Implementing PCCF+ in R (or SQL Server) for Geographic Allocation

APHEO GIS Interest Group

Allison Maier

Research Associate

Knowledge Management

allison.maier@kflaph.ca

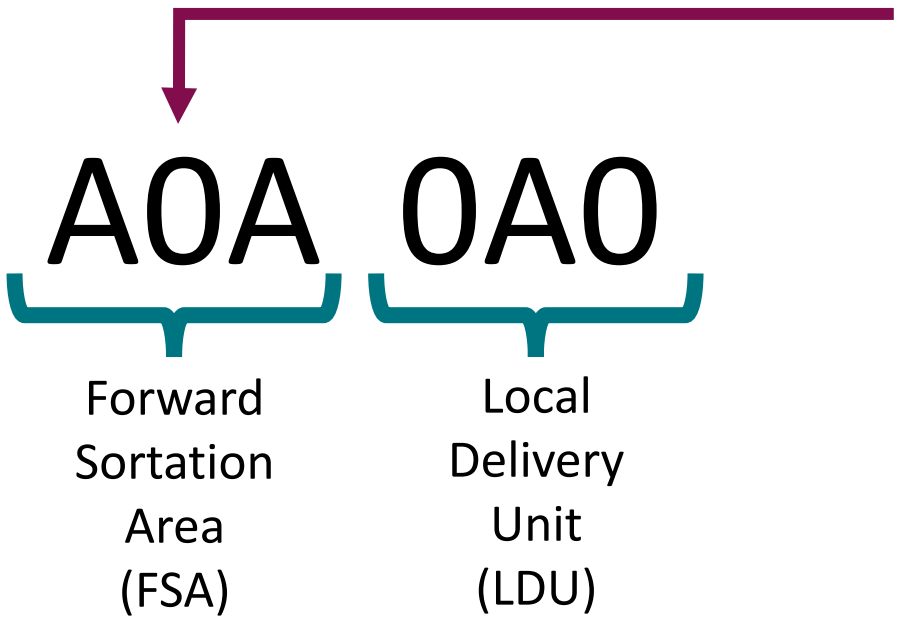
September 26th, 2023

KFL&A Public Health

Outline

- The ins and outs of postal codes
- The problem of geographic allocation using postal codes
- The problematic solution - PCCF
- The proper solution – PCCF+
- KFLAPH's PCCF+ implementation (R and SQL Server)

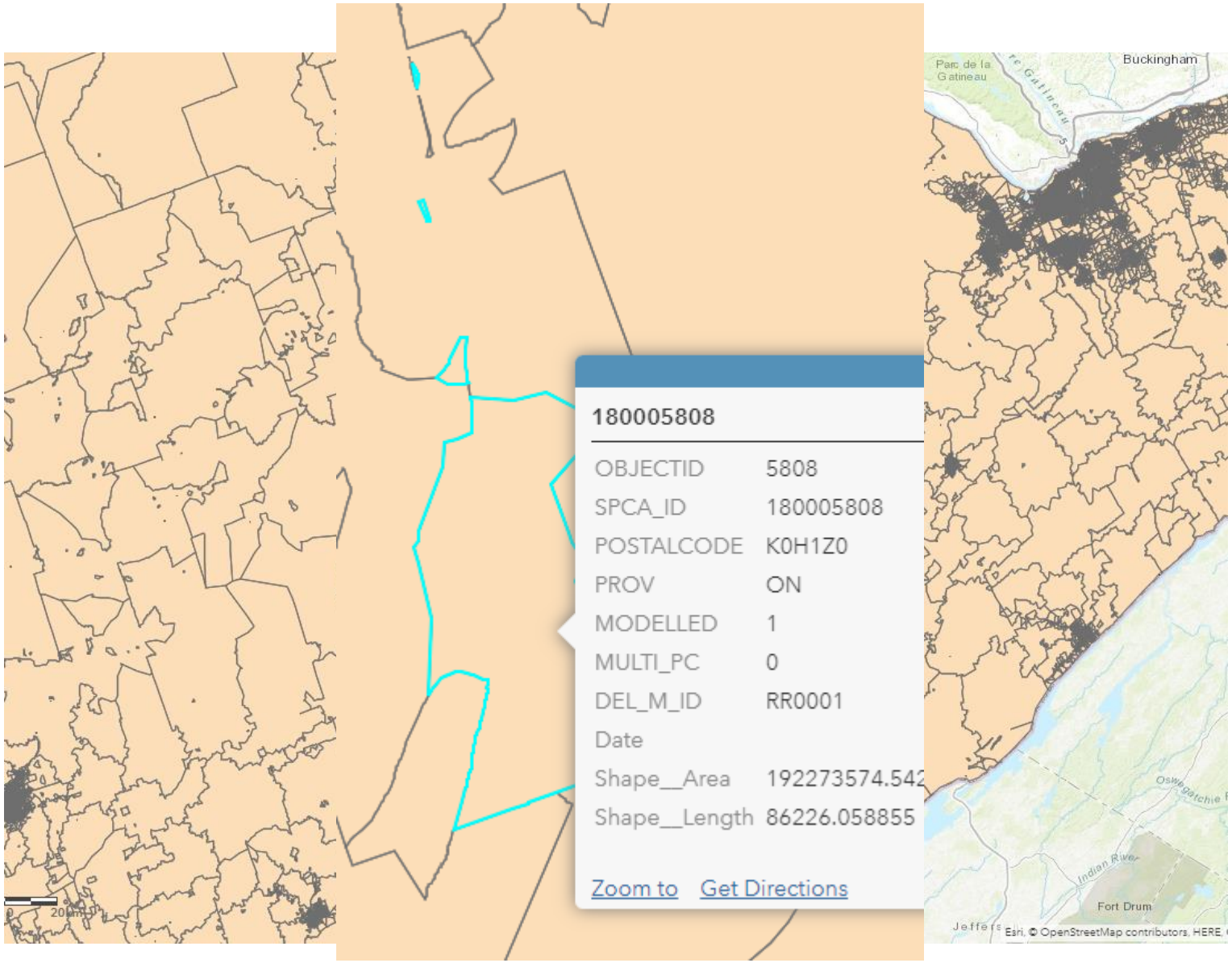
What are postal codes really?



- Around 900,000 postal codes in Canada
  - Over 200,000 in Ontario
- Fundamentally different between urban and rural areas

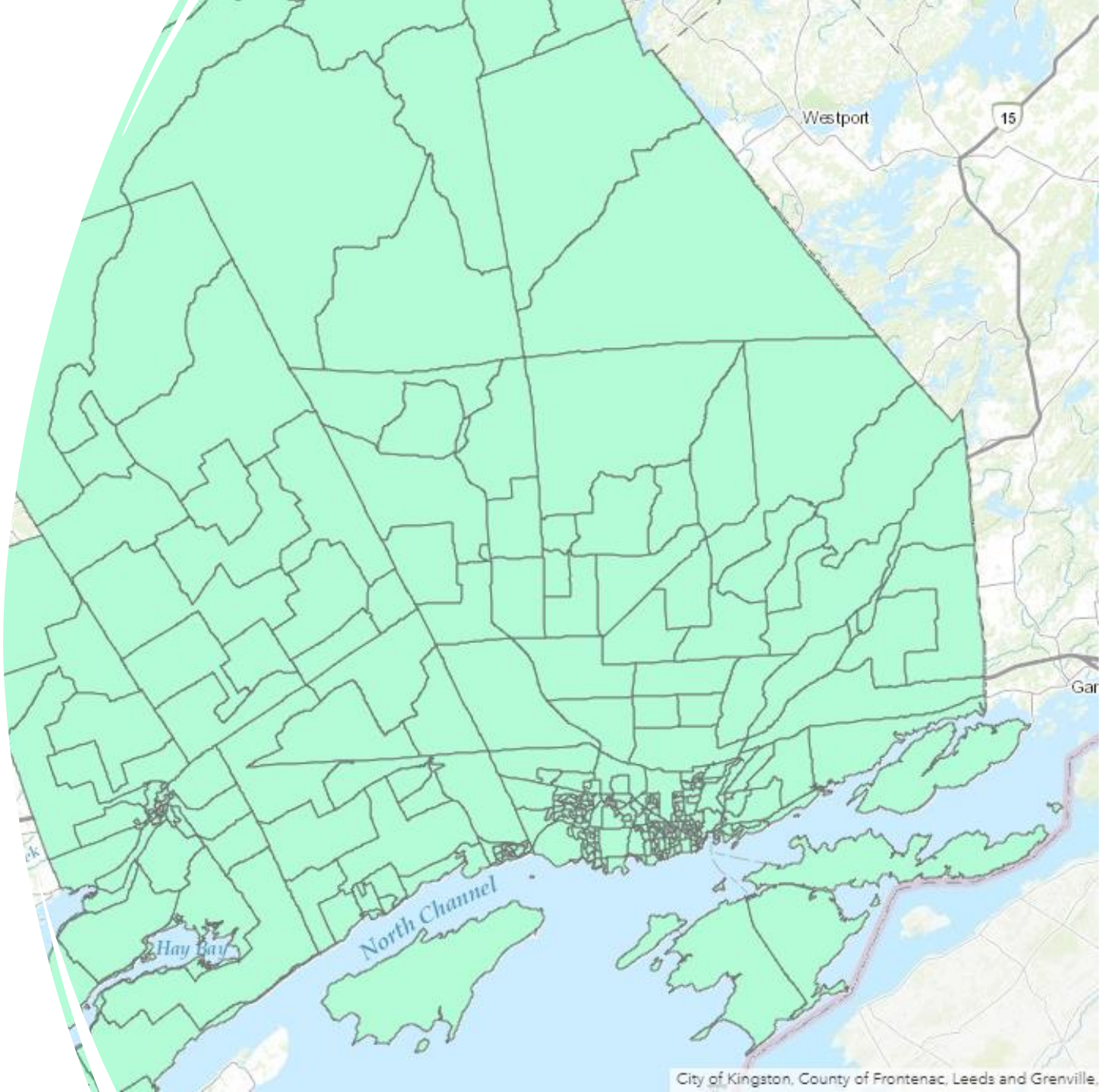
What are postal codes really NOT?

A meaningful unit of geography



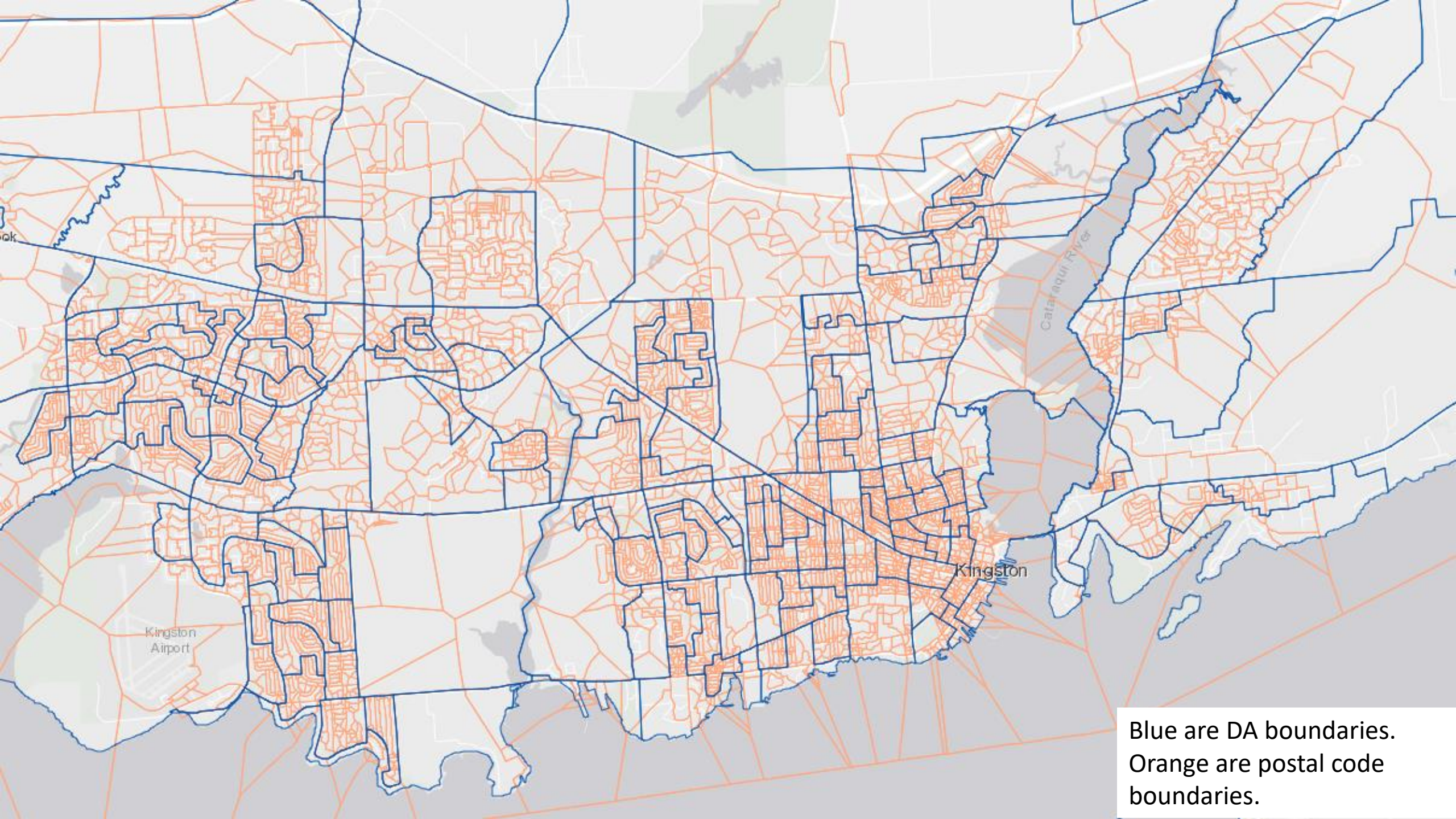
Statistics Canada is the source of meaningful geographies

- Dissemination Area (DA): a small, relatively stable geographic unit. It is the smallest standard geographic area for which all census data are disseminated. DAs cover all the territory of Canada.

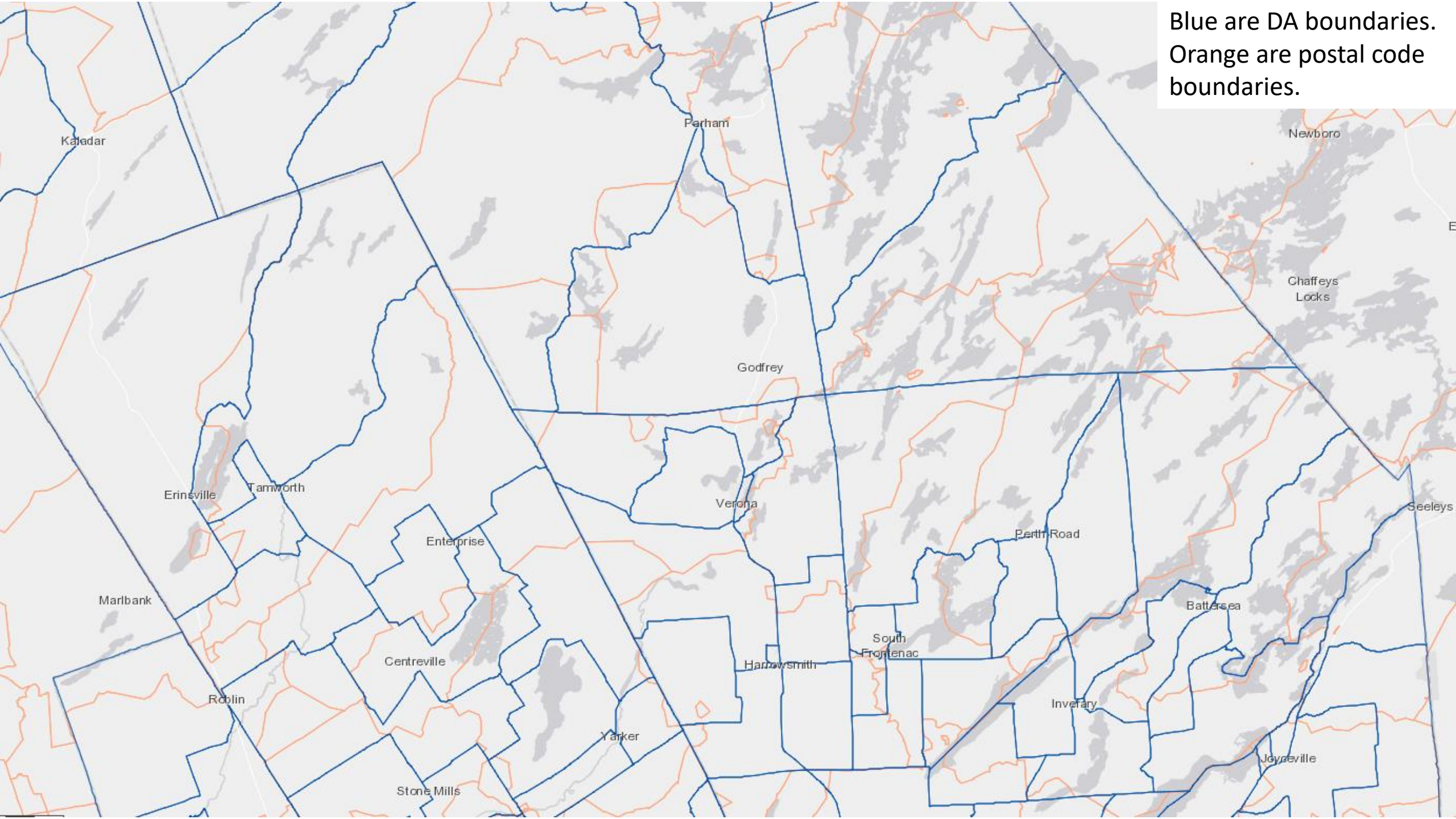


So now we need a way to use someone's postal code to get their DA

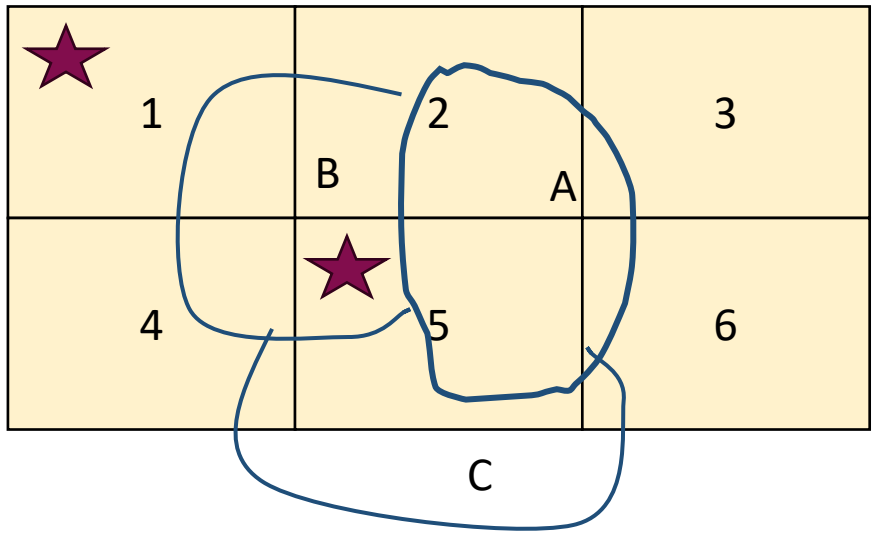
The problematic solution – PCCF (Postal Code Conversion File) – a 1:1 association between postal code and DA



Blue are DA boundaries.
Orange are postal code boundaries.


Blue are DA boundaries.
Orange are postal code boundaries.
Kaladar
Parham
Newboro
Chaffeys Locks
Godfrey
Erinsville
Tamworth
Verona
Enterprise
Perth Road
Seeleys
Marlbank
Battersea
South Frontenac
Centreville
Harrowsmith
Roblin
Inverary
Yarker
Joyceville
Stone Mills

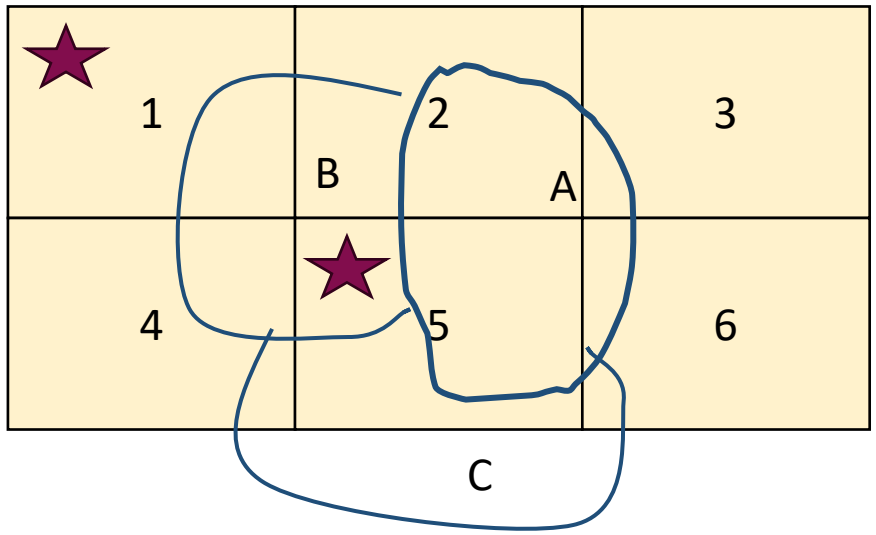
Example



Postal codes
DAs
Towns/villages

- There are 6 postal codes (1-6)
- There are 3 DAs (A, B,C)
- There are two small population centres, one in postal code 5 and one in postal code 1.

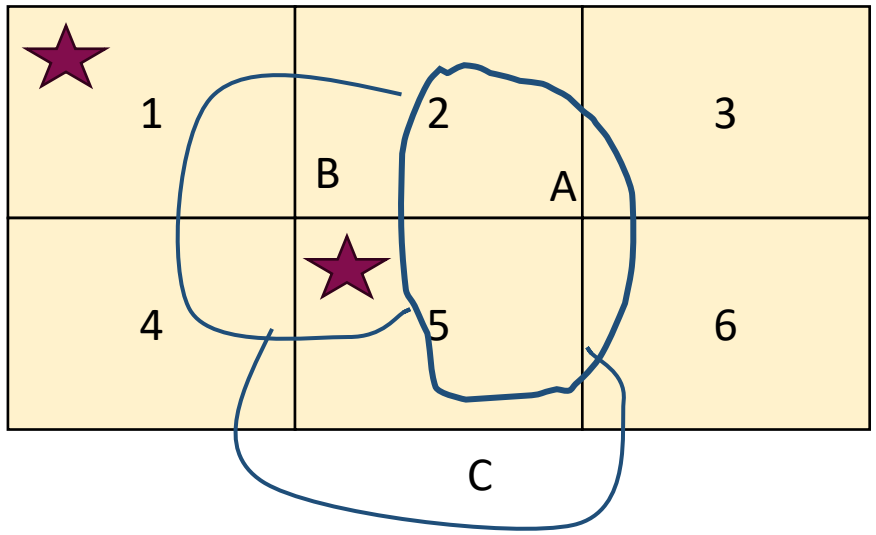
Example



Postal codes
DAs
Towns/villages

Postal Code	DA	Population Weight
1	B	70%
1	Other DAs	30%
2	A	40%
2	B	30%
2	Other DAs	30%
5	A	15%
5	B	65%
5	C	20%

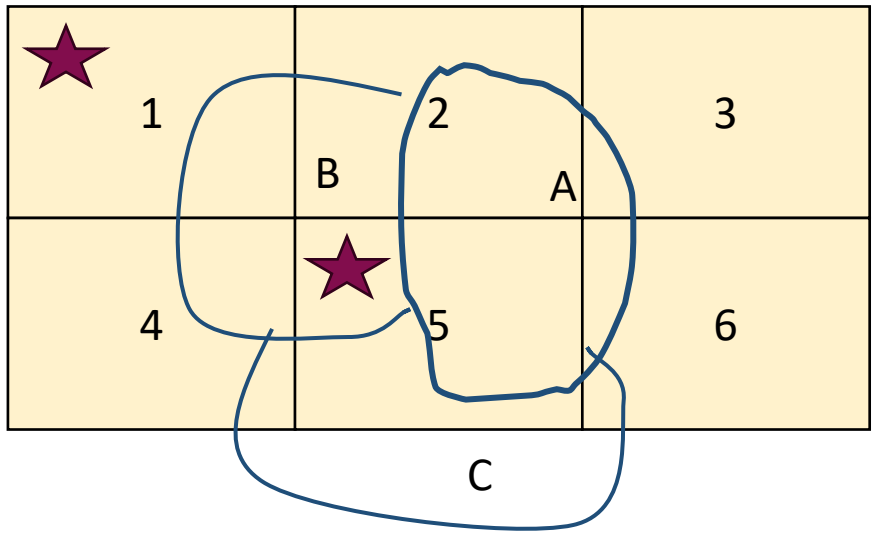
Example



Postal codes
DAs
Towns/villages

Postal Code	DA	Population Weight	Cumulative Population Weight	# People Assigned
5	A	15%	0.15	
5	B	65%	0.80	
5	C	20%	1	

Example

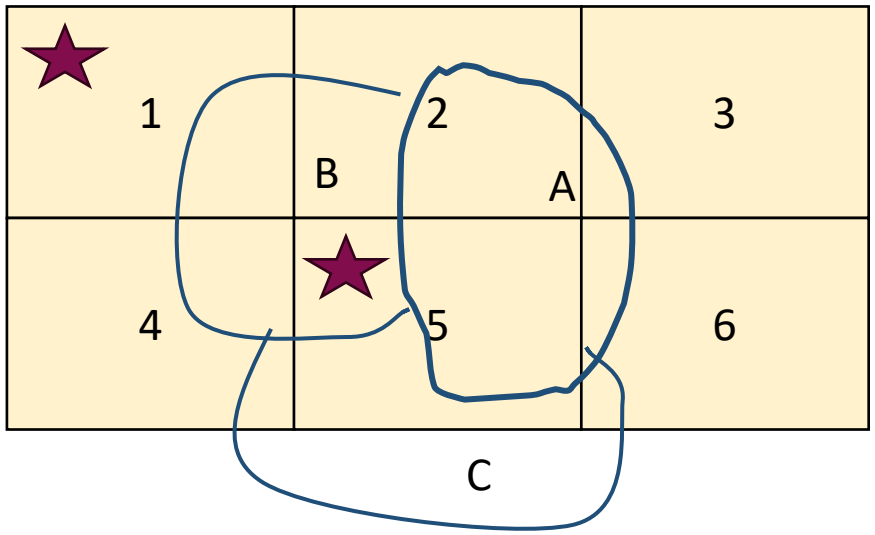


Postal codes
DAs
Towns/villages

Postal Code	DA	Population Weight	Cumulative Population Weight	# People Assigned
5	A	15%	0.15	
5	B	65%	0.80	1
5	C	20%	1	

- Random number 1: 0.7 → DA B

Example

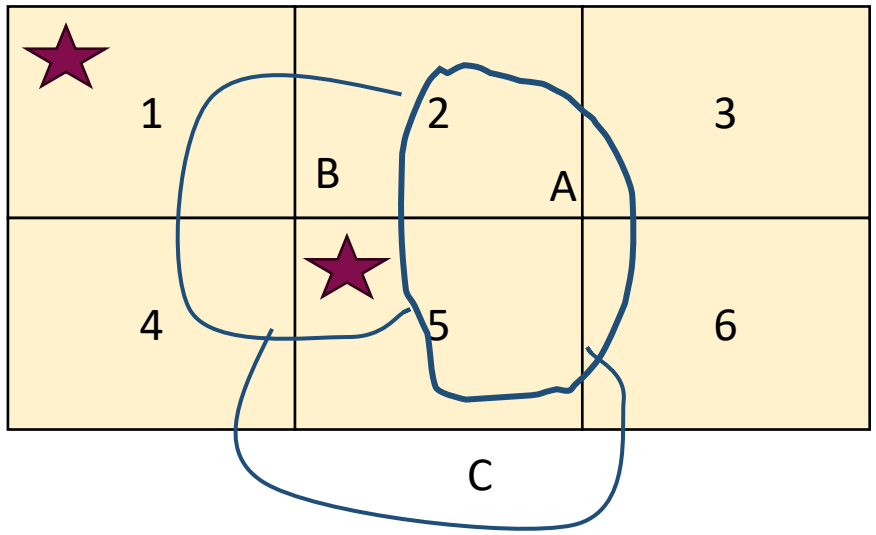


Postal codes
DAs
Towns/villages

Postal Code	DA	Population Weight	Cumulative Population Weight	# People Assigned
5	A	15%	0.15	1
5	B	65%	0.80	1
5	C	20%	1	

- Random number 1: 0.71 → DA B
- Random number 2: 0.09 → DA A

Example

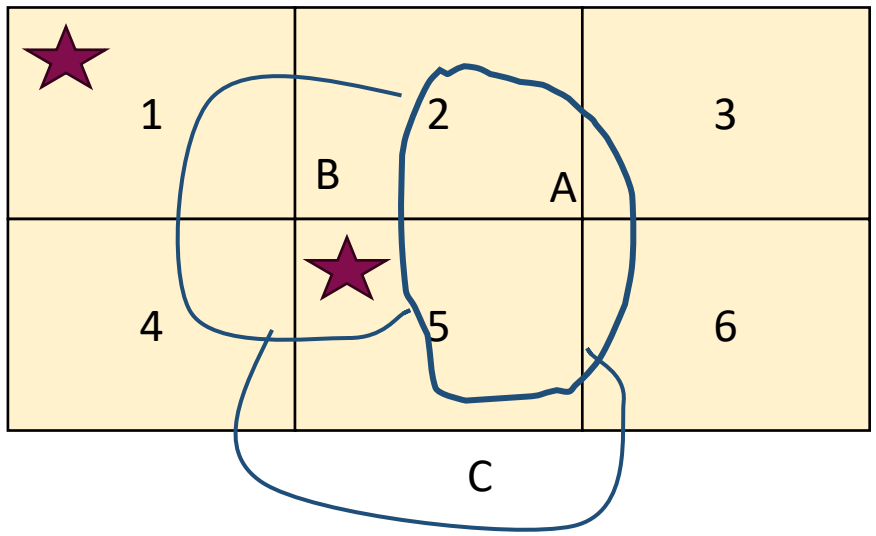


Postal codes
DAs
Towns/villages

Postal Code	DA	Population Weight	Cumulative Population Weight	# People Assigned
5	A	15%	0.15	2
5	B	65%	0.80	1
5	C	20%	1	

- Random number 1: 0.71 → DA B
- Random number 2: 0.09 → DA A
- Random number 3: 0.13 → DA A

Example



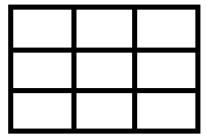
Postal codes
DAs
Towns/villages

Postal Code	DA	Pop Weight	Cumulative Population Weight	# People Assigned (n =3)	# People Assigned (n =100)	# of People Assigned (n=1000)
5	A	15%	0.15	2	13	163
5	B	65%	0.80	1	60	639
5	C	20%	1		27	198

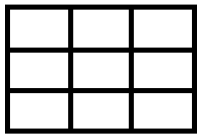
So now we have logic to do this, but how do we implement this?

The real solution – PCCF+, a Statistics Canada many:many implementation

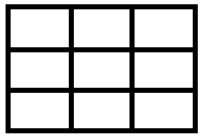
The R implementation - Contents



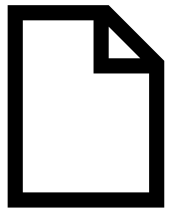
PCCF+ population weight



Base PCCF



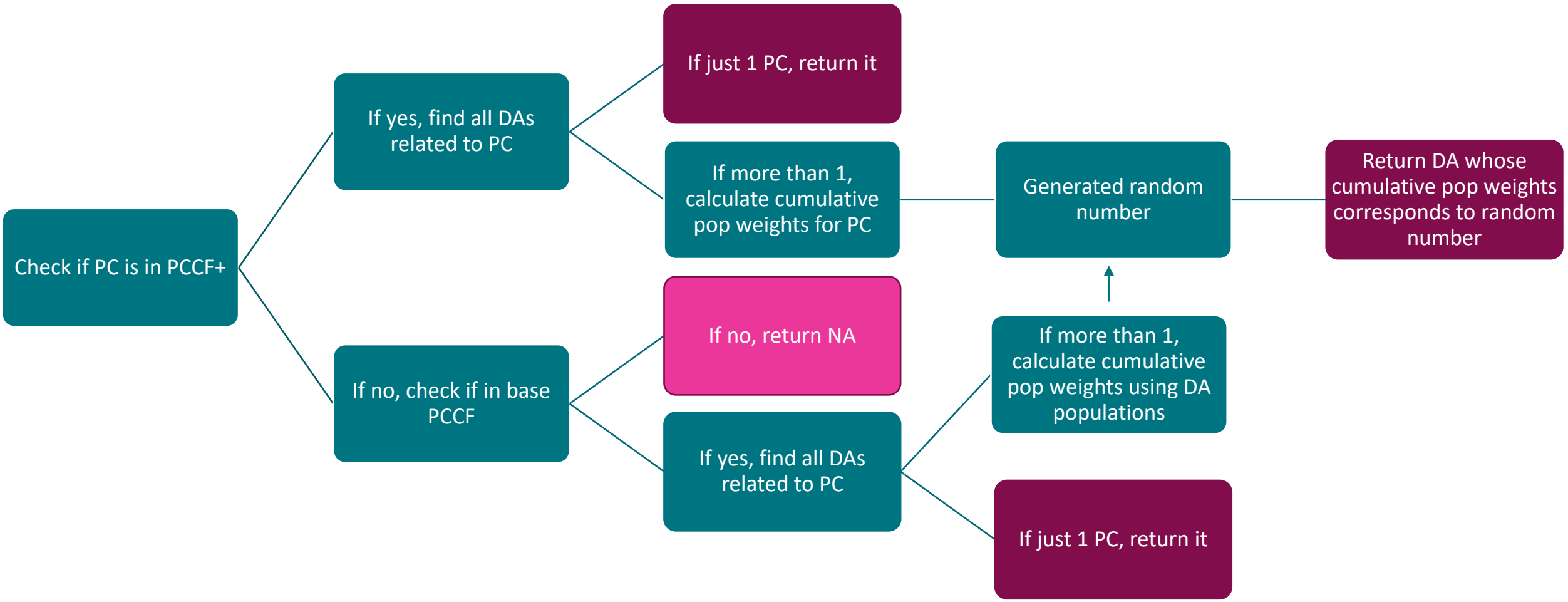
DA populations



R syntax file:

- Data file read
- Function
- Sample code

The R implementation - Function



R examples

```
13  #Remember in the version to be shared pull PCCF from basicPCCF2020
14
15  #Import libraries
16  library(readr)
17  library(dplyr)
18
19  #In this version import the datasets from csvs: filenames are basicPCCF, pccfPlus, pop
20  #Specify the folder where the files located:
21  folder <- "J:/Projects/Population Health Assessment/Census_Geo_DI_Data/PCCF+ in R March 2022/"
22  pccf  <- read_csv(paste0(folder,"basicPCCF.csv"), col_types = cols(DAUID = col_character()))
23  pop <- read_csv(paste0(folder,"pop.csv"), col_types = cols(DAUID = col_character()))
24  pccfPlus <- read_csv(paste0(folder,"pccfPlus.csv"), col_types = cols(DAUID = col_character()))
25
26
```

R examples

```
101
102 #Example code to Running the algorithm en masse:
103 #Create a random dataset sampling from both pccf files (for just one small county to better show probabilities)
104 testData <- data.frame(ID=1:10000,PC=c(sample(pccf$FSALDU[substr(pccf$DAUID,0,4)=="3511"],10000, replace=TRUE)))
105
106 #Run the data (use Sys time to track how long it take - with 40,000 takes 4 mins)
107 #Use sapply to run data on records simulatenously and return as a vector
108 Sys.time()
109 testData$DAUID <- sapply(testData$PC,runPccfPlus)
110 Sys.time()
111
```

```
> runPccfPlus("K7K1P8")
[1] "35100175"
> runPccfPlus("K7K1P8")
[1] "35100175"
> runPccfPlus("K7K1P8")
[1] "35100175"
> runPccfPlus("K7K1P8")
[1] "35100175"
>
```

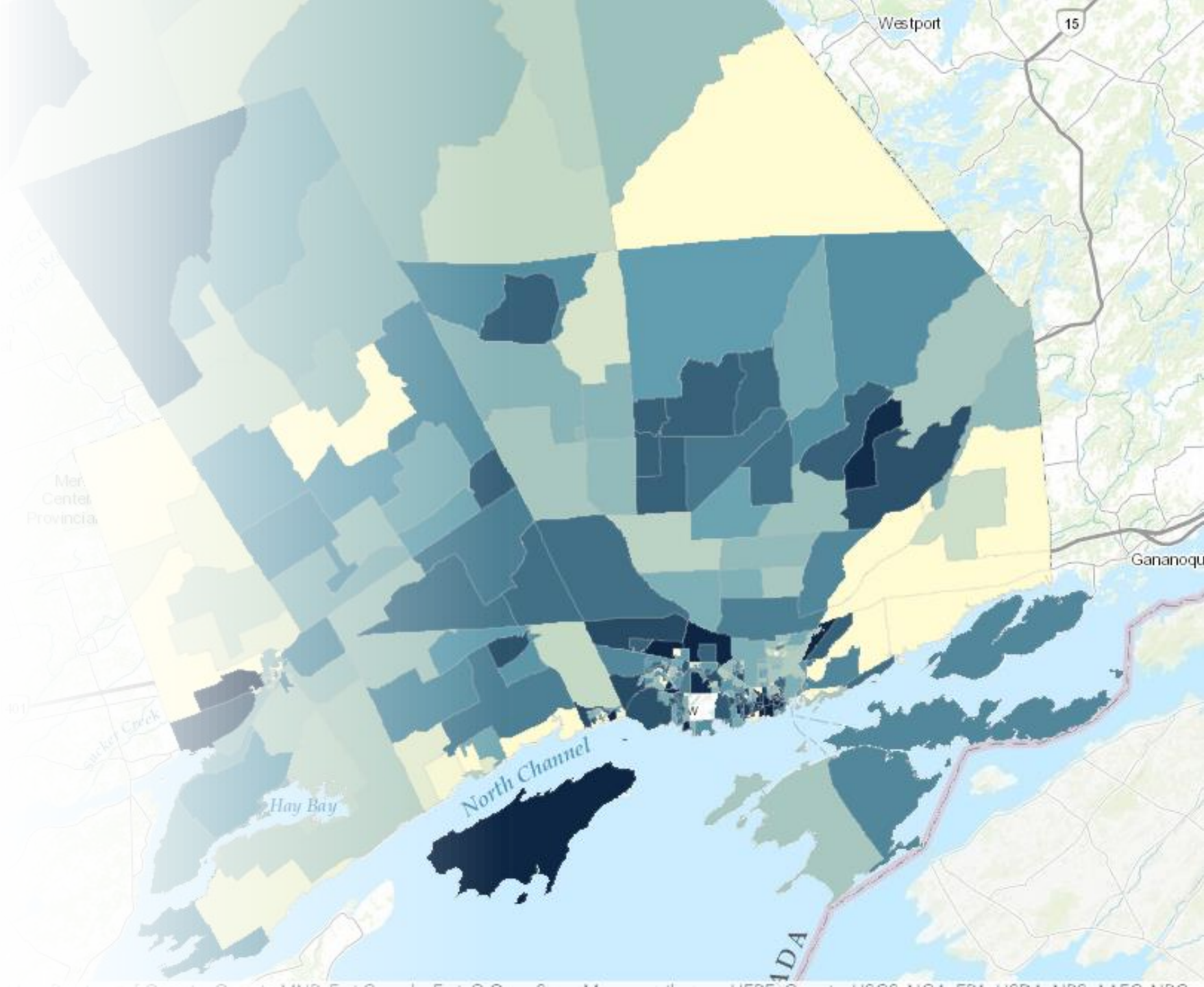
```
> runPccfPlus("K0H2H0")
[1] "35110080"
> runPccfPlus("K0H2H0")
[1] "35110079"
> runPccfPlus("K0H2H0")
[1] "35110078"
> runPccfPlus("K0H2H0")
[1] "35110081"
> runPccfPlus("K0H2H0")
[1] "35110078"
>
```

The SQL Implementation and Examples

```
SELECT [ID]
      ,[Postal_Code]
      ,[Reason_Imm]
      ,[Age]
      ,[Gender]
      ,[Census_SD]
      ,[PHU]
      ,[Max_Dose]
      ,[Last_Administer]
      ,[Last_6_Months]
      ,[Bivalent]
      ,[data_science].[dbo].udfpccfPlusAdjustedDAUID(Postal_Code) AS DAUID
      into [data_science].[dbo].[HE_Vax_DA2]
  FROM [covid].[dbo].[Vaccinated_KFLA_By_Last_Dose]
```

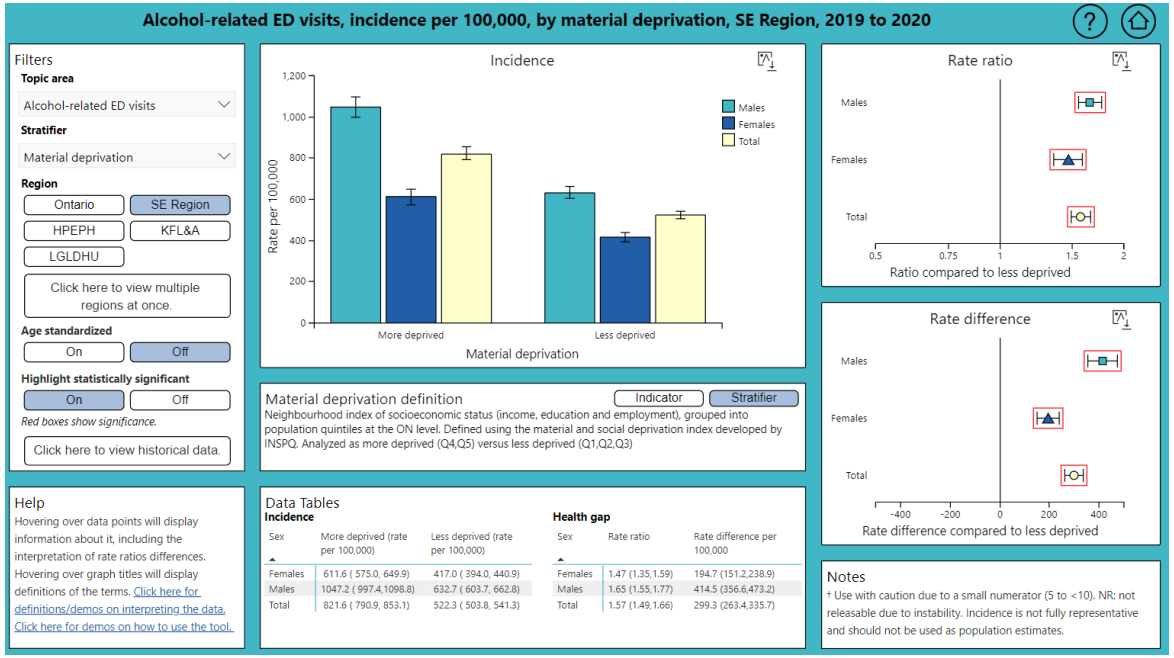
Use cases

- COVID-19 vaccine coverage mapping and program targeting
- Vision screening mapping
- Mapping DoPHS diseases in monthly dashboard
- ACES* PHU allocation



Use cases – health equity analyses

Developing neighbourhood level stratifiers for material and social deprivation



What the KFLAPH solution is not

Institutional postal codes

Non-residential postal codes

Error identification

Urban postal codes with multiple DAs treated slightly differently

So now what?

ACCESS TO KFLAPH SOLUTION

OTHER SOLUTIONS

QUESTIONS?