Big Data: Big today, normal tomorrow

ITU-T Technology Watch Report
November 2013

This Technology Watch report looks at different examples and applications associated with the big data paradigm, identifies commonalities among them by describing their characteristics, and highlights some of the technologies enabling the upsurge of big data. As with many emerging technologies, several challenges need to be identified and addressed to facilitate the adoption of big data solutions in a wider range of scenarios. Big data standardization activities related to the ITU-T work programme are described in the final section of this report.

The rapid evolution of the telecommunication/information and communication technology (ICT) environment requires related technology foresight and immediate action in order to propose ITU-T standardization activities as early as possible.

ITU-T Technology Watch surveys the ICT landscape to capture new topics for standardization activities. Technology Watch Reports assess new technologies with regard to existing standards inside and outside ITU-T and their likely impact on future standardization.

Acknowledgements

This report was written by Martin Adolph of the ITU Telecommunication Standardization Bureau. Please send your feedback and comments to tsbtechwatch@itu.int.

The opinions expressed in this report are those of the author and do not necessarily reflect the views of the International Telecommunication Union or its membership.

This report, along with other Technology Watch Reports can be found at http://itu.int/techwatch.

Cover picture: Shutterstock

Technology Watch is managed by the Policy & Technology Watch Division, ITU Telecommunication Standardization Bureau.

Call for proposals

Experts from industry, research and academia are invited to submit topic proposals and abstracts for future reports in the Technology Watch series. Please contact us at tsbtechwatch@itu.int for details and guidelines.

Big data:
Big today, normal tomorrow

November 2013

Table of contents

1. Introduction

In early 2013 several European countries were rocked by a food scandal which uncovered a network of fraud, mislabeling and sub-standard supply chain management.

This was not the first food scandal, and will surely not be the last. For restaurant chains with thousands of branches and hundreds of suppliers worldwide, it is nearly impossible to monitor the origin and quality of each ingredient. Data and sophisticated real-time analytics are means to discover early (or, better yet, prevent) irregularities.

The events leading to the discovery and resolution of the scandal point to the promises and challenges of data management for multiparty, multidimensional, international systems. Billions of individual pieces of data are amassed each day, from sources including supplier data, delivery slips, restaurant locations, employment records, DNA records, data from Interpol's database of international criminals, and also customer complaints and user-generated content such as location check-ins, messages, photos and videos on social media sites. But more data does not necessarily translate into better information. Gleaning insight and knowledge requires 'connecting the dots' by aggregating data and analyzing it to detect patterns and distill accurate, comprehensive, actionable reports.

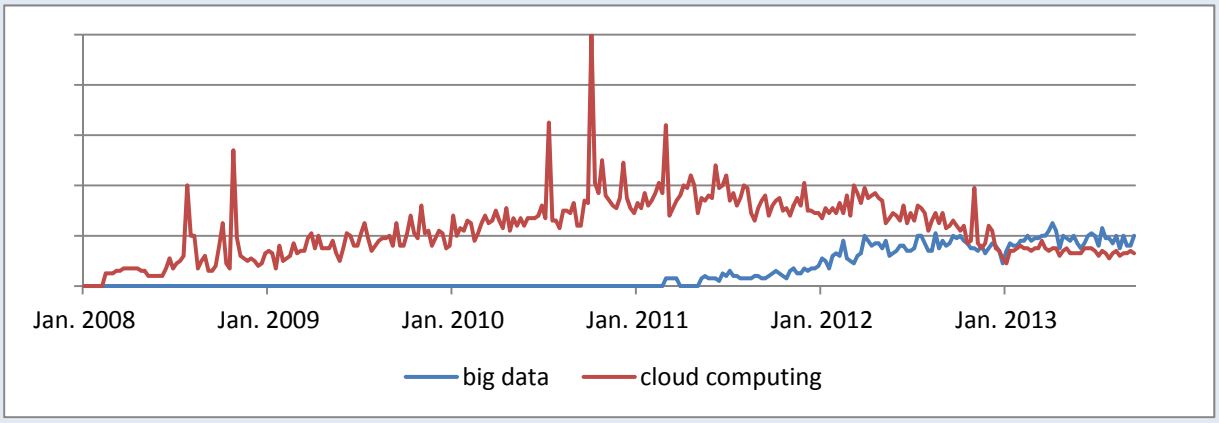
Big data – a composite term describing emerging technological capabilities in solving complex tasks – has been hailed by industry analysts, business strategists and marketing pros as a new frontier for innovation, competition and productivity. *"Practically everything that deals with data or business intelligence can be rebranded into the new gold rush"*[1], and the hype around big data looks set to match the stir created by cloud computing (see Figure 1) where existing offerings were rebranded as 'cloud-enabled' overnight and whole organizations moved to the cloud.

Putting the buzz aside, big data motivates researchers from fields as diverse as physics, computer science, genomics and economics – where it is seen as an opportunity to invent and investigate new methods and algorithms capable of detecting useful patterns or correlations present in big chunks of data. Analyzing more data in shorter spaces of time can lead to better, faster decisions in areas spanning finance, health and research.

This Technology Watch report looks at different examples and applications associated with the big data paradigm (section 2), identifies commonalities among them by describing their characteristics (section 3), and highlights some of the technologies enabling the upsurge of big data (section 4). As with many emerging technologies, several challenges need to be identified (section 5) and addressed to facilitate the adoption of big data solutions in a wider range of scenarios. Global standardization can contribute to addressing such challenges and will help companies enter new markets, reduce costs and increase efficiency. Big data standardization activities related to the ITU-T work programme are described in the final section of this report.

[1] Forbes: "Big Data, Big Hype: Big Deal," 31 December 2012, http://www.forbes.com/sites/edddumbill/2012/12/31/big-data-big-hype-big-deal/

Figure 1: News interest over time: big data vs. cloud computing



Note: Numbers represent search interest relative to the highest point on the chart.

Source: Google Trends, http://www.google.com/trends/

2. Big data everywhere – applications in health, science, transport and beyond

Data is critical in the healthcare industry where it documents the history and evolution of a patient's illness and care, giving healthcare providers the tools they need to make informed treatment decisions. With medical image archives growing by 20 to 40 per cent annually, by 2015, an average hospital is expected to be generating 665 terabytes of medical data each year.[2] McKinsey analysts predict[3] that, if large sets of medical data were routinely collected and electronic health records were filled with high-resolution X-ray images, mammograms, 3D MRIs, 3D CT scans, etc., we could better predict and cater to the healthcare needs of a population; which would not only drive gains in efficiency and quality, but also cut the costs of healthcare dramatically. Applications of big data analytics in the healthcare domain are as numerous as they are multifaceted, both in research and practice, and below we highlight just a few.

Remote patient monitoring, an emerging market segment of machine-to-machine communications (M2M), is proving a source of useful, quite literally lifesaving, information. People with diabetes, for instance, are at risk of long-term complications such as blindness, kidney disease, heart disease and stroke. Remote tracking of a glucometer (a blood sugar reader) helps monitor a patient's compliance with the recommended glucose level. Electronic health records are populated with data in near real time. Time series of patient data can track a patient's status, identify abnormalities and form the basis of treatment decisions. More generally, exploiting remote patient monitoring systems for chronically ill patients can reduce physician appointments, emergency department visits and in-hospital bed days; improving the targeting of care and reducing long-term health complications.

[2] Forbes, http://www.forbes.com/sites/netapp/2013/04/17/healthcare-big-data/

[3] McKinsey, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

It is not only the ill who use technology to monitor every detail of their biological processes.[4] The term *Quantified Self* describes a movement in which people are exploiting wearable sensors to track, visualize, analyze and share their states, movements and performance.[5] Fitness products and sleep monitors are some of the more popular self-quantification tools, and their users populate real-time data streams and global data factories.

Which treatment works best for specific patients?

Studies have shown that wide variations exist in healthcare practices, providers, patients, outcomes and costs across different regions. Analyzing large datasets of patient characteristics, outcomes of treatments and their cost can help identify the most clinically effective and cost-efficient treatments to apply. Comparative effectiveness research has the potential to reduce incidences of 'over-treatment', where interventions do more harm than good, and 'under-treatment', where a specific therapy should have been prescribed but was not. In the long run, over- and under-treatment both have the potential for worse outcomes at higher costs.

Scaling-up comparative effectiveness research can change how we view global health and improve the way public health crises are managed. Consider pneumonia, the single largest cause of child death worldwide. According to WHO data[6], each year the disease claims the lives of more than 1.2 million children under the age of five – more than AIDS, malaria and tuberculosis combined. Pneumonia is preventable with simple interventions and can be treated with low-cost, low-tech medication and care. However, the growing resistance of the bacterium to conventional antibiotics does underline an urgent need for vaccination campaigns to control the disease. Health data is vital in getting this message across to policy makers, aid organizations and donors, but, no matter how accurate and complete raw statistics and endless spreadsheets may be, their form is not one that lends itself to easy analysis and interpretation. Models, analytics and visualizations of deep oceans of data work together to provide a view of a particular problem in the context of other problems, as well as in the contexts of time and geography. Data starts 'telling its life story', in turn becoming a vital decision making tool. The Global Health Data Exchange[7] is such a go-to repository for population health data enriched by a set of tools to visualize and explore the data.[8]

Analyzing global disease patterns and identifying trends at an early stage is mission critical for actors in the pharmaceutical and medical products sector, allowing them to model future demand and costs for their products and so make strategic R&D investment decisions.

High-throughput biology harnesses advances in robotics, automated digital microscopy and other lab technologies to automate experiments in a way that makes large-scale repetition feasible. For example, work that might once have been done by a single lab technician with a microscope and a pipette can now be done at high speed, on a large scale. It is used to define better drug targets, i.e., nucleic acids or native proteins in the body whose activity can be modified by a drug to result in a desirable therapeutic effect.[9]

---

[4] NY Times, http://bits.blogs.nytimes.com/2012/09/07/big-data-in-your-blood/

[5] TED, http://www.ted.com/talks/gary_wolf_the_quantified_self.html

[6] WHO, http://www.who.int/mediacentre/factsheets/fs331/en/index.html

[7] Global Health Data Exchange, http://ghdx.healthmetricsandevaluation.org/

[8] BBC, http://www.bbc.com/future/story/20130618-a-new-way-to-view-global-health

[9] Wikipedia, http://en.wikipedia.org/wiki/Biological_target#Drug_targets

Automated experiments generate very large amounts of data about disease mechanisms and they deliver data of great importance in the early stages of drug discovery. Combined with other medical datasets, they allow scientists to analyze biological pathways systematically, leading to an understanding of how these pathways could be manipulated to treat disease.[10]

Data to solve the mysteries of the universe

Located just a few minutes' drive from ITU headquarters, CERN is host to one of the biggest known experiments in the world, as well as an example of big data, *par excellence*. For over 50 years, CERN has been tackling the growing torrents of data produced by its experiments studying fundamental particles and the forces by which they interact. The Large Hadron Collider (LHC) consists of a 27-kilometer ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way. The detector sports 150 million sensors and acts as a 3D camera, taking pictures of particle collision events at the speed of 40 million times per second.[11]

Recognizing that this data likely holds many of the long-sought answers to the mysteries of the universe, and responding to the need to store, distribute and analyze the up to 30 petabytes of data produced each year, the Worldwide LHC Computing Grid was established in 2002 to provide the necessary global distributed network of computer centers. A lot of CERN's data is unstructured and only indicates *that* something has happened. Scientists around the world now collaborate to structure, reconstruct and analyze *what* has happened and *why*.

Understanding the movement of people

Mobility is a major challenge for modern, growing cities, and the transport sector is innovating to increase efficiency and sustainability. Passengers swiping their RFID-based public transport pass leave a useful trace that helps dispatchers to analyze and direct fleet movements. Companies, road operators and administrations possess enormous databases of vehicle movements based on GPS probe data, sensors and traffic cameras, and they are making full use of these data treasure chests to predict traffic jams in real time, route emergency vehicles more effectively, or, more generally, better understand traffic patterns and solve traffic-related problems.

Drivewise.ly and Zendrive are two California-based startups working on data-driven solutions aimed at making drivers better, safer and more eco-friendly. The assumption is that driving habits and commuting patterns can be recognized or learned by collecting the data captured with the sensors of a driver's smartphone (e.g., GPS, accelerometer) and referencing it to datasets collected elsewhere. Taken in the context of data derived from a larger community of drivers, drivers gain insights such as "*leave 10 minutes earlier to reduce your commute time by 20 minutes"*, and adapting one's driving style can in turn help reduce fuel consumption and emissions. Data collected and analyzed by such apps can attest you for a defensive driving style, which could help in renegotiating your insurance premium.[12]

[10] University of Oxford, http://www.ox.ac.uk/media/news_stories/2013/130503.html

[11] CERN, http://home.web.cern.ch/students-educators/updates/2013/05/exploration-big-data-frontier

[12] http://drivewise.ly/ and http://zendriveblog.tumblr.com/

Your mobile phone leaves mobility traces too, and this is exploited as a resource for transport modeling. This is of particular interest where other transport-related data is scarce. City and transport planning was one of the themes of the 'Data for Development' challenge[13] launched by telecommunication provider Orange in summer 2012. Participants were given access to anonymized datasets provided by the company's Côte d'Ivoire branch which contained 2.5 billion records of calls and text messages exchanged between 5 million users over a period of 5 months.[14]

Situated on a lagoon with only a few bridges connecting its districts, Abidjan, the capital of Côte d'Ivoire is experiencing major traffic congestion. As it drafts a new urban transport plan for individual and collective means of transportation, call records offer an informative set of data on the mobility of the population. Selecting the calls made from a residential area during the evening hours (i.e., when people are at home), and monitoring the locations of the calls made on the same phones throughout the following day, produces data which reveals how many people commute, as well as where and at what times – resulting in mobility maps which inform decisions on road and transport investment.[15] Box 1 details a case where Korea Telecom helped the City of Seoul determine optimal night bus routes. Box 2 showcases a similar analysis in Geneva, Switzerland.

On a larger geographical scale, cell phone data contributes to analysis of migration patterns and is invaluable in crisis management. Launched by the Executive Office of the United Nations Secretary-General in the wake of "The Great Recession", Global Pulse[16] is an innovation initiative established in response to the need for more timely information to track and monitor the impacts of global and local socio-economic crises. The initiative is exploring how new, digital data sources and real-time analytics technologies can help policymakers understand human well-being and emerging vulnerabilities in real time, in the interests of better protecting populations from the aftershock of financial and political crises. Global Pulse is a strong advocate of big data for development and humanitarian purposes.[17]

Monetizing network data assets

Some telecommunications operators have started exploiting aggregated customer data as a source of income by providing analytics on anonymized datasets to third parties. Long used exclusively for network management, billing and meeting lawful intercept requirements[18], communications metadata – data containing information on who sent a message, who received it, and when and where it was sent – may represent yet another way for telecoms players to capitalize on big data during planning, rollout, operation and upgrade phases of network infrastructure deployments.

[13] Orange, http://www.orange.com/en/D4D/Data-for-Development/

[14] http://arxiv.org/abs/1210.0137

[15] OmniTRANS, http://www.omnitrans-international.com/en/general/news/2013-07-04-using-big-data-in-transport-modelling-

[16] United Nations Global Pulse, http://www.unglobalpulse.org/

[17] NY Times, http://www.nytimes.com/2013/08/08/technology/development-groups-tap-big-data-to-direct-humanitarian-aid.html

[18] ITU, http://www.itu.int/oth/T2301000006/en

By extracting detailed traffic information in real time, network analytics help providers to optimize their routing network assets and to predict faults and bottlenecks before they cause any harm. Based on customer value and behavior metrics, the customer may dynamically be offered personalized solutions to respond to such situations. Combined real-time network insights and complete customer profiles add value with tailor-made offerings that increase revenue opportunities and attract and retain customers. Network analytics are also an important means to detect and mitigate denial of service (DoS) attacks.

Box 1: Big data to revisit the late night bus routes

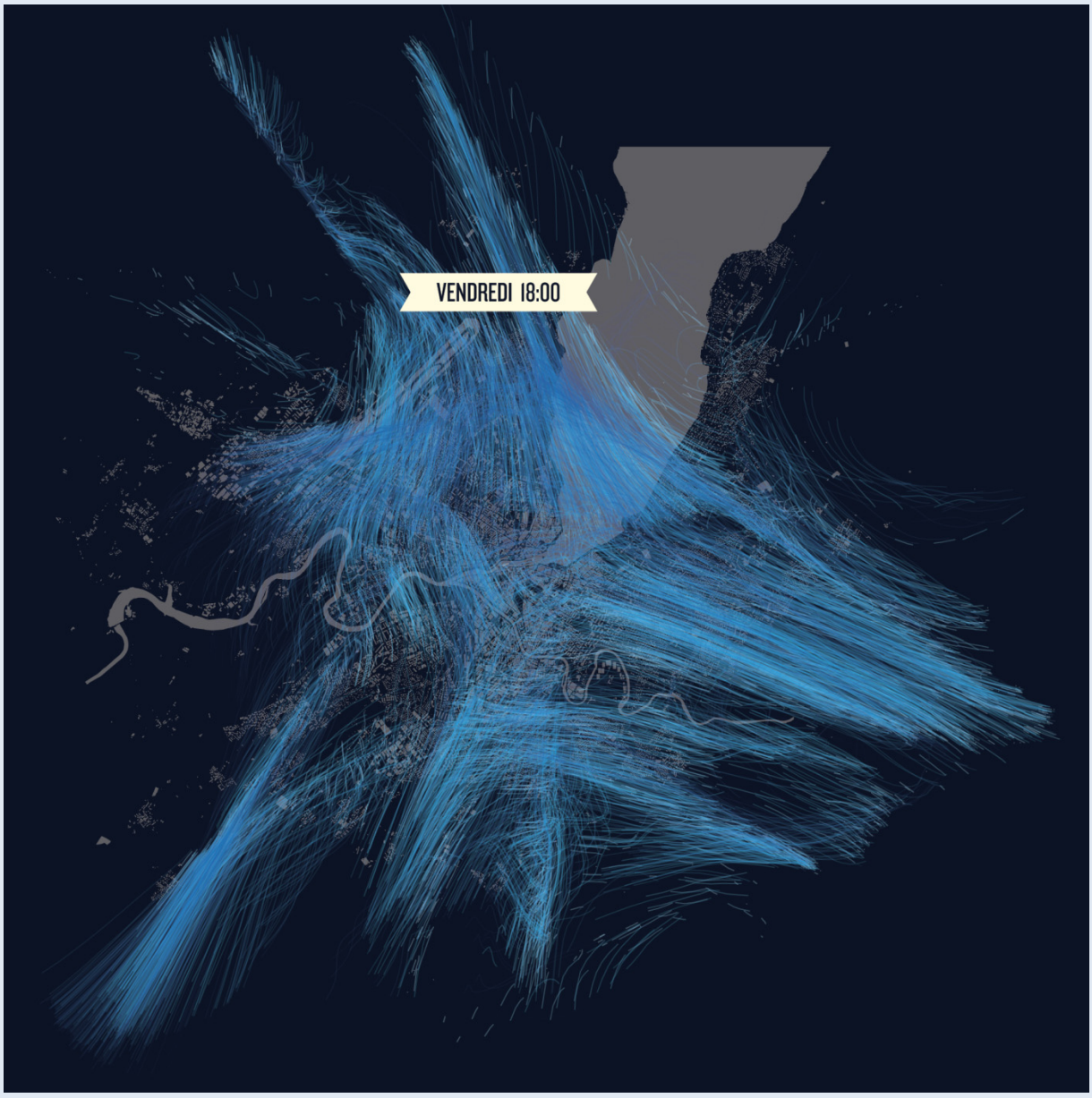
Korea Telecom and the City of Seoul have worked together to enhance the quality of public services using KT's Big Data and the city's public data, a project awarded recognition by President Park Geun-hye. Seoul is seeking to satisfy growing demand for public transport. Previously, night bus routes were designed by reference to daytime bus timetables, but did not reflect population movements by night. KT analyzed the movement of citizens around the city at night based on localized call data, and found the specific areas most frequented at night. In terms of volume, over 300 million Call Detail Records (CDR) data were analyzed for this project, combined with a variety of Seoul's public data. Weighted distances were calculated between the center points of populated areas, with the relative popularity ranking determining the primary stops. These results were then related to a heat map of the floating population, grouped by zones. This analysis established the optimal location of night bus stops that satisfy the most number of citizens, ensure citizens' safe journeys, provide economical transportation, and maximize the usage of public transportation. Based on the results, bus routes were changed to include popular new stops (e.g., Konkuk University), avoid stops little used at night (e.g., Seoul Art Center) or use routes that are congested by day (e.g., Namsan Tunnel is easily used at night).

Big data is suitable for use with public services, because it is based on mass analysis of public transport, avoiding issues with privacy and the use of personal data. Better decisions on public transport can be made, justified by evidence for improving the efficiency of the service, transparency, choice and accountability. As a result, seven more night-bus routes have been added to the original city's plan, so citizens can pay only $2 to travel home, rather than $20 for a taxi. As a result of this project, it is hoped public transport can be made more useful and efficient, and that consumers will reap real savings.

Source: Korea Telecom

Box 2: The dynamic dimension of Geneva

Each day Swisscom subscribers in the city of Geneva generate approximately 15 million connections from 2 million phone calls. These *digital traces* offer new insights into the city's movements, which are of great interest both from an economic and political perspective. An impressive visualization of the data is available online, at http://villevivante.ch/, and on display in ITU's ICT Discovery museum in Geneva.



The blue lines represent mobility traces of mobile subscribers in Geneva on a Friday evening

Source: http://villevivante.ch/

3. What makes data big – characteristics of big data

Definitions of big data are somewhat vague, but as different as the application areas described above may be, there exist common characteristics which help in describing big data. Four *V*s are often used to characterize different aspects of big data[19]:

1) Volume: *Data anytime, anywhere, by anyone and anything*

Volume may be the most compelling attraction of big data analytics. Comparing the effectiveness of a treatment on a population-wide base, considering thousands of factors, yields far better results than would the same analysis for a dataset of 100 patients.

How '*big*' exactly the data has to be to qualify as 'big data' is not specified. It is estimated that 90 per cent of the data in the world today has been created in the last two years[20], with machines and humans both contributing to the data growth.

The example of CERN has demonstrated that volume can present an immediate challenge to conventional resources, and that volume calls for scalable storage and capacity for distributed processing and querying.

2) Velocity: *Every millisecond counts*

The speed of decision making – the time taken from data input to decision output – is a critical factor in the big data discussion.[21] Emerging technologies are capable of processing vast volumes of data in real or near real time, increasing the flexibility with which organizations can respond to changes in the market, shifting customer preferences or evidence of fraud. Big data systems also need to be capable of handling and linking data flows entering at different frequencies. Long championed by high-frequency traders in the financial services market, the race for velocity and tight feedback loops is a key part of gaining competitive advantage in a number of industries.

3) Variety: *The reality of data is messy*

Big data includes any type and structure of data (see Box 3) – text, sensor data, call records, maps, audio, image, video, click streams, log files and more. Source data can be diverse, and it may require time and effort to shape it into a form fit for processing and analysis. The capacity of a system to analyze a variety of source data is crucial as it can yield insights not achievable by consulting one type of data in isolation.

4) Veracity: *Data in doubt*

How accurate or certain is the data upon which we intend to build crucial decisions? Is some data (e.g., sensor data) more trustworthy than other data (e.g., social media data such as a tweet)? Influenced by the three previous *V*s, big data tends to hold a lot of uncertainty attributed to data inconsistency, incompleteness, ambiguities and latency. Of course, the level of uncertainty and imprecision may vary, but it must be factored in. Poor data quality constitutes a cost factor. A system therefore needs capabilities to distinguish, evaluate, weigh or rank different datasets in order to maintain veracity.

[19] Gartner refers to 3 *V*s (volume, velocity, variety), others include *value* as fourth or fifth *V* in order to highlight the increasing socioeconomic value obtained from exploiting big data as a factor of production, like physical or human capital.

[20] IBM, http://www-01.ibm.com/software/data/bigdata/

[21] O'Reilly, http://strata.oreilly.com/2012/01/what-is-big-data.html

Box 3: Variety of data: Structured, semi-structured, and unstructured data

An often-cited statistic is that 80 per cent of data is unstructured[22], be it in emails; word processor, spreadsheet and presentation files; audio, video, sensor and log data; or external data such as social media feeds. Unstructured means there is no latent meaning attached to the data in a way that a computer can understand what it represents. In contrast, structured data has semantic meaning, making it easier to be understood. For instance, databases represent data organized as rows and columns and allow computer programs to understand the meaning of the data with a schema, i.e., a specification of the facts that can enter the database, or those of interest to the possible end-users.

Semi-structured data is a form of structured data that does not conform to the formal structure of data models associated with relational databases, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. XML and other markup languages are often referred to as semi-structured. There is structure in the relationships of the data, but the data is not structured with regard to the meaning of that data. Only a schema attaches meaning to the data.

Structured data is simpler to process because more information is available to the program beforehand in order for it to determine the data's meaning. This approach is more efficient as opposed to spending compute cycles to figure it out. Much of the growth of data in today's age, however, is that of unstructured data, making it critical for systems to be able to process it efficiently and to correctly determine the meaning contained within it. For example, emails and text messages as well as audio and video streams are some of the largest categories of unstructured data today. This type of unstructured data continues to grow unabated, making the efficient processing of it critical to the continued success of business analytic processing systems.

Adopted from http://www.ibm.com/developerworks/xml/library/x-datagrowth/

The appearance and weighting of any of the Four *V*s is highly application- and purpose-specific. Some applications may focus on only a small amount of data but process and analyze real-time streams of many different data types. In another scenario, insight may be gained by, on occasion, processing batches of vast volumes of unstructured data. Combined, these characteristics represent big data's transformational capabilities, but also point to some of the challenges to be discussed later in this report.

Table 1 attempts to summarize the key features of big data's Four *V*s: volume, velocity, variety and veracity.

[22] Breakthrough Analysis, http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

Table 1: Summary of four big data characteristics

Characteristic	Description	Attributes	Drivers
Volume	The amount of data generated or data intensity that must be ingested, analyzed and managed to make decisions based on complete data analysis.	- Exabyte, zettabyte, yottabyte, etc.	- Increase in data sources - Higher resolution sensors - Scalable infrastructure
Velocity	How fast data is being produced and changed and the speed at which data is transformed into insight.	- Batch - Near real-time - Real-time - Streams - Rapid feedback loop	- Improved throughput connectivity - Competitive advantage - Precomputed information
Variety	The degree of diversity of data from sources both inside and outside an organization.	- Degree of structure - Complexity	- Mobile - Social media - Video - Genomics - M2M / IoT
Veracity	The quality and provenance of data.	- Consistency - Completeness - Integrity - Ambiguity	- Cost - Need of traceability and justification

Adapted from TechAmerica: Demystifying big data[23]

4. What makes data big – enablers

Not a technology in itself, big data's uptake has been driven by all components of the innovation value chain. A very simplified big data value chain would include at least four stages:

1) Data is collected where it originates. During the data-generation stage, a stream of data is created from a variety of sources: sensors, human input, etc.

2) The raw data is combined with data from other sources, classified and stored in some kind of data repository.

3) Algorithms and analytics are applied by an intelligence engine to interpret and provide utility to the aggregated data.

4) The outputs of the intelligence engine are converted to tangible values, insights or recommendations.[24]

[23] TechAmerica Foundation: "Demystifying big data," http://www.techamericafoundation.org/official-report-of-the-techamerica-foundations-big-data-commission

[24] Innosight, http://www.innosight.com/innovation-resources/strategy-innovation/winning-within-the-data-value-chain.cfm

Stage 1: The origins of data

Ubiquitous broadband connectivity and M2M are among the key drivers of data volume and variety. Open data can contribute to our fourth *V*: Veracity.

Big today is normal tomorrow

According to the vision of network equipment manufacturer Ericsson, more than 50 billion devices will be connected by 2020, independent of what and where they are, enabling an Internet of Things and Places. All of these devices will measure, sense, generate and communicate data of some size and structure (e.g., a smart meter sends an energy value to the utility provider once a month; a CCTV camera sends a 24/7 live video feed to the control room).

ITU estimates almost 7 billion mobile-cellular subscriptions worldwide (see Figure 2), and each of these subscribers is both a data creator and consumer. Today almost 3 billion people use the Internet and mobile-broadband subscriptions, in particular, have rocketed upwards from 268 million in 2007 to 2.1 billion in 2013. Each of these consumers contribute to the data deluge, with SMS; calls; photos, videos and messages posted on social media sites; emails; searches; clicks on links and ads; online shopping and mobile payments; or location traces left by GPS-enabled smartphones and WiFi network logins.

The rapid uptake of broadband also enables the use of data-heavy services (video, video telephony, etc.) and is reflected in the amount of data being transported over IP networks. And these figures are quickly evolving. Proclaimed only five years ago, the *exabyte era*[25] (10^{18} bytes or over 36,000 years' worth of HD video) was quickly superseded by the *zettabyte era*[26] (10^{21} bytes or a trillion gigabytes), and it is safe to assume that it is just a matter of time before the dawn of the *yottabyte era*. Software-defined networking (SDN) – a space currently playing host to much innovation and standardization – involves the abstraction of lower-level network functionality through algorithms and virtual services, and it is a promising approach to satisfying the demands of big data.

Data philanthropy and open data

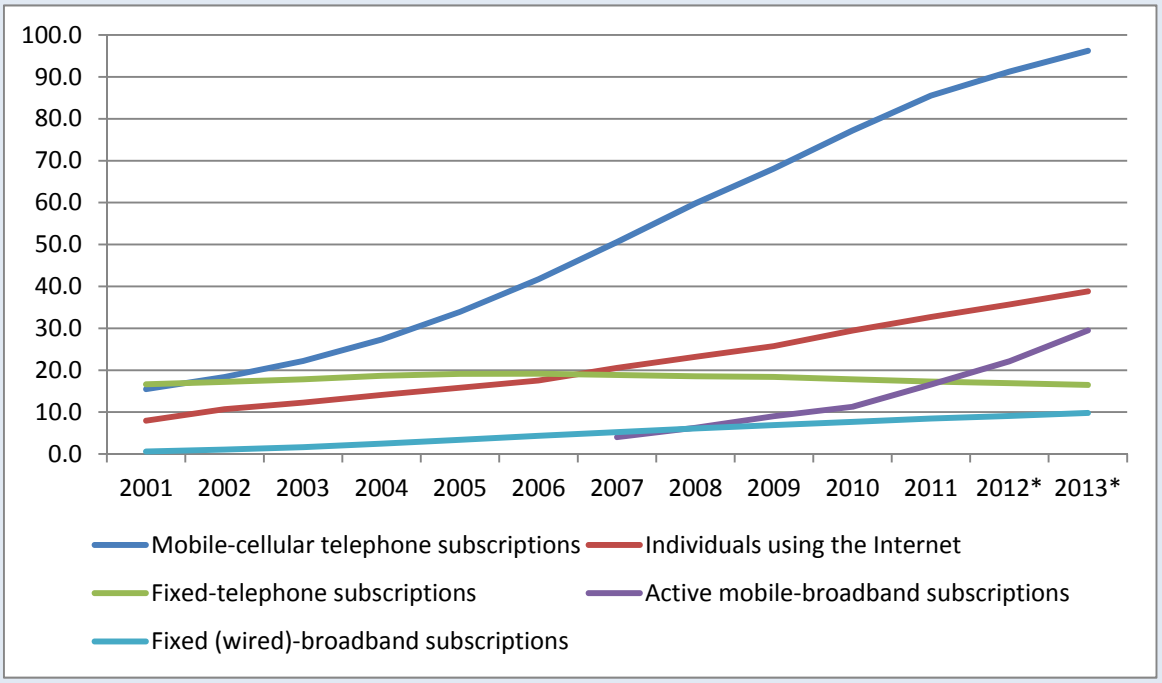
Global Pulse, the UN 'big data for development' initiative, relies on *data philanthropy* where telecommunications providers share large customer datasets. Data philanthropy requires precise and suitable legal frameworks, ethical guidelines and trusted technology solutions to safeguard the highest standards for the preservation of privacy[27] – a critical challenge we will address in the following chapter.

[25] Cisco SP360 Blog: "Welcome to The Exabyte Era," 27 February 2008, http://blogs.cisco.com/sp/welcome_to_the_exabyte_era/

[26] Cisco SP360 Blog: "The Zettabyte Era is Upon Us," 31 May 2012, http://blogs.cisco.com/sp/the-zettabyte-era-is-upon-us/

[27] United Nations Global Pulse, http://www.unglobalpulse.org/data-philanthropy-where-are-we-now

Figure 2: Global ICT takeoff (2001-2013)



*Per 100 inhabitants, * estimate*

Source: ITU World Telecommunication/ICT Indicators database, http://itu.int/en/ITU-D/Statistics/

Governments, public-sector institutions and UN organizations can make significant contributions to the public interest by advocating for open data and releasing their datasets by default. The Open Knowledge Foundation defines openness in the context of open data: *"A piece of data or content is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike."*[28]

The political leaders of the G8 – Canada, France, Germany, Italy, Japan, Russia, the United Kingdom and the United States of America – have set out five strategic principles aimed at unlocking the economic potential of open data, supporting innovation, and providing greater accountability. These principles include an expectation that all government data will be published openly by default, alongside principles to increase the quality, quantity and re-use of the data released from 14 high-value areas – from education to transport, health, crime and justice.[29] Government data portals have emerged in all parts of the world and a number of intergovernmental organizations have responded to calls for open data with foresight and initiative (see Table 2).

[28] http://opendefinition.org/

[29] https://www.gov.uk/government/publications/open-data-charter

Table 2: Examples of open data portals

State / organization	Website
Belgium	http://data.gov.be/
Ghana	http://data.gov.gh/
India	http://data.gov.in/
Kenya	https://www.opendata.go.ke/
Morocco	http://data.gov.ma/
Russia	http://opengovdata.ru/
United Arab Emirates	http://www.government.ae/web/guest/uae-data
United Kingdom	http://data.gov.uk/
United States of America	http://www.data.gov/
European Union	http://open-data.europa.eu/
Organization for Economic Co-operation and Development (OECD)	http://stats.oecd.org/
United Nations (UN)	http://data.un.org/
United Nations High Commissioner for Refugees (UNHCR)	http://data.unhcr.org/
World Bank	http://data.worldbank.org/

Note: This list is by no means exhaustive.

Stage 2: Aggregating, classifying and storing data

At stage 2, volume and velocity can quickly escalate into major challenges, as we have indicated with the example of CERN. Part of the solution is cloud computing, a model for delivering seemingly infinite computing resources on demand. Cloud computing has grown exponentially over the last five years and its pay-per-use plans have been adopted by organizations of all sizes. Cloud services give their customers significantly more flexibility and scalability, not only in terms of data storage, but also in how processing capacity can be scaled up or down as needed.

Distributed file systems, programming models and scalable high-performance databases are big data core technologies exploited in stages 2 and 3 of the simplified big data value chain.

Free and open source software Hadoop, MongoDB, Spark and Storm can be counted among the most prominent enablers. Hadoop, a free and open-source implementation of the Map/Reduce programming model[30], is used to process huge datasets (petabytes of data) in a scalable manner by commissioning parallel processing capabilities which move data subsets to distributed servers (see Box 4). In addition, Hadoop provides a distributed file system (HDFS) that can store data on

[30] The name Hadoop was inspired by a toy elephant, http://www.nytimes.com/2009/03/17/technology/business-computing/17cloud.html

thousands of compute nodes, providing very high aggregate bandwidth across the whole cluster. Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework to re-assign the tasks among other nodes.[31]

CERN uses MongoDB as the primary back-end for the data aggregation of one of the LHC particle detectors.[32] Data and metadata for the detector come from many different sources and are distributed in a variety of digital formats. It is organized and managed by constantly evolving software, which makes use of both relational and non-relational data sources. The back-end allows researchers to query data via free text-based queries, then it aggregates the results from across distributed providers and represents the data in a defined format, all the while preserving integrity, security policy and data formats. MongoDB is a so-called 'NoSQL' database which refers to mechanisms for storage and retrieval of data in less constrained consistency models than traditional relational databases. This aids simplicity of design, scalability, control, and performance in terms of latency and throughput.

Stage 3: Making data useful

The intelligence engine is highly application-specific and generates real intelligence that can be used to make decisions, either by graph analysis, mathematical modeling or through some form of simulation. In the transport sector, for instance, it would have the task of interpreting traffic data supplied by road-side sensors and probes in order to provide traffic information usable by GPS devices. Amazon's patented recommendation engine[33], in basic terms, aggregates data about your browsing, purchasing and reading habits and then, based on a dataset of all the other customers with similar histories, makes extrapolations about what you would like to read next. Your favorite social network site analyses your social graph, a mathematical structure used to model pairwise relations between objects, to suggest friends and content to you.

Stage 4: Information presented intuitively

Sticking to the example of the book seller, the intelligence is presented in different forms and on various occasions, such as in notifications that "*customers who bought this item also bought*" or in personalized emails sent out with tailored purchase recommendations. The last stage in the simplified big data value chain can take many other forms. It may represent a direct user experience, such as a warning, notification or visualization[34], or could result in communication with a smart system, for example by triggering a financial transaction or by adjusting traffic signals to reflect the actual traffic flow.

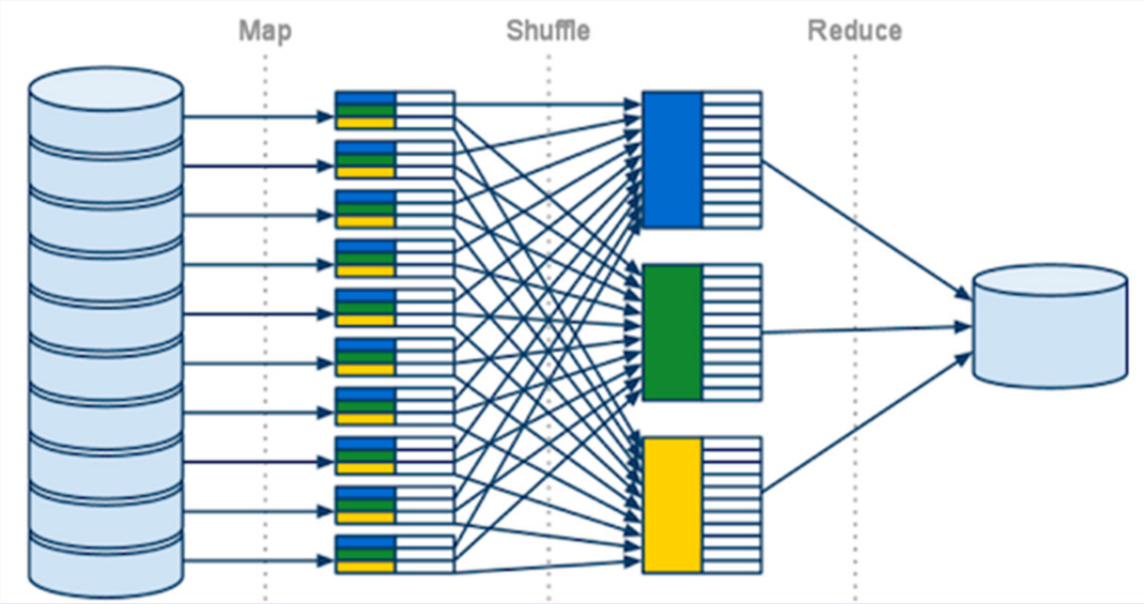
[31] http://wiki.apache.org/hadoop/ProjectDescription

[32] mongoDB, http://blog.mongodb.org/post/660037122/holy-large-hadron-collider-batman

[33] EPO, http://worldwide.espacenet.com/publicationDetails/biblio?CC=US&NR=7113917B2&KC=B2&FT=D

[34] See, for instance, Information is Beautiful, a website *"dedicated to distilling the world's data, information and knowledge into beautiful, interesting and, above all, useful visualizations, infographics and diagrams"* http://www.informationisbeautiful.net/

Box 4: Processing datasets in stages: Map/Reduce model in Google App Engine



- *Reading input:* An input reader reads data from storage (or some other source) and passes it to the next stage. You can choose a pre-defined reader from a list of choices. E.g., one input reader reads all datastore entities of a specified kind, passing each entity as input to the next stage.
- *Map:* You write a map function that runs once for each input value. It returns a collection of name-value pairs which are passed on to the next stage. If there are many, many input values, then this function runs many, many times. The framework divides up the input so that subsets are handled in parallel on multiple instances of your application. A typical map function could count things that occur in each input value that matches some filter.
- *Shuffle:* The Map/Reduce framework "shuffles" together the values returned by the map function. The shuffler groups together name-value pairs that have the same name and then passes those groups on to the next stage.
- *Reduce:* You write a reduce function that runs once for each "name" used in the name-value pairs. If there are many, many names then this function may run many, many times. It returns a collection of values that are passed along to the next stage.
- *Writing output:* An output writer concatenates together the outputs of the reduce functions in arbitrary order and writes them to persistent storage. You can choose a pre-defined writer from a list of choices.

Source: Google App Engine MapReduce Python Overview, https://developers.google.com/appengine/docs/python/dataprocessing/, licensed under the Creative Commons Attribution 3.0 License (http://creativecommons.org/licenses/by/3.0/)

5. Challenges and opportunities for big data adoption

The year 2013 has been a revealing year in the realm of big data analysis and visualization in mass surveillance.

Recent events have called into question the secrecy of telecommunications[35] and the efficacy of security mechanisms and standards. And while these privacy and data protection concerns affect all parts of the ICT ecosystem, they are perhaps especially relevant to big data and cloud computing.

The final section of this report outlines some of the challenges commonly associated with big data, in parallel highlighting related opportunities within the scope of ITU's activities.

Data protection, privacy and cybersecurity

Big data stands in stark contrast to data avoidance and data minimization, two basic principles of data protection. Big data facilitates the tracking of people's movements, behaviors and preferences and, in turn, helps to predict an individual's behavior with unprecedented accuracy, often without the individual's consent. For instance, electronic health records and real-time self-quantification may constitute an enormous step forward in streamlining the prescriptions of drugs or diet and fitness plans. However, that same data is viewed by many consumers as very sensitive.

Large sets of mobile call records, even when anonymized and stripped of all personal information, can be used to create highly unique fingerprints of users, which in combination with other data such as geo-located tweets or "check-ins" may help to reveal the individual.[36] Communications metadata can be useful for telecom network management and billing, but, simply put, exploiting communications metadata on people is a form of surveillance. It not only reveals fine-grained details about people, but it also exposes the relationship between interacting entities.[37, 38]

As the amount of personal data and global digital information grows, so does the number of actors accessing and using this information. Assurances must be given that personal data will be used appropriately, in the context of the intended uses and abiding by the relevant laws.

A closely related concern is cybersecurity. Hardly a week goes by without serious breaches of data, where personal data often falls into the wrong hands.[39] A range of technical solutions (e.g., encryption, VPNs, firewalls, threat monitoring and auditing) can help in managing data privacy and mitigating security risks. Threats and risks need to be reassessed in view of big data, adapting technical solutions in response. The time is ripe to review information security policies, privacy guidelines, and data protection acts.

Legal and regulatory considerations

Who owns my call records? What rights come attached with data, or who and what defines "fair use"? Who is liable for breach of personal data, or decisions based on inconsistent or incomplete datasets that give rise to negative consequences for the individual? These are just a few of the questions to be addressed by policymakers, regulators, industry and consumers if we are to grasp the full potential of big data.

---

[35] See Constitution of ITU: Chapter VI - General Provisions Relating to Telecommunications, Article 37 "Secrecy of Telecommunications," http://www.itu.int/net/about/basic-texts/constitution/chaptervi.aspx#_Toc36522688

[36] Fast Company, http://www.fastcompany.com/3007645/location-location-location/mobile-phones-have-fingerprints-too

[37] Bruce Schneier, https://www.schneier.com/blog/archives/2013/09/metadata_equals.html

[38] NY Times, http://bits.blogs.nytimes.com/2013/06/09/intelligence-agencies-and-the-data-deluge/

[39] See http://spectrum.ieee.org/blog/riskfactor for reports of cybercrime and IT hiccups

Big data and standards

Achieving the big data goals set out by business and consumers will require the interworking of multiple systems and technologies, legacy and new. Technology integration calls for standards to facilitate interoperability among the components of the big data value chain. For instance, UIMA, OWL, PMML, RIF and XBRL are key software standards that support the interoperability of data analytics with a model for unstructured information, ontologies for information models, predictive models, business rules and a format for financial reporting.[40]

The standards community has launched several initiatives and working groups on big data. In 2012, the Cloud Security Alliance established a big data working group[41] with the aim of identifying scalable techniques for data-centric security and privacy problems. The group's investigation is expected to clarify best practices for security and privacy in big data, and also to guide industry and government in the adoption of those best practices. The U.S. National Institute of Standards and Technology (NIST) kicked-off its big data activities with a workshop in June 2012 and a year later launched a public working group. The NIST working group intends to support secure and effective adoption of big data by developing consensus on definitions, taxonomies, secure reference architectures and a technology roadmap for big data analytic techniques and technology infrastructures.[42] ISO/IEC JTC1's data management and interchange standards committee (SC32) has initiated a study on next-generation analytics and big data.[43] The W3C has created several community groups on different aspects of big data.[44]

High-throughput, low-latency, secure, flexible and scalable network infrastructure

We have characterized broadband, M2M, cloud computing, advances in data management and the rise of social media as drivers of big data – not one technology in itself, but a natural outcrop of ICT innovation and development that cuts across its many sub-disciplines. There is rapid progress in all these areas, and we see that the latest generations of products and solutions feature unprecedented performance and scalability in handling, processing and making sense of data while meeting the growing demands of Volume, Velocity, Variety and Veracity.

At present, ITU's standardization activities address individual infrastructure requirements, noting existing work in domains including optical transport and access networks, future network capabilities (e.g., software-defined networks), multimedia and security. A review of this work from the angle of data-driven applications has yet to be undertaken but could yield significant results in the big data context.

A work item has been initiated to study the relationship between cloud computing and big data in view of requirements and capabilities.[45] The recently determined Recommendation ITU-T X.1600, "Security framework for cloud computing"[46], matches security threats with mitigation techniques,

[40] IBM, http://www.ibm.com/developerworks/xml/library/x-datagrowth/

[41] CSA, https://cloudsecurityalliance.org/research/big-data/

[42] NIST, http://bigdatawg.nist.gov/

[43] JTC1 SC32, http://www.jtc1sc32.org/doc/N2351-2400/32N2388b-report_SG_big_data_analytics.pdf

[44] W3C, http://www.w3.org/community/custexpdata/ and http://www.w3.org/community/bigdata/

[45] ITU, http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9853

[46] ITU, http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9404

and the future standardization of the described threat-mitigation techniques is expected to incorporate big data use cases. A previous report in the Technology Watch series advocated the use of privacy-enhancing technologies as means to implement the 'privacy by design' principle, which is of course of great interest to big data applications.[47]

Aggregated datasets and anonymization

With a membership comprising governments, telecommunications operators, equipment manufacturers and academia and research institutes from around the world, ITU is ideally positioned to review current practices in the use of aggregated datasets such as mobile phone records and, in so doing, to develop related technical standards and policies. Progress is being made in techniques for data anonymization, the process of altering data in a way that prevents the identification of key information (e.g., personally identifiable information).[48] These techniques need to be studied and evaluated considering current (and future) data aggregation and analysis capabilities.

Network data analytics

It is further recommended to study, in greater detail, the requirements, architectures, interfaces and types of communications metadata and network metrics best suited to real-time network analytics and related value-added services and use cases.

Platform interop in the verticals

ITU has been accelerating its efforts to increase interoperability in electronic health applications, in areas such as the exchange of health data[49] and the design of personal health systems.[50] With the boom of personal and wearable 'connected health' and fitness products in mind, if a smart wristband could exchange data with a smartwatch of a different make (uninhibited by vendor or manufacturer boundaries), big data would benefit with the ability to pull together and securely integrate data collected on different devices.[51]

The home automation sector faces a similar dilemma. When Amazon launched its marketplace for home automation products[52], such as programmable thermostats, light bulbs and connected door locks, it had to guide customers to ensure that they bought interoperable products. This concerns communications protocols, gateways and data integration platforms alike, and care should be taken to avoid vendor lock-ins that deny consumers the use of the latest data-driven third party applications to reduce energy consumption or heat the house more efficiently.

[47] ITU, http://www.itu.int/en/ITU-T/techwatch/Pages/cloud-computing-privacy.aspx

[48] Intel, http://www.intel.com/content/dam/www/public/us/en/documents/best-practices/enhancing-cloud-security-using-data-anonymization.pdf

[49] ITU, http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9637

[50] ITU, http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9636

[51] VentureBeat, http://venturebeat.com/2013/09/03/future-wearables-samsung/

[52] Amazon.com, http://www.amazon.com/b/?node=6563140011&ref=ha_van

Multimedia analytics

After doubling the efficiency of its Emmy-winning predecessor, ITU-T H.265 is on track to become the Web's leading video codec. Considering multimedia's significant share in total IP traffic, the automatic analysis of digital image, audio and video data is an area to be monitored closely. Multimedia content analysis allows for automatic and rapid extracting of events, metadata or other meaningful information, e.g., for in-video or image search, or motion analysis. Real-time multimedia content analysis is a challenging big data task – will the future generations of codecs be ready to support it?[53]

Open data standards

The open data movement is maturing, in highly industrialized as well as emerging economies. With a number of interoperability and policy issues at hand, ITU is in an opportune position to embrace and advance the cause of open data in partnership with the many open data champions within and outside its membership. From the standards angle, this could include, *inter alia*, the development of requirements for data reporting and mechanisms for publication, distribution and discovery of datasets.

[53] GoPivotal, http://blog.gopivotal.com/features/large-scale-video-analytics-on-hadoop

ITU-T Technology Watch surveys the ICT landscape to capture new topics for standardization activities. Technology Watch Reports assess new technologies with regard to existing standards inside and outside ITU-T and their likely impact on future standardization.

Previous reports in the series include:

ICTs and Climate Change
Ubiquitous Sensor Networks
Remote Collaboration Tools
NGNs and Energy Efficiency
Distributed Computing: Utilities, Grids & Clouds
The Future Internet
Biometrics and Standards
Decreasing Driver Distraction
The Optical World
Trends in Video Games and Gaming
Digital Signage
Privacy in Cloud Computing
E-health Standards and Interoperability
E-learning
Smart Cities
Mobile Money
Spatial Standards

http://www.itu.int/ITU-T/techwatch

Printed in Switzerland
Geneva, 2013
Photo credits: Shutterstock®