

Bridging Complex Data and Generative AI for Data Retrieval

Methods, Challenges, and Opportunities



IASSIST 2025

Vassilis Routsis
Senior Research Fellow
Department of Information Studies

v.routsis@ucl.ac.uk



Economic
and Social
Research Council



CORDIAL AI

CORDIAL-AI is a next generation system to help researchers retrieve subsets or whole data from complex census datasets



UCL Department of Information Studies

UCL Advanced Research Computing

UK Data Service

Project partners



- Project is led by the UCL Department of Information Studies
 - Dr Vassilis Routsis, Prof Antonis Bikakis, Dr Yujian Gan, Dr Sagar Uprety
- In collaboration with the UCL Advanced Research Computing
- Directly affiliated with the UK Data Service
- Funded by the ESRC Future Data Services scheme
- Project still under development



UK Census Flow Data

Flow data enumerate movements of people from one place to another. They include migration, journeys to schools or work, and travel to second residences.

These data are characterised by their complexity, volume, and sparsity.

Data are available from censuses from 1981 onwards. Data from 2011 and 2021 have different levels of access.

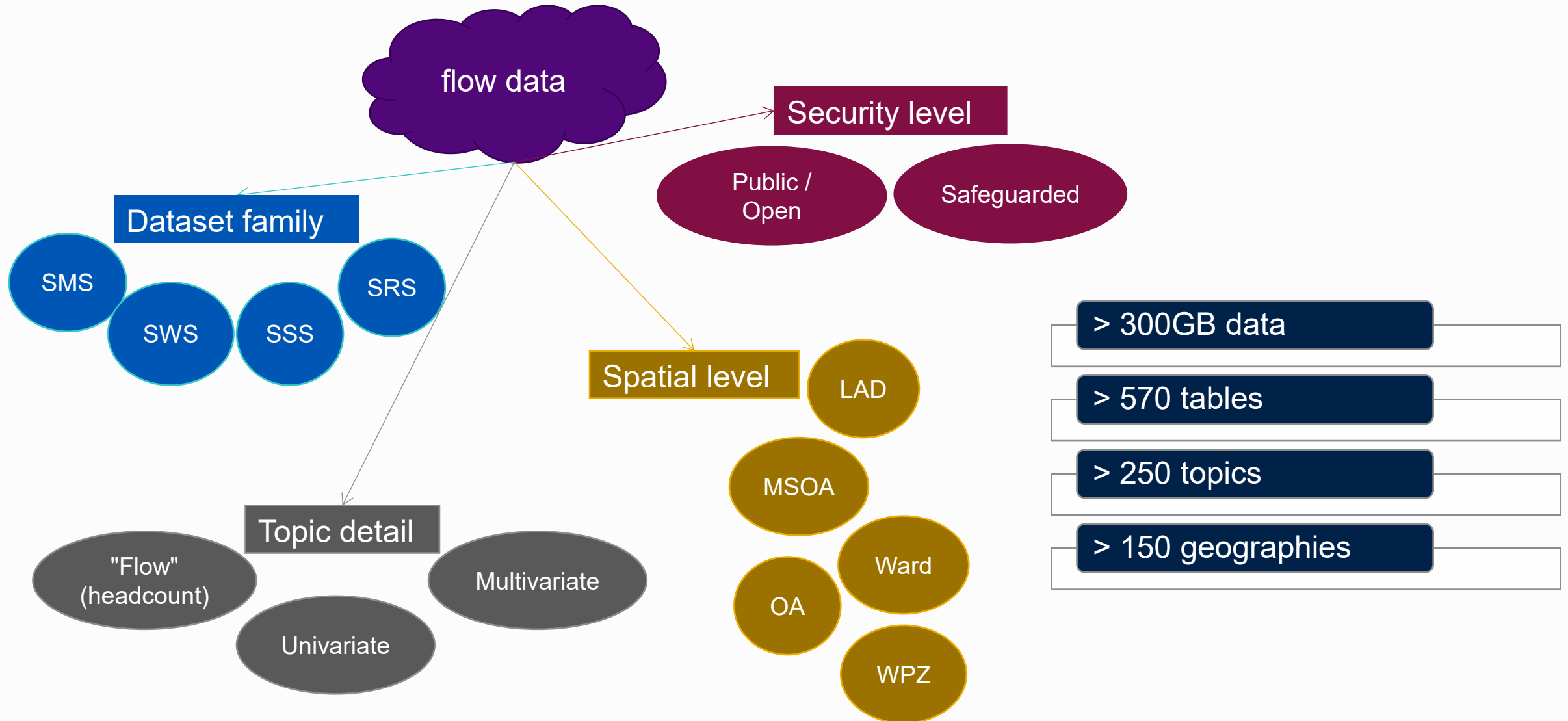


Barriers to using census flow data

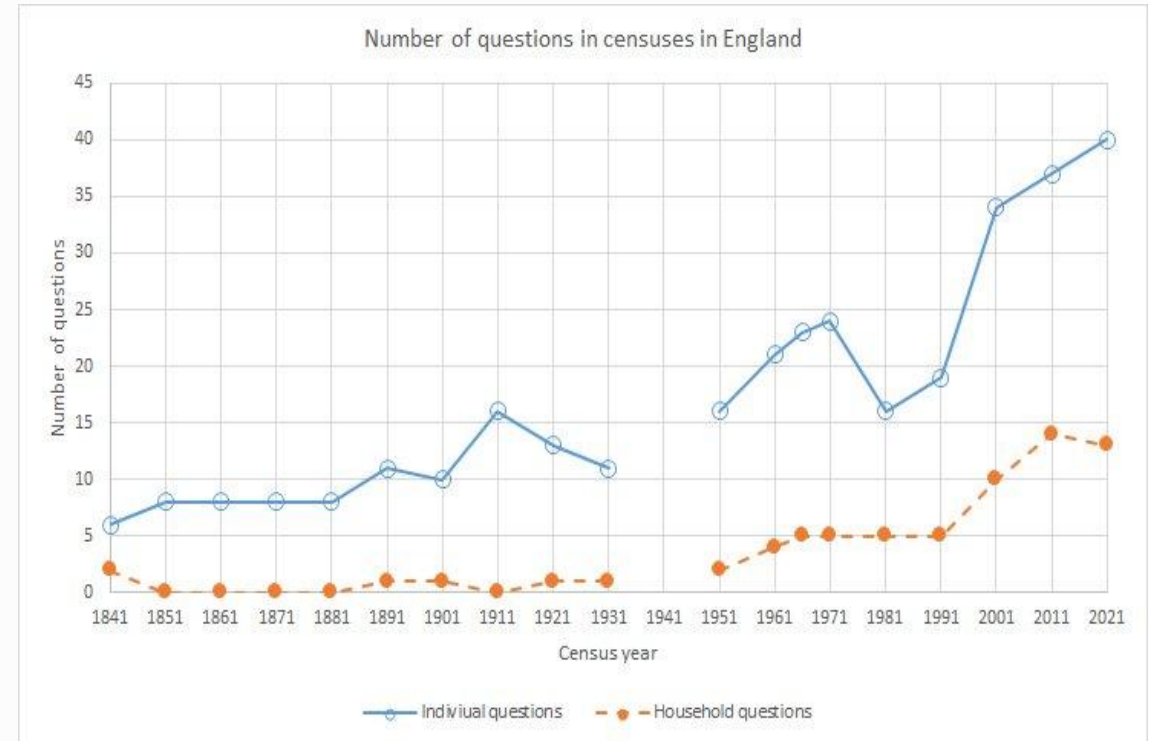
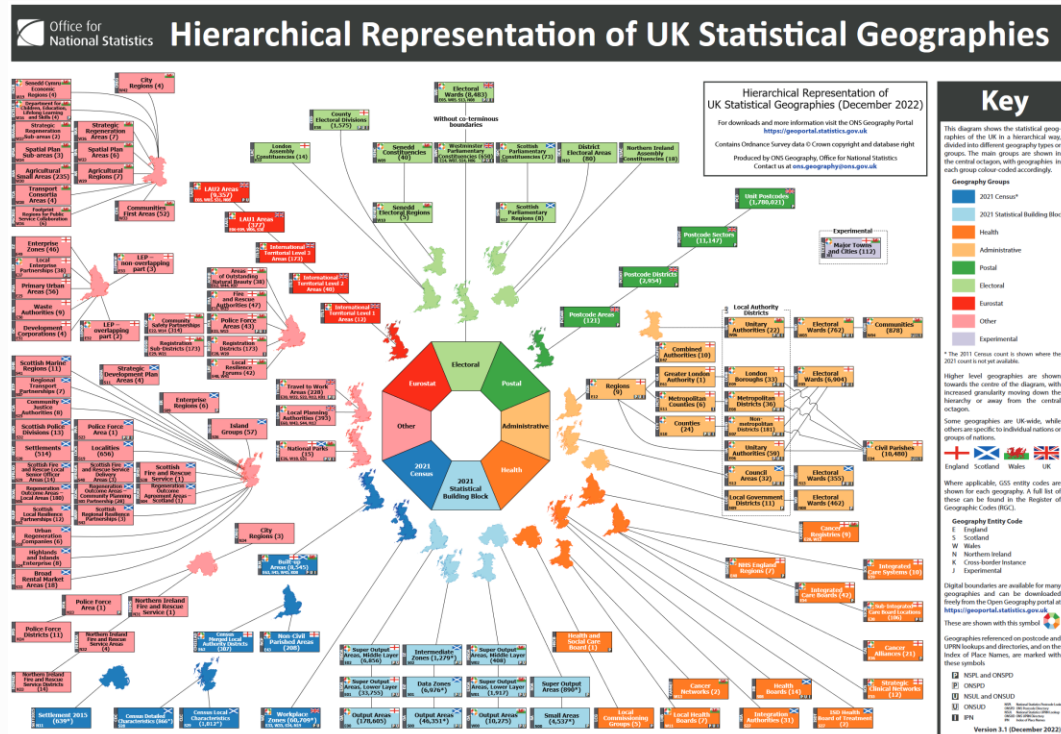
A vital part of the UKDS mission is to provide user support and training to researchers. This engagement allows us to gather valuable feedback and have direct conversations with users, providing unique insights into the barriers they face when using census flow data. Some of the barriers are:

- **Lack of visibility**
 - Researchers are unaware of the existence or depth of census flow datasets
- **Lack of domain knowledge or technical expertise**
 - The complexity of the data can be overwhelming for some users, making it difficult for them to locate relevant information or correctly extract the data they need
 - These challenges can also impact the data analysis and subsequent interpretation of these datasets

Census flow data: Layers of complexity



Challenges: Geographies and topics



<https://geoportal.statistics.gov.uk/documents/ons::hierarchical-representation-of-uk-geographies-december-2022/explore>

Source: https://x.com/oliver_dw/status/1164521314360733701?s=20

Current legacy system - WICID



wicid WICID → Query → Data → Select by table (Select table) Help
Logged in as: guest Customise Logout now

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

0 data items currently selected [Unselect all] [Edit list]

Table 1 [All migrants: age \(5 broad age groups\) by sex](#)
Table 2 [Wholly Moving Households and residents in Wholly Moving Households: counts](#)
Table 3 [All migrants: age \(5 year groups\) by sex](#)
Table 4 [All migrants: marital status by sex](#)
Table 5 [All migrants: ethnic group](#)
Table 6 [All migrants: whether resident in households by whether suffering from limiting long term illness](#)
Table 7 [All migrants aged 16+: economic position](#)
Table 8 [Wholly Moving Households: tenure](#)
Table 8S [Wholly Moving Households: tenure](#)
Table 9 [Wholly Moving Households: sex and economic position of head](#)
Table 10 [Residents in Wholly Moving Households: sex and economic position of head](#)
Table 11S [All migrants: Gaelic speakers](#)
Table 11W [All migrants: Welsh speakers](#)

What now?

You must select some data before you can produce any output!

- Select one of the above tables
- OR go back to the [dataset selection](#) list to choose a different data set
- OR go back to the general [data selection](#) page
- OR go back to the general [query interface](#)

Online since 2000

Received significant upgrades over the years

Still the only web tool to provide bespoke data of four decades data

UK Data Service
Census Support

Flow Data

Flow data links
Flow data home
Information and Publications
Available Data
Downloads
Contact us
News
FAQs
Credits

Support
Accessibility statement
Feedback
Help

Census Support: Flow Data

Flow Data is part of UK Data Service Census Support, providing data and user support for anyone interested in using flow data from the [Census](#).

Census flow data (also referred to as *interaction data*) relate to the movement of people between places. These can be regular short distance moves such as the daily commute to work, or less frequent - and sometimes longer distance - moves such as migration to a new home or an extended stay in a second address.

Flow data differ from other Census data in that they link two locations - an origin and a destination. With two geographies, flow datasets can be much larger than other Census data products.

Getting the data

There are three principal ways of getting to the data. In order to use most of the data sets, it is necessary to be registered.

- WICID - the flexible query builder
- WICID Downloads - flow data tables ready to be downloaded
- WICID Flow Summaries - a quick summary of the most significant flows associated with a chosen area

New data for un-registered users

Un-registered users and interested registered users can now access a time series of UK inter-district migration estimates, which have been produced by CIDER. These data are available directly from the Office for National Statistics - for more information about the dataset and how to gain access, click [here](#).

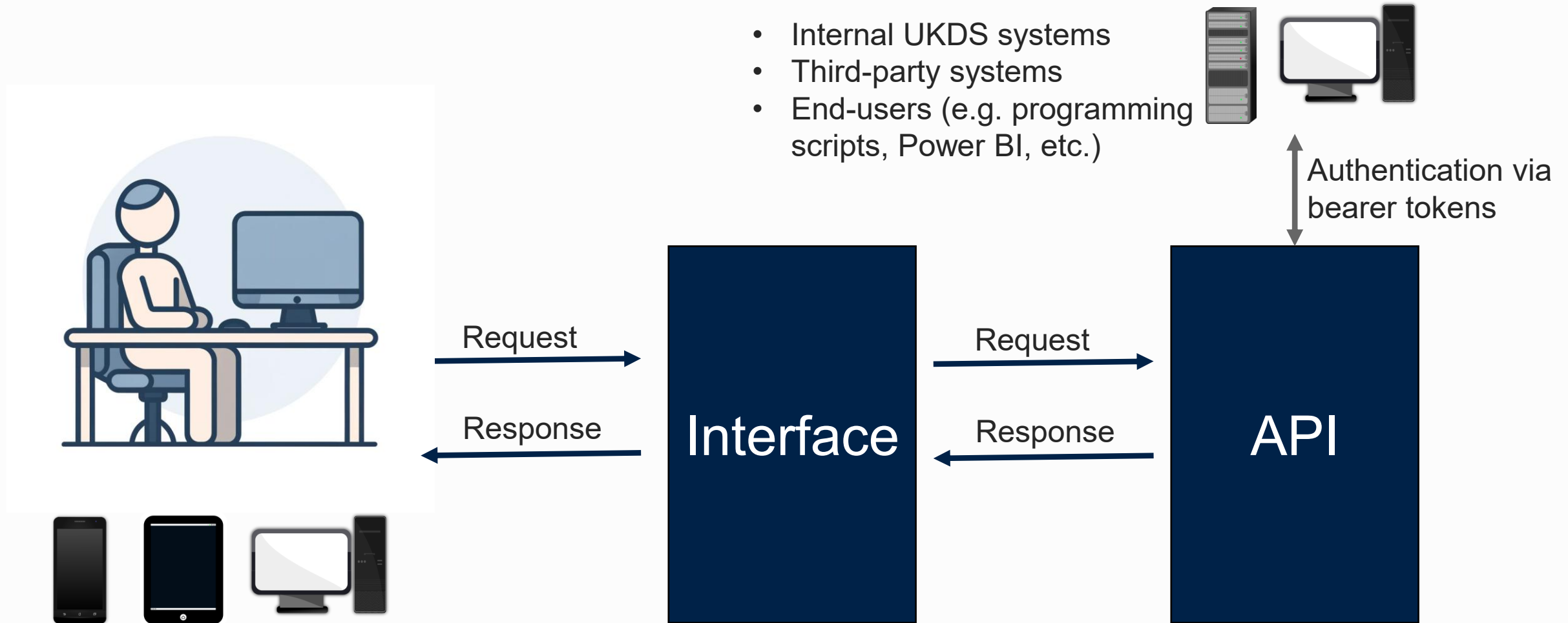
About the data

Currently migration and commuting data from the 1981, 1991, 2001 and 2011 Censuses are available through WICID. Additional non-census datasets are also available from the system.

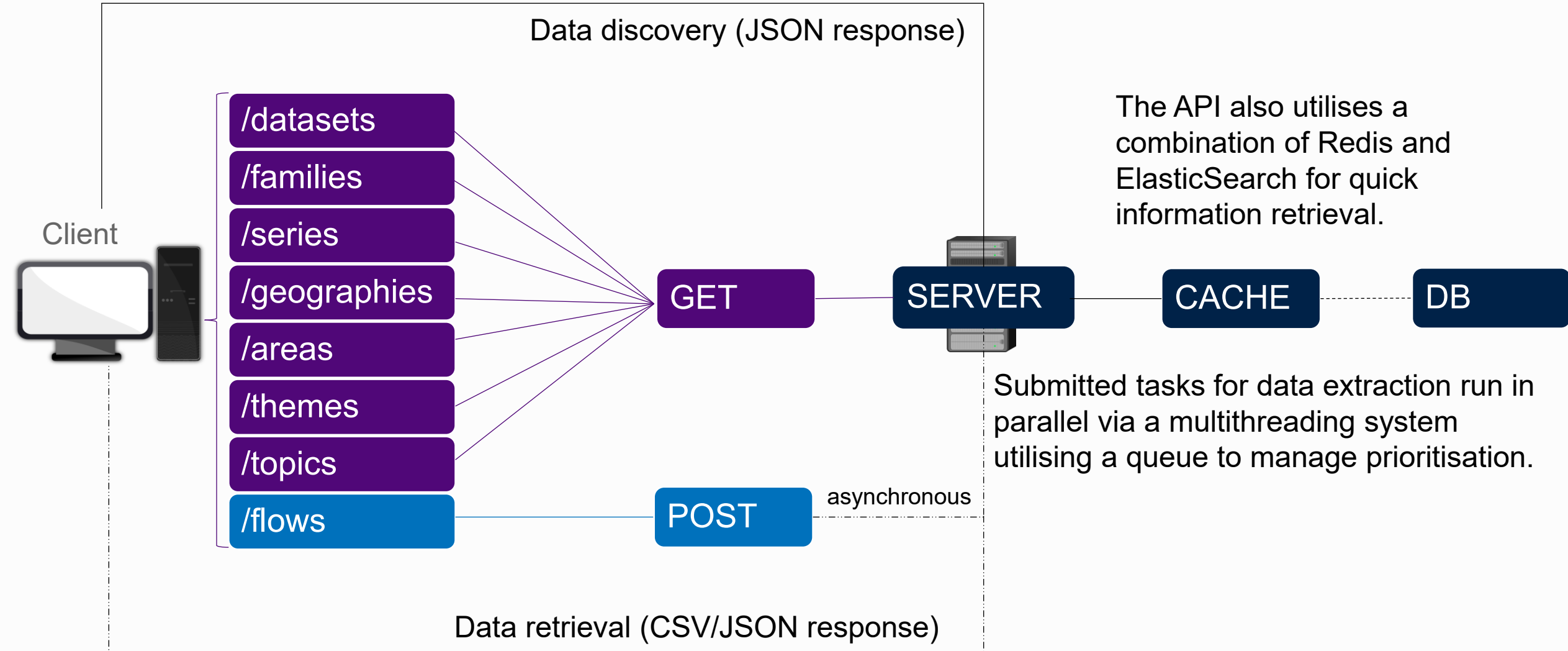
[Detailed information](#) about our data sets
[Tutorial information](#) is available describing how to use the interface

<https://wicid.ukdataservice.ac.uk/>

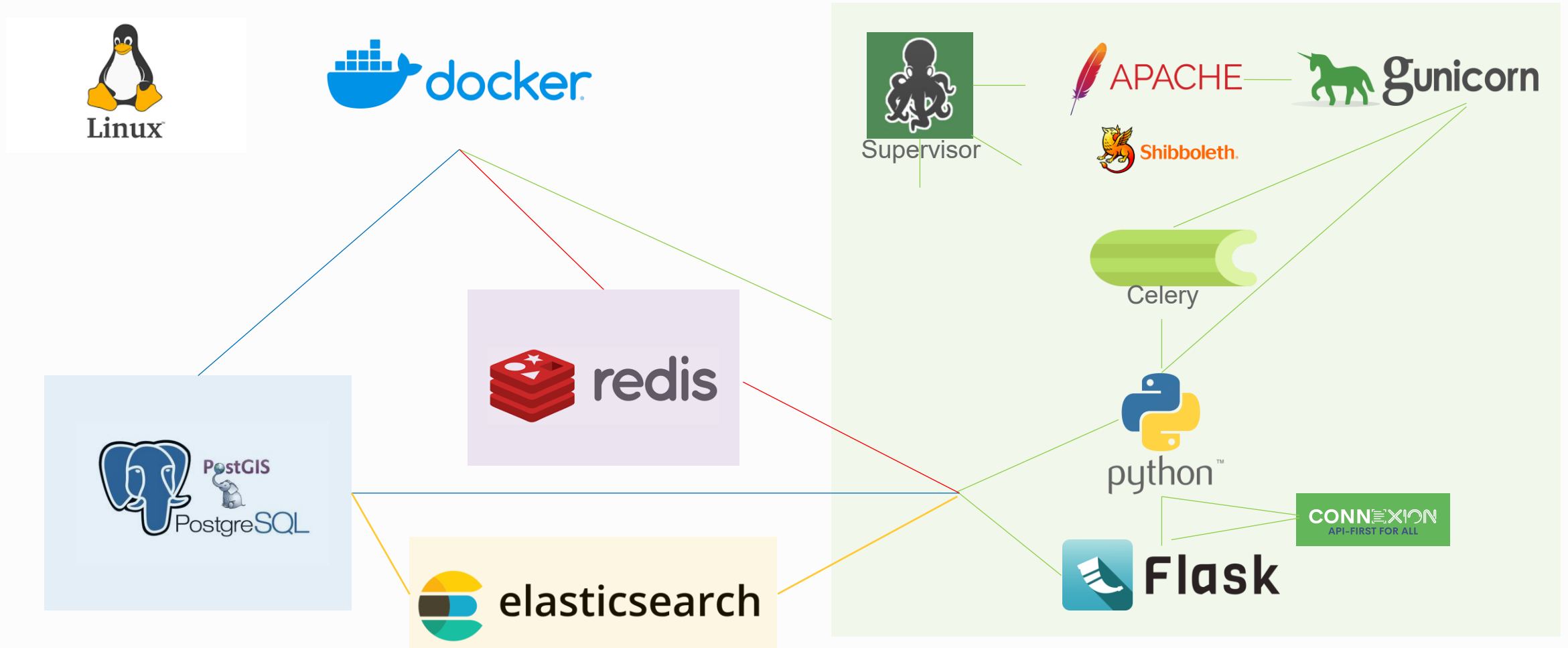
New API-driven platform - Access



Flow data API endpoints



Open-source technology stack - API



New UI (under development)



UK Data Service

MYWORKSPACE

Census Flow Data

WORKSPACE

PUBLICATIONS

HELP

ACCESSIBILITY

CONTACT

Navigation

Home

Services

Workspace

Layout

Data Types

Datasets

Browse available datasets

Families

Browse data families

Series

Browse data series

Geographies

Browse geographies

Themes

Browse themes

Selected Items

1 items selected

Account

Settings

API Keys

Logout

DATASETS

Search

	Dataset ID	Table ID	Dataset Label
<input type="checkbox"/>	330	ODMG01EW	2021 SMS MSOA [Origin and destination of migrants] - O
<input type="checkbox"/>	331	ODMG01EW	2021 SMS LTLA [Origin and destination of migrants] - OD
<input type="checkbox"/>	332	ODMG01EW	2021 SMS UTLA [Origin and destination of migrants] - OT
<input checked="" type="checkbox"/>	333	ODMG01EW	2021 SMS RGN [Origin and destination of migrants] - OD
<input type="checkbox"/>	334	ODMG01EWi	2021 SMS LTLA [Origin and destination of international r

Records per page: 10 1-10 of 66

THEMES

Search

	Theme Alias	Theme Label
<input type="checkbox"/>	age	Age
<input type="checkbox"/>	car	Car ownership
<input type="checkbox"/>	ethnic	Ethnicity
<input type="checkbox"/>	health	Health
<input type="checkbox"/>	hourw	Hours of work

Records per page: 10 1-10 of 12

Census Flow Data

UK Data Service

Discover and download UK Census origin and destination data

Find origin and destination statistics about the UK population including data on housing, health, migration and the labour market. We provide access to data for England, Northern Ireland, Scotland and Wales from the 1971, 1981, 1991, 2001, 2011 and 2021/2022 UK censuses.

Browse our catalogue by:

FAMILIES

DATASETS

GEOGRAPHIES

SERIES

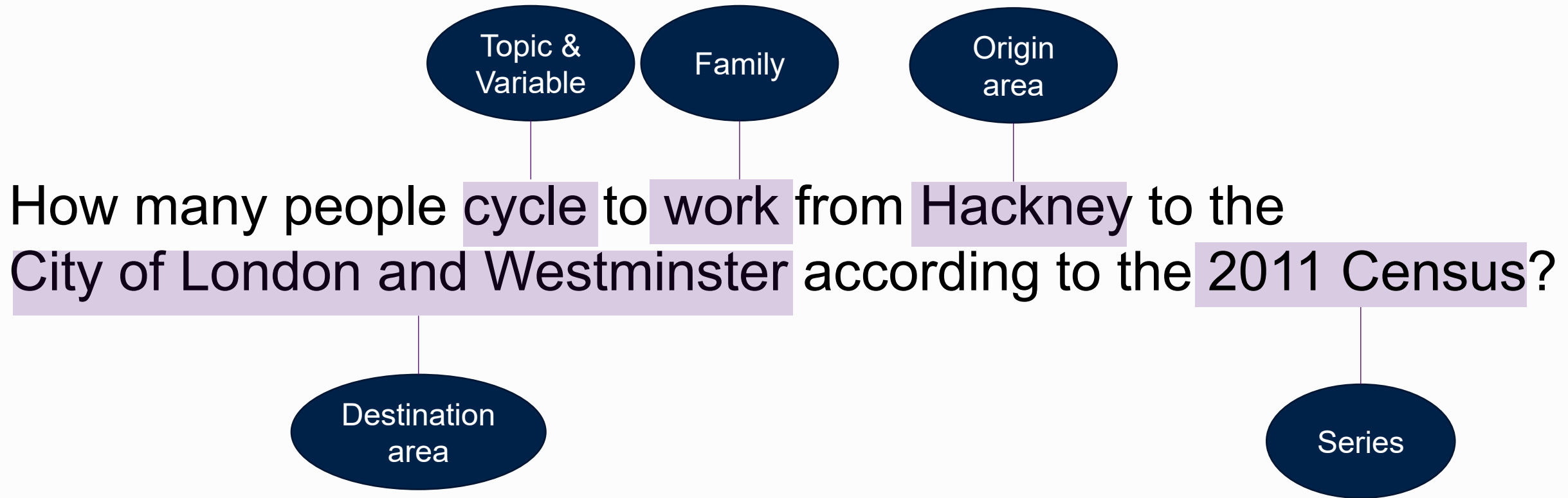
THEMES AND TOPICS

Welcome to the new interface for UK Census Flow Data! This is a beta version that will be improved over time and eventually replace our old platform. Please let us know what you think by reaching out to us. Use the contact button at the top menu.

Use our Workspace environment to discover and construct flow data

WORKSPACE

Example natural language query

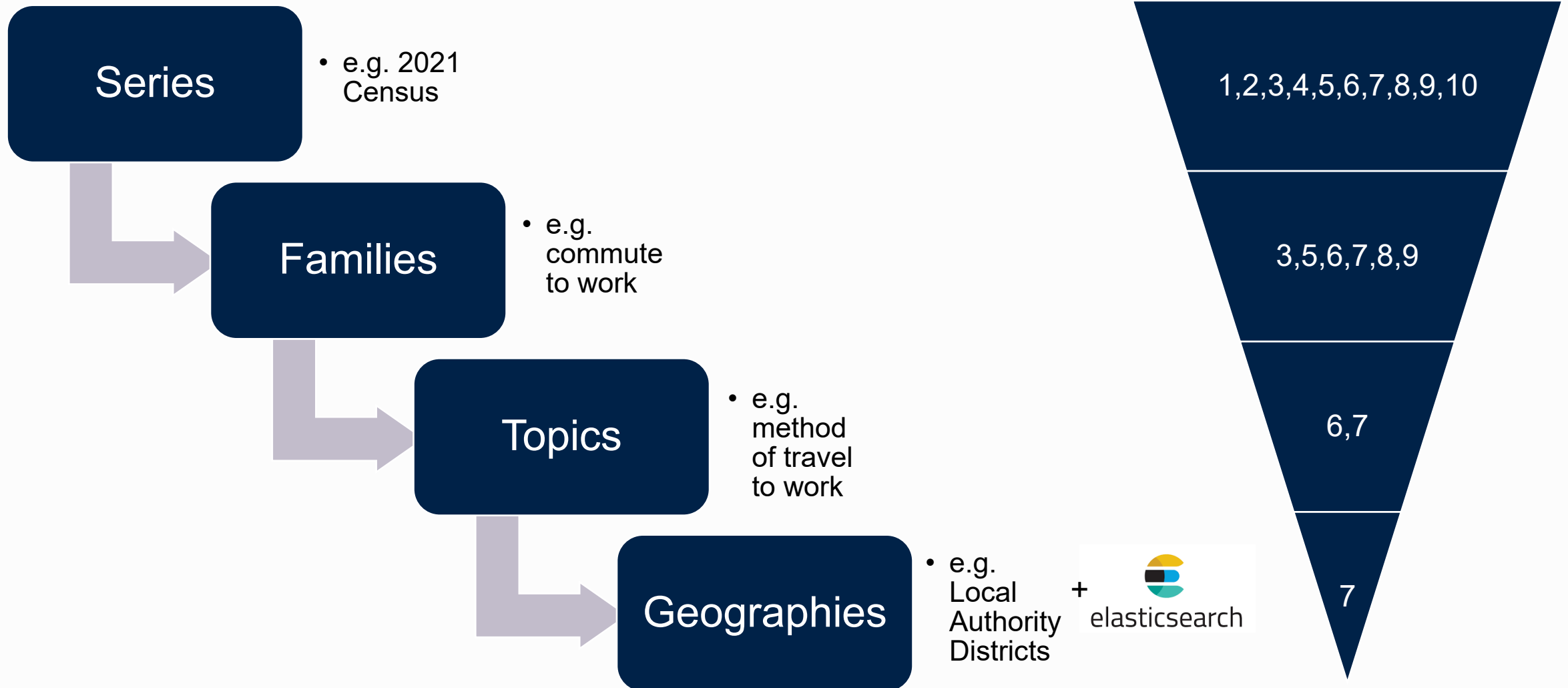


...but there's still some ambiguity left!

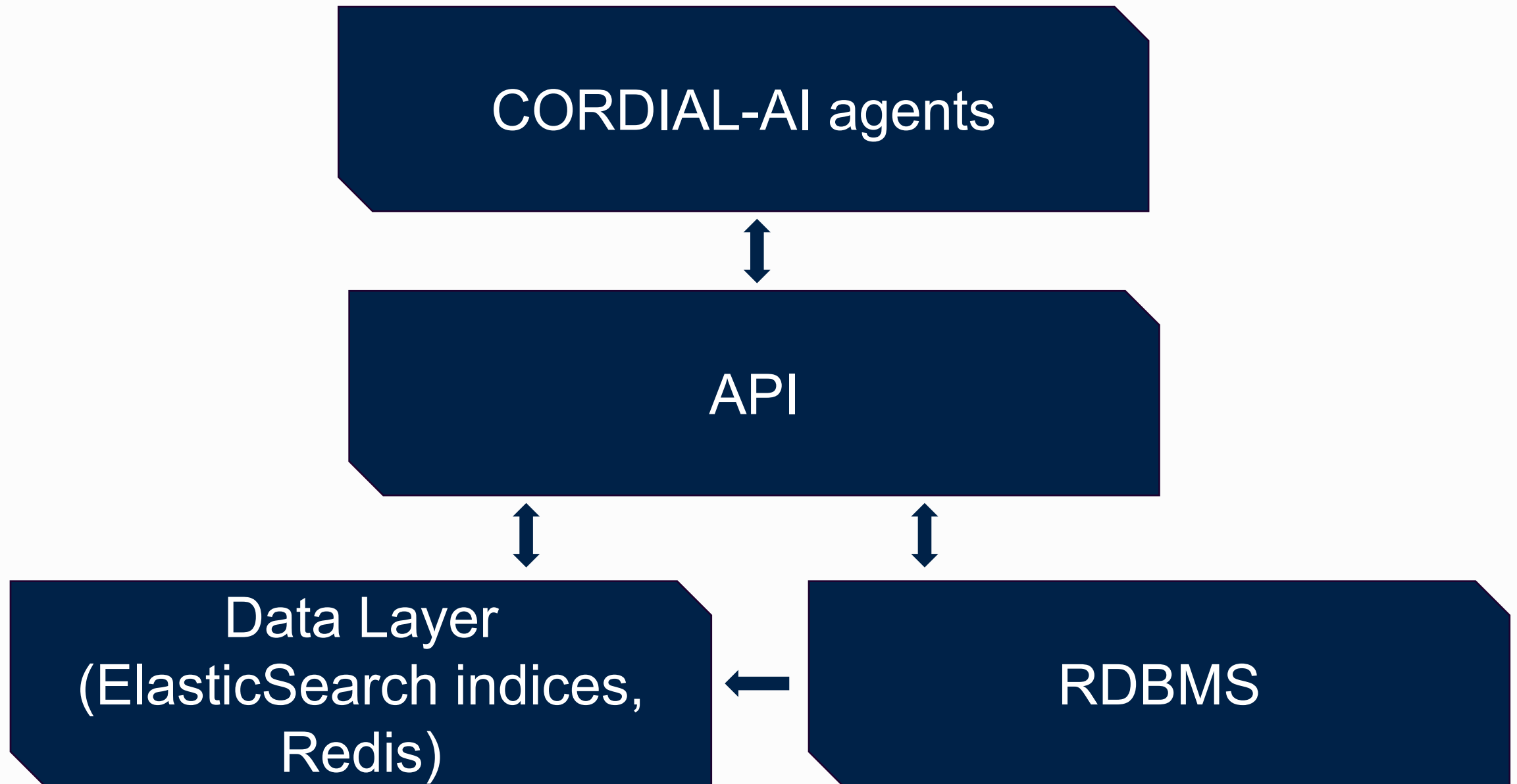
We can guess, but we are not entirely sure on the geography type the user wants the flow in.

Multi-agent system

Multiple agents sequentially filter potential census tables that might be appropriate for the user query. This is done by calculating the intersection of the table ids for each step. Sequential filtering may also allow to provide more insightful feedback to the user.



Information flow



A large, glowing yellow question mark is centered on the left side of the slide, set against a dark, textured background.

Handling ambiguities

1. Make assumptions and inform users on the decision

- Allow the LLM to take initiatives based on the most probable scenario

2. Ask users to clarify on the missing information

- Identify missing information and ask users to clarify. Wherever possible inform users on the available supported options

Sample API request to retrieve data

```
1 {
2   "datasets": [
3     {
4       "dataset_id": 293,
5       "table_id": "WU03UK",
6       "variables": [
7         10
8       ]
9     }
10  ],
11  "geographies": {
12    "origin": [
13      {
14        "geography_id": 96,
15        "areas_type": "sequence_number",
16        "areas": [
17          302
18        ],
19        "labels": [
20          "supplied_code",
21          "label"
22        ]
23      }
24    ],
25    "destination": [
26      {
27        "geography_id": 96,
28        "areas_type": "sequence_number",
29        "areas": [
30          324
31        ],
32        "labels": [
33          "supplied_code",
34          "label"
35        ]
36      }
37    ]
38  },
39  "explaners": [
40    "geography_label"
41  ],
42  "output": {
43    "format": "csv",
44    "layout": "pairlist",
45    "mode": "normal"
46  }
47 }
```

table

variable

geography type

area

geography type

area

2011 Census Special Workplace Statistics Merged LA/LA [Location of usual residence and place of work by method of travel to work] - WU03UK - Open

UK Local Authorities (merged) 2011

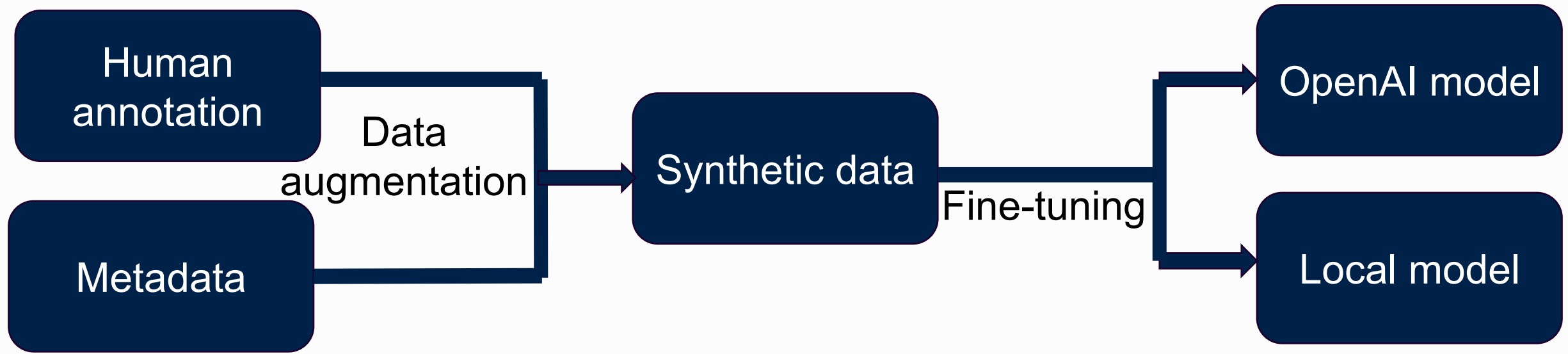
Hackney

UK Local Authorities (merged) 2011

City of London, Westminster

"All categories: Method of travel to work; Persons",
"Work mainly at or from home; Persons",
"Underground, metro, light rail, tram; Persons",
"Train; Persons",
"Bus, minibus or coach; Persons",
"Taxi; Persons",
"Motorcycle, scooter or moped; Persons",
"Driving a car or van; Persons",
"Passenger in a car or van; Persons",
"Bicycle; Persons",
"On foot; Persons",
"Other method of transport to work; Persons"

Training pipelines



Text to census flow data - UI



CORDIAL-AI

Welcome, Vassilis

Model

gpt-4o

API Formatting Mode

In API mode, you can view and modify the API request before execution.

New Chat

Archived Chats

Active

2025-05-20 14:49:52

How many people who moved from.

gpt-4o

Logout

Starting new conversation.

14:49:19

How many people who moved from Colchester to the region of London, rent their home based on the 2021 census?

14:49:34

This API retrieves data from the Census 2021 about people aged 1 and older in England and Wales who lived at a different address one year prior. It specifically focuses on the movement of these individuals between different areas, classified by their housing situation (either social rented, private rented, or living rent-free). The data provides insights into the migration patterns from Colchester to London, detailing how many people moved between these areas and their housing tenure.

14:49:52

Data retrieved for table **ODMG07EW**. Origin: **Colchester**. Destination: **London**.

Hide Preview

Download

</> API Details

Chain of Thought

ext_42936f07337d4527_349.csv

Showing 1 of 1 rows

"Origin Lower Tier Local A...	"Destination Region name"	"Origin Geography Label"	"Destination Geography L...	"Social rented; Persons [34...	"Private rented or lives ren...
Colchester	London	2021 E&W LTLAs	2021 E&W Regions	230	502

Type your message here... (Shift+Enter for new line, Enter to send)

Send

Explainability – Reference validation

How many males aged 16-34 from international origins moved to Camden according to the 2021 Census?

This API retrieves data from the Census 2021 about people aged 1 and older in England and Wales who lived at a different address one year prior. It specifically focuses on migrants moving to Camden, detailing their age and sex. The dataset includes information on males aged 16 to 24 and 25 to 34. It also identifies whether these individuals originated from outside the UK. The data helps understand migration patterns into Camden, categorized by age and gender.

Data retrieved for table ****ODMG10EW****. Origin: ****Person with migrant origin outside the UK****. Destination: ****Camden****.

Preview CSV

Download Data

Hide API Details

datasets: [

0: {

dataset_id: 355

table: ODMG10EW

table_id: ODMG10EW

variables: [

2021 SMS LTLA [Origin and destination of migrants by age and sex] - ODMG10EW - Open

Dataset ID: 355 Table ID: ODMG10EW

Description

Provides Census 2021 estimates on usual residents aged 1+ in E&W living at a different address 1 year before. Classifies residents in each LTLA by age, sex and previous LTLA. Residents outside UK are counted as "Outside UK".

GEOGRAPHIES

TOPICS & THEMES

VARIABLES

METADATA

Supported Geography Types

Legend

Origin - Intrinsic

Origin - Aggregated

Destination - Intrinsic

Destination - Aggregated

Origin Geographies

Intrinsic

2021 England and Wales Lower Tier Local Authorities (LTLA)
2014 Northern Ireland Local Government Districts (compatible with 2021 E&W data)
2019 Scottish Council Areas
2021 England and Wales SMS (Migration) special codes
2021 England and Wales Does Not Apply special code

Aggregated

2021 England and Wales Local Authority Districts (LAD)
2021 England and Wales Upper Tier Local Authorities (UTLA)
2021 England and Wales Regions
2021 England Local Enterprise Partnerships (LEP)
2021 England Inner and Outer London
2021 United Kingdom Countries
2021 England and Wales
2021 Rest of the UK regions/countries (Northern Ireland and Scotland) except England and Wales

Destination Geographies

Intrinsic

2021 England and Wales Lower Tier Local Authorities (LTLA)

Aggregated

2021 England and Wales Local Authority Districts (LAD)
2021 England and Wales Upper Tier Local Authorities (UTLA)
2021 England and Wales Regions
2021 England Local Enterprise Partnerships (LEP)
2021 England Inner and Outer London
2021 United Kingdom Countries
2021 England and Wales

Geography Details

2021 England and Wales Lower Tier Local Authorities (LTLA)
ID: 137

331 areas

2021 England and Wales Local Authority Districts (LAD)
ID: 138

318 areas

Explainability - Chain of Thought



How many people with managerial jobs moved to London as region at UTLA level based on the 2021 census? 12:34:05

This API retrieves data from the 2021 Census about people aged 16 and over in England and Wales who moved to a different address within the past year. It classifies these individuals based on their socio-economic status (NS-SEC) and the area they moved from, specifically focusing on those in higher and lower managerial and administrative occupations. The data covers all Upper Tier Local Authorities (UTLA) in England and Wales, and includes people who moved from outside the UK, categorized as 'Outside UK'. 12:34:32

- Preview CSV
- Download
- API Details
- Chain of Thought

Step 1: Filtered IDs after series and families selection

Based on the series and families criteria, these dataset IDs were filtered:

330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356



Step 2: Filtered IDs after geography selection

After applying geography filters, these dataset IDs remained:

345 346 347 348 330 331 332 334 335



Step 3: Final Dataset IDs

The final dataset IDs determined by the system:

345 346 347 348 330 331 332 334 335



Unselected Dataset IDs

These dataset IDs were available but not selected:

345 347 348



Step 4: Selected Dataset IDs

These dataset IDs were selected for the final API request:

346

Flexibility – API format mode

Give me the number of people with not good health who moved from Bristol to London. I want 2021 census data. 12:12:49 AM

API Request

```
{
  "datasets": [
    {
      "dataset_id": 353,
      "table_id": "ODMG09EW",
      "variables": [
        5,
        6
      ]
    }
  ],
  "explaners": [
    "geography_label"
  ],
  "geographies": {
    "destination": [
      {
        "areas": [
          "London"
        ],
        "areas_type": "label",
        "geography_id": 142,
        "labels": [
          "label"
        ]
      }
    ],
    "origin": [
      {
        "areas": [
          "Bristol, City of"
        ],
        "areas_type": "label",
        "geography_id": 138,
        "labels": [
          "label"
        ]
      }
    ]
  }
}
```

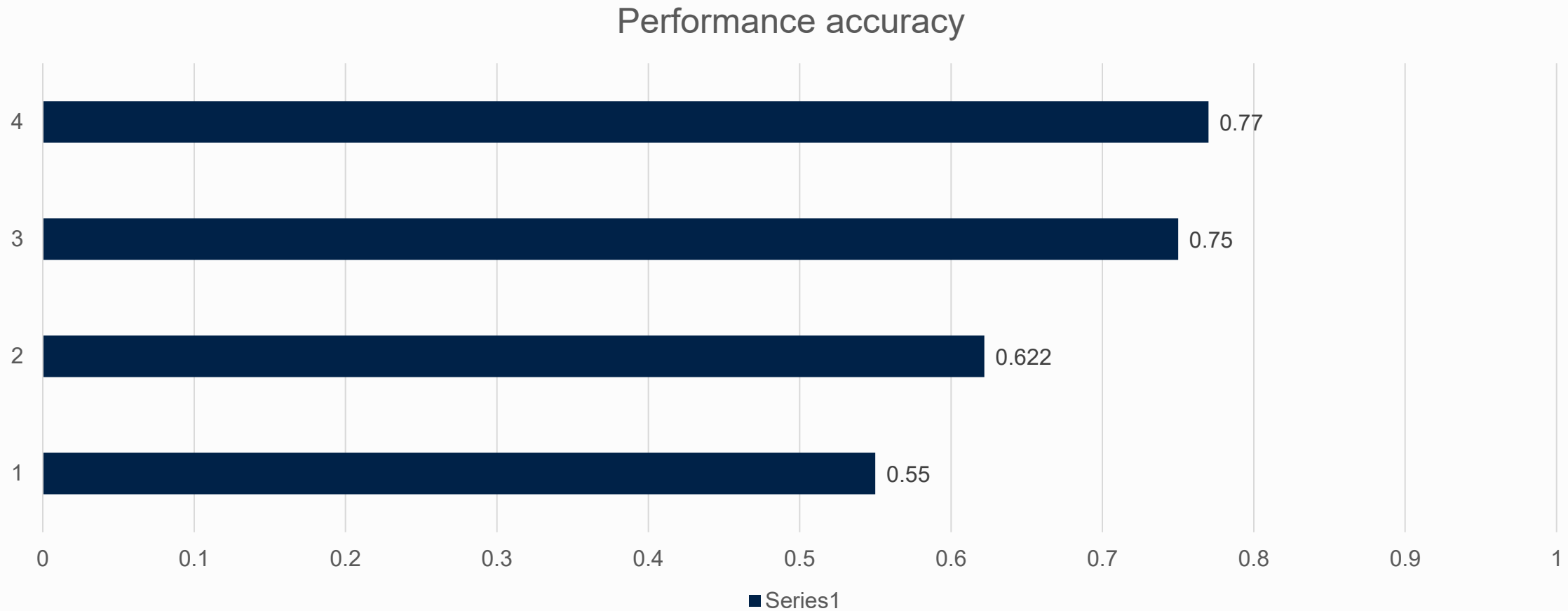
Submit API Request

Cancel

Type your message here... (Shift+Enter for new line, Enter to send)

Send

Preliminary benchmarks



Overall Score = $0.5 \times \text{Dataset Score} + 0.2 \times \text{Variables Score} + 0.3 \times \text{Geography Score}$

Integration of CORDIAL-AI to the new platform



MYWORKSPACE ▾

Cens...[WORKSPACE](#)[PUBLICATIONS](#)[HELP](#)[ACCESSIBILITY](#)[CONTACT](#)

Navigation

- Home
- Services
- Workspace

Account

- Settings
- API Keys
- Logout

Dashboard

[ACCOUNT PRIVILEGES](#)

[MANAGE ACCESS KEYS](#)

This is the main area from where you can start exploring census flow data.

You can explore or download full or subsetting data.

Census Flow Data Workspace →

Use the workspace to build your own data subsets and download them.

Explore Census-Related Data →

Use this to explore census-related data such as geographical areas, variable data, and more.

Use Generative AI to retrieve Flow Data →

Use our experimental generative AI tool to retrieve census flow data.

Use the API to retrieve Flow Data →

Discover and download census flow data using the API.

Thank you.

Vassilis Routsis, v.routsis@ucl.ac.uk