The World Bank Group

BIG DATA AND URBAN MOBILITY - CAIRO JUNE 2, 2014

BACKGROUND DOCUMENT

Jacqueline Dubow

6/2/2014

ICT Information and Communication Technologies

INNOVATE · CONNECT · TRANSFORM

www.worldbank.org/ict

Contents

Tables

1.Introduction

Between 2010 and 2050, the number of people living in urban areas worldwide is expected to grow by 80%—from 3.5 billion to 6.3 billion. This growth will create problems for urban mobility by increasing congestion, raising Green House Gas (GHG) emissions, and accelerating the deterioration of transportation infrastructure. Can urban planners and transportation policy decision makers curb some of these trends by better planning? What information is available for them to enhance the relevance and accuracy of their projections? Beyond the traditional sources of information (such as the expensive and complex origin/destination surveys), the influx of Big Data is creating a paradigm shift toward new modes of transportation, such as ride sharing, pedestrianism, and public transport.

Big Data is the result of the many innovations in technologies and greater affordability of digital devices. We are witnessing an explosion in the quantity and diversity of high frequency digital data. These data—still largely untapped in developing countries—have the potential to allow decision makers to track development progress, improve social protection, and understand where existing policies and programs require adjustment.

Big Data may come from mobile phone call logs, mobile-banking transactions, online user-generated content—such as blog posts and Tweets, online searches, and satellite images. Turning Big Data into actionable information is based on computational techniques that bring trends and patterns within and between extremely large datasets.

The private sector has already started investing in Big Data as a new way to stimulate innovation and productivity growth. The public sector is also an important data user, and a source of data that can generate benefits across the economy. Evidence from Europe's 23 largest governments shows that by fully exploiting public sector data, governments could reduce their administrative costs; some estimate potential savings of 15% to 20%—the equivalent of EUR 150 billion to EUR 300 billion in new value. Such benefits can be obtained from weather forecasts, traffic management, crime statistics, improved transparency of government functions (e.g. procurement) and educational and cultural knowledge for the wider population[1].

This paper presents the challenges—data relevance, reliability and privacy issues, as well as the opportunities—the ability to map unstructured data , to make better informed decisions- obtained by the use of Big Data in developing countries, and more specifically within the transport sector. Big Data are offering new opportunities to use powerful tools to support the fight against poverty and increase growth.

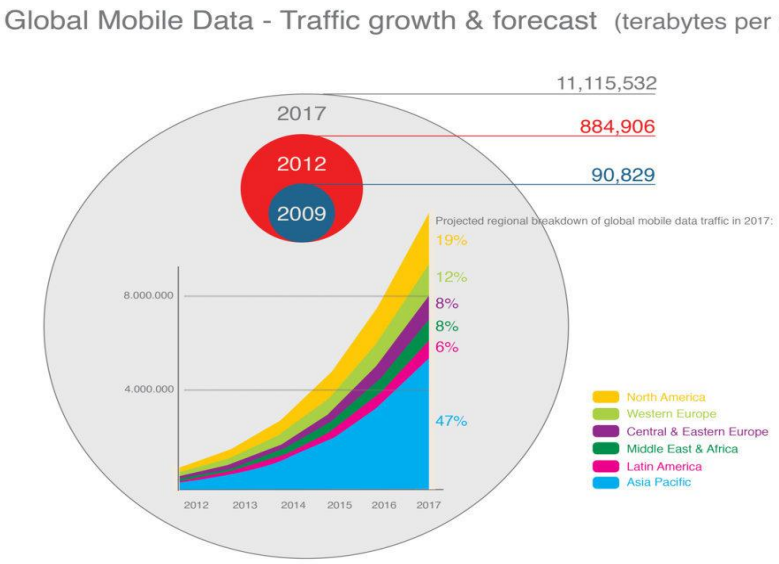
[1] Exploring Data-driven innovations as a New Source of Growth, OECD, June 2013

2.What is Big Data?

The Four "V": Volume, Velocity, Variety and Veracity

The digitization of our everyday activities—such as using mobile phones (seven to eight billion mobile phones worldwide), travel, shopping, downloading music, uploading videos to YouTube , posting on Facebook, and sending tweets—is creating a massive amount of data. At the same time, the dependence on electronic devices is increasing, and every time we are using a device or a digital service, we are leaving digital footprints. This massive amount of digital information (figure 1) represents the main source of Big Data.

Figure 1: Global Mobile Data 2014 - Traffic growth and forecast



Big Data can be defined as a vast collection of structured and unstructured data sets that have become difficult to process using traditional data processing tools due to the volume and complexity of the data. They are characterized by complex high-volume, high-velocity, high-variety and high-veracity information that require innovative information processing. Not all of the four "V" are necessary to be present to result in Big Data. For example, a data set low in volume but high in veracity and complexity, still qualifies as Big Data[2]. Equally important, Big Data is not just about the size of the data, it has more to do with the way data are being processed.

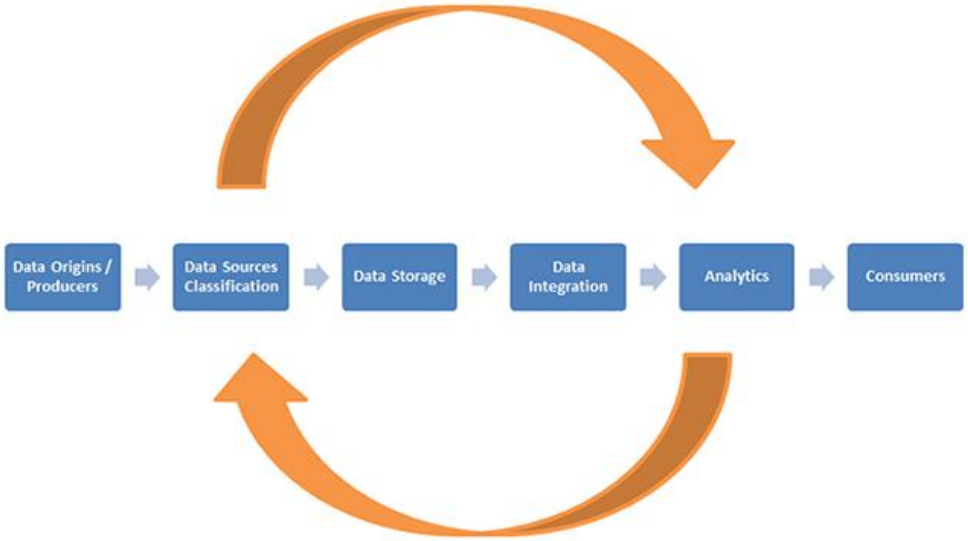
Big Data help in making better decisions by enabling data scientists and other users to analyze huge volumes of transaction data, as well as other data sources that may be left untapped by conventional business intelligence programs. They are quite different from what a statistics

[2] http://www.business2community.com

department does. However, it is important to note that Big Data is still at its beginning, especially in the area of development—but it is gaining ground.

Big Data value Chain

Figure 2: Big Data Value Chain



Source: B2C[3]

As shown in figure 1, Big Data is a new way to convey knowledge. The data set needs to be gathered from the "producers", such as the Internet—click-stream data, social media activity reports, mobile-phone call detail records, and information captured by sensors. The data need to be classified according to their sources, images, videos, and texts, and then stored using technologies supporting data storage. The next process—data integration—involves combining data residing in different sources, and providing users with a unified view of these data. The analytics supports the mining of the data, the determination of what is relevant, and the discovery of patterns and relationships to help decision making. More importantly, data analytics are predictive analytics. They give the probability of different outcomes, and are future-oriented. Big Data analytics can be done with the software tools commonly used as part of advanced analytics disciplines, such as predictive analytics and data mining. However, the unstructured data sources used for Big Data analytics may not fit in traditional data warehouses. Furthermore, traditional data warehouses may not be able to handle the processing demands posed by Big Data. The technologies associated with Big Data analytics include—among others—NoSQL, Hadoop, and MapReduce. These technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems. Consumers and citizens—often through applications—are the ultimate beneficiaries of Big Data, but businesses' processes could also be the beneficiaries.

[3] idem

Big Data can help governments, businesses and individuals to make better informed decisions. They assist in identifying trends, allowing businesses to gain competitive advantages and the public sector to better serve citizens. The use of Big Data leverages technologies' opportunities, and, as a result, a new class of Big Data has emerged.
Table 1 shows how rapidly the volume of these data is growing.

Table 1: Data 'inflation'

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data—including text, numbers, images, videos, etc.
Byte (B)	8 bits	Enough information to create a number or an English letter in computer code. It is the basic unit of computing.
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB.
Megabyte (MB)	1,000KB, or 2^{20}, bytes	From "large" in Greek. The MP3 file of a typical song is about 4MB.
Gigabytes (GB)	1,000MB, or 2^{30}, bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB. A 1GB text file contains over 1 billion characters, or roughly 290 copies of Shakespeare's complete works.
Terabyte (TB)	1,000GB, or 2^{40}, bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB. All the tweets sent before the end of 2013 would approximately fill an 18.5TB text file. Printing such a file (at a rate of 15 A4-sized pages per minute) would take over 1200 years.
Petabyte (PB)	1,000TB, or 2^{50}, bytes	The NSA is reportedly analyzing 1.6 per cent of global Internet traffic, or about 30PB, per day. Continuously playing 30PB of music would take over 60,000 years, which corresponds to the time that has elapsed since the first *Homo Sapiens* left Africa.
Exabyte (EB)	1,000PB, or 2^{60}, bytes	1EB of data corresponds to the storage capacity of 33,554,432 iPhone 5 devices with a 32GB memory. By 2018, the total volume of monthly mobile data traffic is forecast to be about half of an EB. If this volume of data were stored on 32GB iPhone 5 devices stacked one on top of the other, the pile would be over 283 times the height of the Empire State Building.
Zettabyte (ZB)	1,000EB, or 2^{70}, bytes	It is estimated that in 2013, humanity generated 4-5ZB of data, which exceeds the quantity of data in 46 trillion print issues of *The Economist*. If that many magazines were laid out sheet by sheet on the ground, they would cover the total land surface area of the Earth.
Yottabyte (YB)	1,000ZB, or 2^{80}, bytes	The contents of one human's genetic code can be stored in less than 1.5GB, meaning that 1YB of storage could contain the genome of over 800 trillion people, or roughly that of 100,000 times the entire world population.

The prefixes are set by the International Bureau of Weights and Measures.
Source: Adapted and updated from The Economist by Emmanuel Letouzé and Gabriel Pestre, using data from Cisco, the Daily Mail, Twitter (via quora.com), SEC Archives (via expandedramblings.com), Bitesizebio.com, and the book Uncharted: Big Data as a Lens on Human Culture (2013) by Erez Aiden and Jean-Baptiste Michel.

The proportion of digital data produced recently is growing ever faster—up to 90% of the world's data was created over just two years (2010–2012)[add source].

3.Importance of Open Data

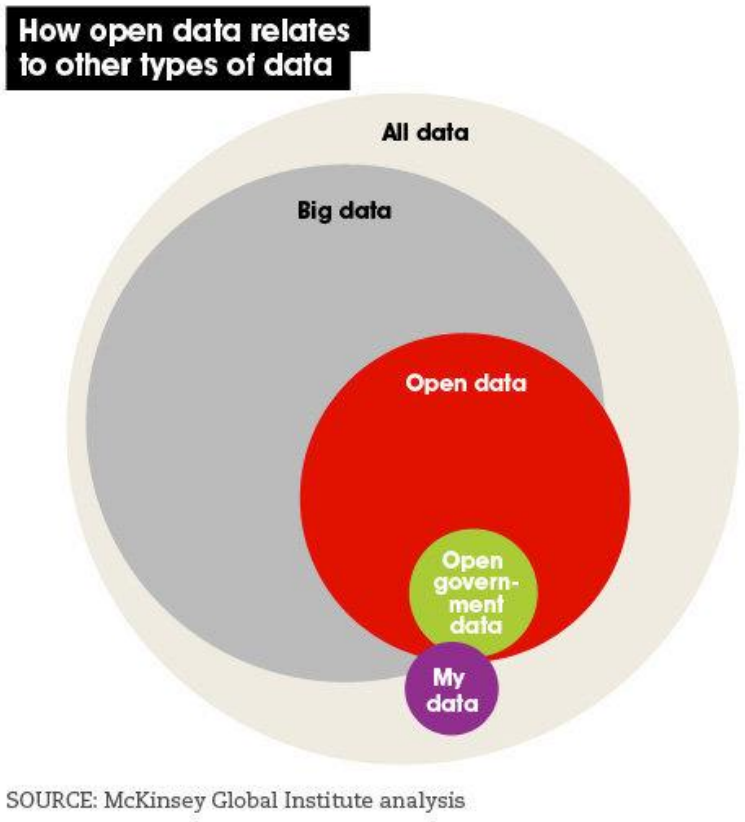
Definition

Open Data supports and enhances Big Data's availability and potential. The concept behind Open Data is that data should be freely available to everyone to use as they wish, without restrictions from copyrights, patents, or other mechanisms of control. It is already changing the way that governments address issues domestically and internationally. The European Commission defines Open Data as all the data and information produced, collected or purchased by government agencies—such as statistics and spatial data—as well as data resulting from projects financed by those government agencies. The concept highlights the initiatives by governments around the world to allow government Open Data for stakeholders in a way that enables researchers to obtain information that can be used safely. Such data should be provided in a machine-readable file format. Several national governments have created websites to distribute a portion of the data they

collect. It is a concept for a collaborative project in municipal Government to create and organize Culture for Open Data or Open Government Data. A list of over 200 local, regional, and national Open Data catalogues is available on the catalog[4]. The World Bank[5], the United Nations[6] and the European Union[7] are leading this effort, and give access to large amount of data.

Benefits

Figure 3: How Open Data relates to other types of data



SOURCE: McKinsey Global Institute analysis

For the foreseeable future, the Big Data and Open Data movements will be the two main pillars of a larger 'data revolution'. Both rise against a background of increased public demand for more openness, agility, transparency, and accountability for public data and actions. Open Data initiatives are promoting high-value data projects inside and outside of government—as opposed to Open Data for the sake of Open Data. Some cities in the US are reaching out to potential data users locally—businesses, academics, entrepreneurs and others who might engage—to

[4] http://datacatalogs.org
[5] http://data.worldbank.org
[6] http://data.un.org
[7] http://open-data.europa.eu/en/data/

understand what is most useful. A 'true' Big Data revolution should be one where data can be leveraged to change power structures and decision-making processes, not just create insights[8].

Data types

There are different types of Big Data. The first one is made of structured data containing numbers of facts. The most frequently used of this type of hard data is Call Detail Records (CDRs) collected by mobile phone operators. CDR data inform about the location of the phone tower (where the call was made), and the time and duration of the call. Large operators collect over six billion CDRs per day[9]. Table 2 describes the data contained within a CDR.

Table 2: Data contained in a CDR

Variable	Data
Caller ID	X76VG588RLPQ
Caller ID tower location	2°24' 22.14" , 35°49' 56.54
Recipient phone number	A81UTC93KK52A81UTC93KK52
Recipient cell tower location	3°26' 30.47", 31°12' 18:01"
Call time	3013-11-07T15:15:00
Call duration	01:12:02

Note: only the phone tower location is given for privacy reasons. Source: New primer on mobile phone network data for development. (UN Global Pulse, 5 November 2013)

The second type of Big Data is unstructured data linked to social media content, such as videos, music, online purchases etc. Because of their unstructured and subjective nature, they are more difficult to analyze.

The third type of Big Data is gathered by digital sensors on road, satellite imagery, videos at toll roads, or water meters.

Big Data can be descriptive, simply providing documentation, mainly a function quite similar to statistics. They can also give a sense of what may happen, such as early warning applications for bad weather—this is the predictive function. The predictive use of Big Data is based on algorithms that learn from and react to data, identifying and using patters. This would be the case of the optimization of a complex system in real-time, such as commuter train services. This use also makes it possible to reduce the decision time.

Even if Big Data are still in their infancy, some challenges have already been identified.

[8] Open data: unlocking innovation and performance with liquid information (McKinsey Global Institute, October 2013)
[9] *New primer on mobile phone network data for development. (UN Global Pulse, 5 November 2013)*

4.Challenges

The main challenges associated with Big Data are access to data, going beyond the numbers, getting the right skills, and privacy issues.

Access to data

Access to data is probably the most crucial challenge. In many countries, access to data is still severely regulated by government bodies and is highly restricted. However, the Open Data experiences in a number of leading countries in the area of open government are showing the impact of Open Data on the performance of these governments and the value they deliver to customers and citizens. Open Data also provides an economic boost and increased job creation. The EU's move toward Open Data directive is expected to create 58,000 jobs in the UK through 2017, and add £216 billion to the country's economy. Open Weather Data in the US has created 400 companies employing 4,000 people. A Spanish study found an increase of about €600m of business from Open Data—with the creation of over 5000 jobs. An Australian study found a return on investment of 500% from Open Data.[10] However, in many developing countries, access to data—despite clear benefits—is still lagging behind, even if some countries have the skilled workforce and the technology to use Big Data.

Consider more than just the numbers and Visualize

A key danger in Big Data is a 'selection bias' in which the samples do not represent a cross section of the population. For example, mining data on Twitter will be biased in favor of young people most likely living in urban areas, because they make up most of Twitter's users. So analyses based on Big Data may lack 'external validity'.

Another risk is that analyses based on Big Data will focus too much on correlation and prediction—at the expense of cause, diagnostics, or inference. Since 2010, some US cities have utilized 'Predictive policing', meaning that police and law enforcement entities mine and analyze data to assess the likelihood of increased crime in certain areas, predicting rises based on historical patterns[11]. Forces dispatch their resources accordingly, and in most cases this has reduced crime. However, this type of knowledge cannot replace preventive policy that tackles the root causes and contributing factors.

The visualization of the data is important as it will enhance the understanding of the findings. It provides a way to determine where to look and what questions to ask. Figure 4 presents a good example of visualization of complex data.

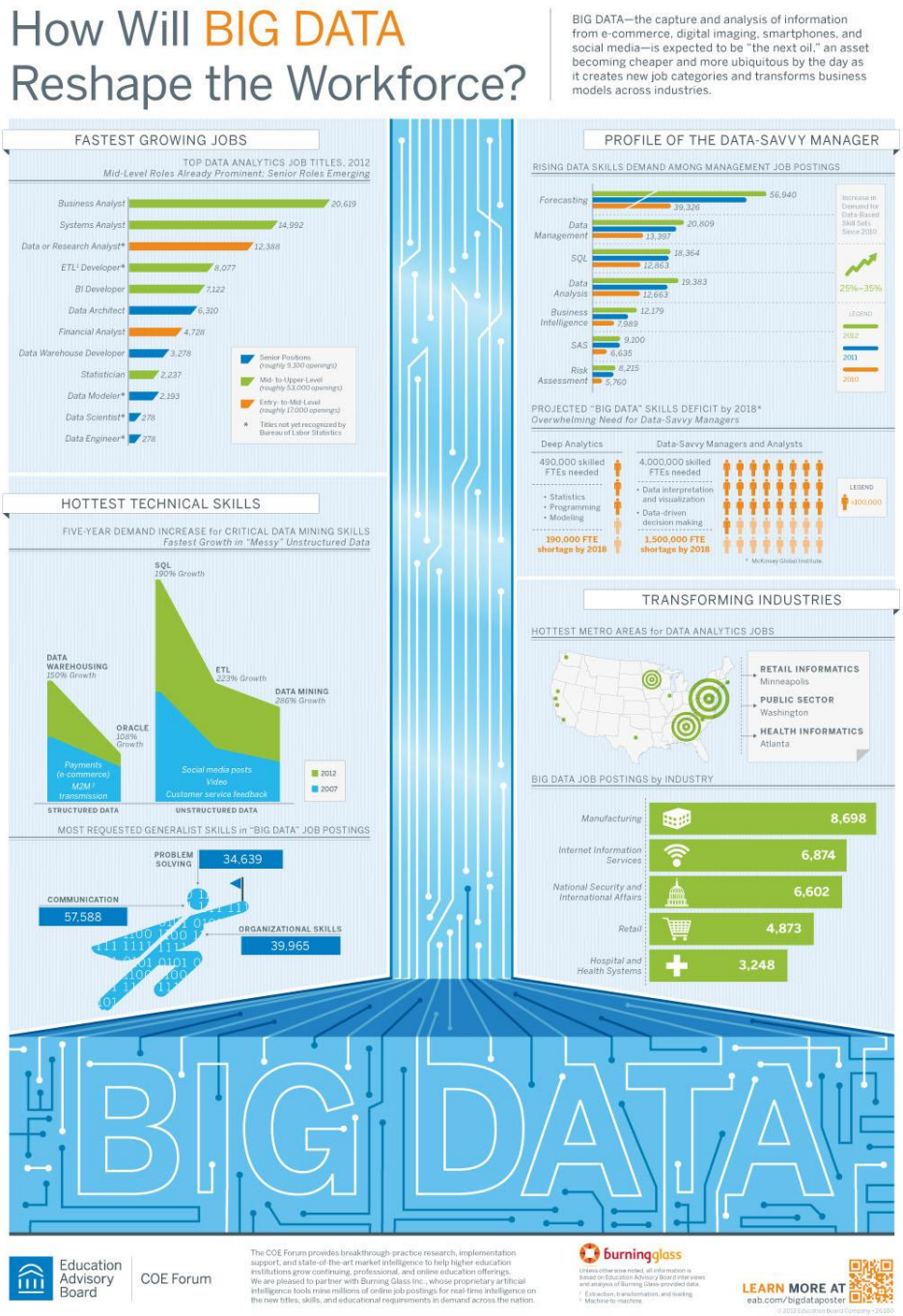
[10] World Bank open Data
[11] http://www.cityofchicago.org/city/en/depts/doit.html

Finding skilled workers, educating the workforce

Big Data are different from statistics; therefore, their treatment requires new skills. Several government bodies and private sector companies, understanding the benefits of Big Data, are now recruiting Chief Data Officers (CDO), and these positions are on the rise. For example, in San Francisco, the first CDO is charged with standardizing the city's data policies across departments and making the city's data more user-friendly and accessible. The CDO is also tasked with establishing and managing data coordinators within individual city departments to develop and monitor data efforts.

The treatment of Big Data requires Data cleaners; they are the ones that ensure consistently clean and accurate data. The data "explorers" go through data to discover what information is needed. The business solution architects compile and structure data for analysis. The data scientists create analytic models. And the campaign experts are in charge of analyzing and executing models for optimal results. Big Data are reshaping the workforce, as job related to Big Data are among the hottest and fastest growing jobs, creating new job categories and transforming business models across industries. Figure 4 shows the evolution and the rapid evolving demand for analytics job in the US market.

Figure 4: Big Data are changing the workforce

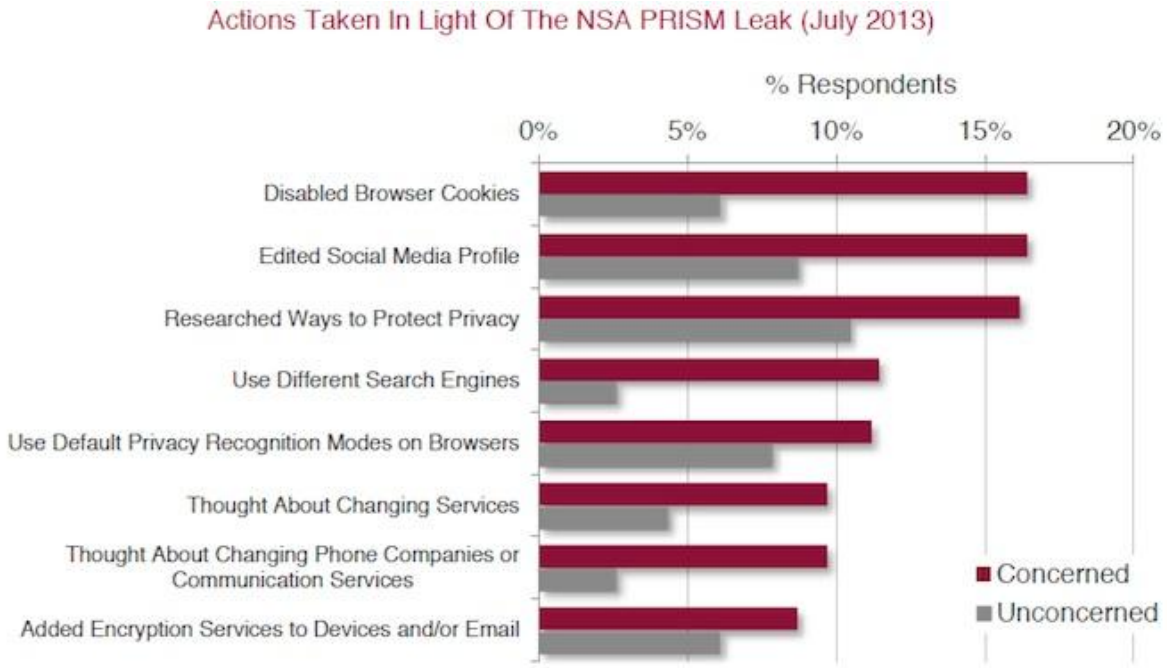


Source: Education Advisory board

Privacy and security

Concerns have been raised regarding how Big Data may erode privacy and promote inequality. Perhaps the most severe risks—and most urgent avenues for research and debate—are to individual rights, privacy, identity, and security. More and more algorithms are infiltrated within the life of the citizens, making decisions for them[12]. Anonymization has its limits. A study of movie rentals showed that even 'anonymized' data could be 'de-anonymized' — linked to a known individual by correlating rental dates of as few as three movies with the dates of posts on an online movie platform[13]. As a result of the US National Security Agency (NSA) intelligence leak, studies are showing that people are changing their privacy and tracking settings on the Internet[14] (see figure 5).

Figure 5: Users are changing their Internet habits



"Data is assumed to accurately reflect the social world, but there are significant gaps, with little or no signals coming from particular communities"[15]. Big Data may create a new "digital divide", because the communities that are supposed to be the beneficiaries are often not captured by the data—they do not use twitter, for example. Thus, disadvantaged communities may be left behind in the Big Data revolution.

[12] https://medium.com/futures-exchange/4bc7c8dcd5
[13] http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
[14] http://www.fastcoexist.com/3015860/people-are-changing-their-internet-habits-now-that-they-know-the-nsa-is-watching
[15] http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html

The Responsible Data Forum[16] is addressing the ethical, legal, privacy and security challenges surrounding the use and sharing of data in development. It is a collaborative effort to develop tools and strategies for managing responsible data for specific projects, to describe best practices and ethical controls for data sharing, and to provide guidelines for data-driven projects and policy interventions to create greater awareness. The group is currently working on draft policies for sharing data in humanitarians emergencies, a "do not harm"checklist for collecting data on vulnerable populations, as well as educational materials on data life-cycles within organizations.

5.Big Data in the Transport Sector

Most of the data generated and stored in all of human history have been produced in the last two years. These data have been collected by different operators from different users and for different purposes, including within the transport sector. Not only are these data available, but the ability to analyze these data has increased dramatically. With finite budgets and a growing urgency on limiting energy usage and pollution, transportation is rapidly becoming the critical issue for cities around the world. Transportation is also a primary example of the potential for Big Data, sensors, and social media to help cities meet modern-day challenges. From optimizing multimodal transport to managing traffic flow, the infrastructure conversation is rapidly moving from roads to data. Data enables a smarter, mobile lifestyle; however, this shift creates entirely new challenges as entrepreneurs and policymakers alike are forced to work within legacy policy frameworks for everything from funding allocation to privacy and open government.

Big Data and analytics can help to more accurately model and optimize demand, capacity, schedules, pricing, and better customer feedback. Leveraging on Big Data can improve customer experience in mass transit systems, improve operational efficiency, and maximize the availability of assets and infrastructure.[17]

The transport sector has been a long time user and beneficiary of new technologies and mobile devices that increase the ability to track fleets and monitor traffic to save time and reduce congestion, as well as the provision of new location-based services. Some examples:

- **Navigation data points.** In 2012, TomTom, a leading provider of navigation hardware and software, had more than 5000 trillion data points in its databases, coming from its navigation devices and other sources. These data points describe time, location, direction and speed of travel of individual anonymised users. Overall, estimates suggest that the global pool of personal geo-location data represented at least one petabyte in 2009, with

[16] https://responsibledata.io/forums/responsible-development-data
[17] Big Data and analytics in travel and transportation, IBM White paper, 2013

growth of about 20% per subsequent year. By 2020, this data pool could provide USD 500 billion in value worldwide in the form of time and fuel savings, or 380 megatonnes (million tons) of CO_2 emissions saved.[18].

- **Real-time traffic information.** Google Traffic Alerts provide information to consumers on their daily commute using a mix of data sources—public data (such as construction schedules), private data (such as telecom companies tracking individual user devices to calculate time to work), and some passively-generated data (for example, a cluster of calls made from a similar location might indicate a traffic jam).
- **Instruments.** These include smart phones, sensors, and onboard vehicle hardware that enable continuous collection, communication, and processing of mobility data—anything from traffic and weather conditions, to parking spots and rideshares.

Open Data are now seen as one of the biggest enabler for intelligent transportation. Most cites in the US manage their traffic and transit data on private databases—accessed exclusively by municipal staff to monitor system performance and implement improvements. By sharing the data, cities could tap into a larger pool of creativity, and many have in fact done it. For example, 4% of the 400,000 monthly trips on Bay Area Rapid Transit (BART) are planned using Embark[19], a multi-modal trip planning IPhone app started by three college students that has since spread nationwide. ParkMe[20], another smartphone app that specializes in predictive algorithms to direct drivers to the best available parking locations, has built an entire business case around providing parking data. While this app is free to customers, developers are willing and eager to pay for ParkMe's data, because they can transform it into their own revenue stream from drivers in search of parking.

Some Experiences

Big Data Fast Tracks China's Urban Transportation Management[21]

In 2012, the Chinese Ministry of Transport launched a national initiative to develop a management platform designed to improve the timeliness and security of public transportation, and to deliver better taxi service for citizens in Beijing, Kunming, Chongqing, and Tianjin. The piloted platform is to address the overall demands of urban transportation management organizations, analyzing 6-36 billion data records per year, and empowering a substantial improvement in performance under various application scenarios—such as complex analysis and inquiry. With the addition of this real-time data, transportation management organizations can efficiently analyze traffic situations—including passenger flow and taxi operations. For example, the application for lost-and-found intends to reduce the inquiry process from days to minutes, and has enabled an increase of up to nearly 1,000 times in performance, which helps improve the service responsiveness of the urban management department to passengers. The piloted platform has enabled urban transportation

[18] Exploring Data-driven innovations as a New Source of Growth, OECD, June 2013
[19] http://letsembark.com/
[20] http://www.parkme.com/
[21] http://en.prodigynetwork.com/big-data/how-big-data-is-transforming-public-transportation-management/

management organizations to improve management from the macro, meso and micro levels. Benefits have included:

- Urban transportation planning: trip characteristics, traffic zone division, and key area path analysis
- Comprehensive traffic management: real-time insights into current urban situations and prediction of traffic conditions; real-time timing analysis of traffic lights, monitoring and prediction of traffic congestion; and overall traffic index, reachable time analysis and prediction for areas and paths
- Dedicated management application groups: operational analysis for buses, taxis and rail transit, comprehensive analysis on capacity and pricing subsidy of public transportation, and labor intensity analysis.

Case Study: City of Da Nang, Vietnam, Traffic Management System

Da Nang is the biggest seaport and fourth largest city in Vietnam with close to 1 million inhabitants. It is also Vietnam's fastest growing metropolitan sprawl. Since 2013, as part of IBM's Smart Cities Challenge, Da Nang's traffic control center has had the tools it needs to predict and prevent congestion on the city's roads, and to better coordinate responses to situations caused by adverse weather or road accidents. Data are aggregated from multiple streams, which city planners can then analyze to detect anomalies and control Da Nang's flow of traffic. The system also has given the Department of Transport access to real-time information for its 100-strong fleet of buses—allowing it to view details, such as the location of each bus, and their current speeds and predicted journey times. Software and sensors are embedded in roads, highways, and buses, and synchronize stop lights to minimize traffic jams. This data can then be shared with passengers—either through video screens at bus stations or via mobile apps—thereby encouraging more people to use the buses and reduce the number of cars on the road. The €37 million project is already reaping benefits by bringing:

- Real-time information on city buses, such as driving speed, location and predicted journey times
- Successful implementation of sensors that monitor traffic on roads, as well as water levels on the flood-prone Han River, to help regulate Da Nang's Port.

Data Challenge on transportation

Cities can benefit from investing, at a fairly low cost, in standardizing and disseminating their transportation data. Software developers can help them to make the most of these data to improve existing roads and parking lots, and to expand transportation options through shared cars or better taxi management, thereby enhancing mobility within the cities.

The following example describes how a city makes available transport data sets and asks developers to create solutions:

MIT Challenge in Boston

The first Massachusetts Institute of technology (MIT) Big Data Challenge was launched in 2013, in partnership with the City of Boston. The focus was on transportation in downtown Boston22. Multiple data sets were made available—including transportation data from more than 2.3 million taxi rides, local events, social media and weather records—with the goal of predicting demand for taxis in downtown Boston, and creating visualizations that provide new ways to understand public transportation patterns in the city.

The interest for the City of Boston is to get new insights into how people use all modes of transportation travel in and around the downtown Boston area. The overall objective is to manage all modes of transportation more efficiently, and to use real-time data to facilitate better trip-planning between the various modes of transportation. With urban congestion on the rise, city planners are looking for ways to improve transportation—such as providing people with more options to get from one place to another (e.g., walking, biking, driving, or using public transit), and reducing the volume and more efficiently routing vehicles in the city.

This MIT Big Data Challenge focused primarily on one mode of public transportation: Taxi Cabs. A better understanding of patterns in taxi ridership can provide new insights for city planners, such as:

- How to get more cabs where they are needed, when they are needed?
- What are the ideal locations for cab stands?
- When and where should the city add or remove cab stands?
- How many cabs should be waiting around a specific location at a specific time of day?
- Are there viable alternatives to taking a cab?
- Are there easy ways to 'link trips' between cabs and other forms of transportation?
- How do taxi ridership patterns differ on weekdays vs. weekends? Seasonally? During different types of events?
- Where should you go at 1am to catch a cab downtown?
- Whether Bruin fans or Celtic fans take more cabs? Whether—and if so, how—the results of games impact transportation patterns?

Open Data Project in San Francisco

The city of San Francisco is using the same model as the one in Boston described above. The experience with the Open Data Project in San Francisco has indicated a three-step process for municipal Open Data programs. It begins when a city experiments successfully with Open Data and publishes a few data sets. This success leads to a second effort that engages the public through hackathons and other events. The process matures when Open Data starts to be used to answer tangible real-world problems in communities. The main lesson to cities and other localities

[22] https://calendar.csail.mit.edu/events/124797

interested in Open Data is to simply experiment—not to be overwhelmed—and gain gradual momentum[23].

6.Conclusion: Big future

Because the growth of Big Data will continue, we can expect more papers and controversies about Big Data's potential and perils for development. The future of Big Data will likely be shaped by three main strands: academic research, legal and technical frameworks for ethical use of data, and larger societal demands for greater accountability.

Research will continue to examine whether and how methodological and scientific frontiers can be pushed, especially in two areas: drawing stronger inferences, and measuring and correcting sample biases.

Policy debate will develop frameworks and standards—normative, legal and technical—for collecting, storing and sharing Big Data. These developments fall under the umbrella term 'ethics of Big Data'. Technical advances will help—for example, by injecting 'noise' in datasets to make re-identification of the individuals represented in them more difficult. But a comprehensive approach to the ethics of Big Data would ideally encompass other humanistic considerations—such as privacy and equality—and champion data literacy.

A third influence on the future of Big Data will be how it engages and evolves alongside the 'Open' Data movement and its underlying social drivers—where 'Open Data' refers to data that is easily accessible, machine-readable, available for free or at a negligible cost, and with minimal limitations on its use, transformation, and distribution.

[23] http://opencityapps.org/

7.Some definitions

These definitions are part of an article published by Emmanuel Letouzé, "Big Data for Development: Key Resources »[24]

Algorithms (and Algorithmic Future)	In mathematics and computer science, an algorithm is a series of predefined instructions or rules written in a programming language designed to tell a computer how to sequentially solve a recurrent problem through calculations and data processing. The use of algorithms for decision-making has grown in several sectors and services, such as policing and banking. This has led to hopes—and worries—about the advent of an 'algorithmic future' where algorithms may replace human functions, or even become an instrument for repression.
Big Data	An umbrella term that, simply put, stands for one or more of three trends: the growing volume of digital data generated daily as a by-product of people's use of digital devices; the new technologies, tools and methods available to analyze large data sets that are not designed for analysis; and the intention to extract policymaking insights from these data and tools.
Call Detail Records (CDRs)	The technical name for mobile phone data recorded by all telecom operators. CDRs contain information about the locations of those sending and receiving calls or text messages through operators' networks, as well as data on time and duration.
Data Revolution	A common term in development discourse since the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda called for a 'data revolution' to "strengthen data and statistics for accountability and decision-making purposes". It refers to a larger phenomenon than Big Data or the 'social data revolution'—defined as the shift in human communication patterns towards greater personal information sharing, and the implications of this shift.
Data Scientist or Data Science	A professional or a field that focuses on solving real-world problems using large amounts of data by combining skills from often distinct areas of expertise: maths, computer science (for example, hacking and coding), statistics, social science, and even storytelling or art.
(New) Digital Divide	The differential access and ability to use information and communications technologies between individuals, communities and countries—and the resulting socioeconomic and political inequalities. The skills and tools required to absorb and analyze the growing amounts of data produced by such technologies may lead to a 'new digital divide'.
False Positives Versus False Negatives (or	A false positive or type I error refers to a prediction or conclusion that turns out to be false—for example, a fire alarm going off when there is no

[24] http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html

Type I Versus Type II Errors)	fire, or an experiment indicating a medical treatment has worked when it has not. A false negative or type II error refers to cases in which a study or a monitoring system fails to identify an event or effect that has occurred. Attempts to predict rare events, such as political revolutions, using increasingly rich data and powerful tools are expected to lead to more false positive than false negative results (also known as over-prediction).
Internal Versus External Validity	Internal validity refers to the extent to which a causal relationship can be confidently established between two phenomena—a reduction in the speed limit and a reduction in the number of road deaths, for example. This requires taking into account all other factors that may affect the outcome, and offering alternative explanations. In the case of a reduction in the number of road deaths, an alternate explanation might be a change in people's drinking habits. External validity refers to the extent to which a study's conclusions can be confidently generalized to other situations and people. In other words, whether the conclusions would hold beyond the area and time for which they were established.
Statistical Machine Learning	A subset of data science, falling at the intersection of traditional statistics and machine learning. Machine learning refers to the construction and study of computer algorithms—step-by-step procedures used for calculations and classification—that can 'learn' when exposed to new data. This enables better predictions and decisions to be made based on what was experienced in the past, as with filtering spam emails, for example. The addition of "statistical" reflects the emphasis on statistical analysis and methodology, which is the main approach to modern machine learning.

8. References

OECD Report (June 2013)
http://search.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP(2012)9/FINAL&docLanguage=En

UN Global Pulse
http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

European Union Open Data Portal http://open-data.europa.eu/

World Bank Report, City of Stockholm's Congestion Charging project:
http://siteresources.worldbank.org/INTTRANSPORT/Resources/StockholmcongestionCBAEliassonn.pdf

http://www.economist.com/news/briefing/21585002- enthusiasts-think-data-services-can-change-cities-century-much-electricity

IBM White Paper
http://public.dhe.ibm.com/common/ssi/ecm/en/gbw03215usen/GBW03215USEN.PDF

Harvard Business Review Blog Network
http://blogs.hbr.org/cs/2013/06/what_the_companies_winning_at.html

http://eu2013.ie/ireland-and-the-presidency/the-eu-and-policy-areas/transport,-telecommunications-and-energy/

How automotive companies use Big Data:
http://www.livemint.com/Specials/P6e4ijI7XVxKKhyEEzzqMO/Auto-makers-bet-on-big-data-for-business-insights.html?ref=mr

http://www.fastcoexist.com/3017102/a-new-underclass-the-people-who-big-data-leaves- behind

Kirkpatrick, Robert, Big Data in Real Time Toward a New Evidence Base for Impact, UN Global Pulse, 2013

http://data.worldbank.org/open-government-data-toolkit

http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html

http://opencityapps.org/

http://letsembark.com/

http://www.parkme.com/

http://en.prodigynetwork.com/big-data/how-big-data-is-transforming-public-transportation-management/