London Borough Data Partnership #8

29 March 2018

City Hall
GLA

GREATERLONDONAUTHORITY

Smart London

Dr Stephen Lorimer
Greater London Authority
E: stephen.lorimer@london.gov.uk
T: @slorimer44 @mayoroflondon #smartlondon
W: medium.com/@smartlondon

MAYOR OF LONDON

Creative
Creative industries
Cultural institutions

Scientific
World-class universities & research
MedTech

Data economy
40+ tech clusters 47k tech businesses
Growth sectors (CleanTech, GovTech)

CITY GOVERNMENT AND SMART LONDON

MAYOR OF LONDON
Strategies for Transport for London, policing, fire, environment, economic development, skills and planning

33 LONDON BOROUGHS
Social services and other local services like waste, public health, libraries

HEALTH AND EDUCATION
Hospitals, GPs, universities, and research institutions

UTILITIES
Water, Gas, Electricity
Telecoms and broadband providers,
Train Operating Companies

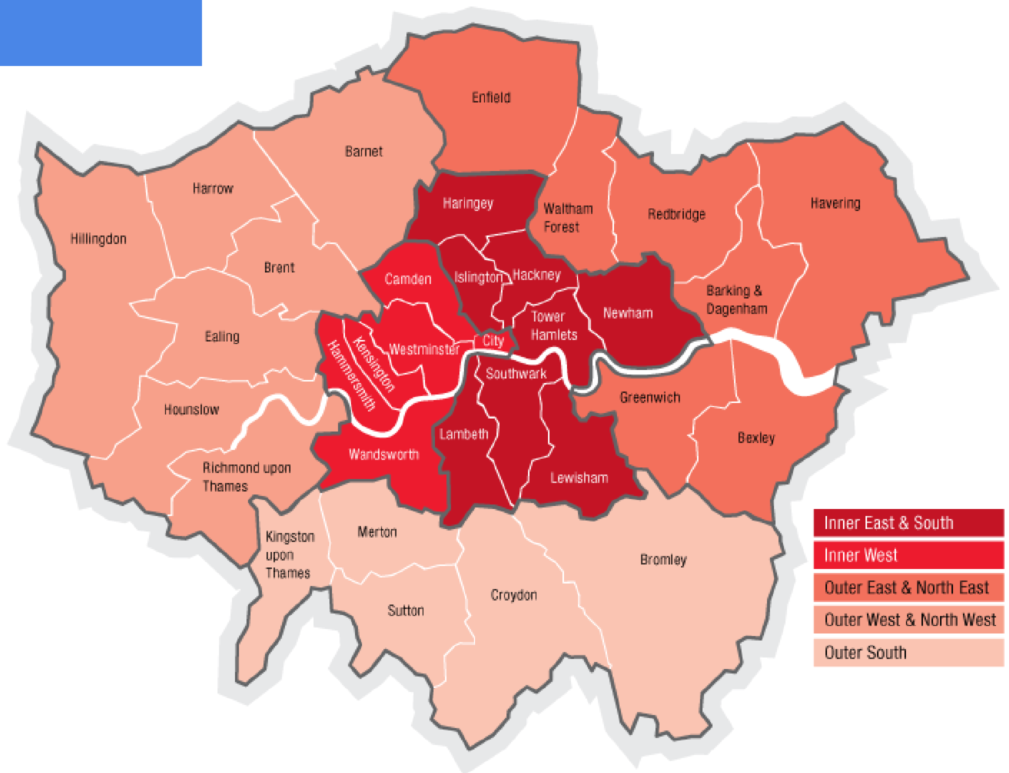
Planning
New London Plan
Future guidance

Delivery
Transport
Policing
Skills
Health

Civic
Non-stat plans
'Convening power'

Budgetary
Grant-making

THAT MAP....



MAYOR OF LONDON

FCC study of 21 cities globally

- ➔ Local governments lack the capacity to understand, develop and implement smart city strategies - requires senior leadership.
- ➔ Need to be embedded in statutory frameworks and plans.
- ➔ For a smart city that is adopted by people and therefore impactful, a collaborative approach is required.
- ➔ Most smart city funding is still derived from innovation pots and not linked to core city funding.
- ➔ Smart city leads should consider how to give clarity to potential private sector partners.

Big systems

↓

Citizen-centric services

Chief Digital Officer & Smart London

Leadership

Shared digital vision, outcomes & language across GLA
Capacity-building at GLA and councils

Collaboration

Scoping foundations with boroughs (LOTI)
Mobilising public sector assets

Innovation

Innovation story
London Tech Week 2018 - Smart City & Women-in-Tech themes

New Smart London Plan

Delivery-focus - 20 actions for London
Partnership with Bloomberg Associates
Driven by Smart London Board
Starts with major Listening Exercise
Launch Q2 2018

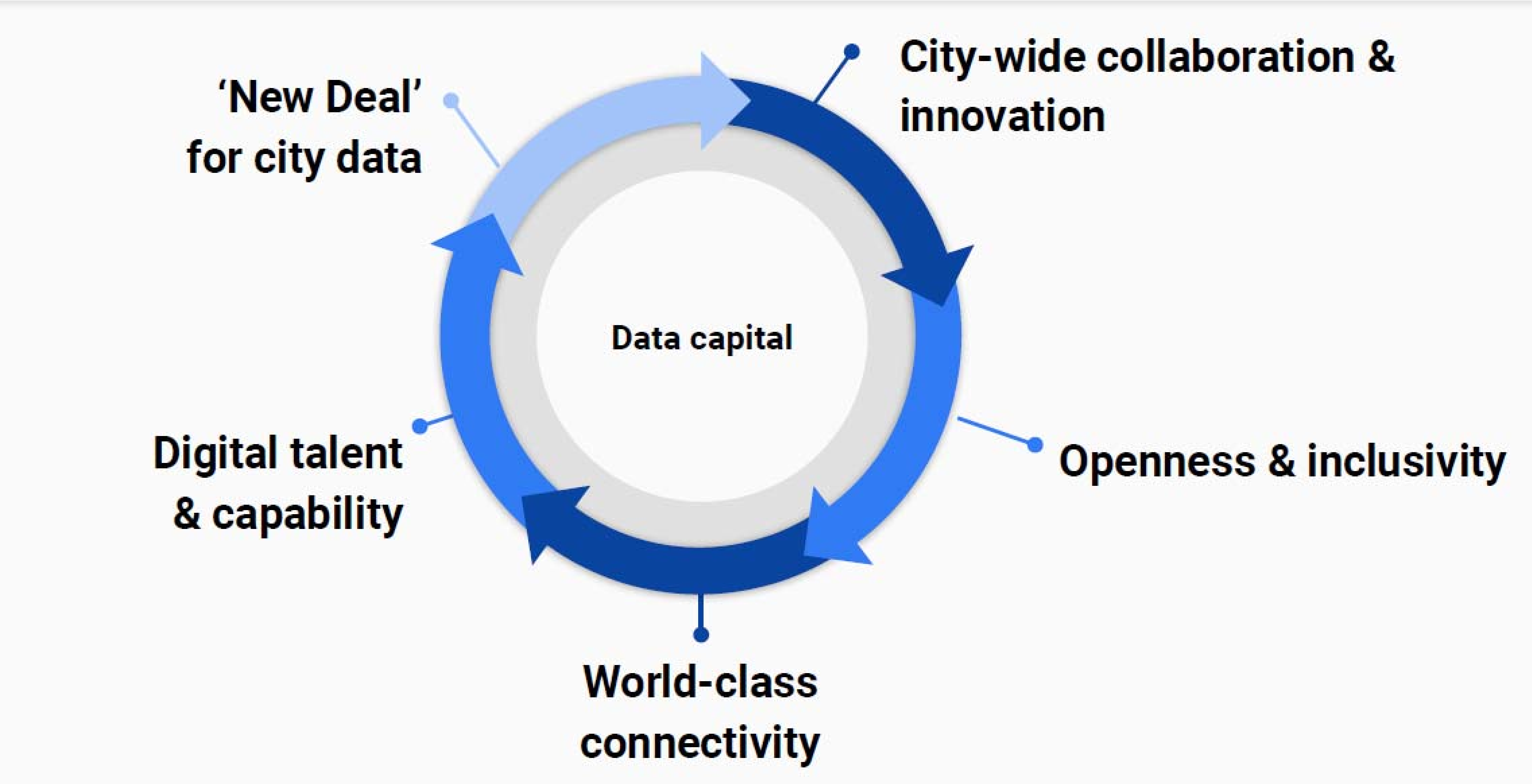
Theo Blackwell
Chief Digital Officer for London
Greater London Authority

Outline Vision

*A 'Smarter London' uses data and technology together for the **good growth** of our city. It mobilises the **power of data** as the fuel for innovation to design and develop safe, open and inclusive solutions for city growth challenges London faces over the next decade and beyond.*

*To stay ahead of the technology curve, rather than follow it, a Smarter London needs new **city-wide collaboration** between public institutions, utilities, our world-class creative, scientific research and tech communities by and for Londoners.*

Smart ‘enablers’



Smart London
Dr Stephen Lorimer
Greater London Authority
E: stephen.lorimer@london.gov.uk
T: @slorimer44 @mayoroflondon #smartlondon
W: medium.com/@smartlondon
LIFEBUOY
MAYOR OF LONDON

Barking and Dagenham
Council of the Year 2018

" The search for HMOs : Can machine learning help?"

Barking & Dagenham's model to assist in the identification of Houses in Multiple Occupation

Pye Nyunt & Phil Canham
Corporate Insight Hub
28th March 2018

Artificial Intelligence

An intelligent agent that perceives its environment and makes decisions to maximise chances of achieving its goal.

Sub-fields of AI:

Machine Learning *Robotics* *Natural Language Processing*

Machine Learning

Giving computers the ability to learn without being explicitly programmed

Supervised Learning

Classification, Regression

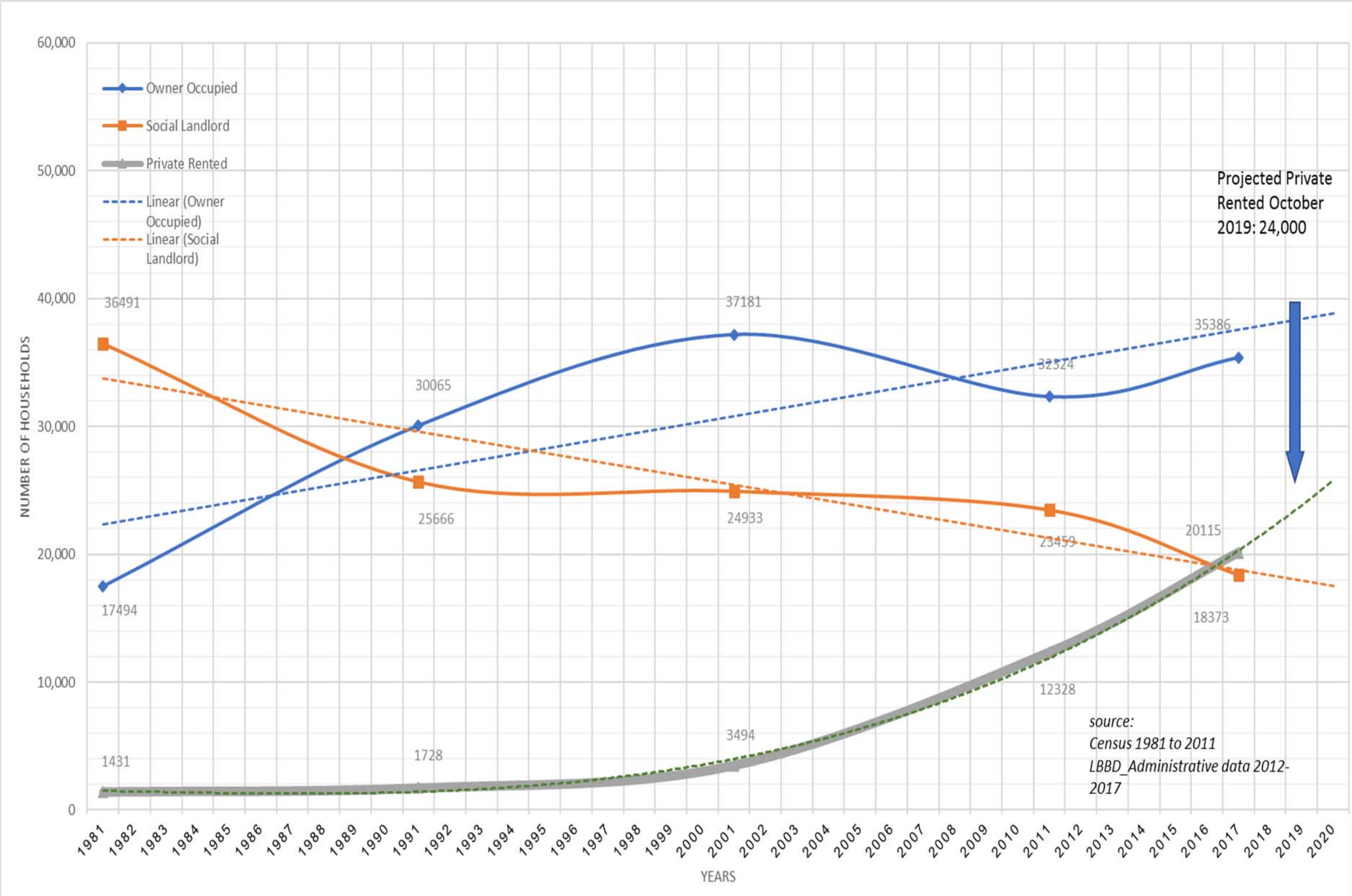
Unsupervised Learning

Clustering, dimensionality, reduction, recommendation

Reinforcement Learning

Reward Maximisation

Change in Private Rented Sector since 1981: LBBD



Aims and definitions

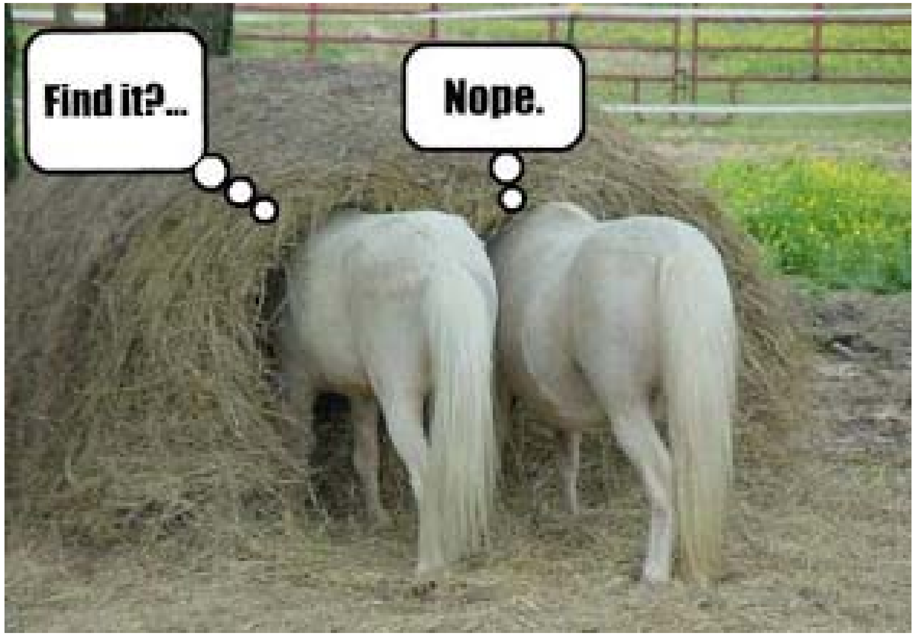
- To provide a robust model which will enable the identification of potential HMOs
- Significantly improve the chance of discovering a new HMO compared to random property selection.
- The aim is not to replace but rather to strengthen the current procedures used by enforcement to identify HMOs
- The definition of HMOs is defined here to coincide with that used by our enforcement service to serve both Mandatory and Additional licences.
- The model will not distinguish between Mandatory and Additional HMOs
- A HMO is a property rented out by at least 3 people who are not from 1 '**household**' (e.g. a **family**) but share facilities like the bathroom and kitchen

Steps

- Define, as far as possible, likely tenure for each household in the borough. The definition of tenure is split here into:
  - Social
  - Owner Occupied
  - Private Rented
- Identify all known and licenced HMOs
- Identify and test a range of potential indicators that are more likely to occur for HMOs compared to non HMOs
- Design and evaluate machine learning model which combines the strongest predictors , in the most effective way, enabling an HMO probability/likelihood to be assigned to every household.
- The programming language R was used to undertake all stages in a reproducible way.
- This also allows changes to be made at any stage, as more information is acquired.

Context

- At the time of the initial model building there were 428 licenced HMOs in the borough, comprising 64 Mandatory and 364 Additionally licenced HMOs
- About 75,000 households in total on our Council Tax Register
- So currently licenced HMOs make up only 0.6% of all households!!!!
- Needle in a haystack!

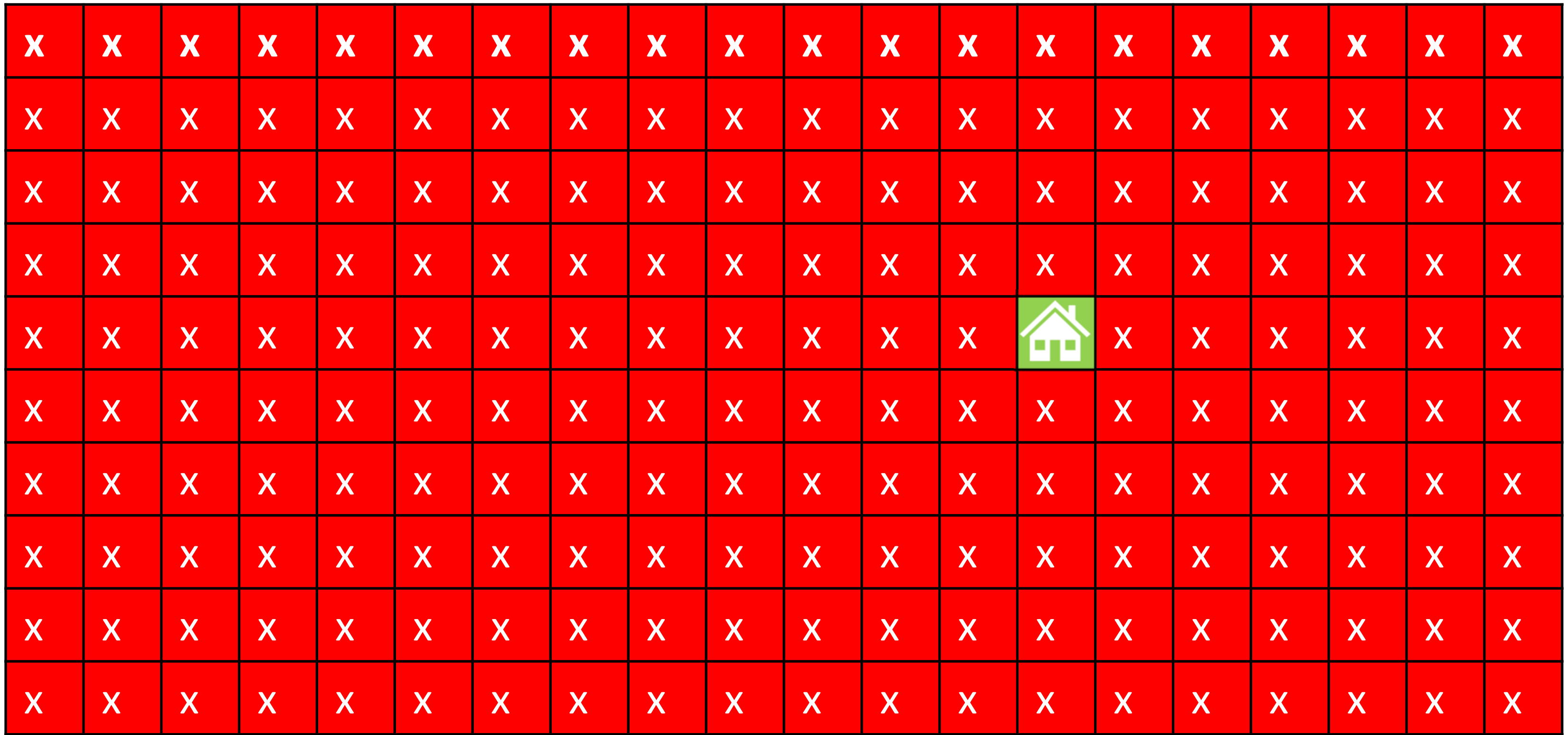


London Borough of Barking & Dagenham
lbbd.gov.uk

Context

Based on our current knowledge

If we visited properties totally at random we would have visit about 200 properties to find just 1 HMO



Preparation

- The identification, collection, cleaning, and preparation stages by far take the most time
- It is essential that these stages are rigorously undertaken
- Equally essential that persons collating these data have good domain knowledge for each dataset.
- Need to understand the strengths and weakness and sources of bias/error in each dataset
- The fundamental building block is the Unique Property Reference Number (UPRN)
- If the UPRN is not available, the data will need to be address matched, adding a further stage.
- We have developed a bespoke algorithm to undertake this: **" Insight Hub Address Matching Algorithm"**

London Borough of Barking&Dagenham
lbbd.gov.uk

Preparation

- Once data is cleaned it is combined into one set with unique records for each UPRN (75000) – so the UPRN is used as the primary and unique key.

- If a dataset has any duplicate UPRNs the data must be grouped and counted or summed to achieve a set of unique records

- For the combined dataset: the first column records the UPRN, the next contains a binary flag identifying a household as either a known HMO or not, the third a known tenure flag

- All other columns include information or predictor variables recorded, at this stage as either numeric, or factor variables. At this stage we have not chosen which variables to use in the final model.

- The next stage is to create "dummy" variables for all factor levels.

- For example, for tenure, the tenure variable would be split into three new variables
  1. Social Rented : Yes/No (1,0)
  2. Private Rented: Yes/No (1,0)
  3. Owner Occupied: Yes/No (1,0)

- This permits greater flexibility in the models that can be created.

Preparation

- We now have a dataset with about 75,000 unique records and ***30 plus*** columns
- Key Variables included

  - ASB incidents (taken from Enforcement Database) : Property Yes/No over 3 years
  - Sidewaste(6 week recording with Refuse service) Extra Rubbish left out with bin: (Yes/No)
  - Occupancy: Residents Matrix 2015- cross matching Councils Admin datasets: numeric
  - Electors 2015: numeric
  - Electors 2017: numeric
  - Change in Electors 2015 – 17: numeric
  - Habitable Rooms (Energy performance certificates or imputed from Ctax band) numeric
  - Change in Surname of Council Tax primary occupant 1 year (Yes/No)
  - HB recipient? 2017: Yes/No
  - Ctax reduction? Yes/No
  - Flat (Yes/No)

- Some variables may covary – not a problem in tree based models

Preparation

- A key point in the analysis was to select only currently licenced and compliant private rented accommodation for the training of the model. This brought the overall total of households down to about 10,000. Not doing this, we believe, may have caused the LODA/NESTA model to underperform.

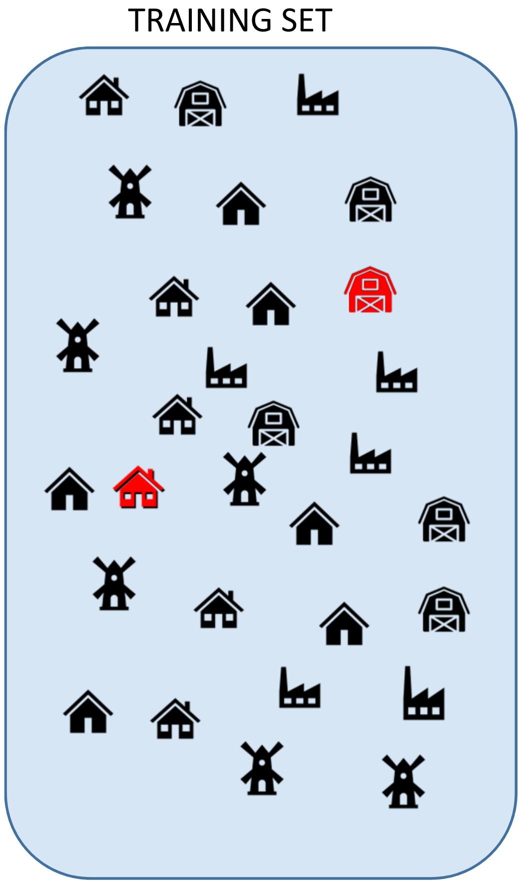
- Training and Test sets were created from this Private Rented subset: 70% of data in a training set, 30% in test set (data were randomly selected with stratification to ensure an equal balance of the target HMO variable. This is an important step, given the large imbalance in the data) – used the R Caret Package

- Random Seeds were set for reproducibility

- Exploratory analysis was then undertaken to establish potential predictors.

- All exploratory analysis was undertaken only on the training set.

Here is the idea !

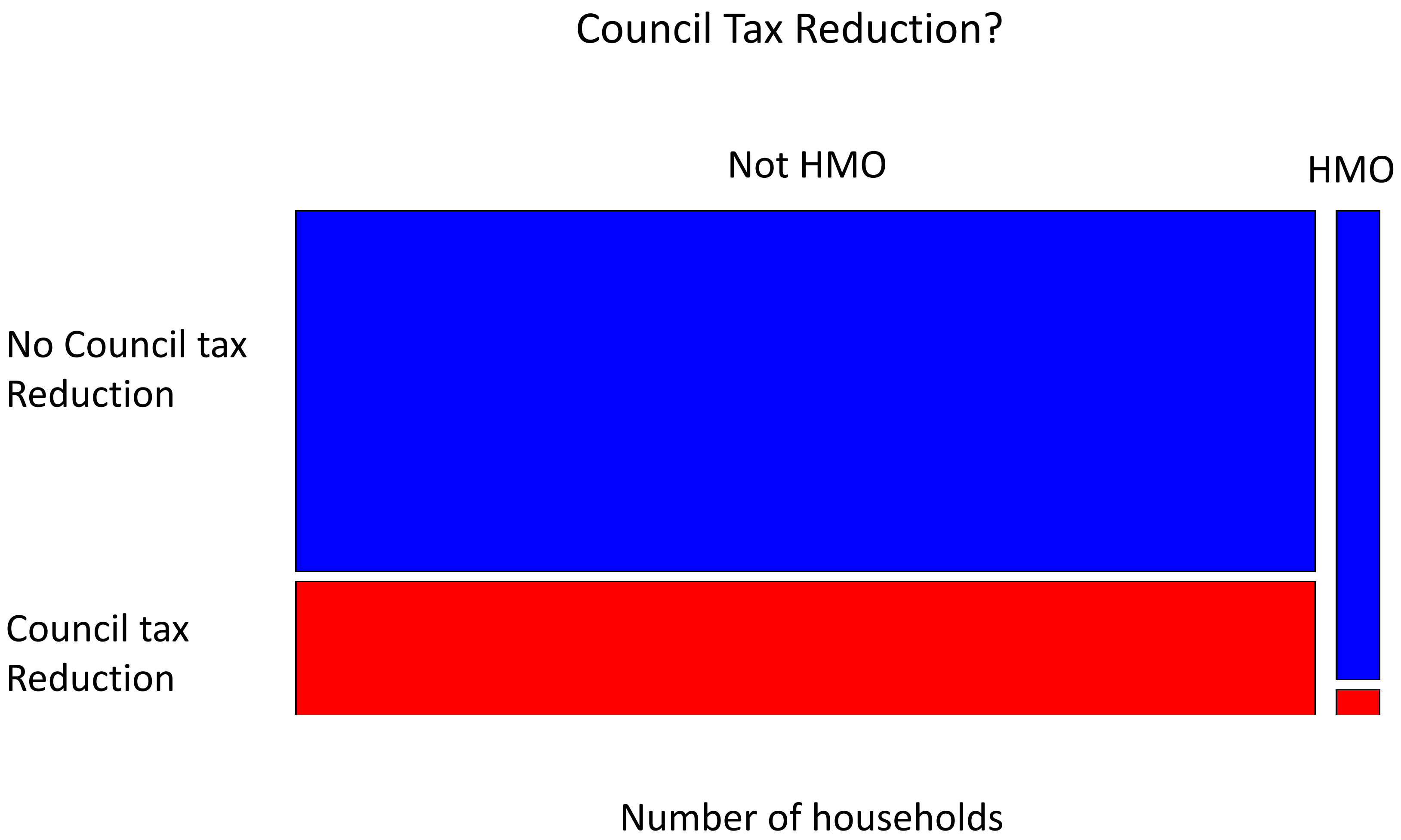


Exploratory data analysis

- Each Variable was tested individually between known HMO and non HMO properties
- The tests were undertaken on the training set only – to avoid overfitting
- Chi Squared test done on each for significance.
- Simple Mosaic Plots were used for visualising differences in factor variables
- Here are some of the key variables we ended up using in the models

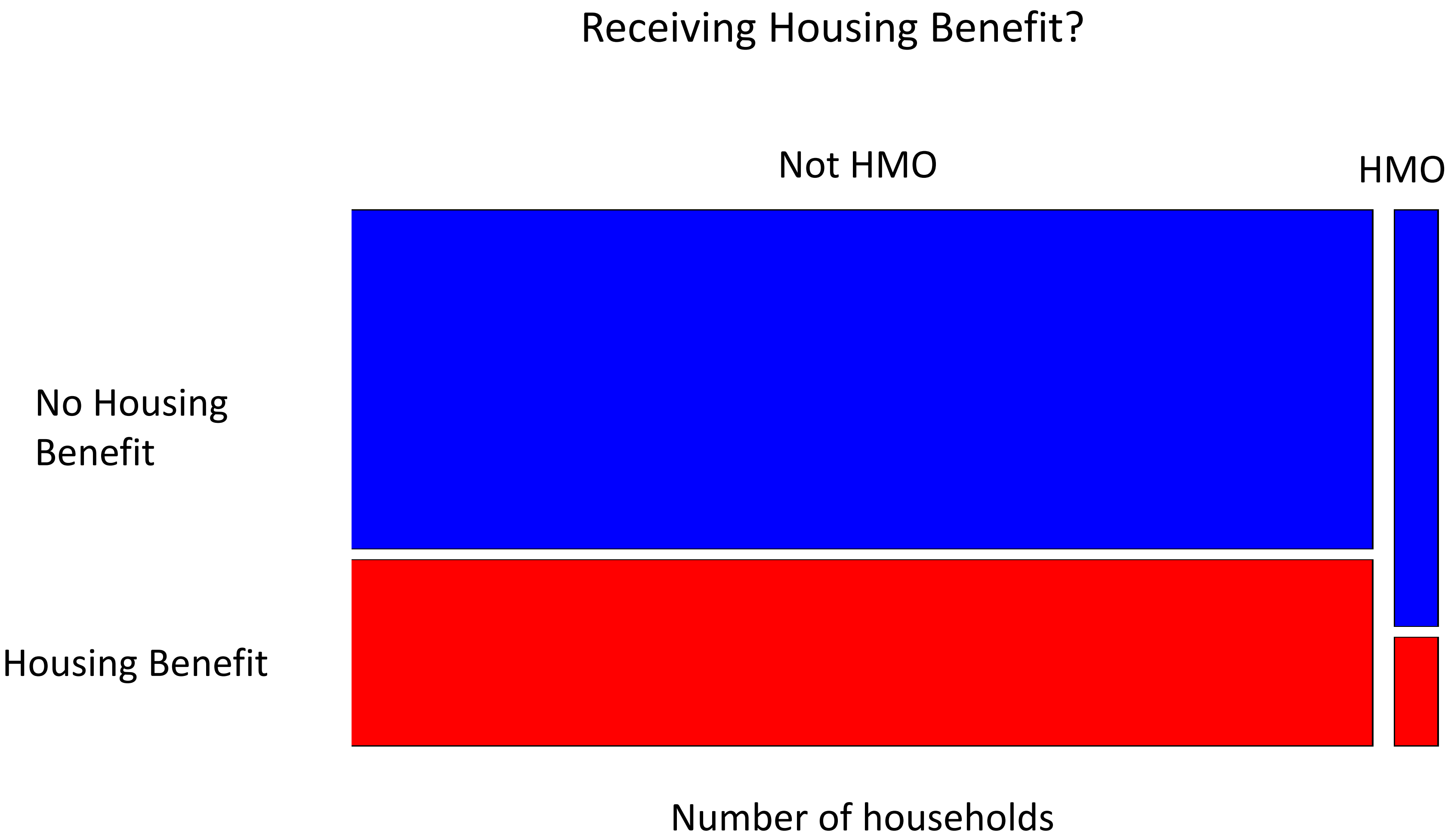
London Borough of Barking&Dagenham lbbd.gov.uk

Exploratory data analysis

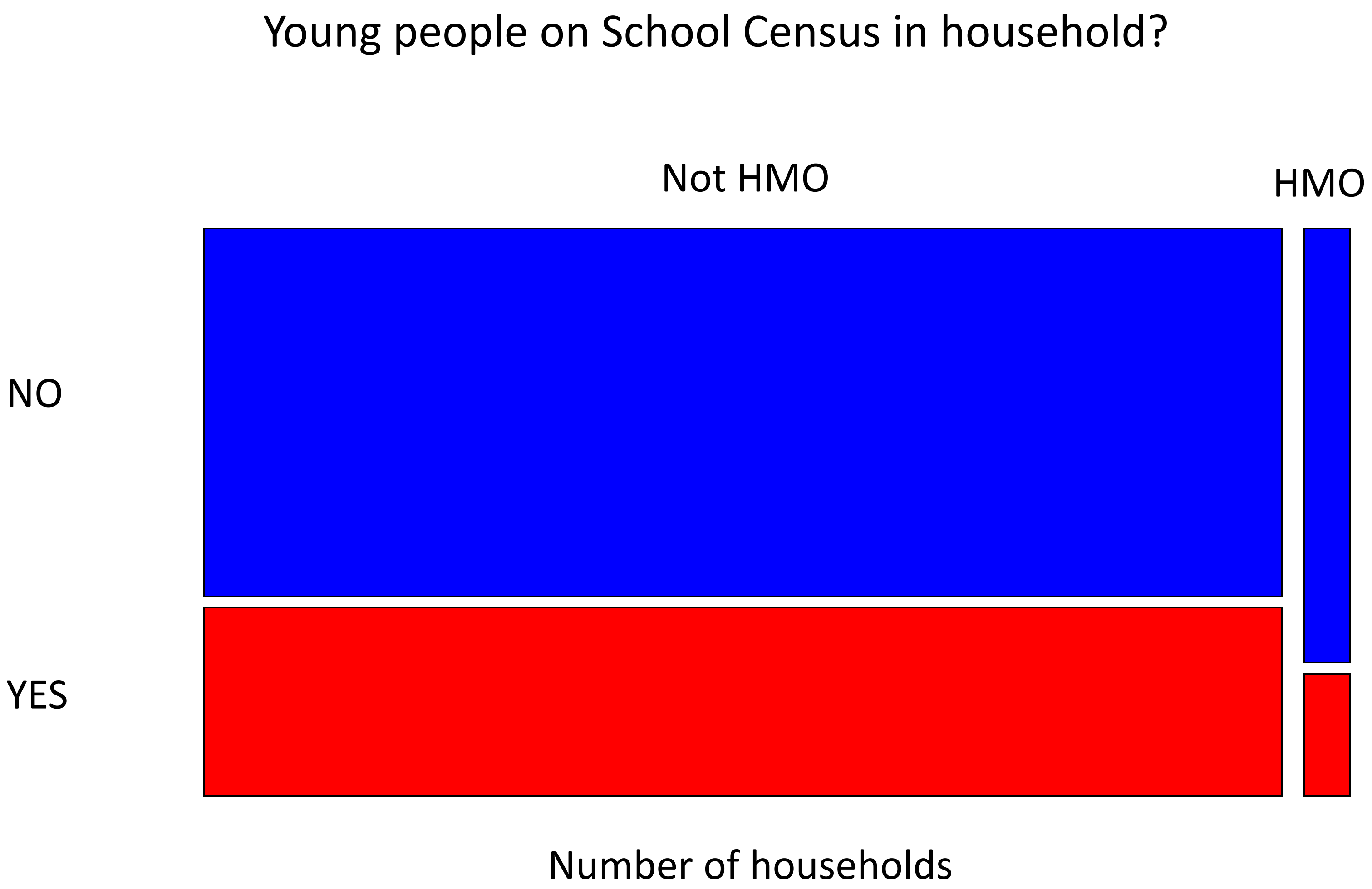



```
Pearson's Chi-squared test with Yates' continuity correction data: table(training$HMO,
training$Ctax_Disc) X-squared = 67.776, df = 1, p-value < 2.2e-16
```

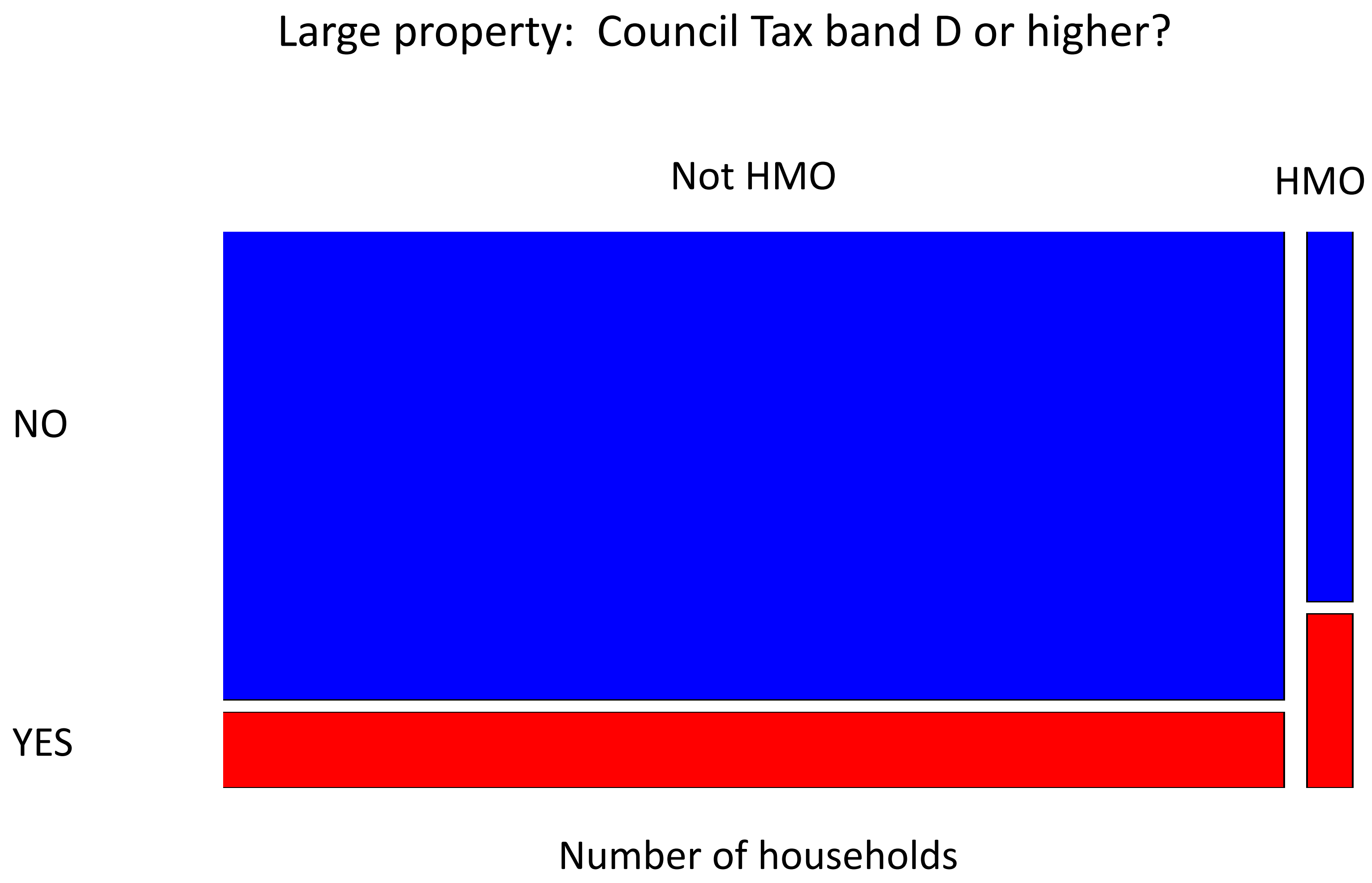
Exploratory data analysis



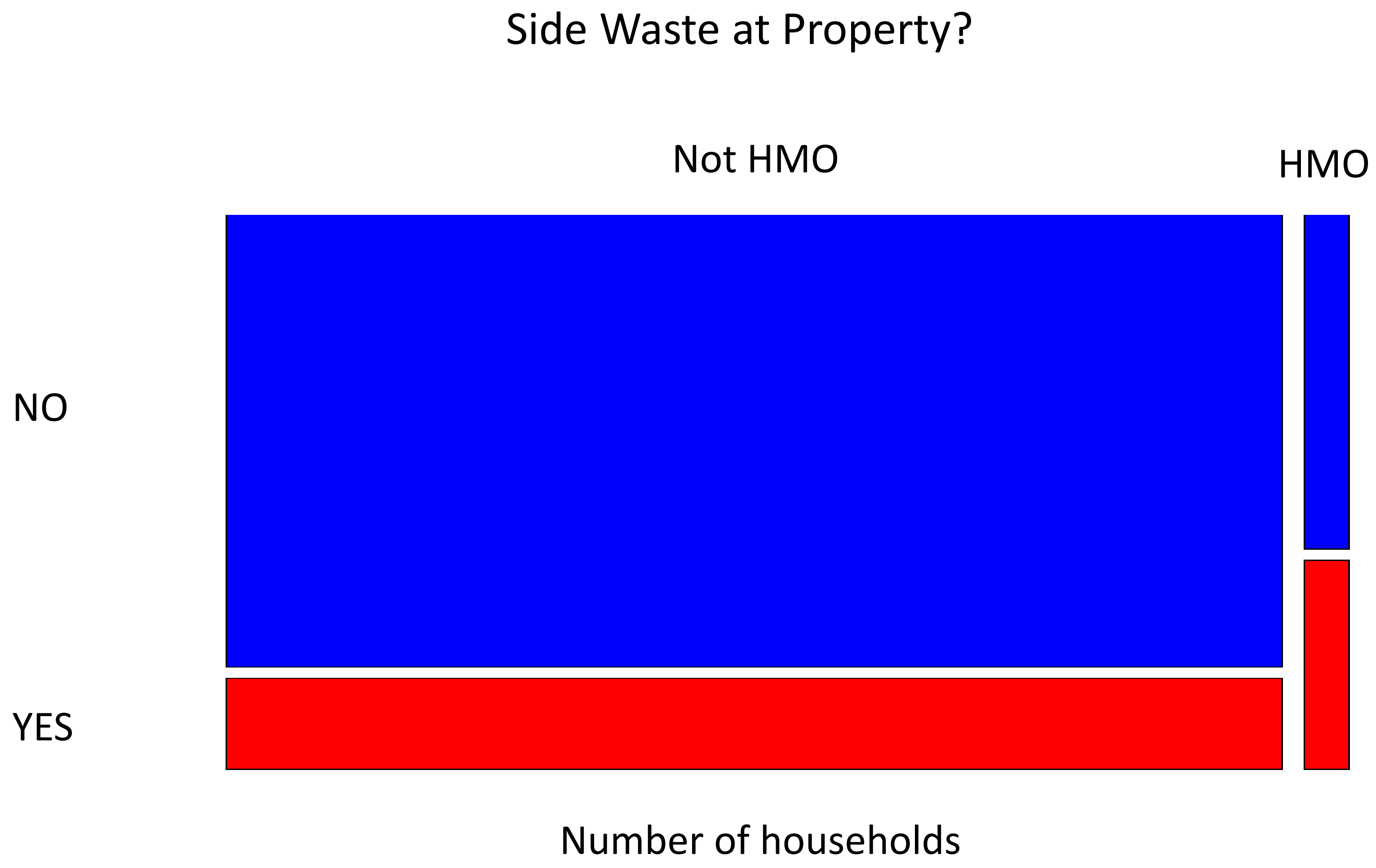
Exploratory data analysis



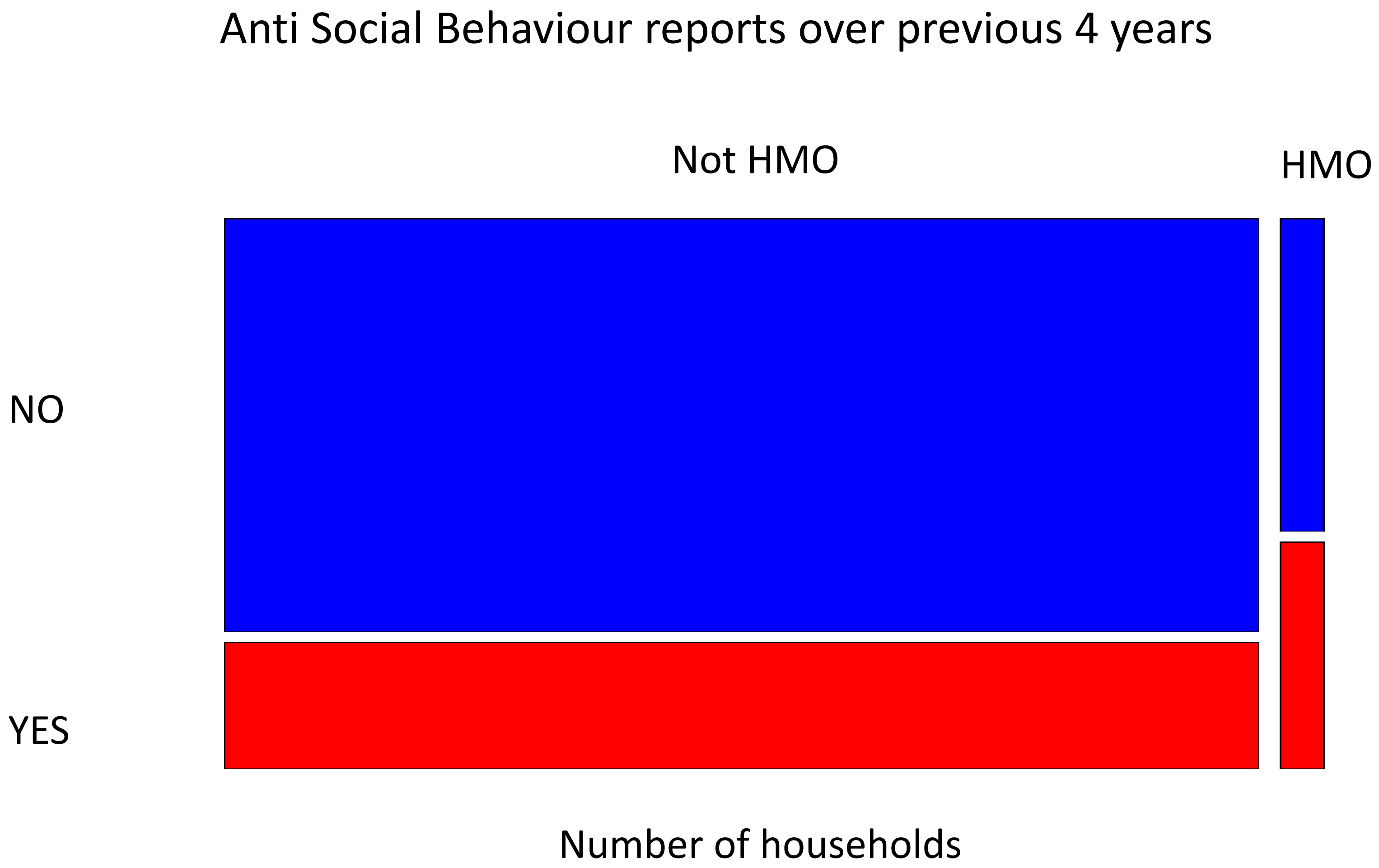
Exploratory data analysis



Exploratory data analysis



Exploratory data analysis



Exploratory data analysis

We also used other predictors including

- Number of moves on electoral register between two points in time
- Number of Habitable Rooms: These were taken from Energy Performance certificate data available from MHCLG. Missing values were imputed from Council Tax band
- Most recent date property sold: Land Registry data

May add further geographical variables in the future, for example to define most likely parts of the borough to have private rented properties

London Borough of Barking & Dagenham
lbbd.gov.uk

Model Building

- Tree based algorithms were primarily used to build the models partly because there is no assumptions needed about normality of data nor the effect of covariance

- We did undertake multiple logistic regression, which had reasonable results, but feel the data may need to be transformed and more carefully chosen

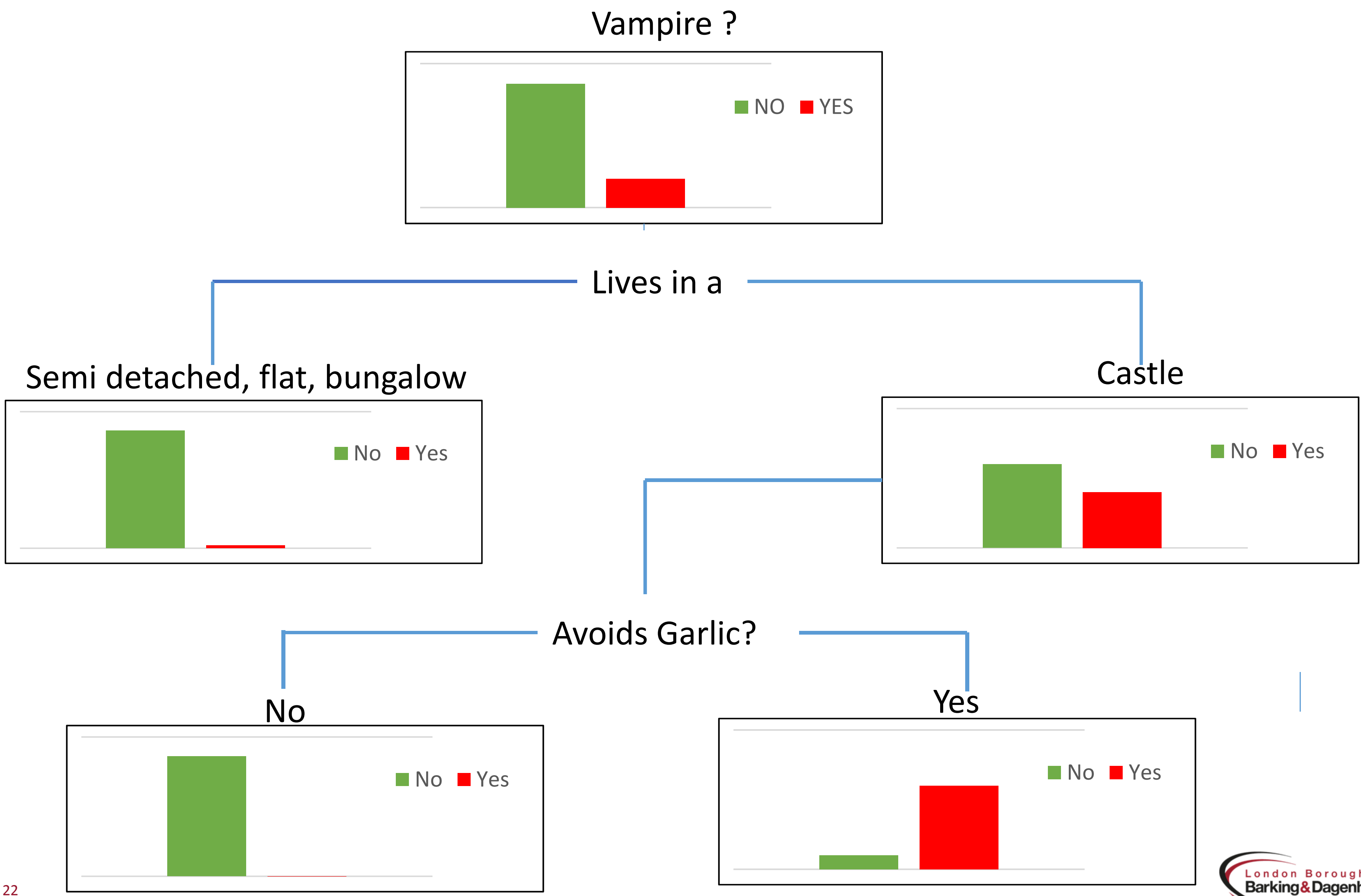
- We used several models including:
- Extreme Gradient Boosting (Xgboost)
- Random Forest
- Balanced Random Forest

- Each model was run via the R Caret package and was trained and checked using 10 fold cross validation, and some other tuning parameters, including tree size, tree number

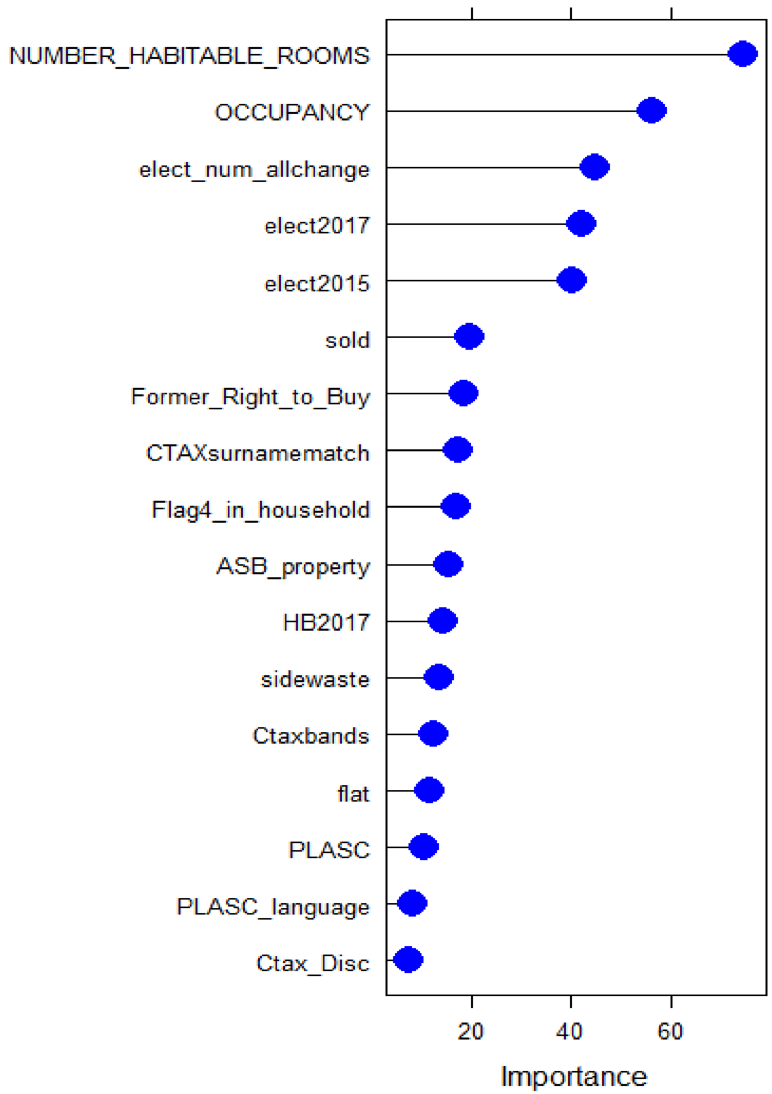
- The models were built to include a probability score so optimal thresholds could be ascertained.

- Key aim in setting threshold was to balance the number of visits required by staff to find an HMO whilst maximising the number of households found

Example of a decision tree : Identifying Vampires



Training Results: Standard Random Forest: 5000 trees



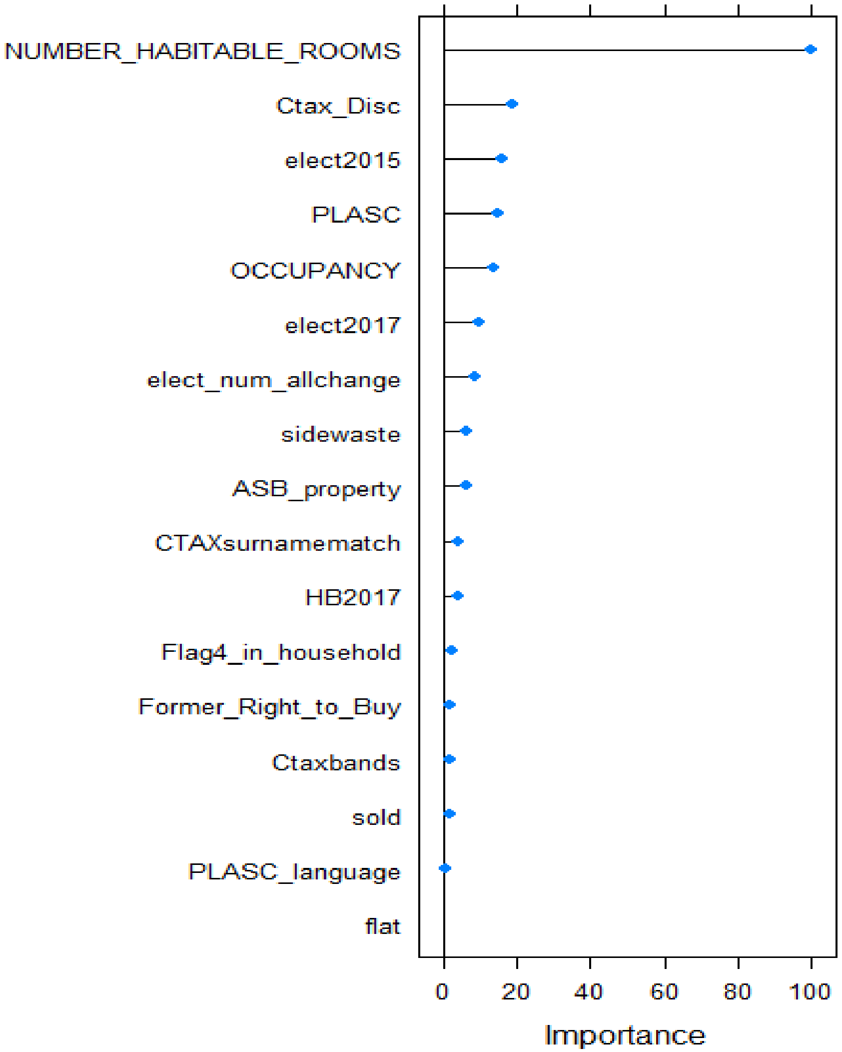
Random Forest 5000 trees

		Actual	
		NOT HMO	HMO
Predicted	NOT HMO	2033	36
	HMO	570	65

Precision: 65/(65+570) = **10%**
Visit 10 households find one HMO

Recall: 65/(65+36) = **64%**
Identify 64% of all HMOs

Training Results: XGBoost



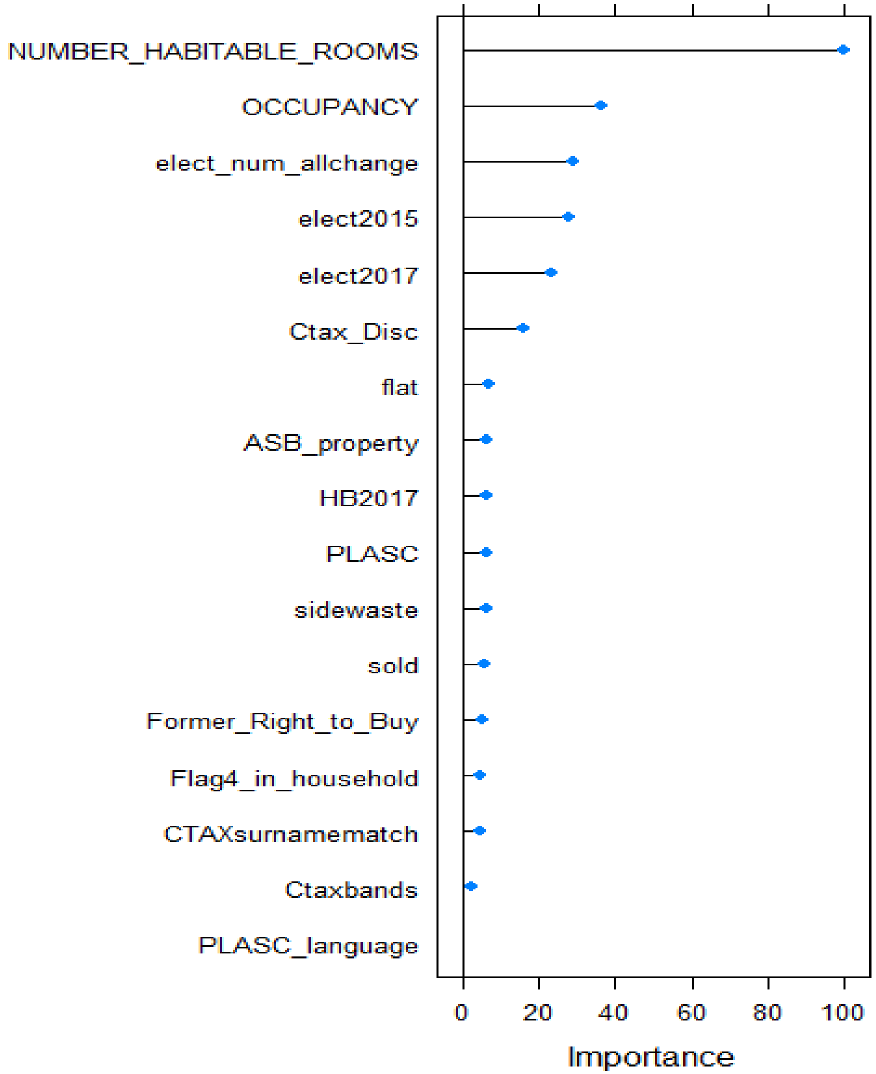
XG boost

		Actual	
		NOT HMO	HMO
Predicted	NOT HMO	1852	18
	HMO	751	83

Precision: 83/(83+751) = **10%**
Visit 10 households find one HMO

Recall: 83/(83+18) = **82%**
Identify 82% of all HMOs

Training Results: Balanced Random Forest 5000 trees



Balanced Random Forest (Up-sampled)

		Actual	
		NOT HMO	HMO
Predicted	NOT HMO	2364	18
	HMO	293	83

Precision: 83/(83+293) = **22%**

Visit 5 households find one HMO

Recall: 83/(83+18) = **82%**

Identify 82% of all HMOs

Training Results: Ensemble method

		Actual	
		NOT HMO	HMO
Predicted	NOT HMO	2351	18
	HMO	252	82

Precision: 82/(82+252) = **25%**
Visit 4 households find one HMO

Recall: 82/(82+18) = **82%**
Identify 82% of all HMOs

London Borough of Barking&Dagenham lbbd.gov.uk

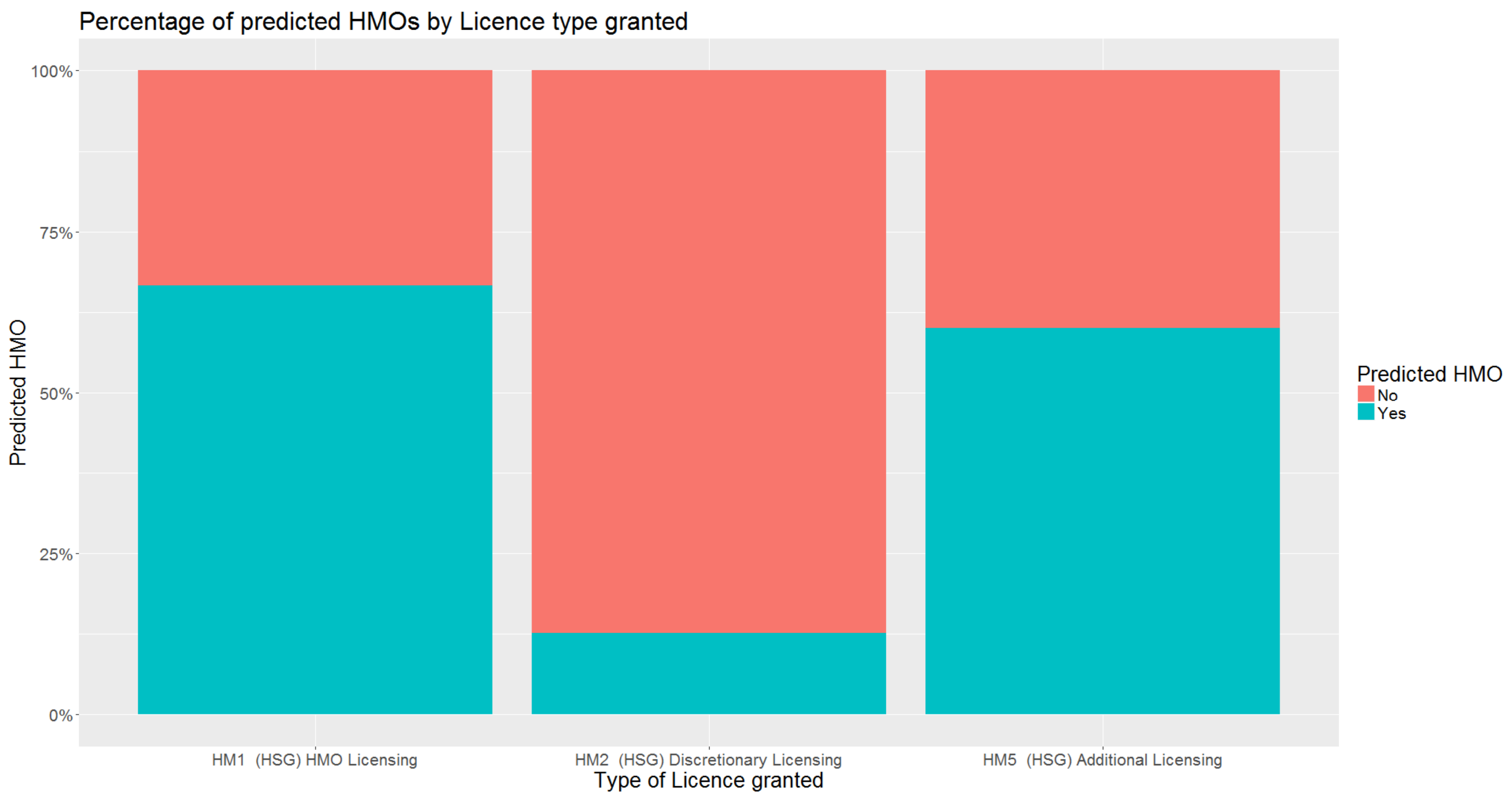
Real Results: Between October 2017 and March 2018

		Actual	
		NOT HMO	HMO
Predicted	NOT HMO	582	12
	HMO	87	20

Precision: 20/(20+87) = **17%**
Visit 6 households find one HMO

Recall: 20/(20+12) = **62.5%**
Identify 62.5% of all HMOs

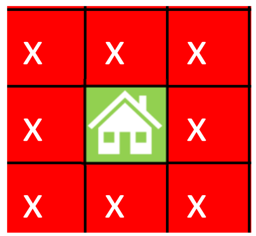
Results: Licences granted since October 2017



And the Punchline!!

Based on results so far

We have made improvements from this



To This

If we are lucky!!!

THANK YOU

ANY QUESTIONS?

London Borough Data Partnership #8

29 March 2018

City Hall
GLA

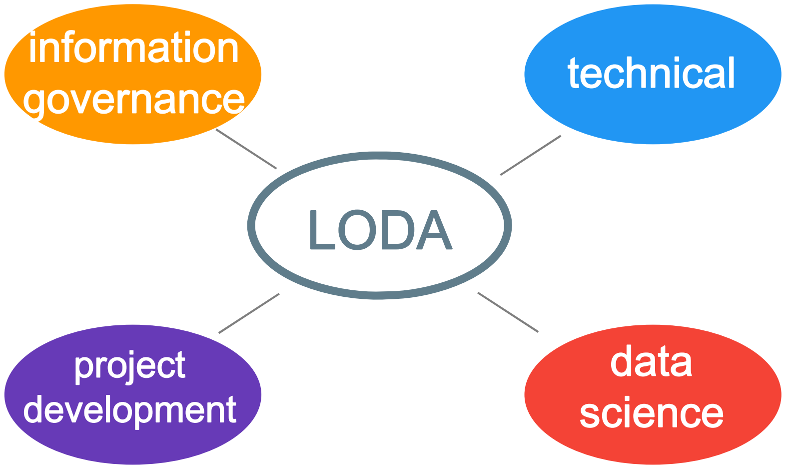
GREATERLONDONAUTHORITY


London Office of Data Analytics
GREATERLONDONAUTHORITY

LODA pilot aims

- Test the policy or service impact of data science
- Show that data-sharing is possible and has tangible benefits
- Develop data sharing protocols useful for the longer term
- Identify barriers to collaborative working and develop solutions
- Contribute to the development of a culture of data-sharing within London

London Office of Data Analytics



a **hub** for the

development,

commissioning and

implementation

of **data science** projects aimed at

addressing public services and urban challenges

which are better tackled together, and which may affect more than one agency

MAYOR OF LONDON

NEWS

Home | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Technology

Cambridge Analytica: The story so far

By Zoe Kleinman
Technology reporter, BBC News

21 March 2018

Share

Cambridge Analytica: What we know so far

It's a sensational story containing allegations of sleaze, psychological manipulation and data misuse that has provoked an internationally furious response.

Tech giant Facebook and data analytics firm Cambridge Analytica are at the centre of a dispute over the harvesting and use of personal data - and whether it was used to influence the outcome of the US 2016 presidential election or the UK Brexit referendum.

v

NEWS

Home | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Politics | Parliaments | Brexit

Brexit: MPs warn of multiple obstacles to EU security deal

21 March 2018

Share

Brexit

A longer transition period after Brexit may be needed to guarantee continued security co-operation, MPs have said.

information governance

technical

data science

project development

MAYOR OF LONDON

Information Sharing Gateway (ISG)

1. A catalogue of data sharing agreements
   - For a given dataset – who am I sharing it with (and under what terms)?
   - For a given organisation – what are we already sharing with them?

2. Create 'high-level' agreements between organisations
   - Signed off by Information Governance lead
   - Can be referenced by the individual data flows

3. Electronic signoff
   - Delegated authority

4. Organisation assurance

Organisation Name	Single Lic	ICO Number	Category	Contact Email	Assurance	Admin Group	Lic	Setup %
GREATER LONDON AUTHORITY		Z4760661	Local Authority	Paul.Hodgson@london.gov.uk	Significant	London Information Sharing Alliance		100%
LONDON FIRE & EMERGENCY PLANNING AUTHORITY (LFEPA)		Z7122455	Fire Service	andrew.mobbs@london-fire.gov.uk	Significant	London Information Sharing Alliance		86%
LONDON LEGACY DEVELOPMENT CORPORATION		Z3138681	Local Authority	DannyBudzak@londonlegacy.co.uk	Not submitted	London Information Sharing Alliance		57%

information governance

technical

data science

project development

MAYOR OF LONDON

Information Sharing Gateway (ISG)

1. A catalogue of data sharing agreements
   - For a given dataset – who am I sharing it with (and under what terms)?
   - For a given organisation – what are we already sharing with them?

2. Create 'high-level' agreements between organisations
   - Signed off by Information Governance lead
   - Can be referenced by the individual data flows

3. Electronic signoff
   - Delegated authority

4. Organisation assurance

Organisation Name	Single Lic	ICO Number	Category	Contact Email	Assurance	Admin Group	Lic	Setup %
GREATER LONDON AUTHORITY		Z4760661	Local Authority	Paul.Hodgson@london.gov.uk	Significant	London Information Sharing Alliance		100%
LONDON FIRE & EMERGENCY PLANNING AUTHORITY (LFEPA)		Z7122455	Fire Service	andrew.mobbs@london-fire.gov.uk	Significant	London Information Sharing Alliance		86%
LONDON LEGACY DEVELOPMENT CORPORATION		Z3138681	Local Authority	DannyBudzak@londonlegacy.co.uk	Not submitted	London Information Sharing Alliance		57%

information governance

technical

data science

project development

MAYOR OF LONDON

Information Sharing Gateway (ISG)

1. A catalogue of data sharing agreements
    - For a given dataset – who am I sharing it with (and under what terms)?
    - For a given organisation – what are we already sharing with them?

2. Create 'high-level' agreements between organisations
    - Signed off by Information Governance lead
    - Can be referenced by the individual data flows

3. Electronic signoff
    - Delegated authority

4. Organisation assurance

Organisation Name	Single Lic	ICO Number	Category	Contact Email	Assurance	Admin Group	Lic	Setup %
GREATER LONDON AUTHORITY		Z4760661	Local Authority	Paul.Hodgson@london.gov.uk	Significant	London Information Sharing Alliance		100%
LONDON FIRE & EMERGENCY PLANNING AUTHORITY (LFEPA)		Z7122455	Fire Service	andrew.mobbs@london-fire.gov.uk	Significant	London Information Sharing Alliance		86%
LONDON LEGACY DEVELOPMENT CORPORATION		Z3138681	Local Authority	DannyBudzak@londonlegacy.co.uk	Not submitted	London Information Sharing Alliance		57%

information governance

technical

data science

project development

MAYOR OF LONDON

Information Sharing Gateway (ISG)

1. A catalogue of data sharing agreements
   - For a given dataset – who am I sharing it with (and under what terms)?
   - For a given organisation – what are we already sharing with them?

2. Create 'high-level' agreements between organisations
   - Signed off by Information Governance lead
   - Can be referenced by the individual data flows

3. Electronic signoff
   - Delegated authority

4. Organisation assurance

Organisation Name	Single Lic	ICO Number	Category	Contact Email	Assurance	Admin Group	Lic	Setup %
GREATER LONDON AUTHORITY		Z4760661	Local Authority	Paul.Hodgson@london.gov.uk	Significant	London Information Sharing Alliance		100%
LONDON FIRE & EMERGENCY PLANNING AUTHORITY (LFEPA)		Z7122455	Fire Service	andrew.mobbs@london-fire.gov.uk	Significant	London Information Sharing Alliance		86%
LONDON LEGACY DEVELOPMENT CORPORATION		Z3138681	Local Authority	DannyBudzak@londonlegacy.co.uk	Not submitted	London Information Sharing Alliance		57%

information governance

technical

data science

project development

MAYOR OF LONDON

Information Sharing Gateway (ISG)

5. Below each data sharing agreement, you can record individual 'data flows'
   - How are you transferring the data
   - How are you storing it?
   - For a given organisation – what are we already sharing with them?

ISGID	Status	Summary Name	Data Flow Name / Identifier	Risk Rating	Added	First Signed	Export	Copy	Archive / Delete
DF002687		PSMA data for contractors	OS AddressBase data to Energy Saving Trust	Not Assessed	23/01/2018				
DF002686	1/1	PSMA data for contractors	OS AddressBase data to RetrofitWorks	Not Assessed	23/01/2018	23/01/2018			
DF002690		SafeStats: LFB	GLA disclosure to LFB (via SafeStats)	Not Assessed	25/01/2018				

6. GDPR
   - ISG is ICO compliant and will be GDPR compliant
   - Can use ISG to respond to requests such as "what are you doing with data about me?"
   - Can carry out Privacy Impact assessment as part of setting up a data flow

information governance

technical

data science

project development

MAYOR OF LONDON

5. Below each data sharing agreement, you can record individual 'data flows'
    - How are you transferring the data
    - How are you storing it?
    - For a given organisation – what are we already sharing with them?

ISGID	Status	Summary Name	Data Flow Name / Identifier	Risk Rating	Added	First Signed	Export	Copy	Archive / Delete
DF002687		PSMA data for contractors	OS AddressBase data to Energy Saving Trust	Not Assessed	23/01/2018				
DF002686	1/1	PSMA data for contractors	OS AddressBase data to RetrofitWorks	Not Assessed	23/01/2018	23/01/2018			
DF002690		SafeStats: LFB	GLA disclosure to LFB (via SafeStats)	Not Assessed	25/01/2018				

6. GDPR
    - ISG is ICO compliant and will be GDPR compliant
    - Can use ISG to respond to requests such as "what are you doing with data about me?"
    - Can carry out Privacy Impact assessment as part of setting up a data flow

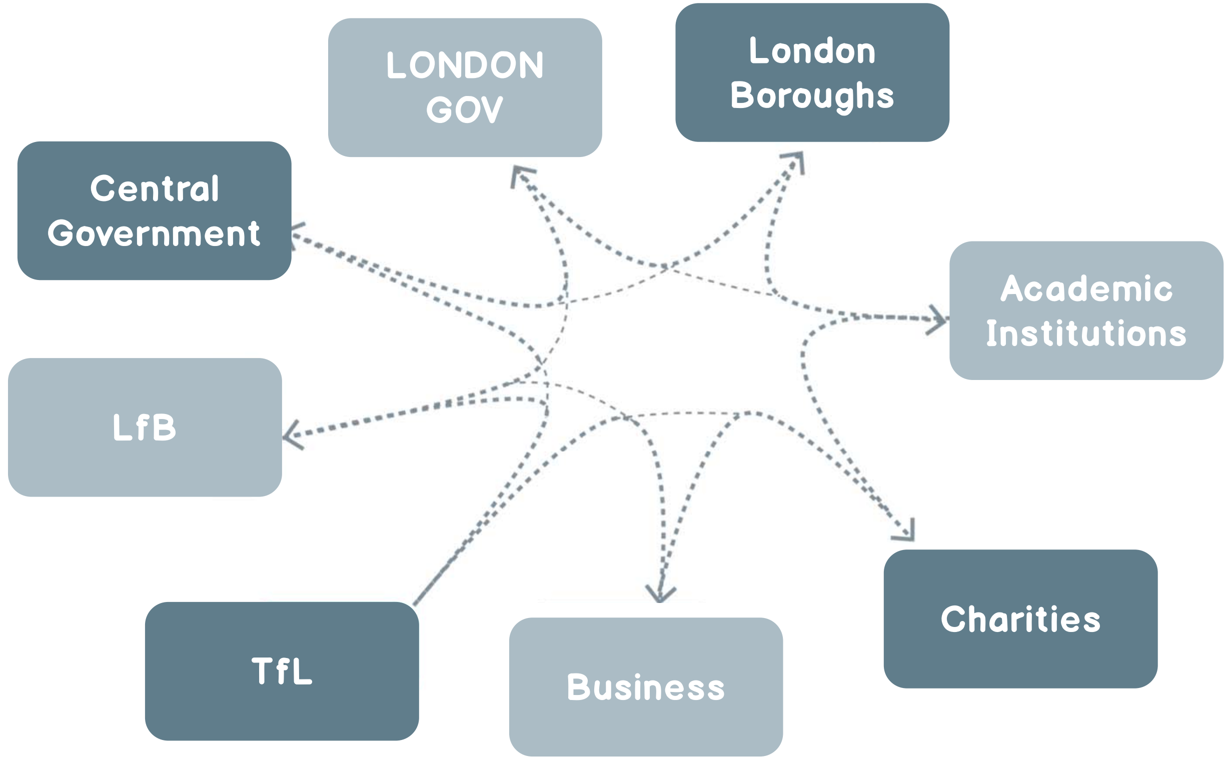
7. Organisations

- 100 initial licences
- Full use of ISG
- Generally defined as an entity that has its own Information Governance set up

8. Sponsored Organisations

- Unlimited
- Can only ISG to share data with the sponsoring organisation – sign up to data sharing
- Smaller organisations (e.g. schools, community groups, etc)

There will never be a single warehouse for all of London's data, so we need to connect..



Principles:

- Open Source
- Cloud-based
- Open APIs
- Sharing knowledge with other cities

- **secure sharing of catalogues &/or data**

information governance

technical

data science

project development

MAYOR OF LONDON

Data Science questions
(in approximate order of difficulty):

- **Descriptive** (can't be generalised)
- **Exploratory** (discover new connections, define future studies)
- **Inferential** (estimate a quantity & uncertainty – e.g. opinion surveys)
- **Predictive** (link not cause – e.g. people who buy lots of strong mints have a higher risk of cancer)
- **Causal** (needs subject knowledge, probably only show average effects e.g. on a neighbourhood rather than a particular household)
- **Mechanistic** (scientific or engineering systems – not applicable to GLA work)

information governance technical data science project development

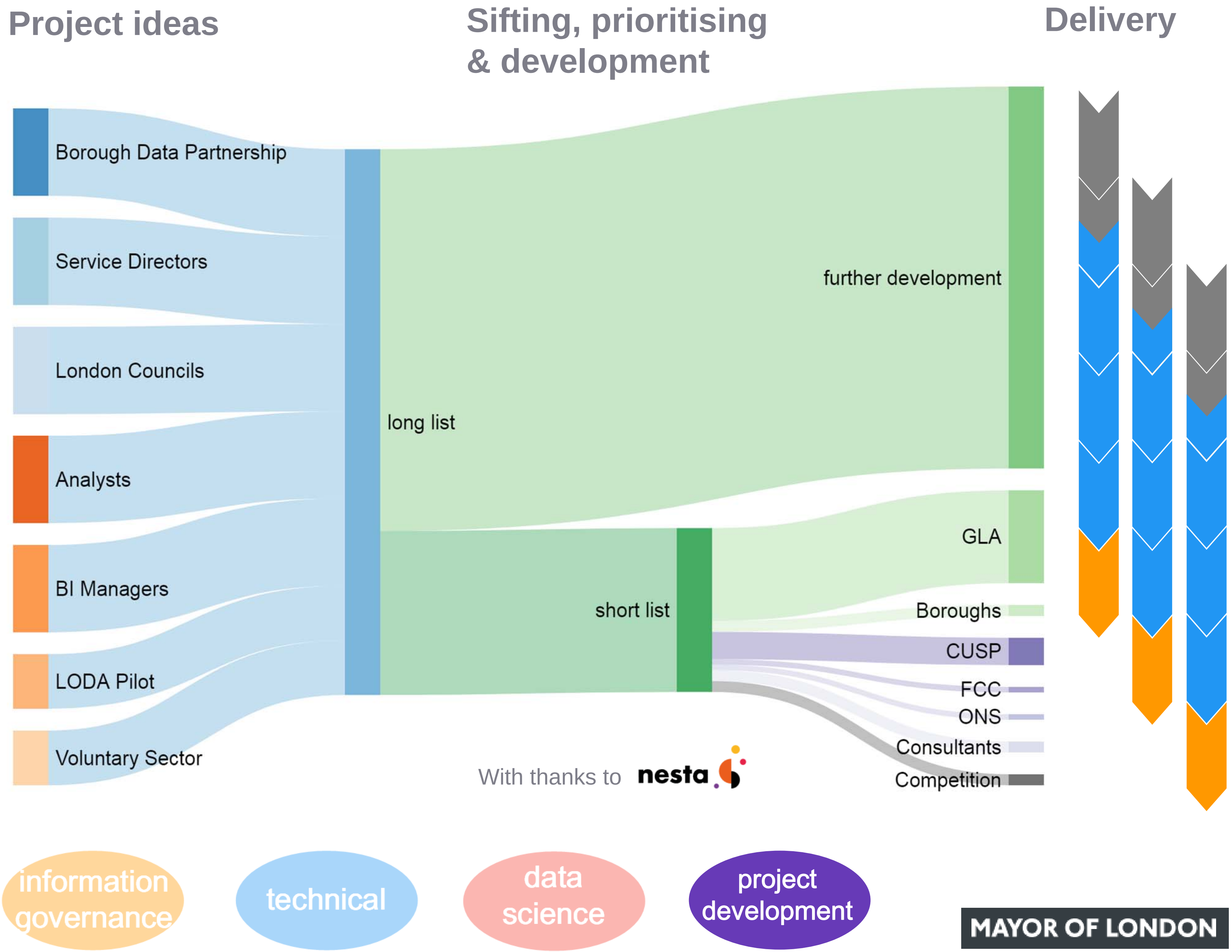
MAYOR OF LONDON

Data Science questions
(in approximate order of difficulty):

- **Descriptive** (can't be generalised)
- **Exploratory** (discover new connections, define future studies)
- **Inferential** (estimate a quantity & uncertainty – e.g. opinion surveys)
- **Predictive** (link not cause – e.g. people who buy lots of strong mints have a higher risk of cancer)
- **Causal** (needs subject knowledge, probably only show average effects e.g. on a neighbourhood rather than a particular household)
- **Mechanistic** (scientific or engineering systems – not applicable to GLA work)

information governance
technical
data science
project development

MAYOR OF LONDON


Project ideas
Sifting, prioritising & development
Delivery
Borough Data Partnership
Service Directors
London Councils
Analysts
BI Managers
LODA Pilot
Voluntary Sector
long list
further development
short list
GLA
Boroughs
CUSP
FCC
ONS
Consultants
Competition
With thanks to nesta
information governance
technical
data science
project development
MAYOR OF LONDON

Project ideas

Sifting, prioritising & development

Delivery

- Data for Service Improvement
- Data for Better Decision Making and Policy
- Urban Experimentation

information governance

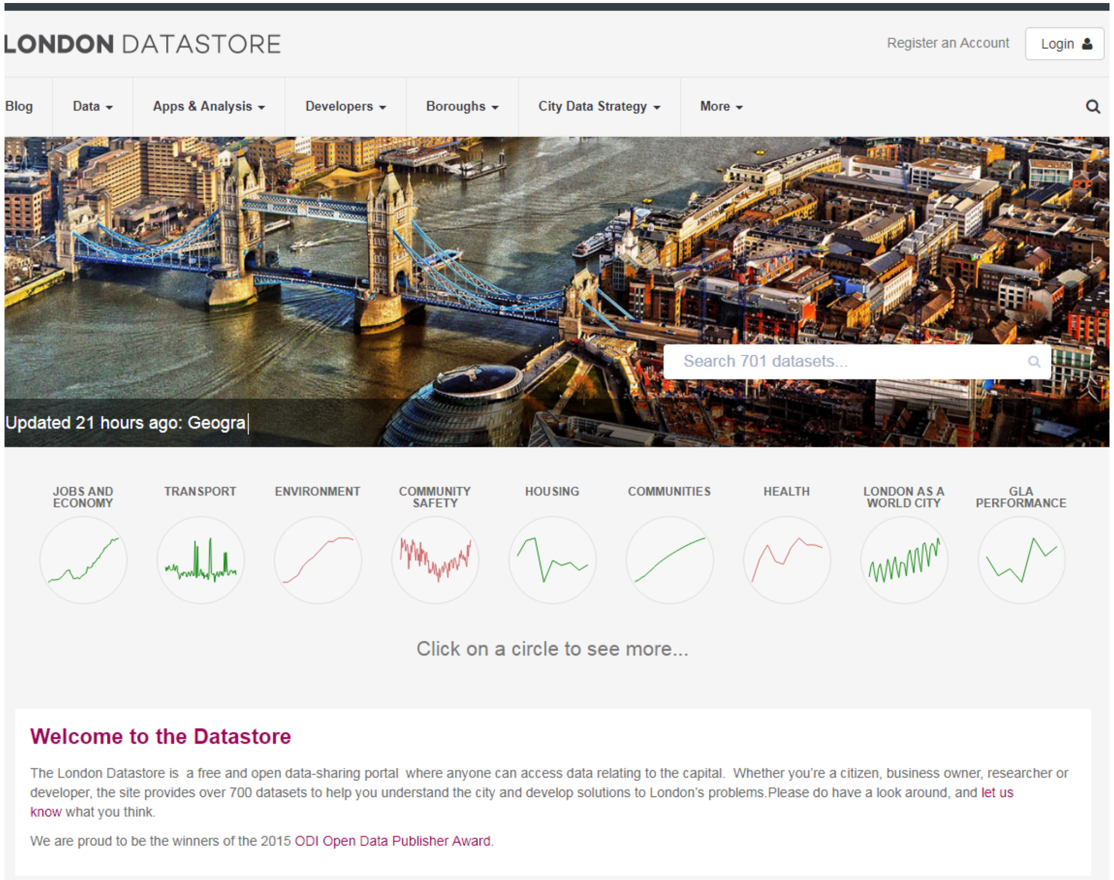
technical

data science

project development

MAYOR OF LONDON

What next?



1. Project Development Board
2. Information Governance Officer (jointly with LfB)
3. Further developments to City DataStore
4. Use London DataStore – resources, data academy, communication

- **1st year of operations**

information governance

technical

data science

project development

MAYOR OF LONDON

Paul Hodgson
GIS & Infrastructure Manager
Greater London Authority
City Hall
The Queen`s Walk
London SE1 2AA

paul.hodgson@london.gov.uk

data.london.gov.uk

MAYOR OF LONDON