Data Technology — As a New Concept for Application of Statistics

Sung H. Park\textsuperscript{1} and Moon W. Suh\textsuperscript{2}

\textsuperscript{1}Seoul National University and \textsuperscript{2}North Carolina State University

Abstract: Data technology (DT) is newly defined apart from Information Technology (IT). Major territory for DT is outlined along with its roles for the information society. DT is concerned primarily with data collection, analysis of data, generation of information and creation of knowledge. On the other hand, IT is mainly concerned with transmission and communication of data, and development of engineering devices for information handling.

The roles of DT for knowledge creation from raw data are explained step-by-step. In order to exploit the DT concept for solving practical problems in a process or a product, a 6 step working flow is suggested. Loss due to poor DT is mentioned with two examples. In addition, e-Statistics is proposed as one of major vehicles for promoting the roles of DT. Overall, DT is explained as a new concept for broadened application of statistical science as a key technology for impacting global competition in the 21st century socio-economic environment.

Key words: Data mining, data technology, e-Statistics, information technology, knowledge triangle, working flow of DT.

1. Introduction

The authors are of the view that the contribution of statistical methods during the last three decades to the socio-economic development and human welfare has not been widely recognized due to flawed packaging and marketing of what statistics is all about. Many statisticians, particularly those with strong mathematical background, often treat their research as mathematical exercises, which results in inevitably imposing limitations on development of statistics. Also many practitioners, on the other hand, are happy with existing software packages. If we continue to do the business as usual, the statistical science will remain as an invisible science to the public. A view similar to this was expressed by Marion L. Straf, the Past President of American Statistical Association in his Presidential
address delivered on August 13, 2002 in New York City. The explosive growth in the number and speed of decision mechanisms and the accompanying analytic needs have stunned the statistical community as a whole. If data-analytic techniques become dominant in other fields, traditional statistics may become unnecessary. In spite of the rapid advance in the speed of computing, the data explosion and complexities of information-driven society made the classical tools in statistics almost powerless at the presence of infinite amount of data generated in continuous time domain, often as a form of signal. At the same time, complex and yet practical decisions are demanded instantaneously for all aspects of modern life for economic, political, and personal gains.

The phenomenal growths in data storage, retrieval, communication and graphical representation capabilities have revolutionized the use of numbers and data aided by the enormous increase in the computing speed. The classical theories on sampling and estimation, for example, are often found inapplicable at the presence of the entire population in a signal form or when the exact parameters, instead of their estimates, are readily available. Statisticians and others in analytic disciplines are overwhelmed by the abundance of data and lack of tools for converting the massive amount of data and signals into useful information and decisions. Friedman (2001) mentioned that most statisticians seem to agree that Statistics is becoming relatively less influential among the information sciences, and if data analytic techniques originating in other fields become dominant, our field will correspondingly suffer. For these reasons, we propose to define a new technology by repackaging what we have been offering to the society and scientific communities. The contents and their qualities are always important but selling the product must be given a serious re-evaluation. In retrospect, we have not done the packaging and marketing of data analysis correctly during the last three decades. The term technology always has commercial implications with applications as the key. Statistics has been understood and treated adequately as a serious discipline of science and has been applied to all aspects of science and society likewise. However, it was much like the way physics and mathematics have been applied to all sciences and societal issues, but not as a technology engine for solving practical problems of the society in a systematic fashion. For this reason, we propose Data Technology as a new and integrated concept for application of statistics in order to impact the society in the future.

Key technological engines for moving the 21st century are said to be IT (information technology), BT (bio-technology), NT (nano-technology), ET (environment technology), ST (space technology) and CT (culture technology). Note that all the XTs (IT, BT, NT, ET, ST and CT) have their own hardware technology as fundamental developing unit. What we have yet to recognize is a technology so deeply imbedded in every technology that deals with quantitative decision pro-
cess whether it is individual, organizational, societal, or scientific. This is what we will call Data Technology, or DT. Even though DT is a software oriented discipline without the hardware core, the authors believe that DT should be a leading technology in this knowledge-based information century. While some had inferred DT to electronic means for communicating and transporting data (such as “Mobile Data Technology”), it was without reference to a broad definition of what this paper intends to promote. Park (2001a, 2001b) first discussed the economics of DT and its applicability.

DT is defined as a scientific methodology that handles the following by use of statistical and probabilistic techniques:

- systematic collection, storage and retrieval of required data,
- refinement and object-oriented analysis of data,
- relations between the system of phenomena under study and the possible collection mechanisms,
- conversion of data into information,
- statistical and computational modeling based on data,
- formation of inferences and relevant knowledge, and
- diagnosis of the present states and prediction of future events.

Note that the definition of DT essentially reflects the flow of knowledge creation through data collection and analysis, information creation, statistical modeling, and future prediction.

We believe that DT, as defined, would be an essential technology for promoting application of existing and new statistical methods imbedded on data for all aspects of modern society. They include organizational management, economic decisions at all levels and dimensions, creation and management of political and social systems, and forecasting for the enhancement of global competitiveness for business, industry and government. The importance of DT is expected to grow fast in this knowledge-driven information society.

2. Data Mining and Next Generation of Statistics

Data Mining (DM) has emerged as technology of a sort during the last decade and it has served a good purpose as an application tool. DM is a software technology that deals with the exploration and analysis of large quantities of data to uncover useful patterns. DT is actually a process for retrospectively
exploring useful information from business data by automatic or semi-automatic means. Of the application papers dealing with DM, many have appeared in journals outside the traditional domains (for instance, Clark (1997), Domingo et al. (2002), Kohavi et al. (2001), Matheus et al. (1993) and Ridgeway et al. (2003)) and have become a fashionable session topic for almost all scientific conferences. Chatfield (1995) made a serious effort in expanding the traditional statistical inferences to DM under uncertainties in picking valid models. Sen (2002) mentioned that knowledge discovery and DM is becoming a dominating force with bioinformatics as a notable example, and coping with data analysis may be the future goal of statistical science.

Yet, DM has not been properly blessed as an area of technology encompassing the broad spectrum of DT as we have defined above. Besides, the connotation that “mining” provides is too narrow and misleading. We believe that DM has not been promoted well as an array of application tools using statistical methods.

About 30 years ago, Healy (1978) said “Statistics may itself be considered as a technology rather than as a science”. Recently, Straf (2003) mentioned that “Statistics is the generation and effective use of knowledge from data”, and “Statistics, as a technology, is a fundamental and invaluable part of the infrastructure of other sciences”. We agree with Healy and Straf. We believe that DT will play a major role for application of statistics, and in that viewpoint, it is evident that the next generation of statistics will have more weight on technology rather than on science.

3. DT in Relation with IT

In information age with abundance of data, DT thrives with rapid expansion of IT. Many people may think that DT is a subset of IT. DT is not a part of IT although there has been much confusion as to what IT is and where it should stop. In order to promote DT, it is important to clarify what IT means. Our belief is that IT can be defined as an engineering system that handles:

- transmission and communication of data and information,
- presentation and control of information and knowledge,
- manufacturing of electronic systems/devices for information transmission and communication,
- manufacturing of computer-based networking instruments, and
- systems management dealing with data, information and networks.
Table 1: Comparison of DT and IT

<table>
<thead>
<tr>
<th>Contents</th>
<th>DT</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Characteristics</td>
<td>Software-oriented, Scientific methods for data analysis, Modeling and forecasting</td>
<td>Systems-oriented, Engineering system for transmission and forecasting communication of data and Information</td>
</tr>
<tr>
<td>Major Products</td>
<td>Software such as DBMS, CRM, SPC, ERP, Statistics, Data-mining, Simulation, Cryptography</td>
<td>Electronic devices and auxiliary devices for communication, measurement and control</td>
</tr>
<tr>
<td>Major Study Fields</td>
<td>Statistics, Mathematics, Information science, Computer science, Decision sciences</td>
<td>Computer engineering, Electronic engineering, Communication engineering, Control and systems engineering</td>
</tr>
</tbody>
</table>

IT should not be viewed to include any analytic systems or decision support mechanism imbedded on data although communication or transmission of data and decision support functions often occur simultaneously. Mixing the two does not benefit IT as it blurs its territory. The differences between DT and IT can be seen in the information flow as shown in Figure 1.

DT is concerned primarily with data collection, statistical analysis of data, generation of useful information for decision, and creation of necessary knowledge from information. On the other hand, IT is mainly concerned with transmission and communication of data, information and images, and development of engineering devices and computers for information handling. Also IT is concerned with engineering tools for knowledge management. Ideally speaking, DT should be the “working engine” of IT. Naturally, the two thrives on each other as a growth in IT creates new needs in DT, and the vice versa. In addition, DT is software oriented as IT is more hardware and systems oriented. Table 1 shows the
differences between DT and IT in terms of their characteristics, major products, and study fields.

To explain the differences between DT and IT clearly, Figure 2 is provided, where DT is the root of a big tree and IT the trunk. The fruits are the final products of DT and IT. The key message is that DT and IT combined could produce good final products, and DT is an essential infra-structure of IT.

![Figure 2: Relationship and Differences between DT and IT](image)

4. Knowledge Triangle

Life we live in the 21st century would be characterized by the knowledge-based information society. We construct a “knowledge triangle” as shown in Figure 3 in which DT and IT simultaneously play important roles. The concept of knowledge triangle was briefly introduced in Park (2003a). To create knowledge, we need 4 steps of DT, in which DT and IT help each other.

Note that in the initial step, raw data (facts) may be created by a mechanism which is designed into the system under study. At this step, the acquisition, storage and retrieval of data are managed. This step can be called 'management DT'. For collection of data, some statistical methods such as sampling design and design of experiments are necessary. In the second step, we may need Data Mining, statistical data analysis, and statistical prediction. In this step, since information is generated and multiplied, this step can be called 'multiplication DT'. In the third step, knowledge generation is visualized through the use of
IT and DT. Hence this step can be called ‘execution DT’. Finally, knowledge valuation is necessary to create wisdom, which results in profit generation. This last step can be called ‘valuation DT’.

Figure 3: Knowledge triangle

In each step, the following activities are to be implemented.

Table 2: Major Activities in Each Step of Knowledge Triangle

<table>
<thead>
<tr>
<th>Steps</th>
<th>Basic Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>System of phenomena and mechanism study</td>
</tr>
<tr>
<td>(management DT)</td>
<td>Measurement acquisition of data, Storage and retrieval of data, Data refinement, Sampling design, Design of experiments, Meta data management</td>
</tr>
<tr>
<td>Step 2</td>
<td>Data analysis and statistical modeling, Data mining, Application scope refinement, Diagnosis and control, Prediction modeling</td>
</tr>
<tr>
<td>(multiplication DT)</td>
<td></td>
</tr>
<tr>
<td>Step 3</td>
<td>Knowledge generation, Decision support review, Enhanced modeling</td>
</tr>
<tr>
<td>(execution DT)</td>
<td></td>
</tr>
<tr>
<td>Step 4</td>
<td>Wisdom generation from information clustering, Valuation and profit generation</td>
</tr>
<tr>
<td>(valuation DT)</td>
<td></td>
</tr>
</tbody>
</table>
5. Loss Due to Poor DT

Poor and inadequate DT can result in a big loss to a business, to a society or to a nation. Two well-known examples are described below.

(1) Quality and Efficiency of Political Decisions

Wrong policies and inefficient legislative and administrative processes are mainly due to lack of proper data, information and projection into the future. Often, the issues are misrepresented and impacts exaggerated due to lack of statistical inference. Economic policies, trade agreement, planning the infrastructure for the future (for example, the number of elementary school classrooms needed for the next 30 years). The political decision process remains perhaps a most primitive form in many parts of the world including the most developed countries like USA. Non-productive political disputes, as prevalent in Korea as in other countries, hamper advances in such other areas as trade and commerce, industry, education, transportation and culture.

(2) Cost of Poor Quality

The cost of poor quality (COPQ) is the total cost incurred by poor quality and poor management. The COPQ can be divided into visible COPQs consisting of prevention costs, appraisal costs and failure costs, and the hidden COPQs that originate from lost opportunity costs and lost goodwill costs, and costs associated with frequent design change and project rework. Juran (1988) estimated that the COPQ for most companies in the world is about 20-40% of total sales value. Conway (1992) even claimed that in most organizations 40% of the total effort, both human and mechanical, is wasted, and the waste can be eliminated or significantly reduced. For textile industry, Suh (1992) found that net profit for the publicly owned companies in U.S. ranged 1.5 to 3.5

In a typical quality control environment, it is often the case that either there exist adequate amount of data with no analytic capabilities for corrective actions or no useful data at all for effective decision making. If DT is properly employed along with creation of useful data by statistical methods such as designed experiments and sampling schemes, the COPQs can be reduced in industry as evidenced by the success stories behind Six-Sigma applications.

6. Working Flow of DT

In order to exploit the DT concepts for solving practical problems in a process or a product, we can use the following 6 step working flow, which is DMAMPC as follows.
1. D (Define): Identification of the process or product that needs analysis, usually aimed at improvement.

2. M (Measure): Collecting the data (dependent and independent variables) and making the necessary evaluation.


5. P (Predict): Predicting future values of dependent variables at a set of independent variables.

6. C (Control): Controlling the dependent variables.

The DMAMPC flow is a structured and time-sequential way used for solving a practical problem in industry. Note that in Six Sigma, the most well known management strategy, the so-called DMAIC (Define, Measure, Analyze, Improve, Control) flow is used. The difference between DMAMPC and DMAIC is that the Model (M) and Predict (P) steps are used in DT instead of the Improve (I) step in Six Sigma, which implies that the activities of modeling and predicting are emphasized in DT. We do not elaborate on Six Sigma here. However, for references on Six Sigma, see Harry (1998), Magnusson et al. (2000), Pyzdek (2001), Park (2003b) and so on. Note that in the ‘Model’ step, we can think of two cultures of statistical modeling. One is stochastic data modeling, and the other is algorithmic modeling. The former mainly utilizes regression analysis, but the latter uses decision trees and neural nets. Breiman (2001) explained the difference in detail.

7. e-Statistics and DT

Recently, the term e-Statistics has emerged (Devillers (2002), Park and Suh (2002)) in a loose way with implications for applying statistical methods under data-rich electronic environment. As a rough cut, e-Statistics may be defined as a statistical methodology dealing with all aspects of electronic processes for data storage and retrieval, data refinement and statistical analysis, statistical model building, simulation and statistical prediction, and generation of knowledge base and inference.

From this definition, we may see clearly that e-Statistics is a most important subset of DT. With sound application of e-Statistics, DT and IT can expand their scope and utility in order to attain their full values. Needless to say, e-Statistics should play an important role for the software industry.
Proliferation of e-Statistics within DT would be an important new vision for the application of old and new statistical tools under a data-rich, on-line, quick-response environment for service, manufacturing, government and all scientific activities in the future. Optimal decisions made electronically under dynamic environment will no doubt enhance and revolutionize the business practice, reform in government affairs, enhance R&D efficiencies for all sectors, and improve quality of human life in all aspects. In sum, we believe DT and e-Statistics are the exciting new companions, one as a tool and the other as a newly blessed domain of major technology for the information age.

8. Concluding Words

We have proposed and defined Data Technology (DT) as a major technology to enhance the quality and efficiency in all aspects of human life through expanded use of statistical methods, and to promote global competitiveness of business and government accelerated by electronic (automatic) decision processes under data-rich, information-driven environment. The border-less and limit-less economic competition with free-trade agreements certainly warrants a new technology as a winning tool, namely, Data Technology. We believe that, if statistics is well equipped with the DT concept explained above, statistics can become a leader in the modern science and technology to promote human quality of life.

Acknowledgement

We thank the referees who gave us valuable suggestions and comments that contributed a lot to improve the quality of this manuscript significantly. This work was partially supported by the second stage Brain Korea 21 Project.

References


