SOCIETY for INDUSTRIAL and APPLIED MATHEMATICS

3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
Phone +1-215-382-9800
Fax +1-215-386-7999
www.siam.org • siam@siam.org

March 1, 2013

In response to a request from the StatSNSF committee, on January 8, 2013, SIAM sent an email message to its members in the United States seeking comments related to support the data science, including several specific questions that had been formulated by the committee. The email is appended below. SIAM received fifteen written responses to the request, which are also included below, after removal of identifying information.

The responses came from a very diverse group, including not only academics, but also respondents from industry and government laboratories. The primary disciplines of those responding include mathematics, statistics, computer science, and systems biology. The comments themselves are diverse as well, but a few themes arose repeatedly.

- The most persistent theme is an emphasis on the highly interdisciplinary nature of data science. Several responses suggest offering grants that require the formation of teams involving mathematicians, statisticians, computational scientists, biologists, etc. Several emphasize the importance of better integration of computing and statistics. Various conclusions concerning the support of data science at NSF are drawn from this:

  "Clearly, a solution cannot be found just within the MPS directorate," referring especially to CISE.

  "It would be more effective to maintain, at least for a while, the existing structure but stimulate interaction among the program directors and encourage a much more pro-active attitude toward collaborations across disciplinary boundaries."

  "Small projects that pair say computational and statistical scientists, each moving halfway toward the other domain, should be strongly encouraged."

- Several respondents emphasize the importance of NSF support for educational efforts related to data science. Some view this as the most important potential contribution of NSF to the area. They suggest that NSF promote the simultaneous education of students in applied mathematics, statistics, and numerical computation, for example, by encouraging applied mathematics programs to incorporate more statistical and probabilistic ideas into their curricula.

  "Boundaries between disciplines need to be de-emphasized as much as possible, and the students need to be encouraged to reach out to other disciplines."

- This quote conveys a view shared by several respondents:

  "Data science is rapidly evolving and far from mature, and it is premature to institutionalize it. What needs to be done is to encourage collaborations across disciplinary boundaries within the existing structures."

- Several respondents stress that current developments in data science represent an exceptional scientific opportunity. As one wrote: "Let's not waste this opportunity."

Respectfully submitted,

Irene Fonseca, SIAM President
Douglas Arnold, SIAM President 2009–2010

*Email of 8 January 2013 to U.S. SIAM members*

From: "Statistical Sciences Survey" <statsnsf@siam.org>
To: SIAM members
Subject: NSF asks SIAM members for commentary on statistical sciences
Date: Wed, 8 Jan 2013 09:03:32 -0500

To SIAM members:

SIAM has received a request from the recently formed StatSNSF committee of the National Science Foundation (NSF). The committee 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. It 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 detailed charge can be found at:

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

SIAM and other professional organizations have been asked to collect and summarize comments from our members to 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 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?

Should you wish to have your comments considered, PLEASE SEND THEM TO <statsnsf@siam.org> BY FEBRUARY 1 AS PLAIN TEXT EMAIL WITHOUT ATTACHMENTS.

Email headers will be removed from comments before they are transmitted. Thus, senders' identities will not be known unless they choose to identify themselves in the message body.

Thank you.

Irene Fonseca, SIAM President

To whom it may concern:

I am a member of both SIAM and ASA and have read the communications from the presidents of both societies with great interest. Although the request from SIAM has slightly different language than the request from ASA, using the term "data science" wherever the ASA communication uses "statistical science", I am writing the same response to both requests. I live and work in the Washington, DC area and have been involved with both statistic and mathematical communities through research, consulting, teaching at the undergraduate and graduate level, and several tours as program director at the National Science Foundation.

In my view, the current developments in data science (used broadly, including data collection in various disciplines, computational treatment, statistical analysis, mathematical modeling, and communication and visualization) represent an abundance of scientific opportunities that I have not seen in three decades of professional activity. Here is an opportunity to find new challenging and relevant research problems for all, to attract the best student talent to a field that is suddenly very "hot", and perhaps to bridge some of the gaps which have arisen, in my view often unnecessarily, between mathematics, statistics, and computer science. Let's not waste this opportunity.

Regarding the questions put forth to the community:

1. Perhaps NSF's most important contribution in "further promoting and facilitating the appropriate development of statistical science /data science" could be in graduate education. This is a very exciting field for students. There is also a well identified talent gap. Create and support opportunities where students can learn applied mathematics, statistics, and computing (the emphasis will necessarily fall on the last two fields), with a view towards data science, and the best researchers in these fields will work on topics that can attract these students. The current management structures at NSF allow for supporting statistical and mathematical work from within DMS, but computing is mainly supported in a separate directorate that has a very different culture. A solution cannot just be found within MPS (the directorate of mathematical and physical sciences).

2. I cannot answer the question whether "research support in the statistical sciences / data science is often not requested from NSF because it lacks a home there". I note that the culture of support in statistics is somewhat different, with statisticians often obtaining partial research support from agencies other than NSF (EPA, USGS, NIH, to name a few). With respect to data science, the field is too young to even make such a statement, since, well, there seem to be very few data. So I regard this as a somewhat loaded question.

My responses to the following three questions necessarily have some selection bias and are colored by my own experiences.

3. "Are there complex or massive data problems that might be amenable to joint attack by several disciplines?" Yes, absolutely. All these massive data problems come from real applications. To understand large data sets from astronomy / climate science / marketing, the mathematical scientist needs to be a member of a team that has astronomers / climate scientists / marketing experts. The mathematical scientist may have to give up the intellectual driver's seat for such projects and should accept this. The boundaries between very high-level consulting and original research may be vague and undefined.

— *continued on next page* —

— *continued* —

4. "Disciplinary areas that could benefit from data science / statistical science methodologies": I'm not aware of large gaps of this nature in the scientific landscape. On the other hand, I'm convinced that new and exciting problems for mathematicians and statisticians will continue to emerge from these interactions. An example is the recent progress in recommendation systems. Questions that were originally formulated by marketing analysts have led to general progress in computational linear algebra (low rank matrix completion) and combinatorial optimization. And this is just one example. I'm sure there'll be more in the future.

5. "Simultaneous development of data science / statistical science methods for different fields": Ideas that were developed for recommendation systems also find applications in image processing (inpainting), due to the connection to matrix completion. Again, this is just one example. The EM (expectation maximization) algorithm continues to find new applications and can even influence researcher's thinking about his or her problem area. And all this is only possible due to the continuing progress in computing methods and computing power.

Thank you for creating the opportunity to contribute to this debate.

One upcoming data intensive challenge of immense complexity is storing, analyzing, and searching in data generated by next-generation sequencing (NGS) technologies. These data include sequence reads from single genomes, from metagenomics, and from RNA-Seq. The initial analysis of each NGS data set in isolation is a computational and statistical challenge that we are only barely keeping up with. Storing so many immense data sets in a manner amenable to further analysis and storage is a daunting task. The most important challenge is to be able to search across data sets to glean biologically meaningful information; there are no current projects with the potential to effectively address this challenge.

Success in this challenge will require meaningful collaborations among biologists, computer scientists, and statisticians. Seeking funding for such collaborations always raises the issue of which program in NSF to apply to. NSF generally does a good job of funding multidisciplinary collaborations, but it would help to have programs that eliminate some of the potential confusion of who funds what.

One recurring idea that has increased effective analysis and software is access to distinguished test problems. Statistics research and software development would likely be helped if the major data base vendors defined an interface for access to data bases that they license. NSF could ideally fund this and make the details widely available. That way researchers would see the challenges.

I should first give a little relevant background. When I was an undergraduate my predilection was for pure mathematics. I loved probability theory but I disliked statistics even though I was taught by such luminaries as Barnard and Cox. When I joined IBM research in 1990 after spending 20 years at the University of Waterloo I recall being in a colleague's office (a statistician) and seeing a book on robust regression on his shelf. It was mainly about l_1 minimization, an area I was familiar with from a numerical analysts point of view, and I knew books in the area. The overlap in references was zero! Although the situation may have improved somewhat I still think many similar scenarios still exist. For example, a young person might believe that the data compression folks were the first to discover l_1 minimization.

In the applied area in recent years I have been mainly working on upstream problems in petroleum. An area of current major interest there is dealing with uncertainty and looking at, for example, ensemble methods, data assimilation and stochastics. I see so much nonsense based on tiny sampling and untenable assumptions and again, an apparent ignorance of related existing work. Of course this is mainly not publications by Statisticians but somehow we are not having the right impact on the applied fields [the same issue arises in the influence of optimization mathematicians (like myself) on engineers who are applying optimization]

I always though that too much statistics was based upon what nice things could be proved rather than what assumptions are realistic. I have the impression that today (at least in an industrial research environment) there is more attention paid to the latter. I do not know if the same is true in statistics departments at universities but there were considerable more industrial research environments, that emphasized research rather than development, in 1990 than there is today. Moreover, the importance of statistics in attacking complex problems has increased considerably but I am not sure that the influence of first-class statisticians on the applied fields has increased commensurately.

I am not sure what is meant by data science but using your definition of it as being defined as the science of planning, acquisition, management, analysis of, and inference from data, it would seem to include data mining and this breath only serves to emphasize my lament.

As to 'What should NSF do to further promote and facilitate the appropriate development of this multidisciplinary field of data science? ' my first though would be to have more meetings where the applied practitioners (for example in application in petroleum) engaged more directly with, for example numerical analysts, data miners and statisticians. I would actually like them to take some published papers, being critical about the inadequacies (it could be in both directions). I actually would love to see something similar in optimization. I realise that this might be as unrealistic as assuming that all distributions are Gaussian. I suspect that the most effective solution may be better marketing of statisticians (and similarly by optimizers) by themselves rather than anything that NSF can do directly.

Concerning 'Is research support in data science or related fields not requested from NSF because it lacks a home? ' Probably, but it seems to be very difficult to handle cross-discipline work. When I despaired at the out of date references to optimization in IEEE publications I tried to expand on the optimization description in several co-authored papers. The journals always cut that material out as not being sufficiently relevant.

For about 10 years I have been working on systems biology problems involving the organization and dynamics of proteins in cell membranes. This requires a solid foundation in probability theory, particularly random variables and random walks. Would be good to know about anomalous diffusion also. To analyze the data produce by the biologists one needs to know about probability distributions and about statistical test used to see if two data sets come from the same distribution or not.

Despite its obvious importance in the country's decision-making, the quality and timeliness of economic data at the State and National level is horrible. This is also true in other countries. While it may be that nothing can be done about it, the area should be examined by a good, cross-disciplinary committee to find out. In particular, I believe there is a place for Barnsley's Iterated Function Systems in the generation of economic data and models.

As one who has been involved in image processing and inverse acoustic problems, I would suggest that statistics is essentially an inverse problem -- to take data, possibly flawed, and extrapolate from that an underlying probability distribution.

Two remarks from a member of both ASA and SIAM:

1) Data science involves a much broader community than just statistics; the statistics community should not be solely in charge of any initiative, or any piece of any initiative -- all decisions should involve others.

2) Data science should be allowed to develop and evolve on its own, without the NSF trying to dictate which directions of research are most promising. The committee's consideration of items (3)-(5) in the list of questions it posed (duplicated below) is not encouraging. Individual researchers and research groups are best equipped to push the frontiers of data science, not a government committee (even if composed of researchers). The NSF should fund good researchers, not pet projects.

(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?

ANSWER: Data Science is part of the mathematical science, as such we could have a special call for it like it is done for Algebra & Number theory, Applied Mathematics etc.

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

ANSWER: No it has a home, it is the mathematical science division. (3) Are there complex or massive data problems that might be amenable to joint attack by several different disciplines?

ANSWER: Yes, absolutely, we have seen that Statistics is not the only tool to analyze data, one requires plenty of linear algebra (e.g. in Google search engine) and more recently optimization and topology!

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

ANSWERS: Linguistics, homeland security, library sciences, operations research analytics, etc

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

ANSWER: I think the methods from geometry and topology will play a bigger and bigger role in image processing and analysis.

Some thoughts on data mining and statistics:
1. Data mining is already a part of computer science and should be evaluated by the same criteria as other parts of computer science. [Is it important, is it successful, etc]

2. 'Data science' is a bad idea. The old claim that statistics is the science of giving meaning to data lacked a satisfactory definition of 'meaning'. Semantics has not succeeded in giving a scientifically credible criteria for 'meaning'. Chomsky on natural language versus formal language and 'meaning' in automated language translation should be considered here.

3. A scientific criteria for 'meaning' might need the neuroscientists to actually define 'meaning' as patterns of firing of neurons observed when someone claims to understand the meaning of something.

3. Statistics is good tool for analyzing data from controlled experiments under lab conditions [standard statistics meaning consequences of the Neyman-Pearson Lemma] and some social scientists claim Savage's Bayesian mathematics is a useful tool, but too often data manipulation is social science palming off opinion behind inappropriate procedures [Medical research does double blind experiments, do the education researchers also?]. Marxists, Keynesians and monetarists all have access to the same economic data. Should they be expected to agree on its 'meaning'?

4. The responsibility for analyzing data lies with the disciplines which generate it, not with 'data science'. Biologists bear responsibility for interpreting DNA data, astronomers for computerized photon collections. If either develop mathematical tools to analyze it, they will probably be funded through the existing disciplines. [Lots of interesting statistical techniques come out of social science, even if the use lacks credibility.]

5. If the NSF believes there is new mathematics to be developed to handle massive sets of numbers then fund mathematics, or computer science. 'Data science' sounds too much like 'statistical science', which wants the credibility of mathematics without admitting it is mathematics.

6. The differing styles of funding for research in mathematics [money for release from teaching] and funding of large teams in research labs [equipment and lab rats] will need reworking if the problems are to be considered in a way in which mathematicians are to interact, as equals, with the people who design and conduct experiments which generate huge numbers of zeroes and ones.

Finally, for over a generation too many in academic math have come to believe they carry a permanent sign saying 'kick me' [eg the February PCAST report]. People go into math as a discipline because the challenges are worthwhile. They leave because the disparity between mathematics and other careers [Wall Street or law] is too large. When academic management replaces mathematicians with NTT retired high school teachers in college math courses, the message to mathematicians is get lost. Most jobs in academic math are for teaching first year students. The people who might have the talent and background in mathematics to develop new mathematics for analysis of massive data sets are being pushed out of research by the same academic management which supports pseudo disciplines such as 'Decision Sciences' and 'Data Sciences'.

If the NSF wants people with mathematical ability to tackle the mathematical challenges of finding credible patterns in large data sets it should oppose the attacks on mathematics. If the NSF does not want to fund mathematical careers at a level with the lab scientists, it should consider paying the data miners on Wall Street to share their trade secrets.

Hello,
Personally, my main field is mathematical modeling in biology and medicine, so naturally I am more aware of the large data formed in this area.

General comment:

While it may be very obvious, I think it should be noted that applied statistics and data-intensive science are multidisciplinary by nature, as to produce data-intensive mostly some other aim outside of mathematics is the cause for the data collection. Simply put, mathematicians may help in analyzing data but they don't produce it.

As far as my understudying goes, since the field of large data sets analysis is relatively new, one direction that need to be promoted is analysis based on understudying of the nature of the data, which requires a multidisciplinary team.

Thus I would think that one way in this direction is to form a collaborative grant, directed to solving science questions - within or outside the mathematical division (e.g., image/video can also create large data sets), but my motivation is mostly biology. It could have either collaborative structure with other divisions at NSF or even NIH, and an independent one to the mathematical division that requires active collaboration with active projects - so that the data analysis can effect the data collection process. The additional funds should be given in a way that will encourage a genuine and active collaboration.

As a statistician, who is a member of both ASA and SIAM, I find myself being funded by the computer science arm of NSF (CDI) to develop computational infrastructure for statistics. This is something that the statistics program should support. Unfortunately its limited resources require it to focus on core mathematical statistics areas.

The dichotomy between 'little science' and 'big science' rings very true in my experience. Statistics is lacking in many 'big science' projects where it could make significant impact. Not many (if any) big science projects are being lead by statisticians. It takes too many years for new statistical results (from 'little science' investigators) to enter practice and then they are often misapplied.

For example, it required a statistics-lead team of two statisticians, a mathematician, and a computer scientist, (with input from many computational science colleagues) to develop big data infrastructure for the R programming language. While the ideas for this were in my head 10 years ago, it was only after I assembled the team less than a year ago that major results became quickly possible.

The field of uncertainty quantification (within the computational mathematics community) is viewed by statisticians as naive in its understanding of uncertainty but advanced in its understanding of computational models. The opposite opinion of statisticians might be held by computational mathematicians. Some recent headway in cross-fertilization is being made by SAMSI.

I find that some areas in numerical analysis are lacking interaction with the statistics community. For example, current numerically stable solutions in linear algebra are lacking support for common statistical uses of least-squares.

I find that supercomputing reliability could benefit from statistical reliability. Much is being reinvented and misapplied, and could benefit from broader interaction with statistics.

Here are a few comments concerning the commentary about StatsNSF:

First, there is nothing in this document on education. It seems to me that encouraging programs in Applied Mathematics to incorporate more statistics and probabilistic ideas into programs would help. At my institution, Maryland, as Dave will know, several years ago we added an Applied Statistics track to our graduate progam. My impression however is that this program is still not fully integrated into our Applied Math program, in the sense that not all students take a course in a statistical topic. NSF could encourage more integration of material in curricula; a similar thing could be done in an opposite direction, also, for example to encourage integration of computational methods into statistical programs.

Second, the document puts emphasis on the "big science" vs. individual projects. I'm not sure why this point is emphasized as strongly as it is. I think it is important to incorporate statistical methods into Applied Mathematics, for both types of endeavors, and small projects that pair say computational and statistical scientists, each moving halfway toward the other domain, should be strongly encouraged.

(1) For automated processing of large data sets the following kinds of models must be developed. They are often not mathematically sophisticated in themselves (often could be developed by students under supervision), but useful models must be thoroughly tested and refined (again, good jobs for students) to be sufficiently accurate and reliable. This modeling is naturally multidisciplinary, depending on the application.

DATA MODELS: to connect raw data (numbers, text, pictures, etc) to (usually numerical) refined data that can be mathematically processed. Example: have the patient report his pain on a scale from 1 to 10, or have students take a physics test that results in a score.

QUERY or INQUIRY MODELS: what is the user/analyst looking for in the data? And what does that look like mathematically? Example: if a government analyst wants to search internet text data to see if anyone in Iraq wants to assassinate the president, what does that question 'wants to assassinate the president' look like in this context?

CONTEXT MODELS: to provide mathematical model of a certain situation of interest. (Microsoft, etc try to do this; when you start typing in Word a cartoon pops up and asks 'It looks like you are typing a letter; do you want me to help?') Example: in the 'does anyone want to assassinate the president' question, a context model might find (automatically from available data) whether all or most elements necessary for a successful assassination are present (people and equipment in same place at the same time with president).

MODULAR MODELING TOOLS should be developed (almost like a programming language??) so that new models can be quickly prototyped for new applications and timely needs.

HIERARCHICAL AND META STATISTICAL METHODS, it seems to me, would be an important area of statistical research in connection with large data problems. This because the data may get combined/fused in layers. Example: temp data gets combined, as does wind, humidity, and barometric pressure data, and then they all get combined to predict tomorrow's weather. In automated analysis, estimates and confidence regions need to be passed from one layer to the next.

— *continued on next page* —

— *continued* —

Important fields: Natural Language Processing, Artificial Intelligence, Machine Learning, Automatic Target Recognition (eg, imagery), data mining.

(3) Several government documents address needs for automated data and information processing of large data sets, often to give decision makers a complete, unambiguous picture of the situation.

DOD

*Joint Warfighting Science and Technology Plan. May, 1996. OSD (USD A&T). Washington D.C.: Department of Defense.

*Science and Technology (S&T) Priorities for Fiscal Years 2013-17 Planning. April 19, 2011. OSD. Washington D.C.: U.S. Department of Defense.

*Department of Defense Research & Engineering Strategic Plan. 2007. Director of Defense Research & Engineering (DDRE). Washington D.C.: U.S. Department of Defense.

Navy

*The U.S. Navy's Vision for Information Dominance. May 2010. USN (CNO). Washington D.C.: U.S. Department of Defense.

*Naval S&T Strategic Plan. September 1, 2011. ONR. Washington D.C.: U.S. Department of Defense.

*ASN Grand Opportunities (NLCCG). April 30, 2010. ASN RDA. Washington D.C.: U.S. Department of Defense.

DHS

*Homeland Security High-Priority Technology Needs - Version 3.0 (2009). May, 2009. DHS. Washington D.C.: U.S. Department of Homeland Security.

(4) The success of Watson on Jeopardy and Google search are great success stories. They (at least Google) can be used 'as is' but to really make this technology useful in a big way means (re-developing it or) the developer making his methods and experience public. (I don't know what the state of this is.)

> (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?

Much less bureaucracy! It should be easy to apply and manage and report on grants. The NSF puts a huge burden on researchers currently.

Along with larger grants there should be many small and medium sized grants that allow more researchers to be involved.

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

Certainly. Without a home in NSF people feel they will not be fairly dealt with and the go elsewhere. Naturally we feel if there is an area in-between the referees will skewer us and give grants to the established specialists. Thus assuring never to have any real innovation.

That is a major, major consideration.

Moreover, there is going to be Data and stats out there that is done through people who are not naturally "NSF" clientele and they will not apply either.

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

Numerical analysis and computational math is close to statistics and there is traditionally not much collaboration. This is the fault of the numerical analyst 100\%, the statisticians have carried on and done good computational math without the other group helping much.

But then again, a computational mathematician doing statistics would never get a serious NSF grant from DMS Comput. Math. This is because numerical analysts will judge the applied stats to be trivial and dismiss it. So why is it surprising that they don't do stats?

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

As a numerical analyst I see a big need for more work in "simulation" and courses too are needed. "Out there" there is a big lack of such quality courses. Researchers too have done some good things but generally the special sessions and conferences are always separated and simulation experts do not talk to numerical analysts nor do computational statisticians.

The future involves HPC computational stats too, in the way the HPC is related to numerical analysis. A big effort needs to be put to that end.

On education.... I went through an excellent numerical analysis education at Univ of Wisconsin Madison in its hey day of Math Research Center, etc. I now am in a dept. that includes statistics, where at Madison Stats is a famous BUT SEPARATE department. We in math had nothing every to do with statistics. Even at the Math Research Center no one cared a whit about statistics.

— *continued on next page* —

I think massive efforts should be made to enhance statistics education:

1. Market that as a good major

2. Bring stats and applied math, and computational math, and computer science TOGETHER any way possible

3. Bring stats to deal with applications better. I mean, there are a lot of depts that do their own statistics because the regular experts cannot teach it the way they need it. Examples, Econ, business, ed Psych, Information science, africology, psychology, social work, public health.... that is an amazing list of complete lack of attention from the statistician educators.

4. Link statistics and data better with operations research and industrial engineering.

5. Create more graduate ms and phd fellowships for statistics to encourage activity.

6. Get more money in it so Dean's will hire more statisticians.

7. Involve high schools like with science fairs to work in the area of data, which will encourage more students to choose that major.

S> (1) What should NSF do to further promote and facilitate the
S> appropriate development of this multidisciplinary field of data
S> science?

Large interdisciplinary projects would help. In my experience (15 years at a mid sized mid quality state university and 8 years at a national laboratory) disciplinary insularity impedes progress. The problems I know about require creativity in science, modeling, computation and statistics. Specialists in each area are professionally encouraged to push the frontier in their area alone. So each specialist uses terrible tools from other areas.

S> (2) Is research support in data science or related fields not
S> requested from NSF because it lacks a home? If so, what might be
S> a possible remedy? I don't know.
S> (3) Are there complex or massive data problems that might be
S> amenable to joint attack by several different disciplines?

Yes. There are many. I am particularly impressed by the PECOS program at UT Austin. It is supported by a DOE program designed to address the kind of social/academic problem you are addressing.

S> (4) Are there disciplinary areas that could benefit [...] (5) Are
S> you aware of simultaneous development of data science methods
S> [...]

Yes. But most of the problems I know well are not in areas that the NSF should fund. However, your effort is important because to execute the missions of the national laboratories, they will need staff members who can team up and exploit and develop the best methods and tools in what are now several disciplines.

At the risk of oversimplification, the applied mathematics/physical sciences have traditionally built detailed models with great effort spent capturing physical causalities at the expense of considering data (and other) uncertainties (and sometimes also variability). On the other hand, the statistical sciences have traditionally built simpler relational models with great effort spent on quantifying data (and other) uncertainties and variability. Disciplines such as Inverse Problems have already faced, and largely overcome, this dichotomy, but I think that it must be more broadly addressed now that we have the computational power to tackle problems with both detailed models and large data sets with uncertainty/variability. Unfortunately, the above "separation of concerns" seems to be even more stubbornly adhered to in our educational systems. I also think that the ongoing standardization of measurement science (e.g. the ISO GUM) represents a microcosm of the combination of the above two concerns. However, I don't think the "big data" wave has yet hit the metrology community.

It would be better if traditional and statistical computational/numerical methods were better integrated. I think too few traditional numerical methods courses include stochastic modeling/simulation, require error analyses that enable uncertainty quantification, or include statistical simulation/computation methods (e.g., Monte Carlo integration or and Markov chain Monte Carlo).