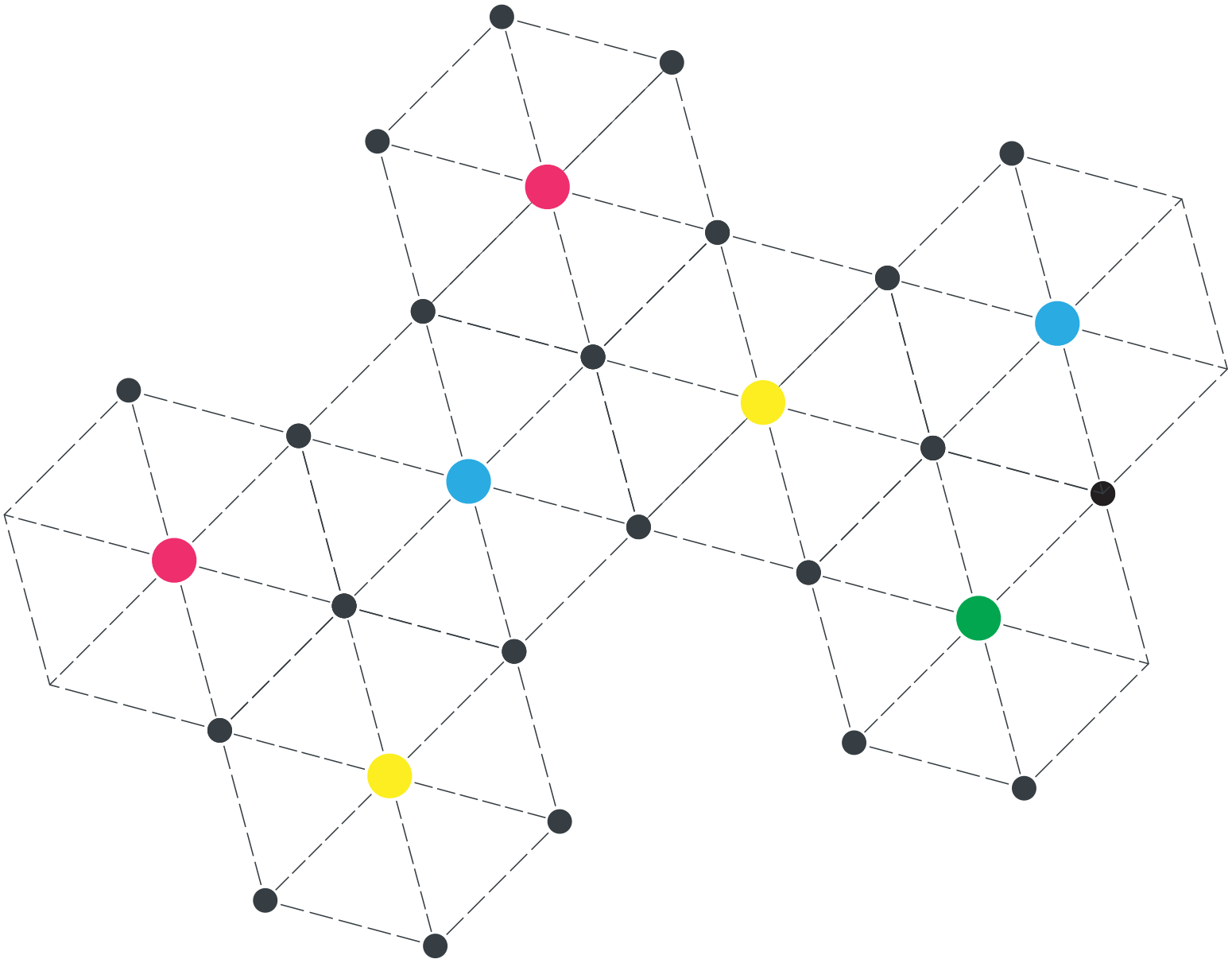
MAYOR OF LONDON



Piloting the London Office of Data Analytics

FEBRUARY 2018

nesta

About the GLA
The Greater London Authority is the strategic regional government for London. Through the Mayor and the London Assembly the GLA is responsible for delivering strategic policy across a full range of policy areas from housing and transport to planning and infrastructure. The GLA also has an important convening role in facilitating collaborative working among the 33 London Boroughs.

About Nesta
Nesta is a global innovation foundation. We back new ideas to tackle the big challenges of our time. We use our knowledge, networks, funding and skills - working in partnership with others, including governments, businesses and charities. We are a UK charity but work all over the world, supported by a financial endowment. To find out more visit www.nesta.org.uk

Authors
This short report has been written by Nevena Dragicevic and Eddie Copeland from Nesta's Government Innovation Team, and Andrew Collinge, Paul Hodgson, Wil Tonkiss and Alan Lewis at the GLA.

Acknowledgements
The GLA and Nesta wish to thank the LODA pilot boroughs - Barking and Dagenham, Brent, Bexley, Camden, Hackney, Islington, Kingston, Lambeth, Lewisham, Sutton, Waltham Forest and Westminster - for all their contributions and efforts throughout the project. We would also like to thank ASI Data Science for its services and guidance.

CONTENTS

Introduction

What would happen if London could source, analyse and act upon its public sector data at a city scale?

This question formed the basis of a year long pilot of a London Office of Data Analytics (LODA), a collaboration between the GLA, Nesta, twelve London boroughs and ASI, a data science firm. In this report, we outline the pilot's origins, methods and what we have learned to date.

The impetus for piloting LODA began with a recognition that on many issues, London's public sector data is like a jigsaw that has never been put together. Every team has their little piece of the puzzle, but no one has the ability to put those pieces together, take a step back and see the big picture. Given the current pressure on public services, that fragmentation is a serious problem as it hinders many of the tried and tested ways of delivering more and better with less. How can boroughs intelligently design shared services if they don't have data on the scale of the problem, demand or opportunity beyond their boundaries? How can they coordinate the actions of different teams if those teams don't have data on what each other is doing? How can they target resources at areas of greatest need if they lack the data on where that need lies? And how can they predict and prevent problems from occurring if they don't have the data that could collectively point to cases of highest risk?

"There are estimated to be up to 15,000 HMOs in some London boroughs[1], yet only 10-20% are believed to be correctly licensed."

Taking inspiration from the Mayor's Office of Data Analytics (MODA) in New York City, the aim of LODA has been to overcome precisely these challenges by putting in place the technical, data and organisational resources to apply data science at a cross-borough level. To test the concept and see how it might work in the very different political and administrative setting of London, the pilot focused on identifying unlicensed HMOs – houses in multiple occupation. There are estimated to be up to 15,000 HMOs in some London boroughs[1], yet only 10-20% are believed to be correctly licensed. Could data help local authority building inspectors find more of these properties?

1 Local Authority Housing Statistics data returns, England 2014-15, DCLG

About this report

This report highlights both the potential benefits of being able to use data at a truly pan-London scale, but also the many barriers that stand in the way of realising that vision, and what it would take to overcome them. The aim of this report, therefore, is to:

- Provide a detailed account of Phase 1 of the LODA pilot which brought together 12 boroughs to develop and test a machine-learning model to find unlicensed HMOs;
- Identify the challenges and barriers to delivery when undertaking a pan-London data-driven project of this type;
- Identify a series of key recommendations which will be used to inform future LODA projects, including the continued development of the HMO model;
- Set out the next steps for LODA including a review of Phases 2 and 3 of the pilot programme. Phase 2 saw the Borough of Barking and Dagenham (one of our partner boroughs) continue to iterate on the model developed in Phase 1, with promising early results. Phase 3, led by the GLA, is currently underway to build on the analysis and lessons learned in earlier phases to create and test a third iteration of a model to identify unlicensed HMO properties in London.
- Present a roadmap for the establishment of a permanent LODA within the Intelligence Unit at the GLA.

We share these findings both to consolidate our own learning about the future shape of LODA, but also in the hope that they will help other cities in their own attempts to make smarter use of data to deliver better, more responsive and effective public services.

LODA Pilot Boroughs

LODA pilot aims

The overarching aim of the LODA pilot was test whether there is value in multiple public sector bodies collaborating with their data to tackle a public service challenge in the capital.

The three key objectives for the pilot were to:

1. Test the hypothesis that if London's boroughs and public sector bodies share and analyse their combined data, services can be improved in ways that would not be possible if each acted alone;

2. Determine whether the Office of Data Analytics (ODA) methodology can be adapted to work in London, and if so, under what conditions;

3. Help inform the structure, operating model and requirements of a permanent London Office of Data Analytics.

Choosing an issue to tackle

A crucial step for the pilot was identifying a public service challenge where data analytics might offer new and actionable insights. Following an initial presentation about LODA to the London Borough Data Partnership, held at Nesta on 12 April 2016, boroughs interested in taking part were invited to submit suggestions for potential challenge areas. Twenty ideas were crowdsourced. This list was reviewed by Nesta and the GLA to create a short-list of the six most promising suggestions.

These six challenge areas became the focus of a workshop held with fifteen London boroughs on 21 June 2016. The workshop began with a presentation by Mike Flowers, Chief Analytics Officer at Enigma, and the creator of the New York MODA model[2].

The workshop aimed to help the London boroughs understand the principles on which the LODA model might work; think through the six suggested challenge areas and flesh out the details of each; and identify which ideas had the greatest potential for the pilot. Split into groups, the participants were taken through a series of rapid exercises to explore each of the challenge areas. The boroughs were asked to assess and score each one according to the extent that it would be likely to:

- Save significant money;
- Have good data available;
- Lead to actionable insights that could deliver results in around 2 months; and
- Be tackled mostly with non-personal data.

2 The advice Mike Flowers provided is outlined in the following article: http://www.nesta.org.uk/blog/three-lessons-city-data-analytics-mike-flowers.

The six shortlisted challenge areas were:

HOUSING

Identify houses not registered as a House in Multiple Occupation (HMO) so that correct charges can be made or fines issued.

EDUCATION

Optimise routes / provision of Special Educational Needs (SEN) transport to schools.

SOCIAL CARE

Join up records from across local authority boundaries to ensure that Troubled Families are identified.

PROCUREMENT

Collate demand for specific goods and services across public sector bodies to enable bulk procurement / reveal where the same goods are being bought for the lowest price across the city.

HEALTH AND WELLBEING

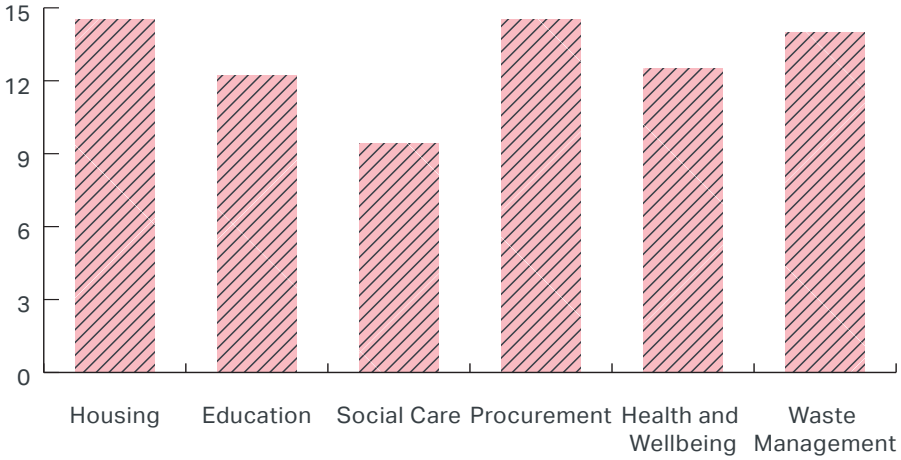
Improve public health by overlaying datasets concerning patterns of obesity / green space / fast food outlets.

WASTE MANAGEMENT

Identify levels of recycling across London to target interventions that increase recycling rates to avoid landfill costs and meet local authority targets.

The average aggregate scores for each challenge were then used to determine which would be taken forward. There were three preferred challenges. This was mostly due to the availability of non-personal datasets relating to each and their potential to have a defined and measurable impact within the short timescale of the pilot.

Final average scores for the 6 potential challenge areas



Following a further consultation with a core local authority steering group, the housing project - identifying unlicensed HMOs - was chosen as the most viable challenge for the first pilot. However, several of the other challenge areas may hold promise and could become future LODA projects, or form the basis for additional LODA pilot schemes[3].

3 A detailed report on the Data Analytics Challenge Workshop, produced by Nesta and the GLA, can be downloaded from the London Datastore at: https://data.london.gov.uk/dataset/london-office-of-data-analytics.

Defining the problem - unlicensed HMOs

The challenge selected concerned unlicensed Houses in Multiple Occupation (HMOs). HMOs are properties that are three or more storeys high, occupied by five or more people forming two or more households, and where the occupants are unrelated and share bathroom, toilet and kitchen facilities. (In spring 2018, new regulation will extend the definition of a mandatory HMO licence to include any property meeting this criteria, irrespective of the floor levels)[4]. Storeys counted can include commercial units at ground floor level, e.g. shops, habitable basements used as living accommodation where amenities are shared, and attics that are occupied.

Landlords who own such a property are required to have a specific licence. Those licences are important for two key reasons. First, unlicensed HMOs are linked to some of the most dangerous and exploitative living conditions in the capital. Second, HMO licence fees are used to locate more HMOs, resource prosecutions against rogue landlords, and enforce compliance. Any properties they fail to license therefore represent missed revenue that could be used to improve standards in the private rented sector. There are estimated to be up to 15,000 HMOs in some London boroughs, yet only 10-20% are believed to be currently licensed[5].

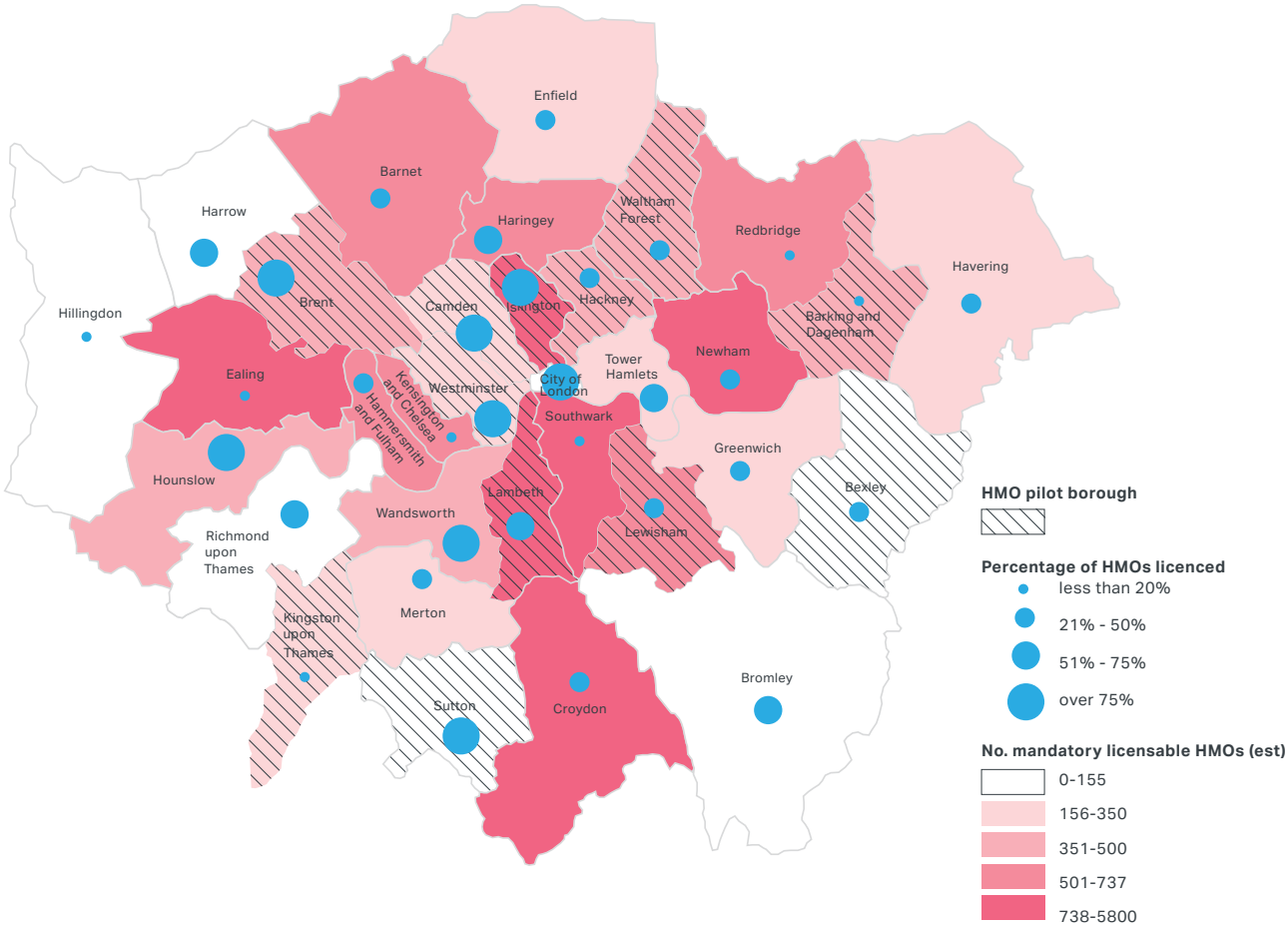
4 “House in multiple occupation licence” https://www.gov.uk/house-in-multiple-occupation-licence
5 Local Authority Housing Statistics data returns, England 2014-15, DCLG

The number of borough housing officers available for private sector housing enforcement varies significantly across London. Some boroughs, with large teams, are able to be proactive in identifying and acting on unlicensed properties. In other boroughs, resources are so restricted that efforts are concentrated to reactive inspections in response to reported rogue landlord activity, and monitoring HMOs subject to mandatory licensing. The desired outcome for the pilot was therefore to see if data could be used to create a predictive model that identified likely unlicensed HMOs for proactive inspections.

Doing so was expected to deliver both short- and long-term benefits. In the short-term, a predictive model could result in the improvement of living conditions for Londoners, increased revenue for boroughs, and raise the efficiency of inspections (performance varies widely across the capital). Longer-term, there are a range of secondary impacts which could result. Improvements in housing can have benefits on the health and well-being of individuals, thus reducing the burden on the health service. There are social benefits such as reduced anti-social behaviour and instances of fly-tipping associated with improvements to housing stock. Also, through engagement with tenants, vulnerable people who may not know about, or have access to, the local services that they need could be provided with additional support.

“There are social benefits such as reduced anti-social behaviour and instances of fly-tipping associated with improvements to housing stock.”

Distribution of known and estimated HMOs



Source: DCLG - Local_Authority Housing Statistics data returns 2015 to 2016

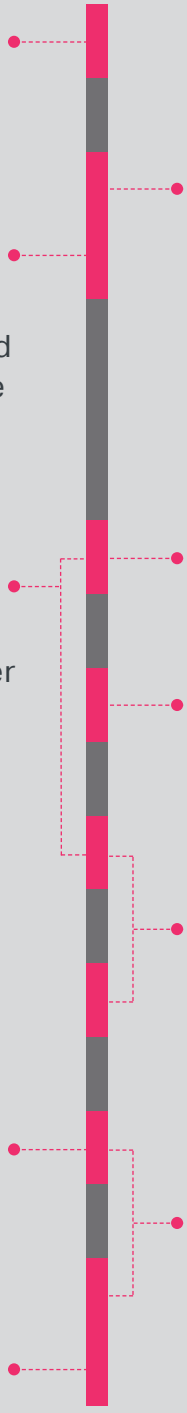
Appointing data scientists

It was determined that, for the purpose of the pilot, the development and implementation of the HMO model algorithm should be undertaken by an external data science consultancy. This decision was made based on a recognition that the capacity of boroughs' in-house data analysts is extremely limited. Most are either already fully assigned with the requirements of their day job, or would be unable to commit to the full length of the pilot.

Proposals were therefore sought from several data science SMEs. From those, ASI Data Science was selected based on the company's demonstration of having a strong understanding of what the pilot was aiming to achieve. The role of ASI in developing and defining the project was significant and their work with the boroughs in the initial stages was valuable in ensuring engagement and understanding in the data process. Similarly, their work to maintain momentum in the development and implementation of the model (covered in detail below) was key to ensuring that the pilot was able to reach the field test stage. One of the key outcomes from the pilot has been a greater understanding from all involved about the cultural barriers to public-private sector partnerships in the data science domain.

Pilot timelines

The key steps involved conducting the pilot are summarised below.



April 2016
Consultation
Ideas for the LODA pilot discussed at meeting of the Borough Data Partnership

July 2016
Start of pilot
ASI Data Science commissioned as data science partners for the project

November 2016 - March 2017
Data Collection
Remaining pilot boroughs gather data to supply to ASI

July 2017
Evaluation of Phase 1

October 2017
Phase 3
GLA begin work on a third iteration of the HMO model

June 2016
Problem Selection Workshop
15 boroughs attend a workshop to determine the challenge to be tackled by the pilot - finding unlicensed HMOs

November 2016
Proof of Concept
First prototype of the predictive model is developed with Westminster's data

January 2017
Information Sharing Protocol Finalised

March 2017 - May 2017
Testing
Boroughs test prioritised inspections lists generated by the predictive algorithm with housing teams.

July 2017 - September 2017
Phase 2
Barking and Dagenham work on a second iteration of the HMO model.

Six step methodology

This section describes how we assessed the viability of using data analytics to find unlicensed HMOs, and how the predictive algorithm was designed and tested in the real world.

Step 1: Interrogating HMO features
Once the challenge area was confirmed, ASI Data Science worked alongside the GLA and Nesta in a series of visits to borough housing and data teams. We began by engaging two boroughs, Westminster and Lambeth, to help us better understand the HMO licensing challenge and define the problem in a way that could be translated into a data science exercise. We started with just two boroughs as it was deemed impractical to understand the processes of 12 separate local authorities simultaneously.

The two boroughs worked with frontline workers and other stakeholders to investigate the process of HMO detection and interrogate the likely features of these properties, which could include factors such as the height of the building, its age, and whether it is located above a commercial premise. As with many frontline workers, housing inspectors can provide a long list of risk criteria, honed over many years of experience.

Step 2: Identifying relevant datasets
ASI Data Science engaged with data analysts at Lambeth and Westminster to identify datasets held by the boroughs that related to the criteria suggested by building inspectors in Step 1, and explore the type and structure of that data. These included physical property features as well as records on anti-social behaviour, noise complaints, council tax bands, housing benefits recipients, and improper waste disposal, among many others. At this point, however, only Westminster was able to proceed with supplying the relevant data: 40 datasets in total.

Step 3: Analysing the problem from a machine learning perspective

Following preliminary analysis on Westminster's data, ASI Data Science determined that the HMO challenge shared much in common with financial fraud detection, given the rarity of such properties. Both issues are classic 'needle in a haystack' problems.

A complicating factor was that the HMO problem was only 'half-labelled', meaning the data showed properties that definitely were HMOs, but not those which were 'definitely not HMOs.' In the case of financial fraud, historical transactions data would be used to determine which purchases were 'fraudulent' or 'not fraudulent', as people would report having been victims of fraud. For HMOs however, while boroughs know which properties in their jurisdiction are 'definitely HMOs', few hold data on those that are 'definitely not' (i.e. properties that are owner-occupied or in ineligible building types), leaving us with a huge range of 'probably not HMOs'.

Based on this analysis, an adapted balanced random forest method, often used in anomaly detection and in cases where data is half-labelled, was selected to train the data and build the machine learning model.

Step 4: Creating the first prototype
Armed with a good amount of data and a better understanding of the problem from a data science perspective, it was time to test the plausibility of a predictive algorithm for HMOs. Initial results were encouraging. ASI identified two metrics by which the operation of the model should be judged: does the model identify at least 50% of know HMOs and, is the total number of HMOs identified between 1-4% of the authority's total housing stock. The second metric is important because HMOs will only ever account for a small proportion of the total houses in an authority and so a relatively low number of positive hits is an early indication that the model is able to distinguish the characteristics of HMOs from the wider stock.

The prototype model had a 50% recognition rate of known HMOs and served up just 671 properties for inspection in a borough of over 200,000. Out of the 40 features Westminster had shared, just four accounted for 90% of the predictive power of the model.

Step 5: Expanding the model - adding data from other boroughs
We had initially assumed that once a model was developed for the first borough, the rest would have to provide identical datasets in the same format. Yet, given the extreme rarity of known HMOs, combined with variations in building stock and other local features, bespoke models were instead developed to avoid over-fitting and capture important differences across boroughs. As a result, boroughs were encouraged to share any data held on all properties that could potentially correlate with unlicensed HMOs.

Most boroughs' data required substantial processing, cleaning and merging. Major limitations in the data included a lack of separation between flats and houses, no distinction between properties in the private rental sector and owner-occupied homes, and difficulty in matching the Unique Property Reference Number (UPRN - the unique identifier we used to match and merge data) to addresses.

Step 6: Testing and evaluation
As described in Step 3, the data for the HMO pilot was only 'half-labelled'. This not only made the problem more difficult, but also meant that a full cross-validation could not be performed. Cross-validation is a technique used in machine learning to estimate the accuracy of a model. Instead, it was necessary to perform a Randomised Control Trial (RCT) to obtain statistically significant results.

Designing an RCT with multiple boroughs, all of which follow different inspection procedures, was challenging. Some conducted only reactive inspections, responding solely to tip-offs and finding HMOs mostly by chance, while others employed more proactive and targeted measures. Some boroughs carry out fewer than 30 inspections per month, while others inspect more than 100 properties.

By design, RCTs require a 'control' (i.e. properties referred by an existing process) and 'treatment' (i.e. properties referred by algorithm) group. This meant that, in some cases, boroughs had to double the number of inspections to visit a property from each list. Good RCTs also aim to minimise any variables that could influence results, such as seasonal variance or the level of expertise of individual inspectors. In our case, housing managers were asked to conceal the source of the referral from inspectors to eliminate any bias. Once the trial began, however, these requirements proved too onerous to enforce consistently, and boroughs shifted to validating whether properties from the algorithm list were licensable HMOs.

The results of our experiment and other outcomes of the LODA pilot are discussed in the next section.

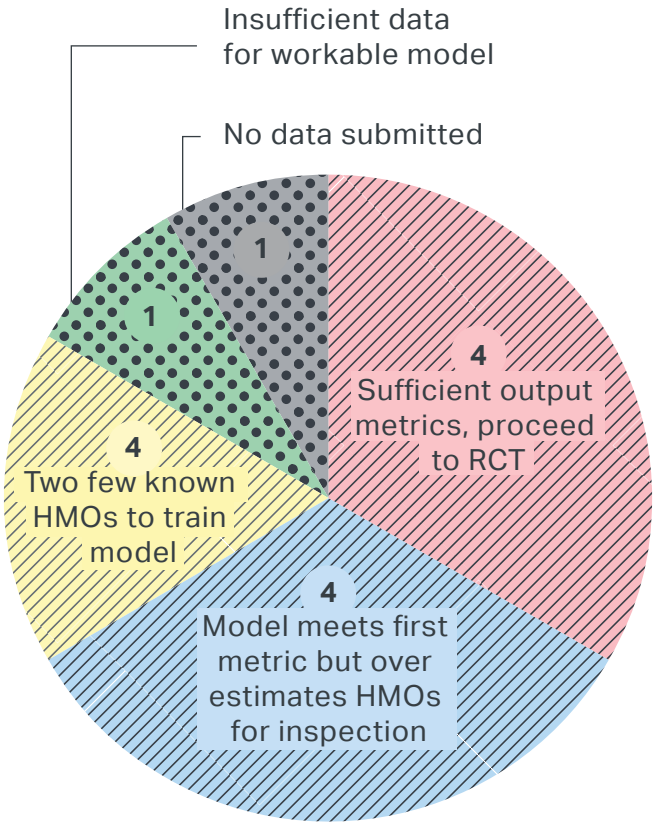
Results

The complex nature of the HMO problem, coupled with significant data challenges, were ultimately too intractable to produce successful predictions in this first phase. Overall, just four of 12 pilot boroughs provided data which generated two output metrics sufficiently high to continue with the RCT. These metrics were 1) how good the algorithm is at recognising known HMOs, and 2) what fraction of a borough's properties it recommends for inspection.

Clearly, if the model only recognises 5% of known HMOs (i.e. the first metric), this would not inspire confidence. The point at which the recognition rate is deemed sufficiently high to proceed is not a hard and fast rule, and a layer of human judgement is required. Similarly, if the algorithm recommends 20% of properties for inspection (i.e. the second metric), it has significantly shrunk the search space but inspecting a fifth of all properties in a borough is not practically actionable or sufficiently powerful to run an RCT. Again, human judgement is required, but ideally the algorithm would recommend 1% of the borough's properties for inspection. Since Westminster was the only borough to meet this criterion, the number was relaxed to allow more councils to participate in the RCT.

Chart: Breakdown of LODA model outcomes for 12 boroughs (figures refer to no. of boroughs)

Metric 1: The model recognises at least 50% of known HMOs
Metric 2: The model flags 1-4% of properties for inspection



Another four boroughs also supplied reasonably good data, but while the machine learning model recognised at least half of known HMO properties, it flagged too many properties for inspection. This could be for a number of reasons. For example, the homogenous housing stock in more suburban boroughs could make it difficult to isolate unique features of HMOs. In all, no licensable HMOs were found.

Despite the null results, conducting the LODA pilot has led to a number of positive outcomes and important discoveries about the nature of joining up and sharing data across local authorities. Over a couple of months, a core group of five boroughs co-produced an information sharing protocol to responsibly share data between 12 boroughs and ASI. This document could serve as a template for future data sharing projects between multiple local authorities and third parties. The cross-cutting nature of the HMO problem has also created more opportunity for collaboration and data sharing within organisations, and identified important gaps and weaknesses in current systems.

The participating boroughs also reported finding value in taking part in the process and having the chance to work collaboratively with their peers inside their own organisations and in other local authorities. A small sample of their feedback has been:

"...conducting the LODA pilot has led to a number of positive outcomes and important discoveries about the nature of joining up and sharing data across local authorities."

"We have more buy-in from other teams willing to share data"
LB Housing Manager

"This felt like real collaboration with partners, I feel I have grown our network with GLA and our own public health and GIS teams"
LB Housing Manager

"We are now embarking on work to address some of the key weaknesses in how we manage and use data. I will use the LODA pilot as an example to justify this."
LB Data Analyst

"Hopefully, the pilot will spur [another iteration] in the future as I really do think numerous boroughs working together would be better than boroughs working in isolation – borough boundaries are simply arbitrary."
LB Project Manager

The value in undertaking a pilot programme is identifying where the barriers to success lie and finding ways to mitigate them for future projects and future iterations of the HMO model. In the sections that follow, we unpack specific challenges and make a number of suggestions for how to improve predictions. We also share our key lessons learned and what that could mean for the future of LODA and, more generally, for innovating with data in the public sector.

Challenges

This section outlines the key challenges we encountered in conducting the LODA pilot. Most difficulty was experienced in the areas of data acquisition and processing, while other technical and capacity issues also added to the project's complexity.

- **Data quality:** Data submitted by most boroughs required significant cleaning, processing and merging. In particular, accurately linking different types of housing and property data to a unique identifier (the UPRN) was one of the biggest challenges. Property related data held in different systems was referenced in various ways, including with UPRNs, but also by address (often incomplete or inconsistent across systems), or by northings/eastings. This made it very time consuming to match various data held on a single property, and influenced the quality and quantity of data individual boroughs were able to provide, which ultimately limited the quality of analysis.
- **Data availability:** Data on private rental sector properties, which could have helped filter out owner-occupied and other ineligible property types, was a critical missing piece of the puzzle. Additionally, commercial data on physical property features (i.e. height of buildings) made only a modest improvement in the modelling, due to incompatible formatting.

- **Data warehousing:** Boroughs with centralised business intelligence teams and data warehouses had an easier time pulling together data from across the organisation, while others were often met with long delays in collecting data held by different departments.
- **Known HMOs:** Known licensed HMOs turned out to be even rarer in some places than we had thought. In a couple of boroughs, despite the excellent quality of data, a lack of known HMOs (as few as 30 in one borough) meant the model had too few cases to train on to reliably predict other HMOs.
- **Precise requirements:** As mentioned, boroughs were free to supply data on the housing features they deemed relevant for their area. In some cases, this made the process more challenging for boroughs, potentially making it harder to explain to colleagues what datasets they required. Precise requirements could have helped boroughs prioritise certain datasets, especially in boroughs with less capacity to work on the pilot. That said, due to the variation in local characteristics, certain types of data may not have been sufficient or relevant across all local authorities.

  Having completed the pilot, ASI suggest that, as a starting point and keeping the local context in mind, the ideal datasets would be those highlighted on the following page.

What does the 'ideal' dataset look like?

- A list of all properties in the borough, listed by UPRN, but also by address (in a clean, standardised format), as well as latitude/longitude, and by northings/eastings.
- A way to distinguish properties between commercial, owner occupied, privately rented and social housing.
- Council tax: for each property, information on which council tax band it is in, whether the property has been late on payments, whether the bill is sent to a different address or to the property.
- Electoral register data: for each property, the number of occupants, number of surnames, number of changes made in last several years.
- The number of housing benefit claimants for all properties with at least one benefit claimant.
- Structural data: For each property, the number of stories in the building, age of the building.
- Complaints data: Complaints data must be in a form in which it is attributable to a property or if not possible, to an ONS Output Area. This is of course difficult in cases such as fly-tipping. Noise complaints are of particular interest however.

- **In-house expertise:** The range of technical expertise available in-house - and to support this particular pilot - varied across boroughs. For example, in one case, a borough stated that it would have had to contract a third party supplier to extract data related to its housing benefits, and would be charged for doing so.
- **Staff capacity:** From legal advice on data sharing, to data collection across the organisational boundaries, input on evaluation design, and increased inspections, the pilot required a larger number of staff and resources to implement than originally anticipated. In a couple of cases, key staff that had been spearheading the project inside a borough had also moved on to different posts, leaving significant gaps in organisational capacity. As the project increased in complexity, our data science partner was equally challenged to provide on-going and in-depth guidance to 12 boroughs.

Lessons learned and recommendations

As well as identifying the set of challenges above, the pilot has surfaced a number of important lessons that will be fundamental to the design of a future London Office of Data Analytics. Those lessons, and the recommendations we derive from them below, will be also be relevant to many other public sector data initiatives.

TECHNOLOGY

Lesson: Local authorities that do not have the ability to join up and match records held in different IT systems within their own organisation will find it extremely challenging to collaborate with other organisations with their data. It is simply too difficult and time consuming to conduct this process manually, especially during experiments where different datasets need to be explored.

Recommendation 1: **Public sector organisations should prioritise future IT investment in data matching tools that enable them to link records across their different IT systems**. Two groups of records need to be linked: those about places, and those about people.

Recommendation 2: Organisations that outsource any of their IT functions should ensure that supplier contracts allow them to access all their data. Currently, some suppliers place heavy restrictions or costs on accessing data outside day-to-day business needs. That arrangement is not acceptable as having only partial access to their own data limits organisations' ability to benefit from data insights.

Recommendation 3: Future initiatives that aim to harness public sector data from multiple sources should begin by conducting data maturity assessments of each participating organisation. Those assessments should focus on evaluating each organisation's technical readiness to ensure obstacles are identified early. Nesta, with support from the Local Government Association, is developing an interactive Data Maturity Framework tool, which will be available mid-2018.

DATA

Lesson: A lack of standardisation in frequently used field names (e.g. lines of an address) in different IT systems makes joining up data for analysis much more difficult than it needs to be.

Recommendation 4: For place-based data, public sector organisations should commit to using Unique Property Reference Numbers (UPRNs). Matching and merging records (as per Recommendation 1) can be made significantly easier if a standard common identifier is used across systems. Significant gains can be made to improve the usability of data at relatively little cost, as only an organisational commitment to recording data consistently is required.

Lesson: Where multiple partners are involved, the data acquisition process is complex and can take a long time. Being as precise as possible about data requirements, for example by developing a data schema or sharing sample datasets, can help minimise delays. However, given that identifying the ideal combination of datasets may take several iterations, it is always likely to be a slow process.

Recommendation 5: **Projects like LODA should allocate significant time to the data acquisition phase, possibly two to three times more than might be expected.** An essential prerequisite for running a successful data acquisition phase is for each organisation to offer a single point of contact for data requests.

Recommendation 6: In addition to using public sector data, datasets from national government, businesses, universities and third sector organisations should also be considered. These other sources of data can help fill important gaps and are often available in consistent formats across a wider geographic area than most public sector datasets.

PEOPLE

Lesson: Projects like LODA are more about getting people to collaborate in new ways across organisations than they are about doing new things with data and technology. Data projects can never just be delegated to data science teams; they must be organisation-wide efforts. In particular, frontline staff must be involved in the process. In the case of LODA, we found it was not enough to engage with building inspectors solely at the start of the pilot to interrogate the likely features of an HMO.

Recommendation 7: Data projects should aim to let data scientists, project managers and other team members work in the same physical space. This co-location, even if only done on an ad-hoc basis, can help promote better communication and learning between those with the data science knowledge, and those with experience of the service challenge being tackled. This is helpful for developing skills, and is also likely to speed up the execution of the project.

Recommendation 8: In-house data analysts should be given the opportunity to work with service managers to tackle public service challenges. Many organisations already have in-house staff who are skilled in the analysis of data, but their time is typically taken up with creating reports for monthly dashboards or reporting on Key Performance Indicators. These staff should be given the time and opportunity to work on higher-value activities for service improvements.

Recommendation 9: It is vital to spend time in the field with frontline staff to understand how their day-to-day operations work, and involve them in every step of the process. Frontline workers are likely to be the greatest source of expertise on a given public service challenge. Data has little value without local context; frontline staff can help validate findings, spot biases in the data, and errors in the output of algorithms.

Recommendation 10: Public sector leaders need to create a culture where it is unacceptable to make major decisions or try to reform a service without at least being aware of what the data says. An organisation's willingness and ability to engage in the use of data is led by the attitude of its leaders. In particular, leaders need to promote a more measured appetite to risk in the use of data, moving from a default mode of assuming that all data must be protected, to recognising that there are significant downsides to not sharing data in appropriate circumstances.

PROCESS

Lesson: Data analytics projects that are technically challenging, have changing data requirements, involve many partners and are culturally new require a more adaptive approach to project management to deal with higher levels of complexity and uncertainty.

Recommendation 11: Collaborative data projects should adopt an agile approach to project management, which builds in regular opportunities for analysis, testing and reflection. Such an approach will allow for better risk assessment, enable small scale testing, and challenge assumptions on an ongoing basis. This will help surface problems and solutions much earlier in the project process.

LEGAL

Lesson: Different organisations have varying levels of risk appetite for sharing the same type of data, even in cases where mostly non-personal and non-sensitive data is involved. While this is difficult to overcome, consistent legal advice and common interpretation of the same data legislation could provide assurance and speed up the process of data sharing.

Recommendation 12: A future London Office of Data Analytics should include a legal function that can provide consistent guidance on data sharing. Such a function would work with Information Governance teams in each participating organisation to identify appropriate legal gateways for sharing data, and ensure all data was handled in an ethical, legal and secure manner. It could also help create a single repository of data sharing agreements, privacy impact assessments and other documentation that could be made available for reuse in future data initiatives by other organisations.

Not the end of the story...

Phase 1 of the LODA pilot has produced a number of valuable insights and recommendations responding to the technical and cultural challenges of joining-up data across local authorities to perform analysis. The pilot has also encouraged a number of London boroughs to take stock of how they work with data and to continue experimenting with analytics.

Phase 2: Local authority-led HMO model
Another important outcome has been the development of a second iteration of the HMO model by Barking and Dagenham. Their model confirms the importance of local context to finding HMOs. The adjustments made by the borough have produced promising early results and will inform GLA's work on Phase 3 of the project.

The Barking and Dagenham Model

Barking and Dagenham's experience in Phase 1 of the LODA pilot highlighted the need for greater familiarity with local characteristics and data idiosyncrasies, which influence analysis. For example, while there are roughly 300 HMOs in the borough, further investigation revealed that about 50 of these were newly-built and recently occupied flats on the same site, for which little data had been collected. This significantly skewed analysis in Phase 1.

In Phase 2, these properties were removed, and the borough reviewed the data it deemed relevant to HMOs. It analysed ASB complaints, side waste reports, occupancy numbers, changes in electors listed at the property, the number of habitable rooms, changes in council tax surnames, council tax reductions received, and housing benefits recipients.

Employing the same analytic method as in Phase 1 (balanced random forest), Barking and Dagenham determined that the number of habitable rooms (taken from energy performance certification data), occupancy, elector turnover and council tax reductions were the strongest predictors of HMOs. An unlicensed HMO property - which the model picked out as having the highest probability of being an unlicensed HMO - has been confirmed. The team is now getting ready to test the remaining properties generated by the model.

Phase 3: GLA development
Building on both the work undertaken by ASI during Phase 1 of the pilot and on the insights gained by Barking & Dagenham in Phase 2, the GLA have committed to taking the HMO model forward into a third phase. Working directly with the code developed by ASI the GLA will initially work with public data, and data held at City Hall, to establish whether a working predictive model can be developed at the London level.

In addition, the GLA will look to evolve ASI borough-level algorithm with the aim providing boroughs with a more powerful version of the original model. One of the key lessons taken from Phases 1 and 2 of the pilot was that quality is better than quantity when incorporating datasets into the model. In Phase 1, much of the predictive power of the model came from two or three key datasets and in Phase 2 Barking and Dagenham showed how the inclusion of particular dataset which correlate strongly with HMOs can significantly increase predictive power. The GLA will work to identify which datasets are most valuable in order to provide boroughs with a clear set of guidelines for prioritising and collating datasets for use in the model.

The future of LODA

The establishment of the London Office of Data Analytics is an opportunity to draw together projects, ideas, initiatives, expertise and resources from across the public sector in London to answer the most important questions our city faces. The role of data in informing and shaping policy is increasingly being recognised and the time is right to establish a central co-ordinating office for data discovery, exploration and application.

The pilot programme - and indeed the full range of other data-driven collaborative projects being undertaken in boroughs, at the GLA and through groups like the Borough Data Partnership - gives valuable insight into how LODA will operate and ad value.

The GLA is currently developing proposals for the establishment of LODA within the Intelligence Unit at City Hall. In its initial form LODA will be a forum for innovation and collaboration with data across London.

The LODA operating model

In its initial form LODA will provide:

A central project management and delivery facility will ensure that projects are supported by city data and policy-literate project officers.

- A core resource of data science expertise will undertake data projects, contribute to the development of proposals, and undertake data discovery.
- A data science academy will help to build data science capacity across the public sector in London.
- The London Datastore act a as convening point where ideas can be exposed, explored and formed into viable projects. The GLA is also set to launch a secure data sharing platform which, among other things, will facilitate LODA projects through the safe exchange of data.
- By providing legal support and information governance advice LODA will make the process of sharing and collaboratively working with data safer and more efficient.

“The London Datastore act a as convening point where ideas can be exposed, explored and formed into viable projects”

The launch of the London Office of Data Analytics is the next step in the continued move towards data-informed policy and decision making in London. Over the coming years the function and offer of LODA will evolve to meet the changing needs of London and the organisations it serves. It will ensure that London stays at the cutting edge of something and retains its place among the most forward-thinking cities in the world.

This pilot was the first step on this journey, the launch of LODA in early 2018 is the next.

Other formats and languages

For a large print, Braille, disc,
sign language video or audio-tape version of this document, or if you would like a summary of this document in your language please contact us at this address:

Public Liaison Unit
Greater London Authority
City Hall
The Queen's Walk
More London
London SE1 2AA

Telephone 020 7983 4000
Minicom 020 7983 4458
www.london.gov.uk

You will need to supply your name, your postal address and state the
format and title of the publication
you require.