Applications of Machine Learning in Real Estate Markets Literature Review

David Veitch
University of Toronto
david.veitch@mail.utoronto.ca

June 2019

Contents

1 Introduction

A literature review is presented below on machine learning methods that are applicable to the real estate market. The literature is organized into several sections corresponding to the general themes which emerged from the literature. Each article includes an abstract, summary, and citation. As well, for articles where it is not immediately obvious what the application to real estate is, I have included my thoughts on how it could be applied. The literature review includes journal articles, conference papers, and working papers from both the economics and machine learning literature.

2 Literature Overview

With respect to applying machine learning to real estate markets, two main themes emerge from the literature. The first theme is the use of machine learning to generate new features about real estate that were previously time consuming or expensive to acquire. Computer vision and natural language processing are two techniques that are used extensively in the literature. The second theme is the use of machine learning when modelling housing prices. In many cases this takes the form of a machine learning algorithm such as a random forest directly predicting the price of a house, and in other cases it involves machine learning techniques such as LASSO performing automatic variable selection.

One of the main criticisms of machine learning is its lack of explainability. We see authors attempt to sidestep this problem by utilizing econometric models with high degrees of explainability (such as hedonic regression which is extensively used in modelling real estate prices), while relying on machine learning for feature generation or variable selection.

One final note is that it is clear from the literature that real estate problems lend themselves very well to machine learning. Machine learning is particularly well-suited to datasets that are large and of high dimension. We see many instances of these types of datasets in the literature, which are generally

collected by local governments or taxing authorities. As well, satellite images and real estate related news and advertisements can be obtained without prohibitive cost. All together, there appears to be a bright future for the application of machine learning to real estate related research questions.

3 Areas of Research

3.1 Computer Vision

Some of the most promising areas where machine learning can be applied to real estate involve computer vision. Generally these applications utilize street view images or satellite data to assess a property's value.

3.1.1 Vision-Based Real Estate Price Estimation [Poursaeed et al., 2017]

Abstract: *Since the advent of online real estate database companies like Zillow, Trulia and Redfin, the problem of automatic estimation of market values for houses has received considerable attention. Several real estate websites provide such estimates using a proprietary formula. Although these estimates are often close to the actual sale prices, in some cases they are highly inaccurate. One of the key factors that affects the value of a house is its interior and exterior appearance, which is not considered in calculating automatic value estimates. In this paper, we evaluate the impact of visual characteristics of a house on its market value. Using deep convolutional neural networks on a large dataset of photos of home interiors and exteriors, we develop a method for estimating the luxury level of real estate photos. We also develop a novel framework for automated value assessment using the above photos in addition to home characteristics including size, offered price and number of bedrooms. Finally, by applying our proposed method for price estimation to a new dataset of real estate photos and metadata, we show that it outperforms Zillow's estimates.*

Authors present an algorithm which incorporates home appearance into real estate price estimation. Authors also release a dataset of photos/metadata of 9,000 homes from Zillow. Authors attempt to estimate 'luxury level' from the photos in real estate listings. First a classifier is trained to classify which room of the house a photo is, or if it is the exterior of the house. Crowdsourcing is used to determine rooms of comparable luxury.

3.1.2 Image Based Appraisal of Real Estate Properties [You et al., 2016]

Abstract: *Real estate appraisal, which is the process of estimating the price for real estate properties, is crucial for both buys and sellers as the basis for negotiation and transaction. Traditionally, the repeat sales model has been widely adopted to estimate real estate price. However, it depends the design and calculation of a complex economic related index, which is challenging to estimate accurately. Today, real estate brokers provide easy access to detailed online information on real estate properties to their clients. We are interested in estimating the real estate price from these large amounts of easily accessed data. In particular, we analyze the prediction power of online house pictures, which is one of the key factors for online users to make a potential visiting decision. The development of robust computer vision algorithms makes the analysis of visual content possible. In this work, we employ a Recurrent Neural Network (RNN) to predict real estate price using the state-of-the-art visual features. The experimental results indicate that our model outperforms several of other state-of-the-art baseline algorithms in terms of both mean absolute error (MAE) and mean absolute percentage error (MAPE).*

Authors explore whether pictures of a house can aid in the task of real estate price estimation. Authors employ a random walk method to generate house sequences based on location to turn the problem into a sequence prediction problem. Microsoft Bing Map API was used to obtain latitude and longitudes for each house given its address. An RNN based model incorporating image data outperforms other methods on average, and the best RNN model significantly outperforms other model architectures.

3.1.3 Measurement and Valuation of Urban Greenness: Remote Sensing and Hedonic Applications to Lisbon, Portugal [Franco and Macdonald, 2018]

Abstract: *This paper explores the role of remote sensing techniques in capturing urban environmental data in the form of tree canopy coverage and measures of urban greenery. Using a classification algo-*

rithm, we identify tree canopy coverage in Lisbon, Portugal, to be approximately 8%. Our results have an accuracy rating of approximately 90% highlighting the benefits of this technique in capturing novel forms of data. Using these measures aggregated to the neighborhood level, we explore the impact of open space accessibility and urban greenness on the residential property market in Lisbon. We capture the heterogeneity of open spaces through their size and average vegetation level, and further explore how the greenness of a resident's neighborhood may elicit complementary or substitutability behavior in house pricing relative to proximity to urban open spaces and other urban ecological variables. Our results indicate that proximity to both large urban forests and smaller neighbourhood parks are capitalized through residential prices. These effects are dependent on neighborhood green composition with neighborhoods which have a higher proportion of sparse or low lying vegetation willing to trade-off proximity to parks (where this type of vegetation is abundant) and have a preference for being closer to urban forests (where there is greater diversity in vegetation from the neighborhood). Overall tree canopy coverage is positively valued with a square kilometer increase in the relative size of tree canopy valued at 0.20% of dwelling prices, or approximately €400 per dwelling. These results highlight the importance of capturing the heterogeneity of urban greenery and the interacting effects with the local ecology and the built environment.

Authors use a one-class SVM supervised machine learning algorithm to analyze spectral information in satellite data to locate tree canopies in Lisbon. One hurdle to overcome in this algorithm was enabling it to differentiate between tree canopies and other vegetation such as grass or lawns. Authors use these measures of tree canopies in a hedonic regression model to assess the impact of tree cover on the price of two bedroom apartments. The authors find the machine learning generated tree canopy features seems to perform better compared to other measures of urban greenness.

3.1.4 Daily Accessed Street Greenery and Housing Price: Measuring Economic Performance of Human-Scale Streetscapes via New Urban Data [Ye et al., 2019]

Abstract: *The protective effects of street greenery on ecological, psychological, and behavioral phenomena have been well recognized. Nevertheless, the potential economic effect of daily accessed street greenery, i.e., a human-scale and perceptual-oriented quality focusing on exposure to street greenery in people's daily lives, has not been fully studied because a quantitative measuring of this human-scale indicator is hard to achieve. This study was an attempt in this direction with the help of new urban data and new analytical tools. Shanghai, which has a mature real estate market, was selected for study, and the housing prices of 1395 private neighborhoods in its city center were collected. We selected more than forty variables that were classified under five categories—location features, distances to the closest facilities, density of facilities within a certain radius, housing and neighborhood features, and daily accessed street greenery—in a hedonic pricing model. The distance and density of facilities were computed through a massive number of points-of-interest and a geographical information system. The visible street greenery was collected from Baidu street view images and then measured via a machine-learning algorithm, while accessibility was measured through space syntax. In addition to the well-recognized effects previously discovered, the results show that visible street greenery and street accessibility at global scale hold significant positive coefficients for housing prices. Visible street greenery even obtains the second-highest regression coefficient in the model. Moreover, the combined assessment, the co-presence of local-scale accessibility and eye-level greenery, is significant for housing price as well. This study provides a scientific and quantitative support for the significance of human-scale street greenery, making it an important issue in urban greening policy for urban planners and decision makers.*

Authors use Baidu street view to determine the 'greenness' of a given neighbourhood in cities in China. SegNet, a deep convolutional neural network, is used to extract greenery from street-view images which is then used as a feature in a model predicting the average housing price in a neighbourhood.

3.1.5 Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks [Glaeser et al., 2018]

Abstract: *How much does the appearance of a house, or its neighbors, impact its price? Do events that impact the incentives facing homeowners, like foreclosure, impact the maintenance and appearance of a home? Using computer vision techniques, we find that a one standard deviation improvement in the appearance of a home in Boston is associated with a .16 log point increase in the home's value, or about $68,000 at the sample mean. The additional predictive power created by images is small relative to location and basic home variables, but external images do outperform variables collected by in-person*

home assessors. A home's value increases by .4 log points, when its neighbor's visually predicted value increases by one log point, and more visible neighbors have a larger price impact than less visible neighbors. Homes that went through foreclosure during the 2008-09 financial crisis experienced a .04 log point decline in their appearance-related value, relative to comparable homes, suggesting that foreclosures reduced the incentives to maintain the housing stock. We do not find more depreciation of appearance in rental properties, or more upgrading of appearance by owners before resale.

Authors use images from Google Street View of Boston houses to assess the external attractiveness of a home. The Resnet-101 CNN trained on Imagenet is used. Authors also look at the effect which the appearance of neighbouring homes have on a home's price.

3.1.6 Peer Effects in Water Conservation: Evidence from Consumer Migration [Bollinger et al., 2018]

Abstract: *Social interactions are widely understood to influence consumer decisions in many choice settings. This paper identifies causal peer effects in water conservation during the growing season, utilizing variation from consumer migration. We use machine learning to classify high-resolution remote sensing images to provide evidence that conversion to dry landscaping underpins the peer effects in water consumption. We also provide evidence that without a price signal, peer effects are muted, demonstrating a complementarity between information transmission and prices. These results inform water use policy in many areas of the world threatened by recurring drought conditions.*

Authors study the causal peer effects in water consumption using water billing and housing transaction data from 300,000 households in Phoenix. Authors attempt to identify houses which make the switch to dry landscaping (landscaping requiring minimal water use) and measures the effects on neighbours' water use. Machine learning was used to identify green space via satellite imagery.

3.1.7 Do People Shape Cities, or Do Cities Shape People? The Co-evolution of Physical, Social, and Economic Change in Five Major U.S. Cities [Naik et al., 2015]

Abstract: *Urban change involves transformations in the physical appearance and the social composition of neighborhoods. Yet, the relationship between the physical and social components of urban change is not well understood due to the lack of comprehensive measures of neighborhood appearance. Here, we introduce a computer vision method to quantify change in physical appearance of streetscapes and generate a dataset of physical change for five large American cities. We combine this dataset with socioeconomic indicators to explore whether demographic and economic changes precede, follow, or co-occur with changes in physical appearance. We find that the strongest predictors of improvement in a neighborhood's physical appearance are population density and share of college-educated adults. Other socioeconomic characteristics, like median income, share of vacant homes, and monthly rent, do not predict improvement in physical appearance. We also find that neighborhood appearances converge to the initial appearances of bordering areas, supporting the Burgess "invasion" theory. In addition, physical appearance is more likely to improve in neighborhoods proximal to the central business district. Finally, we find modest support for "tipping" and "filtering" theories of urban change.*

Authors use a computer vision method to quantify the physical appearance of streetscapes. Images extracted from Google Street View, and after unsuitable images removed, the authors used the Streetscore algorithm to predict what a crowdsourced rating of an image would be. Authors find that education and population density strongly predicted changes in the environment.

3.1.8 Exploring the Potential of Machine Learning for Automatic Slum Identification from VHR Imagery [Duque et al., 2017]

Abstract: *Slum identification in urban settlements is a crucial step in the process of formulation of propoor policies. However, the use of conventional methods for slums detection such as field surveys may result time consuming and costly. This paper explores the possibility of implementing a low-cost standardized method for slum detection. We use spectral, texture and structural features extracted from very high spatial resolution imagery as input data and evaluate the capability of three machine learning algorithms (Logistic Regression, Support Vector Machine and Random Forest) to classify urban areas as slum or no-slum. Using data from Buenos Aires (Argentina), Medellin (Colombia), and Recife (Brazil),*

we found that Support Vector Machine with radial basis kernel deliver the best performance (over 0.81). We also found that singularities within cities preclude the use of a unified classification model.

The literature shows physical characteristics of slums differ markedly from formal settlements, and mapping is an important area of study as many governments do not fully acknowledge their existence. Authors make use of VHR (very high spatial resolution) RBG Google Earth imagery, and find that the top performing slum identification algorithm is a SVM with a radial basis kernel. The authors note that it is important to take into account the differences across cities, precluding a unified classification model.

3.1.9 Inferring Home Location from User's Photo Collections based on Visual Content and Mobility Patterns [Zheng et al., 2014]

Abstract: *Precise home location detection has been actively studied in the past few years. It is indispensable in the researching fields such as personalized marketing and disease propagation. Since the last few decades, the rapid growth of geotagged multimedia database from online social networks provides a valuable opportunity to predict people's home location from temporal, spatial and visual cues. Among the massive amount of social media data, one important type of data is the geotagged web images from image-sharing websites. In this paper, we developed a reliable photo classifier based on the Convolutional Neutral Networks to classify photos as either home or non-home. We then proposed a novel approach to home location prediction by fusing together the visual content of web images and the spatiotemporal features of people's mobility pattern. Using a linear SVM classifier, we showed that the robust fusion of visual and temporal feature achieves significant accuracy improvement over each of the features alone.*

Authors analyze geotagged photos from Flickr. Based on the images a user posts the authors attempt to identify a user's home location (within 100m x 100m square). Train an image classifier to determine if a photo is 'home' or 'non-home'. Combines the insights derived from these images with temporal features to predict where an individual's home is.

3.1.10 The View from Above: Applications of Satellite Data in Economics [Donaldson and Storeygard, 2016]

Abstract: *The past decade or so has seen a dramatic change in the way that economists can learn by watching our planet from above. A revolution has taken place in remote sensing and allied fields such as computer science, engineering, and geography. Petabytes of satellite imagery have become publicly accessible at increasing resolution, many algorithms for extracting meaningful social science information from these images are now routine, and modern cloud-based processing power allows these algorithms to be run at global scale. This paper seeks to introduce economists to the science of remotely sensed data, and to give a flavor of how this new source of data has been used by economists so far and what might be done in the future.*

In this paper the authors discuss applications of machine learning to satellite data. These include land use classification and prediction of household income (at much higher frequencies than traditional surveys) based on images.

3.2 Natural Language Processing

These articles involve using natural language processing to analyze text documents. In a real estate setting the documents analyzed are generally news articles or home advertisements.

3.2.1 Textual Analysis in Real Estate [Nowak and Smith, 2017]

Abstract: *This paper incorporates text data from MLS listings into a hedonic pricing model. We show that the comments section of the MLS, which is populated by real estate agents who arguably have the most local market knowledge and know what homebuyers value, provides information that improves the performance of both in-sample and out-of-sample pricing estimates. Text is found to decrease pricing error by more than 25%. Information from text is incorporated into a linear model using a tokenization approach. By doing so, the implicit prices for various words and phrases are estimated. The estimation focuses on simultaneous variable selection and estimation for linear models in the presence of a large*

number of variables using a penalized regression. The LASSO procedure and variants are shown to outperform least-squares in out-of-sample testing.

Authors use LASSO technique to determine impact of words in real estate listings to houses' sale prices. To illustrate the results, bigrams (pairs of two words) associated with higher prices included: 'serious offers', 'lake front', and 'high end'; on the other hand, bigrams associated with lower prices were: 'auction terms', 'cash only', and 'needs work'.

3.2.2 Text as Data [Gentzkow et al., 2017]

Abstract: *An ever increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications.*

Generic overview of text analysis methods in the context of the social sciences. Mentions text regression as a good choice for predicting a single attribute, for multiple attributes generative models work best. Some applications of this analysis in the social sciences include: inferring authorship, predicting stock prices, gauging central bank sentiment, economic nowcasting, policy uncertainty, and media bias.

3.2.3 Dynamic Interpretation of Emerging Risks in the Financial Sector [Hanley and Hoberg, 2019]

Abstract: *We use computational linguistics to develop a dynamic, interpretable methodology that can detect emerging risks in the financial sector. Our model can predict heightened risk exposures as early as mid-2005, well in advance of the 2008 financial crisis. Risks related to real estate, prepayment, and commercial paper are elevated. Individual bank exposure strongly predicts returns, bank failures, and return volatility. We also document a rise in market instability since 2014 related to sources of funding and mergers and acquisitions. Overall, our model predicts the buildup of emerging risk in the financial system and bank-specific exposures in a timely fashion.*

Authors use computational linguistics applied to bank 10-K filings and daily stock returns to identify emerging risks in the financial sector. Latent Dirichlet Allocation and Semantic Vector Analysis methods are used to create interpretable risk factors that are stable over time. Specifically one of the risk factors that emerges is related to real estate. The authors present two models, one in which the researcher selects the risk factors, and one where they are identified automatically.

3.2.4 News Implied Volatility and Disaster Concerns [Manela and Moreira, 2017]

Abstract: *We construct a text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal. News implied volatility (NVIX) peaks during stock market crashes, times of policy-related uncertainty, world wars, and financial crises. In US postwar data, periods when NVIX is high are followed by periods of above average stock returns, even after controlling for contemporaneous and forward-looking measures of stock market volatility. News coverage related to wars and government policy explains most of the time variation in risk premia our measure identifies. Over the longer 1890–2009 sample that includes the Great Depression and two world wars, high NVIX predicts high future returns in normal times and rises just before transitions into economic disasters. The evidence is consistent with recent theories emphasizing time variation in rare disaster risk as a source of aggregate asset prices fluctuations.*

Authors estimate relationship between option prices/realized volatility and words using support vector regression (good for large feature spaces). Authors identify concerns about war and government policy are particularly relevant for explaining variation in risk premia. An application to real estate could include trying to explain variations in housing sales data based on disaster concerns.

3.2.5 News-Based Sentiment Analysis in Real Estate: A Machine Learning Approach [Hausler et al., 2018]

Abstract: *This paper examines the relationship between news-based sentiment, captured through a machine learning approach, and the US securitised and direct commercial real estate markets. Thus, we*

contribute to the literature on text-based sentiment analysis in real estate by creating and testing various sentiment measures by utilising trained support vector networks. Using a vector autoregressive framework, we find the constructed sentiment indicators to predict the total returns of both markets. The results show a leading relationship of our sentiment, even after controlling for macroeconomic factors and other established sentiment proxies. Furthermore, empirical evidence suggests a shorter response time of the indirect market in relation to the direct one. The findings make a valuable contribution to real estate research and industry participants, as we demonstrate the successful application of a sentiment-creation procedure that enables short and flexible aggregation periods. To the best of our knowledge, this is the first study to apply a machine learning approach to capture textual sentiment relevant to US real estate markets.

Authors look at the relationship between news sentiment (captured via a machine learning model) and the US securitized and direct commercial real estate market. SVM algorithm aggregates news article sentiment to create a monthly measure of market sentiment. A pessimism indicator drives returns, and leads the public market by 1 month, and direct market by 2, 3, and 8 months. The data sources include: S&P Global Market Intelligence Platform, CoStar Commercial Repeat-Sale Index, and FTSE/NAREIT All Equity REIT Total Return Index.

3.2.6 The Potential of Big Housing Data: An Application to the Italian Real-Estate Market [Loberto et al., 2018]

Abstract: *We present a new dataset of housing sales advertisements (ads) taken from Immobiliare.it, a popular online portal for real estate services in Italy. This dataset fills a big gap in Italian housing market statistics, namely the absence of detailed physical characteristics for houses sold. The granularity of online data also makes possible timely analyses at a very detailed geographical level. We first address the main problem of the dataset, i.e. the mismatch between ads and actual housing units - agencies have incentives for posting multiple ads for the same unit. We correct this distortion by using machine learning tools and provide evidence about its quantitative relevance. We then show that the information from this dataset is consistent with existing official statistical sources. Finally, we present some unique applications for these data. For example, we provide first evidence at the Italian level that online interest in a particular area is a leading indicator of prices. Our work is a concrete example of the potential of large user-generated online databases for institutional applications.*

Authors use a C5.0 classification tree algorithm to clean a real estate advertisement dataset by removing duplicates. Dataset shrinks by half after applying this algorithm.

3.2.7 Is There a Real Estate Bubble in Switzerland? [Ardila et al., 2013]

Abstract: *We have analyzed the risks of possible development of bubbles in the Swiss residential real estate market. The data employed in this work has been collected by comparis.ch, and carefully cleaned from duplicate records through a procedure based on supervised machine learning methods. The study uses the log periodic power law (LPPL) bubble model to analyze the development of asking prices of residential properties in all Swiss districts between 2005 and 2013. The results suggest that there are 11 critical districts that exhibit signatures of bubbles, and seven districts where bubbles have already burst. Despite these strong signatures, it is argued that, based on the current economic environment, a soft landing rather than a severe crash is expected.*

Authors use support vector machines and string distant measures to identify duplicate real estate ads. Features which are used to determine if ads duplicate include zip code, quarter, title, description, number of rooms, etc.

3.2.8 Precise Localization of Homes and Activities: Detecting Drinking-While-Tweeting Patterns in Communities [Hossain et al., 2016]

Abstract: *There has been an explosion of research on analyzing social media to map human behavior relevant to public health, such as drinking or drug use. However, it is important not only to detect the local regions where these activities occur, but also to analyze the degree of participation in them by local residents. We develop powerful new methods for fine-grained localization of activities and home locations using Twitter data. We apply these methods to discover and compare alcohol use patterns in a large city,*

New York City, and a more suburban and rural area, Monroe County.

Authors use a SVM on Twitter data to detect when an individual is consuming alcohol. This information is combined with the predicted location of these individuals' homes to determine the prevalence of alcohol consumption by local residents in a given area.

3.3 Feature Generation

This sections looks at applications of machine learning to generate new features or labels for a dataset, beyond what was discussed in the Computer Vision and Natural Language Processing sections. One type of feature generation that many authors have looked at is determining a buyer's ethnicity from their name.

3.3.1 Superstition and Real Estate Prices: Transaction-Level Evidence from the US Housing Market [Humphreys et al., 2019]

Abstract: *We investigate the impact of superstition on prices paid by Chinese-American home buyers. Chinese consider 8 lucky and 4 unlucky. Lacking explicit buyer ethnicity identifiers, we develop a binomial name classifier, a machine learning approach applicable to any data set containing names, that allows for falsification tests using other ethnic groups, and mitigates ambiguity from the transliteration of Chinese characters into the Latin alphabet. Chinese buyers pay 1–2% premiums for addresses including an 8 and 1% discounts for addresses including a 4. These results are unrelated to unobserved property quality; no premium exists when Chinese sell to non-Chinese. The persistence of superstitions reflects the extent of cultural assimilation.*

Authors use Seattle real estate data to determine if Chinese individuals pay a premium or discount for properties based on specific numbers in the property's address. Authors use a binomial classifier trained on rosters from the Olympic games to determine ethnicity.

3.3.2 Homeowner Preferences After September 11th, a Microdata Approach [Nowak and Sayago-Gomez, 2018]

Abstract: *The existence of homeowner preferences – specifically homeowner preferences for neighbors – is fundamental to economic models of sorting. This paper investigates whether or not the terrorist attacks of September 11, 2001 (9/11) impacted local preferences for Arab neighbors. We test for changes in preferences using a differences-in-differences approach in a hedonic pricing model. Relative to sales before 9/11, we find properties within 0.1miles of an Arab homeowner sold at a 1.4% discount (approximately $4133) in the 180 days after 9/11. The results are robust to a number of specifications including time horizon, event date, distance, time, alternative ethnic groups, and the presence of nearby mosques. Previous research has shown price effects at neighborhood levels but has not identified effects at the micro or individual property level, and for good reason: most transaction level data sets do not include ethnic identifiers. Applying methods from the machine learning and biostatistics literature, we develop a binomial classifier using a supervised learning algorithm and identify Arab homeowners based on the name of the buyer. We train the binomial classifier using names from Summer Olympic Rosters for 221 countries during the years 1948–2012. We demonstrate the flexibility of our methodology and perform an interesting counterfactual by identifying Hispanic and Asian homeowners in the data; unlike the statistically significant results for Arab homeowners, we find no meaningful results for Hispanic and Asian homeowners following 9/11.*

Authors use machine learning to label homeowners as Arab or non-Arab based on their names. This algorithm was trained on Summer Olympic rosters for all countries from 1948-2012. The output of the algorithm is the probability a given name found in a real estate transaction data set would be found on the Olympic roster of any Arab league country. Uses L_1 regularization to improve out-of-sample performance. The authors find a discount in prices paid for properties near Arab homeowners after 9/11.

3.3.3 Surname-Based Ethnicity and Ethnic Segregation in the Early Twentieth Century U.S. [Xu, 2019]

Abstract: *In this paper, I discuss a new measure of ethnicity in historical U.S. census data, and apply it in segregation studies. In the early twentieth century U.S., three major sending countries of Central and Eastern European immigrants—namely, Germany, Poland, and Russia—had high degrees of ethnic and cultural diversity. The population in all three countries comprised largely of a mixture of German, Polish, Russian, and Jewish ethnic groups. Consequently, there might be significant heterogeneity in ethnicity among U.S. immigrants born in the same home country. Focusing on the above three sending countries in the 1920 and 1930 U.S. census, I construct an ethnicity variable based essentially on the linguistic origin of the surname. Employing this variable, I examine ethnic segregation within each immigrant group defined based on the country of birth. Results suggest high degrees of within-group ethnic segregation. In particular, ethnic majorities within each group generally resided in areas with significantly more compatriots.*

Author uses machine learning techniques to classify surnames of unknown ethnicity to the correct origin. Author first applies a deterministic algorithm based on previously studied issues of assigning ethnicity to surnames (e.g. common for speakers of Yiddish and Hebrew to be Jewish). In the second step of the algorithm the author uses a Naïve Bayes algorithm. Ultimately the author finds that just using a probabilistic algorithm achieves results similar to using this two-step process. An application to real estate could be assigning ethnicity to the names on transaction records.

3.3.4 Unstructured Data in Marketing [Balducci and Marinova, 2018]

Abstract: *The rise of unstructured data (UD), propelled by novel technologies, is reshaping markets and the management of marketing activities. Yet these increased data remain mostly untapped by many firms, suggesting the potential for further research developments. The integrative framework proposed in this study addresses the nature of UD and pursues theoretical richness and computational advancements by integrating insights from other disciplines. This article makes three main contributions to the literature by (1) offering a unifying definition and conceptualization of UD in marketing; (2) bridging disjoint literature with an organizing framework that synthesizes various subsets of UD relevant for marketing management through an integrative review; and (3) identifying substantive, computational, and theoretical gaps in extant literature and ways to leverage interdisciplinary knowledge to advance marketing research by applying UD analyses to underdeveloped areas.*

General overview of what unstructured data is, how it differs from structured data, and what questions it can address in marketing. Lays out different machine learning methods and what type of unstructured data it can address, and conducts a literature review of some cases where it has been applied. Text analysis and image analysis are particularly relevant for real estate.

3.3.5 Differentially Private M-Estimators [Lei, 2011]

Abstract: *This paper studies privacy preserving M-estimators using perturbed histograms. The proposed approach allows the release of a wide class of M-estimators with both differential privacy and statistical utility without knowing a priori the particular inference procedure. The performance of the proposed method is demonstrated through a careful study of the convergence rates. A practical algorithm is given and applied on a real world data set containing both continuous and categorical variables*

Differential privacy provides a mathematically rigorous privacy guarantee for the data, and requires the presence or absence of any individual record will not effect the outcome greatly. Preserves privacy by introducing randomness to data set, but this makes statistical analysis hard since the output of inference becomes random for fixed dataset. Author looks at M-estimators (a generalization of the MLE) under a differential privacy framework using 'perturbed histograms'. Author applies their method to real estate transaction data in San Francisco. Using linear regression they compare the difference in coefficients between a model run without any perturbations, and those with perturbations. Most coefficients change by less than 10%.

3.3.6 Coming Apart? Cultural Distances in the United States Over Time [Bertrand and Kamenica, 2018]

Abstract: *We analyze temporal trends in cultural distance between groups in the US defined by income, education, gender, race, and political ideology. We measure cultural distance between two groups as the ability to infer an individual's group based on his or her (i) media consumption, (ii) consumer behavior, (iii) time use, or (iv) social attitudes. Gender difference in time use decreased between 1965 and 1995 and has remained constant since. Differences in social attitudes by political ideology and income have increased over the last four decades. Whites and non-whites have converged somewhat on attitudes but have diverged in consumer behavior. For all other demographic divisions and cultural dimensions, cultural distance has been broadly constant over time.*

Authors attempt to measure the cultural distance between groups in the US over time. Distance defined as ability to predict whether an individual is in group A or B based on some covariates (e.g. media consumption), and a machine learning approach (elastic net, regression tree, and random forest) is used to conduct prediction. That is, if it is easy to classify people into groups than distance is large, if it is easy than the groups are fairly similar. An application to real estate could be measuring groups' cultural distance with respect to real estate behaviour.

3.3.7 Social Capital and Labor Market Networks [Asquith et al., 2017]

Abstract: *We explore the links between social capital and labor market networks at the neighborhood level. We harness rich data taken from multiple sources, including matched employer-employee data with which we measure the strength of labor market networks, data on behavior such as voting patterns that have previously been tied to social capital, and new data – not previously used in the study of social capital – on the number and location of non-profit sector establishments at the neighborhood level. We use a machine learning algorithm to identify important potential social capital measures that best predict neighborhood-level variation in labor market networks. We find evidence suggesting that smaller and less centralized schools, and schools with fewer poor students, foster social capital that builds labor market networks, as does a larger Republican vote share. The presence of establishments in a number of non-profit oriented industries are identified as predictive of strong labor market networks, likely because they either provide public goods or facilitate social contacts. These industries include, for example, churches and other religious institutions, police departments, fire and rescue services including volunteer fire departments, country clubs, mayors' offices, chamber music groups, hobby clubs, and museums.*

Authors use machine learning to examine whether higher social capital in a neighbourhood is associated with stronger labor market networks among neighbours. Authors construct four neighbourhood-level measures of social capital: demographic homogeneity, external parental involvement in schools, ideological homogeneity, and civic institutions. LASSO is used to select social capital covariates associated with the strength of local labour market networks. An application to real estate could include determining if certain features of a neighbourhood's real estate contribute to strong labour networks.

3.3.8 Automated Census Record Linking: A Machine Learning Approach [Feigenbaum, 2016]

Abstract: *Thanks to the availability of new historical census sources and advances in record linking technology, economic historians are becoming big data genealogists. Linking individuals over time and between databases has opened up new avenues for research into intergenerational mobility, assimilation, discrimination, and the returns to education. To take advantage of these new research opportunities, scholars need to be able to accurately and efficiently match historical records and produce an unbiased dataset of links for downstream analysis. I detail a standard and transparent census matching technique for constructing linked samples that can be replicated across a variety of cases. The procedure applies insights from machine learning classification and text comparison to the well known problem of record linkage, but with a focus on the sorts of costs and benefits of working with historical data. I begin by extracting a subset of possible matches for each record, and then use training data to tune a matching algorithm that attempts to minimize both false positives and false negatives, taking into account the inherent noise in historical records. To make the procedure precise, I trace its application to an example from my own work, linking children from the 1915 Iowa State Census to their adult-selves in the 1940 Federal Census. In addition, I provide guidance on a number of practical questions, including how large*

the training data needs to be relative to the sample.

Author finds relatively good out-of-sample performance of a machine learning algorithm, specifically probit regression, matching individuals across censuses. Results compared to random forest, SVM, logit, and OLS classifier models. Author uses baseball player biographical information to test if algorithm is creating incorrect links when an individual has died. An application to real estate could include attempting to link individuals across transaction datasets to see what type of properties people buy over time.

3.4 Prediction Tasks

Machine learning is particularly well-suited for prediction tasks. These articles look at instances where the prediction task is directly related to real estate, or algorithms which could be used in a real estate setting.

3.4.1 Machine Learning Methods for Demand Estimation [Bajari et al., 2015]

Abstract: *We survey and apply several techniques from the statistical and computer science literature to the problem of demand estimation. To improve out-of-sample prediction accuracy, we propose a method of combining the underlying models via linear regression. Our method is robust to a large number of regressors; scales easily to very large data sets; combines model selection and estimation; and can flexibly approximate arbitrary non-linear functions. We illustrate our method using a standard scanner panel data set and find that our estimates are considerably more accurate in out-of-sample predictions of demand than some commonly used alternatives.*

Authors compare machine learning methods for predicting consumer demand with traditional econometric models. Goal to find practical tools for datasets with large number of observations and covariates. Uses sales data of 'salty' snacks from one grocery store chain for six years. Find machine learning models produce better out-of-sample fits than linear models (particularly random forests and SVMs). Ensembling models performs better than any individual model (while retaining standard asymptotic properties). An application to real estate of this approach could include predicting housing demand (via new home sales) or housing construction.

3.4.2 Predictably Unequal? The Effects of Machine Learning on Credit Markets [Fuster et al., 2017]

Abstract: *Innovations in statistical technology, including in predicting creditworthiness, have sparked concerns about differential impacts across categories such as race. Theoretically, distributional consequences from better statistical technology can come from greater flexibility to uncover structural relationships, or from triangulation of otherwise excluded characteristics. Using data on US mortgages, we predict default using traditional and machine learning models. We find that Black and Hispanic borrowers are disproportionately less likely to gain from the introduction of machine learning. In a simple equilibrium credit market model, machine learning increases disparity in rates between and within groups; these changes are primarily attributable to greater flexibility.*

Authors examine risk that gains from better statistical modelling may not be evenly distributed. Machine learning delivers higher out-of-sample predictive accuracy of defaults than simpler logistic models, but predicted default propensities across race and ethnic groups look very different under the more sophisticated methods than the simpler methods. Authors estimate that only 8% of accuracy gains from machine learning comes from the algorithms triangulating which group an individual is in. Compare linear and nonlinear logistic models to random forest models, and random forest outperforms.

3.4.3 A First Estimation of the Proportion of Cybercriminal Entities in the Bitcoin Ecosystem Using Supervised Machine Learning [Sun Yin and Vatrapu, 2017]

Abstract: *Bitcoin, a peer-to-peer payment system and digital currency, is often involved in illicit activities such as scamming, ransomware attacks, illegal goods trading, and thievery. At the time of writing, the Bitcoin ecosystem has not yet been mapped and as such there is no estimate of the share of illicit*

activities. This paper provides the first estimation of the portion of cyber-criminal entities in the Bitcoin ecosystem. Our dataset consists of 854 observations categorised into 12 classes (out of which 5 are cybercrime-related) and a total of 100,000 uncategorised observations. The dataset was obtained from the data provider who applied three types of clustering of Bitcoin transactions to categorise entities: co-spend, intelligence-based, and behaviour-based. Thirteen supervised learning classifiers were then tested, of which four prevailed with a cross-validation accuracy of 77.38%, 76.47%, 78.46%, 80.76% respectively. From the top four classifiers, Bagging and Gradient Boosting classifiers were selected based on their weighted average and per class precision on the cybercrime-related categories. Both models were used to classify 100,000 uncategorised entities, showing that the share of cybercrime-related is 29.81% according to Bagging, and 10.95% according to Gradient Boosting with number of entities as the metric. With regard to the number of addresses and current coins held by this type of entities, the results are: 5.79% and 10.02% according to Bagging; and 3.16% and 1.45% according to Gradient Boosting.

Authors train a classifier on a labelled dataset of Bitcoin entities, some of which are engaged in fraudulent activities. Authors then go on to estimate the proportion of illicit activity in the Bitcoin market by using the classifier on unclassified entities. An application to real estate of this approach could include attempting to estimate the amount of illicit activity that occurs in the housing market, especially in light of recent reports of the high level of criminal activity in British Colombia's real estate market [Maloney et al., 2019].

3.4.4 Retention Futility: Targeting High-Risk Customers Might Be Ineffective [Ascarza, 2018]

Abstract: *Companies in a variety of sectors are increasingly managing customer churn proactively, generally by detecting customers at the highest risk of churning and targeting retention efforts towards them. While there is a vast literature on developing churn prediction models that identify customers at the highest risk of churning, no research has investigated whether it is indeed optimal to target those individuals. Combining two field experiments with machine learning techniques, the author demonstrates that customers identified as having the highest risk of churning are not necessarily the best targets for proactive churn programs. This finding is not only contrary to common wisdom but also suggests that retention programs are sometimes futile not because firms offer the wrong incentives but because they do not apply the right targeting rules. Accordingly, firms should focus their modeling efforts on identifying the observed heterogeneity in response to the intervention and to target customers on the basis of their sensitivity to the intervention, regardless of their risk of churning. This approach is empirically demonstrated to be significantly more effective than the standard practice of targeting customers with the highest risk of churning. More broadly, the author encourages firms and researchers using randomized trials (or A/B tests) to look beyond the average effect of interventions and leverage the observed heterogeneity in customers' response to select customer targets.*

Author uses uplift random forests to estimate customer response to a promotion. Tests using this metric (i.e. sensitivity to intervention) as opposed to targeting customers based on those at the highest risk of churning. Suggests it is important for firms to conduct field experiments to test customer response to retention incentives. An application to real estate could include modelling which tenants are most likely to leave a rental building.

3.4.5 Homogeneous Feature Transfer and Heterogeneous Location Fine-Tuning [Guo et al., 2018]

Abstract: *Most existing real estate appraisal methods focus on building accuracy and reliable models from a given dataset but pay little attention to the extensibility of their trained model. As different cities usually contain a different set of location features (district names, apartment names), most existing mass appraisal methods have to train a new model from scratch for different cities or regions. As a result, these approaches require massive data collection for each city and the total training time for a multi-city property appraisal system will be extremely long. Besides, some small cities may not have enough data for training a robust appraisal model. To overcome these limitations, we develop a novel Homogeneous Feature Transfer and Heterogeneous Location Fine-tuning (HFT+HLF) cross-city property appraisal framework. By transferring partial neural network learning from a source city and fine-tuning on the small amount of location information of a target city, our semi-supervised model can achieve similar or even superior performance compared to a fully supervised Artificial neural network (ANN) method.*

Authors cite insufficiency of current econometric models which assume the value of property attributes is constant. Authors note that to transfer trained models from one city to another one can transfer what a machine learning system learns with respect to building characteristics, but not location characteristics (since each city has different districts/residential communities). The model used includes two separate neural networks, one which is transferable, which takes the homogeneous (i.e. apartment level) features as the inputs. The output of this neural network is then fed into a separate neural network alongside the heterogeneous location characteristics to ultimately predict a price. The authors use the first neural net across cities, and then fine tune the second one on each city (with considerably less data). Model achieves similar or superior performance to models without transfer learning.

3.4.6 Predictive Modeling of Surveyed Property Conditions and Vacancy [Martin et al., 2016]

Abstract: *Using the results of a comprehensive in-person survey of properties in Cleveland, Ohio, we fit predictive models of vacancy and property conditions. We draw predictor variables from administrative data that is available in most jurisdictions such as deed recordings, tax assessor's property characteristics, and foreclosure filings. Using logistic regression and machine learning methods, we are able to make reasonably accurate out-of-sample predictions. Our findings indicate that housing professionals could use administrative data and predictive models to identify distressed properties between surveys or among nonsurveyed properties in an area subject to a random sample survey.*

Authors apply machine learning techniques to predict findings of in-person property surveys, specifically whether a property is vacant or distressed, based on administrative data. Algorithms employed include: logistic regression, random forests, and gradient boosting. Authors find that the most informative variables in the model were complaints from neighbours and the tax assessor's record of the property's condition.

3.4.7 Human Decisions and Machine Predictions [Kleinberg et al., 2017]

Abstract: *Can machine learning improve human decision making? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released. The concreteness of the prediction task combined with the volume of data available makes this a promising machine-learning application. Yet comparing the algorithm to judges proves complicated. First, the available data are generated by prior judge decisions. We only observe crime outcomes for released defendants, not for those judges detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. Second, judges may have a broader set of preferences than the variable the algorithm predicts; for instance, judges may care specifically about violent crimes or about racial inequities. We deal with these problems using different econometric strategies, such as quasi-random assignment of cases to judges. Even accounting for these concerns, our results suggest potentially large welfare gains: one policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates. Moreover, all categories of crime, including violent crimes, show reductions; these gains can be achieved while simultaneously reducing racial disparities. These results suggest that while machine learning can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.*

Machine learning's ability to find complex structures/patterns in data make implementing data-driven decision aids increasingly attractive. Authors use gradient boosted trees on dataset of bail decisions in NYC between 2008-2013. Predict whether defendants will be rearrested or attempt to flee. Find judges struggle most with high risk cases. Identifies one issue of using machine learning is the selective labels problem where labels missing in a non-random way. Another lesson is importance of considering decision maker's full payoff function; bad performance may reflect different goals. An application to real estate could be assessing buyers' ability to assess the true value of houses with unique features (i.e. 'high risk' cases).

3.4.8 Machine Learning for Survival Analysis: A Survey [Wang et al., 2017]

Abstract: *Accurately predicting the time of occurrence of an event of interest is a critical problem in longitudinal data analysis. One of the main challenges in this context is the presence of instances whose event outcomes become unobservable after a certain time point or when some instances do not experience any event during the monitoring period. Such a phenomenon is called censoring which can be effectively handled using survival analysis techniques. Traditionally, statistical approaches have been widely developed in the literature to overcome this censoring issue. In addition, many machine learning algorithms are adapted to effectively handle survival data and tackle other challenging problems that arise in real-world data. In this survey, we provide a comprehensive and structured review of the representative statistical methods along with the machine learning techniques used in survival analysis and provide a detailed taxonomy of the existing methods. We also discuss several topics that are closely related to survival analysis and illustrate several successful applications in various real-world application domains. We hope that this paper will provide a more thorough understanding of the recent advances in survival analysis and offer some guidelines on applying these approaches to solve new problems that arise in applications with censored data.*

Authors provide overview of survival analysis and how machine learning can be incorporated into it. Examples include survival trees, Bayesian methods, neural networks, SVMs, and random survival forests. Authors mention duration analysis in economics as one area where survival analysis is successfully applied. An application to real estate could be duration that a house sits on the market before being sold.

3.5 Measuring Price Changes

This section looks at using machine learning methods to measure what covariates influence real estate prices. Machine learning is particularly well-suited for this area of research due to its ability to uncover non-linear relationships.

3.5.1 Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach [Chin et al., 2018]

Abstract: *Urban rail transit investments are expensive and irreversible. As people differ with respect to their demand for trips, their value of time, and the types of real estate they live in, such projects are likely to offer heterogeneous benefits to residents of a city. Defining the opening of a major new subway in Seoul as a treatment for apartments close to the new rail stations, we contrast hedonic estimates based on multivariate hedonic methods with a machine learning approach. This machine learning approach yields new estimates of these heterogeneous effects. While a majority of the "treated" apartment types appreciate in value, other types decline in value. We cross-validate our estimates by studying what types of new housing units developers build in the treated areas close to the new train lines.*

Authors build upon previous studies which use panel estimation strategies to recover estimates of causal effects of new transit access on housing prices. Authors use machine learning (regression trees specifically) to test how houses appreciate near a new subway relative to other houses, allowing effects to vary based on unit-level and community-level factors. As a validation of the machine learning model the paper tests whether new construction near the subway contains features that the machine learning model predicts will have the highest marginal revenue (that is the type of units that benefit most from being close to the subway).

3.5.2 Learning from Man or Machine: Spatial Fixed Effects in Urban Econometrics [Sommervoll and Sommervoll, 2019]

Abstract: *Econometric models with spatial fixed effects (FE) require some kind of spatial aggregation. This aggregation may be based on postcode, school district, county or some other spatial subdivision. Common sense would suggest that the less aggregated, the better inasmuch as aggregation over larger areas tends to gloss over systematic spatial variation. On the other hand, low spatial aggregation results in thin data sets and potentially noisy spatial fixed effects. We show, however, how this trade-off can be substantially lessened if we allow for more flexible aggregations. The key insight is that if we aggregate over areas with similar location premiums, we obtain robust location premiums without glossing over too*

much of the spatial variation. We use machine learning in the form of a genetic algorithm to identify areas with similar location premiums. The best aggregations found by the genetic algorithm outperform a conventional FE by postcode, even with an order of magnitude fewer spatial controls. This opens the door for spatially sparse FEs, if economy in the number of variables is important. The major takeaway, however, is that the genetic algorithm can find spatial aggregations that are both refined and robust, and thus significantly, lessen the trade-off between robust and refined location premium estimates.

Generally when fitting a spatial fixed effects model one aggregates by block, census tract, etc. There is a trade-off between very coarse aggregation and fine aggregation (e.g. coarse aggregation gives robust estimate of average location premium, that is the residual for the fixed effects model, but may miss some finer spatial variation). Using machine learning, specifically genetic algorithms, the authors create a flexible aggregation to find areas spatially far apart, but with similar location premiums (i.e. belonging to the same submarket). At a high level this genetic algorithm begins by randomly creating many models with each one assigning each area to one of 12 submarkets (down from 53 which is the total number of submarkets from two digit Oslo postcodes). The best submarket assignments are kept, and mixed together with some lower ranked assignments. The resulting model gives an out-of-sample performance similar to a model which uses dummy variables for all 53 submarkets. A similar analysis is run which uses a grid instead of postal codes as a spatial aggregator. The authors find that performance increases up to a point as the grid becomes finer, before levelling off.

3.6 Variable Selection

Variable selection, typically done in machine learning via regularization, is frequently used for problems where the dataset and number of covariates is large. This section looks at some methods that could be applicable to variable selection in real estate.

3.6.1 Choosing among Regularized Estimators in Empirical Economics: The Risk of Machine Learning [Abadie and Kasy, 2018]

Abstract: *Many settings in empirical economics involve estimation of a large number of parameters. In such settings methods that combine regularized estimation and data-driven choices of regularization parameters are useful. We provide guidance to applied researchers on (i) the choice between regularized estimators and (ii) data-driven selection of regularization parameters. We characterize the risk and relative performance of regularized estimators as a function of the data generating process, and show that data-driven choices of regularization parameters yield estimators with risk uniformly close to the risk attained under the optimal (unfeasible) choice of regularization parameters. We illustrate using examples from empirical economics.*

The authors look at which regularization methods to use in an applied economics setting. The authors find that there is not one method that is universally optimal. The authors illustrate different regularization methods in three applications: a study examining the effect of locations on intergenerational mobility, an event study for weapon-producing companies' stocks, and estimating the Mincer equation with survey data. An application to real estate could be dimensionality reduction through variable selection in a hedonic pricing model.

3.6.2 How to Separate the Wheat from the Chaff: Improved Variable Selection for New Customer Acquisition [Tillmanns et al., 2017]

Abstract: *Steady customer losses create pressure for firms to acquire new accounts, a task that is both costly and risky. Lacking knowledge about their prospects, firms often use a large array of predictors obtained from list vendors, which in turn rapidly creates massive high-dimensional data problems. Selecting the appropriate variables and their functional relationships with acquisition probabilities is therefore a substantial challenge. This study proposes a Bayesian variable selection approach to optimally select targets for new customer acquisition. Data from an insurance company reveal that this approach outperforms nonselection methods and selection methods based on expert judgment as well as benchmarks based on principal component analysis and bootstrap aggregation of classification trees. Notably, the optimal results show that the Bayesian approach selects panel-based metrics as predictors, detects several nonlinear relationships, selects very large numbers of addresses, and generates profits. In a series of post*

hoc analyses, the authors consider prospects' response behaviors and cross-selling potential and systematically vary the number of predictors and the estimated profit per response. The results reveal that more predictors and higher response rates do not necessarily lead to higher profits.

Authors use a Bayesian variable selection model and compare it with standard parametric/non-parametric models for selecting prospects. Issues commonly encountered in this field include variables from many sources, some of which are highly correlated; this leads to need for good variable selection methods. Authors use Bayesian variable selection with spike-and-slab prior, and model compared against classification trees. An application to real estate could be using spike-and-slab priors for dimension reduction in hedonic pricing models.

3.6.3 Hierarchical Penalization [Szafranski et al., 2008]

Abstract: *Hierarchical penalization is a generic framework for incorporating prior information in the fitting of statistical models, when the explicative variables are organized in a hierarchical structure. The penalizer is a convex functional that performs soft selection at the group level, and shrinks variables within each group. This favors solutions with few leading terms in the final combination. The framework, originally derived for taking prior knowledge into account, is shown to be useful in linear regression, when several parameters are used to model the influence of one feature, or in kernel regression, for learning multiple kernels.*

Authors conduct variable selection where variables can be split into several distinct groups. In this paper the authors perform soft selection at the group level, and then shrink variables within groups to favour solutions with fewer terms within each group. Authors apply method to Delve Census dataset, predicting median house prices based on demographic variables (final dataset included 37 variables in 10 groups). In general the hierarchical penalization performed better than LASSO. Since many real estate variables could fall into groups (e.g. square footage, lot size, number of rooms) this method could be applicable.

3.7 General Overviews of Machine Learning in Economics

This section includes articles that provide broad overviews of how machine learning is used in economic settings.

3.7.1 The Impact of Machine Learning on Economics [Athey, 2018]

Abstract: *This paper provides an assessment of the early contributions of machine learning to economics, as well as predictions about its future contributions. It begins by briefly overviewing some themes from the literature on machine learning, and then draws some contrasts with traditional approaches to estimating the impact of counterfactual policies in economics. Next, we review some of the initial "off-the-shelf" applications of machine learning to economics, including applications in analyzing text and images. We then describe new types of questions that have been posed surrounding the application of machine learning to policy problems, including "prediction policy problems," as well as considerations of fairness and manipulability. We present some highlights from the emerging econometric literature combining machine learning and causal inference. Finally, we overview a set of broader predictions about the future impact of machine learning on economics, including its impacts on the nature of collaboration, funding, research tools, and research questions.*

Broad overview of areas where machine learning is, or has the potential to, have an impact on the field of economics. First theme of the paper is that machine learning does not help much estimating causal effects, but can help with semi-parametric estimation when lots of data exists in a dataset. The second theme is data-driven model selection which helps avoid p-hacking. The third theme is that machine learning works well for simple classification or prediction tasks, but these are problems not generally tackled by economists who are more interested in causal inference. The fourth theme is that many algorithms must be modified in order to provide valid CIs when model selected in data-driven way. Finally, the author comments on how machine learning poised to change how economics research is actually conducted.

3.7.2 Machine Learning Methods Economists Should Know About [Athey and Imbens, 2019]

Abstract: *We discuss the relevance of the recent Machine Learning (ML) literature for economics and econometrics. First we discuss the differences in goals, methods and settings between the ML literature and the traditional econometrics and statistics literatures. Then we discuss some specific methods from the machine learning literature that we view as important for empirical researchers in economics. These include supervised learning methods for regression and classification, unsupervised learning methods, as well as matrix completion methods. Finally, we highlight newly developed methods at the intersection of ML and econometrics, methods that typically perform better than either off-the-shelf ML or more traditional econometric methods when applied to particular classes of problems, problems that include causal inference for average treatment effects, optimal policy estimation, and estimation of the counterfactual effect of price changes in consumer choice models.*

Authors go through a variety of machine learning methods in the supervised, unsupervised, causal inference, reinforcement learning, matrix completion, and text analysis domains.

3.7.3 Artificial Intelligence, Economics, and Industrial Organization [Varian, 2018]

Abstract: *Machine learning (ML) and artificial intelligence (AI) have been around for many years. However, in the last 5 years, remarkable progress has been made using multilayered neural networks in diverse areas such as image recognition, speech recognition, and machine translation. AI is a general purpose technology that is likely to impact many industries. In this chapter I consider how machine learning availability might affect the industrial organization of both firms that provide AI services and industries that adopt AI technology. My intent is not to provide an extensive overview of this rapidly-evolving area, but instead to provide a short summary of some of the forces at work and to describe some possible areas for future research.*

Chief economist of Google provides an overview of how AI poised to impact firms providing AI services and industries adopting AI. Mentions data's decreasing returns to scale. The author also discusses differential pricing being a promising area of application.

3.7.4 How Artificial Intelligence and Machine Learning Can Impact Market Design [Milgrom and Tadelis, 2018]

Abstract: *In complex environments, it is challenging to learn enough about the underlying characteristics of transactions so as to design the best institutions to efficiently generate gains from trade. In recent years, Artificial Intelligence has emerged as an important tool that allows market designers to uncover important market fundamentals, and to better predict fluctuations that can cause friction in markets. This paper offers some recent examples of how Artificial Intelligence helps market designers improve the operations of markets, and outlines directions in which it will continue to shape and influence market design.*

The authors cite examples from online marketplaces where 'grade inflation' occurs for seller ratings. AI can be used to assess the quality of sellers via NLP of text reviews on websites, and can design mechanisms to encourage users to leave feedback. Also discusses using AI to reduce search frictions. Particularly this can be done through automatically learning a customer's intent, and also better segmenting customers.

3.7.5 Big Data: New Tricks for Econometrics [Varian, 2014]

Abstract: *Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools. First, the sheer size of the data involved may require more powerful data manipulation tools. Second, we may have more potential predictors than appropriate for estimation, so we need to do some kind of variable selection. Third, large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships. In this essay, I will describe a few of these tools for manipulating and analyzing big data. I believe that*

these methods have a lot to offer and should be more widely known and used by economists.

Author identifies causality as one of the main areas where machine learning can contribute to econometrics. Mainly by creating a strong predictive model, and conducting an experiment with some treatment, one can take the difference between the actual and estimated outcome to forecast a causal effect. As well the author notes that the averaging over many machine learning models generally produces a better out-of-sample prediction than a single model; this method helps with model uncertainty, which is sometimes large.

References

Alberto Abadie and Maximilian Kasy. Choosing among Regularized Estimators in Empirical Economics: The Risk of Machine Learning. *The Review of Economics and Statistics*, 2018. doi: 10.1162/rest_a_00812. URL https://doi.org/10.1162/rest_a_00812.

Diego Ardila, Peter Cauwels, Dorsa Sanadgol, and Didier Sornette. Is There A Real Estate Bubble in Switzerland? Papers, arXiv.org, 2013. URL https://EconPapers.repec.org/RePEc:arx:papers:1303.4514.

Eva Ascarza. Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018. doi: 10.1509/jmr.16.0163. URL https://doi.org/10.1509/jmr.16.0163.

Brian J Asquith, Judith K Hellerstein, Mark J Kutzbach, and David Neumark. Social Capital and Labor Market Networks. Working Paper 23959, National Bureau of Economic Research, October 2017. URL http://www.nber.org/papers/w23959.

Susan Athey. *The Impact of Machine Learning on Economics*, pages 507–547. University of Chicago Press, January 2018. doi: https://doi.org/10.7208/chicago/9780226613475.001.0001. URL http://www.nber.org/chapters/c14009.

Susan Athey and Guido Imbens. Machine Learning Methods Economists Should Know About. Papers, arXiv.org, 2019. URL https://EconPapers.repec.org/RePEc:arx:papers:1903.10075.

Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. Machine Learning Methods for Demand Estimation. *American Economic Review*, 105(5):481–85, May 2015. doi: 10.1257/aer.p20151021. URL http://www.aeaweb.org/articles?id=10.1257/aer.p20151021.

Bitty Balducci and Detelina Marinova. Unstructured Data in Marketing. *Journal of the Academy of Marketing Science*, 46(4):557–590, Jul 2018. ISSN 1552-7824. doi: 10.1007/s11747-018-0581-x. URL https://doi.org/10.1007/s11747-018-0581-x.

Marianne Bertrand and Emir Kamenica. Coming Apart? Cultural Distances in the United States Over Time. Working Paper 24771, National Bureau of Economic Research, June 2018. URL http://www.nber.org/papers/w24771.

Bryan Bollinger, Jesse Burkhardt, and Kenneth Gillingham. Peer Effects in Water Conservation: Evidence from Consumer Migration. Working Paper 24812, National Bureau of Economic Research, July 2018. URL http://www.nber.org/papers/w24812.

Seungwoo Chin, Matthew E. Kahn, and Hyungsik Roger Moon. Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach. *Real Estate Economics*, 0(0), 2018. doi: 10.1111/1540-6229.12249. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.12249.

Dave Donaldson and Adam Storeygard. The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives*, 30(4):171–98, November 2016. doi: 10.1257/jep.30.4.171. URL http://www.aeaweb.org/articles?id=10.1257/jep.30.4.171.

Juan C. Duque, Jorge E. Patino, and Alejandro Betancourt. Exploring the Potential of Machine Learning for Automatic Slum Identification from VHR Imagery. *Remote Sensing*, 9(9), 2017. ISSN 2072-4292. doi: 10.3390/rs9090895. URL https://www.mdpi.com/2072-4292/9/9/895.

James J. Feigenbaum. Automated Census Record Linking: a Machine Learning Approach, 2016. URL https://open.bu.edu/handle/2144/27526.

Sofia F. Franco and Jacob L. Macdonald. Measurement and Valuation of Urban Greenness: Remote Sensing and Hedonic Applications to Lisbon, Portugal. *Regional Science and Urban Economics*, 72:156 – 180, 2018. ISSN 0166-0462. doi: https://doi.org/10.1016/j.regsciurbeco.2017.03.002. URL http://www.sciencedirect.com/science/article/pii/S0166046216302708. New Advances in Spatial Econometrics: Interactions Matter.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably Unequal? The Effects of Machine Learning on Credit Markets. 2017.

Matthew Gentzkow, Bryan T Kelly, and Matt Taddy. Text as Data. Working Paper 23276, National Bureau of Economic Research, March 2017. URL http://www.nber.org/papers/w23276.

Edward L Glaeser, Michael Scott Kincaid, and Nikhil Naik. Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks. Working Paper 25174, National Bureau of Economic Research, October 2018. URL http://www.nber.org/papers/w25174.

Yihan Guo, Shan Lin, Xiao Ma, Jay Bal, and Chang-Tsun Li. Homogeneous Feature Transfer and Heterogeneous Location Fine-tuning for Cross-City Property Appraisal Framework. *CoRR*, abs/1812.05486, 2018. URL http://arxiv.org/abs/1812.05486.

Kathleen Weiss Hanley and Gerard Hoberg. Dynamic Interpretation of Emerging Risks in the Financial Sector. *The Review of Financial Studies*, 02 2019. ISSN 0893-9454. doi: 10.1093/rfs/hhz023. URL https://doi.org/10.1093/rfs/hhz023.

Jochen Hausler, Jessica Ruscheinsky, and Marcel Lang. News-based Sentiment Analysis in Real Estate: a Machine Learning Approach. *Journal of Property Research*, 35(4):344–371, 2018. doi: 10.1080/09599916.2018.1551923. URL https://doi.org/10.1080/09599916.2018.1551923.

Nabil Hossain, Tianran Hu, Roghayeh Feizi, Ann Marie White, Jiebo Luo, and Henry A. Kautz. Precise Localization of Homes and Activities: Detecting Drinking-While-Tweeting Patterns in Communities. In *ICWSM*, 2016.

Brad R. Humphreys, Adam Nowak, and Yang Zhou. Superstition and Real Estate Prices: Transaction-level Evidence from the US Housing Market. *Applied Economics*, 51(26):2818–2841, 2019. doi: 10.1080/00036846.2018.1558361. URL https://doi.org/10.1080/00036846.2018.1558361.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 08 2017. ISSN 0033-5533. doi: 10.1093/qje/qjx032. URL https://doi.org/10.1093/qje/qjx032.

Jing Lei. Differentially Private M-Estimators. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 361–369. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4376-differentially-private-m-estimators.pdf.

Michele Loberto, Andrea Luciani, and Marco Pangallo. The Potential of Big Housing Data: an Application to the Italian Real-estate Market. Temi di discussione (Economic working papers) 1171, Bank of Italy, Economic Research and International Relations Area, April 2018. URL https://ideas.repec.org/p/bdi/wptemi/td_1171_18.html.

Maureen Maloney, Tsur Somerville, and Brigitte Unger. *Combatting Money Laundering in BC Real Estate*. Mar 2019. URL https://www2.gov.bc.ca/gov/content/housing-tenancy/real-estate-bc/consultations/money-laundering.

Asaf Manela and Alan Moreira. News Implied Volatility and Disaster Concerns. *Journal of Financial Economics*, 123(1):137 – 162, 2017. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2016.01.032. URL http://www.sciencedirect.com/science/article/pii/S0304405X16301751.

Hal Martin, Isaac Oduro, Francisca Richter, Apirl Hirsh Urban, and Stephan Whitaker. Predictive Modeling of Surveyed Property Conditions and Vacancy. Working Papers (Old Series) 1637, Federal Reserve Bank of Cleveland, December 2016. URL https://ideas.repec.org/p/fip/fedcwp/1637.html.

Paul R Milgrom and Steven Tadelis. How Artificial Intelligence and Machine Learning Can Impact Market Design. Working Paper 24282, National Bureau of Economic Research, February 2018. URL http://www.nber.org/papers/w24282.

Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Do People Shape Cities, or Do Cities Shape People? The Co-evolution of Physical, Social, and Economic Change in Five Major U.S. Cities. Working Paper 21620, National Bureau of Economic Research, October 2015. URL http://www.nber.org/papers/w21620.

Adam Nowak and Juan Sayago-Gomez. Homeowner Preferences After September 11th, a Microdata Approach. *Regional Science and Urban Economics*, 70:330 – 351, 2018. ISSN 0166-0462. doi: https://doi.org/10.1016/j.regsciurbeco.2017.10.001. URL http://www.sciencedirect.com/science/article/pii/S0166046217302661.

Adam Nowak and Patrick Smith. Textual Analysis in Real Estate. *Journal of Applied Econometrics*, 32(4):896–918, 2017. doi: 10.1002/jae.2550. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2550.

Omid Poursaeed, Tomas Matera, and Serge J. Belongie. Vision-based Real Estate Price Estimation. *CoRR*, abs/1707.05489, 2017. URL http://arxiv.org/abs/1707.05489.

Avald Sommervoll and Dag Einar Sommervoll. Learning from Man or Machine: Spatial Fixed Effects in Urban Econometrics. *Regional Science and Urban Economics*, 2019. ISSN 0166-0462. doi: https://doi.org/10.1016/j.regsciurbeco.2019.04.005. URL http://www.sciencedirect.com/science/article/pii/S016604621830303X.

H. Sun Yin and R. Vatrapu. A First Estimation of the Proportion of Cybercriminal Entities in the Bitcoin Ecosystem Using Supervised Machine Learning. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3690–3699, Dec 2017. doi: 10.1109/BigData.2017.8258365.

Marie Szafranski, Yves Grandvalet, and Pierre Morizet-mahoudeaux. Hierarchical Penalization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1457–1464. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3338-hierarchical-penalization.pdf.

Sebastian Tillmanns, Frenkel Ter Hofstede, Manfred Krafft, and Oliver Goetz. How to Separate the Wheat from the Chaff: Improved Variable Selection for New Customer Acquisition. *Journal of Marketing*, 81(2):99–113, 2017. doi: 10.1509/jm.15.0398. URL https://doi.org/10.1509/jm.15.0398.

Hal Varian. Artificial Intelligence, Economics, and Industrial Organization. Working Paper 24839, National Bureau of Economic Research, July 2018. URL http://www.nber.org/papers/w24839.

Hal R. Varian. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28, May 2014. doi: 10.1257/jep.28.2.3. URL http://www.aeaweb.org/articles?id=10.1257/jep.28.2.3.

Ping Wang, Yan Li, and Chandan K. Reddy. Machine Learning for Survival Analysis: A Survey. *CoRR*, abs/1708.04649, 2017. URL http://arxiv.org/abs/1708.04649.

Dafeng Xu. Surname-based Ethnicity and Ethnic Segregation in the Early Twentieth Century U.S. *Regional Science and Urban Economics*, 77:1 – 19, 2019. ISSN 0166-0462. doi: https://doi.org/10.1016/j.regsciurbeco.2019.01.005. URL http://www.sciencedirect.com/science/article/pii/S0166046218302114.

Yu Ye, Hanting Xie, Jia Fang, Hetao Jiang, and De Wang. Daily Accessed Street Greenery and Housing Price: Measuring Economic Performance of Human-Scale Streetscapes via New Urban Data. *Sustainability*, 11(6), 2019. ISSN 2071-1050. doi: 10.3390/su11061741. URL https://www.mdpi.com/2071-1050/11/6/1741.

Quanzeng You, Ran Pang, and Jiebo Luo. Image Based Appraisal of Real Estate Properties. *CoRR*, abs/1611.09180, 2016. URL http://arxiv.org/abs/1611.09180.

Danning Zheng, Tianran Hu, Quanzeng You, Henry Kautz, and Jiebo Luo. Inferring Home Location from User's Photo Collections Based on Visual Content and Mobility Patterns. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '14, pages 21–26, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3127-2. doi: 10.1145/2661118.2661123. URL http://doi.acm.org/10.1145/2661118.2661123.