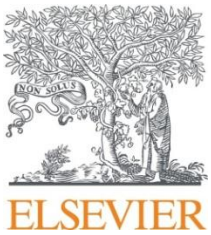
2015 IEEE International Conference on Big Data

October 29-November 1, 2015 • Santa Clara, CA, USA

Sponsored by



ODBMS.ORG
The Resource Portal for Big Data and New Data Management Technologies.
http://www.odbms.org

2015 IEEE International Conference on Big Data

IEEE Big Data 2015 Program Schedule

Santa Clara, CA, USA
October 29—November 1, 2015

Keynote Lecture: ***60 minutes*** (about 45 minutes for talk and 15 minutes for Q and A)
Main conference regular paper: ***25 minutes*** (about 20 minutes for talk and 5 minutes for Q and A)
Main conference short paper: ***15 minutes*** (about 11 minutes for talk and 4 minutes for Q and A)

All conference activities take place at the Hyatt Regency Santa Clara.

Wednesday, 28-October	
4:00 – 8:00 pm Location:	**Registration** *Lobby West*

Thursday, 29-October			
7:30-6:00 pm Location:	**Registration** *Lobby West*		
10:00-10:20 am and 3:30 – 3:50 pm Location:	**Coffee Break** *Grand Ballroom Foyer/Courtyard*		
2:00 – 6:00 pm Location:	**Poster Session (Set up only)** *Grand Ballroom Foyer/Corridor*		
Time	**Session/Workshops**	**Session Chair**	**Location**
Special Session I 8:00 – 6:30 pm	From Data to Insight: Big Data and Analytics for Advanced Manufacturing Systems	Sudarsan Rachuri Ronay Ak Tina Lee Rumi Ghosh Srinivasan Soundar Steve Eglash Sagar Kamarthi	*Ballroom A*
Full Day Workshops 8:00 – 6:30 pm	Leveraging High Performance Computing Resources for Managing Large Datasets (Hands on)	Ritu Arora	*Ballroom B*
	Scalable Cloud Data Management (SCDM)	Felix Gessert, Norber Ritter	*Ballroom C*
	Advances in Software and Hardware for Big Data to Knowledge Discovery (ASH)	Weijia Wu	*Ballroom D*
	Mining Big Data in Social Networks	Bin Zhou, Jian Pei	*Ballroom E*
	Big Data in the Geosciences	Tom Narock, Marshall Ma, Peter Baumann	*Ballroom G*

Thursday, 29-October - continued			
Time	**Sessions/Workshops**	**Session Chair**	**Location**
Sessions 8:00 – 12:00 pm	Intelligent Mining	Uraz Yavanoglu	*Napa*
	Big Data and the Humanities	Mark Hedges	*Ballroom F*
	Introducing the NSF Big Data (BD) Reginal Innovation Hubs	Ashok Krishnamurthy	*Sonoma*
	Advances in High Dimensional Big Data	Sotiris Tasoulis	*Ballroom H*
	Mining Big Data to Improve Clinical Effectiveness	Doug Talbert,, Bill Eberle, Russ Waitman, Mei Liu	*Bayshore*
	Deriving Value from Big Data in Healthcare	Vahid Taslimi	*Mendocino*
12:00 - 1:30 pm	**Lunch (on own)**		
Time	**Sessions/Workshops/Tutorials**	**Session Chair**	**Location**
Tutorial 1:30 – 6:00 pm	Tutorial 1: Optimization Big Data Analytics on Heterogeneous Processors (1:30-3:30 pm)	Mayank Daga, Mauricio Breternitz, Junli Gu	*Ballroom F*
	Tutorial 2: The Era of Big Spatial Data (4:00 -6:00 pm)	Mohamed F. Mokbel, Ahmed Eldawy	*Ballroom F*
Sessions 1:30 – 6:30 pm	Big Data Methodologies and Tools to Improve Big Data Projects	Jeffrey S Saltz	*Sonoma*
	Joint Program Schedule: First Workshop on Data-Centric Infrastructure for Big Data Science & 3rd Workshop on Distributed Storage Systems and Coding for Big Data	Claris Castillo, Bing Zhu	*Ballroom H*
	Big Data for B2B Marketing and Sales	Lei Tang, Shipeng Yu, Hui Xiong	*Bayshore*
	Data and Computational Science Technologies for Earth Science Research	Dan Crichton	*Napa*
	Privacy and Security of Big Data	Alfredo Cuzzocrea	*Mendocino*

Friday, 30-October			
7:30-6:00 pm Location:	**Registration** *Lobby West*		
Time	**Sessions**	**Session Chair**	**Location**
8:30-08:45	Opening and Welcome	Vipin Kumar, Laura Haas Howard Ho, Beng Chin Ooi Mohammed J. Zaki Morris Hui-I Hsiao Jian Li, Sudarsan Rachuri Shipeng Yu, Xiaohua Hu	*Ballroom ABC*
8:45-09:45	Keynote Session 1: How Big Data Changes Statistical Machine Learning *Dr. Léon Bottou, Facebook AI Research, New York*	Mohammed Zaki	*Ballroom ABC*
9:45 – 10:05 am Location:	**Coffee Break** *Grand Ballroom Foyer/Courtyard* **Poster Session (Set up only)** *Grand Ballroom Foyer/Corridor*		

Friday, 30-October - continued			
10:05 am -12:35 pm	**Sessions**	**Session Chair**	**Location**
	S1: Big Data Analytics	Vahid Taslimi Tehrani	*Ballroom ABC*
	S2: System Performance	I-Jen Chiang	*Ballroom D*
	S3: Optimizations for Efficient Big Data Processing	Marian Vajtersic	*Ballroom E*
	I&G1: Learning and Analytics	Nilesh Jain	*Ballroom F*
	Special Session I : From Data to Insight: Big Data and Analytics for Advanced Manufacturing Systems	Sudarsan Rachuri, Ronay Ak Tina Lee, Rumi Ghosh Srinivasan Soundar Steve Eglash, Sagar Kamarthi	*Bayshore*
12:35 – 2:00 pm Location:	**Lunch** *Poolside (outdoors) and the Magnolia Room* **Poster Session Sets Up and Displays** *Grand Ballroom Foyer*		
2:00 – 4:05 pm	**Sessions**	**Session Chair**	**Location**
	L1: Stream Processing	Jeff Saltz	*Ballroom ABC*
	L2: High Performance Computing Platforms	Weijia Xu	*Ballroom D*
	L3: Link and Graph Mining	Mohammad Al Hasan	*Ballroom E*
	I&G2:Search and Social Network	Ricardo Baeza-Yates	*Ballroom F*
	Special Session I: From Data to Insight: Big Data and Analytics for Advanced Manufacturing Systems	Sudarsan Rachuri, Ronay Ak Tina Lee, Rumi Ghosh Srinivasan Soundar Steve Eglash, Sagar Kamarthi	*Bayshore*
4:05 – 4:25 pm Location:	**Coffee Break** *Grand Ballroom Foyer/Courtyard* **Poster Session Sets Up and Displays** *Grand Ballroom Foyer/Corridor*		
4:25 -6:30 pm	**Sessions**	**Session Chair**	**Location**
	L4: Data Integration, Quality and Protection	Ahmed Metwally	*Ballroom ABC*
	L5: Social Web Search and Mining	Yi Fang	*Ballroom D*
	L6: Modern Data Processing Platforms	Umit Catalyurek	*Ballroom E*
	I&G3: Marketing	Shipeng Yu	*Ballroom F*
	Special Session I: From Data to Insight: Big Data and Analytics for Advanced Manufacturing Systems	Sudarsan Rachuri, Ronay Ak Tina Lee, Rumi Ghosh Srinivasan Soundar Steve Eglash, Sagar Kamarthi	*Bayshore*

Saturday, 31-October			
7:30-6:00 pm Location:	**Registration** *Lobby West*		
8:30 -9:30 am	**Sessions**	**Session Chair**	**Location**
	Keynote Speech 2: Moving Past the "Wild West" Era for Big Data *H. V. Jagadish, Bernard A Galler* *Collegiate Professor of Electrical Engineering and Computer Science, University of Michigan*	Laura Haas	*Ballroom ABC*
9:30 - 9:40am Location:	**Coffee Break** *Grand Ballroom Foyer/Courtyard* **Poster Session Displays** *Grand Ballroom Foyer/Corridor*		
9:40 - 12:15 pm	**Sessions**	**Session Chair**	**Location**
	Key Challenges for Future Big Data to Knowledge (BD2K) Technologies	Vijay Raghavan	*Ballroom ABC*
	I&G 4: Platform/Applications	Morris Hsiao	*Ballroom D*
	Tutorial 6: *Tutorial on Predictive Maintenance*	Zhuang Wang	*Ballroom E*
12:15 - 1:40 pm Location:	**Lunch** *Poolside (outdoors) and the Magnolia Room* **Poster Session Displays** *Grand Ballroom Foyer*		
Time	**Sessions**	**Session Chair**	**Location**
1:40 – 3:45 pm	**L7**: Crowdsourcing, Sampling and Visualization Analytics	Anand Tripathi	*Ballroom ABC*
	L8: SpatioTemporal Data Processing	Petros Zerfos	*Ballroom D*
	L9: Efficiency and Scalability	Rajdeep Bhowmik	*Ballroom E*
	I&G 5: System and Hardware	Jian Li	*Ballroom F*
	Tutorial 3: Platforms and Algorithms for Big Data Analytics	Chandan K. Reddy	*Bayshore*
3:45 – 4:00 pm Location:	**Coffee Break** *Grand Ballroom Foyer/Courtyard* **Poster Session Displays** *Grand Ballroom Foyer/Corridor*		
Time	**Sessions**	**Session Chair**	**Location**
4:00- 6:30 pm	**L10:** Algorithms and Systems for Big Data Search and Analytics	Hanghang Tong	*Ballroom ABC*
	L11: Complex Big Data Applications	Kavita Ganesan	*Ballroom D*
	L12: Large-scale Recommendation Systems and Social Media Systems	Marian Vajtersic	*Ballroom E*
	I&G6: Anomaly Detection	Srivathsan Srinivas	*Ballroom F*
	Tutorial 5: The World is Big and Linked: Whole Spectrum Industry Solutions towards Big Graphs	Toyotaro Suzumura, Ching-Yung Lin, Yinglong Xia, Lifeng Nai	*Bayshore*
7:30 – 9:00 pm Location	**Banquet (Ticket required)** *Ballroom ABCD* 1. *Big Data in Huawei,* Dr. Ziang Hu, Huawei 2. *Best Paper Award,* Howard Ho, Beng Chin Ooi, Mohammed J. Zaki 3. *Best Application Paper Award ,* Morris Hui-I Hsiao, Jian Li, Sudarsan Rachuri, Shipeng Yu		

Sunday, 01-November			
07:30-6:00pm **Location**:	**Registration** *Lobby West*		
Time	**Session**	**Session Chair**	**Location**
8:40 - 09:40 am	*Keynote Speech 3*: Conquering Big Data with Spark *Professor Ion Stoica, University of California, Berkeley*	Howard Ho	*Ballroom ABC*
9:40 - 10:00 am and 3:30 – 3:50 pm Location:	**Coffee Break** *Grand Ballroom Foyer/Courtyard* **Poster Session Displays** *Grand Ballroom Foyer/Corridor*		
Time	**Sessions/Tutorial/Workshop**	**Session Chair**	**Location**
Sessions/Tutorial 10:00 am - 12:30 pm	**S4:** Complex Big Data Applications	Nikunj Oza	*Ballroom ABC*
	S5: Integration, Usability and Visualization	Uwe Glasser	*Ballroom D*
	S6: Data Mining and Learning	Christos Anagnostopoulos	*Ballroom E*
	I&G S: I&G short papers	Sudarsan Rachuri	*Ballroom F*
	Tutorial 4: Optimal Connectivity on Big Graphs: Measures, Algorithms and Applications	Hanghang Tong	*Ballroom G*
Workshop 10:00 am -1:00 pm	***Data Quality Issues in Big Data***	GudivaaVenkat	*Ballroom H*
12:30- 1:30 pm	**Lunch** *Poolside (outdoors) and the Magnolia Room* **Poster Session Displays** *Grand Ballroom Foyer*		
Time	**Sessions/Workshops**	**Session Chair**	**Location**
Workshops 1:30 – 6:30 pm	Big Data for Sustainable Development	Aki-Hiro Sato	*Ballroom ABC*
	High Performance Big Graph Data Management, Analysis,and Mining (BigGraph 2015)	Mohammad Al Hasan, Kamesh Madduri, Fengguang Song	*Ballroom D*
	Big Data Analytics in Manufacturing and Supply Chains	Zhang NengSheng	*Ballroom E*
	Special Session III: Granular Computing and Big Data	T.Y. Lin	*Ballroom F*
	Big Data Startup Showcase	Duc Thanh Tran	*Ballroom H*
6:30 pm	***CONFERENCE ADJOURNS***		

Keynote Lectures

Keynote 1: How Big Data changes Statistical Machine Learning

Speaker:
Dr. Léon Bottou, Facebook AI Research, New York

Abstract:
This presentation illustrates how big data forces change on algorithmic techniques and the goals of machine learning, bringing along challenges and opportunities:
1. The theoretical foundations of statistical machine learning traditionally assume that training data is scarce. If one assumes instead that data is abundant and that the bottleneck is the computation time, stochastic algorithms with poor optimization performance become very attractive learning algorithms. These algorithms quickly became the backbone of large-scale machine learning and are the object of very active research.
2. Increasing the training set size cannot improve average errors indefinitely. However this diminishing returns problem vanishes if we measure instead the diversity of conditions in which the trained system performs well. In other words, big data is not an opportunity to increase the average accuracy, but an opportunity to increase coverage. Machine learning research must broaden its statistical framework in order to embrace all the (changing) aspects of real big data problems. Transfer learning, causal inference, and deep learning are successful steps in this direction.

Short Bio:
Léon Bottou received the Diplôme d'Ingénieur de l'École Polytechnique (X84) in 1987, the Magistère de Mathématiques Fondamentales et Appliquées et d'Informatique from École Normale Supérieure in 1988, and a doctorat from Université de Paris-Sud in 1991. His research career took him to AT&T Bell Laboratories, AT&T Labs Research, NEC Labs America and Microsoft. He joined Facebook AI Research in 2015. The long term goal of Léon's research is to understand how to build human-level intelligence. Although reaching this goal requires conceptual advances that cannot be anticipated at this point, it certainly entails clarifying how to learn and how to reason. Leon Bottou best known contributions are his work on neural networks in the 90s, his work on large scale learning in the 00's, and possibly his more recent work on causal inference in learning systems. Léon is also known for the DjVu document compression technology.

Keynote 2: Moving Past the "Wild West" Era for Big Data

Speaker:
H. V. Jagadish, Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science, University of Michigan

Abstract:
The potential of Big Data is widely recognized and many are seeking fortunes with Big Data today, just as they once sought fortunes by heading West in America. While success was initially limited only by creativity and passion, over time we need civilization, with all its accompanying benefits and constraints. As the field of Big Data matures, it is approaching the end of the "Wild West" era. In this talk, I will suggest what the "civilized" era may look like.

Short Bio:
Hosagrahar Visvesvaraya Jagadish (Jag) is a computer scientist in the field of database systems research. He is the Bernard A. Galler Collegiate Professor of Electrical Engineering and Computer Science at the University of Michigan at Ann Arbor and a Senior Scientific Director of the National Center for Integrative Biomedical Informatics established by the National Institutes of Health. Prior to joining the Michigan faculty, he spent over a decade at AT&T Bell Laboratories as a research scientist where he would eventually become head of the Database division. Jagadish earned his bachelor's degree from the Indian Institute of Technology, Delhi and a doctorate in Electrical Engineering from Stanford University in 1985. He was elected fellow of the Association for Computing Machinery in 2003 and trustee of the VLDB Endowment in 2004. He was the founding editor of the Proceedings of the VLDB Endowment (PVLDB) in 2008.

Keynote 3: Conquering Big Data with Spark

Speaker:
Prof. Ion Stocia, UC Berkeley, USA

Abstract:
Today, big and small organizations alike collect huge amounts of data, and they do so with one goal in mind: extract "value" through sophisticated exploratory analysis, and use it as the basis to make decisions as varied as personalized treatment and ad targeting. To address this challenge, we have developed Berkeley Data Analytics Stack (BDAS), an open source data analytics stack for big data processing.
In this talk I'll focus on the execution engine in BDAS: Apache Spark. Apache Spark is a cluster computing engine that is optimized for in-memory processing, and unifies support for a variety of workloads, including batch, streaming, and iterative computations. Spark is now the most active big data project in the open source community, and is already being used by over one thousand organizations

Short Bio:
Ion Stoica is a Professor in the EECS Department at University of California at Berkeley. He received his PhD from Carnegie Mellon University, and his B.S. from Polytechnic Institute Bucharest. He does research on cloud computing and networked computer systems. Past work includes the Dynamic Packet State (DPS), Chord DHT, Internet Indirection Infrastructure (i3), declarative networks, replay-debugging, and multi-layer tracing in distributed systems. His current research focuses on resource management and scheduling for data centers, cluster computing frameworks, and network architectures. He is an ACM Fellow and has received numerous awards, including the SIGCOMM Test of Time Award (2011), and the ACM doctoral dissertation award (2001). In 2006, he co-founded Conviva, a startup to commercialize technologies for large scale video distribution, and in 2013, he co-founded Databricks a startup to commercialize technologies for Big Data processing.

Conference Paper Presentations

L1: Stream Processing	
Regular	BigD477 "Elastic Complex Event Processing exploiting Prediction" Nikos Zacheilas, Vana Kalogeraki, Nikolas Zygouras, Nikolaos Panagiotou, and Dimitrios Gunopulos
Regular	BigD353 "Online and On-demand Partitioning of Streaming Graphs" Ioanna Filippidou and Yiannis Kotidis
Regular	BigD318 "ScaleJoin: a Deterministic, Disjoint-Parallel and Skew-Resilient Stream Join" Vincenzo Gulisano, Yiannis Nikolakopoulos, Marina Papatriantafilou, and Philippas Tsigas
Regular	BigD283 "Data Streaming Algorithms for the Kolmogorov-Smirnov Test" Ashwin Lall
Regular	BigD304 "Workload Scheduling in Distributed Stream Processors using Graph Partitioning" Lorenz Fischer and Abraham Bernstein

L2: High Performance Computing Platforms	
Regular	BigD268 "ScaleGraph: A High-Performance Library for Billion-Scale Graph Analytics" Koji Ueno and Toyotaro Suzumura
Regular	BigD274 "System and Architecture Level Characterization of Big Data Applications on Big and Little Core Server Architectures" Maria Malik and Houman Homayoun
Regular	BigD410 "Distributed Frank-Wolfe under Pipelined Stale Synchronous Parallelism" Nam-Luc Tran, Thomas Peel, and Sabri Skhiri
Regular	BigD581 "Performance Characterization and Acceleration of In-Memory File Systems for Hadoop and Spark Applications on HPC Clusters" Nusrat Islam, Md. Wasi-ur- Rahman, Xiaoyi Lu, Dipti Shankar, and Dhabaleswar K. Panda
Regular	BigD310 "Evaluating Different Distributed-Cyber-Infrastructure for Data and Compute Intensive Scientific Application" Arghya Kusum Das, Seung-Jong Park, Jaeki Hong, and Wooseok Chang

L3: Link and Graph Mining	
Regular	BigD347 "Inferring Crowd-Sourced Venues for Tweets" Bokai Cao, Francine Chen, Dhiraj Joshi, and Philip S. Yu
Regular	BigD349 "Core Decomposition in Large Temporal Graphs" Huanhuan Wu, James Cheng, Yi Lu, Yiping Ke, Yuzhen Huang, Da Yan, and Hejun Wu
Regular	BigD292 "Scalable Classification for Large Dynamic Networks" Yibo Yao and Lawrence Holder
Regular	BigD492 "Modeling Graphs Using a Mixture of Kronecker Models" Suchismit Mahapatra and Varun Chandola
Regular	BigD559 "Toward Precise User-Topic Alignment in Online Social Media” Jiejun Xu and Tsai-Ching Lu

L4: Data Integration, Quality and Protection	
Regular	BigD348 "Big Data Entity Resolution: From Highly to Somehow Similar Entity Descriptions in the Web" Vasilis Efthymiou, Kostas Stefanidis, and Vassilis Christophides
Regular	BigD350 "Parallel Meta-blocking: Realizing Scalable Entity Resolution over Large, Heterogeneous Data"

	Vasilis Efthymiou, George Papadakis, George Papastefanatos, Kostas Stefanidis, and Themis Palpanas
Regular	BigD498 "Data Quality Assessment and Anomaly Detection Via Map / Reduce and Linked Data: A Case Study in the Medical Domain" Stephen Bonner, A. Stephen McGough, Ibad Kureshi, John Brennan, Georgios Theodoropoulos, Laura Moss, David Corsar, and Grigoris Antoniou
Regular	BigD495 "AccountableMR: Toward Accountable MapReduce systems" Huseyin Ulusoy, Murat Kantarcioglu, Erman Pattuk, and Lalana Kagal
Regular	BigD485 "TrustMR: Computation Integrity Assurance system for MapReduce" Huseyin Ulusoy, Murat Kantarcioglu, and Erman Pattuk

L5: Social Web Search and Mining	
Regular	BigD379 "Modelling Cascades Over Time in Microblogs" Wei Xie, Feida Zhu, Siyuan Liu, and Ke Wang
Regular	BigD236 "Dynamic theme tracking in Twitter" Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan
Regular	BigD443 "Full Diffusion History Reconstruction in Networks" Zhen Chen, Hanghang Tong, and Lei Ying
Regular	BigD557 "Identifying Smallest Unique Subgraphs in a Heterogeneous Social Network" Yen-Kai Wang, Wei-Ming Chen, Cheng-Te Li, and Shou-De Lin
Regular	BigD276 "The Roles of Network Communities in Social Information Diffusion" Cheng-Te Li, Yu-Jen Lin, and Mi-Yen Yeh

L6: Modern Data Processing Platforms	
Regular	BigD375 "A Scalable Parallel XQuery Processor" Eldon Carman, Vassilis Tsotras, Till Westmann, Vinayak Borkar, and Michael Carey
Regular	BigD560 "PortHadoop: Support Direct HPC Data Processing in Hadoop" Xi Yang, Ning Liu, Bo Feng, Xian-He Sun, and Shujia Zhou
Regular	BigD397 "Slingshot: A Modular Framework for Designing Data Processing Systems" Bogdan Simion, Daniel Ilha, Suprio Ray, Leslie Barron, Angela Demke Brown, and Ryan Johnson
Regular	BigD326 "When Computing Meets Heterogeneous Cluster: Workload Assignment in Graph Computation" Jilong Xue, Zhi Yang, Shian Hou, and Yafei Dai
Regular	BigD399 "LabBook: Metadata-driven Social Collaborative Data Analysis" Eser Kandogan, Mary Roth, Peter Schwarz, Joshua Hui, Ignacio Terrizzano, Christina Christodoulakis, and Renee Miller

L7: Crowdsourcing, Sampling and Visualization Analytics	
Regular	BigD566 "Visual Interface for Exploring Caution Spots from Vehicle Recorder Big Data" Masahiko Itoh, Daisaku Yokoyama, Masashi Toyoda, and Masaru Kitsuregawa
Regular	BigD267 "Visual Analysis of Bi-directional Movement Behavior" Yixian Zheng, Wenchao Wu, Huamin Qu, Chunyan Ma, and Lionel M. Ni
Regular	BigD320 "Revealing the Fog-of-War: A Visualization-directed, Uncertainty-aware Approach for Exploring High-dimensional Data" Yang Wang and Kwan-Liu Ma
Regular	BigD441 "Effectively Crowdsourcing the Acquisition and Analysis of Visual Data for Disaster Response" Hien To, Seon Ho Kim, and Cyrus Shahabi

Regular	BigD469 "AdaM: an Adaptive Monitoring Framework for Sampling and Filtering on IoT Devices" Demetris Trihinas, George Pallis, and Marios Dikaiakos

L8: SpatioTemporal Data Processing	
Regular	BigD254 "Parallel In-Memory Trajectory-based Spatiotemporal Topological Join" Suprio Ray, Angela Demke Brown, Nick Koudas, Rolando Blanco, and Anil Goel
Regular	BigD275 "Recommending Missing Sensor Values" Chung-Yi Li, Wei-Lun Su, Todd G. McKenzie, Fu-Chun Hsu, Shou-De Lin, Phillip B. Gibbons, and Jane Yung-jen Hsu
Regular	BigD526 "TKSimGPU: A Parallel Top-K Trajectory Similarity Query Processing Algorithm for GPGPUs and Multicore CPUs" Eleazar Leal, Le Gruenwald, Jianting Zhang, and Simin You
Regular	BigD293 "Techniques for Fast and Scalable Time Series Traffic Generation" Jilong Kuang, Daniel Waddington, and Changhui Lin
Regular	BigD291 "A Scalable Approach for Data-Driven Taxi Ride-Sharing Simulation" Masayo Ota, Huy Vo, Claudio T. Silva, and Juliana Freire

L9: Efficiency and Scalability	
Regular	BigD303 "Energy-Efficient Acceleration of Big Data Analytics Applications Using FPGAs" Katayoun Neshatpour, Maria Malik, Mohammad Ali Ghodrat, and Houman Homayoun
Regular	BigD433 "Towards Green Cloud Computing: Demand Allocation and Pricing Policies for Cloud Service Brokerage" Chenxi Qiu, Haiying Shen, and Liuhua Chen
Regular	BigD600 "PANOPTICON: A lock broker architecture for scalable transactions in the datacenter" Serafettin Tasci and Murat Demirbas
Regular	BigD391 "Computing Load Aware and Long-View Load Balancing for Cluster Storage Systems" Guoxin Liu, Haiying Shen, and Haoyu Wang
Regular	BigD563 "Machine Learning at the Limit" John Canny, Huasha Zhao, Ye Chen, Jiangchang Mao, and Bobby Jaros

L10: Algorithms and Systems for Big Data Search and Analytics	
Regular	BigD279 "Angular Quantization Based Affinity Propagation Clustering and its Application to Astronomical Big Spectra Data" Wang Ke, Guo Ping, and Luo A-Li
Regular	BigD300 "CINTIA: a Distributed, Low-Latency Index for Big Interval Data" Ruslan Mavlyutov and Philippe Cudré-Mauroux
Regular	BigD522 "SigCO: Mining Significant Correlations via a Distributed Real-time Computation Engine" Tian Guo, Jean-Paul Calbimonte, Hao Zhuang, and Karl Aberer
Regular	BigD224 "Bandwidth-Efficient Distributed k-Nearest-Neighbor Search with Dynamic Time Warping" Chin-Chi Hsu, Perng-Hwa Kung, Mi-Yen Yeh, Shou-De Lin, and Phillip B. Gibbons
Regular	BigD360 "Learning to Accurately COUNT with Query-Driven Predictive Analytics" Christos Anagnostopoulos and Peter Triantafillou
Regular	BigD401 "Fast Decentralized Gradient Descent Method and Applications to In-situ Seismic Tomography" Liang Zhao, WenZhan Song, and Xiaojing Ye

L11: Complex Big Data Applications	
Regular	BigD255 "Cluster-based Aggregate Forecasting for Residential Electricity Demand using Smart Meter Data" Tri Kurniawan Wijaya, Matteo Vasirani, Samuel Humeau, and Karl Aberer
Regular	BigD266 "Spatially Clustered Join on Heterogeneous Scientific Data Sets" Bin Dong, Suren Byna, and Kesheng Wu
Regular	BigD395 "EveryoneCounts: Data-Driven Digital Advertising based on Uncertain Demand Models in Metro Networks" Desheng Zhang, Ruobing Jiang, Shuai Wang, Yanmin Zhu, Bo Yang, Tian He, Jian Cao, and Fan Zhang
Regular	BigD472 "Scientific Computing Meets Big Data Technology: An Astronomy Use Case" Zhao Zhang, Kyle Barbary, Frank Austin Nothaft, Evan Sparks, Oliver Zahn, Michael J. Franklin, David A. Patterson, and Saul Perlmutter
Regular	BigD416 "Evaluating Cloud Frameworks on Genomic Applications" Michele Bertoni, Stefano Ceri, Abdulrahman Kaitoua, and Pietro Pinoli
Regular	BigD534 "An Interactive Learning Framework for Scalable Classification of Pathology Images" Michael Nalisnik, David Gutman, Jun Kong, and Lee Cooper

L12: Large-scale Recommendation Systems and Social Media Systems	
Regular	BigD366 "Recommending Forum Posts to Designated Experts" Jason H.D. Cho, Yanen Li, Roxana Girju, and Chengxiang Zhai
Regular	BigD373 "Accelerating Collaborative Filtering Using Concepts from High Performance Computing" Mark Gates, Hartwig Anzt, Jakub Kurzak, and Jack Dongarra
Regular	BigD409 "CSFinder: A Cold-Start Friend Finder in Large-Scale Social Networks" Yasser Salem, Jun Hong, and Weiru Liu
Regular	BigD272 "User-Curated Image Collections: Modeling and Recommendation" Yuncheng Li and Jiebo Luo
Regular	BigD248 "SyntacticDiff: Operator-Based Transformation for Comparative Text Mining" Sean Massung and Chengxiang Zhai

S 1: Big Data Analytics	
Short	BigD321 "Predicting the Location of Users on Twitter from Low Density Graphs" Sofia Apreleva and Alejandro Cantarero
Short	BigD514 "City users' classification with mobile phone data" Lorenzo Gabrielli, Barbara Furletti, Roberto Trasarti, Fosca Giannotti, and Dino Pedreschi
Short	BigD582 "Efficient Distributed Maximum Matching for Solving the Container Exchange Problem in the Maritime Industry" Li-Yung Ho, Fei Shao, Jan-Jan Wu, and Pangfeng Liu
Short	BigD271 "America Tweets China: A Fine-Grained Analysis of the State and Individual Characteristics Regards Attitudes towards China" Yu Wang and Jiebo Luo
Short	BigD302 "A MapReduce Based k-NN Joins Probabilistic Classifier" Georgios Chatzigeorgakidis, Sophia Karagiorgou, Spiros Athanasiou, and Spiros Skiadopoulos
Short	BigD415 "Smog Disaster Forecasting using Social Web Data and Physical Sensor Data" Jiaoyan Chen, Huajun Chen, Daning Hu, Yalin Zhou, and Ming Wu
Short	BigD488 "Spatio-Temporal Asynchronous Co-Occurrence Pattern for Big Climate Data towards Long-Lead Flood Prediction" Chung-Hsien Yu, Dong Luo, Wei Ding, Joseph Cohen, David Small, and Shafiqul Islam

Short	BigD429 "MMC-Margin: Identification of Maximum Frequent Subgraphs by Metropolis Monte Carlo Sampling muyi liu and Michael Gribskov
Short	BigD450 "KeyLabel Algorithms for Keyword Search in Large Graphs" Yue Wang, Ke Wang, Ada Wai-Chee Fu, and Raymond Chi-Wing Wong
Short	BigD305 "Scalable k-NN based text clustering" Alessandro Lulli, Thibault Debatty, Laura Ricci, Matteo Dell'Amico, and Pietro Michiardi
Short	BigD609 "Considerations and Recommendations for Data Contributions for Predictive Analytics for Manufacturing" Don Libes, Seungjun Shin, and Jungyub Woo

S 2: System Performance	
Short	BigD398 "Chronos: Failure-Aware Scheduling in Shared Hadoop Clusters" Orcun Yildiz, Shadi Ibrahim, Tran Anh Phuong, and Gabriel Antoniu
Short	BigD238 "ATOM: Automated Tracking, Orchestration, and Monitoring of Resource Usage in Infrastructure as a Service Systems" Min Du and Feifei Li
Short	BigD561 "Regular Expression Acceleration on the Micron Automata Processor: Brill Tagging as a Case Study" Keira Zhou, Jack Wadden, Jeffrey Fox, Ke Wang, Donald Brown, and Kevin Skadron
Short	BigD473 "An Architecture for Stream OLAP Exploiting SPE and OLAP Engine" Kosuke Nakabasami, Toshiyuki Amagasa, Salman Shaikh, Franck Gass, and Hiroyuki Kitagawa
Short	BigD378 "Octopus: A Multi-tenant Scheduler for Graphlab" Srikant Padala, Dinesh Kumar, Arun Raj, and Janakiram Dharanipragada
Short	BigD390 "G-Storm: GPU-enabled High-throughput Online Data Processing in Storm" Zhenhua Chen, Jielong Xu, Jian Tang, Kevin Kwiat, and Charles Kamhoua
Short	BigD519 "BigFUN: A Performance Study of Big Data Management System Functionality" Pouria Pirzadeh, Michael Carey, and Till Westmann
Short	BigD584 "Benchmarking Key-Value Stores on High-Performance Storage and Interconnects for Web-Scale Workloads" Dipti Shankar, Xiaoyi Lu, Md. Wasi-ur-Rahman, Nusrat Islam, and Dhabaleswar K. Panda
Short	BigD387 "Spark Deployment and Performance Evaluation on the MareNostrum Supercomputer" Ruben Tous, Anastasios Gounaris, Carlos Tripiana, Jordi Torres, Sergi Girona, Eduard Ayguadé, Jesús Labarta, Yolanda Becerra, David Carrera, and Mateo Valero
Short	BigD555 "TPS: A Task Placement Strategy for Big Data Workflows" Mahdi Ebrahimi, Aravind Mohan, Shiyong Lu, and Robert Reynolds

S 3: Optimizations for Efficient Big Data Processing	
Short	BigD442 "Cost-Efficient Partitioning of Spatial Data on Cloud" Afsin Akdogan, Saratchandra Indrakanti, Ugur Demiryurek, and Cyrus Shahabi
Short	BigD494 "Two-Mode Data Distribution Scheme for Heterogeneous Storage in Data Centers" Wei Xie, Jiang Zhou, Mark Reyes, Jason Noble, and Yong Chen
Short	BigD338 "Composable and Efficient Functional Big Data Processing Framework" Dongyao Wu, Sherif Sakr, Liming Zhu, and Qinghua Lu
Short	BigD502 "A Predictive Scheduling Framework for Fast and Distributed Stream Data Processing" Teng Li, Jian Tang, and Jielong Xu
Short	BigD358 "Hybrid Active Learning for Non-stationary Streaming Data with Asynchronous Labeling" Hyunjoo Kim, Sriganesh Madhvanath, and Tong Sun
Short	BigD521 "Edge Importance Identification for Energy Efficient Graph Processing" S M Faisal, G. Tziantzioulis, A. M. Gok, S. Parthasarathy, N. Hardavellas, and S. Ogrenci-Memik

Short	BigD222 "Toward Locality-aware Scheduling for Containerized Cloud Services" Dongfang Zhao, Nagapramod Mandagere, Gabriel Alatorre, Mohamed Mohamed, and Heiko Ludwig
Short	BigD216 "ACURDION: An Adaptive Clustering-based Algorithm for Tracing Large-scale MPI Applications" Amir Bahmani and Frank Mueller
Short	BigD510 "A Scalable Implementation of Information Theoretic Feature Selection for High Dimensional Data" Anthony Kleerekoper, Michael Pappas, Mikel Lujan, Gavin Brown, and Adam Pocock
Short	BigD335 "Rewriting Complex SPARQL Analytical Queries for Efficient Cloud-based Processing" Padmashree Ravindra, HyeongSik Kim, and Kemafor Anyanwu

S 4: Complex Big Data Applications	
Short	BigD290 "A Data-Driven Approach to Extract Connectivity Structures from Diffusion Tensor Imaging Data" Yu Jin, Joseph JaJa, Rong Chen, and Edward Herskovits
Short	BigD307 "An Ensemble Learning Based Approach for Building Airfare Forecast Service" Yuwen Chen, Jian Cao, Shanshan Feng, and Yudong Tan
Short	BigD313 "Next-Term Student Grade Prediction" Mack Sweeney, Jaime Lester, and Huzefa Rangwala
Short	BigD447 "Large scale support vector regression for aviation safety" Kamalika Das, Kanishka Bhaduri, Bryan Matthews, and Nikunj Oza
Short	BigD515 "Spaler: Spark and GraphX based de novo genome assembler" Anas Abu-Doleh and Umit Catalyurek
Short	BigD537 "Traffic Forecasting In Complex Urban Networks: Leveraging Big Data and Machine Learning" Florin Schimbinschi, Xuan Vinh Nguyen, James Bailey, Chris Leckie, Hai Vu, and Ramamohanarao Kotagiri
Short	BigD565 "Prediction of Physiological Subsystem Failure and its Impact in the prediction of Patient Mortality" Karla Caballero Barajas and Ram Akella
Short	BigD602 "Cell Analytics in Compound Hit Selection of Bacterial Inhibitors" Robert P. Trevino, Steve A. Kawamoto, Thomas J. Lamkin, and Huan Liu
Short	BigD388 "How not to drown in a sea of information: An event recognition approach" Elias Alevizos, Alexander Artikis, Kostas Patroumpas, Marios Vodas, Yannis Theodoridis, and Nikos Pelekis
Short	BigD467 "You Can Promote, But You Can't Hide: Large-Scale Abused App Detection in Mobile App Stores" Zhen Xie and Sencun Zhu
Short	BigD595 "Human Mobility and Economic Development" Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti

S 5: Integration, Usability and Visualization	
Short	BigD328 "Matisse: A Visual Analytics System for Exploring Emotion Trends in Social Media Text Streams" Chad Steed, Margaret Drouhard, Justin Beaver, Joshua Pyle, and Paul Bogen
Short	BigD598 "An Iterative Methodology for Big Data Management, Analysis and Visualization" Roberto Tardío Olmos, Alejandro Maté Morga, and Juan Carlos Trujillo Mondéjar
Short	BigD218 "Time Maps: A Tool for Visualizing Many Discrete Events Across Multiple Timescales" Max Watson

Short	BigD439 "Brown Dog: Leveraging Everything Towards Autocuration" Smruti Padhy, Greg Jansen, Jay Alameda, Edgar Black, Liana Diesendruck, Mike Dietze, Praveen Kumar, Rob Kooper, Jong Lee, Riu Liu, Ricard Marciano, Luigi Marini, Dave Mattson, Barbara Minsker, Chris Navarro, Marcus Slavenas, William Sullivan, Jason Votava, and Kenton McHenry
Short	BigD417 "DSDQuery DSI - Querying Scientific Data Repositories with Structured Operators" Roee Ebenstein and Gagan Agrawal
Short	BigD269 "A Transaction Model for Management of Replicated Data with Multiple Consistency Levels" Anand Tripathi and BhagavathiDhass Thirunavukarasu
Short	BigD577 "Improving Transaction Processing Performance By Consensus Reduction" Yuqing ZHU, Yilei WANG, and Fan WANG
Short	BigD554 "A Flexible QoS Fortified Distributed Key-Value Storage System for the Cloud" Tonglin Li, Ke Wang, Shiva Srivastava, Dongfang Zhao, Kan Qiao, Iman Sadooghi, Xiaobing Zhou, and Ioan Raicu
Short	BigD359 "Quadtree-Based Lightweight Data Compression for Large-Scale Geospatial Rasters on Multi-Core CPUs" Jianting Zhang, Simin You, and Le Gruenwald
Short	BigD344 "Concept Hierarchies and Human Navigation" Salvador Aguinaga, Aditya Nambiar, Zuozhu Liu, and Tim Weninger

S 6: Data Mining and Learning	
Short	BigD423 "Improving EEG Feature Learning via Synchronized Facial Video" Xiaoyi Li, Xiaowei Jia, and Aidong Zhang
Short	BigD405 "Multi-modal Learning for Video Recommendation based on Mobile Application Usage" Xiaowei Jia, Aosen Wang, Xiaoyi Li, Guangxu Xun, Wenyao Xu, and Aidong Zhang
Short	BigD403 "Task-based Recommendation on a Web-Scale" Yi Zhang and Yongfeng Zhang
Short	BigD380 "A Community Driven Social Recommendation System" Deepika Lalwani, Somayajulu D. V. L. N., and Radha Krishna Pisipati
Short	BigD336 "Robust Crowd Bias Correction via Dual Knowledge Transfer from Multiple Overlapping Sources" Philip S. Yu and Sihong Xie
Short	BigD242 "Learning Relevance from Click Data via Neural Network based Similarity Models" Xugang Ye and Zijie Qi
Short	BigD544 "Super-CWC and Super-LCC: Super Fast Feature Selection Algorithms" Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto, and Dave Shepard
Short	BigD452 "Towards Scalable Quantile Regression Trees" Harish Bhat, Nitesh Kumar, and Garnet Vaz
Short	BigD451 "Iteratively Refining SVMs" Enric Junqué de Fortuny, Theodoros Evgeniou, David Martens, and Foster Provost
Short	BigD259 "Practical Message-Passing Framework for large-Scale Combinatorial Optimization" Inho Cho, Soya Park, Sejun Park, Dongsu Han, and Jinwoo Shin

Industry and Government Paper Presentations

I&G 1: Learning and Analytics	
N216	Tanay Saha, Mohammad Hasan, Chandler Burgess, Md Ahsan Habib, and Jeff Johnson, *Batch Mode Active Learning for Technology-Assisted Review*
N229	Sauptik Dhar, Congrui Yi, Naveen Ramakrishnan, and Mohak Shah, *ADMM based Scalable Machine Learning on Spark*
N230	Dapeng Dong and John Herbert, *Record-aware Compression for Big Textual Data Analysis Acceleration*
N232	Alekh Jindal, Samuel Madden, Malú Castellanos, and Meichun Hsu, *Graph Analytics using Vertica Relational Database*
N240	Vinay Deolalikar, *How Valuable is Your Data? A Quantitative Approach using Data Mining*
N245	Sreenivas Sukumar, *Open Research Challenges with Big Data - A Data-Scientists Perspective*

I&G 2: Search and Social Network	
N238	Viet Ha-Thuc, Ganesh Venkataraman, Mario Rodriguez, Shakti Sinha, Senthil Sundaram, and Lin Guo, *Personalized Expertise Search at LinkedIn*
N202	Ahmed Metwally, Jia-Yu Pan, Minh Doan, and Christos Faloutsos, *Scalable Community Discovery from Multi-Faceted Graphs*
N211	Anjan Goswami, Wei Han, Zhenrui Wang, and Angela Jiang, *Controlled Experiments for Decision-Making in e-Commerce Search*
N224	Fang-Hsiang Su, Manas Somaiya, Shrish Mishra, and Rajyashree Mukherjee, *EXOS: EXpansion On Session for Enhancing Effectiveness of Query Auto-Completion*
N237	Mohammed Korayem, Camilo Ortiz, Khalifeh AlJadda, and Trey Grainger, *Query Sense Disambiguation Leveraging Large Scale User Behavioral Data*

I&G 3: Marketing	
N201	Xiuqiang He, Wenyuan Dai, Guoxiang Cao, Huyang Sun, Mingxuan Yuan, and Qiang Yang, *Mining Target Users for Online Marketing based on App Store Data*
N241	Kang Li, Vinay Deolalikar, and Neeraj Pradhan, *Mining Lifestyle Personas at Scale in E-commerce*
N250	Jayasimha Katukuri, Tolga Konik, Rajyashree Mukherjee, and Santanu Kolay, *Post-Purchase Recommendations in Large-scale Online Marketplaces*
N252	Hong-Han Shuai, Chih-Ya Shen, Hsiang-Chun Hsu, De-Nian Yang, Chung-Kuang Chou, Jihg-Hong Lin, and Ming-Syan Chen, *Revenue Maximization for Telecommunications Company with Social Viral Marketing*

I&G 4: Platform and Applications	
N205	Ernesto Diaz-Aviles, Fabio Pinelli, Karol Lynch, Zubair Nabi, Yiannis Gkoufas, Eric Bouillet, Francesco Calabrese, Eoin Coughlan, Peter Holland, and Jason Salzwedel, *Towards Real-time Customer Experience Prediction for Telecommunication Operators*
N212	Jenny Williams, Paul Cuddihy, Justin McHugh, Kareem Aggour, and Arvind Menon, *Semantics for Big Data Access & Integration: Improving Industrial Equipment Design through Increased Data Usability*
N221	Mayank Kejriwal, Qiaoling Liu, Ferosh Jacob, and Faizan Javed, *A Pipeline for Extracting and Deduplicating Domain-Specific Knowledge Bases*
N233	Andre Luckow, Ken Kennedy, Fabian Manhard, Emil Djerekarov, Bennie Vorster, and Amy Apon, *Automotive Big Data: Applications, Workloads and Infrastructures*
N234	Goktug Cinar, Jeffrey Thompson, and Soundar Srinivasan, *Cost-sensitive optimization of automated inspection*

N248	Levente Klein, Fernando Marianno, Conrad Albrecht, Marcus Freitag, Siyuan Lu, Nigel Hinds, Hendrik Hamann, and Sergio Bermudez, *PAIRS: A scalable geo-spatial data analytics platform*

I&G 5: System and Hardware	
N206	I. Stephen Choi, Weiqing Yang, and Yang-Suk Kee, *Early Experience with Optimizing I/O Performance Using High-Performance SSDs for In-Memory Cluster Computing*
N210	Hyunsik Choi, Yong In Lee, Jongyoung Park, Kangho Roh, and Kwanghyun La, *An Evaluation of Alternative Shared-nothing Architecture for Analytical Processing Systems*
N235	Nicolas Poggi, Josep Lluís Berral, David Carrera, Aaron Call, Rob Reinauer, Nikola Vujic, Daron Green, José Blakeley, and Fabrizio Gagliardi, *From Performance Profiling to Predictive Analytics while Evaluating Hadoop Cost-Efficiency in ALOJA*
N243	Petros Zerfos, Hangu Yeo, Brent Paulovicks, and Vadim Sheinin, *SDFS: Secure Distributed File System for Data-at-Rest Security for Hadoop-as–a-Service*

I&G 6: Anomaly Detection	
N213	Laura Rettig, Mourad Khayati, Michal Piorkowski, and Philippe Cudre-Mauroux, *Online Anomaly Detection over Big Data Streams*
N215	Aungon Nag Radon, Ke Wang, Uwe Glaesser, Hans Wehn, and Andrew Westwell-Roper, *Contextual Verification for False Alarm Reduction in Maritime Anomaly Detection*
N226	Gergely Acs, Jagdish Prasad Achara, and Claude Castelluccia, *Probabilistic km-anonymity (Efficient Anonymization of Large Set-Valued Datasets)*
N246	Hamed Yaghoubi Shahir, Uwe Glässer, Amir Yaghoubi Shahir, and Hans Wehn, *Maritime Situation Analysis Framework: Vessel Interaction Classification and Anomaly Detection*

I&G S: Short papers	
N204	Stephanie Rosenthal, Scott McMillan, and Matthew Gaston, *Developer Toolchains for Large-Scale Analytics: Two Case Studies*
N208	Harshal Godhia, Bibek Panda, Swarnim Narayan, and Ramakrishna Vadakattu, *Enterprise Subscription Churn Prediction*
N219	Joshua Seeger, Aron Culotta, Jason Keller, Patrick van Kessel, and Michael Jugovich, *Data Deidentification in Medical Transcriptions using Regular Expressions and Machine Learning*
N223	Qinlong Luo, Meng Zhao, Faizan Javed, and Ferosh Jacob,*Macau: Large-Scale Skill Sense Disambiguation in the Online Recruitment Domain*
N225	Wei Yi Liu, Morris H. Hsiao, and Shih Yao Dai, *DNA Analysis with MapReduce*
N227	Chaitali Gupta, Ranjan Sinha, and Yong Zhang, *Eagle: User Profile-based Anomaly Detection in Hadoop Clusters*
N231	Manuel Diaz-Granados, Javier Diaz-Montes, and Manish Parashar, *Investigating Insurance Fraud using Social Media*
N236	Luca Cazzanti, Leonardo Millefiori, and Gianfranco Arcieri, *A Document-based Data Model for Large Scale Computational Maritime Situational Awareness*
N249	Jhao-Yin Li, Mi-Yen Yeh, Ming-Syan Chen, and Jihg-Hong Lin, *Modeling Social Influences from Call Records and Mobile Web Browsing Histories (Extended Abstract)*
N251	Christian Seebode, Matthias Ort, Peter Hufnagl, and Christian R. A. Regenbrecht, *Next Generation Biobanks*

Panel:
Key Challenges for Future Big-Data to Knowledge (BD2K) Technologies

Panelists:

Moderator: Vijay Raghavan

1) Ricardo Baeza-Yates, VP, Yahoo!Labs
2) Anil Goel, VP and Chief Architect, SAP Hana Data Platform
3) H.V. Jagadish, Professor, University of Michigan, USA
4) Laura Pullum, Research Scientist, Oak Ridge National Laboratory
5) Ranjan Sinha, Head of Data Science eBay
6) Kristin Tolle, Director Microsoft, USA

Bios of Moderator and Panelists

Moderator:

Vijay Raghavan is the Alfred and Helen Lamson/ BoRSF Endowed Professor in Computer Science at the *Center for Advanced Computer Studies* and the Director of the NSF-sponsored *Industry/ University Cooperative Research Center for Visual and Decision Informatics*. His research interests are in data mining, information retrieval, machine learning and Internet computing. He has published over 250 peer-reviewed research papers- appearing in top-level journals and proceedings- that cumulatively accord him an *h-index* of 32, based on citations. He has served as major advisor for 24 doctoral students. He has also directed industry-sponsored research, on projects pertaining to Neuro-imaging based dementia detection and literature-based biomedical hypotheses generation, respectively. Dr. Raghavan serves as a member of the Executive Committee of the IEEE Technical Committee on Intelligent Informatics (IEEE-TCII), the Web Intelligence Consortium (WIC) Technical Committee and the Web Intelligence and Intelligent Agent Technology Conferences' Steering Committee. He is one of the Editors-in-Chief of the Web Intelligence journal and an Associate Editor of the ACM Transactions on Internet Technology. He is an ACM Distinguished Scientist and served as an ACM Distinguished Lecturer from 1993 – 2006. In addition, he served as a member of the Advisory Committee of the *NSF Computer and Information Science and Engineering* directorate *(CISE-AC)* during 2008 – 2010.

Panelists:

Ricardo Baeza-Yates is VP of Research at Yahoo Labs, Sunnyvale, USA. He is a Fellow of both ACM and IEEE. His research interests are Web Search and Data Mining as well as scalability in general. He received his Ph.D. from the University of Waterloo, Canada. He is co-author of the best-selling textbook, Modern Information Retrieval, published in 1999 by Addison-Wesley with a second enlarged edition in 2011 that won the ASIS&T 2012 book of the year award. He can be reached at rbaeza@acm.org.

Anil K. Goel is a Chief Architect at SAP where he works with the globally distributed HANA Platform and Database engineering team to drive forward looking architectures, vision, strategy and execution for all SAP data management products and technologies. He oversees data platform related co-innovation projects with hardware and software partners as well as collaborative research and internship programs with many universities in North America and Europe. His interests include database system architecture, in-memory and large scale distributed computing, self-management of software systems and cost modelling. Anil earned a PhD in CS from University of Waterloo. He also holds M.Tech in CS from the Indian Institute of Technology, Delhi, and B.E. (Electronics and Communications Engineering) from the University of Delhi.

H. V. Jagadish (Jag) is a computer scientist in the field of database systems research. He is the Bernard A. Galler Collegiate Professor of Electrical Engineering and Computer Science at the University of Michigan at Ann Arbor and Distinguished Scientist at the Michigan Institute for Data Science. Prior to joining the Michigan faculty, he spent over a decade at AT&T Bell Laboratories as a research scientist where he would eventually become head of the Database division. Jagadish earned his bachelor's degree from the Indian Institute of Technology, Delhi and a doctorate in Electrical Engineering from Stanford University in 1985. He was elected fellow of the Association for Computing Machinery in 2003 and trustee of the VLDB Endowment in 2004. He was the founding editor of the Proceedings of the VLDB Endowment (PVLDB) in 2008.

Laura Pullum is a research scientist in the Computer Science and Engineering Division at Oak Ridge National Laboratory (ORNL), a Department of Energy laboratory. She has led numerous research projects in industry, non-profit organizations, academia and government laboratories. Laura's research has focused on the dependability of software-intensive systems that incorporate state-of-the-art technology. Laura is the author of *Software Fault Tolerance: Techniques and Implementation* (Artech House) and lead author of *Guidance for the Verification and Validation of Neural Networks* (IEEE Computer Society Press/Wiley). She has authored numerous book chapters and peer-reviewed papers; holds a patent, serves on technical advisory boards and international science review panels, and serves on the standards working group for the *IEEE Standard for System Verification and Validation*. She has served as a reviewer for numerous international journals, conferences and books; served on the organizing or program committees for international conferences; and is a senior member of the IEEE Computer Society. Laura's research interests include software dependability and intelligent systems. She currently conducts research in the evaluation, verification and validation of predictive analytics and machine learning systems, as well as using "big data" and learning to provide insights into disease dynamics. She holds a BS in Mathematics; Master's degrees in Operations Research, Business Administration, and Geology; and a doctorate in Systems Engineering and Operations Research.

Ranjan Sinha is Head of Data Science Engineering & Technology for Customer Analytics and Personalization at eBay, San Jose. Earlier, as lead scientist at eBay, he led several business-impacting projects in recommendations and personalization, which have significantly enhanced consumers' shopping experiences. He has also contributed in domains such as infrastructure availability, security, and identity linking.Prior to joining eBay, Ranjan was a research academic and chief investigator at the University of Melbourne. He earned his PhD in computer science from RMIT University and has published over 30-refereed works, including in top-tier venues such as ACM JEA, ACM SIGMOD, Bioinformatics journal, IEEE Big Data, and VLDB journal. He was amongst the Top 12 Asia-Pacific Young Inventors and appeared in the article on *Cutting-Edge Crusaders* in the WSJ. He was awarded the Sort Benchmark medals for both JouleSort and PennySort in 2009. He presented a tutorial on *E-commerce Personalization at Scale* at the 2014 ACM CIKM conference and has been recently interviewed by KDnuggets. He is also a co-organizer of the *Bay Area Search Meetup* consisting of over 2,000 members. Ranjan's current interests include scaling data science solutions, introducing engineering practices in data science pipelines, developing real-time predictive analytic solutions, and in the application of semantic relationships between words in large text corpus.

Kristin M. Tolle, Ph.D. is the Director of the Data Science Initiative in Microsoft Research Labs. Since joining Microsoft in 2000, Dr. Tolle has acquired numerous patents and worked for several product teams including the Natural Language Group, Visual Studio, and the Microsoft Excel Team. Since joining Microsoft Research's outreach program in 2006, she has initiated and managed several major initiatives from Biomedical computing and environmental science to more traditional computer and information science programs around natural user interactions and data curation. She was also directed the development of the Microsoft Translator Hub and the Environmental Science Services Toolkit, among other successful research development projects. Dr. Tolle is a co-editor with Tony Hey and Stewart Tansley, as well as one of the authors of one of the earliest books on data science, The Fourth Paradigm: Data Intensive Scientific Discovery. She has two major areas in which she focuses, reducing time to scientific discovery by leveraging cloud computing and machine learning services, like Microsoft Azure and Azure ML and enabling the education of the next generation of data scientists.

Big Data Start-up Showcase

Chair: Duc Thanh Tran

With this special event, we aim to bring together entrepreneurs, startup enthusiasts, technologist and researchers to discuss the latest trends in technologies and to showcase cutting edge Big Data applications. Across verticals and covering major areas of Big Data applications, we have selected a handful of startups to demonstrate innovations in the usage of Big Data and/or good practice in the application of Big Data technologies.

The list of participants include:

1) **Tech (Storage & Querying):** SYSTAP: https://www.blazegraph.com/, a high performance graph database platform that provides supports RDF/SPARQL and the Tinkerpop stack (Blueprints/Gremlin) with scalable solutions including embedded, HA, scale-out, and GPU-acceleration.
Speaker: Brad Bebee, CEO

2) **Tech (Analytics):** Macrodata Lab, presenting http://www.bigobject.io/#/, an analytics database that transforms data tables to hierarchical objects, where In-Place Programming can directly apply to deliver a 100x to 1,000x speed improvement over traditional analytics methods.
Speaker: Jocelyn Chen, Co-founder Macrodata Lab

3) **App (Smart Data Linking):** https://urx.com/, presenting AppViews, giving app users a window into related content inside other app. AppViews API uses information such as keywords, location and time, and returns images, descriptions and deep links from a large index of apps extending across verticals like Music, Events and Rides and including services like SeatGeek, Spotify and Lyft.
Speaker: Delroy Cameron, Data Scientist

4) **App (Health Data Platform):** https://www.humanapi.com/, an integrated health data platform, enabling your users to securely share their health data with you, regardless of how that data was recorded, processed, or stored. Human API enables you to connect with the top wearable devices, health trackers, wireless scales, and fitness and nutrition apps.
Speaker: Ola Wiberg, CTO

5) **App (Fashion Data Platform):** http://www.tulemod.com/, a platform that uses cutting edge technology, from 3D scanning to database analytics, to change the way users, retailers and designers connect with fashion.
Speaker: Daniela Florescu, CEO

6) **App (Climate Forecast):** http://www.climformatics.com/ uses comprehensive data on current and predicted state of regional climate to assess future risk to your particular enterprise.
Speaker: Subarna Bhattacharyya, CEO

7) **App (Social Data Analytics)** http://www.451degrees.com/, the next-generation application delivery network that automatically optimizes content in real-time by analyzing user generated comments to enhance engagement, SEO, search traffic, and online advertising sales.
Speaker: Patrick Giblin CEO

Workshops

Big Humanities Data		
Time	**Title**	**Presenter/Author**
8:00am – 8.05am	**Introduction**	
8:05am - 8:25am	Mining Local Gazetteers of Literary Chinese with CRF and Pattern based Methods for Biographical Information in Chinese History	Chao-Lin Liu, Chih-Kai Huang, Hongsu Wang, and Peter K. Bol
8:25am - 8:45am	Metaphor Mining in Historical German Novels: An Unsupervised Learning Approach	Stefan Pernes
8:45am - 9:05am	The Coding of Literary Form: Data mining and the information structure of historical texts	Dallas Liddle
9:05am - 9:25am	Plot Arceology: a vector-space model of narrative structure	Benjamin Schmidt
9:25am – 9.45am	Scaling Out For Extreme Scale Corpus Data	Matthew Coole, John Mariani, and Paul Rayson
9:45am – 10.05am	A Method for Cross-Document Narrative Alignment of a Two-Hundred-Sixty-Million Word Corpus	Ben Miller, Jennifer Olive, Shakthidhar Gopavaram, Yanjun Zhao, Ayush Shrestha, and Cynthia Berger
10:05am - 10:20am	**Coffee Break**	
10:20am - 10:40am	Predicting Social Trends from Non-photographic Images on Twitter	Mehrdad Yazdani and Lev Manovich
10:40am - 11:00am	Towards a Mobile Social Data Commons	Giles Greenway, Leonard Mack, Tobias Blanke, Mark Cote, and Tom Heath
11:00am - 11:20am	Mixed-Initiative Social Media Analytics at the World Bank: Observations of Citizen Sentiment in Twitter Data to Explore "Trust" of Political Actors and State Institutions and its Relationship to Social Protest	Nadya Calderon, Brian Fisher, Jeff Hemsley, Billy Cescavich, Greg Jansen, Richard Marciano, and Victoria Lemieux
11:20am – 12.05am	**Panel Session: Stanford: CESTA, Big Data and Digital Humanities** **Panel members: Gabriel K. Wolfenstein, Elaine Treharne, Celena Allen, Benjamin Albritton**	

Workshop on Scalable Cloud Data Management		
Time	**Title**	**Presenter/Author**
8:00am-8:50am	Keynote: Purity and the future of scalable storage	Russel Sears (Microsoft)
8:50am-10:35am	**Session I: Data Management**	
	Workload-Driven Adaptive Data Partitioning and Distribution – The Cumulus Approach	Ilir Fetai (University of Basel)
	Comparison of Eager and Quorum-based Replication in a Cloud Environment	Ilir Fetai (University of Basel)
	Priority Register: Application-defined Replacement Orderings for Ad Hoc Reconciliation	Sathiya Prabhu Kumar
10:35am-10:50am	**Coffee Break**	
10:50am-12:35pm	**Session II: Cloud Systems**	
	A Generalized Flow for Multi-class and Binary Classification Tasks: An Azure ML Approach	Matthew Bihis (University of Washington)
	Fine-tuning the Consistency-Latency Trade-off in Quorum-Replicated Distributed Storage Systems	Wojciech Golab (University of Waterloo)
	Account Clustering in Multi-tenant Storage Management Environments	Gabor Madl (IBM Research)
12:35pm-1:35pm	**Lunch Break**	
1:35pm-3:55pm	**Special Session: Smart Data**	
	The Smart Data Technology Program and Open Challenges	Stefan Jaehnichen (Accompanying Research, FZI)
	Towards a Taxonomy of Smart Data Standards	Alexander Lenk

		(Accompanying Research, FZI)
	Standardization in the field of Big Data and Cyber Physical Systems (Industry 4.0)	Filiz Elmas (DIN e.V.)
	SAHRA – Smart Analyses Health Research Access	Matthieu-P. Schapranow (Hasso Plattner Institute)
3:55pm-4:10pm	**Coffee break**	
16:10-17:55	**Session III: Big Data**	
	Indexing Historical Spatio-Temporal Data in the Cloud	Chong Zhang (National University of Defense Technology)
	Push-based System for Molecular Simulation Data Analysis	Yicheng Tu (University of South Florida)
	Marlin: Taming the Big Streaming Data in Large Scale Video Similarity Search	Nan Zhu (McGill University)

Analysis, and Mining (BigGraph 2015)		
Time	**Title**	**Presenter/Author**
1:30 pm - 1:40 pm	Opening remarks	Fengguang Song
1:40 pm - 2:00 pm	DISTINGER: A Distributed Graph Data Structure for Massive Dynamic Graph Processing	Guoyao Feng, Xiao Meng, and Khaled Ammar
2:00 pm - 2:20 pm	A Fast Parallel Algorithm for Counting Triangles in Graphs using Dynamic Load Balancing	Shaikh Arifuzzaman, Maleq Khan, and Madhav Marathe
2:20 pm - 3:30 pm	Workshop Keynote: Distributed Graph Mining in Massive Networks	Mohammed J. Zaki
3:30 pm - 3:50 pm	**Coffee break**	
3:50 pm - 4:10 pm	Scalable Storage Structure for Pattern Matching on Big Graph Data	Janani Balaji and Rajshekhar Sunderraman
4:10 pm - 4:30 pm	LiteMat: a scalable, cost-efficient inference encoding scheme for large RDF graphs	Olivier Cure, Hubert Naacke, Tenfry Randriamalala, and Bernd Amann
4:30 pm - 4:40 pm	**Break**	
4:40 pm - 5:00 pm	Learning Classifiers from Remote RDF Data Stores Augmented with RDFS Subclass Hierarchies	Harris Lin, Ngot Bui, and Vasant Honavar
5:00 pm - 5:20 pm	MQuery: A Query Language for Scientific Meshes	Alireza Rezaei Mahdiraji and Peter Baumann

2nd International Workshop on Privacy and Security of Big Data (PSBD2015)		
Time	**Title**	**Presenter/Author**
1:30pm-1:45 pm	Opening	Alfredo Cuzzocrea, Ernesto Damiani
1:50pm-2:50pm	Keynote Speech	Ernesto Damiani
2:55 pm -4:10 pm (25mins for each)	**Session 1: Algorithms for Supporting Privacy and Security of Big Data** Chair: Alfredo Cuzzocrea, Jinoh Kim	
	Multi-Probe Random Projection Clustering to Secure Very Large Distributed Datasets	Lee C. Carraher, Philip A. Wilsey, Anindya Moitra, and Sayantan Dey
	Fast Summarization and Anonymization of Multivariate Big Time Series	Dymitr Ruta, Ling Cen, and Ernesto Damiani
	Heterogeneous k-Anonymization with High Utility	Katerina Doka, Mingqiang Xue, Dimitrios Tsoumakos, Panagiotis Karras, Alfredo Cuzzocrea, and Nectarios Koziris
4:15 pm:5:05 pm (25mins for each)	**Session 2: Big Data Privacy and Security in Cloud and Service-Oriented Environments** Chair: Lee Carraher	
	Current Security Threats and Prevention Measures Relating to Cloud Services, Hadoop Concurrent Processing, and Big Data	Ather Sharif, Sarah Cooney, Drew Vitek, and Shengqi Gong

	Security for Scientific Data Services Framework	Jinoh Kim, Bin Dong, Surendra Byna, and John Wu
5:10 pm:6:25 pm (25mins for each)	**Session 3: Big Data Privacy and Security Frameworks** Chair: Santosh Aditham	
	A Distributed Framework for Supporting Adaptive Ensemble-based Intrusion Detection	Alfredo Cuzzocrea, Gianluigi Folino, and Pietro Sabatino
	Simplifying Web Analytics for Digital Marketing	Dippy Aggarwal
	A Novel Framework for Mitigating Insider Attacks in Big Data Systems	Santosh Aditham and Nagarajan Ranganathan
	Concluding Remarks	Alfredo Cuzzocrea, Ernesto Damiani

IEEE Big Data Methodologies And Tools to Improve Big Data Projects		
Time	**Title**	**Presenter/Author**
1:00 pm-2:30 pm	**Session 1: Methodologies in Practice**	
	Business Information Modeling: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance	Torsten Priebe
	Exploring the Process of Doing Data Science Via an Ethnographic Study of a Media Advertising Company	Ivan Shamshurin
	Towards Methods for Systematic Research On Big Data	Manirupa Das
2:30pm-3:00 pm	**Break**	
3:00 pm- 4:30pm	**Session 2: Foundational Exploration**	
	Towards A Big Data Theory Model	**Marco Pospiech**
	Three Critical Matters in Big Data Projects for e-Science	Kerk Kee
	The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectivenes	Jeffrey Saltz
4:30 pm- 5:00pm	**Roundtable Discussion - Future Research and Next Steps**	

Data and Computational Science Technologies for Earth Science Research		
Time	**Title**	**Presenter/Author**
1:00-1:20 pm	Embedded Computing for the Data Lifecycle	Richard Doyle
1:20-1:40 pm	Taming Big Data Variety in the Earth Observing System Data and Information system	hristopher Lynnes Jeffrey Walter
1:40-2:00 pm	EROS Architecture Study and Roadmap	im Nelson Chris Engebretson
2:00-2:20 pm	Strategic Roadmap for Earth System Grid Federation	Dean Williams Michael Lautenschlager, etc
2:20-2:40 pm	In-Situ Analytics for Tomographic Imaging in Sensor Network	Goutham Kamath WenZhan Song
2: 40-3:00 pm	Optimization of system architecture for Big Data analysis in climate science	Huikyo Lee Luca Cinquini,etc
3:00-3:20 pm	sing Cloud Computing to Expediting Big Earth Data Processing	Chaowei (Phil) Yang
3:20-3:40 pm	**Break**	
3:40-4:00 pm	DeepDive: Building Knowledge Base Construction Systems for Macroscopic Sciences	Ce Zang
4:00-4:20 pm	Constrained Region Selection method based on Configuration Space for Visualization in Scientific Data Set Search	Shinichi Takeuchi Komei Sugiura, etc
4:20-4:40 pm	NEXUS: A Framework for Earth Science Data Analysis	Thomas Huang Michael Gangl, etc
4:40-:5:00 pm	Ontology-driven data access at the NASA Earth Exchange	Beth Huffer Marc Cotnoir, etc
5:00-5:20 pm	Enhancing Scientific Support in SQL	Peter Baumann Dimitar Misev
5:20-5:40 pm	Advanced Information Systems Big Data Study for Earth Science	Daniel Crichton Michael Little

Advances in high dimensional big data *(Venus: Ballroom H)*		
Time	**Title**	**Presenter/Author**
8:15am – 10:10am	**Session I**	
8:15am – 8:55am	Industry Keynote talk	Dimitris Tasoulis *Senior Execution Researcher, Winton Capital Management, UK*
8:55am-10:10am	Modeling Community Detection using Slow Mixing Random Walks	Ramezan Paravi Torghabeh, etc
	Dimensional Scalability of Supervised and Unsupervised Concept Drift Detection: An Empirical Study	Jorge David Destephen Lavaire, etc
	Efficient Change Detection for High Dimensional Data Streams	Spiros Georgakopoulos, etc
10:10 am -10:30 am	**Coffee Break**	
10:30 am – 12:25pm	**Session II**	
10:30am-11:10am	Keynote Talk: Academic Keynote Talk	Bin Yu *(UC Berkeley, USA)*
11:10am-12:25pm	Big Data Analytics for Demand Response: Clustering Over Space and Time	Charalampos Chelmis, Jahanvi Kolte, etc
	Finding Banded Patterns in Big Data Using Sampling	Fatimah Abdullahi, Frans Coenen,etc
	Scalable preference queries for high-dimensional data using map-reduce	Gheorghi Guzun, Joel Tosado
12:25am-12:55am Tutorial Talk	Hashing Algorithms for Large-Scale Search and Learning	Ping Li *(Rutgers USA)*

The First International Workshop on Mining Big Data in Social Networks (MBD-SONET)		
Time	**Title**	**Presenter/Author**
8:50am-9:00am	Workshop Overview	Bin Zhou
9:00am-9:20am	Discovering Time-evolving Influence from Dynamic Heterogeneous Graphs	Chuan Hu
9:20am-9:40am	Measuring Influence Across Multiple Social Networks	Adithya Rao
9:40am-10:00am	Efficient Approximation Algorithms to Determine Minimum Partial Dominating Sets in Social Networks	Traian Marius Truta
10:00am-10:20am	**Coffee Break (Meeting Room Foyer)**	
10:20am-10:40am	Sentiment Expression via Emoticons on Social Media	Hao Wang
10:40am-11:00am	Identifying Actionable Messages on Social Media	Nemanja Spasojevic
11:00am-11:20am	How much is your information worth - A method for revenue generation for your information	Ng Wee Keong
11:20am-11:40am	Combining Activity-evaluation Information with NMF for Trust-link Prediction in Social Media	Kanji Matsutani
	Lunch Time	
1:30pm-1:50pm	Efficient Large Scale Distributed Matrix Computation with Spark	Rong Gu
1:50pm-2:10pm	The value of analytical queries on Social Networks	Michel De Rougemont
2:10pm-2:30pm	A Set Theoretical Approach to Big Social Data Analytics of Real-World Events	Ravi Vatrapu
2:30pm-2:50pm	On Compressing Massive Streaming Graphs with Quadtrees	Michael Nelson
2:50pm-3:10pm	Characterizing super spreading in microblog: an epidemic-based model	Yu Liu
3:10pm-3:30pm	A Community Detection Method Based on K-shell	Yang Wang
3:30pm-3:50pm	**Coffee Break (Meeting Room Foyer)**	
3:50pm-4:10pm	Finding Community Structure via Rough K-Means in Social Network	Yunlei Zhang
4:10pm-4:30pm	Dynamic Community Detection based on Game Theory in Social Networks	Fei Jiang

Joint Program Schedule First Workshop on Data-Centric Infrastructure for Big Data Science & 3rd Workshop on Distributed Storage Systems and Coding for Big Data		
Time	**Title**	**Presenter/Author**
1:30pm-1:40pm	**Plenary**	
1:40pm-2:20pm	Keynote Speech : Data Federation and Data Management for the LHC Experiments ATLAS and CMS)	Frank Wuerthwien *University of California San Diego/San Diego Supercomputing Center*
2:20pm-2:45pm	Network-Aware Resource Management for Scalable Data Analytics Frameworks Thomas	Renner, Lauritz Thamsen, Odej Kao
2:45pm-3:10pm	Lambda Architecture for Cost-effective Batch and Speed Big Data Processing	Mariam Kiran, Peter Murphy, Inder Monga, Jon Dugan, Sartaj Singh Baveja
3:10pm-3:35pm	On a New Approach to the Index Selection Problem using Mining Algorithms	Parinaz Ameri, Jorg Meyer and Achim Streit
3:35pm-3:50pm	**Coffee Break**	
3:50pm-4:15pm	RADII: Resource Aware Datacentric collaboratIon Infrastructure	Claris Castillo, Fan Jiang, Charles Schmitt, Ilya Baldin, Arcot Rajasekar
4:15pm-4:40pm	A Comprehensive Evaluation of NoSQL Datastores in the Context of Historians and Sensor Data Analysis	Arun Kumar Kalakanti, Vinay Sudhakaran, Varsha Raveendran, and Nisha Menon
4:40pm-5:05pm	On the Implementation of Zigzag Codes for Distributed Storage System	Lijia Lu, Hui Li, Jun Chen, Bing Zhu, Weijuan Yin
5:05pm-5:30pm	Challenges and Opportunities on Network Resource Management in DCN with SDN	Guan Xu, Jun Yang, and Bin Dai

2nd Workshop on Advances in software and hardware for big data to knowledge discovery (ASH)		
Time	**Title**	**Presenter/Author**
8:00am - 8:15am	**Welcome and Workshop Introduction**	
8:15am - 8:40am	Regularized and Sparse Stochastic K-Means for Distributed Large-Scale Clustering	Vilen Jumutc, Rocco Langone, and Johan Suykens
8:40am - 9:05am	Join Algorithms on GPUs: A Revisit After Seven Years	Ran Rui, Hao Li, and Yicheng Tu,
9:05am - 9:30am	Performance Evaluation of Enabling Logistic Regression for Big Data with R	Ruizhu Huang and Weijia Xu
9:30am - 9:55am	A novel symbolization technique for time-series outlier detection	Gavin Smith and James Goulding
9:55am - 10:20am	**Coffee Break**	
10:20am - 10:45am	Shaping Data: Visualization Under Construction	Oliver Bieh-Zimmert and Carsten Felden
10:45am - 11:10am	Visual Analysis of Large-scale LiDAR Point Clouds	Hui Zhang
11:10am - 11:35am	Immersive Visualization for Materials Science Data Analysis using the Oculus Rift	Margaret Drouhard, Chad Steed, Steven Hahn, Thomas Proffen, Jamison Daniel, and Michael Matheson
11:35am – 12:00 am	Texture-Based Edge Bundling: A Web-Based Approach for Interactively Visualizing Large Graphs	Jieting Wu, Lina Yu, and Hongfeng Yu
12 :00pm- 1:30pm	**Lunch Break**	
1:30 pm- 1:55pm	A database-based distributed computation architecture with Accumulo and D4M: an application of eigensolver for large sparse matrix	Yin Huang, Yelena Yesha, and Shujia Zhou

1:55pm -2:20pm	Wrangler's User Environment	Christopher Jordan, David Walling, Weijia Xu, Stephen Mock, Dan Stanzione, and Niall Gaffney
2:20pm - 2:45pm	Big Data Provenance: Challenges, State of the Art and Opportunities	Jianwu Wang, Daniel Crawl, Shweta Purawat, Mai Nguyen, and Ilkay Altintas
2:45pm - 3:10pm	Scalable Dental Computing on Cyberinfrastructure	Hui Zhang and Riqing Chen
3:10pm - 3:30pm	**Coffee Break**	
3:30pm – 3:55pm	Spatio-Temporal Similarity Search Method for Disaster Estimation	Hideki Hayashi, Akinori Asahara, Natsuko Sugaya, Yuichi Ogawa, and Hitoshi Tomita
3:55pm - 4:20pm	Volatility Matrix Inference in High-Frequency Finance with Regularization and Efficient Computations	Jian Zou, Yunbo An, and Hong Yan
4:20pm - 4:45pm	Skill Grouping Method: Mining and Clustering Skill Differences from Body Movement BigData	Shinichi Yamagiwa, Yoshinobu Kawahara, Noriyuki Tabuchi, Yoshinobu Watanabe, and Takeshi Naruo
4:45pm - 5:30pm	**Closing Remark / Discussions and Networking**	

Big Data for B2B Marketing and Sales (Oct 29)		
Time	**Title**	**Presenter/Author**
1:30pm - 5:30pm	From single node MySQL to cluster running Impala/Hadoop/HDFS: creating a scalable data pipeline	Nisheeth Ranjan *CTO and co-founder, BrightFunnel*
	Learning Safe Strategies in Digital Marketing	Mohammad Ghavamzadeh *Senior Analytics Researcher, Adobe Research*
	Mitigating selection bias in social media for explanatory analytics	Arnau Tibau Puig *Data Scientist, Quantfind*
	Application of predictive analytics for sales pipeline evaluation and forecasting	Venkat Rangan *CTO and co-founder, Clari*
	Modeling a sales funnel and lead behavior for predictive lead scoring	Brendan Duncan *Data Scientist, Fliptop & LinkedIn*
	Predictive Analytics for B2B Customer Success	Sherif Botros, *Chief Data Scientist, Gainsight*
	Panel Discussion	All participants

Mining Big Data to Improve Clinical Effectiveness		
Time	**Title**	**Presenter/Author**
8:00am-8:25am	A Data-Driven approach towards Patient Identification for Telehealth Programs	Martha Ganser
8:25am-8:50am	Ensemble Prediction of Vascular Injury in Trauma Care	Max Metzger
8:50am-9:00am	**Break**	
9:00am-10:00am	Invited Talk	Keith Marsalo
10:00am-10:10am	**Break**	

10:10am-10:35am	Using Clinical Data, Hypothesis Generation Tools and PubMed Trends to Discover the Association between Diabetic Retinopathy and Antihypertensive Drugs	Katherine Senter
10:35am-11:00am	Enabling Graph Appliance For Genome Assembly	Rina Singh
11:00am-11:25am	M-SEQ: Early Detection of Anxiety and Depression via Temporal Orders of Diagnoses in Electronic Health Data	Jinghe Zhang

Data Quality Issues in Big Data		
Time	**Title**	**Presenter/Author**
10:00am - 10:30 am	**Opening remarks**	**Vijay Raghavan**
10:00 am - 10:30 am	Mining Telecom Data: Lessons Learned in the Data Cleaning Phase (Keynote Talk)	Hui Zang, Chief Data *Scientist at American Software Laboratory of Huawei Technologies, USA*
10:30 am - 10:47 am	A Framework for Consensual and Online Privacy Preserving Record Linkage in Real-Time	Daniel Müller, Stefan Mau, and Irena Pletikosa Cvijikj, ETH Zurich, Switzerland
10:47 am - 11:04 am	A Memory Capacity Model for High Performing Data-filtering Applications in Samza Framework	Tao Feng, Zhenyun Zhuang, Yi Pan, and Haricharan Ramachandra, *LinkedIn Corp, Mountain View, California, USA*
11:04 am - 11:21 am	Robust and Distributed Web-Scale Near-Dup Document Conflation in Microsoft Academic Service	Chieh-Han Wu and Yang Song, Microsoft Research, *Redmond, Washington, USA*
11:21 am - 11:38 am	Evaluation of Data Quality of Multisite Electronic Health Record Data for Secondary Analys	Alicia Nobles, Ketki Vilankar, Hao Wu, and Laura Barnes *University of Virginia, USA*
11:38 am - 11:50 am	**Break**	
11:50 am - 12:07 pm	CrowdMD: Crowdsourcing-based Approach For Deduplication	Soror SAHRI, Mourad OUZIRI, and Salima BENBERNOU, Universit´e Paris Descartes Sorbonnes Paris Cit´e, France
12:07 pm - 12:24 pm	Data Veracity Estimation with Ensembling Truth Discovery Methods	Laure Berti-Equille, *Qatar Computing Research Institute, Doha, Qatar*
12:24 pm - 12:41 pm	Distributed Life Cycle Scheduling For Cascaded Data Processing	Lavanya Sainik, *Ericsson India Global Services Pvt. Ltd., India*
12:41 pm - 12:58 pm	Big Data, Big Data Quality Problem	David Becker, Bill McMullen, and Trish King *The MITRE Corporation, USA*
12:58 pm - 1:15 pm	Data Quality Issues in Big Data	Dhana Rao, Venkat Gudivada *(East Carolina University), USA, Vijay Raghavan, University of Louisiana at Lafayette, USA*
1:15 pm - 1:20 pm	Closing Remarks	Vijay Raghavan, *University of Louisiana at Lafayette, USA*

Big Data for Sustainable Development		
Time	**Title**	**Presenter/Author**
1:30pm-1:40 pm	Opening remarks	Prof. Aki-Hiro Sato *Kyoto University, Chair*

	Facilitator	Dr. Hidefumi Sawai *NICT, Co-Chair*
1:40pm-2:40 pm	**Statistical Methods**	
	Microdata Analysis of the Accommodation Survey in Japanese Tourism	Aki-Hiro Sato *Kyoto University*
	Using Fisher Information in Big Data	Nasir Ahmad, etc *University of Illinois at Chicago*
2:40pm-2:45pm	**Short Break**	
2:45pm -3:45 pm	**Social Media and its Platform**	
	Detecting Rumor Patterns in Streaming Social Media	Shihan Wang *Tokyo Institute of Technology*
	A Spatio-temporal Multimedia Big Data Framework for a Large Crowd	Bilal Sadiq,etc *Umm Al Qura University*
3:45pm -3:50pm	**Short Break**	
3:50pm-5:20pm	**Specific Fields**	
	A Collaborative Framework for Annotating Energy Datasets	Hông-Ân Cao,etc *ETH Zurich*
	The Relation between Firm Age Distributions and the Decay Rate of Firm Activities in the United States and Japan	Shouji Fujimoto,etc *Kanazawa Gakuin University*
	An Epidemic Simulation with a Delayed Stochastic SIR Model Based on International Socioeconomic-Technological Databases	Aki-Hiro Sato, etc *Kyoto University*
5:20pm-5:30pm	Closing Remarks	Aki-Hiro Sato *Kyoto University, Chair*

Big Data in the Geosciences		
Time	**Title**	**Presenter/Author**
8:00am - 8:10am	**Introduction**	
8:10am - 8:30am	High Performance Dynamic Analysis of Big Spatial Data	David Haynes Suprio Ray, etc
8:30am - 8:50am	WDCloud: An End to End System for Large-Scale Watershed Delineation on Cloud	In Kee Kim Jacob Steele etc
8:50am - 9:10am	Climate Model Diagnostic Analyzer	Seungwon Lee Lei Pan, etc
9:10am - 9:30am	Light-Weight Parallel Python Tools for Earth System Modeling Workflows	Kevin Paul, Sheri Mickelson, etc
9:30am – 10:00am	Spatio-temporal Queries in HBase	Xiaoying Chen, Chong Zhang, etc
10:00am - 10:20am	**Coffee Break**	
10:20am - 10:40am	Optimizing Apache Nutch For Domain Specific Crawling at Large Scale	Luis Alberto Lopez, Ruth Duerr Siri Jodha Khalsa
10:40am - 11:00am	Is Apache Spark Scalable to Seismic Data Analytics and Computations?	Yuzhong Yan, Lei Huang, etc
11:00am - 11:20am	On the Efficient Evaluation of the Array Joins	Peter Baumann Vlad Merticariu
11:20am - 11:40am	Enabling Scientific Data Storage and Processing on Big-data Systems	Saman Biookaghazadeh, Yiqi Xu, etc
11:40am- 12:00pm	Estimating the Impacts of Extreme Earthquakes on Road Networks	Amirhassan Kermanshah, Alireza Karduni, etc
12 :00pm- 1:30pm	**Lunch**	
1:30 pm- 1:50pm	Component Based Dataflow Processing Framework	Andrey Vakhnin, Ricardo Oyarzun,etc
1:50pm -2:10pm	Earth Science Data Fusion with Event Building Approach	Constantine Lukashin, Aron Bartle, etc
2:10pm - 2:30pm	An Optimized Interestingness Hotspot Discovery Framework for Large Gridded Spatio-temporal Datasets	Fatih Akdag Christoph F. Eick,etc
2:30pm - 2:50pm	SciSpark: Applying In-memory Distributed Computing to Weather Event Detection and Tracking	Rahul Palamuttam, Renato Marroquín Mogrovejo,etc

2:50pm - 3:10pm	Detecting environmental disasters in digital news archives	Amelia Yzaguirre, Robert H. Warren,etc
3:10pm - 3:30pm	A Hadoop-Based Visualization and Diagnosis Framework for Earth Science Data	Shujia Zhou, Xi Yang, etc
3:30pm - 3:50pm	**Coffee Break**	
3:50pm - 4:10pm	International Standard OGC Moving Features to address "4Vs" on locational BigData	Akinori Asahara, Hideki Hayashi,etc
4:10pm - 4:30pm	Integrating Big Geoscience Data into the Petascale National Environmental Research Interoperability Platform (NERDIP): successes and unforseen challenges	Lesley Wyborn Benjamin Evans
4:30pm - 5:30pm	**Panel Discussion - Big Data Needs and Future Drivers**	
5:30pm - 6:30pm	**Group Discussion, Demos, and Wrap Up**	

Deriving Value from Big Data in Healthcare		
Time	**Title**	**Presenter/Author**
8:15 am-8:20 am	Opening remarks	Vahid Taslimitehrani *Wright State University, Kno.e.sis*
8:20 am -9:10 am	Watson in the Age of Discovery - How cognitive computing helps accelerate breakthroughs (Keynote)	Dr. Meena Nagarajan *IBM Watson Innovation*
9:10 am -9:35 am	Machine Learning for Stress Detection from ECG Signals in Automobile Drivers	*Neha Keshan, etc*
9:35 am -10:00 am	Sequential Pattern Mining of Electronic Healthcare Reimbursement Claims: Experiences and Challenges in Uncovering How Patients are Treated by Physicians	Kunal Malhotra, etc
10:00 am -10:20 am	**Coffee break**	
10:20 am -11:10 am	Transfer Learning and Survival Regression Methods for Patient Risk Prediction (Keynote)	Dr. Chandan Reddy *Wayne State University*
11:10 am -11:35 am	SQL-Like Big Data Environments: Case study in Clinical Trial Analytics	Akshay Grover, etc
11:35 am -12:00 am	Exploring Spatio-Temporal-Theme Correlation between Physical and Social Streaming Data for Event Detection and Pattern Interpretation from Heterogeneous Sensors	Minh-Son Dao, etc

IEEE Workshop on Big data Analytics in Manufacturing and Supply Chains		
Time	**Title**	**Presenter/Author**
1:30pm - 1:40pm	**Workshop Overview**	
1:40pm - 2:00pm	Forecast UPC-Level FMCG Demand, Part I: Exploratory Analysis and Visualization	
2:00pm - 2.20pm	Forecast UPC-Level FMCG Demand, Part II: Hierarchical Reconciliation	
2:20pm - 2:40pm	Dynamic Aggregation for Time Series Forecasting	
2:40pm - 3:00pm	Big Data Analytics for Empowering Milk Yield Prediction in Dairy Supply Chains	
3:00pm - 3:20pm	**Break**	
3:20pm - 3:40pm	Sparsity Adjusted Information Gain for Feature Selection in Sentiment Analysis	
3:40pm - 4:00pm	A Data Fusion Framework for Large-Scale Measurement Platforms	
4:00pm - 4:20pm	Sensor Event Mining with Hybrid Ensemble Learning and Evolutionary Feature Subset Selection Model	
4:20pm - 4:40pm	Graph-Based Analysis of Resource Dependencies in Project Networks	
4:40pm - 5:00pm	Profit Estimation Error Analysis in Recommender Systems based on Association Rules	
5:00pm - 6:30pm	Networking	

Second Hands-On Workshop on Leveraging High Performance Computing Resources for Managing Large Datasets		
Time	**Title**	**Presenter/Author**
8:30am-8:45 am	**Morning Session**	
8:35am -8:45am	Opening remark	
8:45 am -9:15 am	Managing large datasets - some challenges and solution strategies	
9:15am – 9:45 am	The national cyberinfrastructure and resources related to managing large datasets:	
9:45am -10:15 am	Introduction to Linux and user environment on Stampede supercomputer, part-1 (hands-on session)	
10:15 am – 10:30 am	**Break**	
10:30 am – 11:10 am	Introduction to Linux and and user environment on Stampede supercomputer, part-2 (hands-on session):	
11:10 am – 12:10 am	Downloading and installing the DROID tool for metadata extraction	
12:10 pm – 01:30 pm	**Lunch Break**	
1:30 pm – 2:00 pm	Overview of the protocols for data transfer	
2:00 pm – 03:15 pm	Hands-on exercises on data transfer, metadata extraction and checksum calculation on a remote HPC/HTC resource:	
3:15 pm – 3:30 pm	**Break**	
3:30 pm – 4:30 pm	Hands-on exercise on information visualization using Tableau	
4:30 pm – 6:00 pm	Introduction to databases and how to use them on remote HPC/HTC/high-end storage resources	

Specials Sessions

SPECIAL SESSION I:
From Data to Insight: Big Data and Analytics for Advanced Manufacturing Systems

October 29, 2015

Panel I: Research, Technology and Development Challenges and Deployment.

Time	Event
9:00am – 9:15am	**Opening remarks**: Dr. Sudarsan Rachuri, NIST, Gaithersburg, MD.
9:15am – 9:45am	**Keynote Speech 1**: Data Analytics for Smart Manufacturing Systems *Prof. Kincho Law, Stanford Univ., Stanford, CA.*
9:45am – 10:15am	**Keynote Speech 2**: Industrial Big Data and Predictive Analytics for Future Smart Systems *Prof. Jay Lee, Univ. of Cincinnati, Cincinnati, OH.*
10:15am – 10:40am	**Coffee Break**
10:40am – 11:00am	Panelist presentation (4x5 minutes each)
11:00am – 12:45pm	Discussion
12:45pm – 2:15pm	**Lunch**

Panel Two – Industry use cases, standards and best practices

Time	Event
2:15pm – 2:45pm	**Keynote Speech 3**: TBD *Dr.* Michael Zeller *CEO*, Zementis
2:45pm – 3:15pm	**Keynote Speech 4**: Big Data and Analytics for Advanced Manufacturing Systems – Network Enabled Manufacturing (NEM) at Boeing**.** *Dr. Al Salour, The Boeing Company, St. Louis, Missouri.*
3:15pm – 3:40pm	**Coffee Break**
3:40pm – 4:00pm	Panelists' Presentations (4x5 minutes each)
4:00pm – 5:35pm	Discussion
5:35pm – 6:00pm	Wrap-up, Dr. Sudarsan Rachuri & Dr. Rumi Ghosh
6:00pm – 8:30pm	Poster Session & Reception (TBD)

Details for Panel I:

Moderator: Dr. Sudarsan Rachuri (NIST)

Details for Panel I:
Moderator: Dr. Sudarsan Rachuri (NIST)
Panelists:
Prof. Soundar Kumara (Penn State), Prof. Robert Grossman (Uni. of Chicago), Prof. Jay Lee (University of Cincinnati), Prof. Kincho Law, Prof. Jingshan Li (Univ. of Wisconsin-Madison), Wade Shen (DARPA).

Details for Panel II
Moderator: Dr. Steve Eglash
Panelists:
Dr. Michael Zeller (Zementis), Neil Eklund (Schlumberger), Hauke Schmidt (Bosch), Dr. Al Salour (Boeing) , Brinda Thomas (Tesla).

- **Keynote 1: Data Analytics for Smart Manufacturing Systems**

Speaker: **Prof. Kincho Law, Stanford Univ., Stanford, CA.**

Abstract:

This presentation will discuss a joint research activity with the Laboratory for Manufacturing and Sustainability directed by Prof. David Dornfeld at UC Berkeley and the National Institute of Standards and Technology towards the development of data-driven modeling methodologies and associate tools to assess, predict, and optimize operational performance and energy consumption of machineries and manufacturing systems. Taking advantage of the input and output measurement data collected directly from an operating machine, data driven models can be built to characterize the machine. In this talk, I will describe the development of machine monitoring system that collects machine operation data and converts the raw data into "derived" data with features and process parameters. Using milling machine as an example, I will discuss the use of input-output measurement data to construct predictive model for energy consumption and toolwear analysis.

Short Bio:

Kincho H. Law received his B.Sc. Civil Engineering and B.A. Mathematics (1976) from the University of Hawaii, and M.S. in Civil Engineering (1979) and Ph.D. Civil Engineering (1981) from Carnegie Mellon University. He joined Stanford University in 1988 and is currently Professor of Civil and Environmental Engineering. Prof. Law's professional and research interests focus on computational and information science in engineering. His work has dealt with various aspects of computational science and engineering; sensing, monitoring and control of complex systems; legal and engineering informatics; engineering enterprise integration; web services and internet computing.

- **Keynote 2: Industrial Big Data and Predictive Analytics for Future Smart Systems**

Speaker: **Prof. Jay Lee, Univ. of Cincinnati, Cincinnati, OH.**

Abstract:

As more software and embedded intelligence are integrated in industrial products and systems, predictive technologies can further intertwine intelligent algorithms with electronics and tether-free intelligence to predict product performance degradation and autonomously manage service needs. The presentation will address the trends of industrial transformation in big data environment as well as the readiness of smart predictive informatics tools to manage big data to achieve resilient and smart self-aware systems. First, industry transformation including Germany Industry 4.0, Industrial Internet, and Cyber-Physical System (CPS) will be introduced. Second, advanced predictive analytics technologies for smart product manufacturing and service systems with case studies will be presented. In addition, research advances in designing cyber-physical model for smart product service systems with several case studies will be discussed. Finally, dominant innovation® methodology for service innovation will be discussed.

Short Bio:

Dr. Jay Lee is Ohio Eminent Scholar, L.W. Scott Alter Chair Professor, and Distinguished Univ. Research Professor at the Univ. of Cincinnati and is Founding Director of National Science Foundation (NSF) Industry/University Cooperative Research Center (I/UCRC) on Intelligent Maintenance Systems (IMS www.imscenter.net) which is a multi-campus NSF Industry/University Cooperative Research Center which consists of the Univ. of Cincinnati (lead institution), the Univ. of Michigan, Missouri Univ. of S&T, and Univ. of Texas-Austin. The Center has developed partnerships with over 80 companies from 15 countries since its inception in 2001. In addition, he has mentored his students and developed a spin-off company Predictronics with support from NSF Innovation ICorps Award in 2012. In 2013, he served as an advisory committee member for White House Cyber Physical Systems (CPS) American Challenge Program.

His current research focuses on Industrial Big Data Analytics, Cyber-Physical Systems, as well as Prognostics and Health Management (PHM). He developed the well-known Watchdog Agent® (s systematic platform for industrial data analytics toolbox

used by over 85 global companies) as well as the Dominant Innovation™ (a methodology for product and service innovation design and has been used by many Fortune 500 companies).

Currently, he also serves as advisor to a number of global organizations, including a member of the Manufacturing Executive Leadership Council, member of International S&T Committee of Alstom Transport, France, Scientific Advisory Board of Flanders' MECHATRONICS Technology Centre (FMTC) in Leuven, Belgium, Scientific Advisor Board of SIMTech, Singapore, Member of Advisory Committee of MIRDC Taiwan, etc. In addition, he serves as editors and associate editor for a number of journals including IEEE Transaction on Industrial Informatics, Int. Journal on Prognostics & Health Management (IJPHM), etc.

He received a number of awards including the most recent NSF Alex Schwarzkopf Technological Innovation Prize in Jan. 2014 and the Prognostics Innovation Award from National Instruments in 2012. He is also an honorary advisor to the Heifer International-a charity organization working to end hunger and poverty around the world by providing livestock and training to struggling communities.

- **Keynote 3:** Faster insights through open standards for predictive analytics.

Speaker: **Dr.** Michael Zeller**, CEO,** Zementis

Abstract:
Fast and efficient operational model deployment is a key enabler for driving adoption of Big Data and Internet of Things applications. The Predictive Model Markup Language (PMML) industry standard accelerates the process of moving machine learning models from the data scientists' desktop into enterprise IT systems. It reduces cost and complexity typically associated with such projects while allowing us turn big data into even bigger business value through more intelligent decisions.

Short Bio:

Dr. Michael Zeller is the CEO of Zementis, a software company focused on the operational deployment of predictive analytics. Zementis was recognized by CIO Review as one of the "Top 20 most promising Big Data companies in 2013" and named "Cool Vendor in Data Science" by Gartner in 2014. Michael currently also serves on the Board of Directors of Software San Diego and as Secretary/Treasurer on the Executive Committee of ACM SIGKDD, which is the premier international organization for data mining researchers and practitioners from academia, industry, and government.

Previously, Michael served as CEO of OTW Software, a company focused on implementing structured software engineering processes and delivering object-oriented analysis and design services. Prior to his engagement at OTW, he held the position of Director of Engineering for an aerospace firm, managing the implementation of IT solutions for major aerospace corporations. Michael received a Ph.D. in Physics from the University of Frankfurt (Germany), with emphasis in the development of neural networks, robotics, and human-computer intelligent interaction. Michael also received a visiting scholarship from the Department of Physics at the University of Illinois at Urbana-Champaign and was the recipient of a Presidential Postdoctoral Fellowship from the Computer Science Department at the University of Southern California.

Details for Panel II:

Moderator: Brinda Thomas (Tesla) /Dr. Steve Eglash

- **Keynote 4: Big Data and Analytics for Advanced Manufacturing Systems – Network Enabled Manufacturing (NEM) at Boeing.**

Speaker: **Dr. Al Salour, the Boeing Company, St. Louis, Missouri.**

Abstract:

The Boeing Company has initiated the network enabled manufacturing (NEM) development strategy for its production systems since 2007. NEM is a systems approach to manufacturing controls that will rely on real-time data from sensors and devices to drive factory floor operations. It connects materials, assembly parts/kits, tools, equipment, facilities, location systems, metrology systems, and people to an integrated network to efficiently manage aircraft production. In most cases, real time machine or process data is collected, analyzed, and interfaced with other legacy systems to simplify decisions and reduce/eliminate non-value added tasks. As an example many of the transactions for material receipts and point of use kit item deliveries have been automated to streamline inventory management. At the same time critical machinery at the fabrication and assembly centers are monitored in real time to report their overall equipment effectiveness (OEE) and detect faults via on-board sensors. Machine and process health data are integrated together in advanced portal systems to help the first line managers in scheduling their daily work and determining their priorities. As the industry and academia explore new research opportunities for utilizing sensor based smart systems and continue developing solutions for the future factories, implementation costs and risks will drastically go down. Technology initiatives are often triggered through service providers who can offer solutions in hardware and software for machine learning, process health monitoring, situational awareness, data analytics, simulation modeling, and integration/visualization techniques. This presentation will cover the roadmap plans and trends in manufacturing data acquisition and how this evolution can influence our current and future production systems.

Short Bio:

Dr. Salour is a Boeing Technical Fellow and the enterprise leader for the Network Enabled Manufacturing technologies. He is responsible for the systems approach to develop, integrate, and implement sensor based strategies and plans for Boeing's current and future production systems. His contributions are in machine monitoring, RFID systems, product definition, work instructions optimization, asset management, and networking. Dr. Salour focuses on real time actionable data to significantly reduce daily decisions associated with the build process, quality, and delivery. His sphere of influence spans toward automating and mistake proofing production data and creating self-aware systems that will lead to cost savings and promote manufacturing as a competitive advantage. Dr. Salour is the research investigator with national and international premiere universities and research labs and holds numerous patents and disclosures in manufacturing technologies.

- **Keynote 5: Big Data and Industry 4.0**

Speaker: **Scott C. Hibbard, Vice President of Technology – Factory Automation Bosch Rexroth Corporation Hoffman Estates, Illinois.**

Abstract:
Robert Bosch was one of two chairs of the committee appointed by the German Federal Ministry of Education & Research to develop the concept of what would become the Industry 4.0 initiative. Every business unit within Bosch is committed to the advancement of a connected industry, and incorporates the principals of Industry 4.0 into their daily business activities. Bosch Rexroth, as the provider of drive and motion control solutions to the market, as well as to sister business units within Bosch, is no exception. The control system on a piece of manufacturing equipment plays a key role in effecting Industry 4.0 solutions. On modern sensor-driven machinery, the control is also in an ideal position to both analyze Big Data to improve manufacturing productivity, as well as act as a conduit between the data sources and higher level analytic engines. This presentation will provide insight into the activities occurring within Bosch Rexroth to advance Industry 4.0 and Big Data analytics.

Short Bio:

Scott Hibbard has been associated with the industrial motion control industry for the past 37 years, 34 of those with Bosch Rexroth, including the former Indramat Division of the Rexroth Corporation. Scott played a key role in the pioneering introduction of the first practical industrial brushless AC servo drives and multi-processing CNCs used in high production manufacturing, as well as moving printing and packaging automation technology from mechanical systems to precision electronic motion control. His current responsibilities include management of Rexroth's facility in Hoffman Estates, as well as leading product management and development for Bosch Rexroth's factory automation markets. Scott is currently active in the development of the Industrial Internet (Industry 4.0) focusing on standards enabling meaningful interchange of information across the manufacturing environment.

Scott serves on the board of directors for Sercos North America, the trade association that stewards the ongoing development and adoption of Sercos, the international real time interface standard. Scott is also chairman of the Association for Manufacturing Technology's Technical Issues Committee, and chairs the MTConnect Technical Advisory Group.

Scott graduated from DeVry Technical Institute in 1977. He is the author of numerous technical articles and papers on AC servo, digital drive control, and open control system technologies, as well as energy efficiency in factory automation.

October 30, 2015

Technical Papers' Presentations Session

Time	Title	Author/Presenter
	Session I **Session Chair: Dr. Ronay Ak**	
10:05am-10:35am	**Keynote Speech**: Big Data and Industry 4.0	Scott Hibbard *Vice President of Technology – Bosch Rexroth Corp., Hoffman Estates, IL*
10:35am – 11:00am	Business Understanding, Challenges and Issues of Big Data Analytics for the Servitization of a Capital Equipment Manufacturer	Mikel Niño, José Miguel Blanco, and Arantza Illarramendi,
11:00am– 11:25am	Analysis and Optimization in Smart Manufacturing: Toward Standards on Reusable Knowledge Base of Process Performance	Alexander Brodsky, Guodong Shao, Mohan Krishnamoorthy, Anantha Narayanan, Daniel Menascé, and Ronay Ak
11:25am – 11:50am	Real-Time Energy Prediction for a Milling Machine Tool Using Sparse Gaussian Process Regression	Jinkyoo Park, Raunak Bhinge, Mason Chen, Kincho Law, David DorJinkyooarsan Rachuri
11:50am – 12:15pm	Big Data Process Analytics for Continuous Process Improvement in Manufacturing	Nenad Stojanovic, Marko Dinic, and Ljiljana Stojanovic
12:15pm – 12:40pm	Data Driven Predictive Analytics for A Spindle's Health	Divya Sardana, Raj Bhatnagar, Radu Pavel, and Jonathan Iverson,
12:40pm –2:00pm	**Lunch**	
	SessionII **Session Chair: Prof. Sagar Kamarthi**	
2:00pm–2:25pm	Automated Uncertainty Quantification Analysis Using System Model and Data	Saideep Nannapaneni, David Lechevalier, Anantha Narayanan, Sankaran Mahadevan, and Sudarsan Rachuri
2:25pm – 2:50pm	A Neural Network Meta-Model and its Application for Manufacturing.	David Lechevalier, Ronay Ak, Steven Hudak, Y.Tina Lee, and Sebti Foufou,
2:50pm - 3:15pm	Performance Assessment and Uncertainty Quantification of Predictive Models for Smart Manufacturing Systems	Luca Oneto, Ilenia Orlandi, and Davide Anguita,
3:15pm - 3:40pm	Time complexity and architecture of cloud based prognostics system for a multi-client condition monitoring activity".	Ashwin Thillai Natarajan and Sagar Kamarthi
3:40pm - 4:05pm	A Smart Component Data Model In PLM".	Yunpeng Li, Utpal Roy, Seungjun Shin, and Y.Tina Lee
4:05 pm– 4:25pm	**Coffee Break**	

	Session III **Session Chair: Dr. Rumi Ghosh**	
4:25pm – 4:50pm	Distributed Dynamic Elastic Nets: A Scalable Approach for Regularization in Dynamic Environments	Naveen Ramakrishnan and Rumi Ghosh
4:50pm – 5:15pm	Fast Detection of Material Deformation through Structural Dissimilarity	Daniela Ushizima, Talita Perciano, and Dilworth Parkinson
5:15pm – 5:40pm	Outlier Detection for Large Scale Manufacturing Processes	Abhinav Jauhri, Bradley McDanel, and Chris Connor
5:40pm – 6:00pm	**Closure**	

SPECIAL SESSION 2: Intelligent Mining

Oct 29 Thursday 2015

Session Organizer: Uraz Yavanoglu

Summary:

Recent developments in processing, storing, and sharing huge amount of data become problematic due to the lack of new approaches, techniques, methods, algorithms and technologies. Researchers try to find proper solutions based on their experiences and make contributions to current data mining and classification knowledge. This approach actually causes new problems due to the missing theoretical notions, lack of necessary disciplines and insufficient awareness on data security, information retrieval, social networking within behavioral and social issues on human nature. This special session traces the gap between big data, artificial intelligence (AI), machine learning (ML) and data mining.

Nowadays, researchers use interdisciplinary way to understand knowledge among all types of resource including data, document, tool, device, experience, process and people. This approach may help to understand biological evolution to propose robust and powerful approaches between human nature and big data processing.

Intelligent Mining term is not only related to Computer Science. This special session opens to every researcher as well as industrial partners to make contribution.

Schedule		
Time	**Title**	**Presenter/Author**
7:30am-8:00am	**Registration**	
8:00am-8:20am	**Session Opening Speech** Uraz Yavanoglu	
8:20am-8:40am	Parallel Particle Swarm Optimization (PPSO) Clustering for Learning Analytics	Kannan Govindarajan David Boulanger,etc
8:40am-9:00am	An Efficient Map-Reduce Algorithm for Computing Formal Concepts from Binary data	Raj Bhatnagar Lalit Kumar
9:00am-9:20am	High quality clustering of big data and solving empty-clustering problem with an evolutionary hybrid algorithm	Jeyhun Karimov Murat Ozbayoglu
9:20am-9:40am	QueRIE Reloaded: Using Matrix Factorization to Improve Database Query Recommendations	Magdalini Eirinaki Sweta Patel
9:40am-10:00am	Scalable Adaptive Label Propagation in Grappa	Golnoosh Farnadi Zeinab Mahdavifar,etc
10:00am-10:20am	**Coffee Break**	
10:20am-10:40am	Profiling Subscribers According to Their Internet Usage Characteristics and Behaviors	Kasim Oztoprak
10:40am-11:00am	Agile text mining with Sherlok	Renaud Richardet Jean-Cédric Chappelier,etc
11:00am-11:20am	Monitoring Adolescent Alcohol Use via Multimodal Data Analysis in Social Multimedia	Ran Pang Agustin Baretto,etc
11:20am-11:40am	Analysis and Prediction of E-Customers' Behaviour by Mining Clickstream Data	Gokhan Silahtaroglu Hale Donertasli
11:40am-12:00pm	Learning Relaxed 3-clusters from Pairs of Related Datasets	Jagadeesh Patchala Raj Bhatnagar
12:00pm-	**Discuss Further Collaborations**	

SPECIAL SESSION 3: Granular Computing and Big Data

Session Organizer: T.Y. Lin

It is our pressure to welcome you to this special session. Superficially granular computing (GrC) means "granulate and compute". It is a generalization of classical divide and conquer. To divide, one may directly granulate the Turing machine or indirectly granulate the data. A granule of data is a piece of knowledge or a "seed" of uncertainty. However, this is not what actually happened. Historically, Lotfi Zadeh suggested to us (this session organizer), the concept of granular mathematics. To limit the scope, we proposed granular computing. The central points in GrC have been knowledge approximations, uncertainty management, implicitly complexity and parallelism; Big Data shares the same issues. Such views are aware by many industrialists, even before Big Data became a buzz word. The Encyclopedia of Complexity and System Science (Springer) does include a GrC section (this chair was the editor). Plainly Big Data is a playpen for GrC; a brilliant data mining application will be illustratd.

A granule is often more than a subset. In the language of computer science, a granule is a variable that takes values in subsets of the universe U of discourse. Mathematically speaking the domain of such a variable is the granule, which is a member of P(P(U)), the power set of U. We will adopt both views; both variables and domains are granules. For examples, the topological neighborhood system of a point is a granule. Of course a granule can be a subset of U, a singleton in P(U). A class of very powerful examples is the class of simplicial complexes in algebraic topology. The geometric view provides a very efficient way to find the frequent itemsets. For a real world database from the Hospital of National Taiwan University (1257 columns and 65, 536 rows) a GrC algorithm, called simplicial complex method, is nearly 300 times faster than the FP-growth (FP-growth ran 1474.574 seconds (based on original program), while our ran only 5.142 seconds) . Many powerful granulation have been found and applied; you will see many reports in this section. Please attend this session, you will hear many interesting applications.

Finally I would like to express many thanks to conference organizers for their supports. Most importantly, we thank all the authors for their contributions.

Nov 1, 2015

13:30-13:45	Tsau-young Lin, Chinese Wall Security Policies: Information Flows in Business Cloud
13:45-14:00	Shusaku Tsumoto and Shoji Hirano, *Data Decomposition and Dual Clustering for Clinical Care Management*
14:00-14:15	I-Jen Chiang, *Agglomerative Algorithm to Discover Semantics From Unstructured Big Data*
14:15-14:30	Marcin Szczuka, Łukasz Sosnowski, and Dominik Ślęzak, *Granular Modeling with Fuzzy Comparators*
14:30-14:45	Alexander Denzler, Marcel Wehrle, and Andreas Meier, *A Granular Approach for Identifying User Knowledge*
14:45-15:00	Liang Wu, Teng-Sheng Moh, and Natalia Khuri, *Twitter Opinion Mining for Adverse Drug Reactions*
15:00-15:15	Christ Tseng, Tien Ngyuen, Chetan, Sharma. Cost and Data Exploration Considerations for Big Data Prediction on the Cloud
15:15-15:30	Maria Pershina, *Holistic Entity Matching Across Knowledge Graphs*
15:30-15:45	Patrick G. Clark and Jerzy W. Grzymala-Busse, *Mining Incomplete Data with Many Attribute-Concept Values and "Do Not Care" Conditions*
15:45-16:00	Shusaku Tsumoto and Shoji Hirano, *Granular Formalization of Medical Diagnostic Process*
16:00-16:15	Karan Khare and Teng-Sheng Moh, *Mobile Gesture-Based iPhone User Authentication*
16:15-16:30	Jun Meng, *Parallel Information Fusion Method for Microarray Data Analysis*
16:30-16:45	Yunyun Yang, Gang Xie, Yingping Liang, and T.Y. Lin, *Mining maximum matchings of directed complex networks*
16:45-17:00	Zehua Chen, *GrC-based Statistic Optimization Algorithm for Big Truth Table*
17:00-17:15	Xuezhi Ji, *A-Star Algorithm Based On-Demand Routing Protocol for Hierarchical LEO/MEO Satellite Networks*
17:15-17:30	Tsau Young Lin, *Towards New Numerical Analysis \\ Approximate Computing in Big Data I*

Tutorials

TUTORIAL 1: Optimizing Big Data Analytics on Heterogeneous Processors

Presenters: Mayank Daga, Mauricio Breternitz, Junli Gu

Summary:
Increasing computational and storage capacities have unleashed the field of big data analytics which necessitates novel architectures and tools to extract meaning from gargantuan volumes of data. The launch of AMD APUs with compliance to the Heterogeneous System Architecture (HSA) addresses some specific architectural challenges for big data analytics. In general, the community has focused on enhancing tools on discrete GPUs (dGPUs). However dGPUs bring a performance limitation due to the high overhead associated with data copies from CPU memory to GPU memory as well as constraints on dGPU memory size. HSA eliminates these performance limitations. This tutorial will explore the optimization of tools for big data analytics on the first HSA-enabled APU. We will cover the following topics:

- Programming APUs with HSA with a focus on using C++ and OpenCL programming languages
- Enhancing the programming model with a focus on accelerating Hadoop and Apache Spark
- Enhancing data operations with a focus on optimizing breadth-first search (BFS) and deep neural networks (DNN)

Short Bio.
Mayank Daga leads the efforts on performance optimization of scientific and irregular applications on heterogeneous computingsystems, particularly those that employ GPUs as an accelerator, at AMD Research. Recent research has focused on the algorithmic exploration of high-performance, energy-efficient graph traversals, database searches and sparse linear algebra on AMD accelerators. He also investigates the optimal mapping of deep neural network (DNN) frameworks and applications on CPU+GPU systems. He has developed the fastest single-GPU implementation of Graph500, a popular big data benchmark in high performance computing (HPC). Previously, he was adjudicated as the Outstanding Graduate Student for his work on parallel programming on heterogeneous systems.

Mauricio Breternitz, Ph.D. has research interests and activities spanning the areas of cloud workloads and cluster systems, application of systems research to large data sets and analytics, compilers, compilation techniques and (micro)architecture, (processing) system architecture at the node and cluster level, power-efficient computing, heterogeneous processing systems (CPUs, GPUs). Recent projects include work on understanding and optimizing cloud applications and cloud workload analysis (MapReduce, Hadoop, GraphLab), focusing on system-level characterization and innovation for power efficient computing and dense servers. He has guest lectured for courses on Dynamic Optimization and Code Generation at Carnegie Melon University and University of Texas, Austin. In collaboration with university researchers he investigated efficient acceleration for Hadoop Map-Reduce applications on GPU-accelerated systems. He also investigates efficient scale-out implementation of Deep Neural Network algorithms on CPU+GPU accelerated systems and developed novel algorithms for CPU/GPU code migration.

Junli Gu, Ph.D. has research interests in Heterogeneous Systems and Machine Learning applications. When DNN started to emerge two years ago, she started a small team to implement DNN frameworks in OpenCL and studied how to build heterogeneous systems for DNN + Big Data applications. Optimizations on GPUs and APUs have been a research focus. She also worked to design the Heterogeneous System Coherence for future heterogeneous processor and propose new power model for multi-core CPUs. Her earlier research experience includes a high performance computer architecture design project and multi-media DSP development.

TUTORIAL 2: The Era of Big Spatial Data

Presenters: Mohamed F. Mokbel and Ahmed Eldawy

Summary:
In this tutorial, we present the recent work in the big data community for handling Big Spatial Data. This topic became very hot due to the recent explosion in the amount of spatial data generated by smart phones, satellites and medical devices, among others. This tutorial goes beyond the use of existing systems as-is (e.g., Hadoop, Spark or Impala), and digs deep into the core components of big systems (e.g., indexing and query processing) to describe how they are designed to handle big spatial data. During this 90-minute tutorial, we review the state-of-the-art work in the area of Big Spatial Data while classifying the existing research efforts according to the implementation approach, underlying architecture, and system components. In addition, we provide case studies of full-fledged systems and applications that handle Big Spatial Data which allows the audience to better comprehend the whole tutorial.

Short Bio.
Mohamed Mokbel has delivered eight successful tutorials as one about big spatial data in IEEE MDM 2015, two about mobility and social networks in VLDB 2013 and MDM 2013, two about location-based query processing in EDBT 2006 and MDM 2007,

and three about location privacy in MDM 2007, ACM CCS 2007, and IEEE ICDM 2009. Also, Mohamed Mokbel has delivered seven workshop keynotes as three about Big Spatial Data in BigSpatial 2014, Korean GIS 2014, and IQmulus 2014, one about spatio-temporal data uncertainty at QUeST 2009, one about location privacy at SPRINGL 2009, and two about location-based social network at MobiDE 2011 and LBSN 2013.

Mohamed Mokbel and **Ahmed Eldawy** have recently presented a 90-minute tutorial about the same topic in IEEE Mobile Data Management conference (MDM) 2015, which focused more on the mobility and applications aspects of spatial data, without delving much in the system internals. In contrast, this tutorial is directed to the Big Data community as it puts more focus on query processing over different systems architectures, and describes how to adapt existing Big Data systems to support spatial data.

TUTORIAL 3: Platforms and Algorithms for Big Data Analytics

Presenters: Chandan K. Reddy

Summary:
This tutorial consists of two parts: (i) Big data platforms and their characteristics (ii) Large-scale classification and clustering algorithms.
The first part will provide an in-depth analysis of different platforms available for studying and performing big data analytics. It will survey different hardware platforms available for big data analytics and assesses the advantages and drawbacks of each of these platforms based on various metrics such as scalability, data I/O rate, fault tolerance, real-time processing, data size supported and iterative task support. In addition to the hardware, a detailed description of the software frameworks used within each of these platforms is also discussed along with their strengths and drawbacks. Some of the critical characteristics that will be discussed here can potentially aid the audience in making an informed decision depending on their computational needs. Using ratings table, a rigorous qualitative comparison between different platforms is also discussed. The second part of the tutorial will consist of big data classification and clustering algorithms. In order to provide more insights into the effectiveness of each of the platform in the context of big data analytics, specific implementation level details of the widely used k-nearest neighbor and the k-means clustering algorithm on various platforms will also be described in the form of pseudocode. In addition, recent advances in large-scale linear classification and map-reduce based classification algorithms will be discussed. In the context of clustering, some of the well-known one-pass clustering algorithms and other parallel and distributed clustering solutions will be briefly mentioned.

Short Bio.
Chandan Reddy is an Associate Professor in the Department of Computer Science at Wayne State University. He received his Ph.D. from Cornell University and M.S. from Michigan State University. His primary research interests are Data Mining and Machine Learning with applications to Healthcare Analytics, Bioinformatics and Social Network Analysis. His research is funded by the National Science Foundation, the National Institutes of Health, the Department of Transportation, and the Susan G. Komen for the Cure Foundation. He has published over 50 peer-reviewed articles in leading conferences and journals including TPAMI, TKDE, SIGKDD, ICDM, SDM, and CIKM. He received the Best Application Paper Award in ACM SIGKDD conference in 2010, and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is a senior member of the IEEE and a life member of the ACM. More details about his current work are available at his website http://www.cs.wayne.edu/~reddy/

TUTORIAL 4: Optimal Connectivity on Big Graphs: Measures, Algorithms and Applications

Presenters: Hanghang Tong

Summary:
Graph mining has been playing a pivotal role in many disciplines, ranging from computer science, sociology, civil engineering, physics, economics/marketing, to biology, life science, management science, political science, etc. Among others,a common and fundamental property of the graphs arising from these domains is connectivity. The goal of this tutorial is to (1) provide a concise review of the recent advances in optimizing graph connectivity and its applications; and (2) identify the open challenges and future trends. We believe this is an emerging, high-impact topic in graph mining, which will attract both researchers and practitioners in the big data research community. Our emphasis will be on (1) the recent emerging techniques on addressing graph connectivity optimization problem, especially in the context of big data; and (2) the open challenges/future trends, with a careful balance between the theories, algorithms and applications.

Short Bio.
He is currently an assistant professor at School of Computing, Informatics and Decision Systems Engineering at Arizona State University. Before that, he was an assistant pro- fessor at Computer Science Department, City College, City University of New York, a research staff member at IBM T.J. Watson Research Center and a Post-doctoral fellow in Carnegie Mellon University. He received his M.Sc and Ph.D. degrees from Carnegie Mellon University in 2008 and 2009, both majored in Machine Learning. His research interest is in large scale data mining for graphs and multimedia. He has received several awards, including best paper award in CIKM 2012, best paper award in SDM 2008 and best research paper award in ICDM 2006. He has published over 100 referred articles and more than 20 patents (filed + pending). He has served as a program committee member in top data mining,

databases and artificial intelligence venues (e.g., BigData, SIGKDD, ICDM, SIGMOD, AAAI, WWW, etc). For more details, please refer to his homepage at tonghanghang.org

TUTORIAL 5: The World is Big and Linked: Whole Spectrum Industry Solutions towards Big Graphs

Presenters: Toyotaro Suzumura, Ching-Yung Lin, Yinglong Xia, Lifeng Nai

Summary:
The importance of graph needs no emphasis, since many big data applications consist of entities with internal links, naturally forming a graph. However, graph computing and storage is notorious for its low efficiency, resulting in performance barriers in reality, especially when the graph volume becomes huge. Albeit many efforts have been made to improve the efficiency, it remains lack of systematic study on the characteristics of big graph computing and storage on commodity platforms. For example, Spark GraphX and Apache Giraph address merely the computing framework, sort of skipping the integration with a full functional graph storage layer; while existing graphDBs, e.g. Titan, takes a naive way to organize graph data, with little consideration on computational behaviors, not to mention visualization and scaling out/up techniques.
In this tutorial, we provide full perspective investigations into big graph technology, addressing the overall architecture of industry solutions for big graphs. The tutorial consists of three mutually related parts, all under the umbrella of IBM System G, a real industrial high performance solution towards big graphs. The three parts are: 1) An overview of Graphen, a sister of IBM System G and possibly to be in Apache Incubator, based on open source technologies and compatible with de facto graph standards. 2) An exploration to ScaleGraph, an open source package built with IBM's PGAS and helps win the 1st in the latest Graph500, aiming at scaling out the graph computation and addressing both software performance and productivity. 3) An introduction to GraphBIG, an open source suite with an unified in-memory graph representation and graph analytics on multicore CPUs/GPUs, aiming at scaling up graph computing. We will present industry solutions and experimental results for shedding lights onto the essential behaviors of the big graph processing.

Short Bio.
Dr. Toyotaro Suzumura is currently a research staff member at IBM T.J. Watson Research Center, the headquarter of IBM Research and working on high performance graph analytics platform. He holds his Ph.D. in Computer Science from Tokyo Institute of Technology. His Ph.D. thesis focused on Grid computing and large-scale distributed computing systems. His currentresearch topics include big data processing middlewares, large-scale graph analytics, supercomputing / high performance computing, and microscopic simulation platforms. Tutorial Proposal, IEEE Big Data 2015 He joined IBM Research – Tokyo in 2004 as a research scientist and has also continued to serve professorship in academia since 2009. Since April 2009, he had served as a visiting associate professor at Tokyo Institute of Technology in Japan and led his own research laboratory named Suzumura laboratory. Since October 2013, he has moved to IBM Research – Ireland and its Smarter Tenology Cities Center to lead his traffic simulation project in Dublin, Ireland. He also serves as a visiting associate professor at University College Dublin as well as an adjunct professor at Tokyo Institute of Technology in Japan. He is now a co-principal investigator for two Japanese government projects funded by Japan Science Foundation that aim at building next-generation big data processing middlewares and libraries. One of his notable achievements is that he and his team won the first place at the world competition of "Big Data" processing on supercomputers called "Graph500" in June 2014 by proposing a novel and highly scalable graph search algorithm on massively parallel and distributed computers and successfully implementing it on the Japanese K supercomputer. He has also led a project "ScaleGraph" of developing an open source library for billion-scale graph analytics. He has published 50 reviewed paper for top-quality international conferences and workshops, 42 domestic conferences, and 12 international patents, and also served invited speakers, workshop chairs and/or program committee members for top-tier conferences such as ACM/ IEEE Supercomputing and ACM International World Wide Web Conference.

Dr. Ching-Yung Lin is the Manager of the Network Science and Big Data Analytics Department in IBM T. J. Watson Research Center. He is also an Adjunct Assistant/Associate/Full Professor in Columbia University since 2005, in NYU since 2014, and in University of Washington 2003-2009. He received his Ph.D. from Columbia Univ in 2000, M.S. and B.S. from National Taiwan University in 1993 and 1991, respectively, all in Electrical Engineering. His interest is mainly on fundamental research of multimodality signal understanding, network science, and computational social & cognitive science. Since 2011, Lin has been leading a team of more than 40 Ph.D. researchers in worldwide IBM Research Labs and more than 20 professors and researchers in 10 universities, including Columbia, Northeastern, Northwestern, CMU, U Minnesota, Rutgers, U New Mexico, UC-Berkeley, Stanford Research Institute, and USC. He is an author of 160+ publications and 20+ issued patents. His team recently won the Best Paper Award in IEEE BigData 2013, Best Paper Award in ACM CIKM 2012, and Best Theme Paper Award in ICIS 2011. In 2011,

he was the first IEEE Fellow elevated for contributions to Network Science. He is also an IEEE CASS Distinguished Lecturer 2015-2016.

Dr. Yinglong Xia is currently a research staff member and tech lead in the IBM T.J. Watson Research Center, working on Big Data technology for graph analytics and storage from the perspective of High Performance Computing. He also works on parallel probabilistic inference on graphical models. Before joined IBM, Yinglong received his PhD in Computer Science from the University of Southern California (USC) in 2010. Prior to that, he received his MS and BS from Tsinghua University and the University of Electronic Science and Technology (UESTC) in 2006 and 2003, respectively. Yinglong received the IBM Research Division Awards in 2013 and 2015. He is selected for the IBM Emerging Talents Program for Research and is the recipient of the IBM Corporate Service Corp (CSC) in 2015. He is a director of the Linked Data Benchmark Council (LDBC) since 2014 and he is a founder member on the IEEE Big Data Initiative Standardization Committee and co-chairs the work group on Big Data metadata and management. He was a CRA/NSF Computing Innovation Fellow (CIFellow) in 2010~2012. He has published 48 papers in international journals, conferences and workshops, such as SC, IPDPS, IEEE Trans. TPDS, JPDC, etc. He serves as organizers in numerous academic events, such as the program chair for IEEE CBDCom'15, Tutorial Chair for IEEE Big Data'15, Publicity Chair for IPDPS'15, Track Chair for CSE'15, etc. He also serves on the editorial board of the Journal of Big Data published by Springer. He serves as a Guest Editor on the Journal of Tsinghua Science and Technology, Special Issue on Cloud and Big Data Computing (2016), and a Guest Editor on the IEEE Internet of Things Journal Special Issue on Big Data Analytics and Management in Internet of Things (2014).

Mr. Lifeng Nai is a PhD student and a Technology Enthusiast at the Georgia Institute of Technology, working on graph technology and GPU from the perspective of computer architectures. He has been working with IBM Research as a Joint Study Agreement for about three years and made solid contribution to IBM System G. He received his MS in ECE from Georgia Institute of Technology in 2011 and MS in EE from Shanghai Jiao Tong University in 2009, where he also received his BS in EE in 2008. He was working withAlpscale Microelectronic Corp. and SJTU-Altera Student Innovation Lab. He has published extensively and submitted five patent disclosures.

TUTORIAL 6: Tutorial on Predictive Maintenance

Presenters: Zhuang Wang

Summary:
Predictive maintenance strives to anticipate equipment failures to allow for advance scheduling of corrective maintenance thereby preventing unexpected equipment downtime and improving service quality for the customers. There is a tremendous interest in industry to leverage recent advances in machine learning and data mining to tackle this problem. Whereas the key enabling techniques (such as failure diagnostics and prediction) for predictive maintenance have been of considerable emphasis in the community, the design of practical predictive maintenance systems has not enjoyed the same attention. This is partially because the lack of access to real-world use cases becomes an obstacle for researchers to consider all the characteristics of data and the nature of the problem for the practical design. In this tutorial, we aim to fill the gap between the business needs and technology offerings by a detailed study on the nature and requirements of the real-world predictive maintenance problems as well as a comprehensive survey of the techniques tacking the problems. We will survey the underlying data sources and feature engineering techniques, the learning scenarios and model creation and selection techniques, and will also present several real-world case studies and lessons learned.

Short Bio.
Dr. Zhuang Wang is a Member of Technical Staff at Skytree, a California-based machine learning startup, where he works on the R&D of enterprise-grained machine learning platform. He was an Application Architect in Big Data and Analytics at IBM Global Business Services, where he was dedicated in bridging science and business by developing big data analytics solutions for business innovation in various industries. Prior to IBM, he was a Research Scientist with Siemens Corporate Research and led/worked on a wide variety of projects building predictive maintenance, anomaly detection and decision support systems for servicing fleets of industrial and medical equipments. He earned his Ph.D. in CS at Temple Univ., PA in 2010 and his B.A. in Electronic Commerce at Wuhan Univ., China in 2006. Dr. Wang's research interests are in large-scale machine learning and its applications. He is the author/coauthor of 20+ papers published at JMLR, ICML, KDD, AISTATS et al. He is the project lead of BudgetedSVM, a highly optimized toolbox for SVM approximations when data cannot fit into memory (with 2000+ downloads at sourceforge.net), and the designer of the world first log-based predictive maintenance system.

Posters

Paper ID	Accepted Poster
P202	Ranjeet Devarakonda, *Preparing, Storing and Distributing Multi-dimensional Scientific Data*
P203	Ranjeet Devarakonda, *Metadata Documentation Tool for Large Volumes of Data*
P209	Erica Yang, *Data Optimised Computing for Heterogeneous Big Data Computing Applications*
P210	Vasilis Efthymiou and Kostas Stefanidis, *Top-k Computations in MapReduce: A Case Study on Recommendations*
P211	Kai Chen, Yi Zhou, and Fangyan Dai, *A LSTM-based method for stock returns prediction: A case study of China stock market*
P212	Kazuya Uesato, Hiroki Asai, and Hayato Yamana, *Predicting Various Types of User Attributes in Twitter by Using Personalized PageRank*
P215	Asmelash Teka Hadgu, Aastha Nigam, and Ernesto Diaz-Aviles, *Large-Scale Learning with AdaGrad on Spark*
P216	Sudip Mittal, Karuna Joshi, Claudia Pearce, and Anupam Joshi, *Parallelizing Natural Language Techniques for Knowledge Extraction from Cloud Service Level Agreements.*
P217	Christian Beecks, Merih Seran Uysal, and Thomas Seidl, *Gradient-based Signatures for Big Multimedia Data*
P218	Dimitrios Rafailidis and Stefanos Antaris, *Indexing Media Storms on Flink*
P219	Anoop Kumar, Connor Stokes, Frederick Choi, and Ralph Weischedel, *Scaling NLP Algorithms to Meet High Demand*
P220	Bonnie Dorr, Craig Greenberg, Peter Fontana, Mark Przybocki, Marion Le Bras, Cathryn Ploehn, Oleg Aulov, and Wo Chang, *The NIST Data Science Evaluation Series: Part of the NIST Information Access Division Data Science Initiative*
P221	Alexei Samoylov and Jason Schlachter, *Flexible Ingest Framework - A Spark-based Extensible Architecture for Data Ingest*
P223	Priya Govindan, Ruobing Chen, Katya Scheinberg, and Soundararajan Srinivasan, *A Scalable Solution For Group Feature Selection*
P224	Luna Zhang, *Genetic Deep Neural Networks Using Different Activation Functions for Financial Data Mining*
P227	Ryota Takei and Ayahiko Niimi, *Performance of Graph Reconstruction Method for Large-Scale Web Graph Analysis*
P228	Altti Maarala, Mika Rautiainen, Miikka Salmi, and Jukka Riekki, *Low latency analytics for streaming traffic data with Apache Spark*
P229	Divya Rao and Wee Keong Ng, *How to make money from your information and keep your privacy*
P235	Kezia Rani B and Vinaya Babu A, *Scheduling of Big data Application Workflows in Cloud and Inter-cloud environments.*
P236	Peter Li, Simon Yates, Jenna Lovely, and David Larson, *Patient-Like-Mine – A Real Time, Visual Analytics Tool for Clinical Decision Support*
P237	Samuel Johnson and Kang-Yu Ni, *A Pricing Mechanism Using Social Media and Web Data to Infer Dynamic Consumer Valuations*
P238	Yifan Hao, Huiping Cao, and Yan Qi, *Efficient Keyword Search on Graphs using MapReduce*
P239	Yuqing ZHU, *Non-Blocking One-Phase Commit Made Possible for Distributed Transactions over Replicated Data*
P241	Daisaku Yokoyama and Masashi Toyoda, *A Large Scale Examination of Vehicle Recorder Data to Understand Relationship between Drivers' Behaviors and Their Past Driving Histories*
P242	Yoshitaka Yamamoto and Koji Iwanuma, *Online Pattern Mining for High-Dimensional Data Streams*
P243	Zhenhui Liu, Jingjing He, and Zhihui Du, *Modeling the Learning Behaviours of Massive Open Online Courses*
P245	Jian yin, *Data Confidentiality Challenges in Big Data Applications*

P246	Anh-Phuong TA, *Factorization Machines with Follow-The-Regularized-Leader for CTR prediction in Display Advertising*
P247	Aakash Deep Singh, Wei Wu, and Shonali Krishnaswamy, *Taxi Trip Time Prediction Using Similar Trips and Road Network Data*
P248	Long Ma and Yanqing Zhang, *Using Word2Vec to Process Big Text Data*
P249	Longbiao Chen and Jérémie Jakubowicz, *Inferring Bike Trip Patterns from Bike Sharing System Open Data*
P251	Xiaobing Zhou, Tonglin Li, Ke Wang, Dongfang Zhao, Iman Sadooghi, and Ioan Raicu, *MHT: A Light-weight Scalable Zero-hop MPI Enabled Distributed Key-Value Store*
P253	Hangu Yeo and Catherine Crawford, *Big Data: Cloud Computing in Genomics Applications*
P255	Maoyuan zhang, Fang yuan, and Jianping zhu, *Integrating Semantic Knowledge into Tag-LDA Model through Cloud Model*
P256	Yunkai Liu and Christopher Magno, *A Case Study to Apply Mobile Technology into Individual's Local Community*
P257	Kajanan Sangaralingam, Vivek Kumar Singh, Biying Tan, Chandra Sekhar Saripaka, and Giuseppe Manai, *Clairvoyant-Push: a Real Time News Personalised Push Notifier which uses Topic Modeling and Social Scoring to drive readers engagement*
P258	Hua Fang, Honggang Wang, Chonggang Wang, and Mahmoud Daneshmand, *Using Probabilistic Approach to Joint Clustering and Statistical Inference: Big Investment Data*
P260	Jing Wang, Nikos Ntarmos, and Peter Triantafillou, *Towards a Subgraph/Supergraph Query Index*
P261	Ratna Madhuri Maddipatla, Dr.Mirsad Hadzikadic, Dr.Dipti Patel Misra, and Dr.Lixia Yao, *30 Day Hospital Readmission Analysis*
P262	Mehrdad Yazdani and Larry Smarr, *Using Pairwise Difference Features to Measure Temporal Changes in the Microbial Ecology*
P263	Ardi Imawan and Joonho Kwon, *A Timeline Visualization System for Road Traffic Big Data*
P265	Gaël Chareyron and Berengère Branchet, *A new area tourist ranking method*
P266	Maoyuan zhang, Jianping zhu, Lijun hua, and Fang yuan, *Text Retrieval based on the Feature Conversion of Vector Space*
P267	Kang Li, Vinay Deolalikar, and Neeraj Pradhan, *Big Data Gathering and Mining Pipelines for CRM using Open-source*
P268	Jay Gholap, Vandana Janeja, and Yelena Yesha, *Unified Framework For Clinical Data Analytics (U-CDA)*
P269	Chanpaul Jin Wang, Hua Fang, Chonggang Wang, Mahmoud Daneshmand, and Honggang Wang, *A Novel Initialization Method for Particle Swarm Optimization-based FCM in Big Biomedical Data*
P270	Chandra Khatri, Suman Voleti, Sathish Veeraraghavan, Nish Parikh, Atiq Islam, Shifa Mahmood, Neeraj Garg, and Vivek Singh, *Algorithmic Content Generation for eBay Products*
P271	Ismini Lourentzou, Graham Dyer, Abhishek Sharma, and ChengXiang Zhai, *Hotspots of News Articles: Joint Mining of News Text \& Social Media to Discover Controversial Points in News*
P273	Khalifeh AlJadda, Mohammed Korayem, and Trey Grainger, *Improving the Quality of Semantic Relationships Extracted from Massive User Behavioral Data*
P274	Maruthi Prithivirajan, Vivian Lai, Kyong Jin Shim, and Ping Shung Koo, *Analysis of Star Ratings in Consumer Reviews: A Case Study of Yelp*

Conference WiFi Instruction

Instructions on connecting to Wireless Internet in meeting space

here's the log in and applies ONLY TO THE GRAND BALLROOM and not the entire hotel

Network Name (SSID): IEEE 2015 BIG DATA
PW: ieee2015
The password is CASE SENSITIVE:

IEEE BIGDATA 2016

November, 2016, Washington DC, USA

The IEEE Big Data conference series is a leading forum for disseminating the latest advances in big data research , development and application. We solicit high-quality original research papers (including significant work-in-progress) in any aspect of Big Data with emphasis on 5Vs (Volume, Velocity, Variety, Value and Veracity): big data science and foundations, big data infrastructure, big data management, big data searching and mining, big data privacy/security, and big data applications. Relevant topics include but are not limited to:

1. Big Data Science and Foundations

a. Novel Theoretical Models for Big Data
b. New Computational Models for Big Data
c. Data and Information Quality for Big Data
d. New Data Standards

2. Big Data Infrastructure

a. Cloud/Grid/Stream Computing for Big Data
b. High Performance/Parallel Computing Platforms for Big Data
c. Autonomic Computing and Cyber-infrastructure, System Architectures, Design and Deployment
d. Energy-efficient Computing for Big Data
e. Programming Models and Environments for Cluster, Cloud, and Grid Computing to Support Big Data
f. Software Techniques and Architectures in Cloud/Grid/Stream Computing
g. Big Data Open Platforms
h. New Programming Models for Big Data beyond Hadoop/MapReduce, STORM
i. Software Systems to Support Big Data Computing

3. Big Data Management

a. Advanced database and Web Applications
b. Novel Data Model and Databases for Emerging Hardware
c. Data Preservation
d. Data Provenance
e. Interfaces to Database Systems and Analytics Software Systems
f. Data Protection, Integrity and Privacy Standards and Policies
g. Information Integration and Heterogeneous and Multi-structured Data Integration
h. Data management for Mobile and Pervasive Computing
i. Data Management in the Social Web
j. Crowdsourcing
k. Spatiotemporal and Stream Data Management
l. Scientific Data Management
m. Workflow Optimization
n. Database Management Challenges: Architecture, Storage, User Interfaces

4. Big Data Search and Mining

a. Social Web Search and Mining
b. Web Search
c. Algorithms and Systems for Big Data Search
d. Distributed, and Peer-to-peer Search
e. Big Data Search Architectures, Scalability and Efficiency
f. Data Acquisition, Integration, Cleaning, and Best Practice
g. Visualization Analytics for Big Data
h. Computational Modeling and Data Integration
i. Large-scale Recommendation Systems and Social Media Systems
j. Cloud/Grid/Stream Data Mining- Big Velocity Data
k. Link and Graph Mining
l. Semantic-based Data Mining and Data Pre-processing
m. Mobility and Big Data
n. Multimedia and Multi-structured Data- Big Variety Data

5. Big Data Security & Privacy

a. Intrusion Detection for Gigabit Networks
b. Anomaly and APT Detection in Very Large Scale Systems
c. High Performance Cryptography
d. Visualizing Large Scale Security Data
e. Threat Detection using Big Data Analytics
f. Privacy Threats of Big Data
g. Privacy Preserving Big Data Collection/Analytics
h. HCI Challenges for Big Data Security & Privacy
i. User Studies for any of the above
j. Sociological Aspects of Big Data Privacy

6. Big Data Applications

a. Complex Big Data Applications in Science, Engineering, Medicine, Healthcare,
Finance, Business, Law, Education, Transportation, Retailing, Telecommunication
b. Big Data Analytics in Small Business Enterprises (SMEs),
c. Big Data Analytics in Government, Public Sector and Society in General
d. Real-life Case Studies of Value Creation through Big Data Analytics
e. Big Data as a Service
f. Big Data Industry Standards
g. Experiences with Big Data Project Deployments

INDUSTRIAL and GOVERNMENT Track

The Industrial and government Track solicits papers describing implementations of Big Data solutions relevant to industrial settings. The focus of industry track is on papers that address the practical, applied, or pragmatic or new research challenge issues related to theuse of Big Data in industry