The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXIX, NO. 2 • APRIL 2024

Informed Trading Intensity

VINCENT BOGOUSSLAVSKY, VYACHESLAV FOS, and DMITRIY MURAVYEV*

ABSTRACT

We train a machine learning method on a class of informed trades to develop a new measure of informed trading, informed trading intensity (ITI). ITI increases before earnings, mergers and acquisitions, and news announcements, and has implications for return reversal and asset pricing. ITI is effective because it captures nonlinearities and interactions between informed trading, volume, and volatility. This data-driven approach can shed light on the economics of informed trading, including impatient informed trading, commonality in informed trading, and models of informed trading. Overall, learning from informed trading data can generate an effective informed trading measure.

INFORMED TRADING IS AN IMPORTANT feature of financial markets that various academic literatures attempt to account for. For example, informed investors help keep security prices close to fundamental values and thus are crucial for theories of market efficiency (e.g., Fama (1970); Grossman and Stiglitz (1980)). Models of asset prices also incorporate the effect of informed trading on a firm's information structure or liquidity risk (e.g., O'Hara (2003); Kelly and Ljungqvist (2012)). Informed trading is difficult to identify empirically, however, because investors' information sets are not directly observable and because informed investors usually hide behind uninformed order flow. To overcome these challenges, the literature has developed several theory-based measures of informed trading and/or adverse selection. Among the most

*Bogousslavsky and Fos are with Boston College, Carroll School of Management, and Muravyev is with Michigan State University, Eli Broad College of Business. We thank Shuaiyu Chen (discussant), Tarun Chordia (discussant), Kevin Crotty, Jefferson Duarte (discussant), Jiacui Li (discussant), George Malikov (discussant), Stefan Nagel (the Editor), Andriy Shkilko, Elvira Sojli (discussant), the Associate Editor, and two anonymous referees for many helpful comments. We also thank seminar participants at NBER Big Data and High-Performance Computing for Financial Economics Conference, CFEA, 5th SAFE Market Microstructure Conference, Microstructure Asia Pacific Online Seminar, SFS Cavalcade, 5th FFIC, Arizona State University, Boston College, Florida State University, Michigan State University, Morgan Stanley Quantitative Research Colloquium, and Rice University for helpful comments and suggestions. We thank Daniel Goodman for excellent research assistance. We also thank Kevin Crotty for sharing his informed trading measure and Patrick Augustin for sharing data on illegal insider trading. The latest version of the ITI measure can be downloaded from the authors' website. We have read *The Journal of Finance* disclosure policy and have no conflicts of interest to disclose.

Correspondence: Vincent Bogousslavsky, Finance Department, Carroll School of Management, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA; e-mail: bogoussl@bc.edu.

DOI: 10.1111/jofi.13320

well-known measures, which have been extensively used in empirical finance, economics, and accounting, are the price impact of Kyle (1985), the bid-ask spread of Glosten and Milgrom (1985), and the probability of informed trading of Easley and O'Hara (1987). But recent studies show that these measures perform poorly in capturing realized informed trading.[1]

In this paper, we introduce a novel data-driven approach to construct a measure of realized informed trading. Specifically, we train a Gradient Boosted Trees (GBT) algorithm on informed trading data. The algorithm is trained to identify days with informed trading. Under this standard classification problem, a daily indicator for informed trading is predicted by a set of same-day variables related to liquidity, return, volatility, and volume. After the model is estimated on the training data of observed informed trades, we extrapolate it to the entire stock-day universe, where informed trading is not directly observed. This procedure produces a new measure of informed trading, which we refer to as informed trading intensity (ITI). ITI combines two key ingredients: informed trading data and a machine learning (ML) method.

With respect to the first, we initially train ITI on Schedule 13D trades to study the basic properties of our methodology, as we expect the Schedule 13D sample to have a higher "signal-to-noise ratio" than other informed trading samples (i.e., opportunistic insiders and short sellers) that we use later.[2] As part of Schedule 13D filing, an investor must disclose all trades made in the 60 calendar days before the filing date. Collin-Dufresne and Fos (2015) are the first to collect these trades and show that they are informed. Our methodology does not require all trades to be informed, but we refer to these trades as informed for simplicity.[3] Informed trading may also occur on days without Schedule 13D trading. Our methodology assumes that more informed trading occurs on days with Schedule 13D trading than on days near those with Schedule 13D trading, on average.

The second key ingredient in the construction of our ITI measure is the GBT algorithm. We train the model with a set of 41 concurrent daily variables motivated by microstructure theory that capture liquidity, return, volatility, and

[1] See Collin-Dufresne and Fos (2015), Kacperczyk and Pagnotta (2019), Augustin, Brenner, and Subrahmanyam (2019), Duarte, Hu, and Young (2020), and Ahern (2020).

[2] Cohen, Malloy, and Pomorski (2012) show that nonroutine, or opportunistic, insider trades are informed in the sense that they predict future returns. An extensive literature shows that short sellers are informed and that their trading predicts future stock returns (see, e.g., Senchack and Starks, 1993; Boehmer, Jones, and Zhang, 2008).

[3] An investor is required to file a Schedule 13D if she becomes a beneficiary owner of at least 5% of any class of equity securities in a publicly traded company and intends to influence management (i.e., engage in activism). Several studies document large average positive abnormal returns around Schedule 13D filings (e.g., Holderness and Sheehan (1985), Brav et al. (2008), and Klein and Zur (2009)). Consistent with these results, Section I shows an cumulative abnormal return of about 3% in the 60 calendar days before the filing date, followed by a two-day jump in excess returns of about 2% around the filing date. Importantly, Collin-Dufresne and Fos (2015) show that the positive prefiling abnormal returns occur primarily on days when Schedule 13D filers trade, suggesting that trades by Schedule 13D filers transmit information into prices.

volume.[4] The results are robust to using a Random Forest algorithm instead of GBT. In contrast, linear regression and the Lasso algorithm are subsumed by the Random Forest and GBT algorithms. Thus, nonlinearities and interactions are important to detect days with informed trading. We therefore use the GBT algorithm.

We establish several main results. We first ask whether ITI detects days with Schedule 13D trading. We show that ITI is a much stronger out-of-sample detector of days with Schedule 13D trading than existing measures of liquidity and informed trading. As in most of our tests, we control for stock turnover, returns, realized volatility, order imbalance, and absolute order imbalance, as well as standard measures of liquidity such as effective spread, price impact, depth, and Kyle's lambda. We further find that while markets have changed over time, the explanatory power of ITI for Schedule 13D trading is stable across the sample period, whereas the explanatory power of standard liquidity variables declines. In addition, while the algorithm is not trained on Schedule 13D trading volume, ITI increases with the volume traded by Schedule 13D filers, which suggests that ITI does indeed capture the intensity of informed trading.

What contributes to ITI's effectiveness? While most input variables contribute significantly to ITI, volume-related variables are the most important. However, even when we match days with informed trading to days without informed trading on volume, ITI continues to strongly detect days with informed trading, which indicates that the measure's effectiveness cannot be attributed solely to trading volume. We examine nonlinearities with partial dependence plots and find that ITI is increasing and concave in volume and decreasing and convex in volatility. Surrogate trees, a popular method to interpret ML models, indicate that variable interactions are also important for ITI. Specifically, ITI is particularly high if volume is high but volatility is low. Similarly, ITI is particularly low if volume is low and illiquidity is high.

Having presented ITI's key properties, we extrapolate the measure to the full sample of U.S. common stocks from 1993 to 2019 (about 17 million stock-days). The model parameters are estimated on about 60,000 stock-days with data on Schedule 13D trades. The estimated model then computes ITI for each stock-day with same-day input variables. We exclude from the full sample the training sample of Schedule 13D trades (0.35% of the sample). The extrapolation assumes that the relations between intraday variables and realized informed trading learned by ITI largely hold in the full sample. We acknowledge that the signal-to-noise ratio could be lower outside the restricted sample of Schedule 13D trades. Thus, in addition to capturing informed trading, ITI could reflect uninformed trading when applied to larger unrestricted samples. While the issue of false positives applies to all ML problems, it is more challenging to evaluate here since informed trading is unobservable, even ex post. We therefore conduct a variety of validation tests and show that ITI consistently outperforms existing measures.

[4] These variables are listed in the Internet Appendix, which may be found in the online version of this article.

We next run several additional analyses. First, using the full sample of U.S. common stocks, we show that ITI predicts the strength of price reversal. A fundamental difference between price changes due to realized informed and uninformed trading is that price changes due to uninformed trading are expected to be transient (Hasbrouck (1988, 1991)). We therefore expect returns on days with high realized informed trading to exhibit less reversal than returns on other days. We find support for this prediction, in line with ITI capturing informed trading. This result is robust to controlling for the interactions between returns and turnover (Campbell, Grossman, and Wang (1993)), returns and volatility, and returns and the effective spread. Hence, ITI is not simply a proxy for turnover, volatility, or liquidity.

Second, we show that ITI increases around several types of information events, even when controlling for standard liquidity, volume, and volatility measures. In particular, ITI increases before earnings announcements. ITI further predicts large abnormal returns on earnings announcement dates, suggesting that ITI increases when informed trading is more likely prior to earnings surprises. ITI also increases ahead of unscheduled informational events, such as unscheduled news releases and merger and acquisition (M&A) announcements, and in the days following announcements. The disclosure of news could increase information asymmetry due to heterogeneity in information-processing capacity (Kim and Verrecchia (1994)). Informed investors could also take advantage of increased volume to camouflage their trades.[5]

One may be tempted to conclude that ITI is an aggregate liquidity measure. The increase in ITI ahead of earnings announcement disproves this conclusion, as liquidity tends to worsen before earnings announcements (e.g., Lee, Mucklow, and Ready (1993)). In addition, recall that ITI increases with the size of activist trade, controlling for liquidity variables. Thus, while we expect ITI to correlate with liquidity variables because informed investors time liquidity (Collin-Dufresne and Fos (2016)), ITI is not equivalent to a liquidity measure.

Having established the effectiveness of ITI in detecting informed trading, we explain how our data-driven approach can shed light on the economics of informed trading. We first take advantage of a useful feature of Schedule 13D trading data to show that there is an important distinction between patient and impatient trading. Specifically, after crossing the 5% ownership threshold, a Schedule 13D filer must file with the Securities and Exchange Commission (SEC) within 10 days. Schedule 13D filers therefore trade more aggressively closer to filing (Collin-Dufresne and Fos (2015)). We use this feature of the data to decompose ITI into a "patient" ITI and an "impatient" ITI, where ITI(patient) and ITI(impatient) are trained on the first 40 days and the last 20 days of the 60-day filing window, respectively. We find that impatient informed trading is easier to detect. This result supports theories

[5] A postannouncement increase in informed trading is consistent with the results of Lee, Mucklow, and Ready (1993) Back, Crotty, and Li (2018), and Brennan, Huh, and Subrahmanyam (2018) among others.

such as Kaniel and Liu (2006) where informed traders tend to use more aggressive orders as their information horizon shortens. In particular, while ITI(impatient) and ITI(patient) are both positively correlated with turnover, the relation for ITI(impatient) is much stronger. Moreover, ITI(impatient) is positively associated with realized volatility, whereas ITI(patient) displays the opposite pattern, and ITI(impatient) increases strongly in the two days before an earnings announcement, whereas ITI(patient) does not. These findings are consistent with ITI(impatient) detecting days with aggressive informed trading. ITI(impatient) is also able to detect illegal insider trades (Ahern (2020)).

Our methodology also sheds light on commonality in informed trading. Specifically, we show that ITI is effective in detecting other classes of informed trading as proxied by opportunistic insider trades and spikes in short selling. Fluctuations in the trading environment are likely to generate commonality in informed trading, but controlling for standard liquidity measures explains only part of the commonality. Moreover, ITI continues to detect opportunistic insider trades and spikes in short selling even when we control for ITI measures trained on these data sets. Hence, incremental information can be gained about one type of informed trading from studying other types of informed trading. The strength of commonalities in informed trading, however, is not clear ex ante. Our methodology provides a first pass at this question by comparing the ITI measure trained on Schedule 13D data to ITI measures trained on other data sets.

Finally, our data-driven measure indicates that a fruitful avenue to shed light on the economics of informed trade is to develop models in which volume and absolute order imbalance are both positively related to realized informed trading, even when controlling for one another, but volatility is negatively related to realized informed trading. Based on simulated data from a range of informed trading models, we show that this dimension of informed trading is challenging to capture. Across all models that we consider, none is able to jointly match these three relations. Notably, our ITI measures remain significant detectors of days with informed trading when we control for five well-known measures of informed trading.

We provide a simple application of ITI by examining whether informed trading predicts future stock returns. Increased realized informed trading, as measured by higher ITI, is associated with higher future monthly returns in panel regressions. In portfolio sorts, the Fama-French four-factor (FF4) alpha for the difference between the top and bottom decile portfolios based on ITI is 52 basis points (bps) per month, or 6.4% annualized, with a t-statistic of 6.2. Other ITI measures (except for ITI trained on short-selling data) also positively predict returns. Our results imply a positive relation between realized informed trading and future returns, which supports theories in which stock purchases are more informed than stock sales. In contrast, the results lend only limited support to theories in which information risk is priced.

This paper contributes to four strands of literature. First, an extensive microstructure literature develops measures of stock liquidity and adverse selection (e.g., Glosten and Milgrom (1985); Easley and O'Hara (1987); Hasbrouck

(1988, 1991); Amihud (2002)). Our key contribution to this literature is to show that a data-driven approach performs remarkably well in detecting various types of informed trading. We further show that information learned from one type of informed trading helps to detect other types of informed trading. Finally, we show that our approach generates new insights about informed trading that can guide theory.

Second, our paper contributes to recent studies that use informed trading data to evaluate the performance of stock liquidity and adverse selection measures as well as the trading choices of informed investors (e.g., Collin-Dufresne and Fos (2015); Gantchev and Jotikasthira (2018); Augustin and Subrahmanyam (2020); Cookson, Fos, and Niessner (2021); Akey, Gregoire, and Martineau (2022)). Back, Crotty, and Li (2018) combine return and order flow information to detect informed trading and show that their measure increases when Schedule 13D filers trade. In this paper, we use three classes of informed trades not as a laboratory to evaluate the performance of a particular measure, but rather as a training set for an ML method. Our contribution is to show that the obtained measures perform remarkably well in detecting informed trading in various settings. Moreover, we show that volume is the most important factor for all of the ITI measures, implying that the informed trading literature has focused perhaps too much on order imbalance relative to trading volume.

Third, we contribute to the active debate in the asset pricing literature on whether informed trading affects asset prices (e.g., Easley, Hvidkjaer, and O'Hara (2002); Duarte and Young (2009); Yang, Zhang, and Zhang (2020)). For example, Easley and O'Hara (2004) find a positive relation between the probability of informed trading (PIN) and future stock returns and justify it theoretically. Hughes, Liu, and Liu (2007) explore whether information risk is priced. Duarte and Young (2009) distinguish between the liquidity and information asymmetry components of PIN. We contribute to this debate by showing that ITI measures of realized informed trading are positively priced in the cross section, and we propose asymmetry in the information content between stock purchases and sales as a likely explanation.

Finally, only a handful of studies apply ML techniques to market microstructure. Easley et al. (2021) show that microstructure-based measures predict changes in liquidity and volatility. Kwan, Philip, and Shkilko (2021) use reinforcement learning to study price discovery. In contrast, we use ML techniques to identify informed trading. We borrow some ideas from a more developed literature on ML in asset pricing (see Gu, Kelly, and Xiu (2020), Karolyi and Van Nieuwerburgh (2020), Goldstein, Spatt, and Ye (2021), and Nagel (2021) for reviews).

The paper is organized as follows. Section I describes our data. Section II introduces the ITI approach to detecting informed trading. Section III explains how our data-driven approach can shed light on the economics of informed trading. Section IV provides an asset pricing application by examining the relation between ITI and future stock returns. Finally, Section V concludes.

I. Data

We use three sets of data on informed trading: trades by Schedule 13D filers, opportunistic insider trades, and short sales. Table IA.II in the Internet Appendix provides an overview of the data sets that we use and how we construct them.

Data on trades by Schedule 13D filers come from Schedule 13D filings available on SEC EDGAR. Our sample construction procedure closely follows Collin-Dufresne and Fos (2015).[6] The 13D sample includes the 60-day disclosure period up to the filing date of 1,593 Schedule 13D filings between 1994 and 2018. We extract the following information from each Schedule 13D filing: Committee on Uniform Security Identification Procedures (CUSIP) code of the underlying security, date of every trade, trade type (purchase or sell), size, and price. In this sample, Schedule 13D filers trade on about 36% of days.

Schedule 13D filers know they can increase the value of the firms they hold through their own effort. Their effort level is, of course, conditional on achieving a large stake in a given firm (Back et al. (2018)). In our setting, their actions and shareholdings constitute private information. Up until their holdings reach the 5% threshold, when this information becomes public due to the disclosure requirement. Announcement returns allow us to measure the extent to which the market believes their future actions have value over and above what is already reflected in prices. Figure IA.1 in the Internet Appendix plots the average buy-and-hold return in excess of the buy-and-hold return on the CRSP value-weighed index, from 40 days prior to the filing date to 10 days after. Like Collin-Dufresne and Fos (2015), we find a run-up of about 3% from 40 days to one day prior to the filing date. The two-day jump in excess return at the filing date is around 2%.

Cohen, Malloy, and Pomorski (2012) show that opportunistic insider trades are informed.[7] We follow their methodology to construct a sample of opportunistic insider trades (both buys and sells). The data come from table 1 of the Thomson Reuters Insider Database. We drop from the sample transactions associated with derivative securities and observations that contain cleanse indicators "S" and "A." Classification of routine versus nonroutine trades follows Cohen, Malloy, and Pomorski (2012), who define routine traders as insiders who placed a trade in the same calendar month for at least a certain

[6] First, using an automatic search script, we identify all Schedule 13D filings from 1994 to 2018. Next, we manually check the sample to identify events with information on trades. Since the trading characteristics of ordinary equities might differ from those of other assets, we retain only U.S. common stocks (Center for Research in Security Prices [CRSP] share code 10 or 11). We exclude stocks with price below $5 or market capitalization below $100 million at the beginning of the filing window. Moreover, we exclude events that involve derivatives, such as options, warrants, and swaps (see Collin-Dufresne, Fos, and Muravyev (2021)). Finally, we exclude Schedule 13D/A filings (i.e., amendments to previously submitted filings) that are mistakenly classified as original Schedule 13D filings.

[7] Other recent studies that identify informed insider trading include Akbas, Jiang, and Koch (2020), Alldredge and Cicero (2015), Biggerstaff, Cicero, and Wintoki (2020), and Cziraki and Gider (2021).

number of years in the past, and opportunistic traders as "everyone else," that is, insiders who have traded in the same years as the routine insiders, but for whom we cannot detect an obvious discernible pattern in the past timing of their trades." Our sample of insider trades starts in January 1993 and ends in December 2012.

We obtain short interest data which include daily data on securities borrowing and lending activity from Markit. Markit obtains information from more than 100 equity loan market participants, who together account for approximately 85% of U.S. securities loans. The data start in July 2006 (when the daily updates begin) and end in December 2019. Short interest is defined as the quantity on loan from Markit divided by shares outstanding from CRSP. We identify large spikes in short interest by comparing daily changes in these variables to their 90th percentiles over the entire sample. Markit reports the date when short sales are settled, which is three days (two days starting September 4, 2017) after trades take place (Richardson, Saffi, and Sigurdsson (2017)). We adjust for this date shift before merging the short-sale data with CRSP and Trade and Quote (TAQ). The Markit data are widely used in academic research on short selling. Furthermore, a large literature documents that short sellers are informed and that their trading predicts future stock returns (e.g., Senchack and Starks (1993); Boehmer, Jones, and Zhang (2008); see Reed (2013) for review of this literature).

We also employ several additional data sources. Earnings announcement dates are obtained from Compustat. We adjust for after-hours announcements by comparing trading volume on the reported announcement day to volume on the following business day. The announcement date is set to the day with the highest volume. Data on M&A announcements are obtained from the Thomson Reuters Securities Data Company (SDC) database, and we require that both target and acquirer are public U.S. companies. Stock returns, volume, and prices come CRSP. Intraday transactions data (trades and quotes) come from the TAQ database. Table IA.III lists data sources for the variables used to train the ML model.

Control variables are defined in Table IA.IV and their descriptive statistics are reported in Table I. We report descriptive statistics for the full sample, which includes common stocks from January 1993 to July 2019. We focus on common stocks and exclude stocks with price below $5 or market capitalization less than $100 million at the beginning of each month to limit the influence of microcaps and penny stocks on the results. Table I also reports descriptive statistics for trading indicators in the 13D, insider, and short samples, as well as for the informed trading measures constructed based on these samples. We discuss these numbers after we introduce our methodology.

II. Informed Trading Intensity

In this section, we introduce the ITI approach to detecting informed trading. Empirical measures of informed trading are widely used in many contexts. However, unlike liquidity measures such as the bid-ask spread, informed

Table I

Descriptive Statistics

The table reports the mean, standard deviation (*SD*), within-stock *SD* (SD_w), and 1st, 5th, 25th, 50th, 75th, 95th, and 99th percentiles for the main set of variables. Control variables are described in Table IA.IV in the Internet Appendix and include effective spread (ES), lambda, depth, realized volatility (rvol), turnover (turn), order imbalance (OI), absolute order imbalance (|OI|), and return (ret). These variables are winsorized at 0.05% and 99.95%. 13D trade is an indicator variable that takes the value of one on days with Schedule 13D trades. Insider trade is an indicator variable that takes the value of one on days with opportunistic insider trades. Δ short is an indicator variable that takes the value of one on days when the daily change in short interest exceeds the 90th percentile. Short interest is defined as total short-sale demand divided by the number of shares outstanding. The full sample consists of common stocks from January 1993 to July 2019, excluding any stock-day with a missing value for one of the control variables and excluding any stock-day in the 13D sample. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. The 13D sample consists of the filing period for 1,593 13D filings between 1994 and 2018 (58,197 observations). For this sample, SD_w denotes the within-filing *SD*. The insider sample consists of two days surrounding opportunistic insider trades between 1993 and 2012, which includes 95,464 days with at least one opportunistic buy trade and 260,366 days with at least one opportunistic sell trade (779,007 observations). The short sample consists of two days surrounding 100,000 randomly selected spikes in short interest from June 2006 to December 2010 (216,979 observations).

Variable Name	Mean	*SD*	SD_w	1%	5%	25%	50%	75%	95%	99%	*N*
ES	0.0042	0.0054	0.0046	0.0002	0.0004	0.0009	0.0022	0.0052	0.0144	0.0256	16,823,151
lambda	0.0036	0.0101	0.0097	−0.0159	−0.0036	0.0001	0.0014	0.0046	0.0184	0.0421	16,823,151
depth ($*10^2$)	0.0082	0.0187	0.0172	0.0002	0.0004	0.0016	0.0035	0.0079	0.0286	0.0743	16,823,151
rvol	0.0227	0.0198	0.0180	0.0026	0.0059	0.0113	0.0175	0.0275	0.0563	0.0997	16,823,151
turn	0.0088	0.0134	0.0132	0.0004	0.0008	0.0025	0.0052	0.0102	0.0273	0.0601	16,823,151
OI	0.0001	0.0026	0.0026	−0.0069	−0.0025	−0.0005	0.0000	0.0006	0.0029	0.0073	16,823,151
|OI|	0.0012	0.0024	0.0024	0.0000	0.0000	0.0002	0.0006	0.0013	0.0042	0.0103	16,823,151
ret	0.0005	0.0307	0.0279	−0.0849	−0.0432	−0.0123	0.0000	0.0126	0.0456	0.0942	16,823,151
ITI(13D)	0.2745	0.1595	0.1558	0.0313	0.0660	0.1539	0.2472	0.3696	0.5883	0.7540	16,823,151
ITI(patient)	0.2067	0.1452	0.1427	0.0156	0.0360	0.0967	0.1712	0.2802	0.5014	0.6845	16,823,151
ITI(impatient)	0.4067	0.1375	0.1322	0.1375	0.1998	0.3033	0.3953	0.4999	0.6492	0.7653	16,823,151
ITI(insider)	0.4761	0.1578	0.1548	0.1271	0.2110	0.3647	0.4732	0.5868	0.7467	0.8356	16,103,808
ITI(short)	0.4074	0.0787	0.0663	0.2433	0.3082	0.3648	0.4003	0.4447	0.5284	0.6126	16,664,285
13D trade (13D sample)	0.3603	0.4801	0.4191	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	58,197
Insider trade (insider sample)	0.4390	0.4963	0.4901	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	779,007
Δ short (short sample)	0.4323	0.4954	0.4908	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	216,979

trading is directly observed only in limited cases. Informed investors also usually hide behind uninformed order flow to avoid being detected. To overcome these limitations, empirical measures are typically motivated by theories of informed trading. For example, the PIN measure of Easley et al. (1996) is a nonlinear function of order imbalance and a sufficient statistic for identifying informed trading in their sequential trading model.

We propose a data-driven approach to detect realized informed trading. Using the data, this approach learns how days on which informed investors trade differ from days on which they do not trade. We use ML techniques to account for nonlinearities and interactions between variables, while cross-validation and regularization, which are standard in ML, prevent overfitting the data. After the model is estimated on a training sample of informed trades, we extrapolate it to the stock universe and compute ITI by applying the model parameters to the set of input variables for a given stock-day. ITI is bounded between zero and one and is higher when informed trading is more likely.

A. Estimating ITI

Economists rely primarily on linear regression, a simple model that identifies individual effects and their significance. We also rely on classic regression models after computing the ITI, but we show that more flexible methods are required to detect informed trading because the underlying relations are likely complex, including nonlinearities and variable interactions. For example, informed trading could be more likely if trading volume is above, say, the 90^{th} percentile and stock volatility is below the median. Of course, nonlinearities and interactions can be hard-wired into a linear regression, but ML methods discover them naturally.

We have to make several design choices. First, we pick a data set in which informed trading is directly observed (a training sample in ML terminology) so that an ML method can learn from contrasting model predictions with observed outcomes (supervised learning). We begin with trades disclosed in Schedule 13D filings, which report activist trading dates and trade sizes in a 60-day window prior to filing. After developing the ITI measure based on Schedule 13D trades, we introduce ITI measures based on opportunistic insider trades and spikes in short selling. This variety of informed trading data allows us to study commonalities across various types of informed trading in Section III.B. We focus on ITI trained on Schedule 13D trades because this data set is likely to have the highest signal-to-noise ratio. In particular, when Schedule 13D filers trade in our sample, they trade an average of 28% of daily volume.

Second, we pick an ML method. We outline the main idea here and refer the reader to the classic textbook discussion of Hastie, Tibshirani, and Friedman (2017) for more details. Lasso is popular in economics because it resembles a linear regression except that it shrinks coefficients, making many of them exactly zero (Tibshirani (1996), and Chapter 3 in Hastie, Tibshirani, and Friedman (2017)). Lasso is easy to interpret but allows only for prespecified

nonlinearities and interactions, and it can behave poorly if predictors are highly correlated. The two other methods rely on decision trees (see Chapters 9 and 10 in Hastie, Tibshirani, and Friedman (2017)). Consider a simple tree example, if volume is above the 90th percentile, split on whether the bid-ask spread is above or below the median, and otherwise split on order imbalance's 30th percentile. Each of the four leaves is assigned an expected frequency of activist trading (the historical average).

Decision trees have many desirable features. For instance, they are invariant to variable scaling and they are robust to outliers. Random Forest takes an average over many random decision trees (Breiman (2001)). Each of these many trees is constructed on a sample bootstrapped from a random subset of all predictors from the original data set (e.g., 10 out of 100; see Chapter 15 in Hastie, Tibshirani, and Friedman (2017)). Our preferred method is eXtreme Gradient Boosting (XGBoost, Chen and Guestrin (2016)), which efficiently implements Gradient Tree Boosting (see Chapter 10 in Hastie, Tibshirani, and Friedman (2017)). While Random Forest averages over random trees ("bagging"), in Gradient Tree Boosting each new tree focuses on examples that previous trees find problematic ("boosting"). In general, boosting produces better forecasts than bagging but is slower to estimate.

XGBoost makes Gradient Tree Boosting almost as fast as Random Forest. It also recognizes that trees are prone to overfitting and penalizes trees with many leaves in favor of simpler, shorter trees (i.e., regularization). Regularization makes the models perform worse in-sample, but improves out-of-sample performance, which is the goal. We use the scikit-learn package in Python that implements Lasso and Random Forest and provides an XGBoost interface. While we focus on XGBoost because it yields the best performance, our main results continue to hold with Random Forest.[8]

Third, we pick 41 predictors (or "features") as described in Table IA.III. Four predictors come from CRSP daily files: stock price, return, absolute return, and trading volume. The remaining predictors are based on intraday data from TAQ. The Wharton Research Data Services (WRDS) Intraday Indicators database aggregates TAQ data at the stock-day level into 289 variables (many of which are near duplicates), from which we select 21 unique variables motivated by the microstructure literature. We supplement these factors with 16 variables that are missing from WRDS Indicators, such as depth, morning and afternoon returns, and realized volatility. We pick a limited number of predictors motivated by microstructure theories instead of directly estimating an ML method on every quote update and trade in TAQ. We prefer fewer predictors because statistical power is limited in our setting with a training sample of about 60,000 stock-days and with informed investors actively trying to avoid detection. Also, having interpretable predictors such as volume and volatility helps us explain how ITI measures work. All features are standardized by

[8] Chen and Guestrin (2016) introduced XGBoost and accumulated over 10,000 Google citations in just a few years. When it comes to small to medium structured/tabular data, XGBoost and similar decision tree-based algorithms are considered best-in-class at the moment. About 500 developers in XGBoost's GitHub community translated the method to all major programming languages.

subtracting their average and dividing by their standard deviation over the prior year. The standardization makes features comparable across stocks as well as easier to interpret: if today's volume is three standard deviations above the average, what does that mean for (today's) informed trading? Note that since our focus is on detecting realized informed trading rather than estimating expected informed trading, we use contemporaneous rather than lagged predictors of informed trading.[9]

Cross-validation helps us avoid overfitting and keeps the analysis out-of-sample. Specifically, we follow a standard approach in ML and split the Schedule 13D events into five nonoverlapping parts in calendar time, setting one part (20% of the data) aside for evaluation.[10] With the remaining 80% of the data, we use standard cross-validation to find the set of parameters that balances in-sample and out-of-sample performance. That is, we again split the 80% into five parts, estimate a model on four parts, and evaluate its performance on the remaining part. Once we settle on the model, we return to the 20% of the data that we set aside at the beginning and evaluate the model on a set of observations that it had not previously touched. We then rotate the last part of the data that we set aside and repeat the analysis. This allows us to evaluate model performance out-of-sample on the entire Schedule 13D sample.

In the second part of our analysis, we estimate the model on the full training data set and extrapolate it to the entire cross section of stocks between January 1993 and July 2019. Variables are processed exactly as for the training sample and are then supplied to the estimated model that yields in turn the ITI for a given stock-day. This extrapolation exercise makes implicit assumptions that we discuss in Section II.C.

B. Properties of ITI

In this section, we discuss the properties of ITI trained on trades by Schedule 13D filers. We regress an indicator for Schedule 13D trading on ITI and other liquidity measures. All of our specifications include filing fixed effects. Most specifications control for the effective bid-ask spread, price impact, Kyle's lambda, market depth, realized volatility, turnover, order imbalance, and absolute order imbalance, which are described in Table IA.IV. We control for order imbalance in addition to absolute order imbalance to capture simple "lean against the wind" effects. For example, activists mostly buy and thus could mostly trade against sell imbalances. Table IA.V reports descriptive statistics for the 13D sample.

[9] Predicting informed trading is likely much harder than detecting it ex post. Informed traders want to hide their trades. As a result, the set of variables that predicts informed trading likely varies over time. To illustrate, we construct ITI to predict today's informed trading using only yesterday's data (i.e., $t-1$ data instead of t data). We find that ITI($t-1$)'s power to detect Schedule 13D trades on day t is about 13 times lower than that of ITI(t). This is not surprising, as by analogy one can easily tell ex post whether a day had high volatility, but the ex ante volatility forecast will perform much worse.

[10] Other alternatives for splitting the sample yield similar results.

Table II

Does ITI Detect Schedule 13D Trading Days?

In columns (1) to (3), an indicator for days with Schedule 13D trading over the filing windows (60 days before the filing date to the filing date) is regressed on a set of liquidity variables and filing fixed effects. In columns (4) to (6), 13D filer turnover (share volume traded by the filer divided by total shares outstanding), is regressed on a set of liquidity variables and filing fixed effects, conditional on Schedule 13D trading on that specific stock-day. ITI is informed trading intensity. Effective spread (ES), lambda, depth, realized volatility (rvol), order imbalance (OI), absolute order imbalance (|OI|), and return are winsorized at 0.05% and 99.95%. The sample consists of 1,593 Schedule 13D filings between 1994 and 2018. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. Standard errors are clustered by filing, and associated *t*-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

	Day with Schedule 13D Trading			Schedule 13D Turnover		
Dep. Variable:	(1)	(2)	(3)	(4)	(5)	(6)
ITI	0.697***		0.604***	0.007***		0.004***
	(43.572)		(36.648)	(22.184)		(13.657)
ES		−3.412***	−0.807		0.099***	0.108***
		(−4.053)	(−1.098)		(3.179)	(3.505)
lambda		−1.186***	−0.589***		−0.024***	−0.020***
		(−4.822)	(−2.599)		(−5.275)	(−4.673)
depth		34.360***	21.323***		−0.096	−0.132
		(8.996)	(8.037)		(−1.135)	(−1.603)
rvol		−0.832***	−0.361***		0.000	0.001
		(−6.149)	(−3.475)		(0.111)	(0.329)
turn		1.377***	0.640***		0.039***	0.037***
		(9.845)	(6.110)		(8.284)	(7.885)
OI		0.379	0.294		0.073***	0.071***
		(0.748)	(0.703)		(3.065)	(3.092)
|OI|		6.732***	2.613***		0.180***	0.161***
		(10.534)	(4.781)		(6.329)	(5.811)
ret		−0.083	−0.030		−0.008***	−0.007***
		(−1.258)	(−0.511)		(−3.375)	(−3.064)
Filing FE	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.0986	0.0461	0.1073	0.0692	0.2579	0.2756
Obs.	58,197	58,197	58,197	20,969	20,969	20,969

Column (1) of Table II shows that ITI predicts Schedule 13D trades, with a *t*-statistic greater than 43 and R^2 of 9.86%. While this number is much lower than 100%, it is large in this context. Indeed, many microstructure models assume that informed investors hide behind uninformed trading to avoid detection, and a high R^2 would directly contradict this basic assumption. Thus, ITI measures should be benchmarked against other alternative measures.[11] We benchmark ITI against alternative measures in column (2)

[11] To provide another perspective, if we consider an accuracy metric such as the area under the receiver operating characteristic curve (ROC AUC), ITI has a 71% probability of ranking a randomly selected day with 13D trading above a randomly selected day without 13D trading (reported in Table IA.VI). A random classifier has a 50% ROC AUC.

of Table II, which reports estimates of the same regression with our set of common liquidity variables. The adjusted R^2 of the regression is 4.61%, less than half that achieved by ITI alone. Effective spreads, Kyle's lambda, depth, and realized volatility all suggest improved liquidity on days with Schedule 13D trading. Thus, these common liquidity variables are not able to detect Schedule 13D trades (Collin-Dufresne and Fos, 2015). Again, a natural explanation for this result is that Schedule 13D traders strategically time their trades (Collin-Dufresne and Fos, 2015, 2016). Column (3) shows that the inclusion of common liquidity variables in addition to ITI increases the R^2 from 9.86% (with ITI alone) to 10.73%. The coefficient on ITI decreases only slightly when we control for common liquidity measures.

While we use ITI as a measure of ITI, does it also pick up Schedule 13D trading intensity? To answer this question, we use Schedule 13D trading volume data to compute the share turnover of the Scheduler 13D filer (share volume traded divided by total shares outstanding). That is, Schedule 13D turnover. We then regress Schedule 13D turnover on ITI and control variables. To avoid restating the results in columns (1) to (3), we restrict the sample to days when Schedule 13D filers trade. Hence, we focus on the intensive margin.

The last three columns of Table II report the results. Column (4) shows that ITI is strongly positively associated with Schedule 13D turnover, with a t-statistic greater than 22. The positive and significant relationship between ITI and Schedule 13D turnover remains when we control for common liquidity measures. This result is not a mechanical result because we do not incorporate Schedule 13D trading volume information when training the model. While we could use this information, we choose to keep the classification problem as simple as possible. Overall, ITI is associated with informed trading on both the extensive and intensive margins, even when controlling for common liquidity, volume, and volatility measures.

In summary, Table II shows that ITI captures informed trading not detected by standard liquidity measures. We discuss additional measures of informed trading in Section III.C and reach similar conclusions. Also, even if part of ITI is subsumed by, say, turnover, this is a feature of our methodology. ITI can be interpreted as a measure of ITI, whereas it is difficult to interpret turnover in a similar way (e.g., Duarte, Hu, and Young, 2020).

Markets are changing over our sample period due to technological innovations such as algorithmic trading, lower transaction costs, and higher competition. Is the explanatory power of ITI driven mostly by specific parts of our sample? To address this concern, Figure 1 plots the out-of-sample R^2 when predicting Schedule 13D trades with ITI year by year.[12] The figure compares the R^2 produced by ITI with the R^2 produced by a set of common liquidity variables. ITI's ability to predict this class of informed trades is relatively stable over time and does not display any trend. In contrast, common variables' ability to predict activist trades declines slowly over the sample period. Hence, the predictability gap between ITI and common measures tends to be larger

[12] Figure IA.2 shows the distribution of observations over time for the Schedule 13D filings.

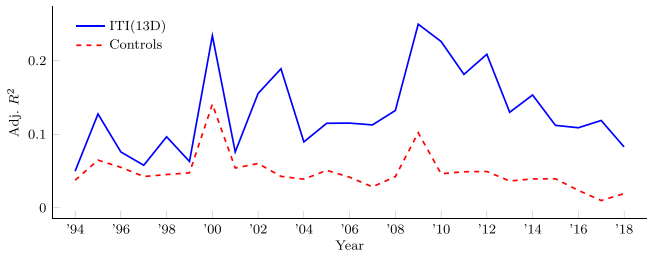

Figure 1. Stability of the algorithm over time in the Schedule 13D sample. This figure reports the adjusted R^2 estimated each year from two specifications: an indicator for Schedule 13D trading over the filing windows (60 days before the filing date to the filing date) is regressed on either informed trading intensity trained on Schedule 13D data (solid line) or a set of control variables (dashed line). Control variables are effective spread, price impact, lambda, depth, realized volatility, turnover, order imbalance, absolute order imbalance, and return. The sample consists of 1,593 Schedule 13D filings between 1994 and 2018. (Color figure can be viewed at wileyonlinelibrary.com)

post-2001. While several factors may be at play, one interesting possibility is that this trend originates from the rise of algorithmic trading postdecimalization. Algorithmic trading makes it easier for informed traders to camouflage their trades, which are therefore harder to detect with a simple linear specification. Overall, the predictive power of the ITI measure is not driven by a part of our sample period.

It is often difficult to explain why ML methods work (e.g., Nagel (2021)). To address this concern, we identify ITI's most important components. We split the model's features, listed in Table IA.III, into four groups, namely, liquidity, return, volatility, and volume variables. We then train the model using only one subset of the variables to predict Schedule 13D trades. We therefore obtain four subset-specific ITI variables. Table III reports the results. The volume grouping dominates all other groupings. Return and volatility variables have low explanatory power. On their own, liquidity variables such as spread, price impact, and depth achieve out-of-sample explanatory power that is roughly half that of ITI constructed from all the variables. Nevertheless, column (6) of Table III shows that none of the groupings is subsumed by the other groupings. This result suggests that one type of information cannot fully capture the occurrence of Schedule 13D trades. In particular, as we show below, the *interaction* between volume and volatility plays an important role in detecting Schedule 13D trades.

We also confirm the results above by performing traditional ML procedures to rank features by importance. First, we use XGB's internal procedure that ranks features based on their gain (a default option based on the average gain across all splits in which a feature was used). Figure IA.3 reports the ranking and shows that volume is by far the most important variable followed by volatility. The other variables are not far behind. Another popular ranking method, SHapley Additive exPlanations (SHAP) (Lundberg and Lee (2017)) is inspired by cooperative game theory. For each observation x and feature f, a

Table III
What Variables Help Detect Informed Trading?

An indicator for days with Schedule 13D trading over the filing window (60 days before the filing date to the filing date) is regressed on informed trading intensity (ITI) measured from a subset of the variables in Table IA.III and filing fixed effects. Schedule 13D trade is an indicator variable that takes the value one on days when the filer trades. ITI(liquidity), ITI(return), ITI(volatility), and ITI(volume) are versions of ITI that are trained using a subset of the explanatory variables. The sample consists of 1,593 Schedule 13D filings between 1994 and 2018. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. Standard errors are clustered by filing, and associated *t*-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

	Day with Schedule 13D Trading					
Dep. Variable:	(1)	(2)	(3)	(4)	(5)	(6)
ITI	0.697***					
	(43.572)					
ITI(liquidity)		0.596***				0.289***
		(24.561)				(17.945)
ITI(return)			0.510***			0.239***
			(18.963)			(14.256)
ITI(volatility)				0.498***		0.208***
				(19.266)		(12.912)
ITI(volume)					0.671***	0.526***
					(39.235)	(36.533)
Filing FE	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.0986	0.0412	0.0264	0.0278	0.0769	0.1043
Obs.	58,197	58,197	58,197	58,197	58,197	58,197

SHAP value measures a weighted-average gain from adding *f* to all possible feature subsets. A separate model must be trained for each possible subset of features, which is computationally expensive. Figure IA.4 ranks features according to SHAP and plots the distribution of SHAP values over observations in the 13D sample. Volume is again the top variable, but other variables are not far behind.

As explained in Section II.A, our ITI measure relies on a GBT algorithm. Table IA.VI shows that GBT outperforms Lasso and Random Forests when predicting Schedule 13D trades. The out-of-sample R^2 increases from 6.44% with Lasso to 9.74% with Random Forests, and increases further to 13.68% with GBT. Standard linear regression performs very similarly to Lasso in our sample. As mentioned above, our main results are robust to using Random Forests to construct ITI. Table IA.VI also shows that Lasso is subsumed by Random Forests and GBT, which suggests that nonlinearities matter for detecting informed trading.

In Figure 2, we examine nonlinearities with partial dependence plots (see Section 10.13 in Hastie, Tibshirani, and Friedman (2017)). These plots show how ITI depends on a variable of interest marginalizing over the values of all other input variables. We pick volume and volatility for the plots because the XGBoost algorithm ranks them as the most important determinants of ITI in

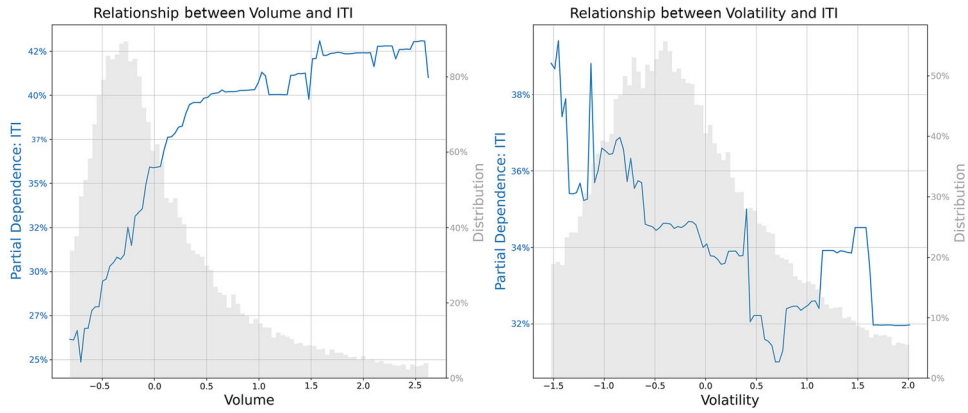

Figure 2. Partial dependence plots of ITI on volume and volatility. The top panel shows that ITI is increasing and concave in volume. The bottom panel shows that ITI is decreasing and convex in volatility. The left *Y*-axis reports ITI, which is shown in a solid line. The right *Y*-axis reports the distribution of a given variable, which is shown in gray bars. The range of values in the *X*-axis spans the distribution of a given variable. Variables are standardized before computing ITI; hence, zero volume means that volume on this day matches the previous one-year average. Partial dependence for ITI if volume $= x$ is computed in two steps. First, volume is set to x for each observation in the Schedule 13D sample and ITI is computed. Second, ITI is averaged over all observations. Partial dependence for ITI if volatility $= x$ is computed in a similar way. Variables are defined in Table IA.III in the Internet Appendix. (Color figure can be viewed at wileyonlinelibrary.com)

Figure IA.3. ITI is increasing and concave in volume and is decreasing and convex in volatility.

We next examine how variable interactions affect ITI with the help of a surrogate tree, another standard ML technique (Craven and Shavlik (1995)). ITI is determined by a complex model, but it can be approximated by a simple three-level tree, a piecewise constant function with just eight values (leaves). Figure 3 shows the results. The tree aims to capture ITI's most important properties and explains about one-third of ITI's variation. It produces two insights. First, how often a variable enters a tree and how high it is in the tree can help assess its importance. Figure 3 shows that volume enters the tree three times, including at the top level. Volatility and illiquidity determine the second level. Second, ITI is particularly high if volume is high but volatility is low. Similarly, ITI is particularly low if volume is low and illiquidity is high. This illustrates how ITI takes into account various nonlinearities and interactions between input variables.

As another way to assess the importance of interactions, we regress ITI on the ITI subset measures that we use in Table III. The results are reported in Table IA.VII. ITI(volume) explains about 35% of the (within-filing) variation in ITI. Furthermore, regressing ITI on all of the subset measures achieves an adjusted R^2 of about 47%. Thus, interactions across types of explanatory variables are likely important to explain variation in ITI.

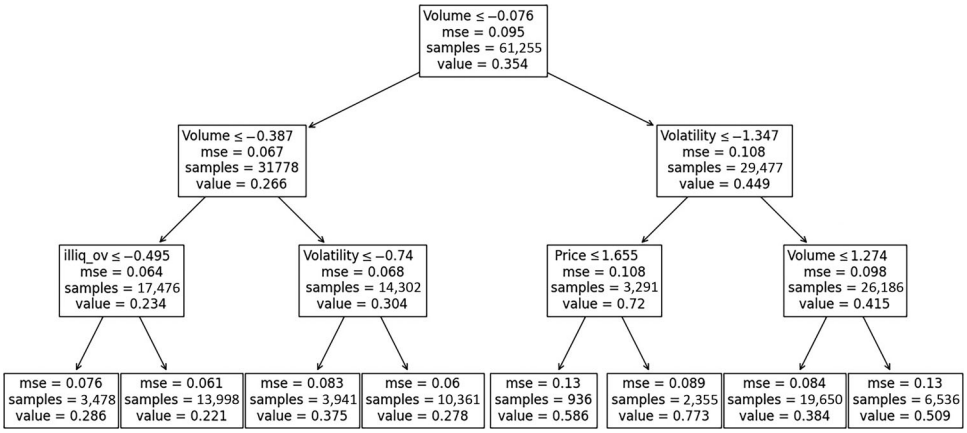

Figure 3. Simple tree that approximates ITI. A regression tree with three levels predicts ITI for the sample of Schedule 13D trades. At each step, a tree picks a variable and a split level that minimizes mean squared error, a greedy algorithm, while encouraging equally sized splits. Each node reports the splitting variable and criteria followed by a measure of fit (MSE), number of observations, and the predicted value of ITI. The tree explains about one-third of the total variation in ITI. To apply the tree, we start at its top and move to the right if a splitting criterion is satisfied until we reach one of the terminal nodes. Variables are defined in Table IA.III in the Internet Appendix.

We conclude this section by providing an additional piece of evidence regarding the role of trading volume. Duarte, Hu, and Young (2020) show that trading volume is a key driver of several existing measures of informed trading. While the results in Table III show that volume-related variables are not the only drivers of ITI, we next report results of a matching test based on the Schedule 13D sample. Specifically, within each 13D filing, each day with a Schedule 13D trade ("treated") is matched to a nontreated observation based on turnover. Only paired observations whose absolute difference in turnover is less than 0.00002 are retained. This threshold is selected such that the within-filing difference in turnover between treated and matched days is statistically insignificant.

Table IV presents the results. Column (1) of Panel A shows that treated and matched days have identical levels of turnover. However, the positive and significant coefficient of ITI in column (2) indicates that ITI is higher on days with informed trading even across days matched on turnover. In Panel B, an indicator for days with Schedule 13D trading is regressed on a set of liquidity variables and filing fixed effects in the matched sample. ITI detects days with informed trading across days matched on turnover, whereas none of the control variables is statistically significant. We therefore conclude that ITI carries information not captured by trading volume.

Table IV
Turnover Matching in the 13D Sample

Within each 13D filing, each treated observation (day with Schedule 13D trade) is matched to a nontreated observation based on turnover. Only paired observations whose absolute difference in turnover is $\leq$0.00002 are retained. This threshold is selected such that the within-filing difference in turnover between stock-days with 13D trade and stock-days without 13D trade is statistically insignificant, as shown in Panel A. In Panel B, an indicator for days with Schedule 13D trading over the filing windows (60 days before the filing date to the filing date) is regressed on a set of liquidity variables and filing fixed effects in this restricted sample. ITI is informed trading intensity. Effective spread (ES), lambda, depth, realized volatility (rvol), order imbalance (OI), absolute order imbalance (|OI|), and return are winsorized at 0.05% and 99.95%. The unrestricted sample consists of 1,593 Schedule 13D filings between 1994 and 2018. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. Standard errors are clustered by filing, and associated *t*-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

Panel A: Difference in Turnover and ITI

Dep. Variable:	Turnover	ITI
Day with Schedule 13D trading	0.000	0.033***
	(1.027)	(5.253)
Filing FE	Yes	Yes
Adj. R^2	−0.0004	0.0163
Obs.	2,388	2,388

Panel B: What Variables Help Detect Informed Trading?

Dep. Variable:	Day with Schedule 13D Trading		
ITI	0.509***		0.507***
	(5.445)		(5.284)
ES		−2.987	−0.380
		(−0.549)	(−0.072)
lambda		−0.767	−0.505
		(−0.689)	(−0.448)
depth		23.136	24.205
		(0.549)	(0.770)
rvol		−2.099	−1.398
		(−1.211)	(−0.827)
OI		−2.444	2.237
		(−0.200)	(0.187)
|OI|		2.532	−11.174
		(0.156)	(−0.697)
ret		−0.319	−0.422
		(−0.441)	(−0.596)
Filing FE	Yes	Yes	Yes
Adj. R^2	0.0163	−0.0003	0.0149
Obs.	2,388	2,388	2,388

C. Extrapolation to the Full Sample of Stocks

The main advantage of our approach is that ITI can be extrapolated beyond the training sample of informed trades to the full sample of U.S. common stocks. Specifically, ITI can be computed from the model estimated on a particular class of informed trades as long as the input variables are observed, which is limited only by the availability of TAQ data. We next perform this extrapolation exercise for ITI trained on Schedule 13D trades. In all of the results for the full sample of stocks, Schedule 13D trading periods are excluded since they are used to train the model. This step results in 0.35% of the sample being dropped.

The extrapolation assumes that the relations between intraday variables and realized informed trading learned by ITI largely hold in the full sample. Since ITI is trained on a restricted sample of Schedule 13D trades, a natural concern is that ITI could capture uninformed trading when applied to larger unrestricted samples. The issue of false positives applies to all ML problems, but it is more challenging to evaluate here since informed trading is unobservable even ex post. We acknowledge that, in addition to capturing informed trading, ITI could also reflect uninformed trading. We conduct a variety of validation tests and show that ITI outperforms existing measures.

Column (1) of Table IA.VIII shows how ITI correlates with standard liquidity measures for the full sample of common stocks. ITI is negatively related to the effective spread and lambda, while it is positively related to depth, turnover, and absolute order imbalance. These relations echo those between Schedule 13D trades and liquidity measures in Table II. The R^2 of the full-sample regression is around 12%. Hence, within-stock variation in ITI is not well captured by standard liquidity measures.

D. Informed Trading ahead of Information Events

We next study the dynamics of ITI ahead of information events. We first investigate the behavior of ITI around earnings announcements, an example of a prescheduled information event. ITI is regressed on indicator variables for each of the 10 days before and 10 days after an earnings announcement and stock fixed effects. Panel A of Figure 4 plots the coefficients on the indicator variables with 95% confidence intervals that can be interpreted as average changes in the value of ITI on the days around an earnings announcement relative to the average stock-specific ITI value outside of these days.

As can be seen in Panel A of Figure 4, ITI is statistically higher two days ahead of an earnings announcement, spikes on the day of the announcement, and then remains elevated for several days. In Table II, ITI is negatively associated with most liquidity measures. One may be tempted to conclude that ITI is simply an aggregated liquidity measure. The increase in ITI ahead of earnings announcements disproves this view. As is well known, market makers increase spreads and lower depth before earnings announcements

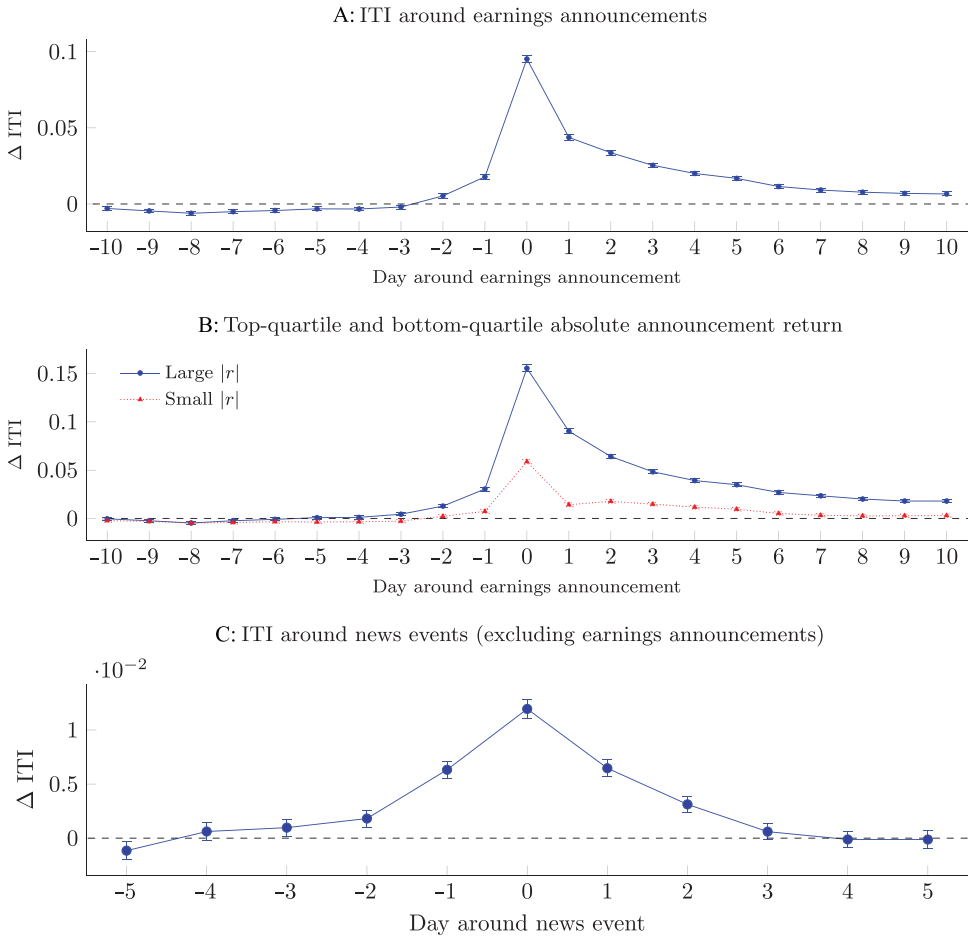

Figure 4. ITI around earnings announcements and news events. In Panel A, informed trading intensity is regressed on indicator variables for days around earnings announcements and stock fixed effects. Panel B reports results separately for announcements that are in the top and bottom quartiles of absolute announcement-day return. The sample includes common stocks from January 1993 to July 2019. In Panel C, informed trading intensity is regressed on indicator variables for days around news events (excluding earnings announcements) and stock fixed effects. News data are obtained from Boudoukh et al. (2019) and cover S&P 500 common stocks from 2000 to 2015. The figure reports 95% confidence intervals based on standard errors that are double-clustered by stock and date. (Color figure can be viewed at wileyonlinelibrary.com)

(e.g., Lee, Mucklow, and Ready, 1993). Hence, while ITI picks up the properties of informed trading, it is not equivalent to a liquidity measure.

Campbell, Ramadorai, and Schwartz (2009) measure daily institutional trading from trade size combined with a buy-sell classification algorithm and find increased institutional trading in the direction of future earnings surprises 40 days prior to earnings announcements. This is not inconsistent

with our results. Figure 4 shows the difference in realized informed trading across days for a given stock, which suggests that ITI remains constant on average until a few days prior to earnings announcements. This is broadly consistent with Campbell, Ramadorai, and Schwartz (2009), who show that cumulative institutional flows increase linearly in the direction of future earnings surprises.

The pattern in ITI is consistent with informed trading prior to the announcement but also following the announcement. This result is consistent with the model of Kim and Verrecchia (1994), in which the disclosure of public news increases information asymmetry among investors since some investors can better process new information than other investors. Another possibility is that informed traders are better able to camouflage their trades following earnings announcements thanks to the higher postannouncement trading activity, as in the model of Collin-Dufresne and Fos (2016). Nevertheless, since the pattern is robust to controlling for turnover, ITI does not simply pick up higher turnover around earnings announcements. Empirically, an increase in informed trading following an earnings announcement is consistent with the results of Lee, Mucklow, and Ready (1993), Back, Crotty, and Li (2018), and Brennan, Huh, and Subrahmanyam (2018), among others.

We next investigate whether the dynamics of ITI depend on the content of earnings announcements and consider announcement-day returns. Panel B of Figure 4 plots the results for announcements that are in the top and bottom quartiles of absolute announcement-day returns. ITI increases most strongly ahead of earnings announcements associated with large absolute surprises. Higher informed trading could be associated with a smaller announcement surprise as more information is incorporated into price ahead of the announcement. Our results suggest otherwise. In a robustness check, we obtain qualitatively similar results using a measure of earnings surprise based on the median analyst forecast (which does not depend on the announcement return).

Earnings announcements are prescheduled and may not be representative of information events in general. We next examine the days prior to news releases, the timing of which is generally unknown to uninformed investors. News data are obtained from Boudoukh et al. (2019) and cover S&P 500 common stocks from 2000 to 2015.[13] ITI is regressed on indicator variables for days around news as well as stock fixed effects. To ensure that we are not picking up the results in Panel A, we exclude five-day windows centered on earnings announcement days. Panel C in Figure 4 plots the coefficients with 95% confidence intervals. ITI increases before news, and this increase is statistically significant. ITI then spikes on the day of the announcement and remains elevated for several days.

We find similar evidence using M&A announcements. Figure IA.5 plots ITI in the 10 days preceding M&A announcements. ITI starts increasing five days before M&A announcements. Like Brennan, Huh, and Subrahmanyam (2018), we also find a dramatic increase in informed trading after M&A

[13] We thank the authors for making this data set available.

announcements. Overall, ITI increases ahead of both scheduled and unscheduled information events. This finding supports the view that ITI is effective in detecting informed trading.

To conclude this section, we use one of the most fundamental intuitions on how informed trades and liquidity trades affect prices. The price impact of informed trades should be permanent, whereas the price impact of liquidity-motivated trades should be transient (Hasbrouck (1988, 1991)). Liquidity providers set prices to reflect the expected intensity of informed trading. If the realized intensity of informed trading (as measured by ITI) is below the expected level, then prices will partially revert. To illustrate, in the seminal model of Glosten and Milgrom (1985) with a risk-neutral market maker, price changes are uncorrelated, independent of the probability of informed trading *expected* by the market maker. However, conditional on knowing that a specific price change is associated with an informed (uninformed) trade, the following price change exhibits on average continuation (reversal). We expect returns on days with high realized informed trading to reverse less than returns on other days, which we can test with ITI in the panel regression

$$r_{i,t+1} = a_t + b_1 * r_{i,t} + b_2 * \mathrm{ITI}_{i,t} + b_3 * \mathrm{ITI}_{i,t} * r_{i,t} + \text{controls} + e_{i,t+1}, \quad (1)$$

where $r_{i,t}$ is the return of stock i on date t.[14] We expect $b_3 > 0$ if there is less reversal on days with higher ITI. Table V reports the results. Column (1) shows that a higher return on day t is followed by a lower return on day $t + 1$. In columns (2) and (3), we interact ITI with the stock return. Higher ITI is associated with lower return reversal. This result is consistent with ITI capturing days with informed trading. In terms of economic magnitude, the within-date standard deviation of ITI is 0.15. Using the values in column (2), a two (within) standard deviation increase in ITI reduces return reversal by about two-thirds.

In column (3), we add interactions between return and turnover, return and volatility, and return and effective bid-ask spread. According to inventory models, return reversal should increase with turnover and volatility (e.g., Campbell, Grossman, and Wang (1993), Llorente et al. (2002), Nagel (2012)). To further control for liquidity effects such as the bid-ask bounce, we also include an interaction between return and effective spread. Importantly, the ITI-return interaction is not affected by including these liquidity controls. Moreover, the ITI-return interaction is positive while the other interactions are zero or negative. These results suggest that ITI captures informed trading rather than inventory shocks or microstructure effects. ITI does not simply proxy for volume, volatility, or liquidity.[15] Hence, the results in Table V support the view that ITI is effective in detecting days with informed trading.

[14] Duarte, Hu, and Young (2020) conduct a similar test with other informed trading measures.

[15] Table IA.IX uses as the dependent variable the future return up to 10 trading days after the initial return. Coefficients are positive and statistically significant for the first two days, positive but generally insignificant between three and six days, and a mix of positive and negative without statistical significance after six days. This pattern does not support the idea that persistent liquidity shocks drive the results.

Table V
Return Reversal

Daily return is regressed on lagged return, ITI, control variables, date fixed effects, and interactions of lagged returns with ITI, turnover, realized volatility, and effective spread. Control variables are effective spread, price impact, lambda, depth, realized volatility, turnover, order imbalance, and absolute order imbalance. Control variables are winsorized at 0.05% and 99.95%. The sample includes common stocks from January 1993 to July 2019. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. Standard errors are double-clustered by stock and date, and associated *t*-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

	ret$(t+1)$		
Dep. Variable	(1)	(2)	(3)
ret	-0.004^{*}	-0.014^{***}	-0.006^{**}
	(-1.774)	(-4.477)	(-1.977)
ITI(13D)		0.002^{***}	0.002^{***}
		(23.553)	(20.001)
ret*ITI(13D)		0.031^{***}	0.034^{***}
		(6.445)	(5.830)
ret*turn			0.008
			(0.541)
ret*rvol			-0.029^{*}
			(-1.652)
ret*ES%			-0.756^{***}
			(-4.936)
Controls	No	No	Yes
Date FE	Yes	Yes	Yes
Adj. R^2	0.0000	0.0002	0.0004
Obs.	16,814,168	16,814,168	16,814,168

III. The Economics of ITI

In this section, we explain how our data-driven approach can shed light on the economics of informed trading. We first show that there is an important distinction between patient and impatient trading, in line with theory. We next show that the methodology can shed light on commonality in informed trading. In particular, ITI is effective in detecting various types of informed trading beyond what can be achieved with standard measures. We then discuss what ITI suggests standard models of informed trading are missing and compare ITI to existing measures of informed trading.

A. Patient and Impatient Trading

An important theoretical distinction can be made between impatient trading and patient trading. Foucault, Kadan, and Kandel (2005) show that traders' patience is a key determinant of limit order book dynamics in a model without asymmetric information. Kaniel and Liu (2006) show that the horizon of

private information is crucial for informed traders' choice between market orders and limit orders. As the horizon increases, the probability of informed traders using limit orders increases, which affects liquidity measures such as the bid-ask spread. Caldentey and Stacchetti (2010) study a version of the Kyle (1985) model in which the asset value is publicly disclosed at a random time. In this model, the insider trades more aggressively when the expected time until disclosure decreases.[16]

In the 13D setting, after an investor reaches 5% ownership in a security, the investor must disclose their positions within 10 days. The first part of the 13D window can be thought of as insider trading with an (almost) infinite expected horizon of information disclosure. We therefore expect 13D filers to trade more aggressively after reaching the 5% ownership threshold. Indeed, the cumulative return pattern in Figure IA.1 is consistent with this idea as the bulk of the abnormal return is earned close to the filing date. We use this feature of the data to decompose ITI(13D) into a "patient" ITI and an "impatient" ITI. To do so, we estimate ITI(patient) using the first 40 days of the filing window and ITI(impatient) using the last 20 days of the filing window.[17] Consistent with greater impatience, activists trade on average 37% of daily volume on 49% of days in the last 20 days of the filing window versus 20% of daily volume on 30% of days in the first 40 days.

Both ITI(impatient) and ITI(patient) are effective in detecting days with Schedule 13D trading, as reported in Table IA.X. Moreover, both measures are higher when Schedule 13D filers purchase a larger number of shares on a given day. Importantly, ITI(patient) and ITI(impatient) are not the same measure. Their unconditional correlation is 0.47. Our set of control variables explains 19.38% of the variation in ITI(impatient) but only 8.68% of the variation in ITI(patient), as reported in Table IA.VIII. This is consistent with the idea that patient (strategic) trading is less dependent on market conditions.

In Table VI, we contrast the relation of ITI(impatient) and ITI(patient) with different variables and informational events. In columns (1) and (2), we regress turnover and realized volatility on ITI(impatient) and ITI(patient). Whereas ITI(impatient) and ITI(impatient) are both positively correlated with turnover, the relation for ITI(impatient) is stronger. Furthermore, ITI(impatient) is positively associated with realized volatility, whereas ITI(patient) displays the opposite pattern. These findings are consistent with ITI(impatient) detecting days with aggressive informed trading. In line with the theory of Caldentey and Stacchetti (2010), an impatient informed trader trades more aggressively and thus generates increased volatility relative to a patient informed trader.

[16] Bolandnazar et al. (2020) provide evidence consistent with the Caldentey and Stacchetti (2010) model using information accidentally disclosed to some investors a few seconds to a few minutes ahead of the public.

[17] We use the last 20 days instead of the last 10 days to ensure that we have enough data points to train the measure.

Table VI
ITI(Patient) and ITI(Impatient)

Different variables are regressed on ITI(patient) and ITI(impatient) and control variables. ITI(patient) (ITI(impatient)) is the informed trading intensity estimated using the first 40 days (last 20 days) of the 60-day Schedule 13D filing window. News-2 and News-1 are indicators for the two days before a news announcement, excluding earnings announcements. News data are obtained from Boudoukh et al. (2019) and cover S&P 500 common stocks from 2000 to 2015. EA-2 and EA-1 are indicators for the two days before an earnings announcement. The earnings announcement day and the 10 days after the announcement are excluded from the regression. Illegal trade is an indicator variable for days with illegal insider trades ahead of information events. We consider a window of 120 days before the event, excluding the day before the event and the day of the event. This regression includes event fixed effects and event day fixed effects as in Ahern (2020). The sample includes 417 insider trades (column (5)). The sample includes common stocks from January 1993 to July 2019. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. Control variables are winsorized at 0.05% and 99.95% and include effective spread, price impact, lambda, depth, realized volatility, turnover, order imbalance, absolute order imbalance, and return. Standard errors are double-clustered by stock and date, and associated t-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

	Turnover (1)	Realized Volatility (2)	EA-2 (3)	EA-1 (4)	News-2 (5)	News-1 (6)	Illegal Trade (7)
ITI(patient)	0.006^{***}	-0.003^{***}	0.001^{***}	-0.001^{*}	-0.004	0.002	-0.010
	(56.727)	(-11.955)	(2.581)	(-1.920)	(-0.847)	(0.463)	(-1.249)
ITI(impatient)	0.028^{***}	0.014^{***}	0.007^{***}	0.021^{***}	0.006	0.027^{***}	0.021^{*}
	(98.857)	(37.125)	(9.693)	(21.614)	(0.871)	(3.700)	(1.803)
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Fixed effects	Stock	Stock	Stock	Stock	Stock	Stock	Event, event-day
Adj. R^2	0.1394	0.0100	0.0002	0.0020	0.0056	0.0092	0.0010
Obs.	18,062,838	18,022,699	14,159,702	14,159,702	1,652,880	1,652,880	18,695

We also conduct ML interpretability procedures that generally confirm the results above.[18] First, XGB ranks volume and volatility as the top two features for both ITI(impatient) and ITI(patient). Absolute order imbalance ranks third for ITI(impatient), while number of nontrading half-hours ranks third for ITI(patient), which is consistent with impatient investors being more likely to take liquidity. Second, the surrogate trees that approximate ITI(impatient) and ITI(patient) have a lot in common and resemble the ITI(13D) tree. However, the second level of the tree uses illiquidity for ITI(impatient) whereas it uses volume for ITI(patient).

We consider how ITI(impatient) and ITI(patient) change ahead of earnings announcements. Columns (3) and (4) in Table VI report the results. In the two days before an earnings announcement, ITI(impatient) increases strongly whereas ITI(patient) does not. Furthermore, columns (5) and (6) show that ITI(impatient) is significantly higher one day before unscheduled news events but no similar increase is observed for ITI(patient). These results support our methodology as ITI(patient) and ITI(impatient) behave in ways that are consistent with economic intuition.

We conclude by studying whether ITI measures detect *illegal* insider trades. We use a data set of 417 illegal insider trades and closely follow the specification in Ahern (2020).[19] Column (7) shows that ITI(impatient) is significantly associated with illegal insider trading. The next section further illustrates the power of ITI(impatient) to detect other classes of informed trading.

Overall, the results above indicate that the distinction between patient trading and impatient trading is important in the data, in line with theory. We argue that our data-driven approach allows us to effectively measure the distinction between patient trading and impatient trading.

B. Commonality in Informed Trading

Does ITI trained on a class of informed trading detect other classes of informed trading? In this section, we address this question to assess the external validity of our measure.

We consider two other types of informed trades: opportunistic insider trades and short sales. Prior work suggests that on average these trades are informed. Cohen, Malloy, and Pomorski (2012) find that nonroutine, or opportunistic, insider trades are informed in the sense that they predict future returns. We follow Cohen, Malloy, and Pomorski (2012) and use a sample of nonroutine insider trades, over the period 1993 to 2012, to build ITI(insider). For each insider trade, we pick the day before and the day after the trade as comparisons. This

[18] Features importance according to XGB's internal ranking and SHAP values for ITI(impatient) and ITI(patient) are shown in Figures IA.3 and IA.4. Surrogate trees for ITI(impatient) and ITI(patient) are shown in Figure IA.6.

[19] Part of this data set comes from Kenneth Ahern's website and is described in Ahern (2020). We thank him for making these data available. We also thank Patrick Augustin for sharing his insider trading data.

short window is motivated by the idea that broad market conditions change relatively little over three days. The final insider sample includes 779,007 stock-day observations.[20] Short sellers specialize in trading on negative information and tend to be informed. For example, Boehmer, Jones, and Zhang (2008) show that highly shorted stocks underperform less shorted stocks by 1.16% on average over the next month. We focus on days with a large increase in short interest to capture when short sellers establish their positions. Short interest equals the total quantity on loan divided by the number of shares outstanding. The daily short-sale data are from Markit and cover the period July 2006 to July 2019. To identify days with substantial short selling, we create an indicator that is set to one if the daily change in short interest exceeds the 90th percentile for the full sample. We randomly select 100,000 stock-days with short interest spikes (indicator equals one). Like for the insider sample, we add to the data set the day before and the day after the spike as comparisons.

Although opportunistic insider trades and spikes in short interest are taken to be driven by informed trading, these measures do not perfectly capture informed trading. For example, some short selling is motivated by hedging motives. Similarly, informed trading can occur on days around spikes in short interest. Our methodology assumes that realized informed trading is *on average* higher on days with insider trades and on days with spikes in short interest than on adjacent days.

In our main specification, we regress indicators for insider trades and spikes in short selling on ITI(13D) and stock fixed effects. The results are reported in Table VII. Columns (1) and (6) show that ITI(13D) detects opportunistic insider trades and spikes in short selling.[21] This result provides an external validation test of the measure.[22] To give some additional perspective, the relation between liquidity variables and trading indicators is not always consistent across indicators. As shown in Table IA.XIII, both effective spread and realized volatility are positively related to insider trading but negatively related to Schedule 13D trading. Hence, ITI captures a commonality across trading indicators that is not obvious from the relation between trading indicators and liquidity variables.

Fluctuations in the trading environment are likely to generate commonality in informed trading. Kadan and Manela (2020) show that the value of information to a trader is a function of volatility and liquidity. Hence, informed investors can trade in a common way simply due to fluctuations in liquidity. To assess the importance of this channel, we add liquidity controls to the regression. Columns (2) and (7) show that these liquidity controls explain only about one-quarter to one-third of ITI(13D)'s ability to detect insider and short trades.

[20] Because insider trades can occur on successive days, insider trades account for more than a third of stock-days (i.e., 44%) in this sample (Table I).

[21] In Table IA.XI, we show that ITI(13D) is able to separately detect insider purchases and insider sales.

[22] In Table IA.XII, we show that ITI(13D) detects Schedule 13D trades and insider trades in the full stock-day sample, which is much noisier than the subsamples considered in Tables II and VII. We thank an anonymous referee for suggesting this test.

Table VII

Do ITI Measures Detect Various Classes of Informed Trading?

Indicator variables for days with 13D trading, days with opportunistic insider trading, or days with a spike in short interest are regressed on informed trading intensity measures and control variables. ITI(13D) is trained on Schedule 13D data. ITI(patient) (ITI(impatient)) is trained using the first 40 days (last 20 days) of the 60-day Schedule 13D filing window. ITI(insider) is trained on opportunistic insider trading data. ITI(short) is trained on short-selling data. The 13D sample consists of the filing period for 1,593 13D filings between 1994 and 2018. The insider sample consists of two days surrounding opportunistic insider trades between 1993 and 2012. The short sample consists of two days surrounding 100,000 randomly selected spikes in short interest from June 2006 to December 2010. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. Control variables are winsorized at 0.05% and 99.95% and include effective spread, price impact, lambda, depth, realized volatility, turnover, order imbalance, absolute order imbalance, and return. Standard errors are clustered by filing for the Schedule 13D sample and by stock for the other samples, and associated t-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

	Insider Trade					Δ Short Interest				
Dep. Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ITI(13D)	0.150***	0.116***	0.073***			0.178***	0.118***	0.069***		
	(42.182)	(30.674)	(19.470)			(27.422)	(16.758)	(9.634)		
ITI(insider)			0.354***		0.325***					
			(47.882)		(43.406)					
ITI(short)								0.452***		0.387***
								(25.921)		(20.973)
ITI(patient)				0.032***	0.019***				0.025***	0.016**
				(7.325)	(4.503)				(3.173)	(2.001)
ITI(impatient)				0.178***	0.117***				0.211***	0.134***
				(34.951)	(23.038)				(22.194)	(13.360)
Controls	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.0025	0.0046	0.0082	0.0058	0.0086	0.0033	0.0077	0.0108	0.0094	0.0114
Obs.	779,007	779,007	779,007	779,007	779,007	216,979	216,979	216,979	216,979	216,979

For example, the estimated ITI coefficient when detecting insider trades declines from 0.150 without controls to 0.116 with controls.

ITI(13D)'s ability to detect insider and short trades beyond control variables could stem from the fact that ITI(13D) accounts for nonlinearities in the relation between informed trading and liquidity variables, which are not picked up by the linear specification. It is also possible that ITI(13D) detects insider trades because the way activists trade contains incremental information that can help detect other types of informed trading. To test these explanations, we repeat the procedure described in Section II to obtain two additional ITI measures: ITI(insider) and ITI(short). Tables IA.XIV and IA.XV show that the methodology works on these other data sets: ITI(insider) (ITI(short)) strongly detects insider trades (spikes in short selling) out of sample even when controlling for standard liquidity variables.[23] We note, however, that R^2 is lower in these alternative data sets than in the Schedule 13D data. This is not surprising. We expect the Schedule 13D data to have a higher signal-to-noise ratio since activist investors typically trade much larger quantities than insiders.

Column (3) reports results of regressing the insider trade indicator on ITI(13D), liquidity controls, and ITI(insider). Even when controlling for ITI(insider), ITI(13D) strongly detects insider trades with an estimated coefficient of 0.073 and a t-statistic of 19.47. A 10 percentage points (pp) increase in ITI(13D) thus leads to 0.7 pp increase in the likelihood of opportunistic insider trading. Column (8) shows a similar result in the short-selling regression: even when controlling for ITI(short), ITI(13D) is a strong detector of spikes in short selling. Therefore, ITI(13D) contains information that can help detect insider and short trades beyond what can be achieved with ITI measures estimated on these trades. This finding implies that incremental information can be gained about one type of informed trading from studying other types of informed trading.

ITI(patient) and ITI(impatient) also detect opportunistic insider trades and spikes in short selling (columns (4), (5), (9), and (10)) even with the full set of controls. In all cases, the coefficient on ITI(impatient) is more than five times as large as the coefficient on ITI(patient). Since these variables have about the same standard deviation (Table I), opportunistic insider trading and spikes in short selling seem better characterized as impatient. When we split insider trades into purchases and sales, we find that both ITI(patient) and ITI(impatient) detect insider purchases, but only ITI(impatient) detects insider sales, as shown in Table IA.XI.

It is not obvious how strong commonalities in informed trading are ex ante. Our data-driven approach allows us to provide a first pass at this question. Table IA.XVI reports estimates of regressions of each ITI measure on the other two ITI measures, control variables, and stock fixed effects. ITI measures

[23] Table I reports summary statistics for the informed trading measures. The average level of a given ITI measure is close to the average level of informed trading in the corresponding sample. Hence, one should be careful not to overinterpret the *level* of the measure for a given stock when it is averaged over a long period of time.

remain correlated even when we control for various liquidity measures and shut down cross-stock variation with stock fixed effects. For example, the coefficient on ITI(short) is 0.608, indicating a high partial correlation between ITI(13D) and ITI(short). Moreover, explanatory power increases by more than 50% when ITI(insider) and ITI(short) are included in the regression relative to the specification with controls only. At the same time, the highest R^2 that we achieve is only about 31%, which indicates a sizable fraction of unexplained variation in each specific ITI measure.[24]

In summary, ITI(13D) detects other types of informed trading, which does not appear to be explained by fluctuations in the trading environment. The results indicate a common pattern in the trading of three different groups of informed investors.

C. Models of Informed Trading

This section explains how our data-driven approach informs theories of informed trading and compares ITI to existing theory-based informed trading measures.

The groupings analysis in Table III shows that volume-related variables are the most important group of variables to understand fluctuations in ITI. To gain more intuition, we specifically consider turnover and absolute order imbalance. Table IA.XVII reports regressions of ITI on turnover and absolute order imbalance. There are two key takeaways. First, both variables are strongly positively associated with ITI. Second, these variables do not subsume each other.

In the multiperiod strategic trade model of Kyle (1985), volume in a given period is driven mostly by noise trading. Hence, a measure of ITI—informed volume over total volume—is negatively associated with volume. In the multiperiod sequential trade model of Easley and O'Hara (1992), volume is strongly positively associated with informed trading. Conditional on an information event occurring, informed traders always trade whereas uninformed traders sometimes do not trade. The data, however, indicate that informed trading is related to volume not only because informed traders drive volume but also because they react to uninformed volume. Table IA.XVIII shows that a 1% increase in non-13D volume is associated with increased informed trading on both extensive and intensive margins. For example, a 1% increase in non-13D volume is associated with an increase of 0.54% in 13D turnover. Hence, the endogeneity of informed volume relative to uninformed volume is an important point for models to consider.

To make the above settings more realistic, we consider simulations of informed trading models where we assume that one realization of a model occurs

[24] The univariate correlation between ITI(13D) and ITI(short) is 0.35. The correlations between ITI(13D) and ITI(insider) as well as ITI(insider) and ITI(short) are 0.16 and 0.15, respectively. The highest univariate correlation between ITI(13D) and any of the other liquidity variables is with turnover (0.26). Similarly, the two most highly correlated variables with ITI(insider) are ITI(13D) and ITI(short), ahead of turnover, which has a correlation of 0.10 with ITI(insider).

Table VIII
Statistics from Simulations of Informed Trading Models

We use simulated data to estimate regressions of informed trading intensity (ITI) on daily volume (V_t), absolute daily order imbalance ($|OI_t|$), and absolute daily return ($|r_t|$). The methodology and associated calibrations are described in Section II in the Internet Appendix. The label <0 (>0) indicates that the median estimated coefficient across calibrations is lower (greater) than zero and statistically significant at the 1% level. Avg R^2 is the average adjusted R^2 across simulations. T denotes the number of trading periods within a day. Kyle-tv refers to the strategic trade model in which we allow noise trading volatility to vary across days. Kyle-stochastic refers to the strategic trade model with stochastic noise trading volatility.

	$\text{ITI}_t = a + b_1 V_t + b_2 \lvert OI_t \rvert + e_t$			$\text{ITI}_t = c + d_1 V_t + d_2 \lvert OI_t \rvert + d_3 \lvert r_t \rvert + u_t$			
	b_1	b_2	Avg R^2	d_1	d_2	d_3	Avg R^2
Data	>0	>0	10.2%	>0	>0	<0	10.4%
PIN ($T = 40$)	>0	>0	76.5%	0	>0	>0	81.6%
PIN ($T = 100$)	>0	>0	87.8%	0	>0	>0	92.2%
PIN ($T = 400$)	>0	>0	96.6%	0	>0	>0	99.2%
APIN ($T = 40$)	0	>0	71.1%	<0	>0	>0	78.2%
APIN ($T = 100$)	<0	>0	83.6%	<0	>0	>0	90.2%
APIN ($T = 400$)	<0	>0	93.2%	<0	>0	>0	95.9%
Kyle ($T = 40$)	<0	>0	75.6%	–	–	–	–
Kyle ($T = 100$)	<0	>0	82.3%	–	–	–	–
Kyle ($T = 400$)	<0	>0	91.1%	–	–	–	–
Kyle-tv ($T = 40$)	<0	>0	57.0%	0	0	>0	74.2%
Kyle-tv ($T = 100$)	<0	>0	63.6%	0	0	>0	78.3%
Kyle-tv ($T = 400$)	<0	>0	67.8%	0	0	>0	83.0%
Kyle-stochastic	0	>0	32.5%	>0	0	>0	37.4%

over the course of a day. Model values can then be aggregated to the daily level to match our regressions. This procedure allows us to vary model parameters across days to better understand what drives informed trading. We present a detailed description of the simulations in Section II of the Internet Appendix. In a nutshell, we use simulated data to estimate regressions of ITI on daily volume, absolute daily order imbalance, and absolute daily return. Table VIII reports the results for various models of informed trading. For each model, we employ a range of different calibrations. The label <0 (>0) indicates that the median estimated coefficient across calibrations is lower (greater) than zero and statistically significant at the 1% level.

In both sequential trade and strategic trade models, daily absolute order imbalance is strongly related to ITI. Variation in the difference between the initial price (p_0) and the fundamental value ($\tilde{v}$) across days drives this relation. When $|\tilde{v} - p_0|$ is high, informed investors trade actively to benefit from their information, which leads to $|OI| \propto |\tilde{v} - p_0|$. Informed trading drives absolute imbalance since noise trading tends to average out in these models. This strong relation implies that regressing ITI on volume and absolute order imbalance results in large R^2; for example, in the 70% to 97% range for the PIN-type models that we consider.

Controlling for the absolute daily return strongly weakens the role of absolute order imbalance. The reason is that the daily return is close to $\tilde{v} - p_0$, as most private information is incorporated into the price by the end of the day. This conclusion holds in all extensions of the sequential and strategic trade models that we consider, such as stochastic noise trading volatility in the strategic trade model (Collin-Dufresne and Fos (2016)), and is inconsistent with our empirical results since ITI has a much stronger association with absolute order imbalance than with any volatility measure. Table IA.XVII shows that the absolute daily return is negatively related to ITI and barely affects the relation between absolute order imbalance and ITI. Intuitively, order imbalance is unlikely to be driven solely by informed trading since this would imply that informed traders are always on the aggressive side of a trade. In fact, Schedule 13D filers' order imbalance is negatively associated with the Lee and Ready (1991) order imbalance.

To summarize, the data suggest that a fruitful avenue is to develop models in which volume and absolute order imbalance are both positively related to realized informed trading but volatility is negatively related to it. This dimension of informed trading appears difficult for extant models to capture, as shown in Table VIII. Across all models and calibrations that we consider, none is able to jointly match these three relations.

In our main specification, we control for several measures that are likely associated with informed trading (e.g., spread, Kyle's lambda). We compare the performance of ITI to that of several existing measures of informed trading. We consider the conditional probability of informed trading obtained from the following models: the PIN model (PIN) of Easley et al. (1996), the adjusted PIN model (APIN) of Duarte and Young (2009), the generalized PIN model (GPIN) of Duarte, Hu, and Young (2020), the Odders-White and Ready (2008) model (OWR), and the Back, Crotty, and Li (2018) model (BCL). For simplicity, we refer to these measures as PIN measures below. We perform this analysis in this subsection rather than across all specifications because PIN measures are not available in the full sample. Specifically, the availability of PIN measures restricts our sample to NYSE-listed stocks from 1994 to 2012.[25] Each of these models is estimated for every stock-year. Then, for each stock-day, the conditional probability of informed trading is computed as the probability of an information event given the estimated structural parameters and specific variables for the stock-day (e.g., the number of buy trades and the number of sell trades in the case of PIN). Following Duarte, Hu, and Young (2020), we add an additional control variable that equals one for days with above-average number of trades and zero otherwise.

Table IX reports how the various measures detect patient Schedule 13D trades, impatient Schedule 13D trades, insider trades, and spikes in short selling. As before, we control for ITI(insider) in the insider sample and for ITI(short) in the short-sale sample. ITI(impatient) consistently detects

[25] These measures are obtained from Edwin Hu's website and described in Duarte, Hu, and Young (2020). We thank him for making the data available.

Table IX

ITI Measures and Other Measures of Informed Trading

This table compares ITI measures to the conditional probability of informed trading obtained from several models: the PIN model (PIN), the adjusted PIN model of Duarte and Young (2009) (APIN), the generalized PIN model (GPIN) of Duarte, Hu, and Young (2020), the Odders-White and Ready (2008) model (OWR), and the Back, Crotty, and Li (2018) model (BCL). ITI(patient) (ITI(impatient)) is the informed trading intensity estimated using the first 40 days (last 20 days) of the 60-day Schedule 13D filing window. ITI(insider) is trained on opportunistic insider trading data. ITI(short) is trained on short-selling data. These data sets are described in Table IA.II. In columns (1) and (2), indicators for days with Schedule 13D trading in the first 40 days (patient trade) and last 20 days (impatient trade) of the 60-day Schedule 13D filing window are regressed on the above measures, control variables, and stock fixed effects (filing fixed effects in the Schedule 13D sample). In columns (3) and (4), the dependent variables are indicators for days with opportunistic insider trades and spikes in short selling. Control variables are winsorized at 0.05% and 99.95% and include effective spread, price impact, lambda, depth, realized volatility, turnover, order imbalance, absolute order imbalance, return, and an indicator variable that takes the value of one for days with above-average number of trades (Duarte, Hu, and Young (2020)). Adj. R^2 ITI(x) is the adjusted R^2 from a regression that only includes the ITI measure trained on the specific data set and fixed effects. The sample consists of NYSE-listed stocks from 1994 to 2012. To be included, a stock must have a price greater than \$5 and a market capitalization greater than \$100 million at the end of the previous month. Standard errors are clustered by filing and associated t-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

Dep. Variable	13D Patient Trade (1)	13D Impatient Trade (2)	Insider Trade (3)	Δ Short (4)
ITI(patient)	0.104**	0.062	−0.015*	0.018
	(2.133)	(0.988)	(−1.767)	(1.240)
ITI(impatient)	0.330***	0.386***	0.081***	0.127***
	(5.732)	(5.945)	(8.318)	(6.778)
ITI(insider)			0.184***	
			(12.011)	
ITI(short)				0.320***
				(9.637)
PIN	0.015	0.044	−0.001	0.005
	(0.839)	(1.341)	(−0.266)	(0.729)
APIN	0.013	0.053**	0.010***	0.014***
	(0.936)	(2.458)	(3.974)	(3.294)
GPIN	−0.021*	0.005	−0.002	−0.001
	(−1.727)	(0.276)	(−0.626)	(−0.157)
OWR	0.001	0.068	0.002	0.014*
	(0.030)	(1.141)	(0.496)	(1.677)
BCL	−0.002	0.043**	0.009***	0.007
	(−0.113)	(2.018)	(3.823)	(1.638)
Controls and FE	Yes	Yes	Yes	Yes
Adj. R^2	0.0391	0.0575	0.0042	0.0109
Adj. R^2 ITI(x)	0.0173	0.0430	0.0023	0.0081
Obs.	6,250	2,979	199,109	71,243

informed trading across all specifications even when controlling for PIN measures. It is the only measure to achieve a significance level of 1% in all four samples. In contrast, ITI(patient) is generally insignificant.[26] Perhaps insider and short trades tend to be impatient.

Some of the PIN conditional probabilities also detect informed trades in Table IX. Among PIN measures, APIN performs best in that it detects impatient trades, insider trades, and spikes in short selling at a significance level of 5%. This is encouraging since it suggests that different approaches provide complementary information to detect informed trading. However, the results indicate substantial differences across PIN models in detecting informed trading. For completeness, we also report the regression results using one PIN measure at a time in Tables IA.XIX and IA.XX since PIN measures are correlated with each other. APIN stands out, followed by BCL and PIN.

To conclude, we note two key differences between ITI and PIN measures. First, by construction, PIN measures are reestimated for each stock-year whereas ITI is estimated on a single sample of informed trades and then extrapolated to the full sample with parameters that are constant across all stocks and years. Second, PIN's calculation is computationally challenging for actively traded securities (e.g., Griffin, Oberoi, and Oduro, 2021), whereas ITI is easy to calculate for the full cross section of stocks once model parameters are estimated on the desired set of informed trades.

IV. Application: Informed Trading and Asset Prices

ITI measures can be useful in many settings. In this section, we provide an asset pricing application by examining the relation between ITI and future stock returns. We first distinguish predictions that concern expected informed trading from predictions that concern realized informed trading.

Several theories imply conflicting predictions about how *expected* intensity of informed trading affects expected returns. Easley and O'Hara (2004) study how the total amount of information in the market affects asset prices. They show that if a higher fraction of information is private, the risks for uninformed investors increase, and thus expected returns *increase*. In contrast, Hughes, Liu, and Liu (2007) and Lambert, Leuz, and Verrecchia (2007) argue that the effect of asymmetric information on returns is diversifiable, and thus is *not priced*, in a competitive market. Empirically, Duarte and Young (2009) show that PIN reflects not only asymmetric information but also illiquidity, and only the component of PIN related to illiquidity is priced. Finally, if informed trading reduces information uncertainty, more efficient prices should lead to higher stock valuations and *lower* future returns. For example, Roll, Schwartz, and Subrahmanyam (2010) argue that option trading makes the underlying stock more informationally efficient and show that higher option volume predicts

[26] One may wonder why ITI(impatient) appears to be a stronger detector of patient trades than ITI(patient). This is due to the inclusion of filing fixed effects in the regression. Filing/stock fixed effects do not materially affect other results in Table IX and the paper.

lower future stock returns. In these theories, prices are set conditional on expected ITI, and expected returns are not affected if investors are risk-neutral. ITI can help test these theories' conflicting return predictions assuming that ITI averaged over a period of time proxies for expected informed trading.

Deviations of *realized* ITI from its expected level can also affect expected returns. Intuitively, stock prices increase after informed buys and decrease after informed sells. If buy and sell trades are equally likely to be informed, then unsigned trading intensity (such as ITI) should not predict average returns. However, several studies suggest that stock purchases are more informed than stock sales (Kraus and Stoll (1972), Chan and Lakonishok (1993), and Campbell, Ramadorai, and Schwartz (2009)). Also, the long-only focus of most institutional investors and short-selling costs limit these investors' ability to exploit negative information. If informed purchases outnumber informed sales, high realized ITI should lead to higher average future returns.[27]

Motivated by these hypotheses, we estimate a (stock-by-week) panel regression of next-month stock returns on ITI measures averaged over a prior week, standard return predictors, and week fixed effects to account for stock return commonality. Stock returns are computed from CRSP daily returns and adjusted for delistings as in Shumway (1997). We skip a day between predictors and returns to avoid confounding effects as today's closing price is an input in tomorrow's return. Other predictors include idiosyncratic volatility (computed from abnormal daily returns from the Fama-French-Carhart four-factor model over the prior month), momentum (stock returns from six months to one month prior to the current date), monthly reversal (previous month return), log market capitalization, CAPM beta, Amihud (2002) illiquidity measure, Kyle's lambda, and the effective bid-ask spread. Predictors are winsorized at the 0.5% and 99.5% levels to avoid outliers. We rely on a stock-by-week panel, which results in overlapping monthly return observations, to avoid turn-of-the-month effects documented by Etula et al. (2020). To account for overlapping returns and cross-stock dependencies, standard errors are clustered by stock and week.[28] Table IA.XXI reports summary statistics for the panel used in this section.

We find several results. Higher realized informed trading over the prior week, as measured by higher ITI(13D), is associated with higher realized returns over the next month. Column (1) in Table X reports that ITI(13D) has a t-statistic of 6.82 for predicting monthly returns. A one-standard-deviation increase in ITI(13D) raises the next-month return by 14 bps on average.

[27] To see why, consider an extreme case with informed buying but no informed selling. A market maker sets the price knowing this difference in expected informed intensities of buys and sells. If no informed buying occurs, realized informed intensity is below the expected level and the price will subsequently decrease. Similarly, if most buys turn out to be informed, the price will increase.

[28] The main results continue to hold if a Fama-MacBeth regression is estimated instead of a panel regression, if nonoverlapping monthly returns are used, or if alternative specifications are used. For example, if monthly returns, order imbalance, or turnover are included as controls. We find similar results for weekly returns and report them in the Internet Appendix.

Table X

Informed Trading Intensity and Future Returns

Monthly returns are regressed on ITI measures (weekly averages), other predictors, and weekly fixed effects. ITI(13D) is trained on Schedule 13D data. ITI(insider) is trained on opportunistic insider trading data. ITI(insider, buy) and ITI(insider, sell) are trained on insider buy and sell trades, respectively. ITI(short) is trained on short-selling data. ITI(patient) (ITI(impatient)) is trained on the first 40 days (last 20 days) of the 60-day Schedule 13D trading window. Column (5) in Panel A controls for other predictors including log market capitalization, CAPM beta, last month return, two- to six-month return, idiosyncratic volatility, Amihud illiquidity, the effective bid-ask spread, and Kyle's lambda. Column (6) in Panel A adds PIN. The PIN sample comes from Brown and Hillegeist (2007) and covers 1993 to 2010. The main sample includes common stocks from January 1993 to July 2019. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million. *t*-Statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. Standard errors are clustered by stock and month.

Panel A: ITIs and Stock Returns

	Monthly Return					
Dep. Variable:	(1)	(2)	(3)	(4)	(5)	(6)
ITI(13D)	0.0126***	0.0117***		0.0053**	0.0067***	0.0096***
	(6.82)	(4.67)		(2.22)	(3.51)	(4.06)
ITI(insider)		0.0119***		0.0107***	0.0083***	0.0093***
		(4.82)		(4.42)	(5.28)	(4.82)
ITI(short)		−0.0084		−0.0201***	−0.0054	0.0038
		(−1.29)		(−2.79)	(−1.05)	(0.60)
ITI(patient)			0.0006	−0.003	−0.0037*	−0.0049*
			(0.28)	(−1.43)	(−1.81)	(−1.82)
ITI(impatient)			0.0158***	0.0163***	0.0157***	0.0178***
			(6.51)	(5.83)	(4.99)	(4.67)
PIN						0.0019
						(0.35)
Controls	No	No	No	No	Yes	Yes
Fixed effects	Week	Week	Week	Week	Week	Week
R^2	0.0001	0.0002	0.0002	0.0002	0.0014	0.0016
Obs.	3,484,923	3,484,923	3,484,923	3,484,923	3,484,923	2,338,675

Panel B: ITI Estimated on Insider Buys or Sells

	Monthly Return			
Dep. Variable:	(1)	(2)	(3)	(4)
ITI(insider, buy)	0.0120***		0.0112***	0.0070***
	(2.91)		(3.45)	(3.46)
ITI(insider, sell)		−0.0065	−0.0018	0.0029
		(−1.62)	(−0.60)	(1.35)
Controls	No	No	No	Yes
R^2	0.0001	0.0000	0.0001	0.0012
Obs.	3,484,923	3,484,923	3,484,923	3,484,923

ITI's ability to predict returns is not affected by controlling for standard return predictors.

After studying ITI(13D), we study other ITI measures. As shown in Section III.B, ITI measures have a common component. To the extent that this common component identifies general trading patterns by informed investors, we expect other ITI measures to also positively predict returns in individual regressions. We confirm with portfolio sorts below, all ITI measures positively and significantly predict stock returns except for ITI(short), which is not a significant predictor. Moreover, column (2) in Table X shows that when we include ITI(13D), ITI(insider), and ITI(short) in the same regression, ITI(13D) and ITI(insider) are strong predictors of future returns with similar *t*-statistics, and ITI(short) is negatively but insignificantly related to future returns. Hence, ITI measures are ordered in line with the directional information that they are estimated on. Schedule 13D filers trade on positive private information, corporate insiders' buys are on average more informed than their sells (e.g., Jeng, Metrick, and Zeckhauser (2003)), while short sellers trade on negative information.

The result that ITI is positively associated with future returns is consistent the buy-sell asymmetry as well as some theories of informed trading risk, such as Easley and O'Hara (2004). However, the result that ITI(short) is a weaker (and sometimes even negative) predictor of returns compared to other ITI measures favors the buy-sell asymmetry explanation. ITI(short) reflects general informed trading that positively predicts returns but also captures unique features of short selling that are likely to negatively predict returns. Taken together, these two effects tend to cancel out. In contrast, expected informed trading risk theories do not distinguish between informed buying and selling and focus on the combined amount of informed trading.

We perform another test that further explores the asymmetry between the information content of buy and sell trades. Specifically, we train ITI(insider) on insider purchases and sales separately and then test the prediction of the buy-sell asymmetry that ITI(insider buy) should predict returns more positively than ITI(insider sell). Panel B of Table X confirms this prediction. In univariate regressions, ITI(insider buy) has a *t*-statistic of 2.9 versus −1.6 for ITI(insider sell). In a joint regression, ITI(insider buy) continues to strongly predict returns while ITI(insider sell) is not significant. The difference between the two is economically and statistically significant (*p*-value of 2% for the zero coefficient difference hypothesis). This test further supports the buy-sell asymmetry explanation.

We study ITI(patient) and ITI(impatient). Intuitively, informed investors trade more aggressively and impatiently if they are more certain about their private signal. Consistent with this hypothesis, column (3) shows that when the regression includes ITI(patient) and ITI(impatient), ITI(impatient) is a strong predictor of future returns, whereas ITI(patient) is not significantly related to future returns. This result is consistent with the results in Section III.A, where ITI(impatient) performs better than ITI(patient) in various validation tests.

In column (4), we include all five ITI measures in the regression. ITI(13D), ITI(insider), and ITI(impatient) are strong positive predictors of future returns, ITI(short) is a negative predictor of future returns, and ITI(patient) is not significant. The results continue to hold once we add control variables to this regression in column (5), except that the coefficient on ITI(short) becomes statistically insignificant.

Controlling for PIN does not affect the results. We use annual PIN estimates over the period 1993 to 2010 from Brown and Hillegeist (2007). In a univariate regression, PIN positively predicts returns with a t-statistic of 2.13. However, column (6) of Table X shows that the coefficient for PIN becomes insignificant once we control for other variables. These results are consistent with Duarte and Young (2009), who show in their table 10 that PIN is a positive but not significant return predictor. In contrast, ITI measures continue to predict returns in the PIN sample.

Portfolio sorts complement panel regressions. Table XI reports returns (Panel A) and alphas from the FF4 model that includes the momentum factor (Panel B) for equally weighted decile portfolios sorted on ITI measures averaged over the prior week. Portfolio sorts confirm the results from panel regressions. When we sort on ITI(13D), the FF4 alpha for the difference between the top and bottom decile portfolios is 0.52% per month, or 6.42% annualized, with a t-statistic of 6.2. Alpha increases gradually from −3 bps for the bottom portfolio to 49 bps for the high portfolio. Consistent with the regression results, ITI(impatient) predicts returns more strongly than ITI(patient), with FF4 monthly alphas for the ten-minus-one portfolio of 0.54% and 0.34%, respectively. ITI (insider) has an FF4 alpha of 0.22%, or 2.67% annualized, lower than for ITI(13D), while ITI(short) yields an insignificant 0.05% alpha. Finally, the results are qualitatively similar for value-weighted portfolio sorts and alternative subsamples.[29]

The horizon of return predictability can further help us distinguish between the buy-sell asymmetry and informed trading risk explanations. To the extent that stock-specific information risk is persistent, the informed trading risk explanation implies that ITIs should predict not only short-term returns but also long-term returns. In contrast, according to buy-sell asymmetry, stock prices underreact to realized informed trading, which suggests short-term return predictability. We repeat portfolio sort analysis but instead of the next-month return, we use returns from the second-next month. Table IA.XXIV shows that

[29] First, ITI(13D) generates a value-weighted monthly alpha of 0.25% with a t-statistic of 2.2. Second, Table IA.XXIII shows that ITI(13D) robustly predicts returns across various subsamples. For each splitting variable, we split the full sample into two equal parts based on the median value of the splitting variable. We then sort stocks into decile portfolios based on ITI(13D) within each part and compute monthly alphas. We consider the following splitting variables: market capitalization, stock turnover, idiosyncratic volatility, effective spread, Kyle's lambda, and PIN. ITI(13D) remains a positive and significant predictor of returns in all subsamples, with monthly alphas for the difference between the top and bottom decile portfolios ranging from 0.31% to 0.85%. As expected, the predictability is stronger for stocks for which informed trading is likely to be more important, such as smaller stocks, stocks with higher trading costs, and stocks with higher volatility.

Table XI

Portfolio Sorts Based on ITI Measures

For each ITI measure, we sort stocks into decile portfolios based on their average ITI measures during the prior week. We compute equally-weighted average return during next month for each decile and the top-minus-bottom difference. We report raw returns and alphas from three factor Fama-French models with a momentum factor. We consider separately five ITI measures. ITI(13D) is trained on Schedule 13D data. ITI(insider) is trained on opportunistic insider trading data. ITI(short) is trained on short-selling data. These data sets are described in Table IA.II. ITI(patient) (ITI(impatient)) is trained on the first 40 days (last 20 days) of the 60-day Schedule 13D trading window. "0.0071" corresponds to 0.71% per month. The sample includes U.S. common stocks from January 1993 to July 2019. To be included, a stock must have a price greater than $5 and a market capitalization greater than $100 million at the end of the previous month. *t*-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. Standard errors are computed using Newey-West adjustment with eight lags.

Panel A. Average monthly returns

	Low	2	3	4	5	6	7	8	9	High	H-L
ITI(13D)	0.0071**	0.0079***	0.0083***	0.0088***	0.0089***	0.0087***	0.0099***	0.0099***	0.0106***	0.0115***	0.0044***
	(2.4)	(2.8)	(2.9)	(3.1)	(3.2)	(3.1)	(3.5)	(3.5)	(3.7)	(4.2)	(4.9)
ITI(insider)	0.0072**	0.0078***	0.0085***	0.0090***	0.0089***	0.0096***	0.0094***	0.0103***	0.0103***	0.0109***	0.0037***
	(2.5)	(2.8)	(3.0)	(3.2)	(3.2)	(3.4)	(3.4)	(3.6)	(3.6)	(3.6)	(3.3)
ITI(short)	0.0085***	0.0086***	0.0086***	0.0089***	0.0093***	0.0088***	0.0095***	0.0098***	0.0099***	0.0098***	0.0014
	(3.2)	(3.2)	(3.1)	(3.2)	(3.3)	(3.1)	(3.3)	(3.4)	(3.3)	(3.1)	(1.0)
ITI(patient)	0.0079***	0.0082***	0.0085***	0.0088***	0.0090***	0.0091***	0.0097***	0.0095***	0.0101***	0.0109***	0.0030***
	(2.7)	(2.9)	(3.0)	(3.1)	(3.2)	(3.2)	(3.5)	(3.4)	(3.5)	(3.9)	(3.9)
ITI(impatient)	0.0064**	0.0077***	0.0080***	0.0088***	0.0093***	0.0094***	0.0099***	0.0103***	0.0104***	0.0115***	0.0051***
	(2.2)	(2.7)	(2.8)	(3.1)	(3.2)	(3.4)	(3.5)	(3.6)	(3.7)	(4.2)	(5.1)

(*Continued*)

Table XI—*Continued*

Panel B. Four-factor Fama-French monthly alphas

	Low	2	3	4	5	6	7	8	9	High	H-L
ITI(13D)	−0.0003	0.0005	0.0007	0.0014***	0.0017***	0.0013***	0.0027***	0.0027***	0.0035***	0.0049***	0.0052***
	(−0.5)	(1.0)	(1.4)	(2.9)	(3.6)	(2.6)	(5.0)	(4.6)	(5.2)	(7.0)	(6.2)
ITI(insider)	0.0003	0.0010*	0.0016***	0.0020***	0.0018***	0.0024***	0.0021***	0.0028***	0.0025***	0.0026***	0.0022***
	(0.6)	(1.6)	(2.8)	(3.5)	(3.2)	(4.4)	(4.1)	(5.9)	(5.1)	(4.4)	(2.9)
ITI(short)	0.0017***	0.0016**	0.0015**	0.0017***	0.0020***	0.0014***	0.0021***	0.0024***	0.0025***	0.0022***	0.0005
	(2.6)	(2.5)	(2.5)	(3.1)	(3.7)	(2.8)	(3.9)	(4.5)	(4.4)	(3.2)	(0.5)
ITI(patient)	0.0008	0.0007	0.0011**	0.0014***	0.0016***	0.0018***	0.0024***	0.0023***	0.0027***	0.0041***	0.0034***
	(1.1)	(1.5)	(2.3)	(2.9)	(3.1)	(3.7)	(4.4)	(4.3)	(4.9)	(6.2)	(4.2)
ITI(impatient)	−0.0006	0.0004	0.0006	0.0015***	0.0018***	0.0021***	0.0025***	0.0030***	0.0032***	0.0048***	0.0054***
	(−1.0)	(0.8)	(1.2)	(3.2)	(3.6)	(4.6)	(4.9)	(5.3)	(4.8)	(6.4)	(5.7)

none of ITI measures predicts returns beyond the next month. This lack of long-term predictability is not consistent with the informed trading risk channel. This conclusion must be interpreted with caution, however, because it assumes that ITI, which reflects current realized informed trading, correlates with expected informed trading in the future.

Overall, ITI measures are positively associated with next-month returns, and this predictability is most consistent with a buy-sell asymmetry explanation.

V. Conclusion

In this paper, we develop a new measure of informed trading by directly learning from informed trading data. We use an ML algorithm to identify days with informed trading. In this standard classification problem, a daily indicator for informed trading is predicted by a set of same-day variables related to volume, volatility, and liquidity. After the model is estimated on the training data of observed informed trades, we extrapolate it to the entire stock-day universe, where informed trading is not directly observed. This procedure produces a new measure of informed trading—the "ITI."

We show that ITI significantly predicts informed trading out-of-sample and is a significant predictor of various information events. In particular, ITI increases before earnings announcements, M&A announcements, and unscheduled news releases. Moreover, returns on days with high ITI exhibit less reversal than returns on other days, in line with the intuition that the price impact of informed trades is permanent. All of these results validate our use of ITI as a measure of ITI.

We show that our data-driven approach can shed light on the economics of informed trading. First, we show a strong distinction between impatient trading and patient trading, consistent with theory. Second, we show that incremental information can be gained about one type of informed trading from studying other types of informed trading. Third, our methodology highlights specific features of informed trading that existing models struggle to capture. Indeed, ITI is not subsumed by existing theory-based measures. These stylized facts provide a fruitful avenue for future research.

ITI can be applied in many settings. We provide an application to the asset pricing literature and ask whether informed trading is priced in the cross section of stock returns, as prior work documents conflicting results. We show that an increase in ITI is associated with higher future monthly returns in the cross section, but this predictability is most consistent with an asymmetry in the informational content of purchases and sales.

The main implication of this paper is that a data-driven ML approach combined with data on informed trading can generate an effective measure of informed trading and improve our understanding of the economics of informed trading.

Initial submission: January 30, 2022; Accepted: June 28, 2023
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

REFERENCES

Ahern, Kenneth R., 2020, Do proxies for informed trading measure informed trading? Evidence from illegal insider trades, *Review of Asset Pricing Studies* 10, 397–440.

Akbas, Ferhat, Chao Jiang, and Paul D. Koch, 2020, Insider investment horizon, *Journal of Finance* 75, 1579–1627.

Akey, Pat, Vincent Gregoire, and Charles Martineau, 2022, Price revelation from insider trading: Evidence from hacked earnings news, *Journal of Financial Economics* 143, 1162–1184.

Alldredge, Dallin M., and David C. Cicero, 2015, Attentive insider trading, *Journal of Financial Economics* 115, 84–101.

Amihud, Yakov, 2002, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets* 5, 31–56.

Augustin, Patrick, Menachem Brenner, and Marti G. Subrahmanyam, 2019, Informed options trading prior to takeover announcements: Insider trading? *Management Science* 65, 5697–5720.

Augustin, Patrick, and Marti G. Subrahmanyam, 2020, Informed options trading before corporate events, *Annual Review of Financial Economics* 12, 327–355.

Back, Kerry, Pierre Collin-Dufresne, Vyacheslav Fos, Tao Li, and Alexander Ljungqvist, 2018, Activism, strategic trading, and liquidity, *Econometrica* 86, 1431–1643.

Back, Kerry, Kevin Crotty, and Tao Li, 2018, Identifying information asymmetry in securities markets, *Review of Financial Studies* 31, 2277–2325.

Biggerstaff, Lee, David Cicero, and M. Babajide Wintoki, 2020, Insider trading patterns, *Journal of Corporate Finance* 64, 101654.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang, 2008, Which shorts are informed? *Journal of Finance* 63, 491–527.

Bolandnazar, Mohammadreza, Robert J. Jackson, Jr., Wei Jiang, and Joshua Mitts, 2020, Trading against the random expiration of private information: A natural experiment, *Journal of Finance* 75, 5–44.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson, 2019, Information, trading, and volatility: Evidence from firm-specific news, *Review of Financial Studies* 32, 992–1033.

Brav, Alon, Wei Jiang, Frank Partnoy, and Randall Thomas, 2008, Hedge fund activism, corporate governance, and firm performance, *Journal of Finance* 63, 1729–1775.

Breiman, Leo, 2001, Random forests, *Machine Learning* 45, 5–32.

Brennan, Michael J., Sahn-Wook Huh, and Avanidhar Subrahmanyam, 2018, High-frequency measures of informed trading and corporate announcements, *Review of Financial Studies* 31, 2326–2376.

Brown, Stephen, and Stephen A. Hillegeist, 2007, How disclosure quality affects the level of information asymmetry, *Review of Accounting Studies* 12, 443–477.

Caldentey, Rene, and Ennio Stacchetti, 2010, Insider trading with a random deadline, *Econometrica* 78, 245–283.

Campbell, John Y., Sanford J. Grossman, and Jiang Wang, 1993, Trading volume and serial correlation in stock returns, *Quarterly Journal of Economics* 108, 905–939.

Campbell, John Y., Tarun Ramadorai, and Allie Schwartz, 2009, Caught on tape: Institutional trading, stock returns, and earnings announcements, *Journal of Financial Economics* 92, 66–91.

Chan, Louis K. C., and Josef Lakonishok, 1993, Institutional trades and intraday stock price behavior, *Journal of Financial Economics* 33, 173–199.

Chen, Tianqi, and Carlos Guestrin, 2016, Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cohen, Lauren, Christopher Malloy, and Lukasz Pomorski, 2012, Decoding inside information, *Journal of Finance* 67, 1009–1043.

Collin-Dufresne, Pierre, and Vyacheslav Fos, 2015, Do prices reveal the presence of informed trading? *Journal of Finance* 70, 1555–1582.

Collin-Dufresne, Pierre, and Vyacheslav Fos, 2016, Insider trading, stochastic liquidity and equilibrium prices, *Econometrica* 84, 1441–1475.
Collin-Dufresne, Pierre, Vyacheslav Fos, and Dmitry Muravyev, 2021, Informed trading in the stock market and option-price discovery, *Journal of Financial and Quantitative Analysis* 56, 1945–1984.
Cookson, Anthony, Vyacheslav Fos, and Marina Niessner, 2021, Does disagreement facilitate informed trading? Evidence from activist investors, Working paper, Boston College and University of Colorado Boulder.
Craven, Mark, and Jude Shavlik, 1995, Extracting tree-structured representations of trained networks, *Advances in Neural Information Processing Systems* 8, 24–30.
Cziraki, Peter, and Jasmin Gider, 2021, The dollar profits to insider trading, *Review of Finance* 25, 1547–1580.
Duarte, Jefferson, Edwin Hu, and Lance Young, 2020, A comparison of some structural models of private information arrival, *Journal of Financial Economics* 135, 795–815.
Duarte, Jefferson, and Lance Young, 2009, Why is PIN priced? *Journal of Financial Economics* 91, 119–138.
Easley, David, Marcos Lopez de Prado, Maureen O'Hara, and Zhibai Zhang, 2021, Microstructure in the machine age, *Review of Financial Studies* 34, 3316–3363.
Easley, David, Soeren Hvidkjaer, and Maureen O'Hara, 2002, Is information risk a determinant of asset returns? *Journal of Finance* 57, 2185–2221.
Easley, David, Nicholas M. Kiefer, Maureen O'Hara, and Joseph B. Paperman, 1996, Liquidity, information, and infrequently traded stocks, *Journal of Finance* 51, 1405–1436.
Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial Economics* 19, 69–90.
Easley, David, and Maureen O'Hara, 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 577–605.
Easley, David, and Maureen O'Hara, 2004, Information and the cost of capital, *Journal of Finance* 59, 1553–1583.
Etula, Erkko, Kalle Rinne, Matti Suominen, and Lauri Vaittinen, 2020, Dash for cash: Monthly market impact of institutional liquidity needs, *Review of Financial Studies* 33, 75–111.
Fama, Eugene F., 1970, Efficient capital markets: A review of theory and empirical work, *Journal of Finance* 25, 383–417.
Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.
Gantchev, Nickolay, and Chotibhak Jotikasthira, 2018, Institutional trading and hedge fund activism, *Management Science* 64, 2930–2950.
Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.
Goldstein, Itay, Chester S. Spatt, and Mao Ye, 2021, Big data in finance, *Review of Financial Studies* 34, 3213–3225.
Griffin, Jim, Jaideep Oberoi, and Samuel D. Oduro, 2021, Estimating the probability of informed trading: A Bayesian approach, *Journal of Banking and Finance* 125, 106045.
Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.
Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.
Hasbrouck, Joel, 1988, Trades, quotes, inventories, and information, *Journal of Financial Economics* 22, 229–252.
Hasbrouck, Joel, 1991, Measuring the information content of stock trades, *Journal of Finance* 46, 179–207.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).
Holderness, Clifford G., and Dennis P. Sheehan, 1985, Raiders or saviors? The evidence on six controversial investors, *Journal of Financial Economics* 14, 555–579.

Hughes, John S., Jing Liu, and Jun Liu, 2007, Information asymmetry, diversification, and cost of capital, *Accounting Review* 82, 705–729.
Jeng, Leslie A., Andrew Metrick, and Richard Zeckhauser, 2003, Estimating the returns to insider trading: A performance-evaluation perspective, *Review of Economics and Statistics* 85, 453–471.
Kacperczyk, Marcin, and Emiliano Pagnotta, 2019, Chasing private information, *Review of Financial Studies* 32, 4997–5047.
Kadan, Ohad, and Asaf Manela, 2020, Liquidity and the strategic value of information, Working paper, Washington University in St. Louis.
Kaniel, Ron, and Hong Liu, 2006, So what orders do informed traders use? *Journal of Business* 79, 1867–1913.
Karolyi, G. Andrew, and Stijn Van Nieuwerburgh, 2020, New methods for the cross-section of returns, *Review of Financial Studies* 33, 1879–1890.
Kelly, Bryan, and Alexander Ljungqvist, 2012, Testing asymmetric-information asset pricing models, *Review of Financial Studies* 25, 1366–1413.
Kim, Oliver, and Robert E. Verrecchia, 1994, Market liquidity and volume around earnings announcements, *Journal of Accounting and Economics* 17, 41–67.
Klein, April, and Emanuel Zur, 2009, Entrepreneurial shareholder activism: Hedge funds and other private investors, *Journal of Finance* 64, 187–229.
Kraus, Alan, and Hans R. Stoll, 1972, Price impacts of block trading on the New York Stock Exchange, *Journal of Finance* 27, 569–588.
Kwan, Amy, Richard Philip, and Andriy Shkilko, 2021, The conduits of price discovery: A machine learning approach, Working paper, University of Sydney and Wilfrid Laurier University.
Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.
Lambert, Richard, Christian Leuz, and Robert E. Verrecchia, 2007, Accounting information, disclosure, and the cost of capital, *Journal of Accounting Research* 45, 385–420.
Lee, Charles, and Mark J. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733–746.
Lee, Charles M. C., Belinda Mucklow, and Mark J. Ready, 1993, Spread, depths, and the impact of earnings information: An intraday analysis, *Review of Financial Studies* 6, 345–374.
Llorente, Guillermo, Roni Michaely, Gideon Saar, and Jiang Wang, 2002, Dynamic volume-return relation of individual stocks, *Review of Financial Studies* 15, 1005–1047.
Lundberg, Scott M., and Su-In Lee, 2017, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems* 30.
Nagel, Stefan, 2012, Evaporating liquidity, *Review of Financial Studies* 25, 2005–2039.
Nagel, Stefan, 2021, *Machine Learning in Asset Pricing* (Princeton, NJ: Princeton University Press).
Odders-White, Elizabeth R., and Mark J. Ready, 2008, The probability and magnitude of information events, *Journal of Financial Economics* 87, 227–248.
O'Hara, Maureen, 2003, Presidential address: Liquidity and price discovery, *Journal of Finance* 58, 1335–1354.
Reed, Adam V., 2013, Short selling, *Annual Review of Financial Economics* 5, 245–258.
Richardson, Scott, Pedro A. C. Saffi, and Kari Sigurdsson, 2017, Deleveraging risk, *Journal of Financial and Quantitative Analysis* 52, 2491–2522.
Roll, Richard, Eduardo Schwartz, and Avanidhar Subrahmanyam, 2010, O/S: The relative trading activity in options and stock, *Journal of Financial Economics* 96, 1–17.
Senchack, Andrew J., and Laura T. Starks, 1993, Short-sale restrictions and market reaction to short-interest announcements, *Journal of Financial and Quantitative Analysis* 28, 177–194.
Shumway, Tyler, 1997, The delisting bias in CRSP data, *Journal of Finance* 52, 327–340.
Tibshirani, Robert, 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 267–288.
Yang, Yung Chiang, Bohui Zhang, and Chu Zhang, 2020, Is information risk priced? Evidence from abnormal idiosyncratic volatility, *Journal of Financial Economics* 135, 528–554.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.