Association for Computing Machinery
Advancing Computing as a Science & Profession

KDD

CONTACT:
Jim Ormond
212-626-0505
ormond@hq.acm.org

KDD 2017 Showcases the Latest in Data Science and Machine Learning

World's Leading Experts Explore Urban Computing, Algorithmic Transparency, Social Applications of Big Data and Many Other Issues

New York, NY, July 26, 2017 – The Association for Computing Machinery's (ACM) Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) today announced highlights of KDD 2017, the group's flagship annual conference. KDD 2017 will be held in Halifax, Nova Scotia, Canada on August 13-17, 2017. Recognized as the longest-running and largest conference on knowledge discovery and data mining, KDD brings together leading researchers and practitioners in the fields of data science, data mining, knowledge discovery, large-scale data analytics, and Big Data.

KDD 2017 will feature exciting discussions, tutorials and workshops as well as showcase cutting-edge research papers.

"As the field of data science continues to expand rapidly, increasingly intertwining with our day-to-day lives, the KDD conference prides itself on being the place where several concepts such as Big Data, predictive analytics, and crowdsourcing originated," said Stan Matwin of Dalhousie University, General Co-Chair for KDD 2017. "A melting pot for entrepreneurs, data science industry leaders, leading professors and other thought leaders, KDD 2017 promises to be the breeding ground for new ideas and technological innovations resulting in both theoretical advances and new technologies having real-world impact."

KDD 2017 HIGHLIGHTS

Keynote Speakers *(All times local to Halifax, GMT-4)*

What's Fair?
Cynthia Dwork, Distinguished Scientist - Microsoft Research / Harvard University
Tuesday, August 15, 8:00 a.m. - 9:30 a.m.
Data, algorithms, and systems have biases embedded within them reflecting designers' explicit and implicit choices, historical biases, and societal priorities. They form, literally and inexorably, a codification of values. "Unfairness" of algorithms – for tasks ranging from advertising to recidivism prediction – has attracted considerable attention in the popular press. The talk will discuss the nascent mathematically rigorous study of fairness in classification and scoring.

Three Principles of Data Science: Predictability, Stability, and Computability
Bin Yu, Professor - University of California, Berkeley
Wednesday, August 16, 8:00 a.m. - 9:30 a.m.
Making prediction as its central task and embracing computation as its core, machine learning has enabled wide-ranging data-driven successes. Good prediction implicitly assumes stability

between past and future. Stability (relative to data and model perturbations) is also a minimum requirement for interpretability and reproducibility of data driven results. Obviously, both prediction and stability principles cannot be employed without feasible computational algorithms, hence the importance of computability. The three principles will be demonstrated through analytical connections, and in the context of two on-going projects, for which "data wisdom" is also indispensable.

The Future of Data Integration
Renée J. Miller, Professor - University of Toronto
Wednesday, August 16, 8:00 a.m. - 9:30 a.m.
The value of data explodes when it is integrated. This talk will present some important innovations in data integration over the last two decades. These include data exchange, which provides a foundation for reasoning about the correctness of transformed data, and the use of declarative mappings in integration. Discussions will also how data mining has been used to facilitate data integration and present some important new data integration challenges that arise in data science.

<u>Research Papers (Highlights)</u>

Accelerating Innovation through Analogy Mining
Tom Hope (Hebrew University of Jerusalem); Joel Chan (Carnegie Mellon University); Aniket Kittur (Carnegie Mellon University); Dafna Shahaf (Hebrew University of Jerusalem)
The availability of large idea repositories (e.g., the U.S. patent database) could significantly accelerate innovation and discovery by providing people with inspiration from solutions to analogous problems. In this paper, the authors explore the viability and value of learning structural representations, which specify the purpose of a product and the mechanisms by which that purpose is achieved. The proposed approach combines crowdsourcing and recurrent neural networks to extract purpose and mechanism vector representations from product descriptions. The analogies retrieved by the proposed models significantly increase people's likelihood of generating creative ideas compared to analogies retrieved by traditional methods.

TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams
Chao Zhang (University of Illinois at Urbana-Champaign); Liyuan Liu (University of Illinois at Urbana-Champaign); Dongming Lei (University of Illinois at Urbana-Champaign); Quan Yuan (University of Illinois at Urbana-Champaign); Honglei Zhuang (University of Illinois at Urbana-Champaign); Tim Hanratty (U.S. Army Research Lab); Jiawei Han (University of Illinois at Urbana-Champaign)
Detecting local events (e.g., protest, disaster) at their onsets is an important task for a wide spectrum of applications, ranging from disaster control to crime monitoring and place recommendation. The authors propose TrioVecEvent, a method that leverages multimodal embeddings of the location, time, and text to achieve accurate online local event detection. The proposed method improves the detection precision of the state-of-the-art method from 36.8% to 80.4% and the pseudo recall from 48.3% to 61.2%.

Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data
David Hallac (Stanford University); Sagar Vare (Stanford University); Stephen Boyd (Stanford University);Jure Leskovec (Stanford University)
Subsequence clustering of multivariate time series is a useful tool for discovering repeated

patterns in temporal data. Once these patterns have been discovered, seemingly complicated datasets can be interpreted as a temporal sequence of only a small number of states, or clusters. In this paper the authors propose a new method of model-based clustering called Toeplitz Inverse Covariance-based Clustering (TICC). Based on a correlation network based representation, TICC simultaneously segments and clusters the time series data. The paper demonstrates on an automobile sensor dataset how TICC can be used to learn interpretable clusters in real-world scenarios.

Applied Data Science Papers (Highlights)

HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Informat

Yanfang Ye (West Virginia University); Shifu Hou (West Virginia University);Yangqiu Song (West Virginia University)

With explosive growth of Android malware and due to the severity of its damages to smart phone users, the detection of Android malware has become an increasingly important topic in cyber security. In this paper, to detect Android malware, instead of using Application Programming Interface (API) calls only, the authors further analyze the different relationships between them and create higher-level semantics which require more efforts for attackers to evade the detection. The paper demonstrates that the developed system HinDroid system outperforms other Android malware detection techniques.

Prognosis and Diagnosis of Parkinson’s Disease Using Multi-Task Learning

Saba Emrani (SAS Institute Inc); Anya McGuirk (SAS Institute Inc.); Wei Xiao (SAS Institute Inc.)

Parkinson’s disease (PD) is a debilitating neurodegenerative disease excessively affecting millions of patients. Early diagnosis of PD is critical as manifestation of symptoms occur many years after the onset of neurodegenration, when more than 60% of dopaminergic neurons are lost. In this paper, the authors employ a multi-task learning regression framework for prediction of Parkinson’s disease progression, where each task is the prediction of PD rating scales at one future time point. The model is then used to identify the important biomarkers predictive of disease progression. The results confirm some of the important biomarkers identified in existing medical studies.

A Data Mining Framework for Valuing Large Portfolios of Variable Annuities

Guojun Gan (University of Connecticut); Jimmy Huang (York University)

A variable annuity is a tax-deferred retirement vehicle created to address concerns that many people have about outliving their assets. In the past decade, the rapid growth of variable annuities has posed great challenges to insurance companies especially when it comes to valuing the complex guarantees embedded in these products. In this paper, the authors propose a data mining framework to address the computational issue associated with the valuation of large portfolios of variable annuity contracts. The experimental results show that the proposed framework is able to produce accurate estimates of various quantities of interest and can reduce the runtime significantly compared to the state-of-the-art approaches.

Tutorials (Highlights)

Making Better Use of the Crowd

Jenn Wortman Vaughan (Microsoft Research)

This tutorial will showcase innovative uses of crowdsourcing that go beyond the collection of data. It will also dive into recent research aimed at understanding who "crowdworkers" are, how they behave, and what this should teach us about best practices for interacting with the crowd.

IoT in Practice: Case Studies in Data Analytics, with Issues in Privacy and Security
Albert Bifet (Telecom ParisTech); Latifur Khan (University of Texas at Dallas); Joao Gama (University of Porto); Wei Fan (Baidu Research Big Data Lab)
This tutorial is a gentle introduction to mining IoT big data streams. It introduces data stream learners for several learning tasks including distributed algorithms; presents applications for predictive maintenance, prediction for renewable energies, and social network analysis for telecommunications data streams; and dwells upon security concerns regarding IoT data streams containing sensitive and confidential data when predictive analytics is performed over a third-party cloud service.

Urban Computing: Enabling Intelligent Cities with AI and Big Data
Yu Zhen (Microsoft Research; Editor-in-Chief of ACM Transactions on Intelligent Systems and Technology)
This will provide an overview of the framework of urban computing, discussing its key challenges and methodologies from data science's perspective (particularly data mining). It will also present a diversity of urban computing applications, ranging from big data-driven environmental protection to transportation, from urban planning to urban economy.

Workshops (Highlights)

Data Science for Intelligent Food, Energy and Water
Monday, August 14, 8 a.m. to 5 p.m.
Machine Learning for Prognostics and Health Management
Monday, August 14, 1 p.m. to 5 p.m.
Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)
Monday, August 14, 8 a.m. to 12 p.m.

About SIGKDD

ACM SIGKDD, which stands for Special Interest Group for Knowledge Discovery and Data Mining (kdd.org), is a professional society comprising of world-renowned data scientists from industry and academia. KDD is the annually held, premier international conference that brings together researchers and practitioners from both academia and industry to deep-dive into novel ideas, latest research results and share in-the-trenches experiences and innovations.

About ACM

ACM, the Association for Computing Machinery (www.acm.org), is the world's largest educational and scientific computing society, uniting computing educators, researchers and professionals to inspire dialogue, share resources and address the field's challenges. ACM strengthens the computing profession's collective voice through strong leadership, promotion of the highest standards, and recognition of technical excellence. ACM supports the professional growth of its members by providing opportunities for life-long learning, career development, and professional networking.