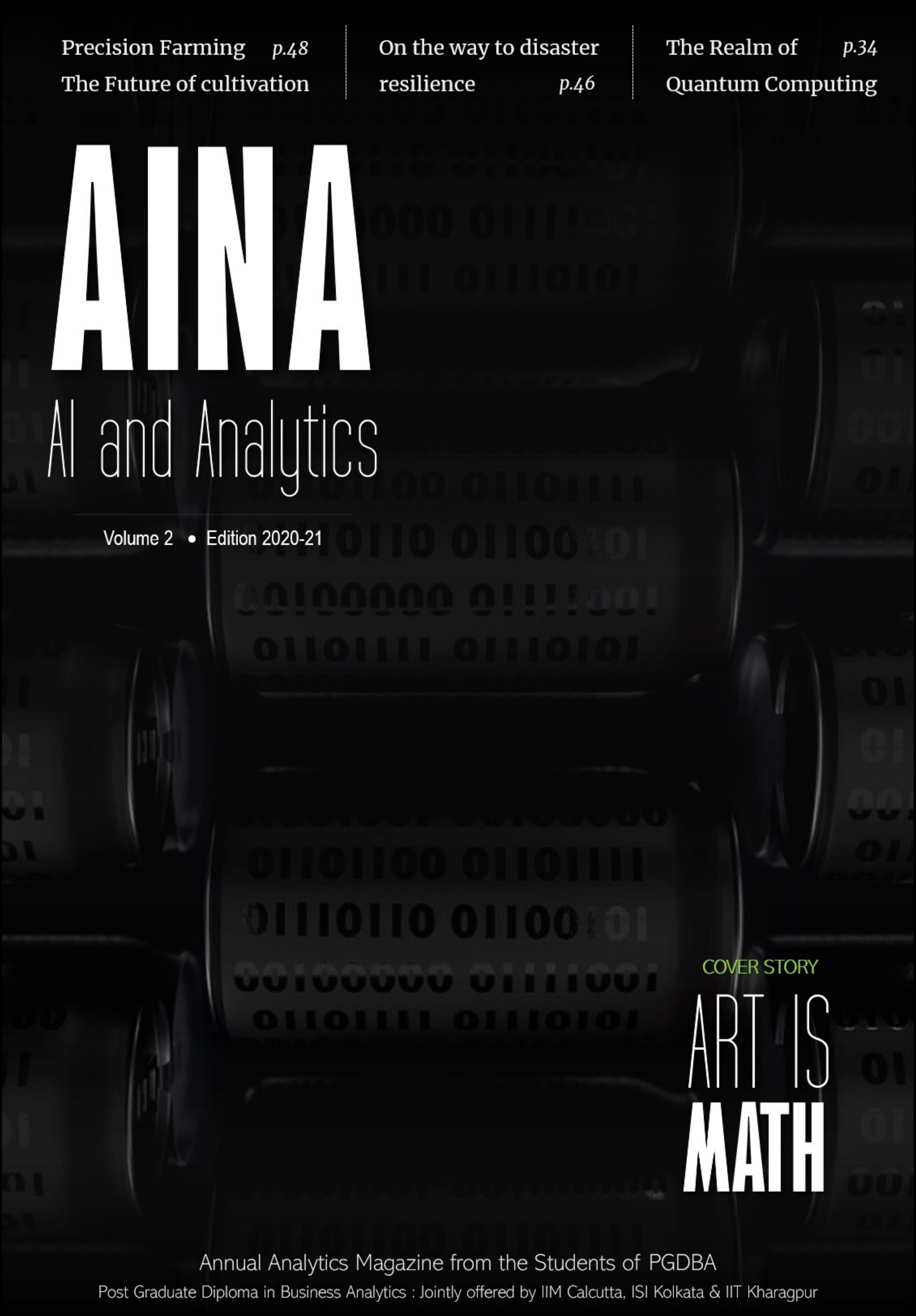
Precision Farming *p.48*
The Future of cultivation

On the way to disaster resilience *p.46*

The Realm of *p.34*
Quantum Computing



# AINA

## AI and Analytics

Volume 2 • Edition 2020-21

COVER STORY

## ART IS MATH

Annual Analytics Magazine from the Students of PGDBA
Post Graduate Diploma in Business Analytics : Jointly offered by IIM Calcutta, ISI Kolkata & IIT Kharagpur

# From the Team

Dingari Sreeram

Editor-in-Chief

Utkarsh Yadav

Content Strategist

Raj Chauhan

Art Director
& Lead Designer

Venkata Ravi Teja

Features Editor

Prasun Kumar

Managing Editor

Thomas Preetham

Creative Director
& Sub Editor

The pandemic has disrupted both the lives and livelihoods across the globe. Though vaccination drives have been a welcome relief, we are still reeling from the effects of this pandemic, which made collaborating, learning and working from home the 'new normal'. Digital readiness has been proven to be a lifesaver in these testing times. Industries and countries that could adjust to and afford this transformation handled this chaos better than their counterparts. The use of artificial intelligence and analytics has produced many new insights in diverse fields. As a result, the need for individuals with these skillsets has seen a meteoric rise in recent years.

Industries are scaling their efforts to leverage the latest technologies to stay a step ahead of the competition. Insights from academia and industry shall always be incredibly beneficial in staying relevant and making intelligent business decisions in this rapidly evolving landscape. In line with this thought, last year, India's first student-driven analytics magazine, 'AINA – Artificial Intelligence and Analytics', was published, with an objective to bring to its readers the latest developments in AI.

We are delighted by your response to the previous edition and would like to present to you the 2021 edition of the AINA magazine. In this edition, as part of 'Interviewing the Masters,' we bring to you intriguing conversations with a Kaggle Grandmaster, an NLP researcher scholar, and a Senior Professor of Computer Science, each offering their expert perspectives. The articles within this edition explore a broad spectrum of themes like Environment, Health, Finance and Art. We are positive that many readers would find the sections on infographics and research publications quite intriguing.

We express our deepest gratitude to the chairperson, the directors, the deans, and the faculty of ISI Kolkata, IIT Kharagpur, and IIM Calcutta for their continued support. We are highly obliged to Prof. Pabitra Mitra, Sudalai Raj Kumar, and Bodhisattwa Majumder for sharing their valuable insights. We take this opportunity to thank Aditya Gadepalli, Kolli Parasuram, Anudeep Immidisetty, Chandu V Grandhi, Srijan Gupta, and Harish Paturu for their constant guidance and encouragement. A special thanks to Jaihind Sawant, Hitesh Kashyap and Abhinav Ranjan and our Alumni for their valuable contributions. Finally, we remain indebted to our family, friends, mentors for being our strength and motivation throughout this journey.

# Contents

## Features

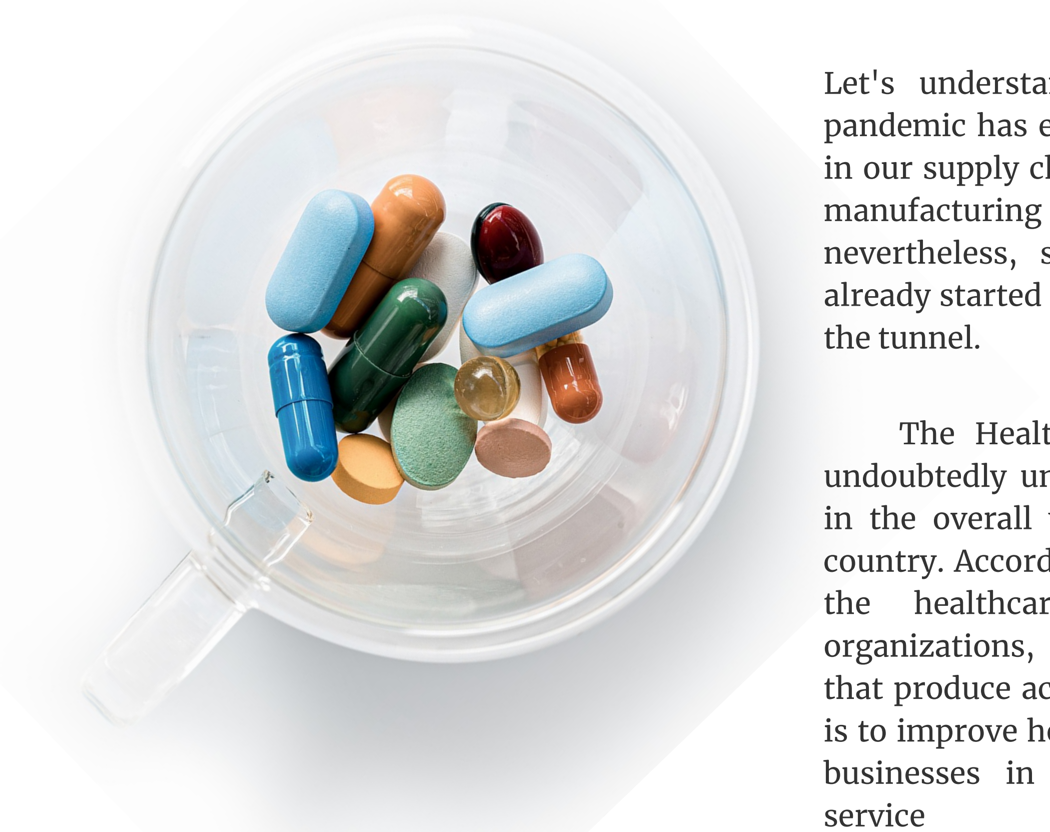
## Expertise

## Quick Reads

## Innovations

Note: An effort has been made to ensure that the images used alongside the text are copyright free with few exceptions of vector images from unsplash.com and freepik.com. We request the concerned people to contact us incase of any objection to their usage @ magazineteampgdba@gmail.com

The content presented in this magazine represents the views of the respective authors developed in due course of going through several online articles, reports, research papers, blog posts, etc.



# A New Hope

## The emerging world of digital healthcare and analytics

Aditya Gadepalli
Anudeep Immidisetty

The next big thing in AI has always been a matter of discussion in several corporate and research communities. With AI spreading its wings into nearly every domain that we come across, this question has gained much more prominence than ever. While Artificial General Intelligence (AGI) might seem too far, recent happenings have shifted the focus towards something more indispensable. The COVID-19 pandemic might be here to stay, yet the disruption it has caused has been bringing several developments of the past decade into the spotlight.

Let's understand the scenario first. This pandemic has exposed serious shortcomings in our supply chains, consumer markets, the manufacturing sector among many, nevertheless, sectors like healthcare have already started showing us light at the end of the tunnel.

The Healthcare system, as it is now undoubtedly understood, plays a major role in the overall well-being of a society or a country. According to the definition of WHO, the healthcare system comprises all organizations, institutions, and resources that produce actions whose primary purpose is to improve health. The major categories of businesses in healthcare include medical service providers (hospitals), pharmaceuticals, insurance providers, and medical equipment manufacturers. This ecosystem comprises one of the largest markets in the world, with the global healthcare market valued at nearly $8.5 trillion in 2018 and is expected to grow much faster at a CAGR of 8.9% to nearly $11.9 trillion by 2022.

In the Indian context alone, this market was valued at $140 billion in 2017 and is projected to reach $372 billion by 2022. Currently, this market is largely skewed towards the western hemisphere with North America alone accounting for 41.9% of the global market in 2018. However, going forward, the fastest growing regions will be the Asia Pacific and Africa, where growth will be at CAGRs of 13.4% and 13.1% respectively.

Large and encompassing industries like these are often burdened with challenges. For example, aging populations, emerging diseases and more recently the COVID-19 pandemic have given rise to advanced focus areas of research and development. These efforts would need additional Infrastructure and R&D spend, often with little to no guarantee of profitable outcomes.

Let's consider CureVac's mRNA research for developing their vaccine shot. A simple preliminary analysis of their under-development mRNA-shot discovered its inefficacy in comparison to its parallels, leading to a whopping 52% plunge in share prices, wiping around $9.6 billion in market value. Such high levels of risk and uncertainty are discouraging for the other players in the ecosystem. We saw how until the peak of the first wave of COVID-19, neither logistics partners nor governments have actively invested in building cold-storage supply chain networks. Was it a failure to forecast medical demand patterns or was it entirely the failure of pharmaceutical firms to prove the profitability of such ventures to these players remains a mystery.

There is a definite need to optimize costs and provide confidence to businesses to garner investment and growth of this much needed sector and AI has been constantly proving its worth in this regard. Be it supporting clinical research, capacity planning of supply chains, demand forecasting for pharmaceuticals or claims validation for healthcare insurance providers, AI has been strengthening the ecosystem from multiple fronts. The most popular upgrades can be witnessed in the areas of automated diagnosis. Google's diabetic retinopathy analysis, MIT's research on melanoma identification, automated MRI/CT scan analysis, and also the Kaggle community's contribution towards these developments prove how healthcare has caught the attention of the AI fraternity. Drug discovery, imaging and diagnostics, genomics, remote monitoring, mental health research and even fitness studies are being assisted or even disrupted in a big way. The advent of wearable tech like the Apple Watch reinforces why AI can be lifesaving. All these forces combined are currently driving progress and demand simultaneously across the sector.

The entire digital revolution is also changing the industry dynamics. Numerous startups are venturing into the digitization of health records, offering tele-consultation and extending e-commerce services. There are entire firms catering to healthcare analytics and related products. Developers have also come forward with apps to leverage wearable tech and actively monitor health. These simple acts of digitization have facilitated the creation of ginormous data lakes, which can be leveraged on the rails of AI to generate insights for every player in the ecosystem. Hospitals can improve their decision support systems, pharmaceuticals can understand disease and immunity trends, supply chains and equipment manufacturers can forecast demand and insurance firms can verify claims better.

The progress so far looks extremely promising until we start analyzing the caveats surrounding these developments. The growing concern around data privacy is a matter of concern, especially when it comes to something as personal as data in patient profiles. Also, the healthcare industry alone contributes at least 4.4% of the global emissions and therefore adoption of AI must assist in reducing this and not make matters worse.

Governments globally are yet to establish policy making centered around technology and especially AI in crucial sectors like healthcare. These would not only reinforce development but also enhance economic performance by minimizing vulnerabilities and creating new-age employment opportunities.

Amidst the entire ocean of advancements, it is important to acknowledge the primal purpose of survival and longevity. In the emerging world with health adversities, people require assurance, however small, that there is hope, and AI might be the exact hope we are looking for.

# AI
# & Mental Health

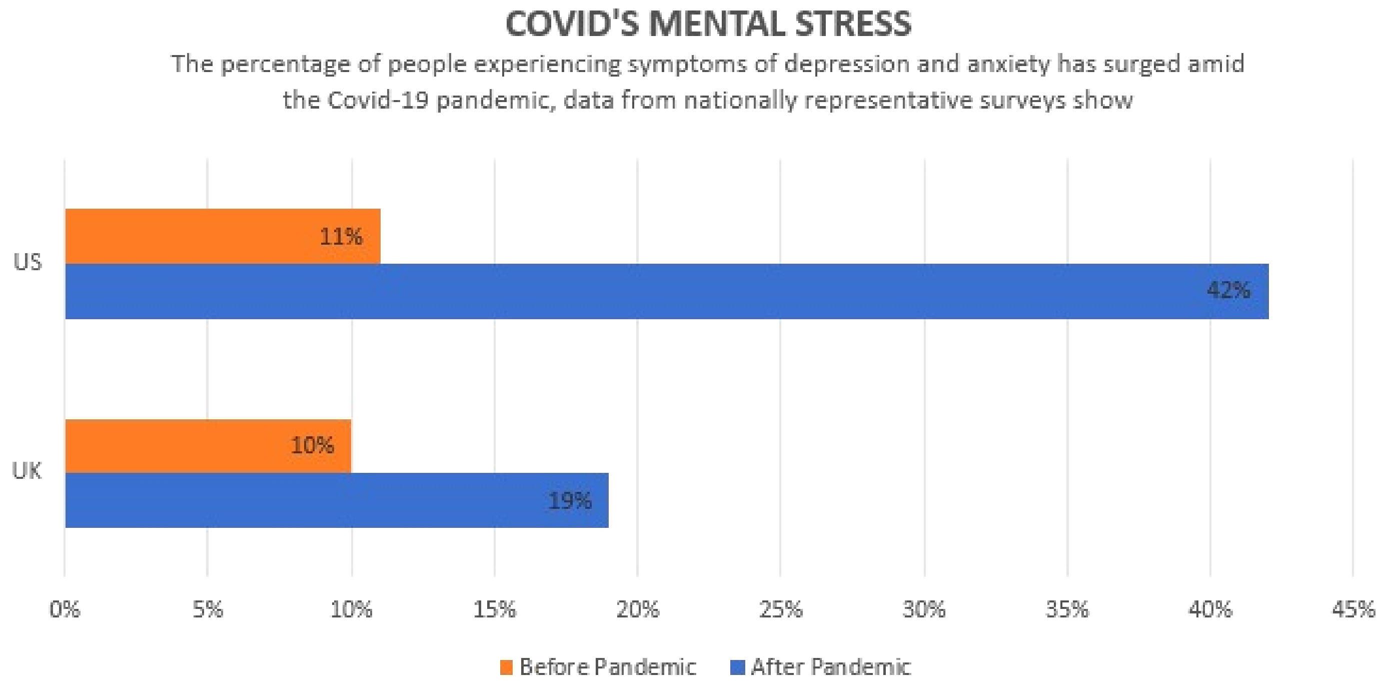
Prasun Kumar

According to a WHO report, 7.5% of Indians, or 97 million people out of 1.3 Billion, have some mental disorder. It also states that there is a massive scarcity of medical care for such patients in India. For the 100k population, there are only 0.3 psychiatrists, 0.12 nurses, and 0.07 psychologists, which is far from the recommended 3 number of psychiatrists and psychologists. Such a crippling healthcare system, and the growing number of people struggling with mental health problems, suggest that a crisis is looming over India. According to an estimate by WHO, between 2012-2030, India will incur an economic loss of USD 1.03 Trillion due to the mental health issues with its young population. Disability-adjusted life year (DALY) is expressed as the number of years lost due to disability, ill health, or early death. The mental health burden for India is 2443 disability-adjusted life years per 10000 population and 21.1 age-adjusted suicide rate per 100,000 population.

The COVID-19 pandemic has worsened the situation further. A survey conducted by US Census Bureau revealed that 42% of people reported symptoms of anxiety and depression in December 2020, while it was 11% the previous year (see Figure-1). Similar observations were made from other surveys worldwide.

This sudden influx of patients suffering from mental health issues can be attributed to the limited social interactions, fear of illness, financial distresses due to job losses, and others. Amidst all these, one major problem has emerged. Many psychiatrists are reporting burnout due to the increased number of patients and the emotional nature of their work.

In such a grim situation where the mental health workforce is in short supply, the advent of AI and Machine Learning has brought extreme hope to solve this task, which has been considered very difficult to tackle. Healthcare is one of the most challenging paradigms for machine learning methods because of the risk associated with the wrong prediction and the lack of suitable data. The human brain has always been a hard nut to crack for scientists. But recently, researchers have made significant progress in applying machine learning techniques to help patients struggling with mental health problems.



COVID'S MENTAL STRESS

The percentage of people experiencing symptoms of depression and anxiety has surged amid the Covid-19 pandemic, data from nationally representative surveys show

Source: Office for National Statistics (UK data); Centers for Disease Control and Prevention (US data)

## Why is a new method needed?

The traditional methods of diagnosing mental illness include physical examination, lab tests, and psychological evaluations. They have not proved very helpful as far as the treatment of mental diseases is concerned. Physical examination includes observing facial expressions, eyebrow shapes, sudden changes in pupil dilation, and different blinking rates. The person who suffers from such problems might show some peculiar physical characteristics. Psychological evaluations include Interviewing the individual and asking questions which gives us a deeper view inside the person's brain. For preliminary detection, these are great tools, but it becomes difficult to find the right path of treatment due to inherent complexity.

What makes mental illness different from other health conditions is the stigma associated with it. These stigmas affect patients at their workplace, home as well as educational institutions. People who had dealt with mental diseases said that the stigmas were more disabling than the disease itself. One of the worst consequences of these stigmas is that the patients are reluctant to seek medical treatment, and thus lead to worse situations day by day. As patients are hesitant to seek medical advice, it exacerbates their problem with time.

Here comes the power of anonymity of AI to our rescue. A conversational chatbot designed to talk to a person suffering from mental issues can help them express themselves freely. They can live without the fear of being judged by the people. The ease of access to these bots and its lower cost compared to psychiatrists make it a lucrative choice in the initial days of the problem.. The importance of psychiatrists and psychologists cannot be replaced; instead, the new AI systems add to what psychiatrists can already do. These advantages help us find the undiagnosed person, speed up the recovery process, and increase the probability of being cured at the right time. Also, the patient needs to be constantly monitored when they are struggling with mental diseases as the panic episodes can come at any time. Technological progress has made it possible by constantly monitoring the patients by sensors and devices that can continuously read the brain wave signals and alert concerned people by sending an alert message if there are significant changes in the brain wave signals.

## Data Collection

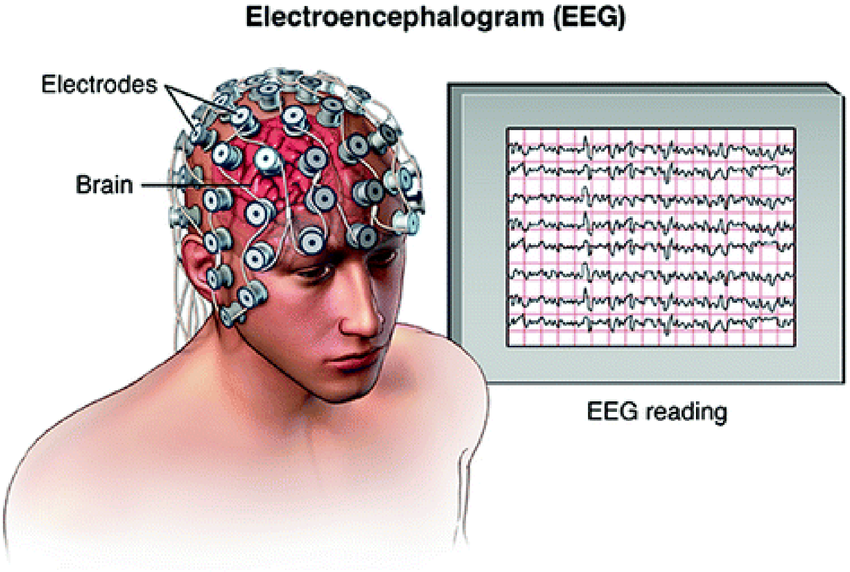
The backbone of any machine learning model is suitable and relevant data. Let us look at some of the ways in which we can collect and use data for treating patients.

## Electronic Health Record (EHR)

In recent times, the promise of machine learning methods in helping doctors to predict, diagnose and treat mental health issues have attracted many researchers. Thanks to the structured healthcare system in the US, researchers can easily access vast amounts of data sets in the form of Electronic Health Record (EHR). EHR data are collected routinely when an individual takes medical consultation. It comprises two types of data: structured data such as medical reports, diagnostic tests, and unstructured data such as prescriptions, clinical notes, and other text -based data.

According to the article "the future of electronic health record" published in the journal Nature, the USA started a significant push to digitize its health record in 2009. In 2008, only 9% of hospitals in the US had electronic health records, but in 2017, it had reached 96%.

The big challenge in using EHR data is the underrepresentation of healthy people in the dataset. Electronic Health Record exists mainly for people suffering from health issues, but the data does not accurately represent healthy people's features. This problem is tackled by using the standardized values of health measures already available with the clinicians. The man-made errors in data entry and poor labeling are among other issues with EHR data.



Source: Siuly S., Li Y., Zhang Y. (2016) Electroencephalogram (EEG) and Its Background.

In: EEG Signal Analysis and Classification. Health Information Science. Springer, Cham

## Smartphone and Sensor Data

Some of the most crucial pieces of information which is helpful to analyze the mental health of an individual are obtained using smartphone sensors, which are difficult to obtain otherwise. Sleep quality, location, exercise, heart rate, communication pattern, and language use are generally extracted to analyze an individual's mental health. Among sensor data collection, the contribution of smartphones is immense because most people own a smartphone these days, and they keep it with them most of the time. Smartphones also have various measurement sensors already in place. Mobile phone sensors can also allow researchers to collect data at regular intervals for monitoring purposes, which can prove to be very useful in the case of mental illness. Smartphone also seems to be very promising in terms of collecting cognitive data at large scale.

## Social Media Data

Social media data can mostly provide the language and activity-related information that can prove to be very useful in assessing the state of mind of someone struggling with mental health issues. But we don't have many instances where this was used by researchers. The primary issue with social media data in this context is the presence of a lot of noise and irrelevant information, which are not very useful from the mental health research point of view. The noise with such data can hinder the ability of the model to predict well and thus can give erroneous results. Researchers have tackled this issue, in some cases, by taking repeated samples for testing the hypothesis since sample size can compensate for the noise in the data.

## Genetic Data

Our genes contain vital information about our characteristics. The genetic data can be very helpful in predicting the outcome of a particular treatment given to a patient. Pharmacogenomics is the field in which scientists try to find the relation between medical treatment and genetic background. A study carried out by Tansey et al. over ~3000 patients suffering from depression found that variation in genetic patterns can explain up to 42% of the differences observed in the antidepressant treatment response of individuals. This further suggests that genetic data can be useful in predicting the treatment response. However, researchers still need to find some substantial success in using genetic data for treating mental issues.

## Brain Measurement Data

Even after multiple attempts, researchers faced a lot of difficulties in working with genetic data. Thus, the brain measurement came into the picture, which has shown promising results. There are various ways in which brain activities can be measured.

## Electroencephalogram (EEG)

EEG is a technique to detect the electrical activities of our brain using metal electrodes attached to our brain. Our brain cells are active all the time and communicate with each other using electrical signals. These electrical signals come as wavy lines on EEG, which contain information about brain activities.

## Magnetoencephalography (MEG)

At cellular levels, each neuron has electrochemical properties, which causes charged ions to flow through the cell. This current creates quite a weak electromagnetic field. The strength of the electromagnetic field becomes measurable only when around 50k neurons are excited together in a specific area. Such a weak electromagnetic field is detected using Superconducting Quantum Interference Device (SQUID).

## Functional Magnetic Resonance Imaging (fMRI)

fMRI measures the variation of blood flow in the parts of the brain as different parts get activated due to different stimuli. It is used to determine which parts of the brain handle critical functions and the effects of diseases or stroke on our brain.

## Application of Machine Learning in Mental Health

For a novel method like Machine Learning to be applicable in the mental health domain, it is important for us to know the applicability and feasibility as well as actionable outcome of these methods. Researchers have already tried applying these techniques and have obtained promising results. Recently, researchers have started applying deep learning models for predicting depression using fMRI, EEG, MEG, voice and visual data, webcam video data, audio during the psychiatric interview, and Reddit self-reported depression diagnostic data.

At present, it is very difficult to cure mental illness. Most of the traditional methods are symptomatic treatments, where doctors try to treat the visible symptoms such as restlessness, anxiousness, irritability, irrational thoughts by giving medicines that subdue the adverse effects of these symptoms. Understanding the root cause behind these problems is the inherent difficulty associated with finding the right treatment. Doctors use a trial-and-error approach while deciding the dose of medicines. In some cases, the trial and error approach works well, but since every person's response to these treatments is different, there are many patients who get incorrect treatment.

One interesting and widely used application of Machine Learning is to predict who will drop out after initiating an antidepressant or any kind of medication. In these applications, EHR data have shown promising results, modest but not clinically actionable. Researchers like Hayes, Pradier, and Hughes tried to find out the patients who are likely to stop taking medicines, the patients whose condition will improve upon treatment, and the patients who will transition to a bipolar diagnosis.

When Pradier, in 2018, tried to predict the patients who will drop out of the treatment process only using the demographic data, he did not obtain good results, but on incorporating structured EHR data along with demographic data, the model performance increased by 13%, which proves the importance of EHR data.

The genetic background gives us very useful information about the way a person will respond to the medication. A polygenic risk score, which is a single value estimate of an individual's genetic liability to a trait or disease, is calculated using the genetic data. In 2017, Garcia et al. found that polygenic score for major depression was not a very suitable predictor for finding antidepressant efficacies. There have been various efforts to use genetic data to predict the effectiveness of the medicine for mental health issues, but we are yet to get any promising results. Although still in its infancy state, personal sensing holds great promise for its application in monitoring at-risk populations.

## Challenges

All types of issues that exist with a machine learning approach translate directly to the above application. Data bias, model interpretability, poor validation schemes are among some of the issues. In recent times, we increasingly see the use of more and more deep learning methods being applied to solve the problems related to mental health. With the advent of such methods, we face the problem of model interpretability. The deep learning models work like a "black box," which takes data and gives us predictions without providing us information about the reason behind the prediction. In other domains, such as object detection, this can be accepted without any issues, but in the case of clinical applications, doctors and medical researchers need to understand the reason behind the predictions so that they can take action based on the recommendations. These problems occur when the domain knowledge is lacking during a study. The current biomedical knowledge bases, clinical research paper databases in addition to medical experts can be incorporated to address the limitation posed by data quality, model generalizability, and interpretability.

Sometimes, researchers do not create their cross-validation strategy properly and report overly enthusiastic results, which is a big challenge for machine learning practitioners. These models, although performing well on the dataset used, fails to generalize well and can give erroneous results when applied to new patients.

Patients concerned about privacy issues are also a big challenge. Collecting mental health data is very sensitive and requires some personal information to be extracted and analyzed, which makes certain people reluctant for opting such treatments.

## Conclusion

Machine Learning has come a long way in the last decade, but despite its progress, its potential in psychiatry has just started to be explored. Advanced techniques like autoencoders and deep belief networks have shown promising results and are widely used in this domain. The application of ML in mental health has demonstrated exciting results and has also signified the immense potential that it carries. However, the current status of the work is limited and there is a lot of scope for researchers to identify the potential benefits and application of machine learning to help patients struggling with mental issues. Given the kind of research and innovations being done in machine learning daily, the field will keep on growing, and novel approaches will solve the current problem one day.

# Covid Vaccine Trials

## How comparing Apples with Apples has gone wrong

Venkata Ravi Teja

# Covid 19

with its high infection rate, choked the health care systems around the world. It severely impacted the economies of almost all the countries. Many developed and developing countries had actively partaken in the vaccine race. Medical science teams worldwide had successfully come up with vaccines within a short span, which is an impossible feat a few decades back. The vaccine's effectiveness is typically tested in multiple phases, and it enters its first phase only after it is proven harmless on animals.

The initial phase involves a small group of participants, typically less than 100, administered with the vaccine in small quantities. They will be monitored to determine whether the vaccine is safe, tolerable, and causes any side effects. In Phase-2, the number of volunteers would be in hundreds. Medical teams would determine the dosage of vaccine required to trigger the immune response during this phase.

Phase-3 trial is the macro version of the first two trials, the number of volunteers would be in thousands, and the volunteers are from different geographies to capture the demographic variability. The experiment should be randomized controlled where the volunteers on whom the experiment is conducted are split into two groups, vaccine and placebo. The assignment into these groups should be completely random, and the entire experiment should be conducted under similar conditions. The vaccine group receives the vaccine, and Placebo groups won't receive the vaccine.

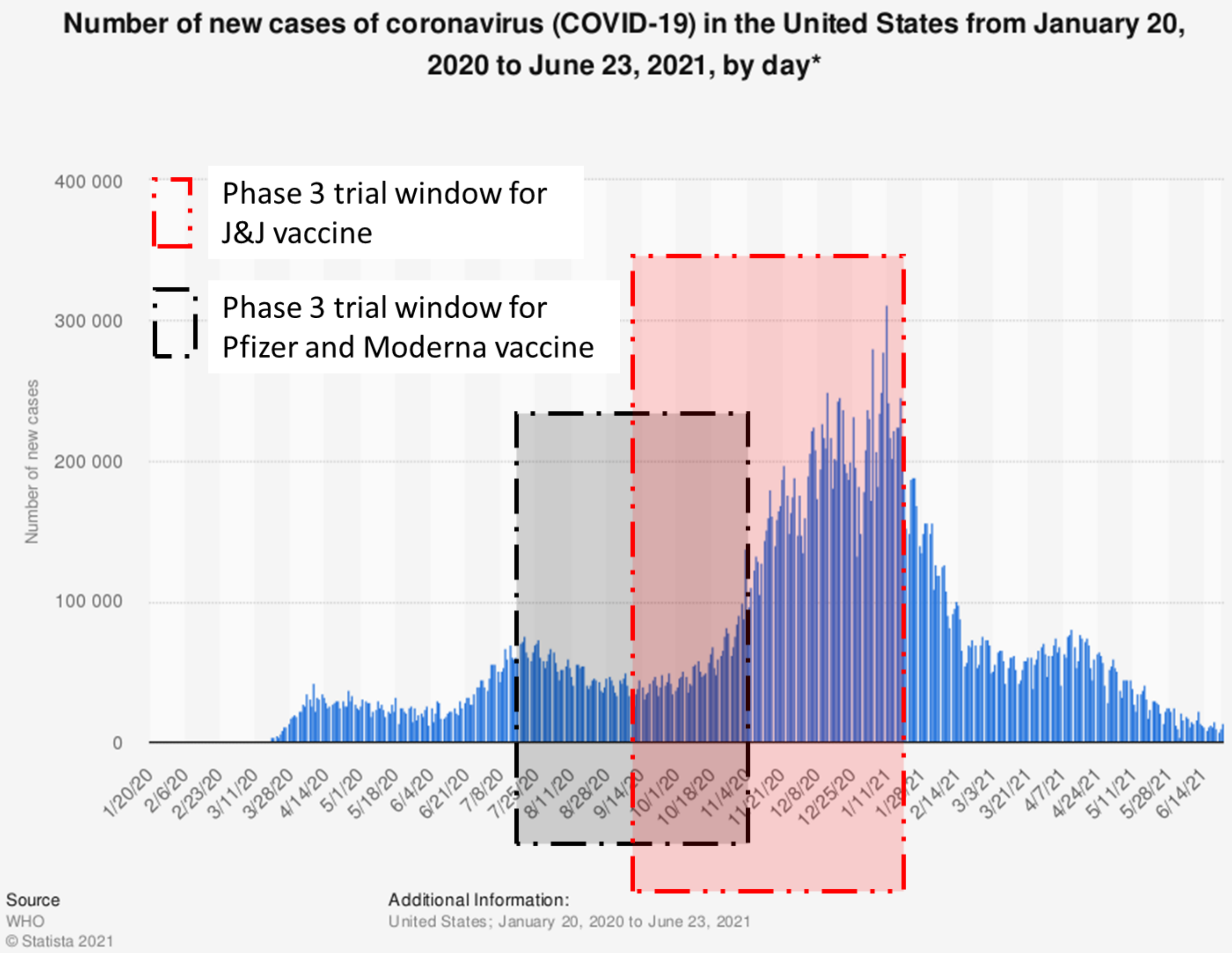
The experiment should be double-blinded, i.e., both volunteers themselves and doctors treating them don't know whether they belong to the vaccine group or placebo group. A Double-blinded experiment ensures that there is no bias from the doctor and puts the placebo groups under the impression that they are receiving the vaccine. By comparing these two groups, we can get to know whether the vaccine has any impact on curing a particular disease.

Food and Drug Administration (FDA) agency of USA has initially approved three vaccines for Covid treatment: Pfizer, Moderna, and Johnson and Johnson(J&J). Pfizer is the first-ever covid vaccine to receive FDA approval on December 11th, 2020, followed by Moderna after a week. Both vaccines are mRNA-based and administered in two doses. The former claimed efficacy of 95%, and the latter claimed 94.1%. Johnson and Johnson (J&J vaccine) is the third vaccine that got its approval from FDA on February 27th, 2021. Unlike the previous two vaccines, it is a single shot carrier vaccine that uses a different approach than the mRNA vaccine to provide immunity. This vaccine claimed an overall efficacy of 72% in the USA and 66% worldwide. The general perception towards these vaccines going purely by their efficacies is that the J&J vaccine is comparatively less effective than Pfizer and Moderna. There are lot more things that have to be deliberated before comparing these vaccines.

The general public has a misapprehension about the efficacy of the vaccine claimed over trials in translating it into the actual scenario. The vaccine efficacy can be calculated as:

$$Vaccine\ Efficacy = 1 - \frac{Attack\ rate\ of\ vaccinated\ group}{Attack\ rate\ of\ placebo\ group}$$

The attack rate of a group is the ratio of the number of people infected with the disease to the total number of people in that group. The ratio of attack rates i.e., the ratio of probabilities of getting infected in the vaccinated group to the placebo group is called Relative risk.



Number of new cases of coronavirus (COVID-19) in the United States from January 20, 2020 to June 23, 2021, by day*

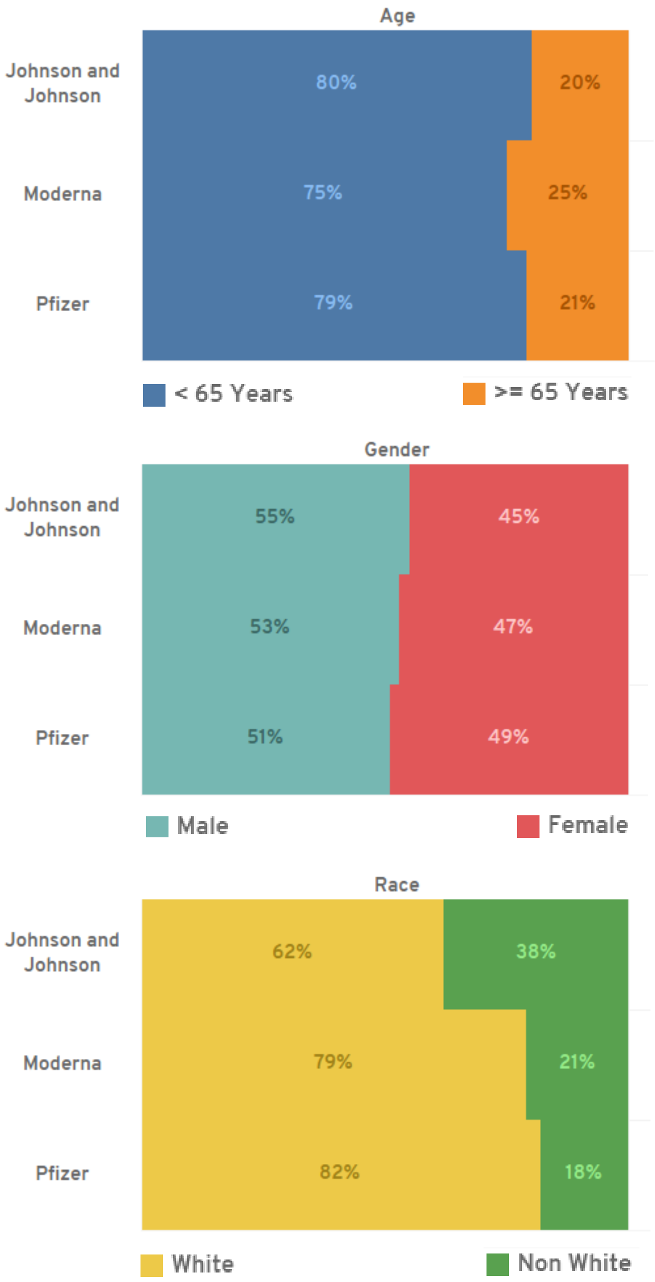
Source
WHO
© Statista 2021

Additional Information:
United States; January 20, 2020 to June 23, 2021

Vaccine trial windows and daily Covid19 cases in the USA

Let's say the claimed efficacy of the vaccine is 95%; this means the relative risk of vaccinated to placebo group is 0.05. This indicates the risk of infection in the vaccinated group is 20 times less compared to the unvaccinated group. This should not be mistaken to 95% of those who got vaccinated will be resistant to the infection. The question is whether the efficacy is a rightful metric to compare the vaccines? Absolutely yes, provided if the vaccines have had their trials at the same time and with a similar volunteer pool. Vaccine efficacy does depend on the time of trials and volunteer pool.

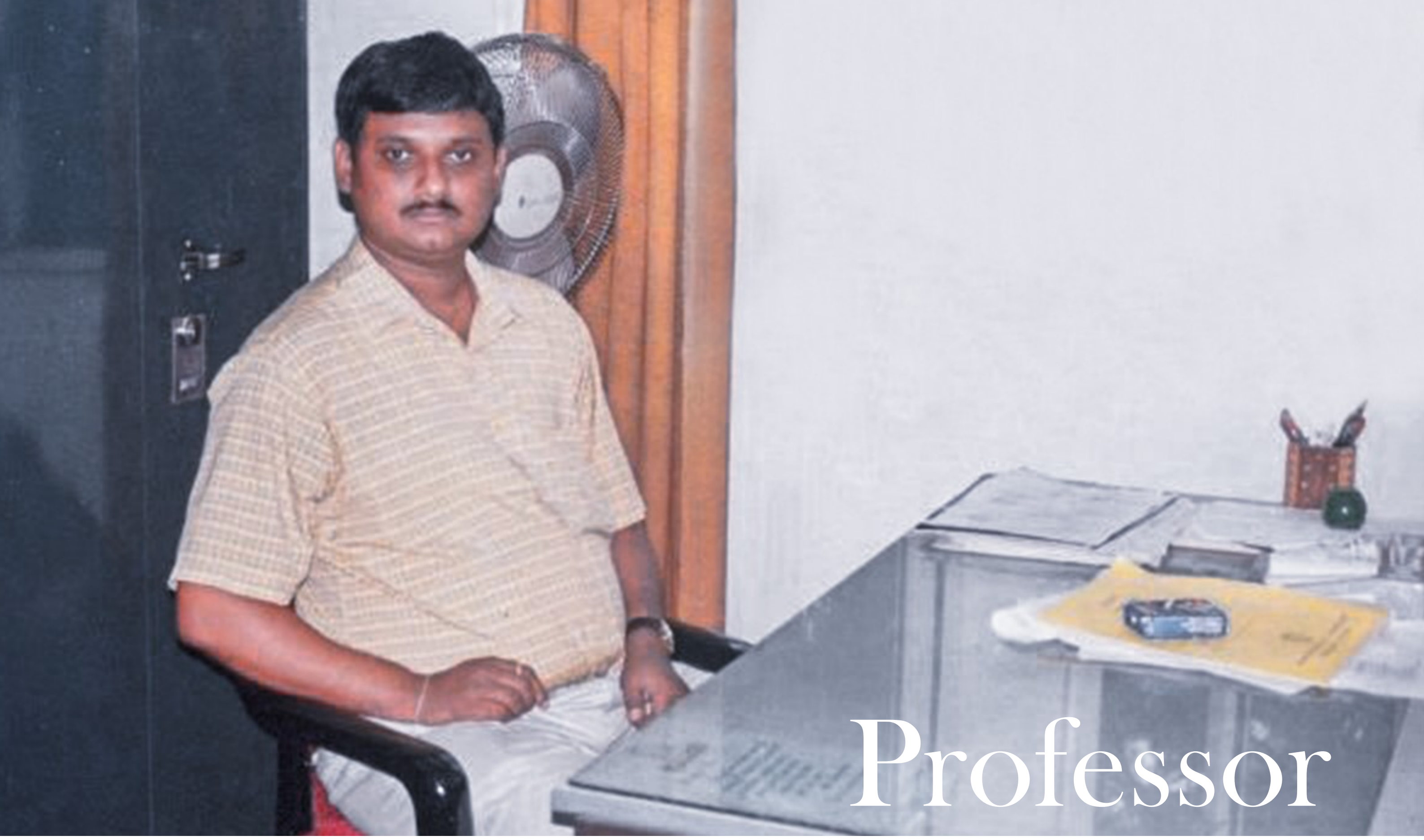
The Phase-3 trials of Pfizer and Moderna were conducted from the end of July 2020 to mid of November 2020. For Johnson and Johnson vaccine, it was conducted from late September 2020 to the end of January 2021. From the figure, it is visible that the vaccine trials of Pfizer and Moderna have started when the daily cases are reducing, whereas for the J&J vaccine, the trials have taken place when the daily cases are increasing until towards the maximum peak. The average daily cases during the trials of the former two vaccines are less compared to the latter, i.e., J&J Vaccine.



Volunteer pools for the three vaccines

The volunteer pool used for the Johnson & Johnson vaccine trial is significantly diverse compared to that of Pfizer and Moderna. Moreover, Pfizer and Moderna vaccines are tested far before the emergence of troublesome variants. Therefore, it is not sure how well these vaccines perform on the new variants. On the other hand, J&J was conducting its vaccine trials during the period of new variants.

The three vaccines cannot be compared based on efficacies as they had their trials at different times of pandemic, used dissimilar protocols, and tested on different pools of people across the world. Here the important metric is the number of hospitalizations, which all three vaccines could equivalently reduce. The efficacies can be directly compared only when there is a head-to-head clinical trial. In the USA, J&J has witnessed a decline in demand for their vaccines. The number of people who opted for Pfizer and Moderna is 11 times more than the number of people who opted for the J&J vaccine. The possible reason is majorly due to the inherent number bias among the people, comparing the vaccines purely through efficacy score.



# Professor Pabitra Mitra

Dr.  Pabitra Mitra is a senior professor of Computer Science Department at IIT Kharagpur. His research interests include Artificial Intelligence and Machine Learning, Data and Web Mining, Communication Medium and Technologies: Language, Speech and Human-computer interaction. He is a recipient of various prestigious awards like Yahoo Faculty Award 2013, IBM Faculty Award 2010, Indian National Academy of Engineering Young Engineer Award 2008 and  Royal Society UK India Science Network Award 2006

**AINA:** After your Graduation from IIT Kharagpur, you have worked for the Centre for Artificial Intelligence and Robotics (CAIR). What nudged you towards a research career in Artificial Intelligence?

**Prof:** I decided that I will pursue research as a career during my B. Tech days itself. But to gain some experience related to how things work in practice, I worked for two years in the industry. Being an electrical engineer, I was interested in control systems and joined CAIR, a DRDO Lab. We used to work on flight control problems and others. At that time, Professor Vidhya Sagar and others were originally control system people started working on artificial neural networks. I got interested in this topic and decided to do a Ph.D. in ISI Calcutta as it had a strong group in neural networks at that time. My work was not only on the neural networks but also on Statistical pattern recognition, which is now known as Machine Learning.

**AINA:** Which is the most disruptive development in the field of analytics in the last decade, according to you?

**Prof:** Of course, it is Deep learning. This development improved the accuracy or range of problems it could tackle and led to a paradigm shift in the way we think to develop a solution to the problems. Earlier, people used to do feature engineering, but the neural network takes care of those things with the advent of deep learning. But this has brought up another set of considerations, such as whether this approach is feasible, how much computational resources are required, and how to deal with overfitting. Solving the ML problem in 1990 is different from 2020 as we have different tools. It's like how planning a travel route by bicycle is different than by a car. The bike has the advantage of not needing any fuel but has limits to the speed it can reach.

**AINA:** In this digitized era of a surplus of resources, what mindset should people have to navigate through the vast ocean of knowledge in their analytics journey?

**Prof:** There are many good resources on particular topics that are widely available but what is lacking is a learning path and organization of resources. It is crucial to know where to start and what sequence of things to be studied to learn a subject well. This goal can be achieved by following materials curated by teachers and experts of the respective field, like standard courses in universities, summer schools, Bootcamp, etc. Once you learn a new subject in this way, you can explore any other topics you like or require for your work from the internet.

> "Anyone can enter this field but to thrive in it, I believe they should have liking to statistics and have ability to map a real-life problem to mathematical problem."

**AINA:** You have been teaching numerous students and researching for around two decades in this rapidly evolving field of AI. What suggestions do you have for those who want to enter and thrive in this field?

**Prof:** Anyone can enter this field, but to thrive in it, I believe they should like statistics and have the ability to map a real-life problem to a mathematical problem. In this field, programming or algorithmic efforts required are not as complex as operating systems or computer networks. Here programs are based on principles of mathematics and statistics. Also, you should be open-minded and have a knack for applying these theories to practical problems. These qualities are also traits of a successful statistician. These capabilities do not come in a day or only by studying books. It requires seeing how algorithms work in the actual field. As these are empirical problems, we need to run them and understand what works and what does not.

**AINA:** As said in chaos theory, a flap of a butterfly's wings can cause a tornado on the other side of the world. So, even a minor change can have an immense impact on climate. In such a complex problem scenario, how can we make a model to study and predict climate? What role could Artificial intelligence play in Climatology?

**Prof:** People use machine learning or any statistical technique when the person solving the problem does not have complete knowledge about it. On the other hand, when we solve a physics problem, our goal is to have no unknowns. For example, if you are solving the mechanics of an airplane, you will know most of the things and can write equations, but on the other hand, if you want to know whether someone likes a movie, you cannot write equations for it.

Though climate science lies on the physics side, equations in climate science are too complex to solve without error. We have climate models, and even now, all the real-life applications are never based on machine learning but are based on physics models. ML is not a primary technique but complements the physics model and fills the lacuna in it.

**AINA:** AI systems are currently very good at specific tasks, but do you think machines can exhibit general intelligence?

**Prof:** The current state of AI is that we are building successful systems in a single domain or a small area. General is a relative term. We, as engineers, feel that if you integrate several small AI subsystems, you get general intelligence. A phase has already started where we are trying to integrate several small AI systems into a larger AI system that we can call a general intelligence system. The more components you integrate, the more general it becomes.

**AINA:** Social media has become an inseparable part of modern life. In your book "Link Prediction in Social Networks", you have discussed various techniques for predicting which nodes in Social networks might be similar. Can you discuss some applications of this area of Link prediction?

**Prof:** I think one area you are already aware of is friend recommendation on Facebook. You can think of Facebook as a big graph or network where each user is a node, and the friendship denotes a link between two nodes. So, you can call the link prediction in this case as friend prediction or friend recommendation. There are many other applications where link prediction can be used, such as in biology, you can predict which genes interact, where the interaction between genes is a link. It is very much used in social network problems and recommendation tasks. For example, in an eCommerce site, where the user is a node and products are other nodes, link prediction helps recommend the product the user is likely to buy.

**AINA:** You have many exciting publications in Health Informatics. Can you kindly let us know what drove you to research this area?

**Prof:** Health is one of the frontiers of human knowledge. It is our health which we know the least about. The research and

“One of the goals of AI in healthcare is to make it more affordable and widespread so that it can reach even the remote regions of the country.”

applications in this field can help humanity a lot.

I was very excited to work in this area because many applications and challenging problems in this field could be solved using AI. Health is a serious topic and cannot be done at a superficial level as it will not be meaningful. It requires a lot of time and effort. Understanding the clinical and biomedical things is needed before we apply any AI techniques. So, working in collaboration with a biologist or a doctor will help a lot.

We are primarily looking into two types of problems in this field. The first type is biomedical imaging, where we analyze and get inferences from the images. The second type of problem includes genomic analysis and metabolic analysis.

For example, when looking at the patient record, a reputed cardiologist can look at 20 to 25 factors, but an AI-trained model can look into 100’s of factors. Human performance varies from person to person, and there can be locations where cardiologists are not present at all. So, the introduction of AI-trained models can revolutionize the diagnosis industry.

Currently, ML is not that much developed to capture every complexity of health, but there is hope that in the future, as more people research in this area, this field will improve.

**AINA:** During the second covid wave in India, we observed a significant disparity in healthcare across the country. What role could AI play in supporting the administration to address this issue?

**Prof:** Currently, AI is mainly used for screening while a doctor still makes the final decision. One of the goals of AI in healthcare is to make it more affordable and widespread so that it can reach even the remote regions of the country. But the digital divide poses a severe challenge as it will be unfair to many people who don’t have a mobile or any other digital device. As we overcome this digital divide, AI would make healthcare more accessible.

These are very challenging problems, and the solution which works in our country won’t come in a day. Many small problems need to be solved before achieving these broad goals. It would take effort from a community of several researchers and technologists to solve these problems, and one should strive to be part of such an effort.

“If an expert human being employed to do a job could have improved the service or a product, there is a scope for AI in that application.”

**AINA:** A common issue these days is that people look at AI/ML as a one-stop solution to every problem. Is there a framework to evaluate the potential of AI in an application?

**Prof:** There is a nice thumb rule for this problem. If an expert human being employed to do a job could have improved the service or a product, there is a scope for AI in that application. For example, in medical imaging, a doctor can identify pneumonia while looking at chest X-rays. We can approximate an AI system to do this task. The whole purpose of AI is to try to mimic a human. I would also like to caution about the false expectation that any problem can be solved with a lot of data. If the underlying phenomenon does not have a pattern, then no amount of data would help you.

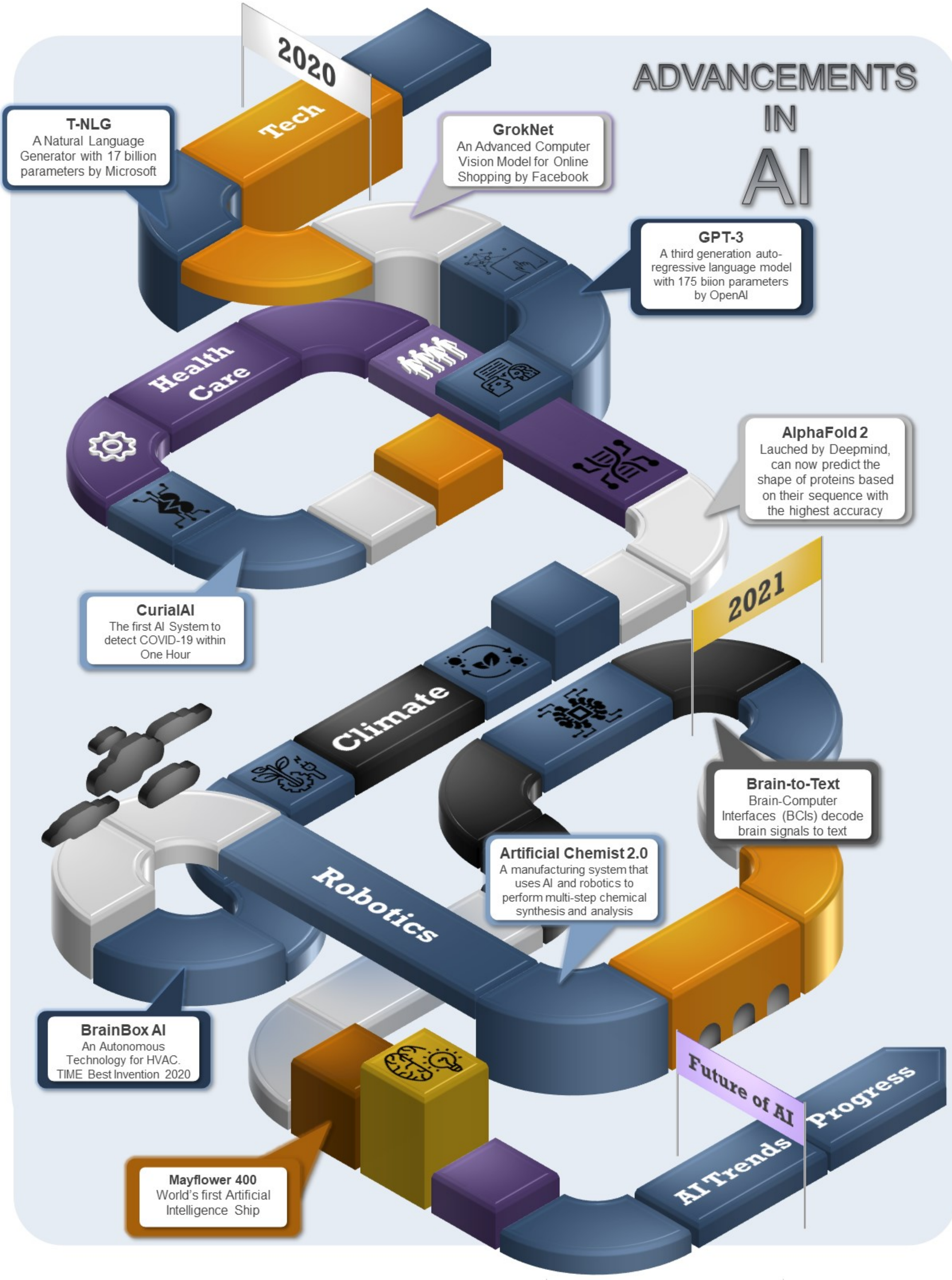
**AINA:** Which subfield of AI do you think will be the hot area in research and industry in the next few years?

**Prof:** All areas of AI! In the near future, many new developments are opening up in all areas of AI. I feel healthcare, education, telecom, and transportation would be a big thing, social media may be saturated, but I may be wrong. Many devices would have a lot of AI built into them. These are some of the immediately visible areas. In the next ten years, AI will be more widespread. I foresee that it will facilitate new research and developments in algorithms, systems, edge devices, IoT, and many other areas.

**AINA:** Thank you so much, sir, for sharing your valuable insights and experiences. We have got to learn a lot from you.

**Prof:** I also enjoyed interacting with you. Thank You & take care.

# ADVANCEMENTS IN AI



**T-NLG**
A Natural Language Generator with 17 billion parameters by Microsoft

**GrokNet**
An Advanced Computer Vision Model for Online Shopping by Facebook

**GPT-3**
A third generation auto-regressive language model with 175 biion parameters by OpenAI

**AlphaFold 2**
Lauched by Deepmind, can now predict the shape of proteins based on their sequence with the highest accuracy

**CurialAI**
The first AI System to detect COVID-19 within One Hour

**Brain-to-Text**
Brain-Computer Interfaces (BCIs) decode brain signals to text

**Artificial Chemist 2.0**
A manufacturing system that uses AI and robotics to perform multi-step chemical synthesis and analysis

**BrainBox AI**
An Autonomous Technology for HVAC. TIME Best Invention 2020

**Mayflower 400**
World's first Artificial Intelligence Ship

**Disclaimer**

- The information in the infographic is not an exhaustive list of recent AI advancements. We highly encourage the readers to explore further depending on one's interest.
- The **Sources of Information** are given in the **References** section at the end of this magazine

https://ainapgdba.ml
magazineteampgdba@gmail.com

# Parts of Speech Tagging using Hidden Markov Model

Prasun Kumar

In this article, we are going to learn one of the most important parts of a natural language processing pipeline, Parts of Speech tagging. Due to the complexity of the English language, it's very important that computers learn the context of each word. Parts of speech tagging is used to tag the parts of speech of the words in a sentence based on the context. For example, consider two sentences:

1. The computer is not able to understand languages because **it** is too dumb.
2. The computer is not able to understand languages because **it** is too complex.

What does **it** refer to in the above two sentences? In the first sentence, **it** refers to the computer, while in the other, **it** refers to the language English. In this example, the part of speech of **it** is the same, but due to a different context, the meaning changes. Take, for instance, the following examples:

1. Snorlax is sleeping
2. Sleeping is a boon

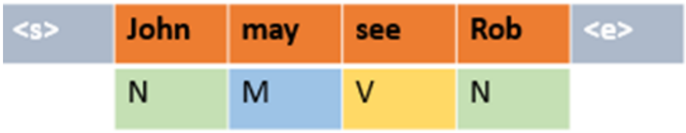
In these examples, the same word **sleeping** has been used as a verb (in the first sentence) as well as a noun (in the second sentence). So, it is important for a computer to understand the parts of speech (PoS) of a word. There are various methods of PoS tagging such as lookup tables, n-grams, Hidden Markov Model, and Viterbi algorithms.

Before understanding these methods, let's look at some of the terminologies. To start with the PoS tagging problem, we need to have a training set of many sentences in which we know the parts of speech of each word priorly. The collection of these sentences is known as text corpus.

Lookup tables and N-grams have some serious limitations. In lookup tables, each word will get tagged to one and only one part of speech each time, irrespective of the context. In the case of n-gram, it is possible that we will get some new combination of words in our test set, and thus n-gram will not be able to tag the PoS for these cases. These limitations make lookup tables and n-grams relatively less popular. Hidden Markov Model and Viterbi Algorithm take care of this problem, and we will understand how they work in this article.

## Hidden Markov Model

We will take an example to understand the working principle of these methods. Let us consider a sentence: ‘John may see Rob’. In this sentence, let us say that a way of tagging PoS is as follows: **John** is a Noun (N), **may** is a modal verb (M), **see** is a verb (V), and **Rob** is a noun (N).



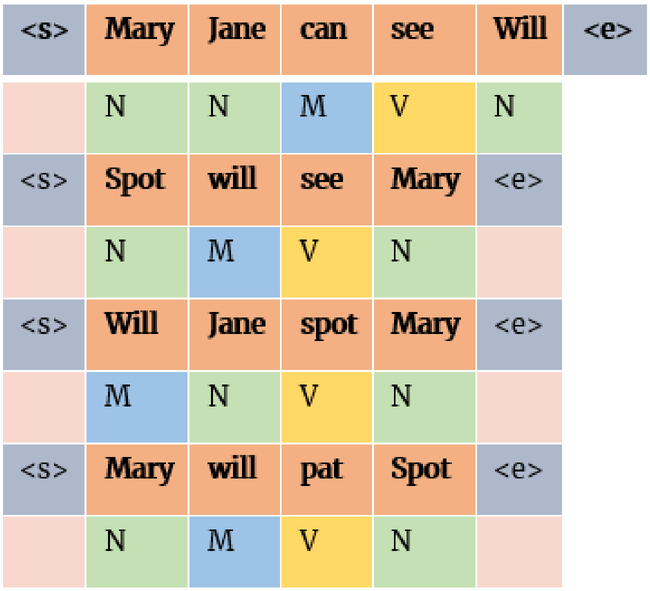
| | | | | | |
|---|---|---|---|---|---|
| <s> | John | may | see | Rob | <e> |
| | N | M | V | N | |

Figure 1: Example of a sentence with parts of speech tags

Our aim is to calculate the probability associated with the above tagging. To find this out, we need two sets of probabilities: transition probability and emission probability. Transition probability tells us about the chance of occurrence of a part of speech after another part of speech, while emission probabilities tell us about the chance of occurrence of a particular word corresponding to a part of speech.

In the above example, transition probabilities include what is the probability that a Modal is coming after a Noun, a Verb is coming after a Modal and a Noun is coming after a Verb. Emission probabilities include the chance that a Noun will be the word **John**, and a Verb will be the word **see**, etc. For the correct tagging, we want overall probabilities (multiplication of all prob.) to be higher.

## Emission Probabilities



| | | | | | | |
|---|---|---|---|---|---|---|
| <s> | Mary | Jane | can | see | Will | <e> |
| | N | N | M | V | N | |
| <s> | Spot | will | see | Mary | <e> | |
| | N | M | V | N | | |
| <s> | Will | Jane | spot | Mary | <e> | |
| | M | N | V | N | | |
| <s> | Mary | will | pat | Spot | <e> | |
| | N | M | V | N | | |

We need our training corpus to have all the PoS tags so that we can find emission and transition probabilities. Let us consider four sentences, as well as their parts of speech associated with each word. The example has been taken from NLP course at Udacity. <s> and <e> denote starting and ending tags.

We will find the probability of each word being a particular part of speech using the above information. For example, **Mary** is occurring 4 times as Noun, in the above corpus, and there are 9 occurrences of words which are Noun, so the probability that a Noun will be the word **Mary** is equal to 4/9.

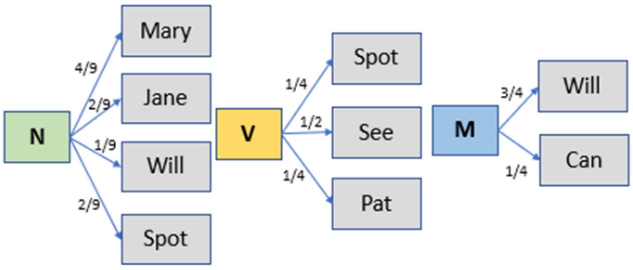
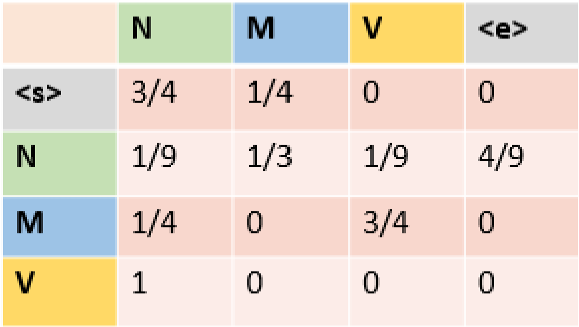


Figure 2: Emission Probabilities

In the same way, we find all the probabilities and represent them as follows, which is called emission probabilities.

## Transition Probabilities

This is the set of probabilities of one part of speech following another. For example, in the above corpus, 'Noun followed by Modal' occurs three times (in first, second, and fourth sentences). In total, Noun is followed by Noun once, by Modal thrice, by Verb once and by end-of-sentence four times. Thus, the probability that Modal occurs after Noun is 3/9 = 1/3. In the same way, we calculate probabilities for all combinations, and summarize them in the following diagram, which represents transition probabilities:



| | N | M | V | <e> |
|---|---|---|---|---|
| <s> | 3/4 | 1/4 | 0 | 0 |
| N | 1/9 | 1/3 | 1/9 | 4/9 |
| M | 1/4 | 0 | 3/4 | 0 |
| V | 1 | 0 | 0 | 0 |

Figure 3: Transition Probability

Now we have both emission and transition probabilities. We will proceed to see them in action. The words which are there in the corpus are called observations because these are the things that we can observe. But, the parts of speech of each word are hidden to us and not directly observable, so we call them hidden states. There are 9 observations and 3 hidden states in our corpus. Each hidden state (PoS) is connected to every other hidden state, with the transition probability, and each hidden state is also connected to every observation (words) by emission probabilities. The following diagram describes this relation:

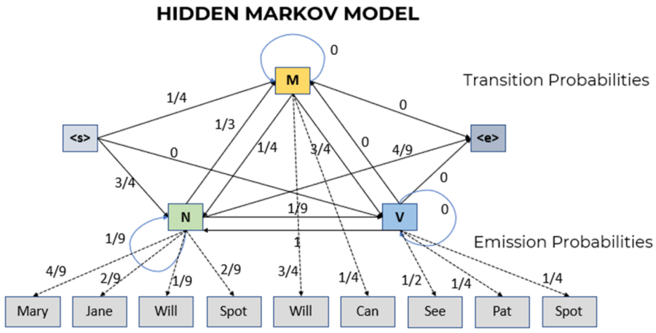


Figure 4: Hidden Markov Model

In the above diagram, values on solid arrows show the probability of a part of speech coming after another part of speech (transition prob.), while the numbers on the dashed arrow show the probability that a noun is the given word (emission probability).

The Hidden Markov Model can generate all sentences based on the sequence in which we travel from one state to another. Here state refers to the parts of speech N (noun), M (modal verb) and V (verb), start-of-sentence (<s>) and end-of-sentence (<e>). Let's consider an example where we want to generate the sentence 'Jane will spot Will.' Let us see in how many ways we can generate this sentence using the above model.

We will start from <s>. One of the ways is to reach Noun (N) with probability 3/4. We can pick the word **Jane** with probability 2/9, and can move to Modal (M) with probability 1/3, and can pick **will** with probability 3/4. We can move to Verb (V) with probability 3/4, and can pick **spot** with probability 1/4, and can move to Noun (N) with probability 1. Then, we can choose **Will** with probability 1/9. Finally, we can reach the end-of-sentence with a probability 4/9. The above few sentences might be complex to understand, so let's have a look at the following diagram to understand the flow. This is one of the many ways to generate the sentence:

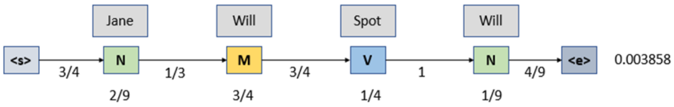


Figure 5: First possibility of occurrence of the above sentence

Moving from one state to another is independent of other states, so we can multiply all these probabilities to calculate the probability of the above combination of words and parts of speech. For the above case, we obtain 0.0003858 after multiplying all the probabilities. There are other ways in which we can generate the same sentences. Let's have a look at one of them:

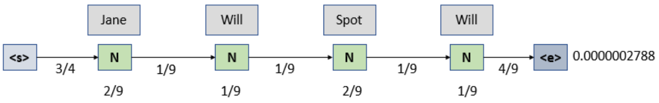


Figure 6: Second possibility of occurrence of the above sentence

The above possibility occurs when all of the words are Noun. This sentence makes no sense in the real world, and thus we have also obtained very low probability. We can check all possibilities in which the above sentence can be generated from our Hidden Markov Model and calculate the likelihood for each one of them. We will ignore those paths in which there is at least one 0 probability edge because these paths will not be possible. Apart from the above two already discussed paths, there are two more paths, whose likelihood is also shown below:

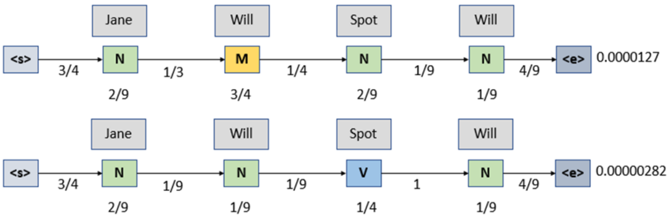


Figure 7: Third and fourth possibility of occurrence of the above sentence

Out of all 4 possibilities, we find that the likelihood of the first possibility is highest. Thus the PoS tags in that sentence will be reported as the correct PoS tags. Choosing that combination that has the highest likelihood is called the maximum likelihood principle, which is widely used in many machine learning algorithms. Based on the above values, we will report that, in the given sentence, **Jane** is Noun, **will** is Modal Verb, **Spot** is a Verb, and **Will** is a Noun. Thus, we are able to find the correct parts of speech of each word in a sentence.

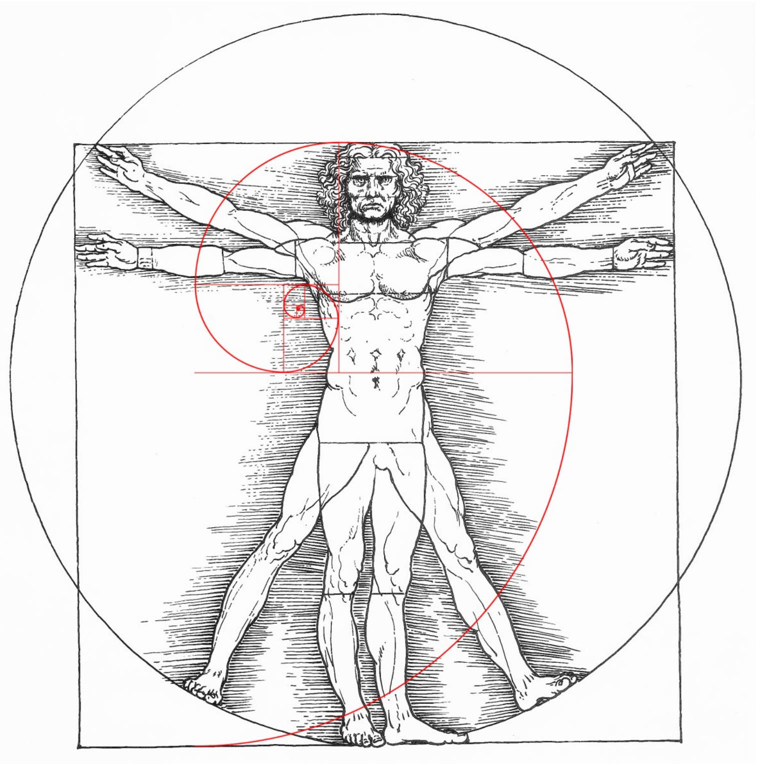
In this article, we learned the application of Hidden Markov Models in Parts of Speech tagging in simple words. To further improve the Hidden Markov Model, we use the Viterbi algorithm, which uses dynamic programming to reduce the calculations required in the above method.

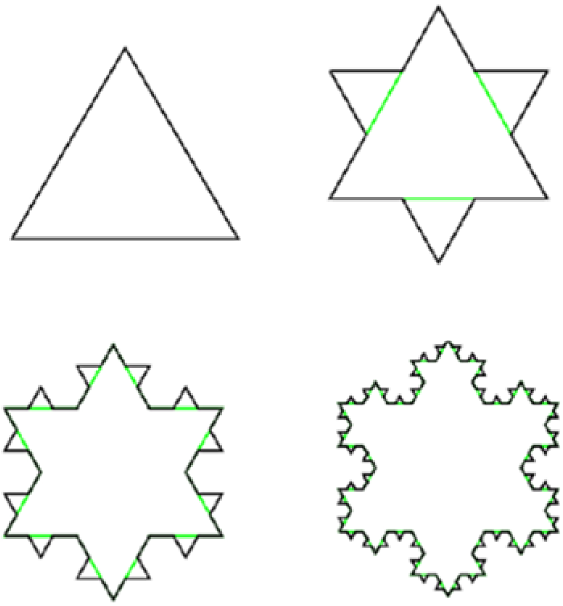
# Art is Math

Raj Chauhan

Memes questioning the use of math in real life are well-liked, and while I did understand that they were nothing more than harmless jokes poking fun at math, I would always be on the lookout for numbers in real life to verify if there was any truth to those memes. While reading the ever-popular ‘The Da Vinci code’, we all stumbled upon the influence of math on a not so glaringly obvious subject – Art.

**And not just influence it, math through AI is now creating art.**





# Math in Art

By observing patterns in nature, we identified the mathematics behind them to understand what catches our attention. The golden ratio (phi or 1.618) occurs in nature in several patterns, and artists used this ratio in their works, from Leonardo da Vinci to the builders of the great pyramids of Giza.

The golden ratio known as the divine proportion during the renaissance appeared in several Da Vinci works. He also created illustrations for the book De Divina Proportions by Luca Paoli.

A 2021 research study by the University of Oregon claimed that babies by age 3 prefer fractal patterns found in nature. Fractals are complex repetitive patterns that maintain a spatial symmetry at all scales.

Children up to this age live in structures devoid of these fractal patterns. Their houses employ straight lines in their architecture. The study suggests that the human brain may have evolved to consider fractals as pleasing to the eye and not something it learns to admire.

An example, the Koch Snowflake is a fractal curve created by repetitions of equilateral triangles. It is developed by Helge von Koch.

Mathematics is prevalent in art and aesthetics. Can we use mathematics to enhance art?

## Engineering Aesthetics

Today people spend a lot of time staring at screens. User Interface or UI designers follow a few mathematical rules while designing an app or any webpage; to compose interface elements design for users to perceive the information efficiently. The golden ratio also appears in UI design to create rectangular design elements and font size selection. To select two different font sizes: for a heading and a subheading, designers first fix one of the two according to visibility requirements. For a bigger font, the other font size is multiplied by 1.618, and for a smaller font, it is divided by 1.618.

The rule of thirds is another practice that enhances designs. Four lines divide the digital canvas into three equal-width horizontal and vertical segments. Designers place the most impactful elements along these lines or their intersections.

Creative use of aesthetics can enhance even everyday PowerPoint presentations. One such application is colour palette generation. An average analyst has limited design experience and uses application suggested colours or presets to make his presentations. Automatic colour palette generators help these users to choose from a seemingly infinite variety of colour palettes without being an expert in creative disciplines.

The online generators are programmed to bundle colours that look good together. A few of these are AI-powered generators trained with datasets of images of photographs, modern art and movies. A few of these are AI-powered generators trained with datasets of images of photographs, art and movies. Mathematics, and now AI provide good design practice guidelines to enhance the user experience. But we can use design to enhance the presentation of mathematics too.

## Visualising Mathematics

Data visualisation or infographics make it easier for a user to understand statistical or quantitative results in a report. Generic visualisations of the data achieve the objective of explaining the data but a well-designed and planned presentation results in more attention paid to the reports.

Several companies have now identified the need of having a structured design approach. They have formulated detailed guidelines on designing their infographics. These guidelines include having their colour palettes to maintain uniformity over the designs of presentations and brochures. Design principles, like the use of whitespace, and layout styling, helps companies to publish visually pleasing reports.

Apart from aesthetics, art is a medium for conveying a message from the artist to the audience. The artist is a storyteller through his/her art. Storytelling is also an aspect of data visualisation. Every effective visualisation must contribute to the explainability of the problem highlighted by the report, rather than simply acting as a graph in it.

The infographic by Charles Minard on the invasion and retreat of Russia by the French army led by Napoleon (flip page to view the infographic) is a masterpiece in storytelling through an infographic.

In his book 'The Visual Display of Quantitative Information', Edward Tufte called the infographic the best statistical graph to have ever been produced.

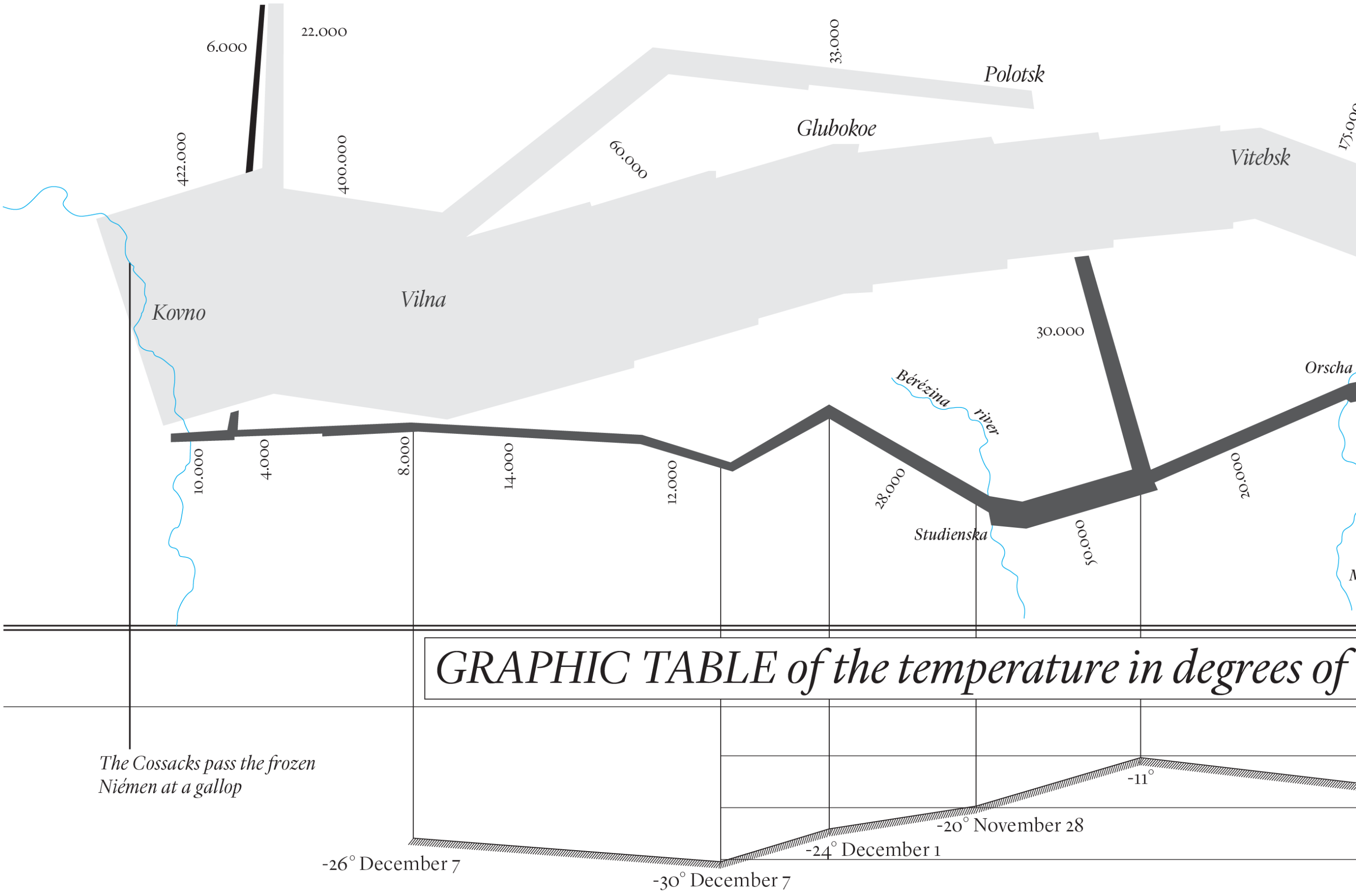
At first glance, the infographic appears to be two line graphs with varying width. When the significance of this width dawns on the user, the infographic transcends into a work of brilliance...

# The greatest infographic

## Napoleon’s march of  Russia by Charles Minard

*FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAM*

*Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement. Paris, 20 November 1869. The numbers of men present are represented by the widths of the co men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the arm under Marshal Davoust, who were sent to Minsk and Mobilow and who rejoined near Orscha and Witebsk, had always marched w*



# Legend

⬜ French army moving into Russia

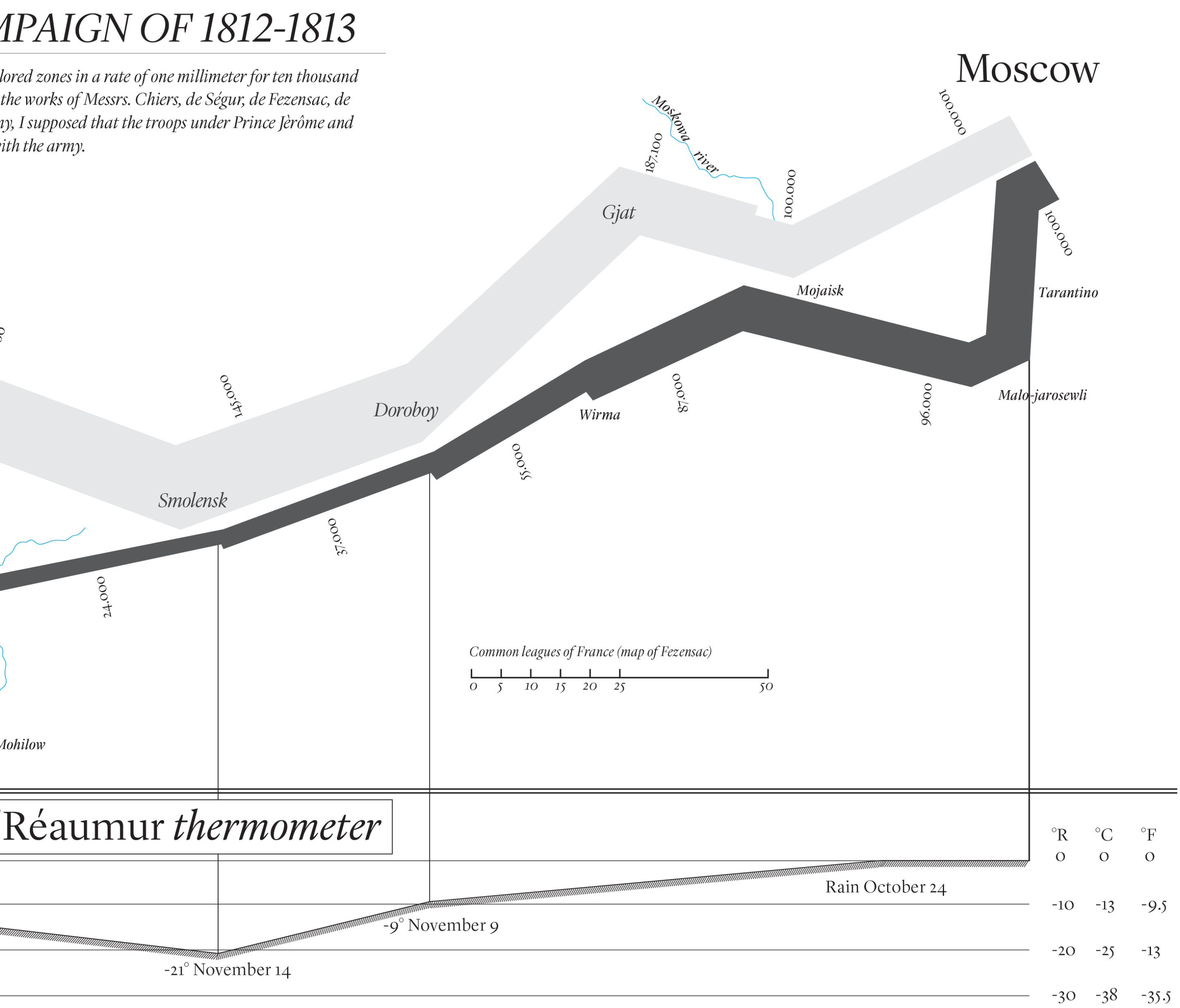
⬛ Retreating French army

width ↔ Number of soldiers

When Napoleon’s French army begun its campaign to Russia with 475,000 men, the Russian army of about 200,000 men stood no chance of a victory. Russia employed an attrition warfare of scorched earth policy leaving Napoleon to rely on his supplies which was incapable of supporting his large army.

By the time the French reached Moscow, their numbers had dwindled to just a 100,000 men owing to lack of supplies, the harsh Russian weather, and two battles.

Napoleon had reached Moscow victorious in his battles, but the Russians left Moscow, burning it down while retreating. Napoleon could not keep up his supplies and



had to begin his retreat as Russian winter was setting in. The temperatures are displayed in the chart below the army movement graph.

By the time they retreated back to France, only 10,000 of the original 475,000 soldiers survived.

Our mind often finds it difficult to interpret large numbers. The infographic, by using the width of the lines for army strength, tries to overcome this difficulty and portray the true nature of this campaign.

The infographic displays 6 variables:

- The size of the army
- The direction of movement
- The relative location of places during the campaign
- Temperature
- The dates of the campaign.

With the amount of information conveyed, Edward Tufte was correct to call it *'the best statistical graph ever drawn'*

## AI as an artist

In 2017, the auction house Christie's auctioned Salvator Mundi by Leonardo da Vinci for $450.3 million at their New York premises, making it the most expensive painting. Garnering equal discussion the very next year, Christie's at their London premises sold a portrait of a man named Edmond de Belamy for $432,000.

Edmond de Belamy is a fictional persona and the work of Artificial Intelligence. The signature on the painting is a mathematical formula: the loss function used to create the artwork. A French group called Obvious, who study interactions of art and AI, used Generative Adversarial Networks or GANs to create the portrait. They trained the network with 15000 renaissance portraits for the GAN to create the artwork.

## How do GANs create art?

A GAN architecture consists of two parts: a generator and a discriminator. The generator is fed with random data or noise and generates an image. This image is then looked at by the discriminator, which uses the training dataset given to it to classify whether the generated image looks like the dataset or not. It discriminates the true art (in the training dataset) from the fake art (from the generator).

The generator then takes this as a feedback and tries to create a better fake image to confuse the discriminator. It treats the discriminator as an adversary which it has to fool. The network training is a min-max game: the generator tries to maximise the discriminator loss. The discriminator tries to minimise it.

The result is an equilibrium where the discriminator can no longer identify fake images given by the generator. As the discriminator cannot distinguish the GAN made art from the training set provided, it retains the art style of the training dataset and does not create its art style.

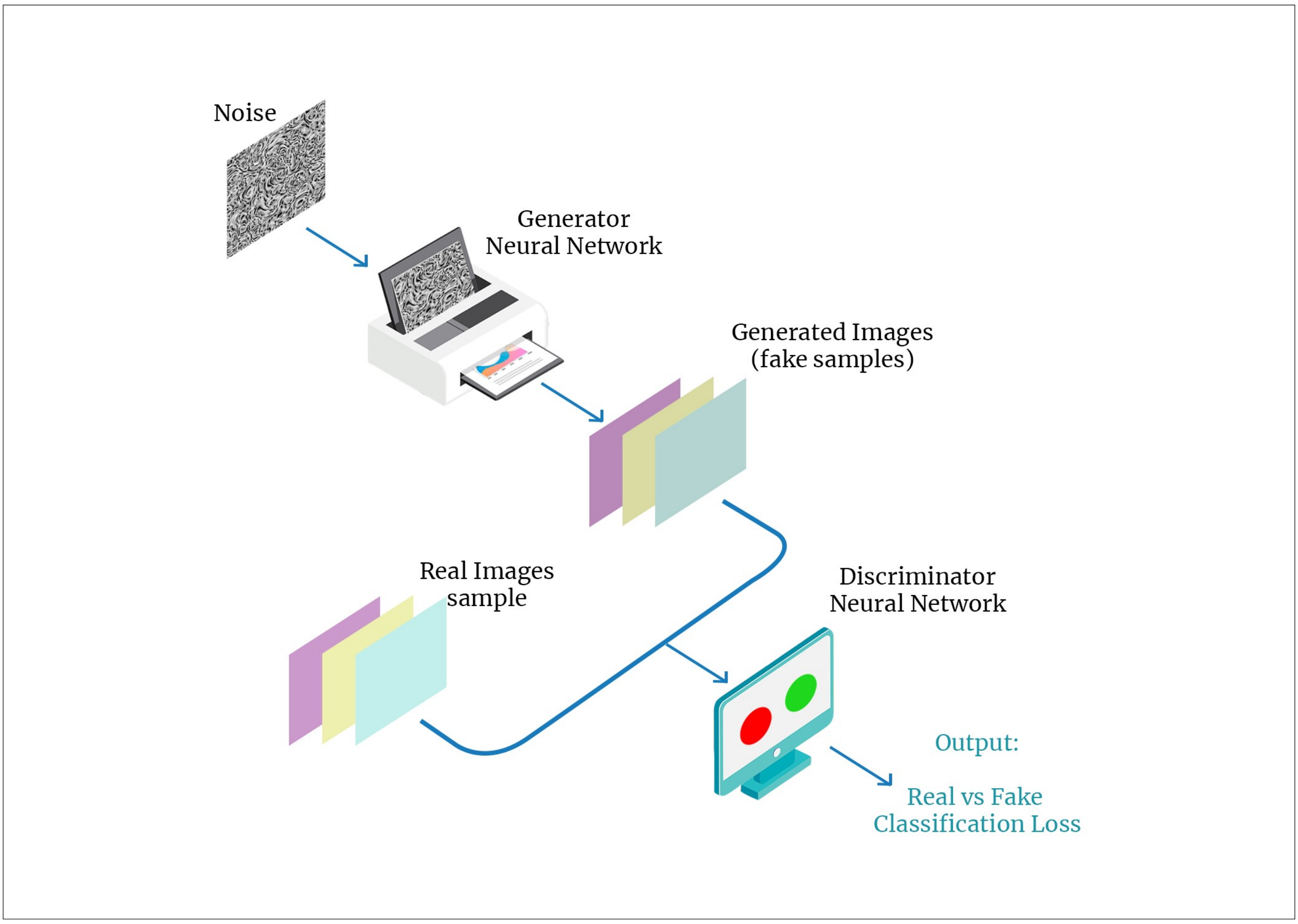
Ahmad Elgammal, a Rutgers University computer science professor, calls GAN made artwork simply repainting. GANs, while trying to mimic an art style, lose the essence of creativity that humans possess.

Elgammal, along with his fellow researchers at Rutgers University AI and Art gallery, came up with an improved architecture for AI-based art, Creative Adversarial Networks or AICAN.

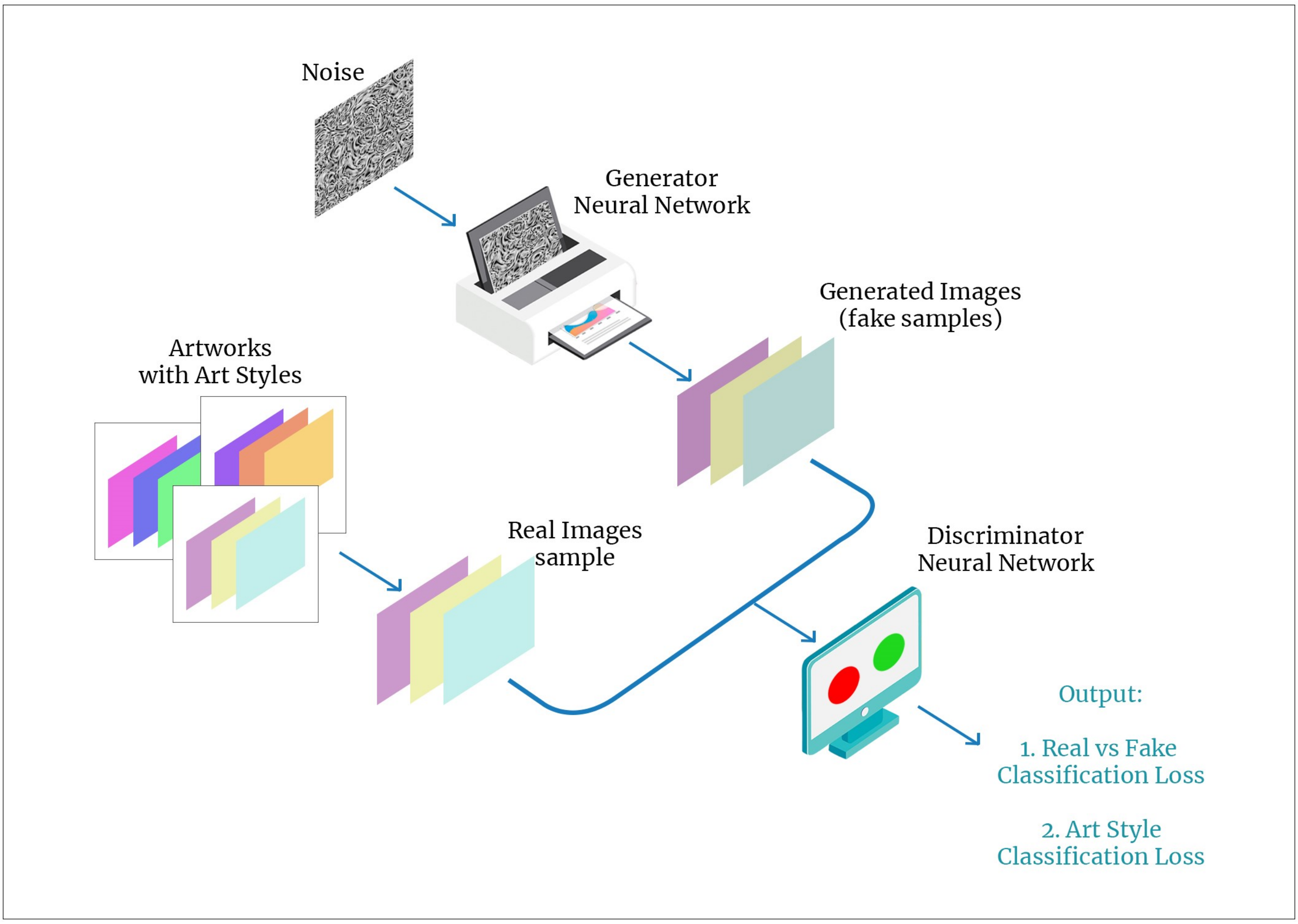
The AICAN, like GAN architecture, involves the generator and the discriminator. But the discriminator in AICAN improves on GAN by providing more feedback. A dataset involving artworks of more than one art style is the training input to the discriminator. When the discriminator receives the generated image, it not only tells the generator whether the image is true art or not, but it also classifies the art style. The generator tries to maximise two losses of the discriminator:

1. Loss of classifying whether the image is true art or fake art, and
2. Loss of classifying art style of the image.

The AICAN successfully confuses the discriminator, to create an artwork classified as true art like the training distribution yet having a unique art style unlike the styles in the training dataset. Essentially, AICAN creates a new art style that looks as real as if humans painted it.



**GAN Architecture**



**AICAN Architecture**

# Can AI replace human creativity?

In his 2019 paper ‘Can Computers Create Art’, Aaron Hertzmann of Adobe Research argues that AI and other computer algorithms are mere tools for artists to use. Even if AIs create new art styles, these AIs have to be trained by humans to generate these styles.

Memo Akten is a London based artist and creator of GHCQ - a painting developed using Google Deepdream generator. He calls Deepdream to be just a better paintbrush developed by google and that the human artist is essential. Not just any image given to the computer algorithm can produce artwork that rivals artists, but human interference in this generation process can produce masterpieces, as Akten himself has shown. Aaron Hertzmann believes: Computers do not create art. The people using computers do. Art is a social practice; artists use their works to express their thoughts and interact with their audience.

While currently, the several variations and ideas in an artist’s mind remain confined within their realm of imagination, AI generators can enable these potential variations to take a physical form on a digital screen, thereby providing the artists with a gallery of exhibits of their creativity.

Some believe that such a tool undermines the human touch behind

**The ceiling of Sistine Chapel, where Michelangelo painted 'Creation of Adam'**

creativity, but art has undergone massive technology-assisted developments throughout its history. Be it the use of colours from looking for sources to extract them from nature to now synthesizing even their tints and shades; from using the material available to now selecting the best suitable material for the artwork. Technological developments have helped artists better express themselves, and AI is one such technological development for artists to use.

When you look at 'The Creation of Adam' by Michelangelo, you do not just see Michelangelo's art style but notice that the shape behind God in the painting resembles a human brain. One of its interpretations is how human sentience is the medium through which God in the artwork connects to Adam – humans. Computer algorithms fail to create this contextual subtext without the intervention of artists.

This sentient creativity separates current AI art and artists. But, a partnership of AI systems and human artists can spark a new direction of co-creativity, where technology learns as an apprentice of these top artists. But with the rapid pace of advancements in AI, the question will forever loom if the apprentice can one day become the master.

# The Realm of Quantum Computing

Utkarsh Yadav

With each passing day, computers are becoming smaller and more powerful than ever before. There has been an exponential rise in computing power in recent decades. But as the size of transistors will reach their physical limits, the number of transistors in a dense integrated circuit cannot be increased further, violating Moore's law in the future. You might be thinking, why is there a physical limit on the size of a transistor? Let us try to understand this.

Transistors are the fundamental building blocks of classical computers. They are electrical switches that allow the passage of electrons. We already have transistors having the size of the order of 10 nanometres. As the transistors become smaller and their size approaches molecular levels, some very non-trivial problems can arise due to quantum effects. The electrons might exhibit quantum tunnelling, i.e., it just disappears from one side of the transistor and reappears on the other side, defeating the whole purpose of the electric switch! This physical limit puts a constraint on the computing power of classical computers, and to overcome it, scientists and researchers are trying to exploit quantum properties and create Quantum Computers.

A classical computer uses 'Bit', which takes a binary value of 0 or 1, as the smallest unit of information. Alternatively, a Quantum computer uses 'Qubit' instead of Bit. A qubit is a two-state quantum system that is a superposition or a mixture of 0 and 1 in a given proportion. One can think of it as a flipped coin that can either be head or tail with a specific pair of probabilities. An example of a qubit can be the polarization of a photon (vertical or horizontal) or the spin of an electron in a magnetic field (spin down or spin up). The state of a qubit collapses either to the values of 0 or 1 only when it is observed. Due to the superposition of states in qubits, quantum computers can represent far more data than a classical computer with bits. For example, 3 qubits represent 8 possible states while 3 bits represent only one among these 8 states. Hence with each unit increase in the number of qubits, we get an exponential increase in the data represented and processed in a quantum computer.

A quantum computer with N qubits can represent $2^N$ states, so with N=300, we have $2^{300}$ states which are already more than the number of particles in the observable universe! Another property of quantum physics that plays an essential role in quantum computing is entanglement. If two particles are in quantum entanglement, the state of one particle determines the state of another particle even if they are far apart.

Classical computers use logical gates to do one computation at a time. Similarly, quantum gates in quantum computers can manipulate the probabilities of qubits using the principles of superposition and entanglement to do multiple computations simultaneously. This way of simultaneous computation in quantum computers can prove to be exponentially more efficient when compared to classical computers.

Quantum computers hold a lot of promise for certain types of computation, which cannot be performed on classical computers. But, such systems require extremely low temperatures and the absence of even minute disturbances to exhibit quantum properties of superposition and entanglement. These conditions make the use of quantum systems very difficult for many practical applications. Another critical aspect of a quantum computer is that it is not universally faster as it does not perform an individual operation fast. Instead, it can perform several operations simultaneously, which is required for certain types of calculations. So, we can say quantum computers will not wholly replace classical computers, but they will coexist and complement each other.

# To be or not to be?

Quantum computers could be used in an extensive range of applications. Its potential use cases could include drug discovery by simulating the interaction of drug molecules and modelling complex systems like climatic conditions or brain activity. They could also be used in computer security by creating a secure channel for communication over long distances by exploiting the principles of quantum entanglement.

Many tech companies like Google, IBM, Microsoft, etc., are already building quantum computers. Google in 2019 demonstrated quantum supremacy on a 54-qubit processor by running a target computation in 200 seconds which would have run on the world's fastest supercomputer in 10,000 years.

Research and applications in quantum computing are still in their inceptive stages, and we are anticipating many more breakthroughs in the coming years. Recent successes have motivated different nations and technology organizations to accelerate their efforts towards the quantum computing race. It would be fascinating to see what the future holds for this technology.

# Sudalai Rajkumar

Sudalai Rajkumar, popularly known as the SRK of data science, needs no introduction. He is among 15 data scientist all over the world who have won prestigious Kaggle Grandmaster title 3 times, in notebooks, datasets and competition section. He has consistently been ranked 1st in Analytics Vidhya datahacks. He is currently working as a Data Scientist at H2O.ai. Before H2O.ai he had worked at various other companies in key positions: as a Lead Data Scientist at Fresh works, Tiger Analytics and lead of R&D at Global Analytics. Its an honor for us to talk to him. Here, we present the excerpt of conversation between AINA team and SRK.

**AINA**: Can you tell us about your journey?

**SRK**: I am working in data science for more than 11 years now. I started my journey as a data analyst back in 2010, and I worked in different types of roles in the data science stream. Data science has not got the level of attraction at that time as it has got now. I have worked in various different domains such as healthcare, retail, financials, customer support, etc. Apart from the professional side, I consistently participate in many Kaggle competitions, from where you guys know me. I started participating in Kaggle completions from 2012-13, and till now, I have won the Grandmaster title in the notebook, dataset, and competitions section of Kaggle. Apart from Kaggle, I participate in other hackathons such as Analytics Vidhya datahacks where I was number 1 for quite some time.

**AINA**: Over these years, you have participated in many Kaggle competitions. Any beautiful memories which you have from these competitions?

**SRK:** That's a good question. There are several memorable competitions, and I have different memories associated with each competition. Rainwater prediction competition is very close to my heart because that was the first competition in which I finished in the top 10. Before this competition, I was not very sure whether I will get good positions at a global level. Honestly speaking, I had some self-doubt, and I am sure everyone has that at some point in time. Before that, I was hesitant to reach out to others to team up. But I reached out to Marios Kazanova (another 2X Kaggle grandmaster and currently ranked 6th in competition section all over the world), & he was top 10 or top 5 at that time also. He was very humble, and he accepted my request. I also learned that people were generally nice and humble, and you should go ahead and ask for help, and they would definitely help you. If we work with people from diverse backgrounds or different countries, we learn different ways of looking at a problem.

**AINA:** SRK, you have stayed in this field for more than a decade. From a beginner to Kaggle grandmaster, how has your approach evolved during this time while solving a problem?

**SRK:** The problems have also evolved a lot during these times. At that time, competitions were relatively easy, to be honest, as we did not have many algorithms. Random Forest was something which we were usually using. Slowly, from 2015 -16, many algorithms for tabular data came out, such as XGBoost, LightGBM, and other variants of

GBM. With time, lots of codes for EDA and modeling are widely available from other people. We had some set of codes ready to do the usual stuff, and most of the time, you just pull a function or class to get things done. But, that part of the code keeps on growing with time. There are always some methods coming up which perform better than the currently existing method, so you add it to your codebase and try it out on other problem to check if it works. After 2016-17, it's all and mostly deep learning models, which are very useful for vision and NLP problems. These days, most of the problems on Kaggle are related to computer vision and NLP, and very less tabular data problems because we have solved most of them already, and people have some standard procedure which can solve them very well in a very short span of time.

> “Feature engineering is more about self-curiosity and how we can help others to understand the data well.”

**AINA:** You had mentioned that you are very interested in doing projects that are very impactful for society. Once, you collected water data from Chennai city when there was a drought to understand the water levels in different reservoirs and suggested that a simple forecast can help mitigate such situations. Your vision of giving back to the community is quite inspirational. What challenges do you face during such projects?

**SRK:** Especially this is very helpful for people who are not already in the industry and want to contribute to society or students who want to do something new. These projects are very helpful for the general public, as well. Datasets like titanic and iris are easily available and good for learning EDA, feature engineering and model building. It's more about self-curiosity and how we can help others to understand the data well. When we do such projects, it's not only about model building. There are a lot of useful steps required for such projects. Picking up the problem, setting up the business context, and then collecting data are crucial steps. I took data from the Chennai metro website, which was not well organized, so I spent a good amount of time exploring the website and scraping the data using Python. In industries, we face such problems quite frequently, and so it's good to have such projects.

**AINA:** You said that data collection is something that is very important. We have been closely following your Covid 19 dataset, which you uploaded last year on Kaggle. We have also done some of our projects using that dataset. It is currently the third most voted dataset on Kaggle and has helped the entire data science and medical research community.

**SRK:** Thanks for the nice words. I am delighted that it helped others.

**AINA:** At H2O.AI, you are working primarily in the NLP domain, and your main task is to stay updated with the recent developments in this field and integrate them into H2O Driverless AI. What are some of the developments in NLP which interest you the most?

**SRK:** If you see, in the last couple of years, there have been many changes in this field due to transformer-based methods. You guys might be learning it or using it in the projects. The BERT-based architectures are transformer-based models; it really changed the landscape of NLP with respect to traditional classification problems. With the advent of these new models, the performance of many tasks has improved as compared to word embeddings and the bag of words approach.

Since language information is already present in these models, the amount of data needed to build a model is also reduced. Apart from categorization, there are a lot of improvements in other areas. Earlier, most of the applications of NLP were for the English language, but now many models are coming up for Hindi, Japanese, Chinese, and others.

**AINA:** In one of your recent tweets, you said that the most underrated quality that a good data scientist has is structural and critical thinking. Many people are struggling with improving this skill. Do you have any suggestions to improve this skill?

**SRK:** For me, it improved gradually by solving many problems on Kaggle. One should try to think of the end objective at each step while solving any problem. It helps in structuring our thoughts. Often, because of not focusing on the end results, we end up spending a lot of time on other things. Another thing one should try is to solve many different problems, which helps develop a better perspective. Rather than directly jumping to solve a problem, take some time, and think about whether it will be helpful to do what you are going to do.

**AINA:** We all have benefited a lot from your starter notebooks. What, according to you, has helped you the most in creating such great EDA notebooks? What are some of the steps that you follow while doing EDA?

**SRK:** It is a kind of a Code Base I have. For Example, if there is a categorical problem, there is something that we need to think through and see the magnitude of categories. If there are around 5 to 10 categories, you can use a countplot and try to understand the underlying intricacies. Else, if there are a large number of categories, we need to understand their distribution. Similarly, for numerical variables, you try to see the number of unique values, missing values, any outliers in the data, etc. These are basics to any EDA, and if you try to work with different datasets, you will get a fair understanding of all the things we look at and most of the time, the frameworks used can be utilized to other datasets too.

**AINA:** People often face a problem where they score well on the public leader board, but when the competition ends, their ranks drop significantly. What are some of the things to keep in mind that help a model in generalizing well?

**SRK:** Cross-validation can help us perform better on the private leader board. But it entirely depends on how we do the cross -validation; and how we are doing the splits so that it matches the dataset. We can do random splits, stratified splits, time-based cross-validation, or take the variable distribution and then split based on it. So, the idea is to create a cross-validation set that mimics the real-world test set. This is something we learn through mistakes. If you don't get it right on the first go or the second go, you don't have to worry about it. This is one of the best learning I had learned from Kaggle. It teaches you structured and critical thinking which are really helpful at every step of the process. So, it's not just the cross -validation that helps but the way we do it matters.

**AINA:** There are various other aspects of Full Stack Data Science like databases, cloud, etc. What areas should one delve into, in terms of model deployment?

**SRK:** There are two ways to go about it. In the first stage, from the Business point of view, if anyone wants to build prototypes to show the management, they use R Shiny. Here at H2O.ai, we have a tool called Wave, an open-source Python development framework that makes it fast and easy for data scientists, machine learning engineers, and software developers to develop real-time interactive AI apps with sophisticated visualizations. The next stage is diving deep into building an ML architecture using Flask so that these models can be deployed in production. We will get help from engineers with such implementation.

> "If tomorrow, someone comes up to you and says that you are an inspiration for them, it's a beautiful feeling."

**AINA:** In recent times, we see a lot of emphases given on data security and privacy which led to a more anonymized data collection process. How would the machine learning models evolve when we have more anonymized data in the future?

**SRK:** Data Privacy is irrefutably important. Most companies don't use private information while modeling, and it has been standard practice at least since 2010. Such information is not being used actively because that creates some bias as well. There is a famous example for this where a credit card company predicted based on gender, and the results weren't accurate. So, in addition to privacy, it is also a concern of working with unbiased data. We need to be careful about data privacy as well as data security.

**AINA:** Enterprise AI is different from AI techniques we apply in our project and Kaggle Competitions. How can one be well prepared before entering the industry?

**SRK:** There are various aspects of Enterprise AI. For Kaggle, we probably start with feature engineering, modeling, and so on, but in the industry, we identify the business problem first and gather relevant data for it. Doing projects that involve gathering data and formulating the business problem will help in this regard. In Kaggle, we aim to improve the metrics, but in industrial analytics, we present the results and recommendations to the board. This will help us gain knowledge on storytelling – the way we communicate the results in the simplest manner possible.

**AINA:** What is your larger vision towards the Data Science Community?

**SRK:** My vision is to try and help others as much as I can. This community will flourish if we help each other grow in whichever capacity possible. If tomorrow, someone comes up to you and says that you are an inspiration for them, that something everyone loves to hear. So, try to impact as many lives as possible positively. If everyone works towards that common goal, the community will grow on its own, is what I believe.

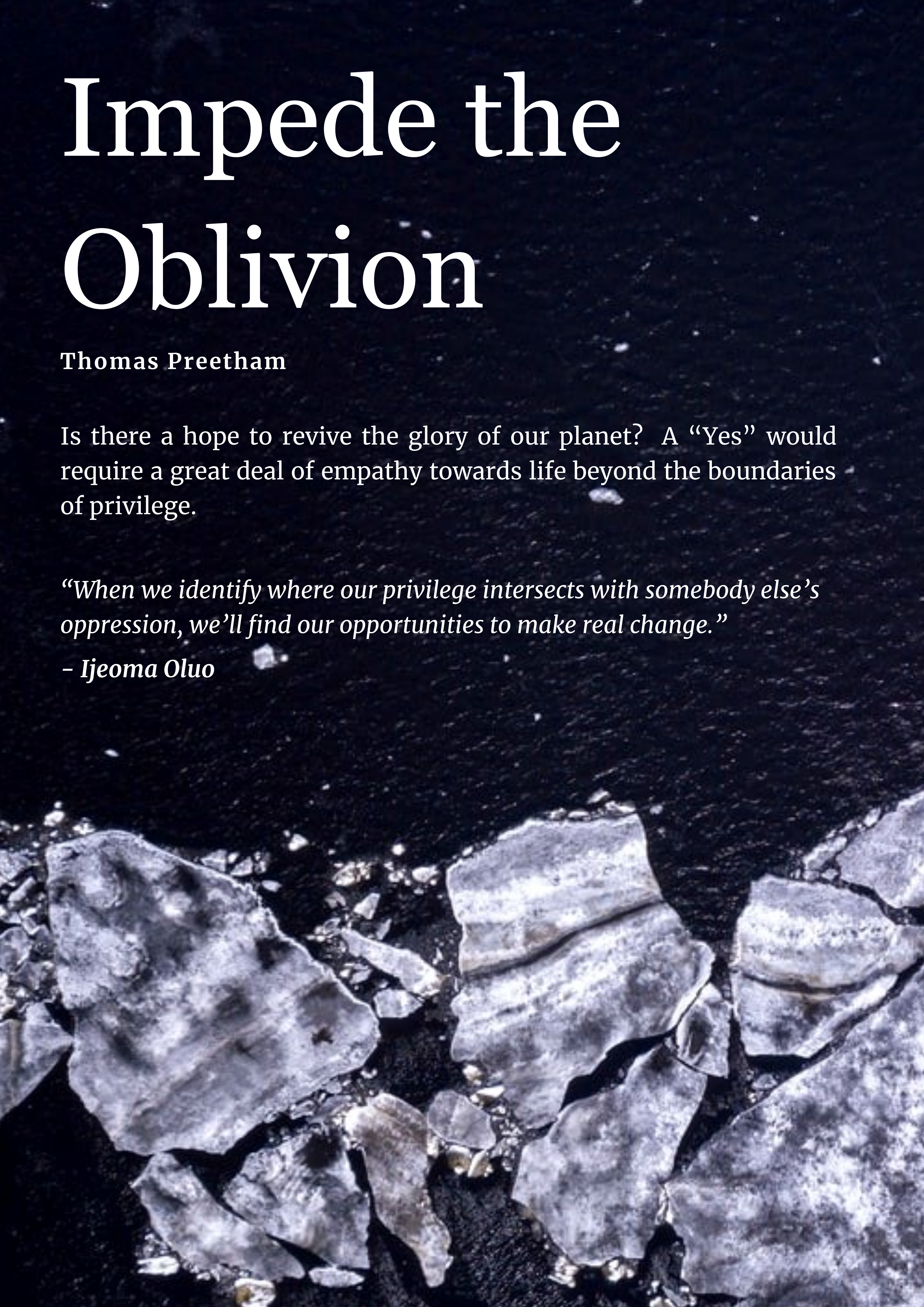
# Impede the Oblivion

**Thomas Preetham**

Is there a hope to revive the glory of our planet? A "Yes" would require a great deal of empathy towards life beyond the boundaries of privilege.

*"When we identify where our privilege intersects with somebody else's oppression, we'll find our opportunities to make real change."*

***- Ijeoma Oluo***



# Earth

is the only planet in our Solar System that is remarkably conducive to life, making it an ideal home for a diversity of life forms. Interpreting the formation of such a habitable planet, after enduring phases of magmatic hellish and a frozen snowball through billions of years, is nearly obscure. No wonder the details of abiogenesis are yet, fuzzy. Researchers believe that life could be possible where there is water, and its origin has been one of the greatest mysteries to humans. In August 2020, researchers at the *Centre de Recherches Pétrographique et Géochimiques*, *France* suggested that Enstatite Chondrite (EC) meteorites may have been responsible for water formation on Earth. EC's isotopic composition resembles that of the Earth's and contains enough Hydrogen and Oxygen that could have supplied a gargantuan amount of water. Ergo, Earth has been blessed with approximately 326 million trillion gallons of water but ironically just 1% of it is relied upon for survival which can quench our needs only for so long.

Over the centuries, owing to the significance of water for the sustenance of life on Earth, the transition from utilizing to exploiting it escalated rather quickly. We are in an era that's facing the worst water crisis across the globe and just a few decades away from reaching an inflection point where most of the world will be confronting its "Day Zero" due to lack of water.

The rate of consumption of water has been increasing exponentially over time and eventually depleting the freshwater reserves. According to a 2016 research article published by Mesfin M. Mekonnen and Arjen Y. Hoekstra, 4 billion people face water scarcity at least one month a year with half a billion facing severe water scarcity throughout the year.

Besides irresponsible exploitation, water scarcity is an unfortunate consequence of changing climatic conditions. With the rising global temperatures, the melting glacial water returns to the dynamic water cycle. The increased rate of evaporation leads to an increased amount of water vapour, a potent greenhouse gas, in the atmosphere. These gasses trap the heat of the Earth within its atmosphere, leading to further increase in temperature & this vicious cycle continues.

**How bad is Climate Change?** According to a report published on 19 April 2021 by the World Meteorological Organization, 2020 has been recorded as one of the three warmest years on record and the temperature is increasing mainly due to the rising concentration levels of the major greenhouse gases. Even if humanity stopped releasing greenhouse gases today, NASA claims that global warming will continue for at least several decades, if not millennia. Carbon dioxide is responsible for two-thirds of the present global warming induced by human activities, largely the combustion of fossil fuels. CO2 is known to be a "long-lived greenhouse gas" as it can remain for decades to centuries in the atmosphere.

Every year, over half of the carbon generated by anthropogenic activities stays in the atmosphere. It settles as black soot, called Black Carbon, on the glaciers and thus reduces their sunlight reflectivity from 70% to 20%. It leads to the ice absorbing more heat from the Sun, which increases the rate of melting of the glaciers by five times the global average. Consequently, the global mean sea level continues to rise at a higher rate through 2020, increasing the risk of quite a few cities getting submerged soon.

Terrestrial ecosystems remove the other half from the atmosphere. The oceans absorb approximately 23% of annual anthropogenic CO2 emissions and act as a climate change buffer. However, the CO2 reacts with seawater, lowering its pH and turning the marine ecosystems acidic at such a rate that the marine organism might not be able to adapt. According to an article published by the Scripps Institution Of Oceanography, “Acidification trends might begin to cause net erosion of coral reefs in this century.” and if this happens, many coastal lines will be vulnerable to deeper waters and violent storms.

It will require a book to explain how bad climate change is affecting the Earth, and hence, this is just the tip of the iceberg. It has reached a point where over the past decade, researchers found strong evidence of climate change’s relationship with the increase in unprecedented extreme events like cyclones, wildfires, hurricanes, floods, etc.

> “The current scenario insinuates the urgency of every individual becoming mindful of the impending adverse effects of climate change and act responsibly towards building a better future.”

## Generation AI

You must be wondering why I have taken a detour towards Artificial Intelligence all of a sudden but bear with me while I connect the dots. AI has been a buzzword for quite some time now. With the advent of newer technologies, AI has become an integral part of our daily lives. We often hear news about how AI has helped increase the revenues of companies, saved lives through smart gadgets and defeated the world champion of “Go” in his own game, just to name a few. What if the tool that has such high potential to completely change the dynamics of technology be leveraged to tackle climate change? According to a 2020 Forbes report, approximately 84% of global energy demand is supplied by burning fossil fuels. Industries and buildings account for 27%, Electricity and Heat Production accounts for 35%, Transportation accounts for 14% and Agriculture accounts for 24% of global greenhouse gas emissions.

**Energy:** Industrial Energy consumption has recently seen a revolutionary optimization with the support of AI. DeepMind and Google teamed up in 2016 to create an AI-powered recommendation engine to help Google's data centers save energy. They announced their next stage of this work in 2018: a safety-first AI system that would automatically regulate cooling in Google's data centers while still being supervised by the data center operators. This ground-breaking technology has uncovered various novel cooling strategies, several of which have now been included in data centers operators' rules and heuristics. Despite only being in existence for a few months, the system has already shown to save roughly 30% on average in energy, with further expected improvements.

To boost the predictability and value of wind power, DeepMind and Google applied machine learning to 700 megawatts of wind generating capacity in the central United States in 2018. They built a system that could anticipate wind power output about 36 hours prior to the actual production employing a neural network trained on historical turbine data and promptly available weather forecasts. The model then suggests making optimal hourly delivery commitments to the power system a full day ahead of time based on these estimates. This genre of machine learning approach can encourage more adoption of carbon-free energy on electric grids and enhance the business case for wind power globally.

**Water:** There are plenty of techniques that help with Water Recycling and Waste Water Management but considering the environmental conditions a mere repair will not suffice. We need more efficient ways to leverage the most abundant resource on our planet, not only cater to our needs but also bring back the glory of Mother Earth. Fortunately, the technology of Desalinating Ocean Water using Reverse Osmosis has been a ray of light and has more than doubled over the last decade but the amount of treated water made in a year still adds up to less than 1% of the water we use. Currently, there are over 20,000 Desalination Plants across the world that produce 25,000 million gallons of water per day.

However, the efficiency of ultrafiltration depends on various external factors like weather, temperature, change in the quality of water collected, etc. To overcome these uncertainties, the process of ultrafiltration can be optimized using Reinforcement Learning which is already being implemented since October 2020 by the researchers of the University of Alberta’s AI4Society in Drayton Valley. The reinforcement learning algorithm collects data from the environment, takes decisions, learns from it and further improves the filtration process resulting in an efficient usage of energy, chemicals and manpower. With the help of AI, desalinating ocean water could not only quench the thirst of many but also replenish the aquifers if done at a large scale by every country.

AI is also being leveraged to cultivate low-carbon materials, improve transportation systems, advanced battery technologies, and a slew of other cutting-edge technologies that promise to make considerable progress toward a more sustainable economy. It can aid researchers in examining spatial data to detect deforestation, tropical cyclones, weather fronts, and atmospheric rivers with an accuracy of 89 to 99 percent, the latter of which can bring heavy precipitation and is frequently difficult for humans to spot on their own. Building better models help people stay safe by improving weather forecasts. It can also help optimize the orientation of solar panels to get the maximum out of solar energy, only to name a few.

## Isn't it too nice to be true?

“It is predicted that by 2025, 20% of the global electricity could be consumed by the devices that store data.”

Implementing AI solutions requires high energy and lots of data. Machine Learning produces significantly more carbon emissions than we can imagine. AI systems, from speech recognition to self-driving automobiles, require a lot of energy and produce high amounts of carbon emissions. Research done at Stanford suggests that an off-the-shelf language-processing AI system produces approximately 635 kgs of emissions, which is almost the same as flying a person on a roundtrip between San Francisco and New York. Depending on the source of power, the whole suite of experiments required to create and train an AI language system from square one can yield up to 35,380 kgs. That's more than twice what an average American exhales in a lifetime. Machine Learning systems will almost certainly require more energy during

production than during training. So, what is the use of the state-of-the-art AI solutions to optimize energy consumption by powerhouses, data centers and desalination centers if it is compensating for the carbon emissions, if not worse, that it is designed to reduce in the first place?

The key is to keep a check on the emissions and choose the right location for the data centers. A team at Stanford built a tool to measure both the energy consumed and the amount of carbon emission by a machine learning project. Consequently, it leads to a movement known as Green AI where the developer is compelled to identify ways to make machine learning cleaner and greener. Research has been going on to explore the plausible ways to make machine learning algorithms more energy efficient whilst preserving their performance. Furthermore, moving the training and production activities to a site that is mostly powered by renewable energy sources can significantly reduce carbon emissions.

## The Verdict

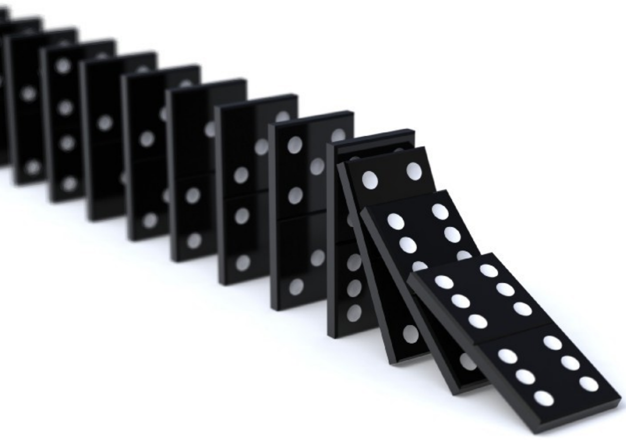
AI irrefutably exhibits prodigious potential to combat climate change; however, there are still some intricacies of the energy sources that need to be taken care of for it to serve the purpose. The scientific studies done by the prestigious British Medical Journal – Lancet suggests that nuclear power is the safest and cleanest source of energy and requires the minimum amount of raw materials. Top countries, according to the Environmental Performance Index of 2020 reveal a paradigm of relying more on the combination of nuclear and renewable energy over fossil fuels. The faster we migrate to a harmonious mix of all the cleaner energy sources available, the greener our technologies develop and the better control we can gain over climate change. All the aforementioned technological advancements will aid us only as long as we don’t repeat the mistakes committed in the past.

The repercussions of climate change are probably the most exorbitant price we might have to pay in the near future unless we put conscious efforts on a systemic and an individual level to mitigate them. It’s high time we take stern measures before it’s too late for any contingency to rescue the living beings from the wrath. We need to educate ourselves on its gravity and let go of selfish motives for the greater good of life on Earth. Every time we decide not to act against nature, we add a few pages for our future. We might not have fallen behind in time to change our lifestyle that is detrimental to this planet. Climate change is palpable and is drastically affecting the Earth. Although we are the ones to blame, we can make sure to be the solution too.

# On the way to Disaster Resilience

## Role of AI in Disaster Mitigation

**Abhinav Ranjan**
**Dingari Sreeram**



On 2nd July 2021, a town in British Columbia was burnt down due to a fast-moving wildfire following an intense heatwave. One of the survivors said, "I ran towards the town, and the fire was pretty much following me, and in about 15 minutes, the whole town was gone.". It is not the first time that disasters of this scale have occurred and wreaked havoc. Globally, on average natural disasters are the cause of more than 60,000 deaths every year. The Haiti earthquake in 2010 has killed more than 3,16,000 people and affected the lives of more than 3 million people. Being the poorest country in the Latin America and Caribbean region, Haiti did not have the capabilities to handle a disaster of this scale, and the country's GDP dropped by 8%.

Natural disasters result in catastrophic damage and substantial economic losses, especially for low to middle-income countries that are not equipped with the proper infrastructure to deal with them, and their recovery may even take decades. With increasing climate changes, the number of disasters and their damage is showing an increasing trend. In 2020, natural disasters threatened more than 160 million people, and global economic loss stood at US$268 billion.

It is the responsibility of disaster managers to develop new methodologies and utilize technology to mitigate these disasters. With the advent of Artificial Intelligence, we now can analyse voluminous data to extract valuable insights that can help in rapid response during all four phases of disaster management (Mitigation, Preparedness, Response, and Recovery). This article will discuss applications of AI models and their role in effectively supporting decision makers in the disaster mitigation phase.

Disaster mitigation includes identifying the hazards, predicting the impact, assessing the vulnerable areas, and developing strategies to make the community more disaster resilient. For example, once a cyclone or hurricane is identified, Hazard Zone maps can be created using various data sources such as terrain conditions, weather, previous disaster data, and human activities in that region. These maps help in evacuation and relief operations and identify potentially hazardous sites such as chemical industries or waste storage locations. Identifying them is crucial to isolate or contain them to avoid the release of toxic chemicals, which could have otherwise hampered the rescue efforts.

Though traditional methods help identify hazards, as we have observed, they are prone to false alarms and are labour intensive and complex. With the help of AI models and their application in zone maps, we can identify high-risk regions. For instance, many researchers have used old disaster data (or simulated data) and applied logistic regression, support vector machine, and neural networks in problems such as snow avalanche prediction, landslide susceptibility, forest fire susceptibility, and other disasters with excellent performance. Let us look at few such research works.

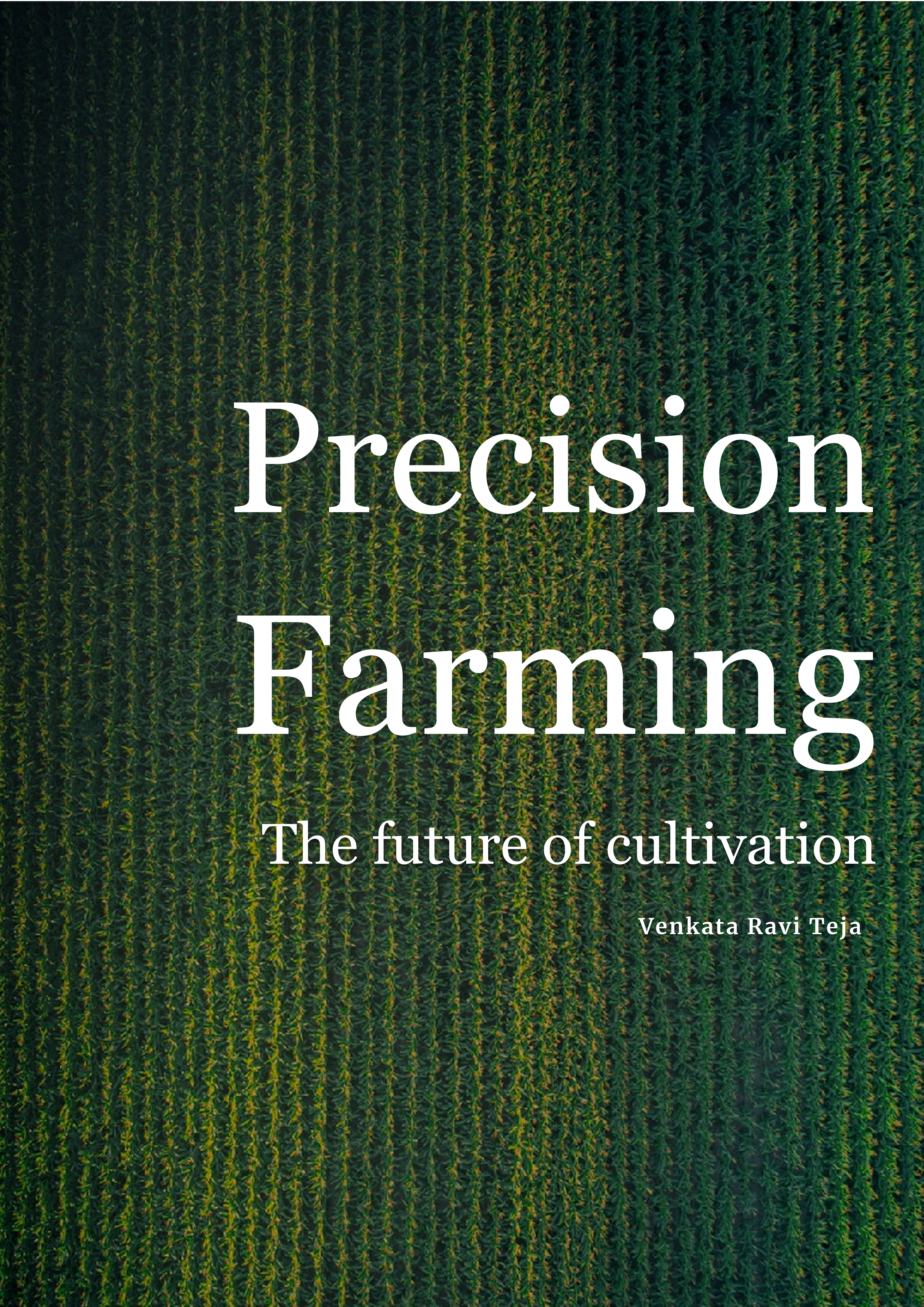
In collaboration with Tohoku University and The University of Tokyo, Fujitsu used the power of the Fugaku supercomputer to generate 20,000 possible tsunami scenarios based on high-resolution simulations and used them as training data for building an AI model to predict flooding before landfall at high resolution. On a similar track, Google has also developed an AI platform that used the data collected from rainfall records and flood simulations to predict floods and warn users via Google Maps and Google search.

Scientists at Google and Harvard have built a neural network to predict the aftershocks of an earthquake by studying more than 131,000 earthquakes and aftershocks. On testing this model on 30k events, the neural network predicted the aftershock more precisely than the traditional models. Feyera Hirpa, Data Scientist of One Concern Inc., tested his prediction model during the 2019 Chikuma River flood in Japan caused by typhoon Hagibis. The flooding model was comparable to the actual flood & validated the power of prediction AI Models.

Satellite Image Analysis can play a critical role in mitigating the impact of the disaster. In Australia, Researchers at Monash Data Futures Institute have used the Sentinel-2 Satellite image data to analyse 4300+ high-resolution images to generate a vegetation map of Victoria state. They built a model using time series classification that could annotate moisture and atmospheric temperature data to understand the vegetation better and its purpose of usage. These hazard maps can also play a crucial role in bushfire prevention, agricultural planning, pollution management, and rehabilitation efforts.

Few other AI research areas during the disaster mitigation phase include identifying community influencers to calm the people during the event, estimating the needs of people, applying optimization algorithms for best plans, and comparing different mitigation strategies.

To summarize, Proper disaster mitigation plans made with the collaboration of physics and AI models help achieve a high level of preparedness and disaster resilience. These insights enable authorities to proactively plan related to evacuation in high -risk identified zones, rehabilitation centres selection, transportation routes, warehouses (to store the aid resources), and distribution routes. These actions help to save many lives, reduce property damage, and implement successful recovery operations.



# Precision Farming

## The future of cultivation

Venkata Ravi Teja

According to Food and Agriculture Organization (FAO), the arable land per person has reduced from 0.38 hectares to 0.23 hectares from 1970 to 2000. It is expected to decline to 0.15 hectares by the year 2050. Excess use of chemical fertilizers and pesticides has a detrimental effect on the environment, reducing soil fertility and polluting the air and water. Extreme weather conditions like cyclones and droughts are increasing in the past years due to climate change. Water scarcity is posing a great challenge, and it is going to worsen in the coming years. United Nations have estimated that by 2050, the farmers have to ramp up the food production by 70 percent. The present challenge is to produce more food with fewer resources & extreme conditions.

$$\mathrm{GCP} = \mathrm{PCP} \times \mathrm{N} \qquad \mathrm{PCP} = \eta \times \mathrm{ANR} \times \mathrm{SAR}$$

For increasing the Gross crop production (GCP), we have to either increase number of crop cycles (N) or Per crop production (PCP). We can't increase ‘N’ in a year beyond a certain number as the vital nutrients in the soil are limited and takes time to rejuvenate through the decomposition of organic matter. ‘PCP’ is constrained by the available natural resources (ANR) in the field, and neither can we increase the amount of Supplied Artificial resources (SAR) like chemical fertilizers as it severely damages soil fertility in the future. So, all we can do is maximizing 'η' the farming process efficiency.

## Concept of Precision farming

The farming process efficiency can be maximized through precision farming. Precision farming manages nonuniform farm fields following correct practices in the right place and time. It increases profitability and sustainability by optimizing available natural resources and protecting the environment.

Precision farming through information enables farmers to monitor crop and soil conditions in a heterogeneous farm field and provide crops with the required resources specific to the site. It also helps farmers in planning the activities from the time of seed plantation to harvesting. This approach requires access to real-time data about the crop, soil, and other pertinent information.

In Precision farming, Satellites, aerial vehicles, and sensors mounted on ground vehicles or handheld collect the required data from the entire field. GPS (Global Positioning System) receivers integrated with field data gathering equipment collect the location and time data. The GPS enables the farmer to navigate to a specific location accurately in the farmland for soil sampling and inspect the stressed crops.

A geographic information system (GIS) represents the field data collected by remote sensing methods through maps that are created using GPS information. GIS is a powerful visualization tool. It creates multiple maps for yield, crop and soil health, weed density, plantation planning, etc., with their respective geographic coordinates.

## Advantages of remote sensing over ground-based methods

Precision Farming includes ground-based methods using sensors planted at various regions in the field and remote sensing methods through which farm fields will be monitored in a no-contact mode. The limitation of ground-based methods is that they will give point-wise information specific to where the sensor is placed. Point-wise sensors are not economical in the case of large and heterogeneous fields. Remote sensing, on the other hand, provides continuous information across wide regions. The satellite data can also be obtained at affordable costs.

## Remote Sensing in precision farming

Remote sensing methods are used in crop and soil monitoring. It detects the abnormalities in the crop much before the human eyes can recognize it. The early warnings on crop and soil health help the farmers in avoiding the loss in crop yield. Optical remote sensing with spectral imaging is the most commonly used technique. It collects the information from the earth's surface through visible, near-infrared, and short-wave infrared sensors. It depends on the reflected light from the target object for identifying them through their spectral signatures, which are unique to that material. The Spectral signature is the ratio of reflected to incident radiant energy of the material. Regular RGB cameras can capture light across the three visible wavelength bands of red, green, and blue, whereas spectral imaging captures the light in the visible spectrum and beyond it. The bands beyond visible are represented as False color composites of Red, Blue, and Green.

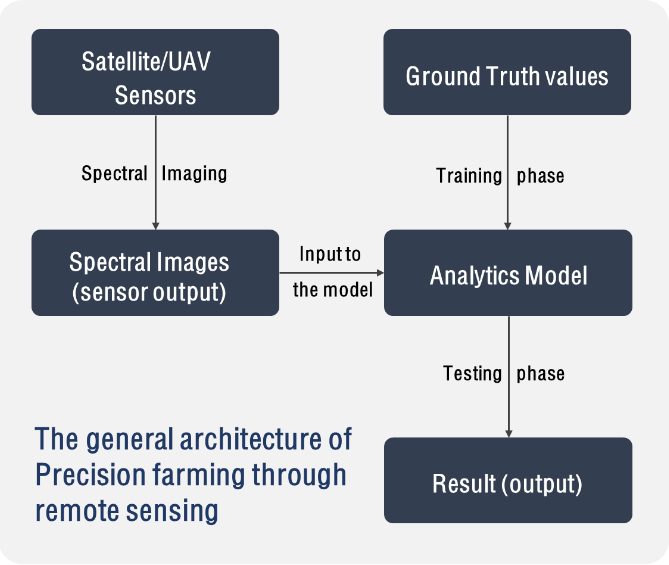
Spectral imaging is broadly divided into Multispectral and Hyperspectral Imaging. Multispectral imaging captures the spectral information from the specified wavelength range of the electromagnetic spectrum in discrete spectral bands. Hyper Spectral imaging collects spectral data in the form of several continuous narrow wavelength bands. The choice between these two techniques depends on the task.

## Concept of Band Ratios

The reflectance values are obtained through spectral imaging for various wavelengths over all the pixels. The limitation with Reflectance values is they change with the angle of sun rays, time of the day, and year. Due to this variation, we can't rely on pure reflectance values for our analysis. Band ratioing is used as a solution to the variability of reflectance with local factors. Different materials exhibit different levels of reflectance at various bands. Band ratios enhance these spectral differences between the bands and curtail the influence of topography and other factors on reflectance. Let’s say there are two types of plants, among which Type1 has high reflectance compared to Type2 at particular wavelength bands A and B. In this case, we can't distinguish between the two types of plants based on absolute reflectance values, As the reflectance of Type 2 plants on the sunlit side will be more than that of Type1 in the shadow. Here we will instead take the ratio of bands A and B to distinguish between the two varieties of plants.

## Spectral imaging methods vs. deep learning methods

All the spectral imaging use cases in precision farming can also be done using deep learning. But, the deep learning model on regular RGB images has disadvantages in terms of its requirement of high processing power, time, and high amount of data to fit the model. A neural network is a black-box model, thus limiting its explainability. Through the band ratio techniques, we can build explainable models which are computationally less intensive, making them economically viable for farmers.



The general architecture of Precision farming through remote sensing

# Soil Monitoring

## Soil organic carbon determination

Soil organic carbon (SOC) is a key indicator in assessing soil health. SOC helps in nutrient and moisture retention and stabilizes the soil structure. SOC controls the C:N(carbon to nitrogen) ratio. An optimal C:N ratio enables the microbes to release the nutrients into the soil by decomposing organic matter. SOC levels are dynamic and require constant monitoring. Spectral imaging can be used for soil mapping over varying levels of SOC. Soils with high SOC content are darker and have low reflectance in the visible range. So, Regression models can be developed using the spectral band ratios as independent variables and actual measurement of SOC from Soil samples as dependent variables in the training phase. This developed model is used on new data to obtain SOC.

## Crop Health Monitoring using Vegetation indices

Band ratioing is a simple ratio that indicates crop and soil health, but it has a couple of problems. The first is the issue with division by zero when red reflectance becomes zero. The second being its wide range of values, which makes it difficult to compare. Hence Spectral indices are derived from band ratios to address these problems. These indices are combinations of reflectance properties from two or more wavelength bands. Spectral indices used for assessing crop health are called Vegetation indices.

Normalized Difference Vegetation Index-NDVI also referred to as the measure of the greenness of the area, is the most commonly used among all vegetation indices. A healthy plant reflects more in the near-infrared band and absorbs more in the red band, whereas stressed vegetation reflects red light and absorbs near-infrared light.

$$\text{NDVI} = \frac{(\text{NIR—Red})}{(\text{NIR + Red})}$$

**NIR** - Reflectance in the Near-Infrared

**Red** – Reflectance in the range of red

| NDVI value | Indication |
|---|---|
| -1 to 0 | Inanimate object |
| 0 to 0.33 | Stressed or unhealthy vegetation and appears in orangish-red |
| 0.33 to 0.66 | Moderate health and appears in tint green |
| 0.66 to 1 | Healthy crop and appears green |

The crops are classified into one of the above classes depending on the NDVI value.

The above thresholds could have been obtained using the general architecture where the analytics model used is a Decision tree on a single variable. But, agronomists defined these thresholds empirically .

## Crop calendar derivation using NDVI time series

The temporal profile of NDVI gives information on various growth stages of the crop and the overall performance of the crop in a particular season or period. Different physiological stages of the crop in its growth cycle would sequentially occur, maintaining the exact chronology over time. The smoothed NDVI time series would provide information on the dates of green-up (beginning of cycle) & harvesting, maximum NDVI date, & length of crop growth.

Crop calendars can be developed based on this, which provide information on the optimal time window for seeding date, duration of the crop Life cycle, and optimum date for harvesting. This would help the farmer plan to purchase inputs, finances, and labor requirements.

## Crop yield estimation

Accurate crop yield estimation is of much interest to policymakers for the country's food security. They plan the food imports and exports depending on the yield estimates. Monteith method is a popular model which uses biomass for crop yield estimation. The net primary productivity (NPP) is the biomass accumulation in the plant.

$$\text{NPP} = \text{fAPAR} \times \text{PAR} \times \text{RUE} \times \text{Wstress} \times \text{Tstress}$$

fAPAR, which is the fraction of absorbed photosynthetically active radiation (PAR), can be estimated by developing a regression model using NDVI values as the independent variable. The model should be trained with existing ground truth values, and the trained model can be used to estimate fAPAR values for different NDVI values

$$\text{fAPAR} = \text{A} \times \text{NDVI} + \text{B}$$

**A** and **B** are regression coefficients

**HI** – Harvest Index

$$\text{Estimated Grain Yield} = \sum_{\text{Sowing}}^{\text{Harvest}} (\text{NPP} \times \text{HI})$$

RUE (Radiation use efficiency) is relatively constant for a specific crop when calculated over a complete growth cycle. Water stress can be calculated through Land Surface Water Index (LSWI). Temperature stress can be computed through weather data. It depends mainly on the difference of mean daily temperature from the minimum, maximum, and optimum temperature of photosynthesis.

$$\text{Wstress} = \frac{(1\text{-LSWI})}{(1\text{-LSWImax})} \qquad \text{LSWI} = \frac{\text{NIR—SWIR}}{\text{NIR + SWIR}}$$

**SWIR** - Reflectance in the range of Short Wave Near Infrared.

## Drought Monitoring in the field

Drought in agriculture is referred to as a soil moisture deficiency. It badly affects the production and growth rate of the plants. Meteorological drought indicators were developed in the past for drought monitoring. A new methodology using spectral indices was developed, addressing the issues of the past methods. Standardized Vegetation Index (SVI), Standardized Water Index (SWI), and Evaporative Stress Index (ESI) are the remote sensing-based indices used in drought monitoring.

SVI, SWI, and ESI are calculated from the anomaly or respective Z scores for their time series, i.e., the deviation of their values in the time series from their individual mean values per standard deviation of their values.

$$\text{Z score} = \frac{\text{Value - Value mean}}{\text{Value std}} \qquad \text{f} = \frac{\text{ET}}{\text{PET}}$$

Individual values in the above time series, for Vegetation Index (VI), is usually NDVI or EVI (Enhanced Vegetation Index), and for Water index, it is NDWI, which is equivalent to LSWI.

Evapotranspiration is the combined evaporation and transpiration to the atmosphere from the surface. ESI can be calculated from the z score of the ratio of evapotranspiration (ET) to the Potential evapotranspiration (PET).

For all three indices, a value less than -1 represents a dry condition. Conversely, high negative values of these indices indicate a severe drought. Recent studies show that a synthetic index combining these three indices includes the information from vegetation growth from SVI, vegetation water content from SWI, and evapotranspiration from EVI, giving more accurate predictions.

SVDI = w1 x SVI + w2 x SWI + w3 x ESI

**SVDI** - Synthetic Vegetation Drought Index

**w1**, **w2,** and **w3** are weights of three indices.

## Thermal Remote Sensing

Another Precision farming technique that gained popularity off late is Thermal imaging which utilizes the long-wave infrared portion of the spectrum. Thermal remote sensing has its advantage when it comes to its ability to operate at night. Similar to optical remote sensing, it can also monitor crops and soil. It captures the radiant (emitted) temperature from the crops, which can be used to detect crop stress and plan irrigation cycles. Thermal Images are either displayed as greyscale images or color composites. In greyscale images, brighter areas indicate warmer regions and darker areas indicate cooler regions. In fact, Thermal and optical remote sensing techniques are not competing; instead, they complement each other.

## Conclusion

Precision farming is the perfect solution to avoid the upcoming food crisis by increasing the efficiency(ŋ) of farming practices. But, the bottleneck for its implementation is the lack of expertise and technical knowledge in processing the available raw data for real-world applications. This can be addressed when the number of educated farmers increases. Also, the big companies which are competent to implement precision farming techniques need to support the farmers and make this practice widespread. Today we are witnessing more and more companies and startups providing state-of-the-art farming technologies and solutions. It is quite evident that precision farming is the next big thing, especially when a big player like Microsoft are aggressively investing in Precision farming.

# Sports Analytics

**Hitesh Kashyap**

### Box Score

Henry Chadwick, a sportswriter, developed box score metric which presented the baseball player's performance in a tabular form. It helped the statisticians measure players' and team's performance quantitatively.

1858

### Basketball Abstracts

Bill James' Baseball Abstracts, collection of annual baseball data won public's attention. Later, he coined a term called "Sabermetrics" to define the science behind a baseball game.

1977

1950s

### Multiple Attempts

Till the middle of the 20th century, many others made unsuccessful attempts to show some real usage of analytics in sports.

"Dhoni finishes off in style. A magnificent strike into the crowd. India lifts the world cup after 28 years". We still fondly remember the words of Ravi Shastri. Observe that the detail '28 years' emphasizes the rarity of this event and the effort invested to attain such a victory.

Numbers have a special relationship with sports. We still get glued to those startling match statistics and graphics that appear on the screen while our favourite match of cricket or football is on. Don't they add a different dimension to our analysis of the game? What is the underlying technology behind all of this? Well, it is all analytics.

Let us explore how analytics started changing the way we enjoy sports and why it is becoming hugely popular.

We know that human beings have limited capacity to process and generate immediate insights from massive data present in raw form. Processing and presenting this data in tabular or graphical constructs help us observe trends and find valuable insights empowering our decision-making abilities. Advancements in technology and computational capabilities have simplified this process of data analysis. For example, in sports, parameters like weather conditions, recent win/loss

2003

### Book

This book written by Michael Lewis focused how Oakland Athletics Manager built a competitive baseball team with minimum budget to clinch American League West title.

2011

### Movie

Bennett Miller directed a movie, Moneyball, starring Brad Pitt, inspired by Moneyball book. The film was a big hit on the box office and served as an eye-opener to many sports analysts.

2021

### Wide Acceptance

Though some of the organizations started using data analytics in sports as early as 1960s, it is now being adopted by many companies.

statistics, and players' performance are used to make predictive machine learning algorithms that aid managers in making game-winning decisions. Predicting future scenarios also involves the use of much sophisticated deep learning and cognitive algorithms.

Each team's win speaks volumes and brings added perks like increasing fan base, attracting more sponsors, retaining top players, increasing merchandise sales, and getting concessions in high-quality sports equipment. It also increases the team's confidence and local pride. These are some of the reasons why analytics is being accepted widely by many companies, and the sports analytics market is growing at a rapid rate.

"The frontier of analytics is just beginning, and there is no end in sight to the potential" - Dr. Lynn Lashbrook, Sports Management Worldwide President, and Founder.

## Sports Analytics Market

A study published by Grand View Research Inc states that the global sports analytics market size will expand at a CAGR of 31.2% and reach $4.6 billion by 2025. Analytics has also helped in the proliferation of the sports gambling industry. The gambling industry is valued at around $800-$1,000 billion, out of which sports gambling contributes about 13% of its share. It helps gamblers to analyze massive amounts of data and information to place the right bet.

Many teams and clubs have collaborated with big companies to develop analytical products to help managers in their decision-making process. Real Madrid, one of the most renowned and revered football clubs, utilizes Microsoft Analytical tools to manage its operational activities and players' performance. This tool also helps to maintain clubs' relationships with more than 550 million global fans. Also, Manchester United trusts Aon for planning their game strategy to stay one step ahead in the competition. Some of the works of sports analytics have been so accurate that they have been written in history books.

## Bull's eye

There have been many instances in sports where analytics has performed outstandingly. One such incident is when Daryl Morey, General Manager of the Houston Rockets, an American basketball team, found three-pointer shot attempts from corners had a higher chance of success than trying two-pointers shots. The result was that the Rockets broke the record for most 3-point attempts during NBA 2018-19 season. Similarly, ScoreWithData of IBM predicted seven hours before the first quarter -final of the World Cup that Imran Tahir, South African spinner, would become the power bowler. This prediction came out to be accurate, and Tahir won the match for South Africa against Sri Lanka. Today, many professional sports like cricket, basketball, football, hockey, etc., use analytics to maximize their team performance and improve their chances of winning. These sports use different kinds of metrics to measure players' and team performance. Although many sports are utilizing the potential of analytics, some are still critical of adopting this technology.

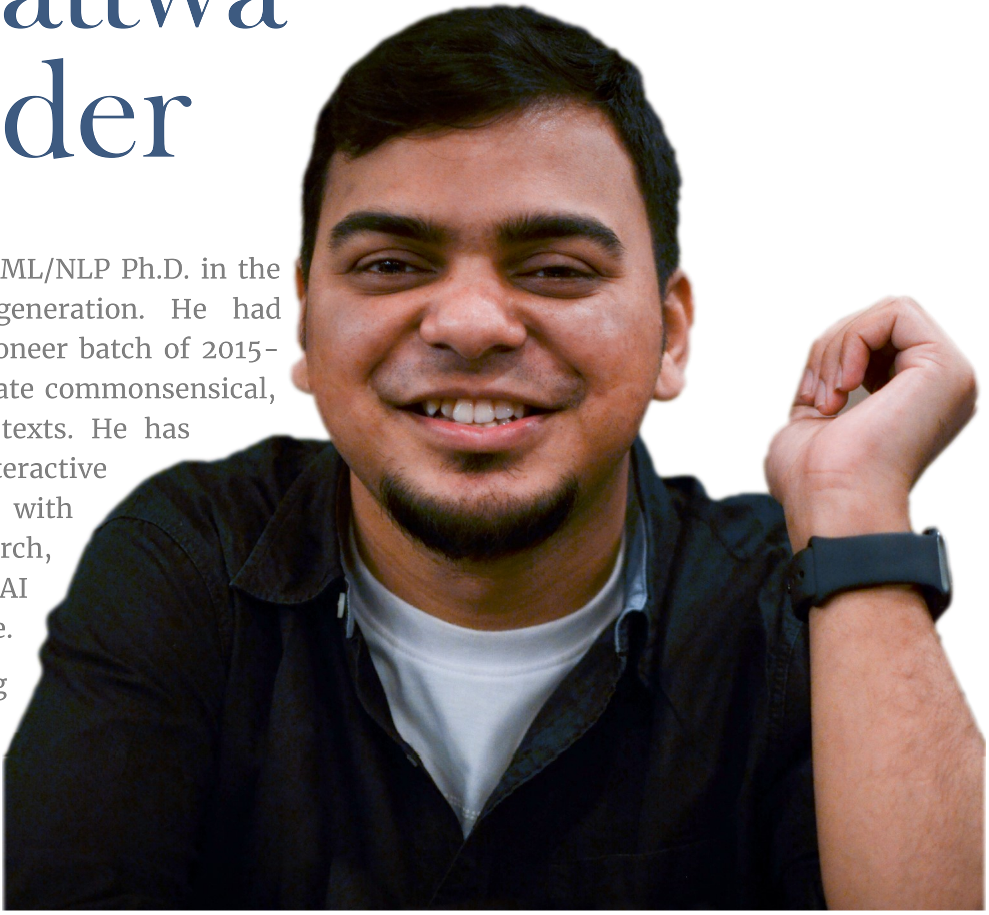
## Challenges for Sports Analytics

Despite growing rapidly, Sports Analytics still faces many challenges. Critics point out that there are certain factors that analytics is not capable of capturing, like player diving in the game, intentionally misleading the opponent, or riling up opponents by yelling/ sledging. They argue that such things can only be captured and processed by humans. However, to a certain extent, analytics can still handle such kinds of unstructured data. Such things are documented using text analytics models, and this unstructured information is converted into standard structured data with rows and columns for processing. Rule-based categorization or machine learning-based models are used to gauge the frequency of words and generate insights. The efficiency of these models can be improved by collecting data from various sources. For example, using scouting reports from different scouts reduces the bias towards any single opinion. Increased research and developments in artificial intelligence would surely help in addressing many of these challenges.

## Next in Sports Analytics

Sports analytics has led to a breakthrough revolution in the sports industry, but it still has a long way to go. The day is not far, with the integration of technology and wearables, when analytics would assess the mental and emotional makeup of the player and how it correlates to the player's on-field performance. With recent advancements in technology, sports analytics will evolve manifold in the years to come, and we will experience some of the never-explored dimensions of sports.

# Bodhisattwa Majumder



Bodhisattwa Majumder is an ML/NLP Ph.D. in the field of natural language generation. He had pursued PGDBA during the pioneer batch of 2015-17. He is researching to generate commonsensical, personalized, and subjective texts. He has developed user-facing interactive systems in collaboration with Google AI, Microsoft Research, Facebook AI Research, Oxford AI team, and Alan Turing Institute.

He co-authored a best-selling O'Reilly book on NLP 'Practical Natural Language Processing' that is being adopted in universities worldwide.

**AINA:** Let's start with the famous question on this dilemma of students pursuing analytics, whether to choose the business side with business analytics or the technical side with pure data science as their career paths. How did you decide to choose a technical path after your master's?

**Bodhisattwa:** It is important not to draw a distinctive line between the business and the technical paths. From an overall perspective, there are specific inputs, and then there is an output. What you eventually do is analyse the characteristics of the system. You can do that from a technical standpoint, understanding how changing the core modules or implementations of the system changes the output of the system. Another thing is looking at it from a business perspective. If you have more flair towards coming up with business ideas, it shows you have good domain knowledge and understand the essential input variables that drive the business. From an overall perspective, I do not see these two as different; it depends on what you are more comfortable with or on what you have more knowledge.

**AINA:** You have been working in natural language generation. What interested you to choose this particular field?

**Bodhisattwa:** I was always interested in social science. And one of the main components of social science is its participants - humans. What makes us different from other living beings is our ability to communicate through language. To understand the language components, you analyse the language or look at the origin of language, how the language emerges. That could have led me to both natural language understanding or natural language generation. Generation is more of a newer task, and what motivated me to choose generation is the ability of the model to come up with the language. Humans communicate with a language, but it's through an interface when you're interacting with the machine. I thought that language generation is an important problem as a missing piece in these accessible interactive systems.

**AINA:** Noam Chomsky on GPT3 commented that language models are not related to language and cognition as they do not understand the language even if they could produce the language. What is your take on it?

**Bodhisattwa:** All of the Machine learning models are predominantly supervised models learned on data. There are certain patterns and a mapping between the input and the output that the machine should learn. It discovers those patterns from the training data. In Deep Learning models, often this mapping is not intuitive to the user, but the model somehow finds this mapping greedily. So, in a way, that statement is correct that it does not actually understand the language. If you break down all languages, there is nothing but an underlying pattern. That is the current state-of-the-art of the language models. They can come up with plausible sentences understanding these patterns, but it does not mean that the model understands the language. Can these models perceive the world the way humans do? There is a famous lecture by Richard Feynman named 'Can machines think' where he says that it is fundamentally not possible for machines to think like humans. There is a fundamental difference in how a task is accomplished by machines, even if they produce the same output.

> “There is a fundamental difference in how a task is accomplished by machines even if they produce the same output. ”

**AINA:** From a researcher's point of view, what is the thought process to select a particular topic to publish on?

**Bodhisattwa:** I think it's important that you choose a domain, like when I started my Ph.D. in 2018, I always wanted to understand what common sense is. Then you look at different progresses in that domain. It is possible that there are certain questions to answer, but there are not enough tools, so you start from scratch. Or some people have already developed the tools, and you can apply these tools to a number of domains that people haven't tried yet. It is a more applied way of doing research. These were the two guiding thoughts that drew me to select common sense as a topic, and in 2018 there were not enough tools, not enough knowledge pieces available for common sense. I did not want to start from scratch, which has its own risks and advantages. After 2019, this new model started coming in from the University of Washington, and I quickly picked it up and applied it to my research. That constituted my recent series of work on how you can efficiently use common sense in natural language generation.

**AINA:** You have done your Ph.D., and you have worked with tech giants like Google and Facebook, so you have both academic exposure and industrial exposure. What proportion of industrial research is going on as compared to academics, and how are they collaborating?

**Bodhisattwa:** I think research has two different components. Firstly, it is coming up with Ideas, building theoretical foundations, and coming up with a prototype that works on a small amount of data. Secondly, making it work on a large scale. What is seen is that often academia is focused on the first part. What industry is very useful in, especially in this machine learning, is that they have the resources like compute power, scientists, etc. They can scale up this solution with huge amounts of data and use them in their product offering process. Now the assumption is, the model that works on small-scale data will get scaled similarly to largescale data. But often, we see that when you try to scale up an algorithm, it has weird peculiarities with this large amount of data. Companies are contributing a lot to fixing these peculiarities. This is very difficult to do in an academic environment due to resource constraints. The industry then asks academia whether academia can use these models to address the unanswered things. If they can, the industries would deploy them or try to see what they can do next.

**AINA:** Can you give a walk-through of how you authored the O'Reilly book on NLP?

**Bodhisattwa:** When I was in Bangalore after PGDBA, I was dabbling through the intricacies of NLP as my master’s research. I, along with my co-authors, started with the idea of writing a book on NLP that specifically targeted industry people because we have seen in the industry either you have large data or only limited data. There is also a budget constraint. You can not collect data randomly, but you have to do it intelligently. You convert a business problem to a set of technical problems and try to understand the system. What certain steps do you follow when you have data and when you don't have data? You may start with data collection, then how do you do that? What are the right questions to ask? These were the questions we four authors had in our minds. We thought that there is no repository that provides this amount of knowledge to the people who work in the industry, and often people make mistakes that are very costly in terms of time and budget. This is what motivated us to write the book.

**AINA:** You have published many papers and even your book with different co-authors. How do you develop this networking?

**Bodhisattwa:** I would like to thank my mentor Harshit Surana, who is also a co-author of my book. He motivated me to come up with ideas on how to collaborate with people. Often in academia, you see people are open to collaboration if you have a good idea. When I'm reaching out to a professor or someone of that stature, I should not ask him to help with trivial things like fixing up the code. It's probably not worth his time. Do some base work, recognize the missing pieces, find out experts who could help you with those missing pieces, and try to reach out to them to collaborate or for advice. This has always been my approach when I have been looking for a collaboration. Thankfully it worked very well for me, and I'm proud of my collaborators. With many papers being published regularly in today's machine learning research, it's almost impossible to catch up to that speed while being a single author. If you open up to collaboration, you will develop your ideas and combine ideas from different perspectives you're probably unaware of. It always saves time.

> “If you have machines as human assistants, with the ability of interacting and critiquing, that is the future of interactive systems.”

**AINA:** As you have worked on user interactive systems and have sufficient exposure to them, how would you envisage their future?

**Bodhisattwa:** The current trend in the recommendations field is interactive recommendation or conversational recommendation. How would you provide feedback to the machine if you don't like a recommendation? For example, the machine is trying to provide a recommendation, and you, too, are helping the machine provide useful information so that you can arrive at the perfect movie. It could also be adversarial where you argue with the machine, and then you settle on a final decision. The interactivity or ability to accept critique is what makes machines human-like. These are the fundamental qualities of human communication, especially in a societal setting. If you have machines as human assistants, i.e., with this ability of interactions and critiquing and getting back, that is the future!

**AINA:** What are some things that helped you in your journey as a researcher?

**Bodhisattwa:** My Statistics background, which I learned at ISI Kolkata, helped me, which I use on a daily basis. It is important to understand the system characteristics. And second is the intricacies of the domain. For example, if you want to learn about NLP, a certain linguistic background can be important. You don't have to go much deeper, but understanding how grammar works. What have people done previously? Then it is your research problem and literature. Once you are in Ph.D., you have to read many papers because you need to understand what people are doing currently in that particular problem. It is difficult to come up with a concept that no one has thought of, especially in machine learning. You have to understand the current state-of-the-art, and for that, you need to read good papers. Read papers from top conferences. These things can help you in your journey as a researcher.

**AINA:** How would you guide a person who wants to do research in machine learning and natural language processing? Does one need to know the end goal beforehand?

**Bodhisattwa:** It depends on your interests. Do you want to write papers or to research and understand the characteristics of the system? If you want to do research in the industry, then you don't have the pressure to write a paper. In that case, you should do more experiments like an empirical evaluation by changing input and output and observing the outcomes. For publishing papers, there are strategies like coming up with a problem that has not yet been solved intelligently or could be done better with the current tools. The second would be that you build a new model for a known task that beats all state-of-the-art models across different datasets. You have to build an analytical story of why the current model is better than previous models. It is not always possible to know our end goal.

Let's say you don't know what Problem A is but see if you can solve problem B, which is a subproblem of A, and you may know the end goal of problem B. If you don't, break it again until you get a specific problem that you can solve. Now you are actually bringing it to the state where you know the output.

**AINA:** Thank you so much for this wonderful session.

**Bodhi Da:** My Pleasure!

# Credit Worthiness

Dingari Sreeram, Utkarsh Yadav

Did your dream of buying a car or starting a business got shattered at the beginning of your career as your loan was rejected due to the absence of credit history? You must be thinking, how do financial institutions decide who is creditworthy and who is not. It may be surprising to know that ability to pay is not the only factor for getting a loan. What determines your loan application approval is the evidence that you can pay your obligations on time. This evidence can be your past EMI payments, loan history, credit card payments, etc. In this article, we would like to discuss creditworthiness in the Covid era and how alternate data is changing the traditional credit scoring systems and enabling financial inclusion.

## Are You Worthy?

Now, let us try to understand what creditworthiness is and how do we measure it. The underlying principles of determining an individual's creditworthiness use the Five C's of Credit, which helps the lender decide whether to issue a credit or not.

Credit bureaus collect repayment history across different credit lines to measure an individual's financial discipline and assign a credit score that quantifies the customer's creditworthiness. The credit score thus calculated mainly assesses the Character and the Capacity of the customer. For example, in India, CIBIL's credit score is calculated based on Payment History (30%), Credit Exposure (25%), Credit type and Duration(25%), and other factors such as hard inquiries for a loan (20%). Financial institutions acquire this credit score from the bureaus and use it to evaluate loan applications. Ironically, these institutions ask for credit history to give a loan, but they deny customers the chance to create credit history by rejecting a loan. It is similar to the dilemma of a fresh graduate who is asked for work experience but not given jobs to gain it in the first place.

In recent years, alternate data has augmented existing methods to assess the borrower's character and determine the conditions for the credit. The use of alternate data has been proven valuable in credit scoring. As per the Experion survey of 2019, 65% of companies already use alternate data in some form or other in credit decision making. For a developing country like India, where a majority of the population is still untouched by the formal credit system, the unavailability of credit history poses a great hurdle in assessing creditworthiness. Startups such as CreditVidhya, Perfios, Creditwatch, Capital Float & others have started harnessing the power of alternative data and advanced analytics in their credit scoring and have shown how they can be vital tools in assessing this new customer base. Before discussing the impact of these hidden customers in detail, let us first see how the retail loan market has performed in the last decade and its future growth potential.

# 5C's

| | |
|---|---|
| CHARACTER | Borrower's Financial Reputation based on history of on time payments |
| CAPACITY | Ability to repay loans on time measured by Debt to Income ratio |
| CAPITAL | Higher down payments reduces probability of defaults and increases lenders confidence |
| COLLATERAL | Lenders can get back some value from collateral in case of default |
| CONDITION | Depends on Principal amount, Interest rate and purpose of loan. |

## Retail Loans in Indian Economy

According to bank credit data released by the Reserve bank of India, outstanding retail loans as of 31st march 2021 stood at 27.74 lakh crore, which is 14% annual growth over the last decade. This growth can be attributed to an increase in confidence of the lenders due to infrastructure improvement, the presence of data due to digitization, and customers' openness to take credit. All these changes show that the cultural stigma associated with taking loans is slowly fading away. Leveraging alternate data to correctly identify and rate the risk of an untapped segment of customers with no previous credit history can further increase the business for lenders and provide loans to people in need.

## Creditworthiness in the Covid era

It is not an exaggeration to say that the Covid pandemic has affected each of us directly or indirectly. As our lives have seen ups and downs in this period, the global economies and markets also faced their fair share of difficulties. Faced with this challenge of the pandemic and threat to our lives, various governments have focussed on safety and health as their priority which is justifiable. Despite being effective in curbing the spread of the virus for some period, lockdowns have led to a rise in unemployment and a slump of cash flow. It is alarming that there is a resurgence of the third Covid wave in many countries. Although health and safety are paramount in this period, we also need to maintain good financial stability as post Covid recovery depends on it.

Due to the pandemic, many businesses and individuals are failing to pay bills and loans on time. Payment history is one of the main factors in credit rating, failure to pay the loans or credits on time will result in a drop in the individual's credit rating. As per financial institutions, the lower is the rating, the higher is the risk of customer default, and the more is the interest charged for his future loans. As of 1st March 2021, SBI charged a 6.7% interest rate for a customer with an 800+ credit rating, while others with a lower credit rating were charged higher. Due to the rise in unemployment & medical emergencies during the Covid pandemic, many customers could have missed their credit payments. Some credit card issuers and banks have offered deferred payments or late fee waivers, either voluntarily or under the direction of government policies. The resurgence of the second and now the third wave of covid-19 has made even the revised payment dates challenging for many.

Keeping in mind the impact of Covid 19, agencies cautioned people to maintain their financial health and credit scores. The credit scoring methods cannot be altered due to the pandemic as it will lead to improper credit risk assessment. The bureaus have increased access to reports and encouraged talks with lenders for deferred payments and waive-offs to handle the finances more effectively. We must stay up to date with the latest credit reports and make sure there are no errors. If any payments are missed during the pandemic, agencies have suggested adding consumer statements in the credit reports. Adding some facts like "Was ill due to Covid 19 and could not pay bills on time" in the consumer statements section of the credit report can highlight the problem.

## Alternatives to the "conventional" credit scoring

Various financial institutions & fintechs are experimenting with novel approaches to look beyond the conventional ways of assessing creditworthiness and rate customers with no past credit history. Many customers don't have a long credit history, but even such customers pay post-paid mobile bills, electricity bills, rentals, and other utility bills. Some other data sources include records of owned assets, information about payments made for alternative lending options such as informal short-term loans, and transaction activity of checking accounts (withdrawals, recurring payments, and average balance). The above-mentioned alternative data sources can give insights about customers' financial discipline and spending habits to the lenders and are termed 'Alternative data'.

While issuing loans, lenders use a range of credit scores and their associated default rates. Lenders often use a cut-off point on these default rates for deciding credit scores that are lendable. This decision is based on the economic value of the loan by comparing the profit made on non-defaulters vs. the loss made on defaulters. Considering a hypothetical example, a lender may decide not to give a loan to a customer with a score in the range of 600-650 because the default rate for this score range is 4% and is not profitable for the lender. Even though 96% of these customers will not default, they still miss out on the loans. Alternative sources of data can be used to have a more refined look at these scores. Augmenting credit scores with alternative data will move many customers from non-lendable to lendable arena and provide access to reduced interest rates on loans. Thus, alternative data is incredibly beneficial for customers. It is also helpful for lenders as it increases the number of good loans while lowering transaction costs & losses.

The Covid pandemic has accelerated an already fast-moving digital economy by making digital payments a common sight across the country. Avenues of alternative data are increasing exponentially due to the uptick in online transactions on various payment apps and eCommerce websites. This trend is creating a lucrative opportunity for lenders to assess and include more potential customers. A vast stream of this alternative data is being generated, and with the increased availability of techniques to handle this Big data through machine learning and AI, we can find insights that would play a crucial role in banking the unbanked. This increased lending activity will provide access to funds for many individuals and help them in starting a new business, continuing education, or buying productive assets.

## A word of caution

Another possible way of assessing creditworthiness would be to use psychometric tests to understand customers' personality insights and behaviours. However, it is prone to some elements of subjectivity. Some of the other alternative sources of data could be social media data, which could give a glimpse of an individual's lifestyle, traveling, and spending habits. The lenders should be very cautious while assessing data from psychometric tests and social media. This data is highly amenable to manipulation, especially when borrowers become aware of its usage in determining their creditworthiness. Alternative data is only useful when it is timely and accurately available. It should have wide coverage and should be available for most of the customers. Not all forms of alternative data are created equal; some have more predictive powers than others. The predictive power depends on the type and amount of loan for which it is used. While training with available past data, algorithms can be biased if training data fed to it is biased against gender, sexual orientation, age, race, and other characteristics. It is necessary to make sure that algorithms can reverse or minimize the prejudices in training data.

These alternative Big data sources can be used to train Machine learning algorithms and find valuable information for classifying worthy customers. Though a wide variety of models could be used to score or cluster customers based on these alternative data sources, one must be careful while selecting these algorithms. There is often a need for explainable and simpler models like decision trees compared to black-box models like deep learning (though the latter may be much more powerful!). Explainable models are preferable as they help lenders to explain to the customers why they were rejected. Similarly, lenders can explain to their top management why were the new business opportunities not seized. These models also help in explaining to the regulators why a specific business decision was taken.

The guiding principle of assessing the creditworthiness of credit-invisible people is not to provide credit to all but is to identify worthy individuals who have the ability and willingness to pay back. This sort of financial inclusion would have multiple benefits like providing credit to deserving customers, increasing the customer base for lenders, and promoting economic activity for the nation.

# Nowcasting

## An economist’s golden bow

Kolli Parasuram

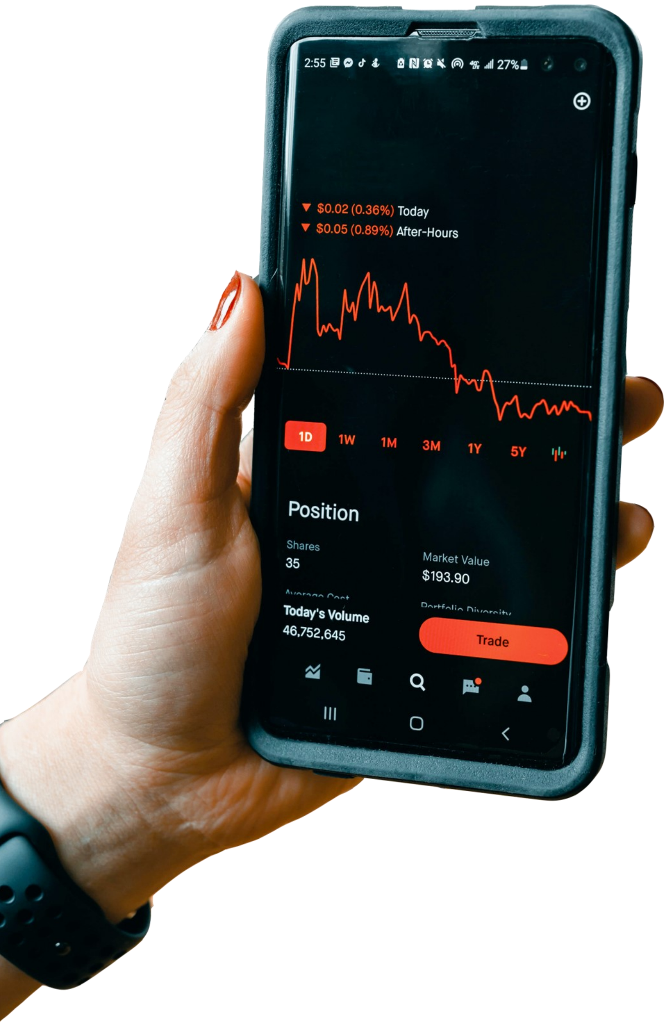
“India in a historic technical recession, RBI signals in first-ever *nowcast*” was the News in November 2020.

What does ‘nowcast’ mean here? It is the RBI’s forecast value for Indian GDP in Q3 of 2020, which they did for the first time using the Nowcasting Technique. Gross Domestic Product (GDP) is the most important measure to assess the economic state of the country. The country’s GDP is released 45 days (almost half a Quarter) post completion of the Quarter. This lag is impeding the Government from assessing the impact of its policies and updating them at the right time. This prompted economists to take more interest in accurately forecasting the country’s GDP.

Traditionally, economists used simple forecasting techniques such as VAR, ARIMAX, VECM to forecast GDP. But since the last two years, nowcasting techniques rose to fame for forecasting macroeconomic indicators in Econometrics.

The term ‘Nowcasting’ is derived from two words, i.e., ‘now’ and ‘forecasting’. Nowcasting is defined as the prediction of the present, the very near future, and the very recent past. It is extensively used in meteorology for weather predictions and recently became popular in economics as typical measures such as GDP used to assess the state of an economy are only determined after a long delay. In addition, Nowcasting helps in real-time forecasting due to its ability to handle high-frequency and varied-frequency input variables to nowcast Macroeconomic indicators.

Recently traditional nowcasting methods in combination with Deep Learning techniques have revolutionized weather prediction. In this article, we restrict ourselves to discuss the revolution caused by Nowcasting in the field of Economics and how it is able to out-perform traditional forecasting techniques used in Econometrics.



The macro-economic indicators are released at different frequencies in a country. For example, in India, CPI and IIP are released monthly, whereas GDP is released Quarterly. Often, traditional forecasting methods have failed to efficiently capture the information from high-frequency (monthly-CPI, IIP) indicators in forecasting the low-frequency (Quarterly - GDP) indicators. Let us say we want to forecast GDP for the Jan-Mar quarter of 2020. In the traditional forecasting techniques, we use all the data (data of both monthly and quarterly frequency) available till Dec '2019 and generate the forecast. This forecast will remain constant throughout the Quarter from Jan to March. Even though monthly indicators are released during this period, they remain unused. On the other hand, the nowcasting techniques enable using these monthly indicators to update its forecast.

The technique of Nowcasting used in econometrics that outperformed traditional simple linear forecasting techniques is **Dynamic Factor Modelling (DFM)**.

In DFM-Nowcasting, multiple input time series are normalized and transformed using Principal Component Analysis (PCA) to obtain factors (Components) on which the Vector AutoRegression (VAR) is applied. Actually, the PCA step enables the model to take more input time series without the problem of overfitting, which we observe in traditional forecasting techniques. Also, these factors are updated dynamically as more data points become available with time. So, the nowcasting technique keeps learning similar to Reinforced Learning (RL) in analytics.

The DFM-Nowcasting is able to outperform the traditional forecasting techniques used in econometrics due to its ability to take in data from more input variables with varied frequencies as soon as they become available.

The success of this technique in meteorology & econometrics can be attributed to the presence of multiple cross-correlated variables, which can be benefited by the PCA step as it retrieves the filtered information.

DFM-Nowcasting updates its nowcast value as soon as any of the input economic indicators are released. This enables us to calculate the impact of particular News (i.e., economic indicator released) on the forecast value. Hence, the nowcasting technique is effectively replicating an expert economist following News regularly.

Economics is the toughest subject to master because the level of impact of one macroeconomic indicator on another is very different across countries. For example, In the US, the spending nature of people depends a lot on the stock-market performance, but in India, the impact of stock-market performance on the spending behaviour of people is less. At present, the Nowcasting technique is limited to forecasting GDP in most cases of Economics, but in the coming future, this technique will be used to nowcast other macroeconomic indicators also. Nowcasting techniques will be able to help economists to Quantify the impact of the News on macroeconomic indicators. The Nowcasting technique can be used in many other fields where you need to handle many cross-correlated time series in the future.

## Glossary:

**VAR:** Vector Auto Regression

**ARIMAX:** ARIMA with Explanatory Variable

**VECM:** Vector Error Correction Model

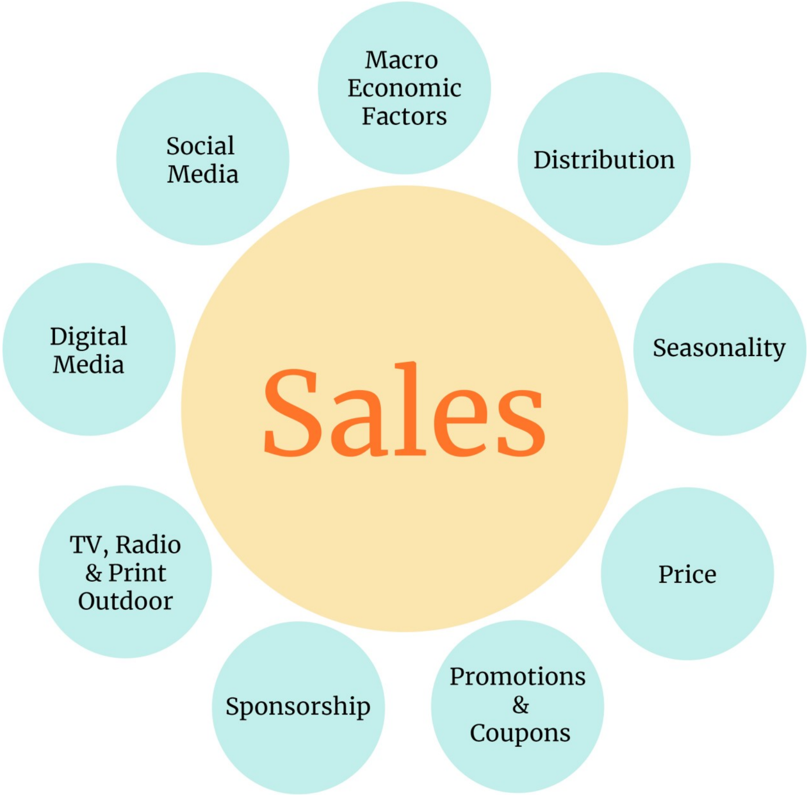
**CPI:** Consumer Price Index

**IIP:** Index of Industrial Production

# Revolutionizing Market Mix Models - Robyn

Jaihind Sawant

Kolli Parasuram



Marketing refers to activities that promote the buying or selling of goods or services. The promotion is done for the product to reach its target customers. The end goal of marketing is to increase sales with a limited budget allotted to marketing. At the start of the century, the number of channels to promote a product are limited, such as TV, Radio, Billboards, etc., which can be humanly manageable. But in today's digital world, more channels were added to promote a product, such as Facebook, YouTube, Instagram, etc. Hence, when a specific Marketing budget is allotted, how much money has to be spent on different channels to maximize a firm's sales is the burning question for any company's marketing team. This is where AI/ML helped them make data-driven decisions by creating the Market Mix Modeling (MMM) technique.

MMM technique helps understand how much each marketing input contributes to sales and the amount to spend on each marketing input to maximize sales. The initial MMM techniques employed Multi-Linear Regression with sales or Market share as the dependent variable, and independent variables are marketing inputs such as Price, TV spends, Digital platform spends, etc. This model lacks the ability to consider Diminishing Returns & Carry-over effects by advertising inputs on the sales. Diminishing Returns mean the reduction in increment of sales for a unit increase in market input amount (i.e., some advertising inputs such as TV advertising spend do not have a linear impact on sales). It becomes zero after a certain threshold. The carry-over effect is the impact on future sales caused by the amount spent today on advertising.

To include the impact of Diminishing Returns and Carry-over effect on marketing campaigns, the Weibull or geometric ad-stock functions are used in the model.

Recently, Facebook introduced the Robyn MMM technique, which revolutionized the MMM techniques by using the time series concept to model the impact of marketing inputs on sales.

Robyn technique uses Facebook's Prophet library to decompose time series into the components of Trend, Seasonality, and Holiday. Robyn time series forecasting is based on an additive model where the Trend component consists of ad-stock and linear functions. The remaining nonlinear component consists of the Seasonality and Holiday effect. The Seasonality component is modeled as a periodic function of time using the Fourier series. The use of the Fourier series enables the model to quickly adapt to change in the seasonality component by increasing the number of Fourier components. The Holiday component is independent and helps to incorporate the effect of holidays and events into the model.

Fourier Series Equation:

$$s(t) = \sum_{n=1}^{N} \left( a_n \cos \frac{2\pi nt}{P} + b_n \sin \frac{2\pi nt}{P} \right)$$

Where 0 ≤ t < P, P is the period of the time series & N is the number of sine and cosine components in the Fourier series.

In MMM, the sales are divided into two components (i.e., Baseline Sales and Incremental Sales) which will help us to calculate Marketing Return on Investment (MROI) most effectively. Baseline variables are created to include the impact of non-media variables (such as temperature, unemployment, etc.) on sales. The expected volume of sales in the absence of in-store and online promotions is called Baseline Sales. The expected additional sales volume above Baseline Sales generated by marketing activities is called Incremental Sales.

Robyn MMM uses the Bhattacharya coefficient of statistics to split data into train and test with high similarity. Robyn MMM also applies Ridge regression to avoid Multi-collinearity and Over-fitting problems. Ridge regression is a regularization technique used to reduce variance by introducing a small amount of bias. The optimal value for penalizing tuning parameter (λ) of Ridge regression is also obtained automatically.

Normally, the MMM model contains high cardinality (number) of parameters (i.e., thetas, alphas, gammas, shapes, and scales), these increase further as the number of marketing channels increases. The high dimensions (cardinality) of parameters makes the model more complex, and it takes more time to obtain optimal point using a gradient optimization algorithm. But in Robyn MMM, the value of parameters (near to optimal point) is obtained by using a gradient-free optimization algorithm called Latin Hypercube Sampling (LHS). LHS is a statistical method to generate a near-random sample of parameter values from a multi-dimensional distribution. The MROI response functions of all the market and non -marketing inputs are obtained from the parameter values given by LHS. A Nonlinear optimization problem is designed using the obtained parameters with an objective function to maximize the sales or Market share under the Marketing Budget and other constraints.

Robyn MMM is not just limited to providing the optimal market mix inputs that intend to maximize MROI within the allotted Market budget. It helps to understand the effects of holidays, weather, functioning of institutions, and the market on advertising. It enables the marketing teams to compare varying marketing techniques by stimulating their impact on future sales. In addition, it reduces human bias and helps to understand the lag and decay effect of advertising to take control of cross-media budget allocation.

# Publications by the PGDBA Alumni

## Interior Point Solving for LP-based prediction+ optimisation

**Jayanta Mandi (Batch 2015-17)**
Advances in Neural Information Processing Systems, 2020

**Abstract:** Solving optimization problem is the key to decision making in many real-life analytics applications. However, the coefficients of the optimization problems are often uncertain and dependent on external factors, such as future demand or energy- or stock prices. Machine learning (ML) models, especially neural networks, are increasingly being used to estimate these coefficients in a data-driven way. We investigate the use of the more principled logarithmic barrier term, as widely used in interior point solvers for linear programming. Instead of differentiating the KKT conditions, we consider the homogeneous self-dual formulation of the LP and we show the relation between the interior point step direction and corresponding gradients needed for learning. Finally, our empirical experiments demonstrate our approach performs as good as if not better than the state-of-the-art QPTL (Quadratic Programming task loss) formulation of Wilder et al. and SPO approach of Elmachtoub and Grigas.

https://proceedings.neurips.cc/paper/2020/hash/51311013e51adebc3c34d2cc591fefee-Abstract.html

## Entropy-Based Consumption Diversity—The Case of India

**Jayanta Mandi (Batch 2015-17)**
Book Chapter, Springer, Singapore

**Abstract:** Applied demand analysts have observed that with increasing income, there is an increase in spending on other non-food commodities, implying a hierarchical structure of consumption pattern. Evidences also supported positive correlation between household income and the dispersion of household spending both at cross-country-level analysis and at household-level analysis. These findings make the case for consumption diversity as indicators of household welfare. In this paper, we examine the stylized facts of behavioural heterogeneity across disaggregated commodity groups by employing entropy-based Theil's measure. Using National Sample Survey household expenditure data of urban sector of four major states of India for the year 2011–2012, we show the extent to which income and other demographic characteristics such as number of children explain the variation in consumption diversity.

https://link.springer.com/chapter/10.1007/978-981-13-9981-7_24

## Smart Predict-and-Optimize for Hard Combinatorial Optimization Problems

**Jayanta Mandi (Batch 2015-17)**
Proceedings of the AAAI Conference on Artificial Intelligence

**Abstract:** Recently, Smart Predict and Optimize (SPO) has been proposed for problems with a linear objective function over the predictions, more specifically linear programming problems. It takes the regret of the predictions on the linear problem into account, by repeatedly solving it during learning. We investigate the use of SPO to solve more realistic discrete optimization problems. The main challenge is the repeated solving of the optimization problem. To this end, we investigate ways to relax the problem as well as warm-starting the learning and the solving. Our results show that even for discrete problems it often suffices to train by solving the relaxation in the SPO loss.

https://ojs.aaai.org/index.php/AAAI/article/view/5521

## Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support

**Jayanta Mandi (Batch 2015-17)**
TENCON 2018-2018 IEEE Region 10 Conference

**Abstract:** Assessing risk for voluminous legal documents such as request for proposal, contracts is tedious and error prone. We have developed "risk-o-meter", a framework, based on machine learning and natural language processing to review and assess risks of any legal document. Our framework uses Paragraph Vector, an unsupervised model to generate vector representation of text. This enables the framework to learn contextual relations of legal terms and generate sensible context aware embedding. The framework then feeds the vector space into a supervised classification algorithm to predict whether a paragraph belongs to a pre-defined risk category or not. The framework thus extracts risk prone paragraphs. This technique efficiently overcomes the limitations of keyword based search. We have achieved an accuracy of 91% for the risk category having the largest training dataset. This framework will help organizations optimize effort to identify risk from large document base with minimal human intervention and thus will help to have risk mitigated sustainable growth. Its machine learning capability makes it scalable to uncover relevant information from any type of document apart from legal documents, provided the library is pre-populated and rich.

https://ieeexplore.ieee.org/abstract/document/8650382

## Hybrid Classification and Reasoning for Image-Based Constraint Solving

**Jayanta Mandi (Batch 2015-17)**
International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research, 2020

**Abstract:** There is an increased interest in solving complex constrained problems where part of the input is not given as facts, but received as raw sensor data such as images or speech. We will use 'visual sudoku' as a prototype problem, where the given cell digits are handwritten and provided as an image thereof. In this case, one first has to train and use a classifier to label the images, so that the labels can be used for solving the problem. In this paper, we explore the hybridisation of classifying the images with the reasoning of a constraint solver. We show that pure constraint reasoning on predictions does not give satisfactory results. Instead, we explore the possibilities of a tighter integration, by exposing the probabilistic estimates of the classifier to the constraint solver. This allows joint inference on these probabilistic estimates, where we use the solver to find the maximum likelihood solution.

https://link.springer.com/chapter/10.1007/978-3-030-58942-4_24

## GAIM: Game Action Information Mining Framework for Multiplayer Online Card Games (Rummy as Case Study)

**Deepanshi Seth (Batch 2017-19)**
Pacific-Asia Conference on Knowledge Discovery and Data Mining 2020

**Abstract:** We introduce GAIM, a deep-learning analytical framework that enables benchmarking and profiling of players, from the perspective of how the players react to the game state and evolution of games. In particular, we focus on multi-player, skill-based card games, and use Rummy as a case study. GAIM framework provides a novel and extensible encapsulation of the game state as an image, and uses Convolutional Neural Networks (CNN) to learn these images to calibrate the goodness of the state, in such a way that the challenges arising from multiple players, chance factors and large state space, are all abstracted.

https://link.springer.com/chapter/10.1007/978-3-030-47436-2_33

## Online Fashion Commerce: Modelling Customer Promise Date

**Preethi V (Batch 2017-19)**
SIAM International Conference on Data Mining (SDM21) - DSRE

**Abstract:** In the e-commerce space, accurate prediction of delivery dates plays a major role in customer experience as well as in optimizing the supply chain operations. Predicting a date later than the actual delivery date might sometimes result in the customer not placing the order (lost sales) while promising a date earlier than the actual delivery date would lead to a bad customer experience and consequent customer churn. In this paper, we present a machine learning-based approach for penalizing incorrect predictions differently using non-conventional loss functions, while working under various uncertainties involved in making successful deliveries such as traffic disruptions, weather conditions, supply chain, and logistics. We examine statistical, deep learning, and conventional machine learning approaches, and we propose an approach that outperformed the pre-existing rule-based models. The proposed model is deployed internally for Fashion e-Commerce and is operational.

https://arxiv.org/abs/2105.00315

## AI Based Information Retrieval System for Identifying Harmful Online Gaming Patterns

**Deepanshi Seth (Batch 2017-19)**
Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval 2021

**Abstract:** In this proposal, we present an automated, data driven, AI powered, Responsible Game Play (RGP) framework cum tool which has been integrated in our online skill gaming platform. RGP pipeline is a combination of: a) a couple of anomaly detection Rule Based Engines; b) a Deep Learning Pipeline which models the game play characteristics of healthy and engaged players to identify potentially risky players, and c) a ML based Local Expert which leverages users' longitudinal behavioural patterns and constructs new features using the adjacent AI OPS and Signal Processing Domains. We integrate the psychometric assessment to nudge and coarse correct at-risk players proactively, ahead of time.

https://doi.org/10.1145/3404835.3464921

## Game Action Modeling for Fine Grained Analyses of Player Behavior in Multi-player Card Games (Rummy as Case Study)

**Deepanshi Seth (Batch 2017-19)**
Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2020

**Abstract:** We present a deep learning framework for game action modeling, which enables fine -grained analyses of player behavior. We develop CNN-based supervised models that effectively learn the critical game play decisions from skilled players, and use these models to assess player characteristics in the system, such as their retention, engagement, deposit buckets, etc. We show that with a carefully constructed input format, that efficiently represents the game state and history as a multi-dimensional image, along with a custom architecture the model learns the strategies of the game accurately.

https://doi.org/10.1145/3394486.3403316

## A Deep Learning Framework for Ensuring Responsible Play in Skill-based Cash Gaming

**Deepanshi Seth (Batch 2017-19)**
2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)

**Abstract:** In this paper we present a deep learning model that helps identify players who are on the verge of displaying irresponsible or addictive game play on Skill based Real money gaming platforms. We use a combination of long short-term memory and adversarial auto-encoder networks to analyse game play along three tell-tale dimensions of immoderation, namely, money, time and despair. Our model provides a state of the art solution for identifying a precise set of problem gamers in skill-based cash games well ahead of time, effectively addressing the challenges of (i) extreme class imbalance, (ii) sparse and incomplete ground truth, (iii) overlapping behavioural patterns between risky and non-risky but highly engaged players.

https://ieeexplore.ieee.org/document/9356268

## Scalable Database Normalization Powered by the Crowd

**GSS Aditya Sairam , Parasuram Kolli, Anudeep Immidisetty, Pranav Kumar, Madhu Sudan B (Batch 2019-21) , Malay Bhattacharyya(Assisstant Professor at ISI Kolkata)**
Proceedings of 8th ACM IKDD CODS and 26th COMAD

**Abstract:** The use of such collective and collaborative intelligence can simplify few, if not most, of the difficult database management tasks. This paper, addresses one such problem, precisely determining functional dependencies (FDs) in a database. Our motivation is to pursue scalable normalization of a database through a gamified crowdsourcing approach.

https://dl.acm.org/doi/10.1145/3430984.3431032

## An Improved Type 2 Fuzzy C Means Clustering for MR brain image segmentation based on Possibilistic Approach and Rough Set Theory

**N T J Preetham Kumar (Batch 2020-22)**
2018 International Conference on Communication and Signal Processing (ICCSP)

**Abstract:** It is necessary to extract various attributes from an image especially in the field of neurological pathology. Magnetic Resonance Imaging (MRI) is a popularly used scanning technique for soft tissues like brain as it provides a detailed view of the tissue. It requires highly accurate segmentation algorithms to cluster a brain image into its constituent tissue regions. In consideration to this necessity, fuzzy set theory proves to be suitable to achieve tissue clustering on the brain MR images. However, the need to obtain better segmentation makes clustering efficiency more demanding. This paper proposes an advanced clustering algorithm known as Improved Rough Possibilistic Type-2 Fuzzy C Means that includes Skull Stripping and Median Filtering to enhance the performance. The proposed algorithm addresses various issues experienced by several other clustering algorithms and its superiority over them is quantitatively validated through authentic performance metrics like Jaccard Index, Accuracy and Adjusted Rand Index.

https://ieeexplore.ieee.org/document/8524438

# References

## AI In Mental Health

1. https://www.who.int/india/health-topics/mental-health#:~:text=WHO%20estimates%20that%20the%20burden,estimated%20at%20USD%201.03%20trillion.
2. https://economictimes.indiatimes.com/magazines/panache/mental-health-in-india-7-5-of-country-affected-less-than-4000-experts-available/articleshow/71500130.cms
3. https://www.worldometers.info/gdp/gdp-by-country/
4. https://thediplomat.com/2020/03/how-committed-is-india-to-mental-health/
5. https://www.nature.com/articles/s41398-020-0780-3.pdf
6. Tansey KE, Guipponi M, Hu X et al. Contribution of common genetic vari-ants to antidepressant response. Biol Psychiatry 2013;73:679-82.
7. Hayes JF, Marston L, Walters K et al. Lithium vs. valproate vs. olanzapine vs. quetiapine as maintenance monotherapy for bipolar disorder: a population-based UK cohort study using electronic health records. World Psychiatry 2016; 15:53-8
8. Pradier MF, McCoy TH Jr, Hughes M et al. Predicting treatment dropout after antidepressant initiation. Transl Psychiatry 2020; 10:60.
9. Hughes MC, Pradier MF, Ross AS et al. Assessment of a prediction model for antidepressant treatment stability using supervised topic models. JAMA Netw Open 2020;3: e205308.
10. A Review of Challenges and Opportunities in Machine Learning for Health
11. Marzyeh Ghassemi, PhD,1 Tristan Naumann, PhD,2 Peter Schulam, PhD,3 Andrew L. Beam, PhD,4 Irene Y. Chen, SM,5 and Rajesh Ranganath, PhD6
12. García-González J, Tansey KE, Hauser J et al. Pharmacogenetics of antidepressant response: a polygenic approach. Prog Neuropsychopharmacol Biol Psychiatry 2017;75:128-34
13. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning David C. Mohr,1 Mi Zhang,2 and Stephen M. Schueller1

## Art is Math

1. CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms [https://arxiv.org/pdf/1706.07068.pdf] Ahmed Elgammal et. al.
2. Can Computers Create Art? [https://arxiv.org/pdf/1801.04486.pdf] Aaron Hertzmann
3. A Neural Algorithm of Artistic Style [https://arxiv.org/pdf/1508.06576.pdf] Leon A. Gatys et. al.
4. Creativity and Artificial Intelligence: A Digital Art Perspective [https://arxiv.org/ftp/arxiv/papers/1807/1807.08195.pdf] Bo Xing and Tshilidzi Marwala
5. Google Deepdream - https://deepdreamgenerator.com/
6. https://www.washingtonpost.com/news/innovations/wp/2016/03/10/googles-psychedelic-paint-brush-raises-the-oldest-question-in-art/
7. https://www.fastcompany.com/90253470/75-of-people-think-this-ai-artist-is-human
8. https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans
9. https://www.nationalgeographic.com/culture/article/charles-minard-cartography-infographics-history
10. https://around.uoregon.edu/content/study-finds-age-3-kids-prefer-natures-fractal-patterns
11. https://theconversation.com/the-price-of-ai-art-has-the-bubble-burst-128698
12. https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134/
13. https://www.forbes.com/sites/bernardmarr/2020/09/04/3-predictions-for-the-role-of-artificial-intelligence-in-art-and-design/?sh=4374e9025bea
14. https://www.mathnasium.com/the-math-behind-art-some-of-our-favorite-artists

# References

## Precision Farming

1. https://www.hindustantimes.com/india-news/extreme-weather-events-on-the-rise/story-6fUAeJnnibsb0ExziRk5BK.html
2. https://www.edmundoptics.com/knowledge-center/application-notes/imaging/hyperspectral-and-multispectral-imaging/
3. http://www.fao.org/3/am859e/am859e01.pdf
4. https://eos.com/make-an-analysis/ndvi/
5. https://www.theguardian.com/environment/2011/nov/28/un-farmers-produce-food-population
6. https://www.ncfc.gov.in/publications/p11.pdf
7. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-3-W6/217/2019/isprs-archives-XLII-3-W6-217-2019.pdf
8. Toward Precision in Crop Yield Estimation Using Remote Sensing and Optimization Techniques - Mohamad M. Awad
9. https://www.downtoearth.org.in/blog/agriculture/why-farmers-today-need-to-take-up-precision-farming-64659
10. DERIVING CROP CALENDAR USING NDVI TIME-SERIES - Jayesh H. Patel, and Markand P. Oza
11. Mapping crop seasonality parameters using NDVI time-series derived from HJ-1 A/B data - Zhuokun Pana, Jingfeng Huanga, Qingbo Zhoud, Limin Wangd, Yongxiang Chenga, Hankui Zhange, George Alan, Blackburnf, Jing Yang, Jianhong Liuh
12. https://www.degruyter.com/document/doi/10.1515/geo-2020-0037/html
13. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/soil-organic-carbon
14. https://www.agriculture.com/crops/cover-crops/heres-why-carbonnitrogen-ratio-matters_568-ar48014
15. Applications of Remote Sensing in Precision Agriculture: A Review Rajendra P. Sishodia, Ram L. Ray and Sudhir K. Singh

## Credit Worthiness

1. https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair
2. https://www.oliverwyman.com/our-expertise/insights/2017/may/alternative-data-and-the-unbanked.html
3. https://qz.com/india/1841163/ai-big-data-can-help-indian-banks-move-beyond-just-credit-score/
4. https://www.livemint.com/industry/banking/credit-score-takes-centre-stage-in-post-covid-home-loan-boom-11614791740517.html
5. https://www.bloombergquint.com/business/retail-and-industry-loans-come-neck-and-neck-for-the-first-time-ever-shows-rbi-data
6. https://deepsense.ai/using-machine-learning-in-credit-risk-modelling-to-reduce-risk-costs/
7. https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling/
8. https://www.earlysalary.com/blogs/the-convergence-of-machine-learning-and-big-data-in-credit-risk-management
9. https://analyticsindiamag.com/top-credit-scoring-startups-in-india-that-use-ai/

## The Real of Quantum Computing

1. https://www.ibm.com/quantum-computing/what-is-quantum-computing/
2. https://ai.googleblog.com/search/label/Quantum%20Computing
3. https://azure.microsoft.com/en-in/overview/what-is-quantum-computing
4. https://hbr.org/2021/07/quantum-computing-is-coming-what-can-it-do
5. https://towardsdatascience.com/quantum-computing-and-ai-789fc9c28c5b

# References

## Impede The Oblivion

1. https://www.techexplorist.com/origin-water-earth/34976/
2. https://advances.sciencemag.org/content/2/2/e1500323.full
3. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
4. https://public.wmo.int/en/media/press-release/climate-change-indicators-and-impacts-worsened-2020
5. https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data
6. https://scripps.ucsd.edu/research/climate-change-resources/carbon-dioxide-and-climate-change
7. https://scripps.ucsd.edu/research/climate-change-resources/faq-ocean-acidification
8. https://reasons.org/explore/blogs/todays-new-reason-to-believe/black-carbons-link-to-climate-change
9. https://news.climate.columbia.edu/2018/06/05/artificial-intelligence-climate-environment/
10. https://deepmind.com/about/deepmind-for-google
11. https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control
12. https://hai.stanford.edu/news/ais-carbon-footprint-problem
13. https://doi.org/10.1016/S0140-6736(07)61253-7
14. https://epi.yale.edu/epi-results/2020/component/epi

## A New Hope

1. Transforming healthcare with AI: The impact on the workforce and organizations | McKinsey
2. The $11.9 Trillion Global Healthcare Market: Key Opportunities & Strategies (2014-2022) - ResearchAndMarkets.com | Business Wire
3. https://www.un.org/en/development/desa/population/publications/pdf/ageing/WP A2017_Highlights.pdf
4. https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf
5. Chapter7 (who.int)

## Advancements in AI

1. https://achievements.ai/
2. https://www.digitaltrends.com/features/2020-ai-major-milestones/
3. https://doi.org/10.1002/aisy.202000245
4. Fazel Bateni, Robert W. Epps, Kameel Abdel-latif, Rokas Dargis, Suyong Han, Amanda A. Volk, Mahdi Ramezani, Tong Cai, Ou Chen, Milad Abolhasani, Ultrafast cation doping of perovskite quantum dots in flow, Matter, 10.1016/j.matt.2021.04.025, (2021). https://doi.org/10.1002/aisy.202000245
5. Willett, F.R., Avansino, D.T., Hochberg, L.R. et al. High-performance brain-to-text communication via handwriting. Nature 593, 249–254 (2021). https://doi.org/10.1038/s41586-021-03506-2
6. https://www.jagranjosh.com/current-affairs/mayflower-400-worlds-first-artificial-intelligence-ship-all-you-need-to-know-1620212363-1

## Sports Analytics

1. https://en.wikipedia.org/wiki/Moneyball_(film)
2. https://www.analyticsinsight.net/top-5-big-data-and-ai-sports-companies-of-2019/
3. https://towardsdatascience.com/scope-of-analytics-in-sports-world-37ed09c39860
4. https://en.wikipedia.org/wiki/Sports_analytics
5. https://blogs.sas.com/content/sascom/2020/06/08/going-beyond-the-box-score-text-analysis-in-sports/

# References

## Nowcasting

1. https://www.business-standard.com/article/economy-policy/india-in-historic-technicalrecession-rbi-signals-in-first-ever-nowcast-120111200122_1.html
2. https://www.newyorkfed.org/research/policy/nowcast
3. https://medium.com/pocasi/superiority-of-ai-in-the-precipitation-nowcasting-2b7cd445ec59

## Revolutionizing Market Mix models

1. https://blog.clairvoyantsoft.com/market-mix-modeling-mmm-introduction-methodology-and-use-case-dc5ae68820f8
2. https://static.googleusercontent.com/media/research.google.com/en//pub s/archive/45998.pdf
3. https://facebookexperimental.github.io/Robyn/docs/about
4. https://research.google/pubs/pub46000/
5. https://facebookexperimental.github.io/Robyn/

## Parts of Speech Tagging using Hidden Markov Model

1. https://www.udacity.com/course/natural-language-processing-nanodegree--nd892
2. Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language, 6, 225-242.

## On the way to Disaster Resilience

1. https://www.independent.co.uk/climate-change/news/canada-heat-wave-lytton-fire-b1877594.html
2. https://www.statista.com/statistics/510894/natural-disasters-globally-and-economic-losses/
3. https://www.livescience.com/33316-top-10-deadliest-natural-disasters.html
4. https://www.preventionweb.net/news/view/78713
5. https://www.nature.com/articles/s41586-018-0438-y
6. https://www.itu.int/en/myitu/News/2020/10/20/14/54/AI-for-Good-Disaster-Risk-Reduction-artificial-intelligence
7. https://link.springer.com/article/10.1007/s11069-020-04124-3
8. https://www.fujitsu.com/global/about/resources/news/press-releases/2021/0216-01.html

## Covid vaccine Trials

1. https://www.statista.com/statistics/1102816/coronavirus-covid19-cases-number-us-americans-by-day/
2. https://www.nyas.org/news-articles/academy-news/new-phase-3-trial-data-suggest-johnson-johnson-s-covid-19-vaccine-could-be-efficacious-against-newly-emerging-variants/
3. https://www.healthline.com/health/adult-vaccines/johnson-and-johnson-vaccine-efficacy#asymptomatic-covid-19
4. https://www.ama-assn.org/delivering-care/public-health/what-doctors-wish-patients-knew-about-johnson-johnson-vaccine
5. https://www.nbcboston.com/news/national-international/pfizer-moderna-johnson-johnson-how-effective-are-they-and-how-do-they-compare/2345756/
6. https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(21)00075-X/fulltext
7. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html
8. https://www.thehindu.com/sci-tech/health/the-hindu-explains-why-is-phase-3-of-covid-19-vaccine-trial-complicated/article32590533.ece

01101100 01101111
01110110 01100101
00100000 01111001
01101111 01110101