Journal of the Chinese Statistical Association
Vol. 53, (2015) 186–202

STATISTICS IN BIG DATA

James J. Chen[†1], Eric Evan Chen[2], Weizhong Zhao[1 3] and Wen Zou[1]

[1]Division of Bioinformatics and Biostatistics National Center for Toxicological Research US Food and Drug Administration Jefferson, Arkansas, U.S.A.

[2]Department of Psychology and Social Behavior University of California, Irvine, CA, U.S.A.

[3]College of Information Engineering Xiangtan University, Xiangtan, Hunan Providence, China.

ABSTRACT

Technological advances in biomedicine, computing, and storage have led to an explosion of digital information and present new challenges in data acquisition, processing, management, transferring, and analysis. The value of big data lies in the analytical use of its information to generate knowledge and action. The goal of big data analytics is to extract knowledge from the data to draw conclusions and make decisions. The purpose of this article is to present a view of prospects of statistics in the context of big data analytics. Statistics is a very old discipline for data analysis and data inference using methods based on probability theory. Statistics and data mining techniques that are useful for big data analytics include: significance testing, classification, regression/prediction, cluster analysis, association rule learning, anomaly detection, and visualization. Statistical analysis provides a scientific justification to move from data to knowledge to action, and is essential to big data analytics. In addition, big data analytics requires good computer skills in information processing and programming skills as well as knowledge expertise that can be applied to the domain of applications. Statisticians can serve a leadership role in the big data movement.

Key words and phrases: data analytics, data mining, data science, Google Flu Trends.
JEL classification: C60, C80, D80.

∗The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration

[†]Correspondence to: James J. Chen
E-mail: jamesj.chen@fda.hhs.gov

1. Introduction

Advances of technologies in biomedicine, computing, and storage have led to explosion of digital information and present new challenges in data acquisition, processing, management, transferring, and analysis. Recently, "Big Data" have received extensive coverage in press and attention among statisticians and many other professions. Big data are data of a massive scale in terms of volume and complexity such that the traditional data processing techniques and software tools become powerless or even useless for some cases, and classical data analysis methods may be ineffective. Applications of big data can be found everywhere nowadays. For example, Kroger uses big data from the customer's shopping patterns and mails coupons specifically designed to an individual household. Walmart uses big data from the customer's behavior and preference to develop models for pricing, predicting, and stocking what products will sell. Facebook and LinkedIn analyze patterns of friendship relationships to suggest other people you may know or might like to know. Netflix and Amazon save customer searches and preferences, and make recommendation of movies to watch and books to read. The National Security Agency uses big data to improve security and to detect and prevent cyber-attacks. Big data algorithms are developed to make decisions in high frequency stock trading as well as in traffic and network management.

Statistics is the science of learning from or making sense out of data. There are many definitions of statistics in the Web dictionaries and books (Hahn and Doganaksoy, 2011). The American Statistical Association offers the following definition: "Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances" (Davidian and Louis, 2012). Statistics has been used around the world by governments, industries, businesses, and academics. Statisticians have made significant contributions to many areas of science, including biology, medicine, agriculture, environmental, etc. In 2008, Dr. Hal Varian, chief economist at Google, said that the statistician is the dream job in the next decade (http : //www.youtube.com/watch?v = D4FQsYTbLoI).

Recently, there are numerous blogs published on the Web discussing big data, data science, and statistics, such as "data science versus statistics", "the end of statistics?", "why big data is in trouble: they forgot about applied statistics", etc. The purpose of this article is to present a view of prospects of statistics and data science in the context of big data.

2. Big Data

Big data is defined as "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications" (http : //en.wikipedia.org/wiki/Big data). However, the term "big data"

is not merely "big", big data can be characterized by the four V's: volume, variety, velocity and veracity.

Volume. Big data relates to data creation, storage, and retrieval. The name "big data" contains the word "big" implicating the amount of data. Therefore, the amount of data should determine whether data can actually be considered as big data or not. The amount of data involves both the number of observations (size) and the number of variables (dimensionality). Currently, big data often refers to the size of data; however, big data in personalized medicine application typically involves a small number in size and a large number in dimension, for example, the UK10k genome project (http : //www.uk10k.org/goals.html) sequenced 1,000 humans around the world, each contained up to 250 million nucleic acids. As data storage technology advances, data volumes are increasing steadily and substantially - ranging from megabytes, gigabytes, terabytes, petabytes, exabytes, to zettabytes of data. It is noted that there is no minimum amount of data to be qualified as a big data. Furthermore, given the speed of increase in storage capacities and software efficiencies, a big dataset today may be just a dataset tomorrow.

Velocity. Velocity refers to the speed of data generation and the analytics process required meeting the demands. Data are streaming in at unprecedented speeds. Big data must be able to capture millions of events per second, and the modelling and algorithms must be fast to reflect the rapid change in a timely manner. For example, stock trading transitions or real-time customer supports require timely retrieval of the right data, manipulation, and execution.

Variety. Big data come in all types of formats: structured numeric data, text documents, audio, video, social media, etc. These data are stored in traditional relation databases and unstructured databases. Data processing, link, transferring, and management among different data formats and systems are challenges and that need to be addressed. In addition, data come from multiple sources. The analytics algorithms, models, and methods for integrative analysis of different data sources need to be developed.

Veracity. Veracity refers to the issues of validity, reliability, and uncertainty. Data come from various sources and in various formats. There are noises, errors, abnormalities and obsolescence in data. To create value from big data, the data must be clean and accurate before modelling and analysis. This is especially important in automated decision-making where the data need to be correct.

There are other characteristics that have been proposed to define big data, such as variability, validity, complexity, visualization, and value. Readers who are interested knowing various V's can search "Big Data V's." It should be noted that the original 3V's (volume, velocity, and variety) are original characteristics of big data that characterize data management information technology (Laney, 2001). Other V's such as veracity and validity, which characterize the value of a dataset, are not big data specific. Big

data is not just about "big"; it is more about insight discovery and value creation from the data. The notion of big data consists of two components 1) data characteristics and 2) data analytics. The above V's describe important data characteristics of big data. Big data analytics refers to the process of uncovering hidden patterns and extracting useful information from the data in order to draw conclusions and make decisions.

Data analytics can be divided into two parts: 1) handling the data and 2) analyzing the data. Handling data typically involves acquisition, processing, and management. Data acquisition refers not only to the retrieval of data, but frequently also includes determining what data to acquire and how to find it. Data processing can include inspecting, cleaning, transforming, normalizing, etc. Data management refers to data organization, administration, and governance to ensure a high level of data quality and accessibility for analytics applications. Since big data involves collecting of massive structured and unstructured data from a wide variety of sources, the traditional database systems and data processing methods are ineffective to develop application tools. New data operation systems and technologies, such as NoSQL and massively parallel processing database systems, Hadoop and MapReduce, have been developed for big data processing/management. These are so called "high-performance analytics" methods using parallel processing with many nodes by breaking down data in different places and often in different systems and combining individual results to generate application outputs.

The value of big data lies in the analytical use of its information to generate knowledge and action. Big data is used to develop business strategies and models in various applications such as analyzing consumer behavior to improve service and predict product demands and trends in retail, detect fraud in banking and insurance, and analyze patient history, risk factor, and real-time monitoring data in healthcare. Big data analytics use high-performance techniques and software tools for analytical applications including predictive modelling, data mining, text analytics, statistical analysis, etc.

3. Statistics

Statistics is a very old discipline for data analysis and data inference using methods based on probability theory. Statisticians have developed theories and new methods for application to almost all scientific disciplines, such as biology, medicine, epidemiology, environmental science, engineering, economy, education, psychology, computer, and many others. Statistical consultants have helped organizations and companies that do not have in-house expertise relevant to their particular projects. Traditionally, statistical analysis involves experimental design, data collection, analysis, interpretation, and drawing conclusions for the study by using probabilistic models and mathematical techniques. Statistics provides the basis for learning from data while taking account of the inherent uncertainty.

Traditionally, statistical analysis can be divided into two types: confirmatory and exploratory. Confirmatory analysis is to test whether the data support a hypothesized model. The hypothesized model is based on theory or previous analytic/experimental research. Statistical analysis of a confirmatory study should be pre-specified in details, such as the hypothesis, test statistic (t-test, non-parametric test, etc.), level of significance, multiple comparison adjustment, etc. Exploratory analysis aims at discovering new knowledge or generating new hypothesis for further or subsequently confirmatory studies. Confirmatory analysis is a top-down structured approach; it starts with a hypothesis aiming to make a conclusive decision. On the other hand, the exploratory analysis is a bottom-up speculative approach into ideas for a hypothesis. A scientific study may start with one or more exploratory experiments and end with a final conclusion by a confirmatory experiment. For example, Phase 1 clinical trials explore safety of a new treatment; Phase 2 trials explore its efficacy; and Phase 3 trials confirm the efficacy in well-planned studies with pre-specified methodology.

Big data analytics typically involve exploratory analysis, such as to investigate the correlations among the explanatory variables or to establish causal relationships between target and explanatory variables.

The hypothesis testing approach is an established standard scientific method for inference on the findings of a scientific study. The analysis is supported by the p-value computed from the data. It is used in randomized clinical trials as the gold standard approach for proving efficacy of a new drug/treatment in regulatory decision. If the p-value is less than the significance level, say 0.05, then the data support the efficacy. In scientific studies, including big data analytics, hypothesis testing is used to determine if an observed finding, say 2 percent increase in sale by a new marking promotion strategy or new modeling strategy, is significant or just a random fluctuation. One problem encountered in big data is that it is virtually impossible NOT to conclude statistical significance since the sample size is huge. Therefore, in addition to the p-value, its associated estimate (e.g., effect size or correlation coefficient) needs to be evaluated. For example, a small, say 0.01 percent, increase in sale may not be meaningful or useful. Big data analytics often involve many explanatory variables, which leads to the well-known multiple testing problem. Many hypotheses are performed to identify or establish relationships between two variables; therefore, the significance level needs to be adjusted so that the overall false positive error rate is controlled.

The recent advancements in high throughput molecular and computational technologies have prompted statisticians to develop data mining techniques for discovery of underlying patterns and trends in large data sets and to make decisions and predictions. Data mining combines the use of computational and statistical techniques to systematically analyze large scale datasets to discover hidden patterns and unexpected occurrences or to develop models and algorithms for predictive analytics. Predictive analytics deals with determining patterns and structures in data and developing models to predict future outcomes and trends. Examples for applications of

predictive analytics include customer relationship management, clinical decision support systems, collection analytics, cross-sell, customer retention, direct marketing, fraud detection, portfolio/product management, risk management, underwriting (Wikipedia, http : //en.wikipedia.org/wiki/Predictive_analytics).

Predictive analytics is a data analytics technique using statistical modeling, data mining, machine learning, and artificial intelligence. Machine learning, a methodology developed to uncover complex patterns and build predictive models by learning from data, is the key component in big data analytics. Machine learning is a branch of artificial intelligence - research to create intelligent algorithms that imitate the step-by-step reasoning that humans use. Machine learning further enables computers the ability to learn from experience.

Statistical analysis focuses on interpretation of model and confidence of inference based on the randomness and uncertainty of the data. Machine learning involves developing models and algorithms for predicting outcomes by learning from the data. Unlike statistics, it is generally not interested in model parameter estimates; it focuses more on predictive accuracy and computational efficiency. Data mining utilizes statistical and machine learning techniques to discover insights in the data or to predict future outcomes accurately. Data mining focuses more on the practical aspects of deploying statistical methods and machine learning algorithms. One major characteristic beyond statistics and machine learning is that data mining involves data management. Data mining and machine learning methods focus on the discovery of knowledge in data and prediction, and less concerns about variability in the data; these two are suitable for large data sets and can be more readily automated. Both machine learning and data mining use basic methods from artificial intelligence and statistics. Artificial intelligence is a logic based approach to reasoning about the data without statistics randomness component and choosing the action that is best to achieve the goal. In sum, data mining, machine learning, artificial intelligence, and statistics, are closely interrelated from methodological principles to theoretical tools. They have a considerable overlap in terms of methodologies for data analytics, e.g. clustering analysis, regression, classification, and prediction. The following section presents commonly used data mining methods, which would include machine learning, artificial intelligence, and statistics, for predictive analytics.

4. Data Mining Methods

Data mining methods are commonly grouped into two categories: supervised learning and unsupervised learning.

Supervised Learning. In supervised learning, each sample is a pair consisting of a set of feature variables and a target outcome variable. The goal of a supervised learning algorithm is to find a function or mapping to connect the set of feature (predictor) variables and the target variable by learning the relationships from the data.

- **Classification** is to assign samples to a pre-defined class. The target variable is labelled representing the class membership of samples. A classification algorithm identifies a predictor set of feature variables and develops a model to predict class membership of new samples with unknown label, based on the predictor set identified. Commonly used classification algorithms are classification trees and random forests, support vector machines, k-nearest neighbors, and logistic regression.

- **Regression/Prediction** is to establish a functional relationship between a set of explanatory variables and target variable. The target variable is an observed outcome variable; it can be binary, frequency count, continuous, and survival time. Regression analysis identifies a set (or subset) of feature variables and builds a model describing a relationship between the target variable and that can be used to predict the outcome of new samples, based on the set of predictors identified. In statistics, the maximum likelihood estimation method is used to estimate model parameters based on the distribution of the target variable. Common regression models are linear regression for continuous data, logistic regression for binary data, Poisson regression for count data, and Cox regression for survival data (McCullagh and Nelder, 1989; Cox, 1984). Logistic regression provides a predictive probability of "success" for each sample. In classification, a threshold value of 0.5 is commonly used as the criterion to classify sample as either success or failure. The main difference between classification and regression is that in classification the samples are assumed to be correctly labelled while in regression the sample class labels are observations from a binary random variable.

Note that the aim of classification and regression is to build a model that can be used to predict or to describe future samples of similar characteristics. They generally do not concern the probabilistic aspect describing the random variation between the predictor and the target variable. More details on supervised learning and common classification algorithms for high-dimensional data can be found in Brieman et al., (1995), Vapnik (1998), Guyon et al. (2002), Hastie et al. (2001), and Kotsiantis (2007).

The primary goal of supervised predictive analytics is to develop a model that can accurately predict future samples of similar characteristics. The most important consideration in model development is to unbiasedly evaluate its "performance." To obtain unbiased estimates, the current sampled data are divided into a training set and a separate test set. The training set is used for model development, and the test set is used for performance evaluation. The key principle is that the test data should never be

used in the model development. A binary classification model commonly is evaluated in terms of sensitivity (the proportion of correct identification of positive samples), specificity (the proportion of correct identification of negative samples), and accuracy. There are several measures and statistical tests for assessment of quality of model (goodness-of-fit), and model comparison and model selection. The most commonly used measures are deviance, Akaike information criterion, and Bayesian information criterion (http : //en.wikipedia.org/wiki/Goodness_of_fit.) More details on the development of classification model and performance evaluation are given by Baek et al. (2009).

Unsupervised Learning. In unsupervised learning, all samples are unlabeled. The objective is to discover structure in the data. In unsupervised learning, there is no target variable and all variables are explanatory variables (http : //en.wikipedia.org/wiki/Unsupervised_learning).

- **Cluster analysis** is the process to identify hidden patterns and structures in data by assigning samples into meaningful and useful clusters and see how they are related. Samples are partitioned into disjoint clusters based on their similarities or differences among attribute variables. Subjects in a cluster should be similar to one another and different from (or unrelated to) the subjects in other clusters. For example, in social network analysis, group mining (user clustering) can identify groups of users sharing common interests and more friendly recommendation can be performed based on the user clustering results. In marketing, customers can be grouped together based their shopping patterns and more interested advertisements can be targeted to specific user groups.

- **Association rule learning** is a method for discovering relationships between samples that occur together among variables in large database. For example, based on the transaction records in super-markets, items which are usually bought together can be found. The most famous example is tales of beers and diapers (http : //blog.patternbuilders.com/2011/03/02/tales − of − beers − and − diap− ers/) which shows that beer and diaper sales are strongly correlated. Based on the results of association rule learning, better rearrangement of items to improve the sales can be conducted in super-markets. The more interested bundle sales can be proposed as well.

- **Anomaly detection** identifies samples that deviate significantly from the general average within a dataset or a combination of data. For example, banking and credit companies develop effective strategies to screen financial transactions for fraud while not disrupting the normal activities.

Visualization uses images, diagrams, or animations to present extracted patterns, structures, and models from the data analytics. Classical graphical techniques such

as boxplots, histograms, charts, scatterplots, 3-D plots, and maps, have served as major tools for exploring data structure (Tukey, 1977; Cleveland, 1994). The matrix visualization techniques for systematically presenting data structures and relationships have been utilized in high-dimensional data analysis over the past few decades (Jacoby, 1998; Chen, 2002; Chen et al., 2008). Dimension reduction techniques are also commonly used for displaying structural information from high-dimensional data to low-dimensional displays. Many visualization techniques, tools and examples for big data analytics are available in Websites using the search words "big data visualization".

Many other data mining, machine learning, and statistics techniques, algorithms, functions, and applications, such as artificial neural networks, pattern recognition, topic modelling, text mining, network analysis, spatial and temporal analysis, are also useful for big data analytics.

Challenges in Data Analytics: Statistics Perspective

Big data is a buzzword that has the potential to help organizations improve operations and make faster, more intelligent decisions. Data science has risen alongside big data and emerged as the science designated for big data. Data science can be defined as the science of applications of computational and statistical techniques for extraction of knowledge from data to make decisions and predictions. Data science is the science of computer-related, data-intensive research, which includes machine learning, statistical learning, artificial intelligence, pattern recognition, visualization, information process and retrieval, high performance computing, etc. The applications are not restricted to big data; nevertheless, big data is an important aspect of data science. On the other hand, statistics has been a major component of data analysis for centuries. In the recent years, statisticians have developed data mining methodologies for the analysis of high-dimensional genomic data and next generation sequencing big data. In November 1997, Professor Jeff Wu characterized statistical work as a trilogy of data collection, modeling and analysis, and decision making. Wu advocated that statistics be renamed data science and statisticians data scientists. Statistics is essential to the big data analytics. There are concerns about the findings of spurious association in the big data analysis due to lack of statistical consideration, Beware the Big Errors of 'Big Data' (http : //www.wired.com/2013/02/big − data − means − big − errors − people/).

In 2009, Google researchers published a remarkable method, known as "Google Flu Trends (GFT)", in Nature (Ginsberg et al., 2009). GFT was able to track the spread of influenza across the US based only on the searches about flu-related queries. The method was much faster than the CDC (Centers for Disease Control and Prevention) flu surveillance system. In other words, the Google researchers showed that GFT was quick, "accurate", and cheap; most impressive, the method was not hypothesis and theory free. The "success" of GFT generated many cheerleaders for big data.

Four years later, there were several reports about the predictability of GFT, such as "Google Flu Trends Gets It Wrong Three Years Running","Why Google Flu Is a Fa−ilure", "Data Fail! How Google Flu Trends Fell Way Short", "Google Flu Trends Fai−lure Shows Drawbacks of Big Data", etc. A recent publication in Science (Lazer et al., 2014) reported that GFT wildly overestimated the number of flu cases in the United States in the last few years. There were discussions about what went wrong with GFT. One problem is the predictive model was based on correlation analysis, that is, the correlation between the search query and report case. The GFT method did not select only those queries that were associated with the spread of flu; rather it used all queries about flu, such as flu symptoms and flu vaccines, resulting in spurious relation. Since the publication of the Science's article, there were concerns about big data and the role of statistics in big data, e.g., "Google Flu Trends: The Limits of Big Data" (http : //bits.blogs.nytimes.com/2014/03/28/google − flu − trends − the − limits − of −big − data/?_r = 1);"Big data: are we making a big mistake?" (http : //www.ft.com/cms/s/2/21a6e7d8 − b479 − 11e3 − a09a − 00144feabdc0.html), "Why big data is in trouble: they forgot about applied statistics" (http : //simplystatistics.org/2014/05/07/why − big − data − is − in − trouble − they − forgot − about − applied − statistics/).

Statistics is a science for knowledge extraction and making sense out of data. Statistical research focuses on data collection, modelling, and interpretation. Almost all statisticians are trained with good mathematical skills and logical thinking. Applied statisticians can contribute to various stages of an experiment in either exploratory or confirmatory studies, including the development of a study protocol, experimental design, measurements for target and exploratory variables, data collection, statistical modelling/procedure, data monitoring, data analysis and interpretation, and statistical report. Big data analytics typically involves exploratory analysis of retrospective or observational data, such that the study hypothesis may be vague or none, non-randomization data, uncontrolled confounding factors, etc. The analysis often is to investigate correlations among explanatory variables or to establish causal relationship between explanatory and target variables. The GFT method conceptually is appropriate. Building a prediction model from a given dataset requires several important considerations: data collection process, selection of informative variables, model predictability (ability to predict the samples generated from a similar process), model generalizability (ability to generalize the model to predict samples generated from different locations and time points), heterogeneity and variability of current and new samples, etc. The GFT model included un-related variables (search queries) in the predictive model, there were temporal effects and unexpected confounding factors (the widespread flu epidemics in December 2012 provoked searches by heathy people). These factors will hinder the performance of a predictive model. The model was a "success" given that no CDC surveillance data were used. A further analysis reported that combining CDC's and Google data have improved predictive accuracy.

Big data tends to be complex and messy. The sampling schemes and data collection process are unknown or unclear. There are considerable challenges in the methodology for big data analytics. Over the years, statisticians have developed methods to address many problems and issues encountered in data analysis. These problems and issues are also present in big data analytics. Common known issues include sample representation of the target population, identification of confounders and stratification, robust procedures, outlier identification and treatment, statistical significance versus practical significance, etc. In high-dimensional data, the issues are over-fitting, spurious clusters, spurious correlations, multiple testing, feature selection, etc. Statistical methodology provides a scientific justification to move from data to knowledge to action.

Challenges for Statisticians

Big data analytics requires methodologies different from common statistical methods. Big data requires efficient computer skills in information processing and programming. Although some statisticians have done great work in algorithm and programing, most statisticians are not equipped with information skills and dealing huge data. The information skills involve data extraction, transformation, loading, data structures and storage, cloud computing technology, and programming, which involves algorithm development, system implementation, and web application. The tools include Python, Git, Flask, Javascript, MySQL, and Hadoop. Furthermore, statistical analysis typically is a top-down or goal-oriented processing. "Statisticians have grown accustomed to well-structured problems, often using a single technique in a delimited domain" (Jordan and Lin, 2014). However, problems in big data are vaguely defined, and analytics strategies frequently use bottom-up processing. For example, development of marketing strategies to increase product sales and predict product demands requires piecing together several small prediction models involving inventory, cost analysis, competing products, customer behavior, etc.

What computing skills should a statistician learn?

There are four critical areas of computing skills, roughly corresponding to the lifecycle of a big data project: data acquisition, data processing, data management, and data mining/analysis. Each of these areas can be further divided into two complementary components: a *technical expertise* component and a *functional knowledge* component. Technical expertise refers to specific skills such as learning a particular programming language, database language, or distributed computing framework. The functional knowledge level involves understanding broader principles such as communication protocols, search algorithms, and concurrent algorithms. For those looking to acquire just the skills needed to execute a specific project, the focus can be almost solely on the expertise component, given that, realistically, time, attention, and energy are often

limited. Some degree of knowledge will likely still be acquired anyway.

In general, however, flexibility is key. Given the variety of data-each with its own acquisition, processing, management, and analytical considerations-both within a single big data project as well as between projects, ultimately, one can expect that the expertise and knowledge associated with each of these components will vary widely.

Data acquisition. Data acquisition is arguably the area that requires the most content knowledge. Here the data collection should be driven by substantive questions and some degree of content knowledge is a necessity. Especially given the volume and velocity that define big data, data acquisition must be selective. Storage capacity, acquisition rate limits, and other constraints prevent acquiring every piece of data that may potentially be of interest.

Acquiring social media data from sites like Twitter or Facebook requires knowledge of programming/scripting languages, such as Python or R. Generally speaking, programs written in almost any programming language can implement the steps required to communicate with the servers hosting the data to be acquired. However, some languages have active communities that have developed open-source software packages that already implement common tasks. Crucially, because many of these software packages are widely used, they are almost invariably more robust-previous users have already done a great deal of bug fixing and validation.

Knowledge of web communication protocols (e.g., HTTP), JSON, and web scripting/coding (e.g., html, PHP, JavaScript) is needed. This knowledge is important to the understanding of how to communicate with online servers, etc. For example, given the velocity of big data, acquiring data across networks may reach bandwidth limits and encounter random network errors that, although minor for smaller datasets, become problematic at the big data scale. Accounting for these sorts of issues requires understanding core communication protocols.

Data Processing. Data processing involves making acquired data storable and useable. A key challenge here is to transform unstructured data (e.g., text, video) into numeric data. Technical skills and functional knowledge are closely intertwined in data processing. For example, in processing text data, programming languages such as Perl and Python are well-equipped to process text. A variety of open source Python-based natural language processing software packages are available.

Some of data processing work will be customized to a particular project. For example, with natural language processing, approaches such as Latent Dirichlet Allocation (LDA) (Blei, et al., 2003) operate on a text level of analysis. Thus, accumulated data must be divided into texts. The definition of "text" for a given research project must be defined theoretically, which requires content domain knowledge. As a technical matter however, one must also understand how to practically create these texts. The popular LDA program Mallet requires that each text be a separate text file in the file

system. Traditionally, this could be done manually and menially by research assistants. However, given its sheer volume, this would be essentially impossible with big data. Creating these separate text files thus requires the technical knowledge to program a script to execute this task.

Data management. Storing and retrieving big data is a big challenge. Broadly speaking, this involves the technical knowledge of database management. Database technologies such as NoSQL are frequently used. Managing data often involves a number of hardware concerns, such as whether to store the data in the cloud or on one's own hard drives. In addition, understanding how the data will be analyzed is essential. When running an LDA analysis on individual text level of analysis, for example, for extremely large volumes of data, storing each text as a separate file may not be feasible. Accessing storage media such as hard drives and retrieving entries from databases both typically take much longer than accessing data already loaded into memory. Here, as with other big data projects, an effective solution requires understanding how the theoretical and analytical requirements of delineating the data into individual texts must be combined with the technical expertise of how to achieve this given the typical big data challenges.

Data mining/analysis. The main difference in the analysis of big data is that it may require the use of distributed processing. One of the main challenges in big data analytics is in handling the scale of the required computing power through techniques such as distributed processing. In many ways, the required technical expertise is an extension of that required to analyze the given type of data in a non-big data context. These parallel, distributed algorithms often require understanding how the analysis can be divided and run on separate processors, after which these results must then be combined. These implementations are often based on the MapReduce model (Apache Hadoop is a popular open-source implementation) or SQL, in combination with programming languages such as R or Python.

In practice, all areas overlap and work together simultaneously. In designing data acquisition programs, one must already have established how the data will be processed (extracted and transformed) and how the data will be managed (where to put the data). Processing data requires understanding where the data came from and how it will be stored. Data processing and data management roughly correspond to "extract-transform-load" issues in more traditional data management-turning data into the preferred format and storing them in an efficient manner.

5. Discussion and Conclusion

The skills necessary to handle big data require more than proficiency in computer science, statistics, and other data technical skills. The most important component is the knowledge expertise that can be applied to the domain of applications. Big

data analytics requires domain expertise to identify important problems and effectively deliver the solution. Big data scientists must be innovative thinkers and modelers, and skilled programmers, who understand the problem well and can formulate key questions and implement the solution to guide decision. In addition, they should have the communication and interpersonal skills to be able to effectively present the findings and recommendations to the management and practitioners, and to work successfully with team members.

Statisticians should be the leaders of the big data movement. However, the role of statistics in big data appears to be minimal to non-existent. Besides lack of information processing expertise, several reasons have been discussed for this disconnect (Davidian and Louis 2012; Jordon and Lin, 2014). First, the contribution of statistics to the society has not been widely recognized due to a lack of understanding by the general public regarding what statistics is about. Many statisticians often treat their research as mathematical exercises without addressing its applicability. Also, many statistics practitioners use available software packages to perform statistical analysis and generate p-values without full insight into the model and the process of data analysis. Of course, there are statisticians who have done great work in applied statistics and omic technologies, but their contributions are not sufficiently recognized. Second, many big data projects are high impact and involve a large team, most statisticians work on smaller projects as collaborators or consultants. Statisticians can take leadership roles as principal investigators and formulate problems in a statistical framework, instead of as co-investigators who only respond to problems identified by other disciplines.

There are challenges for those working on so-called applied statistics in academic institutes. A pioneering work of novel methodology that can solve the most critical issues in a discipline other than statistics would have difficulty to be published in a top statistical journal without making strong efforts to evaluate other methods (even if other methods may not be appropriate for the problem) or to convince the referees of the significance of the work in the subject's area. This problem is not encountered in other fields. Methodologies involving theoretical constructions for narrowly defined problems regarding popular statistics topics can be found in all the major journals. In addition to publication challenges, it may be difficult to fund applied statistical work through grants under the statistics/mathematics program. Most importantly, statisticians who focus on applied problems like big data may experience difficulties with tenure/promotion/hiring because the evaluation criteria do not favor interdisciplinary work. Committees or members might not understand the potentially broad and significant impact (this is not a problem for tenured full professors). On the other hand, Master's degree programs in statistics have been very successful for students in job placement. There is a lack of statisticians and data scientists with the computational skills and statistical training to analyze and interpret data in the world of big data. In closing, "Statistics has been the sexy job of the last 30 years. It has just taken awhile for organizations to catch on." (Goodnight, 2011).

References

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 996-1022.

Breiman, L. (2001). Random forest. Mach. Learning, **45**, 5-32.

Brieman, L., Friedman, J., Olshen, R., Stone, C., Steinberg, D., et al. (1995). CART: Classification and Regression Trees: Stanford, CA.

Chen, C. H. (2002). Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. *Statistica Sinica*, **12**, 7-29.

Chen, Chun-houh., Wolfgang, H., ardle. and Antony, Unwin. (2008). Handbook of Data Visualiza-tion. Berlin, Germany: Springer.

Cleveland, William S. (1994). The Elements of Graphing Data (Revised Edition). Summit, NJ: Hobart Press.

Cox DR, Oakes D. Analysis of survival data. Chapman and Hall, London, UK (1984).

Davidian, M. and Louis, T. A. (2012). Why statistics? Science 336, 12 (Apr 6), doi: 10.1126/science.1218685.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature 457, 1012-1014 (19 February), doi:10.1038/nature07634).

Goodnight, G. (2011). Executive Edge: Statistics make the world work better. http://www.analytics-magazine.org/july-august-2011/354.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning 46: 389-422.

Haha, G. J. and Doganaksoy, N. (2011). A Career in Statistics: Beyond the Numbers. John Wiley & Sons.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer.

Uesaka, H. (2007). Sample size allocation to regions in a multiregional trial. *Journal of Biopharmaceutical Statistics*, **19**, 580-594.

Jacoby, William G. (1998). Statistical Graphics for Visualizing Multivariate Data. Thousand Oaks, CA: Sage.

Jordan, J. M. and Lin, D. K. J. (2014). Statistics for Big Data: Are Statisticians Ready for Big Data? ICSA Bulletin 26, January, 2014, 58-65.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification Techniques Informatica, **31**, 249-268.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner. (6 February 2001).

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science 343, 1203-1205. (14 March 2014)/ doi: 10.1126/science.1248506).

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Model, 2nd Edition. Chapman Hall, London.

Tukey, John W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley Publishing Company.

Vapnik, V. (1998). Statistical learning theory: Wiley, New York.

[Received March 2015; accepted May 2015.]

Journal of the Chinese Statistical Association
Vol. 53, (2015) 186–202

大數據時代的統計學

陳章榮[1] 陳逸凡[2] 趙衛中[1 3] 鄒文[1]

[1]美國食品和藥物管理局 國家毒理研究中心, 阿肯色州, 傑斐遜市
[2]美國加州大學歐文分校, 心理和社會行爲系
[3]湖南湘潭大學信息工程學院

摘 要

生物醫學計算機科學以及數據儲存技術的發展引發了數據信息的大爆炸, 也帶來了數據的獲取處理管理傳輸以及分析方面的挑戰。大數據的價值在於對其信息的分析結果產生的新的認識和行動。大數據分析的目標是爲了從數據中獲取知識來得出結論和作出決定。本文展示了統計學在大數據分析時代的應用和展望。統計學是一門古老的科學, 它採用基於概率論的方法進行數據的分析和推論。對大數據分析有用的統計學和數據挖掘方法包括: 顯著性測試, 分類, 回歸/預測, 聚類分析, 關聯式規則, 異常檢測和視覺化。統計學分析爲從數據到知識再到行爲的過程提供了科學證明, 是大數據分析所不可或缺的。另外, 大數據分析需要較好的處理信息的計算機技能, 程式編程能力, 以及具有各種應用領域的專業知識。統計學家能夠勝任大數據浪潮中的領導作用。

關鍵詞: 數據分析, 數據挖掘, 數據科學, 穀歌流感趨勢。
JEL classification: C60, C80, D80.