Workshop on "Theoretical Foundations of Data Science (TFoDS)"

Organizers: P. Drineas and X. Huo

1. Executive Summary

The workshop on "Theoretical Foundations of Data Science (TFoDS): Algorithmic, Mathematical, and Statistical" took place on Thursday, April 28 through Saturday, April 30, 2016, at the Hilton Arlington Hotel in Virginia. The workshop's objective was the identification of important research challenges that strengthen and broaden the mathematical, statistical, and algorithmic foundations of data science; the discussion of potential opportunities for collaboration between these communities; and the investigation of workforce challenges and infrastructure development.

This white paper summarizes the discussions of the workshop and has three major objectives.

- To produce a brief survey of fundamental areas in the emerging discipline of data science where collaboration between computer scientists, mathematicians and statisticians is necessary to achieve significant progress. This white paper will articulate specific examples and point to tools that are potentially needed for such progress.
- To assess how collaboration between computer scientists, mathematicians and statisticians could potentially contribute to workforce development by advancing and transforming the data science research training of Ph.D. students and post-doctoral researchers. This white paper will describe and assess the possibility and relevance of Industry involvement, in view of the fundamental character of these collaborative efforts.
- To suggest different infrastructure modalities that could significantly promote and advance such collaborations. Objectives and reliable measures of success for such enterprises will be considered. Synergy with existing infrastructures, such as Mathematical Sciences Research Institutes and the Big Data Hubs, as well as opportunities for industry involvement, will be discussed.

The workshop was highly interactive. After four keynote presentations by Piotr Indyk, Michael W. Mahoney, Anna Gilbert, and David Donoho, the remainder of workshop was structured around parallel sessions of working groups, followed by panel discussions that summarized and distilled the contents of each working group. At the end of the workshop, each working group put up their discussion summaries online, so that all workshop participants could add comments and thoughts. Subsequently, the organizers and members of the steering committee (which included David Donoho of Stanford University, David Dunson of Duke University, Robert Ghrist of the University of Pennsylvania, Ilse Ipsen of North Carolina State University, Michael W. Mahoney of the University of California Berkeley, Gary Miller of Carnegie-Mellon University, and Xiaotong T. Shen of the University of Minnesota) drafted/commented this white paper, which was then posted for suggestions from a subset of the participants. Feedback from all the aforementioned stages was taken into account in forming the final version of the white paper.

Prior to summarizing the several common themes that emerged from the discussions of the workshop, it should be emphasized that most of the focus of the workshop was on the Theoretical Foundations of Data Science, or TFoDS for short. At a high level, the workshop participants converged that TFoDS broadly refers to the subfield of data science that focuses on its theoretical foundations. Such theoretical foundations are necessary in all aspects of data science, from the generation and collection of data to the analysis and the decision making processes. In light of the breadth of TFoDS implied by the above characterization, the term TFoDS will be repeatedly used in this report without attempting a more detailed and precise definition.

Main common themes that were discussed by the workshop participants are:

- Theoretical Foundations of Data Science (TFoDS) are fundamental for industrial applications and scientific understanding and there is demand for training students in this area. Academic institutions must seriously focus on how to define and shape this emerging field. It is well-known that data science problems do not respect the disciplinary boundaries that academic disciplines often tend to draw around particular domains. Therefore, TFoDS will be intrinsically inter-disciplinary, in the sense that many different scientific domains will need to work together and develop novel theories that transcend disciplinary boundaries. Particular emphasis should be placed in interdisciplinary collaborations between computer scientists, mathematicians, and statisticians, since these three disciplines are at the heart of TFoDS.
- Data science is a broad discipline, covering everything from the experimental design phase and the data collection stage, all the way to the data analysis process and the inevitable decision making phase at the very end of the "data to knowledge to action" paradigm. As such, data science is likely to grow into a new discipline, with TFoDS being at its heart, since theoretical foundations are of paramount importance in the development of any new scientific field.
- TFoDS should have strong interfaces to application domains. Domain experts must be made aware of the development and advances that TFoDS has to offer. This process will require strong communication mechanisms by TFoDS researchers. Algorithms developed in a vacuum for theoretical purposes only will typically fail to take into account the peculiarities and incompleteness properties of real data. Success of TFoDS 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.
- Interdisciplinary collaboration between theoretical computer science, mathematics, and statistics is necessary to achieve significant progress in TFoDS. This report will provide numerous examples later. There are certainly barriers to such collaborations: for example, the lack of a common language between the different disciplines, as well as the rigidity of institutional structures within universities or the potential existence of funding silos within funding agencies such as the NSF. To break such barriers, multiple mechanisms will be discussed later: examples include funding that emphasizes multi-departmental educational initiatives, TFoDS challenge problems with real data applications, etc.
- Data provenance, reproducibility, privacy, and algorithmic fairness are all fundamental topics that TFoDS should actively investigate. Incentivizing researchers to work in the aforementioned areas would help TFoDS make progress and impact beyond academic environments.

This report also articulates core concepts in TFoDS education and discusses ways in which a TFoDS center or institute, funded by an agency such as the NSF, would be well-poised to achieve the above goals. To

date, no major center exists that emphasizes the foundational aspects of data science. Such an institute or center is an absolute necessity that could play an enabling role in merging relevant sub-disciplines of mathematics, statistics, and computer science to achieve major progress and breakthroughs. More specifically, such a center would be critical in nurturing the convergence of foundational efforts in these communities. Such a TFoDS center would have the resources to bring together industry partners, academic researchers, and other interested parties to work together on data science problems having broad societal impact.

The remainder of this white paper is organized as follows. Section 2 introduces Data Science and its Theoretical Foundations, emphasizing that TFoDS is at the intersections of theoretical computer science, mathematics and statistics. Section 3 surveys a few fundamental areas in TFoDS; the discussion in Section 3 is not exhaustive and mostly reflects topics raised by workshop participants. Workforce development is discussed in Section 4, while possible modalities for TFoDS centers are discussed is Section 5. Logistical and attendance details of the workshop are available at the workshop web site (Section 6). Brief conclusions are presented in Section 7.

2. Introduction

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. For example, from addressing fundamental questions in physics and biology to beginning to test long-held social science theories to creating new economic and social modes of interaction, data and the technologies enabling the creation of massive quantities of data promise to transform existing industries and academic research areas as well as enable the creation of new industries and research areas. As a catch-all term, the term "data science" has been used to represent this general area. As such, data science is a blend of old and new. Much of "the new" is driven in response to new technological challenges such as the generation of massive data, and much of "the old" is ideas and methodologies that have been developed in existing domains.

Much of the Theoretical Foundations of Data Science lies at the intersection between computer science, statistics, and mathematics; we will discuss several examples of research topics for TFoDS in Section 3. Each of those disciplines however, has been built around particular ideas and in response to particular problems that existed several generations ago. Thus, building a foundation for modern data science requires rethinking not only how those three foundational areas interact with the data and with each other, but also how each of those areas interact with implementations and applications. For example, historically, computer science and scientific computing have each carved out different use cases, which have led to different formalization of models, questions to consider, and different computational environments (such as single machine versus distributed data centers versus supercomputers). The design requirements of business, internet, and social media applications lead to questions that tend to be very different from those that arise in scientific and medical applications. There are many similarities between these different areas, but there are also many differences. Designing theoretical foundations of data science requires paying appropriate attention both to problems that domain scientists who generate the data have as well as to the computational environments and platforms where computations are to be done.

a. Aspects of Data Science

Data science is already a reality in industrial and scientific enterprises and there is ever-increasing demand from students to get more training in this field. As we already mentioned, data science is often used as a catch-all term to describe models, methods, and processes created in response to the enormous quantities of data being generated. It is a natural human tendency to try to categorize and silo a novel phenomenon, e.g., is it "just statistics," is it "just artificial intelligence," etc. One remarkable aspect of data science is that (as opposed to buzzwords that appear every few years that are of interest to a narrow group) nearly every research community and traditional area identifies with the term. Of course, each area has different interpretations for this concept, which represents an opportunity and a challenge.

Should we think of data science as a new research discipline, which should have its own conferences and journals, or just a different educational take on existing disciplines and ways of applying them? What is new is not the abstraction and computational modeling, but rather the presence at the interface between data and analysis. This task may be similar to data assimilation, but at very different scales and levels of complexity. The search for general principles brings together areas traditionally considered separate, such as high performance computing and statistics. While working at scale is a key part of data science, it is not its only feature. Data science incorporates experiment design, data collection and analysis as an end-to-end process, while each individual aspect could be seen as part of another field. Data science is an intrinsically interdisciplinary field, involving data collection, models and methods, and careful evaluation of tools.

b. The Need for TFoDS

Every scientific field needs its theoretical foundations. While asking the right foundational questions is a slow process and will have to develop over time, there clearly are many important theoretical issues in data science. 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 real impact in applications.

In order for this to happen, TFoDS has to tackle many challenges. For example, it is well-known among data scientists that a large fraction of the total data analysis time is spent in data preparation and preprocessing, which is often ignored in TFoDS. Data preparation and preprocessing pose many intellectually challenging problems that are related to deep mathematical issues that cannot be easily be formalized. It is not "just engineering", but rather a critical part of deploying a model in production. It's the before and after of many glamorous machine learning problems. Quantitatively-inclined disciplines, such as genetics and computational biology, theoretical chemistry and physics, computational social science, etc., often ignore these steps. There are obvious and non-obvious reasons for this. For example, it is sometimes just convenience or laziness that leads one to start with "Let A be a data matrix," but in other cases it has more to do with how problems are formulated and parameterized. For example, in databases, one does not typically want to make inferential claims and instead one wants to perform computations on the data in the database, regardless of how the data were generated, and this leads to focusing on data movement. Alternatively, in theoretical computer science, it is common to formulate computation per se as a function transforming inputs to outputs, ignoring noise characteristics in the data, etc. These are two examples of viewing computations on data from a computational perspective; and thus developing an improved foundational understanding of how computation interacts with the noise

properties in the input data as well as how the output of computation interacts with inference and other downstream goals is also of central importance.

Another important feature of data science is that it is iterative, with a dynamic feedback loop. Typically, the formulations of so-called online algorithms in machine learning are not particularly well-suited to understanding the iterative aspect of data science more generally. Targets can change as more data are acquired; instead of restricting our attention to idealized systems under restrictive assumptions, dynamic data collection is general, heterogeneous, and messy. Latency is an important issue, and there are humans in the loop. TFoDS should investigate how data analysis techniques, user behaviors, and the data collection process fit together.

Predicting the evolution of TFoDS is somewhat akin to predicting the evolution of Theoretical Computer Science as it started to emerge (and separate itself from mathematics) in the 1950's. The rise of Computer Science departments and the strong presence of Theoretical Computer Science within such departments, came from the arrival of a major trail blazer, namely the wide availability of the Personal Computer (PC) and all the important societal implications to which it gave birth, e.g., the creation of new industries. Researchers interested in these topics found it more fruitful to create their own departments than to work within the existing structures provided by, say, electrical engineering and mathematics departments. The proliferation of large datasets seems to be another major "forcing function" that will catalyze the creation of Data Science departments, with TFoDS being a fundamental constituent of such departments.

c. Connections between TFoDS and Theoretical Computer Science (TCS), Statistics, and Mathematics

There was much discussion in the workshop regarding connections between Theoretical Computer Science (TCS) and TFoDS. TCS addresses large data challenges through algorithm design in various models that go beyond the traditional time and space measures of resource usage e.g., dynamic algorithms, streaming algorithms, sublinear algorithms, distributed algorithms in the synchronous and asynchronous models, etc. These areas offer a wealth of open problems with potential impact in practical applications. On the other hand, data mining and scientific computing (e.g., computational meteorology) often involve applications with dynamically involving inputs. It is unclear whether these more applied fields are aware of the work in TCS. Vice versa, TCS can inform and tune existing models, or devise new models that cater to actual applied problems. For example, the streaming model in TCS was originally motivated in the mid-90s by networking applications. As such, it considered empirical statistics on data that were typically discrete, and there have been challenges in extending some of those ideas to matrix-based and graph-based machine learning and data science. As an example of how to bridge this gap and develop methods that are better suited to today's machine learning and data science problems, Randomized Numerical Linear Algebra (RandNLA) uses many of the sketching ideas developed in streaming and other areas and applies statistical ideas to them, as evidenced by IBM's "Rand NLA for Large Scale Data Analysis". (http://tinyurl.com/j5px9ob).

Most algorithms and heuristics in academic research ignore the fact that many applications give rise to data that are noisy and/or incomplete. Instead, they make extremely idealized assumptions about noise and incompleteness. The area of "matrix completion" is designed to remedy incompleteness. However, matrix completion has not yet become a natural part of algorithm design for other concrete problems, and the behavior of existing algorithms is poorly understood when missing data are imputed in some way. Related to this is the problem of data amplification: the training of machine learning algorithms on scarce

data sets. If training data sets cannot be amplified, then employment of supervised methods is of the question. Hence the resort to "ad-hoc" data amplification methods. A theoretical study of amplification methods, for supervised methods such as deep learning, say, represents an important foundational topic.

Since TFoDS is tightly connected to mathematics (perhaps with tighter connection to computational mathematics), a careful analysis is required for the trade-off among statistical accuracy, approximation quality, and computational effort. Presently, algorithms are designed to produce arbitrarily good statistical approximations, in ever faster asymptotic running times. In practice, this is often an overkill, mainly due to the limitations posed by floating-point arithmetic, etc. A promising ansatz for an analysis is the study of algorithms under resource restrictions: Improving the approximation accuracy within a fixed running time, or a fixed accuracy in less time. Unfortunately, quality-of-approximation guarantees are still too coarse, and cannot even provide qualitative guidance in practice. Whether methods perform well in practice or poorly often depends on implicit regularization properties of the embeddings associated with worst-case algorithms. More theoretical work is needed here; it is worth noting that many of the remarks in this paragraph also apply to connections between TFoDS and TCS.

TFoDS is intrinsically related to statistics and probability. Algorithms for data science problems are often designed under the assumption that the entire data set can be accessed in any fashion desired. In practice, however, data access is often severely restricted, to specific types of queries (e.g., egonets or frontier-based methods in large, sparse graph mining) or to environments where communication dominates computation. Of the essence, thus, are sampling algorithms that are constraint-aware. Related to this is work on communication-aware optimization methods in machine learning, and on communication-avoiding algorithms in numerical linear algebra. However, these are only very specific cases. A broad understanding of the trade-off between statistical properties and communication constraints is lacking in algorithm design. There are strong connections to the study of coresets, sampling and sketching methods in RandNLA, graph property testing, the imputation of missing data, and 'frame sampling' in statistics. All the above topics are not addressed adequately by just TCS, just Mathematics, or just Statistics: an interdisciplinary effort among all three fields would be necessary in order to make sufficient progress.

Finally, the huge success of deep learning brings up obvious questions for the foundations of data science, and to what extent a sharper theoretical understanding of deep networks can be developed. This represents an interesting class of problems, and there has been very little foundational work to explain when and why these methods work, as opposed to developing highly-engineered algorithms that sometimes do impressively well in certain applications. Deep learning is being viewed through various lenses: as a probabilistic problem, as a group invariant representation learning problem, as a kernel learning problem, or through computational complexity and approximation theory. There is hope that general tools for understanding deep networks will be applicable to specific learned networks. This, in turn, will promote better understanding of their function (for instance through invariants), and lead to improvements in performance, compression, and security.

3. Fundamental research areas in TFoDS

a. Areas of Interdisciplinary Collaboration

In this section, we identify several core areas of TFoDS that can benefit by a close collaboration between statistics, theoretical computer science, and applied mathematics.

Combinatorial Inference on complex structures. Classical statistical inference aims at inferring parameters that lie in the Euclidean space. The main technical tools are the laws of large numbers and the central limit theorem. However, there are many important inferential problems, where the central object is a combinatorial object: a graph, a set, or a matching. Combinatorial inference is a new area which aims at developing a unified inferential theory for combinatorial objects. More generally, the question has to do with how data are modeled, how computations are efficiently performed, and how one obtains inferential control.

Take graph-based data and inference on that type of data as an example. The first question is: does the graph represent a single data point or information about many data points? For example, is a graph representing the web or the internet of friendship connections on a social network one data point or many? Many algorithms in computer science assume the former, and it is more common in statistics to assume the latter. Deciding appropriate inferential goals and what even makes sense to compute depends on the answer to this question. Importantly, even if one assumes the graph is a single data point, the methods one develops have formal similarities with algorithms developed by thinking of the graph as information about many data points. Importantly, also, is that while the methods have similarities, they are often used in very different ways, and this can have important implications for the usefulness of algorithms developed in one area to be applied in another area as well as for establishing a unified foundation of data science. For example, in theoretical computer science and applied math, researchers have developed very sophisticated theory on spectral graph theory and graph property testing. These research efforts focus more on understanding the computational aspects (e.g., graph sparsification) and theoretical properties. In Statistics, researchers have developed very sophisticated methods on inferring graphical models or random graph models (e.g., stochastic block model). These research efforts focus more on developing valid inferential theory for uncertain assessment and understanding their fundamental information-theoretic limit. It is rare in data science practice that one wants to sparsify a data graph, i.e., the graph sparsification methods are not obviously-useful, but in many cases they can be made to be useful if they are used in ways that are different than they were developed for. While this is a sign of a robust research area, it does present challenges to establishing the foundations of data science. A merging of these efforts would be a very high impact field. Possible interdisciplinary topics include algorithmically-scalable, statistically-principled algorithms for extremely sparse graphs, graphical model property testing, fundamental limits of graph property testing, graph linkage prediction, etc.

Computation-Statistics Tradeoff. Statistics uses data (information) as resources and the goal is to develop inferential procedures to minimize the population risk (or maximize the population entropy) using the data. In contrast, computational science considers time as a resource and the goal is to develop efficient algorithms to solve the computational task with as little CPU time (and memory space) as possible. Traditionally, the statistics community focuses mainly on the inferential aspect, while the theoretical computer science (TCS) and mathematics communities focus more on the computational aspects. A potentially very high impact research area is to integrate both statistics and computation in a united theoretical framework. This could require us to sharply characterize computation using formal computational models. Alternatively, this could require methods to characterize the statistical properties implicit in worst-case algorithms. Treating data as a resource, and performing TCS-style analysis, or understanding at a much finer level the effect of relaxations of intractable statistical methodologies will be important here. Exciting computation models include the Turning machine, convex relaxation hierarchy,

and statistical query models. More collaborations between statistics, computer science, and mathematics are expected to create very fruitful results.

Randomized Numerical Linear Algebra. Randomized numerical linear algebra (RandNLA) plays an essential role in large scale computation. Moreover, it is an excellent example of how to do truly interdisciplinary research in the data science arena. It has its roots in applied mathematics, theoretical computer science and convex analysis, but that strong worst-case theory is not why it has been so impactful, both with implementations and applications. Instead, it has been so impactful since the underlying theory could be modified to fit problem formulations that are of interest to the different scientific communities that perform research or use matrix computations. For example, TCS is interested in worse-case running time and accuracy bounds that, while useful theoretically, do not lead to numerically accurate algorithms or immediately-useful results in applications. Scientific computing researchers design numerically robust algorithms with efficient implementations (such as LAPACK) that are typically deterministic and do not leverage the power of randomization. Using randomization could lead to improved theory as well as implementations that, at least in special cases, beat even highly optimized libraries like LAPACK; additionally, randomization could scale many numerical linear algebra methods to peta-sized matrices. Machine learning researchers and statisticians are typically much more interesting in downstream inferential applications that use matrix methods as a black box; however, these matrix methods often have regularization properties that could actually be beneficial to such downstream applications.

While there has been a large body of work on RandNLA, there are still many interdisciplinary and critical directions that merit investigation. For example, the issue of sparsity in the input matrices has not been properly explored. There is still much work to be done in order to understand the combination of randomization with sparsity patterns that are either arbitrary or follow a particular structure. Additionally, linear equation solvers for systems of equations with Laplacian matrices as input have received much attention in TCS, but their usefulness for scientific computing applications (where many other, perhaps more heuristic, methods for the same task exist) is still to be determined. Finally, element-wise sparsification of large-scale matrices has received considerably less attention than row/column sampling of matrices and many more exciting results remain to be discovered. Many workshop participants recognized RandNLA as a nice model area for true interdisciplinary research between the mathematics, statistics, and computer science.

Signal processing/harmonic analysis on graphs. There has been a surge of interest (both in the signal processing and the harmonic analysis communities) in extending classical notions from computational harmonic analysis, such as wavelets, multiresolution analysis, etc., to graphs/networks, especially large graphs induced from data. This nascent field is relevant to a range of standard data science problems, such as learning on graph structured data, semi-supervised learning (the standard approach is to turn it into a graph problem), community detection/clustering and understanding the structure of large social and technology networks in general. This field is truly interdisciplinary and progress is conditioned on a confluence of ideas from (i) mathematics (abstract convex analysis and harmonic analysis and spectral graph theory), (ii) computer science (strongly and weakly local spectral algorithms, fast algorithms for computing wavelet transforms in this less structured domain plus machine learning for solving actual learning problems), (iii) statistics (manifold-based machine learning, stochastic block-models, statistical analysis of wavelet-based estimation, theoretical analysis of network models, phase transitions, etc.).

Importantly, for many realistic data graphs, low-rank assumptions as well as popular manifold hypotheses are extremely poor approximation of processes generating the data, and instead the data often consist of small pockets of structure embedded in large-scale noise. This is most obvious in the community detection literature, but similar properties hold in many large-scale networks. This presents serious fundamental challenges. For example, it is common to let some parameter go to infinity to establish measure concentration needed for inferential guarantees, but in these cases the data have empirical properties of non-asymptotic behavior. Relatedly, structures of domain interest, e.g., high degree nodes or small clusters around an individual often have empirical signatures that would be flagged by traditional False Discovery Rate methodology. Large sparse data graphs such as large social and information networks are of applied interest, but they are also excellent testbeds for foundational questions, precisely since they so manifestly violate many of the overly simplified assumptions that are typically made by computer scientists, statisticians, and mathematicians.

This field is also related to the following areas: (i) computer science work on distributed paradigms (Map-reduce, Scala, GraphX) for large-scale graph computations, (ii) algebraic multigrid methods for high performance computing, (iii) recent work on convolutions/correlations on graphs in deep learning, (iv) data science problems involving data with combinatorial structure, for example predicting properties of drug molecules.

Nonconvex statistical optimization. Both statistics and optimization are two corner stones of modern data analysis. Though statistics and optimization have a lot of overlap, most of these overlapping topics are related to convex optimization. However, a salient feature of modern data science is non-convexity, examples include neural network and deep learning, reinforcement learning, spectral methods etc. It is very critical to develop the theoretical foundations of nonconvex statistical optimization. This new field lies at the intersection of modern statistics and large-scale optimization. In particular, one may apply the model-based thinking from modern statistics to solve large and complex optimization problems. It is important to develop a rigorous theoretical framework to sharply characterize the interaction between informational and computational complexity.

There are different sorts of non-convexity. For example, while spectral methods are not convex in their most popular presentation, i.e., as vector optimization problems, they are convex when viewed as semidefinite programs. Alternatively, the interaction structure of DNNs, while not naively convex, seem to exhibit some sort of soft convexity. In many scientific data applications, one could exploit this by using a temperature parameter or an annealing schedule, as in simulated annealing. In machine learning applications, low-precision SGD methods perform a similar function. It is difficult to prove TCS and mathematical statistics theorems about this, but it is particularly important. That these methods perform so well for certain applications, and that their theoretical basis is so weak, suggests that this is a ripe area for methodological work.

Combining physical and statistical models. There are two cultures for building models for complex systems: physical models and statistical models. These two cultures are fundamentally different. For example, the physical model simulates the dynamics by expressing them in terms of partial differential equations or stochastic processes that obey physical laws (like conservation of energy). In contrast, the statistical model exploits powerful probability tools (e.g., probabilistic graphical models) to provide an "exploratory" model to fit the data. In contrast to physical models, statistical models do not require that

the true dynamics satisfy the statistical model. Instead, they use the highly regularized "substitute model" to discover hidden structure in the data or make predictions.

A foundational question is whether one is interested in obtaining results that are "better than random" or "approximately right." While the well-known mantra "All models are wrong..." certainly shows that that no model is right, a difficult-to-quantify objective is how wrong is wrong, and when can that still be useful. This distinction is often ignored in applications, and this too leads to confusion in establishing the foundations of data science. For example, in social network analysis, it is common to posit stochastic block models with strong properties, and then the claim is made that the assumptions hold since some prediction in a machine learning pipeline is better than prior work. But, it is well known that combining many results that are better than random but not approximately right, i.e., weak learners that are not strong learners, can lead to strong learners and strong predictive quality. This success of high-quality prediction is of interest if prediction quality is of interest, but it is of much less interest when one wants to interpret the results for some other goals, as is typical for the use of clusters and communities, e.g., in econometric counterfactual analysis, etc.

Traditionally, the applied math community focuses more on the physical model, while the statistics community focuses more on the statistical model. For a lot of new scientific applications, it would be beneficial to integrate these two modeling approaches and get the best of both worlds. While this research area exists under the name "data assimilation", more research is required.

Mixed type and multi-modality data. An important feature of modern data science is that the data are of mixed types. For example, a typical dataset may be aggregated from many data sources, including imaging data, numerical data, graph data, text data, etc. Even though intensive research has been conducted for each specific data type, we are still lacking a unified framework to study the mixed data in a systematic fashion. This field has both theoretical and applied interest. It would benefit from a close collaboration between statistics, theoretical computer science, and Mathematics. Further research would lead to breakthroughs and important progress in science and engineering.

Applied representation theory and non-commutative harmonic analysis. Representation theory is usually considered a part of pure mathematics, but it plays an important role in data science problems in multiple potential aspects, including when we consider (i) the invariance to rotation/translation and scaling in computer vision, object detection, autonomous driving, and so on; (ii) spatial invariances in physical models, e.g., the potential energy functions learned by machine learning techniques in molecular dynamics. The representation theory of the symmetric group is relevant to the statistical analysis of rankings, rank aggregation algorithms, statistical rank tests, voting theory, game theory and certain problems in computational biology. Fourier analysis and optimization on spatial groups such as SO(3), ISO(3), etc., (and their various types of products) has appeared in robotics, especially in planning for multiply articulated robots, e.g., surgical robots. Convolution over symmetry groups, and hence representation theory, is a key ingredient of convolutional neural networks and scattering networks.

The application areas are not just consumers of existing theory, but raise new questions in fundamental research, such as (i) the development of fast Fourier transforms on non-commutative groups and their homogeneous spaces (for the symmetric group a swathe of works has been done, but even there, basic questions still remain); and (ii) the further development of the theory of Fourier analysis on non-compact groups (e.g., SL(2, C) is relevant to computer vision).

Topological Data Analysis (TDA) and Homological Algebra. They are fields that naturally sit at the intersection of computer science, mathematics and statistics, as they combine discrete structures, with computation. They deal with noise and uncertainty, which require probabilistic/statistical analysis.

Security, privacy, and algorithmic fairness. A particularly important foundational issue in data science is the ability to address security and privacy concerns. Answering such concerns in general is a hard, if not impossible task: much will depend on the particular application domain. It is obvious that aggregation of data would be essential to big data research, so TFoDS will have to answer questions that focus on crediting and compensating the sources, as well as guaranteeing that any privacy considerations that the sources might have will be properly addressed and resolved.

A separate issue that was raised and discussed by the workshop participants had to do with the need to avoid biased models and biased decision making; this raises the question of assessing the fairness of an algorithm as well as the potential bias of training data. Indeed, the growing prevalence of algorithmic decision-making - using models built from data to aid in decision - has forced us to ask if these processes suffer from the same biases that human decision-makers do. Indeed, the White House Office of Science and Technology Policy just put out a new report[1] highlighting the many ways in which the use of big data and machine learning can both avoid human biased decision-making but also introduce new kinds of biases that might be even more opaque and difficult to detect. The foundational challenge here is to define what it means for algorithmic decision-making to be biased, and how one might quantify this. Is the algorithm itself biased? Is the data used to train the algorithm biased? How can we design systems that have some notion of algorithmic fairness built into them without compromising the effectiveness of these systems? And how can we assess existing systems to determine the key influencing variables that go into a model.

b. Barriers to Interdisciplinary Collaboration for TFoDS

The workshop participants agreed that considerably more interdisciplinary collaboration is needed to address important challenges for TFoDS. It was noted that while there exist many examples of pairwise collaborations, research combining ideas from all three fields (theoretical computer science, mathematics and statistics) is rare. For example, time series analysis is one example of a predominantly pairwise endeavor involving mathematicians and statisticians; however, even in this area, more collaboration is necessary. TFoDS would greatly benefit from collaborations spanning all three fields.

A common language is important for interdisciplinary collaborations. This language may be built, at least in part, through important subject areas. These same subject areas would also hold an important place in the TFoDS curriculum. To facilitate a common language, targeted NSF solicitations that fund primarily interdisciplinary collaborations will stimulate such activities. More outreach activities (such as online videos, white-papers, nuggets) may help break the language barriers. Tools and software could also greatly benefit all three groups that are fundamental for TFoDS.

Institutional structures within academic institutions are very rigid. Funding that emphasizes multi-departmental research and educational initiatives could alleviate this problem. Additionally, well defined open problems and challenges that span all three disciplines framed in a way that is community agnostic, but can be addressed by methods from different communities could help break such barriers. A complementary option would be to focus on applications and let individual researchers from other areas

[1] https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

figure out the best way to match their experiences to compelling problems that can serve as use cases for TFoDS.

c. TFoDS and Data Provenance and Reliability

Data provenance is of paramount importance in data science. In modern times, the Web, online databases, and data warehouses have made data sharing and transferring much easier than before. Online databases are valuable, especially for researchers without direct access to raw data. Data provenance is the process of tracing and recording the origins of data and their movements between places where data are generated and stored, which is important for data validation and results reproducibility.

Data provenance is challenging in the following aspects:

- Data origin may not be obvious, and it is not always easy to trace the history.
- To date, there is not a developed standard for citing online data and documents (other than published papers) or acknowledging raw data contributors; most citations do not include their DOIs, even if they have one.
- As data are transferred and copied, and as their formats are changed, errors may be introduced to the data or their metadata.
- Querying is expensive or time-consuming within a large data warehouse.
- Data privacy issues arise, including both legal and ethical ones.

The workshop participants discussed the urgent need for efficient, reliable, and cost-effective techniques for data provenance, which could all be addressed and researched within the TFoDS community. Specific suggestions are:

- The development of systems to make data self-traceable and self-maintainable automatically when data are copied or transferred.
- The design of unified formats for citing the origins of online data and documents via machine learning methods to track data. The automatic generation of metadata to automatically discover the original paper/context that the data initially appear.
- The creation of a “GitHub for data” that tracks data versions and the changes applied.
- The development of efficient matching and querying systems to locate datasets in data warehouses based on keywords.
- The design and analysis of fast and accurate querying techniques for large databases.

Federal funding agencies can play a pivotal role towards this effort in the following ways:

- Build and maintain a centralized and trusted system for scientific data warehouses for online data storage, sharing, version tracking; consider joint actions with NIH, which has experience in maintaining large medical and genetic online databases.
- Provide incentives for PIs to share their data through such online databases.
- Fund and support research effort and activities in developing cutting-edge data provenance techniques.
- Fund PIs with cloud computing/storage fees, in addition to computer purchases, so that they can store and analyze very big data in the cloud, for which local storage is impossible.

The data provenance and reliability issues should certainly be included in the planning phases of a center for TFoDS.

d. TFoDS and its Role in Reproducibility for Data Science

In order for data science to become a rigorous science, reproducibility and benchmarks are of paramount important. It is well-recognized that reproducibility takes effort (e.g., to annotate the data and comment the code), and thus there is a need to incentivize researchers to make their work reproducible. In addition, some researchers resist sharing their data or code because they want to protect their ideas and results in order to publish before their colleagues. Federal funding agencies could help by requiring reproducibility sections (perhaps similar to the Data Management Plan that NSF requires) for proposals and annual reports. Additionally, funding agencies could, at the end of a grant, reward with a small supplement PIs whose results were reproducible and verified by other members of the community.

The TFoDS community can encourage reproducibility through its publication mechanisms. For example, several statistical journals have made the archiving of a paper's data and code a requirement for publication. However, for massive datasets this can cause issues about who hosts and covers the cost of hosting such data. A related issue concerns the nature of the massive dataset: is it distributed and should it remain distributed? Finally, reproducibility is hard to achieve as it can be dependent upon the machine, platform, etc., that are used.

Benchmark problems and families of problems have been very valuable in related fields, such as applied mathematics. Yet for data science, we recognize the need for the benchmark problems and datasets to grow and update over time. Static benchmark problems will not suffice for data science; dynamic benchmark problems are needed.

We believe that reproducibility in data science is itself a research topic that deserves attention from the TFoDS community. Federal funding agencies could draw attention to this area of foundational research in data science with a specialized funding opportunity. Lastly, funding agencies should encourage the production of tools that abstract away details of cloud computing and running algorithms across distributed datasets. Such details are barriers to entry and barriers to collaboration, particularly interdisciplinary collaboration.

e. TFoDS and the Future of Big Data

Big Data will undergo fundamental changes in both the ways that they are used as well as in the ways that they are stored and organized. Such changes will have implications for the TFoDS research needed to enable a principled approach to this evolution, an evolution that will be gated by both new technologies as well as new societal norms for the data access and sharing. Such evolutionary examples include:

- **Big data on mobile platforms and in the internet of things.** As with much of computing, we envision that the dominant form of data access will shift from desktops and laptops to smartphones, tablets and other environment facilities such as wall displays. Users will come to expect access to data anytime/anywhere and that data they access is both location-aware and temporally current.
- **Personalization/customization.** Users will themselves be big data creators through a variety of smart devices connected to their bodies, their vehicles, and their homes and offices. Furthermore, they will come to expect customization of the data they access to their specific interests and needs at the time and pace of access.

- **Distributed provenance.** While currently most big data repositories are curated, centralized and owned by a single organization, we envision that in the future big data collections will form in a dynamic fashion out of smaller collections of data owned by many individuals or organizations. This will require new trade-offs to be made between privacy and usability and new ways to decide which data "belong together."

All of these changes pose deep foundational questions. We must develop algorithms that deal with streaming and highly dynamic data much better than we do today. Real-time performance will be critical for interactive applications of big data. Multimodal data will have to be better integrated. Finally, we need new principles and ideas on how to organize distributed and heterogeneous data collections into meaningful aggregates that are profitable to analyze jointly. In other words, relevant data will have to learn how to seek each other out in a sea of potentially irrelevant noise.

4. Workforce Development

a. Core Concepts in Data Science Education

A successful foundational undergraduate and post-graduate data science education program should capitalize on cultural differences between disciplines that intersect the science of collection, conditioning, modeling, compression and analysis of data. These disciplines include physics, biology, chemistry, astronomy, economics, computer science, electrical engineering, mathematics, statistics, etc. Such a program should combine the best practices of training in unprogrammed learning (empirical experimentation labs, Kaggle competitions) and programmed learning (fundamental principles, analytical tools, formal methods, etc.).

There are obvious core areas that underpin data science and that should be part of a successful program. These include linear algebra and optimization (Math), programming languages, data structures and complexity theory (CSE), information theory and signal processing (ECE), probabilistic models, machine learning, and statistical inference (Stat). However, the workshop participants recognized the need for exposure to other areas including methods of experimental validation and ethics.

The obvious challenge is to fit all the core areas into one finite curriculum; this challenge can potentially be met if the core knowledge is re-thought and re-organized and distilled into a hierarchical structure that is feasible and accessible to the data science students. For example, several core courses can be aggregated into one by keeping only essential materials from each course. Data-driven and/or cap-stone and/or lab-driven courses can also cover basic concepts from core areas, while giving students hands-on experience with data.

Workshop participants noted that teaching to students the principles of experimental validation is crucial. Students should learn to not be afraid of failure of an experiment to validate a hypothesis or of the failure of an experiment to yield reproducible results. Recognizing the value of being able to predict failure is one of the most important aspects of experimental data science. This is a statistical question (e.g., computing the p-value of the experimental outcome relative to the model) as well as a mathematical question (verifying the validity of the model for the data). Experimental validation should be taught at all levels of the data science curriculum (from the early exposure to exploratory data analysis to design of experiments to using sophisticated models to test scientific hypotheses) and also be central even in its theoretical foundations.

Teaching ethical behavior in dealing with data is also crucial, but often an overlooked aspect of training in the mathematical and physical sciences. Teaching good ethical behavior in the data sciences includes teaching the concept that it is okay to fail, as discussed in the previous paragraph, in addition to other aspects of ethical behavior in dealing with data. These include the following: first, teaching the value of honest reporting of results as well as the need for clarity of the data and of the procedures used. Second, teaching the consequences of fraud and misrepresentation through case studies. Third, the importance of obtaining subject consent and IRB approvals when human subject data are collected. Fourth, the importance of protecting privacy through anonymization of data, in addition to awareness of limitations of privacy protecting measures, e.g., nefarious inference attacks to re-identify individuals from anonymized data.

Communication skills (both spoken and written) are crucial to data science programs. Data science curricula should include a communications component either in a separate course or sprinkled in many courses including data-driven or capstone courses where regular lab reports are written and regular oral presentations are practiced. Another route to acquiring such skills include consulting courses that are already common in statistics department. Data science programs should include such a course as well.

New data science programs are great opportunities to attract students who do not fit into traditional programs. The new data science courses are often best co-taught by faculty members from different traditional programs. Incentives include new materials in these courses, access to a new group of non-traditional students, human connections with colleagues in other departments that could be co-writers for future data science grants.

NSF and other federal agencies could incentivize or encourage new data science programs by putting out calls for proposals to co-design data science curricula, by masking business data to use in data science courses, by providing platforms for data-sharing. NSF could also give training grants on data science and do out-reach to K-12 teachers and students with data science workshops and development of data science AP course curriculum.

b. Collaboration and Partnerships in Data Science Education

A successful data scientist should be able of independent out-of-the-box creative thinking and have an understanding of the technical capabilities of analytical tools. However, the most impactful data science will result from teams of people from diverse disciplinary backgrounds to identify novel types of data and models that can be used to for scientific, engineering and societal problem solving. Data science students will come from varied backgrounds, and this creates a good environment for teaching teaming skills. Promotion of new non-traditional thinking and novel approaches to problem solving can be jump started by creating curricula for data science programs in which foundational courses are co-taught by methodological faculty from computer science, statistics, and mathematics and applied courses are co-taught by methodologists and domain specialists in different application fields (e.g., astronomy, health, medicine, social science, engineering, transportation, etc.). Furthermore, for workforce development it will be advantageous for academic programs to partner with industry and government. Such partnerships will help define the relevant competencies, identify industry workforce needs, and define deficiencies in the curriculum. They can take the form of data sharing, sponsored challenge competitions and hackathons, or reaching out to faculty for student internships and projects of interest to the sponsor.

c. Ways that a government-funded Center or Institute Can Help Data Science Education

Data science intrinsically involves many disciplines that offer different cultures and environments for learning and teaching different core competencies. Several centers and institutes have been established recently at Michigan, UC Berkeley, Columbia, and Duke, to name a few, that cover niche areas of data science. Some of these, like MIDAS at the University of Michigan, represent a major investment by universities and their sponsors to connect domain scientists and data scientists, enhance large scale computing infrastructure, and/or launch new degree programs in data science. These activities are creating new and exciting "lateral" connections between existing data-related activities on campus. However, to date no major center exists that emphasizes the foundational aspects of data science that can move the field forward. An NSF or government-funded institute or center can play an enabling role in merging relevant sub-disciplines of mathematics, statistics and computer science to achieve major progress and breakthroughs.

More specifically, such a center can help nurture the convergence of foundational efforts in these communities. It can provide the physical space for multi-disciplinary networking and collaborative research. It can enable inter-disciplinary teams to develop new courses and to band together on data science research projects of relevance to society. Center level funding would go beyond the standard individual or small collaborative grants that do not scale to larger and broader efforts. Center level funding can also enable coordinated efforts to develop research-oriented degree programs and non-degree programs in foundational data science. Such a center could serve as a neutral "honest broker," moderating the common data science interests between nominally competing departments, in particular the departments of mathematics, statistics, and computer science. Such a center can enhance networking and cross-fertilization by providing funding to host visitors, seminars, workshops, summer courses, and other regional or national activities.

A government-funded center on foundational data science would have the resources to bring together industry partners as well as academic researchers and to work on data science problems having broader societal impact.

5. TFoDS Centers: Possible Modalities

a. General Ideas on Potential TFoDS Centers

The workshop participants discussed a number of possible models for TFoDS centers, whose main function is to stimulate and facilitate research into data science, with a particular focus on its theoretical foundations. The three main modalities that emerged from the workshop discussions are:

1. Discrete, uniform centers (possibly one up to three of them) that facilitate interactions between the fundamental disciplines that are involved in the foundations of data science. Such centers would organize workshops, host a number of short- and long-term visitors, as well as organize focus groups to address a particular topic of interest in the broad area of data science foundations. The overall structure of such centers could revolve around semester-long or year-long programs and could be similar to the many successful NSF-funded Math research institutes (e.g., MSRI, IPAM, SAMSI, ICERM, etc.). The advantage of such discrete, uniform centers is that they have been tested in the past and are considered a successful modality, promoting research in a particular area. A potential disadvantage is that such sustained collaboration after the end of the semester- or year-long programs is often problematic.

2. Discrete, uniform centers (possibly one up to three of them) that set an agenda and play a leadership role in theory for data science and its theoretical foundations, rather than just facilitating interactions. In this second modality, the center would consist of multiple researchers from diverse areas that are considered fundamental in the development of data science foundations. The involved researchers would spend a significant amount of their time at the proposed center for the duration of the project, collaborating on fundamental problems in the various aspects of TFoDS. A potential advantage of such a center would be a sustained interdisciplinary research collaboration on fundamentals of data science, given the long-term interaction between participants from complementary research domains. A potential disadvantage would be the lack of sustained and consistent involvement of participants that are not located in physical proximity to such centers as well as the fact that examples of such centers in other fields were not identified.
3. A modality favoring heterogeneous centers that might include a coordination center, a center to meet infrastructure needs, and multiple research centers, all charged to coordinate with each other and form an integrated enterprise to tackle problems in TFoDS. This modality could encompass the first two modalities, thus offering significant advantages. However, the necessary collaboration and coordination would be considerably more involved; also, the budgetary needs of such a center would be more significant than in the previous two cases.
4. Virtual centers. Similar to the above, but with a virtual instead of physical presence. There might have to exist a physical coordinating entity that provides some infrastructure for all the activities under the overall center. All other activities (workshops, visitors, programs, focused groups) would be distributed among multiple places and modern technologies would permit to all locations to virtually participate to an event.
5. One final idea that was discussed by the participants had to do with the use of existing (or novel) NSF funding mechanisms and frameworks to fund pilot programs to design the centers by supporting ideas labs and using planning grants to articulate some of the aforementioned modalities. These pilot projects could then be used to determine the appropriate configuration of the data science centers.

While there was no clear consensus in terms of the best way to organize such centers, a number of recommendations emerged. As a matter of fact, the majority of the participants felt that that virtual centers are not conducive to the type of collaboration needed for successful TFoDS centers. Bringing people from different yet complementary disciplines together for relatively long visits (weeks or a few months) seems to be the predominant way of facilitating better collaboration. There was particular interest in the small-group model, as seen for example at institutes like AIM (the American Institute of Mathematics), where groups of two to four investigators could assemble for periods of a week or longer to collaborate on a specific project.

There was certainly consensus among the participants of the workshop that each potential center should be charged with bringing together mathematicians, statisticians, and computer scientists to collaboratively advance the field of data science and its theoretical foundations. It would be the center's responsibility to design and budget for mechanisms that would incentivize, facilitate, and promote such collaborations. Such centers should also leverage existing investments by various government agencies in order to improve their data infrastructure and data repositories, if necessary. One particular example would be a collaboration with the NSF Big Data regional innovation hubs, which might help create (in a cost efficient manner) new data repositories.

Additionally, the participants in this workshop recommended that such centers should not solely become distributors of small-scale funding to various groups through "internal" proposal solicitations. While such a mechanism could potentially be a small part of a center, it should not be a predominant way of distributing funds, since it seems difficult for a center to hold NSF-style and quality panels to judge proposals.

Finally, there was limited discussion regarding how such a center would complement existing funding mechanisms for Big Data, such as the Big Data initiative or the Computational and Data-Enabled Science and Engineering initiative. The participants concluded that such mechanisms are complementary to the objective of a TFoDS center, but cannot substitute it: a TFoDS center would have the ability to catalyze progress in TFoDS much faster than any of the existing mechanisms that typically involve smaller scale projects and much more limited collaborative actions.

b. Promoting Interdisciplinarity in Data Science Centers

The workshop participants also suggested a number of strategies that could potentially be used to promote true interdisciplinarity. Such strategies included (i) pilot grant programs that could serve a seed money to explore preliminary ideas in TFoDS, (ii) training grants in data science with an emphasis on bridging gaps across different areas, (iii) sponsored faculty exchange, with a particular emphasis in mid-career scientists who are interested in getting training and starting collaborations outside their particular area of expertise, (iv) encourage pairwise interactions, especially if not exclusively between researchers in different TFoDS basic areas, (v) involve industrial affiliates in order to guarantee that the participants will cross disciplinary boundaries in order to solve problems in the "real world", (vi) sharing data and participating in competitions like Kaggle or KDD cup, etc. (vii) facilitate joint advising of post-doctoral researchers, who could naturally move from their PhD training domain to a new research area, (viii) encourage and facilitate domain diversity by involving domain scientists who are interested in solving "real data" problems that almost invariably cross disciplinary boundaries.

A number of challenges were also identified by the participants, especially from people involved in the organization and day-to-day management activities of centers similar to the ones envisioned by TFoDS. The first challenge is that it is hard to get domain scientists to come for more than few days, since they run labs where their physical presence is critical (this was based on the experience of SAMSI). The second challenge is that centers might tend to be conservative in terms of the research that they seek to promote and might steer funding away from individuals coming up with transformational ideas (e.g., breakthrough ideas such as the development of the Support Vector Machines might not have been supported by a center). Both challenges should be mitigated by appropriate strategies that a center for TFoDS should design and implement.

We now discuss in more detail a few of the above mechanisms that were particularly appealing to the workshop participants. First of all, the pilot grants program should involve multiple disciplines, provide small funding for initiatives at the heart of TFoDS (yet this funding should be large enough to break down disciplinary barriers). It could involve institutions that are not geographically close, however some initial collaboration involving face-to face contact would probably be necessary. Second, the training grants in data science mechanism should involve more than one of the disciplines that are central in TFoDS; therefore, applications at the University level involving more than one departments and multiple domains should be encouraged. Third, the sponsored faculty exchanges could involve student rotations and

funding mechanisms for joint advising of students; this mechanism should emphasize broader impact to TFoDS and could potentially be quite expensive.

Finally, the participants discussed the need of a cultural change in terms of promotion and tenure (and other reward mechanisms) within academia for interdisciplinary research. There was a clear consensus that true interdisciplinarity should be strongly rewarded at all levels in order to guarantee that silos within and across departments are broken. NSF could contribute to such cultural changes by providing reward mechanisms for individuals who break down such barriers with their research.

c. Interacting with Practitioners and Industry

The workshop participants also discussed three different but highly interconnected modes of interaction between the center and practitioners through research, through education and workforce development, and through broader outreach. The group identified several existing centers that could serve as models for such interactions, as well as existing entities that could complement the proposed centers.

Starting with research, the participants made several suggestions on how to stimulate the research interaction between TFoDS centers and the industry: host industrial partners (that play key roles in the projects) in-house, hire software engineers to help develop quality code, focus on concrete (vertical) application domains that utilize data science tools, e.g., the development of autonomous vehicles, use students as "intermediaries" between the PIs and industrial collaborators, as well as between PIs working in theoretical and more applied disciplines, use competitions (e.g., Netflix challenge, Kaggle) as the means of developing tools to deal with concrete tasks and data sets.

Continuing with education, there was strong consensus that another important mode of interaction between the center and the industry would be through educational programs for students who join the workforce. The participants identified several approaches towards achieving this goal. First, by integrating education with practical training in the industry and/or government. For example, the center could require that involved students (as a part of their curriculum) should complete a practical project pursued in collaboration with an enterprise or a government entity. Second, by rotating students between different units and research groups (theoretical and practical) to help them develop the ability to deal with applied questions and scenarios.

Finally, the web and the Internet provide multiple and essentially unlimited opportunities for broader outreach and dissemination of research products. For established collaborations, videoconferences can reduce the need for physical collocation of participants. Further (and broader) outreach can be achieved through the following means: broadcasting talks on-line (technical as well as targeted towards general audience), publishing blog posts, developing quality technical articles, written by professional journalists, targeted towards a general audience (e.g., articles in Quanta Magazine), and using standard code hosting sites (e.g., Github) to disseminate software.

Some of the aforementioned mechanisms have been implemented in other centers, such as the Alan Turing Institute, the Simons Center for Data Analysis, the Simons institute for Theoretical Computer Science, the Michigan Institute for Data Science (MIDAS), the Statistical and Applied Mathematical Sciences Institute (SAMSI), the Big Data Institute at the Computer Science and Artificial Intelligence Laboratory (CSAIL), the Information Initiative at Duke (iiD), the NSF Big Data Innovation-Hubs, etc.

The workshop participants also discussed how a TFoDS Center could serve as a resource for the scientific community. In computational biology, after the early success of genome sequencing, several labs became the dominant centers in sequencing technology. A data science center could avoid this by opening access and raising the playing field for everyone to "get into the game" instead of a relative few. Workflow systems, data sharing, code sharing – improving and building on existing codes, documentation of pipelines can increase the participation of researchers in data science. Finally, to catalyze collaborative efforts within the data science center, a competition model in a way similar to industrial challenges and prizes was proposed.

6. Workshop web site

A list of participants, workshop agenda, and a detailed workshop schedule, is available at:

http://www.cs.rpi.edu/TFoDS/

7. Conclusions

Theoretical Foundations of Data Science (TFoDS) broadly refers to the subfield of data science that focuses on its theoretical foundations. These theoretical foundations are necessary in all aspects of data science, from the generation and collection of data to the analysis and the decision making processes. TFoDS will be intrinsically inter-disciplinary; data science problems do not respect the disciplinary boundaries that academic disciplines often tend to draw around particular domains. There are numerous fundamental areas in the emerging discipline of data science where collaboration between computer scientists, mathematicians and statisticians is necessary to achieve significant progress. Data science is likely to grow into a new discipline, with TFoDS being at its heart, since theoretical foundations are of paramount importance in the development of any new scientific field. TFoDS should have strong interfaces to application domains. A TFoDS center or institute, funded by an agency such as the NSF, would be well-poised to enable and stimulate the formation and growth of TFoDS.