Distinguished
LECTURE SERIES
ONR 70TH ANNIVERSARY EDITION

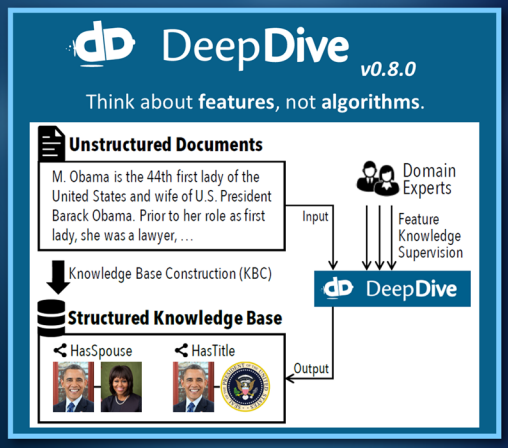
DeepDive:
A Dark Data System

Presented by:

Dr. Christopher Ré

Assistant professor in the Department of Computer Science at Stanford University; 2015 MacArthur Genius Award recipient

SEPT 6
2016
11 AM
IN THE JUNKER CENTER



Office of Naval Research
875 N. Randolph St., Arlington, Virginia
Bobby Junker Executive Conference Center, 14th Floor

Stream: **https://conference.apps.mil/webconf/onrdls**

DeepDive: A Dark Data System

Many pressing questions in science are macroscopic, as they require scientists to integrate information from numerous data sources, often expressed in natural languages or in graphics; these forms of media are fraught with imprecision and ambiguity and so are difficult for machines to understand. Here I describe DeepDive, which is a new type of system designed to cope with these problems. It combines extraction, integration and prediction into one system. For some paleobiology and materials science tasks, DeepDive-based systems have surpassed human volunteers in data quantity and quality (recall and precision). DeepDive is also used by scientists in areas including genomics and drug repurposing, by a number of companies involved in various forms of search, and by law enforcement in the fight against human trafficking. DeepDive does not allow users to write algorithms; instead, it asks them to write only features. A key technical challenge is scaling up the resulting inference and learning engine. I will describe our theoretical work in computing without using traditional synchronization methods for Gibbs sampling and stochastic methods for non-convex optimization problems. In addition, this has led to new ways to analyze convergence for sequential Gibbs sampling. I will also discuss the brand-new successor to DeepDive, called Snorkel, that allows lightweight extraction tasks but without requiring and hand-tuned feature engineering.

ABOUT

Dr. Christopher Ré

Christopher (Chris) Ré is an assistant professor in the Department of Computer Science at Stanford University. His work's goal is to enable users and developers to build applications that more deeply understand and exploit data. He spent four wonderful years on the faculty of the University of Wisconsin, Madison, before moving to Stanford in 2013. He helped discover the first join algorithm with worst-case optimal running time, which won the best paper at PODS 2012. He also helped develop a framework for feature engineering that won the best paper at SIGMOD 2014. He also helped understand the fundamental limits of asychrony for Gibbs Sampling, which won best paper at ICML 2016. In addition, work from his group has been incorporated into scientific efforts including the IceCube neutrino detector and PaleoDeepDive, and into Cloudera's Impala and products from Oracle, Pivotal, and Microsoft's Adam. He received an SIGMOD Dissertation Award in 2010, NSF CAREER Award in 2011, an Alfred P. Sloan Fellowship in 2013, a Moore Data Driven Investigator Award in 2014, the VLDB early Career Award in 2015, the MacArthur Foundation Fellowship in 2015, and an Okawa Research Grant in 2016.