Urban Design and Planning
Volume 166 Issue DP1

Data dimension: accessing urban data and making it accessible
Nabian, Offenhuber, Vanky and Ratti

ice | **proceedings**

Proceedings of the Institution of Civil Engineers
Urban Design and Planning 166 February 2013 Issue DP1
Pages 60–75 http://dx.doi.org/10.1680/udap.12.00011
Paper 1200011
Received 01/03/2012 Accepted 16/07/2012
Keywords: information technology/knowledge management/social impact

ice Institution of Civil Engineers publishing

Data dimension: accessing urban data and making it accessible

1 **Nashid Nabian** MArch, MUDES, DDES
Lecturer, Harvard Graduate School of Design, Cambridge, MA, USA; Research Affiliate, MIT, SENSEable City Laboratory, Cambridge, MA, USA

2 **Dietmar Offenhuber** MSc
Research Fellow, MIT, SENSEable City Laboratory, Cambridge, MA, USA

3 **Anthony Vanky** MSc, MA
PhD Student, MIT, SENSEable City Laboratory, Cambridge, MA, USA

4 **Carlo Ratti** MPhil, PhD
Director, MIT, SENSEable City Laboratory, Cambridge, MA, USA




New technologies allow for new ways to sense the city. Thinking of urban data as substance, this paper criticises a certain approach in dealing with urban-related data analysis when it comes to identifying certain patterns and deriving narratives based on these patterns. In this approach, coined here as the 'big data' approach, having access to large-volume datasets is considered sufficient to study a phenomena and its very dynamics that the data refers to. In contrast to this, this paper examines ways in which data can be produced, modified and delivered; in any of these steps, there is an ongoing cost–benefit analysis, based on which a series of necessary decisions in terms of resolution and quality of data through the lens of filtering has to be made with the goal of accessing data and making it accessible most effectively. Projects from MIT SENSEable City Lab are used to better illustrate these ideas.

1. Introduction

New technologies are allowing new ways to 'sense the city'. For instance, systems already in place that were developed for other reasons but can also function as a source of information on how our cities operate can be utilised. The premise of such sensing practices is that contemporary subjects involuntarily leave digital traces on various networks that are juxtaposed over urban areas. Every time a credit card is used, a text message or an email is sent, an internet search query is submitted, a phone call is made or a purchase is processed on a major online store, an entry with the time and location of this action is added to a dataset on a central server, administered and maintained by the organisational entity providing the platform for these day-to-day operations. This data can be used to make sense of urban dynamics and the flow of materials, capital, information and human resources within a city (Calabrese *et al.*, 2011). As a brief aside, in this article, the term data and its derivatives, such as datasets, are used to refer to raw or filtered and formatted material in digital format. A frequent issue is whether the word 'data' should be used as a singular or plural word. Although the word is plural, it is now common to treat it as a mass noun in singular form.

Apart from tapping into existing networks, customised sensor networks can also be harnessed to decode various flows within cities. The metaphor of 'smart dust' refers to small autonomous sensors dispersed in the environment for the purpose of data collection, forming emergent communication networks that are in many ways different from traditional communication infrastructures (Boustani *et al.*, 2011). Instead of a top-down approach such as implementation of sensor networks, this paper will also consider more grassroots, bottom-up systems for sensing the dynamics of cities. One approach to this is thinking of each urbanite as a human sensor – an agent for sensing and reporting on his or her individual experience through tapping into data generated by user-contributed content on content-sharing platforms (Girardin *et al.*, 2008).

This raises the question, however, of why is it even important to be able to sense the dynamics of a city. Being able to sense a city allows for consideration of the application of operating mechanisms at an urban scale, which actuate the city based on what is sensed about its dynamics in real time. Embedded actuators within the constituting elements of a city and its physical infrastructure are the components of the space that can be controlled by the output of the operating system, based

on changes registered by the sensing mechanism and reported to the operating system as an input.

However, manipulating space through embedded actuators is not the only possible means of spatially regulating urban systems. The inhabitants of cities can be considered possible agents of regulation and actuation: once the datasets generated through different sensing mechanisms are spatially and temporally attached to entities and phenomena in a physical terrain, visualisation of urban data can be cross-referenced with the geographical terrain to allow revelation of urban dynamics in real time, thus democratising access to helpful urban information that offers people more control over their environment by 'allowing them to make decisions that are more informed about their surroundings, reducing the inefficiencies of present day urban systems' (Calabrese *et al.*, 2011).

In this paper, thinking of urban data as substance, the authors would like to criticise an approach in dealing with urban-related data analysis when it comes to identifying certain patterns and advancing narratives based on these patterns. In this approach – coined in this paper the 'big data' approach – having access to large-volume datasets is considered sufficient to study a phenomena and its dynamics that the data refers to.

In contrast to this approach, here the authors examine ways in which data can be produced, modified and delivered, while noting that in any of these steps there is an ongoing cost–benefit analysis based on which a series of decisions (in terms of resolution and quality of data through the lens of filtering) has to be made with the goal of accessing data and making it accessible most effectively. In doing this and to better illustrate these ideas, projects from MIT SENSEable City Lab (http://senseable.mit.edu/) are considered.

2. Classification of approaches to data analysis based on data access or data generation

In terms of generating new data or accessing already existing data, let us start with a simple categorisation that is less concerned with the data structure, but more with the means of its initial generation or the nature of its primary source and effects. Much of the current research under the rubric of big data analytics is largely uncritical regarding the origin of urban datasets. The data is simply taken as a given and the assumptions, processes and transformations embedded in these datasets are not part of the analysis. However, datasets do not appear out of nowhere and the conditions of their generation need to be accounted for. Three different cases of data acquisition are thus distinguished.

- The first group comprises datasets that come from a single system. They are usually by-products of large information infrastructures such as wireless telecommunication systems. The datasets are uniform, follow a consistent logic and reflect the properties of the system that generated them. To analyse them means to take the generating system as a proxy for the phenomenon of interest.
- Datasets from the second group are less consistent and are generated through aggregation of data by multiple authors, generated for many different purposes. Examples are aggregated datasets from social media platforms. In this case, analysis involves appropriation and the extraction of salient features that are consistent over the dataset.
- The third category concerns datasets generated from scratch, probably because no prior data existed about the specific phenomenon of interest and there was no existing system with an informatic by-product that could be leveraged to extract information about the topic under investigation.

The following subsections describe some of the MIT SENSEable City Lab projects that each epitomises one of the above categories.

2.1 Borderline and Connected States of America: accessing data from a single system

The near saturation of wireless communication systems around the world has had a profound impact on global society. People all over the planet can communicate more easily with individuals in all corners of the world, unencumbered by the often-expensive process of installing wires as was previously the case. This pervasiveness of communication has the added benefit of connecting more people and thus provides more detail in communication patterns at the individual level in their communities, regions and around the world. This process creates call detail records (CDRs) that identify the caller and receiver, and spatio-temporal information based on the location of the cellular towers carrying the phones' signals and the time when the call was made.

The data is often from a major telecom provider in the country, which, due to its operational value, can be considered clean and precise. In creating spatial narratives based on data analysis performed on such datasets, involving only one system, the narratives from the research are often self-referential. The burden of being novel must be from within the data itself. In this case, it is a compelling resource for understanding human connections.

By analysing CDRs provided by partnering telecommunications companies, the authors were able to analyse connections between different individuals or aggregate them to see patterns across regions. In research into Great Britain and the USA, a new, more fine-grained approach to regional delineation was used, based on analysing networks of billions of individual

human transactions. This process allows one to measure whether regional boundaries defined by historic and geographic divisions between people respect the more natural ways that people interact across space in contemporary society. The research looked at the human network as a topological entity with no geographical constraints, trying to illustrate emerging communities based on social interactions between their members, mediated via telecommunication networks.

The Borderline project (Figures 1 and 2) analysed twelve billion calls over a one-month period, aggregated into a grid of 3042 square pixels across the map of Britain, each with dimensions of 9·5 km by 9·5 km. Each pixel was treated as a spatial node and its connection strength was measured (calculated using total call time, hence taking into account the local population density) and correlated to every other pixel, thereby deriving a matrix of communications traffic between pairs of pixels in the geographic network. An optimisation algorithm was applied to calculate the strength of these nodal connections and it was found that patterns deviated from traditional boundaries (Ratti *et al.*, 2010).

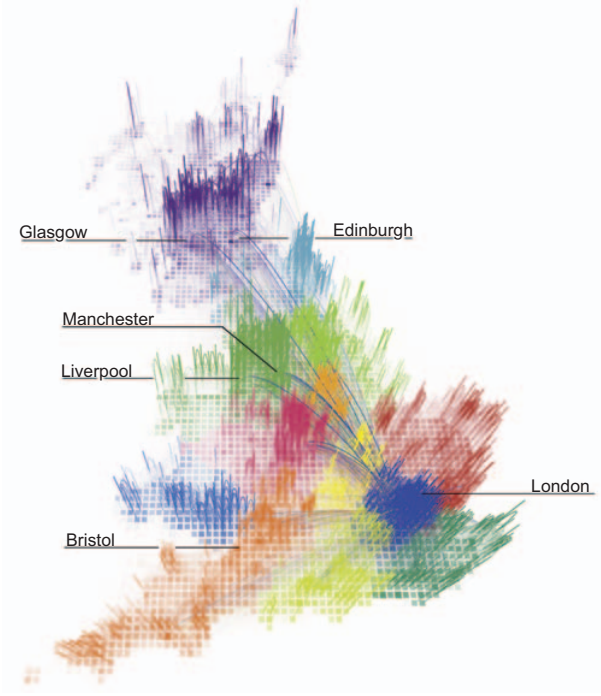

Figure 1. The geography of talk in Great Britain. This figure shows the strongest 80% of links, as measured by total talk time, between areas within Britain. The opacity of each link is proportional to the total call time between two areas and the different colours represent regions identified using network modularity optimisation analysis (©SENSEable City Lab)

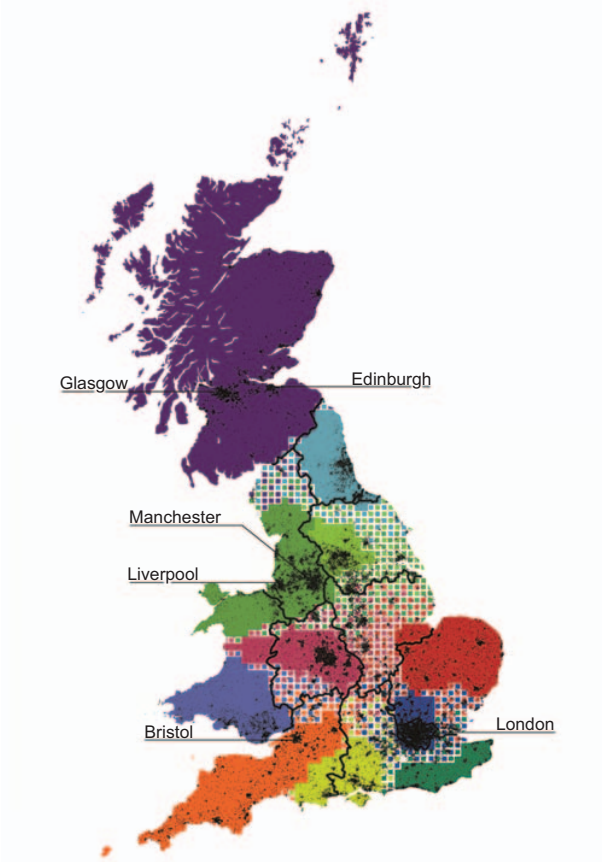

Figure 2. The core regions of Britain, obtained by combining the output from several modularity optimisation methods. The thick black boundary lines show the official government office regions partitioning, together with Scotland and Wales. The black background spots show Britain's towns and cities, some of which are highlighted with a label (©SENSEable City Lab)

Looking at data filtered and massaged as elaborated above, human interaction could be captured more accurately than the official statistical regions (Nomenclature of Territorial Units for Statistics (NUTS)). In some cases, some of the regions created from this analysis – those corresponding to Scotland, South West, London and the East of England – closely match the forms of historically and administratively important regions. In others, such as Wales, the region seems to have been incorporated into areas dominated by the major cities of the West of Britain and its new-found boundaries based on mapping human connections deviate from long-held borders of the official statistical regions.

Comparatively, Scotland appeared to be loosely coupled with the rest of Great Britain in a way that Wales emphatically is

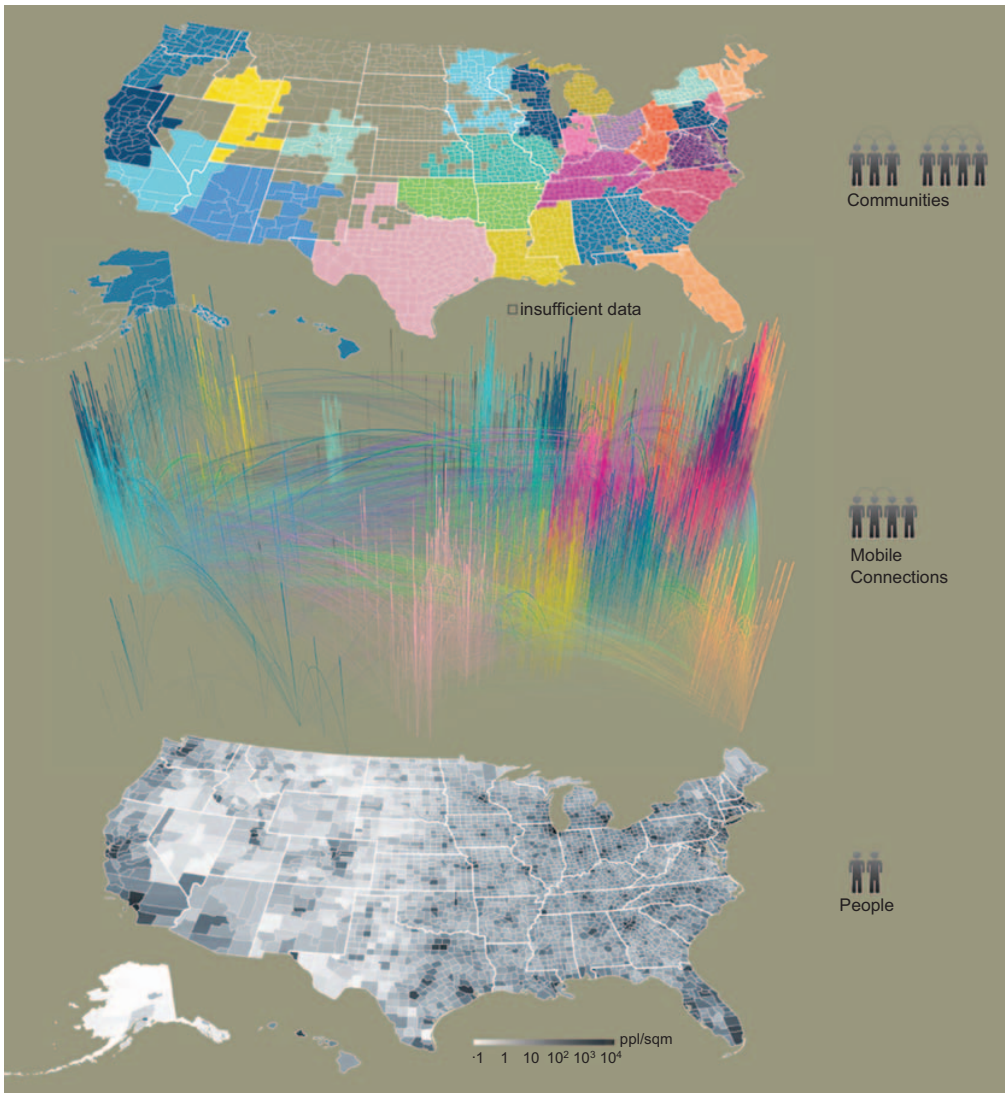

Figure 3. Visualisation of Connected States of America project (©SENSEable City Lab)

not. In other terms, if Scotland and Wales were to become independent from the UK and if the detrimental effect of the secession were considered proportional to the number of external connections, the effect on people would be approximately twice more disruptive in Wales than in Scotland.

Following the same logic of operation, in the Connected States of America (Figure 3) – a similar project undertaken using records from the USA and partitioned county-by-county – the data analysis revealed interesting relationships between and within the familiar map of distinct states. For instance, in Cincinnati, Ohio's neighbouring communities in the neighbouring state of Kentucky related more with Ohioans than fellow Kentuckians. In Chattanooga, Tennessee's residents connected more with those in Georgia, possibly due to its proximity to the economically larger metropolitan area of Atlanta, Georgia.

We can also speculate on the struggles of identity in the Union within and beyond existing state borders. Californians have considered at many times in the state's history dividing San Francisco and Sacramento from Los Angeles and San Diego. The research showed this tension manifested in the mobile communications between northern and southern Californians.

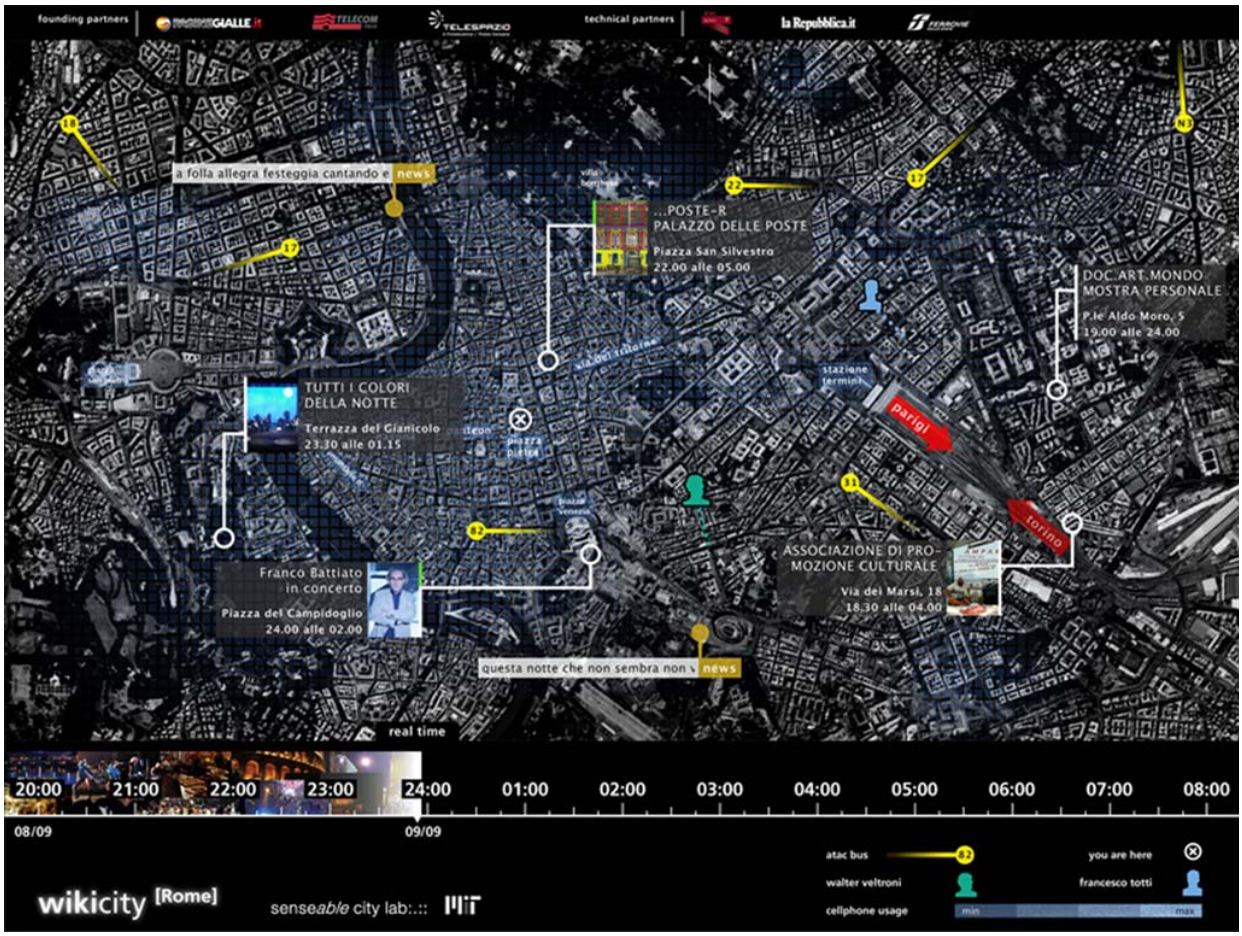

Figure 4. Dynamic map juxtaposing public transport information and current location of urban crowds within Rome (©SENSEable City Lab)

In a similar case, New Jersey was divided between the areas closer to Philadelphia and those closer to New York City. While these instances illustrate the discrepancy of the reality of human connections on the one hand and geopolitical divisions of states on the other, it might also be speculated that the opposite is true with the Texan identity, with its reciprocal connections largely respecting its traditional borders.

Aside from devising the right methods to massage raw telecommunications data to arrive at meaningful and appealing narratives, a serious challenge to this type of study is privacy, as is the case with all uses of CDRs. The records document connections between people indiscriminately and reveal intimate details of individuals of which the callers are unusually unaware. To safeguard personal privacy, individual phone numbers were anonymised by the operator before leaving their storage facilities. The privacy of the individual caller is protected further through geographic aggregation, making it impossible to pinpoint a customer's address, neighbourhood or village. Even anonymised data can, however, reveal the identity of individuals through its correlation with places visited. Geographic aggregation of the data of multiple individuals is an effective way of mitigating this threat, but this is at the expense of accuracy and granularity of the data. To utilise the data without this kind of aggregation in a responsible and ethical way requires the informed consent of individual users, who have to make a conscious decision to what extent they want to share their data with the public or the academic community. Such opt-in/opt-out modes of privacy management are gaining traction, operating under the principle of data-ownership of the individual who generated the data. Such policies make it possible to create open-access platforms that

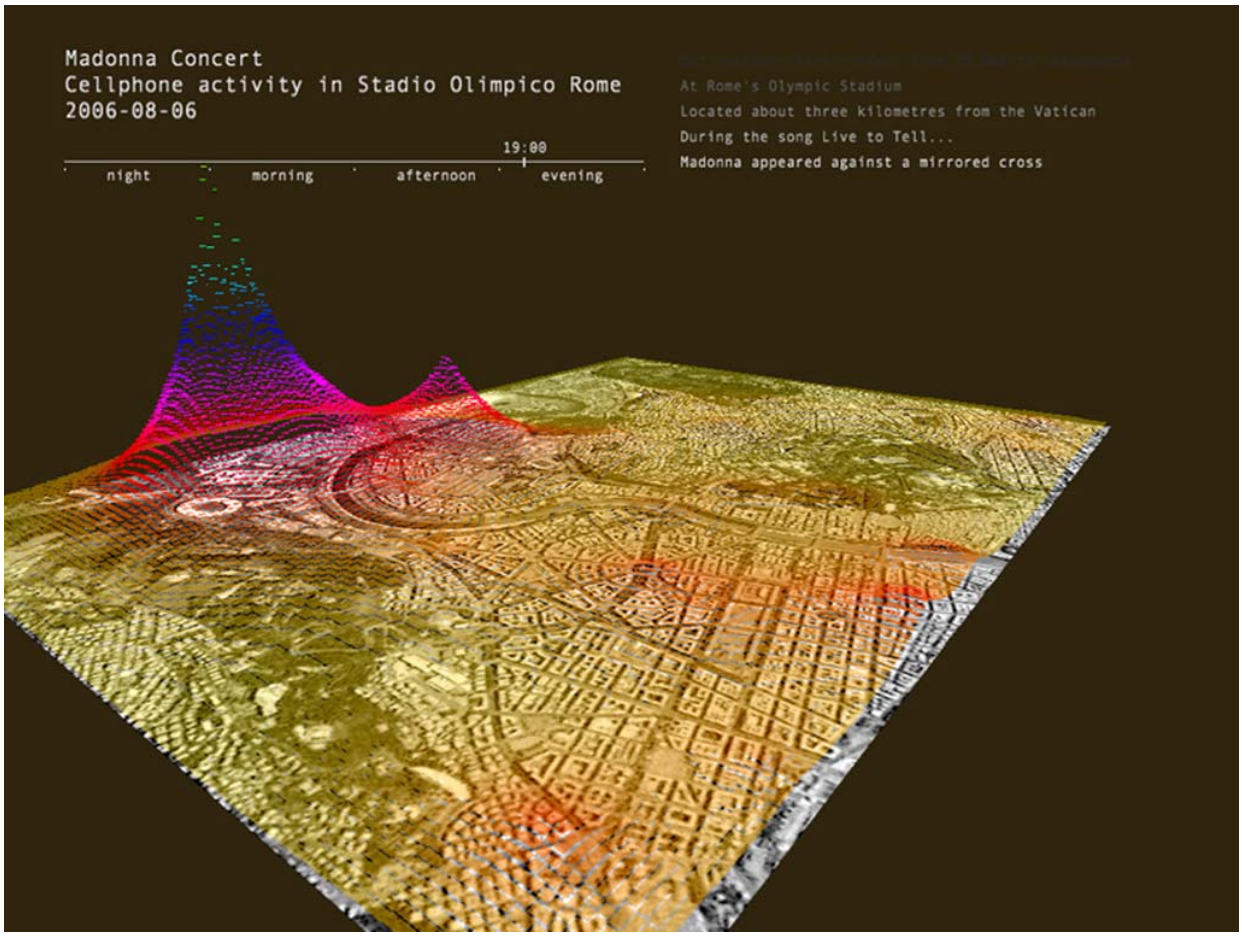

Figure 5. Dynamic map illustrating the real-time levels of activity within Rome through interpolation of cell phone usage of the urban population (©SENSEable City Lab)

aggregate urban information and make it accessible to the public while filtering out private information. Live Singapore! is an example of such an open-access platform (http://senseable.mit.edu/livesingapore/index.html).

2.2 Wikicity Rome: aggregating data from multiple sources

Data has the potential not only to inform citizens about the city, but also to provoke action. To this effect, the Wikicity project (Figures 4 and 5) is concerned with the real-time mapping of urban dynamics, how information could serve as a means of allowing city inhabitants to better inform their decisions and how information can be used by the distributed physical actuators able to act upon the system to realise the desired control strategy (vis-à-vis the actions of the citizens and the urban infrastructure). The ultimate goal of this process is increased overall efficiency and sustainability in making use of the city environment.

In short, the project aimed to close the loop between urban sensing, of varying types, and the multitude of decisions made by actuators of the city (individuals, streetlights, vehicles, etc.). The key component to this was the idea of aggregation of a diverse collection of information uploaded from citizens, local authorities, service providers and businesses. As urban systems now contain isolated sensing capabilities (e.g. global positioning system (GPS) devices to collect the movement of people and vehicles and mobile networks as described previously), the challenge is to combine and overlay these various layers of information in a comprehensible manner to incite action by individuals. This could result in possibilities such as users of the system changing their transportation options to account for traffic from a future event or wanting

to be updated about events that are happening or are going to happen in their present surroundings.

The original Wikicity project in Rome combined mobile phone data, GPS tracks of buses (location, direction and speed) and news and event information during the 2007 *Notte Biachi* onto several urban screens throughout the city. This event implementation allows people access to real-time data on the dynamics that are occurring in the very place they find themselves in, in that moment, creating the intriguing situation that the map is drawn on the basis of dynamic elements of which the map itself is an active part. In this context, the project investigated how people would react to this new perspective of their own city as they began to see various urban elements at play and how this access to real-time data in situ altered the decision-making processes of these observers.

The various feeds provided an added challenge because data formatting standards vary between the various industries, providers and platforms, and the urban platform developed by the MIT Lab has to be able to not only parse, extract and process this information, but it must do so in real time to maintain relevance to the real-time decision-making process of users. To be useable across the different data feeds, an ontology was developed to describe the data, which considered

- location (coordinates system, latitude and longitude)
- time
- data category
- data format (single value, matrix, vector, text, image, etc.)
- data representation (e.g. measurement unit)
- semantics of the data
- the raw data itself.

In a sense, the project depends on both legibility of the usability of the final interface and usability of the data inputs.

Although the Rome example largely remained a series of projected visualisations, the authors believe that the modality of the interface is important because individuals make use of such an information delivery platform. The information sought and what individuals expect from that information is different at different spatio-temporal situations, for example at a transit station or near a major cultural centre that functions as an events venue. Therefore, how the information is aggregated from different sources at the stage of data acquisition and how it is filtered at different delivery points to be responsive to the context within which it is being delivered becomes a filtering challenge.

2.3 Los ojos del mundo: aggregating data from multiple sources

In Wikicity Rome, the multiple sources of data that provided the research foundation came from a multitude of service providers who used the data in their own internal decision-making processes. Because of its operational value, that data has to be accurate and verifiable within precise tolerances and thus its proprietary value limits access to the information to company technicians and a few research institutions. However, the proliferation of mobile devices has led to a massive increase in the volume of records of where people have been and the ease by which they can share this information. This data – from foursquare check-ins to Twitter posts – allows the individual to share their experiences and locations through electronic 'breadcrumbs', whose aggregate begins to tell the story of communities at large.

In the Los ojos del mundo (translation World's Eyes) project (Figures 6 and 7), aggregation strategies allowed the project researchers to understand the dynamics of tourism in Spain. Currently, cities rely on hotel occupancy rates and surveys to understand what tourists see and do in a city when they travel. The best study results using these methods are coarse at best as tourists leave minimal tangible traces of their stay. Los ojos del mundo gained insight into the patterns and trends of tourism by analysing a quintessential activity of travel – photography, specifically the digital photos publicly shared on the internet by people visiting Spain.

This project utilised Flickr, an online photographic community where people share and organise their photographs. Through its community building and technical aspects, this social network provided three sources of user-submitted data as a means to understand patterns of movement and use by various population groups within Barcelona (Figure 6) and Spain as a whole (Figure 7).

- First, as a repository of individuals' photographs, access could be gained to a rich catalogue of people's experiences and the underlying exchangeable image file (EXIF) data that provides metadata created by the camera, including date and time information and sometimes GPS geo-located longitude and latitude coordinates. When the latter are not available from the camera used, Flickr provides a means of manually connecting one's photos to a geographic location, assigning longitude, latitude and an accuracy attribution derived from the zoom level used when the user positioned the photograph.
- Second, Flickr contains profiles voluntarily created by users of and submitters to the site, revealing the countries and cities from where the photographer came.
- Third, the use of tags and descriptions provided by users to sort and organise their photographs provided contextual narratives and words by which the researchers could target and search for specific types of subjects from where the photos came.

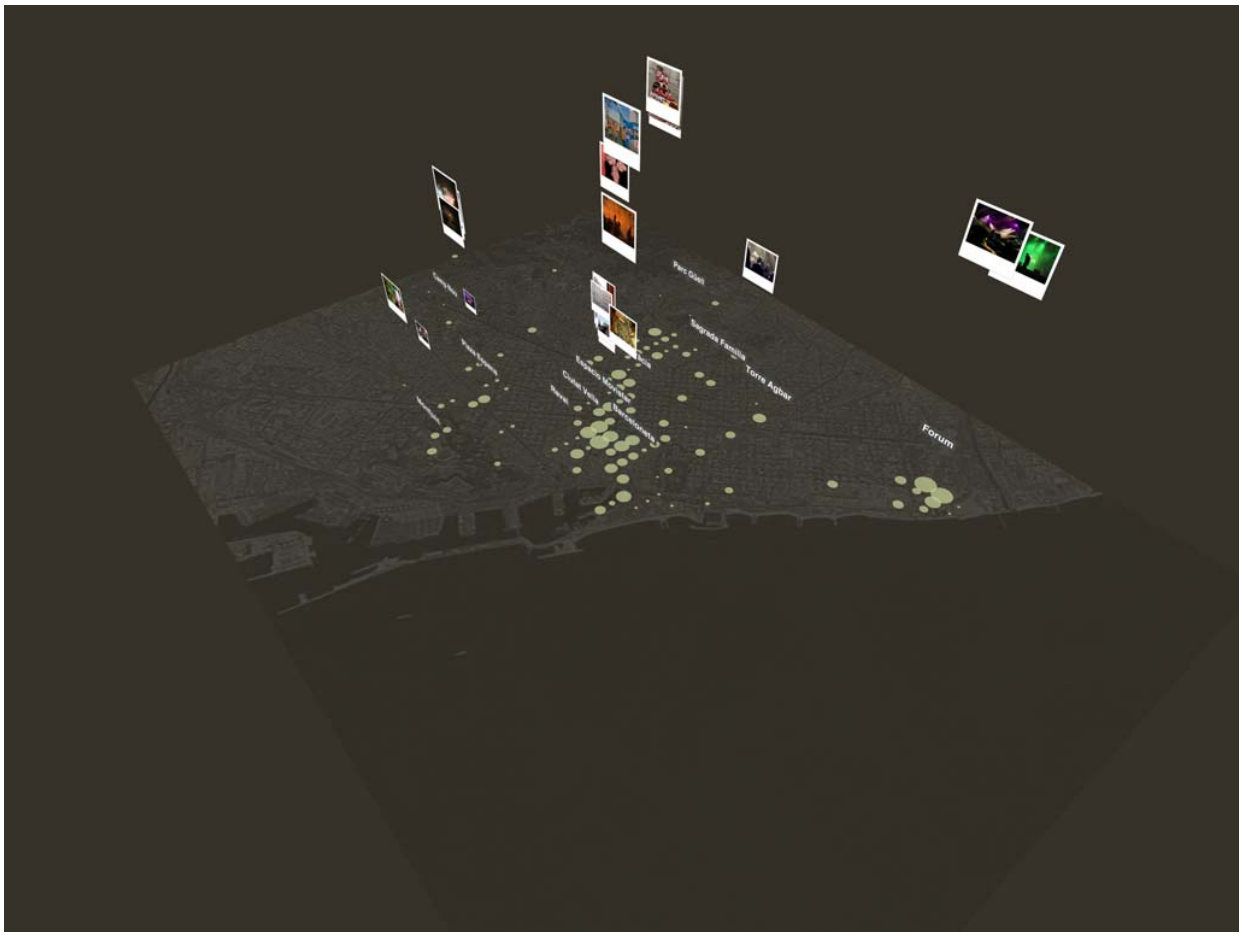

Figure 6. Snap shot of the animation of photos geo-tagged to different neighbourhoods of Barcelona with descriptive tags that relate to 'partying' in the summer of 2007 shows that Barcelona's old town (Ciutat Vella) is where people go to have fun (©SENSEable City Lab)

In each of these three data entries from Flickr, use of crowd-sourced data could be considered opportunistic as the data was created as a service to ease use of the site and made available by the site's public application programming interface (API). Hence, as in any crowd-sourced information, the question of reliability of the responses is always an issue. For instance, the temporal information of the photographs may be incorrect due to an incorrect setting on the individual's camera (few people reset their camera's clocks when travelling). In this case, adjustments can be made when taking the users' home countries into account. Terminology also becomes an issue in collecting such data; for example, a 'party' for a Briton would be a 'fiesta' for a Spaniard, discounting problems of spelling errors and idiosyncrasies in language. Similar issues can also be found when considering the profile information of the user. As a result, the research team was forced to manually process some information in many cases. In a previous study considering Rome in a similar manner, 144 501 geo-referenced photos from 6019 users were analysed and it was found that only 59% of users had disclosed meaningful origin information (Girardin *et al.*, 2008).

The analysis and visualisations considered photographs taken spatio-temporally. The result allowed viewers and researchers to understand where tourists at large travelled to within Spain and Barcelona. By overlaying areas where individuals photographed over a year, locations that attracted photographers' attention (monuments, churches, public spaces, etc.) were

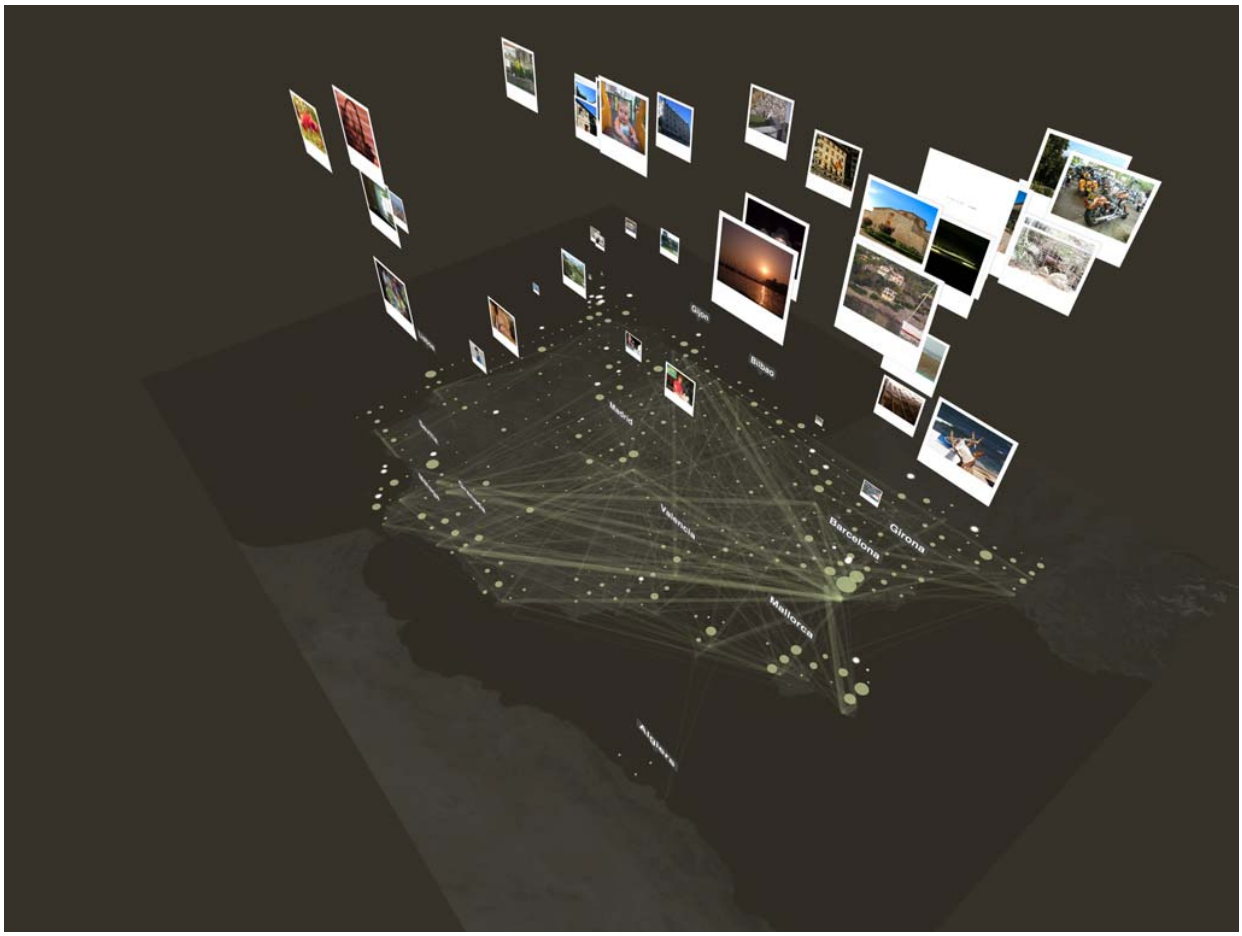

Figure 7. Visualisation showing how Spain is photographed by tourists over the course of one year. While the photos overlap in certain locations and expose places that attract photographers, in other locations the absence of images is eye-catching, revealing the more introverted parts of Spain (©SENSEable City Lab)

exposed. The absence of images in other locations revealed the opposite – locations that could be considered more 'introverted'.

The profile information also revealed a different dynamic when these patterns in photography and tourism were analysed. With information of where travellers had originally come from, geographic presence and trails over time between various nationalities could be delineated. For example, Spaniards visited smaller cities across Spain in greater numbers than foreigners. When just considering Barcelona, the analysis revealed that Britons who visited the city in autumn 2007 stayed close to the city's main landmarks such as Gaudi's Parc Güell and la Sagrada Família, with the Passeig de Gràcia and la Rambla serving as main arteries. In mining the descriptions and tags attached by photographers, insight into the types of activities captured by the photographs could be gained. For example, searches for 'art' revealed the cities that attracted the most artistically inclined visitors; filtering for 'parties' and 'fiestas' across the country revealed cities that hosted the most memorable fêtes over the course of a year and the hottest destinations for nightlife in Barcelona. Unsurprisingly, Barcelona concentrates its fun in the old town (Ciutat Vella), known for its high density of tourists, the bohemian district of Gracia and the Forum area, which hosts music events. All these findings, although not surprising, reinforce intuitive notions we may hold about tourism that would not easily be captured otherwise without these digital traces. However, the

digital traces need to be filtered, checked for accuracy and cross-associated to become reliable if they are to create narratives such as those covered in this section.

2.4 Trash | Track: custom-generated data

The Trash | Track project (Figure 8) is an example of where the generation of urban data proved to be a difficult and expensive process. Custom-made sensors were developed and produced in the required number by a team of researchers at the MIT Lab and in collaboration with members of the industry. Various iterations of the prototype were administered by the project team in order to refine the functionality of the sensors. More than 50 people and 8 different entities (MIT SENSEable City Lab, Waste Management, Qualcomm, Sprint, The Architectural League of New York, Seattle Office of Arts and Cultural Affairs, Seattle Public Utilities and Seattle Public Library) were involved in the design and production process of the sensors, development of the project concept and deployment of the project in two cities. Public participation was solicited for the deployment stage in which volunteers were directed to tag their trash and dispose of it in the urban removal chain.

To track the trash, in Seattle, the MIT team with a group of local volunteers attached more than 3000 custom-developed smart tags (Figure 9) to waste objects discarded by households and schools across the Seattle metropolitan area. The tags' trajectories were then monitored in real time on a central server at MIT. After each deployment, data generated by the network of self-reporting tags was rigorously studied in terms of what filtering algorithms needed to be applied to the raw dataset to create meaningful subsets for the visualisation engine. The project investigated the geographic dimension of urban waste systems by following the movement of individual trash items, thus tracing the flows of an urban infrastructure that is usually hidden from plain sight.

While most citizens have little understanding of the structure and processes of waste management, a lack of reliable data regarding individual material flows in the removal system also remains a challenge in the professional field. Tracking trash raised a number of conceptual and technical challenges, starting with the choice of an appropriate sensing technology. In supply chain management, the technology of choice for tracking the path of objects is usually radio frequency identification (RFID). These tags are small and inexpensive, but can only be detected at close range, requiring a separate infrastructure for their identification. In the waste system, no such infrastructure exists and, even if it were available, the necessary detectors could be easily avoided.

It was considered important not to limit the experiment to assumptions about possible waste destinations and therefore an active sensing technology capable of autonomously reporting back from any location was required. Active location sensing means that an electronic location sensor is attached to an object, the sensing device being slightly smaller than a cell phone. Location is acquired and reported using the cell phone network infrastructure. If available, GPS is used to increase the accuracy of the sensed location.

The deployment of the sensors relied heavily on the involvement of volunteers. Initially, Trash | Track was not designed as a participatory project, yet this aspect soon became the most important part. Volunteers were eager to learn about the structure of the waste system and contributed their ideas, time and materials to the project. Focusing on refuse from private households, these local volunteers donated trash objects,

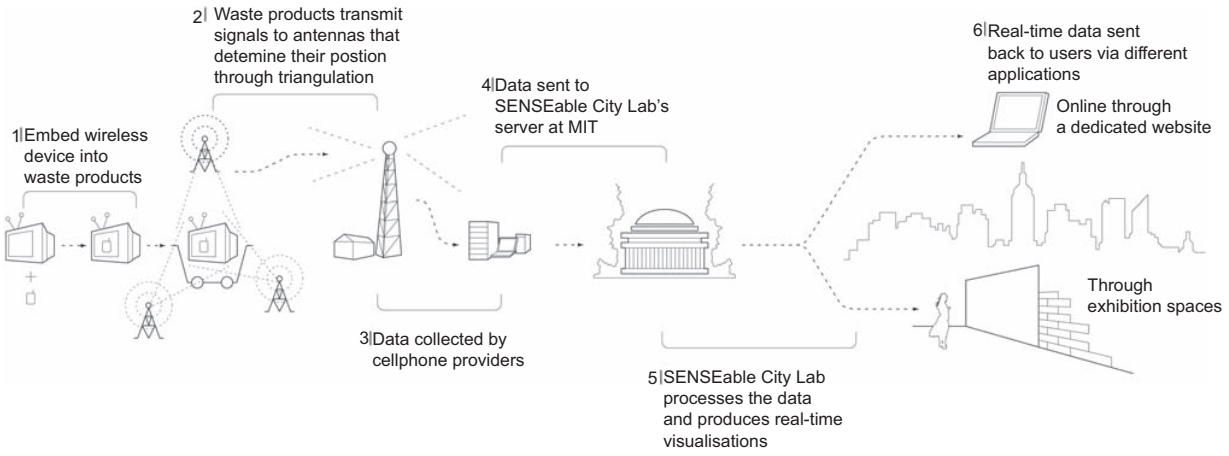

Figure 8. Information flow in Trash | Track (©SENSEable City Lab)

Figure 9. Trash | Track tag (©SENSEable City Lab)

helped with tagging, recruiting and transportation in the city using their own cars. Among the many tracked objects were packaging made from metal, glass, paper or plastic, cell phones, TVs and computers, books, clothing, furniture and toys.

Since the waste stream is a hostile environment for the operation of electronic sensors, the tagging process itself turned out to be the main challenge of the project. Sensors can be crushed and compacted in rubbish trucks, and can be exposed to liquids or buried under material, obstructing radio signals. These issues were addressed by encapsulating each sensor in a protective hull made from durable epoxy foam, while at the same time taking precautions that this hull would not change the appearance of the object too much and lead to discovery of the sensors in material recovery facilities.

After the tagged objects had entered the waste stream, the sensors reported their movement at regular intervals via the cellular network. Real-time maps of the recorded traces were on display in Seattle Public Library and the New York Architectural League, meeting the interest of the many volunteers who were curious to see where their objects ended up (Figures 10 and 11). The movement of the discarded objects was traced over a period of six months, until the batteries of most sensors had expired. The aggregated traces conveyed a rich picture of the waste removal chain; facilities such as transfer stations, recycling centres and landfills could clearly be made out as frequented nodes in the network.

In order to understand this network of facilities and transportation connections, the reported locations were matched with a geographic database of waste facilities maintained by the US Environmental Protection Agency. This allowed the project team to infer the topology of the removal chain – the companies, facilities and transport routes associated with different materials and processes.

A most interesting observation was that electronic and household hazardous waste travelled surprisingly long distances, compared with all other items. Cell phones, batteries and printer cartridges from Seattle arrived at facilities in Florida, Georgia or Mexico, often along routes that seemed convoluted. This is due to the fact that E-waste recycling facilities are sparsely distributed across the country, due to economies of scale, specialisation of recycling processes and the small volumes involved compared with other waste categories. The experiment has shown that the environmental benefit of recycling such objects does not always justify their long transportation. In one case, the trajectory of one printer cartridge covered a distance of over 6000 km, partly by air freight (Offenhuber *et al.*, 2012).

2.5 Challenges of each category: filtering is an answer

The three different data categories illustrated by the case studies come with their own challenges. For existing data, questions of privacy and data ownership needs to be addressed. In appropriation of data, issues concerning resolution, scope and quality have to be dealt with because the datasets are generated for different things and the intended resolution, scope quality and completeness may differ from the researcher's ideas of how to make sense of the data. To this effect, datasets from multiple sources may need to be combined to achieve the required completeness of the viewpoint that a data-driven narrative is to offer. In the generation of data, technical challenges related to the infrastructure of data collection need to be addressed. In the first two categories – appropriation and aggregation of existing data – the infrastructure of data collection has already been implemented and is operational, but in creating one's own data from scratch, an infrastructure needs to be implemented.

Furthermore, a typical caveat of research using datasets acquired from a single source (such as real-time data from a telecommunications provider) is its external validity as a proxy for the phenomenon in question. The dataset might provide a complete account of how a large group of subscribers accesses a specific telecommunications service but, to the researcher, this is usually not the issue of interest. The researcher typically uses this data to answer a more general question beyond the literal meaning of the data. For example, the data might be used as proxy for investigating interactions and activities in a city or country, as is the case in the Borderline and Connected States of America projects. In such situations, if the data is not cross-associated with the reality of the geography that it pertains to as well the specificities of the actions that it records, there is great temptation to confuse the map with the territory

Figure 10. Example of data visualisation of Trash | Track, illustrating the route an aluminium can travelled within the removal chain of the city (©SENSEable City Lab)

and take the data as a true representation of how people communicate in general.

Obviously, the same issues arise when datasets from heterogeneous sources are appropriated, including data generated by users of different social media platforms. In addition to the external validity of the data as a proxy, the internal validity of the dataset – its inconsistencies and biases – also needs to be considered.

When data is generated specifically in an attempt to answer a research question, many of these caveats can be easily avoided. Yet all sensing technologies come with their own set of technical constraints that will almost certainly have an impact on the data. For example, in the Trash | Track project, the team had to deal with the technical challenge of making the tracking devices last as long as possible. But the even more challenging aspect was that, due to lack of prior similar experiments, at the time the sensors were being attached to trash and transported environments far away from the controlled environment of the MIT Lab, there was no way of knowing for how long the researchers would be able to capture the data broadcast from them and when the most useful data would be received: Would it be in the first few days or the last few days? When would those last few days actually be? As a result of all technical constraints involved in the sensing

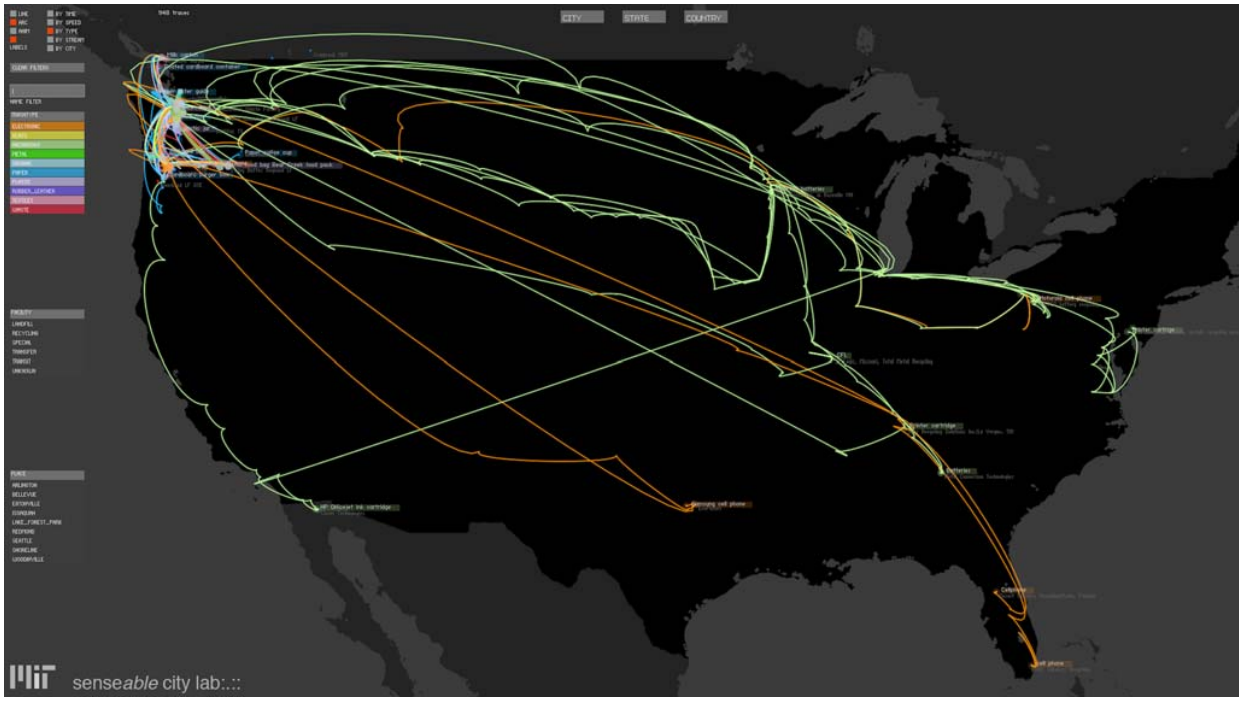

Figure 11. Visualisation of aggregate data regarding how recyclable waste travels throughout the USA (©SENSEable City Lab)

process, the acquired dataset contained a large number of artefacts and biases that had to be addressed through filtering. As a consequence, the dataset used in the final analysis shrank to almost half its raw size.

These are all counter-examples to the popular assumption that 'data is all around us, we just have to use it'. Data is socially constructed (Bijker *et al.*, 1987): we should not see data as a given and hope to learn something from it, but rather look at how and under what conditions it is generated and take this information into account when analysing data and creating narratives about the dynamics to which the data pertain to.

3. Making data accessible

In examining the data dimension, we need to go well beyond our enquiry into accessing data and into the realm of making it accessible to the public as well. However, making data accessible is not an end in itself. We actually intervene between the accumulation of data and its release for public consumption. This act empowers the data and makes it more salient and useful. We normally access data about our environment in order to generate knowledge about its logic of operation. A viable worldview that is reminiscent of the Enlightenment era supports the concept of witness in this regard – an analogy used by Christian Nold in his bio-mapping experiments (Nold, 2009) – that one cannot just produce knowledge without having other people to see it and use it. This explains the proliferation of publicly viewed or witnessed scientific experiments during this era. The same thinking is salient to today's situation: data about the environment that is accessed has to be fed back to those who inhabit it and, in making this data accessible to its target audience, there are challenges to be addressed and potentials to be counted for.

3.1 Potentials of human dimension

In making urban data accessible, the human dimension is added to the equation. In this regard, it would be interesting to use some of the very same projects cited in providing a categorisation of generating new data or accessing already existing data and step back and show how they can actually have an impact on public behaviour. In other words, how making data accessible to citizens can actuate them in certain ways, providing a technologically enhanced platform for collective participation. We can build a very nice symmetry here because these projects (and many others of the same nature) not only generate data but they also create a feedback loop between the environment, its inhabitants and the system for data collection and dissemination – a sort of immediate participatory aspect. Once the acquisition of data from the environment and the broadcasting of data back to the people who inhabit the environment are integrated as a feedback

mechanism, a certain level of humanism is encountered that is integral to this cycle of data acquisition and broadcasting.

In putting citizens in the loop, we can improve data volume and quality by throwing many eyes at them. This in turn offers citizens a way to critique the data and its embedded assumptions, leading to better methods for generation, acquisition and aggregation of datasets. We can also instigate a sense of responsibility for the city and its shared public goods – meaning the urban data that allows for decoding of daily dynamics. Furthermore, the data can support the decisions of individuals and make their lives easier. By democratising access to the data in real time and in high resolution, a new mode of governance that is not 'command and control' but is based on cooperation and voluntary action will emerge once individuals start acting upon the information about the city that is provided to them. For example, Trash | Track was initially a technical study into how a system (urban waste removal in this case) actually operates. However, both the project execution and the results revealed information about the human element and the unpredictability of what people do, how they choose to behave, what they choose to throw away, how they choose to respond to instructions and how they react to results.

Before delving into this subject, it would be useful to define what the authors mean by 'participation'. Participation can be voluntary, where members of an urban population contribute to the data generation/acquisition process consciously with some level of involvement and dedication to the project. At the same time, participation can be involuntary – citizens in the heavily networked cities of today involuntarily leave digital traces from their day-to-day activities as they benefit from various services offered to them on different networks (cellular networks, banking networks, etc.). In tapping into digital traces that the urban population involuntarily leaves on these networks, we make the assumption that the human actor is acting in a certain manner as a whole (calling people they know, uploading photos and geo-tagging for their own reference) to achieve its very own goals, but at the same time the data that is generated as a by-product of these actions can be aggregated with other sources of data and cross-referenced, helping researchers to deduce certain narratives with regard to the dynamics of a city. When it comes to voluntary participation, there are projects that are set up as participatory projects from the beginning – where participation is celebrated as a value in its own right. For others, this is not the case. For example, Trash | Track was not set up as such in the beginning but it manifested a truly emergent participatory quality when people started to develop concerns of their own regarding the experiment, such as asking for a way to be able to track their own trash within the system.

In the Borderline project, just feedback to the audience that there are geopolitical boundaries that do not necessarily represent the boundaries specific to the network of human connections can have an impact on how they organise the social space in relation to geopolitically delineated space, or how they will be opinionated about the exactitude of the geopolitical institutions that divide the space through delineation of different zones.

Going back again to its Enlightenment variation, contemporary understanding of the concept of the witness changes slightly with regard to these projects so that the witness is not just a spectator or a passive observant. The witness, or the audience, of these information delivery platforms is provided with a platform for observing the environment and its dynamics as well as analysing and criticising it. This ideally allows them to act upon their analysis. Hence we can assume that people are actuated based on real-time information that they receive about their environment.

In short, in projects such as Borderline or Trash | Track, a platform is provided that democratises access to large-scale datasets that store the specific dynamics of the environment. The offered viewpoints into how certain systems operate put the observer at a spot that she has the potential to critique the logic of operation of what she is looking at. In Trash | Track, having a holistic view of a field of operation that is at the scale of a country is going to change – at the very least – our understanding of the cost–benefit analysis of this operation, which ultimately can have an impact on how we ourselves operate in relation to the system. In Borderline, access to information regarding the match between geopolitical divisions within a physical space and how social networks operate across these borders may provide a discursive platform that allows for criticism of how the geopolitical boundaries are laid out.

3.2 Challenges to which filtering is an answer

In line with evaluating the potentials, in examining challenges integral to making data accessible, a question to be addressed is: How much data transparency is too much information? For example, in Wikicity Rome, if everybody suddenly gets access to the same added layer of information regarding the real-time spotting of crowds within the city and traffic information, the bottlenecks will just move from more crowded areas to less crowded ones once everybody starts acting on the very same set of real-time information.

4. Conclusion

In many ways, current approaches to urban data analysis resemble the search for extraterrestrial intelligence by listening to terabytes of meaningless signals in the hope of identifying a pattern that makes sense or manifests characteristics of an intelligent extraterrestrial source. In this approach (the 'big data' approach), having access to large-volume datasets is considered sufficient to study the phenomena and its dynamics

that the data refers to. Data is taken at face value and conclusions are drawn based on how well we can identify and predict statistically significant patterns. However, urban data does not come from outer space – it is generated in a specific situation and subjected to technical and non-technical constraints and biases. If we exclusively rely on algorithms for scrutinising large datasets, we run the danger of falling for pareidolia: the human inclination to see patterns everywhere, even where no corresponding phenomenon exists (Riegler, 2007).

In this paper, it is argued that the approach to urban data analysis should be driven by the properties of the data: its structural qualities, the social context of its generation and its inherent technical and non-technical biases. Categorisation focuses on the conditions under which the data is generated and the effects of that data.

Access to information allows the urban public to see hidden patterns that are not observable otherwise. People deduce conclusions about various matters based on recognising how they fit into or create certain patterns of information, by looking at data generated as a by-product of their internal or external operations. Occasionally, the recognised patterns reveal 'unexpected' aspects of the phenomena under study. Yet, as the legacy of John Tukey, an American statistician and author of the book *Exploratory Data Analysis* affirms, 'We need good pictures to force the unexpected upon us' (Tukey, 1972: p.110).

Although this paper has presented examples that make urban data accessible through dynamic visualisations, this is not the only way of delivering such information back to the urban public. Urban data can be connected to constituting elements of public spaces to transform such spaces into interaction fields where social interaction is mediated by digitally enhanced artefacts installed in the public realm and where behaviour is accordingly informed by urban data. An example of this is Lars Spuybroek's D-tower, constructed in 1998–2003 in the city of Doetinchem, The Netherlands, which provided a playful representation of the collective mood of the citizens based on what was voluntarily reported by them on a web-based platform. In one extreme scenario, urban spaces can remain invariant to or 'unplugged' from the digital layer that is blanking our urban environments. In the other extreme scenario, all constituting elements of the city become alive and try to respond to dynamic conditions that emerge. More information on different possibilities in this regard as well as the techno-social challenges of the realisation of such a vision are given by Firmino *et al.* (2006).

The fact of the matter is that the contemporary city is a cradle of information about the various spatio-temporal dynamics that it contains, and recognising patterns in this type of information would reveal aspects of urban life that are invisible, or even unexpected, at first glance. The questions to be addressed are

- How can we sense a city?
- How can we collect information about it?
- How can the space of urbanity be experienced by revealing patterns that are not initially visible – patterns that are only recognised by investigating collected information?
- How can such temporal processes (i.e. the recognised patterns) be made available to a given observer?

In answering these questions, data should be dealt with as the primary substance and its filtering and massaging should be incorporated in both the process of its acquisition from the environment and its delivery back to the environment.

This paper has examined the vulnerabilities, weaknesses and potential dangers of the data that we use and broadcast, along with the challenges, primarily through the lens of filtering. It has been claimed that there are two types of filter – one that applies to the generation of data or reception of data from the environment and one that applies to broadcasting of data or dissemination of data back to the environment. In accessing data, filters are applied to massage data in terms of its resolution, quality and scope as well as to access and cross-associate data that comes from multiple sources rather than data that is single sourced.

In making data accessible, we do not put everything out there: we filter and we do that because we are trying to make it as accessible as possible while minimising the risk of making systems too transparent. It is interesting to think of data as a substance that we have to modify through various filtering techniques and, in certain ways, produce certain narratives and deliver it to a target audience. Yet, with this view point, one needs to bear in mind that visualisations such as the case studies presented in this paper are not data itself, but one of the many techniques to filter data in order to make it as accessible as possible.

REFERENCES

Bijker WE, Hughes TP and Pinch TJ (1987) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA, USA.

Boustani A, Girod L, Offenhuber D *et al.* (2011) Investigation of the waste-removal chain through pervasive computing. *IBM Journal of Research and Development* **55(1/2)**: 11:1–11:11.

Calabrese F, Colonna M, Lovisolo P, Parata D and Ratti C (2011) Real-time urban monitoring using cell phones: a case study

in Rome. *IEEE Transactions on Intelligent Transportation Systems* **12(1)**: 141–151.

Firmino R, Aurigi A and Camargo A (2006) Urban and technological developments. Why is it so hard to integrate ICTs into the planning agenda? *Proceedings of CORP'2006, 11th International Symposium on ICTs in Urban and Spatial Planning and Impacts of ICT on Physical Space, Vienna, Austria*. Medieninhaber and Verleger, Vienna, Austria, pp. 143–152.

Girardin F, Calabrese F, Fiore FD, Ratti C and Blat J (2008) Digital footprinting: uncovering tourists with user-generated content. *IEEE Pervasive Computing* **7(4)**: 36–43.

Nold C (2009) *Emotional Cartography. Technologies of the Self*. See http://emotionalcartography.net/ (accessed 27/11/2012).

Offenhuber D, Lee D, Wolf MI *et al*. (2012) Putting matter in place. *Journal of the American Planning Association* **78(2)**: 173–196.

Ratti C, Sobolevsky S, Calabrese F *et al*. (2010) Redrawing the map of Great Britain from a network of human interactions. *PLoS ONE* **5(12)**: e14248, http://dx.doi.org/10.1371/journal.pone.0014248.

Riegler A (2007) Superstition in the machine. In *Anticipatory Behavior in Adaptive Learning Systems* (Butz MV *et al*. (eds)). Springer, Berlin, Germany, pp. 57–72.

Tukey JW (1972) Exploratory data analysis: as part of a larger whole. In *Proceedings of the 18th Conference on Design of Experiments in Army Research and Development I, Washington, DC,* p. 110.