“Will the Market Fix the Market?”

Eric Budish
AEA/AFA Joint Luncheon Talk
January 6, 2017
Chicago

ROTH INTRODUCTION

ALVIN ROTH: I’m glad to see so many of you here to hear Eric Budish speak to us this afternoon. I’d like to start by introducing the head table. There’s Bob Shiller, the president of the American Economic Association, from Yale University, on my far right; and then there’s Jessica Wachter from Penn; and Canice Prendergast from Chicago; and Oliver Hart, one of the most recent Nobel Prize winners, from Harvard; and Larry Glosten from Columbia; and David Scharfstein, the incoming president of the AFA, also from Harvard; and Stefan Nagel from the University of Michigan; and Cam Harvey from Duke, the current [AFA] president; and Eric Budish, our speaker. And on my left, there’s David Hirshleifer, from UC Irvine; and Richard Thaler, from Chicago; and Robert Hall, from Stanford; and Ariel Pakes, from Harvard; and Susan Athey, from Stanford; and John Campbell, from Harvard; and Pete Kyle, from Maryland; and Peter DeMarzo, from Stanford; and Heidi Williams, from MIT; and Patrick Bolton, from Columbia. We’re all here to help celebrate Eric today.

And so, I’d like to introduce you to Eric, who’s a professor at the University of Chicago Booth School of Business, and I’ve known him for a long time. He was hard to supervise as a PhD student, so he needed a very eclectic committee that included a computer scientist as well as economists. He studied matching, both in theory and in application, and has since done quite a bit of successful applied matching. One of his most successful and earliest matches was to Emma, who is here. But after that—I mean, that’s a hard act to follow—he talked about some stuff that may not look very practical. On the slide, I don’t know how much you can read, but he’s speaking about the combinatorial assignment problem. But, if you were a student at Wharton, you would be getting your courses through the system that he designed to match students to courses. What they say about it is that “Course Match applies the latest academic theory to optimize your schedule based on your preferences and course availability.” So, those of you who may have long ago gotten your MBAs at Wharton, think how much easier it would be if you were there now and you could use Eric’s system.

But what he’s been doing lately is casting a long shadow on the high-frequency trading arms race, and that’s what he’s going to talk to us about today. Eric?

BUDISH INTRODUCTORY REMARKS (02:42)

Al, thank you so much for the warm introduction and the invitation. And more importantly, thank you for being such a wonderful role model over the years, in the research that you do, in the way that you bring your research out from the ivory tower into the world, and as a person. You're the ultimate mensch and I'm so grateful for your mentorship. I'd also like to thank my family for being with me here today, and so many colleagues, collaborators and friends. It's a true honor and a thrill.

SLIDE 2: The Efficient Markets Hypothesis

I've followed the stock market since I was a little kid, as I imagine is the case for many of us here in this room. I distinctly remember reading the newspaper with my father, looking at stock quotes the morning after Black Monday in 1987, and we were debating whether the market was going to go up the next day, that Tuesday. As an optimistic, somewhat confident nine-year-old, I thought the market would go up. I'm pretty sure I told him to "buy IBM"—I checked; actually it went up 11 points that day, and another 7¾ points the day after!

As I'm sure is also the case with many of us here in the room today, the first serious academic idea I was exposed to was the efficient markets hypothesis, first via Burton Malkiel's classic *A Random Walk Down Wall Street*, and more rigorously in John Campbell's asset pricing class in graduate school.

Gene Fama famously distinguished between three forms of the efficient markets hypothesis: the weak form, the semi-strong form, and strong form. The weak form says you can't beat the market—make risk-adjusted excess returns—using just past prices as your information. The semi-strong form says you can't beat the market using both past prices and other forms of public information, like dividends or company earnings. The strong form says you can't beat the market even using private information. Fama concluded that there was evidence against this strong form, but that the weak and semi-strong forms held up pretty well. The plain-English translation is that to beat the market you have to know something that the rest of the market doesn't know. My market call as a nine-year-old was <u>not</u> skill.

SLIDE 3: Modern Understanding of the EMH

Our modern understanding of market efficiency, as summarized by the 2013 Nobel Committee, is that stock prices are "next to impossible" to predict in the short run, but that there is meaningful predictability in the longer run, for instance using the market's price-earnings ratio, or by favoring value stocks over growth stocks.

The main debate is over how to interpret this long-run predictability: to what extent does it reflect variation in risk, across time and across stocks, that we may not fully understand but that is perfectly consistent with market efficiency, and to what extent does it reflect behavioral biases which are not consistent with market efficiency? That the market's hard to beat in the short run, on the other hand, is relatively uncontroversial—the Nobel committee even writes that “such a situation would reflect a rather basic malfunctioning of the market mechanism.”

SLIDE 4: The HFT Arms Race

In 2010, a company called Spread Networks invested $300 million to dig a high-speed fiber optic cable connecting markets in New York and markets in Chicago.

The important feature of this cable is that it was dug in a straight line, whereas previous cables zigzagged around mountains, along railroad tracks, and so forth.

The straightness of this cable shaved round-trip data transmission time between New York and Chicago by 3 milliseconds. That’s 3 thousandths of a second, or roughly 1/100th of the time it takes to blink your eye.

Industry observers described 3ms as an “eternity.” They joked at the time that someone was going to dig a tunnel so it would go through the earth, rather than around the earth, to get an even straighter line.

And this joke sort of became true! The Spread cable quickly became obsolete; no one dug a tunnel, but microwaves are another way to get a faster line between Chicago and New York, because light travels faster through air than through glass.

When I first read about this cable—and I still remember this moment kind of viscerally—I really didn’t understand it. It was pretty obvious that it was some kind of arms race for speed, but what wasn’t at all obvious to me—and if this makes me sound unsophisticated, so be it—was what the speed was for.

Three milliseconds seemed too short to be about fundamentals—again, this is 1/100th of the blink of an eye. Companies issue their earnings releases once a quarter; that’s every eight billion milliseconds, and usually after the market is closed.

My academic, efficient market instinct was to be deeply skeptical of trading strategies that are short-run and purely technical. When academics hear phrases like 200-day moving averages, or head-and-shoulders patterns, or “support points,” we kind of roll our eyes. As Burton Malkiel wrote, “Technical strategies are usually amusing, often comforting, but of no real value.”

SLIDES 5-6: The HFT Arms Race: Market Design Perspective (07:42)

My collaborators Peter Cramton, John Shim and I tried to make sense of this HFT arms race by approaching the topic from the perspective of market design.

What I mean by this is that we assume that participants in a market are acting rationally in their self-interest given market rules. But we wanted to take seriously the possibility that the market rules were themselves flawed, and do so by taking seriously institutional details at a quite detailed level because, all too often, such seemingly unimportant small details of the market end up having a dramatic effect on market performance. Al Roth has called this "economic engineering." I'll note parenthetically, especially since we're in Chicago, that even Milton Friedman himself emphasized the importance of getting the right rules of the game in place before letting market competition do the rest.

Indeed, our research identified a simple structural flaw in the design of modern financial exchanges—essentially, a glitch introduced in the transition from humans to computers. The flaw is that exchange computers treat time as a continuous variable, and process requests to trade serially—that is, one at a time in order of receipt. Essentially, when markets transitioned from being run by humans to being run by computers—which on the whole has been a good thing—we made a mistake and forgot to tell computers to put time into units.

[SLIDE 6] This flaw causes a violation of the efficient market hypothesis, built directly into the market design. In an efficient market, to make money, you either have to (i) take risk, or (ii) have private information. But what we're going to show is that "built in" to the market design are profitable arbitrage opportunities that are (i) riskless, and (ii) use only symmetric public information—information revealed to the whole market at exactly the same time, with economically obvious implications.

<u>I want to underscore this: the details of the market design cause a violation of efficient markets theory, built right in.</u>

These arbitrage opportunities aren't supposed to exist in a well-designed market, and they harm the market in two ways. First, they make markets less liquid—as you'll see, they are like a tax on liquidity provision. Second, they induce a never-ending arms race for speed.

Our paper then proposes an alternative market design that directly addresses the problem. The basic idea is pretty simple: it's to put time into units. Then, if multiple trade requests arrive at "the same time," award the trade to whoever offers the best price—run an auction.

This market design, which we call frequent batch auctions, fixes the failure of the efficient market hypothesis, and in doing so, improves liquidity and stops the arms race.

SLIDE 7: Plan for Talk (10:24)

In the first part of this talk, I want to tell you about this research with Peter Cramton and John Shim in some detail. I'll give you new empirical facts that show how the current market design behaves at high-frequency time horizons. I'll then go over a simple theory model that shows what's wrong with the current market design. Last, I'll show how moving from the current market design to our new design, from continuous serial to discrete batch, directly solves the problem.

I'll then organize the second part of the talk around the main question I've gotten about this research, across academia, industry, and regulators, and that's <u>will the market fix the market.</u> That is, will market forces on their own fix the problem we identified in our research, or is some kind of regulatory intervention necessary.

In this part of the talk I'll discuss some new research on the IO of stock exchange competition that is joint with Robin Lee and John Shim, and which provides a framework for thinking about the private and social returns to market design innovation. I'll also make a few brief remarks at the end that reflect on some of the lessons I've taken from what, to be frank, has been an unusual few years—as a young, then-untenured academic, I published a paper that caused me to find myself in the middle of a high-stakes, highly charged policy debate.

SLIDE 9: Brief Description of the Continuous Limit Order Book

Just to make sure we're all on the same page, let me spend one slide describing the current status quo market design used all around the world—in stock markets, futures markets, currency markets, and so forth— often called the "Continuous Limit Order Book." The basic building block of this market design is called the limit order; it specifies a price, quantity and direction. Traders submit limit orders to the market at any time during the trading day, and they can cancel their outstanding limit orders also at any time during the trading day. These orders and cancelations are processed by the exchange, really by an exchange computer, one at a time in order of receipt, so that is, as a *continuous-time, serial process.* Trade occurs whenever a new order is submitted that is either an offer to buy at a price higher than some outstanding offer to sell, or vice versa. For example, in the figure I have illustrated, an order to buy Microsoft at 62.36, or to sell at 62.35, would initiate a trade in the order book that we show there.

Our data is, roughly, the "play by play" at millisecond granularity of the continuous limit order book.

SLIDES 10-13: Market Correlations Break Down at High Frequency

I want to begin with some empirical facts that show that continuous-time markets don't work as we expect them to literally in continuous time. This is a plot of prices over the course of a trading day for the two most liquid assets that track the S&P 500 Index: the E-mini future that trades here in Chicago on the Mercantile Exchange and the SPY exchange fund that trades in New York on the various equities markets. This is millisecond-level direct-feed data, the same data that high-frequency trading firms subscribe to and parse in real time, and we analyze retrospectively.

Over the course of a trading day, these two assets, as you'd expect, are very highly correlated. They move together, which makes sense because they have an arbitrage relationship; they track the same index. [SLIDE 11] Here's an hour of data, [SLIDE 12] here's a minute of data, [SLIDE 13] and this is what the market looks like if you zoom into high frequency. This is a quarter of a second, 250 milliseconds, about a blink of an eye.

When you zoom into high frequency, the correlation between assets falls apart. The correlation is basically 1, which it's supposed to be, at a minute, an hour, or a day, but the correlation is basically 0 at a millisecond.

The reason this matters economically is that it creates obvious mechanical arbitrage opportunities. For example, when the price in Chicago jumps and the price in New York hasn't reacted yet, you buy cheap in New York and sell expensive in Chicago, or vice versa.

Now, the usual efficient market story with obvious arbitrages is that, once discovered, they get competed away. But I want to show you over the next several slides that that's not what happens under our current market design.

SLIDES 14-16: Arbs and Correlation Breakdown Over Time

We do find that the durations of arbitrage opportunities, how long they last, has come down significantly over time, from over 100ms at the start of our data in 2005 to about 7ms on average at the end of our data in 2011.

]SLIDE 15] The profits per opportunity, however, have stayed pretty flat, [SLIDE 16] and the frequency of opportunities, it does fluctuate over time, but the frequency is mostly driven by market volatility. When the market is jumping around a lot, like during the financial crisis, there's more arbitrage opportunities.

This figure [SLIDE 17] is a complementary way of showing the same phenomenon. On the right-hand side of the graph, you see the market getting faster year by year—this shows up as the 100ms correlation between assets getting higher and higher each year. But on the left-hand side of the graph, you can see that in all years, at high-enough frequency, the correlation is always approximately zero.

SLIDE 18: Latency Arb and Arms Race are "Constants" (15:20)

So, to summarize, competition does raise the bar for how fast you have to be to capture obvious arbitrage opportunities. But it doesn't seem to eliminate the arbs or even reduce their total size. This suggests that we should think of latency arbitrage and the resulting race for seed as something like a "constant" of the market design—I'm going to come back to this point in the theory.

SLIDE 19: Total Size of the Arms Race Prize

In our data, we compute the size of this one arbitrage opportunity to be about $75 million per year. We think this is an underestimate for reasons we go into in detail in the paper, both the way we treat transactions costs and some details of the way the CME provides data. But more important to emphasize is that this is really just the tip of the iceberg in the race for speed. This is where we could get data, but it's just the tip of the iceberg.

Conceptually, all continuous electronic markets around the world are similar, with lots of highly correlated instruments, and nothing in the market architecture to enable prices to move at exactly the same time.

SLIDES 20-26: Highly Correlated Pairs

Here are pairs of US Treasury Bond instruments, here are other equity indices, here are some currency pairs, here are commodities, oil and gas, and here in Al's honor is coffee, and here's simply a long list of obviously correlated pairs that we could write down without much effort.

SLIDE 27: Total Size of the Arms Race Prize

In addition to the highly correlated assets, there are in US equity markets arbitrage opportunities that are even simpler, because the same symbol trades on as many as 13 different exchanges and 50+ dark pools. These within-symbol arbitrage opportunities were at the heart of the Michael Lewis book *Flash Boys*.

There is a race to respond to public news, such as company news or government numbers, and this race to respond to public news was at the heart of the New York Attorney General's interest in high-frequency trading, what he called Insider Trading 2.0.

And then fourth, there is a race to the top of the book, that is an artifact of minimum price ticks.

I still don't think it's possible to put a precise number on the total value of the prize at stake in the speed race, given the data that's currently available to academics. <u>But commonsense extrapolation from our estimates and the other numbers that academic researchers have been able to put out for various aspects of the speed race suggests that the sums are quite substantial—easily single-digit billions per year in US equities alone. Michael Lewis reports on a figure from an industry study of as much as $20bn per year in US equities—which would be a bit more than 1 penny per share traded—though this is higher than can be extrapolated to based on the academically available data</u>.

If you extrapolate from US equities to all financial instruments around the world that trade on continuous limit order books, it's easy to get to double-digit of billions of dollars per year.

If you take an NPV of these annual amounts—and since latency arbitrage is kind of like a tax on trading, the size of the prize grows with market cap, so if you're thinking of a classic Gordon growth model, *r* minus *g* in the denominator is small—it's easy to get to in excess of $100bn NPV.

I don't think this is the #1 policy problem in finance, but it's big, and worth trying to solve.

SLIDE 29: Model: Preliminaries (18:30)

I now want to move on to the theory model. The model is really quite simple, and I think it's pretty teachable, so I want to go over the setup and main ideas in some detail. It's going to show, in I think a quite transparent way, that the mechanical arbitrages I documented empirically are "built in" to the market.

The model is a descendant of the famous Glosten Milgrom (1985) model. If anything, it's simpler in some ways.

There is a security, I'm going to call it *x*, that trades on a continuous limit order book. There is a public signal, we'll call it *y*, of the value of this security *x*.

And I'm going to make a purposefully strong assumption, which is that *x* is worth *y*. The fundamental value of *x* is *perfectly* correlated to the public signal *y*, and can always be costlessly liquidated at this value.

The goal of this assumption is to create a best case scenario for the performance of this market design, assuming away issues like asymmetric information and inventory costs that usually make liquidity provision expensive.

This public signal *y* jumps around; it's what's called a compound Poisson jump process, which is just a fancy way of saying that with some probability, denoted lambda_jump, jumps occur, and the jumps have distribution *J*. What matters is not whether the markets jump up or down, but the absolute value of jump sizes.

SLIDES 30-31: Players: Investors and Trading Firms

There's two kinds of players in the model, Investors and Trading Firms. Investors we can think of as representing end users of financial markets, like mutual funds. We are going to model them very simply. They arrive to the market needing to either buy or sell 1 unit of *x*, with arrival rate lambda_invest. When they show up, they trade immediately at market, and that's it. They're mechanical.

The other kind of player is Trading Firms, or high-frequency traders, you could call them as well. They don't have an intrinsic demand to buy or sell *x*; their goal is simply to buy low and sell high.

Initially, I'm going to assume that there's N trading firms exogenously in the market. Later, I'll endogenize the number of high-frequency trading firms in the model with investment in speed technology.

SLIDE 32: Latency

Initially, I want to assume away all forms of latency. When *y* changes, everybody sees this immediately, so zero time delay, for free. There is no latency in sending messages to the exchange. Again, I'm trying to create a best-case scenario for the continuous market, to try to isolate the core problem.

SLIDES 33-35: "Sniping"

Given the way we've set up the model, and again, there's no asymmetric information, there's no inventory costs, there's no latency, you might conjecture that competition is going to lead to effectively infinite liquidity for investors.

That is, in this model, economic logic would suggest that as an investor, you could buy or sell *x* at price *y*, in effectively unlimited quantity. We've turned off all of the usual sources of costly liquidity provision in financial markets.

But that is not what happens, due to an issue with the market design, and we call this phenomenon "sniping."

Suppose that *y* jumps, say from *y1* to *y2*. Think of this as the moment at which the correlation between *y* and *x* momentarily breaks down, like we saw in the data. Trading firms providing liquidity in the market for *x* send a message to the market, to withdraw their old quotes, and replace them with new quotes. The market's here [using hands to illustrate], there's a jump, I send a message to withdraw my old quotes, replace with new quotes.

<u>However</u>, at the exact same time, *other* trading firms send a message to the market, attempting to trade at the stale quotes before they are cancelled. So I send a message to cancel, but all of you send a message to try to buy from me at my stale price.

Since the market design processes requests to trade <u>serially</u>—so, one at a time in order of receipt—it is possible that one of your messages to snipe my stale quote will get processed before my message to cancel my stale quote

In fact, it's not only possible, but it's <u>probable</u>, because for every one Trading Firm trying to cancel, <u>every other</u> Trading Firm will try to snipe them.

SLIDE 36-41: Sniping (graphics)

So I've been explaining this sniping phenomenon with my hands for several years now, but given that this is a big room I thought we should make some graphics, just to go over the argument, quickly, one more time. So here's the market, there's fundamental value and a bid-and-ask around fundamental value, the fundamental value jumps. If I'm a Trading Firm with quotes in the book, I'm going to send a message to the market to withdraw those old quotes and replace them with new quotes. But at the exact same time, any other trade firm paying attention is going to send a message to the market trying to snipe my stale quotes. So this is the set of messages trying to interact with the stale quote. The likelihood that, as a liquidity provider, my request to cancel reaches the market before one of your requests to snipe is small; it's 1 out of *N*.

SLIDE 42: "Sniping"

So, in a continuous market, symmetric public information creates arbitrage rents. The obvious mechanical arbitrages like we saw in the data are "built in" to the market design.

I want to underscore, again, this just isn't supposed to happen in an efficient market. Symmetric public information is supposed to get into prices for free, and asset prices are supposed to be hard to predict in the short run.

In equilibrium, these arbitrage rents are ultimately paid by investors.

SLIDE 43: Equilibrium Effect on Liquidity

There are some details of the equilibrium analysis that I'm going to skip, but the basic economic ideas are pretty simple.

In equilibrium, trading firms are going to be indifferent between liquidity provision and stale quote sniping.

Liquidity provision earns revenues, as investors show up and pay the bid-ask spread, and has costs, when jumps occur and the liquidity provider gets sniped, in this case, with probability (*N*-1)/*N*.

Sniping has benefits when jumps occur. If you'll notice, in these expressions, the cost to every one liquidity provider, which has the (*N*-1)/*N* at the end, is equal to the benefits to all the *N*-1 other firms trying to snipe him.

Making Trading Firms indifferent between these two activities yields an equation that characterizes the equilibrium cost of liquidity for investors. What it says is that the profits Trading Firms earn from sniping come out of the pockets of investors via the cost of liquidity.

SLIDE 44: Equilibrium, Endogenous Entry

Now, I want to quickly endogenize entry, and we're going to do this in a simple way. We're going to assume that Trading Firms can see innovations in the public signal *y* slowly for free—we'll call that speed delta-slow—or quickly, with speed delta-fast, if they pay a cost, which I'll call c-speed.

Equilibrium now has a very similar structure to above. It is uniquely characterized now by two conditions: the first one is the same condition as before, where there's indifference between liquidity provision and sniping, and the second one's a free entry condition: Trading Firms keep entering until the marginal Trading Firm earns zero profits. This characterizes the amount of entry.

The new equation we get has a nice economic interpretation; it says all of the expenditure by trading firms on speed technology ultimately is borne by investors in the form of more expensive liquidity, an equivalence between the prize in the race, expenditures on speed, and the ultimate cost to investors.

SLIDES 45-46: What's the Market Failure? (25:43)

So as a Chicago guy, one of the questions I ask myself a lot and get asked a lot is, what exactly is the market failure here? Isn't the arms race just healthy competition?

This paper is really a tale of two market failures. The first market failure is sniping—arbitrage rents from symmetric public information, in violation of efficient markets theory. And the second market failure is that these sniping rents then induce an arms race for speed, which mathematically boils down to a prisoners' dilemma.

SLIDE 47: Remark I: Role of HFTs

I want to make two quick remarks about the model.

First, the model does not say that high-frequency trading per se is bad for markets. Rather, what it says is that high-frequency traders endogenously choose do two kinds of activities: they provide liquidity, and they snipe stale quotes. Providing liquidity is useful for investors, and sniping stale quotes is bad for markets.

Don't be misled by the fact that the sniping looks like zero-sum HFT on HFT combat. This misses the economics. The economics is that sniping is like a tax on liquidity provision, and this in turn harms investors.

SLIDE 48: Remark II: Arms Race is a "Constant"

The second remark is that, if you'll notice, the size of the sniping problem in our analysis has nothing to do with the speed technology per se—it didn't matter whether the speed improvements are seconds, or milliseconds, or microseconds, etc.

It also doesn't matter whether the speed technology is expensive or cheap—if it's cheap, there will be a lot of entry; if it's expensive, there will be little entry.

This tells us that the arms race is an <u>equilibrium</u> never-ending feature of the market design.

And, to illustrate, here are some highlights from the high-frequency trading arms race since we first started this research, in fall of 2010.

SLIDES 49-62: HFT Arms Race: Continued

This is the first microwave network connecting Chicago and New York City. It launched in 2011. It wasn't very straight, but since light travels 50% faster through air than glass, it was straight enough to be faster.

[SLIDES 50-52] This first connection actually starts around the corner from this hotel, at the corner of Randolph and Columbus, at the Aon Center. If you look at the top of this building in the center here, you'll see a nice microwave tower.

[SLIDES 53-59] Here then is all the progress since that first connection between Chicago and New York City since 2011, in putting in new microwave links between Chicago and New York. You'll notice that in this diagram are also connections to Washington, D.C. That's because government numbers—<u>which are perhaps the purest form of symmetric public information that there is</u>—are disseminated from D.C. The location of the server farm is literally on K Street.

[SLIDE 60] Here is a microwave path currently in construction—this is data as of December 2016—that appears to be aimed at sending info from Chicago to Seattle by microwave, and then the rest of the way to Asia by underwater cable.

[SLIDES 61-62] And here's a poster I saw at a recent industry conference, for a hardware device. The tagline is, "Nanoseconds matter." A nanosecond is a <u>billionth</u> of a second. Every time you blink your eyes, 400,000,000 nanoseconds have passed.

SLIDE 63: The Case for Frequent Batch Auctions (28:52)

I now want to explain why moving from continuous time to discrete time directly solves the problem.

SLIDE 64: Frequent Batch Auctions; Overview

At a high level, our market design, frequent batch auctions, makes just two changes to the incumbent market design, the continuous limit order book. First, we put time into units. Time is treated as a discrete variable, not continuous.

Second, we process orders in a batch, using an auction rather than processing them one at a time, that is, serially.

SLIDE 65: Frequent Batch Auctions: Definition

Here's a bit more detail on the proposed market design: there's a batch interval which we'll call tau. And to fix ideas for the second part of the talk, let's think 1 millisecond, but it could be longer like 100 milliseconds, and during this batch interval, traders can submit bids and asks, and these are just like standard limit orders. They can be freely modified or canceled at any moment in time. If an order isn't executed in one discrete interval, it carries forward to the next discrete interval until it is either executed or canceled. Again, just like a standard limit order.

At the end of each batch interval, the exchange batches together all outstanding orders—new orders that arrived this interval and outstanding orders from previous intervals—and computes market-level supply and demand curves.

Supply and demand either don't cross or they do; if they don't cross no trade occurs, and all orders just carry forward to the next auction. If they do cross, then the market clears where supply equals demand. It's a uniform-price auction, a design that may be familiar from the US treasury market.

Priority is still price then time, but with time discrete. This means that if my order has been resting in the book for several batch intervals, and yours is new this interval, I have priority over you, but if we both entered in the same interval, we have the same priority.

The information policy is that the same information is disseminated as in the continuous market—trades, outstanding orders, cancels, and so forth—but with the information disseminated in discrete time. This is an important detail to underscore, it's economically important that information is disseminated in discrete time, rather than continuously.

SLIDE 66: Frequent Batch Auctions: 3 Cases (Case 1)

To explain the market design a bit further, I want to go through three cases.

The first case is that in the millisecond, nothing really happens. This is a very common case, because in most stocks and other financial instruments, in most seconds there is no trade, let alone in fractions of seconds. Even the most active symbols in the market have activity in only about 5% of milliseconds, trades in only about 1% of milliseconds. Even for a stock like Google, which is pretty heavily traded, there's activity in less than 0.5% of all milliseconds, and trade in less than 0.1% of milliseconds.

In this case, all outstanding orders simply carry forward to the next time interval, and the state of the book is displayed publicly, flashing on your screen or to your algorithm in discrete time.

This is just like displayed liquidity in a limit order book market. The bottom of the supply curve is the ask; the top of the demand curve is the bid.

SLIDE 67: Frequent Batch Auctions: 3 Cases (Case 2)

A second case is a small amount of trade happens in the interval; an investor shows up and wants to buy a small amount at market.

This case, too, is just like current practice. Investors can "buy at the ask" or "sell at the bid" just like in the limit order book market.

SLIDE 68: Frequent Batch Auctions: 3 Cases (Case 3)

The third case is if there's a burst of activity in the interval, for instance in our model if there's a jump in the public information *y*.

This case is where discrete time and continuous time are importantly different. And there's two reasons.

SLIDES 69-70: Why FBA Solves the Problem

The first, and more obvious, reason is that discrete time reduces the economic relevance of tiny speed advantages.

As long as the batch interval is long relative to the difference between fast and slow traders—and don't think of fast and slow as high frequency traders versus you and me, but more like cutting-edge high-frequency traders versus other sophisticated market participants—then most information arrives at a time during the batch interval when all traders see it equally.

It's only a small sliver of the interval, of proportion delta over tau, where if public information arrives then, the speed advantage is relevant.

The second, more subtle, and I think more important, reason is that the auction itself changes the nature of competition. Instead of competing on speed, trading firms compete on price.

The easiest way to see this is to suppose that public information actually does arrive in this critical window, and there are some slow traders with stale quotes in the book.

In the continuous market, this is going to lead to a race by the fast traders to snipe the stale quotes. Whereas in the discrete auction market, the fast traders compete on price. So if the information is truly public and truly obvious, the auction is going to compete away the arbitrage profits.

SLIDE 71: Takeaways from Equilibrium Analysis

In the interest of time I'm going omit the details of the equilibrium analysis and instead just give you the takeaways.

If we treat the number of high-frequency traders as exogenous—which is probably the right way to think about initial entry of a frequent batch auction exchange, or a modest pilot test—then Frequent Batch Auctions eliminates sniping for *any* batch interval tau. I interpret this "any"—mathematically it's a discontinuity between zero and something positive—but I interpret this "any" as meaning long enough, given computational and communications technology, to enable genuine batch processing if multiple trading firms act on essentially the same signal at essentially the same time. This is probably on the order of a millisecond or less given modern technology.

If we treat the number of high-frequency traders as endogenous—this is the right case for thinking about a more market-wide reform aimed at stopping the speed race—then the equilibrium analysis says that the interval has to be long relative to the speed advantages in play. A rough calibration, given the scale of the modern speed race, points to intervals that are a tenth of a second or potentially significantly less.

SLIDE 72: Computational Benefits of Discrete Time

In addition to stopping sniping, enhancing liquidity, and stopping the arms race, discrete time is also computationally much simpler than continuous time.

The basic conceptual point is that computers and communications technology are not infinitely fast, whereas a continuous-time market implicitly assumes that they are.

This manifests in lots of bizarre ways. To give one example, you have to take relativity into account when trying to parse the market's paper trail or trying to analyze the market as a researcher or as an investor.

Another example is that in continuous markets, there is an intrinsic tradeoff between error-checking and speed. Every line of code that you add to check robustness of a trading decision reduces speed by economically relevant amounts.

SLIDE 73: Alternative Responses to the HFT Arms Race

There have been numerous alternative responses to the high-frequency trading arms race that have been discussed in recent years; we discuss these in some detail in the paper. I'll just make two brief points here. One is that Bans seem to misunderstand cause and effect. The other is that my thoughts on IEX's market design are fairly mixed and nuanced, and if you're interested I'd encourage you to look at the letter I wrote to the SEC on their exchange application, which is on my website.

SLIDE 74: Summary of Budish Cramton and Shim (35:55)

So, to summarize Budish, Cramton and Shim, we look at the high-frequency trading arms race from the perspective of market design. The root problem is not "evil high-frequency traders," it's a market design glitch—continuous-time, serial process trading.

This glitch causes a built-in failure of the efficient markets hypothesis; it's simple to fix by moving to a discrete-time, batch process design. This market design eliminates sniping, it enhances liquidity, and it stops an industrial arms race. It also simplifies the market computationally.

SLIDES 75-76: Response to BCS

The paper was released publicly in working paper form in July 2013, and pretty quickly took on a life of its own, which I guess isn't that surprising in retrospect, but was surprising at the time.

Attention came not just from within academia, but from many different kinds of stakeholders, including exchanges, high-frequency trading firms, the large investment banks and broker-dealers, institutional investors, trade groups, and many different kinds of regulatory bodies around the world.

My approach to the attention that the paper got—and I should say I have no idea if this was the best approach, personally or professionally—was most centrally to invest a lot of time not just giving academic seminars and attending academic conferences, but also in lots of private meetings with stakeholders, lots of industry conferences, lots of panel discussions where sometimes it felt a little bit like I was in the lion's den. I probably visited or at least spoke to half a dozen each of the largest high frequency trading firms, exchanges, broker-dealers, institutional investors, and so forth, and numerous regulatory bodies around the world. Essentially, if a credible stakeholder expressed sincere interest in the work, I made time and often got on a plane. My colleague Austan Goolsbee called this "shoe leather" costs.

I learned an enormous amount from these interactions: first about institutional details; second about how to communicate the work in a language that would translate and that would resonate. I actually think a lot of the improvements to the paper itself, between the first working paper version and the final published version, were shaped by these discussions.

SLIDES 77-80: Reactions

The work of course had a range of reactions. Some of the response was quite positive. The highest-profile public support came from the New York Attorney General, Eric Schneiderman, who among other things noted that as a University of Chicago economist, I am, quote, "not an enemy of free markets," which I of course agree with. [SLIDE 78] Bloomberg and Goldman Sachs provided support. Goldman, full-disclosure, I worked at for a few years before graduate school as an extremely junior investment banker.

[SLIDE 79] Support came from within academia, including, to my delight, from the full spectrum of the University of Chicago lunch table. [SLIDE 80] And my favorite compliment for the work came from the quant hedge fund manager Cliff Asness.

Some of the response of course was also quite caustic, as you can probably infer from Cliff's compliment. I actually got called "communist" a lot. Which I found kind of odd as an economist talking about auctions, so keep that thought in mind.

One prominent exchange executive said at a conference that if US markets adopted discrete time, that he would take continuous markets and all of his technology to North Korea, so that he could, quote, "show them how they could become the center of finance in about three weeks."

SLIDE 81: Most Common Question: Private vs. Regulatory

But I think by far the modal response, and this spans academia, industry, and regulators, was some version of the question that is the title for today's talk. The modal sentiment was something like, "You're probably right … but how do we get from here to there? Is a regulatory mandate required (which is often interpreted as communist)? Or, can market forces alone fix the problem? Will the market fix the market?"

So what I want to do in this part of the talk is to try to answer this question. I'll tell you about some new research, still very much in progress, with Robin Lee and John Shim.

SLIDES 83-87: SEC Chair White, June 2014

To frame the issue, I want to start with a quote from the Chair of the Securities and Exchange Commission, Mary Jo White. This quote is from a speech she gave in June of 2014, in the wake of the release of *Flash Boys*.

She first acknowledges the possibility of an arms race, which is notable in and of itself. She then says she's personally wary of a mandate, but is receptive to exchanges innovating on market design, including frequent batch auctions.

Last, she says that the SEC's job, as the market regulator, is to ensure that it doesn't inadvertently stand in the way of innovation.

SLIDES 88-89: Private vs. Social Returns to Market Design Innovation (40:29)

There's an implicit presumption in Chair White's remarks, which is that market forces correct inefficiency, so long as regulators don't get in the way. This of course is a natural instinct; it's surely the standard case in economics, but as we all know it's not the only case.

The goal of this paper is to build a model of stock exchange competition so that we can understand the private and social returns to innovation, and ultimately try to answer the question "will the market fix the market?"

The paper's got three parts: the first part tries to build a theory model of the status quo, of competition amongst continuous markets. The second part of the paper uses a variety of data—and I'll use the word "validate" the model empirically—it's a very simple and parsimonious model of a very complicated industry, so I want to show you that it's sensible.

Last, we're going to use the model to study the incentives to innovate. Think of this as running a counterfactual in IO. The bulk of the scientific contribution of the paper is in the first two parts, but for the purpose of today's talk, I'm going to try to go through the first two parts quite quickly and get to the third part.

SLIDE 90: Exchange Competition Game

Our model departs from the BCS model. The main departure is that we move from there being a single passive exchange that doesn't really do anything in the BCS model to multiple exchanges who are strategic players. And exchanges, initially, all use the continuous market design, and get to strategically choose two prices. They choose trading fees (as little *f*). Fees are per-share traded, paid by each side of the trade.

And as you'll see, this abstracts from some of the complexity of modern fee schedules, but gets the economics pretty right in the data.

And then second are exchange-specific speed technology fees, which we'll denote big *F*. I didn't mention this in my discussion of the earlier paper, but many of the key technologies in the speed race are sold by exchanges themselves: co-location of servers—the right to put your computer near the exchange computers—the right to get a fast connection between your computer and the exchange's computer, called connectivity; proprietary fast data feeds, that are faster than the regulatorily available data feeds. And to be fastest at any given exchange, trading firms need to buy both general-purpose speed technology as well as this exchange-specific speed technology.

SLIDE 91: Key Institutional Details

There are two key institutional details that shape how we model the status quo. The first is called Unlisted Trading Privileges or UTP, which basically says that stocks are fungible across exchanges. You can buy any particular stock on one exchange, sell it on another exchange, even while it's listed on some third exchange. Where a stock is listed isn't as economically important as many people seem to think.

The second is Regulation National Market System, which is a long and complicated piece of regulation but economically has two aspects that are important for us. One is that it requires that traders, on an order-by-order basis, send their trade to whatever exchange or exchanges have the best price. This obligation can be fulfilled by the trader or by someone acting on behalf of the trader. Second, it requires that exchanges make data about these best prices easily electronically accessible. We're going to model this complicated regulation by assuming that investors and Trading Firms can frictionlessly search all exchanges any time they wish to act—search is zero cost.

I want to note parenthetically that the precise language of Reg NMS didn't anticipate the importance of latency in modern financial markets; that's a story for a different day, for today's talk I want to model Reg NMS as frictionless search.

I'm also going to assume for today's talk that Frequent Batch Auctions is allowed under Reg NMS, but at the moment that's in a regulatory limbo, there's some ambiguity. But a recent ruling by the SEC in June 2016 was a good step.

SLIDE 92: UTP + Reg NMS → Virtual Single Platform → Perfectly Competitive Trading Fees

This combination of Fungibility and Frictionless Search means that, even though stock exchanges look like a platform market that you might think to study with the traditional IO tools of platform competition, in fact we should think about stock exchanges as what we're going to call a "virtual single platform."

Investors and Trading Firms use fungibility and frictionless search to "stitch together" a synthetic single market out of what look like many disparate markets. Once we have single market, we can start to think about the economics in terms of traditional supply and demand—in this case, the supply of liquidity and the demand for liquidity—and exchange trading fees can be thought of as like a tax wedge between supply and demand. Frictionless search then drives this fee towards marginal cost, which in this case is zero.

SLIDE 93: Description of Equilibrium (45:13)

The equilibrium of our model is a bit involved notationally; I am going to omit the math, but the economics are fairly straightforward. There are four key features.

The first is this virtual single platform idea; on this virtual single platform, market shares coordinate behavior and the marginal unit of liquidity is indifferent across all exchanges.

Second, trading fees are perfectly competitive, as I mentioned.

Third, where exchanges do have market power is in the sale of exchange-specific speed technology. If an exchange has 20% share, this means that 20% of the latency arbitrage prize happens on this exchange.

This exchange, uniquely, can sell exchange-specific speed technology to access these sniping opportunities.

And then last is a money pump constraint. An exchange with low market share might be tempted to cut price to even lower than the competitive level to try to gain market share, and hence be able to earn more from the speed game. But this bumps up against a zero lower bound because if an exchange charges a negative fee to trade, this creates a money pump—you trade with yourself and extract money from the exchange.

All four features of the equilibrium are borne out in the data, which I'll go over quite quickly.

SLIDE 94: Empirical Validation: Virtual Single Platform

First, the virtual single platform theory: the volume-depth relationship holds up robustly in the data; at least for the most actively traded stocks, the modal number of exchanges at the best price at any given time is, roughly speaking, "all of them."

SLIDE 95: Empirical Validation: Competitive Trading Fees

Trading fees, as I mentioned, are quite complicated, but when you cut through the complexity and try to compute what is the average price to trade a share of stock during regular trading hours on the main exchanges, the answer is about 0.01 pennies. Now that's not zero, but it's pretty small. If you add 0.01 pennies up across all shares traded in the United States, that's less than the annual revenue of my employer, Chicago Booth.

SLIDES 96-98: Empirical Validation: Profitability of Colo/Data

Third, regarding market power for Exchange-Specific Speed Technology, exchanges in the US do indeed make a large fraction of their revenue from Colocation, Connectivity, and Data. On the left-hand side is a revenue pie chart for the BATS family of exchanges, for which the data is the best, and on the right-hand side is the same revenue pie but for the Chicago Mercantile Exchange, a large futures market. For the CME, Data and Access is around 15% of total revenue, whereas for BATS it's nearly 70%.

[SLIDE 97] In fact, for BATS' US Equities business, trading fees alone are below operating costs, and this is consistent with trading fees not only being competitive, but being like loss leaders. With Colocation revenue, they're quite profitable; without, they lose money. NASDAQ and NYSE look similar, though their data are less good. In total, the three main exchange families make what looks like over $1bn per year from these various kinds of fees.

[SLIDE 98] Here's Nasdaq's current four different tiers of co-location and connectivity, the different prices, the latency differences among which are measured in millionths of seconds.

SLIDE 99: Empirical Validation: Money Pump Constraint

Last, you can see the money-pump constraint bind in the data. If you look all the way to the right, what we call a "Max User"—this is basically a high-volume Trading Firm or Broker—on 8 of 10 exchanges, their fees are actually slightly negative. This is right around where the money pump constraint binds in actuality—it's a bit less than zero, because of per-share taxes that you have to pay to the SEC or FINRA.

SLIDE 100: Incentives for Market Design Innovation

The data suggests that the simple model is pretty sensible, and I now want to turn to the main question at hand, which is the will the market fix the market? What are the incentives for market design innovation?

SLIDE 101: What Happens if FBA Enters (49:02)

So first the good news: Suppose a startup exchange enters and adopts Frequent Batch Auctions—I'll call this Discrete—and charges a fee of zero. All other exchanges use the incumbent Continuous market design, and also still charge a trading fee of zero; they can charge positive prices for colo and data and whatnot.

A reasonable guess—and this was our guess before we started the study—is that there's basically the usual coordination problem of getting a new market off the ground. One equilibrium is nobody pays attention. But in fact, there's a unique equilibrium in our analysis, which is that the Discrete market gets 100% share.

The key reason is the frictionless search. Trading Firms strictly prefer to offer liquidity on the Discrete market than to offer liquidity on Continuous market to avoid the cost of getting sniped. Again, getting sniped is like a tax on liquidity provision. But this only holds if investors notice that they're providing better liquidity on the Discrete market than on the Continuous market.

The key thing is, because search is frictionless, investors notice. In a sense, Reg NMS mandates that they notice.

So, the unique equilibrium is 100% Discrete. Essentially, if you've got two otherwise identical markets operating in parallel, one that's got a tax and the other one doesn't, and there's no frictions, the one without a tax is going to win.

This same argument works if Discrete charges a positive fee, provided that the fee is smaller than the savings from sniping. The economic interpretation is that the market design innovator is getting paid for fixing a problem. And I don't want you to take 100% literally, parenthetically.

SLIDE 102: But…Incumbent Response → Bertrand Trap

That's the very good news. The bad news is that, if an Entrant innovates, Incumbents are going to have incentive to copy.

Suppose an initial market design innovator adopts Discrete and charges a positive fee. Then an incumbent is going to want to switch to Discrete and charge a slightly lower fee and this process continues until fees get competed down towards the perfectly competitive level.

This would be great for the market—you'd have a better market design, at competitive fees—but bad for the entrant, whose profits get competed down to zero.

Conceptually, this is just the classic Bertrand trap. Innovators are unable to capture the social value of their innovations if other firms can easily copy.

SLIDE 103: And...Incumbent Rents from Speed Technology

A second source of tension between private and social value is found by considering the entry problem from the perspective of an Incumbent.

If an Incumbent adopts the Discrete market design, the analysis is similar to before where the market is going to uniquely tip towards the market design that eliminates the sniping tax and the speed race. The sniping tax and Bertrand competition competes fees down to zero.

But there's a key difference, which is that the incumbent's profits used to be positive—they used to have rents, under the old equilibrium, from the exchange-specific speed technology.

Formally, you can model this game among incumbents as a repeated prisoner's dilemma in which the status quo is in equilibrium.

SLIDE 104: Bertrand Trap and Prisoners' Dilemma Among Incumbents

These theoretical ideas—Bertrand Trap, and the idea that incumbents are in a repeated Prisoners' Dilemma—might sound a bit fanciful, but here is a quote from a senior executive at Nasdaq, at an academic event a few years ago. He said:

"Technologically, we could adopt Frequent Batch Auctions. But it would cost time and it would cost effort to get the SEC to approve it, and if it got approved, it would be immediately copied, so there'd be no first-mover advantage. And hence, no incentive to innovate. "

SLIDE 105: Summary of Analysis

So, to summarize:

<u>If</u> a Frequent Batch Auctions exchange enters, the new market design wins significant share, in our stylized model, 100% share, though of course that's an abstraction. However, the fierce competition on fees that ensues drives the entrant's profits down towards zero. And an incumbent's incentive to adopt is even worse, because it foregoes the rents from speed technology. So, things don't look great for the market fixing the market!

An optimistic spin on this analysis is to think mathematically about zero. The great thing about zero is it's pretty close to being positive. So, it's hard to squint at the model and see Frequent Batch Auctions as being an entrepreneurial home run, but you can certainly squint and make the case that there's a positive incentive. For example: if the Bertrand trap doesn't bind literally, or immediately, or adoption is pretty cheap.

And this tells you where to look for a private solution; it could come from a new entrant, like an IEX, or from an incumbent with low market share and hence little to lose from losing the speed race, like the Chicago Stock Exchange, or, as of this fall actually, IEX, which entered with a continuous market design, and currently has share on the order of about 1.5-2%.

A private sector solution could also come from outside of US equities markets—for example, in futures markets, where the Bertrand trap is not an issue because futures contracts aren't fungible across exchanges in the same way that stocks are, so they may be better able to capture the value created by eliminating sniping, either through fees or through market share.

Least likely would be a large incumbent stock exchange with large revenues from the speed race.

So, I don't want to rule out all hope for a private sector solution; rather, what I conclude robustly is that the private incentives to innovate are dramatically lower than the social incentives to innovate, and, for the largest incumbents, they may be negative.

SLIDE 106: Regulatory Response

The analysis suggests a potential role for regulatory intervention. There's a well-documented market failure, and the status quo of the market, and the model suggests that private market incentives alone might not fix the problem.

I first want to describe what I think of as a quite modest regulatory response that the model suggests would shape the nature of private competition to facilitate the market fixing the market.

First, reduce adoption costs for the first entrant. Proactively clarifying that innovation is allowed would be a step in this direction. Second, reduce tick-size constraints. I didn't emphasize tick sizes in today's presentation of the theory, but a fat tick size can be a constraint against the market tipping towards a more efficient design because sniping is often less than a penny per share traded. Third, such a regulatory response might have to find some way to insulate an initial entrant from the Bertrand trap—for instance, a modest exclusivity period to provide some modest first-mover advantage.

That said, the most direct regulatory response to the research I've presented today would be to simply put time into discrete units. Such a policy, properly designed, would fix latency arbitrage, stop the speed race, and would also dramatically simplify the market. While it's hard with the data that's available to academics to express 100% confidence, I personally would support such a policy.

SLIDE 107: Political Economy of Regulation

As we often do in economics talks, I spent a long time trying to convince you that there is some kind of market failure, and then somewhat quickly described a potential regulatory response. But if there's anything that I've learned at the Chicago lunch table, it's to also think about the incentives of regulators. And, in this instance, the political economy of solving the problem is tricky.

Most centrally, addressing sniping and the speed race is a classic example of dispersed benefits and concentrated harm: the harm to High-Frequency Traders with a comparative advantage at speed, to exchanges that sell speed, to speed technology providers, and so forth.

Compounding the concentrated/dispersed problem is that the subject matter is technical and it's nuanced, both of which make it more difficult to organize dispersed interests.

The regulatory risk-reward of fixing the problem isn't great. By this I mean that the benefits of fixing the market are subtle—improvements in liquidity and computational simplicity—but the potential cost is that if you're the regulator who changes market rules and something bad happens, you're going to get blamed, even if one had nothing to do with the other.

Last, but certainly not least, there's a bit of a chicken-and-egg problem, where each country's regulators would kind of like some *other* country to try it first. Even a pilot test would be a major undertaking. My one bit of progress to date on this is with the Financial Conduct Authority in the United Kingdom, which is gathering new kinds of data from exchanges to enable more direct measurement of latency arbitrage.

Bottom line, the political economy for regulators to address the problem isn't great, and, to be honest, I'm a little bit sympathetic.

SLIDES 108-109: Conclusion: Role of Academics: Theory → Practice (57:24)

So where are we? We've got a well-identified market failure that strikes at the core of what we mean by an efficient market, at the core of how efficient markets are supposed to work. It's causing an industrial arms race, it harms investors, it makes markets unnecessarily complicated. It's probably greater than $100 billion in Net Present Value, and there's a reasonably simple solution.

The case for market forces fixing the market failure isn't great. I don't want to say it will never happen, but the economics aren't great. It's not like our paper came out and the large incumbent exchanges CEOs started calling and saying, oh, this is great idea, let's try this. And this kind of makes sense; if you take our model at face value, the private incentives for an Entrant to fix the problem are zero, and the private incentives for an Incumbent to fix the problem are negative.

The political economy for the regulator to fix the market also isn't great. Again, I don't want to say it will never happen, but the political economy is tricky, because of concentrated interests, the technical subject matter, and so forth.

So, something I've been thinking a lot about lately is what to do about this. And I want to emphasize, this particular problem, I think it's important, and I'm of course quite proud of the work we've done, but there are much bigger ideas from economics with the same basic structure, where you have a well-defined market failure, a nice solution from economics, and the political economy for fixing the problem is lousy. Carbon taxes is probably by far the leading example of this.

I have two thoughts I'd like to end with.

SLIDE 110: Friedman on Theory → Practice

My first thought is just be patient and keep doing the work. There's plenty of work left to do on the theory, getting more data, continuing to talk with stakeholders. This is the sentiment of a famous quote of Milton Friedman; he talks about the "inertia of the status quo" and how it "takes a crisis" to get change, but my favorite part of the quote is how he describes our job as economists, as keeping good ideas alive, keep pushing, until an opportunity for constructive change comes along.

SLIDE 111-115: Roth and Zingales on Theory → Practice

My second thought's more of a question, and it's whether there is more we can do as a profession—whether there are institutions that we can build, or professional norms and incentives that we can shape—to bring our good ideas from research into the world. Al Roth, in "The Economist as Engineer"—and this is without a doubt the single paper that's had the biggest influence on my professional life—wrote that we need to foster a new kind of literature in economics, akin to the relationship between engineering and physics, that's considerably more applied, considerably more tailored to institutional details than we're used to and that some of us may be comfortable with.

My colleague Luigi Zingales argued in his AFA presidential address that we should be more accepting of and encouraging of engagement in policy debates, as long as the engagement is disciplined by theory and data. Both Roth and Zingales are thinking about a gap between what academics naturally do in their research, and their published research especially, and what academics might be able to do, and need to do, to get good research ideas implemented in the world.

Reflecting on the unusual experience I've had over the past few years as a relatively young researcher, being thrust into a high-stakes, fast-moving, highly charged debate, and at times, frankly, feeling outgunned and outmanned, no matter how clever my talking points were, by the interests on the other side—the Budish PR machine is no match for the HFT PR machine—I've come to the I guess obvious conclusion that the changes Roth and Zingales are encouraging are especially important for ideas that are high social value, but where concentrated interests are opposed. That is, when our ideas benefit the dispersed.

When social value and private value are aligned, then natural economic forces build the bridges from research to practice. I'm not saying this happens automatically or without serious effort, but at least the winds of change are with us. I especially have in mind some famous examples in finance, like index funds or derivatives or sophisticated portfolio management. These are all ideas that are extremely socially valuable, and there were private incentives to bring the idea out into the world, and the process kind of worked.

But the other case, where social value is positive but private interests are opposed, then not only is there not a natural magnet pulling the good idea from research to practice, but there's active opposition.

In the end I'm kind of an optimist, and believe that, eventually, good ideas win; I'll wager with anyone in this room—please come see me afterwards—that we'll see discrete-time trading eventually, though we might have to measure time in decades.

But I'm left wondering if there's anything we can do as a profession to speed up. If high-frequency traders are investing billions of dollars on improving the speed of information transmission between markets by literally billionths of seconds, is there more we can do as a profession to speed up the transmission of our ideas, from decades to something that's a bit more high-frequency.