Data Science at NSF

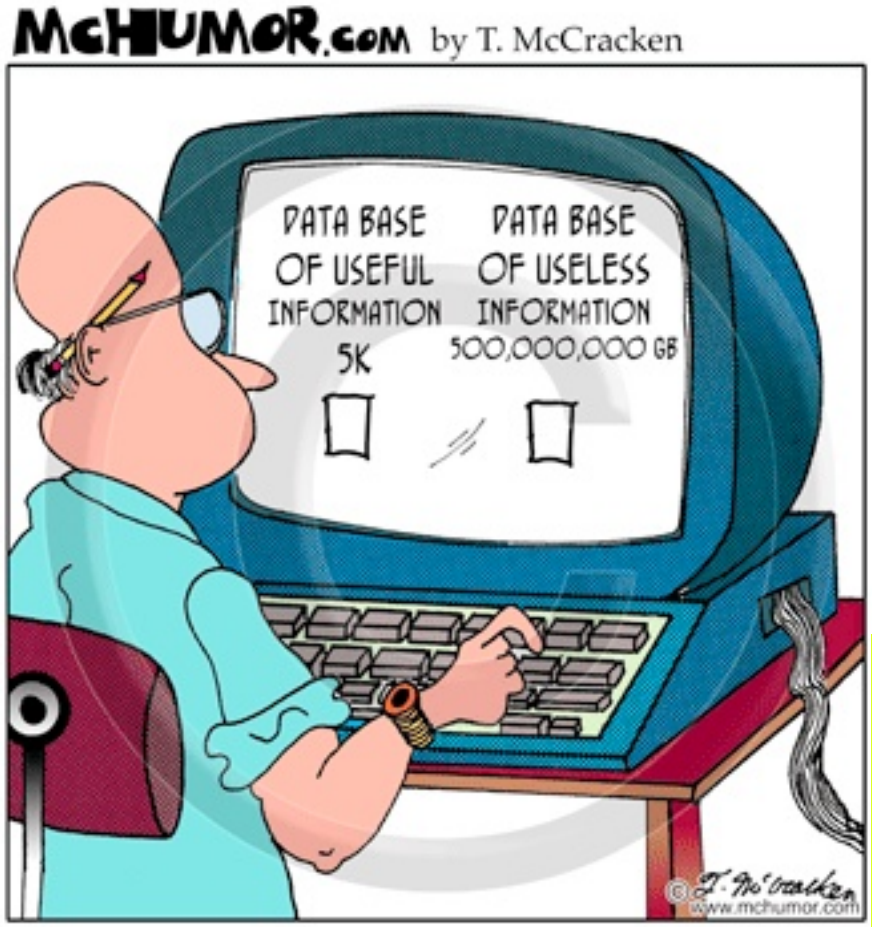
Draft Report of StatSNSF committee: Call for input from NSF A.C.s

Iain Johnstone, Fred Roberts, Co-chairs

Jan 2014

The Context

- Data is central to NSF research
- Statistical sciences + computational resources + disciplinary developments
- Heightened attention to data analysis, prediction
- Focus on reproducibility, reliability of inferences


MCHUMOR.COM by T. McCracken
DATA BASE OF USEFUL INFORMATION 5K
DATA BASE OF USELESS INFORMATION 500,000,000 GB
©T. McCracken mchumor.com


www.VADLO.com
Science
nature
THE FUTURE IS BRIGHT
Daily Journal Of Negative Data

Report Structure

Executive Summary

1. Introduction

2. Data Science in the NSF context

3. Overview of Underlying Challenges

4. Recommendations

5. Research and Data Gathered

Appendices

1. Introduction

- Subcommittee of MPS AC [17 members]

- Charged by MPS AD [w. support of all ADs] to

  "to examine the current structure of support of the statistical sciences within NSF and to provide recommendations for NSF to consider"

- *Charge mandates NSF-wide scope:*
  - *Membership and input from each Directorate AC*
  - AC input sought before report is finalized [May-June]

2. Data Sciencein NSF context

Motivated by NSF Strategic Plan and initial discussions with ADs

Our definition:

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

Our context:

Data science and the enhanced application of data science at NSF

2. Data Science at NSF ctd.

- requires broad set of skills & perspectives
  - Mathematics, statistics, computer science, domain specific expertise

- Challenges at all scales of data
  - 'Big data' is a vast ongoing arena, but
  - NSF should also embrace the 'long tail' of projects of smaller size: new/complex data types

3. Some underlying challenges

- Growth of Data Science
  - McKinsey forecast of shortage
- Fragmentation of Data Science at NSF
  - duplication, 'cracks',…
- Research quality
  - use the best data science, reproducibility,…
- Multi-disciplinarity of Data Science
  - effective collaboration and training

4. Draft Recommendations

- Recommendations in four categories:

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

- Input sought before report is finalized [May-July]

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

2. Create new initiatives that embrace and address the cross-cutting challenges of data science.

- Examples in Section 4

3. Provide mechanisms for enhancing the participation of data scientists in data science activities in interdisciplinary settings

III. Workforce Development

4. Initiate a major thrust to support

- graduate, postdoctoral and early career fellowships and awards,

and develop appropriate programs to expand

- undergraduate exposure to, and
- K-12 awareness of data science.

IV. Proposal and Review Cycle

5. When appropriate:

- in proposals, require a data analysis plan and a disclosure management plan, and
- in review, ensure that there is adequate data science representation on panels.

DISCUSSION

- Recommendations in four categories:

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

- Input sought before report is finalized [May-July]

Supplementary Slides

- Slides giving more details

Data Science in the NSF Context

- Projects focusing on data science as fundamental research area vs. use of existing DS tools to advance a domain research project.
- Committee believes NSF engagement with DS should address both

Data Science in the NSF Context: Examples of Research Topics within DS

- Methods to design data collection
- Cleaning and preparing data
- Data reduction
- Biases in data collection
- Estimation/inference/learning/discovery from data
- Identifying/exploiting similarity in data methodologies across disciplines
- Causal inference
- Reliability of inferences from data
- Integrating network analysis with other methods of data analysis
- Searching through large/complex data sets
- Multiple media/distributed sources

Data Science in the NSF Context: Examples of Research Topics within DS

- Decision analysis in presence of uncertainty
- Data-driven modeling and simulation
- Uncertainty quantification
- Reproducibility in analyses of data
- Privacy and bounded information disclosure
- Data sharing
- Computational tools and software infrastructure for all of the above

Data Science in the NSF Context: Examples of Research Topics Involving Application of Data Science

- Adapting and deploying a DS method in a non-routine way
- Transference and adaptation of existing methodology across fields
- Development of widely-applicable community software
- Development of new DS methods within context of a particular domain, with potential for transference to other domains
- Promotion of interdisciplinary teams
- Enabling the sharing of software
- Presentation of data that enables/simplifies alternative analyses

NSF Organization

Recommendation: **Coordinate Data Science across NSF in a way that engages all Directorates.**

- Goals:
  - Address unevenness in addressing DS across NSF
  - Aid in effective communication of DS successes from one area to another
  - Avoid duplication of effort
  - Provide cross-directorate leadership
  - Create organizational structure thru which the leadership can act to achieve goals
- Specific mechanisms: The Committee believes that people on the ground at NSF are in best position to decide how to organize such an effort

New Research Initiatives

Recommendation: **Create new initiatives that embrace and address the cross-cutting challenges of data science.**

- Examples of potential new initiatives:
  - Cross-directorate initiative: identifying overlap in DS methods in different fields, and finding mechanisms to enhance tech transfer between fields
  - Reproducibility of computational science
  - Theory and analysis of massive data sets and streams
  - Programs in computational statistics and/or applied statistics – methodologies useful in more than one discipline
  - Uncertainty quantification

Enhancing Participation of Data Scientists in Interdisciplinary Settings

Recommendation: **Provide mechanisms for enhancing the participation of data scientists in data science activities in interdisciplinary settings**

- Some examples of potential new/enhanced programs:
  - New initiatives to support direct involvement of data scientists in interdisciplinary teams
  - More joint programs between NSF and other agencies needing new DS (NIH, DoD, DHS, Census Bureau, …)

Enhancing Participation of Data Scientists in Interdisciplinary Settings

Recommendation: **Provide mechanisms for enhancing the participation of data scientists in data science activities in interdisciplinary settings**

- Some examples of potential new/enhanced programs:
  - Programs to enhance community awareness of possible funding thru cross-cutting initiatives at NSF:
    - Workshops org by professional societies, multi-disciplinary institutes
    - Educating the community about DS opportunities at NSF outside of specific programs in divisions
    - Infrastructure building programs
    - More effective use of Dear Colleague Letters
  - Different models of research support (e.g., contests to address DS problems cutting across fields)

Workforce

Recommendation: **Initiate a major thrust to support graduate, postdoctoral and early career fellowships and awards, and develop appropriate programs to expand undergraduate exposure to, and K-12 awareness of data science.**

- Some examples of potential new/enhanced programs:
  - DS fellowships/awards with strong interdisciplinary component and significant mentorship
  - Undergrad summer programs like NIH Summer Institutes for Training Biostatisticians or dedicated DS REU program
  - Science grant supplements for developing teacher guidebooks, discovery-based learning modules, expanding research experiences for teachers
  - Joint programs between EHR and other directorates to develop science grant supplements based on fundamental learning principles

Workforce

Recommendation: **Initiate a major thrust to support graduate, postdoctoral and early career fellowships and awards, and develop appropriate programs to expand undergraduate exposure to, and K-12 awareness of data science.**

- Some examples of potential new/enhanced programs:
  - Programs engaging data scientists with other scientists:
    - Summer conferences, immersive workshops
    - Training programs aimed at transferring DS methodology between fields
    - One-semester programs/sabbaticals for data scientists in a subject matter discipline
    - Short courses on proper archiving of data

Proposal and Review Cycle

Recommendation: **When appropriate, in proposals, require a data analysis plan and a disclosure management plan; and in review, ensure that there is adequate DS representation on panels.**

- Some examples of potential new/enhanced initiatives:
  - The Data Analysis Plan and Disclosure Management Plan would:
    - Expand proposal Data Management Plan
    - Avoid duplicative development of methods in multiple areas
    - Enhance reproducibility of science by detailing how and when data, software, etc. will be made available.

BY VASANT DHAR

Data Science and Prediction

USE OF THE term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry?

including our confidence in the inference. Why then do we need a new term like data science when we have had statistics for centuries? The fact that we now have huge amounts of data should not in and of itself justify the need for a new term.

The short answer is data science is different from statistics and other existing disciplines in several important ways. To start, the raw material, the "data"