Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

Over one terabyte of data released to help researchers benchmark distributed learning algorithms in critical research

New York – June 18, 2015 – Criteo (NASDAQ: CRTO), the performance marketing technology company, today announced the release of the largest public machine learning dataset ever issued to the open source community, with the goal of supporting academic research and innovation in distributed machine learning algorithms.

With the increasing prevalence of large-scale data problems across industries, including performance advertising, the release of datasets such as this are necessary to advance research in the academic space and drive industry progress. Anonymized datasets pulled from real-world applications allow academic researchers to test, refine and advance the machine learning platforms that so many enterprises now rely on. Criteo, for example, relies on its own proprietary distributed learning algorithms to accurately predict when a consumer is most likely to click on a particular ad, thereby increasing the return on an advertiser's investment in ad delivery.

"Accuracy and speed of machine learning algorithms are critical to the success of our business and many others, but they would be impossible to achieve without publicly available datasets," said Olivier Chapelle, Principal Research Scientist at Criteo. "As an active participant in open source projects, Criteo is committed to supporting the machine learning community to drive open source innovation. Recognizing the essential role that independent research plays, we released this unique data set in order to facilitate as much innovation as possible, benefiting the academic community and the industry as a whole, as we collectively work towards improving machine learning technology."

Criteo is able to provide the independent research community with the largest dataset ever released by an organisation thanks to its massive reach, direct relationships with publishers and advertisers, and high consumer engagement with the performance-focused advertisements it powers. The company sees 30 billion HTTP requests per day (including as many as two million requests per second), delivers three billion unique banner advertisements per day, and stores 20 terabytes of new data daily with a capacity for 37 petabytes of raw storage.

With more than four billion lines totalling over one terabyte in size, the newly-released dataset builds on Criteo's click prediction dataset, originally released as part of its Display Advertising Challenge conducted with Kaggle. The latest dataset has already been put to use as a benchmark by researchers at Carnegie Mellon University, with more academic and research applications forthcoming.

"Criteo's one terabyte dataset has proven invaluable for benchmarking the scalability of the learning algorithms for high throughput click-through-rate estimation, which we are developing as part of our Marianas Labs project," said Alexander Smola, Professor at Carnegie Mellon University. "We look forward to continuing to partner with Criteo on additional datasets in the future."

Criteo's terabyte dataset is hosted on Microsoft Azure, and details on how to access, utilize and download it can be found at Criteo Labs. For more information about Criteo and its technology please visit www.criteo.com.

###

About Criteo
Criteo delivers personalized performance marketing at an extensive scale. Measuring return on post-click sales, Criteo makes ROI transparent and easy to measure. Criteo has over 1,500 employees in 24 offices across the Americas, Europe and Asia-Pacific, serving over 7,000 advertisers worldwide with direct relationships with over 10,000 publishers.

For more information, please visit http://www.criteo.com

Contact:
Emma Ferns
Criteo
e.ferns@criteo.com