RTI INTERNATIONAL
3040 E. Cornwallis Road ▪ PO Box 12194 ▪ Research Triangle Park, NC 27709-2194 ▪ USA
Telephone 919.541.6000 ▪ Fax 919.541.5985 ▪ www.rti.org

# Response to Request for Input (RFI)—National Big Data R&D Initiative

**Paul P. Biemer, Alan R. Blatecky, Alan F. Karr (RTI International)**

**Introduction.** Data are everywhere. They being produced by virtually all scientific, educational, governmental, societal and commerce organizations, and being generated by surveys, mobile and embedded systems, sensors, observing systems, scientific instruments, publications, experiments, simulations, evaluations and analyses. In this response, we highlight three central Big Data issues—curation, data quality, and privacy/confidentiality, especially the need to balance confidentiality with usage.

**Experience.** RTI International is a world leader in the collection, integration, analysis, interpretation and visualization of complex data, in contexts ranging from longitudinal surveys to physical measurements to social media to program evaluation, and beyond. The Institute's interests and activities span the full pathway from data to information to knowledge to decisions and practice.

**Curation.** Data that are useful for analysis must be available in ways that allow them to be searched, analyzed, and manipulated, across disciplines, domains, and geographic boundaries. Unless data are adequately curated, it will be very difficult for science, government or business, to use—let alone re-use—them. Federal agencies spend a considerable amount of funding and efforts generating data, but much less time and effort ensuring that the data will be useful to others. Data that are not adequately described with metadata will be very difficult to locate and interpret. Data not identified by permanent identifiers such as an OID (Object Identifier) cannot be effectively registered. Absent information on how the data have been modified or altered over time (addressing issues of provenance and versioning), problems of reproducibility may be insurmountable.

**Comment [BPP1]:** And interpret

**Comment [BPP2]:** Do you mean file name or field identifier or what?

The point is this: if data are not adequately curated, it is questionable whether they should have been generated in the first place, as it will be very difficult to use them in the future or in conjunction with other data. Since a major assumption of the National Big Data R&D Initiative is re-use of all types data, it is strongly recommended that Federal agencies specifically, and more forcefully, address data curation issues (e.g.,metadata, provenance, versioning, permanent identifiers) in their Big Data programs, projects and activities.

**Data Quality**. Some of the errors that plague Big Data are well-known. As they are created, Big Data are often selective, incomplete and erroneous. New errors can be introduced downstream as the data are cleaned, integrated, transformed, and analyzed. Data munging (wrangling) steps comprise 50-80% of the work involved in getting a dataset ready for analysis, but these steps can add both variable and systematic errors to the data, resulting in unreliability, invalidity, and biased or even incorrect inference. To illustrate, Big Data are often portrayed as allowing detailed examination of extremely rare phenomena. But, even very small false identification rates (on the order of one hundredth of one percent or less) can contaminate these data, rendering them essentially useless for analytical purposes. Even in the absence of errors, Big Data pose new challenges to inference, such as noise accumulation, spurious correlations, and incidental endogeneity (Fan, Han, and Liu, Challenges of Big Data Analysis, *National Science Review*, doi:10.1093/nsr/nwt032, 2014). Data errors can exacerbate these problems.

To illustrate, in high-dimensional data, spurious correlations are not rare: in simulated 800-deminsional data having *no correlations* and relatively small sample sizes, the analyst has a 50% chance of observing an absolute correlation that exceeds 0.4. In the presence of data errors, these risks increase substantially even for sample sizes in the hundreds of thousands. Systematic errors common to two variables *magnify* the observed correlations (Biemer and Trewin, A Review of Measurement Error Effects on the Analysis of Survey Data. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), *Survey Measurement and Process Quality*, New York: Wiley and Sons, 603-632, 1997), thereby exacerbating the problem. It is incumbent upon the Federal government as a major user and supplier of Big Data to be cognizant of the issues associated with Big Data errors and their consequences. Failure to recognize the limitations of errors in these data will lead to false discoveries, failed policies and errant decisions.

**Balancing Privacy, Confidentiality and Data Availability.** For many years, official statistics agencies such as the Census Bureau have struggled to balance legal and ethical obligations to protect confidentiality of their datasets, yet make data available for research and policy purposes. Today, several developments are converging that make the balancing increasingly challenging. First is the scale of the data: the more that is collected, the easier it is to identify records, especially by linking to other datasets. Second is the increasing extent to which data subjects are unaware that data about them are being collected, to the point that extant foundations of informed consent being questioned. Repurposing of data creates further issues: protection in one context does not imply protection in all contexts. Finally, "big computation" threatens all existing methods of disclosure limitation, by casting disclosure risk as a problem of computational complexity—not whether confidentiality can be broken, but how much computational effort is needed to do so.

Despite changing attitudes regarding privacy, not protecting data will not become an acceptable, or even the prevailing solution. Indeed, in light of near-daily reports of breaches in the security of corporate information systems, there may a backlash. For multiple reasons, every government data collection in every country in the world is facing severe problems with declining response rates. Not just new techniques, but even new abstractions of disclosure risk, disclosure harm and data utility are needed.

Adding to these pressures are recent mandates (some unfunded) to make research data available, especially given the seemingly excessive rate of "failure to replicate." Failure to pay proper attention to curation, data quality and privacy and confidentiality will attenuate or even negate entirely the benefits of making data available. It is all too easy to make data available but unusable.

**Conclusion.** It is important to realize that data are merely the first step—albeit an essential one—in the pathway from data to information to knowledge to decisions. The value of data is not inherent, but rather is derived from the inferences, insights and decisions drawn from the data. Today, principled ways of trading off cost and data quality are only emerging; the Federal government can play a leadership role by moving attention to decision quality. More data are not necessarily better, if they are not properly curated, their quality is not understood, or they are not protected: bigger is not necessarily better!

Nor do Big Data reduce, let alone eliminate, the crucial need to characterize and quantify uncertainties in inferences. Indeed, as noted above, Big Data can make the need more urgent. There are serious risks of a society drowning in data and starved for principled research and sound policy derived from the data. Federal government initiatives may be the only way to prevent this from happening.