UHERO

THE ECONOMIC RESEARCH ORGANIZATION
AT THE UNIVERSITY OF HAWAI'I

Sources and Types of Big Data for Macroeconomic Forecasting

BY

Philip ME Garboden

Working Paper No. 2019-3

July 8, 2019

UNIVERSITY OF HAWAI'I AT MANOA
2424 MAILE WAY, ROOM 540 • HONOLULU, HAWAI'I 96822
WWW.UHERO.HAWAII.EDU

WORKING PAPERS ARE PRELIMINARY MATERIALS CIRCULATED TO STIMULATE DISCUSSION AND CRITICAL COMMENT. THE VIEWS EXPRESSED ARE THOSE OF THE INDIVIDUAL AUTHORS.

Chapter 1
Sources and Types of Big Data for Macroeconomic Forecasting

Philip ME Garboden

Abstract This chapter considers the types of Big Data that have proven useful for macroeconomic forecasting. It first presents the various definitions of Big Data, proposing one we believe is most useful for forecasting. The literature on both the opportunities and challenges of Big Data are presented. It then proposes a taxonomy of the types of Big Data: 1) Financial Market Data; 2) E-Commerce and Credit Cards; 3) Mobile Phones; 4) Search; 5) Social Media Data; 6) Textual Data; 7) Sensors, and The Internet of Things; 8) Transportation Data; 9) Other Administrative Data. Noteworthy studies are described throughout.

1.1 Understanding What's Big About Big Data

Nearly two decades after the term 'Big Data' first appeared in print, there remains little consensus regarding what it means (Lohr, 2012; Shi, 2014). Like many a scientific craze before it, the term Big Data quickly became an omnipresent buzzword applied to anything and everything that needed the gloss of being cutting edge. Big Data was going to solve many of society's most complex problems (Mayer-Schonberger & Cukier, 2013), while simultaneously sowing the seeds of its self-destruction (O'Neil, 2017). Of course, neither of these predictions was accurate; Big Data do have the potential to advance our understanding of the world, but the hard problems remain hard and incrementalism continues to dominate the social sciences. And while critics have expressed legitimate concerns, specifically at the intersection of data and governance, the consequences are by no means as dire as some would make them out to be.

Hyperbole aside, Big Data do have enormous potential to improve the timeliness and accuracy of macroeconomic forecasts. Just a decade ago, policymakers needed to wait for periodic releases of key indicators such as GDP and inflation followed by a

Philip ME Garboden at University of Hawaii Manoa, 2424 Maile Way, Honolulu, HI 96822, e-mail: pgarbod@hawaii.edu

series of subsequent corrections. Today, high frequency economic time series allow researchers to produce and adjust their forecasts far more frequently (Baldacci et al., 2016; Bok, Caratelli, Giannone, Sbordone, & Tambalotti, 2018; Einav & Levin, 2014a, 2014b; Swanson & Xiong, 2018), even in real-time (Croushore, 2011). Not only are today's time series updated more frequently, but there are more of them available than ever before (Bok et al., 2018) on a much more heterogeneous set of topics (Einav & Levin, 2014b), often with near population-level coverage (Einav & Levin, 2014a).

In this chapter, we consider the types of Big Data that have proven useful for macroeconomic forecasting. We begin by adjudicating between the various definitions of the term, settling on one we believe is most useful for forecasting. We review what the literature has presented as both the strengths and weaknesses of Big Data for forecasters, highlighting the particular set of skills necessary to utilize non-traditional data resources. This chapter leaves any in-depth discussion of analytic tools and data structures to the rest of the volume, and instead highlights the challenges inherent in the data themselves, their management, cleaning, and maintenance. We then propose a taxonomy of the types of data useful for forecasting, providing substantive examples for each. While we are neither the first nor the last researchers to take on such a categorization, we have structured ours to help readers see the full range of data that can be brought to bear for forecasting in a Big Data world.

1.1.1 How big is big?

Not surprisingly, there is no specific threshold after which a dataset can be considered 'Big.' Many commentators have attempted to describe the qualitative differences that separate Big Data from traditional data sources. Two themes emerge: First, Big Data are generally collected for purposes other than academic research and statistical modeling (Baldacci et al., 2016; Einav & Levin, 2014a). Second, they generally require processing beyond the capabilities of standard statistical software (Hassani & Silva, 2015; Shi, 2014; Taylor, Schroeder, & Meyer, 2014). As much as neither of these assertions is wholly correct, there is some value in both.

Administrative data have been used for forecasting well before the advent of Big Data (Bok et al., 2018). Data from Multiple Listing Services, for example, has played a key role in forecasting future housing demand for decades, well before anyone thought to call it 'Big' (Vidger, 1969). Nor are all Big Data administrative data. Data collected for scientific purposes from the Sloan Digital Sky Survey consist of over 175,000 galaxy spectra, prompting researchers to develop novel data management techniques (Yip et al., 2004). Despite these numerous exceptions, however, much of the data we now consider 'Big' is collected as a part of regular business or governing processes and thus, as we will describe below, present a number of technical challenges related to data management and cleaning.

The second argument, that Big Data are 'Big' because of the computational requirements associated with their analysis, is similarly resonant but emerges more from the discipline's origins in computer science than from its uses in macroeconomics. Most of the definitions in this vein focus on whether or not a dataset is "difficult to process and analyze in reasonable time" (Shi, 2014; Tien, 2014) or more flippantly when a dataset is so large that "you can't use STATA" (Hilger quoted in Taylor et al., 2014). And while these definitions clearly present a moving target (just think about STATA from 10 years ago), they reflect the fact that Big Data are an inherently intra-disciplinary endeavor that often requires economists to collaborate with computer scientists to structure the data for analysis.

Rather than focusing on precise definitions, we believe it is more important to distinguish between the types of Big Data, specifically the dimensions by which a dataset can be considered 'Big.' Many canonical attempts have been made to define these parameters, from "volume, velocity, and variety" (Laney, 2001) to its more recent augmented form "volume, velocity, variety, and veracity" (Shi, 2014), to "tall, fat, and huge" (Buono, Mazzi, Kapetanios, Marcellino, & Papailias, 2017).

But most valuable for forecasting is the idea that time series data are big if they are huge in one or more of the following dimensions: the length of time (days, quarters, years) the data is collected (T), the number of samples per unit time that an observation is made (m),[1] and the number of variables that are collected at this rate (K) for an X matrix of dimensions ($mT \times K$) (Diebold, 2016b). Time series data are thus Big if they are tall (huge T), wide (huge K), dense (huge m), or any combination of those. The value of this approach is that it distinguishes between the types of data that are big because they have been collected for a very long time (US population, for example), those that are big because they are collected very frequently (tick-by-tick stock fluctuations), and those that contain a substantial number of variables (satellite imaging data). From a forecasting perspective, this differentiation provides a common language to determine the strengths of a dataset in forecasting volatility versus trend in both the short and the long term. As Diebold points out, dense data (huge m) are largely uninformative for forecasting long term trends (for which one wants tall data) (Diebold, 2016c), but can be quite useful for volatility estimation (Diebold, 2016a).

1.1.2 The challenges of big data

Big Data, however defined, present a unique set of challenges to macroeconomists above and beyond the need for new statistical tools. The data sources themselves require a level of technical expertise outside of the traditional methods curriculum. In this section, we outline those challenges.

[1] The value of m may not be constant for a particular time series and, in some cases such as tick data, may be dependent on the data generating process itself.

Undocumented and changing data structures

Most Big Data are not created for the benefit of economic researchers but exists as a byproduct of business or governmental activities. The Internal Revenue Service keeps tax information on all Americans so that it can collect taxes. Google stores billions of search queries so that it can improve its algorithms and increase advertisement revenue. Electronic medical records ensure continuity of care and accurate billing. None of these systems were designed with academic research in mind. While many businesses have begun to develop APIs and collaborate with academic institutions, these projects exist far outside of these firms' core business model. This has several consequences.

First of all, Big Data generally come to researchers uncleaned, unstructured, and undocumented (Baldacci et al., 2016; Einav & Levin, 2014b; Laney, 2001). The structure of "Big Data" can be quite complex often incorporating spatial and temporal elements into multi-dimensional unbalanced panels (Buono et al., 2017; Matyas, 2017). While traditional survey data include metadata that can help expedite analysis, most business and governmental data are simply exported from proprietary systems which store data in the way most convenient to the users (Bok et al., 2018). Larger more technologically friendly companies have invested resources in making their data more accessible for research. Data from the likes of Zillow, Twitter, and Google, for example, are made available through APIs, online dashboards, and data sharing agreements, generally with accompanying codebooks and documentation of database structure. But not all organizations have the resources necessary to prepare their data in that manner.

Second, traditional longitudinal data sources take care to ensure that the data collected is comparable across time. Data collected in later waves must be identical to or backwards compatible with earlier waves of the survey. For Big Data collected from private sources, the data change as business needs change, resulting in challenges when constructing time series over many years (huge T data). Moreover, much of technology that could realistically allow companies to store vast amounts of data was developed fairly recently, meaning that constructing truly huge T time series data is often incompatible with Big Data. To make matters worse, local government agencies are often required to store data for a set period of time, after which they can (and in our experience often do) destroy the information.

Third, many sources of Big Data rely on the increasing uptake of digital technologies and thus are representative of a changing proportion of the population (Baldacci et al., 2016; Buono et al., 2017). Cellular phone data for example have changed over just the last 20 years from including a small non-representative subset of all telecommunications to being nearly universal. Social networks, by comparison, go in and out of popularity over time as new competitors attract younger early adapters out of older systems.

Need for network infrastructure and distributed computing

A second core challenge for macroeconomists looking to utilize Big Data is that it can rarely be stored and processed on a personal computer (Bryant, Katz, & Lazowska, 2008; Einav & Levin, 2014a, 2014b). Instead, it requires access to distributed cluster computing systems connected via high-speed networks. Ironically, access to these technologies has put industry at an advantage over academia, with researchers' decision to embrace big data coming surprisingly "late to the game" (Bryant et al., 2008).

Fortunately, many campuses now host distributed computing centers to which faculty have access, eliminating a serious obstacle to big data utilization. Unfortunately, usage of these facilities can be expensive and require long-term funding streams if data infrastructure is to be maintained. Moreover, hardware access is necessary but not sufficient. The knowledge to use such equipment falls well outside of the usual Economics training and requires a great deal of specialized knowledge generally housed in computer science and engineering departments. While these challenges have forced Big Data computing to become a rare "interdisciplinary triumph" (Diebold, 2012; Shi, 2014), they also incur costs associated with the translational work necessary to link expertise across disciplines.

Costs and access limitations

While some agencies and corporations have enthusiastically partnered with macroeconomists, others have been far more hesitant to open their data up to outside scrutiny. Many corporations know that their data have value not just to academics but for business purposes as well. While many companies will negotiate data sharing agreements for non-commercial uses, some big data are, quite simply, very expensive.

A complementary concern is that because much of Big Data is proprietary, private, or both, organizations have become increasingly hesitant to share their data publicly, particularly when there are costs associated with deidentification. A security breech can not only harm the data provider's reputation (including potential litigation) but can greatly reduce the amount of data the provider is willing to make available to researchers in the future. In best case scenario, data owners are requiring increasingly costly security systems to be installed to limit potential harm (Bryant et al., 2008). In the worst, they simply refuse to provide the data.

Data snooping, causation, and big data hubris

One of the main challenges with Big Data emerges from its greatest strength: there's a lot of it. Because many of the techniques used to extract patterns from vast datasets are agnostic with respect to theory, the researcher must remain vigilant to avoid overfitting (Baldacci et al., 2016; Hassani & Silva, 2015; Taylor et al., 2014). Although few macroeconomists would conflate the two, there is a tendency for journalists and

the general public to interpret a predictive process with a causal one, a concern that is amplified in situations where variables are not pre-selected on theoretical grounds but instead are allowed to emerge algorithmically.

Big Data generation, moreover, rarely aligns the sampling logics, making traditional tests of statistical significance largely inappropriate (Abadie, Athey, Imbens, & Wooldridge, 2014, 2017). The uncertainty of estimates based on Big Data is more likely to be based on the data generating process, design, or measurement than on issues associated with random draws from a theoretically infinite population.

When ignored, these issues can lead to what some researchers have referred to as "big data hubris" (Baldacci et al., 2016), a sort of overconfidence among researchers that having a vast amount of information can compensate for traditional econometric rigor around issues of selection, endogeneity, and causality.

Perhaps the most common parable of Big Data hubris comes from Google's attempt to build a real-time flu tracker that could help public health professionals respond more quickly (Lazer, Kennedy, King, & Vespignani, 2014). The idea behind the Google Flu Tracker (GFT) was to use search terms such as 'do I have the flu?' to determine the spread of the flu much more rapidly than the CDC's traditional tracking, which relied on reports from hospital laboratories. The GFT was released to enormous fanfare and provided a sort of proof-of-concept for how Big Data could benefit society (Mayer-Schonberger & Cukier, 2013). Unfortunately, as time went on, the GFT became less and less effective, predicting nearly double the amount of flu than was actually occurring (Lazer et al., 2014).

The reason for this was fairly simple. Google had literally millions of search terms available in its database but only 1152 data points from the CDC. Because machine learning was utilized naively - without integration with statistical methods and theory - it was easy to identify a set of search terms that more or less perfectly predicted the CDC data. Going forward however, the algorithm showed itself to have poor out-of-sample validity. According to Lazer et al. (2014), even 3-week-old CDC data do a better job of predicting current flu prevalence than Google's tracker, even after extensive improvements were made to the system.

While there is no reason to give up on what would be a useful public health tool, the shortcomings of GFT illustrate the risks of looking for patterns in data without expertise in economic forecasting and time series analysis. Moreover, Google Search is an evolving platform both in terms of its internal algorithms but also its usage, making it particularly hard to model the data generating function in a way that will be reliable across time.

1.2 Sources of Big Data for Forecasting

Having outlined the challenges of Big Data, we now turn to various forms of Big Data and how they can be useful for macroeconomic forecasting. They are: 1) Financial Market Data; 2) E-Commerce and Credit Cards; 3) Mobile Phones; 4) Search; 5) Social Media Data; 6) Textual Data; 7) Sensors, and The Internet of

Things; 8) Transportation Data; 9) Other Administrative Data. There are a number of existing taxonomies of Big Data available in the literature (Baldacci et al., 2016; Bryant et al., 2008; Buono et al., 2017) and we do not claim any superiority to our structure other than a feeling for its inherent logic in the forecasting context. In each section we briefly describe the types of data that fall into each category and present exemplary (or at least interesting) examples of how each type of data has been used in forecasting. While all examples are related to forecasting, we admittedly look outside of macroeconomics for examples of the less common data types. Our goal was to err on the side of inclusion, as many of these examples may have relevance to macroeconomics.

1.2.1 Financial market data

Many core economic indicators such as inflation and GDP are released several months after the time period they represent, sometimes followed by a series of corrections. Because forecasts of these measures with higher temporal granularity are enormously valuable to private firms as well as government agencies, the issue of how to forecast (or nowcast [2]) economic indicators has received significant attention.

The sheer number of hourly, daily, weekly and monthly data series that can be applied to such analyses is staggering. Nearly all aspects of financial markets are regularly reported including commodity prices, trades and quotes of both foreign and domestic stocks, derivatives, option transactions, production indexes, housing starts and sales, imports, exports, industry sales, treasury bond rates, jobless claims, and currency exchange rates, just to name a few (Bańbura, Giannone, Modugno, & Reichlin, 2013; Buono et al., 2017). The challenge is then how to manage such abundance, particularly when the number of items in each series (mT) is less than the number of series (K), and when the different data series relevant to the forecast are reported at different frequencies (m) and different release lags creating a ragged edge at the end of the trend (Bańbura et al., 2013).

It is well beyond the scope of this chapter to summarize all the examples of how high (or at least higher) frequency financial data have been used to improve macroeconomic forecasts (for some examples see Andreou, Ghysels, & Kourtellos, 2013; Angelini, Camba-Mendez, Giannone, Reichlin, & Rünstler, 2011; Aruoba, Diebold, & Scotti, 2009; Baumeister, Guérin, & Kilian, 2015; Giannone, Reichlin, & Small, 2008; Kim & Swanson, 2018; Monteforte & Moretti, 2013; Pan, Wang, Wang, & Yang, 2018; Stock & Watson, 2002). But a few examples are worth highlighting.

Modugno (2013) examines the issue of constructing an inflation forecast that can be updated continuously, rather than waiting for monthly releases (as is the case in the US). Modugno uses daily data on commodity prices from the World Market Price of Raw Materials (RMP), weekly data on energy prices from the Weekly Retail Gasoline

[2] We generally favor the use of the word 'forecast' throughout the chapter even for predictions of contemporaneous events. In a philosophical sense, we see little difference in predicting a number that is unknown because it has not yet occurred or because it has not yet been observed.

and Diesel Prices data (WRGDP) from the Energy Information Administration, monthly data on manufacturing from the Institute for Supply Management (released two weeks prior to inflation), and daily financial data from the US dollar index, the S&P 500, the Treasury constant maturity rate, and the Treasury-bill rate. He finds that for zero and one month horizons the inclusion of these mixed frequency data outperforms standard models but exclusively due to an improvement in the forecasting of energy and raw material prices.

Degiannakis and Filis (2018) attempt to use high-frequency market data to forecast oil prices. Arguing that because oil markets are becoming increasingly financialized, there are benefits in looking beyond market fundamentals to improve forecasts. Their model combines traditional measures on the global business cycle, oil production, oil stocks, and the capacity utilization rate with "ulta-high" frequency tick-by-tick data on exchange rates, stock market indexes, commodities (oil, gold, copper, gas, palladium, silver), and US T-bill rates. They find that for long term forecasts, the fundamentals remain critical, but the inclusion of highly granular market data improves their short term estimates in ways robust to various comparisons.

1.2.2 E-commerce and scanner data

In order to construct the consumer price index, the census sends fieldworkers out to collect prices on a basket of goods from brick and mortar stores across the country. While this represents a sort of gold standard for data quality, it is not without its limitations. It is both expensive to collect and impossible to monitor in real time (Cavallo & Rigobon, 2016). Nor is it able to address bias related to the substitution of one good for another or to provide details on how quality may shift within the existing basket; factors that are critical for an accurate measure of inflation (Silver & Heravi, 2001).

To fill this gap, economists have begun collecting enormous datasets of prices, relying either on bar-code scanner data (Berardi, Sevestre, & Thébault, 2017; Silver & Heravi, 2001) or by scraping online listings from e-commerce retailers (Cavallo & Rigobon, 2016; Rigobón, 2015). Perhaps the most famous of these is the MIT Billion Prices Project which, in 2019, was collecting 15 million prices every day from more than 1000 retailers in 60 countries (Project, 2019). The genesis of the project came from Cavallo (2013) and his interest in measuring inflation in Argentina. Cavallo suspected, and would later prove, that the official releases from Argentine officials were masking the true rate of inflation in the county. By scraping four years worth of data from supermarket websites in Argentina (and comparison data in Brazil, Chile, Columbia, and Venezuela), Cavallo was able to show an empirical inflation rate of 20 percent, compared to official statistics which hovered around 4 percent. Building on this methodology, the Billion Prices Project began scraping and curating online sales prices from around the world allowing not only for inflation forecasting, but additional empirical research on price setting, stickiness, and so forth. Other similar

work has been done using data from Adobe Analytics, which collects sales data from its clients for the production of business metrics (Goolsbee & Klenow, 2018).

While e-commerce has increased its market share substantially (and is continuing to do so), its penetration remains dwarfed by brick and mortar businesses particularly in specific sectors (like groceries). For this reason, researchers have partnered with particular retailers to collect price scanner data in order to construct price indexes (Ivancic, Diewert, & Fox, 2011), more precisely measure inflation (Silver & Heravi, 2001), and to examine the influence of geopolitical events on sales (Pandya & Venkatesan, 2016). While is seems like that the relative ease of collecting online prices, combined with the increasing e-commerce market share will tend to push research into the online space, both forms of price data have enormous potential for forecasting.

1.2.3 Mobile phones

In advanced economies, nearly 100 percent of the population uses mobile phones, with the number of cellphone accounts exceeding the number of adults by over 25 percent (Blondel, Decuyper, & Krings, 2015). Even in the developing world, mobile phones are used by three quarters of the population; in a country such as Cote d'Ivoire, where fewer than 3 percent of households have access to the web, 83 percent use cell phones (Mao, Shuai, Ahn, & Bollen, 2015). This has pushed many researchers to consider the value of mobile phone data for economic forecasting, particularly in areas where traditional demographic surveys are expensive or dangerous to conduct (Blondel et al., 2015; Ricciato, Widhalm, Craglia, & Pantisano, 2015).

The mobile phone data that is available to researchers is fairly thin, generally consisting solely of the Call Detail Records (CDR) that simply provide an unique identifier of the mobile device, the cell tower to which it connected, and type of connection (data, call, etc.), the time, and duration of the call (Ricciato et al., 2015). Some datasets will also include data on the destination tower or, if the user is moving, the different connecting towers that he or she may utilize during the route. To compensate for this thinness, however, is the data's broad coverage (particularly in areas with a single cellular carrier) and the potential for near real-time data access.

Mobile phone data have been used to forecast demographic trends such as population densities (Deville et al., 2014), poverty (Blumenstock, Cadamuro, & On, 2015; Mao et al., 2015; Smith-Clarke, Mashhadi, & Capra, 2014), and unemployment (Toole et al. 2015). In general, these papers look at the distribution of the number of cell phone calls and the networks of connections between towers (including density and heterogeneity) to improve forecasting at smaller levels of temporal and spatial granularity than is typically available.[3]

In one of the more ingenious analyses, Toole et al. (2015) looked at patterns of cell phone usage before and after a mass layoff in an undisclosed European country.

[3] This approach has particular value in times of disaster, upheaval or other unexpected events for which data are necessary for an effective response.

Using a Bayesian classification model, they found that after being laid off, individuals significantly reduced their level of communication. They made fewer calls, received fewer calls, and spoke with a smaller set of people in the period after the layoff than they had prior to the layoff (or than a control group in a town that did not experience the employment shock). The researchers used the relationship identified in this analysis to train a macroeconomic model to improve forecasts of regional unemployment, finding significant improvements to their predictions once mobile phone data were included.

Mobile phones and other always-on devices have also allowed economists to increase the frequency with which they survey individuals as they go about their daily routines. MacKerron and Mourta's Mappiness project, for example, utilizes a mobile phone app to continually ping users about their level of happiness throughout the day resulting in over 3.5 million data points produced by tens of thousands of British citizens (MacKerron & Mourato, 2010). While this particular project has yet to be employed for forecasting, it's clear that this sort of real time attitudinal data could contribute to the prediction of multiple macroeconomic time series.

1.2.4 Search data

Once the purview of marketing departments, the use of search data has increased in forecasting, driven by the availability of tools such as Google Trends (and, more recently, Google Correlate). These tools report a standardized measure of search volume for particular user-identified queries on a scale from 0, meaning no searches, to 100, representing the highest search volume for a particular topic during the time window. Smaller areas of spatial aggregation, namely states, are also available. While keywords with a search volume below a particular threshold are redacted for confidentiality reasons, Google Trends provides otherwise unavailable insight into a population's interest in a particular topic or desire for particular information.

It is hardly surprising given the nature of this data, that it would be used for forecasting. Although attempts have not been wholly successful (Lazer et al., 2014), many researchers have attempted to use search terms for disease symptoms to track epidemics in real time (Ginsberg et al., 2009; Yuan et al., 2013). Similar approaches have used search engine trend data to forecast hotel room demand (Pan, Chenguang Wu, & Song, 2012), commercial real estate values (Alexander Dietzel, Braun, & Schäfers, 2014), movie openings, video game sales, and song popularity (Goel, Hofman, Lahaie, Pennock, & Watts, 2010). The various studies using search data have indicated the challenges of making such predictions. Goel et al. (2010), for example, look at a number of different media and note significant differences in search data's ability to predict sale performance. For "non-sequel games," search proved a critical early indicator of buzz and thus first month sales. But this pattern did not hold for sequel games for which prequel sales were a much stronger predictor, or movies for which the inclusion of search in the forecasting models provided very little improvement over traditional forecasting models.

One of the places where search data have proved most valuable is in the early prediction of joblessness and unemployment (Choi & Varian, 2009, 2012; D'Amuri & Marcucci, 2017; Smith, 2016; Tefft, 2011). D'Amuri and Marcucci (2017), for example, use search volume for the word 'jobs' to forecast the US monthly unemployment rate finding it to significantly outperform traditional models for a wide range of out-of-sample time periods (particularly during the Great Recession). In an interesting attempt at model falsification, the authors use Google Correlate to find the search term with the highest temporal correlation with 'jobs' but that is substantively unrelated to employment (in this case, the term was 'dos,' referring to the operating system or the Department of State). The authors found that despite its strong in-sample correlation, 'dos' performed poorly in out-of-sample forecasting. This combined with several other robustness checks, provides evidence that unlike early attempts with Google Flu Trends, online searchers looking for job openings are a sustainable predictor of unemployment rates.

1.2.5 Social network data

Since the early days of the internet, social networks have produced vast quantities of data, much of it in real time. As the use of platforms such as Facebook and Twitter have become nearly ubiquitous, an increasing number of economists have looked for ways to harness these data streams for forecasting. Theoretically, if the data from these networks can be collected and processed at sufficient speeds, forecasters may be able to gather information that has not yet been incorporated into the prices on stock or the betting markets, presenting opportunities for arbitrage (Arias, Arratia, & Xuriguera, 2013; Giles, 2010).

Generally speaking, the information provided by social media can be of two distinct kinds. First, it may provide actual information based on eyewitness accounts, rumors, or whisper campaigns that has not yet been reported by mainstream news sources (Williams & Reade, 2016). More commonly, social media data are thought to contain early signals of the sentiments or emotional states of specific populations which is predictive of their future investment behaviors (Mittal & Goel, 2012). Whether exogenous or based on unmeasured fundamentals, such sentiments can drive market behavior and thus may be useful data to incorporate into models for forecasting.

The most common social network used for forecasting is Twitter, likely because it lacks the privacy restrictions of closed networks such as Facebook and because of Twitter's support for the data needs of academic researchers. Each day, approximately 275 million active Twitter users draft 500 million tweets often live tweeting events such as sporting events, concerts, or political rallies in close to real time.

Several papers have examined how these data streams can be leveraged for online betting, amounting to a test of whether incorporating social media data can improve outcome forecasting more accurately than the traditional models used by odds makers. For example, Brown, Rambaccussing, Reade, and Rossi (2018) used 13.8

million tweets responding to events in UK Premier League soccer[4] matches. Coding each tweet using a microblogging dictionary as either a positive or negative reaction to a particular team's performance, the researches found that social media contained significant information that had not yet been incorporated into betting prices on the real time wagering site Betfair.com. Similar research has looked at Twitter's value in predicting box office revenue for films (Arias et al., 2013; Asur & Huberman, 2010) and the results of democratic elections (Williams & Reade, 2016).

Not surprisingly, a larger literature examines the ability of social networks to forecast stock prices, presenting itself as challenges to the efficient market hypothesis (Bollen, Mao, & Zeng, 2011; Mittal & Goel, 2012). Bollen et al. (2011), in a much cited paper, used Twitter data to collect what they term "collective mood states" operationally defined along six dimensions (calm, alert, sure, vital, kind, and happy). The researchers then used a self-organizing fuzzy neural network model to examine the non-linear association between these sentiments and the Dow Jones Industrial Average (DJIA). They found that the inclusion of some mood states (calm in particular) greatly improved predictions for the DJIA, suggesting that public sentiment was not fully incorporated into stock prices in real time. Similar work with similar findings has been done using social networks more directly targeting potential stock market investors such as stock message boards (Antweiler & Frank, 2004), Seeking Alpha (Chen, De, Hu, & Hwang, 2014) and The Motley Fool's CAPS system, which crowdsources individual opinions on stock movements (Avery, Chevalier, & Zeckhauser, 2015).

1.2.6 Text and media data

If social media is the most popular textual data used for forecasting, it is far from the only one. Text-mining is becoming an increasingly popular technique to identify trends in both sentiment and uncertainty (Bholat, Hansen, Santos, & Schonhardt-Bailey, 2015; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). The most popular text data used in such analyses come from online newspapers, particularly business related newspapers like the *Wall Street Journal* or the *Financial Times* (Alanyali, Moat, & Preis, 2013; Baker, Bloom, & Davis, 2016; Schumaker & Chen, 2009; Thorsrud, 2018). But other sources are used as well such as minutes from the Fed's Federal Open Market Committee (FOMC) (Ericsson, 2016, 2017) and Wikipedia (Mestyán, Yasseri, & Kertész, 2013; Moat et al., 2013).

In one of the field's seminal works, Baker et al. (2016) develop an index of economic policy uncertainty (EPU) which counts the number of articles using one or more terms from each of the following three groups: 1) 'economic' or 'economy', 2) 'uncertainty,' or 'uncertain', and 3) 'Congress,' 'deficit,' 'Federal Reserve,' 'legislation,' 'regulation,' or 'White House.' For the last two decades in the United States, the researchers constructed this measure using the top 10 daily newspapers, with

[4] Football.

other sources used internationally and further back in time. While this approach will assuredly not capture all articles that suggest policy uncertainty, the index strongly correlates with existing measures of policy uncertainty and can improve economic forecasts. When the index of uncertainty increases, investment, output and employment all decline.

A wholly different use of textual data is employed by Moat et al. (2013). Rather than focus on the content of news sources, their research considers how investors seek information prior to trading decisions. Collecting a count of views and edits to Wikipedia pages on particular DJIA firms, the authors find a correlation between Wikipedia usage and movements in particular stocks. Fortunately or unfortunately (depending on your view), they did not find similar associations with more generic Wikipedia pages listed on the *General Economic Concepts* page such as 'modern portfolio theory' or 'comparative advantage.'

1.2.7 Sensors, and the internet of things

The Internet of Things, like many technology trends, is more often discussed than used. Nevertheless, there is no doubt that the technology for ubiquitous sensing is decreasing dramatically in cost with potentially profound implications for forecasting. There are currently few examples of sensor data being used for economic forecasting, and that which does exist comes from satellite images, infrared networks, or sophisticated weather sensing hardware (rather than toasters and air conditioners).

One of the primary uses of sensor data is to collect satellite information on land use to predict economic development and GDP (Keola, Andersson, & Hall, 2015; Park, Jeon, Kim, & Choi, 2011; Seto & Kaufmann, 2003). Keola et al. (2015), for example, use satellite imagery from the Defense Meteorological Satellite Program (DMSP) to estimate the level of ambient nighttime light and NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) to determine whether non-urbanized areas are forests or agricultural land. The authors find that these two measures combined can be useful for forecasting economic growth. This technology, the authors argue, is particularly valuable in areas where traditional survey and administrative measures are not yet reliably available, specifically the developing world.

In a much more spatially limited implementation, Howard and Hoff (2013) use a network of passive infrared sensors to collect data on building occupancy. Applying a modified Bayesian combined forecasting approach, the authors are able to forecast building occupancy up to 60 minutes into the future. While seemingly inconsequential, the authors argue that this one-hour-ahead forecast has the potential to dramatically reduce energy consumption as smart heating and cooling systems will be able to pre-cool or pre-heat rooms only when occupancy is forecast.

Looking to the future, an increasing number of electronics will soon be embedded with some form of low-cost computer capable of communicating remotely either to end users or to producers (Fleisch, 2010; Keola et al., 2015). Because this technology is in its infancy, no economic forecasts have incorporated data collected from these

micro-computers (Buono et al., 2017). It is not hard, however, to imagine how such data could function in a forecasting context. Sensors embedded in consumer goods will be able to sense the proximity of other sensors, the functioning of the goods themselves, security threats to those goods, and user behavior (Fleisch, 2010). Data that suggest high levels of product obsolescence could be used to forecast future market demand. Data on user behavior, may some day serve as a proxy for sentiments, fashion, and even health, all of which could theoretically forecast markets in a way similar to social network data.

1.2.8 Transportation data

Over the last few years, there has been an enormous increase in both the quantity and quality of transportation data, whether through GPS enabled mobile phones, street level sensors, or image data. This has led to a sort of renaissance of forecasting within the transportation planning field. To date, the bulk of this work has been detached from the work of macroeconomists as it has primarily focused on predicting traffic congestion (Lv, Duan, Kang, Li, & Wang, 2015; Polson & Sokolov, 2017; Xia, Wang, Li, Li, & Zhang, 2016; Yao & Shen, 2017), electric vehicle charging demand (Arias & Bae, 2016), or transportation-related crime and accidents (Kouziokas, 2017).

One paper that suggest a locus of interaction between macroeconomists and transportation forecasters comes from Madhavi et al. (2017). This paper incorporates transportation data into models of electricity load forecasting and finds that, in addition to traditional weather variables, the volume of traffic into and out of a particular area can be highly predictive of energy demands.

While little work exists, there are a number of plausible uses of transportation data to forecast large-scale trends. Of course, transportation inefficiencies are likely predictive of gasoline prices. The pattern of peak travel demand could be indicative of changes of joblessness, sector growth or decline, and other labor force variables. And finally, the movement of vehicles throughout a metropolitan area could be used to construct a forecast of both residential and commercial real estate demand within particular metropolitan areas.

1.2.9 Other administrative data

The reliability and consistency of some government, nonprofit, and trade association data has made it difficult to construct reliable time series, particularly at a national level. In the presence of undocumented changes in the data collection process (either because of regulation or technology), it becomes difficult to disentangle the data generating process from changes in data coverage and quality. One exception to this trend is use of data on housing sales, collected through the National Association of Realtors' Multiple Listing Service (MLS) in the United States (for examples outside

of the US, see Baltagi & Bresson, 2017). These data, with track all property sales for which a real estate agent was involved, have been invaluable to those attempting to forecast housing market trends (Chen, Ong, Zheng, & Hsu, 2017; Park & Bae, 2015).

Park and Bae (2015), for example, use a variety of machine learning algorithms to forecast trends in housing prices in a single county in Fairfax, Virgina. Using the MLS data, they obtained 15,135 records of sold properties in the county, each of which contained a rich set of property-level attributes. In similar work in the international context, Chen and colleagues use a Support Vector Machine approach using administrative sales records from Taipei City. In both cases, the Big Data forecasting approach improved over previous methods.

One place where big administrative data have significant potential is in the area of local and national budget forecasting. At the national level, the approach to revenue and expenditure forecasting has followed traditional methods conducted by up to six agencies (The Council of Economic Advisers, The Office of Management and Budget, The Federal Reserve Board, The Congressional Budget Office, The Social Security Administration, and the Bureau of Economic Analysis) (Williams & Calabrese, 2016). While some of these agencies may be including Big Data into their forecasts, there is a dearth of literature on the subject (Ghysels & Ozkan, 2015) and federal budget forecasts have traditionally shown poor out-of-year performance (Williams & Calabrese, 2016). It is almost certain that the vast majority of local governments are not doing much beyond linear interpolations, which an abundance of literature suggests is conservative with respect to revenue (see Williams & Calabrese, 2016, for review).

1.2.10 Other potential data sources

We have attempted to provide a fairly comprehensive list of the sources of Big Data useful for forecasting. In this, however, we were limited to what data *have* been tried rather than what data *could* be tried. In this final section, we propose several data sources that have not yet bubbled to the surface in the forecasting literature often because they have particular challenges related to curation or confidentiality.

The most noticeable gap relates to healthcare. Electronic medical records, while extremely sensitive, represent enormous amounts of data about millions of patients. Some of this, such as that collected by Medicare and Medicaid billing, is stored by the government, but other medical data rest with private providers. While social media may be responsive to rapidly spreading epidemics, electronic medical records would be useful in predicting healthcare utilization well into the future. In a more futuristic sense, the steady increase of 'wearables' such as Apple watches and FitBits, has the potential to provide real time information on health and wellness, potentially even enabling researchers to measure stress and anxiety, both of which the literature suggests are predictive of market behavior.

Second, while the use of textual data has increased, audio and video data has been largely ignored by the forecasting literature. Online content is now produced and consumed via innumerable media ranging from YouTube videos, to podcasts, to animated gifs. As speech recognition software improves along with our ability to extract information from images and video recordings, it is likely that some of this data may prove useful for forecasting.

Finally, as noted above, there appears to be a significant under-utilization of administrative data, particularly that which is collected at the local level but relevant nationally. The challenges here have more to do with federalism and local control than they do with data science and econometrics. Data that are primarily collected by state and local governments are often inaccessible and extremely messy. This limits, for example, our ability to use education or court data to forecast national trends; the amount of local autonomy related to the collection and curation of this data presents a simply insurmountable obstacle, at least for now.

1.3 Conclusion

This chapter has outlined the various types of Big Data that can be applied to macroeconomic forecasting. By comparison to other econometric approaches, the field is still relatively new. Like all young fields, both the challenges and opportunities are not fully understood. By and large, the myriad publications outlined here have taken the approach of adding one or more big datasets into existing forecasting approaches and comparing the out-of-sample performance of the new forecast to traditional models. While essential, this approach is challenged by the simple fact that models that fail to improve performance have a much steeper hill to climb for publication (or even public dissemination). And yet knowing what does not improve a model can be very valuable information. Moreover, it remains unclear how approaches combining multiple Big Data sources will perform versus those that select one or the other. Over a dozen papers attempt to use Big Data to forecast unemployment, all using clever and novel sources of information. But are those sources redundant to one another or would combining multiple sources of data produce even better forecasts? Beyond simply bringing more data to the forecasting table, it is these questions that will drive research going forward.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2014). *Finite population causal standard errors*. National Bureau of Economic Research.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2017). Sampling-based vs. design-based uncertainty in regression analysis. *arXiv Preprint arXiv:1706.01778*.

Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports*, *3*, 3578.

Alexander Dietzel, M., Braun, N., & Schäfers, W. (2014). Sentiment-based commercial real estate forecasting with google search volume data. *Journal of Property Investment & Finance*, *32*(6), 540–569.

Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, *31*(2), 240–251.

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L., & Rünstler, G. (2011). Short-term forecasts of Euro area GDP growth. *The Econometrics Journal*, *14*(1), C25–C44.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259–1294.

Arias, M. B., & Bae, S. (2016). Electric vehicle charging demand forecasting model based on big data technologies. *Applied Energy*, *183*, 327–339.

Arias, M., Arratia, A., & Xuriguera, R. (2013). Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 8.

Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, *27*(4), 417–427.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 ieee/wic/acm international conference on web intelligence and intelligent agent technology-volume 01* (pp. 492–499). IEEE Computer Society.

Avery, C. N., Chevalier, J. A., & Zeckhauser, R. J. (2015). The CAPS prediction system and stock market returns. *Review of Finance*, *20*(4), 1363–1381.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Baldacci, E., Buono, D., Kapetanios, G., Krische, S., Marcellino, M., Mazzi, G., & Papailias, F. (2016). Big data and macroeconomic nowcasting: From data access to modelling. Luxembourg: Publications Office of the European Union.

Baltagi, B. H., & Bresson, G. (2017). Modelling housing using multi-dimensional panel data. In L. Matyas (Ed.), *The econometrics of multi-dimensional panels* (pp. 349–376). Springer.

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. In *Handbook of economic forecasting* (Vol. 2, pp. 195–237). Elsevier.

Baumeister, C., Guérin, P., & Kilian, L. (2015). Do high-frequency financial data help forecast oil prices? the MIDAS touch at work. *International Journal of Forecasting*, *31*(2), 238–252.

Berardi, N., Sevestre, P., & Thébault, J. (2017). The determinants of consumer price dispersion: Evidence from french supermarkets. In L. Matyas (Ed.), *The econometrics of multi-dimensional panels* (pp. 427–449). Springer.

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks. *Available at SSRN 2624811*.

Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ data science*, *4*(1), 10.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, (0).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1–8.

Brown, A., Rambaccussing, D., Reade, J. J., & Rossi, G. (2018). Forecasting with social media: Evidence from tweets on soccer matches. *Economic Inquiry*, *56*(3), 1748–1763.

Bryant, R., Katz, R., & Lazowska, E. (2008). *Big-data computing: Creating revolutionary breakthroughs in commerce, science, and society*. Computing Community Consortium.

Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, *1*(2017), 93–145.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152–165.

Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, *30*(2), 151–78.

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1367–1403.

Chen, J.-H., Ong, C. F., Zheng, L., & Hsu, S.-C. (2017). Forecasting spatial dynamics of the housing market using support vector machine. *International Journal of Strategic Property Management*, *21*(3), 273–283.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1–5.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, *88*, 2–9.

Croushore, D. (2011). Frontiers of real-time data analysis. *Journal of Economic :iterature*, *49*(1), 72–100.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801–816.

Degiannakis, S., & Filis, G. (2018). Forecasting oil prices: High-frequency financial data are indeed useful. *Energy Economics*, *76*, 388–402.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, *111*(45), 15888–15893.

Diebold, F. X. (2012). *On the origin(s) and development of the term 'big data'*. Unpublished paper.

Diebold, F. X. (2016a). Big data for volatility vs. trend. https://fxdiebold.blogspot.com. Accessed: 2019-21-03.

Diebold, F. X. (2016b). Big data: Tall, wide, and dense. https://fxdiebold.blogspot.com. Accessed: 2019-21-03.

Diebold, F. X. (2016c). Dense data for long memory. https://fxdiebold.blogspot.com. Accessed: 2019-21-03.

Einav, L., & Levin, J. (2014a). Economics in the age of big data. *Science*, *346*(6210), 1243089.

Einav, L., & Levin, J. (2014b). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1–24.

Ericsson, N. R. (2016). Eliciting GDP forecasts from the FOMC's minutes around the financial crisis. *International Journal of Forecasting*, *32*(2), 571–583.

Ericsson, N. R. (2017). Predicting Fed forecasts. *Journal of Reviews on Global Economics*, *6*, 175–180.

Fleisch, E. (2010). What is the internet of things? an economic perspective. *Economics, Management & Financial Markets*, *5*(2).

Ghysels, E., & Ozkan, N. (2015). Real-time forecasting of the US federal government budget: A simple mixed frequency data regression approach. *International Journal of Forecasting*, *31*(4), 1009–1020.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676.

Giles, J. (2010). Blogs and tweets could predict the future. *New Scientist*, *206*(2765), 20–21.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National academy of sciences*, *107*(41), 17486–17490.

Goolsbee, A. D., & Klenow, P. J. (2018). Internet rising, prices falling: Measuring inflation in a world of e-commerce. In *Aea papers and proceedings* (Vol. 108, pp. 488–92).

Hassani, H., & Silva, E. S. (2015). Forecasting with big data: A review. *Annals of Data Science*, *2*(1), 5–19.

Howard, J., & Hoff, W. (2013). Forecasting building occupancy using sensor network data. In *Proceedings of the 2nd international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications* (pp. 87–94). ACM.

Ivancic, L., Diewert, W. E., & Fox, K. J. (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics*, *161*(1), 24–35.

Keola, S., Andersson, M., & Hall, O. (2015). Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth. *World Development*, *66*, 322–334.

Kim, H. H., & Swanson, N. R. (2018). Methods for backcasting, nowcasting and forecasting using factor-MIDAS: With an application to Korean GDP. *Journal of Forecasting*, *37*(3), 281–302.

Kouziokas, G. N. (2017). The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment. *Transportation Research Procedia*, *24*, 467–473.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*(70), 1.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Lohr, S. (2012). How big data became so big. *New York Times*, *11*.

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 865–873.

MacKerron, G., & Mourato, S. (2010). Lse's mappiness project may help us track the national mood: But how much should we consider happiness in deciding public policy? *British Politics and Policy at LSE*.

Madhavi, K. L., Cordova, J., Ulak, M. B., Ohlsen, M., Ozguven, E. E., Arghandeh, R., & Kocatepe, A. (2017). Advanced electricity load forecasting combining electricity and transportation network. In *2017 north american power symposium (naps)* (pp. 1–6). IEEE.

Mao, H., Shuai, X., Ahn, Y.-Y., & Bollen, J. (2015). Quantifying socio-economic indicators in developing countries from mobile phone communication data: Applications to Côte d'Ivoire. *EPJ Data Science*, *4*(1), 15.

Matyas, L. (2017). *The econometrics of multi-dimensional panels*. Springer.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, think*. Taylor & Francis.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, *8*(8), e71226.

Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports*, *3*, 1801.

Modugno, M. (2013). Now-casting inflation using high frequency data. *International Journal of Forecasting*, *29*(4), 664–675.

Monteforte, L., & Moretti, G. (2013). Real-time forecasts of inflation: The role of financial variables. *Journal of Forecasting*, *32*(1), 51–61.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653–7670.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pan, B., Chenguang Wu, D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196–210.

Pan, Z., Wang, Q., Wang, Y., & Yang, L. (2018). Forecasting US real GDP using oil prices: A time-varying parameter MIDAS model. *Energy Economics*, *72*, 177–187.

Pandya, S. S., & Venkatesan, R. (2016). French roast: Consumer response to international conflict—evidence from supermarket scanner data. *Review of Economics and Statistics*, *98*(1), 42–56.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934.

Park, S., Jeon, S., Kim, S., & Choi, C. (2011). Prediction and comparison of urban growth by land suitability index mapping using GIS and RS in South Korea. *Landscape and Urban Planning*, *99*(2), 104–114.

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1–17.

Project, B. P. (2019). Billion prices project website. http://www.thebillionpricesproject.com. Accessed: 2019-03-21.

Ricciato, F., Widhalm, P., Craglia, M., & Pantisano, F. (2015). *Estimating population density distribution from network-based mobile phone data*. Publications Office of the European Union.

Rigobón, R. (2015). Presidential address: Macroeconomics and online prices. *Economia*, *15*(2), 199–213.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, *27*(2), 12.

Seto, K. C., & Kaufmann, R. K. (2003). Modeling the drivers of urban land use change in the Pearl River Delta, China: Integrating remote sensing with socioeconomic data. *Land Economics*, *79*(1), 106–121.

Shi, Y. (2014). Big data: History, current status, and challenges going forward. *Bridge*, *44*(4), 6–11.

Silver, M., & Heravi, S. (2001). Scanner data and the measurement of inflation. *The Economic Journal*, *111*(472), 383–404.

Smith-Clarke, C., Mashhadi, A., & Capra, L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 511–520). ACM.

Smith, P. (2016). Google's MIDAS touch: Predicting UK unemployment with internet search data. *Journal of Forecasting*, *35*(3), 263–284.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167–1179.

Swanson, N. R., & Xiong, W. (2018). Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics*, *51*(3), 695–746.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, *1*(2), 2053951714536877.

Tefft, N. (2011). Insights on unemployment, unemployment insurance, and mental health. *Journal of Health Economics*, *30*(2), 258–264.

Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 1–17.

Tien, J. (2014). Overview of big data, a US perspective. *The Bridge–Linking Engineering and Society*, *44*(4), 13–17.

Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C., & Lazer, D. (2015). Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, *12*(107), 20150185.

Vidger, L. P. (1969). Analysis of price behavior in san francisco housing markets: The historical pattern (1958–67) and projections (1968–75). *The Annals of Regional Science*, *3*(1), 143–155.

Williams, D. W., & Calabrese, T. D. (2016). The status of budget forecasting. *Journal of Public and Nonprofit Affairs*, *2*(2), 127–160.

Williams, L. V., & Reade, J. J. (2016). Prediction markets, social media and information efficiency. *Kyklos*, *69*(3), 518–556.

Xia, D., Wang, B., Li, H., Li, Y., & Zhang, Z. (2016). A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting. *Neurocomputing*, *179*, 246–263.

Yao, S.-N., & Shen, Y.-C. (2017). Functional data analysis of daily curves in traffic: Transportation forecasting in the real-time. In *2017 computing conference* (pp. 1394–1397). IEEE.

Yip, C.-W., Connolly, A., Szalay, A., Budavári, T., SubbaRao, M., Frieman, J., ... Okamura, S., et al. (2004). Distributions of galaxy spectral types in the sloan digital sky survey. *The Astronomical Journal*, *128*(2), 585.

Yuan, Q., Nsoesie, E. O., Lv, B., Peng, G., Chunara, R., & Brownstein, J. S. (2013). Monitoring influenza epidemics in china with search query from baidu. *PloS one*, *8*(5), e64323.