

# Data Science at NSF

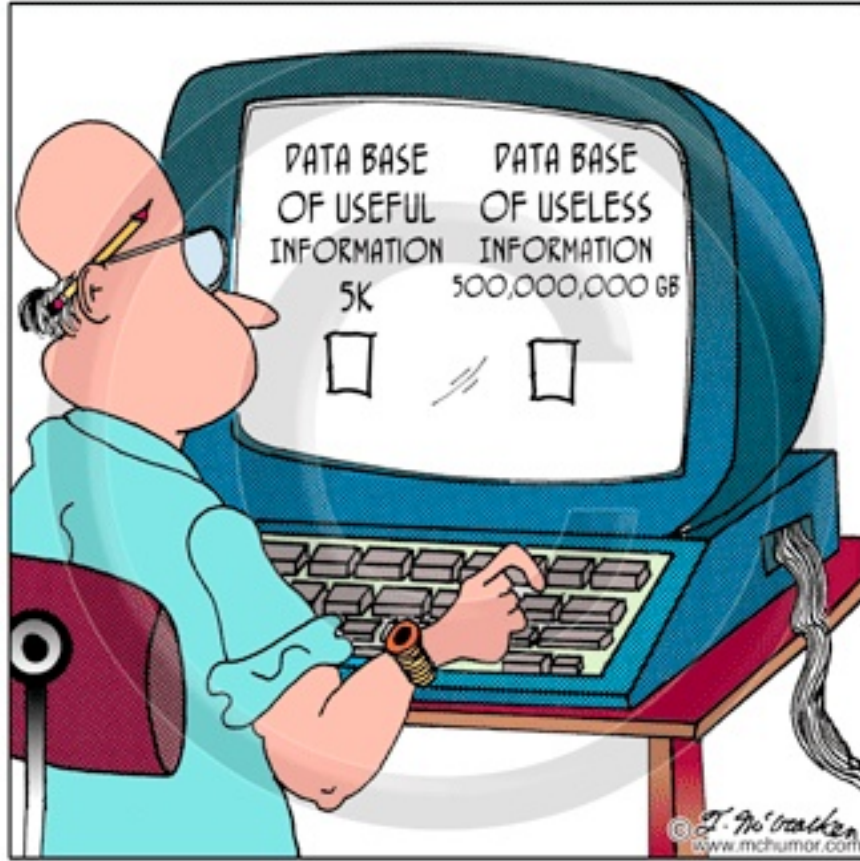
**Draft Report of StatNSF committee:  
Call for input from NSF A.C.s**

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



# 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 Science in 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]