Studies in Rhythm Engineering

Anirban Bandyopadhyay
Kanad Ray *Editors*

Rhythmic Advantages in Big Data and Machine Learning

Springer

Studies in Rhythm Engineering

This is a multi-disciplinary book series, ranging from astrophysics to biology, chemistry, mathematics, geophysics and materials science. Its primary scope is the fundamental science and associated engineering wherever cyclic and rhythmic oscillations are observed.

Time neither being an entity nor a process is unmeasurable and undefined, although a clock only measures the passage of time. The clock drove recurring processes are observed in the biological rhythms, astrophysical and geophysical environments. Always, the clocks are nested, arranged in a geometric shape to govern a phenomenon in nature. These clocks are made of atoms, molecules, their circuits, and complex networks. From biology to astrophysics, the clocks have enriched the science and engineering and the series would act as a catalyst and capture the forthcoming revolution of time cycles expected to unfold in the 21st Century. From the cyclic universe to the time crystal, the book series makes a journey through time to explore the path that time follows.

The series publishes monographs and edited volumes.

More information about this series at https://link.springer.com/bookseries/16136

Anirban Bandyopadhyay · Kanad Ray
Editors

Rhythmic Advantages in Big Data and Machine Learning

Springer

Editors
Anirban Bandyopadhyay
National Institute for Materials Science
Tsukuba, Japan

Kanad Ray
Department of Physics
AMITY University Rajasthan
Jaipur, India

ISSN 2524-5546 ISSN 2524-5554 (electronic)
Studies in Rhythm Engineering
ISBN 978-981-16-5722-1 ISBN 978-981-16-5723-8 (eBook)
https://doi.org/10.1007/978-981-16-5723-8

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The current book in the series of Systems in Rhythm Engineering, SRE, compiles rhythms from Big data in pure computation, astrophysics to basic biological structures. In the earlier book edited by us, we concentrated primarily on the biophysics related studies, because most rhythm related researches are carried out in biophysics. Then Tanusree Dutta edited a book on economics, before getting into the field of rhythms in economic data, it was an effort to dwell on various aspects of rhythms where we could advance rhythm engineering. In future, we have a plan to discuss on share market rhythms and economic disaster predictions. Here we have made an effort to the field of astrophysics and big data. Our target is to create world's most authentic and rigorous database of rhythms in every single field possible. This book is the first comprehensive message that we wish to deliver to the world.

In Chap. 1, Rajdeep and Somesh have made an effort to discuss big data and how rhythm engineering could be used in understanding and fathoming the intricate dynamic patterns in a big data. An architecture of rhythm that we advocate as the fundamental information structure of nature could play pivotal role in the understanding of big data.

In Chap. 2, Joseph Singuinetti has described how brain rhythms could be perturbed to understand its true nature. Especially by concentrating on the transcranial ultrasound perturbation to neural oscillation, the author has made an effort to understand how brain rhythms are organized in an internal architecture that would regulate human cognition and behavior. Focused ultrasound is getting popular day by day and Joseph's documentation would help us understanding how rhythms organize in a human brain.

In Chap. 3, Mario Pinehiro has described vortex dynamics as a foundation of energy minimization process and entropy maximization process. It is amazing to think that loop of fields is engaging to create this universe. Mario has gone to the fundamental details of this mechanism to understand how rhythms could alone construct the foundation of this universe.

In Chap. 4, Shiva Kumar Singh and Marcos a Avila have taken thermal flow as a foundation to understand fractal superstructures contributing to the thermoelectric applications. Thermoelectric materials are very important because they resemble the biomaterials that harvest electrical energy from thermal noise using fractal structures. Thermal signals are electromagnetic and electrical signals are part of electromagnetic field. Such conversions are key governing features of energy harvesting and splitting the two parts of electromagnetic energy.

In Chap. 5, we return to brain modeling again with Taakaki Musha, he has been working on superluminal particles that travel faster than the velocity of light. Are they really produced in the brain? Do they take part in the brain cognition? Explore it here with Taakaki.

In Chap. 6, Pushpendra Singh and others have reported their work on building experimentally an artificial cortex with more than 10000 cortical columns. Real brain has roughly 200000 cortical columns constituting the cortex where in 47 Brodmann's regions, brain's major cognitive information processing takes place.

In Chap. 7, Max Rempel has taken an effort to look into the DNA resonance in a very different way. Normally, we look at dielectric property of a material directly, however, different symmetries have different contributions to the formation of a resonance band. Max has looked into the DNA as an assembly of a tiny semi-circular block and made an effort to understand the mystery of junk DNA. Do they really have codes that we really don't understand?

In Chap. 8, Max Rempel has returned with a topic on quantum consciousness, we have always made an effort to keep one topic on consciousness in our edited compilation of rhythm engineering. Quantum consciousness is interesting, exciting, and extremely debated in the scientific community.

In Chap. 9, Max Calliguri takes us one step further into the domain of microtubules and he advocates a hypothesis that brain carries out hypercomputation using water crystals inside the microtubules. This is where we think our book should end.

We invite all scholars to have a critical debate on the special selections we have made to explore the systems in rhythm engineering.

Jaipur, India Kanad Ray
Tsukuba, Japan Anirban Bandyopadhyay

Contents

Editors and Contributors

About the Editors

Anirban Bandyopadhyay is Senior Principal Scientist at the National Institute for Materials Science (NIMS), Tsukuba, Japan. He received Ph.D. in Supramolecular Electronics at the Indian Association for the Cultivation of Science (IACS), Kolkata, on 2005. From 2005 to 2008, he was ICYS Research Fellow at the ICYS, NIMS, Japan, and worked on the brain-like bio-processor. In 2008, he joined as a permanent scientist at NIMS, working on the time crystal model of human brain and design-synthesis of brain-like organic jelly, written a book "Nanobrain: The making of an artificial brain from a time crystal," on 2020. From 2013 to 2014, he was a visiting scientist at the Massachusetts Institute of Technology (MIT), USA. He has received Hitachi Science and Technology Award, 2010, Inamori Foundation Award, 2011–2012, Kurata Foundation Award, Inamori Foundation Fellow (2011), and Sewa Society International Member, Japan.

Kanad Ray (Senior Member, IEEE) received the M.Sc. degree in Physics from Calcutta University and the Ph.D. degree in Physics from Jadavpur University, West Bengal, India. He has been Professor of Physics and Electronics and Communication and is presently working as Head of the Department of Physics, Amity School of Applied Sciences, Amity University Rajasthan (AUR), Jaipur, India. His current research areas of interest include cognition, communication, electromagnetic field theory, antenna and wave propagation, microwave, computational biology, and applied physics. He has been serving as Editor for various Springer book series. He was Associate Editor of the Journal of Integrative Neuroscience (The Netherlands: IOS Press). He has visited several countries such as Netherlands, Turkey, China, Czechoslovakia, Russia, Portugal, Finland, Belgium, South Africa, Japan, Singapore, Thailand, and Malaysia for various academic missions.

Contributors

B. Aswathy Materials Science and Technology Division, CSIR-National Institute for Interdisciplinary Science and Technology (CSIR-NIIST), Thiruvanathapuram, Kerala, India

Marcos A. Avila CCNH, Universidade Federal do ABC (UFABC), Santo André, SP, Brazil

Anirban Bandyopadhyay International Center for Materials and Nanoarchitectronics (MANA), Research Center for Advanced Measurement and Characterization (RCAMC), NIMS, Tsukuba, Ibaraki, Japan

Rajdeep Banerjee Data Scientist, Strategic Growth Realisation, Bangalore, India

Somesh Kr. Bhattacharya National Institute for Materials Science, Tsukuba, Japan;
United Network of Professionals Pvt. Ltd, FD Block, Kolkata, India

Luigi Maxmilian Caligiuri Foundation of Physics Research Center (FoPRC), Cosenza, Italy

Daisuke Fujita International Center for Materials and Nanoarchitectronics (MANA), Research Center for Advanced Measurement and Characterization (RCAMC), NIMS, Tsukuba, Ibaraki, Japan

Subrata Ghosh Chemical Science and Technology Division, CSIR-North East Institute of Science and Technology, NEIST, Jorhat, Assam, India;
Academy of Scientific and Innovative Research (AcSIR), CSIR-NEIST Campus, Jorhat, Assam, India

Richard Alan Miller OAK, Inc., Grants Pass, OR, USA

Takaaki Musha Advanced Science-Technology Research Organization, Yokohama, Japan;
Foundation of Physics Research Center (FoPRC), Cosenza, Italy

Max Myakishev-Rempel DNA Resonance Research Foundation, San Diego, CA, USA;
Localized Therapeutics, San Diego, CA, USA

Mario J. Pinheiro Department of Physics, Instituto Superior Técnico-IST, Universidade de Lisboa, Lisboa Codex, Portugal

Kanad Ray Amity School of Applied Sciences, Amity University Rajasthan, Kant Kalwar, Jaipur, Rajasthan, India

Pathik Sahoo International Center for Materials and Nanoarchitectronics (MANA), Research Center for Advanced Measurement and Characterization (RCAMC), NIMS, Tsukuba, Ibaraki, Japan

Joseph L. Sanguinetti Center for Consciousness Studies, University of Arizona, Tucson, Arizona, USA

Ivan Viktorovich Savelev DNA Resonance Research Foundation, San Diego, CA, USA;
Localized Therapeutics, San Diego, CA, USA

Pushpendra Singh International Center for Materials and Nanoarchitectronics (MANA), Research Center for Advanced Measurement and Characterization (RCAMC), NIMS, Tsukuba, Ibaraki, Japan;
Amity School of Applied Sciences, Amity University Rajasthan, Kant Kalwar, Jaipur, Rajasthan, India

Shiva Kumar Singh CCNH, Universidade Federal do ABC (UFABC), Santo André, SP, Brazil

Chapter 1
Data: Periodicity and Ways to Unlock Its Full Potential

Rajdeep Banerjee and Somesh Kr. Bhattacharya

1 Introduction

A set of features or information collected through observation is known as "**Data**" [1]. Technically speaking, data is a set of numerical values, or images, or videos, or even text of about one or more persons or objects. In 2006, British mathematician Clive Humby coined the phrase "Data being the new oil" and we are increasingly seeing how true it is. Data is a vital resource that powers the information economy. Additionally, data isn't just abundant, it is cumulative. Finally, data is a non-rival good as the same can be used elsewhere for beneficial purposes. However, the data to be used should be prepared in a timely manner and must be reliable, accurate, complete, and relevant. However, raw data isn't much meaningful and requires processing. Data when processed and presented in a contextual format becomes information [2]. An important property of data is the presence of periodicity. Anything which doesn't have periodicity is "*noise*". In almost in every field of science, engineering, finance, and even music, periodicity is observed. The planetary motion, phases of the moon, solar bursts are a few examples of periodicity in astronomical data. Similarly, meteorological as well as seismological data have periodicity too. The highs and lows of the stock market have both seasonality and cyclical factors embedded in it giving rise to periodicity. Even in health care, one can see the periodic nature. The appearance of diseases related to seasonal changes is a very good example of that. Lastly, periodicity can be seen easily in music. The nodes repeat themselves at regular intervals giving rise to periodicity. Detecting periodicity in data is a field in itself

R. Banerjee
Data Scientist, Strategic Growth Realisation, Cerner India, Bangalore 560064, India

S. Kr. Bhattacharya (✉)
National Institute for Materials Science, 1-2-1 Sengen, Tsukuba 3050047, Japan

United Network of Professionals Pvt. Ltd, FD Block, Sector-III, Bidhan Nagar, Kolkata 700106, India

A. Bandyopadhyay and K. Ray (eds.), *Rhythmic Advantages in Big Data and Machine Learning* , Studies in Rhythm Engineering,
https://doi.org/10.1007/978-981-16-5723-8_1

and includes methods like Discrete Fourier Transform (DFT) for frequency domain analysis [3] or computing the periodogram of a signal and using autocorrelation to find the periodicity [4]. We will discuss the periodicity later in this chapter.

In today's world, anything and everything is data and we generate voluminous data almost every passing second which actually gave birth to "Big Data". The phrase "Big Data" which buzzes around us everywhere, is applied to a specific type of data that has certain traits. However, this phrase has been over used and often incorrectly, which is why itis difficult to gauge its true meaning. For commoners, it is very difficult to understand if big data is a tool or a technology or just a buzzword used by data scientists to scare us. Another concern is if big data really has the potential to usher in dramatic changes or will the hype fade away with time. In any case, over the past few years, big data has become an integral part of several industries and in many cases has shown the potential to be a game-changer.

2 What is Big Data?

To put it simply, big data describes the massive volume of both structured and unstructured data which is so enormous that processing this voluminous data requires special techniques. In short, big data is simply a whole lot of data. This concept is a relatively new one and it constitutes not only the increasing amount but also the diverse types of data that gets collected. As more and more information moves online and gets digitized, it paves way for the analysts to start using it as data. For instance, social media posts, online books, music, videos, along with their reviews, as well as the increased number of sensors have all added to the staggering increase in the amount of data that has become available for analysis. Everything we do now gets stored and tracked as data. Starting from online shopping to browsing movies generates data about our requirements, choices, and lifestyles. Our smartphone constantly uploads data about our location, movement as well as the apps we use.

So far, we have talked about big data from the perspective of its volume. However, big data isn't just the volume of data we generate, it's also about the different types of

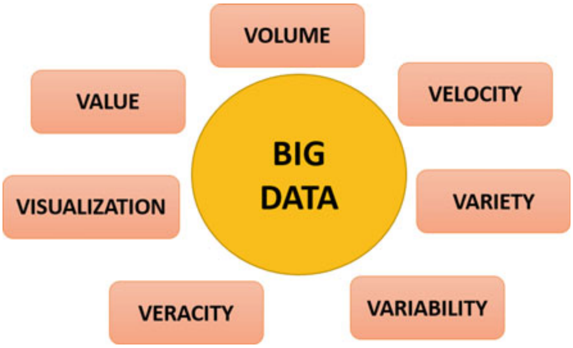

Fig. 1 The seven Versus of big data

data, viz. *text, video, search logs, sensor logs, customer transactions,* etc. Formally, big data is the data that satisfies the following ***seven Vs***: (Fig. 1)

- **Volume**: In every sense, big data is really big! With the dramatic growth of the Internet, mobile devices, social media, and Internet of Things (IoT) technology, the amount of data generated from all these sources has grown accordingly.
- **Velocity**: In addition to volume, the generation of data and the organizations' ability to process it is accelerating.
- **Variety**: In earlier times, most data types could be neatly captured in rows on a structured table. In the big data world, data often comes in unstructured formats like social media posts, server log data, lat-long geocoordinates, photos, audio, video, and free text.
- **Variability**: The meaning of words in the text (unstructured data) can change based on the context.
- **Veracity**: With many different data types and data sources, data quality issues invariably pop up in big data sets. Veracity deals with exploring a data set for data quality and systematically cleansing that data to be useful for analysis.
- **Visualization**: Data visualization is a great tool for data analysis. Exploratory data analysis (EDA) can provide quick and in-depth insight to the end-users.
- **Value**: Data must be combined with rigorous processing and analysis to be useful.

3 Big Data Terminologies

The uncanny feeling with big data comes from the variety of new terms that are associated with it. Here is a quick run-down of the most popular ones:

A. Algorithm: This is the backbone of big data analytics. The underlying mathematical intricacies help in gaining insights and forms the core of decision science.

B. Cloud computing: As big data is voluminous, processing it using the algorithms often specializes in structure. Cloud computing has emerged as a front runner in this and these remote servers have an edge over local machines.

C. Data scientist: A breed of professionals who analyze data and extract vital insights from it.

D. Hadoop: A collection of programs that allow for the storage, retrieval, and analysis of very large data sets.

E. Internet of Things (IoT): IoT refers to objects like sensors that collect, analyze and transmit their data (often without human interference).

F. Predictive analytics: The science of using data analytics to predict trends for future events.

G. Structured versus unstructured data: Structured data is anything that can be organized in a table so that it relates to other data in the same table. Unstructured data is everything that isn't structured like texts, images, and videos.

H. Web scraping: The tool of collection and structuring of data from websites usually through writing scripts.

4 Significance of Big Data

The importance of big data has grown proportionally with the volume of data being produced every day. Businesses have leveraged big data in a big way to understand a number of factors to drive their pursuit of success. Companies have used big data analytics to learn their customer's demand, choices, their best customers, and what they are missing out on conversions. Organizations gather useful insights into their sales and marketing questions so that they can optimize their campaigns and learn more about customer behavior.

Another area where big data analytics is handy for businesses is to make confident decisions. Big data can help companies make choices with confidence, based on an in-depth analysis of what they know about your marketplace, industry, and customers. One of the biggest benefits of big data is its accuracy. Big data analytics provide a complete overview of everything that has been learned so far as organizations grow and suggest positive changes for the company.

Knowledge is power. This concept is at the heart of big data analytics. Big data technologies like cloud computing and machine learning help companies to stay ahead of the curve by identifying inefficiencies and opportunities in company practices.

Finally, as a new generation of technology leaders enter the marketplace, big data delivers the agility and innovation top-tier talent needs from their employer. For instance, millennials are natural technology natives. The younger people in the team will expect access to technology that allows them to make useful decisions rapidly. By constantly collecting and analyzing information, one can create an agile culture that's ready to evolve to suit the latest trends.

5 Applications of Big Data

Big data technologies are leveraged to boost efficiency and design data-driven decisions. Big data finds applications in several industries where it is mandatory to analyze historical data to predict/forecast future performances. Though big data has penetrated almost every field, we mention few domains which leverage big data regularly:

Health Care

As discussed in the previous sections, big data involves data that is high in the four Vs, viz., volume, velocity, variety, and veracity. Since the digitization of data in the form of electronic medical records has begun, health care, because of its inherent need in society, has become one of the major domains of big data, both in terms of source and utilization. In healthcare, the source of the data, in general, determines its utilization. When we talk about utilization it generally means how this data is gobbled up using models and metrics to provide better insights and predictions. In general, the use-cases involve two types of analysis of the data—first, algorithmic––simple logic-based strategy development. These mainly involve a lot of domain knowledge and a limited number of parameters to consider. On the other hand, when the number of parameters at hand is large and in some cases their dependencies are not well defined, we turn to the second approach of machine learning and deep learning—artificial intelligence based predictions and strategies.

In the following few paragraphs, we would like to focus on different sources of big data in health care and see how the utilization differs based on the source. The primary sources of big data in health care can be broadly classified (non-exhaustive) into the following categories:

1. Electronic medical/health records (EMR/EHR),
2. Diagnostic tests data,
3. Social survey data,
4. Insurance claims data, and
5. Inventory management data.

We will briefly describe the kind of data each of these categories holds, and will then go into the possible use-cases in each category, that the data can provide insights and solutions to.

1. Electronic medical/health records (EMR/EHR):
   This is the type of data that the hospitals and the primary healthcare centers (PHCs) collect from patients, which include specifics about admission records, past illness history, patient demographics, present illness, diagnosis, doctor's prescription, etc. In general, these kinds of data are filled in through client (PHCs)-specific software, and follow some general schema, and therefore are structured, and easy to use for further processing.
   EMR/EHRs can help discover phenotype-genotype associations, improve clinical trial protocols, automate adverse drug event detection and prevention, and accelerate precision medicine research [5], using either an algorithmic or machine learning approach. One of the major use cases that has come up recently in this domain is to use machine learning or deep learning models to predict or reduce patient readmissions in health centers [6].
2. Diagnostic tests data:
   These types of data can range from tabular formats, such as lab tests, to image data such as X-ray, CT-scans, MRI, USG, ECG, EEG, etc. Generally, lab test results are inputted through specific software are tabular and structured. But image data sources need special care, especially images of the same use-case, from different sources differ a lot. It should also be noted that in cases where results are hand-written/filled in a form, digitization may include handwritten

text based natural language processing methods, before can be used for any kind of machine learning models.

Recently, deep learning based models are being used extensively for medical image processing and classification [7]. Some of these models are already showing potential to prove as an aide to the medical professionals in understanding and diagnosis of ailments. The use of chat-bots is also growing as primary detection and identification of diseases.

3. Social survey data:

   A large chunk of these types of data are mainly social determinants of health (SDOH) data, resulting from various government and non-government surveys, which are conducted based on geographic locations. SDOH data, in general, come into the following five categories: (i) economic stability; (ii) education access and quality; (iii) social and community context; (iv) neighborhood and built environment; and (v) healthcare access and quality.

   The SDOH has numerous use cases, out of which it has already shown significant impact in improving mortality, morbidity, life expectancy, healthcare expenditures, health status, and functional limitations [8]. Detection of disease prevalence based on age, living conditions, social or community status, etc., can also help in developing a more targeted approach to health care.

4. Insurance claims data:

   As the name suggests, these types of data contain mainly the financial records, such as hospital charges, bills paid, dues, amount claimed through insurance, type of insurance, and so on.

   The primary use cases include predicting claims rejection ratio and risk adjustment by insurers for efficient care management for high-risk individuals [9]. A better understanding of such data can improve insurance policies. Sometimes, where the is a mass insurance policy available, the claims data can be merged with SDOH data to improve the predictive power of the models.

5. Inventory management:

   Types of inventory management include the storage, utilization, and order of medicines, medical equipment, or any kind of resources related to patient care or hospitality.

   Inventory management use-cases can range from reducing medicine wastage, supply-chain management, to predicting high-demand scenarios beforehand, understanding the seasonality of diseases to stockpile specific medications and healthcare equipment, predicting medication usage based on historical data [10], etc.

As we can see, based on the sources, how the use-cases of the data differ. This leads to the variety of approaches that we can take advantage of. In the past few decades, the human race has made great strides in medicines and health care which is evident from the increase in average life expectancies over the years. But at the same time, we are also witnessing more and more disease outbreaks due to global warming, man-wild interactions and conflicts, etc. Therefore, we must make use of the huge power of the data that is ever-growing to tilt the balance toward us, humans, a little more.

Banking and Finance

Banking and Financial Service (BFS) is one of those domains which heavily employs and utilizes big data technology and its associated platforms regularly. Among various applications of big data in BFS, the two most important ones are: (a) Credit card fraud detection, and (b) Stock price prediction (forecasting). Credit card fraud detection provides a unique challenge in big data analytics. Previously, credit card companies use to call up customers to verify their transactions, a method that is both inefficient and expensive. Tracking customer spending patterns and identifying fraud-affected areas, big data technology can help to minimize these frauds.

Similarly, the stock price forecast is also critical for companies involved in the investment business. Taking into account the historical data and latest company news, big data can make reasonable predictions of the trend in the stock market. This allows investors to make strategic decisions about their investments.

Meteorology

Meteorology (weather forecast) is another critical area where big data is used extensively. Rainfall forecast is critical for countries that are heavily dependent on agriculture. Additionally, the weather forecast is also essential for several businesses for daily operations. It also helps in reducing losses caused by natural calamities like floods.

Seismology

The prime objective of seismic exploration is to develop an image of the subsurface geology. Seismographs can help us to understand the intricacies of subsurface geology. Additionally, using big data platforms like Hadoop, Hive, we can forecast seismological events using time series analysis [11].

Public sector.

Federal, as well as local governments, can use big data to plan their projects. It can be used in several public sectors like education, hospital, insurance, etc. Big data can be used to analyze the requirements in public sectors, and understand the need to implement projects.

Sports

Big data has been used to improve training and understanding competitors, using sport sensors. Big data has also been leveraged to predict winners in several sports like NBA, soccer, baseball, American soccer, etc. Additionally, the future performance of players can be predicted too. This helps in determining a players' value and salary by collecting data throughout the season [12].

In Formula One races, the use of technology goes a notch up. Racing cars are fitted with hundreds of sensors that generate terabytes of data. These sensors collect data from tire pressure to fuel burn efficiency [13]. https://en.wikipedia.org/wiki/Big_ data-cite_note-134 Based on these data, engineers and data analysts decide whether adjustments should be made in order to win a race. Besides, using big data, race teams try to predict the time they will finish the race beforehand, based on simulations using data collected over the season [14].

6 Dealing with Numeric Data, Images, and Videos

As discussed in Sect. 3, data can be in form of numbers, images or videos, and even text. Today in the realm of big data, these different types of data are analyzed to extract deep insights. There is a huge volume of work that deals with different algorithms used to analyze these data types. These algorithms collectively form the most talked about phrase of our time “Artificial Intelligence”.

The field of AI began in the 1950s with Alan Turing’s question: “Can machines think?” [15]. The quest to answer this led to the foundation of AI which since then has been progressing rapidly with applications to a wide variety of areas. The main aim of AI is to enable a machine to think and take decisions, perform repetitive tasks, and eliminate human errors by imitating human behavior and abilities. AI is broadly classified into three categories:

(a) **Artificial Narrow Intelligence (ANI)**: ANI is the most common and the weakest type of AI present today which is programmed to perform only a single specific task. Few examples of ANI are Siri, Autopilot in airplanes, chatbots, self-driving cars, and so on [16] (Fig. 2).

(b) **Artificial General Intelligence (AGI)**: AGI machines, conceptualized but not realized yet, will have the ability to learn, understand, and act in a way that is indistinguishable from a human in a given situation. AGI machines can perform multitasking under certain conditions [17].

(c) **Artificial Super Intelligence (ASI)**: ASI is a hypothetical AI where machines will exhibit intelligence superseding human capabilities. ASI will be a multifaceted intelligent machine with far superior power to solve and decide compared to human beings.

One most common and widespread subset of AI is Machine Learning (ML), which uses statistical learning algorithms that have the ability to learn and improve from experiences without explicit programming. From recommender systems on Netflix or YouTube or Spotify, to search engines like Google or Yahoo to even voice assistants like Amazon Alexa, all employ ML to perform their respective tasks. The ML algorithms are trained with a lot of data, which allows them to process information and learn from it.

ML algorithms can be broadly classified into three types:

(a) Supervised Learning: In supervised learning, the data has two components—an input variable/feature and an output variable. The ML algorithm learns to map between the input and output variables. In other words, a supervised learning

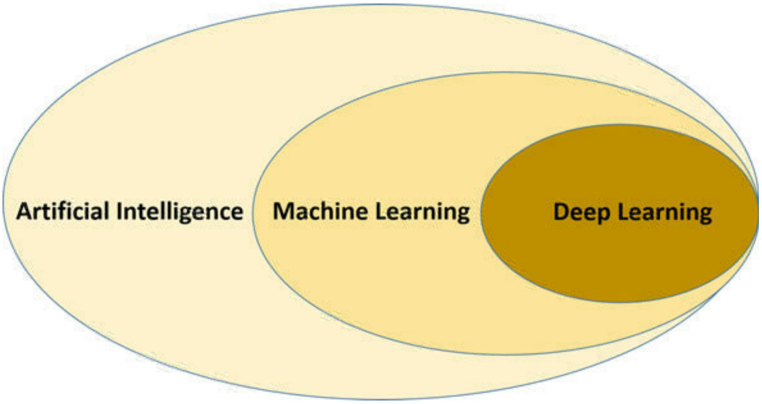

Fig. 2. Schematic representation of Artificial Intelligence and its subsets, viz., Machine Learning and Deep Learning

algorithm takes a known set of input datasets and its known responses to the data (output) to learn and perform the regression or classification model.

(b) Unsupervised Learning: Contrary to supervised learning, unsupervised learning does not have a labeled data. In this case, the algorithm tries to find similarities, differences and infer patterns in the data without any specific reference to outputs.

(c) Reinforcement Learning: The third type of ML which is completely distinct from the previous two is reinforcement learning. In reinforcement learning, the algorithm learns continuously by interacting with the environment. There is an agent which learns from the environment via trial and error by using feedback from its previous actions and experiences. Thus, using this reward-penalty method the algorithm learns to perform the correct actions.

Another subset of machine learning is “**deep learning**” which is inspired by the way our brain processes information. The theoretical base of deep learning is based on the concept of neural networks proposed independently by Alexander Bain [18] and William James [19]. In 1958, psychologist Frank Rosenblatt created the “perceptron” algorithm for pattern recognition based on a two-layer learning computer network using simple addition and subtraction [20]. The backpropagation algorithm created by Werbos allowed the neural networks to learn the relationship between complex, non-linear, and process parameters and update themselves from their failures [21]. Using the formulations of perceptron and backpropagation, Yann LeCun, YoshuaBengio, and Geoffrey Hinton formulated the “deep learning” architecture [22]. With the advancement in algorithms, deep learning has gained tremendous popularity over time. Deep learning has penetrated almost every domain like the field of computer vision, speech recognition, natural language processing, bioinformatics, drug design, etc. It is expected to match human intelligence or even surpass it although at present it is still limited to artificial narrow intelligence.

7 Periodicity and Time Series

At the beginning of the chapter, we stated briefly discussed periodicity in data. Spatiotemporal data, which has both information of space and time, has gained popularity with the advancement in tracking systems like GPS, smartphones, sensor tags attached to animals, RFID tags on merchandise, and even social media. As this type of data becomes widely popular, a great deal of effort has been put forward to analyze them as it can provide valuable insights about the customer behavior, wildlife, etc. The most common analysis of spatiotemporal data involves mining periodicity. Mining periodicity is a challenging problem as it involves bridging the raw data and its semantic understanding. For example, migratory birds never follow the exact same path every year. Weather conditions can affect their migration routes strongly. Similarly, it's quite common to have sparse or even incomplete observations as all the migratory birds are not attached to sensors. Among the existing periodicity mining methods, Discrete Fourier Transform (DFT) and autocorrelation are the most common types. Fourier Transform maps a function of time into a new function whose argument is frequency. In the case of a periodic function, DFT can be simplified to the calculation of a discrete set of complex amplitudes, called Fourier series coefficients. Given a sequence $x(n), n = 0, 1, ..., N - 1$, the normalized DFT is a sequence of complex numbers $X(f)$ is given by

$$X(f) = \frac{1}{\sqrt{N}} \Sigma_{n=0}^{N-1} x(n) e^{-\frac{i2\pi kn}{N}}$$

where the subscript k/N denotes the frequency that each coefficient captures. Vlachos et al. proposed a non-parametric approach to predict the periodicity by introducing a new periodic distance measure technique for time-series sequences [23]. Li et al. designed Fourier Transform based method to mine periodicity in spatiotemporal data of moving objects [24]. Yuan et al. proposed a Bayesian-based non-parametric approach to detect periodic patterns in spatiotemporal data of social media users [25]. For biological systems which are fast, ultradian, circadian, circalunar, and yearly time domains, determining periodicity can be troublesome. Amariei et al. proposed a Fourier-based approach that generates de-noised waveform from multiple frequencies [26]. Though we have discussed briefly periodicity and techniques, we haven't discussed time series yet (Fig. 3). Below we discuss time series data and its analysis.

In sectors like banking and finance, healthcare, meteorology, seismology, and even retails, we encounter a special type of data that is indexed in time order. This type of data is known as time series where the data points are sequenced taken at successively spaced points in time giving rise to a sequence of discrete-time data. Time series data have a natural temporal ordering and the analysis of these time series is significantly different from cross-sectional studies in which there is no natural ordering of the observation points.

A time series data can exhibit a repetitive pattern with elapsed time which is known as periodicity. For instance, data that has a sinusoidal character is periodic.

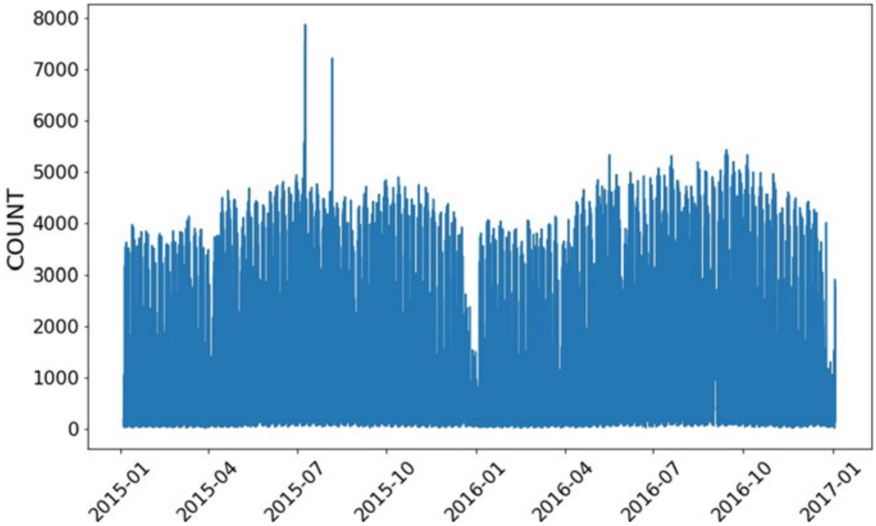

Fig. 3. An example of a time series plot

The period can vary from few minutes (as in the case of an electrocardiograph) to years (solar activity). However, the important point is that the pattern must repeat itself after the time period as shown in the Fig. 3.

For the sake of convenience, periodicity has been divided into two broad categories: (a) Seasonal and (b) Cyclical. Both are extremely important as vital insights can be obtained by analyzing the underlying periodicity. However, before diving into this, let us look at a much simpler and general concept of data called "***trend***". Almost all data exhibit a trend. A trend occurs if there is a tendency in the time series data to move upward over time or downward. Thus, we say that the data has a general tendency to either move up (increasing trend) or down (decreasing trend). However, the term "trend" is not well defined mathematically and may not give a lot of insight. On the other hand, periodicity is mathematically well defined. Seasonality occurs when there is a regular pattern in the data related to the calendar. A seasonal data can exhibit patterns which can either be daily, weekly, monthly, quarterly, or even annually. For example, sales of air conditioners or ice cream may peak every year during the summer months and this observation can be observed every year. Similarly, people who are prone to allergies may show allergic reactions with seasonal changes. Whenever the time series is influenced in a periodic manner by the calendar, we call such time series to have a seasonality. On the other hand, a cyclic pattern exists when the data exhibit highs and lows over a period of time that is not fixed. Figure 4 shows the plot corresponding to seasonality, upward and downward trend as well as a time series with none of these characteristics. The duration of these peaks and troughs may vary for years. For instance, business cycles may usually last several years, but in this case, the length of the current cycle is unknown beforehand.

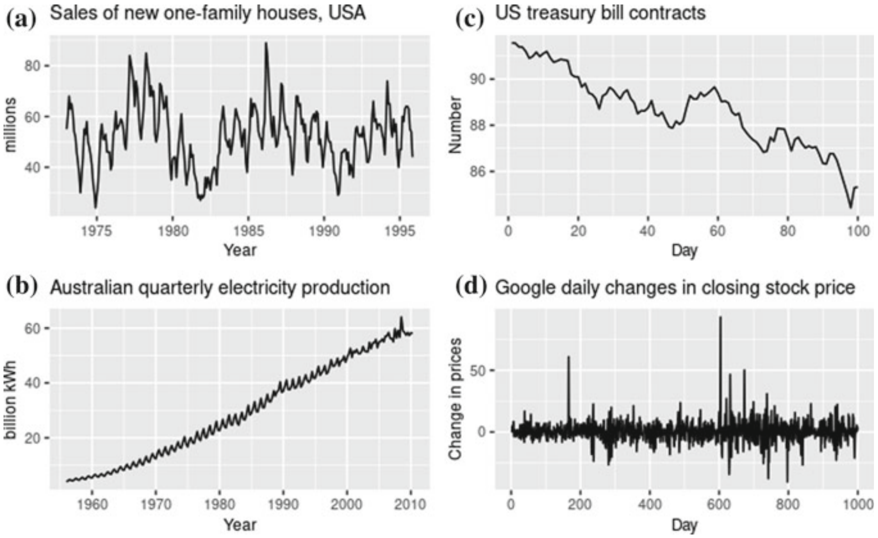

Fig. 4 Plots for **a** seasonality, **b** upward trend, **c** downward trend, and **d** time series with any trend or seasonality

Although the word seasonal and cyclic may be confusing at the first glance, there are important differences between them. While the seasonal pattern is for a constant length of time, a cyclic pattern occurs over a variable length of time. The average length of a cycle is longer than the length of a seasonal pattern. Moreover, the magnitude of a cyclic patternis more variable than the magnitude of a seasonal pattern. Finally, while the timing of peaks and troughs are predictable with seasonal data, it is much harder to predict a cyclic event.

In order to discover the periodicities in a time series data, it is possible to use the periodogram to estimate the spectral density of the signal. The periodogram P can be obtained by squaring the length of each Fourier coefficient as.

$$P\left(f_{\frac{k}{N}}\right) = \left|X_{\frac{k}{N}}\right|^2$$

for $k = 0, 1, 2 \ldots \frac{(N-1)}{2}$. At this point it is also imperative to briefly discuss the subtle difference between prediction and forecast. Prediction is everything about guessing a number (value) of something. This "*something*" can be anything ranging from the price of a house in a city as is done using the Boston Housing or Airbnb dataset to the label of a photo to label a photo as cat or dog etc. On the other hand, forecasting is a special type of prediction where the data has a temporal component. In forecasting, one looks at the past/historical data and makes a prediction for a future time. The weather for future days/weeks, sales of a product in upcoming days/months, the stock market in coming days can be clubbed under the category forecasting as the future is predicted using the historical data.

The statistical art of dealing with time series data is called time series analysis. There are several techniques that are bundled together in time series analysis. Next, we will briefly look into some of these classical techniques of time series analysis for forecasting followed by more modern approaches which are based on deep learning.

7.1 *Classical Time Series Forecasting*

The traditional techniques of time series forecasting include:

(i) Moving average(MA),
(ii) Autoregression (AR),
(iii) Autoregressive Moving Average (ARMA),
(iv) Autoregressive integrated moving average (ARIMA),
(v) Seasonal autoregressive integrated moving average (SARIMA),
(vi) Seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX),
(vii) Vector autoregression (VAR),
(viii) Vector autoregression moving average (VARMA),
(ix) Vector autoregression moving average with exogenous regressors (VARMAX),
(x) Simple exponential smoothing (SES),
(xi) Holt Winter's exponential smoothing (HWES).

Depending on the type of data, domain, and requirement, these abovementioned techniques are used. However, it is important to note that each of these models has their own limitations depending on the nature of the series, viz., seasonality, stationarity, etc.

To keep things simple, we will briefly outline only a few of these techniques without diving deep into the mathematical formulations. A detailed description of each of these techniques is beyond the scope of the present work. For details, we refer to the book by Brockwell and Davis on Time Series [27, 28].

(i) **Moving Average (MA)**.

A moving average or rolling average or running average is a way to analyze data points by creating a series of averages of different subsets of the full dataset. It is a type of finite impulse response filter and simple moving average, cumulative, weighted, and exponential moving average are different types used depending on the type of dataset.

(ii) Autoregressive Mode

As the name suggests, the autoregressive (AR) model is a regression model where the forecasting is performed using the historical data. The autoregressive model is an extremely simple yet effective model where the output or the forecasted value depends linearly on its own previous values. Mathematically speaking, the AR model of order p can be written as:

$$X_t = \beta_0 + \sum_{i=1}^{p} \alpha_i X_{t-1} + \varepsilon_t \tag{1}$$

where X_t and X_{t-1} are the observations at time *t* and *t-1*, respectively, *c* is a constant and ε_t is the white noise with mean zero and standard deviation …

A fundamental principle of autoregression is that the series has a correlation with a delayed (lagged) copy of itself. There exists two important correlations in this context : (a) Autocorrelation [29], and (b) Partial Autocorrelation [29]. Let us understand this by taking a simple example related to the sales of ice cream. Let us denote the volume of sales of ice cream as $X =\{X_1, X_2, \ldots\ldots, X_N\}$ for time stamps $\{t_1, t_2, \ldots., t_N\}$. It is possible that the volume of sales of any given month may have an effect on the volume of sales for the next month or the month later. In other words, we can think of having a relationship like $\{X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5 \rightarrow X_6 \ldots andsoon\}$, where the sales of the any given time point t is directly influenced by the sales of the previous time point $t - 1$. Additionally, we can have a scenario where the present month's sales are influenced by the sales from two time points earlier. For example, the sales at time point t_3 (X_3) can be influenced by the sales at time point $t_1(X_1)$. In this case, there are two ways the sales X_3 can get affected by the sales in X_1. One way is the direct influence that X_1 has on X_3. while the other is the indirect influence v X_2 denoted by $X_1 \rightarrow X_2 \rightarrow X_3$. The first case is what is known as autocorrelation or serial correlation while the latter is known as partial autocorrelation. The autocorrelation is a bit intuitive as it is mathematically given by the Pearson correlation function. If $\{X\}$ is a series of observable at time $\{t_i\}$, then the lag p autocorrelation function (ACF) is defined as

$$ACF = \frac{\sum_{i=1}^{N-k} \left(X_i - \underline{X}\right)\left(X_{i+p} - \underline{X}\right)}{\sum_{i=1}^{N} \left(X_i - \underline{X}\right)} \tag{2}$$

ACF is a correlation coefficient, but instead of being the correlation coefficient between two different variables, the correlation is calculated between two values of the same variable at times X_i and X_{i+k}. Hence, the name autocorrelation.

For the partial autocorrelation function $(PACF)$, we can write the linear regression expression

$$X_t = \Phi_{21} X_{i-1} + \Phi_{22} X_{t-2} + \varepsilon_t \tag{3}$$

where X_t is the value of time t while X_{t-1}, and X_{t-2} are the lagged values at time $t-1$ and $t-2$, respectively, and ε_t is the error term. The coefficient "Φ_{22}" is the $PACF$ in this case and it gives the effect of X_{t-2} on X_t. Overall, the $PACF$ help us to build a good time series forecasting model. It not only helps us to decide if the AR model is appropriate for the data, but also in deciding the order of the AR model.

To apply the AR model, it is critical for the time series data to be stationary. The stationarity condition is satisfied if the mean, the variance, the autocorrelation does not change over time and there are no periodic fluctuations (seasonality). If the series is non-stationary, then there will be a non-declining effect of the previous values on the current value. It simply means that the correlation with increasing lags should ideally decrease.

Going back to our ice cream sales data, the series will have non-stationarity if the sales at time point $t_6(X_6)$ is more influenced by the sales at time point $t_1(X_1)$ or $t_2(X_2)$.

(iii) Autoregressive Integrated Moving Average (ARIMA).

Consider the hypothetical sales data shown in Fig. 5. It shows a clear upward trend in sales with time. In strict mathematical terminology, this data is not stationary. Thus, if we want to predict the sales for the year 2020, we cannot apply the AR model as in this case the mean is not constant. To apply autoregression, the first step is to make the data stationary. Stationarity can be achieved by calculating the difference between successive time points. If the sales data time series $\{X\}$ mentioned earlier has stationarity, we can perform a transformation and define a new series Z_i such that

$$Z_i = X_{i+1} - X_i \tag{4}$$

where $\{Z_i\}$ satisfies the condition of stationarity. This is because if the original time series has linearity, then the difference between two such timestamps will be linear barring the fluctuations around a constant mean. Once we have the stationarity in our data, we can freely use the AR model with more confidence. We can combine the AR and MA models and create a new model called Autoregressive Integrated Moving Average (ARIMA) for forecasting. ARIMA models are characterized by three parameters, (p, d, q),where p is the lag of the AR model and q is the order of the MA model. The parameter d is the order of the difference which we perform to make the data stationary. Thus, we can define the ARIMA model with parameters (p, q, d)

$$Z_i = \Phi_1 Z_{i-1} + \theta_1 \varepsilon_{i-1} + \varepsilon_i \tag{5}$$

In Eq. 5, term $\Phi_1 Z_{i-1}$ is the AR part while the term $\theta_1 \varepsilon_{i-1}$ is the MA part. The final term ε_i is the error term. The part corresponding to "integrated" in ARIMA comes from the fact that Z_i was created by taking the differences of successive terms in the series $\{X\}$. Finally, we would like to predict X_i and not the difference Z_i. If

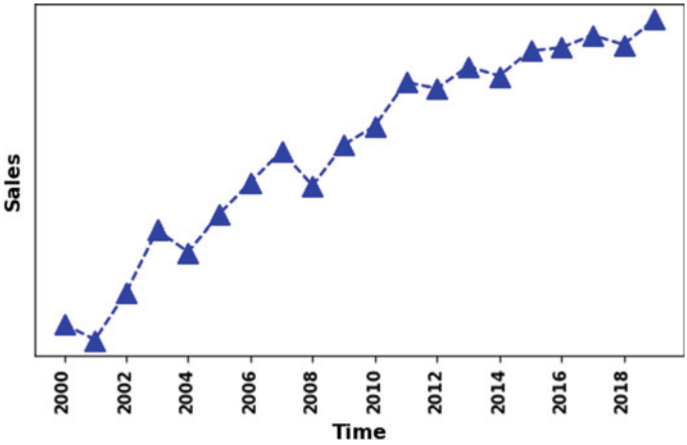

Fig. 5 Seasonality in sales data

X_l is the last term, in the series $\{X\}$,. then for a future time point τ, the value X_τ is given by

$$X_\tau = \sum_{i=1}^{k-l} Z_{k-i} + X_l \tag{6}$$

Finally, if we want to handle time series data with periodic characteristics, then we can use the SARIMA model which combines the seasonal differencing with the ARIMA model. SARIMA model is described by seven parameters $(p, d, q)(P, D, Q)m$.hile (p, d, m) are the same as the ARIMA model; the $(P, D, Q)m$ are the parameters that incorporate the seasonal elements.

7.2 Modern Techniques

Apart from the classical time series analysis, there are few modern recurrent neural network (RNN)-based techniques that have gained popularity in recent times. We will briefly discuss it here. RNN-based [30] methods are capable of identifying the nonlinear and complex structures and patterns of data in time series forecasting. One such RNN architecture is the Long Short-Term Memory (LSTM) [30] (Fig. 6).

LSTM networks have an internal memory (like RNN) making them efficient to remember past events/data. The vanishing gradient problem associated with RNN is resolved in LSTM. These features make LSTM well suited for classification and time series forecasting with time lags of unknown duration. Figure 6 shows the schematic diagram of LSTM.

Three gates constitute the LSTM network architecture: