## Preface

### 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications

http://bigdata-mining.org/

The aim of this workshop is to bring together people from both academia and industry to present their most recent work related to big-data issues, and exchange ideas and thoughts in order to advance this big-data challenge, which has been considered as one of the most exciting opportunities in the past 10 years.

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter, Weibo, etc., that allow users to create contents freely and amplify the already huge Web volume. Furthermore, with mobile phones becoming the sensory gateway to get real-time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR (call data record)-based processing for billing purposes only. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, to buses, railway stations and airports) are all loosely connected. Trillions of such connected components will generate a huge data ocean, and valuable information must be discovered from the data to help improve quality of life and make our world a better place. For example, after we get up every morning, in order to optimize our commute time to work and complete the optimization before we arrive at office, the system needs to process information from traffic, weather, construction, police activities to our calendar schedules, and perform deep optimization under the tight time constraints. In all these applications, we are facing significant challenges in leveraging the vast amount of data, including challenges in (1) system capabilities (2) algorithmic design (3) business models.

August 2014

*Wei Fan, Albert Bifet, Qiang Yang and Philip Yu*
BigMine 2014 Program co-Chairs
http://bigdata-mining.org/

## BigMine 2014 Workshop Organization

### Workshop Chairs

Wei Fan
Huawei Noah's Ark Lab
E-mail: wei.fan at gmail.com

Albert Bifet
Huawei Noah's Ark Lab
E-mail: abifet at cs.waikato.ac.nz

Qiang Yang
Huawei Noah's Ark Lab,
E-mail: qiang.yang at huawei.com

Philip Yu
University of Illinois at Chicago
E-mail: psyu at cs.uic.edu

### Organizers

- Albert Bifet, Huawei Noah's Ark Lab
- Wei Fan, Huawei Noah's Ark Lab
- Jing Gao, University at Buffalo
- Le Gruenwald, University of Oklahoma
- Dimitrios Gunopulos, University of Athens
- Geoff Holmes, University of Waikato
- Latifur Khan, University of Texas at Dallas
- Dekang Lin, Google
- Deepak Turaga, IBM T.J. Watson Research
- Qiang Yang, Huawei Noah's Ark Lab
- Philip Yu, University of Illinois at Chicago
- Kun Zhang, Xavier University of Louisiana
- Xiatian Zhang, Tencent
- Yuanchun Zhou, Chinese Academy of Sciences

### Publicity Chairs

- Albert Bifet, Huawei Noah's Ark Lab
- Erheng Zhong, Hong University of Science of Technology

### Treasury

- Xiaoxiao Shi, University of Illinois at Chicago
- Jing Gao, SUNY Buffalo

## Program Committee

- Vassilis Athitsos, University of Texas at Arlington
- Roberto Bayardo, Google
- Francesco Bonchi, Yahoo! Research Barcelona
- Liangliang Cao, IBM
- Hong Cheng, The Chinese University of Hong Kong
- Alfredo Cuzzocrea, ICAR-CNR & University of Calabria
- Ian Davidson, SUNY
- Gianmarco De Francisci Morales, Yahoo Labs Barcelona
- Nan Du, Georgia Institute of Technology
- Joao Gama, University Porto
- Ricard Gavaldà, Universitat Politècnica de Catalunya
- Fosca Giannotti, ISTI-CNR
- Bart Goethals, University of Antwerp
- Jiawei Han, University of Illinois at Urbana-Champaign
- Marwan Hassani, Aachen University
- Steven C.H. Hoi, Nanyang Technological University
- Dino Ienco, UMR TETIS, Irstea, Montpellier
- Siddhartha Jonnalagadda, Mayo Clinic
- Murat Kantarcioglu, University of Texas at Dallas
- George Karypis, University of Minnesota

- Steve Ko, SUNY at Buffalo
- Vipin Kumar, University of Minnesota, Twin Cities
- Jianhui Li, Computer Network Information Center,Chinese Academy of Sciences
- Cindy Xide Lin, University of Illinois at Urbana-Champaign
- Shou-De Lin, National Taiwan University
- Michael May, Fraunhofer IAIS
- Themis Palpanas, University of Trento
- Fernando Perez-Cruz, University Carlos III
- Bernhard Pfahringer, University of Waikato
- Jesse Read, Universidad Carlos III
- Chandan K. Reddy, Wayne State University
- Cyrus Shahabi, USC
- Ashok Srivastava, NASA
- Frederic Stahl, University of Reading
- Jian-Tao Sun, Microsoft Research Asia
- Jie Tang, Tsinghua University
- Hanghang Tong, Carnegie Mellon University
- Joaquin Vanschoren, Eindhoven University of Technology
- Haifeng Wang, Baidu
- Bo Wang, Nanjing University of Aeronautics & Astronautics
- Yi Wang, Tencent
- Xian Wu, Microsoft
- Tian Wu, Baidu
- Zhenghua Xue, Chinese Academy of Sciences
- Gui-Rong Xue, Shanghai Jiao Tong University
- Xifeng Yan, University of California at Santa Barbara
- Rong Yan, Facebook
- Aden Yuen, Tencent
- Demetris Zeinalipour, University of Cyprus
- Xingquan Zhu, University of Technology, Sydney

## List of Subreviewers

- Alexandre Faria de Carvalho
- Santosh Kabbur
- Shuji Hao
- Asmaa Elbadrawy
- Elaine Faria
- Xingyu Gao

## Invited Keynote Speakers

### Jimmy Lin, University of Maryland and Twitter

### Title: Scaling Big Data Mining Infrastructure: The Twitter Experience

The analytics platform at Twitter has experienced tremendous growth over the past few years in terms of size, complexity, number of users,and variety of use cases. In this talk, I'll discuss the evolution of Twitter's infrastructure and the development of capabilities for data mining on "big data". One important lesson is that successful big data mining in practice is about much more than what most academics would consider data mining: life "in the trenches" is occupied by much preparatory work that precedes the application of data mining algorithms and followed by substantial effort to turn preliminary models into robust solutions. In this context, I'll discuss two topics: First, schemas play an important role in helping data scientists understand petabyte-scale data stores, but they're insufficient to provide an overall "big picture" of the data available to generate insights. Second, we observe that a major challenge in building data analytics platforms stems from the heterogeneity of the various components that must be integrated together into production workflows—we refer to this as "plumbing".

This talk has two goals: For practitioners, I hope to share our experiences to flatten bumps in the road for those who come after us. For academic researchers, I hope to provide a broader context for data mining in production environments, pointing out opportunities for future work.

**Bio:** *Jimmy Lin is an Associate Professor and the Associate Dean of Research in the College of Information Studies (The iSchool) at the University of Maryland, with a joint appointment in the Institute for Advanced Computer Studies (UMIACS) and an affiliate appointment in the Department of Computer Science. He graduated with a Ph.D. in Electrical Engineering and Computer Science from MIT in 2004. Lin's research lies at the intersection of information retrieval and natural language processing; his current work focuses on large-scale distributed algorithms and infrastructure for data analytics. From 2010-2012, Lin spent an extended sabbatical at Twitter, where he worked on services designed to surface relevant content to users and analytics infrastructure to support data science.*

**Katharina Morik, TU Dortmund University**

**Title: Big Data and Small Devices**

How can we learn from the data of small ubiquitous systems? Do we need to send the data to a server or cloud and do all learning there? Or can we learn on some small devices directly? Are smartphones small? Are navigation systems small? How complex is learning allowed to be in times of big data? What about graphical models? Can they be applied on small devices or even learned on restricted processors?

Big data are produced by various sources. Most often, they are distributedly stored at computing farms or clouds. Analytics on the Hadoop Distributed File System (HDFS) then follows the MapReduce programming model. According to the Lambda architecture of Nathan Marz and James Warren, this is the batch layer. It is complemented by the speed layer, which aggregates and integrates incoming data streams in real time. When considering big data and small devices, obviously, we imagine the small devices being hosts of the speed layer, only. Analytics on the small devices is restricted by memory and computation resources.

The interplay of streaming and batch analytics offers a multitude of configurations. In this talk, we discuss opportunities for using sophisticated models for learning spatio-temporal models. In particular, we investigate graphical models, which generate the probabilities for connected (sensor) nodes. First, we present spatio-temporal random fields that take as input data from small devices, are computed at a server, and send results to -possibly different — small devices. Second, we go even further: the Integer Markov Random Field approximates the likelihood estimates such that it can be computed on small devices. We illustrate our learning models by applications from traffic management.

**Bio:** *Katharina Morik is full professor for computer science at the TU Dortmund University, Germany. She earned her Ph.D. (1981) at the University of Hamburg and her habilitation (1988) at the TU Berlin. Starting with natural language processing, her interest moved to machine learning ranging from inductive logic programming to statistical learning, then to the analysis of very large data collections, high-dimensional data, and resource awareness. Her aim to share scientific results strongly supports open source developments. For instance, RapidMiner started out at her lab, which continues to contribute to it. Since 2011 she is leading the collaborative research center SFB876 on resource-aware data analysis, an interdisciplinary center comprising 12 projects, 19 professors, and about 50 Ph D students or Postdocs.*

*She was one of those starting the IEEE International Conference on Data Mining together with Xindong Wu, and was chairing the program of this conference in 2004. She was the program chair of the European Conference on Machine Learning (ECML) in 1989 and one of the program chairs of ECML PKDD 2008. She is in the editorial boards of the international journals "Knowledge and Information Systems" and "Data Mining and Knowledge Discovery".*

**Jieping Ye, Arizona State University**

**Title: Exact Data Reduction for Big Data**

Recent technological innovations have enabled data collection of unprecedented size and complexity. Examples include web text data, social media data, gene expression images, neuroimages, and genome-wide association study (GWAS) data. Such data have incredible potential to address complex scientific and societal questions, however analysis of these data poses major challenges for the scientists. As an emerging and powerful tool for analyzing massive collections of data, data reduction in terms of the number of variables and/or the number of samples has attracted tremendous attentions in the past few years, and has achieved great success in a broad range of applications. The intuition of data reduction is based on the observation that many real-world data with complex structures and billions of variables and/or samples can usually be well explained by a few most relevant explanatory features and/or samples. Most existing methods for data reduction are based on sampling or random projection, and the final model based on the reduced data is an approximation of the true (original) model. In this talk, I will present fundamentally different approaches for data reduction in that there is no approximation in the model, that is, the final model constructed from the reduced data is identical to the original model constructed from the complete data. Finally, I will use several real world examples to demonstrate the potential of exact data reduction for analyzing big data.

**Bio:** *Jieping Ye is an Associate Professor of Computer Science and Engineering at the Arizona State University. He is a core faculty member of the Bio-design Institute at ASU. He received his Ph.D. degree in Computer Science from University of Minnesota, Twin Cities in 2005. His research interests include machine learning, data mining, and biomedical informatics. He has served as Senior Program Committee/Area Chair/Program Committee Vice Chair of many conferences including NIPS, KDD, IJCAI, ICDM, SDM, ACML, and PAKDD. He serves as an Associate Editor of Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, and IEEE Transactions on Pattern Analysis and Machine Intelligence. He won the NSF CAREER Award in 2010. His papers have been selected for the outstanding student paper at ICML in 2004, the KDD best research paper honorable mention in 2010, the KDD best research paper nomination in 2011 and 2012, the SDM best research paper runner up in 2013, and the KDD best research paper runner up in 2013.*

**Jing Gao, University at Buffalo**

### Title: Inferring Information Trustworthiness from Multiple Sources of Heterogeneous Data

Big data leads to big challenges, not only in the volume of data but also in its variety. Multiple descriptions about the same sets of objects or events from different sources will unavoidably lead to data or information inconsistency. Then, among conflicting pieces of data or information, which one is more trustworthy, or represents the true fact? Facing the daunting scale of data, it is unrealistic to expect human to label or tell which data source is more reliable or which piece of information is correct. In this talk, I will discuss our research on integrating data of multiple sources to detect trustworthy information. We have developed a series of optimization-based methods that can automatically infer reliability of sources and facts by correlating and comparing multiple data sources. The effectiveness of the proposed methods is demonstrated on real data sets including weather predictions, multi-choice question answers and online hotel reviews.

**Bio:** *Jing Gao is currently an assistant professor in the Department of Computer Science at the University at Buffalo (UB), State University of New York. She received her PhD from Computer Science Department, University of Illinois at Urbana Champaign in 2011, and subsequently joined UB in 2012. She is broadly interested in data and information analysis with a focus on information integration, ensemble methods, mining data streams, transfer learning and anomaly detection. She has published more than 60 papers in referred journals and conferences and her work has received over 1300 citations. She has served as program committee member of many conferences including KDD, ICDM, SDM, ECML/PKDD, CIKM, ASONAM and BigData. More information about her research can be found at: http://www.cse.buffalo.edu/ jing.*