New and Emerging Methods

Big Data as a Source of Statistical Information[1]

Piet J.H. Daas and Marco J.H. Puts

Abstract
Big Data is an extremely interesting data source for statistics. Since more and more data is generated in our modern world and is digitally stored, it could certainly be used to replace traditional sources or provide additional information for official statistics. Especially given declines in survey response rates, information gathered from Big Data is an interesting addition. However, extracting statistically-relevant information from Big Data sources is not an easy task. In this paper the current state of the art of research on the use of Big Data for official statistics at Statistics Netherlands is described. The paper is based on the real world experiences of the authors obtained during these studies. The topics discussed are related to Big Data methodology, privacy and security concerns and the skills required for successfully employing Big Data.

Introduction
Big Data is a term that one hears more and more often at conferences, meetings and seminars. Since its first introduction in 1997, in a conference paper by Cox and Ellsworth (1997), it has really become a hot topic. This is understandable if one realizes that between the introduction of the term Big Data and the present, the world has changed from a 'data-poor' environment to a world in which data is abundant (Global Pulse, 2012). This is mainly due to the fact that during this period increasing amounts of data have been generated on the web and by sensors in the ever growing number of electronic devices surrounding us. Because of the ongoing decline in the costs of disk storage this data is no longer thrown away but remains stored. As such, Big Data has the potential to provide information on statistically-relevant populations at high frequency, at a high degree of granularity, and from a wide range of angles, narrowing both time and knowledge gaps. This enables the production of more relevant and timely statistics and can result in proxy indicators that enable richer, deeper insights into human experience than traditional sources of official statistics can (Glasson et al., 2013; Global Pulse, 2012).

Anyone who is able to access and analyze Big Data could – potentially – extract meaning from them and gain a competitive edge. This realization has prompted many commercial companies to write white papers and blogs on the huge potential of Big Data. These stories, however, do not always withstand a rigorous scientific analysis and – unfortunately – tend to place the use and potential of Big Data near the edge of the scientific realm. We agree with Glasson et al. (2013), that Big Data has serious potential as it is a very interesting (secondary) data source for official

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

statistics. However, a rigorous scientific approach is needed to uncover its true potential.

This paper is based on the real world experiences of the authors obtained during the Big Data case studies performed at Statistics Netherlands. An overview of the case studies and their initial findings can be found in the papers of Daas and van der Loo (2013) and Daas et al. (2013). Extracting statistically-relevant information from Big Data sources is not an easy task. In this paper the current state of the art of research on the use of Big Data for official statistics is described. It is the start of the development of Big Data methodology, which is data driven due to the lack of a proper scientific foundation. In addition, the skills needed to perform this work and deal with privacy and security issues to enable this research are briefly discussed.

Big Data characteristics

Big Data is often characterized as data of increasing Volume, Velocity and Variety; the famous 3 V's (Manyika et al., 2011). The volume of what is considered 'big' depends on the capabilities of the organization managing the data and on the capabilities of the applications that are traditionally used to process and analyze it. For a survey oriented organization, the limits of processing Big Data are quickly approached, since in most surveys the sample size is optimized, minimizing the amount of data needed. In this respect, statistical organizations that also process administrative data have a head start. At our office Big Data sources that comprise of up to 100 million records a day are studied. The latter constitutes a file of 25 Gigabytes. Compared to the size of the files routinely used by astronomers and climate researches, i.e. 1 or more Terabytes, these are not enormous amounts of data. However, for a statistically oriented organization they certainly are.

The frequency at which Big Data is generated is the second V (velocity). Statistical offices dealing with a continuous stream of data may initiate the production of more speedy statistics (think of weekly and daily figures) and perhaps even of 'real-time' statistics (Glasson et al., 2013; Struijs and Daas, 2013). It should be noted that, in combination with the data volume, velocity can place high demands on the communication bandwidth available.

The third V refers to variety. This results from the increase in the many different sources that could potentially be used and the variability of the data in these sources. Big Data is often largely unstructured, meaning that it has no predefined data model and/or does not fit well into conventional relational databases. However, the lack of structure can also refer to the fact that little or no information is available on the relationship between data elements. In most cases, this V will increase computational complexity.

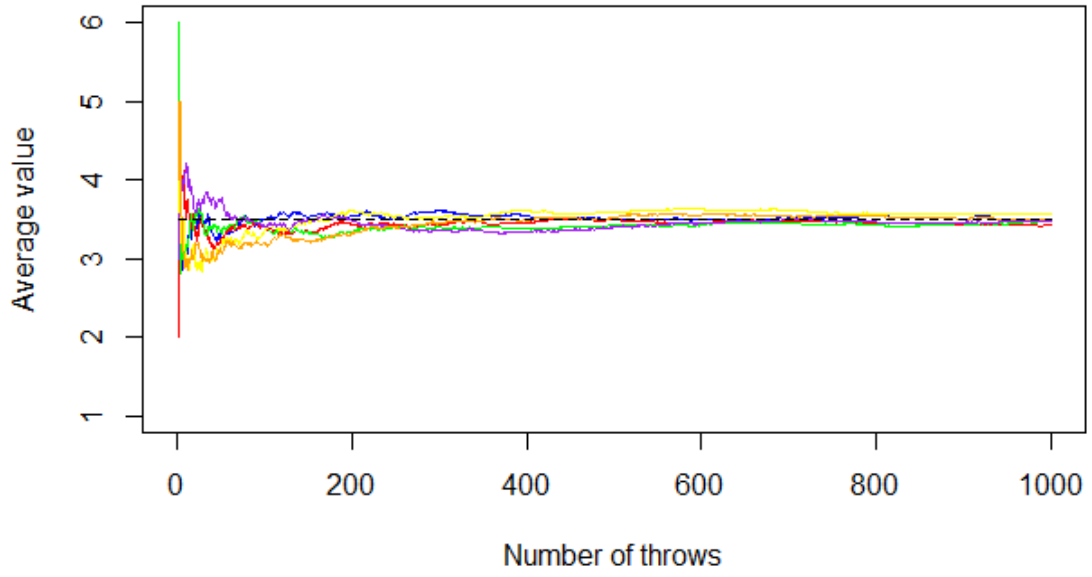
The magnitude of a Big Data source can be seen as the product of the three V's and hence analyzing it can be quite a daunting task. Even more important, however, is the need for a different mindset when performing this task. Working with Big Data requires an open mindset and the ability not to see all problems a priori in terms of sampling theory, e.g. the reduction of variance. Particularly for survey statisticians, who have become accustomed to a data poor environment and – as a result – have developed a focus on extracting the maximum amount of information from (very) small data sets, this may take some time to get used to. However, survey statistics and its methods have adapted to external changes before (Groves, 2011). In our opinion Big Data initiates the need for a change of such magnitude that it truly represents a 'paradigm shift' within the field of statistics (Kuhn, 2012) or even the emergence of a new field of statistical science. Others have a slightly more nuanced view (Walker and Fung, 2013).

Differences between survey and Big Data

The difference between the mindset of using sample surveys and Big Data is illustrated by a plot ordinarily used to demonstrate the law of large numbers. In figure 1 the average of a particular run of throws of a single dice is shown. As the number of throws increases, the average of the values of all the results approaches 3.5; the expected value. Different runs will show a different shape over a small number of throws but over a large number of throws they start to behave very similar. It may take a considerable number of throws to achieve this. The leftmost side of the graphs in figure 1 corresponds to the traditional survey sampling approach.

When small amounts of data are present it is quite difficult to accurately estimate the expected value of the population variable under study. Methods to assure estimates of high accuracy and precision are sought after. When one attempts to use Big Data for statistics the situation more resembles that on the right of figure 12. Large amounts of data are present that, at first, suggest an accurate and precise estimate of the expected value. Note that even after 1000 throws the expected value of 3.5 is not yet reached in several of the runs. The situation will improve when even more runs are performed. This, however, describes the situation in an ideal world.

Figure 1. Development of the average value of 1000 subsequent single dice throws for six different runs. The dotted line represents the expected value (3.5)



Another difference between sample surveys and Big Data becomes clear when looked at the approach traditionally used to derive estimates for population quantities. Suppose a source consists of billions of records, like social media. Big Data sets provide measurements of phenomena at a level of detail far exceeding that of sample surveys. The variability of the phenomena at this level of detail is often found to be large. Sampling of the Big Data set to reduce its size, with the aim to speed up of processing and analysis, only increases the variance of the estimates of the phenomena, thereby forfeiting the benefits Big Data has to offer. In addition, efficient sampling schemes aimed at minimizing the increase in variance are difficult

[2] This comparison is probably not entirely correct, but kept for simplicity. The Big data situation more likely resembles throwing hundreds of dice of all different shapes, with different numbers of sides, while attempting to estimate the average value of the one perfectly cube-shaped dice.

if not impossible to accomplish with Big Data, since stratified or complex sampling is not possible because of a lack of background characteristics of the data records. Simply casting Big Data as a – potentially large – sample is therefore rarely a suitable approach. Alternative data reduction techniques are needed, which optimally reduce the data files in size while retaining as much of their information content as possible.

During our studies we ran into several issues that seriously affect the usability of Big Data for official statistics. Some are methodological, some are practical and some are related to other issues. They are all considered important. A recent UNECE paper confirms these findings and provides a structured overview (Glasson et al., 2013).

The essential issues identified by us are:

1. the ability to process and analyze large amounts of data
2. dealing with noisy, dirty and unstructured data
3. dealing with selectivity
4. ways to go beyond correlation
5. skills needed to perform these tasks
6. legal issues when studying Big Data.

We have not solved all issues yet, but we will describe how we deal or plan to deal with them in the remainder of this paper.

Big Data issues

Processing large amounts of data

When one wants to analyze Big Data, one needs to have a computing environment that enables the rapid processing of large amounts of data. We found – and this is important – that for some data sources analyzing samples or parts of data did not suffice. In many Big Data sources the information content is low, meaning that not all data included is relevant for the research question under study. The low information content is the result of a combination of reasons, the most important of which are: the fact that the data is not generated for the particular purpose the user has in mind (it is secondary data); the data is often unstructured and noisy (more on this below); only a limited number of variables are available; the phenomenon in which the researcher is interested does not occur often. To make successful use of Big Data large amounts of data need to be processed (Boyd and Crawford, 2011). Analyzing a small portion of data may be a good way to start your 'Big Data' studies but the findings derived (if any) certainly need to be verified by those obtained from analyzing the whole dataset. It is through using all data that the true value of Big Data comes out, certainly when you are interested in rare events.

Analyzing Big Data is routinely performed in our office with R or Python, but if a researcher is more comfortable in another programming environment this is no problem. The most important skill here is knowing how to write a program that is able to access all the data in a Big Data set within a reasonable amount of time. Having a secure computer environment with many fast processors, large amounts of RAM and fast disk access certainly helps. Several important considerations are described in Scannapieco et al. (2013) and NAS (2013). Parallel processing could be a way to speed things up. We are currently using (multi-core) general-purpose computing on graphics processing units and are looking at distributed computing, such as (our own) secure private cloud or a local cluster. When studying large amounts of data, creating visual representations is a good way to start to get an idea of their content (Frankel and Reid, 2008).

Dealing with noisy, dirty and unstructured data

As is the case for many secondary data sources and certainly for sources containing huge amounts of data, not all records are relevant for the purpose the researcher has in mind. The irrelevant records, the 'noise', may even negatively affect the relevant ones. We have analyzed quite a number of Big Data sources and have noticed that the signal-to-noise ratio is often rather low. In general the data can be considered as mostly noise; only a mere fraction of the data is of interest – the signal (Silver, 2012). As such, one could consider a lot of Big Data studies as finding a needle in a haystack. However, sometimes the needle also resembles hay. Finding ways to reduce the noise in Big Data, thereby increasing the signal-to-noise ratio, is vital for obtaining a successful result. Aggregating or applying a filter, such as a Kalman or a filter based on Poisson distributed noise (Manton et al., 1999), have been successfully applied, as are constructing queries that prefer the inclusion of relevant records. Another approach routinely used is removing all records with data that are clearly corrupt; dirty records. If one analyzes 100 million records, removing 100 erroneous records is easier than considering the relevance of each 'outlier', their effect on the estimate, and attempting to correct them. One needs to be pragmatic when analyzing Big Data. Applying dimension reduction methods, such as principal component analysis, factor analysis or self-organizing maps, are other ways to decrease the size of the data without losing much information (Hastie et al., 2009).

For unstructured data, such as texts or pictures on web pages, the first step of analysis differs. This type of data is usually transformed into a form more appropriate for statistical analysis; such as word frequencies and their distribution or the clustering of pictures into similar groups. Methods for performing this initial step can be found in areas of science more accustomed to dealing with this kind of data; such as machine learning (Murphy, 2012; Breiman, 2001). We can also learn a lot from Google or other internet giants here (Scott et al., 2013; Spector et al., 2012).

Information theory (Shannon, 1948) can provide more insight into the nature of Big Data. Besides noise, as mentioned above, redundancy also contributes significantly in the total volume of Big Data. Whereas noise reduction can lead to a 'lossy' compression, i.e. part of the data is removed, redundancy reduction leads to a lossless compression of Big Data. In the latter case the data can be reconstructed perfectly. By defining the signal-to-noise ratio in terms of bit-reduction, we could get some appreciation of target information hidden in Big Data.

In all cases, information is extracted from Big Data, usually reducing its size and bringing it more in line with the size of the files statisticians are accustomed to dealing with. In subsequent steps many familiar statistical methods can be used.

Dealing with selectivity

Despite the huge amounts of data present in Big Data sources they may still not cover the complete target population considered. Big Data may therefore be selective. This is an important issue in relation to using Big Data for official statistics where it is – especially when viewed from the perspective of sampling theory – a prerequisite for valid inference. It is therefore essential that this issue is addressed in the context of Big Data. There are several important points to consider.

The first one is the realization that selectivity may vary tremendously per Big Data source and per target variable. Although for some sources this is indeed a major issue, for others coverage is almost or essentially complete, due to the nature of the process through which the data come about. Even when coverage is partial, this may still result in a considerable amount of data for that particular group. Methods that could be used likely resemble post-stratification without having variance issues. This

seems to suggest that the complete absence of a particular part of the population is a more important issue. It will be a challenge to accurately determine this.

The second point is related to the first one. As there is no sample design for Big Data, ways to correct for missing data are needed and will very likely require a model-based or an algorithmic based approach (Buelens et al., 2012). Model-based methods require estimating the model parameters. This pivotal task is challenging for data sources with hardly any auxiliary variables. When models are not tenable, approaches inspired by non-probability correction methods (Baker et al., 2013) or by the area of data-mining and machine-learning (Breiman, 2001) may provide solutions.

The third point is – when applicable – an issue that needs to be addressed first. A lot of Big Data sources actually register events or – more correctly – aggregates of events. This is the reason why many of these sources are big. Examples of (aggregates of) events are: the content and time when a social media message is written, the start and end times and location(s) of a call made by a particular phone and the number of vehicles passing a road sensor at a particular location at a particular point in time. In fact, a major part of the Big Data sources studied by the authors were found to be event-based and hardly contained information on the statistical units of interest. Since the events stored are (indirectly) caused by the statistical units, e.g. people or businesses, dealing with them in a traditional way requires these events to be converted to the corresponding units first. This may not be easy, as a limited amount of identifying information is available. Perhaps the additional use of other (Big Data) sources may assist here. Considering the above, this suggests a two-step approach in which the first step consists of deriving profiles from events to identify units or subpopulations (groups of units). The subsequent step uses inference methods based on the information provided by these. In this context, it is however also important to realize that the part of the population included in Big Data sets might be representative for the whole population for a particular variable studied. Or that, even when it is not, Big Data might still be used to produce (biased) estimates providing that they strongly correlate with existing statistics, for example to improve accuracy and speed of the existing statistics or merely to reduce the sample size.

Correlation and causation

Because of the huge amounts of data available, comparing it to data from another (survey) source may very likely result in a correlation between a particular Big Data variable and a survey variable. However, a high correlation does not always imply causation. In fact having access to more data increases the chance that correlations are found. It is the difficult task of the researcher to fully investigate this 'relation' and try to distinguish a true from a false correlation which may also be referred to as coincidental or spurious. It is best to attempt to falsify the relationship (before anyone else does) by performing additional analysis. Because of its sheer volume this may take some time for Big Data. Patience can be a virtue here, as a longer data series may provide clues to the stability of the relation observed. However, one also needs to realize that the correlation observed may provide a hint at something very interesting. If the correlation cannot easily be falsified, more rigorous analysis aimed at confirming a causal relationship needs to be performed (Pearl, 2009). Be aware that this can be challenging as it requires analyzing lots and lots of data and may even require combining several (Big) data sources. We use cointegration and structural time series studies as our first next step (Krieg and van den Brakel, 2012).

Competencies needed

All of the above merely describes part of the skills currently assigned to survey statisticians. Some of these skills are also new to statisticians experienced in the study of administrative data. Certainly knowledge of high performance computing approaches and algorithmic ways of inference are not types of expertise routinely observed in official statistical environments. Usually survey statisticians, 'register' statisticians and IT-personnel are found in such surroundings. This, in combination with the need for a different mindset and a data-driven (pragmatic) way of analyses prompts us to the sexy "new kid on the block" of statistics; the elusive Data Scientist. This jack of all trades does indeed harbor many of the skills mentioned above. The data science skills usually mentioned are mathematics, statistics, machine learning, computer science and high performance computing, visualization, communication and business/domain expertise (Schutt and O'Neill, 2013). Combined with problem solving skills, perseverance, creativity and an open mind set, these scientists certainly seem suited for the task at hand. However, despite the lack of muscle strength, they seem more to resemble supermen or superwomen. One may seriously consider if a single person is able to gain sufficient expertise in all of these fields. We think that there is a need to create multidisciplinary teams in which people are involved that, as a group, cover all data science skills mentioned above. Such a group should pragmatically tackle Big Data – with an open mind set – from various viewpoints to extract information with the aim of creating statistics. At our office we are currently at the verge of constructing such a group. Depending on its success this group may increase in size.

Privacy and security issues

The Dutch data protection act allows scientific and statistical research on data sources such as Big Data provided appropriate security measures are taken when dealing with privacy-sensitive data. This enables us to perform our research studies on the potential use of Big Data for official statistics. However, the routine production of statistics based on Big Data is another matter. There are several issues, real or perceived, that may impede its routine use (Struijs and Daas, 2013). Data ownership and copyright may be an issue, along with the purpose for which data are registered. Even if data are publicly accessible, for instance on websites or as social media messages that do not have access restrictions, questions of ownership and purpose of publication can be raised. Even the collection of internet data via web robots can be negatively perceived as it causes a burden on the providers of the sites. And even if there are no legal impediments, the perception of the public is a factor to take into account. These concerns have to be taken seriously. Fortunately, there are measures that can be taken to overcome at least some of the obstacles, for example, by anonymizing unique identifiers, removing the privacy sensitive part of a Global Positioning System track (e.g. the first and last 100 meter) or by using informed consent. If a reduction of response burden can be offered, this can be very helpful, also in getting the support of the general public. For the long run, changes in legislation may be considered, to ensure continuous data access for official statistics. But it remains important to stay in line with public opinion, because credibility and public trust are important assets. Within the European Union, changes in the European legislation must also be considered. In addition to the national laws, European laws or regulations can impede the collection of data, even if the Dutch legislation does not present any problem.

Final conclusions

From the above, it is clear that the use of Big Data as a data source for official statistics has considerable implications. Most of all it will certainly affect the work field of statisticians engaged in that area. For them, new skills that go beyond the ones traditionally considered for statisticians, are needed to unleash the true potential of

Big Data. With this, data science becomes introduced in statistical organizations. In addition, and to assure the valid use of Big Data, the need emerges to develop Big Data methodology (West, 2013). Methods specific for the statistical analysis of Big Data must be developed to solve the important methodological challenges describes in this paper. Without such methods, only a limited number of Big Data sources could be used for statistics in an appropriate way. In our opinion, this requires two major changes. The first is shifting the focus of statisticians towards secondary data. The second is changing the mindset to a state that enables collaboration with experts in areas of science with a different statistical culture (Breiman, 2001; Kass, 2011). There are a number of research areas in which considerable expertise on the analysis of large data sets has been developed already. In this respect it is encouraging to see that Big Data gets increasing attention within the international statistical community. At the recent Joint Statistical Meeting and the World Statistics Congress a number of Big Data sessions were held. In addition, several international taskforces have been formed, in which plans for cooperation in this exciting field of research are emerging. Together, the statistical community can certainly face the future with confidence, provided there is a willingness to adapt. Exciting times lie ahead indeed!

Acknowledgments
The authors gratefully acknowledge their Statistical Netherlands colleagues Bart Buelens, Chris de Blois, Rob Willems, Elke Moons, Jan van den Brakel, Peter Struijs, Barteld Braaksma and Marton Vucsan for stimulating discussions and constructive remarks. Mick Couper is gratefully acknowledged for inspiring us to write this paper.

References

Baker, R., Brick, J.M. Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J.,

Tourangeau, R. (2013). Report on the AAPOR Task Force on Non-Probability Sampling. AAPOR report, May.

Boyd, D., Crawford, C. (2011). Six Provocations for Big Data. Paper for A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 21.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science 16 (3), pp. 199-231.

Buelens, B., Boonstra, H.J., van den Brakel, J., Daas, P. (2012). Shifting paradigms in official statistics: from design-based to model-based to algorithmic inference. Discussion paper 201218, Statistics Netherlands, The Hague/Heerlen.

Cox, M., Ellsworth, D. (1997). Application-Controlled Demand Paging for Out-of-Core Visualization. Proceedings of the 8th IEEE Visualization '97 Conference, pp 235-244.

Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2013). Big Data and Official Statistics. Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium.

Daas, P., van der Loo, M. (2013). Big Data (and official statistics). Paper for the 2013 Meeting on the Management of Statistical Information Systems (MSIS 2013), Paris - Bangkok, France - Thailand.

Frankel, F., Reid, R. (2008). Big Data: Distilling meaning from data. Nature 455, p. 30.

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., Khan, A. (2013). What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for

the Modernization of Statistical Production and Services, United Nations Economic Commission for Europe, March 10.

Global Pulse (2012). Big Data for Development: Opportunities & Challenges. White paper, Global pulse, New York, USA, May.

Groves, R.M. (2011). Three Eras of Survey Research. Public Opinion Quarterly 75, pp. 861- 871.

Hastie, T., Tibshirani R., Friedman J. (2009). The elements of statistical learning: data mining, inference, and prediction. Second Ed., Springer.

Kass, R.E. (2011). Statistical Inference: The Big Picture. Statistical Science 26 (1), pp. 1-9.

Krieg, S., Van den Brakel, J.A. (2012). Estimation of the Monthly Unemployment rate for Six Domains through Structural Time Series Modelling with Cointegrated Trends. Computational Statistics and Data Analysis 56 (10), pp. 2918-2933.

Kuhn, T. (2012). The Structure of Scientific Revolutions: 50th Anniversary Edition. University of Chicago Press, Chicago, USA.

Manton, J.H., Krishnamurthy, V., Elliott, R.J. (1999) Discrete Time Filters for Double Stochastic Poisson Processes and Other Exponential Noise Models. Int. J. Adapt. Control Signal Process. 13, pp. 393-416.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A. (2011). Big Data: The next frontier for innovation, competition, and productivity. Report of the McKinsey Global Institute, McKinsey & Company.

Murphy, K.P. (2012). Machine Learning: a Probabilistic Perspective. The MIT Press, USA.

NAS (2013). Frontiers in Massive Data Analysis. The National Academies Press, Washington, D.C., USA.

Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys 3, pp. 96–146.

Scannapieco, M., Virgillito, A., Zardetto, D. (2013). Placing Big Data in Official Statistics: A Big Challenge? Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium.

Schutt, R., O'Neill, C. (2013). Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Sebastopol, CA, USA.

Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., McCulloch, R.E. (2013). Bayes and Big Data: The Consensus Monte Carlo Algorithm. Bayes 250, to appear. Available at: http://static.googleusercontent.com/media/research.google.com/nl//pubs/archive/41849.pdf

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October.

Silver, N. (2012). The Signal and the Noise: Why So Many Predictions Fail — but Some Don't. Penguin Group, New York, USA.

Spector, A., Norvig, P., Petrov, S. (2012). Google's Hybrid Approach to Research. Communications of the ACM 55 (7), pp. 34-37.

Struijs, P., Daas, P.J.H. (2013). Big Data, Big Impact? Paper for the Seminar on Statistical Data Collection, Geneva, Switzerland.

Walker, D., Fung, K. (2013). Big Data and big business: Should statisticians join in? Significance 10 (4), pp. 20-25.

West, G. (2013). Wisdom in Numbers. Scientific American, May, p. 14.

New and Emerging Methods – Call for Volunteers

If you're interested in contributing an article to the "New and Emerging Methods" section of a future edition of *The Survey Statistician*, please contact Mick Couper at mcouper@umich.edu.
