22ND ANNUAL PACIFIC-RIM REAL ESTATE SOCIETY CONFERENCE
SUNSHINE COAST, QUEENSLAND, AUSTRALIA 17-20 JANUARY 2016

TRANSFORMATIONS IN REAL ESTATE RESEARCH: THE BIG DATA REVOLUTION

Kimberly Winson-Geideman and Andy Krause

University of Melbourne

THEME: General (Technology)

STRUCTURED ABSTRACT

Problem/Purpose: Recent developments in technology and improved data gathering techniques have brought many changes to the evaluation of real estate from applied and theoretical perspectives. Large volumes of data are being collected, transformed and analysed to predictively assess market trends and evaluate consumer sentiment by companies such as RealEstate.com.au (Australia), Zillow (US) and Zoopla (UK) as well as researchers around the world. Social media and search engines such as Twitter, Google and Facebook have begun harnessing the data available on their respective sites and are using it to identify trends and patterns among users. The use of Big Data, as it is commonly known, is at the cutting edge and has the potential to transform decision-making in real estate.

Big Data, however, varies greatly in both scale and type by industry and much of the broader discussion ignores issues critical to the analysis of property. This paper introduces the reader to the what, why, and how of Big Data in a property context—what it is, why is it important, and how it is transforming real estate research. We look at recently published articles in academic journals to provide specific instances of how Big Data is used, and we also identify potential issues with sharing, analysis, and interpretation.

Design/methodology/approach: The methodology is a case study of current literature on Big Data applications.

Findings: The findings show that Big Data is having an increasing impact on the evaluation and understanding of real estate markets. We create a typology of real estate data including Core, Static Spatial and Peripheral

Research limitations/implications: While many of the initial applications of Big Data in the property sphere have focused on valuation and residential property sales, real estate investment and development are primed to benefit (as well as suffer disruption) from the explosion of data and the growth of advanced analytical techniques.

Takeaway for practice: This paper introduces the reader to Big Data applications that can and are being used by real estate practitioners.

Originality/value: To our knowledge this is the first research that comprehensively addresses Big Data applications and techniques in the context of real estate.

Keywords: big data, technology, real estate research

Email contact: kimberly.winson@unimelb.edu.au

INTRODUCTION

Data is a key component – perhaps the key component – of our modern economy. This reliance on data has pervaded our personal and social lives as well. As tools and systems have been developed over the past decade to more efficiently collect, store, manage and analyse data, the sheer volume of digital information has expanded exponentially. Once confined to stand-alone servers and databases, these disparate sources of information are increasingly being linked together to offer the possibility for greater understanding, surveillance and recording than ever before. The moniker 'Big Data' has arisen as a catch-all term to describe the large, interconnected databases that are being developed as well as the many of the processes that seek to distil useful knowledge from the flood of digital data that exists.

The precise definition of 'Big Data' is not readily agreed upon (Crampton et al. 2013) and opinions of what constitutes it vary widely across disciplines and major players in the field (Ward & Barker 2013). In this paper we avoid, to the extent possible, the nuanced, semantic debate on what is, is not or should be 'Big Data'.[1] Instead, we will aim to provide a broad survey of potential classes, properties, uses and types of Big Data with a bias toward inclusiveness and generality. The goal here is to introduce the broad concept of Big Data and to explore the extent to which current real estate research utilizes Big Data and where, if any, there are additional possibilities for its expansion into the real estate discipline – in research, industry and education.

To provide background, we begin by discussing some of the more common definitions that are currently applied to Big Data as well as some of its characteristics (good and bad). With Big Data in general as a framework, we then take a specific look at real estate data to see how existing information sources in real estate fit into the Big Data scheme. Particular focus is given to the evolution of real estate data over time as well as the idiosyncrasies of the real estate industry in regards to data-driven analysis. Following this, we offer a survey of existing uses of Big Data in the real estate discipline. We conclude with thoughts regarding the future of Big Data in real estate analysis and education.

WHAT IS BIG DATA?

Big Data is new, novel, and exciting. It is characteristically complex and not without fault, yet it is a phenomenon that needs to be fully explored as a great deal of emphasis has been placed on the anticipated benefits of mining data that will generate new understanding of human behaviour, lifestyle and habits.

Madden (2012) neatly summarises Big Data as data that is "too big, too fast, or too hard for existing tools to process"— too big as organisations are now collecting petabytes of data, too fast as processing applications must provide nearly instantaneous results, and too hard when new technologies are required to analyse it. Boyd and Crawford (2012) liken the emergence of Big Data and the associated computational techniques to the system of mass production devised by Henry Ford. Ford's vision set the standard for technological process in his time, much as Big Data is setting the standard today. The information extracted from Big Data, they propose, has the potential to change the way we think about many different areas of research, from economics to the social sciences to medicine and disease.

As Madden indicates, Big Data refers not only to big sets of data—sets so large they are redefining the boundaries of knowledge—it also refers to changes in the computational tools that are used to analyse data (Burkholder, 1992). Therein lies the conundrum—access to the data is meaningless without the proper tools to manage, analyse and store it (Ward and Barker 2013; Jacobs 2009). New technologies are required as datasets ranging from terabytes to exabytes are often too large for traditional data processing systems.

[1] Throughout this article we will capitalise Big Data but will not use quotations.

Furthermore, the data is complex, including sources such as sensors and social media (Provost and Fawcett, 2013; Chen, Chiang, Storey, 2012). As important as analytics is the availability of experts to work with, analyse and provide reliable interpretations.

Technical characteristics of Big Data

The origins of Big Data are rather complex and the result of years of technological improvements in computing, analysis, data storage and, as some would argue, the introduction of social media. Doug Laney, an analyst at the META Group (now Gartner) was the first to coin the term when he recognised the unique opportunities and challenges associated with the increasing volumes of data being generated globally. In a 2001 research report Laney theorised that there were three things that characterised the growth of this new class of information, making it markedly different from conventional data—"'Big Data' is high-volume, -velocity and –variety information assets that demand cost-effective, innovative forms of processing for enhanced insight and decision making" (Laney, 2001). Laney's 3 Vs definition— volume, velocity, and variety—has morphed substantially to include other "Vs". Some researchers include veracity, validity, value and volatility as well, but it is the 4 Vs model—volume, velocity, variety and veracity—that is now widely recognised as the standard, technical definition of Big Data (IBM).

Any discussion of Big Data typically begins with volume (or scale of data), and if there's one thing we know, it's that Big Data isn't just big, it's really, really big. Approximately 2.5 Quintillion bytes of data are created in a day, and by 2020, the total amount of data is projected to equal 40 zettabytes (IBM). Large amounts of data are captured from everyday uses such as mobile phones, supermarket rewards cards, Internet search engines, and social media sites. Most US companies are storing about 100 terabytes of data and some much more. Consider Twitter, for instance. Twitter boasts about 302 million active monthly users, tweeting about 500 million times each day (Oreskovic, 2015), the equivalent of 183 billion tweets a year. It's the combined challenge of storing, managing and analysing all this information that makes Big Data big.

The second V, velocity, refers to the data that are constantly being created, termed "streaming data" or "complex event processing" (O'Reilly Media, 2012). There are two primary challenges associated with data streams: analysing the data in real-time for immediate use and batching and storing for later analysis. Imagine, for instance, the terabyte of information that flows into the New York Stock Exchange (NYSE) each day. Much of this information is instantaneously fed into algorithms that support high-frequency trading strategies. Investors rely on the immediate evaluation of data and specialised trading tools to conduct high volume trades and profit in the very short-term. The risk associated with this type of investing is quite high and not without issues. High frequency trading is thought to have contributed substantially to the Flash Crash of 6 May 2010 when equity investors rapidly pulled out of the market, exposing investors to significant losses in a matter of seconds.

Analysing streams on a real-time basis and storing streams for later research is of particular importance to online retail establishments. Amazon serves as an excellent example of this. When a shopper selects a certain item, immediate recommendations for other items are suggested based on the customer's purchase history, purchases made by other customers, and often the customer's Internet search history. The resulting product recommendations are based on both historic data and immediately collected information.

Variety refers to the different types of data that are collected including structured and unstructured data, audio, video, web applications and text. Structured data is data that conforms to a set of rules, e.g. bank statements, sale prices expressed as currency, dates—all follow a specific pattern. Unstructured data does not conform to any rules. Twitter, for instance, limits Tweets to 140 characters, however it does not limit the content of the Tweets, which can include photos, videos, weblinks and various combinations of characters. Each has different storage, scaling and analytical requirements, often providing a unique perspective on the same topic.

The veracity of Big Data has to do with its uncertainty—how accurate and dependable the data is. Controlling for abnormalities, bias and noise is just as important with Big Data as it is with conventional data sets, but can be more difficult to address given the scale of the data. Data cleaning needs to be at least as reliable as data collection and furthermore, researchers must ensure that the data being used is relevant to the question that is asked. For example, textual data such as Tweets, product descriptions on Ebay, and status updates on Facebook introduce problems with misspelled words, lack of information, informal language, acronyms and subjectivity. When a Tweet includes words such as "interest", "rate", "increase" and "home", is it referring to interest rate increases and home purchases, or is it referring to the rate of increased interest in home purchases? These issues are often difficult to address, requiring sophisticated software and analytics.

Other Characteristics of Big Data

Big Data and related techniques have evolved over a relatively short time period and as awareness has grown, so has the perceived potential. Boyd and Crawford (2012) address this in a recent paper by approaching the Big Data definition much more theoretically and from a far less technical perspective. Their focus is on analysis and expected benefits, and they posit that limiting discussions of what Big Data is to purely technical aspects misses the more important parts of it. They challenge many of the running assumptions that have been made, defining Big Data as the intersection of technology, culture and scholarly phenomenon, specifically

1. "Technology: maximising computational power and algorithmic accuracy to gather, analyse, link, and compare large data sets.
2. Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
3. Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."

It is the mythological that contributes primarily to the biases and assumptions commonly embedded in Big Data research, they argue, with some contribution from the others. The authors underscore the need for researchers to critically assess their understanding of Big Data as well as the intent and interpretation of their research. They present six common assumptions that emphasise the need for a critical framework for researchers to use to critically evaluate the contribution of Big Data to the greater body of knowledge. We summarise those assumptions here.

Big Data changes the definition of knowledge. The authors argue that the introduction of Big Data to research is not the paradigm shift in knowledge that so many believe it might be. Burkholder (1992) and Lazer (2009) respectively refer to Big Data as a "computational turn in thought and research" and as possessing "unprecedented breadth and depth and scale" (p. 722) as evidence of the dramatic potential of Big Data. Wired magazine's editor Chris Anderson (2008) goes as far as to suggest that enough data and computation will render scientific theory obsolete.

But is volume and scale alone enough to shift an entire research paradigm? Boyd and Crawford argue that it is not. Regardless of size, researchers are still bound by data and research limitations that force questions like: are the data appropriate for the question that is being asked, are the data reliable, has it been entered and processed correctly, and what are the limitations to the data? As Miller (2010) notes, true knowledge discovery requires the context and understanding that can only come from domain expertise. In short, the addition of the word "big" only changes the size; it's still "data" at its core.

Claims to objectivity and accuracy are misleading. Big Data, especially that collected from social media, have been perceived as giving social scientists (who, in the past were often limited to qualitative data), a chance to redefine their discipline quantitatively. No longer would the social sciences be relegated to second tier status as Big Data introduced objective, factual research, Boyd and Crawford argue that Big Data just provides another avenue for research, and not a particularly dependable one at that, as it is subject to the many bias limitations inherent in human discovery. Those who think that the results of quantitative research are objective fact and qualitative research, subjective fact, are mistaken. All research is subjective to a point, regardless of whether one is dealing with quantitative or qualitative data. Even data that is quantitative is collected, cleaned and interpreted—all of which can be deemed subjective exercises.

Bigger data are not always better data. Big Data does not always provide better, more insightful results just because of its size. Big Data is prone to many of the same methodological problems smaller sets are such as user input errors and sampling error. Big Data is also subject to issues not generally associated with conventional data including source identification, particularly given the automated nature of some Big Data sources (e.g. is it a human or a robot that is generating the information?). Big Data is often combined with other Big Data sets, which can serve to compound problems. Additionally, an interesting thing about most Big Data, as it currently exists, is that it captures those observations that are obvious and easy to record, such as transit boarding, mouse clicks and credit card transactions (Kitchen 2014). These are often end results, not motivation and causes. The 'why' of any phenomenon remains critically important to science and one that is not easily explained by mere correlations. Towards this same end, big, messy datasets are not always as useful as small, clean and well understood ones. Data curation is still an important, and often necessary, exercise (Sowe and Zettsu 2014). Furthermore, Big Data can't be perceived as the answer to everything. Data should be reflective of the question that is asked, and sometimes small data provide the best answer.

Taken out of context, Big Data loses its meaning. Real estate data are particularly vulnerable to this issue as markets are very localised and if the information extrapolated from large data sets is not reflective of local market conditions, then it is meaningless. Pinpointing exact data points can be difficult and just because information is collected *from* a certain place, it does not mean the data is *about* that place. Real estate markets have submarkets within markets making it difficult to precisely identify context.

Just because it is accessible does not make it ethical. The authors cite as an example a Harvard study that tracked the interests and friendships of 1,700 Facebook users who were unaware they were being observed. The data were anonymised and then publicly released for other researchers to access. It did not take long to realise that parts of the data could be de-anonymised, compromising the privacy of the users. This leads to several questions that need to be raised about publicly available data that can be attributed to an individual. With the advent of social media researchers have (in some cases) access to a large amount of personal information that is publicly accessible. While it may be legal to use, is it ethical? Public availability is not the equivalent of consent, so is it ethical to showcase public information on an individual without their express consent? Real estate data are often analysed at the level of the individual household though the results are released in the aggregate. Is this enough to ensure anonymity?

Limited access to Big Data creates new digital divides. In the Big Data realm, publicly visibility does not necessarily equate to public accessibility. Much of the data that is generated belongs to large corporations that are under no obligation to release what they have. Some companies completely restrict access to their data whilst others charge a fee for it. Those who do release their information are responsible for deciding who gets what and by doing so create the "haves" and "have-nots" in Big Data research. This introduces a great deal of bias into the findings that are generated from such data. If the data are only released to a select few, under what circumstances are they released? Is there a particular type of research that the company supports to the detriment of contradictory studies? What evidence are researchers within the company privy to that researchers outside are not? These are the types of questions that should be part of a critical evaluation of studies using Big Data.

Technical skills and understanding can create barriers as well, especially for solo or small groups of researchers. The Big Data sphere is highly dependent on advanced coding, database management and statistical or machine learning skills. These new skills are often not easily gained and render access to and analyse of Big Data all but impossible for researchers and team lacking advanced technical know-how. One silver lining to this problem could be that technical requirements of Big Data may engender (or force) new interdisciplinary research collaborations (Ruppert 2013; Arribas-Bel 2014). In such cases one group may provide the domain specific knowledge (sociologists) while another data collection and management skills (geographers and computer scientists) and a third the analytical capabilities (statisticians and applied mathematicians). Perhaps Big Data will finally force academics out of their silos and into cross-campus and industry-related cooperative projects.

ON REAL ESTATE (BIG) DATA

For many (most) individual consumers, real estate holds some degree of emotional attachment or reaction. And, as any agent or broker will admit, emotions can play a large role in determining real estate transactions. Nevertheless, real estate as an industry, academic discipline and asset class has always been centred on data – lease and vacancy rates, home price indices, lot sizes, interest rates, REIT returns, etc. How then, can Big Data provide anything new to a field well steeped in data collection, management analysis? To help answer this question, it is instructive to examine the traditional forms of real estate data, their interrelationships and their deficiencies.

Three typologies of real estate data

Real estate data has traditionally come in three broad types – 1) financial, 3) transactional and 2) physical. Financial data includes information on REIT shares and real estate-related stocks. Transactional data refers to information on real estate purchases, mortgages, leases, expenses, taxes and general financial returns to a development or portfolio of developments. Physical data, conversely, includes information as to the actual real estate land or structure itself, such as property structural characteristics and locational data. REITs and stocks are markedly different from the other broad types as they trade on a continuous basis on share markets such as the NYSE and Australian Stock Exchange. The other two are often combined or linked based on property addresses or parcel identification numbering systems. Tax assessors and local multiple listing services offer common examples where the two data types are combined and utilized in traditional real estate situations. These three types of data, financial, transactional and physical, can be considered 'Core' real estate data.

With the introduction and proliferation of Geographic Information Systems (GIS) in the 1990s, new forms of data could be efficiently utilized in the real estate industry. GIS allowed for extra-locational spatial data to be collected, quantified and analyzed. In this case, "extra-locational" means information on spatial phenomena outside of the physical boundaries of a property. This differs from the transactional and physical data discussed above which contains information on only the property(s) in question. With extra-locational information spatial analysis and business geography (Thrall 2002) became viable real estate undertakings and have slowly become the norm in real estate analysis.

Examples of extra-locational data include neighbourhood information from the census, traffic sheds and street flow patterns, analyses of proximity to amenities and disamenities (externalities), viewsheds, accessibility metrics, etc. In other words, this type of data quantifies how the property itself relates to existing and external physical realities. Prior to GIS, most analysis could only look quantitatively at the property itself (its own characteristics, financial returns, etc), while externalities and extra-locational information was consumed and analysed somewhat qualitatively (tacitly). In short, externalities were accounted for qualitatively and this made up part of the expertise of real estate professionals.

Extra-locational data has been around for nearly two decades and have continued to improve and expand over time. However, this data often remains somewhat limited in terms of temporal resolution (it is often updated slowly, i.e. is not real-time) and usually possesses a well-defined spatial extent – the census block group, the road segment or the school district boundary, for example. As such, this extra-locational data can best be considered as 'Static Spatial' (SS) data.

Whereas GIS broadened our abilities to measure and analyse the static, physical world surrounding our real estate, Big Data (as defined above) seeks to pull in the fugacious, human and psychological aspects as well. For example, where GIS and Static Spatial data allowed us to calculate the nearest distance to bus stops and public transit, government transportation agencies can measure exact ridership numbers by the minute and how these patterns are changing in near real time. GIS helped us understand demand for a neighbourhood based on an index of proximate amenities and disamenties, companies like Zillow and Trulia can count how many people searched for homes in a given neighbourhood last week (and from where they searched). Static Spatial data provided us with likely capture zones for shopping malls, now retailers transactional information systems know exactly (to ZIP/Postal Code or better) where shoppers are coming from and what they are buying.

Table 1 provides a list (certainly not exhaustive) of some common examples of what constitutes each of the three types of real estate data currently used or practically available to be used.

Table 1: Example of three data types

Core	Static Spatial	Peripheral
Sale Transactions	Census information	Internet Searches
Lease Transactions	Road Network Data	Transit Boarding Data
Mortgage Information	Geographic information	Live Traffic Information
Tax Assessment Values	Aggregated Spatial Core Data	Point of Sale (POS) Data
Property Physical Characteristics	Urban Planning Forecasts	Geo-Located Tweets
REIT & Real Estate Stock Data	Spatial Economic Indicators	Pedestrian Foot Counts

These novel data collection schemes, data management systems and data analysis companies organization represent a much wider and broad set of phenomena and entities than has been traditionally represented in real estate research. We label these new systems and sources as 'Peripheral' data – in contrast to Core and Static Spatial data. The Peripheral data is varied, disparate and often real-time. It may be human-focused instead of property-centred. It is often remotely-sensed (gathered mechanically) instead of directly collected by industry professionals. In short, it is a new form of information for use in real estate research and analysis.

In the real estate industry, a large portion of resources – manpower, data collection and analysis – are employed for the purposes of prediction or forecasting. Knowledge about what is likely (or unlikely) to happen in the future is incredibly powerful information for decision makers in real estate markets. With Core data, forecasting methods are purely statistical; i.e. they use past trends in Core data to predict what is likely to occur in the future. The additional of Static Spatial data allows for (cor)relational forecasting, or incorporating the influences of localized, external factors into the analysis. Finally, the incorporation of Peripheral data open up the potential to improve forecasting through the ability to observe and analyse basic, causal relationship – like the number of home searches to the demand for homes in an area or analysing Twitter sentiment to predict movements in REIT shares. This could be considered 'structural' forecasting and offers the potential for more accurate and timelier information on which to base real estate decisions.

Is Real Estate data 'Big'?

As discussed above, 'Big Data' may not be so much a matter of size, as it is of inter-connectivity, variation, veracity and the ability to derive novel information from it. While the case can be made that REIT and real estate-related stock data may qualify as Big Data, neither the rest of the Core nor Static Spatial data would likely meet any of the myriad of definitions. Even merged – as has been done for the past two decades – the case for `Big Data-ship` is better but still fairly weak. Where we see the biggest potential and the best case to be made for Big Data in real estate is when merging all three types, Core, Static Spatial and Peripheral into a searchable and analysable database or collection of databases.

There has been a slow, but steady, movement in real estate towards more connected sets of data. Most Core data is hand collected, meaning that at some point the initial data is/was hand entered. Often this process of data collection and entry falls to local agents, lenders and tax officials. This information, being highly localized, collected at great cost and of great value to the holders of it, was not always readily shared. However, as multiple listing services created an Internet presence and government entities like the Office of Federal Housing Enterprise and Oversight (OFHEO) in the United States became more involved in local lending the need for digitized data (Core) in standardized formats grew rapidly. While some of this information, such as access to Multiple Listing Service data, remained on a subscription fee basis, other such as tax assessor (statutory valuer) property characteristics data was released free to the public.

Industry analysts, agencies and interested third-parties were then able to combine this Core data with the growth in Static Spatial data spurred on by the rise of Geographic Information Systems as a standard tool. At the same time, data collectors and creators like the U.S Census Bureau (and its international equivalents), city and provincial governments as well as large industry players started producing geographically referenced datasets[2] to augment their traditional tabular outputs. Concurrently, large data collectors like CoreLogic began to assemble regional or nationwide datasets of Core data, and greatly increasing the data standardization process along the way.

Fast-forward to today and we see the continued use of Geographic Information Systems, evidenced by the fact that most county and/or state governments have free, interactive GIS applications available on the web. The creation of more consumer friendly products such as Google Earth and OpenStreetMap have also furthered this trend. Many Multiple Listing Services are expanding into regional providers like the Metropolitan Regional Information System (MRIS) operating on the east coast of the United States and providers like Corelogic and Black Knight have grown and represent near total coverage of the large U.S. market. Finally, many open Peripheral data sources like Twitter, Google, Four-Square, Instagram, as well as internal, industry ones such as Amazon shipping details and Zillow home searches, are geo-coded, meaning the location of the activity can be pinpointed relatively well. The connection of these Peripheral sources to Core and Static Spatial and just beginning; in that we see potential.

Who is currently using Big Data and how?

So, who is 'doing' Big Data in real estate? In other words, what researchers, firms and agencies are actually incorporating these new Peripheral data sources into real estate analyses?

On the academic side, the primary use to date of Big Data has been to forecast housing prices based on web searches and media sentiment. Analysing Google search trends in the United States, Wu and Brynjolfsson (2015) were able to build a model that out-performed the National Association of Realtors home price predictions by greater than 20%. Similar analyses in China using search terms and media output (Sun et al 2014) and Big Data computation methods such as Machine Learning (Mu et al. 2014) highlighted the promises of Big Data in predicting future movement in residential real estate markets. Additional work on the Chinese market by Wu and Deng (2015) showed how information and interest in various markets flows from China's superstar cities of Beijing, Shanghai and Shenzen down through secondary and tertiary locations.

There is currently a lack of research into Big Data's role in better understanding commercial and industrial markets as well as into real estate development or investment trends. While residential markets are often the easiest in which to gather large datasets on which to test hypothesis, other facets of real estate research could certainly benefit from expanded research into the potentials of Big Data.

In industry, the uses are far more widespread, but also – as a function of proprietary business –can be more difficult to identify who is using what and how. Home realty and information firms such as Redfin, Zillow and Trulia (US), Domain and RealEstate.com.au (Australia), and Zoopla (UK) offer some of the more common and obvious uses of Big Data in the real estate industry. Data aggregators such as CoreLogic (RPData) and Black Knight have been very successful at compiling, standardizing and selling the raw Big Data necessary for these firms to improve their day-to-day product. Somewhat related, Windemere Real

[2] Most commonly in the form of ESRI shapefiles (.shp) or, more recently, geodatabases (.gdb).

Estate (brokers) utilise traffic and road network information to provide buyers with detailed information on commute times from potential home purchases (Sonka 2014).

A number of the large Chinese developers, such as Vanke and Fantasia Group, leverage their existing databases of past buyers to optimize current marketing programs for new projects (Du et al 2014). For companies that have sold multiple millions of dwellings in the past decade, this can mean a substantial improvement in performance. Real Capital Analytics has employed a massive data gathering and analysing system to better understand the overall real estate market sentiment, both global and localized (Kudyba and Kwatinetz 2014). This information is then used to provide more accurate advice regarding potential real estate decisions made by its clients.

Overall, the uses of Big Data in real estate research are fledgling, but growing quickly as new data source become available and as researchers gain the skills (and often the connections) necessary to access and analyse the data. Within the industry itself, it can be difficult to tell what firms are using what data and techniques due to concerns over sharing this type of business-competitive information. Nonetheless, the few examples that have been made public do show the breadth and depth to which Big Data is being put to further the efficiency and profitability of real estate firms and ventures.

FUTURE THOUGHTS AND CONCLUSIONS

The debate over the definition of Big Data is, in most ways, a distraction from the underlying growth of the importance of data (Big or small) in both research and industry. As a result, we argue that the answer to the question of whether real estate data is Big Data depends on who is asking, and further, we also contend that the answer doesn’t matter. What does matter is that, like many other fields, data-driven research (both academic and industry) is the wave of the future, at least in some part. This means that whether or not what we do in real estate technically qualifies as Big Data is irrelevant, since we will use the same or similar: 1) techniques to collect data; 2) systems to manage data; 3) processes to clean and prepare data; 4) models and algorithms to analyse data; and 5) methods to visualize and present results.

The introduction of Peripheral data, which many construe as Big, augments our ability to provide greater insight into the questions that real estate researchers have asked over the years, but it in no way serves as a paradigm shift in knowledge that some claim it may be. Researchers are subject to the same constraints and subjective decision-making in Big Data analysis as they are with small sets, which are arguably more critical given size, scope and source. While the use of Big Data is important, it is no more important and reliable than other types of data and should be treated with scepticism when not rigorously assessed and thoroughly examined as applicable to the problem at hand.

REFERENCES

Arribas-Bel, D, 2014, Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, vol. 49, pp. 45-53.

Burkholder, L (ed.), 1992, *Philosophy and the Computer*, Westview Press, Boulder, San Francisco and Oxford.

Bollier, D, 2010, The promise and peril of big data, *The Aspen Institute*.

Bowker, G C, 2005, Memory practices in the sciences, MIT Press, Cambridge, MA.

Boyd, D & Crawford K, 2012, Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, and Society*, vol. 15, no. 5, pp. 662-679.

Chen, H, Chiang, R H L, & Story, V C, 2012, Business intelligence and analytics: from Big Data to big impact. *MIS Quarterly*, vol. 36, no. 4, pp. 1165-1188.

Crampton, J, Graham, M, Poorthuis, A, Shelton, T, Stephens, M, Wilson, M, & Zook, M, 2013, *Beyond the geotag; Situating ’big data’ and leveraging the potential of the geoweb*, vol. 40, no. 2, pp. 130-139.

Diakopoulos, N, 2015, Algorithmic accountability: journalistic investigation of computational structures, *Digital Journalism, vol*. 3, no. 3, pp. 398-415.

Du, D, Li, A, Zhang, L & Li, H, 2015, Review on the applications and the handling techniques of big data in Chinese realty enterprises, *Annals of Data Science*, vol. 1, no. 3-4, pp. 339-357.

IBM Big Data Hub, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, accessed on 5 August, 2015.

Jacobs, A, 2009, The pathologies of big Ddata. *Communications of the ACM*, vol. 52 no. 8, pp. 36-44.

Kitchen, R, 2014, The real time city? Big data and smart urbanism, *GeoJournal*, vol. 79, no. 1, pp. 1-14.

Laney, D, 2001, 3-D data management: controlling data volume, velocity, and variety, *Application Delivery Strategies*, META Group, 6 February.

Kudyba, S & Kwatinetz, M, 2015, Introduction to the big data era, In *Big Data, Mining and Analytics: Components of Strategic Decision Making*. Ed. Kudyba, S. pp 1-15, CRC Press, Boca Raton, FL.

Lazer, D, Pentland, A, Adamic, L, Aral, S, Barabasi, A, Brewer, D, Christakis, N, Contractor, N, Fowler, J, Gutmann, M, Jebara, T, King, G, Macy, M, Roy D, & Van Alstyne, M, 2009, Computational social science. *Science*, vol. 323, no. 5915, pp. 721-723.

Madden, S, 2012, From databases to big data, *IEEE Internet Computing*, May/June, 4-6.

Miller, H, 2010, The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, vol. 50, no. 1, pp. 181-201.

Mu, J, Wu, F and Zhang, A, 2014, Housing value forecasting based on machine learning models. *Abstract and Applied Analysis*, 2014, Article ID 648047.

O'Reilly Media, 2012, Volume, velocity, variety: what you need to know about big data, *Forbes*, 19 Jan, accessed from Forbes.com on 5 August 2015.

Oreskovic, A. (2015) Here's another area where Twitter appears to have stalled: tweets per day. *Business Insider Australia*, 16 June, accessed from businessinsider.com.au on 5 August 2015.

Provost, F, & Fawcett, T, 2013, Data science and its relationship to big data and data-driven decision making, *Big Data*, vol. 1, no. 1, pp. 51-59.

Rey, S J, 2014, Open regional science, *The Annals of Regional Science*, vol. *52*, no. 3, pp. 825–837.

Ruppert, E, 2013, Rethinking empirical social sciences, *Dialogues in Human Geography*, vol. 3, no. 3, pp. 268-273.

Sonka, S, 2014, Big data and the ag sector: more than lots of numbers, *International Food and Agribusiness Management Review*, vol. 17, no. 1, pp. 1-20.

Sowe, S, & Zettsu, K, 2014, Curating big data made simple, *Big Data*, vol. 2, no. 1, pp. 23-33.

Sun, D, Du, Y, Xu, W, Zuo, M, Zhang, C & Zhou, J, 2014, Combining online news articles and web search to predict the fluctuation of real estate market in big data context. *Pacific Asia Journal of the Association for Information Systems*, vol. 6, no. 4, pp. 19-37.

Tambe, P, 2014, Big data investment, skills and firm value. *Management Science*, vol. 60, no. 6, pp. 1452-1469.

Thrall, G, 2002, Business geography and new real estate market analysis, Oxford University Press, New York, NY.

Ward, J & Barker, A, 2013, Undefined by data: a survey of big data definitions, *arXiv*, 1309.5821.

Wu, J & Deng, Y, 2015, Intercity information diffusion and price discovery in housing markets: evidence from Google searches, *Journal of Real Estate Finance and Economics*, vol. 50, pp. 289-306.

Wu, L & Brynjolfsson, E, 2015, The future of prediction: how Google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy*, eds. GoldFarb, A, Greenstein, S and Tucker, C. pp 89-118, University of Chicago Press, Chicago, IL.