Subchapter 1.1

Alternative Data and ML for Macro Nowcasting

Apurv Jain[1]

1 – MacroXStudio, San Francisco, CA. Email: apurvj@macroxstudio.com. Parts of this book chapter were written when the author was a visiting researcher at Harvard Business School (HBS).

Worldwide macroeconomic data suffer from three fundamental problems - high dimensionality, a staggered release schedule, and poor data quality. Nowcasts are a popular set of tools that address the first two problems, and the advent of alternative or Big Data offers a chance to address the poor data quality. In this chapter, I provide an overview of nowcasting techniques, discuss the need for an ex-ante hypothesis to guide alternative data selection, and compare typical alternative datasets to traditional data on several quality dimensions such as timeliness and granularity. Finally, I present a case study that establishes that search data can statistically and economically significantly improve US government employment data along the timeliness and accuracy dimensions - a novel result. The case study nowcasts revisions to Non-Farm Payrolls (NFP) three months in advance of the government data, proves these revisions are news and not noise in the framework of Mankiw et al. (1984), controls for Wall Street analyst predictions, and finds that machine learning techniques such as random forest and elastic net provide a substantial improvement over traditional linear regression methods.

1.1.1 The Fundamental Problems of Macro Data

Imagine having the difficult job of a Fed chair in April 2020: COVID-19 is spreading worldwide and we are not yet sure of the economic impact of the disease on the

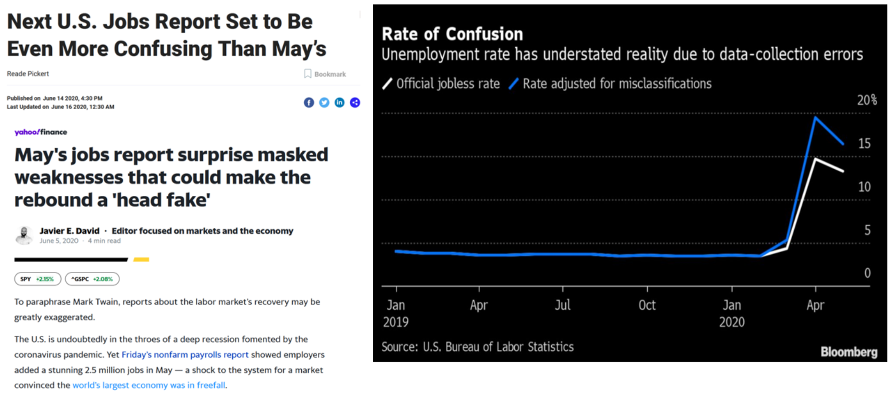

Figure 1.1.1: News headlines illustrate the confusion of real-time noisy economic data.Used with permission of Bloomberg Finance L.P. Source: Bloomberg L.P., June 14, 2020. and Yahoo finance, June 5 2020.

United States. We wait for the employment situation report (Non-Farm Payrolls or NFP), an important and trusted indicator for the month of March 2020 that will be released on April the 3^{rd}. The March number of -701,000 suggests a large magnitude of job losses which is a big change from the +273,000 jobs gained in February. As if that were not bad enough, the data reported for the March 2020 employment situation worsens over time. The May employment report released in the first week of June reveals that the labor market was worse than expected: the actual jobs lost in March were -1,373,000 - considerably higher than the first estimate of -701,000. A similar issue plagues the unemployment rate that declined from 14.5% to 13.3% in the month of May, but after adjusting for data errors appears to have been 16.4% - an *increase*. Fulfilling the dual mandate of "stable prices and maximum sustainable employment" would be easier with better data.[1]

This thought experiment gives us a sense of the fundamental problems with measuring and managing such an enormously complex mechanism as the economic activity of a country. The data collected to support this difficult task are imperfect: *they are many, staggered, slow, incomplete, subject to methodological change, and noisy (hence frequently revised),*[2] *coarse, and maybe potentially biased.*[3] [4] We can organize these problems into are three broad categories:

[1]Orphanides (2001) and Orphanides and Williams (2007) cover the issues of data revisions and their applications to central banking. Orphanides documents how the Fed overreacted to bad measures of inflation in the 1970s, causing the monetary policy to be too easy.

[2]Kliesen (2014) notes some of these fundamental problems as does Jain (2019). Croushore (2011a) provides an overview for how impactful data revisions are especially when it comes to forecasting, policy-making, business cycle turning point predictions, and other macro research. It seems that the list should include asset allocation as a practical matter.

[3]Philadelphia Fed's "Real-Time Data Research" is a valuable resource to get the appropriate data to better test any forecasting methodology after accounting for revisions.

[4]For example, Aruoba (2008) documents predictable biases in US data.

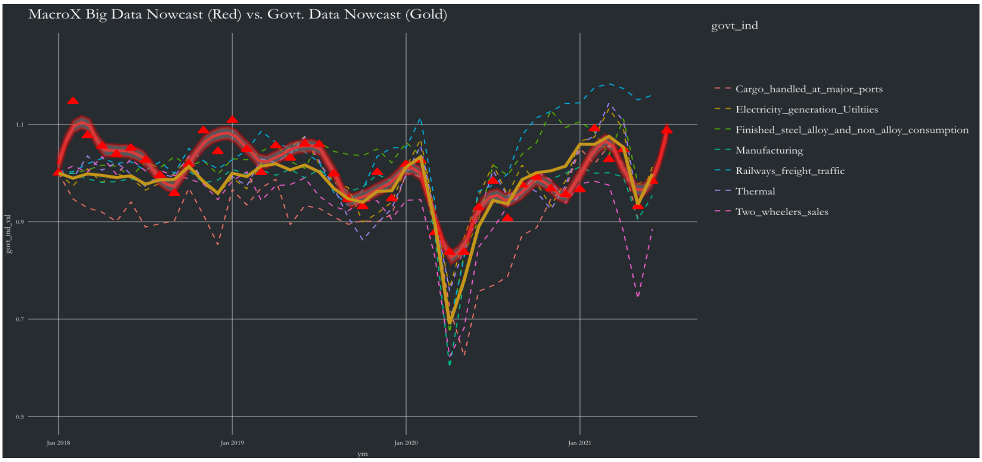

Figure 1.1.2: An alternative or Big data based nowcast can be quite useful in making real-time decisions. The thick red line with embedded triangles shows a satellite and social data based nowcast for India that is at least one month ahead of the typical government data releases and captures the big economic recovery in July 2021 ahead of the various government data (dotted lines) and a traditional nowcast (yellow line) of those government data. Nowcasts and data source MacroXStudio.

1. The Big Data or high dimensionality problem where the number of features (p) is significantly greater than the number of observations (T), also written in mathematical shorthand as $p \gg T$. Thousands of statistical variables get measured, not all of which are relevant for our purposes. For example, FRED a database maintained by the St. Louis Fed has more than 786,000 series and the highest number of variables utilized for a nowcast is around 500.

2. Economic data releases are measured at various frequencies - some such as real GDP may be quarterly and some such as Initial Jobless Claims may be weekly. The data are also released at different times, with various lags and revisions. This is known as the staggered or "jagged edge" problem.

3. The poor data quality problem which includes lagged availability - typically ranging from 1 to 3 months, imprecise measurements and sampling issues, a lack of detail, biases, and substantial revisions. As recent technological progress has reduced the measurement cost for some of the data and generated entirely new types of data not available before (Jain (2019); Krishnamachari (2017), one hopes that alternative data can contribute to reducing the impact of this problem.

In the following sub-section, we will discuss how the first two problems are addressed in the macro nowcasting context. Typically, the high dimensionality problem is tackled by dimension reduction methods and the "jagged edge"

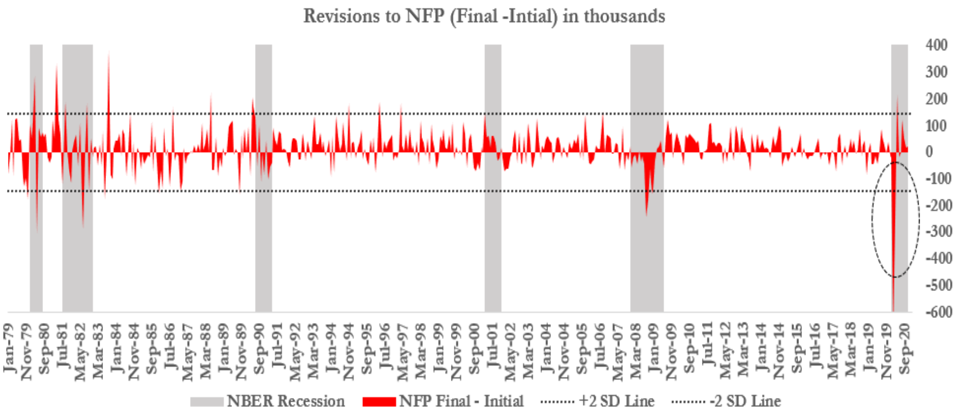

Figure 1.1.3: NFP revisions = Final number (released up to three months later) - the Initial number, cluster and appear at least "locally biased" in the fashion Croushore (2011a) postulates - with better data, expansions get revised upwards and downturns downward. Also, we can see from the figure that the biggest revisions tend to happen when the business cycle shifts which is exactly when accurate information is the most valuable to market participants per Jain (2019). Data Source: FRED and Haver Analytics.

Table 1.1.1: Data revision for some major US economic releases.* indicates significant difference in "good" (data release > series average) and "bad" (data release < series average) economic times via 1000 bootstrapped confidence intervals. The computations *exclude* COVID-19 affected periods which would increase the revision magnitudes substantially, to focus on the long-term mean and patterns. Data source Bloomberg and calculations author's.

Data	**Mean**	**SD**	**%Revision in "good" times**	**%Revision in "bad" times**	**SD in "good" times**	**SD in "bad" times**
US GDP growth (QoQ)	2.2%	2.4%	26%	80%∗	1.35%	2.13%∗
US NFP change (mom)	96.4	212.1	26%	46%∗	77.5	214.6∗
US Retail sales (mom)	0.33%	0.42%	47%	136%∗	0.25%	0.26%∗

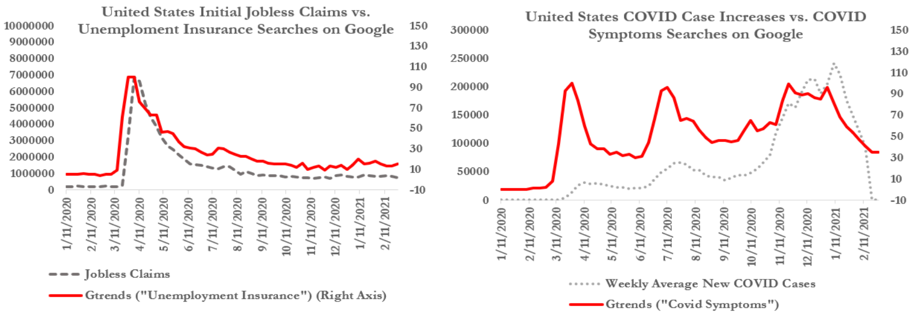

Figure 1.1.4: Search data seem faster than the government data. Data source Google and FRED.

problem by ingenious data release calendar construction and Kalman Filtering methods.

1.1.2 High Dimensionality problem

A high-dimensional($p \gg T$) problem is where the number of potential explanatory variables p far exceeds T the total data points available. This problem has been exacerbated in the last two decades as technology has provided means to collect an almost unlimited number of features which increases p. However, the number of observations T stays limited due to cost concerns or the inherent nature of the problem - for example, one cannot typically generate more economic recessions. Even for the simplest linear (OLS) case [5] when p is less than T (or N) but increasing, the expected prediction error or EPE that scales linearly with the p/T ratio can become a serious issue.[6] The two major concerns in the high-dimensional set-up are overfitting and high variance of the resulting model.

To reduce variance, typically "simple, highly regularized approaches often become the method of choice" (Ch 18, Hastie et al. (2009)). We hope that constraining the estimated coefficients reduces the variance substantially while only minimally increasing the bias. There are many methods to do so which may be categorized into:

- Parameter reduction. Unimportant dependencies are assumed away to reduce the number of parameters to be estimated. For example, diagnolizing the with-in-class covariance matrix.

[5] The simple OLS model being unbiased is a strong assumption. For instance, a regime-switching model such as one implemented by Diebold (2000) can better model business cycle dynamics vs. the OLS.

[6] Friedman et al. (2001) section 2.5 has a good overview including the "Curse of Dimensionality" and the bigger VCV matrix estimation.

- Regularization. Penalty terms are added to the error minimization objective function formulation. This penalty can either eliminate features, reduce their coefficient estimates or provide a combination of elimination and reduction. For example, with either a L_1 Lasso regressions penalty "selects" variables by making many coefficients *exactly* 0, and an L_2 Ridge regression penalty tends to smoothly "shrink" the coefficients of the correlated variables towards each other and 0. Elastic net models are a weighted average of lasso and ridge.

- Dimensional reduction. Variables are projected on a smaller dimensional subspace. For example, Principal Components Regression reduces the dimensionality of features via PCA and then regresses the principal components on the Y variable, Partial Least Squares reduces the dimension of X_p along the direction with a high correlation to Y, and Supervised Principal Components performs dimension reduction only of variables highly correlated with Y or the outcome of interest, and then regresses.

While variance reduction methods, carefully and methodically applied, can mitigate the increased variance from the $p \gg T$ problem, they cannot *solve* it. As the number of features increases exponentially the performance of almost all variance reduction techniques declines. To illustrate the point, James et al. (2013) (chapter 6) perform lasso on a data set with number of observations, $n = 100$, and choose number of features or $p = 20, 50$, and 1000. They find that while high regularization can help in the $p = 50$ case, for the $p = 1000$ case the results are poor across *all* values of the regularization penalty or λ. Even in PCA (which is conceptually similar to a Ridge regression in that it shrinks all standardized coefficients towards 0), the presence of many idiosyncratic components can produce poor results. Additionally, the assessment of the significance of features or the "multiple-testing problem" becomes even more challenging in the $p \gg T$ context.[7] [8] In the case of macroeconomic research, the small (about 195) cross-section of countries as well as the limited time series of accurately recorded business cycles or other economic phenomena of interest can be serious limitations.

Complementing machine learning methods with domain expertise and frameworks can produce better results than pure machine learning. Hastie et al. (2009) suggest the "use of scientific contextual knowledge" to figure out the most appropriate forms of regularization. The macro literature with a long tradition of analyzing economic fluctuations [9] has provided us with a rich set of methods that combine domain and statistical expertise.

[7]Friedman et al. (2001) (chapter 18) discuss the family-wise error rate computation - which includes well-known the Bonferroni correction (that divides α by M the number of multiple trials). For $p \gg T$ cases they recommend computing the false discovery rate instead of the Bonferroni correction that tends to be "too conservative" or finds too few variables significant.

[8]The time series context where the flow of time is meaningful and one cannot arbitrarily mix past and future values can increase the complexity.

[9]Burns and Mitchell modeled co-movements in economic data and aggregates in the 1930s and 40s and identifying recessions and expansions (Burns and Mitchell (1946)), Sargent and Sims's

1.1.3 Nowcasting the Big and Jagged Data

Nowcasting or "forecasting the present" is a term borrowed from meteorology relating to forecasting the current and near-future weather (Bok et al. (2017)). Let us first define it precisely in the macro context and then elaborate and expand upon the definition to achieve a deeper understanding:

An economic nowcast is a current, model-based and replicable, continuously updated, time granular, information rich summary that approximates a key concept like growth.

An economic nowcast is *current* - as its main purpose is to incorporate the latest information, *model-based and replicable* - since is no human judgement is used when the model is functioning, *continuously updated* - for it is typically updated at daily and weekly frequencies, *time granular*- changes in GDP are measured every day or week rather than every quarter, *information rich* - we can see the specific marginal impacts of various data releases, *summary* - there is *one* number representing the many economic data, typically via statistical techniques traditionally belonging to the family of dynamic factor models(DFMs), and finally an *approximation* since even the best measurement of the complex and dynamic economic machine is partial. As table 1.1.3.1 shows DFMs are indeed at the heart of many of the traditional nowcasting applications. Stock and Watson (2016) outline five main reasons for why DFMs have been so popular - the observation of a few latent factors summarizing many data seems to hold empirically, this factor structure is theoretically appealing, can accommodate lots of data, enables real-time monitoring, and they (DFMs) can span a wide variety of macro shocks. The dynamic factor model structure can also handle the "jagged edge" as well as dimension reduction in a coherent state space framework to which a Kalman filter can be applied. More recently, non-linear machine learning techniques ranging from random forest to deep learning have also been used effectively for nowcasting.

To move from many noisy variables to the few signals, one combines appropriate variable selection via domain expertise and dimension reduction methods discussed previously. The NY Fed selects 37 variables (table 1.1.3.1) which are "widely followed by market participants" and the Chicago Fed selected 85 indicators based on an academic article by professors James Stock and Mark Watson. The Chicago Fed's *original* monthly CFNAI Index Fed (2020) listed in table 1.1.3.1 is also a good example of the transformation to summarize by using principal components analysis (PCA). The basic decomposition is expressed simply as:

$$X = \Gamma F + \epsilon \tag{1.1.1}$$

Here $X = N\text{x}T$ [10] stacked matrix of $N\text{x}1$ standardized (stationary, mean 0,

modeled common shocks via dynamic factor structure in the 1970s (Sargent and Sims (1977)), and Stock and Watson (1989) used hundreds of variables. Bok et al. (2017) and Diebold (2000) are a good reference for an overview.

[10] We are switching notations a bit from p features to N to be consistent with the literature.

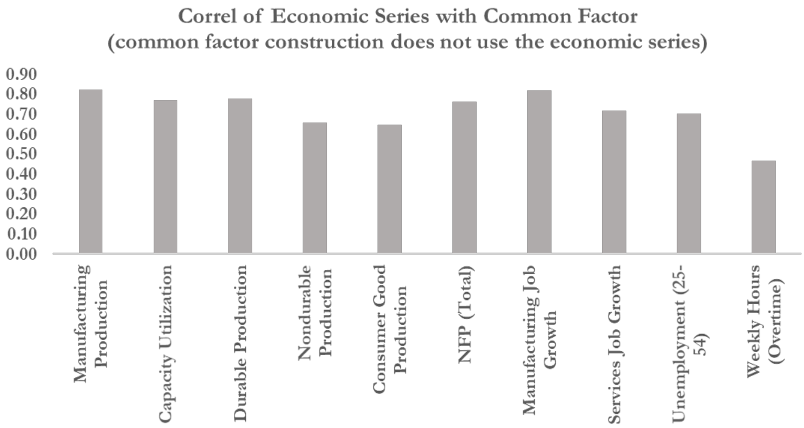

Figure 1.1.5: Correlation of some important economic series with the latent factor. Similar to Stock and Watson (2016) methodology, the specific economic series being correlated has not been used in making the latent factor to avoid spurious correlation. Data source FRED and calculation author's.

and variance 1) data vectors x_t, and Γ is an NxR time invariant matrix loading on a "few" ($R << N$) factors F (RxT) and ϵ is typically modeled as homoskedastic mean 0 random variables with variance covariance matrix $\sigma^2 I$. In the CFNAI application, $N = 85$ and $R = 1$ and the estimation method follows an eigen vector decomposition of the sample variance covariance matrix $\frac{X^T X}{N}$ and using only the *first* principal component (the one associated with the largest eigen value, indicating that it explains the most variance) to generate the weights w that can subsequently be multiplied with the economic data generated at time t, x_t to provide a "less noisy summary" or economic activity index $T_t = wx_t$.

This index "closely tracks periods of economic expansion and contraction, as well periods of increasing and decreasing inflationary pressure", thus revealing regime switching dynamics via the common factor patterns in the data that might be hidden if we examined only the individual observations since each data series has its own idiosyncratic noise. Though the CFNAI was originally designed to track inflation per Stock and Watson (1999), it has had more success in aligning with U.S. expansions and recessions as Brave et al. (2019) note.

The Kalman filter that recursively generates new predictions by updating the factor estimate using the projections of the current data innovations can systematically accommodate *all* the variables into one coherent state-space framework. As an example, consider the static factor F that was invariant in 1.1.1 has been made dynamic into f_t and given its own transition dynamics represented by the matrix A. Once again $X = N$ x T stacked matrix of *measured* Nx1 data release vectors x_t. This is a typical "state space" representation with f_t as the latent or unobserved factor. A can be as simple as an $AR(1)$ process, but 3 lags are more typical. A benefit of this dynamic structure is that it utilizes the time-series variation in

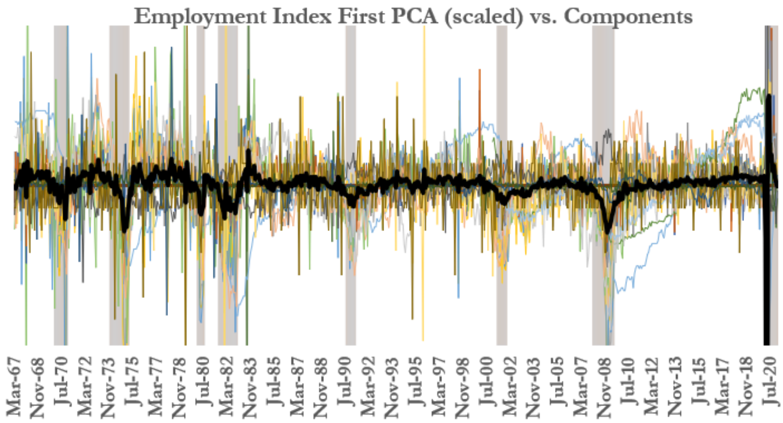

Figure 1.1.6: Latent factor for US employment as the first principal component (scaled) of more than 35 economic series. The less noisy principal component appears better at signaling economic cycle shifts. Data from Haver analytics.

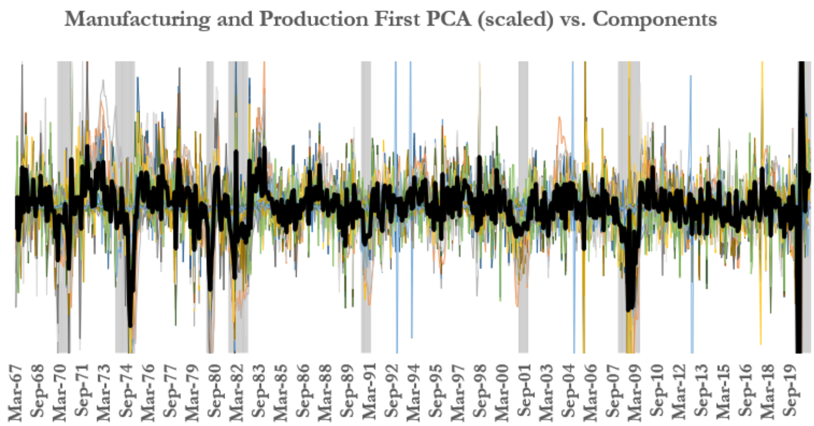

Figure 1.1.7: Latent factor for US manufacturing and production as the first principal component (scaled) of 25 economic series. Data from Haver analytics.

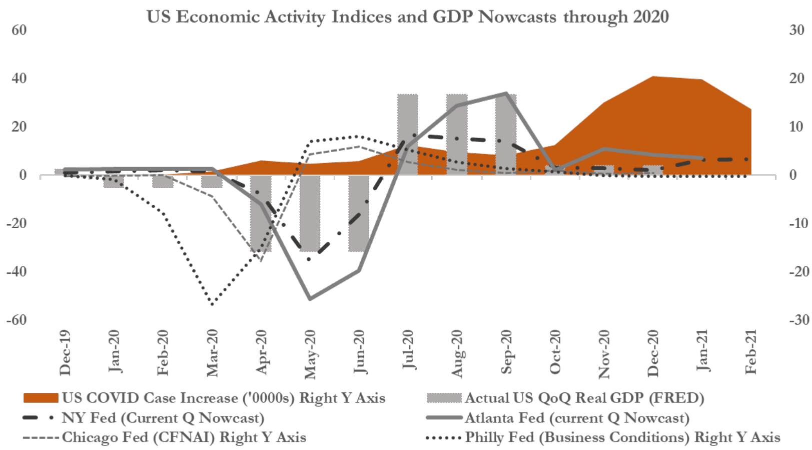

Figure 1.1.8: Comparing the US economic activity indices (typically mean 0) and GDP nowcasts through the recent past. Data sources in the final row of 1.1.3.1.

addition to the cross-sectional variation handled by the principal components.

$$x_t = \Gamma f_t + \epsilon_t \tag{1.1.2}$$

$$f_t = A f_{t-1} + \nu_t \tag{1.1.3}$$

The literature (Bok et al. (2017)) also generally assumes $Var(\nu_t)$ =1 and $Var(\epsilon_t) = H$ and NxN diagonal matrix or each ϵ_t can be decomposed into an AR(1) process an with idiosyncratic component and uncorrelated white noise. The parameters can be estimated using the Kalman smoother and EM algorithm. In the first step, the algorithm is initialized using sample PCA to estimate the factor f_t and then OLS to derive the initial parameters. The second step consists of using the Kalman smoother that uses the entire prior sample to generate an updated estimate of the common parameter. MLE estimators that can accommodate the uncertainty of estimated factors and arbitrary missing data can also be handled.

1.1.3.1 Using the real GDP to reduce noise.

To achieve the goal of noise reduction, there are opportunities for the creative uses of many statistical learning techniques such as partial least squares, supervised principal components (dimensionality reduction), elastic net (variable selection and shrinkage) as well as unsupervised learning. A recent example is Brave et al. (2019) who utilize "Collapsed Dynamic Factor Analysis" pioneered by Bräuning and Koopman (2014) that *simultaneously* models the target variable (real GDP in this case), the principal components, and the unobserved dynamic factors. The idea behind Bräuning and Koopman (2014) is that only the variables *relevant* to predicting or nowcasting the GDP should be selected instead of estimating the dynamics of the large number of other macro variables x_t. In the first step, they carry out a principal component analysis of x_t to lower the number of parameters maximum likelihood has to estimate. In the second step, they reduce the noise related to the principal components being estimates by jointly modeling the principal components with the target variable y_t in the smaller parameter space.[11] For this example, we chose y_t to be the real GDP. However, we can also use other series related to growth or inflation. Another use for this methodology could be to improve the poorly measured growth estimate of country A if we have access to another country B's better-measured growth estimate that happens to be highly correlated to country A's.

[11] Brave et al. (2019) use this technique to estimate the common components as well as generate a monthly nowcast of the real GDP. They write y_t as the hidden or latent real GDP growth that along with x_t depends on two common factors μ_t and f_t as well as a long term trend α and noise η.

$$y_t = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_t \\ f_t \end{bmatrix} + \alpha_t + \eta_t \tag{1.1.4}$$

$$x_t = \begin{bmatrix} \mathbf{1} & 1 \end{bmatrix} \begin{bmatrix} \mu_t \\ f_t \end{bmatrix} + \epsilon_t \tag{1.1.5}$$

Following Bräuning and Koopman (2014) Brave et al. (2019) collapse the panel of 500 monthly time series by pre-multiplying by the factors obtained from PCA of x_t.

$$\begin{bmatrix} l' \\ \hat{\Gamma}' \end{bmatrix} x_t = \begin{bmatrix} l' & \hat{\Gamma}' \end{bmatrix} \begin{bmatrix} \mathbf{1} & \Gamma \end{bmatrix} \begin{bmatrix} \mu_t \\ f_t \end{bmatrix} + \begin{bmatrix} l' \\ \hat{\Gamma}' \end{bmatrix} \epsilon_t \tag{1.1.6}$$

They set $l'\mathbf{1} = 1$, d $\hat{\Gamma}'\mathbf{1} = 0$, $\hat{\Gamma}'\Gamma = 1$, and let $l'\Gamma = \gamma$ - a scalar and obtain the collapsed measurement equation describing the dynamics of $\bar{x}_t$ and $\hat{f}_t$. After this they add the real GDP as an instrument that will correct for the bias in the *estimation of* the principal components that can arise from weak factors. Equation 1.1.7 emphasizes treating the principal components *explicitly* as an errors-in-variables problem - note that $\hat{f}_t$ on the left indicates *estimated* f_t. Thus the overall measurement equation can be written as:

$$\begin{bmatrix} y_t \\ \bar{x}_t \\ \hat{f}_t \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \gamma & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_t \\ f_t \\ \alpha_t \\ \eta_t \end{bmatrix} + \begin{bmatrix} 0 \\ \bar{\epsilon}_t \\ \hat{\epsilon}_t \end{bmatrix} \tag{1.1.7}$$

And the factor dynamics are represented as:

$$\begin{bmatrix} \mu_t \\ f_t \\ \alpha_t \\ \eta_t \end{bmatrix} = \begin{bmatrix} \rho & 0 & 0 & 0 \\ 0 & \phi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ f_{t-1} \\ \alpha_{t-1} \\ \eta_{t-1} \end{bmatrix} + \begin{bmatrix} \nu_t \\ \xi_t \\ \nu_t \\ \eta_t \end{bmatrix} \tag{1.1.8}$$

Other standard assumptions such as diagonal covariance matrices and the common factors having stationary dynamics are also made. Finally, they perform "triangle averaging" of the various monthly GDP estimates to match the quarterly real GDP. Brave et al. (2019) report that for understanding the business cycle, the two components μ_t and f_t interpreted as the leading and coincident/slightly lagging indicators respectively, are quite useful. The coincident factor f_t is the most predictive, but the trend (α_t) and the irregular component (η_t) are not relevant.

Table 1.1.2: Some Traditional Nowcasts.

Field	**NY Fed** **Suspended since Sep 2021*	**Atlanta Fed**	**Chicago Fed**	**Philadelphia Fed**
"Y" or Nowcast Variable	Current Q and Previous Q Real GDP growth rate.	Current Q and Previous Q Real GDP growth rate .	Overall Economic Activity and Inflationary Pressure with a 0 value indicating trend growth. Also estimates monthly GDP growth.	Real Business Conditions with a 0 value indicating average business conditions.
Update frequency	**Weekly**. Every Friday at 11:15 a.m.	**6-7 times a month** corresponding to major economic releases.	**Monthly** with one month lag.	**Daily** in real time.
"X" or Data Used	**37 monthly and quarterly time series** "that move the market and make front page news" grouped into the following categories: Housing and construction, Manufacturing, Labor, Income, Retail and consumption, Income, Surveys, International Trade, and Other (parameter and data revisions. Local blocks are used to control for idiosyncrasies in sub-groups of these variables.	**124 monthly time series from 26 data releases** are used to generate a single common latent factor that summarizes economic activity. Some series like consumer confidence are used only to generate the common factor but not to predict the GDP. GDP is further subdivided into 13 real quantity components like the BEA and the various monthly series corresponding to the 13 components are forecast and then combined using "bridge equations."	**500 monthly time series of U.S. real economic activity**. They start with the 85 time series traditionally used by CFNAI that focus on 4 major categories - production and income; employment, unemployment and hours; personal consumption and housing; and sales, orders, and inventories, and add additional series from the Conference Board, FRED MD, Atlanta Fed and NY Fed Nowcast models.	**6 series of varying frequency**: **weekly** initial jobless claims; **monthly** payroll employment, monthly industrial production, monthly real personal income less transfer payments, monthly real manufacturing and trade sales; and **quarterly** real GDP.

Methodology Summary	**Dynamic factor model** estimated by Kalman filtering and likelihood based methods that directly predict GDP. Accommodate the "jagged edge" or staggered nature of data releases. First all the macro series are summarized using PCA, and the factor loadings, as well as dynamics are initialized by OLS parameters. Then, using Kalman smoother, an updated estimate of common factors is obtained.	**Bridge equation methods** that combine **dynamic factor models'** "top down" approach (direct prediction of GDP with no sub-components predicted like the NY Fed) with a "bottom-up" approach of predicting the 13 GDP components using bridge equations that regress the growth rate of a component on one or more related monthly series as well as their own lags.	**Collapsed dynamic factor model** This is an interesting application of jointly modeling the target variable Y_t - real GDP in this case - in a small-scale dynamic factor space. Thus, real GDP acts as an instrument or supervising variable Y_t to help correct for potential biases from many weak factors. The details are in the footnote in the previous section.	**Dynamic factor model** where business conditions are treated as the latent variable. They explicitly include different and high frequency daily data in their estimation procedure.
References	Nowcasting Report NY Fed. *Suspension due to pandemic induced volatility*	Atlanta Fed GDPNow	Brave-Butters-Kelly Index with 500 variables (newer). CFNAI Index with 85 variables (older).	Aruoba-Diebold-Scott Business Conditions Index

1.1.3.2 Non-linearity and ML for Nowcasting

An interesting strand of macroeconomics forecasting literature that has been successfully using newer machine learning techniques has emerged recently. The Y variables, as it were, remain standard - unemployment, inflation, industrial production and such, but more and more non-linear and computationally intensive modeling techniques are being used on X variables of both types - the traditional macro data and newer alternative data. Researchers are also successfully dealing with the challenge of applying these methods to time series analysis where the *order* of data is crucial. The ML techniques being used cover a variety ranging from tree-based methods like *Random Forest* (Medeiros et al. (2021), Chen et al. (2019)), *Gradient Boosted Trees*(Döpke et al. (2017)) - especially XGBoost, to *elastic net* techniques (Hall et al. (2018)) such as Lasso and Ridge regressions with various regularization schemes as well as time variation (Yousuf and Ng (2021)). Other successful techniques include *Support Vector Machines* (Sermpinis et al. (2014)) that resemble logistic regression, as well as more all the way to neural networks like *Back-Propagation (BPN)*[12] and *Encoder Decoder* (Smalter Hall and Cook (2017)) models.

A striking conclusion of this strand is that the newer machine learning techniques seem to not only pick up on business cycle turning points with more success than the more traditional linear techniques but also outperform the standard linear techniques at both short and long[13] forecasting horizons.

A natural question to ask is what *aspect* of machine learning lies behind this outperformance? Work such as Coulombe et al. (2020) who perform a "meta-analysis" of various machine learning techniques for macroeconomic forecasting are investigating this question. They design experiments to identify which features of ML models - nonlinearity, regularization, cross-validation, or alternative loss function are impactful for macro prediction in both data-rich and data-poor environments. Coulombe et al. (2020) report that nonlinearity (especially Random Forest and Kernel Ridge Regression techniques) has the most positive impact, followed by cross-validation to pick appropriate hyper parameters, but regularization and alternative loss functions do not seem to improve much upon existing linear methods. The effects are even more pronounced when using alternative data, predicting economic cycle turning points, and for longer horizon forecast problems.

Specifically for nowcasting, a study by the Reserve Bank of New Zealand and Bank of England researchers (Richardson et al. (2021)) finds that machine learning models perform better than both a simple AR benchmark as well as the dynamic factor model and also add value when ensembled with official forecasts. The study notes - *"The top-performing models - boosted trees, support vector machine regression and neural networks - can reduce average nowcast errors by approximately 20-23 percent relative to the AR benchmark. The majority of the ML algorithms also outperform the dynamic factor model."*

Meteorology, from which we have borrowed the very term 'nowcasting' is far ahead of us in terms of model sophistication as well as performance. Meteorologists use deep learning models such as convolutional LSTM (Xingjian et al. (2015)) with millions of variables and achieve more than 95% predictive accuracy for the next 3 days weather (Bauer et al. (2015)). Additionally, many other fields such as public health and epidemiology also combine and utilize alternative data effectively for nowcasting.[14]

As we learn from these fields, we should remember that the underlying phenomenon being modeled can be quite different; the weather though complex, does not exhibit reflexive behavior unlike humans who might change their behavior based on the knowledge of others' behavior whether noted by Keynes (1936),

[12]Moshiri and Cameron (2000)) is an example of the early work in the field.

[13]For example, Datar et al. (2020) who utilize accounting data to forecast US GDP find that Random Forest algorithms are effective forecasting 3 to 4 quarters ahead.

[14]Venkatramanan et al. (2021) use mobility data and machine learning to forecast influenza activity.

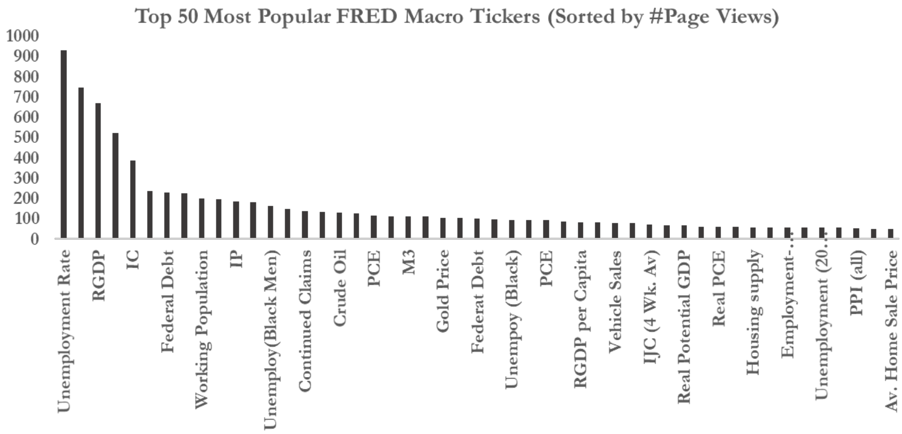

Figure 1.1.9: The top 50 most popular FRED macro tickers sorted by number page views on a Relative Basis (with a 1000 being an imaginary maximum). This seems in line with the power law in empirical data per Clauset et al. (2009). Metadata and usage statistics retrieved from FRED Data Desk team, Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org/, February 9, 2021.

(chapter 12) in his famous beauty context example or other financial market phenomenon as noted by Soros (2009). Popper (1961) in his book - The Poverty of Historicism, notes not only the philosophical differences between physical and social sciences but the negative social impact of narrow minded and idealistic pursuits based upon these ideas. More recently, Lo and Mueller (2010) caution us against the enthusiasm for Physics like "immutable laws" in explaining economic and social phenomenon which can have unintended consequences and O'neil (2016) reports examples of algorithms exhibiting unintended biases and causing economic and social harm in multiple fields.

1.1.4 Dimensions of Alternative Data Quality

1.1.4.1 A Crowd-Sourced Experiment

In practice, domain expert input is important to improving data selection, cleaning, and hypothesis generation. As a proxy for expert input we carried out an experiment with the cooperation of the St. Louis Federal Reserve Bank. We obtained the top 1,000 most popular macroeconomic data releases as ranked by FRED show in Figure 1.1.9. This is similar in spirit to the NY Fed's variable selection criterion of the ones that "move markets and make front page news" (Bok et al. (2017)). Immediately, we can see that all macro data are not created equal: the amount of attention is distributed in a "power law" sense with the top two - unemployment rate in the first place and inflation (CPI) in the second getting 930 and 748 views respectively on a 1-1000 imaginary relative scale. The 15^{th} most popular release - continued unemployment insurance claims - has only 137 views and the 1000^{th} most popular series which happens to be the labor force participation rate specific to 20 year and over white men merely 3 views. We found that the top 50-100 variables obtained by this method corresponded to the variables generally used by the various nowcasts - for example, the variables obtained with this method for the employment sector were the unemployment rate, NFP, initial claims, and labor force participation.

1.1.4.2 The Need for a Hypothesis

Millions of features becoming available through new technologies (figure 1.1.10) is very exciting but it brings its own set of challenges. The St. Louis Fed maintains a huge collection of macroeconomic data FRED with more than 788,000 data series *already* available. We saw in table 1.1.3.1, that the number of variables utilized for nowcasts range from 6 to 500. A simple calculation - ignoring the fact that some of these series are sub-series of each other - shows that even for a 500 variable nowcast, the ex-ante probability of *any one data series* being used is 0.06%, which is orders of magnitude smaller than the probability of not being used at 99.94%. When the statistical learning expert is making the model and considers new features, she faces the ex-ante trade-off of a certain increase in variance vs. the uncertain reduction in bias. Even if a feature appears individually quite related to the 'y' variable in question, it is difficult to guess its noisy marginal predictive benefit after accounting for other variables' predictive power.

Since all data are not equally important, merely adding more data series to the large quantity of already available ones is not an optimal method to increase information about the state of economy. In fact, when the $p \gg T$ problem is severe, with the number of relevant features vs. total possible feature being quite low at (0.06%), the problem *worsens* as we add more features. Additionally, pure statistical techniques such as regularization cannot solve this problem for any value of the hyper-parameter.

Faced with these obstacles, we have arrived at the need for a prior hypothesis. In other words, a structure or set of constraints that can guide variable selection. This variable selection methodology should address our *ex-ante need for economic intuition or even provide a theoretically motivated assurance that this new data series –by construction–* can substantially improve a *particular dimension of data quality* as compared to traditional data. For instance, we may begin with an initial hypothesis about the data set in question being beneficial along the increasing timeliness dimension if it is collected in real-time, or along the lowering bias dimension if it consists of the entire population's response as compared to a sample of the population, and the subsequent analysis can act as a "posterior updating" about the interaction of that data source and the economic phenomenon in question. Asset management researchers like Lopez de Prado and Lipton (2020) are recommending that one "develop theory, not trading rules", similar to the demand for structural models in macroeconomics.

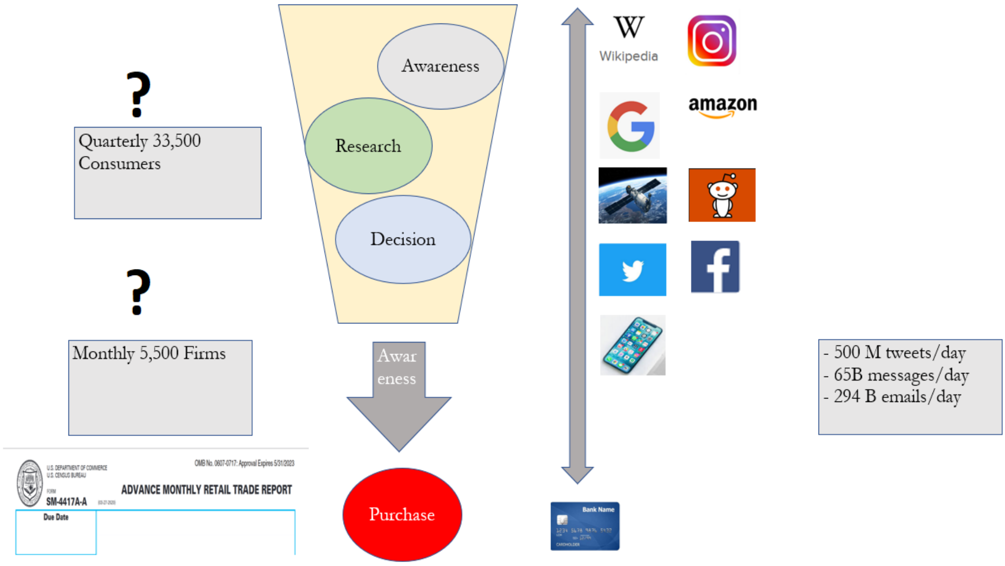

Figure 1.1.10: Traditional data vs. Big or alternative data along the marketing consumer funnel. Consumer becomes conscious of the need in Awareness phase, finds information in the research phase, defines the purchase set in the decision phase, and finally procures product in the purchase phase. Traditional data typically have little information on awareness and decision phase as indicated by question marks. Alternative data typically occupy more than one spot in the funnel, as indicated by the arrow on the right stretching across all phases, in support of Jansen and Schuster (2011) who conclude that the process is more complex than the simple linear representation.

Thinking through a common alternative data source - web search to nowcast government data may illustrate the idea. Many people use web search to seek information before they act, so by construction search data might have the potential to be more timely than the slower government statistics; also it is well known that the sample size of the search population is much higher than that of standard government surveys so search data are likely to be more precise. However, ex-ante we may also be concerned about the urban and tech-centric bias as well as inherent noise in web search data, since heavy tech usage tends to be urban and trending searches can become more popular without any chance in the economic phenomenon. While this is not a full-fledged theory about the economic phenomenon in question, these initial hypotheses act as informed priors that can afford a more rigorous evaluation beyond the merely statistical model fit. If subsequent analysis reports that the nowcast based on search data excels in older rural populations but not in urban tech-centers, one would be skeptical regardless of the statistical fit. Table 1.1.3 considers common characteristics of data that a researcher might wish to focus on and compares some traditional and alternative

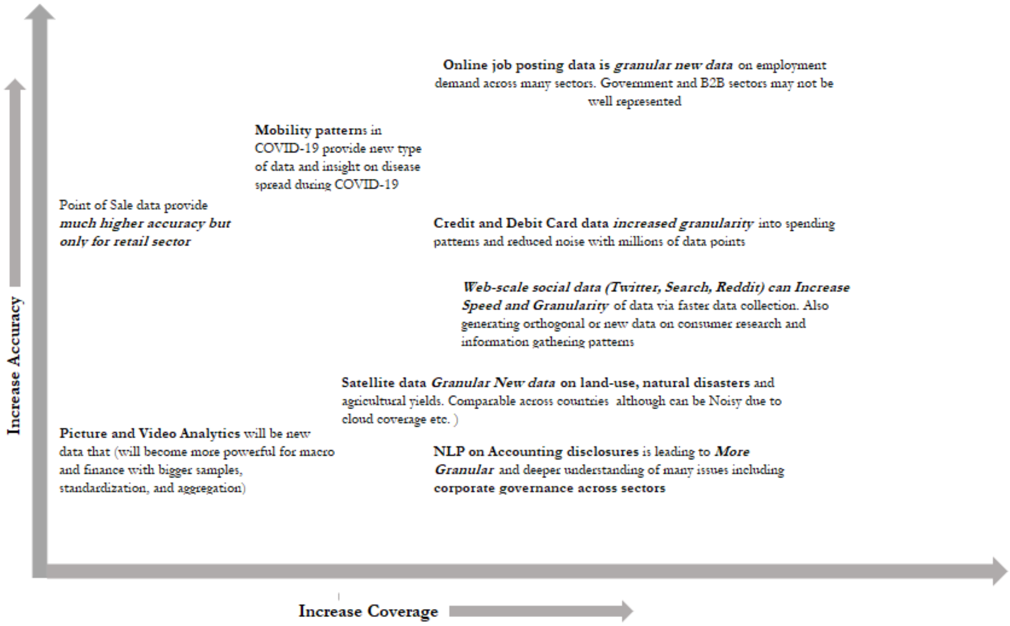

Figure 1.1.11: Sample data quality hypotheses for a few commonly used alternative datasets.

data sources along those dimensions.

1.1.5 NFP and Web Search Case Study

We identified three big problems with macro data - the Big ($p \gg T$) data, the staggered releases, and the poor data quality. The previous sections discussed how nowcasting addresses the first two and in this section we will present a nowcasting case study to improve the timeliness and accuracy dimensions of government data. We pick a popular alternative data set - web-search and the government release we will focus on is Non-Farm Payrolls (NFP) - one of the major data releases in the United States(Baumol (2013)). The question we wish to answer is:

Do search data contain economically important, unique, and more accurate information faster as compared to the official government employment data?

Table 1.1.3: Some Typical Data Quality Dimensions for Comparing Alternative and Traditional Data

Data Quality Dimension	**Traditional Data**	**Alternative Data**	**Remarks**
Timeliness	Typical **lag of 1 to 3 month** even for the United States. Can be worse for developing countries. For example, for India, the current lag is more than **1 year** with the latest *official* labor market data in March 2021 is for June 2018-2019 annual survey.[15]	Social data such as Twitter, Web search, Reddit **can be practically instantaneously collected *and* be leading indicators.** These indicators provide information on intention or the awareness, research, decision phases in figure 1.1.10 that was typically unavailable previously and are likely to **precede action such as purchase**, which was measured by traditional indicators like retail sales.	High frequency (daily, weekly) availability can also mitigate the "jagged edge" problem with traditional data. Some alternative data sources like mobility may be delayed by up to a week and credit card data may be delayed by one to three days until the transaction settles but can still be *leading indicators*.
Wide coverage (high recall)	Carefully designed and **covers many traditional sectors** in a balanced demographic way.	Typically alternative data sources are **stronger in certain sectors** such as consumer behavior and spending, labor market behavior in technology driven and gig-economy sectors. Sensor data such as satellite and mobile phones can be used to better understand mobility patterns during COVID-19 etc. However, typically alternative sources tend to have less visibility into the government and B2B sectors.	Traditional data sources may find it difficult to measure gig-economy activity (Kässi and Lehdonvirta (2018)). Even combining many alternative data sources to measure government or B2B activity is not easy due to high costs and ineffective due to potential sampling biases.

[15] Recently, India is attempting to reduce this interval to three months.

Accuracy (high precision)	The **accuracy of traditional data is low** with up to 50% revisions (Jain (2019)) of the most important data releases even for the United States. This issue partly springs from the high cost of information gathering in the past where it was necessary to make a trade-off between speed and accuracy via frequent small samples (50,000 people once a month) and infrequent full census (5 to 10 years).	Alternative data can measure cheaply and frequently at scale, but **bring their own noise from automation, usage pattern changes, as well as sample biases, data collection methodology changes etc.** (Jain (2019)). Careful curation and cleaning can help lower that noise in certain cases for example in inflation measurement for Argentina (F. Cavallo and Rigobon (2016)).	Survey responses declined during COVID-19 times (Rothbaum and Bee (2020) making it challenging even for well designed traditional data sources. The other challenge is measuring and weighting new sectors appropriately in case of fast technological change.
Long time series	Traditional data tend to have **much longer time series**, typically 30 years or more, and cover multiple economic cycles.	A key issue with alternative data is the **shorter time series** - anywhere from 3 to 15 years which may not cover many economic cycles.	Researchers like Adler et al. (2019) are showing creative approaches in combining official (and potentially under reported) census data to the newer and granular search data which does not have as lengthy a time series.
High granularity	Traditional data tend to offer **limited granularity** since they are sampling based and utilize the same set of questions. For example, the household survey for NFP polls 50,000-60,000 respondants per month.	Naturalistic alternative data can **offer immense granularity** on a minute by minute basis. For example, there are 500 million self generated tweets per day of which even if 5,000,000 are relevant to a particular economic topic is still substantial.	While minute by minute granularity is interesting and extremely useful, it does not increase predictive power of economic phenomenon linearly (Jain (2019))

Low bias	**Traditional data sources tend to be less biased** since they tend to be carefully designed with sophisticated statistical techniques for the explicit purpose of gathering that particular data. However, these data collection are not immune to agency incompetence, government intervention and politics. Worryingly,Aruoba (2008) **documents a relatively large bias - a non 0 mean deviation that may be predictable - in revisions to major economic statistics such as GDP for United States** and concludes that "these are not rational forecasts."	**Alternative data sources tend to be more biased** since they are typically not designed for analysis purposes and may appeal to a very specific population using them - whether they are large corporations, small businesses, or individuals. For Most social data such as Web-search, Reddit, Twitter tend to have an urban and tech-centric bias.	For traditional data bias issues, read (Coyle (2014) for an account of Nigeria, Ghana and a few countries' GDP increasing dramatically or F. Cavallo and Rigobon (2016) and Cavallo (2013) for using a Big Data online inflation index to find misreporting on the part of the Argentinean government. Brave et al. (2020) contains good examples of potential biases in alternative data, for instance they cite potential issues with Homebase whose customers were serviced oriented small businesses and hence showed a larger decline in hours worked during COVID-19 as compared to the balanced BLS panel.
Low Cost	Typically **free in dollar terms for users** since government agencies like the Bureau of Labor Statistics (BLS) for the United States make the data available	Typically **more expensive** with datasets costing anywhere from \$30,000 to more than \$100,000 and with the annual spend by asset managers continually growing per the JPM Big Data Guide (Krishnamachari (2017)).	The **privacy cost and risk** is perceived to be low for traditional data since the user typically volunteers information and it is just a sample but the privacy cost and risks for alternative data can be high (Jain and Seeman (2019)

In accordance with the previous section, first we examine which data dimensions or attributes might we reasonably expect search data to have as compared to the government data, second we carefully clean the data to maximize that attribute, and finally we perform some simple analysis to test our hypothesis. In terms of data dimensions, we hypothesize based on figure 1.1.11 that search is likely to be more timely as compared to government data, widely applicable across multiple sectors and more accurate since the sample size of search data is much bigger than the government NFP surveys. A concern is the potential urban and technologically capable user bias of search. The careful search data cleaning will take the form of generating and matching detailed lists of job titles to the

BLS categories and the final analysis will examine the statistical and economic significance of search data in nowcasting NFP revisions.

1.1.5.1 Background and Related Work

Over the last decade researchers working in this field have looked to web search activity as an attractive information option due to its high volume (nearly 80% of people use web search to look for a new job Smith (2015)), continuous data generation, and real time availability. Web search data being naturalistically generated as a result of people addressing their true interests and concerns (Dumais et al. (2014); Goel et al. (2010)) also makes them potentially different source of information as compared to the official survey based govt data.

There are many notable papers in the field (Choi (2009); Choi and Varian (2012); Concha Artola and Hernández (2015); Einav and Levin (2013); Goel et al. (2010); Hellerstein and Middlethorp (2012); Suhoy (2009)) that use web search to predict a variety of employment related statistics across the world in addition to other official macro statistics such as consumer confidence, tourist inflows, mortgage re-financing, auto sales, and home sales. There are two main stylized facts in this literature about search data - first that there are only small improvements by using search data[16] and the second that good existing data sources can overcome benefits from search data.[17]

[16]Choi and Varian (2012) is a seminal paper in nowcasting or predicting near term economic indicators such as automobile sales, unemployment claims, consumer confidence and others using web search data from Google. Their paper follows several other authors who predict unemployment in various countries such as US, Germany and Israel Askitas et al. (2009); DÁmuri and Marcucci (2010). They tend to agree with Goel et al. (2010) that search data do not provide dramatic improvements in forecast performance but advance the case that these small improvements are economically significant. Specifically, in the field of unemployment they predict initial claims for jobless benefits (IJC) which is very close to a random walk. In fact, they find that the prediction error in their holdout sample increases from 3.68% to 5.95% upon including the search variables, but they suggest that the search model fits better in recessions or turning points where the mean average error decreases from 3.98% to 3.44%. While interesting, Choi and Varian's claim is weakened by the fact that their sample only has one recession and hence a limited number of turning points. Also, the difference in mean average error of 3.98% and 3.44% could just be sampling noise. Other researchers such as Bortoli and Combes Clement and Stephanie (2015) also focus on forecasting household consumption in France, which have high trending and autoregressive components, and similarly find that search data do not add much predictive power over the autoregressive or trend components.

[17]Goel et al. (2010) present empirical results in several domains and pose an important question regarding the incremental information content of search data in the presence of other data sources. They find that "the utility of search data relative to a simple autoregressive model is modest." Additionally, they find that search data add value "in absence of other data sources, or where small improvements in predictive performance are material." In our case predictions of NFP statistic by Wall Street analysts in advance of the release of the statistic itself form a very good existing data source. Most of the literature including Choi (2009); Choi and Varian (2012) does not explicitly control for the baseline knowledge other than controlling for previous trend. While some of the prediction by Wall Street analysts may be based on trend, it seems naive to assume that all of the predictions by highly paid professionals with large incentives to outperform will be simply captured by a linear trend model. DÁmuri and Marcucci (2010) come the closest to

There are three major areas of concern with the prior literature:

First, most of the macro-economic variables predicted such as unemployment rate or unemployment insurance claims (also called Initial Jobless Claims or IJC) have a high (>90%) trend component (Choi (2009); Choi and Varian (2012)) which makes it difficult to judge the importance of the new web search data and tend do not impact the financial markets much.

Second, most studies do not explicitly control for existing experts' forecasts so it is not clear if the information is truly different and hence incremental to existing knowledge.

Third, as both government data and search data are noisy proxies for true and unknowable economic growth,[18] it is unclear how to ascertain the veracity of the information content.

Using a novel data set of more than 830 million employment related searches from a popular search engine (Bing), and leveraging the methodological innovation of nowcasting *revisions to the same official NFP number*, we address these three concerns.

We focus on Non-Farm Payrolls (NFP) rather than the unemployment rate or IJC used in other studies such as Choi (2009); Choi and Varian (2012). As *"the most important data release in the world"* (Baumol (2013)), NFP is economically much more relevant to market participants than other statistics such as unemployment rate or IJC. Using *changes* to NFP as our variable of concern rather than levels of unemployment rate or IJC lowers the trend component from 90% to 60%.

We use a live and realistic antagonist to search data in the form of consensus of highly trained, professional Wall Street economists who get ranked on how well they forecast. This is in contrast to imposing a simple auto-regressive structure or arbitrarily selecting one or two variables relevant to the labor market forecast. It seems reasonable that professional strategists would use all the relevant information from other government data and private vendors to make sophisticated statistical models, as well as leverage their considerable experience to forecast such an important number.

We use the novel methodology of nowcasting the *revisions* in the official NFP release itself, which typically happen 1 to 3 months *after* the initial NFP data are released. As more employers finally respond to the voluntary employment surveys upon which the official NFP statistics are based, the revised NFP numbers, by construction, contain more information than the initial NFP numbers. And if search data can predict the revisions in NFP 1 to 3 months ahead, the case for

our idea by controlling for the survey of professional forecasters as do Antenucci et al. (2014). but they predict the highly trending unemployment rate and jobless claims with low financial market impact. If a data release does not impact financial markets then it is quite likely that the information was already known to the participants and hence reflected in the prices per a standard efficient markets model (Fama (1970).)

[18] Croushore (2011b) has an example of the US real output growth for 1977 Q1 changing until 2010.

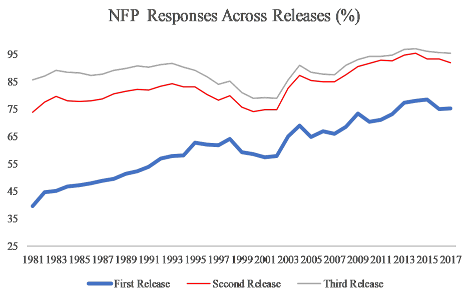

Figure 1.1.12: Response rates to the NFP Employment Survey for the first release vs. the second and third release. Data Source: BLS https://www.bls.gov/osmr/response-rates/

search having superior information becomes much stronger. We also verify the existence of extra information in revisions as compared to the initial releases via news and noise regressions and formal Granger causality tests.

In using the Google search data researchers have been cautious about the the lack of transparency and heavy aggregation in normalization of searches, as well as search term relevance declining over time as Suhoy (2009) and, Chancellor and Counts (2018) report. By using the entirety of our proprietary 830 million employment search related dataset we eliminate the normalization issue, and by cleaning and curating the employment web search dataset with the help of economics experts, we better focus the dataset on the main macroeconomics problem. Encouragingly, we find that the overall performance with web search is better than what the literature has reported thus far, and it improves further when we focus on sampling areas that government surveys do not get the best information from.

As Figure 1.1.12 shows for the establishment survey, typically only 60%-70% of the responses make it back in time for the first release. Phipps and Toth (2012) investigate which types of firms have low response rates to BLS survey and find that four main factors are important: First, the firm size measured in number of employees, second whether a firm is a "white-collar service firm" as indicated by inclusion in three super-sectors: 1) Information 2) Finance and 3) Professional and Business Services, third, whether it has multiple offices, and fourth, the size of the metropolitan service area (MSA) in terms of the population in the area where the official has to collect the employment data. The authors find that "White-collar service establishments with a larger number of multi-units have the lowest response rates." Additionally, small and medium sized businesses, which are more sensitive to cyclical changes of the economy, are also slow to respond since they might lack a specific department to respond to such surveys. Being located in a metropolitan service area (MSA) with a large population is associated with lower BLS responses. Plausibly, the firms with lower initial response rates might be responsible for the most changes from revisions. White-collar service firms in metro areas are the ones quite likely to use internet search and other web-scale data for content production. Mislove et al. (2011) find that twitter users are “more likely to live within populous counties than would be expected from Census data”, The potential *under-representation* of urban area, service firm bias in BLS data may thus be naturally addressed by web-scale data which tend to be biased in the direction of over-representation of such entities.

As has been widely documented in the literature, for example by Goel et al. (2010) and by Chancellor and Counts (2018) more recently, web-searches reflect the users’ *demand for information.* This demand driven data generation is in contrast to government statistics that rely on users’ voluntary *supply of information.* In troubled economic times, users’ demand for information could increase while the supply of information might decline. This differential between demand and supply of information can be utilized. For instance, for NFP, we believe that the differential of demand and supply for information is potentially highest among white-collar service firms located in large metro areas in sectors such as architecture, business, finance.

1.1.5.2 Government Non-Farm Payrolls (NFP) Data Overview

For this case study we focus on the non-farm payrolls (NFP) data release, which estimates the number of jobs added to the U.S. Economy each month and is released by the Bureau of Labor Statistics (BLS). This release consists of two surveys, the first is the household survey of around 60,000 households, and the second known as the establishment payroll survey is a survey of more than 600,000 corporate and government sites. Here we rely on the establishment survey, which further categorizes employment numbers into about 15 sectors and sub-sectors such as “Information”, “Retail trade”, “Professional and business service.” Changes to NFP can help signal the beginning or end of a recession if the economy starts losing or gaining more jobs as compared to the past trend.

Typical Wall Street consensus for NFP data release has around 80 estimates from qualified economists from various investment banks and other institutions, and is usually available about a week before the data release. Given the importance of the data release, these economists at institutions such as Goldman Sachs have the resources and a substantial stake in getting the number right. Bloomberg tracks the previous predictions of these economists and ranks them based on their performance. In our analysis we simply use the average of all the economists’ estimates.

There are two main uses of economic information - the first is to predict the short run impact and the second is to understand the long run trend to make policy decisions Kliesen (2014). The short run impact is typically a few hours or days and can be analyzed using “event studies” which involve the analysis of price impact of the incremental information on liquid financial instruments. The logic behind analyzing the movements in liquid instruments such as government

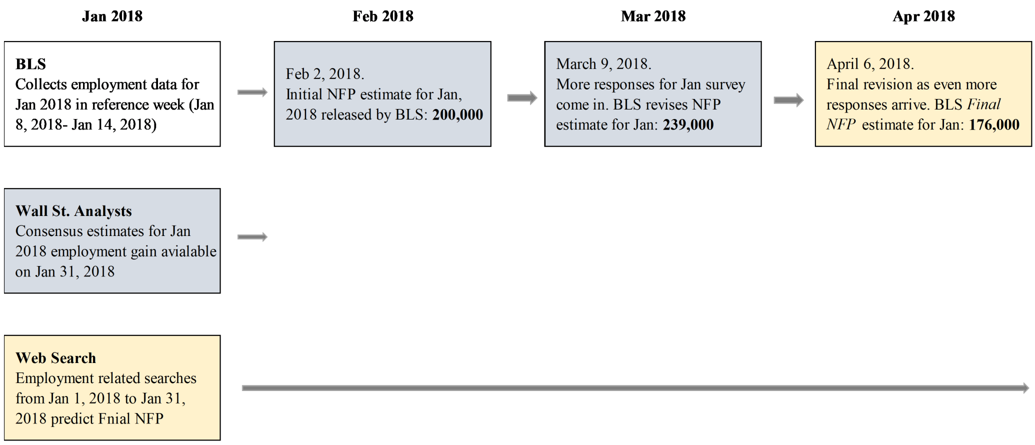

Figure 1.1.13: An indicative timeline of the Initial NFP and Final NFP data releases along with the prediction from Wall Street analysts' consensus and search-based variables.

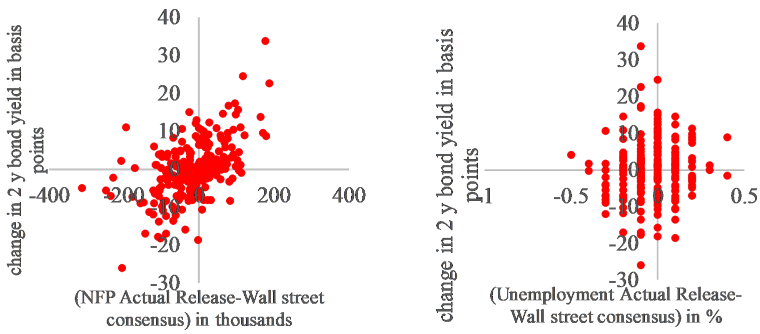

Figure 1.1.14: Movement in the 2-year bond yield for. NFP surprises (Actual release – Wall Street consensus) on the left and Unemployment surprises on the right. The NW corrected t-stat for NFP is 9.4 vs. the t stat for unemployment at 0.32 in a horse race. Data Source: Bloomberg.

bonds is based on the efficient markets hypothesis of Fama (1970) according to which prices adjust in accordance of the revelation of as new information. Thus, the more informationally important data releases should have higher price impact and the less informationally important data releases should have a lower price impact. Taking the particular example of the 2 year bond, calculations show that the typical NFP release explains around 28% of the variance of the US 2 year bond for the day, and a 1 unit change in surprise in the NFP release results in a 0.53 unit move in the 2 year bond yield which is significant at the 0.001 level. In contrast, the surprise in the unemployment rate accounts for *almost none (0.8%)* of the variation in the 2-year bond yield and a 1-unit surprise in unemployment rate only accounts for 0.06 units of movement in the US 2-year bond yield.[19]

The longer-term economic purpose of the NFP report is to help estimate the trend growth of the economy rather than just predict the one day moves in the asset markets. As noted before, this trend growth in the U.S. labor markets is of vital importance to the Federal Reserve for conducting monetary policy, as well as corporations, investors, and citizens who wish to understand where the economic cycle is for making longer term decisions. Baumol (2013); Jain (2019); Kliesen (2014); Lin et al. (2019). Substantial revisions to NFP and other macroeconomic data, make it difficult to truly gauge the current state of the economy and hence make optimal policy decisions in real time.[20]

1.1.5.3 Information Content of NFP Revisions.

Mankiw et al. (1984) suggest two possible outcomes for data revisions - news or noise. In their context, news is defined as extra information uncorrelated with the initial release that was revealed in the final numbers and noise as a forecastable correction that is unrelated to the final number. They test these two competing hypothesis by performing correlations of the revisions with the initial number (noise hypothesis) and final number (news hypothesis) as well as regressions. We find that both the correlation [21] and regression results correspond closer to the news hypothesis for NFP. [22]

We further formalize the intuitive notion illustrated by figure 1.1.15 and the

[19]Since both data releases occur on the same day, we also run a regression with both variables (not shown here) and find that indeed the t statistic associated with NFP surprises is 9.4 vs. the t stat for unemployment at 0.32. The results for IJC are similar.

[20]Specifically, for NFP, the absolute revision is about 102,000 jobs in bad economic times vs. 57,000 jobs in good economic times. Thus, the initial NFP estimates underestimate extent of job losses at the starting of downturns and also consistently underestimate the job gains when the recession ends. Figure 1.1.15, shows two lines- the blue line which represents the analysis conducted using the data available in real time and the dashed red line which represents the analysis conducted with the benefit of hindsight using revised data that are available 1 to 3 months after the initial data. As we can see, during the 2007-2008 financial crises, the Initial NFP data would be delayed by a period of 6 months as compared to the Final NFP data for determining the downturn as well as the upturn.

[21]The correlation between NFP revisions and the Initial NFP is -0.043 with (-0.16,0.10) as the bootstrapped 95% confidence interval, whereas Final NFP and NFP revisions are correlated at

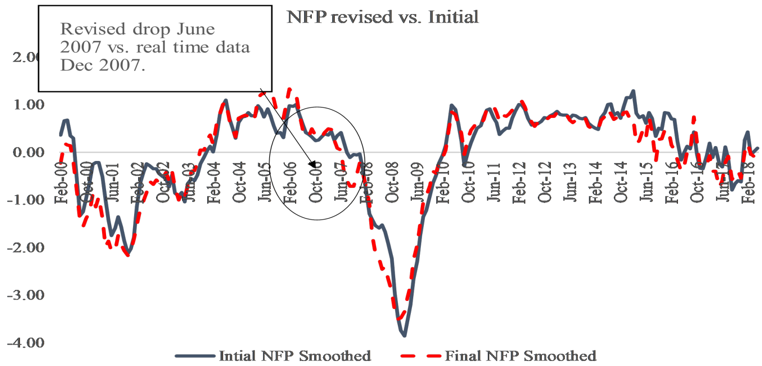

Figure 1.1.15: An intuition for how estimating economic cycle turning points using Final NFP vs. Initial NFP estimates can yield different results.

Table 1.1.4: News vs. Noise Regressions for NFP. Standard Errors are NW corrected. The β for news regressions is clearly close to 1 and for noise is significantly different than 1.

News vs. Noise Regressions

	Noise Hypothesis (Initial ∼ Final)	News Hypothesis (Final ∼ Initial)
Intercept	1.73	14.89**
(NW tstat)	(0.39)	(3.17)
Slope	0.89	0.99***
(NW tstat)	(38.6)	(38.1)
R^2	0.87	0.87
N	492	492

Table 1.1.5: Granger Causality Tests for Information in Initial NFP vs. Final NFP

Granger Causality Wald Tests

Equation	*Excluded*	*Chi2*	*df*	*Prob > Chi2*
NFP Initial_{t+1}	NFP Final_{t+3}	46.38	3	0.00
NFP Final_{t+3}	NFP Initial_{t+1}	5.83	3	0.12

analysis in table 1.1.4 that the Final NFP releases do contain *more information than the Initial NFP releases*, rather than a mere rebasing, with a Granger (1969) causality test . The tests in Table 1.1.5 show that the Final NFP data "Granger cause" the Initial NFP data but not vice versa. In other words, knowing the Final NFP numbers is useful in predicting the Initial NFP numbers but knowing the Initial NFP numbers is not useful in predicting the path of future Final NFP numbers. Thus, it seems reasonable to interpret that Final NFP numbers contain more forward-looking and correct information about the path of the economy rather than the Initial NFP numbers.

1.1.5.4 Web Search Data

We collected employment related searches on MSFT's search engine Bing[23] from January 2012 to March 2018 - a total of 74 months. These searches are from both mobile and desktop devices and constrained to only be in the English language for the country region of United States. We follow the same technique Chancellor and Counts (2018) use in cleaning the data by only using four keywords: "job", "jobs", "career", and "careers" as reflecting the best trade-off for actual results and minimizing false positives, and remove queries with co-occurring irrelevant words ("Steve Jobs"). This yields a total of around 830 million employment related queries. Before constructing various search related variables, we check if search data are substantially demographically or geographically biased across states and we find that is not the case, as shown in Figure 1.1.16.

There are two main types of employment related searches- broad and sector specific. The first type or broad searches are employment related searches typed into MSFT's Bing search engine that did not have any specific job sector associated with them. This broad category is captured by "Overall" employment related

+0.31 (+0.19, 0.44) as the 95% confidence interval.

[22]Similar to Aruoba (2008) we find a significant intercept for the news hypothesis but we believe the bias to be more "local" due to a lack of speed in the fashion Croushore (2011a) suggests - when we have recessions, the initial numbers are slow to reflect how bad the data are and vice versa.

[23]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 were gathered historically; there was no interaction with users by changing search results. No session information is used in our dataset. All data was anonymized and aggregated to county level and to national level for the final analysis. Our use and storage of this data is in agreement with Bing's End User License Agreement and Privacy Policy

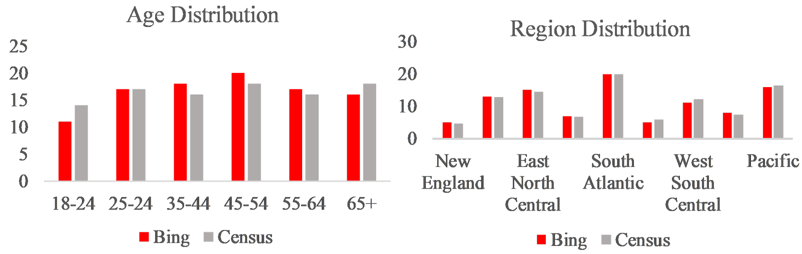

Figure 1.1.16: Demographic and geographic distribution of Bing searches vs. U.S. population. Source: MSFT and US Census.

Finance
insurance jobs in nc- sales financing jobs
Baltimore- medical billing supervisor job
description- collections jobs loisvelle
Retail
Simon outlet mall in norfoplk va job fair- small gift
shop job openings nj- Ducks sportings goods
careers- Job Description for an retail pharmacy
technician
Technology
"digital channels manager" job -Electrical
systems technology jobs alabama -altec careers
php- computer forensic jobs chandler az

Figure 1.1.17: Examples of hand curation of job searches into employment sectors

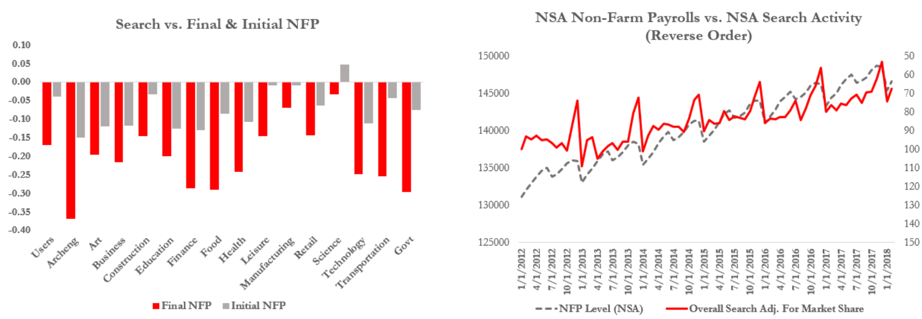

Figure 1.1.18: On the left, Correlation of employment related search categories vs. Initial NFP (gray) and Final NFP (red) release and on the right, Not Seasonally Adjusted (NSA) overall employment related searches adjusted for Bing market share changes vs. NSA Non-Farm Payrolls levels. Data Source: MSFT, FRED, Bloomberg.

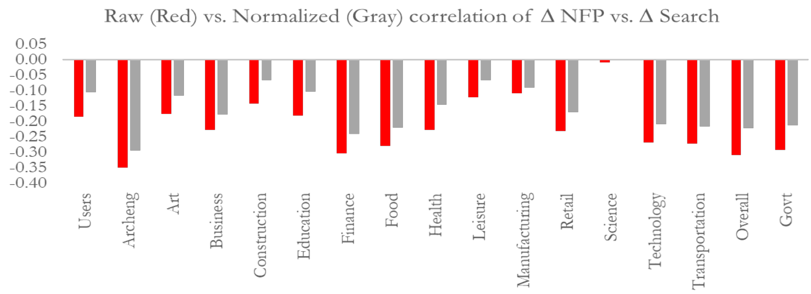

Figure 1.1.19: NFP (diffs) vs. Search (diffs) correlations seasonally adjusted raw diffs(red) and seasonally adjusted normalized differences (gray). These delta correlations do not change much with normalization unlike levels relationships.

searches. An example would be a search like "Jobs in Seattle." The second type of employment related searches are sector specific; their construction follows the technique used by Chancellor and Counts (2018) where job searches are carefully curated and matched to economic sectors obtained from about 15 BLS categories specified in the NFP report. Figure 1.1.17 provides an example of the hand-curation of searches with examples of specific searches that got categorized into various job categories corresponding to the NFP data releases. Figure 1.1.18 shows the plot of the "Overall" employment related searches. We notice that as the unemployment rate declined, the searches for employment also decline from Jan 2012 to March 2018.

To ensure our results are robust to normalization, we compute the correlations (and the regressions) *both ways - normalized by total searches and not normalized* and find very similar results. The distribution of month over month changes in search is substantially more robust than the distribution of levels - which can change dramatically depending on the denominator. Jain (2019) discusses the normalization robustness issue in greater detail. The added intuitive benefit of relying more on the month over month changes in search is that at the monthly time-scale they are more driven by changes in population search behavior rather than gains or losses in the search browser market share or another normalization factor.

1.1.5.5 Search and NFP Correlation

After seasonally adjusting both series, we perform a correlation analysis of the month over month *changes* in the 16 search signals - the 15 curated employment sectors and the broad employment search category of 'Overall' with the *changes* in Initial and Final NFP data releases for the period from 2012 to 2018.[24] The results are displayed in 1.1.6, which also contains the bootstrapped 5% and 95%

[24] Correlation of raw or NSA employment search first differences with NSA NFP first differences is 0.70, but that overestimates the economic relationship since seasonal patterns are predictable.

Table 1.1.6: Correlations of employment relates searches in various sectors with the Initial NFP and Final NFP. Median correlations are in bold and the bootstrapped 5$^{\text{th}}$percentile and the 95$^{\text{th}}$percentile correlations for both are displayed.

Search Category	***Initial NFP*** 50^{th} ***%ile***	*Initial NFP* 5^{th} *%ile*	*Initial NFP* 95^{th} *%ile*	***Final NFP*** 50^{th} ***%ile***	*Final NFP* 5^{th} *%ile*	*Final NFP* 95^{th} *%ile*
Overall	**-0.17**	-0.36	0.03	**-0.30**	-0.46	-0.12
Users	**-0.04**	-0.23	0.15	**-0.17**	-0.37	0.03
Arch.& Engg.	**-0.15**	-0.35	0.06	**-0.37**	-0.52	-0.20
Art	**-0.12**	-0.33	0.09	**-0.20**	-0.41	0.04
Business	**-0.12**	-0.32	0.09	**-0.21**	-0.39	-0.03
Construction	**-0.03**	-0.26	0.20	**-0.15**	-0.35	0.07
Education	**-0.12**	-0.33	0.09	**-0.20**	-0.41	0.04
Finance	**-0.13**	-0.31	0.07	**-0.29**	-0.44	- 0.11
Food	**-0.08**	-0.25	0.09	**-0.29**	-0.47	-0.10
Healthcare	**-0.11**	-0.28	0.08	**-0.24**	-0.40	-0.07
Leisure & Hospitality	**-0.01**	-0.21	0.21	**-0.14**	-0.36	0.09
Manufacturing	**-0.01**	-0.19	0.28	**-0.07**	-0.26	0.13
Retail	**-0.06**	-0.25	0.13	**-0.14**	-0.39	0.11
Science	**0.05**	-0.19	0.28	**-0.03**	-0.31	0.23
Technology	**-0.11**	-0.30	0.10	**-0.25**	-0.42	-0.06
Transportation	**-0.04**	-0.25	0.18	**-0.25**	-0.46	-0.03
Government	**-0.07**	-0.30	0.16	**-0.30**	-0.50	-0.07

confidence intervals for the correlation metric. The search delta correlations follow the same pattern as the levels in the previous section and are negatively correlated with the number of jobs created. There are two potential intuitions behind the negative correlation of *net* jobs created (NFP) and the employment related searches or demand for employment information: the first is that when economic times are good, as signalled by a low unemployment rate, it may cost less effort to find jobs which lowers the number of searches *per potential applicant*, and the second is that in good economic times less people may be looking to change jobs, thus lowering the number of applicants.

We also find that the search data are more correlated across categories to the Final or the revised NFP numbers that are only available *1 to 3 months after* the Initial NFP number. The stronger correlation with the Final NFP numbers rather than the Initial NFP numbers suggests that search data may contain forward looking information relevant to the revised Final NFP numbers. As evidence of robustness, correlations with the Final NFP numbers are negative, as well as higher in magnitude and the result is consistent across the 15 hand-curated sectors in addition to the Overall broad category of generic job searches.[25]

We now investigate the efficacy of search data in nowcasting the Final NFP numbers more formally. We would like to test if search data truly contain information not completely captured by the Initial NFP as well as the existing baseline estimates of the sophisticated Wall Street economists. Towards that objective we run several regressions of the form:

$$\begin{aligned}\text{Final NFP}_{t+3} = \alpha &+ \beta_1 \cdot \text{InitialNFP}_{t+1} + \beta_2 \cdot \text{LaggedInitialNFP}_t \\ &+ \beta_3 \cdot \text{EmploymentSearch}_t + \beta_4 \cdot \text{SectorSpecificSearch}_t \\ &+ \beta_5 \cdot \text{EconomistsSurvey}_t + \epsilon_t \end{aligned} \quad (1.1.9)$$

1.1.5.6 Regression Results

The results of the regressions formulated above are displayed in Table 1.1.7. We compute the standardized coefficients to understand the impact of a 1 unit move in the independent variable on the dependent variable by multiplying the actual ordinary least squares (OLS) regression coefficient by the standard deviation of the independent variable and dividing the result by the standard deviation of the dependent variable. All the t-statistics are adjusted for heteroskedasticity and autocorrelation using Newey and West (1987) standard errors. In the context of NFP, a 1-unit change is about 74,000 jobs. The first regression shows that indeed Wall Street analysts' predictions are significantly (at 0.001 level) related to the final NFP number. The second regression shows that adding Overall employment related searches improves the prediction by increasing the R sq. by 4% with both Wall Street analysts and Overall employment related searches being statistically significant. However, we can see that analysts have more impact (0.45 units or 33,000 jobs) on the Final NFP prediction vs. search data (-0.2 units or 14,500 jobs). The results in these and the following pairs of regressions are very similar when comparing adjusted R. sq that accounts for the increase in the number of predictors.

The third and fourth regressions show that when the noisy Initial NFP number is released, the R sq. jumps from 24% to 68%, and interestingly while the Wall Street analyst information is completely absorbed and not incrementally useful

[25] Granger analysis of Search Data vs. NFP. We have only 72 monthly data points which does not have enough power for serious Granger tests. We used weekly search data that increases the number of search data points, which while not ideal (since government data are not weekly), did reveal some interesting patterns. Broadly, search data are useful in nowcasting or predicting Final NFP, Initial NFP and Wall Street economists' consensus but these data are not useful in predicting the signals based on search. We also find that Wall Street economists' consensus is helpful in predicting the Initial NFP numbers but not the Final NFP Number. Interestingly, it is the employment related searches in the *second week* of the month under consideration - which typically happens to be the reference week for NFP- that has the most power. These second week searches are helpful in predicting both the economists survey and the Final NFP numbers but not the other way around.

Table 1.1.7: Regressions to nowcast Final NFP numbers typically available 2 to 3 months after the relevant month the reference using various predictors. All coefficients $\beta(i)$ standardized as $\beta(i)$*sd(X(i)/sd(Y) to enable easy comparisons across formulations. The time subscripts indicate when the data become available. Final NFP is marked as Final NFP$_{t+3}$ because the Final NFP data typically become available 2 to 3 months after the reference month$_t$ and search data used are from the same reference month$_t$ as are the Wall Street economists survey$_t$. Note that previous lag of NFP is expressed at NFP$_t$ since it becomes available earlier in the same month. Robust Newey-west t statistics that account for serial correlation and heteroskedasticity are shown in parenthesis.

Nowcasting Final NFP_{t+3} Regressions

Initial NFP_{t+1}	Intial NFP_t (lagged)	Overall Employment Search$_t$	Sector-Specific $Search_t$	Economists $Survey_t$	Rsq	N
				0.45*** (+4.63)	0.20	73
		-0.20* (-2.15)		0.45*** (+4.63)	0.24	73
0.81*** (+9.47)				0.01 (+0.1)	0.66	73
0.79*** (+9.74)		-0.14* (-2.05)		0.02 (+0.19)	0.68	73
0.80*** (+9.44)	-0.08 (-0.97)			0.04 (0.45)	0.66	73
0.79*** (+9.69)	-0.07 (-1.11)	-0.14 (-2.38)		0.05 (0.66)	0.69	73
0.78*** (+9.55)			-0.22 (-3.89)	0.05 (0.68)	0.71	73

anymore, the search information is still relevant with only a marginal decrease in magnitude from -0.20 to -0.14. Upon adding Overall employment search data, the R sq. increasing by about 2%.

In the fifth and sixth regressions, we add the prior lags of NFP to account for the trending nature of the macroeconomic data. We observe a similar pattern of Wall Street analyst information being no longer useful after the initial NFP number comes out but the Overall employment search data retaining its statistical significance (-2.38) and economic (-0.14 units) significance. Again, adding the Overall employment search data increases the R sq. by 3% which suggests that the information in search data is not fully captured by the previous macroeconomic trends, Initial NFP numbers or Wall Street Analyst consensus, and hence potentially different.

We also check the impact of focusing on the portions of employment web search where government data are not as good by using the architecture and engineering employment related searches and find that those are highly significant (-0.22 units) with a t statistic of (-3.89) that is meaningful at the 1% level. Also, the R sq. increases by 5% to 71% as compared to the previous model that uses only Wall Street analysts, lagged macroeconomic trends and the noisy Initial NFP release. This result suggests that web search for cyclical sectors such as architecture might contain useful information for the entire economy not captured by the government.

1.1.5.7 Robustness

We perform an out of sample forecast of Final NFP by keeping 30% of the sample as hold out and perform two separate next month forecasts - the first with the Initial NFP release and Wall Street analyst consensus and the second by adding *Overall* employment related search data to the first regression to check the impact of web search data. We find that MSE decreases by 5.8% in the test sample and a similar 3.8% in the holdout set. Interestingly, results improve out of sample - with the MSE decreasing by 11.6% in sample and 21.2% out of sample when we use Architecture and Engineering employment searches instead of the Overall employment related searches in the model. Architecture and Engineering, (as explained in the previous section), represents an urban-centric, cyclical, white-collar service category to which search data appear better suited to estimating than the BLS data. While some of the 21.2% decrease in the holdout MSE could be attributed to a small sample size, the results encourage us towards doing more research on curated sectors.

In alternate regressions not included here for space reasons, we formulate the dependent variable directly as *Revisions = Final NFP - Initial NFP* and obtain very similar results. The variable *Revisions* can be robustly predicted by the *Overall search* variable after controlling for its own prior lags and Wall Street expert consensus. We also perform regressions by using the median instead of the mean and find similar and slightly better results potentially because medians are more stable for noisier search data.

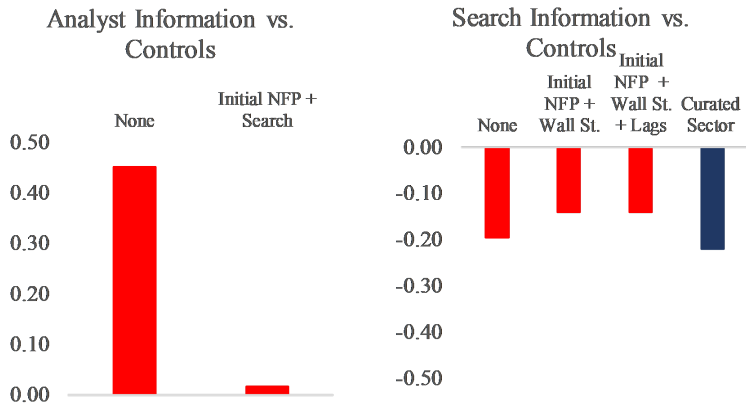

Figure 1.1.20: Generic employment related web search information (right) remains relevant after controlling for Initial NFP and Wall Street analysts' information. But Wall Street analyst information (left) that disappears after Initial NFP is released. Curated web searches (in blue) for sectors (such as Architecture and Engineering) with imbalance of information demand and supply are more predictive than generic employment web searches.

Occasionally, using particular subsets of alternative data can potentially misrepresent how powerful the overall data are. For instance, Altenburger and Ho (2019) find that the reported results of Yelp reviews outperforming food safety inspections fall apart upon comprehensive analysis. When the entire sample of restaurants was used instead of 'Extremely Imbalanced Sampling' of restaurants with no violations or very high violations and the problem changed from a classification into those extreme categories to that of score prediction for the entire sample the results were far less impressive. To avoid such errors, this case study already uses a regression setup that analyzes *all* NFP observations, not just the ones at turning points. However, we obtain significant results with the overall employment searches and not just with one of the sectors. While, sectors where the information demand via search and supply via government surveys is the highest such as architecture yield more significant results, reassuringly our results are similar across all sixteen categories.[26]

[26] Models based on these insights were put in production and the results for NFP improved MSE by about 10% over 4 years which is broadly in line with the reported theoretical performance. Another conceptual and methodological verification was to use similar models to predict other volatile and crucial macroeconomic metrics such as retail sales for which the models outperformed Wall Street analysts by about 20% on the downturns and by about 5% on the upturns.

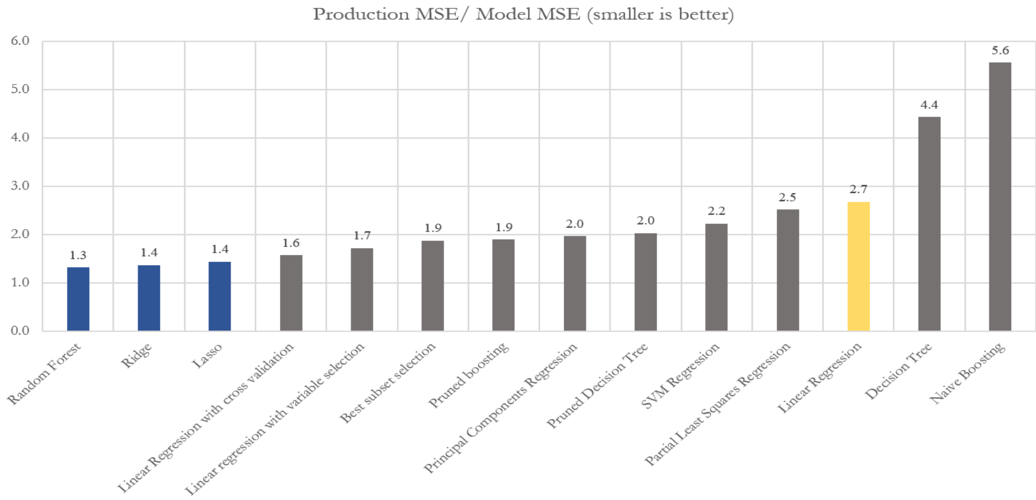

Figure 1.1.21: A MSE based comparison of various simple ML techniques used to build an employment nowcast using the search NFP features in the case study. Each vertical bar in the figure is the ratio of MSE in production to the MSE of the model. The production or out of sample time period was about 24 months.

1.1.5.8 Machine Learning for NFP Revisions

Thus far, the analysis has tested the ex-ante hypotheses about *orthogonal, more accurate, and more timely* information in the search-based features. We find that NFP revisions do contain news or information that is released with up to three months lag, and a powerful and intuitive relationship exists between employment search and NFP that can help meaningfully improve data timeliness and accuracy. We did not find a substantial age and geographic bias in the data, but the sector-specific results suggest search data do measure tech-centric and white-collar firms even better. How much of the difference comes from the under-reporting of sectors such as architecture by the government data and how much from the urban and technology bias of the alternative data will need more investigation.

Having verified the hypotheses, the next stage is to use various machine learning techniques that minimize MSE to effectively incorporate these search-based features into an employment nowcast. Generally, multiple alternative data sources are also combined at this stage. As figure 1.1.21 shows, for nowcasting NFP from web search, we compared various machine learning techniques to the traditional linear regression. Each vertical bar in the figure is the ratio of MSE in production to the MSE of the model and hence a test of overfitting. The actual MSE also follow the same pattern and act as a test of model fit. Random Forest, ridge and lasso cluster together with the best performance of out of sample MSE being 130% to 140 % of the in-sample model fit. SVM, partial least squares, decision trees, boosting and linear regression had out of sample MSE that ranged from 220% to more than 400%. Adding cross-validation and stepwise variable selection improves linear regression performance but for this particular case, it still could not out perform the best machine learning techniques. This finding of higher effectiveness of random forest and elastic net as compared to the traditional linear regression techniques are consistent with literature cited previously in the chapter, as well as the broader machine learning field work such as Caruana and Niculescu-Mizil (2006).

Search and most other social data like twitter and reddit have "spiky" distributions - meaning they do not obey normality and can have a non-zero skew as well as high kurtosis. This can make asymptotic inference tricky and bootstrapping and its many variants should be used to provide a more realistic idea of the distribution. Relatedly, these data almost always have extreme observations which one must decide whether to exclude ("clean") or include and deal with as an "influential point" - this is best done in close consultation with a domain expert and can have a dramatic impact on the model. Ensembling or combining various internally consistent models built on different philosophies that are naturally be stronger in different parts of the parameter space is a useful common practice that also increases robustness.

1.1.6 Conclusion and Discussion

This chapter presented three fundamental problems with the current macroeconomic data - high dimensionality, a staggered release schedule, and poor data quality. It provided an overview of nowcasting techniques that are used to address the dimensionality and the staggered release schedule problems. Traditionally, nowcasting techniques were based on dynamic factor models and state-space frameworks, but more recently have started incorporating machine learning techniques. The arrival of alternative or Big Data with the potential for addressing the data quality problem has made macro nowcasting an exciting field. But alternative data availability is a double-edged sword - just as it promises to improve data quality, it simultaneously worsens the dimensionality problem by dramatically increasing the potential feature set. To address this problem, we discussed the need for an ex-ante hypothesis to guide alternative data selection, provided some sample hypotheses for popular alternative data sources, and noted the typical data quality dimensions such as improving the timeliness and increasing granularity, where alternative data are having success.

The chapter culminated in a case study that utilized a novel dataset of 830 million employment-related searches to nowcast NFP revisions three months ahead of government data with statistical (t statistic of -2.4) and economic (-0.14 standardized unit impact) significance - a result in contrast to prior literature. The case study also established that NFP revisions contain relevant information or "news" per the framework in Mankiw et al. (1984), and controlled for the information content of Wall Street analyst expectations as well as prior lags of NFP. After verifying the improved data quality hypotheses about the increased timeliness and accuracy dimensions, various machine learning models were fit to maximize predictive power. Random forest and elastic net models performed best, generating only half the MSE of linear regression in both the in-sample model fit and out-of-sample tests.

I suggest three related but distinct strands of development for the field - the first an increase in the number and sophistication of statistical learning techniques being used, the second dealing better with the more complex workflow for the insight generation process involving the data engineer, data scientist and domain expert, and the third an increased focus on privacy concerns and an increased investment in security. The first - machine learning techniques - concern specialists more, the more complex workflow concerns the entire firm, and the third - privacy and security - concerns all of us.

The number of machine learning techniques is too vast to comment specifically, but deep learning and other computationally intensive, feature-rich, and non-linear techniques will find more success when applied towards *"feature generation"* - say via NLP of accounting statementsor analysis of pictures and videos - where the number of features (p) vs. the number of observations (N or T) problem is not primary. That said, broader use of machine learning methodologies has great potential ranging from cross-validation for picking the optimal number of parameters, building the next generation of non-linear regime-switching models, or generating better economic activity clusters.

Moving from the concerns of the machine learning specialist to those of the entire firm, the more complex workflow problem that now has to account for everything from data selection and purchase frameworks to technical aspects like data pipelines and visualization is non-trivial in the organizational and technical aspects. The different strengths and weaknesses of the players (Jain (2019); Krishnamachari (2017)) can create organizational challenges - for instance, a data scientist may find it hard to write production-level code just as data engineers may find it hard to clean the data without the scientist or the domain expert. Making fast, relevant and efficient machine learning pipelines is a technical problem, but the tremendous activity in open-source tools has been helpful.[27]

And finally, as a society - the emerging privacy and ethical concerns are a big question for all of us.[28] Acquisti et al. (2016) provide a good overview of the economic value and the consequences of privacy and personal information. Kerry et al. (2020) highlight the gap between existing laws and the data reality of today where most people simply do not and cannot read the various privacy policies. It is possible to "re-identify" individuals with incomplete datasets - for instance Rocher et al. (2019) estimate that 99.98% of the individuals can be re-identified with any dataset containing 15 demographic attributes, which is concerning and might make one argue for very strict laws and rules against data collection. On

[27] Krishnamachari (2017) provides an overview of the industry-standard open-source tools useful for asset management.

[28] Nature magazine penned an editorial about it in July 2019

the other hand the COVID-19 pandemic showed us that having *some aggregate information*, for instance at the zip code level [29] could be helpful for policymakers and society. Finding the right balance where we violate the least privacy and provide the most social benefit is our task as citizens. [30]

Bibliography

Acquisti, A., Taylor, C., and Wagman, L. (2016). The economics of privacy. *Journal of economic Literature*, 54(2):442–92.

Adler, N., Cattuto, C., Kalimeri, K., Paolotti, D., Tizzoni, M., Verhulst, S., Yom-Tov, E., and Young, A. (2019). How search engine data enhance the understanding of determinants of suicide in india and inform prevention: observational study. *Journal of medical internet research*, 21(1):e10179.

Altenburger, K. M. and Ho, D. (2019). Is yelp actually cleaning up the restaurant industry? a re-analysis on the relative usefulness of consumer reviews. In *The Web Conference (WWW)*, pages 2543–2550.

Antenucci, D., Cafarella, M., Levenstein, M. C., Re, C., and Shapiro, M. D. (2014). "using social media to measure labor market flows.".

Aruoba, S. B. (2008). Data revisions are not well behaved. *Journal of money, credit and banking*, 40(2-3):319–340.

Askitas, N., Zimmermann, K., and Askitas, N. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly (formerly: Konjunkturpolitik)*, 55:107–120.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.

Baumol, B. (2013). *The Secrets of Economic Indicators:Hidden Clues to Future Economic Trends and Investment Opportunities*. Pearson Education, Inc.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2017). Macroeconomic nowcasting and forecasting with big data (2017-11-01). frb of ny staff report no. 830.

Brave, S., Cole, R., and Kelley, D. (2019). A "big data" view of the u.s. economy: Introducing the brave-butters-kelley indexes. *Chicago Fed Letter*.

[29] Chetty et al. (2020) show, among other things, the differential impact of COVID-19 in various populations using anonymized data in a live updateble tool that was widely reported by newspapers and utilized by academics.

[30] Jain and Seeman (2019) take the illustrative example of a fictional asset manager who invests money to obtain "extra features" that do not just violate privacy but provide no marginal predictive benefits. They suggest including potential privacy costs in assessing a dataset for purchase.

Brave, S. A., Butters, R. A., and Fogarty, M. (2020). The perils of working with big data and a small framework you can use to avoid them.

Bräuning, F. and Koopman, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, 30(3):572–584.

Burns, A. F. and Mitchell, W. C. (1946). *Measuring business cycles*. Number burn46-1. National bureau of economic research.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

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

Chancellor, S. and Counts, S. (2018). Measuring employment demand using internet search data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 122:1–122:14, New York, NY, USA. ACM.

Chen, J. C., Dunn, A., Hood, K., Driessen, A., and Batch, A. (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., and Team, T. O. I. (2020). *The economic impacts of COVID-19: Evidence from a new public database built using private sector data*. Number w27431. National Bureau of Economic Research.

Choi, H. (2009). Predicting initial claims for unemployment benefits.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1):2–9.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Clement, B. and Stephanie, C. (2015). Contribution from google trends for forecasting the short-term economic outlook in france: limited avenues. article. Département de la Conjoncture.

Concha Artola, Pedraza, P. and Hernández, F. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower*, 36.

Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting?

Coyle, D. (2014). *GDP: A Brief but Affectionate History*. Princeton University Press.

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

Croushore, D. D. (2011b). Real-time forecasting.

DÁmuri, F. and Marcucci, J. (2010). 'google it!' forecasting the us unemployment rate with a google job search index. *MPRA Paper, No. 18248*.

Datar, S., Jain, A., Wang, C. C., and Zhang, S. (2020). Is accounting useful for forecasting gdp growth? a machine learning perspective. *A Machine Learning Perspective (December 1, 2020)*.

Diebold, F. (2000). Big data" dynamic factor models for macroeconomic measurement and forecasting.

Döpke, J., Fritsche, U., and Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting*, 33(4):745–759.

Dumais, S., Jeffries, R., Russell, D., Tang, D., and Teevan, J. (2014). *Understanding User Behavior Through Log Data and Analysis*, pages 349–372.

Einav, L. and Levin, J. (2013). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14.

F. Cavallo, A. and Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research †. *Journal of Economic Perspectives*, 30:151–178.

Fama, E. F. (1970). "efficient capital markets: A review of theory and empirical work". *The Journal of Finance*, 25(2):383–417.

Fed, C. (2020). Background on the chicago fed national activity index.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Goel, S., M Hofman, J., Lahaie, S., M Pennock, D., and 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.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Hall, A. S. et al. (2018). Machine learning approaches to macroeconomic forecasting. *The Federal Reserve Bank of Kansas City Economic Review*, 103(63):2.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hellerstein, R. and Middlethorp, M. (2012). Forecasting with internet search data. Liberty Street Economics.

Jain, A. (2019). *Macro Forecasting Using Alternative Data*, chapter 11. Elsevier.

Jain, A. and Seeman, N. (2019). Privacy vs. alpha: A conversation. *Alpha: A Conversation (October 24, 2019)*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jansen, B. J. and Schuster, S. (2011). Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research*, 12(1):1.

Kässi, O. and Lehdonvirta, V. (2018). Online labour index: Measuring the online gig economy for policy and research. *Technological forecasting and social change*, 137:241–248.

Kerry, C. F., Morris Jr, J. B., Chin, C. T., and Lee, N. E. T. (2020). Bridging the gaps. *Governance Studies at Brookings*.

Keynes, J. M. (1936). *Keynes, The General Theory of Employment, Interest and Money*, volume 7.

Kliesen, K. L. (2014). A guide to tracking the u.s. economy. article. Federal Reserve Bank of St. Louis Review. https://files.stlouisfed.org/files/htdocs/publications/review/2014/q1/kliesen.pdf.

Krishnamachari, R. T. (2017). Big data and ai strategies.

Lin, A. Y., Counts, S., and Cranshaw, J. (2019). Forecasting u.s. domestic migration using internet search queries. In *The Web Conference (WWW)*. IW3C2.

Lo, A. W. and Mueller, M. T. (2010). Warning: physics envy may be hazardous to your wealth! *arXiv preprint arXiv:1003.2688*.

Lopez de Prado, M. and Lipton, A. (2020). Three quant lessons from covid-19 (presentation slides). *Available at SSRN 3562025*.

Mankiw, N. G., Runkle, D. E., and Shapiro, M. D. (1984). Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary economics*, 14(1):15–27.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Mislove, A., Lehmann, S., yeol Ahn, Y., pekka Onnela, J., and Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *In Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Moshiri, S. and Cameron, N. (2000). Neural network versus econometric models in forecasting inflation. *Journal of Forecasting - J FORECASTING*, 19:201–217.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–08.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

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

Orphanides, A. and Williams, J. C. (2007). Robust monetary policy with imperfect knowledge. *Journal of Monetary Economics*, 54(5):1406–1435.

Phipps, P. and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics - ANN APPL STAT*, 6.

Popper, K. R. (1961). The poverty of historicism new york.

Richardson, A., van Florenstein Mulder, T., and Vehbi, T. (2021). Nowcasting gdp using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2):941–948.

Rocher, L., Hendrickx, J. M., and De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9.

Rothbaum, J. and Bee, A. (2020). Coronavirus infects surveys, too: Nonresponse bias during the pandemic in the cps asec. *Social, Economic, and Housing Statistics Division Working Paper*, (2020-10).

Sargent, T. and Sims, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis.

Sermpinis, G., Stasinakis, C., Theofilatos, K., and Karathanasopoulos, A. (2014). Inflation and unemployment forecasting with genetic support vector regression. *Journal of Forecasting*, 33(6):471–487.

Smalter Hall, A. and Cook, T. R. (2017). Macroeconomic indicator forecasting with deep neural networks. *Federal Reserve Bank of Kansas City Working Paper*, (17-11).

Smith, A. (2015). Searching for work in the digital era.

Soros, G. (2009). *The crash of 2008 and what it means: The new paradigm for financial markets*. PublicAffairs.

Stock, J. and Watson, M. (1999). Forecasting in action. *Journal of Monetary Economics*, 44:293–335.

Stock, J. and Watson, M. (2016). Chapter 8 - dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. volume 2 of *Handbook of Macroeconomics*, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, 4:351–394.

Suhoy, T. (2009). "query indices and a 2008 downturn: Israeli data". *Research Department, Bank of Israel. http://www.boi.org.il. Discussion Paper Np. 2009.06*.

Venkatramanan, S., Sadilek, A., Fadikar, A., Barrett, C. L., Biggerstaff, M., Chen, J., Dotiwalla, X., Eastham, P., Gipson, B., Higdon, D., et al. (2021). Forecasting influenza activity using machine-learned mobility map. *Nature communications*, 12(1):1–12.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810.

Yousuf, K. and Ng, S. (2021). Boosting high dimensional predictive regressions with time varying parameters. *Journal of Econometrics*, 224(1):60–87.