Office for National Statistics

ONS Big Data Project

[redacted]

Office for National Statistics

Plan for today

- Introduce the ONS Big Data Project
- Provide a brief overview of our work to date
- Provide information about our future work plan

What is Big Data?

"Big data are high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Gartner 2012)

Volume
- exceeds limits of traditional column and row databases
- constantly growing

Velocity
- arrives rapidly, often in real time

Variety
- does not have a standard structure, e.g. text, images

How is big data generated?

Sensors gathering information: e.g. Climate, traffic etc.

Social media: posts, pictures and videos

Google Site Search

Digital satellite images

Purchase transaction records

GPS

Mobile phone GPS signals

High volume administrative & transactional records

What is the ONS Big Data Project?

- A project which aims to:
  - investigate the potential for big data in official statistics while understanding the challenges
  - establish an ONS policy and longer term strategy which incorporates ONS's position within Government and internationally in this field
  - Recommend next steps to support the strategy going forward
- First phase from January 2014 – March 2015
- Now extended for another year

Big Data Project work packages

- Management and Strategy
- Stakeholder Engagement
- Communication
- Analysis and infrastructure:

Smart meters

Pilots

#Twitter

Prices

TESCO Every little helps

Waitrose

Sainsbury's

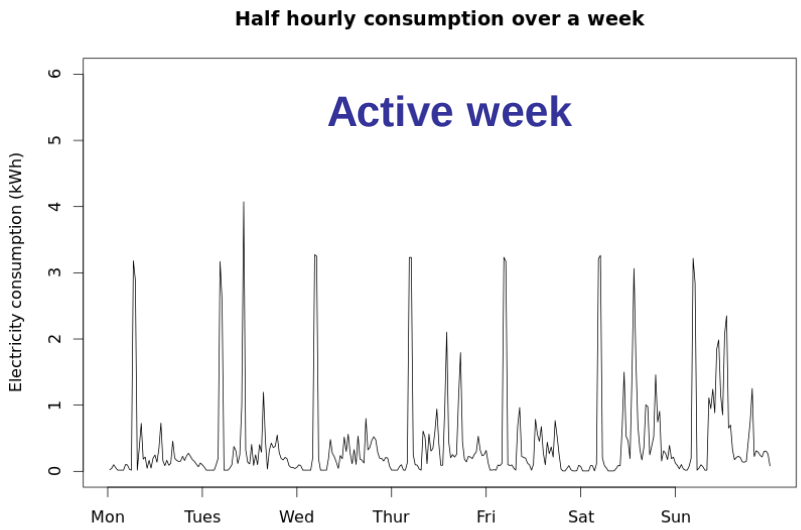
Mobile phones

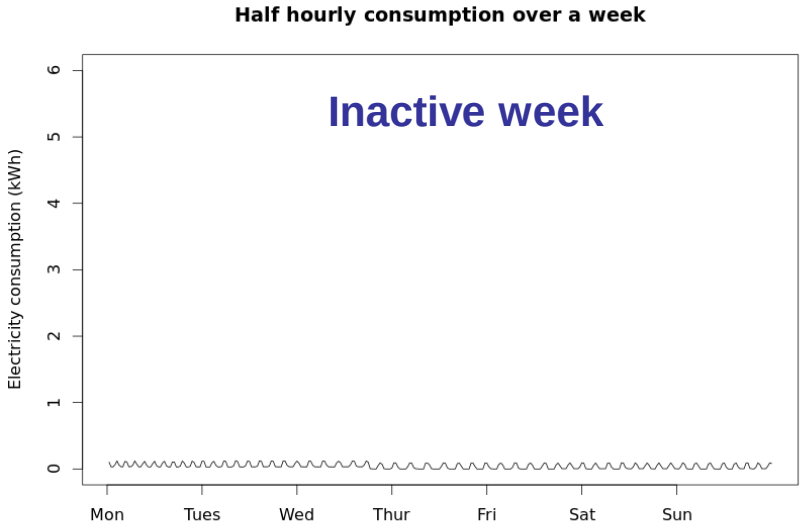
Pilot 1: Smart-type meters

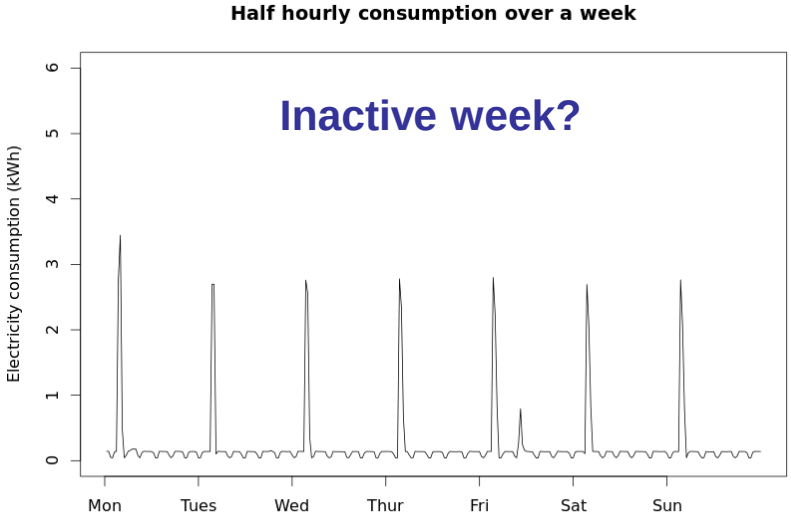
Research Question: Investigate the potential of smart-type meter electricity data (high frequency – 30 mins) to model likelihood of household occupancy patterns

- More efficient response chasing
- Data from smart-type meter trials in Great Britain and Republic of Ireland
- A range of potential methods identified
- Need to be careful of privacy and ethics

Smart-type Meter Energy Use Profiles







Pilot 2: Mobile Phones

Mobile phone data to model population flows, e.g. Commuting statistics

- Building relationships with mobile network operators and other parts of UK Government
- No data yet. Seeking better coordinated data access for Government
- Privacy and ethics (again)

Pilot 3: Prices Project

Research Question: To investigate how we can scrape prices data from the internet and how this data could be used within price statistics

- ONS prices collection is manual
- Web scraping promises more detailed, more frequent and cheaper data
- Prototype web scrapers:
  - 35 CPI/RPI item categories
  - 3 supermarkets
  - Daily collection since April (around 6,500 a day)

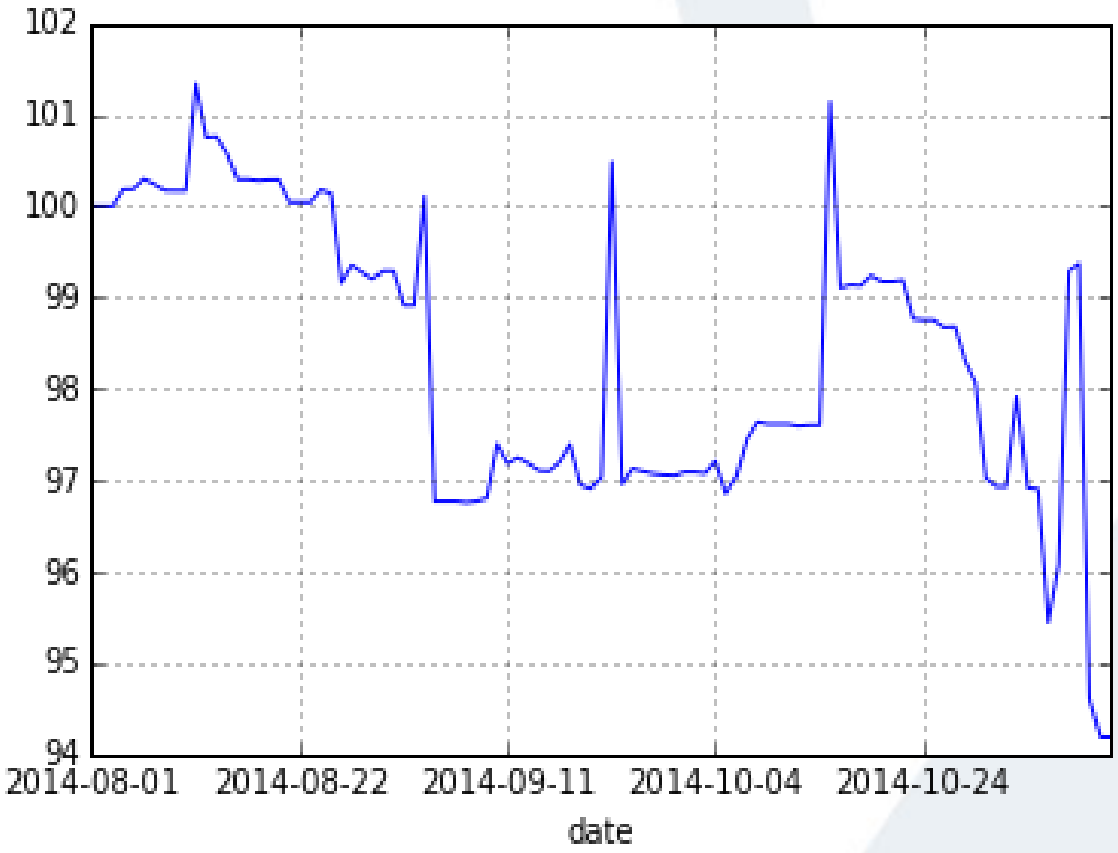
Pilot 3: Prices by webscraping

Rendered webpage:

HTML code:

```
.....
</div><div class="productLists" id="endFacets-1"><ul class="cf products line"><li id="p-254942348-3" class=" first"><div
      class="desc"><h3 class="inBasketInfoContainer"><a id="h-254942348" href="/groceries/Product/Details/?id=254942348"
      class="si_pl_254942348-title"><span class="image"><img
      src="http://img.tesco.com/Groceries/pi/121\5010044000121\IDShot_90x90.jpg" alt="" /><!----></span>Warburtons Toastie Sliced
      White Bread 800G</a></h3><p class="limitedLife"><a href="http://www.tesco.com/groceries/zones/default.aspx?name=quality-and-
      freshness">Delivering the freshest food to your door- Find out more &gt;</a></p><div class="descContent"><!----><div
      class="promo"><a href="/groceries/SpecialOffers/SpecialOfferDetail/Default.aspx?promoId=A31234788" title="All products
      available for this offer" id="flyout-254942348-promo-A31234788--pos" class="promoFlyout"><span class="promoImgBox"><img
      src="/Groceries/UIAssets/I/Sites/Retail/Superstore/Online/Product/pos/2for.png" class="promoFlyout promo" alt="Special Offer"
      id="flyout-254942348-promo-A31234788--posimg" /></span><em>Any 2 for £2.00</em></a><span> valid from 21/1/2014 until
      10/2/2014</span></div><div class="tools"><div class="moreInfo"><a href="/groceries/Product/Details/?id=254942348"
      class="midiFlyout" id="flyout-254942348-midi-0-"><img class="midiFlyout hd"
      src="http://ui.tescoassets.com/groceries/UIAssets/I/../Compressed/I_635209615845382232/Sites/Retail/Superstore/Online/Product/i
      nfoBlue.gif" alt="" title="View product information" id="flyout-254942348-midi-1-" /></a></div><!----><div
      class="links"><ul><li><a
      href="http://www.tesco.com/groceries/product/browse/default.aspx?notepad=white%20sliced%20loaf%20800g&amp;N=4294793217"
      class="shelfFlyout active plaintooltip" id="s-tt-254942348" title="Premium White Bread"> Rest of <span class="hide">Premium
      White Bread <!----></span>shelf </a></li></ul></div></div></div></div><div class="quantity"><div class="content addToBasket"><p
      class="price"><span class="linePrice">£1.45<!----></span><span class="linePriceAbbr"> (£0.18/100g)</span></p><h4
      class="hide">Add to basket</h4><form method="post" id="fMultisearch-254942348"
.....
```

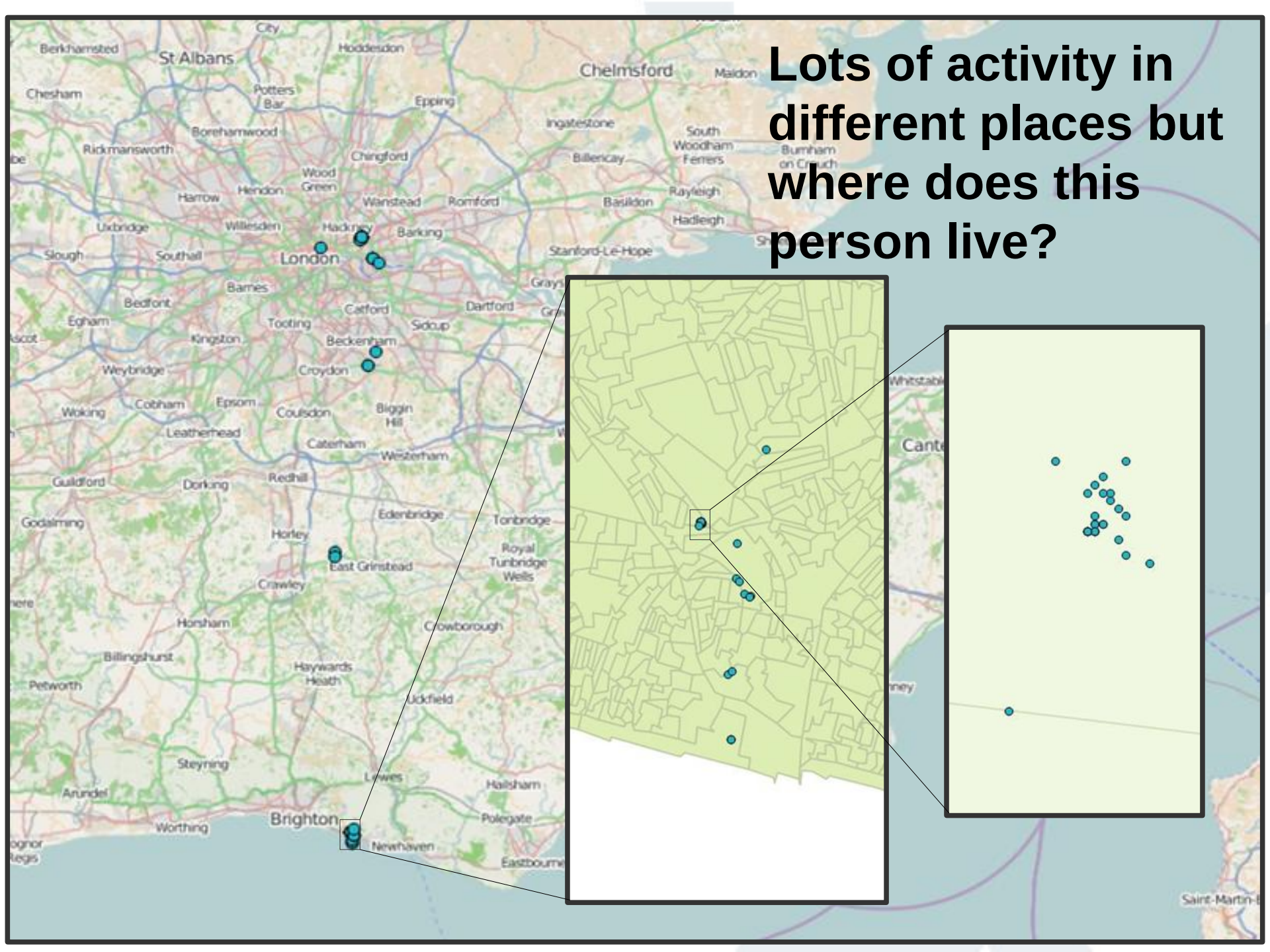
Daily Price Index (Whiskey)

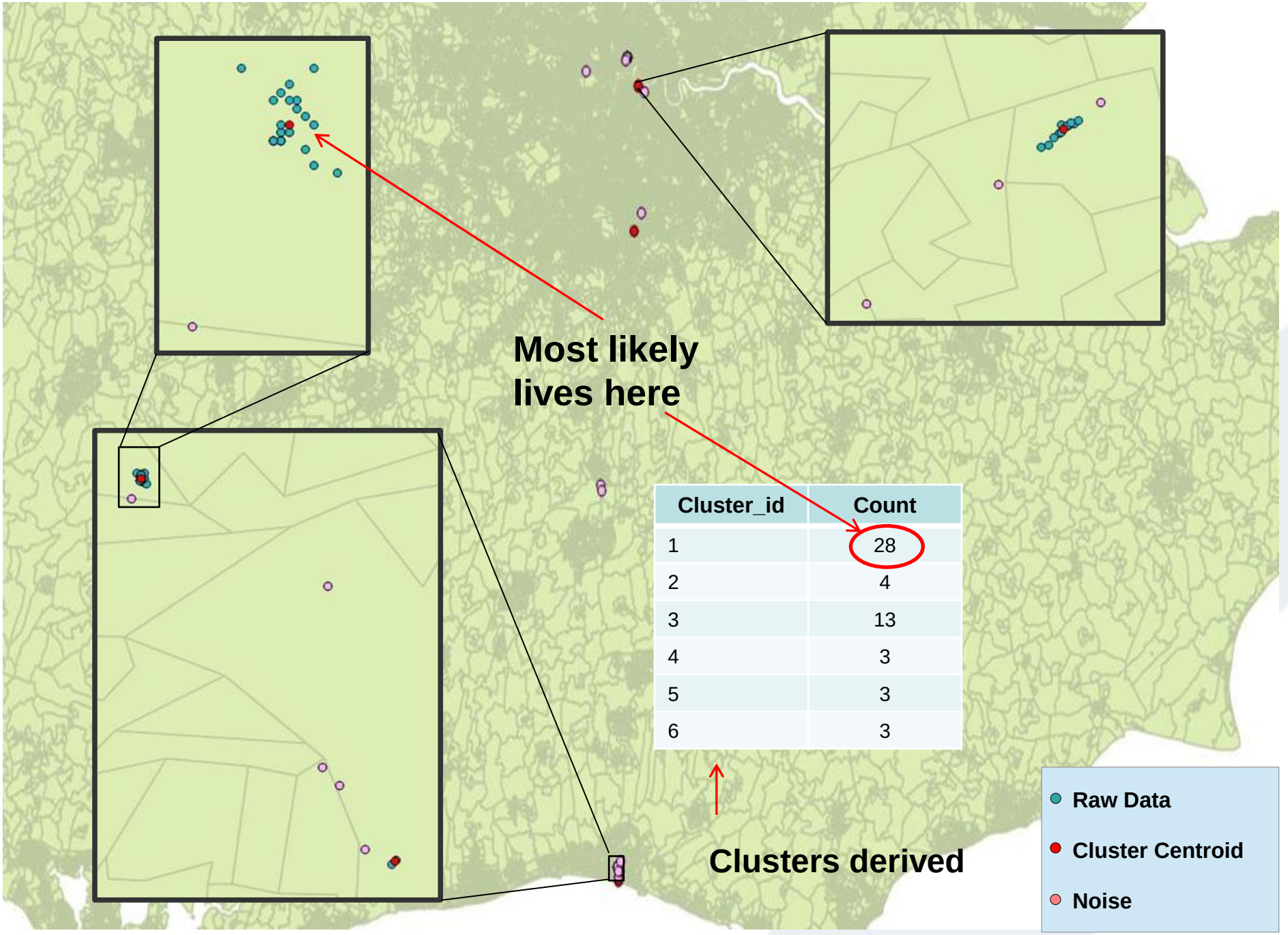


Pilot 4: Twitter Project

Research Question: To investigate how to capture geo-located tweets from Twitter and how this data might provide insights on internal migration

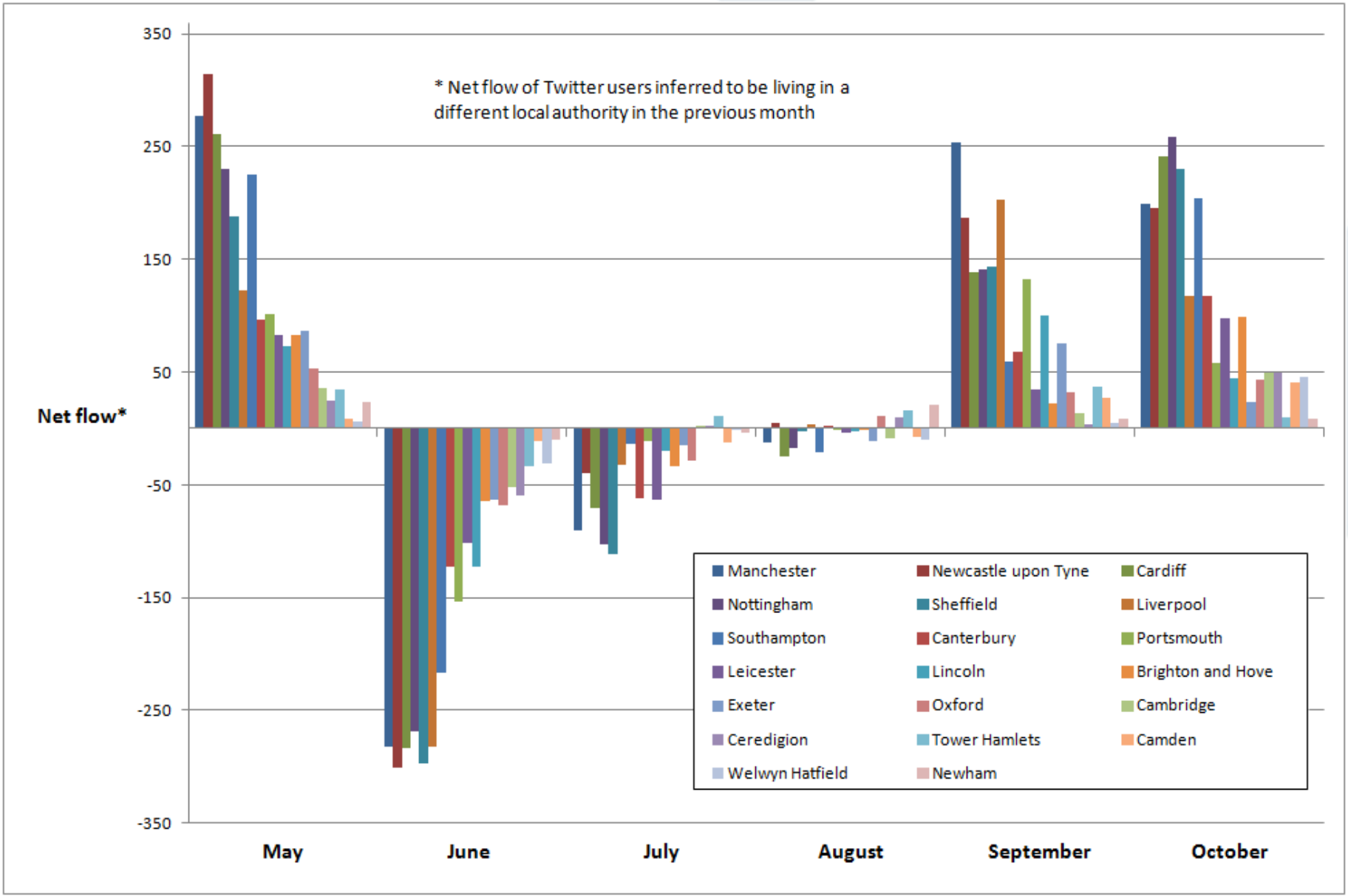
- 7 months of geo-located tweets within Great Britain (about 100 million data points)
- Methodology to infer place of usual residence:
  - Identify user ‘anchor points’ by clustering tweets
  - Identify residential anchor points using AddressBase and nearest neighbour analysis


Lots of activity in different places but where does this person live?
Berkhamsted
St Albans
City
Hoddesdon
Chelmsford
Maldon
Chesham
Potters Bar
Epping
Ingatestone
Borehamwood
South Woodham Ferrers
Rickmansworth
Chingford
Billericay
Burnham on Crouch
Wood Green
Harrow
Hendon
Wanstead
Romford
Basildon
Rayleigh
Hadleigh
Uxbridge
Willesden
Hackney
Barking
Slough
Southall
London
Stanford-Le-Hope
Barnes
Bedfont
Catford
Dartford
Egham
Tooting
Sidcup
Kingston
Beckenham
Weybridge
Croydon
Cobham
Epsom
Biggin Hill
Woking
Coulsdon
Leatherhead
Caterham
Westerham
Guildford
Dorking
Redhill
Godalming
Edenbridge
Tonbridge
Horley
Royal Tunbridge Wells
East Grinstead
Crawley
Horsham
Crowborough
Billingshurst
Haywards Heath
Petworth
Uckfield
Steyning
Lewes
Hailsham
Arundel
Brighton
Polegate
Worthing
Newhaven
Eastbourne



Cluster_id	Count
1	28
2	4
3	13
4	3
5	3
6	3

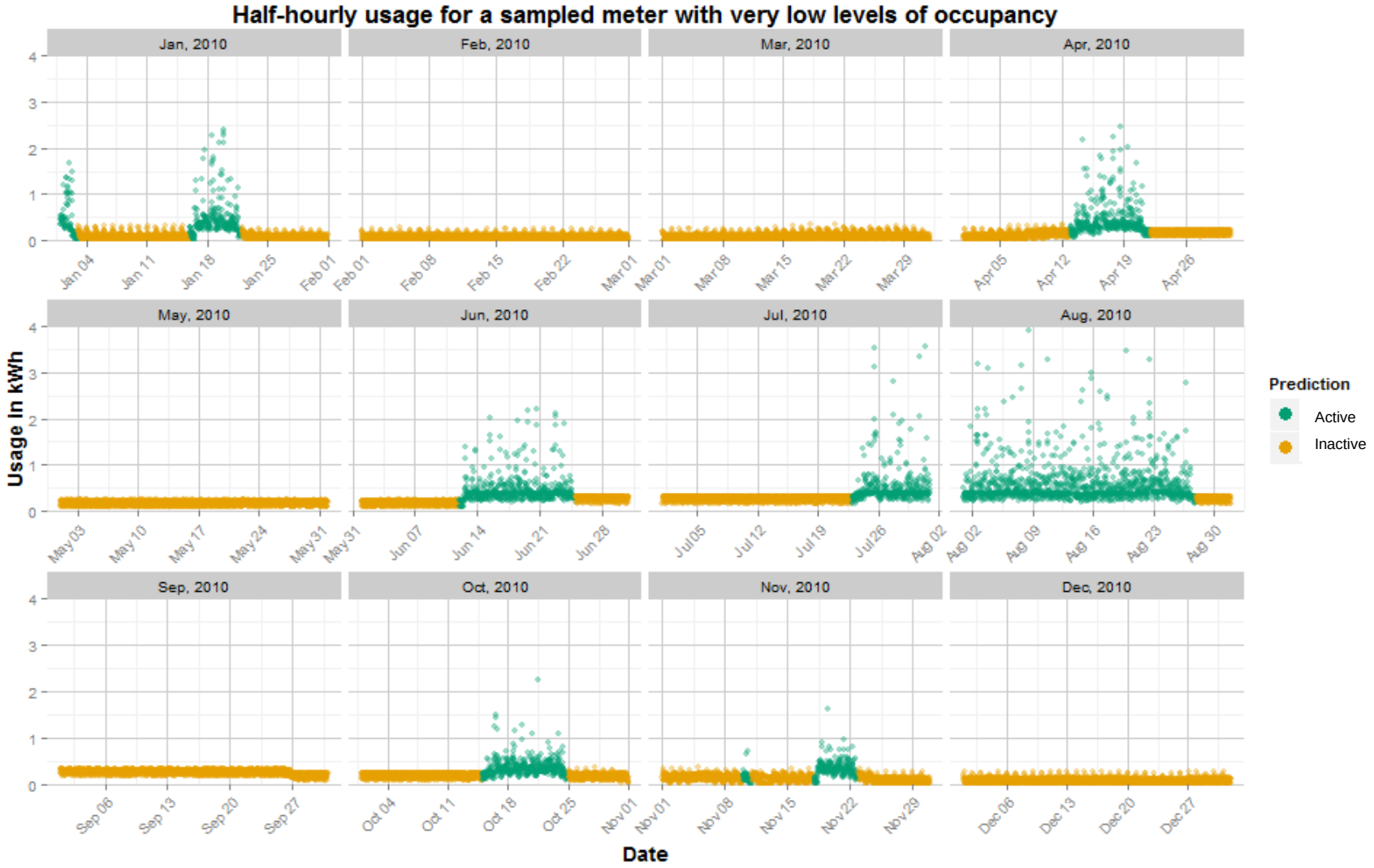
Use case: Student mobility



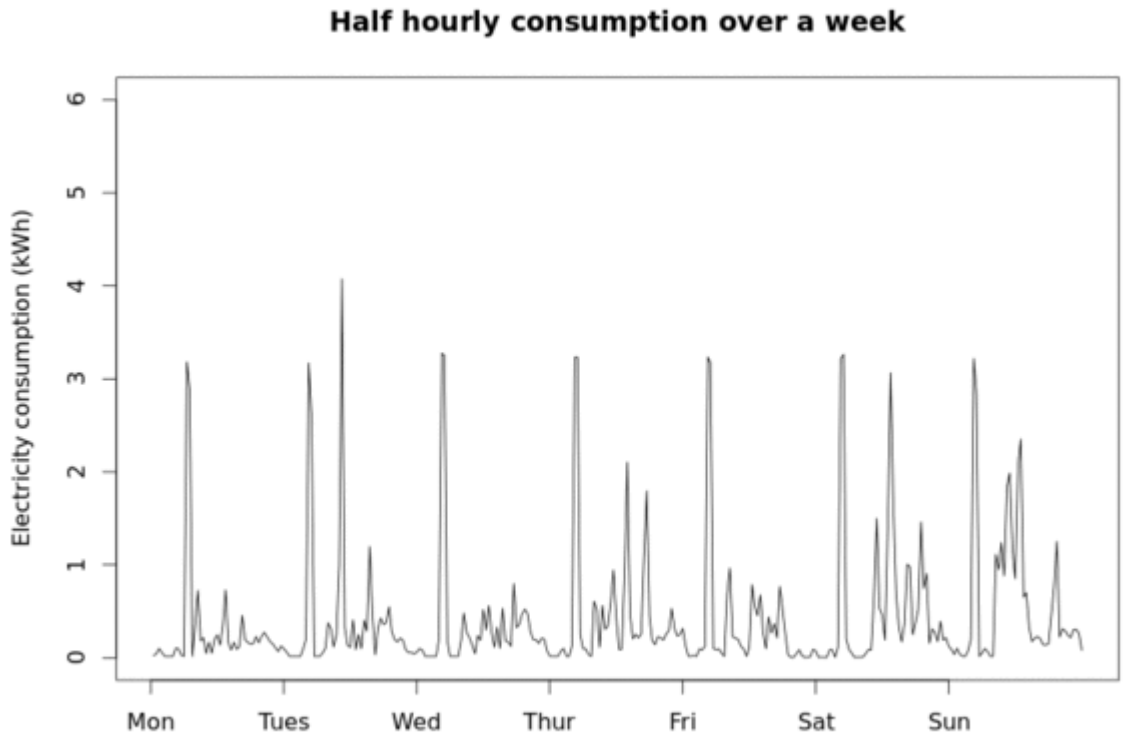
Presentation and dissemination challenges

- Volume of data presents challenge to presentation
- Used R software to automate visualisation for smart-type meter pilot
- Visualising summary statistics (like with ‘small data’)
- Could use interactive visualisations (eg. by using free software Tableau)
- Appropriate / ethical use of data needs to be clearly communicated

Use of R to display half hourly electricity data for a year



Use of R animations



Where to from here?

- Funding – 1 year more
- Prioritisation of pilots and other activities (Continue? New?)
- Improve our understanding of technology
- Team expansion (TBA)
- Establishment of Government Data Science Partnership to coordinate delivery and further development of data science in government

Questions

- ?