Investigation and evaluation of data techniques used on small and large data sets

Using Examples from the UK Charity Sector

By 2020:

- 1.7 Mb of data will be created every second for every human being on earth
- Accumulated data will grow to approximately 44 trillion gigabytes *(Marr, 2018).*

- Greater capacity to store and process this information

PROBLEMS & OPPORTUNITIES

CASE STUDIES:

SMALL DATA

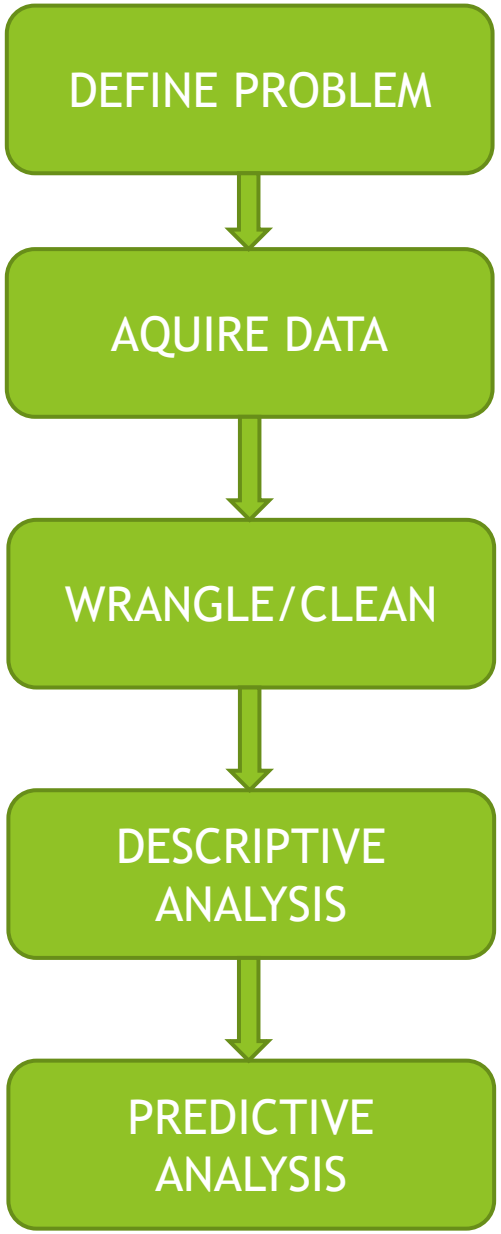
the trussell trust
Stop UK Hunger

LARGE DATA

SMALL DATA

- Structured in nature
- Held locally
- Databases or spreadsheets
- Smaller volumes

BASIC DATA SCIENCE MODEL USED FOR THIS ANALYSIS



Step 1

News

English charities 'get almost 50 per cent less lottery cash per head than Scottish counterparts'

English charities get almost 50 per cent less Lottery cash per head than their Scottish counterparts, it has been claimed.

The Big Lottery Fund handed English causes £510 million last year, the equivalent of £9.32 a head, while in Scotland, they received a total of £76 million, or £14.04 a head.

Scots win the lotto: English charities get 50% less money per head than those north of the border

- **Organisers gave £76million to Scottish causes last year, which is £14.04 a head**
- **But in England the figure was less than £10 a head, with a total of £510million**
- **Tory MP Nadine Dorries is demanding a review over allocation of money**
- **Fellow Conservative MP Andrew Bridgen says the figures are shocking**

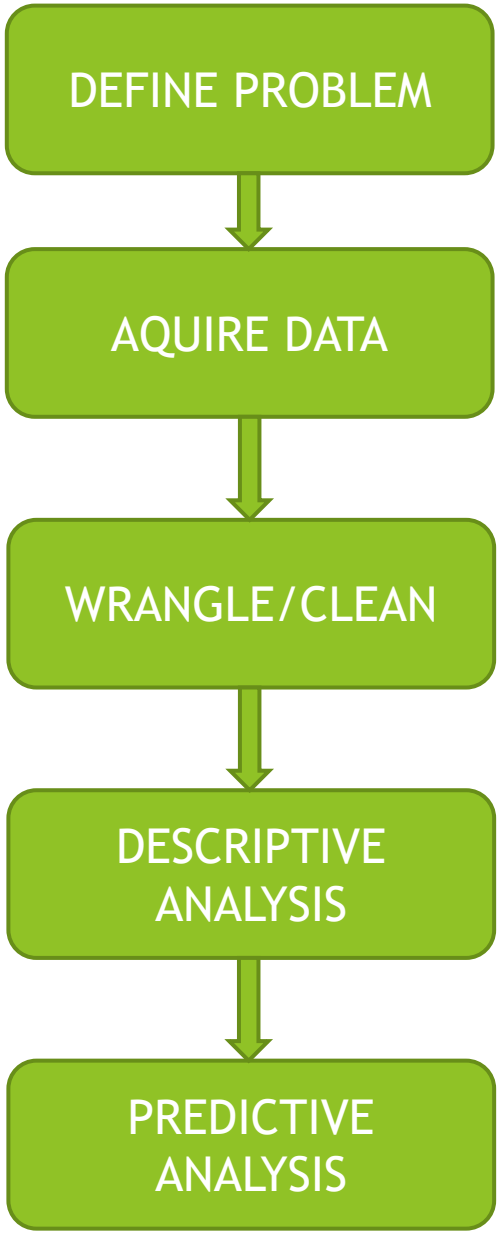
By DANIEL MARTIN POLICY EDITOR FOR THE DAILY MAIL

The annual report from the Big Lottery Fund also shows that spending in Wales, at £14.29 a head, and in Northern Ireland at £14.21 per person, is on a par with Scotland.

Question:

- Does the BBC Children in Need Appeal give most money per head to Scotland (Wales & Northern Ireland) or if not, who gets the most?

BASIC DATA SCIENCE MODEL USED FOR THIS ANALYSIS

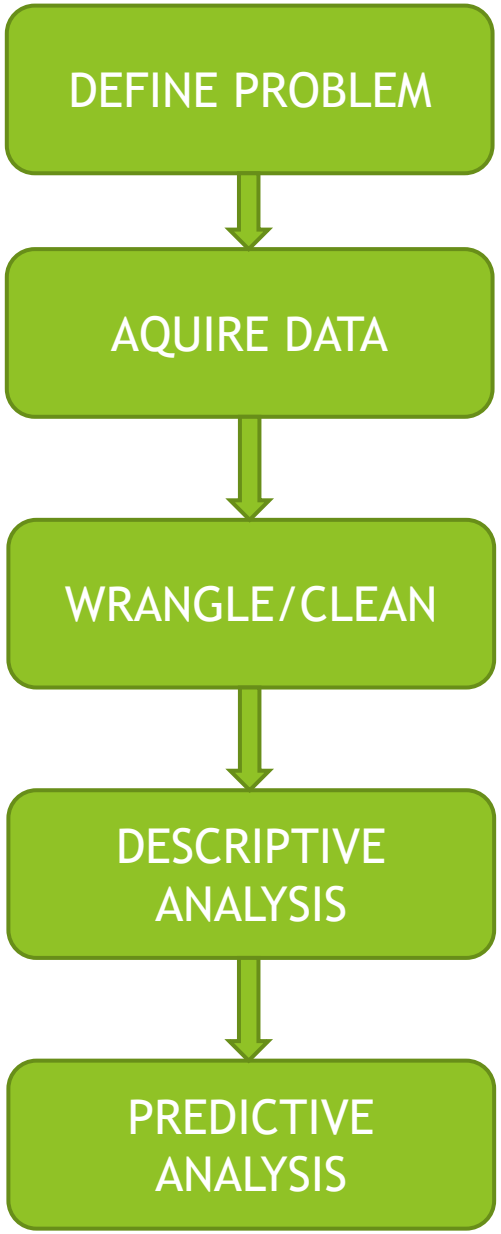


Step 2

Step 2: Acquire Data

- Publishes basic grant data from 94 charities in the UK
- Open and available for people to download and use
- BBC data - 1700 lines, 15 fields in a CSV file

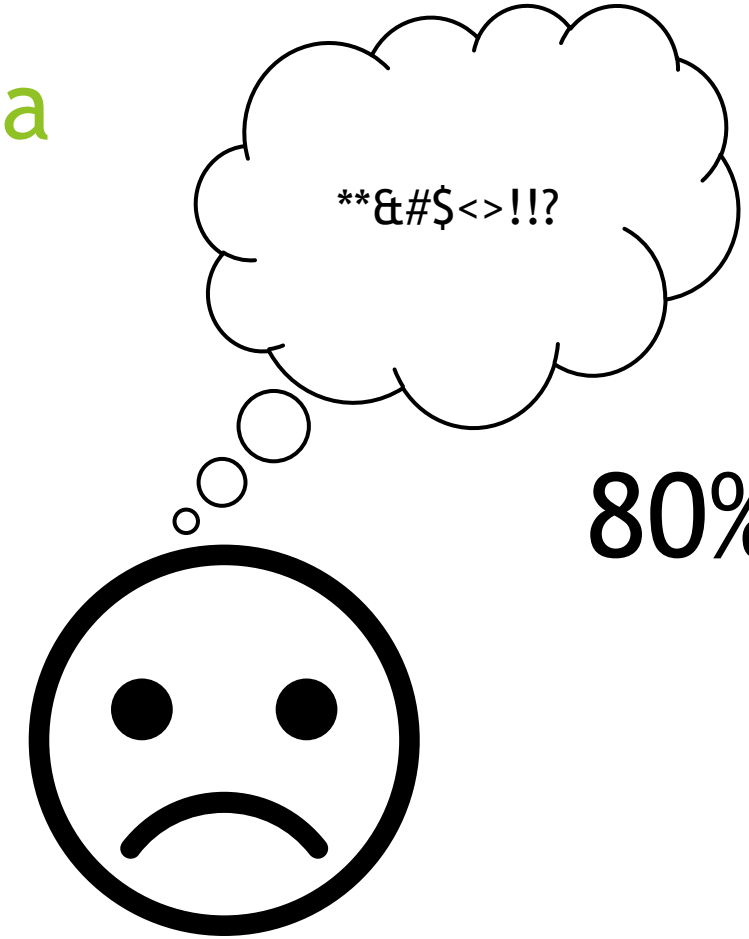
BASIC DATA SCIENCE MODEL USED FOR THIS ANALYSIS



Step 3

Step 3 Cleaning Data

- Remove duplicates
- Find errors
- Missing values
- Formatting problems
- Identify outliers
- Remove unwanted data



80%

```
['Identifier', 'Recipient Org:Identifier', 'Recipient Org:Name', 'Title', 'Recipient Org:Location:Geographic Code Typ
e', 'Recipient Org:Location:Name', 'Recipient Org:Location:Geographic Code', 'Description', 'Amount Awarded', 'Curren
cy', 'Grant Programme:Title', 'Planned Dates:Duration (months)', 'Award Date', 'Funding Org:Name', 'Funding Org:Ident
ifier', 'Last modified 360G-CIN-64988', '360G-CIN-nipitinthebud', 'Nip It In The Bud', 'Grant to Nip It In The Bud',
'UA', 'Rhondda Cynon Taf', 'W06000016', 'This project will provide a programme of activities for children and young p
eople in an area of high deprivation. This will develop their personal and social skills and provide positive opportu
nities and role models.', '26468', 'GBP', 'Small Grants', '36', '01/11/2011', 'BBC Children in Need', 'GB-CHC-802052'
, '2016-10-02T18:00:00Z 360G-CIN-68854', 'GB-SC-SC038627', 'Boomerang Community Centre', 'Grant to Boomerang Communit
y Centre', 'UA', 'Dundee City', 'S12000042', 'This project will run youth clubs for children who live in an area of l
ittle opportunity and could otherwise be involved in anti-social behaviour or substance abuse. The project will incre
ase their confidence and integrate them into the community.', '80340', 'GBP', 'Main Grants', '36', '01/02/2013', 'BBC
Children in Need', 'GB-CHC-802052', '2016-10-02T18:00:00Z 360G-CIN-68857', '360G-CIN-connorstoylibraries', 'Connors T
oy Libraries', 'Grant to Connors Toy Libraries', 'UA', 'Portsmouth', 'E06000044', 'This project provides two communit
y based toy library sessions each week for pre-school children in a deprived area. Through play the children are happ
ier and develop better social skills.', '29745', 'GBP', 'Small Grants', '36', '01/02/2013', 'BBC Children in Need', '
GB-CHC-802052', '2016-10-02T18:00:00Z 360G-CIN-68881', '360G-CIN-homestartknowsley', 'Home-Start Knowsley', 'Grant to
Home-Start Knowsley', 'MD', 'Knowsley', 'E08000011', 'The project will train volunteers to undertake home visits to s
upport families with young children.', '53694', 'GBP', 'Main Grants', '36', '01/02/2013', 'BBC Children in Need', 'GB
-CHC-802052', '2016-10-02T18:00:00Z 360G-CIN-68914', '360G-CIN-newcastleunitedfoundation', 'Newcastle United Foundati
on', 'Grant to Newcastle United Foundation', 'MD', 'Newcastle upon Tyne', 'E08000021', 'The project will provide a 36
week programme of football fun clubs in 10 different venues  including fitness and mobility work throwing passing bal
l control and shooting skills for young disabled people', '75000', 'GBP', 'Main Grants', '36', '01/02/2013', 'BBC Chi
ldren in Need', 'GB-CHC-802052', '2016-10-02T18:00:00Z ']
```

- Python code used to clean data
- Removing whitespaces with Strip()
- For loops and If statements to clear strings like ‘GBP’
- Regular expressions (REGEX) to clear indicator codes
- Test set 2000 lines
- Also problems with commas and apostrophes

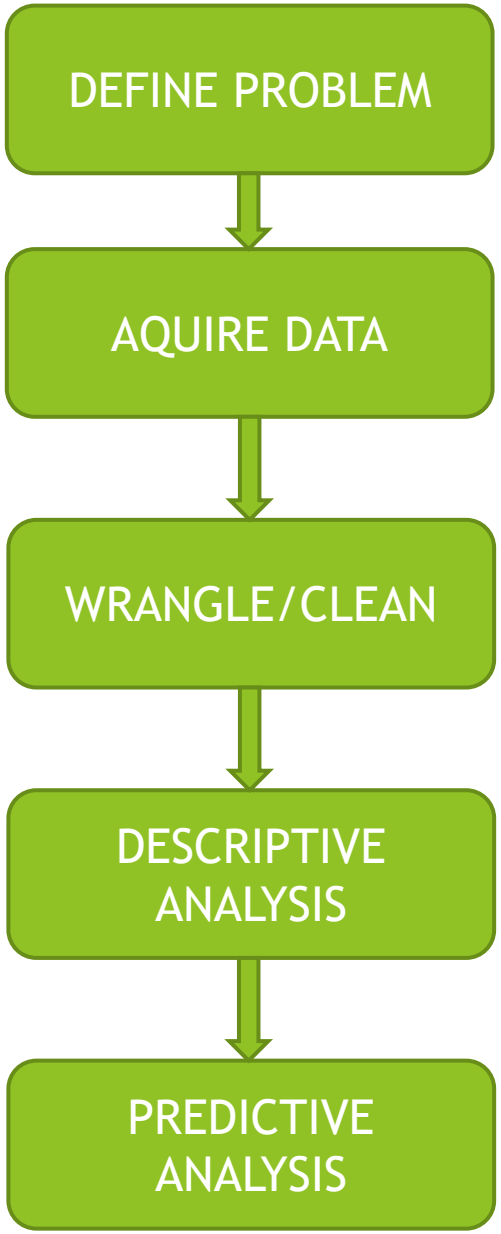
Regular Expressions Examples

- **Target 1: Identifier**
- Pattern – GB-CHC/COH/CDC/EDU/NIC-000000
- Regular Expression:
- [r'GB\-[CHC|COH|EDU|NIC|] \-[\w] +')

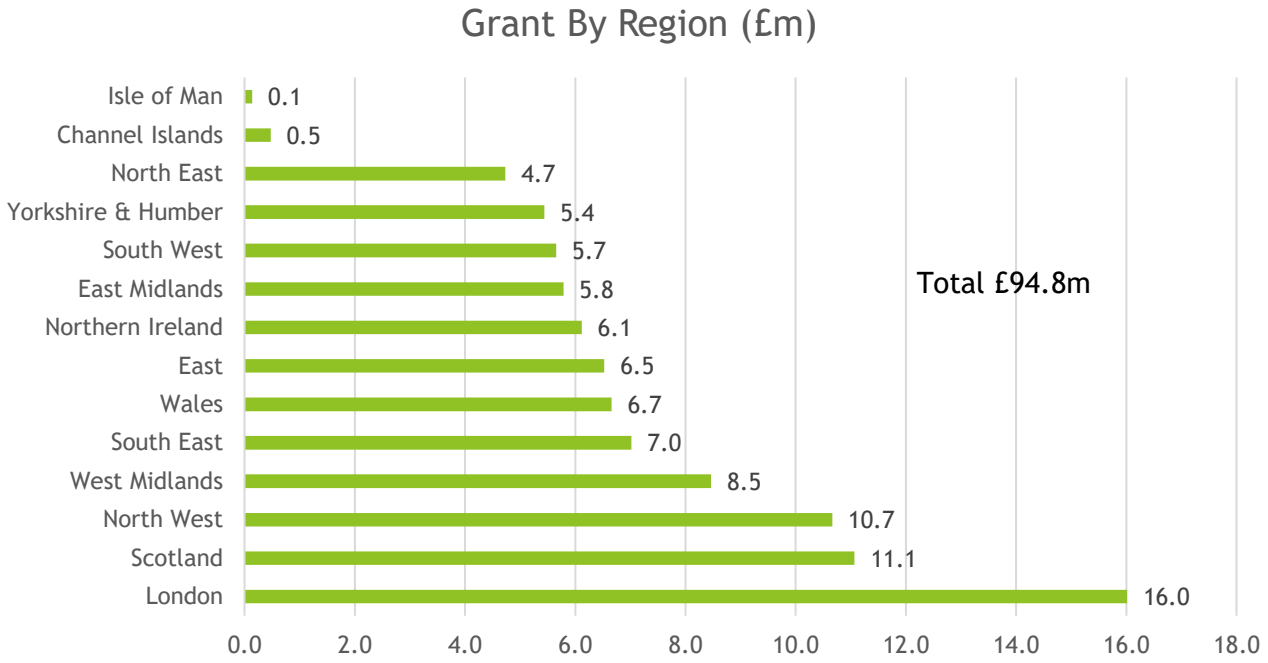
- **Target 2: Government Location Code**
- Pattern - E0000000
- Regular Expression:
- (r'[A-Z] {1} \d {7}')

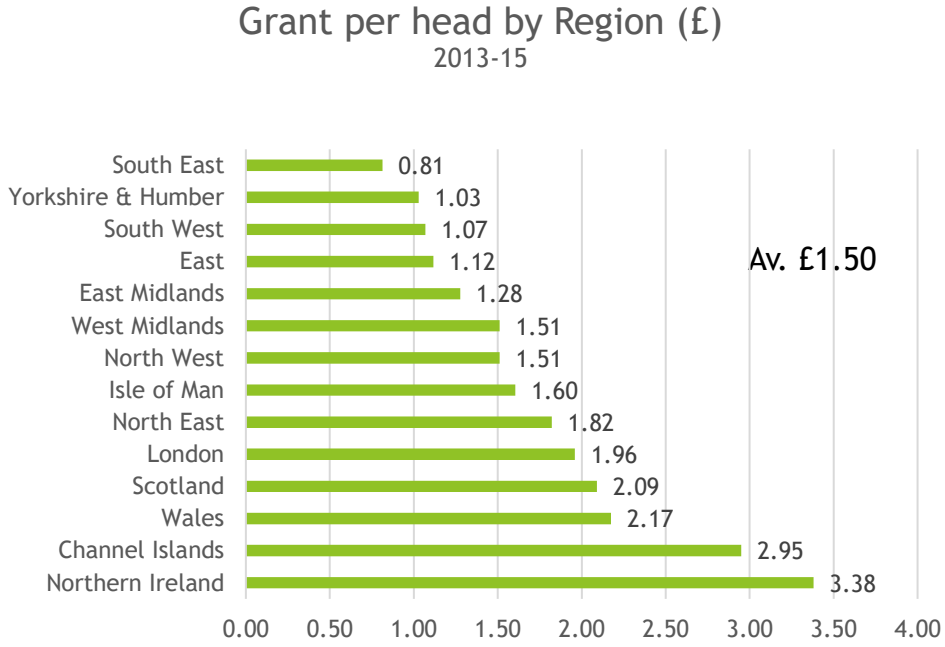
- **Target 3: Local Authority Type**
- Any of MD, LGD, LOND, NMD,UA, #N/A or #N\A
- Regular Expression:
- (r'#N\\A|#N\/A|LGD|LONB|MD|NMD|UA')

BASIC DATA SCIENCE MODEL USED FOR THIS ANALYSIS



Step 4


Grant By Region (£m)
Isle of Man 0.1
Channel Islands 0.5
North East 4.7
Yorkshire & Humber 5.4
South West 5.7
East Midlands 5.8
Northern Ireland 6.1
East 6.5
Wales 6.7
South East 7.0
West Midlands 8.5
North West 10.7
Scotland 11.1
London 16.0
0.0 2.0 4.0 6.0 8.0 10.0 12.0 14.0 16.0 18.0
Total £94.8m


Grant per head by Region (£)
2013-15
South East 0.81
Yorkshire & Humber 1.03
South West 1.07
East 1.12
East Midlands 1.28
West Midlands 1.51
North West 1.51
Isle of Man 1.60
North East 1.82
London 1.96
Scotland 2.09
Wales 2.17
Channel Islands 2.95
Northern Ireland 3.38
0.00 0.50 1.00 1.50 2.00 2.50 3.00 3.50 4.00
Av. £1.50

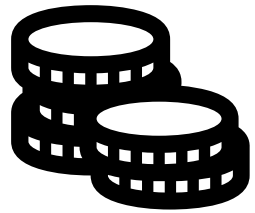
Town/City Level Analysis - Top and Bottom Five Grant/Head

Top Five	£/Head
Camden	**10.86**
Belfast	**10.63**
Islington	**7.69**
Derry and Strabane	**6.73**
Norwich	**6.66**

Bottom Five	
Telford and Wrekin	0.06
Blaby	0.05
Calderdale	0.05
Chiltern	0.05
East Renfrewshire	0.04

Havant	£0.08 – Actually the 8th lowest overall
Chichester	£0.64
Portsmouth	£1.44
Bognor	No grants!

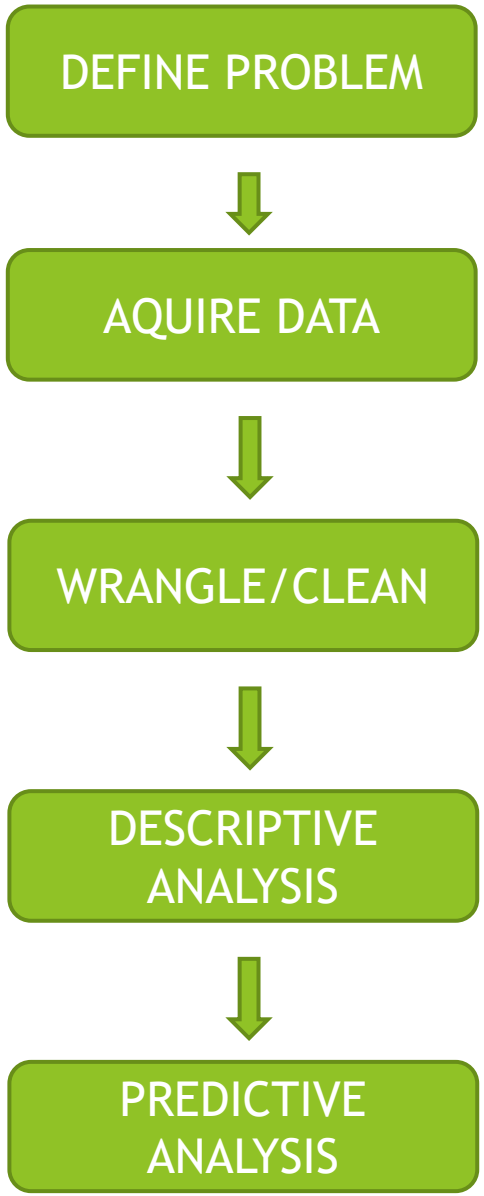
- Does the BBC Children in Need Appeal give most money per head to Scotland (Wales and NI) like the National Lottery or if not, who gets the most?



- Northern Ireland £3.38
- Channel Islands £2.95
- Wales £2.17
- Scotland £2.09
- England £1.32

- Wales £14.29
- Northern Ireland £14.21
- Scotland £14.04
- England £ 9.32

Most to Northern Ireland, picture is worse for England



BASIC DATA SCIENCE MODEL USED FOR THIS ANALYSIS

Predictive Analysis:

Using historic data to predict or forecast future outcomes

Step 5

Evaluation

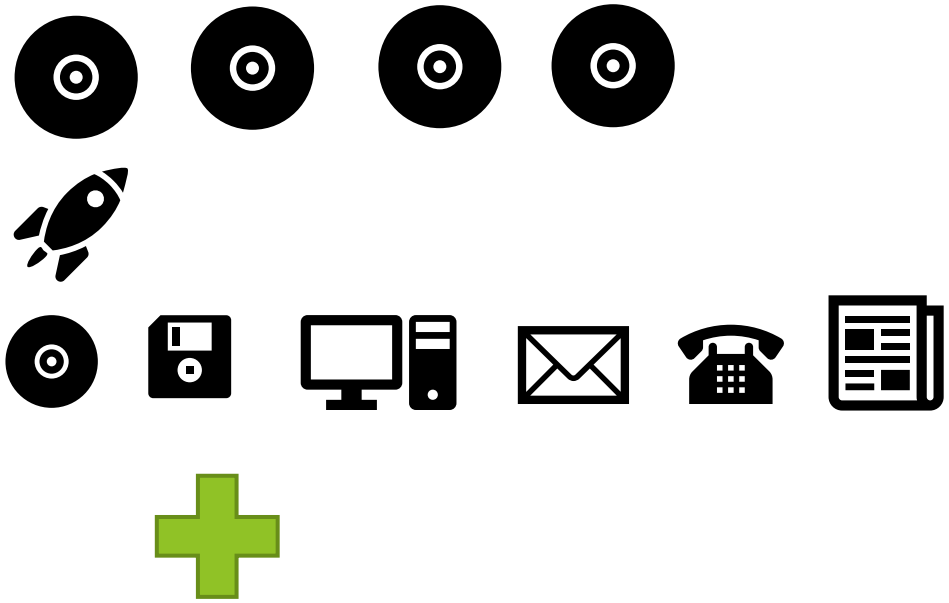
- Traditional approach to grant allocation using local assessment of bids not transparent.
- Projects that write good bids get more
- Some organisations don’t apply
- Predictive analysis could enable them to be proactive and target resources effectively
- Have advertised for Data Analysts so likely to be on their minds.
- Combine data with Government socio-economic data get more insights
- Share data with other child action charities to identify best causes to fund and build models

BIG DATA

VOLUME

VELOCITY

VARIETY



VERACITY

VOLATILITY

VALIDITY

BIG DATA

Support Centres

MACMILLAN CANCER SUPPORT Information and support Get involved

Online Community

Supporter Care Hub
0300 1000 200
Monday to Friday, 9am-5pm

Local Groups

IN YOUR AREA

MACMILLAN CANCER SUPPORT

Need to talk?
0808 808 00 00
Monday to Friday, 9am-8pm
Call us free* >

TAKE PART WHEREVER YOU ARE

THIS IS A NATIONAL EVENT

facebook

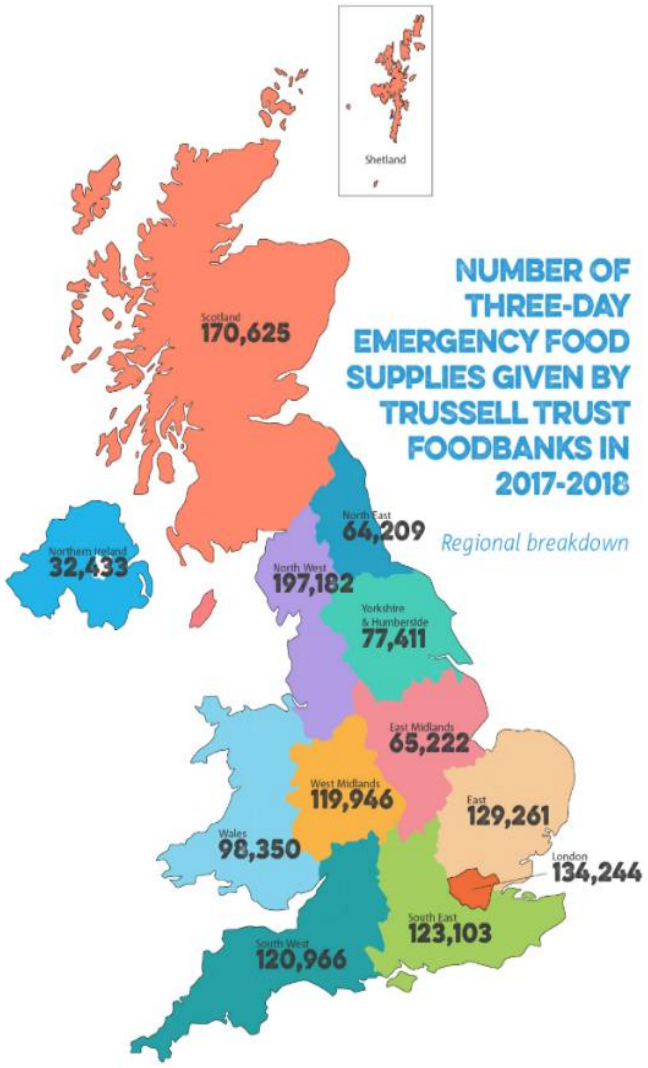
WE ARE MACMILLAN. CANCER SUPPORT
WORLD'S BIGGEST COFFEE MORNING

THE SHOP MACMILLAN CANCER SUPPORT

Trussell Trust

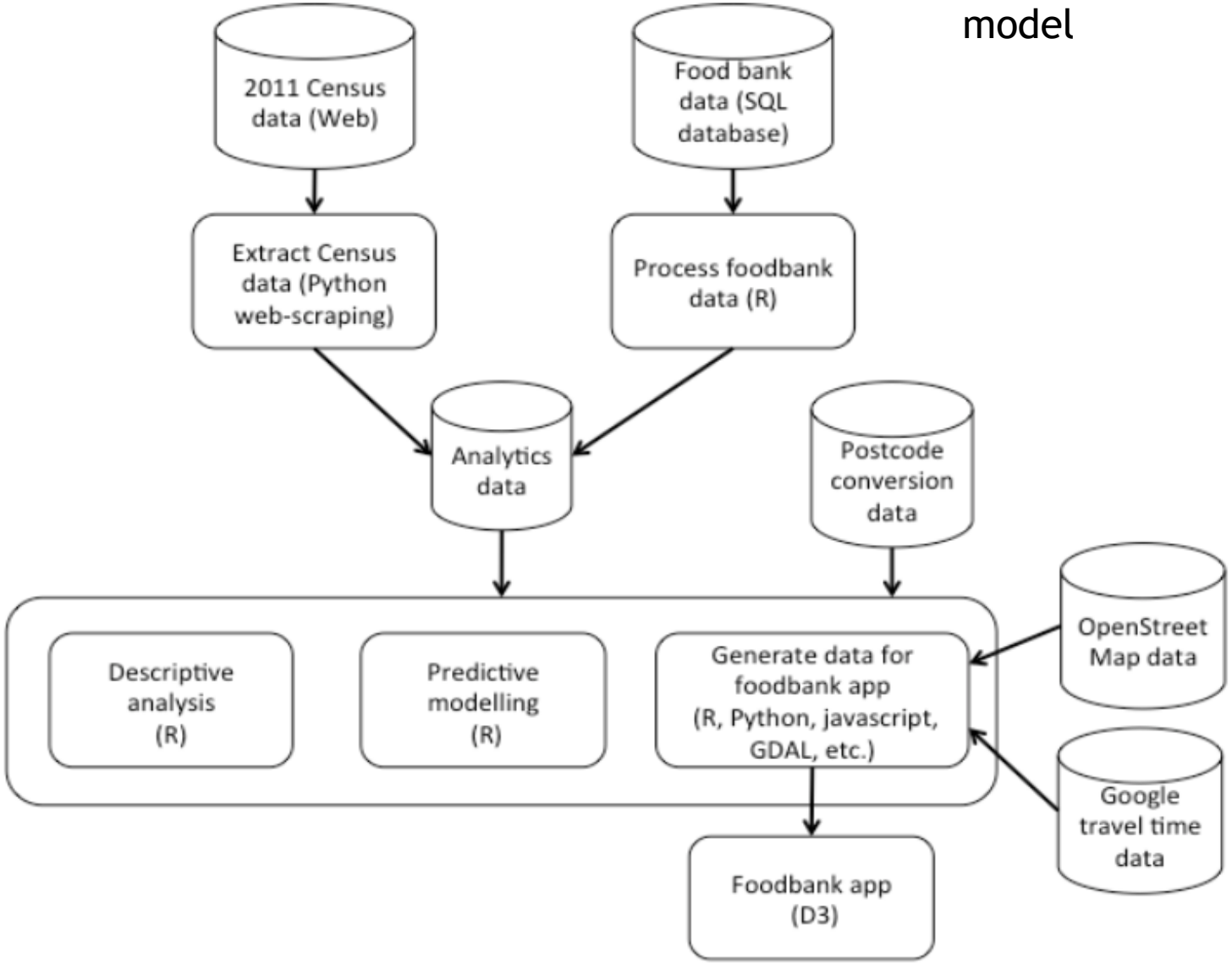
Building a Predictive Model

1,332,952 three-day emergency food supplies given to people in crisis 2017/18



- 420 food banks
- 1200 distribution centres
- Increase in demand of 13% in 2018
- The largest food bank organisation but there are others!

Building a predictive model

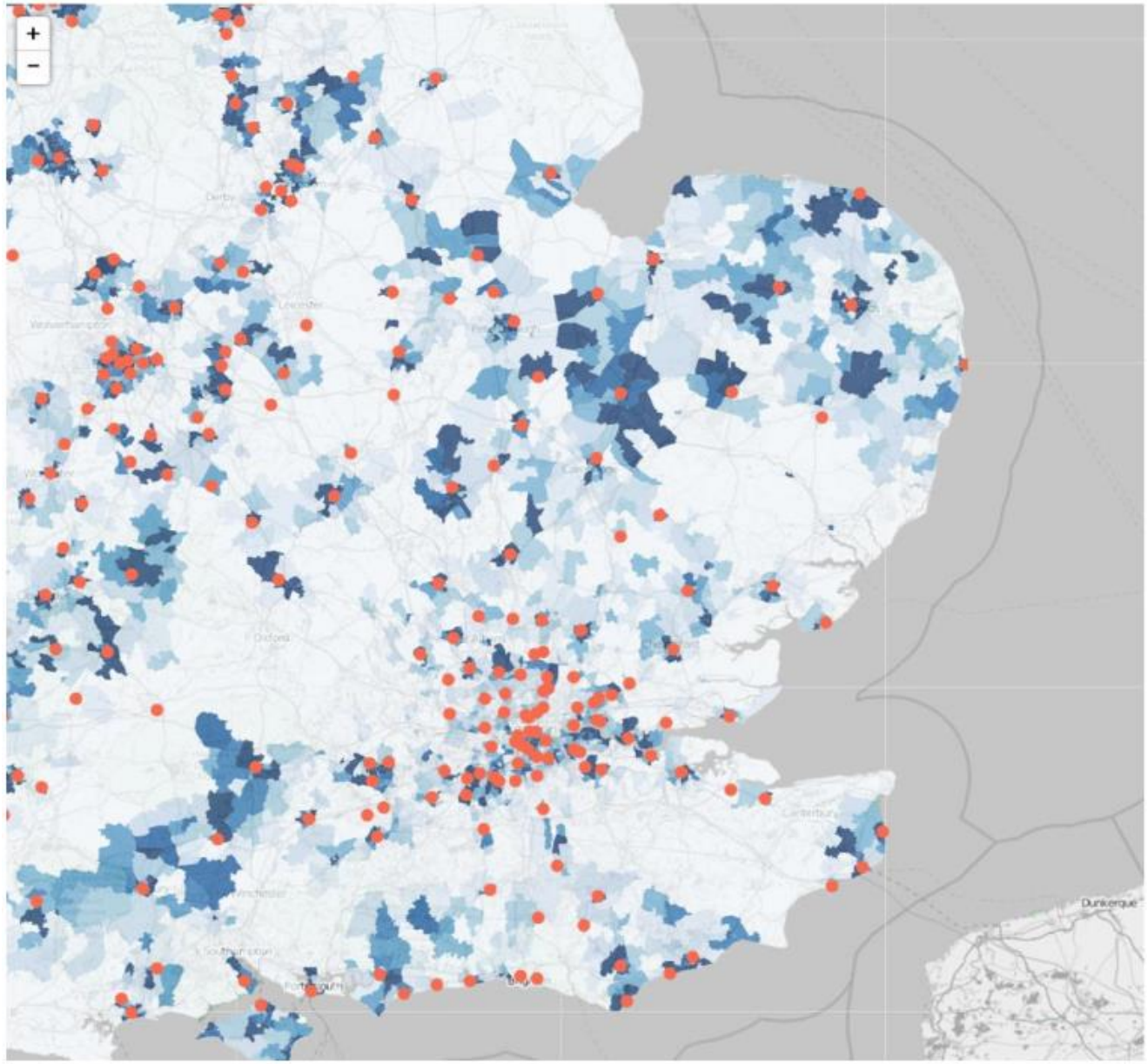


Trussell Food Bank Model Part 1

- Worked with Hull University
- Identified 7 main drivers of food bank use one of which maturity of a food bank
- Developed a standard model for a typical food bank based on Bayesian Statistical Model
- Typical food bank serves 50 customers a week, takes 6 months to maturity and over Christmas and New Year will provide double the yearly average

Trussell Food Bank Model Part 2

- Built a logistic regression model
- Based on 11 food bank data variables
- Also used 30 Census socio-economic variables
- Used the data to build an interactive map for predicting food bank demand


+
−
Heat map
Actual use
Predicted need
Crisis
Select the crisis type
Benefit delays
Opacity
Foodbanks
Visible
Reach
(At the moment we only have actual customers)
Actual customers
Drive time
Walking
Transit
Legend
Explanation goes here
0.0
0.53
1.1
1.6
2.1
2.6
3.2
3.7
4.2

Trussell Interactive Map

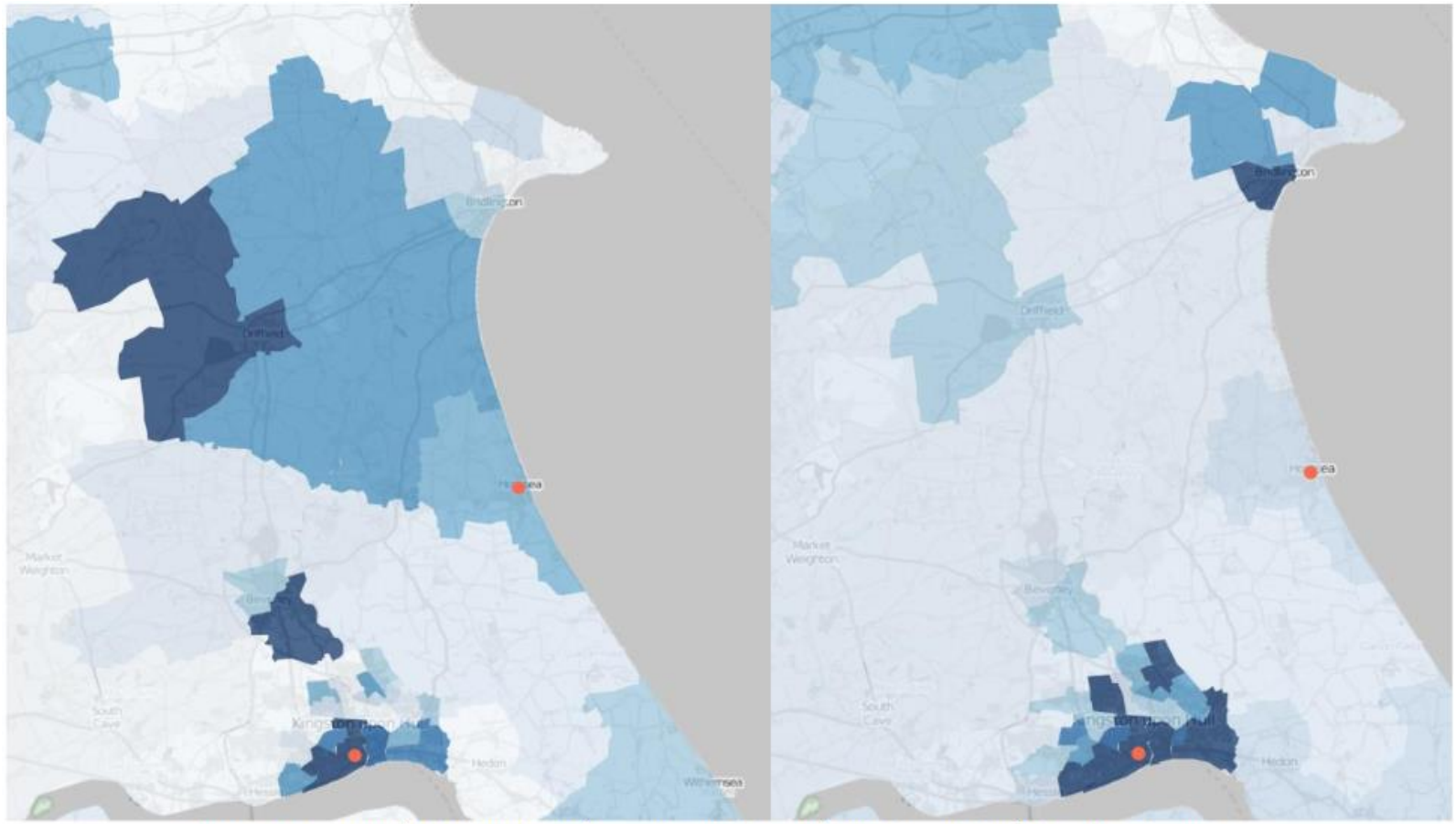


Figure 12: actual (left hand side) versus predicted (right hand side) food bank usage

Evaluation of the Model

- Works well when the environment is stable but food bank variables can be volatile
- Other issues: weather, seasonal migration, local economic factors not included in the model
- Government policy : Universal Credit
- Multicollinearity if variables are related to each other - Census variables likely to be!
- Overfitting if it models the training data too well and can increase with greater complexity if number of parameters is too high
- Need to refine the model, test and adjust it if necessary
- Support with professional knowledge in this area

Sharing Data – The Issues

- Sharing data with other charities e.g. single elderly known to have food poverty, elderly charities have data?
- Need to take account of GDPR – retention of data, right to be forgotten, consent to hold data
- 11 charities recently fined for breaching Data Protection Act in respect of donor lists
- Also ethical concerns – high profile cases like Olive Cooke who received 3000 mailings per year from charities
- So need to be clear about what data is shared, higher level operational is probably ok but donor data that is not anonymised is not
- Need to comply with the law!

Need For National Statistics

- Need for National Food Insecurity or Poverty Index like the Fuel Poverty Indicator? Other countries do this
- National food bank database, lots of organisations delivering and no one source of data to inform policy or build models
- Government take the lead? Committed to UN Sustainable Development Goal to reduce hunger

Targeting Donors

- Another whole area of predictive analytics to target high value donors
- Big business in US!
- Not yet in UK but again GDPR relevant here and the Data Protection breaches likely to have had a cooling effect on this area

Summary

- Recommended that:
  - Charities move towards using predictive methods to identify needy groups and target funds accordingly
  - As predictive models can be very complex and those involving socio-economic indicators especially so – charities exercise caution in their use, test and monitor the models and support with professional knowledge of the sector!
  - Work together to inform and lobby Government to develop a national measure of food insecurity and a database of foodbank usage

HELP CHARITIES, HELP GOVERNMENT AND HELP THOSE IN MOST NEED

Thank You For Listening