SPECIAL CONVOCATION

and the

FIFTY-NINTH ANNUAL CONVOCATION

24 DECEMBER 2014

ADDRESS BY

GUEST-IN-CHIEF

PROF. SANKAR K. PAL

Distinguished Scientist and Former Director
Indian Statistical Institute, Kolkata

JADAVPUR UNIVERSITY

KOLKATA 700 032

INDIA

SPECIAL CONVOCATION

and the

FIFTY-NINTH

ANNUAL CONVOCATION

24 December 2014

ADDRESS BY

GUEST-IN-CHIEF

Prof. Sankar K. Pal

Distinguished Scientist and Former Director

Indian Statistical Institute, Kolkata

JADAVPUR UNIVERSITY
KOLKATA 700032
INDIA

Convocation Address

by

Prof. Sankar K. Pal*

BIG Data:
Challenges, Opportunities and National Relevance

His Excellency, Shri K N Tripathi, Hon'ble Governer of West Bengal, and Chancellor of Jadavpur University; Prof. Abhijit Chakrabarti, Vice Chancellor of the University; Members of the Court and Distinguished Faculty; Degree Recipients and their Parents; Beloved Students; Guests; Representatives of the Media, Ladies and Gentlemen:

It is a great privilege and honour for me to participate in the Special Convocation and 59th Annual Convocation of Jadavpur University, Kolkata, as the Guest-in-Chief and to deliver the Annual Convocation Address. This event is especially significant to me as it is my first convocation address to be delivered in my own state, and that too in an internationally renowned university of the country which has been rated as a "Five Star University" in India by the National Assessment and Accreditation Council.

I take this also as an opportunity to acknowledge the contribution that this university has made to my research career, although I am not a member of its alumni. For

*Distinguished Scientist and Former Director Indian Statistical Institute, Kolkata example, some of the PhD students that I have graduated from ISI are the product of your university, while several others have obtained their PhD degree from this university under my supervision.

Before I go to the main part of my speech, let me first offer my heartiest congratulations to the graduating students on this important occasion when they have completed their studies at Jadavpur University, and are going to start a new phase in their professional lives. I also congratulate the distinguished faculty members of the Institute for nurturing the students and shaping the young minds. In these days when excellence is increasingly becoming a rare commodity, those of us who are able to be a part of institutions of excellence should consider ourselves to be fortunate indeed.

When I was requested by your university to deliver the Annual Convocation Address, the date of the event clashed with one of my pre-scheduled overseas programmes. Despite this, I accepted the invitation by curtailing my trip, because it was from a University that I have always been proud of, as an Indian, in general, and a Bengali and a Calcuttan in particular. However, I was not sure what the theme of my address would be. After some deliberation, I decided to share some thoughts on my research experience of about four decades, including some recent efforts which could be beneficial to the prospective scientists, technologists and engineering researchers today. Moreover, we are in the midst of what is popularly called the Information Revolution and are living in a so-called World of Knowledge where great volumes of data are constantly being generated all around us. Accordingly, I have decided upon a contemporary topic like "Big Data: Challenges and National Relevance" which is a forefront research area, has enormous relevance in the context of national development, and is a "must-know" subject to any applied scientist, technologist or practitioner dealing with data. It is also likely to be a subject of interest to a common man. I shall be mentioning, in brief, the A, B, Cs of Big data, challenges and issues, uncertainty analysis, relevance for social developments, opportunities, and our national initiatives. These will be followed by certain issues that may concern you, and some advice. My thoughts are based on my experiences as a researcher in pattern recognition, machine intelligence and soft computing, in a broader sense, and working in different premier institutes in India and abroad including the Imperial College, London ; NASA Johnson Space Center, Houston; University of California, Berkeley; University of Maryland, College Park, and US Naval Research Lab, Washington, DC.

What is Big Data?

It is similar to data as we know it, except that it is substantially bigger in scale, diversity and complexity. However, having the data bigger requires new architectures, techniques, algorithms, tools and analytics to manage it and extract hidden knowledge from it. In doing so, one needs to solve new problems that have cropped up, as well as old problems in a better way. Big Data is characterised by the four Vs, namely, *volume* (terabytes to zettabytes, 10^{12}-10^{21}), *variety* (structured, semi-structured and unstructured; heterogeneous), *velocity* (high rate of change, dynamic, streaming) and *veracity* (uncertainty and incompleteness).

Why Growth of Big Data?

The main reasons are

- Increase in storage capacities
- Increase in processing power
- Availability of data

Data storage has grown significantly after 2000 due to digitalization of analog data, that is, consequent to a shift from analog storage to digital storage. Computation (processing) capability/power has increased sharply mainly because of the use of mobile phones and video game consoles. According to a recent report in *Science* (2011) by Hibert and Lopez, there has been a sharp rise from 25% in 2000 to 94% in 2007 in digital storage among overall storage. Handling-capacity of information by global installed computation has increased from <0.001x10^{12} million instructions/sec in 1986 to 6.379x10^{12} million instructions/sec in 2007.

The different sectors from where the Big Data is available include: government, communication and media, discrete/ process manufacturing, banking, health-care providers, securities and investment services, education, transportation, insurance, resource industries, and construction. Type of data generated and stored varies with sector. Text/numerals are high in most sectors while video and audio components are in some. For example, in sectors like communication and media, and government, the data has *high* content of video, audio and text/numbers, and *medium* amount of image, whereas in manufacturing, it is *high* for text/numbers, *medium* for video and image, and *low* for audio. As expected, in health care, image and text/ numerals are *high*, and video and audio components are *low*. All these signify the heterogeneity of the data.

Another source of Big Data is social networks and mobiles. While the use of social networking applications through PCs and smart phones is increasing day by day, the growth is more prominent in case of latter. Interestingly, the number of frequent users has been increasing significantly within the set of all users.

Furthermore, the data sets grow in size in part because they are increasingly being gathered by - Ubiquitous information-sensing mobile devices, Aerial sensory technologies (remote sensing), Software logs, Digital cameras, Microphones, RFID (radio-frequency identification) readers, and Wireless sensor networks.

Dealing with Big Data: Challenges and Issues

So, Big Data refers to a collection of datasets that grow so large in volume (scalability), variety and velocity (dynamic) and becomes complex that it is difficult to capture, store, manage, share, analyze and visualize it with the conventional data analysis tools. It requires exceptional technologies to efficiently process within *tolerable elapsed times*. In other words, it needs new forms of processing to enable enhanced decision-making and knowledge discovery, and deliver accurate predictions of various kinds in agile platforms. Here new forms mean new approaches, challenges, techniques, architectures to solve new problems. Accordingly, it demands a revolutionary change both in research methodologies and tools. Existing computational intelligence techniques may need to be completely re-hauled.

The existing data mining and knowledge discovery processes mainly involve issues related to data modalities like ontologies, structured, networks, text, multimedia and signals, and issues related to data operators like collect, prepare, represent, model, reason and visualize. In case of Big Data, the additional issues include usage, quality, context, streaming and scalability. Typical research areas involved are - information retrieval, pattern recognition, data mining, knowledge discovery in data base, machine learning, natural language processing, semantic web etc. And challenges lie with tasks like - Capturing, Pre-processing, Storage, Search, Retrieval, Analysis and Visualization.

Dealing with Big Data: Technologies and Uncertainty Analysis

Suitable technologies that may be used include crowd sourcing, data fusion and integration, machine learning, signal processing, natural language processing, simulation, time series analysis and visualization. The soft computing (SC) paradigm too appears to have a strong promise in developing methodologies for handling Big Data, as it has been successful in several of these tasks for pattern analysis, data mining and knowledge discovery.

What is Soft Computing?

Unlike conventional hard computing, soft computing exploits the tolerance for imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness, low-cost solutions, and close resemblance to human-like decision-making. It is characterized by the use of inexact and intelligent solutions, keeping human mind as the role model, to handle computationally hard tasks such as those associated with the 4Vs involved in Big Data management. At this juncture, the key players of SC are Fuzzy Logic (FL), Artificial Neural Networks (ANN), Evolutionary Algorithms (EA) and Rough Sets (RS), and their different symbiotic integrations. These have proven roles in uncertainty handling, reasoning, learning, optimization and searching, which any design engineer would love to exploit for producing efficient, robust and flexible methodologies.

One may note that managing uncertainty in decision-making is very crucial for mining any kind of data, no matter small or big. While FL is well known for modelling uncertainty arising from vague, ill-defined or overlapping concepts/ regions, RS models uncertainty due to granularity (or limited discernibility) in the domain of discourse. Their effectiveness, both individually and in combination, has been established worldwide for mining audio, video, image and text patterns, as present in the Big Data generated by different sectors. FS and RS can be further coupled, if required, with (probabilistic) uncertainty arising from randomness in occurrence of events in order to result in a much stronger framework for handling real life ambiguous applications. In case of Big-data the problem becomes more acute because of the manifold characteristics of some of the Vs, like high varieties, dynamism, streaming, time varying, variability and incompleteness. This possibly demands the judicious integration of the three aforesaid theories for efficient handling.

Big Data for Social Development: UN Initiative, Scenario and Analytics

The recent waves of global shocks – food, fuel, and financial – have revealed a wide gap between the onset of a global crisis and the availability of actionable information that can help protect the world's most vulnerable populations against further regressions. Actionable information means the ability to quickly and as accurately as possible, profile and respond to crises that have the potential to undo the social development gains. In other words, it demands for more agile systems. It has been realized that the traditional statistics, household surveys and census data may not be effective in generating the kind of real-time picture that decision makers need in order to develop timely responses to ongoing issues.

The United Nations (UN) Global Pulse, since its inception in 2009, has been investigating the viability of using new and alternative data sources, such as Online Content, Data Exhaust, Physical Sensors and Crowd-sourced Reports, to support the development goals. They are in the process of designing various approaches for harnessing Big Data and real-time analytics for monitoring the development progress, emerging vulnerabilities and overall well-being of the populations the UN Serves.

Innovative private companies are finding ways of utilizing real-time analytics to efficiently analyze these new data to better understand the changing needs of their customers and to respond with more agile platforms. Here typical applications include: product recommendation, segmentation of customers, fraud detection or churn prevention where the emphasis is on real-time and highly scalable predictive analytics.

While analytics over Big Data is playing a leading role, there has been a shortage of deep analytical talent globally. It has been observed that the demand in USA is 50 to 60 percent greater than its projected supply by 2018.

Big Data Potentiality in India

Big Data has a great future and promise in India, although the current picture is not as rosy as in many other nations. With the increase in social media usage and adoption of information technology by different sectors, for example, banking, financial services, insurance, retail, and hospitality, Big Data has drawn the attention of Indian enterprises and so has Big Data analytics. Though realization is there, expertise is lacking. Several large enterprises are either in the process of starting or contemplating the use of Big Data analytics, whereas the small and medium businesses are not there yet. Interestingly, as per a report in Analytics India Magazine, the analytics organizations in India (that provide services externally around analytics and related fields) have grown recently both in number and size. Bangalore is the hub of analytics in India, though other cities are coming up.

A significant application of Big Data analytics in India is that it can be leveraged by the Central and State Governments for reform and implementation of the various policies and government schemes from time to time. For example, analysis of the large data periodically collected about delivery, outputs, outcomes and impact of the education initiatives and health care initiatives at primary, secondary and tertiary level can be made useful in formulating the education and health care policies respectively. Similarly in Direct Benefit Transfer scheme, the Government can decide the funding policies and keep a track of improvement and the growth in a particular region, data in Election and Voting systems to help people and growth of the country, and the AADHAAR information for monitoring the citizen-related initiatives. Other prominent application areas in India include Township planning, Tax administration, River network optimization and Unemployment analysis.

It may be mentioned here that Big Data industry is expected to be worth USD 25 billion globally by 2015. NASSCOM predicts that Indian Big-data industry will be worth 1 billion by then (source: DST Report 2014).

DST Initiative

Anticipating the fast growth of Business Analytics, its various applications, and tremendous significance and prospects in India, the Department of Science and Technology, GOI has recently started a Big Data Initiative (BDI) Programme to chalk out a strategic Road Map for promoting the Big Data Science, Technology and Applications in order to derive the benefits towards the overall development of the nation. For fostering research in this high-potential emerging area, a BDI programme support scheme has been launched to provide financial support primarily for R & D projects, establishment of Centers of Excellence, organizing national-level conferences/ workshops, and in-house training programs for faculty and students.

In summary, my dear new graduates, Big Data analysis or analytics research over Big Data has high potential from the R & D perspective in both academia and industry. It is a highly inter-disciplinary research area comprising statistics, mathematics, computer science, information technology and various emerging applications in scientific domain and Industry. The objective is, for example, to develop complex procedures running over large-scale, enormous-sized data repositories for extracting useful knowledge hidden therein, discover new insights from Big Data, and deliver accurate predictions of various kinds as required. Uncertainty analysis plays a crucial role in prediction. In India the scenario is extremely encouraging.

Conclusions

Before I conclude, I wish to draw your attention to a few issues.

As we all know, Jadavpur University was established in 1955 and it is the proud successor of the National Council of Education, Bengal, which was established as a result of the nationalist movement in Higher Education, as a part of the Indian freedom struggle. It is therefore our collective responsibility to ensure that this great university will continue to bear with pride this rich inheritance as it enters its Diamond Jubilee next year. At present, the university is a highly acclaimed and internationally reputed one, with several achievements and laurels. It is one of the few top-ranked Engineering and Technology Universities/ Institutes in India, and a UGC-recognised centre of excellence. Attaining this level of excellence is no doubt a simple task for any institution. However, even more challenging and tough is the task of sustaining this excellence over years, particularly in today's rapidly changing knowledge-based world. It requires a consolidated effort of teachers, students and non-teaching staff, who should work together constantly as a consortium towards this goal.

In this context, one may note that accordingly to international rankings, academic institutions in India are certainly not in any enviable position; ours are even much behind compared to those in China, a neighbouring third-world country with similar population-related issues, opportunities, limitations, knowledge-based society and socio-economic problems. So you all have a big challenge, ahead of you, to upgrade the index of excellence.

Dear new young graduates, this convocation is a red-letter day for you all, and a joyful occasion for your parents, friends, relations and your teachers. Please thank your parents and be grateful to them, by virtue of whose sacrifices, encouragement and nurturing you are here today. You will now be stepping into different professions and new spheres of life, whether you take up jobs or opt for higher study. Wherever you go, please keep in mind, as advice, the following:

- Work hard
- Define your goal
- Set Delhi as your target; Kanpur will lie on the way
- No short-cuts
- Be honest in your profession
- Finish your work ahead of deadline
- Respect your teachers

With this, I congratulate you once again on your worthy performance and wish you all success in life. I pray that you are endowed with the necessary strength, courage and wisdom to bring welfare to yourselves and your fellow beings. I am sure you will contribute to the progress of our nation, and make your Alma Mater and the country proud of your achievements.

I wish you all merry Christmas and happy new year!

*****

Printed by : Sri Sanat Kumar Basu
Jadavpur University Press (Printing), Kolkata - 700032