



THE HONG KONG  
POLYTECHNIC UNIVERSITY  
香港理工大學



DEPARTMENT OF  
LAND SURVEYING AND GEO-INFORMATICS  
土地測量及地理資訊學系



# Big Data & The City

## Redefining Big, Recasting Small

**Michael Batty**

[m.batty@ucl.ac.uk](mailto:m.batty@ucl.ac.uk)

 @jmmichaelbatty

21 April, 2016

<http://www.complexcity.info/>  
<http://www.spatialcomplexity.info/>



Centre for Advanced Spatial Analysis



# Outline

- Some Ideas about Smart Cities and Big Data
- A Short History of Big Data: How Big is Big?
- Mobility, Transit, & Real-Time Streaming: The Oyster Card Data Set
- Learning about Mobility from the Data

*Variabilities – Heterogeneity and Travel Profiles*

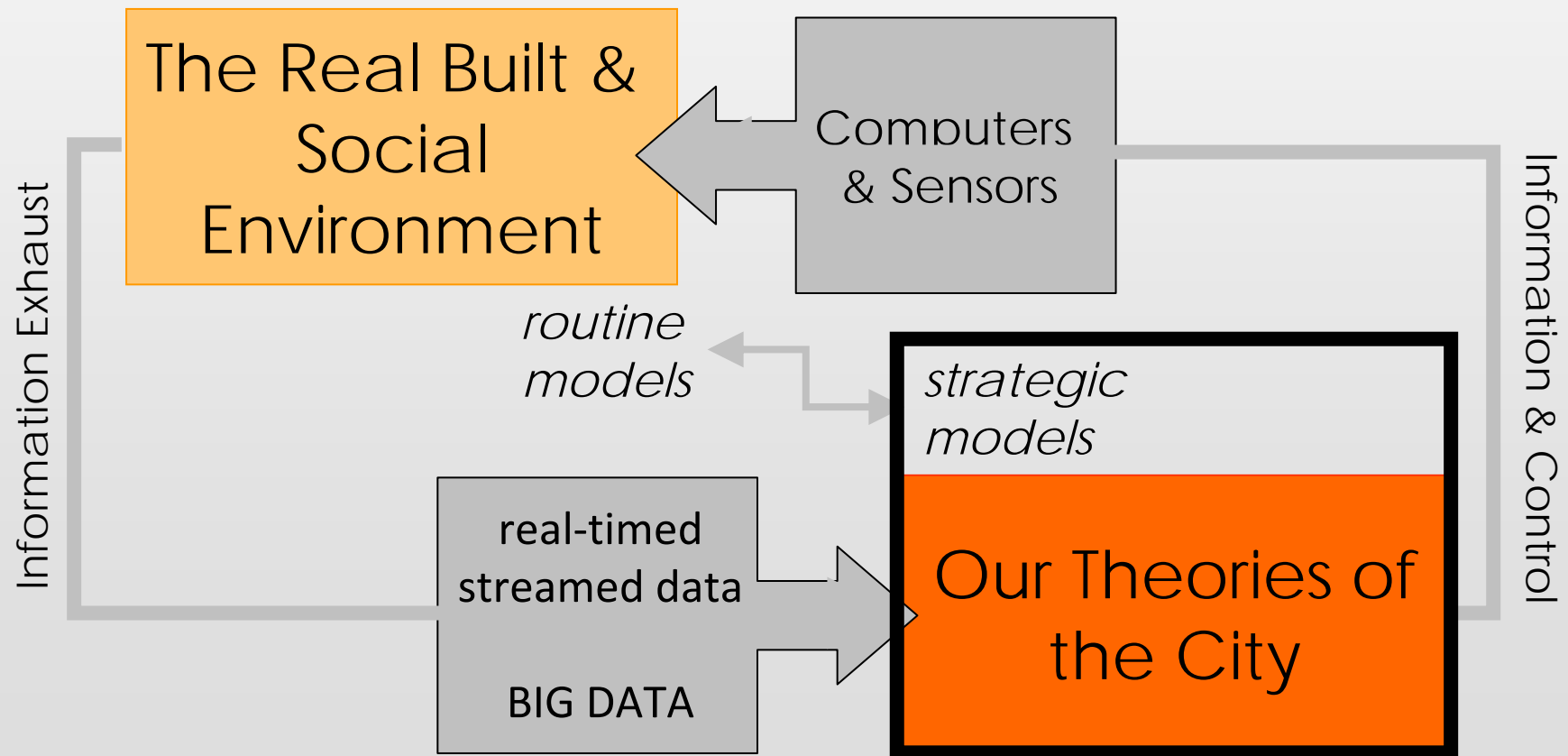
*Disruptions – Signal Failures, Stalled Trains*

*Variable Locational Dynamics of Demand*

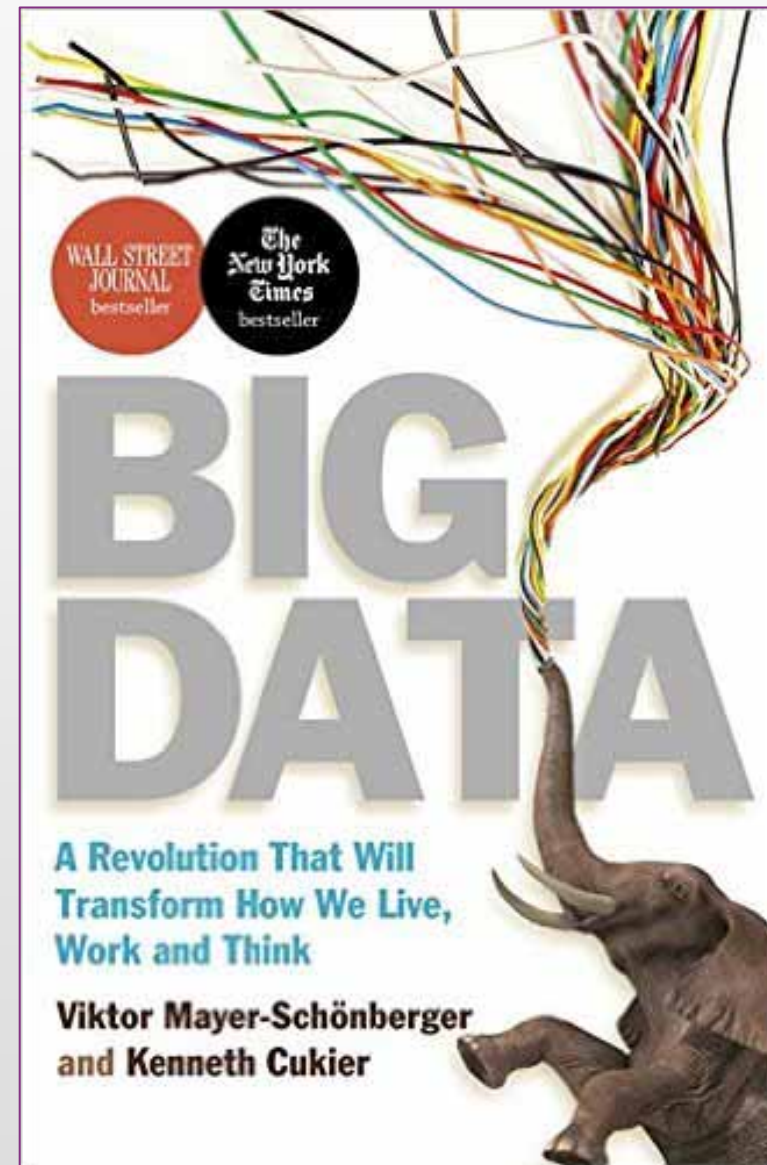
- Related Real -Time Data: Bikes, Social Media
- What Can We Learn: The Limits to Big Data

# What Are Future Cities? Smart Cities?

The spreading out of computers into public places & the built environment and all their consequences



- The way we access the smart city is through technologies that let us generate and use data and its useful equivalent – **information** (data) is key
- Access through **mobile** and **fixed devices** like phones, smart cards, through fixed sensors which record transactions and so on
- These usually complement rather than substitute for data which we collected and used in the past
- This has **introduced time into our thinking** – in the past most urban planning for future cities was timeless – think of garden cities, new towns, master plans
- This is all part and parcel of increasing complexity; more time scales, more opportunities, more diversity



# How Big is Data? Big Can Be Small & Small Big

- Data is big with respect to its volume. I know there are other definitions – velocity, variety etc. but to me, data is big if it requires large use of computer memory implying volume.
- In cities, data usually implies numbers of locations and their attributes but locations imply interactions.
- Thus data are relations between locations and in essence if we have  ***$n$  locations***, we have  ***$n^2$  interactions***. Thus small data can become big. EG:


# Examples: Dublin 1837, Ireland 1888, London 1953


[Posts](#) [About](#) [Complexity](#) [Fractals](#) [Networks](#) [Simulations](#) [Media](#) [Books](#) [Articles](#) [Editorials](#)

[← Visualising Fast Flows](#) [Movies Are Now Online →](#)

## The Oldest Flow Map





Posted on [June 27, 2011](#) by [Michael Batty](#)






 [Tweet](#) 3




.... according to the great cartographer Arthur Robinson, the [two maps of traffic](#) between Dublin and the rest of Ireland by Lt. Harness of the British Army in 1837, are the oldest. Mapped for the Irish Railway Commissioners prior to construction of the railway. Are these actually the first?

Be Sociable, Share!

 [Tweet](#) 3  [Like](#) 0  [+1](#) 0  [Share](#)  [Submit](#)

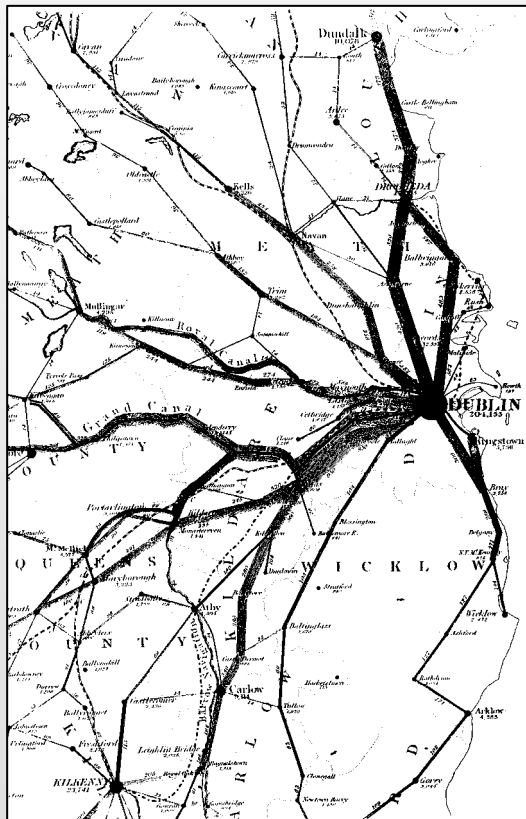


### About Michael Batty

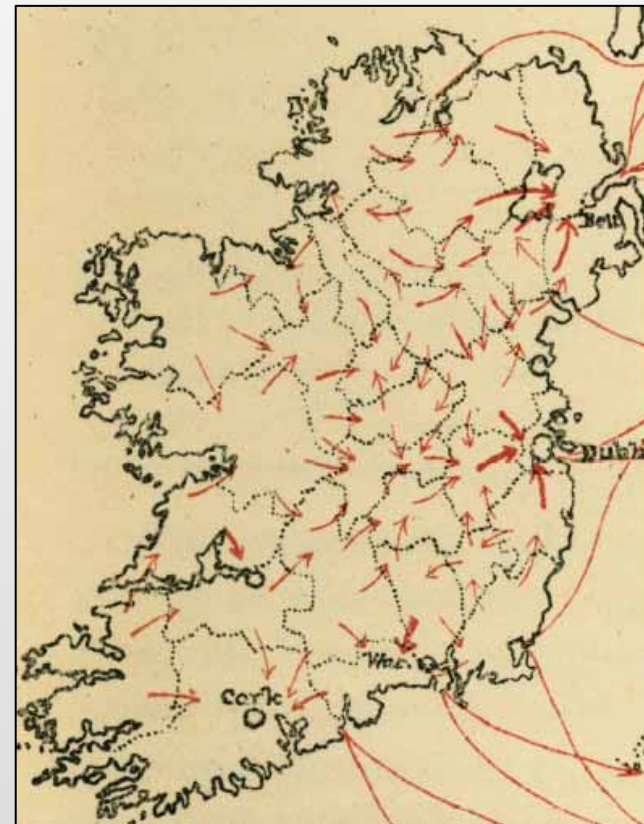
I chair CASA at UCL which I set up in 1995. I am Bartlett Professor In UCL.  
[View all posts by Michael Batty →](#)



# Examples: Dublin 1837, Ireland 1888, London 1955



Harness, 1837



Ravenstein 1888



blog.bigdatatoolkit.org

# BigDataToolkit



## Big Data Problems have been around longer than you think

The Strata Conference is in town and one presentation that caught my eye was titled The Great Railway Caper: Big Data in

big data, data processing, problems, shortest path

[Read More](#)

<https://www.youtube.com/watch?v=pcBJfkE5UwU>

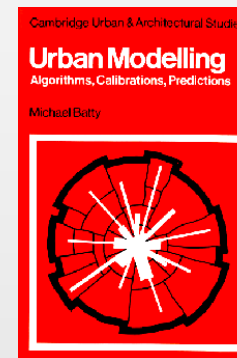
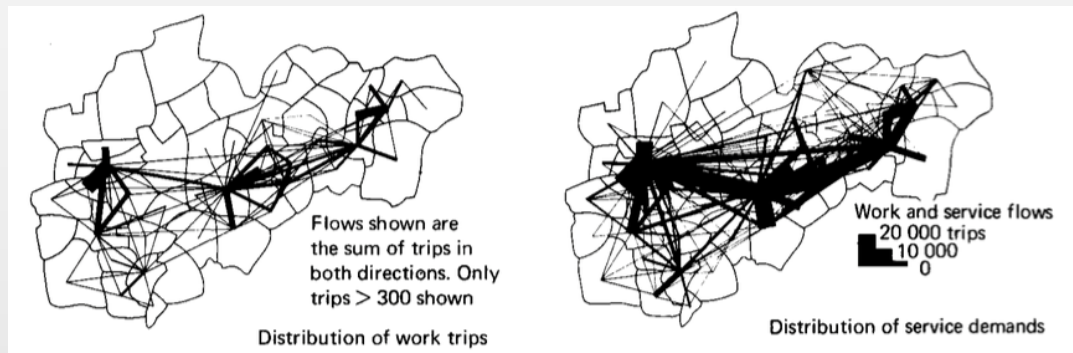
# Locations and Interactions: Flow Systems in Cities

Elsewhere I have argued that we should treat cities as flow systems – as networks. This has been a focus for a long time in transport and land use and we have always been up against the problem of big data.

So let me begin my illustration of this dilemma and how we are thinking about it with some problems that have very small data. Problems of spatial interaction where our numbers of locations is small  $< 100$ ,  $\sim 50$

# Understanding and Visualising Flows

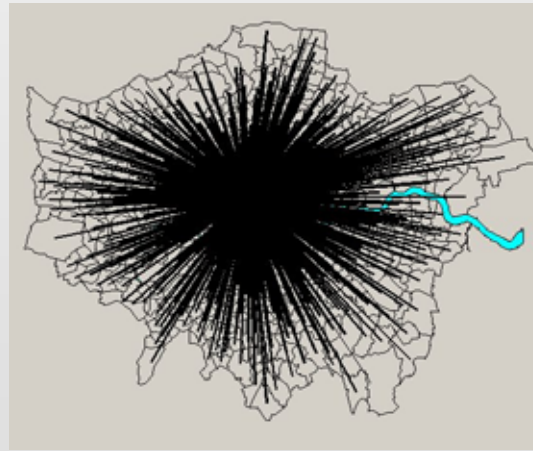
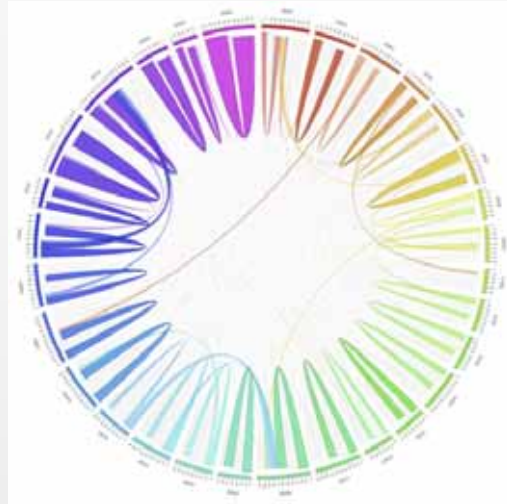
An early model circa 1967-8 Central and NE Lancs



M. Batty (1976)  
**Urban Modelling**  
Cambridge UP

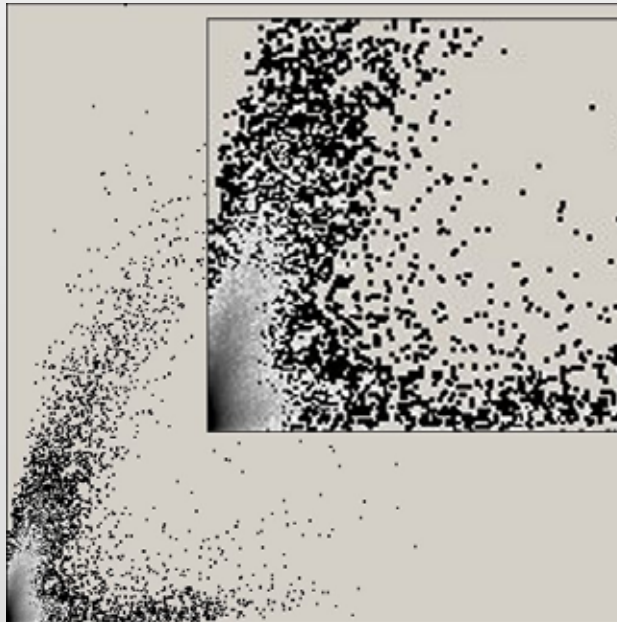
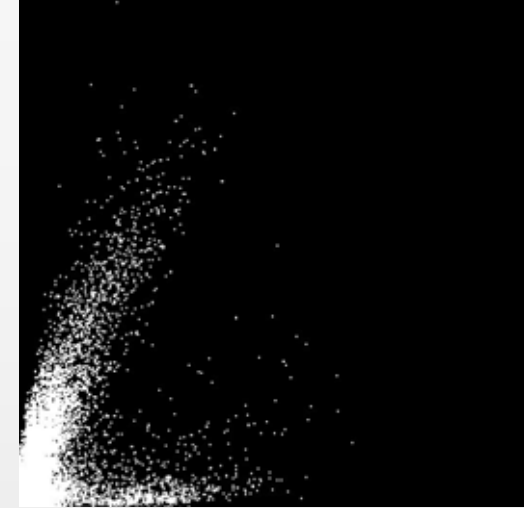
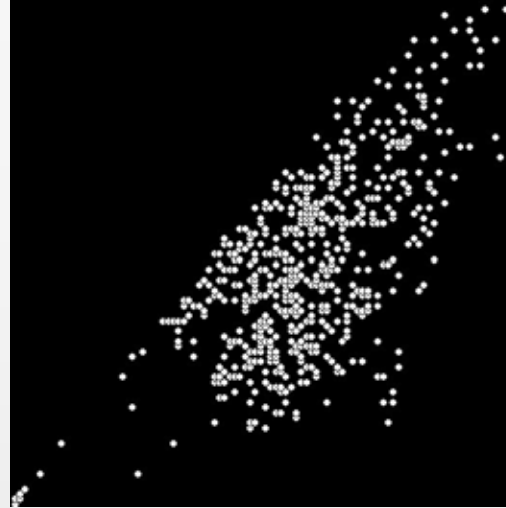
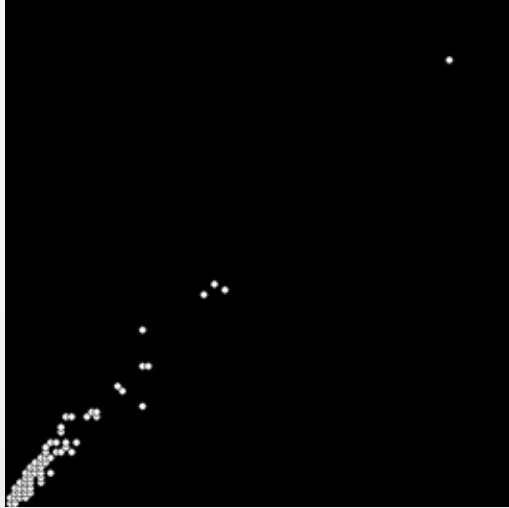


$n^2=33^2=1089$ , not so big but hard to visualise



$n^2=633^2=400,689$ , bigger but impossible to visualise





Even our statistics breaks down when we get large numbers like over several thousand as you can see on the left and above right for 400K data points where the pattern is highly convoluted. This is from a gravity model.

Now what happens when we really do scale up to the level of MSOAs of which there are 7201 in the UK – do we partition and argue we don't need to scale up to  $n^2=7201^2=51,854,401$ .

Circa 52 million points is an issue but our models run in a matter of seconds but that is a lot of data to store – ok it is sparse but sparsity isn't structured so we can't easily partition and in any case we want to compute any possible flows between central London say and Newcastle. Here is the problems scaled up and this is what we are grappling with at present.







(a) MSOA (A=7201)



(b) LSOA (A=34753)



(c) OA (A=181408)

Figure 8.2: ONS Geographies for MSOA, LSOA and OA levels.

## The Web and the Desktop: Users are also Data

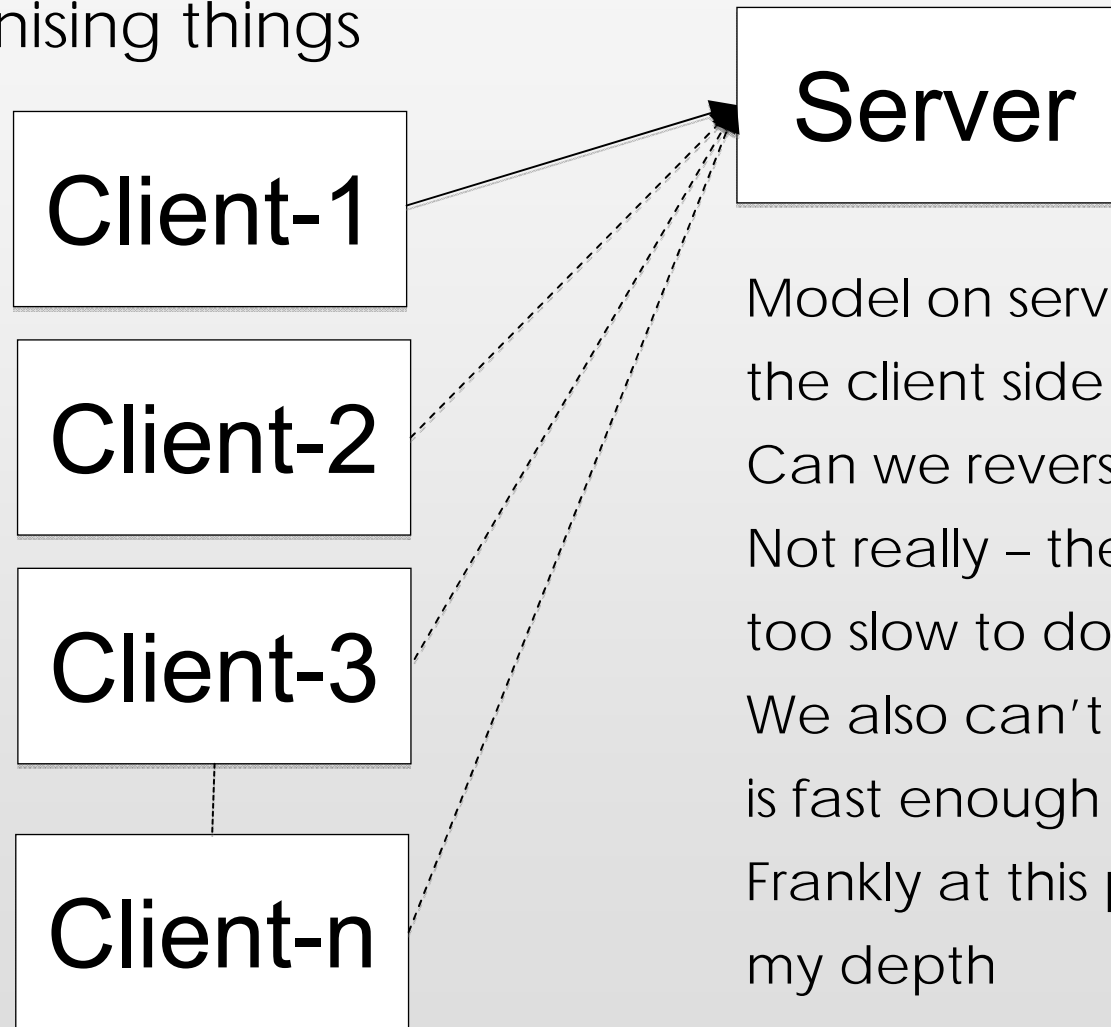
We are building a model of the UK – well E&W at present – we will add Scotland before long – which is of the nature we have been implying – Without going into details, the model takes a few seconds to run – it will take a lot longer when finished as we will add sectors and of course the number of big data we have to hold in RAM might be very large – currently we need to hold 4 such 52 million sized matrices – we may need to go up to 8 in time and that will involve a lot of packing and moving in and out of core, I think

But the real issue is users – if our model is this large, and we have many users, then our data problem is exploded by the users –

Our big data is our original and predicted data from the model, times the number of users. Why are users data ? Well because they are using data differently – they are making their own predictions and thus scaling up the data.

We could have one model for each users but we don't know who the users are? We thus want them to access this on the web. This is where it all hits the fan ..

Here is a block diagram of how we are currently organising things



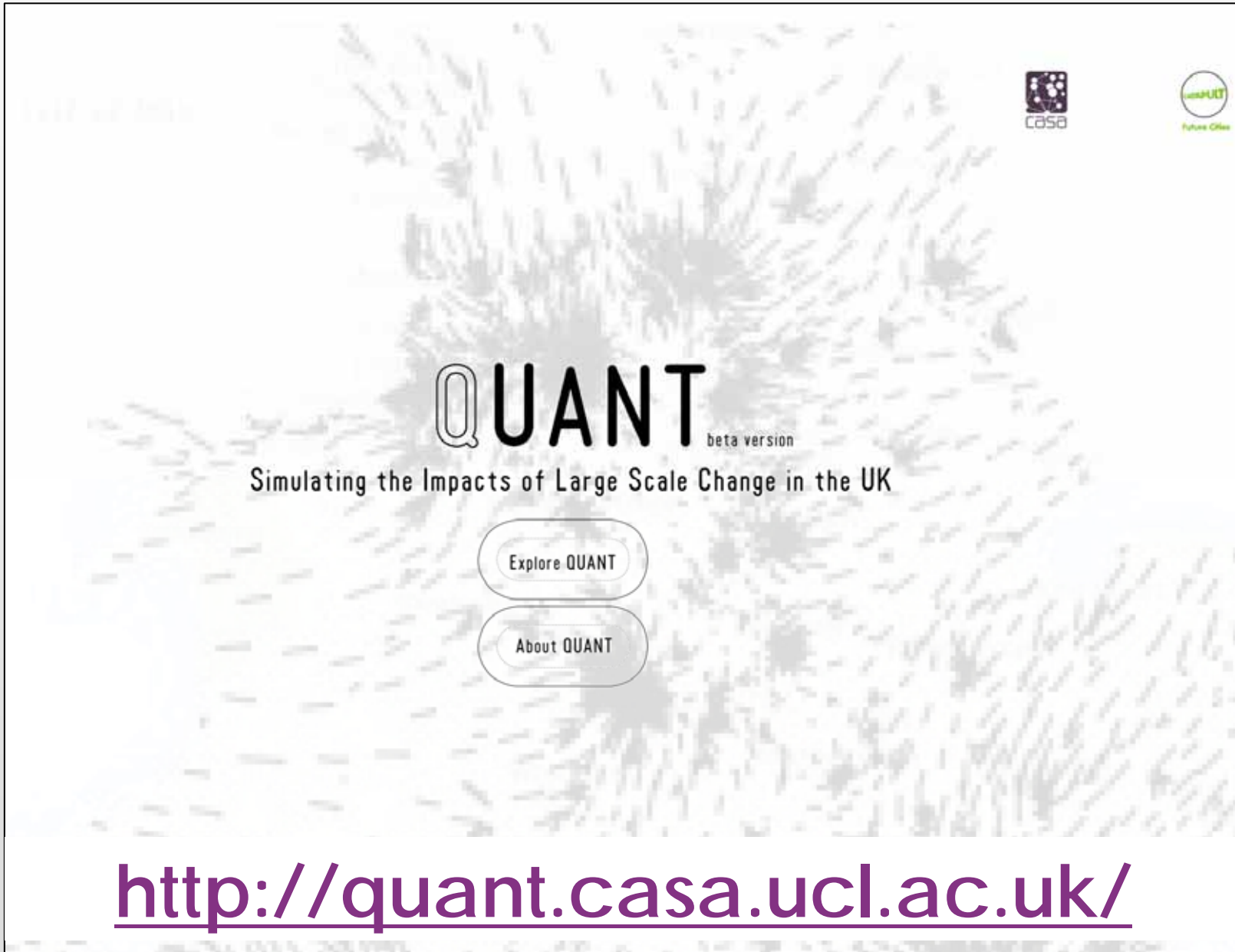
Model on server side; Maps on the client side

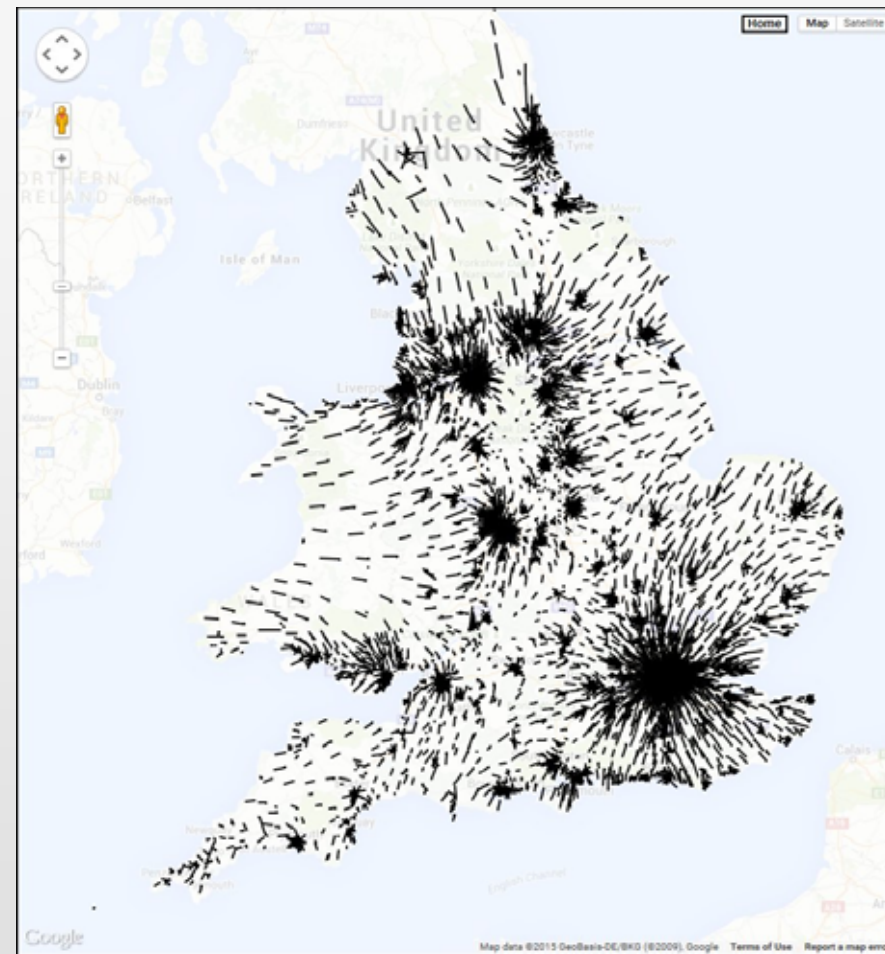
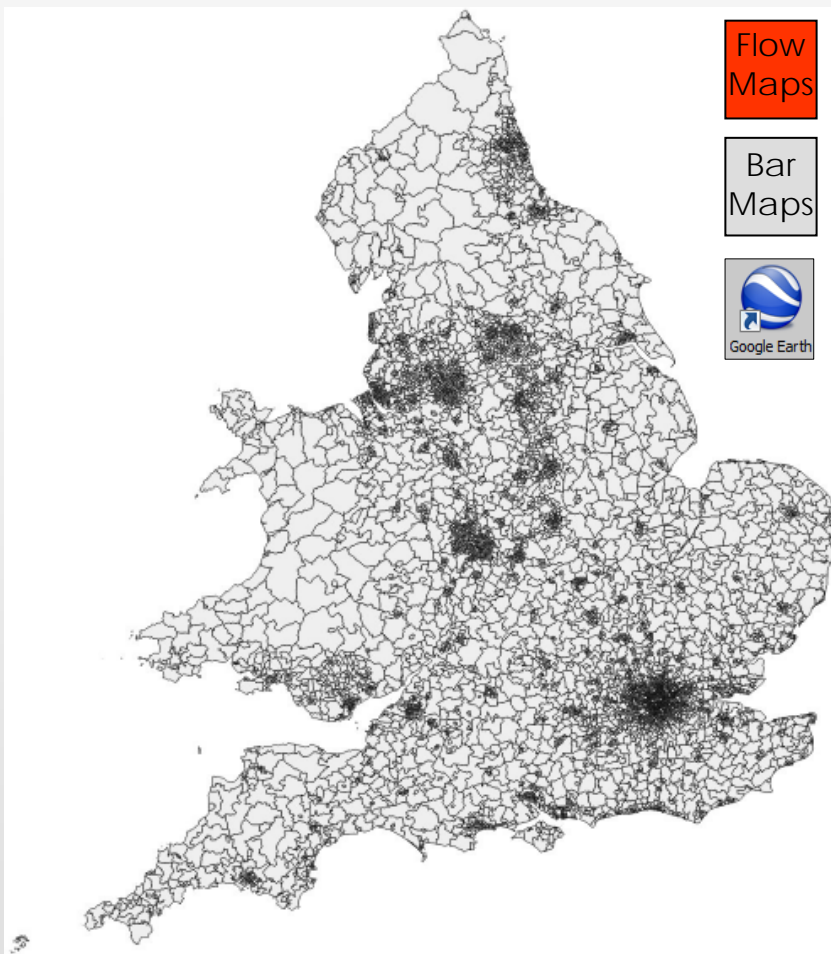
Can we reverse this?

Not really – the matrices are too slow to download to client?

We also can't assume the client is fast enough for computation.

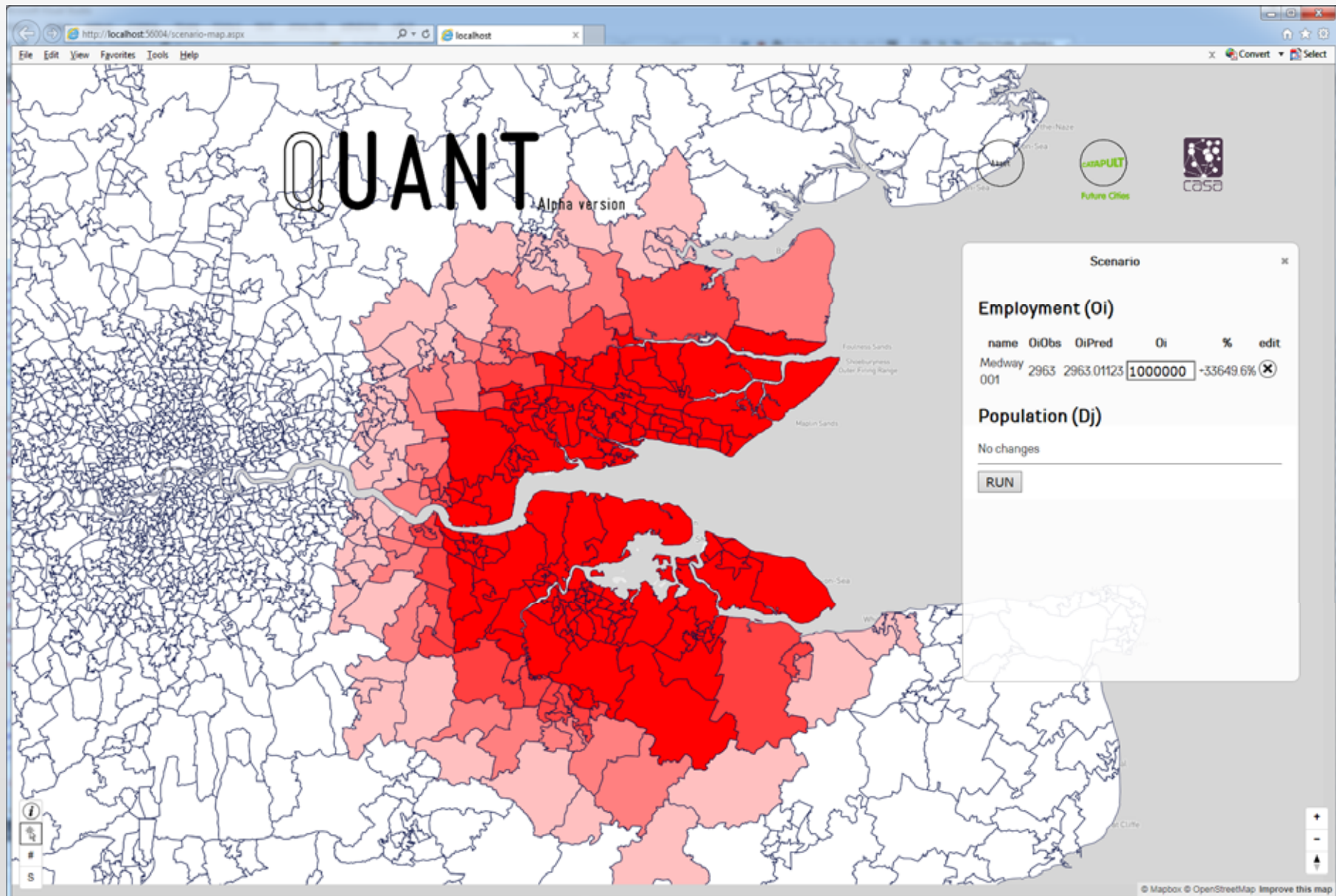
Frankly at this point, I am out of my depth





$$[x_i, y_i] = [x_i, y_i], \left[ \left[ x_i + \frac{\sum_j T_{ij} [x_i - x_j]}{n} \right], \left[ y_i + \frac{\sum_j T_{ij} [y_i - y_j]}{n} \right] \right]$$







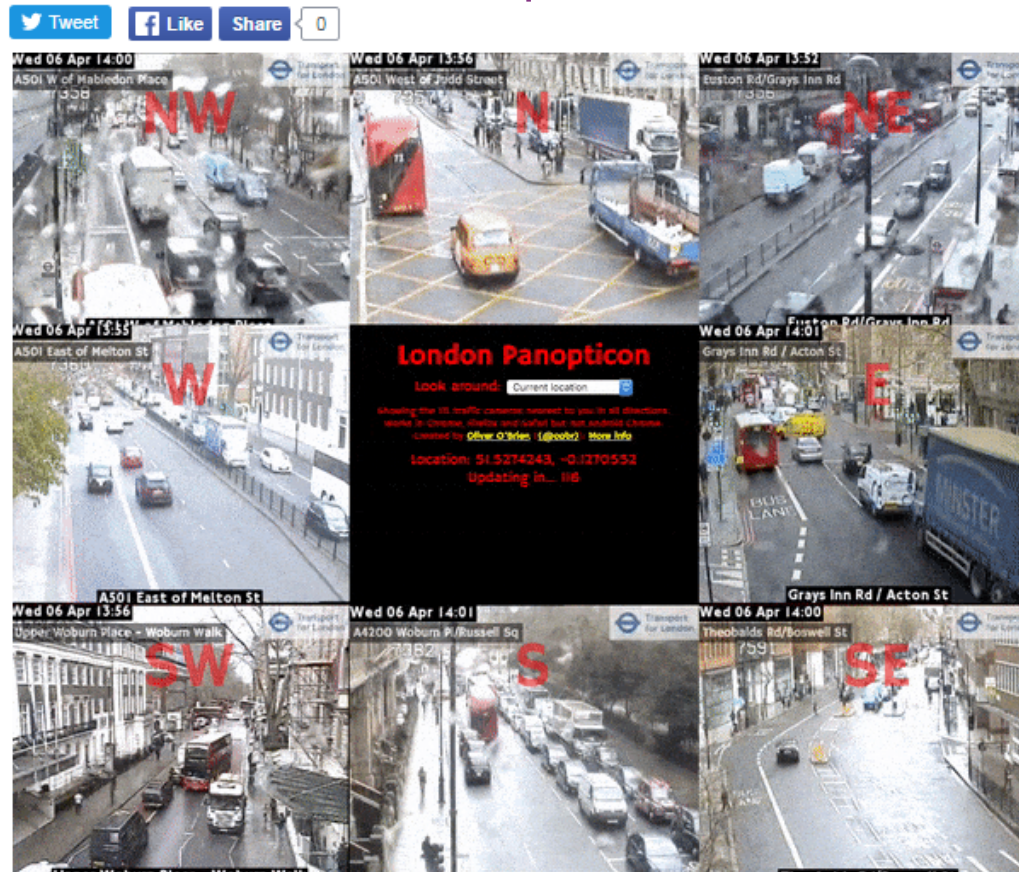
# Real-Time Streaming: What Sort of Data ?



# London Panopticon

🕒 6 April 2016 📍 London

<http://vis.oobrien.com/panopticon/>



# Real-Time Streaming: The Oyster Card Data Set

- Tap at **start** and **end** of train journeys
- Tap at **start only** on buses
- Accepted at 695 Underground and rail stations, and on thousands of buses
- **Many Variants of the Data Sets**
- **991 million** Oyster Card taps over Summer 2012 – this is big data
- Quality of Data
- What Can We Use It For
- Missing Data and Noise







Tube, Overground and National Rail Networks in London  
where Oyster cards can be used

## OYSTER GIVES UP PEARLS

How studying millions of Oyster Card journeys reveals London's 'polycentres'



Researchers from UCL have analysed millions of Oyster Card journeys in a bid to understand how, why and where we travel in London.

Professor Michael Batty (UCL Centre for Advanced Spatial Analysis) and Dr Soong Kang (UCL Management Science and Innovation) applied the techniques of statistical physics to their mountain of raw data.

The pair joined forces with a computational social scientist and a physicist, both based in Paris, to explore patterns of commuting by tube into central London.



They used Transport for London's database of 11 million records taken over one week from the Oyster Card electronic ticketing system.

### Latest news from UCL Engineering

New web privacy system could revolutionise the safety of surfing

UCL host Google Girls Coding Programme with Generating Genius and University of West Indies

Professor Polina Bayvel to Give Royal Society Lecture

### Twitter feed

RT @markmiodownik: Am giving a ENGins seminar today for @UCLEngineering @UCLENGins all UCL engineers welcome - Roberts G06, 6:30pm. [http://...](#)  
8:58am Thu 9th October 2014

RT @Centre4EngEdu: We're hiring! Multi-talented Centre Administrator required to help us launch and expand! [bit.ly/2sERSM](#)  
10:54am Wed 8th October 2014

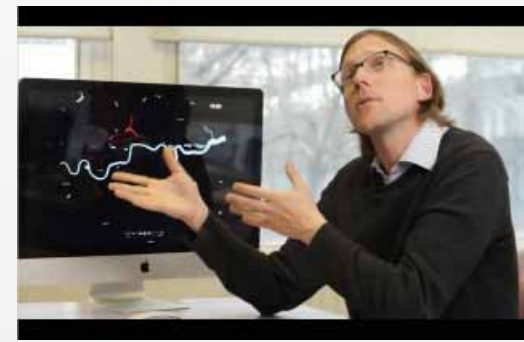
### Join our mailing list

And how can we make sense of this



<http://www.simulacra.info/>



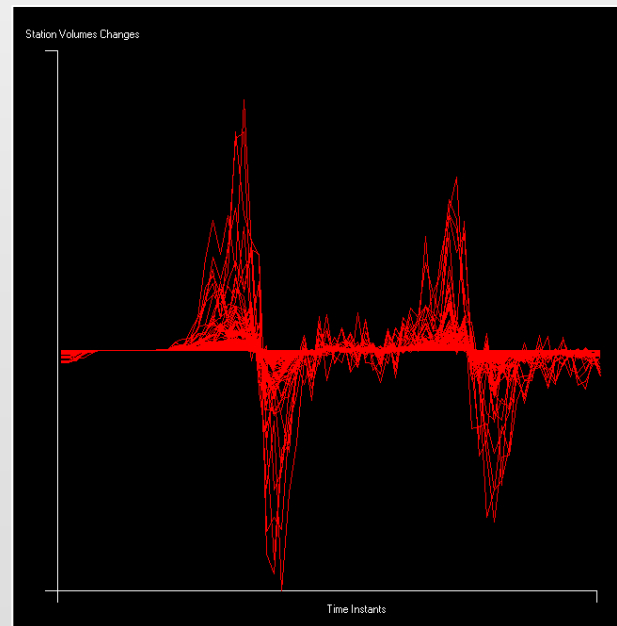
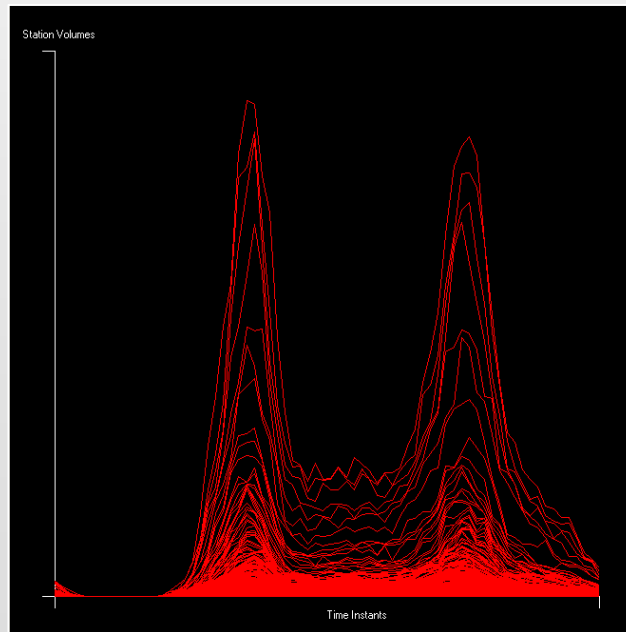


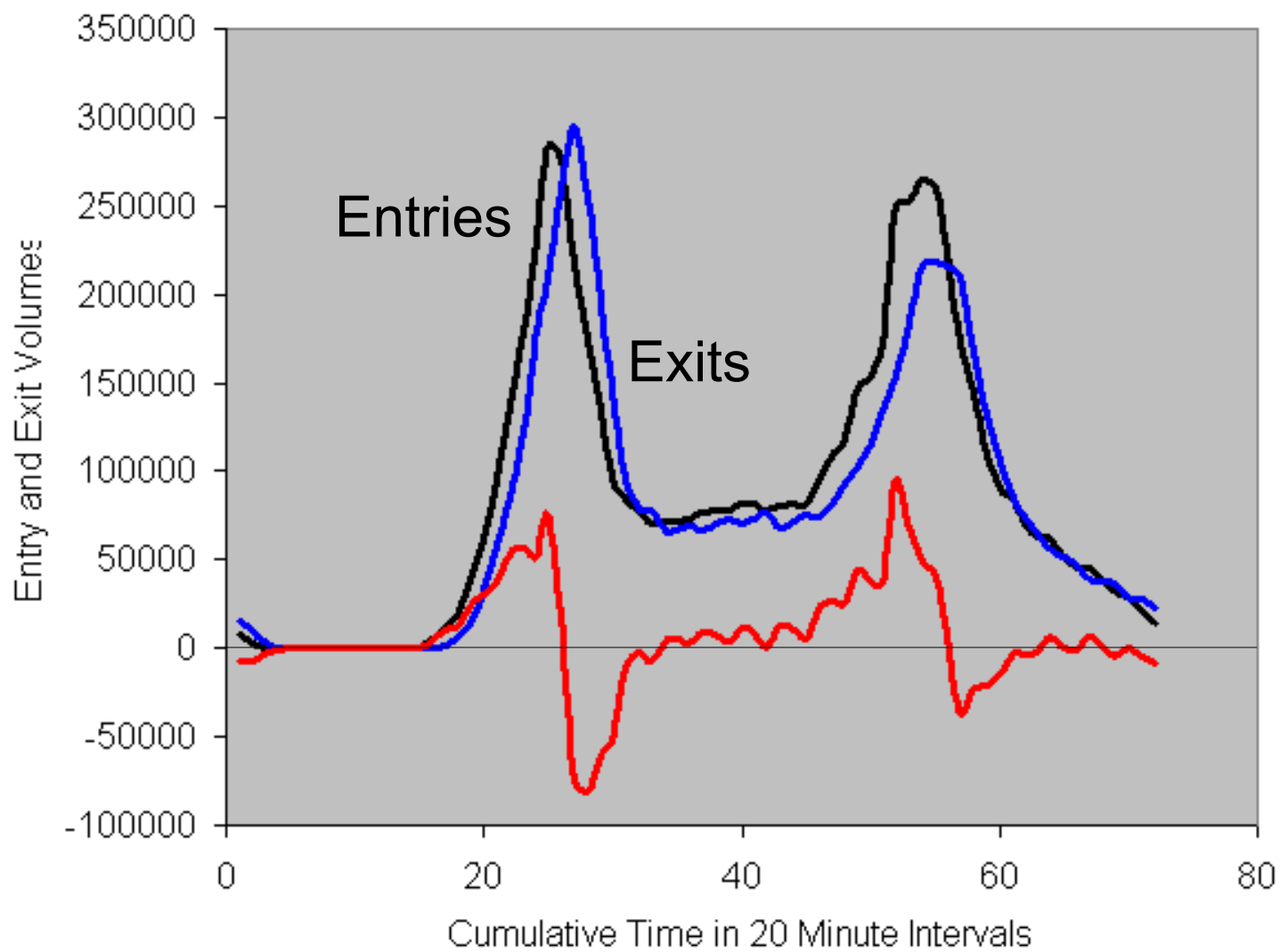
This of course was the thing that Lt Henry Harness did in Dublin in 1837 and what Minard et al. did a little later. In our LUTI models, this is an enormous problem as the scale of this assignment to networks is different



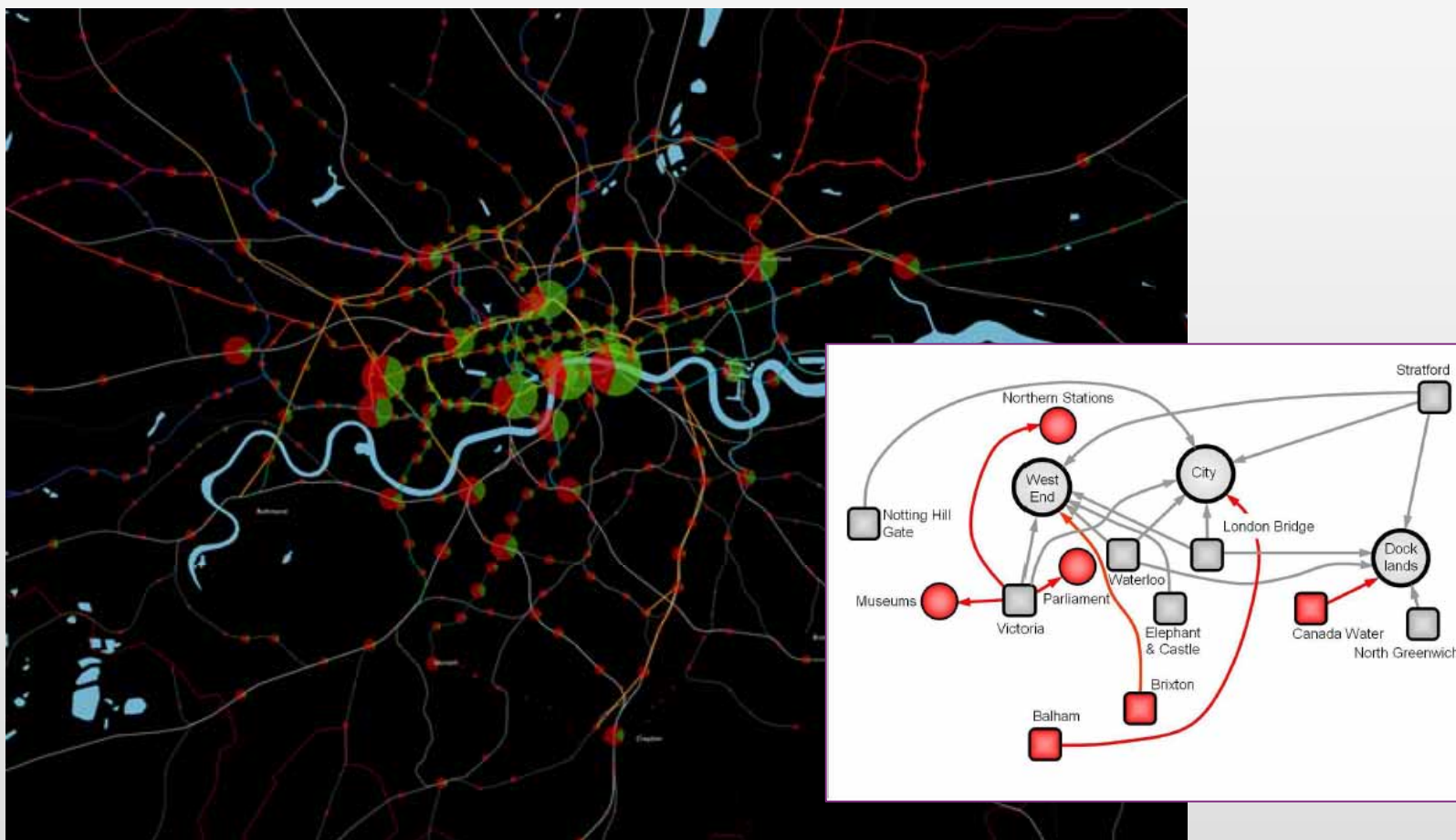
# Variabilities – Heterogeneity and Travel Profiles

First we will look at some of the data and how it varies in terms of the diurnal flows usually morning and evening peaks, with a small blip (peak) around 10pm at night



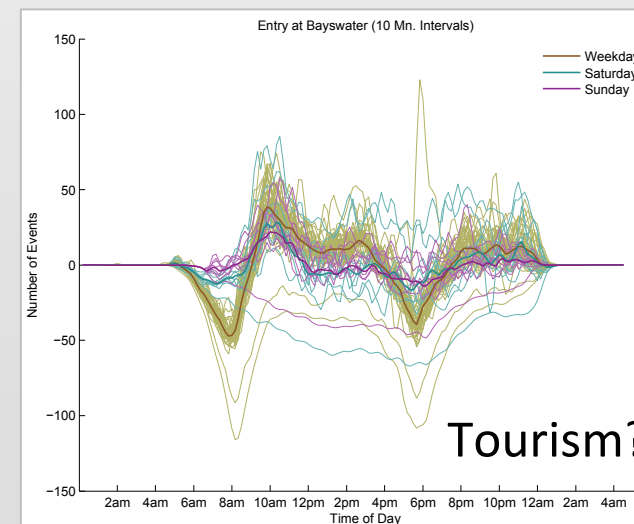
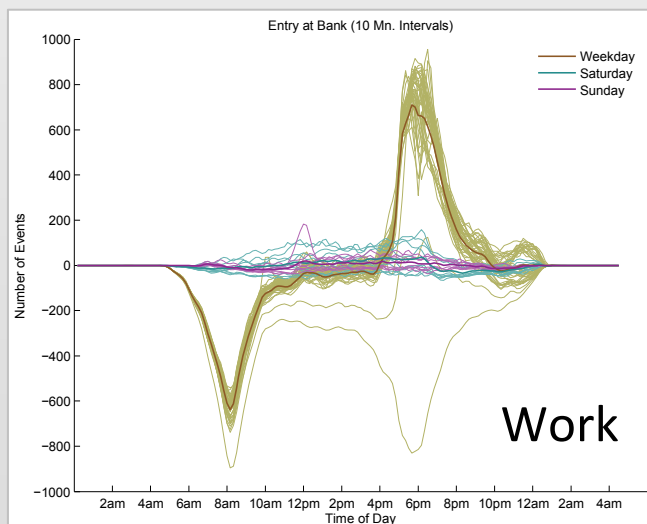
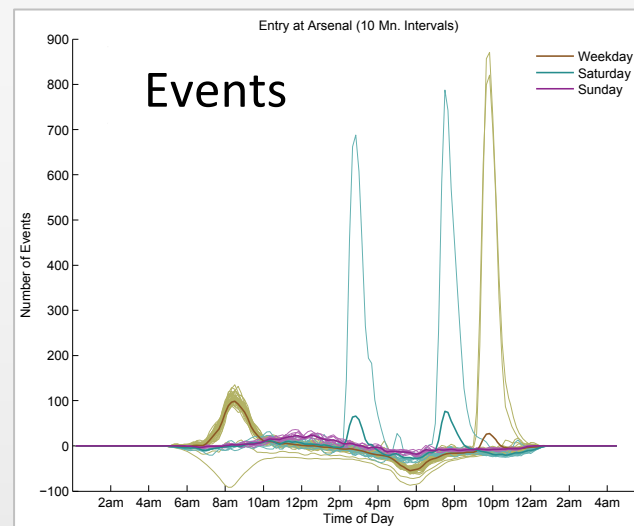
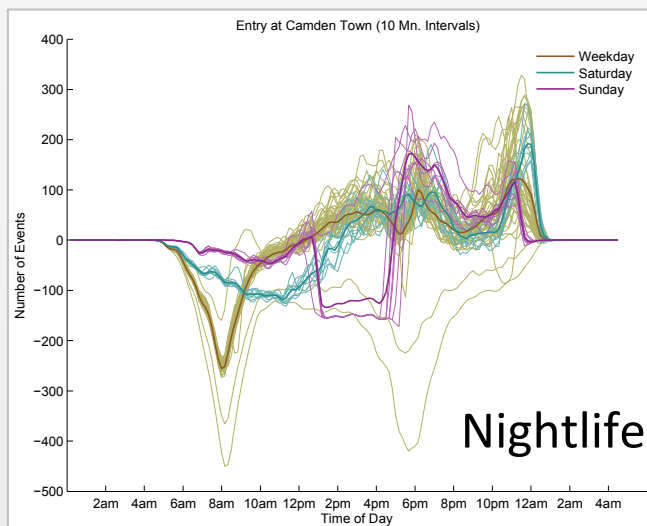


# Oyster Card Data – interpreting urban structure, multitrips, etc.



Roth C., Kang S. M., Batty, M., and Barthelemy, M. (2011) Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. **PLoS ONE 6(1)**: e15923. doi:10.1371/journal.pone.0015923

# Particular Events: Weekdays, Saturdays and Sundays



# Comparing Variability for different time intervals for Three World Cities: London, Beijing and Singapore

*Table 1. Summary statistics of one-week of smart-card data (metro trips only)*

	London	Singapore	Beijing
<b>Monday</b>	3,457,234	2,208,173	4,577,500
<b>Tuesday</b>	3,621,983	2,250,597	4,421,737
<b>Wednesday</b>	3,677,807	2,277,850	4,564,335
<b>Thursday</b>	3,667,126	2,276,408	4,582,144
<b>Friday</b>	3,762,336	2,409,600	4,880,267
<b>Number of stations (1)</b>	400	130	233
<b>Number of tube line</b>	13	4	17
<b>Area (2)</b>	1,572 km <sup>2</sup>	718.3 km <sup>2</sup>	2267 km <sup>2</sup>
<b>Total population (3)</b>	8.63 million	5.3 million	21.15 million
<b>Ridership of Metro</b>	20%	35%	21%
<b>Length of metro lines</b>	402km	182km	465 km
		(MRT+LRT)	

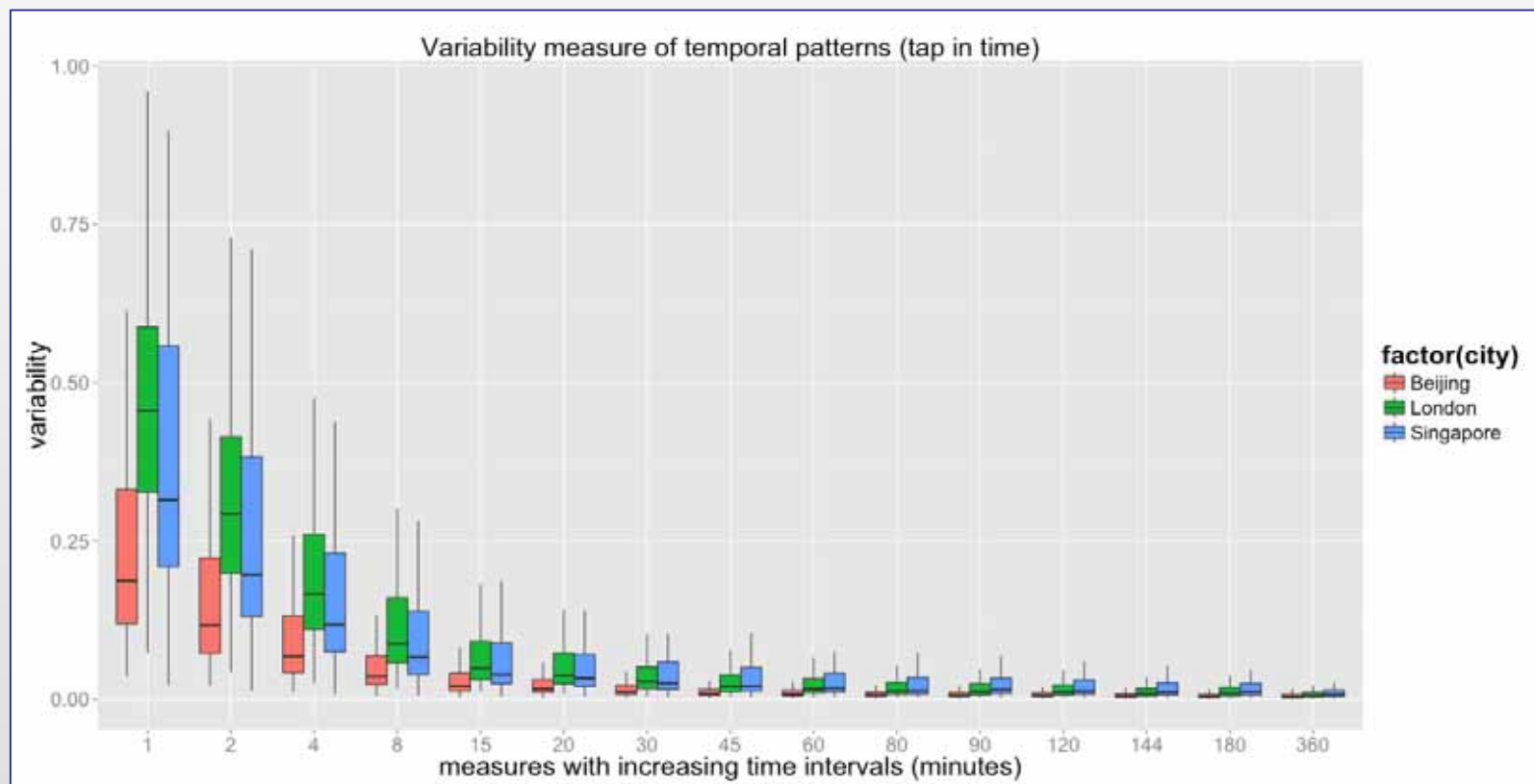
(1) Number of stations is the number of stations with smart-card records generated.

(2) The area of Beijing only counts the area enclosed by the 6th ring road for a fair comparison.

(3) From the World Population Review, <http://worldpopulationreview.com/world-cities/> accessed 17 January 2016

Zhong, C., Batty, M., Manley, E., Wan, J., Wang, Z., Che, F., and Schmitt, G. (2016) Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing using Smart-Card Data., **PLOS One**, <http://dx.doi.org/10.1371/journal.pone.0149222>

## From 1 minute intervals to the whole day



# Comparing Variability for different time Intervals over the day

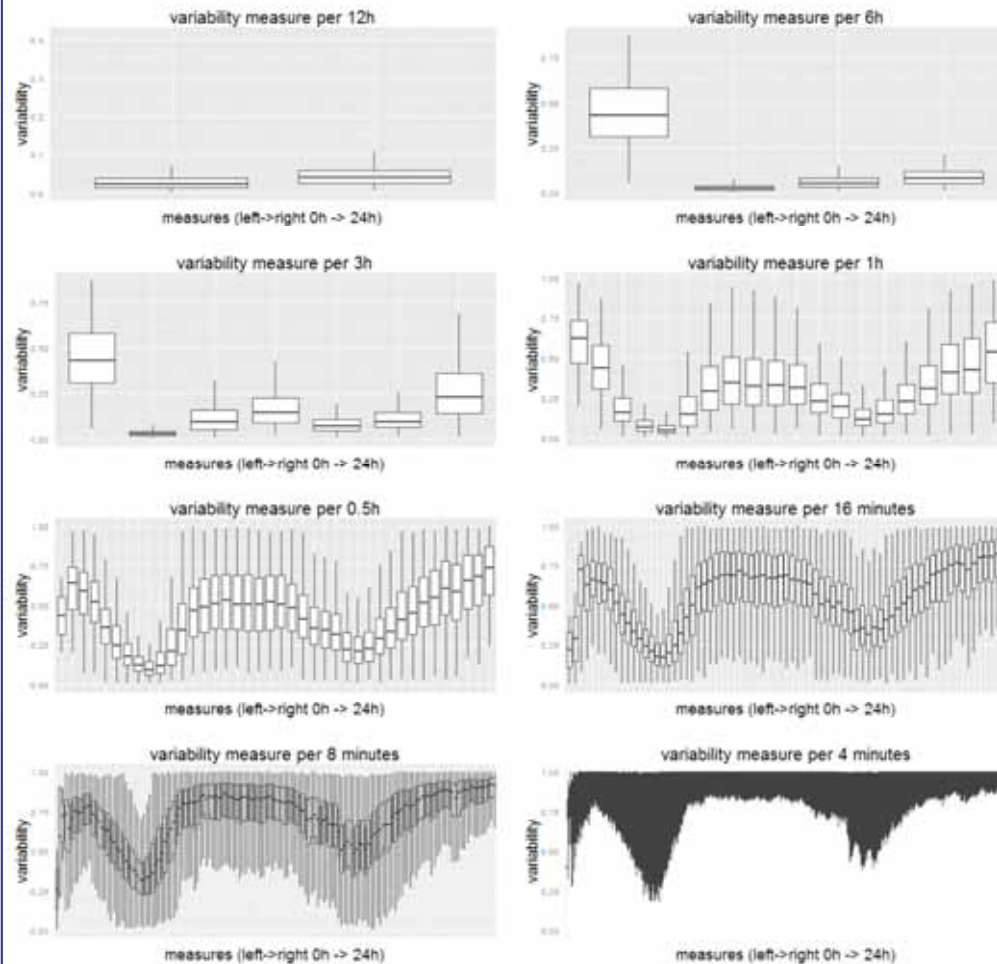
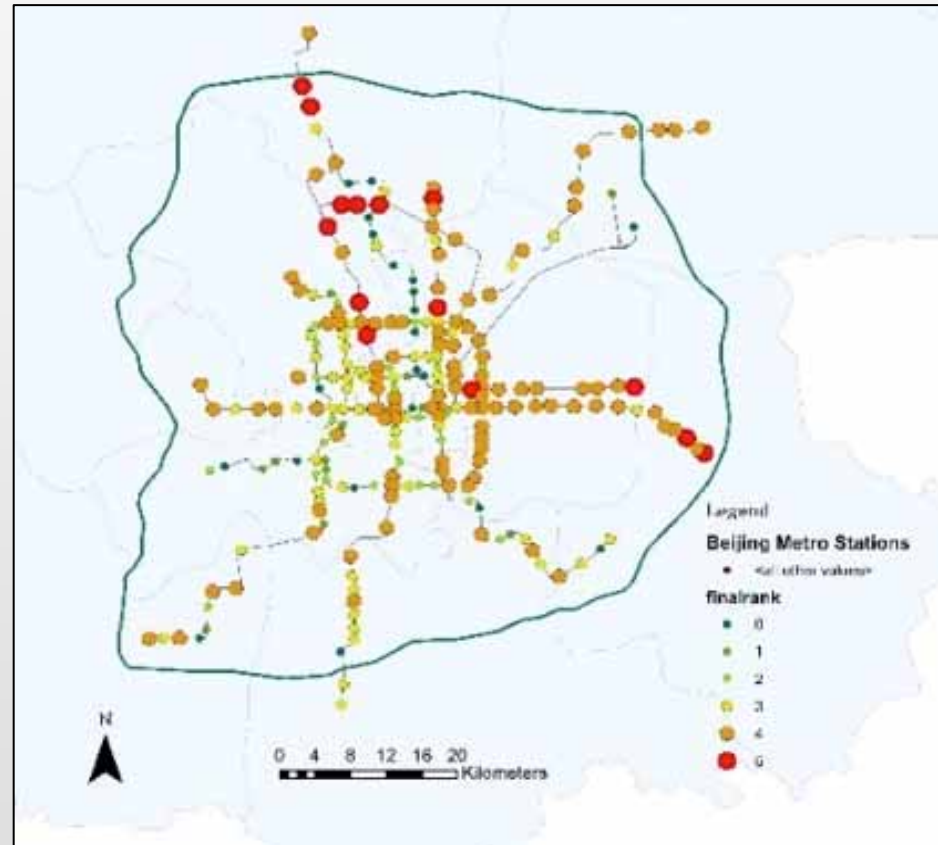
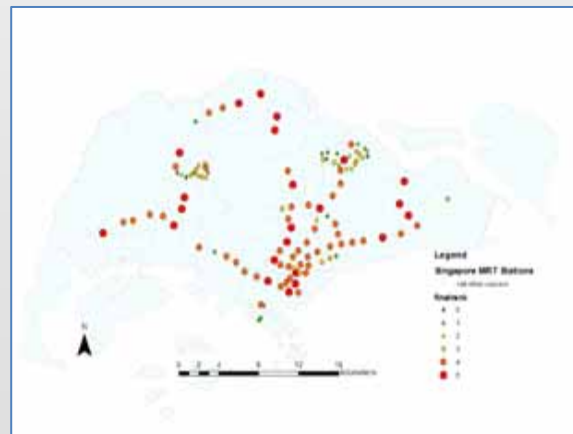
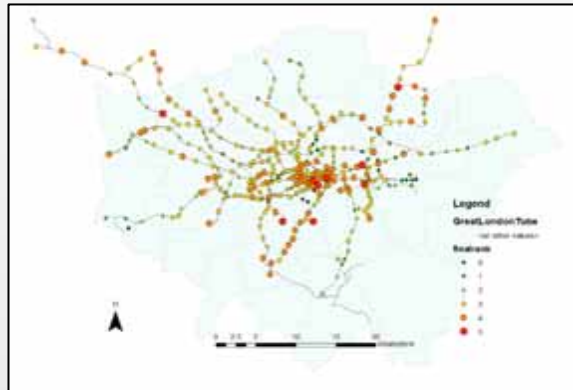


Figure 1. Variability of regularity in the trip matrix over time.

**Note:** Each box plot shows the variability of 400 stations over time measured at different temporal scales. Overall, eight subplots give a similar trend where lower variability appears during peak hours (around 9 am in the morning and 6pm in the evening). More details can be captured as differences of variability between each time unit are magnified as we decrease the temporal scale from 12h to 4 minutes.



# Comparing Variability for different time intervals for Three World Cities: London, Beijing and Singapore





# Disruptions – Signal Failures, Stalled Trains

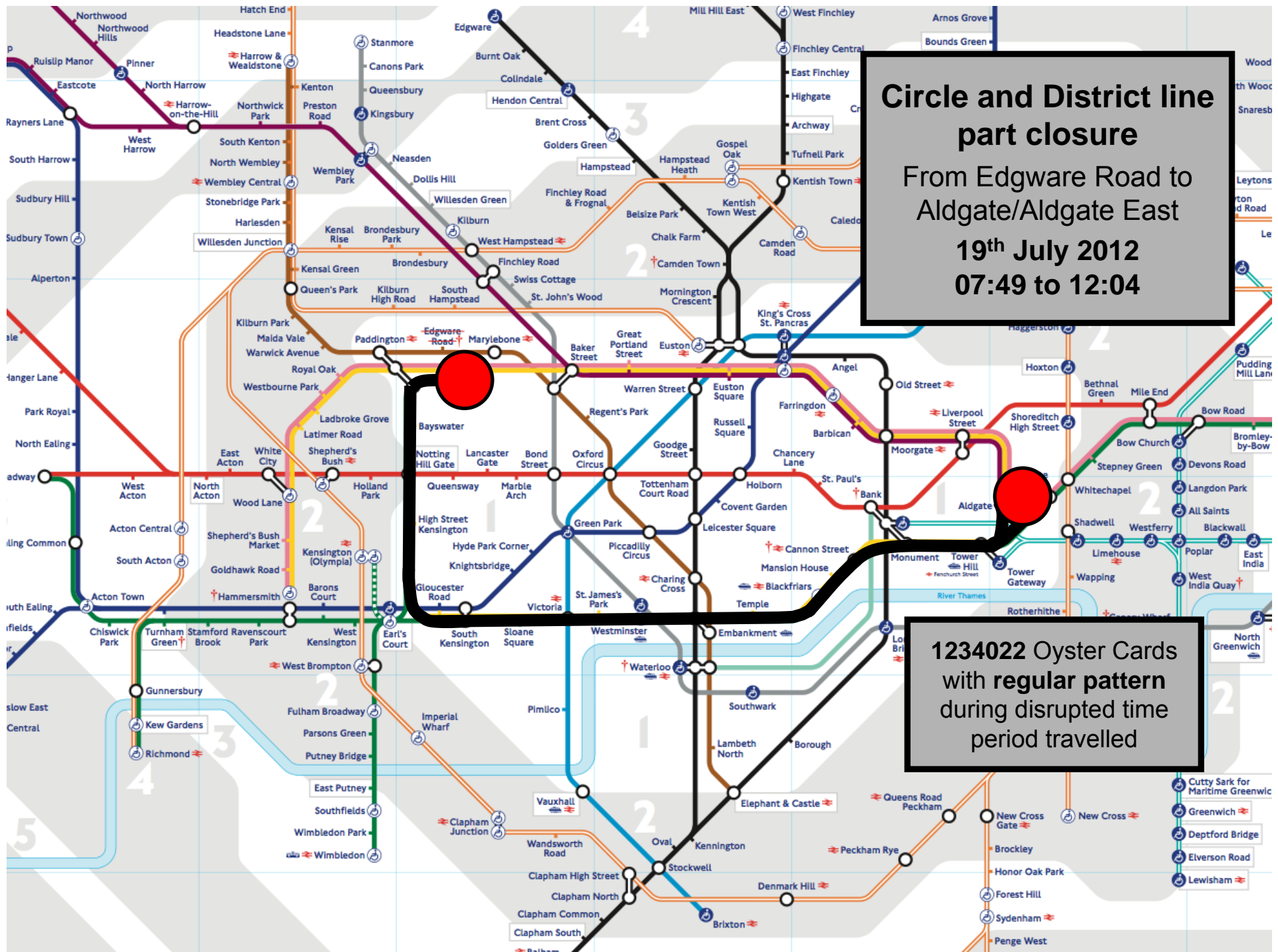
- We will look at three disruptions – the Circle and District Lines which had a 4 hour stoppage on July 19<sup>th</sup> 2012
- And a Bus Strike in East London and how this shows up in the data
- And typical pattern of delay on all modes visualised for Greater London

## Circle and District line part closure

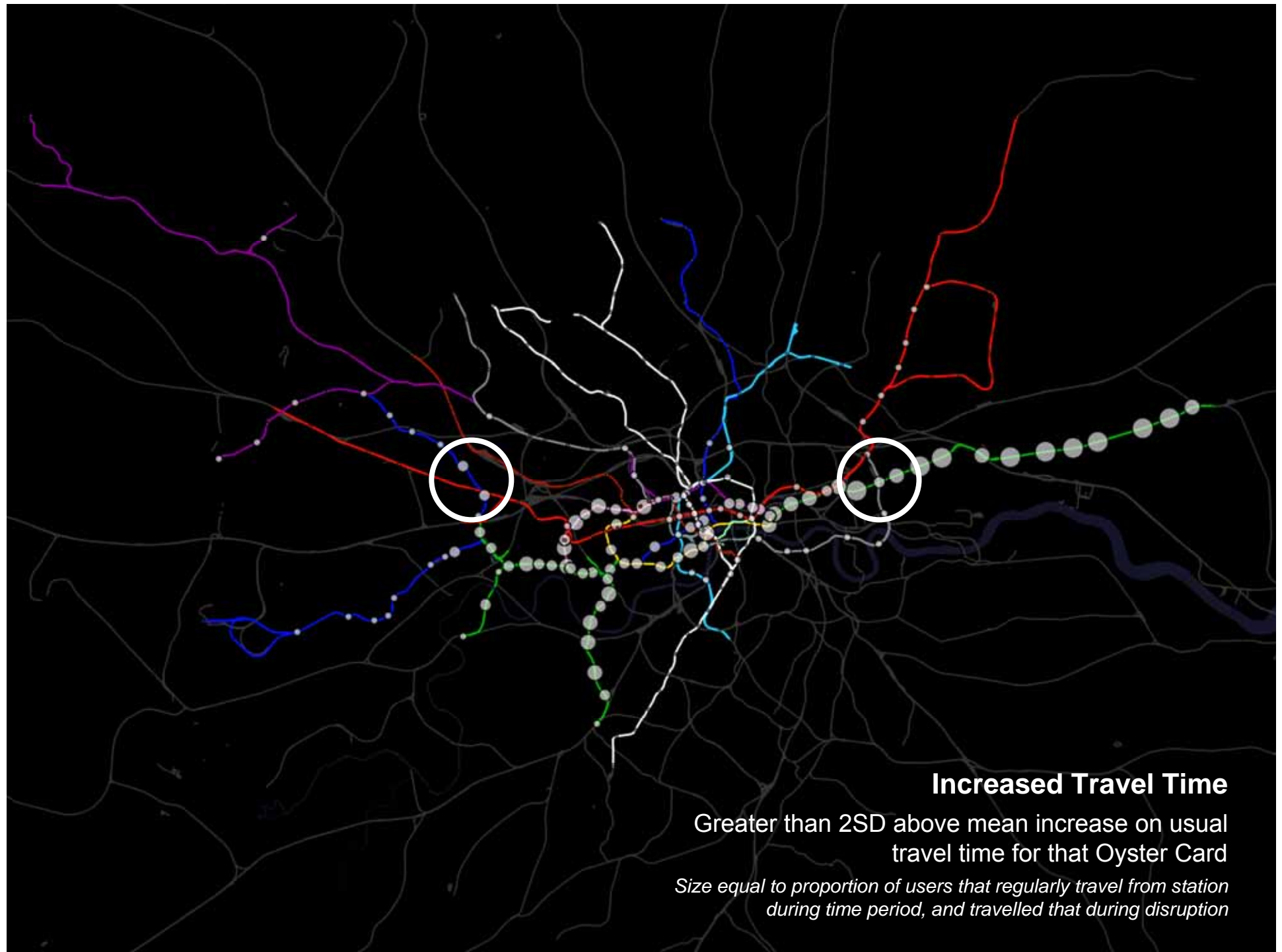
From Edgware Road to  
Aldgate/Aldgate East

**19<sup>th</sup> July 2012**  
**07:49 to 12:04**

**1234022 Oyster Cards**  
with **regular pattern**  
during disrupted time  
period travelled

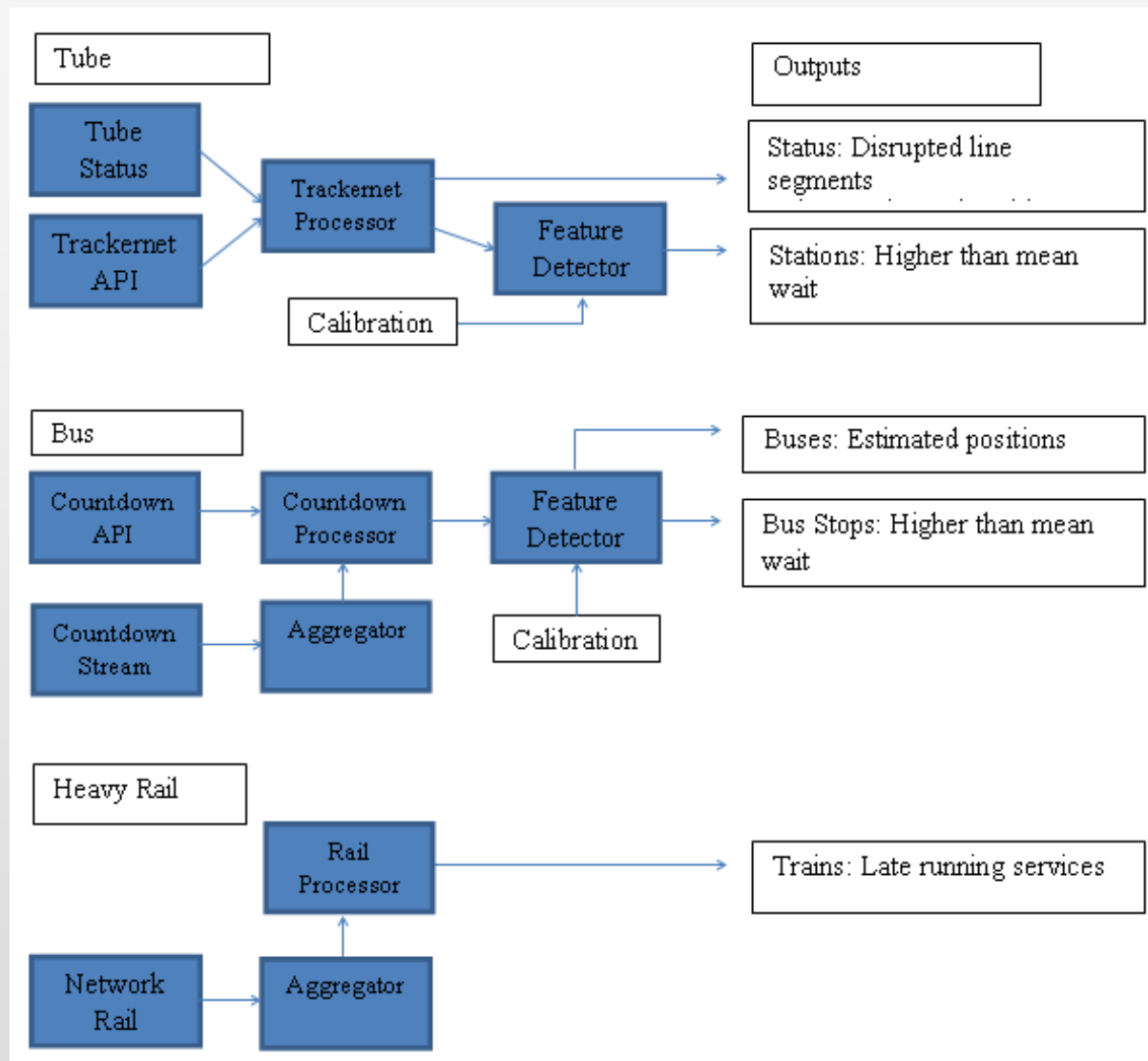


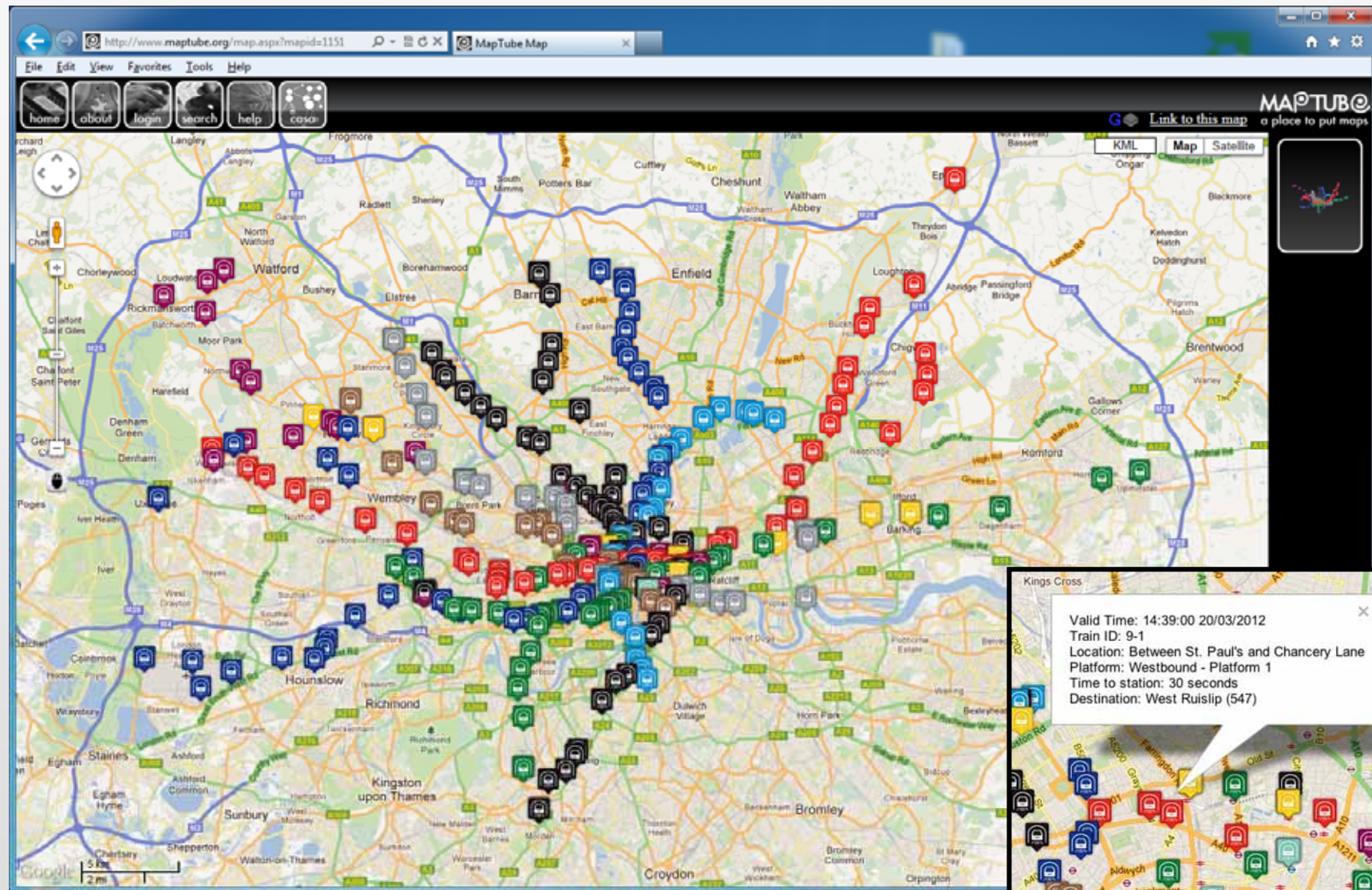


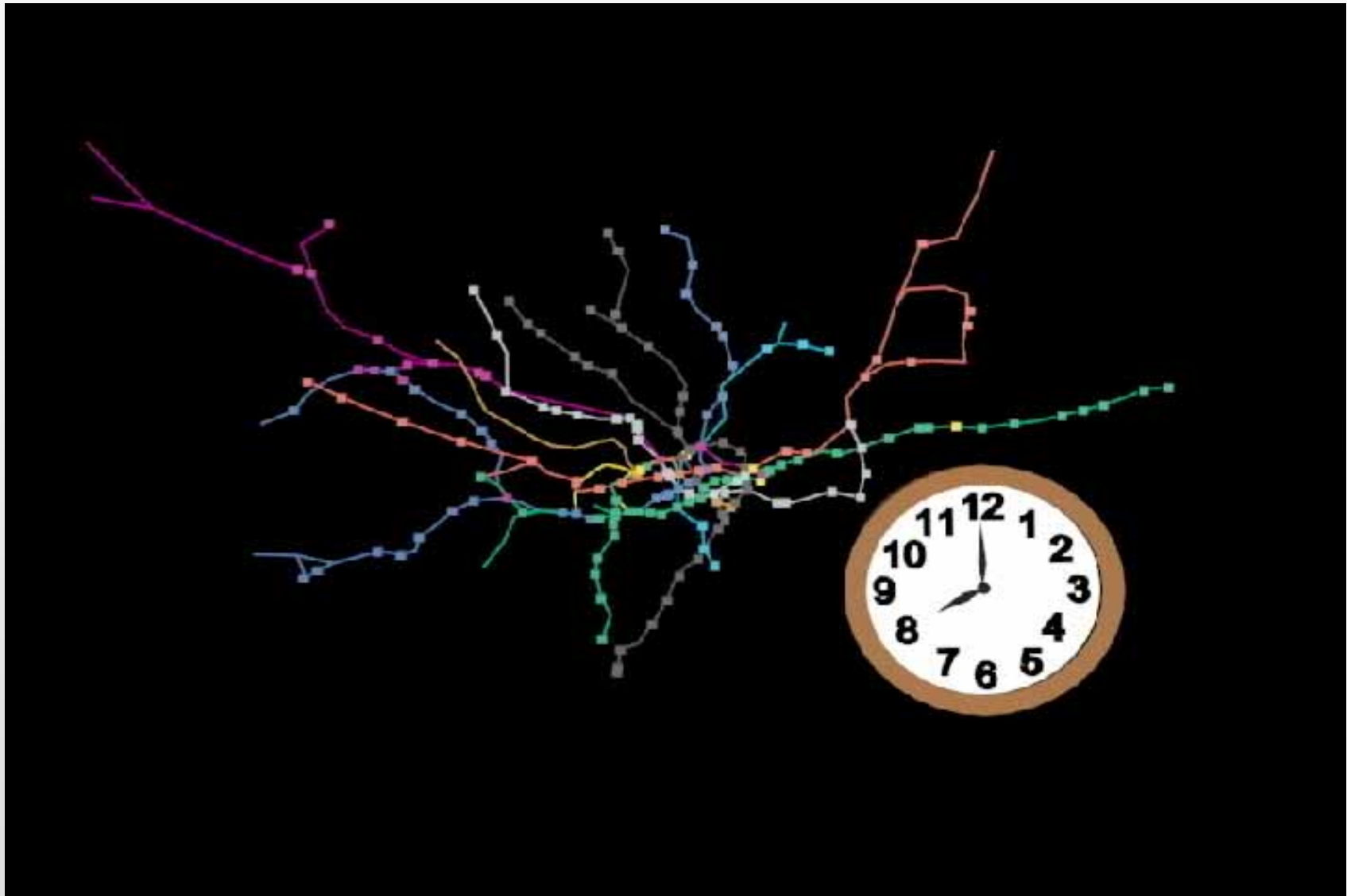




# The Public Transport System in Terms of Vehicle Flows

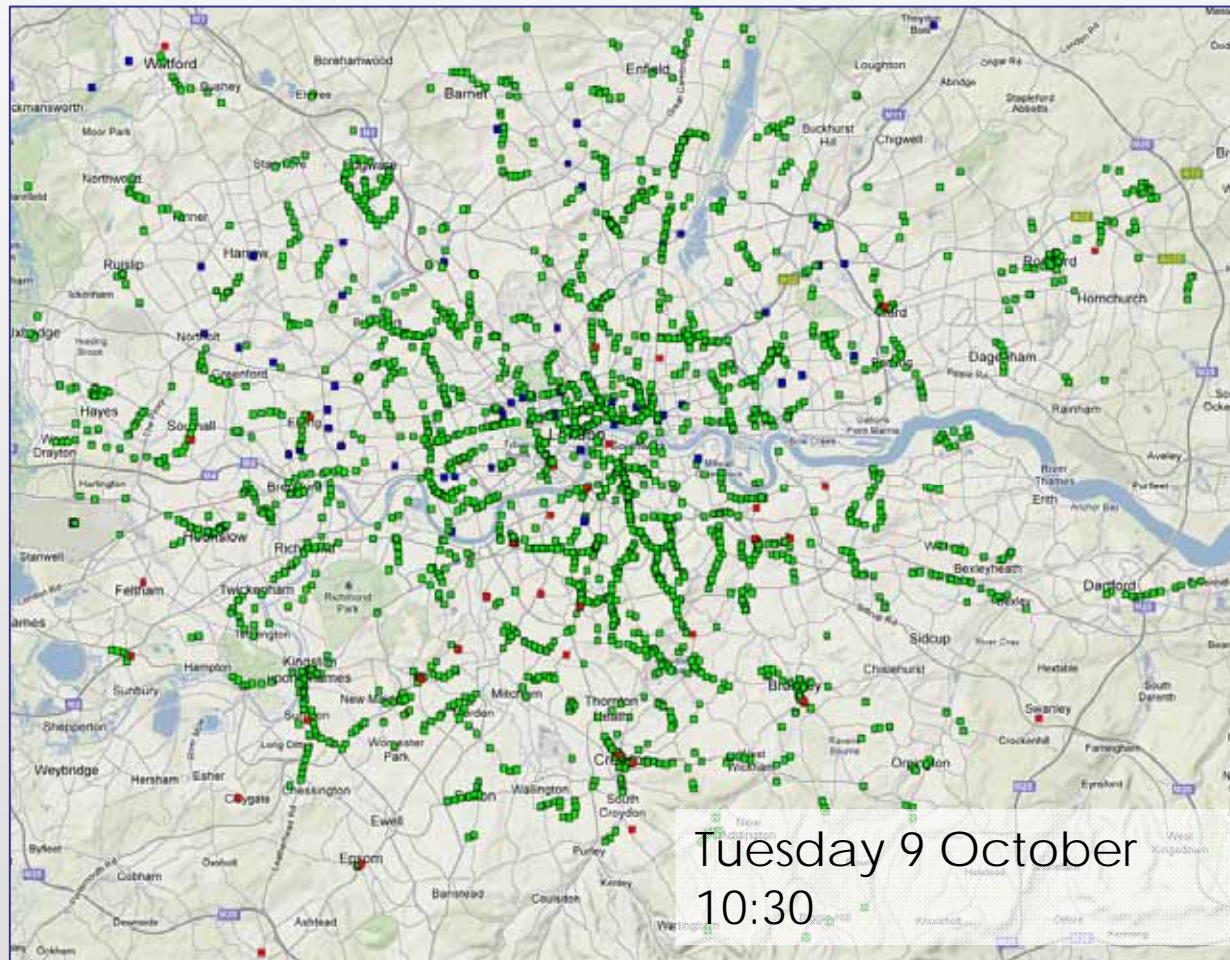











# Delays from Tube, National Rail and Bus Fused



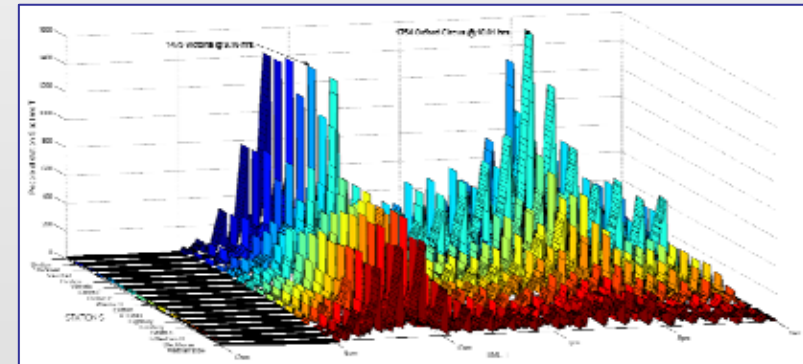
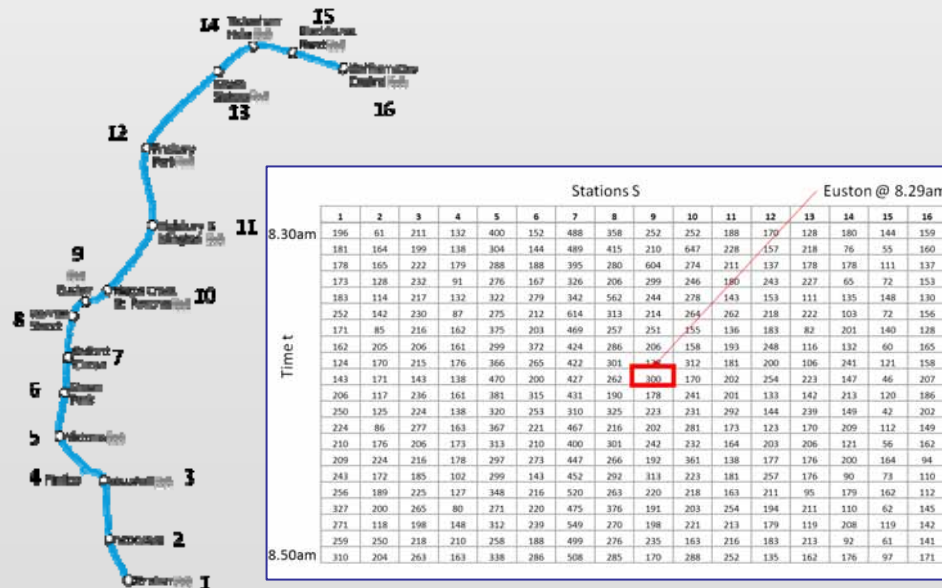
## Key

-  National Rail more than 5 minutes late
-  Tube stations showing a wait time 15% above expected
-  Bus stops showing a wait time 20% above expected

Tube delays from the TfL status feed are also plotted as lines

# Locational Dynamics of Demand

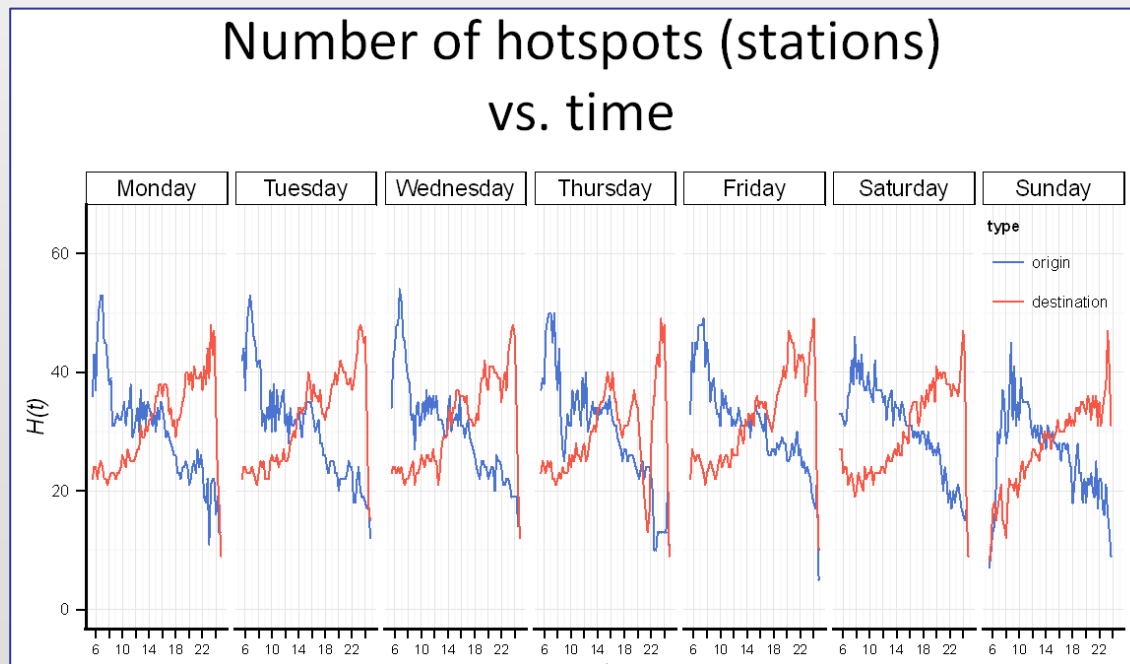
We are currently using information theory to figure out how much information from trips is transmitted from station to station through time by working out how many passengers are in stations or on trains in stations over time. We are using the concept of **transfer entropy** to do this. I don't have time to say much about this but here is a picture about this for one line



$$T_{YX} = \sum_{t=1} p(y_{t+1}, y_t, x_t) \log \frac{p(y_{t+1}|y_t, x_t)}{p(y_{t+1}|y_t)}$$

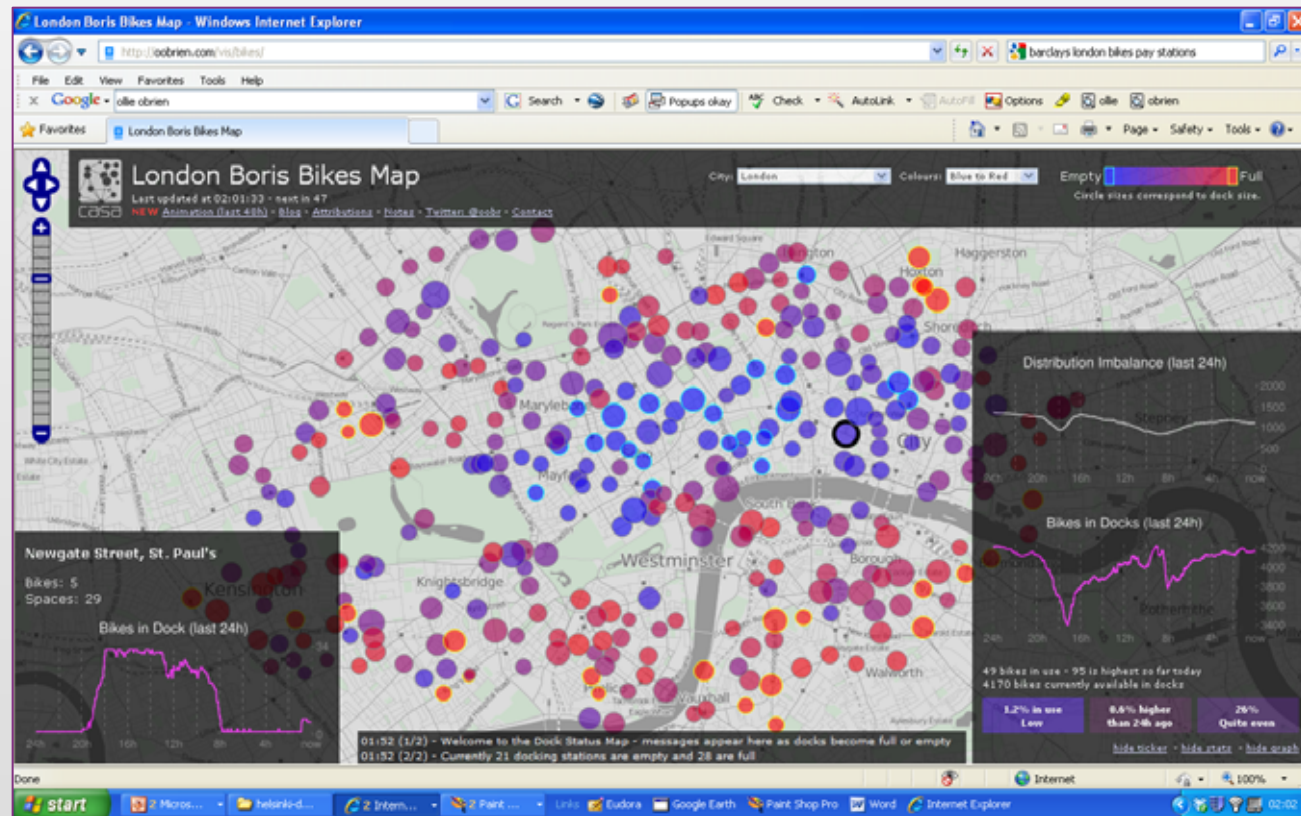


Second we are working with the Oyster data again with Melanie Bosredon in our group and Marc Barthelemy in Paris on extracting clusters from the travel data using a new method of defining intensity. I will show this as a simple movie of origin and destination intensities as they change over time of day.



# Related Real-Time Data: Bikes, Social Media

*A lot of data is now coming online for travel and one of our group Oliver O'Brien has some 97 bike schemes world wide for which he has online data in real time - Bikes Data – 4200 bikes, started Nov 2010, all the data- everything – all trips, all times, all stations/docks*





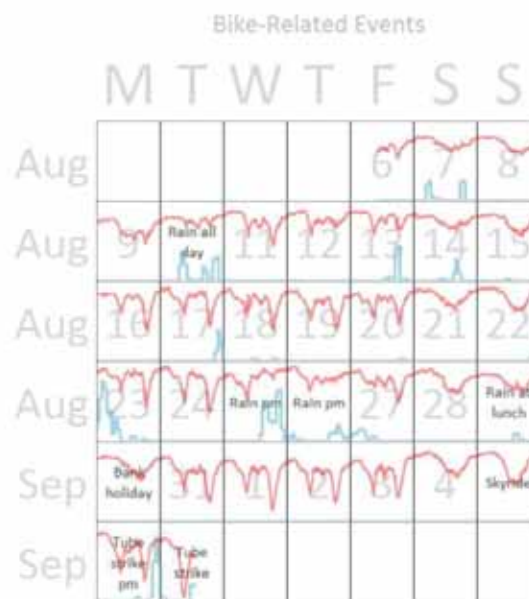
*Animations of Public Bike Movements*



*Animations of Changes in the Bike Nodes: Docking*

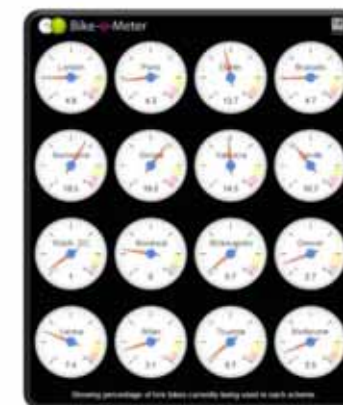
## More Analysis

- **London**
- Graph shows number of bikes available to hire
- Effect of rain
  - Using the CASA weather station
- Effect of the tube strikes



## Bike-o-Meter [casa.ucl.ac.uk/bom](http://casa.ucl.ac.uk/bom)

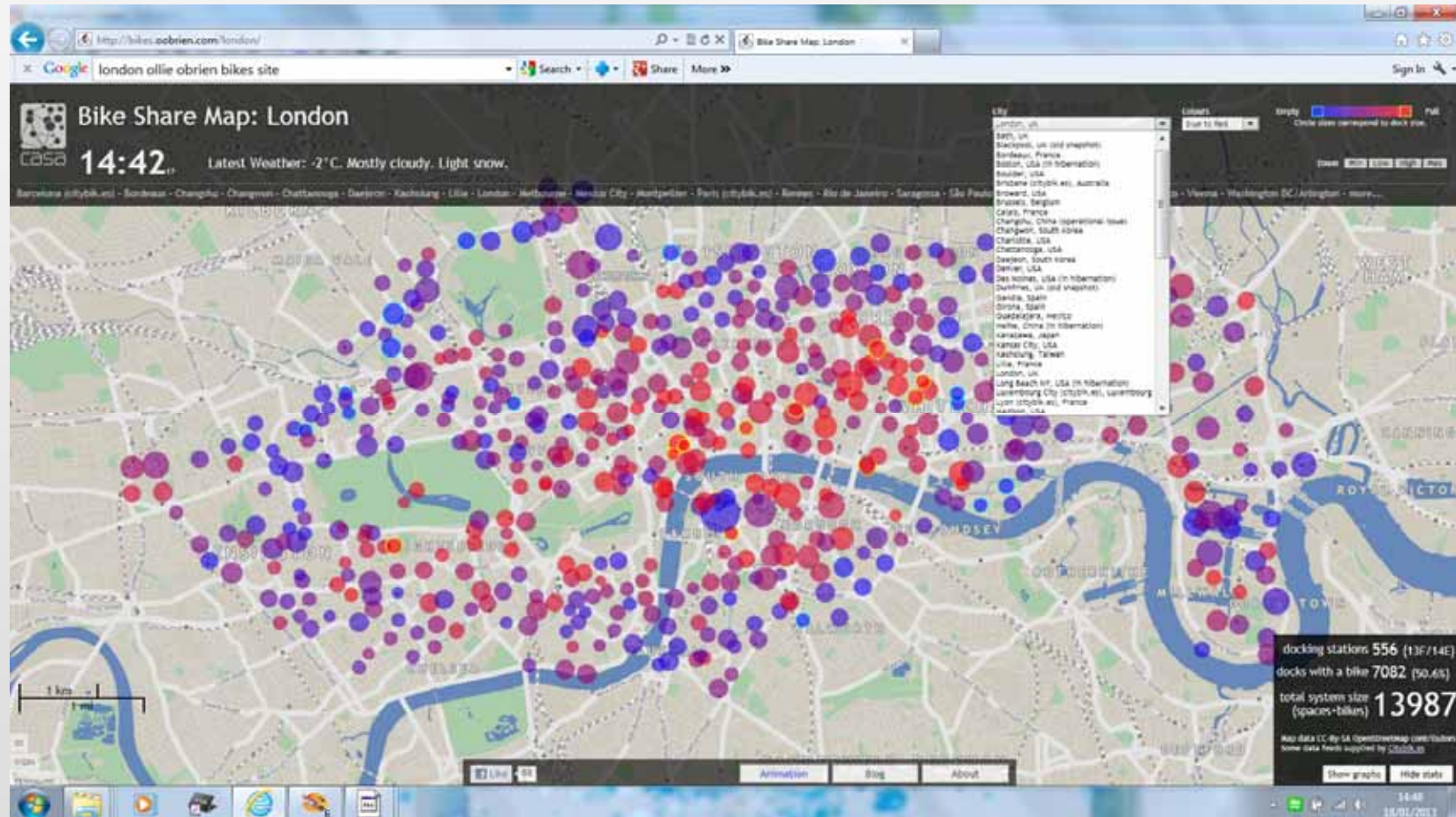
- Tweet-o-Meter for bikes
  - Steven Gray (@frogo)
  - Using Google Gauges
- See the real life Tweet-o-Meters at the new British Library "Growing Knowledge" exhibition
  - Should be easy to hack to show the Bike-o-Meters instead ☺



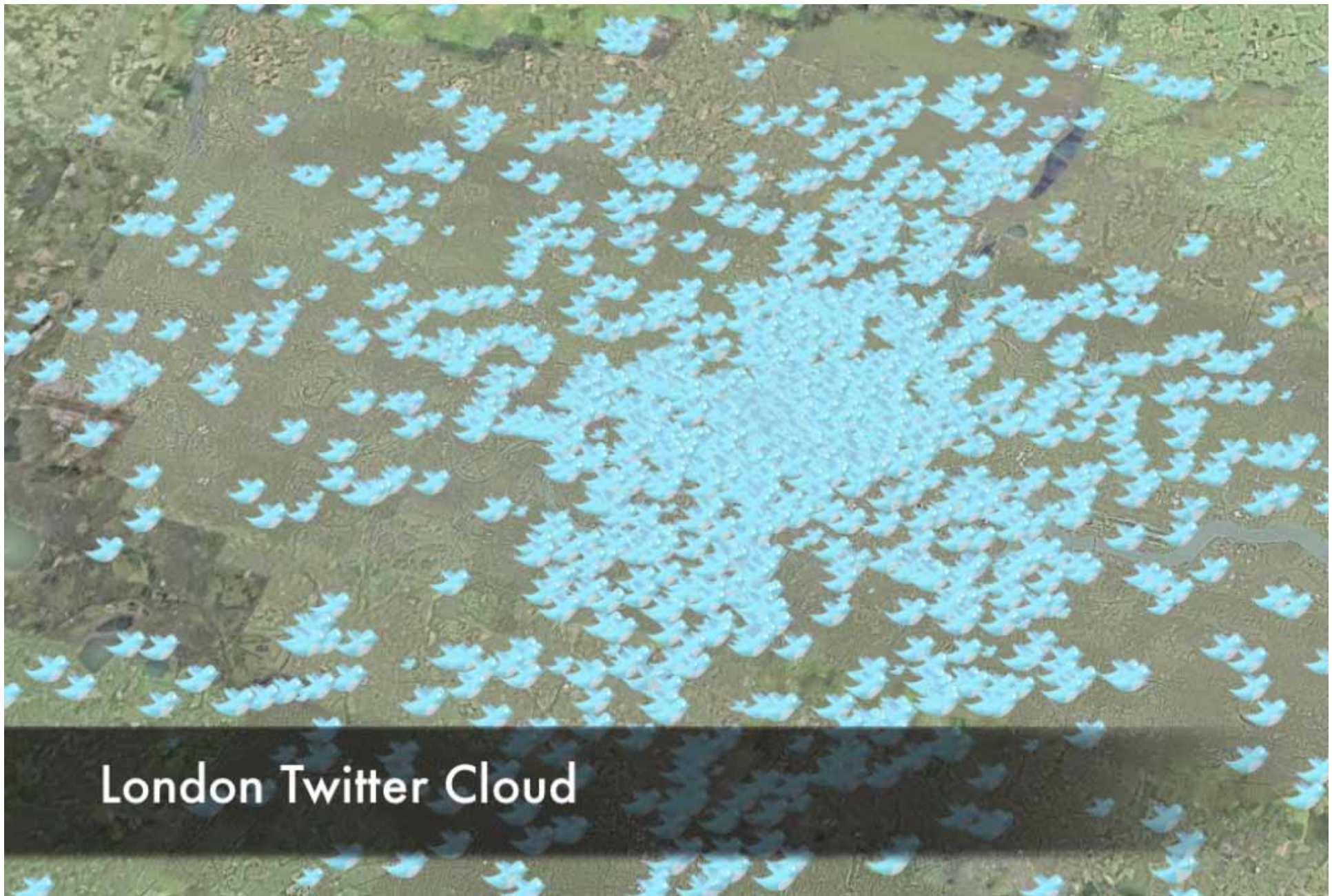


# The Website: Real Time Visualisation of Origins and Destinations Activity

<http://bikes.oobrien.com/london/>

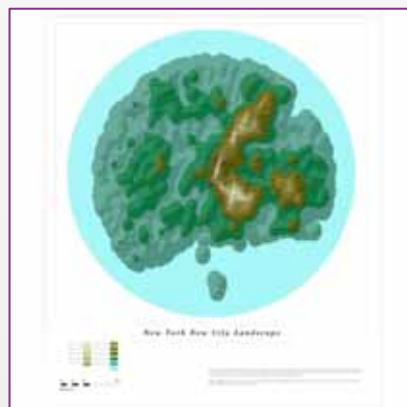




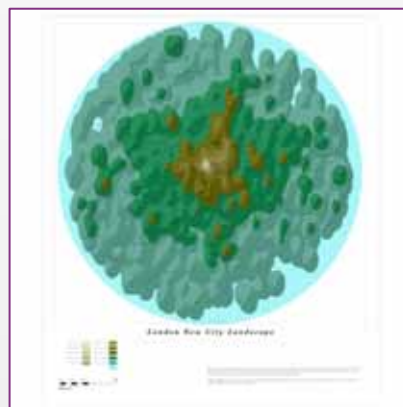


## London Twitter Cloud

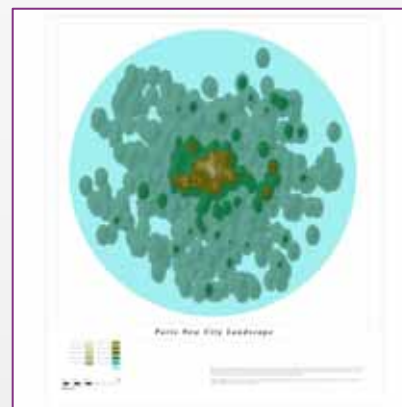




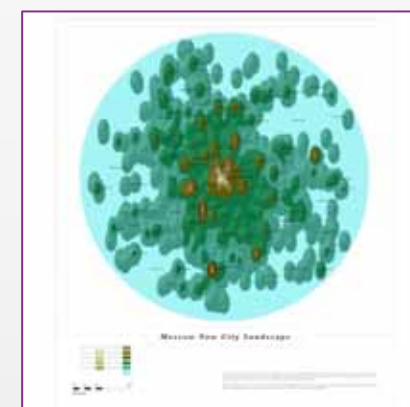
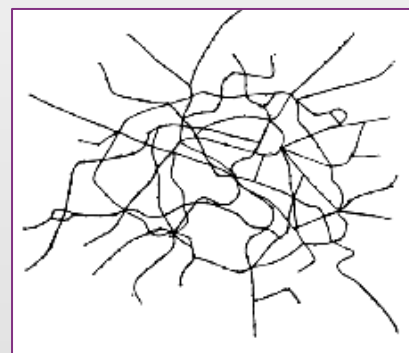
New York



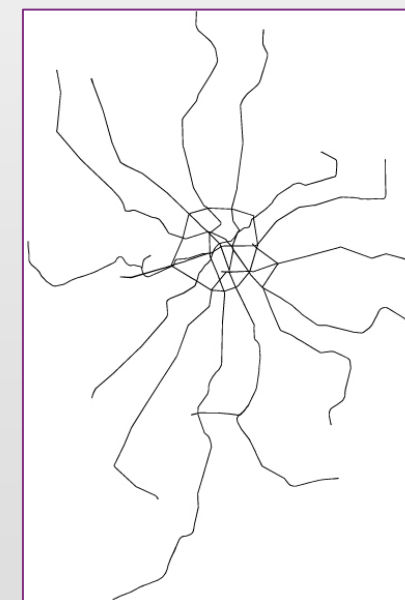
London

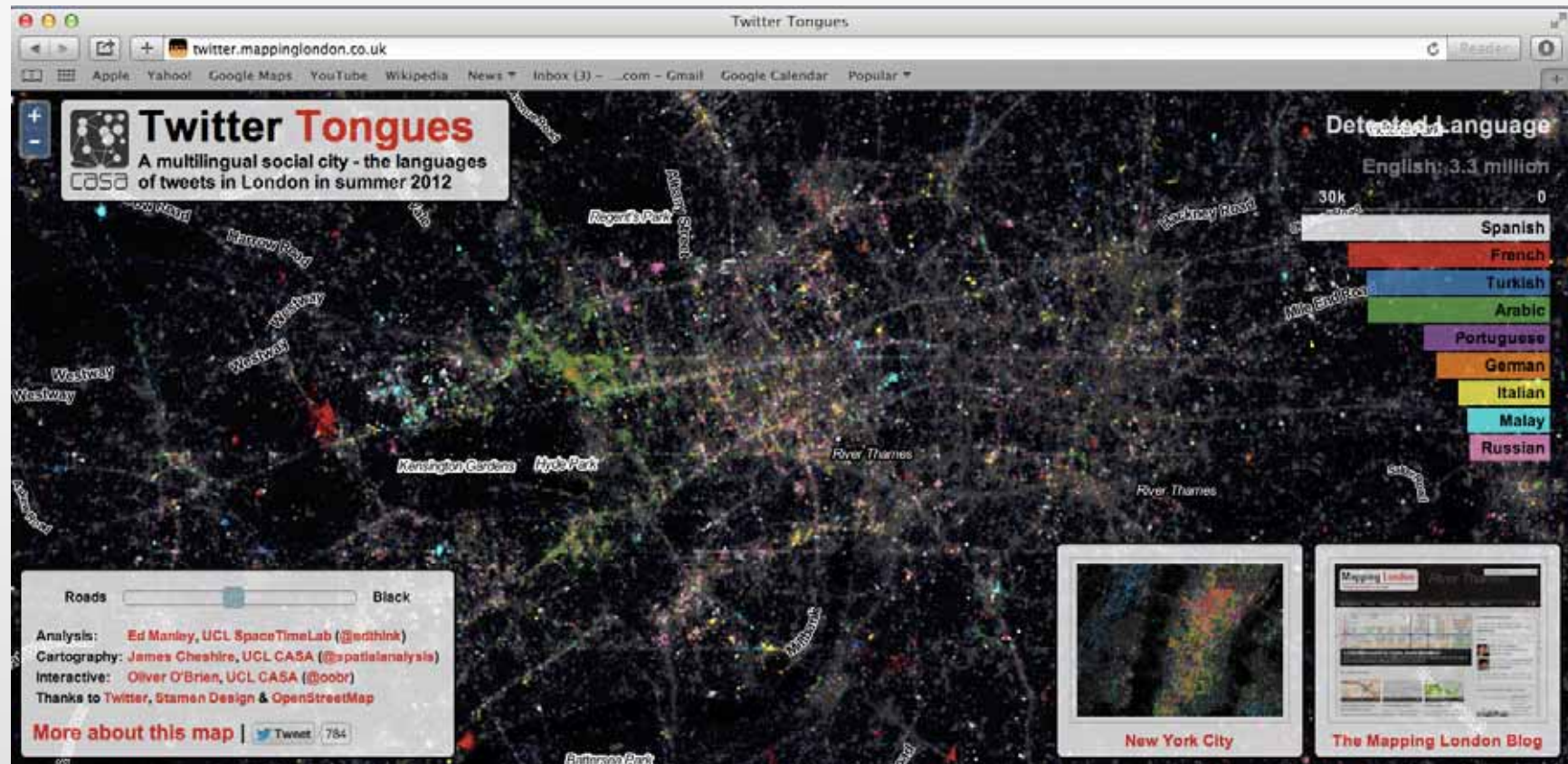


Paris

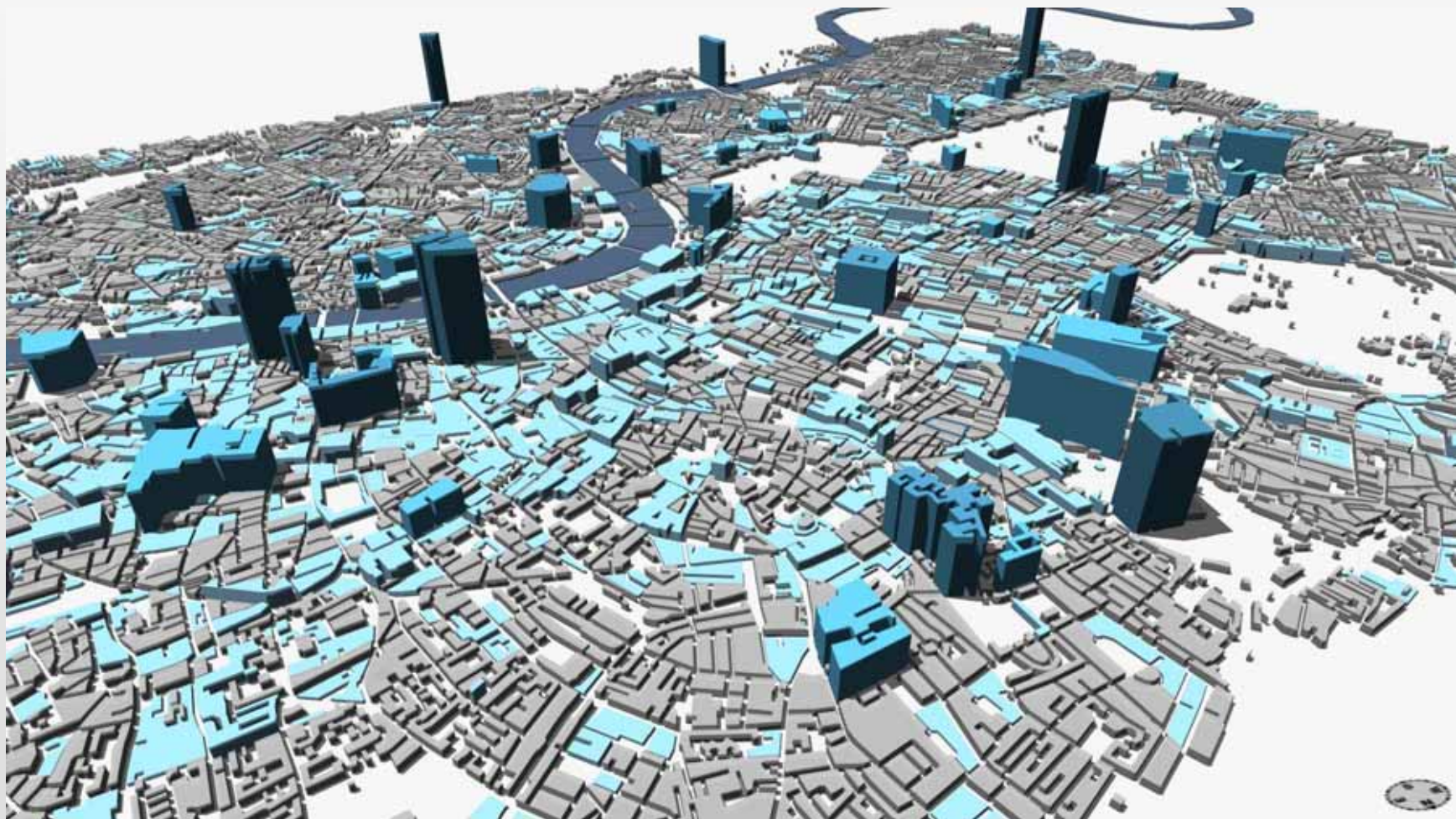


Moscow









# What Can We Learn: The Limits to Big Data

We need to add geo-demographics to this data – how  
– we barely have any possibility of doing this  
because of confidentiality

We only have a difference between young and old in  
terms of the card data

Chen Zhong my post doc has done a lot of work on this  
relating to extracting such data from related data  
sets producing synthetic results –our paper in IJGIS

*International Journal of Geographical Information Science*, 2014  
<http://dx.doi.org/10.1080/13658816.2014.914521>



## **Detecting the dynamics of urban structure through spatial network analysis**

Chen Zhong<sup>a\*</sup>, Stefan Müller Arisona<sup>a,b</sup>, Xianfeng Huang<sup>c</sup>, Michael Batty<sup>d</sup>  
and Gerhard Schmitt<sup>a</sup>

# References

Manley, E., Chen, Z., and Batty, M. (2016) Spatiotemporal Variation in Travel Regularity through Transit User Profiling, to be submitted.

O'Brien, O, Cheshire, J. and Batty (2014) Mining Bicycle Sharing Data for Generating Insights in Sustainable Transport Systems, **Journal of Transport Geography**, **34**, 262–273

Roth C., Kang S. M., Batty, M., and Barthelemy, M. (2011) Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. **PLoS ONE 6(1)**: e15923. doi:10.1371/journal.pone.0015923

Zhong, C., Arisona, S. M., Huang, X., Schmitt, G. and Batty, M. (2014)) Detecting the Dynamics of Urban Structure through Spatial Network Analysis, **International Journal of Geographical Information Science**, <http://dx.doi.org/10.1080/13658816.2014.914521>

Zhong, C., Batty, M., Manley, E., Wan, J., Wang, Z., Che, F., and Schmitt, G. (2016) Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing using Smart-Card Data., **PLOS One**, <http://dx.doi.org/10.1371/journal.pone.0149222>

Zhong, C., Huang, X., Arisona, S. M., Schmitt, G., and Batty, M. (2014) Inferring building functions from a probabilistic model using public transportation data, **Computers, Environment and Urban Systems**, **48**, 124–137

Zhong, C., Manley, E., Stefan Muller Arisona, S., Batty, M., and Schmitt, G. (2015) Measuring Variability of Mobility Patterns from Multiday Smart-card Data, **Journal of Computational Science**, doi.org/doi:10.1016/j.jocs.2015.04.021





# Thanks

## Acknowledgements

Melanie Bosredon, Gareth Simons, Roberto Murcio, Richard Milton,  
Oliver O'Brien, Stephen Gray, Fabian Neuhaus, Jon Reades, Ed  
Manley, Chen Zhong, Flora Roumpani & Stephan Hugel

<http://www.complexcity.info/>  
<http://www.spatialcomplexity.info/>  
<http://blogs.casa.ucl.ac.uk/>

[m.batty@ucl.ac.uk](mailto:m.batty@ucl.ac.uk)



@j michaelbatty