Big Data Analysis: The Next Frontier

Nii Ayi Armah, Canadian Economic Analysis

- Current analysis is heavily dependent on data. The more timely, accurate and relevant the data, the better our assessment of the current state of economic activity.
- Technological advancements have provided an opportunity to exploit digital data from business transactions, social media and networked computers. The combination of all of these data is called "big data."
- Analysis of the vast quantities of digital information contained in big data can offer fresh insight for the monitoring of economic activity and inflation. Moreover, the timeliness of big data could improve real-time decision making for monetary policy.
- The potential of big data is, however, limited by challenges related to methodological constraints, a lack of easy access to the data and privacy concerns.

By providing an assessment of the present state of the economy, current analysis[1] contributes to the Bank of Canada's long-term macroeconomic projections, which in turn help to inform monetary policy decisions. Immediate and complete information about every economic and financial transaction within a country would improve current analysis by facilitating accurate and timely measurement of important macroeconomic indicators. Unfortunately, this ideal data set does not exist. The macroeconomic data produced by official statistical agencies are published with a lag and are subject to revision. Gross domestic product (GDP), for example, is a quarterly series that is published with a two-month lag and revised over the next four years. The consumer price index (CPI) is a monthly series that, although not subject to revision, is published three weeks after the end of the reporting month.

These issues with official data have led some researchers to explore the possibility of complementing official data with the use of "non-official" data that may be more timely.[2] An example of the early use of such data for current analysis is Lamont (1997), who finds that counting the frequency of appearances of the word "shortage" in print newspapers can be a good predictor of inflation in the United States. The Bank of Canada also uses non-official data to monitor the economy. For example, the Bank's regional

[1] See Coletti and Kozicki (this issue) for a discussion of the role of current analysis in monetary policy.

[2] There is a trade-off between the timely publication of official data and the accuracy of those data.

offices collect and analyze data obtained from quarterly consultations with businesses across Canada to gather their perspectives on such topics as demand and capacity pressures, as well as their views on economic activity in the future. These data, summarized in the *Business Outlook Survey*, provide a source of timely information that augments views gleaned from official data. With advances in technology, the proliferation of digital data and the declining cost of digital storage, another type of non-official data has recently emerged and is growing quickly—“big data.”

Big data is a large-scale collection of information, some of which, such as business transactions, has always existed in corporate ledgers in the form of daily sales or inventory levels. Rich micro-level administrative data maintained by government agencies have also existed for some time. The high cost of retrieving and organizing all this information en masse has slowed the exploitation of these complementary sources of data for current analysis. However, digitization of previously paper-based data sources has made data much more accessible and easier to organize and analyze. Data emanating from services offered by public institutions or government agencies are a rich source of information on the behaviour of citizens. In addition, the rapid development of computer networking and the Internet has led to new sources of information in social media and web searches, as well as in electronic payments data such as credit card and debit card transactions. Since these data are ubiquitous and can be gathered quickly, they could provide more timely and detailed information about economic and financial transactions. Big data could therefore be another non-official resource and represents the next frontier in advancing current analysis. By providing an opportunity to exploit vast quantities of digital information, big data offers fresh insight that is relevant to the monitoring of economic activity and inflation. Its timeliness could augment official data to improve real-time decision making in current analysis, and it could also be an input in the construction of official statistics (Daas and van der Loo 2013).

This article describes the most important attributes of big data and discusses its possible applications and advantages for current analysis. Challenges that limit the full potential of big data, as well as initiatives to address these challenges, are then explored. Finally, the article concludes with the prospects for the use of big data in future current analysis.

What Is Big Data?

Big data refers to large and diverse digital data sets generated from economic transactions and social media interactions of billions of people around the world.

◂ *Big data refers to large and diverse digital data sets generated from economic transactions and social media interactions*

The four Vs of big data

Big data has four main defining attributes: volume, variety, velocity and value. The **volume** of big data is typically much larger than that of traditional data sets. Manyika et al. (2011) describe the size of these data sets as being beyond the ability of typical database software tools to capture, store, manage and analyze. **Box 1** provides a sense of the magnitude of big data.

The types of information that constitute big data come from a **variety** of sources. Only about 10 per cent of big data is structured data (Gens 2011), the type that fits neatly into the rows and columns of relational databases. To be processed by traditional data-management tools and warehouses, and meaningfully interpreted by analysts, data must be in structured form. Examples of structured data are the transactional data that companies

Box 1

The Magnitude of Big Data

- A 2011 study by International Data Corporation (IDC) indicates that 1.8 zettabytes (1.8 trillion gigabytes) of data would be created that year (Gantz and Reinsel 2011). This amount of data would fill 57.5 billion 32-gigabyte iPads (EMC^2 2011).
- Brands and organizations featured on Facebook receive 34,722 "likes" every minute of the day (Wikibon 2012).
- IDC estimates that the number of transactions between firms and those between firms and consumers will reach 450 billion per day by 2020 (Wikibon 2012).
- Walmart processes more than 1 million customer transactions every hour. These transactions are stored in databases that are estimated to contain more than 2.5 petabytes (2.5 million gigabytes) of data. This information would fill 167 times the number of books in the U.S. Library of Congress (Talend 2012).
- The Canadian Payments Association processed 6.3 billion individual retail payments in 2011 alone (Canadian Payments Association 2012).

collect on their customers, and the time-series data that statistical agencies collect on various macroeconomic and financial indicators. Unstructured data, which make up the remaining 90 per cent of big data, include emails, tweets, Facebook posts, road traffic information and audiovisual data. Traditional data warehouses strain under the load of unstructured data and typically cannot process them.

Velocity refers to the fact that data generated from some big data sources such as social media, mobile devices, Internet transactions and networked devices are updated very quickly. This creates an avalanche of data flows that overwhelms most traditional data-analysis hardware and software. Extracting value in real time from rapidly generated data requires specialized skills and data-analysis systems.

The ability to leverage insights and create significant **value** is the most important attribute of big data. Combining big data with sophisticated analytics could provide novel insights into household behaviour, firm expectations, financial market stability and economic activity that would support effective decision making. For example, these advanced methodologies are capable of analyzing patterns in a social network (which may be interconnected in highly complex ways) to determine how these interconnections could influence consumer expectations about inflation or other economic variables (see Einav and Levin 2013 for more details).

◂ *Combining big data with sophisticated analytics could provide novel insights into household behaviour, firm expectations, financial market stability and economic activity*

Big Data and Current Analysis: A Glimpse into the Future

Since accurate and timely information about the current state of economic activity is important for monetary policy decisions, big data provides the opportunity to improve current analysis by exploiting digital data from economic transactions as well as by measuring consumer sentiment from social media and Internet searches. For example, existing monthly indicators could be combined with big data to predict GDP growth before official National Accounts data are released for a given quarter.[3]

An advantage to using big data is the ability to construct metrics that evolve quickly over time. The Billion Prices Project (BPP)[4] at the Massachusetts Institute of Technology, led by economists Alberto Cavallo and Roberto Rigobon,

[3] Binette and Chang (this issue) describe a forecasting tool that uses a data set that, although large, is not of the magnitude of big data.

[4] For more information on the Billion Prices Project, see http://bpp.mit.edu/.

calculates a daily inflation index from a continuously evolving basket of goods. Data for the BPP are collected with software that scours the websites of online retailers for their prices on a wide array of products.[5] The index is then calculated as an average of individual price changes. This virtual real-time inflation index could offer policy-makers and statistical agencies a glimpse of what is happening to inflation in real time. For example, BPP data show that, after Lehmann Brothers collapsed in September 2008, businesses started cutting prices almost immediately, suggesting that aggregate demand had weakened (Surowiecki 2011). In contrast, the official inflation numbers released by the statistical agencies did not show this deflationary pressure until that November, when October CPI data were released.

Canadians are increasingly moving away from traditional methods of payment, such as cash and cheques, toward a variety of electronic payment methods (Canadian Payments Association 2012). Analysis of these timely electronic data could help to predict economic activity and assess possible revisions to official retail and consumption data. Other research provides some evidence that payment-system data could be useful for studying the economic effects of occasional extreme events. For example, Galbraith and Tkacz (2013) use daily data on Canadian debit transaction volumes, as well as data on cheque transaction volumes and values, to investigate the impact on personal consumer expenditures of the 11 September 2001 terrorist attacks, the Severe Acute Respiratory Syndrome (SARS) epidemic in the spring of 2003, and the August 2003 electrical blackout in Ontario and in parts of Northeastern and Midwestern United States. Contrary to initial perceptions of these events, the authors find only small and temporary effects.

Big data could also be used to study developments in the labour and housing markets. Assessments of these markets have been carried out using data on the number of Internet searches. Choi and Varian (2009) find that unemployment and welfare-related searches can improve predictions of initial claims for unemployment benefits. Askitas and Zimmermann (2009), D'Amuri (2009), and Suhoy (2009) also find that Internet searches can be relevant for predicting labour market conditions in Germany, Italy and Israel, respectively. Choi and Varian (2011) as well as Wu and Brynjolfsson (2009) find that housing-related searches can improve on traditional models for predicting housing sales in the United States. Furthermore, Webb (2009) suggests that the high degree of correlation between the number of searches for "foreclosure" and the actual number of foreclosures can be the basis for an early-warning system to predict problems in the U.S. housing market.

McLaren and Shanbhogue (2011) examine the importance of online searches for predicting activity in the labour and housing markets in the United Kingdom. The authors specify two separate models in which either the growth in U.K. unemployment or growth in house prices is a function of previous growth rates. Their results indicate that the inclusion of Internet searches in these models improves the models' forecasting performance. McLaren and Shanbhogue (2011) point out that these data are particularly helpful for analyzing the impact of unexpected developments, such as temporary plant closures, epidemics and labour strikes. While survey data must be collected based on predetermined questions, Internet search data are more flexible and can be used to assess these special circumstances.

Finally, big data could be an input in the construction of official statistics. For example, some European countries are using point-of-sale scanner data in the compilation of their CPIs. Statistics Norway exploits scanner data to

◂ *Big data could be an input in the construction of official statistics*

[5] The prices of services are not included in this data set.

compute a subindex for food and non-alcoholic beverages (Rodriguez and Haraldsen 2006). In June 2002, Statistics Netherlands introduced supermarket scanner data into its CPI (Schut 2002), and the Swiss Federal Statistical Office replaced the prices formerly collected in retail outlets with prices taken from scanner data to calculate its price indexes (Müller et al. 2006).

Challenges and Initiatives

Despite the innovation that has materialized from big data thus far, several factors limit its full potential, chief among them methodological constraints, the lack of easy access to data sets and privacy concerns.

Methodological constraints

Although strides have been made in developing methodologies for extracting value from big data, the implementation of these methodologies for current analysis is still evolving. Specifically, it remains unclear how best to select, organize and aggregate unstructured data so that they provide meaningful signals about economic conditions, and what analytical tools need to be developed to integrate those signals with information from conventional data sources. In addition, subsets of populations covered by big data are at times not necessarily representative of a relevant target population used for official statistics. Assessing how representative big data samples are could prove to be problematic for standard methodologies.

◂ *Although strides have been made in developing methodologies for extracting value from big data, the implementation of these methodologies for current analysis is still evolving*

Lack of access to the data

Much of the data that constitute big data currently exist in silos. To unleash the full potential of big data, there is the need to first integrate the fragmented data sets so that they can be accessed easily and quickly by interested parties. The advent of cloud computing has enabled the creation of data centres that house massive amounts of data in one location. Since pooling of data sets is of paramount importance, a number of initiatives are under way to enhance access to big data. For example, the Government of Canada and the Ontario government have collaborated with IBM and a consortium of seven universities to establish a new Ontario-based, $210 million research project and data centre to help university and economic researchers use high-performance cloud-computing infrastructure to better exploit big data. Another initiative is an agreement between the U.S. Library of Congress and Twitter in 2010 to release 170 billion archived tweets to researchers and other interested parties exploring topics ranging from tracking vaccination rates to predicting stock market activity (Osterberg 2013).

Big data or Big Brother?

Information at the level of individual households and businesses can provide important insights into current economic conditions. By uncovering hidden connections between seemingly unrelated pieces of data, big data analysis can reveal personal information that some might deem too sensitive to share. The reasons for collecting and the need to protect these data are becoming more prominent issues in the debate over privacy and the appropriate use of personal data. Nevertheless, as much as institutions and individuals need to be careful about how invasive they are in their efforts to collect big data, it is difficult to deny that big data analysis has the potential to offer valuable information on economic growth and to improve current analysis. A balanced regulatory framework is therefore necessary to

◂ *A balanced regulatory framework is necessary to effectively address concerns about privacy, while still benefiting from technological advances* effectively address concerns about privacy and the use of personal information, while still benefiting from technological advances and a thriving data-driven economy.

Conclusion

Reliable information about the current state of the economy is an important component for conducting monetary policy and, since data are the main resource driving current analysis, their accuracy and timeliness are key. A better real-time gauge of current economic conditions can improve assessments of economic momentum and forecasts of future growth. Digitization and the advent of the Internet have exponentially increased the amount of data available and also created new, viable sources. Practical applications of these data include the construction of timely price metrics from online retail prices, the use of electronic payment methods to help predict economic activity, and the use of Internet searches to assess labour and housing markets. Harnessing the full potential of this profusion of data is challenging. While some progress in unlocking the value of these data has been made with traditional data-analysis methods, the use of big data for current analysis is still in its infancy.

Literature Cited

Askitas, N. and K. F. Zimmermann. 2009. "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly* 55 (2): 107–20.

Binette, A. and J. Chang. 2013. "CSI: A Model for Tracking Short-Term Growth in Canadian Real GDP." *Bank of Canada Review* (Summer): 3–12.

Canadian Payments Association. 2012. "Examining Canadian Payment Methods and Trends." (October).

Choi, H. and H. Varian. 2009. "Predicting Initial Claims for Unemployment Benefits." Google Inc. Available at http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/fr//archive/papers/initialclaimsUS.pdf.

———. 2011. "Predicting the Present with Google Trends." Google Inc. Available at http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf.

Coletti, D. and S. Kozicki. 2013. "Introduction: Tools for Current Analysis at the Bank of Canada." *Bank of Canada Review* (Summer): 1–2.

Daas, P. and M. van der Loo. 2013. "Big Data (and Official Statistics)." Statistics Netherlands. Working Paper presented at the Meeting on the Management of Statistical Information Systems, Paris and Bangkok, 23–25 April.

D'Amuri, F. 2009. "Predicting Unemployment in Short Samples with Internet Job Search Query Data." Munich Personal RePEc Archive Paper No. 18403.

Einav, L. and J. D. Levin. 2013. “The Data Revolution and Economic Analysis.” National Bureau of Economic Research Working Paper No. 19035.

EMC^2. 2011. “World’s Data More Than Doubling Every Two Years—Driving Big Data Opportunity, New IT Roles.” Press Release, 28 June.

Galbraith, J. W. and G. Tkacz. 2013. “Analyzing Economic Effects of September 11 and Other Extreme Events Using Debit and Payments System Data.” *Canadian Public Policy* 39 (1): 119–34.

Gantz, J. and D. Reinsel. 2011. “Extracting Value from Chaos.” International Data Corporation (IDC) Digital Universe. Available at http://www.emc.com/leadership/programs/digital-universe.htm.

Gens, F. 2011. “IDC Predictions 2012: Competing for 2020.” IDC Analyze the Future (December). Available at http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf.

Lamont, O. 1997. “Do ‘Shortages’ Cause Inflation?” In *Reducing Inflation: Motivation and Strategy*, 281–306. Edited by C. D. Romer and D. H. Romer. Chicago: University of Chicago Press.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. Hung Byers. 2011. “Big Data: The Next Frontier for Innovation, Competition, and Productivity.” McKinsey Global Institute (May).

McLaren, N. and R. Shanbhogue. 2011. “Using Internet Search Data as Economic Indicators.” Bank of England *Quarterly Bulletin* (Q2): 134–40.

Müller, R., H. M. Herren, S. Röthlisberger and C. Becker Vermeulen. 2006. “Recent Developments in the Swiss CPI: Scanner Data, Telecommunications and Health Price Collection.” Swiss Federal Statistical Office. Paper presented at the 9th meeting of the Ottawa Group, London, 14–16 May.

Osterberg, G. 2013. “Update on the Twitter Archive at the Library of Congress.” Library of Congress, 4 January.

Rodriguez, J. and F. Haraldsen. 2006. “The Use of Scanner Data in the Norwegian CPI: The ‘New’ Index for Food and Non-Alcoholic Beverages.” *Economic Survey* 4: 21–28.

Schut, C. (ed.). 2002. “Gebruik van Scannerdata van Supermarkten in de Consumentenprijsindex.” Statistics Netherlands, 4 July.

Suhoy, T. 2009. “Query Indices and a 2008 Downturn: Israeli Data.” Bank of Israel Discussion Paper No. 2009–06.

Surowiecki, J. 2011. “A Billion Prices Now.” *The New Yorker*, 30 May.

Talend. 2012. “How Big Is Big Data Adoption? Survey Results.” Available at http://info.talend.com/rs/talend/images/WP_EN_BD_Talend_SurveyResults_BigDataAdoption.pdf.

Webb, G. K. 2009. "Internet Search Statistics as a Source of Business Intelligence: Searches on Foreclosure as an Estimate of Actual Home Foreclosures." *Issues in Information Systems* 10 (2): 82–87.

Wikibon. 2012. "A Comprehensive List of Big Data Statistics." Wikibon Blog, 1 August. Available at http://wikibon.org/blog/big-data-statistics/.

Wu, L. and E. Brynjolfsson. 2009. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales." Sloan School of Management, Massachusetts Institute of Technology.