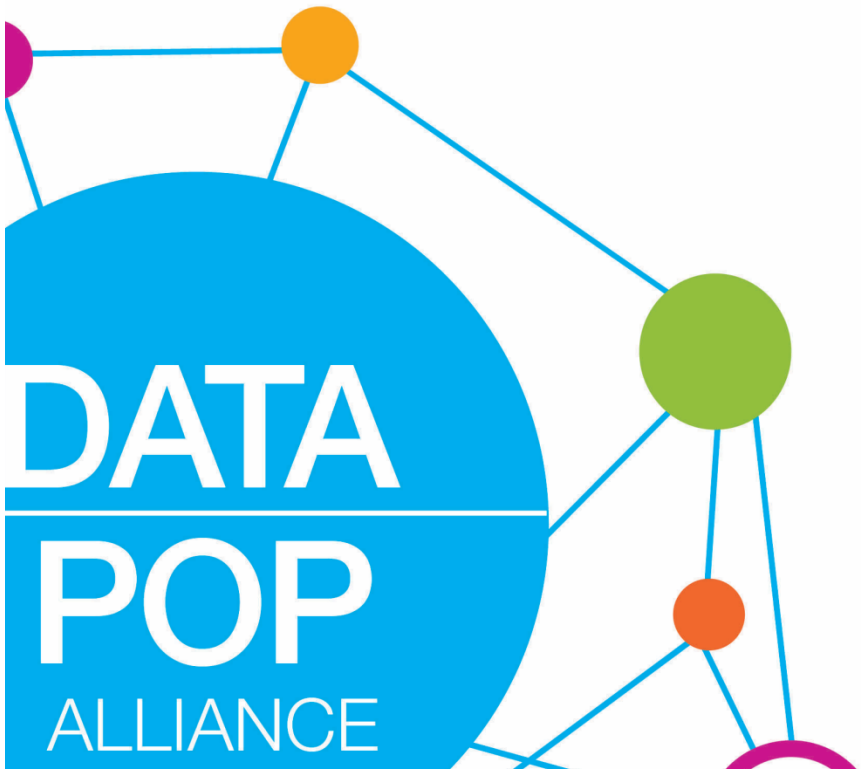
DATA-POP ALLIANCE
PRIMERS SERIES

BIG DATA & DEVELOPMENT
AN OVERVIEW

March 2015

Written by

Emmanuel Letouzé



WORLD BANK GROUP
Leadership, Learning and Innovation

SciDevNet

mit media lab

HARVARD HUMANITARIAN INITIATIVE

ODI

About Data-Pop Alliance

Data-Pop Alliance is a research, policy and capacity building think-tank on Big Data and development, jointly created by the Harvard Humanitarian Initiative (HHI), the MIT Media Lab, and the Overseas Development Institute (ODI) to promote a people-centred Big Data revolution.

About the Author

This Primer was written by Emmanuel Letouzé.

Emmanuel Letouzé is the Director and co-founder of Data-Pop Alliance. He is a Fellow at the Harvard Humanitarian Initiative, a Visiting Scholar at MIT Media Lab, a Research Associate at ODI, a PhD Candidate (ABD) at UC Berkeley, and the author of UN Global Pulse's report "Big Data for Development: Challenges and Opportunities" (2012).

Acknowledgement

This is the inaugural Primer in Data-Pop Alliance's 'Primer Series' developed in collaboration with the World Bank Leadership, Learning and Innovation group and other partners.

This Primer's content draws on articles written by the author published on SciDev.Net in April 2014 as part of a Big Data for Development Spotlight: http://www.scidev.net/global/data/spotlight/big-data-for-development.html, used with SciDev.Net's authorization.

This document benefited from ideas and feedback from SciDev.Net staff (Anita Makri, Kaz Janowski) and World Bank staff (Adarsh Desai, Tariq Khokhar, Amparo Ballivian, Trevor Monroe, and Isabelle Huynh) and numerous interactions and discussions with the Data-Pop Alliance leadership and team, affiliates and partners. All errors and omissions remain those of the author.

Letouzé E, "Big Data and Development: General Overview Primer." Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative. 03/2015.

DATA-POP ALLIANCE
PRIMERS SERIES

BIG DATA & DEVELOPMENT
AN OVERVIEW

March 2015

Written by

Emmanuel Letouzé

Data-Pop Alliance
World Bank Group
SciDev.Net

Table of Contents

1 Facts, Figures and the Big Picture

Despite the buzz, 'Big Data for development'—i.e. the field of research and practice about the applications and implications of Big Data for policymaking and development—remains in its intellectual and operational infancy. Is this "new oil" poised to be a blessing or a curse for human development and social progress?

Optimists have called it a revolution that will change "how we live, think and work", even expressed the hope that "Africa's statistical tragedy" may be partly fixed by Big Data. But skeptics and critics have been more circumspect, or plainly antagonistic, referring to Big Data as a big ruse, a big hype, a big risk as well as, of course, 'Big Brother'. The Big Data buzz could just be a bubble, some observers point out - automated analysis of large datasets is not new. So what is new here?

The early years of a recent phenomenon

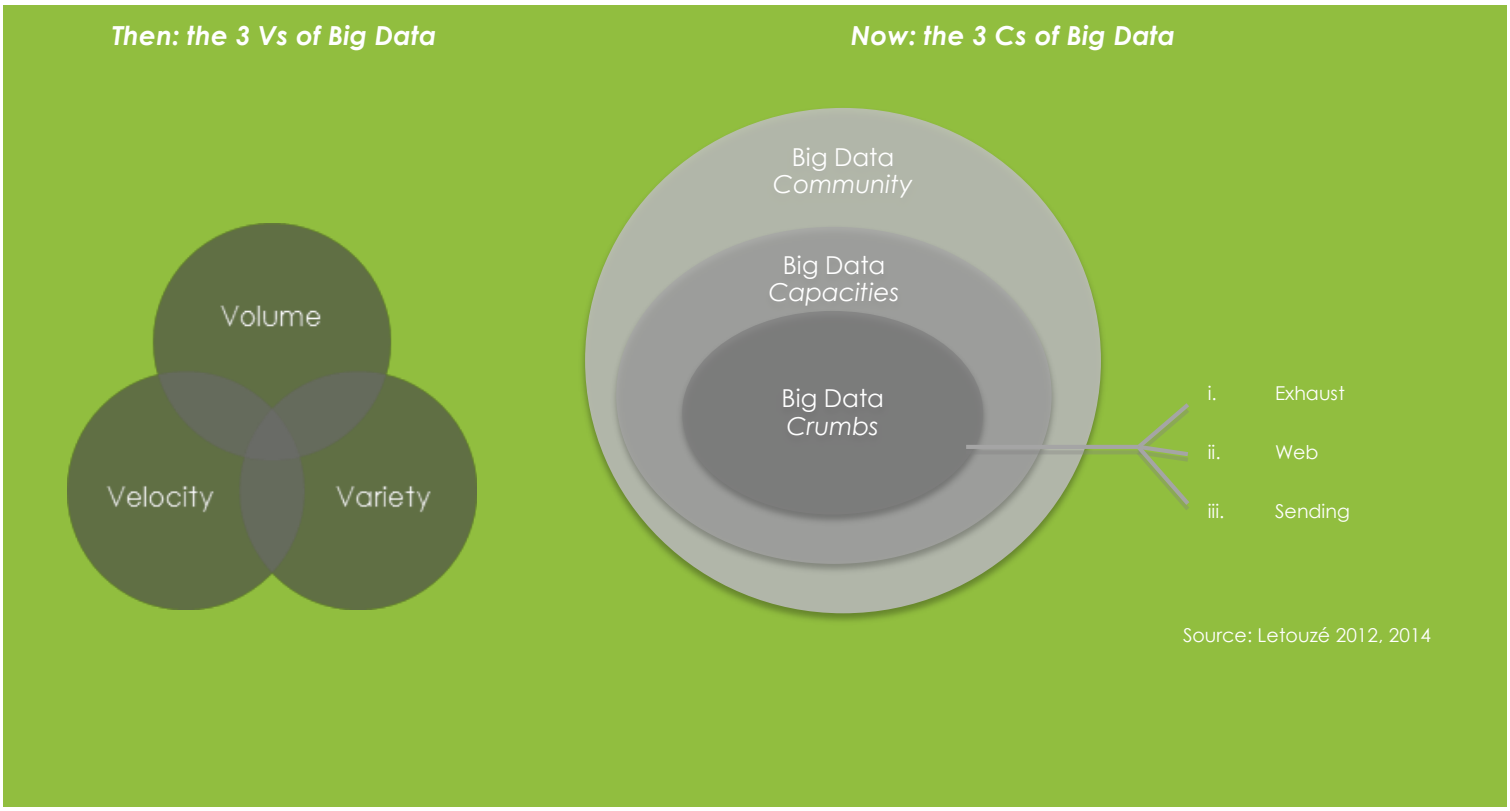
Joe Hellerstein, computer scientist at the University of California, Berkeley, United States, made an early mention of an upcoming "Industrial Revolution of data" in November 2008, while The Economist talked about a "data deluge" in early 2010. 'Big Data' itself became a mainstream term only a couple of years ago. Google searches are one metric that reflect this: the number of searches that include the term did not take off until 2011–12. In those two years, four major reports were published: by the UN Global Pulse, the World Economic Forum, the McKinsey Global Institute and danah boyd and Kate Crawford, researchers at Microsoft and academic institutions.

1.1 From the 3 Vs to the 3 Cs of Big Data

Although there is no single agreed upon definition of Big Data, it must be approached as a new ecosystem that is part and parcel of a larger social phenomenon driven by digital technology. This Big Data ecosystem can be characterized as the union of 3 Cs: Big Data Crumbs, Capacities and Communities. This characterization is more accurate and less confusing than the 3 Vs of Big Data (for Volume, Velocity and Variety) used in the early years of Big Data, circa 2010-12. The major limitation of the 3 Vs was their exclusive focus on Big Data as being just 'big data'. Another was their emphasis on Big Data being essentially a quantitative, rather than qualitative, shift. It cannot be said loudly and clearly enough: Big Data cannot be constrained to big data, and the 3Vs ought to belong to history.

What are the 3 Cs? The first C, 'Crumbs', refers to "digital breadcrumbs", as put by MIT Professor and Data-Pop Alliance Academic Director Alex 'Sandy' Pentland. Contrary to traditional survey data, these data are not produced for the purpose of statistical inference; instead they are for the most part, passively left behind by humans using digital devices and services, many of which were unavailable 5 or 10 years ago.

Each of these actions leaves a digital trace, which, added up, makes up the bulk of Big Data as data. Each year since 2012, over 1.2 zettabytes of data have been produced —1021 bytes, enough to fill 80 billion 16GB iPhones that would circle the earth more than 100 times. The volume of these data is growing fast, and just as a population with a sudden outburst of fertility gets both larger and younger, the proportion of digital data produced recently is growing ever faster —it is commonly reported that up to 90 per cent of the world's data was created in the last year alone, although the exact source of the claim and estimation methodology are unclear.



These data come in 3 main types. One is small, 'hard', structured data that can be easily quantified and organized (in columns and rows for instance) for systematic analysis, and that cannot be edited by their emitters. Examples include Call Detail Records (CDRs) and credit card transactions, as well as EZ pass or subway records. Some argue that this kind of data constitute the real novelty and promise of Big Data; as elaborated by Alex 'Sandy' Pentland, "*the power of*

Big Data is that it is information about people's behavior instead of information about their beliefs".

A second kind of data includes videos, online documents, blog posts and other social media content. These are 'unstructured' data—harder to analyze in an automated fashion. They are also more subject to their authors' editorial choices: someone may blog about boycotting a certain product, but his/her credit card statement may tell a different story. A third type is gathered by digital sensors that pick up human actions, such as electric meters or satellite imagery that can pick up deforestation.

Some consider the universe of Big Data as data to be much wider, to include administrative records, price or weather data, for instance, or books that have been previously digitized—which, taken collectively, may constitute a fourth kind.

Importantly, as mentioned, what these data have in common is that they were not collected or sampled with the explicit intention of drawing conclusions from them. So the term Big Data is fundamentally misleading: size isn't a defining feature, it is only a corollary of their nature. Even a small 'Big Data' dataset can be Big Data if it doesn't stem from fully controlled processes like surveys and statistical imputations undertaken by official bodies.

Call Detail Records (CDRs)

Call Detail records (CDRs) are metadata (data about data) that capture subscribers' use of their cell-phones — including an identification code and, at a minimum, the location of the phone tower that routed the call for both caller and receiver — and the time and duration of call. Large operators collect over six billion CDRs per day. [15]

CALLER ID	CALLER CELL TOWER LOCATION	RECIPIENT PHONE NUMBER	RECIPIENT CELL TOWER LOCATION	CALL TIME	CALL DURATION
X76VG588RLPQ	2°24' 22.14", 35°49' 56.54"	A81UTC93KK52	3°26' 30.47", 31°12' 18.01"	2013-11-07T15:15:00	01:12:02

http://www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev

If these data constitute the core of Big Data as an ecosystem, they do not constitute its whole. The second C of Big Data stands for Capacities—tools, methods, software and hardware: as put by Harvard University professor Gary King, *"Big Data is not about the data"*. These capacities include powerful computers, parallel computing systems, as well as statistical machine-learning techniques and algorithms that are able to look for and unveil patterns and trends in vast amounts of complex data.

The third C is for Community. Big Data is also comprised of the 'movement' of individual and institutional actors that operate largely outside of traditional policy and research spheres;

multidisciplinary teams of social and computer scientists with a "*mindset to turn mess into meaning*", as data scientist Andreas Weigend puts it. It also encompasses regular people using Google maps to decide whether they will take their car or the subway to go to a meeting. In the age of Big Data more than ever, everyone is a decision maker.

This characterization suggests that Big Data is a complex system, with feedback loops. New methods will yield new data; new data will give someone the idea of creating a data science start-up, etc. It also shows that the phrase "using Big Data" is fundamentally missing the point, unless it is explicitly meant to say that one considers using the Big Data ecosystem to achieve certain goals. Rather, the question and challenge is why and how to *engage with* Big Data—to try and become part of it, affect its evolution and/or benefit from its innovations.

Data 'inflation'

Unit	Size	Significance
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), BistesizeBio.com, and "Uncharted: Big Data as a Lens on Human Culture" (2013) by Erez Aiden and Jean-Baptiste Michel.

1.2 The promise: supply and demand factors

The excitement over Big Data has stemmed from two sets of factors: supply of ever-more data and analytics capacities, and demand for better, faster and cheaper information — in other words there has been and remains both a push for and a pull towards Big Data.

Availability of reliable and up-to-date data has been improving significantly over time; however in many instances gaps remain. For instance, a good indicator of a region's poverty or underdevelopment is a lack of poverty or development data. Some countries (most of them with a recent history of conflict) haven't had a census in four decades or more. Their population size, structure and distribution can often only be approximated with triangulation of information from different sources. Even though official figures exist, they are often based on incomplete data. Poor data also mean that some countries' official GDP figures get an overnight boost — 40 per cent for Ghana in 2010 or 60 per cent for Nigeria in 2014 — when changes in the structure of their economies, such as the rise of the technology sector, are finally taken into account.

This lack of reliable data has presided over the call for a 'Data Revolution' that led to the publication of a report by an UN-appointed experts group (although the term 'data revolution' predated this call by many years). The basic and somewhat simplistic rationale is that, in the age of Big Data, economies should be steered by policymakers relying on better navigation instruments and indicators that let them design and implement more agile and better targeted policies and programmes. Big Data has even been said to hold the potential for national statistical systems in data-poor areas to 'leapfrog' ahead, much as many poor countries skipped the landline phase to jump straight into the mobile phone era.

The appeal of potentially leaping ahead is also shaped by the 'supply side' of Big Data. There is early practical evidence and a growing body of work on Big Data's novel potential to help understand and affect human populations and processes.

For example, as has been reported in many previous publications, Big Data has been used to track inflation online, estimate and predict changes in GDP in near real-time, monitor traffic or the outbreak of epidemics. Monitoring social media data to analyse people's sentiments is opening new ways to measure welfare, while email and Twitter data could be used to study internal and international migration. And an especially rich academic literature uses CDRs to study migration patterns, socioeconomic levels, and disease spread, among others. As smartphones will soon overtake regular cellphones around the globe, CDR analysis will recede and new 'crumbs' will become the next frontiers.

Meanwhile, taxonomies have been proposed to clarify how Big Data could benefit development. One taxonomy distinguishes the 'early warning' uses from 'real-time awareness', or from 'real-time monitoring' of the impact of a policy or programme. Another contrasts its descriptive function (such as a real-time map) from predictive and prescriptive applications.

The predictive use can be understood in two senses of the term; either as inference or nowcasting—predicting what is happening right now (such as when cell-phone activity is used to predict socioeconomic levels), or forecasting (in a fashion very similar to what meteorologists do). The prescriptive use requires making causal inferences; i.e. establishing the existence, direction and magnitude of a causal link between some treatment or variable X and some effect or variable Y. Because Big Data as data do not result from controlled processes, making causal inferences with Big Data through randomized control trials for example will be very difficult to near impossible—but other techniques are being developed as discussed further below.

The following table provides examples of applications falling under each use category.

Taxonomies of actual and potential uses of Big Data for development

	Applications	Explanation	Examples	Comments
UN Global Pulse report Taxonomy[1] (Letouzé, 2012)	1. *Early warning*	Early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis	Predictive policing is based upon the notion that analysis of historical data can reveal certain combinations of factors associated with greater likelihood of crime in an area; it can be used to allocate police resources. Google Flu trends is another example, where searches for particular terms ("runny nose", "itchy eyes") are analyzed to detect the onset of the flu season — although its accuracy is debated.	This application assumes that certain regularities in human behaviours can be observed and modelled. Key challenges for policy include the tendency of most malfunction-detection systems and forecasting models to over-predict — i.e. to have a higher prevalence of 'false positives'.
	2. *Real-time awareness*	Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies	Using data released by Orange, researchers found a high degree of association between mobile phone networks and language distribution in Ivory Coast — suggesting that such data may provide information about language communities in countries where it is unavailable.	The appeal for this application is the notion that Big Data may be a substitute for bad or scarce data; but models that show high correlations between 'Big Data-based' and 'traditional' indicators often require the availability of the latter to be trained and built. 'Real-time' here means using high frequency digital data to get a picture of reality at any given time.

Alternative taxonomy (Letouzé et al., 2013)	3. *Real-time feedback*	The ability to monitor a population in real time makes it possible to understand where policies and programmes are failing, and make the necessary adjustments	Private corporations already use Big Data analytics for development, which includes analysing the impact of a policy action — e.g. the introduction of new traffic regulations — in real-time.	Although appealing, few (if any) actual examples of this application exist; a challenge is making sure that any observed change can be attributed to the intervention or 'treatment'. However high-frequency data can also contain 'natural experiments' — such as a sudden drop in online prices of a given good — that can be leveraged to infer causality.
	1. *Descriptive*	*Big Data can document and convey what is happening*	This application is quite similar to the 'real-time awareness' application — although it is less ambitious in its objectives. Any infographic, including maps, that renders vast amounts of data legible to the reader is an example of a descriptive application.	Describing data always implies making choices and assumptions — about what and how data are displayed — that need to be made explicit and understood; it is well known that even bar graphs and maps can be misleading.
	2. *Predictive*	*Big Data could give a sense of what is likely to happen, regardless of why*	One kind of 'prediction' refers to what may happen *next* —as in the case of predictive policing. Another kind refers to proxying prevailing conditions through Big Data—as in the cases of socioeconomic levels using CDRs in Latin America and Ivory Coast.	Similar comments as those made for the 'early-warning' and 'real-time awareness' applications apply.
	3. *Prescriptive*	*Big Data might shed light on why things may happen and what could be done about it*	So far there have been few examples of this application in development contexts.	Most comments about 'real-time feedback' apply. An example would require being able to assign causality. The prescriptive application works best in theory when supported by feedback systems and loops on the effect of policy actions.

1.3 The grey side of Big Data: risks and challenges

The promise of Big Data's applications to real-world problems has been met with warnings about its perils, and more broadly active discussions about its societal implications. Perhaps the most severe risks — and most urgent avenues for research and debate — are to individual and group rights, privacy, identity, and security. In addition to the obvious intrusion of surveillance activities and issues around their legality and legitimacy, there are important questions about 'data anonymisation': what it means and its limits. An early study of movie rentals showed that even 'anonymised' data could be 'de-anonymised' —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. Other research has found that CDRs that record location and time, even when free of any individual identifier could be re-individualised –- which is referred as re-identification. In that case, four

data points were theoretically sufficient to uniquely single out individuals in an entire dataset with 95 per cent accuracy. Recent research using credit card transactions yield very similar conclusions: our behaviors are unique and predictable enough to make it very hard for any given individual to hide in the digital crowd.

Critics also point to the basic risks associated with basing decisions on analyses lacking either external or internal validity.

Big Data – risks to drawing valid conclusions

A key challenge in Big Data is that the people generating it have selected themselves as data generators through their activity. In technical terms this is a 'selection bias' and it means that analysis of Big Data is likely to yield a different result from a traditional survey (or poll), which would seek out a representative cross section of the population. For example, trying to answer the question "do people in country A prefer rice or chips?" by mining data on Twitter would be biased in favour of young people's preferences as they make up more of Twitter's users. So analyses based on Big Data may lack 'external validity', although it is possible that individuals that differ in almost all respects may have similar preferences and display identical behaviours (young people may have the same preferences as older people).

Another risk comes from analyses that are flawed because they lack 'internal validity'. For instance, a sharp drop in the volume of CDRs from an area might be interpreted, based on past events, as heralding a looming conflict. But it could actually be caused by something different, such as a mobile phone tower having gone down in the area.

Another risk is that analyses based on Big Data will focus too much on correlation and prediction — at the expense of cause, diagnostics or causal inference, without which policy is essentially blind. A good example is 'predictive policing'. Police and law enforcement forces in some US and UK cities have crunched data to assess the likelihood of increased crime in certain areas for years, predicting rises based on historical patterns. Forces dispatch their resources accordingly, and this has reduced crime in most cases. However, unless there is knowledge of why crime is rising, it is difficult to put in place preventive policy that tackles the root causes or contributing factors. At the same time, proponents argue that cracking down on crime in an area may have a cumulative structural effect.

Yet another big risk that is receiving growing attention is Big Data's potential to create a 'new digital divide' that may widen rather than close existing gaps in income and power worldwide. One of the 'three paradoxes' of Big Data is that because it requires analytical capacities and access to data that only a fraction of institutions, corporations and individuals have, it may disempower the very communities and countries it promises to serve. People with the most data

and capacities are in the best position to develop Big Data systems to their economic and political advantages, even as they claim to use them to benefit others.

A last basic challenge is that of putting the data to use—fundamentally to understand how data has affected societies historically. Most discussions about the 'data revolution' assume that 'data matter' and that poor data are to blame for poor policies. But lack of data has historically played only a marginal role in the decisions leading to bad policies and poor outcomes. And a blind 'algorithmic' future may undercut the very processes that are meant to ensure that the way data are turned into decisions is subject to democratic oversight. At the same time, there is tremendous potential for societies to understand and affect processes that have grappled with for centuries. Fulfilling that promise will require profound reframing and rewiring of our political, ethical, technological and legal systems.

1.4 Big Data's future or the future's Big Data?

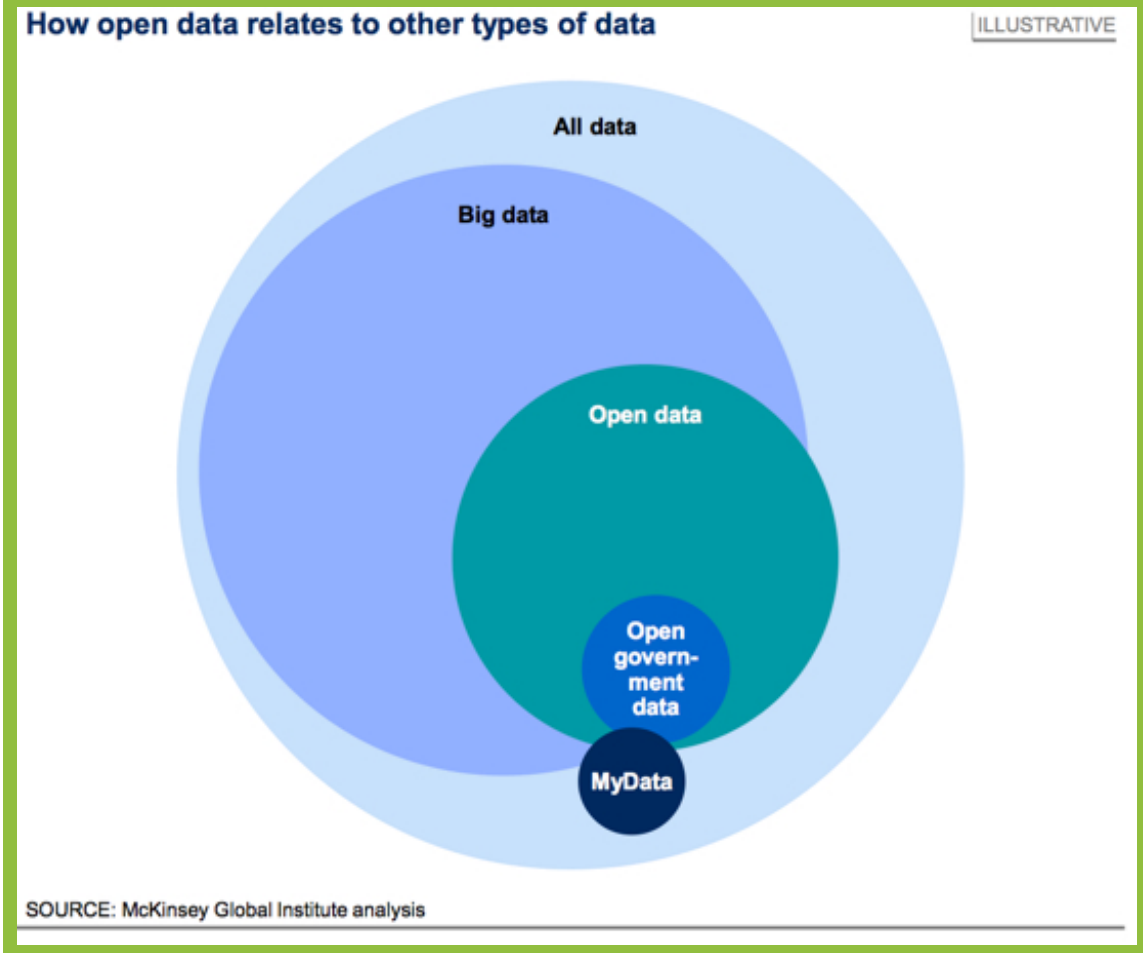
Since the growth in data production is highly unlikely to abate, and human creativity and curiosity are almost limitless, the 'Big Data bubble' is unlikely to burst in the near future. The world can expect more discussions and controversies about Big Data's potential and perils for development and societies at large. The future of Big Data will likely be shaped by three main related strands: academic research, legal and technical frameworks for ethical use of data, and larger societal demands for greater accountability and participation.

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

Policy debate will develop frameworks and standards — normative, legal and technical — for collecting, storing and sharing Big Data streams and sets. 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—although they will probably never make it impossible. But a comprehensive approach to the ethics of Big Data would ideally encompass other humanistic considerations such as privacy and equality, as well as champion data literacy and human-centered design.

A third related 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, accessible for free or at negligible cost, and with minimal limitations on its use, transformation, and distribution.

Classifications of Data



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, accountability and participation. The political overtones — so easily forgotten — are clear. 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.

2 Key Resources and Actors

An early mention of the upcoming "Industrial Revolution of data" can be found in a blog by UC Berkeley computer scientist Joe Hellerstein. It was published in November 2008, a few months after Wired had claimed that data deluge would signify "the End of Theory", where numbers would speak for themselves. Then in 2009, a group of leading computer and social scientists published a commentary in Science describing the advent of a new academic field leveraging data to reveal patterns of individual and group behaviors: "Computational Social Science". In early 2010, The Economist ran an article as part of a special report on the "Data Deluge" that stirred significant interest and remains highly informative today. The same can be said of the Wall Street Journal's "The Really Smart Phone" feature published in 2011, and the New York Times' "The Age of Big Data" opinion article published in 2012.

Since these early years, there has been an explosion in the number of publications about Big Data and international development, but three reports published within a few months in 2011-2012 can be considered seminal pieces in the field: the McKinsey Global Institute's "Big Data: the Next Frontier for Innovation, Competition and Productivity", The World Economic Forum's "Big Data, Big Impact: New Possibilities for International Development" and UN Global Pulse's "Big Data for Development: Challenges and Opportunities". Other noteworthy contributions include Martin Hilbert's literature review "Big Data for Development: From Information- to Knowledge Societies" and the International Peace Institute's report "Big Data for Conflict Prevention".

Several recent books that popularize the topic merit mention — one is Viktor Mayer-Schönberger and Kenneth Cukier's *Big Data: a Revolution That Will Transform How We Live, Work, and Think*. Others include MIT Professor Alex 'Sandy' Pentland's *Honest Signals* and the more recent *Social Physics*, a collection of essays edited by Lisa Gitelman titled *Raw Data is an Oxymoron*, as well as *Unchartered*, a book focusing on semantic analysis by Erez Aiden and Jean-Baptiste Michel.

Harvard Professor Gary King, one of the contributors to the 2009 Science article, has written many seminal articles about Big Data, notably a commentary on the "data-rich future of the social sciences" in addition to several technical papers. He has also talked about the "social science data revolution" to describe the opportunities and requirements of conducting social science research in the age of Big Data. On the specific issue of the impact of Big Data on social science research, a paper titled "The data revolution and economic analysis" by Liran Einav and Jonathan D. Levin is an engaging read, as is a post by Angus Whyte on the LSE blog. A few websites provide links to academic publications about Big Data or documenting research that makes use of Big Data. Notable examples include the Harvard Big Data for Social Good group and the MIT Human Dynamics Lab webpage.

Over the past couple of years, thousands of media articles and editorials have covered Big Data's impact on society. One of the most comprehensive is a Harvard Business Review feature published in 2014.

The Guardian online has a very good dedicated section (a 'data store') on Big Data. Many more informative articles can be found using a key word search for "Big Data" on the websites of major newspapers such as Forbes, the Wall Street Journal, the New York Times and, for French speakers, Le Monde. Specialized publications that offer more technical articles include the MIT Technology Review and Wired.

Critical summaries of key resources and challenges about Big Data and development can be found in a 2013 post on the website of European think-tank Bruegel, and in a post on the specific case of poverty monitoring on the website developmentprogress.com, facilitated by the UK think-tank ODI. Other good resources for an overview of the field include a Flipboard and a timeline curated by human rights activist Sanjana Hattotuwa.

Many organizations provide great resources through their work and websites. The UN Global Pulse, an innovation unit located in the Executive Office of the UN Secretary-General, has published two useful primers — a post on mobile-phone network data for development, hosted on its regular blog, and a primer on Big Data for Development. Another organization with interesting resources is the Qatar Computing Resource Institute.

In academia, leading universities and programmes where valuable resources can be found include Harvard University's Institute for Quantitative Social Sciences (IQSS), UC Berkeley's D-Lab, Columbia University's Institute for Data Sciences and Engineering, the Oxford Internet Institute, and the Harvard School of Public Health's Big Data for Social Good group.

Several universities have started offering study programmes in data science — some are available online and many are listed here, while free offerings are available on Coursera. As of today, there does not seem to be any offering focusing specifically on development and Big Data or data science. The University of Chicago offers a Data Science for Social Good Fellowship.

The website of the Data for Development (D4D) group — an informal consortium of institutions led by the UK mobile phone operator Orange — offers numerous resources related to a Big Data research competition it organized in 2012-13. GSMA, a global association of mobile operators and a D4D member, also provides resources on Big Data and development through its work on personal data. The World Economic Forum's own work on personal data is also worth considering. Other non-profit organizations of interest include, Flowminder, Gapminder, DataKind, and Bayes Impact.

Institutes and partnerships that have been created more recently include Data & Society and Data-Pop Alliance.

Anyone interested in Big Data and statistics can join groups such as Stanford University's Statistics for Social Good working group and Google's Data Science for Social Good group. A few bloggers are especially active. One is Patrick Meier on his iRevolution blog with a series of posts on Big Data, particularly applications related to humanitarian assistance and crises; another is Jay Ulfelder's on his Dart-Throwing Chimp blog, especially on issues of forecasting.

Another source of information on Big Data is of course Twitter — notably the hash tags #bigdata, which will yield close to one post per second, #bigdata4dev or simply #data4dev.

Several large foundations — such as the Knight Foundation, The Rockefeller Foundation and The Gates Foundation — have already showed interest in Big Data. Many of them, and probably more to come, are positioning this work under the larger umbrella of the data revolution agenda as part of the post-2015 framework of development goals — a good source of information on the topic is the post2015 site. A topic attracting growing attention is the impact of Big Data on official statistics, with useful information provided on the websites of UNECE and the UN Statistical Division.

There is a plethora of events and forums on Big Data with direct relevance to development: NetMob conferences on the analysis of cellphone data, Strata conferences, ICCM conferences, and TED talks on the topic. Other noteworthy videos include Professor Pentland's interview for the games-industry magazine The Edge, Kate Crawford's publication at the Strata 2013 conference and The Economist's data editor Kenneth Cukier's intervention at the latest TNW conference. Readers/viewers with a bit of time and an interest in technical aspects of Big Data can watch this presentation by Nathan Eagle as well as this Unconference on the Future of Statistics, or listen to a recent interview by one of its participants, Daniela Witten.

Issues of individual privacy, ethics and human rights around the use of Big Data are getting increasing attention. A good summary of the main positions and contributions in the 'privacy' debate can be found in a recent post on the NGO Privacy International's website, which contains many other valuable articles as well. Among the most prominent critical voices on Big Data are researchers danah boyd and Kate Crawford, who expressed skepticism in their 2011 essay "6 provocations for Big Data" and since then in other publications independently or with co-authors. Useful summaries of key challenges of using Big Data in crisis contexts can be found on The Human Rights Data Analysis Group's website, notably blog posts by Patrick Ball, and on the website of tech-training organisation Techchange. Various events are also dedicated to the responsible use of data, including the Responsible Data Forum series and a recent Ethics of Data conference held at Stanford in September 2014.

3 Definitions of Key Terms and Concepts

Big Data: an umbrella term signifying one or more of three trends: the growing volume of digital data generated as a by-product of the use of digital devices by people on a daily basis; new technologies, tools and methods available to analyse large datasets not originally designed for analysis; and the intention to extract from these data and tools insights that can be used for policymaking. In this primer, Big Data is distinguished from big data, where Big Data refers to a new ecosystem made up of its 3Cs for Big Data Crumbs, Capacities and Community.

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, and implications of, human communication patterns towards greater personal information sharing.

Data scientist or data science: a professional or a field that requires solving real-world problems using large amounts of data by combining (or blending) skills from often distinct areas of expertise: math and statistics, computer science (e.g. hacking and coding), statistics, social science and even storytelling or art.

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 for calculations and/or classification — that can teach themselves to grow and change when exposed to new data. It represents the ability to 'learn' to make better predictions and decisions based on what was experienced in the past, as with spam filtering, for example. The addition of "statistical" reflects the emphasis on statistical analysis and methodology, which is the predominant approach to modern machine learning.

Call Detail Records (CDRs): the technical name of cell-phone data recorded by telecom operators. CDRs contain information about the time, duration, as well as the locations of senders and receivers of calls or text messages transmitted through their networks. They are the automated equivalent of paper toll tickets that were written and timed by operators for long-distance calls in a manual telephone exchange.

(New) Digital divide: commonly defined as the differential access to and ability to use new information and communication technologies between individuals, communities and countries, and the resulting inequality in socioeconomic and political opportunities and outcomes. The skills and tools required to absorb and analyse large amounts of data resulting from the growing use of technology may provide the foundation for the creation of "new digital divide".

Algorithms (and algorithm future): in computer science, an algorithm is a series of predefined instructions or rules written in a programing language designed to tell a computer how to sequentially solve a recurrent problem, especially as it involves making calculations and processing data. The growing use of algorithms for decision-making purposes in an increasing array of activities and industries—e.g. policing and banking—has led to hopes and worries about the advent of an 'algorithmic future' where algorithms may come to 'rule the world'.

4 Selected Bibliography

Data data everywhere. Kenneth Cukier interviewed for The Economist

Emmanuel Letouzé. Big data for development: opportunities and challenges. (UN Global Pulse, May 2012)

Big data, big impact: new possibilities for international development. (WEF, 2012)

James Manyika and others. Big data: the next frontier for innovation, competition and productivity. (McKinsey Global Institute May 2011)

danah boyd and Kate Crawford. Six provocations for Big Data. (A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011)

Christopher Frank. Improving decision making in the world of Big Data. (Forbes, 25 March 2012)

Reinventing society in the wake of Big Data. A Conversation with Alex (Sandy) Pentland (Edge, 30 August 2012)

Social impact through satellite remote sensing: visualising acute and chronic crises beyond the visible spectrum. (UN Global Pulse, 28 November 2011)

Michael Horrigan. Big Data: a perspective from the BLS. Column written for AMSTATNEWS, the magazine of the American Statistical Association. (1 January 2013)

Gary King. Big Data is not about the data! Presentation (Harvard University USA, 19 November 2013)

Sanjeev Sardana Big Data: it's not a buzzword, it's a movement (Forbes blog, 20 November 2013)

Melamed C. Development data: how accurate are the figures? (The Guardian, 31 January 2014)

Laura Gray. How to boost GDP stats by 60% (BBC News Magazine, 9 December 2012)

The billion prices project. Massachusetts Institute of Technology

Measuring economic sentiment (The Economist, 18 July 2012)

Piet Daas and Mark van der Loo, Big Data (and official statistics) Working paper prepared for the Meeting on the Management of Statistical Information Systems. (23-25 April 2013)

Rebecca Tave Gluskin and others. Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends. (PLOS Neglected Tropical Diseases, 27 February 2014)

Emilio Zagheni and others. Inferring international and internal migration patterns from Twitter data. (World Wide Web Conference, April 7-11, 2014, Seoul, Korea)

New primer on mobile phone network data for development. (UN Global Pulse, 5 November 2013)

Joshua Blumenstock and others. Motives for mobile phone-based giving: evidence in the aftermath of natural disasters (30 December, 2013)

Michael Wu. Big Data Reduction 3: from descriptive to prescriptive. (Science of Social blog, Lithium 10 April 2013)

Arvind Narayanan and Vitaly Shmatikov Robust de-anonymization of large sparse datasets. Pages 111-125 in Proceedings of the 2008 IEEE Symposium on Security and Privacy (IEEE Computer Society Washington, DC, USA 2008)

Yves-Alexandre de Montjoye and others. Unique in the Crowd: The privacy bounds of human mobility (Nature scientific reports 25 March 2013)

Erica Goode. Sending the police before there's a crime. (The New York Times, 15 August 2011)

It is getting easier to foresee wrongdoing and spot likely wrongdoers (The Economist, 18 July 2013)

Kate Crawford. Think again: Big Data. Why the rise of machines isn't all it's cracked up to be. (Foreign Policy, 9 May 2013)

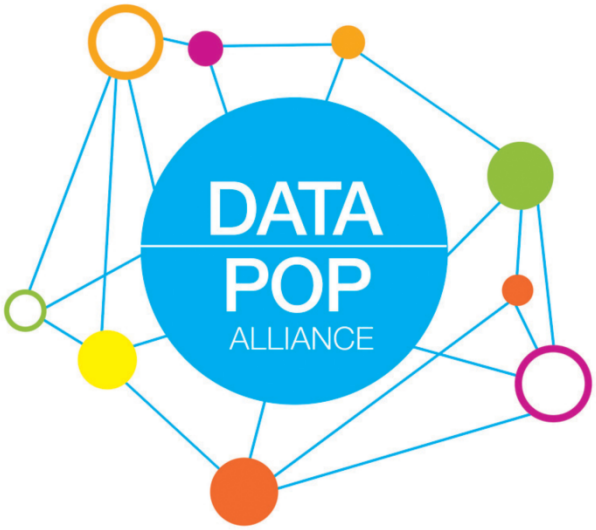
Neil M. Richards and Jonathan H. King. Three paradoxes of Big Data. (Stanford Law Review, 3 September 2013)

Neil M. Richards and Jonathan H. King. Big Data ethics. (Wake Forest Law Review, 23 January 2014)

Rahul Bhargava. Toward a concept of popular data. (MIT, 18 November 2013)

James Manyika and others. Open data: unlocking innovation and performance with liquid information (McKinsey Global Institute, October 2013)

Emmanuel Letouzé. The Big Data revolution should be about knowledge security (Post-2015.org, 1 April 2014



Promoting a people-centered Big Data revolution

mit media lab

HARVARD HUMANITARIAN INITIATIVE

ODI

Data-Pop Alliance is a think-tank on Big Data and development jointly created by the Harvard Humanitarian Initiative (HHI), the MIT Media Lab, and the Overseas Development Institute (ODI) to promote a people-centered Big Data revolution.

www.datapopalliance.org
contact@datapopalliance.org