Association for Computing Machinery

Advancing Computing as a Science & Profession

KDD

NEWS RELEASE

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

KDD 2020 SHOWCASES BRIGHTEST MINDS IN DATA SCIENCE AND AI

World's Largest and Oldest Data Mining Conference Goes Virtual for the First Time

New York, NY, August 21, 2020 – The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD) will hold its flagship annual conference, KDD 2020, virtually, August 23-27. The KDD conference series, started in 1989, is the world's oldest and largest data mining conference, and is the venue where concepts such as big data, data science, predictive analytics and crowdsourcing were first introduced. Continuing this tradition, KDD 2020 will showcase leading-edge research papers in data science, data mining, knowledge discovery, large-scale data analytics and big data. Despite being a fully virtual event, KDD 2020 will include all the same program offerings as previous years, including exciting keynote addresses, topical panels, invited talks, highly selective research and applied data science papers, informative and hands-on tutorials, and workshops.

"KDD is a 'must attend' conference, where the theory and practice in data science, machine learning and artificial intelligence come together in industry-defining innovations," explained KDD 2020 General Co-chair Rajesh K. Gupta, University of California, San Diego. "Initially we had hoped that at least a portion of the conference could be 'in-person,' but ultimately we decided a fully virtual conference would be the safest option for our community. While organizing a fully virtual conference is unchartered territory, we have made sure all of the program facets from previous years will be part of KDD 2020—from fascinating keynote addresses, to engaging research, workshops and panels. We've planned an outstanding program and we are confident we will have record conference registrations."

"Data science has exploded in the last 30 years and is now reshaping so many different disciplines," added KDD 2020 General Co-chair Yan Liu, University of Southern California. "An example of this is KDD 2020's Applied Data Science Invited Speakers track, which we are particularly excited about. This year, we have a roster of 18 leading practitioners in the field, working at companies such as Siemens, Microsoft, Facebook, Google, Amazon and Uber, among many others."

KDD 2020 will feature four keynote talks, 18 applied data science invited talks, and 217 accepted research papers grouped into 43 sessions for oral presentations, workshops and tutorials. A partial listing of highlights follows. The full KDD program is available here.

Keynote Talks

"Explanations that Matter through Meta-Provenance"

Yolanda Gil, University of Southern California

Provenance standards have now been used for many years to generate useful explanations of the data analytic process used to generate a new finding. These explanations convey the details of analytic steps and the original data used in an analysis. In this talk, Gil will discuss the need for explanations that provide the context and rationale for how the data

analysis process was designed. She will also illustrate with examples from several domains the kinds of explanations that can be generated from meta-provenance and discuss important areas of future work.

"AI for Intelligent Financial Services: Examples and Discussion"

Manuela M. Veloso, Carnegie Mellon University

There are many opportunities to pursue AI and ML in the financial domain. In this talk, I will overview several research directions we are pursuing in engagement with the lines of business, ranging from data and knowledge, learning from experience, reasoning and planning, multi agent systems, and secure and private AI. Veloso will offer concrete examples of projects, and conclude with the many challenges and opportunities that AI can offer in the financial domain.

"A Look at State-Space Multi-Taper Time-Frequency Analysis"

Dr. Emery Brown, Massachusetts Institute of Technology, Harvard Medical School, Massachusetts General Hospital

Time series arising for studies of physical, biological, economic and sociological systems are an important data class. The growing interest in this type of data has come about because of significant recent advances in sensor, recording and digitization technologies. In this lecture, Brown will discuss his recent work on the development of a state-space multi-taper (SS-MT) framework for the analysis of non-stationary time series.

"Computational Epidemiology at the Time of COVID-19"

Alessandro Vespignani, Northeastern University

The data science revolution is finally enabling the development of large-scale data-driven models that provide scenarios, forecasts and risk analysis for infectious disease threats. These models also provide rationales and quantitative analysis to support policy making decisions and intervention plans. Vespignani will review and discuss recent results and challenges in the area and focus on ongoing work aimed at responding to the COVID-19 pandemic.

Best Paper Awards

Best Paper: "On Sampled Metrics for Item Recommendation"

Walid Krichene, Steffen Rendle, Google

The task of item recommendation requires ranking a large catalogue of items given a context. Item recommendation algorithms are evaluated using ranking metrics that depend on the positions of relevant items. To speed up the computation of metrics, recent work often uses sampled metrics where only a smaller set of random items and the relevant items are ranked. The Google research team investigates sampled metrics in more detail and shows that they are inconsistent with their exact version, in the sense that they do not persist relative statements, e.g., *recommender A is better than B*, not even in expectation.

Best Student Paper: "TIPRDC: Task-Independent Privacy-Respecting Data Crowdsourcing Framework for Deep Learning with Anonymized Intermediate Representations"

Ang Li, Huanrui Yang, Yiran Chen, Duke University; Yixiao Duan, Jianlei Yang, Beihang University

The research group from Duke University presents TIPRDC, a task-independent privacy-respecting data crowdsourcing framework with anonymized intermediate representation. The goal of this framework is to learn a feature extractor that can hide the privacy information from the intermediate representations; while maximally retaining the original information embedded in the raw data for the data collector to accomplish unknown learning tasks.

Best Paper Runner Up: "Malicious Attacks against Deep Reinforcement Learning Interpretations"

Mengdi Huai, Jianhui Sun, Renqin Cai, Aidong Zhang, University of Virginia; Liuyi Yao, State University of New York at Buffalo

The combination of deep learning and reinforcement learning (RL) and has demonstrated its ability to model dynamics in a plethora of sequential decision-making problems. To improve the transparency, various interpretation methods for DRL have been proposed. However, those DRL interpretation methods make an implicit assumption that they are performed in

a reliable and secure environment, which is not true in practical applications. The University of Virginia team investigates the vulnerability of DRL interpretation methods in the malicious environment. Specifically, the first study of the adversarial attacks against DRL interpretations was introduced. An optimization framework was proposed to address the studied adversarial attacks.

SIGKDD Awards

2020 ACM SIGKDD Innovation Award

Thorsten Joachims, professor of Computer Science and Information Science at Cornell University, is recognized for his research contributions in machine learning, including influential work studying human biases in information retrieval, support vector machines (SVM) and structured output prediction. Notably, Joachims pioneered methods for eliciting reliable preferences from implicit feedback, methods for unbiased learning-to-rank and ranking methods that provide fairness guarantees. The ACM SIGKDD Innovation Award is the highest honor for technical excellence in the field of knowledge discovery and data mining. It is conferred on an individual or group of collaborators whose outstanding technical innovations have greatly influenced the direction of research and development in the field.

"I am greatly honored by this recognition from the KDD community," said Joachims. "KDD is known for innovation—not only as an academic endeavor, but also with an eye towards real-world impact and social good."

2020 ACM SIGKDD Service Award

Michael Zeller, head of artificial intelligence (AI) strategy and solutions at Temasek, is honored for his contributions to the field through dedication to ACM SIGKDD as the volunteer treasurer and secretary of the executive committee. Zeller has served on the executive board for eight years, playing an instrumental role in planning multiple KDD conferences. With a special emphasis on applied AI, his mission as an executive committee member is to foster strong partnerships between research institutions and industry organizations as a key for the continued success of the KDD community. The ACM SIGKDD Service Award is the highest recognition of service awarded in the field. The award honors an individual or group of collaborators for outstanding contributions to professional KDD societies or society-at-large through applications of knowledge discovery and data mining.

"As a longtime member of ACM SIGKDD, I am always incredibly impressed by the contributions of our volunteers," said Zeller. "Without their dedication and belief in our mission, we would never have been able to create such a vibrant data science community, let alone organize a conference of this magnitude and quality year after year."

2020 ACM SIGKDD Dissertation Award

Rediet Abebe, incoming assistant professor of Computer Science at the University of California at Berkeley, earned this year's ACM SIGKDD Dissertation Award for her Ph.D. thesis, "Designing Algorithms for Social Good." Abebe is the first female computer scientist to be inducted into the Harvard Society of Fellows and co-founded Mechanism Design for Social Good (MDSG), a multi-institutional initiative to improve access to opportunity for historically underserved and disadvantaged communities. Jingbo Shang, assistant professor of Computer Science at University of California at San Diego, earned runner-up for his thesis, "Constructing and Mining Heterogeneous Information Networks from Massive Text." The ACM SIGKDD Dissertation Award recognizes outstanding work done by graduate students in the areas of data science, machine learning and data mining.

2020 ACM SIGKDD Rising Star Award

Danai Koutra, Morris Wellman assistant professor of Computer Science and Engineering at University of Michigan, and **Jiliang Tang,** assistant professor of Computer Science and Engineering at Michigan State University, both received the first annual ACM SIGKDD Rising Star Award. Koutra's research in large-scale data mining focuses on principled, interpretable and scalable methods for network summarization and multi-network analysis. Tang's notable work includes research into representation learning, especially on graphs and its applications on the internet and social media domains. New this year,

the Rising Star Award celebrates individual work done in the first five years after earning a PhD. The award aims to celebrate the early accomplishments of the SIGKDD communities' brightest new minds.

2020 SIGKDD Test of Time Award for Research

The SIGKDD Test of Time award recognizes outstanding KDD papers, at least ten years old, which have had a lasting impact on the data mining research community and continue to be cited as the foundation for new branches of research. This year, the Test of Time Award for Research goes to **Victor S. Sheng**, **Foster Provost** and **Panagiotis Ipeirotis** for their approach to selective acquisition of multiple labels featured in the 2008 peer-reviewed paper, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers."

2020 SIGKDD Test of Time Award for Applied Science

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang and Zhong Su received the inaugural Test of Time Award for Applied Science in recognition of their study of mining academic social networks published in the 2008 peer-reviewed paper, "ArnetMiner: Extraction and Mining of Academic Social Networks." SIGKDD introduced this award to honor influential research in real-world applications of data science.

Research Track Papers

(Partial list: The full program of Research Track Papers can be found here.)

"A Novel Deep Learning Model by Stacking Conditional Restricted Boltzmann Machine and Deep Neural Network"

Tianyu Kang, Ping Chen, Wei Ding, University of Massachusetts Boston; John Quackenbush, Harvard T.H. Chan School of Public Health

A real-world system often exhibits complex dynamics arising from interaction among its subunits. In machine learning and data mining these interactions are usually formulated as dependency and correlation among system variables. The collaborative research teams from Harvard University and University of Massachusetts Boston present a novel deep learning model to tackle functionally interactive features by stacking a Conditional Restricted Boltzmann Machine and a Deep Neural Network (CRBM-DNN). The new model can solve both supervised and unsupervised learning problems. Compared to a regular neural network of the same size, CRBM-DNN has fewer parameters so they require fewer training samples.

"Parameterized Correlation Clustering in Hypergraphs and Bipartite Graphs"

Nate Veldt, Cornell University; David F. Gleich, Purdue University; Anthony Wirth, The University of Melbourne

Motivated by applications in community detection and dense subgraph discovery, this paper considers new clustering objectives in hypergraphs and bipartite graphs. These objectives are parameterized by one or more resolution parameters in order to enable diverse knowledge discovery in complex data. The experimental results highlight the flexibility of the new framework and the diversity of results that can be obtained in different parameter settings.

"Multi-level Graph Convolutional Networks for Cross-platform Anchor Link Prediction"

Hongxu Chen, Bogdan Gabrys, Katarzyna Musial, University of Technology Sydney; Hongzhi Yin, Tong Chen, The University of Queensland; Xiangguo Sun, Southeast University

Cross-platform account matching plays a significant role in social network analytics, and is beneficial for a wide range of data mining applications. However, existing methods either heavily rely on high-quality user generated content (including user profiles) or suffer from data insufficiency problem if only focusing on network topology, which brings researchers into an insoluble dilemma of model selection. The Australian research team proposes a novel framework that considers multi-level graph convolutions on both local network structure and hypergraph structure in a unified manner. The proposed method overcomes data insufficiency problem of existing work and does not necessarily rely on user demographic information.

“GHashing: Semantic Graph Hashing for Approximate Similarity Search in Graph Databases”

Zongyue Qin, Peking University; Yunsheng Bai, Yizhou Sun, University of California, Los Angeles

Graph similarity search is an important data mining problem. Existing methods are capable of managing databases with thousands or tens of thousands of graphs. However, how to scale the graph similarity search to databases that have hundreds of thousands or even millions of graphs remains a challenging problem. Inspired by the recent success of deep learning-based supervised hashing, called semantic hashing, in image and document retrieval, this paper proposes a novel graph neural network (GNN) based pruning approach, GHashing, for graph similarity search. Exploiting the powerful learning ability of deep neural networks and the efficiency of hashing methods for approximate nearest neighbor, GHashing demonstrates significantly better performance compared to state-of-the-art methods.

Applied Data Science Track Papers

(Partial list: The full program of Applied Data Science Track Papers can be found here.)

“AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types”

Gabriel Blanco Saldana, Saurabh Deshpande, Xin Luna Dong, Xiang He, Andrey Kan, Xian Li, Yan Liang, Jun Ma, Alexandre Michetti Manduca, Jay Ren, Surender Pal Singh, Fan Xiao, Yifan Ethan Xu, Chenwei Zhang, Tong Zhao, Amazon; Haw-Shiuan Chang, University of Massachusetts Amherst; Giannis Karamanolakis, Columbia University; Yuning Mao, University of Illinois at Urbana Champaign, Yaqing Wang, State University of New York at Buffalo, Christos Faloutsos, Carnegie Mellon University; Andrew McCallum, University of Massachusetts Amherst; Jiawei Han, University of Illinois at Urbana Champaign

Can one build a knowledge graph (KG) for all products in the world? Knowledge graphs have firmly established themselves as valuable sources of information for search and question answering, and it is natural to wonder if a KG can contain information about products offered at online retail sites. There have been several successful examples of generic KGs, but organizing information about products poses many additional challenges, including sparsity and noise of structured data for products, complexity of the domain with millions of product types and thousands of attributes, heterogeneity across large number of categories, as well as large and constantly growing number of products. In this paper, the authors describe AutoKnow, their automatic (self-driving) system that addresses these challenges.

“BusTr: Predicting Bus Travel Times from Real-Time Traffic”

Richard Barnes, UC Berkeley; Senaka Buthpitiya, James Cook, Alex Fabrikant, Andrew Tomkins, Fangzhou Xu, Google Research

The authors present BusTr, a machine-learned model for translating road traffic forecasts into predictions of bus delays, used by Google Maps to serve the majority of the world's public transit systems where no official real-time bus tracking is provided. They demonstrate that their neural sequence model improves over DeepTTE, the state-of-the-art baseline, both in performance (-30% MAPE) and training stability. They also demonstrate significant generalization gains over simpler models, evaluated on longitudinal data to cope with a constantly evolving world.

“Identifying Homeless Youth At-Risk of Substance Use Disorder: Data-Driven Insights for Policymakers”

Maryam Tabar, Stephanie Winkler, Dongwon Lee, Amulya Yadav, Pennsylvania State University; Heesoo Park, Sungkyunkwan University; Anamika Barman-Adhikari, University of Denver

Substance Use Disorder (SUD) is a devastating disease that leads to significant mental and behavioral impairments. Unfortunately, there is no definitive data-driven study on analyzing factors associated with SUD among homeless youth. The authors aim to fill this gap by making the following three contributions: (i) they use a real-world dataset collected from 1,400 homeless youth (across six American states) to build accurate Machine Learning (ML) models for predicting the susceptibility of homeless youth to SUD; (ii) they find a representative set of factors associated with SUD among this population by analyzing feature importance values associated with their ML models; and (iii) they investigate the effect of geographical heterogeneity on the factors associated with SUD.

"Managing Diversity in Airbnb Search"

Mustafa Abdool, Malay Haldar, Prashant Ramanathan, Tyler Sax, Lanbo Zhang, Aamir Manasawala, Shulin Yang, Bradley Turnbull, Qing Zhang, Thomas Legrand, Airbnb

One of the long-standing questions in search systems is the role of diversity in results. From a product perspective, showing diverse results provides the user with more choice and should lead to an improved experience. However, this intuition is at odds with common machine learning approaches to ranking which directly optimize the relevance of each individual item without a holistic view of the result set. In this paper, the authors describe their journey in tackling the problem of diversity for Airbnb search, starting from heuristic based approaches and concluding with a novel deep learning solution that produces an embedding of the entire query context by leveraging Recurrent Neural Networks (RNNs).

Applied Data Science Invited Talks

KDD's Applied Data Science Invited Talks feature highly influential speakers who have directly contributed to successful data mining applications in finance, healthcare, bioinformatics, public policy, infrastructure, telecommunications, social media and computational advertising. The invited talks and speakers include:

- **"Toward Responsible AI by Planning to Fail,"** *Saleema Amershi, Microsoft Research*
- **"Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning,"** *Timnit Gebru, Google*
- **"Fairness, Accountability, and Transparency in Predictive Models in Criminal Justice,"** *Kristian Lum, HRDAG*
- **"Multimodal Machine Learning for Video and Image Analysis,"** *Shalini Ghosh, Samsung Research America*
- **"Unleashing the Power of Subjective Data: Managing Experiences as First-Class Citizens,"** *Wang-Chiew Tan, Megagon Labs*
- **"Geospatial Technologies for Ride-Hailing and Emergency Vehicle Fleets,"** *Dawn Woodard, Uber*
- **"AI: Healthcare's Prescription for Transformation,"** *Taha Kass-Hout, Amazon*
- **"Using Machine Learning to Detect Cancer Early,"** *Jan Schellenberger, Grail*
- **"Artificial Intelligence for Healthcare,"** *Dorin Comaniciu, Siemens Healthineers*
- **"A perspective from the first U.S. Chief Data Scientist,"** *DJ Patil, Devoted Health*
- **"Innovating with Language AI,"** *Ashwin Ram, Google*
- **"Build the State-of-the-art Machine Technology for the Crypto Economy,"** *Michael Li, Coinbase*
- **"Representation Learning, Inference, and Reasoning,"** *Fernando Pereira, Google*
- **"Straddling the Boundary Between Contribution and Solution Driven Science,"** *Daniel Marcu, Amazon*
- **"Data Paucity and Low Resource Scenarios: Challenges and Opportunities,"** *Mona Diab, Facebook AI and The George Washington University*
- **"Next-Generation Frameworks,"** *Anima Anandkumar, Nvidia and Caltech*
- **"Preserving Integrity in Online Social Media,"** *Alon Halevy, Facebook*
- **"How AI Can Help Build Resiliency for Small Business in a Global Economic Crisis,"** *Nhung Ho, Intuit*

Panels

This year's conference will feature three panels:

"Fighting a pandemic: convergence of expertise, data science and policy," a panel of experts from around the globe will address the challenges and opportunities of using data science to fight a pandemic. Panelists will cite real-world cases where using data science helped the fight against the pandemic and cautionary tales of when it hindered that fight. Panelists include Vittoria Colizza, INSERM & Sorbonne University, France; Lauren Gardner, Johns Hopkins University, US; Marcel Salathé, EPFL, Switzerland; Samuel Scarpino, Northeastern University, US; Joseph T. Wu, University of Hong Kong, Hong Kong, China.

The **"Women's Panel"** brings together a diverse group of women scientists and leaders to learn more about their journeys, what it takes to have a successful career in the field of Data Science and Artificial Intelligence, and their crucial work on some of the pressing problems of today. Prakruthi Prabhakar from LinkedIn will moderate the panel, which includes Calandra Moore, Data Scientist, Department of Defense; Elaine O. Nsoesie, Assistant Professor, Boston University School of Public Health; Karin Kimbrough, Chief Economist, LinkedIn; Sihem Amer-Yahia, Research Director, CNRS; Subarna Sinha Engineering Leader, Machine Learning, 23andMe; and Vanessa Murdock, Applied Science Manager, Amazon.

The panel **"The Near Future of Automated Data Science"** will explore the demand for creative, technically-skilled data scientists in the United States. Although the demand for data scientists is booming, our ability to train students to fill those jobs is falling behind. There is also a growing concern that technical skill alone is insufficient for long-term data science career success, partially due to the fact that many data science tasks are being automated. The panel will discuss the skill sets data scientists should focus on, such as business understanding, explanation, and storytelling. Panelists include Danielle Gewurz, Deloitte Consulting; Shubha Nabar, Faras AI; Monica Rogati, Data Science and AI Advisor; Horst Samulowitz, IBM Watson Research Center.

About SIGKDD

The ACM Special Interest Group for Knowledge Discovery from Data (SIGKDD) is a professional society comprising world-renowned data scientists from industry and academia. KDD is the 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. More details can be found at kdd.org.

About ACM

ACM, the Association for Computing Machinery, 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.

###