(https://github.com/HazyResearch/deepdive)

DeepDive

(http://deepdive.stanford.edu/)

- Quick Start (http://deepdive.stanford.edu/quickstart)
- Documentation (http://deepdive.stanford.edu/#documentation)
- Showcase (http://deepdive.stanford.edu/showcase/apps)
- Papers (http://deepdive.stanford.edu/papers)
- Data (http://deepdive.stanford.edu/opendata)
- Chat (https://gitter.im/HazyResearch/deepdive)
- Forum (https://groups.google.com/d/forum/deepdive-users)

undefined

undefined

undefined

undefined

undefined

undefined

undefined

Papers

DeepDive System Overview

- *Best Reference:* DeepDive: A Data Management System for Automatic Knowledge Base Construction. (http://pages.cs.wisc.edu/%7Eczhang/zhang.thesis.pdf) Ce Zhang.Ph.D. Dissertation, University of Wisconsin-Madison, 2015.

Applications

- *Best Reference:* Machine-compiled Macroevolutionary History of Phanerozoic Life. (http://arxiv.org/abs/1406.2963) Shanan E. Peters, Ce Zhang, Miron Livny, and C. Ré. PloS ONE.

- The Mobilize Center: an NIH big data to knowledge center to advance human movement research and improve mobility. (http://jamia.oxfordjournals.org/content/early/2015/08/13/jamia.ocv071.abstract) Joy P Ku, Jennifer L Hicks, Trevor Hastie, Jure Leskovec, C. Ré, and Scott L Delp. AMIA, 2015.

- Large-scale extraction of gene interactions from full text literature using DeepDive. Emily Mallory, Ce Zhang, C. Ré, and Russ Altman. Bioinformatics 2015.

- A Demonstration of Data Labeling in Knowledge Base Construction. (http://cs.stanford.edu/people/chrismre/papers/p2148-shin-vldb-demo.pdf) Jaeho Shin, Mike Cafarella, and C. Ré. VLDB Demo, 2015.

- Elementary: Large-scale Knowledge-base Construction via Machine Learning and Statistical Inference. (http://www.cs.stanford.edu/people/chrismre/papers/elementary_journal.pdf) F. Niu, Ce Zhang, C. Ré, and J. Shavlik. IJSWIS, Special Issue on Knowledge Extraction from the Web, 2012.

- Stanford's 2014 Slot Filling Systems. (http://i.stanford.edu/hazy/papers/2014kbp-systemdescription.pdf) Gabor Angeli, Sonal Gupta, Melvin Jose, Christopher D. Manning, Christopher Ré, Julie Tibshirani, Jean Y. Wu, Sen Wu, Ce Zhang.

- Using Social Media to Measure Labor Market Flow. (http://www-personal.umich.edu/%7Eshapiro/papers/LaborFlowsSocialMedia.pdf) D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, and M. Shapiro. NBER. Selected for NBER Digest (http://www.nber.org/digest/aug14/aug14.pdf).

- Feature Selection in Enterprise Analytics: A Demonstration using an R-based Data Analytics System (http://cs.stanford.edu/people/chrismre/papers/ColumbusDemo-VLDB.pdf) P. Konda, A. Kumar, C. Ré, and V. Sashikanth. VLDB Demo 2013

- GeoDeepDive: Statistical Inference using Familiar Data-Processing Languages. (http://www.cs.stanford.edu/people/chrismre/papers/gdd_demo.pdf) Ce Zhang, V. Govindaraju, J. Borchardt, T. Foltz, C. Ré, and S. Peters. SIGMOD 13 (demo).

- Hazy: Making it Easier to Build and Maintain Big-data Analytics. (http://queue.acm.org/detail.cfm?id=2431055) Arun Kumar, Feng Niu, and C. Ré ACM Queue, 2013. Invited to CACM March 2013

- Building an Entity-Centric Stream Filtering Test Collection for TREC 2102 (http://trec.nist.gov/pubs/trec21/papers/KBA.OVERVIEW.pdf). J.R. Frank, M.Kleiman-Weiner, D. A. Roberts, F.Niu, Ce Zhang, C. Ré, and I. Soboroff. TREC 2013

Feature Engineering

- *Best Reference:* Feature Engineering for Knowledge Base Construction. (http://arxiv.org/abs/1407.6439) DeepDive Group. Data Engineering Bulletin.

- Brainwash: A Data System for Feature Engineering. (http://www.cs.stanford.edu/people/chrismre/papers/mythical_man.pdf) M. Anderson et al. CIDR Conference 2013 (Vision Track)

- Materialization Optimizations for Feature Selection. (http://cs.stanford.edu/people/chrismre/papers /mod539-zhang.pdf) Ce Zhang, Arun Kumar, and C. Ré. SIGMOD 2014. Best Paper Award.
- Parallel Feature Selection Inspired by Group Testing. (http://papers.nips.cc/paper/5296-parallel-feature- selection-inspired-by-group-testing.pdf) Y. Zhou et al. NIPS2014.

The Underlying Engine

- *Best Overall Description:* Incremental Knowledge Base Construction Using DeepDive (http://i.stanford.edu /hazy/papers/inc.pdf) Sen Wu, Ce Zhang, Christopher De Sa, Jaeho Shin, Feiran Wang, and C. Ré. VLDB. 2015.
- *Best Reference for Sampler:* DimmWitted: A Study of Main-Memory Statistical Analytics. (http://arxiv.org /abs/1403.7550) Ce Zhang and C. Ré. VLDB 2014.
- *Best Reference for Grounding:* Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS. (http://www.cs.stanford.edu/people/chrismre/papers/tuffy-vldb11.pdf) F. Niu, C. Ré, A.Doan, and J.W. Shavlik. PVLDB 11.
- Rapidly Mixing Gibbs Sampling for a Class of Factor Graphs Using Hierarchy Width. Christopher De Sa, Ce Zhang, Kunle Olukotun, and C. Ré. NIPS 2015 (Spotlight).
- Taming the Wild: A Unified Analysis of Hogwild!-Style Algorithms. (http://arxiv.org/abs/1506.06438) Christopher De Sa, Ce Zhang, Kunle Olukotun, and C. Ré. NIPS 2015.
- Towards High-Throughput Gibbs Sampling at Scale: A Study across Storage Managers. (http://www.cs.stanford.edu/people/chrismre/papers/elementary_sigmod.pdf) Ce Zhang and C. Ré. SIGMOD 2013
- An Asynchronous Parallel Stochastic Coordinate Descent Algorithm. (http://arxiv.org/abs/1311.1873) J. Liu, S. Wright, C. Ré, V. Bittorf, S. Sridhar. ICML 2014.
- Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. (http://i.stanford.edu /hazy/papers/hogwild-nips.pdf) F. Niu, B. Recht, C. Ré, and S. J. Wright. NIPS, 2011
- Felix: Scaling Inference for Markov Logic with an Operator-based Approach. (http://arxiv.org /pdf/1108.0294v1) Feng Niu, Ce Zhang, C. Ré, and Jude Shavlik.

Some Related Techniques

- Understanding Tables in Context Using Standard NLP Toolkits. (http://cs.stanford.edu/people/chrismre /papers/jointable-acl.pdf) Vidhya Govindaraju, Ce Zhang, and C. Ré. ACL 2013
- Probabilistic Management of OCR using an RDBMS. (http://www.cs.stanford.edu/people/chrismre/papers /HazyOCR_VLDB2012.pdf) Arun Kumar and C. Ré. PVLDB 2012.
- Optimizing Statistical Information Extraction Programs Over Evolving Text. (http://www.cs.stanford.edu /people/chrismre/papers/crflex-icde12.pdf) Fei Chen, Xixuan Feng, C. Ré, and Min Wang. ICDE
- Incrementally maintaining classification using an RDBMS. (http://www.cs.stanford.edu/people/chrismre /papers/hazy-classification-vldb11.pdf) Mehmet Levent Koc and C. Ré. PVLDB Volume 4, 2011, p. 302-313

Studies

- Big Data versus the Crowd: Looking for Relationships in All the Right Places. (http://www.cs.stanford.edu /people/chrismre/papers/acl2012.pdf) Ce Zhang, Feng Niu, C. Ré, and Jude Shavlik. ACL, 2012.

Formal Foundations

- Probabilistic Databases. (http://www.morganclaypool.com/doi/pdfplus/10.2200 /S00362ED1V01Y201105DTM016) Dan Suciu, Dan Olteanu, C. Ré, and Christoph Koch. Morgan Claypool's Synthesis Lectures, 2011
- Transducing Markov Sequences. (http://dl.acm.org/citation.cfm?doid=2630065) Benny Kimelfeld and C. Ré. JACM 2014.
- Queries and materialized views on probabilistic databases. (http://dx.doi.org/10.1016/j.jcss.2010.04.006) Nilesh N. Dalvi, C. Re, and Dan Suciu. JCSS 2011.

Hazy Research Group, InfoLab (http://infolab.stanford.edu), Computer Science Department, Stanford University (http://cs.stanford.edu).

Questions? Chat with us (https://gitter.im/HazyResearch/deepdive) or ask deepdive-users (https://groups.google.com/d/forum/deepdive-users)!
Fork DeepDive on Github (https://github.com/HazyResearch/deepdive) and join deepdive-dev (https://groups.google.com/d/forum/deepdive-dev).