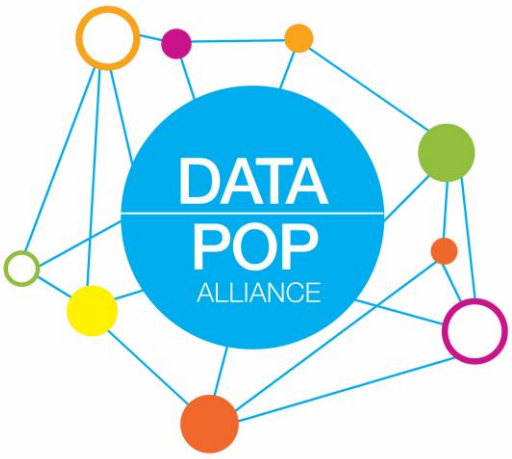


DATA-POP ALLIANCE WHITE PAPERS SERIES

Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach

Emmanuel Letouzé
Johannes Jütting

Draft for discussion

This version: December 17th, 2014

About this document

This is the inaugural piece in Data-Pop Alliance's 'White Papers Series' developed in collaboration with our partners. Data-Pop Alliance is the first research, capacity building and policy 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.

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, one of its co-authors, presented views discussed in more details in this paper.

The second White Paper in the series on "*The Politics and Ethics of Cell-Phone Data Analytics*", written with partial support from the World Bank, is being finalized and is available for comments on www.datapopalliance.com/comments, as is this paper. Another White Paper on "*Data Literacy in the Data Rich Era*" funded by Internews will be available in January.

About the authors

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.

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.

Comments on this document can be provided online at http://www.datapopalliance.org/comments. Note that this document contains hyperlinks that can be accessed when read online as a PDF or on a web browser.

Please do not quote this version.

Table of Contents

Abstract

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

While the current discussions mainly focus on if and how Big Data can contribute to produce faster, cheaper, more frequent and different development indicators for better policies, this paper takes a different starting point. It does so by stressing 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 moved by their own incentives, and should stir serious strategic rethinking and rewiring on the part of the official statistical community.

It contents that the emergence of this new ecosystem provides both an historical opportunity, and a political, 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.

Introduction

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

This question has led to several publications, forums and pilot projects since about the early months of 2013, the vast majority in and about OECD countries.[3] In more recent months and weeks, 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.[4]

The predominant answer to the question above may be summarized as follows: *"Big Data may provide faster, cheaper, and more granular data to complement, certainly not replace, official statistics, to design, implement and implement better policies and programs, but many challenges remain that will require adapted responses"*. This is pretty much the state of affairs.

But the absence of much specifics as to how—and also exactly why—that may or should happen, how that may change for the better the state of the world we live in, has done little to convert those who frame Big Data as mere hype, and argue that scare resources would be best invested elsewhere—for instance making sure that all countries conduct regular censuses and surveys and collect 'basic statistics'.

There is evidently partial truth in these perspectives and criticisms, and good points have been made in most contributions—with notably a growing and welcome recognition that better data won't magically or mechanistically lead to better policies and better outcomes—as if bad data was the primary cause of persistent poverty, rising inequality, environmental degradation, and social oppression in the first instance.

This paper argues that while the current debates have led to an increased understanding and recognition of the opportunities and challenges ahead, *the fundamental question to be asked and answered is not whether and how Big Data as data may or may not facilitate the production of official statistics*. Rather, taking a systems 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 Bata community in order to reap the potential sizeable benefits for human development while avoiding its losing relevance and putting the societies it is mandated to serve at risks. Overall, we think 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. One by revisiting the traditional terms and framing of the 'Big Data' and official statistics' question; two by suggesting a number of principles and steps that official statisticians and statistical systems—including but not only National Statistical Offices[5]—may follow as they start engaging and continue evolving 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 with greater depth and breath; Section 3 offers suggestions about preconditions, principles and policies for action.

1. Roots and state of the 'Big Data and Official Statistics' question

Interest in the 'Big Data and official statistics' question has roots its from two main developments—termed the statistical disillusion and the rise of Big Data.

1.1. The statistical disillusion

The first factor can be termed the statistical 'disillusion'—rather than "tragedy". Although, as mentioned in the introduction, 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 compilation systems of quarterly 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 titled *"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 old 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 on MDG monitoring in general and its poverty indicator in particular.[10] For instance, Ghana's GDP grew over 60% overnight after the rebasing exercise of 2010[11]; Nigeria by over 75% after its own, in April 2014.[12] The root cause is having GDP data based on old consumption and production patterns.

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 growing fast, 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: http://www.huffingtonpost.com/marcelo-giugale/fix-africas-statistics_b_2324936.html

A handful of countries—not all poor, but all with a history of violence—have not conducted a population census in decades, such that their populations and any per capita data can only be estimated, even as 'official', precise, and yet most likely inaccurate figures are provided.[13]

Poor poverty data, in particular, turns its monitoring and forecasting into exercises in 'guesstimation' where sub-continental averages have to be used as crude proxies for county-level data.[14]

A country like Kenya has not produced poverty data in almost a decade.[15] Unemployment figures are close to meaningless in most developing countries. The list goes on.

In addition to the structural challenge of rebasing GDP, well-known and mutually reinforcing determinants include low levels of financial resources, poor human and technical capacities—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)—, lack of trust and dialogue, inadequate institutional organizations, weak or ill-motivated political will, etc.[16]

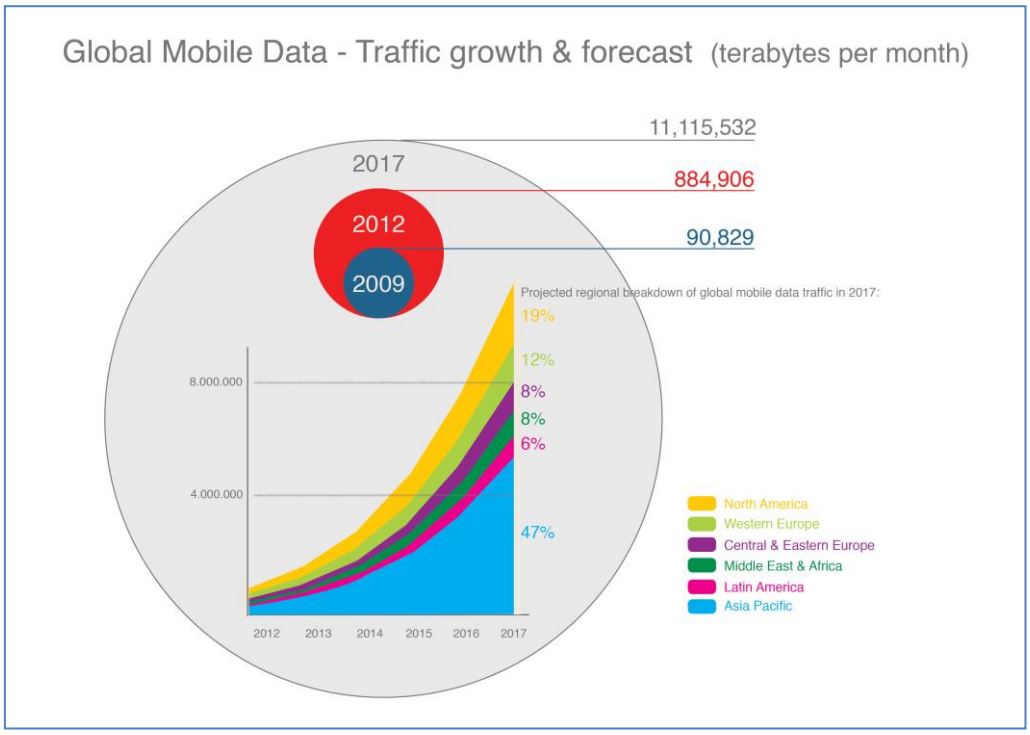
Importantly, as the 'pilot in a plane' analogy described above suggests, much of the focus has been on the data deficit and the subsequent 'measurement' challenge, and the need to get 'better data' in the hands of well-meaning policymakers to devise better policies.

1.2. The rise of Big Data

The second development is what has initially been 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*". Mike Horrigan at the US BLS defined Big Data as "*nonsampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference.*"[18] The absence of a single agreed-upon definition of 'Big Data' is not very problematic; much less than the unfortunate continuing focus on Big Data as 'just' big datasets—or any of the 3 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 further 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 and picked up about humans populations behaviors and beliefs: each year since 2012, over 1.2 zettabytes of data have been produced — 10^{21} bytes, enough to fill 80 billion 16GB iPhones that would circle the earth more than 100 times.[19] The volume of these data is growing fast. And just like a human population with a sudden outburst of fertility gets larger and younger, the proportion of digital data produced recently (i.e. baby data) is growing—it has been said many times that up to 90 per cent of the world's data was created over just two years, although the assertion is almost impossible to source or corroborate.

Data 'inflation'



As noted above, a lot has been said and written about the applications and (much less so) *implications* of Big Data in the public policy and social science arenas[20], including, increasingly, for official statistics.[21]

One perspective posits 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*"[22]. As mentioned, others have also argued that Big Data may partially fix the "*statistical tragedy*" [23] a few developing countries—for example by providing near real-time, fine-grained 'Big Data-based' poverty figures. Big Data, according to some, may even preside over a "*leap frogging*" of poor countries' statistical systems.[24]—the same way some have bypassed fixed telephone lines to go straight to cell-phone, and seem to be heading into the smart-phone and tablet era without going through the PC stage. And indeed, a key source of 'Big Data' is phone phone data known as Call Detail records (CDRs), most of which from cell-phones, recorded by telecom operators. It is estimated that over 80% of Internet traffic will soon transit through hand-held devices, which will lead to further creation of geolocalized, time stamped data.

Called 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

The broad notion that new digital data and tools hold the potential to increase both temporal and geographical granularities while reducing costs of data collection can be traced back to two seminal 2009 papers, the 'GDP and night light emission' paper[25] and the 'Now casting via Google queries' papers.[26] More recent and much-cited examples and applications have given further impetus to this argument—e.g. efforts to track inflation online such as those spearheaded by the BPP project[27], estimate and predict changes in GDP in near real-time[28], monitor traffic[29], etc. Other somewhat less widely-cited papers have used email data and CDRs for instance to study migration patterns, malaria spread, etc. in development contexts.[30] The field of sentiment analysis from social media data is also opening additional avenues to develop alternative measures of welfare.

A first question is how Big Data can be used for development. Two main taxonomies of applications of Big Data have been proposed in subsequent papers. UN Global Pulse's[31] proposed a three-tier taxonomy of uses: "real-time awareness", "early warning" and "real-time feedback"; another distinguished its "descriptive" (e.g. maps), "predictive" (i.e. either proxying or forecasting) and "prescriptive" (the realm of causal inference) functions.[32] Whereas interest in the latter is poised to grow[33], most applications have relied on the first two, and perhaps most visibly on the second—predictions, understood as either better 'forecasting' what may happen next, or 'proxying' (or inferring) some variable of interest via another.

1.3. Pilots and controversies in official statistics

Aware of these changing conditions, as well as interested in exploring most potentially cost-saving avenues,[34] several NSOs from OECD countries have started experiencing with Big Data—as well as leveraging large administrative datasets, which we don't consider as part of Big Data. This has typically been done through pilot projects, initially and still in many cases, for example within the US BLS[35], Statistics Netherlands (with traffic loop and social media data, notably)[36], Statistics Korea[37], Statistics Ireland, Statistics New Zealand, and others. Other initiatives are in the making—most in OECD countries, although the Paris21 initiative and partners such as ODI are planning to increase their support to developing partners and that the UN Data Revolution report has called for increasing support to NSOs in the area of new data.[38]

At the global level, a (non-representative) survey designed by three authors of this paper and administered by the Paris21 initiative in 2013 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.[39]

The latter statistics may find its root in a mix of scepticism echoing concerns over data quality, reliability as well as ethical considerations and fear of "*losing relevance*"[40] within the official statistics community. Big Data is at times described as simply bad data, or just "*the tech world's one-size-fits-all (…) answer to solving the world's most intractable problems*". [41]As such, it would be ill suited to countries that may have more pressing statistical concerns to attend, [42] 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*". [43]

Little has yet been done by NSOs in methodological terms. Hardly any work has been done in the area of sample bias correction for instance; counter examples are a study led by Harvard faculty to estimate the impact of biases in mobile-phone ownership on CDR-based estimates of human mobility[44] and another academic paper proposing a correction factor for email-based estimates of international migration flows[45] — interestingly both largely unmentioned in the current 'official' debate despite their seminal nature. [46]

Whatever their take is, there is also an acute and understandable sense of frustration among official statisticians pressured to open up their data to private sector actors that are concomitantly increasingly locking what they and many consider, 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 views have nonetheless coalesced towards a lose consensus according to which these vast volumes of data may hold the potential, provide an opportunity, to supplement ("but not replace", one is pressed to add immediately) official statistics, by either helping produce faster inflation or more accurate GDP figures, for instance, or coming with alternative measures.

But we argue that the state of affairs isn't satisfactory for two main reasons. One 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; two because strategic and operational considerations and suggestions are often lacking. We turn to both aspects successively.

2. Revisiting the terms and framing of the question

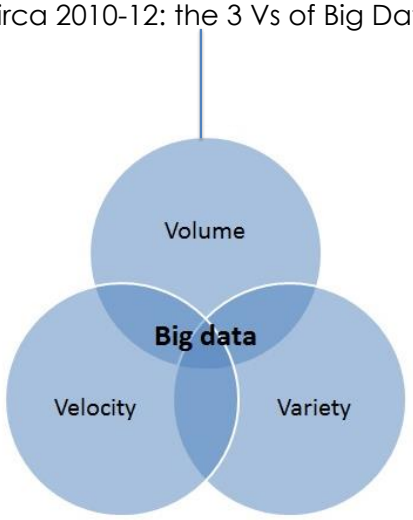
2.1. Big Data isn't big data: from the 3 Vs to the 3 Cs

Let us start by unpacking and clarifying the terms of the question to start revisiting it on firm ground. We start by Big Data. What exactly is 'Big Data'? In this paper, 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. Our starting point is to define Big Data not just as big data characterized by the 3 Vs of Volume, Velocity and Variety, as has long been the case, but through 3 Cs.

The first C of Big Data refers to *Crumbs*[47]—to Big Data as new kinds of passively-generated individual and networked "*traces of human actions picked up by digital devices*" [48]. These "*digital breadcrumbs*" have the potential to paint a picture of some aspects of the social world with unprecedented levels of details and shades. [49]. Their fundamental revolutionary nature is qualitative.

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 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 wrong assumptions and unnecessary controversies—the notion that Big Data is or isn't about providing an "*automated 30,0000-foot view of the world*"[50].

Circa 2010-12: the 3 Vs of Big Data



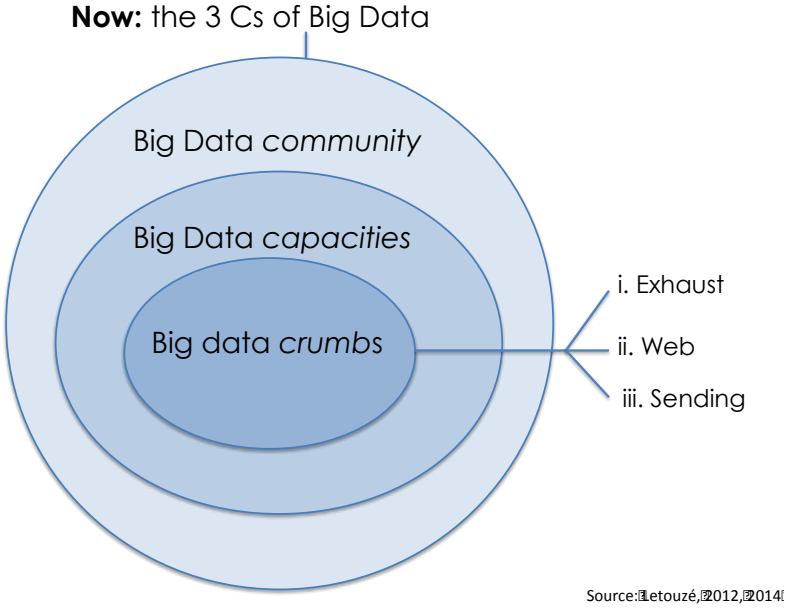
From a systems perspective, these data are not just an exhaust of digital and connected societies that can be used to report on human ecosystems: some of them—such as social media data—have an endogenous effect by shaping preferences and fueling aspirations. They also produce a surplus that goes largely unrecorded in GDP—adding to the traditional measurement challenge.[51] 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 Gary King, "*Big Data is not about the data*"[52]. It is 'about' the intent and capacities[53] to yield and convey what is routinely and vaguely referred to as 'insights'[54] (which appears safer than to use 'information') from these qualitatively new kinds of data. Part of it is advanced storage and computing capacities; another is advanced quantitative and computer science methods and tools—primarily statistical machine-learning techniques, algorithms, etc.

An example is Telefonica researchers' attempt to 'predict' socioeconomic levels (SELs) in a "*major city in Latin America*" (Mexico-City) by matching CDRs and official survey data using supervised machine learning to unveil differential digital signatures of different SELs and create a model applicable to other regions and/or later points in time.[55] This example illustrates the focus on the predictive use of Big Data, here understood as proxying, with yet little in the way of moving towards causal inference. Yet another part of Big Data capacities is visualization techniques and tools that allow presenting complex trends and patterns in appealing and often customizable ways.

The third C of Big Data is for *Community*, or Communities: Big Data must also be considered as referring to 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 R software being an example.

Others, in the private sector or intelligence communities, function in a highly controlled and secretive manner, for obvious commercial and political reasons. It is also the case that at the minute most of the 'Big Data' are held by private corporations—telecom companies, financial institutions, etc.—and only a handful in the public domain, most of them unstructured and hard to work with.



A key point that cannot be stressed enough is this: Big Data is not just big data. It is qualitatively new kinds of data about people's behaviours and beliefs, new kinds of tools, and new kinds of actors. For any discussion of the applications and implications of Big Data for official statistics (and development) to be meaningful, Big Data *must* be approach and conceived as an ecosystem—a complex system—made of these three complex ecosystems—the data, the tools and methods, and the actors.

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

2.2. The dual nature and purpose of official statistics

Most discussions about Big Data and official statistics do provide some definitions of Big Data—too often focused on Big Data being big data, as just noted—but overlook much consideration for 'official statistics' as a concept. This is probably because the term official statistics is so pervasive in our personal and professional lives that we take its (or is it their?) nature and role as givens. And yet, unpacking what we mean by and expect from 'official statistics provides useful insights on the larger question at hand.

In our view, and as previous authors have underlined, including recently[56], 'official statistics' has a dual *nature* and serves a dual *purpose*.

The dual nature of official statistics as an object of analysis is:

1) official statistics as measurement tools, i.e. data, produced by official bodies and systems and
2) these official bodies and systems (which is not restricted to their center, NSOs).

When one wonders or worries about the 'future of official statistics' in the age of Big Data, it is clear that both are referred to. 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: 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 statistic. It is also worth noting that official statistics can, according to these Fundamental Principles, draw on "*all types of sources*". [57]

Fundamental Principles of Official Statistics (revised, 213)

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

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.

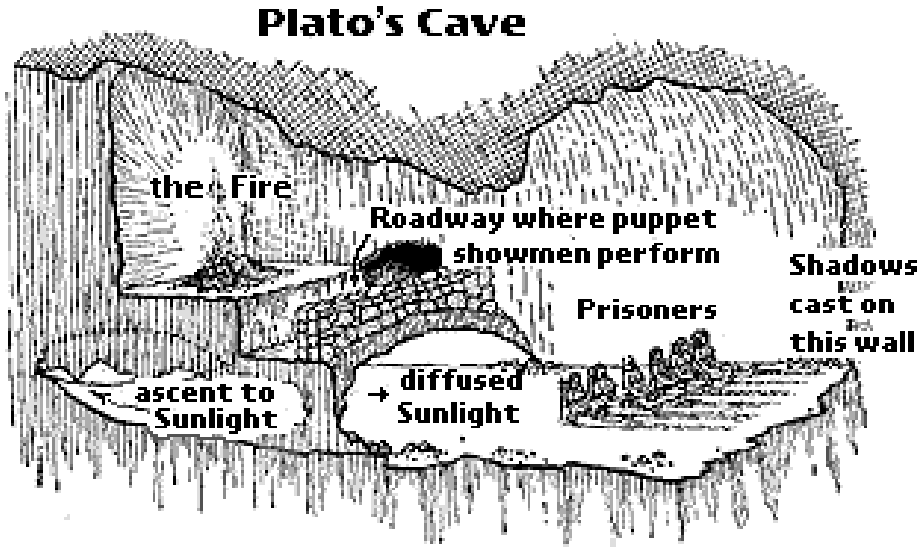
Principle 8. Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

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: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf

But official and real are of course not synonymous terms. 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*".[58] As such, to paraphrase Plato, they are, by necessity—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.



On the other hand, the latter—official statistics as bodies and systems—are not or should not be solely defined by the former—official statistics as data. In other words, their defining feature, their essential role, it *not* to produce official statistics as data—it is not even to produce information. What is it?

Official statistics not only has a dual nature, but it also serves two main functions:

1) Its 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"[59] is to provide society with "*knowledge of itself, on which to base its own choices and evaluate the effects of political decisions.*"[60].
2) The second function of official statistics is provides a deliberative space where what is worth measuring, how it is measured, and for which purpose it is measured is freely and openly debated—to act as "*a debated public institution*".

These points stresses how official statistics isn't merely an—often poor—mirror of the world that benevolent and enlightened policymakers use to craft policies, but fundamentally a public industry and space that exist 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 put by Enrico Giovannini during a recent 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, it remains the alpha of omega of economic policymaking and ranking.

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

2.3. Why engaging 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 pose non-trivial conceptual, theoretical, and technological challenges as well as deep anthropological, ethnographic and ethical ones; these are discussed in papers and contributions that are seldom referred to in much level of details if at all in talks about the 'opportunity' provided by Big Data for official statistics. [61]

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

Last they point to the distinction and competition between creating insights and information within modern techno-infused societies—and their difference with enhancing a society's "knowledge of itself".

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 see that the role they have been mandated with be fulfilled: that societies have 'knowledge of themselves' that reflects, and a public space to discuss, what they care to know and mean to impact.

This directs attention to the crux of the issue: its essentially political dimension (as the term 'data revolution' suggests). A former President 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 reflects and serves 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*"[62] is uncontroversial, and overall unproblematic. But it is clear that Big Data alters and accelerates these dynamics: it is presiding over the fast 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.[63]

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 what 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 behaviours*". Another recent example showed that the results of 2010 census results population in the UK are being disputed on the basis of sewage data.[64] 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 devise and release near real-time alternative measures of wellbeing, leveraging the tremedous potential of Big Data.

Last, an obvious risk on non-engagement is the creation of a new digital divide that may result, in less than a decade, in a situation where all data on sub-Saharan African economies and societies may be derived remotely, with little to no inputs from local official institutions 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 there seems to be little in the way of addressing them in the Big Data community.

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.

3. Towards a new conceptual and operational approach

3.1. 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 fist argue that societies must thrive to be knowledge secure—the same way they thrive to be food secure. Next, 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*"[65]. These conditions are translated into the 4 pillars of food availability, food 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 make the argument clearer, in what follows, the only major changes to the (italicized) description of the four pillars of the official FAO food security framework[66] is the substitution of 'data' or 'knowledge' for 'food' as appropriate[67], plus a few minor edits that have been left apparent. We also discuss some of the implications of these preconditions.

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 2nd Fundamental Principle of Official Statistics[68], 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 nonsampling errors that can produce biases in our estimates and threats to valid inference.*" [69]

2) ***Data access***—*i.e. the access by individuals to adequate resources (entitlements) for acquiring appropriate data for a nutritious diet 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.*"[70] 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 diet, clean water, sanitation and health care individual and collective processing to reach a state of nutritional well-being knowledge where all physiological information needs are met. This brings out the importance of non-data inputs in knowledge security.*

This critical point notably 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 (or 'dataracy') and add that of graphic literacy (or 'graphicacy'). [71]

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—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 CDRs in recent months—apart perhaps from the counterexample of the first and second Orange D4D challenges[72]--although the 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 4 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 IEAG group.

3.2. Proposed operational principles to create a deliberative space

We propose 4 operational principles to facilitate the creation of a deliberative space for societies to debate issues of measurement and policy objectives.

1) A first principle is ***shared responsibility.***

Official statisticians and statistical systems are not solely 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 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 do have a responsibility, as the keeper of an increasing share of data. The idea of compelling private companies to systematically share their raw data with official institutions and researchers is 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. [73]

Private companies and their research arms—such as Microsoft or IBM research, SAS, R&D of telecom companies would also work on project could also directly partner with official statistical systems. Researchers and policymakers also bear significant responsibility—in particular, ad hoc requests and pressures to get data should be avoided to instead support efforts to find standardized, ethical and stable, data sharing tools and protocols with private companies.

The media do too, who could "*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 has occurred on a number of international newspapers, with the task of overseeing the evaluation of the quality of the data published, would enable a clear qualitative leap in terms of information disseminated to citizens.*" [74]

2) A corollary principle is ***institutional collegiality.***

It should be clear that what needs to be done will require the active involvement, support and good will of many actors. This principle reflects concerns and arguments made in support of the 'data philanthropy' movement.[75] 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 the fact that private corporations and even academic institutions are unlikely to be the driving forces.

3) Yet another key principle is ***strategic incrementality*.**

This refers to the need to start small—including through pilot projects, as has been 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—in that sense we are more likely facing an evolution than a revolution.

The time horizon is not the next 2 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, building plus devising and setting standards and norms for the long term are absolutely essential ingredients and objectives.

1) The fourth and last principle is ***context specificity*.**

Of course, some standard and norms as well as 'good practices' with respect to data sharing arrangements and protocols as well as 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[76]).

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 sources of data and skills. International and regional statistical actors will play a central role in helping shape these common processes and indicators.

But in general, the steps and ways through 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. Each society will have to find its rhythm and path, but the message is that it is high time to start moving.

Concluding remarks: sketching the contours of an action plan

Since its creation over 2 centuries ago, official statistics as a public industry has long 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*"[77]*;* or rather, beyond geeking and bean-counting. It's fundamentally about empowering people. And in this endeavor, official statistical systems and their personel have an instrumental role to play; to make Hal Varian's famous prediction from 2009 that "*statistician will be the next sexiest job of the next decade*"[78] reality—on par and working with that of data scientists.

Not only the Fundamental Principles of Official Statistics do 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.

It is easier said than done, and many NSOs and other members of the official statistics community do have too scarce resources to allocate and hard choices to make. National Development Strategies are a good natural entry points, 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 seized.

Big Data will not partially help remedy any "statistical tragedy" unless dedicated systems and capacities are incrementally built on the basis of a solid understanding of what Big Data is about, why official statistical community has a political obligation to engage and of the context in and ways through which is it to be deployed and reported on. That is also a political imperative that the donor community needs to rise up to.

In light of the above, a general action plan may be follow, inspired by the model of agile software development,[79] considering each phases as part of a loop rather than as steps in a linear process. These steps would be:

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 in official statistical systems and institutions long term modernization strategy, not as side projects;
4) Evaluation and adjustments.

Points (1)-(2)-(3) need to be the subject of additional research and thinking.

Annexes

Applications of Big Data to societal questions: 2 taxonomies and examples

	Applications	Explanation	Examples	Comments
UN Global Pulse report Taxonomy[1] **(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 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.
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 proxing 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.

How Big are 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 ofa 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.

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 97487
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?
CDRs
and using computers and algorithms, researchers try to find the strongest correlations
between CDRs and survey data to build a "predictive model."
SURVEY DATA
PREDICTIVE MODEL?
CDRs
"that can then take CDRs from a later time or different area"
"and turn them into estimates of socioeconomic levels without a survey!"
SURVEY DATA?
PREDICTIVE MODEL
MANU.

Bibliographical endnotes

[1] HLP report, 2013 and http://www.un.org/apps/news/story.asp?NewsID=48594#.VCG3Ced8GeQ and http://www.undatarevolution.org/
[2] Devarajan, 2013, Guigale, 2013
[3] Including but not limited to Daas et al, 2013, Scannapieco et al, 2013, UNECE HLG, 2012, Chang, 2012, Horrigan 2013.
[4] http://www.undatarevolution.org/report/
[5] We use NSOs and NSIs (National Statistical Institutes) interchangeably.
[6] http://en.istat.it/istat/eventi/2010/10_conferenza_statistica/Relazione_pres_10conf.pdf
[7] http://www.imf.org/external/pubs/ft/wp/2013/wp1360.pdf and http://www.iariw2012.com/wp-content/uploads/2012/08/IARIW_revisions_Moulton_Fixler.pdf and http://www.bruegel.org/nc/blog/detail/article/1059-blogs-review-big-data-aggregates-and-individuals/
[8] http://www.wired.com/business/2013/10/next-big-thing-economic-data
[9] http://www.bruegel.org/nc/blog/detail/article/1059-blogs-review-big-data-aggregates-and-individuals/
[10] http://unstats.un.org/unsd/statcom/doc13/2013-21-Indicators-E.pdf and recent WB data.
[11] http://www.reuters.com/article/2010/11/05/ozatp-ghana-economy-idAFJOE6A40BG20101105
[12] http://www.bloomberg.com/news/2014-04-06/nigerian-economy-overtakes-south-africa-s-on-rebased-gdp.html
[13] http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm.
[14]http://www.brookings.edu/research/interactives/2013/ending-extreme-poverty
[15]http://www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare and http://blogs.worldbank.org/africacan/big-data-and-development-the-second-half-of-the-chess-board, notably
[16] http://www.paris21.org/sites/default/files/MUMPS-full.pdf
[17] http://strata.oreilly.com/2008/11/the-commoditization-of-massive.html
[18] http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/
[19] Scidev
[20] For an overview http://www.bruegel.org/nc/blog/detail/article/1059-blogs-review-big-data-aggregates-and-individuals/
[21]http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html
[22] http://www.lemonde.fr/idees/article/2013/01/07/les-donnees-puissance-du-futur_1813693_3232.html.
[23] Devarajan, 2013, Guigale, 2013
[24] Guigale, 2013, Fengler, 2013
[25] http://www.nber.org/papers/w15199.pdf?new_window=1
[26] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302
[27] http://bpp.mit.edu
[28]http://www.economist.com/blogs/graphicdetail/2012/07/measuring-economic-sentiment
[29]http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf
[30] For more examples and full references see http://www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare
[31] Letouzé, 2012.
[32] Letouzé, Meier and Vinck, 2013. http://www.ipinst.org/media/pdf/publications/ipi_epub_new_technology_final.pdf; fopr an overview & comparison of both taxonomies http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html
[33] https://gigaom.com/2014/09/11/google-has-open-sourced-a-tool-for-inferring-cause-from-correlations/ http://www.nasonline.org/programs/sackler-colloquia/upcoming-colloquia/Big-data.html and http://blogs.worldbank.org/impactevaluations/big-data-causal-inference-and-good-data-mining citing Sandy
[34] 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."
[35] http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/
[36]http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf
[37]http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_3_Korea.pdf
[38] http://www1.unece.org/stat/platform/display/Collection/Draft+HLG+Project+Proposal+on+Big+Data
[39]https://docs.google.com/forms/d/1lcclo0bty-oUm8AyN_zl-0I4KhwwnSWQB4VY4Wfuso4/viewanalytics#start=publishanalytics
[40] http://www.statistics.gov.hk/wsc/STS027-P1-S.pdf
[41] Crawford, 2013.

[42] http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm
[43] 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
[44] http://www.ncbi.nlm.nih.gov/pubmed/23389897
[45] Zagheni and Weber
[46] For more examples and full references see http://www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare
[47] Pentland, 2012, Letouzé et al 2013, Letouzé 2014
[48] Letouzé et al. 2013
[49] http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data
[50]http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1926431_code1210838.pdf?abstractid=1926431&mirid=1
[51]For an overview http://www.bruegel.org/nc/blog/detail/article/1044-blogs-review-gdp-welfare-and-the-rise-of-data-driven-activities/
[52] (King, 2013).
[53] Toyama
[54] http://www.ibm.com/analytics/us/en/what-is-smarter-analytics/big-data-analysis.html
[55] http://www.vanessafriasmartinez.org/uploads/umap2011.pdf
[56] http://www.laviedesidees.fr/IMG/pdf/20141003_desrosieres.pdf and Ted Porter.
[57] Fundamental Principles of Official Statistics
[58] http://www.history.ucla.edu/people/faculty/faculty-1/faculty-1?lid=384
[59] Enrico Giovannini, 2010
[60] Enrico Giovannini, 2010
[61] (Burrel, 2012, UN Global Pulse, 2012),
[62] Enrico Giovannini, 2010
[63] Surveillance activities by governements are not discussed here.
[64] http://www2.ljworld.com/weblogs/town_talk/2013/mar/14/census-rejects-citys-appeal-of-2010-popu/
[65] ftp://ftp.fao.org/es/ESA/policybriefs/pb_02.pdf
[66] ftp://ftp.fao.org/es/ESA/policybriefs/pb_02.pdf
[67] 'alimentary security' would be more appropriate than "food security", such that the concepts of food in food security and in food access for instance differ as data and knowledge do)
[68] "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."
[69] Horrigan
[70] Giovannini
[71] Giovannini
[72] http://www.d4d.orange.com/home
[73] See Letouzé and Vinck for the D4D, forthcoming.
[74] Giovannini
[75] (Kirkpatrick, 2011; Meier, 2012).
[76] See Letouzé and Vinck, 2014
[77] http://www.theguardian.com/global-development/poverty-matters/2013/oct/03/data-revolution-development-policy
[78] http://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/
[79] See UN Global Pulse (2012?) and Letouzé, Meier and Vinck (2013)