I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

Does Big Data Lead to Smarter Cities? Problems, Pitfalls and Opportunities

MICHAEL BATTY*

Abstract: In this paper, we describe the emergence of "big data" in cities and argue that an appropriate definition of such data, to make it distinct from the many other data that are relevant to urban research, is that such data is streamed. What makes it big is that it streams from sensors that operate in real time and that such data is only finite when the sensor is switched off. We illustrate some of these data through dashboards and portals that are fast appearing to synthesize the various different streams and feeds that define this data. The coincidence of the managing and controlling urban functions in cities in real time if often associated with big data and we explore these parallels, introducing a number of technical issues pertaining to making bit data useful, particularly its integration with other big data sets. We then sketch the difficulties of such integration using the data we have been working with relating to smart cards and geo-temporal positioning in the pubic transport systems that define various travel networks in greater London, and we then conclude with some suggestions pertaining to how such data might be best exploited for realizing the potential of managing the city in real time.

I. THE RISE OF BIG DATA: HISTORICAL ANTECEDENTS, CONTEMPORARY DEVELOPMENTS

There is a wonderful story that shows big data to be entirely dependent on our ability to process it, which in turn depends on the

* Bartlett Professor of Planning, University College London, 90 Tottenham Court Road, London W1N 6TR, UK.

capacity of the computers that we have available. Back in 1953, Joe Lyons and Company operated a popular chain of some 250 tea shops (and many other outlets such as hotels) in British towns and cities and, as part of their operation, they were at the forefront of digital computing in business. They teamed up in the late 1940s with Maurice Wilkes[1] at Cambridge, who was working on one of the world's first digital computers, EDSAC. In return for funding some of Wilkes' research, the company began the production of its own computer closely fashioned around the Cambridge model (Ferry, 2012). LEO1 was first produced in 1951 and LEO2 a couple of years later. By then, the embryonic computer group at Lyons had begun to take in computing jobs to pay for their idle cycles from organizations as diverse as the UK Meteorological Office and the Ford Motor Company.[2]

In 1955, they were approached by British Railways, who wanted to compute all the distances between some 5000 stations in Great Britain, so that they could price their freight (and passenger services) efficiently and consistently. Now 5000 stations – a number that implies that there are some 5000*4999/2 symmetric link distances to be calculated – is a tiny data problem today, but in 1953 it was enormous. This was "big data" by any standards and, although the LEO machine could certainly compute this matrix of distances, the process was painful. Remember that this was an era when computers were still based on valve technology and when everything had to be inputted to the machine through an intermediate media – through punched tape and then punched cards in this case. Greening-Jackson (2012) says of the machine:

> "It had 2K _ 35-bit words of memory, implemented using mercury delay lines. Input was from punched card, and output could be either to card or to a printer. Programs were hand assembled (i.e. there was no separate assembler program) and the (decimal) op-codes were written on coding sheets. These coding sheets were then keyed as Binary Coded Decimal (BCD)

[1] Wilkes' group pioneered digital computers with strong links to the US efforts and Bletchley Park where Turing amongst others developed the first digital computer Colossus in 1943. In fact the Lyons group were amongst the first non-scientific group anywhere to make contact with the founders of digital computation (Ferry, 2012).

[2] *Wikipedia*, s.v. "LEO," last modified March 28, 2015, http://en.wikipedia.org/wiki/LEO_(computer)/.

> on to punched-tape, which was subsequently converted to pure binary on punched cards by LEO itself. Each punched card could hold 16 instructions."

And so on. It is amazing that anything worked, but it did, and it provided the kind of discipline in problem-solving that still drives us to search for better solutions to computable problems.

What the team did was to break the problem down regionally into parts, solve the shortest routes problem for each geographical bit and then stitch the system back together. For example, Scotland had only three rail lines connecting the rest of Britain, and it was easy to see that this kind of partition could minimize the calculations. It is a completely manageable problem today without any partition, and there are applications out there on your phone and GPS device that use similar procedures to compute shortest routes in microseconds. But there was a problem. No one had solved shortest routes problems before. In fact, it was not until 1959 that Dijkstra[3] invented his famous algorithm but this was 1953. So the Lyons team simply invented it on the job, so-to-speak. When this story was discussed with a key member of the team, Roger Coleman, in 2012, he admitted he had never heard of Dijkstra![4]

There is another punch line to this story. When the job was completed, the team delivered a stack of printout to British Railways, so they could use this as a lookup table to price their freight. And in typical British fashion they never heard from them again. Whether the results were ever used we will never know but we suspect not. Yet this is still a very sobering story about big data. Big data is relative in terms of its volume, and this relativity depends on how quickly and easily it can be processed. In this sense, big data always stretches the limits of our computing power. What is small data today is what was big data yesterday, and definitions based on volume – and there are many – are thus rather limited. Currently we are able to process data volumes that are measured in terabytes – a thousand gigabytes – which can now be stored on an external hard drive which costs about

---

[3] Edgar Dijkstra was the first person to publish an algorithm that enabled one to compute all the shortest routes between the nodes in a network given knowledge of the direct segments that linked the nodes.

[4] A much fuller presentation of this story was given by John Graham-Cumming in a keynote entitled "The Great Railway Caper" at the Strata Conference London 2012 and his talk is available at "John Graham-Cumming keynote Strata Conference London 2012 "The Great Railway Caper," YouTube video, posted by "O'Reilly," October 2, 2012, http://www.youtube.com/watch?v=pcBJfkE5UwU.

$50. Of course, it is limits on random access memory that really determine what we can process as the LEO team back fifty or more years ago was really well aware of. Once we pass beyond a terabyte, we are up against quite hard limits for most routine computing, and thence we move into the very specialized territory of big data and its processing where data volumes are at petabyte or even exabyte level. This requires very specialized skills as well as new forms of analytical modeling as the data volumes get greater. Volumes are thus still important as they imply qualitative change in how we go about working with such data. In fact, Kitchin (2013) has a much fuller definition of big data that incorporates several key issues identified in the big data community. In his conception big data is: huge in volume, high in velocity, diverse in variety, exhaustive in scope, fine-grained in resolution, relational in nature, flexible in terms of its extensionality and scalability.

The definition of what constitutes big data is thus relative to our abilities to process it, as our example clearly shows. In fact, it is processing rather than storage that is key to its definition. A simple rule of thumb currently suggested as being a definition of big data is any volume that will not fit in an Excel spread sheet or rather any volume of data that cannot be manipulated with such a spread sheet. If you filled the currently available spread sheet from Microsoft Office which has 1,048,576 (2^{20}) rows and 16,384 (2^{14}) columns, it would freeze and no processing could be done, illustrating that it is not simply storage that is at issue, it is processing relative to storage. In fact, my own predilection for a definition of big data pertains to data that is streamed in real time. There are many data sets that are not temporally streamed that seem "big" in the volume sense, but it is only data that is truly incomplete at any point in time that can be big data in the streaming sense because that data can continue to be collected. That is, at any instant, what is already available can be difficult to understand without future streams of the same data. Big data in this definition has no potential bounds; it is data that, once the sensor is turned on so that it can be streamed, is bounded in volume only by the point in time at which the sensor is switched off. Usually this is not known in most of the applications that we will have recourse to explore here.

In fact, data that originates from sensors is not new but its translation into digital form is relatively new or rather, its availability to interpretation, access, and exploration is relatively new in that the sensors themselves are now likely to be linked to devices that enable the storage of their data as well their analysis. Much of this data is used for control purposes, as we will note in the next section, but it is

very different in structure from that which is collected through conventional non-digital means. In fact, big data tends to lack significant structure as it is streamed from devices that may add a little structure to it, but that structure is likely to be rudimentary. It is likely to focus simply on obvious properties of location, time, and any other feature that is associated with the object in the system that is being monitored.

There are several different types of sensors, which are rapidly being deployed into urban environments, that have recently raised the profile of these kinds of streamed data and indeed have given rise to the very term big data. We can make at least three distinctions between different types of sensor although our treatment is not all-inclusive but simply notes those most relevant to environmental applications. The most basic sensors are devices that simply record the position, time and basic attributes of a mechanical or electrical device that is performing some routine function in time and space. Good examples from the past are loop counters in roads that record volumes of vehicles passing over them at a different time periods. Digital versions of the same are now available from arrays of sensors that provide a much more complete picture for several kinds of moving vehicle; in a later section, we will explore how such data can be used to explore the position of tube trains and buses in Greater London taken from data that is available in the public domain.

The second kind of sensor is more flexible and this relates to devices such as smart cars associated with individuals and can be used to activate devices which themselves are fixed. For example, smart cards which record information about trip-making and monies spent and are activated to travel on fixed route systems such as subways produce data that is much more varied than data associated with fixed position sensors. The user of the smart card has mobility and this makes the system able to respond to changes in demand and usage. Loyalty cards which accumulate points and can also be used as credit cards are even more flexible because they generate data that pertains to different kinds of purchase, and this provides enormous potential for profiling users and adapting the targeting and display of goods linked to these profiles.

A third type of sensed data pertains to individual devices such as mobile phones which contain a mass of functionality that can link users in terms of the position and timing to many different applications. Mobile phone call usage (from call data records – CDRs) can be used to generate patterns of movement in space. Associated with the profiles that mobile phone operators compile on their users, they can provide patterns of usage and behavior in terms of

movement, purchasing and related tasks in different places. Related information can also be extracted from social media, which is activated from phones and related devices, although this kind of data is in its infancy with respect to being directly useful for purposes of both understanding the city better and for control.

In the past, most socio-economic data has been collected using personal surveys from questionnaires that seek to elicit directly the responses of individuals to a series of pre-planned questions. The "gold standard" in this respect is the Population Census, which in most western countries is taken every tenth year, sometimes every fifth. It usually entails a complete enumeration of the population, which picks up considerable detail concerning the activities of every individual in each household, measured at the location where the individual resides. This kind of data is highly structured. It may lack structure for particular purposes because it is designed chiefly as a detailed head count; because the data is regarded as a general resource, the precise usage of each category of data is not finely tuned to serve other objectives. Certain movement and migration patterns can be extracted, and there is some effort devoted to making the data comparable between each decade, constructing a time series of the data, despite the fact that the time interval is rather aggregate and thus much change is missed. Generally with such data, it is aggregated to categories that make identification of any individual impossible, and privacy and confidentiality issues dominate the construction and release of such data sets. There are many data sets of this kind and, in the case the Population Census, the number of attributes concatenated against the multiple attribute categories that might be formed can elevate this kind of data to the status of "big" in volumetric terms. In fact, although only 56 questions are asked, there are many categories of answer, and the UK Population Census thus generates something in the order of 10^{10} different cell counts, which is between gigabytes and terabytes in terms of storage. Usually data is never available from such sources at this level of detail although, in principle, it is easy to see how this kind of data might become "big" under certain conditions.

There are data sets now appearing which categorize populations at the same level of detail as the Population Census but are being created on a much finer temporal cycle – usually weeks and months rather than seconds and minutes. In this sense, some of these data sets might be construed as being "big" as they scale temporally. Geo-demographic data sets that reflect point-of-sale data are typical examples especially where sales are made up of very many individual items as in supermarket shopping; many online web resources such as *Amazon* are now generating enormous data sets that can be linked to

particular consumer attributes in terms demand profiling. Google, of course, represents that most obvious of big data for its search data is enormous, and increasingly associated with profiling. There are now some 2 million searches per minute, only slightly more than the 1.8 million posts to Facebook. In fact, there are around 200 million emails sent each minute and some 20 million photo views on Flickr. These are very dramatic data volumes. To an extent this represents the cutting edge of big data, and these social media and internet related data might even be classed as yet another variant of big data.

II. UNDERSTANDING AND CONTROLLING THE SMART CITY

To an extent, the emergence of big data is directly associated with new ways of monitoring and thence controlling routine functions that occur in time and space where their control is on a second-by-second basis. Indeed, the emergence of the idea of the "smart city" is based on the notion that computers have reached the point where they can be deployed for many kinds of public function associated with making the routine management of the city more effective – more efficient, of course – but possibly more equitable and certainly more sustainable. In short, big data is a consequence of this deployment of computers through networks of sensors that are computer-controlled and whose data captures the operation of these systems, if not their management. Such data is used in real time to control and steer in various ways the routine functions that such systems use.

In fact, the smart city movement is largely driven by the extension of computation into the public domain, namely, public spaces, the urban commons and the public sector which have been dominated hitherto by non-automated forms of activity. Many aspects of this domain are now being sensed in various ways. There is considerable overlap between the public and private realms, particularly where sensing technologies are deployed in private spaces but monitor what essentially is the interface with public spaces or public spaces per se. Closed circuit TV is the classic example. For a long time, information technologies have been penetrating private spaces at work or in retail outlets, for example, as well as some public spaces, but it is only very recently that city-wide systems of sensors have been deployed. This is particularly evident as we will illustrate here in transport where their control in terms of passenger ticketing is now largely automatic in one form or another, despite the link between big data associated with such systems and their control still being rudimentary.

At a very basic level, much individual behavior in cities, which manifests itself in patterns of movement and location, is being

informed by new information technologies, particularly data that is being communicated via mobile devices, such as smart phones. For example, transport information can be picked up across the mobile internet, although so far it is extremely difficult to get any sense of how users are reacting to such information in terms of their behavior patterns. That behavior is being influenced by the availability of information in a mobile context is easy enough to demonstrate, but the scale and impact of this is almost impossible to track and measure. Currently there are very few systems that exercise overall control over city-wide systems. London's Surface Transport and Traffic Operations Centre (STTOC) (Theophilus 2014), for example, is a mixture of traditional data and media communicated using analogue and digital devices but largely coordinated by operators who use their judgment on the basis of these media to make decisions about traffic signal timing controls, the deployment of accident and emergency services, and coordination between different transport modes. The notion of a city control center is a long way from reality despite some high profile exemplars such as that pioneered by IBM in Rio de Janeiro (Singer 2012).

For a long time there has been real time data streamed from different urban locations but, until the last couple of decades, most of this has not been coordinated in any fashion. Weather information is some of the oldest, but local data is still at a relatively coarse spatial scale as sensors are rarely deployed in any systematic coverage, although this is changing. In many cities, there are an increasing number of portals for accessing such streamed data, although much of the data remains unstructured. The openness of these kinds of data depends very much on how valuable the operators consider it to be. In public systems, for example, where the profit motive is low-key or in systems, which are highly controlled with a monopoly value, streamed data would appear of less value than data that pertains to any marketing function. What are emerging very rapidly, however, are portals in the form of dashboards that take streamed data and display it in an-easy-to-absorb format, putting such data into a context that more general users can reflect upon. Currently, these are really of only general interest in cities but as their data improves in terms of information that might pertain to active decision-making and management, then such dashboards will become important.

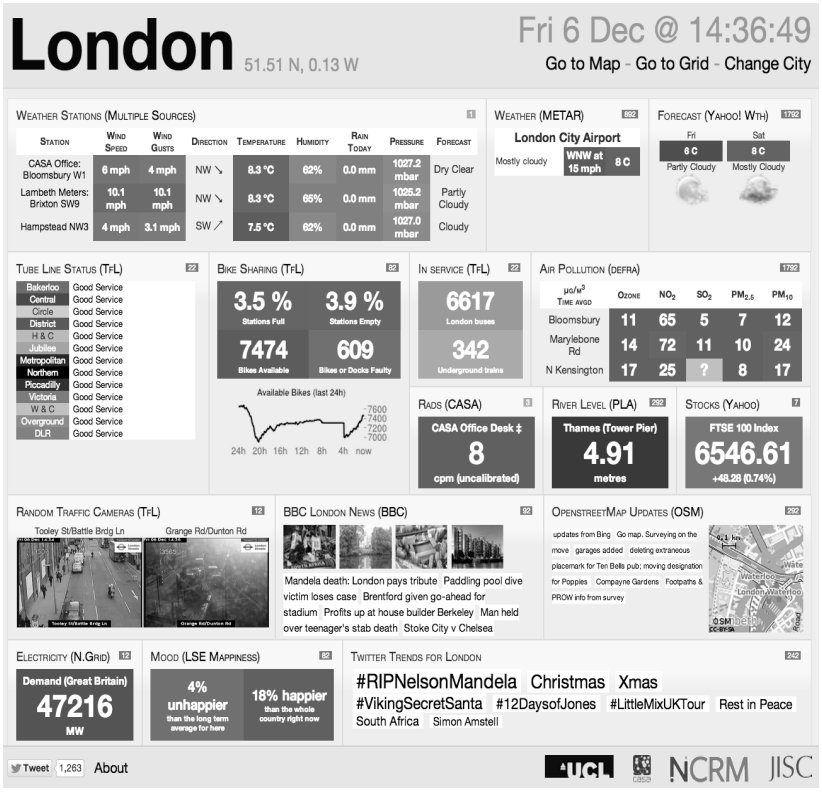


Figure 1: Live Data Feeds into A Dashboard
(from http://citydashboard.org/london/)

In our own context, we have developed simple dashboards for several UK cities. An example of one is shown in Figure 1 for London where we have taken a combination of the 27 real-time live traffic and transport data feeds from the London Data Store built by the Greater London Authority (GLA) to make such data open. Through these feeds, our center (CASA at UCL) has created a City Dashboard as a means of viewing a key number of these live data feeds. This essentially is a dumb interface to a visualization of these data streams that is updated in real time, and delivered in a web-based manner through http://www.citydashboard.org. In fact, we simplify the data feeds in the dashboard which collates about 20 such feeds from air pollution through to energy demand, river flow, the FTSE 100, the number of buses in service, the status of the subway networks, and so on, which we illustrate in Figure 1. The dashboard is an early example of collating and visualizing data feeds to provide a view of how a city is

currently performing.

Not limited to London, the dashboard has also been built for Birmingham, Brighton, Cardiff, Edinburgh, Glasgow Leeds, and Manchester with a version for Venice under development. But in these different cities, the types of streaming data can be a little different for the dashboard highlights the variability in the availability of feeds from city to city. In the UK, London, at the present time, is the location for a majority of data feeds with their number updated on a second-by-second basis. The majority of this data is either collected via a web Application Programming Interface (API) which is usually a web site where a user can query the status of the system with the live data being delivered to the user (or client) or the data can be delivered off-line and mined in accordance with a data provider's terms and conditions. The ability to tap into these API feeds allows the city dashboard to provide a view of the particular city at a glance with the use of simple color coding to indicate the positive or negative connotations of the current state of the data.

London and Amsterdam both have dashboards that go beyond our own physically orientated data streaming technology. Amsterdam's dashboard is organized in terms of socio-economic data and is an archive rather than a live stream of outputs. Divided into eight sections -- transport, environment, population, culture, social-political, sport, security and the economy -- it displays trends for these counts shown over the working day as in Figure 2(a). The speed of refresh is hours not seconds, and thus this dashboard is more like a periodic "state of the nation" report. In fact, users can plot the data in mappable form. The dashboard has simple GIS functionality, and users can thus get a picture of how these categories are changing spatially across the city, which is divided into 50 or more small zones. In a sense, this implies what might be possible in the not too distant future as new sources of open data come on-stream such as house prices, rents, or migration statistics, These potentially might be delivered and updated on a day-by-day basis or at least on a cycle which is much shorter than the typical year, approaching the second-by-second focus of dashboards based on streamed physical data. London's official dashboard is a cross between our own and the Amsterdam board for it contains more abstracted information about rates of change and is focused more on socio-economic issues, as Figure 2(b) implies.

The distinction between physical and socio-economic data in urban applications is important. Although many urban functions deal with physical data and much physical data has significant social and economic implications, the kind of data that is captured and displayed

in dashboard and portals such as the ones just described deals with routine change at the level of managing rather than planning the city. Big data, because of the strong streaming element to its form, is largely associated with shorter-term management of urban functions and in this sense, it is changing the emphasis in city planning from the longer to the shorter term (Batty 2013). Many of these functions have been managed long before the rise of big data and the automation which generates it, but such management has been routine. The existence of automated data captured in digital terms now provides the opportunities for intelligent control of these systems, and a great array of new data mining and related pattern recognition technologies is being brought to bear on the city in these terms. It is this that has the potential to make the city "smart," although to date, little has been accomplished and much of the smart city movement is involved with realizing this potential. There are many obstacles to transgress if the promise of big data is to be borne out in smarter cities, which will depend as much on smart citizenry as anything else. However there are some important technical issues that dominate the debate and it is to these that we now turn.

III. INTEGRATING DATA SETS: BIG AND SMALL, OLD AND NEW

One of the key prospects for big data involves the supposed opportunities for linking or integrating such data with other related data sets, thus realizing economies of scale, and adding value to the data (Batty, et al. 2012). In fact this is a notion that is not particular to big data but it appears to have become more significant. This is probably because, if one joins two or more big data sets, the increased volume can be more than the sum of the parts; joining data often involves concatenations that explode the number of categories and thus the potential dimensionality of the data. The whole idea of integrating two or more data sets involves finding some common key and in the geospatial world this is invariably some spatial referent – a zip or post code – or some spatial metric such as latitude-longitude. This is the most neutral of keys but any field which two or more data sets have in common can be used to make the join.

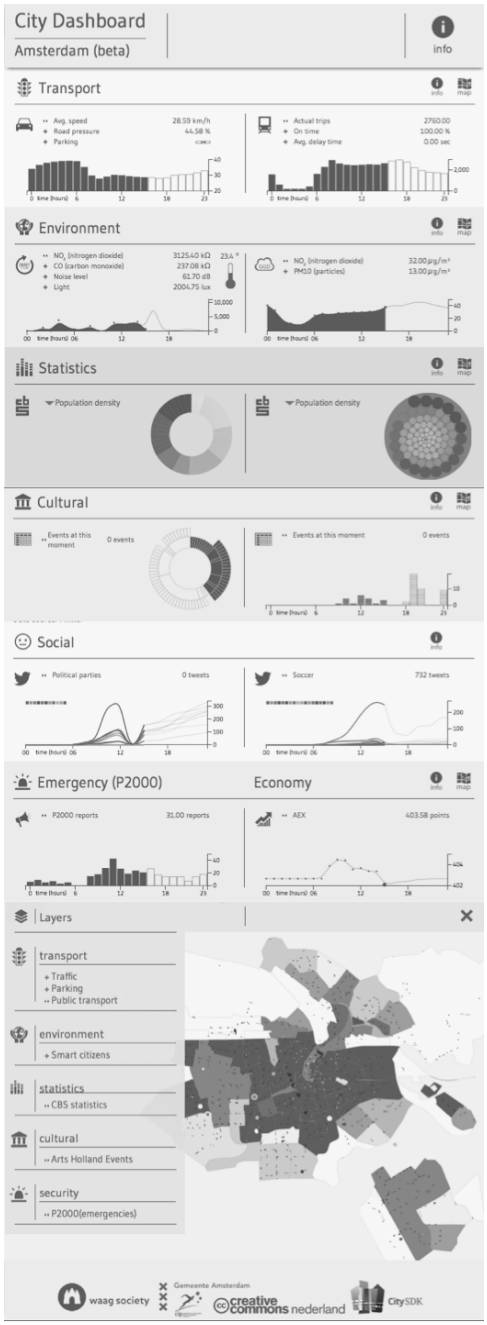


Figure 2a: Previous page
The Amsterdam Dashboard as a Living Archive
(from http://citydashboard.waag.org/)

Figure 2b: Above
The London Dashboard from the London Data Store
(from http://data.london.gov.uk/london-dashboard/)

To illustrate how problematic is the integration of two or more big data sets, we will draw an example from our work with data from travel on the London underground (Tube) system. Most passengers using the London Tube and overground rail network use a stored value card called Oyster on which they can load money for travel on the network and which is activated using a smart card reader when they enter and leave the system. This deducts the cost of their journey as well as recording a variety of other information, most of which pertains to where they have traveled. Data about the monies left on the card and the status of the traveller, relating to the type of card, are also available. No other attribute data is available but any user can display a record of their recent journeys on the ticket machines in any rail station. An illustration of the system is given in Figure 3.

Transport for London (TfL), which operates the system, make much of its data available as open data. One of the sets that we are working with comprises twelve weeks of data for all travelers on the rail network from 16 June to 9 September 2012. This covers the period

of the London Olympics, and we are assuming that the data at the start of the period and the end is the most representative of weekday and weekend travel on the system. A traveller on the system taps in and then taps out; in total during this period, there are some 544 million tap events on the network which implies that there are on average some 7.5 million taps each weekday and 3.63 million each weekend day. As a very crude rule of thumb, we might assume that, if a person makes a journey in one direction, they will make it in the other, and thus we need to divide the number of taps by 4 to provide some sense of how many commuters are in the data. We estimate between about 1.5 and 2 million commuters per day use the network. But part of the analytic problem with this kind of data is guessing or estimating what these individual trips are for. In fact there are quite a lot of travelers who do not tap in or out; for example, those with special cards – free cards for those over 60 years of age, and those who have a season ticket card – need not tap in or out at open barriers because the cost of individual journeys is not relevant. In fact, there can be quite a leakage of travelers from the data set because of the fact that barriers are left open, particularly in mainline stations late at night. There are thus many issues with the data such as these, even though the data provides an excellent record of personal travel.

Figure 3: The Oyster Card Tap-In-Tap-Out System

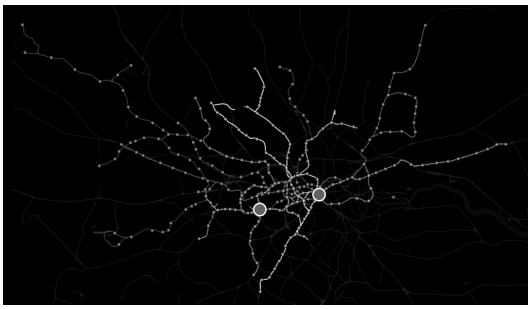
The data set that we are working with is delivered to us offline by TfL, and in general, this data has never been used to date for any real time control of the system. All analysis has been accomplished from the archive and so far it has been used to examine strategic rather than routine issues. The data set is structured as an XML file of 32 fields starting with *date_key*, *time_key*, *oyster_id* and so on from which we can extract data pertaining to where a user taps in and out, the length and cost of the journey, and the number of segments traveled where a segment is a journey from the point of tap-in to tap-out. Data can of course be aggregated for any set of segments or numbers of travelers using rail stations, and it is straightforward to generate an origin and destination matrix of travelers where origins are stations where the individuals tap in and destinations stations where they tap out. What we cannot derive from the data is the actual subway or rail line used which make up the segment. In short, when a passenger enters the system, there are many ways in which they can travel to their destination without leaving the system, that is, without tapping out, and there is no record of what this is. There is no information technology to date that enables anyone or anything to track the passenger from tap-in to tap-out, although in time this might be possible.

However, what we have done to assign travelers to rail lines is use the standard Dijkstra algorithm referred to in the first section of the paper. This enables us to compute the most likely route. The picture we have of the system's behavior in terms of flow volumes through time between stations is computed in this way. The rail network however is extremely complicated and travelers who do not know it will take longer to travel and find the right lines than those who are veteran travelers. Those who know what is above ground or outside the network can also use this knowledge to know where to travel. There are multiple shortcuts within stations themselves -- that is, more than one way to get to a line from the point of tap-in or to the point of tap-out from the platform. All of this adds a high degree of uncertainty as to how travelers actually use the network.

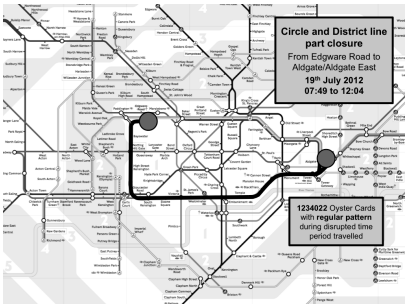
In one sense, this is not as relevant as inferring trip purposes from this data and mapping the volumes in stations to related land use and activities in the immediate vicinity. Research is beginning in attempting to link location to trip movements, but the work is in its infancy and involves linking data sets that are related at different aggregations, thus introducing considerable uncertainty into such analyses (Zhong et al. 2014; Munizaga and Palma 2012). In fact, our focus on this data is to examine disruption. During the twelve- week period which this data covers, we are able to examine the impact of a

part closure on the Circle and District lines which lasted for over four hours on July 19, 2012 from 07:49 to 12:04. This closure affected some 1.23 million Oyster card holders who were on the system during this period and whose trip pattern was impacted by this closure in various ways, such as adding to their travel time, causing them to switch from tube to bus, divert to other lines, or leave the system to complete their journey by walking. We show the configuration of the system and disruption in terms of stations and rail lines in Figures 4(a) to 4(f) where we show some of the computations made with respect to travelers impacted by this disruption. In and of itself, this kind of analysis is useful and innovative in that in a wider context, it can lead to new strategies to deal with disruption. But in fact the problems that it exposes are much more deep-seated than these pictures imply.

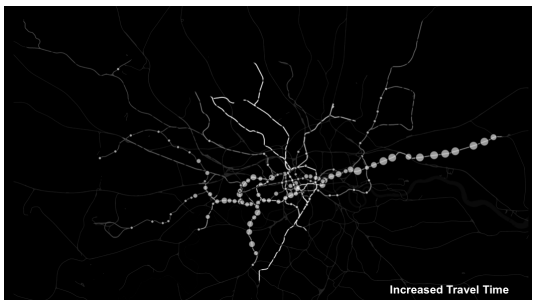
The key issue in measuring disruption is that the disruption occurs on trains and thus passengers must be linked to trains. As we have argued, this is difficult but not impossible as various assumptions might be made about how passengers once tapped in then move to platforms and enter trains. In fact, TfL has several other sources of open data which measure the supply of their trains. The Trackernet application (web API) produces data (with a three-minute latency) that gives the position and time of every train on the network. By querying the URL (http://cloud.tfl.gov.uk/TrackerNet/), the user can extract various positional, temporal and other basic attributes of each train on the network, thus being able to position it quite accurately as the data in Figure 5 reveals. Thus, in principle, if we know how a passenger is moving from the point where they enter the system to the point where they board a train on the platform, we can associate passengers with trains: in short, we can integrate demand data from Oyster with supply data from Trackernet.



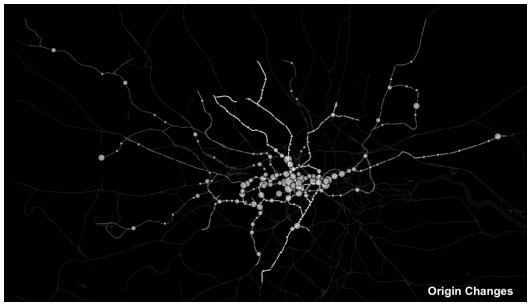
a) The London Underground Network



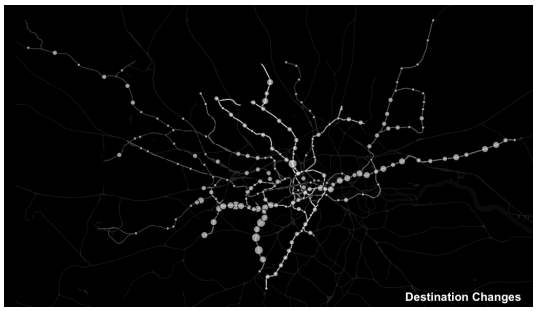
b) Schematic of the Disruption



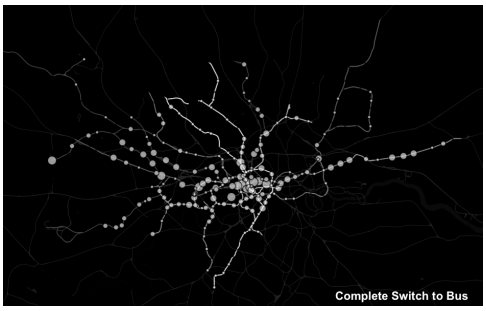
c) Increased Travel Times



d) Changes in Origins



e) Changes in the Destinations



f) Complete Mode Switch to Bus

Figure 4: Measuring Disruptions in Passenger Journeys

This integration of two big data sets is essential to making sense of disruption, largely because on any train which is or has been subject to disruption (which we assume is delay), there will be passengers with different degrees of disadvantage. Those who boarded the train before the disruption will have been disadvantaged while those after the disruption has cleared will not, unless they have come from another related area of the system. In short, we need to integrate these two data sets, but there is no common key. Again, although we have position and time for passengers and trains, we cannot unambiguously link these – we are missing the common key, which is the actual position and time when the train doors open and the passengers step onto or out of the train and on the platform. Whether we will ever get this is an open question. In the far future, it may be that we will all be so wired and monitored that our every move will be recorded, but this is unlikely for many reasons relating not only to feasibility, but probably more to personal freedoms. (These concerns,

although relevant to this application, do not pose any threats as yet.) The point here of course is to demonstrate just how difficult it is to integrate big data sets, which is clearly necessary when problems of the kind we have just sketched are to be considered. Indeed these are exactly the problems, which many consider to be key to making cities smart.

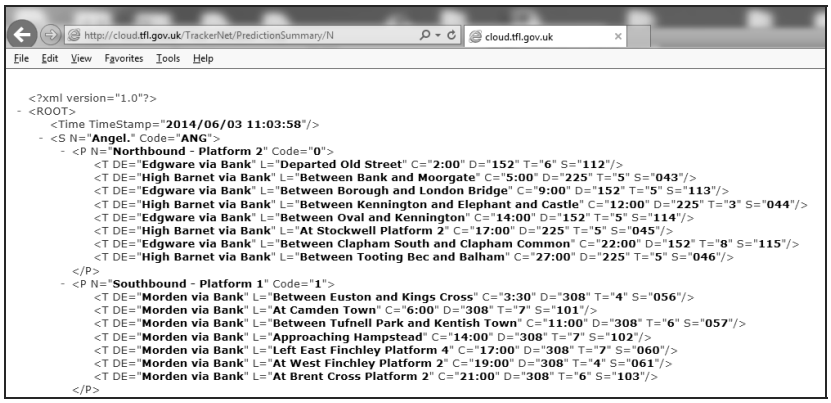


Figure 5: A Typical Query Showing the Position And Time Stamp of Trains Approaching the Subway Station Angel at 11:03, 2014/06/03

Before we move onto considering how this big data might be employed for real time control, it is worth noting many other features that limit its usefulness. In the example of disruption we have just illustrated, we have omitted the overground railways from the Tube largely because the Trackernet does not cover the overground, as the Network Rail API covers this. Moreover Trackernet also misses the Docklands Light Railway (and some minor lines like the Wimbledon tram). There are also glitches in both data sets with respect to accuracy and it is difficult to guess the causes of these errors but the Oyster card data has revealed impossible situations where the same oyster-id is used at the same time in different places while Trackernet sometimes produces data for trains heading in the wrong direction.

IV. THE REAL TIME CITY

A generation ago before the era of big data and the smart city, most formal activities in thinking about the future city related to longer term change, changes that were the subject of physical planning and management involving plans and policies that took years rather than hours or weeks to implement and achieve. The

longest plans were predicated over time horizons of decades. In fact, such plans still exist today although very long-term plans are now just one of many examples of plans and management across a wide spectrum of time horizons. During the last half-century, these time horizons have begun to shorten, probably as a result of ever rapid change and aging populations which have accumulated enough experience to see how different time horizons can complicate our picture of the future. Moreover, during this period, the role of prediction in social decision-making has been markedly revised. With the rise in complexity theory, the notion that the future can be predicted in any sense at all has come under enormous scrutiny.

There have always been routine services in cities that have been subject to automation, and emergency services have been at the forefront of these activities for 50 years or more. In the late 1960s, the New York City RAND Institute was set up to develop various urban operations research tools for scheduling emergency services, particularly police and fire, but it is widely agreed that the problem they faced was mapping the actual nature of these urban services, the working practices of those empowered to implement them, and the politics of resource allocation through to successful delivery (Flood, 2011). Automated tools for solving such problems had to be embedded into this wider context, as ever a painful, painstaking process of organizational change. In short, the basic problem of implementing new information technologies in urban contexts is that it is essential to learn about the actual and optimal organizational structures of how such technologies need to be developed.

Indeed, the development of computer services in any kind of complex organization is fraught with difficulties, and the experience with large IT projects in the public domain is poor, almost everywhere. Fifty years ago, the RAND Institute failed in its mission in New York City largely because it was impossible for the bureaucrats, politicians, consultants, civil servants and operators to work with one another and the imposition of a layer of expertise on top of existing practices was simply too much for the system. Townsend (2013) documents this rather nicely in his recent book on *Smart Cities,* and it is clear from any considered analysis that these organizational issues are uppermost in the development of computer tools and infrastructures for making cities smarter (Batty 2014).

The most high profile example of the real time city at present is the "Operations Center" which IBM has set up to monitor short-term responses to crises in the city of Rio de Janeiro (Singer 2012). In response to recent crises – particularly landslides due to flooding and to the international events either staged or proposed for the next 2

years (the World Cup and Olympic Games) -- the city and its mayor proposed that the Center might monitor and control emergencies based on weather (of course), traffic, police and crime, and public health. The Center is designed as a large visual interface to each of these subsystems of the city with a trained staff of analysts and observers watching how these systems perform in real time and taking action if and when these systems begin to fail. In one sense, these functions already exist and have simply been collected together under one roof, but as they are being informed by networks of sensors across the city, there are clear advantages to some integration. However, the real value of such a center must be in how it is linked to actual physical responses and the control center is simply the interface to many other intricate networked systems where physical and human resources are delivered to enable appropriate change to be engendered.

In fact, there are remarkably few infrastructure systems in cities that are subject to real time computer control. Traffic is one example where progress is rapid but, as we noted earlier with respect to the London control center, this is still largely dominated by human image processing and manual styles of management because there are few algorithms other than those that are used to control wide area traffic signaling for dealing with the system on a comprehensive basis. The fact that well-organized agencies such as Transport for London are as of yet unable to link their demand data to their supply for purposes of control is not only because of the difficulty of finding common keys. The state of the art in real time traffic control is still quite primitive. Despite the fact that there is considerable hyperbole now about autonomous vehicles, the feasibility of such control and management is problematic any time soon. What is in fact happening is that an ecology of related but not integrated tools and methods is being fashioned, often around open data, that allows developers to build applications that individual users can employ to find out information about such systems and act on this information in personal ways. There are promising developments in terms of real-time control such as in individualized navigation systems. An example is Waze (https://www.waze.com/), which combines crowdsourcing with real time updates from in-car devices and Sat-Nav, but the movement to automation in real time is slow. There are simply too many human issues that make such automation difficult. In Figure 6, we show the example of TfL's open data interface and a small sample of some of the applications that have been developed for users making use of this data to inform intelligent travel decisions.

Figure 6: The Open Data Interface Provided by Transport for London with a Sample of Independent Apps – Tube, Bikes, Bus – from this Data

In other domains, progress is even slower. Although utility systems are fast being automated initially for maintenance purposes and smart metering is in prospect, the notion of acting intelligently in controlling energy distribution in real time in buildings, for example, is limited by the fact that most of the routine functions in cities are operated by individuals who are not coordinated. The coordination comes from emergent behaviors that are structured according to principles of individual competition under differential resource constraints. Indeed, even where providers such as municipalities are automating routine services that they are mandated to deliver, the responses they seek are not coordinated for they are individual actions that drive the use of such systems. This is what makes participation partial. These limits extend to any form of crowdsourcing, i.e., any form of social media that is doubtless generating massive volumes of data about usage and preferences but which is and never can be representative.

These then are the limits to the real time city. There is some prospect that new forms of data of a more abstracted nature and useful for more strategic planning is becoming available in real time but often on cycles that are much slower than the sort of data that is streamed incessantly (Batty 2013). A lot of new data is being streamed, but the frequency of change is much slower than faster systems involving continual movement of people or energy. For example, house prices and related transactions, migration into and out of cities, updates to the geometry of cities in terms of maps and other physical content – all these are on the horizon and rudimentary forms of application and their data are now available. In the next

decade, we will see substantial progress in this area as big data streamed in real time begins to reveal insights into much broader and perhaps more important questions as to how cities and their quality of life are evolving.

V. THE PROSPECTS FOR BIG DATA IN THE SMART CITY

We have talked about the all pervasiveness of computers and computation in modern life for the last 50 years, but changes in digital technology never cease to surprise. Yet the problems of actually implementing these technologies appear extremely problematic in comparison with simpler technologies developed in the mechanical and to an extent electrical eras. Most technologies that we have invented during the last 200 years are robust in that they admit levels of tolerance in terms of their workability that do not lead to widespread breakdown. But it is the social consequences of these technologies that are the most significant for the subject matter of this paper, cities. There are many important challenges that are posed by the spreading out of computers into the public domain, and these need urgent resolution if big data is to yield the kind of promise that has gripped the field. By way of conclusion, it is worth summarizing the key challenges that we have raised.

The first challenge is the question involving the lack of structure in many sets of big data. Because such data is usually collected for purposes other than the kinds of analysis implied here, big data often lacks the kind of structure that analysis requires so that the pattern and structure in such data can be exploited in terms of our understanding. Data which is streamed in real time often has no filters placed on its form and thus it is highly descriptive of the operation of some system. For example, the Oyster data that we have described here gives time and position and fare status of a traveler but cannot be linked in any way to other personal attributes. Such linkage would be necessary if the data were to be used to target passengers in different ways, for example, making their experience of the trip more pleasant and advising them on how they might improve their travel. It is often remarked -- most significantly by Anderson (2007) in a highly controversial article where he argued that the rise of big data heralded the end of theory -- that all one needs to do is search for patterns in big data and that once these are found through exhaustive data mining techniques, then all will be revealed. In fact, this is quite false; if one approaches data with no prior conceptions about what it is and what it means, then it is unlikely that one will derive any appropriate meaning from it (West 2013). Lazer et al. (2014) in their comment on

the failures of *Google Flu Trends*, once hailed as the great example of how one can mine "big data" -- in this case, using Google search terms correlated with "flu" -- point to the inability of the algorithm to distinguish between things like "winter" and "flu." They conclude that in terms of big data, "we are far from a place where they can supplant more traditional methods or theories."

This leads to our second challenge and this involves integration, the search for a common key that will link more than one data set together. In the geospatial world, the common key has been the address point or the coordinate reference. For many years, geographic information system technologies have continued to refine such systems so that diverse data sets can be linked together. In fact, such address coding has advanced to the point where names rather than numbers and seemingly out-of-sequence number sets used in different countries and cultures can now be dealt with effectively. But if there is no common spatial key between two data sets, then, other than manufacturing a synthetic key from independent data, which is occasionally possible, there is simply no way such data can be integrated. In unstructured data, there are far fewer possibilities for integration anyway, and this is likely to remain one of the major obstacles to the use of big data in the context of the smart city.

Our third challenge relates to the organizational structures that are needed in cities to exploit big data and the analytics that is able to unpack them. In the development of science in human affairs, particularly in urban and social policy analysis, the organizational structures that determine how decisions are made and how human systems function definitely fall under the banner of complex systems. Complex systems are inherently unpredictable in that they are built and function from the bottom up with coordination and often their sustainability a complex web of political, social, competitive, and conflicting actions. There is the tendency to assume that every new technological development will not follow the same path as before, but it appears that many of the efforts in developing big data for smart cities are likely to face the same problems as those faced 50 years ago when the technologies produced for the space program and the Cold War were imported into municipal government (Light 2003; Szanton 1981). Although problems of integration and structure in big data are legion, problems of using it and related analytic technologies are even more significant. What is urgently required is a mapping of these technologies onto the practices and structures in which decisions are made in municipalities and city government and how these interface with the many other agencies that have a stake in the smart city.

These are the challenges and it is no accident that the most successful developments to date are those that are bottom up – for example, Transport for London's opening of their data to provide developers with free data for Apps – rather than the top-down control centers and the new smart towns where it is not easy to map these new technologies to the organizational and political nexus that cities depend upon. The challenge with big data for the smart city is not simply technological, but more organizational and political. It requires developments on both fronts for progress to be made and for the potential of big data to be realized.

REFERENCES

Anderson, C. 2007. "The End of Theory: Will the Data Deluge Make the Scientific Method Obsolete?," *Wired Magazine,* July 16. http:// www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Batty, M. 2013. "Big Data, Smart Cities, and City Planning," *Dialogues in Human Geography* 3, no. 3: 274-279.

Batty, M. 2014. Commentary. "Can It Happen Again? Planning Support, Lee's Requiem and the Rise of the Smart Cities Movement," *Environment and Planning B* 41, no. 3: 388–391.

Batty, M., Axhausen, K., Fosca, G., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., and Portugali, Y. 2012. "Smart Cities of the Future," *European Physical Journal Special Topics* 214, no. 1: 481–518.

Ferry, G. 2012. *A Computer Called LEO: Lyons Tea Shops and the World's First Office Computer*. London: Harper.

Flood, J. 2011. *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City—and Determined the Future of Cities*. New York: Riverhead Trade.

Greening-Jackson, T. 2012. "LEO I and the BR Job," available at https://docs.google.com/file/d/0BwUohGCPTAlANlFrQ3M3TnIyZjg/edit?pli=1.

Kitchin, R. 2013. "The Real-Time City? Big Data and Smart Urbanism," *GeoJournal* 79, no. 1: 1–14.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. 2014. "Big Data: The Parable of Google Flu: Traps in Big Data Analysis," *Science* 343, no. 6176: 1203–5.

Light, J. S. 2003. *From Warfare to Welfare: Defense Intellectuals and Urban Problems in Cold War America*. Baltimore, Maryland: Johns Hopkins University Press.

Munizaga, M., and C. Palma. 2012. "Estimation of a Disaggregate Multimodal Public Transport Origin-Destination Matrix from Passive Smartcard Data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies* 24, 9-18.

Singer, N. 2012. "Mission Control, Built for Cities: I.B.M. Takes "Smarter Cities" Concept to Rio de Janeiro," *The New York Times*, March 3, 2012, available at http://www.nytimes.com/2012/03/04/business/ibm-takes-smarter-cities-concept-to-rio-de-janeiro.html.

Szanton, P. 1981. *Not Well Advised: The City as Client—An Illuminating Analysis of Urban Governments and Their Consultants*. New York: Universe Publishers.

Theophilus, M. 2014. "Surface Transport and Traffic Operations Centre (STTOC)," available at http://spatialcomplexity.blogweb.casa.ucl.ac.uk /files/2014/06/ Surface-Transport-and-Traffic-Operations-Centre-STTOC_Marc-Theophilus.pdf, accessed at 02/06/2014.

Townsend, A. 2013. *Smart Cities: Big Data, Civic Hackers and the Quest for a New Utopia*. New York: W.W. Norton & Co.

West, G. F. 2013. "Big Data Needs a Big Theory To Go With It," *Scientific American*, May 15, http://www.scientificamerican.com/article.cfm?id=big-data-needs-big-theory, accessed 04/06/14.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., and Schmitt, G. 2014. "Detecting the Dynamics of Urban Structure through Spatial Network Analysis," *International Journal of Geographical Information Science* 28, no. 11: http://dx.doi.org/10.1080/13658816.2014.914521.