International Transport Forum

CPB Corporate Partnership Board

Data-Driven Transport Policy

Corporate Partnership Board Report

OECD

Data-Driven Transport Policy

Corporate Partnership Board Report

About the International Transport Forum

The International Transport Forum at the OECD is an intergovernmental organisation with 57 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. It is administratively integrated with the OECD, yet politically autonomous.

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. ITF's Annual Summit is the world's largest gathering of transport ministers and the leading global platform for dialogue on transport policy.

Our member countries are: Albania, Argentina, 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, Israel, Italy, Japan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Montenegro, Morocco, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom and United States.

Foreword

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. They are launched in areas where CPB member companies identify an emerging issue in transport policy or an innovation challenge to the transport system. Led by the ITF, work is carried out in a collaborative fashion in working groups consisting of CPB member companies, external experts and ITF staff.

Many thanks to the members of the Corporate Partnership Board companies involved in this work: Ford, Google, HERE, INRIX, Kapsch, Michelin, PTV, Uber. The principal authors of this report were Philippe Crist, Tom Voege and Diego Canales. The report draws also from contributions and discussions during an expert's workshop, organised 09 and 10 November 2015 in Paris. Many thanks also to Ande Monier of the International Transport Forum for her assistance in the organisation of this workshop.

Participants in this workshop included:

Michael Batty, University College London, UK
Carla Bonina, University of Surrey, UK
Andrew Byrd, Conveyal, US
Diego Canales, World Resources Institute
Ann Cavoukian, Ryerson University, Privacy and Big Data Institute, Canada
Dorothy Chou, Uber
Olivier Esper, Google
Frank Felten, Planung Transport Verkehr AG (PTV AG)
Yves-Alexandre de Montjoye, Massachusetts Institute of Technology, Media Lab, US
Rosina Howe-Teo, Land Transport Authority, Singapore
Patricia Hu, Transport Statistics, Department of Transport, US
Paulo Humanes, Planung Transport Verkehr AG (PTV AG), Germany
Demosthenes Ikonomou, European Union Agency for Network and Information Security (ENISA)
Gilbert Konzett, Kapsch TrafficCom AG, Austria
Sabina Lindström, Unit for Transport Data, Ministry of Transport and Communications, Finland
Peter Miller, ITO World, UK
Scott Nelson, HERE,
Jonah Ong, Land Transport Authority, Singapore
Alex Pentland, MIT, Connection Science and Human Dynamics Labs, US
John Polak, Urban Systems Laboratory, Imperial College London, UK
Laura Schewel, StreetLight Data, Inc., US
Scott Sedlik, INRIX
Andrew Stott, Department of Transport Transparency Board
Jasja Tijink, Kapsch TrafficCom AG, Austria
Kevin Webb, Conveyal, US
Jesper Wibrand, DTU Transport, Denmark
Lin Zhang, Tsinghua University, Department of Electronic Engineering, China

The project was coordinated by Philippe Crist and Sharon Masterson of the International Transport Forum.

Table of contents

Tables

Figures

Executive summary

Background

What issues arise with the increased use of location data and require special attention regarding privacy, trust and security? Given that much of this data is produced by commercial actors and is often central to their business strategies, what new models for accessing and sharing data would allow collaboration between public and private sector in sourcing, accessing or co-creating data for better managing transport operations and improve the planning transport networks?

Building on our 2015 study *Big Data and Transport: Understanding and Assessing Options*, this report presents the findings of an extensive exploration of these two broad themes at a workshop on "21st Century Public Interest Data Sharing" in Paris in November 2015, which involved a wide range of experts and stakeholders brought together under the auspices of the International Transport Forum's Corporate Partnership Board.

Findings

Data are essential to the planning, delivery and management of transport services and infrastructure – whether data covering home and work locations, leisure destinations and demand for travel between these and others. Data are also necessary for ensuring the safe operation of traffic, to respond to incidents in real-time and to understand and address crash patterns and trends. Increasingly, vehicle and map data will become essential for supporting higher and higher levels of automated driving.

Much of transport-related data has a geospatial component that allows for a more detailed understanding of where people are, where they are travelling, in what conditions and in some cases how and for what purpose. This data is being sensed in new ways, from a broadening array of sensing platforms and in a wide range of formats with several recognised advantages over traditional data-collection methods, notably scale (coverage of entire transport networks) and latency/frequency of data collection (24 hours a day; 365 days a year; in many cases in real-time).

Policy insights

Data is being collected in ways that support new business models in transport but challenge existing regulation

Infrastructure-generated data is quickly being replaced by sensor-generated data, largely via the proliferation of mobile phones, on-board navigation devices and vehicle-to-vehicle communication. Sensor-generated data has given rise to the development of new business models that deliver services linked to the location of an individual or that can be enhanced with this data. Public authorities often lack the ability to monitor and control the use of this data.

Transport data is shifting to the private sector and away from the public sector

The share of mobility-relevant data collected by the private sector is growing. The private sector collects millions and millions of data points in the context of commercial activity or as a by-product of location-based services, and a considerable gap with the public sector is emerging. Yet data from location-based services would allow governments real insights and the benefits of closing the data gap are potentially quite large, allowing potential efficiencies in public sector performance through new data-based services or streamlined operations.

The shift of data ownership from the public to the private sector may ultimately imply a shift in control

The ever-increasing accumulation of data by the private sector could lead to a future where most traffic operations and control responsibilities are effectively outsourced to those that hold the data. In a not too distant future, navigation services providers who are already layering traffic information, digital mapping and navigation algorithms over the road infrastructure, might take control over traffic flows. Ultimately, fully automated vehicles will create and use a high-definition and seamless representation of transport infrastructure that may surpass in quality that held by public authorities. This shift from public to private control is already happening, e.g. in traffic control centres managed by commercial operators.

Transport authorities should account for biases in the data they use and encourage use of adequate metadata

Data generated from location-based services have inherent representativeness biases, i.e. they only reflect behaviour by social groups that have access to the data-generating technologies. Knowledge of these biases is extremely important to make an informed decision on whether data is usable for a specific task, whether it needs to be corrected and how, or supplemented with other data. Thus, some level of metadata or statistical information accompanying any dataset is needed to make the data potentially useable.

Mandatory private-public data sharing should be limited. Only where clear benefits to all parties exist and public authorities have capacity to handle the data should they be considered

Public authorities can compel regulated entities to provide data. They should do so when mutual benefits exist – for example, establishing data sharing schemes in return for transport service licensing – or when data sharing is required to deliver on public policy objectives. Simply requiring regulated parties to provide data may not be sufficient for authorities to extract useable information from it. The skills to understand, format, clean, parse and analyse large data streams are not typically found in the public sector. Public authorities with limited budgets will have to compete with high-paying private-sector companies for data scientists and statisticians.

Data sharing does not necessarily mean sharing raw data

Public authorities may benefit from a sliding scale of data access that reflects their needs and capabilities. A focus on access to data may overlook new possibilities of bespoke data management and analysis services tailored by private-sector firms to the needs and capacity of a government agency. Some agencies might be well served through a dashboard overview of key indicators while others might want access to the raw data feed to carry out their own analyses. In the former case, data and output auditing will be necessary to ensure that the output can be trusted.

Whatever data is collected and whoever holds it, dats should be an integral part of more flexible regulation of emerging transport services

The way data is collected, processed and stored is likely to fundamentally change in the near future. Decision makers now have the opportunity to influence and shape this development process. New forms of data collection and new data types can help support more flexible regulation. In particular, better, more timely and finer data can better target regulatory interventions to achieve specific outcomes.

1. Introduction: Location-based mobility data

Data is essential to the planning, delivery and management of transport services and infrastructure - data covering home and work locations, leisure destinations and demand for travel between all these and others as well. Data is also necessary for ensuring the safe operation of traffic and understanding and addressing crash patterns and trends, as well as responding to such incidents in real-time. Increasingly, vehicle and map data will become essential for supporting higher and higher levels of automated driving. Transport agencies have been collecting data from an array of sources. The data generally falls into four known applications: traffic volumes and flows (counts), network travel times and traffic speeds (historic and in real-time), incident detection and trip origin-destination matrices.

In order to produce these information applications, governments have at their disposal a number of instruments and mechanisms for collecting transport data. However, collection of this type of data has usually been time-consuming and not immune to the characteristic trade-offs between collection costs, coverage and accuracy. Construction of origin-destination matrices, which are an essential input for planning the development of the transport network, or even modifications to it, are generally based on traditional household travel surveys. These surveys are complex, requiring the calculation of large samples and a logistical setup for its distribution and collection, which can be costly and time consuming. Examples like these abound, where the collection mechanisms traditionally known and utilised by the governments are not always keeping up the new technological innovations, nor are adapted to capture the rapid evolution of trends and behaviours within cities.

Much of this data has a geospatial component , as well as a temporal component, that allows for a more detailed understanding of where people are, where they are travelling, in what conditions and in some cases how and for what purpose , all this throughout different times of the day. Much of this data is being gathered in new ways, from a broadening array of sensing platforms and in a wide range of formats, with several recognised advantages over traditional methods, such as: scale (coverage of entire transport networks - e.g. road and public transport), data collection latency and frequency (24 hours a day for 365 days a year, and in many cases real-time collection). This data is collected, stored and exploited by a diverse set of actors that extends well beyond the field of transport, and, especially to the private sector. All of these developments enable the delivery of location-based services (LBS).

As the private sector continuously collects millions and millions of data points as part of their business models, or as a by-product of the location-based services they provide, the share of mobility-relevant data collected by the private, as opposed to the public sector, is growing and starting to create a considerable gap. But if these gaps are to be closed (and this is debatable), new relationship models and partnerships will be needed. Prior work undertaken by the at the ITF in the context of its Corporate Partnership Board (CPB - see Box 1) discussed these changes and noted that one of the key challenges facing authorities was how to manage the delivery of public policy with increasingly privately sourced and owned data concerning the location and movement of individuals.

The benefits of closing this gap are potentially quite large, given efficiencies which could be leveraged through new services with this data. The limited number of applications listed before is mostly constrained by the type of data collected and the mechanisms used to collect it. New applications abound, e.g. using indoor location fixes to determine pedestrian flow patterns or waiting times at stations, or using vehicle occupants' mobile device accelerometer data to help identify pothole locations through vibrations patterns. In the near future it is foreseeable that many other applications will emerge as new business models are built around them, and if these benefits are to be realised, public and private sector incentives towards sharing of this data should be better aligned.

This type of location data can be very useful for managing transport networks and planning for new capacity. It can complement existing data collection and, in some cases, even replace several of the traditional data collection methods at a fraction of the costs. But location data is highly personal and difficult to robustly anonymise, and there are real questions as how to balance data privacy and the benefits that can be derived from innovative uses for this data for managing and helping plan for transport activities.

As a next step of the data-related work of the CPB, a workshop entitled "21st Century Public Interest Data Sharing" was held in Paris in November 2015, in order to analyse some key issues in more detail and to involve a large variety of experts and stakeholders in this domain.

Box 1: **CPB Report "Big Data and Transport"**

The ITF Corporate Partnership Board (CPB) Report "Big Data and Transport: Understanding and assessing options", published in May 2015, examined 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. The report gives an overview of relevant issues, broadly characterises Big Data, and describes regulatory frameworks that govern data collection and use.

Main findings:

- The volume and speeds at which data today is generated, processed and stored is unprecedented. It will fundamentally alter the transport sector.
- Sensors and data storage/transmission capacity in vehicles provide new opportunities for enhanced safety.
- Multi-platform sensing technologies are now able to precisely locate and track people, vehicles and objects.
- 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.
- Location and trajectory data is inherently personal in nature and difficult to anonymise effectively.
- Data protection policies are lagging behind new modes of data collection and uses. This is especially true for location data.

Policy insights:

- Road safety improvements can be accelerated through the specification and harmonisation of a limited set of safety-related vehicle data elements.
- 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.
- More effective protection of location data will have to be designed upfront into technologies, algorithms and processes.
- New models of public-private partnership involving data sharing may be necessary to leverage all the benefits of Big Data.
- Data visualisation will play an increasingly important role in policy dialogue.

Framing questions for the workshop

This workshop addressed the management of location data privacy. It explored whether there is a need for new models framing access to, and use of, mobility-relevant location data and if so, what they might be. It also looked at aspects of public policy as they relate to access and control of transport-relevant data by addressing the following questions:

- "Privacy-by-Design" principles are unevenly or not at all incorporated into location data collection. Should this change and how might this impact the usability of location data for traffic operations, planning and safety applications?
- What strategies exist to durably protect sensitive personal location data? Should more personal control of individuals' location and mobility-related data be offered or mandated? If so, how?
- What are the broader public policy implications of a switch to more and more private control and ownership of transport-relevant data?
- Is there a need to move beyond the current supplier-client relationship governing public authority access to most privately collected location and mobility data? If so, what form might this relationship take?
- Is there a benefit from minimum public interest data sets for transport operations, planning and safety applications? If public interest data sets were to be operationalised, how might they be specified and what are the technical challenges in establishing them?
- Currently, there is increasing pressure for public data sets to be open for public use. Can open location and mobility data be reconciled with increased data and privacy protection?

Background

There has been an explosion of data resources in the transport sector in the past decade, particularly data resources relating to location and activities of individuals. These datasets are already being used by innovative commercial actors, transport authorities and other government agencies for a variety of purposes. Whilst many understand the vast potential of using these data resources, many are also searching for a more precise understanding how this potential can be achieved. In particular, how to overcome the many "small data" and "big data" challenges inherent in conjoining multiple disparate data sets in such a way as to extract trustable and useable information. Major challenges to this exist in the fields of privacy, trust, and security around this data from the point of view of individuals, organisations, and governments.

The workshop examined two broad discussion themes:

- What are unique issues arising in the context of location and mobility-specific data that require special attention regarding privacy, trust, and security?
- Much of this data is produced by commercial actors and is central to many, but not all, of their core business strategies. This leads to a situation where the most timely, accurate and helpful data to carry out public policy is no longer held by the public sector mandated to carry out this mission. What then is the new data access and sharing model that should emerge to allow public and private interaction in sourcing, accessing or otherwise co-creating data necessary to manage transport activities and plan for transport networks?

Traditional relationships amongst industry, technology, government and citizens are changing. Consequently, roles separating production, service provision, regulation, labour, and consumption are blurring as well. Technology is oftentimes driving these changes, particularly by enabling the emergence of

new, often disruptive, in many sectors, but also by allowing more traditional services to innovate and further develop.

In this broad context, the importance of citizen-led change is growing. Some governments increasingly feel they no longer have the right tools or sufficient information to accompany these changes and to deliver on public policy objectives. Large quantities of data are being generated and are increasingly available to governments from the commercial sector crowding out more traditional data collection methods employed by transport authorities. With this data comes new and augmented challenges. These include ensuring sufficient in-house technical capacity to use these disparate and oftentimes unstructured data sets and ensuring adequate and inviolable privacy protection for both individuals and commercially sensitive information. Furthermore, the very rapid pace of innovation in the private sector often outstrips regulators' attempts to keep up with changes in technological developments and new services.

Some directly or indirectly transport-related start-ups begin quite small and with little budget, then rapidly acquire external funding through venture capital and other sources and grow into large companies. Once established, these companies have a significant and lasting impact on transport behaviour and behaviours that impact transport demand. Transport data emerging from and around these ventures, particularly geo-localised data, must be seen within a broader context, especially in view of the wider digital enablement of society as a whole. Data fuels both innovation and disruption. A key challenge to be addressed revolves around data ownership and use. Data relating to specific individuals should be treated with respect and care by those who collect them, but also by all other entities in the chain that act upon or conjoin this with other data.

2. Location-based data: Technological and sociological trends

There are a number of broad trends that give rise to the emergence of novel forms of data just as these data also drive a number of societal trends. Policy-makers should be aware of the wider technological, social and business drivers that are promoting greater production, use and sharing of location data. In particular, they should be aware of the multiple sources of potentially transport-relevant location data including the emergence of smartphone sensor platforms, GPS, Wi-Fi and other location techniques. This also includes new sources of location-sensing data including audio-video streams, the development of connected and autonomous vehicles, mobility services, the "Internet of things", etc. This is important because not all data is "born" equal and data provenance, including the technical parameters of data sensing, collection and processing, has non-trivial impacts on the fitness for purpose and representativeness of the data. In particular, policy should account for the growing importance of data emerging from the sharing economy and peer-to-peer exchanges, as well as changing attitudes to the sharing and use of personal location data.

Box 2: **"Data about cites: Redefining big, recasting small"**

The development of data with respect to its use in understanding and planning cities is intimately bound up with the development of methods for manipulating such data, in particular digital computation. Although data volumes have dramatically increased as has their variety in urban contexts, largely due to the development of micro devices that enable all kinds of human and physical phenomena to be sensed in real time, big data is not peculiar to contemporary times. It essentially goes back to basic notions of how we deal with relationships and functions in cities that relate to interactions.

Big data is thus generated by concatenating smaller data sets, and in particular if we change our focus from locations to interactions and flows, then data has faced the challenges of bigness for many years. Thus one needs to be more careful about defining what is big data; for this it could be useful to look at traditional interaction patterns, i.e. flows of traffic in cities and show some of the problems of searching for pattern in such data.

Furthermore, examining much more routine travel data which is sensed from using smart cards for fare-charging and relating this to questions of matching demand and supply in the context of understanding the routine operation of transit can gives a sense of the variety of big data and the challenges that are increasingly necessary in dealing with this kind of data in the face of advances in digital computation. Thus there are large and relevant changes in the data landscape and their effects. But these are perhaps not that new, radical or have as big potential as the current hype surrounding this topic might suggest.

Source: Presentation by Michael Batty, University College London,at workshop (Paris, 9-10 November 2016)

The workshop touched upon a number of issues related to data trends and specifications:

- What are the key trends driving growth in spatial big data and the related technological (platforms) and sociological (sharing economy, generational changes to perception) developments?
- In what ways are these trends likely to change over time (e.g. in a 10-year time frame)? What potential disrupting trends (e.g. vehicle automation or automation technology more generally, new ways of organising transport services that are real-time responsive) can be identified? What might be the effects of an unauthorised release of data, causing distress or damage?
- How to define transport or transport-relevant data? Is it limited to data generated by the transport sector, e.g. through operation of systems? Transport is a derived demand, i.e. travel to carry out a variety of activities (shopping, work, recreation, etc.), thus should behavioural data also be

included here, and how should these be linked? How does transport relate to the wider city, in the narrow as well as the wider sense of transport data, now and in the future, and in terms of stakeholders?

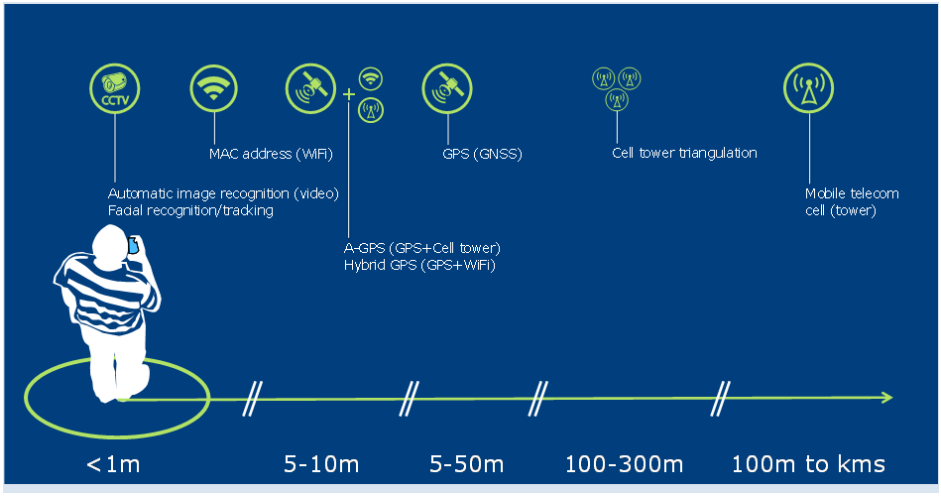
New technologies

The sourcing of data is quickly evolving from infrastructure-generated data to sensor-generated data, largely via the proliferation of mobile-based devices, such as smartphones, on-board units (fleet management) and automobile-based (portable navigation devices, in-vehicle navigation devices and connected vehicles). These data sourcing devices and techniques contrast greatly with past data collection methods, such as mobile floating car data (data for speed and travel times) or embedded sensor technology, (data on volume and that can infer travel speeds and times).

The main difference between fixed sensors, hard-wired or wireless, and mobile-based sensors, is that the former ones are permanently installed and collect data about the activity of the surrounding terrain, which can be the speeds at which vehicles pass by, distance between vehicles, or similar activity, and the transmitted data has embedded into it the location coordinates of the sensor collecting the data. Whereas, mobile-based sensors instead transmit the device's location coordinates along with a timestamp attached of when the location was recorded or sent to the server (and often a timestamp for when the data was received at the server to capture latency). Using the set of location coordinates collected by the mobile sensor analysis can then derive vectors that are attributable to the sensor, such as speed, distance between two positions or direction of movement.

In addition, mobile sensor-generated data has given rise to the development of new business models, particularly ones that leverage this data to deliver services which are dependent on the location of the individual, or that can be enhanced with this data. One of the early enablers of LBSs was the 1996 United States Communication Commission's mandate to implement wireless position systems to allow 911 emergency callers to be located using cell tower triangulation with a precision of 125 meters. Today, cell-tower triangulation is still used to produce location-based data from mobile devices but given that most LBSs require higher levels of location accuracy than can be delivered by cell towers, other, more precise localisation techniques are commonly used (see Figure 1).

Figure 1. **Geo-location technologies and accuracy**



Source: ITF (2015)

A particularity with GPS-enabled mobiles and smartphones, is that location data is often being collected passively and almost ubiquitously by any smartphone that is turned on, basically working as a sensor that individuals carry around, tracking their movements and activities day and night. For example, a 2014 research study from Carnegie Mellon (Dwoskin, 2015) (Almuhimedi, et al., 2014) tracked and monitored participants' download and use of their own choice of apps with bespoke software that recorded app requests for a variety of permissions including location readings. The study revealed that a dozen of the most popular Android apps collected location data from users 6 200 times with accuracies of 50 meters over the two-week study period – or roughly every three minutes. These apps included the Weather Channel, which uses location data collected every 10 minutes to deliver localised weather reports as well as pre-installed software from Google that collected location data more frequently. At the time of the study the Android platform granted share location permissions en masse to the rest of the installed apps, making it difficult for the consumer to have a choice over which apps to allow and which ones not. But perhaps the most relevant finding was that collected location data was shared 73% of the time with a third party advertising network.

While much more is known and communicated about the development of new and innovative business models, than about the technology development itself, there are a fair amount of known developments taking place that will have a large effect on data collection. Wi-Fi is an example of a technology which is getting some attention, mainly because of the ubiquity of mobile devices with such capabilities, but also because the number of Wi-Fi networks that can be detected and used to identify locations over time. Wi-Fi location-based systems are composed of access points and data services (internet). The access point will collect the mobile device's media access control (MAC) address, which is transmitted by the mobile device when pinging Wi-Fi access points as it seeks to connect to known networks.

Singapore's Land Transport Authority (LTA) has been using Wi-Fi -based localisation in order to monitor and analyse location data within the Mass Rapid Transit (MRT) stations. The pilot started through the deployment of free Wi-Fi within 33 MRT stations as a mechanism to improve the customers' waiting experience at stations. In parallel LTA used the same Wi-Fi network to get the commuters' devices locations and use that data to monitor dwelling times at different stations and to better design the physical layout of MRT platforms. LTA plans to expand the program so that by 2020 all MRT stations will have free Wi-Fi services.

New business models

The proliferation of GPS-enabled mobile devices - numbering approximately one billion in 2013, with smartphones representing the majority (GSA, 2015) - has led to the emergence of new business models leveraging, collecting and processing location data to deliver a wide range of LBS related to transport. These principally concern vehicle navigation and multi-modal routing services, but some services also relate to the provision of transport services supporting the movement of people and goods.

To put things in perspective, the 2015 GNSS Market report for the European Union (GSA, 2015) estimates that there are approximately three billion apps currently relying on location-based data to power their services, with apps for the Android operating platform representing the majority.

Another related LBS trend enabled by on-board units or vehicle-navigation systems (portable or in-vehicle), is the growth of the use of data generated by these devices for road transport services. On-board units were traditionally were used to support logistics for fleet operations and monitor its performance. Now, on-board units are being used by transport authorities to implement country-wide electronic tolling solutions for heavy goods vehicles weighing more than 12 tonnes, with Germany and Switzerland acting as early pioneers followed by Hungary and Slovakia. Increasingly, however, for-vehicle navigation systems are being supplanted by smartphone-based platforms linked to individuals, rather than to vehicles. But in-

vehicle location-based technologies remain an important source of data and their numbers (but not their relative share) is growing with automotive manufacturers becoming the integrators for such systems and the owners of the generated data.

The range of commercial actors collecting location data and the wide spectrum of use of this data is considerable. The discussions at the workshop centred around a number of data services and providers that revealed a broad diversity of that can be classified according to the transport 'purpose' for which data is collected and the kind of information product/application created from processing the raw data (see Box 3).

Box 3: **Location-based data sources and analytics**

Data sources

1) Data services for transportation - GPS-based
   - a) Road navigation: TomTom, Waze, Garmin
   - b) Multi-modal routing services: Google, HERE
   - c) Transit-based routing: CityMapper, Moovit, The Transit App
   - d) Bicycle and running: Strava (also sells the aggregate and anonymised data to governments)
   - e) Provision of transport services (use of the data for running their business)
     - i) Taxi-hailing apps:Didi Kuaidi, Easy Taxi, Gett, GrabTaxi, Ola Cabs
     - ii) Ride-hailing apps:Cabify, Lyft and Uber
2) Non-transport location-based services
   - a) Social networks:Instagram, Twitter, Facebook
3) Non-GPS tracking technologies
   - a) Automatic number plate recognition (ANPR)
   - b) Transponders/tags: electronic toll collection systems
   - c) Bluetooth
   - d) Wi-Fi: Bitcarrier
   - e) Mobile data: Telecom companies
   - f) Payment systems: smartcards, contactless payments

Data analytics

4) Transport management and operations (traffic and flow data, traffic density):INRIX, TomTom, Bitcarrier
5) Transport planning data analytics (origin-destination matrices): Streetlight data, INRIX, TomTom
   - a) Telecom data as a service (TDaaS) using Telco cellular data: Airsage

A key message in relation to these business models is that the more vertically-integrated the business model is, the more easily it is to collect location data. Firms, which provide a location-based service through multi-modal routing apps (Tier 2), but that also participate as a Tier 1 supplier for the mobile-device software platform will have the ability to collect data from multiple entry points. The first data collection entry point is through the mobile-device itself and its operating system. In this case, data coverage will depend on the number of people using such a device and there is some evidence of income and geographic stratification by operating system that could bias analysis based on this data. The second entry point is through Apps offered on the platform. In these cases, population coverage may be lower, though this data

potentially provides more information about the type of trip and the users' demographics through links to accounts, social network accounts or favourite addresses (oftentimes home and work locations).

On-board devices have also been leveraged by transport agencies in places where the latter have jurisdiction over certain types of vehicles they regulate such as taxis. In these instances (such as Beijing's or Seoul's taxi fleets) the vehicles serve as data-collecting probes. Deriving useable knowledge from such data streams is nonetheless not a trivial task since raw data has to be processed and transformed, and issues such as data cleaning (e.g. stripping vehicle IDs), sample selection (not all the network links will be represented equally by the GPS traces, therefore statistical work to compensate for this bias is necessary), the need to assign data points to the road network and other data transformation may be necessary before use. Data transformation may be challenging for many authorities given the high ongoing costs and specialist staff necessary for the sole purpose of having probe vehicle data. Cities like New York, Beijing and Seoul derive this kind of data from their taxi fleet, but not always uniformly. For instance, in New York, location and time stamps are generated only when a transaction takes place, therefore only on the pick-up and drop-off, and not during the trip. Though there is clearly an appetite for data on the part of public authorities, there are real limits to the technical capacity of many of these to extract meaningful insight from large, unstructured data sets.

Data may also be used in support of infrastructure pricing, transport service delivery, and performance-driven planning and contracting. In 2014, Australia's Bureau of Infrastructure, Transport and Regional Economics (BITRE) organised a workshop to address the viability of using new technologies to improve data collection mechanisms. They specifically looked at new traffic data sources and their ability to improve current regulatory and operational functions for road authorities (BITRE, 2014). The main motivation for the BITRE was to use this data to improve toll road revenue forecasts and projections and potentially lower bidding costs, given that this issue has been financially damaging to Australia over the last decade. The workshop's main messages were:

- Establishing what data is really useful and in what format.
- Communicating what data is available, who owns it and how it can be accessed.
- There are still technical issues to resolve as to the best data fusion to deploy and how to extract difficult components from trip data such as modal information.
- It was agreed that the transport sector is unlikely to provide a commercial business case for accessing the data, a range of other potential data users is needed.
- Informed express consent would satisfy the existing regulatory framework, however, it is recognised that there are still challenges.
- One of the challenges is to make data no longer Personal Identifiable Information (PII) through a de-identification process and yet still maintain the data's value to the data analyst.

Business model innovation, including innovative uses of data and novel data sources, is also happening separately from the development of new technologies, and these processes are accelerating at a much faster pace than what governments and transport agencies are typically used to. By the time evolving business models and the technologies that accelerate or otherwise facilitate their adoption converge and start to play out on city streets and in public space, it is often difficult or even too late for regulators to act effectively. New models for anticipatory but flexible policy-making are necessary but have yet to be developed.

Much more is known and communicated about the development of new and innovative business models, than about technology development itself. Nonetheless many technologies are being developed that will have a large effect on data generation and collection. In particular, the trend to miniaturisation is an important one that will multiply data monitoring possibilities through, for example, the use of embedded sensors in transport infrastructure and use of novel kinds of embarked sensors in clothes and objects, etc.

The deployment of new sensing and monitoring technologies will, in turn, lead to new business and public use cases that are either anticipated by technology developers or will increasingly be unanticipated - resulting from novel ways of combining data from different sources with value-adding data analytics.

Many participants in the workshop expressed confidence that the greater good of society will dominate in the trend of data production and access. In the case of a health crisis for instance, public authorities seeking to understand and control disease vectors are likely to request, and gain, access to data enabling them to do so. However, this access will be conditioned by the need for commercial service providers to ensure the robust protection of their customers' data. This tension is precisely what came to a head in early 2016 when the US Federal Bureau of Investigation sough to compel Apple to provide access to an encrypted device in the context of an anti-terrorism investigation.

The issue of commercial trust vs. public good notwithstanding, participants saw promise for combining new forms of cross-sectorial data emerging as part of the sharing economy and individual ownership of data relating to citizens. The latter enables monetising this data for personal gain and the emergence of new forms of services, including data brokers that also manage privacy and trust profiles for individuals. All of this will require new forms of regulation (private or self-regulation) enabling citizens to confidently open up their personal location and other data. The risk is that, absent robust data protection, a breach of personal location data will occur thus leading to a radical change in the public perception of the adequacy of data protection and a reduction in the willingness of individuals to give access to personal data.

Location-based data: Assessing fitness for purpose and other challenges

Smartphone penetration rates are not uniform around the world, or even within countries or cities. As ubiquitous as LBSs are, with almost complete coverage for those with a GPS-enabled mobile device, the data generated still is just a sample of the whole population. Sampling bias is further exacerbated by different penetration rates for specific apps and smartphone operating systems, both of which can lead to other significant biases. Transport authorities must carefully consider issues with non-representative sampling when looking to use this type of data. There is arguably a need for a sliding scale of policies or guidelines related use of this type of data for different parts of the world or even within different parts of the same country (urban vs. rural) depending on the presence and strength of these biases.

Depending on the purpose and task, biases in the data source matter to different extents. For example, biases in the data source when generating traffic speed statistics can matter less, than in instances where the task is to prepare analysis for planning purposes. In the latter case, biased representativeness can lead to severe underrepresentation of certain groups within the population. For this purpose focusing on building a well-structured sample will be very important. But even for traffic speed statistics, biases may be present based − for instance, speed profiles generated from GPS-equipped taxi fleets may not represent a good proxy for real traffic speeds when taxis are allowed to operate in separated and less-congested bus lanes. Likewise, data derived from commercial delivery fleets may be biased from vehicle use patterns that differ greatly from the passenger vehicle fleet.

While poor representativeness may in many instances not be an obstacle to using data from smartphones to calculate traffic speeds for motor vehicles on specific road segments, the same cannot be said for calculating traffic density or flows between different parts of the city. There is also the question of how to incorporate data on the movement of those not generating data (e.g. pedestrians or cyclists not carrying a mobile phone) into traffic management and transport planning. Hybrid models of combining big data analytics with more conventional survey data are, and likely will remain, necessary for the near-term future.

For these reasons knowledge of these biases is extremely important to make informed decisions on whether the data needs to be corrected and how, supplemented with other data sources, or deemed unusable for that specific task. Therefore some level of metadata or statistical information accompanying any dataset, whether raw or post-processed is needed to inform the potential uses of the data.

Going beyond transport data

In order to infer information useful for transport planning purposes, much more than just transport data is needed. Additional data sources covering demographics, land-use (location of business, services and activities) and historical patterns of flows through time and space are required to characterise future transport demand. Knowledge of flow data allows authorities to better plan activities and manage flows, reducing congestion, ultimately mapping the life of citizens. But transport data alone only covers part of the spectrum of mobility-related data. Data on environmental impacts, road safety (crash rates, locations, etc.), housing prices, shopping patterns and other uses of public space could be better aggregated to provide a more holistic view of the city ecosystem that extends beyond what is happening, or is observed, on roads and in transport systems.

New possibilities are emerging for this aggregation to take place but fundamental questions remain as to how "smart" cities can and should be. Successful cities are so because of managed chaos that gives rise to serendipitous encounters and random events that generate knowledge, opportunity and economic activity. Designing away this chaos could also lead to a decrease in the creative function of cities and to their attractiveness and effectiveness as engines of prosperity.

Some participants raised the question of what important data is unobserved and unavailable to authorities. For example, observed behaviour does not account for suppressed demand – that is all trips not currently taken because they are perceived as too impractical or uncomfortable or are simply not possible under current conditions. In order to capture such data, other information sources (e.g. quantitative or qualitative surveys and other market research techniques) are still necessary in addition to purely using sensor based data.

Participants also pointed out that with the proliferation of multiple different sensing platforms there are new possibilities for merging data from these within the public sector (e.g. between health and transport) just as there are new possibilities for data merging amongst different commercial operators (e.g. between commercial weather services and app-based ride-sourcing platforms). Of course there is the potential for novel types of public-private data analysis. In some cases, big data analytics are superfluous when small carefully crafted data sets can provide sufficient insight at a fraction of the data collection and analysis load required for big data.

Data visualisation has added new dimensions to the ease with which sometimes complicated and intractable challenges can be quickly and easily understood and acted upon. Good data visualisations are helpful in communicating issues and in seeking to motivate or compel action. In some cases, they can also act as part of the analytical process and allow the public and authorities to discover previously unsuspected issues or solutions. However, it is important to understand that data visualisation often involves a number of embedded assumptions regarding the types of indicators to illustrate and their relative importance. In this respect, they can be seen as data "journalism" that presents a curated selection of specific parts of the bigger picture. Visualisation should therefore be seen for what it is – as an aide to comprehension – and not as a stand-alone analytical framework replacing validated scientific approaches like statistical analysis and modelling.

3. Collecting and using personal location-based data: Privacy and other risks

Participants discussed the nature of data-related risks and threats faced by different entities including individuals, transport operators, information service providers, the automotive sector, infrastructure owners and operators, non-transport data collectors, public institutions, regulators and international organisations. In particular, the workshop touched on specific risks related to location data and covered different types of risks and threats. These could include, for example, the risk of discovery of sensitive personal information and mass surveillance. It could also include the threat of access to highly personal data by government agencies, criminals, or terrorists. Discussions touched on a number of potential technological and institutional responses which could avoid, mitigate and where necessary repair integrity and trust. Participants also noted that authorities should critically evaluate both existing and emerging approaches and identify gaps in capability regarding access to and use of personal location data.

Box 4: **Privacy-by-Design as a framework guiding the use of geo-location data**

Why is there a need for Privacy-by-Design (PbD)? Most privacy breaches remain undetected, unchallenged and unregulated, as regulators only see the tip of the iceberg. Regulatory compliance alone is unsustainable as the sole model for ensuring the future of privacy. The approach of PbD aims to change the paradigm from a zero-sum to a "positive-sum" model, creating a win-win scenario, not an either/or one involving unnecessary trade-offs and false dichotomies.

There are currently nine PbD application areas, including Closed Circuit Television (CCTV) in mass transit, biometrics in gaming, smart meters/grid, mobile communication, near field communication, Radio-Frequency Identification (RFID)/sensor technologies, redesigning Internet Protocol (IP) geo-location, remote home health care, and big data/data analytics. PbD is based on seven core principles that address data minimisation and de-identification. Though big data and open data should be encouraged, not all data is the same, and when it comes to personal data, protection becomes more challenging.

There are many fears associated with location data. The US Federal Trade Commission has cautioned that "location data can quickly become sensitive personal information". And privacy advocates argue that location tracking via mobile devices is "the deepest privacy threat, and is often completely invisible".

Data minimisation is the most important safeguard in protecting personally identifiable information, including for a variety of research purposes and data analysis. The use of strong de-identification techniques, data aggregation and encryption techniques, in particular, are absolutely critical. Public datasets use de-identified data to gain the most from location and mobility data.

The claim that de-identification has no value in protecting privacy due to the ease of re-identification is a myth. If proper de-identification techniques and re-identification risk management procedures are used, re-identification becomes a very difficult task. While there may be a residual risk of re-identification, in the vast majority of cases de-identification will strongly protect the privacy of individuals when additional safeguards are in place.

Therefore, there are considerable risks in abandoning de-identification efforts, including the fact that individuals and organisations may simply cease disclosing deidentified information for secondary purposes, even those seen to be in the public interest.

Source: Ann Cavoukian, Ryerson University

In particular, the discussion of threats, risks and responses in relation to location-based data addressed the following issues:

- What are the specific risks and threats to privacy security? Whose interests could be imperilled and in what way by various breaches enabled by the new data environment? Also, what are "real" risks as opposed to mere irritants?
- Is Privacy-by-Design really a positive-sum approach, is it possible to build privacy into innovation without hindering it - or are they mutually antagonistic?

Threats and risks

Threats and risks can materialise themselves in two forms; e.g those directly affecting individuals and private citizens or those affecting commercial entities. In both instances, data security can be compromised through misuse irrespective of whether the data was collected legally or illegally. Misuse here includes compromising trade secrets, intellectual property, and espionage in the case of companies, the illegal use or publication of personal data relating to individuals and, more generally, hacking, cyber-terrorism and intelligence agencies illegally gathering data.

The problem lies less in how data is collected, but more in how it is then consequently processed and used. Issues relating to data collection could be overcome more easily by encryption. An example is GPS data currently being stored and transmitted in simple text formats. In this case, up-front encryption could be integrated into the data sensor/processor directly, rather than in post-processing, thus minimising the possibility of raw data falling into the wrong hands. Post-processing is generally well regulated by governments and this is also an area where law is more easily applicable. There is also a large risk of data discovery due to third parties handling data in a negligent way. This includes companies not implementing the appropriate procedures to guarantee that security breaches cannot occur but also may involve cases of specific members of staff accessing data for personal purposes or giving access to data to others in an unauthorised way.

Anonymisation

Each information application that is generated (traffic speeds, traffic incidents, traffic volumes and origin-destination trips) from location data present particular challenges for anonymisation. Traffic speeds and travel times represent the less challenging of the three, and there is a relatively good source of academic research tackling this problem. For example, the Mobile Millennium research project launched in 2008 by UC Berkeley, the Nokia Research Centre, and NAVTEQ, used location data generated from the participants' GPS-enabled phones to generate traffic and speed data (UC Berkeley, 2011). The project's main breakthrough, was that not only were they able to create a reliable traffic-monitoring system from sparse data, but that they developed a methodology, named the "virtual trip lines" to protect the privacy of the participant individuals, and transmit the data using bank-rated encryption (Hoh, et al., 2008).

Firms like INRIX, StreetLight Data and TomTom also have addressed these issues and commercialise traffic data from a number of data sources and providers. Other companies like Bitcarrier, also produce traffic data with the aid of Bluetooth and Wi-Fi sensors.

Road incidents and event data are perhaps the easiest of the applications to deal with, given that only one location is needed and the risk of discovery of individual's identity and trip-making patterns is eliminated. Unlike other sequential location-based data, incident data concerns traffic or road events that have occurred, or that will occur in the future, but are not necessarily connected to an individual's movements. The Bay Area Metropolitan Transportation Commission (MTC) is working with other Canadian partners on the development of an open standard called Open511 (MTC, 2015). The standard will represent this type of

data which can be retrieved through an application programme interface (API). Specifically it will include: road incidents (accidents), construction, special events, weather conditions and road conditions. The goal behind its development was to enable interoperability to use this data, through an open standard that would allow jurisdictions and agencies to share this amongst each other, but also to share it with the open data community. MTC currently shares this data, bundled with their real-time traffic feed, but will shortly open an API supporting Extensible Markup Language (XML) and JavaScript Object Notation (JSON) specifically for road incident data.

Traffic volumes and flow data present a different set of challenges given that they cannot be solely derived from speeds since other factors such as road capacity have to be taken into account. In such cases data fusion from different sources such as road sensors or non-GPS mobile data from telecommunication service providers may be an option.

Origin-destination (OD) matrices are probably the most problematic use case because the most valuable insights are drawn from analysing highly disaggregated data for individual travellers. There is no clear consensus as to what is the best method for anonymising data for this purpose (for a broad explanation of the different methods, see the International Transport Forum's Corporate Partnership Board "Big Data and Transport" report released in 2015). But a key issue when choosing a method is how to strike a balance between anonymising the data while still retaining the data's value despite this simplification. Along these lines, in 2015, Uber reached out to the city of Boston and offered to provide them with quarterly anonymous data about the trips that were originating or finishing in the city, but aggregated by zip code. Specifically the data consisted of the zip code locations for pickups and drop-offs, timestamps, and duration for each of those trips. This approach taken by Uber addresses some of the privacy issues around the sharing of this data, but it also highlights the need to pay careful attention to the purpose and the uses that authorities wish for the data, and a need to evaluate if the nature of the data is deemed adequate for the policy issue at hand. In the case of the Boston data, one might surmise that the data represents those relatively well-off individuals with access to a smartphone, subscription to data services and enough income to use Uber's services.

For planning purposes more disaggregate information about users and trips is usually better, but this approach will not always be feasible given the privacy constraints around it. This is an open question with no clear consensus as how it should be addressed, but given its importance it deems to be highlighted. Further work needs to be done as how to make this function for all parties.

Other possibilities, already being undertaken by some firms, entails creating these OD matrices or any information application requested on behalf of the government, and providing the results to the government or interested parties, such as transport modelers, but without the underlying data. In other words, these companies serve as intermediaries, and bear the risk of managing the data and brokering the raw data with the necessary players. Companies like INRIX, Streetlight Data, TomTom and Airsage, this last one using cellular signal data from Telecommunication Companies, are all working in this space.

Much of the discussion surrounding privacy of personal data, location-based or otherwise, relates to models where entities access actual data or otherwise transmit it amongst themselves. This has certainly been the traditional model of data access within the transport field but it is not the only model, nor necessarily the most desirable model, for deriving useable knowledge from data. The value from data comes not from the data itself but from the knowledge derived from the data and so alternative models for generating that knowledge made possible by advances in data science may be preferable in the long run. The OpenPDS/SafeAnswers framework developed by MIT (see Box 5) represents one potential pathway towards generating trusted knowledge from data without ever revealing the data. Other strategies rendered possible by computing advances include stronger encryption and blockchain technology that increase the inviolability of the data itself.

Box 5: **New frameworks required for location data privacy**

Geo-localised data of the type that is most useful for transport applications presents specific and robust challenges with regards to privacy protection. Crucially, traditional strategies of anonymisation fail to deliver adequate levels of privacy protection. Multiple re-identification attacks on anonymised geo-localised data have highlighted this vulnerability - for example, in a mobile phone based dataset with 1.5 million people, only four external location references (e.g. from other data sources) are sufficient to re-identify individuals in 95% of the cases.

Individual patterns of movement and behaviour are quite unique, and adding "noise" to datasets makes the process of re-identifying specific individuals more difficult, but it does not prevent it. Further, conjoined datasets allow the discovery of highly personal details not necessarily contained in one or the other original data – joining mobile phone records and other available data one can lead to reliable prediction of age and gender. A new framework for securing big data via up-front integration of security features and protocols into data formats and collection – "privacy through security" – can help to address the vulnerabilities inherent in traditional data collection and privacy methods.

Freely releasing data which has been imperfectly anonymised runs the risk of de-identification and discovery by third parties. The SafeAnswers framework (below) transmits code to act on data rather than releasing data for analysis by second- or third-party data processors.

The OpenPDS/ SafeAnswers Framework*, developed at Massachusetts Institute of Technology, allows users to collect, store, and give fine-grained access to their data all while protecting their privacy. SafeAnswers allows applications to ask questions that will be answered using the user's personal data. In practice, applications will send code to be run against the data and the answer, not the data, will be sent back. By transferring code, not data. OpenPDS/SafeAnswers turns a very hard anonymisation problem into an easier security challenge.

SafeAnswers uses two separate layers for aggregating the user's data: first sensitive data processing takes place within the user's Personal Data Store (PDS) allowing the dimensionality of the data to be safely reduced on a per-need basis; second data can be anonymously aggregated across users without the need to share sensitive data with an intermediate entity through a privacy-preserving group computation method.

With SafeAnswers generic computations on user data are performed in the safe environment of the PDS, under the control of the user: the user does not have to hand data over to receive a service. Only the answers, summarised data, necessary to the app leaves the boundaries of the user's PDS. Rather than exporting raw accelerometer or GPS data, it could be sufficient for an app to know if a person is active or which general geographic zone the person is currently in. Instead of sending raw accelerometers readings or GPS coordinates to the app owner's server to process, that computation can be done inside the user's PDS by the corresponding analytic module.

*Adapted from http://openpds.media.mit.edu/)
Source: Presentation by Yves-Alexandre de Montjoye, Massachusetts Institute of Technology, at workshop (Paris, 9-10 November 2015).

Solutions for consent and terms of use

Eliciting consent from citizens to share their data will require that the former trust data collectors to safeguard privacy. But this trust is not sufficient, or rather, it must be built on an unambiguous understanding of what kinds of data are being collected, how the data will be processed, shared, conjoined with other data and used. This understanding allows citizens to make informed decisions and to internalise the trade-offs between sharing location data for location-based services and the payback to individuals from location-based services. At present, the consent process is unwieldy and oriented to cover the legal obligations of the data collector and thus overly exhaustive for citizens who rarely read, let alone understand, the conditions attached to their data sharing. In order for trust to build, terms of consent and the ways in which these are communicated must improve and should place the citizen, rather than the collector, at the heart of the consent process. Without this approach, it is plausible that people will consent less and may choose to opt out of data sharing entirely. This could affect representation bias for the remaining data and could impact business models dependent on accessing individuals' locations.

Clarity in what people are consenting to (the "terms of use") and flexibility in allowing individuals to opt in to various parts of services (rather than a "one-stop" or all-encompassing consent framework) provides individuals with greater control over, and comfort with, their data sharing. This "customisation" of consent can perhaps be communicated in a more graphic way, using icons, for example, to explain the details of the consent process.

Another solution discussed could be for a privacy service industry to develop acting as trusted third-party intermediaries between private individuals and data-related service providers. These service providers could manage consent based on privacy profiles defined by individuals thus allowing only user-defined data contents to be shared. The privacy profiles can act as a proxy for individuals' comfort with sharing their location data and would be centralised at the level of the third-party consent platform thus eliminating the need to specify sharing permissions for each unique service accessed. Greater uptake of data privacy profiles can also lead data collectors to ensure that their practices conform to the most popular or widespread of the profiles − thus aligning data protection efforts on the part of service providers with preferences expressed by individuals. The development of this industry has only just begun and, even so, principally in the realm of business-to-business transactions. Customer-facing third party privacy management has yet to take off for business-to-consumer transactions (with the exception of privacy related to payments for services as offered by intermediaries like Visa and PayPal) leading some participants to ask if there is an unmet need for policy and regulations to encourage this development.

Participants noted however that it is nonetheless difficult to judge how far concern over privacy protection by consumers goes beyond a diffuse sense of worry about privacy. This may be because such a concern is not very present and that citizens value the services derived from revealing their location enough to outweigh any general concerns they have about privacy. It could also be that the scope of location data usage − both consented to and the potential risk of misuse − are deemed acceptable. Should the possibilities of individualisation and customisation of location data services continue to progress (for example, by serving hyper-individualised advertisement), citizens' unease with data sharing may increase as well. It was also noted that privacy concerns and the importance assigned to privacy protection varies geographically as well as over generations.

As with other regulated industries, there could also be an expectation by consumers that regulation will keep up with technology and that safeguards will be put into place thus obviating the need for individuals to actively take steps themselves to ensure their own privacy. However, regulatory responses in the face of innovation can often stifle what they seek to regulate and many pointed that this is a real risk in terms of encouraging new services.

How then can regulators allow industry to innovate, without harming the whole industry? A solution for this could be for industry to put self-regulation into place by giving their users more liberty as to what they wish to share and with whom. In Android's earlier platform version, the default interface for the app permissions functionality (e.g., what allows the smartphone to tap into GPS capabilities or the camera) used to work under a take-it-or-leave-it scheme, where the user was not allowed to install an app if they wanted to deny specific permissions that were requested by the app, such as the use of the GPS. Under the new Android Marshmallow platform, the interface changed and users are now allowed to manage these permissions independently for each app, similar to iOS practice. The generalisation of user control over permissions represents progress but more can be done to inform average users about how many apps have access to their data. This is true not only for access to location data, but also to photos, camera, microphone, and information about their contacts. Apps permissions have drawn a lot of attention lately, and it seems that they are more important than ever, mainly because of these issues.

Data is not cost-free. Costs include putting in place the technical infrastructure (including deploying apps on smartphones), deploying the software to collect, aggregate and transmit the data, paying for the analytical

capacity to process and extract knowledge from the data, or, alternatively, purchasing data or analytical products derived from the data). In this context, authorities must weigh the costs of data collection and analytics against the magnitude and scope of the expected benefits. Applying benefit-cost analysis to data acquisition and use, however, is complicated as many impacts of data use are difficult to quantify and completely eliminating risk is probably not possible.

In this context, credit card and other banking information data is particularly important in terms of the need for safeguarding them and the very direct threats from malicious use, rather than solely the invasion of privacy as in other examples. In addition, there could be schemes to provide compensation for victims of privacy breaches.

Safeguarding privacy and trust: Allocation of risks and roles

Is there anything fundamentally different from geo-localised data and its privacy risks and threats compared to other data formats and contents? Participants did not feel this appeared to be the case, other than the scale and immediacy of discovery of patterns of behaviour related to location data. There is a difference between the nature and the impact of and individual weighting of security risks. Anonymising location-based data is not as straightforward as it sometimes appears, as after removing parts of the data it is often still possible by establishing specific patterns to re-identify large parts of the datasets. With the growing amounts of data being collected and being accessible, it thus gets increasingly easy to re-identify individuals in anonymised datasets.

It is necessary to individually segment the organisations that have access to data, including enterprises and government agencies, and allocate the different inherent risks to each segment. This should include government agencies asking private data providers for access to data, which was perceived to be increasing even more with location-based data. Should governments abide by different rules than the private sector (perhaps due to perception), even if 'informal', when asking for access to location data? Origin-destination surveys provide data that is roughly analogous to that collected by some smartphone apps or operating systems, but governments have a direct communication channel with respondents asking for their consent. With this kind of data, the consent comes from the private sector, not directly from the user.

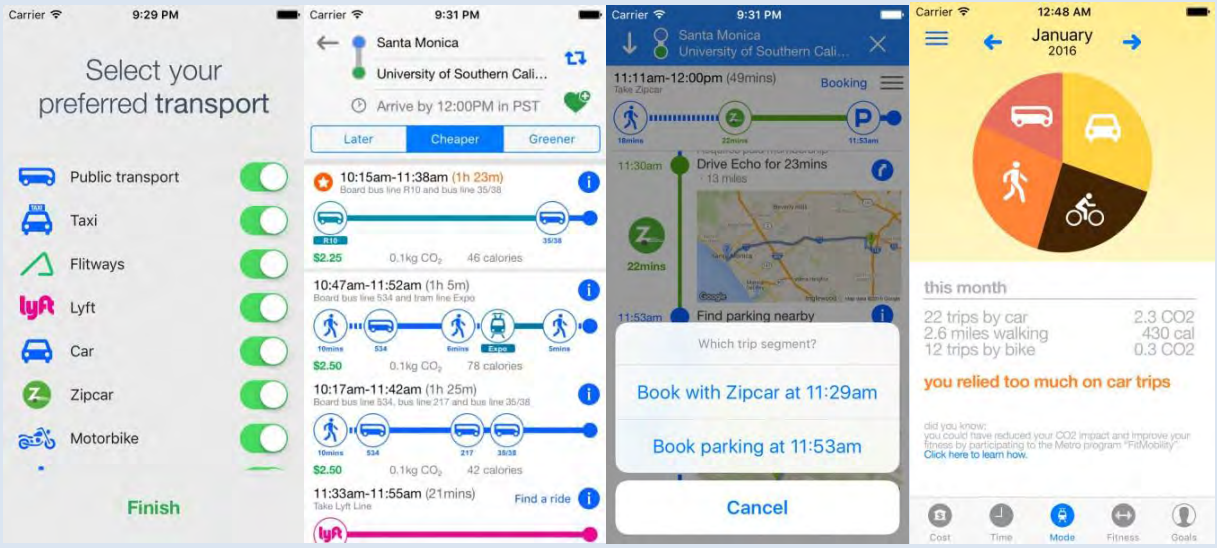
There is also a cost benefit trade-off of large investments, e.g. fitting black boxes into vehicle fleets or mandatory multi-service on-board units for road tolling and fleet monitoring. It increasingly enables employers to monitor their workforce, for logistics, scheduling, but also both individual behaviour and health related information. Along this line companies like Driversiti are developing software, to detect a broad range of driving behaviours and road conditions by using a smartphone's sensors and replace or complement a vehicle's active safety systems. Such companies are gathering not only data about drivers' behaviour time-stamped to specific locations, but also about road conditions, which helps fleet operator companies, ride-hailing services and insurance companies. In aggregate all this data can have an incredible value to governments.

While this type of data sharing might be acceptable or even unavoidable in the work domain, in the more private and personal areas of citizens' lives this will be a very different issue. Consumers are generally willing to share data with private businesses if the benefits of this are immediately clear, e.g. improved customer care, more personalised service, simplified processes or reduced fees, etc, or if the trade-offs are unknown. If the public sector increasingly accesses data from private individuals, they too need to demonstrate the added bonus for citizens of allowing access to this data. Clearly data could conceivably be used by both commercial actors and governments to change behaviour but in the case of the former, the result may be more manipulative (e.g. getting individuals to change their consumer behaviour) whereas in the latter case, the result may be more coercive given the powers wielded by governments.

Box 5: **Xerox GoLA and GoDenver App**

A good example of the public and the private sector joining efforts for collecting this type of data, and at the same time providing a benefit to the user, was Xerox's partnership with the city of Los Angeles and the city of Denver to develop a multi-modal routing and booking app. The app aggregates several modes such as biking, transit, taxi and driving, and additional shared mobility services provided locally such as bikesharing, carsharing, ride-hailing. It provides routing options according to the user's parameters such as the cheapest, shortest or most sustainable (measured as CO_2 produced) route. It also allows the user to book those services through the app, and in the future it looks to integrate the payments as well.

Another interesting feature that is included within the app is the ability of creating a profile after enough data is collected. This feature will allow users to create goals and track progress on issues related to fitness, finances and commuting/travelling time. Finally, as per the user's agreement the app collects this location data and shares it with the city planners in an anonymised fashion. The shared data entails trip origins and destinations, and preferred travel mode. The construction of these kind of apps relate closely to the concept of Mobility as a Service (MaaS), currently being pushed in several parts of Europe.



Source: http://appadvice.com/review/go-la

Government agencies here have two roles, regulating how industry handles data, but also how governments themselves handle data. Citizens have sometimes expressed skepticism that publicly-held data is sufficiently protected. Part of the lack of trust for sharing government-held data relies on the lack of mechanisms to assure the private sector, and citizens perhaps even more, that the shared data would be used for that purpose for which it was collected and that purpose solely. Whenever an agreement for sharing data is made, it should be accompanied with a binding agreement which clearly states the purpose of the arrangement and the uses that will be given to the data. It should perhaps also delineate the actions that the government will take to ensure the privacy and protection of the data and make sure that the data is not used for other purposes outside what is stated. While this may seem easy at first sight, in practice it can be quite difficult for agencies and governments to establish partnership agreements around the data than to buy it directly from the private sector. In this sense innovation should not only come from the sharing of the data itself, but on the binding agreements and how these are structured.

In terms of policy and regulation, there are cases where different sectors (e.g. telecommunications) are regulated much more heavily than others (e.g. apps), but in some instances both collect and provide quite similar location-based data content, leading to inequalities in the market and potential regulatory loopholes

allowing or enabling data discovery and potentially breaches. This also leads to a false sense of security and also may give financial burdens to some companies and sectors. In relation to this, there is often also an asymmetry between who wins and who loses in the area of big data.

This might lead to heavy-handed inopportune regulatory interventions which will have additional adverse effects rather than improving the situation. Furthermore, regulations suffer from geographical constraints and often fail to capture multi-national enterprises, particularly in view of platforms and operating systems. Another issue is that though few people have the skill to access data illegally, the results of that access can be incredibly damaging and widespread.

Privacy-by-design is a self-certification approach, where service providers state that they have implemented its principles and thus privacy is being addressed in a systemic way from the beginning in all stages of the service design and delivery, but this might give a false sense of security, as the actual risk level is difficult to judge. Another approach is data minimisation, but the question with big data is, should only minimum data elements be collected or, alternatively, should no constraints be placed on data collection and control exercised only in the processing and use of the data? This approach introduces the issue of data security and privacy breaches arising from the handling of the data, often by third parties.

Data minimisation approaches seem not to be the current trend. On the contrary, the advent of Big Data indicates a general willingness to collect as much data as possible and to allow new, sometimes unforeseen, use cases to develop on the grounds that this innovation provides new efficiencies and opportunities. And it is particularly the innovative use of data collected for often completely different purposed that has the potential to generate real added value. But there can also be the potential of privacy-by-design leading to more innovation, rather than hindering it, as technology might be a more powerful tool to ensure privacy, rather than policies, laws, and regulations.

4. New data sharing models and partnerships

The discussion around models for commercialisation and of using data centred on a number of questions. Is there a need to move beyond the current supplier-client relationship governing public authority access to most privately collected location and mobility data? What models for success are currently being developed or are already in operation, including emerging hybrid-models? What is the public authority mandate for traffic control and management and how is it changing; what are current and potential future approaches for shifting control and management functions from public to private; what are the concerns and what are the potential benefits?

Box 6: **Thinking about minimum standards for public interest data sets**

Do we need to think about setting minimum public interest data sets? Even if the data collection, processing, and analytics are increasingly not done by the public sector, should there nevertheless still be a mandate in this respect? In this context, authorities should consider three broad types of data:

- Type-1 data: includes data relating to traffic flows and use of infrastructure, that public authorities are still mandated to manage, as they also own much of the infrastructure. This data conventionally is collected through sensors embedded in the infrastructure, e.g. loops in the road surface and through the use of probe vehicles.
- Type-2 data: behavioural data on activities and more socio-economic information, which is essential for long-term planning to deliver adequate infrastructure and adequate levels of service. Traditionally this was done through very time and labour consuming detailed travel surveys and official censuses.
- Type-3 data: additional information of interest to decision makers, based on access to novel data sources and analytics, which did not exist before. One example here could be access to accelerometer data from vehicles and mobile devices.

Not only has the loci of data collection started to shift from the public to the private sphere, there is an emerging possibility that control of what occurs on public roads and transport networks may also start to shift away from the public sector as well. This will particularly be the case when envisaging private enterprises operating fleets of automated vehicles in public areas or when the provision of privately sourced and distributed travel and routing information leads to observed changes in traffic volumes and behaviour. Different models for relationships between private and public sector are possible:

- Client-supplier relationship: these already exist, where public authorities are clients to private data providers for access to data sets and/ or data analytics and visualisation.
- Regulatory trades: where the private sector as part of the licensing process for providing services has to give access to data (e.g. taxi services in Beijing, Singapore, and Seoul and app-based platforms in some cities).
- Other models, e.g. co-creation of data, data collection and processing partnerships, etc.

The scope and purpose of data sharing in this context can be limited, unlimited, on a case-by-case basis, or linked to a specific purpose. The skill and capacity of public authorities to manage data received from the private sector is another issue. Tensions also arise in view of open data requirements mandated by the public sector. There is an asymmetry of data access between public and private sector. Therefore, will authorities still be able to have the overview and vision over all relevant urban data for planning purposes, or will (or should) this be devolved to the private sector as well?

Source: Philippe Crist, ITF

Specific issues that were discussed in the context session included:

- Relating to the current relationships between private and public sector regarding the provision and use of data, particularly spatial data, are there any other models being used, other than the ones that were discussed previously?
- What are the implications of these changes in the data environment and how services are delivered? Is there a transition not only in the data, but also in the service layer?
- What should be the response in terms of developing the architecture for appropriate solutions in this context? What should those relationships look like? Should these evolve or should there be active management of the development and creation of these partnerships, in order to bring about desirable outcomes?

Several models for data sharing can already be found across different regions and countries, with various degrees of success. Most of these models are fairly young, given the nature of the data, so it may be too early to draw any conclusion from the information available, or even recommend one over the other. Nevertheless, there is an emerging trend in this area, and it is important to recognise the existence of innovative arrangements so that governments or transport agencies can at least have some elements to assess which could fit their context. The models that are identified below can be divided into: partnerships, which can be either public-private or public-citizen, mandatory schemes and new paradigms.

Public-private data partnerships

Within this category we will present three initiatives worth mentioning as examples of public-private partnerships harnessing this type of data to assist cities: Flow (Sidewalk labs, Google Mobility and U.S. DOT), Open Traffic (World Bank) and Waze (Connected Citizens Program).

Flow: Urban transportation co-ordination platform

A few weeks after the U.S Department of Transportation announced their 2016 Smart City Challenge: Transforming Transportation, Google announced its "new urban innovation firm" Sidewalk Labs as a partner to the initiative. The Challenge invited mid-size cities to demonstrate how data and emerging technologies could be applied to solve problems such as congestion and traffic safety, while protecting the environment and supporting economic vitality. The partnership between the U.S. DOT, participating cities and Sidewalk Labs centres on the development and testing of a data platform called Flow (Sidewalk Labs, 2016). The platform will bring together location-based data from multiple sources and sensors, and in particular, anonymised data generated by smartphones. According to Sidewalk Labs this platform will allow cities to gain a better understanding of citizens' travel patterns and desired destinations, as well as bring solutions so that citizens' access can be delivered more efficiently, equitably and safely.

The platform will have the following capabilities:

- Integrate aggregated, anonymised smartphone data from billions of kilometres of trips (starting with Google's Urban Mobility programme) along with sensor data (via LinkNYC Wi-Fi kiosks -- https://www.link.nyc/) to create a real-time view of road and curb use.
- Select and analyse specific road segments to understand what's driving congestion based on the type of trip being made and the neighborhoods where traffic originates.
- Simulate the impact of new roads, transit routes, mobility services, and incentives on traffic by asking "what if" questions and sharing data across Flow cities.
- Test new technologies like autonomous vehicles by deploying sensors and assessing mobility impacts on the overall system.

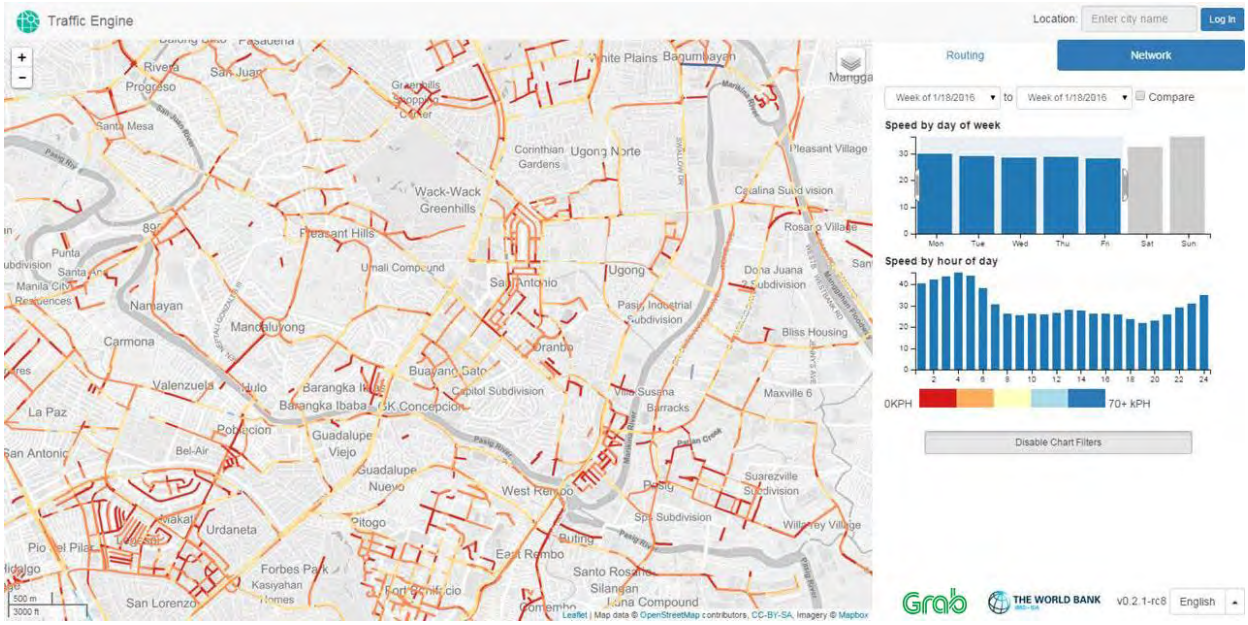
Google's Urban Mobility programme, which is linked to the Flow platform, has also begun sharing aggregated and anonymised data about historical traffic statistics like average speed, relative traffic volumes and traffic flow, with research institutions. These institutions use this data to improve city-wide mobility, or for specific projects, such as tunnel closures for Södra Länken, Sweden's national road. Other applications include helping the Netherlands forego physical road sensors in favour of more cost-effective approaches for collecting data, while retaining the same level of service.

The traffic statistics are derived from aggregate location data which is being collected through Google's smartphones apps when users proactively choose to share this data through their permissions settings. When aggregating the data, Google uses differential privacy algorithms which are based on solid mathematical foundations and which have proven to be more effective than simple techniques such as hashing that have are vulnerable to re-identification efforts.

Open Traffic: Open source, global traffic speed data set

A team from the World Bank, led by Holly Krambeck, recently created Open Traffic, an open source web-based platform that uses real-time and historic GPS location data and transforms it into anonymised traffic speed statistics (World Bank Group, 2015). The aim of Open Traffic is to improve global access to critical transport datasets. By giving governments access to the platform and the statistics generated within it, resource-constrained agencies can make better, evidence-based decisions that previously had been out of their technical and budgetary reach.

Figure 2. **Open Traffic speed profile map based on Grab Taxi real-time and historic data**



The platform's first partnership was with Grab Taxi, a taxi hailing app with more than 250 000 drivers operating in 30 cities in Southeast Asia, as well as the Philippines Department of Transportation and Communications. The goal of the partnership is to use Grab Taxi's drivers' smartphone data (producing GPS location fixes at six-second intervals) to generate traffic speeds, flows and intersection delays for the cities of Cebu and Manila. As part of the partnership the project team will train transportation planners in Cebu and Manila to use the platform, and apply the analysis in three initial applications: peak-hour analysis along

key streets, travel time reliability analysis and vulnerability of specific streets to bad weather and traffic crashes.

On the technical side, the platform applied the concept of "virtual trip lines", previously developed by the Berkeley team on the Mobile Millennium project, for anonymising the data and turning it into speed statistics for individual road segments. Open Traffic also uses OpenStreetMap (OSM) tiles to represent and link the traffic calculations to the road the segments within OSM's road network.

The Connected Citizens Program by Waze: Citizen-government data exchange platform

Waze is another business which has been forging public private partnership with a number of cities around the world through their "Connected Citizens Program" (Waze, 2016). The first of these cities was Rio de Janeiro, and since then it has expanded to almost 16 cities and six states. The partnership has consisted of a mutual data sharing agreement, where Waze gives cities access to its data in real-time. This data enables authorities to identify congestion and bottlenecks based on analysis of users' GPS data and measured by the statistical deviation from the baseline speed for that specific network link. The data also includes user-reported incidents such as crashes, traffic jams, hazards, construction, potholes, stopped vehicles, objects on road and missing signs.

Data about speeds for the whole network are currently not shared. Rather, speed is only shared whenever a road link displays congestion past a certain threshold according to degraded travel speeds. In exchange governments are expected to give Waze data regarding major events that are planned which will result in road closures, such as sport events, construction, and holidays/festivals or VIP visits. This data can be uploaded through a variety of forms, such as a data API feed (formats accepted are JSON, XML or KML) or through the "road closure tool", but the data should at least contain: coordinates, street names, description, and start and end time for the closures. Waze has developed the closure and incident specification (CIFS - http://tinyurl.com/zbj84kd) for importing this type of data from government or third parties sharing data with them.

Waze allows the integration of the data into cities' traffic management centres through an XML or JSON API feed updated every two minutes (which is similar to the CIFS specification), or viewed through the "Traffic View" portal, a web-based interface which shows incidents. Waze does not share historical data with partners however, though they have occasionally worked with some authorities to conduct specific research. Waze also allows opening communication channels between their users and the government (Stern, 2016). This feature started when Superstorm Sandy hit the grand New York area, and the U.S. Federal government turned to Waze for help with the motor fuel refueling plan for the area. They needed to identify which gas stations were still functioning and where fuels were most needed. Waze set up an online form that their users could fill, and in a matter of hours 10 000 responses had been provided.

Public-citizen data partnerships

Another model of data sharing concerns the combination of government-sourced data with crowd-sourced data where members of the public directly share information with authorities or do so indirectly through intermediaries. In order for this to happen though, citizens must perceive a clear benefit to sharing. The example of Xerox's partnership with the cities of LA and Denver mentioned earlier is an example of this kind of citizen-led data sharing since Xerox acts as an intermediary between the app users and public authorities. This intermediation can be seen as a manner with which to build assurance or trust for citizens while allowing authorities to benefit from the private sector's expertise and experience in developing and supporting the business model behind the app. Further, the participation of the private sector intermediary can also deliver marketing benefits that are essential to the app's success - in order for the data generated

by the app to be valuable to the city, it must be built on a large user base, which the commercial partner can deliver through marketing and quality assurance.

The open transport data movement has pushed many public transport agencies to stop developing their own routing platform and apps, and instead focus on their comparative advantage, which is to run their services and open access to their data. This allows app developers to innovate and provide services that compete on quality and utility to users. Overall this has had a good effect, but the downside has been that transport agencies stopped having control of the data generated by those app users. However, this does not mean that authorities cannot reach out to the private sector to seek partnerships to access this data in order to maintain or improve service quality. However, in these instances permission to share user data with public authorities is enacted at the time of app installation and not later. This means that the terms of the data sharing partnership must be fixed in advance and included, whenever possible, within the user agreement.

Mandatory data sharing

Agencies have mandates for planning and managing transport and road networks. As noted previously, in order to carry out that mandate, they have had to collect data mostly through physical road sensors, manual vehicle counts, or any of the other methods described before. Some cities, however, have started to look at which other data sources and collection methods could be used in order to make this process more efficient and cost-effective, and using the authority to compel data sharing.

Authorities have the possibility to compel data sharing on the part of individuals (e.g. data on income) and commercial parties (e.g. data on compliance with established laws).For obvious reasons, however, this power should be limited to those instances where the data is necessary to carry out a public policy mandate. Even in those instances, the scope of data required should be minimised to that which is just necessary to carry out the public policy mandate. Following their rapid and sometimes disruptive deployment, many authorities have sought to compel app-based ride-sourcing platforms to share data relating to their services. These data requests have fallen under general regulatory reporting requirements, specific regulatory reporting requirements (such as airport access) and law enforcement requests.

Uber inventories these requests in the United States in their 2016 Transparency Report (Uber, 2016). Governments may request information about trips, trip requests, pickup and dropoff areas, fares, vehicles, and drivers in their jurisdictions for a set time period. Uber notes that in many instances government agencies request overly broad (in Uber's view) data in relation to the regulatory task at hand. In those instances, Uber reports that they have sought to negotiate with authorities to narrow the scope of the data requested.

Table 1. **Compliance with Regulatory Reporting Requirements, Uber, July-December 2015 (United States)**

General Regulatory Reporting Requirements				
Total Requests	Riders affected	Drivers affected	Compliance	
33	11 644 000	583 000	As required	21.2%
			As required, after negotiating narrower scope	42.4%
			As required, unsuccessfully tried to narrow scope	36.7%
Airport Reporting Requirements				
34	1 645 000	156 000	As required	100%
Law Enforcement Requests				
Rider accounts requested	Driver accounts requested	Percentage of requests where some data was produced	Compliance	
			Fully complied	31.8%
			Partially complied	52.8%
			Withdrawn/ no data found	15.4%

Mutual benefits need to be clear to both parties when establishing mandatory data sharing schemes in return for transport service or operational licensing since there may often be a discrepancy of expectations amongst parties. In addition, agreements would benefit from being bound to general standards and codes of practice. The concepts of public-private data sharing including novel data sources has great potential in less developed countries, which might use these to "leap-frog" ahead to the most promising and efficient forms of data collection. Finally, the utility value of mandating open access to specific public interest data sets seemed to be less apparent in the discussion than was previously expected, but it will be important to pay close attention to the initiatives below, as these can shed light about the future of open data within this context.

Participants also underscored that simply requiring regulated parties to provide data may not be sufficient for authorities to extract useable information from that data. The particular skill sets to understand, format, clean, parse and analyse large, unstructured or differently structured and high velocity data are not typically found in the public sector. Furthermore public authorities will have to compete with the private sector for data scientists, including statisticians, and this competition will be complicated in light of pressure on public budgets – especially in light of the highly remunerative wages offered by the private sector in this area.

New York City Taxi and Limousine Commission: Use of taxi and for-hire vehicle data for transport policy

In 2015 New York City started mandating via its Taxi and Limousine Commission (TLC) that for-hire vehicle companies submit trip log data on a monthly basis. Required data elements include: date of trip, time of trip, pickup location coordinates, driver's for-hire license number, and vehicle's for-hire license number. The data is similar to the one TLC has been collecting for taxis over a number of years now, though taxi data also includes drop-off locations and drop-off timestamps. The city uses the TLC taxi data mostly for planning purposes. For instance, the Department of Transportation (Weeks, Parfenov, & Muthuswamy, 2014) uses the data to produce travel times, origin-destination patterns and measures of economic activity, which feed into a number of their studies. Since only pick-up and drop-off locations are collected (for taxis) and not the complete trip route, algorithms are used to infer the routes taken by taxis in order to estimate network link travel times. Taxis in New York have very high penetration rates and according to researchers (Kamga & Ukkusuri, 2013) this compensates for any potential bias on using taxi data for measuring network link travel times.

The City of São Paulo: Data-driven rules governing for-hire regulation

São Paulo is another city looking to collect data from app-based platforms facilitating for-hire transport activity though this data will be used to calculate a congestion charge for such traffic generated by the platforms (Prefeitura de São Paulo, 2015). The level of the charge will be directly linked to distances driven and location. The justification behind this regulation is that the city is looking to charge vehicles operating in partnership with app-based platforms for the commercial use of public road infrastructure. This approach is roughly analogous (in spirit) to the truck tolling systems commonly found in several European countries. The proposal envisions auctioning kilometres as credits which the app-based platforms would buy in order to carry out their activity, supplemented by a surcharge if the original allotment of credits is exceeded. To operationalise all this and to calculate the price of the kilometres auctioned, the city government plans to collect data from the app-based platforms – in particular 30-second interval GPS data from the smartphones running the apps. The data collected would include: pick-up and drop-off location, timestamps, distance and route of travel, price paid and service evaluation.

The city government also plans to incentivise the platforms' "behavior" by changes in the price of the kilometre in accordance to: service provided outside of the peak hour, service provided in the outskirts of the city, where public transport accessibility is low, vehicles serving more than one occupant (load factor), and vehicles adapted for people with disabilities.

New data sharing paradigms

The concept of public authorities subcontracting the private sector to carry out data collection is not new or limited to the advent of the concept of big data and the use of novel data sources. Private sector involvement in carrying out traditional transport related data collection, including for origin-destination travel surveys, was and still is quite common. But an ever-increasing accumulation of data is taking place within, and sourced by, the private sector. This is probably even more so in the developing world.

Is there also a threat (perceived or real) that the private sector is inadvertently creating 'regulatory capture' which will lead to a future where most traffic operations and control responsibilities are effectively outsourced to those that hold the data? In a not too distant future could it be possible to see navigation services providers, which are already layering traffic information, digital mapping and navigation algorithms over the road infrastructure, to take over the traffic signals? Ultimately, fully automated vehicles will create and use a high-definition and seamless representation of the infrastructure that may surpass in quality that was held by public authorities. There are many companies already working on the different building blocks of these scenarios for the future, particularly in terms of capabilities for real-time and evolutive mapping.

This shift from public to private control is already happening – effective control is being outsourced in some instances in the case of commercial operators managing traffic control centres. These developments could lead to a not completely implausible scenario where high-income neighborhoods pay navigation service providers to tweak their algorithm or install geofences such that traffic is redirected outside of their areas, and diverted to less well-off neighborhoods instead.

Increasingly, phenomena of supplier lock-in are manifesting themselves, where it becomes very difficult to change suppliers and systems due to scale, compatibility or learning effects. But with the increase of suppliers offering services there could also be a fusion of inputs from different sources, where authorities pay only the marginal costs of subscribing to services, alleviating the risk of lock-in. Also, since, the whole process is very much innovation driven, with new providers and new solutions constantly entering the market. In this area, open standards may provide some assurance of cross-compatibility and decrease learning and other transaction costs associated with changing suppliers thus reducing the risks from technology or system lock-in.

Data auditing

There may also be an emerging need for a framework to allow third party auditing of data in terms of its quality as this will ensure public authorities get a fair deal out of the relationships they establish with the private sector. This will also allow them to compare different suppliers in a balanced fashion. This is particularly important for contractual agreements where remuneration is based on service levels, and validity of data therefore needs to be guaranteed.

In 2011 the San Francisco Bay Area Metropolitan Transportation Commission (MTC) began a process to review its own traffic data collection, noticing that increasing budget constraints on the maintenance of their road sensors were having an impact on the quality of the data and that new alternatives would need to be sought (Banner, 2013). This prompted a review of the industry, which found that GPS vehicle probe data would meet their needs and goals. As part of this work MTC had one of its subcontractors review and evaluate traffic data that would be procured from the selected vendor, INRIX. The evaluation showed that the data complied with the standards required by the MTC. The final agreements of the contract called for INRIX to provide anonymised traffic data covering the Bay Area network, guarantee an uptime of more than 99.5%, and licenses for historic and real-time data extending to public agency partners which include: Caltrans, SFCTA, NCTPA, Santa Clara County and the City of San Jose.

In addition to this more traditional data auditing approach, it is also important to evaluate whether using a combination of traditional and novel data sources actually delivers better quality information to public authorities at similar or lower prices. A number of different business models are currently being deployed. But there have to be specific pre-agreed parameters for the cooperation to be beneficial and successful. The issue of data bought by the public sector from the private sector then becoming open data also needs to be managed, as this involves commercial information, thus requiring remuneration or some type of agreement.

Furthermore, in some countries where legal obligations require governments to make available data held by them, as in the United States, there could be a legal mandate to disclose all data purchased or procured by the public sector. This could lead to damaging disclosures of commercially sensitive information. Different countries have differed in whether or not information given by third parties in confidence would fall under this regulation. New York City, which in 2015 started collecting trip data from for-hire vehicle companies, was served with a Freedom of Information Act request (FOI) which obliged the city to grant access to data held by the TLC. The data was anonymised by stripping out the driver's identification number and the vehicle license number.

Open data

Access to open data either in isolation or in combination with other commercial datasets has driven innovation and new business models. Public-domain crowd-sourced mapping data available through OpenStreetMap has been a catalyst to many transport applications. Likewise, the array of existing public transport apps was greatly facilitated by the proliferation of operators opening their data. These developments may or may not spread to other data sources depending on the resolution of privacy concerns and the protection of commercially sensitive information. Nonetheless, despite these concerns, some authorities have taken initial steps to deploy open data services.

San Francisco Bay Area 511 (Transportation services information platform)

The San Francisco Bay Area's 511 transportation services information one-stop shop is operated by the Metropolitan Transportation Commission (MTC) in partnership with the California Highway Patrol and the California Department of Transportation (Caltrans). It has an open data feed which delivers traffic data including traffic speeds and incidents. Traffic speed data comes from INRIX procured data, and data on

incidents are collected through Caltrans' sensors. The traffic data feed that MTC provides to developers currently bundles incident data and traffic speed data together. Even though the traffic data that MTC receives is only licensed to be used by partner agencies, the agency was able to negotiate permission to open that data bundled in aggregate bins of five minutes instead of the one minute data that MTC receives. This represents a significant advance over what other agencies are doing with privately acquired data. While MTC has the financial power and leverage to negotiate these types of arrangements, it isn't at all clear how smaller local governments can overcome asymmetric relations with commercial data providers in order to achieve these types of economies of scale in bargaining power.

The workshop discussions touched on the possibility that the business case for achieving added value could be developed by the private sector not only on aspects of data aggregation but could also include a large share of bespoke data management and analysis tailored to the needs and internal capacity of the agency. Given that the level of in-house capabilities varies within different public agencies, some might only need a dashboard overview of key performance statistics, while others might want access to the raw data feed to carry out their own analytics. Another important aspect is the level of training of in-house staff in terms of procedures for privacy protection to avoid breaches through negligent behaviour. Here, privacy-by-design, data auditing principles and data standards, could be built into the process and the wider data value chain.

An important context for establishing these partnerships is that many countries and cities around the world cannot further extend their physical transport infrastructure in response to rising demands, either for financial or environmental reasons. Innovative uses of novel data sources and creative services, that come along, can have the capacity to increase infrastructure capacity without the need for physically expanding, by better deploying services or distributing demand more efficiently. In a way a better understanding of our transport systems and its demand, which is enabled through good use of this data, can help authorities more efficiently employ underutilised resources and assets.

5. Concluding discussion

At the close of the workshop, the Chair and the Secretariat offered a summing-up of the main points emerging from workshop discussions. Data is a fundamental part of transport providing core value in terms of facilitating transport services. Gaining access to, and sharing this data, will imply trade-offs between benefits and risks but should tend towards allowing safe innovation.

Minimum sets of data and data formats collected and shared that should be open are not easy to identify, but this should be part of an evolving negotiation process. Data collected by the public sector should by default be shared, except when damages that could emerge from sharing are expressly identified and found to outweigh the benefits of sharing.

The fact that data is collected upon public thoroughfares and within public spaces, maintained and managed by public authorities, gives an opening for the latter to negotiate innovative data sharing or data co-creation partnerships with the private sector. This data sharing could, at a minimum, be used in-house by authorities to better manage infrastructure, but in some cases could also be shared more broadly as a way of leveraging new efficiencies.

Current trends suggest large changes in the nature of transport service delivery in the near future. The extent and nature of these changes are unclear and so it is difficult to envisage the right regulatory response to these. Data can and should be part of a more flexible regulatory environment, however, that allows better alignment between rules and outcomes.

Authorities should make clear the real benefits that better data, better access to, and better use of data can lead to – especially with respect to perennial transport challenges including congestion management, incident management, multi-modal and seamless wayfinding, equity and environmental impacts. These benefits must be balanced against protection of privacy and commercial data.

The focus of public action should be, as much as possible, to create the right environment for innovation (rather than regulation) based on public private partnerships and co-operation with academia, allowing the piloting of services and products.

Funding schemes, e.g. seed-money, to help researchers and entrepreneurs get innovative ideas off the ground could be helpful but it is unclear at this stage if public funding (other than funding for research) is better than private capital in getting ideas turned into effective and successful business models. In any case, for publicly-funded projects, some form of outcome-based incentive – including in the funding – should be the norm.

The way data is collected, processed, and stored is likely to fundamentally change in the near-term future from how it is done today. Decision-makers now have the opportunity to influence and shape this development process and should not assume that today's situation will be tomorrow's status quo. New forms of data collection and new data types can help support more flexible and experiment-based regulation.

Bibliography

Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., . . . Agarwal, Y. (2014),"Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging". School of Computer Science, Carnegie Mellon University. http://reports-archive.adm.cs.cmu.edu/anon/isr2014/CMU-ISR-14-116.pdf (accessed 22 April 2016).

Banner, J. (2013), "Traffic Data Collection in the San Francisco Bay Area". Presentation at the MTC Tech Transfer Seminar. http://mtc.ca.gov/sites/default/files/4_AOC_Tech_Transfer_Seminar_Banner_06032013.pdf (accessed 22 April 2016).

BITRE. (2014). "New Data Sources for Transport Workshop". Bureau of Infrastructure, Transport and Regional Economics (BITRE), https://bitre.gov.au/events/2014/new_data_sources.aspx (accessed 22 April 2016).

Dwoskin, E. (2015), "Apps Track Users—Once Every 3 Minutes".*Wall Street Journal*: http://www.wsj.com/articles/apps-track-usersonce-every-3-minutes-1427166955. (accessed 22 April 2016).

GSA. (2015, March). "GNSS Market Report", Issue 4. European Global Navigation Satellite Systems Agency. DOI: 10.2878/251572

Hoh, B., M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, . . . Q. Jacobson (2008), "Virtual trip lines for distributed privacy-preserving traffic monitoring". Proceeding of the 6th international conference on Mobile systems, applications, and services - MobiSys '08. DOI:10.1145/1378600.1378604.

ITF/OECD (2015). *Big Data and Transport: Understanding and assessing options.* Corporate Partnership Board. http://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf.

Kamga, C., & S. Ukkusuri (2013), "The Use of Large-scale Datasets for Understanding Network State". Final Report. (U. T.-R. 2, Ed.) New York City: The City College of New York/CUNY.

MTC (2015), "An open format for publishing your road event data". Open511: http://www.open511.org/ (accessed 22 April 2016).

Prefeitura de São Paulo (2015), "Regulação do uso intensivo do viário urbano". Consulta Pública – Proposta de Decreto Municipal 29 de dezembro de 2015. http://www.capital.sp.gov.br/static/2015/12/A6cbf1rCqjYF6VZWaAjihg.pdf (accessed 22 April 2016).

Sidewalk Labs. (2016). "Flow: The Transportation Coordination Platform for Cities". Sidewalk Labs: http://www.sidewalklabs.com/flow/index.html (accessed 22 April 2016).

Stern, N. (2016). "Waze's Drive Towards Successful Public Partnerships". (H. K. Innovation, Editor), Data-Smart City Solutions, http://datasmart.ash.harvard.edu/news/article/wazes-drive-towards-successful-public-partnerships-786 (accessed 22 April 2016).

Uber. (2016). "Transparency Report". Uber Technologies, https://transparencyreport.uber.com/ (accessed 22 April 2016).

UC Berkeley. (2011). "Mobile Millenium". Traffic: University of California, Berkeley: http://traffic.berkeley.edu/ (accessed 22 April 2016).

Waze. (2016). "Connecting Citizens and Governments through Data". Waze Connected Citizens: https://www.waze.com/ccp (accessed 22 April 2016).

Weeks, A., S. Parfenov, & S. Muthuswamy (2014) "NYCDOT's Experience with Big Data and use in Transportation Projects". http://www.utrc2.org/sites/default/files/Andrew%20Weeks%20and%20Stanislav%20Parfenov.pdf (accessed 22 April 2016).

World Bank Group. (2015). "opentraffic". opentraffic, http://opentraffic.io/ (accessed 22 April 2016).

International Transport Forum

Data-Driven Transport Policy

Data are essential to the planning, delivery and management of mobility services and transport infrastructure. These data are being generated in new ways, e.g. through sensors, and much of it contains location information. This report examines the privacy, trust and security issues created by the omnipresence of location-specific data. It also explores new approaches for collaboration between the private and public sector to access, share or co-create relevant data to help manage transport operation and planning. The report draws on a workshop involving industry leaders, policy makers and academics.

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
F-75775 Paris Cedex 16
T +33 (0)1 45 24 97 10
F +33 (0)1 45 24 13 22
Email: contact@itf-oecd.org
Web: www.itf-oecd.org