verian

Strand 1 Data Source Prioritisation Log

R&D Science and Analysis Programme: Capturing Engagement Numbers

This workbook contains 6 tabs:

Tab 1: This tab contains one row per greenlit data source with a final outcome from their initial prioritised status, including if and how they have been used in the depth phase to conclude Strand 1 work.

Tabs 2-6: Each of these tabs contain the outcome for each data source in a category assessed against the prioritisation criteria set out in sections 5-9 of the Strand 1 report.

verian

Data Category	Data Source	Description	Granularity and Frequency	Access	Attendance and Participation	Prioritisation for Depth Phase	Depth Phase Outcome & Future Use
Transport	**Strava Metro**	Aggregated information about human-powered travel (bikes, running, walking). Includes purpose of travel (leisure vs. commute). Includes routes taken between 2 destinations.	Street-level. Hourly, with a 1-2 week delay. Oversamples young, male audiences.	Access to features can be secured via an application to Strava Metro setting out how data users are directly involved in active transportation infrastructure planning. This project used a letter of recommendation from DCMS to secure access cost-free as the project is contributing to the evidence base on the use of Strava data for research purposes. It is unclear if this would be the case more generally or for DCMS more widely moving forward.	May be used to estimate participation in active travel and sport. Also may be used to estimate attendance to spaces.	Green Successful use for similar purposes identified in the literature reviewed in the evidence synthesis during the breadth phase.	Application for testing in Edinburgh, following DCMS letter of recommendation, was successful. Depth phase testing was successful, producing moderately positive correlation. Carried forward to Strand 2 for further experimentation at 5 case study sites.
Transport	**TomTom Traffic API**	Traffic data, including flow, incidents, origin destination analysis and historical crowding data.	Road-level and user-definable areas. Real-time, with multiple years of historical data.	Pre-paid. Access is secured via long-term subscriptions.	Could be used to determine attendance at an event. Could be used to determine demographics of attendees based on their origin of travel.	Green High granularity and frequency, alongside wide UK coverage, making it notably useful for events where attendees travel by car (e.g., rural events). Its usefulness in urban areas with public transport options will be limited.	Decision not to advance due to prohibitive cost model. This was true for both access directly via TomTom and by third party providers such as AtkinsRealis.
Social Media	**Pulsar**	Mining of social media streams (Twitter/Facebook/Instagram /Threads/Tiktok) for mentions, as well as blogs, forums and Reddit. Event-related keywords have to be identified in advance.	Since this source can combine content data from several platforms it can reflect various sources of data, thereby allowing for more diversity, larger volumes of data and a richer overall granularity.	Costs are based on the desired contract length and desired volume of 'mentions' per month. As an indication, this project has negotiated a bespoke agreement with Pulsar for a 6 month contract allowing access to up to 200,000 'mentions' per month. Pulsar usually insists on a 12 month contract, which was beyond the scope of this project but could be desirable for a future user.	Event-related keywords must be selected in advance and evidence of participation depends on the success/quality of the search inputted. Access is through the Pulsar dashboard. Audience segmentation can be provided through another tool available from Pulsar.	Green Faculty.AI have an existing contract with Pulsar (as do colleagues in DCMS).	Following delays in the DPIA data approvals process, Depth phase testing has been completed for 3 test sites: Scotland Rugby International Games, Farnborough International Airshow and Lewes Bonfire Night. These sites were chosen to provide balance of events that would have a high level of in-person attendance, and those likely to have high levels of virtual-only participation.
Mobile App	**Huq**	Individual-level GPS data collected from the use of up to 300 mobile phone applications on GPS enabled devices. Huq use this data to create dashboards where use of specific points of interest can be examined.	The raw data is near real-time. The company offers aggregate products based on this. Coverage varies by geographic area but enables coverage of around 1-5% of the population.	Cost is generally to access a dashboard, based on number of points of interest. However, this project through University of Glasgow's existing working relationships with Huq has enabled a bespoke arrangement to be agreed, allowing privileged access to datasets beyond what is provided through the dashboard. This agreement includes data science support from Huq to further understand their data.	Could be used to estimate attendance for fixed sites. Could be used to estimate demographics or reach.	Green This data benefits from demographic and visitor origin information. Furthermore, members of the consortium have a working relationship with the company. Data should be prioritised for utilisation in combination with regression and machine learning. It can also be used to estimate event demographics and reach in combination with sample weighting.	Depth phase testing has produced estimated visitation numbers that correlate strongly with trends in the official visitation data for the British Museum, including through the peaks and troughs of various lockdown restrictions. The strong positive correlation of 0.77 for the monthly estimate (with some month-on-month variation) is a positive outcome for a first attempt at this type of analysis. Further experimentation will now take place in Strand 2.
Mobile App	**ActiveXChange**	Individual-level GPS movement data targeting health and wellbeing. The data captured is provided primarily as an 'activity index' that reflects the level of use in the specified timespan and geographic space.	Aggregated at an hourly level.	Data provided directly by the company. Cost is based on number of points of interest requested.	Could be used to estimate attendance for fixed sites.	Green Data should be prioritised in combination with regression and machine learning but does not contain reach or demographic information.	The data did not contain any information about reach or user demographics. This limited its value over alternatives such as Huq, which we have progressed with.
Mobile App	**Unacast**	This data shows the number of device visits by day of the week and time of the day for residents, workers and other type of visitor profile.	Aggregates daily and hourly patterns at monthly intervals using quadgrids or census areas.	Data available indirectly via Carto.	Could be used to estimate attendance for fixed sites.	Green Data should be prioritised in combination with regression and machine learning but does not contain reach or demographic information.	This data sources was available through the Carto geospatial dashboard. Decision taken not to advance due to prohibitive cost model of Carto. Attempts to access the data directly from Unacast were unsuccessful.
Mobile App	**CKDelta**	Using mobile app data: (1) Footfall by dwell times (2) Catchment Areas (3) Footfall by age bands	Aggregated (300m * 300m grid or 1km * 1km grid) every 30 minutes.	Data available from Carto, linking three data products together.	Could be used to estimate attendance for fixed sites. Could be used to estimate demographics or reach.	Green This data may benefit from some demographic and visitor origin information. Data should be prioritised in combination with regression and machine learning. It may also be used to estimate event demographics and reach.	This data sources was available through the Carto geospatial dashboard. Decision taken not to advance due to prohibitive cost model of Carto. Attempts to access the data directly from CKDelta were unsuccessful.
Mobile App	**Echo Analytics**	Develops footfall estimates to points of interest from mobile app location data.	Weekly, monthly, or quarterly trends in footfall counts. Echo Analytics do not provide absolute numbers.	Access available via bundles or subscription model from Echo Analytics directly.	Could be used to estimate trends in attendance for fixed sites.	Green Data should be prioritised in combination with regression and machine learning but does not contain event catchment or demographic information.	Following negotiations with Echo Analytics decision made not to advance due to prohibitive cost model.

Deployable-Sensing	**Drones**	Drones can be used to capture images of the entire area of an outdoor event (e.g., music festivals). ML models can then be used to provide counts.	Can count individual attendees.	Can only be deployed for outdoor events and would require collaboration of events' organisers. Adverse weather conditions may hinder the deployment of the data source. Agreement to share this data, if required, is rarely straightforward.	N/A	**Green** The accuracy of the counts will be highly dependent on how well the pictures capture the entire crowd, but initial testing and literature show promise.	Unable to access for Depth Phase testing but hope to trial in Strand 2 alongside methods developed for other data sources via footage provided from Bradford City of Culture. Access to footage of the homecoming parade via Merseyside Police was not pursued owing to a lack of baseline data for validation.
Deployable-Sensing	**Conventional Cameras**	Conventional cameras can be used to capture images of the crowd at an event. ML models can then be used to provide counts.	Can count individual attendees.	Would require collaboration of events' organisers but can be used for both indoor and outdoor events. Agreement to share this data, if required, is rarely straightforward.	N/A	**Green** The accuracy of the counts will be highly dependent on how well the pictures capture the entire crowd, but initial testing and literature show promise.	Despite being greenlit initially logistical challenges in sending out project members to take photographs along with unresolved GDPR concerns on consent of data subjects and privacy means this option has not been pursued further at this stage.
Deployable-Sensing	**CCTV** **(based on UBDC CCTV Object Detection)**	Fixed, 24/7 cameras can be used to track attendance over time to a given space. ML models can then be used to provide counts.	Can count individual attendees.	Would require collaboration of events' organisers or CCTV provider. Agreement to share this data, if required, is rarely straightforward.	N/A	**Green** Highly dependent on the availability and positioning of CCTV cameras.	The team from University of Glasgow have worked on a separate Glasgow based project and the approach has shown potential to estimate footfall at various sites of interest. Not taken forward by this project at this time.
Event and Space	**OpenStreetMap**	Geographic database, including roads and points of interest.	UK-wide, street-level. Database updated daily.	Open and free. Can be accessed through multiple APIs, including Overpass API.	N/A	**Green** Can be used to obtain relevant Points of Interest (POIs) like sports or cultural spaces to track participation.	This data is being used to retrieve polygons of cultural and sports sites.
Event and Space	**Ordnance Survey Points of Interest**	Geographic database including points of interest by category.	UK-wide, per point of interest. Updated quarterly.	Accessible through an Ordnance Survey partner. Released as a PDT file.	N/A	**Green** Can be used to obtain relevant POIs to track participation.	Being used in the mobile phone app framework to help get location boundaries and potentially feature as part of estimate calculations.
Event and Space	**England & Wales Census**	Census 2021 for England and Wales.	Output areas of 40-250 households.	Open and free. Can be accessed through multiple APIs, including Overpass API.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.	Potential for use in validation efforts as part of Strand 2 experimentation
Event and Space	**Scotland Census**	Census 2022 for Scotland.	Output areas of 50 households.	Open and free Staggered release of data through 2023-2024.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.	Potential for use in validation efforts as part of Strand 2 experimentation
Event and Space	**Northern Ireland Census**	Census 2021 for Northern Ireland.	Variable (3,780 Data Zone outputs of different size).	Open and free. Can be accessed through multiple APIs, including Overpass API.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.	Potential for use in validation efforts as part of Strand 2 experimentation
Event and Space	**Indices of Multiple Deprivation (IMD)**	IMD data for all UK nations is available from the Consumer Data Research Centre.	Variable depending on UK nation, detailed geographic data zones in Scotland and lower super output area level in England.	Open and free.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.	The team has obtained this open dataset and are using it in combination with mobile phone app data to estimate demographics of attendance to cultural sites.
Event and Space	**CACI's Acorn**	Geo-demographic segmentation.	Segments for UK postcodes, 7 categories, 22 groups and 65 types.	Public sector version available, pay for access.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.	Remained an option but Index of Multiple Deprivation data is being used in preference as a geo-demographic dataset.
Event and Space	**Experian's Mosaic**	Geo-demographic segmentation.	Segments for UK postcodes, 15 summary groups, 66 detailed types.	Central government version available, pay for access.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.	Potential use for sampling or weighting but ultimately not taken further as not event specific so less helpful than other sources for generating estimates.

verian

Data Source	Description	Granularity and Frequency	Access	Attendance and Participation	Prioritisation for Depth Phase
Huq	Individual-level GPS data collected from the use of up to 300 mobile phone applications on GPS enabled devices. Huq use this data to create dashboards where use of specific points of interest can be examined.	The raw data is near real-time. The company offers aggregate products based on this. Coverage varies by geographic area but coverage of around 1-5% of the population.	Cost is generally to access a dashboard, based on number of points of interest. However, this project through University of Glasgow existing working relationships with Huq have negotiated a bespoke agreement with Huq allowing privileged access to datasets requested for testing and experimentation through the Depth Phase and Strand 2 activity. This agreement includes data science support from Huq to further understand their data.	Could be used to estimate attendance for fixed sites. Could be used to estimate demographics or reach.	**Green** This data benefits from demographic and visitor origin information. Further, members of the consortium have a working relationship with the company. Data should be prioritised to be used in combination with regression and machine learning. It can also be used to estimate event demographics and reach in combination with sample weighting.
ActiveXChange	Individual-level GPS movement data targeting health and wellbeing. The data captured is provided primarily as an 'activity index' that reflects the level of use in the specified timespan and geographic space.	Aggregated at an hourly level.	Data provided directly by the company. Cost is based on number of points of interest requested.	Could be used to estimate attendance for fixed sites.	**Green** Data should be prioritised in combination with regression and machine learning but does not contain reach or demographic information.
Unacast	This data shows the number of device visits by day of the week and time of the day for residents, workers and other type of visitor profile.	Aggregates daily and hourly patterns at a monthly interval using quadgrids or census areas.	Data available indirectly via Carto.	Could be used to estimate attendance for fixed sites.	**Green** Data should be prioritised in combination with regression and machine learning but does not contain reach or demographic information.
CKDelta	Using mobile app data: (1) Footfall by dwell times (2) Catchment Areas (3) Footfall by age bands	Aggregated (300m * 300m grid or 1km * 1km grid) every 30 minutes.	Data available from Carto, linking three data products together.	Could be used to estimate attendance for fixed sites. Could be used to estimate demographics or reach.	**Green** This data may benefit from some demographic and visitor origin information. Data should be prioritised in combination with regression and machine learning. It may also be used to estimate event demographics and reach.
Echo Analytics	Develops footfall estimates to points of interest from mobile app location data.	Weekly, monthly, or quarterly trends in footfall counts. Echo Analytics do not provide absolute numbers.	Access available via bundles or subscription model from Echo Analytics directly.	Could be used to estimate trends in attendance for fixed sites.	**Green** Data should be prioritised in combination with regression and machine learning but does not contain reach or demographic information.
Airsage	Individual-level GPS data collected from the use of mobile phone applications on GPS enabled devices.	Aggregated daily or hourly to 10, 100 or 1000m grids.	Unclear - sample data requested but not provided.	Could be used to estimate attendance for fixed sites.	**Amber** Given greenlit data sources already in place, kept in reserve for this project.
Geolytix	Develops mobility analysis based on mobile phone app data.	Aggregated to grid or census area.	Provided directly by the company. Cost unknown.	Could be used to estimate attendance for fixed sites.	**Amber** Given greenlit data sources already in place, kept in reserve for this project.
Tamaco	Individual-level GPS data collected from the use of mobile phone applications on GPS enabled devices.	The raw data is near real-time. The company offers aggregate products based on this.	Discontinued in UK but University of Glasgow had a partnership and have historic access to raw data for non-commercial purposes.	Could be used to estimate attendance for fixed sites. Could be used to estimate demographics or reach.	**Red** Although data available to UBDC, license does not permit use and company no longer service UK.
Cuebiq	Uses mobile phone app data to provide metrics on visits and dwell time for points of interest.	Real-time.	From the information we have been able to assess not clear if they operate in the UK.	Could be used to estimate attendance for fixed sites.	**Red** Given lack of clarity around accessibility.

Factori	Develops consumer insights from mobile app location data.	Monthly. Aggregated (polygon/Uber hex).	Provided directly by the company or through Carto platform.	Could be used to estimate attendance for fixed sites.	**Red** Attempts to access data directly through Factori have been unsuccessful and cost through Carto platform is prohibitively expensive for this project (approx. £50,000 per license). Methods developed using Huq in preference would be transferable should DCMS pursue this source in future.
Citymapper Mobility Index	Aggregated navigation and mobile data providing mobility trends.	Daily	Discontinued		**Red** Discontinued.
Apple Maps Mobility Trends Reports	Aggregated navigation and mobile data providing mobility trends.	Daily	Discontinued		**Red** Discontinued.
Google Community Mobility Reports	Aggregated navigation and mobile data providing mobility trends.	Daily	Discontinued		**Red** Discontinued.

verian

Data Source	Description	Granularity and Frequency	Access	Attendance and Participation	Prioritisation for Depth Phase
Strava Metro	Aggregated information about human-powered travel (bikes, running, walking). Includes purpose of travel (leisure vs. Commute). Includes routes taken between 2 destinations.	Street-level. Hourly, with a 1-2 week delay. Oversamples young, male audiences.	Access to features can be secured via an application to Strava Metro setting out how data users are directly involved in active transportation infrastructure planning. This project has used a letter of recommendation from DCMS to secure access cost-free as it is contributing to the evidence base on the use of this data. It is unclear if this would be the case more generally or for DCMS more widely moving forward.	May be used to estimate participation in active travel and sport. May be used to estimate attendance to spaces.	Green Used in the literature reviewed for similar purposes as this project.
TomTom Traffic API	Traffic data, including flow, incidents, origin destination analysis and historical crowding data.	Road-level and user-definable areas. Real-time, with multiple years of historical data.	Prepaid. Access is secured via long-term subscriptions.	Could be used to determine attendance at an event. Could be used to determine demographics of attendees based on their origin of travel.	Green High granularity and frequency, alongside wide UK coverage, making it notably useful for events where attendees travel by car (e.g., rural events). Its usefulness in urban areas with public transport options will be limited.
Transport for London (TfL)	TfL transport data. Includes occupancy of car parks and bike points, as well as crowding information for most Underground stations. Includes some traffic data for roads managed by TfL.	Varies. Most live data is updated every 5 minutes for individual stations or vehicles. Static data sources are also available. No raw counts available for crowdedness information, presented as a percentage relative to the maximum occupation ever registered in the station.	Free and open.	May be used to estimate attendance to events in a fixed location in London.	Amber Crowdedness information is provided for underground stations only and in percentage form. This source could only be used for London events in proximity to underground stations and in combination with other data sources.
National Highways WebTRIS	Traffic data for England's highways.	English highways. Updated every 15 minutes.	Free and open. Service will stop in early- to mid-2024 for a 6-month maintenance.	Could be used to determine attendance to an event by car.	Amber Only usable for rural events. Difficult to link these road types with specific events and spaces. Service will be temporarily stopped in 2024. Some roads suffer from very poor data collection.
Glasgow CCTV Object Detection Counts	Level of use of a geographic space in a specified time span.	Hourly. Selected locations in Glasgow.	Free access through API. Owned by University of Glasgow	Could be used as a proxy to determine attendance to events.	Amber Data currently restricted to selection of cameras in Glasgow which limits scope. There is potential to extend the sample of cameras for this project, but it requires agreements with camera operators.
DfT Road Traffic Statistics	Baseline road traffic statistics for motorways, the 'A' road network and some minor roads.	Street-level. Daily sample collected on a yearly basis.	Free and open.	Could be used to estimate baseline road usage to determine attendance to an event by car.	Red Very difficult to link these road types with specific events and spaces. A 1-day sample may not be representative enough for visitation baseline data.
Waze for Cities	Traffic data, including alerts, jams and unusual traffic (irregularities).	Road-level. Real-time, with 1 year of historical data.	Very restricted access, requiring sharing road data back to Waze.	May be used to determine attendance to an event.	Red Does not provide raw vehicle counts, but rather a qualitative estimation of traffic, limiting its usefulness. Access is very restricted.
Uber Movement	Aggregated data on Uber trips.	Unclear.	Discontinued since October 2023.	Could have been used to determine attendance at an event. Could have been used to determine demographics of attendees based on their origin of travel.	Red Discontinued.
Tier Data Sharing	Information about how Tier vehicles (e.g. e-bikes, e-scooters) are being used for transportation, including live vehicles data.	Real-time. Individual Tier vehicles.	Unavailable to this project – only provided to city councils where they operate.	Could have been used to determine demographics of attendees based on their origin of travel.	Red Unavailable to this project and very small user base compared to other methods of transport.

Lime GBFS	Standardized data feed for shared mobility systems, including location and status of scooters and bikes.	Real-time. Individual Lime vehicles.	Unclear – only provided to "municipal transit operators".	Could have been used to determine demographics of attendees based on their origin of travel.	**Red** Very small user base compared to other methods of transport.
Otonomo	Real-time data for traffic and parking.	Real-time. Individual vehicles and streets in major UK cities.	Unknown - sample data for assessment requested but no response from provider.	Could have been used to determine attendance to an event.	**Red** Unknown sample size of vehicles. Unknown usefulness of data due to no response from data provider meaning this project unable to assess potential use or viability.
AppyWay	Parking and kerbside dataset.	Kerbside level.	No longer pursued after call held with representative clarified that they do not have real-time parking tracking data, but rather focus on line markings and road contours for parking purposes.	Data collected means not possible to use for estimating attendance.	**Red** Does not provide data that would support this project.
Office of Rail and Road's ODM	Journeys recorded between all pairs of mainline stations in financial year 2021-2022.	Static. UK mainline stations.	Free and open.	May be used as a baseline for travel to and from locations.	**Red** Baseline passenger numbers were negatively affected by COVID-19 and are not representative of current travel trends. Only containing mainline stations limits its application.
Skyscanner API	Traveller search demand data for flights.	Daily. Individual airports.	Data is accessible only through a partner program.	Could be used as a proxy to determine attendance to events and origin of attendees.	**Red** Difficult to correlate interest in travel with attendance at events. Transport within the UK is primarily done with other methods of transport.
Bus Open Data Service	Collation of data feeds mandatory for all bus operators in England. An optional field is 'crowdedness', which tracks how many people are on a given bus.	Information on passenger density ('crowdedness') is only available for Go-Ahead East Yorkshire buses. Real-time.	Free and open.	Could be used to determine attendance to events.	**Red** Very limited availability. Individual bus services do not share information on passenger density ('crowdedness') data outside of this service either.
Network Rail Daily Concourse Footfall	A daily file containing the people movement counts in and out of the concourse area of Network Rail managed rail stations.	Daily. Individual Network Rail managed stations.	Free and open.	Tested to determine origin of UK visitors to an event with no success.	**Red** Only containing mainline stations limits its application. Not possible to correlate travel patterns with events.
Via+Citymapper	Citymapper aggregated user data, including demand patterns and bus stop usage. Includes journey origins for a given place.	No details given by provider.	Unavailable: company would not like to participate in the project after being contacted.	Could have been used to determine attendance at events. Could have been used to determine demographics of attendees based on their origin of travel.	**Red** Unavailable to this project.
Fluctuo City Dive	Collated data of shared mobility services, including origin/destination.	Daily. 0.1 km2 hexagons.	Unknown.	Could be used to determine demographics of attendees based on their origin of travel.	**Red** Extremely limited sample size means it's not usable for analysis.

verian

Data Source	Description	Granularity and Frequency	Access	Attendance and Participation	Prioritisation for Depth Phase
Pulsar	Mining of social media streams (Twitter/Facebook/Instagram/Threads/Tiktok) for mentions, as well as blogs, forums and Reddit. Event-related keywords have to be identified in advance.	Since this source can combine content data from several platforms, it can reflect various sources of data thereby allowing for more diversity & volumes of data and a richer overall granularity.	Costs are based on the desired contract length and desired volume of 'mentions' per month. As an indication, this project has negotiated a bespoke agreement with Pulsar for a 6 month contract allowing access to up to 200,000 'mentions' per month. Pulsar usually insist on a 12 month contract which was beyond the scope of this project but could be desirable for a future user.	Event-related keywords must be selected in advance, and evidence of participation depends on the success/quality of the search inputted. Access is through the Pulsar dashboard. Audience segmentation can be provided through another tool available from Pulsar.	**Green** Faculty.AI have an existing contract with Pulsar (as do colleagues in DCMS).
X (formerly known as Twitter) - Mentions	Mining of Twitter streaming API for mentions of keywords. Those mentions can be further classified into those that are more likely to be event-specific, either manually or through automatic classifiers	Streams of tweets can be accessed in real-time. Twitter users can mention their hometown in their handle. Any further identification of demographics would require suitable classifiers.	Requires expensive Twitter API access. Access to the Twitter API starts at $5,000 per month for 10,000 mentions per month. This cost is prohibitive for the scope of this project, but also does not represent value for money when compared to more affordable aggregators such as Pulsar.	Much less accurate than geolocated tweets.	**Amber** The alternative of using a commercial aggregation service such as Pulsar is lower risk, avoiding the handling of the complexities and costs of Twitter API access.
Flickr	Public photo sharing site with geotagged photographs available.	Granularity is very low – at best, access over yearly timeframes could be captured.	API (public data accessible by tag or location).	Given the low granularity of the data, it is not very accurate for estimating attendance.	**Amber** Unlikely to be useful for sufficient use cases.
Facebook check-ins / tagged mentions	Mining of several check-ins over time for selected Facebook Pages using Facebook API.	Counts would need to be frequently accessed in order to gain understanding of access.	Permission would need to be sought from Facebook for API credentials with such privileges.	Check-ins can help to measure attendance; tagged mentions might only be virtual attendance.	**Red** The alternative of using a commercial aggregation service such as Pulsar is more operationally feasible and with a lower risk of failing to access data.
Twitter / X - Geolocations	Mining of Twitter streaming API for the geolocations of tweets – can be used to identify density at locations.	Live streams of tweets in particular locations could be obtained.	Requires an expensive Twitter API access. Twitter access directly is $5000/month, for just 10k tweets maximum: https://developer.x.com/en/docs/twitter-api/getting-started/about-twitter-api	In the past, Twitter data was able to show people tweeting within geographical areas. From June 2019 Twitter users have not been able to tag their exact location. Twitter removed this functionality over privacy concerns and to simplify the user experience of tweeting.	**Red** Accurate geolocations are seldom recorded in Tweets anymore. While deploying a geolocation prediction process is possible, it adds both complexity and uncertainty.

Data Source	Description	Granularity and Frequency	Access	Attendance and Participation	Prioritisation for Depth Phase
Drones	Drones can be used to capture images of the entire area of an outdoor event (e.g., music festivals). ML models can then be used to provide counts.	Can count individual attendees.	Can only be deployed for outdoor events and would require collaboration of events' organisers. Adverse weather conditions may hinder the deployment of the data source. Agreement to share this data, if required, is rarely straightforward.	N/A	**Green** The accuracy of the counts will be highly dependent on how well the pictures capture the entire crowd, but initial testing and literature show promise.
Conventional Cameras	Conventional cameras can be used to capture images of the crowd at an event. ML models can then be used to provide counts.	Can count individual attendees.	Would require collaboration of events' organisers but can be used for both indoor and outdoor events. Agreement to share this data, if required, is rarely straightforward.	N/A	**Green** The accuracy of the counts will be highly dependent on how well the pictures capture the entire crowd, but initial testing and literature show promise.
CCTV **(based on UBDC CCTV Object Detection)**	Fixed, 24/7 cameras can be used to track attendance over time to a given space. ML models can then be used to provide counts.	Can count individual attendees.	Would require collaboration of events' organisers or CCTV provider. Agreement to share this data, if required, is rarely straightforward.	N/A	**Green** Highly dependent on the availability and positioning of CCTV cameras.
Radio Sensors **(including wifi e.g. CrowdScan)**	Radio sensors can be deployed throughout an event space to track attendance. The quality of the signal can give an indication of the number of people present.	Measures are taken at regular intervals. Although it attempts to count individual attendees, there are concerns with accuracy.	Would need to be set up in a bespoke manner for each event, including training it without attendees to set a baseline for the counting model.		**Red** Potential high cost and complexity of setting up the sensors. Accuracy is not very reliable.

verian

Data Source	Description	Granularity and Frequency	Access	Attendance and Participation	Prioritisation for Depth Phase
OpenStreetMap	Geographic database, including roads and points of interest.	UK-wide, street-level. Database updated daily.	Open and free. Can be accessed through multiple APIs, including Overpass API.	N/A	**Green** Can be used to obtain relevant Points of Interest (POIs) like sports or cultural spaces to track participation.
Ordnance Survey Points of Interest	Geographic database including points of interest by category.	UK-wide, per point of interest. Updated quarterly.	Accessible through an Ordnance Survey partner. Released as a PDT file.	N/A	**Green** Can be used to obtain relevant POIs to track participation.
England & Wales Census	Census 2021 for England and Wales.	Output areas of 40-250 households.	Open and free. Can be accessed through multiple APIs, including Overpass API.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.
Scotland Census	Census 2022 for Scotland.	Output areas of 50 households.	Open and free Staggered release of data through 2023-2024.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.
Northern Ireland Census	Census 2021 for Northern Ireland.	Variable (3,780 Data Zone outputs of different size).	Open and free. Can be accessed through multiple APIs, including Overpass API.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.
Indices of Multiple Deprivation (IMD)	IMD data for all UK nations is available from the Consumer Data Research Centre.	Variable depending on UK nation, detailed geographic data zones in Scotland and lower super output area level in England.	Open and free.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.
CACI's Acorn	Geo-demographic segmentation.	Segments for UK postcodes, 7 categories, 22 groups and 65 types.	Public sector version available, pay for access.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.
Experian's Mosaic	Geo-demographic segmentation.	Segments for UK postcodes, 15 summary groups, 66 detailed types.	Central government version available, pay for access.	Can be linked with other data sources to retrieve demographic information.	**Green** Can be used to retrieve demographics of attendees and participants.
TfL GIS Layers	Collection of ArcGIS maps created by TfL, including bus routes and stations.	London, street-level.	Open Downloading the maps requires an ArcGIS account.	N/A	**Red** Does not provide information relevant to crowd counting, and other information is obtainable in other, easier to access sources (e.g. TfL API).
ParkRun	A recording of persons taking part in weekly running events.	Weekly	Sport England have Parkrun data and may be a potential way of accessing it. Parkrun themselves no longer offer an open access API, but most sites do publish basic weekly finisher information.	The number of attendees at each running event is available and can act as a ground truth source. Demographics (gender, age group, event location) are available.	**Red.** Access through Parkrun directly is no longer available and using Parkrun data on its own is unlikely to be useful given the small numbers of attendees at individual sites and the fact it is limited to running events. It should be noted, however, that parkrun data is '**green**' rated for use as baseline information to support our Strava modelling approach. Here, basic published information is sufficient to make this data source useful.
Meetup	Used by large numbers of events and can be used to discover events.	Per-event	Use of the Meetup API for commercial use requires approval from Meetup.	The number of tickets and attendees is available for each event and can be used as a ground truth source.	**Red** Documentation is sufficiently vague to make it unclear if Meetup would be suitable either for estimating attendance at events generally, or for cultural/sporting activities too.

Evenbrite	Used by large numbers of events and can be used to discover events.	Per-event	Searchable API was removed in 2019.	N/A	**Red** No longer available
Facebook Events	Used by some types of events for increasing interest.	Per-event	Facebook events API was made non-public in 2019.	N/A	**Red** No longer available