7th International Conference
Big Data Analytics
December 17 - 20, 2019

www.bda2019.org

Supported by
sprinklr

This conference is organised by the School of Engineering and Applied Science, Ahmedabad University
Ahmedabad University is a liberal arts driven research university. It is dedicated to rigorous academic pursuit with a focus on building inquiry as a value through interdisciplinary learning. We prepare students to think deeply and creatively across fields, and emerge as independent thinkers and compassionate leaders who can innovatively engage with the complex challenges of our society. Our research and teaching spans humanities and languages, social sciences, mathematical and physical sciences, biological and life sciences, performing and visual arts, engineering, and management.

The University was established in 2009 by the Ahmedabad Education Society at the turn of the last century, to reimagine higher education and to build the foundations of liberal education and research for benefit of society.

An Ahmedabad University education is premised on three unique characteristics: interdisciplinarity, experiential learning, and research thinking. Instead of the traditional lecture-based method, our emphasis is on teaching students to think for themselves - learning to link theoretical developments to living contexts and bringing one’s learning to bear on the world at large. Our system of education is broad and deep – flexible enough to help a young student discover her own passion and area of interest yet deep to build enduring skills and knowledge. The University’s signature Foundation Programme builds the fundamentals of a holistic application of theoretical learning. Our graduate programmes and particularly the doctoral programmes are at the frontiers of knowledge and their application.

At Ahmedabad, you learn from the best of faculty who have been trained at some of the best universities globally and in some of the more cutting edge laboratories. Our faculty research is published at high quality international outlets. One third of our students go for higher education, one third back to family business or start a new enterprise and the remaining find jobs in Indian and global corporations. Located in the centre of the city amidst an upcoming urban forest, the university’s campus comprises heritage and contemporary buildings designed but some of the most exciting global architects. They house the university’s four Schools and four Centers.

The Amrut Mody School of Management has built the country’s most advanced undergraduate and MBA programmes with strong grounding in liberal arts and industry consulting. The School of Arts and Sciences aims to create an intellectual environment with a difference and be at the frontiers of cutting-edge global research across the natural, physical and social sciences, humanities, and performing and visual arts. It offers interdisciplinary programmes at undergraduate and graduate levels. The School of Engineering and Applied Science is one of the most unique engineering schools in India with an engineering foundation spanning several disciplines (based on the premise that all engineers must know many disciplines to be able to solve complex problems) and a unique hands-on experiential learning environment (i.e., project based learning and building of products) that is strongly centered on design.

The Centre for Learning Futures focuses on innovation in teaching and learning and runs the reading and writing programme of the University. The Centre for Heritage Management is one of its kind in India. It focuses on the management of tangible, and intangible heritage as well as on art history and aesthetics. The Global Centre for Environment and Energy addresses research needs at the interfaces of three vital global challenges - the environment, energy and sustainability by interlinking Policy, Entrepreneurship, and Science and Technology. VentureStudio is the University’s incubator that helps students start an innovation driven enterprise. It also offers end to end support to startups, including assistance with funding.
About the Conference

**CHIEF PATRONS**

<table>
<thead>
<tr>
<th>Pankaj Chandra</th>
<th>Sudhir Jain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vice Chancellor</td>
<td>Director</td>
</tr>
<tr>
<td>Ahmedabad University</td>
<td>IIT, Gandhinagar</td>
</tr>
</tbody>
</table>

**HONORARY CHAIRS**

<table>
<thead>
<tr>
<th>K. S. Dasgupta</th>
<th>Sunil Kale</th>
<th>Rekha Jain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Director</td>
<td>Dean, School of Engineering and</td>
<td>Professor</td>
</tr>
<tr>
<td>DA-IICT, Gandhinagar</td>
<td>Applied Science</td>
<td>IIM Ahmedabad</td>
</tr>
<tr>
<td></td>
<td>Ahmedabad University</td>
<td></td>
</tr>
</tbody>
</table>

**GENERAL CHAIR**

Sanjay Chaudhary
Associate Dean, School of Engineering and Applied Science
Ahmedabad University

**STEERING COMMITTEE CHAIR**

P. Krishna Reddy
IIIT Hyderabad

**STEERING COMMITTEE**

- **S K Gupta** (IIT Delhi)
- **Srinath Srinivasa** (IIIT Bangalore)
- **Krithi Ramamritham** (IIT Bombay)
- **Sanjay Kumar Madria** (Missouri University of Science and Technology, USA)
- **Masaru Kitsuregawa** (University of Tokyo, Japan)
- **Raj K Bhatnagar** (University of Cincinnati, USA)
- **Vasudha Bhatnagar** (University of Delhi)
- **Mukesh Mohania** (IBM Research, Australia)
- **HV Jagadish** (University of Michigan, USA)
- **Ramesh Kumar Agrawal** (Jawaharlal Nehru University)
- **Divyakant Agrawal** (University of California at Santa Barbara, USA)
- **Arun Agarwal** (University of Hyderabad)
- **Subhash Bhalla** (The University of Aizu, Japan)
- **Jaideep Srivastava** (University of Minnesota, USA)
- **Anirban Mondal** (Ashoka University)
- **Sharma Chakravarthy** (The University of Texas at Arlington, USA)
PROGRAMME COMMITTEE CHAIRS

<table>
<thead>
<tr>
<th>Sanjay Madria</th>
<th>Philippe Fournier-Viger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missouri University of Science and Technology, USA</td>
<td>Harbin Institute of Technology (Shenzhen), China</td>
</tr>
</tbody>
</table>

ORGANISING CHAIRS

<table>
<thead>
<tr>
<th>Kumar Shashi Prabh</th>
<th>Mehul Raval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmedabad University</td>
<td>Pandit Deendayal Petroleum University</td>
</tr>
</tbody>
</table>

FINANCE CHAIRS

<table>
<thead>
<tr>
<th>Dhaval Patel (Ahmedabad University)</th>
<th>Amitava Ghosh (Ahmedabad University)</th>
</tr>
</thead>
</table>

SPONSORSHIP CHAIR

<table>
<thead>
<tr>
<th>Anil Roy (DA-IICT, Gandhinagar)</th>
<th>P Krishna Reddy (IIIT Hyderabad)</th>
</tr>
</thead>
</table>

WORKSHOP CHAIRS

<table>
<thead>
<tr>
<th>Srikrishnan Divakaran (Ahmedabad University)</th>
<th>Punam Bedi (University of Delhi)</th>
</tr>
</thead>
</table>

TUTORIAL CHAIRS

<table>
<thead>
<tr>
<th>Mukesh Mohania (IBM, Australia)</th>
<th>Takahiro Hara (Osaka University, Japan)</th>
</tr>
</thead>
</table>

PUBLICITY CHAIRS

<table>
<thead>
<tr>
<th>Shelly Sachdeva (National Institute of Technology Delhi)</th>
<th>M. T. Savaliya (Vishwakarma Government Engineering College)</th>
</tr>
</thead>
</table>

WEBSITE CHAIR

<table>
<thead>
<tr>
<th>Pratik Trivedi (Ahmedabad University)</th>
<th></th>
</tr>
</thead>
</table>
ORGANISING COMMITTEE

- C K Bhensdadia (Dharmsinh Desai University)
- Amit Ganatra (CHARUSAT)
- Sanjay Sampat (Government Polytechnic College)
- Aditya Patel (Ahmedabad University)
- George Varughese (Ahmedabad University)
- Kangna Bagani (Ahmedabad University)
- Rahul Bhathiji (Ahmedabad University)
- Barbara Moraswka (Ahmedabad University)
- Anurag Lakhani (Ahmedabad University)
- Kuntal Patel (Ahmedabad University)
- Shefali Naik (Ahmedabad University)
- Jaydeep Raolji (Ahmedabad University)
- Kunjal Gajjar (Ahmedabad University)
- Hiral Vegda (Ahmedabad University)
- Kinjal Arya (Ahmedabad University)
- Govind Prajapati (Ahmedabad University)
- Amar Gajjar (Ahmedabad University)
- Aman Dave (Ahmedabad University)
- Kruti Yadav (Ahmedabad University)
- Manav Shah (Ahmedabad University)
- Jainam Shah (Ahmedabad University)
- Mohit Vaswani (Ahmedabad University)
- Tirth Jivani (Ahmedabad University)
- Neha Shukla (Ahmedabad University)
- Vinay Vachcharajani (Ahmedabad University)
- Siddhi Shah (Ahmedabad University)

PROGRAMME COMMITTEE

- Amartya Sen (Oakland University, MI, USA)
- Akhil Kumar (Penn State University, USA)
- Alok Singh (University of Hyderabad)
- Anirban Mondal (Ashoka University)
- Arkady Zaslavsky (CSIRO, Australia)
- Engelbert Mephu Nguifo (Université Blaise Pascal – Clermont Ferrand, France)
- Himanshu Gupta (IBM Research, India)
- Jerry Chun-Wei Lin (Western Norway University of Applied Sciences, Norway)
- Ji Zhang (University Southern Queensland, Australia)
- Jose Marie Luna (University of Cordoba, Spain)
- Ladjel Bellatreche (ENSMA, France)
- Lakshmish Ramaswamy (The University of Georgia USA)
- Lukas Pichl (International Christian University, Japan)
- Morteza Zihayat (Ryerson University, Canada)
- Naresh Manwani (IIIT Hyderabad)
- Nazha Selmaoui Folcher (Université de Nouvelle Calédonie, New Caledonia)
- Pradeep Kumar (IIM Lucknow)
- Praveen Rao (University of Missouri - Kansas City, USA)
- Prem Jayaraman (Swinburne University of Technology, Australia)
- Roger Nkambou (Université du Québec à Montréal, Canada)
- Santhanagopalan Rajagopalan (IIIT Bangalore)
- Satish Narayana Srirama (University of Tartu, Estonia)
- Sebastian Ventura (University of Cordoba, Spain)
- Srinath Srinivasa (IIIT, Bangalore)
- Tin Truong Chi (University of Dalat, Vietnam)
- Tzung-Pei (Hong National University of Kaohsiung, Taiwan)
- Uday Kiran (University of Tokyo, Japan)
- Vasudha Bhatnagar (University of Delhi, India)
- Vimal Kumar (University of Walkato, New Zealand)
- Zeyar Aung Masdar (Institute of Science and Technology, UAE)

Venue: Auditorium
GICT Building, Central Campus

09:00 - 09:30  Registration (Venue: Lobby, Auditorium, GICT Building)
09:30 - 09:45  Inauguration
09:45 - 10:45  Invited Talk: “Video Analytics” by Mehul Raval, PDPU, Gandhinagar
10:45 - 11:15  Tea / Coffee Break (Venue: Open Space, First Floor, GICT Building)
13:15 - 14:00  Lunch Break (Venue: Open Space, First Floor, GICT Building)
14:00 - 15:00  Invited Talk: “Data Clustering with Applications in Computational Biology”, by Srikrishnan Divakaran, Ahmedabad University, Ahmedabad, India.
15:00 - 15:30  Tea / Coffee Break (Venue: Open Space, First Floor, GICT Building)
15:30 – 16:30  Invited Talk: “Big Data and Machine Learning for Disaster Management”, by Sanjay Madria, Distinguished Professor, Missouri University of Sc. & Tech., USA
16:30 - 17:00  Invited Talk: “Stemmers for Gujarati Language” by Nakul Dave, VGEC (Ahmedabad) and Dr. Mayuri Mehta, SCET(Surat)
17:00 - 17:30  Closing Remarks
TUESDAY
17 DEC 2019

Workshop on “Data Science for Agriculture and Natural Resource Management”

Venue: Room 106, GICT Building, Central Campus

09:00 - 09:30  
Registration (Venue: Lobby, Auditorium, GICT Building)

09:30 - 09:45  
Inauguration and Introduction by P. Krishna Reddy, IIIT Hyderabad

09:45 - 10:45  
Keynote Talk: “Experiences of Building IT-based Systems to Deliver Actionable Agricultural Information to Indian Farmers” by Krishna Reddy, IIIT Hyderabad

10:45 - 11:15  
Tea / Coffee Break

11:15 - 12:15  
Invited Talk: “Predictive Soil Modelling - Development of Machine Learning Approach” by Ranendu Ghosh, Professor and Dean (Students), DA-IICT, Gandhinagar

12:15 - 13:15  
Invited Talk: “Developing Big Data Analytics Architecture for Spatial Data” by Sanjay Chaudhary and Purnima Shah, School of Engineering and Applied Science, Ahmedabad University

13:15 - 14:00  
Lunch Break

14:00 - 15:00  
Invited Talk: “Information Technology and Natural Resource Management” by Dr. G.P. Obi Reddy, National Bureau of Soil Survey and Land Use Planning, Nagpur, India

15:00 - 15:30  
Tea / Coffee Break

15:30 - 16:30  
Discussions

16:30 - 17:00  
Closing Remarks
08:00 - 08:30  
**Registration** (Venue: Lobby, Auditorium, GICT Building)  
**Breakfast** (Venue: Open Space, First Floor, GICT Building)

Venue: Auditorium, GICT Building

08:30 - 9:00  
**Welcome by General Chair and Programme Introduction by Programme Committee Chairs**

09:00 - 9:45  
**Inauguration and Welcome Address**  
**Pankaj Chandra, Vice-Chancellor, Ahmedabad University**

**Inaugural Talk by the Chief Guest**  
**Professor Errol D'Souza, Director, Indian Institute of Management Ahmedabad**

**Session Chair: Sanjay Chaudhary, Ahmedabad University**

9:45 - 10:45  
**Keynote Talk 1: Blockchain issues and solutions, how to let your big data sing**  
Speaker: Elizabeth Chang, Professor, Australian Defence Force Academy (ADFA), Australia  
Session Chair: Philippe Fournier-Viger, Harbin Institute of Technology (Shenzhen), China

10:45 - 11:15  
**Tea/Coffee Break** (Venue: Open Space, First Floor, GICT Building)

11:15 - 12:15  
**Invited Talk 1: Elections & Social Media**  
Speaker: Ponnurangam Kumaraguru, Professor, IIIT Delhi

**Invited Talk 2: Interactive Web-Based Geospatial Big-Data Analytics for Vegetation Monitoring**  
Speaker: Shashikant Sharma, Scientist-SG and Group Head, Space Applications Centre, ISRO, Ahmedabad  
Session Chair: Mukesh Mohania, IIIT Delhi

12:15 - 13:15  
**Lunch** (Venue: Open Space, First Floor, GICT Building)
13:15 - 14:45  
**Web Data and Information Extraction**  
Session Chair: Goce Trajcevski, Iowa State University, USA  

**Paper 1: Data Cube is Dead, Long Life to Data Cube in the Age of Web Data**  
Selma Khouri, Nabila Berkani, Ladjel Bellatreche and Dihia Lanasri. ENSMA, France; Poitiers University, France  

**Paper 2: Improving Result Diversity using Query Term Proximity in Exploratory Search**  
Vikram Singh and Mayank Dave. National Institute of Technology, Kurukshetra  

**Paper 3: Pairing Users in Social Media via Processing Meta-data from Conversational Files**  
Meghna Chaudhary, Ravi Sharma and Sriram Chellappan, University of South Florida, USA  

14:45 - 15:15  
**Tea/Coffee Break** (Venue: Open Space, First Floor, GICT Building)
15:15 - 17:15

**Predictive Analytics Applications**
Session Chair: R. Uday Kiran, Aizu University, Japan

**Paper 4: Suśruta: Artificial Intelligence and Bayesian Knowledge Network in Health Care - Smartphone Apps for diagnosis and differentiation of anemias with higher accuracy at Resource Constrained Point-of-Care settings**
Asoke Talukder, Shubham Yadav, Debajnan Das, Venkanna U, Rajarshi Mahapatra, Ankur Shrivastava, Prantar Chakrabarti and Sakthi Ganesh. IIIT Naya Raipur, Chhattisgarh, India; SRIT India Pvt. Ltd., Bangalore, India; AIIMS, India; NRS Medical College and Hospital, Kolkata

**Paper 5: Market Intelligence for Agricultural Commodities using Forecasting and Deep Learning Techniques**
Swapnil Shrivastava, Supriya N Pal and Ranjeet Walia. Centre for Development of Advanced Computing (CDAC), Bangalore, India; National Institute of Technology, Hamirpur

Paper 6: TKG: Efficient Mining of Top-K Frequent Subgraphs
Philippe Fournier-Viger, Chao Cheng, Jerry Chun-Wei Lin and Unil Yun, Harbin Institute of Technology (Shenzhen), China; Western Norway University, Norway; Sejong University, Seoul, Republic of Korea

**Paper 7: Recent Advances and Challenges in design of Non-Goal-Oriented Dialogue Systems**
Akanksha Mehndiratta and Krishna Asawa. Jaypee Institute of Information Technology, Noida

**Tutorial 1 (Continue from previous session)**

**Tutorial 2: Streaming Data Processing**
Speaker: Prasad Despande, Co-founder, Alpha Analyst, India.
Session Chair: Ladjel Bellatreche, ENSMA, France
Venue: 107, GICT Building

17:30 - 18:00

**Free time for Networking**
Venue: 113, GICT Building

18:00 - 19:00

**Cultural Program by Students of Ahmedabad University**
Venue: Auditorium, GICT Building

19:00 - 20:00

**Conference Reception with Dinner**
Venue: Open Space, First Floor, GICT Building

Steering committee meeting (invited members only)
Venue: 113, GICT Building
08.15-09.00  Tea / Coffee with Breakfast (Venue: Open Space, First Floor, GICT Building)

09.00-10.00  Keynote Talk 2: From GPS, Google Maps & Uber to Spatial Computing
Speaker: Shashi Shekhar, McKnight Distinguished University Professor, University of Minnesota, USA
Session Chair: Sanjay Madria, Missouri University of Sc. & Tech., USA
Venue: Auditorium, GICT Building

10.00-10.30  Tea/Coffee Break (Venue: Open Space, First Floor, GICT Building)

10.30-11.30  Invited Talk 3: Turning Sensing Data into Smart Data for Smart Sustainable Cities
Speaker: Koji Zettsu, NIST, Japan

Invited Talk 4: The potential of Big Data Analytics in making sense of Urban Heat Island and Air Quality Monitoring data
Speaker: Girish Agrawal, Professor, Jindal School of Art and Architecture (JSAA), Jindal Global University (JGU), NCR of Delhi
Session Chair: Anish Mathuria, DA-IICT, Gandhinagar

11.30-13.00  Graph Analytics
Session Chair: Anirban Dasgupta, IIT Gandhinagar

Paper 8: Gossip Based Distributed Real Time Task Scheduling with Guaranteed Performance on Heterogeneous Networks
Moumita Chatterjee, Anirban Mitra, Sudipta Roy, Somasis Roy, Hirav Shah, Sanjit Kr. Setua, Calcutta University, India; Ganpat University, Mehsana, India; Academy of Technology, India; Washington University, USA

Paper 9: Data-Driven Optimization of Public Transit Schedule
Sanchita Basak, Fangzhou Sun, Saptarshi Sengupta and Abhishek Dubey, Vanderbilt University, USA.

Paper 10: Discovering Spatial High Utility Frequent Itemsets in Spatiotemporal Databases,
P.P.C Reddy, R. Uday Kiran, Koji Zettsu, Masashi Toyoda, P. Krishna Reddy, and Masaru Kitsuregawa, IIIT, Hyderabad, India; National Institute of Information and Communications Technology, Tokyo, Japan; National Institute of Informatics, Tokyo, Japan
7th International Conference on Big Data Analytics

Venue: School of Engineering and Applied Science
Ahmedabad University
GICT Building, Central Campus, Navrangpura,
Ahmedabad 380009 | www.bda2019.org

11:30 - 13:00
Tutorial 3: Emergence of Artificial Intelligence and Blockchain in Healthcare
Speakers: Mayuri Mehta, SCET, Surat and Anuj Kumar Garg, IBM India
Session Chair: Arup Dasgupta, SAC, India.
Venue: 106, GICT Building

Tutorial 4: Knowledge Graphs: Representation, Management and Applications
Speaker: Raghava Mutharaju, IIIT Delhi, India.
Venue: 107, GICT Building
Session Chair: Alka Mahajan, Institute of Technology, Nirma University

13:00 - 14:00
Lunch (Venue: Open Space, First Floor, GICT Building)

14:00 - 16:00
Deep Learning
Session Chair: Philippe Fournier-Viger. Harbin Institute of Technology (Shenzhen), China

Paper 11: Deep Learning in the Domain of Near-Duplicate Document Detection
Rajendra Kumar Roul and Jajati Keshari Sahoo, Thapar Institute of Engineering and Technology, Patiala, India; BITS-Pilani, Goa

Paper 12: Real Time Static Gesture Detection using Machine Learning
Kalpdrum Passi and Sandipgiri Goswami, Laurentian University, Canada

Paper 13: Efficient Algorithms For Flock Detection in Large Spatio-Temporal Data
Jui Mhatre, Harsha Agrawal, and Sumit Sen Veermata Jijabai Technological Institute, Mumbai, India; Indian Institute of Technology - Bombay

Paper 14: Local Temporal Compression for (Globally) Evolving Spatial Surfaces
Xu Teng, Prabin Giri, Matthew Dwyer, Jidong Sun, and Goce Trajcevski. Iowa State University, USA

16:00 - 16:15
Tea/Coffee Break (Venue: Open Space, First Floor, GICT Building)

16:15 - 17:45
Panel: Big Data Analytics is different from AI
Moderator: Anirban Mondal, Ashoka University
Panelist: Goce Trajcevski. Iowa State University, Philippe Fournier-Viger. Harbin Institute of Technology (Shenzhen), China, Koji Zettsu, Director, NIST, Japan, Shashi Shekhar, McKnight Distinguished University Professor, University of Minnesota, USA, Sriram Chellappan, University of South Florida, USA
Venue: Auditorium, GICT Building

18:00 - 21:00
Banquet Dinner (Venue: Lawns, GICT Building)
Keynote Talk 3: Deep Learning Models for Medical Image Analysis: Challenges and Future Directions  
Speaker: Ramesh Agrawal, Dean, School of Engineering, Jawaharlal Nehru University, New Delhi, India.  
Session Chair: Sunil Kale, Ahmedabad University, India.  
Venue: Auditorium, GICT Building

Tea/Coffee Break (Venue: Open Space, First Floor, GICT Building)

Data Mining  
Session Chair: Rekha Jain, IIM Ahmedabad

Paper 15: An Explicit Relationship between Sequential Patterns and their Concise Representations  
Hai Duong, Tin Truong, Bac Le and Philippe Fournier-Viger. University of Science, Ho Chi Minh, Vietnam; University of Dalat, Dalat, Vietnam, and Harbin Institute of Technology (Shenzhen), Shenzhen, China

Paper 16: A Novel Approach to Identify the Determinants of Online Review Helpfulness and Predict the Helpfulness Score Across Product Categories  
Debasmita Dey and Pradeep Kumar. Indian Institute of Management, Lucknow

Paper 17: Analysis and Recognition of Hand-drawn Images with Effective Data Handling  
Mohit Gupta and Pulkit Mehndiratta. Jaypee Institute of Information Technology, Noida

Paper 18: Interpreting Context of Images using Scene Graphs  
Himangi Mittal, Ajith Abraham, and Anuja Arora, Jaypee Institute of Information Technology, India; MIR Labs, USA

Lunch Break (Venue: Open Space, First Floor, GICT Building)

Invited Talk 5: Segment-search vs Knowledge Graphs: Making a Keyword Search Engine for Web Documents  
Speaker: Subhash Bhalla, Professor, The University of Aizu, Japan  
Session Chair: S. K. Patra, Director IIIT Vadodara

Local City Tour
KEYNOTE 1

Enterprise on Blockchain
Let your big data sing

Professor Elizabeth Chang (UNSW Canberra, Australia)

Abstract: This keynote gives an overview of 3 core components of Big data analytics, namely frontier AI enabled structured, semi and unstructured big data screening, reasoning and predictive analytics, known as applications; heterogenous data source integration known as backend or clouds; and data representation, visualisation and user interaction known as front-end. There is a world-wide effort for solution development on how to utilization of Blockchain for back-end enterprise databases modernization including distributed trust, single south of truth, transparency, immutability, traceability and anonymity etc. In addition, how to produce low-cost yet high dynamic systems that utilise variety of open source AI components for data intelligence and how to move the UX to mobile enabled platforms rather than back-office. To date, there has been no ready used solutions for large corporate and enterprises.

This talk presents two real world project experiences and the issues and challenges facing to the adoption of the disruptive technology like Blockchain in the large corporate or private enterprises, where 1000-10,000 + legacy system co-exists and how blockchain can be integrated in such a complex and continue evolving environment. Particularly we present our patented work on Blockchain enabled big data analytics, data migration, data lake, combined with AI and open source technologies.

This keynote also presents the technical issues of Blockchain including Blockchain ecosystems with legacy ERPs, real time performance of Blockchain, data lake on Blockchain, data integration and analytics between multi-Blockchains, private keys and certificates for multi-channel private Blockchains, and limitations of standards and policy. We provide possible solutions including Blockchain configuration, plug n play Blockchain, enterprise private and public Blockchains, network of partners, speed, space and security, building trusted data sources for visibility, transparency, immutability, accountability, and enablement of productivity, efficiency and effectiveness.

Biography: Professor Elizabeth Chang is Professor of Logistics and Canberra Fellow at UNSW at Australian Defence Force Academy (ADFA). Professor Chang leads the Defence Logistics research group at UNSW Canberra, targeting the key issues in Logistics ICT, Defence Logistics and Sustainment, Big Data Management, Predictive Analytics, IoT and Cyber-Physical Systems, Trust, Security, Risk and Privacy.

She has delivered over 50 Keynote/Plenary speeches largely at major IEEE Conferences and most recently in the areas of Semantics, Business Intelligence, Big Data Management, Data Quality and the like. Her academic achievement includes 24 Competitive Research Grants, including 12 Australian Research Council (ARC) Grants worth over $15 million. She has supervised/co-supervised 42 PhD theses to completion, 21 Master theses and 16 Post-docs. She has published 7 authored books, over 500 international journal papers and conference papers with an H-Index of 47 (Google Scholar) and over 12,000 citations.
KEYNOTE 2

From GPS, Google Maps & Uber to Spatial Computing

Professor Shashi Shekhar (University of Minnesota, USA)

Abstract: Spatial computing and spatial big data have already enriched billions of lives via pervasive services (e.g., navigation and ride-sharing apps), ubiquitous systems (e.g., geographical information system, spatial database management system), and pioneering scientific methods (e.g., spatial statistics). With 2 billion receivers in use for location and time services, the GPS has become a critical infrastructure for the world economy for use cases ranging from precision agriculture to navigation to ride sharing to smart cities. Government and industry have recently started major initiatives such as NASA Earth Exchange, Amazon Earth on AWS, Google Earth Engine, Microsoft AI for Earth, and NSF Navigating the New Arctic for meeting grand challenges facing our changing planet such as climate change and environmental sustainability. However, many spatial data science questions need to be probed to realize the transformative potential. For example, how may modern economy survive wide-spread GPS-jamming (or spoofing)? How may one continuously monitor our changing planet at high spatial resolution even during nights? How may spatial big data (e.g., smart-phone trajectories) be mined without violating privacy? How may machine learning methods be generalized to address spatial challenges (e.g., auto-correlation, multi-scale, modifiable areal unit problem such as Gerrymandering)? How may we address spatial bias in data even when social feedback loops increase it? How may algorithms scale up to spatial big data to learn unbiased models? How may we leverage vehicle big data (e.g., on-board diagnostics with high spatio-temporal resolution) for eco-routing to model edge dependence of energy-use and emissions? This presentation shares a perspective on the societal accomplishments, opportunities, and research needs in spatial computing, spatial big data and spatial data science.

Biography: Shashi Shekhar, a McKnight Distinguished University Professor at the University of Minnesota, is a leading scholar in Geographic Information Systems (GIS). He co-edited an Encyclopedia of GIS and co-authored a Spatial Databases textbook. He received the IEEE-CS Technical Achievement Award, the UCGIS Education Award and was elected a Fellow of the IEEE and the AAAS. Shashi is a co-Editor-in-Chief of Geo-Informatica journal (Springer) and has served on the Computing Community Consortium Council, National Academies' committees (Mapping Sciences, GEOINT Workforce). He co-organized the "From GPS and Virtual Globes to Spatial Computing 2020" workshop to catalyze community research visions.
Deep Learning Models for Medical Image Analysis: Challenges and Future Directions

Professor Ramesh Agrawal (Dean, School of Engineering, Jawaharlal Nehru University, New Delhi)

Abstract: Medical imaging modalities such as X-rays, CT scan, ultrasound, magnetic resonance imaging, functional magnetic resonance imaging (fMRI) exist to measure physical property of the human body. These modalities have raised the capabilities of clinicians by providing visualization of the internal details of human body with no or minimal invasion and has thus reduced the surgical procedures. However, manual analysis of these medical images is challenging due to several reasons such as extensive medical data, the bias involved in subjective assessment by clinicians, presence of the fine details that ordinarily cannot be observed through visual inspection. Machine learning methods have been employed to analyse medical images, which supports the clinician in the diagnosis of peculiar diseases that are difficult to identify otherwise. Recently, deep learning method is proposed which has the capability to learn or extract high-level, complex abstractions to represent samples through a hierarchical learning process from training samples. The power of deep learning lies in its capacity to analyse and learn patterns from large volume of data in unsupervised manner, which can be a valuable tool for Big Data Analytics where most of the available raw data is largely unlabeled and uncategorized. Additionally, the availability of GPU allows us to analyse unprecedented amounts of data. In my talk, I will discuss state of the art deep learning models for medical image segmentation and classification for a wide variety of health application areas. Finally, the talk will also focus on current challenges and future research directions for medical image analysis.

Biography: Dr. R. K. Agrawal is Professor in the School of Computer & Systems Sciences of Jawaharlal Nehru University, New Delhi. He is presently Dean of School of Engineering and also served as Dean of School of Computer & Systems Sciences between 2015-2017. He holds M.Tech degree in computer science from IIT Delhi and a PhD from University of Delhi. His research topics include data mining, pattern recognition, machine learning and medical imaging. Dr. Agrawal is an author of more than 100 scientific publications in refereed international journals and conferences, most of them on machine learning and medical imaging related. He has served on the Steering Committees, Program Committees and organizing Committees of many international/national conferences. Dean, School of Engineering Professor, School of Computer & Systems Sciences Jawaharlal Nehru University, New Delhi
INVITED TALK 1

Elections & Social Media

Dr Ponnurangam Kumaraguru (IIIT Delhi)

Abstract: Social Media, today, is playing a very important role in the politics of almost every nation. In India, there has been a boom in social media political campaigns, which was more marked in the 2014 General Elections. All major political parties used social media to promote their manifestos and interact in the public domain in a one-to-one manner. They analysed different views of the people, and worked on them. The Home Minister, in a seminar after the 2014 general elections, also said that “Through social media, the government is outlining its plan, its vision”. After assuming power, the NDA government has been using Twitter, Facebook and blogs to outline its plans, vision, showcase the progress, and other updates. One of the major reason for this was to involve and cater to the interest of the youth of the country, that are active on most social media websites. Owing to this, political parties have invested huge amounts of funds into social media campaigning, hence also increasing their reach. In this talk, I will discuss the data we collected from various social media services for Lok Sabha Elections 2019, including WhatsApp, and the analysis that we have done with the data. I will specifically discuss some interesting insights on suspected accounts, #MainBhiChowkidar campaign, privacy leaks, comparison between 2014 & 2019 data, hashtag analysis. We have made many of our datasets public at http://precog.iiitd.edu.in/resources.html More information can be found at http://precog.iiitd.edu.in/

Biography: Professor Ponnurangam Kumaraguru (PK), is an Associate Professor at IIIT-Delhi and is currently on sabbatical as a Visiting Associate Professor at IIIT Hyderabad. He is a TEDx and an ACM Distinguished speaker & ACM India Eminent Speaker. PK received his Ph.D. from the School of Computer Science at Carnegie Mellon University (CMU). His Ph.D. thesis work on anti-phishing research at CMU contributed in creating an award-winning startup - Wombat Security Technologies. Wombat was recently acquired for USD 225 Million. In addition to his contributions to academia, PK is on advisory role on various government organizations and a Fortune 500 company. He is also part of various Board of Studies of different academic institutes across the country. He has co-authored research papers in the field of Privacy and Security in Online Social Media, Cyber Security, Computational Social Science, Data Science for Social Good, amongst others. He is passionate about solving societal problems using technologies, especially contributing to the domain of social computing. PK and his students have played an integral role in developing a technology used by over 80+ State and Central Government agencies in India. He has spent time making notable contributions at organizations like IBM India Research Labs, Adobe Research Labs – India, Universidade Federal de Minas Gerais (UFMG) – Brazil, and Max Planck Institute for Software Systems – Germany. PK has served as a Program Committee member at prestigious conferences like WWW, CHI, PETs, ICWSM, CSCW, AsiaCCS. He was the Founding Head of Cybersecurity Education and Research Centre (CERC) at IIIT-Delhi. PK started and successfully manages PreCog (precog.iiitd.edu.in), a research group at IIIT-Delhi. PK can be reached on Insta, on Facebook, on Twitter, on LinkedIn and at pk@iiitd.ac.in. Search for #ProfGiri for PK’s social media presence.
INVITED TALK 2

Interactive Web-Based Geospatial Big-Data Analytics for Vegetation Monitoring

Mr Shashikant Sharma (Scientist-SG and Group Head Space Applications Centre, ISRO, Ahmedabad)

Abstract: Monitoring of vegetation health has important applications in the areas of agriculture, climate change and biodiversity conservation. Satellite imagery plays a crucial role in monitoring vegetation health and assessment of physical conditions that may impact the health of vegetation. Ever increasing availability of datasets with high temporal and spatial resolution and ever-growing size of historical long-term datasets presents an opportunity for deriving critical insights for improving agricultural productivity, planning targeted biodiversity conservation programs and accurately assessing climate change impact. However, handling and analysing such huge volume and variety of data is challenging and requires specialized expertise and hardware/software infrastructure. In this paper, we present an interactive analysis system which is delivered as a web-application, for interactively analysing large multi-temporal datasets for vegetation monitoring applications. This system is openly accessible and requires only a web-browser on the part of the user. We discuss key aspects for backend design and implementation that enable very low latency interactive processing and present several useful applications of the system.

Biography: Shashikant Sharma is a Scientist-SG and Group Head, VRG/EPSA at Space Application Center (ISRO). He has thirty years of Research and Development experience in the field of Satellite data processing and GIS and its application in mapping and management of natural resources. He has published 40 scientific papers in various national & international Journals. He was awarded bronze Prize for CATCON award at XXth ISPRS congress held in July 2004 at Istanbul, Turkey and he was awarded ‘Vikram Sarabhai Award in Information Technology 2001-02’ by Department of Science and Technology, Government of Gujarat.
INVITED TALK 3

Turning Sensing Data into Smart Data for Smart Sustainable Cities

Dr Koji Zettsu (Director, NIST, Japan)

Abstract: With recent advances in the Internet of Things (IoT), a wide variety of sensing data are disseminated, shared and utilized in smart cities to improve their efficiency and quality of life of citizens. The key is to turn sensing data into actionable information called “smart data”, used for planning, monitoring, navigation and intelligent decision making. In order to manipulate smart data, advanced data analytics is indispensable for detecting valuable events from sensing data and discovering and predicting latent associations among different kind of events. Their optimization in collaboration between a variety of observation data and application-specific data collected from users is also a crucial. In NICT Real World Information Analytics Project, an ICT platform called xData (cross-data) platform is constructed for developing smart applications with harnessing the above technologies toward realization of smart sustainable cities. For example, association discovery from a variety of meteorological and traffic data is performed to create and distribute a map that predicts various transport disturbance risks due to heavy rain, heavy snow and other abnormal weather conditions and to navigate safe, risk-free routes. The purpose of this talk is to introduce and discuss the latest challenges of novel methods and practices on smart data analytics and utilization.

Biography: Dr. Koji ZETTSU is a Director General of Big Data Integration Research Center of National Institute of Information and Communications Technology (NICT). He has been doing research and development of data analytics technology in NICT, and now leading Real Space Information Analytics Project since 2016 to implement smart data platform based on data mining and AI. For promoting industry-academia-government collaboration on the platform, he is also a leader of Cross-Data Collaboration Project of Smart IoT Acceleration Forum in Japan. Mr. Zettsu received Ph.D. in Informatics from Kyoto University in 2005. His research interests are databases, data mining, information retrieval and software engineering. He has serviced on numerous academic societies, conference committees and working groups. National Institute of Information and Communications Technology Tokyo, Japan.
INVITED TALK 4

The potential of Big Data Analytics in making sense of Urban Heat Island and Air Quality Monitoring data

Dr. Girish Agrawal (Professor, Jindal School of Art and Architecture (JSAA), O.P. Jindal Global University (JGU), NCR of Delhi)

Abstract: Deteriorating ambient air quality and increasing surface and near-surface air temperatures are worsening problems in urban areas. While ambient air quality (AAQ) is understood in absolute terms, as the suspended pollutant load in the breathable zone, surface and near-surface air temperatures are most commonly talked about in relative terms – as the difference in temperature between “urban” and “rural” areas, the so-called urban heat island (UHI) effect. Historically, AAQ and UHI have been studied as separate issues by climate scientists and geographers, respectively. Both are major urban environmental issues and of prime importance for urban planning, but have not been on the radar screen of urban planners, city engineers, and policy makers. This is largely due to the lack of good empirical data and models which could be used in the planning process. Ambient air quality and near-surface air temperatures are intricately linked and are both influenced by the form of the built-up area and the functioning of the population using those areas. In other words, both UHI and AAQ are influenced by various characteristics of the urban morphology. These characteristics include the shape, size and layout of the physical infrastructure, the form, and the population density, layout of streets, available modes of transport, land use and land cover, and spatial growth patterns, the function.

The purpose of this presentation is not to talk about an application of big data analytics (BDA) to modeling air quality or urban heat stress. But to look at AAQ and UHI data collected by the authors and their collaborators, and ask the question: How can we use the techniques of BDA to integrate large volumes of heterogeneous data and discover the linkages among a host of physical and economic, cultural and social parameters to extract the underlying patterns? Until recently, remotely sensed land surface temperature (LST) was the only reliable source for studying the UHI phenomena. Satellite technology provides spatially continuous observations of LST over large areas, but the urban configurations of large cities in India vary spatially and functionally, and given their rapid spatial growth and increasing population density, it has been difficult to understand and link the UHI to various urban morphology characteristics. Similarly, the concentration of particulate matter in air is affected by a large number of interlinked factors: the urban morphology, the local and regional land use patterns, and climate factors such as temperature, humidity, sunlight, and air movement. Modelling and understanding the contributions of these factors to air quality has not seen much success as it has been difficult to extract patterns from the available data which is both sparse, in terms of spatial coverage, and too much, in terms of sheer volume.
Data scarcity is no longer an issue. With inexpensive, ground-based, static and mobile sensors and automated systems for recording data from them becoming ubiquitous, we are now drowning in a sea of data riches. A determined researcher can access terabytes of data from their desk. But, ironically, this has made modelling even more difficult due to the sheer volume and heterogeneity of data. This is where the role of big data analytics comes in.

**Biography:** Girish Agrawal is a Professor in the School of Art & Architecture, O.P. Jindal Global University (JGU). His research interests are eclectic – he has authored and co-authored papers on applications of computations neural networks to geotechnical data analysis, road safety law and policy, earthwork optimization, public infrastructure, informal transport networks, public health, urban heat island effect, and urban development policy, among other areas. His current focus is on understanding urban morphology in three contexts: how roadway design and urban layout influences driver behavior (and so road safety); urban heat island effect; and air quality in urban and peri-urban areas. At present, Girish is working with researchers from four countries, doing large-scale data collection using networks of low-cost sensors, in their quest to study various aspects of natural and built systems.

Girish holds a B. Tech. (Hons) in Civil Engineering (1985) from IIT Delhi, an MS (1987) and a PhD (1992), both in Geotechnical Engineering, from Purdue University, and a Juris Doctor (2006) from the UC Berkeley School of Law. He is a licensed Civil Engineer, licensed Geotechnical Engineer, and a licensed Attorney at Law in California, and is also enrolled as an Advocate with the Bar Council of Delhi. Before coming to JGU, he was Professor and Head of the Department of Civil Engineering at Shiv Nadar University, where he also served as the Associate Director for Strategy and Planning for the School of Engineering. In the past, he has been a visiting member of the faculty at IIT Delhi and Purdue University.
INVITED TALK 5

Segment-search vs Knowledge Graphs: Making a Keyword Search Engine for Web Documents

Dr Subhash Bhalla (Professor, The University of Aizu, Japan)

Abstract: It is becoming increasingly popular to publish data on the web in the form of documents. Segment-search is a semantic search engine for web documents. It presents a query language. It is suitable for skilled and semi-skilled domain experts, who are adept at the use of a specific collection of documents. It returns suitable documents selected by using document fragments, that satisfy user's query. In contrast to knowledge graph approach, the technique is based on performing web page segmentation as per user perceived objects. Thus, it allows users' to query without the knowledge of complex query languages or learning about the data organization schemes. The proposed system is scalable and can cater to large scale web document sources.

Biography: Dr. Subhash Bhalla teaches at University of Aizu since 1993. His area of study is Distributed Information Systems. He completed studies up to PhD from IIT Delhi. He started teaching at School of Computer and Systems Sciences at JNU in 1986. After that, he briefly did Post-doctoral study at Sloan School of Management at MIT during 1987-88. His current interests include-standards in Electronic Health Records Databases, data modeling in Big data archives in Time-domain Astronomy, Polystore database systems and semantic keyword searches over knowledge graphs.
Web Data and Information Extraction

RESEARCH PAPER 1

Akanksha Mehndiratta and Krishna Asawa
Recent Advances and Challenges in design of Non-Goal Oriented Dialogue System

Abstract: Human being is pervasively surrounded by smart devices that provide numerous services to them. These devices are equipped with interfaces that are natural and intuitive with the aim of providing user an effortless and organic interaction with the devices. A dialogue agent is one such interface that interacts with the user in natural language. Recent paradigm classifies them as goal-oriented and non-goal-oriented dialogue systems. The aim of goal-oriented dialogue systems is to assist the user in completing a task. Evidently, the design of goal-oriented-dialogue agents has made a lot of progress. But the interactions with non-goal-oriented dialogue systems are reasonable, open-domain and more applicable to real-world applications. This paper reviews the state-of-the-art in Non-goal-oriented dialogue systems. The design of such systems has advanced due to Big Data hence most of the recent models are data-driven. This paper is a comprehensive study of data driven systems - advances in learning models and recent frontiers. Also provide an insight on datasets and evaluation methods and the limitations that the data-driven methods, datasets and evaluation methods present.

RESEARCH PAPER 2

Selma Khouri, Nabila Berkani, Ladjel Bellatreche, and Dihia Lanasri
Data Cube is Dead, Long Life to Data Cube in the Age of Web Data

Abstract: In a short time, the data warehouse (DW) technology took an important place in the academic and industrial landscapes. This place materialized in the large majority of engineering and management schools that adopted it in their curriculum and in the small, medium-size and large companies that enhanced their decision making capabilities thanks to it. The 1990s saw the advent of conferences such as DaWaK and DOLAP that carried the acronyms DW and OLAP in their titles. Then, all of a sudden, this technology has been upset by the arrival of Big Data. Consequently, those actors have replaced DW and OLAP by Big Data Analytics. We are well placed to assert that this brutal move may have a negative impact on schools, academia, and industry. This technology is not dead, today's context, with the connected world and Web of Data, is more favorable than when building DW merely stemmed from company internal sources. In this invited paper, we attempt to answer the following question: how does DW technology interact with Linked Open Data (LOD)? To answer the question, we provide a complete vision to augment the traditional DW with LOD, to capture and quantify the added value generated through this interaction. This vision covers the main steps of the DW life-cycle. This value is estimated through two different perspectives: (i) a source-oriented vision, by calculating the rate of the DW augmentation in terms of multidimensional concepts and instances, and (ii) a goal-oriented vision where the value is calculated according to the ability of the DW to estimate the performance levels of defined goals that reflect the strategy of a company, using the defined DW of the case study of a leading Algerian company.
RESEARCH PAPER 3

Vikram Singh and Mayank Dave
Improving Result Diversity using Query Term Proximity in Exploratory Search

Abstract: In the information retrieval system, relevance manifestation is pivotal and regularly based on document-term statistics, i.e. term frequency (tf), inverse document frequency (idf), etc. Query term proximity within matched documents is mostly under-explored. In this paper, a novel information retrieval framework is proposed, to promote the documents among all relevant retrieved ones. The relevance estimation is a weighted combination of document statistics and query term statistics, and term-term proximity is a simply aggregates of diverse user preferences aspects in query formation, thus adapted into the framework with conventional relevance measures. Intuitively, QTP is exploited to promote the documents for balanced exploitation-exploration, and eventually navigate a search towards goals. The evaluation asserts the usability of QTP measures to balance several seeking tradeoffs, e.g. relevance, novelty, result diversity (Coverage and Topicality), and overall retrieval. The assessment of user search trails indicates significant growth in a learning outcome.

RESEARCH PAPER 4

Meghna Chaudhary, Ravi Sharma and Sriram Chellappan
Pairing Users in Social Media via Processing Meta-data from Conversational Files

Abstract: Massive amounts of data today are being generated from users engaging on social media. Despite knowing that whatever they post on social media can be viewed, downloaded and analyzed by unauthorized entities, a large number of people are still willing to compromise their privacy today. On the other hand though, this trend may change. Improved awareness on protecting content on social media, coupled with governments creating and enforcing data protection laws, mean that in the near future, users may become increasingly protective of what they share. Furthermore, new laws could limit what data social media companies can use without explicit consent from users. In this paper, we present and address a relatively new problem in privacy-preserved mining of social media logs. Specifically, the problem here is the feasibility of deriving the topology of network communications (i.e., match senders and receivers in a social network), but with only meta-data of conversational files that are shared by users, after anonymizing all identities and content. More explicitly, if users are willing to share only a) whether a message was sent or received, b) the temporal ordering of messages and c) the length of each message (after anonymizing everything else, including usernames from their social media logs), how can the underlying topology of sender-receiver patterns be generated. To address this problem, we present a Dynamic Time Warping based solution that models the meta-data as a time series sequence. We present formal algorithm and interesting results in multiple scenarios wherein users may or may not delete content arbitrarily before sharing. Our performance results are very favorable when applied in the context of Twitter. Towards the end of the paper, we also present interesting practical applications of our problem and solutions. To the best of our knowledge, the problem we address and the solution we propose are unique, and could provide important future perspectives on learning from privacy-preserving mining of social media logs.
RESEARCH PAPER 5

Rajendra Kumar Roul
Deep Learning in the Domain of Near-Duplicate Document Detection

Abstract: Increasing of web users due to the popularity of the internet increases the digital documents on the web and among them many are duplicates and near-duplicates. Identifying duplicate and near-duplicate documents in a huge collection is a significant problem with widespread application and hence, detection and elimination of those documents are the need of the day. This paper proposes a technique to detect the near-duplicate documents on the web which has four main aspects: the first aspect is related to the selection of the important terms from a corpus of documents by developing a new correlation-based feature selection (CBFS) mechanism which enhances the performance of the classifier. The second aspect is to compute the similarity scores between each pair of documents of the corpus. The third aspect concerns with combining these similarities scores with the class label of each pair of documents to generate the feature vector for training the Multi-layer ELM (deep learning architecture) and other established classifiers and the fourth and final aspect introduces a heuristics method to rank the near-duplicate documents based on their similarity scores. The empirical results on DUC datasets witness the effectiveness of the proposed approach using Multi-layer ELM as highly appreciable compared to other state-of-the-arts classifiers including the deep learning classifiers.

Predictive Analytics in medical and agricultural domains

RESEARCH PAPER 6

Shubham Yadav, Sakthi Ganesh, Debanjan Das, Venkanna U, Rajarshi Mahapatra, Ankur Kumar Shrivastava, Prantar Chakrabarti, Asoke K Talukder. Suśruta

Artificial Intelligence and Bayesian Knowledge Network in Health Care – Smartphone Apps for diagnosis and differentiation of anemias with higher accuracy at Resource Constrained Point-of-Care settings

Abstract: Anemia in India carries a major disease burden. This includes both nutritional anemias, of which iron deficiency anemia (IDA) is the commonest and inherited hemolytic anemias like β thalassemias (β-TT). In Eastern and North-Eastern India about 25-40% of the public blood bank blood is consumed by patients with β Thalassemia. Moreover, 51% of Indian women in the reproductive age suffer from IDA. Most vulnerable group of anemic patients is in the rural underserved regions. This underserved population can only be served through the use of artificial intelligence (AI), automation, supported by telemedicine. To combat the problems of both IDA and β-thalassemia by early diagnosis at the point-of-care– we have developed Suśruta – an Artificial Intelligence (AI) driven robust smartphone-based health care application. This App uses five major components of AI; namely: (1) Natural language processing (NLP) to analyze the unstructured clinical data and translate it into computer understandable 3rd Generation SNOMED and ICD10 ontologies; (2) Speech Synthesis; (3) Artificial Neural Network (ANN) with Machine Learning
and Deep Learning (ML/DL) on 60,283 labelled common blood counts (CBC) and High Performance Liquid Chromatography (HPLC) data collected over 8 years by a teaching hospital in Kolkata for β-TT screening; (4) Computer Vision and Image Processing techniques to interpret hemoglobin content in blood through non-invasive analysis of conjunctiva and nailbed images; and (5) NoSQL and Big-data Graph database-driven Bayesian Knowledge Network for Evidence Based Medicine and Bayesian Outcome Tracing for Predictive Medicine. Unlike previous systems, the ML/DL technique of β-Thalassemia carrier screening with CBC improved the accuracy of screening by two folds compared similar approaches analyzing CBC with Mentzer Index. Moreover, the uniqueness of Suṣruta is that the App is robust and works both in an offline and online mode at resource constrained point-of-care. This is the first time AI is used for comprehensive anemia care by early diagnosis, which empowers the ordinary health workers in rural underserved communities. Furthermore, it will introduce the concept of patient empowerment and person centered care by changing the definition of point-of-care in rural India.

**RESEARCH PAPER 7**

**Swapnil Shrivastava, Supriya N Pal and Ranjeet Walia**

Market Intelligence for Agricultural Commodities using Forecasting and Deep Learning Techniques

**Abstract:** About two third of the Indian population continues to live in villages and depends on agriculture as main source of livelihood. The agrarian distress that took decades to build up, surfaced in the form of decline in the proportion of rural population and increase in the number of farmer’s suicide cases. There is a strong hypothesis that in addition to production techniques, the widespread availability of Market Intelligence would bring substantial improvement in financial condition of farmers. Hence there is a need for comprehensive system, which would fetch, interlink, transform and analyze relevant data from various ministries / departments /organizations spread across the country to generate precise, appropriate and timely Market Intelligence. We took a step in this direction by design and implementation of Market Intelligence System Proof of Concept (PoC) using available datasets for few agricultural commodities. This PoC takes daily market price and weather data as input, transforms it into information and generates actionable intelligence by applying forecasting and deep learning techniques. The system provides trend analysis, short term as well as long term commodity price prediction and market selection as insights for farmers. The Auto Regressive Integrated Moving Average (ARIMA) forecasting technique and Recurrent Neural Network (RNN) deep learning techniques are applied for short term and long term agricultural commodity price prediction respectively. The study results demonstrate intended utility of forecasting and deep learning techniques for generating Market Intelligence System. The paper concludes with benefits of comprehensive Market Intelligence system, challenges and future work.
Graph Analytics

RESEARCH PAPER 8

Philippe Fournier-Viger, Chao Cheng, Jerry Chun-Wei Lin, Unil Yun, and R. Uday Kiran

TKG: Efficient Mining of Top-K Frequent Subgraphs

Abstract: Frequent subgraph mining is a popular data mining task, which consists of finding all subgraphs that appear in at least minsup graphs of a graph database. An important limitation of traditional frequent subgraph mining algorithms is that the minsup parameter is hard to set. If set too high, few patterns are found and useful information may be missed. But if set too low, runtimes can become very long and a huge number of patterns may be found. Finding an appropriate minsup value to find just enough patterns can thus be very time-consuming. This paper addresses this limitation by proposing an efficient algorithm named TKG to find the top-k frequent subgraphs, where the only parameter is k, the number of patterns to be found. The algorithm utilizes a dynamic search procedure to always explore the most promising patterns first. An extensive experimental evaluation shows that TKG has excellent performance and that it provides a valuable alternative to traditional frequent subgraph mining algorithms.

RESEARCH PAPER 9

Moumita Chatterjee, Anirban Mitra, Sudipta Roy , Somasis Roy, Hirav Shah, Sanjit Kr. Setua

Gossip Based Distributed Real Time Task Scheduling with Guaranteed Performance on Heterogeneous Networks

Abstract: This paper considers the scheduling of distributable real time tasks in dynamic networks which are prone to failures and do not have a fixed network infrastructure. We propose a distributable scheduling algorithm using gossip called GBTS-F for reliable and dynamic discovery of appropriate nodes which can execute the tasks. GBTS-F uses the slack time of the tasks for optimizing the gossiping duration and thus satisfies the each task timing constraints with probabilistic guarantee. Even though gossip protocols are usually fault tolerant but to handle byzantine faults and to control the high message complexity incurred during gossiping we propose to use an expander graph. Performance analysis and simulation results show that GBTS-F performs better than other state of art algorithms in terms of message complexity and task success probability.
Sanchita Basak, Fangzhou Sun, Saptarshi Sengupta and Abhishek Dubey
Data-Driven Optimization of Public Transit Schedule

Abstract: Bus transit systems are the backbone of public transportation in the United States. An important indicator of the quality of service in such infrastructures is on-time performance at stops, with published transit schedules playing an integral role governing the level of success of the service. However there are relatively few optimization architectures leveraging stochastic search that focus on optimizing bus timetables with the objective of maximizing probability of bus arrivals at timepoints with delays within desired on-time ranges. In addition to this, there is a lack of substantial research considering monthly and seasonal variations of delay patterns integrated with such optimization strategies. To address these, this paper makes the following contributions to the corpus of studies on transit on-time performance optimization: (a) an unsupervised clustering mechanism is presented which groups months with similar seasonal delay patterns, (b) the problem is formulated as a single-objective optimization task and a greedy algorithm, a genetic algorithm (GA) as well as a particle swarm optimization (PSO) algorithm are employed to solve it, (c) a detailed discussion on empirical results comparing the algorithms are provided and sensitivity analysis on hyper-parameters of the heuristics are presented along with execution times, which will help practitioners looking at similar problems. The analyses conducted are insightful in the local context of improving public transit scheduling in the Nashville metro region as well as informative from a global perspective as an elaborate case study which builds upon the growing corpus of empirical studies using nature-inspired approaches to transit schedule optimization.
Pattern Mining

RESEARCH PAPER 11

P.P.C Reddy, R. Uday Kiran, Koji Zettsu, Masashi Toyoda, P. Krishna Reddy, and Masaru Kitsuregawa

Discovering Spatial High Utility Frequent Itemsets in Spatiotemporal Databases

Abstract: Spatial High Utility Itemset Mining (SHUIM) aims to discover all itemsets in a spatiotemporal database that satisfy the userspecified minimum utility (minUtil) and maximum distance (maxDist) constraints. The popular adoption and successful industrial application of SHUIM suffers from the following two limitations: (i) Since SHUIM determines the interestingness of an itemset without taking into account its support within the data, SHUIM facilitates sporadic itemsets with high utility to be generated as SHUIs. In particular, items in long transactions can combine with each other and be generated as SHUIs. (ii) SHUIM is a computationally expensive process because the generated itemsets do not satisfy the downward closure property. This paper introduces Spatial High Utility Frequent Itemset Mining (SHUFIM) to address these two issues. A SHUI in a spatiotemporal database is said to be a SHUFI if its support is no less than the user-specified minimum support (minSup) constraint. The usage of minSup not only facilitates the proposed model to be tolerant to the long transactions within the data but also facilitates us to employ additional pruning techniques to reduce the computational cost. A single scan fast algorithm has also been proposed to discover all SHUFIs in a spatiotemporal database. Experimental results demonstrate that the proposed algorithm is efficient. We also demonstrate the usefulness of the proposed model with a real-world application.

RESEARCH PAPER 12

Jui Mhatre, Harsha Agrawal, and Sumit Sen

Efficient Algorithms For Flock Detection in Large Spatio-Temporal Data

Abstract: Increasing availability of location-based applications and sensor devices have necessitated quicker analysis of moving object data streams in order to identify patterns. The efficiency of currently available algorithms used in pattern detection is not adequate to handle large scale data streams that are increasingly available. We focus on the particular problem of flock detection in moving object data and our goal is to detect flocks quickly and using fast algorithms. Firstly, we employ a triangular grid to reduce the search space of clustering algorithms which has a significant effect in case of dense objects. As a second step, we implement a modified flock membership function and pipeline creation that ensures better memory and time performance during cluster detection. We show that this refinement also improves the rate of flock detection. Finally, we parallelize our algorithm to further enhance the handling of massive data streams. Based on an extensive empirical evaluation of these algorithms across a variety of moving object data sets, we show that our method is significantly faster than the existing comparable methods over sliding windows. In particular, it requires lesser time to identify flocks and is 2-4 times faster thus confirming the efficiency and effectiveness of our approach.
RESEARCH PAPER 13

Xu Teng, Prabin Giri, Matthew Dwyer, Jidong Sun, and Goce Trajcevski

Local Temporal Compression for (Globally) Evolving Spatial Surfaces

Abstract: The advances in the Internet of Things (IoT) paradigm have enabled generation of large volumes of data from multiple domains, capturing the evolution of various physical and social phenomena of interest. One of the consequences of such enormous data generation is that it needs to be stored, processed and queried along with having the answers presented in an intuitive manner. A number of techniques have been proposed to alleviate the impact of the sheer volume of the data on the storage and processing overheads, along with bandwidth consumption and, among them, the most dominant is compression. In this paper, we consider a setting in which multiple geographically dispersed data sources are generating data streams, however, the values from the discrete locations are used to construct a representation of continuous (time-evolving) surface. We have used different compression techniques to reduce the size of the raw measurements in each location, and we analyzed the impact of the compression on the quality of approximating the evolution of the shapes corresponding to a particular phenomenon. Specifically, we use the data from discrete locations to construct a TIN (triangulated irregular networks), which evolves over time as the measurements in each locations change. To analyze the global impact of the different compression techniques that are applied locally, we used different surface distance functions between raw-data TINs and compressed data TINs. We provide detailed discussions based on our experimental observations regarding the corresponding (compression method, distance function) pairs.

RESEARCH PAPER 14

Hai Duong, Tin Truong, Bac Le and Philippe Fournier-Viger

An Explicit Relationship between Sequential Patterns and their Concise Representations

Abstract: Mining sequential patterns in a sequence database (SDB) is an important and useful data mining task. Most existing algorithms for performing this task directly mine the set of all frequent sequences in an SDB. However, these algorithms often exhibit poor performance on large SDBs due to the enormous search space and cardinality of frequent sequences. In addition, constraint-based mining algorithms relying on this approach must read an SDB again when a constraint is changed by the user. To address this issue, this paper proposes a novel approach for generating frequent sequences from the two sets of frequent closed sequences and frequent generator sequences, which are concise representations of sequences. The proposed approach is based on a novel explicit relationship between frequent sequences and these two sets. This relationship is the theoretical basis for a novel efficient algorithm named GFS-CR that directly enumerates frequent sequences from frequent closed sequences and frequent generator sequences rather than mining them from an SDB. Experimental results show that GFS-CR outperforms state-of-the-art algorithms in terms of runtime and scalability.
Machine Learning

RESEARCH PAPER 15

Debasmita Dey and Pradeep Kumar
A novel approach to identify the determinants of online review helpfulness and predict the helpfulness score across product categories

Abstract: The proliferation in the number of available online reviews provides an excellent opportunity to use this accumulated enormous information of any product in a more strategic way to improve the quality of the product and services of the e-commerce company. Due to the non-uniform quality of online reviews, it is crucial to identify those helpful reviews from the pile of a large amount of low quality and low informative other reviews. This system will help the customers to form an unbiased opinion quickly by looking at its level of helpfulness. The e-commerce companies measure the helpfulness of a review using the number of votes it gets from other customers. This situation arises problems to newlyauthored potentially helpful reviews due to lack of votes. Thus it is essential to have an automated process to estimate and predict helpfulness of any review. This paper identifies the essential characteristics of online reviews influencing the helpfulness of it. This study categorized all characteristics of reviews collected from previous literature in four main categories and then study the combined effect of the four aspects in predicting the helpfulness of a review. The product type (Search or Experience) acts as a control variable in the factors identification model of helpful prediction of a review. An analysis of total 14782 reviews from Amazon.com across five different product category shows the factors influencing the helpfulness of a review varies across product categories. Then a comparative study of two widely used machine learning, Artificial Neural Network and Multiple Adaptive Regression Spline are presented to predict the helpfulness of online review across five different categories and a better method of predicting helpfulness of online reviews are suggested based on the type of product. This study solves the starvation problem of potential newly-authored or infamous reviews without any manual votes along with high accuracy of helpfulness prediction.

RESEARCH PAPER 16

Mohit Gupta and Pulkit Mehndiratta
Analysis and Recognition of Hand-drawn Images with Effective Data Handling

Abstract: Email is the most frequently used web application for communication and collaboration due to its easy access, fast interactions, and convenient management. More than 60% of the email traffic constitutes business to consumer (B2C) emails (e.g., flight reservations, payment reminder, order confirmations, etc.). Most of these emails are generated by filling a template with user or transaction specific values from databases. In this paper we describe various algorithms related to extracting important information from these emails. Unlike web pages, emails are personal and due to privacy and legal considerations, no other human except the receiver can view them. Thus, adapting extraction techniques used for web pages, such as HTML wrapper-based techniques, have privacy and scalability challenges. We describe end-to-end information extraction system for emails data collection, anonymization, classification, building the information extraction models, deployment, and monitoring.
To handle the privacy and scalability issues, we focus on algorithms which can work with minimum human annotated samples for building classifier and extraction techniques. Similarly, we present algorithms to minimize samples for human inspection to detect precision and recall gaps in the extraction pipeline.

**RESEARCH PAPER 17**

**Kalpdrum Passi and Sandipgiri Goswami**

Real Time Static Gesture Detection Using Deep Learning

**Abstract:** Sign gesture recognition is an important problem in human-computer interaction with significant societal influence. However, it is a very complex task, since sign gestures are naturally deformable objects. Gesture recognition contains unsolved problems for the last two decades, such as low accuracy or low speed, and despite many proposed methods, no perfect result has been found to explain these unsolved problems. In this paper, we propose a deep learning approach to translating sign gesture language into text. In this study, we have introduced a self-generated image data set for American Sign language (ASL). This dataset is a collection of 36 characters containing A to Z alphabets and 0 to 9 number digits. The proposed system can recognize static gestures. This system can learn and classify specific sign gestures of any person. A convolutional neural network (CNN) algorithm is proposed for classifying ASL images to text. An accuracy of 99% on the alphabet gestures and 100% accuracy on digits was achieved. This is the best accuracy compared to existing systems.

**RESEARCH PAPER 18**

**Himangi Mittal, Ajith Abraham, and Anuja Arora**

Interpreting Context of Images using Scene Graphs

**Abstract:** Understanding a visual scene incorporates objects, relationships, and context. Traditional methods working on an image mostly focus on object detection and fail to capture the relationship between the objects. Relationships can give rich semantic information about the objects in a scene. The context can be conducive in comprehending an image since it will help us to perceive the relation between the objects and thus, give us a deeper insight into the image. Through this idea, our project delivers a model which focuses on finding the context present in an image by representing the image as a graph, where the nodes will the objects and edges will be the relation between them. The context is found using the visual and semantic cues which are further concatenated and given to the Support Vector Machines (SVM) to detect the relation between two objects. This presents us with the context of the image which can be further used in applications such as similar image retrieval, image captioning, or story generation.
TUTORIALS

Tutorial 1: Data for AI | Duration: 3 hours
Dr Anush Sankaran, Dr Hima Patel, and Sameep Mehta, IBM Research – India

Abstract: It is well understood in the literature that the performance of an AI model is upper bounded by the quality of the data. While researchers and practitioners have focused on improving the quality of the model (such as neural architecture search and automated feature selection), there is limited amount of efforts towards improving the data quality. To overcome this challenge, we propose the “Data for AI” - a research agenda around data understanding, data quality measurement, data cleaning and preparation, and making it ready for the corresponding upstream AI application. The aim of this tutorial is to talk about tools and techniques to make data ready for an AI application. We will discuss different dimensions along which the quality of data for AI can be measured such as: Data Trust, Data Richness, Data Labels and Data Cleanliness. Specifically, we will deep dive into three specific issues in terms of data quality:

• Data Richness: This problem encompasses the fundamental question of whether “Is the given data is rich enough to perform analytics or classification using an AI model?”. In plenty of industry settings, this is the problem that is being addressed by domain experts and data scientists. Here, we aim to automatically analyze and provide a score and explanation for a provided dataset, based on its richness value, even before the data science lifecycle starts. This can allow a data scientist to either ask for additional data or trim unnecessary data.

• Noisy Data Labels: In many largescale classification problems, labelled data is often obtained through crowd sourcing or by automatically using information on the web. These systems are prone to human labelling errors, measurement errors and subjective biases of labelers. We discuss about methods for detection and cleaning of label noise present in the data.

• AutoLabeling of Documents: Most of the data that we collect or gather through sensors/ internet are unlabeled. With minimal SME effort, we can get a few samples of them labelled. “How to label the rest of the large volume of unlabelled data?”. The aim of this research work is to automatically tag and extract keywords from documents scraped through web, so that the documents could be intelligently indexed. The enhanced indexing provides better results for any kind of upstream search and retrieve applications as well as question answering applications.

Biography:

Anush Sankaran is a researcher with IBM Research AI. He completed his Ph.D. degree from IAB lab IIIT Delhi in the area of biometrics and machine learning under the guidance of Dr. Mayank Vatsa and Dr. Richa Singh. His research interests include deep learning, image processing, human cognition, and their applications. He is now primarily leading the technical efforts in project DatAI - a completely automated solution for preparing complex data to be consumed by AI applications.

Prior to that he has successfully led and technically delivered projects such as IBM Deep Learning IDE and Machine Learning for Creativity. He was a recipient of the TCS Ph.D. Research Fellowship from 2010 to 2015. He has written more than 25 peer-reviewed conferences and journals and also has the Best Poster Awards in the IEEE BTAS 2013 and the IEEE IJCB 2014. Anush received the B.Tech. degree (Gold Medal) in computer science from the Coimbatore Institute of Technology, Coimbatore, India, in 2010.
**Hima Patel** is a research manager at IBM Research Lab and leads a team of researchers that work in problems at the intersection of machine learning and NLP. Her passion is to work on research areas that help solve real business problems. Prior to IBM, she has spent time in research groups at Shell, GE and brief stint at Visa where she has worked on varied research problems spanning from object detection in medical images to anomaly detection from multivariate sensor data collected from machines. She is also actively involved with the organization of Grace Hopper India conference for last four years and is also an executive committee member for IEEE Computer Society, Bangalore Section. She holds a B.E. in Computer Science from GCET and M Tech in Information and Communication Technology from DA-IICT.

**Sameep Mehta** is a senior manager and STSM at IBM Research India. He joined IBM Research after completing his Ph.D. from The Ohio State University in 2006. His current research interests are in AI & Machine Learning, He spends lot of time to develop algorithms and systems for data lifecycle management.

**Tutorial 2: Streaming Data Processing | Duration: 1.5 hours**  
Dr Prasad Despande, Co-founder, Alpha Analyst

**Abstract:** In this tutorial we will introduce Stream data processing. We will cover the general principles of designing a scalable streaming data processing engine. We will look at the different kind of guarantees on data processing that different stream engines give. We will delve into a couple of streaming platforms such as Apache Storm and Spark Streaming in more detail. Further, we will compare some popular stream processing engines on various dimensions. At the end of the tutorial, the audience should have a fair idea of Stream processing concepts and be able to make an informed choice while picking the right engine for their use case.

**Biography:** Prasad’s expertise lies in data management and analytics, specifically in database internals, data integration, OLAP, data mining and text analytics. He received a B. Tech in Computer Science and Engineering from IIT, Bombay and MS and PhD degrees in Database Systems from the University of Wisconsin, Madison. He is an ACM Distinguished Scientist and a former member of the IBM Academy of Technology. Currently he is working as VP Engineering and Co-founder at Cube Corp, that aims to bring automation and intelligence to enterprise application testing. Prior to this, he founded Alpha Analyst and has worked as CTO at IQLECT, leading the technical team to deliver the first version of their streaming data analytics product. Before that he spent 12 years at IBM Research, where he was a Senior Technical Staff Member and Manager of the Information Analytics group. He has more than 45 publications in reputed conferences and journals and 35 patents issued. He has served on the Program Committees of many conferences, as a PC member and PC Co-Chair.
**Tutorial 3:** Disrupting Healthcare Industry with Integration of Blockchain, AI and Analytics | **Duration:** 3 hours

Dr Mayuri Mehta, SCET, Surat and Dr Anuj Kumar Garg, IBM India

**Abstract:** Healthcare industry across the world has progressed very rapidly over last two decades. It is likely to rise 5.4% annually to just over $10 trillion. The newer research and innovations in medical domain have helped increasing the life expectancy in men from 51 to 66 years and from 55 to 71 in women in last three decades. However, the healthcare industry is behind other sectors in adopting the newer IT technologies. This tutorial focuses on imbibing the latest technologies, specifically Artificial Intelligence (AI) and Blockchain, with medical domain innovations.

Blockchain has emerged from a conventional crypto world to a business world with many emerging solutions in Healthcare domain. The Blockchain represents the immutable and append only database that is written post consensus from each party who is participant in the network. The primary scope of Blockchain in this proposal will focus on storing electronic medical records securely over the Blockchain and using the date for auto claim settlements, medical research and providing support to AI engines to quick and better diagnosis.

Healthcare using AI is amongst the fastest growing research area across the globe. AI along with big data can improve and reduce the cost of healthcare. A massive amount of heterogeneous data generated in healthcare sector offers opportunities for big data analytics. Such analysis transforms big data into real and actionable insights to healthcare practices, thus provide new understanding and ways for better and quicker treatment. Due to increasing availability of electronic healthcare data (structured as well as unstructured data) and rapid progress of analytics techniques, a lot of research is being carried out in this area.

The first session of tutorial focuses on advent of blockchain in healthcare. In this session, we will discuss some practical applications of blockchain technology in record sharing, auto claim settlements, incentivizing patients and providing authentic information to the research organizations. During the second session of tutorial, we shall discuss the significance of artificial intelligence for healthcare sector. Next, we will demonstrate two AI- powered healthcare applications. The discussion of each application will include precise problem statement, proposed solution, data collected & used, experimental analysis and challenges faced to achieve this success. Finally, we will discuss why do we need explanations, what are the explanations and what is the future of Explainable AI (XAI) in healthcare.

**Biography:** Dr. Mayuri is currently working as a Professor and PG In-Charge of the Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Surat. She received doctorate in Computer Engineering from NIT, Surat. She has 19 years of teaching experience including 10 years of research experience. Her areas of teaching and research include Data Science, Machine Learning/Deep Learning, Computer Algorithms, Python Programming and Research Methodology. More than 25 publications in National/International Conference and Journal are credited to her, which reflect her research outlook. She has received approval from Indian Council of Medical Research (ICMR) for the project titled “Craniofacial Measurements: Facial Index and Nasal Index of Population of South Gujarat and its Relevance in Reconstructive Surgery”. She has also received funds several times from Gujarat Council on Science and Technology (GUJCOST) for organizing various workshops. She has conducted workshop/tutorial in several prestigious International conferences. She has also delivered expert talks and conducted hands-on sessions in STTPs/FDPs. She has served as a Session Chair and as a Member, Advisory Committee/Technical Program Committee for several International Conferences. With the noble intention of applying her technical knowledge for societal impact, she is working on AI-powered and healthcare data driven projects in association with the doctors doing private practice and the doctors of Medical Colleges and their Local Research Units (LRU). Her completed healthcare projects are ‘A Deep Learning based Craniofacial Distance Measurements for Facial Reconstructive Surgery’, ‘An Approach to Detect Dry Eye Disease using Deep Learning’ and ‘Racial Categorization of Population of South Gujarat using Deep Learning’.
Tutorial 4: Knowledge Graphs: Representation, Management and Applications
Duration: 1.5 hours
Dr Raghava Mutharaju, IIIT Delhi and Dr. Srikanta Bedathur, IIT Delhi

Abstract: Knowledge Graphs (KG) are graph structures that capture knowledge in the form of entities, relationships between them, and additional information including provenance and context of relationships. They are now extensively used by several enterprises across many different domains such as the Web, e-commerce, healthcare, geoscience, manufacturing, aviation, power, oil and gas. Private KGs are used by commercial enterprises such as Google, Microsoft, Yahoo, LinkedIn, Amazon, eBay, GE, etc., and on the other hand, DBpedia, Wikidata, YAGO, NELL etc., are available publicly. Semantic Web technologies such as RDF (Resource Description Framework), and OWL (Web Ontology Language), which are also W3C standards, play a crucial role in the representation of KGs. In this introductory tutorial on KGs, we will discuss RDF and how they are used to represent a KG, how one can store and operate over knowledge graphs at scale, and finally present a few important applications of knowledge graphs including their uses in information retrieval and NLP for entity linking, entity retrieval, question answering, and conversational systems.

Biography: Raghava Mutharaju is an Assistant Professor in the Computer Science and Engineering department of IIIT-Delhi, India. He got his PhD in Computer Science and Engineering from Wright State University, Dayton, OH, USA, in 2016 working under the supervision of Prof. Pascal Hitzler. He has worked in Industry research labs such as GE Research, IBM Research, Bell Labs, and Xerox Research. His research interests are in various aspects of Semantic Web such as knowledge graph construction, ontology modeling, reasoning, querying, and its applications to specific domains. He is also interested in Big Data applications. He has published at several venues such as ISWC, ESWC, ECAI, and WISE. He co-organized workshops at WWW 2019, WebSci 2017, ISWC 2015 and tutorials at ISWC 2019, IJCAI 2016, AAAI 2015 and ISWC 2014. He is/has been on the Program Committee of several (Semantic) Web conferences such as WWW, ISWC, ESWC, K-CAP and SEMANTiCS.

Tutorial 5: Large-Scale Information Extraction from Emails under Data Constraints
Duration: 1.5 hours
Dr. Rajeev Gupta and Ranganath Kondapally Microsoft R&D, India

Abstract: In this tutorial, we describe various information extraction techniques in data constrained environments. We focus on extracting information from emails, specifically business-to-consumer (B2C) emails (e.g., order confirmations, flight reservations, etc.). Unlike web pages, emails are personal and due to privacy and legal considerations, no other human except the receiver can view them. Thus, adapting extraction techniques used for web pages, such as HTML wrapper-based techniques, have privacy and scalability challenges. We describe end-to-end information extraction system for emails—data collection, anonymization, classification, building information extraction models, deployment, and monitoring. We particularly focus on various extraction algorithms which either require just a small set of donated (original) emails or work well with (larger) anonymized data. Besides describing efficient anonymization techniques, we also describe several machine learning techniques where different textual and visual features from emails are used. We conclude the tutorial by outlining open research problems in data-constrained information extraction.
Biography: Rajeev Gupta is a principal researcher at Microsoft, AI & Research, India. He got his PhD from Indian Institute of Technology (IIT) Bombay in Computer Science. He has more than 25 publications and 20 patents in the areas of data management, information extraction, and distributed computing in reputed conferences and journals such as TKDE, ICDE, VLDB, WWW, SIGMETRICS, CIKM, KDD, etc. He is currently working in applying AI for information extraction and search.

Ranganath Kondapally is a senior researcher at Microsoft, AI & Research, India. He got his PhD in Computer Science from Dartmouth College in the area of proving bounds for certain class of streaming algorithms. His areas of interest include information extraction, machine learning algorithms, and complexity theory. He has numerous conferences, journal publications, and patents in his name in the areas of information extraction, streaming algorithms, and virtual reality. Currently, he is working on information extraction and inferencing problems on big-data, powering delightful personal assistant experiences.

**Tutorial 6: Big Data Analysis Using Multilayer Networks**  
**Duration:** 3 hours  
Dr Sharma Chakravarthy, Professor, University of Texas at Arlington

Abstract: In this tutorial, we argue that graph analysis techniques are extremely important and is receiving renewed attention as the data sets are becoming complex and their sizes are increasing. There is also renewed interest in answering graph queries in both exact and approximate way due to the presence of large graph based such as FreeBase and entity-based very large graphs. Although aggregate analysis techniques (such as communities, hubs, subgraphs, etc.) exist for single and simple graphs, extending them to attribute graphs is not easy. As an alternative, multilayer Networks (or MLNs) are being explored and new analysis approaches are being developed. In this tutorial, we start with single graph analysis, establish their limitations for complex data sets analysis and introduce MLNs. We will discuss the notions of communities and hubs and their relevance to aggregate analysis. We will discuss extending these notions to multilayer networks, challenges, and current approaches and solutions. We will apply the techniques discussed to a number of case studies to appreciate their need, utility, applicability, efficiency, and scalability.

Biography: Professor Chakravarthy is an ACM Distinguished Scientist and Distinguished speaker. He is also an IEEE Senior Member. He is also a Fulbright specialist. He organized (General Co-Chair) the 13th international Conference on Distributed Event-Based Systems (DEBS 2013). He has spent several summers at the Rome Air Force Research Laboratory (AFRL) as a Faculty Fellow working in continuous query processing over fault-tolerant networks and video stream analysis. Sharma Chakravarthy is Professor of Computer and Engineering Department at The University of Texas at Arlington, Texas. He established the Information Technology Laboratory at UT Arlington in Jan 2000 and currently heads it. Sharma Chakravarthy has also established the NSF funded, Distributed and Parallel Computing Cluster (DPCC@UTA) at UT Arlington in 2003. He is the recipient of the university-level “Creative Outstanding Researcher” award for 2003 and the department level senior outstanding researcher award in 2002. He is well known for his work on stream data processing, semantic query optimization, multiple query optimization, active databases (HiPAC project at CCA and Sentinel project at the University of Florida, Gainesville), and more recently scalability issues in graph mining, social network analysis, and graph analysis of multilayered networks. His group at UTA is currently adapting map/reduce and other paradigms for scaling graph mining algorithms to very large graphs and for answering graph queries. He has applied machine learning techniques to rank answers, identify general- and topic-based experts in a Question-Answer (or Q-A) social network. His work on InfoSift – a classification system for text, email, and web – has used graph mining techniques.
His current research includes big data analysis using multi-layered networks, stream data processing for disparate domains (e.g., video analysis), scaling graph mining algorithms for analyzing very large social and other networks, active and real-time databases, distributed and heterogeneous databases, query optimization (single, multiple, logic-based, and graph), and multi-media databases. He has published over 200 papers/book chapters in refereed international journals and conference proceedings. He has supervised 15+ PhD theses and 90+ MS thesis. He has given tutorial on a number of database topics, such as graph mining, active, real-time, distributed, object-oriented, and heterogeneous databases in North America, Europe, and Asia. He is listed in Who's Who Among South Asian Americans and Who's Who Among America's Teachers.

Prior to joining UTA, he was with the University of Florida, Gainesville. Prior to that, he worked as a Computer Scientist at the Computer Corporation of America (CCA) and as a Member, Technical Staff at Xerox Advanced Information Technology, Cambridge, MA. Sharma Chakrvarthy received the B.E. degree in Electrical Engineering from the Indian Institute of Science, Bangalore and M.Tech from IIT Bombay, India. He worked at TIFR (Tata Institute of Fundamental Research), Bombay, India for a few years. He received M.S. and Ph.D degrees from the University of Maryland in College park in 1981 and 1985, respectively.

**PANEL**

**Big Data Analytics is different from AI**

**Moderator:** Dr. Anirban Mondal, Ashoka University

**Description:** Currently, there is a research trend towards considering both Big Data Analytics (BDA) and Artificial Intelligence (AI)/ Machine Learning (ML) as part of the same research area. Preceded by database research, during the last few decades, the field of data mining has emerged for extracting interesting patterns from large amounts of data. Subsequently, due to explosive growth of data, BDA has emerged based on the requirement of processing of large databases by considering the 5Vs of Big Data as well as aspects of NoSQL. On the other hand, starting from rule-based engines, the areas of AI and ML have emerged for facilitating automation as well as for improving the accuracy of predictions by learning from massive amounts of data. Although historically BDA and AI/ML have evolved differently, large-scale systems, such as Google, Facebook and Amazon, employ both BDA and AI/ML. The goal of this panel discussion is to understand the ramifications of whether BDA and AI/ML would essentially converge as a single area or will they grow as separate areas in their own right. How would such ramifications impact students, researchers and practitioners?

**Questions:**

1. Should CS students learn theory and skills related to both BDA and ML?
2. Should researchers work across both BDA and ML or specialize in any one of these areas?
3. In the future, will the industry have separate roles for BDA and ML specialists?
4. From a long-term perspective, do you see BDA and ML converging as a single research area or will they grow independently?