Samuel Hopkins

I study algorithms and computational complexity for statistical problems involving data sets which are large, high-dimensional, and very noisy. Extracting useful information from such data represents a major computational challenge. Traditional statistics is inappropriate to understand computation with massive data sets: it offers algorithms which require computation time scaling exponentially with the amount of data and its dimension. On the other hand, a recent revolution in machine learning demonstrates that some high-dimensional statistics problems have computationally-efficient algorithms (i.e. consuming reasonable computational resources): enormous, noisy data sets are key to recent extraordinary progress in machine learning.

The traditional theory of efficient algorithms cannot explain these new methods in high-dimensional statistics, nor does it suggest the correct algorithms to use for such problems. In fact, it incorrectly predicts that many statistical tasks cannot be solved by efficient algorithms. The trouble is that it only analyzes algorithms which must succeed *in the worst case:* when the devil chooses the parameters of the problem to be solved as disadvantageously as possible. Nature, unlike the devil, often behaves *randomly* but rarely *adversarially.*

My work develops a theory of computationally-efficient algorithms for large-scale statistical problems, to understand which are solvable by efficient algorithms, which are not, and why the border between efficiently and inefficiently solvable lies where it does. I propose to tackle the challenge by developing *optimal meta-algorithms*—generic algorithmic strategies which capture the best efficient algorithms for many statistical problems at once[1]—and using these meta-algorithms to solve new challenging statistical tasks and as a lens to understand why some problems remain infeasible to solve with limited computational resources even when nature, not the devil, generates the data.

The resulting theory will offer key advantages now enjoyed by practitioners using big data sets and by the worst-case theory of algorithms, but lacking today in theory of algorithms for statistical data. At present, algorithms with rigorous analyses for statistical problems are usually invented on a problem-by-problem basis; each problem requires a new algorithm and a new analysis. By contrast, practitioners commonly use successful generic techniques (examples include Markov-chain Monte-Carlo/Metropolis and stochastic gradient descent) which transfer from one problem to the next. While these are generally unsuitable for rigorous analysis, meta-algorithms for theorists will come with a mathematical toolkit; specializing meta-algorithms to solve particular problems offers a means to transfer mathematical insights from one problem to the next.

The worst-case theory of algorithms has identified algorithm-independent structures—such as *convexity* and *NP-completeness*—which explain why some problems are solvable by efficient algorithms and some are not. By contrast, in statistical settings today often the best explanation for non-existence of an algorithm for particular problem is just that we have not found one yet. Early progress on theoretical meta-algorithms for statistical problems already demonstrates that the meta-algorithms lens can identify both the structures that make efficient algorithms possible and those which make them impossible, even though such structures are necessarily quite different from their worst-case counterparts. This offers the first broadly-applicable means to predict and explain which statistics problems allow efficient algorithms and which do not.

I have investigated the *sum of squares method* (and refinements thereof) as a meta-algorithm for *planted problems.*[2] Planted problems are a special class of statistical problems in which nature is as random (hence as benevolent) as possible: one assumes that the structure to be found in the data and the noise which obscures it follow known and well-behaved distributions. Planted problems are nonetheless sufficiently broad to capture many nontrivial statistical tasks. Some can be solved by surprising and sophisticated algorithms, while others seem not to allow efficient solutions. Beyond developing many new algorithms for nontrivial statistical problems (for example in community detection, graph clustering, and factor analysis) by application of the sum of squares meta-algorithm, my work has identified a new and simple structural feature of any planted problem which allows an efficient algorithm. This leads to a new classification of hard and easy planted problems via their low-degree Fourier spectrum, and the first broadly-applicable and rigorous explanation of the structure underlying computationally-hard planted problems.

Planted problems make strong assumptions—e.g. that noise is normal and samples are truly independent—which preclude studying algorithmic consequences of systematic biases common in statistics. Such biases might arise from corrupted data, unknown covariates, model misspecification, or lack of true independence. Developing a meta-algorithmic theory in the face of such biases would be one of my main projects as a fellow. These settings expose new algorithmic phenomena, and require the theory to explore the ground between the planted and worst-case settings.

Another of my main projects will be to study connections between meta-algorithms using convex programming and Fourier analysis (such as the sum of squares method) and *belief propagation*, a very different family of meta-algorithms proposed by statistical physicists; this has potential to reveal deep connections between algorithms and the geometry of physical phase transitions.

Prasad Raghavendra and Luca Trevisan are experts on the intersection of randomness and algorithms. Both are also experts in the sum of squares method, giving us a foundation for collaboration. Berkeley EECS is also home to many leaders in machine learning with enormous data sets; it is an ideal environment for the research described above.

[1] While it is not obvious that optimal meta-algorithms should exist at all, there is already strong evidence that they do exist for some worst-case algorithms settings, such as combinatorial optimization.

[2] These joint works have appeared in FOCS, STOC, SODA, and COLT; several have been invited to special journal issues.