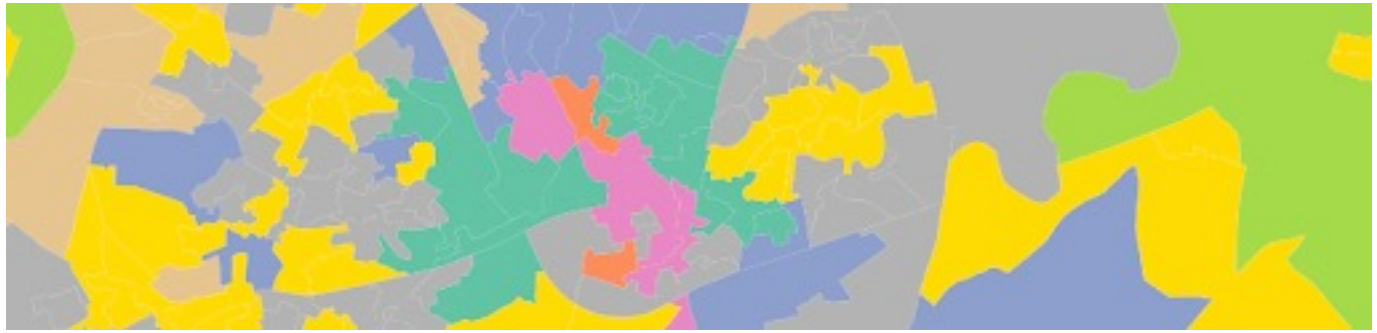
Principles of Spatial Analysis

WEEK 07: GEODEMOGRAPHICS



UCL

This week

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

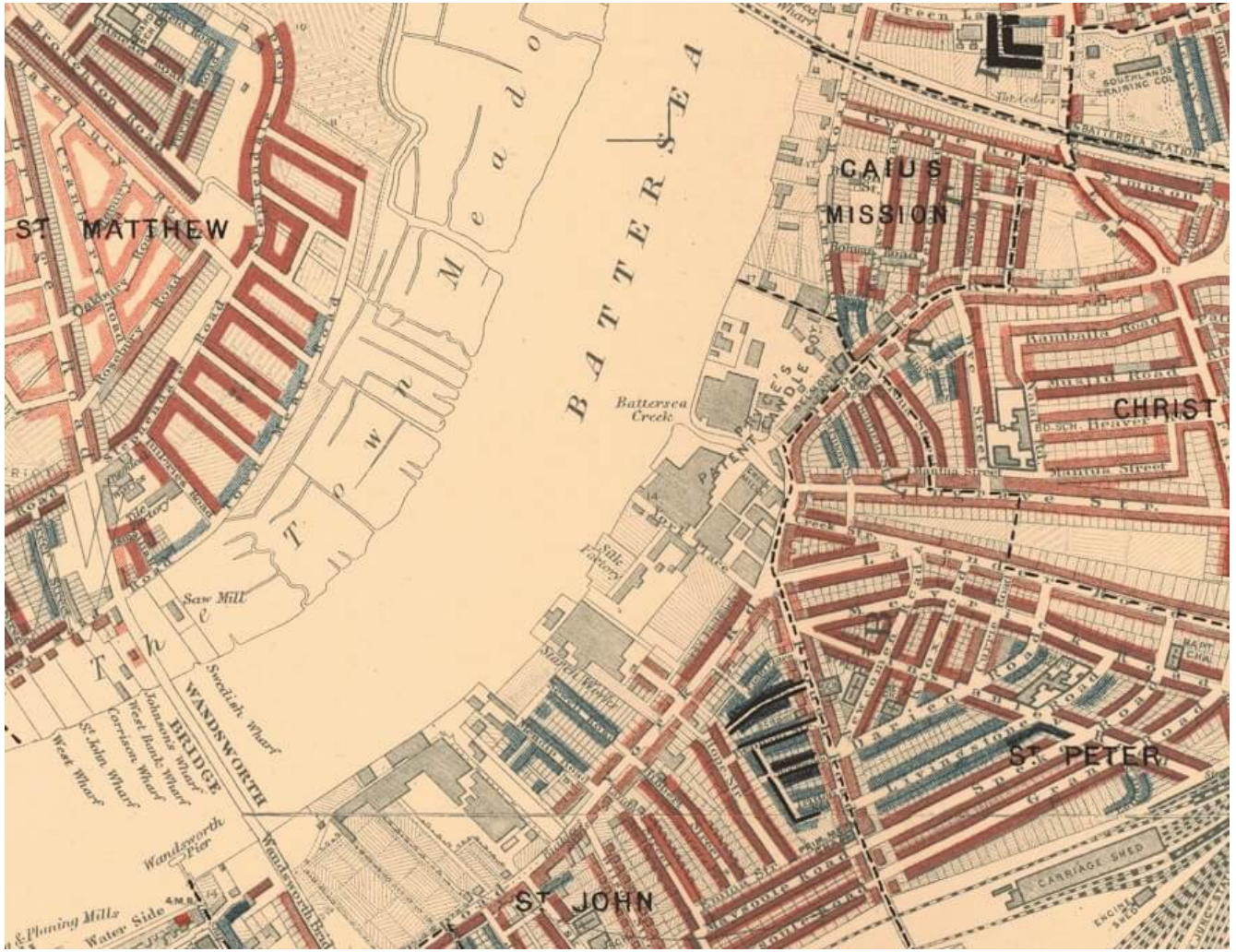
Geodemographics

- 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 / Experia MOSAIC).
- Office for National Statistics' Output Area Classification (2001, 2011, 2021).
- ONS' Output Area Classification completely open using Census data.
- All classification have a similar structure, typically hierarchical.

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:

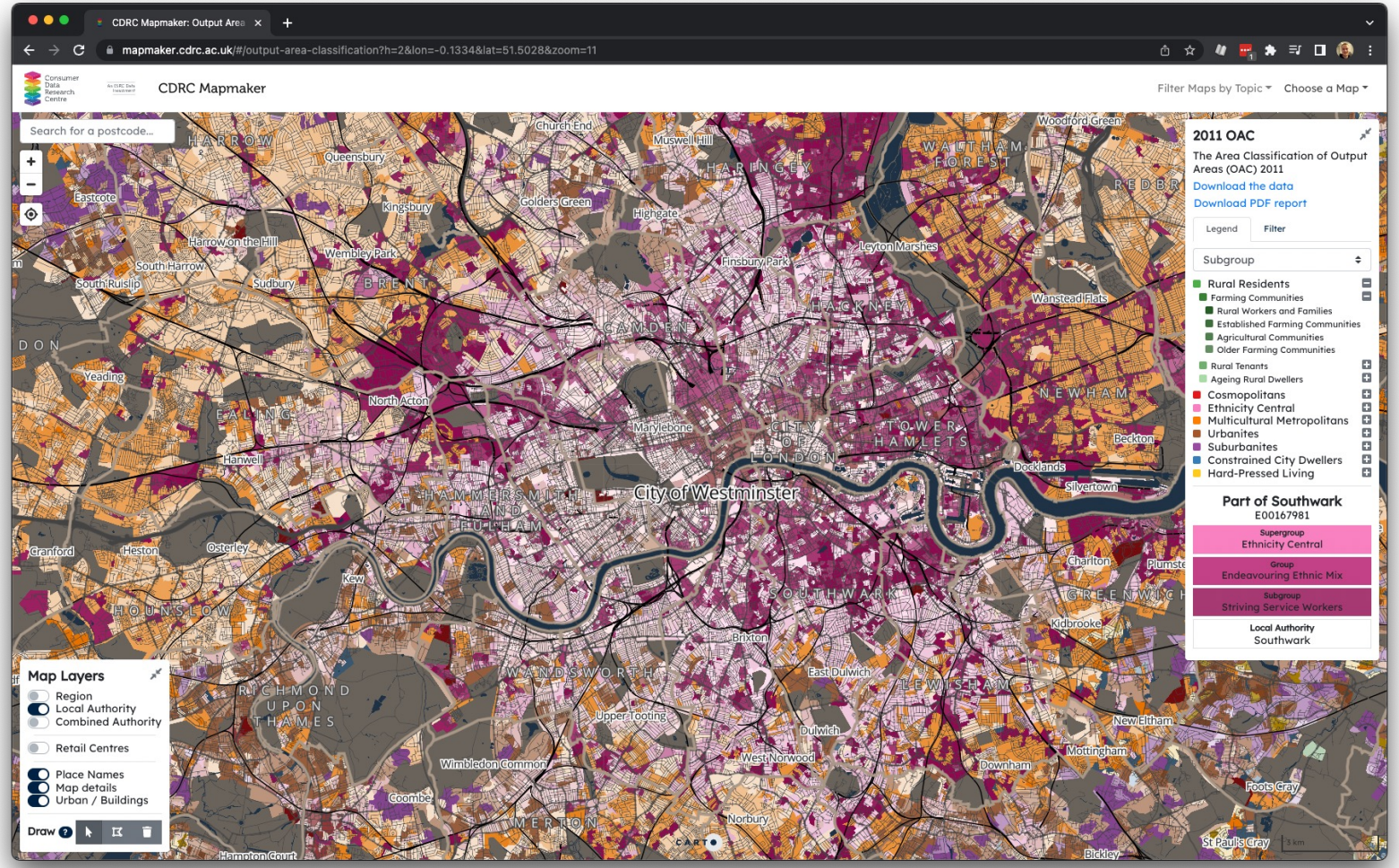
- Output Area Classification: 60 census variables, focused on holistic understanding of area characteristics.
- Workplace Zones Classification: 48 census variables across 4 domains, focused on workers and workplaces.
- Internet User Classification: built from a range of consumer, survey and open (census) 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:

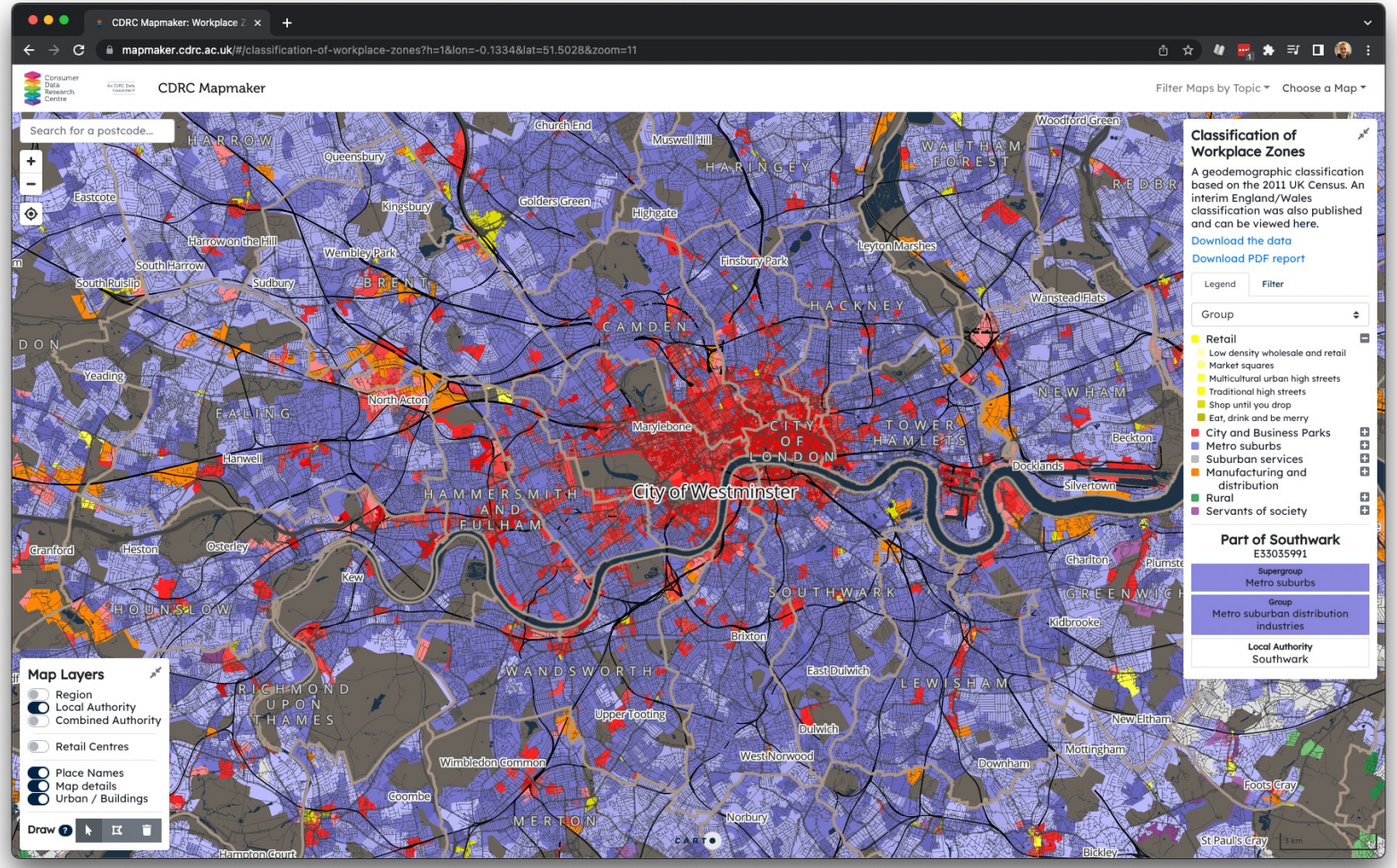
- Sometimes there is a hierarchy of clusters (e.g. supergroups, clustergroups).
- Each cluster will have a name and description (pen portrait).

Geodemographics



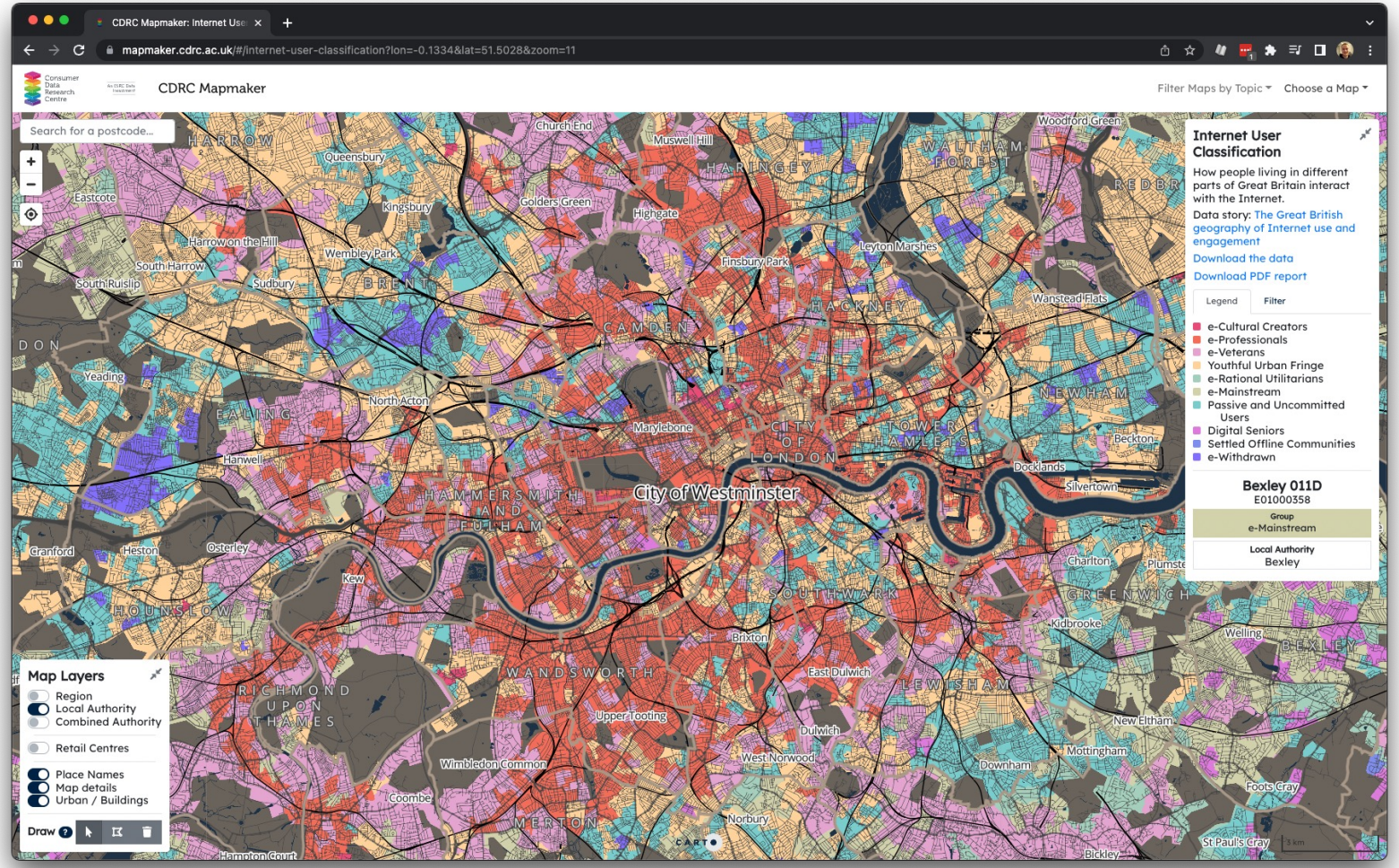
2011 Output Area Classification on mapmapker.cdrc.ac.uk

Geodemographics



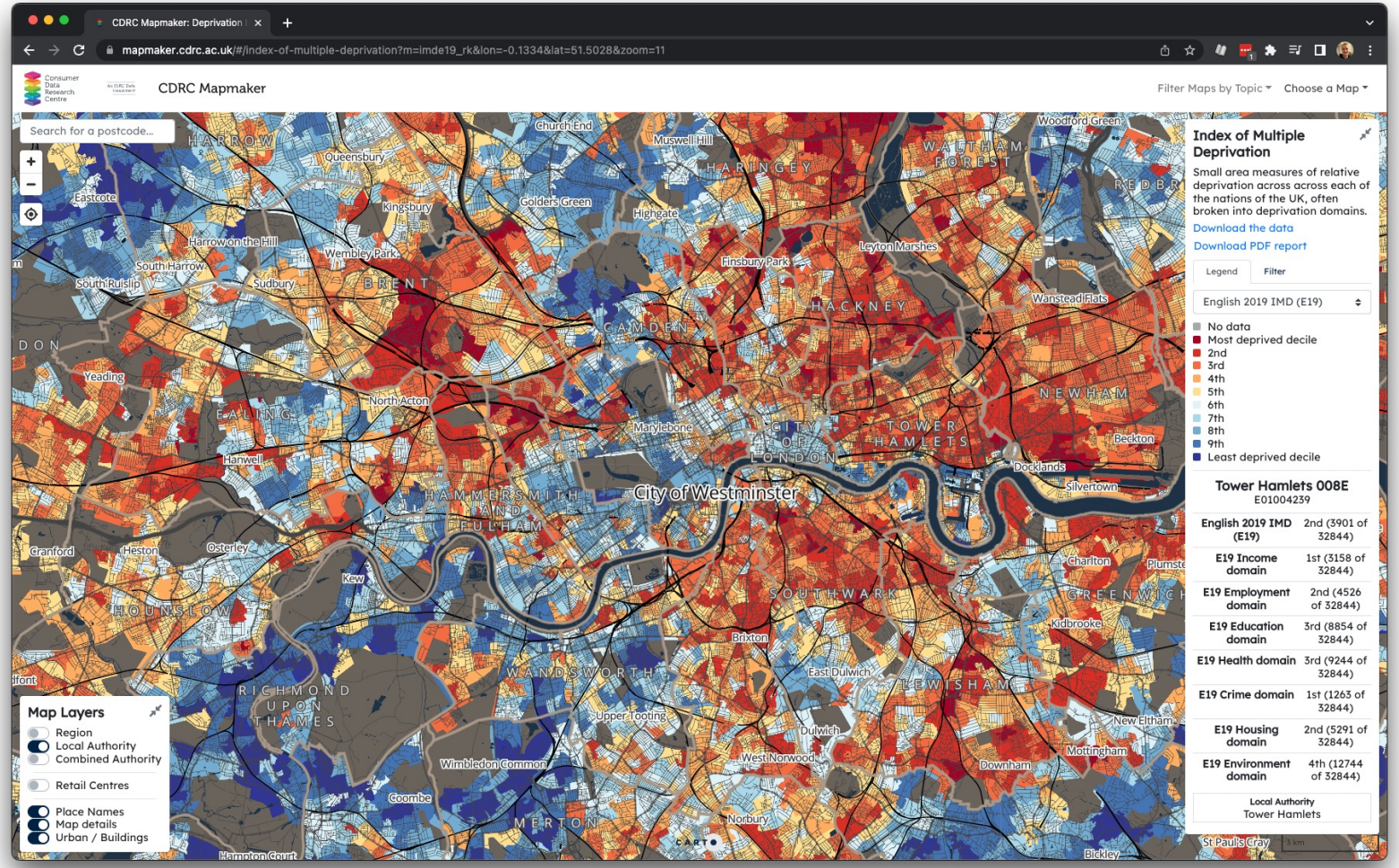
Workplace Zone Classification on mapmapker.cdrc.ac.uk

Geodemographics



Internet User Classification on mapmapker.cdrc.ac.uk

Indices



2019 Index of Multiple Deprivation on mapmapker.cdrc.ac.uk

Limitations

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

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

Internet user classification

Singleton *et al.* 2020:

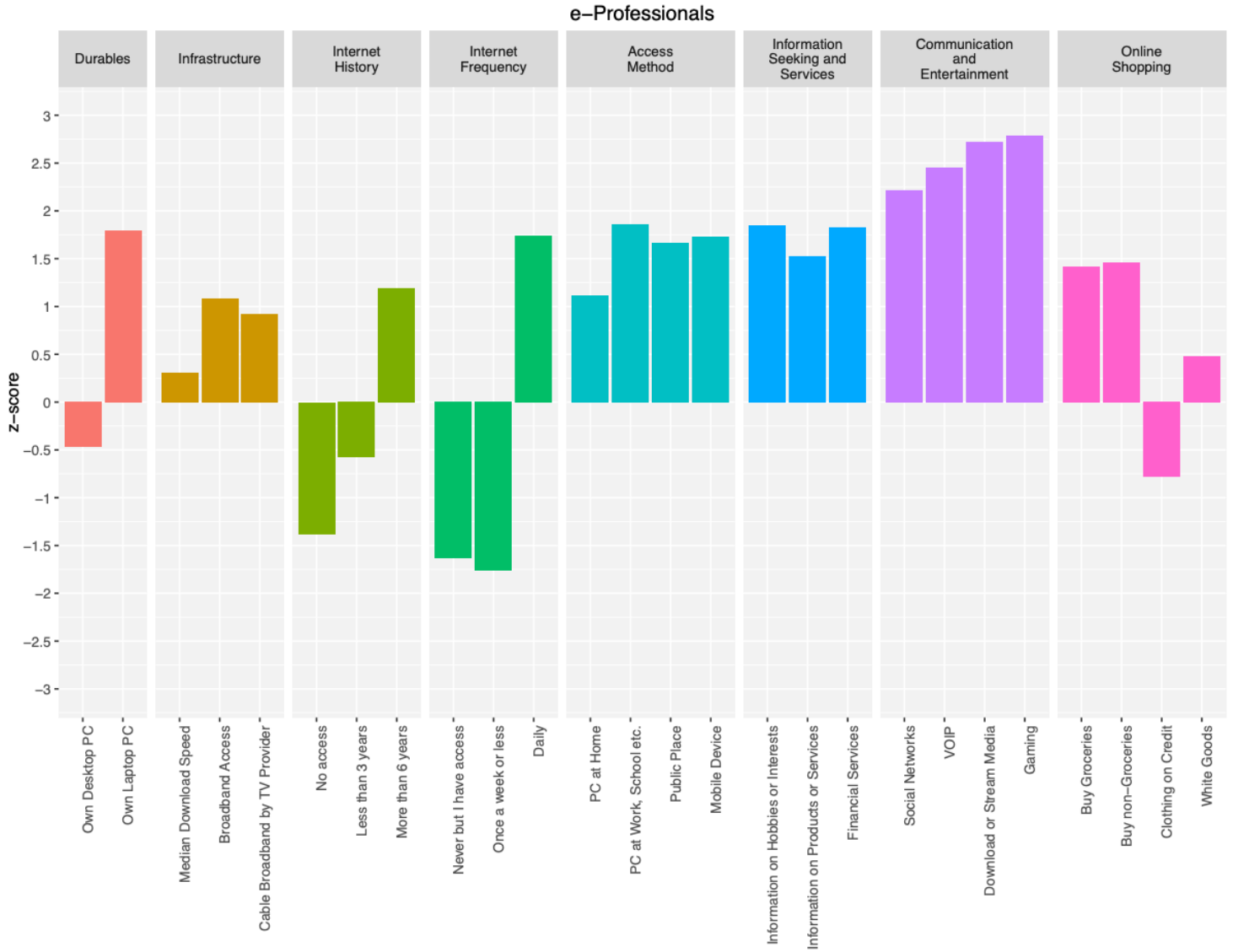
- bespoke classification created by CDRC researchers
- how do populations interact with the internet
- ‘profiles of internet use and engagement’
- built from a range of consumer data, survey data, and open data
- classification is available as open data available

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

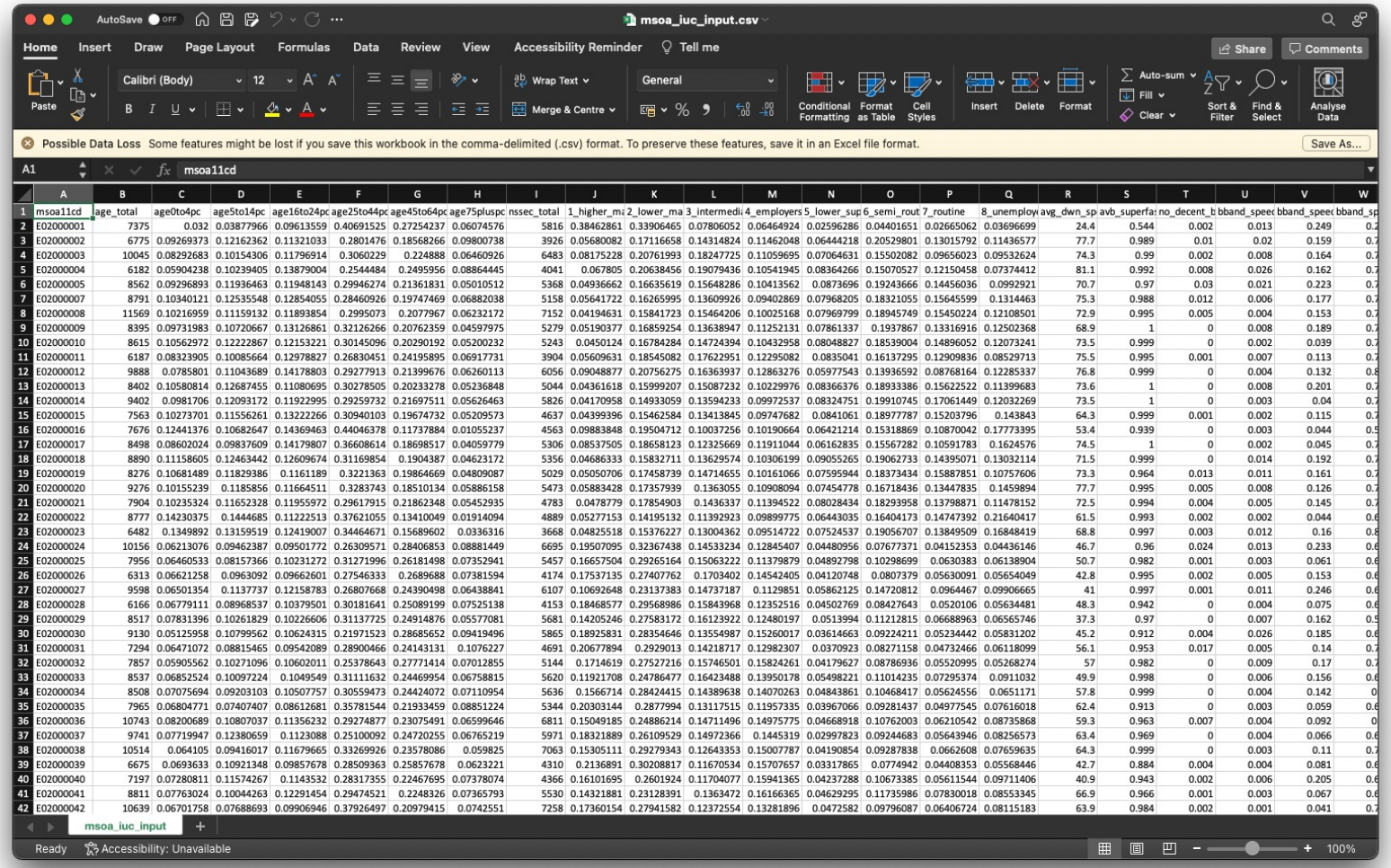
e-Professionals:

“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”

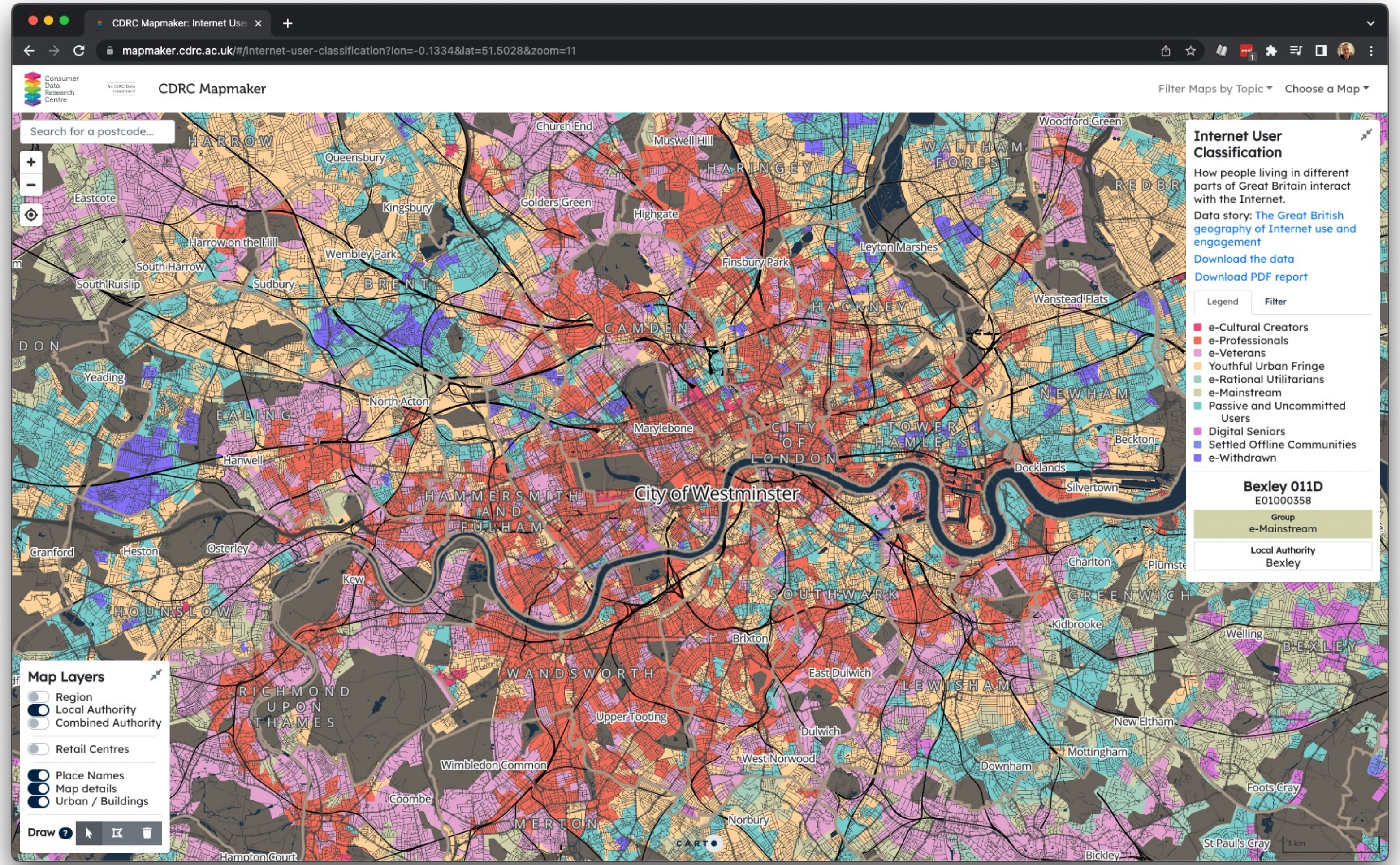
Internet user classification

- 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_iuc_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.
Save As...
A1
msoa11cd
msoa_iuc_input
Ready
Accessibility: Unavailable
100%

Workflow



CDRC Mapmaker: Internet User
mapmaker.cdrc.ac.uk/#/internet-user-classification?lon=-0.1334&lat=51.5028&zoom=11
CDRC Mapmaker
Filter Maps by Topic
Choose a Map
Search for a postcode...
Internet User Classification
How people living in different parts of Great Britain interact with the Internet.
Data story: The Great British geography of Internet use and engagement
Download the data
Download PDF report
Legend
Filter
e-Cultural Creators
e-Professionals
e-Veterans
Youthful Urban Fringe
e-Rational Utilitarians
e-Mainstream
Passive and Uncommitted Users
Digital Seniors
Settled Offline Communities
e-Withdrawn
Bexley 011D
E01000358
Group
e-Mainstream
Local Authority
Bexley
Map Layers
Region
Local Authority
Combined Authority
Retail Centres
Place Names
Map details
Urban / Buildings
Draw
City of Westminster

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?

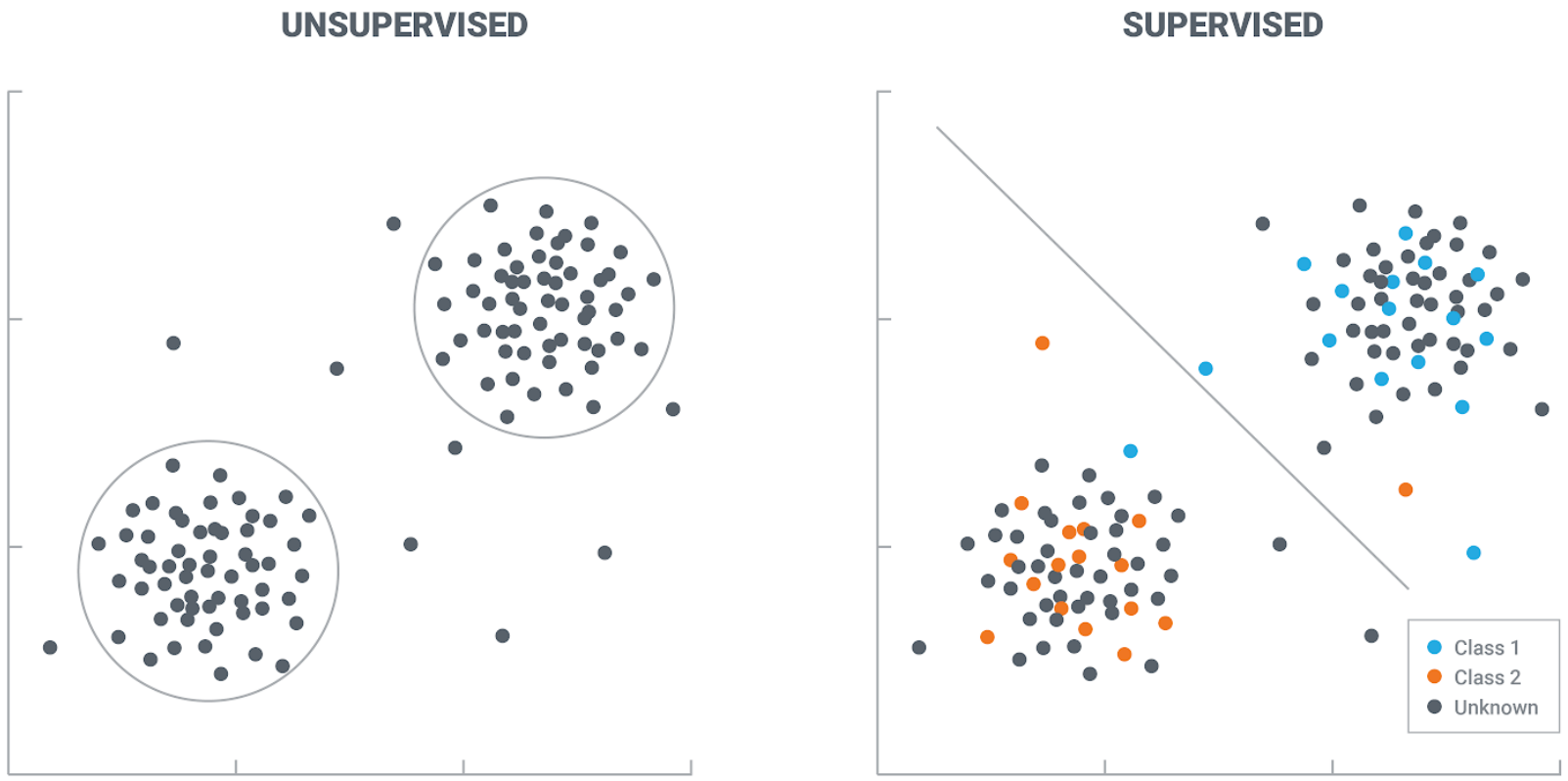
- Coefficient (or r-square value) is presented on a scale between -1 and 1.
- The greater the value, the greater the association between the two variables.
- Set an appropriate threshold - as a rule of thumb, two variables with coefficients greater than ±0.8 can be considered to be highly correlated.
- 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



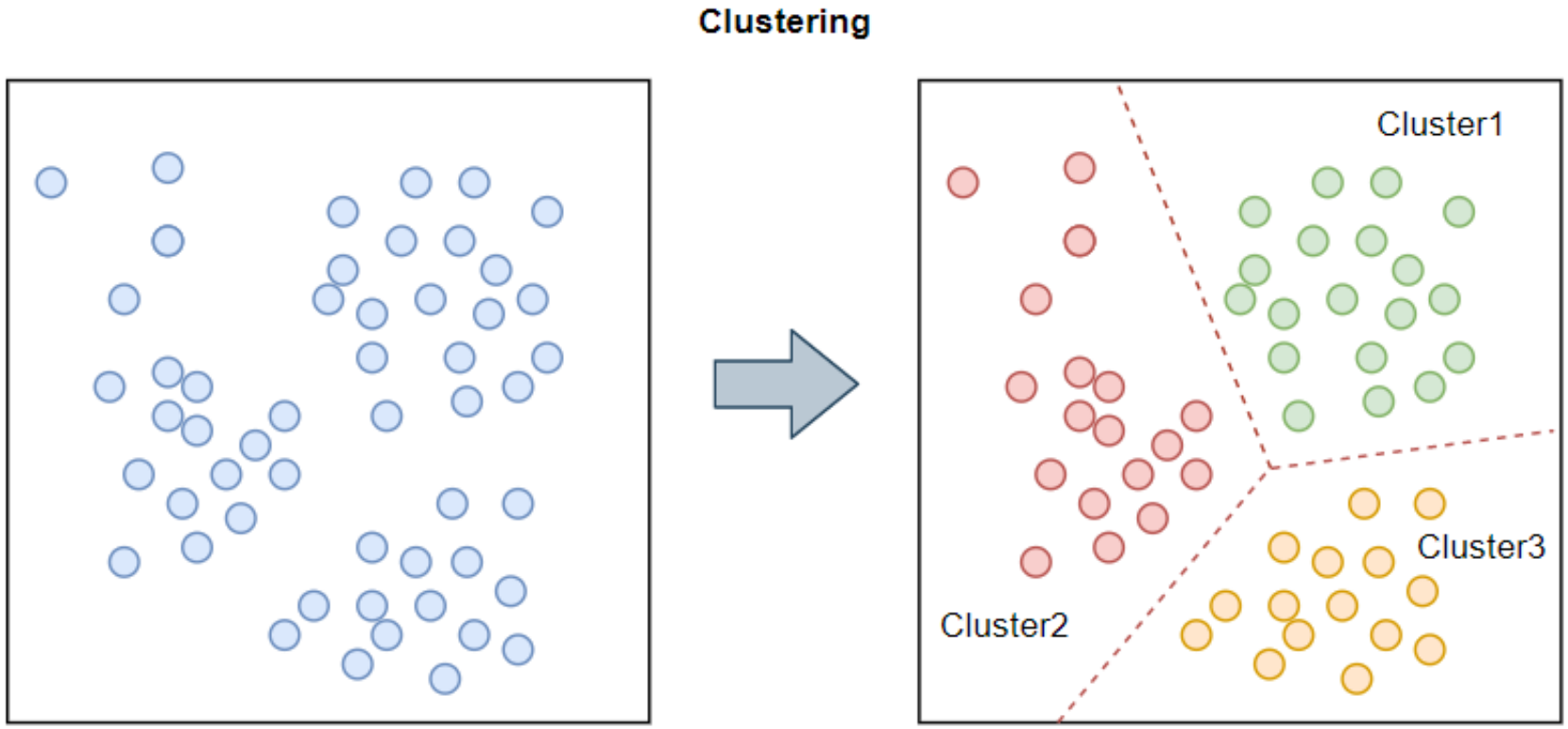
k-means

- Assign geographic areas with common underlying attributes to similar classification groups.
- Used in: Internet User Classification, ONS 2011 Output Area Classification.

k-means

- `k` clusters (pre-defined) of `n` individual observations.
- Each observation can have any number of attribute data.
- Choice of data is a balancing act: theory, available data, statistical considerations.

k-means



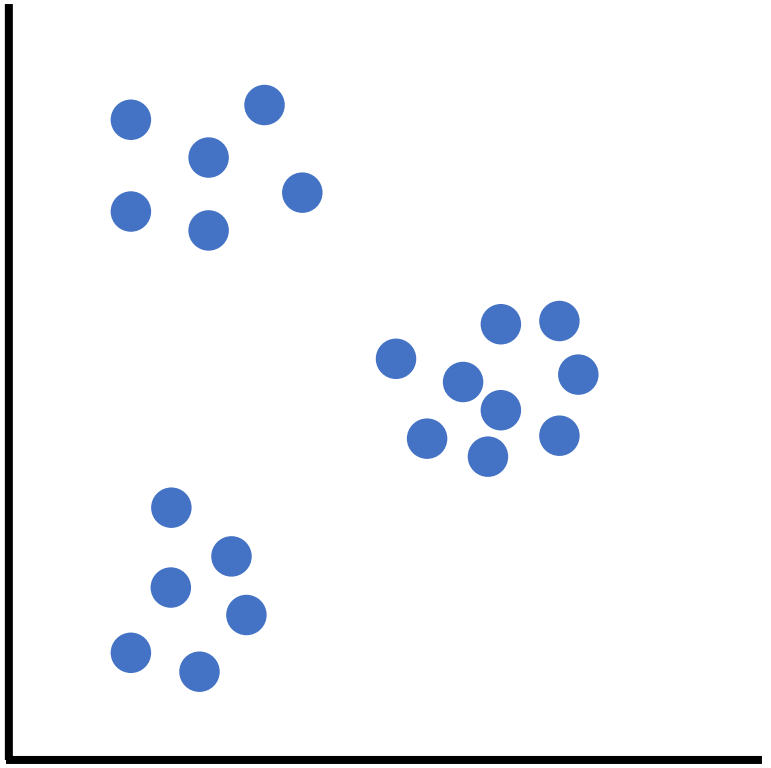
k-means

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

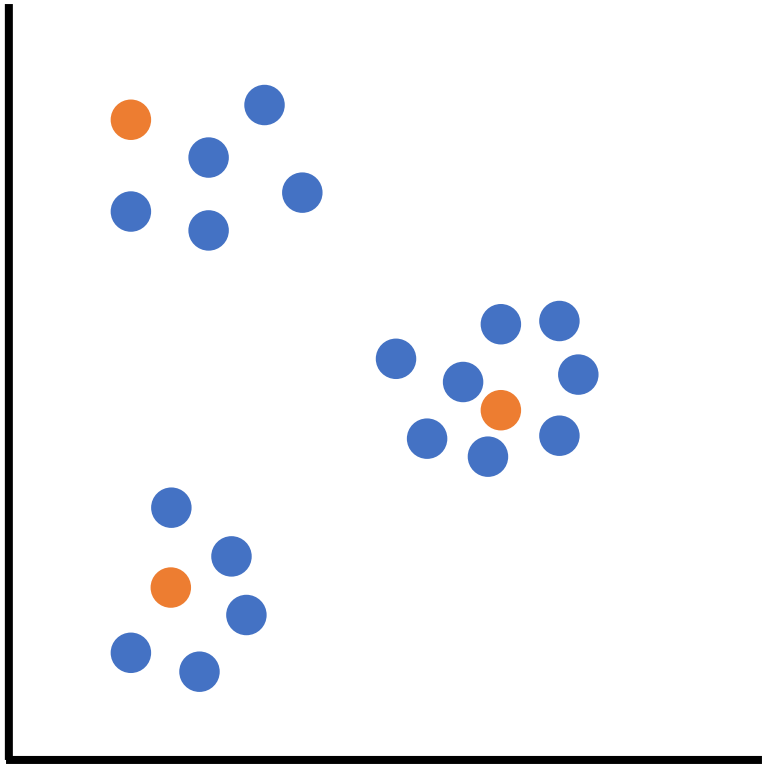
Process

- 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

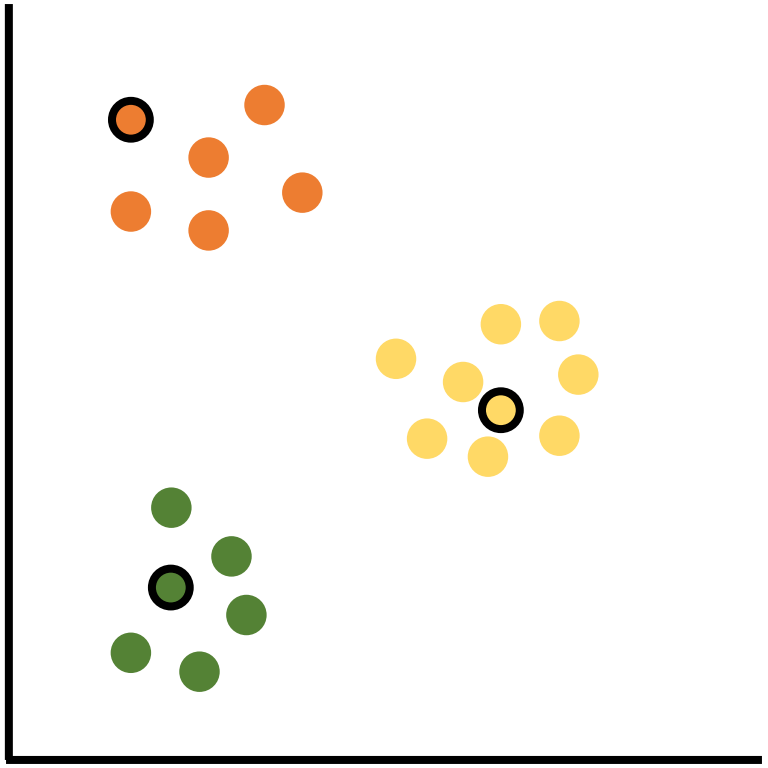
Step 1



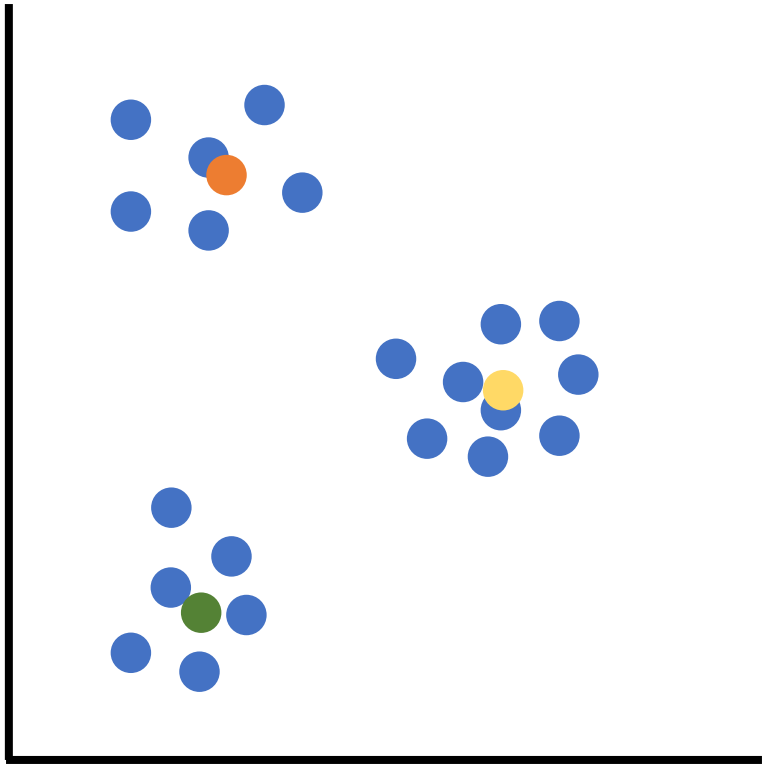
Step 2



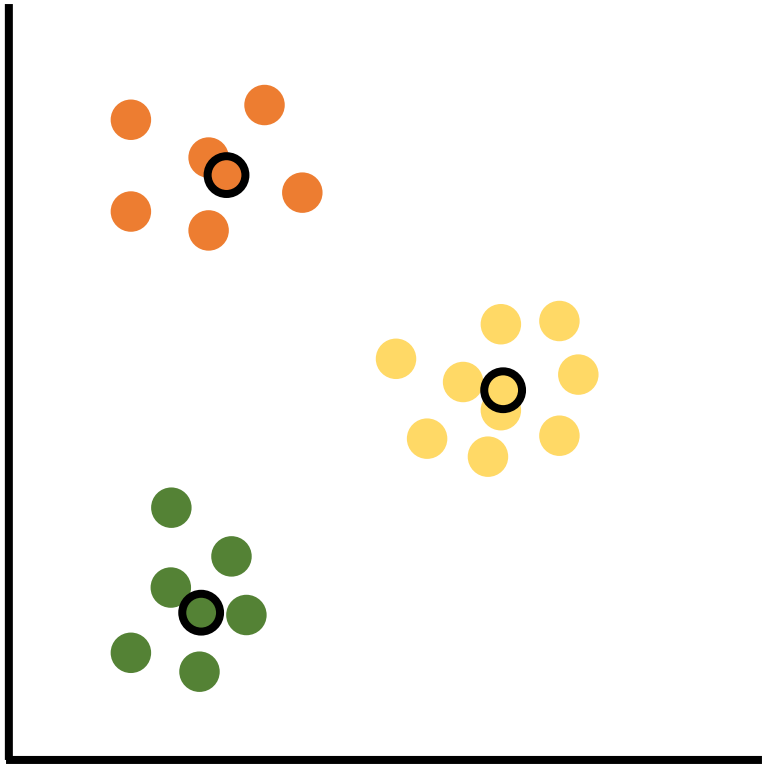
Step 3



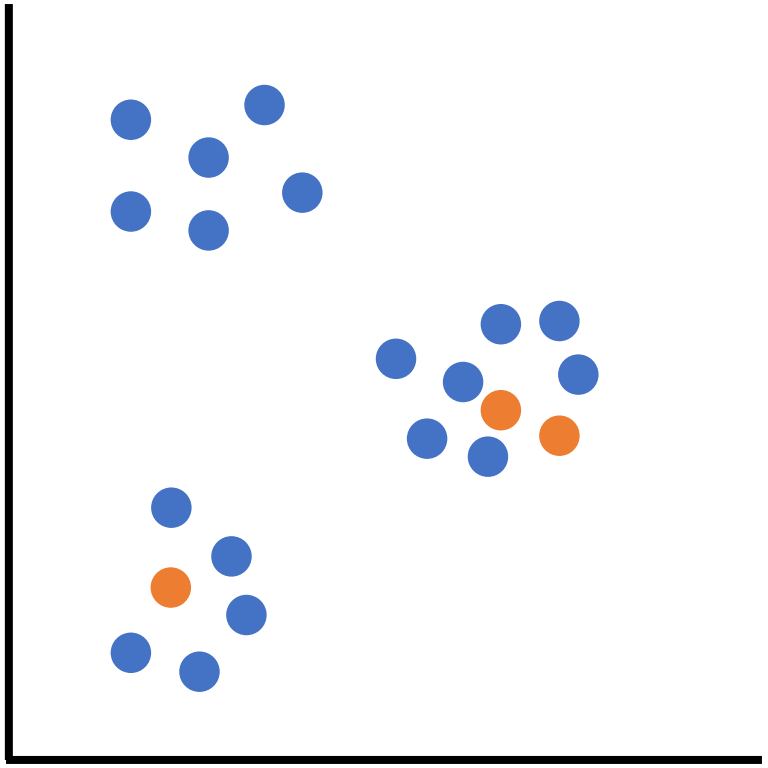
Step 4



Step 5



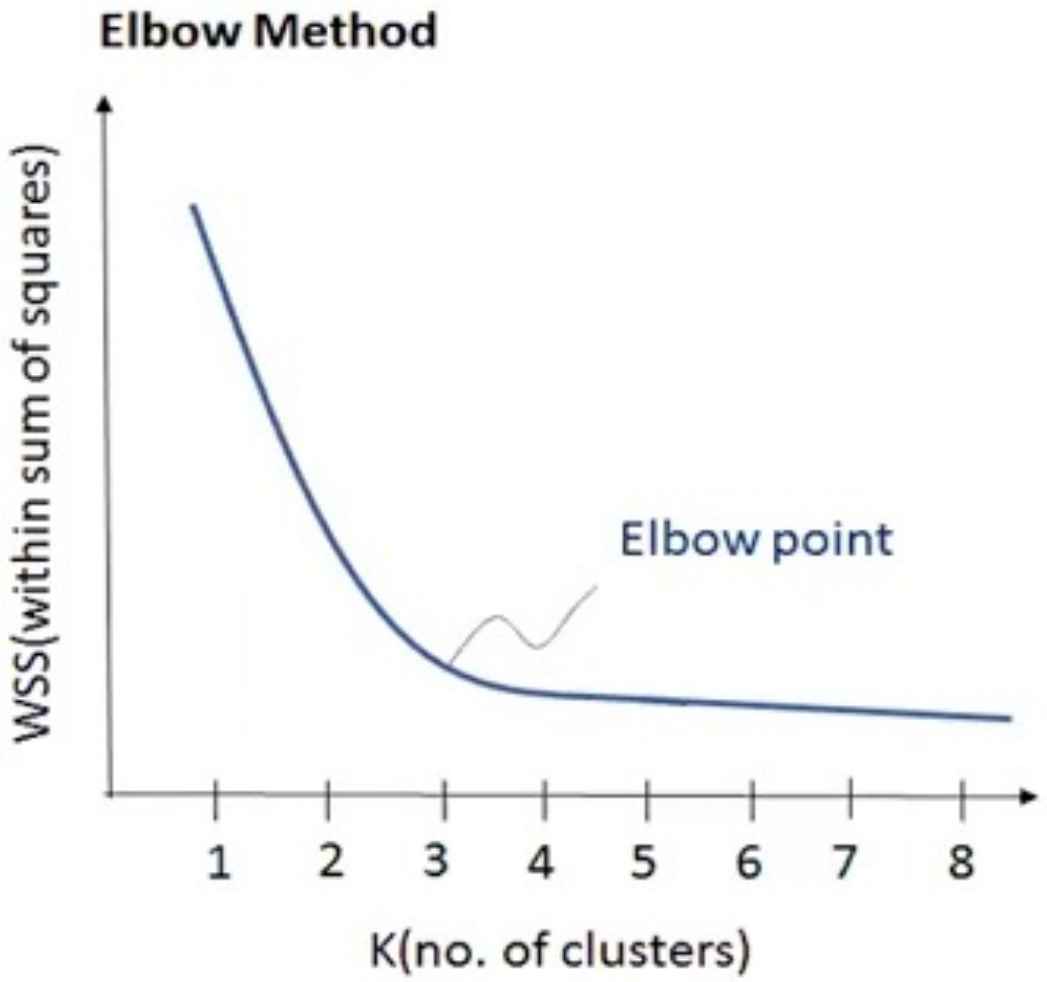
Multiple iterations



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.


BONUS
BONUS
BONUS
BONUS

Automation

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

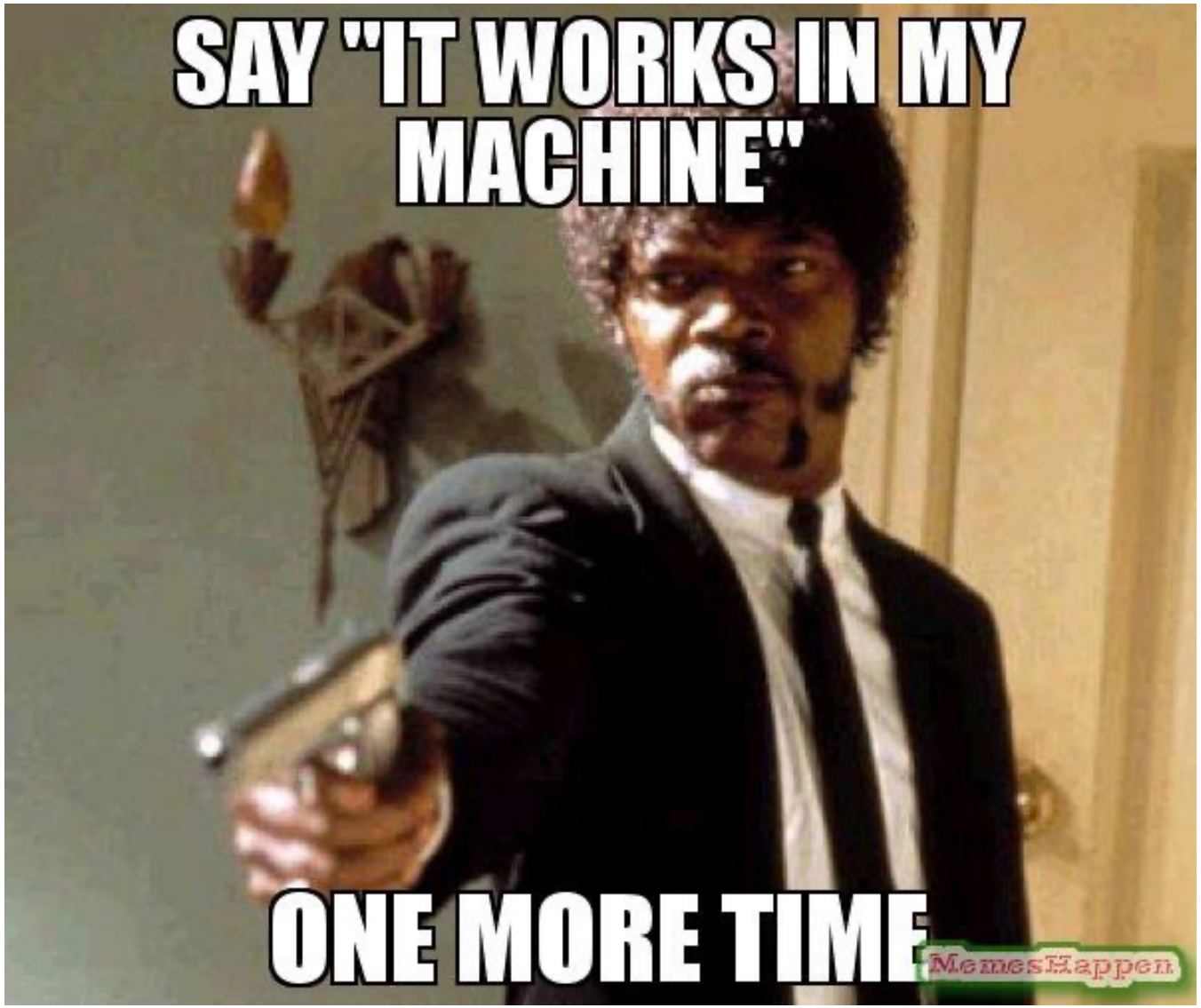
… but you can combine these into functions and use your own arguments and parameters.

… but if you use some workflows repeatedly and across projects you could think of create 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
quickmeme.com


SAY "IT WORKS IN MY MACHINE"
ONE MORE TIME
MemesHappen

```r
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

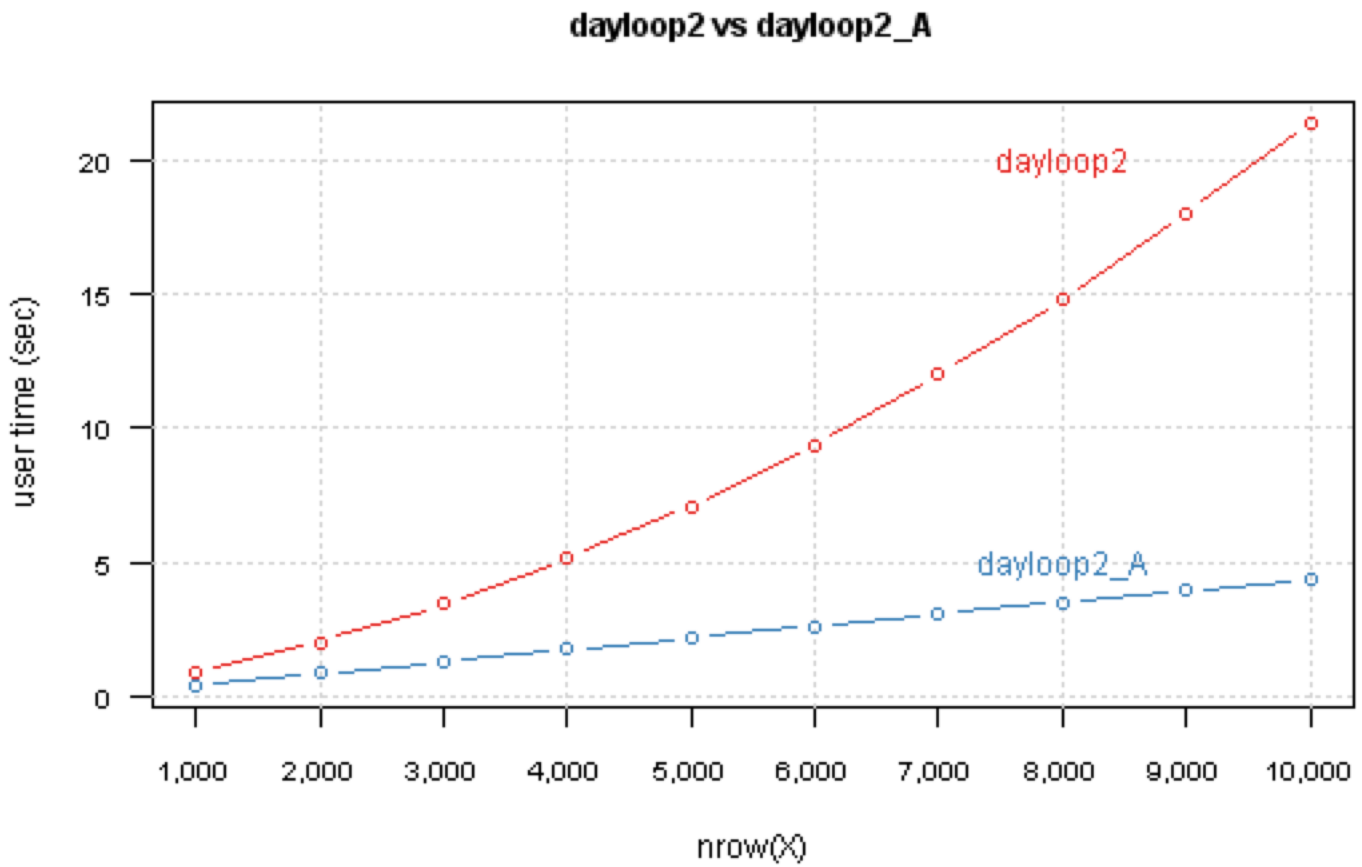
~ 850k rows

```
dayloop2 <- function(temp){
    for (i in 1:nrow(temp)){
        temp[i,10] <- i
        if (i > 1) {
            if ((temp[i,2] == temp[i-1,2]) & (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)
}
```

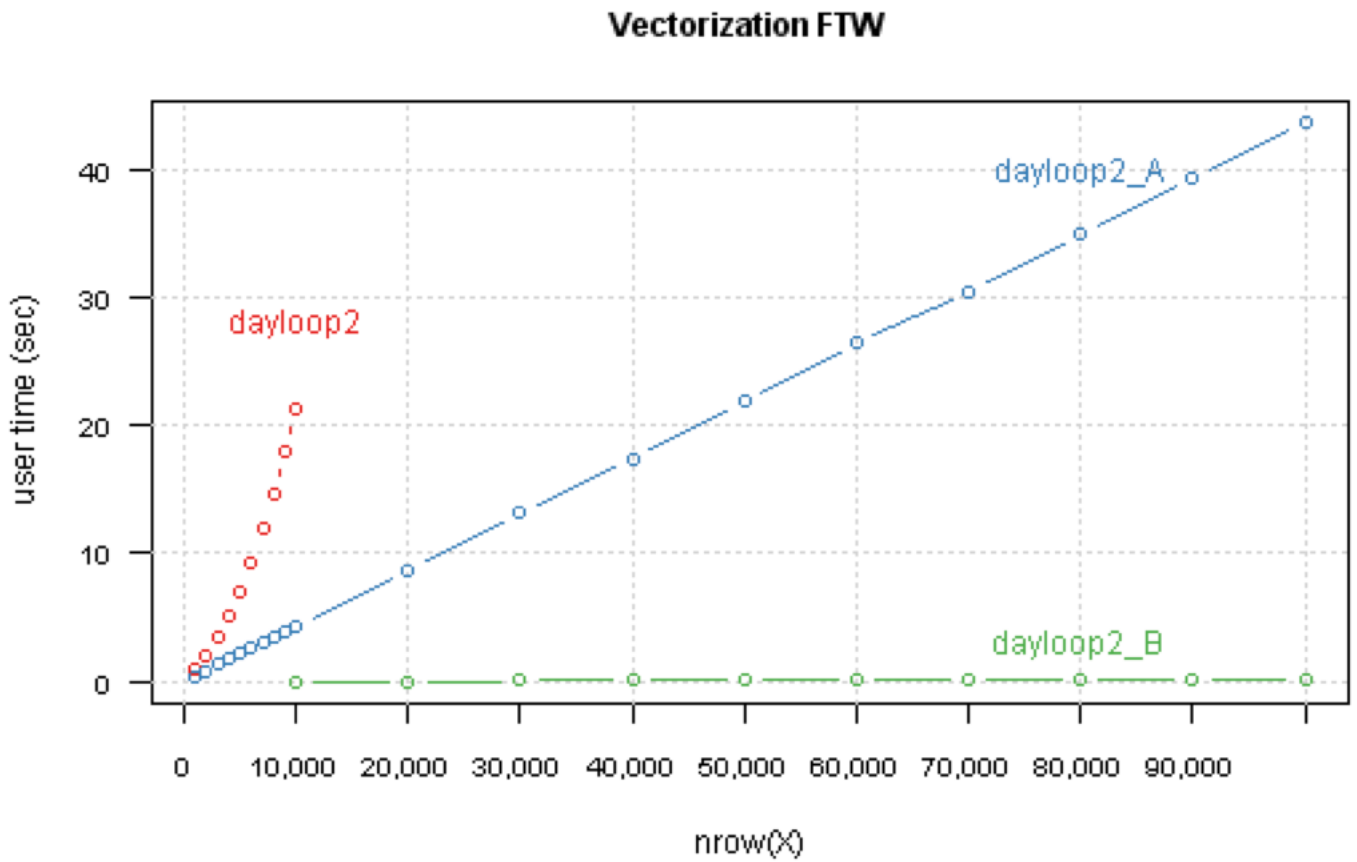
> 10 hours

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

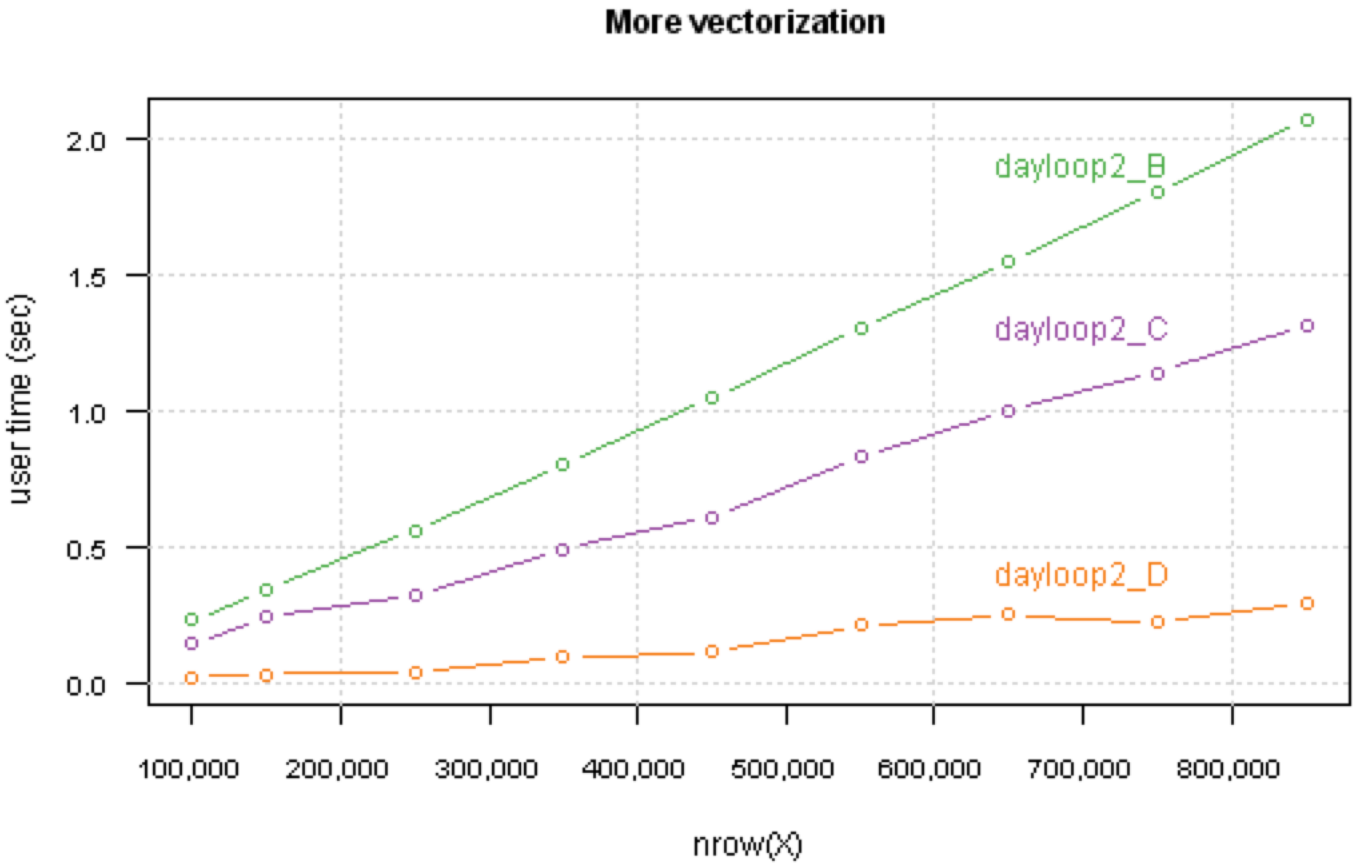
~ 850k rows



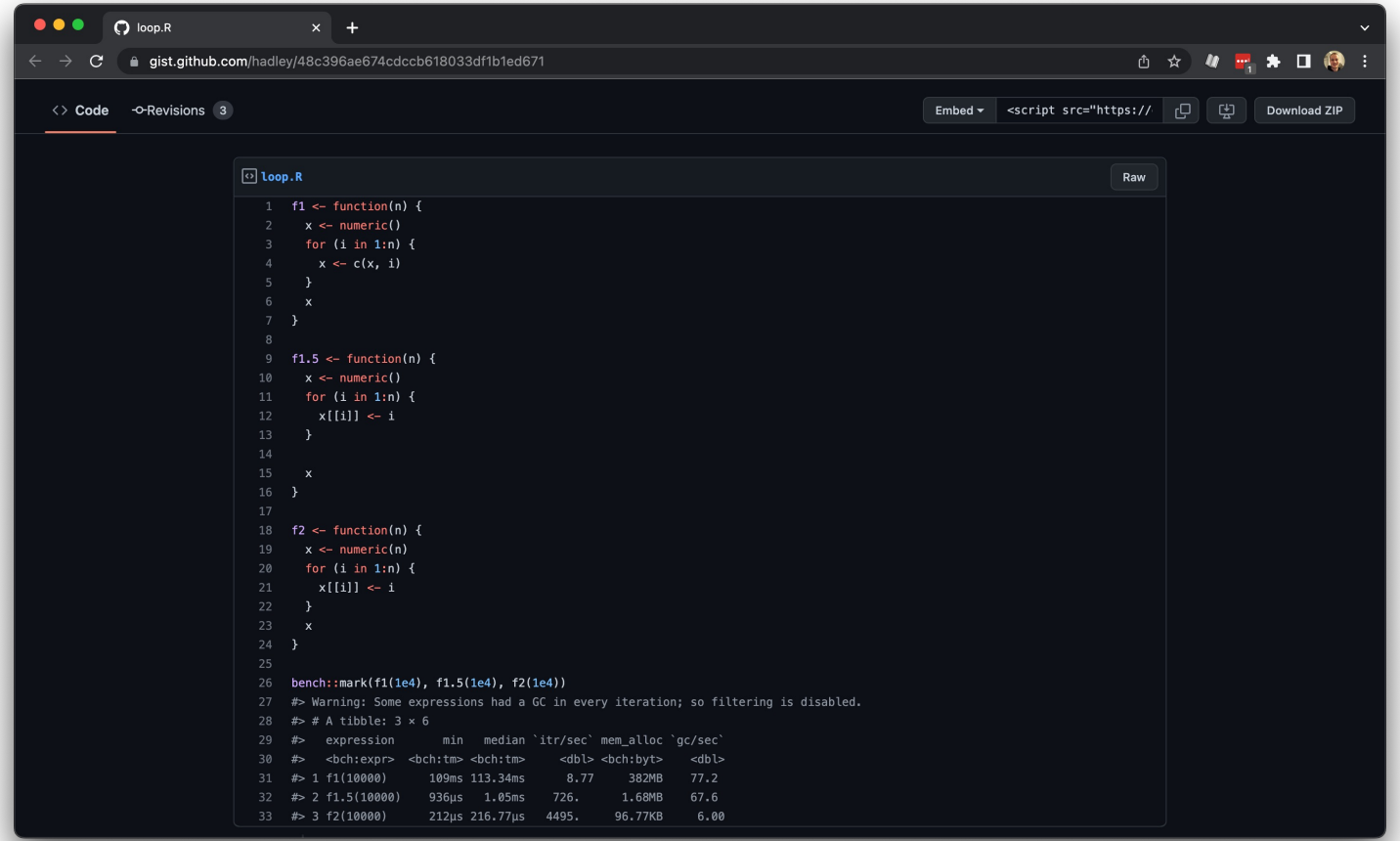
reduce indexing



vectorisation



more vectorisation


loop.R
gist.github.com/hadley/48c396ae674cdccb618033df1b1ed671
Code
Revisions 3
Embed
<script src="https://
Download ZIP
loop.R
Raw
f1 <- function(n) {
x <- numeric()
for (i in 1:n) {
x <- c(x, i)
}
x
}
f1.5 <- function(n) {
x <- numeric()
for (i in 1:n) {
x[[i]] <- i
}
x
}
f2 <- function(n) {
x <- numeric(n)
for (i in 1:n) {
x[[i]] <- i
}
x
}
bench::mark(f1(1e4), f1.5(1e4), f2(1e4))
#> Warning: Some expressions had a GC in every iteration; so filtering is disabled.
#> # A tibble: 3 × 6
#> expression min median `itr/sec` mem_alloc `gc/sec`
#> <bch:expr> <bch:tm> <bch:tm> <dbl> <bch:byt> <dbl>
#> 1 f1(10000) 109ms 113.34ms 8.77 382MB 77.2
#> 2 f1.5(10000) 936µs 1.05ms 726. 1.68MB 67.6
#> 3 f2(10000) 212µs 216.77µs 4495. 96.77KB 6.00

Speeding up

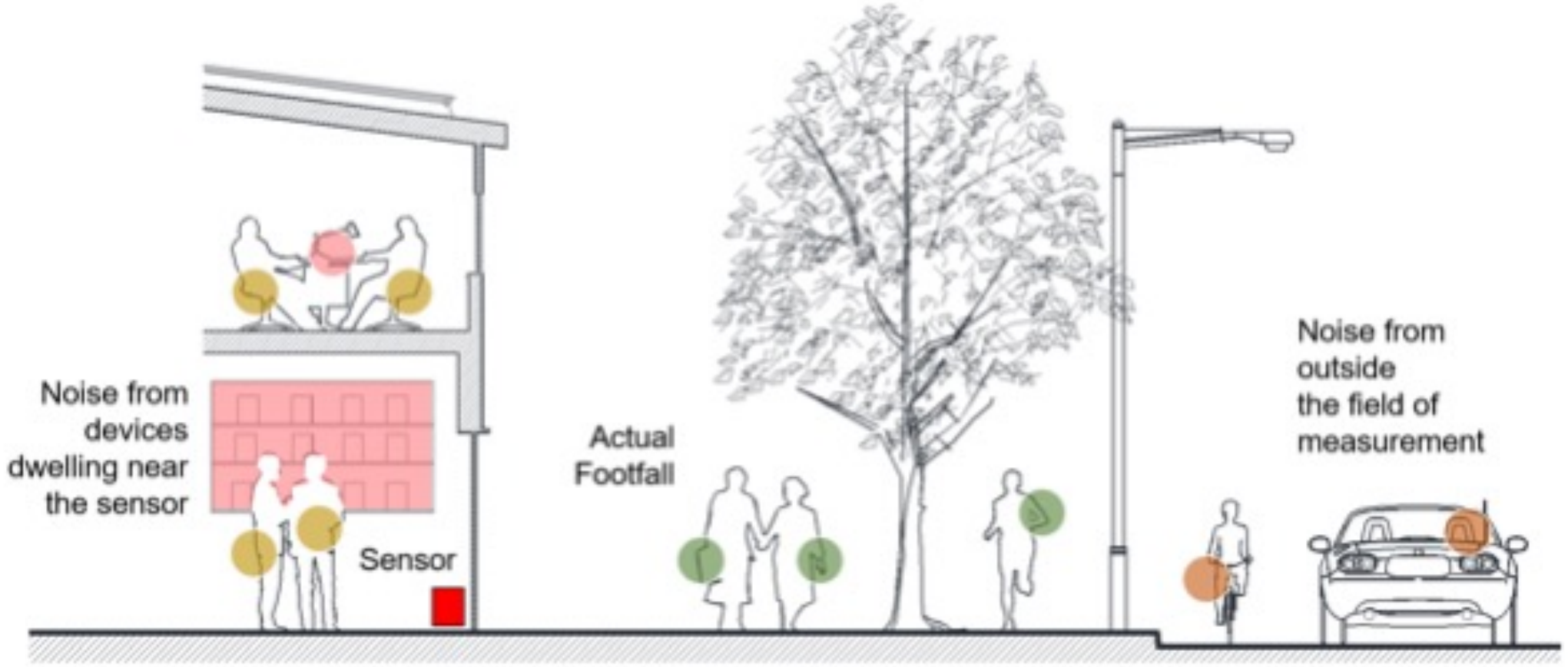
First things to check:

- 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

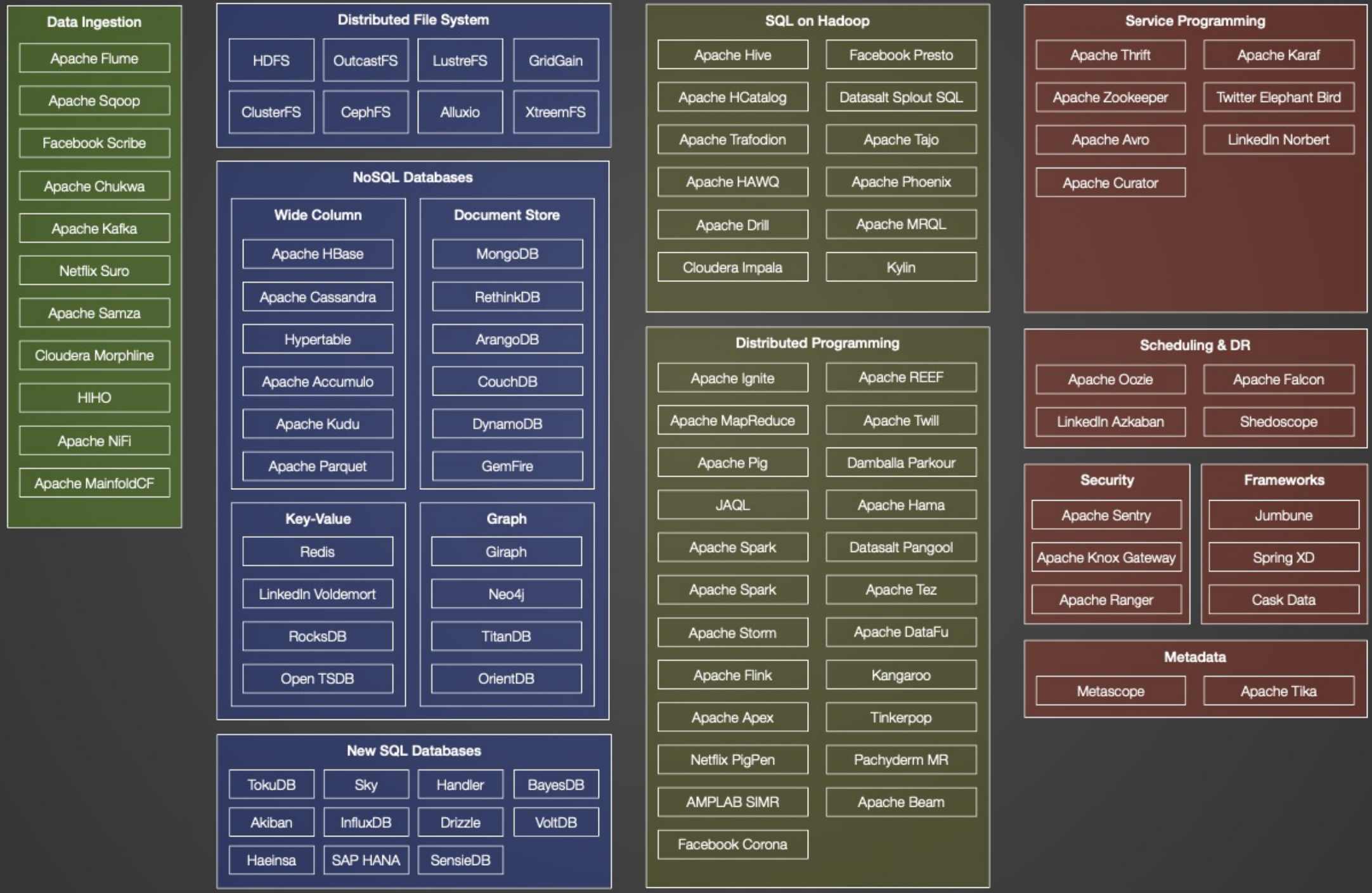
Smart street sensor project CDRC

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


Data Ingestion
Apache Flume
Apache Sqoop
Facebook Scribe
Apache Chukwa
Apache Kafka
Netflix Suro
Apache Samza
Cloudera Morphline
HIHO
Apache NiFi
Apache MainfoldCF
Distributed File System
HDFS
OutcastFS
LustreFS
GridGain
ClusterFS
CephFS
Alluxio
XtreemFS
NoSQL Databases
Wide Column
Apache HBase
Apache Cassandra
Hypertable
Apache Accumulo
Apache Kudu
Apache Parquet
Document Store
MongoDB
RethinkDB
ArangoDB
CouchDB
DynamoDB
GemFire
Key-Value
Redis
LinkedIn Voldemort
RocksDB
Open TSDB
Graph
Giraph
Neo4j
TitanDB
OrientDB
New SQL Databases
TokuDB
Sky
Handler
BayesDB
Akiban
InfluxDB
Drizzle
VoltDB
Haeinsa
SAP HANA
SensieDB
SQL on Hadoop
Apache Hive
Facebook Presto
Apache HCatalog
Datasalt Splout SQL
Apache Trafodion
Apache Tajo
Apache HAWQ
Apache Phoenix
Apache Drill
Apache MRQL
Cloudera Impala
Kylin
Distributed Programming
Apache Ignite
Apache REEF
Apache MapReduce
Apache Twill
Apache Pig
Damballa Parkour
JAQL
Apache Hama
Apache Spark
Datasalt Pangool
Apache Spark
Apache Tez
Apache Storm
Apache DataFu
Apache Flink
Kangaroo
Apache Apex
Tinkerpop
Netflix PigPen
Pachyderm MR
AMPLAB SIMR
Apache Beam
Facebook Corona
Service Programming
Apache Thrift
Apache Karaf
Apache Zookeeper
Twitter Elephant Bird
Apache Avro
LinkedIn Norbert
Apache Curator
Scheduling & DR
Apache Oozie
Apache Falcon
LinkedIn Azkaban
Shedoscope
Security
Apache Sentry
Apache Knox Gateway
Apache Ranger
Frameworks
Jumbune
Spring XD
Cask Data
Metadata
Metascope
Apache Tika

Medium data toolkit

Soundararaj 2019:

- Unix philosophy: pipes and text streams
- Command-line tools: `cat`, `find`, `sort`, `uniq`, `sed`, `grep`, `awk`
- Parallelisation: using `xargs` 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