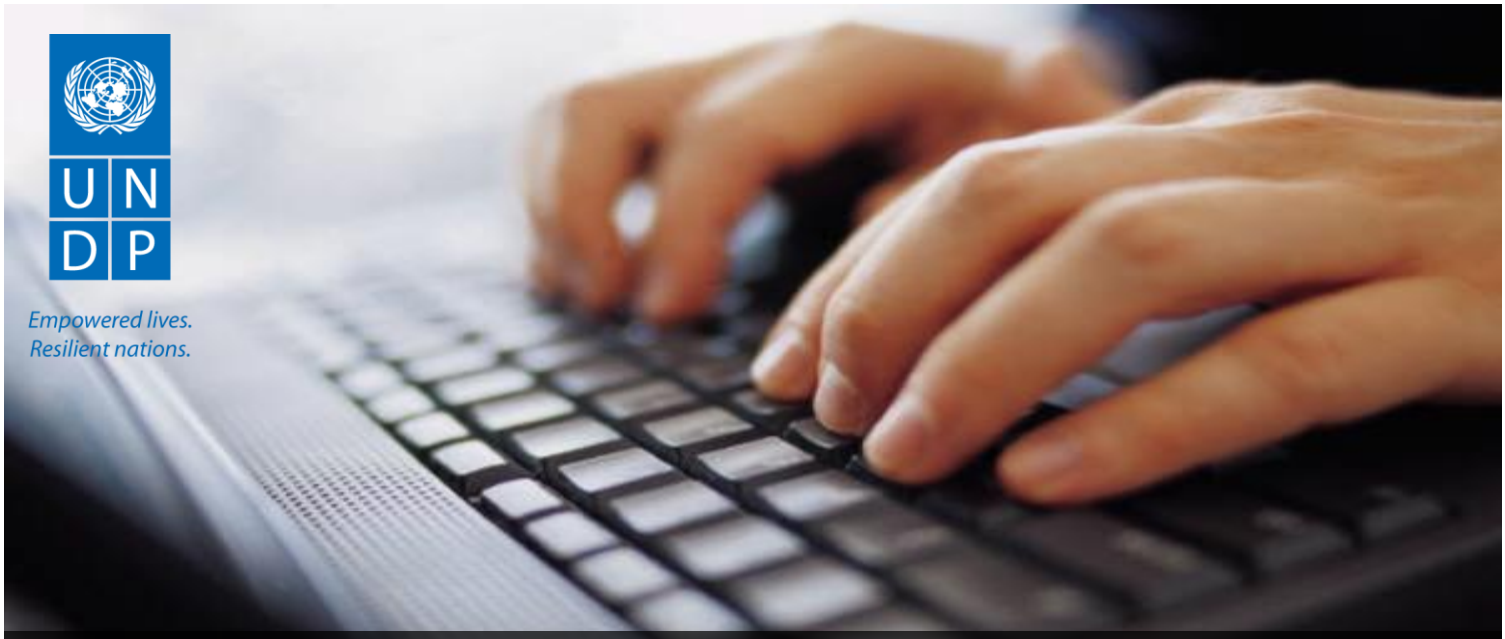


Big Data for Development in China

Abstract

The purpose of this paper is to draw the attention of development practitioners to the potential of Big Data for Development, i.e. identification of sources of Big Data relevant to policy and planning of development programmes, in China.

The paper reviews the existing literature on Big Data for Development. After a brief overview of the background and key concepts related to Big Data for Development, the paper argues that there is a case for development practitioners to explore the potential of Big Data for Development in China.

It recommends two levels of Big Data for Development work for development practitioners to engage with in China. The first level of work is related to creating an enabling environment for Big Data for Development. The second level of work concerns tackling particular development challenges with the Big Data approach. As illustrative examples, the paper discusses how Big Data could be used to promote sustainable e-waste disposal practices, improve productivity of the public sector, understand socioeconomic trend, map poverty, improve urban transport planning, and identify pollution hotspots in cities.

The paper also discusses the challenges that China faces in Big Data for Development application on 3 fronts – data, analytical, and operational/systemic. In particular, the issue of data protection and privacy requires strong attention in order to protect the fundamental rights of the public.

The paper concludes by highlighting how Big Data could contribute to result-based management, the importance of development practitioners to understand its limitations, and its link to inclusive decision-making process.

Author: Jackie Hoi-Wai CHENG, National Economist of UNDP China (Contact e-mail: jackie.cheng@undp.org).

The author is grateful to Hannah Ryder, Giulio Quaggiotto, Ramya Gopalan, Peng WU, Benjamin William Mason, and Zenobia CHAN for their valuable comments. All errors remain those of the author.

Table of Contents

Introduction of Big Data

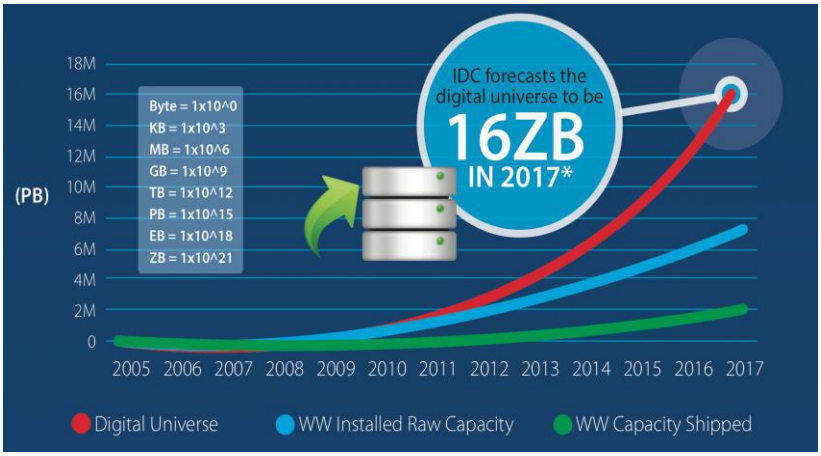
Across the world, Big Data has become one of the most discussed topics in recent years. Broadly speaking, it is an umbrella term that refers to the "explosion in the quantity and diversity of high frequency digital data."[1] Big Data has gathered attention in all places, ranging from private to public sector, from developing to developed world, and reaching every sector in the global economy.

The growth of interest in Big Data could be at least partly attributed to the so-called "data revolution" or "data deluge" that the world is experiencing. The size of the digital size is growing 40 percent a year into the next decade. It is estimated that digital data is doubling in size every 2 years, and by 2020 its size will reach 44,000 exabytes.[2] In other words, the digital universe will contain nearly as many digital bits as there are stars in the universe.[3]

While there are many definitions of Big Data, there is a general consensus that Big Data can be characterized by the 3 Vs, i.e. volume, velocity, and variety. It comes from the now widely-accepted definition by the IT research and advisory firm Gartner, which describes Big Data as "high volume, high velocity and/or high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making and process optimization." Another 2 Vs are also commonly discussed in the context of Big Data: veracity (or reliability) of the data, and volatility of data (e.g. rapid changes in the availability and validity of data).

Figure 1

Global digital data created and storage capacity



Source: Where in the World is Storage. International Data Corporation (IDC) Infographic.
*Source: http://www.emc.com/leadership/digital-universe/index.htm

Examples of Big Data sources include: administrative data (e.g. electronic medical records, insurance records, hospital visits, bank records), transactional data (e.g. credit card transactions, on-line transactions), sensor data (e.g. satellite imaging, climate sensors), tracking devices (e.g. tracking data from mobile phones, GPS), behavioral data (e.g. online searches), and opinion data (e.g. comments on social media such as Facebook, Twitter, Weibo).

1 UN Global Pulse. 2012. Big Data for Development: Opportunities and Challenges

2 According to Cisco, an exabyte has the capacity to hold over 36,000 years worth of High-Definition quality video. Source: http://blogs.cisco.com/news/the-dawn-of-the-zettabyte-era-infographic/

3 IDC. April 2014. Executive Summary of "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things". (sponsored by EMC)

Big Data for Development

Big Data has certainly captured the attention of the private sector. According to a 2011 McKinsey Global Institute report, it is estimated that Big Data can generate significant financial value across sectors.[4] For example, major applications of personal location data have the potential to create value of USD 100 billion or more for service providers alone in 10 years' time. Another example is the global manufacturing industry, where the use of Big Data could lead to up to 50 percent decrease in product development and assembly cost and up to 7 percent reduction in working capital.

Outside the private sector, there is also increasing evidence that shows Big Data can be applicable to the development context. Big data could be used to more precisely assess the poverty situation in rural areas, improve the productivity of the public sector, create a more responsive social protection system, strengthen the resilience of cities against climate change and provide information of higher granularity for urban transportation planning, amongst other applications.

Understanding that there is great potential for Big Data to contribute to international development, the concept of "Big Data for Development" has gradually emerged in the development community. The concept refers to the "identification of sources of Big Data relevant to policy and planning of development programmes" and is different from both "traditional" development data and the aforementioned Big Data used in the private sector. This developmental concept of Big Data is championed by the UN Global Pulse, which is an initiative launched by the UN Secretary-General in 2009 to leverage innovations in digital data, with rapid data collection and analysis to help decision-makers gain a real-time understanding of how crises impact vulnerable population.

Compared with their counterparts in the developed world, developing countries are in general less developed in terms of IT infrastructure, supporting services and human resources. Nevertheless, Big Data also has high relevance to the developing world. One key development is the rapid spread of mobile phone technology to billions of individuals over the past decade. By 2013, there were over 6.8 billion mobile-cellular subscriptions globally, with penetration rate at 89% in developing countries.[5] The UN Global Pulse calls this surge of mobile phone use the most significant event in the developing world since the decolonization movement and the Green Revolution.[6] As mobile phones now very often serve multiple purposes – ranging from communication to banking, from trading to data transfer – it provides real time insights into human behavior that were previously not attainable.

Context: Rising global volatility

The importance of Big Data for Development is further accentuated by the fact that the world is increasingly volatile, in terms of socioeconomic and weather conditions, among others.[7] This rising volatility is exemplified by the series of crises that have unraveled over the past several years, starting with the food and fuel crises during 2007-2008 which were followed by the Great Recession – the worst financial crisis since the Great Depression –

[4] McKinsey Global Institute. 2011. Big Data: The Next Frontier for Innovation, Competition and Productivity

[5] Source: ICT Facts and Figures: The World in 2013. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf

[6] UN Global Pulse. 2012. Big Data for Development: Opportunities and Challenges

[7] It should be noted that it was in the wake of the global financial crisis that the leaders of G-20 and the UN Secretary-General called for the establishment of the Global Pulse Initiative.

that started in the second half of 2008. A famine then hit the Horn of Africa in 2011. And even after almost 6 years since the start of the Global Financial Crisis (GFC) in 2008, the global economy still remains fragile, with multiple areas in the world marred by major instability.

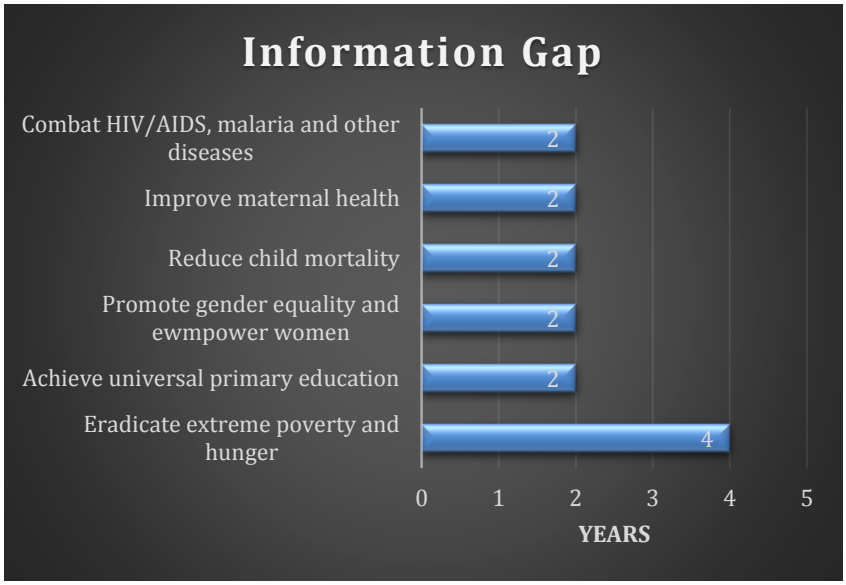
What is more, the sobering fact is that global volatility is not likely to improve in the foreseeable future. The OECD states that "[d]isruptive shocks to the global economy are likely to become more frequent and cause greater economic and societal hardship."[8] Global warming is also posing increasing uncertainty to the global environmental conditions and hence well-being of the global population. According to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, climate change is projected to "progressively increase inter-annual variability of crop yields in many regions." The report also has a high level of confidence that all aspects of food security will potentially be affected by climate change, including food access, utilization, and price stability.[9]

The increasing interconnectivity of the global economy has made this growing volatility an even bigger threat to global sustainable development than otherwise, as disruptive shocks occurring in a certain part of the world or to a certain population group can now also easily affect others. For example, the GFC originating in the US and Europe had major impacts on the global economy for reasons such as the heavy reliance by emerging economies of exporting to advanced economies. On the other hand, even though the world is now more interconnected than ever, impacts of disruptive shocks might still not be visible and trackable immediately, despite the fact that they might be severe and long lasting. Consequences of such impacts might not always be detected by the traditional monitoring system, and even for those impacts that are detectable it is often too late for one to react, given that many official statistics are generated annually. For example, Figure 2 illustrates the information gap for the Millennium Development Goals, which shows that the MDG data at the global level are suffering multi-year lags.

In this regard, the introduction of Big Data could complement the traditional official statistics and help development practitioners to better capture data on development situations, at a larger scale and in a more up-to-date fashion, through the use of mobile phones, internet, radios, and other geospatial technologies, etc.

Figure 2

Information gap for the MDGs in 2014



N.B. Information gap is defined as the age of the most recent data that are available for evaluating the overall progress of the MDGs

Source: Millennium Development Goals Report 2014

8 "Economy: Global Shocks to Become More Frequent, Says OECD." Organisation for Economic Cooperation and Development. 27 June. 2011. http://www.oecd.org/document/15/0,3746,en_21571361_44315115_48252559_1_1_1_1,00.html

9 IPCC, 2014: Summary for policymakers. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken,P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1-32.

Sources of Big Data for Development

According to the UN Global Pulse, sources of Big Data for Development are those that "can be analyzed to gain insight into human well-being and development." These sources in general share the following features:

- Digitally generated: Data is created digitally, not digitized manually, and can be manipulated by computers
- Passively produced: Data is a by-product of interactions with digital services
- Automatically collected: A system is in place that automatically extracts and stores the relevant generated data
- Geographically or temporally trackable
- Continuously analyzed: Information is relevant to human well-being and development, and can be analyzed in real time.

The UN Global Pulse has noted that the concept of "real time" in the context of development is different from that in private sector, as "real time" data for development purposes does not necessarily mean the data is made available immediately after the occurrence of "data-generating" events. Instead, "real time" data can be understood as information that is produced and made available within a relatively short time frame so that it can be processed and allow actions to be taken in response, i.e. creating a feedback loop. Both the intrinsic time dimensionality and that of the feedback loop define the "real time" characteristic of Big Data for Development.[10] And this "real time" characteristic allows Big Data to help tracking development progress better.

Digital Divide

In the context of international development, one needs to be aware of the so-called "Digital Divide," which refers to the disparity in IT infrastructure, general services, capacities and skills between the developing and the developed world. To illustrate such a divide, it is estimated that in 1986 the top performing 20 percent of the world's storage technologies were capable of holding 75 percent of the world's technologically stored information, and this figure grew to 93 percent by 2007.[11]

Another dimension of the Digital Divide is in the general supporting services. In addition to hardware structures, analyzing Big Data also requires heavy reliance on software services. Basic capacities in production, adoption, and adaptation of software products and services are all crucial to the development of a Big Data-enabling environment. Evidence has shown that countries that are already lagging behind in terms of spending on ICT in absolute terms also have fewer capabilities for software and computer services in relative terms.[12]

Professionals who are able to process and analyze the data are also essential for Big Data development. Statisticians and awareness about the importance of statistical capabilities are relatively lacking in developing countries, but there are also incidences where some countries with comparatively low-income levels are able to achieve high graduation rates for professionals with strong analytical skills (e.g. Romania and Poland).[13]

In essence, the Digital Divide reflects the uneven development across countries at the global level and across regions at the national level. In order to ensure that less developed areas can exploit the potential of Big Data as

---

10 UN Global Pulse. 2012. Big Data for Development: Opportunities and Challenges

11 Hilbert, Martin, Big Data for Development: From Information- to Knowledge Societies (January 15, 2013). Available at SSRN: http://ssrn.com/abstract=2205145 or http://dx.doi.org/10.2139/ssrn.2205145

12 Ibid.

13 Ibid.

effectively as developed areas and prevent the Digital Divide from perpetuating or even expanding, tailored support for less developed areas in adopting the Big Data approach would need to be carefully planned.

The case for exploring Big Data for Development in China

Careful review of the existing literature indicates that the current discussion related to Big Data in China is firmly centered on its business potential or the technological advances in the field. The largest Big Data conferences took place or are going to take place in China, including the Big Data Technology Conference, Big Data & Analytics Innovation Summit, China Legal Big Data Symposium, Big Data Asia Showcase, and Big Data World Forum. All these have either a strong business or technology focus (or in some cases both) and certain objectives, such as investigating "the key scientific and technological issues of Big Data," deliberating on "advantages of using Big Data with Business Intelligence," creating "a platform for all legal, forensics, investigation and information management professionals to come together and extend collaboration opportunity," or discussing how to use "the power of Big Data to drive business strategy."

In contrast, there have been limited discussions on how Big Data can be used for development purposes. However, it is the position of this paper that there is considerable potential for Big Data in the development sector and it is time to start exploring how to exploit such potential.

First of all, China has the world's largest mobile phone market, with over 1.2 billion mobile subscriptions.[14] With over 600 million Internet users, it also has the biggest Internet user population in the world.[15] Moreover, China has the world's most active environment for social media, with the government estimating that over 250 million people use social media – be it in the form of blogs, social-networking sites, microblogs or other online communities. In fact, estimates of non-governmental entities are often higher, putting the number over 590 million.[16] It is also estimated that the digital universe in China will continue to grow at a rapid rate, with the country's share of global digital data expected to rise to 18 percent by 2020, up from 13 percent in 2014.[17]

The massive numbers of real-time information streams and people who use mobile phones, Internet, and social media in China creates a favorable environment where the Big Data approach could be effective in providing insights on emerging concerns that are highly relevant to China's development.

There are signs that the government has also started to examine the potential of Big Data in the public sector. In June 2014, the Chinese People's Political Consultative Conference (CPPCC) – the country's political advisory legislative body – held a consultation forum in Beijing on how to use Big Data technology to enhance governance capability. The forum – chaired by CPPCC Chairman Yu Zhengsheng – focused on how government can support and promote the application of Big Data technology, establish a China-specific legal system for data matters, set

[14] Source: http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/a

[15] Source: http://www.washingtonpost.com/blogs/the-switch/wp/2014/01/31/china-has-almost-twice-as-many-internet-users-as-the-u-s-has-people/; http://it.21cn.com/itnews/a/2014/0731/15/27931731.shtml (in Chinese)

[16] There are likely "double-counting" issues here, i.e. people have more than one social media accounts. Source: http://www.go-globe.com/blog/social-media-china/; http://www.socialmediatoday.com/content/understanding-social-media-china-2014

[17] EMC Digital. April 2014. Digital universe country brief: China. Link: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014-china.pdf

up a cross-ministry sharing mechanism for Big Data, develop a national Big Data strategy and refine existing mechanisms for sharing and utilization of data.[18]

At the local government levels, the Province of Guizhou has rolled out a range of preferential policies for the development of Big Data development and application.[19] It is also reported that power, data center and communications infrastructure has been established in Chongqing to enable the city to support Big Data, drastically accelerating the city's journey towards urbanization.[20] All these developments indicate that government interest at both the central and local levels in Big Data are burgeoning.

Two proposed levels of work in relation to Big Data for Development in China

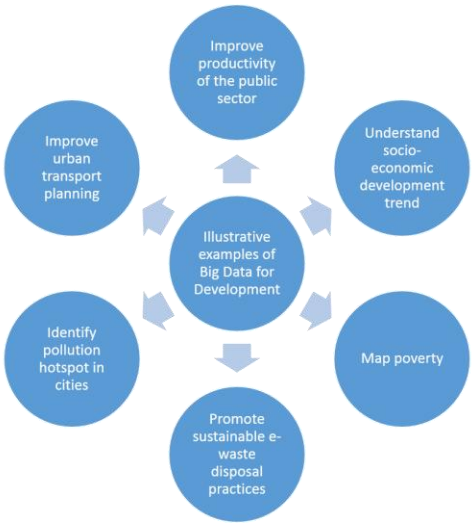
In order to take advantage of the considerable potential of Big Data for Development, development agencies could take a leading role in initiatives to support the Chinese government. In this view, this paper has identified 2 levels of work in relation to Big Data for Development:

1. **The first level is to create an enabling environment for Big Data for Development:** In order to create such an environment, development agencies should work with the public and private sectors to address many of the challenges that were outlined in the previous section, including data, analytical, operational/systemic challenges.

2. **The second level is to tackle particular development challenges with the Big Data approach:** Possible options here include assess poverty situations, improve productivity of the public sector, refine urban transport planning and strengthening understanding of socioeconomic trends, promote sustainable e-waste disposal practices, and identify pollution hotspots in cities, amongst others.

Figure 3

6 illustrative examples of Big Data for Development



[18] Source: http://gb.cri.cn/42071/2014/06/13/2225s4576641.htm (in Chinese)

[19] Source: http://www.datacenterdynamics.com/focus/archive/2014/03/china%E2%80%99s-new-big-data-hub

[20] Source: http://blog.pacnet.com/big-data-enables-non-traditional-opportunities-for-urbanization-development/

The First Level of Work: Create an enabling environment for Big Development for Development

Promote Data Philanthropy

As the private sector possesses most of the data out there, development agencies could work with the public and private sectors to promote the idea of "Data Philanthropy," i.e. private sector sharing data to support more timely and targeted policy actions. In order to put this concept into implementation in China, development agencies can initiate conversations with Chinese business leaders on the advantages of engaging in Data Philanthropy. In fact, the UN Global Pulse has already engaged in such conversations at international level and what they found out is that private sectors agree that this concept is not only a new form of charity or corporate social responsibility, but also "a sensible strategy for business risk mitigation, particularly when investing in emerging markets." The UN Global Pulse gave an illustrative example where data on calling patterns collected by mobile phone companies could help to reveal signs of brewing development challenges, such as a low-level food crisis. By sharing such data with policy makers, it could help the government speed up its responses to the crisis. Without the data, the government might not be able to react promptly to the crisis, with the result that the population of the area could suffer lasting harm and many of them might not be able to continue to afford the mobile phone services, thence cutting into profits for the companies.

There are a wide range of issues related to Data Philanthropy, which need to be addressed, including data privacy, creating incentives for data sharing, constructing a national architecture for data philanthropy and how a public-private partnership agreement would look like. Therefore, conversations on Data Philanthropy would also need to involve experts on legal, ethical, technological, methodological and political issues, with development agencies taking advantage of their convening power in creating such multi-disciplinary discussions.

It should be noted that in November 2013 China's National Bureau of Statistics (NBS) signed a series of agreements with 11 major Chinese enterprises, aiming to build long-term collaborative relationships on using Big Data.[21] These enterprises – including Baidu, Alibaba, China Unicom, and Fanya Metal Exchange, etc. – have indicated their willingness to share data with NBS to maximize the effect of Big Data application. For example, the cooperation between NBS and Baidu focuses on 3 main aspects:

1. Generalizing the official statistical data and programs through the Baidu website;
2. Improving the predicting model of the macro economy by combining the Big Data on the web with survey data collected by NBS;
3. Grasping more meaningful statistical requirements and completing the survey programs by following netizens' paths on the web platform.

Development agencies could seek to be involved in these established partnerships or even create new partnerships with these enterprises to identify and share data that would be relevant for development purposes.

Work with the government to develop Big Data strategy

As discussed in an earlier section, the central and local governments have shown considerable interest in Big Data. Given that different levels of government are still at the stage of exploring the options for Big Data application, it

[21] National Bureau of Statistics. 2014. Big Data and Official Statistics in China. Working Paper. For Meeting on the Management of Statistical Information Systems (MSIS 2014).

would be a good opportunity for development agencies to work with central and local governments to explore how Big Data can be used for development.

At the international level, a good example of exploring how Big Data can be used in the public sector is the Public Service Big Data Strategy that the Department of Finance and Deregulation of the Australian Government produced. In this strategy paper, opportunities and benefits of Big Data application in providing public services are presented. The paper also discusses the principles that government agencies should follow while using Big Data and proposes a series of actions that will help to facilitate the application of Big Data, including developing practice guidance, identifying and reporting on barriers to Big Data analytics, and enhancing skills and experience in Big Data analysis, etc. Development agencies could work with central and local government to produce similar Big Data strategy papers for the public service sector and other development areas where the agencies have considerable expertise.

The process of preparing such papers should involve broad consultations across different government agencies, which will also help to sensitize government officials to the notion of Big Data and related matters. The papers could help to communicate a clear vision of what the government seeks to achieve in specific sectors with Big Data application, and how the objectives would be achieved using Big Data.

Among the local governments, Guizhou and Chongqing have demonstrated strong interest in Big Data and these are some of the local governments that development agencies could consider working with.

The Second Level of Work: Tackle specific development challenges with the Big Data approach

Promote sustainable e-waste disposal practices

China, as the world's second biggest e-waste producer and biggest e-waste importer, is facing tremendous challenges in e-waste management. From 2009 to 2013, national electronic waste grew at an annual average of 21.6%. However, out of the over 3.6 million tonnes of e-waste being generated domestically, the actual amount of e-waste processed from formal channels only accounts for 40% of the theoretical scrap.

One of the biggest challenges of e-waste management in China, as well as in other developing countries, is that the informal sector dominates e-waste collection. The informal sector dismantles e-waste with enormous negative impacts on the environment and human health. As a result, the informal sector also has much bigger profits than the formal sector, which is more tightly regulated and has to put in resources to minimize the negative environmental or health impacts.

Smartphone apps can be developed to promote sustainable e-waste disposal practices by quantifying the damage caused by incorrect disposal, assisting in correct disposal and locating the e-waste collection points that are nearest to the application user. This would support the formal sector with more competitive logistical arrangements and better interactions with consumers. This could also help to streamline the recycling process and cut down on these 'informal recycling stations' where unaccredited entities reclaim precious metals from within electronic equipment but then dispose of the toxic materials incorrectly, contributing to severe ground and water pollution.

In fact, UNDP is already working with the Chinese search-engine leader Baidu on developing such app, as the inaugural product of a Big Data joint laboratory between the two organizations. The Joint Lab is designed as an open platform that brings together big data and development experts from UNDP and Baidu as well as partners

from government, academia, CSOs and the private sector in both traditional and new technology industries across the country to tap into new insights and produce idea prototypes for testing and implementation.

The joint Lab will leverage Baidu's Big Data engine to identify valuable data, which can contribute to formulating and implementing development strategies. Also, the public will be engaged to raise awareness of the specific challenges that the Joint Lab will be addressing, in order to encourage participation and foster behaviour change.

Improve productivity of the public sector

It is estimated by the McKinsey Global Institute that by using Big Data to improve its productivity, Europe's public sector could reduce the costs of administrative activities by up to 15-20 percent, creating the equivalent of USD 223 billion to 446 billion – or even higher – in new value.[22] Creation of such value comes from both efficiency gains and a reduction in the gap between actual and potential collection of tax revenue. The report identifies several Big Data levers that would lead to such results, including increasing transparency and applying advance analytics. These levers could increase annual productivity growth in the public sector by up to 0.5 percentage points through 2020. The report points out that Big Data can also play a similar role in other countries and regions.

The findings of the McKinsey report are echoed by a 2012 report produced by the UK-based think tank Policy Exchange.[23] The report, entitled "The Big Data Opportunity: Making Government Faster, Smarter and More Personal," contends that Big Data gives the public sector "new ways to organize, learn and innovate." The report estimates that achieving cutting-edge performance, with the help of Big Data, could save the UK public sector up to £16 billion to £33 billion a year – equivalent to £250 to £500 per head of the population.

Development agencies can work with the Chinese government to improve public sector productivity through using Big Data. The first step could be an econometrics exercise to estimate the value of cost-saving that the adoption of a Big Data approach could bring to the government, which would help to create incentives for Big Data application in the public sector. For actual application, development agencies could cooperate with central or local governments to set up pilots for testing the potential of Big Data in improving public sector productivity.

The McKinsey report suggests that Big Data could be used to uncover variability in the performance of different parts of a government entity that are performing broadly similar functions, which is not detectable at the aggregate level. Comparison of performance across units could help to identify units that are of low productivity and allow governments to subsequently address the issue. Automated algorithms – a crucial element of Big Data application – could be used to analyze large datasets and detect anomalies, such as in tax collection or benefits payments from labor or social security departments.

Understand socioeconomic development trend

One of the first and more prominent examples of the UN Global Pulse in demonstrating the potential of Big Data for Development is the semi-automated sentiment analysis of social media streams in Indonesia, which was shown to have significant correlation with official statistics on food and fuel prices. The results have shown that

[22] McKinsey Global Institute. 2011. Big Data: The Next Frontier for Innovation, Competition and Productivity

[23] Policy Exchange. 2012. The Big Data Opportunity: Making Government Faster, Smarter and More Personal.

even a basic analysis of the volume of tweets related to food price rises shows a correlation with official statistics on consumer price index.[24]

Another example is the UN Global Pulse-SAS International partnership on examining the role of social media as an early indicator of an unemployment hike in Ireland. The initiative collected digital data on social media, blogs, forums and news articles, etc., that were related to unemployment and performed sentiment analysis to categorize the mood of these online conversations. They were able to find that increased social media conversations about work-related anxiety and confusion provided a 3-month early warning indicator of an unemployment hike.[25]

In the case of Indonesia, it was reported that there were 20 million Twitter accounts (1 in every 12 persons is an active user of Twitter). In comparison, it is estimated that there are over 250 million general social networking users (over 1 in every 6 persons). The potential of China's social media stream in revealing socioeconomic development trend is clearly tremendous.

Map poverty

Social surveys and censuses periodically collected by national statistical institutes contain valuable information describing the social and economic wellbeing of a country and the relative health of different areas. However, this form of data collection is well known to be an onerous task due to the cost involved. In a paper by Christopher Smith and his colleagues (2013), they propose the use of ubiquitous sensing as an alternative to the traditional method of collecting socio-demographic data.[26] Ubiquitous sensing, closely related to Big Data, refers to the passive collection of people's digital footprints (e.g., location based social networking check-ins, phone calls, etc.), which can provide a detailed picture of human mobility and communication. In this paper, Call Detail Records (CDRs) are mined in order to generate proxies for poverty indicators, which can then be used to estimate poverty on a continuous basis and at low cost, as opposed to the slow iteration of census survey cycles.

The authors have demonstrated the potential of CDR data to provide an invaluable source of poverty estimates, even without knowledge of individual behavior. They have uncovered several features of communication patterns among mobile phone users in Cote d'Ivoire that track poverty of regions as defined by the Multidimensional Poverty Index.

With over 1.2 billion mobile phone subscription and over 83 percent of rural Internet users using mobile phones to access the Internet, China is in a prime position to use mobile phone data to map poverty in rural areas.[27] The cost of producing estimates from passively and automatically collected communication data is negligible compared to that of manual surveying, which makes it possible to obtain estimates of poverty levels on a continuous basis.

[24] UN Global Pulse. 2014. "Mining Indonesian Tweets to Understand Food Price Crises" Methods Paper.

[25] UN Global Pulse. 2013. Big Data for Development: A Primer.

[26] Smith, Christopher, Afra Mashhadi, and Licia Capra. 2013. "Ubiquitous sensing for mapping poverty in developing countries." Paper submitted to the Orange D4D Challenge.

[27] Source: http://www.cww.net.cn/tech/html/2014/6/17/20146171015356336.htm (in Chinese)

Improve urban transport planning

Traffic congestion and associated air pollution are major challenges that many Chinese cities are facing.[28] Big Data can be used to improve urban transport planning, thus addressing these two issues. At the moment, municipal governments and city planners alike are lacking credible, up-to-date traffic data that they can use to identify underlying problems with cities' transportation networks and to develop strategies that best address traffic-related issues. Some of the major shortcomings of Chinese transport databases include the lack of comprehensive urban transport databases and the lack of data uploading and downloading mechanisms.

Introducing a cloud-based platform in vehicles and a smart data center would allow the monitoring and collection of instant traffic behavior which can help to identify anomalous behavior (indicating underlying problems in cities' road networks) and also share instant updates on traffic to drivers in order to avoid traffic congestion.

For example, there were earlier projects that used GPS systems to track the locations of taxis in Beijing to identify the underlying problems with the city's transportation network. Using data from the GPS system, researchers from Microsoft were able to find out at what times the network of roads and subway lines between two districts in the city was unable support the number of people traveling between them.[29]

The GPS system, however, does not allow for instant processing and sharing of traffic information. By introducing a cloud-based system, drivers whose vehicles are connected to the cloud network can be instantaneously informed of the current traffic information and the system can use the latest information of current traffic conditions to recommend routes that take the shortest time. It provides a more dynamic approach to addressing traffic congestion and, consequently, the related air pollution issue.

The main development objectives of this proposed initiative is to use an inclusive, participatory approach to capture accurate, up-to-date citywide information and address two main challenges that are plaguing Chinese cities, i.e. air pollution and traffic congestion. Through addressing these challenges, this initiative could improve the well-being of the selected city's residents, in particular their health and time saved from being stuck in traffic, which they can spend on pursuing productive matters that they value.

Identify pollution hotspots in cities

Air pollution in China is making headlines around the world with hazardous haze blanketing Beijing for extended period of time. There is increased recognition of the impacts of air pollution in China. It is estimated that 350,000 to 1.2 million premature deaths linked to air pollution each year in China.[30] In fact, the most recent estimates made by the World Health Organization show that air pollution might now be the world's largest environmental health risk, with low- and middle-income countries in the Western Pacific region, including China, having the largest number of death per capita from air pollution in 2012.[31]

---

[28] Data on air pollution in China is complex and dynamic with new studies continuously emerging. In addition, air quality varies from region to region. For example, a study by the Chinese Academy of Sciences released in December 2013, indicated that vehicle emissions accounted for only 4% of Beijing's PM2.5 (http://www.atmos-chemphys.net/13/7053/2013/acp-13-7053-2013.html). However, previous studies had indicated that vehicle emissions may contribute 20-30% of PM2.5. See here for a discussion of the controversy http://www.rsc.org/chemistryworld/2014/01/pollutionresearch-sparks-car-control-debate-china

[29] Source: http://research.microsoft.com/en-us/projects/urbancomputing/

[30] Source: http://www.scmp.com/news/china/article/1399671/ex-health-minister-endorses-finding-chinas-smog-kills-350000-year

[31] Source: http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

Big Data can be used to identify the so called pollution hotspots in cities, which allow citizens to plan a less-polluted route for their daily jog or bike ride, etc. Sensory devices that measure air quality and particulate-matter level can be used to create a real-time pollution map. Combining such a map with data on weather pattern, a real-time short-range forecast of where the pollution hotspots will be blown could then be developed to allow end users to plan less-polluted routes while travelling within the cities.

Challenges for application of Big Data for Development in China

Review of existing literature and also previous experiences of adopting Big Data approach in other countries shows that there are clearly a multitude of issues that need to be addressed in order to ensure effective application of Big Data for Development. China will also need to address these challenges while using Big Data for development purpose. These challenges can be broadly grouped into 3 categories:

1. Operational/systemic challenges
    a. **Privacy.** One of the biggest challenges for Big Data application is the issue of privacy. The UN Global Pulse considers it "the most sensitive issue, with conceptual, legal, and technological implications." As a fundamental human right, privacy has its intrinsic values. It also has tremendous instrumental values as a modern society needs privacy to flourish and "without privacy, safety, diversity, pluralism, innovation, and basic freedoms are at risk."[32] As discussed earlier, privacy is a major reason why firms are reluctant to share the data of their clients, users and their own operations. Well-specified national data disclosure policies and a proper regulatory environment that ensures data privacy will be essential to provide the assurance that firms need in order to make their data available for development purpose.

    In particular, one can refer to the Resolution adopted in the 36th International Conference of Data Protection and Privacy Commissioners, which calls for, inter alia, respect of the principle of purpose specification; limiting the amount of data collected and stored to the level that is necessary for the intended lawful purpose; obtaining valid consents from data subjects in connection with use of personal data for analysis and profiling purposes; transparency about which data is collected, how the data is processed, for what purposes, and whether or not the data will be distributed to third parties; and giving individual appropriate access to the data collected about them, as well as access to information and decisions made about them.[33]

    b. **Changes in decision-making process.** Effective application of Big Data for Development would also require changes in the decision-making process, which customarily relies on traditional statistics. Given the high frequency of Big Data, a more responsive mechanism will need to be put in place that allows the government to process the information and act quickly in response. Also, since Big Data is often unstructured and relatively imprecise (compared to official statistics), government officials also have to learn how to effectively interpret and make use of the information provided by Big Data. This requires capacity building to turn decision makers into more sophisticated data users.

    c. **Administrative barriers.** The full potential of Big Data can only be realized when different institutions, public and private, work together seamlessly beyond their administrative boundaries. This requires development agencies to take advantage of their convening power to gather influential

[32] UN Global Pulse. 2012. Big Data for Development: Opportunities and Challenges

[33] Full text of Resolution adopted in 36th International Conference of Data Protection & Privacy Commissioners. http://www.privacyconference2014.org/media/16427/Resolution-Big-Data.pdf

partners of Big Data to create new models for different institutions to work together to create value in both social and economic development.

2. Data challenges
   a. **Accessibility.** From the technical perspective, it is not the case that all data produced by Big Data sources are easily accessible and storable, especially for data generated on social media platforms. There are also considerable difficulties with accessing data from mobile phone carriers. Another major challenge related to data accessibility is the unwillingness of private firms that are in possession of the majority of Big Data to share data about their clients and users, as well as their own operations. The reasons behind such reluctance could be multi-fold, including concerns regarding the loss of competitiveness by giving up firm data, the risk of breaching privacy law and regulations, and the possibility of damaging the firm's reputation, etc.

   b. **Availability.** The use of mobile phone, Internet, and social media tend to be more concentrated in the cities than in the rural areas, meaning that data collected from those sources most likely have urban-bias and are not representative of the entire population in China.

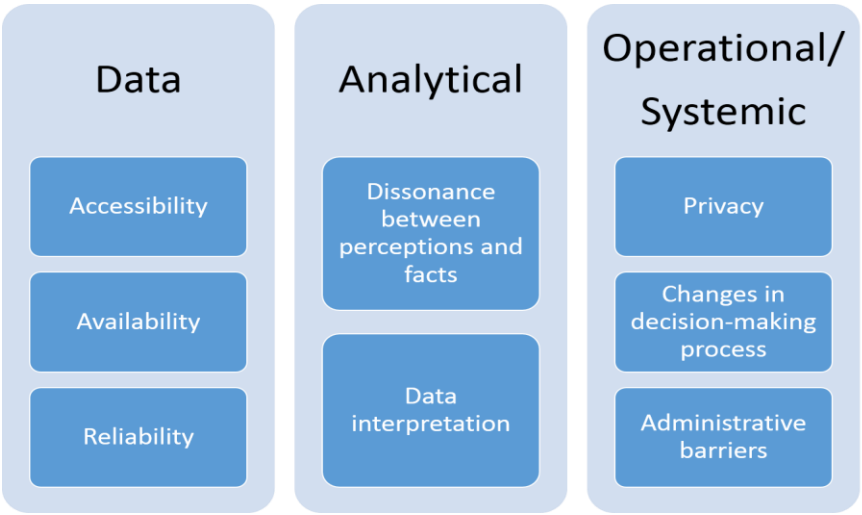
   c. **Reliability.** There might be attempts at interfering with the operation of systems or suppressing signals. The highly competitive social media landscape in China means that the user groups of different social media platforms are very dynamic in terms of their sizes and compositions, which would have strong implications on how data from such sources is being interpreted.

3. Analytical challenges
   a. **Dissonance between perceptions and facts.** This is a particular challenge when dealing with behavioral or opinion data such as online searches or comments on social media platforms, which are basically derived from users' own perceptions. The example of Google Flu Trends has shown that while the tool was very good at predicting the spread of nonspecific respiratory illness that are similar to flu, based on Goggle searches for flu-like symptoms, it did not do a very good job at predicting the actual flu. The issue is largely a result of people confusing flu with other illnesses because they have similar symptoms. Another related analytical challenge is the dissonance between expressed intentions and actual intentions. It is a particular problem while dealing with text-based data, which leaves data analysts with considerable room for interpretation. Accurate interpretation of such data would require good understanding of the contextual background against which the data is generated.

   b. **Data interpretation.** The nature of Big Data sources means that there is a range of data interpretation issues that require stronger attention while dealing with Big Data, as compared to dealing with official statistics. The first of these is the issue of sampling bias, where samples are not representative of the entire population of interest. The aforementioned urban bias of Big Data generated by social media platforms is an example of this. Another issue is the misinterpretation of correlation as causation. Such mistakes could happen when data analysts fail to consider the existence of confounding factors. Other data interpretation challenges include the difficulty of defining and detecting anomalies in human ecosystems.

Figure 4

Major challenges confronting Big Data for Development



Conclusion

Across the world, there has been budding interest in the potential of Big Data in contributing to development and indeed there have been some initial successes in demonstrating such potential. This paper argues that there is a case for development practitioners to explore how Big Data can be used for development purpose in China and proposed potential work that could be done in this area.

Three additional points should be made before concluding this paper.

First, the Big Data approach contributes to results-based management. Collection and analysis of data collected through Big Data sources allow development actors, who contributing directly or indirectly to achieving a set of development results, to ensure that their processes, products, and services contribute to the achievement of desired results. Given the "real time"-ness of Big Data, actors could use the information and evidence of actual results to promptly inform decision-making on the design, resourcing, and delivery of programmes and activities as well as for accountability and reporting.

Second, even though Big Data has great potential for supporting development work, development practitioners have to understand the limits of the Big Data approach and the conclusion that one can draw from analyzing Big Data. In many ways, Big Data approach faces the same limitations as traditional statistical and numerical methods. While the size of Big Data allows it to be very good at detecting correlation, one still requires good understanding of the development context and the dynamics between variables of interest, control variables and unobservable variables in order to make causal interpretation. There are certainly situations where it would be useful to identify correlation alone, for example, identification of correlation between the volume of tweets related to food price and official statistics on inflation suggests that social media stream might be used to provide a more real-time assessment of inflation. However, in the situation where one wants to determine which policy instruments are most suitable for addressing certain development challenges, it would then require one to tease out the causal relationship between the policy instruments and the outcome of interest. In such situations, mere identification of correlation is insufficient.

Also, there are things such as happiness, livability, trust and many others that by nature are difficult to measure. Evaluations of these matters are inherently vague that even the adoption of Big Data approach would not make them more precise.

Third, the Big Data approach – by definition – would be most effective when the data are voluminous and are representative of the population or subjects that are of interest. In other words, the approach would be most effectual when there is a broad participation by different members of the society in contributing to the collection of data. For example, in the exercise of tracking inflation or unemployment rate through analyzing social media comments, the results would certainly be more accurate should the pool of social media users are sizeable and representative of the whole society, as opposed to representing only a subset of the population (hence leading to a biased estimation). In the case of improving the productivity of the public sector, broad public participation in evaluating the wide range of services that public sector provide to the general population would also help to more accurately identify areas where improvement in productivity are most needed.

In this sense, the promotion of the Big Data approach requires and in many ways incubates a more inclusive decision-making process, and the creation of an enabling environment for all social partners to contribute to a greater extent towards national development. ■