CMSC 35401: The Interplay of Learning and Game Theory (Autumn 2022)

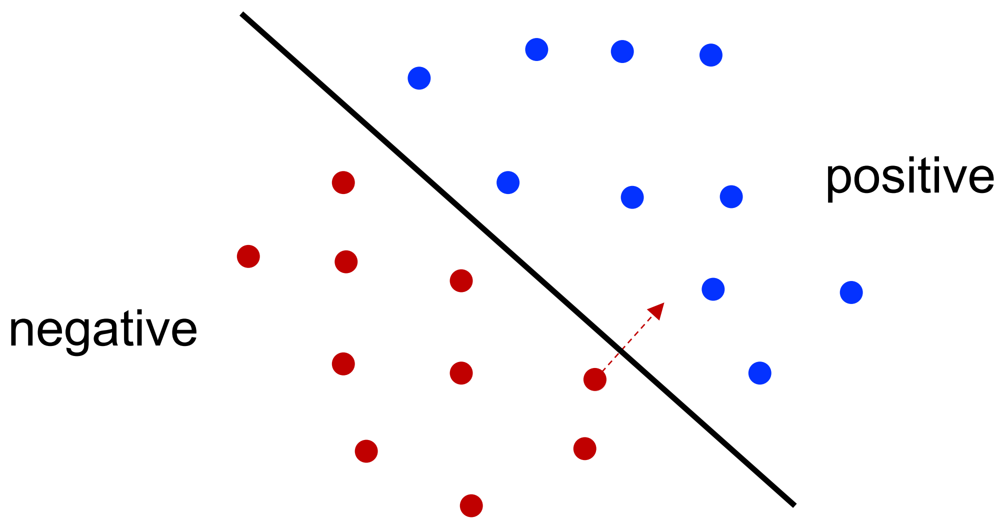
Learning From Strategic Data Sources

Instructor: Haifeng Xu

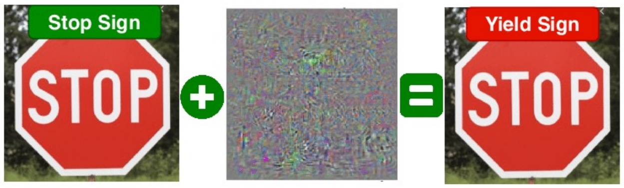
Outline

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

Classification

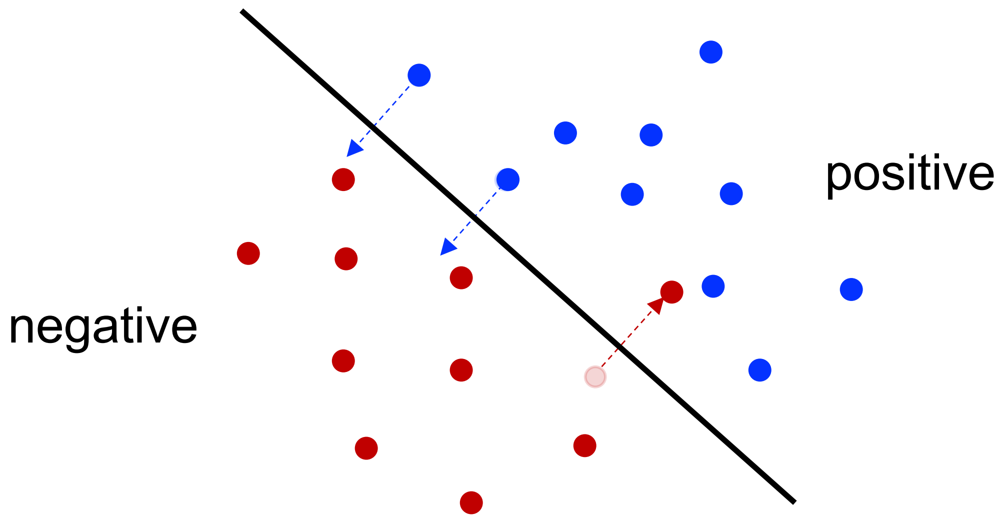


Data points' features may be manipulated

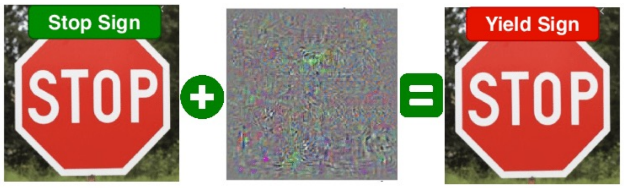


Adversarial attack

[Goodfellow et al.’15]
[Eykholt et al.’18]
[Cullina et al.’18]
……



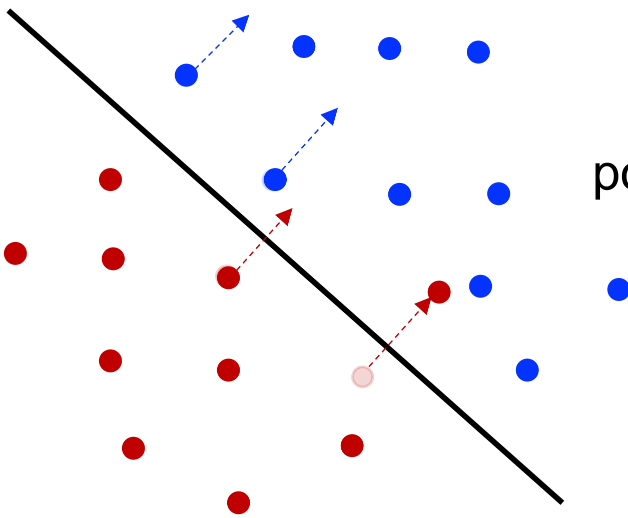
Data points’ features may be manipulated



Adversarial attack

Strategic manipulation

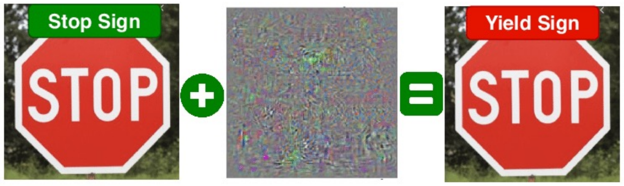
[Hardt et al.’16]
[Hu et al.’19]
[Ghalme et al.’21]
……



positive

negative

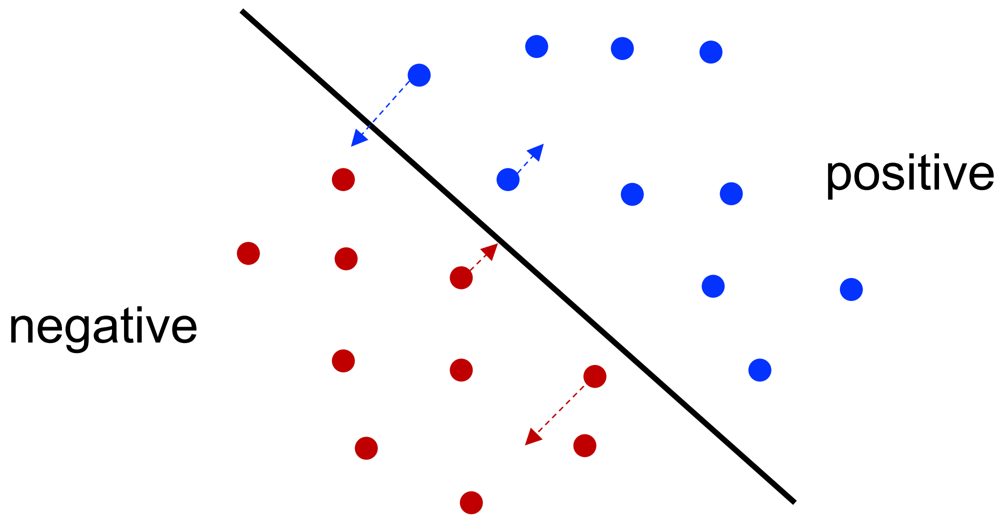
Data points’ features may be manipulated



Adversarial attack

[SVXY'21]

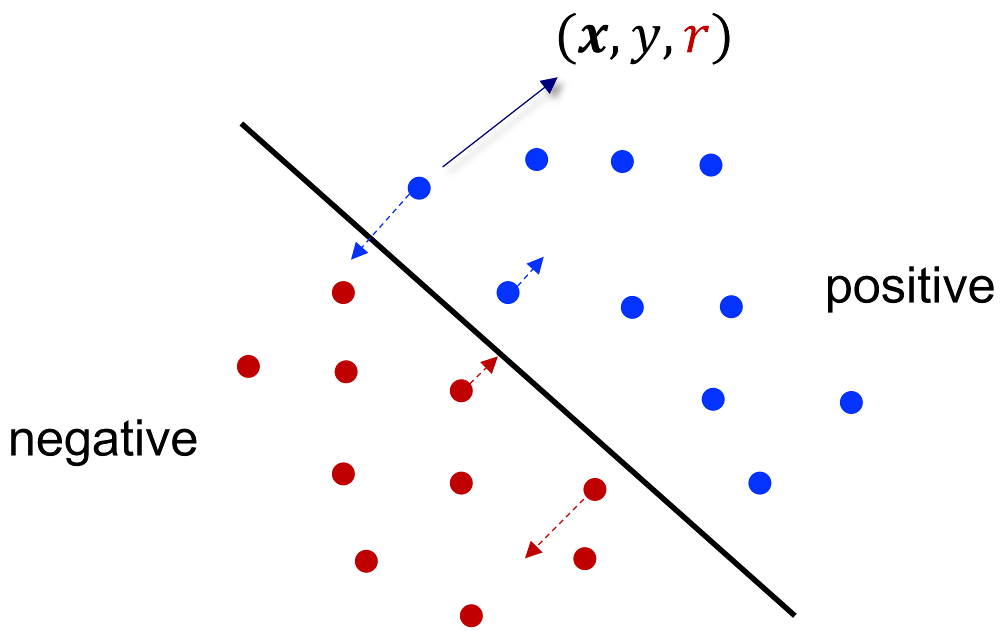
Strategic manipulation



Data points' features may be manipulated

A Unified Model of Strategic Classification

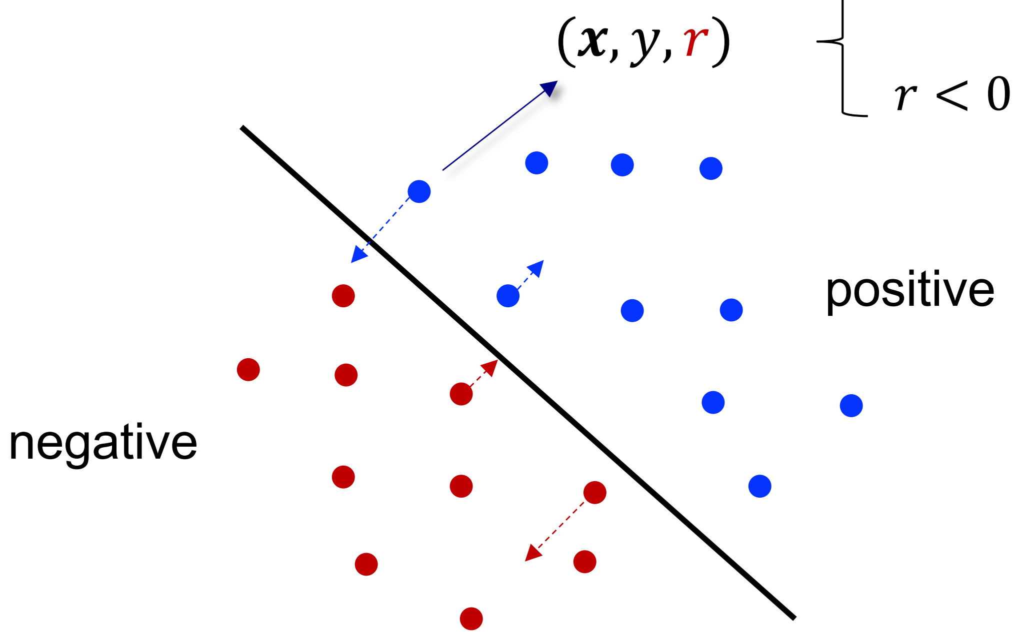
- Each data point is an individual agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive



A Unified Model of Strategic Classification

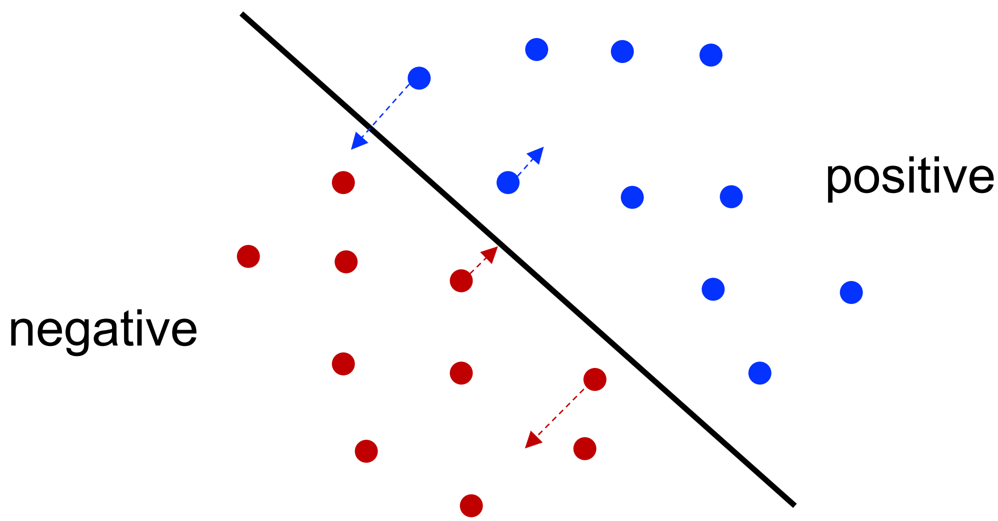
- Each data point is an individual agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive

$$\begin{cases} r > 0 \rightarrow \text{prefers positive} \\ r < 0 \rightarrow \text{prefers negative} \end{cases}$$



A Unified Model of Strategic Classification

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



A Unified Model of Strategic Classification

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

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

A Unified Model of Strategic Classification

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

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

This is a game now!

A Unified Model of Strategic Classification

The Strategic Classification Problem

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

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

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

A Unified Model of Strategic Classification

The Strategic Classification Problem

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

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

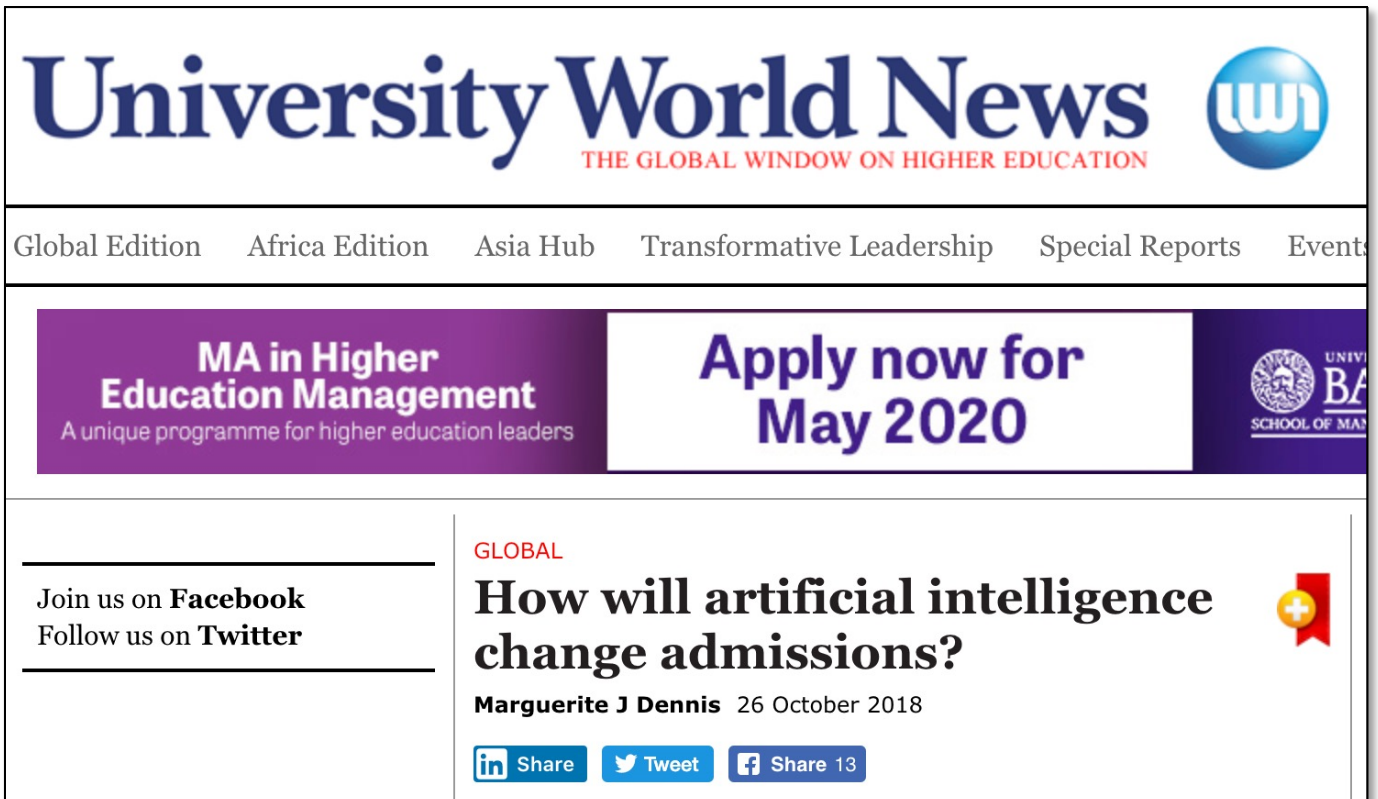
Some notably special cases

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


University World News
THE GLOBAL WINDOW ON HIGHER EDUCATION

Global Edition | Africa Edition | Asia Hub | Transformative Leadership | Special Reports | Events

Join us on **Facebook**
Follow us on **Twitter**

GLOBAL

How will artificial intelligence change admissions?

Marguerite J Dennis 26 October 2018

Share Tweet Share 13

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

Slides adapted [from Xintong Wang]

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 Tschoenberg22 Suzi Ring journosooz Janan Hanna

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

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

9 0 4 13 Email

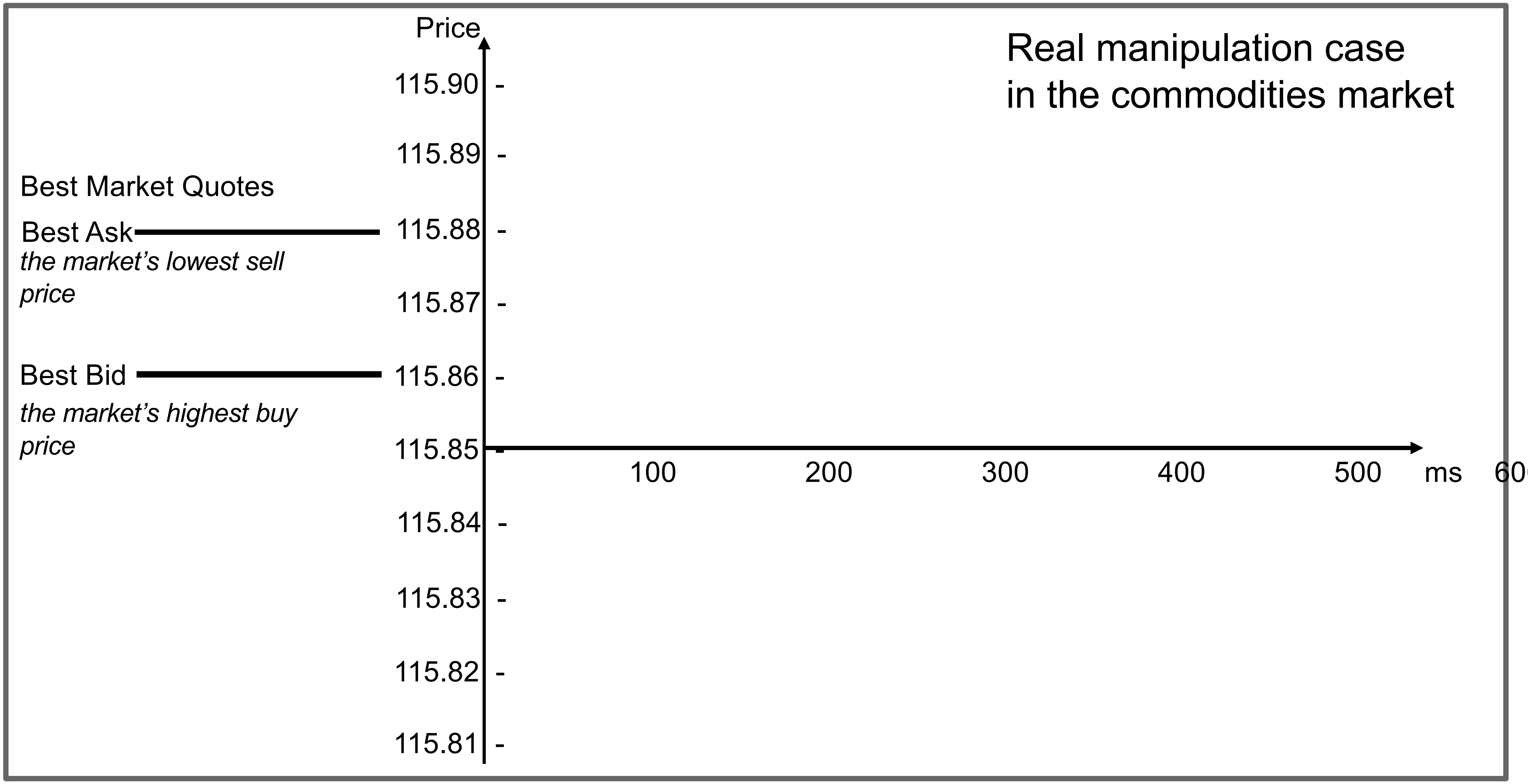
Prosecutors said Michael Coscia wanted to lure other traders to markets by creating an illusion of demand so that he could make money on smaller trades Photo: AP

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

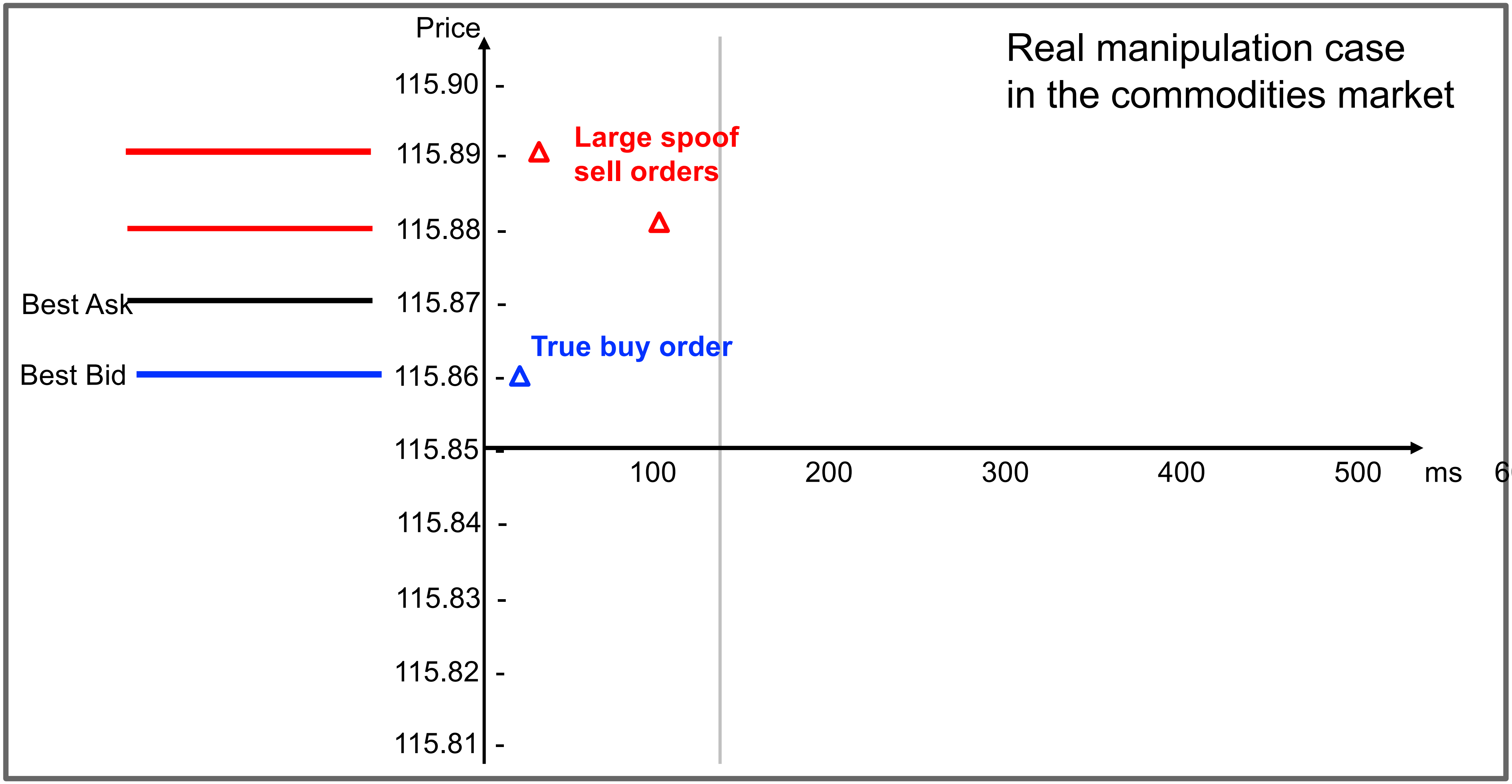
Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

Slides adapted from Xintong Wang

Manipulation in Stock Trading

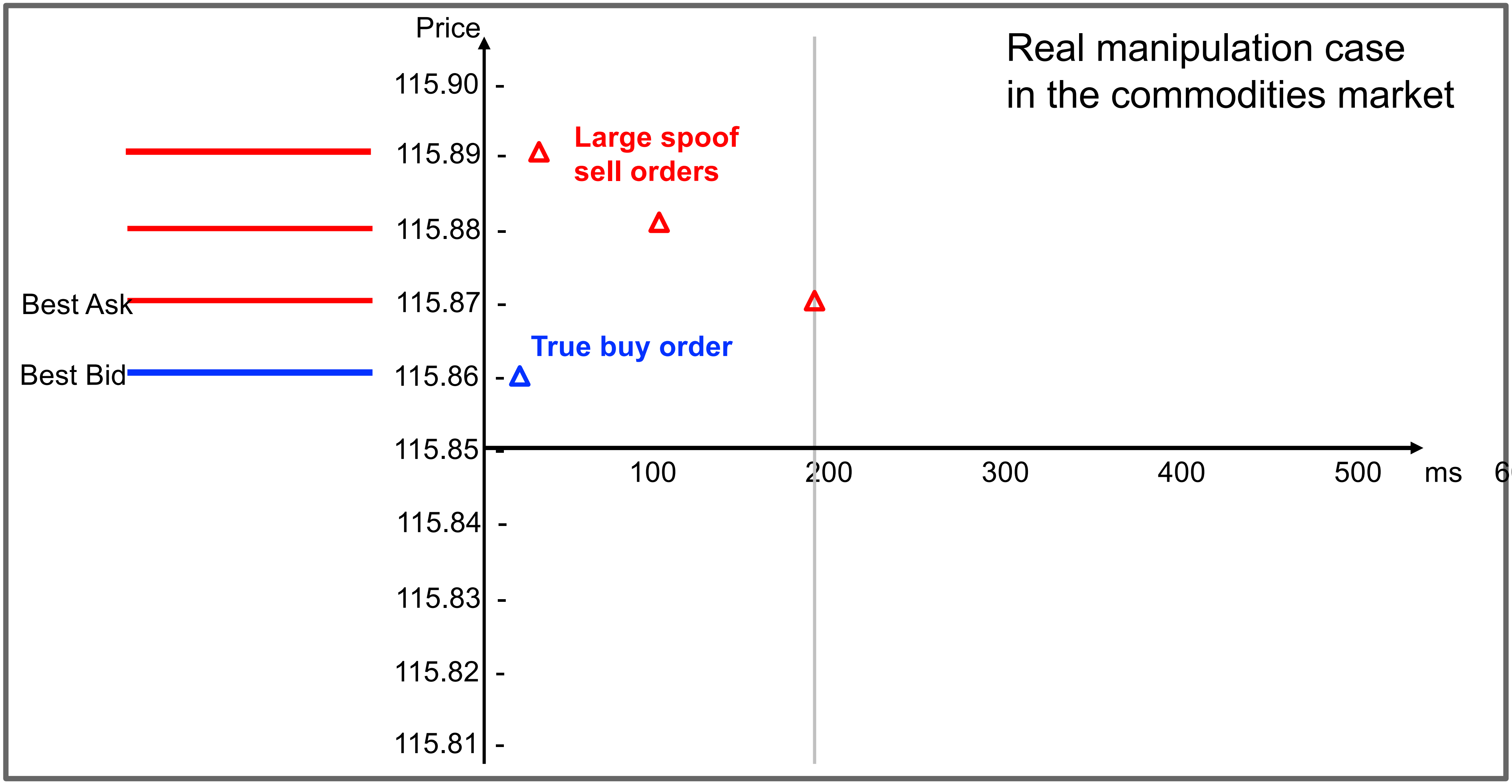
Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

Manipulation in Stock Trading



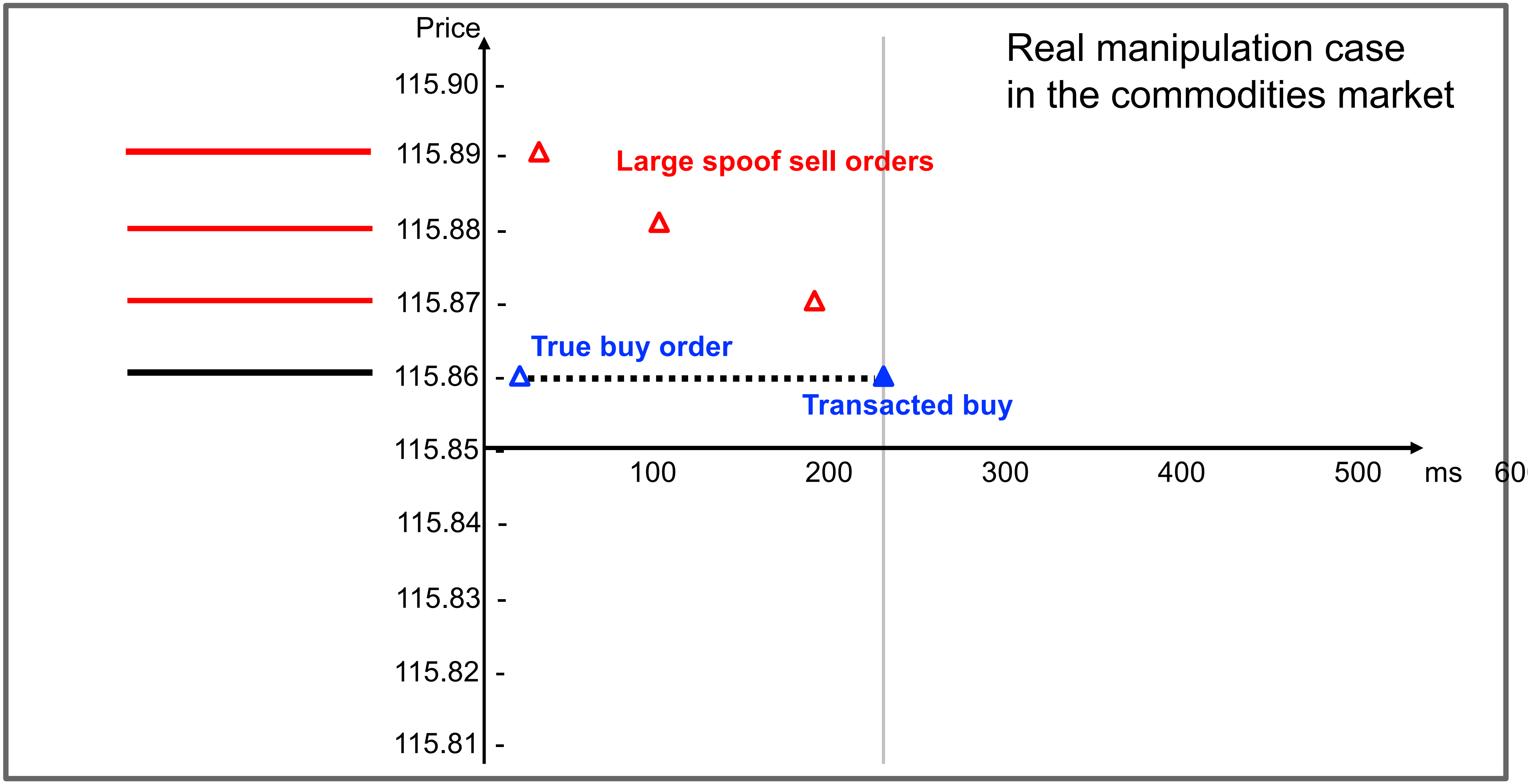
Source: Financial Conduct Authority, Animated Example of Mr. Coscia's Trading

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 $(\boldsymbol{x}_1, y_1, r_1), \cdots, (\boldsymbol{x}_n, y_n, r_n) \sim \mathcal{D}$

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

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)

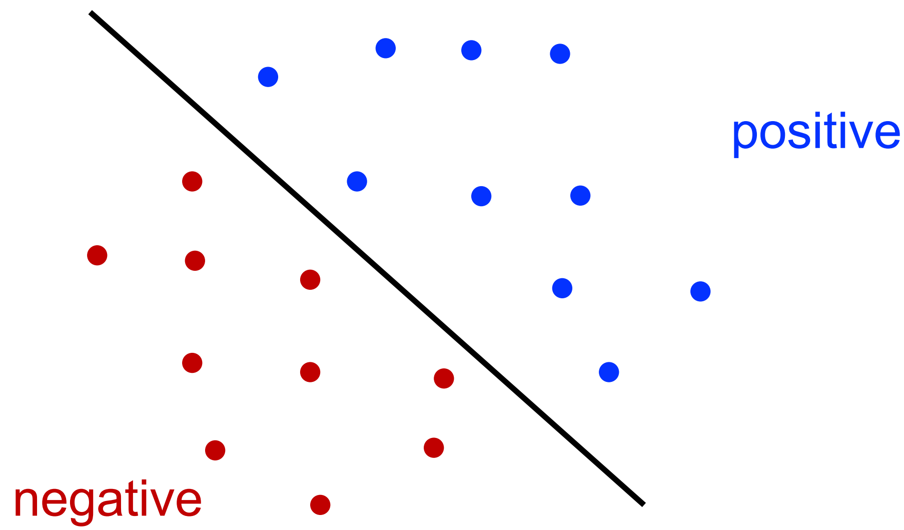
The Learnability of Strategic Classifiers

… 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

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

where ϵ is accuracy loss and δ 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

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

where ϵ is accuracy loss and δ 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

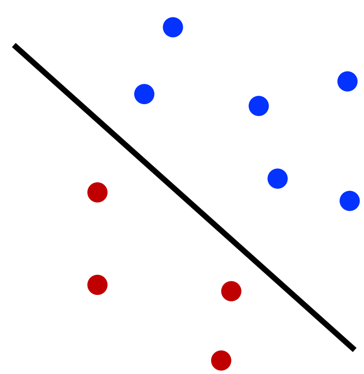
Input: n uncontaminated training data $(\boldsymbol{x}_1, y_1, r_1), \cdots, (\boldsymbol{x}_n, y_n, r_n) \sim \mathcal{D}$

Output: a classifier $h(\boldsymbol{z})$ that minimizes "strategic risk"

SERM: $$\min_h \sum_{i=1}^n \mathbb{I}[h(\boldsymbol{z}_i) \neq y_i]$$

$$\text{s.t.} \quad \boldsymbol{z}_i = \arg\max_{z \in X} \left[r_i \cdot \mathbb{I}(h(\boldsymbol{z}) = 1) - c(\boldsymbol{x}_i - \boldsymbol{z}_i) \right], \forall i$$

- Strategic ERM minimizes empirical risk by accounting for manipulation



Strategic Empirical Risk Minimization

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

SERM: $\min_{h} \sum_{i=1}^{n} \mathbb{I}[h(\boldsymbol{z}_i) \neq y_i]$

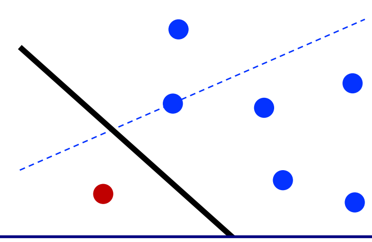
$$\text{s.t.} \quad \boldsymbol{z}_i = \arg\max_{\boldsymbol{z} \in X} \left[r_i \cdot \mathbb{I}(h(\boldsymbol{z}) = 1) \; - \; c(\boldsymbol{x}_i - \boldsymbol{z}_i) \right] \; , \forall i$$

- 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

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

[SVXY, ICML'21]

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*.

$$\mathit{min}^- = \min\{r : (\boldsymbol{x}, y, r) \textit{ with } y = -1\} \quad \textit{and}$$
$$\mathit{max}^+ = \max\{r : (\boldsymbol{x}, y, r) \textit{ with } y = +1\}$$

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

Outline

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

Vignette 1: Manipulation in Multi-Armed Bandits

$\text{Reward}_1 \sim$ 

$\text{Reward}_2 \sim$ 

$\text{Reward}_3 \sim$ 

$\text{Reward}_4 \sim$ 

Paper: [The Intrinsic Robustness of Stochastic Bandits to Strategic Manipulation]

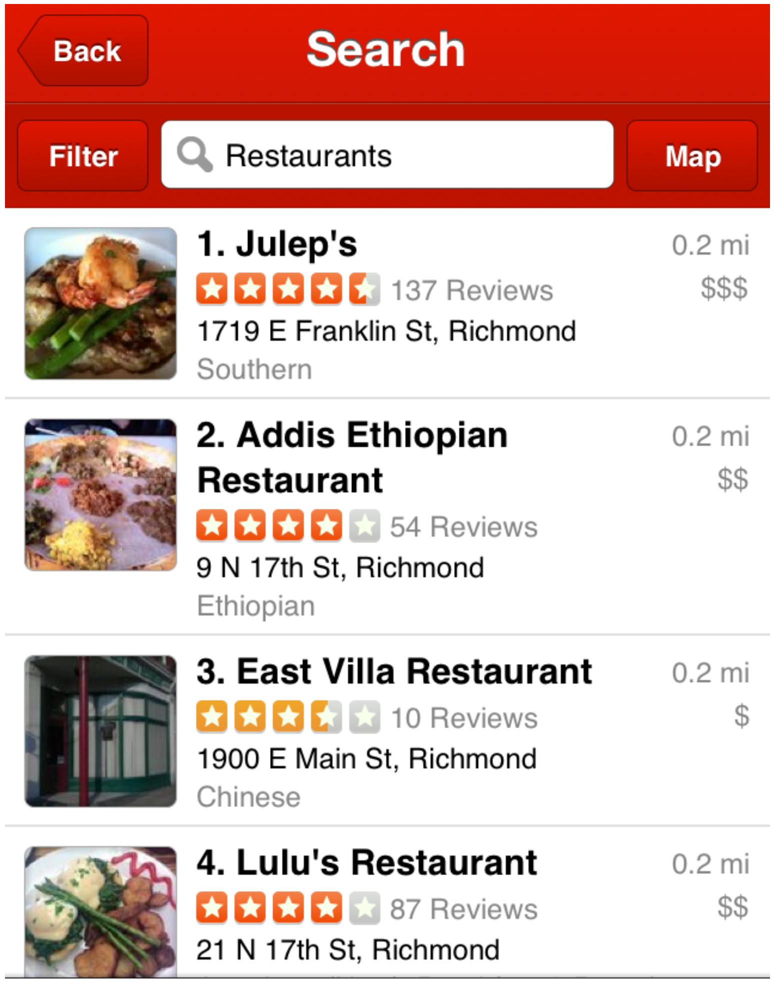
Vignette 1: Manipulation in Multi-Armed Bandits

$\text{Reward}_1 \sim$

$\text{Reward}_2 \sim$

$\text{Reward}_3 \sim$

$\text{Reward}_4 \sim$



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

Paper: [The Intrinsic Robustness of Stochastic Bandits to Strategic Manipulation]

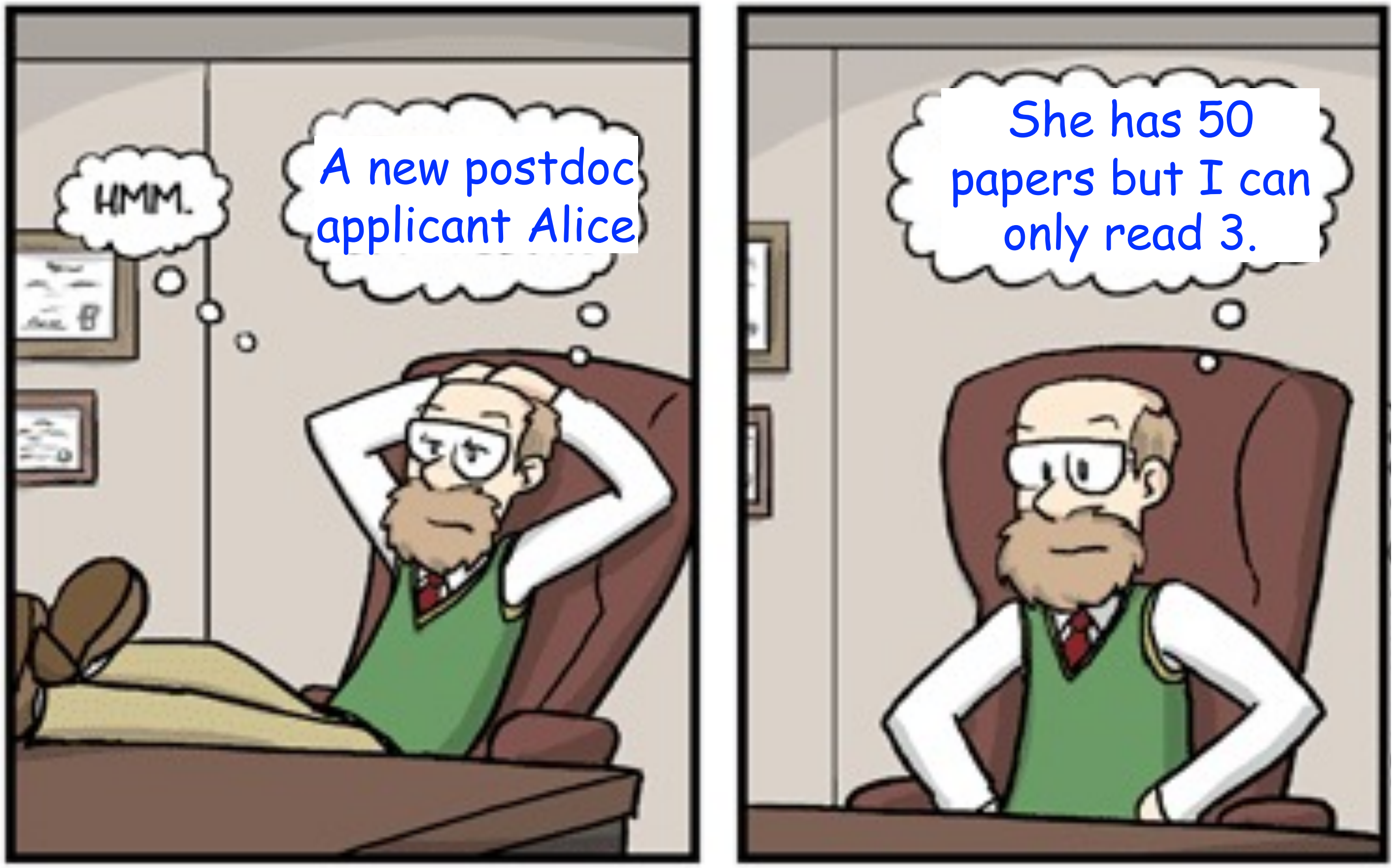
Vignette 1: Manipulation in Multi-Armed Bandits

Theorem. Most standard stochastic bandit algorithms (including UCB, ϵ-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

Paper: The Intrinsic Robustness of Stochastic Bandits to Strategic Manipulation

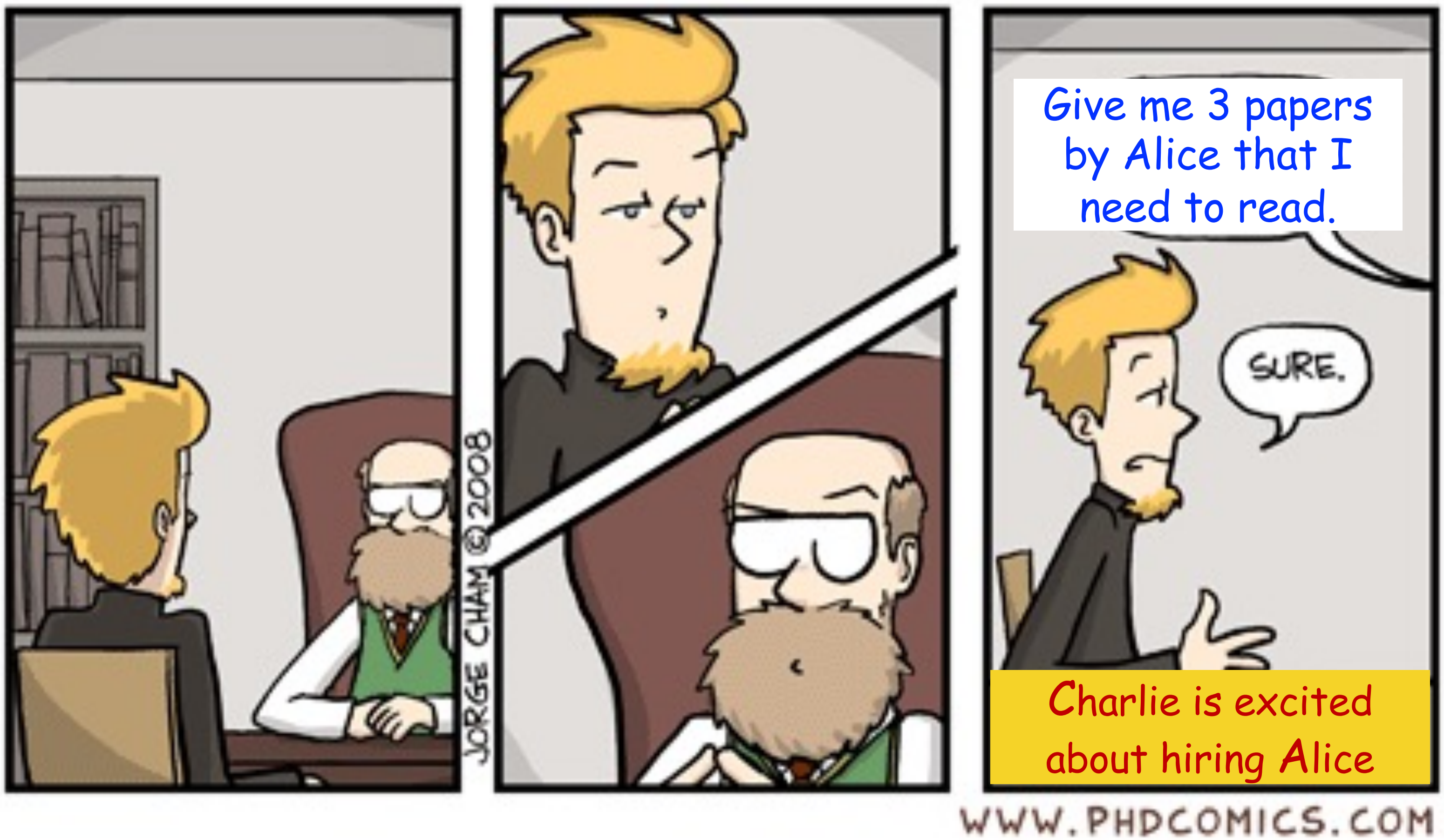
Vignette 2: Distinguish Distributions from Strategic Samples



The Trouble of Professor Bob

Paper: When Samples Are Strategically Selected

Vignette 2: Distinguish Distributions from Strategic Samples



Current postdoc Charlie is happy . . .

Paper: [When Samples Are Strategically Selected]

Vignette 2: Distinguish Distributions from Strategic Samples

I got to pick best 3 papers to persuade Bob, so that he will hire Alice.

Charlie shall pick best 3 papers by Alice – I need to calibrate for that

They know what each other is thinking…

Paper: [When Samples Are Strategically Selected]

Vignette 2: Distinguish Distributions from Strategic Samples

- A distribution $l \in \{g, b\}$ arrives, which can be a good distribution (g) or a bad one (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

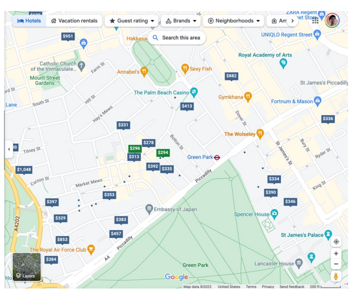
vs

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