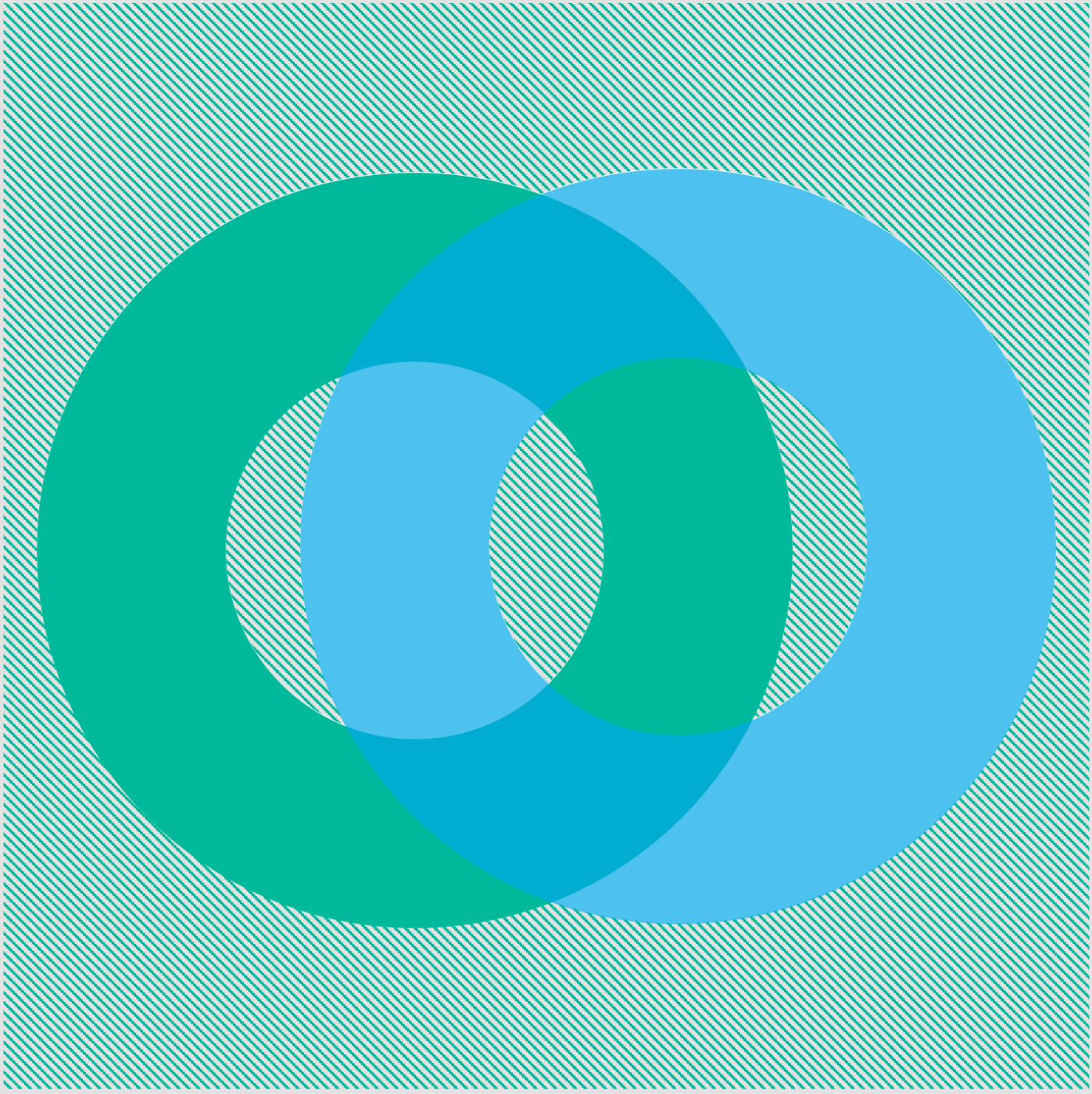
DEMOS

Exploring how to measure social integration using digital and online data

Alex Krasodomski-Jones
Elliot Jones
Hettie O'Brien
Andrew Gloag

December 2018



Executive Summary

In the last few years, the focus of social integration work in the UK has been on social contact between people from different backgrounds, often emphasising ethnicity and nationality as points of difference between people. While useful, this approach is incomplete. Understanding social integration, especially in an urban context, means looking at the extent to which people positively interact and connect with others who are different to themselves, across age, social class, sexuality, gender, disability and other social cleavages. It is determined by the level of equality between people, the nature of their relationships, and their degree of participation in the communities in which they live.

Existing surveys on social integration relating to London are often patchy in their coverage, and overly focused on migrants and ethnic minorities, rather than the central question of how social integration occurs across the population as a whole. This means understanding how individuals integrate across generations, social classes, educational differences and many other societal cleavages. Understanding alternative online and digital data sources could help to build a more comprehensive evidence base on social integration within London.

This report:

- Outlines an understanding of social integration
- Examines the existing work on measuring urban life through digital and online data
- Highlights the potential opportunities and pitfalls in using digital and online data
- Provides an analysis of a diverse selection of potential digital and online data sources
- Provides an outline of potential use cases for this data across the breadth of social integration measures.

There are four main takeaways from this report:

1. There are significant sources of digital and online data with the potential to inform policymakers' understanding of social integration.
2. Big digital data is best when used in conjunction with census and survey data that can verify observed behaviours and identify causal mechanisms.
3. Digital data can be an additional tool to "nowcast" the present between traditional active data collection methods.
4. Awareness of the representativeness problems inherent to digital data is essential for social research in this area, but these biases can be useful when properly considered.

Acknowledgements

We would like to thank the academics and subject specialists who generously gave up their time to be interviewed for this research and engaged in discussions which informed the final content of the report.

This project was commissioned by the Greater London Authority (GLA). We would also like to extend our thanks to Barry Fong, Vivienne Avery and Spencer Thompson at the GLA for their guidance, support and feedback through the project.

Introduction

Do people in London feel lonely? How likely are they to help their neighbours? How many volunteer, or engage in political causes? Social integration has emerged as a central policy concern in London. Measuring integration requires asking these questions, and more, but finding answers can prove difficult. Building an accurate picture of social integration in London would typically involve interviewing lots of people, undertaking surveys or inferring answers from aggregate statistics. Yet large datasets often arrive after many months have elapsed - meaning they are already out of date. Aggregate statistics rarely provide granular detail. Surveys only document self-reported attitudes rather than observed behaviour and are costly and time consuming to design and conduct.

Data from digital platforms have the potential to improve researchers' understanding of complex social science questions by providing insights into the observed behaviours of urban residents and real-time information about how cities are changing. The evolution of "big data", extremely large datasets that can be computationally analysed to reveal trends and human behaviours, along with the ubiquity of networked devices across the city, has produced a wealth of new information about the behaviours and attitudes of urban residents. For researchers trying to understand social integration across London's population, these sources could prove ground-breaking.

However, big data is far from a panacea. As with traditional forms of social science research, it possesses limitations. Users of social media platforms and internet enabled devices are often younger and from higher income backgrounds, presenting problems with selection bias and representativeness. Big data is plagued with the same issues of causal inference as traditional statistical methods, can be incomplete or poorly structured, and rarely solves identification problems on its own. Finally, while big data offers researchers a potential deluge of new information, accessing this information raises pertinent ethical questions about privacy and re-publication that are coming under increasing public scrutiny.

The goal of this report, authored by Demos and commissioned by the Greater London Authority (GLA), is to present a clear roadmap of the types of online and digital data that exist, and how the GLA could apply these sources to evaluate and measure social integration in London. In 2018, the GLA developed a three-pronged definition of social integration in their strategy outline, *All of Us*. Taking this as a starting point for this research, Demos assessed the categories of different online and digital data and analysed the extent to which they would be useful for gauging specific measures of social integration. This report reviews data sources that could bring greater insight to the GLA's understanding of social integration in London and aims to complement the work the GLA is currently doing to improve London's social evidence base through its social integration measures and Survey of Londoners.

The report begins with an overview of the area and of existing research and work taking place in this field. The report reviews 47 possible data sources that could be used by the GLA as part of their work, before offering some tentative conclusions and a number of areas we believe may be worth exploring further.

Background

The social integration strategies of urban and national policymakers have often focused on integrating migrants and refugees into a homogenous majority. Yet in a diverse city like London, where there is less of an obvious “majority” group, it’s clear that this model doesn’t fit. In the Mayor of London’s flagship report on their social integration strategy, published in March 2018, the GLA defined social integration as the extent to which people “positively interact and connect with others who are different to themselves”, the “level of equality between people”, the “nature of their relationships”, and “their degree of participation in the communities in which they live”. This definition is geared towards understanding social integration within a highly diverse context. It encompasses three categories; relationships, participation and equality. Within each category, the GLA pinpoints a number of measures that together build a comprehensive picture of social integration in London.

The GLA analysed how it could draw on existing data sources, including census and survey data, to create an evidence base for studying social integration in London. They concluded that existing surveys on social integration relating to London are often patchy in their coverage, and overly focused on migrants and ethnic minorities, rather than the central question of how social integration occurs across the population as a whole. This report aims to explore how online and digital data could help build a more comprehensive evidence base for social integration. Government and academic survey/administrative data were excluded from this report, as the GLA has greater knowledge of and access to these sources already.

Using big data to "nowcast": existing research

An emergent body of research exists that uses live digital data to answer social scientific questions. At the urban level, research from Glaeser, Kominers, Luca and Naik[1] show how Google Street View can be used to map processes of gentrification. By training a computer-learning algorithm to identify trends in the physical architecture of the city, researchers were able to predict levels of household income from images stored on Google's street navigation platform. Where urban economics has tended to be "detached from many physical aspects of the city" due to a lack of data on the physical attributes of urban space, they find that 'big data'-driven tools like Google Street View make it possible to drill down into the social and economic implications of the built environment, using computational learning to measure how social outcomes "may be influenced by urban space".

Other work by Glaeser, Kim and Luca[2] makes use of Yelp data to "nowcast" urban social change and map processes of gentrification. When combined with data on gentrification from the US Census, Federal Housing Finance agency and Streetscore (an algorithm using Google Street View), they find that neighbourhood gentrification can be predicted by the number of local grocery shops, cafes, restaurants and bars within an area. Broad-stroke social processes can be inferred from granular insight; they find that the entry of a new coffee shop into a zip code within a year is associated with a 0.5 per cent increase in housing prices - deducing significant social implications from small changes at the local economic level that are registered in Yelp data.

It's important to note that big data alone frequently cannot answer social science questions. Researchers often combine this data with existing census and survey datasets to yield greater predictive power and interpret results where correlations are partial. While big data can produce a real-time picture of urban trends and social processes, sources like social media data also suffer from selection bias, where particular age or demographic groups are more likely to make use of digital platforms like Facebook, Twitter and Instagram. As one interviewee from a geographic information software firm told us:

"You absolutely have to use census data or survey data in addition to open data, particularly when you're dealing with social media, where there are biases that you can't get rid of. That's why you need other data sets, like large-scale surveys, to enhance and verify your dataset".

This is why the concept of "nowcasting" is central to understanding the potential uses and pitfalls of using online and digital data to understand social and demographic trends. "Nowcasting" refers to forecasting on a very short timescale, where researchers use available data to predict the present or near future. Where survey and census data can

[1] Edward Glaeser, Scott Duke Kominers, Michael Luca and Nikhil Naik, *Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life*. Working Paper no. 21778, December (2015). Accessed at: https://www.nber.org/papers/w21778

[2] Edward Glaeser, Hyunjin Kim and Michael Luca, *Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity.* Working Paper no. 24010, November (2017). Accessed at: https://www.nber.org/papers/w24010

provide robust evidence of how social and demographic patterns change over time, using digital data to nowcast equips researchers with an “ear to the ground” and an extra tool for shedding light on the present. Big data does not become a replacement for large-scale surveys and census research, but an additional means of gauging and reflecting on the current state of play, and anticipating what may happen in the future. In this regard, big data is likely to be most useful as an interim means of measuring social integration at intervals between survey research.

In their second work on nowcasting influenza, Preis and Moat[3] make use of Google search queries to provide real-time estimates of influenza-like illness across a population. Their model, which combines real-time Google flu search trends with historic data on influenza levels, allows them to estimate current levels of influenza before the release of official data one week later. In this way, nowcasting provides an additional means of estimating current levels of influenza in the interim of official data, reflecting the contention that nowcasting is most useful as an interval measurement that supplements official data sources.

Building on another example[4] of using Google search terms to nowcast influenza-like illnesses across a population, researchers at the University of Bristol use Twitter data to nowcast the “mood of the nation”.[5] Twitter possesses another advantage for nowcasting; tweets are “of-the-moment” and “sent on impulse”, the researchers note, with an “immediacy” that reflects what the sender is feeling at the time, in contrast to the considered opinions that are conveyed in survey data. Drawing on a series of tweets sampled from the 54 largest cities in the UK over a period of 30 months, and analysing these for sentiment, the researchers mapped emotional changes over time. Yet “correlations … are not causations”, they warn. Where data analysis proves patchy, the interpretive methods of social scientists must weigh in - again affirming how big data is supplementary to, rather than a replacement for, traditional social science research methods.

Twitter has also been used to predict gentrification across London boroughs[6]. By combining a dataset of 37,000 Twitter users and 42,000 venues in London, researchers built a network of places listed on Foursquare and mapped Twitter social networks through people checking-in to locations. This allowed them to “distinguish between places that bring together strangers versus those which tend to bring together friends, as well as places that attract diverse individuals as opposed to those which attract regulars”. By

---

[3] Tobias Preis and Helen Susannah Moat, *Adaptive Nowcasting of Influenza Outbreaks using Google searches.* Royal Society Open Science 1.2, October (2014). Accessed at: https://royalsocietypublishing.org/doi/full/10.1098/rsos.140095

[4] Jeremy Ginsberg, M.S. Mohebbi, R.S. Patel, Lynette Brammer, Mark Smolinski, L. Brilliant, *Detecting influenza epidemics using search engine query data.* Nature 19.457 (7232), February (2009). Accessed at https://www.ncbi.nlm.nih.gov/pubmed/19020500

[5] Thomas Lansdall-Welfare, Vasileios Lampos and Nello Cristianini, *Nowcasting the mood of the nation.* Significance Royal Statistical Society, August (2012). Accessed at: https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2012.00588.x

[6] D. Histrova, M.J. Williams, M. Musolesi, P. Panzarasa and C. Mascolo. *Measuring Urban Social Diversity using Interconnected Geo-Social Networks.* Proceedings of the 25th Conference on World Wide Web, 21-30, April (2016). Accessed at: https://www.repository.cam.ac.uk/handle/1810/253603

correlating these properties with wellbeing indicators in London neighbourhoods, they discovered “signals of gentrification in deprived areas”. The researchers warned that using social media data in population studies can be problematic due to the inherent bias of such platforms, whose users are more likely to be affluent, with greater social mobility. But the study also demonstrated how such biases can become an advantage when determining whether an area is becoming gentrified, capturing the user demography of social media platforms to researchers’ advantage.

Three salient findings emerge from existing studies that use digital and online data to answer complex social-scientific questions. First, big data works best when combined with aggregate statistical data such as that found in censuses and surveys that enhance and verify observed behaviours. Second, and relatedly, online and digital data is not currently used as a replacement for the role of traditional social-scientific methodologies including interviews and surveys, but as an additional tool to “nowcast” the present. Finally, the biases that run through social media usage, with particular demographic groups disproportionately represented on certain social media platforms, are an obstacle to researchers using big data for social research. As research measuring gentrification using Foursquare and Twitter indicates[7], there are cases where biases can be harnessed to researchers’ advantage. However, being cognisant of the problems with sample representativeness inherent to social media data is essential when making use of online and digital data in social research.

[7] ibid.

Data Source Reviews

This section reviews 47 data sources identified as potentially valuable to the GLA. Sites were chosen by whether analysts felt they supported an analysis of one of the GLA's existing measures of social integration, or if a specific example in the wider literature on measuring social integration through digital sources mentioned a data source. Within each data source category that was analysed, we focused on sources with large user bases to lean towards sites that might provide sample sizes that were more representative of the London population.

In some cases, websites were not reviewed when they were overly similar to sources deemed inappropriate or inaccessible to the GLA. In the case of dating apps, for instance, we could have considered Minder – an app targeting Muslim Londoners, or Grindr, an app for LGBT+ users. However, lessons from Tinder and Bumble above suggest this style of app would not meet the lowest thresholds for utility. Others were not comprehensively analysed as there was no publicly available way of accessing data and any access would be based on a bespoke partnership arrangement between the GLA and the data provider, of which this report cannot provide meaningful analysis.

The team received valuable feedback from a number of academic reviewers regarding sources that were not considered. Some data sources, such as gov.uk and local authority online data, were deemed out of scope for this report as those data sources are more easily accessed and reviewed by the GLA internally. Valuable inclusions for a similar analysis would include alternative platforms to the more mainstream examples discussed below, particularly those used more frequently by minority groups. Although analysts considered including these platforms, there was concern that they and the data they could provide would not be sufficiently large to include in studies of the London-wide population.

Each source was scored against thirteen metrics, with analysts deciding on a score from A down to D. For instance, a data source that provided structured data through a public API would score an A for data, while one that expressly forbade scraping and did not provide an API would likely score a D. An indication of what A and D represent for each category is shown in the table below.

Accessibility & Availability	Data	D is impossible to access, A is easy
	Current Availability	D is very high risk the data will stop being accessible or will be incomparably different, A is it'll likely be there for the foreseeable future
	Future Availability	
Granularity, Bias, Detail and Representivity	Level of Detail	D is a single aggregated statistic across the whole population, A is multiple useful factors at an invidual citizen level
	Geographic Depth	D is UK wide only, A is granularity at an individual or postcode level
	Utility	D is the data is impossible to draw any conclusions from whatsoever, A is a fully representative survey
	Frequency of Updates	D is never, A is hourly+
	Representativeness	D is completely unrepresentative of the GLA, A reflects the GLA with a high degree of completeness.
Ethics	Ethics of Access	D requires theft or a breach of stated terms of service, A is a publicly supported API supported by user agreement.
	Ethics of Publishing	D is publishing highly sensitive personal data without consent, A is publishing data that users have agreed and understood would be made public
	Perceptions of Privacy	D is comparable to medical records, A to a person's name.
Technology Considerations	Skills Required	A is a calculator, D is a PHD and a DeepMind Supercomputer
	Costs of Ongoing Access	Cost/unit

The names of each data source are highlighted based on their potential current utility to the GLA. Green is useful, orange is unclear and red is likely useless. A full list of sources considered is in Appendix A.

Sources are presented in three tiers of estimated usefulness to the GLA, then alphabetically within each tier.

Data Sources

EventBrite	Airbnb	Bumble
Foursquare	CityMapper	Change.Org*
Google Maps	Find a Job	Dice*
Google Street View	Freecycle	DuoLingo
Meetup	Indeed. com	Facebook
Mobile Company Data	JustGiving	Glassdoor
Trustpilot	Monster	Google Search Trends*
Twitter	Mumsnet	Gumtree
Yelp	Netmums	Instagram*
Youtube	Park Run	LinkedIn*
Zoopla	Reddit	Mobike*
		newsapi.org*
		Ofo*
		OpenTable*
		Oyster*
		Quiqup*
		RightMove*
		Snapchat*
		Spareroom*
		Tinder
		TripAdvisor
		Tumblr*
		Uber
		WhatsApp*
		Wikipedia*

Sources with an asterisk were considered but not written up fully given the data was either deemed unusable or access was prohibited.

Eventbrite

Eventbrite is an event planning tool used to advertise and sell tickets. It is widely used by a large variety of organisations, from small book clubs to large music events or conferences.

Data	**A**

Data is returned in a JSON format. Events, User and Venues are all separate returnable objects with different structures.

Current Availability	**A**

Access to Eventbrite data is available through a public API with extensive documentation.

Future Availability	**A**

Eventbrite is an established event service and especially following its Facebook integration, we expect the company to exist for the foreseeable future and they have given no indications of closing or limiting the API. We would expect the data to be accessible on demand over the next few years.

Level of Detail	**A**

Able to access the name, time/date, type, description, venue location and organisation hosting.

Utility	**A**

Likely to be useful for both relationships and participation. Association membership has traditionally been used as a measure of social capital which fits into the GLA wider model of social integration. Further, it can tell the GLA about the type of activity individuals are engaging in over an area - for example, a spike in language classes, particularly English language, might be a sign of increased attempts at integration in a local community. It also gives information about ticketing costs which could give information about economic based participation and equality, e.g. do certain types of events have a high cost barrier which may exclude those of lower socio-economic status?

Geographic Depth	**B**

Not all events provide a venue (either not specified or not publicly visible), but a large proportion do, and those provide the full address of the venue, down to the street level.

Representativeness	**B**

Only a fraction of events on the platform take place in London, but most events can be specified by location, so it should be possible to isolate a sample from London. The types of events are broad but tend to skew towards the young and the tech-minded, e.g. a large number of coding/data science meet-ups.

Ethics of Access	**B**

Available by a public API, However, in the Eventbrite API's Terms of Use it specifies that you may store Site Content relating to future events, but you may not store any Site Content relating to events that have occurred in the past unless the user has given you explicit permission.

Ethics of Publishing	**B**

The API returns mostly anonymised data, e.g. the details of public events but not those who attended. The only concern is around the user data of event hosts, which should probably be anonymised before publishing - though we expect that data to be aggregated and anonymised before it is useful and therefore publishable.

Public Perceptions of Privacy	A

All information accessible through the API is publicly available and the platform is explicitly for attending events with others so users, especially those hosting the events, have strong reason to believe anything they post will be visible to strangers. Additionally, the API wouldn't allow access to unlisted events unless the user was the owner of those events.

Skills Required	A

The primary technical skills needed to interact with this API are moderate - analysts will need an ability to interpret technical specifications and follow documentation, and to store and process returned data. Familiarity with a major programming language (Python, Java, C++, PHP etc) will also be useful.

Frequency of Updates	A

Live but access to the API is limited to 1,000 calls per hour on each OAuth token

Costs of Ongoing Access	Free

See:

Eventbrite API https://www.eventbrite.com/platform/api/

Examples of Possible Use Cases

Measure: Participation

The number and popularity of events in a category taking place on the platform could be a good way to measure political, civic and leisure activity across the GLA. For example, measuring the number of participants and events held in the charities and causes category to measure civic participation, across London and between given geographic areas.

Foursquare

Foursquare is a local search platform that provides search results for its users and recommendations of where to go near users' location based on their previous browsing history, purchases and check-in history

Data	B

Available data is text, photos and user-generated reviews of places across the capital. Foursquare also stores data on where users check in and user-generated lists of favourite places.

Current Availability	A

Foursquare data is available through a public API with a well-supported GitHub repository.

Future Availability	A

We have no reason to suspect that Foursquare wouldn't continue to make its data available.

Level of Detail	A

Foursquare provides individual level data on businesses and attractions within London, including the address along with information about the type of location, business name,

reviews, ratings etc. It also provides individual level data on users' frequently visited places.

Utility	**A**

Foursquare is most likely to be useful in looking at the relationships aspect of social integration, including distinguishing between places that bring together strangers versus those that bring together friends, and places that attract diverse individuals or regulars.

Geographic Depth	**A**

As a mapping tool itself, it has the ability to drill down to individual addresses, which can then be reaggregated to perform analysis at any level across the GLA - a ward area is likely to be the right balance between geographic depth and ability to draw meaningful conclusions about a given area.

Representativeness	**D**

As researchers who have previously made use of Foursquare note, social media platforms like Foursquare tend to attract upwardly mobile and socially affluent users, and are therefore not representative of London as a whole.

Ethics of Access	**A**

Foursquare provides a public API and encourages its use.

Ethics of Publishing	**B**

Although Foursquare is a public site and its users are broadly anonymous, as with most social platforms we would recommend not re-publishing attributable Foursquare content.

Public Perceptions of Privacy	**B**

Foursquare is a culturally public space. However, we would recommend not re-publishing attributable content.

Skills Required	**B**

The primary technical skills needed to interact with this API are moderate. Familiarity with a programming software would be useful.

Frequency of Updates	**A**

Foursquare data is near live.

Costs of Ongoing Access	**Free**

Previous Successful Applications

By combining Foursquare data with a dataset of 37,000 Twitter users, researchers at the University of Cambridge[8] built a network of places listed in Foursquare and mapped Twitter social networks through people checking-in to locations, allowing them to distinguish between places that bring together strangers/ friends and map broader processes of gentrification. .

See:

Foursquare GitHub https://github.com/foursquare

Examples of Possible Use Cases

[8] D. Histrova, M.J. Williams, M. Musolesi, P. Panzarasa and C. Mascolo. *Measuring Urban Social Diversity using Interconnected Geo-Social Networks.* Proceedings of the 25th Conference on World Wide Web, 21-30, April (2016). Accessed at: https://www.repository.cam.ac.uk/handle/1810/253603

Measure: Social mixing

In conjunction with Twitter data, Foursquare could be used to measure the extent to which places in London bring together people from diverse backgrounds without prior connections to one another.

Measure: Participation in leisure activities

The GLA could use Foursquare to add colour to an analysis of leisure activities across the capital and the locations with high proportions of people checking-in to leisure activities.

Measure: Occupational segregation

In conjunction with census data on house prices, Foursquare could be used to measure broader processes of urban gentrification by correlating particular business types to rises in overall rent in order to infer the changing economic makeup of London wards.

Measure: Financial resilience

In conjunction with census data on household incomes, Foursquare could be used to map areas with proportionally higher or lower levels of disposable household income.

Google Maps

Google Maps is the world's most-used and most detailed mapping application by a significant margin. It provides street-level imaging, street maps and details of attractions, retailers, offices and other non-residential locations.

Data	**B**

Requests to the API output an Object with a list of properties in various types, e.g. Arrays, Strings etc. that would need to be extracted and formatted into a JSON or Spreadsheet

Current Availability	**B**

Access to Google Maps data is available through a simple API and is supported with extensive documentation. However, Google explicitly prohibits scraping of data, e.g. copying and storing business names and/or reviews - you can only cache an ID, along with longitude and latitude, so data can be geo-filtered and categorised offline but otherwise will need continual access requests to the API to perform analysis.

Future Availability	**A**

There is no reason to think that Google will close this API or materially reduce the available content from it. However, the API has become steadily more expensive over time and may continue to increase in price.

Level of Detail	**A**

Google Maps provides individual level data on all businesses and attractions within London. This includes the address along with information about the type of location, business name, reviews, ratings etc.

Utility	**A**

Google Maps is most likely to be useful when looking at the relationships aspect of social integration, particularly around social mixing. It could also be used to measure participation by tracking the number of a particular type of business, e.g. sports club, gym, etc. to determine association membership etc.

Geographic Depth	**A**

As a mapping tool itself, it has the ability to drill down to individual addresses, which can then be reaggregated to perform analysis at any level across the GLA - a ward area is likely to be the right balance between geographic depth and ability to draw meaningful conclusions about a given area.

Representativeness	**A**

Google Maps is the most popular mapping tool available. It has near universal coverage of all legitimate physical businesses and attractions, and has equally deep information across all of the GLA

Ethics of Access	**A**

All Google Maps data is offered through a public API and very little of it is associated with a particular individual, so there are relatively low ethical concerns around accessing this data.

Ethics of Publishing	**B**

All information about places is already publicly available and uploaded either by the business owners themselves or visitors with the acknowledgement that what they upload will be public; this content is not directly associated with a particular uploader so it should be fine to republish publicly.

Public Perceptions of Privacy	**A**

Most data is not personal and the public routinely access the same data via the Google Maps application, so there is unlikely to be significant perceptions of privacy except in photos and Street View.

Skills Required	**B**

The primary technical skills needed to interact with this API are moderate - analysts will need an ability to interpret technical specifications and follow documentation but processing the returned data will be more challenging than for other data sources.

Frequency of Updates	**B**

Satellite and Street View imagery is updated every few years. Roads and streets are more frequent, within months of a change, especially in populous urban areas like London. Frequency of change of details about particular places is sporadic and uncertain since a large amount of content is user-generated, though again in urban areas that have high population density changes are likely to be within weeks after real-life changes.

Costs of Ongoing Access	Variable depending on data requested but approximately $20-25 for full information on 8000 individual places (0.0025$/place).

See:

Google Maps Prices https://cloud.google.com/maps-platform/pricing/sheet/#places

Google Map API https://developers.google.com/maps/documentation/

Examples of Possible Use Cases

Measure: Relationships

Google Maps could be used to measure the degree of social mixing in a given area by the looking at the categories of restaurants/food shops in the area. For example, the number of different types of cuisine represented could act as a proxy for ethnic mixing, whereas

looking at the amount of places in different price ranges could indicate class mixing.

Measure: Participation

Looking at the number of sport clubs, community theatres, etc. in a given area could be a good indicator of participation in leisure activities.

Google Street View

Google Street View is a technology featured in Google Maps and Google Earth that provides panoramic views from positions along many streets in the world.

Data	**C**

Google's Street View service photographs and collates imagery from neighbourhoods across the world. The platform enables panoramic views of London streets and captures imagery suggestive of broader demographic traits – including the prevalence of neighbourhood green space and the popularity of particular car brands within neighbourhoods.

Current Availability	**A**

Google Street View provides a public API.

Future Availability	**A**

We have no reason to believe this data would not be available in the future.

Level of Detail	**C**

Google Street View's photographic imagery can be used to infer demographic characteristics about neighbourhoods, but does not contain specific information about the places that appear on the platform.

Utility	**B**

The GLA could make use of Google Street View in conjunction with other data sources to infer demographic characteristics about particular neighbourhoods.

Geographic Depth	**A**

Google Street View provides granular geographic detail at the street and door number level.

Representativeness	**C**

Google Street View is visually representative of the places it depicts. While this computerised imagery can be used to infer broader demographic characteristics about particular neighbourhoods, this can't be taken as an exact representation.

Ethics of Access	**A**

Google Street View data is accessible through a public API.

Ethics of Publishing	**A**

Google Street View's public API allows researchers and users to re-publish non-interactive Street View panoramas.

Public Perceptions of Privacy	**A**

Google Street View data is culturally public and available to everyone.

Skills Required	**C**

Google Street View data can be accessed through a public API. However, to analyse this data at scale, the GLA would need to make use of an algorithm to comb google Street View data for particular characteristics - such as the car brands that appear within neighbourhoods.

Frequency of Updates	**C**

Google Street View is updated every 2 to 3 years.

Costs of Ongoing Access	**Free**

Previous Successful Applications

Researchers at Stanford University built an algorithm that would comb Google Street View data for the car brands that appeared within particular neighbourhoods. By comparing Google Street View data on car type with data from the American Community Survey, researchers were able to predict a host of demographic identifiers, including household income, race, education and unemployment. Researchers at Harvard University also made use of computerised deep learning to estimate the demographic makeup of neighbourhoods across the United States.

See:

Google Street View	https://mapstreetview.com/
Dev manual	https://developers.google.com/maps/documentation/streetview/intro

Examples of Possible Use Cases

Measure: Occupational segregation, educational attainment gap and NEET

Taking inspiration from research at the University of Stanford, the GLA could comb Google Street View data on car brand, comparing this with another large dataset on household income and household education, to map the areas of occupational and educational inequality within London.

Measure: Neighbourhood cohesion

The GLA could make use of deep learning techniques to train a computer to spot patterns in Google Street View imagery to guess household income levels and measure inner-city economic inequality. However, it's worth remembering that this technique has only been tried so far in Boston and New York - and it remains to be seen whether it would work as well in other cities with different visual and cultural signifiers of demography.

Meetup

Meetup is a service used to organise online groups that host in-person events for people with similar interests.

Data	**A**

Data is returned in a JSON format

Current Availability	**A**

Access to Meetup data is available through a public API with extensive documentation

Future Availability	**A**

Meetup is currently owned by WeWork and they have made no indication of plans to shut down the service, so we would expect the data to be accessible on demand for the foreseeable future

Level of Detail	**A**

Able to access the name, time/date, type, description, venue location, attendance numbers and a sample of comments and attendees for a given event. Also able to access groups by city, description, purpose and membership numbers.

Utility	**A**

Likely to be useful for both relationships and participation. Association membership has traditionally been used as a measure of social capital which fits into the GLA wider model of social integration. Further, it can tell the GLA about the type of activity individuals are engaging in over an area - for example, a spike in language classes, particularly English language, might be a sign of increased attempts at integration in a local community.

Geographic Depth	**B**

Not all events provide a venue (either not specified or not publicly visible), but a large proportion do, and those provide the full address of the venue, down to the street level.

Representativeness	**C**

Only a fraction of events on the platform take place in London, but most events can be specified by location, so it should be possible to isolate a sample from London. The types of events are broad but tend to skew towards the young and the tech-minded, e.g. a large number of coding/data science meet-ups.

Ethics of Access	**A**

Available by a public API with no restrictions in Meetup's terms of service, besides those already covered under GDPR.

Ethics of Publishing	**B**

The API returns non-anonymised data associated with a particular person, including their name and potential photo. Therefore, we would not recommend publishing individual comments or posts. However, aggregate and anonymised data, e.g. the number of meetups of a certain type and the number who attended each meetup of a certain type seems ethical.

Public Perceptions of Privacy	**A**

The organiser of the Meetup group hosting can choose to restrict the visibility of their group to only members, in which case it cannot be accessed by the API. Therefore, all information accessible through the API is publicly available and the platform is explicitly for attending events with others so users have strong reasons to believe that the information they post will be viewed by strangers.

Skills Required	**A**

The primary technical skills needed to interact with this API are moderate - analysts will need an ability to interpret technical specifications and follow documentation, and to store and process returned data. Familiarity with a major programming language (Python, Java, C++, PHP etc) will also be useful.

Frequency of Updates	**A**

Live but access to the API is rate-limited.

Costs of Ongoing Access	**Free**

See:

Meetup API https://www.meetup.com/meetup_api/

Examples of Possible Use Cases

Measure: Participation

The GLA may be able to look at the number of Meetups by category to determine civic participation or leisure activity in a given area.

Measure: Relationships

Measure the diversity of social interactions by looking at the number and variety of groups a given individual is a member of. Further, looking at whether users cluster into groups of groups or whether there are cross-cutting connections and whether users of one group tend to be members of a diverse number of other groups

Mobile Company Data

Mobile Networks carry significant amounts of metadata that once aggregated may be used as measures of social integration geographically or by economic class.

Data	**A**

Available data includes information about network users, including their location via GPS, their mobile phone number, address and email, gender, date of birth, and the performance and use of their contract, including the numbers they dial, their handset type, the amount of data they use, and their timeliness in bill payment. Mobile phone companies also automatically generate Call Detail Records (CDRs).

Current Availability	**B**

No mobile phone company currently makes data publicly available, but some have differing levels of analytics which can be shared with third-party partners on a cost basis.

Future Availability	**B**

As companies seek to monetise their data by understanding its policy relevance it is likely that availability will increase in future.

Level of Detail	**A**

The level of detail varies by company, according to their commercial interests and ethical policies. Some provide anonymised aggregated location information, e.g. how many phones are in a given location and how many phones move between locations. Some also provide anonymised demographic data, such as age range, gender and handset type. Some companies mention that they are willing to provide other data they collect to third-party providers, this could be browsing behaviour or traffic data like numbers called and duration of calls, however, they only do this with the explicit consent of users.

Utility	**A**

If data is available, then CDRs, information on handset type, timeliness of bill payments and data usage could be helpful for measuring social mixing, occupational segregation and financial resilience.

Geographic Depth	**A**

Mobile phone operators track users by GPS phone masts and Wi-Fi networks, allowing granular geographic insight at an individual level.

Representativeness	**B**

Mobile phones are ubiquitous across all demographic groups and there is no reason to think that most mobile phone providers customer bases do not reflect the London area at large. The GLA could also make use of specialist phone operators used for international calls, such as Lycamobile, if it wished to focus in greater depth on migrant communities.

Ethics of Access	**A**

There are unlikely to be ethical concerns around accessing data via a partnership with a mobile phone provider as the company itself can directly provide the data in a format that does not require explicit user consent to be processed and allows users to opt out if they choose to.

Ethics of Publishing	**B**

There are unlikely to be ethical concerns around publishing anonymised and aggregated data provided by mobile phone companies. However, if the GLA requests any individualised data, they should not publish it.

Public Perceptions of Privacy	**B**

User data is private and not considered public information, with phone usage and calls likely to be a particularly sensitive area to some people. Aggregated data where attributable characteristics have been removed can be shared freely.

Skills Required	**NA**

As it is unclear what the particular format and method of access for the data will be, e.g. an API or dashboard, it is difficult to say what technical skills will be required to use this data. However, as this is a bespoke partnership and mobile companies generally have a large technical capacity and long history of working with third-parties, there is likely to be technical support available, with less of the burden being placed on the GLA itself.

Frequency of Updates	**A**

Mobile phone data is likely near live.

Costs of Ongoing Access	**Uncertain**

The costs of any data access are likely to be negotiated on a case by case basis depending on the specific request.

Previous Successful Applications

Researchers have previously used CDRs to track emergency migration, map commuter flows, and measure the geographic distribution of users' social connections. Frequently making and receiving calls with contacts outside of one's immediate community is correlated with higher socio-economic class.[9]

Examples of Possible Use Cases

Measure: Use of Digital Networks

The GLA could use data about users' contract performance to measure the amount of data that mobile phone users within particular areas consume, in order to estimate an

[9] United Nations Global Pulse (October 2013) Mobile Phone Network Data for Development. Accessed at: http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf

area's use of digital networks.

Measure: Financial resilience

The GLA could use data on the timeliness of bill payments to infer phone users' financial resilience, with late payments an indicator of poor financial resilience.

Measure: Social mixing

By measuring the number of calls made to people outside of a phone user's community, the GLA could infer high or low socio-economic class within a particular area, with frequent calls made to contacts outside of one's immediate community correlated with higher socio-economic class.

Trustpilot

Trustpilot is a business review site.

Data	**A**

All data is returned in a JSON format, making it easy to work with.

Current Availability	**A**

Access to Trustpilot data is available through a public API with extensive documentation

Future Availability	**A**

Trustpilot uses the same APIs for operating their website and business product and so they are committed to their continued availability, improvement and evolution.

Level of Detail	**A**

You can access the name, address, category, ratings and reviews of a given business. You can also access the name, location, gender and associated reviews for a given user

Utility	**B**

Trustpilot is likely to be of some value in considering equality and outcomes, especially around discrimination. There are potential uses for participation but those are likely to be covered better by more detailed mapping sources.

Geographic Depth	**B**

The business profiles for physical business generally have a specific address associated with them, including house number and postcode. The API allows you to view user's cities, so could be used to pin down reviewers who are native to the London area, however some user profiles only provide their country, e.g. United Kingdom. This could still be useful in identifying some of the reviews of tourists.

Representativeness	**C**

Trustpilot is broadly representative of the breadth of types of businesses in London. However, reviews are likely to only reflect those with a particularly strong feeling about a particular business, either positively or negatively. Therefore, the reviews may not reflect the full range of views about a given business.

Ethics of Access	**B**

The data is available by a public API with the full permission of Trustpilot

Ethics of Publishing	**B**

Company profiles and user reviews are all publicly available, however Trustpilot discourages the sharing of personal information from particular users. Information about

companies with reviews has the potential to be inaccurate or misleading and so likely should not be publicly shared disaggregated.

Public Perceptions of Privacy	**B**

Businesses are unlikely to believe that their profiles are private and individual users post the reviews publicly for strangers to see so it is unlikely there will be privacy concerns.

Skills Required	**A**

The primary technical skills needed to interact with this API are moderate - analysts will need an ability to interpret technical specifications and follow documentation, and to store and process returned data. Familiarity with a major programming language (Python, Java, C++, PHP etc.) will also be useful.

Frequency of Updates	**A**

Data is updated near-live.

Costs of Ongoing Access	**Free**

See:

Trustpilot API https://developers.trustpilot.com/

Examples of Possible Use Cases

Measure: Equality

The GLA could use a keyword search on reviews for businesses to determine the number of reported incidents of discrimination against customers by businesses within an area.

Twitter

Twitter is a US based microblogging site and public discussion platform. Users post short tweets and images, retweet others' content, and engage in conversation with people within and outside of their social networks.

Data	**A**

Available data is largely text posts under 140 characters, searchable through specific hashtags, phrases, key words and locations. Twitter is ideal for observing real-time human behaviours and interactions that are unprompted by researchers. Users can communicate with one another publicly through the @reply command and contribute to broader, active conversations using hashtags.

Current Availability	**A**

Twitter data is available through a public API

Future Availability	**A**

We have no reason to believe this data would not be available in the future.

Level of Detail	**B**

It is possible to infer location and demographic information such as gender, age and ethnicity from Twitter profiles. The sheer volume of real-time tweets allows researchers to gauge conversations as they unfold, and collect time-series data. Location tagging

enables researchers to map location-specific conversations and trends, while the majority of active users have a profile photo that can be coded for basic demographic information (e.g. ethnicity). Through coding for specific words or phrases, or using language-processing software, researchers could also conduct sentiment analyses of Twitter data.

Utility	**C**

Twitter offers an insight into peoples' self-reported attitudes towards topics, and is useful for qualitatively interpreting peoples' reflections on and feelings about particular subjects. Aggregated Twitter data can also be used to map location-specific trending topics within particular parts of the city through geo-tagged tweets.

Geographic Depth	**B**

London specific geographic location can be inferred either directly, where Twitter users have turned on location sharing, or indirectly, by inferring a users' location from the location of their network. Where users have turned on location sharing or tagged tweets to particular locations, researchers can map in depth the geographic location of Tweets.

Representativeness	**C**

In the UK, research[10] shows that men are proportionately more likely than women to use Twitter; the age distribution of Twitter users is younger than that of the UK population; and managerial, administrative and professional occupational groups are more likely to use Twitter.

Ethics of Access	**A**

Twitter's open API platform provides access to public Twitter data. It asks users to respect its Developer Agreement Policy (found here https://developer.twitter.com/en/developer-terms/agreement-and-policy).

Ethics of Publishing	**B**

Although Twitter is a public site, its users are not anonymous. As with most social platforms we would recommend not re-publishing attributable content or deriving inferred sensitive characteristics about Twitter users. Aggregated data can be shared freely as long as it does not store any personal data and complies with Twitter's Developer Agreement Policy.

Public Perceptions of Privacy	**B**

Twitter is a public platform. Tweets that will be relevant for the GLA are those which are already publicly available and searchable through key words and hashtags.

Skills Required	**C**

Tweets are limited to 140 characters, making Twitter data easier to analyse at scale than long-form text. Researchers could make use of a time-based job scheduler (e.g. Cronjobs) to run scripts at designated intervals for time-series data. To analyse sentiments and group together trending topics, researchers could employ a tweet sentiment visualisation app (e.g. Tweet Visualizer).

Frequency of Updates	**A**

Data available through the Twitter API is near-live, and a query will return posts, comments and other metadata from the time the query is made.

Costs of Ongoing Access	**Free**

[10] Luke Sloan, *Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey.* Social Media + Society, March (2017). Accessed at: https://journals.sagepub.com/doi/10.1177/2056305117698981

Previous Successful Applications

Researchers from the University of Washington used Twitter to analyse the political attitudes of a hard-to-reach group of people who self-reported that they decided they were not going to vote in the 2012 Presidential Election.

Northwestern University's Journalism School created a map of different neighbourhoods in Columbus, Ohio, which aggregates topics pertaining to neighbourhood life and shows how many people are tweeting about this subject in a heatmap format.[11]

See:

Developer Agreement: https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

Note:

Data returned from Twitter does not contain information about whether an account belongs to an individual or an organisation (e.g. a charity or business). However, there are ways in which this information could be inferred, e.g. by compiling a list of known and vocal institutional accounts and removing these from the dataset. Demos' Centre for the Analysis of Social Media has used its own algorithms to ascertain individual / organisation status from user metadata, for example an account's description field, and it is possible that paid-for third party tools may offer similar functionality.

Examples of Possible Use Cases

Measure: Political Participation

Twitter could be used to quantitatively estimate the proportion of people who say they are going to vote or not vote. Individual tweets could also be coded to interpret peoples' perceptions, attitudes and feelings about forthcoming elections and political participation.

Measure: Civic Participation

The GLA could use Twitter to evaluate peoples' perceptions, attitudes to and involvement in aspects of civic life by analysing tweets using hashtags and key words. The GLA could also map Londoners' connection with their elected councillors through Twitter interactions between councillors and Twitter users

Measure: Neighbourhood Cohesion

By analysing geo-located tweets in particular areas of London for key words, the GLA could use Twitter to gauge how socially cohesive people feel their local area is.

Measure: Participation in Leisure Activities

The GLA could use Twitter to evaluate peoples' responses to and perceptions of leisure activities and events in the capital by type of activity.

Measure: Loneliness

By analysing tweets using key words, phrases and hashtags relating to social isolation and loneliness, the GLA could investigate peoples' self-reported feelings of and attitudes towards loneliness in London

Measure: Feeling of Belonging

[11] https://neighborhoodbuzz.knightlab.com/tweets/columbus/near-east/

Through an analysis of tweets using particular phrases, hashtags and key words, tweet clusters around particular topics, and geo-located tweets from specific areas in the capital, the GLA could use Twitter to monitor how Londoners are engaging with and reflecting on issues concerning social integration.

Measure: Social mixing

The GLA could infer the extent to which Londoners report frequent contact with people from different demographic backgrounds by mapping their Twitter connection according to inferred demographic characteristics (ethnicity, gender).

Yelp

Yelp is a local search platform powered by a crowd-sourced review forum. Users leave reviews of local businesses, including restaurants and valet services, and data is aggregated to rate businesses out of five stars.

Data	**A**
Yelp stores data on businesses across the world, including their exact geographic location, opening times, photos and reviews posted by users, prices and similar suggestions.	
Current Availability	**A**
Yelp data is accessible through a public API.	
Future Availability	**A**
There is no reason to expect that Yelp data would not continue to be made publicly accessible.	
Level of Detail	**C**
Yelp data that will most likely be useful for the GLA is aggregated data on the number of bars/ restaurants and local businesses within particular areas in London - this provides an overall view of the local economy within specific boroughs, without much granular detail.	
Utility	**C**
Yelp data is most likely to be useful for the GLA as a source of statistical data on the number of local businesses within an area, such as bars and restaurants. This would need to be combined with statistics on house prices or a similar metric to provide a reliable indication of social gentrification.	
Geographic Depth	**A**
Yelp provides granular geographic detail, listing the street address of businesses on the platform.	
Representativeness	**B**
Almost 20,000 London restaurants are listed on Yelp. The site attracts more young users than old; 48% of Yelp visitors are aged 25-44.	
Ethics of Access	**A**
Yelp data is publicly accessible and supported by a public API.	
Ethics of Publishing	**A**

Yelp business listings are publicly accessible. We would not recommend re-publishing Yelp reviews with attributable content.

Public Perceptions of Privacy	A

Yelp is a culturally public space and Yelp listings are intended to reach a broad public audience.

Skills Required	B

The majority of useful Yelp information is categorised by location, making it easy to analyse.

Costs of Ongoing Access	Free

Previous Successful Applications

Researchers at Harvard University have used publicly available Yelp data to nowcast local economic activity. Led by urban economist Edward Glaeser, they also correlated Yelp data to housing price statistics to infer levels of gentrification in New York City and quantify neighbourhood change.

See:

Yelp London https://www.yelp.co.uk/london

Yelp public API
https://www.yelp.com/developers/documentation/v3/business_search

Examples of Possible Use Cases

Measure: Feeling of belonging / neighbourhood cohesion

Taking inspiration from previous research using Yelp data, the GLA could aggregate Yelp data on the existence of restaurants, cafes, bars and local grocery shops and correlate this with available data on house prices to quantify the extent of gentrification within an area. This could be used to infer the negative of "feeling of belonging" and "neighbourhood cohesion".

Measure: Social mixing (weak ties)

The GLA could infer levels of social mixing within particular London boroughs by analysing the proportion of restaurants serving food with diverse ethnic heritage.

Measure: Occupational segregation

The GLA could infer household income and therefore occupational segregation by mapping the proportion of cheap / affordable food options within particular neighbourhoods.

YouTube

YouTube is the internet's most widely-used video sharing site.

Data	A

YouTube offers a large amount of metadata about its content through an API. All data is returned in a JSON format.

Current Availability	**A**

Access to YouTube data is available through a simple API and is supported with extensive documentation.

Future Availability	**A**

There is no reason to think that YouTube will close this API or materially reduce the available content from it, though there is precedent for Google to suddenly start charging for its data services.

Level of Detail	**B**

The API provides access to auto-generated captions, comments and metadata from videos.

Utility	**B**

It could be used to measure equality through analysing mentions of discrimination in captions of videos. It could also be used to measure participation in civic or charitable causes, e.g. measuring the number of Ice Bucket Challenge videos and how they spread through GLA YouTube content creators.

Geographic Depth	**B**

Able to access the country of upload for videos and channels. It was previously possible to access the longitude and latitude of videos but this has been deprecated June 1, 2017 and so is unavailable on many videos including those going forward.

Representativeness	**A**

YouTube is the world's most popular video sharing site and is used by a large proportion of the GLA population. However, the data that is available is from uploaded videos and comments, which are more likely to be representative of younger and more affluent demographics (though this depends on the genre of video). Videos and channels with large follower bases are likely to be representative of influencers within communities in London however.

Ethics of Access	**A**

Accessed by a public API.

Ethics of Publishing	**B**

YouTube videos and channels are publicly available already and generally created with intention of being seen by others. However, caution should be taken if sharing content from small channels or videos with low view counts as exposure could lead to unintended consequences as a result of the increased publicity.

Public Perceptions of Privacy	**A**

YouTube channels are publicly facing and the API does not allow you to access unlisted videos unless those videos are part of a public playlist, so the public is unlikely to view any of the accessible content as private

Skills Required	**A**

The primary technical skills needed to interact with this API are moderate - analysts will need an ability to interpret technical specifications and follow documentation, and to store and process returned data. Familiarity with a major programming language (Python, Java, C++, PHP etc) will also be useful.

Frequency of Updates	**A**

Live - however there is a daily quota on calls to the API on 1 million units, which is equivalent to accessing the metadata of 1 million videos.

Costs of Ongoing Access	Free

See:

YouTube Developer Notes https://developers.google.com/youtube/v3/docs/

Examples of Possible Use Cases

Measure: Relationships

Using subscription data to map connections between channels to see whether they group into clusters or have diverse connections/interests.

Measure: Outcomes

Using keywords to analyse the content of captions to measure instances of reported discrimination. Further, this could provide information for the type of discrimination, e.g. physical, verbal and if was on class, race, etc.

Zoopla

Zoopla is a UK-based property website.

Data	A

Zoopla data includes house price data and information about a house and listing. All the data is output as an XML by default but a JSON version can be requested instead.

Current Availability	A

Access to Zoopla data is available through a public API with extensive documentation

Future Availability	A

There is no reason to believe Zoopla will stop operating or prevent access to their data in the foreseeable future. Every new key is initially issued under a three month free license and Zoopla indicates that they may review their terms of use or follow up cases where additional terms need to be agreed.

Level of Detail	A

Details about the price, size, property type and usage (e.g. semi-detached for sale versus flat to rent) on an individual property level. It has average house prices for a given area and also offers historic listings on a request-basis.

Utility	A

Zoopla provides live and historic data on house prices. This be used to understand the changing economic circumstances of an area over time and the stability of the community (by the number of properties changing hands in a given year).

Geographic Depth	A

As a property website, all listings are at an individual address level

Representativeness	A

Zoopla covers the whole GLA and as one of the largest property sites, it is likely to host property adverts that reflect the breadth and depth of property sales across the GLA in terms of price, location, usage, demographics of the buyers and sellers etc.

Ethics of Access	**A**

Available by a public API with the permission of the platform.

Ethics of Publishing	**B**

All data accessible is publicly available on the Zoopla platform, however the individual details of properties for sales should probably not be published except when the properties are already of public interest.

Public Perceptions of Privacy	**A**

The public generally view details about their homes to be private. However, those who are advertising homes for sale do so with the explicit intention of those details being noticed by strangers, so there is likely to be less concern around privacy for homes up for sale than homes in general.

Skills Required	**B**

The primary technical skills needed to interact with this API are moderate - analysts will need an ability to interpret technical specifications and follow documentation, and to store and process returned data. Familiarity with a major programming language (Python, Java, C++, PHP etc) will also be useful.

Frequency of Updates	**A**

Live, however each API key is limited to 100 calls per hour.

Costs of Ongoing Access	**Free**

See:

Zoopla API https://developer.zoopla.co.uk/

Examples of Possible Use Cases

Measure: Equality

The GLA may be able to 'nowcast' housing affordability in a given area by live monitoring listings in a given area.

Measure: The Pace of Local Change

The GLA may be able to measure the percentage of 'affordable housing' in a given area and the range of house prices and types.

Airbnb

Airbnb is a peer-to-peer tourism platform where users can list and rent short-term tenancies across the world.

Data	**C**

Airbnb stores data on short-term rental properties around the world. This includes their geographic location, photographs, price per rental per night and rating according to Airbnb users. Airbnb also stores data on users - including properties they have booked and/or let and messages they have exchanged with other Airbnb users.

Current Availability	**B**

The data that would be relevant for the GLA is publicly available. While Airbnb does not have a public API, data can be scraped using a web scraper.

Future Availability	**B**

There is no reason to believe that currently available data on property listings would not remain available in the future for as long as Airbnb is a viable platform.

Level of Detail	**C**

Airbnb provides ward-level data on the existence of short-term rentals within the city.

Utility	**C**

The GLA could use Airbnb data to infer gentrification processes within London by mapping the spread of Airbnb property expansion and the characteristics of those properties.

Geographic Depth	**A**

Airbnb properties are categorised by neighbourhood location, offering granular information about the prevalence of Airbnb apartments within particular neighbourhoods.

Representativeness	**C**

Airbnb data on short-term rentals could be used as a representative indicator of gentrification within London neighbourhoods. However, it is still uncertain whether Airbnb has acutely gentrifying effects in London, where multiple factors compound a lack of affordable housing, compared to smaller cities that are more dependent on the tourism economy.

Ethics of Access	**B**

Airbnb data can be accessed through a web scraper, but the platform does not have a public API.

Ethics of Publishing	**B**

Airbnb data includes photographs of individual homes and indicators of their geographic location. Therefore, we would advise not re-publishing non-aggregated Airbnb data that could be attributed to Airbnb users.

Public Perceptions of Privacy	**B**

Airbnb is a publicly accessible platform, but its listings include details about peoples' private domestic spaces. Therefore, the GLA should not re-publish non-aggregated Airbnb data.

Skills Required	**B**

Airbnb data on property listings can be scraped fairly easily using a web scraper.

Frequency of Updates	**A**

Airbnb data on available property listings is near live.

Costs of Ongoing Access	Free

Previous Successful Applications

There is an emerging field of research on whether the prevalence of Airbnb properties in a given city is an indicator of broader gentrification. In smaller cities that are dependent on the tourism economy, including Amsterdam and Barcelona, Airbnb is frequently cited as a problem for social cohesion and availability of affordable housing. However, in London, it's unclear whether Airbnb can be isolated from the broader pressures created by the buy-to-let landlord sector on the availability of affordable housing.

See:

Airbnb scraper	https://github.com/adodd202/Airbnb_Scraping
Airbnb	Airbnb.com

Examples of Possible Use Cases

Measure: Housing Affordability

In conjunction with statistical data on house prices across London, Airbnb could be used to map how property usage (i.e. holiday lets) drive up house prices in the capital, creating a lack of affordable housing.

Measure: Neighbourhood Cohesion

The GLA could use Airbnb data on property listings to map areas of high social transience; neighbourhoods with a concentration of Airbnb listings could potentially predict high transience and therefore lower social cohesion.

Citymapper

Citymapper is a popular routing app in the London area.

Data	C

A call for data returns a link to a Citymapper journey.

Current Availability	A

Access to Citymapper is available through a simple API.

Future Availability	A

There is no reason to believe Citymapper data will become more restricted in future, and in fact, it is likely they will expand the capabilities of their API.

Level of Detail	D

Very limited, only shows distance and time between two places.

Utility	D

Very limited.

Geographic Depth	A

Information is available on a point-to-point basis and so is available at an address level through-out London

Representativeness	A

Citymapper was originally developed specifically in the context of the GLA and covers the entire area with equal depth.

Ethics of Access	A

Access is through a public API and no personal data can be accessed, so there are likely to be no ethical issues with accessing this data.

Ethics of Publishing	A

No personal data is available, only estimates based on aggregates of journeys and Citymapper's own estimations, so there is no concern around publishing any data accessed.

Public Perceptions of Privacy	A

No personal data available and the public themselves routinely access this data through the app itself, so unlikely to be any expectations of privacy.

Skills Required	A

To access the API, analysts will need an ability to interpret technical specifications and follow documentation, and to store and process returned data.

Frequency of Updates	Unknown

Costs of Ongoing Access	Free for a limited quota of access requests; larger quota costed on a case-by-case basis so unclear what the GLA would pay.

See:

Citymapper API https://citymapper.com/tools/1063/api-for-robots

Find A Job

'Find a job' replaced Universal Jobsmatch as the UK Department for Work and Pensions' jobs listing site, available online and at Jobcentre Plus. It aggregates advertised jobs in the UK and categorises them according to salary, location, contract and job type.

Data	C

Find a job includes data on job locations within London, salaries, job and contract categories. Its predecessor service, Universal Jobsmatch, stored personal information about users, including CV, address, age, name, email and skills.

Current Availability	C

There is no official API for the Find a job service, and as such any data would have to be collected through a browser scrape or by contacting the DWP, who run the service, directly. While their Acceptable Use Policy for the site (at https://findajob.dwp.gov.uk) prohibits data from being used commercially, it does not explicitly disallow automated collection.

Future Availability	B

Since Find a job is run by a government department, it is highly likely that any currently accessible data will remain so in future.

Level of Detail	D

Find a job is most likely to produce aggregated statistical information about salary levels and job characteristics in London, and therefore would be useful in measuring occupational segregation and inequality within London's formal economy.

Utility	C

It would be possible to use Find a job data to infer levels of occupational segregation and inequality across the capital. However, because Find a job only includes jobs within the formal economy that employers have chosen to advertise, the data would not be as representative as a survey of individuals' occupations.

Geographic Depth	B

Find a job allows users to search for jobs by area, offering geographic granularity at the borough level.

Representativeness	B

It is not possible to extract demographic information about Find a job users. While the platform advertises a variety of jobs, it cannot be considered representative of all jobs and salary levels in London; many jobs and occupations will not be included on the platform.

Ethics of Access	D

Find a job data on job listings is intended to be publicly accessible.

Ethics of Publishing	A

Find a job data on job listings is considered publicly accessible knowledge and does not contain any attributable personal information.

Public Perceptions of Privacy	A

Find a job's job listings are considered publicly accessible knowledge and does not contain any attributable personal information.

Skills Required	C

As there is no official API, developers would need to build their own methods for collecting data from the Find a job service, or ask the DWP for access

Costs of Ongoing Access	Free

See:

Find a job https://findajob.dwp.gov.uk/

Examples of Possible Use Cases

Measure: Occupational Segregation

Find a job data on jobs listings could provide the GLA with an estimate of occupational segregation across the capital by measuring percentages of high/low paid occupations.

Freecycle

Freecycle is an online gift system where users advertise goods that they are giving away, or looking for, without expectation of anything in return. Founded in Arizona, Freecycle, at the time of publication, has 42 active groups across London.

Data	**C**

Available data is largely short-form free text posts about goods and services procured or offered free of charge

Current Availability	**A**

Freecycle does not have a public API but can be accessed by using a scraper that creates an API.

Future Availability	**C**

There is no reason to suppose that Freecycle would not continue to be accessible in the future. However, it is unknown whether Freecycle will continue to be a popular platform with Londoners due to competition from similar sites.

Level of Detail	**C**

Freecycle is most likely to produce short-form information about the types of goods and services that people seek across London boroughs, and London residents' comparative level of participation in a trust-based gift economy platform. Users are anonymous, although their location can be inferred through participation in local Freecycle groups.

Utility	**C**

Freecycle would be useful as a source of information about the likelihood of Londoners across different boroughs to participate in a trust-based gift economy platform, and could therefore be used to infer or add colour to an analysis of social mixing, neighbourhood cohesion and helping neighbours / social trust.

Geographic Depth	**B**

Users remain anonymous on Freecycle, and are not required to register or provide personal information in order to use the site. However, their geographic location can be inferred through their use of place-specific Freecycle groups in London, of which there are 42.

Representativeness	**D**

Freecycle does not store personal demographic information about its users, and therefore it is impossible to gauge how representative the platform is of Londoners as a whole.

Ethics of Access	**A**

Freecycle is a public platform that is open for all to use.

Ethics of Publishing	**B**

Freecycle users are anonymous and the platform does not require personal data from users. Aggregated Freecycle content can be shared freely.

Public Perceptions of Privacy	**A**

Freecycle is primarily a public platform that does not store personal information. Information that would be most useful to the GLA, e.g. the relative activity of Freecycle groups across London boroughs, is considered public.

Skills Required	**B**

The information that would be relevant to the GLA would likely be a comparative analysis of the aggregate number of users and rate of interaction across Freecycle groups within the capital, which could be gained by using a scraping tool.

Frequency of Updates	**?**

The frequency of Freecycle updates is not known.

Costs of Ongoing Access	**Free**

Previous Successful Applications

Researchers have previously used Freecycle as a case study in social reciprocity. A study from the University of California, Berkeley, revealed that Freecycle had a "viral effect" on community spirit and generosity[12]. As such, we think involvement in Freecycle could be used as an indicator of social trust, helping neighbours and neighbourhood cohesion.

See:

Freecycle website	https://www.freecycle.org/
Freecycle scraper GitHub	https://github.com/mikegreen1995/freecycleScraper

Examples of Possible Use Cases

Measure: Social mixing

The GLA could evaluate comparative levels of participation in Freecycle across different areas in London to add insight to levels of social mixing and neighbourhood cohesion.

Measure: Helping neighbours/ Social trust

As a trust-based gift economy platform where users place adverts for borrowing equipment and giving away items for free, Freecycle data on levels of participation across London could be used to inform a GLA study of the extent to which Londoners in particular areas of the capital are likely to trust others and help neighbours.

Indeed

Indeed is a jobs listing site that aggregates job adverts across London and the UK, and categorises these by job and contract type, location and salary.

Data	**D**

Available data is job listings categorised by salary, job type (full time / permanent / part-time/ temporary and contract), London borough where the job is located, title and company. Indeed.com also stores personal data from user accounts, including CVs, street address, gender, occupation and mobile number. It does not share attributable personal data with third parties unless this data has been aggregated.

Current Availability	**A**

Access to Indeed data is through an API, which is supported with a GitHub repository.

[12] Rob Willer, Francis J. Flynn and Sonya Zak, *Structure, Identity and Solidarity: A Comparative Field Study of Direct and Generalized Exchange.* Administrative Science Quarterly, May 9 (2012). Accessed at: https://www.gsb.stanford.edu/faculty-research/publications/structure-identity-solidarity-comparative-field-study-direct

Future Availability	A

At this time there is no reason to think that Indeed would not continue to support API access to its platform.

Level of Detail	D

Indeed is most likely to produce aggregated statistical information about salary levels and job characteristics in London, and therefore would be useful in measuring occupational segregation and inequality within London's formal economy.

Utility	C

It would be possible to use Indeed data to infer levels of occupational segregation and inequality across the capital. However, because Indeed.com only includes jobs within the formal economy that employers have chosen to advertise, the data would not be as representative as a survey of individuals' occupations.

Geographic Depth	B

Indeed allows for some geographic granularity as it categorises jobs according to London borough.

Representativeness	D

It is not possible to extract demographic information about Indeed users. While the platform advertises a variety of jobs, it cannot be considered representative of all jobs and salary levels in London; many jobs and occupations will not be included on the platform.

Ethics of Access	A

Indeed data on job listings is publicly accessible through their API.

Ethics of Publishing	A

Indeed data on job listings is considered publicly accessible knowledge and does not contain any attributable personal information.

Public Perceptions of Privacy	A

Indeed job listings are publicly available and are intended to reach a broad public audience.

Skills Required	B

Ability to interface with a well-documented API.

Frequency of Updates	B

Indeed is updated daily with new jobs listings.

Costs of Ongoing Access	Free

See:

Indeed	Indeed.com
Indeed.com API GitHub documentation	https://github.com/indeedassessments/api-documentation

Examples of Possible Use Cases

Measure: Occupational segregation

Indeed data on jobs listings could provide the GLA with an estimate of occupational segregation across the capital by measuring percentages of high/low paid occupations.

JustGiving

JustGiving is a public crowdfunding platform. It is free to use for individuals, while charities are charged a 5% commission on each donation.

Data	**C**

Available data is free-text posts and comments about causes that JustGiving users are fundraising for, fundraising targets, and location of fundraising causes. There is little available data on people who are donating to causes as many users remain anonymous.

Current Availability	**A**

Access to JustGiving data is through a public API with a well-supported GitHub repository.

Future Availability	**A**

There is no reason to suspect that JustGiving data would not continue to be available in the future.

Level of Detail	**C**

JustGiving data is long-form nuanced descriptions of fundraising causes with little available data on the platform's donating users.

Utility	**C**

JustGiving data would most likely be useful for gauging the types of charitable causes that Londoners are fundraising for rather than a quantitative assessment of how many Londoners are involved in charitable causes.

Geographic Depth	**D**

JustGiving includes brief geographic information about charitable causes but does not list geographic information about donating users.

Representativeness	**C**

JustGiving has users in over 90% of UK postcodes, but does not publish any information about the number of users in London or their demographic details (e.g. age).

Ethics of Access	**A**

JustGiving provides an open API and encourages its use.

Ethics of Publishing	**B**

JustGiving is a public site, but we would recommend not re-publishing attributable content or personal information.

Public Perceptions of Privacy	**B**

JustGiving's platform is public.

Skills Required	**A**

An ability to interact with an API.

Frequency of Updates	**A**

Near live.

Costs of Ongoing Access	**Free**

Examples of Possible Use Cases

Measure: Civic Participation

The GLA could use JustGiving data to map the number of JustGiving causes by London neighbourhood and give an estimate of the types of causes that Londoners are involved with.

Monster Jobs

Monster is a jobs listing site that aggregates job adverts across London and the UK, and categorises these by job and contract type, location and salary.

Data	**D**

Available data is job listings categorised by salary, job type (full time / permanent / part-time/ temporary and contract), London borough where the job is located, title and company. Monster also stores personal data from user accounts, including CVs, street address, gender, occupation and mobile number. It also stores users' race and ethnicity, where provided, but does not share attributable personal data with third parties unless users have consented or the data has been aggregated to remove personal characteristics.

Current Availability	**A**

Job listings data is available through Monster's public Job Search API

Future Availability	**A**

At this time there is no reason to think that Monster would not continue to support API access to its platform.

Level of Detail	**D**

Monster could produce aggregated statistical information about salary levels and job characteristics in London, and therefore would be useful in measuring occupational segregation and inequality within London's formal economy.

Utility	**C**

It would be possible to use Monster jobs data to infer levels of occupational segregation and inequality across the capital. However, because Monster only includes jobs within the formal economy that employers have chosen to advertise, the data would not be as representative as a survey of individuals' occupations.

Geographic Depth	**B**

Monster could produce aggregated statistical information about salary levels and job characteristics in London, and therefore would be useful in measuring occupational segregation and inequality within London's formal economy.

Representativeness	**D**

It is not possible to extract demographic information about Monster users. While the platform advertises a variety of jobs, it cannot be considered representative of all jobs and salary levels in London; many jobs and occupations will not be included on the platform.

Ethics of Access	**A**

Monster job listing data is made accessible through their public API.

Ethics of Publishing	**A**

Monster data on job listings is considered publicly accessible knowledge and does not contain any attributable personal information.

Public Perceptions of Privacy	A

The data that would be available to the GLA, Monster job listings, is publicly available and is intended to reach a broad public audience.

Skills Required	A

Monster jobs listings data is made accessible through a public API with query parameters that categorise job listings by country, state, city etcetera.

Frequency of Updates	B

Monster is updated daily with new jobs listings.

Costs of Ongoing Access	Free

See:

Monster	Monster.co.uk
Monster Job Search API	http://partner.monster.com/job-search

Examples of Possible Use Cases

Measure: Outcomes

Monster data on jobs listings could provide the GLA with an estimate of occupational segregation across the capital by measuring percentages of high/low paid occupations.

Mumsnet

Mumsnet was founded in 2000 as a social networking site and messaging board for parents across the UK covering topics relating to pregnancy, parenthood, conception and product reviews. It has since expanded to include a jobs board for working parents. Mumsnet Local allows users to search forums for topics relating to their local area within a 1 - 50 mile radius.

Data	C

Available data is largely free-text posts and comments encompassing conversations between users and listings for parenting-related services.

Current Availability	C

Mumsnet has no active API, meaning there is no official channel through which data can be programmatically collected. However, the site's forum is not password protected, and does not require membership to access. The site also does not prohibit automated 'scraping' of content - the process of collecting data by automatically reading it from a browser, though it does specify a number of specific pages which it does not allow to be scraped (see https://www.mumsnet.com/robots.txt) which should be respected. With some technical investment, a web scraping script could likely be configured to gather large numbers of posts and save these to a database for analysis.

Future Availability	B

Given its intended audience and mission, it is likely that Mumsnet will remain free to use and public. However, as this forum is a privately run site which does not rely on maintaining API access, there is a possibility that forum structure or access stipulations could change without notice.

Level of Detail	**D**

Mumsnet, like Netmums, is likely to produce long-form, nuanced reflections on parenting in London, and could therefore form part of an assessment of social integration concerning parents in London that included a qualitative element. Mumsnet has less detail at the borough-level than Netmums, which features forums and groups across London boroughs.

Utility	**C**

Due to its borough-level detail, Mumsnet Local data could be used to study participation in family-related leisure events, loneliness and social cohesion among parents and neighbourhood cohesion

Geographic Depth	**C**

There is little individualised data about Mumsnet users. Compared to Netmums, Mumsnet does not include as much borough-level detail due to a focus on general parenting discussion topics rather than local events and activities. However, it is possible to access subforums pertaining to discussion topics about parenting in London.

Representativeness	**C**

It is not possible to extract demographic data from Mumsnet. Mumsnet users are predominantly female. The majority are already parents, or expecting children. As such, we expect Mumsnet to be most useful in capturing the voices of London mothers. Anecdotally, Mumsnet has come under criticism in the past for being more exclusively middle class than Netmums, although it's not possible to quantify the extent to which this is true.

Ethics of Access	**B**

In the absence of an API, which is usually outlined to people in a signed user agreement, it is likely that users of the forum might not expect their data to be collected and processed at scale. Care should be taken to ensure that forums and users are notified of processing where necessary; we would also recommend that collected data be anonymised at source as far as possible.

Ethics of Publishing	**B**

Although Mumsnet is a public site and its users are broadly anonymous, as with most social platforms we would recommend not re-publishing attributable Mumsnet content.

Public Perceptions of Privacy	**A**

Mumsnet is primarily a public platform; threads are culturally public spaces.

Skills Required	**C**

The majority of useful Mumsnet information is long-form free text, making it difficult to analyse at scale without language classification technology.

Frequency of Updates	**A**

Data available on Mumsnet forums is near-live.

Costs of Ongoing Access	

As data would be scraped from the open internet rather than accessed through an API, no external costs for access will be imposed by the platform themselves. All costs will be

internal - primarily in technical support to continue scraping and storing the data, ensuring GDPR compliance etc.

Previous Successful Applications

Christine Hine from the Department of Sociology at the University of Surrey previously extracted data from Mumsnet to study parental attitudes to headlice. She coded themes including different emotional registers, treatment options and perceived sources of expertise.

Examples of Possible Use Cases

Measure: Participation in leisure

Mumsnet data could provide qualitative colour to a study of parental participation in leisure activities in London boroughs by scraping information about leisure activities and participation from threads tagged to London postcodes in the Mumsnet site.

Measure: Social isolation

By gathering data from subthreads about social isolation and loneliness, the GLA could provide in-depth qualitative data about self-reported loneliness among parents in London boroughs.

Measure: Helping neighbours/Social trust

The GLA could use data from Mumsnet threads tagged to London postcodes on neighbourhood initiatives and/or self-reported feelings of neighbourhood cohesion and social trust.

Measure: Diverse relationships/Social mixing

Mumsnet forums include multiple threads about ethnic diversity and experiences of being from an ethnic minority in relation to parenting and schools in London. Data could provide the GLA with qualitative evidence of the extent to which users report feeling they have diverse relationships or positive experiences of social mixing within London boroughs.

Netmums

Netmums was founded in 2000 as a social networking site and messaging board for parents across the UK. Users discuss a variety of topics in "chat" threads, while a "local" section features threads related to activities in parents' local area.

Data	**C**

Available data is largely free-text posts and comments encompassing conversations between users and listings for parenting-related services at the local borough level.

Current Availability	**C**

Netmums has no active API, meaning there is no official channel through which data can be programmatically collected. However, the site's forum is not password protected, and does not require membership to access. The site also does not prohibit automated 'scraping' of content - the process of collecting data by automatically reading it from a browser. With some technical investment, a web scraping script could likely be configured to gather large numbers of posts and save these to a database for analysis.

Future Availability	**B**

Given its intended audience and mission, it is likely that Netmums will remain free to use and public. However, as this forum is a privately run site which does not rely on maintaining API access, there is a possibility that forum structure or access stipulations could change without notice.

Level of Detail	**D**

Netmums, like Mumsnet, is likely to produce long-form, nuanced reflections on parenting in London, and could therefore form part of an assessment of social integration concerning parents at the London borough level that included a qualitative element.

Utility	**C**

Due to its borough-level detail, Netmums data could be used to study participation in family-related leisure events, loneliness and social cohesion among parents and neighbourhood cohesion

Geographic Depth	**C**

While there is little user-level data about Netmums users, location can be inferred through users' activities on specific borough-level subforums.

Representativeness	**C**

It is not possible to extract demographic data from Netmums. Netmums users are predominantly female. The majority are already parents, or expecting children. As such, we expect Netmums to be most useful in capturing the voices of London mothers.

Ethics of Access	**B**

In the absence of an API, which is usually outlined to people in a signed user agreement, it is likely that users of the forum might not expect their data to be collected and processed at scale. Care should be taken to ensure that forums and users are notified of processing where necessary; we would also recommend that collected data be anonymised at source as far as possible.

Ethics of Publishing	**B**

Although Netmums is a public site and its users are broadly anonymous, as with most social platforms we would recommend not re-publishing attributable Netmums content.

Public Perceptions of Privacy	**A**

Netmums is primarily a public platform; threads are culturally public spaces.

Skills Required	**C**

The majority of useful Netmums information is long-form free text, making it difficult to analyse at scale without language classification technology.

Frequency of Updates	**A**

Data available on Netmums forums is near-live.

Costs of Ongoing Access	

As data would be scraped from the open internet rather than accessed through an API, no external costs for access will be imposed by the platform themselves. All costs will be internal - primarily in technical support to continue scraping and storing the data, ensuring GDPR compliance etc.

See:

Netmums privacy policy https://www.netmums.com/info/privacy-policy

Netmums website https://www.netmums.com/

Examples of Possible Use Cases

Measure: Participation in leisure activities

Netmums data could provide qualitative colour to a study of parental participation in leisure activities in London boroughs by scraping information about leisure activities and participation from the "local" threads in the Netmums site.

Measure: Social isolation

One of Netmums' stated intentions is to "make it unnecessary for any mum to feel lonely or isolated". By gathering data from subthreads about social isolation, the GLA could provide in-depth qualitative data about self-reported loneliness among parents in London boroughs.

Measure: Helping neighbours/social trust

The GLA could use data from Netmums' "local" threads on neighbourhood initiatives and/or self-reported feelings of neighbourhood cohesion and social trust.

Measure: Diverse relationships/social mixing

Netmums' local forums include multiple threads about ethnic diversity and experiences of being from an ethnic minority in relation to parenting and schools in London. Data could provide the GLA with qualitative evidence of the extent to which users report feeling they have diverse relationships or positive experiences of social mixing within London boroughs.

Parkrun

Parkrun is a five kilometre running event that takes place every Saturday in 12 countries across the world. Users register online to participate in their local Parkrun - there are 47 groups in London. Parkrun's website and app store basic demographic information about Parkrun users, including names, age, gender, participation instances and finish times.

Data	**B**

Parkrun collects data on participants' age, gender, postcode and instances of participation in Parkrun events.

Current Availability	**A**

Parkrun has a public API giving researchers access to publicly-visible data, including users' gender, age grade, participation locations and finish times. It also makes anonymised aggregated data on participation in Parkrun events available to third parties, including the Department of Health and Social Care, in line with its mission statement to create "a healthier and happier planet" and to understand the "health and wellbeing of our communities".

Future Availability	**A**

We have no reason to believe this data would not be available in the future.

Level of Detail	**B**

Parkrun stores data about users' age, gender, location by postcode and participation instances.

Utility	**C**

Parkrun offers an insight into peoples' participation in leisure activities across 47 Parkrun groups in London, including a breakdown of participants' ages, genders and the frequency with which they attend Parkrun events. It would be most useful to the GLA for a sample of participation in leisure activities across particular boroughs.

Geographic Depth	**B**

Parkrun counts 47 groups across London, giving granular borough-level detail about participation in leisure activities.

Representativeness	**B**

According to participants, Parkrun attracts participants from a diverse range of ages, ethnicities and social backgrounds. There are 127,151 parkrunners in the UK and 47 parkrun groups in London.

Ethics of Access	**A**

Parkrun does not offer privacy settings, and users' name, age, gender, participation locations and finish times are publicly accessible.

Ethics of Publishing	**B**

Parkrun is a public site and its users are not anonymised. As with most social platforms we would recommend not re-publishing attributable Parkrun content, but aggregated Parkrun data that doesn't store personal information can be re-published freely.

Public Perceptions of Privacy	**B**

Parkrun is primarily a public platform. It does not provide users privacy settings in relation to their participation in Parkrun events and basic demographic information (name, age, gender).

Skills Required	**A**

Parkrun data covers basic demographic information and can be accessed using the website's public API.

Frequency of Updates	**C**

The frequency of Parkrun data updates is dependent on volunteers who log data at Parkrun events. Updates are not as frequent as a live social media platform, and a query will return data about past users and their participation instances.

Costs of Ongoing Access	**Free**

Previous Successful Applications

Parkrun has a research partnership with the Department of Health and Social Care funded research centre at Sheffield Hallam University. The Parkrun research board oversees projects that use Parkrun's dataset to investigate themes relating to social engagement, demography and disability.

See:

Parkrun http://www.parkrun.com/

Parkrun API https://www.parkrun.com/news/2014/10/09/parkrun-api-release/

Examples of Possible Use Cases

Measure: Participation in leisure activities

The GLA could access Parkrun data to comparatively measure the number of Parkrun participants, and their frequency of participation, across London boroughs.

Reddit

Reddit is a US based social news aggregation, content rating, and discussion website. Members post content such as questions and replies, links – both to external sites, and other Reddit feeds, and images, which are then voted up or down by other members.

Data	**A**

Access to Reddit data is through a simple API and is supported with extensive documentation and a well-supported GitHub repository. Available data is largely free-text posts and comments, and structured relationships between users and 'subreddits' or sub-forums on the platform.

Current Availability	**A**

At this time there is no reason to doubt continued support for API access to the platform.

Future Availability	**A**

We have no reason to believe this data would not be available in the future.

Level of Detail	**C**

Of all social media platforms, Reddit is most likely to produce long-form, nuanced reflections on life in London, and could therefore form part of an assessment of social integration that included a qualitative element.

Utility	**C**

As noted above, Reddit is likely to be most useful as a source of longer pieces reflecting on life in London than long-term measures of social inclusion.

Geographic Depth	**B**

There is little user-level data available on Reddit - users are anonymised and demographic or geographic information can only be inferred, most often through language analytics or identifying activity across age- or location-specific subreddits. For instance, a user who is active in /r/London is likely to be in the London area.

Representativeness	**C**

It is not possible to extract demographic information from Reddit data. Predominantly, the site is used by younger audiences and Pew research estimated two-thirds of Reddit users to be male[13]. The same research finds approximately 7 percent of Reddit users to be based in the UK. As such, we deem it highly unrepresentative of Londoners in general. It is most likely to be useful in capturing the voices of younger, technologically-savvy Londoners.

Ethics of Access	**A**

[13] Michael Barthel, Galen Stocking, Jesse Hollcomb and Amy Mitchell, *Reddit news users more likely to be male, young and digital in their news preferences.* Pew Research Centre, February (2016). Accessed at: http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Reddit provides an open API and encourages its use. It asks users to respect the terms and conditions.

Ethics of Publishing	B

Although Reddit is a public site and its users are broadly anonymous, as with most social platforms we would recommend not re-publishing attributable Reddit content. Aggregated Reddit content can be shared freely.

Public Perceptions of Privacy	B

Reddit is primarily a public platform, describing itself as 'the front page of the internet'. Subreddits most likely to be useful to the GLA (/r/London, for instance) are culturally public spaces.

Skills Required	C

The majority of useful Reddit information is long-form free text, making it difficult to analyse at scale without language classification technology.

Frequency of Updates	A

Data available through the Reddit API is near-live, and a query will return posts, comments and other metadata from the time the query is made.

Costs of Ongoing Access	Free

Previous Successful Applications

Tech Nation analysed posts pertaining to careers and career development as part of its assessment of the state of technology-related careers in the UK.[14] The team analysed the contents of conversations pertaining to careers and categorised them by the sector discussed.

See:

Reddit API Documentation	https://www.reddit.com/dev/api/
API GitHub	https://github.com/reddit-archive/reddit/wiki/API

Examples of Possible Use Cases

Measure: Misc

Reddit could be used to bring colour to wider social integration measurement programmes by capturing the 'voice of the Londoner' through long-form reflections on life in London as captured through data collections and manual review of Reddit conversations.

Measure: Feeling of Belonging

Reddit could also be used to understand the wider interests of London-based Reddit users through correlating activity in the London subreddits with wider activity on the platform. Through an analysis of comments reflecting on current affairs and news, the GLA may be able to monitor the ways in which Reddit users in London reflect on issues of social inequality, politics and education.

[14] https://technation.io/talent/future-talent-key-findings/

Bumble

Bumble is a global dating app based in the US. Unlike Tinder, it requires women to take the first move in initiating conversation. Users swipe through images of potential "matches" (dates) and chat through the app's messaging interface.

Data **A**

Stored data includes users' education background, age, location, interests, occupation, political outlook, lifestyle and information about the types of people they are interested in interacting with and a record of conversations between users. Bumble states that it shares aggregated information with third parties but does not share data that can be used to identify individuals.

Current Availability **D**

Bumble does not have a public API. It does share aggregated data with third parties for advertising purposes, however the GLA would need to contact Bumble directly in order to get hold of this aggregated data.

Future Availability **D**

The GLA would not be able to access Bumble data without first negotiating a formal agreement with Bumble.

Level of Detail **B**

Bumble houses detailed information about users, including their age, occupation, educational background and location, and provides detailed information about social relationships and interactions between people living in London. The company makes personal data available for third parties and demographic profiling as long as that data can't be used to identify people directly.

Utility **D**

While Bumble could theoretically provide detailed information on social integration in London through data about users' demography, social interactions and social relationships, scraping this data raises ethical and practical question (see ethics of access and publishing, and public perceptions of privacy).

Geographic Depth **B**

Bumble users share their geographic location in order to use the app, and are able to specify their preferred geographic range for potential matches.

Representativeness **B**

Almost three quarters of Bumble users are under 35. Bumble was launched to "disrupt traditional gender roles in heteronormative dating"; it has a more equal gender split than Tinder.

Ethics of Access **D**

Bumble states that it "may share aggregated information with third parties" including "personal data (but which doesn’t identify you directly) together with other information including log data for industry analysis and demographic profiling". However, Bumble does not have a public API and therefore the GLA would need to negotiate with Bumble directly in order to access its data.

Ethics of Publishing **D**

Bumble data is not public. We would therefore not recommend re-publishing attributable Bumble content or attempting to publish non-attributable content without first negotiating with Bumble.

Public Perceptions of Privacy	D

Anyone can sign up to use Bumble. However, user data is only shared with potential "matches", and Bumble is not considered a public platform.

Skills Required	C

Unknown.

Frequency of Updates	A

Bumble swipes (interactions between users) and conversations are live.

Costs of Ongoing Access	Free

See:

Bumble privacy policy	https://bumble.com/privacy/
Bumble	https://bumble.com/

Examples of Possible Use Cases

Measure: Diverse Relationships

By analysing the regularity of interactions between users with different demographic traits (e.g. educational background, ethnicity and profession), the GLA could hypothetically use Bumble's data as a proportional estimate of the diversity of romantic and social relationships among young Londoners across the capital.

Measure: Social Mixing

The GLA could hypothetically study the frequency of "swipes" and social interactions between Bumble users in London with different ethnicities, social class backgrounds and ages from each other.

Duolingo

Duolingo is a language learning platform which offers 37 languages and a digital language proficiency exam. It has approximately 300 million users and is among the most popular and widely used services of its type.

Current Availability	D

There is no publicly accessible useful data and no examples of them sharing this data.

Previous Successful Applications	A

Duolingo has produced its own analysis across the states of America, looking at rates of language learning, dedication of those learners, and the variety of languages being learnt.

Facebook

Facebook is the world's largest social networking platform, with more active users than any other platform. It holds a huge amount of data about individual users, groups, events and businesses, though is increasingly limiting the extent to which this data is publicly available.

Current Availability	D

Since the Cambridge Analytica scandal, Facebook is undergoing a reassessment of its access protocols and has limited the ability of third parties to collect and analyse its data. This requires applications collecting data to register with Facebook, and is approved on Facebook's discretion.

Examples of Possible Use Cases

Measure: Relationships

Using self-reported demographic data on gender, age, ethnicity etc. on users within London and constructing a network map of their friendships, the GLA could determine the diversity of Londoners social groups across different areas and groups. Looking at the interconnection of those networks could also determine the extent to which Londoners have bonding (intra-group) social capital versus bridging (inter-group) social capital.

Glassdoor

Glassdoor is a website where employees and former employees anonymously review companies and their management.

Current Availability	D

Glassdoor has a public API, however there is next to no documentation or indication of what is available to access. It suggests that there are further private APIs but the nature of and access to these APIs is only available to approved partners. Further, Glassdoor's terms of use specifically prohibit the scraping of data from its site.

See:

https://www.glassdoor.co.uk/about/terms.htm

Gumtree

Gumtree is a public listings website for services, classified adverts, properties, jobs and community events based in the UK.

Data	D

Available data on Gumtree is largely free-text classifieds adverts and posts from users procuring and/or advertising goods and services. Gumtree is predominantly a classified advertising website. Because of this, it's difficult to measure users' self-reported attitudes towards social integration, or to gauge the extent to which people are benefitting from or participating in the services that they are procuring / advertising.

Current Availability	D

Gumtree does not offer a public API, and the site's robots.txt file prevents large portions of the site from being scraped by other means.

Future Availability	D

Pending a change in policy from Gumtree, there is little reason to believe that access to Gumtree will become available in future

Level of Detail	D

Gumtree collects basic information about users' browsing patterns, location, email addresses and mobile numbers, but does not directly collect demographic information at the individual citizen level. Moreover, the Gumtree content that would be useful to the GLA - the "community" section of the website - does not include information about the extent to which citizens participate in community events, services like ridesharing, or volunteering opportunities. Therefore Gumtree could only be used to suggestively infer particular measures of social integration.

Utility	D

Specific Gumtree pages including those on skills exchange, ridesharing, lost and found and sports teams, could be useful in adding colour to research on social trust/ helping neighbours, indicating a proclivity to trust neighbours and local residents.

Geographic Depth	B

Gumtree posts are tagged with specific London boroughs, allowing researchers to infer users' locations within the capital.

Representativeness	D

It is not possible to extract demographic information from Gumtree data. There are no publicly available statistics on Gumtree users - as such we deem it highly unrepresentative of Londoners in general.

Ethics of Access	C

Gumtree is a public space: items and advertising posted there is designed to be seen by the public. However, the site has taken steps to disallow the collection of this content, and users may not expect their posts to remain accessible to third parties once their posts have been taken down.

Ethics of Publishing	B

Gumtree is a public site and its users are broadly anonymous. But as with most social platforms we would not recommend re-publishing attributable content. Aggregated Gumtree content can be shared freely.

Public Perceptions of Privacy	C

Gumtree is primarily a public platform, although public outcry followed in a case involving Gumtree Australia where personal data including email addresses and phone numbers was stolen by hackers.

Skills Required	D

As Gumtree adverts are unlikely to be conducive to automated collection, this question does not apply.

Frequency of Updates	B

If data could be collected, the site is constantly updated, so analysis could proceed on near-live data

Costs of Ongoing Access	N/A

See:

Gumtree website	https://www.gumtree.com/
Gumtree API	https://github.com/GumTreeDiff/gumtree/wiki/GumTree-API

Examples of Possible Use Cases

Measure: Helping neighbours / Social trust

Gumtree classifieds about ridesharing in London could be used within a study of helping neighbours/ social trust to illustrate Londoners' propensity to ride-share with others. Likewise, Gumtree classifieds on language and skills swaps could add detail and colour to a study of helping neighbours / social trust.

Measure: Helping neighbours / Social trust

Data from Gumtree's lost and found section could be used to infer the extent to which residents within particular areas of London perceive there is social trust within their neighbourhood through the proportion of "lost" objects - indicating petty crime / theft - versus the proportion of "found" objects, indicating a propensity towards returning lost objects.

Tinder

Tinder is a US based dating app that is now used around the world. Its users are predominantly urban residents aged 18-34. The app enables people to register potential "matches" - dates - by swiping left (no) or right (yes) on other users whose profile pictures are displayed onscreen.

Data	A

Tinder data includes users' education background, age, location, interests and occupation, in addition to information about the types of people they are interested in interacting with and a record of conversations between users. Tinder states that it shares "non-personal information" and "personal information in hashed, non-human readable form", suggesting that available data would be in an aggregated form that could not be used to identify individual demographic characteristics.

Current Availability	D

Tinder does not have a public API. It shares non-personal and non-human readable personal data with third parties, but the GLA would need to contact the company directly in order to access this data.

Future Availability	D

The GLA would not be able to access Tinder data without first negotiating an agreement with Tinder.

Level of Detail	B

In addition to detailed information about users, including their age, occupation, educational background and location, Tinder provides information about social

relationships and interactions between people living in London through its "swipe" function and messaging interface.

Utility	**D**

While Tinder could theoretically provide detailed information on social integration in London through data about users' demography, social interactions and social relationships, scraping this data raises ethical and practical question (see ethics of access and publishing, and public perceptions of privacy)

Geographic Depth	**B**

Tinder has detailed geographic information about its users. Location sharing is a mandatory aspect of the Tinder app. In addition, Tinder's optional "Places" feature, which is not currently available in the UK, will allow users to log their frequently visited places and see the favourite places of other users.

Representativeness	**C**

The majority of Tinder users are aged between 16 and 34, with few people over the age of 50 reporting they use Tinder. There are also proportionately more men (approximately 60/40) than women on Tinder.

Ethics of Access	**D**

Tinder states that it may "use and share non-personal information" and "personal information in hashed, non-human readable form" with third parties to develop targeted in-app advertising. However, Tinder does not have a public API and therefore the GLA would need to negotiate with Tinder directly in order to access its data.

Ethics of Publishing	**D**

Tinder data is not public. We would therefore not recommend re-publishing attributable Tinder content or attempting to publish non-attributable content without first negotiating with Tinder.

Public Perceptions of Privacy	**D**

Although anyone can sign up to Tinder, the app is not a public platform. In instances where programmers have previously attempted to scrape data, including user pictures that were later published online, public backlash has followed.[15]

Frequency of Updates	**A**

Tinder swipes and conversations are live.

Costs of Ongoing Access	**N/A**

See:

Tinder privacy policy	https://www.gotinder.com/privacy
Tinder website	https://tinder.com/?utm_source=gotinder-nav-flame

Examples of Possible Use Cases

Measure: Diverse relationships

By analysing the regularity of interactions between users with different demographic traits (e.g. educational background, ethnicity and profession), the GLA could hypothetically use

[15] Mary Papenfuss, *Massive Tinder Photo Grab Is Latest Scary Warning to be Careful What You Post.* Huffington Post, April (2017). Accessed at: https://www.huffingtonpost.co.uk/entry/40000-photo-tinder-sweep_us_59052818e4b0bb2d086f0335

Tinder's data as a proportional estimate of the diversity of romantic and social relationships among young Londoners across the capital.

Measure: Social mixing

The GLA could hypothetically study the frequency of "swipes" and social interactions between Tinder users in London with different ethnicities, social class backgrounds and ages from each other.

TripAdvisor

TripAdvisor, Inc. is an American travel and restaurant website company that shows hotel and restaurant reviews, accommodation bookings and other travel-related content.

Current Availability	**D**

TripAdvisor does not allow access to the Content API for purposes of data analysis or academic research.

Uber Taxis

Uber is a peer-to-peer ridesharing and taxi cab company headquartered in San Francisco.

Current Availability	**D**

There is no way of publicly accessing live Uber data. The API provided allows app developers to call an Uber only. However, Uber Movement provides anonymised data from over two billion trips to help urban planning around the world, and may hold a small amount of value and as such be worth investigating.

See:

Uber Developer Front End https://developer.uber.com/

Conclusions

Digital and online data gives researchers new tools for understanding and measuring human behaviour, relationships and social attitudes. But while online data can produce real-time pictures of urban trends and social processes, it also suffers from selection bias and patchiness that researchers must keep in mind when attempting to further their understanding of complex social scientific questions. Big data cannot serve as a replacement for methods like interviews, surveys and censuses that are the bread and butter of social-scientific research.

For the GLA's purposes, however, digital and online data may prove a useful addition to traditional survey methods for understanding social integration in London. Whether using Google Street View to infer levels of occupational and economic segregation from visual signifiers, or drawing on Parkrun data to paint a snapshot of Londoners' involvement in leisure activities, the sources analysed in this report all contain potential avenues for practical application and qualitative colour. Moreover, when placed in conversation with aggregated data from large-scale datasets, such as the annual Integrated Household Survey or School Census, digital and online data can enhance researchers' understanding, add insight to their findings, and contribute to robust evidence bases.

We found the concept of "nowcasting" helpful in understanding how the GLA could make use of digital and online data sources for measuring social integration. This refers to forecasting on a very short timescale, where researchers use available data to predict the present or near future. In this way, online data can equip researchers with an "ear to the ground", or an extra tool for shedding light on the present. Nowcasting may not provide an exact or total picture of social reality, and the partiality of social media data may lead researchers to become overly reliant on inference. However, nowcasting can estimate the present in ways that traditional social-scientific research methods cannot. Where surveys are costly, lengthy to produce, and provide results after many months have already elapsed, digital and online data enables researchers to gauge the pulse or mood of contemporary issues. This is made evident on a platform like Twitter, where users share thoughts instantaneously, repeatedly, and unprompted by a survey design or interviewer.

To make the most of emergent digital data sources, we recommend that the GLA uses big data as an interim means of measuring social integration at intervals between survey research. For instance the GLA could use Foursquare and Yelp to map urban gentrification through the prevalence of particular types of businesses correlated with increasing house prices and household income, and use this to infer increasing social polarisation. In addition, we recommend placing digital sources in conversation with survey and census data to estimate the extent to which key social integration measures might be predictable or estimable through digital sources.

Eight Ideas for Measuring Social Integration

- NHS England publishes monthly **data about the drugs prescribed by GP practices and Clinical Commissioning Groups**. The GLA could gather statistics on the prescription of antidepressants within London boroughs, in conjunction with keyword analysis of self-reported loneliness on Twitter, to infer pockets of social isolation and loneliness within London. Similar analysis using Netmums and Mumsnet could reveal social isolation among mothers who may be particularly vulnerable to loneliness, especially during maternity leave.
- Using **place mapping data**, e.g. Google Maps, the GLA could track the proliferation of chain shops versus independent businesses and markets. A high proportion of chain stores and the homogenisation of an area through the replacement of independent businesses could indicate a loss of local culture and rising rents, which may lead to residents feeling less invested in their local area and having a lower sense of belonging.
- Data on **mobile phone usage,** including handset type, the extent to which users make calls to people located outside of their community, the timeliness of bill payments and data usage could provide useful indices of London residents' socio-economic class and financial resilience, in addition to helping to map commuting times and geographic mobility.
- London's civic infrastructure – its parks, libraries and public spaces – are essential to encouraging social mixing and integration. The GLA could use Google Maps to **infer the opportunities for social integration** within particular wards according to their built environment and the prevalence of public spaces and civic meeting places.
- Using **review data from customers and employees of local businesses**, by collating reviews of given businesses across Trustpilot, Yelp, Google Maps, etc. the GLA could perform keyword searches for mentions of discriminatory activity. This would allow the GLA to map low-level unequal outcomes that may go unreported by law enforcement records and identify where different types of discrimination, e.g. class, gender, race etc., are most prevalent, giving a more granular measure of social integration.
- 12% of workers in London work night shifts[16], creating issues around sleep and disruption that can have a corrosive effect on workers' mental health and social lives. The GLA could use **available data on the number of people working night shifts**, for example by looking at opening times on Google Maps, advertised hours on job listings on Indeed, and mobile phone usage, to indicate the synchronicity of London's neighbourhoods and infer a broader picture of social integration.
- The GLA could make use of existing offline data sources, including **school census data** to measure pupil demography, and **house price data** to measure housing affordability, in conjunction with online data sources that provide an indication of gentrification, such as Yelp, to create a broader evidence base for measuring social mixing and urban segregation.
- Using **longitudinal data like house sales records** from Zoopla, rental advertisements, and sampling of business mapping data, the GLA could measure the rate of churn of residents and businesses within a neighbourhood. Residing in an area for longer increases a resident or proprietor's sense of belonging, as confirmed by data from the Community Life Survey 2016-17, and deep, cross-cutting social bonds within an area take time to develop. By implication, a high rate of churn could suggest that these bonds are constantly having to be refreshed and rebuilt in the neighbourhood, suggesting a lower level of social integration

[16] Trade Union Congress, *Number of People working night shifts up by more than 25,000, new TUC analysis reveals.* October (2016). Accessed at: https://www.tuc.org.uk/news/number-people-working-night-shifts-more-250000-2011-new-tuc-analysis-reveals

Appendix A

Table 1: List of Data Sources Reviewed

Reddit	Find a Job	Gumtree
Trustpilot	Yelp	Freecycle
Citymapper	Google Street View	Indeed.com
Google Maps	JustGiving	Monster
Meetup	Glassdoor	JustGiving
Eventbrite	Uber	Airbnb
Zoopla	Facebook*	TripAdvisor
YouTube	Bumble	Mumsnet
Twitter	Netmums	Park Run
Tinder	Google Street View	

Table 2: List of Data Sources Considered But Not Comprehensively Reviewed

LinkedIn*	newsapi.org	Wikipedia
SpareRoom	OpenTable	Change. Org
Rightmove	Ofo	Oyster
Duolingo*	Mobike	Google Search Trends
Instagram	Quiqup	WhatsApp
Snapchat	Dice	Tumblr

**It would be worthwhile to explore bespoke partnerships with these data sources would be worthwhile exploring bespoke partnerships with*