Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 1- 13

Big Data and Agriculture

B.L.S. PRAKASA RAO
CR Rao Advanced Institute of Research in Mathematics,
Statistics and Computer Science, Hyderabad 500046, India

Received: 25 March 2021; Revised: 25 May 2021; Accepted: 28 May 2021

Abstract

This article is based on the presentation "Big Data and Agriculture" delivered as the M.N. Das Memorial Lecture at the 23rd Annual Conference of the Society of Statistics, Computers and Applications "Visionary Innovations in Statistical Theory and Applications (VISTA-2021), Hyderabad. We present several aspects of what is now known as BIG DATA with its advantages as well as pitfalls in general with special reference for agriculture.

Key words: Big data in Agriculture; Massive data; Handling of Big data; Dimension reduction; Computing issues for Big data.

1. Introduction

This article is based on the presentation "Big Data-Agriculture" delivered as the M.N. Das Memorial Lecture at the 23rd Annual Conference of the Society of Statistics, Computers and Applications "Visionary Innovations in Statistical Theory and Applications (VISTA-2021)" held at Hyderabad, India. We present several aspects of what is now known as BIG DATA with its advantages as well as pitfalls. Far from being an exhaustive review of this emerging field, this is an overview from the point of view of a statistician and it is a compilation of ideas of many researchers, organizations and from online sources. For recent surveys on various aspects of Big data as applied to several other areas beside agriculture, the reader is referred to Govindarajulu *et al.* (2015), Prakasa Rao (2015, 2017), Pyne *et al.* (2016), Kitchin (2015), Struiijs *et al.* (2014) and a comprehensive report on the current trends and the future challenges in statistics of Big data published by the London workshop on the future of Statistical Sciences (2014).

Without any doubt, the most discussed current trend in statistics is BIG DATA. Different people think of different things when they hear about Big Data. For statisticians, how to get usable information out of data bases that are so huge and complex that many of the traditional or classical methods cannot handle? For computer scientists, Big Data poses problems of data storage and management, communication and computation. For citizens, Big Data brings up questions of privacy and confidentiality.

Big Data is *relentless*. It is continuously generated on a massive scale. It is generated by online interactions among people, by transactions between people and systems and by sensor-

Corresponding Author: B.L.S. PRAKASA RAO
Email: blsprao@gmail.com

enabled equipment such as aerial sensing technologies (remote sensing), information-sensing mobile devices, wireless sensor networks etc. Big Data is relatable. It can be related, linked and integrated to provide highly detailed information. Such a detail makes it possible, for instance, for banks to introduce individually tailored services and for health care providers to offer personalized medicines. Big data is a class of data sets so large that it becomes difficult to process it using standard methods of data processing. The problems of such data include capture or collection, cleaning, storage, search, sharing, transfer, visualization and analysis. Big data is difficult to work with using most relational data base management systems, desktop statistics and visualization packages. Big Data usually includes data sets with size beyond the ability of commonly used software tools.

When do we say that a data is a Big Data? Is there a way of quantifying the data? We will come back to answering this question later in this article.

Advantage of studying Big Data is that additional information can be derived from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found. For instance, analysis of a large data in marketing a product will lead to information on business trend for that product. Big Data can make important contributions to international development. Analysis of Big Data leads to a cost-effective way to improve decision making in important areas such as health care, economic productivity, crime and security, natural disaster and resource management. Large data sets are encountered in meteorology, genomics, biological and environmental research. They are also present in other areas such as internet search, finance and business informatics. Data sets are big as they are gathered using sensor technologies. There are also examples of Big Data in areas which we can call Big Science and in Science for research. For Government, Big Data is present for climate simulation and analysis and for national security areas. Decoding human genome which took ten years to process earlier can now be done in a week. This is also an example of a Big Data. Human genome data base is another example of a Big Data. A single human genome contains more than 3 billion base pairs. The 1000 Genomes project has 200 terabytes (200 trillion bytes) of data. Human brain data is an example of a Big Data. A single human brain scan consists of data on more than 200,000 voxel locations which could be measured repeatedly at 300 time points.

Big Data sizes are a constantly moving target. It involves increasing volume (amount of data), velocity (speed of data in and out) and variety (range of data types and sources). Big Data are high volume, high velocity and/or high variety information assets. It requires new forms of processing to enable enhanced decision making, insight discovery and process optimization. During the last twenty years, several companies abroad such as Monsanto dealing with agricultural products are adopting to data-driven approach to conduct more targeted services to reduce risks and to improve performance. They are implementing specialized data analytics to collect, store, manage and analyze large data sets. Analyzing these massive data sets help measuring firms risks as well as systemic risks. Analysis of such data requires people who are familiar with sophisticated statistical techniques dealing with productivity and risk management. Big Data are of various types and sizes. Massive amounts of data are hidden in social networks such as Google, Facebook, LinkedIn, Youtube and Twitter. These data reveal numerous individual characteristics and have been exploited by government or official statistics and other agencies.

2. Practice of Big Data in Agriculture

In order to tackle the increasing challenges of agricultural production, the complex agricultural ecosystems need to be better understood. This can happen by means of modern digital technologies that monitor continuously the physical environment producing large quantities of data in an unprecedented pace. The analysis of the big data would enable farmers and companies to extract values from it improving their productivity. Our discussion here is based on the excellent reviews by Carolan (2016), Kamilaris *et al.* (2017) and Coble *et al.* (2018).

Population growth along with socio-economic factors have historically been associated to food shortage. In the last fifty years, the world population has grown from three billion to more than six creating a high demand for food. The global population would increase by more than thirty percent by the year 2050 according to UN which means that a seventy percent increase on food production must be achieved. Land degradation and water contamination, climate change, socio-cultural development, government policies, market fluctuations add uncertainties to food security. These uncertainties challenge agriculture to improve productivity. Advancements in crop growth modeling and yield monitoring together with global navigation satellite systems (GPS) have enabled precise localization of point measurements in the field so that spatial variability maps can be created, a concept known as “precision agriculture”.

Agricultural practices are now supported by biotechnology and emerging digital technologies such as remote sensing, cloud computing and internet of things leading to the notion of “smart farming”. Smart farming is important for tackling the challenges of agricultural production in terms of productivity, environmental impact, food security and sustainability. To address the challenges of smart farming and sustainable agriculture, the complex multivariate and unpredictable agricultural ecosystems need to be better analyzed and understood. Emerging digital technologies contribute to this understanding by monitoring and measuring continuously various aspects of the physical environment producing large quantities of data in unprecedented pace. This implies the need for large scale collection, storage, pre-processing, modeling and analysis of large amount of data coming from various heterogeneous sources.

Agricultural “big data” creates the necessity for large investments in infrastructure for data storage and processing which need to operate in real-time for some applications such as weather forecasting, monitoring for crop pests and animal diseases. “Big Data Analytics” is the term used to describe a new generation of practices designed so that the farmers and related organizations can extract economic value from very large volumes of data by enabling high-velocity capture, discovery and analysis.

Big data can be used in several agricultural applications. Applications related to estimations of crop production and yields, land mapping, weather forecasting and food security require large amounts of data. Recognition of animals’ diseases, plants’ poor nutrition require high velocity decision making as well decisions on farmers’ productivity, weather forecasting which need to be taken in real time. In these cases, the time horizon for the decisions involved require operational and tactical planning in a short term.

Some applications such as computation of crop insurance premiums, farmers’ productivity improvement require wide variety of data from heterogeneous sources. Furthermore some applications require high reliability of the data involved for studies on diseases of plants/animals.

2.1. Some possible area of applications of Big Data analytics

(1) Platforms enabling supply chain actors to have access to high-quality products and processes enabling crops to be integrated to the international supply chain according to the global needs.

(2) As farmers are sometimes not able to sell harvests due to oversupply and not getting the expected returns from harvests, tools for better predictions for yield and demand have to be developed.

(3) Providing advice and guidance to farmers based on their crops' responsiveness to fertilizers is likely to lead to a more appropriate management of fertilizers use. This could also apply to better use of herbicides and pesticides by the farmers.

(4) Scanning equipment in plants, shipment tracking and retail monitoring of consumers' purchases creates the potential to enhance products' traceability through the supply chain increasing food safety. Since the agricultural production is prone to deterioration over time after harvesting, optimization procedures are essential to minimize losses and maximize quality.

(5) Remote-sensing for large scale land/crop mapping will be critical for monitoring the impact in respect of measuring and achieving the productivity and environmental sustainability targets.

(6) Advanced scientific models and simulations for environmental phenomena could provide a basis for establishing platforms for policy makers assisting in decision making towards sustainability of physical ecosystems.

(7) High throughput screening methods that can offer quantitative analysis of the interaction between plants and their environment with high precision and accuracy.

(8) In advanced countries, self-operating agricultural robots could revolutionize agriculture and its over-all productivity as they may automatically identify and remove weeds and identify and fight pests in the agricultural farms.

(9) Fully automated and data-intensive production systems (such as greenhouses and other indoor led illuminated places) would be on the rise with less use of pesticides, water etc.

Majority of the applications discussed would produce large amounts of big data which could be used by policy makers to balance supply and demand, reduce the negative impact of agriculture on the environment, raise food safety and security and increase productivity. The potential open access of the data to the public can create large number of opportunities for research and development towards smooth and more sustainable farming.

A variety of indicators suggest that the availability of sensors, mapping technology and tracking technologies have changed many farming systems and the management of the food system as it flows from producers to consumers abroad. Big data has significant potential to address the issues of modern societies including the needs of consumers, financial analysts, bankers, marketing agents, producers and decision makers. Sensor technology and data analytics are now applied abroad to agricultural applications. Agricultural companies are investing in robotics, mechanization and other hardware along with farm management software, sensors and internet of things (IoT).

2.2. Applications in agriculture involving Big Data

(i) **Weather forecasting**: Environmental factors like weather influence crop growth and development as well as demand for agricultural and non-agricultural lands. Agricultural production has spatial yield variability partly because of spatial variability in soil and interaction with weather which is also spatially varied.

(ii) **Crop yield production and crop selection**: Machine learning provides many effective algorithms which can identify input and output relationship in crop selection and yield prediction. Using techniques such as artificial neural networks (ANN), one can do crop selection based on factors such as climate, soil, natural calamities, famine and other inputs.

(iii) **Irrigation systems**: Agriculture consumes a major portion of water. Variability in rainfall, climate change and dropping of the water table creates problems. Using smart irrigation systems, the data collected by the sensors can be used to make better decisions regarding water shortages.

(iv) **Crop disease prediction**: ANN is helpful in predicting pest attacks in advance. Such models deal with noisy and wide variety of data and account for possible factors such as historical data, sattelite/sensor data, field conditions, images of leaves etcetra to effectively learn and predict crop diseases.

(v) **Agricultural policy and trade**: A large quantity of data on production output of crops, changes in import and export costs, market demand and supply, market price trends, cultivation costs, payments for workers, transportation costs and marketing costs could be used by ANN algorithms to predict support prices for different commodities produced by farmers.

Big Data has been celebrated as a “revolution”, as “the end of theory” and predictive analytics based on the big data are “crucial” for the end of poverty everywhere. Big Data is considered as the next “Big Thing” in many agricultural circles due to the expanding amount of information collected relating to crop production (big sil data) combined with extensive weather data (big climate data) which together form the backbone of precision agriculture technology. Precision agriculture can play a substantial role in meeting the increasing demand for food and raw materials while ensuring sustainable use of natural resources and the environment.

“Big Data” is still undefined. When do we term a data as Big Data? What was “big” data fifty or sixty years ago is “small” by today’s standards. What is “big” today might be considered “small” several years from now.

Bio-technology can also be considered as “big data” as locating the genes for favourable and profitable traits in plants in order to create new seed varieties involves sifting through billions of base pairs in genomes which is a big data!! Analyzing large amount of data is a difficult issue. One important tool for such analysis is the vector auto-regressive model involving generally at most ten variables and the number of parameters grows quadratically with the size of the model. One solution is to incorporate what is called a sparsity assumption. What it essentially implies that even though there may be infinite number of unknown parameters present in the stochastic model, except for finite number of them, most of the others are negligible and have very little influence on the estimation issues. There are concerns dealing with Big Data such as privacy. We will not discuss this issue in this article.

2.3. When is a data a BIG DATA? (*c.f.* Fokoue (2015); Report of London Workshop (2014))

Big Data comes in various ways, types, shapes, forms and sizes. The dimensionality p of the input space (number of parameters) and the sample size n are usually the main ingredients in characterization of data bigness. Large p small n data sets will require different set of tools from the large n small p sets. Here n is the data size and p the number of unknown parameters/varaibles/covariates. There is no method which performs well on all types of data.

Let us consider a data set $\mathcal{D} = \{(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}), \ldots, (\mathbf{x_n}, \mathbf{y_n})\}$ where $\mathbf{x}'_i = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional vector of characteristics/covariates from the input space $\mathcal{X}$ and y_i is the corresponding response. The matrix $\mathbf{X}$ of order $n \times p$ given by

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np}) \end{pmatrix}$$

is the data matrix.

Five aspects of the **data matrix** are important:

(i) the dimension p representing the number of explanatory variables measured; (ii) the sample size n representing the number of observations/sites at which the variables are measured or collected; (iii) The relationship between p and n measured through the ratio of them; (iv) The type of variables measured (categorical, interval, count, ordinal, real-valued, vector-valued, function-valued) and the indication of scales/units of measurement; and (v) The relationship among the columns of the data matrix to check multi-collinearity in the explanatory variables.

2.4. What is meant by "Massive or Big Data" as a function of p?

Suppose we are dealing with a multiple linear regression problem with p covariates or explanatory variables under a Gaussian noise/error. For a model space search for variable selection, we have to find the best subset from among $2^p - 1$ models/sub-models. If $p = 20$, then $2^p - 1$ is about a million; if $p = 30$, then $2^p - 1$ is about a billion; and if $p = 40$, then $2^p - 1$ is about a trillion. Hence any problem with more than $p = 50$ variables is a massive data problem. It involves searching a thousand trillion models which is a huge task even for modern computers. Hence any problem with more than 50 predictor variables can be called BIG DATA. If the number of predictor variables is more than 100, then it is called a MASSIVE DATA problem.

2.5. What is meant by “Massive or Big Data” as a function of n?

We generally believe that the larger the sample from a population, the better is the inference, due to the law of large numbers. However the computational and statistical complexity in using methods of regression analysis involves inversion of $n \times n$ matrices which is computationally intensive when n is large. It takes $O(n^3)$ number of operations to invert an $n \times n$ matrix. Based on this observation, we might say that the data is observation-massive if $n > 1000$.

2.6. What is meant by “Massive or Big Data” as a function of n/p?

Suppose that we are in a situation with a data where $p > 50$ or $n > 1000$. We have seen that the data can be considered massive in both cases. However the ratio n/p is even more important than n and p taken separately. Let us suppose that we have at least ten observations for each one of the p variables. Hence we have $n > 10p$. Let us also suppose that the information in the data is an increasing function of n. We have the following scenario (c.f. Fokoue (2015)).

(A) $n/p < 1$	Ip	$n << p, n > 1000$	Large p, Large n
(D) $n/p < 1$	Ip	$n << p, n \leq 1000$	Large p, Small n
(B) $1 \leq n/p < 10$	Is	$n > 1000$	Small p, Large n
(E) $1 \leq n/p < 10$	Is	$n \leq 1000$	Small p, Smaller n
(C) $n/p \geq 10$	Ia	$n >> p, n > 1000$	Smaller p, Large n
(F) $n/p \geq 10$	Ia	$n >> p, n \leq 1000$	Smaller p, Small n

(Ip=Information poverty; Is= Information scarcity; Ia= Information abundance)

The BIG DATA problem is with the cases (A) and (D).

For statisticians, Big Data challenges some basic paradigms. The aim is to develop a model that describes how the response variable is related to p other variables or covariates and to determine which variables are important to characterize or explain the relationship. Fitting the model to data involves estimating the parameters from the data and assessing the evidence that they are different from zero indicating the importance of the variable. When $p >> n$, the number of parameters is huge relative to the information about them in the data. Thousands of irrelevant parameters will appear to be statistically significant if one uses small data statistics. Big Data has special features that are not present in the classical data sets. Big Data are characterized by massive sample size and high-dimensionality. Massive sample size allows one to discover hidden patterns associated with small sub-populations. Modeling the intrinsic heterogeneity of Big Data needs better statistical methods. The problems of high-dimensionality in data are noise accumulation, spurious correlation and incidental endogeny. Big Data is often a consequence of aggregation of many data sources corresponding to different sub-populations. Each sub-population might have a unique feature which is not shared by others. A large sample size enables one to better understand heterogeneity. A mixture model for the population may be appropriate for a Big data. Analyzing Big Data requires simultaneous estimation or testing of a large number of parameters. Errors in inferring on these parameters accumulate when a decision on inference from the data depends on these parameters. Such a noise accumulation is severe in high-dimensional data and it may even dominate the true signal. This is handled by the sparsity assumption. High-dimensionality brings in spurious correlation due to the fact that many uncorrelated random variables may have high sample correlation coefficient in high dimensions. Spurious correlation leads to wrong inferences and hence false results. Unlike spurious correlation, incidental endogeny may be present in Big Data. It is the existence of correlation between variable “unintentionally” as well as due to “high-dimensionality”. The former is analogous to finding two persons who look alike but have no genetic relationship where as the latter is similar to meeting an acquaintance by chance in a big city. Endogeny happens due to selection bias, measurement errors and omitted variables. With the advantage of high-tech measurement techniques, it is now possible to collect as many features as possible. This increases the possibility that some of them might be correlated to the residual noise leading to incidental endogeny.

Another reason for incidental endogeny is the following. Big Data are usually aggregated from multiple sources with possibly different data generating schemes. This increase the possibility of selection bias and measurement errors which also leads to possible incidental endogeny. Some statistical methods have been proposed to handle such issues such as penalized quasi-likelihood to handle noise accumulation issue.

Big Data are massive and very high-dimensional and involve large-scale optimization if one wants to use a likelihood or quasi-likelihood approach directly. Optimization with a large number of variables is not only expensive due to computational costs but also suffers from slow numerical rates of convergence and instability. It is also computationally infeasible to apply optimization methods on the raw data. To handle the data both from statistical and computational views, **dimension-reduction** techniques have to be adopted.

2.7. Some issues with Big Data (*c.f.* Fokoue (2015); Buelens *et al.* (2014))

(i) **Batch data against incremental data production :** Big Data is delivered generally in a sequential and incremental manner leading to online learning methods. Online algorithms have the important advantage that the data does not have to be stored in memory. All that is required is in the storage of the built model at the given time in the sense that the stored model is akin to the underlying model. If the sample size n is very large, the data cannot fit into the computer memory and one can consider building a learning method that receives the data sequentially or incrementally rather than trying to load the complete data set into memory. This can be termed as sequentialization. Sequentialization is useful for streaming data and for massive data that is too large to be loaded into memory all at once.

(ii) **Missing values and Imputation schemes:** In most of the cases of massive data, it is quite common to be faced with missing values. One should check at first whether they are missing systematically, that is in a pattern, or if they are missing at random and the rate at which they are missing. Three approaches are suggested to take care of this problem: (a) Deletion which consists of deleting all the rows in the Data matrix that contain any missing values (b) Central imputation which consists of filling the missing cells of the Data matrix with central tendencies like mean, mode or median and (c) Model-based imputation methods such as EM-algorithm.

(iii) **Inherent lack of structure and importance of pre-processing:** Most of the Big Data is unstructured and needs pre-processing. With the inherently unstructured data like text data, the pre-processing of data leads to data matrices, whose entries are frequencies of terms in the case of text data, that contain too many zeroes leading to the sparsity problem. The sparsity problem in turn leads to modeling issues.

(iv) **Homogeneity versus heterogeneity:** There are massive data sets which have input space homogeneous, that is, all the variables are of the same type.There are other types of Big Data where the input space consists of variables of different types. Such types of data arise when the variables can be categorical, ordinal, interval, count and real-valued.

(v) **Differences in measurement:** It is generally observed that the variables involved are measured on different scales leading to modeling problems. One way to take care of this problem is to perform transformations that project the variables onto the same scale. This is done either by standardization which leads all the variables to have mean zero and variance one or by unitization which consists in transform the variables so that the support for all of them is the unit interval [0,1].

(vi) **Selection bias and quality:** When Big Data are discussed in relation to official statistics, one point of criticism is that Big Data are collected by mechanisms unrelated to probability sampling and are therefore not suitable for production of official statistics. This is mainly because **Big Data sets are not representative of a population of interest**. In other words, they are selective by nature and therefore yield biased results. We will not discuss this issue here. Big Data is often a by-product of some process not primarily aimed at data collection.

Analysis of Big Data is data-driven and not hypothesis-driven. **For Big Data, the coverage is large but incomplete and selective.** It may be unclear what the relevant target population is.

(vii) **Comparison of data sources:** Let us look at a comparison of different data sources for official statistics as compared to Big Data.

Comparison between Sample Survey and big data:

Data Source	Sample Survey	Big Data
Volume	Small	Big
Velocity	Slow	Fast
Variety	Narrow	Wide
Records	Units	Events or Units
Generator	Sample	Various Organizations
Coverage	Small fraction	Large/Incomplete

(Ref: Buelens *et al.* (2014))

Comparison between Census and Big Data:

Data Source	Census	Big Data
Volume	Large	Big
Velocity	Slow	Fast
Variety	Narrow	Wide
Records	Units	Events or Units
Generator	Administration	Various Organizations
Coverage	Large/Complete	Large/Incomplete

(Ref: Buelens *et al.* (2014))

Big Data can be the single source of data for the production of some statistic about a population of interest. Assessing selectivity of the data is important. Correcting for selectivity can some times be achieved by choosing suitable method of model-based inference (Breiman (2001)). These methods are aimed at predicting values for missing/unobserved units. The results will be biased if specific sub-populations are missing from the Big Data set. Big Data set can be used as auxiliary data set in a procedure mainly based on a sample survey. The possible gain of such an application for the sample survey is likely reduction in sample size and the associated cost. Using small area models, the Big Data can be used as a predictor for survey based measurement.Big Data mechanism can be used as a data collection strategy for sample surveys. Big Data may be used irrespective of selectivity issues as a preliminary survey. Findings obtained from Big Data can be further checked and investigated through sample surveys.

3. Methods of Handling Big Data (*c.f.* Fokue (2015))

(i) **Dimension reduction:** Dimensionality reduction involves the determination of intrinsic dimensionality q of the input space where $q << p$. This can be done by orthogonalization techniques on the input space which reduces the problem to a lower dimension and orthogonal input space leading to variance reduction for the estimator. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are the methods for dimensionality reduction. However if $p >> n$, then most of these techniques cannot be used directly.

(ii) **Bagging:** As it was observed earlier, it is common in massive data that a single model selected does not lead to optimal prediction. If there is a multicollinearity between the variables which is bound to happen when p is very large, the estimators are unstable and of large variances. Bootstrap aggregation (also called bagging) reduces the variance of the estimators by aggregation of bootstrapped versions of the base estimators.

(iii) **Parelellization:** When the computational complexity for building the base learner is high, the method of bagging becomes inefficient and not practical. One way to avoid this problem is to use parallel processing. Big Data analytics will need parallel processing or parallelization for speeding up computation or to handle massive data that cannot fit into a single computer memory.One way to make statistical procedures more efficient in analysis of Big Data is to paralellize them, that is, to write many algorithms that can run on many computers or many processors at the same time. The method of "Bootstrap" is a standard method for inferring the probability distribution from a sample. It is computationally intensive. It is ideally suitable for paralellization because it involves generating numerous independent rounds of simulated data. One can use "Bag of Little Bootstraps" (BLB) which generates results comparable to the regular bootstrap but much faster.

(v) **Regularization:** With large p and small n, there exist a multiplicity of solutions for any optimization problem involving Big Data and hence the problem becomes ill-posed. Regularization methods are used to find a feasible optimal solution and one method of regularization is Lagrangian formulation of a constrained version of the problem. LASSO (Tibshirani (1996)) is one such method in high-dimensional data analysis.

(vi) **Assumption of sparsity:** As we noted earlier, thousands of irrelevant parameters will appear to be statistically significant if we use small data statistics for Big Data. In classical statistics, if the data implies occurrence of an event that has one-in-a million chance of occurring, then we are sure it is not by chance and hence consider it statistically significant. But if we are considering a Big Data with a large number of parameters, it is possible for the event to occur by chance and not due to significance of the relationship. Most data sets have only a few strong relationships between variables and everything else is noise. Thus most of the parameters do not matter. This leads to sparsity assumption which is to assume that all but a few parameters are negligible.

(viii) **The problem of "Big n, Big p, Little t":** The speed at which one can process is an important element in analyzing Big Data. Classical statistics was always done in an off-line mode, the size was small and the the time for analysis was essentially unlimited. However, in the era of Big Data things are different.

(ix) **Computing issues for Big Data (Fan *et al.* (2013)):** As was mentioned earlier, the massive or very large sample size of Big data is a challenge for traditional computing infrastructure. Big Data is highly dynamic and not feasible or possible to store in a centralized data base. **The fundamental approach to store and process such data is to "divide and conquer".** The idea is to partition a large problem into more tractable and independent sub-problems. Each sub-problem is tackled in parallel by different processing units. Results from individual sub-problems are then combined to get the final result. "Hadoop" is an example of basic software and programming infrastructure for Big Data processing. "MapReduce" is a programming model for processing large data sets in a parallel fashion. "Cloud Computing" is suitable for storing and processing of Big Data. We are not discussing the problems involved in storage and computation connected with Big Data in this article.

3.1. Why Big Data is in trouble?

Answer: They forgot about Applied Statistics (Jeff Leak, May 7, 2014 "Simply Statistics"). There were articles with titles such as "The parable of Google Flu: traps in big data analysis", "Big data: are we making a big mistake ?", "Google Flu trends: the limits of big data", "Eight (No, Nine!) problems with Big Data" *etc.* All of the articles mentioned here and on-line point out the problems of Big Data such as sampling populations, multiple testing, selection bias and over-fitting beside others.

"There is a tendency for Big Data researcher and more traditional applied statistician to live in two different realms. Big Data offers enormous possibilities for understanding human interactions at a societal scale with rich spatial and temporal dynamics and for detecting complex interactions and nonlinearities among variables. However traditional “small data” often offer information that is not contained in Big Data” (Lazer *et al.* (2014)).

Acknowledgements

The author thanks all the colleagues whose ideas were used in preparing this overview. Work of the author was supported under the scheme “INSA Senior Scientist” of the Indian National Science Academy (INSA) at the CR Rao Advanced Institute of Mathematics, Statistics and Computer science, Hyderabad, India. The author thanks Prof. V.K. Gupta for inviting him to deliver the M.N. Das Memorial Lecture at the 23rd Annual conference of SSCA held in Hyderabad, India during February 24-28, 2021 where this paper was presented.

References

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, **16(3)**, 199-231.

Buelens, D., Daas, P., Burger, J. and van den Brakel, J. (2014). Selectivity of Big Data. *Discussion paper, Statistics Netherlands.*

Carolan, M. (2016). Publicizing food: Big data, precision agriculture and co-experimental techniques of addition. *Sociologia Ruralis*, DOI:10.1111/soru.12120.

Coble, K. H., Mishra, A. K., Ferell, S. and Griffin, T. (2018). Big data in agriculture: A challenge for the future. *Applied Economic Perspectives and Policy*, **40**, 79-96.

Fakoue, E. (2015). A taxonomy of Big Data for optimal predictive machine learning and data mining. *arXiv.1501.0060v1 [Stat. ML] 3 Jan 2015.*

Fan, J., Han, F., and Liu, H. (2013). Challenges of Big Data analytics. *arXiv.1308.1479v1 [stat.ML] 7 Aug 2013.*

Govindarajulu, V., Raghavan, V. V. and Rao, C. R. (2015). Big Data Analytics. *Handbook of Statistics*, Vol. 33, Elsevier, Amsterdam.

Kamilaris, A., Karatakoullis, A. and Prenafeta-Boldu, F. X. (2017). A review on the practice of big data in agriculture. *Computers and Electronics in Agriculture*, **143**, 23-37.

Kitchin, R. (2015). Big data and official statistics: Opportunities, challenges and risks. *The Programmable City working Paper 9* .

Lazer, D., Kennedy, R., King, G. and Vespignani„ A. (2014). The parable of Google Flu traps in Big Data analysis. *Science.* **343**, 1203-1205.

Leak, J. (2014). Why big data is in trouble; they forgot about applied statistics. *“Simply Statistics”*, May 7, 2014.

Prakasa Rao, B.L.S. (2015). Brief Notes on Big Data: A Cursory Look. *Lecture Notes*, C R Rao Advanced Institute for Mathematics, Statistics and Computer Science, Hyderabad, India, pp.1.

Prakasa Rao, B.L.S. (2017) Brief Notes on Big Data. *Visleshana*, Computer Society of India, Special Interest Group - Big Data Analytics, **1** (3), April-June 2017, pp.9-12.

Pyne, S., Prakasa Rao, B.L.S. and Rao, S.B. (2016). *Big Data Analytics.* Edited by Saumyadipta Pyne, B.L.S. Prakasa Rao and S.B. Rao, Springer (India) Pvt. Ltd., New Delhi.

Struijs, P., Braaksma, B. and jH daas, P. (2014). Official statistics and Big data. *BIG DATA & Society*, DOI: 10.1177/2053951714538417, SAGE (2014).

Tibshirani, R. (1996). Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society*, **Series B 58**, 267-288.

Current trends and future challenges in statistics: Big Data. (2014). *Statistics and Science: A Report of the London Workshop on future of the Statistical Sciences*, pp. 20-25.