CHAPTER 1

Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data

Serena Ng

This paper seeks to better understand what makes big data analysis different, what we can and cannot do with existing econometric tools, and what issues need to be dealt with in order to work with the data efficiently. As a case study, I set out to extract any business cycle information that might exist in four terabytes of weekly scanner data. The main challenge is to handle the volume, variety, and characteristics of the data within the constraints of our computing environment. Scalable and efficient algorithms are available to ease the computation burden, but they often have unknown statistical properties and are not designed for the purpose of efficient estimation or optimal inference. As well, economic data have unique characteristics that generic algorithms may not accommodate. There is a need for computationally efficient econometric methods as big data is likely here to stay.

1 INTRODUCTION

The goal of a researcher is often to extract signals from the data, and without data, no theory can be validated or falsified. Fortunately, we live in a digital age that has an abundance of data. According to the website Wikibon (www.wikibon.org), there are some 2.7 zetabytes of data in the digital universe.[1] The US Library of Congress collected 235 terabytes of data as of 2011. Facebook alone stores and analyzes over 30 petabytes of user-generated data. Google processed 20 petabytes of data daily back in 2008, and undoubtedly

Financial support from the National Science Foundation (SES-0962431) is gratefully acknowledged. I would like to thank David Weinstein for introducing me to this work and Jessie Handbury for getting me started. I also thank Jushan Bai, Christos Boutsidis, Jean-Jacques Forneron, Matt Shum, and Juan Ospina for helpful discussions. This work would not be possible without the contribution of Rishab Guha and Evan Munro. I am deeply indebted to their help. All errors are mine.

[1] 1024 Megabytes = 1 Gigabyte, 1024 Gigabytes = 1 Terabyte, 1024 Terabytes = 1 Petabyte, 1024 Petabytes = 1 Exabyte, and 1024 Exabytes = 1 Zetabyte.

much more are being processed now. Walmart handles more than one million customer transactions per hour. Data from financial markets are available at ticks of a second. We now have biometrics data on finger prints, handwriting, medical images, and last but not least, genes. The 1000 Genomes project stored 464 terabytes of data in 2013 and the size of the database is still growing.[2] Even if these numbers are a bit off, there is lot of information out there to be mined. The data can potentially lead economists to a better understanding of consumer and firm behavior, as well as the design and functioning of markets. The data can also potentially improve the monitoring of traffic control, climatic change, terror threats, causes and treatment of health conditions. It is not surprising that many businesses and academics are in a big data frenzy. The Obama Administration announced the Big Data Research and Development Initiative in 2012.[3] The National Bureau of Economic Research offered lectures on big data in two of the last three summer institutes. Courses on big data analysis often have long waiting lists.

Many economists have written about the potential uses of big data. New overview articles seem to appear in RePEc every month. Some concentrate on the economic issues that can be studied with the data, as in the excellent articles by Einav and Levin (2013, 2014), Athey (2013). Other surveys take a more statistical perspective. For example, Varian (2014) considers machine learning tools that are increasingly popular in predictive modeling. Fan et al. (2014) warn about the possibility of spurious correlation, incidental endogeneity, and noise accumulation that come with big data and suggests new methods to handle these challenges. While the use of big data in predictive analysis has drawn the most attention, much of economic analysis is about making causal statements. Belloni et al. (2014) discuss how regularization can permit quality inference about model parameters in high-dimensional models. Athey and Imbens (2015) use machine-learning methods to estimate treatment effects that may differ across subsets of the population.

As with these studies, I also consider methods specific to big data. But instead of predictive modeling and taking the data as given, I focus on data preprocessing, perhaps the most time-consuming step of a big data analysis. This paper was initially motivated by the curiosity to learn what makes big data analysis different, how far can our existing econometric tools take us, and getting a sense of what issues need to be addressed if big data are here to stay, using as case study four terabytes of weekly retail sales data collected between 2006 and 2010. A distinctive feature of the dataset is that it has direct measures of prices and quantities. I use the opportunity to analyze the cyclical aspects of the quantity data. This is interesting because the Great Recession of 2008 is in this sample, and official consumption data do not come at higher than a monthly frequency. The project gives me a better understanding of the

[2] The project seeks to find most genetic variants that have frequencies of at least 1 percent in the population.

[3] See www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

limitations of statistics/econometrics in big data analysis, and why methods in the domain of data science are useful.

A gigabyte of data can be easily analyzed on our desktop computers using our favorite statistical software packages. The problem is that methods which we understand and work well with small datasets may not be big data friendly or scalable. Even though I have four terabytes of data, it is impossible to analyze them all at once. The memory requirement is beyond the capacity of our computers even with unlimited financial resources. Aggregation, whether in the time, spatial, or product dimension, would seem to take away features that make the data special. Fortunately, even if we could analyze all the data, it might not be necessary to do so. Studying a subset of the data might suffice, provided that the subset is appropriately assembled. Hence the first part of this paper explores two random subsampling algorithms developed by computer scientists to accurately approximate large matrices. Random subsampling is neither efficient nor necessary when the sample size is manageable. In a big data setting, random sampling not only speeds up the analysis; it is a way to overcome the constraints imposed by the computing environment. However, the subspace-sampling algorithms considered are developed to run fast and have desirable "worst case error bound," quite distinct from the optimality criteria such as mean-squared error and consistency that we typically use. There is thus a need to evaluate these algorithms in terms of quantities that we analyze. This is difficult when the probability structure of the data is not specified.

Business cycle analyses typically use data collected by government agencies that also handle the data irregularities. With the Nielsen data, the task of removing seasonal effects is left to the user. The challenge is that weekly seasonal variations are not exactly periodic. Structural modeling on a series by series basis may deliver a filtered series that is statistically optimal, but this is impractical when we have millions if not billions of highly heterogeneous series to analyze. Hence the second part of this paper explores a practical approach to modeling the seasonal effects and have mixed success. I find that removing the seasonal effects at individual level is no guarantee that the seasonal variations at the aggregate level will be removed. The exercise does, however, suggest promising ways of handling seasonality that need to be further explored.

More generally, the volume, heterogeneity, and high sampling frequency that generate excitement about the data are precisely what make extracting signals from the data difficult. Big data creates a need for econometric methods that are easy to use, computationally fast, and can be applied to data with diverse features. Accomplishing these objectives may entail a change from the current practice of writing down models to fit a narrow class of data, to writing down models that would fit a variety of data types even if they may not be optimal for any particular data type. The difference is a bit like shopping at a general merchandise store versus a specialty store; there is a trade-off between quality and convenience. The non-probabilistic methods developed by data scientists enable efficient computations, but they are not developed

with estimation and inference in mind. It is an open question whether computational efficiency and statistical efficiency are compatible goals. It is also debatable if precision of estimates obtained in a data rich environment can be judged the same way as when the sample size is small. Understanding the statistical underpinnings of computationally efficient methods can go a long way in easing the transition to big data modeling. This can be important as big data are likely here to stay.

2 DATA ANALYSIS IN THE DIGITAL AGE

This section has two parts. Subsection 1 draws attention to the challenges that big data pose for traditional statistical modeling that is also the foundation of econometrics. Subsection 2 highlights some characteristics of big data and summarizes features of the Nielsen scanner data used in the analysis of Section 3 and 4.

2.1 Data Science and Statistics

A lot has been written about "big data" in recent years, but not everyone has the same notion of what big data is. Labor and health economists have long been analyzing big surveys, census data and administrative records such as maintained by the Social Security Administration, Medicare and Medicaid. Increasingly, macroeconomists also turn to big data to study the price determination process, sometimes involving unpublished data. But once access to the data is granted, analysis of these pre-Google big data can proceed using existing hardware and software packages like STATA and MATLAB.

The post-Google data that concern this study are the large and complex datasets that are not collected through surveys, not supported by official agencies, and cannot be stored or analyzed without switching to a new computing environment at some point. If 8 bytes (64 bits) are used to store a real number, a few billion of observations for several variables would be beyond the capacity of most off-the-shelf desktop computers. What makes big data analysis different is not just that the sheer size of the dataset makes number crunching computationally challenging,[4] but also that the observations are often irregularly spaced and unstructured. Indeed, it is quite common to use three-Vs to characterize big data: large in **V**olume, come in a **V**ariety of sources and formats, and arrive at a fast **V**elocity. Some add variability and veracity to the list because the data are sometimes inconsistent in some dimensions and possibly inaccurate. Conventional methods designed to process data with rectangular structures often do not apply. There is no statistical agency to oversee confidentiality and integrity of the data, and the tasks of cleaning and handling

[4] A computational decision problem is NP hard if it is at least as hard as the hardest problems in class NP (whose solutions can be verified in polynomial time). Intuitively, an NP-hard problem is one that admits no general computational solution that is significantly faster than a brute-force search.

the data are in the hands of researchers, many of whom have limited knowledge about database management. PYTHON and R seem to be commonly used to prepare the data for analysis but often, programs written for one dataset are of little use for another because each dataset typically has its own quirky features.

Each of the three Vs pose interesting challenges for statistical analysis because they violate assumptions underlying methods developed for conventional data. Because of variety, it may be difficult to justify a common data generating process. Because of volume, thinking about how to conduct optimal estimation and inference is not realistic when we struggle just to find ways to summarize the massive amount of information. It would also not be useful to have complex models that cannot be estimated, or MCMC methods that cannot be completed within a reasonable time frame. Bayesian estimation would essentially be likelihood-based when sample information dominates the prior. Because of velocity and volume, the standard error of estimates will be tiny. But because the noise level in big data can be high, the assumption that information increases with sample size may be questionable, an issue noted in Granger (1988). A new way of doing asymptotic analysis may well be warranted.

A big data project typically uses methods that are part statistics, part computer science, and part mathematics, and is often associated with the field of *data science*. Cleveland (2001) proposes to expand the areas of technical work in statistics and to call the new field "data science." Wikipedia defines the field as "extraction of knowledge or insights from large volumes of data," thereby directly linking data science with big data. Another characterization is well summarized by how *The Journal of Data Science* defines its scope: "everything to do with data: collecting, analyzing, modeling, . . . , yet the most important part is its application." The emphasis here is the ability to apply what is learned from the data analysis to practical use, such as business analytics and predictions. In a sense, this view treats data analysis as an immediate input of production; what ultimately matters is the value of the final good.

In an influential paper, Brieman (2001) distinguishes data science from traditional statistical analysis as follows. A statistician assumes a model, or a data generating process, to make sense of the data. Econometric analysis largely follows this stochastic model paradigm. The theoretical results are not always well communicated to practitioners and not always taken to the next level after publication of the article. Brieman (2001) argues that the commitment to stochastic models has handicapped statisticians from addressing problems of interest and encourage the adoption of a more diverse set of tools. A data scientist accepts the possibility that the assumptions underlying models may not be correct. He/she therefore uses algorithms, or machine-learning methods, to map data to objects of interest, leaving unspecified the data generating process that nature assigns. Probability models and likelihoods are replaced by random forests, regression trees, boosting, and support vector machines. One looks for clusters and frequent items in the data. The work of a data scientist often has immediate downstream uses (such as for business decisions or in gene mapping).

Big data provides a momentum boost to move away from stochastic modeling as the more data with the three V features we have, the more difficult it is to defend a model that is generally valid. The American Statistical Association (ASA) has a working group to study the future direction of the discipline at large. The group sees collaboration with data scientists as a way for statisticians to contribute to exciting problems of the digital generation.[5] The Institute of Mathematical Statistics also recognizes the challenge that data science poses. In her 2014 presidential address, Bin Yu remarked that data science represents an inevitable (re-)merging of computational and statistical thinking. She suggests calling themselves (i.e., statisticians) data scientists in order to fortify their position in the new era of data analysis, echoing the suggestion the statistician Jeff Wu made at an inagural lecture at the University of Michigan in 1997.[6]

While statisticians are open to the idea that computer science and mathematics will play an important role in the future, economists are slower to react. Most of us have little experience with big data and know little about the computational aspect of data analysis. As will be discussed in Section 3, we may well have to become active in this area of research as we are increasingly presented with opportunities to analyze big economic data, and see that there are data issues that require our knowledge and input.

2.2 Data Types

Most post-Google big datasets are not intentionally collected, hence they are cheap to produce compared to data produced by surveys. The big datasets used in economic analysis are usually in one of two forms. The first is generated by search engines and social media websites such as Google, Facebook, and Twitter. It is no secret that online clicks have been used to target products to potential buyers. Social media data are now more effective than data from loyalty programs in predicting repeated purchases. But web search data have many uses other than advertising, the most famous of which is probably the initial success of prediction of flu outbreaks by Ginsberg et al. (2009). A creative use of social media data is the U-report, a UNICEF project that collects text-messages from young people in Uganda. IBM researchers were able to apply machine learning methods to the tweets to learn about economic, political, and health conditions, and to alert health officials of Ebola outbreaks.[7] Projects of this type are now expanded to other parts of Africa.

[5] See www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf.

[6] See also http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics/, http://magazine.amstat.org/blog/2010/09/01/statrevolution/, http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf, for differences between the two fields. http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/

[7] www.research.ibm.com/articles/textual-analysis-u-report.shtml.

A second type of data comes from web searches. Such data provide information about intent and potential actions, hence can be useful for prediction. Choi and Varian (2012) find that a small number of Google search queries can "nowcast" car sales and shows how proxies for consumer confidence can be constructed from Google Trends data. Preis et al. (2013) compute a Future Orientation index and finds a correlation between online searches and realized economic outcomes.[8] Koop and Onorante (2013) use Google search data to improve short-term forecasts. Antenucci et al. (2014) use Twitter data to produce an index that can predict job loss.

A different type of big data is action-based, arising from the real-time purchases at stores such as Walmart, and from charges processed by, for example, Mastercard. These databases are relatively structured and often have a business value. As an example, Target was reported to form prediction indicators from buying habits of customers going through life changing events, such as divorce and giving birth, and push to them promotional flyers.[9] Based on Matercard transactions, SpendingPulse^{TM} claimed that its near-real-time purchase data can predict spending weeks if not months ahead of other sources.

Data on prices are of particular interest to economists. The Billion Prices project gives real time inflation predictions by aggregating information in five million items sold by about 300 online retailers around the world. Handbury et al. (2013) use a Japanese dataset with five billion observations on price and quantity to construct an ideal (Tornqvist) price index. The authors report a nontrivial difference between their measure and the official measure of inflation. This type of data is valuable when credibility of the official data is in question, as in the case of inflation in Argentina. It is also useful when release of the data is disrupted by unanticipated circumstances, such as in the case of earthquakes in Chile and Japan, see Cavallo (2012) and Cavallo et al. (2013).

2.3 The Nielsen Data

The dataset that motivates this analysis is the Retail Scanner Data collected weekly by the Nielsen marketing group. The database is managed by the Kilts Center for Marketing at the University of Chicago. Through a university license, the data are made available for analysis a couple of years after the actual transactions. The data are collected at 35,000 participating grocery and drug stores, and mass merchandisers affiliated with about 90 participating retail chains across 55 MSA (metropolitan statistical area) in the US. Our data are from 2006 to 2010. The dataset covers three million unique UPC (universal product code) for 1073 products in 106 product groups which are in turn classified into ten categories: *dry groceries, frozen, dairy, deli, meat,*

[8] The Future orientation index is the ratio of the volume of searchers of the future (i.e., 2011) to the past (i.e., 2010).

[9] Tolentino (2013) analyzes loyalty programs, Goel (2014) on Facebook, and Duhigg (2012) on Target.

fresh food, non-food, alcoholic beverage, general merchandise, and *health and beauty*. The data are structured (i.e., in numeric format only, audio and video files are not involved) but highly heterogeneous. There is also information about location (zip and fips county codes) and the retailer code, but retailers are not identified by name. Household-level information is in a companion Nielsen Homescan Consumer Panel database which is not used in this study. The Nielsen data have been widely studied in marketing analysis of specific products.[10]

The variables of interest are price, price multiplier, and number of units sold, from which one can compute the unit price and the week's total dollar sales. Total volume is computed from the quantity of goods in individual packaging, and unit of measure for that quantity. Several features make the data interesting to economists. First, prices and quantities are separately observed. In contrast, conventional price deflators are inferred from observations on value and quantities. Furthermore, these data are recorded at a higher frequency and at more locations than the official data on retail sales. In fact, few economic indicators (on price or quantity) are available at a weekly frequency. Even at a monthly frequency, there is little data available at a local level. However, the Nielsen data also have several drawbacks. The data only cover grocery store purchases and ignore services and durables which tend to be more cyclical. Furthermore, the data are not seasonally adjusted.

An increasing number of researchers are using scanner data to answer interesting economic questions. Broda et al. (2009) conclude from analyzing the Homescan data that the poor actually pays less for food purchases, not more, as poverty analyses based on the CPI suggest. Beraja et al. (2015) construct monthly price indexes to study the impact of local versus aggregate shocks. The indexes are constructed from the bottom up (group by group), keeping memory usage at a manageable level. Coibion et al. (2015) use an IRI database that is similar to the Nielsen data but with fewer products to study the cyclicality of sales. They aggregate the data to monthly frequency and pool the data across markets to run fixed-effect regressions. Cha et al. (2015) aggregate the weekly Homescan data to annual level and find that food consumed at home is countercyclical.

Far fewer studies have looked at the price data at the native (weekly) frequency. Even harder to find are studies that analyze the quantity data. One reason could be that there are not many predictors available at a weekly frequency for a structural demand analysis. Even at a descriptive level, analysis of the quantity data at the weekly level requires separating the cyclical from the seasonal components. I hope to make some progress on this front, given the unique opportunity of having the financial crisis in the sample of 2006–10.

A total of six products will be analyzed: beer, foreign wine, meat, eggs, pet food and baby food. Results for lightbeer and beer, domestic wine and foreign wine are similar and not reported. For a given product, let unit price

[10] Research papers using the data can be found in http://research.chicagobooth.edu/nielsen/research/.

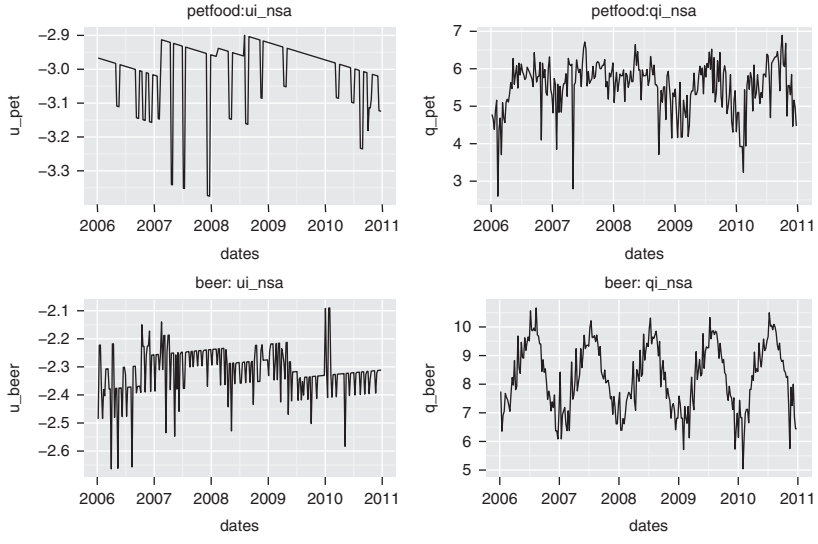

Figure 1 u_{it} and q_{it} (NSA): Pet Food and Beer

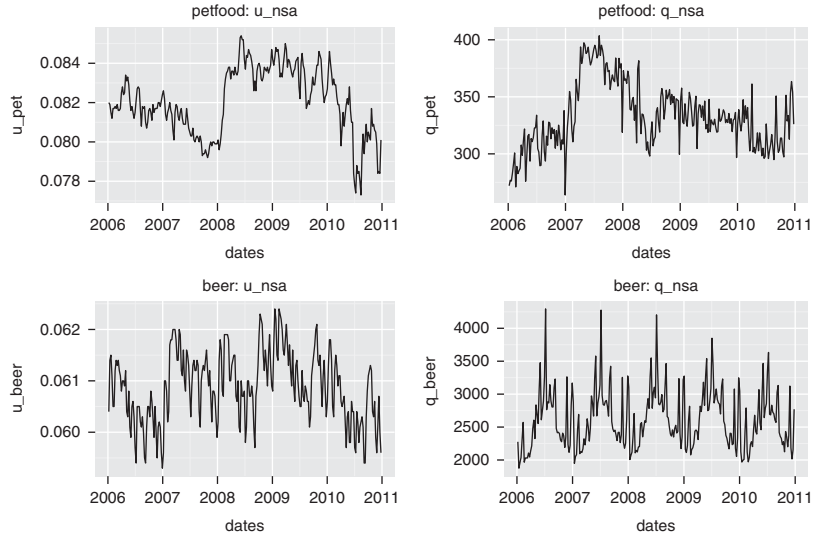

Figure 2 $\bar{u}_t$ and $\bar{q}_t$ (NSA): Pet Food and Beer

at week t be u_{ti} and units sold be q_{ti}, where i is a unique store–UPC pair. For example, Coke Zero and Coca-Cola Light sold at the same store are two different units, as are Coke Zero sold at say, Seven-Eleven and Wawa. To get an idea of features in the data, Figure 1 shows u_{ti} and q_{ti} for one i selected from the pet food and one i selected from the beer group. Figure 2 shows the unweighted mean over all i in the balanced panel (denoted $\bar{u}_t$ and $\bar{q}_t$). The u_{ti}

series for both products are non-smooth, reflecting infrequent price changes. The downward spikes are likely due to discounts. Chevalier et al. (2003) finds evidence of price discounts around seasonal peaks in demand. The seasonal variations in the quantity data for beer are strong at both the individual and aggregate levels.

My goal is to extract the cyclical information in the q_{ti} data. After linearly detrending the data, the first two principal components explain around 15 percent of the variations, suggesting the presence of pervasive variations. In the next two sections, I consider two problems encountered. The first is memory constraint which leads to investigation of random sampling algorithms. The second relates to the goal of extracting the cyclical component, which calls for the need to seasonally adjust the weekly data on a large scale. As we will see, knowledge of tools unfamiliar to economists can go some way in making the analysis more efficient, but many issues remain to be solved.

3 BALANCING AND SKETCHING THE DATA

The time it takes to perform a task on the computer depends not just on how efficiently the program is written, and in what language, but also on the hardware which big data put to a serious challenge. Specifically, the computation speed depends on how frequently the data are brought from physical storage (the hard disk) to RAM, how fast and how much data can be moved from RAM to the CPU, and the queuing time which depends on the amount of the requested RAM. We have almost four terabytes of data and processing them requires a lot of RAM! The original intent was to perform all computations on a cloud server such as Amazon Web Service. Unfortunately, the user agreement restricts data storage to university-owned hardware. It took months to find a feasible and efficient alternative. Eventually, my computing environment consists of a (not very fast) server that allows each job to use up to 256GB of RAM, and a desktop (2011-vintage iMac) upgraded to 24GB of RAM.

To reduce the volume of data in a systematic manner, my student helpers (Rishab Guha, Evan Munro) and I started by constructing a balanced panel for each of the products considered. This is itself a RAM- and time-intensive exercise. We are familiar with MATLAB and somewhat familiar with R but have no prior experience with database management or packages like PANDAS, which we subsequently use. We initially wrote programs in STATA, PYTHON and R for the same task as a way to check bugs but settled on using PYTHON. We experimented with several ways of balancing the panel. The first method keeps only those UPC-stores that are available for every week in the year and then concatenate the five years of data to keep only those UPC-stores that are available for each of the 260 weeks. The second method stacks all 260 weeks of data and selects those store–UPCs with recorded sales in every week. Eventually, we (i) manually sort the data frame by UPC and store code, (ii) loop through the underlying array while keeping track of the number of observations for each unique UPC/store code combination, and (iii) keep only those with 260