Data About Cities: Redefining Big, Recasting Small

Michael Batty[1]

Abstract

In this paper, we argue that the development of data with respect to its use in understanding and planning cities is intimately bound up with the development of methods for manipulating such data, in particular digital computation. We argue that although data volumes have dramatically increased as has their variety in urban contexts largely due to the development of micro-devices that enable all kinds of human and physical phenomena to be sensed in real time, big data is not peculiar to contemporary times. It essentially goes back to basic notions of how we deal with relationships and functions in cities that relate to interactions. Big data is thus generated by concatenating smaller data sets and in particular if we change our focus from locations to interactions and flows, then data has faced the challenges of bigness for many years. This should make us more careful about defining what is 'big data' and to illustrate these points, we first look at traditional interaction patterns – flows of traffic in cities and show some of the problems of searching for pattern in such data. We then augment this discussion of big data by examining much more routine travel data which is sensed from using smart cards for fare-charging and relating this to questions of matching demand and supply in the context of understanding the routine operation of transit. This gives us some sense of the variety of big data and the challenges that are increasingly necessary in dealing with this kind of data in the face of advances in digital computation.

A Paper Prepared for the Data and the City Workshop
The Programmable City Project at National University of Ireland, Maynooth
(http://www.maynoothuniversity.ie/progcity)

August 31-September 1st 2015

[1] Michael Batty is Bartlett Professor of Planning at Centre for Advanced Spatial Analysis (CASA), University College London (UCL), 90 Tottenham Court Road, London W1T 4TJ; m.batty@ucl.ac.uk; @jmichaelbatty; www.complexcity.info

Introduction

Data with respect to cities only became significant after world war 2. This was largely although not exclusively driven by developments in digital computation and statistical theory, first in the interwar years and thence much more aggressively from the 1950s on after the invention of the transistor. This went hand in hand with the development of ever bigger data volumes and related resources, as well as developments in many kinds of electronic media and communications technologies. Historically, as populations grew in western countries during the industrial revolution, it became more and more important to count them, albeit as much for taxation as other purposes of government. By the late 19th century, automated machines were first used to process Population censuses, the most high profile example being Herman Hollerith's development of the punched card tabulator used for the 1890 US Census which led ultimately to the formation of the IBM company.

The Population census was in fact one of the only systematic catalogues of data produced on a continuing basis until national accounts and related economic data began to be collected seriously and routinely in the 1920s (Bos, 2011). But right from the start, data was always big with respect to the available means by which it could be manipulated. There is a wonderful story from the 1950s about the use of spare cycles from the early computers developed for the Lyons Tea Company (Ferry, 2010) where these computers were used to compute shortest routes for freight in the rail system so that British Railways could price these goods accordingly. Dramatic and ingenious manipulations had to be devised to make this possible such as noting that to stuff the data into the needed memory, Scotland needed to be treated separately from the rest of Britain and then stitched back together after separate computation. In the process, those involved actually invented the well-known Dijkstra algorithm a year before Dijkstra did so himself and four years before he published it (Graham-Cumming, 2012). There are countless examples such as these in the history of computing and data during the last 75 years which show how the limits of computation were reached and new algorithms and data mining techniques invented due to volumes of the then big data.

So big is a relative concept and some data has always been big with respect to how it might be manipulated using state-of-the-art computation. But apart from the sheer volume of data, in cities data has always been big in another sense. As I argue elsewhere (Batty, 2013), our concern is no longer and indeed never has been exclusively with location but it is with interactions: relationships between locations which are best measured by flows that have volume and direction. The volume of data contained in flows is in general the square of the elements that define the locations between which the flows are generated. If there are n locations, then there are n^2 possible interactions between them and thus the data associated with interactions increases exponentially as the number of locations increases or as locations get finer and finer in terms of their resolution. In this paper, our contention will be that big data can be generated from small through interactions, and that higher order effects and much of what we might consider traditional data in city systems is in fact big data. Although we will

definitely not conclude that the big data revolution is a red herring, we will conclude that bigness is never what it seems and that bigness in terms of computational time taken to explore data which might be quite small in volume is as important as dealing with massive data volumes. In fact currently the most important challenges involve not big data *per se* but computational schemes for making sense of both small and big, with a focus on extracting meaning from both big and small. As we will see, dealing with quite modestly sized data sets can lead to a proliferation of strategies for simplifying this data, and thus our focus here will be on techniques that enable us to see patterns and order in different volumes of data and their interactions.

Classifying City Data

An early typology that has withstood the test of time was introduced by Berry (1964) whereby he defined what he called a 'geographic matrix'. This consisted of an array of places – locations – and their attributes which he called characteristics. Such a matrix he argued was the essence of geographical analysis in that the two dimensions of place and their characteristics defined the central qualities of location. To this he added a third dimension, that of time with this being the third dimension of the matrix but one which did not have the detail of the other two. In fact, he envisaged these additional time slices to be of limited number but in principle, each of these dimensions could take on any number of categories. In one sense, although he did not use the term, the geographic matrix in its three dimensional form is close if not identical to what in data science is now called the 'data cube'.

Berry then proceeded to use this matrix to explode a spatial data set. In one sense, the focus was on place rather than its characteristics or its temporal positioning but by concatenating these dimensions, one might envisage a series of relationships in single, pairwise or in three-wise fashion. If we label characteristics by their volume as M, places as N, and time slices by T, then there are 7 possible combinations of relations: M, N and T by themselves, $M \otimes N$, $M \otimes T$ and $N \otimes T$, and then $M \otimes N \otimes T$. Unpacking these further[2], then we might consider relations between $M \otimes M$, $N \otimes N$, and $T \otimes T$. Significant for this discussion is the relation between N and itself which essentially is spatial interaction – linkages and flows between locations – but with flows across time between T and itself (noting that time is irreversible of course) and even relations between characteristics also being significant. Indeed traditional multivariate analysis has tended to deal with comparisons and correlations between characteristics in terms of place or places in terms of characteristics.

Berry's data cube as we will now call it (although we are unclear whether or not he ever used the term) is based on categories or types. The notion of continuity is

[2] The operator $\otimes$ is a concatenation symbol that includes several different ways of interrelating the variables that are concatenated. For example, $M \otimes N$ relates M to N through counting the number of instances of M in N, or vice versa, the intensity of M with respect to N and vice versa, and so on.

not particularly significant in this conception although it is entirely possible to think of characteristics as being splayed out on a spectrum – income for example – or places defined from continuous representations of the earth's surface. Time however is another matter. Berry (1964) considered this in geographical terms to be highly discrete with nothing like the richness of the other two dimensions but one of the characteristics that marks out new conceptions and origins of data – big data – in contemporary times is that time has literally exploded. Much new data is generated from devices and people associated with devices that sense the environment in real time and capture data down to the temporal resolution of the device. This in principle can be measured to the number of decimal points embodied in the hardware: 32, 64 bit and variants thereof (extended by parallel processing).

Berry's focus however was another kind of data explosion that comes from generating relationships between the dimensions. We will illustrate these here with respect to relationships between places – spatial interactions – which can also be tagged to quite fine resolutions of time. In fact it is important to be clear as to the way the data cube might be used in the analysis of city data. Even though it is based on three dimensions which can in fact be extended to many more, usually any analysis takes one of these as being the anchor point – place, characteristics or time – and conducts analysis with respect to relationships associated with this anchor. Although the data cube is generic, whenever data is considered in these terms, the problem is usually structured from one of these perspectives and thus it is important to see the size of data, its volume and its variety at least, in terms of the particular perspective adopted.

It is worth indicating how traditional urban data – urban populations collected from traditional sources such as complete Population censuses, for example – can explode into big data. This was possible long before the current era of big data and it is very clear when spatial interaction is considered. In 1964, Lowry built a state of the art urban model for Pittsburgh which divided the region into 456 zones between which the flows of people moving to work, shop and so on were collected. The data was collected from household interviews for traffic studies but the volume when considered with respect to the matrix of interactions was huge by the standards of those times – $456^2 = 207{,}936$ possible interactions (trips, distances, etc.). This was in an era when many mainframe computers could barely store more than 64K numbers and most of the transport models then built always pushed up against these limits. Indeed it was one of the main reasons for the enormous problems that were associated with the earliest urban models. Douglas B. Lee (1973) in his famous paper detailing the experience with these tools, defined the problem as one of data 'hungriness' (Batty, 2014). In fact, right from the beginning of digital computing, indeed even before with weather forecasting in the 1920s, data had always been big relative to the devices we had at our disposal to manipulate them. But the explosion which occurs when one concatenates data has always been there to explore and many of our tools have been developed with respect to such limits. Witness the description earlier of the way what came to be called Dijkstra's shortest routes algorithm and partitioning of networks to effect its use for problems too big for the then computer technology.

In the sequel, we will explore how these kinds of problem have got bigger in terms of their data requirements examining how we might deal with matrices with millions of entries which explode from spatial systems defined by thousands of locations. Before we do so, however, we should bring the story up to date with respect to spatial interactions by introducing recent developments. These truly do broach the question of how big is big and what tools and techniques do we now need that we did not have in the past with respect to the spatial analysis of such city systems.

The Emergence of Big Data In and For Cities

Traditionally urban data sources that measure the various characteristics of places, say, tend to be spatially aggregated in the way they are made available, although as the level of disaggregation increases towards the individual level, temporal considerations do come to the fore. In fact for many years, some individual data has always been collected but only in cases where confidentiality is unimportant is this data available. The best examples pertain to traffic where inductive loop counters and like systems have been embedded in road surfaces to count volumes of traffic as individual vehicles and this data can become quite voluminous. In fact although this is big data in the contemporary sense, in the past it has been aggregated and filtered to meet the restrictions of the analytical methods – transport and traffic flow models for example – that are available for its interpretation.

Before we turn to examples of big data, it is important to get some sense of what this term means for it has only become significant in the last decade or so. This has coincided with the development and dissemination of countless digital devices that sense characteristics of objects in the physical environment with respect to their type, positioning and time when they are observed. These are of course the three dimensions of our data cube and big data thus tends to be data that is dimensioned in at least these three ways – by their attributes or characteristics, by their spatial positioning or location, and by the time instant at which the relevant objects are observed. The objects can be human or physical, indeed of any type as long as they are associated with a relevant sensing device, and very often, if the object is human, then the sensing device has purpose in that it can be activated by the person or it can remain in passive mode.

There are many definitions of big data. The cliché is that big data is defined by its volume, variety, velocity, veracity and value. This simply roots the data in questions of size (bigness), variety (diversity and extent), velocity (temporal frequency of collection or observation), veracity (level of accuracy and/or uncertainty), and value (what it brings to various purposes) but it might be countered that all these criteria apply equally to small data. However the implication is that it is size, scale and scope that pertain to these characteristics (http://www.ibmbigdatahub.com/infographic/four-vs-big-data/). In fact big data is much more than these four or five 'Vs'. Dutcher (2014) has collected together some 40 definitions from 'thought leaders' across the industry at the DataScience Berkeley Blog (http://datascience.berkeley.edu/what-is-big-data/)

and one of the main characteristics that comes from this sampling of expert opinion is that big data is more about the tools that are needed to process it – to understand it – than its size or volume.

Often big data is hard to understand because it has little structure, it is sometimes but not always large, and traditional tools are very difficult to use in its processing. For example, very large quantities of household census data although not any larger in the volumetric sense than they have been in the last half century, often stretch and confuse the traditional multivariate techniques that we are accustomed to. Even plotting a scatter diagram relating, say, population income to level of education at the individual or household level for a country the size of the UK requires visualisations of more than 20 million points and most if not all statistical packages and even statistical interpretations break down with such volumes. These by the way would not be regarded as big data by contemporary standards for the usual rule of thumb is that the data must be giga- and upwards in size for it to be classed as big data. This would rule out most census data at the individual level but as we have argued, special data mining techniques are usually required for data that is in the tens, hundreds, and thousands of thousands.

Big data which is streamed in real time thus represents the cutting edge of new data about the functioning of cities. Much of this data is streamed from devices that are simply embedded in the physical environment and transmit data in continuous fashion with little human interference or management. Loop counters in roads are a classic example but many related analogue devices have been used for many years to record aspects of the weather, the use of energy, breaking news and so on, much of which is captured in the various dashboards that have been set up to pull together such data and make it intelligible to interested observers and policy-makers. These dashboards have mainly been produced so far to demonstrate that by pulling such data together one can gain an immediate impression of the state of the city (O'Brien et al., 2014; Kitchin et al., 2014). In fact the synthesis that is required to make sense of this is very hard to develop as many of the data sources cannot be integrated in any way. Moreover much of this streamed data in real time reflects very different concerns for cities from more traditional data sets. For a long time, there has been concern with routine functioning of cities in terms of traffic, crime, policing, the delivery of emergency services and so on and models that enable predictions of routine and not so routine events have been a major concern to city government. But these operational research types of model and their data do in fact function in real time – usually daily time spans – and in this sense, this sort of data used for these is being enriched by better and more comprehensive sensing.

Real-time data pertaining to the socio-economic structure of the city is much more problematic to collect using sensing devices. Unambiguous answers to queries which involve the human condition are almost impossible to link to real-time sensors. Information on people's choices are fraught with difficulty in terms of collection and interpretation. The reason why so much data in real time is transit data is that travel is a relatively routinised activity whereas collecting

data about unemployment, income, employment activity, migration and so on requires human and related agencies to put in place systems where people are required to answer or register. Some data is being picked up in retailing with respect to sales data from smart, credit, loyalty cards and so on but invariably where this data is collected (and sometimes available) in real time, various sensing devices are used. Data which is compiled from registrations are in fact being made available nearer to real time such as house prices, and related area accounting data. In these cases, the frequency at which such data is produced is hardly real time – monthly at best to date – but this kind of data depends on the frequency of changes – people make changes in these phenomena over matters of days and weeks and months rather than seconds and minutes. A good example of where such data is being linked to dashboards is the Amsterdam Board which does contain such socio-economic data as well as mapping and limited GIS/mapping functionality (Batty et al., 2015).

To illustrate these issues, we will focus on transport where data is intrinsically big in terms of traditional data collected from questionnaires about travel patterns administered to individual travellers or to households, in terms of new sets of travel data gathered from smart card usage for collecting fares, from real-time movement data transmitted from vehicles themselves, and from data captured from monitoring of passengers through automated observations. Not only is transport data big in the sense that much of it deals with how travellers move between origins and destinations thus generating spatial interactions, but it is also big in temporal terms because using automated methods it can be captured continuously. In this sense, too, we are able to look at transport data that pertains to very short as well as much longer time periods, with consequent implications for the use of this data in different types of planning and management. We will now turn to these examples.

Traditional Transport Interaction Data: Big Data Generating Complex Visualisations

Ever since transportation planning formally began in the 1950s, the focus has been on potential interactions or flows between origins and destinations. Different types of traffic form the essence of transport models usually based on different modes but the class of models that we will allude to deal with many other kinds of flow from social networks, to input-output trade relations, to patterns of migration and so on. The concatenations that we are focussing on here are flows between places, that is $N \otimes N$ which generate travel volumes that can be substantial as the number of places N increases as we noted above for the first land use transport models such as Lowry's (1964) model of Pittsburgh. Until quite recently, visualising flows has been stymied by constraints imposed on graphics as much as by the size of the data. To consider the nature of the problem, in Figure 1(a) we show London divided into 33 separate but contiguous zones for which a journey to work matrix – flows from any zone which is a borough to any other – is almost impossible to plot clearly. 33 zones generates a total possible number of trips $33^2 = 1089$ which may not appear to a large number but is very hard to plot clearly. We show this plot in Figure 1(b) where it

is very clear that where one is plotting all links from any zone to another, but excluding the intra-zonal trips and also suppressing the asymmetry of the matrix where a trip T_{ij} which is the flow from zone i to j by adding the flows as $T_{ij} + T_{ji}$, still produces a flow map which is hard to interpret. Plotting individual trips from one origin to all destinations is the only way to make the map clear but we get no sense of the polycentricity of the system from this visualisation and this is what we really need to detect in the data.

Figure 1: Total Two-Way Trips a) The Zoning System b) All Trips Plotted c) Trips Associated with Westminster (The Centre), d) Trips Associated with Hillingdon (Heathrow). Note that Intrazonal Trips are not Plotted.

Now this is a very crude characterisation of journey to work in Greater London. Even 50 years ago, we would not be content with this level of resolution and therefore we will work with a much bigger data set composed of dividing these 33 zones into their constituent wards – local electoral districts which have on average around 10,000 resident populations living within them. There are 633 such zones and immediately the data has exploded to $633^2 = 400689$ potential interactions which is quite large – not quite half a million but a large number for any kind of statistical analysis. In fact we usually calibrate a model to this kind of data so that we predict each of these flows. In fact many of the flows for a system of this size and resolution will be small and quite a few zero in terms of the observations but there will always be a total number of flows predicted no matter however small. And this means that we have to face dealing with the complete matrix.

In Figure 2, we show the new more disaggregate zoning system and in fact we have to make the areas a little larger to show this level of detail on the page. It is not worth showing a plot for the full trip matrix as this is simply a mess with no way of detecting the complexity of the physical form. What we want to do is detect how close different patterns from different parts of the metropolis are and a first way into this problem is using visualisation. The notion of examining trips origin by origin or destination by destination is an obvious way forward and we do this in Figure 2(b) and (c) as we did in Figure 1 for the coarser resolution system. Aggregation and animation are ways of dealing with this data in terms of building up a structured understanding of this complexity but the problem really becomes serious once we wish to test comparisons and compute correlations

between the observed trip matrix and any other matrix such as a predicted one. Then the nature of the problem begins to change as we need to plot all the points to find a single relationship and to show how this kind of problem explodes into big data which need new methods, we will compare the 633 x 633 matrix with one that is predicted by the model.

633 Zones Based on London Wards

Trips from and to the Ward West End

Trips from and to Heathrow Airport

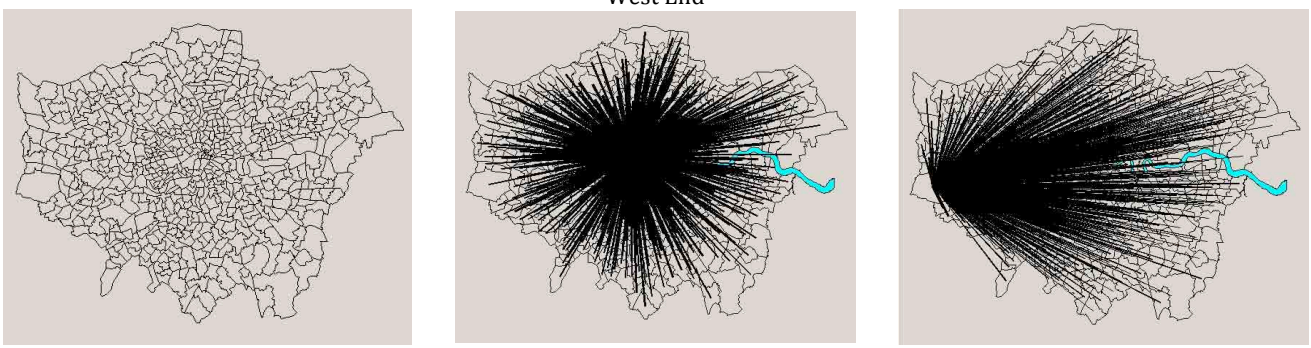

Figure 2: Total Two-Way Trips a) The Fine Scale Zoning System b) Trips Associated with an Inner City Ward c) Trips Associated with Heathrow Airport

We must now say a little about the model that we use to predict the data that we observe in the patterns shown in Figure 2. Data about cities is never far away from analytical techniques and simulation models and without losing the reader, we need to say a little about the nature of the model that we will build to produce predictions which can be compared against the data in Figure 2. The model predicts trips T_{ij}' between origins O_i^{obs} and destinations D_j^{obs} which are then compared against observed trips T_{ij}^{obs}. Observed origin and destination volumes – O_i^{obs} and D_j^{obs} – are computed from the observed data as $O_i^{obs} = \sum_j T_{ij}^{obs}$ and $D_j^{obs} = \sum_i T_{ij}^{obs}$. The model is an unconstrained gravity model that computes predicted trips as a function of the observed origin and destination volumes and an inverse functions of distance d_{ij} between each origin and destination pair. The model is $T_{ij}' = K\, O_i^{obs} D_j^{obs} \exp(-\beta d_{ij})$ where K and β are parameters that meet normalising constraints. From the model, we clearly derive predicted trips but also predicted origin and destination totals $O_i' = \sum_j T_{ij}'$ and $D_j' = \sum_i T_{ij}'$. To measure how good the model fits the data, we need to examine the scatter plots which contain the correlations between O_i' and O_i^{obs}, D_j' and D_j^{obs}, and T_{ij}' and T_{ij}^{obs}.

The scatter plots for origins and destinations are easy enough to visualise as there are 633 observations in each. However for the trips, there are a possible total of 400,689. However in terms of the observed trip data some 64% of these are zero observations. As the data is taken from a 10% sample, this poses a

problem. Should we compare zero cells with predicted ones which will always be positive and should we compare cells with a fractional number with integers? If we exclude the zero cells, then we still have some 142291 to deal with, implying that only 36% of our data matrix is occupied. This does not change the nature of the problem in terms of its visualisation and the search for pattern which we illustrate in Figures 3 and 4.

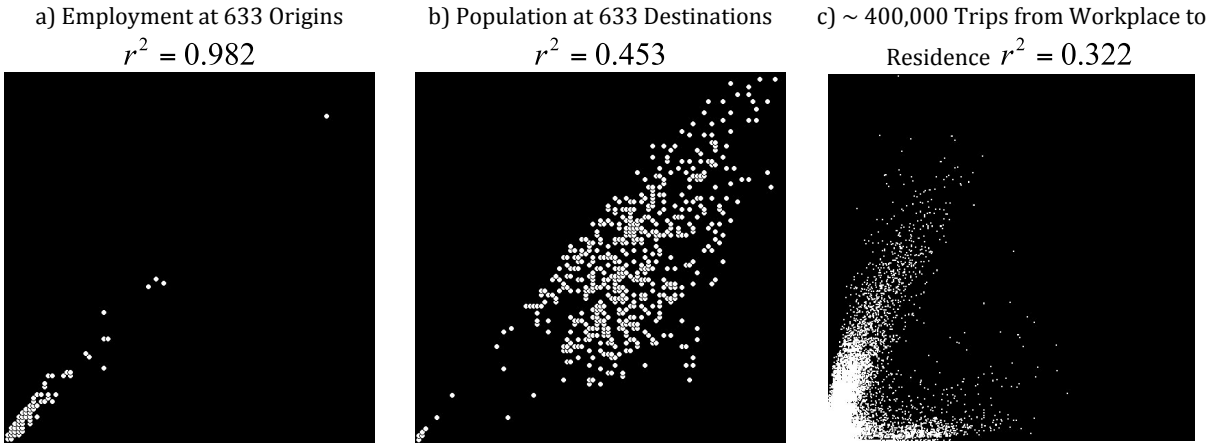

Figure 3: Predicted Against Observed Data a) Origin Employments b) Destination Working Populations, and c) Trips from Work to Home

Figure 3 is revealing. The three scatters are very different with employment being predicted rather well, residential population less well, and trips showing that there are at least two regimes characterising travel in London. In fact, the scatter of trips in Figure 3(c) reveals a clear density map and in Figure 4 we show this as best we can. The intensity of very small trips is much greater than larger ones for the distribution of trip volumes follows some sort of power law. In Figure 4 we have blown up the lower portion of the scatter to reveal this intensity and this reveals that this kind of data mining must be supplemented by many other kinds of visualisation and analysis so that the true patterning of a system with this kind of complexity can be laid bare.

Figure 4: The Density of the Scatter: Different Patterns at Different Scales

Now all this may not look very much like big data but our current extensions of these models are equivalent to an entire systems of cities at the same level of resolution as the Greater London model in Figure 2. We are working on a model with 7201 zones – middle layer super output areas (MSOAs) which have an average population for England and Wales of 10,000 and average number of households of some 4000. Our model is built for all these zones and immediately there comes a problem of visualising the scatter of origins and destinations as well as trips of which there are a total possible cells in the matrix of $7201^2 = 51{,}854{,}401$. Visualising nearly 52 million points on a scatter graph is well beyond our capabilities and although only 10 million or so of these points are likely to be above zero[3], this is still beyond the capabilities of this kind of analysis. We show the zoning system in Figure 5(a) and when we move to flows, it is impossible to use the single origin many destination tool to visualise a set of flows one by one. What we have done here and this illustrates the judicious choice of visualisation is to produce a single flow for each origin to all its destinations using a weighted directional vector. For each origin i, we compute the average vector $[\vec{x}_i, \vec{y}_i] = [(x_i y_i), (\Sigma_j T_{ij}[x_i - x_j]/n, \Sigma_j T_{ij}[y_i - y_j]/n)]$ which gives us a single arrow that computes the average strength and direction of the flow. Much information is lost in our visualisation but in the system we are developing, there is zoom capability that is able to illustrate the overall pattern at a coarse spatial scale and the detail at the finest scale of the zones themselves. We show the coarser visualisation for England and Wales in Figure 5(b).

a) The Zoning System for England and Wales Based on MSOAs

b) Average Directional Flows from Population Centres to Employment in E&W

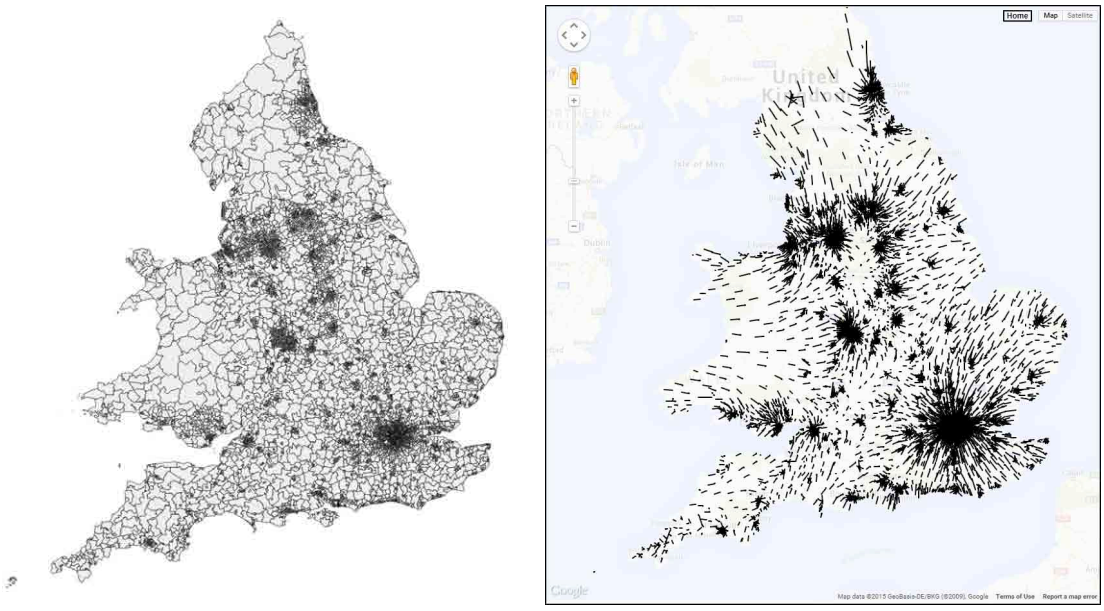

Figure 5: Visualising Big Data Based 10s of Millions of Transport Flows

[3] This is a guess. We have not computed the sparsity of this 52 million cell matrix but by the time of the meeting, I will have these numbers.

This has been possible in terms of data available for the last 30 years or more but only now that we have computers large enough are we able to exploit the bigness of this data. This is very different from the big data that we will present in the next section where the volume comes largely from the temporal and individual rather than spatial dimension. It does reveal, however, that big data has been with us for a while and it is computation more than anything else – the fact that we can now manipulate it in diverse ways – that is one of the ultimate determinants of the size of data set that we can handle, interpret and use fruitfully.

Real-Time Streamed Transportation Data At the Micro Level

Since the 1950s, data has been collected in continuous time specifically for traffic flow analysis. Much of this data has been hard to link to origin-destination data of the kind just examined largely because it is supply-side data pertaining to vehicular movement and not to intentional trip-making. However with the advent of RIFD and related technologies[4], it is now possible to collect data on where people enter and exit a transit system or where they embark and end any journey if the relevant collector is in place. Devices which are specially devised for data collection in question are by far the best as the data that they produce is unambiguous (although there may be substantial noise still to be filtered out). Mobile devices for other purposes such as phones can also be used to extract data from call records which locate the phone when a call is made. Such data is being used as proxies for travel. This data is much closer to origin and destination data of the traditional kind as long as there is some certainty as to what constitute origins and destinations, and many new directions are emerging which seek to transform this kind of data into useful purposes for urban analysis (Chen, Batty and van Vuren, 2015).

Because this data is recorded at the exact time when the smart card or mobile device is linked to the system in question, there is a continuous or at least a continual record of activations which represent real-time collection, if not streaming to some archive that is accessible in real time itself or for *post hoc* analysis. In short, the data is as voluminous as the number of activations. If this is phone calls, then it is the number of calls made from that device per day or over whatever unit of time and space the data is made available or aggregated to. Here we will use data that we have from the RFID smart card which is in use on all public transport in Greater London called the Oyster Card. This card stores the money that travellers use to pay for journeys and the system is designed to recognise the category of payer as well as the time and place where the traveller taps in or out of the system. Travellers tap in and out on trains but only tap in on buses.

We have several tranches of data from this system. Our largest set is for 86 days in the summer of 2012 (16 June to 9 September which covers the period when

[4] RFID – Radio Frequency Identification: wireless devices that transmit data from some networked or standalone system that can generate data.

the 2012 Olympics were held) where there were 9,902,266,857, nearly 10 billion taps. Of these taps, 44% were on buses and 56% on rail which is tube and overground with some mainline network rail. As there is only tap-in on buses, we can guess that if round trips are made by rail, then this is about half of all rail trips meaning that there are about 60% more bus trips than rail. This is notwithstanding the fact that our Greater London transport data from the 2001 Population census records that bus travel is about 60% less than rail travel. But in the last 15 years with congestion charging and differential costing of these two modes, we think these figures show a reversal of patronage. All modes of course have increased due to population growth. The data shows that 11,535,090 different Oyster cards are used for these 10 billion taps which is 86 taps per unique card, on average about 1 tap per card per day. This does not of course account for people using more than one card.

This data is quite unstructured. It comes as a flat file where each tap is recorded by place and time – subway station, location of bus by stop etc., and some classification of the traveller such as whether the card is free (over 60s in age, disabled traveller and so on), and what payment category is active on the card. Generally it is possible to trace the behaviours of an individual card holder through time and space, in this case over 86 days. The degree of heterogeneity in the data set is enormous and this is feature that makes it usable for all kinds of temporal modelling at the level of the card holder conceived of as an agent. However there are critical problems with this data. The analysis of one day's worth of data in November 2010 from a series we have of 3 weeks data for the 660 tube and overground stations revealed that 6.2 million travellers tapped in but only 5.4 tapped out. Essentially this was because barriers were up. A large class of Oyster users with free passes are not fined for not tapping in or out while season ticket holders are also not fined as their cards are loaded with a fixed amount of money for a period. This is quite a large loss of data. If you combine this with travellers using more than one card, then this confounds the data set for transport analysis.

There are 660 rail stations and over 19,000 bus stops and it is possible with some analysis to figure out how many journeys are made by tracing different travellers in terms of the tap-in and -out activity during the working day for rail at least. We have attempted some analysis of buses with respect to travellers who have a unique identifier and who hop onto buses and trains within a certain time interval which we assume captures some multi-modal journeys but our analysis is limited and our confidence in extracting multimodal journeys in general is low. In terms of the rail system, we are able to produce distinct trips in terms of segments although the analysis of round trips is more limited. For example in the 2012 data, we can identify 291 million trips between one station and another in terms of a tap-in and tap-out with the most popular segment in the system the trip from Victoria to Oxford Circus and vice versa. Waterloo to Canary Wharf is the most frequent during the morning and evening peak with Waterloo and Victoria the two biggest volume hubs in the system. There is much data and near data analysis of this kind that we can engage in with these large data sets and one of the tantalising prospects of big data like this is the analysis

of regularity and heterogeneity in such data, notwithstanding the much deeper challenges of connecting this data up to origins and destinations.

In understanding cities, origins and destinations of trips, indeed of any flow, is essential for understanding the rationale of the location where those creating the flow are based. This relates to ongoing activities which are reflected in economic or social features of the location which in turn are represented by land uses, building types and other physical aspects of the composition of the place. One of the problems with smart card data that is orientated to transit systems such as fixed rail is that the locations which anchor these infrastructure do not have the same meaning as origins and destinations in terms of work, shopping, residences, schools and so on which generate trips. It is extremely difficult to tie places where people enter such systems to the comprehensive patterns of locations that are described by traditional data. We can quite easily assemble flow matrices and assign trips to network segments such as lines between stations – although the precise paths of travel have to be inferred, but tying these to places of work, residence and so on is difficult. Some headway has been made using smart card data for Singapore (Zhong et al, 2014) but the problem is perennial and requires additional data to link points of fixed infrastructure to ultimate origins and destinations. Synthesising or integrating such big transit data with origin and destination data from household surveys will always be problematic. The analysis of phone call data has related problems of identification with respect to the functions of the locations from which phone calls are made and the locations to which they are sent. This is not simply a question of linking the cell towers to the actual locations of the phone users but the reason why the phone calls are being made in the first place. This is an ongoing issue with big data which is streamed from sensors and mobile devices.

We have assembled several pictures of transit systems in operation from our Oyster Card data. Jon Reades worked on finding shortest routes between stations identified in the data and pieced together actual flows by assigning origin data from tap-ins to the network by finding the shortest routes on lines linking the origin to the destination – the places where the traveller tapped out. He has produced a computer movie of a typical week from the 2012 data by adding data for several typical weeks – excluding the Olympic Games weeks – and producing an averaged version which shows the peaks and troughs in the data from Sunday to Saturday. The weekend days are very different with much less pronounced morning and evening peaks while typical workdays show very distinct morning and evening peaks that in themselves are very different with a small blip in the central area in the late evening – the entertainment peak. You can see the movie by clicking on the caption to Figure 2 which shows snapshots from the movie made by UCL Engineering which is on **YouTube**. The actual computer animation by Reades (2013) is shown on **Vimeo** at https://vimeo.com/41760845

We are developing several projects using the Oyster Card data but so far these tend to examine very different aspects of the city from those that pertain to traditional flow data. The focus is inevitably on questions of disruption and smooth flowing on a fine scale temporal basis but we are not able to relate these to links between home and work. We are able of course to examine the

variability of the tap-in and tap-out data with respect to the station hubs through two interlocking patterns of entries and exit volumes that reflect two layers of polycentricity which vary through time and are reflected in the peak and off-peak flows patterns. The essential challenge is to tie this to other data such as activity volumes of employment retailing, residential populations and so on that come from more traditional sources. In short the challenge is to relate this kind of short term routine big data sensed in real time and at a relatively coarse spatial scale with cross-sectional data at a finer scale of areal units which are averages for a typical day at a fixed point in time. This is the challenge of merging finer temporal scales based on individual behaviours with cross-sectional data pertaining to aggregate data over wider spatial scales.

Clips from the **YouTube** Movie: *Oyster Gives Up Its Pearls*, made by UCL Engineering from Jon Reades' Movies of the Data

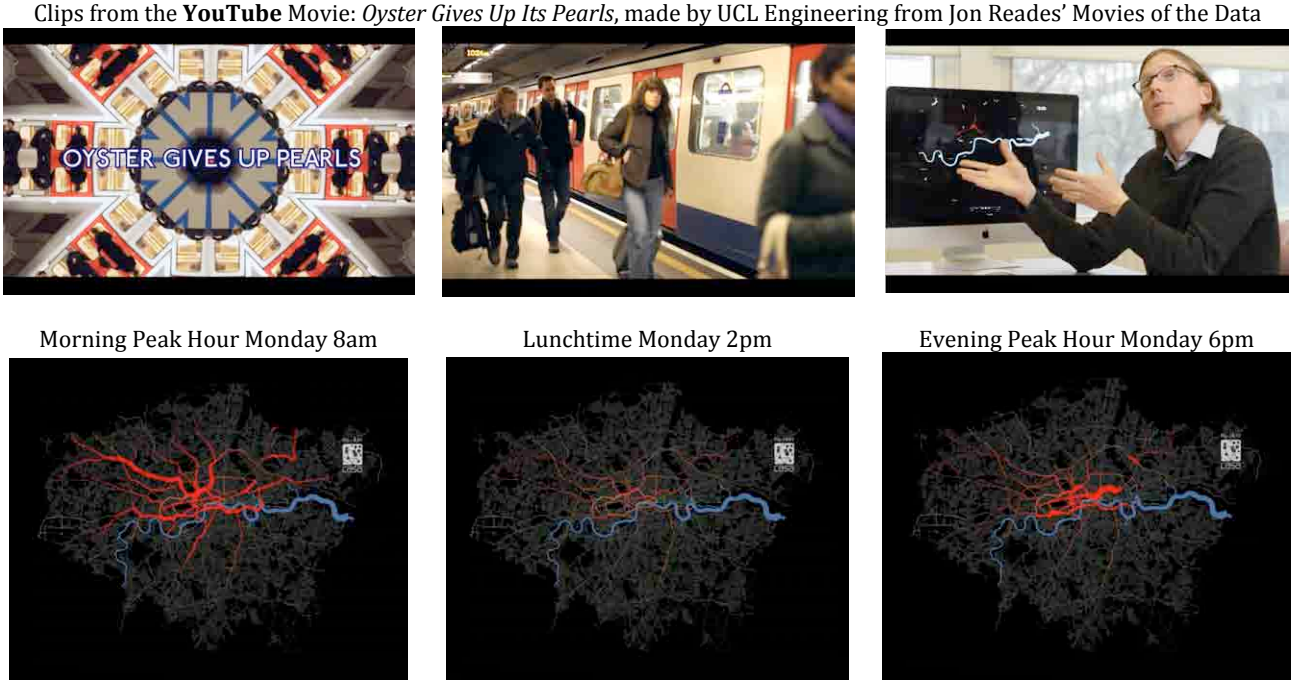

Figure 6: Visualisations of the Flows on the Rail Segments During a Working Day. Movie available at YouTube (https://www.youtube.com/watch?v=9sAugcb2Qj4)

Conclusions and Next Steps

Big data is never what it seems. The multiple V's that have become its signature definition do not capture the fact that quite small data when elaborated into second, third and higher order effects can become big in the sense that conventional techniques and models fail to deal with its extended volumes. Our first illustrations here do focus on quite modest data sets and we are conscious that really big data volumes that come from interaction patterns are hard to measure in terms of their complexity through visualisation. The visualisation of data in countless ways has proceeded in parallel to the big data revolution which is focussed more on data mining through machine learning and in essence involves iterative techniques for searching for patterns in such data that may or

may not have substantive meaning. For example, our illustration of the quality of the fit of our spatial interaction model of journey to work in Greater London which we show in Figures 3 and 4, suggests several features of our model and data that are quite counter to one another. In fact the intensity of points in Figure 4 – the fact that a large proportion points are inside the core of the scatter – probably need to be separated out and in some senses the correlations between this subset of points is likely to have a very different meaning from the overall scatter which we imply in the inset. When we get many thousands of observations, the notion of the system being partitioned into different generic parts comes to the fore and this is a reminder that in both small and big data sets, the data contains these sorts of substantive interpretations that need to be considered whatever size the data is.

Our continuing work on contemporary big data is taking many forms but so far it is mainly dealing with transit data. Data on energy flows and usage in the smart city is not focal as yet while the analysis of big data associated with social media is and may well remain in some preliminary form for many years. Representativeness is the key issue as is meaning in such data and it is not clear as yet the extent to which this social media data pertains to the social and economic functioning of the city which is a prime concern of data about cities. In terms of transit data, stitching together different data sources is of major concern. Erhardt et al. (2016) show how different automated sources can be combined for automated passenger count and vehicle locations on buses in the San Francisco Bay Area and how this data can be scaled up to deal with area wide transit. In this sense, big data is created or rather extended and conflated through techniques like mashups. These kinds of integration are as important as the search for pattern in such data and as the big data revolution proceeds, it is increasingly clear that the pronouncements on the end of theory, made so vociferously at the beginning of this period by commentators such as Anderson (2008) are not being borne out in any sense. The need for theory is of even greater significance that it ever was and as data volumes grow the need to approach such bigness with clear theory has never been more important.

References

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine*, 16-07, 23 June, 2008, available at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Batty, M. (2013) *The New Science of Cities*, The MIT Press, Cambridge MA.

Batty, M. (2014) Can It Happen Again? Planning Support, Lee's Requiem and the Rise of the Smart Cities Movement, *Environment and Planning B: Planning and Design*, **41**(3), 388 – 391.

Batty, M., Hudson-Smith, A., Hugel, S., and Roumpani, F. (2015) Visualising Data for Smart Cities, in Vesco, A., and Ferrero, F. (Editors) *Handbook of Research on Social, Economic, and Environmental Sustainability in the Development of Smart*

Cities, IGI Global, Herschey, PA, 339-362.

Berry, B. J. L. (1964) Approaches to Regional Analysis: A Synthesis, *Annals of the Association of American Geographers*, **54**, 2–11

Bos, F. (2011) *Three Centuries of Macro-Economic Statistics*, Munich Personal, RePEc Archive (MPRA) Paper No. 35391 at http://mpra.ub.uni-muenchen.de/35391/

Chen, C., Batty, M., van Vuren, T. (2015) Editorial, *Transportation* **42**, 537–540.

Dutcher, J. (2014) What Is Big Data? *DataScience Berkeley Blog* available at http://datascience.berkeley.edu/what-is-big-data/

Erhardt, G. D., Lock, O., Arcaute, E., and Batty, M. (2016) A Big Data Matching Tool for Measuring Transit System Performance, in Piyushimita, T., Nebiyou, T., and Zellner, M. (Editors) *Seeing Cities Through Big Data – Research, Methods and Applications in Urban Informatics*, Springer, New York, forthcoming.

Ferry , G. (2010) *A Computer Called LEO: Lyons Tea Shops and The World's First Office Computer*, Harper Perennial, New York.

Graham-Cumming, J. (2012) *The Great Railway Caper: Big Data in 1955*, at https://www.youtube.com/watch?v=pcBJfkE5UwU
and see http://bigdata.blogweb.casa.ucl.ac.uk/2012/10/03/big-data-problems/

Kitchin, R., Lauriaulta, T. P., and McArdleb, G. (2014) Knowing and Governing Cities Through Urban Indicators, City Benchmarking and Real-Time Dashboards, *Regional Studies, Regional Science*, Vol. 2, No. 1, 6–28, http://dx.doi.org/10.1080/21681376.2014.983149

Lee, D. B. (1973) Requiem for Large-Scale Models, *Journal of the American Institute of Planners*, **39**, 163–178.

Lowry, I. S. (1964) *A Model of Metropolis*, RM-4035-RC, The Rand Corporation, Santa Monica, CA, available at
http://www.rand.org/content/dam/rand/pubs/research_memoranda/2006/RM4035.pdf

O'Brien, O., Batty, M., Gray, S., Cheshire, J., and Hudson-Smith, A. (2014) *On City Dashboards and Data Stores*, a paper presented to the Workshop on Big Data and Urban Informatics, August 11-12, 2014. University of Illinois at Chicago, Chicago, IL, http://urbanbigdata.uic.edu/proceedings/

Reades, J. (2013) Pulse of the City, *Vimeo* at https://vimeo.com/41760845 original at http://simulacra.blogs.casa.ucl.ac.uk/2011/08/pulse-of-the-city/

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** (11), 2178-2199.