Update from StatSNSF Subcommittee

Iain Johnstone & Fred Roberts,

Co-Chairs,

July 18, 2013

The Task

From Ed Seidel's August 14, 2012 letter to MPSAC:

"that MPSAC form a subcommittee to examine the current structure of support of the statistical sciences within NSF and to provide recommendations for NSF to consider"

Statistical Science-> Data Science

Motivated by NSF Strategic Plan and initial discussions with ADs

Our definition:

"*Data Science:* the science of *planning*, *acquisition*, *management*, *analysis* of, and *inference* from data"

Our context:

Data science and the enhanced application of data science at NSF

Report Section: "Data Science in the NSF Context"

[Draft!!] Recommendations

- From working groups, community input, calls
- **Still under discussion**, seeking feedback

I. NSF Organization
II. NSF Research Initiatives
III. Workforce Development
IV. Proposal and Panels

I. NSF Organization

1) **Coordinate Data Science across NSF in a way that engages all Directorates.**

Including:

Coordinate current efforts across NSF involving data science
Identify/mitigate fragmentation of data science research.
Develop/lead new cross-directorate initiatives involving DS [Examples]
Develop policies to increase the quality of science through proper use of DS.
Improve representation of DS experts on review panels, …

"Coordinate Data Science across NSF…"

(cont'd):
Develop funding models to include data scientists in cross-disciplinary research.
Connect with emerging education efforts focusing on DS
Study reproducibility issues in NSF funded science
Track data science funding

Some *possible* mechanisms:

- Office of Data Science [e.g. NIH]
- Data Science Working Group [e.g. SEES]
- Cross-foundation leadership group

II. NSF Research Initiatives

1) **Create new initiatives that embrace and address the cross-cutting nature of foundational research in data science**

   Cut across NSF directorates; adequately funded

Some ***possible*** examples:

- Computational stat and/or applied stat – new methodology cutting across more than 1 discipline
- Proposals to synthesize & identify overlap in DS/stat methods in different fields
- Research initiative on reproducibility of computational science – reproducible computational results
- Research in massive data analysis

II. NSF Research Initiatives

2) Provide long-term mechanism for support of DS activities in interdisciplinary settings, e.g. by including support for data scientists in interdisciplinary teams

3). Develop more joint programs between NSF & other agencies that need new DS.

Naturally interdisciplinary; also lead to training in interdisciplinarity; could involve multiple agencies

4). Create programs to enhance community awareness of possible funding thru cross-cutting initiatives at NSF

E.g., Workshops; community building programs (DIBB); infrastructure building programs (Earth Cube, Data Way, BCC); more effective “Dear Colleague” letters

III. Workforce Development

1) **Initiate a major thrust to support grad fellowships, postdoc fellowships, CAREER Awards in DS.**
2) **Introduce undergrads to opportunities & challenges in DS**

E.g., expand undergrad summer programs akin to NIH Summer Institutes in Biostat; DS REU program

3). Sponsor development efforts geared at raising awareness of DS in K-12

4). Develop programs for interaction between data scientists and other scientists

E.g., summer conferences; training programs to transfer methodology between fields; targeted sabbaticals; short courses on proper data archiving

IV. Proposals & Panels

1) **Expand requirement for a data management plan to include a data analysis plan and a disclosure management plan**

   Encourage more careful examination of methodologies to be used. Detail how and when data will be made available.

2) **Ensure adequate data science representation on review panels in which DS is a key component of the science**

QUESTIONS AND DISCUSSION