Chapter 1

MACRO FORECASTING USING ALTERNATIVE DATA

Apurv Jain
Harvard Business School and Microsoft Research
apurvj@gmail.com

Abstract

Traditional macroeconomic data used by economic agents to make decisions are noisy, lack richness, and produced with considerable lag. This chapter explores how alternative, web-scale data sources ("Big Data") can help. We present a case study using a common alternative data source- web search to predict one of the most important data releases- non-farm payrolls (NFP). We discuss the efficacy of various machine learning (ML) techniques, the live performance of alternative data prediction models and the typical problems faced in practice.

Contents

[1] These are mostly mistakes I have either made myself directly or seen firsthand!😀

1.1 The Importance Of Macroeconomic Measurement And Prediction

Accurate measurement of the economic activity of around 161.5 million active labor-force in the United States[2] is a critical and difficult task. To get a handle on such complexity a plethora of indicators is collected by various government and private agencies. The FRED database operated by the St. Louis branch of the Federal Reserve Bank contains around $508,000$ indicators for the United States and other countries[3], with $280,000$ series for regional data in the United States alone.

To describe and understand how the economy functions, generally we make simplifications called economic models. Almost all models, use some typical key concepts such as economic growth, population growth, the labor market, production, consumption, savings, the capital market, inflation, productivity, technology etc. The main difference in various models tends to be in describing how these concepts relate to one another and how they evolve. To evaluate the performance or even to calibrate these models- which are of special interest to policy makers such as central banks, requires accurate and consistent measurement of these key concepts. As a society, we are affected by the macroeconomic expansions and contractions. In contractions, economic growth is lower, it is typically harder to find a job, and unemployment rate increases. In expansions output increases, businesses expand and typically wages rise. Economic agents such as central banks, asset owners, firms, and individuals care about the longer-term trend of the economy for planning purposes and spend considerable efforts to identify turning points as well as the short-term deviations from trend. (Kliesen, 2014). Aruoba, Diebold and Scotti (2008) eloquently state the importance of having accurate and continuous economic measurement- "Of central importance is the constant grappling. Real economic agents, making real decisions, in real time, want accurate and timely estimates of real activity." They also highlight a big opportunity in terms of faster estimates- "Business cycle chronologies such as the NBER's, which proclaims expansions and contractions long after the fact, are not useful in that regard."

The purposes of these agents will be different- for example, the U.S. central bank has a dual mandate of stable prices and maximum sustainable employment[4], and a business owner would be concerned about purchasing equipment – a durable good with high cost if the economy is about to enter a downturn. The trend in the economic data across various sectors can help identify the changes in the entire economy as well as the sector relevant to the economic agent. Figure 1.1 provides an illustration of the economic impact of contraction or recessions via lower returns in the broad S&P 500 stock market index. The shaded gray bars in the figure are periods judged by the National Bureau of Economic Research (NBER) to be recessions.

[2] https://www.bls.gov/news.release/empsit.a.htm

[3] https://fred.stlouisfed.org/

[4] https://www.chicagofed.org/research/dual-mandate/dual-mandate. "The monetary policy goals of the Federal Reserve are to foster economic conditions that achieve both stable prices and maximum sustainable employment."

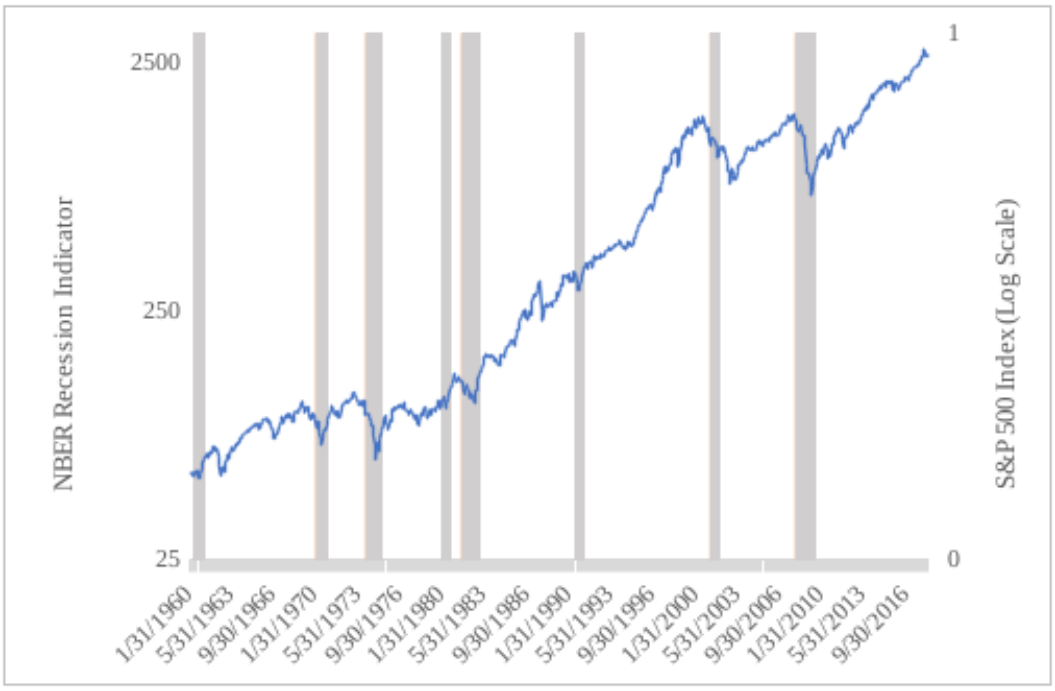


Figure 1.1: NBER recession and stock market

Table 1.1: Returns on stocks and bonds in NBER recession vs. Expansions (Non-Recession). The growth index is a Nowcast index computed similarly to NY Fed and de-trended. The data used is from 1976 until 2018. Data source: Bloomberg, Microsoft and NBER.

Asset Class		**Mean (Annualized)**	**Volatility (Annualized)**	**Skew**	**Kurtosis**
Stocks					
	Unconditional	8.9%	14.2%	−0.2	0.3
	Expansion (Nowcast Growth Index > 0)	10.9%	12.6%	0.1	0.2
	Contraction Nowcast Growth Index < 0)	4.8%	16.9%	−0.2	0.0
	NBER Recession	−4.3%	20.4%	−0.2	−0.5
	NBER Expansion	11.0%	13.1%	−0.2	0.3
Bonds					
	Unconditional	7.2%	5.0%	0.2	1.6
	Expansion (Nowcast Growth Index > 0)	6.4%	4.1%	0.2	0.5
	Contraction (Nowcast Growth Index > 0)	9.0%	6.8%	0.3	2.6
	NBER Recession	12.3%	9.2%	0.9	1.8
	NBER Expansion	6.4%	4.4%	−0.1	0.6

1.2 Important Economic Data Releases and Prediction

Most of these "traditional economic metrics" are collected by government agencies such as the Bureau of Economic Analysis (BEA) or the Bureau of Labor Statistics (BLS) in form of surveys or complex accounting calculations which are time consuming and subject to substantial revisions and potential biases (Baumol, 2012). Amongst these vast number of data series, a few key statistics such as Gross Domestic Product (GDP), the number of jobs added per month called "Non-Farm Payrolls (NFP)" and others noted in Table 1.2, stand out. The relative importance of these statistics can be inferred in a variety of ways such as the higher price impact on the financial markets to the release of these numbers in contrast to insignificant impact of most others, more mentions in speeches by the Federal Reserve officials and other government authorities, more coverage by professional economists on Wall Street, as well as academic references such as Baumol (2012) or Kliesen (2014).

The two key problems: Shocks and Revisions: Understanding and predicting the trend of various economic data collected by the government agencies would be quite useful and helpful. However, this prediction is not easy. In fact, Kliesen (2014) calls it a "daunting challenge, even for trained professional economists." Kliesen goes on to outline two principal challenges: first, the unpredictability of shocks to the economy and the second, the substantial revisions to the data. The following section describes the revision problem in detail.

1.3 Macro Data Are Noisy

1.3.1 The Revision Problem in Traditional Data

Traditional data mentioned above tend to have substantial revisions which makes it difficult to truly gauge the current state of the economy. As Kliesen (2014) comments, "...revisions to data compound the difficulty of correctly identifying shocks and their significance in real time..."

As a motivating example, we consider an important traditional data source- the GDP estimate of the fourth quarter (Q4) in 2007. Figure 1.2 shows that any central banker trying to get a true picture of the economic growth in Q4, 2007- which would be helpful in formulating a policy response to financial crisis of 2007-2008, may not have accurate GDP data until 2013. The estimate of Q4, 2007 GDP for United States varied from $+2.9\%$ to -0.2% which would make it difficult to be confident about one's policy response. Prior academic research (Orphanides, 2001; Orphanides and Williams, 2006) documents that even if the conduct of monetary policy were simplified to a simple linear function[5], the final rule prescription for the interest rate could be different than the current interest rate by up to 1% to 1.5% as compared to the initial rate in each period due to the data revisions!

Revisions are also a problem with standard empirical analysis in the academic literature. In fact, Orphanides (2001) comments "Indeed, standard practice in empirical macro-economics is to employ ex-post revised data for the analysis of historical time series without adequate investigation of the possible consequences of this practice on the results."

[5]For example, the "Taylor rule" that trades off the level of employment (output gap) vs. the inflation for an economy

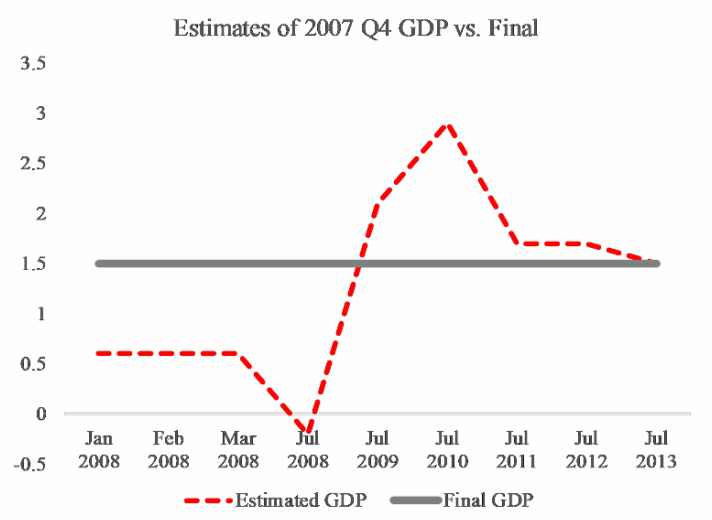


Figure 1.2: Estimates of 2007, Q4 GDP over time. X-axis shows the time when these estimates were published. Source: Kliesen (2014), Bureau of Economic Analysis (BEA) and Bloomberg.

1.3.2 Increased Noise in Times of Low Growth

More formally, in Table 1.2, we analyze the revisions of some the most popular economic statistics across the key sectors of the U.S. Economy such as overall economic growth (GDP), employment (NFP), retail sales and home sales. We select the most important data release for the sector and compute the absolute value of the revision (Final-Initial number) and subsequently divide that by the absolute value of the Final number, which results in a percentage revision and conveys the magnitude of the revision relative to the value of the data itself.

Table 1.2: Important US data releases revision percentage. Revisions are computed using absolute values of revisions and the statistics. * indicates revision values are different in good and bad economic times at 5% or more significance. Data Source: Bloomberg and agencies above. Data from 1998 onwards.

U.S. Data Release	Data Source	Overall mean	Overall volatility	Overall Percent Revisions	Revisions in "good" economic times (% of abs value)	Revisions in "bad" economic times (% of abs value)	Volatility in "good" economic times	Volatility in "bad" economic times
GDP growth (QoQ)	Bureau of Economic Analysis	2.20%	2.40%	37%	26%	80%*	1.35%	2.13%
Non-Farm Payrolls (change mom) in thousands	Bureau of Labor Statistics	96.4	212.1	31%	26%	46%*	77.5	214.6
Retail Sales Ex-Auto and Gas (change mom)	U.S. Census Bureau	0.33%	0.42%	67%	47%	136%*	0.25%	0.26%
Pending Home Sales (mom)	National Association of Realtors	0.05%	3.15%	62%	67%	57%	2.04%	3.27%
Average Revision				**49%**	**41%**	**80%**		

We find that on average these survey-based data have an average absolute revision of 49% of the magnitude of the data release! Typically, when economic growth as measured by GDP is low or when the economy is losing jobs, economists label such time "bad economic times" since the economy as a whole can consume less. Prices for most assets fall in such times since they tend to be positively correlated with growth and consumption, and risk averse investors demand extra return for assets that exhibit such behavior (Cochrane, 2001). For the purposes of Table 1.2, a bad economic time for a sector is defined as the time when that particular sector exhibits below trend growth. In such economic bad times, the government or the central bank may be more likely to attempt mitigating measures, but to respond optimally, it is critical to measure just how "bad" things are. As Table 1.2 shows, the magnitude of the revisions rises to around 80% in bad economic times vs. 41% in the good economic times. Even in absolute terms, the measurement error is higher or about the same across these main categories of economic releases.

For example, the absolute revision in bad economic times for GDP is around 1.3% in an economically bad or lower growth time vs. 1% in good or higher growth times, and for NFP the absolute revision is about $70,000$ jobs in bad economic times vs. $57,000$ jobs in good economic times[6].

1.4 Our Goal: Real Time Macro Data With Less Noise

1.4.1 Nowcasting

Nowcasting is the practice of high frequency contemporaneous forecasting of variables such as a growth rate in terms of other higher frequency and noisy variables such as weekly data releases like unemployment claims, or monthly data releases such as NFP, unemployment rate, consumer confidence surveys or manufacturing indices. These more frequent data releases used, though not direct measures of economic growth, tend to have a statistical and intuitively justifiable relationship with economic growth and the benefit of being available faster than the quarterly official GDP number. Thus, by tying together various data sources in a regularly updating statistical model, nowcasting maps the economic trend in real time.

1.4.2 The Pyramid like Framework of Nowcasting

The typical framework for nowcasting tends to resemble a pyramid. This pyramid shape comes from a few key assumptions: Typically, we assume that there are many noisy inputs that are structurally imperfect measures of the unknowable but critical few- the key metrics such as economic growth, inflation, or employment that capture the "state of the economy" almost completely. Thus the "noisy many" naturally form the bottom of the pyramid and the signal to noise ratio is supposed to increase as we progress upward towards the "meaningful few."

Economic theory deals with these "state variables" at the top of the pyramid that remain hidden in the noise and the task of the empiricists is to "separate the wheat from the chaff" or select the right set of noisy inputs, and then use statistical analysis techniques (say a Kalman Filter type framework) to construct the best estimate for the hidden state variable.

[6]Fed governor John C. Williams also mentions how the differences in estimates of real-time and current (revised) GDP were much larger in 2009- closer to the crisis and have subsequently reverted back. https://www.frbsf.org/economic-research/publications/economic-letter/2017/november/problems-predicting-potential-output/

There are many good approaches to modeling the economy in real time and different research groups focus on different sectors of the economy per their prejudices, needs and data availability. Any reasonable implementation of real time models or "nowcast" needs to deal with differing frequencies of data release, revisions, time lengths, reliability. Aruoba et al. (2008) is a good reference for a nowcast framework as well as an empirical implementation of it.

In our group, for a particular implementation, we focus on constructing a model of the economic growth based on employment, retail, auto, housing, and consumer credit sectors. These sectors were chosen because of the particular frameworks (or models) the researchers find illuminating for a particular task (asset allocation) and are also shaped by what data are available- most alternative data available to us fell in those sectors.

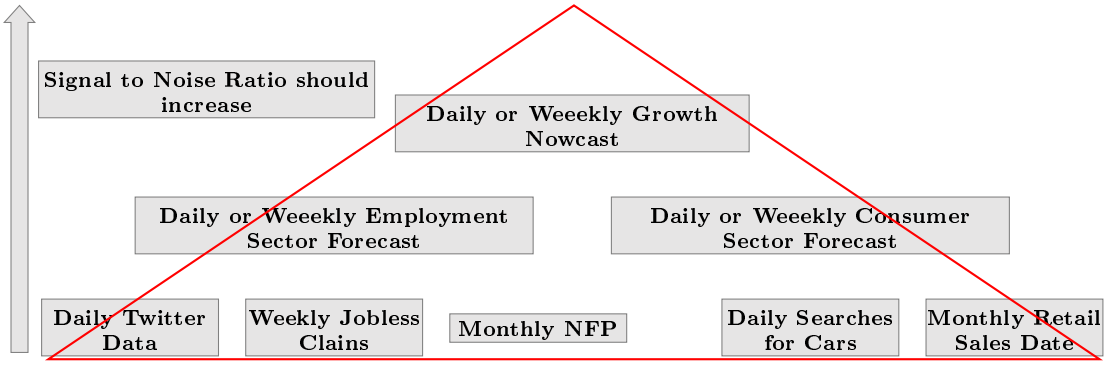


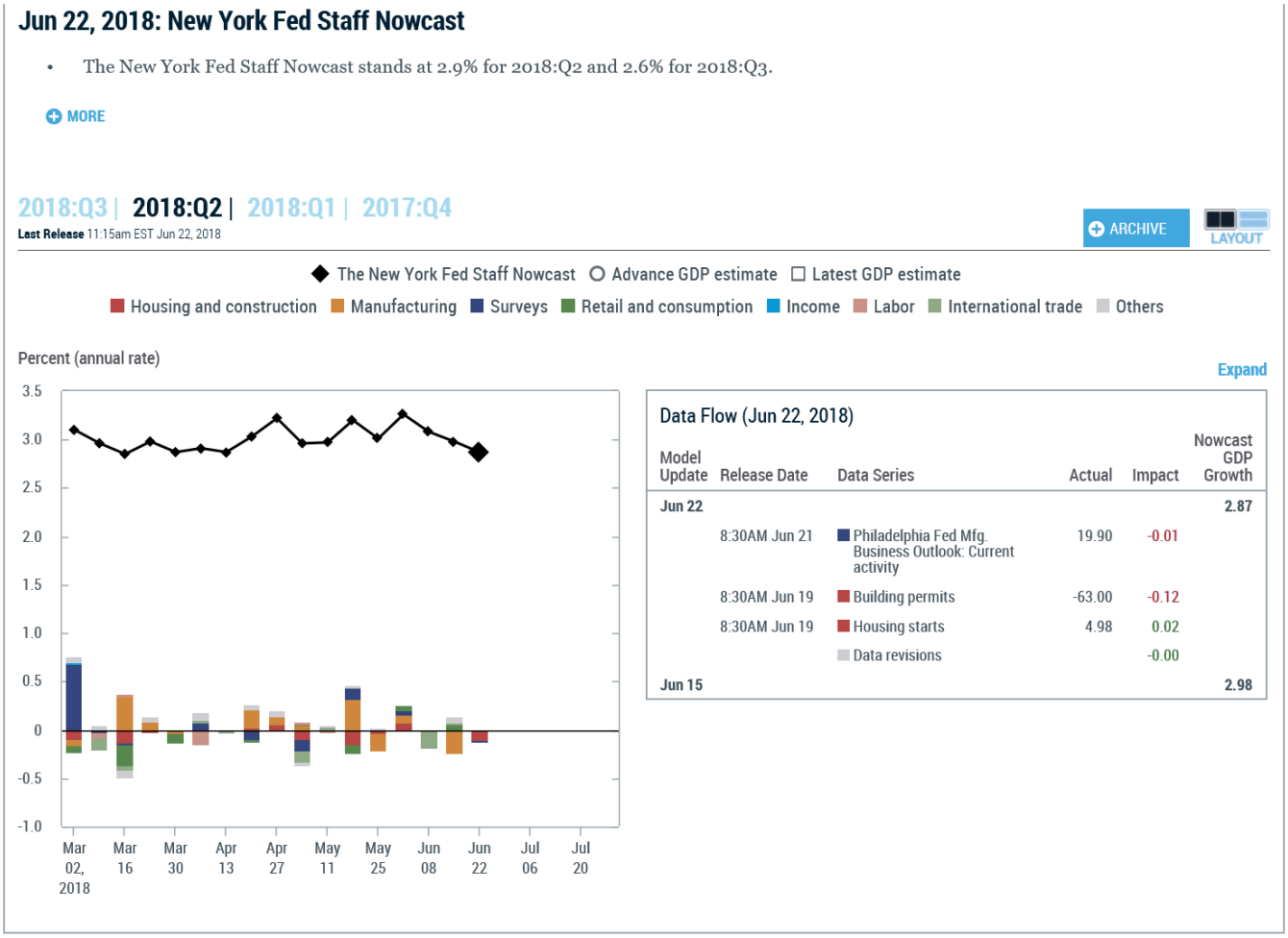
Figure 1.3: A Typical Implementation of a Nowcasting Framework. Noisy data sources at the bottom aggregated into a higher signal to noise ratio "state variable" estimate closer to the top.

Nowcasting vs. Alternative Data: Occasionally, these terms are confused, so we clarify the difference here. Nowcasting is a methodology that provides an on-going and high frequency estimate of an economic concept such as growth by combining noisy estimates of varying frequencies. As discussed previously, there are many ways of making a "nowcast." Alternative data (which we will formally tackle in the next section) are a type of data not yet fully incorporated in the economic mainstream that tend to be web-scale and generated from related activities. Practically, these alternative data in the raw form typically have a wide cross section, a short time series, and can have more sharp jumps as compared to the traditional data. Thus, we can have a nowcast with alternative data or without. Also, some alternative data might have predictive properties, so we can also have a forecast based on a similar pyramid like framework.

Interestingly, various nowcast predictions can vary substantially. For example, on the 22$^{\text{nd}}$ of June, 2018 the New York Fed Nowcast for the GDP was 2.9%[7] vs. the Atlanta Fed Nowcast of 4.7% on June 19$^{\text{th}}$ whereas the average GDP growth rate of the US has been around 2.2% after 1998.

The data revision problem can cause a difference in when we estimate the economic cycle turns even when using frequent and important data releases such as NFP. Jain (2018) shows that using the revised NFP numbers which are releases with a 1 to 3-month delay vs. using

[7] https://www.newyorkfed.org/research/policy/nowcast



Source: Authors' calculations, based on data accessed through Haver Analytics.

Notes: We start reporting the nowcast for a reference quarter about one month before the quarter begins; we stop updating it about one month after the quarter closes. Colored bars reflect the impact of each broad category of data on the nowcast; the impact of specific data releases is shown in the accompanying table.

Figure 1.4: NY Fed's GDP Nowcast page on June 27$^{\text{th}}$, 2018.

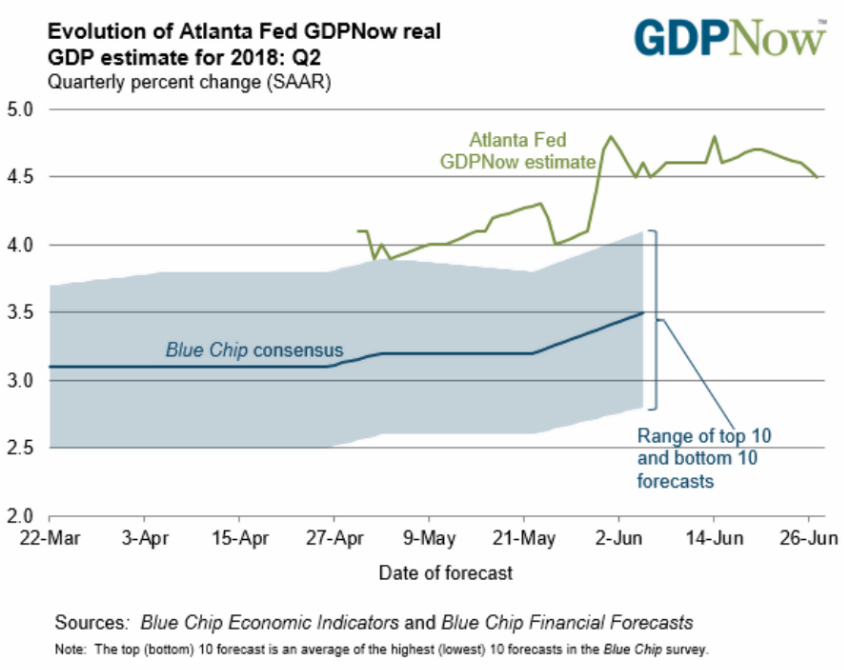


Figure 1.5: Atlanta Fed's GDP Nowcast Page, June 27th, 2018.

the initial releases would cause a difference of around 6 months in estimating the beginning of the 2007-2008 financial crisis. Jain (2018) also formalizes the notion with a Granger causality test which show that the revised NFP numbers have more information about the future path of NFP than the initial NFP numbers.

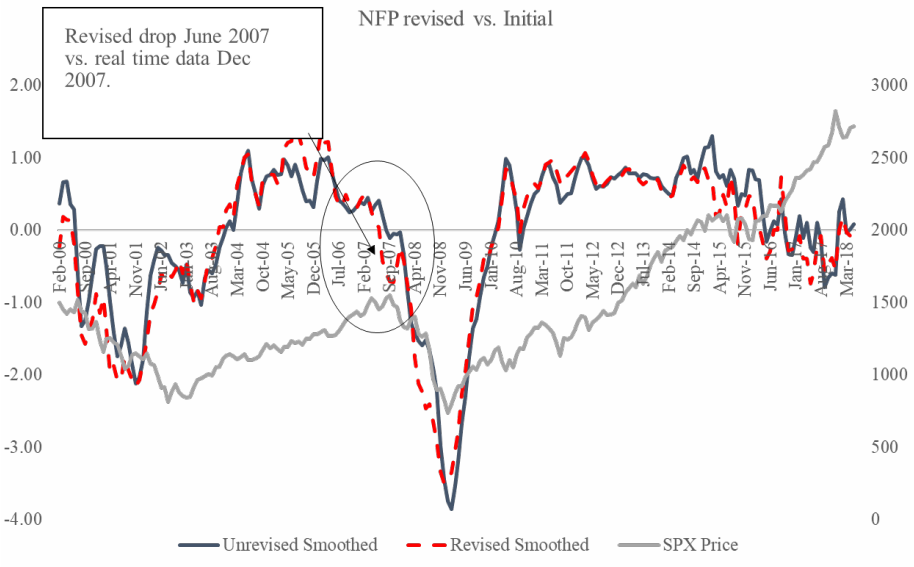


Figure 1.6: Estimating economic cycle turning points using revised vs. unrevised NFP estimates smoothed and detrended. From Jain (2018).

National Bureau of Economic Research (NBER) typically takes between 6 to 21 months[8] after the event to identify that the contraction or recession is beginning or ending. While all the delay cannot be attributed to the noise in the data, the long delay in the context of the result above does suggest that there is a substantial opportunity to do better.

1.5 Alternative Data

Over the last few years several alternative data sources such as browser clicks, search data, twitter, and Instagram data have become popular (Goel, Hofman, Lahaie, Pennock and Watts, 2010). We refer to these data as "web-scale" data due to the high volume of data produced. The desired characteristics of a data series are timeliness, wide coverage, accuracy, and sufficient length which makes comparisons across business cycles and various time periods easy.

Table 1.3: Traditional vs. Web-scale data strengths and weaknesses. The traditional data source is aggregated whereas the web-scale data may or may not be aggregated as per researcher requirement.

Desired Data Characteristic	**Traditional Data**	**Web-scale data**
Example	Household survey for NFP polls $50,000 - -60,000$ homes per month	Self generated 500 million tweets per day of which 5,000,000 may be relevant to economics
Timeliness	1 to 3 month lag normal	Instantaneous
Wide coverage (high recall)	Carefully designed to balance areas and demographics	Urban and tech centric
Accurate (high precision)	Noisy due to small sa1nples and antiquated collection methods	Noise fro1n automation
Long time series	> 20 years for most	~ 5 years
Rich	Difficult to obtain rich data for new questions.	Richness easier to obtain with creative sampling and deeper understanding of data

We would also prefer having as much detail or "richness" as possible so we can investigate any interesting macro phenomenon at a micro level. Table 1.3 compares two data sources – one traditional – the household survey for NFP and the second self-reported employment related tweets to give a sense of the benefits and challenges of both sources.

Traditional data sources such as the household survey for NFP tend to be much smaller in size with about $50,000$ to $60,000$ homes contacted each month. The government conducts interviews which have questions such as "Are you working?", "Did you make an attempt in the last four weeks to find a job?" (Baumol, 2012) The benefits of such traditional data are the careful design, with a stratified sample that ensures balanced demographics and continuity over time. However, the survey is only conducted monthly and due to the low sample size, it may be difficult to tease out if young black males are more discouraged vs. white females who are retiring because of the higher returns in the stock market affecting their wealth portfolio. Additionally, based on the BLS data it seems that the response rate for voluntary household

[8] http://www.nber.org/cycles/recessions_faq.html

surveys such as Consumer Price Index (CPI-Housing) has been declining over time (Figure 1.7) whereas the number of tweets and searches over time as well as the percentage of people who tweet has been increasing substantially (Figure 1.8).[9] In contrast to the monthly household survey, there are about 500 million tweets per day and even if 1% of those are relevant to the U.S. labor economy it is around 5 million responses which is orders of magnitude larger. The other difference is that since people are free to tweet about any topic they like (Proserpio, Counts and Jain, 2016) they would mention some of the issues affecting them when they become important rather than wait for an official to modify the survey to include them.

Quick warning: The alternative data sources that we mention here such as twitter, search, and click data suffer from their own problems such as spam, a high degree of noise, and the lack of independence in the social data sources where popular or trending items receive a higher ranking and hence become positively autocorrelated over time by construction. The fact that this "warnings" paragraph is small should not be taken to mean that alternative data are a panacea, or that the issues stated above are a comprehensive list; it simply reflects the fact that this is an exciting but immature area. We are still learning a lot more about these data and it is too soon to have very firm opinions or even stylized facts that are commonly believed. Additionally, some of the practical difficulties of using such alternative data will become more obvious in the case study that is presented as a major part of this chapter.

Traditional and alternative data sources- comparing information contents and evaluating complementary usage: It appears that we can compare the information content of two datasets in a reasonable way – say we are examining employment data and decide to compare the information about the employment sector in the household survey data and the self-reported employment related tweets. We can check if the alternative data source is statistically significant in predicting the future household employment data release after controlling for Wall Street expectations to account for knowledge in the public domain as well as the previous household employment data releases to account for trend effects. There are two main problems with this- the first a statistical one, the alternative data series are short and noisy rendering the tests not so powerful and if we increased the number of alternative data series- it becomes even more difficult, and the second a philosophical one of the noisy government data (the household employment survey) being held as the gold standard. Additionally, the alternative data contain more detail and rich information from the entire population and we may not be looking at the "right features."

One may argue that the government data move the markets and are important and that is a reasonable argument (in fact we advance this as well) but there may still be a circularity. If we all believe that the central bank responds to the small monthly data collected every month, those data- even if more inaccurate- will continue to have a price impact on the market where central banks can matter and be self-fulfilling[10]. We could imagine a scenario where keeping

[9] Of course, the growth of number of twitter, search and other alternative data sources will slow down and these numbers are not expressed as a percentage of population, but the world population did not increase by 10X from 2010 to 2018, thus the growth rate for active twitter users is a good indicator of the increasing penetration in the population.

[10] One can argue that the efficient market will correct for it, but if there is a large agent (central bank or a government) whose objectives are not purely economic, and there are limits to liquidity that argument may not apply. There are arguments if a central bank is even relevant – in that case, I suggest reading Romer (2016) where he discusses the role of the Federal Reserve in curbing inflation. Of course, the market may eventually converge to better data by yielding excess returns to participants who collect the better data- which is the mechanism by which the market approaches efficiency.

everything else the same, if we all start believing that the government will respond in case a sizeable number of citizens communicate that they find the labor market "difficult" and that the searches for jobs, as well as tax receipts and credit card purchases are decreasing dramatically-those data might have a higher price impact. Thus whatever data we pay attention to as a society also becomes important.

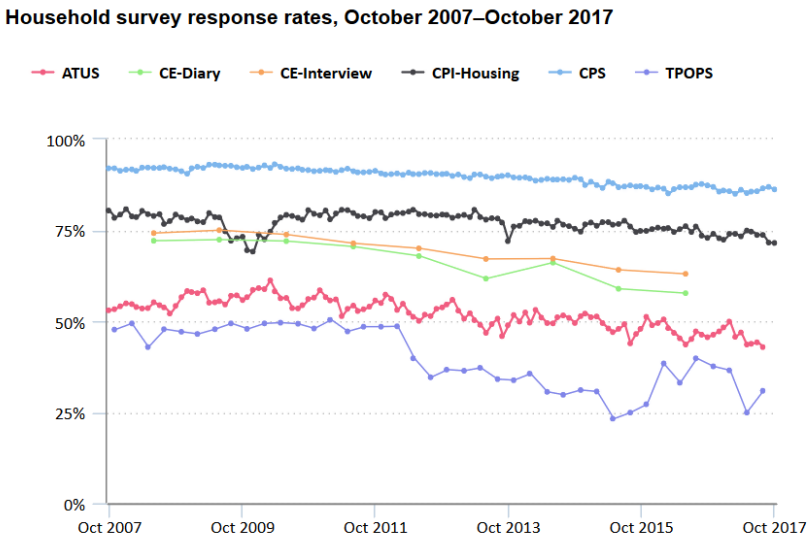


Figure 1.7: BLS Household survey response rates are declining over time. ATUS = American Time Use Survey; CE= Consumer Expenditure Survey, CPI = Consumer Price Index, CPS = Current Population Survey, TPOPS = Telephone Point of Purchase Survey. Source: BLS. https://www.bls.gov/osmr/response-rates/home.htm

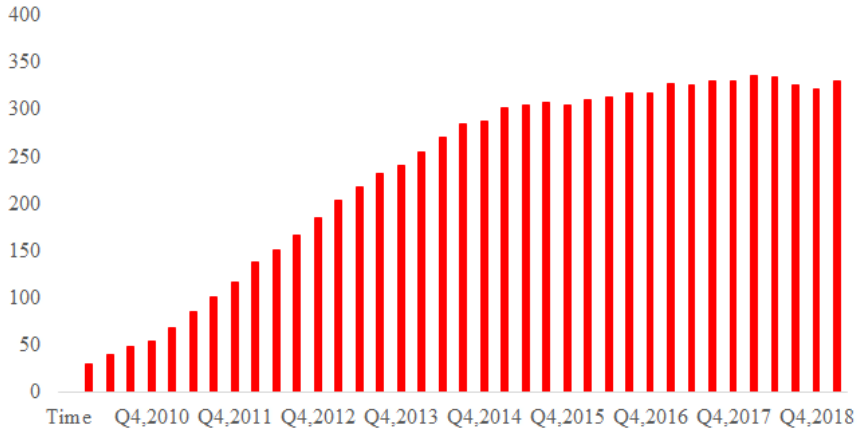


Figure 1.8: Number of active twitter users has grown dramatically over the last 8 years. Source: http://www.internetlivestats.com/twitter-statistics/

Recognizing those limitations, in practice we look for "economic intuition" or a story about why any data matters in addition to the market based tests described above. In our experience, these two types of data end up being complements. Functionally, having the entire population inform us of something (however subjectively) captured as alternative data ends up being more informative and exposes new trends or interesting issues, and rich and frequent detail that the researchers or policy officials may not have thought of, whereas carefully, objectively collected data that are comparable across various cycles (long time series) and cross-sectionally can be helpful in calibrating the policy response. We believe that some of alternative data will become more mainstream as they prove their usefulness over time and thus a better understanding of the substitutability and complementarity between both types of data sources will emerge.

Usefulness of Uncorrelated Noise: The BLS NFP number tends to have a 90% confidence interval of $\pm 115,000$ while the average release is around $90,000$[11] (Table 1.2). Thus, alternative data that do not rely on the same industry surveys for their data could help make the overall growth estimate more accurate – for instance in the nowcasting framework -by providing data about employment from an uncorrelated source. This lack of correlation would cause the standard error to decrease when we combine these sources appropriately. Continuing with the example above, if we could get a similarly noisy but completely uncorrelated data source, the confidence interval for the average NFP release would shrink to $\sim \pm 81,000$ and the release of $90,000$ which was not statistically significant at the 90% level would now be a significant increase in employment, thus helping us have a more accurate picture of the employment sector.

1.5.1 The Micro-foundations of Macro: Alternative Data in the context of the Lucas and Romer Critique

To bring out the potential of alternative data, it might help to take a very brief (and extremely incomplete) tour through the history of economic modeling that nevertheless is useful because it provides a flavor of the challenges faced by economists when modeling our complex economic machine. With macroeconomic crises and technological progress- models may come and models may go but the need for better data goes on forever![12]

There are two main kinds of models in macroeconomics: structural that explicitly use economic theory and nonstructural that "let the data speak". Reiss and Wolak (2007) provide a good framework to understand both. They take the case of a set of observable "endogenous" variables, y, that are related to another set of observable "explanatory" variables, x. In their framework the nonstructural approach would measure both x and y, and then estimate a conditional density using the appropriate statistical methods. The structural approach would seek to clarify how institutional and economic conditions might affect x and y and try to embed that in the model. Thus, the task of the structural model is to clearly make connections "between institutional, economic and statistical assumptions." The benefit of a structural model, is increased context and the inevitable consequence by design is that the structural approach has some embedded theory about how the "economic machine" should work. We human beings exhibit adaptive and complex behavior, which makes the task of modeling our aggregate (economic) behavior challenging if not impossible. For any model there is a tension

[11] https://www.bls.gov/news.release/empsit.tn.htm.

[12] Apologies for the repurposing of Alfred Lord Tennyson's poem- The Brook.

between nailing down what is happening (positive) in all its detail, and what should happen (normative).[13]

Diebold (1998) presents a 10,000-foot view of how macroeconomic forecasting developed over time in an excellent survey paper "The Past, Present, and Future of Macroeconomic Forecasting." He outlines how there is an "interplay between measurement and theory" and how this interplay affects the nonstructural and structural approaches to forecasting. We can see the application of Karl Popper's philosophy of empirically motivated critical thinking in the field of macroeconomic modeling in Diebold's paper. Diebold's narrative follows Popper by showing how theories evolve and become more popular as they better explain current phenomenon than older models and models seem to change more after macroeconomic crises! Keynesian macroeconomic theory was written during and after the great depression of 1930s, as Keynes advocated the use of fiscal and monetary policy to mitigate the adverse effects of downturns in aggregate demand. In response to theory, measurement techniques evolved and many statistical contributions such as those by Fisher, Neyman, and Pearson were made. However, economists were not satisfied with the Keynesian "systems of equations" approach as Prescott (1986) named them. This approach involved ad hoc postulated decision rules about consumption, investment and treatment of expectations, and did not have fundamental justification for "sticky prices" and "sticky wages."

Along with the intellectual dissatisfaction, when empirical facts diverged from the theory in the 1970s, it sent the profession in a new direction of incorporating rational expectations. In the 1970s supply shocks resulted in the economy experiencing both "high inflation" and "low growth"- a phenomenon that was deemed unlikely in the Keynesian aggregate demand driven paradigm involving trade-off between high growth and high inflation. In his now famous "Lucas critique", Robert Lucas (1976) argues that "optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker," where the series refers to the relevant time series that individuals would consider while making their decision. This observation from Lucas is at the heart of the issue he raises about the "clear-cut conflict" between the structural and nonstructural modeling. Lucas outlines the policy maker's problem with nonstructural models by pointing out that purely econometric models that use aggregate past data and are successful at predictions "in principle, can provide no useful information as to the actual consequences of alternative economic policies.[14]" Diebold outlines how economists such as Fair and Taylor incorporated more "realistic behavior" in the "systems of equations" approach in the form of rational expectations into econometric models as well as a better feedback loop by including model assessment and fit into the framework. Thus rescued, these models are used by policy organizations including central banks. The same energy shock that sowed seeds of doubt about the existing theory resulted in more interests in nonstructural approaches. Diebold mentions an important paper by Sargent and Sims (1977) titled "Business Cycle Modeling Without Pretending to Have too Much a Priori Theory."

Then Diebold moves to the theoretical flavor of the day dynamic stochastic general equi-

[13] Modelers are not unaware of the difficulty and hence there is a common saying- "all models are wrong but some are useful" which is attributed to statistician George G. E. P. Box (1976). I personally find two of his phrases quite useful: that a good model should seek to find an "economical description of natural phenomenon" and be "simple but evocative." (1979).

[14] Lucas provides several examples- including one of estimating actual consumption of a consumer via permanent income hypothesis. In this example a purely statistical model to predict consumption based on past data would not capture policy changes that are known in advance, but a theoretical model would have better success.

librium (DSGE) models that allow for rational agents to make optimal decisions rather than rely on an "ad hoc" system of equation. In this neoclassical framework technology shocks by and large explain the business cycle. Fast forward to the global financial crises of 2008, and the profession is again filled with voices for change. It is in this context that Romer's (2016) critique can be examined. Romer makes the point that the current modeling machinery has gotten so cumbersome that model parameters cannot be statistically identified and the attribution of most of business cycle fluctuations to technology shocks reaches "bewildering conclusions".

This path of macroeconomic theory and modeling is normal for any science, whether we understand these events as the paradigms in Kuhn's framework exposited in "The Structure of Scientific Revolutions" or we think about the incremental and better empirical content test of Popper. What is common in all these tremendous intellectual contributions, particularly in the critiques from both the Nobel laureates (Lucas in 1995 and Romer in 2018) is a desire to get to a "fundamental understanding." This fundamental understanding of human behavior and embedding it into an economics framework has been a common dream of researchers often called the "microfoundations of macroeconomics." (For example Barro (1993)). In his critique, one the issues Romer takes with rational expectations and DSGE models is about the lack of the microeconomic evidence: "There is no microeconomic evidence for the negative phlogiston shocks that the model invokes..."

The future? Now imagine, consumers were empowered to communicate real-time and individual statements about their understanding of the government's policy and state their own response! The resulting analysis of policy impact would be unusually rich because of the individual level actions as well as a record of how people got to their decision. People would self-report (to the extent comfortable) their emotions as well as the information they sought in making decisions. Thus, we could actually see how temporary or permanent is do they believe the shock to the income is and we would also get a bit more microeconomic evidence on how people behave. Such research is now a reality and herein lies perhaps the biggest promise of these "web-scale" alternative data which capture for the first time, voluntary, individual user generated, self-reported data at a large and accessible scale. Thus, we can supplement purely theoretical arguments and our assumptions with better data. Over time as we get newer and more detailed micro data that truly describe the behavior of individuals perhaps our theories would improve. These microfoundations of user behavior might eventually help us better predict responses to policies that cannot be modeled purely statistically as Lucas (1976) suggests, or as Soros' reflexivity (2008) might imply, the data will change in response to the actions taken! In any case, these data will lead to interesting research.

Some caution: If we take this idea of more data to its logical conclusion, it seems that we could fulfill the dream of an accurate economic map in the spirit of the humongous cartographic maps that Jorge Luis Borges refers to in "On Exactitude of Science." Here we must be careful and realize that simply more frequent data may not be enough- for example, common sense would suggest that measuring how one particular business performs on a minute by minute basis in an economic expansion one day is unlikely to clarify how that business will fare in a world with different tariffs and an economic recession. However, a detailed model of individual consumer preferences and their sensitivities to price might increase our ability to price a product optimally or predict better than before how they might react to a competitive product. Additionally, it is critical to have consumers understand the value of their data as well as their rights regarding it to prevent abuse of this power – a kind of "data colonialism"

(Couldry and Mejias, 2018).

Using individual level psychological variables to predict national unemployment: An example of taking advantage using alternative micro data to understand more "fundamental behaviors" that lead to macroeconomic prediction can be found in Proserpio et al. (2016) where the paper predicts one month ahead national level unemployment rate for United States using individual twitter data over time. Proserpio et al. (2016) analyze more than individual users from 2010 to 2015 who lost or gained a job using a "differences in differences" estimation. They analyze the effect of job loss and gain on intuitive psychological variables such as anxiety, sadness, and anger and then define a behavioral macroeconomic model that selects the relevant variables using stepwise regression. Using only the psychological variables, their model predicts the national unemployment rate for the U.S. with half the MSE vs. an autoregressive (AR) model. Interestingly, they also find that a significant correlation between the average level of anger in the population to how difficult jobs are to obtain ("Jobs Hard to Get" measure by the Conference Board in their consumer confidence survey).

Figure 1.9: From Proserpio, Counts and Jain (2016). This shows the users' psychological variables before and after the employment economic shock which happens at time 0 (in months).

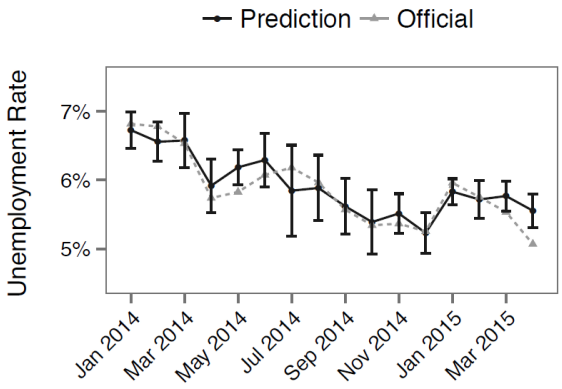


Figure 1.10: From Proserpio, Counts and Jain (2016). This figure shows the rolling 1 month ahead predictions of U.S. unemployment rate from the behavioral macroeconomic model that relies only on the psychological variables from 1.2B tweets.

Such studies were not possible, say a decade before, where not only would the lack of computational power be an issue but these types of self-reported data from large groups of population at such a high frequency were not easily available. This growth in computational power and data has enabled a rich literature in the field of social computational sciences and in the more traditional domains of economics.

1.6 A Framework For Alternative Data

The JPM Big Data and AI Strategies Guide (2017) suggests there are 3 main types of data available: individual activity, business processes, and sensor data. Examples of individual activity data would be data like social media, news and reviews, and web searches, business process data would be data like credit cards or accounting data, and sensor data would be satellite data. The JPM guide estimates the current size of the Big Data, related technology and analytics market at $130B, and it is expected to grow to over $200B by 2020. There are more than 500 data providers and they fall into providers or raw data, semi-processed data and final predictions or signals for investment industry.

The formal framework we will adopt in this chapter is something my team uses in practice. This framework is based on our goal of modeling macroeconomic data and we will provide both practical and theoretical motivation for it. In a practical sense, imagine we have a certain data budget and we are trying to assess how much money should we pay for a source of data a vendor is presenting to us. A natural question we would ask is:

> "*How much incremental information will this data source provide relative to the amount of money we would pay for it?*"

In the context of modeling various sectors of the economy (or the stock market) one can divide this incremental information question into two sub-questions:

Question 1: How much extra information (edge) vs. existing data sources does a dataset provide for a given sector or a stock?

Question 2: How many sectors or stocks (breadth) does it cover?

If we are thinking in terms of cross-sectional stock prediction, these questions will appear to be in the spirit of the "fundamental law of active management" (Grinold, 1989) which measures the "information ratio" or volatility scaled excess active returns achieved by an investment manager that can be described as the information coefficient (edge or skill) multiplied by the breadth (typically the square root of the number of independent forecasts).

More formally, the precision and recall language is used in the field of Machine Learning (ML) and is also related to type I and type II errors of statistics.

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{1.1}$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{1.2}$$

Thus precision (say in a classification problem sense) is the fraction of values identified in a certain class that do in fact belong to that class and recall is the fraction of values of the certain class identified by the classifier. The sense in which we are using precision and recall

in the data context are slightly different than a classification problem. When evaluating a data source, precision and recall refer to our expectation of the incremental value generated from the data source based on how we believe data intrinsically relate to the phenomenon we are trying to model. For instance, we are aware that most people use credit cards, and all firms fulfilling a certain asset size criterion with some equity securities must file accounting statements like 10-K and 10-Q. These time series for the accounting and credit card data also covers several business cycles which is helpful for modeling purposes. In addition, we may have beliefs about the quality of the data: How precisely do they actually measure the phenomenon. Does this credit card data provider accurately capture the spending on most consumers? We may match it to existing data and verify various statistical properties of the data. In practice, substantial due diligence is required before making a data purchase and even with continued usage. After the purchase, we match if our expectations of the information content of the data were "well formed."

.

Table 1.4: Classification of alternative data. Some types of data from JPM Big Data and AI Guide (2017). Estimation of precision and recall from the author's research and judgement.

Data type	Accuracy (Accuracy for the the sector)	Recall (Number of sectors covered)	Timing	Comment
Satellite parking lot	Low	Low	Ex-post	We expect fast improvements. in accuracy in this sector
E-commerce transaction	High	Low	Ex-post	
Footall traffic	High	Low	Ex-post	
Mobile data	Low	High	Ex-post	
Email receipt	High	Low	Ex-post	
Product review	High	Low	Ex-post/Ex- ante	Intial reviews may predict future sales and reviews
Twitter	Low	High	Ex-ante	Investor mental state from tweets can help predict future actions.
Credit and debit card	High	High	Ex-post	
Payment systems	Low	High	Ex-post	
News	High	High	Ex-post	
Web search	Low	High	Ex-ante	People generally search before they act
Microstructure data	Low	High	Ex-post	
Tax return data	High	High	Ex-post	
Accounting data	High	High	Ex-post	
Mortgage application data	High	Low	Ex-post	
Point of sale data	High	Low	Ex-post	
Intercompany payments	Low	Low	Ex-post	
Order and shipment tracking	High	High	Ex-post	

For purposes of exposition we have divided the data into 4 main boxes with high and low precision, and high and low recall based on the author's beliefs. For instance, we believe that currently satellite data is helpful only for a few sectors of the economy such as retail or energy since it can capture the number of vehicles parked in malls or the equipment being used by energy companies.[15] Over time, we expect that technological improvements such

[15] There are always game theoretic aspects to these or almost any data, where we may believe that some

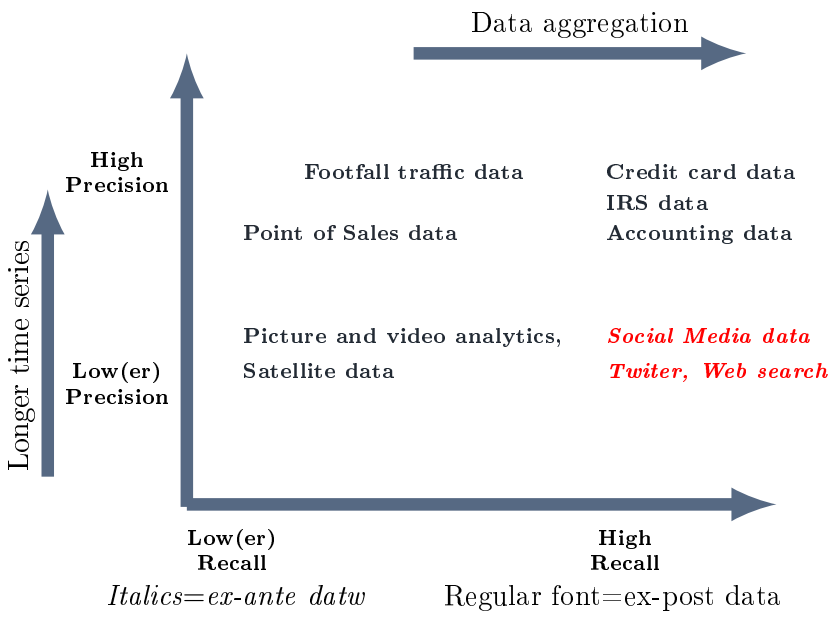


Figure 1.11: Classification of data sources along expected precision and recall axes.

as better satellite pictures of cloudy days will increase the precision of data and the time series will naturally grow in length and cover more business cycles. Additionally, intelligent data aggregation has tremendous potential to increase the recall or the population coverage of alternative data. For instance, if we want to get a better understanding of the overall economic activity we will benefit from matching credit card data, tax data and companies accounting data by each district.

Social or user generated online data such as search, and tweets are different in one crucial aspect- they are more predictive of user's future actions (ex-ante) rather than an extremely fast capturing of the actions that have already occurred (ex-post). To take the example of search and employment, as Chancellor and Counts (2018) note- 80% of job seekers use search engines to find new employment, thus the data source has a high recall. Given the inevitable automation that must occur to process such large quantities of data as a search engine and the arbitrary and ambiguous nature of queries, there is inevitably some noise. However, there an interesting forward-looking quality to search data. A user currently searching for employment is more likely to find employment (everything else equal) in the future and spend more money on consumption: The increase in search for employment is noisily measured but predictive of the future credit card spending that will be quite precisely measured but occur later. Thus, there is a natural complementarity between the quadrants- especially the ex-ante data and noisy data like search and ex-post and precise data like credit cards or accounting.

In this chapter we will mainly explore the application of web search data to predicting macroeconomic variables.

companies could respond to increase the particular metric being measured (like encouraging more cars to park in its lots) rather than the output (actual sales) for short term gains or to diffuse the information content in the data. For now we are ignoring these aspects.

1.7 Predicting Data Releases With Search Data

Web searches are queries users type into a search engine to obtain information. Choi and Varian's (2012) seminal work analyzed aggregated Google Trends (now Google Insights) data using a query index. As they describe it, "the query index starts with a query share: the total query volume for search term in a given geographic region divided by the total number of queries in that region at a point in time. The query share numbers are then normalized so that they start at 0 in January 1, 2004. Numbers at a later data indicate the percentage deviations from the query share on January 1, 2004." Using these queries Choi and Varian predicted the future releases of present economy activity several indicators such as retail sales, automotive sales, home sales, and travel. They insist they are not forecasting the future but "predicting the present" by merely aggregating and counting faster than the government which releases the various economic indicators mentioned above with a lag.

The search data we will use will be based on the Microsoft search engine Bing. Figure 1.12 shows that the Bing search data do not have large geographic or demographic biases by comparing the population of Bing users to the U.S. population. Figure 1.13 shows that the market share of Bing over time.

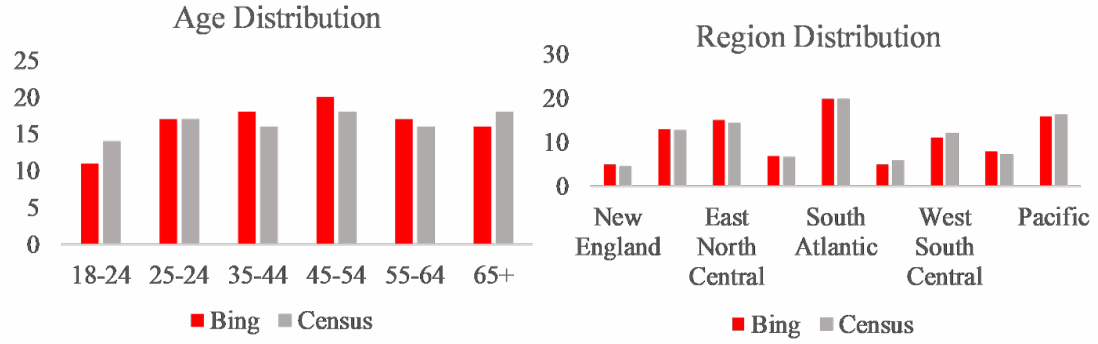


Figure 1.12: Demographic and Geographic distribution of Bing searches vs. U.S. population. Source MSFT and US Census. Numbers are in percentages.

As discussed, employment concerns policy makers as well as investors and individuals. Accordingly, we will model the NFP data release. For modeling NFP we will follow the careful curation procedure of Chancellor and Counts (2018) instead of the aggregation of Choi and Varian (2012) that tends to lose a lot of the detail and introduces considerable noise due to heavy normalization described above. The data contain a sample of English language queries from the years January 2012 to March 2018 from mobile and desktop devices. They filter the queries for the appearance of four keywords: "job", "jobs", "career", and "careers". Other terms such as "employment" generated more false positives and hence were dropped.

Subsequently, a high-level job category classifier that follows the BLS job classification on a by and large basis was used. There are 18 potential categories as well as the generic job query category that just captures the nonspecific queries like "job in Seattle". Two researchers generated an initial set of 500 job titles take from BLS and Census data and searches of job sites such as Glassdoor.com. Subsequently, the two researchers hand-annotated which if the

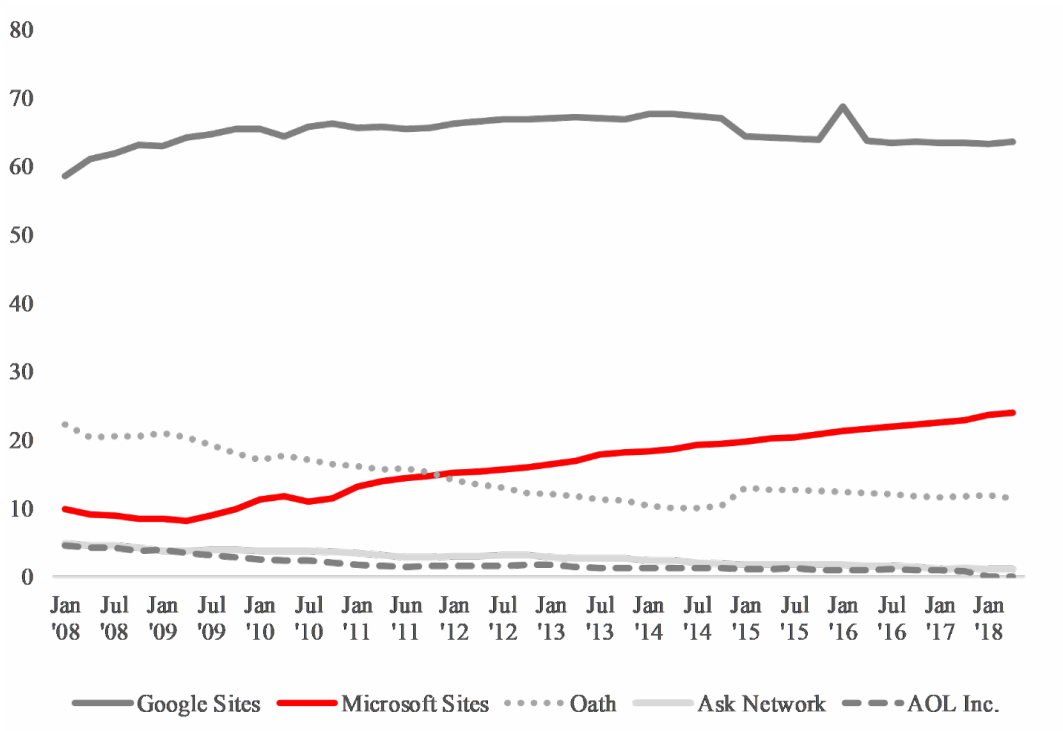


Figure 1.13: Search Engine market share over time. Source: comscore.com. Microsoft sites corresponds to Bing.

18 BLS categories, the random sample of 250 queries matched. These lists were updated iteratively until they found less than 10 new titles per 250 searches. Additionally, Chancellor and Counts validated these searches on a dataset with 10,000 searches by running a string matching system.

Table 1.5: Examples of searches for the 15 job categories in the Employment Sector. Source: Chancellor and Counts (2018).

	Raw #	%	Examples from Our Dataset
Generic Job Queries	184,500,000	82	"careers in.", "fedex careers", "most commonly asked interview questions for a job"
Job-Specific Queries	40,500,000	18	-
Total	225,000,000	100	-
Architecture/Engineering	648,000	1.6	"entry level biomedical engineering jobs", "auto-cad jobs in central IN", "engineering carters firearms"
Art	1,539,000	3.8	"freelance writing jobs for beginners", "voice acting careers", "winterthur museum curator job".
Business	3,118,500	7.7	"vp of operations job", "marketing and product preference jobs", "hr career springfield ma"
Construction	1,134,000	2.8	"construction laborer jobs in reno nv", "welder jobs in wisconsin", "construction inspection jobs 06415"
Education	6,520,500	16.1	"community college professor jobs", "atlanta nanny jobs", "washoe county school district careers"
Finance	4,090,500	10.1	"financial banking jObs in VT", "mortgage lender careers". "medical insurance specialist job"
Food	810,000	2.2	"bartender jobs in minneapolis", "foodservice jobs", "craigslist dishwasher jobs".
Healthcare	12,555,000	31.0	"surgical tech jobs". "mental health jobs westernmass". "clinic job rm jax"
Leisure/Hospitality	1,822,500	4.5	"hollywood casino jobs", "fitness jobs in new hampshire", "laundry jobs in hotel"
Manufacturing	1,377,000	3.4	"machine operator jobs in columbia sc". "jobs in shipfitting in jacksonville", "machinist jobs in nj"
Retail	1,255,500	3.1	"electric boat jobs", "clothing store job applications online", "retail career at outlets near me"
Science	850,500	2.1	"psychology research associate jobs", "jobs in r&d in dc", "boston scientific careers"
Technology	1,701,000	4.2	"computer jobs in the army", "software architect career", "sql dba jobs near me"
Transponation	3,078,000	8	"Chicago airport runway jobs". "cdl jobs in boise id", "railroad jobs in kansas"

1.7.1 Why Curate? The Google flu story

As is clear from the details of the employment data curation mentioned previously, data curation is a painstaking process. Part of the reason we go through this effort is to guard as much as possible against outcomes like Google Flu as elegantly covered by Lazer et al (2014) in "The Parable of Google Flu: Traps in Big Data Analysis."

Google built an indicator- Google Flu Trends (GFT) to predict influenza like illnesses (ILI) that were supposed to predict the future Centers for Disease Control and Prediction (CDC) numbers. After initial success, in February 2013, GFT was predicting more than double the

proportion of doctor visits for ILI. The authors talk about two main reasons for the failure- the first is big data hubris- the idea that big data are a substitute for traditional data collection that ignore than foundational issues of measurement, and the second the algorithm dynamics that were changed as Google changed its main product- the search engine- it changed the data generating process. The way we attempt to address the first issue of hubris is to work with subject matter experts and recognize that the data generating process are not accurate and can change. Constant monitoring of data properties as well as getting as much detail as possible on the data generating process (the search engine in this case) are the main methods we use. To address the second issue of algorithm dynamics changing, we update and select our own keywords – as detailed previously- rather than rely on a passive usage of the Bing search engine.

1.7.2 Modeling differences rather than levels

When making a model to understand an economic sector better, we have the choice of modeling levels or differences. While, there is no strict rule, currently we lean towards modeling differences. The authors Lazer et al raise a conceptual point about the auto-regressive (AR) model capturing 90% of the trend and hence question the potential benefit of using alternative data at higher granularity. They go on to suggest that alternative data may be better for providing details rather than economics metrics of interest. While, we agree with their idea about using search or social data having the ability to provide detail, but believe predicting the deltas or the first difference of $(\text{Value}(t+1) - \text{Value}(t))$, rather than only $\text{Value}(t+1)$ may address their primary concern. More formally, imagine $\text{Value}(t+1)$ which could stand for the level of unemployment has a quite auto-regressive nature

$$\text{Value}(t+1) \sim a + b \cdot \text{Value}(t) + E(t+1) \tag{1.3}$$

Where b is close to but smaller than 1 and the R sq. is close to 85 to 90%, it may be better to define the innovation $R(t+1)$ as:

$$R(t+1) = \text{Value}(t+1) - \text{Value}(t) \tag{1.4}$$

We could model $R(t+1)$ which will have better statistical properties and considerably less predictability than the AR model. Back to our employment sector, it means modeling the monthly increases in jobs such as non-farm payrolls (NFP) rather than the current unemployment rate. In fact, the delta modeling may also be more interest in economics and finance since it is the unpredictable part of the economic data that matters to economic decisions and moves asset prices rather than the largely known trend as Cochrane (2001) explains.

We also find that modeling deltas tends to be more robust. We can imagine that changes in search happen for various reasons discussed previously such as changes in search engine behavior, user technological preferences, seasonality, and the economic phenomenon being modeled as informally summarized in equation (1.5) below.

$$\begin{aligned} \Delta\text{search} \sim\ & \Delta\text{search engine behavior} + \\ & + \Delta\text{user technological preferences} \\ & + \text{seasonality} + \Delta\text{economic behavior} + e(t) \end{aligned} \tag{1.5}$$

As Lazer et al mention that, abrupt search engine behavior change can have major impacts on estimation. We find this to be true when we examine the correlations of NFP to the levels of employment related searches vs. deltas or the month over month changes in searches.

If the level of searches is not adjusted for Bing market share which was growing rapidly, we find very different results than if we do adjust the searches for the Bing market share. In contrast, deltas also tend to be more robust to the misspecification we can see in Figure 1.14. The blue colored bars are the correlations with various employment search categories levels and monthly changes (deltas) and NFP. Both blue bars have negative signs across various categories confirming an intuition that as the labor market improves and more people are employed, the number of searches per job decreases.

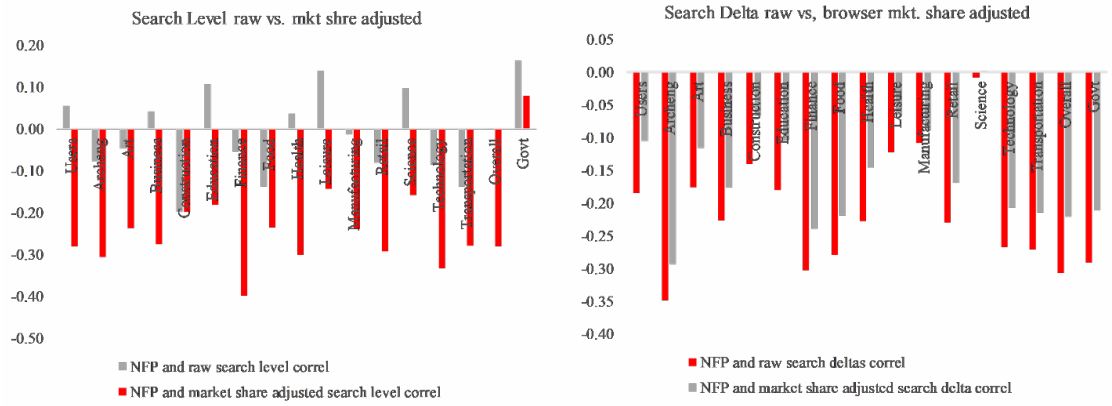


Figure 1.14: Comparison of correlation of NFP with levels and deltas when adjusting (blue) or not adjusting (orange) for market share.

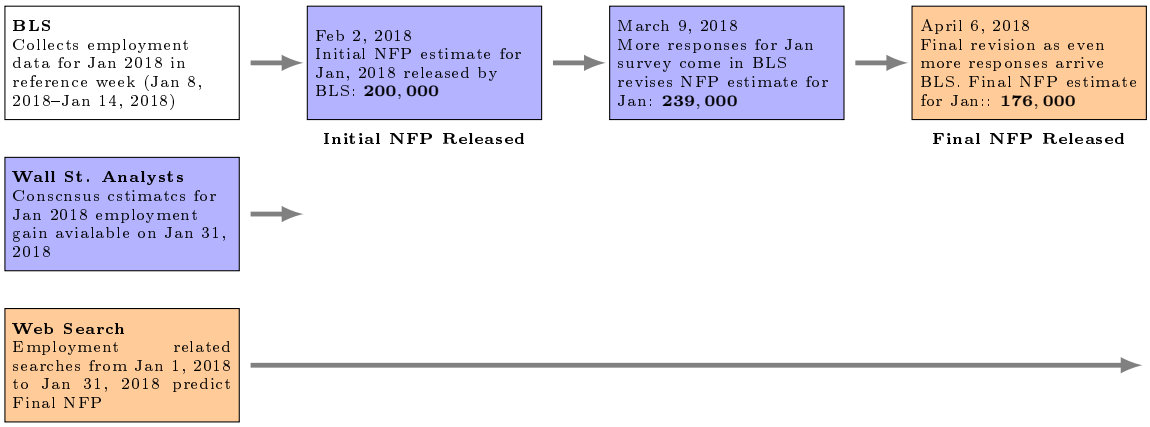


Figure 1.15: An indicative timeline of NFP Initial and Final data release and the prediction from economists' consensus and search-based variables

When we examine the orange bars which represent the raw counts of both search levels and deltas we see that the search levels change sign and the relationship stops having the intuitive sign with about half the categories showing a positive sign and the other half a negative sign, essentially the noise has drowned out the signal for levels. However, for the case of deltas

whether we adjust the employment searches for the Bing market share or not the results are quite close. Of course, the more formal identification would attempt to estimate equation (1.5) or some similar set up.

1.7.3 Housing, Retail, and Auto sectors with alternative data

Like employment, we curate search data[16] for other economic sectors such as home sales, retail sales and auto sales. The various categories and their correlation with the corresponding government data release and the 10^{th} and 90^{th} percentile of the bootstrapped confidence intervals are displayed in Table 1.6.

As an example of the data cleaning process we show the raw Bing searches related to existing homes. The particular terms related to homes were chosen carefully in a manner similar to employment by a team of data scientists, researchers and economists familiar with estimating macroeconomic housing statistics. The raw searches display considerable seasonal patterns as well as "spikes" which indicate an unusually high volume of searches. We adjust the data for the market share of Bing, winsorize it to remove the high search volumes since the trending nature of searches implies that popular results will be more likely to be shown to users and hence the relationship between the volume of searches and the economic phenomena will be different at different levels of search, remove the seasonality and average the data over a month to relate it to the monthly economic releases from the traditional data sources. Figure 1.16 shows the cleaned level of search data corresponding to home searches and we relate it to the actual level of the US home sales. The relationship seems reasonable upon visual inspection.

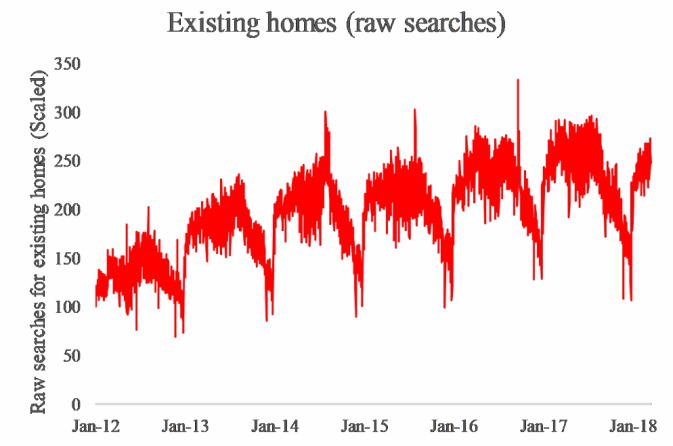


Figure 1.16: Raw Bing searches related to existing homes. Source MSFT. January 2012 is normalized to 100 for graphical purposes.

[16] Ethics Review and Data Protection

This study was found in line with the Common Rule for exemption by the Microsoft Research Ethics Advisory Board under protocol 7. Our data was gathered historically; there was no interactions with users by changing search results. All data was anonymized and aggregated to county level. No session information is used in our dataset. Our use and storage of this data is in agreement with Bing's End User License Agreement and Privacy Policy.

The actual modeling does not use levels but the month over month changes or deltas for reasons described above. Table 1.6 shows the correlation across employment search features with non-farm payrolls (NFP), month over month retail sales changes (excluding autos and gas), pending homes sales (PHS), and auto sales. We see that the correlations follow intuitive signs with searches for employment declining when the economy improves, but searches for cars, housing, and auto features by and large increasing in the same condition.

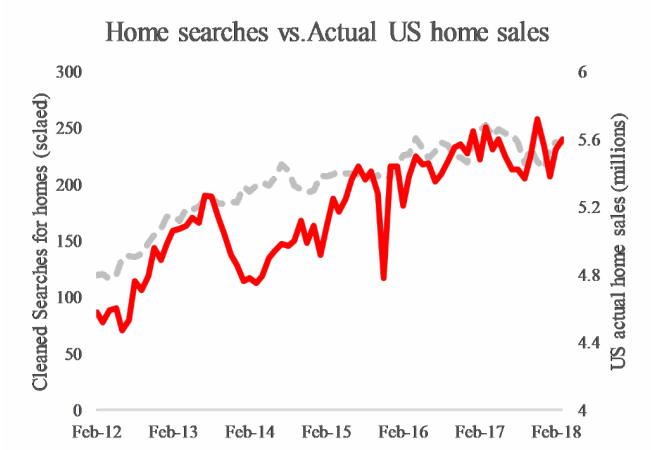


Figure 1.17: Raw searches for existing homes on Bing. (January 2012 is normalized to 100).

Table 1.6: Correlation between curated search features for economic categories and traditional economic data. The bootstrapped 10th and 90th percentile are also presented.

Official Data Category	**Search Field**	**Median (50th percentile) correlation**	**10th percentile**	**90th percentile**
Employment (Non-Farm Payrolls)				
	Overall employment related searches	−0.30	−0.43	−0.19
	All Users	−0.18	−0.32	−0.02
	Architecture and Engineering	−0.37	−0.48	−0.26
	Art	−0.20	−0.36	−0.04
	Business	−0.22	−0.35	−0.08
	Construction	−0.15	−0.29	0.01
	Education	−0.21	−0.35	−0.05
	Finance	−0.29	−0.40	−0.17
	Food	−0.30	−0.42	−0.16
	Health	−0.24	−0.35	−0.13
	Leisure	−0.14	−0.29	0.02
	Manufacturing	−0.07	−0.21	0.06

Continued on next page

Table 1.6 – *Continued from previous page*

Official Data Category	Search Field	Median (50^{th} percentile) correlation	10^{th} percentile	90^{th} percentile
	Retail	-0.15	-0.32	0.02
	Science	-0.04	-0.24	0.16
	Technology	-0.25	-0.38	-0.12
	Transportation	-0.26	-0.41	-0.12
	Govt	-0.30	-0.45	-0.15
Home Sales (Pending Home Sales)				
	Existing Homes	0.24	0.06	0.40
	New Homes	0.15	0.01	0.29
	Real estate website activity	0.05	-0.09	0.20
	Home Financing	0.25	0.04	0.44
	Home Refinancing	0.38	0.19	0.55
	Home Bankruptcy	0.10	-0.08	0.27
	Realtors	-0.10	-0.26	0.10
	Home Inspection and Appraisal	0.17	0.02	0.30
	Home Foreclosure	0.20	0.04	0.34
	Home Closing	0.01	-0.18	0.18
Retail Sales (Retail sales ex-auto and Gas)				
	Beauty Services	0.22	0.05	0.37
	Fast Food	-0.11	-0.35	0.15
	Restaurants	0.21	0.04	0.37
	Noncyclical Services	0.18	0.01	0.35
	Cyclical Services	0.06	-0.08	0.21
	Durable Retail Goods	0.15	-0.01	0.29
	Nondurable Retail Goods	0.14	-0.03	0.31
	Discount Retail Goods	0.26	0.10	0.41
	Shopping Malls	0.21	0.00	0.38
	Utilities	0.33	0.20	0.44
	Mature Entertainment	0.09	-0.10	0.29
	Motor Vehicle	0.12	-0.03	0.26
Auto Sales (Annualized Total US Auto Sales monthly)				
	Auto Websites	-0.61	-0.68	-0.52
	Auto Makers	0.51	0.35	0.65

Continued on next page

Table 1.6 – *Continued from previous page*

Official Data Category	Search Field	Median (50th percentile) correlation	10th percentile	90th percentile
	Auto Reviews	−0.59	−0.70	−0.48
	Auto Price	0.23	0.04	0.40
	Auto Comparison	0.40	0.26	0.51
	Auto Insurance	0.63	0.54	0.74
	Auto Financing	0.91	0.89	0.93
	New Car Dealers	0.73	0.62	0.86
	Used Car Dealers	0.78	0.70	0.84
	Motorcycle Dealers	0.79	0.73	0.84

In the next section we use the curated data to make a model.

1.8 Modeling Case Study: Non-Farm Payrolls (NFP)

For this section we will imagine an analyst with reasonable proficiency in statistical learning as well as macroeconomics who is tasked with making a model from alternative search data in a live prediction environment. Our analyst faces many decisions and the frameworks are well revealed from giving specific examples of these decisions.

Our analyst begins by gathering what are all the possible alternative data sources she is being asked to use. In this case she first starts with what she believes gives her better ex-ante prediction and will then move to complement the noisier ex-ante prediction with the higher accuracy and recall but ex-post data such as credit cards or accounting. Figure 1.18 shows the various data sources available as twitter data where we pick specific tweets similar to Proserpio et al. (2016), and for search data we could use impressions which is the content that appears when we type in a search query, or we could also use actual clicks which show a higher intention by the user, also available are the so called "local searches"- which are searcher near the user's zip code, which can indicate a higher intention for the information sought- whether it is purchasing a local product or applying for unemployment claims at the local office. Internet explorer is the browser that Microsoft offers, and our analyst has access to the number of aggregated, anonymized total clicks on various websites and she can use it to obtain the total clicks on various unemployment claim websites.

Now our analyst must decide which particular government data release related to employment she would choose to model. There are many official data releases such as unemployment rate, non-farm payrolls (NFP), Initial Jobless Claims (IJC), wages, and there are other statistics [Baumol (2012) is a great soource] such as Help-Wanted Online Advertising, Corporate Layoff and Hiring Announcements, Mass Layoff Statistics (MLS), and even the ADP National Employment report.

Among these statistics as Baumol suggests IJC and Unemployment rate and NFP are important- the rest are helpful characterizations but do not move markets or covered by the media as much. The unemployment and NFP series come out in the same releases called the employment situation that is typically available at https://stats.bls.gov/news.release/pdf/empsit.pdf. Now between unemployment rate and NFP we pick NFP per the reasoning

Figure 1.18: The low precision, high recall and ex-ante data for employment available to the analyst.

in the previous section of both predicting and using data that are differenced. NFP are the jobs created each month rather than the overall level of unemployment that has a high AR component to it and will not change much as compared to the last month.

There are two main surveys conducted by the BLS: the household (around 60,000 households) and the establishment survey (around 440,000 corporate and government work sites). The NFP number comes from the establishment survey. Our analyst also checks as suggested in Jain (2018) that the NFP data release does indeed move the market. That market move is a clear indication of the information content in the data release since the financial markets would react to new information per the vast literature on event studies. Figure 1.19 shows the result of the event study analysis of NFP surprises vs. price returns for the dollar index.

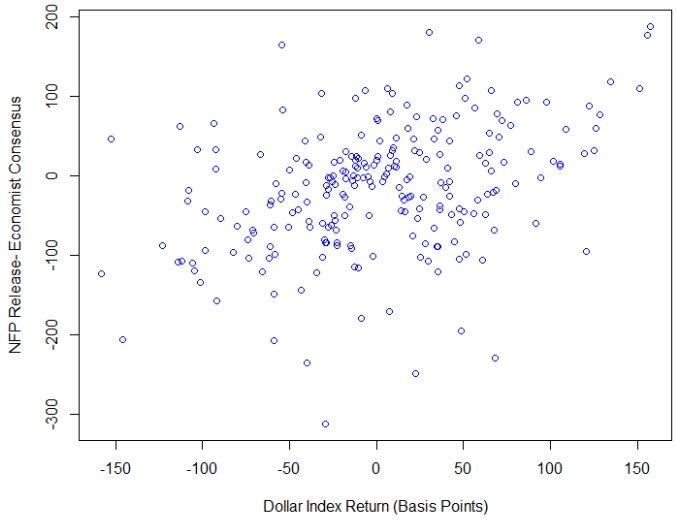


Figure 1.19: Movement in the USD vs. NFP surprises (Actual release – Wall Street consensus). Data: Bloomberg and time period from Jan 1998 until today. Source: Jain (2018).

1.8.1 Interpretable vs. Blackbox or top down vs. bottom up models via Kuhn

Now our analyst must decide which statistical learning procedures to use. Here she faces the trade-offs between exposition or the ease of explaining her model to the team vs. better prediction with more flexible and recent techniques from the machine learning (ML) literature that are harder to explain to non-experts or interpret with a narrative. Figure 1.20 borrowed from James, Witten, Hastie and Tibishirani (2013) shows the trade off that is explained quite well in the book.

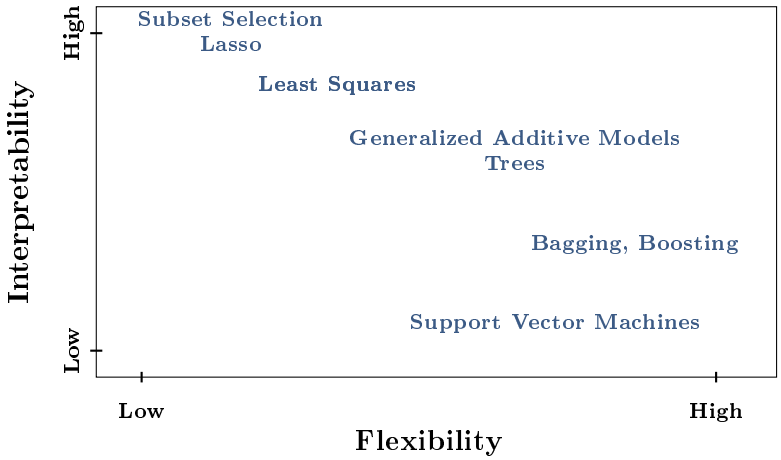


Figure 1.20: Source James, Witten, Hastie and Tibishirani (2013). The interpretability vs. flexibility trade-off for common statistical learning techniques.

There is a deeper philosophical question of why we make models. One answer is "top down" or to develop intuition about the dynamics of something incredibly complex to allow for better decisions. For example many economics models capture some essential part of the complex economic machine in a simple enough framework that allows us to ask scenario based questions: Given this simplification of the central bank into three variables- say employment, inflation and interest rate that are connected in this simplified way such as a Taylor rule, what would be the effect of increased employment on the nominal interest rates? Such models tend to have an underlying economic narrative or paradigm around which the economics profession can coordinate their research efforts, even if the predictive power of these models is low. From reading the previous section on the Lucas critique, we can see the "top down" models are more likely to be structural in nature.

Kuhn (1962) in his seminal work- The Structure of Scientific Revolutions addresses the progress of scientific knowledge within the paradigm and a particular paradigm may not be rejected until the "anomalies" relative to that paradigm become so difficult to handle that the profession moves towards a different paradigm that offers, better and simpler solutions. This dynamic of coordination and communication around a paradigm applies even to the particular group the analyst might work in! And hence the group may be more accepting of a model they can interpret more easily and everyone can agree on rather than the model which they cannot understand.

Another answer of why we make models is "bottom up" to predict the outcome of a complex process better than our intuition! Most humans cannot process differential equations or invert a 50 by 50 matrix in their minds and hence we have computer programs to help us with that. A parallel in the world of modeling is that we may believe NFP is affected by many variables such as weather, people's attitudes about jobs, many kinds of searches, we may also posit that the real relationship between searches and NFP is non-linear and is better modeled empirically using a flexible ML method that can completely fit the data and extract every ounce of predictability from the data rather than have a simple mathematical model of dynamics. Typically, the "bottom up" models tend to be more non-structural but no model is purely devoid of economic or social psychology theory or completely identified with it. So the "top down" or "bottom up" is a useful alternative characterization.

1.8.2 The practical reason: Modeling noise in small datasets

Away from the theoretical concerns above, the practical reason for simpler models in our case is that macroeconomic data are noisy as shown above and just like the parable of Google flu we have almost an infinite number of potential search variables that can fit these noisy data and if we allow the usage of extremely complex non-linear techniques then we will end up fitting noise and may be unable to detect the true data generating process via our model. Thus, the other reason most live and practical prediction scenarios do not prefer Blackbox techniques is because of their concern that the curse of high dimensionality as James et al. (2013) refer to it along with the noisy data and a limited time period (there are no recessions in our data set from 2012 to 2018) would result in worse predictions.

1.8.3 The five keys: Clean data, internal consistency, shrinkage, bootstrapping and ensembling

In this section before we continue with the investigations into modeling NFP with our analyst, we frame the five key principles that we find useful across the many kinds of models we make. Cleaning (and visualizing) the data is critical, and we address the failures associated with not doing it properly in the next section. Visualization of the data is a part of cleaning the data and is also helpful when developing "stories" or hypothesis to be tested. Internal consistency refers to a unified and logical approach from data to model to testing. For example, if we decide to proceed with a structural or a top down model to model growth shocks, then it should apply across all sectors. If we find that increasing searches signal higher propensity to buy cars because we believe people form intention for future actions by gathering information via searches, then the same idea with the same sign should apply to the housing market. A model with a positive coefficient for search on cars but a negative one on housing would find it difficult to pass the internal consistency test.

Diebold (1998) defines Shrinkage as the idea of coaxing or "shrinking" parameter values in certain directions", for instance in the direction of a prior we believe in like 0. Shrinkage (by and large) reduces the MSE by lowering the variance considerably at the potential cost of a small increase in bias (say if the coefficient shrank to 0 is not exactly 0). Shrinkage has long been known to be practically useful as Diebold (1998) mentions the case of vector autoregressions using Bayesian shrinkage producing "drastically superior forecasts over the unrestricted vector autoregressions." In the high dimensional problems that we will tackle,

we find that reducing the number of parameters often turns out to be quite useful and the resulting model with fewer parameters also has the benefit of being more interpretable. Polson (2017) provides an interpretation of Ridge regression as a Bayesian hierarchical model with a normal likelihood and prior.

Alternative data sources tend to have "spiky" distributions, as we can see in Figure 1.16 that plots the searches for new homes. More technically, we are not fully aware of the statistical properties of these data which also have small sample sizes in at least one dimension and hence asymptotic inference may be unreliable in addition to being analytically complicated. Kogan (2010) suggests the use of simulation methods such as bootstrap to deal with analytically challenging problems and to adjust for bias and improve the precision of asymptotic approximations in small samples such as confidence intervals and test rejection regions etc. The bootstrap confidence interval is naturally accurate asymptotically since we are just repeatedly sampling the given distribution and it has advantages such as the one pointed out by Kogan (2010) that for a "t-statistic" the bootstrapped distribution is more accurate than the large-sample normal distribution. Additionally, with these noisy data, the possibility of "influential points" affecting estimates is quite high, especially when considering non-linear models. Thus, repeated sampling or bootstrapping is a key ingredient for inference and understanding how the statistical procedures we are using interacts with the data we have.

Ensembling is simply combining outputs of different models to produce one prediction and the fundamental idea is of variance reduction by combining imperfectly correlated outputs. Better results from ensembling are obtained when combining various internally consistent models built on different philosophies. For example, we believe that ensembling an internally consistent "linear" top down Lasso model and an internally consistent "bottom up", "non-linear" random forest or deep learning model is fundamentally more robust than simply averaging two similar models. Maclin and Optiz (1999) express this idea eloquently- "research has demonstrated that a good ensemble is one where the individual classifiers in the ensemble are both accurate and make their errors on different parts of the input space."

1.8.4 The Model Overconfidence Metric (MOM)

Given the discussion above, our analyst decides to explore most of the techniques in the James et al. (2013) book for modeling NFP and will use the data from 2012 to 2016 as the modeling dataset and the rest of it as a holdout sample to test the efficacy of the techniques. Table 1.7 provides an example of the type of notes the analyst should make while thoughtfully utilizing each technique.

After the analyst utilizes various techniques, she decides to take the ratio of the in-sample error (In Sample MSE) and the out of sample or the mean squared error when taking the model generated to fit the holdout sample to generate the "model overconfidence metric (MOM)." Formally, she defines the MOM metric as

$$\text{Model Overconfidence Metric} = \text{Out of Sample MSE}/\text{In Sample MSE} \tag{1.6}$$

For example, a model that has a low in sample MSE- say due to overfitting of the data and has a high out of sample MSE would be deemed quite overconfident due to a high value of the MOM metric. Similarly, a model that has low out of sample error vs. the in-sample error would score low and hence would be a better performing model. Below, we reproduce the metric for this case study.

Table 1.7: Example of notes for each technique for modeling NFP

Modeling Methodology	Particular Technique	Variables Chosen	Decision criterion	In Sample MSE	Holdout Sample MSE	Over-confidence Measure*	Comments
Linear Regression	Linear Regression	All		3405	9110	2.7	With average correlation of 0.56 among variables and only 59 data points vs. 18 possible variables using all variables will compound the overfitting, lack of interpretation and lower the power of the test. We find that the overall F statistic of the regression is not significant.
Subset Selction	Bootstrapped correlation variable selection	Overall, Archeng, Finance, Technology and Govt.	Bootstrapped correlations and wide coverage	4807	8265	1.7	We examine the bootstrapped confidence intervals and aim for a a wide variety of fields in addition to just the overall job related searches
	Linear Regression with cross validation	Archeng, Overall	MSE and 5 fold cross validation	4630	7275	1.6	We chose between models suggested by manual variable selection above.
	Best subset selection	Archeng, Science, Overall	Min Cp	4117	7723	1.9	
	Forward and Backward selection	Archeng, Science, Overall		4117	7723	1.9	
Elastic Net	Lasso	Overall, Food, Government, Archeng		4559	6543	1.4	
	Ridge	All		4674	6394	1.4	Shrinking parameters seems to help!
	Elastic Net	All		5101	6991	1.4	

Table 1.8: MOM metric for the particular model of NFP using search variables. All techniques used 5-fold validation, the modeling sample was 2012 to 2016, holdout sample was Jan 2017 onwards

Modeling Methodology	**Model Overconfidence Metric (MOM) =MSE Holdout MSE/MSE Model.** **(smaller is better)**	**Holdout MSE**
Random Forest	1.3	6801
Ridge	1.4	6394
Lasso	1.4	6543
Linear Regression with cross validation	1.6	7275
Linear regression with variable selection	1.7	8265
Best subset selection	1.9	7723
Pruned boosting	1.9	6877
Principal Components Regression	2.0	8126
Pruned Decision Tree	2.0	6919
SVM Regression	2.2	7994
Partial Least Squares Regression	2.5	8945
Linear Regression	2.7	9110
Decision Tree	4.4	8888
Naive Boosting	5.6	13 650

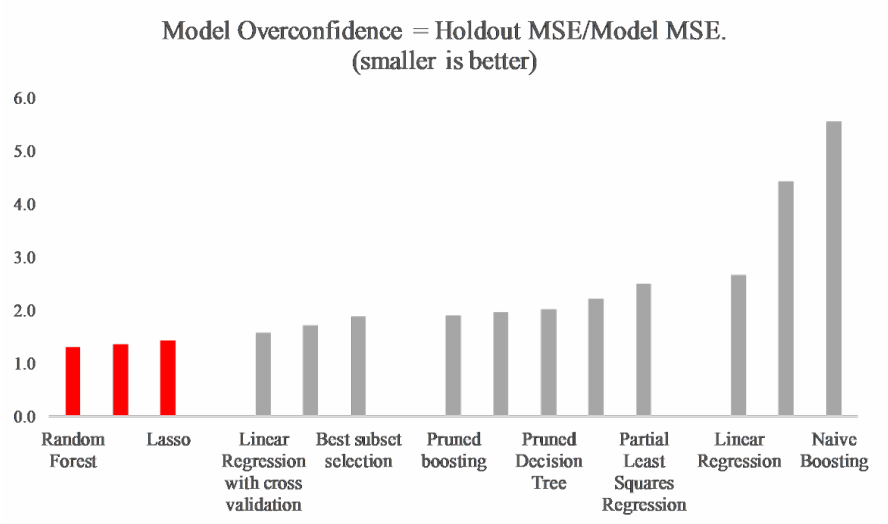


Figure 1.21: Graphical representation of the error in holdout vs. in sample via MOM metric.

1.8.5 Discussion of case study results

Challenges for Linear regression: We find that in this case study simple linear regression which tends to be one of the favorite technique of the economics profession does not perform well, neither do naïve decision trees, or even boosted trees- which typically guard against overfitting by iterating slowly. The boosted trees did perform better after many iterations to find suitable step sizes. We believe there are two chief culprits- first, the high dimensionality of our variables, and second, the noisiness of the macro data which can result in local overfitting. The majority of these results are actually keeping in line with known literature. As far as linear regression is concerned, Polson (2017) actually points out that even in simulated data we do better than simple linear regression by shrinking the coefficients to reduce the variance. The typical MLE (maximum likelihood estimator) or the OLS (Ordinary Least Squares) is designed to have zero bias, which means that it can suffer from high variance. According to him the main advantage of linear regression is interpretability!

The usefulness of cross validation and shrinkage: We find that the typical aggregation techniques like principal components perform a bit better but still the out of sample MSE is almost 2 times the in-sample MSE even after doing 5-fold validation. Combining some economic intuition by selecting some variables was helpful, as was cross validation and best subset selection for linear regression. The improvement of linear regression with the addition of cross validation showcases the importance of bootstrapping as a philosophy, especially when dealing with social data that can be non-normal and also in providing a sense of how dependent models may be on "influential data points" in small samples. The success of best subset selection suggests the other theme of "shrinkage" which reduces the number of parameters to estimate resulting in more robust models. The theme of shrinkage being helpful can also be seen in the fact that ridge and lasso techniques that actively use a penalty term as the number of terms in a model grows, show the best results along with techniques based on random forest.

Random Forest- bootstrapping non-linearity with no theory: We see that random forest techniques perform quite well, and since a random forest is essentially a decision tree-based technique that uses bootstrapping with de-correlated iterations (where new variables have to be selected each iteration). This indicates that there are non-linearities in the data, especially since the random forest-based techniques selected some different variables than the linear methods, but the non-linearities can be modeled robustly only when we use bootstrapping and go through all possible variables. The good performance of a purely empirical non-linear technique such as random forest which does not start ex-ante with some economic theory is also interesting and similar to the with recent encouraging results in the field of deep learning- a very non-linear technique that seems to have useful applications for noisy and complex data where no particular theory is obvious (Heaton, Polson and Witte, 2017). In our case study of modeling NFP, we find that random forest performs better than the linear regression-based techniques and also selects different variables as compared to the linear regression[17].

Ensembling helps: As suggested in the previous section, we find that combining the two approaches of ensembling random forest and lasso-based methods we get a good mix of parsimonious linear and non-linear methods and the resulting out of sample MSE is much lower than with either technique.

Our informal results for which techniques performed best for the particular case of modeling

[17] A comment on tree based methods from the case study is that sorting based on entropy or node purity measures tends to result in substantially different models for noisy and high dimensionality problems.

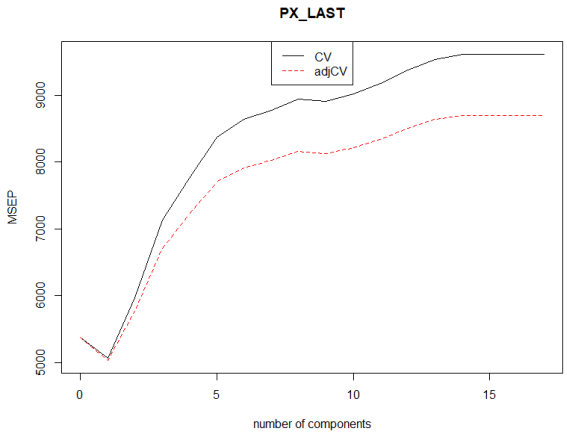


Figure 1.22: Example of selecting the right number of principal components by 5-fold cross validation while modeling NFP.

NFP from 2012 to 2018 using specific curated search variables has been examined much more formally in the statistical learning literature such as Caruana and Niculescu-Mizil (2006) where they perform a formal large-scale empirical comparison between supervised learning methods SVMs, neural nets, logistic regression, naïve Bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps using a variety of performance criterion.

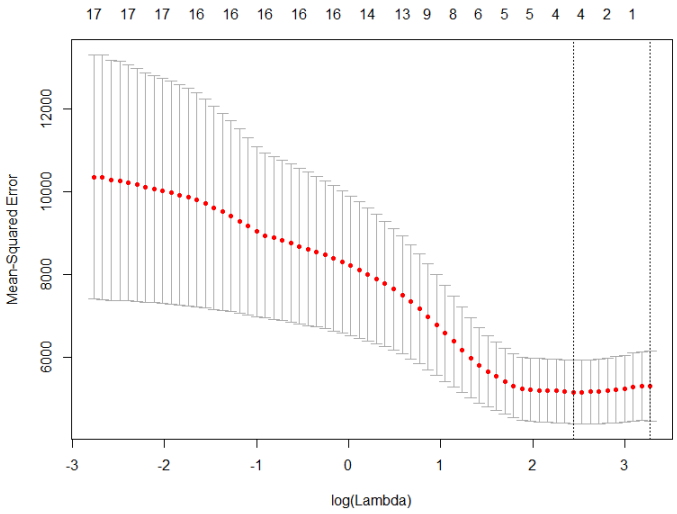


Figure 1.23: Example of picking the optimal value for lambda for the Lasso Technique for modeling NFP

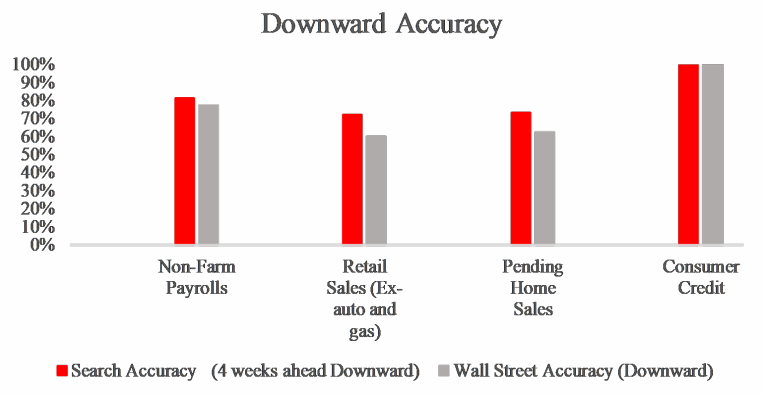
Like our case study, they find that random forests, and bagged trees perform quite well across a variety of problems and do not require as much calibration. Other techniques perform much better for specific problems and under specific types of calibrations.

1.9 Live production results

Having discussed the various possible models for NFP, for the purposes of live production, we find the best results with a forward moving ensemble of LASSO and random forest models, and we find that the model improves with each data point and is able to adapt to the dynamic social data. In addition to NFP, we extend a similar modeling approach to other sectors using some of the features shown in Table 1.6. We note the actual performance of indicators such as employment, retail, home, and consumer credit up to 4 weeks before the data release in Table 1.9. We find that in terms of MSE they are about the same or worse than wall street consensus. However, in terms of directional accuracy these numbers tend to be better than wall street consensus, especially on the downturns but the sample are too small for a high degree of statistical confidence in that result.

Table 1.9: Performance of employment, housing and retail indicators in data upticks and downticks across 1 to 4 weeks before the data release. Data source MSFT and Bloomberg.

Date Release	**Accuracy** (4 weeks ahead Overall)	**Street** Accuracy (Overall)	**Accuracy** (4 weeks ahead Upward)	**Wall Street** Accuracy (Upward)	**Accuracy** (4 weeks ahead Downward)	**Wall Street** Accuracy (Downward)
Non-Farm Payrolls	79%	77%	77%	77%	81%	77%
Retail Sales (Ex-auto and gas)	73%	69%	74%	78%	72%	60%
Pending Home Sales	76%	73%	80%	84%	73%	62%
Consumer Credit	100%	100%	100%	100%	100%	100%



Caption?

What is remarkable, however is the ex-ante or predictive nature of the search data-based measure since they are already at their optimal performance 4 weeks before. More data over time does not seem to help them. - including weeks 3 through 1 does not seem to improve the directional accuracy in our sample. Hence our characterization of search data is noisy (low precision), but wide recall and ex-ante.

1.9.1 Prediction in practice: The main mistakes[18]

Modeling and predictions are conducted in a team environment and the overall process can be sketched out in the loop shown in Figure 1.24 where generating hypothesis, data exploration and cleaning are part of the initial development and testing of the hypothesis partly happens in the holdout sample, but partly in the live production environment. Typically, because of all the data mining concerns, in such a team environment the performance impact of actual live 12 data points and prediction is much higher than the same number of 12 data points of a holdout sample. Strong process controls and awareness of potential mistakes can reduce this performance gap between the holdout sample and the live performance of the model and because of the noise in the series, a high overall process quality is the best insurance a team has of being able to stay committed to the model since judging performance in real time is difficult.

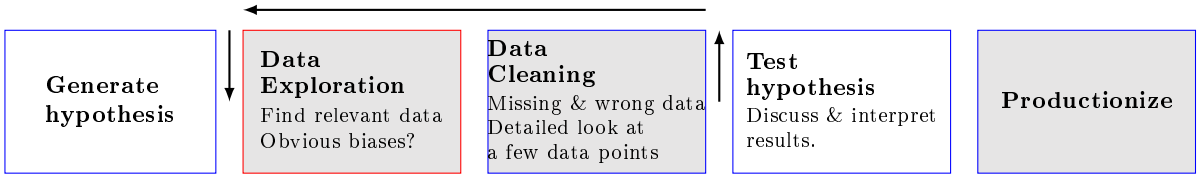


Figure 1.24: The typical model development loop followed.

There are four main categories of mistakes in the model development loop described above:

1. Data processing mistakes
2. Statistical learning mistakes
3. Conceptual mistakes
4. Organizational mistakes

Data processing mistakes such as not cleaning the data properly (say not removing GDP figures like 400% growth in some quarter, or not investigating if 6700 searches in one evening for pizza in a village of population 400 are correct) are fairly common due to the unstructured nature of the data and the domain knowledge required to spot them. Thus, data cleaning is something the entire team should be aware of and help. As Hadley Wickham (2014) states data cleaning is not studied enough when in fact 80% of data analysis is data cleaning. No data is unconditionally clean, each dataset tends to be cleaned relative to the problem being modeled and any decisions on cleaning data are decisions on how to model the data. I believe it is critical to have the person modeling the data be very involved with the cleaning of it hence critiques of the academic economics profession where data cleaning, modeling, and theorizing sometimes seem too detached for practical purposes resonate (Orphanides, 2001; Romer, 2016).

Statistical learning mistakes refer to either technical mistakes like using linear regression to model unit root processes or using ML techniques that one does not conceptually understand and having a high degree of overconfidence in the model.

[18] These are mostly mistakes I have either made myself directly or seen firsthand!😀

Conceptual mistakes are more along the lines of inherent contradiction in our goals and processes – for instance modeling business cycle dynamics and using those to arrive at asset allocation decisions and removing the financial crises such as 2007-2008 from the data. Occasionally such processes are justified by the desire to not have outliers influence statistical inference but there is a philosophical issue that for the purposes of an asset allocation model or making central bank policy these crises are the times with substantial impacts on outcomes and any description of reality that does not contain these is inherent flawed.

Organizational mistakes render teams ineffective for two principal reasons- the first a lack focus- no tangible goal is often a telling symptom of this, and the second because they get mired in organizational frictions. Different skillsets are necessary to perform the various functions in the model development loop above, especially as alternative data source, high computing power and ML techniques have become popular in the finance and economics domains. These different skills come together as a part of a team. In Table 1.10 we list the typical archetypes that might fulfill different roles in the process. Developing common ground and constant communication for these various types of expertise is critical for functioning better as a team.

We represent the ideal process in Figure 1.25 as the intersection of domain expertise, the research mindset where we can keep iterating and one failed experiment does not doom the enterprise, along with newer datasets and artificial intelligence (AI) based techniques such as ML that is supported by a robust technological architecture. The closer the team can get to the center of the diagram, the better the research and product output.

Table 1.10: Archetypes of the desired skills. For illustration purposes, no offence intended!

Archetype	**Expertise**	**Typical Strength**	**Typical Challenge**
Domain Expert	Institutional and markets knowledge.	Good Intuition	May not recognise own cognitive biases.
		Respects how difficult forecasting better than the wisdom of the crowds is.	May be suspicious of ML methods
Technology Expert	Data and Machine Learning (ML) Techniques	Skilled at data cleaning. Efficient and standardized data storage.	Likely to overfit the data.
		Can glean insights from data using visualization and modeling	Might make suboptimal model due to low context about the problem.
Researcher	Systematic Thinking in conseptual framework	Desighns an efficient research agenda.	May not be open to new findings that go against favorite theoretical framework.
		Mitigates overfitting	Prefers to enage in long term projects that take time to pay off.

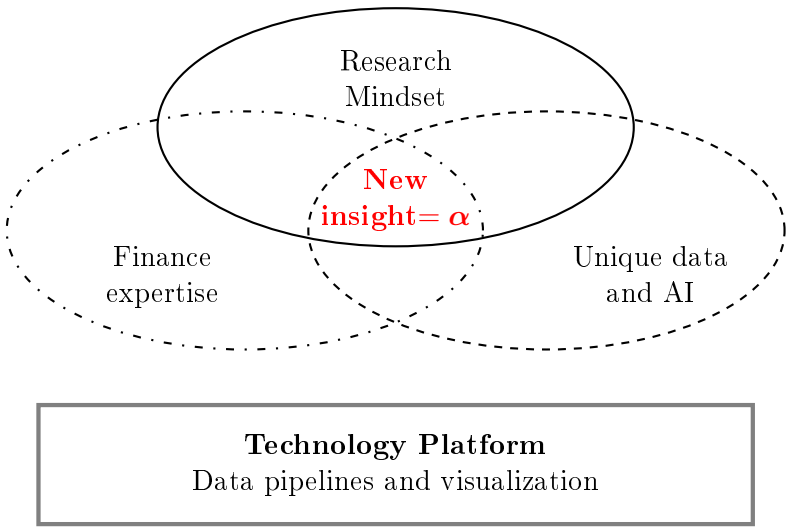


Figure 1.25: The ideal process combines a clear goal, lean organizational structure and as much overlap between different skills as possible. The closer to the center the team process is- the better the results.

1.9.2 Public benefits of microfoundations of macro

Real people plan their lives based on their understanding of the economic opportunities around them which are communicated (at least in part) through economic data. Institutional actions such as central bank policy, fiscal aid, and even bailouts and other aid from -say- the World Bank also depend on data to know the location and the extent of any socio-economic problem. Thus, economic measurement has a big role to play in the social and democratic sphere since it influences the actions and relationships of individuals, institutions, and communities. All actors could use better data and economic noise is not evenly distributed:

The first asymmetry of measurement: Economic data are noisier in times of recessions and low growth.

The second asymmetry of measurement: Poor countries have much worse economic data.

The first asymmetry of measurement was shown in Table 1.2, where we showed that both the volatility as well as the revisions are higher in times of low growth in the United States- one of the most powerful and sophisticated countries in the world. So even as citizens are losing employment in certain states, the central bank or the government cannot act for 3 to 6 months due to lack of data. Poor data have also created longer term problems-for instance the Federal Reserve policymakers' belief that the economy was operating much lower than its potential in the 1960s and 1970s might have contributed to the overheating and high inflation of the economy later (Orphanides, 2002). Another area for improvement is state level GDP which is only released on a yearly basis! State level GDP seems to be a critical input to help the government decide on a response to any economic issue! Now instead of the GDP, we can use other indicators such as NFP, unemployment rate, average house worked in manufacturing, and real wages that are released monthly.

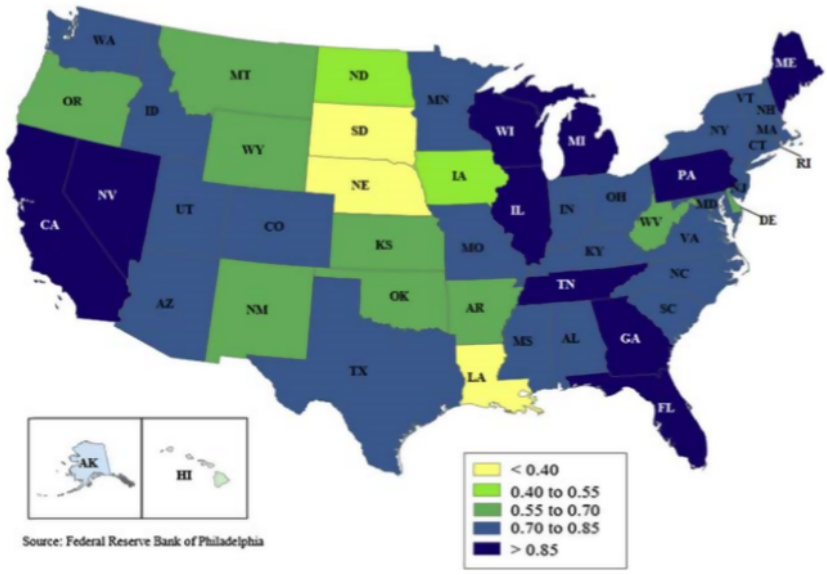


Figure 1.26: Novak (2013). Correlations between state GDP and state coincident indices are not uniform. This might make it harder to spot economic problems particular to the yellow and light green colored states.

Novak (2013) uses these monthly indicators to compute a proxy for the state-level economic activity and finds that while they are reasonably effective on an overall basis with a correlation of 0.69, for states with more agriculture and mining, the correlation can be quite low- below 0.40 and hence would make it difficult to gauge the economic conditions especially in a situation where the national economy is performing well and these sectors are suffering. On average the same states- more reliant on agriculture and mining are the states that tend to be economically challenged, thus the lack of good data would impact them disproportionately more- the second asymmetry of economic measurement.

Given the amount of effort needed for such a complex calculation as GDP, it is not surprising that less developed countries would face challenges. However, the extent and the fundamental nature of the measurement problem in developing countries is surprising. Diane Coyle (2014) presents the case of Ghana whose GDP increased by 60% in November 2010! This happened as Ghana updated the weights used to compute the price index for calculating real GDP for the first time since 1993. The new GDP calculation in turn changed the official classification into a "low-middle-income" country. This is not a one-off issue. Alwyn Young (2012) notes how sub-Saharan living standards have been growing at about 3.5X the rate indicated in international datasets for the last two decades! Young says that for the poorest regions of the world such as 45 sub-Saharan African countries, 24 of those countries did not have any benchmark study of prices in the popular Penn World Tables. Coyle (2014) agrees- "the number of countries for which GDP data were available had reached only sixty as late as 1985.... Thirty years' worth of annual data for sixty countries is not much when it comes to testing detailed causal explanations of growth..." But these data or lack thereof are used by economists and "experts" to make international comparisons, tell stories and most alarmingly, decide policy!

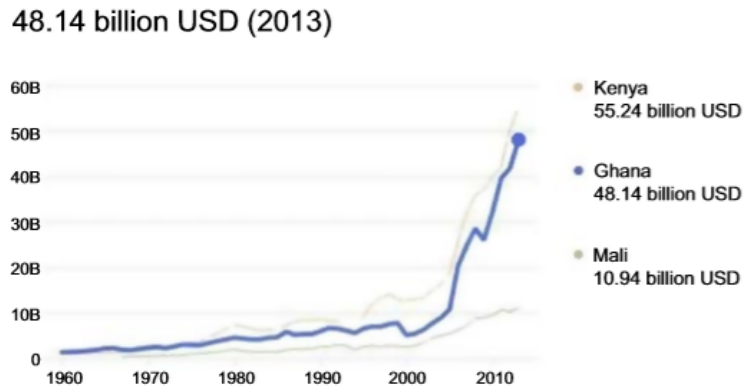


Figure 1.27: The dramatic change in Ghana GDP upon updating weights from Coyle (2016). Data from World Bank.

1.9.3 Two main contributions: Accurate measurement and more detail

Measuring and visualizing real activity directly: Better measurement is not a new idea. In fact, even official GDP measurement apparently came about when America's congress tasked Simon Kuznets in 1932 to estimate the national income over the past few years. Only after seeing the actual data he produced, did the full extent of the depression become clear.[19]

Young (2012) offers a potential solution to the Sub-Saharan country GDP measurement problem in the previous section by estimating their GDP from demographic and health surveys that collect irregular but in-depth data. He uses four areas- 1) ownership of durables, 2) housing conditions, 3) children's nutrition and health, and 4) household time and family economics and relates them to GDP. By using alternative data creatively- perhaps in the spirit of Young we may be able to help direct aid efforts better and complement and sanity check the government statistics. These data would measure and show real activity directly which is frequently measured and does not rely upon price levels or government data collection that could be politically distorted and would provide detailed information in real time[20]. By avoiding the problems associated with measuring price levels, it could provide a complement to the computed GDP numbers. Imagine mapping all mobile phone, search, twitter, web browser, and mobile location data relevant to employment at an aggregated, anonymized level. This map of people's behavior could help policy makers understand the effects of a financial crisis in remote areas in real time and divert aid as needed. This tool could be tremendously useful for both policy makers as well as citizens. Such direct mapping of activity could help mitigate concerns about GDP no longer being valid for "digital world" of Wikipedia and Linux (Coyle, 2014).

Visualization is a key part of this solution since most of us understand pictures much better than words or mathematics. Thus, all citizens could participate in a more informed manner in public debates in the era of "fake news" and such politicization of data where the former chief

[19] https://www.economist.com/briefing/2016/04/30/the-trouble-with-gdp

[20] There is already some excellent work in the field- for example using mobile data to understand poverty in developing countries such as Blumenstock (2016). Billion Prices Project at MIT (http://www.thebillionpricesproject.com) is another great example of the use of such data that offered an alternative to CPI for the public when the Argentinian government was manipulated by the government.

of Greece's statistical agency is being criminally prosecuted "for getting the numbers right."[21]

More detail and better democracy: The second contribution is in providing a platform where every citizen can choose to have an economic voice. Self-reporting of issues that citizens face and collecting granular data will prove more helpful for policy makers rather than simple aggregates. The then chair of the Federal Reserve Ben Bernanke (2012) acknowledged concerns with purely aggregate statistics while dealing with the aftermath of the global financial crisis[22] – "aggregate statistics can sometimes mask important information. For example, even though some key aggregate metrics--including consumer spending, disposable income, household net worth, and debt service payments--have moved in the direction of recovery, it is clear that many individuals and households continue to struggle with difficult economic and financial conditions." It costs very little to tweet something about our employment situation. In Proserpio (2016) et al. these individual level tweets about employment or the lack thereof are "heard," and by providing the kind of authentic, direct and individual feedback to the policy maker could make the entire process more democratic.

1.9.4 Mitigating Data Colonialism?

The current technology platforms where us ordinary citizens lead our normal lives leave a trail of "data exhaust" which helps companies extract value by creating opportunities to know us - the consumers -holistically in a way that did not exist before. If our data a "natural resource" is taken from us without our informed consent and used to the companies' benefit and our detriment at a large scale,[23] and this starts seeming "normal" or a "way of life" it is dangerous and undemocratic. In fact, I will go out on a limb and say that historically the combination of a pure profit motive and vast asymmetry of power (say via technology) seems not to have led to "moral" choices that benefit entire humanity[24]. In a useful and timely article Couldry and Mejias (2018) point out that "unlike oil, data are not a substance found in nature. It must be appropriated." (emphasis ours). In that article the authors explore parallels of today's economic machine with historic colonialism to show how it normalizes resource appropriation and redefines social relation so that dispossession seems natural (emphasis ours). They define data colonialism as follows- "Data colonialism combines the predatory extractive practices of historical colonialism with the abstract quantification methods of computing."

So how can we make appropriation, dispossession and asymmetric exploitation of our data not the normal state of affairs? The topic is too big and complex for any one person to have an effective solution, but too important to not suggest something in the hopes of being at least a small catalyst for discussion, I would like to suggest two keys:

The first key is to provide better information, and protection or rights to the consumer regarding their data. This seems to be increasing with changes in data privacy regulation such as General Data Protection Regulation (GDPR)[25], which is the "most important change in data privacy regulation in 20 years."

The second key lies in doing something positive and communal that brings us all together instead of a purely exploitative practice. We can do this by increasing universal or population

[21] https://www.ft.com/content/31995e48-6073-11e6-b38c-7b39cbb1138a.

[22] https://www.federalreserve.gov/newsevents/speech/bernanke20120806a.htm

[23] for example, to sell us more products made by the same company at higher prices due to an information monopoly via network effect

[24] Reading the history of the East India Company in India and China may be interesting.

[25] https://eugdpr.org/

wide benefits of and access to this technology and information. As a self-serving example- consider creating the detailed map of economic activity suggested above, along making the aggregated underlying data and code freely available in a transparent manner. It will help citizens and central banks in poor countries especially in times of recession. Such public good projects would create goodwill. Platforms like Google, MSFT Bing, Facebook, Amazon, Twitter, Instagram, and Uber have information that is definitely valuable to the owners but combined may be even more valuable to society. As an initial step, by providing aggregated, anonymized and lagged data these firms can keep a reasonable amount (per regulators) of competitive advantage in terms of understanding the micro-level consumer behavior intact, and simultaneously contribute enormously to broader society- central banks as well as ordinary citizens.

Acknowledgements

I believe there are no sole authors and this chapter is no exception. I want to thank amazing colleagues at MSFT- the first being Scott Counts- my co-conspirator on many projects and a great source of humor, enthusiasm, and guidance, Jens Nordvig at ExanteData for critical revisions and encouragement, Geraint Jones for thoughtful comments, Nikita Artizov for defending R programming, and suggesting I put more emphasis on the social good section, Martin Ryan for being a key collaborator over the years, Chris Quirk for his NLP guidance and humor, the researchers at the NY Fed, DC Fed, St. Louis Fed -especially Kevin Kliesen for his encouragement, the Bank of Canada research team- especially Bob Fay at CIGI for his comments, other authors of the book attending the conference, the BingPredicts team for their diligence with the data, Ben Mandel and Andrew Haughwout for excellent comments and editing, my family, and my wife Sara for her patience, many readings and constructive criticism. All errors are my own.

Bibliography

1. Kliesen, K. (2014). A guide to tracking the us economy. *Federal Reserve Bank of St. Louis Review, First Quarter 2014*, *96*, 35–54. Retrieved from https://insurancenewsnet.com/oarticle/A-Guide-to-Tracking-the-US-Economy-a-481934

2. Aruoba, S. B., Diebold, F. & Scotti, C. (2008). *Real-time measurement of business conditions*. National Bureau of Economic Research. doi:10.3386/w14349

3. Baumol, B. (2012). *The secrets of economic indicators: Hidden clues to future economic trends and investment opportunities.* (3rd). Pearson Education, Inc.

4. Orphanides, A. (2001). Monetary policy rules based on real-time data. *American Economic Review*, *91* (4), 964–985. doi:10.1257/aer.91.4.964

5. Orphanides, A. & Williams, J. (2006). Monetary policy with imperfect knowledge. *Journal of the European Economic Association*, *4*, 366–375. doi:10.1162/jeea.2006.4.2-3.366

6. Cochrane, J. H. (2001). *Asset pricing*. Princeton University Press. Retrieved from https://faculty.chicagobooth.edu/john.cochrane/research/papers/samplechapters.pdf

7. Jain, A. (2018). *Does search data know something wall street and government data don't?* Under review.

8. Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424. doi:10.2307/1912791

9. Goel, S., Hofman, J., Lahaie, S., Pennock, D. & Watts, D. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 17486–90. doi:10.1073/pnas.1005962107

10. Proserpio, D., Counts, S. & Jain, A. (2016). The psychology of job loss: Using social media data to characterize and predict unemployment. In *Proceedings of the 8th acm conference on web science* (pp. 223–232). WebSci '16. Hannover, Germany: ACM. doi:10.1145/2908131.2913008

11. Romer, P. (2016). The Trouble with Macroeconomics. *Commons Memorial Lecture of the Omicron Delta Epsilon Society*. Retrieved from https://paulromer.net/wp-content/uploads/2016/09/WP-Trouble.pdf

12. Reiss, P. C. & Wolak, F. A. (2007). Structural Econometric Modeling: Rationales and Examples from Industrial Organization. Elsevier. doi:10.1016/S1573-4412(07)06064-3

13. Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71* (356), 791–799. doi:10.1080/01621459.1976.10480949

14. Popper, C. (1961). *The Poverty of Historicism* (2nd). London: Routledge.

15. Prescott, E. C. (1986). Theory ahead of business-cycle measurement. *Carnegie-Rochester Confer. Series on Public Policy*, *25*(100), 11–44. doi:10.1016/0167-2231(86)90035-7

16. Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Confer. Series on Public Policy*, *1*(100), 19–46. doi:10.1016/S0167-2231(76)80003-6

17. Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Retrieved from https://go.galegroup.com/ps/i.do?id=GALE%7B%5C%%7D7CA60864212%7B%5C&%7Dsid=googleScholar%7B%5C&%7Dv=2.1%7B%5C&%7Dit=r%7B%5C&%7Dlinkaccess=abs%7B%5C&%7Dissn=10475141%7B%5C&%7Dp=AONE%7B%5C&%7Dsw=w

18. Barro, R. J. (1993). *Macroeconomics* (4th). The MIT Press.

19. Soros, G. (2009). *The Crash of 2008 and What it Means*. New York: Perseus Books Group. Retrieved from www.perseusbooks.com

20. Borges, J. L. (1958). On Exactitude of Science. In *Viajes de varones prudentes. libro iv, cap. xlv, l_erida*.

21. Couldry, N. & Mejias, U. A. (2018). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television and New Media*, *20*(4), 336–349. doi:10.1177/1527476418796632

22. Chancellor, S. & Counts, S. (2018). Measuring employment demand using internet search data. In *2018 acm conference on human factors in computing systems (chi)*. ACM. Retrieved from https://www.microsoft.com/en-us/research/publication/measuring-employment-demand-using-internet-search-data/

23. Choi, H. & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*(SUPPL.1), 2–9. doi:10.1111/j.1475-4932.2012.00809.x

24. James, G., Witten, D., Hastie, T. & Tibishirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. doi:10.1007/978-1-4614-7138-7

25. Polson, N. (2017). Teaching Notes for Probability 41901. Retrieved from http://faculty.chicagobooth.edu/nicholas.polson/teaching/41900/beamer41901-2.pdf

26. Kogan, L. (2010). Teaching Notes for 15.450. Retrieved from https://ocw.mit.edu/courses/sloan-school-of-management/15-450-analytics-of-finance-fall-2010/lecture-notes/MIT15%7B%5C_%7D450F10%7B%5C_%7Dlec09.pdf

27. Maclin, R. & Optiz, D. (1999). Popular Ensemble Methods : An Empirical Study. *Journal of Artificial Intelligence Research*, *11*(July), 169–198. doi:10.1613/jair.614

28. Heaton, J. B., Polson, N. G. & Witte, J. H. (2017). Deep learning for finance: deep portfolios. John Wiley and Sons Ltd. doi:10.1002/asmb.2209

29. Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Acm international conference proceeding series* (Vol. 148, pp. 161–168). doi:10.1145/1143844.1143865

30. Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. doi:10.18637/jss.v059.i10

31. Orphanides, A. (2002). Monetary-policy Rules and the Great Inflation. *American Economic Review*, *92*(2), 115–120. doi:10.1257/000282802320189104

32. Novak, J. (2013). The effectiveness of the state coincident indexes. *Special Report. Philadelphia Federal Reserve*. Retrieved from https://www.philadelphiafed.org/-/media/research-and-data/publications/research-rap/2012/the-effectiveness-of-the-state-coincident-indexes.pdf

33. Coyle, D. (2014). *GDP: A Brief but Affectionate History*. Princeton University Press. Retrieved from https://www.google.com/books?hl=el%7B%5C&%7Dlr=%7B%5C&%7Did=t7pKCAAAQBAJ%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PP1%7B%5C&%7Ddq=Diane+Colyle.+2014.+GDP:+A+Brief+but+Affectionate+History.+Princeton+University+Press%7B%5C&%7Dots=aAK-Kvab6g%7B%5C&%7Dsig=DxpfOmdEO3et2vv83D4wXRZpOV4

34. Young, A. (2012). The African growth miracle. *Journal of Political Economy*, *120*(4), 696–739. doi:10.1086/668501

35. Coyle, D. (2016). The trouble with GDP and emerging markets. Retrieved from https://www.weforum.org/agenda/2016/04/the-trouble-with-gdp-and-emerging-markets/

36. Blumenstock, J. E. (2016). Fighting poverty with data Machine learning algorithms measure and target poverty. *Science*, *353*(6301), 753–754. doi:10.1126/science.aah5217

Further Reading

37. Newey, K. W. & West, D. K. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. doi:10.2307/1913610

38. Kolanovic, M. & Krishnamachari, R. T. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. *J.P. Morgan Global Quantitative & Derivatives Strategy Report*.

39. Nallareddy, S. & Ogneva, M. (2017). Predicting restatements in macroeconomic indicators using accounting information. In *Accounting review* (Vol. 92, *2*, pp. 151–182). American Accounting Association. doi:10.2308/accr-51528

40. Box, G. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in statistics* (pp. 201–236). Academic Press. doi:10.1016/b978-0-12-438150-6.50018-2