot shows the homepage of University World News. At the top, the logo "University World News" is displayed in large blue letters, with the subtitle "THE GLOBAL WINDOW ON HIGHER EDUCATION" in smaller red letters below it. To the right of the logo is a blue circular icon with the letters "w" and "n". Below the header, there is a navigation bar with links: "Global Edition", "Africa Edition", "Asia Hub", "Transformative Leadership", "Special Reports", and "Events". A purple banner features the text "MA in Higher Education Management" and "A unique programme for higher education leaders". Next to it is another purple banner with "Apply now for May 2020" and the logo of the University of Bath School of Management. The main content area includes a "GLOBAL" section with the headline "How will artificial intelligence change admissions?", written by Marguerite J Dennis on 26 October 2018. There are social media sharing buttons for LinkedIn, Twitter, and Facebook.

# Strategic Classifications are Everywhere

- University admissions
  - Students academic records are selectively revealed
- Classify loan lending decisions
  - Borrowers will selectively report their features
  - Heterogeneous preferences: not all borrowers prefer the same loan



30-Year  
Fixed

V.S

15-Year  
Fixed

# Strategic Classifications are Everywhere

- University admissions
  - Students academic records are selectively revealed
- Classify loan lending decisions
  - Borrowers will selectively report their features
- We decide which restaurants to go based on Yelp rating
  - Yelp may selectively show you the ratings
- Hiring job candidates in various scenarios

# Strategic Classifications are Everywhere

- University admissions
  - Students academic records are selectively revealed
- Classify loan lending decisions
  - Borrowers will selectively report their features
- We decide which restaurants to go based on Yelp rating
  - Yelp may selectively show you the ratings
- Hiring job candidates in various scenarios
- Note: this problem deserves study even you do classification manually instead of using an automated classifier

# Manipulation in Stock Trading

**Spoofing** is the practice of submitting large **spurious** buy (sell) orders to create artificial demand (supply) and mislead other traders.

UBS, Deutsche Bank and HSBC to pay millions in spoofing settlement, CFTC says

- Deutsche Bank will pay \$30 million, UBS \$15 million and HSBC \$1.6 million to settle civil charges that some of their traders engaged in spoofing in the precious metals market.
- The CFTC charged six individuals, and the Department of Justice charged eight with crimes related to deceptive trading in a wide-ranging investigation.

Liz Moyer

Published 2:29 PM ET Mon, 29 Jan 2018 | Updated 8:32 AM ET Wed, 31 Jan 2018

**CNBC**



Luke MacGregor | Reuters

## Flash Crash Trader E-Mails Show Spoofing Strategy, U.S. Says

by Tom Schoenberg, Suzi Ring, Janan Hanna

September 3, 2015 – 4:03 PM EDT Updated on September 4, 2015 – 9:32 AM EDT

f t ↗



■ Navinder Singh Sarao leaves Westminster Magistrates' Court in London, on Friday, Aug. 28, 2015. Photographer: Chris Ratcliffe/Bloomberg

- ▶ Failed orders are 'costing me,' Sarao said to tell programmer
- ▶ Indictment's new details seen bolstering U.S. extradition case

US seals first prosecution against stock market trader for 'spoofing'

A jury convicts Michael Coscia on six charges of commodities fraud and six charges of spoofing, all of the charges he faced

f 9 t 0 in 4 ↗ 13 Email



Prosecutors said Michael Coscia wanted to lure other traders to markets by creating an illusion of demand so that he could make money on smaller trades Photo: AP

By Reuters  
11:48PM GMT 03 Nov 2015

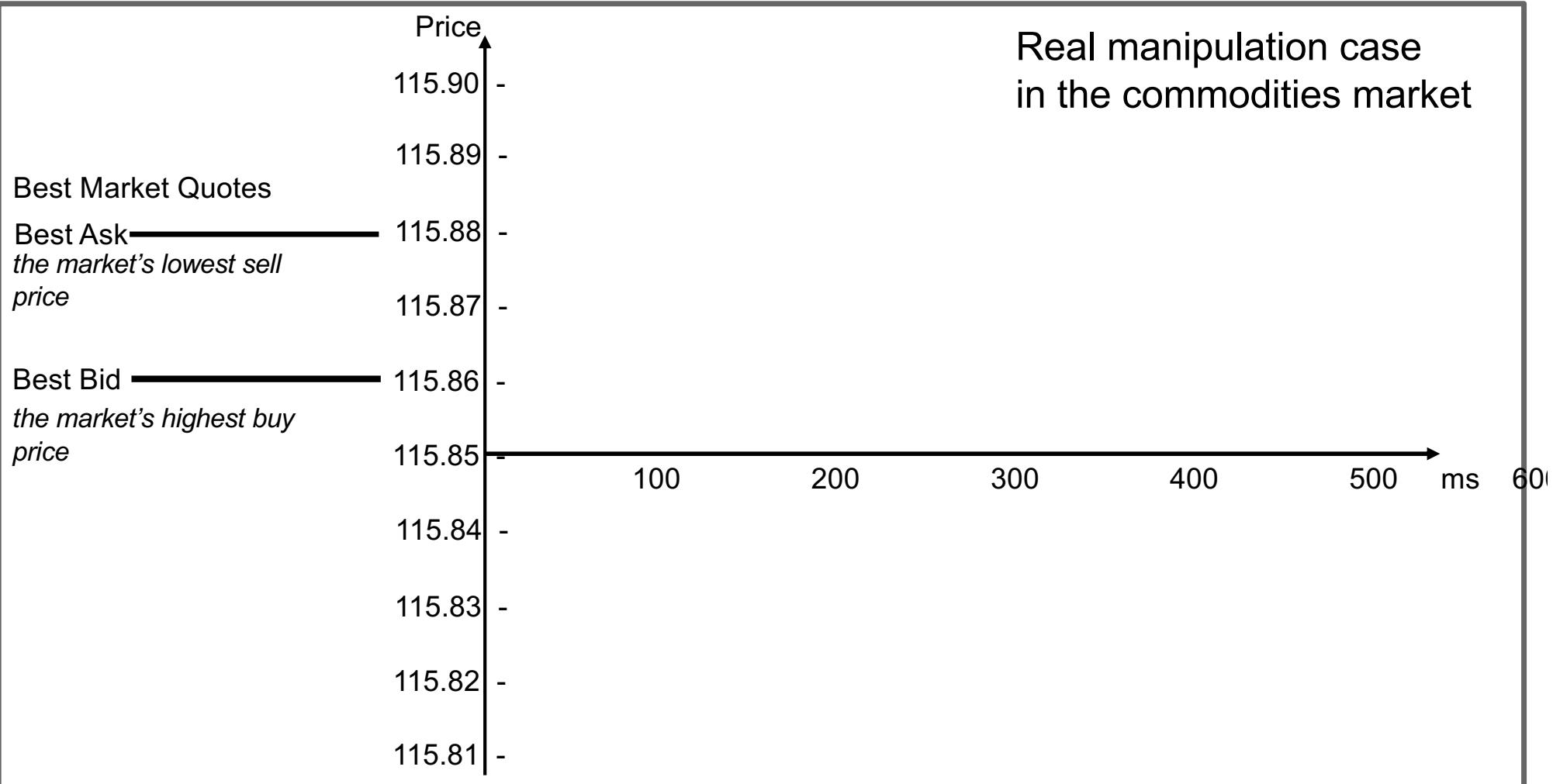
A US jury has found high-frequency trader Michael Coscia guilty of commodities fraud and "spoofing" in the US government's first criminal

## Useful sources

<https://www.fca.org.uk/publication/final-notices/coscia.pdf>

<https://www.fca.org.uk/publication/final-notices/coscia-appendix-1a.pdf>

# Manipulation in Stock Trading



Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Manipulation in Stock Trading



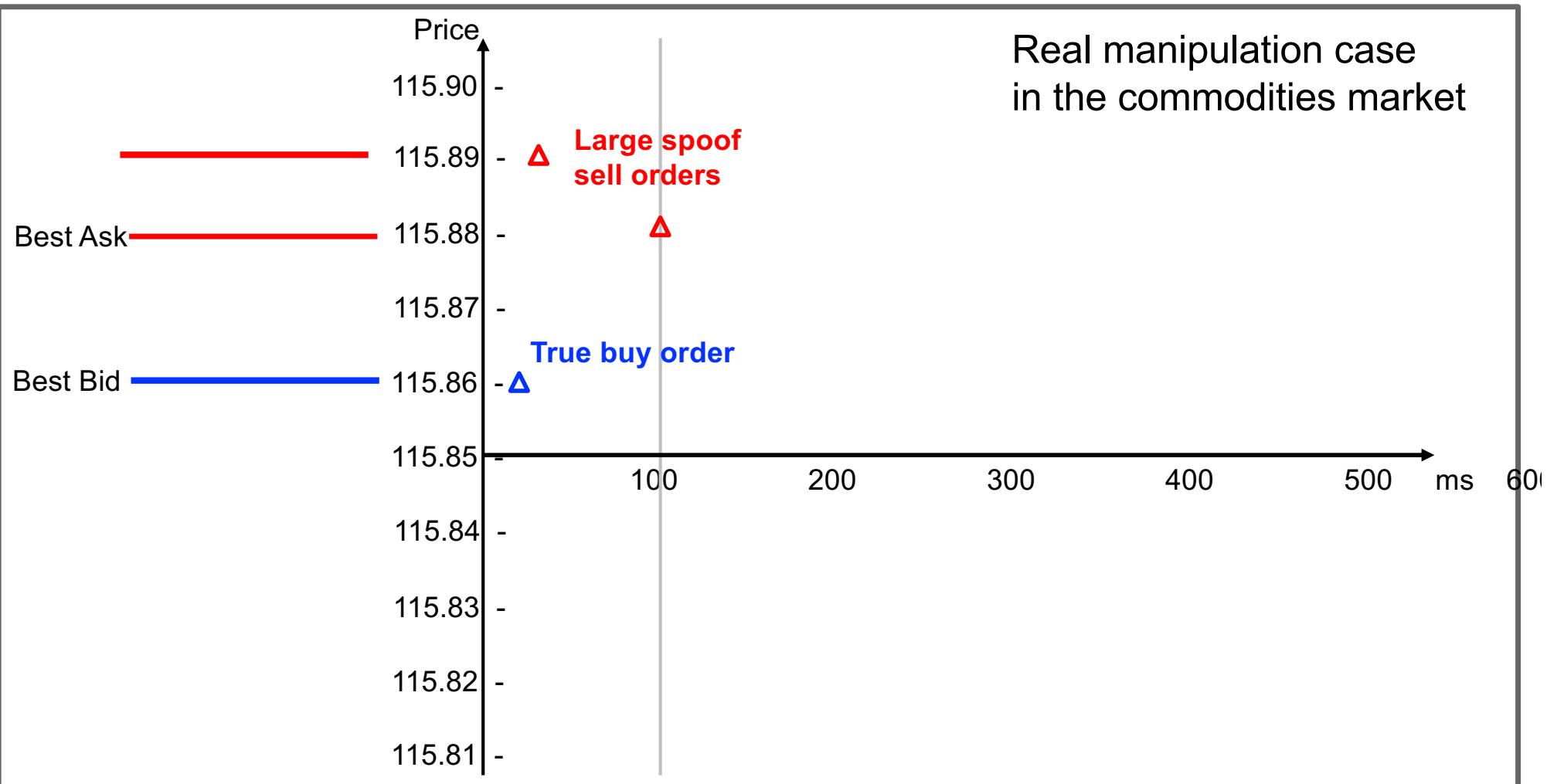
Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Manipulation in Stock Trading



Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Manipulation in Stock Trading



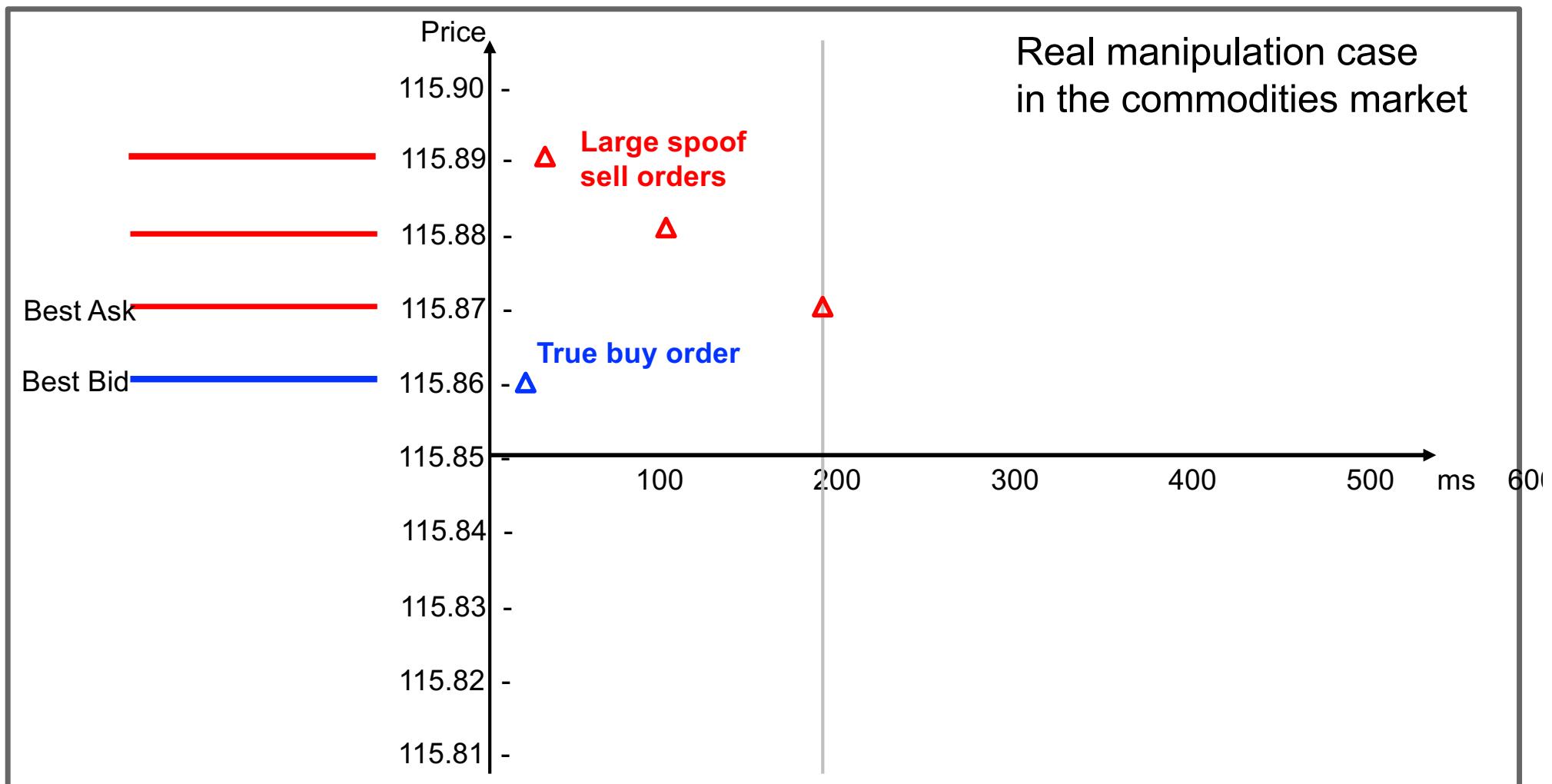
Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Manipulation in Stock Trading



Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Manipulation in Stock Trading



Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Manipulation in Stock Trading



Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

# Outline

- Introduction to Strategic Classification
- Learnability and Computability of Strategic Classifiers
- Beyond Classification

# Recall...

## The Strategic Classification Problem

**Input:**  $n$  uncontaminated training data  $(x_1, y_1, r_1), \dots, (x_n, y_n, r_n) \sim \mathcal{D}$

**Learning goal:** compute a classifier  $f$  that predicts well based only on the manipulated feature  $z^*(x, r; f)$

But will this general problem still be learnable?



In classic ML setup

- ✓ Learnability (sample complexity) of a hypothesis class is governed by its **VC-dimension**
- ✓ The learning algorithm is the **empirical risk minimization (ERM)**

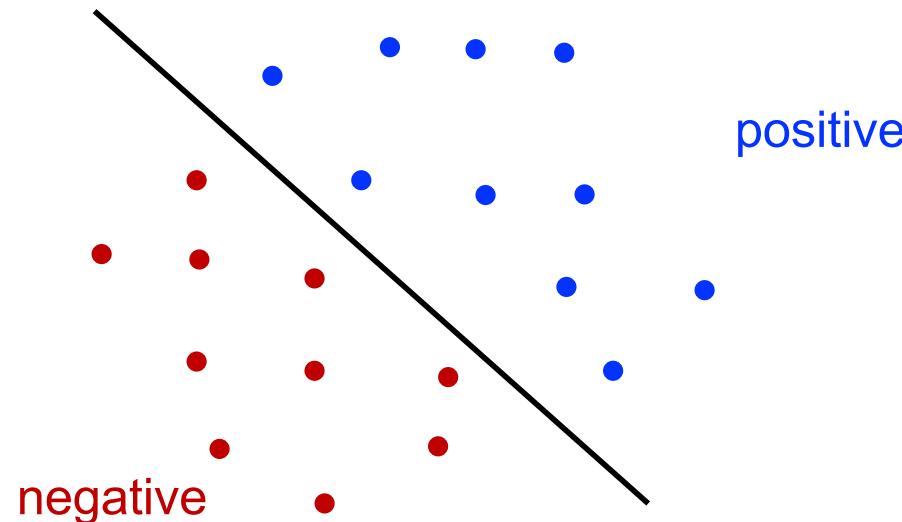
# The Learnability of Strategic Classifiers

... is governed by a variant, coined **strategic VC-dimension (SVC)**

# The Learnability of Strategic Classifiers

... is governed by a variant, coined **strategic VC-dimension (SVC)**

- Defined over the **equilibrium** of the classification outcome



Challenge is to characterize the classification outcomes under strategic manipulation

# The Learnability of Strategic Classifiers

... is governed by a variant, coined **strategic VC-dimension (SVC)**

**Theorem.** Any strategic classification instance is (PAC) learnable via a **strategic variant of ERM**, with sample complexity

$$n(\epsilon, \delta) = \Theta\left(\frac{SVC + \log(1/\delta)}{\epsilon^2}\right)$$

where  $\epsilon$  is accuracy loss and  $\delta$  is the failure probability.

Unifies learnability of all previous special cases

- Generalizes the fundamental theorem of classic PAC learning ( $r = 0$ )
- Recovers the main sample complexity result of [Hardt et al.'16] with  $r = 1$ , for which we show their  $SVC = 3$
- Generalizes learnability of adversarial classification [Cullina et al.'18] with  $r = -y$

# The Learnability of Strategic Classifiers

... is governed by a variant, coined **strategic VC-dimension (SVC)**

**Theorem.** Any strategic classification instance is (PAC) learnable via a **strategic variant of ERM**, with sample complexity

$$n(\epsilon, \delta) = \Theta\left(\frac{SVC + \log(1/\delta)}{\epsilon^2}\right)$$

where  $\epsilon$  is accuracy loss and  $\delta$  is the failure probability.

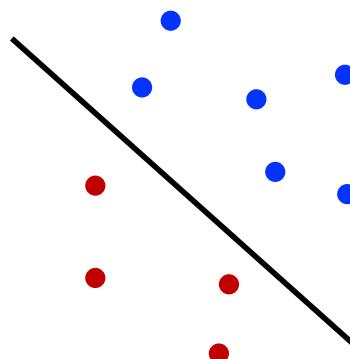
Implies learnability of many new setups with heterogeneous data preferences: loan approval, student admission, classifying job candidates,...

# Strategic Empirical Risk Minimization

**Input:**  $n$  uncontaminated training data  $(\mathbf{x}_1, y_1, r_1), \dots, (\mathbf{x}_n, y_n, r_n) \sim \mathcal{D}$   
**Output:** a classifier  $h(\mathbf{z})$  that minimizes “strategic risk”

$$\begin{aligned} \text{SERM: } & \min_h \sum_{i=1}^n \mathbb{I}[h(\mathbf{z}_i) = y_i] \\ \text{s. t. } & \mathbf{z}_i = \arg \max_{\mathbf{z} \in X} [r_i \cdot \mathbb{I}(h(\mathbf{z}) = 1) - c(\mathbf{x}_i - \mathbf{z}_i)], \forall i \end{aligned}$$

- Strategic ERM minimizes empirical risk by accounting for manipulation



# Strategic Empirical Risk Minimization

**Input:**  $n$  uncontaminated training data  $(\mathbf{x}_1, y_1, r_1), \dots, (\mathbf{x}_n, y_n, r_n) \sim \mathcal{D}$   
**Output:** a classifier  $h(\mathbf{z})$  that minimizes “strategic risk”

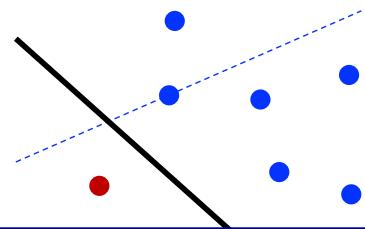
$$\begin{aligned} \text{SERM: } & \min_h \sum_{i=1}^n \mathbb{I}[h(\mathbf{z}_i) = y_i] \\ & \text{s. t. } \mathbf{z}_i = \arg \max_{\mathbf{z} \in X} [r_i \cdot \mathbb{I}(h(\mathbf{z}) = 1) - c(\mathbf{x}_i - \mathbf{z}_i)], \forall i \end{aligned}$$

- Strategic ERM minimizes empirical risk by accounting for manipulation
- This is a bi-level optimization problem (a Stackelberg game with  $n$  followers)
  - Difficult to solve due to non-smooth objective functions
  - An useful approach is to relax objective to use convex loss

# Instantiation to Linear Classification

**Theorem.** The SVC of  $d$ -dimensional linear classifiers is at most  $d + 1$ .

- $d + 1$  is the VC of linear classifiers in classic setup
- Learning strategic linear classifiers is no harder statistically
- Why can SVC be smaller than VC dimension?



## Lessons Learned

Flexibility of manipulating features reduces the “richness” of possible classification outcomes, and may make it easier to learn

# Computing Strategic Linear Classifier

Unfortunately, not all news is good...

**Theorem.** Strategic empirical risk minimization (ERM) is NP-hard for linear classification.

# Computing Strategic Linear Classifier

Unfortunately, not all news is good...

**Theorem.** Strategic empirical risk minimization (ERM) is NP-hard for linear classification. But, strategic ERM can be solved in polynomial time when the instance is *essentially adversarial*.

$$\min^- = \min\{r : (\mathbf{x}, y, r) \text{ with } y = -1\} \text{ and}$$

$$\max^+ = \max\{r : (\mathbf{x}, y, r) \text{ with } y = +1\}$$

Essentially adversarial if  $\min^- \geq \max^+$

# Outline

- Introduction to Strategic Classification
- Learnability and Computability of Strategic Classifiers
- Beyond Classification

# Vignette I: Manipulation in Multi-Armed Bandits

Reward<sub>1</sub> ~



Reward<sub>2</sub> ~



Reward<sub>3</sub> ~



Reward<sub>4</sub> ~



# Vignette I: Manipulation in Multi-Armed Bandits

Search			
	Filter	Restaurants	Map
Reward <sub>1</sub> ~		<b>1. Julep's</b> 137 Reviews 1719 E Franklin St, Richmond Southern	0.2 mi \$\$\$
Reward <sub>2</sub> ~		<b>2. Addis Ethiopian Restaurant</b> 54 Reviews 9 N 17th St, Richmond Ethiopian	0.2 mi \$\$
Reward <sub>3</sub> ~		<b>3. East Villa Restaurant</b> 10 Reviews 1900 E Main St, Richmond Chinese	0.2 mi \$
Reward <sub>4</sub> ~		<b>4. Lulu's Restaurant</b> 87 Reviews 21 N 17th St, Richmond	0.2 mi \$\$

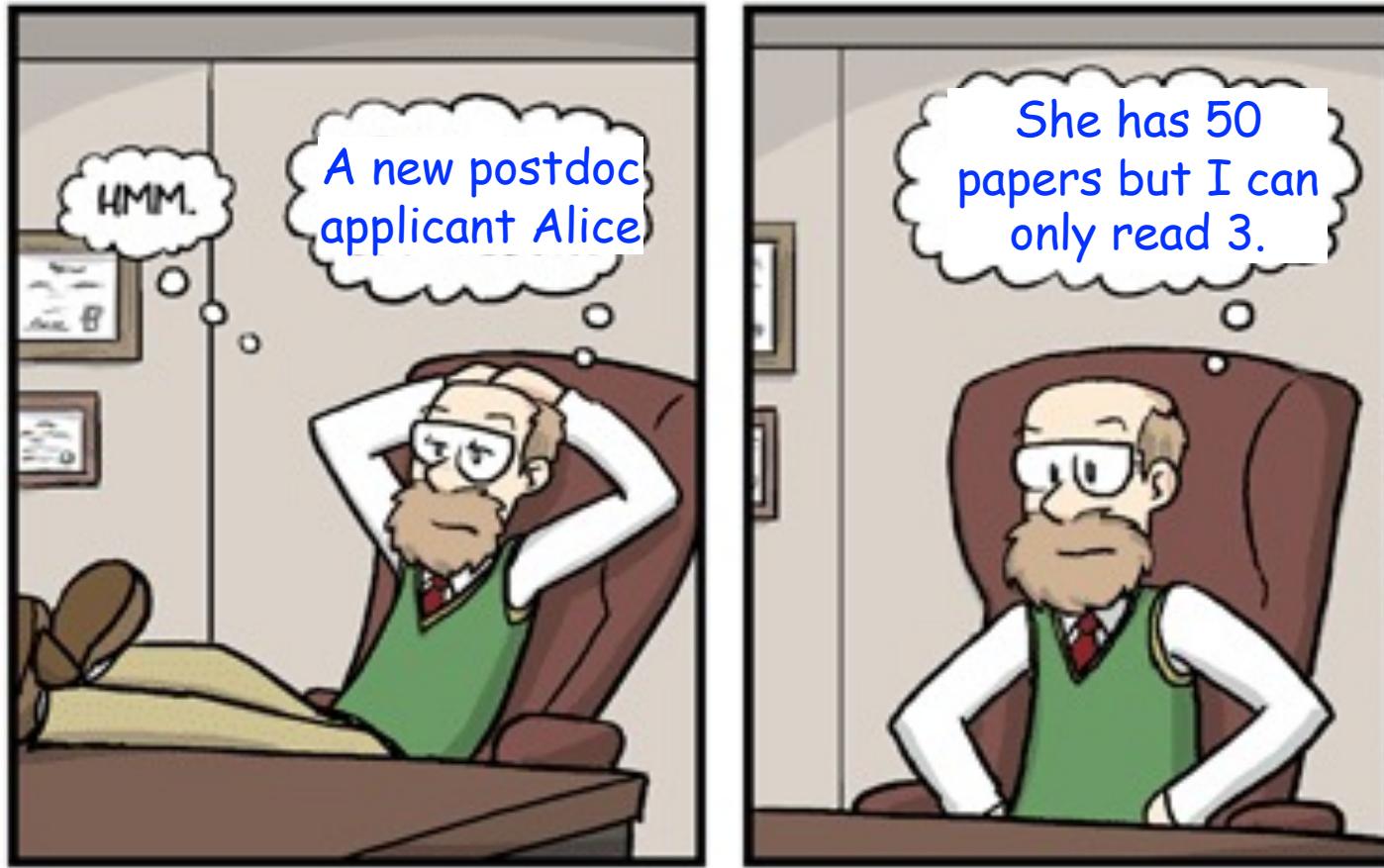
Each arm has incentives to manipulate its rewards to induce more pulls

# Vignette I: Manipulation in Multi-Armed Bandits

**Theorem.** Most standard stochastic bandit algorithms (including UCB,  $\epsilon$ -Greedy and Thomas Sampling) are all robust to selfish arms' strategic reward manipulation.

- A sharp contrast to adversarial reward attacks, which can ruin all these algorithms easily

## Vignette 2: Distinguish Distributions from Strategic Samples



The Trouble of Professor Bob

Paper: [When Samples Are Strategically Selected](#)

## Vignette 2: Distinguish Distributions from Strategic Samples



Current postdoc Charlie is happy . . .

## Vignette 2: Distinguish Distributions from Strategic Samples



They know what each other is thinking...

## Vignette 2: Distinguish Distributions from Strategic Samples

- A **distribution**  $l \in \{g, b\}$  arrives, which can be good ( $l = g$ ) or bad ( $l = b$ )
- An **agent** has access to  $n$  i.i.d. samples from  $l$ , from which he chooses a **subset of exactly  $m$  samples** as his report
  - Agent's goal: persuade a **principal** to accept  $l$
- Principal observes agent's report, and decides whether to accept
  - Principal's goal: accept when  $l = g$  and reject when  $l = b$
  - Want to minimize her **probability of mistakes**

Other applications: e.g., deciding where to hold Olympics based on photographs of different city locations



vs



vs



Paper: [When Samples Are Strategically Selected](#)

## Vignette 2: Distinguish Distributions from Strategic Samples

- A **distribution**  $l \in \{g, b\}$  arrives, which can be good ( $l = g$ ) or bad ( $l = b$ )
- An **agent** has access to  $n$  i.i.d. samples from  $l$ , from which he chooses a **subset of exactly  $m$  samples** as his report
  - Agent's goal: persuade a **principal** to accept  $l$
- Principal observes agent's report, and decides whether to accept
  - Principal's goal: accept when  $l = g$  and reject when  $l = b$
  - Want to minimize her **probability of mistakes**

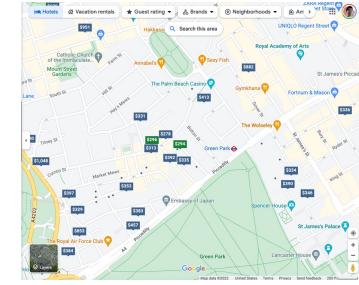
Other applications: e.g., deciding where to hold Olympics based on photographs of different city locations



vs



vs



Paper: [When Samples Are Strategically Selected](#)

# Concluding Remarks

- Very active research area, with motivations from numerous economic applications
- Strategic studies of classification (online or offline, training time vs testing time), regression, bandits, reinforcement learning...
  - Closely related to adversarial attack and algorithm robustness as well
- Today's lecture – manipulation does not change true nature
  - Next lecture – “strategic improvement”

# Thank You

Haifeng Xu

University of Chicago

[haifengxu@uchicago.edu](mailto:haifengxu@uchicago.edu)