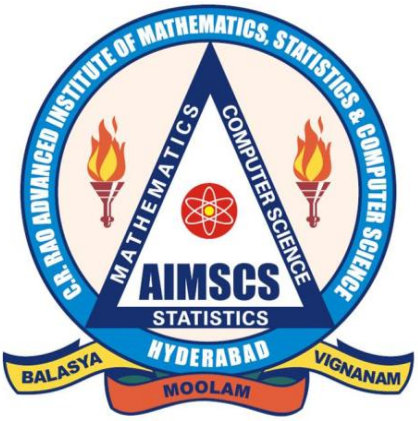
C R RAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

C R RAO AIMSCS Lecture Notes Series



Author (s): **B.L.S. PRAKASA RAO**

Title of the Notes: **Brief Notes on BIG DATA: A Cursory Look**

Lecture Notes No.: **LN2015-01**

Date: **June 29, 2015**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

BRIEF NOTES ON BIG DATA: A CURSORY LOOK

B.L.S. PRAKASA RAO [1]
CR Rao Advanced Inst. of Mathematics, Statistics
and Computer Science, Hyderabad 500046, India

1 Introduction

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. This brief notes gives a cursory look on ideas on several aspects connected with collection and analysis of Big Data. It is a compilation of ideas from different people, from various organizations and from different sources online. Our discussion does not cover computational aspects in analysis of Big Data.

2 What is BIG DATA?

(Fan et al. (2013))
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-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, curation, storage, search, sharing, transfer, visualization and analysis. Big data is difficult

[1] For private circulation only

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?

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. These include"Large Hadron Collision Experiment" which represent about 150 million sensors delivering data at 40 million times per second. There are nearly 600 million collisions per second. After filtering and not recording 99.999%, there are 100 collisions of interest per second. The Large Hadron collider experiment generates more than a petabyte (1000 trillion bytes) of data per year. Astronomical data collected by Sloan Digital Sky Survey (SDSS) is an example of Big Data. Decoding human genome which took ten years to process earlier can now be done in a week. This is also an example of 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.

For Government, Big Data is present for climate simulation and analysis and for national security areas. For private sector companies such as Flipkart and Amazon, Big Data comes up from millions of back-end operations every day involving queries from customer transactions, from vendors etc.

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 fifteen years, several companies abroad 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. For example, available financial data sources include stock prices, currency and derivative trades, transaction records, high-frequency trades, unstructured news and texts, consumer confidence and business sentiments from social media and internet among others. 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 such as portfolio management, stock regulation, proprietary trading, financial consulting and risk management.

Big Data are of various types and sizes. Massive amounts of data are hidden in social net works such as Google, Face book, Linked In , You tube and Twitter. These data reveal numerous individual characteristics and have been exploited. Government or official statistics is a Big Data. There are new types of data now. These data are not numbers but they come in the form of a curve (function), image, shape or network. The data might be a "Functional Data" which may be a time series with measurements of the blood oxygenation taken at a particular point and at different moments in time. Here the observed function is a sample from an infinite dimensional space since it involves knowing the oxidation at infinitely many instants. The data from e-commerce is of functional type, for instance, results of auctioning of a commodity/item during a day by an auctioning company. Another type of data include correlated random functions. For instance, the observed data at time t might be the region of the brain that is active at time t. Brain and neuroimaging data are typical examples of another type of functional data. These data is acquired to map the neuron activity of the human brain to find out how the human brain works. The next-generation functional data is not only a Big Data but complex.

Examples include the following: (1) Aramiki,E; Maskawa, S. and Morita, M. (2011) used the data from Twitter to predict influenza epidemic; (2) Bollen, J., Mao, H. and Zeng, X. (2011) used the data from Twitter to predict stock market trends.

Social media and internet contains massive amounts of information on the consumer preferences leading to information on the economic indicators, business cycles and political attitudes of the society.

Analyzing large amount of economic and financial data is a difficult issue. One important tool for such analysis is the usual vector auto-regressive model involving generally at most ten

variables and the number of parameters grows quadratically with the size of the model. Now a days econometricians need to analyze multivariate time series with more than hundreds of variables. Incorporating all these variables lead to over-fitting and bad prediction. One solution is to incorporate sparsity assumption. Another example, where a large number of variables might be present, is in portfolio optimization and risk management. Here the problem is estimating the covariance and inverse covariance matrices of the returns of the assets in the portfolio. If we have 1000 stocks to be managed, then there will be 500500 covariance parameters to be estimated. Even if we could estimate individual parameters, the total error in estimation can be large (Pourahmadi: Modern methods in Covariance Estimation with High-Dimensional Data (2013), Wiley, New York).

There are concerns dealing with Big Data such as privacy. We will come back to this issue later.

3 When is a data a BIG DATA?

(cf. 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;
(v) The relationship among the columns of the data matrix to check multi-collinearity in the explanatory variables.

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.

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 $n > 1000$.

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 (cf. Fokoue (2015)).

(i)	$n/p < 1$	Information poverty	$n << p, n > 1000$	Large p, Large n	(A)
(ii)	$n/p < 1$	Information poverty	$n << p, n \leq 1000$	Large p, Smaller n	(D)
(iii)	$1 \leq n/p < 10$	Information scarcity	$n > 1000$	Smaller p, Large n	(B)
(iv)	$1 \leq n/p < 10$	Information scarcity	$n \leq 1000$	Smaller p, Smaller n	(E)
(v)	$n/p \geq 10$	Information abundance	$n >> p, n > 1000$	Much smaller p, Large n	(C)
(vi)	$n/p \geq 10$	Information abundance	$n >> p, n \leq 1000$	Much smaller p, Small n	(F)

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

4 Important features of Big Data

(cf. Fan et al. (2013))

For statisticians, Big Data challenges some basic paradigms. One example is the "large p and small n problem". In some human genomics investigations, the researcher might collect data on 100 patients with cancer to determine which genes are responsible for that cancer. Genome-wide studies look at a half-million locations where the variations might or can occur. The number of variables $p = 500,000$ is much grater than the sample size $n = 100$. Similarly in neuroimaging studies, the variables correspond to voxels or regions of interest which often outnumber the participants in the survey. In both situations, 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 subpopulation 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. For example, a mixture probability density of the form

$$\lambda_1 \, p_1(y; \theta_1(\mathbf{x})) + \ldots + \lambda_m \, p_m(y; \theta_m(\mathbf{x}))$$

where $\lambda_j \geq 0$ represents the proportion of j-th subpopulation and $p_j(y; \theta_j(\mathbf{x}))$ is the probability density of the j-th sub-population given the covariate $\mathbf{x}$ with $\theta_j(\mathbf{x})$ as the parameter might be a good fit for a Big Data. In practice, λ_j is very small for some j. If the sample size n is small, then $n\lambda_j$ is also small and hence it is not possible to infer about $\theta_j(\mathbf{x})$.

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. All these issues have been pointed out by Fan, Hau and Lu (2013). 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.

5 Some issues with Big Data

(cf. 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 preprocessing: Most of the Big Data is unstructured and needs preprocessing. With the inherently unstructured data like text data, the preprocessing 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. Examples of such data include audio processing, video processing and image processing. There are other types of Big Data where the input space consists of variables of different types. Such types of data arise in business, marketing and social sciences where 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. When a data set becomes available through some mechanism other than random sampling, there is no guarantee what so ever that the data is representative unless the coverage is full. When considering the use of Big Data for official statistics, an assessment of selectivity has to be conducted. How does one assess selectivity of Big Data?

(vii) No clarity of target population: Another problem of Big Data dealing with official statistics is that many Big data sources contain records of events not necessarily directly associated with statistical units such as household, persons or enterprizes. 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.

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

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

(Ref: Buelens et al. (2014))

For Big Data dealing with the official statistics, there are no approaches developed till now to measure the errors or to check the quality. It is clear that bias due to selectivity has role to play in the accounting of Big Data.

(ix) Use of Big Data in official statistics:

(a) 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 (Leo Breiman (2001), Statistical Science, 16, 199-231). 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.

(b) 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.

(c) Big Data mechanism can be used as a data collection strategy for sample surveys.

(d) 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.

6 Methods of handling Big Data

(cf. 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) Kernelization: In applications such as signal processing, it is always that $p >> n$ in time domain. A ten second audio tape at a 44100 Hz sampling rate generates a vector of dimension $p = 44100$ in time domain and one usually has a few hundred or may be thousand $(= n)$ tracks for analysis. In image processing, there are similar problems of dimensionality with a face of size 640×512 generating a $p = 327680$ dimension input space. In both these cases, it is not possible to use PCA or SVD because $p >> n$. Here one uses the method of kernelization. Given a data set with n input vectors $\mathbf{x_i} \in \mathcal{X}$ from some p-dimensional space, the main component of kernelization is a bivariate function $K(.,.)$ defined on $\mathcal{X} \times \mathcal{X}$ with values in R. The matrix $\mathbf{K}$ given by

$$\begin{pmatrix} K(\mathbf{x_1}, \mathbf{x_1}) & \dots & K(\mathbf{x_1}, \mathbf{x_n}) \\ K(\mathbf{x_2}, \mathbf{x_1}) & \dots & K(\mathbf{x_2}, \mathbf{x_n}) \\ \dots & \dots & \dots \\ K(\mathbf{x_n}, \mathbf{x_1}) & \dots & K(\mathbf{x_n}, \mathbf{x_n}) \end{pmatrix}$$

is called a Gram matrix. The Gram matrix is of order $n \times n$ and does not depend on p. One can compute the eigenvalues and eigenvectors of the matrix K of lower dimension n and analyze the data.

(iii) 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 multi-collinearity 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.

(iv) 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 parallelize 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. However it is ideally suitable for parallelization 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 described 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. This will allow a way of extracting information from a Big Data. One such method is L_1-minimization called LASSO due to Tibshirani (1996). This was used in the field of image processing to extract an image in sharp focus from blurry or noisy data.

(vii) False Discovery Rate (FDR): Another technique that is applied for analysis of Big Data, specially in the genome and neuroimaging research, is the false discovery rate (FDR) suggested by Benjamini and Hochberg (1995). If a study finds 20 locations in a human

genome with a statistically significant association with cancer and it has a false discovery rate of ten percent, then we can expect that two of the 20 discoveries to be false on the average. The FdR does not indicate which discoveries are spurious but that can be determined sometimes by a follow-up study.

(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. For a web company which is trying to predict user reaction and elicit user behaviour such as clicking on an advertisement sponsored by a client, time is important. The web company might have only milliseconds to decide how to respond to a given user's click. Furthermore the model constantly has to change to adopt to new users and new products. The objective of the person who is analyzing the data may not be to deliver a perfect answer but to deliver a good answer fast.

(ix) Privacy and Confidentiality: How to keep privacy and confidentiality in the era of Big Data? Public concerns about privacy, confidentiality and misuse and abuse of individual data is a matter of concern in collection of Big Data. There are ways of masking Big Data. One way is to anonymize the records after they are collected by adding a random noise or to do matrix masking of the data matrix by a known mathematical operation so that individual information is difficult to retrieve. Cryptography is another discipline that applies mathematical transformations to data that are either irreversible or reversible only with a password or reversible only at a great expense that an opponent can ill afford to pay for it.

7 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 presenting the problems involved in storage and computation connected with Big Data in this brief notes.

8 Why Big Data is in Trouble?

Answer: They forgot about Applied Statistics (Jeff Leak, May 7, 2014 "Simply Statistics").
There were articles with titles "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"
"The parable of Google Flu: Traps in Big Data Analysis"

All of the articles listed above and on-line point out the problems of Big Data such as sampling populations, multiple testing, selection bias and over-fitting besides 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)).

References:

Aramiki, E., Maskawa, S., and Morita, M. (2011) Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1568-1576.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

Bollen, J., Mao, H., and Zeng, X. (2011) Twitter mood predicts the stock market. *Journal of Computational Science*, **2**, 1-8.

Buelens, B., Daas, P., Burger, J., Puts, M. and van den Brakel, J. (2014) Selectivity of Big Data, Discussion Paper, Statistics Netherlands.

Fan Jianqing, Han Fang and Liu Han (2013) Challenges of Big Data analytics, arXiv:1308.1479v1 [stat.ML] 7 Aug 2013.

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

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

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

Pourahmadi, M. (2013) *Modern Methods to Covariance Estimation with High-Dimensional Data*, Wiley, New York.

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" *Statistics and Science: A Report of the London Workshop on future of the Statistical Sciences* (2014), pp. 20-25.