AI in Healthcare, Animal Husbandry & Plant Genomics

Hello! Good Afternoon everybody. After about 50 years of largely being in technology labs, pilots and science fiction, artificial intelligence (AI) has taken center stage today. Barely a day goes by, without dozens of new magazine and newspaper articles, blog posts, TV stories, LinkedIn columns, and tweets about cognitive technologies. It shouldn't be at all surprising. The impact of AI has been very upfront and highly personal these days. The technology is beginning to reshape the existing ecosystem.

Initially conceived as a technology that could mimic human intelligence, AI has evolved in ways that far exceed its original conception. With incredible advances made in data collection, processing and compute power, systems are now deployed to assist in a variety of tasks, exhibit intelligence and enhance user experience. Over a period of time, AI capabilities have increased manifold leading to its utility in various facets of life.

I will try to bring out, the significance of AI in Healthcare, animal husbandry & plant genomics and try to explain the nuances in the usage of AI in this session. Our vision should be that we in India should not be a passive participant to the technologies developed elsewhere, but positioned to actively influence AI development path.

The global genomics industry, is worth $16.4 Billion USD as of 2018, and is expected to reach $41.2 Billion USD by 2025. The genomics industry, consists of genomic products and services. The genomic products, are expected to dominate the market, due to the rising number of research programs, undertaken by government and private organizations. The genomics services, includes next-generation sequencing, core genomics, biomarker translations, and many others.

After 30 years, of the launch of the Human Genome Project, a major impediment, in drug development remains: a lack of understanding of, disease biology. The

map of the genome, provided the code for producing proteins, leading to medicines targeting genetic mutations. But that was only 2% of the puzzle.

Many diseases, are driven by abnormal expression of genes, but how genes are turned on, off, up, or down remained a mystery. Something goes on in the 98% percent of the genome, once considered “junk”, because it doesn’t code for proteins. Rapidly growing science is now elucidating how that 98% determines the function of every cell, by controlling which genes are expressed at what time and in what amounts. Understanding this “regulatory genome”, will help scientists create a new generation of medicines.

The human genome comprises more than 3 billion base pairs. Recent technological advances, have increased the mechanistic understanding, of the genome biology to an incredible degree. However, the complexity, and sheer amount of information, contained in DNA and chromatin, remain roadblocks to complete understanding, of all functions and interactions of the genome.

Connecting genotype to phenotype, predicting regulatory function, and classifying mutation types, are all areas in which harnessing the vast genomic information from a large number of individuals, can lead to new insights.

However, working in this large data space is challenging, when conventional methods are used. Therefore, new and innovative approaches are needed, in genome science to enrich understanding of basic biology and connection to diseases.

The combination of artificial intelligence technologies and genomics, has the potential to end poverty, end hunger, protect, restore and promote aquatic and terrestrial ecosystems. The transformation afoot is that – as DNA is understood as a numerical sequence, that can be computationally mapped and manipulated – algorithmic tools can be introduced, to speed up further discoveries.

One exciting and promising approach, now being applied in the genomics field, is deep learning, a variation of machine learning that uses neural networks to automatically extract novel features from input data. Deep learning has been successfully implemented in areas, such as image recognition or robotics (as in self-driving cars), and it is most useful when large amount of data is available. In this respect, using deep learning as a tool, in the field of genomics is entirely apt.

Although, it is still in somewhat early stages, deep learning in genomics, has the potential to inform fields such as cancer diagnosis and treatment, clinical genetics, crop improvement, epidemiology and public health, population genetics, evolutionary or phylogenetic analyses, and functional genomics.

Deep learning, a sub-field of artificial intelligence, combined with computer vision techniques can be used to analyze the growing amount of genomics imagery data. In computer vision, deep learning algorithms that include convolutional neural networks and recurrent neural networks, are solving computer vision tasks such as image classification, semantic segmentation, image retrieval and image captioning.

Convolutional Neural Networks, are the basic blocks of computer vision. In genomics, the technique is related to assimilating a screen of genome sequence as an image. The genome sequence, is a fixed length 1D sequence screen, with four channels such as A, C, G, T. The genome, is made up of 20,000 – 25000 genes, in human beings. Sequencing the genome, is a critical first step to understanding it. CNNs can analyse, single sequence through 1D convolutional kernel. Two class image classification, is performed for identifying protein-binding specificity of DNA sequences. The key feature of CNNs, is the adaptive feature extraction, when the training process is performed. CNNs are used to discover, recurring patterns with small variances. These small variances, can be genomic sequence motifs.

Deep learning, combined with natural language processing techniques, can be used to analyse, the expanding amount of genomics-related text, found in publicly-available research papers. Deep neural networks are solving tasks such as named entity recognition, relation extraction, and information retrieval. Deep learning technologies, are ideally suited to deal with, natural language processing tasks since they offer, state-of-the-art performance and overcome challenges with feature engineering.

While much attention has been paid, to the implications for human health, genetic sequencing and analysis, could also be ground-breaking for agriculture and animal husbandry. Farming is becoming a data-centric business powered by artificial intelligence. China's big tech firms are using neural network-backed computer vision, wearable devices, and predictive analytics algorithms, to reimagine pig, chicken, cow, goose, and cockroach farming.

One major challenge of population genetics is the efficient and precise phenotyping of a large population with replicated tests. Visual assessment is incapable, of capturing small yet critical phenotypic variation in plants and animals and is plagued by the lack of repeatability as well as reliability. With the advances in data analytics, machine learning (ML) has been shown to be critical for **image phenotyping** of stress-related traits.

In Japan, Osaka University's intelligent cow breeding system, can detect, contagious viral diseases in livestock, with up to 99% accuracy. The system is being adopted for cowhouses with automatic milking machines and feeding robots and several Japanese dairy farms are using it along with wearable devices to fine-tune milking and feeding and provide real-time updates. At the same time, computer vision and data manipulating software portals are part of the bigger IoT makeover of food production.

When researchers can sequence and analyze DNA, something that artificial intelligence systems make faster, cheaper and more accurate, they gain perspective on the particular genetic blueprint, that orchestrates all activities of that organism. With this insight they can make decisions about care, what an organism might be susceptible to in the future, what mutations might cause different diseases, and how to prepare for the future.

With the development of genetics as a stream of biological sciences, it is evident that many major advances were first delivered in plants. Plants were considerably easier and faster to manipulate genetically than animals, with inherent advantages such as shorter generation times, the potential to produce very large populations, and an ability to manipulate genetic recombination by selfing, outcrossing or both. In modern agriculture plant breeding is the engine, that drives innovation in crops. Improved harvests, disease resistance, plant vigor, drought tolerance – each of these is the product of Mendelian Laws of Probability.

Prior to the introduction of advanced math and data science, breeders had to establish, large scale growing experiments, and observe, how the genetic crosses performed. Many generations, of trial and error were required, to achieve desirable outcomes. This required not only time, but also significant amount of land, water, energy and other precious natural resources.

Just as the sequencing, of the human genome, has improved our understanding of genetic disorders, the sequencing of crop genomes such as for wheat, barley and canola, will allow us to boost agricultural productivity, and improve global food security. New breeding technologies can identify genetic traits, that increase yields and improve climatic resilience to challenges such as droughts.

To make full use, of new advances in genome technology, we need to understand how variation in genomes, corresponds to changes, in the performance of crops. One challenge, in doing this is the immense scale and volume of data, that must

be analysed. Crop genomes are typically large; much larger than the human genome. Interpretation also requires genome data to be integrated with environmental data and crop performance indicators further expanding the scale of the data challenge.

The public are becoming more and more interested, in where their food comes from and there is a growing need for breeders to tap into this public interest. CRISPR gene editing techniques, are only manipulating an existing plant's genome and do not introduce foreign DNA, the technology is therefore, not subject to regulatory regimes, governing the genetically modified organisms (GMOs). The upshot of all this is thought to be a sort of democratization in the "who and where", further advances will come from in plant genomics. Previously, R&D in the space would have been the exclusive domain, of deep-pocketed agribusiness shops like Monsanto and Syngenta – but no more.

Although deep learning holds enormous promise for advancing new discoveries in genomics, it also should be implemented mindfully and with appropriate caution. Deep learning should be applied to biological datasets of sufficient size, usually on the order of thousands of samples. The 'black box' nature of deep learning neural networks is an intrinsic property and does not necessarily lend itself well to complete understanding or transparency.

Subtle variations in the input data, can have outsized effects and must be controlled for as well as possible. Importantly deep learning methods, should be compared with simpler machine learning models with fewer parameters to ensure that the additional model complexity afforded by deep learning has not led to overfitting of the data. Depending on the type and size of the datasets being analyzed and the questions being asked, deep learning can either offer benefits, or introduce more uncertainty.

Another challenge maybe large amount of annotation requirement. An annotation (irrespective of the context), is a note, added by way of explanation or commentary. Genome annotation, remains a major challenge for scientists, investigating the human genome.

DNA annotation or **genome annotation** is the process of identifying the locations of genes, and all of the coding regions in a genome, and determining what those genes do. Once a genome is sequenced, it needs to be annotated to make sense of it. Generating ground-truth labels for genomics datasets, can be expensive as it is a highly domain specific area. Genome annotation consists of three main steps:

1. identifying portions of the genome that do not code for proteins
2. identifying elements on the genome, a process called gene prediction
3. attaching biological information to these elements

Automatic annotation tools attempt to perform these steps via computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally these approaches co-exist, and complement each other in the same annotation pipeline. Genome annotation is an active area of investigation, and involves a number of different organizations in the life science community, which publish the results of their efforts in publicly available biological databases, accessible via the web and other electronic means.

A simple method of gene annotation relies on homology based search tools like BLAST, to search for homologous genes in specific databases, the resulting information is then used to annotate genes and genomes. However as information is added to the annotation platform, manual annotators become capable of deconvoluting, discrepancies between genes that are given the same annotation.

Some databases use genome context information, similarity scores, experimental data, and integrations of other resources, to provide genome annotations, through

their Subsystems approach. Other databases like Ensembl, rely on curated data sources as well as a range of different software tools, in their automated genome annotation pipeline

Another important point is Data Curation. Many important and complex AI projects require the use of disparate data sources (both structured and unstructured) that are time varying, at various levels of quality (wrt. completeness, accuracy, etc.) and of ambiguous origins. Data curation can enable automated data discovery, advanced search and retrieval, improvement in the overall data quality, and increased data reuse.

The process can be described, using what we call, the "Seven C's" of data curation:

1) Collect—U should Interface to the **data sources** and accept the inputs;

2) Characterize—Capture available metadata;

3) Clean—Identify and correct the data quality issues;

4) Contextualize—Provide context and provenance;

5) Categorize—Fit within framework, that defines the problem domain;

6) Correlate— Find relationships among the various data; and,

7) Catalog—Store and make data, and metadata accessible, with application program interfaces, (APIs) for search and analysis.

The benefits of the data curation process are a reduction in problem-solving time, improved data quality, increased confidence in solutions, reduced time and manual effort to perform the curation itself, and the ability to solve problems that were previously too complex or time-consuming to solve because of data problems.

Computer databases are an increasingly necessary tool, for organizing the vast amounts of biological data currently available and for making it easier for

researchers, to locate relevant information. In 1979, the Los Alamos Sequence Database was established as a repository for biological sequences. In 1982, this database was renamed GenBank and, later the same year, moved to the newly instituted National Center for Biotechnology Information (NCBI), where it lives today. By the end of 1983, more than 2,000 sequences were stored in GenBank, with a total of just under 1 million base pairs. The phenomenal growth of sequence data in GenBank is challenging to manage, and continues unabated. At about the same time, a joint effort between NCBI, the European Molecular Biology Laboratory (EMBL), and the DNA Databank of Japan (DDBJ) created the International Nucleotide Sequence Database Collaboration (INSDC) to collect and disseminate, the burgeoning amount of nucleotide and amino acid sequence data that was becoming available. Since then, the INSDC databases have grown to contain over 95 billion base pairs, reflecting an exponential growth rate in which the amount of stored data has doubled every 18 months. The sheer volume of the raw sequence data in these repositories, has led to attempts to reorganize this information into various kinds of smaller, specialized databases. Such databases include various genome browsers, model organism databases, molecule- or process-specific databases, and others.

As previously mentioned, the INSDC is a collaboration of NCBI's GenBank in the U.S., EMBL in Europe, and the DDBJ in Japan. Each of these databases accepts direct submissions of biological sequences from individual researchers, from sequencing projects and from patent applications from around the world. Sequences are entered into the database, and given a unique identification or accession number.

These submitted entries, are stored in a "library" of records, and each entry is "owned" by—and can only be updated by—its submitter. The data integrated in these entries, include the submitter's name, the originating organism, the

definition, the actual sequence, related references, and more. (Examples of IDs and entries are accessible through these links.)

The submitted entries, are then shared across the three repositories on a daily basis, and releases of the data are made regularly. This has been a boon to the research community facilitating the sharing of sequence data and allowing the advancement of research. These sequence repositories have become the universal, comprehensive, and authoritative resources for the exponentially growing amount of sequence data currently available to researchers.

However those data, that are not accepted by INSDC, is also humongous and if databank of curated data exist for these, then they may contribute as negative example for training AI Models. In AI to remove biases from the models, equal amount of positive and negative data samples are required at the training phase, for the models to perform optimally. You learn from both positive experiences in life as well as negative, same for deep learning neural networks and there is an immediate need to setup such a databank in India and work on curation of these data as well.

As more data become available, better models will be able to be trained, thus resulting in even more precise and accurate predictions of genomic features and functions. These are exciting times, for scientists applying their knowledge and skills to the betterment of everyone in agriculture, food security and development goals .

At a time when **India is striving to rekindle productivity and growth**, AI promises to fill the gap. A **full and responsible implementation of AI** will open new economic opportunities that would not otherwise exist.

The guiding principle should be to create **"people first**" policies, and strategies, centred on using AI to augment, and extend people's capabilities for the benefit of humankind.

With the Collaboration of stakeholders, in a coordinated fashion with common end goal, leveraging the strength of AI developed nationally or internationally, we can truly build AI systems that are not just good for India but can also be replicated by other nations sharing a common vision as that of India.

2nd National Symposium on Database Development and Biocuration (NSDDB-2019)

Department of Plant Molecular Biology, University of Delhi South Campus

December 17 and 18, 2019

Venue: Auditorium, 3rd floor Biotech Center, UDSC

Programme Schedule

Day 1, Dec. 17, 2019

09.30-10.00 am	Registration		
10.00-10.30 am	Opening Ceremony	Lighting of the Lamp	
		Introduction to the Symposium	
Technical Lectures			
Time	**Name**	**Affiliation**	**Lecture Title**
10.30-11.00 am	Dr. Ramesh V Sonti	NIPGR, New Delhi	Journey from 'Big Data' to Knowledge: Role of Biocuration
11.00 -11.30 am	Dr. Yogesh Shouche	NCCS, Pune	Human Microbiome: A Global Snapshot
11.30-11.50 pm	**Tea Break**		
Flash Talks 11.50-12.30	Dr. Divya Tej Sowpati	CCMB, Hyderabad	MSDB: a comprehensive, annotated database of microsatellites
	Ms. Pratibha Gour	UDSC, New Delhi	Impact of digitization of experimental data: A case study in rice
	Dr. Tina Begum	SIB, Switzerland	Bgee database: creating knowledge from gene expression in any animal species
12.30-01.00 pm	Dr. Bratati Kahali	IISc, Banglore	Whole Genome Sequencing (WGS) for the Indian Population: an analytical perspective
01.00-02.00 pm	Lunch		
02.00-02.30 pm	Prof. Maharaj K Pandit	DU, New Delhi	Build-up of India's Plant Biodiversity Through Ages: Unravelling its Evolutionary Uniqueness and Imperatives of Conservation
Flash Talks 02.30-03.00 pm	Dr. Shikha Rani	NIPGR, New Delhi	Systematic identification and curation of Long non-coding RNAs from Solanum lycopersicum leaf during heat stress
	Ms. Deeksha Pandey	DU, New Delhi	In-silico prediction tools for investigation of antimicrobial resistance genes in various –omics dataset
	Dr. Suresh Kumar Rana	G.B. Pant NIHESD	Biodiversity research over 200 years in the Himalaya: Trends, gaps and policy implications
03.00-03.30 pm	Tea Break		
03.30 - 4.00 pm	Ms. Sharmistha Dasgupta	NIC, New Delhi	AI in Healthcare, Animal Husbandry & Plant Genomics
4.00 - 4.30 pm	Dr. Manish Kumar	UDSC, New Delhi	Artificial Intelligence and Machine Learning: Buzzwords that Transformed the Biological Sciences
4.30 - 5.00 pm	Dr. Manoj Kumar	IMTECH, Chandigarh	Application of artificial intelligence and machine learning in biological data analytics
5.00 - 5.30 pm	Mr. Sandesh Chopade	University College London	From genes to data and everything in between

Day 2, Dec. 18, 2019

Technical Lectures

Time	Name	Affiliation	Lecture Title
10.00 - 10.30 am	Prof. Andrew Lynn	JNU, New Delhi	
10.30 -11.00 am	Dr. Pankaj Khurana	DIPAS, New Delhi	Digital Information-Systems for Research in hypoxic-systems
11.00 - 11.20 am	Tea Break		
Flash Talks 11.20 - 11.40 pm	Dr. Dinesh Kumar Jaiswal	GGSIPU, New Delhi	Transcriptomic and protein network analyses of heterotrimeric Gα subunit (RGA1) mutant expands its functional roles in rice
	Dr. Kalpana Singh	CCS Univ. Merrut	WheatQTLdb: A QTL database for wheat
11.40 - 12.10 pm	Dr. Dinesh Gupta	ICGEB, New Delhi	Applications of Artificial Intelligence in biological data mining
12.10 - 12.40 pm	Dr. Deepika Kalo	Redcliffe Life Sciences Pvt. Ltd	Curation and interpretation of genomic variants
12.40 - 01.20 pm	Panel Discussion		
	Topic: The biocuration community in India: Who all can join? **Panelists: Prof. M. K. Pandit** (DU), **Dr. Amitabh Mohanty** (Prog. Director, NGGF), **Mr. Praveen Gupta** (Premas Life sciences Pvt Ltd.), **Dr. Dinesh Gupta** (ICGEB), **Prof. Andrew Lynn** (JNU), **Dr. Manoj Kumar** (IMTECH), **Dr. Saurabh Raghuvanshi** (UDSC)		
01.20 - 01.30 pm	Closing Remarks		


NSDDB 2019


NSDDB


NSDDB
December 17-18, 2019

****