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

https://www.nitrd.gov/bigdata/rfi/02102014.aspx

I. **RFI Responder**: Ivo D. Dinov, PhD, Associate Professor and Director, Statistics Online Computational Resource (SOCR), University of Michigan, Ann Arbor, MI 48109 (734) 615-5087, http://umich.edu/~dinov, statistics@umich.edu.

II. **Responder Big Data Experience**: In my ongoing work, I have done research [1,2], designed innovative data analytics services [3,4], developed distributed software tools [5,6], and integrated the computational, theoretical and applied components of Big Data Science and Infrastructure into the graduate curriculum [7,8]. As an example, a recent article, "*The perfect neuroimaging-genetics-computation storm: collision of petabytes of data, millions of hardware devices and thousands of software tools*" [9], illustrated that the Kryder's law for exponential increase of the volume of data is real. Although Moore's law for exponential increase of computational power (transistor capacity) and Kryder's laws indicate similar exponential expansion over time, the rate of increase of data volume is significantly higher than our ability to manage, process and interpret that data we collect. Relative to the Moore's law, the exponential parameter in the Kryder's law for data increase is bigger.

III. **Comments on Initial Framework**: Although the National Big Data R&D Initiative Framework [10] pinpoints some significant Big Data challenges and opportunities, it falls short of what is necessary to ensure effective, coordinated, agile and sustainable model for proactively addressing the needs for managing the avalanche of data in all aspects of daily living. The 3 most notable shortcomings of this initial plan include *crowd-source engagement*, *agile development* and *training sustainability*. Current and future innovations in Big Data analytics are most likely going to come from disparate resources, small-group initiatives, open-source/open-science community and truly transdisciplinary interactions, not from Big-Business, Big-Academy, or Big-Government. We need a concerted effort to harness the ingenuity of the broader community using large-number of smaller-budget, rapid-development, high-risk, product-oriented projects (including funding mechanisms). The era of Big Data analytics is here and we need to realize continuous-development, rapid agile prototyping, and experience-based (e.g., evidence-based) redesign are the new normal for all innovation, including basic science, computational modeling, applied research or studies of system complexity.

IV. **Gaps and Barriers**: (1) Gaps in technological skills and expertise to conduct high-throughput Big Data analysis, (2) lack of open, interactive and interoperable learning

---

[1] http://dx.doi.org/10.3389/fninf.2014.00041
[2] http://www.socr.umich.edu/people/dinov/publications.html
[3] http://socr.umich.edu/HTML5/
[4] http://socr.umich.edu/HTML5/BrainViewer/
[5] http://pipeline.loni.usc.edu/get-started/acknowledgmentscredits/
[6] http://distributome.org
[7] http://www.socr.umich.edu/people/dinov/SMHS_Courses.html
[8] http://www.socr.umich.edu/people/dinov/courses.html
[9] http://dx.doi.org/10.1007/s11682-013-9248-x
[10] https://www.nitrd.gov/nitrdgroups/images/0/09/Federal_BD_R&D_Thrusts_and_Priority_Themes.pdf

resources, (3) challenges of sharing tools, materials and activities across different disciplines, (4) considerable discipline-specific knowledge boundaries, (5) challenges of fusion of qualitative (e.g., study design) and quantitative (e.g., analytic modeling) Big Data methods, (6) availability of integrated and interoperable Big Data resources. For example, a SOCR study is currently looking at healthcare data (e.g., CMS, WHO), demographic trends (e.g., Census), economics factors (e.g., BLS) and social trends (e.g., Web-traffic, Tweeter) aiming to determine the associations and organization of data elements from many hundreds of data elements in the presence of incongruent sampling, complex, incomplete and heterogeneous data from different sources.

V. **High-Impact Ideas**: Decisively, there are 4 specific and complementary funding directions that could significantly impact the process of extracting information from Big Data, translating that information to knowledge, which can in turn guide our actions: (1) Enforce (for real) open-science principles; (2) Engage and actively participate (e.g., fund) non-traditional high-risk/high-potential-impact studies; (3) Adapt to rapid agile development, testing, redesign, productization and utilization of data, tools, services and architecture; (4) Redesign the data science curricula (from high-school to doctoral level training). Big Data is incredibly powerful, but its *Achilles hill* is time. Its value is in the moment and its importance decreases exponentially with time [11], which makes the effective and rapid response to collected data critically important.

VI. **Investment Targets**: We need to demand open-science, reduce barriers to sharing of data, protocols, and tools, cast a wider funding net and diversify the scope of research and development efforts, and finally, significantly overhaul Data Science education.

VII. **Partnerships**: Enabling new public-private-government partnerships is easier said than done. There are many complexities, conflicts of interests, and divergent visions that affect the agreements establishing effective cooperations between organizations and independent institutions. Long- and short-term goals can collide and misalignment of interests, bottom-lines and scope of impact can bifurcate causing frictions and difficulties in managing extremely diverse partnerships. Albeit not forced, partnerships should be encouraged and facilitated as necessary. The best scientific discoveries are advanced and popularized not by regulations, but by loose interactions, open-sharing and collaboration.

VIII. **Justification of Contribution**: There are 2 ways to deal with the influx of significant disruptive technologies and – reactive response (passive) or proactive action. Government institutions and regulators, funding agencies and organizations involved in generating, aggregating, processing, analyzing, interpreting or managing large, incongruent, heterogeneous and complex data may choose to lead or follow the Big Data revolution. As engaged citizens, energetic researchers and diligent parents, we need to ensure our National priorities position us well for riding the Big Data wave, as opposed to training in its wake. The National Big Data R&D Initiative and the Networking and Information Technology Research and Development group are in a powerful position to gently guide, actively engage, and significantly stimulate future innovation in Big Data science, engineering, applications, knowledge and action.

[11] http://www.aaas.org/news/big-data-blog-part-v-interview-dr-ivo-dinov