REPORT ON THE COMMENTS RECEIVED BY THE INSTITUTE OF MATHEMATICAL STATISTICS IN RESPONSE TO THE REQUEST FOR INPUT ON THE STRUCTURE OF NATIONAL SCIENCE FOUNDATION TO SUPPORT RESEARCH IN STATISTICAL SCIENCES

1. The IMS Community

The Institute of Mathematical Statistics (IMS) is an international society largely concerned with fostering research on the theory and applications of probability and statistics. The IMS has over 4,500 members, approximately 60% of whom are based in the United States. The IMS publishes five leading journals covering a broad range of research in the theory, methodology, and applications of probability and statistics and it co-sponsors a number of others. In addition, IMS sponsors or cosponsors numerous research conferences and workshops in the U.S. and abroad, and is especially active in assisting new researchers as they transition from graduate study to research-related career activities.

The IMS was founded in 1935 to provide a more suitable home for mathematically based statistical research than was available at the time within the American Statistical Association (ASA), and in particular to support the publication of research in statistics and probability. The IMS maintains strong links to the ASA through various cooperative activities, and also conducts joint activities with several other societies, including the Applied Probability Society of INFORMS, the Bernoulli Society for Mathematical Statistics and Probability, the International Biometric Society, the International Statistical Institute, the International Society for Bayesian Analysis, the Statistical Society of Canada.

IMS recently became a member of the International Council for Industrial and Applied Mathematics to further connections between probability, statistics and applied mathematics. IMS has been trying to strengthen those connections because of its recognition that many interdisciplinary projects require a spectrum of quantitative methods, drawing from applied mathematics, probability and statistics, and that other sciences really need a way of connecting with them all.

Most IMS members are employed in academia, and since the Institute represents both statistics and probability, members can be found in departments of computer science, mathematics, and statistics, as well as in other departments and faculties where statistics and probability play significant roles. Indeed, IMS members receive funding from the full spectrum of NSF directorates and divisions, involving a diverse array of statistical theory and methodology, as well as application, and engage in virtually all of the interdisciplinary activities supported by NSF.

2. Request from NSF

The IMS and other professional societies were asked by the Directorate for Mathematical and Physical Sciences for input to a Committee set up to " ... work in concert with a cross-foundation NSF working group, and various communities and societies, to deliver recommendations concerning the statistical sciences at NSF to the MPS Advisory Committee and other NSF advisory committees in Spring 2013." The committee asked for input in response to the following five questions:

(1) What should NSF do to further promote and facilitate the appropriate development of this multidisciplinary field of data science? Are there management structures that should be considered?
(2) Is research support in statistics, data science or related fields not requested from NSF because it lacks a home? If so, what might be a possible remedy?
(3) Are there complex or massive data problems that might be amenable to joint attack by several different disciplines?
(4) Are there disciplinary areas that could benefit from data science methodologies that are already being employed in other areas?
(5) Are you aware of simultaneous development of data science methods for different fields that might benefit from cross fertilization?

3. IMS Member's Response

In response to an e-mail request to the membership, reproduced here in Appendix A, 41 members submitted comments on the five questions listed above as well as on the broader set of circumstances surrounding the state of probability and statistics and the placement of funding for them within the NSF management infrastructure. There were many more responses on the structural issues relating to the placement of statistics within NSF from senior members of the IMS community, and the responses from the more junior members tended to focus on specific areas of application and interdisciplinary collaboration. Many respondents, both junior and senior, noted that statistics is not only a mathematical science, and that the ways of judging the scientific contributions of statistical work differ substantially from the ways of judging mathematical research. Mathematics is, of course, used in a great deal of statistical research, but it is neither necessary or sufficient in many instances.

As a prelude to our organization of the submitted comments, we note that the challenges associated with the placement of statistics and probability at the NSF mirror in some ways the placement of these disciplines within American universities. Many existing statistics groups and departments grew out of mathematics departments, based on the need for statistical scientists to have a home separate from mathematics.[1] In other instances, statistics units and departments were directly linked to a field of application such as agriculture, business and management, economics, or public health, or to a mix of these and other

[1] Much of this history is chronicled in a recently released volume: Alan Agresti and Xiao-Li Meng (eds.) (2013) *Strength in Numbers: The Rising of Academic Statistics Departments in the U.S.* Springer, New York.

disciplines. Thus today one finds statistics departments and/or groups located in business schools, colleges of arts and sciences, colleges of engineering, colleges of humanities and social sciences, colleges of computer science, and often in multiple units simultaneously. By contrast, probability is more often than not a shared activity between departments of statistics and mathematics, though probabilists can also be found in engineering and other units such as business schools. But no matter where statisticians and probabilists are situated, many of them collaborate closely with disciplinary scientists in what has come to be described as interdisciplinary research activities.

In terms of professional identity, statisticians typically associate themselves with the ASA and/or the IMS, as well as regional International Biometric Society groups. On the other hand, many probabilists retain an affiliation with IMS but also link to mathematical societies such as AMS or SIAM. The breadth of IMS itself is exemplified by its newest journal, the Annals of Applied Statistics. Given the foregoing description, it will come as no surprise to anyone familiar with probability and statistics, or with the IMS, to find a diversity of responses on the key questions of the placement of statistics within NSF, and the role of support for statistics viewed more broadly. But with few notable exceptions, the responses the IMS received from its community urged the placement of statistics primarily outside of DMS (though with clear links to DMS in many instances), and with diverse views on the appropriate directorate. Depending on who is making decisions regarding the placement of such a directorate, and whether probability lies within it or within DMS, it would be useful to keep in mind an aphorism associated with the late great statistician John Tukey: “The best is the enemy of the good.” In the context of the NSF, a search for an optimal arrangement for the location of the statistical sciences should not prevent its placement outside of DMS, a development that would parallel the founding of separate departments of statistics in universities that occurred decades ago.

The IMS responses are not simply targeted at increasing, or at least not diminishing, financial support for statistics, although that is a major concern. The strongest impression one takes away from them is an earnest desire within the statistical community to increase the quality of the interdisciplinary work that the NSF supports by bringing to bear the foundations of the field and its systematic ways of thinking about data and inference to work arising in other domains.

3.1. **Summary of Responses to Questions.**

3.1.1. *On a Division of Statistical Sciences.* A number of respondents favored (explicitly or implicitly) the creation of a Division of Statistical Science (DSS) within the NSF that would function as an umbrella for the funding of statistical research. Some felt the DSS should oversee the funding of all statistical research; others felt it should oversee funding of more applied, interdisciplinary research, leaving the funding of more theoretical research to the DMS. Several respondents clearly felt that the evaluation of DMS grant proposals using mathematical criteria was too restrictive, and was counterproductive to interdisciplinary research. The following quotes represent some typical sentiments:

A former IMS Editor, Respondent (18), noted: “[T]he fit of statistics inside mathematical sciences...is sometimes forced and inefficient.”

A former ASA President, Respondent (23), observed: “[W]e are subsumed by the mathematics community (under the current DMS structure)....”

Respondent (21): “I suggest having a stat program for imminent statistical needs, that may not be mathematical needs.”

Respondent (27): “The DMS has been anything but a nurturing home for statistics.”

Respondent (28): “I think one thing that should be pressed is to separate mathematics and statistics.”

Regarding possible restructuring, a Respondent (3), a senior statistician, former IMS President, and member of the NAS, presented one extreme, suggesting a new NSF *Directorate* for Interdisciplinary Sciences that would support work across many fields, including Mathematics, Engineering, and Computer Science. A similar, but less ambitious, suggestion by Respondent (4), was to create a Division of Data Science that would also support work from multiple fields. The individual proposing this hybrid division felt that a division focusing on solely on statistics would be too narrow, and unlikely to receive broad based support. By contrast, another former ASA President, Respondent (20), says that, “[a]t the very least, the NSF should establish a division for statistics research that is separate from mathematics.” This and several other respondents, for example, (33,34), appear to favor the narrower notion of a stand-alone division that is focused primarily on funding statistics research, and that would give statistics its own marquee within the organizational structure of NSF funding.

We expect that many probabilists are comfortable with the existing DMS structure, and are likely to view changes to statistics funding for Big Data, even broadly construed, as being detrimental to probability. This is expressed succinctly by Respondent (1), a senior probabilist and member of the NAS, who says simply “My main concern is that any changes in statistics funding at the NSF not disadvantage support for research in probability.”

The IMS was founded at a time when there was a much closer nexus between statistics and probability, and much of the motivation for research in probability came from statistical problems. A senior probabilist, Respondent (38), notes that “ ... probability is currently enjoying something of a golden age with Fields Medalists such as Okounkov, Smirnov, Tao, Villani and Werner doing some of their research in the subject and with achievements in the discipline being recognized by other major awards such as the Abel and Gauss prizes.”

The IMS still functions remarkably well as an organization that represents the interests of both statisticians and probabilists, but the fields have definitely diverged. The same probabilist, Respondent (38), observes that “ ... much of the work of ... many of the best probabilists in general is motivated by problems from mathematical physics, combinatorics, computer science, biology and finance rather than by statistics as it is traditionally conceived.” The committee believes that the wider mathematical and scientific community is not as aware of this evolving distinction. Thus there is some concern that if a significant portion of funding for statistical research is folded into a program that has “big data” as its primary focus, and if a similarly significant portion of funding for probability research is handled by this same new part of the NSF administrative structure simply because of the historical connection between the two fields, then this may skew funding of probability research towards topics that are neither the most exciting nor the most appropriate.

Among the other oppositions appearing in the discussion are statistics vs. other mathematical sciences, mathematical vs. computational sciences, and mathematical vs. disciplinary sciences.

3.1.2. *Question 1.* **What should NSF do to further promote and facilitate the appropriate development of this multidisciplinary field of data science? Are there management structures that should be considered?**

Many of the responses excerpted in section 3.1.1 above already implicitly or explicitly deal with this question.

Respondent (37) remarks "I don't ... completely grasp the implications of going to the label data science, but it sounds like something that would be positive if done well." Respondent (37) adds that the University of Chicago is following the lead of other Statistics Departments such as Berkeley and CMU as regards to " ... expanding ... in the direction of 'applied and computational mathematics'."

Respondent (3) shares the same concern about the vagueness and perhaps even vacuity of the label *data science*: "Although we are in the age of big data I find the data science designation awkward. After all what self-respecting science is not involved with data?" This respondent argues for the establishment of a new Directorate of Interdisciplinary Sciences or Infrastructure Sciences that would be an umbrella for those fields that " ... generate techniques of organization and analysis which may originate and/or are motivated by subject matter sciences or begin by being freestanding as in the theoretical parts of the fields but are then transferable across subject matter fields. Subsequently the ideas are preserved in more general forms worked on by theoreticians in these sciences, pure mathematicians, theoretical statisticians and computer scientists and then may reappear as novel fields of application present."

Whilst not going as far as Respondent (3) in terms of suggested administrative restructuring, several of the other respondents express similar opinions.

Respondent (17): "Statistics is no more a branch of mathematics than physics is a branch of mathematics."

Respondent (18): " ... statistics is no longer just a branch of mathematics ..." and has become " ... infrastructure for interdisciplinary research."

Respondent (27): "All of chemistry follows from physical principles at the molecular and atomic levels—yet chemistry is not a subfield of physics. We would not leave funding decisions for the core development of chemistry in the hands of physicists—the result would look a lot more like physics than like chemistry."

Respondent (23) points out the inefficiencies and redundancies that are, in part, created by the current disciplinary divisions: "My own work in recent years has overlapped with algorithmic/computational statistics. One consequence of this is that I see snapshot views of some of the work in computer science. It is quite amazing to me how much so-called 'innovative' work in CS is simply re-inventing (badly, too often unfortunately) what statistics did successfully years ago. I've seen 'work' that shows a total unawareness of even elementary statistics." Creating a structure that could more successfully draw upon reviewers with statistical expertise to evaluate proposals that are essentially statistical in nature but have been submitted to programs in other fields could result in savings from not funding work that "re-invents the wheel" by unwittingly developing a pre-existing piece of statistical methodology within the specific context of some discipline.

But some respondents still see the need for the small grants that have been traditional within DMS. Respondent (32) still supports " ... small statistics research grants to qualified and more younger researchers in the field of statistical science along with the major provision to get in to active collaboration with interdisciplinary researchers on an equal footing—not just providing routine data analysis."

3.1.3. *Question 2.* **Is research support in statistics, data science or related fields not requested from NSF because it lacks a home? If so, what might be a possible remedy?**

Respondent (2) who reflects the data orientation of much of IMS's younger membership notes: "I conduct research in several areas where gathering tens of terabytes of data is daily routine. I also deeply care about Statistical Science as our core discipline and I have a fundamental problem about what "Data Science" is. I got into Statistics because it *is* the Data Science, not because I wanted to be in Mathematics and I ask for the respect of people doing real work."

A senior statistician, Respondent (36), claims that the bulk of current big science—big data awards are not supporting the development of new and useful methodology but instead go to " ... proposals that [are] essentially 'here's a data set and here's what we're going to do to it."' This respondent goes on to make the point that the NSF is currently failing to support statistics (and good science) by not requiring that large projects which have a substantial statistical component should involve the active participation of researchers who are aware of the most advanced techniques in the field and can develop appropriate new methodologies as needed. In order to highlight the pitfalls of not following such a policy, Respondent (36) cites the recent Higgs boson papers and asserts " ... that once the multi-billion dollar LHC was built, it's all statistics with nary a statistician or anyone with statistical training in sight." Respondent (36) further states that the crude statistical analyses used resulted in many fewer particles being detected than were actually present in the data and hence required more extensive running of the experiments than was necessary. Insisting that the data rich projects it funds involve personnel from data science is one way that NSF could support the field as an essential piece of the scientific infrastructure without requiring the development of new administrative structures.

Several respondents felt that worthy research was not being funded because of the current structures at NSF. One reason cited several times was the inappropriate application of the "aesthetics" of mathematics to work that drew its strength from different values.

Respondent (27): "Most statisticians are discouraged workers who gave up on DMS long ago because their ideas were not mathematically elegant enough to win grants there. It just wasn't good mathematics."

Similar problems also arise, however, from interdisciplinary work that involves statistics and a "user" discipline being judged too much by the norms of the latter. Of course, much of the evidence in this regard is somewhat anecdotal, but experiences similar to the following were reported by a number of respondents.

A senior statistician with a long track record of DMS funding, Respondent (10), notes: "I do have one example where I applied to a program in the NSF Statistics area that was supposed to support collaborative research. A very famous colleague in biology mentioned a problem to me (after a university committee meeting) and we began to discuss it. I applied to the NSF Stat program (with my colleague) giving a description of the biological background and some ideas for the development of rather complex and quite interesting stochastic models. The idea was a short term grant for me to work on the models and theoretical properties needed to provide statistical analyses while my colleague would do the pilot studies needed to justify a grant in the biological sciences. Unfortunately, the Statistics NSF evaluator apparently felt that the application needed approval by biologists, and sent it there. This was a disaster: without extensive pilot studies the grant was totally inappropriate to be considered as a biology project. It cast very negative aspersions on my colleague, and the negative response from the biologists killed the application in Stat."

There is also some evidence that funding is not being requested from NSF simply because the current structure does not make it clear that certain initiatives will support statistical research.

A former DMS rotator, Respondent (26), observes: "Take NIGMS — 'Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences' — how does a statistician know that this is meant for a statistician? You, of course, have to read the document (go ahead and do it, it will take awhile to get to the word statistics if you don't use a word search). Why can't the title be: math and stat sciences?"

The concern reflected by comments such as this is related to a wider unease about the names of initiatives such as "Big Data" and "Data Science" and the sense in which they may be either construed to be so broad as to be almost meaningless (cf. the above comment by Respondent (3), "After all what self-respecting science is not involved with data?") or, at the other extreme, to be already owned by other communities such as computer science or statistical physics. For example, a former IMS President and Editor, Respondent (41), notes that "new journals such as *EPJ Data Science* often include no articles from what we may describe as members of the statistical community."

Furthermore, there is a sense that funding models appropriate for mathematics don't foster the larger and more expensive research efforts that are appropriate as some parts of statistical data science evolve towards a *modus operandi* that appears more like that of the laboratory sciences.

Yet another former IMS President, Respondent (13): "For example, NSF statistics grants should support post-docs and programmers so that statistics groups could take on high-impact projects and their findings can be turned into software in public domain."

3.1.4. *Question 3.* **Are there complex or massive data problems that might be amenable to joint attack by several different disciplines?**

Respondents who answered this question answered in the affirmative, and many provided examples drawn from their own research, or research initiatives with which they had some familiarity. Target problems and associated fields included:

- astronomy
- genetics (human, animal, and plant)
- genomics (human, animal, and plant)
- imaging
- neuroscience
- proteomics
- record matching for surveys
- remote sensing
- synthetic biology
- web-scraping

3.1.5. *Question 4.* **Are there disciplinary areas that could benefit from data science methodologies that are already being employed in other areas?**

Respondents who answered this question answered in the affirmative. As with Question 3, many provided examples drawn from their own research, or research initiatives with which they had some familiarity. A partial list of disciplinary areas includes:

- astronomy
- biology
- climatology
- genetics
- genomics
- chemometrics
- education
- experimental design
- imaging
- neuroscience
- particle physics
- seismology
- sociology

Associated data science areas and methodologies included compressed sensing, computational mathematics, computer science, economics, mathematics, mathematical finance, and statistics. Many of these have adopted names that reflect the cross-disciplinary aspects of the work: algebraic statistics, astrostatistics, bibliometrics, chemometrics, econometrics envirometrics, machine learning, and psychometrics.

Several respondents stressed the critical need for more statistical input and participation in a number of fields where data analyses are often carried out by disciplinary experts with varying degrees of statistical expertise. These fields included chemistry, climate science, medical imaging, and high energy physics. A number of respondents expressed concerns that computer scientists are viewed by disciplinary scientists as, and indeed sometimes act as, statistical analysts in Big Data applications.

3.1.6. *Question 5.* **Are you aware of simultaneous development of data science methods for different fields that might benefit from cross fertilization?**

This question is open to several interpretations, and evoked a wide variety of responses from respondents. Suggested sets of fields where existing data science methods might benefit from interaction included

- bayesian methods in different fields
- bioinformatics
- climate science
- econometrics
- geographic information sciences and statistical sciences
- health policy and war simulations
- marine biology, seismology, and forestry
- political science
- computational biology, neuroscience, and signal processing

4. Summary

The responses of IMS members to the NSF-posed questions reflected the breadth of its membership. Respondents included many senior probabilists and statisticians, as well as leaders of all of the major statistical societies. With a few notable exceptions, the responses the IMS received from its community urged the placement of statistics primarily outside of DMS (though with clear links to DMS in many instances), and with diverse views on the appropriate directorate.

Respondent (39) observed: "Discussions about changes to NSF funding for statistics and/or statistical sciences invariably lead to (legitimate) concerns among probabilists regarding how such changes would affect funding for probability. In trying to address such concerns, and balance the needs of its diverse membership, it is natural for some within IMS to view initiatives involving titles such as "Big Data" and "Data Science" in the narrow terms of a potential conflict between probability and statistics."

While such conflicts may exist in some cases, our committee and many of the IMS respondents believe that focusing on statistics vs. probability misses the larger and more important issue: How can the statistical sciences, including probability, effectively compete for recognition and funding with computational sciences such as computer science and applied mathematics? A successful strategy for addressing this basic question would have long term benefits for statistics and probability, beyond and including cross-disciplinary initiatives like those associated with big data.

Respectfully submitted,
Steven N. Evans
Stephen E. Fienberg, Chair
Andrew B. Nobel

March 20, 2013.

Appendix A: Message from IMS President Requesting Input from IMS Members

Dear IMS members:

I am writing concerning an important matter for our US members concerning the structure of National Science Foundation (NSF) to support research in statistical sciences. As you will presumably remember, last fall a proposition was made to change the name of the Division of Mathematical Sciences (DMS). This lead to an intense discussion in the community. IMS also collected opinions of members which then were summarized in a report and forwarded to the NSF's Mathematical and Physical Sciences Advisory Committee (MPSAC). This report can be viewed here

http://imstat.org/news/2011/12/27/1325017446389.html

Last summer NSF decided to keep the name of the Division, but MPSAC was asked to form the StatsNSF Committee charged with examining the current structure of support of the statistical sciences within the NSF. The decision letter can be viewed at

http://www.nsf.gov/attachments/124926/public/Response_MPSAC_Subcommittee_Report_on_Name_of_Division_of_Mathematical_Sciences_8-16-2012.pdf

The StatsNSF Committee is co-chaired by Iain Johnstone and Fred Roberts. It has been charged, in part, to provide recommendations for ways to better structure and support statistical sciences across the NSF and recommendations for means of enhancing the role of statistics in data-intensive science as an integral segment of data initiatives being developed within NSF. The Committee has chosen to interpret statistics at NSF broadly, indeed extending it to embrace data science, defined as the science of planning, acquisition, management, analysis of, and inference from data. The full charge to the committee can be viewed here

http://www.nsf.gov/attachments/124926/public/Request_to_form_MPSAC_Subcommittee_StatsNSF_8-15-2012_Final.pdf

IMS and other societies have been asked to collect and summarize comments from their members for the Committee. Although comments on any aspect of its charge are welcome, the Committee is especially interested in hearing comments on questions such as:

(1) What should NSF do to further promote and facilitate the appropriate development of this multidisciplinary field of data science? Are there management structures that should be considered?

(2) Is research support in statistics, data science or related fields not requested from NSF because it lacks a home? If so, what might be a possible remedy?

(3) Are there complex or massive data problems that might be amenable to joint attack by several different disciplines?

(4) Are there disciplinary areas that could benefit from data science methodologies that are already being employed in other areas?

(5) Are you aware of simultaneous development of data science methods for different fields that might benefit from cross fertilization?

The IMS Presidents (current, past and elect) have decided to use a similar procedure as a year ago when we collected opinions about the name-change proposal. That is, we solicit inputs from our members in the US and ask you to send these to statsnsf@imstat.org on or before February 15 at the latest. A small committee consisting of Steve Evans (UC Berkeley), Steve Fienberg (Carnegie Mellon) and Andrew Nobel (UNC Chapel Hill) will review the comments and suggestions received and prepare a summary for the StatsNSF committee which will also be put on the IMS web page. The summary will not reveal the identities of respondents. However, it will be helpful to the committee if, in writing a comment, members indicate their role as an NSF DMS stakeholder. Since the StatsNSF committee would like to receive a preliminary summary by February 1, it would be appreciated if you could send your comments already by January 25.

You can also send your input directly to statsnsf@nsf.gov, but the StatsNSF Committee prefers that societies gather the comments from their members.

The IMS looks forward to receiving your input on this important topic.

Yours sincerely,

Hans R. Kunsch
President
Institute of Mathematical Statistics