ISSUE 1

Brief on ICT Trend

Big Data

UNITED NATIONS
APCICT - ESCAP

Brief on ICT Trend

Issue 1

Big Data

Sanjay Bahl
Senior Consultant, CERT India

APCICT ASIAN AND PACIFIC TRAINING CENTRE FOR INFORMATION AND COMMUNICATION TECHNOLOGY FOR DEVELOPMENT

Brief on ICT Trend

Issue 1: BIG DATA

Contact:
United Nations Asian and Pacific Training Centre for Information and Communication Technology for Development (UN-APCICT/ESCAP)
5th Floor G-Tower, 175 Art center daero,
Yeonsu-gu, Incheon City (22004)
Republic of Korea

Tel: +82 32 458 6650
Fax: +82 32 458 6691
E-mail: info@unapcict.org
http://www.unapcict.org

Design and Layout: Docufriends
Printed in the Republic of Korea

Supported by

KOICA
Korea International Cooperation Agency

PREFACE

Established in June 2006 as a regional institute of the Economic and Social Commission for Asia and the Pacific (ESCAP), the United Nations Asian and Pacific Training Centre for ICT for Development (UN-APCICT/ESCAP) has been strengthening the human and institutional capacities of countries in the region to use ICT for sustainable development. It has developed and implemented flagship capacity building programmes – the Academy of ICT Essentials for Government Leaders targeted to civil servants and the Primer Series on ICTD for Youth for society's future leaders – which have seen widespread adoption among government organizations and universities in Asia and the Pacific, and beyond. As a regional hub on ICT capacity development, the Centre also provides a platform for regional dialogue, exchange of experiences and cooperation among national partners and ICTD stakeholders.

The Centre complements its flagship programmes through research and knowledge-sharing. The ICTD Briefing Notes provides high-level and senior government officials and policymakers with concise and policy-oriented information on key ICT for these trends can be utilized in national development strategies and programmes.

The first issue of the Brief on ICT Trend is on the topic of big data. Improvements in connectivity and mobile subscriptions around the world have resulted in the generation of volumes of data that can be useful in development work. In 2013, the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda recognized that big data, with its tremendous wealth of information, can support evidence-based decision making in critical development areas. Application of big data for development, however, is still in its early stages and the need to understand the possibilities and implications for policymaking is rapidly growing.

This first issue aims to enhance awareness of government officials and policymakers on big data and its potential applications in sustainable development. We hope that this resource will be a useful starting point from which government leaders and other stakeholders can improve their understanding on the topic. It offers an overview of big data, examples of real-world applications, and challenges and opportunities presented to government leaders and policymakers in developing countries.

I hope that you will find this first issue of the Brief on ICT Trend a valuable knowledge resource. I would like to express my appreciation to Mr. Sanjay Bahl for his support in developing this publication. My thanks to our partners and the ICT experts who provided valuable feedback during the multiple rounds of reviews, and to Christine Apikul for her work in editing the manuscript. Special thanks go to the Korean International Cooperation Agency (KOICA) for its financial support to this endeavor.

We look forward to adding more issues to this series as new ICT trends emerge with potential development opportunities.

Hyeun-Suk Rhee, Ph.D.
Director
UN-APCICT/ESCAP

ABOUT

Information and communication technology (ICT) as a driver for social and economic development received global attention since the turn of the century when the United Nations Millennium Development Goals (MDGs) called for making the benefits of ICTs available to all. ICTs—from radio to the Internet—have become widely accepted as critical tools for sustainable development and nowadays, it is impossible to attempt socio-economic development and poverty reduction without ICTs.

When the MDGs were formulated in 2000, the Internet was still in its infancy and the mobile phone revolution had not yet taken off. The ICT landscape has changed dramatically since then and advances in ICTs will continue at an accelerated pace. As we formulate and implement the post-2015 sustainable development framework, it is important to understand the transformations and trends taking place in this fast-changing digital era, in order to fully leverage the potential benefits of ICT advances. This series, entitled, “Brief on ICT Trend” has been created for precisely this reason, to introduce the implications of various ICT trends for sustainable and inclusive development.

This series is aimed at government policymakers and others who would like to better understand how different ICT trends will impact upon the economic, social and environmental arenas. The Brief on ICT Trends synthesizes existing research knowledge on the opportunities and risks of different ICT innovations to sustainable development, and provides recommendations for policy and practice. It is intended to be an introductory reference to support policymaking and programme planning.

Each issue in this series focus on a particular topic, and this first issue introduces “Big Data”. Improvements in connectivity and mobile subscriptions around the world have resulted in the generation of large volumes of data which when captured, stored and analysed can be used to support development efforts. The High-Level Panel of Eminent Persons on the Post-2015 Development Agenda in 2013 recognized that big data, with its tremendous wealth of information, can enhance evidence-based decision making in critical development areas. This issue looks at how government policymakers can maximize the potential of big data by balancing its risks and rewards.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

We are living in a world where data is being generated, captured and stored at an unprecedented scale. The magnitude of data available today and the advent of techniques and technologies to analyzanalyse it are changing nearly every aspect of our lives and the way business is done.

As recognized in the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda in 2013, big data, with its tremendous wealth of information, can strengthen decision-making in critical development areas such as health care, education, economic productivity, resource management, and crime and security, by providing cost-effective solutions.

Application of big data for global development is, however, still in its inception and several challenges, such as analytics, privacy, data access, infrastructure and human capacity should be addressed in order to fulfill its promises. Some key policy implications are suggested for policymakers and practitioners for future big data strategy.

While big data has great potentials to change the world and how we do business, its benefits will depend on how well its risks and challenges are managed. Forward-looking policymakers and practitioners will embrace the opportunities afforded by big data for their national development and provide the necessary support to realize the potential by balancing the risks and rewards of big data.

1. Introduction

1. Introduction

We are living in a world where data is being generated, captured and stored at an unprecedented scale. The magnitude of data (hence the term Big data) available today and the advent of techniques and technologies to analyse it are changing nearly every aspect of our lives and the way business is done. Public decisions that were based on hunches and assumptions can now be made based on facts driven from data. From an economic perspective, decision-making supported with big data is expected to have similar benefits to the increase in 'efficiency' and 'productivity' that ICT has brought about during the recent decade.[1]

As recognized in the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda, the advent of big data is also a historic expansion of the role that data plays in development.[2] It can strengthen decision-making in critical development areas such as health care, education, economic productivity, resource management, and crime and security, by providing cost-effective solutions. Without appropriate access to big data and capacity in data analytics, however, developing countries may lag behind, creating a new kind of digital divide.

1.1. What is Big Data?

So, what is Big Data? Big data is a loosely defined term used to describe "data sets so large and complex that they become difficult to work with using standard statistical software".[3] Big data is characterized using the three 3Vs—as "high-volume, high-velocity, and high-variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization."[4] Because the great majority of big data is unstructured, unfiltered, and incomplete by nature, "'veracity"' is also considered to beas one of the major characteristics of big data. Lastly, we need to extract "value" by By processing and analyzanalysing, we need to extract 'value' from this big data sources. Together, the five Vs are frequently used to describe the characteristics of big data. (See Figure 1)

1 Martin Hilbert, Big Data for Development (pre-published version, 2013).

2 Realizing that availability of data empowers people by arming them with information and encourages data-driven decision-making for development, theHigh Level Panel on the Post-2015 Development Agenda called for a "data revolution," a new international initiative to improve the quality and scope of data available to citizens and policymakers.

3 Chris Sniders, Uwe Matzat and Ulf-Dietrich Reips, "Big Data: Big Gaps of Knowledge in the Field of Internet", *International Journal of Internet Science,* vol. 7 (2012), pp. 1-5.

4 Mark A. Beyer and Douglas Laney, "The Importance of 'Big Data': A Definition", *Gartner*, 21 June 2012. Available from https://www.gartner.com/doc/2057415/importance-big-data-definition.

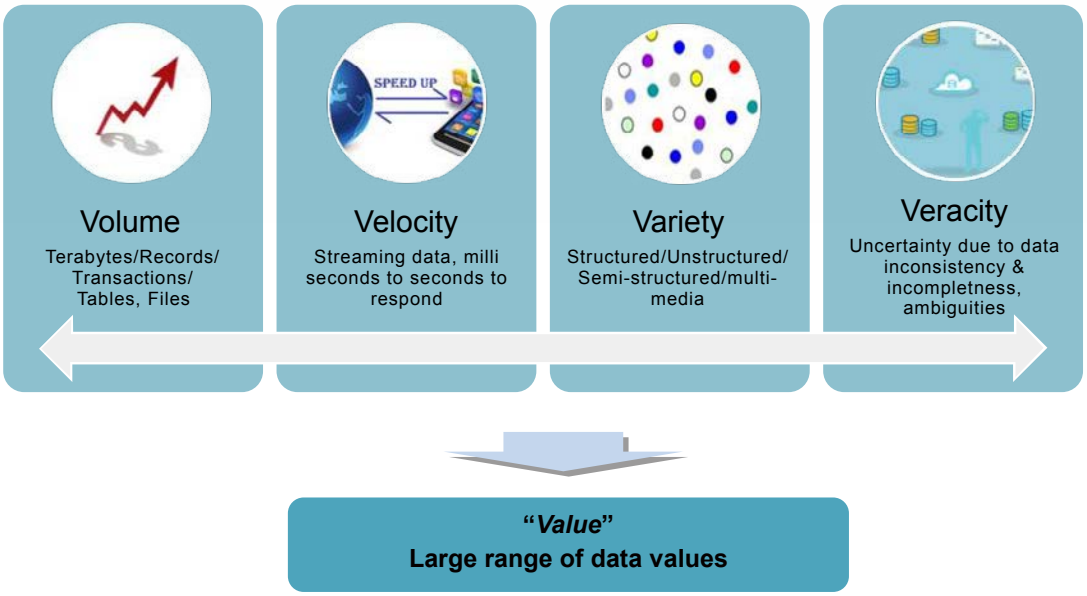

Figure 1. The five Vs of big data

Big data is not just about the massive growth of data volumes. In fact, the extraordinary diversity of data types is one of the critical factors that make it hard to deal with big data.

For instance, traditional database systems are usually designed to process "structural data", which can be quantified or easily tagged, categorized and organized for systematic analysis. Examples include national official statistics. The vast majority of big data is, however, "unstructured data" and does not fit neatly in a database. Such unstructured data files often include multimedia content (e.g., videos, photos, audio files) and texts (e.g., social media content, e-mail messages), both of which are much harder to analyse than traditional databases.

In addition to these data generated through the Internet, there is a third kind of big data gathered by digital sensors such as household appliances, cars and health sensors.[6] These types of observational data reflect human actions and thus, provide ample information on human behaviour. Examples of big data and their characteristics are described in Figure 2.

[5] Other examples include smart meters installed in homes to record electricity consumption, or satellite imagery that can pick up physical information such as vegetation cover as an indicator of deforestation.

Archives: Archives of scanned documents, statements, insurance forms, medical record and customer correspondence, paper archives, and print stream files that contain original systems of record between organizations and their customers

Docs: XLS, PDF, CSV, email, Word, PPT, HTML, HTML5, plain text, XML, JSON, etc.

Media: Images, videos, audio, Flash, live streams, podcasts, etc.

Business Apps: Project management, marketing automation, productivity, CRM, ERP content management systems, HR, storage, talent management, procurement, expense management, Google Docs, intranets, portals, etc.

Social Media: Twitter, Linkedin, Facebook, Tumblr, Blog, SlideShare, YouTube, Google+, Instagram, Flickr, Pinterest, Vimeo, Wordpress, IM, RSS, Review, Chatter, Jive, Yammer, etc.

Public Web: Government, weather, competitive, traffic, regulatory, compliance, health care services, economic, census, public finance, stock, OSINT, the World Bank, SEC/Edgar, Wikipedia, IMDb, and other web services

Data Storage: SQL, NoSQL, Hadoop, doc repository, file systems, etc.

Machine Log Data: Event logs, server data, application logs, business process logs, audit logs, call detail records, mobile location, mobile app usage, clickstream data, etc.

Sensor Data: Medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within devices, video games, cable boxes or household appliances, assembly lines, office buildings, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery, etc.

Figure 2. Types and characteristics of big data

Source: Kapow Software, "Understanding the Various Sources of Big Data – Infographic", BigData Startups. Available from http://www.bigdata-startups.com/BigData-startup/understanding-sources-big-data-infographic/#!prettyPhoto.

In order to extract value from the massive and diverse data, it is required to have scalable and affordable technologies that can store, discover and analyse the huge amount of data for enhanced decision-making process.[6] In this sense, big data is not just about the amount of data, a single technology, technique or initiative but more of a trend in data revolution.

Analysis of big data and its application can provide valuable insights about human behaviour and intentions. This will help development practitioners and policymakers in tackling various development issues by supplementing traditional data sources and providing more timely information.

1.2. Drivers of Big Data

1.2.1. The Emergence of Big Data

In today's society, an enormous amount of data is automatically generated and captured from various sources due to increasing interactions between individuals, businesses and public organizations. Google now processes over 40,000 search queries every second on average, which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide.[7] Also, every minute on Facebook, 510 comments are posted and 136,000 photos are uploaded. Overall, the production of data is growing at an accelerated pace. Experts now point to a 4,300 per cent increase in annual data generation by 2020 (see Figure 3).[8]

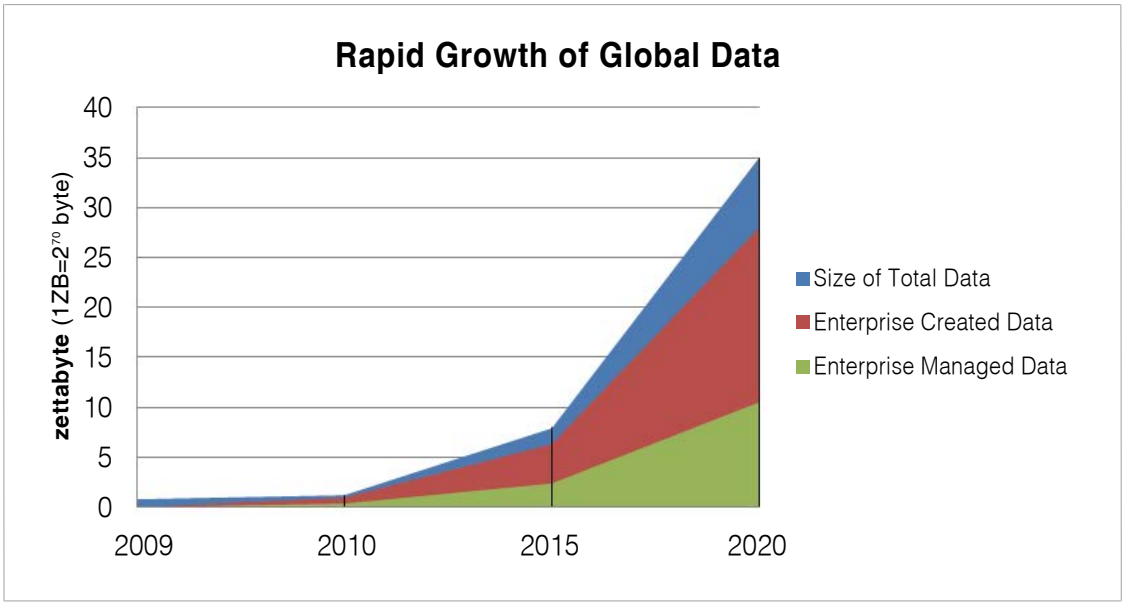

Figure 3. The rapid growth of global data

Source: Computer Sciences Corporation, "Big Data Universe Beginning to Explode", 2012. Available from http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode

6 Martin Hilbert, *Big Data for Development* (pre-published version, 2013).

7 Internet Live Stats, "Google Search Statistics". Available from http://www.internetlivestats.com/google-search-statistics/. Accessed on 7 August 2014.

8 Computer Sciences Corporation, "Big Data Universe Beginning to Explode", 2012. Available from http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode.

With all the data generated and captured, the increase in data storage capacity becomes a tremendously important subsystem that can determine the success of big data and its analytics implementation. Building on established technologies such as the Redundant Array of Independent Disks,[9] individual hard disk drives can be pooled together to form a reliable storage system that is bigger and faster than when the hard drives are used alone.[10] Hard drive capacity has also increased 50-million-fold since 1956. It took 26 years to create a 1 GB hard drive, but between 2007 and 2011, hard drives quadrupled in size from 1 TB to 4 TB. Within the next ten years, 20 TB hard drives may even become commonplace.[11]

The capacity to compute and process data has also expanded sharply. While the capacity to store and communicate data has grown about 25-30 per cent annually over recent decades, the capacity to compute information has grown about 60-80 per cent annually.[12] The consolidation of these computing and storage capacities into globally accessibly data centres, and making them available as either public or private cloud storage and computing services, has become a viable and mainstream solution for large-scale data processing. Such solutions are expected to grow at an aggressive pace in the next few years.

To derive competitive advantages in the world, computer and data scientists are motivated to extract all possible insights available to them. Revolutionizing how data is stored, recovered, queried and analysed with the increased computing power and storage capacities available to them, they are driving big data analytics deeper and further.

1.2.2. Data Revolution in Developing Countries

The increase in connectivity and the increasing use of smart devices ensure that the data revolution will also occur and be experienced in developing countries.

During the last decade, mobile phones and network have spread to billions of people in the developing world. According to the International Telecommunication Union, mobile phone subscriptions worldwide will reach almost 7 billion by the end of 2014.[13] Almost 80 per cent of these subscriptions are in developing countries.

9 Redundant Array of Independent Disks is a way of storing the same data in different places, thus, redundantly on multiple hard disks.

10 Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau, Redundant Arrays of Inexpensive Disks (2014). Available from http://pages.cs.wisc.edu/~remzi/OSTEP/file-raid.pdf.

11 Jenna Dutcher, "Data Size Matters [Infographic]", Datascience @Berkeley Blog, Berkeley School of Information, 6 November 2013. Available from http://datascience.berkeley.edu/big-data-infographic/.

12 Martin Hilbert and Priscila Lopez, "The World's Technological Capacity to Store, Communicate and Compute Information", Science, vol. 332, no. 6025 (2011), pp. 60-65; and Martin Hilbert and Priscila Lopez, "How to Measure the World's Technological Capacity to Communicate, Store and Compute Information? Part I: Results and edu/big-data-infographic/.

13 International Telecommunication Union, *The World in 2014: ICT Facts and Figures* (Geneva, 2014). Available from http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx.

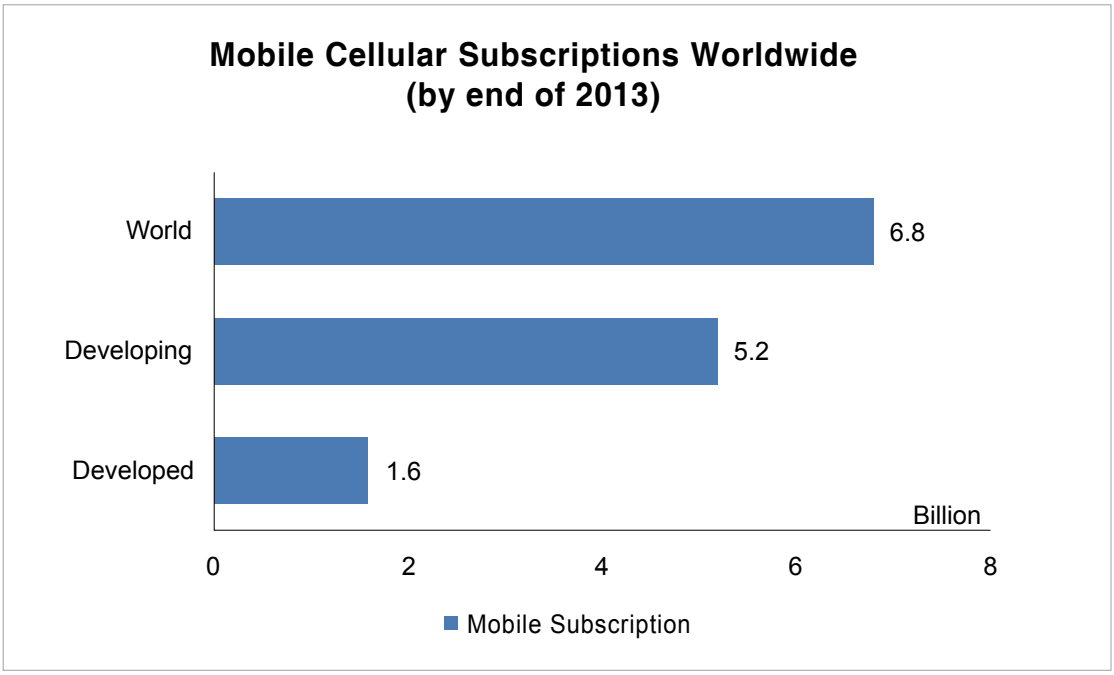

Figure 4. Mobile phone subscriptions worldwide

Source: International Telecommunication Union, World Telecommunication/ICT Indicators Database 2014 (18th Edition), 20 June 2014. Available from http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx.

Mobile phones are not only a tool for personal communication but have also become an indispensible part of daily life across the developing countries. Mobile phones are frequently used for financial and banking services, data transfer such as prices of various commodities, and medical information. Such growth and availability of mobile phones in developing countries have contributed to the emergence of big data, which can be analysed to enable evidence-based decision-making in developing countries.

The use of social media such as Facebook and Twitter is also growing rapidly in the developing world. For example, Egypt, Russia, the Philippines and 14 other developing countries outpace the United States of America in the proportion of Internet users who log on to social sites.[14] As the use of social media becomes widespread in developing countries, analysing trends in social media and its contents could provide valuable information on the emerging issues in the developing world and on ways to improve development programmes.

[14] Stephanie Pappas, "17 Developing Countries That Love Social Media More than the US", *Live Science,* 13 February 2014. Available from http://www.livescience.com/43347-developing-countries-social-media.html.

2. Why is Big Data Important for Development?

2. Why is Big Data Important for Development?

2.1. Call for Data Revolution for Development

2.1.1. Lack of Reliable Data

Policymakers and practitioners often experience difficulties in decision-making with the current data tools and systems available. In particular, many developing countries lack robust official systems to collect and utilize even basic demographic and health data, which is the basis of most development data. For example, in sub-Saharan Africa, only around 12 of the 49 countries have had a population census in the last 10 years.[15]

Even if they had a population survey, not all of them are done in the same way, which makes it very difficult to compare countries or combine data from different countries. When new data is collected, the estimates can also significantly change. For example, in Kenya, the estimated HIV prevalence in 2003 fell from 2.3 million to 1.2 million after conducting a population-based survey. Also, Ethiopia released new figure on HIV prevalence, lowering its estimate from 1 million to 0.5 million.[16]

Such lack of reliable data highly matters for the government in developing countries, which have scarce resources and need to put them to where they will do most good. Besides, they need solid data to evaluate whether the policies and programmes carried out to improve people's lives are actually working.

Recognizing the needs for more reliable data for development, the international community called for a data revolution in the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda in 2013. In particular, on improving the availability, accessibility, quality, transparency and timeliness of data; harnessing diverse sources of knowledge; and bridging the data gaps for development.[17] In this regard, analysis of the massive archive of big data provides new potential to look into the urgent problems and patterns in various development issues. Since big data is generated and may now be analysed in real time using the high performance computing networks, it can also provide more up-to-date information for decision-making.

2.1.2. Demand for "New" Knowledge

Big data not only supplements traditional data sources such as surveys or statistical imputations undertaken by official bodies, it also contributes to understanding human

15 Claire Melamed, "Development Data: How Accurate are the Figures?" The Guardian Poverty Matters Blog, 31 January 2014. Available from http://www.theguardian.com/global-development/poverty-matters/2014/jan/31/data-development-reliable-figures-numbers.

16 Jill Kristen Kresge, "HIV Prevalence Estimates: Fact or Fiction – IAVI Report", *The Publication on AIDS Vaccine Research,* vol. 11, no. 4 (2007).

17 United Nations, "What is Data Revolution", United Nations High-Level Panel: the Post-2015 Development Agenda, August 2013.

behaviour and intentions. For example, big data has been utilized to successfully predict employment, stock prices, election result and changes in gross domestic product in near real time, and also to monitor traffic and outbreak of diseases. Trends in social media are also tracked and analysed to measure people's welfare and socio-economic levels.

Among various uses, the United Nations Global Pulse proposes how big data could benefit development by providing new data roughly in three categories: 1) real-time awareness, 2) real-time monitoring of the impact of a policy, and 3) early warning uses, (see Table 1).[18]

Table 1. Actual and potential uses of big data for development

	Explanation	Example
Real-Time Awareness/ Predictive	Fine-grained and current representation of reality can inform the design and targeting of programmes and policies	Supplementing a census map using call data records in Latin America
Real-Time Feedback/ Prescriptive or diagnostic	The ability to monitor a population in real time makes it possible to understand where policies and programmes are failing so that the necessary adjustments can be made	Analysis of impact of a policy action e.g. the introduction of new traffic regulations – in real time
Early Warning/ Descriptive	Early detection of anomalies in how populations use digital devices and services can enable faster response in terms of crisis	Google Flu Trends analysed to detect the onset of the flu season; increased criminality in a given area

Source: Emmanuel Letouze, "Big Data for Development: Facts and Figures", SciDevNet, 15 April 2014. Available from http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html; SAS and United Nations Global Pulse, Using Social Media and Online Conversations to Add Depth to Unemployment Statistics, Methodological White Paper, 8 December 2011. Available from http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics; and SAS and United Nations Global Pulse, Can a Country's Online "Mood" Predict Unemployment Spikes? 12 March 2012. Available from http://www.sas.com/en_us/news/press-releases/2012/march/un-sma.html.

As discussed so far, big data provides considerable information to improve the models and estimates that inform all sorts of decision-making for development by supplementing traditional data as well as creating new knowledge.

2.2. Potential Uses of Big Data for Development

Big data has been used in various areas to address social and economic issues. Following are some data-intensive sectors that can effectively utilize big data for development, including education, health, agriculture, energy and transport.

2.2.1. Education

Big data has the potential to improve student learning. By studying big data, teachers and school administrators can find patterns and track individual performances that will help to customize the ways students are taught. For example, the International Financial Cooperation, an affiliation of the World Bank Group, mines large data sets for more than 63,000 students, collecting more than 14 million data points on annual student academic

18 Emmanuel Letouze, "Big Data for Development: Facts and Figures", SciDevNet, 15 April 2014. Available from http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html.

19 Mark Maccarthy, "Big Data Improves Education Around the World", *SIIA Digital Disclosure,* 18 April 2014. Available from http://www.siia.net/blog/index.php/2014/04/big-data-improves-education-around-the-world/.

performance that are used to shape instruction and achieve learning objectives.[19] By providing timely information on learning performance through big data analysis, the International Financial Cooperation helps students get the assistance they need much more quickly through the individual learning action plan developed by their teachers.

2.2.2. Health

Health care is one of the most important areas that affect the well-being of the poor. Better health care also promises greater productivity and increasing income opportunities. Big data can support a wide range of medical and health care services such as disease monitoring and surveillance, population health management, and medical decision support.[20] For example, using Google Maps and free public health data, the University of Florida created heat maps for municipalities based on numerous factors including population growth and chronic disease rates. These factors were compared with the availability of medical services in those areas. With this mapping, the university could find areas that were underserved for breast cancer screening and thus, redirected its health care units accordingly.[21] The advances in using big data also help doctors make accurate decisions that lead to better quality health care service.

2.2.3. Agriculture

In agriculture, big data means information can be collected along the whole supply chain including from supermarkets, weather sensing equipment, digital images and research papers.[22] These data sets can then be transformed into actionable information through analytics. For example, Tech Mahindra, an information technology services company in India, has developed a system called “Farm-to-Fork” that allows the monitoring of conditions in food shipping containers. When conditions such as temperature, humidity and oxygen levels change, alerts are sent out and the problem can be rectified either remotely or manually. Also, to help farmers prepare for extreme weather conditions that might damage crops, big data from the fields can be combined with weather patterns.[23]

2.2.4.Energy[24]

Big data has rich potential for the energy sector by helping governments and utilities companies predict energy consumption more effectively. For example, in Iowa, the city government is placing sensors on utility meters in order to collect data on how people and organizations within the city are using water, gas and electricity. By doing so, the government is able to manage energy procurement with greater precision and save on energy costs.[25] Also, in terms of renewable energy production to reduce fossil fuel

20 Wullianallur Raghupathi and Viju Raghupathi, “Review of Big Data Analytics in Healthcare: Promise and Potential”, *Health Information Science and Systems,* vol. 2, no. 3 (2014).

21 Brian Eastwood, “6 Big Data Analytics Use Cases for Healthcare IT”, *CIO,* 23 April 2013. Available from http://www.cio.com/article/2386531/healthcare/6-big-data-analytics-use-cases-for-healthcare-it.html.

22 Gordon Conway, “Big Data: Big Hope or Big Risk?” One Billion Hungry: Can We Feed the World? Blog, 17 April 2014. Available from http://canwefeedtheworld.wordpress.com/2014/04/17/big-data-big-hope-or-big-risk/#more-1093.

23 Gil Allouche, “Big Data is Transforming Every Industry, from Health and Education, to Farming and Energy”, *Betanews,* 31 July 2014. Available from http://betanews.com/2014/07/31/big-data-is-transforming-every-industry-from-health-and-education-to-farming-and-energy/.

24 Anders Quitzau, “Transforming Energy and Utilities through Big Data & Analytics”, I*BM Corporation,* 28 March 2014. Available from http://www.slideshare.net/AndersQuitzaulbm/big-data-analyticsin-energy-utilities.

25 Gordon Conway, “Big Data: Big Hope or Big Risk?” One Billion Hungry: Can We Feed the World? Blog, 17 April 2014. Available from http://canwefeedtheworld.wordpress.com/2014/04/17/big-data-big-hope-or-big-risk/#more-1093.

consumption, China is utilizing big data on weather and geography to predict how much

electricity will be generated at a wind farm or solar power plant. It allows utilities to avoid excessively using expensive and carbon-polluting fossil fuel plants to plug the gaps in supply when the renewable energy supply suddenly drops.[26]

2.2.5. Transport

Transport is another major sector where big data provides ample potential for efficient energy management as well as congestion control, which are critical issues for many cities around the world. For example, Singapore collects, analyses and disseminates data in real time on local road traffic conditions using a unified "i-Transport Platform". The platform analyses the amount for road tolls to optimize the use and efficiency of the country's transportation infrastructure and improve safety. Once the data has been processed into information that is relevant and useful, it disseminates this information via electronic signboards on the roads, web portals, its Twitter account, radio and mobile applications.[27] With the rapid growth of container traffic and size of container vessels, the freight movements and routing are also optimized. This along with data from the engine, weather conditions and other parameters provides information on the health of the ships machinery, thereby helping in fuel efficiency and reduced downtime or payments of demurrage in the port.

As suggested so far, big data is showing potential not only as an alternative source of data but also as a new knowledge depository across various areas for the purpose of socio-economic development. Big data is still in its infancy when it comes to the use for development, but it is likely that the importance will continue to grow given that digital data is on the rise.

[26] Todd Woody, "Big Data is Giving China an Edge in Renewable Energy Production", *QUARTZ,* 8 August 2013. Available from http://qz.com/113547/big-data-is-giving-china-an-edge-in-renewable-energy-production/#/h/4214,2,3/.

[27] Kelly Ng, "Singapore Government Uses Big Data Analytics to Optimise Transport Management", *FutureGov,* 17 April 2014. Available from http://www.futuregov.asia/articles/2014/apr/17/singapore-government-uses-big-data-analytics-optim/.

3. Real-World Applications of Big Data for Development

3. Real-World Applications of Big Data for Development

This chapter introduces real-world applications of big data. In particular, these case studies introduce how big data can be harnessed for development by providing real-time awareness, real-time feedback and early warning.[28] The case studies also intend to provide insight on the key implications of using big data in practice.

3.1. Big Data Analysis for Real-Time Awareness

3.1.1. Computing Cost-Effective Census Maps using Cell Phone Traces[29]

A. Overview

This project, conducted by Telefonica Research in Spain, is an example of how policymakers can formulate, implement and evaluate socio-economic policies by using real-time information provided with large-scale cell phone data.

B. Background

Census maps have large amounts of information on the socio-economic status of households at a national level. They also contain information that characterizes various social and economic aspects such as educational level or the access to electricity. These maps are important because policymakers often make important decisions based upon such information. However, computing and compiling the maps require extensive resources and become highly expensive, especially for developing countries with limited budgets.

In this context, Telefonica Research (Spain) proposed a new tool, “CenCell”, for governments and policymakers. CenCell makes the computation of census maps affordable by utilizing cell phone call data records, which is generating large amounts of digital footprints. These footprints can reveal human behavioural traits related to specific socio-economic characteristics. By using call data records, CenCell decreases the number of geographical areas that need to be interviewed by the enumerators, thus, the budget allocated for the computation of census maps can be reduced.

28 United Nations Global Pulse, *Big Data for Development: A Primer* (2013). Available from http://www.unglobalpulse.org/sites/default/files/Primer%202013_FINAL%20FOR%20PRINT.pdf.

29 Vanessa Frias-Martinez and others, “Computing Cost-Effective Census Maps from Cell Phone Traces”, paper presented at the Second Workshop on Pervasive Urban Applications, Newcastle, UK, 2012; and Vanessa Frias-Martinez and Enrique Frias-Martinze, “Enhancing Public Policy Decision Making using Large-Scale Cell Phone Data”, *United Nations Global Pulse*, 4 September 2012. Available from http://www.unglobalpulse.org/publicpolicyandcellphonedata.

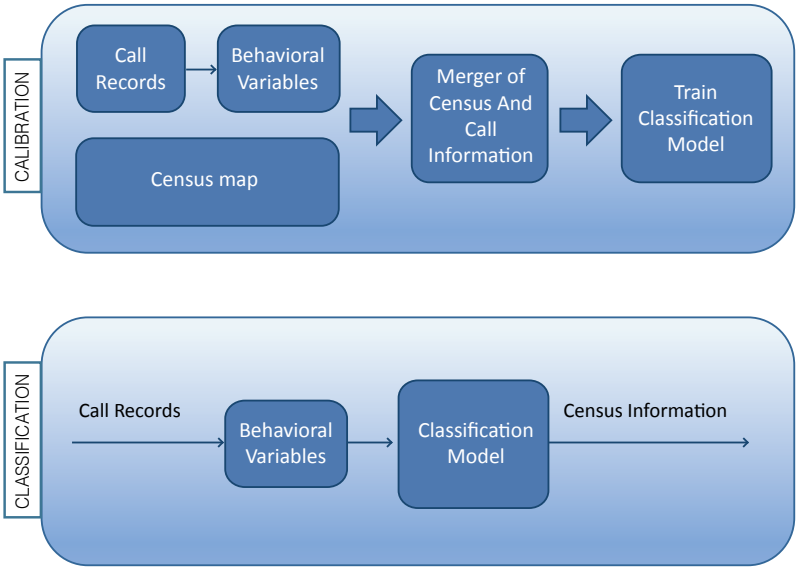

Figure 5. CenCell architecture

Source: Vanessa Frias-Martinez and others, "Computing Cost-Effective Census Maps from Cell Phone Traces", paper presented at the Second Workshop on Pervasive Urban Applications, Newcastle, UK, 2012.

C. Methodology

Figure 5 depicts the general architecture of CenCell. Basically, it consists of two main phases: 1) the calibration phase, which needs to be executed only once to set up the system for a region; and 2) the classification phase, which is executed every time census information is required for a specific geographical area in the region that was not covered by the enumerators through household surveys.

In the calibration phase, two data sets, including the anonymized cell phone call records[30] for the region under study and regional socio-economic levels statistics computed through household surveys, are combined to obtain a map. Then this map is used to create a model that will show the socio-economic levels of the areas not covered by household surveys based on the cell phone behavioural patterns of its citizens in the classification phase. This classification phase, which only requires access to anonymized aggregated calling records, can be executed as many times as needed and allows policymakers to compute affordable census maps without the need to conduct household surveys across all the region.

D.Implications

An experimental evaluation of CenCell was conducted in a large city in an emerging economy in Latin America. It used data from six months of cell phone calls, SMSes and MMSes from over 500,000 pre-paid and contract subscribers. Results from the evaluation reveal that the socio-economic level assigned by CenCell using cell phone traces are very good approximations when compared to the original values captured by the National Statistical Institute of the emerging economy. Moreover, CenCell identifies socio-economic levels at a fraction of the original cost and with a higher frequency, allowing more frequent evaluation of the impact of policy decisions on each specific area.

30 Call data records were used to model behaviour variables, clustered in three group variables—consumption, social and mobility.

3.1.2. Big Data for Disaster Risk Management: Smart Big Board

A. Overview

This project, conducted by National Disaster Management Institute in the Republic of Korea, is an example of how big data can be used for disaster risk management by providing comprehensive real-time monitoring information on disasters.

B. Background

Today's society requires a government that promptly responds to crisis and also prepares for the future. For these reasons, it is necessary to establish a real-time monitoring system that provides useful information for disaster response in the right place and at the right time. Most importantly, a government should be able to communicate effectively with its citizens about managing disaster risks. The Republic of Korea's National Disaster Management Institute developed "Smart Big Board" to manage national disaster risks in the era of big data.

C. Methodology

Smart Big Board is designed to display comprehensive disaster related information including weather information, CCTV records and disaster history on a layered digital map. Smart Big Board analyses and displays the big data collected through social media (mostly from Twitter), as well as the data collected with remote sensing equipment including satellites and unmanned aerial vehicles.[31]

Figure 6. Display of Smart Big Board

Source : Republic of Korea's National Disaster Management Institute, Smart Big Board Brief Note 2014.

31 An unmanned aerial vehicle, commonly known as a drone and referred to as a remotely piloted aircraft by the International Civil Aviation Organization, is an aircraft without a human pilot aboard.

The major functions of Smart Big Board include the following:

i. Smart Big Board provides weather and disaster related information on a digital map. Users can receive disaster alerts on selected areas by designating the areas of concern. The information on geographical features and road condition around the disaster-prone areas provided by Smart Big Board is critical for taking actions when disaster occurs.

ii. Unlike conventional media, social media provides disaster information without the lapse of time. Smart Big Board makes use of Twitter data in disaster risk management including disaster prevention, response and recovery. Analysing real-time Twitter data enables the identification of the causes and effects of disasters.

iii. Smart Big Board also provides real-time video of disasters to help disaster managers make prompt decisions by analysing real-time disaster information collected through social media.

D. Implications

This example shows that using big data collected from social media and other sources could play an important role in managing and communicating disaster risks promptly. Smart Big Board is still in development, with improvements being made as it gains more experience in data selection, analysis and display. For effective disaster risk management using big data, governments and related agencies need to prepare concrete strategies to develop big data analysis techniques, and plan to link and standardize various types of data in an institutional framework.

3.2. Big Data Analysis for Real-Time Feedback

3.2.1. Agenda Setting in the Indian Elections[32]

A. Overview

This project is an example of how policymakers can effectively set policy agendas to pursue inclusive growth and good governance by listening to and validating the voice of the citizens.

B. Background

In 2014, the 16th National Election was conducted in India. Whereas the previous national elections mainly relied on hunches, traditional wisdom, opinion polls and speeches, this election utilized digital technologies that integrate social media and big data analytics to set policy agendas and get real-time feedback from the electorates during the election campaign.

32 Neerja Pawha Jetley, "How Big Data Has Changed India Elections", *CNBC,* 10 April 2014. Available from http://www.cnbc.com/id/101571567.

C. Methodology

By data mining the voter sentiments, emotions and concerns from social media including Facebook and Twitter, the elected government was able to connect with the voters (including over 100 million new young voters) at the grass-roots level. It also drew up a micro-targeting and micro-messaging strategy for each voting cluster and state based on voters' emotions, reactions, sentiments and concerns. This analysis helped the elected government change their election strategy in real time and create innovative models for voter engagement.

The data collected and mined was also used for preparing the political parties manifesto, driving donations for the political party, and enrolling volunteers during the campaign period. In order to carry out such analysis, meticulous attention was paid towards obtaining a representative data sample in every single constituency.

Figure 7. Word cloud of the Bharatiya Janata Party manifeso for the 2014 Indian elections

Source: The Bharatiya Janata Party manifesto for the 2014 Lok Sabha elections.

D. Implications

This project shows how big data can create new value by real-time micro-segmentation of citizens, enable communication between citizens and government leaders, and discover the needs and bring out various citizens' aspirations, while being a platform for competition and growth.

3.2.2. Evaluating the Impact of Epidemic Alerts using Cell Phone Data[33]

A. Overview

This project, also conducted by Telefonica Research in Spain, is an example of using cell phone data to help policymakers evaluate the impact of public policy on epidemic spreading and calibrate their future actions.

B. Background

The outbreaks of pandemics such as H1N1 and SARS have shown that human mobility plays a central role in the spreading of epidemics. In order to control the spread of such diseases, governments often implement restrictions on mobility of its citizens by issuing public health alerts. In order to improve future action plans on epidemics, it is important to understand the impact and effectiveness of such health alerts and interventional actions on the spread of epidemics.

C. Methodology

Telefonica Research tested the impact of the actions taken by the Mexican government on H1N1 during April and May of 2009 by using call data records. The actions taken by the Mexican government included health alerts and regulations on reducing mobility. On 16 April, the authorities raised a medical alert (stage 1), followed by the closing of schools and universities on 27 April (stage 2), and the final shutdown of all basic activities on 1 May (stage 3). To test the impact of the actions, the researchers analysed cell phone call data records for 1,000,000 anonymized customers of the most affected Mexican states obtained from January 2009 to May 2009. Using these data, population mobility analysis and geographic mobility analysis were conducted (see Table 2).

Table 2. Types of mobility analysis

Population Mobility Analysis	Geographic Mobility Analysis
• Focuses on comparing the aggregated mobility of the population during the different alert stages with a baseline intended to characterize typical mobility behaviour	• Evaluates the impact that the government alerts had on specific geographic areas that contain critical infrastructure like airports or hospitals • Aims to understand whether the number of individuals that visited these infrastructures varied as a result of the government mandates

Source : Vanessa Frias-Martinez, Alberto Rubio and Enrique Frias-Martinez, "Measuring the Impact of Epidemic Alerts on Human Mobility", paper presented at the Second Workshop on Pervasive Urban Applications, Newcastle, UK, 2012.

The results showed that medical alerts (stage 1) did not seem to significantly affect human mobility, whereas interventional actions (stages 2 and 3) significantly did, particularly if the intervention took place during regular working days. A direct consequence for the design of epidemic alerts is that the enactment of a total closure of activities during a holiday period is not as effective for slowing down the spread of the epidemic as the partial closing of some activities (typically schools) during regular workdays.

33 Vanessa Frias-Martinez, Alberto Rubio and Enrique Frias-Martinez, "Measuring the Impact of Epidemic Alerts on Human Mobility", paper presented at the Second Workshop on Pervasive Urban Applications, Newcastle, UK, 2012.

Also, they found that the increase in number of visitors that the airport received during the stage 2 alert implies that closing of infrastructures (stage 3 alert) might provoke an increase in the number of individuals that visit transport hubs before its enactment, thus limiting the containment and possibly causing an undesired increase in the spread of the epidemic.

D. Implications

Based on the analytics of call data records and preventive actions of the government, it was able to reduce citizen mobility by a maximum of 30 per cent, and this in turn resulted in a decrease in the number of infected cases by 10 per cent. In addition, these actions helped to postpone the epidemic by 40 hours, thereby allowing government authorities to react faster to control the epidemic.

This project shows how big data created new value by segmenting the population in the country for customized and targeted actions while creating transparency between citizens and government and improving the performance of the government.

3.3. Big Data Analysis for Early Warning

3.3.1. Predicting Unemployment Spike through Social Media Analysis[34]

A. Overview

This project, implemented by the United Nations Global Pulse in partnership with SAS, is an example of how social media and other online user-generated content can be used to provide early warning on unemployment spike.

B. Background

Unemployment rate prediction helps government make decisions and design policies to cope with the effects of unemployment. Traditionally, unemployment figures were published as part of government labour statistics. Recently, the United Nations Global Pulse in partnership with SAS adopted a data mining framework using social media contents to predict unemployment rate.

This project analyses whether and how the feelings and information shared on social media could enrich understanding of the effect of changing employment conditions on people's perceptions and decisions. Also, by comparing the qualitative information offered by social media with official unemployment figures, it tests whether online conversations can provide an early indicator of impending job losses.

C.Methodology

For the project, researchers analysed the cases of the United States of America and Ireland where social media and blogs are widely used and where the ongoing economic crisis has had a significant impact on employment. By comparing the qualitative

34 SAS and United Nations Global Pulse, Can a Country's Online "Mood" Predict Unemployment Spikes? 12 March 2012. Available from http://www.sas.com/en_us/news/press-releases/2012/march/un-sma.html; and SAS and United Nations Global Pulse, Using Social Media and Online Conversations to Add Depth to Unemployment Statistics, Methodological White Paper, 8 December 2011. Available from http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics.

information from social media with official unemployment figures, the project aimed to identify: 1) whether online conversations provide an early indicator of impending job losses, and 2) how they can help policymakers better understand the type and sequence of coping strategies used by individuals better.

In order to address the questions, the United Nations Global Pulse and SAS first automatically retrieved online job-related conversations from blogs, forums and news from the United States of America (430,000 documents) and Ireland (28,000 documents) from 2009 to 2011. Then they assigned a quantitative mood score to each document based on the tone or mood of the conversations (e.g. happiness, depression, anxiety). Those unemployment related documents dealing with other topics such as housing and transportation were also quantified and categorized into pre-defined lists of topics representing potential coping mechanisms. The indicators related to unemployment spike are depicted in figure 8.

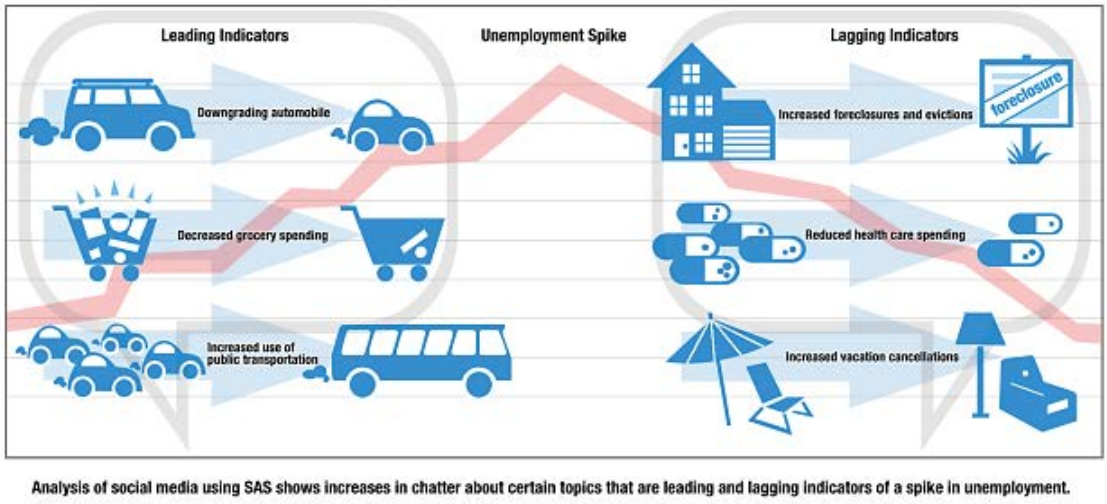

Figure 8. Indicators of unemployment spike

Source : SAS and United Nations Global Pulse, Can a Country's Online "Mood" Predict Unemployment Spikes? 12 March 2012. Available from http://www.sas.com/en_us/news/press-releases/2012/march/un-sma.html.

Finally, these measures—aggregated mood scores and the volume of conversations around different topics—were analysed and compared with official unemployment statistics over time in search of interesting dynamic correlations.

Data Acquisition	→	Data Filtering	→	Data Validation	→	Data Analysis
Relevant data set from public Internet sources retrieved		Eliminate noises and sort the documents by topic		Review representative sample of the data to check data validity		a. Sentiment analysis to know how the unemployed feel about their status b. Cross-correlation analysis to find relationships among mood and coping mechanisms

Figure 9. Data analysis process

D. Implications

The researchers found that increased conversation about cutting back on groceries, increasing use of public transportation and downgrading one's automobile could, indeed, predict an unemployment spike. After a spike, surges in social media conversations about such topics as cancelled vacations, reduced health care spending, and foreclosures or evictions shed light on lagging economic effects.

While official unemployment statistics measures unemployment claims, the information drawn from social media shows aggregated signs of unemployment in the casual online conversation of millions of people. Such information, which might give us more information about the same topic but at a finer scale, could be invaluable for policymakers trying to mitigate negative effects of increased unemployment.

This project shows how big data is creating new value in terms of providing real-time feedback for policymakers and improving the ability to manage disruptive events. By doing so, it can contribute to discovering the needs of the citizens and improving their standard of living.

3.3.2. Google Flu Trends for Monitoring Flu Trends

A. Overview

This project is an example of how big data can be utilized to predict the prevalence of the flu and find better ways to do surveillance for outbreaks of influenza or any other diseases using Google searches.

B. Background

In 2008, Google launched a project called Google Flu Trends, which attempts to predict the prevalence of the flu from searches that users made for about 40 flu-related queries. It currently provides estimates of influenza activity for more than 25 countries. It has recently come under attack from academia for wildly overestimating the rate of flu infection in the United States of America.

C. Methodology[35]

Google Flu Trends monitors health tracking behaviour of millions of users online and analyses them for any flu-like illness presents in a population. Google Flu Trends has found a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. By counting how often flu-related Google search queries appear, it can estimate how much flu is circulating in different countries and regions around the world. Also, it compares these findings to a historic baseline level of influenza activity for its corresponding region and reports the activity level as either minimal, low, moderate, high or intense.[36]

35 Jeremy Ginsberg and others, "Detecting Influenza Epidemics using Search Engine Query Data", *Nature*, vol. 457 (2009), pp. 1012-1014.

36 Roni Zeiger, "Google Flu Trends Overview on Youtube", *Google.org,* 6 October 2009. Available from http://www.youtube.com/watch?v=6111nS66Dpk.

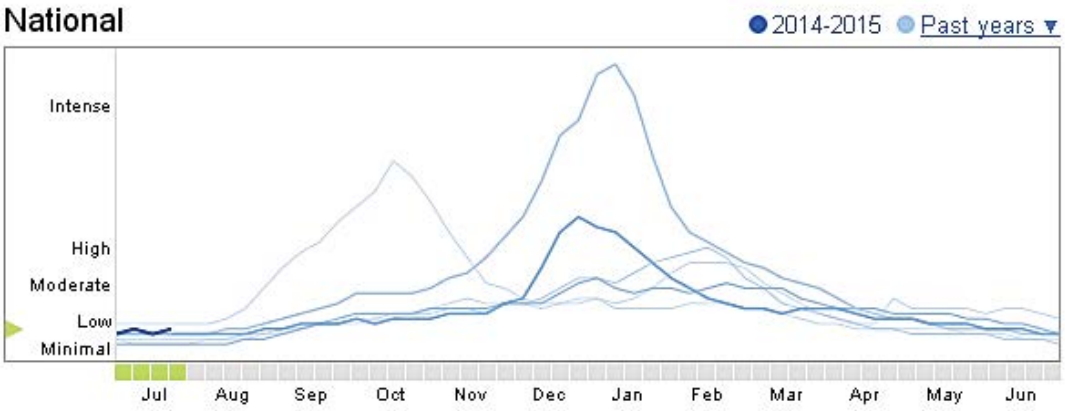

Figure 10. Display of Google Flu Trends

D. Implications

The initial Google paper stated that the Google Flu Trends predictions were 97 per cent accurate when compared with official Centers for Disease Control and Prevention data.[37] It turned out the result of flu predictions by Google Flu Trends, however, were overestimated by more than fifty percent in the 2011-2012 and 2012-2013 seasons. From August 2011 to September 2013, Google Flu Trends over-predicted the prevalence of the flu in 100 out of 108 weeks.[38]

This clearly shows that just having massive amount of data collected over a period of time does not mean that it may be the right data that could help provide a true picture of what is happening. In fact, Google Flu Trends and other big data methods can be useful but only when they are based on accurate data and modelling. This highlights the need for policymakers to question and check the data collection method, modelling and simulation of big data, and ensure that their decisions are based on sound data centric foundation.[39]

37 Jeremy Ginsberg and others, "Detecting Influenza Epidemics using Search Engine Query Data", *Nature,* vol. 457 (2009), pp. 1012-1014.

38 David Lazer, Ryan Kennedy, Gary King and Alessandro Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis", *Science,* vol. 343, no. 14 (March 2014), pp. 1203-1205. Available from http://j.mp/1ii4ETo.

39 Stephen E. Arnold, "Google Flu Trends: How Algorithms Get Lost", *Beyond Search,* 15 March 2014. Available from http://arnoldit.com/wordpress/2014/03/15/google-flu-trends-how-algorithms-get-lost/; and Alexic C. Madrigal, "In Defense of Google Flu Trends", *The Atlantic,* 27 March 2014. Available from http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/.

4. Challenges of Big Data for Development

4. Challenges of Big Data for Development

Big data, with its tremendous wealth of information, promises ample opportunities to enhance our ability to understand the human behaviour and pursue development. Application of big data for global development is, however, still in its inception and several challenges should be addressed in order to fulfil its promises. According to a survey conducted by United Nations Economic and Social Council (ECOSOC) in 2013,[40] the major challenges in using big data for development include analytics, privacy, data access, infrastructure and human capacity.[41]

4.1. Analytics

The majority of big data is unstructured, unfiltered "data exhaust" (passively collected transactional data from people's use of digital services) from digital products such as electronic and online transactions, social media and digital sensors.[42] Hence, one of the major challenges in using big data for development is drawing values out of the bulk of data with suitable methodologies to process, analyse and interpret the data.

4.1.1. Data Quality and Analysis

The first challenge that is likely to face the big data users lies in the quality of data itself. In many cases, big data is incomplete or missing, or stored in a format that is inaccessible to automated processing. The examples include video, social media contents, medical records including MRI pictures, surveillance camera footages and GPS records, all of which are in different formats and need to be processed into a machine readable format prior to data analysis.

Oftentimes, the validity of data inputs is also questioned because they may be missing, falsified or fabricated by its producers in the first place. One can easily imagine a situation where social media users post false or wrong information online. Therefore, cleaning up and verifying the data is an increasingly important but often ignored job that can help prevent costly mistakes in the first place.

Due to the huge volume of data, big data also requires an information extraction process using scalable data analysis techniques. In this information extraction process, however, there is always a possibility of distortion or manipulation of data, and the discarding of useful information. Therefore, it is critical to have in-depth research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not losing useful information.

40 United Nations Economic and Social Council, Big Data and Modernization of Statistical Systems, Report of the Secretary-General, Statistical Commission, 4-7 March 2014.

41 In the survey, methodological, information technology and management challenges were indicated most often as challenges, followed closely by legislative and privacy challenges.

42 Ramon G. Jose Albert, "Challenges, Opportunities and Issues on Using Big Data for Meeting Current and Emerging Demands on Measuring Progress and Development", paper presented at the Meeting on the Management of Statistical Information Systems, United Nations Economic Commission for Europe, 14-16 April 2014.

4.1.2. Data Interpretation

In interpreting the results of big data analysis, one of the key challenges addressed by its users is the representativeness of the data. For example, the people who use mobile or digital services and thus, generate a big portion of real-time big data may just be a certain group of people who have better access to mobile service and/or a relatively young population. In this case, they cannot be a representative sample of the larger population and is likely to lead to a different result from a population survey.

There are also risks and challenges associated with drawing valid inferences and conclusions from the data analysis. For example, people may try to find patterns and correlations in the massive quantities of data when, in fact, they do not exist. Also, even where an actual correlation is found, it does not necessarily show that there is a causation link. Many people, however, rush to judgments without a solid understanding of the deeper dynamics at play.

The issues discussed so far are just a few of the key challenges related to data analysis interpretation of big data. In order to avoid these problems, it is important to work with analysts who are fully aware of these limitations of big data analytics and make decisions within acceptable boundaries. Also, it is critical for a decision maker to be able to interpret the results recognizing the interplay of large-scale data and more detailed, qualitative small data.

4.2. Privacy

Big data contains various types of passively collected data such as Twitter comments, browsing habits in search engines, private information in social networking sites and mobile call records, which raise concerns about threat to privacy. For example, a large set of anonymized mobile call records can be used to create fingerprints of users, which when combined with other data such as geo-located tweets or check-ins could reveal the individual's identity.

The issue of privacy becomes particularly sensitive due to the possibility of personalized data being used for control, which can harm or manipulate democratic processes. For example, the use of big data analysis in election can be a double-edged sword because, while identifying the needs of the electorate can help to set policy agendas, it can also be used to help candidates spin a message to please an identified group of interest.[43]

As the amount of personal data and digital information rapidly increases in this digital era, big data analytics are increasingly being used to make decisions about people who have a right to know on what basis those decisions are made. This phenomenon calls for

43 International Telecommunication Union, "Big Data, Big Deal, Big Challenge", *ITU News,* Issue no. 1, January 2014. Available from https://itunews.itu.int/En/4848-Big-data-big-deal-big-challenge.note.aspx.

the right technical, legal and ethical standards for big data and for individuals. Therefore, it must be assured and guaranteed that personal data will be used appropriately by creating adequate data governance standards and regulations that define how and by whom data are collected, stored and curated for accountability. Governments also need to continuously monitor whether personal data are used in the context of its intended use and abiding by the relevant laws.

4.3. Data Access

In this era of big data, there is an overwhelming amount of useful data publicly available online from user-generated content such as news, blogs and social media, to the structured data being shared through open data initiatives.[44] However, there is a tremendous wealth of data, known as "massive passive data" or "data exhaust" held by corporations, which is not accessible. It is the digital traces that we leave behind in our daily lives and personal information collected by the private sector. These data have proven to be extremely valuable for policymakers and development practitioners to get real-time feedback on how well their development policies and programmes are working.

While there is clear evidence that such real-time data are needed for better policy decision-making for development, private companies and other non-governmental entities have been reluctant to share data about their clients and users due to various reasons, including protection of their customers and their competitiveness, reputation and legal considerations, among others. In order to improve access to data held by the private sector, the United Nations Global Pulse has advocated the concept called "data philanthropy", where corporations are encouraged to share anonymized data for use by the public sector to protect vulnerable populations. Corporations are participating in this initiative recognizing that more effective policy action will lead to greater resilience from economic shocks, and therefore translate into better business continuity.[45]

This United Nations Global Pulse initiative for greater data access is just a beginning for utilizing big data from non-governmental sources. More research and initiatives should be carried out on the business models and appropriate incentives for the private sector to share their data with the public sector, while also protecting both the privacy of their customers and their edge over the competition.

4.4. Infrastructure

Big data represents significant opportunities for policymakers and practitioners to enhance their decision-making process for socio-economic development. However, drawing out the trends and values hidden within big data takes infrastructure as well as human and

44 Robert Kirkpatrick, "Data Philanthrophy: Public & Private Sector Data Sharing for Global Resilience", United Nations Global Pulse Blog, 16 September 2011. Available from http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience.

45 World Economic Forum, *Big Data, Big Impact: New Possibilities for International Development* (Geneva, 2012). Available from http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

technical capacity, which may lead to a new kind of digital divide between developed and developing countries.

Today, it is not possible to store all the data that is produced every second in the existing data storage infrastructure. Moreover, the relational database management system, which has been a predominant choice for the storage of information since the 1980s, is being challenged to scale to the levels to meet the growing demand of storing big data. A new technology architecture, which is capable of supporting both structured and poly-structured data that need different types of storage infrastructure, is required in the era of big data.

Other challenges that policymakers should be aware of for big data analytics is that it requires efficient and scalable algorithms, scientific basis for selection of method or design of the model, data visualization, and human-computer interfaces that can handle big data. The tools used for big data analytics should be able to perform multimedia analytics (i.e., automatic content analysis of digital image, audio and video data allowing for automatic and rapid extraction of events, metadata or other meaningful information). The tools for big data analytics should also be able to handle multiple time scales, discover and model latent causality, and integrate structured and unstructured causality models.

As shown in the investment in information technology around the world, information and computation capacity are mostly concentrated in highly developed countries. In addition to hardware infrastructure, big data also heavily depends on software services to analyse the data. Basic capabilities in the production, adoption and adaptation of software products and services are a key ingredient for thriving in big data environment, and developing countries have even less capabilities for software and computer services.

4.5. Human Capacity

In order to exploit big data for development, human capacities in statistics and data analytics are crucial. However, it is much harder in developing countries to find statisticians, data specialists and data mangers that are equipped with programming and design capacity.

One of the most sought after skills is that of the data scientist in the big data environment. The data scientist is a person who looks for patterns and insights in large data sets and carries out correlations that may not be immediately obvious to real-world issues. The data scientist may find a visible trend in the data set through the exploratory analysis process, which leads to a hypothesis or theory. The data scientist works with a team of data analysts (who decide which data analysis and visualization methods would be the best fit for the project) and data engineers (who create and maintain the big data technical infrastructure). Working with the team, the data scientist produces a summary of the findings and share the learning at a broader level for others to replicate or carry out further work on. The data scientist has overlapping skills in business, programming, analytical

skills and domain knowledge. One of the organizational challenges that would need to be resolved is the creation of trust between the data scientist and the business manager.

Despite such challenges in human resource capacity for big data, several governments in the developing countries such as Kenya and Ghana are working toward the better use of big data for development. Kenya, for example, introduced the Open Data Portal in 2011, making key government data freely available to the public through a single online portal, and has built information technology activity in the private sector through government grants and incubation programmes.[46] Ghana also launched the Open Data Initiative in 2012 to make Government of Ghana data available to the public for re-use. Although their performance is still contested and significant challenges such as limited government buy-in and a weak legal framework lie ahead, some more concrete measures for opening up public data and ICT initiatives are being undertaken, opening up greater possibilities of using big data for development.

Since there is a huge gap in the demand and supply of analytics roles such as data scientists, data analysts and data engineers as well as managers for big data project, the academia needs to gear up for filling the supply gap by ensuring that they produce high quality multi-disciplinary skilled resources. There is also an opportunity for government to collaborate and engage with industry and academia to learn and leverage experience and expertise while investing in public-private partnerships to gain valuable insights.

[46] Jan Piotrowski, "Big Obstacle Ahead for Big Data for Development", *SciDevNet,* 15 April 2014. Available from http://www.scidev.net/global/data/feature/obstacles-big-data-development.html; and Greg Brown, "Why Kenya's Open Data Portal is Failing – and Why It Can Still Succeed", *Sunlight Foundation,* 23 September 2013. Available from http://sunlightfoundation.com/blog/2013/09/23/why-kenyas-open-data-portal-is-failing-and-why-it-can-still-succeed/.

5. Next Steps for Big Data for Development

5. Next Steps for Big Data for Development

The world is increasingly driven by data as more and better data become available. The data revolution has already happened in the private sector and now large-scale investment is flowing into establishing big data capabilities in many governments and various public and private organizations.[47]

While big data has great potentials to change the world and how we do business, its benefits will depend on how well its risks and challenges are managed. Policymakers therefore need to prepare and create an environment that is conducive and supportive to those organizations seeking to benefit from exploiting big data for socio-economic development. Considering the major challenges in promoting big data for development, several policy implications can be drawn for policymakers and practitioners for future big data strategy.

Firstly, policymakers must establish policies and legal frameworks to protect data privacy and security. It will help governments as well as public and private organizations know what types of data they can store and share, and which data are forbidden by regulations. Policymakers also need to pursue a common understanding on the usage of big data with citizens and assure them that the use of big data is consistent with the public interest. Without strong political buy in and public trust, big data initiatives will not be sustained in the long term.
Secondly, policymakers and executives of public organizations can identify the owner for "big data" in the organization and formally establish a "Chief Data Officer/Scientist" for effective data management.[48] By doing so, they will be able to position big data as an integral element of data management in government and establish a data-drive decision culture.

Thirdly, governments need to take the lead in opening public data so that these data can be easily accessed with minimal limitations on its use. Government agencies have the capacity and resources to gather extensive data, which can be utilized to provide major socio-economic implications through its analysis. Such open data initiatives will also be able to provide a basic platform to collect and share big data.

Lastly, in many cases, developing countries have limited resources to gather and utilize big data by themselves. In that case, external data providers such as corporations or international organizations can complement existing data-gathering efforts through data sharing agreements or data collaborations. Meanwhile, the global community could also consider creating large data banks on various development issues, which have the capacity to hold various types of data. For this to happen, multi-sector alliances that promote data sharing on thematic issues will have to be created.[49]

47 Bahjat El-Darwiche and others, "Big Data Maturity: An Action Plan for Policymakers and Executives", in *The Global Information Technology Report 2014,* World Economic Forum and INSEAD (Geneva, 2014). Available from http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf.

48 Ibid.

49 Kevin C. Desouza and Kendra L. Smith, "Big Data for Social Innovation", *Stanford Social Innovation Review,* Summer 2014. Available from http://www.ssireview.org/articles/entry/big_data_for_social_innovation.

Big data is a fast-moving technology that will affect all aspects of our lives. Managing big data requires a significant change in the way that governments and businesses operate and their capabilities. Forward-looking policymakers and practitioners will embrace the opportunities afforded by big data for their national development and provide the necessary support to realize the potential by balancing the risks and rewards of big data.

REFERENCES

Albert, Ramon G. Jose (2014). Challenges, Opportunities and Issues on Using Big Data for Meeting Current and Emerging Demands on Measuring Progress and Development. Paper presented at the Meeting on the Management of Statistical Information Systems. United Nations Economic Commission for Europe, 14-16 April.

Allouche, Gil (2014). Big Data is Transforming Every Industry, from Health and Education, to Farming and Energy. Betanews, 31 July. Available from http://betanews.com/2014/07/31/big-data-is-transforming-every-industry-from-health-and-education-to-farming-and-energy/.

Arnold, Stephen E. (2014). Google Flu Trends: How Algorithms Get Lost. Beyond Search, 15 March. Available from http://arnoldit.com/wordpress/2014/03/15/google-flu-trends-how-algorithms-get-lost/.

Arpaci-Dusseau, Remzi H. and Andrea C. Arpaci-Dusseau (2014). Redundant Arrays of Inexpensive Disks. Available from http://pages.cs.wisc.edu/~remzi/OSTEP/file-raid.pdf.

Beyer, Mark A. and Douglas Laney (2012). The Importance of 'Big Data': A Definition. Gartner, 21 June. Available from https://www.gartner.com/doc/2057415/importance-big-data-definition.

Brown, Greg (2013). Why Kenya's Open Data Portal is Failing – and Why It Can Still Succeed. Sunlight Foundation, 23 September. Available from http://sunlightfoundation.com/blog/2013/09/23/why-kenyas-open-data-portal-is-failing-and-why-it-can-still-succeed/.

Computer Sciences Corporation (2012). Big Data Universe Beginning to Explode. Available from http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode.

Conway, Gordon (2014). Big Data: Big Hope or Big Risk? One Billion Hungry: Can We Feed the World? Blog, 17 April. Available from http://canwefeedtheworld.wordpress.com/2014/04/17/big-data-big-hope-or-big-risk/#more-1093.

Desouza, Kevin C. and Kendra L. Smith (2014). Big Data for Social Innovation. Stanford Social Innovation Review, Summer. Available from http://www.ssireview.org/articles/entry/big_data_for_social_innovation.

Dutcher, Jenna (2013). Data Size Matters [Infographic]. Datascience @Berkeley Blog, Berkeley School of Information, 6 November. Available from http://datascience.berkeley.edu/big-data-infographic/.

Eastwood, Brian (2013). 6 Big Data Analytics Use Cases for Healthcare IT. CIO, 23 April. Available from http://www.cio.com/article/2386531/healthcare/6-big-data-analytics-use-cases-for-healthcare-it.html.

El-Darwiche, Bahjat and others (2014). Big Data Maturity: An Action Plan for Policymakers and Executives. In The Global Information Technology Report 2014, World Economic Forum and INSEAD. Geneva. Available from http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf.

Frias-Martinez, Vanessa and Enrique Frias-Martinez (2012). Enhancing Public Policy Decision Making using Large-Scale Cell Phone Data. United Nations Global Pulse, 4 September. Available from http://www.unglobalpulse.org/publicpolicyandcellphonedata.

Frias-Martinez, Vanessa, Alberto Rubio and Enrique Frias-Martinez (2012). Measuring the Impact of Epidemic Alerts on Human Mobility. Paper presented at the Second Workshop on Pervasive Urban Applications, Newcastle, UK.

Frias-Martinez, Vanessa and others (2012). Computing Cost-Effective Census Maps from Cell Phone Traces. Paper presented at the Second Workshop on Pervasive Urban Applications, Newcastle, UK.

Ginsberg, Jeremy and others (2009). Detecting Influenza Epidemics using Search Engine Query Data. Nature, vol. 457, pp. 1012-1014.

Hilbert, Martin (2013). Big Data for Development (pre-published version).

Hilbert, Martin and Priscila Lopez (2011). The World's Technological Capacity to Store, Communicate and Compute Information. Science, vol. 332, no. 6025, pp. 60-65.

Hilbert, Martin and Priscila Lopez (2012). How to Measure the World's Technological Capacity to Communicate, Store and Compute Information? Part I: Results and Scope. International Journal of Communication, vol. 6, pp. 956-979.

International Telecommunication Union (2014). Big Data, Big Deal, Big Challenge. ITU News, Issue no. 1, January. Available from https://itunews.itu.int/En/4848-Big-data-big-deal-big-challenge.note.aspx.

International Telecommunication Union (2014). The World in 2014: ICT Facts and Figures. Geneva. Available from http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx.

Internet Live Stats (2014). Google Search Statistics. Available from http://www.internetlivestats.com/google-search-statistics/. Accessed on 7 August 2014.

Jetley, Neerja Pawha (2014). How Big Data Has Changed India Elections. CNBC, 10 April. Available from http://www.cnbc.com/id/101571567.

Kapow Software (2013). Understanding The Various Sources of Big Data – Infographic. Big Data Start-up. Available from http://www.bigdata-startups.com/BigData-startup/understanding-sources-big-data-infographic/#!prettyPhoto.

Kirkpatrick, Robert (2011). Data Philanthrophy: Public & Private Sector Data Sharing for Global Resilience. United Nations Global Pulse Blog, 16 September. Available from http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience.

Kresge, Jill Kristen (2007). HIV Prevalence Estimates: Fact or Fiction – IAVI Report. The Publication on AIDS Vaccine Research, vol. 11, no. 4.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science, vol. 343, no. 14 (March), pp. 1203-1205. Available from http://j.mp/1ii4ETo.

Letouze, Emmanuel (2014). Big Data for Development: Facts and Figures. SciDevNet, 15 April. Available from http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html.

Maccarthy, Mark (2014). Big Data Improves Education Around the World. SIIA Digital Disclosure, 18 April. Available from http://www.siia.net/blog/index.php/2014/04/big-data-improves-education-around-the-world/.

Madrigal, Alexic C. (2014). In Defense of Google Flu Trends. The Atlantic, 27 March. Available from http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/.

Melamed, Claire (2014). Development Data: How Accurate are the Figures? The Guardian Poverty Matters Blog, 31 January. Available from http://www.theguardian.com/global-development/poverty-matters/2014/jan/31/data-development-reliable-figures-numbers.

Ng, Kelly (2014). Singapore Government Uses Big Data Analytics to Optimise Transport Management. FutureGov, 17 April. Available from http://www.futuregov.asia/articles/2014/apr/17/singapore-government-uses-big-data-analytics-optim/.

Pappas, Stephanie (2014). 17 Developing Countries That Love Social Media More than the US. Live Science, 13 February. Available from http://www.livescience.com/43347-developing-countries-social-media.html.

Piotrowski, Jan (2014). Big Obstacles Ahead for Big Data for Development. SciDevNet, 15 April. Available from http://www.scidev.net/global/data/feature/obstacles-big-data-development.html.

Quitzau, Anders (2014). Transforming Energy and Utilities through Big Data & Analytics. IBM Corporation, 28 March. Available from http://www.slideshare.net/AndersQuitzauIbm/big-data-analyticsin-energy-utilities.

Raghupathi, Wullianallur and Viju Raghupathi (2014). Review of Big Data Analytics in Healthcare: Promise and Potential. Health Information Science and Systems, vol. 2, no. 3.

SAS and United Nations Global Pulse (2011). Using Social Media and Online Conversations to Add Depth to Unemployment Statistics. Methodological White Paper, 8 December. Available from http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics.

SAS and United Nations Global Pulse (2012). Can a Country's Online "Mood" Predict Unemployment Spikes? 12 March. Available from http://www.sas.com/en_us/news/press-releases/2012/march/un-sma.html.

Sniders, Chris, Uwe Matzat and Ulf-Dietrich Reips (2012). Big Data: Big Gaps of Knowledge in the Field of Internet. International Journal of Internet Science, vol. 7, pp. 1-5.

United Nations (2013). What is Data Revolution. United Nations High-Level Panel: the Post-2015 Development Agenda, August.

United Nations Economic and Social Council (2014). Big Data and Modernization of Statistical Systems. Report of the Secretary-General, Statistical Commission, 4-7 March.

United Nations Global Pulse (2013). Big Data for Development: A Primer. Available from http://www.unglobalpulse.org/sites/default/files/Primer%202013_FINAL%20FOR%20PRINT.pdf.

Woody, Todd (2013). Big Data is Giving China an Edge in Renewable Energy Production. QUARTZ, 8 August. Available from http://qz.com/113547/big-data-is-giving-china-an-edge-in-renewable-energy-production/#/h/4214,2,3/.

Zeiger, Roni (2009). Google Flu Trends Overview on Youtube. Google.org, 6 October. Available from http://www.youtube.com/watch?v=6111nS66Dpk.

World Economic Forum (2012). Big Data, Big Impact: New Possibilities for International Development. Geneva. Available from http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

GLOSSARY

	Definition
Big Data	Data sets so large and complex that they become difficult to work with using standard statistical software.
Call Data Record	A file containing information about recent system usage such as the identities of sources (points of origin), the identities of destinations (endpoints), the duration of each call, the amount billed for each call, the total usage time in the billing period, the total free time remaining in the billing period, and the running total charged during the billing period.
Causation (in statistics)	Causal statistical relationship between conduct and result.
Centers for Disease Control and Prevention	The national public health institute of the United States of America.
Chief Data Officer/ Scientist	A corporate officer responsible for enterprise-wide governance and utilization of information as an asset, via data processing, analysis, data mining, information trading and other means. The Chief Data Officer can have various reporting lines including to the Chief Technology Officer, Chief Information Officer, Chief Executive Officer, Chief Marketing Officer or Chief Strategy Officer.
Cloud Storage	A model of data storage where the digital data is stored in logical pools, the physical storage spans across multiple servers (and often locations), and the physical environment is typically owned and managed by a hosting company.
Correlation (in statistics)	Any of a broad class of statistical relationships involving dependence.
Data Exhaust	Unstructured information or data that is a by-product of the online activities of Internet users.
Data Mining	The computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
Data Philanthropy	A new form of partnership in which private sector companies share data for public benefit.
Data Revolution	Realizing that the availability of data empowers people by arming them with information and encourages data-driven decision-making for development, the High Level Panel on the Post-2015 Development Agenda called for a “data revolution”, a new international initiative to improve the quality and scope of data available to citizens and policymakers
Data Validation	The process of ensuring that a program operates on clean, correct and useful data.
Open Data	The idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

Redundant Array of Independent Disks	A way of storing the same data in different places, thus, redundantly on multiple hard disks.
Remote Sensing Equipment	The acquisition of information about an object or phenomenon without making physical contact with the object. In modern usage, the term generally refers to the use of aerial sensor technologies to detect and classify objects on Earth (e.g. satellite).
Social Networking Service or Social Networking Site	A website with multiple users where they can publish content.
Structural Data	Data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets. Structured data can be quantified or easily tagged, categorized and organized for systematic analysis.
TB (Terabyte)	A multiple of the unit byte for digital information. 1 TB = 1000000000000bytes = 1012bytes = 1000gigabytes.
Unmanned Aerial Vehicle	Commonly known as a drone and referred to as a remotely piloted aircraft by the International Civil Aviation Organization. It is an aircraft without a human pilot aboard. Its flight is controlled either autonomously by onboard computers or by the remote control of a pilot on the ground or in another vehicle.
Unstructured Data	Information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers and facts as well.

UN-APCICT/ESCAP
United Nations Asian and Pacific Training Centre for Information
and Communication Technology for Development
5th Floor G-Tower, 175 Art center daero,
Yeonsu-gu, Incheon City, Republic of Korea

www.unapcict.org

UNITED NATIONS ESCAP
Economic and Social Commission for Asia and the Pacific
/APCICT

Published under programme supported by Korea International Cooperation Agency