

OECD Digital Economy Papers No. 245

The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis

Christian Reimsbach-Kounatze

https://dx.doi.org/10.1787/5js7t9wqzvg8-en

OECD

The OECD Digital Economy Papers series (http://oe.cd/digital-economy-papers) covers a broad range of ICT-related issues, both technical and analytical in nature, and makes selected studies available to a wider readership. It includes policy reports, which are officially declassified by an OECD committee, and occasionally working papers, which are meant to share early knowledge and elicit feedback.

This document is a working paper. OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the authors. The release of this working paper has been authorised by Andrew Wyckoff, OECD Director for Science, Technology and Innovation.

The Directorate for Science, Technology and Innovation (STI) also publishes the OECD Science, Technology and Industry Working Paper series (http://oe.cd/sti-working-papers), which covers a broad range of themes related to OECD's research and policy work on knowledge-based sources of economic and social growth and, more specifically, on the translation of science and technology into innovation.

Comments on STI's Working Papers are welcomed, and may be sent to the Directorate for Science, Technology and Innovation, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France; e-mail: sti.contact@oecd.org.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

© OECD 2015

Applications for permission to reproduce or translate all or part of this material should be made to:
OECD Publications, 2 rue André-Pascal, 75775 Paris, Cedex 16, France
e-mail: rights@oecd.org

FOREWORD

This paper describes the potential of the proliferation of new sources of large volumes of data, sometimes also referred to as "big data", for informing policy-making in several areas. It also outlines the challenges that the proliferation of data raises for the production of official statistics and for statistical policies.

The paper builds on and extends a report on "Official Statistics in the Era of Ubiquitous Connectivity and Pervasive Technologies" [DSTI/ICCP/IIS(2010)15], that was presented to the OECD Working Party on Measurement and Analysis of the Digital Economy (WPMADE; formerly Working Party on Indicators for the Information Society, WPIIS) in June 2010 as well as the report on "The Proliferation of Data and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis" [STD/CSTAT(2012)2], that was presented to the OECD Committee on Statistics and Statistical Policy (CSSP; formerly Committee for Statistics, CSTAT) in June 2012. A version of the paper was also presented to the OECD Committee on Digital Economy Policy (CDEP; formerly the Committee of Information, Computer and Communications Policy, ICCP) in December 2012 [under code DSTI/ICCP(2012)11].

The paper also benefited from discussions at the 2012 OECD Technology Foresight Forum on "Harnessing Data as a New Source of Growth: Big Data Analytics and Policies" held on 22 October 2012 at OECD Headquarters in Paris, France (see http://oe.cd/tff2012). Major insights originated from the OECD (2013) report "Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues" (see http://oe.cd/bigdata1), and the OECD (2014) interim synthesis report on "Data-driven Innovation for Growth and Well-being" (http://oe.cd/bigdata2). This paper contributes to Phase II of the OECD project on *New Sources of Growth: Knowledge-Based Capital*, in particular its pillar on data-driven innovation (DDI, see http://oe.cd/bigdata).

TABLE OF CONTENTS

THE PROLIFERATION OF "BIG DATA" AND IMPLICATIONS FOR OFFICIAL STATISTICS AND STATISTICAL AGENCIES: A PRELIMINARY ANALYSIS 6

1. The proliferation of data 6
2. Early efforts to exploit "big data" and non-traditional on-line sources for statistics 9
   - ICTs and the Internet 10
   - Prices 13
   - Employment and skills 14
   - Output 17
   - Demographics 20
   - Development and natural risk management 21
3. The limits of "big data" 22
   - Poor quality data 22
   - Inappropriate use of data and analytics 24
   - Changing data environment 24
4. Implications for statistical agencies and statistical policy 25
   - NSOs as a trusted 3rd party 26
   - NSOs as a clearing house 27
   - NSOs as a user of statistics from non-traditional sources 27
   - NSOs as an issuer of analytical best practices 30
5. Conclusion 31

BIBLIOGRAPHY 33

Figures

Figure 1. Monthly global IP data traffic, 2005-17 7
Figure 2. Number of unique botnet command and control (C&C) machines by country, 2006-11 11
Figure 3. Google Insights for Search 12
Figure 4. Daily online price index, United States, 2008-2014 13
Figure 5. Unemployment prediction for Germany, January 2004 – May 2009 15
Figure 6. Unemployment vs. number of ads in the United States, May 2005 – May 2009 16
Figure 7. Net job starters by top industries by volume, 1998-2008 17
Figure 8. Forecasting monthly growth in automobile sales in Chile, January 2006 – July 2010 18
Figure 9. Monthly prediction of growth in automobile sales in Chile, January 2006 – July 2010 19
Figure 10. SWIFT Index predictive power: OECD GDP quarterly growth, Q1 2004 – Q1 2011 20
Figure 11. Flu-infection rates in the United States, January 2011 – December 2012 25
Figure 12. Share of data specialist in selected OECD countries, 2011-13 29

Boxes

Box 1. The difficulty of defining "big data" beyond volume, velocity and variety of data 8
Box 2. Microdata 10
Box 3. What is Google Insights for Search? 12
Box 4. The factors affecting data quality 23
Box 5. Copyrights and data analytics 30

THE PROLIFERATION OF "BIG DATA" AND IMPLICATIONS FOR OFFICIAL STATISTICS AND STATISTICAL AGENCIES: A PRELIMINARY ANALYSIS

1. The proliferation of data

A confluence of significant technological, social and economic trends including the growth of smart devices and infrastructure, the growing ubiquity of wired and wireless broadband access, the appeal of social networking sites and the widespread adoption of IT systems are resulting in the generation of huge streams of data (OECD, 2013a). The consultancy, IDC (2012), has estimated that the global volume of digital data will multiply by a factor of 40 by the end of this decade after having exceeded 1 000 exabytes in 2010 (an exabyte is a billion gigabytes). This is partly driven by the fact that nearly all media including books, photos, audio/video is now digitized, up from only 25% in 2000 (see MGI, 2011).

With the additional deployment of radio-frequency identification (RFID) in combination with the deployment of (real world) sensors interconnected through the Internet of Things (IoT), off-line social and economic activities are generating a new tidal wave of data as the physical world is increasingly transformed into processable and quantifiable data. This process, which is sometimes referred to as "datafication"[1], will reach its tipping point once machine-to-machine communication[2] bypasses human data communication, signalling a new phase of data-driven innovation, that today is only at its infancy even in the most advanced economies (OECD, 2014b).

More than 30 million interconnected sensors are estimated to be deployed worldwide today in areas such as security, health care, the environment, transport systems or energy control systems, and their numbers are growing by around 30% a year (MGI, 2011). The growing ubiquity of sensors is reflected in the widespread of smartphones, which accounts for roughly 15% of the 7 billion mobile subscriptions worldwide, each of these devices capable of collecting and transmitting geo-location data related to traffic, the environment or even health care (see ITU, 2012; Cisco, 2013). These new data sources are enabling an increasing number of innovative services that have barely been possible before. Based on its Floating Mobile Data (FMD) technology, for example, mobile telecommunication services firm Orange is able to collect and use anonymized mobile phone traffic data to determine instantaneous speeds and traffic density at a given point of the road network, and deduce for example the travel time or the formation of traffic jams. The anonymized mobile phone traffic data is then sold to third parties including government agencies and private companies.[3]

---

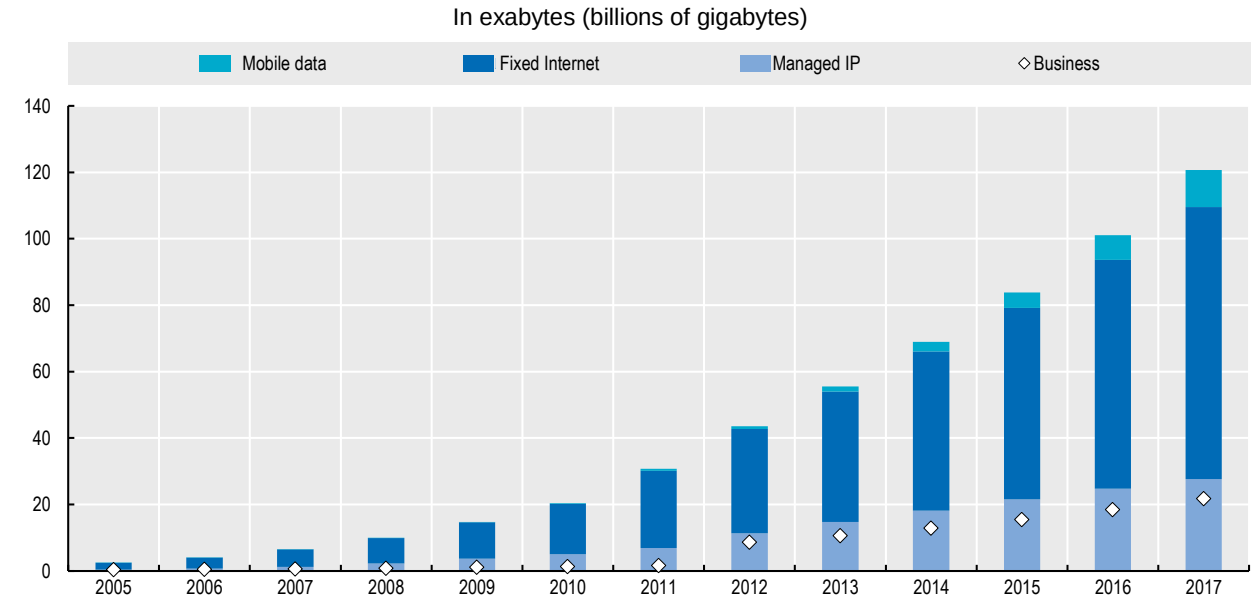
1 "Datafication" is a portmanteau for "data" and "quantification" (Hey, 2004; Bertolucci, 2013; Mayer-Schönberger and Cukier, 2013). As Mayer-Schönberger and Cukier (2013) explain: "To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analyzed".

2 Machine-to-Machine communication (M2M) is a key characteristic of the Internet of Things and describes the process where data is communicated to other machines including a central computer.

3 In January 2012, for example, Orange signed an agreement with Mediamobile, a leading provider of traffic information services in Europe, to use FMD data for its traffic information service V-Trafic (see Orange, 2012).

Given that mobile phone penetration (subscriptions per 100 inhabitants) exceeds 100% in most OECD countries and that wireless broadband penetration is at nearly 70% in the OECD area, this source of data will grow further significantly as smartphones become the prevalent personal device around the world. Already, these multi-purpose mobile devices generated more than 1.5 exabytes (billions of gigabytes) of data every month in 2013 worldwide. This is only the beginning as other smart devices proliferate, including *smart meters* that collect real-time data on energy consumption (see OECD, 2012a) or *smart automobiles* equipped with sensors to monitor and transmit the state of the car's components as well as of the environment in which the car is moving (see OECD, 2012b).[4] Overall, Cisco (2013) estimates that the amount of data traffic generated by all mobile devices will almost double every year to reach more than 11 exabytes (billions of gigabytes) by 2017 (Figure 1). Some of this volume of data represents a new resource for innovators that can build new data-driven goods and services.

Figure 1. Monthly global IP data traffic, 2005-17



Source: OECD (2014a), Measuring the Digital Economy: A New Perspective based on Cisco (2013).

The phenomenon of collecting, compiling, linking and analysing very large flows of data in real-time requires powerful, new analytical techniques and data sharing models to handle the size and complexity of processing the data. The availability of new techniques and the associated shift in how operations within an organisation are organised, signal a shift towards a data-driven or data-centric socio-economic model that is commonly discussed under the umbrella term "big data" (see Box 1 on definitions). In such a data-driven world, data is a core asset that proves a huge new resource for innovation, new industries and applications and competitive advantage (OECD, 2013a; 2014b). While harnessing this new asset is non-trivial, the continued rapid decline in the cost of analytics, including computing power and data storage, as well as the continued expansion of broadband makes it increasingly within reach. Storage costs, for example, have decreased to the point at which data can generally be kept for long periods of time if not indefinitely.

[4] The number of mobile wireless devices connected to the Internet across the globe is estimated to reach 50 billion by 2020 (see OECD, 2011b).

Box 1. The difficulty of defining "big data" beyond volume, velocity and variety of data

There is still no clear definition of "big data". Initially the term "big data" referred to data sets for which volume became an issue in terms of data management and processing. This is consistent with many of today's definitions such as the one suggested by Loukides (2010), who defines "big data" as data for which "the size of the data itself becomes part of the problem" or The McKinsey Global Institute (MGI, 2011), who similarly defines it as data for which the "size is beyond the ability of typical database software tools to capture, store, manage, and analyse". However the emphasis on the volume alone can be misleading, whether this is measured in gigabytes, petabytes (millions of gigabytes), or exabytes (billions of gigabytes). In some cases what is relevant is not the volume, but for example the number of readings, the way data is used and the resulting complexity. For example, managing a day's worth of data from thousands of sensors close to real time is more challenging than managing a video collection of the same size in bytes. This distinction is captured by the three Vs definition[5] of big data, which points to its three main characteristics including:

- The **volume** of the data as covered by most definitions today (see Loukides, 2010 and MGI, 2011, which are cited above; but also McGuire et al., 2012);
- The **variety** of the data, which refers to mostly unstructured data sets from sources as diverse as web logs, social media, mobile communications, sensors and financial transactions. Variety also goes hand in hand with the capability to link these diverse data sets; and
- The **velocity** or the speed at which data is generated, accessed, processed and analysed. Real-time monitoring and real-time "nowcasting" are often listed here as benefits that go along the velocity of "big data".

The problem still with the 3Vs and similar definitions is that they are in continuous flux, as they describe technical properties which depend on the evolving state of the art in data storage and processing. Furthermore, these definitions misleadingly suggest that it is all about data. While this is true in the case of volume, what is behind variety and velocity is primarily data analytics; that is the capacity to process and analyse unstructured diverse data in (close to) real-time. Furthermore the term "big data" does not suggest how the data is used, what type of innovation it can enable, and also how it relates to other concepts such as "open data", "linked data", "data mashups", and so on. These are the reasons why the OECD KBC2: DATA project does not primarily focus on the concept "big data", but rather focusses on "data-driven innovation", which is based on the *use of data and analytics to innovate* for growth and well-being.

Source: OECD (2013; 2014b)

Alongside these developments, we are witnessing the restructuring of industries, governments and academia to take advantage of this new phenomenon as well as the emergence of new data-driven organisational models that are very successful and may portend a broader structural change by harnessing and exploiting huge streams of data in real-time (OECD, 2013a; 2014b).[6] This phenomenon is multi-faceted:

- In business, "big data" techniques can be used in a wide number of operations ranging from optimising the value chain and manufacturing production to more efficiently using labour and

5 This definition originated from the META Group (now part of Gartner) in 2001 (see Laney, 2001).

6 The strong interest of big companies in playing a leading role in "big data" is manifested in the growing number and volume of merger and acquisitions (M&A) deals (see OECD, 2014b). IBM was the most active acquirer with, for example, the acquisition of Netezza, a data warehousing and analytics company, for USD 1.7 billion in 2010.

exploiting the interface with consumers. It allows firms to expose variability that would have otherwise remained hidden, undertake controlled experiments and segment and tailor their products at a low cost (Brynjolfsson *et al*, 2011). As a consequence, the impact could be very large and not restricted to a few industries as has previously been the case. The massive growth of data as an input to production suggests that data is becoming a new factor of production akin to labour and capital (see The Economist, 2010), prompting some to speculate that "the global economy is on the cusp of a new wave of productivity growth enabled by big data" (MGI, 2011). While perhaps an exaggeration, mastering "big data" could impart a new competitive advantage to developed countries that typically have more advanced services, distribution channels, brand management techniques and market knowledge.

- Social interaction has been transformed by social networking sites like Facebook which has over 900 million active participants generating together e.g. 1 500 status updates every second on average (Hachman, 2012; Bullas, 2011). As of 2011, LinkedIn, another social networking site, had data on approximately 150 million professionals around the world, including information about their curriculum vitae, their job applications and shifts in employment. This represents a huge, new source of demographic and employment data.

- The health care sector sits on a growing mountain of data generated from the administration of the health system and from the emergence and diffusion of electronic health records. New data on results of diagnostic tests, medical images and the banking of biological samples are also being generated. For example, there are now vast collections of medical images. 2.5 petabytes – more than a million, billion data units – are stored away each year from mammograms in the US alone.

- Science has always been data driven, but now new instruments such as super colliders or telescopes have fundamentally changed the scale of what is being collected: the Digital Sky Survey, started in 2000, collected more data through its telescope in its first week than had been amassed in the history of astronomy (The Economist, 2010).

The search for new sources of growth and productivity gains that will boost incomes, profits and tax revenues, as well as the desire to improve the efficiency of government during this era of belt-tightening, makes the issue of how best to exploit "big data" especially relevant. With this potential comes a wide array of policy issues, most prominently the protection of privacy. New sources of data, new actors and the increasing ease of linking data on individuals and transferring it, risk undermining many of the frameworks on which privacy protection is based. But the potential implications for policy spill into many other domains including labour, competition, health, government administration and last but not least statistical policy (OECD, 2013a). The purpose of this paper is to briefly sketch out some of the implications for statistics and statistical policy.

2. Early efforts to exploit "big data" and non-traditional on-line sources for statistics

The potential of big data for statistical purposes is summarized by Steve Lohr writing for the New York Times: "It is the size of the data sets on the Web that opens new worlds of discovery. Traditionally, social sciences tracked people's behaviour by interviewing or surveying them. But the Web provides this amazing resource for observing how millions of people interact" (Lohr, 2009).

The Internet has become a potential new source for statistics in addition to micro data sets that are collected and stored by National Statistic Offices (NSOs), and also increasingly used for new analysis approaches (Box 2). The following sections briefly outline the potential of these new Internet-related data sources for generating close to real-time evidence across a number of statistical areas including (*i*) information and communication technologies (ICTs) including the Internet, (*ii*) prices, (*iii*) employment,

(*iv*) economic output, (*v*) demographics, and last but not least (*vi*) development. While at the moment, methods to mine these sources are still in their infancy and need rigorous scientific scrutiny, their rapid take-up by policy makers is a harbinger of an important trend and transition underway. Some implications of this shift are outlined in section 3. They were discussed by Skaliotis (2009, 2010), who highlighted the potential of "ubiquitous connectivity and pervasive technologies" for official statistics, and then at the OECD Working Party on Measurement and Analysis of the Digital Economy (WPMADE, former Working Party on Indicators for the Information Society) in June 2010.

Box 2. Microdata

"Microdata" refers to data that has been collected and stored at the level of individual respondents or business entities. It is sometimes seen as the "true wealth" of National Statistic Offices (NSOs) (Giovannini, 2012). For example, the 2010 Community Innovation Survey (CIS), which is part of the EU science and technology statistics, includes (non-anonymised) microdata about of business innovation activities in 22 countries. The Current Population Survey (CPS), as another example, is a statistical survey conducted by the United States Census Bureau to collect microdata on the employment situation on a monthly basis.

Microdata can be anonymised or not, but they are always disaggregated in contrast to sectoral data and macrodata, which summarize individual details to aggregates at the sectoral or regional level respectively. The aggregation typically results in information loss, which could otherwise provide important insights. In particular, microdata can be used for exploring relationships between two or more different data sets when data linkage is feasible. For example, the OECD Directorate for Science, Technology and Innovation (DSTI) has developed a Micro-Data Lab to integrate microdata sets from its OECD Patent Database (73 million record on patent applications), the Orbis© database (85 million records containing comprehensive data on companies worldwide), the Scopus© database (26 million records on scientific publications) with trademark data (6 million records) and design right data (almost 1 million records). The inter-linkage of these different data sources through the Micro-Data Lab enables STI to have deeper insights into the origins of innovative activities across the economy (see section on data linkage).

Access to microdata provides researchers with much more freedom to investigate complex interactions and perform detailed analysis. Microdata allow for example to better understand industry and macro dynamics, and can be used to better inform policy design and monitoring (de Panizza and de Prato, 2009). However, microdata can also raise issues on confidentiality and privacy given that microdata are collected at the level of individual respondents or business entities. NSOs have developed a number of strategies to give access to microdata, while protecting confidentiality and privacy. For example, Eurostat, the statistical body of the European Commission, provides access to the CIS microdata only to researchers that have successfully applied for access and the full CIS microdata sets can only be accessed in the Safe Center at Eurostat's premises in Luxembourg. DSTI, as another example, has pioneered a "distributed" approach to empirical analysis which draws on confidential micro data. While DSTI provides a common framework (including common research and policy questions, the indicators, the econometric modelling, and the software routines), researchers with access to their own country's micro data compile results that are then compared and analysed by DSTI or lead countries.

ICTs and the Internet

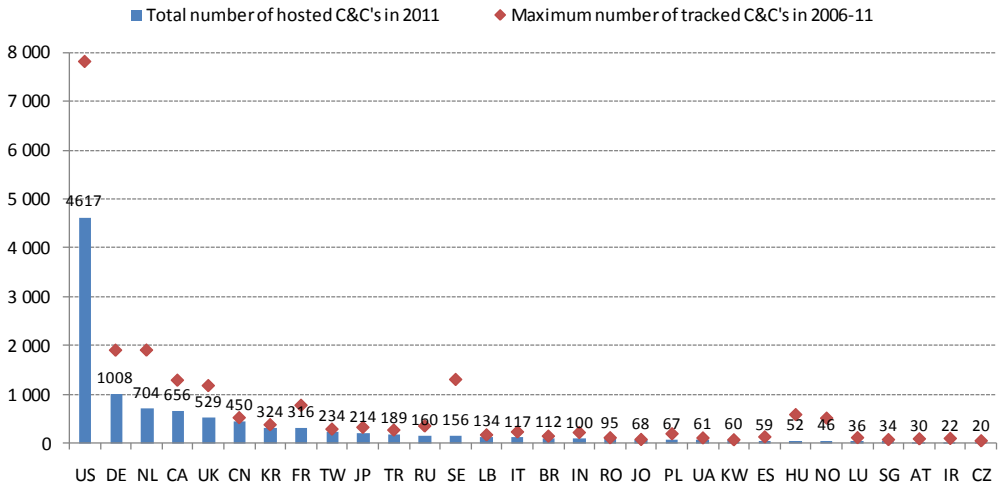
One of the earliest adoptions of "big data" for the creation of official statistics is for statistics on the adoption of ICTs and the Internet. In particular, the use of the Internet as a data source for the creation of Internet-related statistics was an important development. It was largely motivated by the challenge faced by many national statistics offices (NSOs) in developing indicators on the very fast developments on ICTs and Internet adoption. Traditional statistics, when available, generally took several years to prepare, and thus could rarely capture current developments, which however were of high interest to ICT and Internet policy makers. As individuals and organisations increasingly leave behind 'digital footprints' on the Internet, using these footprints for the creation of ICT- and Internet-related indicators seemed a promising approach that a number of NSOs started to explore.

In their report commissioned by the European Commission (EC, 2010), Dialogic Innovatie & Interactie highlighted a number of approaches for using the Internet as data source, which they classified as (i) user centric, (ii) network centric, and (iii) site centric approaches.

- **User centric approaches** measure the utilization of (ICT) systems by the user. These include for example bandwidth monitoring to measure the real network traffic of a household or browser monitoring to measure users behaviour in browsers. In those cases the placement of a physical device at a household or of software executed on the user's device is typically required. For example, the Measurement Lab (M-Lab) is an open distributed server platform on which network researchers deploy active network measurement tools. M-Lab tools send predetermined traffic flows between the user's client and the closest M-lab server. They measure flow behaviour along specific parameters, end-to-end between these points. About 200 000 consumers access one of M-Lab's 10 tools daily.[7] M-Lab data could then be used, for example, to compare actual download throughput with advertised speed data as provided by the OECD.

- **Network-centric approaches** are not targeted towards specific nodes in the internet, but instead measure the massive data communication flow between many nodes. In practice this involves the installation of a rather sophisticated measurement device on the Internet (EC, 2010). Examples in the area of cyber security include data on malware and botnets collected through e.g. *honey nets*[8], that are networks of systems that emulate a set of vulnerable IT services to attract *e.g.* malware (see OECD, 2012c). For example, data collected through honey nets by the Shadowserver Foundation, a "volunteer watchdog group of security professionals", can show trends in the total number of active bot-machines which otherwise could not be collected (Figure 2).

Figure 2. Number of unique botnet command and control (C&C) machines by country, 2006-11
Top 30 countries



Source: OECD (2012c) based on data from the Shadow Server Foundation

7 These tools include three mobile-specific tools, and two hardware-based tools. Real-time usage statistics for all M-Lab tools are visible at http://measurementlab.net/usage.

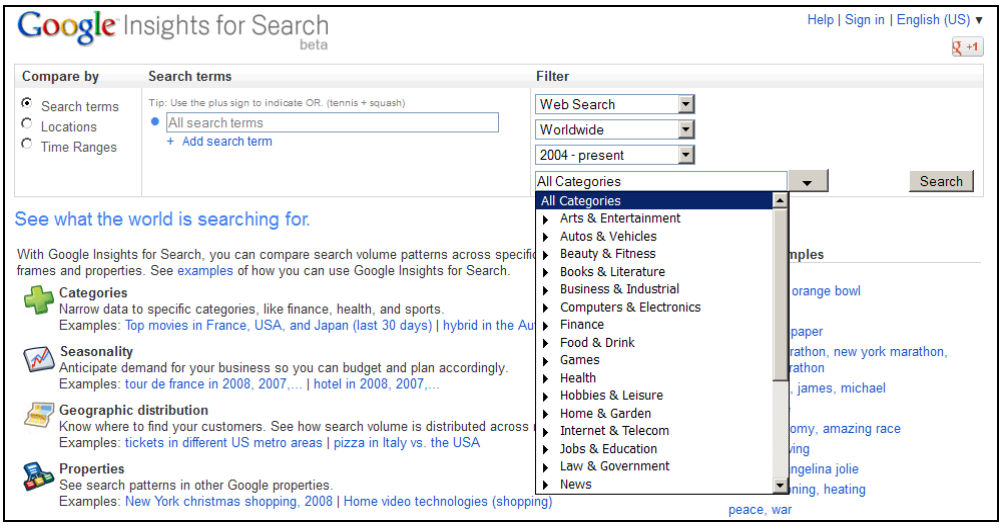
8 A *honey net* is a network of *honey pots*. A *honey pot* is a system that emulates a set of vulnerable IT services. It is usually "isolated, protected and guarded, but gives the appearance that it contains a vulnerable system of value to the attacker. It thus acts as fly-papers for malicious code and other attackers" and gives security experts the possibility to analyse attacks and malicious code used live or ex post (see http://www.cert.se/honeynet).

- **Site-centric approaches** are considered the most mature (EC, 2010). They either involve the analysis of "log files" (logs) of web servers to analyse user behaviour on web sites or the collection of well-defined data on more or less structured websites. Access to logs is sometimes simplified through application programming interfaces (APIs) or dedicated user interfaces such as Google Insights for Search (Box 3), while access to data from structured websites, are collected through site crawlers by automatically "scraping" selected web sites. Both types of site-centric approaches are at the origin of many of the methods presented below for generating close to real-time evidence in statistical domains such as prices, employment, economic output, demographics, and last but not least development.

Box 3. What is Google Insights for Search?

Google Insights for Search (www.google.com/insights/search/) is a service from Google that provides an indicator on the daily popularity of search terms people have entered into the Google search engine (Figure 3). Search queries can be compared across countries and in some cases even across regions within a time frame ranging back to 1 January 2004. The queries are "broad matched", meaning that queries such as "used automobiles" are counted in the calculation of the query for "automobile" (see Choi and Varian, 2011). Search queries are classified into a number of categories (30 at the top level and about 250 at the second level) "using a natural language classification engine". Categories include for example "Arts & Entertainment" and "Autos & Vehicles". The assignment of search terms to categories is probabilistic "in the sense that a query such as 'apple' could be partially assigned to Computers & Electronics, Food & Drink, and Entertainment" (Choi and Varian, 2011).

Figure 3. Google Insights for Search



Before results can be accessed (as csv files), they undergo two major transformations that affect the usability of Google Insight data (see Carrière-Swallow and Labbé, 2010):

1. The raw results are normalized by the total number of search queries in the geographical region of interest. By doing so, any trends from growth in the total number of Internet users of from change in the relative popularity of Google are removed from the data.
2. The normalized data are rescaled to an index with a maximum value of 100 for the most frequent queries. "This means that magnitudes are not directly comparable across series as a measure of relative popularity" (Carrière-Swallow and Labbé, 2010).

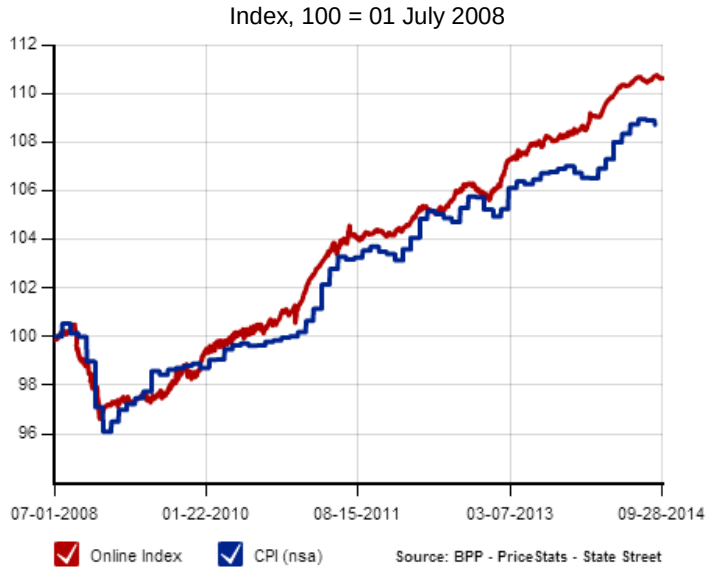
It should be noted that Google only tracks search queries which have a "meaningful volume" because of privacy considerations. Furthermore, the data are computed using a sampling method leading to variations of "few percent from day to day" (Choi and Varian, 2011).

Prices

As products are increasingly sold online, the Internet itself has become a rich source for price information. Online retailers such as Amazon.com and online market places such as eBay are publishing millions of prices on a wide range of goods and services daily. In 2011, for instance, eBay collected data on more than 100 million active users offering approximately 6 million new goods for sale every day.[9] All this price information is a rich source for the creation of real-time price statistics.

The "Billion Price Project", collects price information over the Internet for the creation of "near-time" statistics. More than half a million prices on goods (not services) per day are collected by "scraping" the web of online retailers. The resulting dataset contains daily prices on a wide array of products sold by online retailers as well as information on product descriptions, package sizes, brands, and special characteristics (e.g. "organic") among others (see BPP, 2014). This is not only five times what the US Government collects, but is also less cost intensive, given that price information is not collected manually through researchers visiting thousands of shops as it is the case for traditional inflation statistics. Price information collection via the Internet is then used to compute the *daily online price index*, which is basically an average of all individual price changes across all categories and retailers, used to estimate annual and monthly inflation. Unlike official inflation numbers, which are published monthly with a lag of weeks, the *online price index* is updated every day with a lag of just three days. In addition, the BPP has a periodicity of days as opposed to months. This allows researchers and policy makers to identify major inflation trends before they are visible in official statistics. The web page states that the index "is not designed to forecast official inflation announcements, but to provide real-time information on major inflation trends." For example, in September 2008 when Lehman collapsed, the online price index showed a decline in prices, a movement that was not picked up until November by the CPI (see Figure 4; Surowiecki, 2011). Today, several OECD countries including the United States, the United Kingdom, Germany, France as well as key Partner countries such as Brazil are working with PriceStats, the company managing the index, to contribute and use BPP statistics.

Figure 4. Daily online price index, United States, 2008-2014



Source: bpp.mit.edu

[9] The total gross merchandise volume (GMV), excluding vehicles, in eBay was more than USD 60 billion in 2011.

Employment and skills

Given the severe impact of the economic crisis on employment (the unemployment rate in OECD countries has not been below 7% since the beginning of 2009), tracking employment trends in a timely fashion is of particular interest to the general public, economists, and in particular to policy makers. However, in many OECD countries employment data are only available after several weeks at best, and concerns have been raised that these data may not reflect well the ongoing structural changes in the economy (see Askitas and Zimmermann, 2009).

The Internet provides a number of promising sources for near-time indicators to improve measurement related to employment. The three major sources include: (i) search engine data such as from Google Insights for Search for predicting unemployment trends, (ii) online advertised job vacancy data series such as the Conference Board's *Help Wanted OnLine* (HWOL) for predicting job offers, and (iii) using social networking sites such as LinkedIn, which has employment related data of more than 150 million members around the world. These different approaches are discussed briefly in the following sections.

Search related data on unemployment

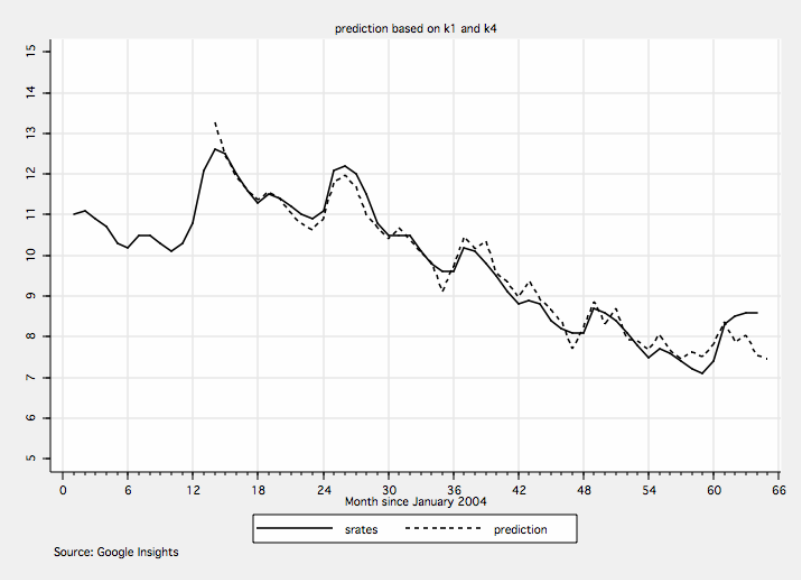
Many search engines track the keywords entered by users searching for web content. Some even provide a service for accessing statistics on the popularity of specific keywords by regions and time. This is for instance the case with Google Insights for Search (see Box 1). Where keywords are related to topics specific to the unemployed, services such as Google Insights can provide a real-time indicator for measuring and predicting unemployment trends.

Google Insights can thus provide a daily indicator on unemployment trends in each country, if the right keywords are specified. However, finding the best keywords requires some analysis, in particular because keywords may vary across countries and cultures (see Askitas and Zimmermann, 2009 for unemployment prediction in Germany; D'Amuri and Marcucci, 2010 for the United States; and Suhoy, 2010 for Israel).

In the case of Germany, Askitas and Zimmermann (2009) analysed the prediction power of the keywords "Arbeitsamt OR Arbeitsagentur" ("unemployment office or agency", in the following k1) and "Stepstone OR Jobworld OR Jobscout OR Meinestadt OR meine Stadt OR Monster Jobs OR Monster de OR Jobboerse" (the most popular job search engines in Germany, in the following k4) among other keywords.[10] The authors found that the forecast based on k1 and k4 indicated much earlier changes in trends compared to official statistics: the prediction for October to December 2008, for instance, anticipated the turning point to the rise in unemployment (see Figure 5). However, after a perfect fit through January 2009, the two trends began to diverge (after month 60). The authors suggest that this is due to changes in labour policy in Germany affecting the role of government support for short-time employment that came into effect in January 2009. As a result of this new policy, the interest in short-time work increased but was not captured in the authors' initial regression models. By replacing the keywords k1 with "Kurzarbeit" ("short-time work") the differences between the forecast and the reality once again disappeared.

[10] Google Insights supports queries for disjunctions of keywords.

Figure 5. Unemployment prediction for Germany, January 2004 – May 2009
Percentage



Source: Askitas and Zimmermann (2009).

Online job vacancy data

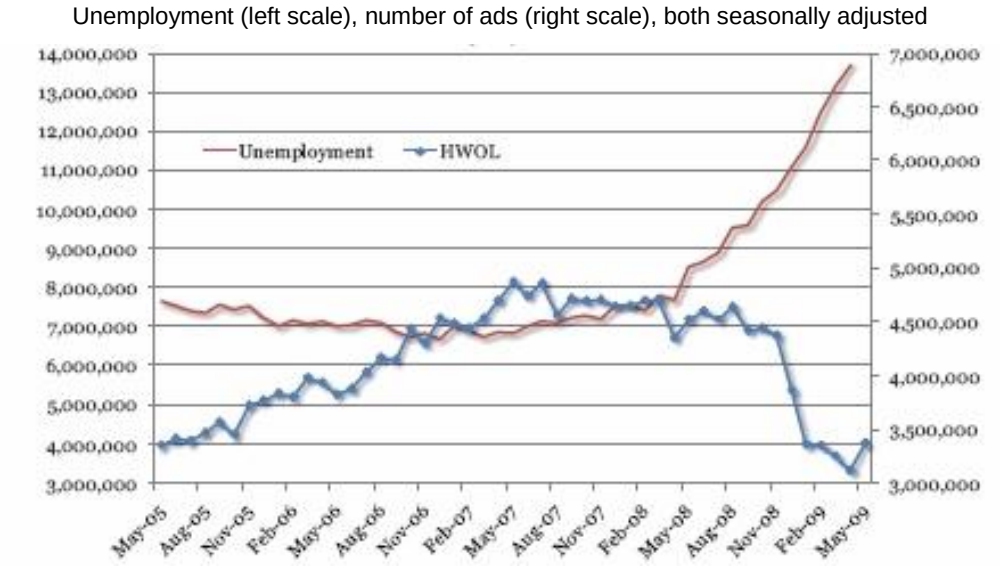
Online platforms used for posting job vacancies (*i.e.* job boards) are another promising source for real-time data on employment, namely on job offers. Experts estimate that 70% or more of US job announcements are posted on the Internet, with the share rising to 95% for jobs outside of retail, food service, building maintenance and construction (Vollman, 2010). Not surprisingly**,** the share of individuals in OECD countries using the Internet for job research is significant and increasing (from 14% in 2007 to 17% in 2010) (OECD, 2011). Given these trends, the importance of these source data can be expected to increase in the future.

A number of programs are collecting job vacancy data per region in order to measure the current level of job offers. For example, the Australian Department of Education, Employment and Workplace Relations (DEEWR) has created the Internet Vacancy Index (IVI) based on vacancies newly lodged on four online recruitment websites used in Australia: SEEK, CareerOne, MyCareer and Australian JobSearch. The IVI is used as complement to the newspaper-based Skilled Vacancy Index (SVI) also created by the DEEWR.

In the United States, the Conference Board, a business organisation and private think tank, has developed the Help Wanted OnLine (HWOL) data series to measure job offers advertised online at the national, regional, state and metropolitan area levels at a detailed (6-digit) occupational level. In total, data of more than 1 200 job boards in the United States are included, covering both online newspaper ads and internet job board ads.[11] Data are provided on a monthly basis and made comparable in timing and geographic detail to the Bureau of Labor Statistics (BLS) monthly unemployment numbers (see Figure 6).

[11] The HWOL data series cover all major sources for online advertised vacancies as posted directly on internet job boards or through newspaper online ads. At present, ads on corporate web sites for their own jobs are excluded from coverage. However, given that a number of job boards scrape these corporate websites, these ads may also appear in the HWOL data count.

Figure 6. Unemployment vs. number of ads in the United States, May 2005 – May 2009



Source: The Conference Board, BLS.

Besides the possibility to automatically count the number of job offers in an economy, data sources such as the HWOL data series also allow to assess skills and education requirements of job offers. This not only allows to identify skills most in demand in real-time, but also to assess the level of skills shortages and skill mismatches where data on education programs and vocational training are available and can be compared.

As is the case for other big data sources used for statistics, data sources such as HWOL are essentially a universe count and are not subject to the typical sampling and non-response error components associated with most statistical surveys. However, there are other (non-sampling) error sources that these data sources are subject to, such as *i*) population under-coverage due to missing portions of the targeted population (*e.g.* a large Internet job board) and *ii*) over-coverage due to the inability to fully eliminate duplicate ads from collected data. Additional potential sources of non-sampling error could include occupational and/or geographic coding errors which would affect the proper classification of individual ads.[12]

This series has led to a wide-range of analyses, ranging from sub-Federal studies (US States) to studies of specific "in-demand" occupations, to using key words to track the emergence of new sectors and the location of innovative activity ("clusters"), such as the development of "Apps" for smart devices like the iPad.[13]

Profiles in social networking sites

Social networking sites provide another rich source for near-time indicators on employment, as individuals are increasingly providing details about their private and professional life. According to the OECD (2011a), nearly 50% of OECD Internet users were active social network users in 2010. Of interest for the creation of employment related statistics are past and current employment status as provided in

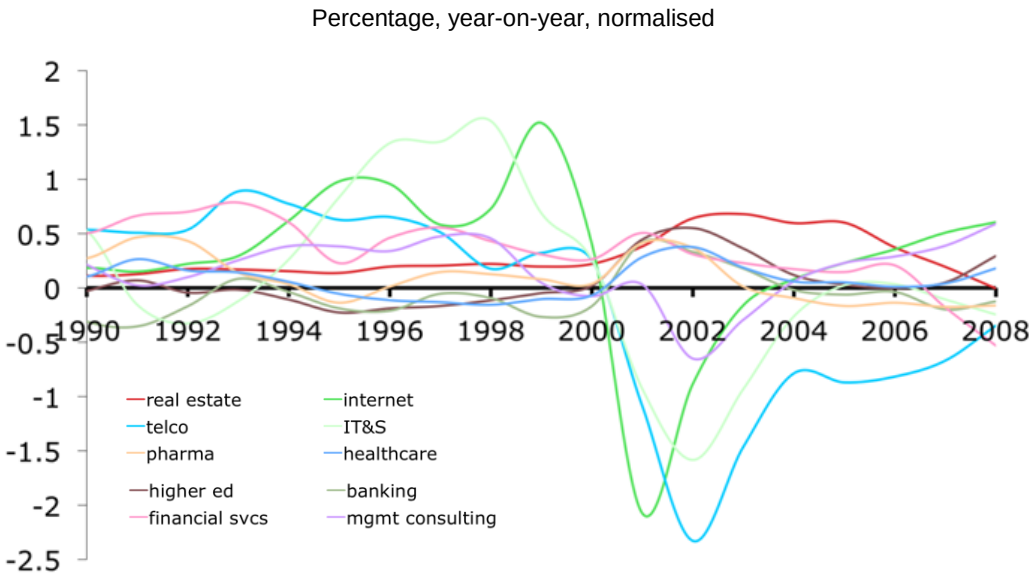
12 See http://www.conference-board.org/pdf_free/HWOLJan11_TN.pdf .

13 See http://innovationandgrowth.wordpress.com/2012/02/07/app-economy-is-job-leader-into-the-future/

particular to social networking sites for professionals such as LinkedIn, but sites such as Facebook also collect relevant data related to the carrier and skills of individuals and thus could be used for the creation of real-time employment-related indicators.

LinkedIn, for example, has data of approximately 150 million members around the world in 2011, providing information about their curriculum vitae and their job applications. By monitoring and analyzing the net change in positions across industries, the impact of business cycles on employment can be assessed in real-time as well as ongoing structural changes in the economy. Figure 7, for example, highlights the increase in the number of job starters in the "Internet" industry during the dot.com bubble in 2000, as well as the impact of the bust in 2001-2002 (normalized by the overall increase in job starters). It also highlights the first employment effects of the financial and economic crisis which was already visible in 2007 with the number of net job starters in financial services decreasing notably (there were more people leaving financial services than starting in this sector).

Figure 7. Net job starters by top industries by volume, 1998-2008



Note: 6 out of 147 industries represent 25% of new positions in 2011.

Source: LinkedIn Analytics.

Output

The Internet can provide a data source for predicting trends on output across the economy. In the following sections two promising sources are discussed, namely: (i) search engine data such as provided by Google Insights, and (ii) transaction based indicators such as the SWIFT Index, which is based on the volumes of SWIFT[14] customer credit transfers.

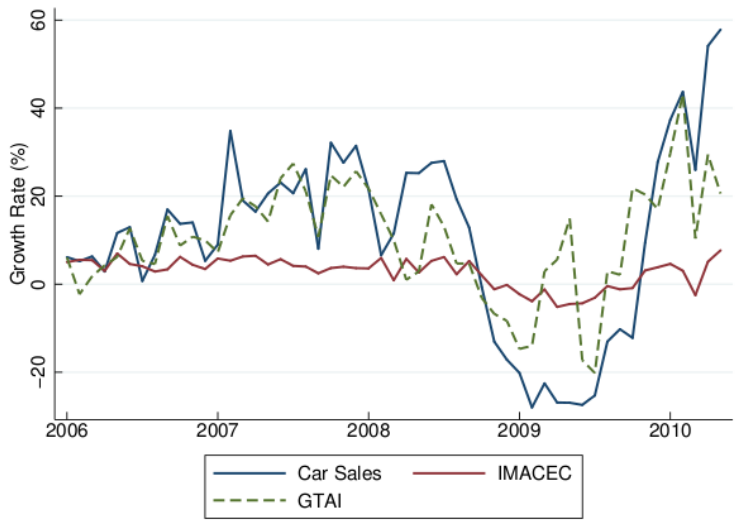
14 The Society for Worldwide Interbank Financial Telecommunication (SWIFT) provides the network that enables more than 10 000 financial institutions and businesses worldwide (210 Countries) to send and receive data on financial transactions. SWIFT first introduced the bank identifier codes (BICs) which became standardized under ISO 9362, and are therefore referred to as "SWIFT codes".

Search engine data

A number of authors such as Choi and Varian (2011), Carrière-Swallow and Labbé (2013), and Della Penna & Huang (2010) use Google Insights data to predict present ("to nowcast") economic metrics including retail good consumption and travel activities to cite a few. Because the potential of Google Insights has already been discussed in the previous section on employment (also see Box 1 for more details on Goolge Insights), this section only briefly describes the research of Carrière-Swallow and Labbé (2013), which is based on Google Insights data for predicting automobile sales in Chile. This case is the more interesting as it highlights that Google Insights data can help predict output despite challenges related to low broadband penetration rates and wealth levels when compared to the OECD average, as is the case for Chile.

Carrière-Swallow and Labbé (2013) use Google Insights to create a Google Trend Activity Index (GTAI) with the names of nine of the most popular automobile manufacturers in Chile (by volume of sales) being the keywords. The output to be forecast was the year-over-year (y-o-y) growth in the volume of car sales in Chile as provided by the national statistics agency of Chile.[15] As Figure 8 shows, the GTAI already provides a relative good fit with car sales.

Figure 8. Forecasting monthly growth in automobile sales in Chile, January 2006 – July 2010



Source: Carrière-Swallow and Labbé (2013).

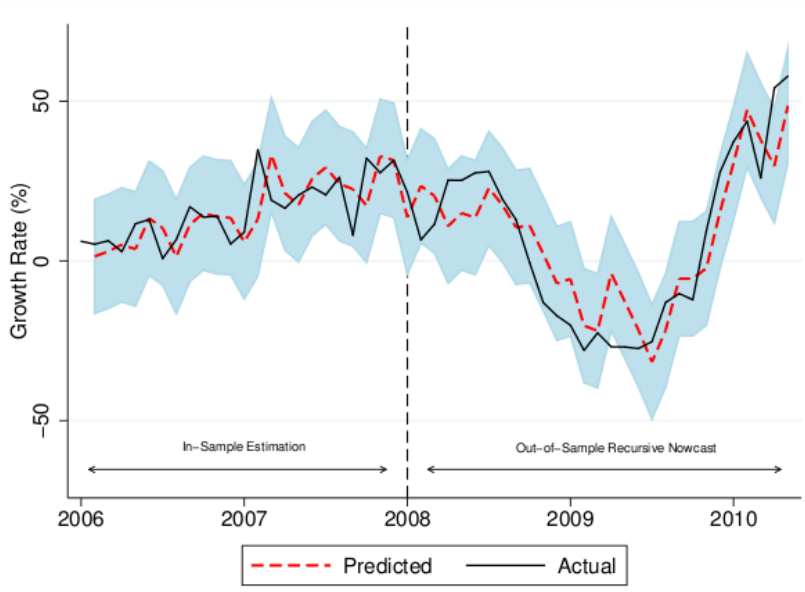
However, Carrière-Swallow and Labbé (2013) do not use the GTAI alone to forecast automobile sales, but rather to augment their existing models.[16] To further improve their forecasting models, they also introduce a backward shifted one-month "window of search". This is shifted by two weeks to reflect the fact that Internet users searching for automobile information do not proceed with their purchase in the very

15 Y-o-y changes in variables were used throughout the analysis to "avoid spurious correlations from seasonal effects that our short sample length does not allow us to reliably model" (Carrière-Swallow and Labbé, 2013).

16 Carrière-Swallow and Labbé (2013) show that using the GTAI significantly improves each benchmark model. It reduces the Root Mean Squared Error (RMSE) by between 9% and 14% and increases significantly the success rate of correctly identifying the direction of the change in growth rates in the ARMA(2,2) model to 65% of months compared to 50% of month when the GTAI was not used.

same week, but rather a couple of weeks afterwards.[17] The window of observation is also used to improve the robustness of the estimated parameters of the model used for out-of-sample forecasting (see Figure 9).

Figure 9. Monthly prediction of growth in automobile sales in Chile, January 2006 – July 2010



Source: Carrière-Swallow and Labbé (2013).

Overall, Carrière-Swallow and Labbé (2013) conclude that Google Insights data improves both the in- and out-of-sample accuracy of models for automobile sales, and that it could be considered a slightly leading indicator.

Financial transaction based indicators

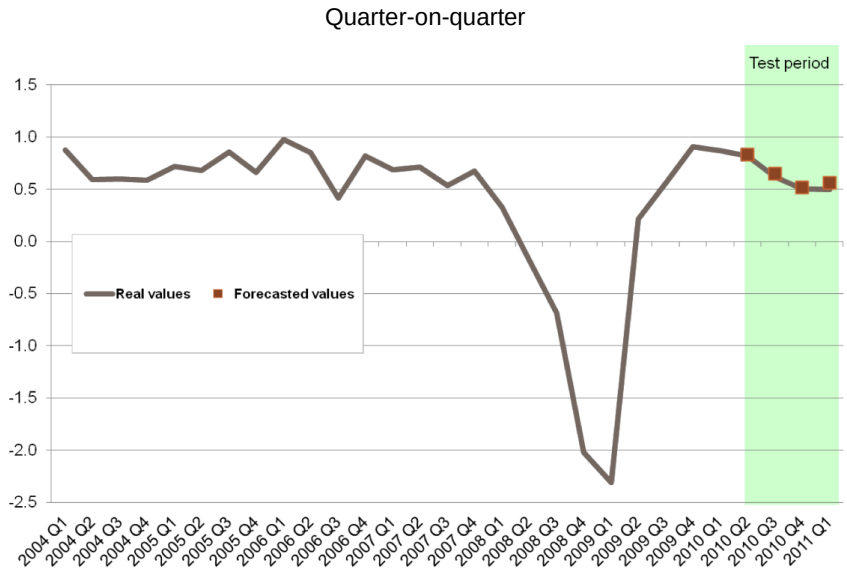
Today, a large share of global financial transactions is supported by IT-systems, and in many cases these systems are monitoring and logging the transactions for security and accountability purposes among other reasons. Based on these logs indicators can be created to measure and forecast global financial transactions as well as world economy output trends.

The SWIFT Index, for example, has been discussed as a predictor for world economy output growth. It is based on the volumes of SWIFT customer credit transfers between financial institutions, as tracked by the so called SWIFT MT103 payment messages. The MT103 payment message is "a specific message format used mainly for transferring information about money between customers of different banks or other similar financial institutions" and generates several million data points daily (Steinert-Threkeld, 2011: SWIFT, 2012). In order to capture as far as possible only those payment messages that are linked to real economy activity, messages need to be filtered to exclude events that are not underlined by a specific economic activity. The MT103 message flows are than aggregated at world and respectively at OECD-level to form the underlying data for calculating the SWIFT Index based on "the changes in these volumes against the base month considered to be January 2005" (SWIFT, 2012).

[17] By doing so, Carrière-Swallow and Labbé (2013) improve their forecasting models (the RMSE decreases and the adjusted R^2 increases for lags of up to two weeks).

SWIFT (2011) already highlighted the high correlation of payments traffic and in particular the volumes of MT103 messages with GDP. In SWIFT (2012), the prediction power of the static ARMA(p,q) models (only based on the SWIFT Index) were compared with the dynamic ARMA(p,q) models (based on the SWIFT Index and the GDP growth in the previous month). The conclusion of these tests was that combining the SWIFT Index with the previous GDP growth rate provides the best predictor. In the case of quarter-on-quarter (q-o-q) growth, the RMSE was reduced by 57% compared to the benchmark model and R^2 increasing from 0.51 to 0.71 (see Figure 10). For Q1 2012 SWIFT (2012b) predicted a 2% y-o-y growth in GDP in the OECD area and for Q2 2012 a 2.3% y-o-y growth.

Figure 10. SWIFT Index predictive power: OECD GDP quarterly growth, Q1 2004 – Q1 2011



Note: Test period forecasts based on OECD GDP data published in May 2011 and SWIFT Index total aggregate series.

Source: SWIFT.

Demographics

As people place more-and-more personal information on social networking sites like Facebook and LinkedIn demographic data such as age, education, marital status and location are becoming more available. This is also true with businesses of all types that are increasingly engaged in targeted marketing. For example, in the UK, roughly half of all households have a Tesco, the large UK supermarket chain, fidelity or loyalty card. As a result, Tesco processes 100 market baskets a second which contain on average 27 products per basket, accounting for 6 million transactions a day (Ryan, 2010). Each product purchased throws off 45 pieces of data which can be analysed to determine the ethnicity of the shopper, the likely family size and whether the household has a pet.

Frequently this data is then linked to data that retailers obtain when the fidelity card is obtained, such as age and address, which is then augmented with data that the firm can purchase from 3rd party "data aggregators" such as job history, credit history and estimated salary, marital status, number of cars owned and the year you bought (or lost) your house (Duhigg, 2012). Firms use this data to perform "predictive analysis" that imputes personal preferences such as vegetarianism, religion and sexual orientation. With more data and improved analytical capabilities, this predictive capability has become increasingly sophisticated and accurate. One example of this comes from the US retail chain, Target, which can predict whether a woman is pregnant or not, and even estimate her due date with extraordinarily good accuracy

based on the purchase data of not more than 25 products (Duhigg, 2012). While this has created an uproar in some communities, it has also created a demand from other firms who would like to purchase Target's know-how (Hill, 2012).

Development and natural risk management

As highlighted already in previous sections, big data provides a cost effective way to create real-time indicators. This can be very relevant for developing economies as well, which may find it challenging to finance and build reliable statistical systems. In particular in the context of natural disasters big data can provide the crucial real-time information needed to save human life. The following examples have in common that they are based on the use of mobile phones, for which penetration rates were still growing with two digit rates in developing countries in 2011 (ITU, 2012). These data sources are being explored by international initiatives such as *Paris21*, the Partnership in Statistics for Development in the 21st century, and *Global Pulse*, an initiative launched by the Executive Office of the United Nations Secretary-General, in response to the need for more timely data to track and monitor the impacts of global and local socio-economic crises (United Nations Global Pulse, 2012).

The first example discusses the use of mobile phones for doing household surveys to gain a better understanding of changes in human well-being[18]. The second example is based on the International Network of Crisis Mappers, an international community of "experts, practitioners, policymakers, technologists, researchers, journalists, scholars, hackers and skilled volunteers" dedicated to the use of mobile and web-based applications for generating participatory maps and crowdsourcing event data.[19] These initiatives are being explored in the following.

Mobile surveys

National statistic offices (NSOs) in developing economies often find it more challenging to do reliable nationwide surveys than in developed economies. To a major extent this is due to the high share of the population living in rural areas. This not only makes sampling more challenging, but may result in higher costs for reaching underrepresented populations than NSOs' budgets allow.

The mobile survey project of the UN Global Pulse, together with IT company Jana, aims at replicating the standards of traditional household surveys, in real-time on a global scale over SMS. It is targeting in particular underrepresented populations that have access to mobile phone technology. Through its proprietary network, Jana has access to over 2 billion mobile subscribers who have opted into answering survey questions "in exchange for a small amount of airtime".[20] So it builds on the pay-as-you-go fee models currently dominating telecommunication markets in developing economies (see OECD, 2011). Analysts can access the demographic data (including information about economic status, gender, age, literacy, etc.) via a web interface including interactive (visual) analytic tools.

It should be highlighted at this point that mobile based surveys still suffer from the same drawbacks as traditional surveys, although they may be less cost intensive to run; namely they assume that the answers provided by respondents are correct. Research has shown, however, that individuals may sometimes not be willing or able to answer the surveys correctly. This can be because respondents either *i*) consider the

18 See http://www.unglobalpulse.org/projects/global-snapshot-wellbeing-mobile-survey

19 See http://crisismappers.net/

20 See http://jana.com/about-us/

question asked too sensitive, or *ii*) do not have the necessary skills to understand and answer the question correctly (see OECD, 2011).

Crowd-sourcing event data

A growing number of initiatives are focussing on using mobile networks and the Internet for generating participatory maps and crowd-sourcing event data. These initiatives rely on the participation of thousands and millions of users to provide information over channels as diverse as telephone, SMSs, and social networking sites such as Twitter. All these information are then provided with their geo-location[21] on an interactive map for practitioners in the field and for policy makers. It is interesting to note that some of these initiatives are even sharing the same underlying technologies.

The Pak Flood Incident Reporting System, for example, was a system for reporting incidents related to the 2010 flood disaster in Pakistan, which killed over 1 600 and displaced over 18 million individuals. The tool gives users a way to report floods via SMS. These incidents are then interactively mapped and provided in combination with reports available in the media as online reports for practitioners and volunteers in the field, as well as to policy makers looking to better respond to the natural disaster. Another initiative that is based on the same underlying technology is the LRA Crisis Tracker. It is used to track criminal activities of the Lord's Resistance Army (LRA), a rebel group most active in Central Africa. "Using information sourced from Invisible Children's Early Warning Radio Network, UN agencies, and local NGOs, this tool allows for better response from governments, policy-makers, and humanitarian organizations"[22] (LRA Crisis Tracker, 2012).

3. The limits of "big data"

The use of non-traditional sources for statistics does not come without limitations, which in the current "big data" hype are even more important to acknowledge. There are considerable risks that the underlying data and analytic algorithms could lead to unexpected false results. This is more the case where the analysis process is automated, as illustrated by the case of the Knight Capital Group, a global financial services firm, which lost USD 440 million in 2012, most of it in less than an hour, because it's algorithmic trading system (ATS) behaved unexpectedly (Mehta, 2012). Users should therefore be aware of the limitations that come with the use of big data; otherwise they may cause social and economic harms to themselves as well as to third parties (e.g. to individuals through privacy violations). Three types of errors can be distinguished: (i) errors caused by poor data quality, (ii) errors that come with the inappropriate use of data and analytics, and (iii) errors that are caused by the unexpectedly changing environment from which data is collected (i.e. data environment). The latter issue is particularly relevant for the automation of data analytics.

Poor quality data

The information that can be extracted from data depends on the quality of the data. Poor quality data will therefore almost always lead to poor results ("garbage in, garbage out"). Therefore, data cleaning (or scrubbing) is often highlighted as an important step before the data can be analysed. And this often involves significant costs as it can account for 50% to 80% of a data analyst's time together with the actual data collection (Lohr, 2014). Because information is context dependent, data quality however depends on the intended use: Data that are of good quality for certain applications can thus be of poor quality for other

21 Geolocation can be generated through cell phone triangulation, GPS, or directly by asking users in the case of phone calls.

22 See www.lracrisistracker.com/#about.

applications. The OECD (2011b) *Quality Framework and Guidelines for OECD Statistical Activities* therefore defines data quality as "fitness for use" in terms of user needs (see Box 4). It highlights that "if data is accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data". The OECD (2013b) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) also provides a number of criteria for the quality of personal data for the purpose of privacy protection. The Recommendation states that "personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date".

Box 4. The factors affecting data quality

The OECD (2011b) *Quality Framework and Guidelines for OECD Statistical Activities* defines data quality as "fitness for use" in terms of user needs, which underlines the context dependency of data. OECD (2011b) in particular suggests that data quality (and thus value) needs to be viewed as a multi-faceted concept. It defines the following seven dimensions of data quality:

1. **Relevance**: "is characterised by the degree to which the data serves to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts";
2. **Accuracy**: is "the degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure";
3. **Credibility**: "the credibility of data products refers to the confidence that users place in those products based simply on their image of the data producer, i.e. the brand image. Confidence by users is built over time. One important aspect is trust in the objectivity of the data";
4. **Timeliness**: "reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon". Real-time data is data with a minimal timeliness";
5. **Accessibility**: "reflects how readily the data can be located and accessed" as discussed in the previous section on data access and sharing;
6. **Interpretability**: "reflects the ease with which the user may understand and properly use and analyse the data". The availability of meta-data plays an important role here as they provide for example "the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any"; and
7. **Coherence**: "reflects the degree to which they are logically connected and mutually consistent. Coherence implies that the same term should not be used without explanation for different concepts or data items; that different terms should not be used without explanation for the same concept or data item; and that variations in methodology that might affect data values should not be made without explanation. Coherence in its loosest sense implies the data are 'at least reconcilable'".

Furthermore, the information that can be extracted from data is not only a function of the data itself, but also a function of the (analytic) capacity to link data and to extract insights. This capacity is not only determined by available (meta-) data, analytic techniques and technologies, but more importantly, is a function of pre-existing knowledge and skills. This means that there are a number of factors beyond the data itself which determine its quality:

1. **Data linkage**: Information depends on how the underlying data is organized and structured and how it can be linked. In other words, the same data sets can lead to different information depending on their structure including their linkages with other (meta-) data.
2. **Data analytic capacities**: The quality of data also depends on the meaning as extracted or interpreted by the receiver. The same data sets can thus lead to different information and is thus depending on the analytic capacities of the "receiver" including her or his skills and (pre-) knowledge, available techniques and technologies for data analysis.

Inappropriate use of data and analytics

As highlighted above, some have suggested that with big data, decision makers could base their actions only on analytical facts without the need to understand the phenomenon, on which they are acting. As correlation would be enough with big data, scientific methods and theories would no longer be important. While it is true that analytics can be effective in detecting correlations in "big data", especially those that would not be visible with smaller sized volumes of data, it is also widely accepted among practitioners that data analysis itself relies on rigorous scientific methods, in order to produce appropriate results.

The rigour starts with how the quality of the data is assessed and assured. But even if data has good quality, which is not trivial, data analytics can still lead to wrong results if the data used is irrelevant or not representative, and does not fit the business or scientific questions it is supposed to answer (Loukides, 2014). Experts recognize that it is often too tempting to think that with big data, one has sufficient data to answer almost every question and to neglect data biases that could lead to false conclusions. The temptation is even bigger when correlations are suggested to be enough to drive decision-making processes, in which case the results could lead to nonsense. This is because in big data analyses correlations can often appear statistically significant even if there is no causal relationship. Marcus and Davis (2014) give the illustrative example, where big data analysis reveals that the United States murder rate was well correlated with the market share of Internet Explorer from 2006 to 2011. Obviously, any causal relationship between the two variables is spurious.

The risk of inappropriate use of data and analytics underlines the need for high skills in data analysis. It also challenges current trends in the democratisation of data analytics, which suggests that everyone and every organisation today can apply data analytics effectively. As O'Neil (2013a) argues, the simplicity of applying data analytics today thanks to software improvements make it easy for non-experts to believe in software generated answers which might not correspond to reality. Furthermore, the need for understanding causal relationship means that sufficient domain specific knowledge is necessary to apply data and analytics. Obviously the availability of high skills in data analysis and the rigorous use of data and analytics do not prevent data and analytics to be wrongly used intentionally for economic, political, or other advantages. Literature is full of cases where sophisticated econometric models have been used to lie with data. O'Neil (2013b) discusses some examples.

Changing data environment

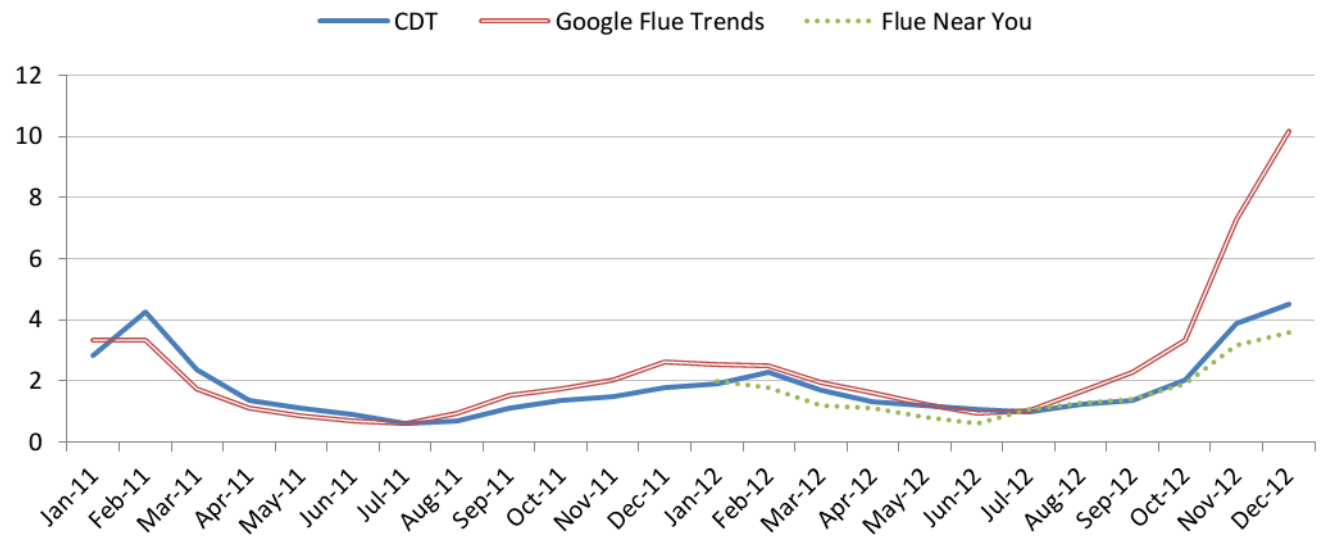
Even when the data and analytics are perfectly used initially, this does not mean that they will always deliver the right results. Data analytics, in particular when used for decision automation, can sometimes be easily "gamed" once the factors affecting the underlying algorithms have been understood, for example, through reverse engineering. Marcus and Davis (2014) present for example the case where essay evaluation analytics that relied on measures like sentence length and word sophistication to determine typical scores given by human graders. These programs were outwitted by students who suddenly started "writing long sentences and using obscure words, rather than learning how to actually formulate and write clear, coherent text". More popular examples (with business implications) are techniques known as "Google bombing" and "spamdexing" where users are adjusting Internet content, links and sites to artificially elevate website search placement in search engines (Segal, 2011; Marcus and Davis, 2014).

Data analytics does not need to be intentionally gamed to lead to wrong results. Often they are just not robust enough to unexpected changes in the data environment. This is because data analytics users (including the developers of autonomous systems) cannot envision all eventualities that could affect the functioning of their analytic algorithms and software, in particular when it is used in a dynamic environment. In other words, data analytics are not perfect and some environments are more challenging

than others. The case of the Knight Capital Group, which lost USD 440 million in financial markets in 2012 due to unexpected behaviour of its trading algorithm, was already mentioned above. A more recent example is Google Flu Trends, which is based on Google Insights for Search and provides statistics on the regional and time-based popularity of specific keywords that correlate with flu infections.[23] Google Flu Trends has been used by researchers and citizens as a means to accurately estimate flu infection trends at a faster rate than the statistics provided by the Centers for Disease Control and Prevention (CDC). However, in January 2013, Google Flu Trends drastically overestimated flu infection rates in the United States (Figure 11). Experts estimate that this was due to "widespread media coverage of [that] year's severe US flu season" which triggered an additional wave of flu-related searches by flu unaffected people (Butler, 2013).

Figure 11. Flu-infection rates in the United States, January 2011 – December 2012

Estimated % of population with influenza-like illness, monthly average



Source: OECD based on Butler (2013)

These incidents, intentioned or not, are caused by the dynamic nature of the data environment. The assumptions underlying many data analytic applications may change over time, either because users suddenly change their behaviour in unexpected ways as presented above (see essay evaluation analytics) or because new behavioral patterns emerge out of the complexity of the data environment (see algorithmic trading). As Lazer et al. (2014) further explains, one major cause of the failures (in the case of Google Flu Trends) may have been that the Internet constantly changes and as a result the Google search engine itself constantly changes. Patterns in data collected are therefore hardly robust over time.

4. Implications for statistical agencies and statistical policy

Torrents of data streaming across public and private networks are a growing reality and increasingly a wide variety of organisations are mining these data to produce statistics in areas that were previously the undisputed domain of national statistical offices (NSOs). While private data suppliers have existed for centuries, what is new is the growth and improved quality. A networked world has almost eliminated the

23 Google Trends now also include surveillance for a second disease, dengue.

gap between collection and publication, allowing continuous data collection and enabling the collection of large samples that approach the population in some cases.

While the displacement of the NSOs as the source of the base data is not (yet) occurring, the use of non-traditional sources to "now cast" this base is becoming increasingly common. The confluence of technological, social and economic trends suggests that this shift is likely to grow quickly in a short period of time. As policy makers begin to experiment with these new sources and statistics and their expectations begin to change as regards to standards of timeliness, detail and frequency, the scenario of "bad data pushing out good data" becomes more likely.

Concurrent with this shift are tightening budgets and declining response rates (Groves, 2011) that compel NSOs to explore how best to harness this phenomenon in their mission to supply quality statistics for improving economic performance and social welfare. There is no turning back.

The proliferation of new providers of statistics, many of them private businesses, raised a number of statistical policy issues for NSOs, including:

- Should NSOs take on a new mission as a trusted 3rd party whose role would be to certify the statistical quality of these new sources?
- Should NSOs become a "clearing house" for statistics from non-traditional sources that meet their quality standards?
- Should NSOs use non-traditional sources to augment (and perhaps replace) their official series?
- Should NSOs issue statistical "best practices" in the use of non-traditional sources and the mining of "big data"?

What follows is a cursory listing of some of these implications intended for discussion. Further study is needed that incorporates discussions that have occurred in other forums as well as exploring some of the issues in greater depth.

NSOs as a trusted 3rd party

With the potential for improving timeliness and lowering the cost of statistics come new issues about how to ensure that the statistics collected via the Internet and other non-traditional sources are of high quality. With this comes the need for a "trusted 3rd party" to certify that certain standards, principles, and norms have been achieved such as defined for example by the OECD (2011b) *Quality Framework and Guidelines for OECD Statistical Activities* (see Box 4). This potential role has been suggested by Skaliotis (2009) of Eurostat, Robert Groves, Director of the US Census Bureau (see COSSA, 2011), and recently the US National Academy (see National Research Council, 2012).[24]

The reliability, statistical validity and generalisability of new forms of data are not well understood (see previous section on the limits of "big data"). For example, while techniques have been developed for accurately sampling some social networks, researchers have not addressed how well this social network represents the larger population, or how to estimate the error incurred in using this subset (Mislove et al, 2007). It is important to understand how the data collected from web sites compares with traditional survey data, particularly because different web sites have very different coverage. While a few organisations have

[24] See also http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF.

begun to publicly disclose their methodologies, and some have released the algorithms and codes used for the calculation of their trends and indices, in the majority of cases the publicly available methodology is insufficient to judge its statistical quality.

As in the case of many other Internet-related activities, the usual signals that provide trust are elusive. This has led to the appearance of 3rd Parties like "TRUST-e" who verify that certain norms and practices are being followed.[25] This compliance is designated by a "trust mark." NSOs could provide a similar service by establishing statistical standards that private providers had to meet to earn a NSO "trust mark." Beyond this methodological role, NSOs could provide a statistical function by testing the accuracy of the non-traditional data against official data, analysing for biases and deficiencies. While clearly a departure from the past for NSOs, this function would build on the core competence of NSOs as the guardians of rigorously derived and accurate statistics. The traditional surveys that have been the pillars of most NSOs would continue as a needed benchmark for non-traditional data, albeit perhaps less frequently, thereby saving resources.

As the number of non-traditional statistics is predicted to rise over time, there is a question as to how NSOs could perform this certification on a potentially large scale. Given that some of the non-traditional statistical series (*e.g.* BPP) cover a number of countries, a question arises as to which NSO should undertake the certification process: the country of origin or the country of observation? This suggests a need for internationally agreed standards and perhaps the need for intergovernmental organisations like the UN, the OECD or Eurostat to consider performing this task.

NSOs as a clearing house

An extension of this trusted 3rd party role could be the role of NSO's as "clearing houses" for non-traditional data series that meet their standards.WEF[26] The idea of a clearing house function would provide a central source for non-traditional statistics and indicators that have met the standards of the NSO and would allow NSOs to guide users in the proper use of these statistics as well as their limitations. By being selective, the NSO would provide a useful filter in a world that is already overrun by "too much information".

While this new role would be in keeping with the role NSOs currently play to disseminate their statistics and provide metadata, extending this model to private sector statistics could raise a number of new problems, including the NSO's liability for these statistics, both in terms of quality and why the NSO selected one series to be part of the clearing house and not another. As mentioned above, the anticipated fast rise of non-traditional statistics may require a quick scaling up of this function.

NSOs as a user of statistics from non-traditional sources

Aside from a certification (trusted 3rd party) or dissemination (clearing house) role, NSOs may make the decision to become active users of these non-traditional sources as a means of augmenting their traditional data (*e.g.* now casting) and in some cases, as a standalone statistical series. Groves (2011) describes it as: a "blended data world by building on top of existing surveys" which uses a multi-modal data acquisition and manipulation of data, including:

25 See http://www.truste.com/about-TRUSTe/

26 The Dialogic (2008) report on "Analysing the Internet as a data source (IaD)" includes a section on policy recommendations in which there is a 'plea' addressed to NSOs to play the function of a *clearing house* for Internet-based statistics).

- Internet behaviours;
- Administrative records;
- Internet self-reporting;
- Telephone;
- Face-to-face;
- Paper surveys;
- Real-time mode switch to fill in missing data; and
- Real-time estimation.

This new "blended data world" could be achieved either through NSOs collecting data from these sources themselves (*e.g.* "scrapping" the web as the BPP does or compiling the help wanted ads from various sites as the Conference Board does), or through purchasing the statistics from other organisations: a decision of "make" versus "buy."

Given the trends described above, the changing expectations from users on timeliness, detail and frequency as well as the growing use of these indicators from policy makers, NSOs will need to carefully consider this option. While there is a clear danger of NSOs becoming less relevant if they do not embrace this idea, there are a number of factors that need to be considered and questions answered before this new approach becomes a reality. Three of these limitations are outlined below:

Skills

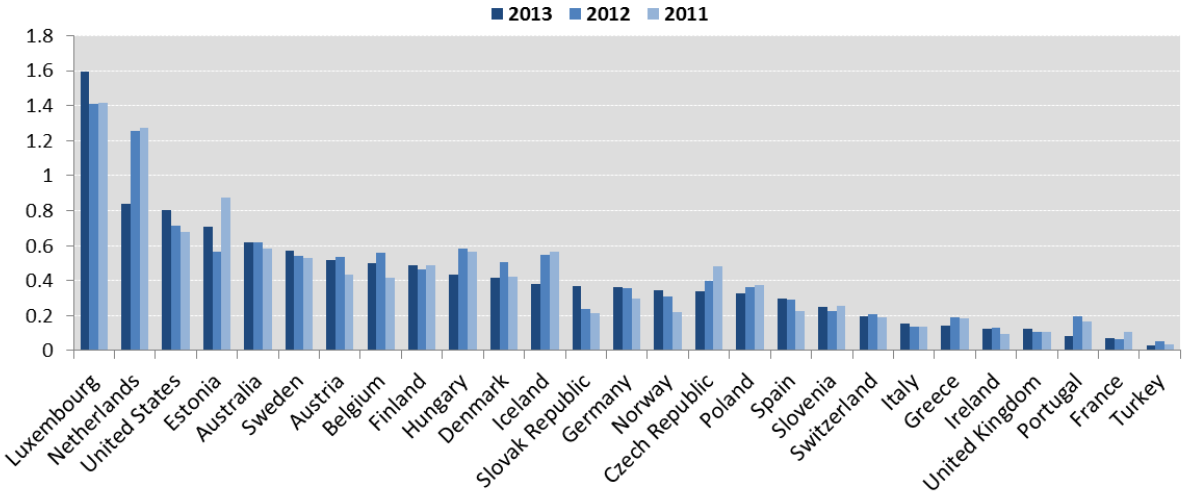
Compilation of data from non-traditional sources and the conversion of them into useful statistics using various analytical techniques require a skill set that while within the typical NSO profile, emphasises decision analysis and computer science besides statistics. Estimates suggest that data specialists[27] in 2013 accounted for around 0.6% of total employment in countries such as the Netherlands, the United States, Australia and Estonia, while in Luxembourg the share of data specialist almost reached 1.6% of total employment (Figure 12). In countries such as Portugal, France, and Turkey, the share of data specialists is far below 0.1% however. Based on estimates of the Bureau of Labor Statistics of the United States, demand for data specialist jobs are expected to grow at 17% between 2012 and 2022.[28] Statisticians, actuaries, and mathematicians are expected to have the fastest growth between 2012 and 2022 (26%). However, the share of statisticians, actuaries, and mathematicians in total employment is decreasing since 2012 suggesting with their further growing relative wages that countries could be facing a shortage in statisticians, actuaries, and mathematicians. This is consistent with MGI (2011) estimates that demand for "deep analytical talent" in the United States could be 50 to 60% greater than its projected supply by 2018.

27 Following the OECD definition ICT specialist (see OECD, 2014a), data-specialists are defined for the purpose of this report as those jobs for which *working with data constitutes a main part of the job*. In an attempt to provide comparable measures across OECD countries, data specialists have been defined according to the 2008 International Standard Classification of Occupations (ISCO-08) to include the following two occupations at 3-digit level: *Mathematicians, actuaries and statisticians* (212) and *Database and network professionals* (252).

28 This is six percentage points faster than the estimated total employment growth for that same period.

Figure 12. Share of data specialist in selected OECD countries, 2011-13

As share of total employment



Note: Data for Ireland and the United Kingdom are underestimated since they only consider ISCO-08 code 212. Detailed data for code 252 is not available.

Source: OECD based on ELFS and US Current Population Survey March Supplement, November 2014.

In the past, there have been considerable mismatches between the supply of and demand for ICT skills in general and for ICT specialist skills in particular such as software engineering skills. Shortfalls in domestic supply (owing to a large share of students leaving compulsory education, lack of educational courses and little training in the industry), restrictions on immigration of highly skilled personnel, or difficulties in international sourcing of development and analytical tasks requiring large amounts of interaction among employees are continuing challenges, as is the relatively low number of female employees in the ICT industry (OECD, 2012d). All this suggests that NSOs would be increasingly bidding against private firms for people who have these skills and could be forced to pay a premium to attract talent.

Data governance

The use of non-traditional data sources raises key challenges related to data governance, including issues such as (i) data access and sharing, (ii) data linkage and interoperability, (iii) data quality and curation, and in particular (iv) data ownership and control. Data governance can be particularly complex for NSOs in the case of non-traditional data sources since many of these new data sources reside on web sites or databases owned and controlled by private actors such as Google, eBay, Amazon, and many other internet service providers, social networking sites, and so on. The legality of web scraping, for example, has been challenged several times in courts both in the United States and abroad and there does not appear to be a consensus.[29] This relates to a broader discussion on the impact of intellectual property rights, and copyright in particular, on the use of data analytics as discussed briefly in Box 5. Furthermore, as consumers and users become more aware of the value of their data that various firms collect from them, a

29 Ryanair, a European airline, initiated a series of legal actions to prevent companies such as Billigfluege and Ticket Point from scraping ticket price data from their website to allow for easier comparison shopping (see Ryanair, 2010). eBay v. Bidder's Edge was a 2000 court case in California (see National Research Council, 2012);

shift in attitude may occur where they demand more explicit ownership or control of their data. The WEF (2011) suggests a model based on "end-user centricity" that recognises that "end users" are "...vital and independent stakeholders in the co-creation and value exchange of services and experiences." WEF envisions a system where individuals know about the data that is captured or inferred about them, the uses it is put to, and the parties that have access to it. This system would allow individuals to manage the extent to which their personal data is shared and make them aware of the compensation that they are receiving.

Box 5. Copyrights and data analytics

Data analytics is leading to an "automation" of knowledge creation, with text mining constituting a key enabling technology (Lok, 2010). Based on early work by Swanson (1986), scientists are now further exploring the use of data analytics for automated hypothesis generation and some have proposed analytical frameworks for standardising this scientific approach. Abedi et al. (2012), for example, have developed a hypothesis generation framework (HGF) to identify "crisp semantic associations" among entities of interest". Conceptual biology, as another example, has emerged as a complement to empirical biology and it is characterised by the use of text mining for hypothesis discovery and testing. This involves "partially automated methods for finding evidence in the literature to support hypothetical relationships" (Bekhuis, 2006). Thanks to these types of methods, insights were possible which otherwise would have been difficult to discover. One example is the discovery of adverse effects to drugs (Gurulingappa et al., 2013; Davis et al. 2013).

The potential for productivity gains in the creation of scientific knowledge are thus huge. However, questions have emerged about whether current copyright regimes are appropriately calibrated with regard to "automatic" scientific knowledge creation. According to the analysis of the JISC (2012) on the value and benefits of text mining, "the barriers limiting uptake of text mining appeared sufficiently significant to restrict seriously current and future text mining in UKFHE, irrespective of the degree of potential economic and innovation gains for society". Copyright has been identified as one of these barriers, which has led to debates between the scientific community and the publishers of scientific journals.

Source: OECD (2014b)

Privacy

Associated with the use of data from non-traditional sources is the question of privacy, which is beginning to plague many "Web 2.0" firms and can be expected to be an issue for NSOs should they actively engage in this "blended data" strategy. This may require new legal provisions which are not covered adequately by current statistical legislation. One particular challenge that deserves to be highlighted explicitly at this point is the blurring distinction between personal and non-personal data. The OECD (2013b) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, for example, define "personal data" as "any information relating to an identified or identifiable individual (data subject)". Any data that are not related to an identified or identifiable individual is therefore non-personal. However, data analytics have made it easier to relate seemingly non-personal data to an identified or identifiable individual, blurring the distinction between personal and non-personal data and as a result challenging any regulatory approach that determines the applicability of rights, restrictions and obligations on the basis of the "personal" nature of the data involved (OECD, 2013b; 2014b).

NSOs as an issuer of analytical best practices

As a wide number of organisations begin to compile statistics from their data streams, NSOs could consider issuing a series of analytical and statistical "best practices" in areas such as sampling, methods for "now casting," metadata standards and dissemination. Adhering to these best practices could be tied to the certification process and the acquisition of a "trust mark" by the NSO. Given the expertise and experience that resides in NSOs, this function would be relatively simple to perform, although it may require a slightly different skill set then the current profile. Given growing concern over the risk of "false discoveries,"

"…the trouble with seeking a meaningful needle in massive haystacks of data, is that many bits of straw look like needles" says Trevor Hastie, a Stanford University statistics professor (see Lohr, 2012).

Big data provides unlimited ammunition for biased fact finding, this guidance could include as well how to properly interpret results. This role may be best provided through partnerships with academics or research organizations who are actively exploring how best to exploit "big data" for statistics. Some research agencies have begun to fund research in this area, providing another area of collaboration with NSOs[30].

5. Conclusion

NSOs are not newcomers to the world of "big data", it is a realm where they bring considerable experience and credibility. But increasingly, this world has a growing number of data suppliers who are constructing their own statistics and indicators across many areas that were here-to-now the exclusive province of NSOs. Many of these indicators from non-traditional sources have received considerable attention because of their timeliness, detail and frequency. Because of these qualities, policy makers have begun to use them and more generally, expectations have begun to rise. Trends suggest that this will become increasingly common, although non-traditional sources will still in most cases complement, rather than substitute, traditional statistics. The BPP is an illustrative case. While it delivers a daily price index, it could never substitute the CPI, if only because the BPP rely on the CPI weights from consumer expenditure surveys to compile its daily index.

The developments discussed in this paper still require that NSOs consider how best to exploit this phenomenon to best fulfill their mission, which is to provide statistics that "underpin transparency and openness of policy decisions […] and provide a basis for the smooth functioning of society" (EC, 2009). This paper has provided an overview of developments with examples of statistics and indicators from a number of different areas. It then drew preliminary implications for statistical policy issues that NSOs may have to face, including:

- Should NSOs take on a new mission as a trusted 3rd party whose role would be to certify the statistical quality of these new sources?
- Should NSOs become a "clearing house" for statistics from non-traditional sources that meet their quality standards?
- Should NSOs use non-traditional sources to augment (and perhaps replace) their official series?
- Should NSOs issue statistical "best practices" in the use of non-traditional sources and the mining of "big data"?

Given the technological, social and economic factors that are propelling this movement forward, NSOs will need to address these questions, and forums such as the OECD offer a useful platform for working towards understanding the new potential roles to be played in this world of "big data". However, the analysis set forth in this paper has not yet profited from insight into the current efforts by NSOs to capture the benefits of big data and non-traditional on-line sources. Some NSOs are already in the process of tackling the benefits of these new data sources.

30 See http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767

In September 2013, for example, the *European Statistical System Committee* (ESSC) adopted the *Scheveningen Memorandum on "Big Data and Official Statistics"* (ESSC, 2013) to encourage the ESSC and its partners to "effectively examine the potential of Big Data sources" and to "adopt and action plan and roadmap by mid-2014". The *High-Level Group for the Modernisation of Statistical Production and Services* (HLG), which was set up by the *Bureau of the Conference of European Statisticians* to promote standards based modernisation in 2010, started in 2014 to assess the potential of "big data" with the following three main objectives:[31]

1. To identify, examine and provide guidance for statistical organizations on the main strategic and methodological issues that Big Data poses for the official statistics industry;

2. To demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts;

3. To facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

Last, but not least, in March 2014 the United Nations Statistical Commission established "a global working group mandated to provide strategic vision, direction and coordination of a global programme on Big Data for official statistics, to promote practical use of sources of Big Data for official statistics, while finding solutions for their challenges, and to promote capacity building and sharing of experiences in this respect" (UNSD-NBS China, 2014).[32]

All these initiatives underline the growing interest and experience of NSOs on "big data" for official statistics, but also the need for further international dialogues and knowledge exchange across NSOs. This paper, which benefited from the OECD horizontal (cross-committee) project on *New Sources of Growth: Knowledge-Based Capital*, in particular its Pillar on *Data-driven Innovation* (DDI, see http://oe.cd/bigdata), is among the first contributions to this international dialogue. Further comments and suggestions from NSOs about their particular experiences in this regard would therefore help improve the analysis and suggestions in the paper.

31 See http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Project.

32 As a first step towards these objectives, the United Nations Statistics Division (UNSD) and National Bureau of Statistics of China (NBS China) organised the "International Conference on Big Data for Official Statistics" on 28-30 October 2014 in Beijing, China.

BIBLIOGRAPHY

Abedi, V., R. Zand, M. Yeasin and F. E. Faisal (2012), "An automated framework for hypotheses generation using literature" *BioData mining*, 5(1)

Arthur, B. W. (2011), "The Second Economy," McKinsey Quarterly, October.

Askitas N. and K. N. Zimmermann (2009), "Google econometrics and unemployment forecasting", Technical report, SSRN 899, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341.

Bekhuis, T. (2006), "Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy", Biomed Digit Libr. 2006; 3: 2. Published online Apr 3, 2006. doi: 10.1186/1742-5581-3-2

Bertolucci, J. (2013), "Big Data's New Buzzword: Datafication", InformationWeek, 25 February, available at: www.informationweek.com/big-data/news/big-data-analytics/big-datas-new-buzzword-datafication/240149288

BIAC (2011), "BIAC Thought Starter: A Strategic Vision for OECD Work on Science, Technology and Industry, 12 October, mimeo.

Bollier, D. (2010), The Promise and Peril of Big Data, The Aspen Institute, Washington DC.

BPP (2014), US Daily Index, available at: http://bpp.mit.edu/usa/ (accessed on 22 October)

Brynjolfsson, E, L. M. Hitt, and H. H. Kim, (2011), "Strength in Numbers: how does data driven decision making affect firm performance?" 22 April, SSRN abstract 1819486, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486 .

Bullas, J. (2011), "50 Fascinating Facebook Facts And Figures", jeffbullas.com, 28 April, available at: www.jeffbullas.com/2011/04/28/50-fascinating-facebook-facts-and-figures.

Butler, D. (2013) "When Google got flu wrong", Nature, February 13, available at: http://www.nature.com/news/when-google-got-flu-wrong-1.12413

Carriere-Swallow and F. Labbe (2010), "Nowcasting with Google Trends in an Emerging Market", Central Bank of Chile Working Papers , No. 588, July, available at: http://ideas.repec.org/p/chb/bcchwp/588.html.

Choi, H. and H. Varian (2011), Predicting the Present with Google Trends, 18 December, http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf.

Cisco (2013), "Cisco Visual Networking Index: Forecast and Methodology, 2013–2018" , White Paper, 10 June, available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.

Commission of European Communities [EC] (2010) "Riding the Wave: How Europe can gain from the rising tide of scientific data", Final report by the High-level Expert Group on Scientific, October, available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

Commission of European Communities [EC] (2009), "Communication from the Commission to the European Parliament and the Council on the production method of EU Statistics: a vision for the next decade," COM(2009)404 final, Brussels, 10 August, available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF.

Commission of European Communities [EC] (2003), Directive 2003/98/EC of the European Parliament and of the Council, available at: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf.

COSSA (2011), Washington Update, Vol. 30, No. 20-B, November 10, 2011, pp. 13-14.

D'Amuri, F. and J. Marcucci (2010), "Google it! Forecasting the US unemployment rate with a Google job search index", SSRN, 2010, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1594132.

Davis, A. P., T.s C. Wiegers, P. M. Roberts, B.L. King, J.M. Lay, K. Lennon-Hopkins, D. Sciaky, R. Johnson, H. Keating, N. Greene, R. Hernandez, K. J. McConnell, A. E. Enayetallah, and C.J. Mattingly (2013), "A CTD–Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions", Database 2013: bat080 doi:10.1093/database/bat080 published online November 28, 2013

Della Penna, N., and H. Huang, (2010), "Constructing consumer sentiment index for US using Google searches", Working paper No. 2009-26, available at: https://ideas.repec.org/p/ris/albaec/2009_026.html

Dialogic (2008), "Go with the dataflow! Analysing the Internet as a data source (IaD)", 28 April, available at: www.umic.pt/images/stories/publicacoes1/main_report.pdf.

Duhigg, C (2012), "How Companies Learn Your Secrets," The New York Times, 16 February, available at: www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

The Economist (2010), "Data, data everywhere", 27 February, available at: www.economist.com/node/15557443.

Edgecliff-Johnson, A. (2011), "FT Launches web-based app to bypass Apple's iTunes" 7 June, The Financial Times, available at: www.ft.com/cms/s/0/8b458e4a-9084-11e0-9531-00144feab49a.html.

European Statistical System Committee [ESSC] (2013), *Scheveningen Memorandum on "Big Data and Official Statistics"*, September, available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version_0.pdf.

Graves, A. (2009), The Price of Everything but the Value of Nothing, in Uhlir (2009).

Groves, R. M. (2011), "A Possible Data Future for the Observational Social Sciences," presented at the COSSA 30th Anniversary meeting, Washington, DC, November 2, 201, COSSA Washington Update, Vol. 30, No. 20-B, November 10, 2011, pp. 13-14.

Gurulingappa, H., L. Toldo, A.M Rajput, J. A. Kors, A. Taweel and Y. Tayrouz (2013), "Automatic detection of adverse events to predict drug label changes using text and data mining techniques", Pharmacoepidemiology and Drug Safety, November, 22(11), pages 1189–1194.

Hachman, M. (2012), "Facebook Now Totals 901 Million Users, Profits Slip", PC Magazin, available at: www.pcmag.com/article2/0,2817,2403410,00.asp.

Harding, R. (2011), "Bank to Publish Inflation Data from the Internet", Financial Times, 4 May, available at: www.ft.com/cms/s/0/f3495070-7659-11e0-b4f7-00144feabdc0.html.

Hey, J. (2004), "The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link", working paper, December, available at: www.dataschemata.com/uploads/7/4/8/7/7487334/dikwchain.pdf.

Hilbert M. and P. Lopez, (2011), "The world's technological capacity to store, communicate and compute information", Science, 10 February, available at: www.sciencemag.org/content/332/6025/60.

Hill, K. (2012), Could Target Sell Its 'Pregnancy Prediction Score'?, Forbes, 16 February, available at: www.forbes.com/sites/kashmirhill/2012/02/16/could-target-sell-its-pregnancy-prediction-score/.

Houghton, J., B. Rasmussen, and P. Sheehan (2010), "Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs", Report to SPARC, July 2010, Centre for Strategic Economic Studies, Victoria University.

IDC (2010), "The Digital Universe Decade – Are You Ready?" May, available at: www.emc.com/leadership/programs/digital-universe.htm .

ITU (2012), Key ICT indicators for developed and developing countries and the world (totals and penetration rates), available at: www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html, last time accessed: 07 December 2011.

JISC (2012), "The Value and Benefits of Text Mining", JISC, available at: www.jisc.ac.uk/sites/default/files/value-text-mining.pdf.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014), "The Parable of Google Flu: Traps in Big Data Analysis", Science , Vol. 343, 14 March, available at: http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

Lok, C. (2010), "Literature mining: Speed reading", 27 January, Nature, 463, 416-418, available at: www.nature.com/news/2010/100127/full/463416a.html.

Lohr, S. (2014), "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights", New York Times, 17 August, available at: www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Lohr, S. (2012), "The Age of Big Data," The New York Times, 11 February, available at: www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all.

Lohr, S. (2009), "For Today's Graduate, Just One Word: Statistics", New York Times, 5 August, available: www.nytimes.com/2009/08/06/technology/06stats.html.

LRA Crisis Tracker (2012), LRA Crisis Tracker: Map Methodology & Database Codebook v 1.2, Invisible Children + Resolve, available at www.lracrisistracker.com/sites/default/files/Map-Methodology-and-Database%20Codebook%20v1.0.pdf.

Loukides, M. (2014), "The backlash against big data, continued", O'Reilly Radar, 11 April, available at: http://radar.oreilly.com/2014/04/the-backlash-against-big-data-continued-2.html.

Loukides, M. (2010), "What is data science? The future belongs to the companies and people that turn data into products", O'Reilly Radar, 2 June, http://radar.oreilly.com/2010/06/what-is-data-science.html.

Lyman P. and H. Varian (2003), "How Much Information?", School of Information Management and Sysytems, University of California at Berekely.

Manyika et al (2011) "Big data: The next frontier for innovation, competition and productivity" McKinsey Global Institute, May, available at: www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Marcus and Davis (2014), "Eight (No, Nine!) Problems With Big Data", The New York Times, 6 April, available at: www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Mayer-Schönberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, 5 March, Eamon Dolan/Houghton Mifflin Harcourt.

McGuire, T., J.Manyika and M. Chui (2012), "Why big data is the new competitive advantage", *Ivey Business Journal*, July/August, available at: http://iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.VCJ7lPnoQjM

McKinsey Global Institute [MGI] (2011), *Big data: The next frontier for innovation, competition, and productivity*, May, available at: www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

Metha, N. (2012), "Knight $440 Million Loss Sealed by Rules on Canceling Trades", Bloomberg, 14 August, available at: www.bloomberg.com/news/2012-08-14/knight-440-million-loss-sealed-by-new-rules-on-canceling-trades.html.

Mislove A., Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, Bobby Bhattacharjee (2007), "Measurement and Analysis of Online Social Networks." in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

National Research Council (2012), Improving Measures of Science, Technology, and Innovation: Interim Report. Panel on Developing Science, Technology, and Innovation Indicators for the Future, Editors R.E. Litan, A.W. Wyckoff, and K.H. Fealing.

OECD (2014a), *Measuring the Digital Economy: A New Perspective*, OECD, Paris.

OECD (2014b), "Data-driven Innovation for Growth and Well-being: Interim Synthesis Report", Background report to the 4th Meeting of the OECD Global Forum on the Knowledge Economy, 2-3 October 2014, available at: http://oe.cd/bigdata2.

OECD (2013a), "Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by 'Big Data'", *OECD Digital Economy Papers*, No. 222, OECD Publishing. doi: 10.1787/5k47zw3fcp43-en.

OECD (2013b), "OECD Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data", 11 July, available at: www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2012a), "ICT Applications for the Smart Grid: Opportunities and Policy Implications", *OECD Digital Economy Papers*, No. 190, 10 January, OECD Publishing. doi: 10.1787/5k9h2q8v9bln-en.

OECD (2012b), "Machine-to-Machine Communications: Connecting Billions of Devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing. doi: 10.1787/5k9gsh2gp043-en.

OECD (2012c), "Improving the Evidence Base for Information Security and Privacy Policies: Understanding the Opportunities and Challenges related to Measuring Information Security, Privacy and the Protection of Children Online", *OECD Digital Economy Papers*, No. 214, OECD Publishing. doi: 10.1787/5k4dq3rkb19n-en.

OECD (2012d), "ICT Skills and Employment: New Competences and Jobs for a Greener and Smarter Economy", *OECD Digital Economy Papers*, No. 198, OECD Publishing. doi: 10.1787/5k994f3prlr5-en.

OECD (2011a), "Enhancing Consumer Policy Making: The Role of Consumer Surveys", DSTI/CP(2011)3/FINAL, 23 May, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=DSTI/CP%282011%293/FINAL&docLanguage=En.

OECD (2011b), "Quality Framework and Guidelines for OECD Statistical Activities", 17 January, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291.

O'Neil, C. (2013a), "K-Nearest Neighbors: dangerously simple", 4 April, available at: http://mathbabe.org/2013/04/04/k-nearest-neighbors-dangerously-simple/.

O'Neil, C. (2013b), "We don't need more complicated models, we need to stop lying with our models", 3 April, available at: http://mathbabe.org/2013/04/03/we-dont-need-more-complicated-models-we-need-to-stop-lying-with-our-models/.

Orange (2012), "V-Trafic becomes the first service to incorporate Orange's Floating Mobile Data as traffic information source", Press Release, 19 January, available at: www.orange-business.com/mnc/press/press_releases/2012/mediamobile.html.

OSTP (2010) "Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Sustainable Economics for a Digital Planet: Ensuring Long Term Access to Digital Information," February., http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Rosen, J. (2010), "The Web Means the End of Forgetting," New York Times Magazine, 21 July, available at: www.nytimes.com/2010/07/25/magazine/25privacy-t2.html.

Ryan, M. (2010), "Tesco Metrics: Every Little Bit of Data Helps", Economics, Psychology and Policy Blog, 16 December, available at: http://economicspsychologypolicy.blogspot.fr/2010/12/tesco-metrics-every-little-bit-of-data.html.

Ryanair (2010), "Ryanair Wins Screenscraper High Court Action", Press release, 01 March, available at: http://www.ryanair.com/en/news/ryanair-wins-screenscraper-high-court-action.

Segal, D. (2011), "The Dirty Little Secrets of Search" *The New York Times*, 12 February, available at: www.nytimes.com/2011/02/13/business/13search.html.

Skaliotis, M. (2009), "Official Statistics in the Era of Ubiquitous Connectivity and Pervasive Technologies", 'STATISTICS – INVESTMENT IN THE FUTURE 2', 14-15 September, www.czso.cz/conference2009/proceedings/data/technology/skaliotis.pdf.

Skaliotis, M. (2010), "Timeliness and Accuracy in Official Statistics 2.0", October 2010, www.nso.gov.mt/docs/MichailSkaliotis%20_Eurostat.pdf.

Steinert-Threkeld, T. (2011), "SIBOS 2011: SWIFT to Roll Out Global Economic Barometer," 19 September in www.securitiestechnologymonitor.com/news/swift-index-gdp-global-economic-barometer.

Stiglitz, J.E., Orszag, P.R., and Orszag, J.M. (2000), "The Role of Government in a Digital Age", Washington, D.C: Computer and Communications Industry Association, available at: http://www.dol.gov/ebsa/pdf/ccia.pdf.

Suhoy, T. (2009), "Query indices and a 2008 downturn: Israeli data", Technical report, Bank of Israel, 2009, www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.

Surowiecki, J. (2011), "A Billion Prices Now", The New Yorker, 30 May, available at: www.newyorker.com/talk/financial/2011/05/30/110530ta_talk_surowiecki.

Swanson D. R. (1986), "Undiscovered Public Knowledge", Library Quarterly, 56:103–118.

SWIFT (2011), "Leveraging your SWIFT traffic information to make better business decisions", SOFE 2011, Business Intelligence breakout session, available at: www.swift.com/resources/documents/sofe2011_BusinessIntelligence.pdf.

SWIFT (2012), "The SWIFT Index: Technical Description", February, available at: www.swift.com/resources/documents/SWIFTIndex_technical_doc.pdf.

Uhlir, P.F. (2009), (Rapporteur) The Socioeconomic Effects of Public Sector Information on Digital Networks: Towards Better Understanding of Different Access and Reuse Policies, Workshop Summary, National Academy of Sciences, Washington DC.

United Nations Global Pulse (2012), "Big Data for Development: Opportunities & Challenges", United Nations Global Pulse, May, available at: www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

United Nations Statistics Division [UNSD] and National Bureau of Statistics of China [NBS China] (2014), "International Conference on Big Data for Official Statistics", Concept Note, UNSD-NBS China, 28-30 October, available at:

http://unstats.un.org/unsd/trade/events/2014/Beijing/Conference%20on%20big%20Data%20-%20draft%20Concept%20Note%20-%2013%20August%202014.pdf.

Vollman, J. (2010), "Real Time Labor Market Information", Presentation at the Brookings Institution LMI Forum, 27 September, available at: www.brookings.edu/research/speeches/2010/09/27-labor-statistics-reamer.

World Economic Forum [WEF] (2011), "Personal Data: The Emergence of a New Asseet Class," 11 January, available at www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf.