Fault-Tolerant, Secure Large Scale Data Management in a New World

Amr El Abbadi
Department of Computer Science,
University of California, Santa Barbara

Abstract

Once upon a time data was structured, clean and resided on machines that were trustworthy and even when they failed, they simply crashed. This era has come and gone. Data is often unstructured, dirty and might even reside on untrusted infrastructure. The rise of the cloud computing paradigm as well as the rapid popularity of blockchains demand a rethinking of our naïve, comfortable beliefs in an ideal benign infrastructure. In the cloud, clients store their sensitive data in remote servers owned and operated by cloud providers. Current solutions face significant scalability and performance challenges when incorporated in large scale data repositories. Novel database designs need to directly address the natural tension between performance, fault-tolerance and trustworthiness. In this talk, I will discuss recent attempts for large scale data management in faulty and potentially untrusted settings, their limitations and potential approaches to mitigate these shortcomings.

Bio:
Amr El Abbadi is a Professor of Computer Science at the University of California, Santa Barbara. He received his B. Eng. from Alexandria University, Egypt, and his Ph.D. from Cornell University. His research interests are in the fields of fault-tolerant distributed systems and databases, focusing recently on Cloud data management and blockchain based systems. Prof. El Abbadi is an ACM Fellow, AAAS Fellow, and IEEE Fellow. He was Chair of the Computer Science Department at UCSB from 2007 to 2011. He has served as a journal editor for several database journals, including, The VLDB Journal, IEEE Transactions on Computers and The Computer Journal. He has been Program Chair for multiple database and distributed systems conferences. He currently serves on the executive committee of the IEEE Technical Committee on Data Engineering (TCDE) and was a board member of the VLDB Endowment from 2002 to 2008. In 2007, Prof. El Abbadi received the UCSB Senate Outstanding Mentorship Award for his excellence in mentoring graduate students. In 2013, his student, Sudipto Das received the SIGMOD Jim Gray Doctoral Dissertation Award. Prof. El Abbadi is also a co-recipient of the Test of Time Award at EDBT/ICDT 2015. He has published over 300 articles in databases and distributed systems and has supervised over 35 PhD students.

Identifying Tipping-points in Financial Markets using Data Science methods

Anirban Chakraborti
School of Computational & Integrative Sciences,
Jawaharlal Nehru University, New Delhi - 110067, India
Email: anirban@jnu.ac.in
URL: http://www.jnu.ac.in/Faculty/anirban/index.html

Tipping-points or critical points, which are usually unpredictable, are ubiquitous in complex systems found in nature and society – ranging from natural hazards, such as earthquakes, volcanic eruptions, hurricanes, extreme weather conditions due to global warming, failure of physical and social structures due to large scale terror strikes, market crashes and economic slowdowns, electric grid failures, traffic breakdowns, disease and epidemic spreads, etc. Such tipping-points often reveal underlying dynamical processes and thus provide ground for better scientific understanding of complex systems like stock markets, fractures or earthquakes, climate, etc. Recently, we have focused our attention on identifying generic indicators that may detect if a complex system like the financial market is close to a tipping-point, using the methods of data science. Here, a brief overview of our recent results will be presented.

References

1. Chakraborti, Anirban, Hrishidev, Kiran Sharma, and Hirdesh K. Pharasi. "Eigen-entropy measure to study phase separation in market behavior." *arXiv preprint arXiv:1910.06242* (2019).
2. Pharasi, Hirdesh K., et al, "Complex market dynamics in the light of random matrix theory." *New Perspectives and Challenges in Econophysics and Sociophysics* (Springer, Cham, 2019) pp 13-34.
3. Pharasi, Hirdesh K., et al. "Identifying long-term precursors of financial market crashes using correlation patterns." *New Journal of Physics* 20.10 (2018): 103041.
4. Chakraborti, Anirban, et al. "Characterization of catastrophic instabilities in complex systems using emerging spectra" (2019) *https://www.researchgate.net/publication/336639750_Characterization_of_catastrophic_instabilities_in_complex_systems_using_emerging_spectra*.
5. Halder, Rajdeep, et al. “Risk estimation using topological information of the market.” *Manuscript in preparation* (2020).

Solving PDEs using Machine Learning

Balaji Srinivasan
IIT-Madras
Dept of Mechanical Engineering,

Abstract

The development of numerical methods for partial differential equations is now a mature subject. Algorithms for several applications have now been understood to the level that commercial solvers regularly utilize canned algorithms. Nonetheless, as applications become more complex and data becomes abundant, important avenues of study in computing differential equations

In this talk, we will discuss the implications that the rapid development of Machine Learning has on the computational solution of problems governed by differential equations. In particular, we will discuss some developments in using techniques from various Machine Learning algorithms towards automating various stages of mathematical modeling – from automated modeling [1], through reusing old solutions [2] to developing automatic numerical discretizarions [3,4]. We will also discuss the future developments possible in this field and their implications

References

1. Ling, Julia, Reese Jones, and Jeremy Templeton. "Machine learning strategies for systems with invariance properties." *Journal of Computational Physics* 318 (2016): 22-35.
2. Guo, Xiaoxiao, Wei Li, and Francesco Iorio. "Convolutional neural networks for steady flow approximation." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
3. Raissi, Maziar, Paris Perdikaris, and George E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations." *Journal of Computational Physics* 378 (2019): 686-707.
4. Dwivedi, Vikas and Srinivasan, Balaji. "Physics Informed Extreme Learning Machine (PIELM)–A rapid method for the numerical solution of partial differential equations", *Neurocomputing* (2019): https://doi.org/10.1016/j.neucom.2019.12.099

Data to Discovery in Information Security

N Balakrishnan
Honorary Professor
Indian Institute of Science
Bangalore 560012

Abstract

The major focus of this talk is to show the use of Artificial Intelligence (AI) and Machine Learning (ML) in discovering from the ever increasing quantity of data both in computer networks and in social media usage to protect networks and to predict useful user behavior in mitigating covert use of the social media.

In the first part of the talk, we focus on our two decades of research in the use of ML in handling large volumes of network data to enhance the performance of Intrusion Detection Systems, detecting structured botnets at infrastructure level traffic, malware detection and classification for offline and for online streaming data.

In the second part of the talk, we focus on topics such as Influence maximization and mobility and churn prediction for a large population of a city which depend on large volumes of data from social media and Call Data Records (CDR). These are of importance in smart cities, smart businesses and smart economy.

We conclude the talk by presenting the future opportunities in big data and AI.

Influence maximization in unknown social networks: Learning Policies for Effective Graph Sampling

Balaraman Ravindran
IIT Madras

Abstract:

A serious challenge when finding influential actors in real-world social networks, to enable efficient community-wide interventions, is the lack of knowledge about the structure of the underlying network. Current state-of-the-art methods rely on hand-crafted sampling algorithms; these methods sample nodes and their neighbours in a carefully constructed order and choose opinion leaders from this discovered network to maximize influence spread in the (unknown) complete network.

In this work, we propose a reinforcement learning framework to discover effective network sampling heuristics by leveraging automatically learnt node and graph representations that encode important structural properties of the network. At training time, the method identifies portions of the network such that the nodes selected from this sampled subgraph can effectively influence nodes in the complete network. The output of this training is a transferable, adaptive policy that identifies an effective sequence of nodes to query on unseen graphs. The success of this policy is underpinned by a set of careful choices for embedding local and global information about the graph, and providing appropriate reward signals during training. We experiment with real-world social networks from four different domains and show that the policies learned by our RL agent provide a 7-23% improvement over the current state-of-the-art method.

Human Centric AI in Autonomous Driving

Bhanu Prakash
Head of Advanced Engineering,
ADAS India at Continental AG

Abstract:

1.2 million People are killed on the world's roads every year. In America alone, 33,000 people are killed each year. To put that in perspective, that's the same as a 747 falling out of the sky every working day. Autonomous Driving (AD) is the solution to reduce fatalities. It has transcended far beyond being a crazy moonshot idea over the last half-decade or so. It is quickly becoming the biggest technology today that promises to shape our tomorrow. This talk would be covering topics on what is AD and how AD is achieved. A brief overview of how Continental is embracing AI to solve problems in AD. How human centered AI is helping to better the performance of AI where safety is critical. Finally, the challenges and pitfalls of AI will be discussed.

About Speaker:

Bhanu prakash has got 19+ years of rich experience in SW Development for automotive and smart Phone products. He is also championing Innovation Programs inside the organization. He is currently the Head of Advanced Engineering, ADAS India at Continental AG, where he is responsible for deploying high performing algorithms (classic & AI) into their Sensors : RADAR, Camera & Fusion.

Prior to joining continental in 2015, he worked at SAMSUNG R&D for more than 12 years in various technical leadership roles in areas of Multimedia DSP, 3D Graphics, IP design. Bhanu Prakash is the inventor of more than 7 patents and patent applications relating to Touch sensing & Driver safety & Autonomous Driving.

Bhanu Prakash holds MTech degree in Communication systems from PSG College of technology, Coimbatore.

Classification of Similar Foodborne Pathogens using Deep Convolutional Neural Network for Raman Spectroscopic data.

Dipak Bhikaji Kumbhar[1], Dhanya Reghu[1], Divya Shrungar[1], Sanchita Sil[2], Siva Umapathy[1,3*]

[1] Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India

[2] Defense Bioengineering and Electromedical Laboratory (DEBEL), Bangalore 560093, India

[3] Indian Institute of Science Education and Research, Bhopal,462066, India

Abstract:

Raman Spectroscopy is a vibrational technique which is widely used in identification and classification of materials and biological samples etc. The accuracy of the classification model increases when the analysis methods includes the concepts of Artificial Intelligence (AI). Most commonly used classification algorithms in Raman spectroscopy are Linear Discriminant Analysis (LDA), Principal component analysis (PCA), Support Vector Machines (SVM) etc. However, all these algorithms need feature extraction to be done manually and are limited by the accuracy, speed and performance. Use of Convolutional Neural Network (CNN) in Raman spectroscopic data analysis automates the feature extraction and improves the classification accuracy. CNN is a bottom- up hierarchical multilayer perceptron that performs pre-processing, feature extraction and classification. Because of this characteristic, CNN has already found its place in Raman spectroscopy for identification and regression and is rapidly dominating other state-of-the-art machine learning methods. We have used deep learning technique to discriminate spectral data of similar type bacteria and it showed very promising result.

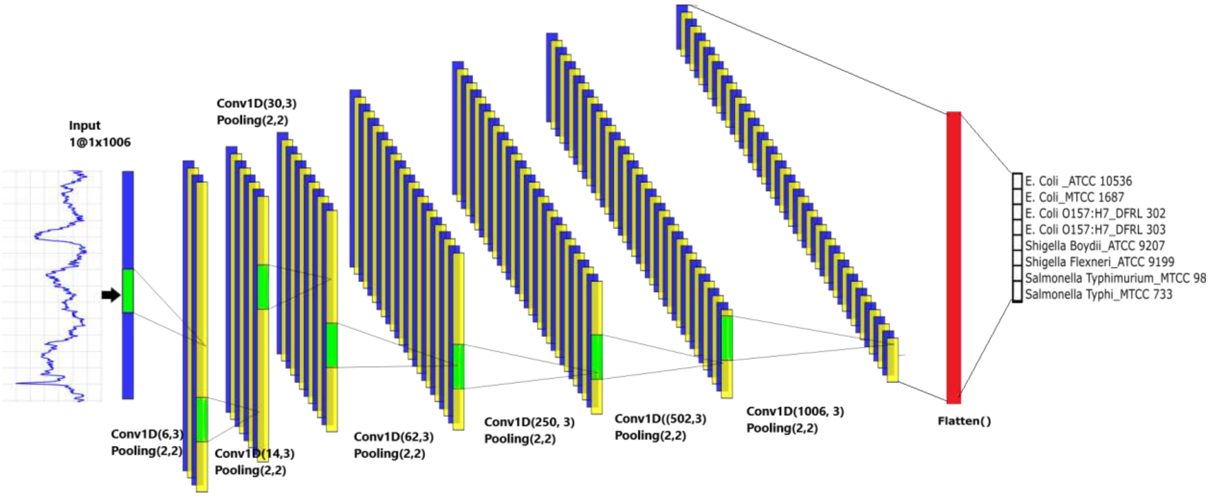

Figure: Schematic of Deep CNN with various convolution layers for feature extraction from spectral data of similar foodborne pathogens.

References:

[1] R. Gautam, S. Vanga, F. Ariese and S. Umapathy. EPJ Techniques and Instrumentation (2015) 2:8.
[2] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, S. J. Gibson, Analyst (2017), 142, 4067−4074.
[3] Acta Geodyn. Geomater., Vol. 8, No. 1 (161), 47–58, (2011).
[4] Yessi Jusman, Siew Cheok Ng, and Noor Azuan Abu Osman ScientificWorldJournal. (2014); 2014: 810368.
[5] A. Krizhevsky, I. Sutskever and G. E. Hinton, Advances in neural information processing systems, (2012), pp. 1097–1105.
[6] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), (2006).

Visualizing Uncertainty In Monte Carlo Experiments

Dootika Vats
IIT Kanpur

Abstract

In Monte Carlo simulations, samples are obtained from a target distribution in order to estimate various features. We present a flexible class of visualizations for assessing the quality of estimation, which are principled, practical, and easy to implement. To this end, we establish joint asymptotic normality for any finite collection of means and quantiles. Using the limiting distribution, we construct $1-\alpha$ level simultaneous confidence intervals, which we integrate within visualization plots. We demonstrate the utility of our visualizations in various Monte Carlo simulation settings including Monte Carlo estimation of expectations and quantiles, Monte Carlo simulation studies, and Bayesian analyses using Markov chain Monte Carlo sampling. The marginal-friendly interpretation enables practitioners to visualize simultaneous uncertainty, a substantial improvement from the current visualizations available in software.

A Generalized Deep Learning Framework for Whole-Slide Histopathology Image Segmentation and Analysis

Ganapathy Krishnamurthi

Department of Engineering Design, IIT-Madras

Abstract:
Histopathology tissue analysis is considered the gold standard in cancer diagnosis and prognosis. Given the large size of these images and the increase in the number of potential cancer cases, an automated solution as an aid to histopathologists is highly desirable. In the recent past, deep learning-based techniques have provided state of the art results in a wide variety of image analysis tasks, including analysis of digitized slides. However, the size of images and variability in histopathology tasks makes it a challenge to develop an integrated framework for histopathology image analysis. We propose a deep learning-based framework for histopathology tissue analysis. We demonstrate the generalizability of our framework, including training and inference, on several open-source datasets, which include CAMELYON (breast cancer metastasis), DigestPath (colon cancer), and PAIP (liver cancer) datasets. We discuss multiple types of uncertainties pertaining to data and model. Simultaneously, we demonstrate our model generalization across different data distribution by evaluating some samples on TCGA data. Our entire framework and related documentation are freely available at GitHub and PyPi.

Data-Intense Computational Plasma Physics

Rajaraman Ganesh

Institute for Plasma Research, Gandhinagar, Gujarat, INDIA

Powered by tremendous growth in accessible compute-power as well as due to the very nature of the physics problems which demand computational resources due to their physical complexity, discoveries via computer simulations has been on fast track. However, due to the very size of the data generated, analyzing and obtaining meaningful information has become a challenge, which I believe has impeded the growth in discoveries via scientific computing in plasma physics.

In this talk, I shall present some simple examples of event-rich data generated from computer simulations of novel plasma phenomena using kinetic, fluid, particle-in-cell and molecular dynamic simulations with a hope to stimulate discussions on applications of machine learning methods and big-data analysis to address problems in data-intense computational plasma physics.

Title: Data Science and Astrology: Is There A Difference?

Speaker: Jayant Haritsa
Dept. of Computational & Data Sciences
Indian Institute of Science, Bangalore

Abstract: Data Science has become the buzzword of choice in recent times, especially in the software industry. The accompanying hoopla has spawned frenetic claims foretelling the development of great and wondrous solutions to all kinds of contemporary local and global challenges. In this talk, we will attempt to provide a more nuanced and balanced view of the area. Specifically, we will make the case that data science is certainly a potent force-multiplier for solving challenging problems, but is equally subject to methodological abuse, leading to grossly erroneous conclusions. Worse, such misuse could also be deliberate to suit preconceived outcomes and biases, concealed under a scientific veneer. In a nutshell, we will advocate the view that data science should be a tool of last resort, not first, and leveraged as a support, not substitute, for domain expertise.

Bio: Jayant Haritsa is on the faculty of the Department of Computational & Data Sciences at the Indian Institute of Science, Bangalore, since 1993. He received a BTech degree from the Indian Institute of Technology (Madras), and MS and PhD degrees from the University of Wisconsin (Madison). He is a Fellow of ACM and IEEE, and a recipient of the Swarnajayanti Fellowship, the Shanti Swarup Bhatnagar Award, and the Infosys Prize.

Photo:

Re-imagining agri-businesses through emerging technologies

Jitendra Singh, Research Scientist, IBM Research -India

Agri-businesses (farming, trading, food supply chain, insurance, finance, retail, etc) are undergoing massive transformation worldwide, thanks to advances in technologies such as AI, Remote Sensing, Block Chain, Weather Predictions, etc. For example, multi-sensor high-spatial/temporal resolution satellite data is now much more easily and quickly available, the fidelity of weather predictions has increased drastically for any location on the globe, advances in DL/ML algorithms and platforms now enable extracting insights from humongous satellite data sets at a field level and Block Chain based platforms are enabling trust, provenance and have the potential to reduce food wastage across end-to-end food supply chains. IBM Research is developing and validating solutions relevant to the above-mentioned businesses with a focus on scalability to different geographies. In this talk, I will present scientific and technical challenges and meaningful results on uses-cases such as crop monitoring (crop health, acreage, type, yield), soil moisture estimation etc. Furthermore, I will share the learnings that we have gained so far working with clients and growers and outline challenges that lies ahead.

THE FOURTH PARADIGM CONFERENCE:
From Data to Discovery,
27th-30th January 2020 @ IISER Bhopal

Observing the changing Atmospheric Composition in the Anthropocene from space and from aircraft.

John P. Burrows
Institute of Environmental Physics, University of Bremen
P.O, Box 330440, 28334 Bremen, Germany

Prior to the Neolithic Revolution around 12 000 Yr BP, the Earth's population was around 4 million hunter gatherers. Population grew to around 1 Billion at the beginning of the Industrial Revolution in the UK (1750-1800). Thereafter, its growth began to accelerate and population is now over 7.6 Billion with more than 50% living in megacities or urban agglomerations, known as Major Population Centres. The energy to fuel this growth has been in large produced from fossil fuel combustion. It has resulted in improved longevity, a rapidly increasing standard of living, supplied by an evolving industrial sectors. However, this has resulted in ever increasing and evolving emissions of pollutants (both short lived climate pollutants and long lived greenhouse gases) and global land use change This is impacting on air and water quality, human health, ecosystem services and climate change. Pollution extends form the local to global scale. The world has now entered a new geological epoch the Anthropocene. To assess accurately the impact of man in the evolving Anthropocene both space based an airborne measurement of atmospheric composition are a pre requisite.

The SCIAMACHY (SCanning Imaging Absorption spectrometer for Atmospheric CHartographY) project was proposed in 1988 to meet this need. SCIAMACHY flew on the ESA Envisat (2002 to 2012). SCIAMACHY, a passive remote sensing double monochrometer, measured the up-welling radiation from the top of the atmosphere. The GOME (Global Ozone Monitoring Experiment) was a smaller version of SCIAMACHY, on ESA ERS-2 (1995 to 2011) viewing only in nadir. GOME-2 is a somewhat improved version of GOME, aboard the ESA/EUMESAT. SCIAMACHY, GOME, and GOME-2 are best known for the retrievals of trace gases in the troposphere and the stratosphere, where measurements of the total and tropospheric columns of the trace gases O_3, NO_2, HCHO, CHO.CHO, BrO, IO, H_2O, CO as well ocean colour, cloud and aerosol optical properties. In addition, SCIAMACHY observations made the first dry column mixing ratios of CO, CH_4 and CO_2.

In addition to space measurements, aircraft are needed to provide high spatial resolution measurements of trace atmospheric constituents to test our understanding of the transport and transformation of atmospheric emissions. EMeRGe (Effect of MEgacities on the Transport and Transformation of Pollutants on the Regional to Global Scales) was proposed to exploit the capabilities of the German HALO research aircraft to study the outflow from MPC. This presentation will provide highlights from both spaced based remote sensing and provisional observations from the EMeRGe campaigns.

Probabilistic modeling to detect heterogeneity in genomic data

Leelavati Narlikar
CSIR National Chemical Laboratory, Pune

Abstract:

Machine learning algorithms are increasingly being used to gain insights into molecular biology. I will talk about one example of using Bayesian mixture modeling to identify diverse regulatory mechanisms hidden in datasets resulting from high-throughput sequencing experiments. Although our algorithm may seem specific to Biology, it is in fact very general, based on standard statistical principles, and therefore can be applied to problems in other domains as well.

Convergenetix: Data, Computing, Technology and Innovation Jointly Advancing Biomedical Science

Madhavi K. Ganapathiraju, Associate Professor, School of Medicine and School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA madhavi@pitt.edu

Abstract:

Biotechnology and lab automation have enabled characterization of unprecedented amounts of data related to biological organisms, including genetic fingerprinting of individuals at population scale. In a parallel development, hospitals are adopting digitization of electronic health records and are moving towards adopting interoperability standards that allow collection of combined bio (genomic, proteomic, trascriptomic) and medical data. Where there is data, there is scope for discovery with innovative computational methods. This talk presents an introduction to interesting problems in biomedical domains that are amenable to computational solutions. We have looked at the biomedical questions by drawing parallels to problems addressed in language technologies and network analyses. We have studied n-gram language modeling to detect genetic signatures and repetitive regions, latent semantic analysis to identify transmembrane helices in proteins, and social network analysis to predict edges in the human protein interactome. These approaches have been translated to experimentally validated results of biomedical significance. Specifically, they led to the discovery of the protein that activates immunity upon viral infection, identified two drugs, which treat allergies and motion sickness respectively, to be potentially repurposable for treating schizophrenia and are under clinical trials, and identified network modules that point to the role of specific cellular organelles in a complex disease called congenital heart disease. While the application domain is biology, the presentation will be suitable for computer scientists and engineers without background in biology.

Speaker's Biography:

Madhavi Ganapathiraju received Ph.D. from Carnegie Mellon University School of Computer Science in Language and Information Technologies and M.Engg. degree from Indian Institute of Science in Electrical Communications Engineering. She is Associate Professor at the Department of Biomedical Informatics of University of Pittsburgh School of Medicine and at Intelligent Systems Program of University of Pittsburgh School of Computing and Information. She is adjunct faculty at Language Technologies Institute of Carnegie Mellon University School of Computer Science, faculty of Biomedical Informatics Training Program (www.dbmi.pitt.edu), Intelligent Systems Program (www.isp.pitt.edu), Joint CMU-Pitt Computational Biology Program (www.compbio.cmu.edu) and Integrated Systems Biology Program (www.isb.pitt.edu) of University of Pittsburgh. Significant research presented in this talk was funded by the Biobehavioral Research Awards for Innovative New Scientists (BRAINS) grant from National Institute of Mental Health of the National Institutes of Health, USA. Her publications may be viewed at http://tinyurl.com/MadhaviCMUscholar.

Accelerating Applied Sciences using Machine Learning

Dr. Manish Gupta
Director, Google Research India

Abstract: While numerical simulations have been extensively used to support scientific discoveries, for many problems, accurate simulations are still computationally infeasible. We describe several problems, including drug discovery, climate modeling and fluid dynamics, where deep learning techniques have been recently applied to achieve high fidelity simulations with currently available computational hardware. We also describe new proposed directions towards transforming medical science and healthcare using machine learning.

Bio: Dr. Manish Gupta is the Director of Google Research India and the Infosys Foundation Chair Professor at IIIT Bangalore. Previously, Manish has led VideoKen, a video technology startup, and the research centers for Xerox and IBM in India. As a Senior Manager at the IBM T.J. Watson Research Center in Yorktown Heights, New York, Manish led the team developing system software for the Blue Gene/L supercomputer. IBM was awarded a National Medal of Technology and Innovation for Blue Gene by US President Barack Obama in 2009. Manish holds a Ph.D. in Computer Science from the University of Illinois at Urbana Champaign. He has co-authored about 75 papers, with more than 7,000 citations in Google Scholar (and an h-index of 45), and has been granted 19 US patents. While at IBM, Manish received two Outstanding Technical Achievement Awards, an Outstanding Innovation Award and the Lou Gerstner Team Award for Client Excellence. Manish is a Fellow of ACM and the Indian National Academy of Engineering, and a recipient of a Distinguished Alumnus Award from IIT Delhi.

Data Driven Applications and Usage of Artificial Intelligence

Priyank Saxena
Co-founder & CEO, Infraeyes, Bengaluru.

Abstract

Data driven decision making is a process that involves collecting data based on measurable goals, analyzing patterns, and utilizing them to develop solutions. There is far more data being generated today than humans can analyze in any meaningful way. Artificial Intelligence along with big data technologies and high-performance computing is creating new opportunities for data intensive science. The combination has found some very interesting and useful applications in Medical Sciences, Agriculture, Wild Life Conservation, Space and even Sports. In this session, we will look at some practical applications where Artificial Intelligence is used to solve live problems.

The fourth paradigm for smarter cities in India

This talk explores the fourth paradigm in the context of smarter cities and describes an effort in this area being pursued by the Rober Bosch Centre for Cyber-Physical Systems, IISc, Bangalore. This effort is focusing on creating a market place of data and knowledge sources (raw, processed, static and dynamic, coming from sensors or via crowdsourcing), which can be accessed (subject to governance rules) by other public and private players to build various types of services and applications for the stakeholders (such as pollution monitoring and bus tracking), facilitate planning activities (such as traffic regulation and emergency response), generate insights for better governance, and even create new (processed) data sources to be consumed by others. The talk will describe examples of applications involving machine learning and video analytics in the areas of pollution monitoring, bus arrival time prediction, and traffic management. The initial experience shows that in emerging markets such as India, many of the assumptions that are made on the nature and quality of data, as well as affordability and availability of infrastructure, are not valid. As a result, we may need to approach the problems using a hybrid of paradigms.

Raghu Krishnapuram

Raghu Krishnapuram is currently a distinguished member of technical staff at the Robert Bosch Centre for Cyber-Physical Systems, IISc, Bangalore. He received his B.Tech. degree from IIT-Bombay in 1978, his M.S. degree from Louisiana State University, in 1984, and his PhD in computer engineering from Carnegie Mellon University in 1987. He worked in the academia in the US until the year 2000, initially at the University of Missouri, Columbia, and later at Colorado School of Mines, Golden. From 2000 to 2013, he held various technical leadership positions at IBM Research India, where he led projects in the area of “Knowledge, Information, and Smarter Planet Solutions”, with a particular focus on emerging markets. During 2014-15, Raghu worked at IBM T J Watson Center, Yorktown Heights, New York, where he was a technical leader for cognitive computing research. Raghu was Program Manager, Financial Services, Xerox Research Centre – India, during 2015-16, and was Head, R&D and IP Cell, at Ramaiah Institute of Technology, Bangalore during 2016-2017.

AI/ML in Weather and Climate Modelling and Forecasting

A. K. sahai, R. Chattopadhyay, S. Joseph and R. Mandal

Indian Institute of Tropical Meteorology, Pune

Abstract

Climate science is being increasingly acquiring huge and abundant data generated by Earth-observing satellites and also very high resolution climate models are producing tremendous amount of data, researchers are increasingly focusing on new tools to understand the current and future state of the climate. Therefore, Artificial Intelligence/ Machine Learning algorithm of data analysis in climate science is getting momentum. The application of artificial intelligence/machine learning (AI/ML) based techniques in climate sciences has been attempted in numerous studies focusing on two broad aspects: (a) classification and clustering of data and subsequent understanding in terms of pattern recognition problems and (b) application of the AI/ML techniques in time series prediction. Several studies have attempted a combination of both the above approaches. Another approach is to use the AI/ML-based technique to improve the output generated from dynamical models.

The life cycle of summer monsoon season over Indian region and its forecast is an area of active research since the last hundred years in the field of climate science. The forecast of rainfall during the monsoon season with lead time for days to month is a very challenging problem owing to the stochastic nature of the system. The understanding of Indian Summer Monsoon has been developed tremendously in the last few decades owing to the large availability of meteorological data both from observation (satellite, ground-based observations) and climate models. The AI/ML-based tools have been used to improve the understanding of the monsoon life cycle as well as its forecast in different lead-time scale.

In the lecture, I will describe the application of a clustering technique (Self Organizing Map or SOM) to identify the dominant patterns of the Indian summer monsoon, its nature of variability and its forecast for to 2-3 weeks lead-time. The first part in the forecast problem is the understanding of large scale patterns in which SOM is used as a pattern recognition tool. In the second part, we use an analog-based prediction technique to forecast the area-averaged rainfall over central India. Also, a probabilistic version of the forecast model is also developed which provides improved information over the deterministic forecasts.

The SOM based technique can also be used to improve the raw climate model output which is used by operational forecasters. The classic downscaling (also called the super-resolution problem in AI/ML parlance) problem of climate forecast is shown as an example in which model the AI/ML-based post-processed operational forecast shows huge improvements as compared to the “raw” model forecasts.

Inferring the "Laws" of Finance from Data

Sitabhra Sinha
The Institute of Mathematical Sciences, Chennai

Abstract:

A distinction often made between physical and social sciences is that the latter are not as amenable as the former to controlled experiments allowing for rigorous verification of theoretical predictions. However, astronomy is a hard science despite being almost exclusively based on observations, as it is almost impossible to do controlled experiments with planets, stars and galaxies. Just as collection and analysis of high-quality data from the period of Tycho Brahe led to the formulation of empirical laws by Kepler and was later followed by the theoretical groundwork of Newton that precisely explains the motion of planetary bodies, it is possible that the use of big data, especially from financial markets, will eventually lead to a well-established set of "laws" of economic activity. This talk will describe the first steps in this direction, focusing on how analysis of price and transaction data from financial markets (including bitcoin as well as more traditional assets such as currencies and equities) suggests the existence of empirical regularities ("stylized facts") that may be universal across space, time and asset classes.In particular, we shall discuss the heavy-tailed distribution of asset price fluctuations (the "inverse cubic/square law"), use of cross-correlation matrix spectral statistics to reconstruct the network relating price movements of different assets, and how the departure of such a network from structural balance can warn us of an impending systemic crises in the economy, such as the 1929 Great Depression or the 2008 Great Recession.

Artificial Intelligence to Biological Intelligence: Data-driven models for cognitive science

Sumeet Agarwal
IIT Delhi

AI and machine learning techniques, in combination with large-scale neural and behavioural data sets, are increasingly being deployed to try and better understand the workings of our own minds across a range of cognitive tasks. Here I will present an overview of some of our recent work in this area, touching upon examples from both the language and vision domains, and seeking to bring out the kinds of cognitive hypotheses we can currently examine via such data-driven models, as well as directions for future investigation.

Predicting Weather-Driven Impacts in the Current and Future Climate

Lloyd Treinish
IBM Distinguished Engineer
Chief Scientist – Environmental Modelling, Climate, Weather and Deep Thunder
Head, Atmospheric Science
IBM Thomas J. Watson Research Center
Yorktown Heights, NY, USA

Weather profoundly impacts the global economy and society at large. Not a week goes by when we don't learn of problems that affect our society and economy connected to severe weather, water or changes in our planet's climate. It can be the latest water crises in India, wildfires in Australia or unexpected impact from extreme precipitation. Such situations stress the resiliency of the infrastructure on which our lives and livelihoods depend. Despite the growing social unrest and economic disruption from such events that are expected to become more frequent, one must begin by understanding the underlying science and what is driving what we are seeing.

Efficient, resilient, and safe operation of many cities, for instance, is dependent on the local weather conditions at the scale of their critical infrastructure (electric, communications and water utilities, transportation, etc.) This includes both routine and severe weather events such as tropical storms, tornadoes, snowstorms, damaging winds and hail. With precipitation events, local topography and weather influence water runoff and infiltration, which directly affect flooding as well as drinking water quality and availability. For agricultural concerns, those same conditions as well as drought influence the effectiveness of irrigation and ultimately crop yield, or the environmental impact on nearby land and water bodies, among other farm operations.

Until recently, many businesses and public institutions largely accepted weather risk as a factor that could not be mitigated. However, with the increasing skill of weather forecasts, weather risk mitigation is a reality. This is especially true if the lead time for the availability of such predictions enables them to be actionable. In other words, it can support proactive allocation and deployment of resources (people and equipment) to minimize time for restoration of damage from severe events, for example.

The initial step to address this gap is the application of state-of-the-art (physical) weather models at the spatial scale of business operations. The results of such a model can then be coupled to additional models that predict the weather (business) impacts via machine learning techniques. While scientific and technical challenges remain to apply these methods, they are sufficiently mature that they are being used for operational decision making in many industries.

I will introduce the science and how we apply it in practice to address different business use cases. I will also discuss our on-going research to further such capabilities, including how we are extending this concept to a climate scale in order to evaluate the potential localized impacts of a warming planet and the effectiveness of strategies being used to mitigate such impacts.

Mathematics for Bigdata

S. Lakshmivarahan
School of Computer Science
University of Oklahoma, Norman, OK-73019, USA

Abstract: This talk will provide a brief overview of the historical evolution of Data Analytics and provide many examples of problem areas where different forms of bigdata – time series, configuration vs. similarity matrices and spatio-temporal type, arise naturally. We provide examples of paradoxes related to volume, distances and measures in high dimensional spaces which in turn provides an impetus for the development of dimension reduction (DR) techniques. DR techniques can be classified at least along two dimensions – deterministic vs. randomized and linear vs. nonlinear. Examples of applications of these techniques will be presented.

From Data Science to Precision Medicine

Vijay Chandru
Co-Founder and Chairman, Strand Life Sciences

Abstract:
The paradigm in modern healthcare is changing from "Population Medicine" to "Precision medicine". Population health assumes that the exact same solution will apply, without discrimination to all people. We now know for sure that this isn't true. While the medical community has accepted precision in more complex diseases like cancer - individual variations can alter treatment course in even much simpler conditions. People differ from each other due to a whole host of reasons - they could be genetic, lifestyle or environmental. It is therefore important for all these to be considered while making a treatment choice.

The decision process in precision medicine has 3 steps: (i) understanding the healthcare risks of the individual, (ii) measuring and monitoring the risk, and (iii) taking the right clinical and non-clinical decisions to help reduce the risk. We will see that it takes a diverse set of capabilities and experience to help in the precision medicine decision making process. The experience in diagnostics with genomics, the comprehensive range of tests, lab operating skills and quality mindset are important to measure and monitor individual risks accurately. Finally, expertise in data analytics, technology and AI is needed to detect patterns and incorporate them into decision making paradigms.

Machine Learning for Software Requirements

Arpit Sharma,
Department of Electrical Engineering and Computer Science
Indian Institute of Science Education and Research Bhopal, India
Email: arpit@iiserb.ac.in

Abstract: Requirements are the basis for every software project, defining what the stakeholders in a potential new system need from it, and also what the system must do in order to satisfy that need. All subsequent steps in software development are influenced by the requirements. Hence, improving the quality of requirements means improving the overall quality of the software product. A major cause of poor-quality requirements is that the stakeholders involved in the development process have different interpretations of technical terms. In order to avoid these issues and to improve the understandability of requirements, it is necessary that all stakeholders of the development process share the same understanding of the terminology used. Specialized terms used in the requirements document should therefore be defined in a glossary. A glossary defines specialized or technical terms and abbreviations which are specific to an application domain. Manual extraction of glossary terms from a large body of requirements is an expensive and time-consuming task.

In this talk we will discuss a fundamentally new approach for automated extraction of glossary terms from large-sized crowd-based requirements documents using machine learning. In particular, we will focus on using the state-of-the art neural word embeddings technique for creating a semantic filter for glossary term extraction. Additionally, we also briefly discuss some other interesting applications of machine learning for improving the quality of software requirements.

References:

1. Siba Mishra & Arpit Sharma. Automatic Word Embeddings-Based Glossary Term Extraction from Large-Sized Software Requirements. Proceedings of the 26th International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ). LNCS Springer (2020)
2. Kushagra Bhatia, Siba Mishra & Arpit Sharma. Clustering Glossary Terms Extracted from Large-Sized Software Requirements using FastText. Proceedings of the 13th Innovations in Software Engineering Conference (ISEC). ACM Digital Library (2020)
3. Siba Mishra & Arpit Sharma. On the Use of Word Embeddings for Identifying Domain Specific Ambiguities in Requirements. 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE). Proceedings of the 27th International Requirements Engineering Conference Workshops (REW). IEEE Computer Society: 234-240 (2019)

Fake News Detection

Kushal Shah

Department of Electrical Engineering and Computer Science, IISER Bhopal.

Email : kushals@iiserb.ac.in

Abstract

There can be two types of fake news, one which is factually incorrect, and the other which is factual, but presented in a manipulative manner and thus, representing fictional style of writing. In our analysis, we are dealing with the latter, which is essentially an opinion disguised as news, and is written using lot of descriptive words in order to evoke an emotional response in the reader. An algorithm for detecting this kind of fake news has been developed by our research group at IISER Bhopal [1, 2, 3].

This algorithm takes an English text (in our case, a news article) as input and, using Logistic Regression, outputs the probability of the text being written in a manipulative (fiction) or straight forward (non-fiction) manner. From this earlier analysis, it was found that fiction texts tend to have a higher value for the ratio of number adverbs to adjectives and non-fiction articles tend to have a higher value for the ratio of the number of adjectives to pronouns.

References :

[1] M. R. Qureshi, S. Ranjan, Rajakrishnan P. R., K. Shah, A Simple Approach to Classify Fictional and Non-Fictional Genre, Proceedings of the Second Storytelling Workshop,Association for Computational Linguistics, pages 81–89 Florence, Italy, August 1, 2019.

[2] Fictometer Web Service : http://fictometer.com/

[Users can upload any PDF file or raw text on this website to analyse its fiction content]

[3] NewsChase News App :

https://play.google.com/store/apps/details?id=com.iiserb.fictonews

Hypoxia-induced changes in intragenic DNA methylation correlate with alternative splicing in breast cancer

Nagarjun Vijay
IISER Bhopal

Abstract:
The tumor microenvironment is marked by gradients in the level of oxygen and nutrients, with oxygen levels reaching a minimum at the core of the tumor, a condition known as tumor hypoxia. Mediated by members of the HIF family of transcription factors, hypoxia leads to a more aggressive tumor phenotype by transactivation of several genes as well as reprogramming of pre-mRNA splicing. Intragenic DNA methylation, which is known to affect alternative splicing in cancer, could be one of several reasons behind the changes in splicing patterns under hypoxia. Here, we have tried to establish a correlation between intragenic DNA methylation and alternative usage of exons in tumor hypoxia. First, we have generated a custom hypoxia signature consisting of 34 genes that are up-regulated under hypoxia and are direct targets of HIF-1α. Using this gene expression signature, we have successfully stratified publicly available breast cancer patient samples into hypoxia positive and hypoxia negative groups followed by mining of differentially spliced isoforms between these groups. The Hypoxia Hallmark signature from MSigDB was also used independently to stratify the same tumor samples into hypoxic and normoxic. We found that 821 genes were showing differential splicing between samples stratified using a custom signature, whereas, 911 genes were showing differential splicing between samples stratified using the MSigDB signature. Finally, we performed multiple correlation tests between the methylation levels (β) of microarray probes located within 1 kilobase pairs of isoform-specific exons using those exons' expression levels in the same patient samples in which the methylation level was recorded. We found that the expression level of one of the exons of DHX32 and BICD2 significantly correlated with the methylation levels, and we were also able to predict patient survival (p-value: 0.02 for DHX32 and 0.0024 for BICD2). Our findings provide new insights into the potential functional role of intragenic DNA methylation in modulating alternative splicing during hypoxia.

Addressing climate change extremes using dynamical models and big data: an integrated approach

Pankaj Kumar

Indian Institute of Science Education and Research Bhopal, India

Abstract

Extreme weather events have increased many folds, especially their intensity and frequency. They have a significant adverse impact on regional to local property and lethality. Climate models are effective tools to investigate the complex systems and associated mechanism in the climate and environmental sciences. They can be used to project future changes and framing of policies based on these projections. The immensity of the climate science data derived from such systems (observations and climate model simulations) is overwhelming. Management and analysis of this big data present a ubiquitous challenge to climate scientists and policymakers alike. The fusion of big data (climate model) and advance data science methods such as machine learning can help climate scientists overcome challenges such as prediction and associated uncertainty of extreme events, climate downscaling, multi-model ensembles, short-range weather forecasting such as the prediction of heatwaves among others. In this study, an example of a prediction of seasonal temperature anomaly has been carried out using Canonical Correlation Analysis. It is a multivariate analysis which is used to derive a spatiotemporal relationship between two climatic parameters. The skill of the model is calculated using the correlation between predicted and original data and the causal relationship is further established using the Granger-causality test.

Uniform Information Density Effects on Syntactic Choice in Hindi

Rajakrishnan P. Rajkumar
IISER Bhopal

Abstract:

According to the UNIFORM INFORMATION DENSITY (UID) hypothesis, speakers tend to distribute information density across the signal uniformly while producing language. We use a variant of the above UID hypothesis in order to investigate the extent to which word order choices in Hindi are influenced by the drive to minimize the variance of information across entire sentences. To this end, we propose multiple lexical and syntactic measures (at both word and constituent levels) to capture the uniform spread of information across a sentence. Subsequently, we incorporate these measures in machine learning models aimed to distinguish between a naturally occurring corpus sentence and its grammatical variants (expressing the same idea). Our results indicate that our UID measures are not a significant factor in predicting the corpus sentence in the presence of lexical surprisal, a competing control predictor.

Iterative reconfiguration algorithm for pruning large uncertain biological networks

Rati Sharma[1] and Erel Levine[2]
[1]Indian Institute of Science Education and Research, Bhopal, Madhya Pradesh, India
[2]Northeastern University, Boston, Massachusetts, USA

Large biological networks such as protein protein interaction networks or gene co-functional networks fall under the category of uncertain graphs because the links in these graphs are given in terms of likelihood scores. These likelihood scores are generated as a result of an analysis of several biological experiments that are often riddled with noisy statistics. Therefore, link prediction in uncertain graphs has emerged as a major area of research in recent times [1]. Recent link prediction methods, such as GraphGAN [2] that is a generative framework and SEAL [3] that is based on defining local subgraphs and learning features through graph neural networks assume the network used to train the algorithm to be the ground truth. However, biological networks are far from complete. Therefore, there is a need for an algorithm that does not assume correctness in the graph network at hand, but still reconfigures, i.e., predicts new links while at the same time prunes the network by removing insignificant links. We have recently developed a new network reconfiguration methodology where we first identify the local subgraphs around each link in the network, determine some of the node and edge features of the subgraph and then use that information to iteratively predict new links and remove some existing links. Eventually, this method reconfigures an uncertain graph where the links that have lower likelihood score and are sparsely connected are removed whereas new links appear in highly connected regions. In this talk, I will present details about this algorithm and also discuss the results of its subsequent testing on gene co-functional networks of two bacterial species, namely, *E. coli* and *P. aeruginosa*. The differentially expressed genes (DEGs) of various RNAseq experiments of the two bacterial species are listed from a literature survey. Comparing the original and the new reconfigured network, it is inferred that the reconfigured network has more links made up of DEGs. Further, the network community quality of the reconfigured network is also much better than the original network.

References

[1] E. Mutlu and T. Oghaz, Proceedings of ACM, NewYork, NY, USA (2018)
[2] H. Wang et al., The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2508 (2018)
[3] M. Zhang and Y. Chen, 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Economists and Data: Happily Married

Sandip K. Agarwal
IISER Bhopal

For economists, the following words of W. E. Deming, “In God we trust; all others must bring data.” are becoming more and more sacrosanct. Increasingly, economists are making use of newly available large-scale administrative data or private sector data that often are obtained through collaborations with private firms, giving rise to new opportunities and challenges. The advent of big data is already allowing for better measurement of economic effects and outcomes and is enabling novel research designs across a range of topics. This talk will briefly talk about different current and potential applications from the use of AI and Machine Learning in Economics – touching examples from the agriculture, judiciary, online retail markets and financial trading. Finally, it will give an economist’s view of automation - particularly AI as a General Purpose Technology (GPT) and its implications for the global economy.

AI-based platforms to predict the metabolism of xenobiotics molecules by human microbiome

Vineet K. Sharma*, Ashok K. Sharma, Aditya Malwe, Shubham K. Jaiswal

IISER Bhopal

The population-specific differences in human microbiome due to diet, geographical location, age, etc. play a key role in determining human health and metabolism. The microbial species residing in the human gut, skin etc. also harbours an enormous metabolic potential, which can alter the metabolism of orally administered drugs or topically applied molecules leading to individual/population-specific differences in their metabolism. The investigation of microbiome-driven xenobiotic metabolism by experimental studies is a challenging task considering the large heterogeneous pool of human bacteria and their metabolic enzymes. Therefore, using a hybrid of AI, machine learning and chemoinformatics approaches, we have developed novel computational approaches such as 'DrugBug' and 'SkinBug' to predict the metabolic enzymes and bacterial species that can potentially carry out the promiscuous biotransformation of a xenobiotic molecules. A substrate database was constructed for metabolic enzymes from all the known/available human-associated bacteria. The structural properties (fingerprints) from these substrates were extracted and used for the development of Random Forest and Deep Learning models, which displayed excellent accuracies on cross-validation and blind set. After the prediction of EC subclass, the specific metabolic enzyme (EC) is identified using a molecular similarity search. The performance was further evaluated on an independent set of FDA-approved drugs and other clinically important molecules. A few selected predictions were also validated using structural and experimental methods. To our knowledge, DrugBug and SkinBug are the only available AI-based tools for the prediction of xenobiotic metabolism by metabolic enzymes present in the human skin and gut microbiome.