Mobility Data: Changes and Opportunities

Participating Partners:
Kapsch TrafficCom, PTV Group,

External Expert:
Carlo Ratti Associati

27 October 2014
### Mobility Data: Changes and Opportunities

#### What we did

- **1 byte**: ~1,000 (10^3) bytes

#### What we found

- **Kilobyte**: ~1000 (10^3) bytes
- **Megabyte**: ~1,000,000 (10^6) bytes
- **Gigabyte**: ~1,000,000,000 (10^9) bytes
- **Terabyte**: ~1,000,000,000,000 (10^12) bytes
- **Petabyte**: ~1,000,000,000,000,000 (10^15) bytes
- **Exabyte**: ~1,000,000,000,000,000,000 (10^18) bytes
- **Zettabyte**: ~1,000,000,000,000,000,000,000 (10^21) bytes
- **Yottabyte**: ~1,000,000,000,000,000,000,000,000,000,000 (10^24) bytes

#### Why

- **Several petabytes**: traffic stored by INRIX to produce traffic analysis for e.g. Google
- **~30 terabytes**: data produced by a Boeing 777 on a transatlantic trip
- **~140 gigabytes**: data gathered per day, Nokia HERE mapping car
- **~60 gigabytes**: data gathered per hour, Google self driving prototype car
- **~25 gigabytes**: data analysed per hour by Ford Fusion Energi plug-in hybrid
- **~25 gigabytes**: data analysed per hour by Ford Fusion Energi plug-in hybrid
- **~60 gigabytes**: data gathered per hour, Google self driving prototype car
- **~140 gigabytes**: data gathered per day, Nokia HERE mapping car
- **~30 terabytes**: data produced by a Boeing 777 on a transatlantic trip
- **~1 terabyte**: total amount of visual content processed by the human race in 2013
- **4.4 zetabytes**: estimated size of the digital universe in 2013
- **44 zetabytes**: estimated size of the digital universe in 2020
- **Several petabytes**: traffic stored by INRIX to produce traffic analysis for e.g. Google
Never before has so much information about so many people, events, and objects been available at such speed, to so many people. One recent estimate puts the total size of the “digital universe” – comprising digital content from photographs, movies and surveillance video feeds, data produced and sent by sensors and connected devices, internet content, email, SMS, audio streams to phone call meta-data – at 4.4 zettabytes in 2013.

Doubling every two months, the size of the digital universe is projected to grow to 44 zettabytes in 2020. These estimates represent a staggering amount of data, a significant portion of which relates to events and people.

There are tremendous upsides from the use and fusion of these data streams. There are also associated risks since much of this data is highly personal in nature. This study looks at the considerations authorities should have regarding the creation, processing, conditions of use and access to data to help them carry out their mandates.

We undertook this study on the basis of meetings and discussions amongst project partners, desktop research and invited the contribution of an external expert group – Carlo Ratti Associati, in conjunction with the MIT SENSEable City Laboratory.
Data Analysis Pipeline

- Acquisition
- Recording
- Extracting
- Cleaning
- Annotation
- Storage
- Integration
- Aggregation
- Representation
- Visualisation
- Analysis
- Modelling
- Interpretation
- Reinterpretation
- Deletion

Data/analysis issues:
- Heterogeneity
- Volatility
- Scale
- Velocity, timeliness
- Traceability, privacy

Value Representativeness
Human input Interpretation
Value
Growing sensor ubiquity and steeply declining data storage costs give rise to big data. New tools including sophisticated algorithms, distributed computing capacity and techniques that allow the efficient processing of very large datasets have fuelled the ability to extract meaningful insights from big data.

Extracting insight from big data is not trivial. The big data pipeline spans from the collection of data, its preparation for storage and analysis, its integration with other sources, modelling, visualisation and analysis to its interpretation, archiving or deletion.

Data can be purposely sensed – for example where data is collected from a customised sensor network for a pre-defined use. Such data is channelled to a central control mechanism where it is aggregated and managed; extracted information used as basis for how the monitored terrain should be regulated and appropriate action is taken.

Data can also be opportunistically sensed – for example where data is collected for one purpose and used for another. This approach to data collection is made possible largely by the fact that mobile phones have become so ubiquitous – citizens replace the need for purpose-built sensors, contributing real-time data through their smart devices.

Increasingly, data can be crowd-sourced by “citizen-sensors” reporting on their individual experience via content-sharing platforms.

In all cases, big data may be “born digital” or “born analogue”. Born digital data is created by users or by a computing device specifically for use in a machine processing environment, whereas “born analogue” concerns data that arises from an imprint of a physical phenomenon upon a sensing device and its subsequent conversion into a digital signal.
Mobility Data: Changes and Opportunities
The fusion of purposely-sensed, opportunistically-sensed and crowd-sourced data generates new knowledge regarding transport activity and flows. It also creates unique privacy risks.

More diverse datasets offer more numerous, accurate and insightful results than smaller, less diverse ones. When combined, these data reveal hitherto unsuspected or un-observed patterns in our daily lives which can be used to benefit both individuals and society. There is also the risk that insights derived from these patterns may open new avenues for misuse and potential manipulation of individuals and their behaviour. The knowledge derived from this fusion may not have been anticipated by data collectors at the time of collection, nor may the use of these insights have been anticipated or communicated to people who are the object of that data.
Big data has not done away with the need for statistical rigour since big data is not only prone to many of the same errors and biases in smaller datasets, it also creates new ones
Massive and near real-time datasets are so large that they may seem to mimic reality. Some maintain that just knowing that there is a link between observable variables and an outcome is sufficient to predict the frequency of that outcome in the future. All that is needed is an algorithm that consistently detects patterns in the data. This view of big data analytics would do away with the need for explanatory theories and models. This viewpoint is simplistic and overlooks the difficulty deriving robust predictive results from big data over the long term.

There are a lot of small data problems that occur in big data and a theory-free approach to big data analysis does not make these go away – it makes them worse. Big data has not freed analysts and policy-makers from the need for statistical rigour since big data is not only prone to many of the same errors and biases in smaller datasets, it also creates new ones.

"Letting the data speak" may be effective in providing new insights but it does not explain which correlations are meaningful or predictive. Even if a correlation may prove to be robust over a given period, it alone cannot provide insight into what might cause the correlation to break down – nor what pattern may emerge in its place. And even big datasets still may fail traditional statistical tests – especially those of sample bias and sample error. Understanding the provenance of data is essential for authorities to gauge its fitness for purpose and this can be helped by maintaining persistent metadata fields.
**129 vehicle-related data elements**

in the European Telecommunications Standards Institute’s (ETSI) common data dictionary for intelligent transport systems. If shared amongst vehicles, infrastructure and services, many could improve safety but some could compromise privacy.
Traffic operations, transport planning and safety are areas where authorities must critically evaluate where and how new, or newly available data and data-related insights, can improve policy.

Transport is an activity that at its heart is about connecting locations with flows. Locations may be proximate, well-connected and displaying high levels of access – as in many urban areas – or not. Flows may concern people or goods and may involve any number of vehicle types – or not, as in the case of walking.

Resolving the location-flow equation requires delivering, and managing the use of, infrastructure – but it may also involve decisions regarding where to site activities so that the need to move is minimised.

All of these decisions require information – a lot of information – regarding places, people and activities. "Big data" holds much promise for improving the planning and management of transport activity by radically increasing the amount or near-real-time availability of mobility-related data.

Likewise, access to more detailed and actionable data regarding the status of vehicles and the environment in which they operate holds much promise for improving the safety of transport. Efforts are underway to develop datasets that, if shared, could improve road safety. If uncoordinated, these requirements may increase compliance costs for manufacturers.

Overall, the collection of mobility-related data is migrating from the public to the private sector. Some efficiency improvements may result but long-term considerations regarding how authorities access the data that informs their decisions must be addressed.
**Mobility Data:** Changes and Opportunities

**What we did**
- Mobility Data

**What we found**
- Mobility Data: Changes and Opportunities

**Why**
- Mobility Data: Changes and Opportunities

**What we did**
- Mobility Data

**Location Sensing Technologies and Precision**

- MAC address (WiFi)
- Automatic image recognition (video)
- Facial recognition/tracking
- GPS (GNSS)
- A-GPS (GPS+Cell tower)
- Hybrid GPS (GPS+WiFi)
- Cell tower triangulation
- Mobile telecom cell (tower)

- <1m
- 5-10m
- 5-50m
- 100-300m
- 100m to kms
Locating and tracking individuals at precisions up to a few centimetres in both outdoor and indoor environments is currently feasible and will likely become standard – at least in urban areas – as current location-sensing technologies become omnipresent. The widespread penetration of mobile (especially smartphone) technology makes this possible. The same location technologies deployed in the current generation of mobile telephones is also migrating to vehicles enabling precise and persistent tracking.

The proliferation of devices able to use global navigation satellite systems signals (GNSS, of which GPS is one) to locate themselves has been fundamental to the growth in location data. Cellular signal processing techniques have augmented the precision and rapidity of location fixes allowing for many new location-based services to emerge.

New location techniques are gaining popularity and further enhance tracking precision. Exploiting real-time and either crowd-sourced or opportunistically sensed WiFi networks is perhaps the most important of these. Because of the structure of the data signals involved, this technique also allows tracking of prior locations visited by individuals. Increasingly, exploitation of video feeds or audio signals will allow even more ubiquitous and precise tracking capability.

The exploitation of sensor data – e.g. from accelerometers on smartphones – can also reveal how the smartphone owner is moving. This capability is already reliable and will likely become more so as new devices combine sensors with on-chip processors.
### Data Use and Privacy: New Perspectives

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

World Economic Forum, 2013
Data protection policies are lagging behind new modes of data collection and uses – this is especially true for location data.

Rules governing the collection and use of personal data – e.g. data that cannot be de-identified – are outdated in two ways. Data is now collected in ways that were not anticipated by regulation and authorities have not accounted for the new knowledge that emerges from data fusion.

A split has emerged among those who would seek to retain prior notification and consent frameworks for data collection and those who would abandon these in order to focus on specifying allowable uses of that data only. Among the former are those who seek to modify and strengthen rules based on the principles outlined in the 1980 OECD Personal Data Protection Framework and Guidelines. Among the latter are many in the private sector and some national administrations who suggest these rules are to a large extent unworkable in their current form.

A new focus on regulating use rather than notification and consent presupposes the presence of a well-funded and competent regulatory authority to oversee data uses and to resolve, and possibly prosecute, conflicts. Furthermore, such an agency should be equipped to address asymmetric and extended legal struggles with large and powerful multi-national corporations. These may be overly optimistic expectations.
'Privacy by design’ calls for data collection systems and practices to be designed (or re-designed) from the ground up in order to include strong and irreversible pro-privacy measures. Privacy by design in the context of big data rests on seven features:

**Full attribution:** Every observation or record should be traceable to its point and time of creation.

**Data tethering:** Additions, modifications and deletions in data systems should be accounted for in real time and should be persistently linked to the data.

**Analytics of anonymised data:** Anonymising or encrypting personally identifiable data as well as quasi-identifiers is an essential step in ensuring that big data analytics do not encroach on individuals’ privacy rights.

**Tamper-Resistant Audit Logs:** Every material interaction upon a dataset containing personal information, identifiers or quasi-identifiers should be logged in a tamper-resistant manner by default and by design.

**False Negative Favouring:** When personal data entailing civil liberties are concerned, it is better to design privacy protection policies that favour false negatives rather than false positives.

**Self-Correcting False Positives:** With every new data point created or presented, prior assertions based on that data should be re-evaluated to ensure they remain correct and, if not, amended.

**Information Transfer Accounting:** Every onward transfer of data either to human observation or to machine systems should be logged to allow stakeholders (data controllers or data objects) to understand how their data is flowing and is being used.
Cybersecurity refers to a set of technologies and practices that define, and attempt to enforce, policies relating to computer use and electronic communications. In particular, cybersecurity seeks to deliver the following:

**Identity and authentication**: Are you who you say you are?

**Authorisation**: What are you allowed to do to which part of the system?

**Availability**: Can attackers interfere with authorised functions?

**Confidentiality**: Can data communications be passively copied by someone not authorised to do so?

**Integrity**: Can data or communications be actively manipulated by someone not authorised to do so?

**Non-repudiation, auditability**: Can actions later be shown to have actually occurred?

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

Both cybersecurity and data privacy focus on potential damage that can be caused via malevolent or dangerous manipulation of computer and communication systems. Poor cybersecurity practices may lead to exposure, gathering and malicious use of personal data. However, privacy risks remain even in fully secured systems. Misuse of personal data in otherwise secure systems by authorised operators represents a violation of privacy policy, not of security policy. Similarly, violations of privacy may emerge as a result of data fusion across multiple, fully secured, systems. These distinctions are important since it is not enough to focus solely on cyber-security in order to ensure personal data protection.
People’s patterns of movement in space and time are repetitive and predictable. These trajectories are powerful identifiers – like fingerprints.
Location data has fuelled growth in the connected economy but citizens are wary of divulging too much information about their whereabouts and their daily behaviour. Surveys indicate that people want to have specific knowledge about when and why applications collect location data and that they are increasingly concerned about the scope of location data collection. At the same time, many individuals will allow access to their location data in return for access to services they value.

Location data is rarely linked to a unique individual – what is being tracked is a sensor platform. However, many of these devices (especially mobile handsets and car-based navigation systems) are intimately linked to one person. Geo-referenced and time-stamped data, when linked to publicly available personal and business registries or other sources, exposes a daily pattern of activity that includes where a person sleeps, where they work and other places they frequent.

These patterns of daily activity are extremely repetitive and predictable. As such they are powerful quasi-identifiers – like fingerprints - but they reveal much more. They can disclose a person’s religion (repeated visits to a place of worship), their political affiliation (visits to political or NGO offices and co-location with demonstrations) and other information that can be inferred from where they go and where they spend time. This data can, in conjunction with similar data on other people, reveal the network of friends, acquaintances or colleagues a person has – especially when cross-referenced with volunteered data social networking sites. This data can also reveal important and potentially significant pattern breaks that can compromise privacy (e.g. visits to an obstetrician’s or an oncologist’s practice).
Research on the privacy bounds of location data has resulted in a number of high-profile re-identification cases that have successfully isolated individual mobility traces from relatively coarse location data (e.g. at the scale of several hundreds of metres) and resulted in re-identification rates ranging from 35% to 88%.

The difficulty with which trajectory data can be adequately and persistently protected has led some to question whether it is worth the effort to do so. On the other end of the spectrum, ‘Privacy by Design’ advocates have stated that the risk of re-identification has been largely overstated and that few real cases of linking individual mobility traces to unique names, addresses, or other personal identifying data have been undertaken. There is truth in both arguments.

With time, the sophistication of de-identification algorithms is likely to grow as is the availability of other sources of information that could compromise the anonymity of location and trajectory data. De-identification efforts are not fully “trivial”, they require algorithmic sophistication, time to “clean” data errors, access to reliable data on personal identifiers (home address, id number, etc.). What they require most of all, though, is sufficient motivation to overcome these hurdles.

At the same time, location-based and trajectory data are difficult to fully and permanently de-identify. Protecting the anonymity of high dimensional data like space-time trajectories is more complicated than anonymising low-dimensional data such as addresses, names or employer.
Multiple de-identification techniques ranging from simple masking techniques to cryptographic protection exist. In the case of high-dimensional location or trajectory-based data, there is a compelling argument to be made for using the most robust of these techniques and even seeking additional data protection methods.

Simple anonymisation or pseudonymisation is only possible when geospatial data are associated with straightforward data identifiers (such as name or address) that can be redacted or given pseudonyms. Neither of these approaches should be considered an adequate basis for de-identification, especially in the context of associated geospatial coordinates.

Efforts to anonymise location and trajectory data by clustering points or traces into larger groupings of like data can improve anonymity. Numerous researchers have studied generalisation-based approaches but there is a risk of re-identification even from aggregate data if the aggregation ignores particular characteristics of the dataset. Obfuscating data by adding spurious data elements or records can help but sufficiently large sets of location-based or trajectory data can overcome the impact of such data “noise”.

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

Protecting location data... opportunities exist but require new approaches
Policy insights

Road safety improvements can be accelerated through the specification and harmonisation of a limited set of safety-related vehicle data elements
E-call, E-911 and vehicle data black boxes provide post-crash data best suited for emergency services and forensic investigation. Much more vehicle-related data is available and, if shared in a common format, could prevent crashes. Further work is needed to identify a core set of safety-related data elements to be publicly shared and to ensure the encryption protocols necessary to secure data elements that could otherwise compromise privacy.

Transport authorities will need to audit the data they use in order to understand what it says and does not say and how it can best be used
Big data in transport is not immune from small data problems – especially those relating to statistical validity, bias and incorrectly imputed causality. Transport authorities will need to ensure an adequate level of data literacy for handling new streams of data and novel data types. Ensuring robust, persistent and harmonised provenance metadata will facilitate data usability audits. Big data is often not clean – issues with data quality may entail significant upfront costs to render it usable and this should be considered early in the decision-making processes.

Data visualisation will play an increasingly important role in policy dialogue
Effective data visualisation can quickly communicate key aspects of data analysis and reveal new patterns to decision-makers and the public. Public agencies will have to be able to handle the visual language of data as effectively as they handle written and spreadsheet-based analysis.
Policy insights

More effective protection of location data will have to be designed upfront into technologies, algorithms and processes. Adapting data protection frameworks to increasingly pervasive and precise location data is difficult largely because data privacy has not served as a design element from the outset. Both voluntary and regulatory initiatives should employ a ‘Privacy by Design’ approach which ensures that strong data protection and controls are front-loaded into data collection processes. Technological advances including the arrival of ‘system on a chip’ sensors can aid by allowing on-the-fly data encryption. Other advances could include protocols allowing for citizens to control and allocate rights regarding their data.

Failing to ensure strong privacy protection may result in a regulatory backlash against the collection and processing of location data. This would hamper innovation, reduce consumer welfare and curb the economic benefits the use of such data delivers.

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

An increasing amount of the actionable data pertaining to road safety, traffic management and travel behaviour is held by the private sector. On the other hand, public authorities are still – and will likely continue to be – mandated to provide essential services that the public prefers private actors not to provide.

Should data access by public authorities continue to be modelled strictly on a supplier-client relationship or can new, more creative partnerships be developed that enable both the private sector to innovate and the public sector to carry out its mandates? Innovative data-sharing partnerships will likely develop though these should not obviate the need for market power tests, benefit-cost assessment and public utility objectives.
ABOUT THE INTERNATIONAL TRANSPORT FORUM

The International Transport Forum at the OECD is an intergovernmental organisation with 54 member countries. It acts as a strategic think tank with the objective of helping shape the transport policy agenda on a global level and ensuring that it contributes to economic growth, environmental protection, social inclusion and the preservation of human life and well-being. Until 2007 the organisation was known as the European Conference of Ministers of Transport (ECMT), then its geographic reach was widened and it became the International Transport Forum.

The International Transport Forum manages the Multilateral Licences for international road haulage on the European continent.

The transport policy related work of the International Transport Forum rests on three pillars

- Annual Summit
- Research Centre
- Corporate Partnership Board

CORPORATE PARTNERSHIP BOARD (CPB)

The CPB is a global network of companies from across all transport modes and closely related areas like energy, finance, IT, who understand the opportunities and challenges to transport and want to work with the ITF to improve policy analysis and advice by adding a corporate perspective to the process. The CPB provides a unique avenue for participating in the debate on the challenges and trends facing global transport, and bringing issues important to businesses to the attention of policy makers, key transport stakeholders in ministries, the business community, and international organisations.

The work started in early 2014 and there are currently four projects underway:

- Autonomous Driving: Regulatory Issues
- Urban Mobility: System Upgrade
- Mobility Data: Changes and Opportunities
- Drivers of Logistics Performance: Case Study

This is a background document, with the final report due January 2015.

Contact:
Philippe Crist
T +33 (0)1 45 24 94 47
E philippe.crist@oecd.org
www.internationaltransportforum.org