International Transport Forum

CPB Corporate Partnership Board



Big Data and Transport

Understanding and assessing options



Corporate Partnership Board Report

OECD

Big Data and Transport

Understanding and assessing options



Corporate Partnership Board Report

International Transport Forum

The International Transport Forum at the OECD is an intergovernmental organisation with 54 member countries. It acts as a think tank for transport policy and organises the Annual Summit of transport ministers. ITF is the only global body that covers all transport modes.

ITF works for transport policies that improve peoples' lives. Our mission is to foster a deeper understanding of the role of transport in economic growth, environmental sustainability and social inclusion and to raise the public profile of transport policy.

ITF organises global dialogue for better transport. We act as a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision makers and civil society.

Our member countries are: Albania, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, China (People's Republic of), Croatia, Czech Republic, Denmark, Estonia, Finland, France, Former Yugoslav Republic of Macedonia, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Italy, Japan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Montenegro, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom and United States.

Disclaimer
This work is published under the responsibility of the Secretary-General of the International Transport Forum. Funding for this work has been provided by the ITF Corporate Partnership Board. The opinions expressed and arguments employed herein do not necessarily reflect the official views of International Transport Forum member countries. 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.

www.internationaltransportforum.org

Table of contents

Executive summary 5

Introduction 7

1. Big Data, big changes 9

2. The Big Data lifecycle 11

Data acquisition and recording **12**

Data production: Digital vs. analogue **12**

Need for transparency and metadata on data provenance **18**

Data extraction, cleaning, annotation and storage **21**

Integration, aggregation and fusion **22**

Analysis, modelling and visualisation **24**

Data mining 24

Modelling 25

Visualisation and dissemination 26

3. Big Data, personal data and privacy 33

Personal data protection frameworks **35**

OECD personal data protection framework and guidelines 35

EU personal data protection frameworks 36

APEC and other Asian personal data protection frameworks 38

United States' personal data protection frameworks 39

Private-sector initiatives 40

"Privacy by Design" 43

Privacy and location-based data **45**

"You are where you've been" 49

"We know where you've been" 51

"It's hard to hide where you've been" 52

Anonymising location and trajectory data: four suggestions **52**

Bibliography 59

Figures

1. Data size scale 10
2. Hype cycle for emerging technologies 11
3. Big Data collection and analysis lifecycle 12
4. Views of Waze, a community-based traffic and navigation application 18
5. Optimisation strategy for taxi sharing 26
6. Pick-up and drop-off points of all 170 million taxi trips over a year in New York City 27
7. Traffic incidents visualised using the Traffic Origins approach 28
8. Analysis and visualisation of labour market access in Buenos Aires 29
9. Individuals' time-space trajectories are powerful identifiers 51

Executive summary

This report examines issues relating to the arrival of massive, often real-time, data sets whose exploitation and amalgamation can lead to new policy-relevant insights and operational improvements for transport services and activity. It is comprised of three parts. The first section gives an overview of the issues examined. The second broadly characterises Big Data, and describes its production, sourcing and key elements in Big Data analysis. The third section describes regulatory frameworks that govern data collection and use, and focuses on issues related to data privacy for location data.

What we found

The volume and speeds at which data today is generated, processed and stored is unprecedented. It will fundamentally alter the transport sector

The combination of low-cost and widespread sensing (much of it involving personal devices), the steep drop in data storage costs and the availability of new data processing algorithms improves our ability to capture and analyse more detailed representations of reality. Today these representations augment traditional sources of transport data collection. In the future they will likely replace them.

Sensors and data storage/transmission capacity in vehicles provide new opportunities for enhanced safety

Work is underway to harmonise standards regarding these technologies and communications protocols in order to accelerate safety improvements and lower implementation costs for conventional and, increasingly, automated vehicles.

Multi-platform sensing technologies are now able to precisely locate and track people, vehicles and objects

Locating and tracking individuals at precisions up to a few centimetres in both outdoor and indoor environments is feasible and will likely become standard – at least in urban areas – as location-sensing technologies become omnipresent. The widespread penetration of mobile, especially smartphone, technology makes this possible in ways not previously achievable. The location technologies deployed in today's mobile phones are increasinglybeing built into vehicles, enabling precise and persistent tracking.

The fusion of purposely-sensed, opportunistically-sensed and crowd-sourced data generates new knowledge about transport activity and flows. It also creates unique privacy risks

When combined, these data reveal hitherto unsuspected or unobserved patterns in our daily lives. They can be used to the benefit of both individuals and society. There is also the risk that insights derived from these patterns may open new avenues for misuse and potential manipulation of individuals and their behaviour. The knowledge derived from this fusion may not have been anticipated by data collectors at the time of collection and the use of these insights may not have been anticipated or communicated to people who are the object of that data.

Location and trajectory data is inherently personal in nature and difficult to anonymise effectively

Tracking and co-locating people with other people and places exposes a daily pattern of activity and relationships that serve as powerful quasi-identifiers. Trajectories are as unique as fingerprints and though many techniques exist to de-identify this data, doing so effectively, while retaining sufficient detail for useful analysis, is not easy.

Data protection policies are lagging behind new modes of data collection and uses. This is especially true for location data

Rules governing the collection and use of personal data (e.g. data that cannot be de-identified) are outdated. Data is now collected in ways that were not anticipated by regulations, and authorities have not accounted for the new knowledge that emerges from data fusion. A split has emerged between those who

would seek to retain prior notification and consent frameworks for data collection and those who would abandon these in order to focus only on specifying allowable uses of that data.

Policy insights

Road safety improvements can be accelerated through the specification and harmonisation of a limited set of safety-related vehicle data elements

Technologies like E-call, E-911 and vehicle data black boxes provide post-crash data well suited for improving emergency services and forensic investigations. Much more vehicle-related data is available and, if shared in a common format, could enhance road safety. Further work is needed to identify a core set of safety-related data elements to be publicly shared and to ensure the encryption protocols necessary to secure data that could compromise privacy.

Transport authorities will need to audit the data they use in order to understand what it says (and what it does not say) and how it can best be used

Big Data in transport is not immune from small data problems – especially those relating to statistical validity, bias and incorrectly imputed causality. Transport authorities will need to ensure an adequate level of data literacy for handling new streams of data and novel data types. Ensuring robust and persistent metadata with harmonised provenance will facilitate data usability audits. Big Data is often not clean. Lack of data quality may mean significant upfront costs to render the data useable. This should be factored into decision making processes.

More effective protection of location data will have to be designed upfront into technologies, algorithms and processes

Adapting data protection frameworks to increasingly pervasive and precise location data is difficult, largely because data privacy has not been incorporated as a design element from the outset. Both voluntary and regulatory initiatives should employ a “Privacy by Design” approach which ensures that strong data protection and controls are front-loaded into data collection processes. Technological advances including the arrival of system-on-a-chip sensors can aid this by allowing on-the-fly data encryption. Other advances could include protocols allowing for citizens to control and allocate rights regarding their data. Failing to ensure strong privacy protection may result in a regulatory backlash against the collection and processing of location data. This could hamper innovation, reduce consumer welfare and curb the social and economic benefits the use of such data delivers.

New models of public-private partnership involving data-sharing may be necessary to leverage all the benefits of Big Data

An increasing amount of the actionable data pertaining to road safety, traffic management and travel behaviour is held by the private sector. Yet public authorities are still, and will likely continue to be, mandated to provide essential services. Innovative data-sharing partnerships between the public and private sectors may need to go beyond today’s simple supplier-client relationship. These new arrangements should not obviate the need for market power tests, cost-benefit assessment and public utility objectives.

Data visualisation will play an increasingly important role in policy dialogue

Effective data visualisations can quickly communicate key aspects of data analysis and reveal new patterns to decision makers and the public. Public agencies will need to be able to handle the visual language of data as effectively as they handle written and spreadsheet-based analysis.

Introduction

This report was written on the basis of desk research and expert input and interviews among practitioners and researchers from both the public and private sectors. It investigates how mobility-related data generation, collection and use are rapidly evolving, reviews existing data protection frameworks and highlights several key strategic areas where data privacy will have to be improved if long-term innovation benefits are to be realised.

The principal author of this report was Philippe Crist of the International Transport Forum with substantial inputs provided by Emma Greer and Carlo Ratti of Carlo Ratti Associati, Paulo Humanes of PTV AG and Gilbert Konzett, Jasja Tijink, Diego Figuero and Richard Lax, all of Kapsch TrafficCom. The report benefitted from valuable inputs provided by José Viegas and Antigone Lykotrafiti. The project was coordinated by Philippe Crist and Sharon Masterson of the International Transport Forum.

1. Big Data, big changes

Never before has so much timely information about events, people and objects been so widely and quickly available. One recent estimate puts the total size of the "digital universe" – comprised of digital content spanning from photographs, movies and surveillance video feeds, data produced and sent by sensors and connected devices, internet content, email, sms, audio streams to phone call metadata – at 4.4 zettabytes (see Figure 1) in 2013. Doubling every two months, the size of the digital universe is projected to grow to 44 zettabytes in 2020 (IDC, 2014). For reference, the total amount of visual information conveyed by the eyes to the brains of the entire global population in 2013 amounted to approximately 1 zettabyte per day. These estimates represent a staggering amount of data, and a significant portion relates to events and people (credit card and payment transactions, surveillance video, vehicle sensor outputs, Wi-Fi access signals, volunteered text and imagery on social networks). This datacan be used to better understand, anticipate or manipulate human behaviour.

The acceleration in both the growth and velocity of exploitable and often open data will trigger significant and disruptive change across a number of sectors – including transport. Compelling cases have been made for the value of Big Data analytics for urban planning (via the convergence of high definition geographic data with information regarding the observed or interpreted use of urban space by citizens), intelligent transport (via visualisation and analysis of the real-time usage of transport networks) and safety (via the processing of real-time data regarding vehicle operation and the surrounding environment to avoid or minimise potentially dangerous conflicts). However, it is not clear that authorities and regulation within and outside of the transport sector have kept pace with the proliferation of new, or newly available, data. Just as a better understanding of how mobility-related data can help resolve policy challenges relating to congestion and safety, for example, failure to account for the changing nature of data collection, use and access can also lead to negative outcomes - in particular regarding an unintended and unwanted erosion of privacy rights.

To be clear, the collection and exploitation of large data sets – so-called "Big Data"– is not new and is not linked to a single technological change. Rather, what has occurred is the confluence of new data collection mechanisms based on ubiquitous digital devices, greatly enhanced storage capacity and computing power as well as enhanced sensing and communication technologies. These technologiesenable near real-time use and transmission of massive amounts of data.

Some of these data streams are purpose-built to address well-defined questions and to resolve specific tasks. For instance data from automatic toll payment transponders broadcast data necessary for the processing and secure payment of road tolls. However, much of the potential value (or damage) from data lies in its combinatory use with other data sources. These data need not be well-defined or purpose-tied – and often are not. They are more akin to "digital dust" that lingers from our interactions with any number of computing systems and digital infrastructure and services.

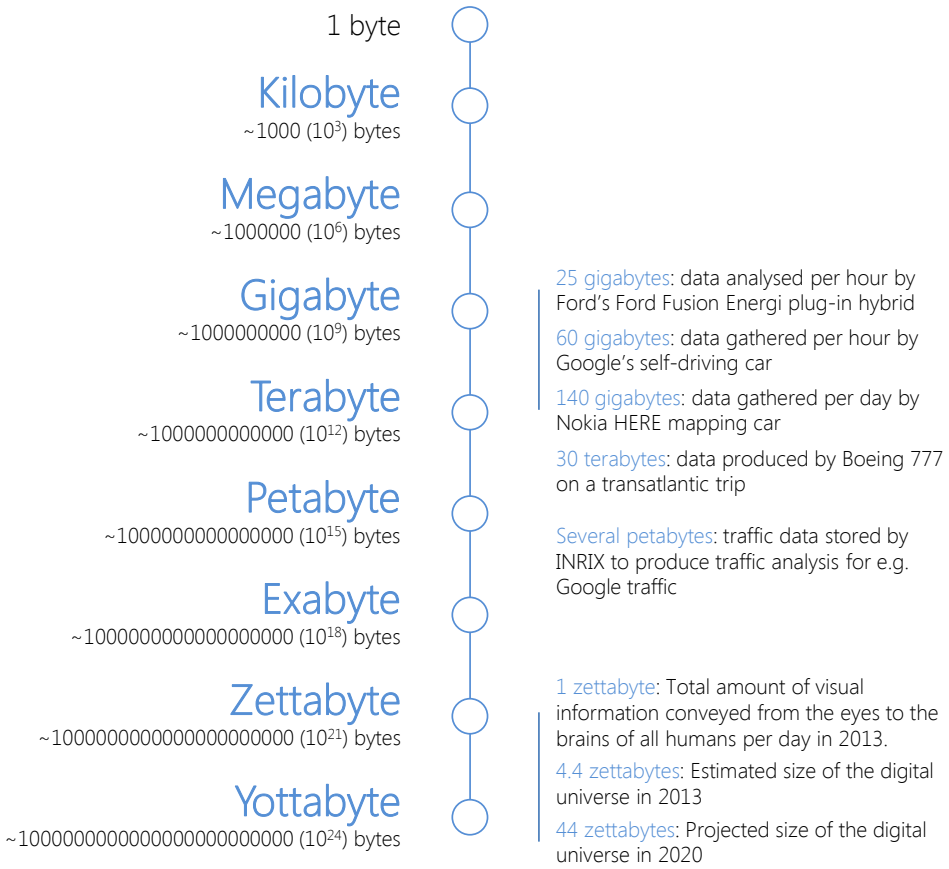
When combined, these data reveal hitherto unsuspected or unobserved patterns in our daily lives which can be used to benefit both individuals and society. There is also the risk that insights derived from these patterns may open up new avenues for misuse of data and potential manipulation of individuals and their behaviour. Big Data is seen as both an opportunity and a challenge. This is especially true for the management and governance of transport-related data.

Transport is a complex activity but at its most basic expression it is simply about connecting locations with flows. These locations may be proximate, well-connected and displaying high levels of access – as in many urban areas – or not. The flows between these locations may concern people or goods and may involve any number of vehicle types – or not, as in the case of walking.

Resolving the location-flow equation requires delivering and managing the use of infrastructure assets such as roads, bridges, tracks, airports, ports, bus stations and cycle paths – but it may also involve decisions regarding where to site activities so that the need to move is obviated. All of these decisions require information – a lot of information – regarding places, people and activities. Big Data holds much promise for improving the planning and management of transport activity by radically increasing the amount or near-real-time availability of mobility-related data. Likewise, access to more detailed and actionable data regarding the operation of vehicles and of the environment in which they operate holds much promise for improving the safety of transport.

These three fields – operations, planning and safety – are areas where authorities must critically evaluate where and how new, or newly available data and data-related insights, can improve transport policy.

Figure 1. **Data size scale**



Source: Nokia HERE, Forbes, Idealab, GE, ITF calculations.

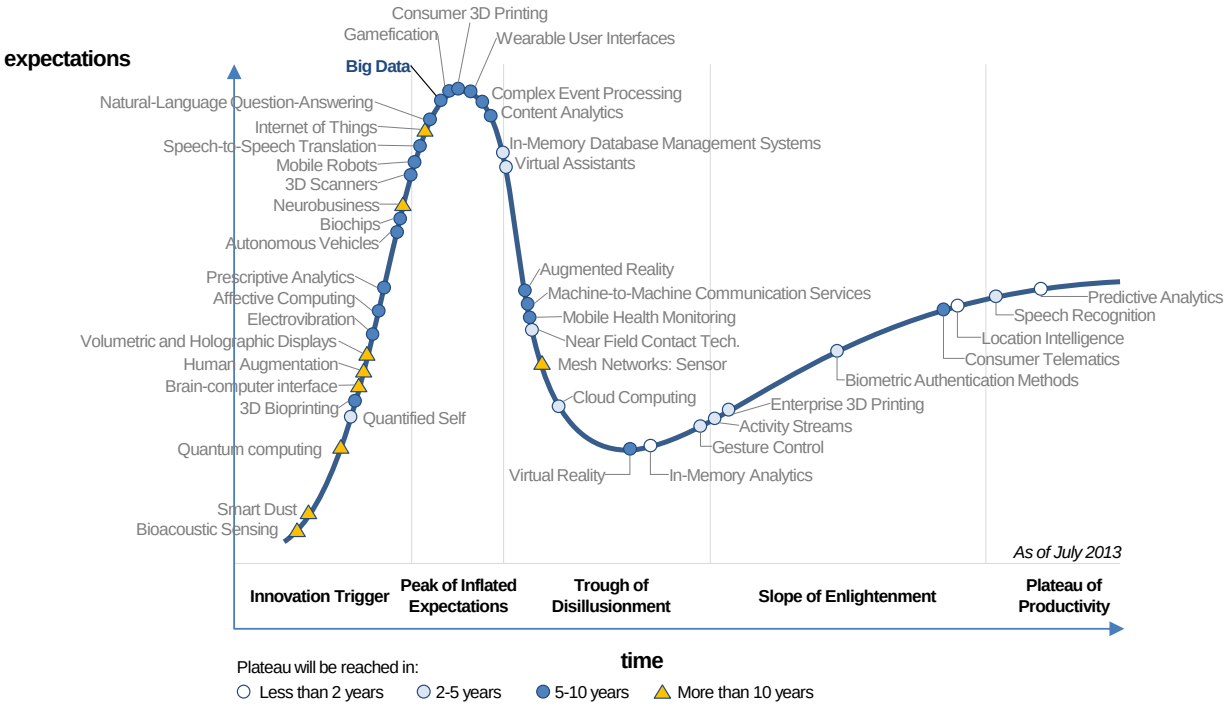
2. The Big Data lifecycle

Big Data broadly refers to extremely large data sets now able to be acquired, stored and interpreted through modern technology. While no broadly agreed definition exists for big data, it is commonly understood to qualify datasets too large to be contained or processed using the resources of a typical personal computer or the analytical capacity of commonly used spreadsheet applications.

Volume is only one attribute of Big Data. Other significant attributes include velocity (the speed at which data is collected and processed) and variety (the range of structured and unstructured elements that comprise the data sets). Overall, volume, velocity and variety are typically used to differentiate Big Data from other data. However, it is important to understand that these are purely descriptive terms. They do not capture the fundamental changes that have occurred in recent years that have given rise to such large and exploitable data sets.

Big Data – in transport and elsewhere – has emerged from the convergence of rapidly decreasing costs for collecting, storing and processing, and then disseminating data. Decreasing costs for sensors has led to a proliferation of sensing platforms transforming large swathes of the analogue world into digitally processed signals. Decreasing data storage costs have allowed the retention of data that had previously been discarded. As noted by science historian George Dyson "Big Data is what happened when the cost of storing information became less than the cost of making the decision to throw it away." (Dyson, 2013)

Figure 2. **Hype cycle for emerging technologies (2013)**



Source: Gartner Research.

At the same time the advent of inexpensive, often open-source analytical software has democratised access to cost-effective and near real-time processing and analysis of large, high-velocity and high variance data sets (Asslett, 2013).

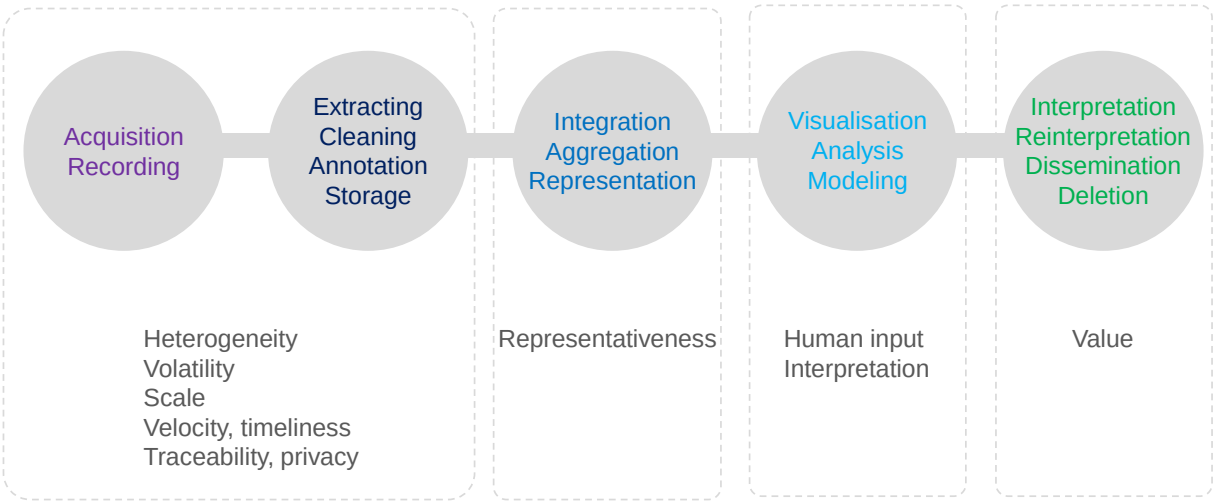
As a term, Big Data is relatively recent. Yet it has already generated considerable interest and discussion, including in the field of transport. While strong interest is generally seen as a positive the enthusiasim or "hype" for a new technology can go too far and lead to inflated expectations. Big Data is nearing the apex of such a "hype" curve (Figure 2) and it remains to be seen how relevant, robust and perennial a concept it proves to be for mobility-related data.

Big Data is not a singular construct; rather, it is a process spanning data acquisition, processing and interpretation (see Figure 3). This lifecycle of Big Data is described in the following sections.

Data acquisition and recording

People increasingly leave a digital trace wherever they go (both voluntarily and involuntarily). The technology utilized in each phone call, text message, email, social media post, online search, and credit card purchase and many other electronic transactions reports on the user's location at a given point in time. This data is then relayed to the central servers of the service providers that enable these actions. When cross-referenced with the geographical terrain, data harnessed at this scale offers a means of understanding, and responding to, the urban dynamics of the city in real-time. Making sense of this data, especially for policy, requires familiarity with the technical aspects of data production methods as well as an understanding of how, or from whom, the data is sourced. We address these issues in the following sections.

Figure 3. **Big Data collection and analysis lifecycle**



Data production: Digital vs. analogue

In general terms, data may be either be "born digital" or "born analogue". (PCAST, 2014) "Born digital" data is created by users or by a computing device specifically for use in a machine processing environment. Examples of "born digital" data include (PCAST, 2014):

- Global Positioning System (GPS) or other geo-localised spatial data stamps.
- Time stamps and process logs.
- Metadata regarding device identity, status and location used by mobile devices to stay connected to various networks (GSM, Wi-Fi, etc).
- Data produced by devices, vehicles and networked objects.
- Public transport card tap-ins or swipes and other data associated with portal access (badge, key cards, RFID tags) or cordon passage (e.g. toll roads, congestion charging systems, etc).
- Commercial transaction data (credit card use and transaction records, bar-code and RFID tag reading, etc.).
- Emails and SMS, metadata relating to phone calls.

"Born digital" data is produced by design to address one or a series of specific needs. Efficiency considerations has meant that only the specific data required for a process was generated and retained, in order to avoid straining storage and processing capacicity or inflating costs. However, the drop in processing and storage costs effectively means that over-collection of data (beyond the stated initial purpose for data collection) is easily possible and has a near zero real cost.

"Born analogue" data is data that arises from an imprint of a physical phenomenon (light, sound, motion, presence of a chemical or biological compound, magnetic impedance, etc) upon a sensing device, and its subsequent conversion into a digital signal. Sensors may include cameras, microphones, magnetic field detecting devices, heart rate monitors, accelerometers, thermal sensors, etc. Costs for sensors have decreased sharply contributing to a rapidly pervasive sensing environment.

Examples of "born analogue" data include:

- Video streams from surveillance, in-vehicle, roadside or other cameras.
- Audio content of voice phone calls, ambient audio from video cameras or microphone networks.
- Motion/inertia (accelerometers, ultrasonic sensors), heading (compass), temperature, infrared radiation, electromagnetic fields, air pressure etc.
- Data relating to heartrate, respiration, gait, and other physical and health parameters.
- Electromagnetic or light (laser) reflectance of objects (e.g. synthetic aperture radar –SAR or laser-based LIDAR systems).

With the proliferation of sensors some may suggest "born analogue" data is rapidly expanding to encompass all potential observations, both now and into the future. However, two filtering mechanisms operate at the initial stage of data collection that limit the scope of data acquired. The first filter is the design specifications (and therefore limitations) imposed on any sensor or recording device. For example, due to their design specifications a heat sensor will not record audio signals while an accelerometer will not record geographic coordinates. Nevertheless, the ability to infer one phenomenon from the observation of another is constantly evolving. For instance, research into the electromagnetic interference (which road authorities have tried to shield roadside data transmission cables from) has shown that the data produced by the interference can be used to infer vehicle movements and provide traffic counts.

The second operative filter relates to the *observation rate* of sensed or monitored events versus the rate of *data retention and transmission*. For instance, engine sensors in a commercial aircraft may process up to 10 terabytes of data per 30 minutes of flying time but most of this data is discarded as soon as it is used. Rates of data retention are orders of magnitude lower and the amount of data transmitted during flight even lower still. Much sensed or generated data is rapidly discarded and what is left may be filtered and compressed before being treated and used. However, the potential value of sensed data may not be recognised at the system design phase. There is a need to ensure filters or compression algorithms do not discard potentially useful information.

Filtering and compression-linked data loss is becoming less of an issue with the advent of lower cost data storage and transmission technologies, coupled with state-of-the-art database platforms such as Hadoop (see Box 1 on Big Data core technologies). These changes, in addition to growing sophistication of, and decreased costs for, dual sensor-computing devices, increase the relevance of retained data. Sharply dropping sensor costs and size, improved sensor performance, lower data storage costs and improved relevance of retained data all contribute to the large-scale increase of exploitable data. However, as noted by the Community Computing Consortium, challenges remain:

> "One challenge is to define these filters in such a way that they do not discard useful information. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artefact that deserves attention? In addition, the data collected by these sensors most often are spatially and temporally correlated (e.g., traffic sensors on the same road segment). We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack. Furthermore, we require "on-line" analysis techniques that can process such streaming data on the fly, since we cannot afford to store first and reduce afterward." (Community Research Association, 2012)

In general, the literature on Big Data classifies sources under three broad categories: opportunistic sensing, purposely sensing and crowdsensing. Opportunistic sensing leverages data running on existing systems, such as a telecommunication network, but can be used to better understand mobility. In other terms, data is collected for one purpose and used for another. This approach to data collection is made possible largely by widespread use of mobile phones (ITU, 2014) – citizens replace the need for purpose-built sensors, contributing real-time data through their portable devices. Other typical data providers include credit card companies recording user transaction and taxi fleets reporting vehicle GPS.

The potential of opportunistic sensing is furthered by recent, and forecasted, increases in the sample size, reporting frequency and processing power of existing networks. Previously, mobile phones generated data only when calls were made. Today, with the transition to smart phones, time and location is communicated to service providers every time a text or email is sent, a photo is uploaded or on-line purchase is made. By 2020, more than 70% of mobile phones are expected to have GPS capability. (McKinsey Global Institute, 2011)

Mozilla, the free software community which produces the Firefox web browser in partnership with Chinese chip maker Spreadtrum, has prototyped a low-cost smartphone device aimed at the developing world for USD 25. (Spreadtrum, 2013) The smartphone will be able to run simple apps and make use of mobile Internet, empowering citizens across social classes with more data producing and receiving capabilities. Furthermore, advancements in telecommunication technologies and data connectivity will lead to better access to real-time updates. When triangulated with signals from several towers, callers can be located within a few dozen metres. (European Commission, 2011)

Of the data types associated with opportunistic sensing, McKinsey's 2011 report on Big Data emphasises the transformative potential of location data from mobile phones for transportation applications. The traditional means of studying city dynamics, such as census surveys and vehicle counting, are both time-consuming and expensive. Meanwhile, mobile phone carriers are routinely collecting location data on all active users. This offers a valuable means by which to monitor activity patterns frequently, cheaply and at an unprecedented scale. (Becker, et al., 2011) Call Detail Records (CDRs) identify the approximate location of the caller and receiver at the time of the call through the cell towers carrying their phone signals. Strung together over an extended period of time, each recorded location contributes to an observed flow of people between different geographical regions, whose precision is tied to frequency of mobile phone use. More accurate still, data from smartphones equipped with GPS and Wi-Fi are able to locate the user within five meters.

Box 1. **Big Data core technologies**

Big Data draws on a range of technologies and system architectures that are designed to extract social or economic value from high velocity, large scale and extremely diverse and heterogeneous data streams. (Cavoukian & Jones, 2012) At their core are four interlinked technological developments.

Ubiquitous data logging and sensor platforms

Extensive software event logging (and storage) and the deployment of millions of sensing devices enable the real-time production of petabytes of data globally.

Real-time in-stream data analysis

Sophisticated algorithms and distributed computing capacity (often hard-wired to sensor platforms) enable the real-time parsing and analysis of data as it is produced. In-memory analysis is especially useful for extracting relevant data from unstructured analogue video or audio streams (see below).

New analytic frameworks

New techniques have emerged that allow efficient processing of very large data sets within the constraints of available runtime computing capacity. Many of these techniques have been released under open-source licenses free of commercial rights. This has greatly accelerated their uptake. Map-reduce work processes (such as Hadoop or its derivatives) leverage parallel processing by breaking up large and complex semi-structured and unstructured data sets into more manageable subsets. They then allocate coordinated processing tasks to multiple distributed servers. These algorithms are fully scalable and are not bound by having to formalise database relationships ahead of storage and analysis. They can be applied to directly to the data, irrespective of size, format and complexity. Nonetheless, they may not be sufficiently reactive to use in the context of in-stream data analysis. Other approaches have emerged that are specifically geared to the analysis real-time streaming data and involve some form of in-memory processing (that is, analysis occurs without data storage).

Advances in data storage

Dropping data storage costs have increased the ratio of retained to generated data. This data includes information that, in the past, had seemed insignificant or trivial (e.g. “digital dust”) and was therefore discarded. However, when analysed by sophisticated algorithms or merged with other sources of contextual data, “digital dust” may provide important new insights. This data is increasingly being stored remotely (away from the systems that produce it) in data centres that may even be in another jurisdiction. Related to the development of remote data centres is the emergence of “cloud” computing capacity that can be used to analyse large and real-time data sets. The “cloud” refers to remote data storage centres as well as the suite of data transfer and networking protocols that allow access to and analysis of distributed data as if it were located on a single server. Not only does “cloud computing” deliver economies of scale in relation to data storage, management and support costs, it also opens up new possibilities for ad-hoc and customisable access to computing capacity on public cloud-based platforms (e.g. Amazon Web Services, Google Cloud Platform, etc.)

In contrast to opportunistic sensing, purposely-sensed data sets are derived from ad hoc sensor networks configured to study a specific phenomenon. Due to advances in microelectronics sensors and computation are becoming increasingly affordable and widely distributed, a phenomenon often referred to as “smart dust”. Hence, networks of remote sensing agents can now be embedded in the city fabric to extract large amounts of information. This data is channeled to central control stations where it is aggregated, analysed and used to make decisions on how the monitored terrain should be regulated and actuated. (Ratti & Nabian, 2010) Here, the resulting data sets tend to be more uniform, and the stated use and actual end-use scenarios are better aligned to decode various flows within the city.

Transportation systems that make use of information from cameras and microcontrollers to optimise public transit, monitor the environment and run security applications are known as intelligent transportation systems (ITS). In general, information for ITS is extracted from two types of customised sensor networks fixed sensor agents and dynamic probes, e.g. mobile sensors. (Calabrese, et al., 2011)

In fixed sensor agents, location coordinates are linked to the identification code of the sensor transmitting the data. Mobile sensors, on the other hand, include updated location coordinates with each transmission. Both sensor types can be configured to transmit information at predefined intervals, or to respond to requests for data from the central server. (Ratti & Nabian, 2010)

Fixed sensors are permanently installed and monitor the real-time dynamics of the surrounding terrain, such as the speed at which vehicles are travelling, distance between vehicles, road surface conditions and CO_2 emissions from car exhausts. They can be hard-wired or wireless and range from roadside traffic cameras connected with optical fibres buried underground, to sensors that operate over radar, ultrasound, or infrared (e.g. tyre overheating). Sensors that are inductive, piezoelectric, or magnetic can be laid under the road. For example, loop detectors embedded in the pavement report on traffic flow through perturbations in circuit conductivity caused by vehicles passing overhead. In each case, sensors are equipped with a control unit, battery system, solar panel and transmission system.

Location-based data accounts for most of the information derived from sensors capable of probing the terrain. The most common example being Global Positioning System (GPS) receivers. GPS-enabled vehicles periodically report their position and the time the message was transmitted. GPS units derive velocity and direction of movement from two or more position measurements. The receivers use a control unit and transmission system to relay this information to a central server via text message or transmission of data packets. Once collected, GPS data needs to be carefully processed to account for communication delays, and extract accurate timestamps for each location reading.

Both fixed sensors and GPS receivers are limited by the considerable investment required to embed and maintain the sensing agents. In order to achieve sufficient sensing capability, these sensors must be deployed on mass. In the case of GPS receivers, their ability to accurately estimate traffic conditions is tied to the density of vehicles reporting on a given area. As the number of probes increase, transmission cost to the central system can become high. The transmission costs associated with wired fixed sensors are especially high given they must be strung together as part of a larger network in order to produce a complete view of mobility conditions in a city, and the cable infrastructure required to connect each one with the central collecting point is expensive.

The constraints outlined above have deterred many municipalities from investing in sufficient monitoring systems required to process purposely-sensed information at the city scale. This is particularly true for large urban areas consisting primarily of smaller streets. Instead, their application is often limited to select locations such as highways, major urban corridors and large intersections. (Calabrese, et al., 2011) One notable exception to this trend is the City of Rio de Janeiro (see Box 2).

The Rio de Janeiro Operations Centre uses mobile application to warn citizens about heavy rain, strong wind, fog, energy shortages, traffic signal malfunctions, mudslides, fire, smoke and points of flooding It also receives information from the public. On an average day, Rio's transportat planners receive aggregated views from 110 000 drivers and reports on 60 000 traffic incidents. Since 2013, Rio de Janeiro has been the first city in the world to collect real-time data both from drivers who use Google's navigation application Waze, and pedestrians who use the public transportation app Moovit. The crowd-sourced data is overlaid with real-time information from various sensors and cameras. In the future, the city plans to start monitoring how cyclists move around the city using cycling app Strava. (Forbes, 2014b)

Google Waze crunches a continuous stream of traffic data from its community to propose what it recognises as the fastest route to the destination. Like TomTom or Google Maps, the app provides users with step-by-step directions to the selected destination. Waze's map differs in its ability to integrate user-generated content. Individuals submit incident reports that mark the precise location of the accident, traffic jam and any other driving hazard. As with many crowd-sourced apps, its success depends on the volume of users; and hence, is more reliable in dense urban areas than in rural ones.

Box 2. **Rio de Janeiro Municipal Operations Centre**

After a series of floods and mudslides claimed the lives of 72 people in April 2010, city officials recognised the need to overhaul city operations more significantly in preparation for the 2014 World Cup and Olympics in 2016. (United States Environmental Protection Agency, 2014) In collaboration with IBM, the City of Rio de Janeiro launched the Rio de Janeiro Operations Centre (ROC) in 2010 with the initial aim of preventing deaths from annual floods. This centre was later expanded to include all emergency response situations in Rio de Janeiro.

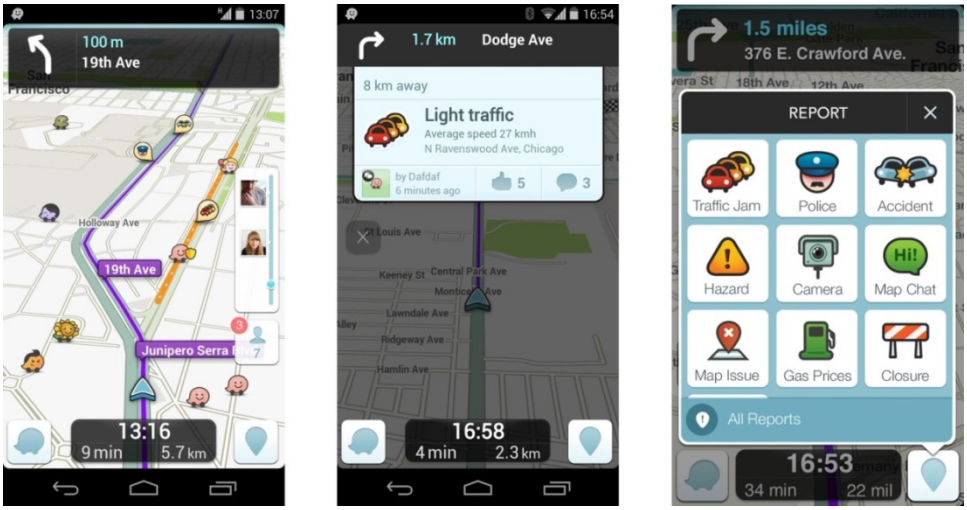
In traditional applications of top-down sensor networks, data from each department operates in isolation. However, ROC's approach to information exchange is based on the understanding that overall communication channels are essential to getting the right data to the right place and can make all the difference in an effective response to an emergency situation. The information-sharing platform they created enables them to tap into various departments and agencies, and look for patterns across diverse data sets to better coordinate resources during a crisis.

The centrally located facility surveys 560 cameras around the city and another 350 from private sector utility concessionaires and public sector authorities (Centro de Operações da Prefeitura do Rio de Janeiro, 2014). The incoming feeds are aggregated on a single server and displayed across a 80-square meter (861 square feet) wall of tiled screens – a smart map comprised of 120 layers of information updated in real-time such as GPS tracking of buses, city officials and local traffic. With over 400 employees working in shifts 24 hours per day, seven days a week ROC performs a variety of functions aimed at improving the efficiency, safety, and effectiveness of relevant government agencies in the city. While much of the attention paid to the centre focuses on emergency monitoring and response, especially related to weather, a significant portion of the work undertaken relates to ensuring the smooth functioning of day-to-day operations like transport.or example, through the centre, Companhia Municipal de Limpeza Urbana (municipal waste corporation) can monitor where its trucks are and better sequence trash collection, minimising fuel consumption and improving waste management services (US Environmental Protection Agency, 2014).

IBM has created similar data centres elsewhere in the world for single agencies (such as crime centres for New York and Madrid, and a congestion fee system for Stockholm) but ROC is the first application of a citywide system to integrate all stages of crisis management from prediction, mitigation and preparation, to immediate response. In Rio de Janeiro, the centre gathers data from 30 government departments and public agencies – water, electricity, gas, trash collection and sanitation, weather and traffic monitoring – in real-time through fixed sensors, video cameras and GPS devices. Data fusion software collates this data using algorithms to identify patterns and trends, including where incidents are most likely. (Open Data Research Network, 2014)

In a statement on the use of sensor-based systems to correlate situational events with historical data at their Intelligent Operations Centre for Smarter Cities, IBM's Director of Public Safety explained "The aim is to help cities of all sizes use analytics more effectively to make intelligent decisions based on better quality and timelier information. City managers can access information that crosses boundaries, so they're not focusing on a problem within a single domain. They can start to think about how one agency's response to an event affects other agencies". (Yasin, 2011) In every city, the complete story cannot be told by traffic figures and meteorological data alone. To fully assess a crisis situation, the voice of the people must be heard. Each urbanite can be thought of as a human sensor, capable of reporting on their experience of the city through content sharing platforms such as Flickr, Twitter, Facebook or Wikipedia. (Ratti & Nabian, 2010) These actions offer a unique view on how people navigate their environment, bringing clarity to points of attraction or spontaneous migrations. This approach describes the third data source known as crowdsensing.

Figure 4. **Views of Waze, a community-based traffic and navigation application**
(interface showing the reported incidents and traffic suggestions)



Source: Google Play.

Need for transparency and metadata on data provenance

Transparency regarding the nature of data and the conditions under which it was collected is crucial for data-driven transport policy making. In this respect, the initial recording and subsequent preservation of metadata plays an essential role in enabling data interpretation and re-interpretation. This metadata may include information on data structure, the context in which it was collected and how it was generated (e.g. its provenance). For sensor-based data, provenance data is especially important as the type of sensor platform may affect the representativeness of the data produced (see Box 3 on current shortcomings in Big Data analytics).

For instance, if a traffic data source is a network of embedded loop detectors, it becomes important to account for the sometimes significant portion of detectors that may be offline or that consistently give implausible and possibly incorrect readings. Likewise, accounting for smartphone or app penetration rates across demographic segments becomes important when analysing smartphone or app-sourced data for designing public transport services. Ensuring a non-degradable provenance metadata is especially important for fused data sets whose analysis will depend on understanding the nature of all of the component data streams. However, the more detailed the provenance metadata (e.g. down to a single identifiable sensor) the more difficult it becomes to manage privacy issues.

Box 3. **Current shortcomings in Big Data analytics**

Proponents of Big Data-driven analysis have predicted that:

> "In the next two decades, we will be able to predict huge areas of the future with far greater accuracy than ever before in human history, including events long thought to be beyond the realm of human inference. The rate by which we can extrapolate meaningful patterns from the data of the present is quickening as rapidly as is the spread of the Internet because the two are inexorably linked. The Internet is turning prediction into an equation… as sensors, cameras, and microphones constitute one way for computer systems to collect information about their—and our—shared environment, these systems are developing perceptions that far exceed our own." (Tucker, 2014)

Massive and near real-time data sets, often based on ubiquitous sensing, are so large that they may seem to mimic reality. Just knowing there is a link between two or more observable variables and an outcome may be sufficient to predict the frequency of that outcome in the future. All that is needed is an algorithm that consistently detects patterns in the data. In this view of Big Data analytics the need for classic statistical tests regarding bias and validity or explanatory theories and models would be eliminated. The data could already be assumed to be an accurate representation of reality. However, this is an overly reductive viewpoint. Furthermore, some early successes in Big Data analysis based on this assumption have failed to provide robust predictive results over the long term.

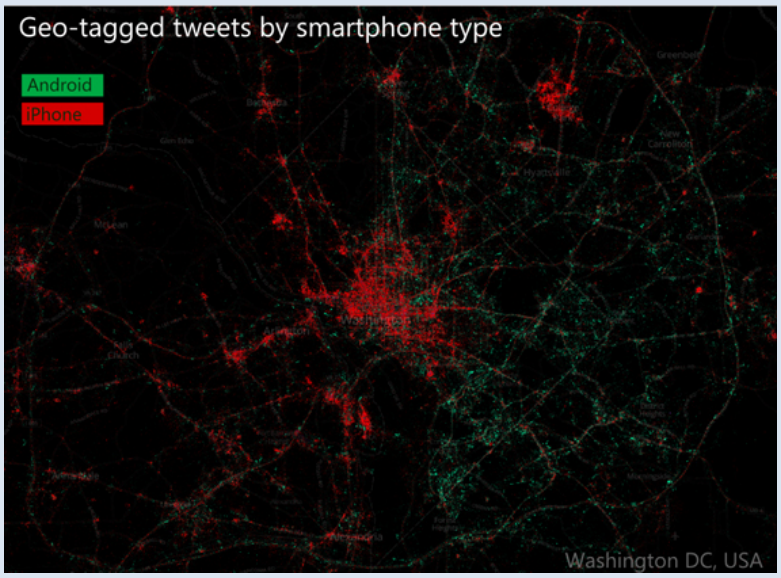
Observers have pointed out that "there are a lot of small data problems that occur in Big Data" (Harford, 2014) and that a theory-free approach to analysing Big Data does not make these go away – in fact, it makes them worse. Big Data has not emancipated analysts and policy makers from the strictures of statistical rigor since Big Data is not only prone to many of the same errors and biases in smaller data sets, it also creates new ones.

Causality, correlation and multiple correlations

Big Data analytics are well suited to the discovery of correlations that were not obvious, or even visible, in the data initially. This is where "letting the data speak" is the most effective method of providing new insights. Correlation and causation are two different things and though correlative variables may reveal the possibility of a causal relationship in the data, they do not explain which correlations are meaningful or predictive. For example, combining low granularity hospital data with geo-localised patient data, traffic flow data and digital maps may reveal that living close to busy urban roads correlates with early mortality. A logical assumption would be that this was due to exposure to air and noise pollution. However, property values adjacent to busy urban roadways may be low and the population living in these areas may have low incomes. These two variables may correlate closely with a number of risk factors (diet, exercise, smoking) that contribute to early mortality. Without a more detailed understanding of the causes of death, authorities may find that policy efforts to reduce pollution (or traffic) may not necessarily reduce early mortality along these corridors. Furthermore, Big Data analytics may amplify the importance of spurious correlations due to the extremely large volumes of data processed. While Big Data analytics are suited to "letting the data speak", understanding what exactly is being said calls for human interpretation.

Even if a correlation may prove to be robust over a given period, Big Data analytics alone cannot provide insight into what might cause the correlation to break down – nor what pattern may emerge in its place. This is a second critique of Big Data-led analytics – they can often be helpful in examining fairly common occurrences but have difficulty in handling less common or outlying events.

Differential smartphone penetration bias in Washington, DC
(visualised through geo-referenced tweets)



Source: https://www.mapBox.com/blog/visualising-3-billion-tweets/

Bias and representativeness

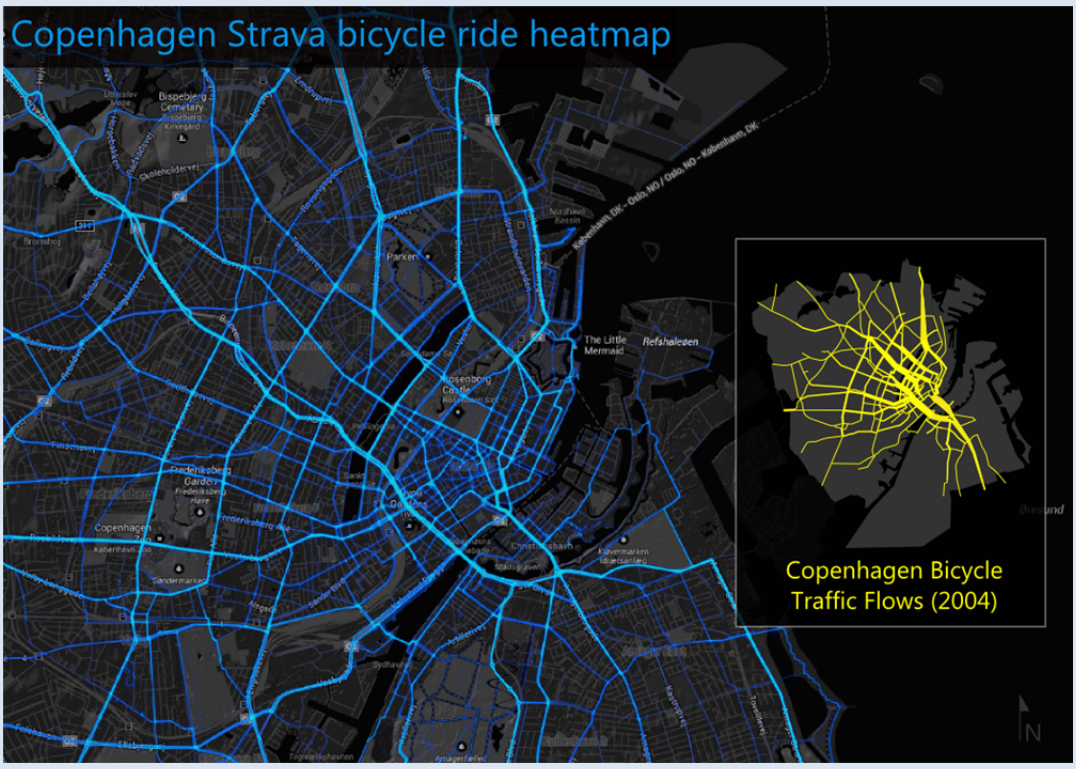
A third critique of Big Data analytics is that the existence of massive data sets does not eradicate traditional statistical traps – especially those of sample bias and sample error.

Claims relating to the ubiquity of sensor networks and other sources of Big Data are often exaggerated or ignore specific and consistent sources of bias that may be relevant for policy analysis. For example, use of smartphone-sourced location data may be subject to biases that result not only from differential rates of smartphone penetration amongst different demographic groups (e.g. the very young and the old, high income vs. low income) but also to differential rates of penetration of operating systems amongst the population of smartphone owners.

Analysis of geo-tagged tweets reveals a stark segregation between iPhone and Android users in many locales (as in illustrated in Figure of Washington, DC). These differences may correlate to income or race (or both) and may have real consequences when considering the use of smartphone-sourced location data. An iPhone-only app, for instance, that provides authorities with automatic pothole detection reports based on signals from the phone's accelerometer and other sensors would direct authorities to repair potholes only in those areas with high levels of iPhone ownership. In the case of Washington, this would ignore nearly half of the city's roads.

In another example, biases may emerge from volunteered location data due to specific characteristics of app-users versus the general population. Strava, for instance, is an app that started as a way for cyclists and runners to compete against each other's times on defined segments. It has amassed a very large data set of geo-localised tracks matched with user profiles (77.7 million bicycle rides, 19.7 million runs, 220 billion GPS data points) around the world based on self-reporting by app users. It markets this data (under the Strava Metro product) to planners and city authorities to help shape urban transport policy. Strava claims a 40% share of commuting routes in its data. However, it seems reasonable to expect, given the app's stated purpose, that a bias in favour of recreational cyclists' and runners' route preferences may exist, though Strava claims a 40% share of commuting routes in its data. A potential secondary bias may also emerge due to the type of bicycle commuter most likely to use the Strava app (e.g. more competitive, longer-distance commuters).

Potential representativity bias: Copenhagen cycling



Source: engineering.strava.com

This potential for bias can be seen in Strava's mapping of Copenhagen (figure above), a city with very high levels of utilitarian cycling. Here Strava accentuates slightly different routes than the city's own bicycle traffic survey (especially the route running from the northwest to the southeast). Strava routes are likely used by recreational and sports cyclists on training rides rather than shorter-distance utilitarian cyclists. Likewise, Strava seems to under-represent short-distance bicycle trips in the city centre. Without further contextual data on the population generating the data – and especially the population of cyclists and pedestrians not using the app – city planners would be hard-pressed to formulate sound policy recommendations.

The use of Big Data by policy-makers should be assessed according to the specific concerns at hand and the level of knowledge relating to a problem to be addressed. For well-defined challenges where many variables and their inter-relationships are understood, small, carefully stratified and representatively sampled data may be as, and perhaps even more, effective at finding solutions than the use of Big Data. However, for situations where knowledge is low, Big Data analytics may help elucidate relevant questions to ask and identify potential new directions for policy.

Data extraction, cleaning, annotation and storage

Beyond questions of availability and collection costs, an important factor to consider when selecting a data source is its fitness for analysis. Data analytics refers to all the ways in which information is extracted from a given data set. Once parsed into relevant fields (e.g. origin and destination time, longitude and latitude), a series of operations can be performed to clean, transform and model the data in pursuit of meaningful conclusions.

A range of techniques and tools have been developed, or adapted, to aggregate, manipulate and visualise Big Data. These draw on expertise from a number of fields including statistics, computer science, applied mathematics and economics. This both adds to the challenge of making Big Data accessible and highlights the need for a multidisciplinary approach.

Within the context of transportation planning, spatial analytics typically extract the topological, geometric, or geographic properties encoded in a data set. Across the various case studies related to mobility, the techniques for data analysis can be grouped into, but are not limited to, the following categories (McKinsey Global Institute, 2011):

- **Data fusion:** techniques to consolidate data produced by multiple sources, such as location data produced by mobile phones and GPS-enabled vehicles.
- **Data mining:** techniques to extract patterns from large data sets, such as the relationships between discrete nodes in a transportation network.
- **Optimisation:** techniques to reorganise complex systems and processes to improve their performance according to one or more parameters, such as travel time or fuel efficiency.
- **Visualisation:** techniques used for generating images, diagrams, or animations to communicate the results of data analysis, such as traffic maps. Visualisation techniques are used both during and after data analytics to make sense of the information.

An important factor to consider when selecting a data source is the scope and quality of the resulting data set. Data extracted from a single source is generally considered clean and precise. However, meaningful analysis of a single source depends largely on the generating system's ability to serve as a proxy for the phenomenon of interest. The reality is that data is often "messy", in that it is heterogeneous, "dirty" (includes incorrect, mislabelled, missing or potentially spurious data) and, in its native format, is incompatible with other data sources. Part of the challenge lies in the fact that some data may be highly structured (for example, GPS latitude and longitude data and commercial transaction data) facilitating rapid analysis while other data may comprise highly unstructured data sets (emails, social media content, video and audio streams) and therefore be more difficult and time consuming to analyse. Advances in data

processing and analysis techniques allow the mixing of both structured and unstructured data in order to elicit new insights, but this requires "clean" data.

Cleaning data and preparing it for analytical use is a non-trivial task that may entail significant cost. Structured data must be parsed and missing or potentially incorrect data accounted for. Unstructured data must be correctly interpreted, categorised and consistently labelled. In some instances, manual "data wrangling", "data munging" or "data janitor work" remains a necessary component of the data collection and analysis stream. This work requires a large investment in time and resources – accounting for 50% to 80% of data scientists' time according to some estimates. (Lohr, 2014) Data preparation may be greatly facilitated by the use of appropriate compensatory algorithms but the choice of algorithm may itself lead to imputing or prediction errors if it incorrectly interprets missing values[1] or minimises outliers that could be important. Just as metadata relating to the data itself may help improve subsequent interpretation, metadata regarding data cleaning and correction methods may be equally helpful – within reason, since each new metadata element may weigh down data processing speed and efficiency.

In 2013, the MIT SENSEable City lab looked to data produced by social media platform Twitter to infer on global mobility patterns. More specifically, the study extracted GPS coordinates from the mobile devices, or IP addresses from the computers, used to send tweets. At present, approximately 1% of tweets can be geo-located. (Morstatter, et al., 2013) This is expected to increase dramatically in the coming years due to the proliferation of smart devices and mobile applications. Common to both homogenous and fused data sets is a cleaning phase': The data set gathered from Twitter's Streaming API service was examined for statistical errors and artificial tweets; namely for users having relocated at an impossible speed and posts made by commercial users who were less likely to reflect human activity. From here, each user's country of residence was identified, based on where they had posted the highest percentage of tweets. All tweets made outside of this country would flag the user as a visitor. The resulting network of tweet flows suggested country-to-country preferences, as well as peak travel times for residents of a particular country. (Hawelka, et al., 2014)

The major challenge associated with analysing a single source is in going beyond the literal meaning of the data to answer more general questions. The Twitter streaming data sets provide an accurate account of where a group of subscribers accessed a specific service at a given point in time. However, the resulting representations are of limited use in elucidatingwhat influences user flows, such as the motivations behind their movement and their preferred modes of travel.

Integration, aggregation and fusion

New insights can emerge from the analysis of single data sets but the real potential for new knowledge rests on the improved ability to apply analytical methodologies to multiple data sources.

Data fusion techniques match and aggregate several heterogeneous data sets creating or enhancing a representation of reality that can be used for data mining (see Box 4). Data fusion is an especially important step in using inputs from multiple sensor platforms. For instance, data fusion algorithms help process inputs from wheel movement sensors, accelerometers, magnetometers, cellular signal sensors, cameras, laser scanners and GPS chips. All these data sources contribute to creating a precise representation of the location of a car on a street. Such data fusion is necessary for the development of autonomous driving vehicles. Here, the final representation of a vehicle (e.g. the precise vehicle location, size, direction and speed) is all that is retained, thus eliminating the need to store every sensor's individual data stream. Mid-level data fusion methodologies merging structured machine-produced data are relatively well advanced. On the other hand, high level data fusion tasks merging multiple unstructured analogue

[1] See for example discussion in (Li & Li, 2013) and (Hutchins, 2010).

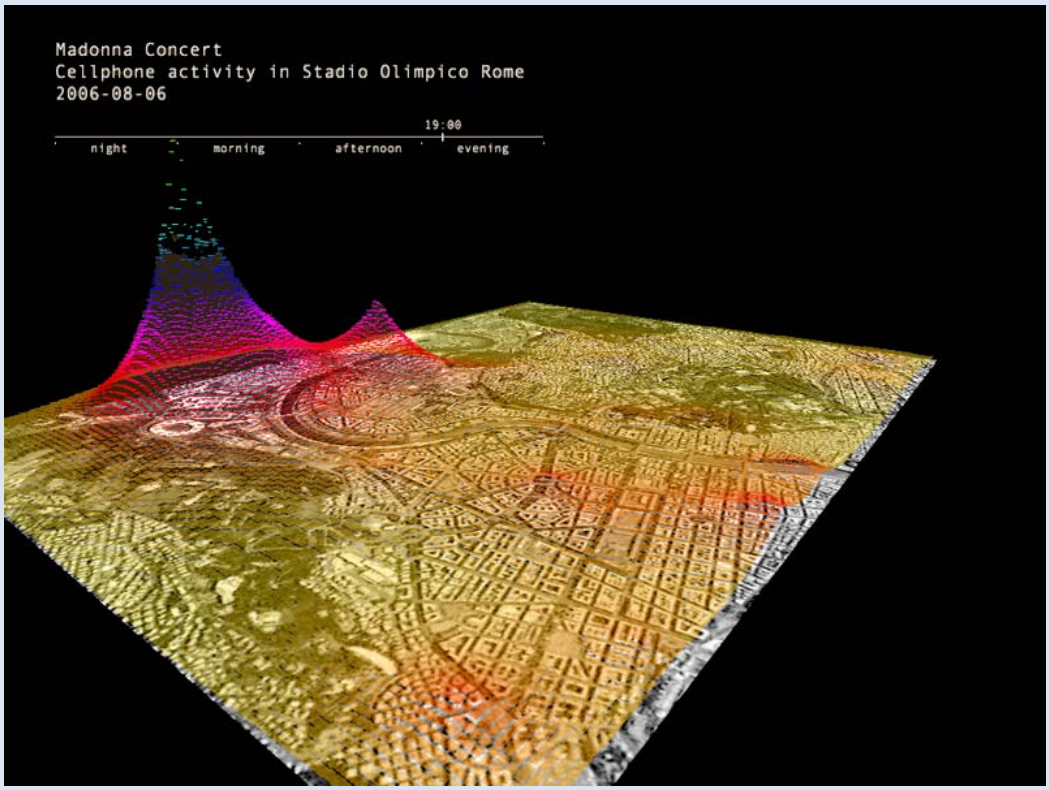
sensor inputs remains challenging and a focus of current research. (Khalegi, et al., 2013) It is this high-level data fusion capacity – one that starts to emulate human capacities – that will be necessary for the large-scale deployment of sensor-centric autonomous vehicles.

In the case of data integration or aggregation, data sets are matched and merged on the basis of shared attributes and variables but the whole of each separate data set is retained. This method is well suited for increasing knowledge discovery via the analysis of contextual data.

Box 4. **WikiCity Rome case study**

In the WikiCity Rome case study, the real-time visualisation of data mined from communication networks was cross-referenced with the geographical terrain. This allowed urban dynamics to be presented in real-time to observers. Such technologically enhanced spectacles - real-time infoscapes projected onto architectural surfaces, or accessed via worn and handheld devices - provoke a temporary displacement of the observer from the physical terrain they inhabit to a distant location, providing them with an overview of the dynamics contained within the urban landscape.

Dynamic map illustrating the levels of activity within Rome during a Madonna concert
(real-time data created through interpolation of mobile phone usage)



Source: WikiCity Rome, MIT SENSEable City Lab.

The WikiCity Rome project tapped into aggregated data from mobile phone usage. The resulting visualisations depicted the pulse points of the city, providing an overview of how the urban landscape is occupied, and where and in which temporal patterns the mobile phone-using crowd is dispersed. Crowdsensing based on mobile phone usage allows for spotting the “hot” locations and congested spots of the city in real-time. This can help authorities to regulate traffic and the flow of resources within the city, based on real-time dynamics.

When the system was exhibited at the 10th International Architecture Exhibition of the Venice Biennale, researchers at MIT SENSEable City Lab supplemented the mobile phone-based evaluation of urban dynamics with data based on the instantaneous positioning of buses and taxis. This provided real-time information about mobility, ranging from traffic conditions to the movements of pedestrians throughout the city. The visualisations provided a qualitative understanding of how the aggregated data of network mobile phone usage and public transit location information can be used to provide valuable services to citizens and authorities. Such information "can give city dwellers a deeper knowledge of urban dynamics and more control over their environment by allowing them to make decisions that are more informed about their surroundings, reducing the inefficiencies of present day urban systems". (Calabrese, et al., 2011)

The challenging aspect of data fusion is in the extraction of salient features across multiple data sets generated for different uses. In the case of the WikiCity Rome case study (see Box 4), the team at MIT had to parse, extract and process the data collected by Telecom Italia, Atac, and Samarcanda. Furthermore, they had to do so in real-time so as to contribute to the decision-making process of its users. In order to create a dialogue between the different feeds, a common ontology was developed to describe the data in terms of the following:

- **Location** (coordinates system, latitude and longitude) of the transmitting device or reported action.
- **Time** of data transmission or reported action.
- **Data category** (location-based mobile phone data, GPS data, news bulletin).
- **Data format** (single value, matrix, vector, text, image, etc.)
- **Data representation** (e.g. measurement unit).
- **Semantics** of the data (e.g. tracking vehicle or mobile phone).

In general, data fusion techniques aim to fuse accuracy and semantics. Data collected from multiple sources often contain inconsistencies in terms of resolution. For example, in the case of WikiCity Rome, Telecom Italia's probes were sampling signal strength every 4.8 seconds, while Atac and Samarcanda's servers received GPS readings at 30-second and 5-minute intervals respectively. Semantics refers to the subject being represented. In one case, the data tracks a vehicle; the other follows the initiator and recipient of a call.

Analysis, modelling and visualisation

One outcome of the availability and relative low-cost of exploiting very large data sets is an evolution in, and a broadening of, the analytic techniques used to extract insights from this data. Traditional approaches involving statistics or optimisation methods are still relevant but run into data processing limitations when considering extremely large and high-velocity data sets[2]. Other knowledge-discovery approaches including data mining (and the contribution of data mining to machine learning, network analysis and pattern recognition) and visualisation techniques are more suited to Big Data.

Data mining

Data mining approaches differ from traditional database analysis methods in that they do not presuppose a model describing relationships in the data nor do they require specific queries on which to base analysis. Rather, these approaches let the data speak for itself, relying on algorithms to discover patterns that are not apparent in the single, or more often, joined data sets.

[2] The fact that traditional statistical approaches are challenged by massive and high-velocity data sets should not discount the fact that statistical principles – especially in relation to data representativeness and bias – remain very relevant for Big Data applications (see Box 3).

Data mining algorithms perform different types of operations (PCAST, 2014):

- **Classification**, where objects or events are classified according to known categories (e.g insurance companies employ classification algorithms to assign crash risk categories to drivers sharing certain characteristics).
- **Clustering**, where patterns of similarity are sought in the raw data.
- **Regression** or numerical prediction, where numerical quantities are predicted according to regression analysis.
- **Association**, where relationships between items in single or joined data sets are identified.
- **Anomaly detection**, where outliers or pattern breaks in data sets are identified.
- **Summarisation**, tabulating and presenting salient features within data sets.

Data mining approaches can be based on examples of relationships that provided by human operators and are used to guide the process or via unsupervised operation where patterns are discovered algorithmically. Especially in the latter case, the relationships exposed amongst data elements are correlative – that is, patterns are revealed but not their value or significance.

Modelling

Building and running models helps test hypotheses regarding the impact and importance of different variables in real-world systems. By simplifying simulating real-world phenomena, models help to characterise, understand, quantify and visualise relationships that are difficult to grasp in complex systems. Building models require data on baseline conditions and insight regarding the nature of relationships, either correlative or causal, between multiple phenomena. The arrival of Big Data has drastically increased the scale, scope and accessibility of modelling exercises though it should be noted that ability of these to accurately track the real world is linked not just, or even principally, to the quantity and quality of baseline data. Model construction and ensuring the right questions are asked remain essential to provide high-value outputs. Well-constructed models built on sparse data may be as or more effective than poorly designed models working on massive, real-time data sets.

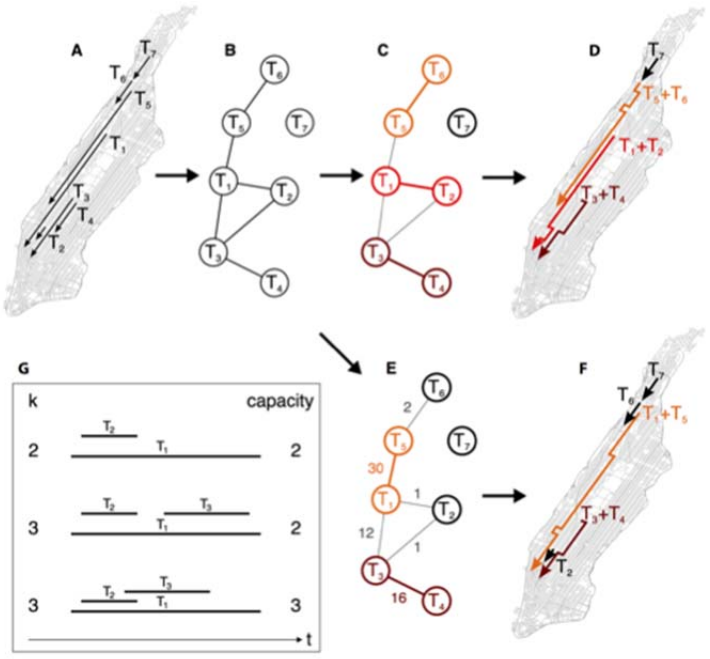
With this caveat in mind, Big Data sources and techniques have allowed for novel models to be constructed that provide new questions to be asked and new insights to be derived – as in the case of the recent HubCab initiative carried out by the MIT SENSEable City Lab in partnership with Audi and General Electric (GE).

HubCab analyses taxi trips to explore the benefits and impacts of vehicle sharing in New York city. The data was derived from the records of over 150 million trips made by 13 586 registered taxis in Manhattan during 2011. (Santi, et al., 2014) The GPS-enabled taxis reported on the geographic coordinates (longitude and latitude) and time of each trip's origin and destination, creating a map of pick-up and drop-off points.

When applying geospatial analysis and modelling techniques to this data, the monitored terrain is either broken into points, lines or polygon boundaries. (de Smith, et al., 2013) Using OpenStreemap, the HubCab team drew an open-licensed world map to obtain the outline of streets. These were then cut into over 200 000 segments of 40-metre lengths. Once footpaths, service roads and other street types unlikely to receive taxi traffic were removed, over one trillion possible routes were identified.

The resulting data set could be used to study more conventional queries, such as the location of the nearest taxi or most efficient route for a single trip. The innovative aspect of the HubCab project, however, is in its capacity to model and optimise trip-sharing opportunities through what is referred to as the "shareability network" (see Figure 5). At a conceptual level, this involves building a network of links between nodes. In the HubCab initiative the nodes represent individual taxi trips and the links connect trips that can be combined. Taxi routes are recalculated in real-time to pick up new passengers based on their current location and desired destination. (Santi, et al., 2014)

Figure 5. **Optimisation strategy for taxi sharing**



Source: Santi, et al., 2014.

Conclusions drawn from this exercise point to the potential impacts of taxi-sharing both at the level of the city and of the individual. In cities like New York, taxi services account for a major share of individual mobility. Hence, reducing the number of taxi trips would lead to dramatic reductions in air pollution and traffic congestion. In the analysis performed by the MIT SENSEable City Lab, substantial benefits were observed in triple trip sharing models over double trip sharing. Logically, the environmental gains associated with reducing the number of taxi trips by a factor of 3 are greater than those achieved by reducing by a factor of 2. However, trip sharing equates to longer wait times at pick-up points and less direct routes to individual destinations. The viability of the triple trip sharing solution depends largely on patient customers. This raises the question of whether cities like New York might push for fare systems that incentivise patience.

Visualisation and dissemination

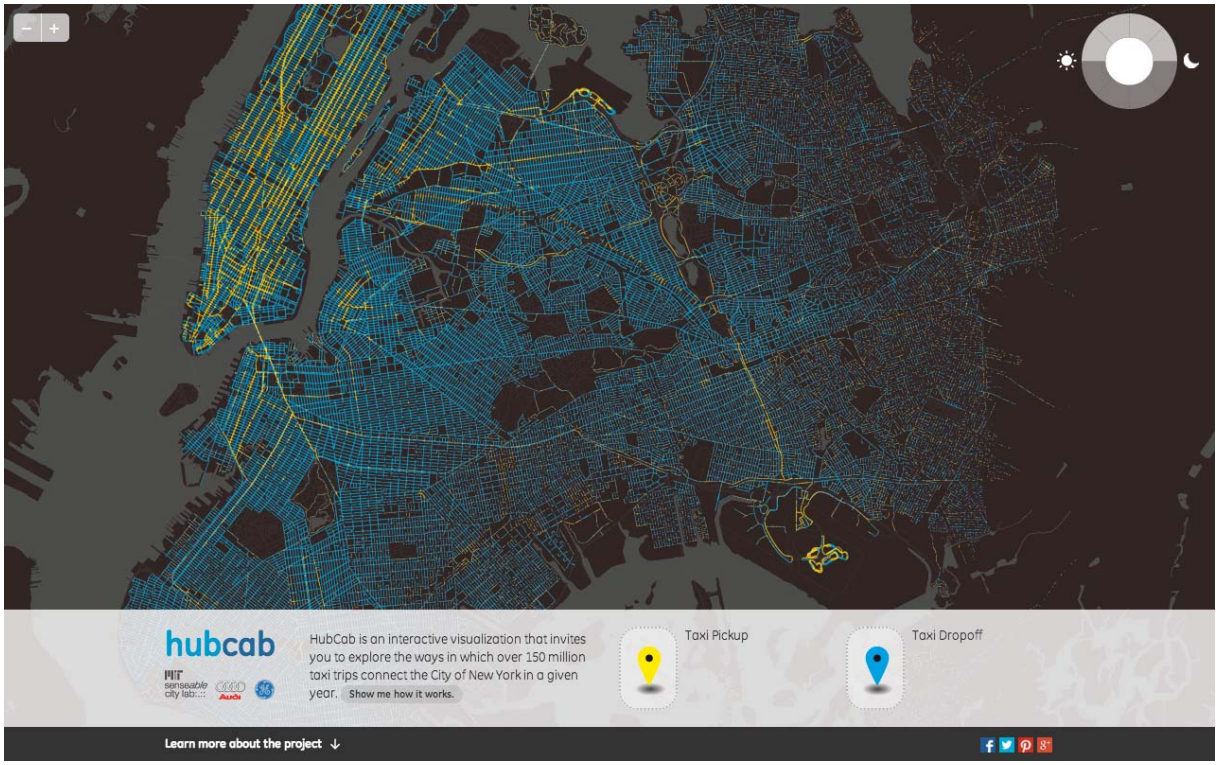
The immediate outcomes exhibited in the WikiCity Rome and HubCab examples are interactive maps that invite people to engage with mobility patterns shaped by their surroundings. The power of these visualisations is in their ability to inspire action from the most cost-effective and readily available urban actuators: citizens. Observing the real-time city becomes a means for people to understand the present and anticipate future urban environments. This could result in users electing to share a cab with a stranger to save on the cab fare or changing their mode of transport to avoid traffic congestion. On an urban scale, information delivery platforms capable of combining layers of information in a comprehensible manner can increase the overall efficiency and sustainability of city planning and regulation.

For centuries, humans have relied on the graphical or pictorial representation of data to make records of information accessible, comprehensible and, most importantly, appealing to the human mind. Consequently, the recent explosion of data has also seen a rise in tools focusing on effectively visualising that data. Many tools excel in depicting information using traditional methods like tables, histograms, pie charts and bar graphs. But bar charts couched in lengthy documents or slide presentations are often not

adapted to an audience beyond professionals. As data sets gets larger and efforts to exploit this data seeks to reach more people, the language of data visualisation must adapt and improve.

Citizens are living in an increasingly visual world, peering into screens of different sizes with incrementally superior resolutions with every device upgrade. As visual literacy rises, more professionals will be expected to know the language of data visulation. Visualisations – like the one of New York City taxi trips in Figure 6 - serve not only for information delivery but also for generating interest and making an impact – they are presentations of information framed at the convergence of art, digital media and information technology.

Figure 6. **Pick-up and drop-off points of all 170 million taxi trips over a year in New York City**

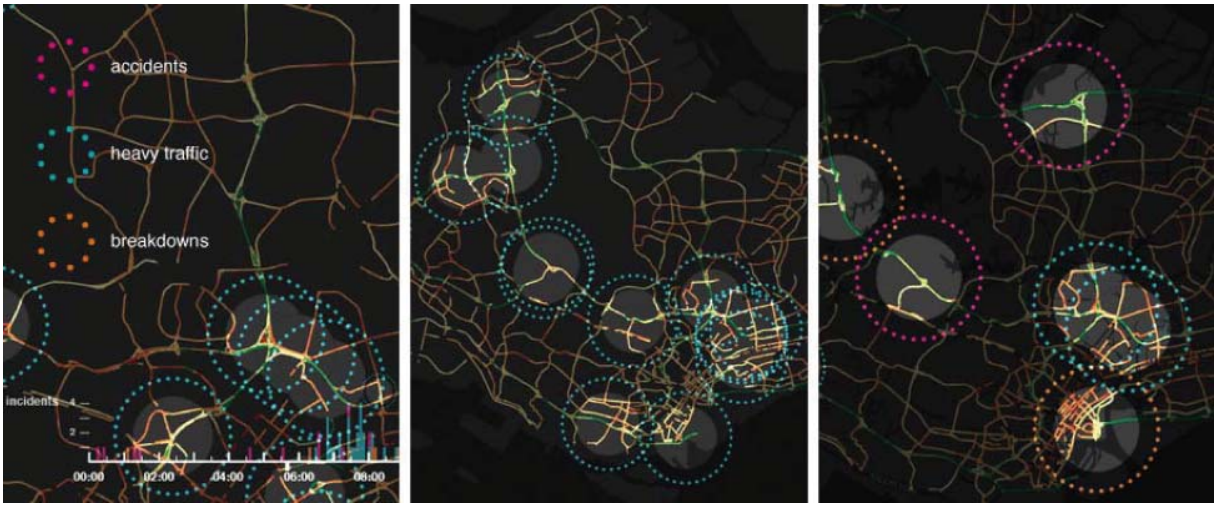


Source: MIT SENSEable City Lab.

The application of real-time visualisation tools to study traffic congestion has increased in recent years in response to the availability of data collected by traffic management centres and telecommunication companies through daily operations. Perhaps the best-known examples are online map services (Baidu, Bing, Google, Here, Naver, etc). These services use color-coded paths to indicate traffic speeds derived from road sensors and GPS-enabled vehicles and mobile devices. The more recent Google Waze maps out traffic incidents and other hazards reported by its 50 million users. (Forbes, 2014a)

In 2014, the researchers at MIT SENSEable City Lab sought to move beyond linking traffic accidents to congestion towards more predictive visualisation tools for traffic management centres. Their “Traffic Origins” study introduces time-lapse visualisations to observe how congestion propagates from a point source, or, when used in after-action reviews, to understand the effectiveness of mitigation measures. Despite having targeted expert users, the study emphasises the effectiveness of simple, aesthetically pleasing visualisations to communicate complex relationships to both traffic controllers and laypersons. (Anwar, et al., 2014)

Figure 7. **Traffic incidents visualised using the Traffic Origins approach**
(Expanding circles indicate incidents and the surrounding traffic conditions)



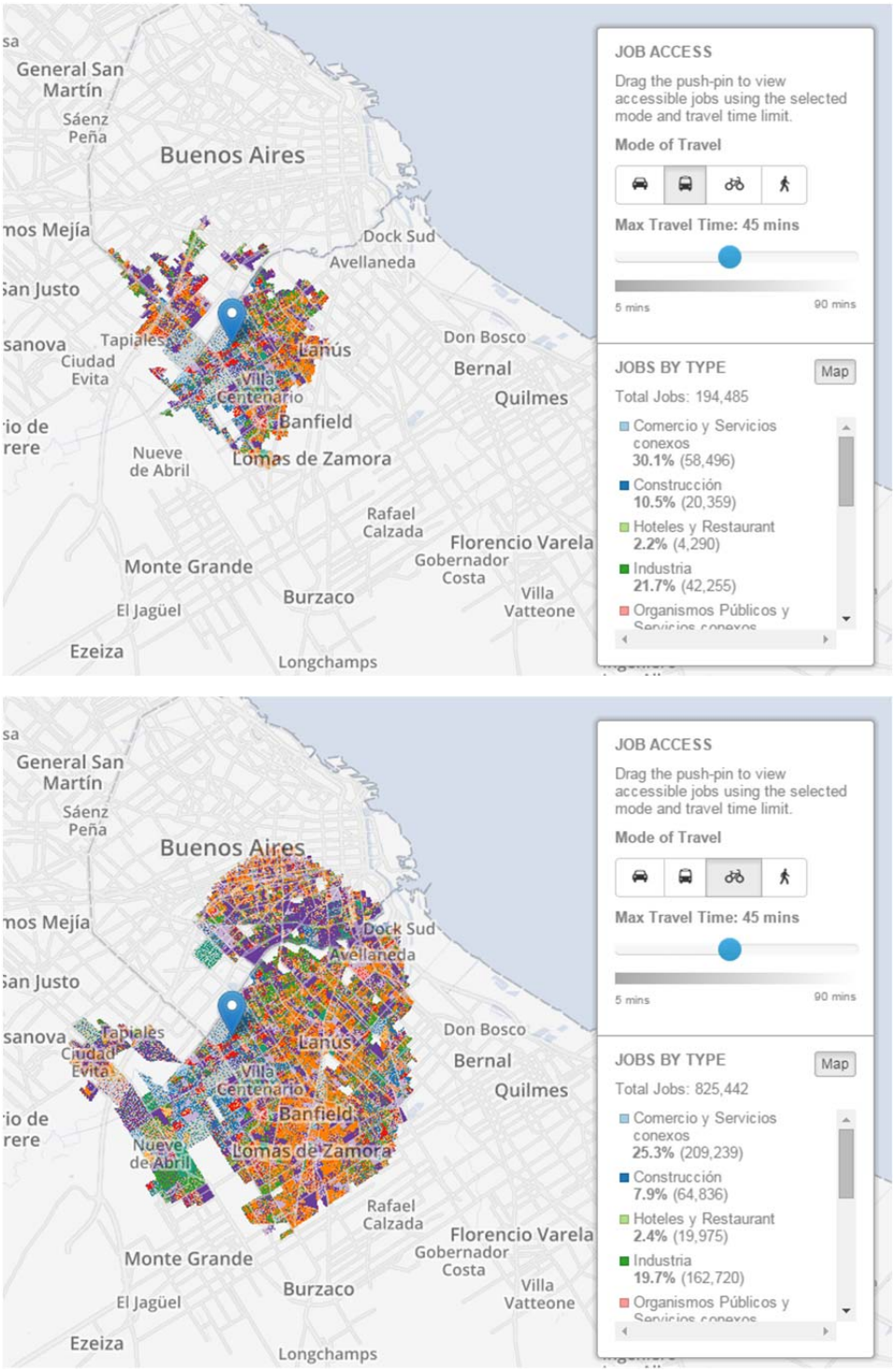
Source: Anwar, et al., 2014.

The next generation of visualisation tools for mobility applications should have a new visual language for data. This language should ot only be scientific, but also accessible, communicative and compelling. Ideally the tools should include the following capabilities:

- **Geo-spatial:** plotting data on customisable maps with additional geographical information.
- **Time resolution:** observing hourly, daily, weekly etc. patterns by easily switching between different time resolutions.
- **3D:** data depicted as 3D objects on a 3D globe for an immersive experience.
- **Animation:** free navigation to different periods of time in the data and comparison capabilities.
- **Interaction:** ability to pan or zoom to particular points and interact with them to display additional information.

These tools should be built to serve as a visualisation platform for experts and the average person alike – a knowledge base and centralised hub for visualising different types of data in unique and innovative ways. They could be made accessible through an Internet browser, eliminating the need to install special software. The interface should be simple and intuitive to encourage data interaction for the average user and eliminate the need for coding. Basic data operations available might include “grouping”, “filtering”, “adding metadata” and “identifying data types”, and would allow users transform and structure raw data into meaningful representations. For quick exploration and storage of massive data sets, the system’s back-end should be connected to a powerful cloud-computing engine, leveraging the open-source distributed computing framework. These frameworks are able to create visualisations from Big Data in a matter of minutes.

Figure 8. **Analysis and visualisation of labour market access in Buenos Aires**
(by mode of transport)



Source: World Bank and Conveyal, 2014.

Figure 8 depicts an analytic and visualisation-based tool developed by Conveyal in partnership with the World Bank. Building on prior work undertaken for the Regional Plan Association of the greater New York metropolitan area, this online tool allows the quantification and visualisation of access to labour markets in Buenos Aires, Argentina. It is built upon a hybrid platform comprising open-source spatial and transport network analysis software (OpenTripPlanner and Transport Analyst)[3], an open source format for encoding public transport schedules and route/stop locations (General Transit Feed Specification – GTFS), a global, open and collaborative cartographic database (OpenStreetMap), and an open source Java library for implementing on-line map-based visualisations (Leaflet[4]). Using this freely open and replicable platform users can dynamically visualise open data on jobs (including job distribution, access and sector) within the city. This in turn allows users to effectively derive insights that had been difficult or nearly impossible to elucidate previously.

For example, Figure 8 shows in the "villa miseria" of Villa Fiorito, in Buenos Aires. "Villas miseria" are low-income informal housing neighbourhoods, or slums. The online tool calculates the number and type of jobs accessible by these low-income residents given actual public transport schedules and real street topography. In a 45-minute (one-way) time window, Villa Fiorito residents can access 194 485 jobs using scheduled public transport. However, as the visualisation shows, they could potentially access four times as many jobs (825 442) by bicycle. This type of information, previously not as easily accessible, could more effectively guide investment in public transport or, alternatively, in safe bicycle infrastructure[5].

To promote universal access and shareability, visualisations should be able to be exported and distributed in a number of formats, from pictures to videos or web pages. Interactive tabletops and large interactive displays are more likely to appeal to novice users (Benko, et al., 2009) and attract attention in public spaces. (Isenberg, et al., 2010) The shareability of the visualisation platform itself becomes equally important as it allows for smart integration with other platforms like government web portals. It also allows new modules to be added by third parties in an open-source environment.

Visualisations based on data need not only be made using the types of "pre-digested" formats described above. Diffusion of the products of data analysis, as well as the distribution of some forms of data, already takes place using paper or digital documents. However, the utility of these channels is waning since these media are generally static. Online access to tabular or geographic data that is directly useable by various software platforms can be valuable for some forms of data analysis. Increasingly, however, diffusion of data will occur via machine-interpretable Application Programming Interfaces (APIs) or data formats that allow the integration of this data directly into different mobile or other applications. App-led data access on mobile handsets based on APIs has contributed to many new services that greatly facilitate navigation, travel, logistics and other transport-related services for individuals and businesses. Data diffusion via APIs will play an increasingly important role in citizen access and use of mobility-related data but the ultimate benefits of this data diffusion channel will depend on the terms of use associated with that data.

Data terms of use run the spectrum: from completely "open" terms with no restrictions on use and redistribution, to highly-constrained terms that allow commercial access to data only under a limited set of conditions. Open data proponents advocate for as much data as possible to be provided under open terms of use. This includes (nearly) all government-collected data, which they argue citzens already pay for via taxes, and therefore should have free access to. There are of course issues with the open data model relating to security and privacy concerns. However, many authorities are moving towards opening access to

[3] Conveyal.com, accessed April 2015.

[4] Leaflet is a service based on leafletjs, an open source JavaScript library for mobile-friendly interactive maps (leafletjs.com).

[5] The same analysis indicates that, in freely flowing traffic conditions, Villa Fiorito residents could potentially reach 25 times more jobs by car in 45 minutes than by scheduled public transport. This serves to explain the compelling attraction of individual motorised mobility. The potential of car travel is unlikely to be reached, though, as it is constrained by household incomes as well as by time losses linked to congestion at peak hours.

much of their data, especially in transport. On the other end of the spectrum, the limitations imposed in accessing commercially collected data are understandable, given that the companies collecting data, and delivering services based on the use and analysis of it, must deliver value to their owners and shareholders.

The debate surrounding “open” versus “closed” data access is one that has emerged alongside the development of Big Data and may prove to be a transitory debate, as data access channels and terms evolve over time. Ultimately, more flexible and perhaps more modular data terms of use will need to evolve to allow individuals, authorities and commercial operators to all make the most of Big Data.

3. Big Data, personal data and privacy

The exponential growth in the production and storage of mobility-related data has been accompanied by rising concerns relating to the adequacy of regulations ensuring privacy. These concerns have been fuelled by the personally identifiable nature of much of the data being collected and the fact that it is often collected without the full knowledge and informed consent of the data object. Even arguably "anonymous" data – for instance unencrypted transmissions from tire pressure monitoring systems – can now be easily cross-referenced with other sources of contextual data to link individuals to vehicles and vehicles to locations and trajectories. This can be seen to compromise reasonable expectations of personal privacy. Location-based data is particularly vulnerable to breaches in privacy. Yet much of the mobility-related data being produced today has a geospatial component.

Big Data analytics raises several issues relating to *generic* privacy threats. These are related to, but different from, threats related to breaches of cyber-security (see box 5). Privacy threats exist in relation to the collection or discovery of personal data by economic agents as well as by governments. In the latter case, recent allegations relating to large-scale data collection and storage by governments has raised acute concerns regarding the extent of state-sponsored "dataveillance". Mobility-related data, especially location-based data, raises a set of specific privacy and data protection concerns that will be addressed in this section.

Box 5. **Privacy vs. cybersecurity threats**

Privacy is "the claim by individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others, and the right to control information about oneself even after divulgating it" Alan F. Westin. (Westin, 1967)

Geo-spatial privacy is "the ability of individuals to move in public space with the expectation that under normal circumstances their location will not be systematically and secretly monitored for later use" Geospatial Privacy and Risk Management Guide, Natural Resources, Canada. (Natural Resources Canada, 2010)

Cybersecurity and privacy are two distinct but related concepts. They define and attempt to enforce policies relating to computer use and electronic communications. In particular, cybersecurity seeks to assess the following (PCAST, 2014):

- **Identity and authentication:** Are you who you say you are?
- **Authorisation:** What are you allowed to do to which part of the system?
- **Availability:** Can attackers interfere with authorised functions?
- **Confidentiality:** Can data communications be passively copied by someone not authorised to do so?
- **Integrity:** Can data or communications be actively manipulated by someone not authorised to do so?
- **Non-repudiation, auditability:** Can actions later be shown to have occurred?

The growing importance of network-based information and other connected services in transport obviously poses increased cyber-security risks, especially when networked-based systems interact directly or indirectly with primary control systems of vehicles.

A recent survey of potential cyber-attack vulnerabilities of US cars identified a number of potential attack surfaces posing variable risks depending on vehicle and sub-system design. It notes that manufacturers' anticipation of risks and design response is uneven, especially for secondary systems – including the distributed network of electronic control units (ECUs) within vehicles. Convergence between sensor networks and vehicle control systems (e.g. those found in automatic cruise control, lane keeping or parking assistance functions) poses particularly strong risks in that sensor inputs can potentially be modified or spoofed leading to degraded or lost control of vehicles (Miller & Valasek, 2014).

One response to the growing complexity of vehicle hardware-software interfaces is to develop comprehensive but variable secure system architecture based on critical risk assessments. For instance, EU project EVITA

outlines trust models and security measures that are based on core hardware security modules, hard-wired into each electronic control unit. These have various levels of strength depending on the mission-criticality of each ECU sub-system. In this case, ECUs controlling speed, forward and backward motion, steering and braking receive the highest level of protection and those controlling the on-board environment receive lower protection. At the core of the system are strong cryptographic methods that ensure the integrity and authenticity of system messages and allow for the detection of tampered, altered or non-authentic messages to or from vehicle systems. (EVITA, 2012)

Other cyber-security vulnerabilities remain. Two recent examples of cyber-attacks on indirect but mission-critical systems involve spoofing of Global Positioning System (GPS) signals used to pilot ships and aircraft. In the first instance, researchers were able to feed spoofed GPS coordinates to the automatic navigation system of a vessel allowing the attackers to gain full directional control. Although the attack was on an automatic navigation system, the method also fed incorrect GPS coordinates to all on-board GPS receivers. This meant the crew only received data indicating the vessel was still on-course. (Bhatti & Humphreys, 2014) In the second instance, an unmanned drone was spoofed into flying off-course. (Kerns, et al., 2014) It has been reported that a similar approach was used to gain access to and divert a US military drone in 2011. (Peterson, 2011)

Cyber-security risks are also a concern for transport systems more broadly, as increasing complexity and connectedness opens up new avenues for malicious interventions. The need for adequate data encryption protocols and practices for handling remote sensor data was highlighted by a remote and passive hack of unencrypted wireless road sensor data. Spoofing or manipulating this data could have important and severe consequences for traffic system operations that depend on these data feeds to coordinate emergency services, signal timing and traffic variable messaging systems, among others. (Cerrudo, 2014)

Both cybersecurity and privacy focus on potential damages that can be caused via malevolent or dangerous manipulation of computer and communication systems. Poor cybersecurity practices may lead to exposure, gathering and malicious use of personal data. However, privacy risks remain even in fully secured systems. Misuse of personal data in otherwise secure systems by authorised operators represents a violation of privacy policy, not of security policy. Similarly, violations of privacy may emerge after data fusion across multiple, fully secured, systems. These distinctions are important and demonstrate it is not enough to focus solely on cyber-security in order to ensure personal data protection. (PCAST, 2014)

Despite concerns over privacy, location-based data enhances services available to individuals and may contribute to significant improvements in safety, traffic operations and transport planning. For instance, E-call or E-911 services that enable vehicles to report their spatial coordinates to a central server in case of a crash improve response times and accuracy.

Likewise, individuals voluntarily contributing their spatial coordinates to applications have the expectation that this data will improve the quality of service they receive. This exchange is the basis for many services that provide real-time traffic information or personalised recommendations for hotels, restaurants and the like. But the value of this data is not limited to the explicitly identified first-use case. Companies can and have aggregated and sold this data – or the results of their analysis of this data – to other companies or public authorities for re-use, merging with other data and re-analysis.

There is a real tension between the value of large-scale flows of what may be weakly or non-anonymised data and the contribution that this same data can make to individuals and society. Part of this tension emerges because uses for this data may emerge only *after* the data has been collected or combined with other data, rendering notification of intent moot. There is also a realisation across many jurisdictions that the regulatory framework concerning data collection and use, for mobility-related data in particular, is poorly adapted to changes occurring in the volume and velocity of data collection. Finally, there is a fear that regulatory backlash against the collection and use of Big Data may hamper yet-undiscovered value in this data and curb the economic and social benefits the use of such data promises.

These challenges are acute and what is at stake is an erosion of personal privacy rights. These rights are ones that some in the private sector believe are, at best, not aligned to current technological developments

or, at worst, irrelevant[6]. Many proponents of this belief point to the very real improvements in service delivery that could be facilitated by data convergence. A world in which, for instance, personal and seamless mobility choices could be offered to citizens based on their individual characteristics needs and behaviours. However, as pointed out at a recent OECD Technology Foresight Forum, Big Data analytics should not be absolved of core ethical principles. Foremost amongst these principals is the understanding that "just because you can, doesn't mean you should". (OECD, 2013)

There is also a risk that regulatory backlash against Big Data fuelled by attacks on personal privacy may hamper innovation and curb the economic and social benefits the use of such data promises. Evolving regulatory approaches will have to simultaneously deliver on the pro-privacy and pro-innovation expectations of citizens.

Personal data protection frameworks

OECD personal data protection framework and guidelines

In 1980, the Organisation for Cooperation and Development (OECD) adopted the first internationally-agreed (but non-binding) guidelines framing the collection and exchange of personally identifiable data – the "Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data". These guidelines were issued in the hope that they would serve to balance privacy concerns with the benefits that derive from the international flow of information. They have served as the underlying framework for data protection and privacy laws throughout the OECD and elsewhere. They are not law and their implementation into national regulations has not been uniform. Yet they have served as a basis for the legal framework surrounding data protection in many jurisdictions. The guidelines are articulated around eight principles addressing (for more detail, see box 6):

- Limitations to data collection.
- Data accuracy and relevance to stated use.
- Communication of the purpose for data collection and limitations of data use to that purpose.
- Restrictions on data disclosure.
- Data safeguard and security measures.
- Transparency regarding the use, and changes in use, of personal data.
- The right of individuals to have access to or control the use of their data.
- Accountability of data controllers regarding the above principles.

As a general guiding framework for data privacy policies, these principles have served relatively well over the past 35 years. However, in practice and in important points of detail, their continued implementation has become increasingly problematic. The environment in which data privacy efforts has evolved since the creation of the 1980 OECD Guidelines, as outlined in previous sections. In particular, the implementation of these principles is challenged by a number of factors (OECD, 2013):

- The growing ubiquity of data collection across multiple platforms.
- The volume and velocity of data produced and collected.
- Data fusion and aggregation efforts that potentially de-anonymise data.
- The range of analytical methods and techniques that reveal information regarding individuals, their behaviour and associations and their interests.

[6] Some leaders in the IT sector have intimated that privacy rights as they have been interpreted in the past are no longer feasible. Eric Schmidt, Executive Chairman of Google, is quoted as saying: "If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place" – (Banks, 2011)). Scott McNealy, CEO of Sun Microsystems is on reecord saying "You have zero privacy anyway … Get over it." (Sprenger, 1999).

- Ex-post data mining and re-use of data in ways that were not originally intended.
- The growing range of threats to the privacy of personal data.
- The number of actors that can either compromise personal data privacy or act to protect it.
- Citizens' insufficient or ill-informed knowledge regarding the complexity of interactions relating to the collection and use of personal data.
- The ease of access to and global availability of personal data.

Furthermore, at the core of the OECD Guidelines (and other initiatives and laws inspired by OECD Guidelines them) is the notion that only personal data – e.g. information that could identify an individual such as a name or a civil registration number, or that could reasonably be used to identify an individual such as an address – should be governed by those principles. Personal data that has been de-identified – e.g. personal data that has had individual identifiers removed or that has been modified in such a way as to make re-identification reasonably unlikely – arguably falls outside of the scope of these principles. However, the combinatory aspect of Big Data blurs the line between personal and "anonymous" data. There is a real risk that the latter may serve to re-identify the former when used in conjunction with other ostensibly anonymous data sources.

In response to these challenges, the OECD reassessed the 1980 principles and adopted a revised set of principless (OECD, 2013) that provided new guidance, notably in the areas of accountability and notification of security breaches. The 2013 guidelines also highlighted the need for further research on the evolving nature of consent, purpose limitation and of the role of the individual in data privacy. Ultimately, however, the expert group preparing the OECD Council document could not reach consensus regarding the modification of the original eight core principles and thus these remain unchanged.

A number of other national or international privacy protection guidelines have been based on the OECD principles each emphasising or de-emphasising certain aspects. (Cate, 2006)

EU personal data protection frameworks

In 1995, the EU Data Protection Directive (Directive 95/46/EC) identified a set of principles that largely incorporates those of the OECD guidelines. Two additional principles were added: one on independent oversight of data controllers and processors; and one outlining the legally enforceable rights of individuals against data collectors and processors.

The ePrivacy Directive of 2002 (and revised in 2009) speaks more specifically to location data. Location data in the Directive refers to:

> "The latitude, longitude and altitude of the user's terminal equipment, to the direction of travel, to the level of accuracy of the location information, to the identification of the network cell in which the terminal equipment is located at a certain point in time and to the time the location information was recorded." (Directive 2009/136/EC)

However, the ePrivacy Directive considers location data only in the context of binding rules for telecoms operators (Cheung, 2014) – this is an important distinction that largely ignores other collectors, aggregators and users of location data (see Box 8 on location accuracy). Companies identified as "information society services" fall outside of the regulatory framework of the ePrivacy Directive, even if the location data they collect is transmitted via telecom operators' networks.

In 2012, the European Commission outlined a new directive – the General Data Protection Directive – to replace that of 1995. This new regulation builds on the prior text but outlines several enhanced privacy protections. These include:

- The "right to be forgotten" to help manage online data privacy risks. Individuals can request that data pertaining to them be deleted if there are no legitimate grounds for keeping it.

- Improved visibility and access to one's personal data and the right to transfer one's personal data from one service provider to another.
- Requirements for clear and explicit consent to collect and use personal data.
- Improved administrative and judicial remedies in cases of violation of data protection rights.
- More robust responsibility and accountability for those collecting and processing personal data – e.g. through data protection risk assessments, data protection officers, and the principles of "Privacy by Design" and "Privacy by Default". (European Commission, 2012)

Crucially, the General Data Protection Directive simplifies the definition of personal data to "any information related to a data subject". This explicitly includes location data. In contrast with recent US discussions on data protection, the proposed EU Directive upholds the need for robust notice and consent prior to the collection and processing of personal data from data subjects. The EU approach also stresses the need to limit data collection to within its stated purpose, to ensure restrictions on automated processing and the need for independent regulatory oversight and robust enforcement mechanisms for transgressions regarding personal data rules (as inscribed in national laws).

European approaches to data privacy are seen by many as the standard in most parts of the world where data privacy laws have been enacted, and this influence seems to be growing. (Greenleaf, 2012) (Schwartz, 2013) The General Data Protection Directive is set to be adopted in 2015 and will represent the strongest implementation of the data protection rules inspired by the OECD Guidelines. This may foreshadow a strengthening of personal data protection rules in a number of jurisdictions – but not all as discussed below.

Box 6. OECD Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data

As adopted by the Council of Ministers of the Organisation for Economic Cooperation and Development on 23 September 1980.

1. Collection limitation principle

There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

2. Data quality principle

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.

3. Purpose specification principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

4. Use limitation principle

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with the "purpose specification principle" except:

a) with the consent of the data subject

b) by the authority of law.

5. Security safeguards principle

Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.

6. Openness principle

There should be a general policy of openness about developments, practices and policies with respect to

personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

7. Individual participation principle

An individual should have the right:

a) To obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him.

b) To have communicated to him, data relating to him:

i) within a reasonable time

ii) at a charge, if any, that is not excessive

iii) in a reasonable manner

iv) in a form that is readily intelligible to him.

c) To be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial.

d) To challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.

8. Accountability principle

A data controller should be accountable for complying with measures which give effect to the principles stated above.

APEC and other Asian personal data protection frameworks

In 2004, the Asia-Pacific Economic Cooperation (APEC) forum adopted the APEC Privacy Framework that, in its 9 principles, consciously builds on the OECD Guidelines. It extends the notice and consent principle to include a call for "clear and easily accessible" statements, along with a call for all reasonably practical steps be taken to provide consent either before or at the time of collection – or soon thereafter. This change in language is a concession to the fact that notice and consent may be difficult to achieve in certain scenarios, including those that involve real-time or machine-to-machine digital data collection.

The APEC Privacy Framework also introduces a principle on preventing harm –in particular, the need for data protection efforts to be commensurate with the potential for harm from personal data release or discovery. Not all personal data poses the same risk for creating negative outcomes for data subjects and data protection efforts should account for this. (APEC, 2004)

Within Asia, many countries have put in place personal data protection frameworks that are inspired by those of the OECD/EU and APEC with some important distinctions. In particular, many countries in the region have adopted an EU-like approach that encompasses formalised notice and consent requirements for data collection and use, the creation of a limited, and often time-bound, set of conditions on data processing and controls on the forward movement of data to third parties or other jurisdictions.

In Japan, for instance, notice of data collection can be provided directly to an individual or via public announcement. In either case, explicit consent is not required if the purposes for use of the data have previously been specified in the personal notice or public announcement. Specific opt-in consent is only required in cases that go beyond the announced data usage. (Rich, 2014) The South Korean Data Protection Act has stringent guidelines in place regarding the need for notice and express consent for collection, use and transfer of personal data. The notice must specify the intended use of personal data and its eventual disclosure to third-parties (Rich, 2014).

United States' personal data protection frameworks

In 1998 and again in 2000, the US Federal Trade Commission (FTC) reported to the US Congress those principles it felt should frame US data protection policy. These related to (Cate, 2006):

- notice given to citizens regarding data collection and intended uses
- the choice framework offered to citizens relating to personal data collection and use and the need to obtain consent
- the possibility for citizen's to access data about themselves and to contest this data on grounds of inaccuracy or incompleteness
- the responsibility for data controllers to keep personal data secure and accurate, and
- the need for enforcement and redress mechanisms to secure the above.[7]

Noticeably absent from the FTC communication to Congress is an explicit data collection limitation principle and one relating to data quality.

In 2012, the Federal government issued a report outlining a Consumer Privacy Bill of Rights (CBRB) which addresses commercial but not public sector uses of personal data. While drawing on precedent set by the OECD Guidelines, the CBRB introduces the concept of context-dependency – that is that citizens should expect that data controllers and processors will collect, analyse and dispose of data in line with the context in which citizens supplied the data. (PCAST, 2014)

Generally, the United States and the EU have differed in their approaches to the protection of personal data in a number of significant ways. While the EU has favoured a single framework for addressing personal data protection, the US (and some other countries) have favoured a sectoral approach and, in particular, distinguish between the private and public sector when making rules governing the collection, processing and use of data.

The way in which personal data is treated by the law in the US also differs according to the entity that controls the data or the type of data recorded. This sectoral approach has tended to place higher restrictions on incumbent industries in existing and more regulated sectors such as telecommunications in contrast with many recent new and emerging businesses in the IT sector. (Schwartz, 2013)

In contrast with the EU, the US generally adopts an approach best characterised by "regulatory parsimony" in the field of personal data protection. For instance, the US approach generally allows data collection and processing unless a specific law prohibits it whereas the EU requires express legal authorisation for data processing. The divergence from an OECD/European-style approach to personal data protection seems likely to increase given recent developments.

In January 2014, the Executive Office of the President of the United States announced an initiative to investigate ways in which Big Data will affect the lives of citizens – including their privacy rights – and to suggest new directions for policies accounting for these changes.[8] The output from that initiative highlights the difficulty traditional notify-and-consent frameworks face in light of evolving data collection and processing practices.

In the first instance, it notes that the context-dependency principle of the CBRB fails to account for the changing nature in which personal data is collected and how personally identifiable data can be inferred from "anonymous" data. Data that can be tied to an individual is increasingly not provided by the data

[7] The enforcement/redress principle was dropped in the guidance issued by the FTC in 2000.

[8] The initiative was led by White House Counsellor John Podesta and informed by an advisory group of the President's Council of Advisors on Science and Technology. The outcome of this work was two reports: *Big Data: Seizing Opportunities, Preserving Values* (Executive Office of the President, 2014), and *Big Data and Privacy: A Technological Perspective* (PCAST, 2014). Both outline future policy directions that could be frame US policy in the matter, although neither are binding.

subject itself but harvested from ambient data sources or emerges ex-post from co-mingling of various data streams. This risk increases with the growth in data volumes and, especially, the growth in data sources like smart phones and other personal mobile devices.

In the second instance, it stresses that the value emerging from data collection and analytics increasingly emerges after the fact and after combining and mining diverse data streams. The President's Council of Advisors on Science and Technology suggests that compelling benefits emerge from retaining personal or potentially personably identifiable data for subsequent re-analysis as analytical techniques evolve. Arbitrarily requiring data to be deleted would, they argue, stifle innovative insights and uses that may otherwise emerge.

Effectively anonymising or otherwise obscuring personal information within retained data would moot privacy concerns relating to data retention. But the proliferation of Big Data sets and the multiplication of data sources erode the effectiveness of anonymisation and de-identification techniques.

The Executive Office report also highlights the challenge of enacting meaningful consent from citizens and consumers. Consumers are often presented with an exhaustive notification text while they are trying to access a particular service. The app or service provider may have spent considerable effort and legal expertise crafting a thorough consent notification document that does not conflict with the service provider's commercial interest. However, the consumer in a hurry to access a service rarely reviews such a notification, rendering any consent they provide as superficial or blind. In this way, data protection efforts may in fact be weakened rather than strengthened by the current notice and consent framework.

Additionally, the intersection between the need to provide notice and consent under current regulatory frameworks and the practice of broadly and open-endedly mining data has resulted in increasingly wide-ranging and permissive privacy notice and consent clauses. This has led to a situation where data subjects may willingly agree to an erosion of their privacy rights because they are neither engaged in a meaningful way and because of overly broad conditions regarding use of their data.

Faced with these challenges, the report of the Executive Office asks "whether a greater focus on how data is *used* and *reused* would be a more productive basis for managing privacy rights in a big data environment". (Executive Office of the President, 2014) In particular, the President's Council of Advisors on Science and Technology notes that:

> "…the non-obvious nature of Big Data's products of analysis make it all but impossible for an individual to make fine-grained privacy choices for every new situation or app. For the principle of Individual Control to have meaning, [we] believe that the burden should no longer fall on the consumer to manage privacy for each company with which the consumer interacts by a framework like "notice and consent." Rather, each company should take responsibility for conforming its uses of personal data to a personal privacy profile designated by the consumer and made available to that company (including from a third party designated by the consumer)". (PCAST, 2014)

This view is shared by many industry actors (see below) and, if implemented, would signal a strong break from existing data protection frameworks like those outlined in the OECD principles and in EU legislation.

Private-sector initiatives

Since the update to the OECD principles, two private-sector initiatives have sought to push the regulatory framework further still. The first initiative, led by Microsoft Corporation and organised by the Oxford University Internet Institute, released the report "Data Protection Principles for the 21st Century". (OII, 2014) The second initiative, organised by the World Economic Forum (WEF) in collaboration with the Boston

Consulting Group (BCG), published its analysis in the report "Unlocking the Value of Personal Data: From Collection to Usage". (WEF, 2013)

Both initiatives share the view that the 2013 update of the OECD Guidelines is anachronous and insufficiently addresses the rapid and fundamental changes surrounding data collection, analysis and use. They detail what they see as fundamental differences with traditional approaches to data protection as exemplified in the OECD rules (see Table 1).

Table 1. **Emerging private sector perspectives regarding data use and privacy**

Traditional Approach	Emerging New Perspectives
Data actively collected with data subject and data user awareness.	Data largely from machine-to-machine transactions and passive collection – difficult to notify individuals prior to collection.
Definition of personal data is predetermined, well-identified and binary (personal/not personal).	Definition of personal nature dependant on combinatory techniques and other data sources or may be contextual and dependent on social norms.
Data collected for a predetermined specific use and for a duration in line with that use.	Social benefits, economic value and innovation come from co-mingling data sets, subsequent uses and exploratory data mining.
Data accessed and used principally by the data subject.	Data user can be the data subject, the data controller and/or third party data processors.
Individual provides consent without full engagement or understanding.	Individuals engage in meaningful consent, understand how data is used and derive value from data use.
Data privacy framework seeks to minimise risks to individuals.	Data protection framework focuses more on balancing individual privacy with innovation, social benefits and economic growth.

Source: (WEF, 2013).

The WEF and Oxford/Microsoft initiatives stress that the current framework offering a binary (yes/no) and one-time consent is out of step with pro-innovation data collection and use practices, as well as ignoring the fact that data subjects are also data producers themselves. In particular, WEF outlines what it has identified as four main shortcomings of the revised OECD Guidelines:

- They fail to account for the possibility that new and beneficial uses for the data will be discovered, long after the time of collection.
- They do not account for networked data architectures that lower the cost of data collection, transfer and processing to nearly zero, and enable multi-user access to a single piece of data.
- The torrent of data being generated from and about data subjects imposes an undue cognitive burden on individual data subjects. Overwhelming them with notices is ultimately disempowering and ineffective in terms of protection.
- In many instances - for example, while driving a car or when data is collected using many machine-to-machine (M2M) methods - it is no longer practical or effective to gain the consent of individuals using traditional approaches. (WEF, 2013)

In many fundamental ways, the conclusions of the WEF and Oxford/Microsoft initiatives echo the recent work of the Executive Office of the President in the United States. This could be because the latter drew heavily on industry representatives involved in the former. They also run counter to the most recent developments in EU personal data protection.

In light of the ambiguous nature of personal data and the difficulty in reconciling the principles of purpose specification, use limitation, notification and consent with evolving Big Data collection and analytic

practices, both the WEF and Oxford/Microsoft initiatives emphasise a need to re-visit some of the core principles of the OECD Guidelines.

Most important is the need to re-formulate personal data protection practices that currently rest on providing notice and consent, towards principles outlining clear rules and sanctions on allowable uses of personal data. Oxford/Microsoft's "Data Protection Principles for the 21st Century" summarises this position:

> "A revised approach should shift responsibility away from individuals and towards data collectors and data users, who should be held accountable for how they manage data rather than whether they obtain individual consent. In addition, a revised approach should focus more on data use than on data collection because the context in which personal information will be used and the value it will hold are often unclear at the time of collection." (OII, 2014)

Under this premise, Oxford/Microsoft proposes a revised set of principles adapted from the 1980 OECD Guidelines. These revised principles make a distinction between principles that should apply to data collection and those that should apply to data use. Furthermore, it broadly expands the boundaries of data use. It calls for potential harms that could stem from discovery of personal data to be more rigorously balanced with the benefits that could stem from personal data collection and use. It also stresses that the principles it proposes are only applicable to personal data that has not been de-identified. Finally, it retains the notion of notice and consent as per the original OECD Guidelines but loosens this requirement by calling on data collectors to "evaluate whether individuals might reasonably anticipate that their data will be collected in determining whether consent is required". (OII, 2014)

WEF notes a number of emerging issues that will have to be addressed if the shift outlined above were to become operationalised:

Firstly, a shift from passive and binary consent to more engaged models of user involvement with their data brings a need to trace data to a source and to an individual. Attaching metadata that is both Persistent and un-purgeable to raw output from sensors or as early as possible in the data collection chain could help. This is an approach promoted by "Privacy by Design" advocates (below and box 7).

Secondly, more meaningful options will have to be made clear to citizens regarding the uses of their personal data. In this context, offering a broad palette of context-specific use consent choices – e.g. allowing use of personal data for medical or emergency services but not for targeting advertisement – may be the way forward. It may also be useful to distinguish between using personal data to generate broad insights (e.g. in support of transport planning) versus use of these insights (e.g. particular daily mobility patterns) applied to an individual.

Finally, in addition to regulatory frameworks, WEF notes that new types of user-centric arrangements (peer networks, privacy "labels", designated privacy profile "managers", etc.) have yet to be explored. These arrangements can help to further engage individuals in ensuring that their preferences regarding the use and re-use of their personal data are met.

It should also be noted that, generally, the drift away from existing notice and consent frameworks to one more focused on use of data presupposes the presence of a well-funded and competent regulatory agency. Such an agency would oversee data uses and resolve, and possibly prosecute, conflicts. Furthermore, such an agency would need to be equipped to address asymmetric and extended legal struggles with large and powerful multi-national corporations. There is evidence that such an approach may work – examples in the field of competition policy come to mind – but it is far from clear that authorities have an appetite for creating such strong and well-resourced agencies given the general move away from expensive and powerful regulatory control in many countries.

"Privacy by Design"

Much of the regulatory discussion regarding the collection and use of personal data says little about the *design* of data collection mechanisms and practices. In most cases it assumes that little will evolve in terms of the way in which data and data systems embed and transmit bits of personably identifiable information. This view is contested, especially by those who advocate the "Privacy by Design" approach (see box 7). This approach holds that data collection systems and practices should be designed (or re-designed) from the ground up to include strong and irreversible pro-privacy measures for data collecting and handling systems – even in the design of machine logging protocols and sensors.

Box 7. **"Privacy by Design"**

"Privacy by Design" is an approach to the design of data collection mechanisms and practices developed by Ann Cavoukian, Executive Director of the Institute for Privacy and Big Data at Ryerson University, Canada, and former Information and Privacy Commissioner for the Province of Ontario. "Privacy by Design" is based on the principle that strong pro-privacy measures should be addressed at the design stage for data collection and analysis, not retro-fitted ex-post once data has been collected and analytic systems developed. The approach is comprised of seven core principles.

1. Proactive, not reactive – preventative, not remedial

The "Privacy by Design" (PbD) approach is proactive rather than reactive. It anticipates and prevents privacy invasive events before they happen. PbD does not wait for privacy risks to materialise, nor does it offer remedies for resolving privacy infractions once they have occurred – it aims to prevent them from occurring. "Privacy by Design" comes before-the-fact, not after.

2. Privacy as the default setting

"Privacy by Design" delivers the maximum degree of privacy by ensuring that personal data are automatically protected in any given IT system or organisational practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy – it is built into the system, by default.

3. Privacy embedded into design

"Privacy by Design" is embedded into the design and architecture of IT systems, organisational and business practices. It is not bolted on as an add-on, after the fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

4. Full functionality: positive-sum, not zero-sum

"Privacy by Design" seeks to accommodate all legitimate interests and objectives in a positive-sum, win-win manner, not through a zero-sum approach, where unnecessary trade-offs are made. "Privacy by Design" avoids the pretence of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.

5. End-to-end security: full lifecycle protection

"Privacy by Design", having been embedded into the system prior to the first element of information being collected, extends securely throughout the entire lifecycle of the personal data involved – strong security measures are essential to privacy, from start to finish. This ensures that all personal data are securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, "Privacy by Design" ensures cradle to grave, secure lifecycle management of personal information, end-to-end.

6. Visibility and transparency: keep it open

"Privacy by Design" seeks to assure all stakeholders that whatever the organisational or business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike.

7. Respect for User Privacy: keep it user-centric

Above all, "Privacy by Design" requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. "Privacy by Design" is user-centric.

Source: Cavoukian, 2012.

"Privacy by Design" calls for privacy-protective measures to be built directly into the design and operation of technology as well as into data management practices surrounding data systems (e.g. work processes, management structures, physical security perimeters, linkages in networked infrastructure) (Cavoukian, 2010). Proponents of "Privacy by Design" feel that this approach enhances the ability for data analytics (and Big Data analytics in particular) to deliver value since fewer conflicts emerge regarding the release of personally identifiable information. Both the EU Data Protection Directive and the WEF report outlined in the previous section note the importance of "Privacy by Design" approaches.

"Privacy by Design" in the context of big data rests on seven features (Cavoukian & Jones, 2012):

1. **Full attribution**: Every observation or record should be traceable to its point and time of creation. This includes data related to the type of machine logging process or sensor platform involved. Merge-purge data processing where some data or metadata fields are discarded when data is combined must be avoided. This allows the fine-grained implementation of user data preferences and ensures their backwards compatibility, even in merged data sets. Full provenance metadata facilitates data accountability, reconciliation and audit and is essential for the application of personalised privacy controls on data. According to "Privacy by Design" advocates, full attribution of data should be made the default setting in systems and should not be allowed to be turned off.
2. **Data tethering:** Additions, modifications and deletions in data systems should be accounted for in real time and should propagate throughout all other data systems that include the data in question. When data is modified in one system, these changes should cascade throughout the shared data ecosystem irrespective of who manages each linked or merged data set. This includes changes in the privacy setting for that data. This feature pre-supposes strong access rights that are linked to the original data collector or user. Data tethering should be the default according to "Privacy by Design" and this should be an inalterable setting of data management systems.
3. **Analytics of anonymised data**: Anonymising or encrypting personally identifiable data as well as quasi-identifiers is an essential step in ensuring that Big Data analytics do not encroach on individuals' privacy rights. Adequate de-identification techniques are essential to pro-privacy data analytics. These de-identification techniques should be adapted to the potential harm that may emerge from re-identification as well as to the potential difficulty or barriers to re-identification. Seemingly anonymous location-based data (e.g. anonymous GPS tracks) requires particular attention since it serves a strong quasi-identifier in combination with other data sources. Data anonymisation or encryption should occur as early in the data collection and analysis chain as possible.
4. **Tamper-resistant audit logs**: Every material interaction upon a data set containing personal information, identifiers or quasi-identifiers should be logged in a tamper-resistant manner by default and by design. System administrators should be subject to, and should not be able to override, this setting.
5. **False negative favouring**: When personal data entailing civil liberties are concerned, it is better to design privacy protection policies that favour false negatives rather than false positives – e.g. it is better to miss a few things than to inadvertently make claims on the basis of personal data that are not

true. Algorithmically-favouring false negatives should be the default in data collection and analysis systems unless there are compelling and transparent reasons for the contrary.

6. **Self-correcting false positives:** With every new data point created or presented, prior assertions based on that data should be re-evaluated to ensure they remain correct and if not, new corrected assertions should be propagated backwards and forwards in real time.

7. **Information transfer accounting:** Every onward transfer of data either to human observation or to machine systems should be logged to allow stakeholders (data controllers or data objects) to understand how their data is flowing and is being used.

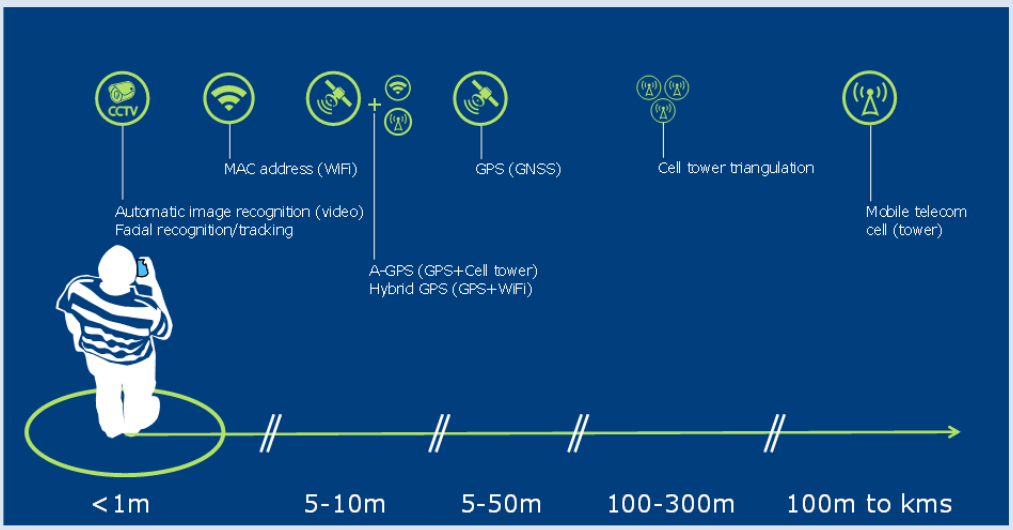
These hardware and system design features could go a long way to improving the systematic definition, tracking and application of individual privacy settings across multiple operators and jurisdictions. However, at present they are not bundled into any of the data protection frameworks outlined earlier. Furthermore, they are not operationalised in national data protection legislation or comprehensively integrated into industry codes of practise. Work on updating data protection practices should draw on these design standards in order to ensure that strong personal data protection settings are at the heart of new data collection and analysis efforts. This will likely entail codifying "Privacy by Design" into national and international personal data protection rules in a way compatible with technological developments and industry practices.

Privacy and location-based data

Personal data includes information such as name, address, sex, employment history, marital status, religion, finances, and unique identifiers such as passport or identity card numbers. This data can reveal facts that individuals may not agree to share broadly. This data can also be used to discover further personal information when combined with other data sources. All of the privacy frameworks outlined above are specifically concerned about these types of data and many efforts have gone into their protection or de-identification. However, location data can also be seen as very personal.

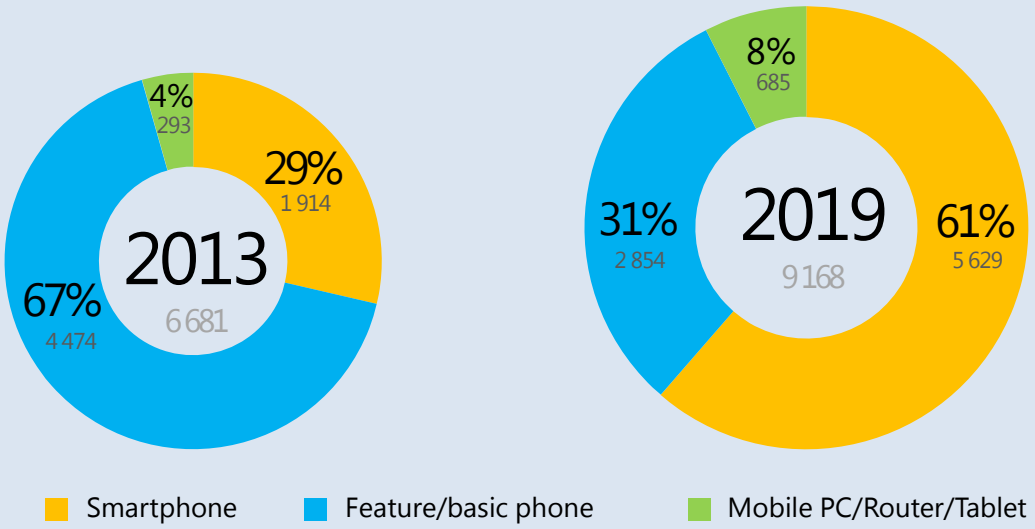
Box 8. **Precision of geo-location technologies**

Precision of geo-location technologies



Locating and tracking individuals at higher than one meter precision (up to a precision of a few centimetres) in both outside and indoor environments is currently feasible. It will likely become standard – at least in urban areas – as current location-sensing technologies become ubiquitous. In a large measure, the widespread penetration of mobile phone technology – and especially that of smartphones – makes this possible. The same location technologies deployed in the current generation of mobile phones are also migrating to vehicles enabling precise and persistent tracking.

Global Mobile Subscriptions (millions)



Source: Ericsson, 2013.

In 2013, Ericsson estimates that there were 6.9 billion mobile subscriptions globally (including smartphones, basic phones PCs, routers, tablets and M2M devices using mobile subscriptions). All of these devices access cellular networks (2G, 3G, 4G LTE and others) and in so doing create a log of geo-localised data used by network operators to provide for seamless call service. In addition, almost all new phones (smartphones and basic phones) have global positioning chipsets and all new smartphones have Wi-Fi capability as well as numerous embedded sensors. These technologies allow operators and application developers to have access to extremely precise location data – an option that many app developers are exercising. In 2014, the Global Privacy Enforcement Network, a group of 39 national and international privacy enforcement authorities, conducted a review of popular mobile apps in their respective countries and regions. It found that 32% of the 1 211 apps investigated sought access to the devices' location data. (OPCC, 2014) Moreover, according to one survey in the United States, half of all mobile phone users and 74% of smartphone users use location-based services. (Zickhur, 2012)

Cellular base station-based localisation

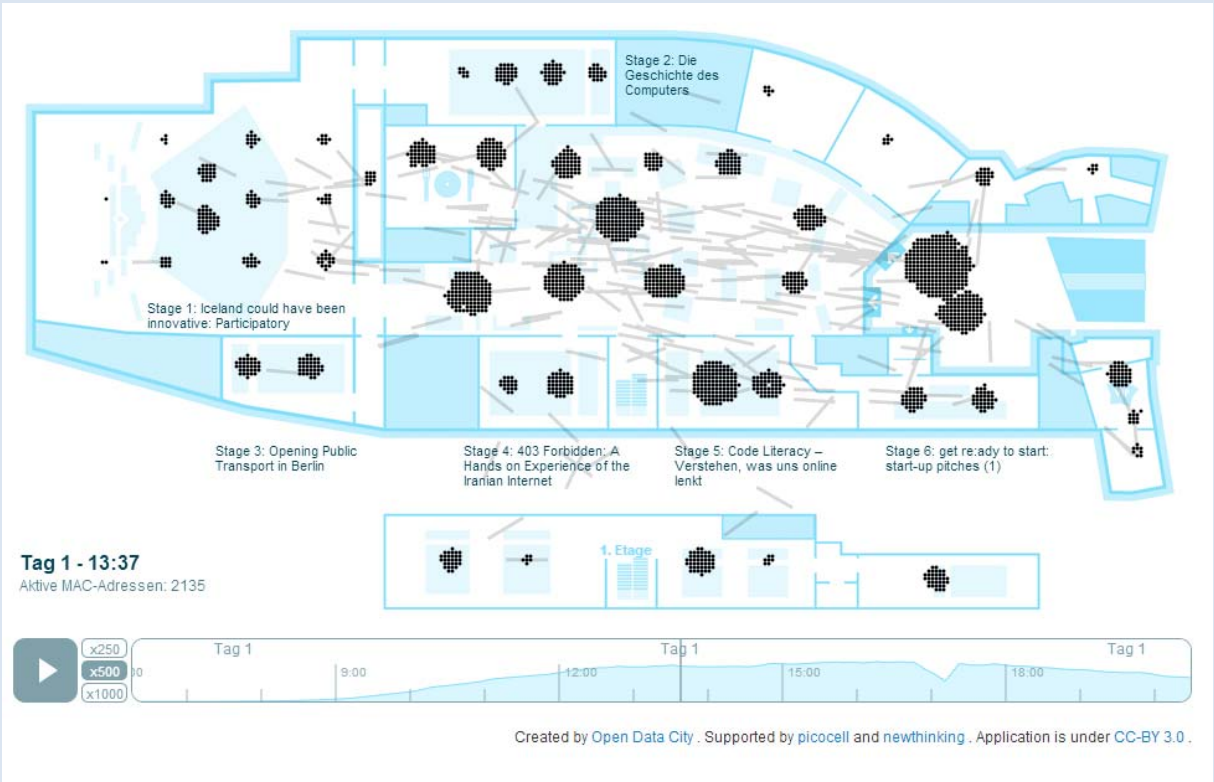
Mobile phone connectivity requires a near-constant series of handshakes between handsets and the network via cellular communication antennas. These antennas, and the cells they define, are densely located in urban areas and less densely located elsewhere. Mobile phones regularly and frequently "ping" cellular networks in order to determine their location, which is calculated by determining the location of the cell antenna closest to the handset. This results in a precision equal to the size of the cell, which can range from a few hundred metres in urban areas to a few kilometres elsewhere. Cellular operators also keep track of hand-offs from one cell to another in order to provide seamless service while the handset is in use. In-call hand-offs and in-call dwell times (e.g. the amount of time an in-use mobile phone remains in a single cell) provide rough indications of movement and immobility. Cell antenna location data is augmented by several other techniques that account for return signal response time, signal strength and angular deflection. When these data are triangulated with signals from several other cell antennas, location precision is improved and can run anywhere from a few dozen to hundreds of meters. (European Commission, 2011)

Cellular location logs constitute large, complex and growing data sets owned and exploited by cellular network operators in the course of ensuring seamless phone communications. Cellular-generated location data – especially when linked to consumer location and demographic profiles – represent a large potential source of revenue for operators, e.g. by selling analyses of their own data or by selling data for analysis by third-parties. This data may also be relevant for certain transport policy applications. For example, by matching triangulated cell data with map data relating to transport networks in order to estimate traffic flows and speeds. However, the differential precision across large-scale areas may be problematic for some applications.

Global Navigation Satellite System-based localisation

Almost all new phones and all smartphones integrate a Global Navigation Satellite System (GNSS) system microchip that allows precision location information to be generated from one of two (and soon three) dedicated satellite networks. The most common of these is GPS. In open areas with clean lines of sight to at least 4 satellites, GPS accuracy can be up to 5 metres. This accuracy degrades, however, in areas where GPS signals are disrupted by tall buildings or trees and inside of buildings. Assisted-GPS (A-GPS) increases location accuracy by combining GPS location signals with cellular location data providing sub-10 metre precision. Other forms of hybridised GPS location systems can provide similar levels of precision by using Wi-Fi network signals.

Indoor localisation and tracking of Wi-Fi-enabled devices inside a conference centre



Source: http://apps.opendatacity.de/relog/

Wi-Fi-based localisation

Wi-Fi-enabled outdoor and indoor location sensing can deliver even greater precision by tracking individual media access control addresses (MAC addresses – unique identifiers allocated to individual devices such as a laptop computers, mobile phones, tablets, Wi-Fi-enabled cars, etc…) within a network of Wi-Fi routers and transponders. Wi-Fi-enabled devices set to automatically connect to one or several networks regularly ping the available networks in order to join to known ones. This ping contains the MAC address unique to each device thus enabling device location and tracking. Sometimes this ping also includes data on previous Wi-Fi networks the device has connected to. With sufficiently dense Wi-Fi router networks, very precise location and movement

data can be inferred (as illustrated in the figure above).

Wi-Fi network configurations and node locations are also collected by numerous commercial operators delivering a suite of geolocation services[9]. These use "found" data from Wi-Fi sweeps, volunteered information from smartphone owners (based on automatic sensing and recording of all Wi-Fi networks detected along with their respective signal strengths) or Wi-Fi network managers. These inputs create very precise Wi-Fi "fingerprints" for individual spatial coordinates. These can then be used for locating new devices as they cross these coordinates. Some devices and apps regularly transmit data on Wi-Fi network "snapshots", which can be used to create context-aware services. This same data, if retained on a device or a central server, can also serve to track the device.

Other localisation technologies

Additional sensors in smartphones and other mobile platforms including accelerometers, gyroscopes and magnetometers enhance location and tracking even when cellular/Wi-Fi connections are insufficient and GPS signals are degraded. For instance, merged sensor data from an accelerometer, gyroscope and magnetometer can help determine location in reference to a last-known GPS-determined position by calculating heading and speed. This type of dead reckoning enhances tracking in areas with low, or no, GPS signals (as in tunnels). Dedicated networks of radio frequency identification signal (tracking the location of radio frequency identification – RFID – tags) or Bluetooth receivers can also provide precision location data, especially inside buildings or structures such as tunnels. The latter is the basis behind Apple's iBeacon location technology.

The ability to extract precise location information from "noisy" and unstructured analogue data feeds, such as those produced by cameras and microphones, has advanced tremendously thanks to sophisticated image and audio recognition algorithms. These, combined with coupled sensing-computing chips and in-stream signal processing techniques, allow machines to view, classify and attach significance to what they "see" and "hear".

Current image recognition-based technologies (and soon, voice recognition-based technologies) ensure sub-1 metre precision. Tracking individual objects and people based solely on face- or voice-recognition technology, especially across a number of sensors, is problematic but is improving and is being deployed in certain commercial or law-enforcement situations. Identifying, classifying and tracking semi-structured visual data from video or still feeds (e.g. automatic license-plate recognition, extraction of street-sign data from a series of geo-referenced photographs or video) is less challenging and has already been deployed at a large scale. Both HERE Maps and Google extract, classify and geo-code traffic signs as recorded by their mapping cars.

Machine logs recording specific and spatially localised interactions with stationary devices such as contactless turnstiles in public transport systems, electronic toll stations and even commercial payment card terminals provide a rich record of data points that can be built up to provide point-based trajectories in both time and space. These data allow for very precise positioning of card-holders or vehicle-based transponders but only at the point of interaction with the terminal. In closed systems such as freeways, travel times and speeds can be approximated as can passenger and vehicle flows at specific points. Additionally, fixed or mobile computers can be located with some level of precision by using their IP address though this data is not as precise as some of the other data sources discussed above

Finally, digital pictures, especially those taken on GPS-enabled devices, often include geographic coordinates of where the picture was taken in the file header. Picture archives and other image posting sites can be mined for location histories that are associated with the pictures an individual has taken. This history can be explicitly tied to an identifiable person when cross-referenced with volunteered information on, for instance, social media sites. In aggregate, this data can help track where persons tend to congregate by examining the co-location of pictures in space and time.

Precise geo-referenced location data represents a large and growing subset of Big Data as mobile devices and location-sensing technologies become ubiquitous (see Box 8). Collection and use of this data helps to provide citizens with cellular phone service, enhanced contextual data (proximity to restaurants, friends,

[9] Examples of geolocation service providers include Apple, Google, Skyhook, Streetlight Data, Euclid Analytics, Insoft, Navizon, Altergeo, Combain Services.

etc.), navigation, traffic, or other services (e.g. meteorological updates). Location-based services are a growing part of the connected economy and are expected to contribute up to USD 700 billion in consumer surplus by 2020. The majority of this surplus (USD 500 billion or 70%) would result from time and fuel savings due to GPS navigation and real-time traffic services. (Manyika, et al., 2011) Location data has fuelled growth in the connected economy but citizens are wary of divulging too much information about their whereabouts and their daily behaviour. One survey indicated that over 70% of those responding want to have specific knowledge about when and why applications collect location data (Balebako, et al., 2013) and another found that citizens were increasingly concerned about the scope of location data collection. This concern has heightened in light of secret, large-scale and trans-border data collection efforts by national security agencies:

> "The legal framework surrounding the use of location-based data and tracking technologies has evolved more slowly than the technologies themselves. Data protection rules have primarily been concerned with the protection of personal data – and only incidentally concerned with other dimensions of privacy including privacy of the person, privacy of personal communications and privacy of personal behaviour (though these areas are often addressed in criminal and civil law)." (Clarke & Wigan, 2011)

Location data are sourced from a number of platforms (see Box 8); mobile phone handset, tablet or computer, GNSS (such as GPS) receiver (e.g. in a car), Wi-Fi enabled device, video or localised machine tracking and logging devices (smartcard public transport turnstiles, tolling systems, etc.). The location accuracy derived from these methods varies in precision, reliability and timeliness. Some of the data can be imprecise but produced in real-time (with a lag of milliseconds to seconds), some can be very precise but only available ex-post via machine logs. Much of the location data produced by personal devices or embedded systems in vehicles, however, is both precise and delivered in nearly real-time. (Clarke & Wigan, 2011)

Advances in sensor platform architectures are likely to increase the amount of location-tagged data produced, just as the ability to process data captured in-stream becomes more prevalent and less costly. Semiconductor chip-maker Broadcom's announcement of a combined GNSS chip and sensor hub into a single system-on-a-chip signals a new powerful class of sensing platforms. (Broadcom, 2014) These platforms will enable on-the-fly fusion and pre-processing of data from Wi-Fi, Bluetooth, GPS, Micro Electro-Mechanical Systems (MEMS) auch as accelerometers, and other technologies as they become more widespread. (e.g. LTE networks) Due to significant claimed power savings and reduced size[10], these types of chipsets open up new possibilities for always-on location sensing and analysis for next-generation portable or wearable devices. Combined sensing-processing chipsets open up the possibility for cryptographically treating data in real-time or for applying privacy-enhancing pre-processing, but they do not in themselves ensure greater privacy protection and could plausibly even erode personal data privacy further still.

"You are where you've been"

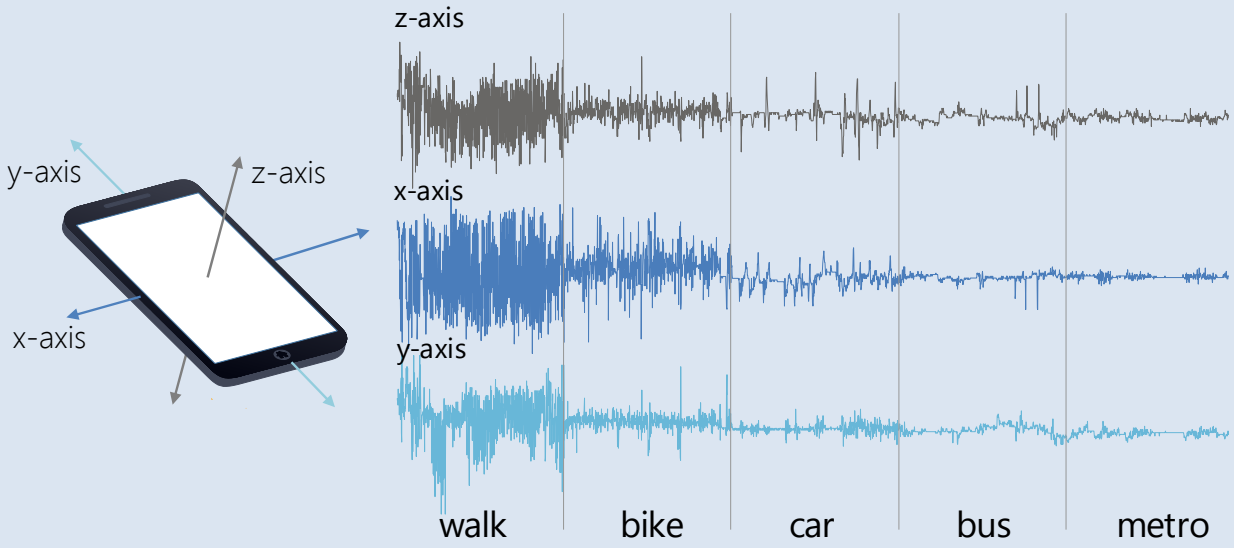
Rarely is the data *directly* linked to a unique individual – what is being tracked is a sensor-based platform. Objects such as these that exist in the real world (a phone, a SIM card, a car) are known as entities. Entities can have identities (Marie's car, Dinesh's phone, Jari's SIM card) that result from the linking of an entity and a specific identifier (Vehicle ID number, IMEA handset number, SIM card number). However, the geo-spatial data collected by these devices reveals a lot about individuals. Many of these platforms (especially mobile handsets and car-based navigation systems) are intimately linked to one person's activity patterns in time and space – not just to a specific identity number. Mobile handsets are almost always on or near to their owners and cars are rarely shared outside of the household.

[10] Broadcom claims an 80% reduction in power usage and a 34% reduction in chipset area.

Box 9. **Transport mode detection via Smartphone sensing**

The great amount of data produced by mobile phones has created new opportunities to infer movement and travel-mode related information for individuals. Extracting clustered speed profiles from cellular positioning data has been found to be relatively robust method for making a first determination of travel mode (Wang, et al., 2010).

Transport mode detection – Tri-axial accelerometer profiles for various travel modes



Adapted from (Feng & Timmermans, 2013)

More recently, research has focused on extracting movement-related information from smartphone and other portable device accelerometers. These accelerometers provide a flow of tri-axial readings measuring acceleration against baseline gravitational pull. Acceleration profiles are created by sampling the data and these profiles can then be analysed in order to isolate unique acceleration "signatures". Because of the close link between smartphones and their owners, these signatures can reveal many things about how individuals move.

In the field of health, acceleration profiles can be used to identify gait-related pathologies such as Parkinson's disease or to track daily activity (e.g. step counting or activity profiling). Gait can also serve as a biometric identification parameter as it is uniquely related to specific individuals.

In the field of transport, considerable research has been undertaken to algorithmically infer mode of travel from acceleration profiles[11]. The predictive accuracy of these methods has improved greatly with some research indicating (much) higher than 90% correct inference across multiple modes. This accuracy has been achieved by using training data sets and archived data but in the near future increasingly reliable, accurate and real-time travel mode inference will be possible thanks to improved algorithms and in-sensor processing.

By looking at patterns of relative inactivity and linking these to publicly available personal and business registries, location data exposes a daily pattern of activity that includes where a person sleeps, where they work and other places they frequent. These patterns of daily activity have been found to be extremely repetitive and predictable. This data can reveal a person's religion (repeated visits to a place of worship), their political affiliation (visits to political or NGO offices and co-location with demonstrations) and other information that can be inferred from where they go and where they spend time. This data can, in

[11] See for example (Feng & Timmermans, 2013), (Hemminki, et al., 2013), (Manzoni, et al., nd), (Shafique & Hato, 2014), (Reddy, 2010), (Susi, et al., 2013), (Stockx, 2014) and (Bernecker, et al., 2012).

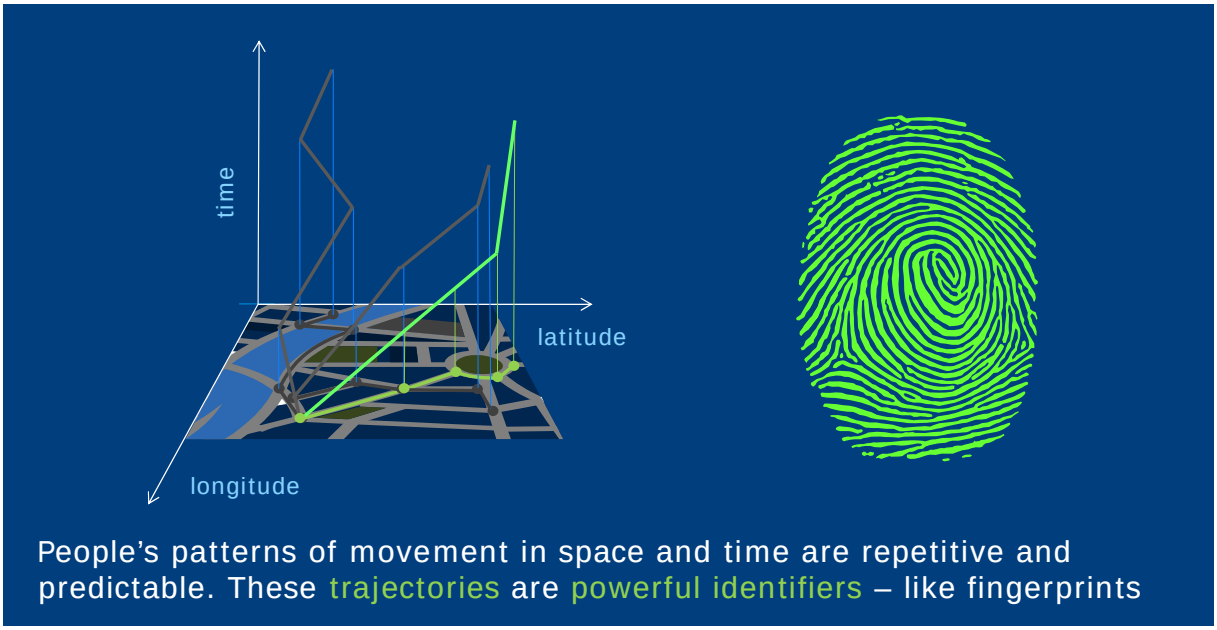
conjunction with similar data on other people, reveal the network of friends, acquaintances or colleagues a person has – especially when cross-referenced with volunteered data on social networking sites. This data can also reveal important and potentially significant pattern breaks that can compromise privacy (e.g. visits to an obstetrician's or an oncologist's practice). (Blumberg & Eckersley, 2009)

Fine-scale analysis of sensor outputs in smartphones can even reveal whether a person is walking, cycling, driving, on a bus or in a plane or other vessel (see Box 9).

"We know where you've been"

The ability to discover personal information grows as geospatial data is accumulated and as other data sources become available for cross-referencing and co-mingling. Though nominally "anonymous", the geo-spatial trail people leave behind is in fact highly personal and unique – it is (nearly) as identifiable as a fingerprint (Figure 9). Trajectory-based and time-stamped location data is a potent **quasi-identifier** for a single person or persons within a single household.

Figure 9. **Individuals' time-space trajectories are powerful identifiers**



The fact that fine-grained trajectory data can be linked to specific individuals seems understandable. However, even coarse-grained and imprecise trajectory data can be re-identified with relatively little effort.

Research on the privacy bounds of location data has resulted in a number of high-profile re-identification cases that have successfully isolated individual mobility traces from low granularity cellular base station data. A team of researchers at the Massachusetts Institute of Technology Media Lab analysed 15 months of mobile phone data for 1.5 million subscribers. They found that even for data with a temporal resolution of one hour and a spatial resolution equal to the cellular network's base tower cells[12], just four spatio-temporal points were sufficient to isolate and uniquely identify 95% of the individuals. Further coarsening of the temporal and spatial granularity only weakly and gradually lessoned the ability to isolate unique

[12] A few hundred metres in an urban setting, much larger distances elsewhere – see Box 8.

individuals. (de Montjoye, et al., 2013) Other research has confirmed the vulnerability of similarly sparse and coarse trajectory data to plausible and relatively straightforward de-identification and inference attacks resulting in re-identification rates ranging from 35% to 88%.[13]

"It's hard to hide where you've been"

The difficulty with which trajectory data can be adequately and persistently protected has led some to question whether it is worth the effort to do so. As seen in the previous section, voices are emerging in the United States and from within the private sector encouraging a move away from regulating data collection – including location and trajectory data – and towards setting robust rules regarding the use of that data. On the other end of the spectrum, "Privacy by Design" advocates have stated that the risk of re-identification has been largely overstated and that few real cases of linking individual mobility traces to unique names, addresses, or other personal identifying data have been undertaken (Cavoukian & Castro, 2014)[14]. There is truth in both arguments.

With time, the sophistication of de-identification algorithms is likely to grow as is the availability of other sources of information that could compromise the anonymity of location and trajectory data. De-identification techniques are a major research interest. Yet the results of this work are not confined to laboratories, but rather serve to generate commercially valuable data. Government agencies also use these techniques to track individuals. Nonetheless, de-identification attacks are not fully "trivial" – they require algorithmic sophistication, time to "clean" data errors, access to reliable data on personal identifiers (home address, ID number, etc.) and, most of all, sufficient motivation to overcome these hurdles.

At the same time, location-based and trajectory data *are* difficult to fully and permanently de-identify. Protecting the anonymity of high dimensional data like space-time trajectories or genetic information *is* more complicated than anonymising low-dimensional data such as addresses, names, blood-type, etc.

Anonymising location and trajectory data: four suggestions

The most robust data protection methods should be applied to location, trajectory and other high dimensional personal data

Data collectors and processors have at their disposal a multitude of de-identification techniques that range from simple anonymisation to cryptographic protection (see Box 10). In the case of high-dimensional location or trajectory-based data, there is a compelling argument to be made for using the most robust of these techniques and even seeking additional data protection methods. (Cavoukian & Castro, 2014)

Box 10. **Anonymisation and de-identification strategies**

De-identified personal data is personal data that has had individual identifiers removed or has been modified in such a way as to make re-identification reasonably unlikely. This approach preserves privacy by ensuring that data cannot be linked to a single individual or entity (such as a car). The best way to irreversibly de-identify data is to delete it but this would negate any possible benefits from post-collection analysis. For this reason, great effort has gone into finding robust ways to anonymise data while retaining value through sufficient granularity. In so doing, anonymised data must be robust to three risks:

- **Singling out:** individuals or unique objects should not be able to be isolated from the anonymised data set.
- **Linkability:** single or groups of data subjects should not be able to be linked via records in the same or separate data sets (some techniques protect against singling out, but not against linkability).

[13] See (Zhang, 2011), (De Mulder, et al., 2008) (Song, et al., 2014) and a full review of recent research on re-identification of trajectory data in (Gambs, et al., 2014).

[14] Of course doing so would pose ethical challenges that researchers may be unwilling to face or overcome.

- **Inference:** attributes of an individual or a group of data subjects should not be possible to deduce from the values of other attributes.

Data de-identification techniques, and in particular techniques for de-identifying location and trajectory data, is a rich and continuously evolving research topic. This evolution is necessary given the growing sophistication of re-identification and inference-based attacks on de-identified location and trajectory data. There are 3 broad approaches to preserving privacy through the anonymisation of location and trajectory data: randomisation, generalisation and pseudonymisation/encryption.

Randomisation-based approaches alter the nature of data in order to reduce its representation of reality. In doing so, these approaches seek to render data sufficiently uncertain so that they cannot be inferred to a single individual.

Randomisation techniques include "salting" data sets with spurious elements. "Salting" is not a robust standalone anonymisation technique. Should the noise added to the data set fall out of a credible range – or be semantically inconsistent with "real" values – then these elements can be identified and stripped from the data set negating the anonymisation effort. At the same time, sufficient noise should be added to adequately degrade the representativeness of the data set but this can reduce the value of the data set. Further, the data set still must have direct identifiers stripped from it prior to processing.

Permutation-based approaches, where some values are shuffled amongst data set records so as to create artificial records, preserve the distribution of values within a data set but degrade the traceability of individual records to unique individuals. However, permutation fails to anonymise data adequately if the some attributes are strongly correlated and can thus be inferred even if values are permuted. Permutation, like data set "salting", is not a sufficiently robust anonymisation technique.

The above techniques insert noise upstream to its use. "Differential privacy" describes a technique whereby noise is added to query responses made on an original data set. In this context, the anonymisation of the data is operated on the fly in response to third-party queries. Repeated and targeted queries, however, can isolate key elements of the data and lead to its re-identification. This approach requires monitoring and actively controlling data queries so as to counter this risk as well as ensuring that sufficient noise is added to query responses.

Generalisation

Generalisation-based approaches de-identify data by grouping attributes in increasing units of magnitude so as to erase their individual identities – e.g. grouping responses by census blocks rather than by individual addresses, or by month rather than by day. This approach is robust to singling out but requires sophisticated and targeted techniques to protect against linking and inference-based re-identification attacks.

Generalisation-based approaches include grouping records such that they are indistinguishable from a defined number of other records or by only identifying records based on interval values (06:00-09:00 hrs, 30-50 kilometres per hour, census blocks regrouping at least 200 households, etc.). However, with sufficiently strong quasi-identifiers that are not themselves generalised and insufficient clustering, generalised data sets can be re-identified. One way of increasing the robustness of generalisation techniques is ensuring that generalised classes of attributes contain a sufficiently broad range of values and distribution. To do so the distribution of the values in each generalised attribute class must mirror the range of attribute values within the entire data set (in order to prevent inference attacks based on differential attribute distribution). Nonetheless, the increasing sophistication of re-identification algorithms and the availability of data sets that can be used in inference attacks continue to erode the ability for even the best generalisation-based techniques to protect anonymity.

Privacy by pseudonymisation / encryption

Pseudonymisation is a security-enhancing rather than an anonymisation technique per se. It consists of replacing one attribute (typically a direct identifier) with another unique value. As such it reduces the linkability of a data set but does not impact the data set's vulnerability to singling out or inference attacks. Pseudonymisation can be independent of the original data as in the case of a random number assigned to the attribute or in the case of a username assigned by the data subject. Pseudo-identities or values can also be directly generated from the original data via the use of a hash function or an encryption key. In the case of single-key pseudonymisation, direct identifiers may be hidden but the data sets remain vulnerable to singling out, linking and inference-based attacks, especially for location or trajectory and other high-dimensional data.

Encryption-based approaches can effectively secure data from unauthorised access and use by, in effect, hiding the data rather than anonymising it. The level of protection offered by encryption is related to the type of encryption employed. Encryption locked by a single key is only as effective as the security of the key. Sophisticated decryption algorithms, vulnerability to brute force attacks and the ability to mobilise large and scalable computing resources to break encryption keys justify the use of extremely secure encryption keys. If the key is compromised, either by breaking or by obtaining it from human operators, then personal or location data is fully accessible.

Data sets encrypted by hashing involve producing a fixed-sized encrypted output from an input variable of differing sizes. Hashing is irreversible but not immune to re-identification attacks if the input values are of a fixed size and the full range of input values are otherwise known. In this case, the attacker simply has to run all known attribute values through the same hashing method to derive a table of corresponding hash-original values. This can then be used to single out, link or infer unique individuals and relationships between attributes. In a recent example of the shortcomings of insufficient hash encryption, the city of New York released an extensive data set of individual taxi trajectories comprised of 173 million geo-referenced records. The principal identifying features, the hack licence and medallion numbers were hashed using a well-known algorithm which encrypted them irreversibly. The rest of the data, including time stamps, location and trajectory data, were provided in open text. Because of the invariant format of both taxi license and medallion numbers, it was trivial to run all possible iterations of these numbers through the same encryption algorithm and find matches in the released hashed data. In less than two hours, all 173 million records had been de-anonymised and linked to full details available in other data sets recording taxi license holders. (Goodin, 2014)

The above example illustrates not that encryption itself is an insufficient technique to protect personal data but rather that poorly designed encryption methods are vulnerable. Adding noise to the attribute to be hashed ("salted" hash), combining the hash function with a secret key, or individually hashing each attribute and deleting the correspondence table would have more effectively protected the database from re-identification.

Though data collectors may argue that location and trajectory data are de-facto anonymous, this is clearly not the case since the geo-spatial component of the data is itself a powerful quasi-identifier. (Gambs, et al., 2014)[15] The next step of location data protection – simple anonymisation or pseudonymisation – is only possible when geospatial data are associated with straightforward data identifiers (such as name, address, etc.) that can be redacted or given pseudonyms. Neither of these approaches should be considered an adequate basis for de-identification especially in the context of associated geospatial coordinates. (Cavoukian & Castro, 2014)

Efforts to anonymise location and trajectory data by clustering points or traces into larger groupings of like data can improve anonymity. Numerous researchers have studied such generalisation-based approaches. However, there is a risk of re-identification even from aggregate data if the aggregation ignores particular characteristics of the data set. Unsophisticated clustering techniques – such as aggregating data by obfuscating data or by adding spurious data – can help but sufficiently large sets of location-based or trajectory data can nullify the impact of such data "noise".

Effective protection of location, trajectory and other high dimensional personal data should combine both anonymisation and encryption

Neither anonymity nor pseudonymity, nor clustering nor obfuscation prevents location-based data from being transmitted, interpreted and exploited. The data is still composed of recognisable "plain text" latitudinal, longitudinal and time-referenced character strings albeit at different levels of coarseness. Cryptographic methods, on the other hand, remove the ability to interpret the geospatial data by transforming it cryptographically. Only those with the appropriate key can convert the cipher back into plain text and then exploit the geospatial elements of the record.

[15] It is for this reason that the proposed EU General Data Protection Directive explicitly defines location data as "personal data". As such, location data would fall under the scope of that Directive (as transcribed into national legislation) and be subject to strong requirements regarding notice and consent before collection.

Cryptographic protection of location and trajectory data seems a promising approach and one that has attracted considerable attention. On-the-fly encryption and de-encryption is facilitated by the emergence of "sensors-on-a-chip" that can sense, compute, encrypt and transmit data in real time. Such encryption allows trusted communication between devices that handle mission-critical tasks (e.g. communicating location, heading, speed, system state between vehicles, or between infrastructure and vehicles, in a complex traffic environment). It can also serve to protect and enforce rights associated with personal data, including location-based data. Encryption-based approaches, however, require sufficiently robust and secure system architecture to manage encryption keys, trusted identities and certificates. Public key infrastructure (see Box 11) ultimately requires buy-in from a broad cross-section of actors and a trusted, likely public body, to manage identity certification in an international context.

Box 11. **Encryption and public key infrastructure**

Public key protocols could contribute significantly to the protection and trustworthiness of essential data whether in the case of encrypted communications between connected devices (e.g. cars, infrastructure, etc.) or in the case of personal data protection and management systems.

In information security, certification refers to the issuing of certificates used for security verification of messages between systems. Those certificates in fact contain a public key, i.e. a key that can be used to verify the electronic signature that is appended to a message.

The different certificates and certifying entities may be addressed in different ways. For example, pseudonym certification authorities are also referred to as authorisation authorities, while pseudonym certificates are referred to as authorisation certificates, authorisation tickets or short-term certificates. Long-term certification authorities also go by the names enrolment authority or enrolment credentials.

Communication and data exchange requires trust and robust anonymity

Cooperative ITS systems (C-ITS) depend on managed and trusted access in order to allow the communication of essential data without compromising the security or privacy of system users. These systems must ensure that, firstly, the messages that are exchanged are authentic messages, i.e. they originate from the source they claim to have come from; and secondly, that the anonymity of the users is assured. This can be done by verifying the message's signature.

To give an example: A car senses a slippery stretch of road and slows down. It warns the vehicles around it that it is slowing down and the road is slippery. The infrastructure also receives this message. The road operator spreads the warning further and looks into causes and remedies to the situation. How do the road operator and the other cars know that this message has not been sent from, say, a non-C-ITS device on the roadside or the source of the message has been hacked? Furthermore, the individual motorist has to be sure that no movement can be tracked by unauthorised actors. In another example, an individual may authorise a data collector to collect and use data for a specific application. How can this authorisation be authenticated and traced to the specific permission granted by the data subject?

The recipient of a message has to trust the source of the message in order to be able to verify the message security. In ITS communications, trust is the confidence that a particular public key belongs to the entity claimed, which ensures that the corresponding signature has been really provided by the entity. This is done by a dedicated arrangement called Public Key Infrastructure (PKI), composed of certification authorities that confirm the ownership of a public key by an entity i.e. by issuing a public key certificate, which is an electronic document that binds a public key with an identity, as stated in the certificate. PKI architecture is a system design safeguard against abuse and a mean to protect the privacy of the user.

The masquerade

The analogy of a masked ball can help to illustrate the trust relationships embedded in public key architecture. The masquerade has certain rules: 1) Guests must stay anonymous at all times; 2) an invitation to the ball entitles guests to collect masks at the wardrobe; 3) guests can only trust other masked people and anybody can listen to the conversations among masked guests; 4) only friends of the host are invited; 5) masks are differentiated but they are all identifiable as having been distributed by the host (e.g. they display different features but they are all red - the identifier for the host's masks). The invitation itself is anonymous –guests'

names are not revealed.

Guests enter the ball and produce their invitation at the wardrobe where they receive a mask. All masked guests can be trusted because they are all friends of the host. Still guests do not reveal their identity to others and may return often to the wardrobe for a new mask. The host of the ball is not the owner of the house where the ball takes place – but the owner of the house has given the host permission to hold the ball and distribute masks in the wardrobe – as long as these are only distributed to trusted friends of the host.

Like the guests at the masquerade C-ITS equipment wears a mask when it communicates. The receiving C-ITS hardware: 1) looks at the mask and 2) believes the message, because it recognises the mask the sender wears (e.g. it is red). It does not identify the sender's identity. The wardrobe where guests collect their masks in this metaphor is the pseudonym certification authority. The host who has issued invitations to the masquerade takes the role of what in C-ITS would be the long-term certification authority. The owners of the house where the ball is held take the role of the root authority that controls the complete system. In C-ITS terms this system would be called a public key infrastructure or PKI.

How can trust be established between the different players in C-ITS? In particular, how, from a technical perspective, can trust be communicated and messages technically masked in a secure manner? Furthermore, how can trust be established institutionally, meaning how should a body trusted be structured?

Certificates mean trust

The masquerade is based on trust. Trust in this analogy is communicated via:

- The invitation; guests received this from the host. The invitation also provides assurance that the owner of the house has authorised the host to hold the masquerade.
- The mask; whoever wears a mask is a friend of the host and can be trusted by other guests. The invitation and the mask represent so-called certificates. The invitation represents the long-term certificate, the mask that is frequently changed during the party represents the so-called pseudonym certificate.
- The glance between guests; those at the party, when they see another masked guest, can be confident that that person is a friend of the host and hence trustworthy.

How does this work technically? When sending a message the C-ITS equipment sends three items: 1) the message (in the analogy: the content of guests' conversation); 2) the certificate itself that will allow any C-ITS recipient to verify the signature (in the analogy: the mask); 3) the signature it generated from the currently valid certificate and that particular message (in the analogy: the glance of the mask and hearing the voice of the masked person). The signature and the certificate are pieces of data generated using mathematical algorithms.

The mask and wardrobe: pseudonym certificates and pseudonym authorities

Pseudonym certificates serve the purpose of anonymising the movement of a piece of C-ITS equipment. They are the mask in the analogy. The C-ITS equipment changes them on a frequent basis. The movement of the C-ITS equipment cannot be tracked by the recipient or a series of recipients, since the C-ITS equipment changes its certificate on a regular basis.

The C-ITS sender receives its pseudonym certificates from the pseudonym certification authority – at the masquerade this would be the wardrobe. The pseudonym certification authority has the task to make sure the C-ITS equipment requesting the certificates is authorised to send C-ITS messages, meaning it is not stolen, hacked or damaged. It needs to check the authenticity of the C-ITS equipment, but it should not know the owner. Only if the pseudonym certificate is not directly linked to the owner of the C-ITS equipment can the C-ITS equipment transmit its messages anonymously. At the wardrobe, this is done by checking the invitations of the guests, rather than by knowing the guests themselves. Guests are trusted if they have an invitation and are thus issued a mask necessary to communicate in a trusted fashion with other guests.

The invitation and the host: long-term certificates and long-term authorities

The long-term certificate has the objective of shielding the identity of the owner of the C-ITS equipment. It stays with the equipment for longer periods of time, just as the invitation is valid for the whole party (whereas you change masks several times during the party). The long-term certificate of the C-ITS equipment is used to request pseudonym certificates from the pseudonym certification authority. The pseudonym certification

authority can use it to check the authenticity of the equipment it deals with. The pseudonym certification authority only knows the long-term certificate. It will not know the owner of the C-ITS equipment or the vehicle it is installed in or, in the case of roadside equipment, the location where it is installed. The long-term certificates are issued by the so-called long-term certification authority. In our metaphor this role is taken by the host. The long-term certification authority is independent from the pseudonym certification authority just like the wardrobe is independent from the host. The long-term certification authority (the host) knows the owner of the C-ITS equipment.

Permission to hold the ball by the owner of the house: authorisation and root certification authority

The C-ITS equipment needs to know if the pseudonym certification authority is itself authorised to issue pseudonym certificates. It does this by checking the status of the pseudonym certification authority's own certificate. In our metaphor this would be trust in the owners of the house. They allow the ball to take place in the first place and condone the distribution of masks in the wardrobe. The root certification authority will issue each pseudonym certification authority a pseudonym certification authority certificate that is used to produce the pseudonym certificates. It does this only if the pseudonym certification authority complies with all its obligations. The root certification authority supervises the various pseudonym certification authorities.

The same counts for the long-term certification authorities. If the long-term certification authority is compliant with its obligations, the root certification authority will issue each long-term certification authority a long-term certification authority certificate that the long-term certification authority uses to generate long-term certificates for the C-ITS equipment. In the masquerade analogy, guests assume that if the host has issued an invitation, she will have done so with the approval of the house owner.

Every PKI can only have one root certification authority. It publishes the requirements for pseudonym certification and long-term certification authorities and checks applicants' suitability to play these roles. It also controls if existing authorities adhere to the rules.

The house owner: policy authority

In our metaphor it is the house owner who decides that the host may invite friends over and to allow the distribution of masks from the wardrobe. These decisions are in fact policy decisions. Here the metaphor does not match exactly, since in a PKI architecture the policy authority is separate from the root certification authority. The policy authority does not issue any certificates. It defines C-ITS PKI policy, the rules to which the root certification authority, the pseudonym certification authority and the long-term certification authority adhere to. It also supervises the root certification authority.

The limits of existing data protection strategies are being reached. The 21st century will require a "New Deal on Data" to fully protect consumers and unleash innovation

The critique of existing data protection frameworks and the divergent set of proposed remedies outlined in this section 3 highlight the need for an updated approach to personal data *ownership* as opposed to protection. Projected changes in data collection technology, coupled with increasing unease over ubiquitous data collection signal the limitation of incremental approaches. In this context, some have called for a fundamental reformulation of data protection efforts – a "New Deal on Data" that re-charts the relationship, respective roles and responsibilities and the nature of interactions between data subjects and producers and data collectors and users. (Pentland, 2009)

The approach outlined by proponents of the "New Deal on Data" emphasises data subjects' ownership of their own data as opposed to ownership granted to data collection entities. In this sense, the "New Deal" approach proposes granting individuals the same rights over the disposition and use of their data as they have over their bodies and their money.

New data collection system architectures would be needed to operationalise this approach to data ownership. Central to these would be the notion of a personal data locker or store. A personal data store centralises all data associated with a single individual. This individual exercises full and granular control over access to that data according to their express preferences. This approach changes the current paradigm of collector-owned data and creates a market for access and use of personal data. Access can be

given to third parties that promise sufficient value but, since ownership and control of the data remains with the data subject, this access can be revoked at any time. This would no-doubt require the development of new business models for monetising use of this data, but would ensure a robust and conscious protection of individual privacy preferences[16].

Proponents of the "New Deal on Data" have developed an open source framework for such a personal data store (openPDS) and are trialling it. Results are promising and include the finding that greater control over data ownership leads to greater data sharing (seemingly in response to heightened trust that data users will conform to specific individual wishes in a transparent and auditable manner). (HBR, 2014)

Novel data protection mechanisms can develop around the concept of personal data stores. One such mechanism – SafeAnswers – promises robust protection of high dimensional personal data, while at the same time allowing open access to the data itself. (de Montjoye, et al., 2013) This mechanism is built around data users submitting code snippets that mediate on individuals' raw data in their personal data store without releasing any of that data itself. Under a personal data store framework, the SafeAnswers approach calls for potential data users to submit a request for information regarding an individual's data. The question could be "is the individual close to my store?", "how much time does the individual spend in traffic on a weekday?" or "does the individual use the underground on weekends? If the individual accepts that request (perhaps granting this acceptance in return for a service or other form of compensation from the data user), the data user submits a standardised snippet of code that then interacts with the user's personal data store, querying GPS log data, accelerometer data, or other form of location/trajectory data required to answer the question. The answer is sent back to the data user without sensitive location or trajectory data ever having been divulged. This approach outlines how novel data ownership rules may be required before conflicting demands for data protection and increased innovation can be reconciled.

New models of public-private partnership involving data-sharing may be necessary to leverage both public and private benefits

Under existing data ownership rules a significant amount of the actionable data pertaining to road safety, traffic management and travel behaviour is held by the private sector. Should data ownership rules change along the lines of "A New Deal on Data", individuals would retain ownership and control use of this data. Under both data ownership frameworks, public authorities will likely continue to be mandated to provide essential services. In this context, much as public authorities have coercive ability to require access to personal data (e.g. on property ownership, personal revenue, criminal records), there may be scope to define public-access data sets that are built on aggregated personal location and trajectory data. These data could relate to traffic flows, crash locations and causes (as reported by embarked-vehicle ITS systems), as well as crowd density, location and movement data. This would entail a move away from the supplier-client relationship that some authorities have with data collectors. Recent moves by some data collectors to share their data with public authorities (e.g. Uber in the Boston area) show that more creative partnerships can be developed that enable both the private sector to innovate and the public sector to carry out its mandates. However, work will be required to define the scope and scale of data access by public authorities and, in particular, to ensure that the collection of such data is in line with public mandates.

[16] For a full discussion of the operational aspects of the "New Deal on Data", see (de Montjoye, et al., 2014) and (HBR, 2014).

Bibliography

Anwar, A., Nagel, T. & Ratti, C., 2014. *Traffic Origins: A Simple Visualization Technique to Support Traffic Incident Analysis..* s.l., IEEE Pacific Visualization Symposium.

APEC, 2004. *APEC Privacy Framework.* s.l.:s.n.

Asslett, M., 2013. *Spotlight: Big data reconsidered: it's the economics, stupid.* [Online]
Available at: https://451research.com/report-short?entityId=79479&referrer=marketing
[Accessed 7 August 2014].

Balebako, R., Shay, R. & Cranor, L. F., 2013. *Is Your Inseam a Biometric? Evaluating the Understandability of Mobile Privacy Notice Categories,* Pittsburgh: CyLab: Carnegie Mellon University.

Banks, C., 2011. *Top 10: The Quotable Eric Schmidt.* [Online]
Available at: http://blogs.wsj.com/digits/2011/01/21/top-10-the-quotable-eric-schmidt/
[Accessed 16 September 2014].

Becker, R. et al., 2011. A Tale of One City: Using Cellular Network Data for Urban Planning. *Pervasive Computing, IEEE,* 10(4), pp. 18 - 26.

Benko, H., Morris, M., Brush, A. & Wilson, A., 2009. *Insights on Interactive Tabletops: A Survey of Researchers and Developers. ,* s.l.: Microsoft Research Technical Report MSR-TR-2009-22.

Bernecker, T. et al., 2012. *Activity recognition on 3D accelerometer data.* München: Institute for Informatics, Ludwig-Maximilians-Universität.

Bhatti, J. & Humphreys, T. E., 2014. *Covert Control of Surface Vessels via Counterfeit Civil GPS Signals.* s.l.:The University of Texas at Austin Radionavigation Laboratory.

Blumberg, A. J. & Eckersley, P., 2009. *On Locational Privacy and How to Avoid Losing it Forever.* s.l.:Electronic Frontier Foundation.

Broadcom, 2014. *Broadcom Announces Industry's First Global Navigation and Sensor Hub Combo Chip.* Irvine, CA: s.n.

Calabrese, F., Colonna, M., Lovisolo, P. & Ratti, C., 2011. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems,* 12(1), pp. 141 - 151.

Cate, F. H., 2006. The Failure of Fair Information Practice Principles. In: *Consumer Protection in the Age of the Information Economy.* s.l.:Social Science Research Network (SSRN).

Cavoukian, A., 2010. *Privacy by design: the definitive workshop. A foreword by Ann Cavoukian, Ph.D.* Springer: s.n.

Cavoukian, A., 2012. *Operationalizing Privacy by Design: A Guide to Implementing Strong Privacy Practices.* Toronto(Ontario): Office of the Information and Privacy Comissioner for the Province of Ontario.

Cavoukian, A. & Castro, D., 2014. *Big Data and Innovation, Setting the Record Straight: De-identification Does Work.* Toronto: Office of the Information and Privacy Commissioner of the Province of Ontario.

Cavoukian, A. & Jones, J., 2012. *Privacy by Design in the Age of Big Data.* Toronto(Ontario): Office of the Information and Privacy Commissioner for the Province of Ontario.

Cerrudo, C., 2014. *Hacking US Traffic Control Systems.* s.l., DEFCON Hacking Conference.

Cheung, A. S., 2014. Location privacy: the challenges of mobile service devices. *Computer Law and Security Review,* 02.30(1).

Clarke, R. & Wigan, M., 2011. You are where you've been: the privacvy implications of location and tracking technologies. *Journal of Location-Based Services,* 5(3-4).

Community Research Association, 2012. *Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States.* s.l.:Community Computing Consortium.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D., 2013. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific Reports: Nature.com*, 25 March, Volume 3.

de Montjoye, Y.-A., Shmueli, E., Wang, S. & Pentland, A. S., 2014. openPDS: Protecting the Privacy of Metadata through SafeAnswers. *PLOS,* 9(7).

De Mulder, Y., Danezis, G., Batina, L. & Preneel, B., 2008. *Identification via location-profiling in GSM networks.* Alexandria (VA), Association for Computing Machinery.

de Smith, M. J., Goodchild, M. F. & Longley, P. A., 2013. *Geospatial Analysis. A comprehensive Guide to Principles, Techniques and Software Tools.,* s.l.: Spatial Analysis Online.

Dyson, G., 2013. *No Time Is There - The Digital Universe and Why Things Appear To Be Speeding Up.* [Online]
Available at: http://longnow.org/seminars/02013/mar/19/no-time-there-digital-universe-and-why-things-appear-be-speeding/
[Accessed 7 August 2014].

European Commission, Article 29 Data Protection Working Party, 2014. *Opinion 05/2014 on Anonymisation Techniques.* Brussels: European Commission.

European Commission, 2011. *Opinion 13/2011 on Geolocation services on smart mobile devices ,* s.l.: European Commission.

European Commission, 2012. *Commission proposes a comprehensive reform of the data protection rules.* [Online]
Available at: http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm
[Accessed 16 September 2014].

EVITA, 2012. *EVITA Project Summary Report.* [Online]
Available at: http://www.evita-project.org/Publications/EVITAD0.pdf
[Accessed 10 September 2014].

Executive Office of the President, 2014. *Big Data: Seizing Opportunities, Preserving Values.* Washington, DC: Executive Office of the President.

Feng, T. & Timmermans, H. J., 2013. Transportation mode recognition using GPS and accelerometer. *Transportation Research,* Volume 37, p. Part C.

Forbes, 2014a. *Four Reasons Google Bought Waze.* [Online]
Available at: http://www.forbes.com/sites/petercohan/2013/06/11/four-reasons-for-google-to-buy-waze/
[Accessed 14 October 2014].

Forbes, 2014b. *Why Google's Waze Is Trading User Data With Local Governments.* [Online]
Available at: http://www.forbes.com/sites/parmyolson/2014/07/07/why-google-waze-helps-local-governments-track-its-users/
[Accessed 15 October 2014].

Fu, H. et al., 2014. Fu, H. et al., 2014. A Field Study of Run-Time Location Access Disclosures on Android Smartphones. *Internet Society*.

Gambs, S., Killijian, M.-O. & del Prado Cortez, M. N., 2014. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences,* December.80(8).

Goodin, D., 2014. *Poorly anonymized logs reveal NYC cab drivers' detailed whereabouts.* [Online]
Available at: http://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/
[Accessed 23 July 2014].

Greenleaf, G., 2012. The Influence of European Data Privacy Standards Outside Europe: Implications for Globalization of Convention 108. *International Data Privacy Law,* April, 2(2), pp. 68-92.

Harford, T., 2014. Big data: Are we making a big mistake?, 28 March 2014. *FT Magazine*, 28 March.

Hawelka, B. et al., 2014. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science,* 41(3).

HBR, 2014. With Big Data comes Big Responsibility (Interview with Alex "Sandy" Pentland, Toshiba Professor of Media Arts and Sciences at MIT). *Harvard Business Review,* November.Issue November 2014.

Hemminki, S., Nurmi, P. & Tarkoma, S., 2013. *Accelerometer-based transportation mode detection on smartphones.* Rome, Association for Computing Machinery.

Holleczek, T. et al., 2014. *Detecting Weak Public Transport Connections from Cell Phone and Public Transport Records.* s.l., The Third ASE International Conference on Big Data Science and Computing.

Hutchins, J. I. A. S. P., 2010. *Probabilistic Analysis of a Large-Scale Urban Traffic Sensor Data Set. Knowledge Discovery from Sensor Data.* Berlin Heidelberg: Springer.

IDC, I. D. C. -., 2014. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things.* [Online]
Available at: http://www.emc.com/leadership/digital-universe/2014iview/index.htm
[Accessed 7 August 2014].

IEA OPEN, nd. *Energy Technology Bulletin.* [Online]
Available at: http://www.iea.org/impagr/cip/archived_bulletins/issue_no23.htm
[Accessed 14 October 2014].

Isenberg, P., Hinrichs, U., Hancock, M. & Carpendale, S., 2010. Digital tables for collaborative information exploration. In: *Tabletops-Horizontal Interactive Displays.* s.l.:Springer, pp. 387-405.

ITU, 2014. *World Telecommunication/ICT Indicators database,* s.l.: ITU.

Janeiro, C. d. O. d. P. d. R. d., 2014. *Centro de Operações da Prefeitura do Rio de Janeiro.* [Online]
Available at: http://www.centrodeoperacoes.rio.gov.br/institucional.
[Accessed 14 October 2014].

Kerns, A. J., Shepard, D. P., Bhatti, J. A. & Humphreys, T. E., 2014. Unmanned Aircraft Capture and Control via GPS Spoofing. *Journal of Field Robotics,* 31(4).

Khalegi, B., Khamis, A., Karray, F. & Razavi, S., 2013. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion,* Volume 14, pp. 28-44.

Krzywinski, M. et al., 2009. Circos: An Information Aesthetic for Comparative Genomics. 19(9).

Li, Y. & Li, Z., 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies,* Volume 34, pp. 108-120.

Lohr, S., 2014. *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights.* s.l.:New York Times.

Manyika, J. et al., 2011. *Big data: The next frontier for innovation, competition and productivity,* s.l.: McKinsey Global Institute.

Manzoni, V., Maniloff, D., Kloeckl, K. & Ratti, C., nd. Transportation mode identification and real-time CO2 emission estimation using smartphones. *SENSEable City Lab, Massachusetts Institute of Technology.*

McKinsey Global Institute, 2011. *Big data: The next frontier for innovation, competition, and productivity,* s.l.: McKinsey & Company.

Miller, C. & Valasek, C., 2014. *A Survey of Remote Automotive Attack Surfaces,* s.l.: Illmatics.

Morstatter, F., Pfeffer, J., Liu, H. & Carley, K., 2013. *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose,* s.l.: ICWSM .

Nabian, N., Offenhuber, D., Vanky, A. & Ratti, C., 2012. Data Dimension: Accessing Urban Data and Making it Accessible. *Urban Design and Planning,* 166(1).

Natural Resources Canada, 2010. *Geospatial Privacy and Risk Management Guide for Federal Agencies.* Ottawa: Canada Privacy Services Incorporated.

OECD, 2013. *Exploring Data-Driven Innovation as a New Source of Growth; Mapping the Policy Issues Raised by Big Data,* Paris: OECD Publishing.

OECD, 2013. *Privacy Expert Group Report on the Review of the 1980 OECD Privacy Guidelines,* s.l.: OECD Publishing.

OECD, 2013. *The OECD Privacy Framework.* Paris: OECD Publishing.

OII, 2014. *Data Protection Principles for the 21st Century.* Oxford: Oxford University.

OPCC, 2014. *Results of the 2014 Global Privacy Enforcement Network Sweep.* [Online]
Available at: https://www.priv.gc.ca/media/nr-c/2014/bg_140910_e.asp
[Accessed 23 September 2014].

Open Data Research Network, 2014. *Transparency and Open Government Data in Rio de Janeiro: The Collateral Effect of the Smart City.* [Online]
Available at: http://www.opendataresearch.org/content/2014/576/transparency-and-open-government-data-rio-de-janeiro-collateral-effect-smart-city
[Accessed 14 October 2014].

PCAST, 2014. *Big Data and Privacy: A Technological Perspective.* s.l.:President's Council of Advisors on Science and Technology: Executive Office of the President.

Pentland, A., 2009. *Reality Mining of Mobile Communications: Towards a New Deal on Data.* s.l.:World Economic Forum.

Peterson, S., 2011. *Downed US drone: How Iran caught the 'beast' [Online],. Available at:.* [Online]
Available at: http://www.csmonitor.com/World/Middle-East/2011/1209/Downed-US-drone-How-Iran-caught-the-beast
[Accessed 21 September 2014].

Price, L., de la Rue du Can, S., Sinton, J. & Worrell, E., 2008. Sectoral trends in global energy use and greenhouse gas emissions. *Energy Policy,* 36(4).

Ratti, C. & Nabian, N., 2010. Virtual Space, The City to Come. In: *Innovation Perspectives for the 21st Century.* s.l.:BBVA, pp. 383-397.

Reddy, S. e., 2010. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN),* 6(2).

Rich, C., 2014. *Privacy Laws in Asia.* s.l.:Bureau of National Affairs, Inc.

Santi, P., Resta, G. & Ratti, C., 2014. Quantifying the Benefits of Taxi Trips in New York through Shareability Networks. *ERCIM News 98.*

Schwartz, P. M., 2013. The EU-US Privacy Collision: A Turn to Institutions and Procedures. *Harvard Law Review,* May.126(7).

Shafique, M. A. & Hato, E., 2014. Use of acceleration data for transport mode prediction. *Transportation,* 2014(8).

Song, Y., Dahlmeier, D. & Bressan, S., 2014. *Not so unique in the crowd: a simple and effective algorithm for anonymising location data.* Gold Coast, SIGIR.

Spreadtrum, 2013. *Spreadtrum and Mozilla Take Aim at Global Smartphone Accessibility with Turnkey Solution for US$25 Smartphones.* [Online]
Available at: http://www.spreadtrum.com/en/news/press-releases/spreadtrum-and-mozilla-take-aim-at-global-smartphone-accessibility
[Accessed 14 October 2014].

Sprenger, P., 1999. *Sun on Privacy: 'Get Over It'.* [Online]
Available at: http://archive.wired.com/politics/law/news/1999/01/17538
[Accessed 14 September 2014].

Stockx, T., 2014. *Going deeper underground: Using accelerometers on mobile devices to enable positioning on underground public transportation systems.* s.l.: Universiteit Hasselt.

Susi, M., Renaudin, V. & Lachapelle, G., 2013. Motion mode detection and step detection algorithms for mobile phone users. *Sensors,* Volume 13.

Tech America Foundation, 2012. *Demystifying Big Data: A practical guide to transforming the business of government,* Washington, DC: Federal Big Data Commission: TechAmerica Foundation.

Tucker, P., 2014. *The Naked Future: What Happens in a World That Anticipates Your Every Move?.* New York: Penguin Group (USA).

UN Populations Fund, 2007. *State of the World Population; Unleashing the potential of urban growth,* New York: United Nations.

United States Environmental Protection Agency, 2014. *U.S.-Brazil Joint Initiative on Urban Sustainability.* [Online]
Available at: http://www.epa.gov/jius/projects/rio_de_janeiro/rio_operations_center.html
[Accessed 2014 September 2014].

Wang, H., Calabrese, F., Di Lorenzo, G. & Ratti, C., 2010. *Transportation mode inference from anonymized and aggregated mobile phone call detail records.,* Funchal: IEEE.

WEF, 2013. *Unlocking the Value of Personal Data: From Collection to Usage.* Geneva: World Economic Forum.

Westin, A., 1967. *Privacy and Freedom.* New York: Anthenum.

World Bank and Conveyal, 2014. *Transport Analyst - Buenos Aires, Job Access.* [Online]
Available at: http://wb-ba-analyst.dev.conveyal.com/
[Accessed 13 February 2015].

Yasin, R., 2011. *IBM puts its 'smart city' technology in one package.* [Online]
Available at: http://gcn.com/404.aspx?404=http://gcn.com/articles/2011/06/07/ibm-intelligent-operations-center-for-%20smarter-cities.aspx
[Accessed 10 September 2014].

Yuan, J., Zheng, Y., X. & Sun, G., 2011. *Driving with knowledge from the physical world..* s.l., 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, ACM.

Zhang, H. &. B. J. 2., 2011. *Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study.* Las Vegas, MobiCom.

Zickhur, K., 2012. *Three-quarters of smartphone owners use location-based services.* [Online]
Available at: http://www.pewinternet.org/2012/05/11/three-quarters-of-smartphone-owners-use-location-based-services/
[Accessed 28 September 2014].

International Transport Forum

Big Data and Transport

Understanding and assessing options

This report examines issues relating to the arrival of massive, often real-time, data sets whose exploitation and amalgamation can lead to new policy-relevant insights and operational improvements for transport services and activity. It is comprised of three parts. The first section gives an overview of the issues examined. The second broadly characterises Big Data, and describes its production, sourcing and key elements in Big Data analysis. The third section describes regulatory frameworks that govern data collection and use, and focuses on issues related to data privacy for location data.

The work for this report was carried out in the context of a project initiated and funded by the International Transport Forum's Corporate Partnership Board (CPB). CPB projects are designed to enrich policy discussion with a business perspective. Led by the ITF, work is carried out in a collaborative fashion in working groups consisting of CPB member companies, external experts and ITF researchers.

International Transport Forum
2 rue André Pascal
75775 Paris Cedex 16
France
T +33 (0)1 45 24 97 10
F +33 (0)1 45 24 13 22
Email : itf.contact@oecd.org
Web: www.internationaltransportforum.org

2015-05 /Photo credit: wavebreakmedia, shutterstock.com