The Market Effects of Algorithms[*]

Lindsey Raymond
MIT

July 14, 2024

Please see here for latest version

Abstract

While there is excitement about the potential of algorithms to optimize individual decision-making, changes in individual behavior will, almost inevitably, impact markets. Yet little is known about these effects. In this paper, I study how the availability of algorithmic prediction changes entry, allocation, and prices in the US residential real estate market, a key driver of household wealth. I identify a *market-level* natural experiment that generates variation in the cost of using algorithms to value houses: digitization, the transition from physical to digital housing records. I show that digitization leads to entry by investors using algorithms, but does not push out investors using human judgment. Instead, human investors shift towards houses that are difficult to predict algorithmically. Algorithmic investors predominantly purchase minority-owned homes, a segment of the market where humans may be biased. Digitization increases the average sale price of minority-owned homes by 5% or $5,000 and nearly eliminates racial disparities in home prices. Algorithmic investors, via competition, affect the prices paid by humans for minority homes, which drives most of the reduction in racial disparities. This decrease in racial inequality underscores the potential of algorithms to mitigate human biases at the market level.

[*] Correspondence to lraymond@mit.edu. I am especially grateful to Sendhil Mullainathan, Danielle Li, Erik Brynjolfsson and Scott Stern for their invaluable feedback and advice. I thank Mert Demirer, Brandon Enriquez, Robert Gibbons, Anh Nguyen, Manish Raghavan, Ashesh Rambachan, David Reshef, Daniel Rock, Frank Schilbach, Eric So, Evan Soltas, Maggie Yellen and various seminar participants for helpful comments and suggestions. I thank Max Feng for providing excellent research assistance and the Stanford Digital Economy Lab for funding. The content is solely the author's responsibility and does not necessarily represent the official views of Stanford University or MIT.

Many consequential decisions depend on predictions — hiring depends on predictions of who will be most productive; extending credit depends on the prediction of default; and investing decisions crucially rely on predictions of returns. The advent of machine learning and digital data has created a great deal of interest in the use of technology in prediction problems. Prediction technologies, or algorithms, could potentially change and improve decision making. In fact, a rich literature has already explored a variety of questions, including whether algorithms do better than humans or when algorithms might inherit biases of people making such decisions. Although much progress has been made, one area remains largely unexplored: *market* level impacts.

In addition to enhancing individual decision making, algorithms could have broader market-level and equilibrium consequences. Algorithms could lead to new entrants, change the nature of competition, and alter market-level outcomes such as prices. These broader market dynamics mean that even if algorithms help people make better decisions, individuals could still be worse off. In contrast, even if algorithms do not uniformly improve individual decisions, sectors of the market where human error is most pronounced could benefit. Studies designed to capture the impact of algorithms on decision quality cannot, by design, account for effects beyond the individual (or firm) level.

I empirically examine the market effects of algorithms in the US residential real estate market. In particular, I focus on investors, who buy houses to rent out or resell. I study the housing market due to the importance of the setting, the centrality of prediction in decision making, and the presence of an elegant natural experiment. First, housing is the largest contributor to the wealth of the median household and the largest asset market and therefore is a substantively interesting application in its own right (Derenoncourt et al., 2022; Malone, 2023).[1] Second, investing in real estate hinges on accurate predictions: investors aim to forecast potential rental income, property appreciation, annual maintenance costs, and overall future value. Yet, prediction is a difficult cognitive task for humans. Various human behavioral biases, such as the influence of weather, sentiment, anchoring, and loss aversion, among other heuristics, have been shown to impact housing prices (Busse et al., 2012; Kermani and Wong, 2021a; Salzman and Zwinkels, 2017).[2] Human investors devote considerable energy to counteracting the influence of these behavioral biases.

To address the central identification challenge—that algorithms are not randomly assigned to markets—I exploit an institutional feature of real estate markets and a simple yet fundamental

[1] A single-family home is a structure designed for a single one household, usually on its own plot of land, and while often associated with homeownership, is the largest single segment of the rental market (Neal et al., 2020; Freddie Mac Economic & Housing Research, 2018).

[2] Throughout the paper, I use "algorithms", "machine-generated predictions" and "ML algorithms" interchangeably. I also use "algorithmic investor", "investor using algorithms" and "investor using algorithmic valuation" to refer to investors using algorithms and "human investors" to refer to those who rely on human expertise.

insight: machine learning algorithms require machine-readable data. Specifically, algorithms need detailed property information, such as the number of bedrooms and bathrooms, yard size, age of the house, historical sale prices, home improvements records, and current market data. In the US, county governments are responsible for collecting this public information for routine administrative tasks like planning, legal processes, and taxation. This information was traditionally stored as paper documents and microfilm records in county offices. However, as part of a broader move towards open and transparent governance, counties began to digitize these archives into electronic database systems. This transition from physical to digital records—a process known as digitization—generates variation in the cost of using algorithms to value property. I analyze the county-level transition to digitized housing data across Georgia, North Carolina, South Carolina and Tennessee, over the period from 2009 to 2021. This natural experiment allows me to contribute some of the first evidence on the market-level effects of algorithms.[3]

First, I investigate whether digitization actually prompts entry by investors making use of predictive algorithms—called algorithmic investors.[4] After digitization, there is a six-fold increase in the number of houses bought by algorithmic investors. This surge is sharp and evident in the raw data and in the event study analysis. On average, algorithmic investors account for roughly 10% of all investor activity after digitization.

Digitization also changes the structure of the investor market. Before digitization, the market was dominated by small "mom-and-pop" entrepreneurs. The average human investor would typically purchase around 1.5 houses annually, operating within two different zip codes. Many of these investors worked as contractors, plumbers, or real estate agents, often buying in their local area. In general, this market was believed to remain localized and fragmented due to the considerable advantage of the mom-and-pop entrepreneurs in local and qualitative information and relevant expertise (Fields, 2018). Algorithmic investors, who depend on automated valuations instead of human expertise, are capable of buying hundreds of homes annually in hundreds of zip codes. Consequently, digitization has resulted in a 23% increase in the size of the average firm and has doubled the number of geographic areas in which the average firm invests.

These initial findings could be misleading if the timing of county digitization is correlated with the activities of algorithmic investors or overall housing market activity. Specifically, there are two principal potential confounders to consider. The first concern is that algorithmic investment

[3] Prior work in this field often concentrates on issues such as the implications of algorithmic pricing for competitive dynamics or analyzes the broad impact of algorithmic trading in financial markets Calvino and Fontanelli (2023); Calder-Wang and Kim (2023); Clark et al. (2023); Brown and MacKay (2023); Aggarwal and Thomas (2014). I am unaware of other work on the market effects of ML-powered algorithms.

[4] I define algorithmic investor as an investor using algorithms to value houses. Investors not using algorithms are "human investors."

firms themselves might influence when a county chooses to digitize its records. For example, if an investor wanted to buy houses within a certain county, they might lobby for or financially support improvements in record management. To address this concern, I check for any suggestion of preexisting investor interest and the timing of digitization. If the timing of digitization were indeed driven by investors, I would expect early-digitizing counties to differ from those digitizing later. However, the timing of digitization is not strongly correlated with any observable characteristics of the counties.

Second, an unobserved factor, such as changing county business policies, could simultaneously affect both digitization and housing market activities. For instance, the construction of a new manufacturing plant could prompt a push for modernization within county administrations—leading to record digitization—and draw interest from algorithmic investors. To test for potential confounding by unobserved variables, I leverage bureaucratic inconsistencies in the timing of digitization for each property. Due to budgetary limitations, counties often digitized their records in batches, resulting in variability when each property became digitally accessible. With houses still awaiting digitization subject to the same potential unobserved variables—like changes in business policies—yet not being algorithmically assessable, these houses provide a natural control group.

To test for any evidence of confounders such as changing economic policies, I conduct a series of falsification tests at the county, house, and neighborhood levels and a triple-difference analysis. In the triple-difference analysis, I compare not-yet-digitized houses to digitized houses, before and after county digitization, in each county. In the falsification tests, I compare the impact of county digitization on digitized and not-yet-digitized houses within the same census tract, block group and block. Across all specifications, the observed effects are concentrated on properties that have been digitized, rather than those yet to be digitized. These patterns do not suggest substantial unobserved county- or neighborhood-level shocks driving the results. Overall, this suggests that algorithmic investors' activity depends on access to machine-readable data.

My second result investigates the natural question around digitization and the subsequent entry of algorithmic investors into the market: do these entities displace the small-scale, individual entrepreneurs who previously dominated the market? In fact, both traditional "mom-and-pop" entrepreneurs and algorithmic investors coexist in the post-digitization period. To make sense of this, I propose a conceptual framework that recognizes that humans and algorithms possess *distinct* comparative advantages. Humans have access to a wealth of non-digitizable information that is inaccessible to an algorithm. Humans see details such as the aesthetics of bathroom tiles, yard sunlight exposure, and ambient neighborhood noise levels. While limited to quantifiable data, algorithms derive their predictions from structured statistical relationships (Mullainathan and Obermeyer, 2021;

Kahneman et al., 2021). Furthermore, algorithms may not be susceptible to some sources of human errors, such as cognitive limitations or explicit prejudices. If the private information available to humans is important, then humans may do better. If humans make systematic mistakes, algorithms may have an advantage. This framework generates two clear predictions: Human investors should specialize where private information is important and where their comparative advantage is strongest. Algorithms should target properties where human errors are most prevalent. I empirically explore each of these hypotheses in turn.

In line with my theoretical framework, digitization leads human investors to shift their focus toward properties that pose difficulties for algorithmic prediction and away from those that are easily predictable. To capture where models do well and where they struggle, I use commonly available house data to predict price with an extreme gradient boosted tree model. Houses vary widely in their predictability. The discrepancy between actual and algorithm-predicted prices is as high as 50% for some houses. For others, the model error is less than 1%. Using this measure, digitization doubles the likelihood that human investors purchase the hardest-to-predict properties and reduces by half their propensity to purchase the most predictable houses. These results are consistent with human investors gravitating toward parts of the market where their informational advantage is highest.

What makes some houses easy to predict with a model and others so difficult? For some properties, important information is missing from the digitized data sets, leading to significant model discrepancies. In others, the unobserved information matters more. Take older homes as an example: they often have hidden issues like lead paint, presenting a serious health risk and entailing additional remediation costs. Since information about the presence of lead paint isn't accessible to algorithms—yet can be inferred or discovered by humans through inspection—models typically have a more difficult time valuing older properties. More recent houses built after the prohibition of materials like lead-based paints present fewer such challenges for algorithmic assessments. As a third example, data inaccuracies, such as errors in recording the number of bedrooms, disrupt accurate valuations with algorithms. While data errors cause problems for algorithms, human investors, who physically inspect properties and can count the number of bedrooms, will not be affected. These insights suggest a testable implication of my framework: Algorithmic investors should steer clear of properties where certain institutional factors enhance the informational advantage of humans. Using data errors, building regulations, and county zoning rules as illustrative examples, I show that algorithmic investors avoid houses with these characteristics, while humans invest in these houses.

In this framework, human errors create opportunities for algorithms. Although a variety of human behavioral biases could possibly generate human error, I will focus on racial bias (Whittle et al., 2014). Prior to the Fair Housing Act, race was explicitly used to determine house values. While it is now illegal to explicitly incorporate race, racial disparities remain and have been the subject of extensive study (Elster and Zussman, 2022; Perry et al., 2018; Freddie Mac Economic & Housing Research, 2021; Cutler et al., 1999). Specifically, prior work has identified a persistent valuation gap: minority homeowners tend to receive lower prices for their homes compared to White homeowners, even after adjusting for house and neighborhood characteristics (Elster and Zussman, 2022; Harris, 1999; Perry et al., 2018). This *race penalty* could be evidence that humans are undervaluing minority homes or driven by omitted variables or preferences.

Before exploring where algorithms buy houses, I estimate the race penalty in my sample prior to digitization and test for alternative explanations. Before digitization, a minority homeowner receives about 5% less than a White homeowner when selling their home, adjusting for house and neighborhood characteristics. While human biases could account for such disparities, this gap could also be driven by omitted variables that are correlated with homeowner race. A leading concern is that minority homeowners are often more cash-constrained or less wealthy, leading to differences in home maintenance or yard care reflected in the appearance of the house (Perry et al., 2018; Harris, 1999). To investigate whether this gap simply reflects omitted variables associated with home appearance, I utilize a deep learning model trained on images of house exteriors, yards, and driveways. After controlling for aspects of house quality captured by house images, the race penalty persists. This suggests that differences in property maintenance or aesthetic factors do not fully explain the lower price received by minority homeowners.

Consistent with the possibility that humans may undervalue minority-owned homes, algorithmic investors disproportionately buy minority-owned homes. The impact of digitization on the likelihood of purchase by an algorithmic investor is six times larger for a minority homeowner compared to a White homeowner in the same census block. In other words, digitization of minority homes leads to a 250% increase in the probability that an algorithmic investor buys that home compared to 40% for a White-owned house. Moreover, areas where algorithmic investors are active do not have much larger shares of minority residents than those where human investors buy houses. This suggests that algorithmic investors target minority-owned houses, rather than just neighborhoods with higher shares of minority residents.

In my third set of results, I investigate how these changes in market composition impact overall prices and racial disparities in prices. Before digitization, both owner-occupiers and human investors typically pay around 5% less for homes owned by minorities compared to those owned by white homeowners. However, algorithmic investors, who enter and buy houses after digitization, do not exhibit a race penalty. In other words, the price algorithmic investors pay for a house does not depend on the race of the homeowner. After digitization, the race-associated price discrepancy decreases significantly and eventually disappears within six years of a county transitioning to digitized records. That is to say, before digitization, observably similar homes sell for different prices based on the seller's race. After digitization, observably similar houses sell for the same price.

Importantly, human investors and owner-occupiers drive much of this reduction in market-level racial disparities. After digitization, the race penalty among owner-occupier purchases decreases from 5% to 3%. Among human investors, the race penalty falls from 5% to 1.5%. This decline can be attributed to two main factors. First, algorithmic investors' presence may drive up house prices through competitive bidding, affecting final sale prices even in transactions they do not win. Second, transaction prices inform listing prices for new homes on the market; higher starting prices lead to higher sale prices for minority-owned homes, regardless of algorithmic investor participation. Given that owner-occupiers represent about 80% of the market, these indirect effects drive the overall reduction in racial pricing disparities. As a result of these changes, the aggregate impact of digitization is a 5% increase in average sale prices for homes owned by minorities, compared to a 1% increase for White-owned homes. These findings highlight how market interactions can amplify the impacts of algorithms in ways that firm-level analysis cannot capture.

Although one explanation for the increasing prices of minority homes might be human mistakes, another possibility is that algorithmic investors are simply overpaying for minority-owned houses. Algorithms do not see all aspects of house quality available to humans—potentially leading to adverse selection. To investigate whether this is occurring, I examine the gross returns—the difference between purchase and resale prices. If algorithmic investors consistently overpaid for minority-owned properties, their gross returns should be lower than the gross returns on White-owned homes. In fact, the gross returns on White and minority-owned homes are not statistically different. This holds true whether I use the resale price or an alternative measure of value, such as tax assessor estimates, to calculate gross margins.

Further undermining the adverse selection hypothesis is the increasing price that human buyers pay for minority-owned homes after digitization. If minority-owned homes were unobservably bad, human buyers should not be willing to pay more for these properties. Instead, the price human buyers are willing to pay for minority homes also rises after digitization. Together, this evidence is more consistent with humans previously undervaluing minority-owned than with algorithms overvaluing such properties.

So far, my analysis has focused on the transaction prices of the sold properties. Nonetheless, it's crucial to consider how these shifts might influence the valuation of unsold homes—assets that constitute a substantial proportion of wealth for the median household. My estimates suggest that digitization leads to a 6% appreciation in the average value of unsold minority-owned homes. This appreciation is considerable when viewed in relation to median households' wealth: a 6% rise in property values corresponds to an increase by roughly 20% and 13% of the median Black and Hispanic family wealth, respectively (Bhutta et al., 2020).

These findings highlight how markets can amplify the effects of algorithms. Here, low average levels of algorithmic investor activity induce significant changes that impact those using algorithms and those not using algorithms alike. Competition and price effects lead to a reduction in racial disparities in property values, affecting homeowners across the market, and change the behavior of human investors and owner-occupiers. My findings are similar in spirit to Becker (1957), where competition penalizes and drives out firms with discriminatory views. The magnitude and patterns of these effects raise questions about how algorithms could be reshaping other parts of the economy.

This paper builds on a growing empirical literature on the impacts of access to algorithmic recommendations. Comparing human decision-makers' choices with predictive models has a long history (Dawes, 1971; Dawes et al., 1989; Hastie and Dawes, 2001). Modern advances in ML, increased computing power, and data availability have renewed interest in these questions. I build on prior work that shows that algorithmic recommendations can lead to, for example, improvements ranging from better heart attack diagnosis, to more efficient bail and hiring decisions.[5] Other work shows that access to algorithms translates into productivity or efficiency.[6] However, not all studies find positive effects.[7]

There is limited work on the effects of ML-powered algorithms at the market level. Calvano et al. (2020); Clark et al. (2023); Calder-Wang and Kim (2023); Brown and MacKay (2023) focus on the impact of ML-powered pricing algorithms on collusive behavior and price levels. Other studies focus on the impacts of automated algorithmic trading on the liquidity and pricing efficiency of financial markets (Hendershott et al., 2011; Chaboud et al., 2014; Upson and Van Ness, 2017). I examine the market-level impacts of algorithmic prediction outside of financial market trading.

[5] For example, see Autor and Scarborough (2008); Li et al. (2020); Raghavan et al. (2020); Frankel (2021); The White House (2022); OECD (2023) for applications in the labor market, Einav et al. (2013); Fuster et al. (2022); Gillis and Spiess (2019); Arnold et al. (2018); Blattner and Nelson (2021) for consumer finance, Mullainathan and Obermeyer (2021); Obermeyer and Emanuel (2016); Kleinberg et al. (2016); Chouldechova et al. (2018); Abaluck et al. (2020); Kleinberg et al. (2017a); Mullainathan and Rambachan (2023) for examples in the criminal justice system, health care, among other areas. See Rambachan (2022); Kleinberg et al. (2017b, 2015) for issues comparing human and machine predictions.

[6] See Brynjolfsson et al. (2023) for the impacts of generative AI on productivity in customer service, Harris and Yellen (2023) for the impact of the adoption of predictive maintenance on repair costs in a trucking company. See Bubeck et al. (2023); Choi and Schwarcz (2023); Peng et al. (2023); Noy and Zhang (2023) for additional effects of AI access on productivity, writing, and test taking capabilities.

[7] For instance, Acemoglu et al. (2022) finds no detectable relationship between AI investments and firm performance, while Babina et al. (2022) finds a positive relationship.

This paper is closely related to a large literature on racial disparities in the housing market. Although centralized discrimination has declined over time, audit studies, surveys, and empirical work continue to find evidence consistent with racial discrimination in the housing market.[8] Racial disparities in house values contribute to the large and persistent racial wealth gap.

Initially, it was hoped that the use of algorithms would help reduce racial disparities. For example, Kleinberg et al. (2018) show that reliance on algorithms to grant bail could simultaneously reduce crime, jail populations, and racial disparities. However, there are many examples of algorithmic bias, or algorithms that disparately direct fewer opportunities or resources toward minorities.[9] This paper is the first to show the indirect effects of algorithms on racial bias that work via market competition.

Understanding the impact of large investors on the housing market is an important policy question. For instance, in December of 2023, Democrats introduced legislation in the House and Senate that would ban hedge fund ownership of single family homes (Kaysen, 2023; Merkley and Smith, 2023). A growing interdisciplinary body of work has examined the impacts of large single-family investors in the US and, more recently, Europe.[10] This paper speaks to policy discussions around algorithms in the housing market and digitization of public records.

1 Background

I provide some background on real estate investment and human and algorithmic investors and elaborate on why real estate investment is fundamentally a prediction task. After setting out the prediction problem, I detail the human and algorithmic investors' approaches to prediction. I describe county real estate records, the timing of the digitization process, and explain why digital county records are significant for algorithmic valuation. Then, I provide some relevant background on racial disparities in the housing market.

[8] For example, see Elster and Zussman (2022); Perry et al. (2018); Perry (2021); Bayer et al. (2017); Kim (2000); Freddie Mac Economic & Housing Research (2021); Zhang and Leonard (2021); Kermani and Wong (2021b); Lewis et al. (2011). See Cutler et al. (1999) for a summary of centralized discrimination.

[9] See Smith (2021) for a summary of empirical work on algorithmic bias. See Rambachan and Roth (2019), Rambachan et al. (2020), Bakalar et al. (2021), Kleinberg et al. (2016) and Cowgill and Tucker (2019) for theoretical work.

[10] Fields (2018, 2022) examine how technology-driven "calculative agency" enabled the financialization of the single-family housing market. Raymond et al. (2016, 2018, 2021) study the impacts of institutional investors in Georgia and housing insecurity. Mills et al. (2019) provides some empirical early-stage analysis of the activities of these firms. Gurun et al. (2023) study the increase in institutional investor ownership and the impacts of investor mergers on rent and neighborhood safety. Buchak et al. (2022) studies the "i-buyer" firms (e.g. Zillow, Offerpad, Redfin and Opendoor) and their impacts on liquidity in the housing market. Francke et al. (2023) examine the impact of a ban on large institutional buyers of housing in the Netherlands.

1.1 Real Estate Investment

1.1.1 Single Family Homes

Residential real estate is the largest asset class in the United States, with a total value of $43 trillion (Malone, 2023). Single-family detached houses comprise 86% of the value of all residential real estate and 66% of the entire housing stock and are common in rural and urban areas (Malone, 2023; Neal et al., 2020). In my sample, single-family houses make up 66% of the occupied housing in urban areas and 72% in rural areas (U.S. Census Bureau, 2021). Single-family houses are purchased by two types of buyers: owner-occupiers, who buy houses to live in, and investors, who buy to rent out or flip these properties. Although most strongly associated with homeownership, 17%, or about 14 million of these homes, are occupied by renters. These houses make up the largest single segment of rental housing (about 41%) and are particularly important in areas less urban and with lower income (Census, 2023; Neal et al., 2020; Freddie Mac Economic & Housing Research, 2018).[11]

1.1.2 Investing is a Prediction Problem

Investing depends on the prediction of net income and asset value. Investors forecast possible house income from asset value and rent against upgrades, repairs, and ongoing maintenance. Investors assess the physical condition of the home, both inside and outside, including the structure, the fixtures in the bathroom, and the electrical systems. They will try to confirm actual square footage and configurations; for instance, does the house have an illegal unzoned bedroom that the owner is to claim value for? What do the neighboring houses look like? Does the physical layout of the house make good use of the space? Officially, the metric investors try to predict is the capitalization rate, or net income divided by asset value. The capitalization rate is a standard metric used by investors in the real estate industry to compare properties.[12]

1.1.3 The Human Informational Advantage and "Mom and Pop" Investor

Local entrepreneurs are best positioned to solve this prediction problem. In fact, it was widely believed that these "mom-and-pop" entrepreneurs would always dominate the single family home market due to their informational advantage (Fields, 2018; American Homes 4 Rent, 2013, 2018). Local residents know where traffic is bad, which neighborhoods have the best parks, recent patterns of gentrification, and closings of manufacturing plants or retail stores. These "mom-and-pops" also often had a background in construction and local real estate, helping them evaluate the costs

[11] In the US, single-family homes are detached dwellings built to be occupied by one household on their own plot of land.

[12] Appendix Figure A.1, shows an example of the capitalization rate information provided for a multifamily property.

and time required to complete each repair accurately. Individuals familiar with local construction practices can estimate how much exposure to moisture will degrade the foundation of the house. The average mom-and-pop owns a single property in their local area and often works in construction or real estate (Fields, 2022; American Homes 4 Rent, 2013).

In principle, a single company could employ a large number of individuals to evaluate and acquire property. However, evaluating single family homes that are scattered and structurally unique with people is prohibitively expensive in terms of money and time (Amherst, 2016). In the early 2000s, Redbrick Partners attempted to assemble a large portfolio of single-family homes using this strategy. The firm amassed 1,000 homes over the next four years, but struggled to acquire and manage individual houses efficiently. Despite the rapidly increasing price of houses, the firm determined that it was too costly to deal with spatially and physically distinct housing units without technology (Mills et al., 2019). As a result, Redbrick Partners decided to exit the business in 2006 (Fields, 2018).

1.1.4 Investing is a Challenging Cognitive Task

Although human investors have access to huge amounts of information, processing all of this into a single estimate is a challenging cognitive task. How to weigh the value of going from a one-car garage to two-car garage while taking into account the nice local park, old bathroom fixtures, and trees that may need to be cut down? In scenarios where humans have to weigh a lot of different information, human cognitive limitations can hinder accuracy. For example, Mullainathan and Obermeyer (2021) show that doctors seem to rely on an overly simplistic model to predict heart attacks and overestimate the importance of physical symptoms such as chest pain. Humans also generally do not have experience in learning from thousands of houses—they are limited to their own experience.

Human investors expend substantial effort to structure their decision processes to avoid errors. An industry standard practice is to develop a "buy box" that guides which houses they *will* buy.[13] A buy box is a list of criteria that outline where an investor buys houses and where they believe they have an advantage in valuing houses. An investor might have a buy box that targets houses in a specific zip code in Fresno, California where "the middle class lives" with a diverse employment base between two and three bedrooms. Another industry standard best practice is a blacklist that *eliminates* houses from consideration. For example, an investor could avoid all houses that require electrical system, roof, or septic tank repair because these construction projects are notoriously unpredictable. These tools are efforts to help investors avoid two well-known pitfalls: buying houses

[13] This is also an algorithm, but not a ML algorithm. For example, see New Investors Must Start with a Buy Box or they are wasting time and money.

that "feel like a great deal" or houses that are aesthetically pleasing but with significant structural flaws. However, doing this well is challenging. Redfin estimates that investors lose money on one in seven homes (Redfin, 2023a).

1.1.5 Using Algorithms to Value Houses

ML algorithms, also known as acquisition, automated valuation, or acquisition engines, use statistical patterns in the data to predict the value of the house. Investors use a wide variety of data sources: population, homeownership, vacancy rates, income, crime index, school quality, recent transactions, type of construction, ongoing capital needs, and employment, among others (Amherst, 2016; Invitation Homes, 2017). The set of possible factors that might be useful in estimating net income and asset value is high dimensional. The high dimensional nature of the data creates considerable risk of in-sample overfitting, leading to bad out-of-sample prediction. ML algorithms, which strike a balance between penalizing model complexity and maximizing accuracy, are crucial to doing this well. ML is also necessary because the house value has no explicit formula, requiring a data-driven mechanism to identify patterns and correlations. These algorithms distill the enormous amount of information available into a single estimate of net income. Although building these technology platforms is expensive and requires specialized teams of data scientists and software engineers, "[w]ithout using technology to filter and deliver automated valuations... it would be extremely time-consuming and inefficient to review and bid on these properties... The entire process uses a vast amount of data that is impossible to distill into actionable information without the use of technology" (Amherst, 2016; Christophers, 2023).

These algorithms are embedded within an "acquisition team" of human analysts who monitor the houses found by the algorithms. Unlike human investors, these buying teams do not drive around neighborhoods looking for houses. Instead, the algorithms filter through all available houses for sale, estimating net income. The most attractive houses are sent to a queue for the buying teams to review from their desktops (Fields, 2022).[14] Many of the acquisition teams are located in New York, California, and Texas and may never visit the neighborhood where they own houses. The buying team takes the list of properties found by the algorithms, reviews them, and manages the process of generating an offer.[15] Although the algorithm is embedded in a human buying team, the buying team does not physically search for properties themselves and the algorithm does not have

[14] According to its IPO prospectus, Invitation Homes, one of the largest single-family investors, underwrote more than a million homes to assemble its portfolio of 50,000 properties.

[15] Many algorithmic firms employ their own internal real estate agents to make offers on properties. Offers are made primarily by real estate agents to homeowners.

access to all of the information available if they were to visit a property in person.[16] I will formalize this trade-off more explicitly in Section 4.1.

1.1.6 Algorithms Enable a "Factory-Like" Production Line

ML algorithms produce a single quantifiable house value estimate that can be interpreted without the need for a local context. In the words of one analyst, "... capital markets cannot get into a home... So, [algorithms] take all expenses, all maintenance, water heaters, roof, and flatten them into a format that can be consumed by capital markets" (Fields, 2018). ML algorithms "flatten" a single-family home into a numerical estimate of net income that can be integrated into formal decision-making processes, without the need for deep knowledge of local construction practices. This reduces the cost of acquiring and managing single family homes faced by Redbrick Partners, who found it too costly and inefficient without such a technology.[17] Algorithms helped create a "factory-like" production line for the acquisition, renovation, and leasing of single-family homes (Fields, 2018).

1.2 County Housing Records

Algorithms depend on housing data produced by county governments. In this section, I describe county records, the digitization process, and the impacts of digitization on investors in the housing market.

1.2.1 The Process of County Record Digitization

County governments' records are the most accurate and up-to-date sources of housing market activity and the characteristics of the housing stock.[18] These records, which were kept in paper books or microfilm, are frequently used in day-to-day county business. Dividing property in a divorce proceeding, building and engineering planning, genealogy research, and verifying property ownership all require access. However, paper and microfilm records are not easy to search for, are expensive to maintain, susceptible to physical damage, and are difficult to access and store. Spurred by the

[16] According to Amherst Residential, about five hundred homes newly for sale are listed daily within its target geographic markets and Amherst Explorer, the firm's algorithm filters these listings and delivers automated valuations by estimating potential rents, refurbishing costs, taxes, insurance, and other expenses to calculate an estimated net operating income. Each morning, the firm has a list of targeted properties with projected returns that run automatically before anyone even has had time to drink coffee Amherst (2016).

[17] Furthermore, unlike small investors, who also oversee the property repairs themselves, algorithmic valuation helps other individuals at the same firm make decisions around repair and maintenance costs without ever seeing the house in person.

[18] By law, County governments are responsible for maintaining public records of property; the Recorder's office maintains and preserves all legal documents affecting title to real property, and the Assessor's office determines the value of real property to collect property taxes. Deed records are public records that date back to county founding; some land records date back to the 1600s.

clear downsides of paper records and the 2009 Obama Administration efforts to promote digital and transparent government, many counties began to digitize. Digitization transferred these paper and microfilm records into a digital database and made them publicly accessible and searchable on the Internet (The White House, 2009).[19] Panel A of Figure 1 shows the share of counties with publicly available digitized records. The share of counties with digitized records rose from 40% to 80% by the mid-2010s.

The time to complete digitization was determined by legislative and budget allocation decisions, as well as by technical difficulties of digitization. First, each state had to ensure that county recorders could legally store their records digitally.[20] Then, each county needed to allocate funding. Digitization required scanning and indexing each paper or microfilm record, a time-consuming and expensive process. Counties tended to work backward from their most recent records, digitizing houses by year of the last sale. Next, each county needed to construct a software database, requiring significant investment in information technology, and connect this database to their website. Finally, each state determined a common standard for computer systems.

The time to complete digitization varied significantly between counties. Panel B of Figure 1 shows the share of counties with publicly available digitized recorder systems by state over time. In general, due to the coordination required to digitize records, there are sharp spikes in the share of digitization within each state. However, the year each county process varied significantly due to unexpected issues with setting up the database, digitizing records, or funding, leading to idiosyncratic variation.

1.2.2 How County Digitization Changes the Housing Market

Digitization affects the housing market through three channels: real-time data availability, training data derived from historical transactions, and comprehensive information on the characteristics of each house. Once a county transitions to digital records, all new housing sales become immediately available online. This real-time, reliable, and accurate information is vital in enabling algorithms to update promptly, learning which houses are on sale and which have recently sold. The real-time availability of digital data was consistently highlighted as the most impactful change.[21] Digitization also makes it easy to download historical transaction data. This digital information serves as training data for the algorithms discussed in Section 1.1.5.

[19] Because this information is public data, digitizing these records also required making them accessible online.

[20] I use the year county Recorder deed records are first available. In practice, property characteristics data also generally become available at this time.

[21] Interview with the Georgia Superior Court Clerks' Cooperative Authority. MLS data and Zillow data are considered unreliable because they depend on accurate data entry from real estate agents and are generally not updated in real time. Private data providers, especially in the early 2010s, either did not have data or failed to provide real-time data updates.

To predict value, algorithms require a digital representation that includes its characteristics—the number of bedrooms, bathrooms, stories, and whether it has a basement. When a house is in the database, digital records of the characteristics of the house are readily available and easy to assess algorithmically. If the house has not yet been added, investors would need to manually collect these data to estimate the value, making it harder to value these houses with an algorithm. [22] I leverage the bureaucratic variation in when each house was digitized to perform robustness checks and estimate house-level effects.

1.3 Racial and Ethnic Price Differences in the Housing Market

Prior to the passage of the Fair Housing Act, race was explicitly taken into account when estimating the value of a house. For example, "The Valuation of Real Estate," a popular textbook for real estate appraisal, claimed that neighborhood decline inevitably results from occupation by ". . . the poorest, most incompetent, and least desirable groups in the city," and described how "... racial heritage and tendencies seem to be of paramount importance" in determining property values (Babcock, 1932; Wheaton, 2023). While it is now illegal to explicitly incorporate race, racial disparities remain in house values. I review the evidence on racial disparities in house values. In Section 4.5.1, I examine racial disparities in my sample.

1.3.1 Evidence on Racial Disparities in House Prices

Racial disparities exist in house prices. Harris (1999) documents that moving from a less than 10% Black to between 10% and 60% Black neighborhood is associated with a 2.3% drop in house value, accounting for house and neighborhood characteristics. Perry et al. (2018) estimate that homes lose 23% of value when moving from a census tract with 0% Black residents to one that is 50% Black. At the building level, Elster and Zussman (2022) find that house prices decrease 2 to 3% after minorities move in.

Price disparities could reflect preferences, omitted variables, or biases. White homebuyers exhibit a strong negative outgroup bias (negative perceptions or prejudices towards those not in their group) toward living in areas with Black and Hispanic neighbors. Minority home buyers do not show strong preferences and are willing to live in a variety of places, including majority White neighborhoods (Lewis et al., 2011). Minority neighborhoods could be associated with higher crime, lower investment, and lower property values (Harris, 1999; Lewis et al., 2011; Howell and Korver-Glenn, 2018; Perry, 2021; Freddie Mac Economic & Housing Research, 2021). Price differences could reflect omitted variables correlated with the race of the homeowner, such as differences in

[22] Collecting this data by hand is possible, but significantly more costly.

neighborhood amenities or house characteristics. For example, levels of pollution and noise are typically higher in minority neighborhoods (Casey et al., 2017; Tessum et al., 2021).

Yet, neighborhood characteristics cannot fully explain price disparities because disparities persist even when considering the value of the same house. Appraisal is the process through which a real estate appraiser estimates the fair market value of a house for property tax or credit purposes. Widespread anecdotal accounts of appraisal undervaluation have been reported for minority buyers. After receiving a low appraised house value, some minority homeowners have tried to "whitewash" their homes by removing all family photos, asking a White friend to stand in as the homeowner, and received higher estimates of house value in a second appraisal (Kamin, 2023; Lilien, 2023; Howell and Korver-Glenn, 2018). A very small audit study of this found that, on average, a White homeowner received a 7% higher appraisal than a minority couple for the same house (Lilien, 2023). In general, minority-owned homes are more likely to receive appraisal estimates below what a buyer has offered to pay, even when considering the characteristics of the house and the neighborhood (Freddie Mac Economic & Housing Research, 2021; Perry, 2021; Howell and Korver-Glenn, 2018). This suggests that the omitted variables may not fully explain racial disparities in prices. In Section 4.5.1, I examine racial disparities in prices in my data.

2 Data and Empirical Strategy

2.1 Empirical Strategy: Digitization

My analysis uses a difference-in-difference (DiD) analysis. I use a dynamic event study with differential timing to isolate the causal impact of digitization on the entry of algorithmic investors and market-level and house-level outcomes:

$$y_{ct} = \delta_t + \alpha_c + \sum_{j \neq -1}^{J} \beta^j \times \mathbb{1}[t = j] \times D_{ct} + \gamma X_{ct} + \epsilon_{ct} \tag{1}$$

The outcome variables y_{ct} capture the results for county c and year t. First, I examine the impact of digitization on algorithmic investor entry. The outcome is $y_{ct} = ln(1 + q_{ct}^{algo})$, the natural log transformation of one plus the number of houses purchased by algorithmic investors (q_{ct}^{algo}) in county c and year t. I estimate the impact of digitization on price using $y_{ct} = ln(price_{ct})$ or the natural log of the county average sale price of houses in year t. D_{ct} is an indicator equal to one if county c has digitized in year t and 0 otherwise. Digitization is an absorbing state; once a county builds a database system, they do not return to paper records. Counties that had not been digitized by 2017 are used as controls. All regressions include year fixed effects (δ_t) to account for factors that vary over time such as interest rates, housing market policy and other macroeconomic variables. I also account for time-invariant factors specific to each county, such as size, income levels, and geography (α_c). Standard errors are clustered at the county level. The β^j vector is the parameter of interest that captures the time-varying treatment effect of digitization. At the county level, I weight the regressions based on the number of property transactions in each county-year.

I use a series of dynamic differences in difference estimators that are robust to the effects of digitization varying over time. The treatment effects of digitization could increase over time as algorithms may become more accurate and organizational processes are established. On the other hand, the effects of treatment could also decrease as competition in the housing market intensifies. To address time-varying treatment effects, I use the Sun and Abraham (2021) interaction weighted estimator (IW) that is robust to the correlation over time and across adoption cohorts. I also present results using a series of additional robust estimators introduced by de Chaisemartin and D'Haultfœuille (2020), Borusyak et al. (2022), Callaway and Sant'Anna (2021) as well as using traditional two-way fixed effects regression analysis. In general, estimates from robust estimators are larger and more stable because they avoid comparisons between already-treated counties.

These estimators require three assumptions: no anticipation, no spillovers between treated and not-yet-treated counties, and parallel trends. First, participants should not change their behavior in anticipation of future treatment. Second, digitization in one county should not impact the housing market in a county that has not yet been digitized. Third, in the absence of treatment, the treatment and control groups would have evolved similarly.[23] For example, there should be no changes in county economic policy that differentially impact treatment and controls. In Section 3.2, I examine the robustness to a series of alternative explanations.

In Figure 1, I plot the share of counties with accessible and digitized county Recorder databases over time. The sharp nature of digitization patterns is important to my empirical strategy. The discrete change in digitization will generate discrete changes in algorithm availability, while other unobservables should evolve smoothly around the threshold.

I also estimate a series of cross-sectional hedonic regressions at the house level. This complements the county-level analysis and allows me to explore the impact of house-level digitization (D_{ict}) on house-level outcomes, accounting for differences in observable house characteristics. We examine the likelihood that an algorithmic investor purchases an available house, denoted $\mathbb{1}[q_{ict}^{algo} = 1]$, and a natural logarithmic transformation of the sale price. At the house level, algorithmic purchase could be correlated with unobserved aspects of the house, the number of bidders, time on the market,

[23] In another way of saying the same thing, the timing of digitization is not correlated with first stage or reduced form outcomes.

or the tech-savviness of the listing real estate agent. To address this, I also perform a Two-Stage Least Squares (2SLS) regression, where purchase by an algorithmic investor is instrumented with digitization (Angrist et al., 1996).

$$\mathbb{1}[q_{ict}^{algo} = 1] = \delta_t + \alpha_g + \beta D_{ict} + \gamma X_{ict} + \epsilon_{ict} \tag{2}$$

The second stage of the relevant house level regression, run using 2SLS to obtain correct standard errors, is:

$$y_{ict} = \delta_t + \alpha_g + \beta \times \widehat{\mathbb{1}[q_{ict}^{algo} = 1]} + \gamma X_{ict} + \epsilon_{ict} \tag{3}$$

2.2 Data

My sample includes data from 400 counties in Georgia, North Carolina, South Carolina, and Tennessee, spanning the period between 2009 and 2021. Information on property records comes from the county governments. I use detailed property-level house characteristics and sales information from ATTOM Data and Zillow. I also rely on aggregated rental and listing data from Zillow and demographic and socioeconomic data from the US Census.

2.2.1 Digitization Data

I hand-collected data on county record digitization from county recorders' offices, the Internet Archive, and ATTOM Data. The primary source of information was direct interviews with county officials. County officials provided the year when their transaction records first became publicly available online. Once counties switched to electronic records, all future property transactions were automatically digitized, and information on recent transactions became immediately available online. Database systems also enabled easy download of historical data and house information.

I supplemented these interviews with snapshots of county websites from the Internet Archive. These snapshots provide verification of when the county website first provides remote access to the county records. Counties did not keep systematic records when each house record was digitized. Instead, I collect this information from ATTOM Data, who tracked when each record was added. I discuss further details on digitization in Sections 1.2 and 1.2.2.

A central concern with hand-collected data is the potential for measurement error. To address this, I use the digitization year provided by ATTOM to corroborate the county information. Although these two series will not align perfectly—-since houses are not all digitized at once and new houses are continually added—the two are similar. To validate the year of digitization of the ATTOM house record, I compared the year provided by ATTOM with the year of digitization from a subset of Georgia counties that maintained more detailed records of house-level digitization. While these records are no longer updated, I collected copies of this information stored by the Internet Archive. Once again, the ATTOM Data year of digitization closely corresponds to county records.

2.2.2 Identifying Investors

Investors are corporate entities that buy houses to rent out or resell homes (Redfin, 2023b). I exclude government entities, banks, credit unions, timeshare operators, securitized mortgage trusts, homeowner associations, churches, corporate relocation services, hotels, vacation rentals, farms, builders, and property owner associations. This definition follows other work on investors in the single-family market (Redfin, 2023b; Mills et al., 2019).

After identifying all investors, I categorize each firm as human or algorithmic. I identify algorithmic investors and their properties using business registration information, public filings, and personnel records. I start with properties owned by corporate entities and identify corporate mailing addresses (Gurun et al., 2023; Mills et al., 2019). To match subsidiaries to the parent firm, I perform two rounds of fuzzy clustering, first on the mailing address and then on public business registration data, properties listed on landlord websites, and known lists of corporate subsidiaries from SEC filings. After this two-round matching procedure, I determine whether each firm's investment strategy is algorithmic or not using SEC filings, news articles and interviews, company websites and personnel records. If the company uses an "algorithmic acquisition engine" or "automated valuation platform" or employs a data science or software engineering team, I code them as algorithmic. Consistent with previous studies, I find about 40 algorithmic investors in my data, which own about 130,000 houses (Mills et al., 2019; Gurun et al., 2023).[24] Although not all companies using algorithmic valuation conduct interviews or file with the SEC, all companies have business registration data, websites, and personnel records available on LinkedIn.[25]

I identify human investors as those using non-algorithmic acquisition strategies. I rely on news articles, interviews, company websites, and personnel records to determine whether a firm relies primarily on human judgment to evaluate investments. As a result of the dominance of the mom-and-pop entrepreneur, no human investors are public firms that file with the SEC. However, most of the human investors have websites or personnel records available on LinkedIn, and all have business registration data. Due to the time-intensive nature of this search process, I only explicitly categorize firms with at least 80 purchases in my sample. Of the entities with less than 80 purchases over the

[24] There are a series of consolidations between the algorithmic investors in the dataset such that at the end of the sample, the total number of firms is smaller.

[25] Only algorithmic investors that are publicly traded REITS or involved sale of securities to investors, must submit SEC filings.

decade in my sample, of those that are not categorized as algorithmic, I assume that these are investors using human judgment.

2.2.3 Housing Market Data

Residential housing market comes from ATTOM Data and Zillow's ZTRAX database. Both sources provide records from county Recorder offices and county property tax assessor records. The recorder office data include detailed property transaction information, including sale price, date, identities of buyers and sellers, the corporate structure of the buyer or seller, any relationship between the two, and indicators for arms-length transactions and sales of newly constructed houses.

The tax assessor records provide detailed property and yard characteristics such as property type, longitude and latitude, year built, architectural style, number of bedrooms and bathrooms, type of air conditioning and roof construction material. The records also include estimates of the house market value, land and improvements over time. As of 2023, all housing records in this sample dating back to the early 2000s have been digitized, enabling historical analysis. I drop non-arms-length and multiparcel transactions. I geocode each house to the corresponding census county, tract, block group, and block, using latitude and longitude. Two percent of houses cannot be geocoded to the census block level, but all houses are geocoded to the census block group level.

I supplement the house-level transaction files with various publicly available information from Zillow on housing market dynamics. These measures include the average sale price to the list price, the share of listings with price cuts, the median sale prices, and the share of sales over the list price at the zip code and county level.

In addition to the information on each house, I scraped exterior and interior house images. I collect these images from Zillow and investor websites where these properties are listed. Images are only available for a subset of the houses in my sample, about 50,000 houses. Then, I process exterior house images into vector embeddings for analysis using a deep learning model, which I describe in Section 4.5.2.

I also use a variety of socioeconomic and demographic variables from the American Community Survey (ACS) and the 2010 and 2020 Decennial Census. Many of the counties in my study have fewer than 65,000 residents and thus do not meet the 1-year ACS inclusion threshold, and I rely on the 5-year ACS waves instead. I use a variety of demographic and socioeconomic data, including factors such as median income, median age, racial composition, education, the fraction of the population that is rent burdened, median rent, household size, share in the labor force, share unemployed in the county, census tract, block group, and block level.

2.2.4 Identifying Homeowner Race

I use the Bayesian Improved Surname Geocoding (BISG) proxy method to infer race and ethnicity from publicly available homeowner names. The BISG model predicts race and ethnicity based on owners' surnames and census block addresses using Bayes' theorem. This approach is widely adopted in fair lending analysis (Elliott et al., 2009). The Consumer Financial Protection Bureau, which uses this algorithm for fair lending analysis, has conducted accuracy tests in mortgage lending, a setting that closely mirrors my own (Consumer Financial Protection Bureau, 2023). Using census block geocoding, BISG exhibits Area Under the Curve (AUC) scores of 0.94 or higher across classifications, including Hispanic, Black, non-Hispanic White and Asian borrowers. This suggests the model can accurately categorize races and ethnicity from geography and surname information.[26]

2.3 Summary Statistics

I present some summary statistics on the houses purchased by owner-occupiers, human investors, and algorithmic investors.

2.3.1 House-Level Descriptive Statistics

Owner-occupiers make up the bulk of the market. Owner occupiers buy about 86% of all houses as shown in column 1 of Table 1. On average, they purchase 2.12 bedroom, 2.14 bathroom houses, 30 year old houses with a garage, a parking space, and a fireplace for about $194,270.

Human investors purchase on average less expensive, older homes. However, these houses are not significantly different from the overall population. Column 2 of Table 1 shows that human investors are more likely to buy slightly smaller, less expensive, and older houses around $127,755 that are less likely to have a garage and a parking space.

Algorithmic investors tend to purchase newer, larger, and more expensive homes. As shown in column 3 of Table 1, these investors focus on properties with 2.76 bedrooms, 2.47 bathrooms, a mean transaction price of $219,130, and almost always include a parking space. The houses they purchase are, on average, only 21 years old and were remodeled 18 years ago.

The most striking difference between human and algorithmic investors is the very low variation in characteristics of houses purchased by algorithms and the very large standard deviations among houses purchased by human investors. The standard deviation on all house characteristics in column 3 of Table 1 are much smaller than column 2. These differences can be illustrated even more clearly in Panels A through D of Appendix Figure A.5. The distribution of houses purchased by algorithmic

[26] AUC scores range from 0 to 1 and represent the model's classification accuracy. A score of 0.5 indicates that the model performs no better than random guessing, while 1 indicates perfect classification.

investors, relative to human investors, is much more concentrated in terms of bedrooms, bathrooms, age, and sale price. I will return to this in more detail in Section 4.

Table 2 shows the county-level characteristics of the houses purchased by human and algorithmic investors. Algorithmic investors are active in slightly larger and wealthier counties with a higher Hispanic population. Otherwise, the characteristics of the county are relatively similar.

2.3.2 Firm-Level Descriptive Statistics

Algorithmic firms also tend to be much larger firms operating in wide geographic areas. As shown in Panel E of Appendix Figure A.5, before digitization, the market is dominated by many small firms. Conditional on participating in the market, the average human investor purchases a single house. Algorithmic firms purchase an average of 2,000 houses a year. As a result, once a county digitizes, the scale of the largest firms in the market increases significantly. Panels F of Appendix Figure A.5 shows algorithmic investors active in close to 300 different zip codes each year. They have less than 5% of purchases in the same zip code as their corporate mailing address. However, 40% of the houses bought by human investors are in the same code as the investor's corporate mailing address.

3 Digitization Leads to Algorithmic Investor Entry

3.1 County Digitization and Entry

The raw data clearly demonstrates the impact of digitization on the buying behavior of algorithmic investors. Panel A Figure 2 shows the natural log transformation of the number of houses purchased by algorithmic investors in each county, by time to digitization. Panel B of Figure 2 shows a sharp increase in market share, which increased from nearly zero before digitization. Following digitization, algorithmic investors buy on average 2% of all houses sold. Panel C of Figure 2 shows that the number of houses bought by human investors does not change.

Figure 3 presents the analysis of the accompanying event study that shows similar large and persistent increases in the number and share of homes bought by algorithmic investors. Panel A of Figure 3 shows that county digitization leads to a 200 log point increase in the number of houses purchased by algorithmic investors following digitization.[27] This increase persists and remains stable until the end of the sample period. Panel B of Figure 3 shows that digitization is associated with an increase in market share of algorithmic investors of 2%. All regressions are adjusted for county- and year-fixed effects and weighted by the number of transactions. Standard errors are clustered

[27] The increase is $e^(2) - 1 = 6.4$

at the county level. County- and year-fixed effects account for time-varying common shocks that impact the housing market and county-specific characteristics.

Alternative estimators show similar results. In Appendix Figure A.3, I show the results are similar using alternative event study estimators: Borusyak et al. (2022), Sun and Abraham (2021), de Chaisemartin and D'Haultfœuille (2020) and the traditional fixed two-way effects model. Robust estimators avoid comparing newly treated units with already treated units, thus delivering larger and more stable estimates than the two-way fixed effects model.[28]

In Table 3, I present the corresponding DiD estimates. Across estimates, I find digitization leads to large increases in the number of houses purchased by algorithmic investors. The Callaway and Sant'Anna (2021) estimates of a 100 log point increase are lower because this estimator cannot be weighted by county size. Taken together, I interpret these results to suggest that county digitization and the subsequent increase in data availability, on average, lead to a sharp and sustained increase in home purchases by algorithmic investors.

The timing of county digitization is not related to observable county characteristics. Table A.1 shows that early and late digitizing counties are balanced in unemployment, income, other demographics, rent, and vacancy rates. Early digitizing counties are larger and have a 1-percent higher Hispanic population than late digitizing counties, but are otherwise similar in socioeconomic and demographics. In Appendix Table A.4, I show how estimates from the standard DiD vary with additional controls. In column 1, I show that, controlling for county and year fixed effects, digitization increases the number of houses purchased by algorithmic investors by 113 log points. In column 2, I shows how estimates vary with additional controls for pre-digitization county socioeconomic status, including demographics, poverty, unemployment, share with young children and educational attainment. In column 3, I add controls for the pre-digitization number of housing units and rent burden. In general, the estimates fall slightly, but remain stable. I interpret these results to suggest that my estimates of the impact of digitization are not driven by systematic differences in observables between counties.

I also see similar strong impacts at the house level. Column 1 of Table 4 shows that digitization results in a 17-fold increase in the likelihood that an algorithmic investor purchases a home compared to a non-digitized house in the same census tract. Column 2 indicates a 16-fold increase compared to a non-digitized house in the same census block group. Column 3 demonstrates a 7-fold increase within the same census block. These results suggest that algorithmic investors are sensitive to the availability of digital information when valuing houses.

[28] Callaway and Sant'Anna (2021) cannot be weighted with the number of transactions, so I only plot the other estimators.

3.2 Within County Triple Difference and Falsification Tests

I next address if there are unobservable factors that affect both algorithmic investor activity and the timing of digitization. For instance, county officials might be working to attract business investment and modernize government processes. To investigate this, I leverage house-level variation in the cost of algorithmically valuing houses to test for evidence of unobserved shocks.

The timing of house-level digitization is not related to house or neighborhood attributes. Appendix Table A.2 reveals that the early and late digitized houses are evenly matched in features such as the number of bedrooms and bathrooms and the presence of a basement or other structures.[29] While houses that are digitized later show a more recent last sale date, newly constructed houses are also digitized later such that there are no substantial disparities. Appendix Table A.3 shows that houses are also similar on neighborhood characteristics.

To investigate common county shocks, I compare digitized and not-yet-digitized houses within the same county before and after digitization. Panel A of Figure 4 plots the raw data, showing the number of houses bought by algorithmic investors separately for digitized and non-digitized houses. Algorithmic investors mostly purchase houses that exist in the county's database. These investors buy very small numbers of non-digitized houses. These outcomes could potentially be attributed to measurement errors in county record keeping, misclassification of algorithmic investors, transactions involving the purchase of multiple houses, or scenarios where algorithmic investors supplement county databases with additional data.

Panel B of Figure 4 displays the corresponding interaction-weighted event study at the county level separately for digitized and non-digitized houses. This panel illustrates that the increase in the number of homes purchased by algorithmic investors is almost entirely confined to houses with digital records. Unobserved county-level shocks, such as changes in housing or foreclosure policy, should impact all houses in a county, regardless of digitization status. County shocks are not consistent with an impact that is so concentrated in digitized houses.

I perform a falsification exercise to assess if county digitization impacts nondigitized houses after adjusting for house-level characteristics, using the regression in Equation 4 run at the census block group level. In the equation, β^{Digit} captures the impact of market digitization on algorithmic investor purchases of digitized houses. $\beta^{NoDigit}$ measures the impacts on non-digitized houses. I

[29] For this analysis, "early digitized" refers to those digitized before the county's median digitization year, and "late digitized" were digitized after.

also include controls for neighborhood and house characteristics.

$$1[q_{igt}^{algo} = 1] = \delta_t + \alpha_g + \beta^{Digit} \times \mathbb{1}[HouseDigitized_{ict} = 1] \times CountyDigit_{ct} + \\ \beta^{NoDigit} \times \mathbb{1}[HouseDigitized_{ict} = 0] \times CountyDigit_{ct} + \gamma X_{igt} + \epsilon_{igt} \quad (4)$$

Column 4 of Table 4, shows that county digitization does not impact non-digitized houses; the impact is solely on digitized houses. These results are not consistent with unobserved, neighborhood-level shocks driving investor activity.

However, suppose that the existence in the county database simplifies the house discovery process for *all* investors. Human investors should then also be more likely to buy digitized houses. I test this by examining the effect of house-level digitization on the propensity to purchase by individual investors in column 5 of Table 4. $\beta^{NoDigit} = 0.0017$ and $\beta^{Digit} = -0.0699$. Digitization reduces the probability of human investment purchase and has no impact on non-digitized houses. I interpret these results to show that digitization affects algorithmic investors differently than human investors.

Together, these results build confidence that digitization and changing data availability drive algorithmic activity. First, if algorithmic investors had some influence on the digitization process, early digitized houses might look different from those that are digitized later. Second, if algorithmic investors were not relying on algorithms to purchase houses, we would not expect their purchases to be so heavily concentrated in digitized houses. Third, if localized neighborhood shocks were driving our results, we would expect both digitized and non-digitized houses in the same local area to be impacted in a similar manner. Lastly, I show that the impact of digitization on the likelihood of purchase is specific to algorithmic investors. Thus, the evidence suggests that algorithmic investment is indeed driven by changes in data availability due to digitization.

4 Allocation and Specialization

In this section, I consider the possibility that algorithms and humans have distinct comparative advantages in prediction problems. I begin with a conceptual framework that illustrates the trade-off between humans and machines.

4.1 Conceptual Framework

Houses are characterized by an observable X and an unobserved Z, seen by humans, and a common value component Y. Although the underlying data is multidimensional, I will use two unidimensional variables $x(X) = E[Y|X]$ and $z(X, Z) = E[Y|X, Z] - E[Y|X]$ and $E[y|X, Z] = E[y|x, z] = x + z$.

Human investors generate a prediction using both x and z, but may be biased ($\delta(x, z) \geqslant 0$). Humans may be biased on some houses, but not on others, or may not make systematic errors. For instance, humans seem to overvalue houses with pools and air conditioning during warm weather (Busse et al., 2012).

$$h(x, z) = E[y|x, z] + \delta(x, z)$$

ML-powered algorithms use patterns in data to make predictions. Algorithms look for patterns in thousands of examples, rather than just being limited to their own experience. They are not subject to the same cognitive limitations as humans. For example, algorithms can quantify the specific value of a two-car garage versus a one-car garage, which is likely outside the scope of most humans. Algorithms are also not explicitly impact by factors like warm weather, how they are feeling at the time and prejudices. However, an algorithm cannot see z.

$$m(x) = E[y|x]$$

For any given house, would an algorithm do better or would a human do better at predicting true y? Given a house with true value y, if $|E[y|x] - E[y|x, z]| >> 0$ or $|E[y|x] - y| >> 0$, then the human-accessible private information is important and a human might do better. For instance, some houses are architecturally complex, and where the information available to an algorithm might not capture the house's aesthetic appeal. Or, a house might have a beautiful view of a nearby farm, such that the value of the house and land is higher than would be predicted by an algorithm. However, proximity to a farm can also have several drawbacks: loud mechanical noises, smell of manure, proximity to pesticides, and a higher than usual number of rodents. These factors may lower the house value compared to what would otherwise be predicted by an algorithm. A human can also walk through the interior of a house, estimating repairs and maintenance costs. However, if $\delta(x, z) > 0$, or human decision making is systematically biased, the value of an algorithm might outweigh the importance of private information.

$$m(x) - h(x, z) = \underbrace{E[y|x] - E[y|x, z]}_{\text{informational advantage, } \mu} - \underbrace{\delta(x, z)}_{\text{human bias, } \delta} \tag{5}$$

Equation 5 highlights the tradeoff between human and algorithmic valuation. if private information z is important, the human informational advantage can outweigh human error. If $\delta(x, z) = 0$ or humans are not biased, humans will do better. If humans make mistakes, the benefits of an algorithm can outweigh the importance of private information.

4.2 Measuring House Predictability

Before showing how human investors respond to digitization, I categorize houses by their degree of algorithmic predictability. I construct this measure of how difficult it is to algorithmically predict each house from commonly available observables. I refer to the difficulty of predicting a house from observables as *predictability.*

To construct this, I use the Extreme Gradient Boosting (XGBoost) algorithm predicting the transaction price (Chen and Guestrin, 2016). Given the high-dimensionality of the data and significance of nonlinear relationships, nonparametric models outperform linear models when modeling houses. For example, even a slight increase in square footage could have a significant impact on price in a densely populated neighborhood, while the same would not be true in a rural area. In these cases, nonparametric models, such as tree-based algorithms, are able to capture nuanced, nonlinear relationships, particularly among the many variables that can influence house pricing – location, size, design, age, local amenities, etc.

The XGBoost algorithm operates on a gradient boosting framework in which new models are generated to correct the errors of pre-existing ones. In essence, it creates a robust overall model by combining multiple weak models to improve the accuracy of the prediction according to the regularized objective shown in Equation 6. l is the differentiable convex loss function, T is the number of leaves in each tree and w is the leaf weights (Chen and Guestrin, 2016). Intuitively, this objective function balances training loss $l(\hat{y}_i, y_i)$ with L1 regularization (γT) and L2 regularization ($\lambda||w||^2$) components, encouraging both simpler and more generalizable models.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \gamma T + \frac{1}{2}\lambda||w||^2 \tag{6}$$

The model is built using pre-digitization data for each county to exclude any impacts from algorithmic investors. I randomly split the data into a training set and 25% held-out test set. Using the training data set, I perform a grid search through the XGBoost hyperparameter space, using 3-fold cross validation with early stopping (Shen et al., 2022; LaValle et al., 2004).

In panel A of Figure 5, I plot the predicted versus actual log price for the held-out sample. The average out-of-sample root mean squared error is 0.903. More intepretably, 40% of the houses in the test set are within 10% of the price. The same measures computed for Zillow's Zestimate, which incorporates demand information from user interactions with their website, reveal that 59% of houses are priced within 10% of the sales price in areas with Zillow coverage (Zillow, 2023).

For each house, I calculate the out-of-sample average—the difference between the actual and predicted price—to capture how easy or hard it is to predict each house. The variation in prediction error is enormous, with the average model varying widely: for some houses, it is close to 50%, while for others it is close to zero. Houses in neighborhoods built by the same builder and in the same style are easier to model, but older houses, built before the introduction of modern building codes and full of architectural distinction, are much less standardized. Features such as sentimental, aesthetic, or historical value or proximity to a noisy highway or pungent agricultural property may also play a role. If qualities that are hard to quantify empirically significantly influence a house's value, algorithmic prediction will be less accurate.[30]

Counties also differ in institutional processes for the collection and quality of their housing data. Counties vary in the frequency with which they update their housing records, the quality of their data control, and the thoroughness of the information they collect on each home or transaction. Together, the less informative or accurate the observable information, the more important human private information becomes.

4.3 Human Investor Shift Towards Hard to Predict Houses

Human investors purchase houses across the entire distribution of model error; in some instances, the predicted price is far lower than the actual price, while in other cases the predicted price significantly exceeds the actual price. In Panel A of Figure 5, I show the predicted versus actual prices of the gradient-boosted tree model described in Section 4.2. In general, the model is best at predicting houses in the middle of the distribution. In Panel B of Figure 5, I plot the actual price versus the predicted price for houses in a held-out test set from 2012 and 2013. The houses are colored in light blue if these houses will be purchased by human investors in the future, and houses in purple are those that will be purchased by algorithmic investors. Human investors purchase houses across the entire distribution of model error; in some instances, the predicted price is far lower than the actual price, while in other cases the predicted price significantly exceeds the actual price. While some of these may reflect poor human decision-making, on average, large differences between predicted and actual prices may reflect private information. Unlike human investors, algorithmic investors only purchase houses where the model-predicted price closely approximates the actual price. In other words, they buy houses where the scope for adverse selection is small.

In Figure 6, I show how human investors react to digitization. Human investors become 50% less likely to buy houses in the lowest decile of model error, where algorithms are most effective. They become almost twice as likely to buy houses that are most difficult to predict.[31] Human investors become less likely to purchase houses in in deciles 1 to decile 7, and more likely to purchase houses

[30] Price someone is willing to pay may also depend on mood, weather, or noise.

[31] In the pre period, likelihood of purchase by an human investor is .12.

in the top two deciles of average model error. These results are consistent with human investors specializing where human comparative advantage is highest.

4.4 Discontinuities: Data errors, zoning rules and lead paint

A testable implication of my conceptual framework is that characteristics that increase the importance of private information, should limit algorithmic investor buying. I investigate this with three specific examples and show that algorithmic investors avoid houses where institutional factors limit algorithm accuracy while human investors do not. These results provide additional evidence that algorithmic investors depend on quantifiable information in the dataset while humans are not so constrained. These results are further evidence for patterns of distinct comparative advantage.

4.4.1 Zoning Rules

I illustrate this first with unusual zoning rules in Wilson County, Tennessee. In Panel A of Figure 5, there is a distinct group of houses in Wilson County where the predicted model price is much higher than the actual price. This is a result of unusual zoning rules for bedrooms in Wilson County, which make it difficult to interpret the number of bedrooms in county data. Wilson County only considered a room a legally zoned bedroom if the room also included a specific type of closet. As a result, tax assessor records list most houses as having zero bedrooms, although the "true" number of bedrooms is higher. Algorithms cannot accurately value houses without access to the true number of bedrooms, and the average algorithm error is large in this county. Although Wilson County is close to Nashville and similar to other places where algorithms buy many houses, algorithmic investment in the county is limited. However, human investors invest heavily.

4.4.2 Lead Paint

Houses constructed before lead paint was banned are more difficult to value algorithmically. In the early 1900s, lead was a commonly used additive in paint and other building materials. During the 1960s and 1970s, detailed studies on the effects of lead poisoning led to concerns about health effects in residential structures. The Consumer Product Safety Commission banned lead paint in residential construction in 1978 (The Department of Housing and Urban Development, 2023). Houses built before 1978 may have lead paint, whereas those built after 1978 do not have lead paint. Human investors, who can physically inspect houses, can determine whether lead is a concern and accurately forecast the additional costs necessary to deal with lead exposure, regardless of when a house was built. Algorithmic investors face uncertainty about lead exposure and face unpredictably higher construction costs to deal with lead exposure in houses built before 1978.[32]

In panel A of Figure 7, I test for a discontinuity in the density of houses bought by algorithmic investors, using a local polynomial density estimator (Cattaneo et al., 2018, 2019). As seen visually in Panel A, the null hypothesis of no discontinuity around 1979 is rejected, with a p-value 0.000. In Panel B, I plot the density of houses purchased by human investors. In this case, with a p-value of 0.295, the null hypothesis that the density shows no evidence of manipulation cannot be rejected. I interpret these results to suggest that algorithmic investors appear to respond to the imposition of lead paint, while human investors do not.

4.4.3 Data Errors

Data entry errors limit the effectiveness of algorithms. There are almost 220,000 houses in the county database dataset with more than 15 bedrooms or 15 bathrooms. These houses reflect data entry errors. Algorithms struggle to accurately value houses with data entry errors because the number of bedrooms and bathrooms is such a crucial piece of information.[33] Panel C of Figure 7 shows the number of houses with data errors sold over time; the series is relatively spiky but without any clear trends, indicating that the availability of houses with data errors does not strongly vary over time. Panel D of Figure 7 shows the natural log transformation of the number of houses with data errors purchased by human and algorithmic investors. Algorithmic investors avoid purchasing houses with data errors, whereas human investors do not. Data entry errors do not pose problems for human investors who do not rely exclusively on hard information.

In all of these instances, institutional details produce variation in the value of private information and create opportunities for human investors. This highlights how institutional details can shape the effectiveness of algorithms use and create opportunities for human judgment.

4.5 Algorithmic Investors Specialize in Minority-Owned Homes

After illustrating human comparative advantage and where human investors focus their efforts after digitization, I now turn to algorithmic investors. I first establish the existence and robustness of a race penalty, suggesting the possibility of human bias, and then show that algorithmic investors disproportionately buy minority-owned homes.

[32] Lead-paint remains the most significant source of lead exposure in the US because many houses were built before 1978 (US EPA, 2014). Any renovation, repair or painting project in a pre-1978 home can easily create dangerous lead dust, requiring special lead-safe contracting procedures and contractors (US EPA, 2013).

[33] In principle, they could collect this information manually, but algorithmic firms are not organizationally set up to do this.

4.5.1 The Race Penalty

Prior to county digitization, I calculate a 5% race penalty—lower sale price received by a minority homeowner compared to a White homeowner—accounting for house and neighborhood characteristics (Elster and Zussman, 2022; Perry et al., 2018; Perry, 2021; Bayer et al., 2017; Kim, 2000; Freddie Mac Economic & Housing Research, 2021; Quillian et al., 2020; Zhang and Leonard, 2021; Kermani and Wong, 2021b; Lewis et al., 2011). In Figure 8, I plot the race penalty in period *prior* to digitization, controlling for various levels of neighborhood characteristics and observable characteristics of the house. All specifications include year and geography fixed effects.[34]

The first bar in Figure 8 shows that minority homes sell at a 14% discount relative to White homes in the same county, adjusting for house characteristics. This gap drops to 7% when adding census tract fixed effects. The 50% decrease in the race penalty suggests that there is a lot of unobserved heterogeneity between houses in the same county.[35] At the census block group level, the implied race penalty is 5%. At the census block level, minority-owned homes sell for about 3% or $4,700 less. All of these numbers are calculated in the years before digitization.

4.5.2 Deep Learning Image Analysis

House characteristics do not fully capture many differences between houses. Houses are structurally unique three-dimensional objects that derive their value from the color of the paint, the landscape, the maintenance, and the cleanliness of the windows. Two houses in the same neighborhood may have completely different architectural styles or states of disrepair. (Pinto and Peter, 2021; Harris, 1999; Choi et al., 2019). Minority homeowners are less wealthy and may invest less in house maintenance and aesthetics (Harris, 1999). The race penalty could simply reflect these differences in house appearance.

I use a deep learning model to calculate the race penalty adjusting for house images. I scraped house images from Zillow and other apartment rental websites. Appendix Figure A.6 shows an example house image. Images are not available for all houses in my sample. I rely on images for a total of 50,000 houses. I use AutoGluon, a deep learning model designed for unstructured data such as images, to convert each exterior image into a high-dimensional embedding vector (Erickson et al., 2020). The position of each image within this vector space is indicative of its visual features or content, ensuring that similar images are close to each other in the embedding space. Adding these

[34] Include year by geography fixed effects yields very similar results.

[35] In our sample, census tracts encompass $4,517$ people or $2,006$ housing units. Census block groups average around $1,610$ people in our sample or 716 housing units. A census block contains around 65 people. I used the 2010 population to calculate these averages.

deep learning embeddings to the race gap regression will control for previously omitted variables the aesthetic features of the house and the yard.

4.5.3 Race Penalty with Deep Learning

Incorporating house exterior images does not significantly change the race penalty. In Table 5, I show the race penalty coefficients, controlling for the quality and appearance of the house with deep learning-generated embeddings. These race penalty estimates are similar to the estimates from Section 4.5.1. For example, at the census block level, the race penalty is 2.1% with the image emeddings and 3.3% without images, including block by year fixed effects. The existence and persistence of this race penalty suggest that the race penalty coefficient may reflect more than simply differences in house quality. For instance, consistent with other evidence that individuals associate lower values with the same home when they perceive it to be owned by a minority, humans could be undervaluing minority-owned homes (Lilien, 2023).

4.5.4 Algorithmic Investors buy Minority-Owned Homes

Algorithmic investors disproportionately buy minority-owned homes. As shown in Table 6, the impact of digitization of house records is twice as strong for minority homeowners than for White homeowners. In column 1 of Table 6, the impact of digitization on a minority-owned home is twice as large relative to a White-owned home in the same census tract or census block group. However, the impact of digitization is six times as large compared to a White-owned house in the same census block. These results suggest that algorithmic investors do not just focus on minority neighborhoods, rather, they specifically target minority houses.

These results are not simply driven by all investors targeting minority-owned homes. In Appendix Table A.5, I demonstrate the effect of digitization by homeowner race within a sample exclusively consisting of investor transactions (human and algorithm), excluding the owner-occupiers (those buying houses to live in). These regressions will illustrate the impact of digitization on the likelihood of purchase by algorithmic investors compared to human investors. Column 1 of Appendix Table A.5 reveals that the effect of digitization on minority-owned homes is five times stronger than on White-owned homes. The impact of digitization is five times as large at the census block group level (column 2), and nine times larger at the census block level (column 3). These results suggest that algorithmic investors are disproportionately likely to buy minority-owned homes, even compared to human investors.

5 Prices, Spillovers and Racial Disparities

In this section, I explore the consequences for market-level prices and racial inequalities.

5.1 Digitization Shrinks the Race Penalty

First, I explore how digitization affects the race penalty. Panel A of Figure 9 plots the race penalty coefficient by time to digitization, including census block group controls.[36] In the year following digitization, the race penalty shrinks from 8% to 4%. The race penalty continues to fall until it disappeared six years after digitization.

In Panel B of Figure 9, I investigate the mechanism behind this change. The first two bars in panel B of Figure 9 plot the race penalty for purchases of owner-occupiers, those who buy houses to live in, and human investors, prior to digitization. Both pay approximately 5% less for the observably similar house in the same census block group with a minority owner compared to a White homeowner. However, as shown in the blue bar, algorithmic investors do not exhibit any race penalty. Algorithmic investors pay the same price for an observably similar house regardless of the race of the homeowner.

Interestingly, digitization also reduces the race penalty among owner-occupiers and human investors. The fourth bar in Panel B of Figure 9 shows that, after digitization, owner occupiers pay only about 3% less for minority-owned homes. The last bar in Panel B of Figure 9 plots the post digitization race penalty for human investors. After digitization, human investors pay only about 1.5% less for minority-owned homes. In Appendix Table A.6, I use a 2SLS analysis is used to address endogeneity concerns around other factors related to bidding behavior that could drive these results. The results are much noisier, but qualitatively similar.

These indirect effects are driven by two mechanisms; price anchoring given higher prices of similar houses and algorithmic investors bidding up the prices of minority homes, even in cases where they do not ultimately acquire the house. Real estate agents set listing prices based on similar, recently sold properties. If properties in minority neighborhoods are priced higher, other houses will have higher listing prices, which subsequently leads to higher sales prices. Furthermore, when algorithmic investors try to buy homes, they can drive up the final price for other buyers. In fact, the share of owner-occupiers is so large that the majority of the impact of digitization on the race penalty is due to these indirect effects.

As a result, Figure 10 shows that digitization leads to a 5% increase in the prices of minority homes. Among White homeowners, it is possible that algorithms may not raise prices. If home-

[36] Digitization varies at the county by year level, so we cannot include geography by year controls.

owners are willing to sell their homes at a discount in exchange for a prompt offer, could lead to a decline in prices. Instead, we also see an increase; digitization leads to a 1.5% increase in the average sale price of White-owned homes.

5.2 Adverse Selection or Human Error?

A natural question is whether algorithms are taking advantage of human mistakes or simply overpaying for unobservably worse homes. Adverse selection has been widely cited as a barrier to the use of algorithms in the housing market and has been widely discussed as the reason why Zillow, an algorithmic investor, decided to stop buying houses (Economist, 2021).

I disentangle adverse selection from human error with two complementary approaches. In the first, I calculate the gross margin on each house sold: the difference between the resale and transaction price. If algorithmic firms overvalue minority homes relative to White-owned homes, then the gross margin on minority homes should be lower compared to White-owned homes. I also calculate the gross margin with the estimates of the house market value from tax assessors. Unlike resale price, which is only available for resold homes, these estimates are available for all homes. However, these are estimates made by the human tax assessor rather than actual transaction prices.[37]

Using my two measures of gross margin, I estimate the following regression for house i bought in year t, resold/assessed in year r in census block c, including purchase year by census tract, block group or block and resale or assessment year fixed effects:[38]

$$log(price^{resale}_{irc}) - log(price^{sale}_{itc}) = \gamma X_{irtc} + \beta_1^{algo} \times SellerMinority_{itrc} + \epsilon_{itrc} \tag{7}$$

The coefficient β_1 indicates if the margin is systematically different for minority homeowners. If houses purchased from minority homeowners are adversely selected, the margin should be lower, or $\beta_1^{algo} < 0$. However, if the higher prices paid by algorithmic investors for minority homes reflect the true value, then $\beta_1^{algo} \approx 0$ or $\beta_1^{algo} > 0$.

I find no significant differences in the gross margin by race of the homeowner. Columns 1 through 3 of Table 7 include census tract, block group, and block by year fixed effects, respectively, among houses bought by algorithmic investors.[39] Columns 1 through 3 of Appendix Table A.7 show similar results using the assessment margin. Among houses purchased by algorithmic investors, the margin on minority-owned homes is not systematically different from that on White-owned homes. In column 4 of Table 7 and column 4 of Appendix Table A.7, I explore whether the resale margin

[37] Note that if tax assessor evaluations are also biased, our results will be more conservative.

[38] The time between purchase and resale or assessment is a linear combination of purchase and resale year, so this would drop from any regression.

[39] Not all houses can be geocoded to a census Block level, but I show all three results.

differs for homes bought in neighborhoods with greater minority shares. In both cases, I find no strong relationship. These results suggest that the gross margin of the algorithmic investor does not vary with the composition of the neighborhood.

Next, I show that the gross margin on minority-owned investors is *higher* that that on White homes. If minority-owned homes are priced too low, the gross margin should be higher due to the discounted acquisition price. Columns 5 through 7 in Table 7 show that the gross margin on minority homes is 10% higher among purchases by human investors. In column 8, I explore whether the margin varies by neighborhood composition. The margin may be higher in more minority neighborhoods that contain a higher share of minority residents, but the estimate is noisy. Column 8 of Appendix Table A.7 shows that the assessment margin is 9% higher in neighborhoods with a higher share of minority residents.

If these results are due to racial preferences or heuristics, the differences may be more pronounced in neighborhoods with more minority residents, where humans may have more trouble accurately valuing houses or biases. These results suggest that the higher prices algorithmic investors pay for minority-owned homes are not driven by adverse selection and may, in fact, reflect algorithm comparative advantage in valuing houses where human biases, prejudices, or cognitive limitations may cloud judgment.

These results are also not consistent with adverse selection among algorithmic firms. If minority-owned houses are unobservably worse, humans should not be willing to pay more for these houses after digitization. Humans can access unosbervable aspects of house quality that are not apparent to algorithms and should not be subject to the same adverse selection concerns.

5.3 Spillovers

Thus far, my analysis has focused on the prices of sold homes. However, house values of occupied houses are a key driver of household wealth and play an important role in credit markets (Guren et al., 2020). If minority-owned houses purchased by algorithmic investors are predominantly located in majority White areas and not structurally similar to other minority-owned houses, algorithmic investor activity would not necessarily have spillover effects on unsold minority homes. However, if algorithmic purchases are similar to other minority-owned unsold houses, their activity could also have large indirect impacts on minority homeowner house values and household wealth.

Following Hirano et al. (2003), using the estimate of the expected price of sold minority homes, $E[P|S = 1]$, I write the inverse propensity weighted unsold minority-owned homes house price impact $E[P|U=1]$ as:

$$\begin{aligned}
E[P|U=1] &= \sum_X p(X|U=1)E[P|U=1,X] \\
&= \sum_X \frac{p(U=1|X)p(X)}{p(U=1)}E[P|U=1,X] \\
&= \frac{1}{p(U=1)}\sum_X p(U=1|X)p(X)E[P|U=1,X]\frac{p(X|S=1)p(S=1)}{p(S=1|X)p(X)} \\
&= \frac{p(S=1)}{p(U=1)}\sum_X E[P|S=1,X]\frac{p(U=1|X)p(X|S=1)}{p(S=1|X)} \\
&= \frac{p(S=1)}{p(U=1)}E\left[\frac{p(U=1|X)}{p(S=1|X)}P|S=1\right] \qquad (8)
\end{aligned}$$

Equation (8) says that I can recover the average impact on the price of minority-owned homes by reweighting the prices of sold minority-owned homes, using a ratio of propensity scores to account for differences in house characteristics. Re-weighting based on the characteristics of the observable houses and the census block, I find an average increase of 6% in the value of minority homes. It is important to emphasize that this analysis relies on a selection on observables assumption when reweighting. Although algorithmic investors may not be able to see unobservable characteristics, part of the impact comes from human investors and owner-occupiers, who have access to unobserved information. If unsold minority houses are very different on unobservables than sold minority houses, this estimate may overstate the impacts.

6 Conclusion

Progress in ML and the widespread availability of digitized data opens up a wide set of economic possibilities. This work illustrates how the availability of algorithmic prediction not only influences individual decisions, but also precipitates a range of changes at the market level, affecting participation, firm organization, and equilibrium outcomes. In the housing market, the availability of machine-generated predictions leads to new entrants using algorithms to value houses. Human investors react by moving towards parts of the market where algorithms are least effective. Algorithmic investors buy disproportionately where human decisions are biased, causing large price increases. Six years after digitization, the race penalty disappears. Much of these impacts are due to indirect effects of algorithmic investors that manifest through the nature of competition. These findings suggest numerous avenues for future research.

First, when algorithms and humans disagree, we cannot assume that the algorithm is correct: unobserved information can lead to algorithm errors. At the same time, we cannot assume that the human is always correct. Instead, the value of the tradeoff depends on the importance of private information and degree of bias in human decisions. A growing number of papers show that human errors can be sufficiently systematic to outweigh the value of private information (Kleinberg et al., 2017b; Mullainathan and Obermeyer, 2019; Rambachan, 2022; Kahneman et al., 2021). More work is needed to better understand how the value of this tradeoff varies across people and prediction problems.

Second, as an evolving technology, the ML tools used by companies employing algorithmic prediction strategies are rapidly changing. In this setting, there is an ecosystem of companies attempting to curate detailed and increasingly accurate datasets, from comprehensive house surveys that measure construction quality to mobile phone data that track neighborhood activities. As data quality improves, the percentage of "predictable" houses that can be targeted for purchase by algorithmic investors may increase.

Furthermore, the efficacy of algorithmic prediction may depend on the specific legal and institutional structures, which vary widely between states. This study examines Georgia, North Carolina, South Carolina, and Tennessee, where housing market transactions and prices are part of the public record. However, in twelve US states, property transaction prices are not automatically included in the public record, thus potentially curtailing the effectiveness of algorithmic prediction. More work is needed to explore how institutional structures impact the potential effects of algorithms.

Finally, this study does not dwell on the potential implications of organizational differences between algorithmic and human investors. Algorithmic investors typically operate as large, formal, arms-length organizations, while human investors often manage their rental properties more informally. Unlike human investors, who frequently choose tenants personally or through their social networks, algorithmic firms may rely more heavily on automated screening procedures for tenant selection. Together, these changes could have lasting effects on the local labor market and communities. Given the rapid adoption of algorithms, these effects deserve further study.

References

Abaluck, Jason, Leila Agha, David C. Chan Jr, Daniel Singer, and Diana Zhu, "Fixing Misallocation with Guidelines: Awareness vs. Adherence," Technical Report, National Bureau of Economic Research 2020.

Acemoglu, Daron, Gary Anderson, David Beede, Catherine Buffington, Eric Childress, Emin Dinlersoz, Lucia Foster, Nathan Goldschlag, John Haltiwanger, Zachary Kroff, Pascual Restrepo, and Nikolas Zolas, "Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey," 2022.

Aggarwal, Nidhi and Susan Thomas, "The causal impact of algorithmic trading on market quality," in "in" 2014.

American Homes 4 Rent, "Form S-11," https://www.sec.gov/Archives/edgar/data/1562401/00011931251324714 2013.

_ , "10K-2017-Q4.Pdf," https://s29.q4cdn.com/671712101/files/doc_financial/quarterly_results/2017/q4/10K-2017-Q4.pdf 2018.

Amherst, "U.S. Single Family Rental - An Emerging Institutional Asset Class," https://www.amherst.com/insights/u-s-single-family-rental-an-emerging-institutional-asset-class/ November 2016.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

Arnold, David, Will Dobbie, and Crystal S Yang, "Racial Bias in Bail Decisions," *The Quarterly Journal of Economics*, 2018, p. qjy012.

Autor, David H and David Scarborough, "Does job testing harm minority workers? Evidence from retail establishments," *The Quarterly Journal of Economics*, 2008, *123* (1), 219–277.

Babcock, Frederick M., *The Valuation of Real Estate*, New York: McGraw Hill Book Company, 1932.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation," May 2022.

Bakalar, Chloé, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao, "Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems," 2021.

Bayer, Patrick, Marcus Casey, Fernando Ferreira, and Robert McMillan, "Racial and Ethnic Price Differentials in the Housing Market," *Journal of Urban Economics*, November 2017, *102*, 91–105.

Becker, Gary S. (Gary Stanley), *The Economics of Discrimination.* Studies in Economics of the Economics Research Center of the University of Chicago, Chicago: University of Chicago Press, 1957.

Bhutta, Neil, Andrew C. Chang, Lisa J. Dettling, and Joanne W. Hsu with assistance from Julia Hewitt, "Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances," *FEDS Notes*, September 2020.

Blattner, Laura and Scott Nelson, "How Costly Is Noise? Data and Disparities in Consumer Credit," May 2021. Comment: 86 pages, 17 figures.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," 2022.

Brown, Zach Y. and Alexander MacKay, "Competition in Pricing Algorithms," *American Economic Journal: Microeconomics*, May 2023, *15* (2), 109–56.

Brynjolfsson, Erik, Lindsey Raymond, and Danielle Li, "Generative AI at Work," *NBER Working Paper No. 31161*, 2023.

Bubeck, Sebastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru, "Why Is Intermediating Houses so Difficult? Evidence from iBuyers," July 2022.

Busse, Meghan R, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso, "Projection Bias in the Car and Housing Markets," Working Paper 18212, National Bureau of Economic Research July 2012.

Calder-Wang, Sophie and Gi Heung Kim, "Coordinated vs Efficient Prices: The Impact of Algorithmic Pricing on Multifamily Rental Markets," July 2023.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, December 2021, *225* (2), 200–230.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello, "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, October 2020, *110* (10), 3267–97.

Calvino, Flavio and Luca Fontanelli, "A Portrait of AI Adopters across Countries: Firm Characteristics, Assets' Complementarities and Productivity," Technical Report, OECD, Paris April 2023.

Casey, Joan A., Frosch Rachel Morello, Daniel J. Mennitt, Kurt Fristrup, Elizabeth L. Ogburn, and Peter James, "Race/Ethnicity, Socioeconomic Status, Residential Segregation, and Spatial Variation in Noise Exposure in the Contiguous United States," *Environmental Health Perspectives*, 2017, *125* (7), 077017.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, "Manipulation Testing Based on Density Discontinuity," *The Stata Journal: Promoting communications on statistics and Stata*, March 2018, *18* (1), 234–261.

_ , _ , and _ , "Simple Local Polynomial Density Estimators," 2019.

Census, "Census Tabulation Detail: Tenure by Units in Structure," https://censusreporter.org/tables/B25032/ 2023.

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega, "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market," *The Journal of Finance*, 2014, *69* (5), 2045–2084.

Chen, Tianqi and Carlos Guestrin, "XGBoost," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM aug 2016.

Choi, Jonathan H. and Daniel Schwarcz, "AI Assistance in Legal Analysis: An Empirical Study," August 2023.

Choi, Jung Hyun, Caitlin Young, Alanna McCargo, Michael Neal, and Laurie Goodman, "Explaining the Black-White Homeownership Gap," *The Urban Institute*, 2019.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan, "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," in "Proceedings of the 1st Conference on Fairness, Accountability and Transparency" PMLR January 2018, pp. 134–148.

Christophers, Brett, “How and Why U.S. Single-Family Housing Became an Investor Asset Class,” *Journal of Urban History*, 2023, *49* (2), 430–449.

Clark, Robert, Stephanie Assad, Daniel Ershov, and Lei Xu, “Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market,” *Journal of Political Economy*, 2023, *0* (ja), null.

Consumer Financial Protection Bureau, “Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity,” https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/ June 2023.

Cowgill, Bo and Catherine E Tucker, “Economics, fairness and algorithmic bias,” *preparation for: Journal of Economic Perspectives*, 2019.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor, “The Rise and Decline of the American Ghetto,” *Journal of Political Economy*, June 1999, *107* (3), 455–506.

Dawes, Robyn M., “A case study of graduate admissions: Application of three principles of human decision making,” *American Psychologist*, 1971, *26* (2), 180–188.

_ , David Faust, and Paul E. Meehl, “Clinical Versus Actuarial Judgment,” *Science*, 1989, *243* (4899), 1668–1674.

de Chaisemartin, Clément and Xavier D’Haultfœuille, “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects,” *American Economic Review*, September 2020, *110* (9), 2964–96.

Derenoncourt, Ellora, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick, “Wealth of Two Nations: The U.S. Racial Wealth Gap, 1860-2020,” Working Paper 30101, National Bureau of Economic Research June 2022.

Economist, The, “A Whodunnit on Zillow,” *The Economist*, 2021.

Einav, Liran, Mark Jenkins, and Jonathan Levin, “The impact of credit scoring on consumer lending,” *The RAND Journal of Economics*, 2013, *44* (2), 249–274.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie, “Using the Census Bureau’s Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities,” *Health Services and Outcomes Research Methodology*, June 2009, *9* (2), 69–83.

Elster, Yael and Noam Zussman, “Minorities and Property Values: Evidence from Residential Buildings in Israel,” *Journal of Urban Economics*, December 2022, p. 103525.

EPA, OAR US, “Lead’s Impact on Indoor Air Quality,” https://www.epa.gov/indoor-air-quality-iaq/leads-impact-indoor-air-quality August 2014.

EPA, OCSPP US, “Lead Renovation, Repair and Painting Program,” https://www.epa.gov/lead/lead-renovation-repair-and-painting-program February 2013.

Erickson, Nick, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola, “AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data,” 2020.

Fields, Desiree, “Constructing a New Asset Class: Property-led Financial Accumulation after the Crisis,” *Economic Geography*, March 2018, *94* (2), 118–140.

_ , “Automated Landlord: Digital Technologies and Post-Crisis Financial Accumulation,” *Environment and Planning A Economy and Space*, February 2022, *54* (1), 160–181.

Francke, Marc, Lianne Hans, Matthijs Korevaar, and Sjoerd van Bekkum, "Buy-to-Live vs. Buy-to-Let: The Impact of Real Estate Investors on Housing Costs and Neighborhoods," June 2023.

Frankel, Alex, "Selecting Applicants," *Econometrica*, 2021, *89* (2), 615–645.

Freddie Mac Economic & Housing Research, "Single-Family Rental: An Evolving Market," Technical Report, Federal Home Loan Mortgage Corporation 2018.

_ , "Racial and Ethnic Valuation Gaps in Home Purchase Appraisals," Technical Report, Federal Home Loan Mortgage Corporation September 2021.

Fuster, Andres, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Gillis, Talia B and Jann L Spiess, "Big Data and Discrimination," *The University of Chicago Law Review*, 2019, p. 29.

Guren, Adam M, Alisdair McKay, Emi Nakamura, and Jón Steinsson, "Housing Wealth Effects: The Long View," *The Review of Economic Studies*, April 2020, *88* (2), 669–707.

Gurun, Umit G, Jiabin Wu, Steven Chong Xiao, and Serena Wenjing Xiao, "Do Wall Street Landlords Undermine Renters' Welfare?," *The Review of Financial Studies*, January 2023, *36* (1), 70–121.

Harris, Adam and Maggie Yellen, "Human Decision-Making with Machine Prediction: Evidence from Predictive Maintenance in Trucking," 2023.

Harris, David, ""Property Values Drop When Blacks Move in, Because...": Racial and Socioeconomic Determinants of Neighborhood Desirability," *American Sociological Review*, 1999, *64* (3), 461–479.

Hastie, R. and R. M. Dawes, *Rational choice in an uncertain world: The psychology of judgment and decision making*, Sage Publications, 2001.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, "Does Algorithmic Trading Improve Liquidity?," *The Journal of Finance*, 2011, *66* (1), 1–33.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71* (4), 1161–1189. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00442.

Howell, Junia and Elizabeth Korver-Glenn, "Neighborhoods, Race, and the Twenty-first-century Housing Appraisal Industry," *Sociology of Race and Ethnicity*, 2018, *4*, 473 – 490.

Invitation Homes, "2017-Annual-Report.Pdf," https://s28.q4cdn.com/264003623/files/doc_financials/2017/ar/2 Annual-Report.pdf 2017.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment*, first edition ed., New York: Little, Brown Spark, 2021.

Kamin, Debra, "Home Appraised With a Black Owner: $472,000. With a White Owner: $750,000. - The New York Times," *New York Times*, June 2023.

Kaysen, Ronda, "New Legislation Proposes to Take Wall Street Out of the Housing Market," *The New York Times*, December 2023.

Kermani, Amir and Francis Wong, "Racial Disparities in Housing Returns," Working Paper 29306, National Bureau of Economic Research September 2021.

_ **and** _ , "Racial Disparities in Housing Returns," September 2021.

Kim, Sunwoong, "Race and Home Price Appreciation in Urban Neighborhoods: Evidence from Milwaukee, Wisconsin," *The Review of Black Political Economy*, December 2000, *28* (2), 9–28.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*, 08 2017, *133* (1), 237–293.

_ , _ , _ , _ , and _ , "Human decisions and machine predictions*," *The Quarterly Journal of Economics*, August 2017, *133* (1), 237–293. tex.eprint: https://academic.oup.com/qje/article-pdf/133/1/237/30636517/qjx032.pdf.

_ , Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein, "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 2018, *10.*

_ , _ , _ , and Ziad Obermeyer, "Prediction Policy Problems," *American Economic Review*, May 2015, *105* (5), 491–495.

_ , Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv:1609.05807 [cs, stat]*, November 2016. arXiv: 1609.05807.

LaValle, Steven M, Michael S Branicky, and Stephen R Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *The International Journal of Robotics Research*, 2004, *23* (7-8), 673–692.

Lewis, Valerie A., Michael O. Emerson, and Stephen L. Klineberg, "Who We'll Live With: Neighborhood Racial Composition Preferences of Whites, Blacks and Latinos," *Social Forces*, 2011, *89* (4), 1385–1407.

Li, Danielle, Lindsey Raymond, and Peter Bergman, "Hiring as Exploration," *NBER Working Paper No. 27736*, 2020.

Lilien, Jake, "Faulty Foundations: Mystery-Shopper Testing In Home Appraisals Exposes Racial Bias Undermining Black Wealth," 2023.

Malone, Thomas, "Residential Real Estate: Largest US Asset Class but Not Biggest Economic Driver," https://www.corelogic.com/intelligence/why-the-uss-largest-asset-class-residential-real-estate-does-not-substantially-contribute-to-the-economic-output/ June 2023.

Merkley, Mr and Adam Smith, "End Hedge Fund Control of American Homes Act of 2023," 2023.

Mills, James, Raven Molloy, and Rebecca Zarutskie, "Large-Scale Buy-to-Rent Investors in the Single-Family Housing Market: The Emergence of a New Asset Class," *Real Estate Economics*, 2019, *47* (2), 399–430.

Mullainathan, Sendhil and Ashesh Rambachan, "From Predictive Algorithms to Automatic Generation of Anomalies," May 2023.

_ and Ziad Obermeyer, "Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error," *NBER WP*, 2019.

_ and _ , "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care*," *The Quarterly Journal of Economics*, December 2021, *137* (2), 679–727.

Neal, Michael, Laurie Goodman, and Caitlin Young, "Housing Supply Chartbook," *The Urban Institute*, January 2020.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Available at SSRN 4375283*, 2023.

Obermeyer, Ziad and Ezekiel J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *The New England journal of medicine*, September 2016, *375* (13), 1216–1219.

OECD, *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*, Paris: Organisation for Economic Co-operation and Development, 2023.

Peng, Baolin, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao, "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback," 2023.

Perry, Andre, Jonathan Rothwell, and David Harshbarger, "The Devaluation of Assets in Black Neighborhoods," November 2018.

Perry, Jonathan Rothwell and Andre M., "Biased Appraisals and the Devaluation of Housing in Black Neighborhoods," November 2021.

Pinto, Edward and Tobias Peter, "AEI Housing Center Response to Perry and Rothwell (2021)," *American Enterprise Institute Housing Center*, December 2021.

Quillian, Lincoln, John J. Lee, and Brandon Honoré, "Racial Discrimination in the U.S. Housing and Mortgage Lending Markets: A Quantitative Review of Trends, 1976–2016," *Race and Social Problems*, March 2020, *12* (1), 13–28.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy, "Mitigating bias in algorithmic hiring," in "Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency" ACM jan 2020.

Rambachan, Ashesh, "Identifying Prediction Mistakes in Observational Data," 2022, (27111).

_ and Jonathan Roth, "Bias In, Bias Out? Evaluating the Folk Wisdom," 2019.

_ , Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig, "An Economic Approach to Regulating Algorithms," Working Paper 27111, National Bureau of Economic Research May 2020.

Raymond, Elora Lee, Ben Miller, Michaela McKinney, and Jonathan Braun, "Gentrifying Atlanta: Investor Purchases of Rental Housing, Evictions, and the Displacement of Black Residents," *Housing Policy Debate*, September 2021, *31* (3-5), 818–834.

_ , Richard Duckworth, Benjamin Miller, Michael Lucas, and Shiraj Pokharel, "From Foreclosure to Eviction: Housing Insecurity in Corporate-Owned Single-Family Rentals," *Cityscape*, 2018, *20* (3), 159–188.

_ , _ , Benjmain Miller, Michael Lucas, and Shiraj Pokharel, "Corporate Landlords, Institutional Investors, and Displacement: Eviction Rates in Singlefamily Rentals," *FRB Atlanta community and economic development discussion paper*, 2016, (2016-4).

Redfin, "Housing Investors Sell 1 in 7 Homes at a Loss—Highest Share Since 2016," https://www.redfin.com/news/investor-homes-sold-at-a-loss/ 2023.

Redfin, "Investor Home Purchases Fell a Record 49% in the First Quarter," https://www.redfin.com/news/investor-home-purchases-q1-2023/ 2023.

Salzman, Diego and Remco C.J. Zwinkels, "Behavioral Real Estate," *Journal of Real Estate Literature*, January 2017, *25* (1), 77–106.

Shen, Ruoqi, Liyao Gao, and Yi-An Ma, "On Optimal Early Stopping: Over-informative versus Under-informative Parametrization," 2022.

Smith, Julie M., "Algorithms and Bias," *Encyclopedia of Organizational Knowledge, Administration, and Technology*, 2021, pp. 918–932.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Tessum, Christopher W., David A. Paolella, Sarah E. Chambliss, Joshua S. Apte, Jason D. Hill, and Julian D. Marshall, "PM<sub>2.5</sub> polluters disproportionately and systemically affect people of color in the United States," *Science Advances*, 2021, *7* (18), eabf4491.

The Department of Housing and Urban Development, "Legislative History of Lead-Based Paint," 2023.

The White House, "Memorandum on Transparency and Open Government," January 2009.

_ , "The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America," Technical Report, The White House December 2022.

Upson, James and Robert A. Van Ness, "Multiple Markets, Algorithmic Trading, and Market Liquidity," *Journal of Financial Markets*, January 2017, *32*, 49–68.

U.S. Census Bureau, "S2504 Physical Housing Characteristics for Occupied Housing Units," 2021.

Wheaton, David, "Fighting Appraisal Bias: How the Government and Housing Industry Can Better Address This Discriminatory Practice," https://www.naacpldf.org/appraisal-algorithmic-bias/ 2023.

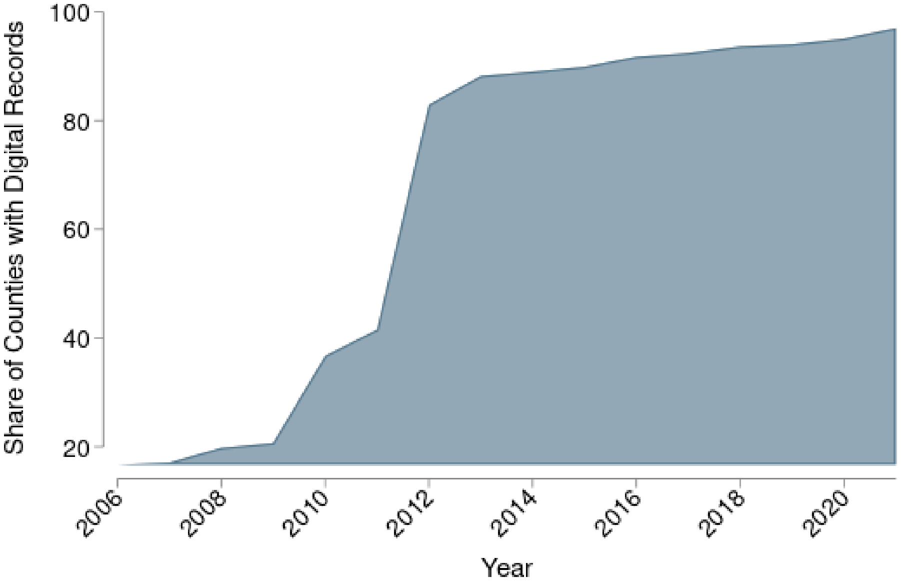
Whittle, Richard, T. Davies, Matthew Gobey, and John Simister, "Behavioural Economics and House Prices: A Literature Review," in "in" 2014.

Zhang, Lei and Tammy Leonard, "External Validity of Hedonic Price Estimates: Heterogeneity in the Price Discount Associated with Having Black and Hispanic Neighbors," *Journal of Regional Science*, 2021, *61* (1), 62–85.

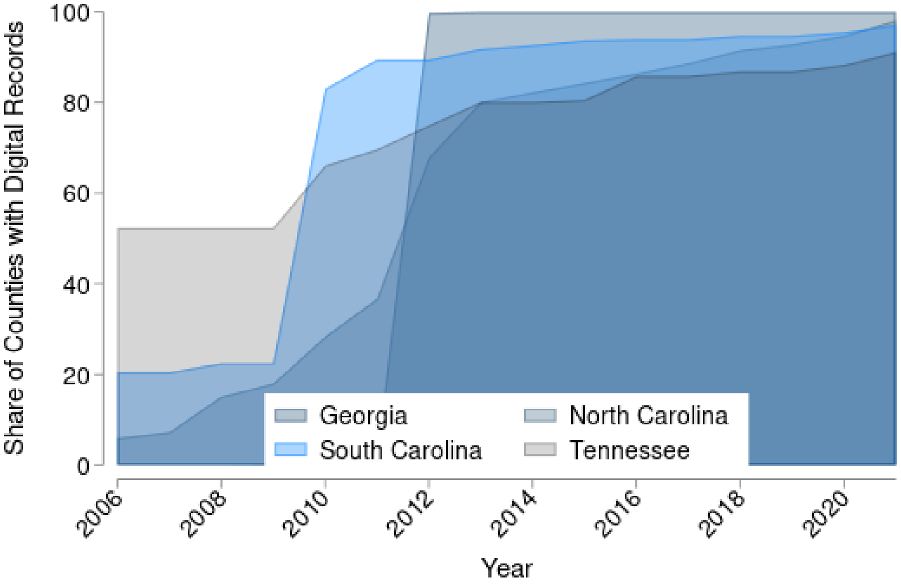
Zillow, "What Is a Zestimate? Zillow's Zestimate Accuracy," 2023.

Figure 1: County Record Digitization

A. Share of Counties with Digitized Records

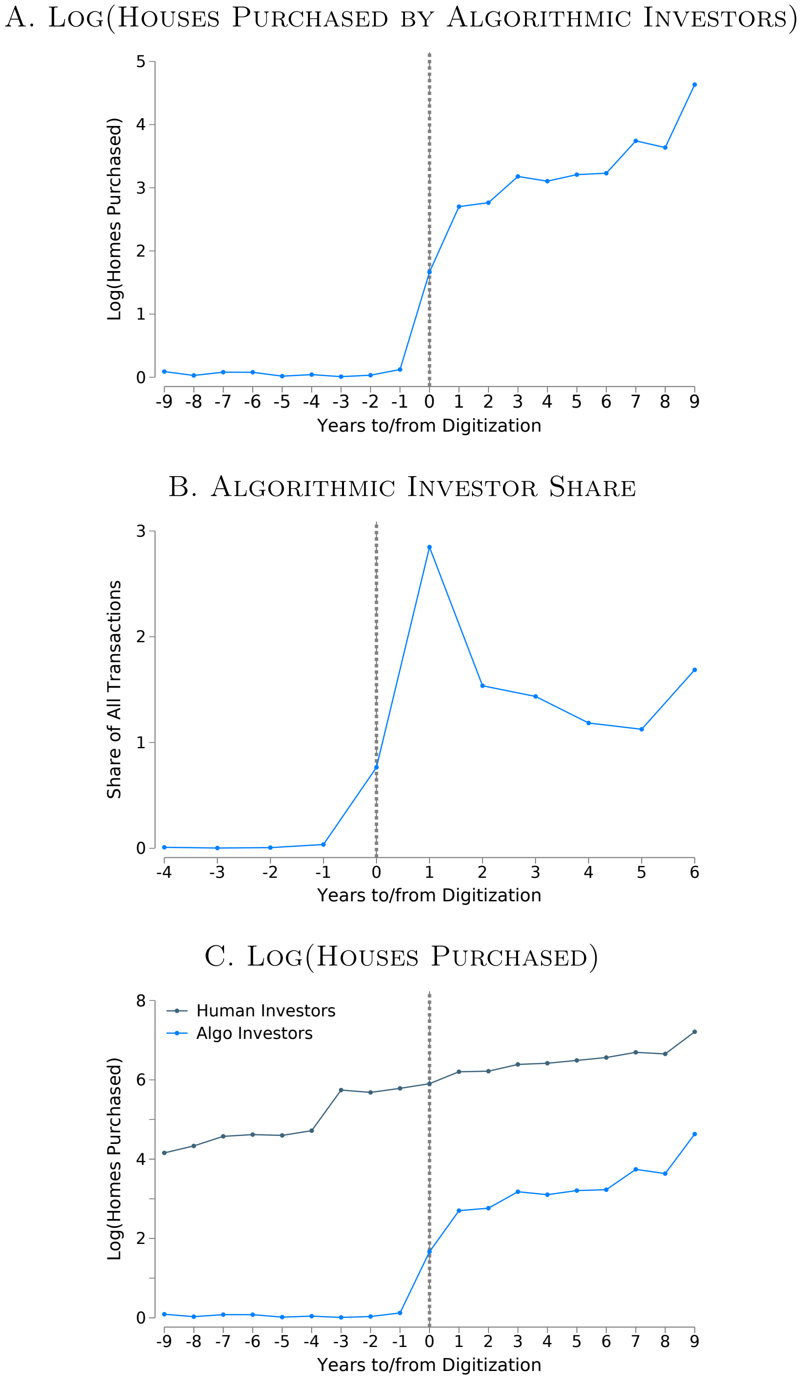


B. Share of Counties with Digitized Records, by State



Notes: This figure shows the share of counties in the sample with digitized and publicly accessible Recorder data over time. Panel B shows the share by state. The graphs are weighted by the number of housing transactions. All data comes from county governments.

Figure 2: Algorithmic investors buying, by time to digitization



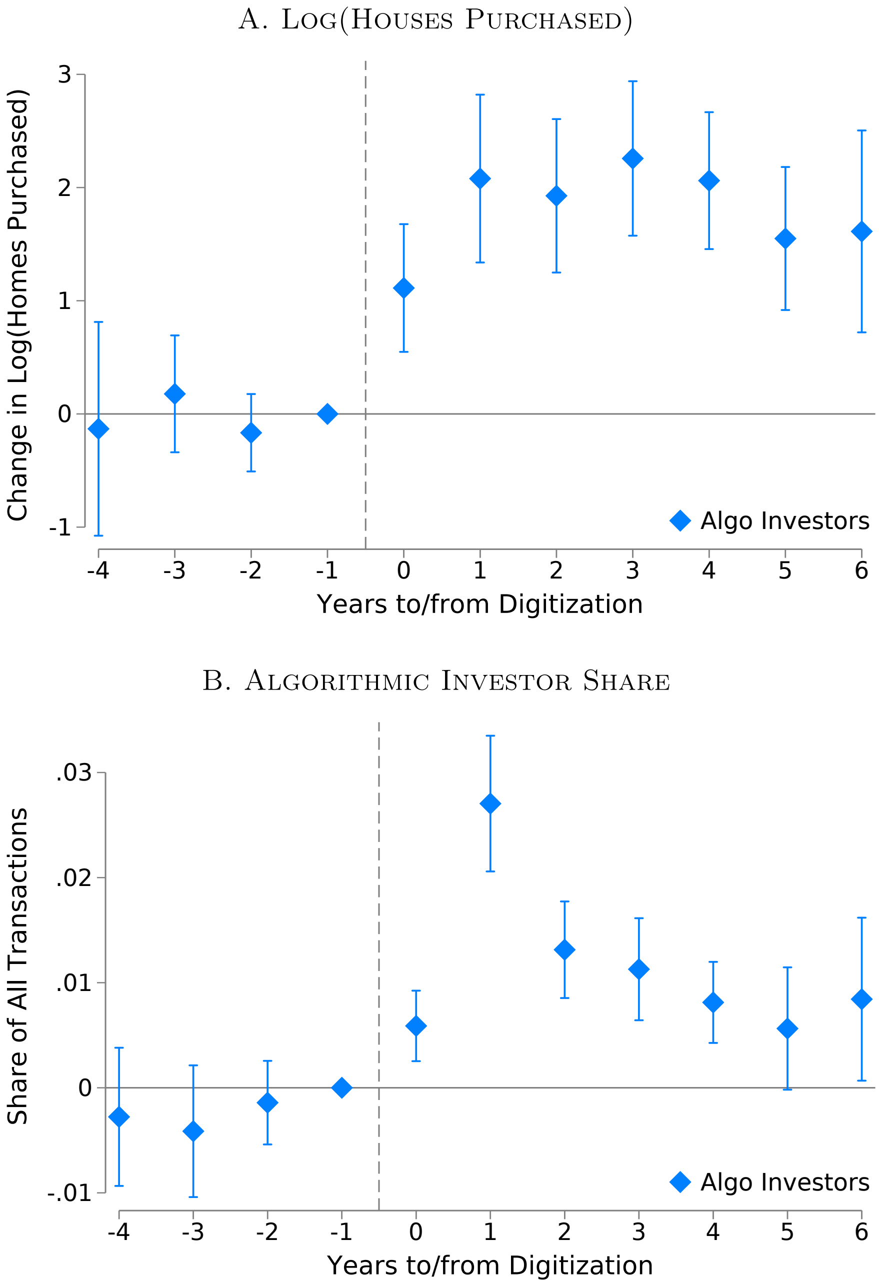
A. Log(Houses Purchased by Algorithmic Investors)

B. Algorithmic Investor Share

C. Log(Houses Purchased)

Notes: This figure shows the number of houses purchased by algorithmic investors in county c and in year t, by time to digitization. Panel A shows the natural log of the number of homes purchased by algorithmic investors. Panel B plots the number of algorithmic investors purchases as a share of all transactions. Panel C adds the natural log of the number of houses purchased by human investors. All data come from ATTOM Data, Zillow and county digitization records.

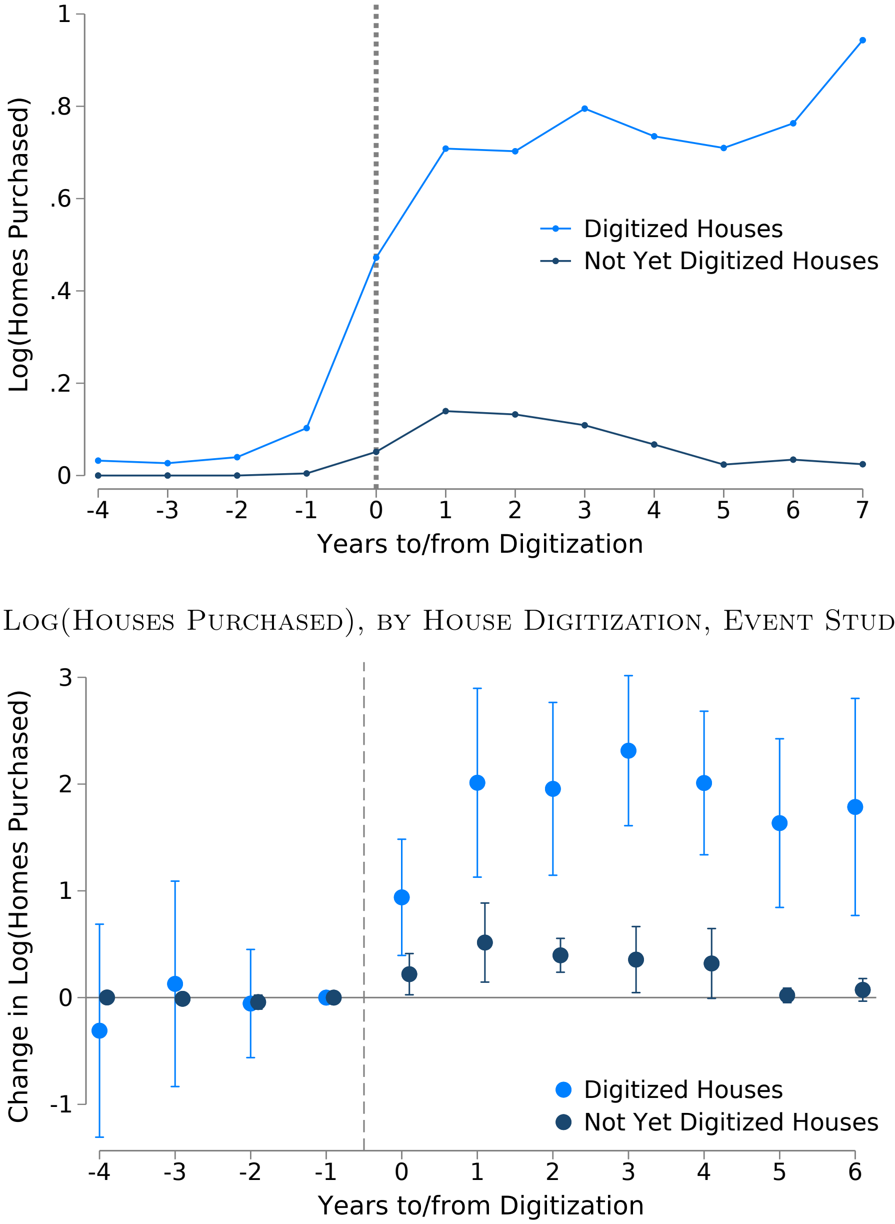
Figure 3: Event Studies, Log(Houses Purchased) by Algorithmic Investors


A. Log(Houses Purchased)

B. Algorithmic Investor Share

Notes: These figures plot the coefficients and 95 percent confidence intervals from Sun and Abraham (2021) interaction-weighted event study regressions of county digitization. Panel A shows the natural log of the number of homes purchased by algorithmic investors. Panel B plots the number of algorithmic investors purchases as a share of all transactions. All specifications include state and year fixed effects, standard errors are clustered at the county level. Regressions are weighted by the number of transactions in each county and year. All data come from ATTOM Data, Zillow and county digitization records.

Figure 4: Houses Purchased by Algorithmic Investors, by House Digitization Status

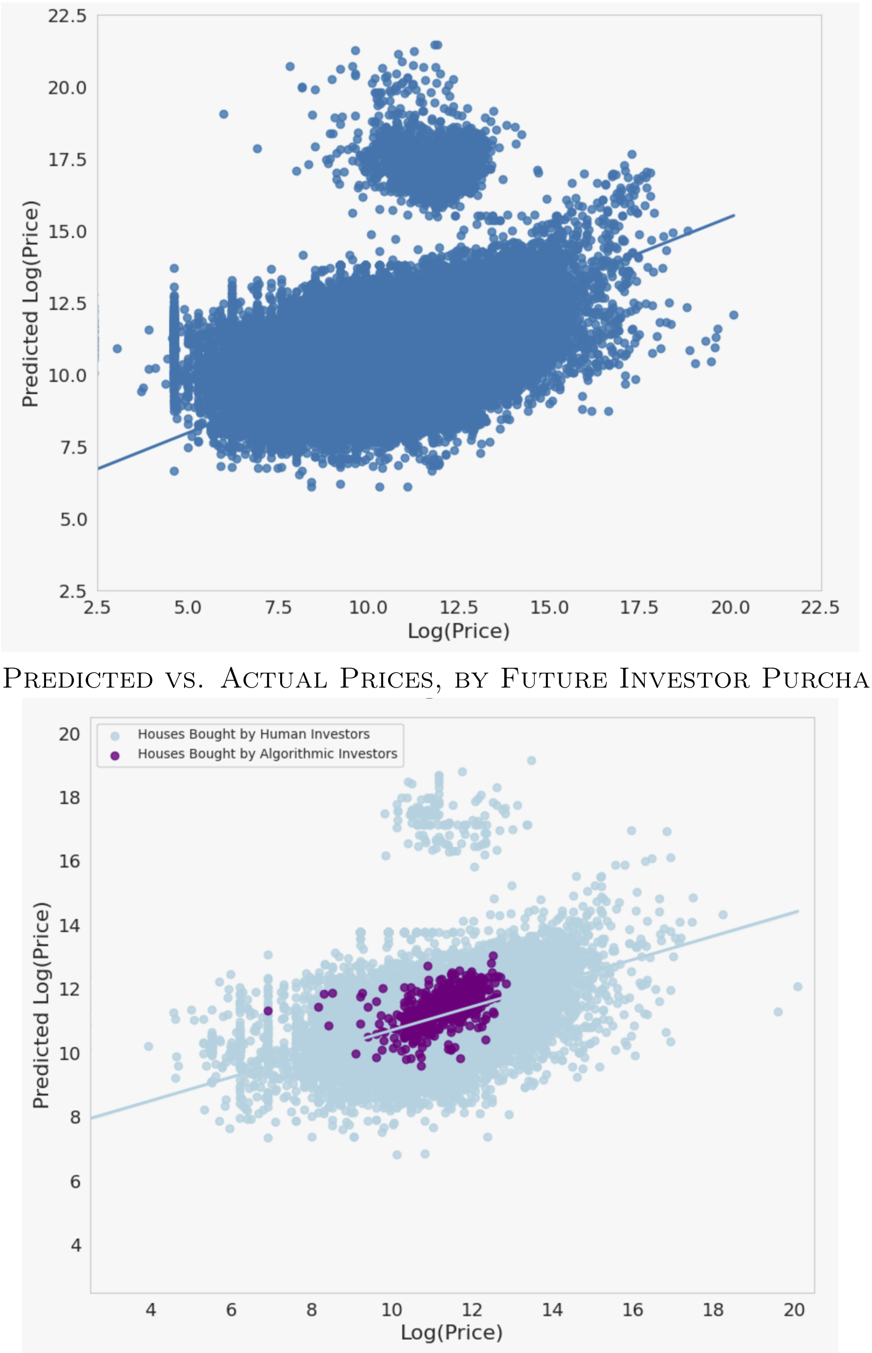


A. Log(Houses Purchased), by House Digitization, Raw Data

B. Log(Houses Purchased), by House Digitization, Event Study

Notes: These figures show the impact of county digitization on the number of homes purchased by algorithmic firms separately estimated for *digitized houses*, houses that have been digitized, and *non-digitized houses*, houses that have not been digitized and only have paper records. Panel A shows the raw natural log of the number of homes purchased by algorithmic firms and Panel B plots the coefficients and 95 percent confidence intervals from Sun and Abraham (2021) interaction-weighted event study regressions. All specifications include state and year fixed effects, standard errors are clustered at the county level and are weighted by the number of transactions. All data come from ATTOM Data, Zillow and county digitization records.

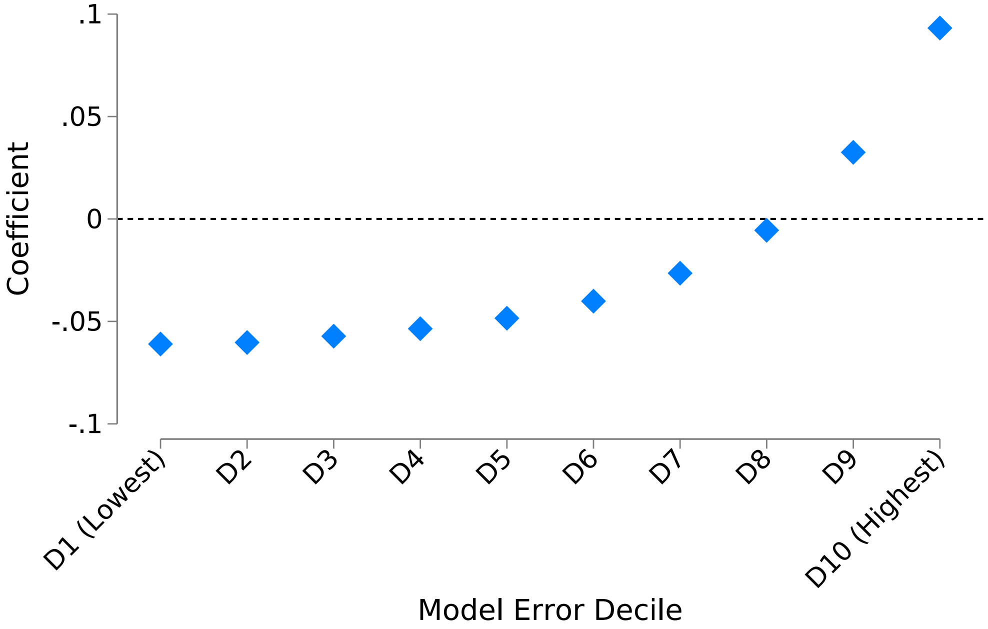
Figure 5: Model Predicted vs. Actual Price


A. Predicted vs. Actual Prices, Out of Sample

B. Predicted vs. Actual Prices, by Future Investor Purchase

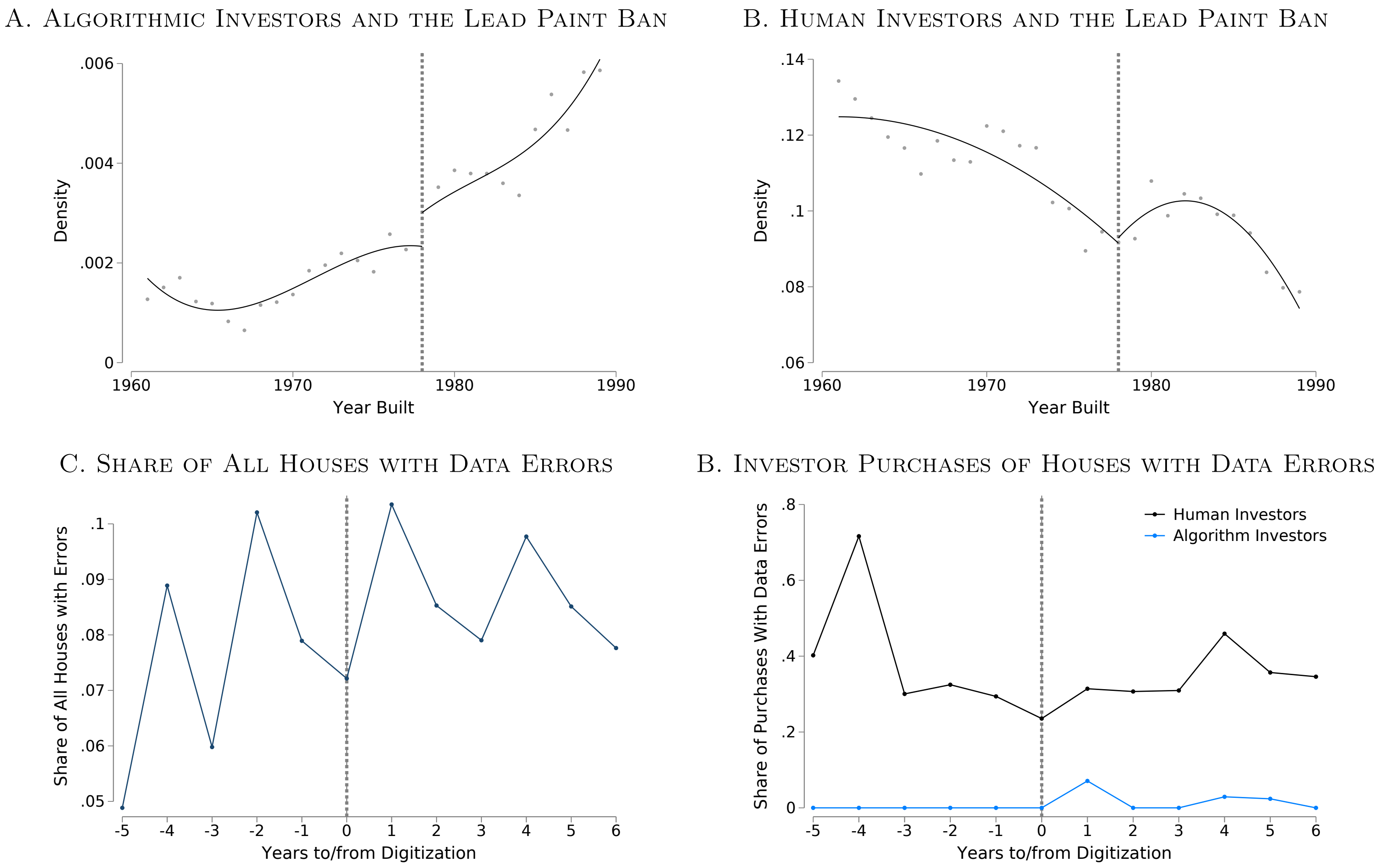
Notes: Panel A plots the plots the model-predicted natural log of sales price and actual sale price on held out sample of housing transactions. Panel B shows the same results separately for houses that will be purchased in the future by human investors and those that will be purchased by algorithmic investors. All data comes from ATTOM Data and Zillow.

Figure 6: Impact of Digitization by House Predictability



Notes: These figures plot the impact of house-level digitization on the likelihood of a purchase by a human investor. The model error is calculated as the average difference between the actual and predicted prices for each house. Errors are residualized to account for year-specific fixed effects. Every house is grouped into a decile of model error, with the houses with the lowest mean absolute error in decile 1 and the houses with the largest error in decile 10. All specifications include census block group and year-fixed effects. All data come from ATTOM Data and Zillow.

Figure 7: Discontinuities



A. Algorithmic Investors and the Lead Paint Ban

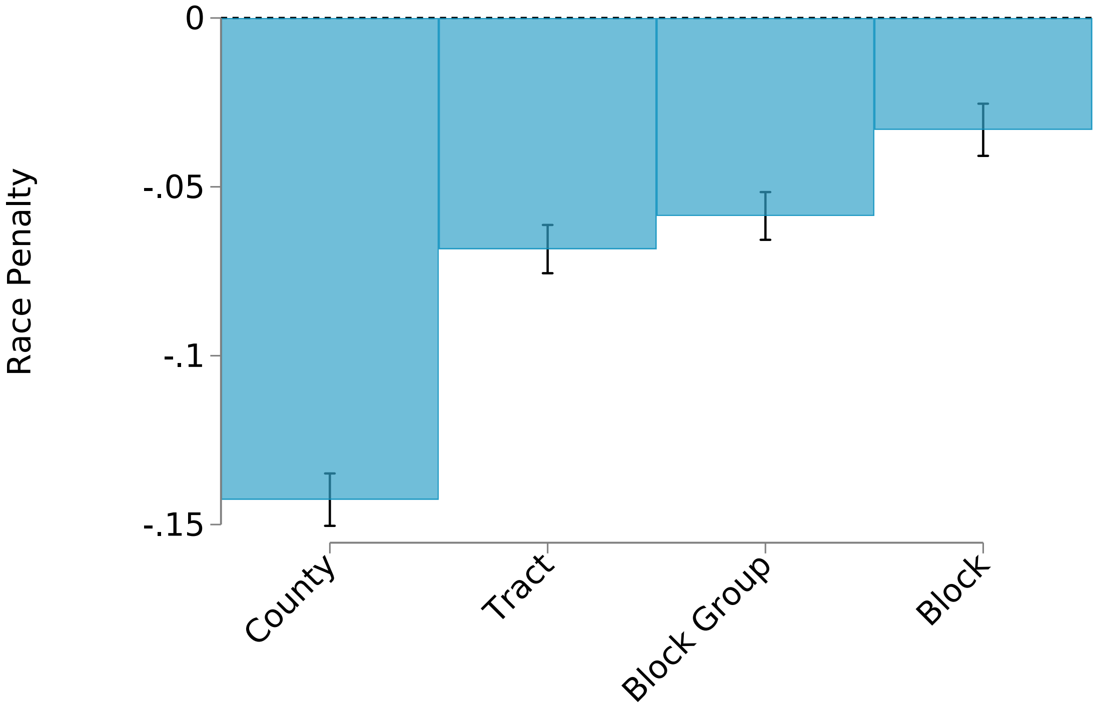
B. Human Investors and the Lead Paint Ban

C. Share of All Houses with Data Errors

B. Investor Purchases of Houses with Data Errors

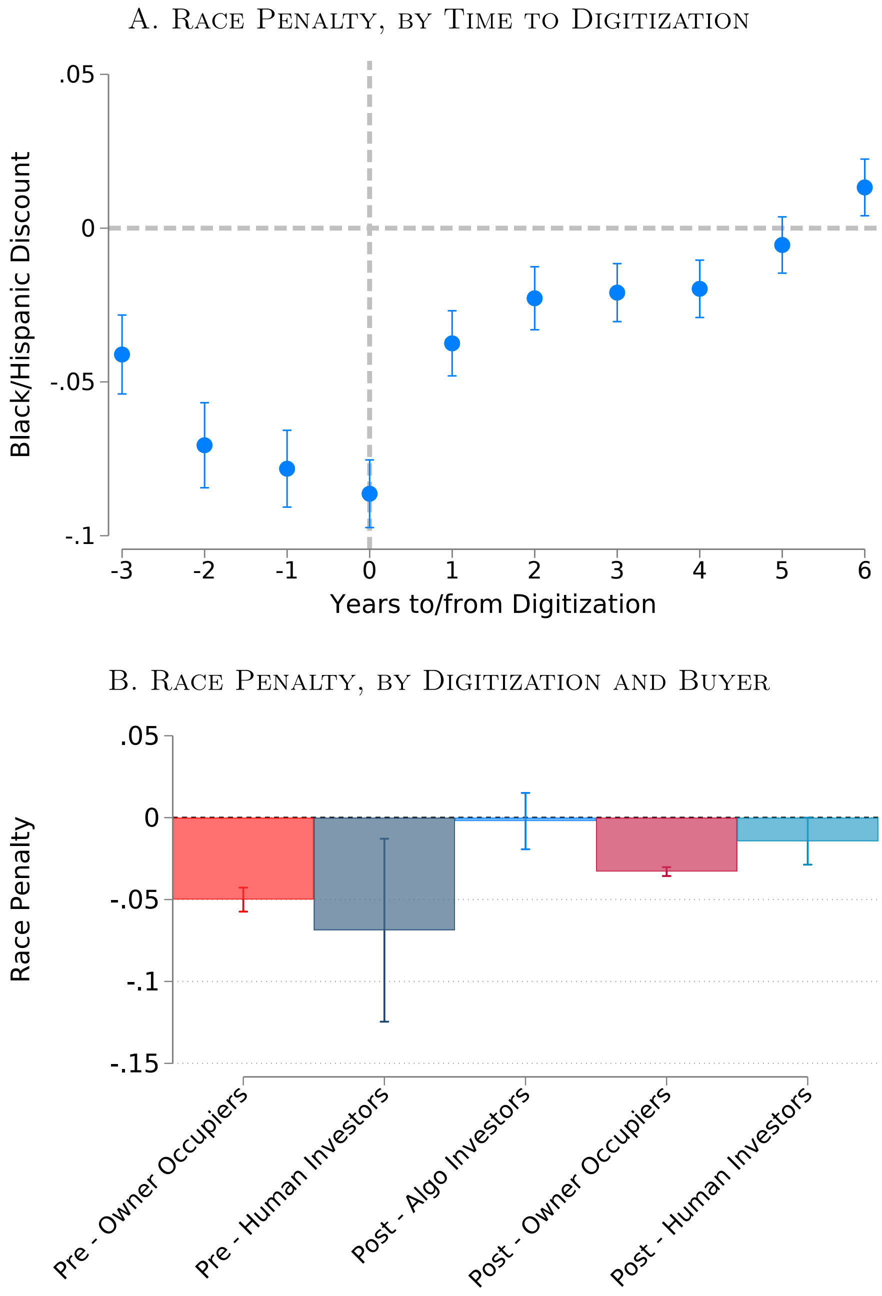
Notes: Panel A plots the distribution of houses purchased by algorithmic investors by year of construction. Panel B plots the same for human investors. Panel C shows the share of houses sold every year with data errors. Panel D plots the share of houses purchased by algorithmic and human investors with data errors. All data comes from ATTOM Data and county digitization records.

Figure 8: Race Penalty before Digitization, by Geography



Notes: This table shows the race penalty or coefficient value that captures the residual difference in sales price between an observably similar house sold by Black or Hispanic homeowners and one sold by a White homeowner. The race penalty is calculated during the time before digitization. The regressions run include geography and year fixed effects along with all available observable characteristics of the house. Standard errors are clustered at the relevant geography. All data comes from ATTOM Data.

Figure 9: Race Penalty, by Time to Digitization

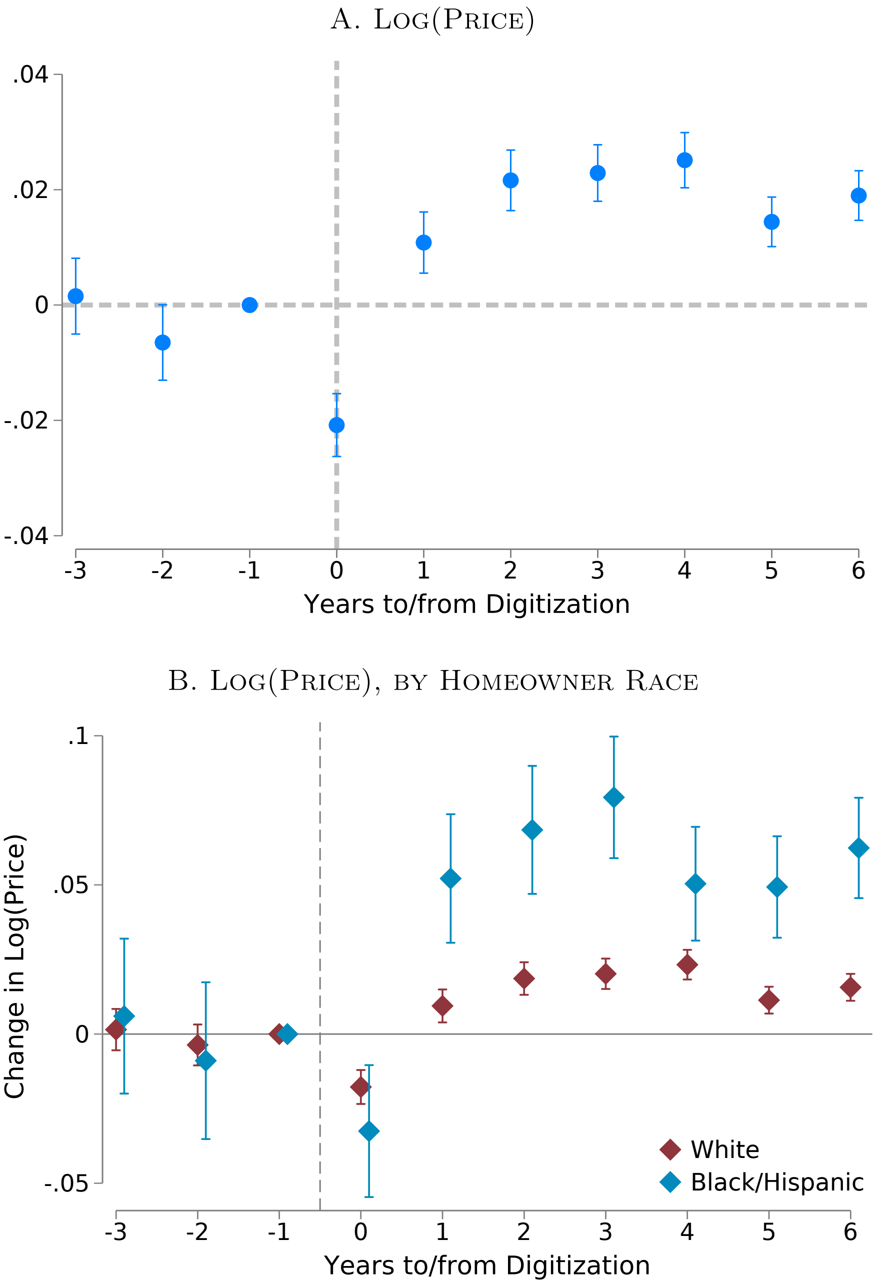


A. Race Penalty, by Time to Digitization

B. Race Penalty, by Digitization and Buyer

Notes: Panel A shows the race penalty, or residual difference in sale price between houses sold by White and minority homeowners by time to digitization. All specifications include census block group fixed effects and year-fixed effects and standard errors are clustered at the block group level. Panel B shows the same coefficient plotted pre- and post-digitization for houses purchased by three different types of buyers: owner-occupiers, human investors, and algorithmic investors. All data comes from ATTOM Data and county digitization records.

Figure 10: Digitization on Price

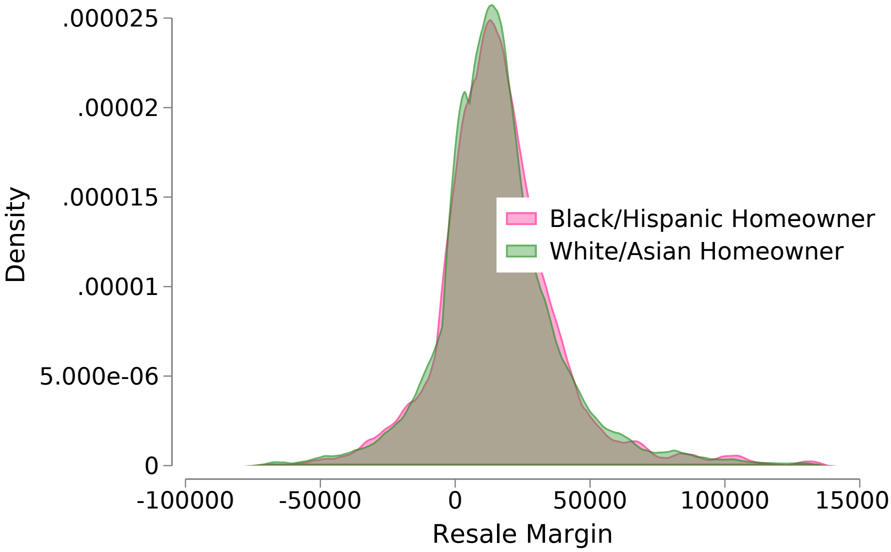


A. Log(Price)

B. Log(Price), by Homeowner Race

Notes: This graph plots the impact of digitization on the natural log of housing transaction prices at the county level in aggregate and separately by White and minority homeowner. All specifications include census block and year fixed effects, standard errors are clustered at the block level. All data comes from ATTOM Data, Zillow and county digitization records.

Figure 11: Resale Margin, by Homeowner Race



Notes: This graph plots the gross margin or difference between the sale price and the purchase price for houses bought by algorithmic investors according to the race of the homeowner. All data comes from ATTOM Data and county digitization records.

Table 1: Transaction Summary Statistics

	(1) Owner Occupiers	(2) Human Investors	(3) Algo Investors
Sale Price	194,270.04	127,755.99	219,130.88
	(158,431.32)	(145,159.96)	(103,655.74)
Bedrooms	2.12	2.27	2.76
	(3.17)	(3.58)	(1.47)
Bathrooms	2.14	2.09	2.47
	(2.38)	(5.30)	(1.01)
Partial Baths	0.27	0.25	0.43
	(0.48)	(0.48)	(0.50)
Stories	1.25	1.18	1.57
	(0.75)	(0.86)	(0.64)
Additional Buildings	0.07	0.12	0.03
	(0.58)	(1.18)	(0.24)
Garage	0.56	0.48	0.82
	(0.50)	(0.50)	(0.38)
Fireplace	0.59	0.55	0.82
	(0.49)	(0.50)	(0.39)
Basement	0.17	0.13	0.17
	(0.37)	(0.34)	(0.38)
Parking Spaces	0.75	0.58	0.91
	(8.72)	(7.40)	(0.99)
House Age	30.94	36.30	21.31
	(25.89)	(29.26)	(15.53)
Age Since Remodel	24.27	28.82	18.85
	(21.18)	(25.10)	(13.96)
Observations	7223587	975776	111027

mean coefficients; sd in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table shows the house characteristics of the transactions in our sample. The sample in column 1 includes all houses, including those purchased by owner-occupiers, those buying houses to live in, and investors. Column 2 includes purchases made by human investors, and column 3 includes purchases by investors using algorithms. Houses with missing or zero transaction prices are removed from the sample. All data come from ATTOM Data and ZTRAX.

Table 2: Investors Purchases, by County Characteristics

Variable	(1) Human Investors	(2) Algorithmic Investors	(3) Difference
County 2010 Population	347,012.69	498,205.72	151,193.02***
	(317,242.69)	(327,782.25)	(0.00)
Total Housing Units	152,197.36	208,074.12	55,876.77***
	(138,687.72)	(141,824.16)	(0.00)
Share Black	27.83	28.58	0.76***
	(16.73)	(14.35)	(0.00)
Share Hispanic	7.77	10.63	2.86***
	(4.20)	(4.66)	(0.00)
Share White	58.71	53.05	-5.66***
	(19.49)	(17.41)	(0.00)
Share Asian	2.98	4.55	1.56***
	(2.29)	(2.93)	(0.00)
Share Persons under 18	24.47	26.39	1.93***
	(2.74)	(2.33)	(0.00)
Median Income	54,298.52	66,305.10	12,006.58***
	(12,705.78)	(12,281.54)	(0.00)
Median Rent	896.24	1,085.91	189.67***
	(191.01)	(182.29)	(0.00)
Share Families in Poverty	11.74	9.31	-2.43***
	(3.82)	(2.83)	(0.00)
Mean Family Size	3.18	3.29	0.10***
	(0.19)	(0.15)	(0.00)
Share Persons under 18	24.47	26.39	1.93***
	(2.74)	(2.33)	(0.00)
Observations	975,776	111,027	1,086,803

Notes: This table shows socioeconomic and demographic characteristics of counties where algorithmic and human investors purchase houses, weighted by the number of purchases. Data is at the house transaction level. All data comes from the US Decennial Census and the American Community Survey.

Table 3: Log(Houses Purchased) by Algorithmic Investors, Difference-in-Difference Estimators

	Point Estimate	Standard Error	Lower Bound 95% Confidence Interval	Upper Bound 95% Confidence Interval
TWFE-OLS	1.130	0.380	0.386	1.874
Borusyak-Jaravel-Spiess	2.451	0.446	1.578	3.325
Callaway-Sant'Anna	1.002	0.021	0.960	1.043
DeChaisemartin-D'Haultfoeuille	2.653	0.325	2.015	3.290
Sun-Abraham	1.988	0.286	1.428	2.549

Notes: This table shows the impact of county data digitization deployment on the log of houses purchased by algorithmic investors. I show results using the robust difference-in-differences estimators introduced in Borusyak et al. (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021) along with a traditional two way fixed-effects. Callaway and Sant'Anna (2021) are cannot be weighted, so I present the unweighted estimates. All regressions include county, year fixed effects, and standard errors are clustered at the county level. Regressions are weighted by the number of transactions.

Table 4: House Digitization on Algorithmic Investor Purchase

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Algorithmic Investors	Algorithmic Investors	Algorithmic Investors	Algorithmic Investors	Human Investors
House Digitized	0.0023**	0.0021***	0.0009**		
	(0.0011)	(0.0008)	(0.0004)		
County Digitization x House **Not** Digitized				-0.0006	0.0017
				(0.0008)	(0.0023)
County Digitization x House Digitized				0.0098***	-0.0069***
				(0.0008)	(0.0023)
Observations	6,895,957	6,890,606	6,817,554	6,890,606	6,890,606
R-squared	0.0550	0.0598	0.1056	0.0600	0.0716
House Characteristics	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Location FE	Tract	Block Group	Block	Block Group	Block Group
Preperiod DV Mean	.00013	.00013	.00013	.00013	0.1247

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the results of cross-sectional difference-in-difference regressions estimating the impact of house record digitization on the purchase by an algorithmic investor. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the geographic level. All data come from ATTOM Data, ZTRAX and county governments.

Table 5: Race Penalty, with House Images

	(1) Log(Price)	(2) Log(Price)	(3) Log(Price)
Seller Black/Hispanic	-0.0557***	-0.0441***	-0.021***
	(0.0023)	(0.0042)	(0.0051)
Observations	30,130	30,130	29,037
R-squared	0.69	0.71	0.83
House + Lot	Yes	Yes	Yes
Year x Geo	Yes	Yes	Yes
Geographic FE	Tract	Block Group	Block
Adjusted R-squared	.571	.598	.688

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the race penalty–the residual difference in sale price between houses sold by White and minority homeowners. House exteriors are captured using a deep learning model to create vector representations of house images and included in the regressions as controls. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the geographic level. All data come from ATTOM Data, ZTRAX, Zillow and investor websites.

Table 6: House Digitization on Algorithmic Investor Purchase, by Homeowner Race

	(1) Algorithm Purchase	(2) Algorithm Purchase	(3) Algorithm Purchase
Seller Minority	-0.0037***	-0.0039***	-0.0049***
	(0.0005)	(0.0005)	(0.0004)
Digitization x Seller White	0.0022**	0.0020**	0.0007*
	(0.0011)	(0.0008)	(0.0004)
Digitization x Seller Minority	0.0044***	0.0042***	0.0043***
	(0.0007)	(0.0006)	(0.0005)
Geography FE	Tract	Block Group	Block
Year FE	Yes	Yes	Yes
Sample	All	All	All
DV Mean	.0018	.0018	.0018
Observations	6895957	6890606	6817554

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the results of cross-sectional difference-in-difference regressions estimating the impact of house record digitization on the purchase by an algorithmic investor. I separately estimate effects by homeowner race. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the relevant geographic level. All data come from ATTOM Data, ZTRAX and county governments.

Table 7: Resale Margin

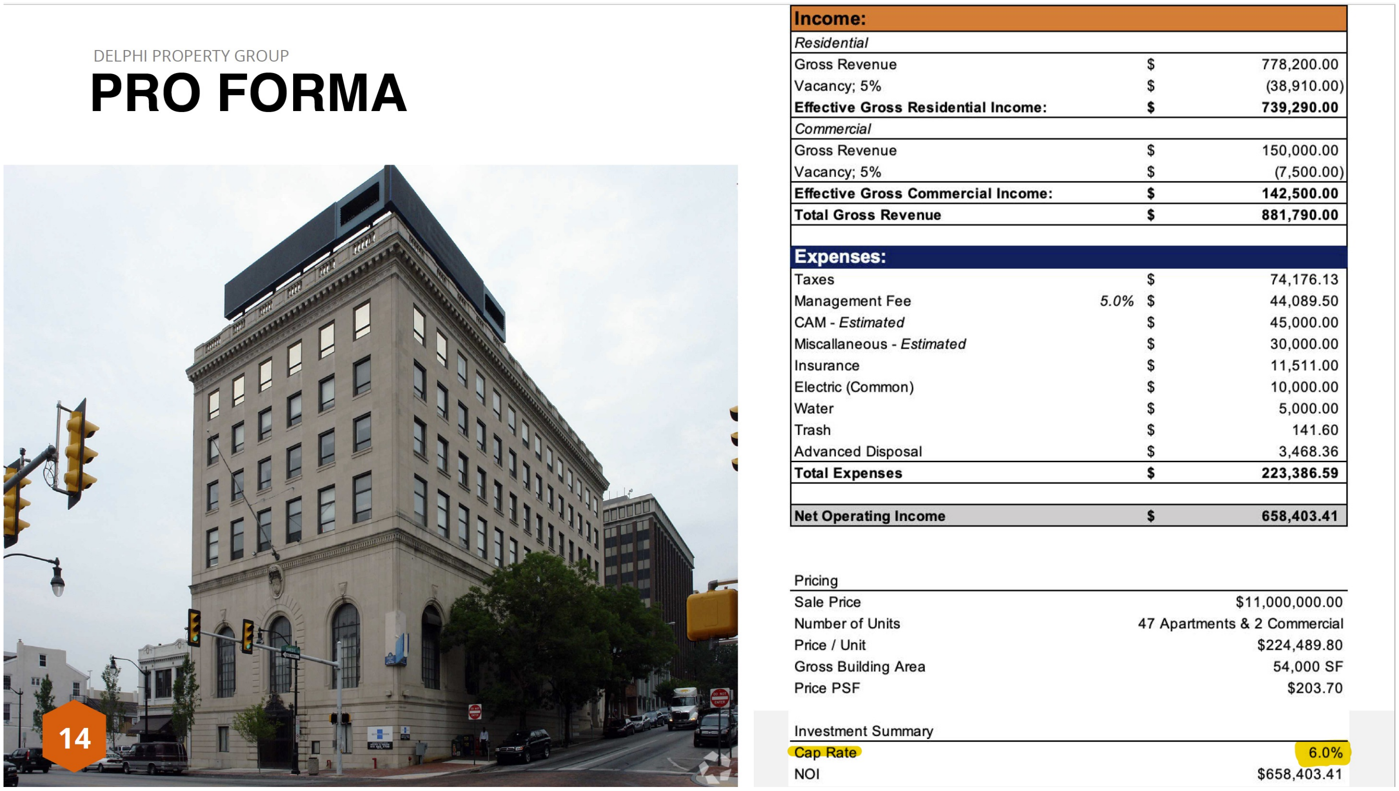
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Log(Resale Margin)	Log(Resale Margin)	Log(Resale Margin)	Log(Resale Margin)	Log(Resale Margin)	Log(Resale Margin)	Log(Resale Margin)	Log(Resale Margin)
Seller Minority = 1	0.010	0.010	-0.003	0.002	0.095***	0.101***	0.121***	0.066
	(0.009)	(0.010)	(0.018)	(0.017)	(0.019)	(0.021)	(0.033)	(0.041)
Seller Minority x Minority Neighborhood = 1				0.012				0.047
				(0.021)				(0.048)
Observations	6,775	5,630	2,212	5,630	56,459	45,527	23,449	45,527
R-squared	0.459	0.515	0.646	0.515	0.403	0.452	0.474	0.452
FE	Year x Tract	Year x Block Group	Year x Block	Year x Block Group	Year x Tract	Year x Block Group	Year x Block	Year x Block Group
Resale Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Buyers	Algorithms	Algorithms	Algorithms	Algorithms	Humans	Humans	Humans	Humans
DV Mean	0.0646	0.0576	0.0549	0.0576	0.369	0.362	0.307	0.362

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table shows the difference in the natural log of the price the house sells for in the future, or the resale price, and the natural log of the price paid, or the *gross margin*. The *Seller Minority* variable indicates if the house was bought from Black or Hispanic homeowners or White homeowners. *Minority neighborhood* indicates if the house is in a census block group with an above average minority resident share. All specifications includes the resale year and sale year by geography fixed effects, and standard errors are clustered at the geographic level. All data comes from ATTOM Data.

Appendix Materials

Figure A.1: Capitalization Rate Example



DELPHI PROPERTY GROUP

PRO FORMA

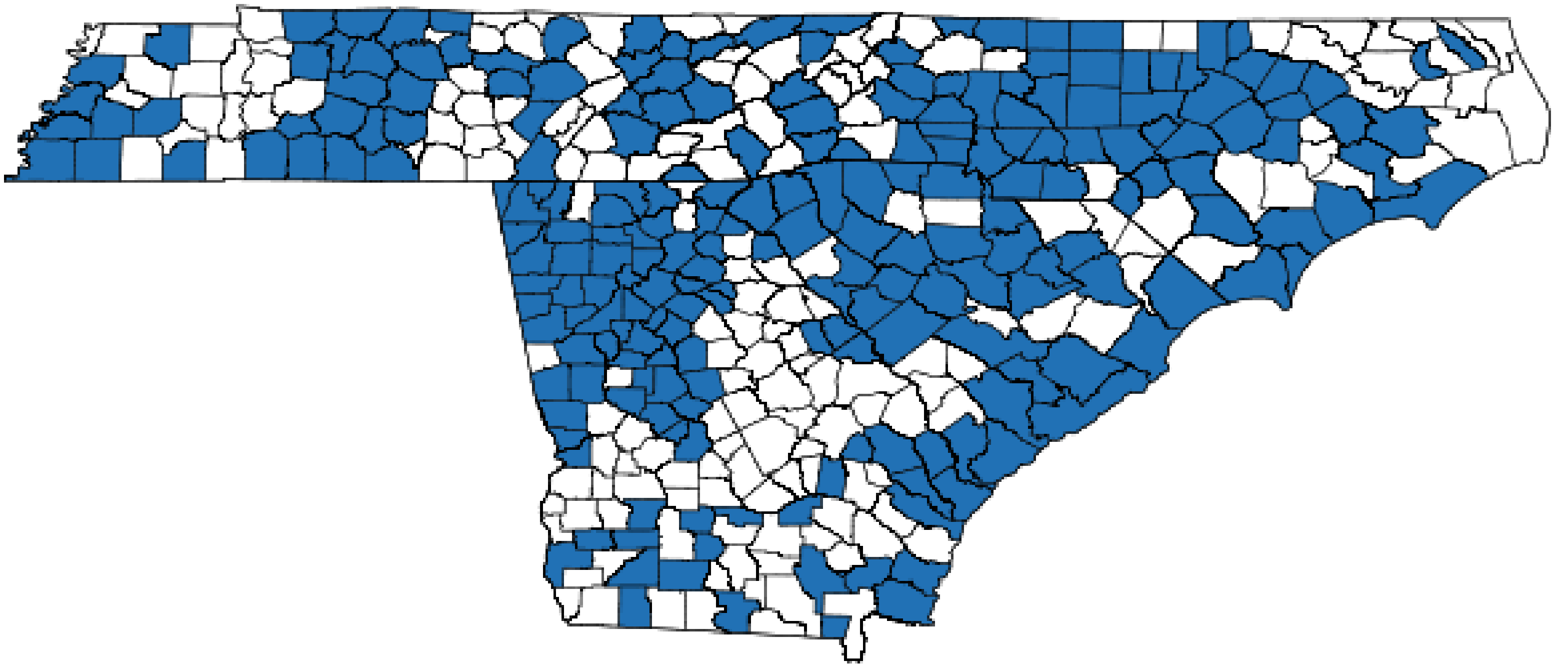
Income:			
Residential			
Gross Revenue		$	778,200.00
Vacancy; 5%		$	(38,910.00)
Effective Gross Residential Income:		**$**	**739,290.00**
Commercial			
Gross Revenue		$	150,000.00
Vacancy; 5%		$	(7,500.00)
Effective Gross Commercial Income:		**$**	**142,500.00**
Total Gross Revenue		**$**	**881,790.00**
Expenses:			
Taxes		$	74,176.13
Management Fee	*5.0%*	$	44,089.50
CAM - *Estimated*		$	45,000.00
Miscallaneous - *Estimated*		$	30,000.00
Insurance		$	11,511.00
Electric (Common)		$	10,000.00
Water		$	5,000.00
Trash		$	141.60
Advanced Disposal		$	3,468.36
Total Expenses		**$**	**223,386.59**
Net Operating Income		**$**	**658,403.41**

Pricing	
Sale Price	$11,000,000.00
Number of Units	47 Apartments & 2 Commercial
Price / Unit	$224,489.80
Gross Building Area	54,000 SF
Price PSF	$203.70

Investment Summary	
Cap Rate	6.0%
NOI	$658,403.41

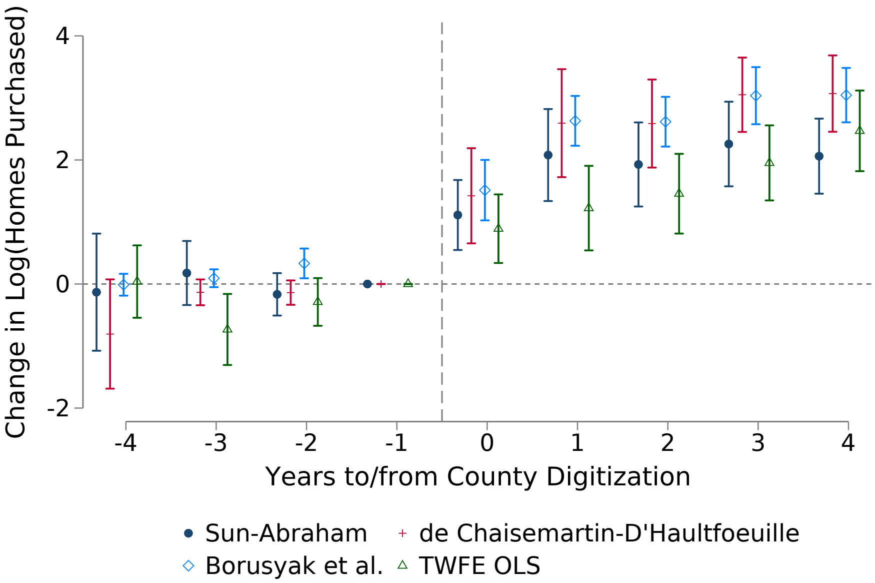
Notes: This shows a sample of the marketing material for 1 West Main Street Norristown, PA, a mixed use multifamily apartment building. This page includes the building capitalization rate.

Figure A.2: Investor Activity



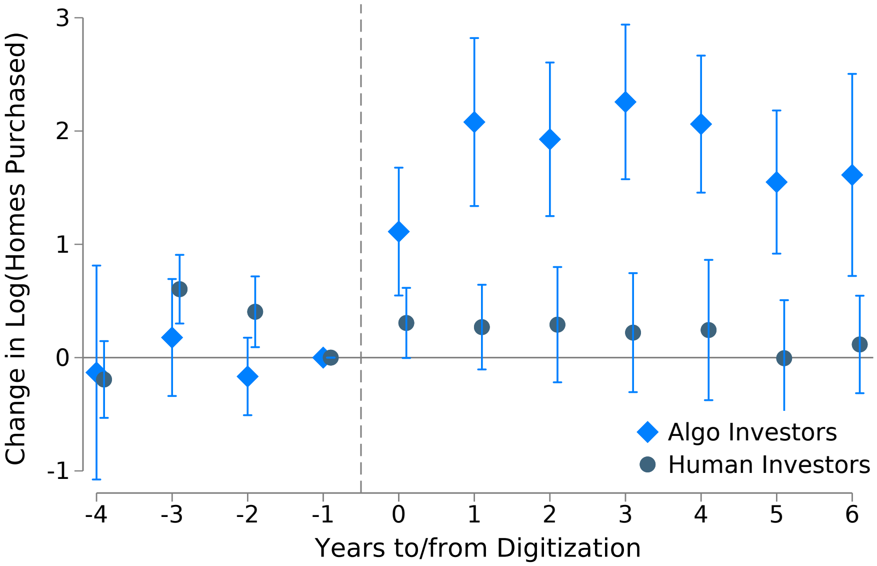
Notes: This graph shows the counties in Georgia, North Carolina, South Carolina and Tennessee where human and algorithm investors are active in blue. Counties in White have only human investors.

Figure A.3: Alternative Event Studies, Log(Houses Purchased) by Algorithmic Investors



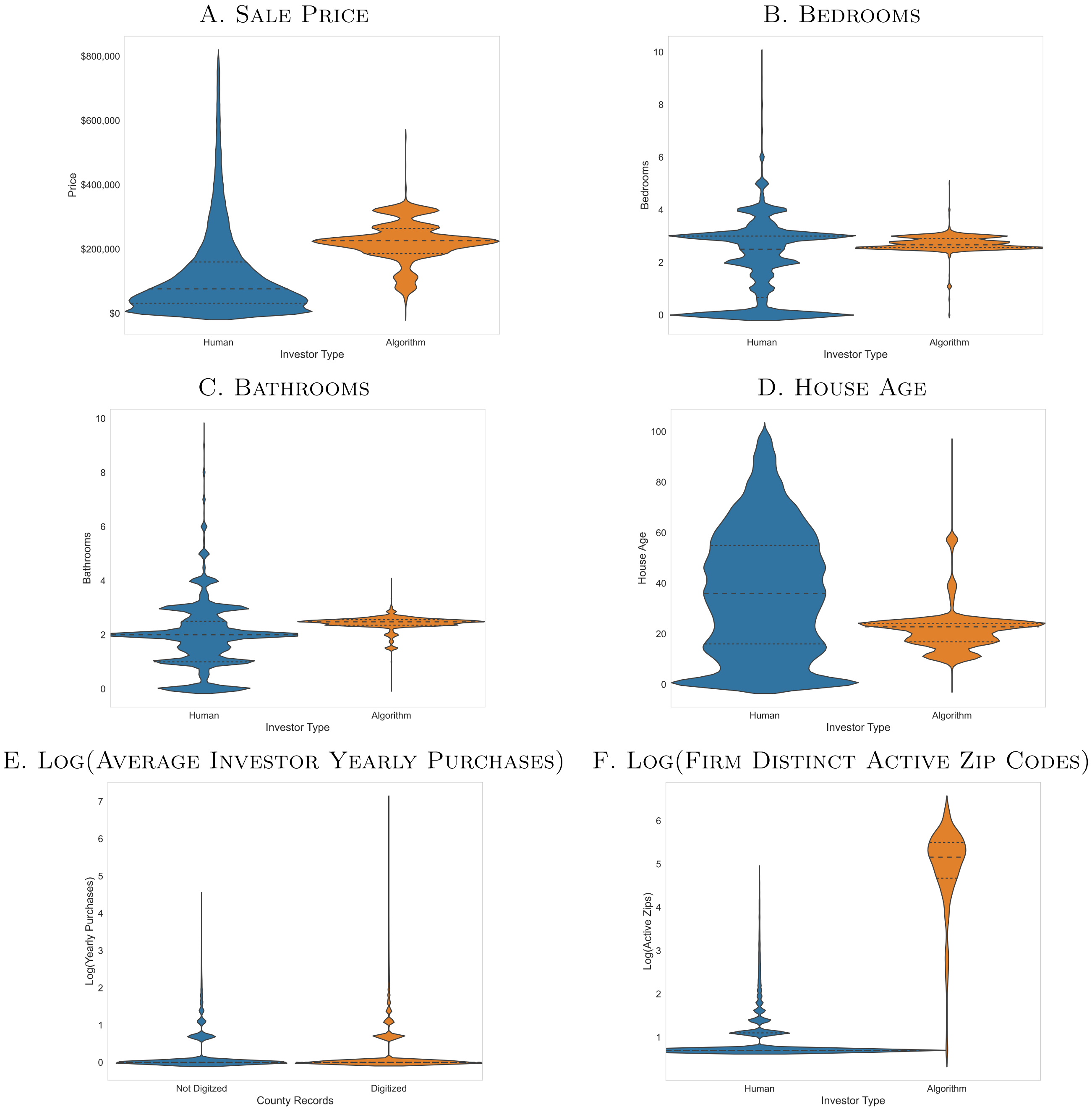
Notes: These figures plot the coefficients and 95 percent confidence intervals using a variety of robust dynamic difference-in-differences estimators introduced in Borusyak et al. (2022), de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021) and a standard two-way fixed effects regression model. All specifications include state and year fixed effects, standard errors are clustered at the county level and are weighted by the number of transactions. All data comes from ATTOM Data, Zillow, and county digitization records.

Figure A.4: Log(House Purchases) by Investor Type



Notes: This figure plots coefficients and 95 percent confidence interval from Sun and Abraham (2021) interaction-weighted event study regressions of county digitization on the natural log of the quantity of homes. I plot these results separately for the number of houses purchased by human or algorithmic investors in each county and year, weighted by the number of transactions. The regression includes state and year fixed effects, and standard errors are clustered at the county level. All data comes from ATTOM Data, Zillow and county digitization records.

Figure A.5: House Characteristics, by Investor


A. Sale Price

B. Bedrooms

C. Bathrooms

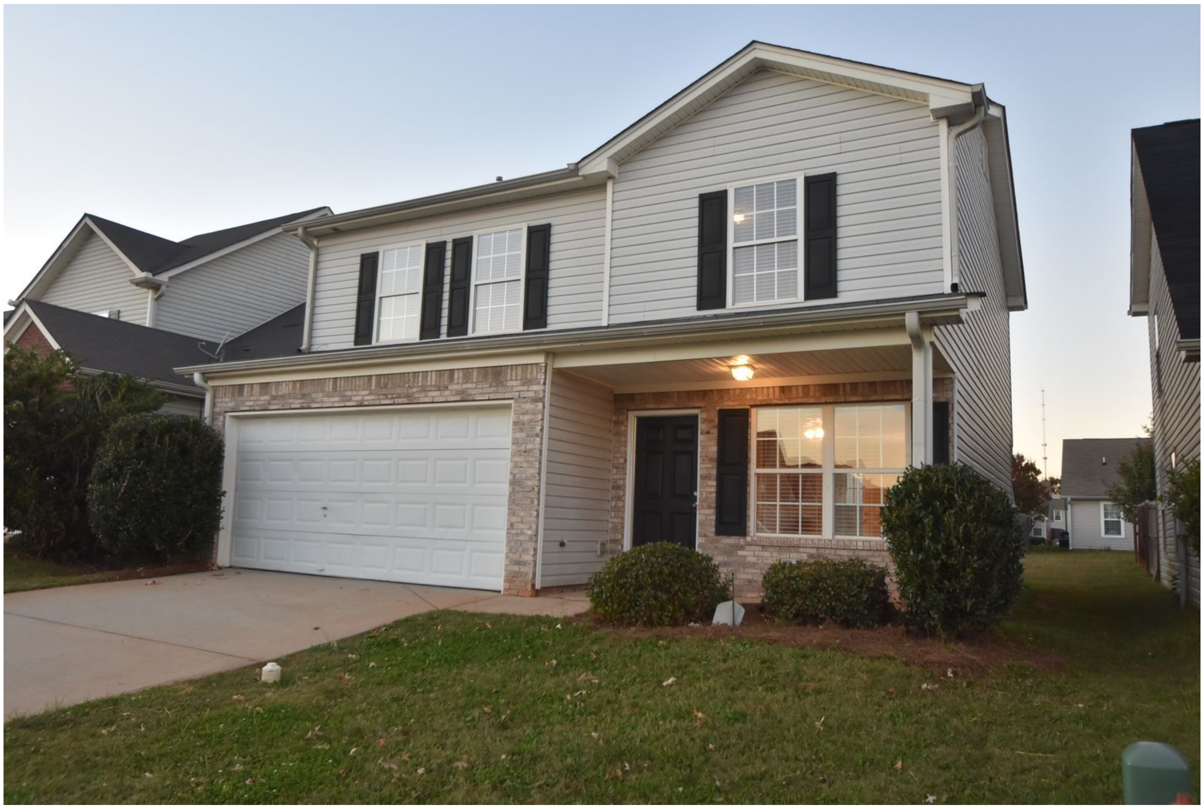
D. House Age

E. Log(Average Investor Yearly Purchases)

F. Log(Firm Distinct Active Zip Codes)

Notes: This figures plots characteristics of houses purchased by human and algorithmic investors. Panel A plots the purchase prices of houses, Panel B plots the number of house bedrooms. Panel C shows the number of bathrooms and panel D shows the age of the house. Panel E plots the natural log of average houses purchased by investors each before and after digitization. Panel F plots the natural log of zip codes investors are active each year. All data come from ATTOM Data and Zillow.

Figure A.6: House Exterior Images



Notes: This shows an example of the exterior images of the house used in the deep learning model.

Table A.1: Balance Table: Counties, by Year of Digitization

Variable	(1) Early Digitizers	(2) Late Digitizers	(3) Difference
Population	84,157.59	49,187.26	-34,970.33***
	(144,805.80)	(51,842.46)	(0.00)
Unemployment Rate	4.69	4.51	-0.19
	(1.81)	(1.69)	(0.36)
Share in Labor Force	56.67	54.77	-1.89**
	(7.01)	(6.24)	(0.01)
Share Units Occupied	82.58	81.33	-1.25
	(8.68)	(8.82)	(0.23)
Share Vacant	2.17	1.90	-0.27
	(1.44)	(2.17)	(0.27)
Median Rent	710.86	679.33	-31.53*
	(164.93)	(161.44)	(0.10)
Share Families in Poverty	14.66	14.62	-0.04
	(5.25)	(4.99)	(0.95)
Mean Family Size	3.14	3.07	-0.07***
	(0.29)	(0.20)	(0.01)
Median Income	44,399.43	42,521.30	-1,878.12
	(11,331.64)	(12,210.29)	(0.19)
Share Black	22.99	19.92	-3.07
	(18.08)	(19.44)	(0.17)
Share Hispanic	5.81	4.63	-1.18***
	(4.61)	(3.64)	(0.01)
Share White	67.19	71.74	4.56*
	(19.80)	(20.63)	(0.06)
Share Asian	1.25	0.85	-0.40***
	(1.31)	(0.95)	(0.00)
Observations	303	97	400

Notes: This table shows the covariate balance table for counties digitized before and after the median. All variables are calculated at the county level. All data come from ATTOM Data, ZTRAX and the US Census.

Table A.2: Balance Table: Houses, by Year of Digitization

	(1)	(2)	(3)
Variable	Early Digitizers	Late Digitizers	Difference
Years since Sale	10.55	9.79	-0.76***
	(9.49)	(8.74)	(0.00)
Sale Price	210,262.55	202,658.52	-7,604.03***
	(959,260.94)	(804,228.75)	(0.00)
Bedrooms	2.19	2.07	-0.12***
	(1.68)	(3.29)	(0.00)
Bathrooms	2.03	2.14	0.11***
	(2.69)	(2.24)	(0.00)
Partial Baths	0.29	0.27	-0.02***
	(0.52)	(0.47)	(0.00)
Stories	1.17	1.26	0.09***
	(0.88)	(0.69)	(0.00)
Buildings	0.05	0.07	0.01***
	(0.42)	(0.53)	(0.00)
Garage	0.55	0.56	0.02***
	(0.50)	(0.50)	(0.00)
Fireplace	0.60	0.58	-0.02***
	(0.49)	(0.49)	(0.00)
Basement	0.18	0.17	-0.01***
	(0.38)	(0.37)	(0.00)
Parking Spaces	0.97	0.69	-0.28***
	(17.96)	(1.77)	(0.00)
House Age	33.12	30.18	-2.94***
	(24.85)	(26.03)	(0.00)
Age Since Remodel	27.87	23.68	-4.19***
	(21.54)	(21.13)	(0.00)
Minority Homeowner	0.04	0.04	-0.00***
	(0.20)	(0.19)	(0.00)
Homeowner Asian	0.02	0.03	0.01***
	(0.15)	(0.17)	(0.00)
Homeowner White	0.88	0.87	-0.01***
	(0.33)	(0.34)	(0.00)
Observations	1,096,423	3,684,075	4,780,498

Notes: This table shows the covariate balance table for houses that digitized before and after the median ("Early Digitizers") or later ("Late Digitizers"). The unit of observation is at the house level. All data come from ATTOM Data, ZTRAX and the US Census.

Table A.3: Balance Table: Neighborhood Characteristics, by Year of Digitization

Variable	(1) Early Digitizers	(2) Late Digitizers	(3) Difference
Population	2,065.07	2,215.21	150.13***
	(1,284.76)	(1,448.16)	(0.00)
Housing Units	928.23	992.07	63.84***
	(540.46)	(600.80)	(0.00)
Share White	68.00	67.66	-0.34***
	(26.67)	(28.04)	(0.00)
Share Black	20.04	21.25	1.21***
	(21.69)	(24.18)	(0.00)
Share Asian	2.50	2.85	0.35***
	(3.12)	(4.33)	(0.00)
Share Under 18	23.66	24.01	0.35***
	(6.34)	(6.22)	(0.00)
Median Earnings	53,533.62	54,046.14	512.52***
	(13,100.19)	(13,607.33)	(0.00)
Rent	864.28	878.85	14.57***
	(200.30)	(199.76)	(0.00)
Age	38.23	38.09	-0.13***
	(4.39)	(4.43)	(0.00)
Mortgage Costs	1,310.06	1,332.71	22.65***
	(236.65)	(256.49)	(0.00)
Median List Price	216,445.45	205,233.47	-11,211.99***
	(75,300.37)	(70,603.58)	(0.00)
Days on the Market	109.73	107.64	-2.08***
	(30.66)	(27.31)	(0.00)
Observations	1,096,423	3,684,075	4,780,498

Notes: This table shows the covariate balance table for houses that digitized before and after the median ("Early Digitizers") or later ("Late Digitizers"). When possible, all statistics are at the census block group level. Information from Zillow is at the zip code level and the unit of observation is at the house level. All data come from ATTOM Data, ZTRAX and the US Census.

Table A.4: County Digitization and Algorithmic Investors Buying

VARIABLES	(1) Ln(Q_Algo)	(2) Ln(Q_Algo)	(3) Ln(Q_Algo)
County Digitization	1.130**	0.780**	0.749**
	(0.380)	(0.221)	(0.229)
Observations	3,962	3,962	3,962
R-squared	0.798	0.812	0.816
Year FE	Yes	Yes	Yes
Location FE	Yes	Yes	Yes
SocioEconomics	-	Yes	Yes
Housing Stock	-	-	Yes
DV Mean	2.597	2.597	2.597

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the results of county-level difference-in-difference regressions estimating the effect of county record digitization on the natural log of houses purchased by algorithmic investors. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the county level. Column 2 includes county population, demographics, poverty, unemployment rate and educational characteristics. Column 3 add housing stock characteristics such as the number of housing units and rent burden. All data comes from ATTOM Data, ZTRAX the US Census and county governments.

Table A.5: Algorithmic Investor Purchase, by Homeowner Race, Investor Sample

	(1) Algorithm Purchase	(2) Algorithm Purchase	(3) Algorithm Purchase
Seller Minority	-0.0133***	-0.0137***	-0.0194***
	(0.0036)	(0.0036)	(0.0045)
Digitization x Seller White	0.0079*	0.0075**	0.0042*
	(0.0045)	(0.0035)	(0.0022)
Digitization x Seller Minority	0.0415***	0.0396***	0.0389***
	(0.0043)	(0.0042)	(0.0050)
Geography FE	Tract	Block Group	Block
Year FE	Yes	Yes	Yes
Sample	Investors	Investors	Investors
DV Mean	.0018	.0018	.0018
Observations	898975	898061	802192

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the results of cross-sectional difference-in-difference regressions estimating the impact of house record digitization on the purchase by an algorithmic investor, by homeowner race. The sample includes only investor purchases, so the coefficients are interpreted as the likelihood of being purchased by an algorithmic investor compared to human investors. All specifications include house characteristics, year, and geography fixed effects, and standard errors are clustered at the geographic level. All data comes from ATTOM Data, ZTRAX and county governments.

Table A.6: IV Analysis: Algorithmic Investors and Race Penalty

	(1) First Stage	(2) 2SLS	(3) First Stage	(4) 2SLS	(5) First Stage	(6) 2SLS
Digitization	0.043***		0.046***		0.067***	
	(0.004)		(0.004)		(0.006)	
Algo Buyer		0.289		0.279		0.291
		(0.434)		(0.320)		(0.209)
AlgoxSellerBlack/Hisp		0.526***		0.529***		0.527***
		(0.127)		(0.116)		(0.107)
Geo Level	Tract+Year	Tract+Year	BG+Year	BG+Year	Block+Year	Block+Year
DV Mean	.002	164167	.002	164167	.002	164167
Adj R-squared	.317	.345	.317	.345	.344	.345
Observations	222666	222772	221537	222686	151452	222686

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the results of cross-sectional 2SLS regressions that estimate the algorithmic investor purchase on the race penalty, instrumenting for the algorithmic purchase with house-level digitization. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the geographic level and use log sale price as the outcome. All data comes from ATTOM Data, ZTRAX and county governments.

Table A.7: Assessment Margin

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Log(Assess Margin)	Log(Assess Margin)	Log(Assess Margin)	Log(Assess Margin)	Log(Assess Margin)	Log(Assess Margin)	Log(Assess Margin)	Log(Assess Margin)
Seller Minority = 1	0.004	0.004	0.002	0.003	-0.013*	-0.005	0.022**	-0.052***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)	(0.008)	(0.010)	(0.012)
Seller Minority x Minority Neighborhood = 1				0.003				0.090***
				(0.006)				(0.015)
Observations	81,387	76,312	46,294	76,312	467,588	440,142	238,501	440,142
R-squared	0.714	0.748	0.820	0.748	0.362	0.470	0.723	0.470
FE	Year x Tract	Year x Block Group	Year x Block	Year x Block Group	Year x Tract	Year x Block Group	Year x Block	Year x Block Group
Assessment Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Buyers	Algorithms	Algorithms	Algorithms	Algorithms	Humans	Humans	Humans	Humans
DV Mean	0.0911	0.0839	0.0577	0.0839	0.698	0.703	0.795	0.703

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the difference in the natural log of estimated house market value and the natural log transformation of the price paid, or *assessment margin*. *Minority neighborhood* indicates if the house is in a census block group with an above average minority resident share. All specifications include tax estimate and sale year by geography fixed effects. Standard errors are clustered at the geographic level. All data comes from ATTOM Data.