POSITION PAPER: Senate Bill No. 688 "Big Data Act"

Prepare for the Senate Committee on Science and Technology

Prepared by Jose Ramon G. Albert, PhD[1]

18 September 2018

1. **The legislature has recognized the importance of big data, i.e., digital data collected from the use of mobile phones, the internet, social media, search engines, digital trade transactions, and sensors.** Big data has arisen from the growing use of technology by our citizenry (and people across the world). As people get to use mobile phones more and more, and use the internet to search for information, and connect with other people, more data is being captured, produced, stored, accessed, analyzed, archived, and re-analyzed, and at an exponential pace. The proposed "Big Data Act" in the Senate (and its counterpart House bill), provides for the establishment of a Big Data Center to serve as the country's research body examining big data. The Big Data Center is proposed to be a unit attached to Philippine Statistical Research and Training Institute, an attached agency of the National Economic and Development Authority. The Big Data Center is being earmarked P200 million an initial operating fund.

2. **While this legislative measure provides a concrete mechanism for testing the use of big data for development, the current bill appears to be putting too much emphasis on technologies that are fast evolving, and on related but not equivalent technological issues for big data analytics.** Retrieving and examining big data streams for analytics require adequate technological infrastructure, both hardware and software. Many data mining tools are neither suitable nor efficiently used for large datasets in a sequential computer. A critical issue that the Big Data Center will require is thus better ICT infrastructure to download these big data sources (bandwidth), as well as to catalogue, organize and process the complex collage of data in a sufficiently timely manner. Recent practice in big data analytics involves utilizing a cluster of physical computers running a framework tool such as Hadoop-MapReduce[2], and/or use cloud computing/processing. The availability of interfaces by some statistical packages such as open sourced R to Hadoop or MapReduce for most used statistical platforms has, however, significantly contributed to the use of big data analytics. Further, the cloud has also emerged as an ideal computing environment for big data. On the infrastructure side, cloud computing, through "infrastructure as a service" in a public cloud or "platform as a service" in a private cloud, provides options for accessing and managing very large data sets as well as for supporting powerful infrastructure elements at a relatively-low cost. Further, an increasing number of "software as a service" in a hybrid cloud are also capable of performing the processing and data integration tasks. **The current draft bill defines various divisions in the Big Data Center**, e.g., Open Data, Partnership, Data Analytics and Storage, Privacy and Data Anonymity, **without seeing how the Big Data Center fits into the current PSRTI units**, i.e., Administration, Training and Research Divisions. The Big Data

---

[1] Senior Research Fellow, Philippine Institute for Development Studies (PIDS)

[2] Hadoop is an open-source software project, managed by the Apache Software Foundation, targeted at supporting the execution of data-oriented application on clusters of generic hardware. The Hadoop project is comprised of four modules: Hadoop Distributed File System (HDFS), the MapReduce model, and Hadoop Common and YARN. The first two modules are critical: HDFS allows data to be stored in an easily accessible format, across a large number of linked storage devices, while MapReduce carries out two basic operations - reading data from the database and putting it into a format suitable for analysis (map); and performing mathematical operations (reduce). Hadoop Common provides the tools (in Java) needed for s user's computer systems (e.g., Windows, Unix, MacOS, *etc.*) to read data stored under the Hadoop file system, while YARN manages resources of the systems storing the data and running the analysis.

Center should be somehow working in tandem with the Research Division to look into obtaining insights from both traditional and new data sources.

- "Open Data" is not quite the same as "Big Data"; the Open Data Initiative is currently a function of a unit in the Department of Information and Communications Technology. Does the current bill want to move this function to the Big Data Center, or focus its efforts on Big Data Analytics?
- Legal protocols will certainly be required to access big data holdings for development purposes (without infringing on data privacy), as well as to prevent misuse of big data. However, putting Data Analytics and Storage in the same rank as Partnership and Privacy (and Open Data) is not technically sound given the partnership is a modality to secure some data holdings (especially in the private sector), while privacy is a functionality to be ensured as one goes about analytics.

3. **If Congress considers it best to establish this new body, it is important to keep the structure at the Big Data Center relatively loose to focus on generating insights on development issues** such as use of social media data (to monitor inflation with social media conversations as has been done in the UN Global Pulse Jakarta, or possibly other concerns such as tourism data), or satellite images (to look into alternative data sources of rice production, or making use of luminosity to generate extra proxy indicators on poverty and welfare), or streams of data from GPS (to monitor traffic). These big data research projects will need proofs of concept and could be approved by the Technical Advisory Committee that provides guidance on the work of the Big Data Center.

4. **While government needs to see the importance of examining big data, it should also understand that digital traces can be incomplete,** and even obscure information, especially if big data are examined inadequately, with bias and with malice. Complementing traditional with innovative data sources to portray conditions regarding development issues should thus be undertaken with much care and preparation to ensure that resulting insights gained are reliable. The big data sources can result in a messy collage of data points. Big data can be categorized largely into three main sources: human-sourced information (e.g., social networks), process-mediated data (e.g., search engines, commercial transactions especially digital transactions), machine-generated data (e.g., mobile phone location). All of these voluminous, fast-paced, and complex data, however, are often by-products of transactions from hyper-connectivity, and as such, do not necessarily involve a target population, much unlike traditional data sources of official statistics (such as censuses and surveys). Various tools have to be used to assess the veracity of big data. Bias does not necessarily disappear in voluminous big data. While statistics using big data may not be completely accurate, they are often viewed as "good enough" and at near real time. The gains in velocity (and cost) in yielding statistics from big data sources, as well as the complexity and the sheer size of big data however requires a different type of data processing and analytic tools from those used for "small data" to yield statistics that are fit for use. New business models must be developed even by government to leverage data resources, human talent, and decision-making capacity. Institutional frameworks and arrangements, such as public private partnerships and linkages with various institutions engaged in data science need to be developed and enhanced by the proposed Big Data Center (once it is established) to further obtain insights about development.

DISCLAIMER: The views expressed herein do not necessarily reflect those of the PIDS. This statement draws largely on the empirical evidence summarized in the Philippine Institute for Development Studies Policy Note Number 14-04 https://dirp4.pids.gov.ph/webportal/CDN/PUBLICATIONS/pidspn1404_rev2.pdf, the UN ESCAP Tech Monitor article http://techmonitor.net/tm/images/d/d8/17apr_jun_sf1.pdf and the ADB blog on big data https://blogs.adb.org/blog/big-data-can-transform-sdg-performance-here-s-how