An Insight on Big Data Analytics

Ross Sparks, Adrien Ickowicz and Hans J. Lenz

Abstract This paper discusses the opportunities big data offers decision makers from a statistical perspective. It calls for a multidisciplinary approach by computer scientists, statisticians and domain experts to providing useful big data solutions. Big data calls for us to think in new ways and communicate effectively within such teams. We make a plea for linking data-driven and model-driven analytics, and stress the role of cause-effect models for knowledge enhancement in big data analytics. We remember Kant's statement that theory without data is blind, but facts without theories are meaningless. A case is made for each discipline to define the contribution they offer to big data solutions so that effective teams can be formed to improve inductions. Although new approaches are needed much of the past learning related to small data are valuable in providing big data solutions. Here we have in mind the long-term academic training and field experience of statisticians concerning reduction of dataset volumes, sampling in a more general setting, data depreciation and quality, model design and validation, visualisation, etc. We expect that combining the present approaches will give incentives for increasing the chances for "real big solutions".

1 Introduction

Generally Big Data involves routinely collected data that is integrated from different sources and joined together. The theory is that the combined data hold more information in it than analysing the separate datasets independently. Combined datasets do not always have all the variables of interest but generally hold more variables of interest than the separate datasets. This suggests that Big Data has the potential to solve many problems we could not by analysing these datasets separately. Generally observation studies need to be carefully planned for them make causal inferences

R. Sparks · A. Ickowicz
CSIRO Computational Informatics, North Ryde, NSW 1670, Australia

H.J. Lenz (✉)
Institut Für Statistik und Ökonometrie, Inst. Für Wirtschaftsinformatik,
Freie Universität Berlin, Boltzmannstr.20, K30 D-14195 Berlin, Germany
e-mail: hans-j.lenz@fu-berlin.de

N. Japkowicz and J. Stefanowski (eds.), *Big Data Analysis: New Algorithms for a New Society*, Studies in Big Data 16, DOI 10.1007/978-3-319-26989-4_2

and unless this happens with Big Data it is not going to solve many of the problems we are interested in as statisticians. More effort is needed in designing Big Data collection processes for specific analytical purposes before its value can be broadened in making reliable inferential judgments.

The quality of the information in Big Data collection processes is an issue. There is the issue of gross recording errors that need to be addressed and the quality of the information used to join datasets from different sources need to be considered in the analytical approach. If joined datasets are used and there are selection bias issues with each dataset used in the join, then the combined Big Data will have compounded selection bias issues if the join is carried out using the intersection principle. If the join includes all the data as well as all information that is missing using the union principle in joining the datasets, then we could be dealing with a massive missing data problem, but the selection bias issue will generally be reduced. There are significant challenges when dealing with such situations. For example if a probabilistic join is used then the join has some uncertainty, and this requires a change to the analytical methods to deal with this uncertainty [7]. This adds to the challenge. All of this fits into the section of the paper that looks at the issue of whether the Big Data are fit for purpose.

Section 3 will discuss some general tools that may be useful in reducing the size of the analytical effort in analysing big datasets. These are often useful in taking the original dataset that may be in peta scale or tera scale say down to the more manageable giga scale. This section is by no means complete but it documents what we have found to be useful.

Section 4 of the paper looks at the issue of analysing massive datasets when it is impossible to include all the data in the routines for their analysis. This will use the divide and rule principle, that is, the big dataset will be divided up into manageable pockets in a way that helps improve the analytical purpose, e.g., inference or predictions (or forecasts). How to divide the datasets up is an open research question which will not be answered in this paper, but some general principles will be discussed in Sect. 3.

Section 5 has another look at ways of reducing the volume of data in certain instances. Section 6 finishes with comments about the tension between data mining and statisticians and a call for a collaborative approach to building knowledge that will help us better manage the future. Section 7 examines the question of whether theory is essential. Section 8 briefly examines intellectual property issues. Section 9 finishes with a discussion of the issues and summarises

2 Is Big Data Fit for Purpose?

2.1 Do We Need Big Data?

It is fashionable to talk about the opportunities that Big Data offer decision makers. Big Data is attracting the interests of industry and resulting in their preparedness to invest money and resources in achieving the related business gains. Business gains can be achieved using both Big Data and small data. Therefore industry should think carefully about what they would like to achieve, and then establish whether the appropriate data are available for achieving their objectives or making their decisions. That is, they should think about investing better not investing more. Often we require the appropriate data to achieve unbiased solutions. Before the Big Data focus showed up, statisticians addressed their problems by carefully thinking whether the available data are adequate for the purpose or whether new data needed to be collected by an efficient experimental design.

Statistics has long been the avenue for answering important research questions. However the computer has increased our ability to deal with larger and larger datasets, and in some sense answering more complex questions. However ensuring that the data is fit for purpose is even more important in the Big Data context. Before computers were available, inverting a 4×4 matrix of reals took a considerable amount of time, while now this is trivial. Therefore statistics has evolved with the advent of computers. The growth in computational statistics research has allowed statisticians to fit more complex models than previously was possible by using MCMC methods, and improving inference using bootstrap and cross-validation methods.

Our view is that Big Data increases opportunities, but much that has been learned in the past is also relevant in the Big Data space, and in fact we argue that it is even more important. Our view is that answering the right question is more important than the appropriate data, but a close second is having the appropriate data to answer these questions. Big Data is sold as the means of solving all questions but we feel this perception is misguided. Savage's book [10], links the development of statistics in the late twentieth century to the British-American school and its view of probability as objectivistic theory of knowledge. According to this view, the mathematical concept (model) by which we understand our problems must be obtained by observing repetition of events, and *from no other source whatsoever*. This is quite enlightening in the Big Data settings. The first point made, is that the modern statistics (as defined by [10]) referred to as statistical inference, is the daughter of the probability theory. Accurate inference lies in the construction of a model to understand the data. We will explore that point further in Sect. 7. The second point is that any information other than the repetition of the event remains clueless in regard to the application of statistical techniques. Big Data implies more data, but it may not imply more information. Big Data may not build on our current knowledge or answer our important questions. An excess of non-relevant information is likely to be misleading or may create confusion with what is important or add to our spurious/false 'discoveries'. However Big Data that is built on a theoretical framework of knowledge discovery (see [3], p. 106) is likely to improve our understanding and build on our current knowledge. The view that Big Data offers all the answer to our quests for knowledge, and all we need to do is discover where it is embedded in the Big Data is dangerous.

2.2 What About Big Data Do We Need?

Big Data is unlikely to solve all problems of interest to the data custodians unless it has been designed to achieve this aim. Most routine datasets collect the measures that are easy to accumulate mostly because they are necessary administrative data such as revenues and expenditures, because they are easy to measure and collect or simply “open data” ready for downloading free of any fees. A typical example is data from social networks. But the question arises whether what we got is what we need or is “N = ALL” perhaps a seductive illusion, Harford [6]?

The first step before using any dataset is to decide whether the dataset is fit for purpose. We break the fit for purpose evaluation down into the answering following questions:

1. Are all the appropriate variables available?
2. Are these variables measured accurately enough to answer these question? Are there potential recording errors?
3. Does the data represent the population we wish to make inferences about or wish to predict? What selection biases are there?
4. Does the data cover the appropriate time frames for the purpose? Is the time between measures and the duration of collection appropriate?
5. Are there any redundancies in the dataset that are worth removing?
6. Are all measures well defined and consistently measured over time?
7. Has measurement accuracy improved over time and therefore what historical data are useful for the purpose?
8. Is there any missing data and if there is, then what is the nature of the missingness?
9. Do any of the measurement suffer detection limits? For example, is the measurement process incapable of measuring values either below or above a certain limit?
10. Is the spatial information adequate for the purpose?

Some of these fit well with the five V’s raised by Megahed and Jones-Farmer [8] as volume, variety, velocity, veracity and value. Veracity refers to the trustworthiness of the data in terms of creating knowledge relating to the purpose. This calls for data management processes for maintaining the veracity of the data. For example in large scale sensor networks, where many measures are collected every 5 min over long periods of time, requires real-time checks on the spatio-temporal consistency of measures as well as checking whether the measures are consistent with related measures collected at the same site (e.g., see [11]). Therefore Big Data increases the need for the appropriate level of data management. Improved accuracy can sometimes be forced by a certain level of aggregation either over space/geography or by temporal aggregation. For example considering the average measurement per 5 min when the data are recorded every minute or averaging measurements made within a spatial grid. This certainly has advantages when 1 min measures are highly autocorrelated and neighbouring measures are almost measuring the same entity. On the other hand, this can result in a loss of either spatial or temporal resolution when aggregating over too large space or too large time periods, respectively. It is therefore better to build in the appropriate level of accuracy into measures by using the appropriate data management techniques and controls on the measurement process.

The challenges with sensor networks is whether consistency of measures checking be done at the location of each sensor before sending the information back to the root node in the network (thus not checking for spatial consistency) or send the information to the root node first and then do the multivariate-spatio-temporal consistency checking. Such decision may not depend on which approach delivers greater accuracy but in wireless solar operation sensors this may be based on power considerations. Nevertheless accuracy of measurement will impact on what analytical approach will be used to analyse the data.

3 Basic Toolbox for Analysing Big Data

Datasets are increasing in size and purchasing memory space in this digital age is becoming cheaper. Therefore the size and complexity of datasets is growing nearly exponentially. Having the appropriate tools for dealing with such complexity is important with both n (sample size) and p (number of variables) being large in the n by p data matrix. The following methods are useful in managing the computational complexity:

1. **Aggregation and Grouping**: There are many common examples of aggregations that are common place to-day:

   - The billions of market transactions per second in the world involving over 1000 TB per annum (PB/a) is aggregated into GDP per year (USD/a) published in the UNO Yearbook by the National Accounts Group of UNO, New York (8 Bytes/a).
   - Instead of singletons like screws, nails etc. these are combined into one category/class called hardware as a larger.
   - It is fairly common to bin peoples ages into groups, e.g. age intervals [0, 18], [18, 65], [65, 120], and to study behaviour within cohorts.

2. **Blocking**: Semantic keys are built so that users can find certain information very fast. As an example the Administrative Record Census 2011, Germany, used attribute ‘address’ for household generation as a main blocking variable. Privacy concerns often result in the lowest level of geography that is released on individuals is postal code, and in many analyses this is used as a blocking variable. This is at times used to define people who are similar in some way, e.g., with similar social disadvantage index.
3. **Compression and Sparsity exploitation**: An example is the sparse matrix storage of images such as that used by ‘jpeg’. Dimension reduction techniques of data compression are fairly common. Examples are multi-dimensional scaling (MDS), Projection Pursuit, PCA, non linear PCA, radial basis functions or

wavelets. Examples of application are image reconstruction using wavelets or PCA.

4. **Sufficient statistics**: Another very common data compression approach is to only store the sufficient statistics for later analysis, such as is commonly used in Meta Analysis. This reduces the full data by only storing and using statistical functions of the data, e.g., the sample mean and sample standard deviation for Gaussian data. In modern control theory this principle is applied by signal filtering techniques like the Kalman Filter.
5. **Fragmentation and Divisibility (divide et impera)**: We fragment a feature in such a way that it preserves its essential features for analysis. For example, a company made up of different stores at different location around a country. Keeping the total sales at each store allows us to calculate the total sales for the company. The maximum or minimum sales at each store still allows us to calculate in minimum or maximum sale for the company. The top ten sales at each store allows us to calculate the top ten sales for the company. Where this fails is with the median sales at each store; this does not allow us to estimate the median sale for the company.

   A good example of divisibility is that a joint multivariate density can be preserved by factorization of densities say using Markov fields or Markov chains/processes, e.g., example if $X \rightarrow Y \rightarrow Z$ is a Markov chain, then $f(x, y, z) = f_x(x) f_{y|x}(y|x) f_{z|y}(z|y)$ where $f(x, y, z)$ is the joint density of x, y and z, $f_x(x)$ is the marginal density of x, $f_{y|x}(y|x)$ is the conditional density of y given the value of x, and $f_{z|y}(z|y)$ is the conditional density of z given the value of y.
6. **Recursive versus global Estimation (parameter learning) procedures/ algorithms**: This could involve Generalised Least Squares (GLS) or Ordinary Least Squares (OLS) estimation versus Kalman Filtering or recursive GLS/OLS. For example: the recursive arithmetic mean estimator is given by

   $$\bar{x}_n = (1 - \lambda_n)\bar{x}_{n-1} + \lambda_n x_n$$

   where $\lambda_n = 1/n$, while the Kalman filter includes a signal to noise (variance) ratio, υ, leading to $\lambda_n = 1/(1/\upsilon + n)$.
7. **Algorithms**: One Pass Algorithm (like Greedy Algorithm) versus Multi Pass Algorithms (cf. backtracking, Iteration)
8. **Type of Optimum**: Local optimum/Pareto optimum/global optimum. Heuristic optimisation often delivers a "practical useful" local optimum with strongly bounded computational efforts, the proof of its optimality may be very CPU-time consuming.
9. **Solution types of combinatorial problems**: Limited enumeration, branch and bound methods or full enumeration. Example: Traversing or exploring game trees or social/technical networks.
10. **Sequencing of operations (for additive or coupled algebraic operations) or parallelisation** Examples: Linking of stand-alone programs for solving one (separable) problem in 1- memory-1 CPU machine. Dividing the task up into parallel streams that can be run in parallel to each other.

11. **Invariant Embedding**: Instead of sampling with a given frequency (time window) we record the time stamp, event and value. An example is the measuring of electricity consumption of private households either using a fixed sampling frequency or recording the triple (time stamp, load (kw), type of electric appliance).

Many of the methods mentioned in this section are used to divide the analytical task into more manageable chunks.

4 Dividing the Analytical Task Up into Manageable Chunks

This section will focus on two applications both involving forecasting. The first application deals with forecasting or inferential generalised linear models with a unique defined response variable. The second deals with forecasting counts in complex tabular settings. As the sample size increases generally the proportion of the error due to the systematic error reduces but the proportion of model error starts to increase. Therefore much more attention needs to be devoted to establishing the appropriate model for Big Data applications.

4.1 Generalised Linear Models Example

When datasets get too large to include all observations in the analysis phase, then subdividing the data is important in managing the analytical task. This has computational advantages and information advantages as well (see the results later). Even in smaller datasets it makes statistical sense to divide the data into a test (learning) sample and a validation sample, cf. cross-validation. This is particularly true of model building which is the main focus of this section. The test sample is used to formulate a useful model for prediction/forecasting or inference where selecting: the model, the explanatory variables, and transformations of the response or explanatory variables (e.g., see projection pursuit by [4]). Furthermore, the splitting helps avoiding overfitting and biased estimates of goodness of fit criteria. In addition the test data are used to validate whether any assumptions may hold approximately. After we have settled on a useful operating model with the test data, then we validate the selected model using the new validation dataset. Here we recheck assumptions and assess the goodness-of-fit for the selected model. In other words, the validation dataset is used to assess the usefulness of the selected model. If there are two comparable models selected at the test phase, then the validation dataset can be used to differentiate them and select the better one, or decide to use both and apply ensemble forecasts or inferences.

Different tasks would involve different ways of dividing up the work and so this section is not going to do justice in providing advice for all different tasks. We will consider the forecasting task as one option and we start by looking at forecasting a continuous variable. The model formulation stage would use test dataset (generally about two thirds to half the data). However since this test data may still be excessively large we split this test data into say 100 datasets that are selected randomly without replacement of roughly equal size (n). This process divides the test data into 100 subsets that are non-overlapping but exhaustive of the test dataset. Assume that the ith test sample subset has response variable observations given by vector y_i with related predictor variables that include the same number of observations as in y_i (some of these explanatory variables could be lag response variables). This matrix of predictor variables is denoted by X_i. Consider the generalised linear model structure as an example where

$$g(E(y_i)) = X_i \beta_i, \quad i = 1, 2, \ldots, 100$$

where g is the link function and β_i is the coefficient for the ith test sample and E is the expectation operator. We expect that if the model was appropriate then $\beta_i = \beta$ for all i. We may want to compare either several g link function options or several distribution options for the response y_i. Assume that the fitted models for the ith test dataset resulting in an estimated model formed by substituting β_i by its estimate $\widehat{\beta_i}$ in the equation above. Then since $\beta_i = \beta$ the ensemble estimate for the component β_j is $\bar{\beta}_j = \sum_{i=1}^{100} \widehat{\beta}_{ij}/100$ where $\bar{\beta}_j$ is the generalised linear model estimate of the regression coefficient derived from the partitioned test dataset. The model fitting algorithms produce estimates of model standard errors for each $\widehat{\beta_i}$ which are denoted s_w and interpreted as the within sample uncertainty in the estimate of the coefficient. However the sample estimated standard errors (s_j) for the between test data subsets estimates of the jth regression parameter in the model is given by

$$s_j^2 = \sum_{i=1}^{100} (\widehat{\beta}_{ij} - \bar{\beta}_j)^2/100$$

which assesses how much the individual estimate differs on average from the ensemble estimate. In addition, the distribution of estimated parameters $\widehat{\beta}_{ij}$ for all $i = 1, 2, \ldots, 100$ would be useful in determining the consistency in the jth regression parameters across the various test subset samples. The s_j^2 value is a reflection of the stability of the model across different random samples and measures the robustness of the model parameter estimates. With highly collinear explanatory variables the regression parameter estimates can be unstable, but prediction is usually stable in such cases. We therefore can compare the variation in model prediction errors by calculating

$$S_i^2 = \sum_{k=1}^{n} (y_{ik} - g^{-1}(X_i \widehat{\beta}_{ik}))^2/n$$

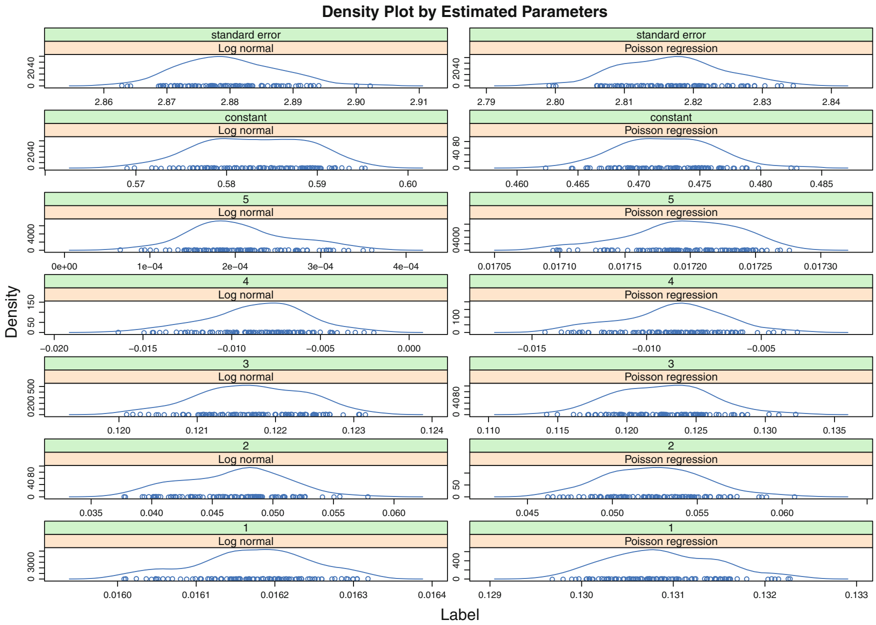

Fig. 1 Comparison of competing models: distribution of estimated parameters and validation standard errors

across all validation samples and test samples. The variation in S_i^2 values provide evidence for the robust performance of the model predictions. These between sample variations could be useful in comparing the robustness of competing models, and therefore help make a decision on the appropriate approximating model (denoted the operating model).

A simulated example is presented in Fig. 1. The data contains 20 million observations generated using the following Poisson regression model $\mu = \exp(0.15 \times x_1 + 0.02 \times x_3 - 0.01 \times x_4) \times as.factor(x_2) \times (\exp(0.06), \exp(0.12), \exp(0.2))$ where $x_1 \sim N(4, 4)$, $as.factor(x_2)$ is a ordinal factor having three levels, $x_3 \sim U(0, 25)$ and $x_4 \sim U(0, 100)$ i.e., uniformly distributed. The response variables were simulated as Poisson with mean μ. The data is split into two 100 validation samples of n = 100,000 observations and the same as training data. The model is fitted using each 100,000 observations in test samples and then the prediction are validated using 100,000 validation dataset. This cycled through each of the 100 training and validation sets. The distributions of the estimated parameters of the model and the prediction standard errors are reported in Fig. 1. Two models were fitted based on no knowledge of the true Poisson regression model used to simulate the data. The Poisson regression model for the counts with expected value:

$$\mu = \exp(\beta_0 + x_1\beta_1 + as.factor(x_2)_{\text{level 2}}\beta_2 + as.factor(x_2)_{\text{level 3}}\beta_3 + x_3\beta_4 + x_4\beta_5)$$

The linear model that is fitted is $log(y+1) = \alpha_0 + x_1\alpha_1 + x_2\alpha_2 + x_3\alpha_3 + x_4\alpha_4 + x_2 \times x_4\alpha_5 + error$. These models are compared in Fig. 1. The regression coefficients in Fig. 1 are in number order (β for the Poisson regression model and α otherwise). Looking at the distribution of the estimated regression parameters and the validation standard errors for the fitted models; the Poisson regression is the better model, and therefore this model is preferred. The estimated regression coefficients generally vary less in the Poisson regression model and the standard errors are on average smaller. The evidence is more clear if the density plot of the differences between the two model matched validation standard errors are plotted, which indicated that the Poisson regression always had a smaller validation standard error. In this way competing models can be compared when faced with large data sets.

A similar approach to the above can be used for fitting Bayesian Hierarchical models (for example). Here we have established credible intervals for model parameters (and forecasts if that is the purpose) for each sample i. These credible intervals could be plotted for all $i = 1, 2, \ldots, 100$ as a way of assessing the validity of the model and the consistency of these intervals. Combining of the Bayesian parameter estimates as mentioned before could provide ensemble estimates for parameters, and the variation of these from the ensemble estimate could be a way of validating the robustness of the model. In addition such empirical evidence can be used to compare different Bayesian hierarchical models and select the model which show the better properties. We believe that in the case of Big Data a validation sample is still necessary because model decisions are still made based on it. This same approach could also be used to compare different burn-in and iteration estimation strategies.

With forecasts, using very large datasets, we wish to avoid refitting the model using all the data each time a new data value is observed. In linear models this can largely be avoided by using some recursive estimation procedure such as the Kalman filter and some state space models [13]. Bolt and Sparks adopted a simpler approach of using a moving window of the same size and exponential weights to give the most recent observation a greater weight, but their approach is only reasonable for one-step-ahead forecasts.

4.2 Forecasting Counts in Complex Tabular Settings

If we are trying to forecast the daily social service needs within a country, then the challenge is a little different. We could still follow the approach designed above, but it is our view that this would not be as efficient as defining cohorts of the population with similar needs and temporal trends. For example, all university students apply for similar support for their university education at the same time of the year. Dividing the population into m different cohorts which have very similar temporal trends and seasonal variation for their demands on the country's social services or geographical regions whose population has homogeneous services needs across time and with the same longitudinal influences seems sensible. The divide of the population into non-overlapping and exhaustive population cohorts is likely to improve the forecasts of needs within cohorts and thus improve the forecasts of the national needs by aggregating up from these cohorts. This approach is not only likely to help make the task more manageable but it will also help improve forecasts.

On the other hand if our interests were in forecasting particular cohort needs and we notice that many cohorts have similar temporal trends, then it may be helpful to decide which cohort counts would be better predicted by forecasting the total counts from the cohorts with similar trends and then proportionally allocate these forecast counts to the respective cohorts. This simplifies the task by aggregating counts to a more manageable level and at times delivers more robust predictions if the cohorts aggregated over all have the same trends.

5 Reducing the Size of the Data that Needs to Be Modeled

The very basic way of reducing the size of the data in space-time applications is by either temporal aggregation thus reducing the number of measures within a unit of time, or spatial aggregation reducing the spatial resolution of the data. An example is the sea surface temperature measured at a fine grid all around Australia with these measures having high spatio-temporal correlations. Assume we were trying to predict the insured costs of floods at 20 locations around Australia given the sea's surface temperatures as explanatory variables. There are several ways of tackling this problem. One is to use technology which exploits Lasso type technology [5, 12] exploiting sparsity, boosting and use ensemble methods. The other approach which we prefer is to create latent variables from sea surface temperature that have physical meaning to the climatologists and are good predictors of flood insured costs at each of the locations of interest. This latent variable (or latent variables) takes the place of these many temperature measurements and therefore reduces the size of the data needed for forecasts.

When we are trying to forecast multi-way tabular counts, e.g., a large array of counts, then at times a drastic reduction in number of cell counts that require forecasts is needed. In such cases it may be worth modelling aggregated cell counts over several dimensions and then proportionally allocate counts to cells that were aggregated over in a way that preserves all interactions. This could be achieved by establishing the cells with the same temporal trends and model the aggregations over these cells counts and then proportionally allocate these forecast totals to the individual cells used to get these total to derive cell forecasts. An example of this is presented in Bolt and Sparks [1]. The only issue with this is if any covariate interacts with time then this model is unlikely to be adequate. Such local errors can quite easily be fixed using temporal smoothing adjustments. Bolt and Sparks [1] approach to forecasting large volumes of counts suited their monitoring applications where early detections of interactions with time were important. Hence this modelling approach will not generally be useful for forecasting applications involving a large number of cells. Another way of reducing the size of the problem is by conditioning, for example, if we condition on age group j and modeled only those in age group j, and repeat this for all other age groups. This could be made more complex by conditioning on age and ethnicity, or by conditioning on three variables. Once the aggregated counts for the conditioned space is found this can be modeled and forecasts established. Forecasts for the whole space is achieved by aggregating over the entire conditional space that makes up the 'whole'. All of these examples lend themselves very well to parallel processing.

6 The Tension Between Data Mining and Statistics

Deming ([3], p. 106) said that "Knowledge comes from theory. Without theory, there is no way to use the information that comes to us on the instant". The Deming quote relating to knowledge may not sit that well with many data mining approaches that search for something interesting in the data. Theory we think is formulated by past observations generating beliefs that are tested by well planned studies, and only then integrated into knowledge when the belief has been "proven" to be true. Data is certainly not information—it has to be turning into information. Many data mining methods are rather short on theory but they still aim to turn data into information. We believe that data mining plays an important role in generating beliefs that needed to be integrated into a theoretical frame which we will call knowledge. When modelling data statisticians sometimes find these theoretical frameworks are too restrictive. At times statisticians make assumptions that have theoretical foundations which are practically unrealistic. This is generally used to make progress towards solving a problem and it is a step in the right direction, but not the appropriate solution. Eventually over time someone builds on this idea and the problem can then be solved without unrealistic assumptions. This is how the theoretical framework is extended to solving the more difficult problems. Non-statistically trained data-miners we believe too often drop the theoretical considerations. Some data-miners attempt to transform data into information using common sense and make judgments about knowledge called learning from the data—sometimes they may get it wrong but often they may be right. Have we statisticians got too hung-up about theory? We do not think so. We may assume too much at first in trying to solve a problem but our foundations are the theory. The current Big Data initiatives are mostly based on the assumption that Big Data is going to drive knowledge (without a theoretical framework). We disagree with this assertion and believe the solution is for data-miners and statistician to collaborate in the process of generating knowledge within a sound theoretical framework. We believe that statisticians should stop making assumptions that remain unchecked and data-miners should work with statisticians in helping discover knowledge that will help manage the future. It is knowledge that helps us improve the management of the future and this should be our focus.

In risk assessment statisticians are generally good at estimating the likelihood, they are trained to evaluate beliefs or hunches and they are trained to building efficient empirical models, but generally they are not adequate trained in the efficient manipulations of massive datasets. Data-miners and computer scientists have the advantage in mining very large volumes of data and extracting features of interest. However there are many issues that data-miners may ignore, e.g., defining the population under study with respect to time, region and subject, defining problem *adequate* variables, utilising background information ("meta data"), paying attention to selection biases when collecting data, the efficient design of observational studies caring about randomness and test/control groups etc.

7 Does the New Big Data Initiative Need No Theory?

The view of Savage on modern statistics raises the question of whether Big Data offers us more information. An interesting question is: does the Big Data current thrust lie outside the modern statistics theory and practice. Alternatively should we define post-modern statistics with Big Data as the main driver. The introduction of this paper questions the current Big Data focus. The ensemble approach of aggregating over the predictions of different models to achieve better predictions may deliver more accurate predictions, but it may not lead to a better understanding than one model. This highlights the importance in selecting the appropriate analytical approach relating to the aim or purpose. However an important question is whether a well thought out model or theory are needed at all?

Statisticians use empirical models to approximate the "real data model" and integrate this with mathematical theory to understand processes and build knowledge. The focus is to understand the sources of variation, and then make conclusions that are supported by the data. Statistical modelling align with Popper [9] view, "*the belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society to be used as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not*". This was true for most of the data we have access to. The model shapes the data in trying to best fit it, and the data shapes the model in that it helps us use models with the appropriate assumptions. The less data we have, the more the appropriate model will help in drawing unbiased-low variance estimators/predictions for our problem. However is the assertion that Big Data reduces the need for developing an operating model? Alternatively can every problem be solved by constructing an appropriate empirical model? Like Breiman [2] we believe statisticians need to be more pragmatic. Breiman [2] notes the existence of two parallel cultures in statistical modelling. The first one assumes the data are generated by a given stochastic data model. The second culture uses algorithmic models and treats the data mechanism as unknown. Breiman [2] accuses the statistical community of having focused too much on appropriate empirical models, leading to the development of "*irrelevant theory and questionable scientific conclusions*". Luckily, since 2001, the discipline evolved through this, and made better use of the available computational resources available. Techniques like Gaussian Processes, Bayesian Non-parametric statistic and machine learning deliver successful outcomes (see [14]). It is probably safe to say that modern statisticians have nowadays a toolbox full of machine learning tricks and data-miners similarly have modern statistical tools in their toolbox. However, as a mathematical discipline, it is unlikely that statisticians will move too far away from their theory-driven techniques to full black-box algorithms.

8 Who Owns Big Data?

Another question of interest is the current shift in the intellectual property from the scientific methodology to the data itself. Until recently the major intellectual property was in building the model/technique/algorithm to extract/infer valuable information from the data. Protection was controlled through patents and publications and ownership was recognised by law. Now there is a view that the intellectual property resides in the data. Companies may trust scientists to use their data to answer research questions, but not without protecting the ownership of their data with confidentiality agreements. Big Data is by essence collected from everywhere. The danger is in every corporate entity protecting their data and this lack of data sharing limits the amount value that integrating data from different sources can offer us in understanding our world. For example understanding the consequence of changes in climate requires insurance companies and companies to share their data on insured costs and losses respectively.

9 Discussion

Big data offers us scientists with numerous challenges, and therefore it demands contributions from computer scientists, data-miners, mathematicians, and statisticians. The greatest difficulty is deciding on what value our various skills offer in solving problems and answering questions using Big Data. We feel that collaboration and co-teaching across each of these disciplines is the best way of deciding on the value we each offer.

The big advantage is that all these disciplines have added to the tools that are needed to manipulate and analyse Big Data. As datasets increase in size we statisticians are going to need to lean on the tools developed by computer scientists and data-miners more and more. In addition, new theoretical frameworks may be needed to ensure that judgment mistakes are not made. The Big Data challenge is extracting information in real-time decision making situations where both n and p are large and there is a real-time dimension to the problem. Often people use simple statistical methods to analyse such data and limit their inference to answering fairly simple questions. However, the challenge for both data-miners and statistician working in this area is to move the questions and analytical methods up to the more complex questions with a particular emphasis on avoiding giving biased solutions.

For large data sets, it is well known now that testing for statistical significance is of limited value and the challenges are more aligned with accurate estimates and confidence intervals. Clearly Big Data research demands diverse skills recognizing that the problems are too difficult and large to be "owned" by one discipline area. Statisticians are lacking in the skills necessary for manipulating these large data sets efficiently, but statisticians have the skills that avoid biases and help divide the analytical tasks into manageable chunks without the loss of information.

The general view is that Big Data is the data miners domain and statistics does not play a key role. However, this view is narrow for the following reasons:

- The data quality challenges for ensuring the data are fit-for-purpose are enormous. It requires statistical skills involving outlier detection that avoids masking and swamping. These would involve:

1. The need for prospective robust statistical quality control methods involving the multivariate spatio-temporal consistency checking of data. The aim being that the measurement process is accurate and that the data are free from influential errors.
2. Planning of the dimension reduction process in a way that preserves all the sufficient statistics for future decisions. In other words, design the aggregation process and data compression process to maximize the information needed for its purpose.
3. Plan for future studies using the data—stratify the population into homogeneous groups to help with sample designs for future analyses. Think about how the data can be used for future longitudinal studies.
4. Propensity score matching should be used to avoid biases in observational studies and planning for potential future designed trials.
5. The whole aspect of assuring that the data are fit for purpose needs careful statistical thought and planning.

- Compressing data is not just about selecting a window over which to aggregate values—it is about compressing the data in a way that retains as much of the necessary information as possible. It is about preserving the sufficient statistics.
- Inference becomes more about mathematical significance (the size of the influence of a variable) and less about statistical significance. Estimation and prediction is all about avoiding biases—there may be selection bias issues.

The challenges listed above are statistical in nature and by no means are complete, but it is important to decide what part each discipline plays in the future development of analytical techniques for large data sets, and what parts are best done in partnership with others.

A quick summary of needs are:

1. Fast and efficient exploratory data analysis
2. Intelligent ways of reducing dimensions (both in the task and the data).
3. Intelligent ways of exploiting sparsity.
4. Intelligent ways of breaking up the analytical task (e.g., stratification and the parallel processing of different strata).

5. Intelligent and efficient visualisation, anomaly detection, feature extraction, pattern recognition.
6. Commitment to unbiased estimation and prediction/forecasting analytics.
7. Effective design—supported by starting with a thinking about what data to collect, how to collect it, and then how to analyse it.
8. Efficient designs for breaking the data into training and validation samples.
9. Real-time challenges—fast processing—estimation, forecasting, feature extraction, anomaly detection, clustering, etc.

References

1. Bolt, S., Sparks, R.: Detecting and diagnosing hotspots for the enhanced management of hospital emergency departments in Queensland, Australia. Med. Inform. Dec. Making **13**, 134 (2013)
2. Breiman, L.: Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statis. Sci. **16**(3), 199–231 (2001)
3. Deming, W.E.: The new economics: for industry, government, education, 2nd edn. The MIT Press, Cambridge (2000)
4. Friedman, J.H., Stuetzle, W.: Projection pursuit regression. J. Am. Statis. Assoc. **76**, 817–823 (1981)
5. Friedman, J.H.: Fast sparse regression and classification. Int. J. Forecast. **28**, 722–738 (2012)
6. Harford, T.: Big data: are we making a big mistake? Significance **11**(5), 14–19 (2014)
7. Lahiri, P., Larsen, M.: Regression analysis with linked data. J. Am. Statis. Assoc. **100**, 222–230 (2005)
8. Megahed, F.M., Jones-Farmer, L.A.: A statistical process monitoring perspective on big data. In: XIth International Workshop on Intelligent Statistical Quality Control, CSIRO, Sydney (2013)
9. Popper, K.: Science as falsification. Conject. Refutat. Readings in the Philosophy of Science, 33–39 (1963)
10. Savage, L.J.: The Foundations of Statistics, Dover edn, 352pp (1972)
11. Sparks, R.S., Okugami, C.: Data quality: algorithms for automatic detection of unusual measurements. Front. Statis. Proc. Control **10**, 385–400 (2012)
12. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Statis. Soc. Series B (Methodological), **24**, 267–288 (1996)
13. West, M., Harrison, P.J.: Bayesian Forecasting and Dynamic Models. Springer, New York (1997)
14. Williams, C., Rasmussen, C.: Gaussian processes for regression (1996). http://eprints.aston.ac.uk/651/1/getPDF.pdf

http://www.springer.com/978-3-319-26987-0

Big Data Analysis: New Algorithms for a New Society
Japkowicz, N.; Stefanowski, J. (Eds.)
2016, XII, 329 p. 63 illus., 35 illus. in color., Hardcover
ISBN: 978-3-319-26987-0