The **IBM Almaden Principles and Methodologies Group** (the theory group), part of the Accelerated Discovery Lab, mainly devises and analyzes algorithms for a variety of problems, using their expertise in areas including optimization, probability, linear algebra, machine learning, and database theory. As well as publishing in top theory conferences and journals, they also collaborate with system builders on a variety of applied projects.

Theoretical work focuses on *provability*, *formality*, and *abstraction*; each of these has practical significance: for example, preservation of secrecy and privacy require the strongest possible guarantees, as given by formal mathematical proofs, while abstraction often implies simplicity, in addition to breadth of application. Sometimes the most significant contribution of theory is simply to put work in an area into a formal setting, which can greatly clarify issues and relations.

A major area for the theory group in recent years has been algorithms for big data, with work on:

- applying *sketching* techniques to matrix computations, including work for DARPA, discussed below;
- understanding how to re-use data in a statistically valid way [4];
- *streaming* algorithms, for handling high-velocity, high-volume data [18];
- putting matrix and database computations in a common framework;
- caching and other data management techniques, discussed below.

The theory group has worked on applications including in social media[13], computer vision [12], data management for Hadoop [5], information extraction from unstructured text[6, 17], neural network chip design, building dictionaries of words using text corpora, and improving detection of extra-terrestrial signals with the SETI project. Here are some specific examples of recent work:

Matrix computations via sketching. The Sketching Linear Algebra Kernel (libSkylark) is a open-source library for matrix computations, suitable for general statistical data analysis and optimization applications, built under the DARPA XDATA program. The library implements the technique of *sketching*, which compresses matrices in a way that preserves key matrix properties. Sketching can accelerate solution of regression, low-rank approximation, and other fundamental matrix computations. A key recent development in this area was sketching methods that take *input-sparsity time*, that is, time that is proportional to the input size [3], a significant breakthrough in this area. Work continues on matrix algorithms based on this technique.

Online decisions under uncertainty for the smart grid. Renewable power sources are key for the next-generation smart gird. However, they are both unpredictable and intermittent, leading to significant challenges in planning, which must shift from offline to online. Theory group member T.S. Jayram and his co-authors have developed simple and efficient planning algorithms in this setting, including some that can make effective use of short-term predictions of future conditions [15, 16, 11].

Caching for tiered storage. Group member Nimrod Megiddo is the co-developer of ARC, the Almaden Replacement Cache, a well-known caching maintenance technique that improves in measurable, provable ways on classical techniques such as LRU [14]. Megiddo has recently worked on other cache maintenance problems, such as managing a storage hierarchy that has both a Solid-State Drive (SSD) and disk. Here the problem is to decide, based on efficiency and cost constraints, which units of data should be stored in the SSD. He was able to formulate this as an integer programming problem that due to its form was easy to solve (in contrast to most such problems).

Learning with statistical queries. A *statistical query* on a dataset is one that requests only aggregate properties, such as correlations between features, or averages. A learning algorithm that makes only statistical queries has significant advantages: it is noise tolerant and privacy-preserving, and can take advantage of parallel access to the data. However, such an algorithm also has limited power, and there are many basic questions about the possible uses of such algorithms, and various tradeoffs. Group member Vitaly Feldman has studied this area extensively [10, 8, 9], designing new learning algorithms using statistical queries, and characterizing the complexity of solving problems by using them.

Information Extraction. A key task of information extraction is the creation and use of relations extracted from text. Ron Fagin *et al.* [6] developed a foundational framework whose central construct is the *spanner*. The *spans* of a string are intervals $[a, b)$ that specify substrings of it. A spanner maps an input string into a relation over its spans. Fagin *et al.* give several ways to represent spanners, with various powers of expressivity, and give ways to clean inconsistent data [7]. Fagin, Phokion Kolaitis and co-authors [1, 2] gave a declarative approach to *linking* entities. Here to link entities means, e.g., to say that "Mary Smith" is the same person as "M. Smith," or that "Mary Smith" works for "IBM".

References

[1] Douglas Burdick, Ronald Fagin, Phokion G Kolaitis, Lucian Popa, and Wang-Chiew Tan. A declarative framework for linking entities. *ACM Transactions on Database Systems (TODS)*, 41(3):17, 2016.

[2] Douglas Burdick, Ronald Fagin, Phokion G Kolaitis, Lucian Popa, and Wang-Chiew Tan. A declarative framework for linking entities. In *Proc. Int. Conf. on Database Theory (ICDT)*, 2017.

[3] Kenneth L. Clarkson and David P. Woodruff. Low Rank Approximation and Regression in Input Sparsity Time. In *Proc. 45th ACM Symp. on Theory of Computing*, STOC '13, pages 81–90, New York, NY, USA, 2013. ACM. Winner, Best Paper Award. To appear, by invitation, JACM.

[4] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[5] Mohamed Y Eltabakh, Fatma Özcan, Yannis Sismanis, Peter J Haas, Hamid Pirahesh, and Jan Vondrak. Eagle-eyed elephant: split-oriented indexing in hadoop. In *Proc. 16th Int. Conf. on Extending Database Technology*, pages 89–100, 2013.

[6] Ronald Fagin, Benny Kimelfeld, Frederick Reiss, and Stijn Vansummeren. Document spanners: A formal approach to information extraction. *J. ACM*, 62(2):12, 2015.

[7] Ronald Fagin, Benny Kimelfeld, Frederick Reiss, and Stijn Vansummeren. Declarative cleaning of inconsistencies in information extraction. *ACM Transactions on Database Systems (TODS)*, 41(1):6, 2016.

[8] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proc. 45th ACM Symp. on Theory of Computing (STOC)*, pages 655–664. ACM, 2013.

[9] Vitaly Feldman and Varun Kanade. Computational bounds on statistical query learning. In *COLT*, pages 16–1, 2012.

[10] Maria Florina Balcan and Vitaly Feldman. Statistical active learning algorithms for noise tolerance and differential privacy. *arXiv preprint arXiv:1307.3102*, 2013.

[11] Vikas K Garg, TS Jayram, and Balakrishnan Narayanaswamy. Online optimization with dynamic temporal uncertainty: Incorporating short term predictions for renewable integration in intelligent energy systems. In *AAAI*, 2013.

[12] Ravindra Kumar, Ting Chen, Marcus Hardt, David Beymer, Karen Brannon, and Tanveer Syeda-Mahmood. Multiple kernel completion and its application to cardiac disease discrimination. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symp. on*, pages 764–767. IEEE, 2013.

[13] Jalal Mahmud, Michelle X Zhou, Nimrod Megiddo, Jeffrey Nichols, and Clemens Drews. Recommending targeted strangers from whom to solicit information on social media. In *Proc. 2013 Int. Conf. Intelligent User interfaces*, pages 37–48. ACM, 2013.

[14] Nimrod Megiddo and Dharmendra S Modha. ARC: A Self-Tuning, Low Overhead Replacement Cache. In *FAST*, volume 3, pages 115–130, 2003. Winner, USENIX Test of Time Award, 2014.

[15] Balakrishnan Narayanaswamy, Vikas K Garg, and TS Jayram. Online optimization for the smart (micro) grid. In *Proc. 3rd Int. Conf. Future Energy Systems*, page 19. ACM, 2012.

[16] Balakrishnan Narayanaswamy, Vikas K Garg, and TS Jayram. Prediction based storage management in the smart grid. In *Proc. 3rd IEEE Int. Conf. Smart Grid Communications (SmartGridComm)*, pages 498–503. IEEE, 2012.

[17] Sudeepa Roy, Laura Chiticariu, Vitaly Feldman, Frederick R Reiss, and Huaiyu Zhu. Provenance-based dictionary refinement in information extraction. In *Proc. 2013 ACM SIGMOD Int. Conf. on Management of Data*, pages 457–468. ACM, 2013.

[18] David P Woodruff. Data Streams and Applications in Computer Science. *Bulletin of EATCS*, 3(114), 2014.