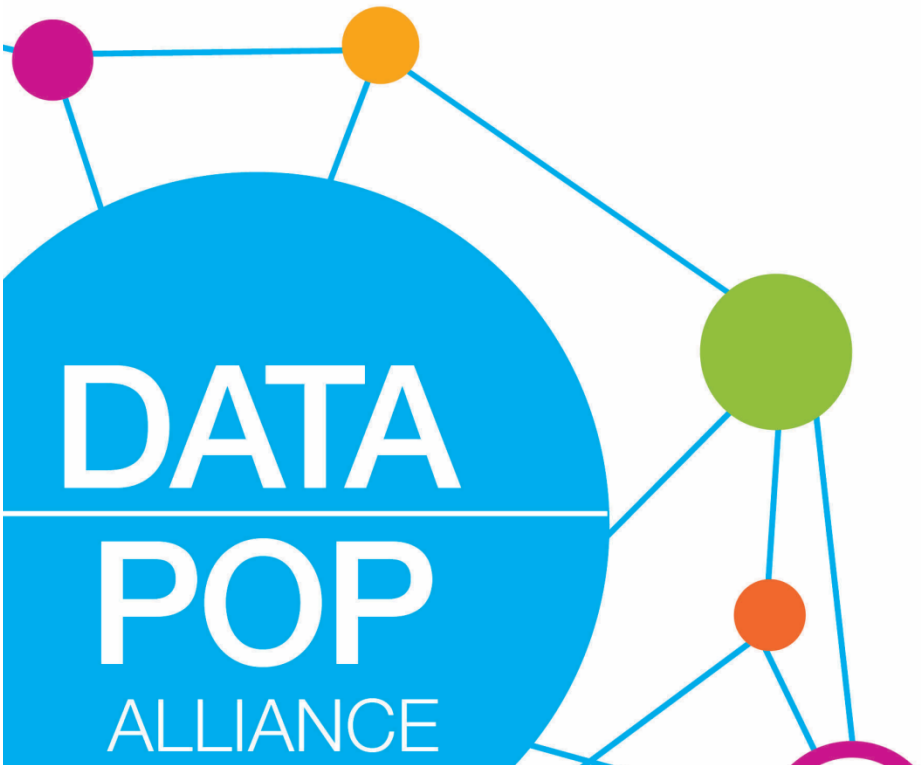
DATA-POP ALLIANCE WHITE PAPERS SERIES

OFFICIAL STATISTICS, BIG DATA, AND HUMAN DEVELOPMENT

March 2015



PARIS21

HARVARD HUMANITARIAN INITIATIVE

mit media lab

ODI

About Data-Pop Alliance

The 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 Authors

This paper was written by Emmanuel Letouzé and Johannes Jütting.

Emmanuel Letouzé (corresponding author) is the Director and co-Founder of Data-Pop Alliance. He is a Visiting Scholar at MIT Media Lab, a Fellow at the Harvard Humanitarian Initiative, a Research Associate at ODI, and PhD Candidate at UC Berkeley.

Contact: eletouze@datapopalliance.org.

Johannes Jütting is the Manager of Paris21 Secretariat, hosted at the OECD. He was a member of the United Nations Independent Expert Advisory Group on the Data Revolution for Sustainable Development. He is a member of Data-Pop Alliance's Steering Committee.

Acknowledgement

This is the inaugural paper in Data-Pop Alliance's White Papers Series developed in collaboration with our partners.

An initial version of this paper was commissioned by Paris21 for a special technical session on "The potential of Internet, big data and organic data for official statistics" at the 59th World Statistical Congress of the International Statistical Institute held in August 2013 in Hong Kong, during which Emmanuel Letouzé presented views discussed in more details in this paper.

The views presented in this paper are those of the authors and do not represent those of their institutions. This version benefited from suggestions and ideas from Gérard Chenais, Michail Skaliotis. It is currently under peer review. All errors and omissions remain those of the authors.

Citations: Letouzé E, Jütting J. "Official Statistics, Big Data and Human Development." Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute. March 2015.

DATA-POP ALLIANCE
WHITE PAPERS SERIES

OFFICIAL STATISTICS, BIG DATA AND HUMAN DEVELOPMENT

March 2015

Written by

Emmanuel Letouzé
Johannes Jütting

Data-Pop Alliance
Paris21

Table of Contents

Abstract 5

1 Introduction 6

2 Roots and State of the "Big Data and Official Statistics" Question 7

2.1 The statistical disillusion 7

2.2 The rise of Big Data 9

2.3 Pilots and controversies in official statistics 11

3 Revisiting the Terms and Phrasing of the Question 13

3.1 Big Data isn't just big data: from the three V's to the three C's 13

3.2 The dual nature and purpose of official statistics 15

3.3 Why engaging with Big Data is not a technical question but a political obligation 18

4 Towards a New Conceptual and Operational Approach 20

4.1 Proposed conceptual pillars for knowledge secure societies 20

4.2 Proposed operational principles to create a deliberative space 22

5 Concluding Remarks: Sketching the Contours of an Action Plan 24

Annexes 26

Annex 1. Application of Big Data to societal questions: two taxonomies and examples 26

Annex 2. How big is Big Data? 27

Annex 3. Predicting socioeconomic levels through cell-phone data and machine learning 28

End Notes 30

Abstract

This paper aims to contribute to the ongoing and future debate about the relationships between Big Data, official statistics and development—primarily by revisiting and reframing the terms and parameters of this debate.

Most current discussions on Big Data mainly focus on if and how it can contribute to producing faster, cheaper, more frequent and different development indicators for better policies. This paper takes a different starting point. It stresses the fundamental political nature of the debate, encouraging us to go and think beyond issues of measurement and stressing the centrality of politics beyond policy.

It argues that in fact Big Data needs to be seen as an entirely new ecosystem comprising new data, new tools and methods, and new actors motivated by their own incentives, and should stir serious strategic rethinking and rewiring on the part of the official statistical community.

It contends that the emergence of this new ecosystem provides both an historical opportunity, and a political and democratic obligation for official statistical systems: to recall, retain or regain their primary role as the legitimate custodian of knowledge and creator of a deliberative public space for and about societies, to discuss and drive human development on the basis of sound democratic and statistical principles.

1 Introduction

One dimension of the ongoing "Data Revolution"[1] discussion is how Big Data may or should impact official statistics. A question that has attracted attention has been: Is Big Data a potential fix, a possible threat, or irrelevant to the dearth of data in poor countries that some have named a "statistical tragedy"?[2]

This question has led to several publications, forums and pilot projects since the beginning of 2013, the vast majority in and about OECD countries.[3] In more recent months, numerous articles about, references to and groups working on the (or a) "Data Revolution" have emerged, with specific or implicit reference to developing countries, culminating in the publication of a report by a UN-appointed Independent Expert Group (IEG).[4]

The prevailing answer to this question may be summarized as follows: Big Data may provide faster, cheaper and more granular data to complement, but certainly not replace, official statistics, and to design and implement better policies and programs, but many challenges remain that will require adapted responses.

But in the absence of many specifics as to how—and also exactly why—this may or should happen, and how this may change the world we live in for the better, it has been difficult to convert critics who frame Big Data as mere hype and argue that resources would be best invested elsewhere—for instance making sure that all countries conduct regular censuses and surveys and collect "basic statistics".

These perspectives and criticisms evidently contain some truth, and most contributions to the debate have made good points. Notably, there has been a growing and welcome recognition that better data won't magically or mechanistically lead to better policies and better outcomes—just as bad data was not the primary cause of persistent poverty, rising inequality, environmental degradation and social oppression in the first place.

This paper argues that while the current debates have led to an increased understanding of the opportunities and challenges ahead, the fundamental question is not whether and how Big Data as data may or may not facilitate the production of official statistics. Rather, taking a systematic approach and looking beyond issues of measurement, this paper argues that the emergence of Big Data as an entirely new ecosystem requires the official statistical community to engage with the Big Data community in order to reap the potential sizeable benefits for human development. This approach is also significant in order to maintain relevance and to avoid putting the societies it is mandated to serve at risk. Overall, there is an urgent need for greater conceptual clarity, technical specificity, political breadth and strategic foresight than has generally been the case so far.

This paper aims to contribute to that goal in two main ways. First, by revisiting the traditional terms and framing of the Big Data and official statistics question; and second, by suggesting a number of principles and steps that official statisticians and statistical systems—including, but not confined to, national statistical offices (NSOs)[5]—may follow as they start engaging in the Big Data era—as we believe should urgently happen.

The rest of this paper is structured as follows. Section 1 summarizes the roots and state of the "Big Data and official statistics" question; Section 2 discusses why and how the question would benefit from being approached in greater depth and breath; Section 3 suggests the preconditions, principles and policies for action.

2 Roots and State of the "Big Data and Official Statistics" Question

Interest in the "Big Data and official statistics" question has its roots in two main developments: the statistical "disillusion" and the rise of Big Data.

2.1 The statistical disillusion

We refer to the statistical 'disillusion', instead of what many have called the statistical "tragedy", defined as poor or scarce data in certain regions. Although this paper is primarily concerned with developing countries, it is worth noting that the statistical disillusion is not restricted to them. Cases in point of the past decade include the statistical dimension of the Greek crisis[6] and the realization that the systems compiling quarterly gross domestic product (GDP) in OECD countries did not function well in times of increased economic volatility.[7]

Many citizens and organizations are unhappy with the current state of official statistical affairs. A recent article, "Big-Data Men [would] Rewrite Government's Tired Economic Models", described a San Francisco-based start-up whose co-founder believed that "we shouldn't have to rely on the creaky wheels of a government bureaucracy for our vital economic data".[8] Another example is the historical discontent with the way human welfare is measured—or not measured—which has led to the development of alternative indicators to GDP.[9]

Of course, the statistical disillusion tends to be greater in developing countries, despite noteworthy progress, including the monitoring of Millennium Development Goals (MDGs) in general and more specifically the tracking of the income poverty target 1.A.[10] For instance, Ghana's GDP grew over 60% overnight after it undertook a rebasing exercise in 2010;[11] Nigeria's

grew by over 75% after a similar exercise in April 2014.[12] The root cause of the inaccuracy lay in GDP data based on old consumption and production patterns.

A handful of countries—not all poor, but all with a history of violence—have not conducted a population census in decades. Their populations and any per capita data can therefore only be estimated, even "official" figures, which are supposedly precise, are most likely inaccurate.[13]
A country like Kenya has not produced poverty data in almost a decade.[14] Unemployment figures are close to meaningless in most developing countries. The list goes on. Poor-quality poverty data, in particular, turns monitoring and forecasting into exercises in "guesstimation" where sub-continental averages have to be used as crude proxies for country-level data.[15]

In addition to the structural challenge of rebasing GDP, well-known and mutually reinforcing determinants of poor-quality data include poor human and technical capacities—both the cause and effect of a brain drain, whereby the best-trained official statisticians in low-income countries are hired by the private sector (or the international development agencies)—as well as low levels of financial resources, lack of trust and dialogue, inadequate institutional organizations, weak or ill-motivated political will, etc.[16]

Importantly, as the "pilot in a plane" analogy below suggests, much of the focus when answering the "Big Data and official statistics" question has been on the data deficit, the subsequent measurement challenge, and the need to get better data into the hands of well-meaning policy makers to devise better policies.

A 36,000-foot perspective on Africa's statistical tragedy

How would you feel if you were on an airplane and the pilot made the following announcement: '*This is your captain speaking. I'm happy to report that all of our engines checked fine, we have just climbed to 36,000 feet, will soon reach our cruising speed, and should get to our destination right on time.... I think. You see, the airline has not invested enough in our flight instruments over the past 40 years. Some of them are obsolete, some are inaccurate and some are just plain broken. So, to be honest with you, I'm not sure how good the engines really are. And I can only estimate our altitude, speed and location. Apart from that, sit back, relax and enjoy the ride.*' This is, in a nutshell, the story of statistics in Africa. Fuelled by its many natural resources, the region is quickly growing, is finally beginning to reduce poverty and seems headed for success. Or so we think, for there are major problems with its data, problems that call for urgent, game-changing action.

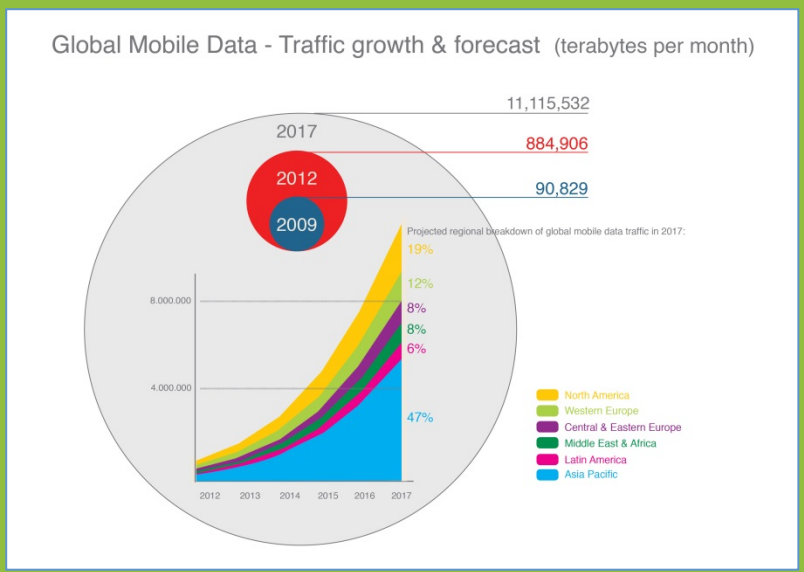
Source: Giugale, M. (4 March 2012), "Fix Africa's statistics", The World Post, www.huffingtonpost.com/marcelo-giugale/fix-africas-statistics_b_2324936.html

2.2 The rise of Big Data

The second development is what was initially termed "the industrial revolution of data"[17] in 2008—and later simply "Big Data". Big Data has been described as "data sets that are impossible to store and process using common software tools, regardless of the computing power or the physical storage at hand"[18]. Mike Horrigan at the United States Bureau of Labor Statistics (BLS) defined Big Data as "non-sampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference." [19] The absence of a single agreed-upon definition of Big Data is much less problematic than the continuing focus on Big Data as "just" big datasets—or simply as the "three Vs" of volume, velocity, and variety used to characterize Big Data in its early years until 2012-13.

Although we discuss definitional considerations in greater detail below, the starting point and central feature of Big Data as a phenomenon is indeed the unprecedented growth in the volume and variety of high-frequency digital data—structured and unstructured—passively emitted by human populations in the course of their activities. Each year since 2012, over 1.2 zettabytes of data have been produced — 1021 bytes, enough to fill 80 billion 16GB iPhones (see Annex 2 for more information about the measurement of data).[20] As the figure below shows, the volume of data is growing rapidly. And, just as a human population with a sudden outburst of fertility gets both larger and younger, the proportion of data that are less than a day or a year old (i.e. which would be akin to 'baby data') is growing—starting in 2013 it has been said many times that up to 90% of the world's data has been created over the two preceding years (2011-2013), but the assertion is almost impossible to corroborate.[21]

Data inflation



A great deal has been discussed and written about the applications and, to a lesser extent, the *implications* of Big Data on public policy and social science,[22] including an increasing focus in official statistics.[23]

One argument is that Big Data could provide faster, cheaper, more granular data and help meet growing and changing demands. It was claimed, for example that "Google knows or is in a position to know more about France than INSEE, [the National Institute of Statistics and Economic Studies]".[24] Others have also argued that Big Data may partially "fix the statistical tragedy"[25] of some developing countries—for example by providing near real-time, fine-grained poverty figures. Some argue that Big Data may even facilitate a "leapfrogging" of poor countries' statistical systems[26]—the same way some regions have bypassed fixed telephone lines to go straight to cell phones, and seem to be heading into the smart phone and tablet era without going through the PC stage. Indeed, a key source of Big Data is phone data known as call detail records (CDRs), most of which comes from cell phones, recorded by telecom operators (see below). It is estimated that over 80% of Internet traffic will soon transit through hand-held devices, which will lead to further creation of geo-localized, time-stamped data.

Call detail records

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 the call. Large operators collect over 6 billion CDRs per day.

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

Global Pulse (2013), Mobile Phone Network Data for Development, Global Pulse, http://www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev .

The broad notion that new digital data and tools hold the potential to increase both temporal and geographical granularities while reducing the costs of data collection can be traced back to two seminal 2009 papers, one on GDP and night light emission[27] and one on "now casting" via Google queries.[28] More recent and much-cited examples and applications have given further impetus to this argument. These include efforts to track inflation online such as those spearheaded by the Billion Prices Project (BPP),[29] to estimate and predict changes in GDP in near real-time,[30] and to monitor traffic.[31] Other somewhat less widely cited papers have used email data and CDRs to study migration patterns, malaria spread, etc. in development

contexts.[32] The field of sentiment analysis from social media data is also opening additional avenues to develop alternative measures of welfare.

A foundational question is how Big Data can be used for development. Subsequent papers proposed two main taxonomies of the applications of Big Data (see Annex 1 for more detail on these taxonomies). A UN Global Pulse paper[33] proposed a three-tier taxonomy of uses: real-time awareness, early warning and real-time feedback. Letouzé, Meier and Vinck distinguished descriptive (e.g. maps), predictive (understood as either better forecasting of what may happen next, or proxying or inferring some variable of interest via another) and prescriptive (the realm of causal inference) functions.[34] Whereas interest in the prescriptive function is poised to grow,[35] most applications have relied on the first two, and perhaps most visibly on the predictive function.

2.3 Pilots and controversies in official statistics

Aware of these changing conditions, and interested in exploring most potentially cost-saving avenues,[36] several NSOs from OECD countries have started experimenting with Big Data—as well as leveraging large administrative datasets, which we don't consider as part of Big Data. This has typically initially been done through pilot projects, for example within the US BLS,[37] Statistics Netherlands (with traffic loop and social media data, notably),[38] Statistics Korea,[39] Statistics Ireland, Statistics New Zealand and others. Other initiatives are in the making, mostly in OECD countries, although the Paris21 initiative and partners such as ODI are planning to increase their support to developing country partners. The UN Data Revolution report[40] has called for increasing support to NSOs in the area of new data.[41]

At the global level, a (non-representative) survey designed by the authors of this paper and administered by the Paris21 initiative in 2013 sent to NSOs and regional statistical offices found that almost all respondents (94%) felt that "Big Data [could] be used to supplement official statistics", and close to 80% said that it "should". Only slightly over half said that Big Data had been talked about, and less than 15% that it had been used, in their institutions.[42]

This low level of current engagement may find its root in a mix of skepticism based on concerns over data quality and reliability as well as ethical considerations and fear of "losing relevance"[43] within the official statistics community. Big Data has been described as simply bad data, or "the tech world's one-size-fits-all (…) answer to solving the world's most intractable problems".[44] As such, these critics claim it to be ill suited to countries that may have more pressing statistical concerns,[45] including those that are unable to collect and produce "basic statistics".

Even in the case of an advanced economy, France, the head of its NSO, INSEE, stated in an official document that Big Data currently seemed largely irrelevant to its work: "INSEE is following big data attentively. However, all articles on the subject refer to very advanced indicators that,

for the time being, present little interest, in that they can only save a few days compared to the production of cyclical statistical indicators and nothing that seems to be operational in that respect".[46]

Little methodological research has yet been done by NSOs. For instance they have done hardly any work in the area of sample bias correction. However studies in this area are being explored in the academic realm. Harvard faculty led a study to estimate the impact of biases in mobile-phone ownership on CDR-based estimates of human mobility[47] and another academic paper has proposed a correction factor for email-based estimates of international migration flows.[48] Interestingly, both have been largely unmentioned in the current "official" debate, despite their seminal nature.[49]

Whatever their take on Big Data, official statisticians express an acute and understandable sense of frustration over pressure to open up their data to private-sector actors, while these same actors are increasingly locking away what they and many consider to be "their" data—although this notion is getting increasing challenged—in order to protect both their consumers' confidentiality and perhaps even more their own commercial interests.

These different points of view have nonetheless coalesced around a loose consensus that these vast volumes of data may provide an opportunity, to supplement ("but not replace", one is pressed to add immediately) official statistics, either by helping to produce faster or more accurate figures on inflation or GDP, for instance, or coming up with alternative measures.

But we argue that the state of affairs isn't satisfactory for two main reasons. First, because the terms of the question should be deepened and its frame broadened to fully account for its complexity and importance—in other words there is a need to revisit and clarify what we are talking about. Second, because strategic and operational considerations and suggestions are often lacking. The following sections consider both aspects in turn.

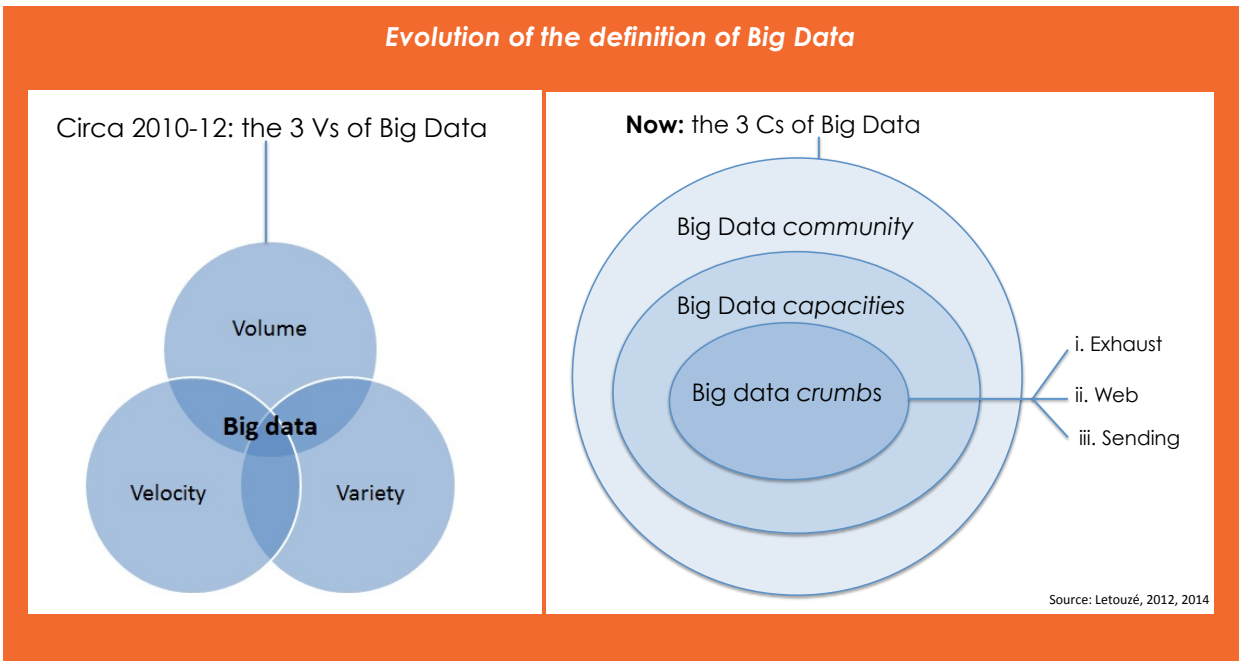
3 Revisiting the Terms and Phrasing of the Question

3.1 Big Data isn't just big data: from the three V's to the three C's

Let us start by unpacking and clarifying the terms of the question to revisit it on firm ground. We start with Big Data. What exactly is Big Data?

In this paper, as in others, we argue that a sole focus on Big Data as (big) data does not provide an adequate basis for thinking about the applications and implications of Big Data for official statistics. An Excel file containing CDRs for 100 cell towers aggregated on a daily basis may be small, and yet, as data, it is Big Data; the World Development Indicators database and all the censuses ever conducted constitute very large files, and yet they are not Big Data. Focusing so much on Big Data as big data has lead to incorrect assumptions and unnecessary controversies—the notion that Big Data is or isn't about providing an "automated 30,000-foot view of the world."[50]

Our starting point is to characterize Big Data not as the three Vs of volume, velocity and variety, as has long been the case, but through three Cs.



The first C stands for "crumbs"[51]— identifying Big Data as new kinds of passively-generated individual and networked "traces of human actions picked up by digital devices".[52] These "digital breadcrumbs" have the potential to paint a picture of some aspects of the social world with unprecedented levels of detail and shades.[53] Their fundamental revolutionary nature is qualitative.

From a systems perspective, these data are not just a byproduct of digital and connected societies that can be used to report on human ecosystems: some of them—such as social media data—have an endogenous effect by shaping preferences and fueling aspirations. They also produce an economic surplus that goes largely unrecorded in GDP—adding to the traditional measurement challenge.[54] Big Data may also change what we can and care to measure in powerful and complex ways.

The second C stands for "capacities". In the words of Harvard social scientist and statistician Gary King, "Big Data is not about the data".[55] It is "about" the intent and capacity[56] to yield and convey what are routinely and vaguely referred to as "insights"[57] (which appears a safer term than "information") from these qualitatively new kinds of data. Part of it involves advanced storage and computing capacities and part entails advanced quantitative and computer science methods and tools—primarily statistical machine-learning techniques, algorithms, etc.

One example is Telefonica researchers' attempt to "predict" socioeconomic levels (SELs) in a "major city in Latin America" (Mexico City). They matched CDRs and official survey data, using supervised machine learning to identify differential digital signatures of different SELs and to create a model applicable to other regions and/or later points in time (see Annex 3 for an illustrated explanation).[58] This example illustrates the focus on the predictive use of Big Data, here understood as proxying, with little movement yet towards causal inference. Yet another aspect of Big Data capacities is visualization techniques and tools that allow complex trends and patterns to be presented in appealing and often customizable ways.

The third C of Big Data stands for "community": Big Data must also be considered as referring to the people and groups "making use" of crumbs and capacities (Andreas Weigend actually defines Big Data as a "mind-set").

Many of these new actors have embraced and indeed spurred the open source movement and new ways of working based on the lessons of agile software development—the development of the statistical computing software, R, being a key example. Others in the private sector or intelligence communities function in a highly controlled and secretive manner, for obvious commercial and political reasons. At the moment most of the "Big Data" is held by private corporations—telecom companies, financial institutions, etc.—and only a handful of datasets are in the public domain, most of them unstructured and hard to work with.

The key point that cannot be stressed enough is this: Big Data is not just big data. It is about qualitatively new kinds of data regarding people's behaviors and beliefs, new kinds of tools, and new kinds of actors. If any discussion of the applications and implications of Big Data for official statistics (and development) is to be meaningful, Big Data must be approached and conceived of as an ecosystem—a complex system—made up of the data, the tools and methods, and the actors.

Asking whether and how Big Data may or should affect official statistics is not only about whether and how official statisticians should or should not use Big Data to produce official statistics. It is about whether, why and how, official statisticians and systems should deal with Big Data as an emerging complex ecosystem. Fully understanding and addressing this question also requires a good understanding of what official statistics are.

3.2 The dual nature and purpose of official statistics

While most discussions about Big Data and official statistics do provide some definitions of Big Data, they overlook much consideration of the concept of official statistics. This is probably because the term "official statistics" is so pervasive in our personal and professional lives that we take their nature and role as givens. And yet, unpacking what we mean by and expect from official statistics provides useful insights into the larger question at hand.

In our view, and as previous authors have underlined,[59] "official statistics" have a dual nature and serve a dual purpose:

1. as measurement tools, i.e. data, produced by official bodies and systems,
2. and as entities and systems producing official statistics (which are not restricted to their core - NSOs).

When one wonders or worries about the "future of official statistics" in the age of Big Data, it is clear that both of these purposes are important to consider. Both components are linked by a non-symmetrical relationship: on the one hand, the former—official statistics as data— are fully defined in reference to the latter—bodies and systems. On the other hand, official statistics as bodies and systems are not or should not be solely defined by their measurement tools.

The only defining feature of official statistics as data is to be provided by official statistical bodies and systems, produced on the basis of other data sources—survey, other—according to professional standards and norms reflected in the *Fundamental Principles of Official Statistics*.

It follows that any data produced or provided by an official statistical body (in such a way) is and would be called official statistics. It is also worth noting that official statistics can, according to these fundamental principles, draw on "all types of sources".[60]

Fundamental Principles of Official Statistics (revised, 2013)

Principle 1. Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Principle 2. To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

Principle 3. To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Principle 4. The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Principle 5. Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

Principle 6. Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Principle 7. The laws, regulations and measures under which the statistical systems operate are to be made public.

Principle 8. Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

Principle 9. The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

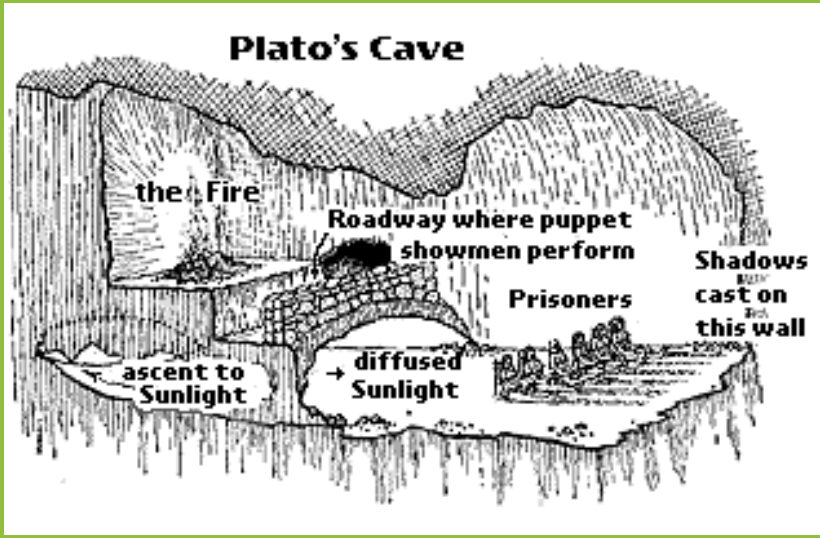
Principle 10. Bilateral and multilateral= cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

Source: United Nations (2014), Fundamental Principles of Official Statistics, UN General Assembly Resolution 68/261, United Nations, http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf

But "official" and "real" are of course not synonymous. Official statistics as data are quantitative constructs meant to measure real-world socioeconomic processes and aggregates via

processes and rules that "confine and tame the personal and subjective".[61] As such, to paraphrase Plato, they are necessarily—by construction, i.e. even the best ones—shadows in the cave. These shadows can be blurry, confusing, misleading. The percentage of people below the poverty line is an example: it is not a true-to-life picture of human reality and deprivation in a given area—it is a very crude proxy. The use of GDP for "the economy" (which grew/expanded or declined/contracted) is another obvious example.

Official statistics as shadows in a cave



Official statistics as bodies and systems, by contrast, are not and should not be solely defined by their official statistics. In other words, their defining feature, their essential role, is not to produce official statistics as data—it is not even to produce information. What is it?

Official statistics not only have a dual nature, but they also serve two main functions.

The first and probably most fundamental function is to produce knowledge. In the words of Enrico Giovanni in 2010, the essential role of modern statistics, referred to as "statistics 2.0",[62] is to provide society with "knowledge of itself, on which to base its own choices and evaluate the effects of political decisions."[63]

The second function of official statistics is to provide a deliberative space to freely and openly debate what is worth measuring, how it is measured and why it is measured—to act as "a debated public institution".

These points stress how official statistics aren't merely an—often poor—mirror of the world that benevolent and enlightened policy makers use to craft policies, but fundamentally a public industry that exists to transform the world by creating knowledge and providing a public space

where deliberative discussions can take place. This poses the question of what it takes and implies for these functions to be met.

It is often argued that NSOs of poor countries should first and foremost focus on producing basic statistics. This begs the question: what are basic statistics? Who gets to decide what they are? As Enrico Giovannini has put it in public debate, no one in their right mind would come up with GDP as a measure of economic activity in the 21st century—an indicator that has most likely led to environmental degradation on a catastrophic scale and yet, despite all its flaws, remains the alpha and omega of economic policy making and ranking.

In addition, for all the talk and hopes about the advent of evidence-based policy making, finding policies firmly rooted in evidence is no simple task; rather, they often have roots in political conundrums and conciliations in which official statistics are often sidelined.

3.3 Why engaging with Big Data is not a technical question but a political obligation

These facts matter for several reasons. First, they suggest that using Big Data to shed light on human societies poses non-trivial conceptual, theoretical and technological challenges as well as deep anthropological, ethnographic and ethical ones. These challenges are discussed in papers and contributions that are seldom referred to in much level of detail if at all in talks about the "opportunity" provided by Big Data for official statistics.[64]

They also suggest that the bulk of the tools and skills required to yield the "insights" offered by Big Data are today being developed outside of the official statistical community, very often using data stored and held by private companies—increasingly so in the case of CDRs, for instance. But it is not clear at all that holding data means ownership. Official statistical bodies must weigh in strongly on the debates over data ownership and control.

Lastly they stress the fact that many different actors are producer of insights and information about and within modern techno-infused societies, and that insights and information are different from knowledge.

So we argue that to a large extent, the fundamental question is not whether the official statistical community should use Big Data as substitutes for or complements to official statistics, but why the official statistical community should engage with the Big Data community to fulfill their mandated role: to provide societies with "knowledge of themselves" that reflects what they care to know and mean to impact, and a public space to discuss what that is.

This directs attention to the crux of the issue: this discussion has at its core an essentially political dimension (as the term "Data Revolution" suggests). Olavi Niitamo, former director general of Statistics Finland, said: "Knowledge is power; statistics is democracy". Twenty-first century official statistical systems must thrive to ensure that societies benefit from knowledge and can deliberate on the objectives and impact of policies in ways that reflect and serve societal aspirations, sound technical standards and democratic principles.

The fact that the technological revolution "has put an end to the monopolistic power that statistical institutes held until around twenty years ago" [65] is uncontroversial, and overall unproblematic. But it is clear that Big Data alters and accelerates these dynamics: it is presiding over the rapid emergence of institutions and individuals with unmatched access to data and resources that are able to report on and influence societies outside the realm and reach of governments and other traditional policy actors—especially in developing countries.[66]

What are some of the risks of non-engagement? The most important question—or answer—is not whether NSOs may or may not lose their relevance. It is, to reassert the point, that societies may not, even less than today, benefit from knowledge that reflects and serves democratic principles and processes, based on information subject to checks and balances, reproduction, verification and contestation. What may follow is a two or more tier system, with a proliferation of alternative "official" statistics—where "official" refers to the sort of figures official statistical systems have traditionally reported on.

The example of the San Francisco-based start-up provides a case in point: what if the non-official "official statistics"—the inflation figures—differ vastly from the official "official statistics"? Who shall the public and investors trust or distrust? Official inflation figures are already largely distrusted in much of Europe since at least the introduction of the euro, which, in Italy for instance, led to and was amplified by the construction and use of "completely unreliable inflation estimates, produced by private institutes using very weak methodologies, producing confusion and encouraging wrong behaviors"[67]. Another recent example showed that the results of 2010 census figures for the population in the United States are being disputed on the basis of sewage data.[68] What if these examples multiply?

There is also a real the risk of observing a growing tension and mismatch between societal demands and official supply. Focusing on producing finer, more accurate estimates of GDP—an indicator devised in a data-poor industrial era that was never intended to be a measure of welfare by its creator—may not be the right approach. Social media companies, academic teams and civil-society organizations will increasingly leverage the tremendous potential of Big Data to devise and release near real-time alternative measures of well-being.

Last, an obvious risk of non-engagement is the creation of a new digital divide that may result, in less than a decade, in a situation where development data on the poorest countries may be

derived remotely, with little to no inputs from local official institutions, communities and societies; where no adequate skills and networks are built locally.

The main message is this: for official statistics, engaging with Big Data is not a technical consideration but a political obligation. It is an imperative to retain, or regain, their primary role as the legitimate custodian of knowledge and creator of a deliberative public space for and about societies to discuss and drive human development on the basis of sound democratic (including ethical) and statistical principles.

The good news is that it is entirely possible. Official statisticians are very well placed to do so—both in terms of their legitimacy and their skills. For example, the fact that statistical properties of stationary, linearity and normal distribution are unlikely to hold with high-frequency data is well known, but the Big Data community seems to be doing little to address this issue.

Similarly, as mentioned above, too much emphasis is currently being placed on prediction capacities and models; official statisticians have much to offer to shape and strengthen the study of causal inference using these new data, not just to predict what has happened or may but to understand why. They don't have to do all of this alone—there is indeed an entire community to tap into.

A number of preconditions and principles can usefully inform future policy actions.

4 Towards a New Conceptual and Operational Approach

4.1 Proposed conceptual pillars for knowledge secure societies

In order to maximize the benefit of Big Data and minimize the risks that may arise from it, we first argue that societies must strive to be knowledge secure much in the same way they need to strive to be food secure. We propose a conceptual framework mirrored on that of food security.

According to the United Nations, a society is food secure "when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life".[69] These conditions are translated into four pillars of food availability, access, utilization and stability. What does it take for a society

to be knowledge secure? We suggest it takes data availability, data access, data utilization and data stability.

To clarify the argument further, in what follows, the only major changes to the (italicized) description of the four pillars of the official FAO food security framework[70] is the substitution of "data" or "knowledge" for "food" as appropriate,[71] plus a few minor edits. We also discuss some of the implications of these preconditions.

Thus, promoting "knowledge secure" societies in the age of data may entail ensuring:

1. ***Data availability***—*i.e. the availability of sufficient quantities of data of appropriate quality, supplied through domestic production or imports (including data aid).*

This principle brings out the importance of producing data that meets societal demands and needs of the time—not what is deemed official statistics. It also stresses how "quality", as characterized in the second of the *Fundamental Principles of Official Statistics*,[72] has to remain a central concern in the production of data by official statistical systems—i.e. official statistics. An example is that "the blending of estimates drawn from traditional statistical methods and the incorporation of larger universe data requires clear statements of how these estimates are developed and a perspective on potential sources of sampling and non-sampling errors that can produce biases in our estimates and threats to valid inference."[73]

2. ***Data access***—*i.e. the access by individuals to adequate resources (entitlements) for acquiring appropriate data to enhance their knowledge. Entitlements are defined as the set of all commodity bundles over which a person can establish command given the legal, political, economic and social arrangements of the community in which they live (including traditional rights such as access to common resources).*

This highlights the importance of transparency, user-friendliness and visibility. For example, knowing that "95% of Google users do not go beyond the first page, it is clear that either institutes of statistics structure their information in such a way as to become easily findable by such algorithms, or their role in the world of information will become marginal".[74] It also stresses how official statistical bodies and systems have to play a central role in all debates over ownership to the rights of control over personal data.

3. ***Data utilization***—*i.e. the utilization of data through adequate individual and collective processing to reach a state of knowledge where all information needs are met. This brings out the importance of non-data inputs in knowledge security.*

This critical point stresses the fundamental importance of "considering how [information] is brought to the final user by the media, so as to satisfy the greatest possible number of individuals (not only members of the government or of an economic or cultural elite), the extent to which

users trust that information (and therefore the institution that produces it), and their capacity to transform data into knowledge (what is defined as statistical literacy)"—to which we prefer the concept of data literacy and add that of graphic literacy (or "graphicacy").[75]

4. ***Data stability***—*i.e. to be knowledge secure, a population, household or individual must have access to adequate data at all times. They should not risk losing access to data as a consequence of sudden shocks (e.g. an economic or climatic crisis) or cyclical events (e.g. seasonal data insecurity). The concept of stability can therefore refer to both the availability and access dimensions of knowledge security."*

This notably suggests the need to put in place legal and policy frameworks and systems that ensure a steady and predictable access to some data—even aggregated, always anonymized—which is currently held by corporations but whose rights and control ought to be put in the hands of their emitters and their representatives. This new data ecosystem would be in stark contrast to the ad hoc way researchers have tended to access data such as CDRs in recent months—apart perhaps from the counterexample of the first and second Orange D4D challenges,[76] although this initiative has its critics.

Critically, this conceptual framework is intended to complement the *Fundamental Principles of Official Statistics* by distilling their core features and implications to sketch the four preconditions (or 1st-order priorities) of "Statistics 2.0" in the Big Data age. It is also consistent with, but at a higher level, than the principles of the UN IEA group.

4.2 Proposed operational principles to create a deliberative space

We propose four operational principles to facilitate the creation of a deliberative space for societies to debate issues of measurement and policy objectives.

1) Shared responsibility

It is not just official statisticians and statistical systems who are responsible for enhancing knowledge within societies, or even meeting the preconditions identified above—many other actors have a responsibility to contribute, be it through the provision of data, information or cognitive and analytical capacities and tools (e.g. private corporations, the media, political actors broadly understood, and the education system).

Further, turning greater knowledge into better public policies is definitely a major goal of the "data revolution". Doing so falls outside the mandate of official statistical institutions and systems. But their responsibility—their mandate and its implications—cannot be stressed enough—and behind theirs those of all major donors.

Private corporations also have a responsibility as the keepers of an increasing share of data. Compelling private companies to systematically share their raw data with official institutions and researchers would be both unrealistic and undesirable. What is needed is to devise mechanisms and legal frameworks for private companies to share their data under formalized and stable arrangements.[77]

Private companies—such as Microsoft or IBM research, SAS, and the research and development (R&D) arms of telecom companies—could also directly partner with official statistical systems. Researchers and policy makers also bear significant responsibility—in particular, they should avoid ad hoc requests for data and instead support efforts to find standardized, ethical and stable data-sharing tools and protocols with private companies.

The media bear responsibility too. They "could undertake not to give space to statistical data on themes which, however curious and potentially interesting they may be, are produced according to methods that are not clearly explained and already covered by official statistics" and hire a "statistical editor", as a number of international newspapers have done, with the task of overseeing the evaluation of the quality of the data published. This would enable a clear qualitative leap in terms of information disseminated to citizens.[78]

2) Institutional collegiality

It should be clear that what needs to be done will require the active involvement, support and goodwill of many actors. This principle reflects the concerns and arguments made in support of the "data philanthropy" movement.[79] Its most basic implication is to develop partnerships and systematic information and knowledge sharing between institutions.

The movement should be initiated from official statistical institutions, in light of their prerogatives and mandate, and because private corporations and even academic institutions are unlikely to be the driving forces.

3) Strategic incrementality

This principle refers to the need to start small—including pilot projects, the predominant model with Big Data so far—but with a clear vision and strategic roadmap. The nature and extent of the future "blending" of the official statistics community and the Big Data community, and of their respective tools and techniques, is anyone's guess, but what is certain is that official statistics as an industry will not change overnight—we are likely facing an evolution rather than a revolution.

The time horizon is not the next two months—the next quarterly report—but the next five to ten years, the next generation. Changing the overall timeframe does change short-term decisions and priorities. Building local awareness, buy-in and capacities, and building, devising and setting standards and norms for the long term are absolutely essential.

4) Context specificity

In general, the ways in which official statistical systems will engage with Big Data should be in great part determined by local aspirations and conditions—some societies may wish to develop and monitor certain indicators sooner than others; some may have more immediate access to Big Data sets and actors, etc.

Of course, some standards and norms—as well as "good practices" with respect to data sharing arrangements and protocols, and new analytical, visualization and diffusion methods—may be near universal, although they may differ according to country and time-specific considerations (for example, telecom companies may be forced to release individual level data in times of acute crisis[80]).

Likewise, the post-2015 framework is likely to highlight non-traditional issues such as social cohesion, well-being and personal security across the board, areas where Big Data will be invaluable both as a source of data and of skills. International and regional statistical actors will play a central role in helping shape these common processes and indicators.

Each society will have to find its rhythm and path, but the message is that it is high time to start moving.

5 Concluding Remarks: Sketching the Contours of an Action Plan

Since its creation over two centuries ago, official statistics as a public industry has moved far beyond its initial mandate of reporting on activities of or for the state ("statistics" is derived from the German *Statistik*—science of the state) to report on the state of societies in light of new available data and demands. This should not change in the Big Data era.

The data revolution should "go beyond the geeks and the bean-counters",[81] or rather, beyond geek-ing and bean-counting. It is fundamentally about empowering people. And in this endeavor, official statistical systems and their personnel have an instrumental role to play in order to make Hal Varian's famous prediction from 2009 – that "statistician will be the next sexiest job of the next decade"[82] – a reality.

Not only do the *Fundamental Principles of Official Statistics* not preclude the use of new data, but democratic considerations command that members of the official statistics community quickly and forcefully engage with the Big Data community to become one of its key actors, leveraging both their political legitimacy and technical expertise to fulfill their dual purpose.

This is easier said than done, and many NSOs and other members of the official statistics community have scarce resources to allocate and hard choices to make. National development strategies are a good natural entry point for this engagement, and the ongoing debates around the "data revolution", and other recent and current projects about Big Data and official statistics certainly provide a momentum that needs to be harnessed.

Big Data will not help remedy any "statistical tragedy" without dedicated systems and capacity. These need to be built incrementally, on the basis of a solid understanding of what Big Data is about, why the official statistical community has a political obligation to engage with it, and of the context in and ways through which it should be deployed and reported on. That is also a political imperative that the donor community needs to meet.

In light of the above, a general action plan may be as follows, inspired by the model of agile software development.[83] Each phase should be considered as part of a loop rather than as steps in a linear process:

1) Awareness raising and advocacy—to which this paper hopes to contribute—to set the whole enterprise on the right path.
2) Partnership building, between public and private sector organizations, with dedicated investments.
3) Piloting and testing, fully integrated into official statistical systems and institutions' long-term modernization strategy, not as side projects.
4) Evaluation and adjustments.

Annexes

Annex 1. Application of Big Data to societal questions: two taxonomies and examples

	Applications	Explanation	Examples	Comments
UN Global Pulse report Taxonomy **(Letouzé, 2012)**	Early warning	Early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis	Predictive policing, 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 behaviors can be observed and modeled. 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'.
	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.
	Real-time feedback	The ability to monitor a population in real time makes it possible to understand where policies and programs are failing, and make the necessary adjustments	Private corporations already use Big Data analytics for development, which includes analyzing 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.
Alternative taxonomy **(Letouzé et al., 2013)**	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.
	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.
	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.

Annex 2. How big is Big Data?

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

Annex 3. Predicting socioeconomic levels through cell-phone data and machine learning (published on scidev.net)


"these digital traces, recorded by every telecom operator, are "call Detail Records", or CDRs, metadata that look like that
CALLER ID
CALLER LOCATION
RECIPIENT ID
RECIPIENT LOCATION
CALL TIME
CALL DURATION
X36872 97489
2°24'22" 35°49'58"
A8C492 TC7364G
3°38'49" 31°12'22"
2014.04.01 ET 17 22
01.12.27
TELECOM OPERATOR DATA CENTER
"and these CDRs will show differences in calling patterns between different areas..."
POOR area
RICH area
their "digital signatures" will differ
maybe people in richer areas will stop calling sooner at night?
and using computers and algorithms, researchers try to find the strongest correlations
between CDRs and survey data to build a "predictive model."
CDRs
PREDICTIVE MODEL?
SURVEY DATA
"that can then take CDRs from a later time or different area"
"and turn them into estimates of socioeconomic levels without a survey!
CDRs
PREDICTIVE MODEL
SURVEY DATA?
MANU.

End Notes

[1] HLP report, 2013; UN News Centre (29 August 2014), "Data revolution advisory group named by UN Secretary-General", UN News Centre, www.un.org/apps/news/story.asp?NewsID=48594#.VCG3Ced8GeQ; Data Revolution Group website www.undatarevolution.org/.
[2] http://blogs.worldbank.org/africacan/africa-s-statistical-tragedyhttp://www.cgdev.org/sites/default/files/direct-dividend-payments.pdf, http://www.huffingtonpost.com/marcelo-giugale/fix-africas-statistics_b_2324936.html and http://mortenjerven.com/statistical-tragedy-in-africa-take-2/
[3] Including but not limited to Daas et al, 2013, Scannapieco et al, 2013, UNECE HLG, 2012, Chang, 2012, Horrigan 2013.
[4] UN Data Revolution (2014), *A World That Counts: Mobilising the Data Revolution for Sustainable Development*, United Nations Independent Expert Advisory Group on a Data Revolution for Sustainable Development, www.undatarevolution.org/report/.
[5] We use NSOs and NSIs (National Statistical Institutes) interchangeably.
[6] Giovannini, Enrico. "Statistics 2.0: The next level", 15-16 December 2010, 10th National Conference of Statistics, Rome, http://en.istat.it/istat/eventi/2010/10_conferenza_statistica/Relazione_pres_10conf.pdf
[7] Shrestha, M. And M. Marini (2013), "Quarterly GDP revisions in G-20 countries: Evidence from the 2008 Financial crisis", *IMF Working Paper*, WP.13/60, IMF (International Monetary Fund), www.imf.org/external/pubs/ft/wp/2013/wp1360.pdf; http://www.iariw2012.com/wp-content/uploads/2012/08/IARIW_revisions_Moulton_Fixler.pdf; Cohen-Setton, J. and E. Letouzé (26 March 2013), "Blogs review: Big Data, aggregates and individuals", Breugel, www.bruegel.org/nc/blog/detail/article/1059-blogs-review-big-data-aggregates-and-individuals/.
[8] Wohlsen, M. (29 October 2013), "The next big thing you missed: Big-Data men rewrite government's tired economic models, *Wired*, http://www.wired.com/2013/10/next-big-thing-economic-data/.
[9] Cohen-Setton, J. and E. Letouzé (26 March 2013), "Blogs review: Big Data, aggregates and individuals", Breugel, www.bruegel.org/nc/blog/detail/article/1059-blogs-review-big-data-aggregates-and-individuals/.
[10] Big Data, for better or worse: 90% of world's data generated over last two years", SINTEF, 22 May 2013, Science Daily, http://www.sciencedaily.com/releases/2013/05/130522085217.htm
[11] Kpodo, K. (5 November 2010), "Data overhaul shows Ghana's economy 60 pct bigger", Reuters, www.reuters.com/article/2010/11/05/ozatp-ghana-economy-idAFJOE6A40BG20101105.
[12] Magnowski, D. (7 April 2014), "Nigerian economy overtakes South Africa's on rebased GDP", Bloomberg Business, www.bloomberg.com/news/2014-04-06/nigerian-economy-overtakes-south-africa-s-on-rebased-gdp.html.
[13] UN Statistics Division (2015), "Census dates for all countries", 2010 World Population and Housing Census Programme, UN Statistics Division, http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm.
[14] Letouzé, E. (11 June 2013), "Could Big Data provide alternative measures of poverty and welfare?", Development Progress Blog, www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare; Fengler, W. (6 February 2013), "Data and development: 'The second half of the chess board'", The World Bank, http://blogs.worldbank.org/africacan/big-data-and-development-the-second-half-of-the-chess-board, notably.
[15] Chandy, L., N. Ledlie and V. Penciakova (24 April 2013), "The final countdown: Prospects for ending extreme poverty by 2030", Brookings, www.brookings.edu/research/interactives/2013/ending-extreme-poverty.
[16] Scott, C. (2005), *Measuring Up to the Measurement Problem: The Role of Statistics in Evidence-Based Policy-Making*, Paris21, http://www.paris21.org/sites/default/files/MUMPS-full.pdf
[17] Hellerstein, J. (19 November 2008), "The commoditization of massive data analysis", O'Reilly Radar, http://strata.oreilly.com/2008/11/the-commoditization-of-massive.html.
[18] Scannapieco, Monica, Antonino Virgillito and Diego Zardetto. (2013), "Placing Big Data in Official Statistics: A Big Challenge?", 2013 New Techniques and Technologies for Statistics conference, Brussels.
[19] Horrigon, M.W. (1 January 2013), "Big Data: A perspective from the BLS", AMStatNews, http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/.
[20] ScidevNet, "Big Data for development: Facts and figures", (April 2014) Spotlight on Data for development, http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html

[21] http://www.sciencedaily.com/releases/2013/05/130522085217.htm
[22] For an overview see: Cohen-Setton, J. and E. Letouzé (26 March 2013), "Blogs review: Big Data, aggregates and individuals", Breugel, www.bruegel.org/nc/blog/detail/article/1059-blogs-review-big-data-aggregates-and-individuals/.
[23]UN Statistical Commission (2013), "Big Data for policy, development and official statistics", Friday Seminar on Emerging Issues, 22 February 2013, United Nations, New York, http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html.
[24] Grumbach, S. and S. Frénot (7 January 2013), "Les données, puissance du futur", Le Monde Idées, www.lemonde.fr/idees/article/2013/01/07/les-donnees-puissance-du-futur_1813693_3232.html.
[25] Devarajan, 2013 ; Guigale, 2013
[26] Guigale, 2013 ; Fengler, 2013
[27] Henderson, J.V., A. Storeygard and D.N. Weil, "Measuring economic growth from outer space", NBER Working Paper Series, 15199, National Bureau of Economic Research, Cambridge, MA, www.nber.org/papers/w15199.pdf.
[28] Varian, H.R. and C. Hyunyoung (2 April 2009), "Predicting the present with Google Trends", Google Research Blog, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302.
[29] The Billion Prices Project @ MIT http://bpp.mit.edu
[30] Economist (18 July 2012), "Falling BRICS: Measuring economic sentiment", *The Economist*, www.economist.com/blogs/graphicdetail/2012/07/measuring-economic-sentiment.
[31] UNECE (United Nations Economic Commission for Europe) (2013), "Big Data (and official statistics)", Working Paper, Meeting on the Management of Statistical Information Systems (MSIS 2013), 23-25 April 2013, Paris, France and Bangkok, Thailand, www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf.
[32] For more examples and full references see Letouzé, E. (11 June 2013), "Could Big Data provide alternative measures of poverty and welfare?", Development Progress Blog, www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare.
[33] Letouzé, E. "Big Data for Development: Challenges and Opportunities", (2012), UN Global Pulse, http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf
[34] Letouzé, E. , P. Meier and P. Vinck (2013), "Big data for conflict prevention: New oil and old fires", in F. Mancini (ed.), New Technology and the Prevention of Violence and Conflict, International Peace Institute, New York, http://www.ipinst.org/wp-content/uploads/publications/ipi_epub_new_technology_final.pdf; for an overview & comparison of both taxonomies see Letouzé (15 April 2014), "Big data for development: Facts and figures", SciDev, www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html
[35] Harris, D. (11 September 2014), "Google has open sourced a tool for inferring cause from correlations", Gigaom, https://gigaom.com/2014/09/11/google-has-open-sourced-a-tool-for-inferring-cause-from-correlations/; National Academy of Science (2015), Drawing Causal Inference from Big Data, Arthur M. Sackler Colloquia, 26-27 March 2015, Washington, DC, www.nasonline.org/programs/sackler-colloquia/upcoming-colloquia/Big-data.html; Galasso, E. (22 April 2014), "Big data, causal inference and 'good data mining'", World Bank blog, http://blogs.worldbank.org/impactevaluations/big-data-causal-inference-and-good-data-mining citing Sandy.
[36] Mike Horrigan: "Given rising costs of data collection and tighter resources, there is a need to consider the creative use of Big Data, including corporate data." [when and where?]
[37] Horrigon, M.W. (1 January 2013), "Big Data: A perspective from the BLS", AMStatNews, http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/.
[38] UNECE (2013), "Big Data (and official statistics)", Working Paper, Meeting on the Management of Statistical Information Systems (MSIS 2013), 23-25 April 2013, Paris, France and Bangkok, Thailand, www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf.
[39] UNECE (2013) "Production of official statistics by using Big Data", Working Paper, Meeting on the Management of Statistical Information Systems (MSIS 2013), 23-25 April 2013, Paris, France and Bangkok, Thailand, www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_3_Korea.pdf
[40] United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG), (November 2014), UN Data Revolution Report: "A World that Counts", http://www.undatarevolution.org/report/
[41] Willis-Nuñez, Fiona, "Draft HLG Project Proposal on Big Data", (2013), UNECE Statistics, http://www1.unece.org/stat/platform/display/bigdata/Draft+HLG+Project+Proposal+on+Big+Data

[42] Results summarised here: https://docs.google.com/forms/d/1lcclo0bty-oUm8AyN_zl-0l4KhwwnSWQB4VY4Wfuso4/viewanalytics#start=publishanalytics
[43] Ploug, N. (undated) "New forms of data for official statistics", Statistics Denmark, www.statistics.gov.hk/wsc/STS027-P1-S.pdf.
[44] Crawford, Kate, "Think Again: Big Data", (10 May 2013), Foreign Policy, http://foreignpolicy.com/2013/05/10/think-again-big-data/
[45] UN Statistics Division (2015), "Census dates for all countries", 2010 World Population and Housing Census Programme, UN Statistics Division, http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm
[46] « L'Insee suit attentivement le big data. Pour autant, tous les articles sur le sujet évoquent des indicateurs très avancés qui ne présentent, pour l'instant, que peu d'intérêt, car ne permettant que de gagner quelques jours par rapport à la sortie d'un indicateur statistique conjoncturel et rien d'autre n'apparaît aujourd'hui opérationnel en la matière »
[47] Wesolowski, A. et al. (2013), "The impact of biases in mobile phone ownership on estimates of human mobility", Journal of the Royal Society, Interface, Vol. 10(81), doi: 10.1098/rsif.2012.0986, www.ncbi.nlm.nih.gov/pubmed/23389897.
[48] Zagheni, Emilio and Ingmar Weber (2012), "You are where you E-mail: Using E-mail Data to Estimate International Migration Rates", WebSci 2012. http://www.demogr.mpg.de/publications/files/4598_1340471188_1_Zagheni&Weber_Websci12.pdf
[49] For more examples and full references see Letouzé, E. (11 June 2013), "Could Big Data provide alternative measures of poverty and welfare?", Development Progress Blog, www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare.
[50] boyd, d. and K. Crawford (2011), "Six provocations for Big Data", A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September, 2011, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.
[51] Pentland, 2012; Letouzé et al., 2013; Letouzé, 2014.
[52] Letouzé et al., 2013.
[53] Pentland, A. (30 August 2012), "Reinventing society in the wake of big data", Edge.org, www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data.
[54]For an overview see Cohen-Setton, J. and E. Letouzé (18 March 2013), "Blogs review: GDP, welfare and the rise of data-driven activities", Breugel, www.bruegel.org/nc/blog/detail/article/1044-blogs-review-gdp-welfare-and-the-rise-of-data-driven-activities/.
[55] King, Gary, "Big Data is Not About the Data!", (14 May 2013), Presentation to New England Artificial Intelligence Meetup, http://gking.harvard.edu/presentations/big-data-not-about-data-1
[56] Toyama, Kentaro, "Can technology end poverty?", (1 November 2010), Boston Review, http://bostonreview.net/forum/can-technology-end-poverty
[57] IBM (undated), "what is analytics?", IBM website, www.ibm.com/analytics/us/en/what-is-smarter-analytics/big-data-analysis.html.
[58] Soto, V. et al. (2011), "Prediction of socioeconomic levels using cell phone records", User Modeling, Adaption and Personalization, Vol. 6787, pp. 377-388, www.vanessafriasmartinez.org/uploads/umap2011.pdf.
[59] Benbouzid, B. (2014), "Quantifier pour transformer", *la Vie des Idées*, www.laviedesidees.fr/IMG/pdf/20141003_desrosieres.pdf and Ted Porter
[60] United Nations (2014), *Fundamental Principles of Official Statistics*, UN General Assembly Resolution 68/261, United Nations, http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf.
[61] Porter, T.M. (undated), "Research interests", UCLA Department of History website, www.history.ucla.edu/people/faculty/faculty-1/faculty-1?lid=384.
[62] Enrico Giovannini, 2010
[63] Enrico Giovannini, 2010
[64] Burrel, 2012 ; UN Global Pulse, 2012
[65] Enrico Giovannini, 2010
[66] Surveillance activities by governments are not discussed here.
[67] Enrico Giovannini, 2010.
[68] Lawhorn, C. (14 March 2013), "Census rejects city's appeal of 2010 population totals; new Census numbers for Douglas County show growth slowed in 2012", *Lawrence Journal-World*, http://www2.ljworld.com/weblogs/town_talk/2013/mar/14/census-rejects-citys-appeal-of-2010-popu/

[69] FAO (Food and Agriculture Organization) (2006), "Food security", *Policy Brief*, Issue 2, ftp://ftp.fao.org/es/ESA/policybriefs/pb_02.pdf.
[70] FAO (Food and Agriculture Organization) (2006), "Food security", *Policy Brief*, Issue 2, ftp://ftp.fao.org/es/ESA/policybriefs/pb_02.pdf.
[71] "Alimentary security" would be more appropriate than "food security", in that the concepts of "food" in food security and in food access, for instance, differ as data and knowledge do.
[72] "To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data." United Nations (2014), *Fundamental Principles of Official Statistics*, UN General Assembly Resolution 68/261, United Nations, http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf.
[73] Horrigan, Michael W., "Big Data: A perspective from the ALS", (1 January 2013), AMSTAT News, http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/
[74] Enrico Giovannini, 2010.
[75] Enrico Giovannini, 2010.
[76] Orange D4D (Data for development) website: www.d4d.orange.com/home
[77] See Letouzé and Vinck for the D4D, forthcoming.
[78] Enrico Giovannini, 2010.
[79] Kirkpatrick, 2011; Meier, 2012
[80] See Letouzé and Vinck, 2014
[81] Glennie, J. (3 October 2013), "A development data revolution needs to go beyond the geeks and bean-counters", Poverty Matters blog, Guardian, www.theguardian.com/global-development/poverty-matters/2013/oct/03/data-revolution-development-policy.
[82] Flowing Data (25 Feburary 2009), "Google's Chief Economist Hal Varian on statistics and data", Flowingdata, http://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/.
[83] See UN Global Pulse (2012) and Letouzé, Meier and Vinck (2013)