THE BIG DATA REVOLUTION

What Does It Mean for Research?

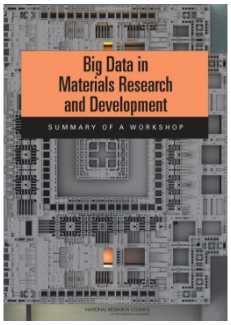
October 14-15, 2014

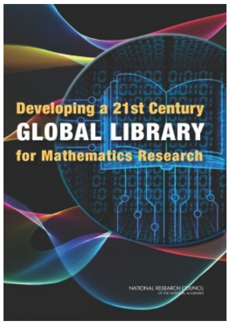
Government-University-Industry Research Roundtable

*List of selected reports from the National Academies related to the topic Big Data

FURTHERING AMERICA'S RESEARCH ENTERPRISE (DBASSE 2014)
Scientific research has enabled America to remain at the forefront of global competition for commercially viable technologies and other innovations. For more than 65 years, the United States has led the world in science and technology. Discoveries from scientific research have extended our understanding of the physical and natural world, the cosmos, society, and of humans - their minds, bodies, and economic and other social interactions. Through these discoveries, science has enabled longer and healthier lives, provided for a better-educated citizenry, enhanced the national economy, and strengthened America's position in the global economy. At a time of budget stringency, how can we foster scientific innovation to ensure America's unprecedented prosperity, security, and quality of life? Although many studies have investigated the impacts of research on society, *Furthering America's Research Enterprise* brings to bear a fresh approach informed by a more holistic understanding of the research enterprise as a complex, dynamic system. This understanding illuminates why America's research enterprise has historically been so successful; where attention should be focused to increase the societal benefits of research investments; and how those who make decisions on the allocation of funds for scientific research can best carry out their task.

BIG DATA IN MATERIALS RESEARCH AND DEVELOPMENT: SUMMARY OF A WORKSHOP (DEPS, 2014)
Big Data in Materials Research and Development is the summary of a workshop convened by the National Research Council Standing Committee on Defense Materials Manufacturing and Infrastructure in February 2014 to discuss the impact of big data on materials and manufacturing. The materials science community would benefit from appropriate access to data and metadata for materials development, processing, application development, and application life cycles. Currently, that access does not appear to be sufficiently widespread, and many workshop participants captured the constraints and identified potential improvements to enable broader access to materials and manufacturing data and metadata. This report discusses issues in defense materials, manufacturing and infrastructure, including data ownership and access; collaboration and exploitation of big data's capabilities; and maintenance of data.



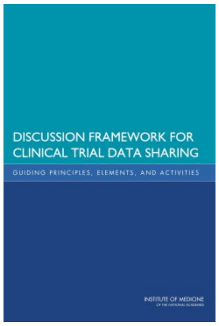


DEVELOPING A 21ST CENTURY GLOBAL LIBRARY FOR MATHEMATICS (DEPS, 2014)

Developing a 21st Century Global Library for Mathematics Research discusses how information about what the mathematical literature contains can be formalized and made easier to express, encode, and explore. Many of the tools necessary to make this information system a reality will require much more than indexing and will instead depend on community input paired with machine learning, where mathematicians' expertise can fill the gaps of automatization. This report proposes the establishment of an organization; the development of a set of platforms, tools, and services; the deployment of an ongoing applied research program to complement the development work; and the mobilization and coordination of the mathematical community to take the first steps toward these capabilities. The report recommends building on the extensive work done by many dedicated individuals under the rubric of the World Digital Mathematical Library, as well as many other community initiatives. *Developing a 21st Century Global Library for Mathematics* envisions a combination of machine learning methods and community-based editorial effort that makes a significantly greater portion of the information and knowledge in the global mathematical corpus available to researchers as linked open data through a central organizational entity-referred to in the report as the Digital Mathematics Library. This report describes how such a library might operate - discussing development and research needs, role in facilitating discover and interaction, and establishing partnerships with publishers.

DISCUSSION FRAMEWORK FOR CLINICAL TRIAL DATA SHARING: GUIDING PRINCIPLES, ELEMENTS, AND ACTIVITIES (IOM, 2014)

Sharing data generated through the conduct of clinical trials offers the promise of placing evidence about the safety and efficacy of therapies and clinical interventions on a firmer basis and enhancing the benefits of clinical trials. Ultimately, such data sharing - if carried out appropriately - could lead to improved clinical care and greater public trust in clinical research and health care. *Discussion Framework for Clinical Trial Data Sharing: Guiding Principles, Elements, and Activities* is part of a study of how data from clinical trials might best be shared. This document is designed as a framework for discussion and public comment. This framework is being released to stimulate reactions and comments from stakeholders and the public. The framework summarizes the committee's initial thoughts on guiding principles that underpin responsible sharing of clinical trial data, defines key elements of clinical trial data and data sharing, and describes a selected set of clinical trial data sharing activities.


DISCUSSION FRAMEWORK FOR
CLINICAL TRIAL DATA SHARING
GUIDING PRINCIPLES, ELEMENTS, AND ACTIVITIES
INSTITUTE OF MEDICINE

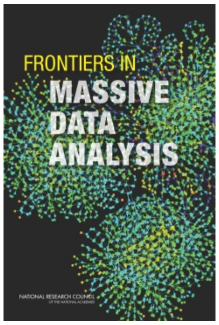
Capturing Change in Science,
Technology, and Innovation
IMPROVING INDICATORS TO INFORM POLICY
NATIONAL RESEARCH COUNCIL

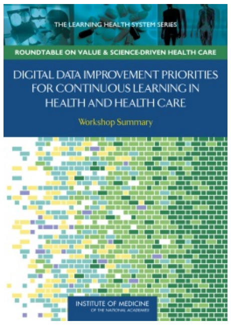
CAPTURING CHANGE IN SCIENCE, TECHNOLOGY, AND INNOVATION (DBASSE/PGA, 2014)

Since the 1950s, under congressional mandate, the U.S. National Science Foundation (NSF) - through its National Center for Science and Engineering Statistics (NCSES) and predecessor agencies - has produced regularly updated measures of research and development expenditures, employment and training in science and engineering, and other indicators of the state of U.S. science and technology. A more recent focus has been on measuring innovation in the corporate sector. NCSES collects its own data on science, technology, and innovation (STI) activities and also incorporates data from other agencies to produce indicators that are used for monitoring purposes - including comparisons among sectors, regions, and with other countries - and for identifying trends that may require policy attention and generate research needs. NCSES also provides extensive tabulations and microdata files for in-depth analysis. *Capturing Change in Science, Technology, and Innovation* assesses and provides recommendations regarding the need for revised, refocused, and newly developed indicators of STI activities that would enable NCSES to respond to changing policy concerns. This report also identifies and assesses both existing and potential data resources and tools that NCSES could exploit to further develop its indicators program. Finally, the report considers strategic pathways for NCSES to move forward with an improved STI indicators program. The recommendations offered in *Capturing Change in Science, Technology, and Innovation* are intended to serve as the basis for a strategic program of work that will enhance NCSES's ability to produce indicators that capture change in science, technology, and innovation to inform policy and optimally meet the needs of its user community.

FRONTIERS IN MASSIVE DATA ANALYSIS (DEPS, 2013)

Frontiers in Massive Data Analysis examines the frontier of analyzing massive amounts of data, whether in a static database or streaming through a system. Data at that scale--terabytes and petabytes--is increasingly common in science (e.g., particle physics, remote sensing, genomics), Internet commerce, business analytics, national security, communications, and elsewhere. The tools that work to infer knowledge from data at smaller scales do not necessarily work, or work well, at such massive scale. New tools, skills, and approaches are necessary, and this report identifies many of them, plus promising research directions to explore. *Frontiers in Massive Data Analysis* discusses pitfalls in trying to infer knowledge from massive data, and it characterizes seven major classes of computation that are common in the analysis of massive data. Overall, this report illustrates the cross-disciplinary knowledge--from computer science, statistics, machine learning, and application disciplines--that must be brought to bear to make useful inferences from massive data.



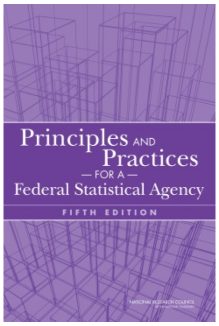

THE LEARNING HEALTH SYSTEM SERIES
ROUNDTABLE ON VALUE & SCIENCE-DRIVEN HEALTH CARE
DIGITAL DATA IMPROVEMENT PRIORITIES FOR CONTINUOUS LEARNING IN HEALTH AND HEALTH CARE
Workshop Summary
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

DIGITAL DATA IMPROVEMENT PRIORITIES FOR CONTINUOUS LEARNING IN HEALTH AND HEALTH CARE: WORKSHOP SUMMARY (IOM, 2013)

Digital health data are the lifeblood of a continuous learning health system. A steady flow of reliable data is necessary to coordinate and monitor patient care, analyze and improve systems of care, conduct research to develop new products and approaches, assess the effectiveness of medical interventions, and advance population health. The totality of available health data is a crucial resource that should be considered an invaluable public asset in the pursuit of better care, improved health, and lower health care costs. In addition to increased data collection, more organizations are sharing digital health data. Data collected to meet federal reporting requirements or for administrative purposes are becoming more accessible. Efforts such as Health.Data.gov provide access to government datasets for the development of insights and software applications with the goal of improving health. Within the private sector, at least one pharmaceutical company is actively exploring release of some of its clinical trial data for research by others. *Digital Data Improvement Priorities for Continuous Learning in Health and Health Care: Workshop Summary* summarizes discussions at the March 2012 Institute of Medicine (2012) workshop to identify and characterize the current deficiencies in the reliability, availability, and usability of digital health data and consider strategies, priorities, and responsibilities to address such deficiencies.

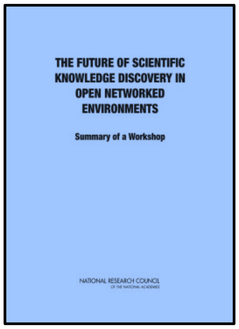
SHARING CLINICAL RESEARCH DATA: WORKSHOP SUMMARY (IOM, 2013)

Pharmaceutical companies, academic researchers, and government agencies such as the Food and Drug Administration and the National Institutes of Health all possess large quantities of clinical research data. If these data were shared more widely within and across sectors, the resulting research advances derived from data pooling and analysis could improve public health, enhance patient safety, and spur drug development. Data sharing can also increase public trust in clinical trials and conclusions derived from them by lending transparency to the clinical research process. This public workshop focused on strategies to facilitate sharing of clinical research data in order to advance scientific knowledge and public health. While the workshop focused on sharing of data from preplanned interventional studies of human subjects, models and projects involving sharing of other clinical data types were considered to the extent that they provided lessons learned and best practices. The workshop objectives were to examine the benefits of sharing of clinical research data from all sectors and among these sectors, including, for example: benefits to the research and development enterprise and benefits to the analysis of safety and efficacy. *Sharing Clinical Research Data: Workshop Summary* identifies barriers and challenges to sharing clinical research data, explores strategies to address these barriers and challenges, including identifying priority actions and "low-hanging fruit" opportunities, and discusses strategies for using these potentially large datasets to facilitate scientific and public health advances.



PRINCIPLES AND PRACTICES FOR A FEDERAL STATISTICAL AGENCY: FIFTH EDITION (DBASSE, 2013)

Publicly available statistics from government agencies that are credible, relevant, accurate, and timely are essential for policy makers, individuals, households, businesses, academic institutions, and other organizations to make informed decisions. Even more, the effective operation of a democratic system of government depends on the unhindered flow of statistical information to its citizens. In the United States, federal statistical agencies in cabinet departments and independent agencies are the governmental units whose principal function is to compile, analyze, and disseminate information for such statistical purposes as describing population characteristics and trends, planning and monitoring programs, and conducting research and evaluation. The work of these agencies is coordinated by the U.S. Office of Management and Budget. Statistical agencies may acquire information not only from surveys or censuses of people and organizations, but also from such sources as government administrative records, private-sector datasets, and Internet sources that are judged of suitable quality and relevance for statistical use. They may conduct analyses, but they do not advocate policies or take partisan positions. Statistical purposes for which they provide information relate to descriptions of groups and exclude any interest in or identification of an individual person, institution, or economic unit. Four principles are fundamental for a federal statistical agency: relevance to policy issues, credibility among data users, trust among data providers, and independence from political and other undue external influence. *Principles and Practices for a Federal Statistical Agency: Fifth Edition* explains these four principles in detail.



THE FUTURE OF SCIENTIFIC KNOWLEDGE DISCOVERY IN OPEN NETWORKED ENVIRONMENTS

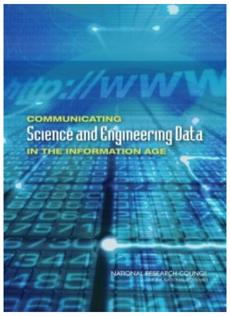
Summary of a Workshop

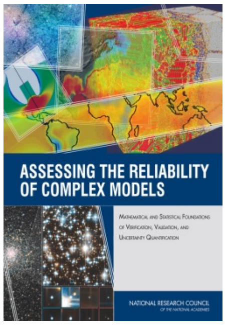
NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE FUTURE OF SCIENTIFIC KNOWLEDGE DISCOVERY IN OPEN NETWORKED ENVIRONMENTS: SUMMARY OF A WORKSHOP (PGA, 2012)
Digital technologies and networks are now part of everyday work in the sciences, and have enhanced access to and use of scientific data, information, and literature significantly. This 2011 workshop focused on computer-mediated or computational scientific knowledge discovery, taken broadly as any research processes enabled by digital computing technologies such as data mining, information retrieval and extraction, artificial intelligence, and distributed grid computing. Participants considered the following questions:

- What are the opportunities over the next 5-10 years in computer-mediated scientific knowledge discovery in the open online environment, and the potential benefits to science and society?
- What are the techniques and methods used to understand these processes, the validity and reliability of their results, and their impact inside and outside science?
- What are the major barriers to computer-mediated scientific knowledge discovery within the scientific community?
- What actions could be taken by sponsors and researchers to understand better the computer-mediated scientific knowledge discovery processes and mechanisms for openly available data?

COMMUNICATING SCIENCE AND ENGINEERING DATA IN THE INFORMATION AGE (DEPS, 2012)
The National Center for Science and Engineering Statistics (NCSES) of the National Science Foundation (NSF) communicates its science and engineering (S&E) information to data users in a very fluid environment that is undergoing modernization at a pace at which data producer dissemination practices, protocols, and technologies, on one hand, and user demands and capabilities, on the other, are changing faster than the agency has been able to accommodate. NCSES asked the Committee on National Statistics and the Computer Science and Telecommunications Board of the National Research Council to form a panel to review the NCSES communication and dissemination program that is concerned with the collection and distribution of information on science and engineering and to recommend future directions for the program. *Communicating Science and Engineering Data in the Information Age* includes recommendations to improve NCSES's dissemination program and improve data user engagement. This report includes recommendations such as NCSES's transition to a dissemination framework that emphasizes database management rather than data presentation, and that NCSES analyze the results of its initial online consumer survey and refine it over time. The implementation of the report's recommendations should be undertaken within an overall framework that accords priority to the basic quality of the data and the fundamentals of dissemination, then to significant enhancements that are achievable in the short term, while laying the groundwork for other long-term improvements.




ASSESSING THE RELIABILITY OF COMPLEX MODELS
Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification
NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES

ASSESSING THE RELIABILITY OF COMPLEX MODELS: MATHEMATICAL AND STATISTICAL FOUNDATIONS OF VERIFICATION, VALIDATION, AND UNCERTAINTY QUANTIFICATION (DEPS, 2012)
Advances in computing hardware and algorithms have dramatically improved the ability to simulate complex processes computationally. Today's simulation capabilities offer the prospect of addressing questions that in the past could be addressed only by resource-intensive experimentation, if at all. *Assessing the Reliability of Complex Models* recognizes the ubiquity of uncertainty in computational estimates of reality and the necessity for its quantification. As computational science and engineering have matured, the process of quantifying or bounding uncertainties in a computational estimate of a physical quality of interest has evolved into a small set of interdependent tasks: verification, validation, and uncertainty of quantification (VVUQ). In recognition of the increasing importance of computational simulation and the increasing need to assess uncertainties in computational results, the National Research Council was asked to study the mathematical foundations of VVUQ and to recommend steps that will ultimately lead to improved processes. *Assessing the Reliability of Complex Models* discusses changes in education of professionals and dissemination of information that should enhance the ability of future VVUQ practitioners to improve and properly apply VVUQ methodologies to difficult problems, enhance the ability of VVUQ customers to understand VVUQ results and use them to make informed decisions, and enhance the ability of all VVUQ stakeholders to communicate with each other. This report is an essential resource for all decision and policy makers in the field, students, stakeholders, UQ experts, and VVUQ educators and practitioners.

Report of a Workshop on Big Data

Committee for Science and Technology Challenges to
U.S. National Security Interests

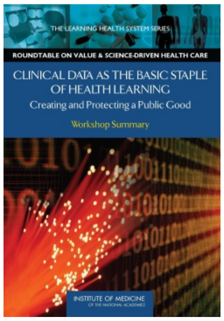
Division on Engineering and Physical Sciences

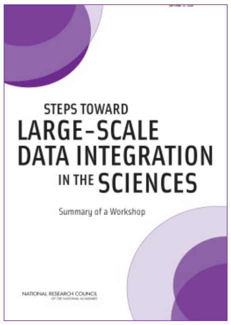
NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

BIG DATA: A WORKSHOP REPORT (DEPS, 2012)
In 2012, the Defense Intelligence Agency (DIA) approached the National Research Council's TIGER standing committee and asked it to develop a list of workshop topics to explore the impact of emerging science and technology. From the list of topics given to DIA, three were chosen to be developed by the Committee for Science and Technology Challenges to U.S. National Security Interests. The first in a series of three workshops was held on April 23-24, 2012. This report summarizes that first workshop which explored the phenomenon known as big data. The objective for the first workshop is given in the statement of task, which explains that that workshop will review emerging capabilities in large computational data to include speed, data fusion, use, and commodification of data used in decision making. The workshop will also review the subsequent increase in vulnerabilities over the capabilities gained and the significance to national security. The committee devised an agenda that helped the committee, sponsors, and workshop attendees probe issues of national security related to so-called big data, as well as gain understanding of potential related vulnerabilities. The workshop was used to gather data that is described in this report, which presents views expressed by individual workshop participants. *Big Data: A Workshop Report* is the first in a series of three workshops, held in early 2012 to further the ongoing engagement among the National Research Council's (NRC's) Technology Insight-Gauge, Evaluate, and Review (TIGER) Standing Committee, the scientific and technical intelligence (S&TI) community, and the consumers of S&TI products.

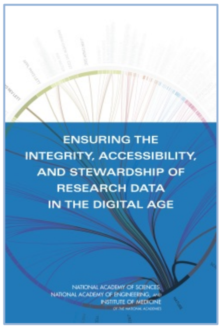
CLINICAL DATA AS THE BASIC STAPLE OF HEALTH LEARNING: CREATING AND PROTECTING A PUBLIC GOOD: WORKSHOP SUMMARY (IOM, 2010)
Successful development of clinical data as an engine for knowledge generation has the potential to transform health and health care in America. As part of its Learning Health System Series, the Roundtable on Value & Science-Driven Health Care hosted a workshop to discuss expanding the access to and use of clinical data as a foundation for care improvement.

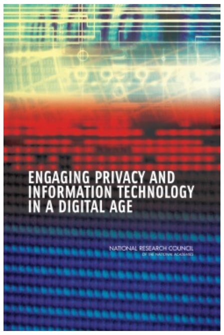

THE LEARNING HEALTH SYSTEM SERIES
ROUNDTABLE ON VALUE & SCIENCE-DRIVEN HEALTH CARE
CLINICAL DATA AS THE BASIC STAPLE OF HEALTH LEARNING
Creating and Protecting a Public Good
Workshop Summary
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES



STEPS TOWARD LARGE-SCALE DATA INTEGRATION IN THE SCIENCES: SUMMARY OF A WORKSHOP (DEPS/PGA, 2010)
Steps Toward Large-Scale Data Integration in the Sciences summarizes a National Research Council (NRC) workshop to identify some of the major challenges that hinder large-scale data integration in the sciences and some of the technologies that could lead to solutions. The workshop was held August 19-20, 2009, in Washington, D.C. The workshop examined a collection of scientific research domains, with application experts explaining the issues in their disciplines and current best practices. This approach allowed the participants to gain insights about both commonalities and differences in the data integration challenges facing the various communities. In addition to hearing from research domain experts, the workshop also featured experts working on the cutting edge of techniques for handling data integration problems. This provided participants with insights on the current state of the art. The goals were to identify areas in which the emerging needs of research communities are not being addressed and to point to opportunities for addressing these needs through closer engagement between the affected communities and cutting-edge computer science.

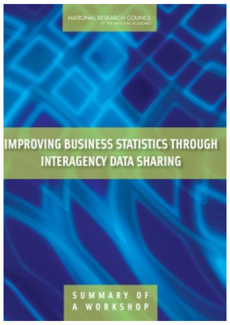
ENSURING THE INTEGRITY, ACCESSIBILITY, AND STEWARDSHIP OF RESEARCH DATA IN THE DIGITAL AGE (PGA/IOM, 2009)
As digital technologies are expanding the power and reach of research, they also create complex issues: complications in ensuring the validity of research data; standards that do not keep pace with the high rate of innovation; restrictions on data sharing that reduce the ability of researchers to verify results and build on previous research; and huge increases in the amount of data being generated, creating severe challenges in preserving that data for long-term use. This report recommends that all researchers receive appropriate training in the management of research data and calls on them to make all research data, methods, and other information underlying results publicly accessible in a timely manner. The stewardship of research data is also a critical long-term task for the research enterprise and its stakeholders. In 2010, committee co-chairs and staff engaged scientists and information professionals on these issues at the annual meetings of the AAAS and the Coalition on Networked Information, as well as the NAS E-Journal Summit and the International Association of Scientific and Technological Libraries.

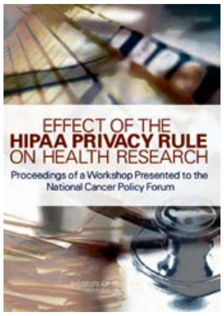

ENSURING THE
INTEGRITY, ACCESSIBILITY,
AND STEWARDSHIP OF
RESEARCH DATA
IN THE DIGITAL AGE
NATIONAL ACADEMY OF SCIENCES
NATIONAL ACADEMY OF ENGINEERING, and
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES



ENGAGING PRIVACY AND INFORMATION TECHNOLOGY IN A DIGITAL AGE (DEPS, 2007)
Privacy is a growing concern in the United States and around the world. The spread of the Internet and the seemingly boundaryless options for collecting, saving, sharing, and comparing information trigger consumer worries. Online practices of business and government agencies may present new ways to compromise privacy, and e-commerce and technologies that make a wide range of personal information available to anyone with a Web browser only begin to hint at the possibilities for inappropriate or unwarranted intrusion into our personal lives. *Engaging Privacy and Information Technology in a Digital Age* presents a comprehensive and multidisciplinary examination of privacy in the information age. It explores such important concepts as how the threats to privacy evolving, how can privacy be protected and how society can balance the interests of individuals, businesses and government in ways that promote privacy reasonably and effectively? This book seeks to raise awareness of the web of connectedness among the actions one takes and the privacy policies that are enacted, and provides a variety of tools and concepts with which debates over privacy can be more fruitfully engaged. *Engaging Privacy and Information Technology in a Digital Age* focuses on three major components affecting notions, perceptions, and expectations of privacy: technological change, societal shifts, and circumstantial discontinuities. This book will be of special interest to anyone interested in understanding why privacy issues are often so intractable.

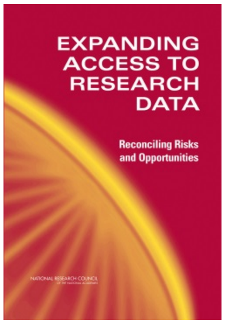
IMPROVING BUSINESS STATISTICS THROUGH INTERAGENCY DATA SHARING: SUMMARY OF A WORKSHOP (DBASSE, 2006)
U.S. business data are used broadly, providing the building blocks for key national—as well as regional and local—statistics measuring aggregate income and output, employment, investment, prices, and productivity. Beyond aggregate statistics, individual- and firm-level data are used for a wide range of microanalyses by academic researchers and by policy makers. In the United States, data collection and production efforts are conducted by a decentralized system of statistical agencies. This apparatus yields an extensive array of data that, particularly when made available in the form of microdata, provides an unparalleled resource for policy analysis and research on social issues and for the production of economic statistics. However, the decentralized nature of the statistical system also creates challenges to efficient data collection, to containment of respondent burden, and to maintaining consistency of terms and units of measurement. It is these challenges that raise to paramount importance the practice of effective data sharing among the statistical agencies. With this as the backdrop, the Bureau of Economic Analysis (BEA) asked the Committee on National Statistics of the National Academies to convene a workshop to discuss interagency business data sharing. The workshop was held October 21, 2005. This report is a summary of the discussions of that workshop. The workshop focused on the benefits of data sharing to two groups of stakeholders: the statistical agencies themselves and downstream data users.


NATIONAL RESEARCH COUNCIL
IMPROVING BUSINESS STATISTICS THROUGH
INTERAGENCY DATA SHARING
SUMMARY OF
A WORKSHOP


EFFECT OF THE
HIPAA PRIVACY RULE
ON HEALTH RESEARCH
Proceedings of a Workshop Presented to the
National Cancer Policy Forum

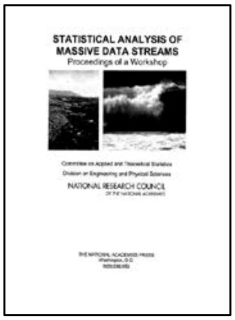
EFFECTS OF THE HIPAA PRIVACY RULE ON HEALTH RESEARCH: PROCEEDINGS OF A WORKSHOP PRESENTED TO THE NATIONAL CANCER POLICY FORUM (IOM, 2010)
These proceedings of a workshop presented to the National Cancer Policy Forum are the result of the Forum decision to examine the HIPAA Privacy Rule and its effects on health research and privacy. In preparation for the June meeting, the Forum invited a group of speakers balanced among those from all sectors, private academic, advocacy, industry, and public, including those who were focused on protecting the privacy of health information, those who had participated in preparation of the Privacy Rule, those who were responsible for both funding and carrying out health research, and those who had studied the Privacy Rule and recommended changes. Also, the North American Association of Central Cancer Registries carried out a short, two-question survey of its members enquiring about HIPAA Privacy Rule generated problems in cancer registry research, and the results of this brief preliminary survey were presented to the workshop.

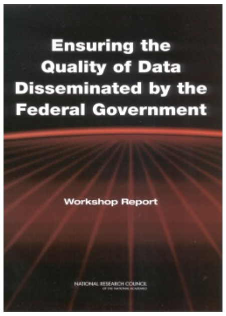
STRATEGIES FOR PRESERVATION OF AND OPEN ACCESS TO SCIENTIFIC DATA IN CHINA: SUMMARY OF A WORKSHOP (PGA, 2006)
Preservation of and open access to digital scientific resources are essential to global research, yet the challenges in storing and maintaining access to these collections are substantial. China faces major hurdles in this regard. A workshop held in June 2004 in Beijing convened scientific information managers, digital archiving experts, national science policy and funding officials, and representatives of development organizations to explore the scientific and technical, legal and policy, institutional and economic, and management aspects of creating sustainable and accessible archives of digital health and environmental data in China.


EXPANDING ACCESS TO RESEARCH DATA
Reconciling Risks and Opportunities

EXPANDING RESEARCH DATA: RECONCILING RISKS AND OPPORTUNITIES (DBASSE, 2005)
Policy makers need information about the nation ranging from trends in the overall economy down to the use by individuals of Medicare in order to evaluate existing programs and to develop new ones. This information often comes from research based on data about individual people, households, and businesses and other organizations, collected by statistical agencies. The benefit of increasing data accessibility to researchers and analysts is better informed public policy. To realize this benefit, a variety of modes for data access including restricted access to confidential data and unrestricted access to appropriately altered public-use data must be used. The risk of expanded access to potentially sensitive data is the increased probability of breaching the confidentiality of the data and, in turn, eroding public confidence in the data collection enterprise. Indeed, the statistical system of the United States ultimately depends on the willingness of the public to provide the information on which research data are based. *Expanding Access to Research Data* issues guidance on how to more fully exploit these tradeoffs. The panel's recommendations focus on needs highlighted by legal, social, and technological changes that have occurred during the last decade.

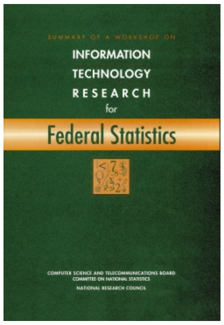
STATISTICAL ANALYSIS OF MASSIVE DATA STREAMS: PROCEEDINGS OF A WORKSHOP (DEPS, 2004)
Massive data streams, large quantities of data that arrive continuously, are becoming increasingly commonplace in many areas of science and technology. Consequently development of analytical methods for such streams is of growing importance. To address this issue, the National Security Agency asked the NRC to hold a workshop to explore methods for analysis of streams of data so as to stimulate progress in the field. This report presents the results of that workshop. It provides presentations that focused on five different research areas where massive data streams are present: atmospheric and meteorological data; high-energy physics; integrated data systems; network traffic; and mining commercial data streams. The goals of the report are to improve communication among researchers in the field and to increase relevant statistical science activity.

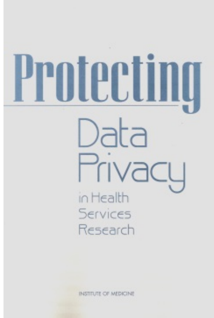

STATISTICAL ANALYSIS OF MASSIVE DATA STREAMS
Proceedings of a Workshop
NATIONAL RESEARCH COUNCIL


Ensuring the Quality of Data Disseminated by the Federal Government
Workshop Report

ENSURING THE QUALITY OF DATA DISSEMINATED BY THE FEDERAL GOVERNMENT: WORKSHOP REPORT (PGA, 2003)
The National Academies Science, Technology, and Law Program convened three workshops focusing on specific aspects of OMB's "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies." The workshops were intended to assist the agencies in developing their agency-specific implementation guidelines. This workshop report details the approaches agencies are considering using to implement the guidelines.

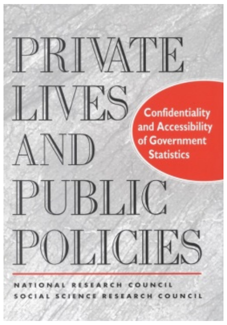
SUMMARY OF A WORKSHOP ON INFORMATION TECHNOLOGY RESEARCH FOR FEDERAL STATISTICS (DEPS/DBASSE, 2000)
Part of an in-depth study of how information technology research and development could more effectively support advances in the use of information technology (IT) in government, Summary of a Workshop on Information Technology Research for Federal Statistics explores IT research opportunities of relevance to the collection, analysis, and dissemination of federal statistics. On February 9 and 10, 1999, participants from a number of communities - IT research, IT research management, federal statistics, and academic statistics - met to identify ways to foster interaction among computing and communications researchers, federal managers, and professionals in specific domains that could lead to collaborative research efforts. By establishing research links between these communities and creating collaborative mechanisms aimed at meeting relevant requirements, this workshop promoted thinking in the computing and communications research community and throughout government about possibilities for advances in technology that will support a variety of digital initiatives by the government.





PROTECTING DATA PRIVACY IN HEALTH SERVICES RESEARCH (DBASSE, 2000)
The need for quality improvement and for cost saving are driving both individual choices and health system dynamics. The health services research that we need to support informed choices depends on access to data, but at the same time, individual privacy and patient-health care provider confidentiality must be protected.

IMPROVING ACCESS TO AND CONFIDENTIALITY OF RESEARCH DATA: REPORT OF A WORKSHOP (DBASSE, 2000)
Improving Access to and Confidentiality of Research Data summarizes a workshop convened by the Committee on National Statistics (CNSTAT) to promote discussion about methods for advancing the often conflicting goals of exploiting the research potential of microdata and maintaining acceptable levels of confidentiality. This report outlines essential themes of the access versus confidentiality debate that emerged during the workshop. Among these themes are the tradeoffs and tensions between the needs of researchers and other data users on the one hand and confidentiality requirements on the other; the relative advantages and costs of data perturbation techniques (applied to facilitate public release) versus restricted access as tools for improving security; and the need to quantify disclosure risks--both absolute and relative--created by researchers and research data, as well as by other data users and other types of data.


PRIVATE LIVES AND PUBLIC POLICIES
Confidentiality and Accessibility of Government Statistics
NATIONAL RESEARCH COUNCIL
SOCIAL SCIENCE RESEARCH COUNCIL

PRIVATE LIVES AND PUBLIC POLICY: CONFIDENTIALITY AND ACCESSIBILITY OF GOVERNMENT STATISTICS (DBASSE, 1993)
Americans are increasingly concerned about the privacy of personal data--yet we demand more and more information for public decision making. This volume explores the seeming conflicts between privacy and data access, an issue of concern to federal statistical agencies collecting the data, research organizations using the data, and individuals providing the data. A panel of experts offers principles and specific recommendations for managing data and improving the balance between needed government use of data and the privacy of respondents. The volume examines factors such as the growth of computer technology that are making confidentiality an increasingly critical problem. The volume explores how data collectors communicate with data providers, with a focus on informed consent to use data, and describes the legal and ethical obligations data users have toward individual subjects as well as toward the agencies providing the data. In the context of historical practices in the United States, Canada, and Sweden, statistical techniques for protecting individuals' identities are evaluated in detail. Legislative and regulatory restraints on access to data are examined, including a discussion about their effects on research.

ABOUT THE GOVERNMENT-UNIVERSITY-INDUSTRY RESEARCH ROUNDTABLE (GUIRR)

GUIRR’s formal mission, revised in 1995, is “to convene senior-most representatives from government, universities, and industry to define and explore critical issues related to the national and global science and technology agenda that are of shared interest; to frame the next critical question stemming from current debate and analysis; and to incubate activities of on-going value to the stakeholders. This forum will be designed to facilitate candid dialogue among participants, to foster self-implementing activities, and, where appropriate, to carry awareness of consequences to the wider public.”

*The reports listed do not include all National Academies’ reports on topics related to big data. To find more on this topic or browse other National Academies reports go to: www.nationalacademies.org.

Government | University | Industry
RESEARCH ROUNDTABLE