NSF TRIPODS Phase II NSF 19-604

Collaborative Research: STRIDES: Southeastern Transdisciplinary Research Institute for Data Engineering and Science

The GT&Duke STRIDES Team

January 18, 2020

Summary

[this is from the LOI] We propose creating a new Institute aiming to explore and advance the synergies between mathematics, theoretical computer science (TCS), statistics, and electrical engineering in laying the theoretical foundations for data science. Proposed activities include national interdisciplinary workshops, collaborative cross-disciplinary projects, research labs with participants from all communities, joint seminars, and co-supervision of Ph.D. students and postdocs by multiple mentors and institutes, in order to catalyze and promote a true synergy among all related disciplines. Georgia Tech and Duke University will build on the existing strengths and prominence in all the research areas (mathematics, TCS, statistics, and electrical engineering) and build on synergistic avenues of research. Georgia Tech has renowned faculty in the College of Computing, the College of Engineering, the College of Sciences, and the Scheller College of Business and is considered the preeminent technology university in the Southeast. Duke is known as an institution for its interdisciplinarity, and Duke faculty and Duke faculty have world class expertise and recognition in data science, in addition Duke also has access to rich data repositories, including those of a top medical school and regional hospital as well as a 21st century social science data infrastructure built by the Social Science Research Institute (SSRI). The coupling of the strengths in the technological aspects of data science at Georgia Tech with the liberal arts, social science, and health aspects of data science at Duke offer a rich educational and research foundation in the Southeast. An important component of the proposal is to foster professional development both for junior data scientists as well as domain experts that want to become more involved in data science. Georgia Tech and Duke will develop innovative programs that will reach out to technology centers, medical centers, as well as social science centers with the goal of increasing the interactions between junior data scientists and domain experts. The two institutions will also promote diversity and representation that reflects the community via collaborations with Under-represented Minorities institutions.

Georgia Tech and Duke have already made significant investments. Georgia Tech has a new Interdisciplinary Research Institute on Data Engineering and Science (IDEaS), a new 21-story building (Coda) co-locating data science industry and academia in Midtown Atlanta, and serving as collaborative lead institution to the NSF South Big Data Hub, which spans 16 Southern states and over a hundred partner organizations. Duke University has made considerable investments in data science research and education through the Rhodes Information Initiative at Duke (iiD), initiated in 2013. The iiD has features including labs, team rooms, classrooms, offices, and the 2-story Ahmadieh Family Atrium, all of which are designed to incubate cross-disciplinary interactions among faculty, postdocs, and students. In addition, Duke has a new data science initiative focusing on information and analysis-driven health innovation entitled AI.Health which has a goal of engaging the biotechnology and biomedical industries in the NC Triangle region including the Duke Hospital system with data science expertise and developing the educational framework to train data scientists in this application space. Duke University and Georgia Tech are prominent research institutes in the southeast. Georgia Tech is conveniently located at the midtown of Atlanta, with easy access to public transportation (e.g., MARTA), which routes directly to Atlanta's international airport. The Hartsfield-Jackson Atlanta International Airport serves the largest number of passengers in the world and offers direct flights to most US cities. Many Fortune 500 companies have headquarters in Atlanta. There are many local companies that are data oriented. Duke University is in the Research Triangle in North Carolina with strong local industrial research presence. Both Duke University and Georgia Tech have strengths in all related research communities. The joint institute (by Duke and Georgia Tech) will directly benefit the entire southeast, while also having a positive national impact.

NSF requirement. Project Description, limited to 30 pages total, consists of each of the following topics:

The **intellectual focus of the proposed institute**; the rationale for the proposed institute, its mission and goals, and its expected impact; plans for future growth and resource development; proposed steps toward developing its role as a national resource; and results of prior NSF support of the institute if applicable. This section is not to exceed 20 pages total including results of prior NSF support, which may take up to 5 pages.

A tentative **schedule** of scientific activities, with plans for Year 1 and a provisional schedule for Years 2 and 3.

Plans for human resource development, including the selection and mentoring of a diverse cohort of students and postdoctoral participants, as appropriate, and the selection and involvement of researchers at all career levels.

Plans for outreach and for dissemination of outcomes.

1 Introduction

The Georgia Institute of Technology and the Duke University propose to create **STRIDES:** *the* **S***outheastern* **T***ransdisciplinary* **R***esearch* **I***nstitute for* **D***ata* **E***ngineering and* **S***cience*. STRIDES will integrate research and education in mathematical, statistical, and algorithmic foundations for data science. The initial research topics of focus include **xxx the following will be updated later.** (T1) transcribing data with new models and mathematics, (T2) creating new paradigms for decentralized and scalable inference, and (T3) designing efficient strategies with theoretical guarantees, harnessing the combined perspectives of statistics, optimization, and numerical methods. The research will be carried out in the context of big datasets from multiple domains including biology, design, manufacturing, logistics, and sustainability.

During its lifetime, the institute will develop methods and tools to address high-impact multidisciplinary foundational challenges in data science. The institute will bring together mathematicians, computer scientists, and statisticians to jointly work with domain specialists with data challenges across engineering, science and computing. Data arising from experimental, observational, and/or simulated processes in the natural and social sciences and other areas have created enormous opportunities for understanding the world in which we live. Data science is already a reality in industrial and scientific enterprises and there is ever-increasing demand for both research and training in this field. Virtually every scientific discipline is expected to benefit from advances in theoretical foundations of data science, which we seek to strengthen through the TRIAD institute.

The spectrum of fundamental problems in data science is vast. We broadly categorize them as addressing 1) how the data is collected and interpreted, and 2) how the data is analyzed. To address the former, we propose to study modeling and analysis techniques for data with new features and in nontraditional formats (T1), as well as develop decentralized modeling and processing techniques which would not require the data to be transferred to a data center (T2). To address the latter aspect, we propose to study modeling approaches with optimal statistical and computational tradeoffs, and develop algorithms that can be accelerated, distributed/parallelized, are asynchronous and/or stochastic (T3). Key to fundamental advances in these research topics is to derive theoretical guarantees in both the asymptotic and finite-sample cases.

Intellectual merits. The emergence of massive computational power via cloud computing and supercomputing infrastructure has given theorists an unprecedented opportunity to join the fray of empirical science and make significant impact on applications. TRIAD will be particularly well placed to address the growing challenges in the foundations of data science. TRIAD's intellectual focus is to design and build transdisciplinary research programs that provide an enabling and cross-fertilizing platform of ideas and stakeholders (including theoreticians/scientists from domain sciences and users of technology).

Broader impacts. TRIAD will enrich careers of participants ranging from undergraduate students to senior researchers from around the nation in due time. We will make prudent efforts to reach out to diverse communities including participants from smaller colleges and institutions serving under-represented minorities. TRIAD team will actively engage in outreach through public lectures, curricular materials, press releases and dissemination via social channels.

1.1 Key Functions of the Institute

TRIAD will lead and administer a fully integrated program of Research, Education and Outreach in foundations of data science. The institute's research agenda will inform educational programs and will be driven by transdisciplinary solutions to pressing data science challenges. The institute will facilitate interactive interactions between theorists and practitioners with the aim of bringing fundamentally rigorous and broadly applicable solutions and tools to support data-driven discovery in sciences, engineering, and beyond.

Proposed activities include national interdisciplinary workshops, collaborative cross-disciplinary projects, research labs with participants from all communities, joint seminars, and co-supervision of Ph.D. students and postdocs by multiple mentors, in order to catalyze and promote a true synergy among mathematicians, statisticians, and theoretical and algorithmic computer scientists.

TRIAD is particularly well placed to address the growing challenges in data science. Nowadays, analysis of massive, dynamic, and complex data is an area of great importance in many domains. To support true understanding of what is feasible through data-driven approaches and develop broadly applicable and insightful solutions, it is imperative to establish theoretical foundations of data science. Much of the underlying intellectual foundations underpinning data science lies at the intersection between computer science, statistics, and mathematics. In Phase I, TRIAD's main mode of operation will be focused on creating and operating working groups, organizing national and international workshops, and innovation labs. Participating individuals will include senior, mid-career, and junior faculty members, early career researchers including research scientists and postdoctoral fellows, undergraduate and graduate students, and data science practitioners. All activities will be planned to include interdisciplinary researchers rooted in the three foundational disciplines: mathematics, theoretical computer science, and statistics. TRIAD will rapidly deploy information technology and communication infrastructure so that research findings can be quickly and effectively disseminated, while the research community at large can easily access and comment/critique TRIAD's choice of research programs and topics. The utilization of contemporary cyber-infrastructure is likely to lead to international impact and collaboration, which enhances the institute's ability to establish a solid foundation for data science research. The institute will create an intellectual atmosphere that connects theoreticians, practitioners, and scientists from across the nation and the world on a regular basis. Findings in TRIAD's activities will lead to presentations at major conferences and publications in refereed journals.

TRIAD's programs will enrich careers of participants ranging from undergraduate students to senior researchers from around the nation. Postdoctoral fellows and graduate students are introduced to collaborative research in the proposed activities and through workshops. TRIAD personnel will make prudent efforts to reach out to diverse communities including participants from smaller colleges and institutions serving under-represented minorities. TRIAD will conduct public outreach through public lectures, press releases, and dissemination via internet and social media. TRIAD will draw upon the nationally acclaimed Georgia Tech online degree programs (e.g., online masters programs in computer science and in analytics) and curriculum development efforts in machine learning and data science, so that students across the nation can learn about state-of-the-art and interdisciplinary research topics that are not typically covered within traditional campus courses. For Phase II, we propose additional activities (such as customized workshops) that will combine interactive projects and field trips to acquaint undergraduate and/or high-school students nationwide with data-science related techniques and the themes of the TRIAD's year-long programs.

1.2 University Support and Infrastructure

Georgia Institute of Technology (abbreviated Georgia Tech hereafter) is a world-class research university with extensive expertise in data analytics, statistics, theoretical computer science, operations research and simulation, and mathematics. All of its computing and engineering departments are ranked in the top 10 by the U.S. News & World Report, with over half in the top five. Most of its statistics, optimization, and operations research faculty are housed in the School of Industrial Systems and Engineering, ranked as the top such department in the nation. Georgia Tech is the largest producer of engineering degrees awarded to women and underrepresented minorities. Research and education at Georgia Tech is known for its real-world focus, and strong ties to government and industry.

Georgia Tech has shown very strong commitment to Data Science through several major investments in recent years. In 2016, Georgia Tech launched the interdisciplinary Research Institute (IRI) for Data Engineering and Science (IDEaS), charged with facilitating, nurturing, and promoting data science research and data-driven discovery across campus. Georgia Tech has both on-campus and on-line M.S. program in Analytics, and is launching a Ph.D. program in Machine Learning with a sizable data science component. It is investing in a $375 million, 21-story, 750,000+ sq. ft. building (termed Coda) devoted to data science and high performance computing, along with a 80,000 sq. ft. data center to host large-scale computing and data repositories. Planned for early 2019 occupancy, the building will be equally shared by Georgia Tech and relevant data science industry, promoting academia-industry interaction. The TRIAD institute will be administratively structured within IDEaS (led by co-PIs Aluru and Randall), enabling it to benefit from staff, infrastructure, space, and other resources provided through IDEaS, amplifying the impact of TRIPODS Phase I funding.

Georgia Tech has established itself as a national leader in the data sciences. Since 2015, it is serving as the collaborative-lead institution for the NSF South Big Data Regional Innovation Hub (led by co-PI Aluru), one of four such Hubs established to serve the nation. In this role, it supports regional and national-scale efforts in research, industry adoption, and training activities across 16 Southern states from Delaware to Texas, and Washington D.C. The Hub has over 150 partner organizations drawn from academia, industry, government labs, and non-profits. TRIAD team members have been at the forefront of the data science revolution, involved in the White House, NITRID, NSF, NIH, DOE, and DARPA big data initiatives, as well as funding from key programs such as NSF Bigdata, NIH BD2K, and DARPA XDATA.

Georgia Tech is located in Atlanta, the eighth largest economy in the nation and third among cities with the most Fortune 500 companies [39]. With the world's busiest airport and single-hop reachability to many destinations that Atlanta provides, Georgia Tech is uniquely positioned to serve meeting needs of the research community such as workshops and short courses. The proposed TRIAD institute aligns with the university's strategic plan, is synergistic with many new initiatives, and will be housed in the new Coda building alongside IDEaS and the South Hub.

2 Research Programs – XH: Need to work on this...

We propose research activities along four major threads. *Transcribing data with new models and mathematics* (**Section 2.1**), integrates cutting edge data-analysis techniques with dynamical modeling methodologies. *New paradigms of inference* take into account de-centralized data and scalability of the corresponding algorithms (**Section 2.2**). *Efficient strategies with theoretical guarantees* are needed for both statistical efficiency and computational complexity, exploiting the combined perspectives of statistics, optimization, and numerical methods (**Section 2.3**). *Connections between foundations of data science with its applications in engineering, industry, biology, and other domains* (**Section 2.4**) enhances our arsenal of approaches with practical relevance. Interdisciplinary collaboration between theoretical computer science, mathematics, and statistics is necessary to achieve significant progress in all four threads.

2.1 Modeling the devil in the data: data-inspired mathematical analyses

Data-driven modeling and dynamics, including stochastic dynamical approaches, is naturally connected to optimization and statistical inference, driving transformative changes across both application-specific settings, as well as generating new foundational principles [12, 82, 87, 58]. Advances in foundations will promote valid and efficient methods across a huge range of approaches including reduced order modeling plus sparse sampling techniques, model-constrained optimization, as well as equation-free and multiple scale perspectives. Transdisciplinary foundational connections generate effective measures for model/parameter identification and validation, guiding feedback between modeling and data for model "learning" in complex and uncertain contexts. Stochastic and statistical methods, together with the use of random structures, allow for distributional approaches that avoid over-specified, inflexible, over-sensitive, or over-fit models.

Dynamical stochastic modeling and stochastic optimization: A combination of dynamics, stochastics, and optimization is critical in data-driven mapping of large datasets onto large scale "full" models, reduced order, or statistical models. Optimization combined with forward modeling gives new algorithms for faster model identification and reduced model error. However, in applications with large scale and data errors, powerful ideas *beyond the existing theory* are required. New algorithms connecting scalable statistical methods (Section 2.2) and computations (forward problems), e.g., in stochastic PDE-constrained optimization, exploit the model's and data's structures together with the dynamics of the algorithm, for improved efficiency that address large problems [122]. New techniques developed in specific application areas need to be generalized at a foundational level for large classes of problems.

Data inspired random matrix analysis and algorithms: Large datasets generated through text classification and stock data appear generally in empirical covariance matrices ([68, 67, 60, 81, 5, 71, 70]). The delicate estimation of the spectrum of the covariance matrix for the (noisy) model exposes underlying structures of the (noisy) model. While PCA works reasonably well for the top eigenvalues, new techniques in estimating the whole spectrum (as well as the eigenvectors) are needed, particularly when the spectrum is continuous. We plan approaches based on highly nontrivial optimization procedures that generate efficient and practical algorithms, paving the road for more generalizations beyond covariance structures. Another topic related to Section 2.2 is understanding large structures using local structures that lead to universal objects usually in some limit, as in the quintessential examples of the central limit theorem, eigenvalue distributions of random matrices, longest common subsequences, and coagulation/fragmentation processes.

Predictive Analytics of Massive Streaming Data: Semantically rich streaming graph data captures the changing relationships in social networks, financial transactions, communication, and data transfers. Keeping up with real-time changing situations requires rapidly analyzing detecting, and predicting unusual behavior in these rich and inter-related datasets. Current methods build models based on investigation of historical data, but adapting to new behavior by re-analyzing old data introduces a delay in response. We have novel methods for high-performance analysis of this rich, streaming graph data, providing results as the data changes rather than waiting for forensic analysis. Using our streaming graph framework, STINGER, we can move beyond current approaches to rapidly update quantitatively predictive metrics, forecasting behavior and tracking the quality of past predictions. Rapid feedback assists both analysts and applications reacting to changing situations.

Dynamic network modeling: Real life network data must be connected to key questions in network evolution exposing the influence of topology (structure), interactions, and individual dynamics, reception, and transmission properties of the network elements [1, 6]. Furthermore, many real life networks are dynamic - e.g., brain function plasticity [40], immune system development and function, and close proximity contacts in disease transmission [110] - with short time scales of activity of the individual nodes, so that few of the foundational tools for static networks extend to them. Studying disease transmission requires connecting massive datasets with intrahost networks, social/population networks, and cross-immunoreactivity networks [99] and advanced nonlinear stochastic dynamics. Efficient algorithms for model and parameter selection will rely on dynamics connected to new compression, optimization, matrix analysis and efficient statistical measures. Other analyses seek reduced networks approximating the fundamental characteristics of the whole network, connecting statistical inference and scalability with recently validated isospectral transformations [13].

Other connections to computation, modeling, and applications. We mention briefly additional fundamental areas of mathematics and computations in the expertise of the TRIAD team, beyond those mentioned in the rest of Section 2. *Nonlinear algebraic methods* view models geometrically as algebraic varieties or analytic manifolds, so that optimizing model selection via a natural distance function from the data to these varieties. *Persistent homology* is at the core of Topological data analysis (TDA) [24, 25]. Emerging topics of multi-parameter persistent homology hold promise for issues arising in sampling from moduli spaces. *Image restoration, enhancement, segmentation and compression* are advancing through connections with random structures and optimization intersecting with sensor data, optimal path planning, and segmentation, and connecting to a range of topics in computer science, optimization, computer vision, as well as specific application areas [119, 94, 44].

2.2 New paradigm of inference: decentralized and scalable

Technological advances in sensing, data storage, robotics, and computing have led to a rapid proliferation of large-scale and distributed data, calling for novel statistical and computational approaches – techniques that enable groups of distributed and mobile platforms to reconstruct signals, estimate parameters, navigate tricky geometries, and make decisions with little to no intervention from a distant central controller, commonly with limited communication capacity. On the other hand, existing statistical inference theory was developed more than half a century ago when data was more manageable in size. Consequently it is mostly separated from the contemporary development of applied math (including optimization), pure math (e.g., topological structure and algebraic geometry), and efficient randomization and other novel algorithmic techniques from theoretical computer science. Our objectives in algorithms, strategies, and principles in **distributable and/or scalable statistical inference** are to design and study i) inference algorithms handling data distributed at different locations, not assembled in a central location; and ii) statistical inference methods that are scalable with large size data, taking into account the modern implementation framework, such as parallelization and the interplay between the computation and CPU/disk/internodal communication. We seek practical models, statistical theory, and computationally efficient and provably correct algorithms that can help scientists to conduct more effective distributed and/or scalable data analysis.

Motivation for distributed inference in data science. In many contemporary applications, the volume of data streams makes it unrealistic to store all the data in a centralized location, so that data are often partitioned across multiple servers. Examples include search engine companies with data coming from a large number of locations and each collecting terabytes of data [27], and high volumes of data (like videos) that must be stored distributively [83]. The societal impact of effectively solving distributed inference is vast and applicable to many scenarios – cost and volume to transfer data across a major supply chain company or superstore giant; public health surveillance, such as undertaken by the CDC, where it is a challenge to provide a warning system limiting false alarms, and potential benefits of using health data from hospitals across the nation, restricted by privacy, legal, and propriety concerns. There are many forces at play here keeping data distributed: (1) Scale: difficulties to manipulate on a single machine, (2) Communication cost: data transfers being costly, (3) Privacy: transfer causes privacy concerns, and (4) Security: reduce risks of loss. These factors provide strong motivation for our planned research focus in developing theoretical foundations for distributed and/or scalable statistical inference.

Motivation for scalable methods in both computing and inference. It is now widely recognized that data access and communication costs dominate computational costs, and technological trends indicate the gap is expected to worsen further. Hence, algorithms must be designed to minimize communication costs, measured in time or energy as appropriate. Theoretical study of these algorithms provides provable lower bounds on communication. Second, massive benefits to statistical inference would be realized with algorithms that integrate statistical, optimization and computational perspectives, rather than commonly used off-the-shelf optimization (see Section 2.3). Statistically and computationally optimal inference requires improved understanding of, and adaptability to, the underlying probabilistic (specifically random matrices) and geometric "structures" of high-dimensional data. Team member expertise includes adaptation to sparsity in statistical models with high-dimensional vector-valued parameters (such as linear regression), to low-rank structures in matrix models (such as trace regression), and to the underlying combinatorial structures in statistical models of large networks, all featured in Section 2.1. These methods rely on complexity penalization and other model selection tools that support adaptation to the underlying structures and are often based on convex optimization. Tight probabilistic bounds on the risk of the resulting estimators show their optimality and adaptivity, complemented by their distributional properties. Planned *Research Outcomes include:*

1) Theoretical Guarantees of distributed and scalable statistical estimation, see e.g., [47];

2) New Computation-and-Communication-Efficient Distributed Statistical Methods, taking advantage of existing literature and adapting existing statistical techniques, to get theoretically supported computation-and-communication-efficient distributed statistical estimators;

3) Software to support reproducibility, available in open-source along with related documentation, together with developing a comprehensive validation plan with alternatives;

4) Experimental Studies using extensive simulations to evaluate the properties of the resulting estimators in finite-sample cases, on both synthetic data and real datasets; comparison with other existing state-of-the-art methods will be made;

5) Applications. We will explore/demonstrate the applicability and need of our methods in applications including collaborations with local large corporations, such as UPS, Home Depot, Delta Airlines, etc., who have distributed data and would like to perform several types of aggregated inference, developing candidate prototypical applications for testing and improving new methods. Georgia Tech already has close relationship with these companies, including some of the company innovation centers located on campus.

2.3 Optimization algorithms for inference and learning

Since its beginning, optimization has been recognized as a vital tool to transform raw data into useful knowledge to support decision-making. A variety of data analysis methods, from classical linear regression and maximum likelihood estimation, to the more recent support vector machine [104, 105, 106], total variation minimization [92], metric learning [116], compressed sensing [18, 36] and matrix completion [21, 22, 89], are built upon optimization. Our research efforts in this domain will fall into two main categories: overcoming the challenges to optimization algorithms posed by modern datasets, and using convex optimization as a tool for statistical inference algorithms with provably near-optimal performance guarantees.

Optimization for modern data. Models with relatively small datasets can be routinely solved by off-the-shelf solvers, e.g., those based on interior point methods (IPM) (see, e.g., monographs [8, 10, 11, 85, 90, 91, 93, 103, 114, 120] and the recent survey [84]). However, as discussed in Section 2.2, the unprecedented growth in the size of datasets has presented significant challenges to the design of optimization algorithms.

1) *High dimensionality:* Optimization problems from large-scale data analysis usually have a design dimension (number of variables) of $n \approx 10^4$ or more. For example, in image processing, n (number of pixels) can easily exceed 10^6. In another interesting matrix completion example by Netflix, n (number of unknown reviews) can be as high as 8.6×10^9. For these values of n, the cost of an IPM iteration, being at least quadratic in n, becomes prohibitively large.

2) *Data uncertainty:* Datasets are often viewed as random samples from a certain unknown distribution. To account for such uncertainty, stochastic programming (SP) has been widely used in data analysis. However, traditional SP solution approaches need to scan the entire dataset during each iteration, impractical for big datasets [102].

3) *Structural ambiguity:* For convex programming models, it becomes difficult to extract global, structural information from big datasets, e.g., smoothness and regularity parameters, are important but difficult to extract in current convex programming approaches. In some cases, we do not even know if the model is convex, and all we have is a large collection of simulated objective values, e.g., in bandit learning [9] and nonparametric regression [33].

4) *Distributed data and computation:* As discussed in Section 2.2, traditional central data collection requires agents to submit their private data to a central provider with little control on how the data will be used and potentially incur high set-up/transmission costs. Decentralized optimization provides a viable approach to deal with these data privacy related issues. Another line of research stems from the assumption of sequential execution in traditional large-scale optimization shifting to distributed (parallelized, asynchronous) optimization, while also accounting for communication costs, gives rise to, e.g., identifying "parallelization and communication friendly" structures of convex optimization models along with related algorithmic design and complexity analysis, investigating and achieving limits of performance on various classes of distributed optimization models.

5) *Online computation:* Real-time inference and decisions demand that optimization for inference is integrated directly into the data pipeline. Then "dynamic" optimization is needed, often approached as minimizing regret [46, 95], together with the requirement that algorithms update in a provably efficient manner.

6) *Randomized optimization algorithms:* Many recent developments in randomized algorithms, providing provably efficient optimization routines, were motivated by specialized problems such as low rank approximations [113, 59] and combinatorial flows [101, 80, 69]. Their generalizations led to new ways of randomizing and accelerating core optimization tools such as accelerated gradient descent [72], mirror descent [2], and second-order optimization methods [73]. Combined perspectives for randomized optimization routines from different areas, such as Gibbs sampling, stochastic gradient descent, and randomized pre-conditioners can lead to improved randomized optimization algorithms for general use.

Statistical Inference via Convex Optimization. Section 2.2 motivates the combination of optimization as a discipline with statistical inference to obtain provably near-optimal results. This is in sharp contrast to the traditional use of optimization as a "number crunching" tool for realizing a statistical estimates. For example, reducing statistical inference to convex optimization exploits its built in computational tractability, so that the inference can be implemented numerically in an efficient and scalable fashion. Examples include denoising signals of unknown local structure [50, 51, 45, 86], hypothesis testing [15, 41, 56, 57], change point detection [14, 52, 23], estimating linear, linear-fractional and quadratic forms [34, 55, 53] and signal recovery [19, 20, 37, 35, 54] from indirect observations. Under rather general structural assumptions on the observed data, these inferences and their risk are computed efficiently, in sharp contrast to traditional high-dimensional and non-parametric statistics on near-optimal inferences and risks presented in closed analytic form, with severe restrictions from the viewpoint of applications. Besides further bridging statistics and optimization with applications to sparsity-oriented signal recovery, signal processing in Gaussian observation schemes, Poisson and quantum imaging, change point detection, and many more, we also seek convex relaxations for inherently nonconvex inference problems. Prominent examples include linear inverse problems with sparsity [37, 16, 17] and low-rank constraints [89, 22, 31], and recent *nonlinear* inverse problems contexts including phase retrieval [3], and structural regularization [4]. Convex geometry plays a fundamental role in creating and analyzing fast algorithms for learning certain properties of unknown distributions [88, 109, 108, 107], complementing classical techniques of probability and statistics, with methods and results from convexity theory, such as Dvoretzky theorem, estimates on the marginals of log-concave distributions, and Brunn-Minkowski theory.

Deep learning as a Tool for Algorithm Design. As illustrated above, many large scale data analytics problems are intrinsically hard and complex, making the design of effective and scalable algorithms very challenging. Domain experts have to perform extensive research, and experiment with many trial-and-errors, in order to craft approximation or heuristic schemes that meet the dual goals of effectiveness and scalability [28, 29]. Very often, restricted assumptions about the data need to be made in order for the designed algorithms to work and obtain performance guarantees. Regardless, previous algorithm design paradigms seldom systematically exploit a common trait of real-world problems: instances of the same type of problem are solved repeatedly on a regular basis, differing only in their data [30, 38]. We will aim to develop a deep learning framework for scalable algorithm design based on the idea of embedding steps of an algorithm into nonlinear spaces, and learn these embedded algorithms from problem instances via direct supervision or reinforcement learning (see [78, 61] for current work). In contrast to traditional algorithm design where every step in an algorithm is prescribed by experts, the embedding design will delegate some difficult algorithm choices to deep learning models so as to avoid either large memory requirement, restricted assumption on the data, or limited design space exploration. We will develop efficient procedures to train such embedded algorithms as well as study the theoretical properties of such new (algorithm) design paradigm using concepts such as pseudo-dimensions of the embedded algorithms. Furthermore, we will demonstrate the usefulness of the new paradigm on real world data analytics problems, such as materials discovery problems, social recommendation problems, and combinatorial problems such as vehicle routing and set cover problems.

2.4 Applications in data science related fields

Symbiotically, data science theory informs new algorithms and methodologies, while the constraints of "real-world" applications define new directions for fundamental exploration. Within Georgia Tech the proximity to top researchers integrating inference and learning algorithms to solve real-world problems will strengthen the impact of our foundational work, finding solutions that address practical barriers.

Advanced manufacturing. Modern advancements of sensor technologies, communication networks, and computing power result in dense data-rich environments for manufacturing systems. Foundational data science supports the integration of massive data readily available throughout product realization with a holistic system-level data fusion approach for effective analysis, design, quality control, and performance improvement. This requires (i) handling of rich data streams ; (ii) extracting knowledge about the dynamics driving these systems, and (iii) translating this knowledge to enhance design, analysis, and control. At the center stage of future engineering research is the integration of theories, tools, and techniques from engineering, statistics, computations and optimization, establishing transformative methodologies for solving engineering problems. Research in manufacturing systems will follow the same trend, expanding significantly in (a) Data-driven modeling to capture complexity beyond first-principle based modeling; (b) Integrated design such as design space characterization and high dimensional nonlinear optimization problems critical for product and service realization; (c) Expansion of multistage systems beyond the existing focus on multistage discrete manufacturing processes. Methodologies of [96, 97, 98, 76, 74, 77, 75, 115] are particularly relevant to this proposal.

Learning at the Sensor. Data is created at the sensor, where the physical world is translated into digital. The importance of array processing has re-emerged when many elements are spread over large spatial regions, as in radar, seismic exploration, and underwater acoustics, as well as when dense arrays with 100s-1000s of elements packed into a small aperture as in millimeter wave communications and imaging, ultrasonic acoustics, photonics, and neural probes. Within the large scale signal processing challenges on these arrays is an opportunity: the structure emerging with the increased number of variables. In the service of next-generation array processes, we focus on these structured matrix or tensor estimation problems, for which there are very few unifying principles, and many fundamental problems left unanswered. A second area is motivated by the recent surge of attention from the integrated circuits community in highly parallelized computing architectures, where many simple computational processors can be implemented directly at the sensor and run at extraordinarily low power. Both mathematical and implementation points-of-view are needed for algorithms for highly-parallelized architectures that solve the programs in a highly-distributed manner with little to no centralized control, in contrast to standard distributed optimization literature. The new fundamental results needed in these contexts bridge the fields of optimization, neural nets, and message passing for graphical inference. While having some features in common with the problems of Section 2.2 (large datasets distributed over multiple geographic locations), there are also differences in the decentralized sensor problems, focusing on small- to medium-sized problems in quick succession.

Inferring dynamical properties from biological data. Innovations in high-throughput measurements across scales enable new inferences about interaction from dynamics in biological systems, traditionally studied through dynamic nonlinear forward models. Examples include the relationships between gene regulatory networks and gene expression, ecological community network structure and ecosystem resilience, and social network structure and the spread of infectious disease. As discussed in Section 2.1, these inferences are based in inverse problems integrating dynamic, statistical, and optimization approaches with massive datasets. Prominent examples include sequencing approaches to infer species interactions in a high-diversity microbiome [100], viral parasites of microbial cells [49], and transmission of influenza during airplane flights [110]. Iterative use of model-based inference has the potential to open new opportunities for understanding and control.

3 Institute Activities and Community Involvement

The key goal of TRIAD institute is to not only establish a research program in the focus areas at the foundations of data science described earlier, but also nurture and grow a vibrant nationwide community around them, as well as undertake activities for community benefit and training. Georgia Tech's excellent reputation and its established research and educational programs in statistics, theoretical computer science (TCS), and mathematics, along with our vast networks of existing collaborations, will help accomplish this goal.

The following activities include workshops, innovation labs, short-term visitors, graduate student and post-doc enabled joint research, course development, and a lecture series.

3.1 Innovation Labs/Thinktanks

We propose to organize innovation labs bringing together researchers from TCS, Math, and Stat, taking into account both applied and foundational perspectives, working together to translate applied questions into foundational questions across all areas. The organizers will determine the themes of the innovation labs per the research programs proposed above, and will approach experienced senior researchers in the field with respect to mentoring and supervision. We will utilize the *Idea Lab* framework that has been developed and experimented successfully in other related proposal-fertilization activities. In particular, these innovation labs will take a form similar to study group sessions in industrial problem identification. These events will bring together senior and junior researchers, from different disciplines, to focus on specific data science related questions with the goal of turning the questions into new directions for novel theory and computations, which in turn will advance the foundations of data science. The focus areas are based on themes in data science, and the outcomes of the proposed focus activities can then be translated into subsequent collaboration and co-supervision of PhD and postdoc projects; they will also form the basis for future focused workshops, visitation of external experts, or research proposals. Teams in different areas can work in parallel during the innovation labs. Some groups will likely include industry and government contacts.

Experience. Georgia Tech already has excellent examples of generating think-tank activities. ARC (Algorithms & Randomness Center) was founded to create interdisciplinary expertise tackling cross-disciplinary and real-world problems with experts in foundational needs. GT-MAP (Georgia Tech Mathematics and Applications Portal) facilitates Mathematics as an effective research partner of the broader community at Georgia Tech, and provides a stable entity where researchers from the campus community can present their work and share ideas.

3.2 Short-term visitation

During Phase I, we will approach potential researchers who are interested in spending short periods of time at TRIAD to help enhance the proposed research. Due to budget limitations, we may start with short-term visits, with each researcher staying 1-4 weeks. We will proactively seek opportunities for co-hosting, tapping into other resources for joint sponsorship. We will also encourage and host sabbatical visitors to fully or partially associate with TRIAD.

3.3 Lecture Series in Mathematical Foundations of Data Science

TRIAD will invite high-profile researchers each year to give a series of lectures related to the foundations of data science. Specific examples of distinguished researchers we plan to invite include:

i) Prof. Roman Vershynin (University of Michigan) who has been working for many years on the development of non-asymptotic theory of random matrices and its applications to a number of problems in data science (compressed sensing, covariance estimation, statistics of networks, etc);

ii) Prof. Gabor Lugosi (Pompeu Fabra University, Barcelona, Spain) will be asked to lecture on Elements of Combinatorial Statistics.

iii) Prof. Boaz Klartag (Tel Aviv University, Israel), an expert in convex geometry and analysis, has agreed to give several lectures for a non-expert audience. Galyna Livshyts and Tetali plan to work with the local colleges and help host the lectures at Spelman college in Atlanta.

The institute will throw open these events to other universities/colleges in the area, e.g., Georgia State University, Emory University, Spelman College, as well as webcast the lectures if speakers so permit to reach nationwide audience. Atlanta is home to the largest concentration of colleges and universities in the South, including historically black colleges and universities (HBCUs).

4 Broader Impacts

4.1 Enabling Interdisciplinary Collaboration

One of the main objectives of the cross-disiplinary center will be fostering new collaborations across traditional disciplinary boundaries. Many of the PIs and senior personnel have strong track records for promoting interdisciplinary research and understand the challenges. We also have a lot of experience with data science and the effort required to ensure the theroetical questions pursued are of actual practical significance in the application domains. Educational modules will be integrated into all of TRIAD's activities, such as prefacing all workshops with tutorials and taking extra efforts to enusre their appropriateness for newcomers to various fields. We will provide means for interdisciplinary groups to have regular contact so that research forms a more solid interdisciplinary foundation.

4.2 Communications

An experienced Director of Communications will play a central role in ensuring the success of TRIAD. During Phase I, TRIAD will be supported by IDEaS communication director Jennifer Salazar. She has a record of success working with researchers on large interdisciplinary and multi-institution projects, possess rich experience leading research-related communication campaigns, digital media, and public relations, and is a talented writer. She will work closely with the Executive Leadership team to establish relationships, and to track achievements both internally and externally. Her primary responsibilities will be to proactively remain apprised of important successes, develop a strategic communication plan and product calendar, and drive awareness of project progress both internally across the wider team, and externally to the research community, stakeholders, and general public. We expect sustaining planned Phase II activity will require TRIAD hiring or at least cost-sharing such a position.

4.3 Recruiting students and faculty to TRIAD programs

Georgia Tech's track record with recruiting students to their online and degree programs (e.g., the 25-year old, highly visible, ACO Ph.D. program) assures us that once advertised these courses and programs will be oversubscribed. We will advertise on TRIAD web portal and other Georgia Tech websites as we develop new short courses and webinars. We will actively recruit students at large, including underrepresented minorities and women, and faculty from smaller colleges, to participate in our programs and/or apply for graduate fellowships and assistantships.

4.4 Interaction with domain experts

Due to strong presence of engineering programs and local data science industry, we have an efficient and effective interface to application domains. Domain experts will be made aware of the developments and advances that theoretical foundations of data sciences have to offer. The proposed institute will incorporate deep and frequent interactions between theoreticians and domain experts. Success of theoretical foundations of data sciences will strongly depend on connections between statistical accuracy and quality-of-approximation as a tradeoff of various computational constraints that are imposed by modern computing infrastructure; The fact that TRIAD will also be co-located with the high-performance computing center will enable such connections.

5 Results from prior NSF support

The PIs are each supported by multiple NSF awards during the prior five years. Some reflect existing collaborative strengths among the PIs and senior personnel. We summarize a few.

Xiaoming Huo is supported by DMS-1613152, and DMS-1106940, Achieving spatial adaptation via inconstant penalization: theory and computational strategies, Aug 2011–July 2014, $140,000.

Intellectual Merit. The PI investigated how to achieve adaptive functional estimation when the underlying model has inhomogeneous roughness. Publications that were partially enabled by this project include [118, 117, 62, 7, 111, 64, 63, 65, 66, 121, 112, 48, 32, 79, 48].

Broader Impacts. Huo disseminated the research through multiple external talks. He trained three female Ph.D. students (one graduated and is now an assistant professor).

Srinivas Aluru has been supported by 11 NSF grants, including 6 from CCF. We report on IIS-1247716/1416259, BIGDATA: Genomes Galore - Core Techniques, Libraries, and Domain Specific Languages for High-Throughput DNA Sequencing, Jan 2013–Dec 2017, $2,000,000.

Intellectual Merit: This project is led by the PI in partnership with Stanford and Virginia Tech. The PI's group developed parallel algorithms for a variety of string and graph based index and data structures prevalent in bioinformatics. The work received multiple recognitions: best student paper (Supercomputing 2015); first selected paper by ACM SIGHPC under the scientific reproducibility initiative; and selection as a benchmark for Student Cluster Competition (Supercomputing 2016).

Broader Impacts: Research results are disseminated as software libraries (github.com/ParBLiSS). The PI ran three international workshops to lead community efforts for high-throughput sequence analytics. He also assisted NSF in organizing a U.S.-Japan Big Data PI meeting.

Prasad Tetali was supported by DMS-1407657 (July 2014–June 2018, $288,000), CCF-1415496 and CCF-1415498 (Mar 2014–Feb 2017, $600,000).

Intellectual Merit. The projects funded novel directions in optimal transport and discrete and continuous optimization, inspired by concrete real-world problems [42, 43]. The team modeled industrial challenges arising from current-day needs, and identified the precise algorithmic and optimization tools needed to solve the corresponding issues [26].

Broader Impacts. Tetali and team trained six Ph.D. students (one female) and two postdocs, and co-hosted an Industry Day for theoreticians. The team developed a pilot expert system *Ask Minmax*, to help non-experts diagnose and identify commonly encountered optimization problems.

Schedule

A tentative **schedule** of scientific activities, with plans for Year 1 and a provisional schedule for Years 2 and 3.

[This needs to be developed.]

Plans for human resource development, including the selection and mentoring of a diverse cohort of students and postdoctoral participants, as appropriate, and the selection and involvement of researchers at all career levels.

5.1 Graduate students and Postdocs

In Phase I, TRIAD has budgeted two graduate research assistantships and a partially sponsored postdoctoral fellowship. We believe that collaboration can be more productive through cross-supervision; the funding will foster cross-collaborations that take a future-looking approach to recruitment and research training.

Cross-disciplinary graduate/postdoc co-supervision. The graduate research assistantship will support students to be co-advised across academic units on campus. Supervisors must commit to be involved in co-supervision. Students will work on trans-disciplinary projects. The students learn a new field, bring that knowledge/connection back to home group, and bring new expertise to the other sponsoring group. Fellows with advanced preparation and degrees will be admitted in a postdoc position. The joint advising structure with be similar, with additional responsibilities.

Georgia Tech already has several interdisciplinary graduate programs. A new Ph.D. program in Machine Learning is recently approved and will start running in 2018. Another interdisciplinary Ph.D. program that has thrived for 25 years is the Algorithms, Combinatorics and Optimization (ACO) program, run jointly by the schools of mathematics, computer science and industrial and systems engineering, cutting across the Colleges of Sciences, Computing and Engineering. TRIAD participants have ample experience with cross-disciplinary training. They have a strong track record of mentoring and placement: e.g., Ruta Mehta and Jugal Garg (ARC postdocs, now faculty in CS and OR at UIUC, Illinois), Phillipe Rigollet (Math postdoc, now faculty at MIT), Will Perkins (NSF postdoc, faculty at Birmingham, U.K., moving to University of Waterloo, Canada), Kevin Costello (NSF postdoc, faculty at UC-Riverside); the first batch of NSF-funded IMPACT postdocs Maryam Yashtini and Christina Frederick soon to start faculty appointments at Georgetown and NJIT, respectively; Stas Minsker (Math/Stats student, faculty at USC, LA), Nayantara Bhatnagar, Adam Marcus, Amin Saberi and Emma Cohen (all ACO students, now faculty at Delaware, Princeton and Stanford, and a researcher at the Center for Communications Research in Princeton, respectively).

5.2 Curriculum Development

TRIAD will support the development of transdisciplinary courses for foundations of data science. Justin Romberg and Mark Davenport will design a new data science related machine learning course. Tetali and others will develop “Mathematics of Data Science” at the undergraduate level. Planned topics include fundamentals from high-dimensional subspaces, singular value decomposition, Markov chains, and machine learning. The institute will support activities whose benefits transcend Georgia Tech through development of tutorials, on-line courses, course modules, lecture notes, tutorials, and sharable slide sets.

This line of activities is aligned with many existing efforts on Georgia Tech campus. We describe a select few. Michael Lacey taught a special topics course on “Mathematics of Compressive Sensing” during Fall 2016, which had a large (50+) attendance. Jeff Wu and Arkadi Nemirovski are currently co-teaching a course on topics at the interface of statistics and optimization. Arkadi Nemirovski, Vladimir Koltchinskii, and others are writing a book on optimization methods in statistics. Georgia Tech also has outstanding on-line degree programs hailed as a national model.

Plans for outreach and for dissemination of outcomes.

5.3 Workshops

TRIAD will hold about three workshops each year, along the lines of the three research themes the institute focuses on. The workshops will be national gatherings of junior and senior researchers, early career researchers such as postdocs, and graduate and senior undergraduate students, with additional international participants. These will be week-long events which will include (besides research presentations) tutorials, poster sessions, panels, working groups, open problem sessions, and tea-time gatherings to encourage collaboration among researchers. TRIAD will use advanced information technology to disseminate findings from these workshops online.

Our team has planned a number of initial workshops to hit the ground running as soon as the institute is formed:

1. With the help of the Algorithms & Randomness Center (ARC), we propose a workshop on *Randomness in Data Science and Optimization*, and feature top experts in TCS and optimization. This will be co-organized by Singh, Vempala, Vigoda and Tetali.

2. Koltchinskii, Nemirovski, and Romberg will co-organize a workshop addressing current challenges in optimization algorithms for modern datasets, high-dimensional statistics and non-convex inference problems, bringing together relevant statisticians and experts in continuous optimization and signal processing.

3. Huo and his colleagues (e.g., Le Song) will organize a workshop on decentralized and scalable statistical inference, as well as deep learning related methodologies.

Topics for other workshops will be decided by the TRIAD team, taking community feedback and evolving data science landscape into account for maximum impact. External researchers will be welcomed and involved in both organizing the scientific program of the workshops as well as plan topic areas and activities. By leveraging support from synergistic institutions at Georgia Tech (IDEaS, ARC, South Hub, etc.), we will be able to organize or co-organize more workshops than the nine budgeted. We will also hold one or two workshops focused on education in foundations of data science, vital to creating future workforce in this economically important area.

Prior experience: The PIs have track record of hosting interdisciplinary workshops, some of which have launched new subfields in recent years. ARC hosted several workshops bringing experts in randomized algorithms, MCMC and phase transitions, submodular optimization and network science. Externally, Tetali co-organized a thematic workshop *Graphical Models, Statistical Inference and Algorithms* (GRAMSIA) twice at the NSF-funded math institutes IPAM (January 2012) and IMA (May 2015). Koltchinskii co-organized a workshop at the NSF-funded institute SAMSI in 2014 on *Geometric Aspects of High-Dimensional Inference*. Huo organized similar events at SAMSI, and is currently involved in the forthcoming Joint Statistics Meeting, and a Banff workshop on data science related topics. These workshops have been instrumental in bringing relevant experts to come together and exchange novel ideas and breakthrough algorithms. The themes have evolved along with the frontier topics at the heart of the foundations of data science.

References

[1] Valentin S Afraimovich and Leonid A Bunimovich. Dynamical networks: interplay of topology, interactions and local dynamics. *Nonlinearity*, 20(7):1761, 2007.

[2] Zeyuan Allen-Zhu and Lorenzo Orecchia. Using optimization to break the epsilon barrier: A faster and simpler width-independent algorithm for solving positive linear programs in parallel. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1439–1456. Society for Industrial and Applied Mathematics, 2015.

[3] S. Bahmani and J. Romberg. Phase retrieval meeets statistical learning theory: A flexible convex relaxation. arXiv:1610.04210 (to appear at AISTATS 2017), October 2016.

[4] S. Bahmani and J. Romberg. Anchored regression: Solving random convex equations via convex programming. arXiv:1702.05327, February 2017.

[5] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[6] Yuri Bakhtin and Leonid Bunimovich. The optimal sink and the best source in a markov chain. *Journal of Statistical Physics*, 143(5):943–954, 2011.

[7] K. Bastani, Z. (James) Kong, W. Huang, X. Huo, and Y. Zhou. Fault diagnosis using an enhanced relevance vector machine (RVM) for partially diagnosable multi-stationassembly processes. *IEEE Transactions on Automation Science and Engineering*, 10(1):124–136, January 2013.

[8] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2000.

[9] D. Bergemann and J. Valimaki. Bandit problems. In V. L. Klee, editor, *The New Palgrave Dictionary of Economics*, volume 7 of *Proceedings of Symposia in Pure Mathematics*, pages 1–11. Macmillan Press, Princeton, New Jersey, 2nd edition, 2008.

[10] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*, volume 15 of *Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1994.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[12] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[13] Leonid Bunimovich and Benjamin Webb. Isospectral transformations. *Springer Monographs in Mathematics, Springer, New York*, 2014.

[14] M. Burnashev. On the minimax detection of an imperfectly known signal in a white noise background. *Theory Probab. Appl.*, 24:107–119, 1979.

[15] M. Burnashev. Discrimination of hypotheses for Gaussian measures and a geometric characterization of the Gaussian distribution. *Math. Notes*, 32:757–761, 1982.

[16] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. on Pure and Applied Math.*, 59(8):1207–1223, 2006.

[17] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Proc. Mag.*, 25(2):21–30, March 2008.

[18] E. J. Candés, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, pages 489–509, 2006.

[19] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51:4203–4215, 2006.

[20] E. J. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

[21] E.J. Candes and B. Recht. Exact matrix completion via convex optimization. Manuscript, California Institute of Technology, Pasadena, CA, 2008.

[22] E.J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. Manuscript, California Institute of Technology, Pasadena, CA, 2009.

[23] Y. Cao, V. Guigues, A. Juditsky, A. Nemirovski, and Y. Xie. Change detection via affine and quadratic detectors. https://arxiv.org/pdf/1608.00524.pdf, 2017.

[24] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[25] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete & Computational Geometry*, 42(1):71–93, 2009.

[26] Henrik I Christensen, Arindam Khan, Sebastian Pokutta, and Prasad Tetali. Approximation and online algorithms for multidimensional bin packing: A survey. *Computer Science Review*, 2017.

[27] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, and Peter Hochschild. Spanner: Googles globally distributed database. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 2012.

[28] H. Dai, B. Dai, and L. Song. Discriminative embeddings of latent variable models for structured data. In *International Conference on Machine Learning (ICML)*, 2016.

[29] H. Dai, B. Dai, Y. Zhang, S. Li, and L. Song. Recurrent hidden semi-markov models. In *International Conference on Learning Representations (ICLR)*, 2017.

[30] H. Dai, Y. Wang, R. Trivedi, and L. Song. Recurrent coevolutionary feature embedding processes for recommendation. In *Recsys Workshop on Deep Learning for Recommendation Systems*, 2016.

[31] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Selected Topics in Sig. Proc.*, 10(4):608–622, 2016.

[32] Debraj De, Wen-Zhan Song, Mingsen Xu, Diane Cook, and Xiaoming Huo. FindingHuMo:real-time tracking of motion trajectories from anonymous binary sensing in smart environments. In *The 32nd International Conference on Distributed Computing Systems (ICDCS'12)*, 2012. (acceptance ratio 13%: 71 out of 515).

[33] J. Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260Ã±1281, 2003.

[34] D. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994.

[35] D. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical report, Department of Statistics, Stanford University, 2004.

[36] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, pages 1289–1306, 2006.

[37] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory*, 47(7):2845–2862, 2001.

[38] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Rodriguez, and L. Song. Recurrent marked temporal point process. In *Knowledge Discovery and Data Mining (KDD)*, 2016.

[39] Geolounge. Fortune 1000 companies list. https://www.geolounge.com/fortune-1000-companies-list-2016/, 2016.

[40] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[41] A. Goldenshluger, A. Juditsky, and A. Nemirovski. Hypothesis testing by convex optimization. *Electron. J. Statist.*, 9(2):1645–1712, 2015. https://arxiv.org/pdf/1311.6765.pdf.

[42] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, Yan Shu, and Prasad Tetali. Characterization of a class of weak transport-entropy inequalities on the line. *arXiv preprint arXiv:1509.04202*, 2015.

[43] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *arXiv preprint arXiv:1412.7480*, 2014.

[44] Minh Ha Quang, Sung Ha Kang, and Triet M Le. Image and video colorization using vector-valued reproducing kernel hilbert spaces. *Journal of Mathematical Imaging and Vision*, 37(1):49–65, 2010.

[45] Z. Harchaoui, A. Juditsky, A. Nemirovski, and D. Ostrovsky. Adaptive recovery of signals by convex optimization. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *JMLR: Workshop and Conference Proceedings*, pages 929–955, 2015.

[46] E. Hazan. *Introduction to Online Convex Optimization*. Foundations and Trends in Optimization. NOW, 2015.

[47] Cheng Huang and Xiaoming Huo. A distributed one-step estimator. ArXiv, November 2015.

[48] Xiaoming Huo and Gabor J. Szekely. Fast computing for distance covariance. *Technometrics*, 2015. Accepted.

[49] Luis F. Jover, Justin Romberg, and Joshua S. Weitz. Inferring phage–bacteria infection networks from time-series data. *Royal Society Open Science*, 3(11), 2016.

[50] A. Juditsky and A. Nemirovski. Nonparametric denoising of signals with unknown local structure, I: Oracle inequalities. *Applied and Computational Harmonic Analysis*, 27:157–179, 2009.

[51] A. Juditsky and A. Nemirovski. Nonparametric denoising signals with unknown local structure, II: Nonparametric function recovery. *Applied and Computational Harmonic Analysis*, 29(3):354–367, 2010.

[52] A. Juditsky and A. Nemirovski. On detecting harmonic oscillations. *Bernoulli*, 21(2):1134–1165, 2015.

[53] A. Juditsky and A. Nemirovski. Estimating linear and quadratic forms via indirect observations. https://arxiv.org/pdf/1612.01508.pdf, 2016.

[54] A. Juditsky and A. Nemirovski. Near-optimality of linear recovery in Gaussian observation scheme under $\|\cdot\|_2^2$-loss. Submitted to Annals of Statistics. https://arxiv.org/pdf/1602.01355.pdf, April 2016.

[55] A. B. Juditsky and A. S. Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5a):2278–2300, 2009. https://arxiv.org/pdf/0908.3108.pdf.

[56] AB Juditsky and AS Nemirovski. On sequential hypotheses testing via convex optimization. *Automation and Remote Control*, 76(5):809–825, 2015.

[57] Anatoli Juditsky and Arkadi Nemirovski. Hypothesis testing via affine detectors. *Electronic journal of statistics*, 10(2):2204–2242, 2016.

[58] Sung Ha Kang, Seong Jun Kim, and Haomin Zhou. Path optimization with limited sensing ability. *Journal of Computational Physics*, 299:887–901, 2015.

[59] Ravi Kannan, Santosh Vempala, and David P. Woodruff. Principal component analysis and higher correlations for distributed data. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1040–1057, 2014.

[60] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, pages 2757–2790, 2008.

[61] E. Khalil, P. Le Bodic, L. Song, G. Nemhauser, and B. Dilkina. Learning to branch in mixed integer programming. In *AAAI Conference on Artificial Intelligence*, 2016.

[62] H. Kim and X. Huo. Locally optimal adaptive smoothing splines. *Journal of Nonparametric Statistics*, 24(3):665–680, September 2012.

[63] H.-Y. Kim, X. Huo, and J. Shi. A single interval based classifier. *Annals of Operations Research*, 216(1):307–325, 2014. http://www.springer.com/alert/urltracking.do?id=L471c28fMeb6377Sae0acc1.

[64] Heeyoung Kim and Xiaoming Huo. Optimal sampling and curve interpolation via wavelets. *Applied Mathematics Letters*, 26:774–779, 2013.

[65] Heeyoung Kim and Xiaoming Huo. Asymptotic optimality of a multivariate version of the generalized cross validation in adaptive smoothing splines. *Electronic Journal of Statistics*, 8(0):159–183, 2014.

[66] Heeyoung Kim, Xiaoming Huo, M. Shilling, and H. D. Tran. A lipschitz regularity based statistical model, with applications in coordinate metrology. *IEEE Transactions on Automation Science and Engineering*, 11(2, ITASC7):327–337, April 2014.

[67] Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *arXiv preprint arXiv:1408.4643*, 2014.

[68] Vladimir Koltchinskii and Karim Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *arXiv preprint arXiv:1504.07333*, 2015.

[69] I. Koutis, G. Miller, and R. Peng. Approaching optimality for solving sdd linear systems. *SIAM Journal on Computing*, 43(1):337–354, 2014.

[70] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

[71] Olivier Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139:360–384, 2015.

[72] Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 147–156. IEEE, 2013.

[73] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $o(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 424–433. IEEE, 2014.

[74] J. Li and J. Jin. Optimal sensor allocation by integrating causal models and set-covering algorithms. *IIE Transactions*, 42(8):564–576, 2010.

[75] J. Li, J. Jin, and J. Shi. Causation-based t^2 decomposition for multivariate process monitoring and diagnosis. *Journal of Quality Technology*, 40(1):46–58, 2008.

[76] J. Li and J. Shi. Knowledge discovery from observational data for process control through causal Bayesian networks. *IIE Transactions*, 39(6):681–690, 2007.

[77] K. Liu and J. Shi. Objectives-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a bayesian network. *IIE Transactions*, 45(6):630–643, 2013.

[78] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[79] Y. Lu, X. Huo, and P. Tsiotras. Beamlet-based graph structure for path planning using multiscale information. *IEEE Trans. Automatic Control*, 57(5):1166–1178, 2012.

[80] Aleksander Madry. Navigating central path with electrical flows: From flows to matchings, and back. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 253–262. IEEE, 2013. Available at http://arxiv.org/abs/1307.2205.

[81] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[82] J Michopoulos, P Tsompanopoulou, E Houstis, J Rice, C Farhat, M Lesoinne, and F Lechenault. Design of a data-driven environment for multiphysics and multi-domain applications. *Dynamic Data Driven Applications Systems*, 2003.

[83] Siddharth Mitra, Mayank Agrawal, Amit Yadav, Niklas Carlsson, Derek Eager, and Anirban Mahanti. Characterizing web-based video sharing workloads. *ACM Transactions on the Web*, 5(2), May 2011.

[84] A. Nemirovski and M. Todd. Interior-point methods for optimization. *Acta Numerica*, 17:191–234, 2008.

[85] Y. E. Nesterov and A. Nemirovski. *Interior point Polynomial algorithms in Convex Programming: Theory and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[86] D. Ostrovsky, Z. Harchaoui, A. Juditsky, and A. Nemirovski. Structure-blind signal recovery. In D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, 2016.

[87] Benjamin Peherstorfer and Karen Willcox. Online adaptive model reduction for nonlinear systems via low-rank updates. *SIAM Journal on Scientific Computing*, 37(4):A2123–A2150, 2015.

[88] Luis Rademacher and Santosh Vempala. Testing geometric convexity. *Proc. of the 24th FST&TCS, Chennai*, 2004.

[89] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. Manuscript, California Institute of Technology, Pasadena, CA, 2007. Arxiv preprint arXiv:0706.4138.

[90] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. SIAM, Philadelphia, 2001.

[91] C. Roos, T. Terlaky, and J.-P. Vial. *Theory and Algorithms for Linear Optimization: An Interior Point Approach*. Wiley, 1997.

[92] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259Ã±268, 1992.

[93] R. Saigal, L. Vandenberghe, and H. Wolkowicz. *Handbook of Semidefinite Programming*. Kluwer Academic Publishers, Boston-Dordrecht-London, 2000.

[94] Berta Sandberg, Sung Ha Kang, and Tony F Chan. Unsupervised multiphase segmentation: A phase balancing model. *IEEE transactions on image processing*, 19(1):119–130, 2010.

[95] S. Shalev-Shwartz. *Online learning and online convex optimization*. Foundations and Trends in Machine Learning. NOW, 2011.

[96] J. Shi. In-process quality improvement: Concepts and methodology. NSF report, 1996.

[97] J. Shi. *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*. CRC Press, Taylor & Francis Group, 2006.

[98] J. Shi and S. Zhou. Quality control and improvement for multistage systems: A survey. *IIE Transactions on Quality and Reliability Engineering*, 41:744–753, 2009.

[99] Pavel Skums, Leonid Bunimovich, and Yury Khudyakov. Antigenic cooperation among intrahost hcv variants organized into a complex network of cross-immunoreactivity. *Proceedings of the National Academy of Sciences*, 112(21):6653–6658, 2015.

[100] Richard R. Stein, Vanni Bucci, Nora C. Toussaint, Charlie G. Buffie, Gunnar Rtsch, Eric G. Pamer, Chris Sander, and Joo B. Xavier. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLOS Computational Biology*, 9:1–11, 12 2013.

[101] Shang-Hua Teng. The laplacian paradigm: Emerging algorithms for massive graphs. In *International Conference on Theory and Applications of Models of Computation*, pages 2–14. Springer, 2010.

[102] UCI. Machine learning repository. `http://archive.ics.uci.edu/ml/index.html`.

[103] R. J. Vanderbei. *Linear Programming: Foundations and Extensions*. Springer, New York, 2007.

[104] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[105] V. Vapnik. *Statistical Learning Theory*. Wiley, Hoboken, NJ, 1998.

[106] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:189–217, 1963.

[107] Santosh Vempala. The random projection method. *Handbook of Randomized Computing Combinatorial Optimization*, pages 651–671, 2001.

[108] Santosh Vempala. Learning convex concepts from gaussian distributions with pca. *Proc. of FOCS*, 2010.

[109] Santosh Vempala. A random sampling based algorithm for learning the intersection of half-spaces. *JACM*, 2010.

[110] Hertzberg Vicki and Weiss Howard. Transmission of droplet-mediated respiratory diseases during transcontinental airline flights. Submitted for publication.

[111] C. Wang and X. Huo. Object tracking under low signal-to-noise-ratio with the instantaneous-possible-moving-position model. *Signal Processing*, 93(5):1044–1055, May 2013.

[112] Huizhu Wang, Seong-Hee Kim, Xiaoming Huo, Youngmi Hur, and James Wilson. Monitoring nonlinear profiles adaptively with a wavelet-based distribution-free CUSUM chart. *International Journal of Production Research*, 53(15):4648–4667, August 2015.

[113] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[114] S. Wright. *Primal-dual interior-point methods*. SIAM, Philadelphia, 1997.

[115] C. F. J. Wu and M. S. Hamada. *Experiments: Planning, Analysis, and Optimization*. John Wiley, New York, NY, second edition (716 pages) edition, 2009.

[116] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.

[117] H. Xu, D. Luo, X. Huo, and X. Yang. World expo problem and its mixed integer programming based solution. In *Workshop on Behavior and Social Informatics (BSIUCBCN2013), in conjunction with the 2013 Pacific-Asia Conference on Data Mining and Knowledge Discovery (PAKDD2013)*, Gold Coast, Australia, April 14 2013. acceptance rate 44%: 16 out of 36.

[118] Zhouwang Yang, Huizhi Xie, and Xiaoming Huo. *Perspectives on Big Data Analysis Contemporary Mathematics*, volume 622, chapter Data-driven smoothing can preserve good asymptotic properties, pages 125–139. American Mathematical Society, Providence, RI, 2014.

[119] Maryam Yashtini and Sung Ha Kang. A fast relaxed normal two split method and an effective weighted tv approach for euler's elastica image inpainting. *SIAM Journal on Imaging Sciences*, 9(4):1552–1581, 2016.

[120] Y. Ye. *Interior Point Algorithms: Theory and Analysis*. John Wiley, Hoboken, NJ, 1997.

[121] Yuanyuan Zhang, Renfu Li, Dinggen Li, Yang Hu, and Xiaoming Huo. Stabilization of the stochastic jump diffusion systems by state-feedback control. *Journal of the Franklin Institute*, 351(3):1596–1614, March 2014.

[122] Fang Zhilong, Chia Ying Lee, Curt Da Silva, Felix Herrmann, and Tristan Van Leeuwen. Uncertainty quantification for wavefield reconstruction inversion using a pde-free semidefinite hessian and randomize-then-optimize method. In *SEG Technical Program Expanded Abstracts 2016*, pages 1390–1394. Society of Exploration Geophysicists, 2016.