**Probabilistic Hashing for Efficient Search and Learning on Massive Data**

Ping Li, Cornell University

Modern applications in the context of search often encounter massive high-dimensional binary data. For example, using n-gram representations, documents are often parsed to be binary (0/1) vectors in billions, trillions, quadrillions, or even higher dimensions. How to efficiently store, transmit, and search these data is a very interesting research topic with numerous applications in the industry. This talk focuses on a probabilistic hashing method named b-bit minwise hashing, which stores only the lowest b bits (for small b) of each hashed value after applying the standard minwise hashing procedure. Theoretically, it can be shown that b-bit minwise hashing improves minwise hashing at least by 21.3-fold when the threshold similarity (i.e., resemblance) is 0.5. More interestingly, we realize that b-bit minwise hashing can be seamlessly integrated with (i) logistic regression and SVM to solve extremely large predictive learning problems, and (ii) locality sensitive hashing (LSH) for sub-linear time near neighbor search. Extensive experiments will be presented.