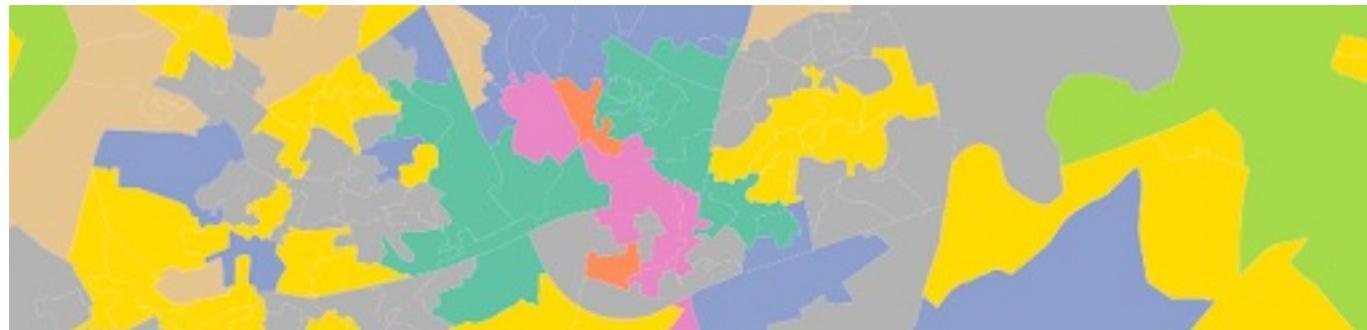


Principles of Spatial Analysis

WEEK 07: GEODEMOGRAPHICS



This week

- Geodemographic classifications
- k-means clustering
- Bonus section on speeding up code in R

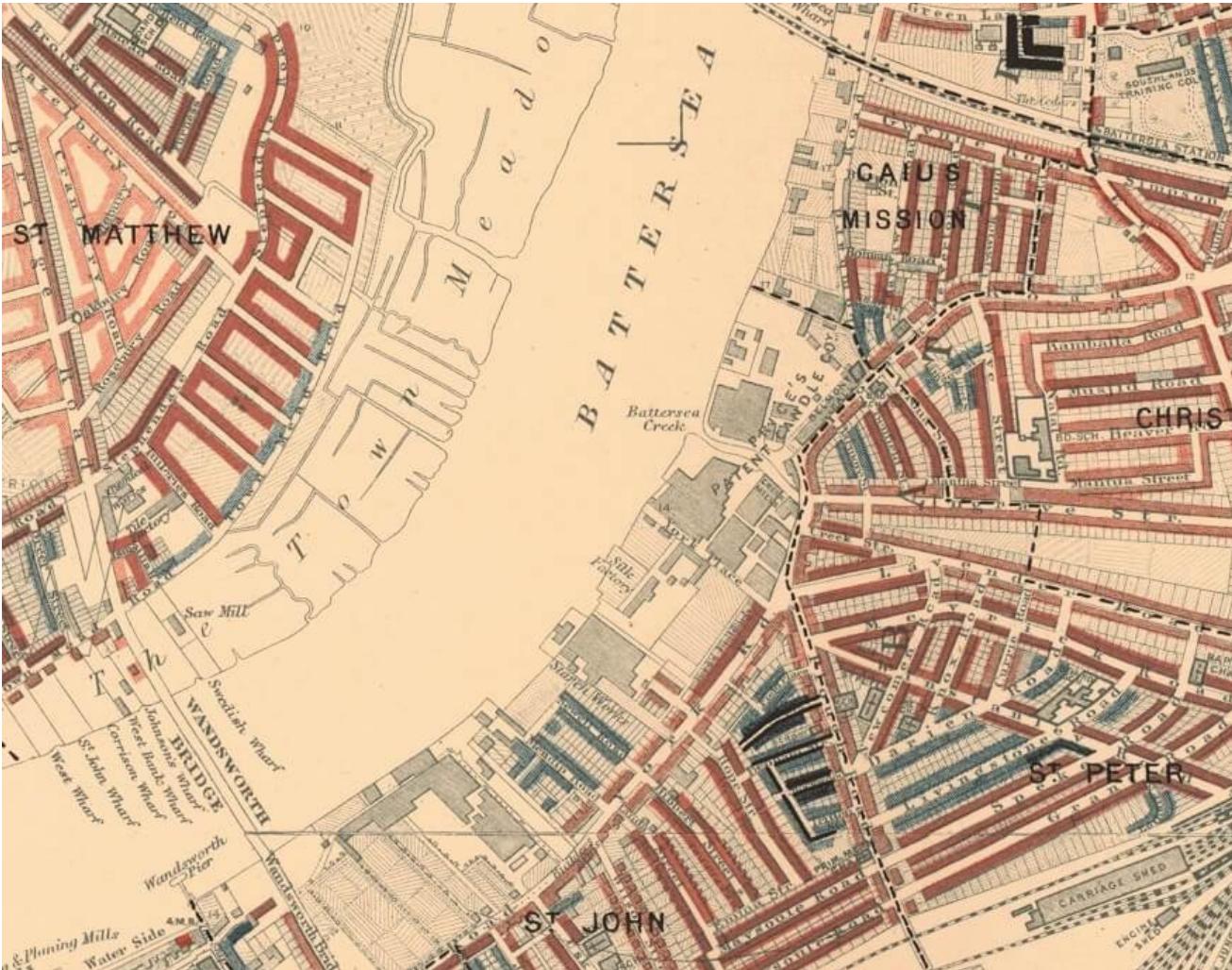
Geodemographics

- Analysis of people (*demographics*) by where they live (*geo*).
- Used to identify similar neighbourhoods or administrative areas.
- Means of multivariate data reduction for the differentiation of areas.
- Been around for many many years, dating back to the late 1800s.

Charles Booth

- Created the first geodemographic-style classification.
- Shipping business owner and philanthropist.
- Survey: *Life and Labour of People in London*.
- Mostly qualitative analysis by walking through areas.
- Books and books and books with notes.
- He noticed that there is a geographical pattern in the distribution of different social categories: people who live in a particular neighbourhood and share similar living conditions, are of similar characteristics and social status.

Charles Booth



Maps Descriptive of London Poverty 1889. Charles Booth's *Inquiry into Life and Labour in London* (1886-1903).

Charles Booth

Classification	Colour	
Lowest class. Vicious, semi-criminal.	Black	
Very poor, casual. Chronic want.	Dark blue	
Poor. 18s. to 21s. a week for a moderate family.	Light blue	
Mixed. Some comfortable others poor.	Purple	
Fairly comfortable. Good ordinary earnings.	Pink	
Middle class. Well-to-do.	Red	
Upper-middle and upper classes. Wealthy.	Yellow	

Geodemographics

- Further developed in 1970s to target urban deprivation funding.
- Commercial sector also got involved (CACI ACORN / Experian MOSAIC).
- Office for National Statistics' Output Area Classification (2001, 2011, 2021).
- ONS' Output Area Classification completely open using Census data.

Geodemographics

Singleton *et al.* 2020:

"A geodemographic classification is created by assembling a wide range of measures that describe the characteristics of areas and/or those people living within them, and then, through the implementation of unsupervised learning (clustering), identifies groups of areas that share common characteristics. Emerging clusters may be divided or aggregated to create a hierarchy, and it is typical that these be accompanied by labels, descriptions, photographs, diagrams and graphs."

Variable selection

Choice of measures can create bespoke classifications:

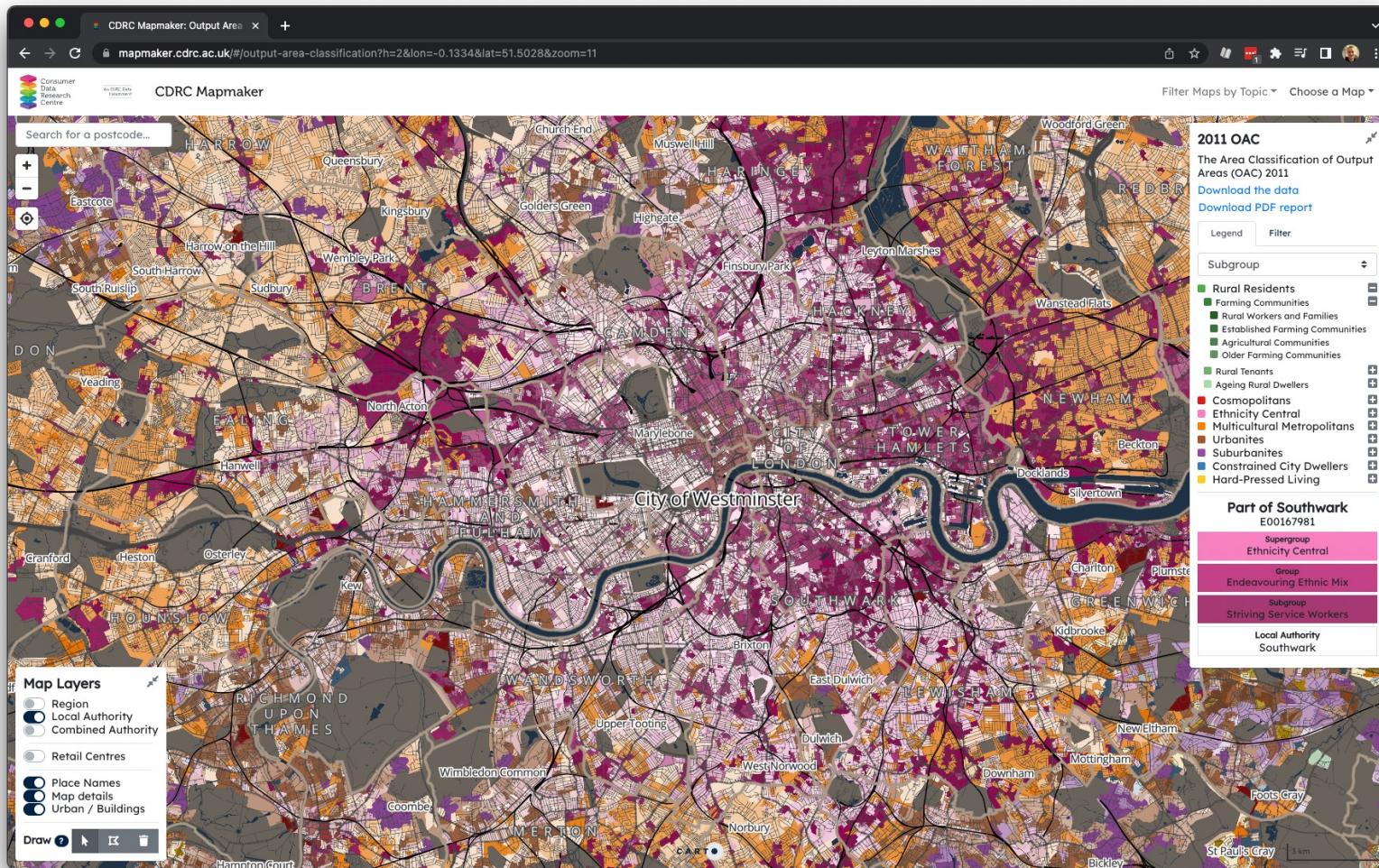
- Output Area Classification: 60 census variables, focused on holistic understanding of area characteristics.
- Workplace Zones Classification: 48 census variables across 4 domains, focused on workers and workplaces.
- Internet User Classification: built from a range of consumer, survey and open (census) data focusing on how populations interact with the internet (use and engagement).

Clustering

Variables are partitioned into clusters using the k-means clustering algorithm and then mapped to their respective areal unit, e.g. Output Area, Workplace Zone:

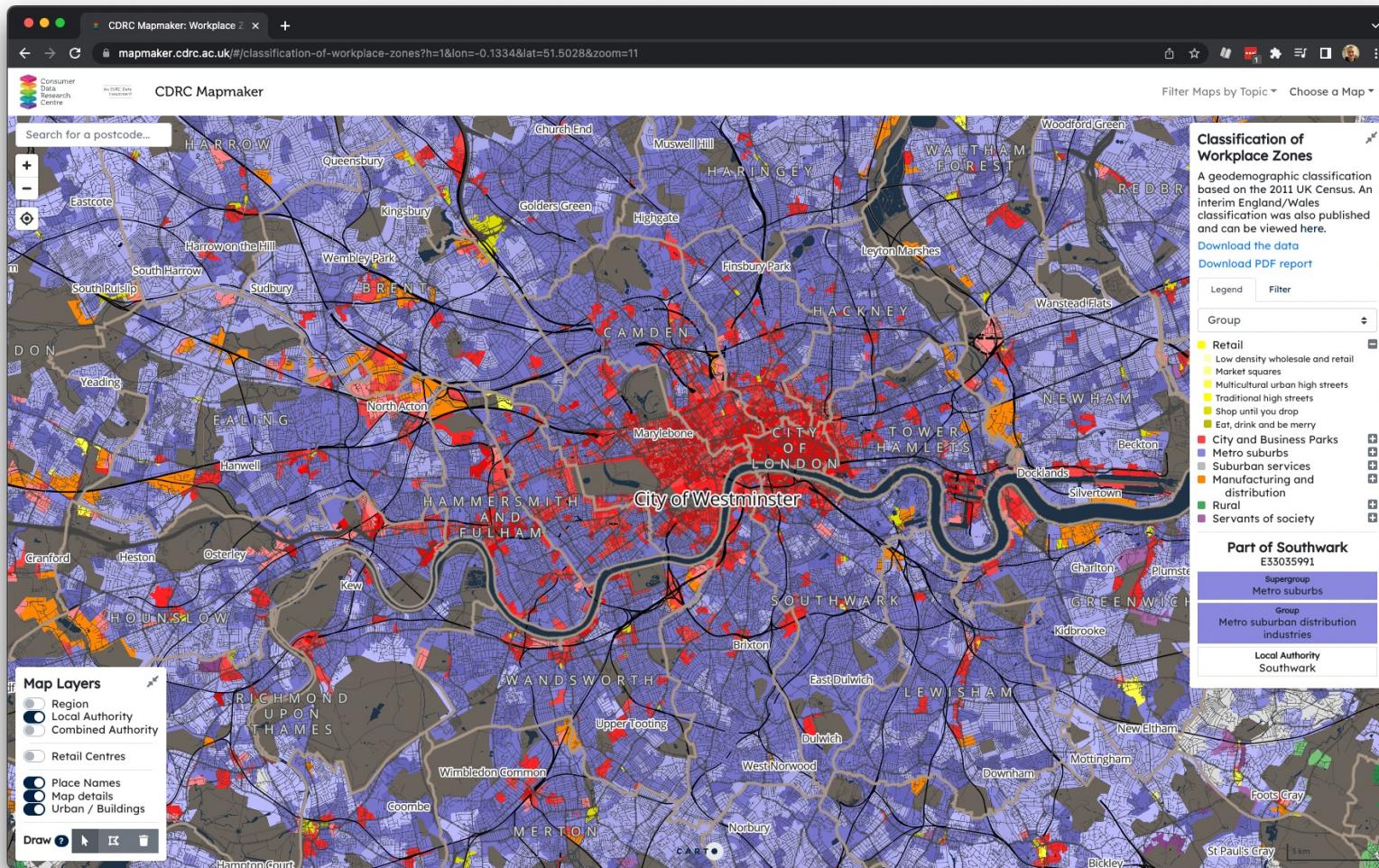
- Sometimes there is a hierarchy of clusters (e.g. supergroups, clustergroups).
- Each cluster will have a name and description (pen portrait).

Geodemographics



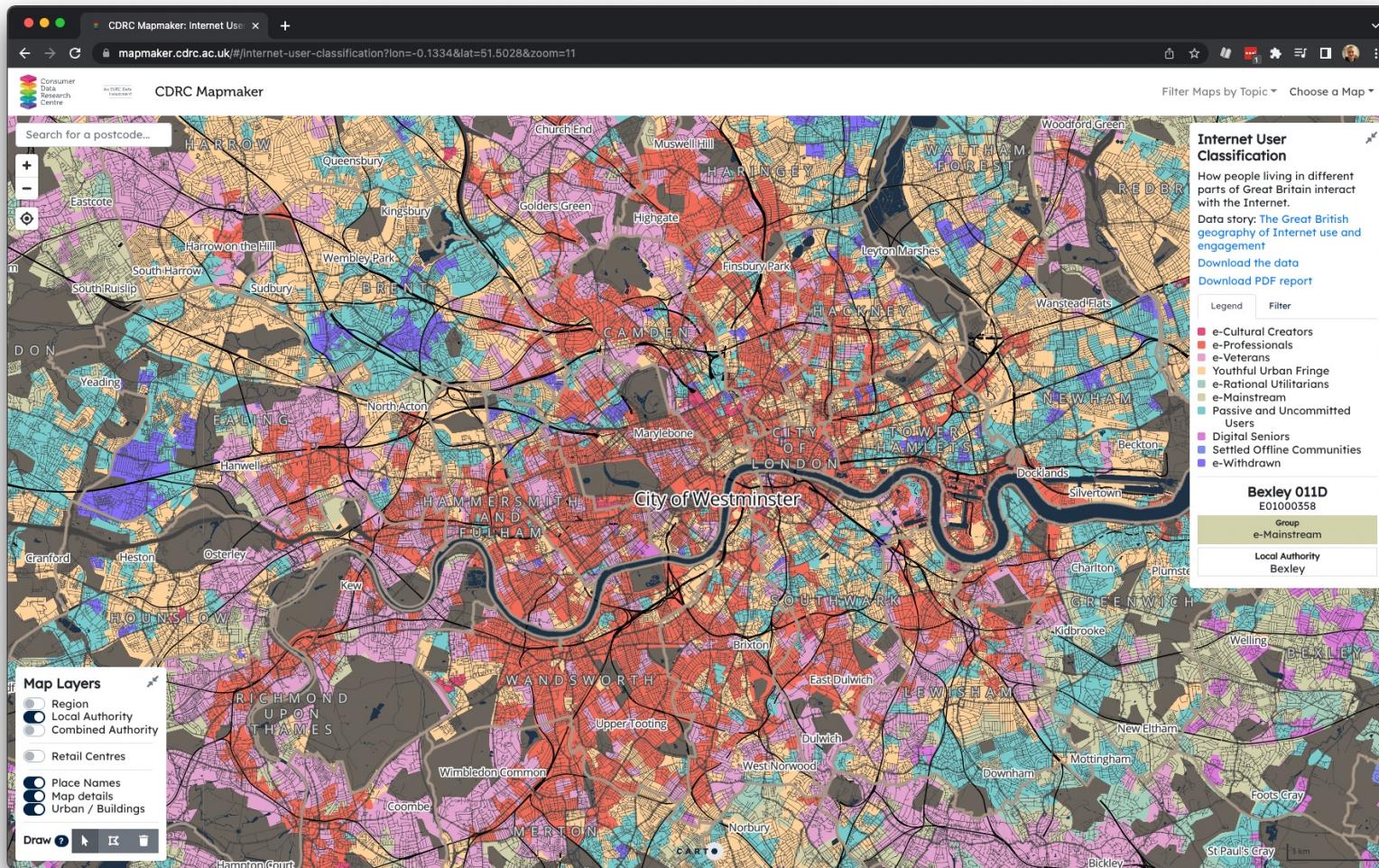
2011 Output Area Classification on mapmaker.cdrc.ac.uk

Geodemographics



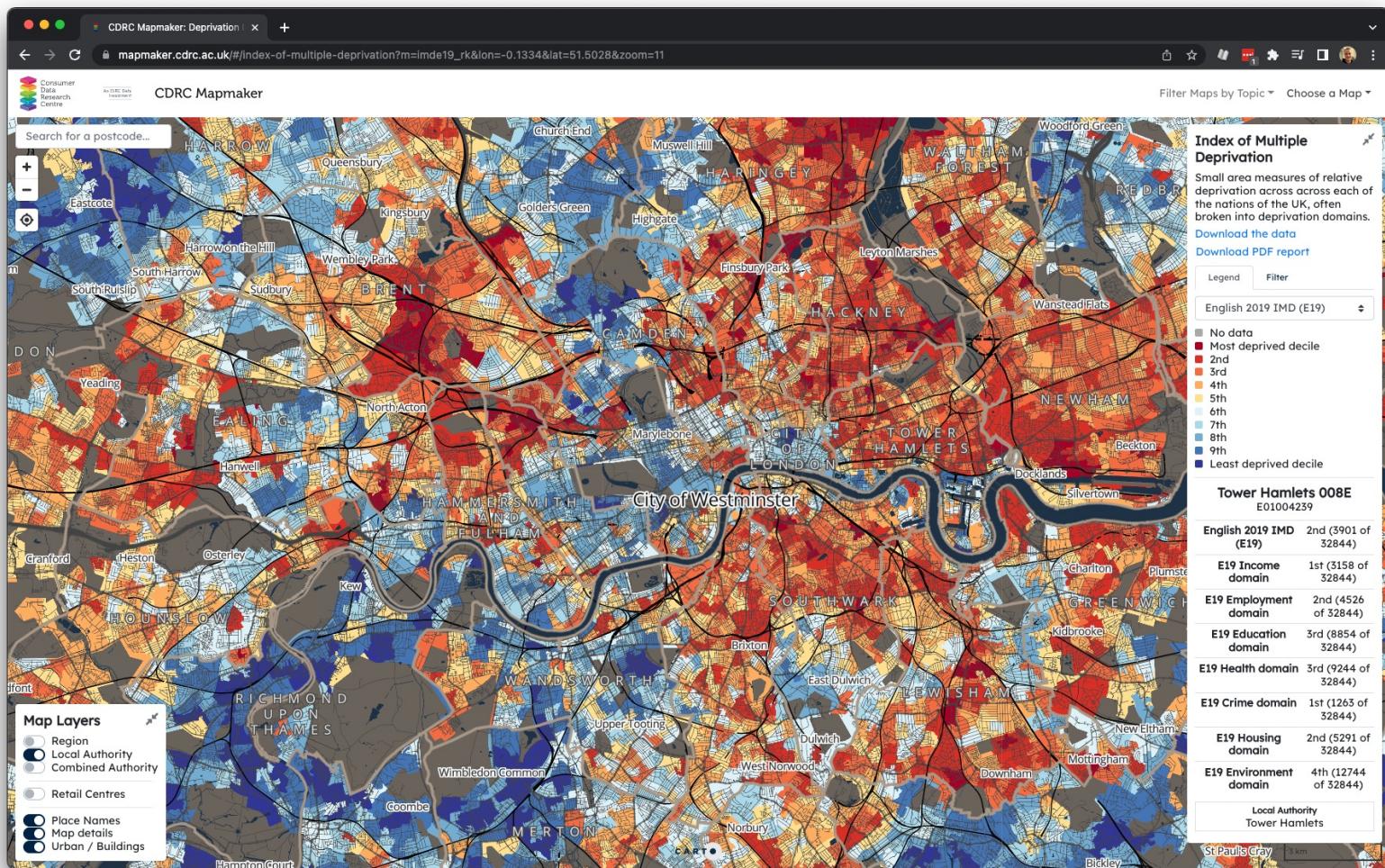
Workplace Zone Classification on mapmaker.cdrc.ac.uk

Geodemographics



Internet User Classification on mapmaker.cdrc.ac.uk

Indices



Limitations

- Highly dependent on the input data.
- Input data can get old very quickly (depending on the topic).
- Inherent biases within the input data – see also the article by Dalton and Thatcher (2015) on the Data, Politics Society reading list.

Applications

Using the geodemographic classification as input for further analysis:

- Harris *et al.* 2007: differences in school choice between social groups
- Brundson *et al.* 2011: participation in higher education
- Martin *et al.* 2018: analysis of travel-to-work flows
- Goodman *et al.* 2011: socio-economic inequalities in exposure to air pollution
- Trasberg and Cheshire 2021: profiling changing activity patterns

Internet user classification

Singleton *et al.* 2020:

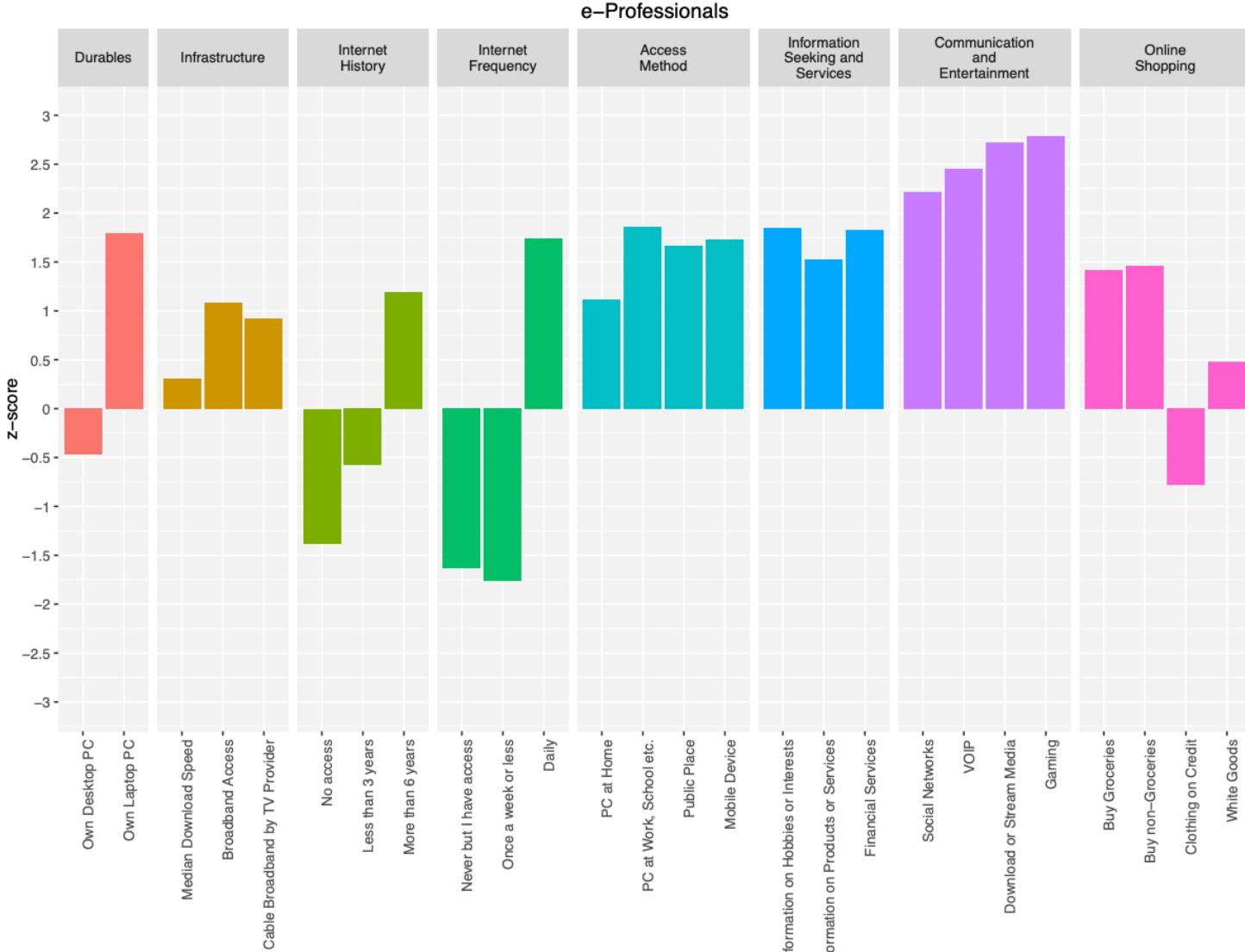
- bespoke classification created by CDRC researchers
- how do populations interact with the internet
- 'profiles of internet use and engagement'
- built from a range of consumer data, survey data, and open data
- classification is available as open data

Internet user classification

Several noteworthy variables:

- British Population Survey: internet access, frequency of internet usage, access to PC, type of internet use
- transactional (consumer) data on online shopping
- average broadband speed
- census variables such as age, ethnicity
- National Statistics Socio-economic classification (NS-SEC)

Internet user classification



Internet User Classification mean attributes of the *e-Professionals*

Internet user classification

e-Professionals:

"This Group has high levels of Internet engagement, particularly regarding social networks, communication, streaming and gaming, but relatively low levels of online shopping, besides groceries. They are new but very active users, with a very high proportion of the population engaging on a daily basis. (...) Geographically, this Group is mainly located close to the city centre or within the proximity of Higher Education Institutes, where infrastructure accessibility, such as cable broadband, is sufficient"

Internet user classification

- Measures of access to and use of internet.
- Identification of areas to target potential interventions.
- Analysis of areas where people are likely to work from home.

Workflow

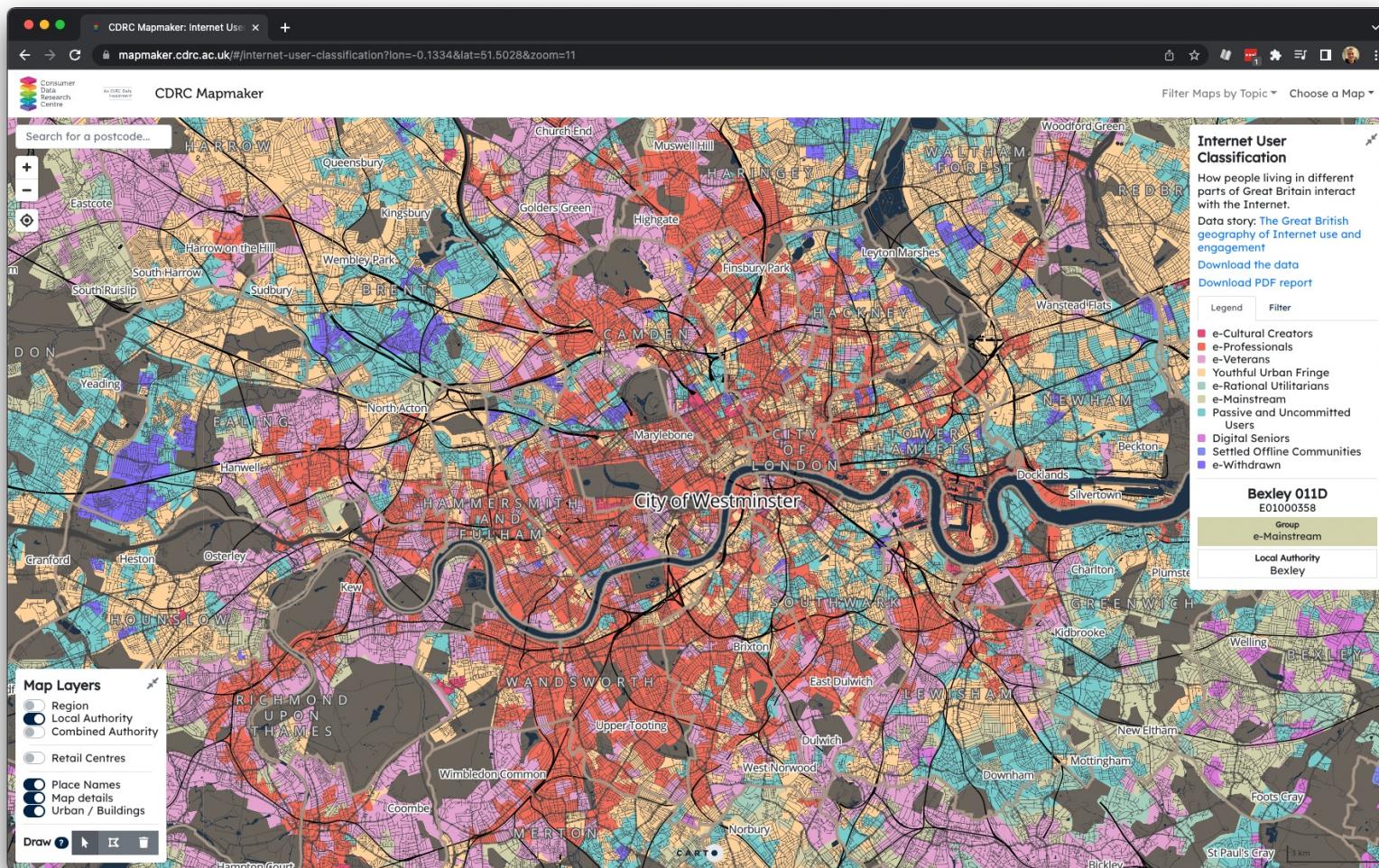
msoa_luc_input.csv

Possible Data Loss: Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

msoa11cd

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	msoa11cd	age_total	age0to4pc	age16to24pc	age25to44pc	age45to64pc	age75pluspc	nssec_total	1_higher_mi	2_lower_mi	ma_3_intermedi_4_employers_5_lowerr_supt_7_routine	8_unemploy	avg_dwn_sp	avb_superfa	no_decent_b	bband_speed	bband_sp	bband_sp	bband_sp	bband_sp	bband_sp	bband_sp	
2	E02000001	7375	0.032	0.38779966	0.09613559	0.40691525	0.27254237	0.06074576	5816	0.38462861	0.33906465	0.07806052	0.06464924	0.02596286	0.0401651	0.02665062	0.03696699	24.4	0.544	0.002	0.013	0.249	0.2
3	E02000002	6775	0.09269373	0.12162362	0.11321033	0.2801478	0.18568266	0.09800738	3926	0.05680082	0.17116658	0.14314824	0.1162048	0.06444218	0.20529801	0.13015792	0.11436577	77.7	0.988	0.01	0.02	0.159	0.2
4	E02000003	10045	0.08292683	0.10154306	0.11796914	0.3060229	0.248888	0.06460926	6483	0.08175228	0.20761993	0.18247725	0.11059695	0.07064631	0.15502082	0.09656023	0.09532624	74.3	0.99	0.002	0.008	0.164	0.2
5	E02000004	6182	0.05904238	0.10239405	0.13879004	0.2544848	0.2495956	0.08864445	4041	0.067805	0.20638456	0.19079436	0.10541945	0.08364266	0.15070527	0.12150458	0.07374412	81.1	0.992	0.008	0.026	0.162	0.2
6	E02000005	8562	0.09296893	0.11936463	0.11948143	0.21361831	0.15648286	0.05010512	5386	0.04936662	0.16635619	0.15648286	0.10413563	0.0873696	0.0992921	0.10702967	0.0992921	70.7	0.97	0.03	0.021	0.223	0.2
7	E02000007	8791	0.10340121	0.12535548	0.12854055	0.28460926	0.19747469	0.06882038	5158	0.05641722	0.16265995	0.13609926	0.09402869	0.07968205	0.18321055	0.15645599	0.13144663	75.3	0.988	0.012	0.006	0.177	0.2
8	E02000008	11569	0.10216959	0.1159132	0.11893854	0.2995073	0.2077967	0.06232172	7152	0.04194631	0.154641723	0.10025168	0.07969799	0.18945749	0.15450224	0.12108501	72.9	0.995	0.005	0.004	0.153	0.2	
9	E02000009	8395	0.09731988	0.10720667	0.13126861	0.32126262	0.20762359	0.04597975	5279	0.05190377	0.16859254	0.13638947	0.11252131	0.07861337	0.1397867	0.13316196	0.12502368	68.9	1	0	0.008	0.189	0.2
10	E02000010	8615	0.10562972	0.12222867	0.12153221	0.3045098	0.20290192	0.05200232	5243	0.04051024	0.16784284	0.14724394	0.10432958	0.08048847	0.15839002	0.14896052	0.12073241	73.5	0.999	0	0.002	0.039	0.2
11	E02000011	6187	0.08323903	0.10085664	0.12978827	0.26830451	0.24195895	0.06911731	3904	0.05609631	0.18545082	0.17622951	0.12295082	0.08330541	0.16137295	0.12909836	0.08529713	75.5	0.995	0.001	0.007	0.113	0.2
12	E02000012	9888	0.0785801	0.11043689	0.14178803	0.29277913	0.21399676	0.06260113	6056	0.09048877	0.20756275	0.16363937	0.12863737	0.05977543	0.13996592	0.08768164	0.12285337	76.8	0.999	0	0.004	0.132	0.2
13	E02000013	8402	0.10580814	0.12687455	0.14096095	0.20787505	0.20233278	0.05236848	5044	0.04361618	0.15992007	0.15087232	0.10229976	0.083636376	0.18933386	0.15622522	0.11399683	73.6	1	0	0.008	0.201	0.2
14	E02000014	9402	0.0981700	0.12093172	0.11922995	0.29259732	0.21697511	0.05626463	5826	0.04170958	0.14933305	0.13594233	0.09972537	0.08324751	0.19910745	0.17061449	0.12032369	73.5	1	0	0.003	0.04	0.2
15	E02000015	7563	0.10273970	0.1156261	0.11222266	0.2994109	0.19674732	0.05209573	4637	0.04393700	0.15646284	0.13143845	0.09747482	0.0841061	0.18977787	0.15203796	0.143843	64.3	0.999	0.001	0.002	0.115	0.2
16	E02000016	7676	0.12441373	0.10682647	0.14369463	0.44460379	0.11737884	0.01055237	4563	0.09883848	0.1590474	0.103725	0.10190664	0.06421214	0.1531886	0.10870042	0.17773395	53.4	0.939	0	0.003	0.044	0.2
17	E02000017	8498	0.08602024	0.09837609	0.14179807	0.36608614	0.05049577	5309	0.08537505	0.18698517	0.1232566	0.1191044	0.06162835	0.15567282	0.10591783	0.1624576	74.5	1	0	0.002	0.045	0.2	
18	E02000018	8890	0.11158605	0.12463442	0.12609674	0.31169854	0.1904387	0.04623171	5356	0.04686333	0.15832711	0.13629574	0.10306199	0.09055265	0.10962733	0.14395071	0.13032114	71.5	0.999	0	0.014	0.192	0.2
19	E02000019	8276	0.10681489	0.11829386	0.1161189	0.3221363	0.19864669	0.04809087	5029	0.05050706	0.17458739	0.14714655	0.10161066	0.07595944	0.18373434	0.15887851	0.10757606	73.3	0.964	0.013	0.011	0.161	0.2
20	E02000020	9276	0.10155239	0.1185856	0.11664511	0.32823743	0.18510314	0.05868156	5473	0.05838428	0.17397399	0.13630513	0.10508094	0.07457487	0.16718436	0.13447835	0.14598964	77.7	0.995	0.005	0.008	0.126	0.2
21	E02000021	7904	0.10235324	0.11652328	0.11955972	0.29617915	0.21862348	0.05452935	4783	0.04787799	0.17854938	0.1436337	0.11394522	0.08028434	0.18293958	0.13798871	0.11478152	72.5	0.994	0.004	0.005	0.145	0.2
22	E02000022	8777	0.14230375	0.1444685	0.11222513	0.31051409	0.1914094	0.0488109	4888	0.05277153	0.14915312	0.11392923	0.09899775	0.06443035	0.16040173	0.14747392	0.21640417	61.5	0.993	0.002	0.002	0.044	0.2
23	E02000023	6482	0.1349899	0.13159519	0.12419007	0.34464671	0.15869602	0.0336316	3666	0.04825518	0.15372627	0.1300436	0.09514722	0.07524537	0.10956707	0.13849509	0.16848919	68.8	0.997	0.003	0.012	0.16	0.2
24	E02000024	10156	0.06213070	0.09462387	0.0950172	0.28406853	0.08881449	0.0595075	5096	0.08537055	0.18698517	0.1232656	0.1191044	0.06162835	0.15567282	0.10591783	0.1624576	46.7	0.96	0.024	0.013	0.233	0.2
25	E02000025	7956	0.06460533	0.08157366	0.10212327	0.31271998	0.26181498	0.05375322	5457	0.16657504	0.29265164	0.15063222	0.11379879	0.0892798	0.10298699	0.0630383	0.16138904	50.7	0.982	0.001	0.003	0.061	0.2
26	E02000026	6313	0.06621258	0.0963092	0.09662601	0.27546333	0.2689688	0.07381594	4174	0.17537135	0.27407762	0.17073402	0.14542405	0.04120748	0.0807379	0.05630091	0.05654049	42.8	0.995	0.002	0.005	0.153	0.2
27	E02000027	9598	0.07075694	0.09203103	0.10507757	0.30595973	0.24424072	0.07110954	5636	0.1566714	0.28424415	0.14389638	0.1407263	0.04843861	0.16406417	0.05624556	0.0651171	41	0.997	0.001	0.011	0.246	0.2
28	E02000028	6166	0.06779111	0.08968589	0.1037905	0.30181641	0.25089199	0.07525138	4153	0.18468577	0.29568896	0.1584396	0.12352516	0.0452769	0.08427463	0.0520106	0.05634481	48.3	0.942	0	0.004	0.075	0.2
29	E02000029	8517	0.12831396	0.10261829	0.10226606	0.3117725	0.24914876	0.05577081	5681	0.14205246	0.27583172	0.16123922	0.12480197	0.051394	0.11218215	0.06688963	0.06565746	37.3	0.97	0	0.007	0.162	0.2
30	E02000030	9130	0.05125958	0.10799562	0.10624315	0.21971523	0.2886562	0.05419496	5865	0.18925831	0.28354646	0.13554987	0.15260017	0.03614663	0.09224211	0.05234442	0.05831202	45.2	0.912	0.004	0.026	0.185	0.2
31	E02000031	7294	0.06471072	0.08815465	0.09542089	0.2890046	0.24143131	0.107622	4691	0.20677894	0.2929013	0.14218717	0.12928307	0.0370923	0.08271158	0.04732466	0.06118099	56.1	0.953	0.017	0.005	0.14	0.2
32	E02000032	7857	0.05905562	0.10217096	0.10602011	0.25378643	0.27771414	0.07012855	5144	0.1714619	0.27527216	0.15746501	0.15824261	0.04179627	0.08786936	0.05520995	0.05268274	57	0.982	0	0.009	0.17	0.2
33	E02000033	8537	0.06852524	0.10097224	0.1049549	0.31116132	0.24469954	0.06758815	5620	0.11921708	0.24786477	0.16423488	0.13950178	0.05498221	0.11014235	0.07295374	0.0911032	49.9	0.998	0	0.006	0.156	0.2
34	E02000034	8508	0.07075694	0.09203103	0.10507757	0.30595973	0.24424072	0.07110954	5636	0.1566714	0.28424415	0.14389638	0.1407263	0.04843861	0.16406417	0.05624556	0.0651171	57.8	0.999	0	0.004	0.142	0
35	E02000035	7965	0.06804771	0.07407407	0.08612681	0.3578154	0.21933459	0.08851224	5344	0.20303144	0.2877994	0.13117515	0.11957335	0.03967066	0.09281437	0.04977545	0.07616018	62.4	0.913	0	0.003	0.059	0.2
3																							

Workflow



Workflow

Typical workflow:

1. Choose input domains
2. Select associated quantifiable variables
3. Standardise variables
4. Measure variables for association (multi-collinearity)
5. Choose clustering method
6. Choose number of clusters
7. Interpret findings, test and finalise them

Input domain

- What type of application will your classification be aimed/themed around? What information is relevant?
- What or who are you segmenting? Individuals / Households / Postcodes / Output Areas / Wards ?
- Scale of data available?
- Difficulties to combine different datasets? Temporal mismatch, spatial mismatch?

Variables

Need to select variables that are associated with application or phenomenon you are trying to classify within the population you are looking to segment:

- Subjective and **arbitrary** – what you choose will affect the results.
- No single or unique set of variables.
- Balancing act of theory, available data and statistical considerations.
- Selection of variables directly impact what is classified.

Standardise variables

Standardise between areas to account for differences between areas:

- expressing variables as percentages or proportions (e.g. using a relevant population denominator)

Standardise between variables to account for value ranges:

- transform all variables to be on the same scale
- z-scores, range standardisation

Multi-collinearity

Use of correlation matrix – correlated or not?

- Coefficient (or r-square value) is presented on a scale between -1 and 1.
- The greater the value, the greater the association between the two variables.
- Set an appropriate **threshold** - as a rule of thumb, two variables with coefficients greater than ± 0.8 can be considered to be highly correlated.
- Remove the variable which correlates the greatest with the other variables in the matrix to leave most unique one.
- May replace or keep variable with "convincing argument".

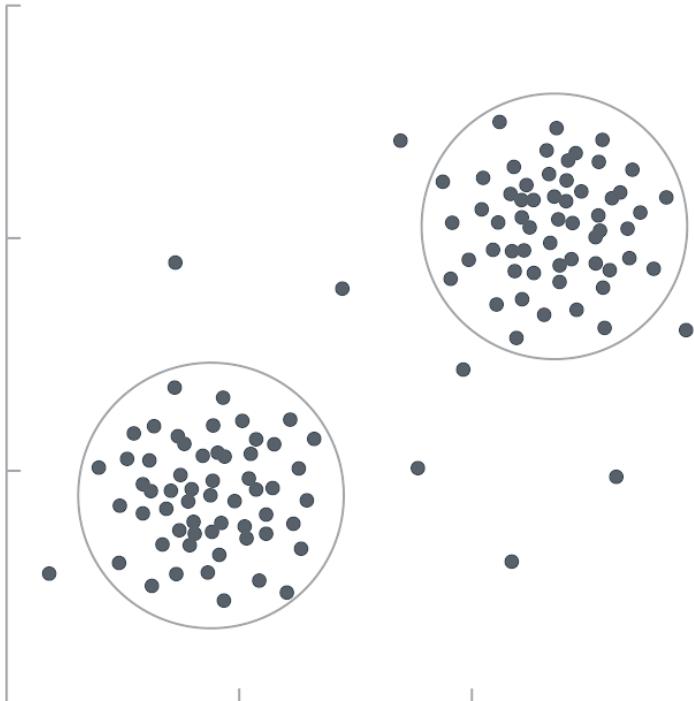
Clustering

Use of clustering algorithms to partition demographic data into groups sharing similar characteristics:

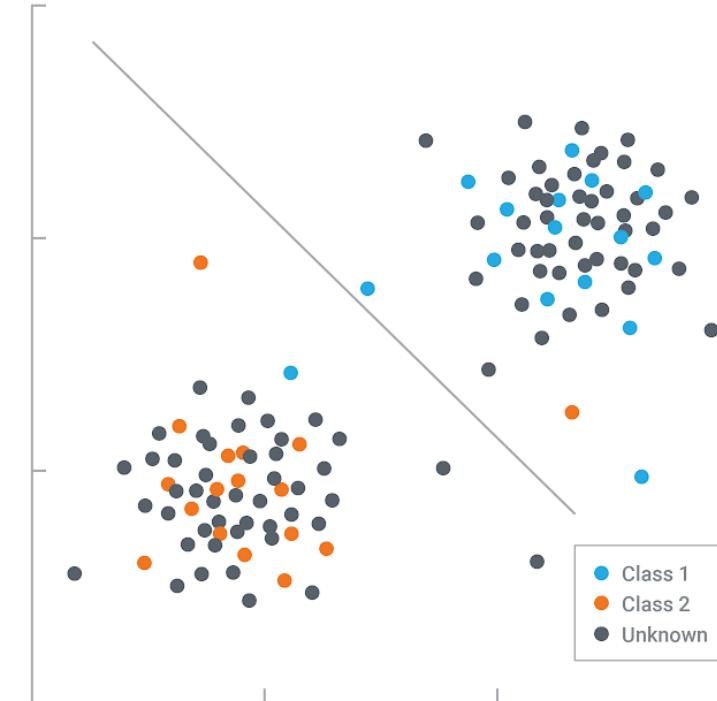
- Different methods often create different groups.
- Requires user intuition to decide best outcome and optimal cluster numbers.
- Some methods (such as k-means) create different groups with each run.
- Creating a geodemographic classification therefore sometimes requires running an algorithm multiple times.
- Can be computationally expensive.
- Unsupervised methods used.

Unsupervised versus supervised

UNSUPERVISED



SUPERVISED



k-means

- Assign geographic areas with common underlying attributes to similar classification groups.
- k clusters (pre-defined) of n individual observations.
- Each observation can have any number of attribute data.
- Used in: Internet User Classification, ONS 2011 Output Area Classification.

k-means



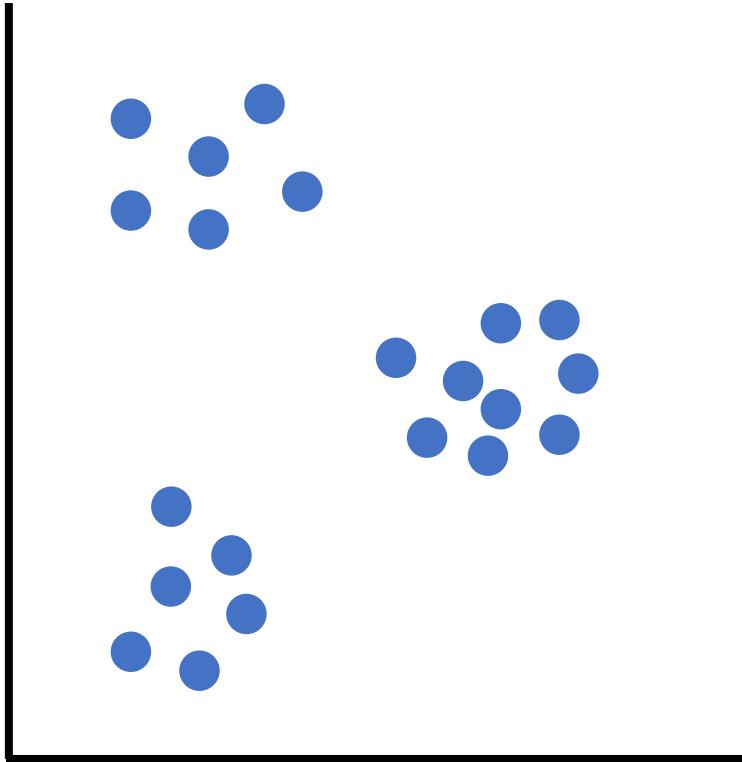
k-means

- Minimising the distance between an observation's input variables to the means of the respective cluster groups.
- Maximising the distance between cluster groups.
- Number of clusters defined a priori.

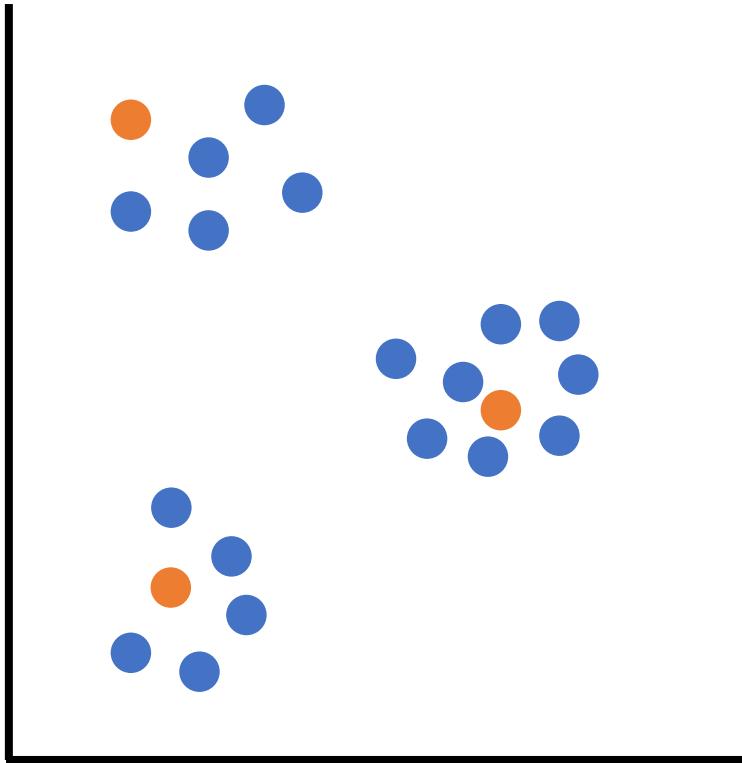
Process

- Step 1: identify your k
- Step 2: randomly identify k distinct data points as initial cluster centre
- Step 3: assign each observations to the nearest cluster
- Step 4: calculate the mean of each cluster
- Step 5: repeat with mean value becoming new cluster centre until no change

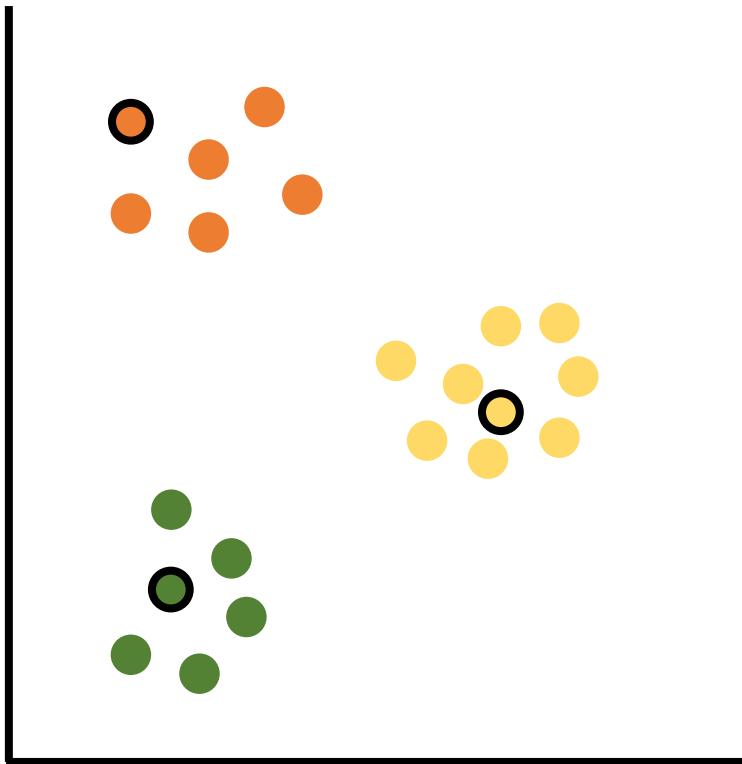
Step 1



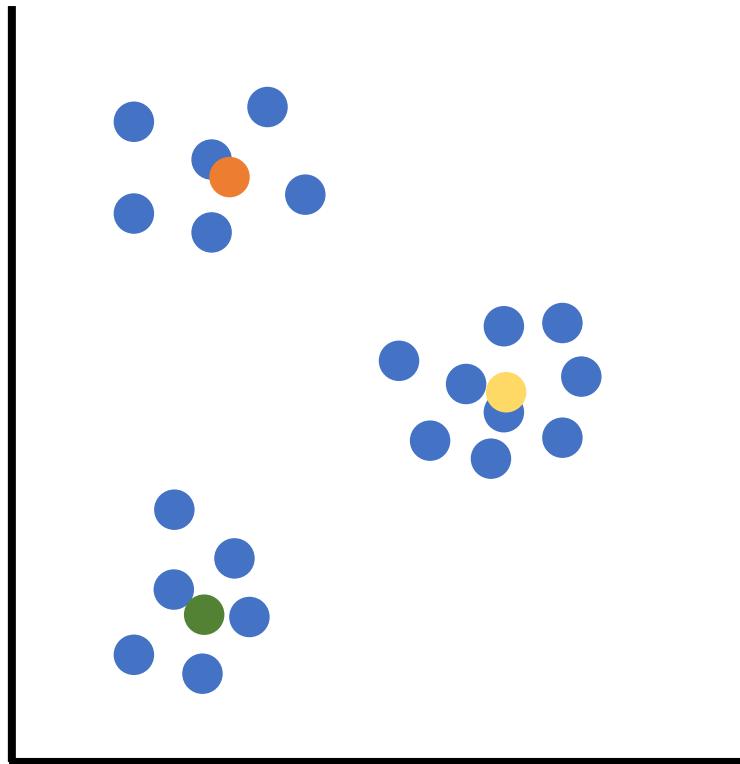
Step 2



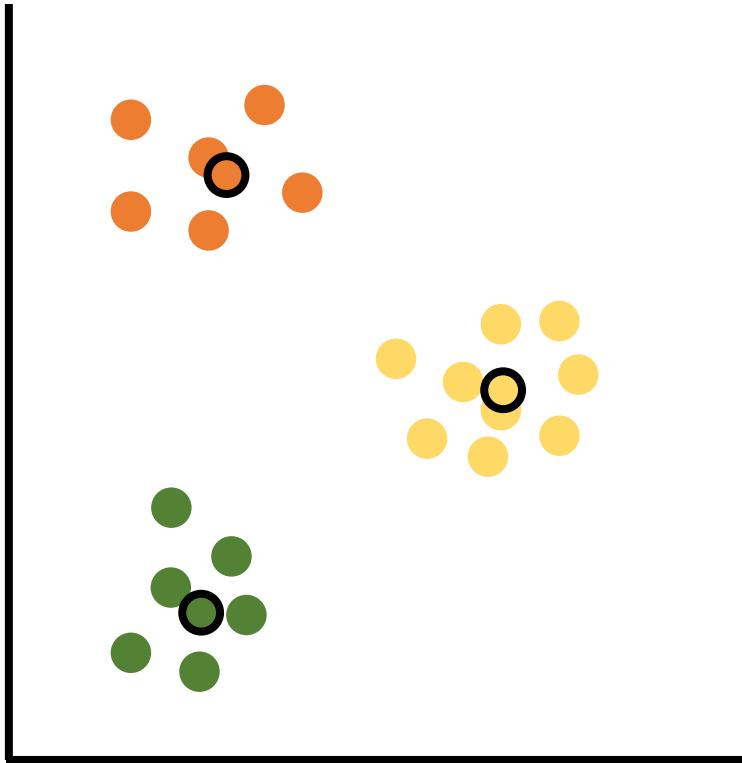
Step 3



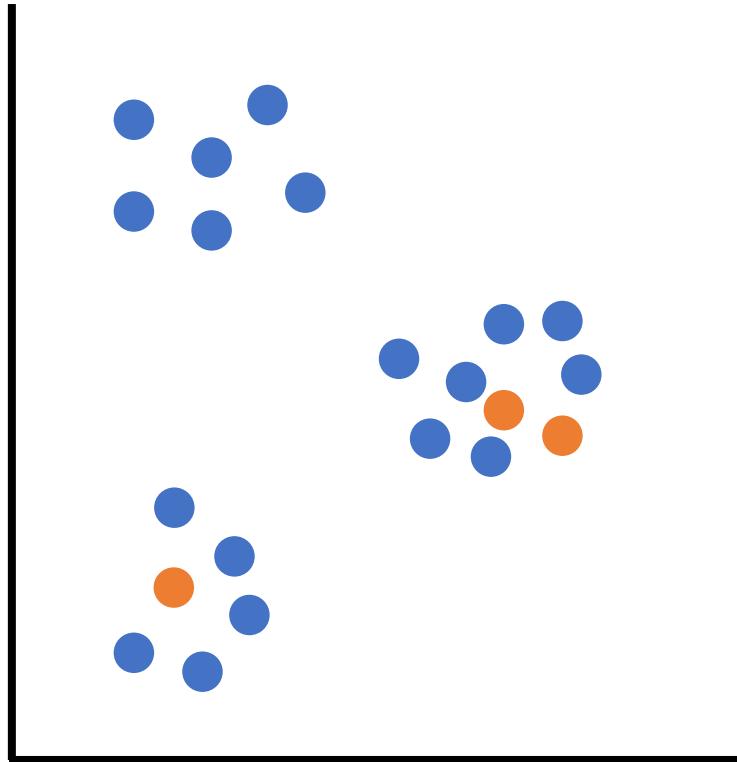
Step 4



Step 5



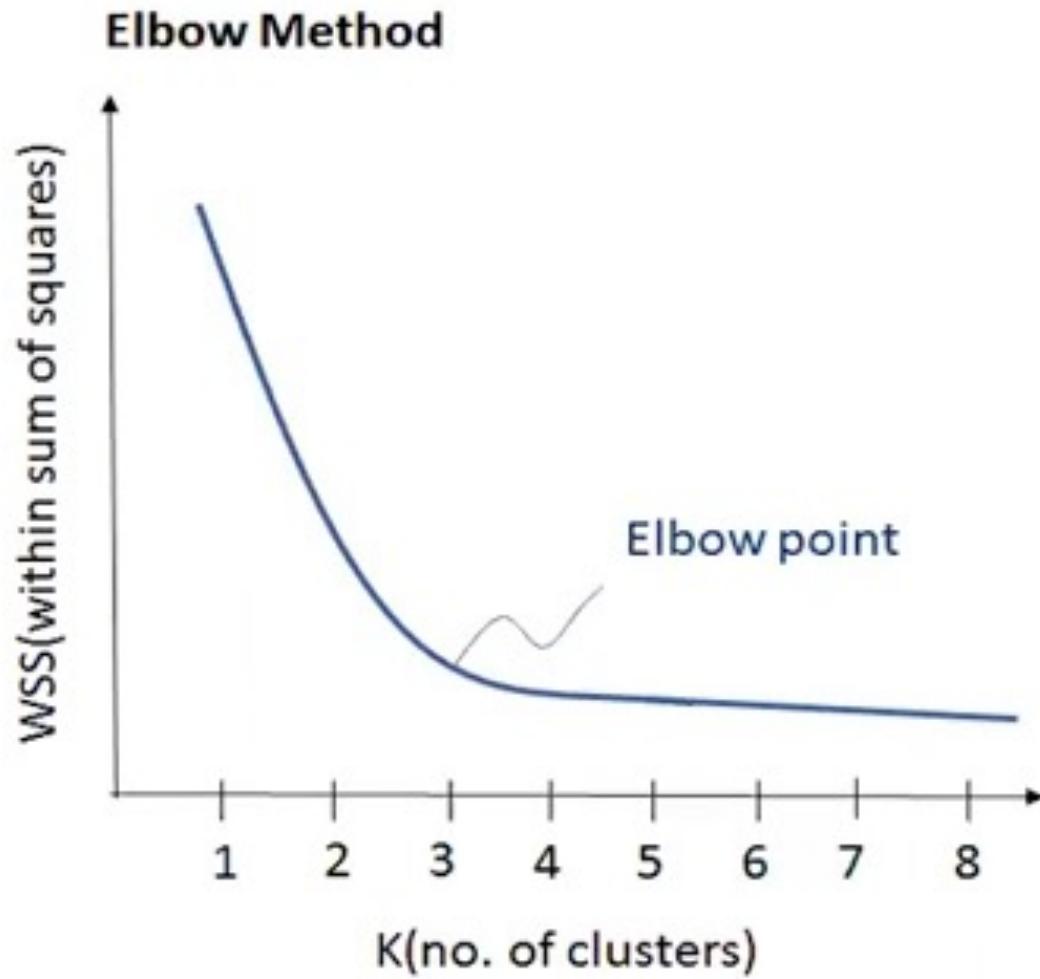
Multiple iterations



Number of clusters

- Too few: too much variation within the groups.
- Too many: overfitting and splitting similar observations.
- Iterate through the model multiple times and try to minimise variation within cluster groups.

Number of clusters



Interpretation of clusters

- Look at the means of the input data for each cluster ('signature').
- Can use radar or bar plots to characterise the cluster.
- Not spatial in nature (different to the DBSCAN).

Cluster labels

- Should generalise the characteristics of each group.
- Should be clear for users unfamiliar with geodemographics.
- Should not cause offense to the inhabitants of each group.
- Should not draw on assumptions that cannot be applied by the data.
- Be aware of the unevenness of the ecological fallacy.
- Create pen portraits for easy interpretation.

Conclusion

- Geodemographics as the analysis of people by where they live.
- Typically, a form of unsupervised machine learning is used; k-means.
- Once created: **benchmarking** against additional quantitative data.
- Sometimes the spatial scale should be re-considered.



Automation

99% of what you have done / will do is using existing functions

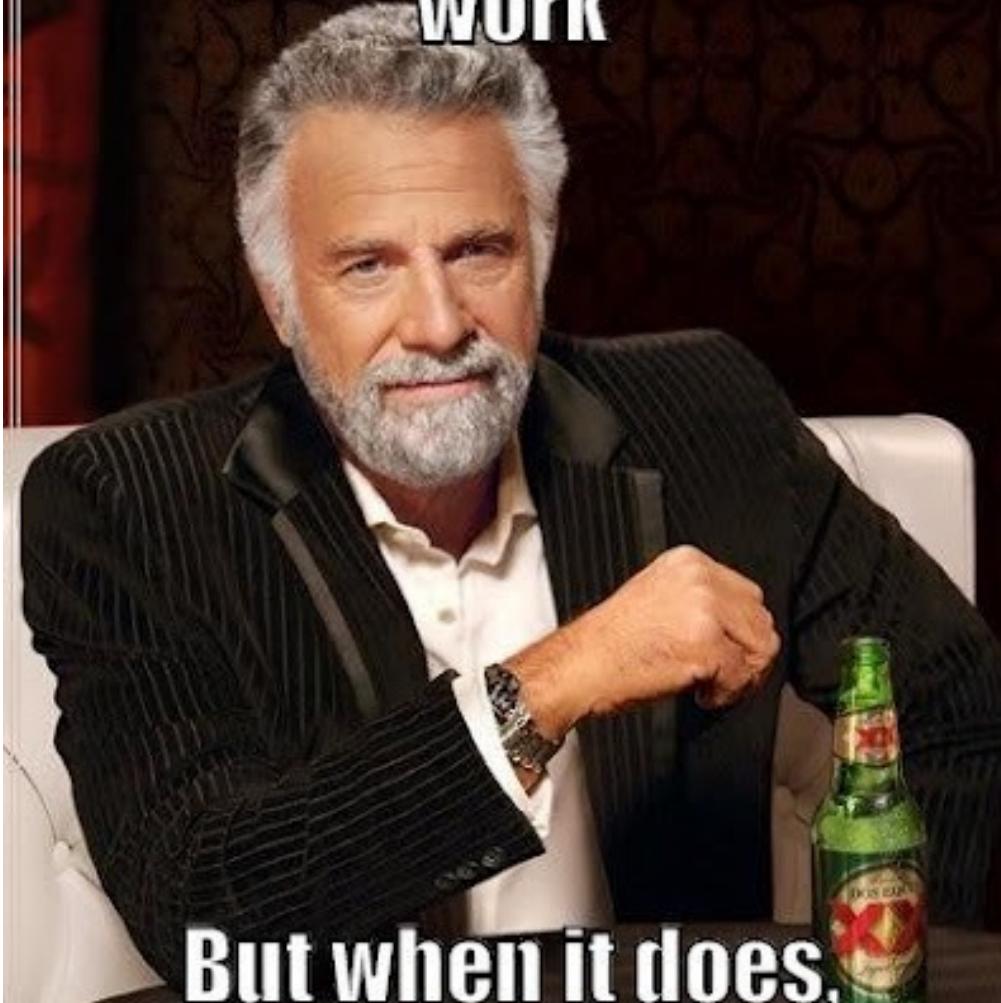
... but you can combine these into functions and use your own arguments and parameters.

... but if you use some workflows repeatedly and across projects you could think of creating a library or package.

No matter what:

R is optimised in certain ways: design of your code is very important.

**My code doesn't always
work**



**But when it does,
it works on my machine**

**SAY "IT WORKS IN MY
MACHINE"**

ONE MORE TIME

MemesHappen

```
dayloop2 <- function(temp){  
  for (i in 1:nrow(temp)){  
    temp[i,10] <- i  
    if (i > 1) {  
      if ((temp[i,6] == temp[i-1,6]) & (temp[i,3] == temp[i-1,3])) {  
        temp[i,10] <- temp[i,9] + temp[i-1,10]  
      } else {  
        temp[i,10] <- temp[i,9]  
      }  
    } else {  
      temp[i,10] <- temp[i,9]  
    }  
  }  
  names(temp)[names(temp) == "V10"] <- "Kumm."  
  return(temp)  
}
```

<https://stackoverflow.com/questions/2908822/speed-up-the-loop-operation-in-r>

~ 850k rows

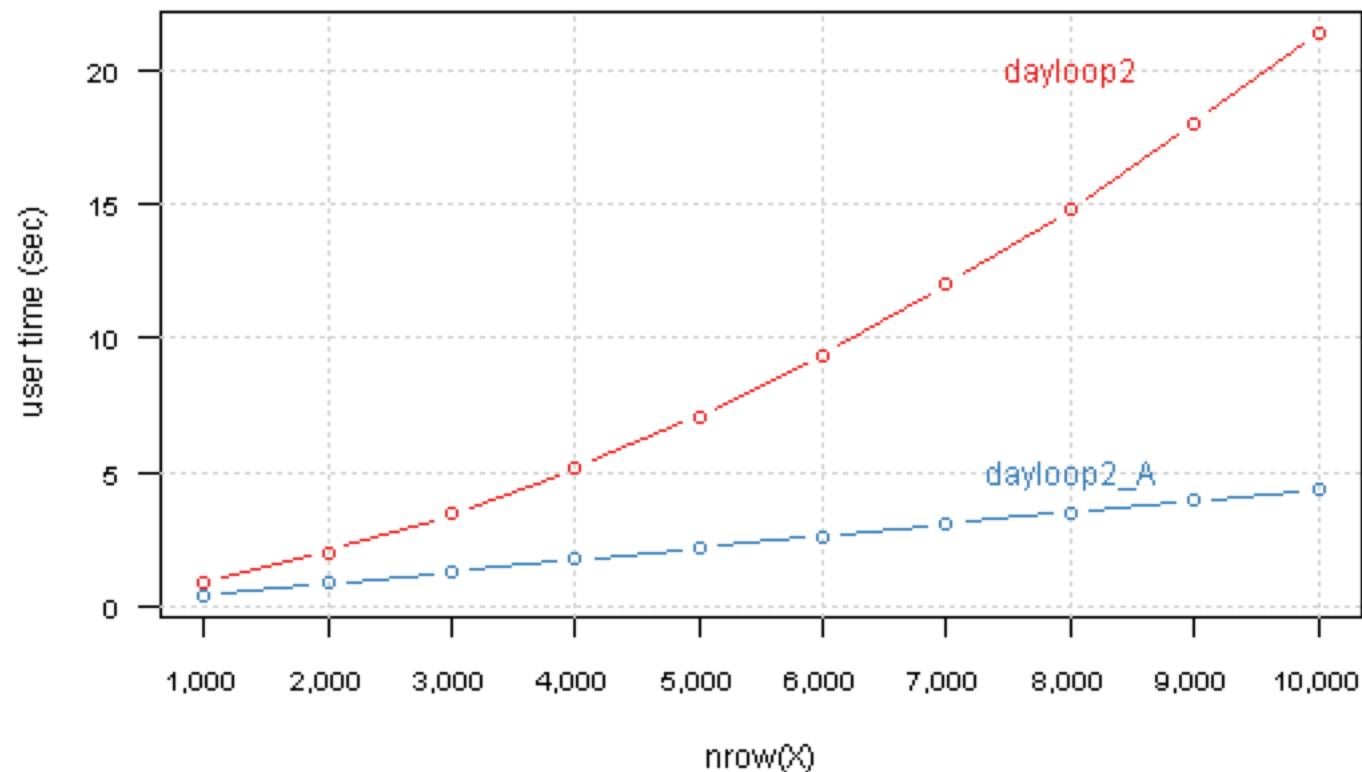
> 10 hours

```
dayloop2 <- function(temp){  
  for (i in 1:nrow(temp)){  
    temp[i,10] <- i  
    if (i > 1) {  
      if ((temp[i,9] <= temp[i-1,9]) & (temp[i,3] == temp[i-1,3])) {  
        temp[i,10] <- temp[i,9] + temp[i-1,10]  
      } else {  
        temp[i,10] <- temp[i,9]  
      }  
    } else {  
      temp[i,10] <- temp[i,9]  
    }  
  }  
  names(temp)[names(temp) == "V10"] <- "Kumm."  
  return(temp)  
}
```

<https://stackoverflow.com/questions/2908822/speed-up-the-loop-operation-in-r>

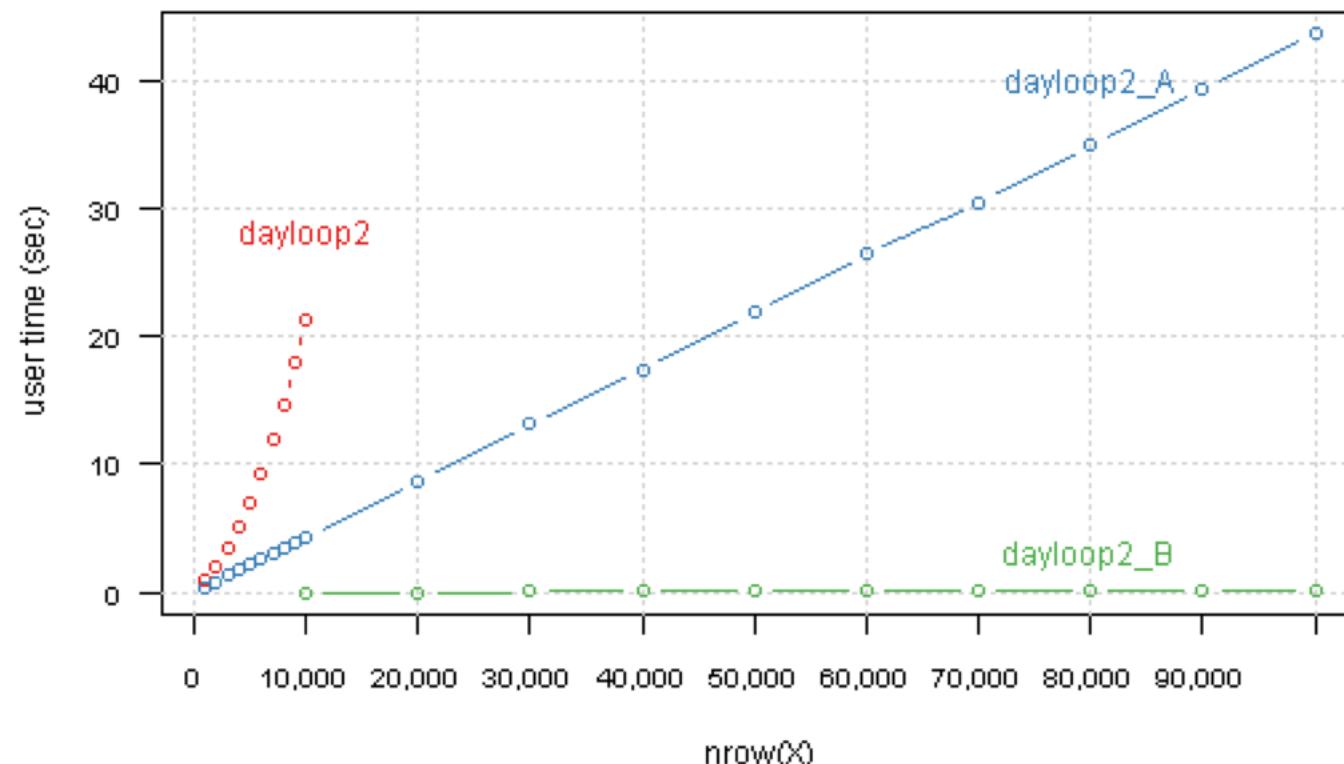
~ 850k rows

dayloop2 vs dayloop2_A



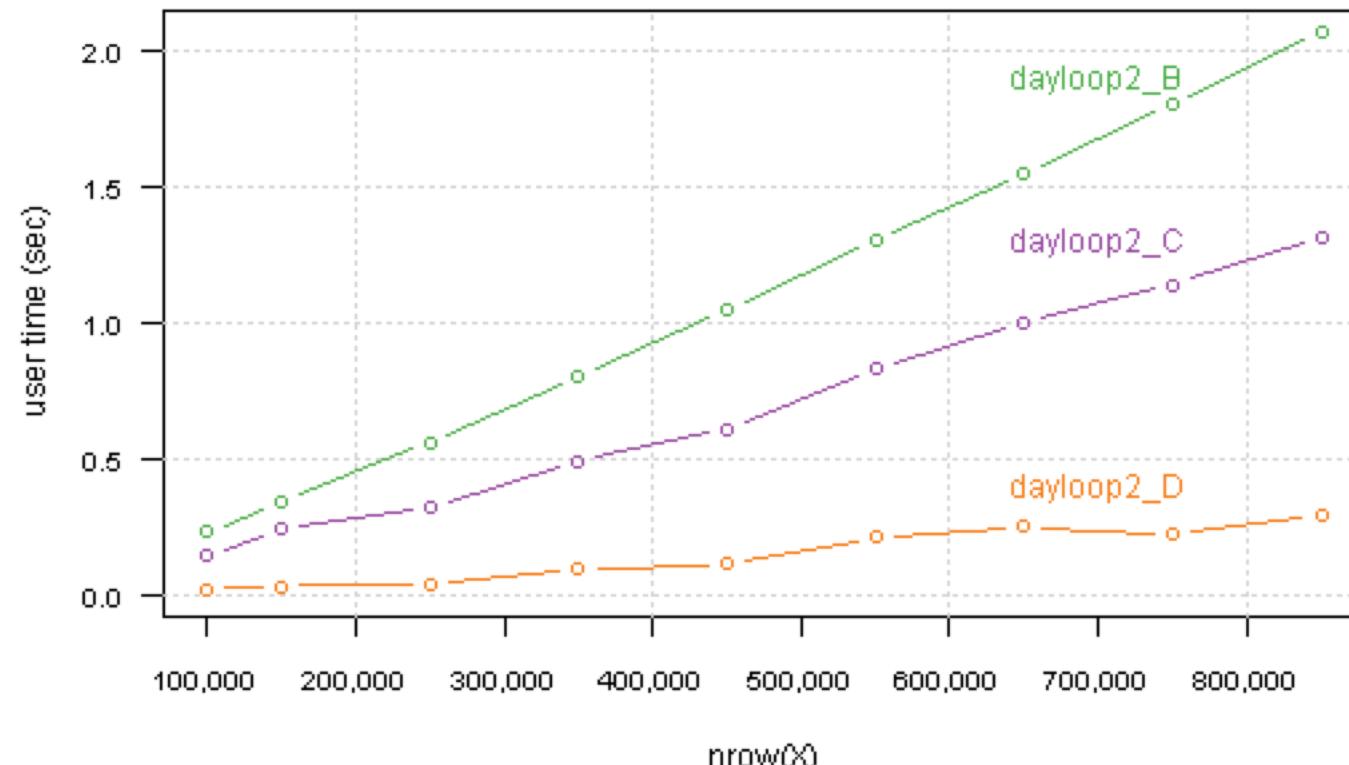
reduce indexing

Vectorization FTW



vectorisation

More vectorization



more vectorisation

A screenshot of a GitHub Gist page titled "loop.R". The page has a dark theme. At the top, there are navigation icons for back, forward, search, and refresh, followed by the URL "gist.github.com/hadley/48c396ae674cdccb618033df1b1ed671". Below the URL, there are tabs for "Code" (which is selected) and "Revisions" (with a count of 3). On the right side of the header, there are buttons for "Embed", "Raw", and "Download ZIP". The main content area contains the R code for "loop.R".

```
loop.R
1 f1 <- function(n) {
2   x <- numeric()
3   for (i in 1:n) {
4     x <- c(x, i)
5   }
6   x
7 }
8
9 f1.5 <- function(n) {
10  x <- numeric()
11  for (i in 1:n) {
12    x[[i]] <- i
13  }
14
15  x
16 }
17
18 f2 <- function(n) {
19   x <- numeric(n)
20   for (i in 1:n) {
21     x[[i]] <- i
22   }
23   x
24 }
25
26 bench::mark(f1(1e4), f1.5(1e4), f2(1e4))
27 #> Warning: Some expressions had a GC in every iteration; so filtering is disabled.
28 #> # A tibble: 3 × 6
29 #>   expression      min    median `itr/sec` mem_alloc `gc/sec`
30 #>   <bch:expr> <bch:tmin> <bch:tmed> <bch:byt>   <bch:byt>
31 #> 1  f1(10000)    109ms 113.34ms    8.77    382MB  77.2
32 #> 2  f1.5(10000)  936µs 1.05ms    726.    1.68MB  67.6
33 #> 3  f2(10000)    212µs 216.77µs  4495.    96.77KB  6.00
```

Speeding up

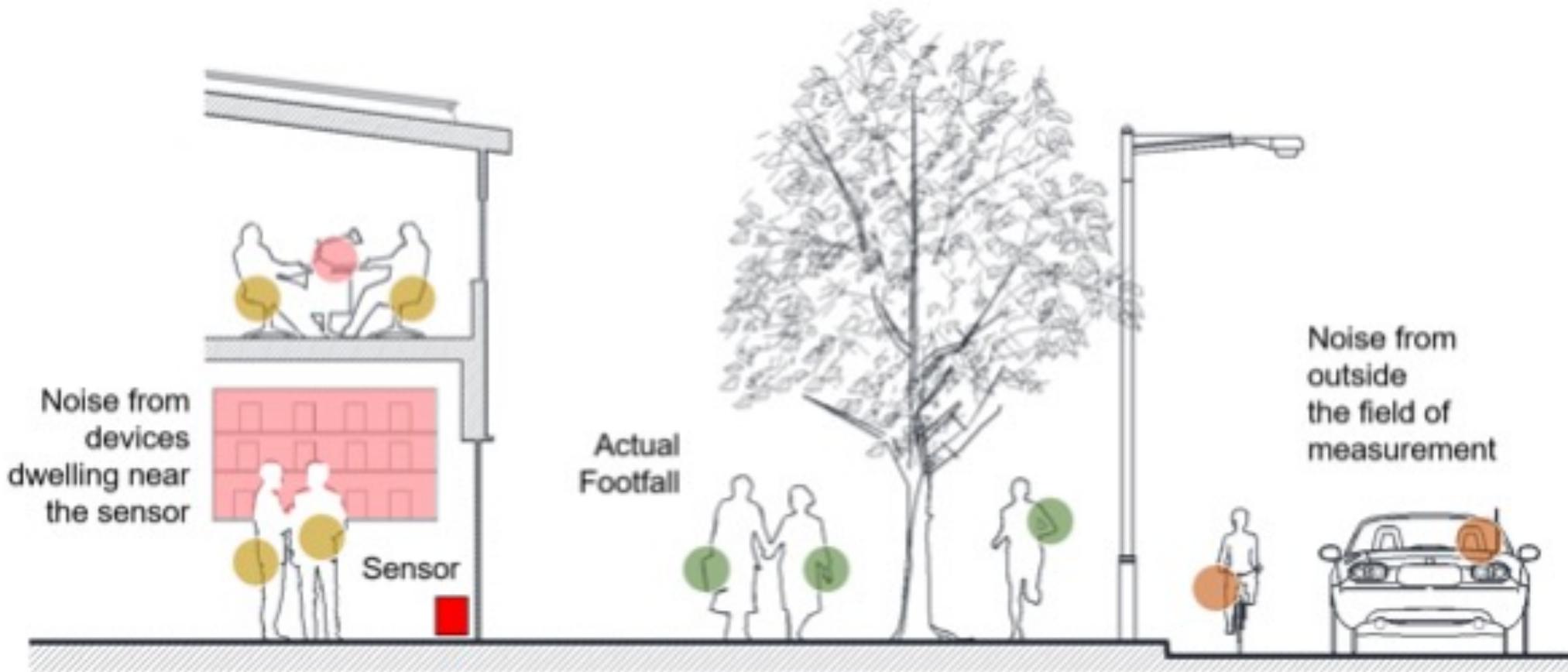
First things to check:

- Store the results and access them rather than repeatedly re-calculating
- Take non-loop-dependent calculations out of loops
- Avoid necessary calculations (e.g. regex or fixed search?)

Not always enough, more advanced strategies:

- Parallel processing
- GPU-accelerated analytics
- Different tools?

Smart street sensor project CDRC



Smart street sensor project CDRC

Start of the project:

- 40 locations equipped with sensors
- 1 million measurements / day
- 100 MB / day
- Download from servers: 20 minutes
- Processing: 30 minutes

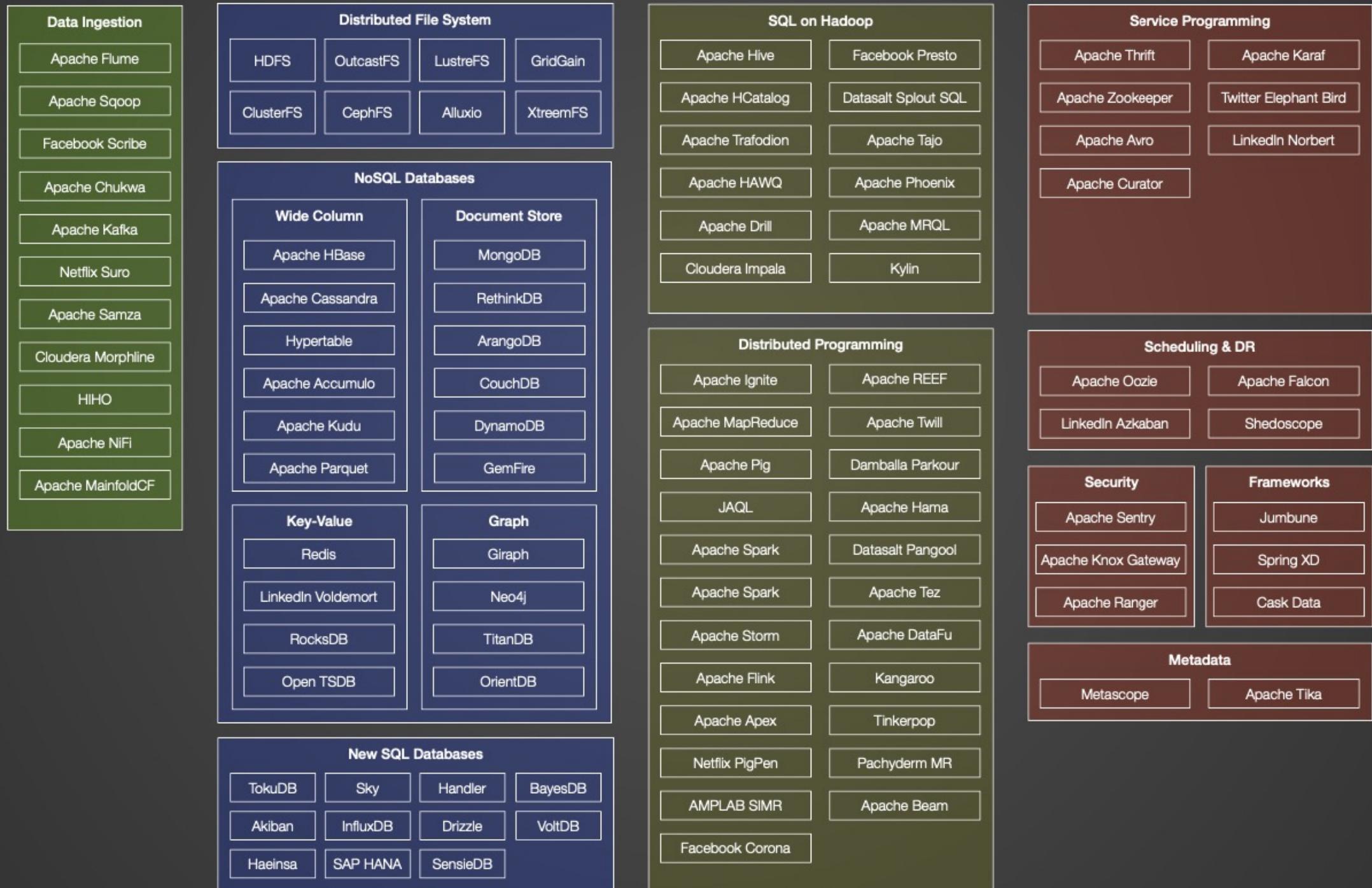
Smart street sensor project CDRC

Growth of the project:

- 675 locations equipped with sensors
- 31 million measurements / day
- 2 GB / day
- Download from servers: 2 hours
- Processing: 5 hours

Smart street sensor project CDRC

- Solution: set-up a Big Data infrastructure?
- Can be very confusing, costly



Medium data toolkit

Soundararaj 2019:

- Unix philosophy: pipes and text streams
- Command-line tools: cat, find, sort, uniq, sed, grep, awk
- Parallelisation: using xargs, GNU parallel to utilise all processing cores

Unix philosophy

- Write programs that do one thing and do it well.
- Write programs to work together.
- Write programs to handle text streams, because that is a universal interface.

Pipelines compared

Benchmarked:

- Full Pipeline in R: 20 minutes, 14 seconds
- Full pipeline in Unix tools: 18 seconds
- Full pipeline in Unix tools (parallelised): 3 seconds

2 hours download, 5 hours of processing got done in 15 minutes.

Pipelines compared

- Lots of data != Big Data
- In some cases: evaluate the data for 'bigness' in each dimension (volume, velocity, variety, etc.) and then decide if there are appropriate tools available that do one thing very well.

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk

