Integration of alternative data into consumer price statistics: the UK approach

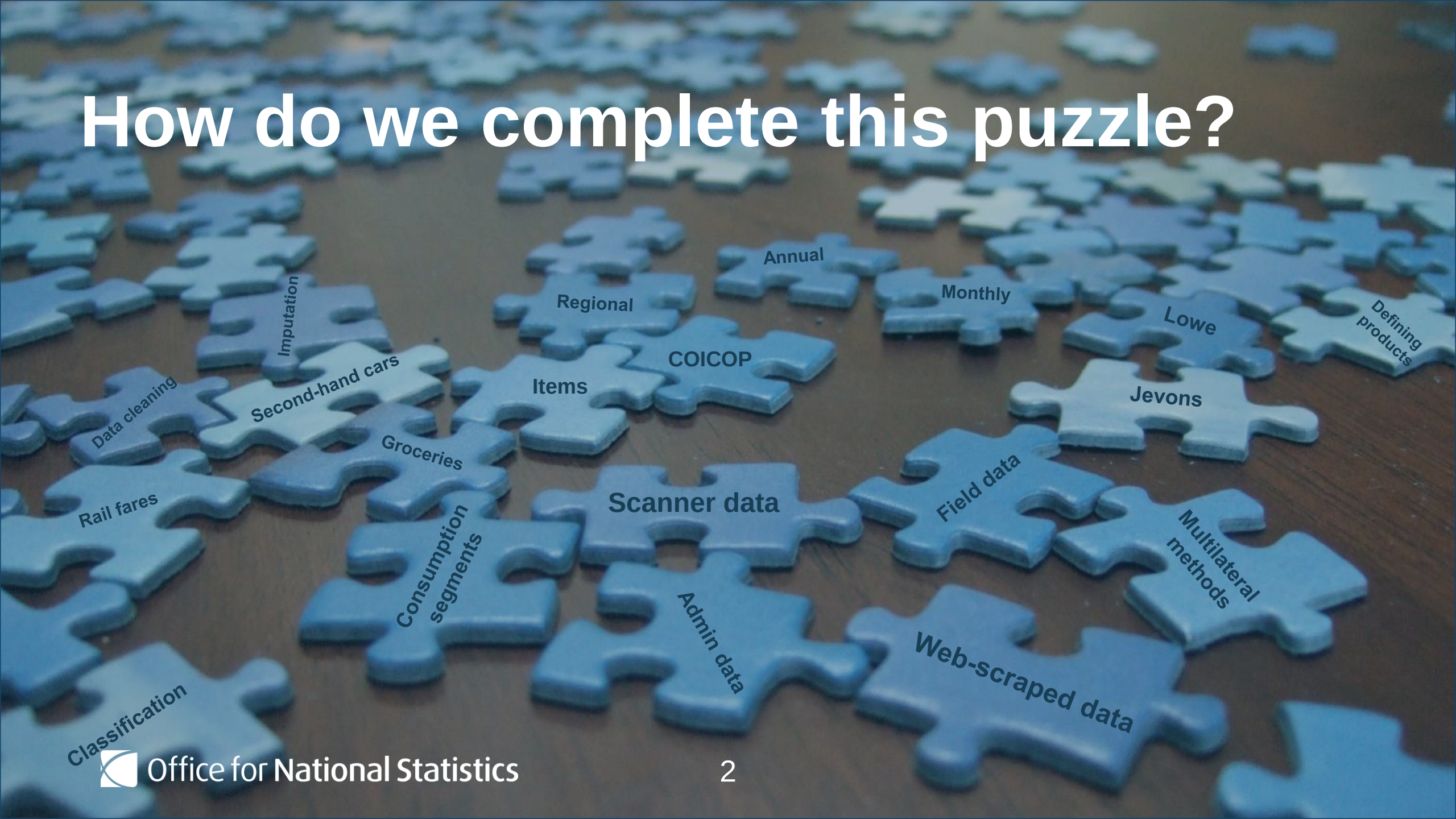
Helen Sands

Lead statistician

Scanner data implementation team

Office for National Statistics

How do we complete this puzzle?



New data

New UK data

Rail fares transaction data
Daily data feeds covering all transactions
One supplier for all of GB
~100% market coverage

Go live in 2023

Second-hand cars web data
Daily data feeds covering all advertisements
One supplier for all of UK
~70% market coverage?

Go live in 2023

Groceries scanner data
Weekly/monthly feeds, some data daily some weekly
3 (+) major supermarket chains
~50% market coverage

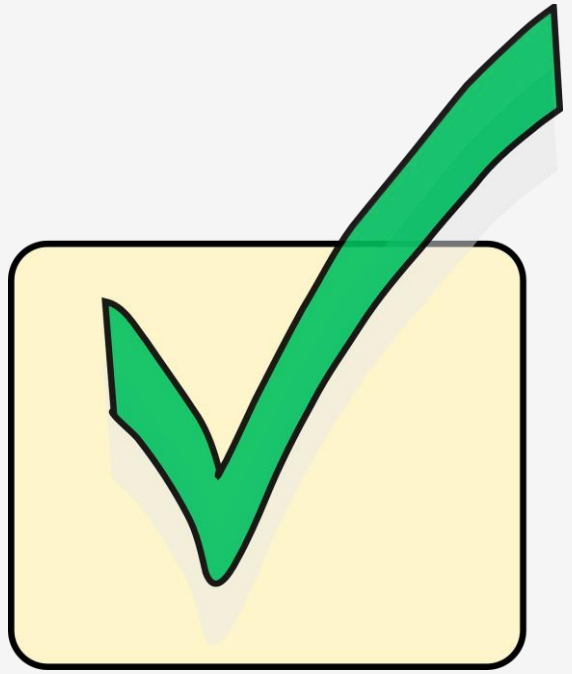
Go live in 2024

Quality assurance and classification

Quality assurance

Series of initial quality checks:

- Have the expected files arrived?
- Are variables correct data type?
- Are values within expected range?
- Are the data size/shape as expected?
- Are there a consistent (low) number of null values?



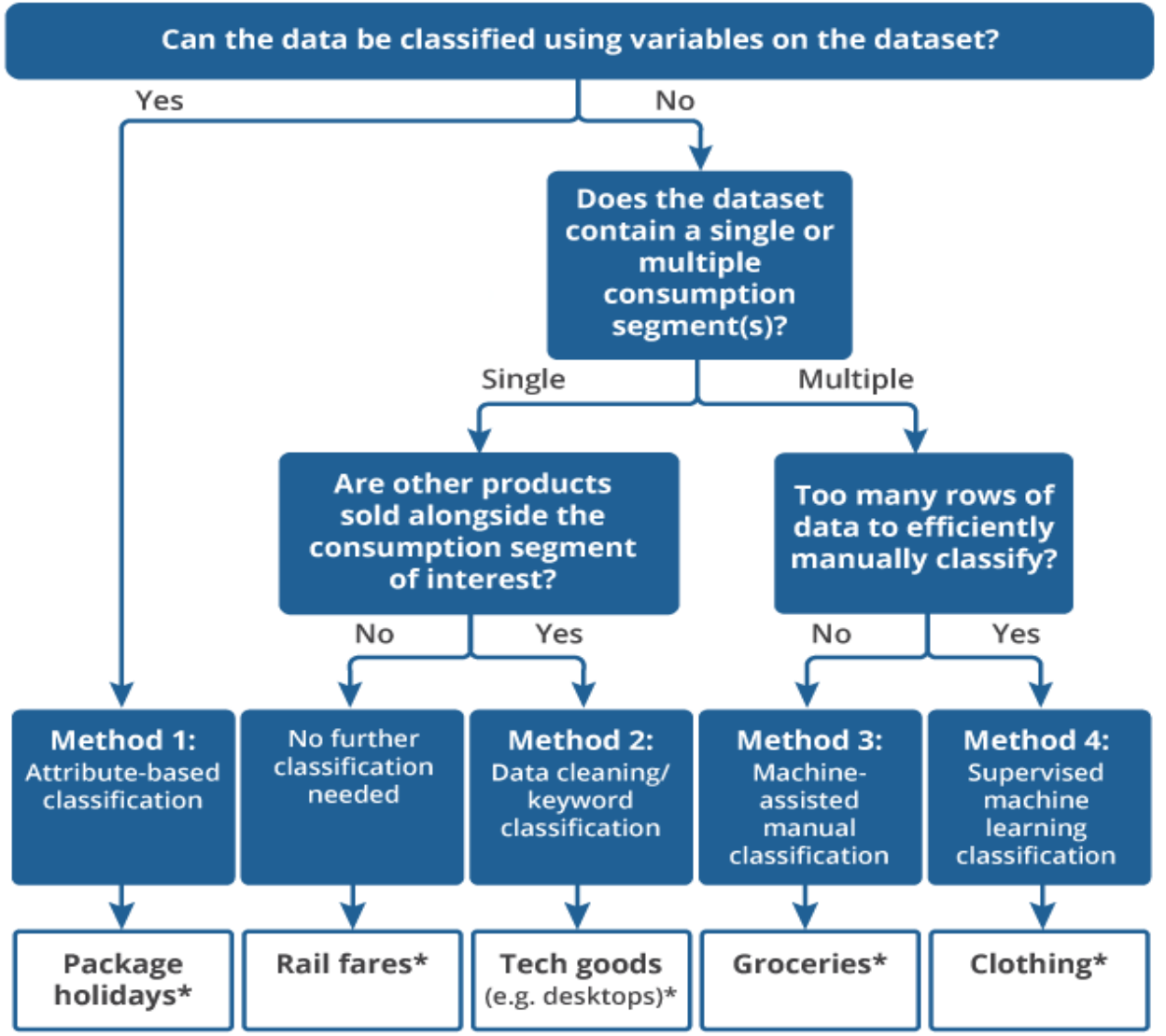
If any failures, issues raised with retailer and redelivery arranged if necessary

Further “curiosity” checks to investigate sensibility of data

Classification/filtering

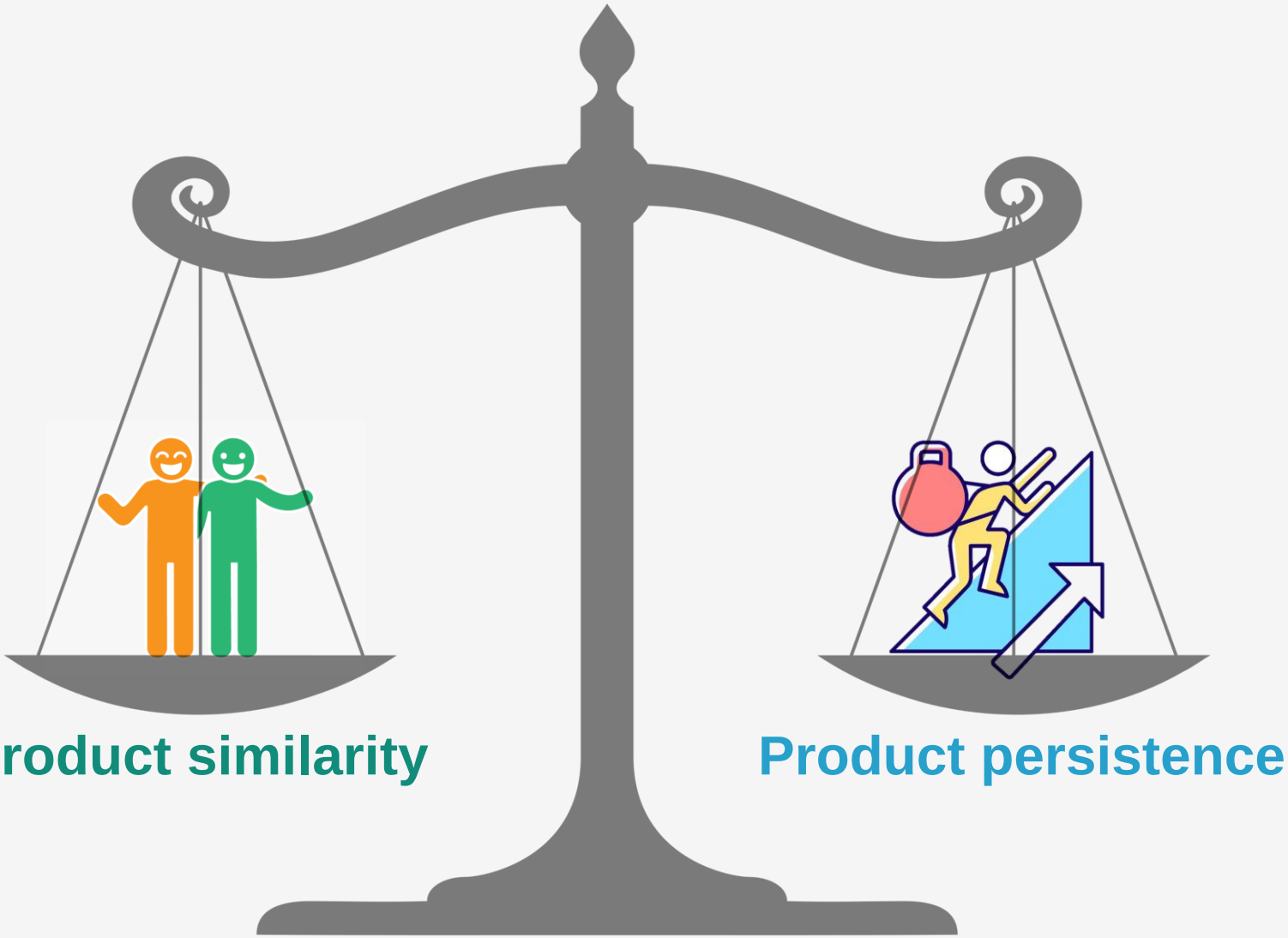
For **rail fares** & **second hand cars**, other products are sold alongside products of interest (e.g. station parking and second-hand motorbikes), therefore a simple data cleaning classification method is applied

Groceries cannot be classified using variables on the dataset and contain multiple consumption segments (COICOP6). For now manual classification has been achievable – in future we look to improve the efficiency of this using machine learning to assist the classification process



*Recommendations for these categories are subject to change due to ongoing research. These categories are given as examples and are not an exhaustive list of all categories being explored.

Defining a product


Product similarity
Product persistence

Groceries

(has unique identifiers)

GTIN

SKU

Relaunch linking

Relaunch linked SKU + unit of measurement

Rail fares

(no unique identifiers)

Origin/destination stations

+ Ticket type

+ Ticket class

+ Journey type

+ Region?

+ Traveller demographic?

Second-hand cars

(can't use unique identifier!)

Make

+Model

+Mark?

+Trim?

+Age?

+Mileage?

+Body type?

+Fuel type?

MARS (Chessa, 2019)

> Helps to identify which variables should be used to define a unique product to ensure balance between product similarity and persistence through time

Aggregation

New UK aggregation structure (2024)

To enable the introduction of new data sources we will move to a new aggregation structure.

This has been developed under four key considerations:

1. We can realise more potential from alternative data sources, while continuing existing practices for our traditional collection to ensure good market coverage
2. We have the flexibility to use alternative data in combination with, or in place of, traditionally collected data, weighted according to our best information on retailer market share
3. We can more readily calculate regional/sub-national consumer price statistics, for which there has been growing demand in the UK
4. We enable a smooth transition towards the latest iteration of Classification of Individual Consumption According to Purpose (COICOP 2018), while also realigning our numerical system for our detailed (COICOP 6) level of the hierarchy coding with higher COICOP levels

Realising more potential?

Traditional: Sample item selected to represent broader group, e.g. Basmati rice chosen to represent all rice. Reduces collection cost/burden & maintains homogeneity within elementary aggregate (EA) indexes

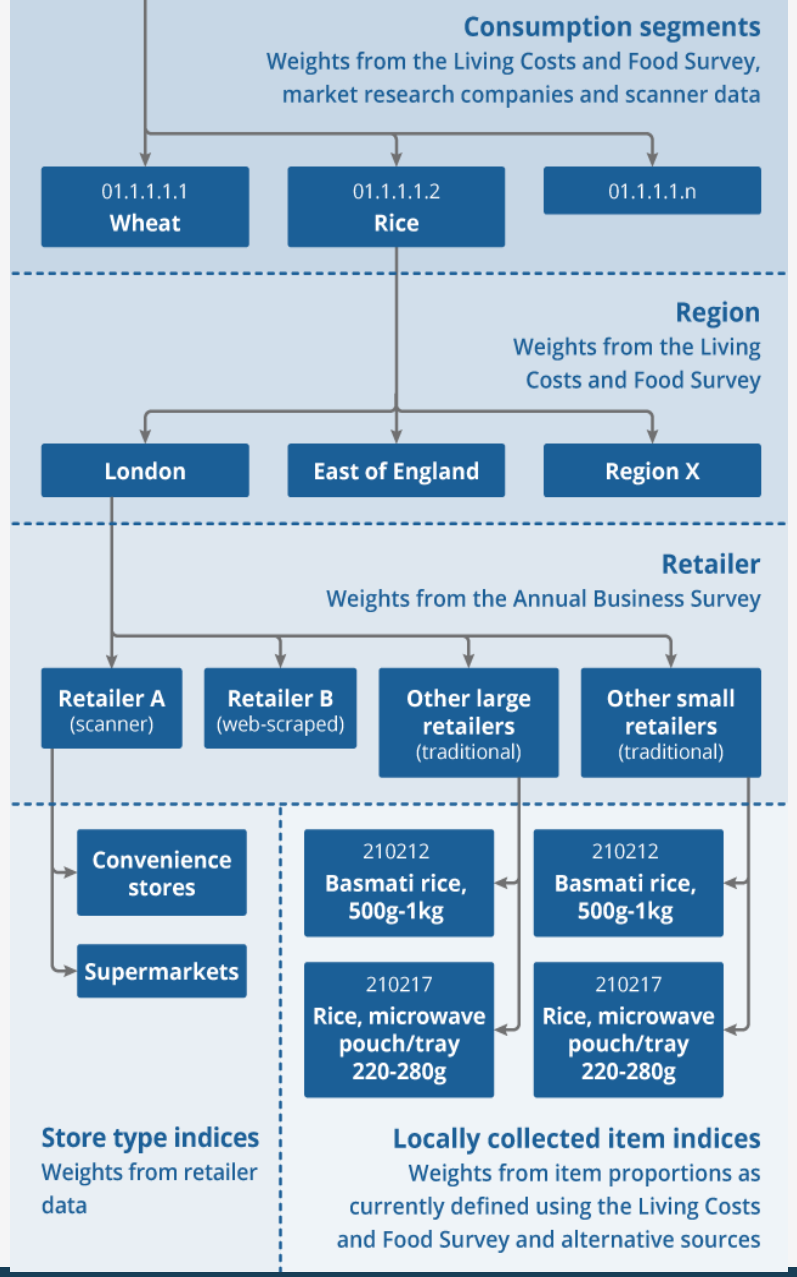
Scanner: Now receive data for all rice varieties, so can include and weight according to economic importance. BUT only from some retailers, how do we then use all these rice data without implicitly exaggerating the weight for the retailers who provide scanner data?

New COICOP 6 level, “consumption segments”, replacing old COICOP6 level “items”. These are typically equivalent or broader, allowing us to use more of the scanner data

New regional level, allowing us to more readily produce sub-national statistics

Retailer level allows us to weight retailers’ according to their market share, independently of how many data are provided – consumption segments are only introduced when we have >70% market coverage through either scanner data, traditional data or a combination

Stratum level indices allow us to maintain item definitions for traditional collections and produce more granular indices for some retailers helping us understand the households experiencing higher inflation. EAs can use either multilateral or bilateral methods



Index methods

Traditional data EA indices:

- Jevons; Dutot; Lowe/Laspeyres
- Fixed base
- Jan(y)=100

Scanner data EA indices:

- GEKS-T (otherwise known as CCDI)
- Mean splice
- 25 month window, re-referenced to Jan(y)=100

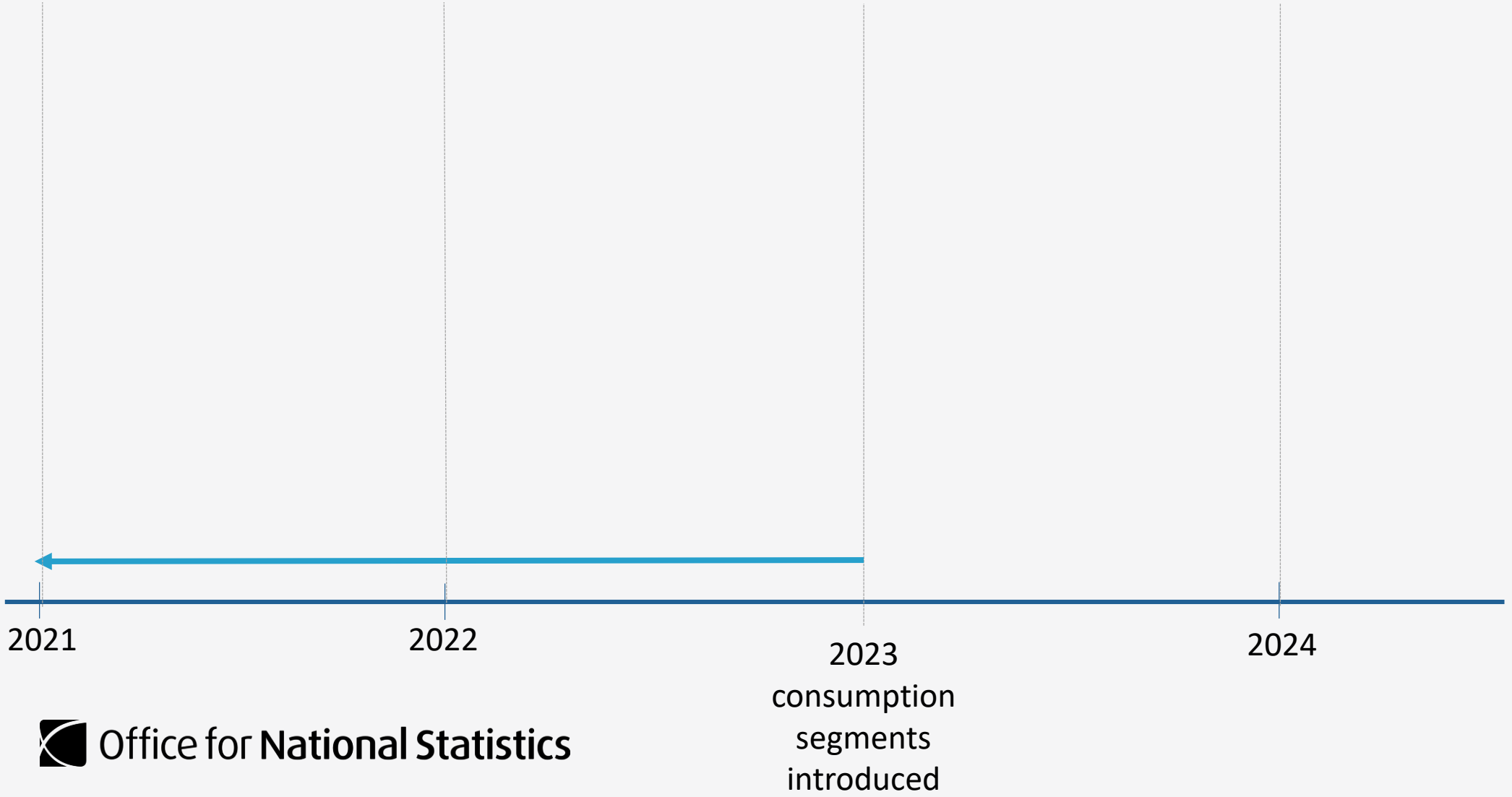
Aggregation of EA indices:

- Lowe
- Fixed-base
- Dec(y)=100, Jan(y)=100

Annual process

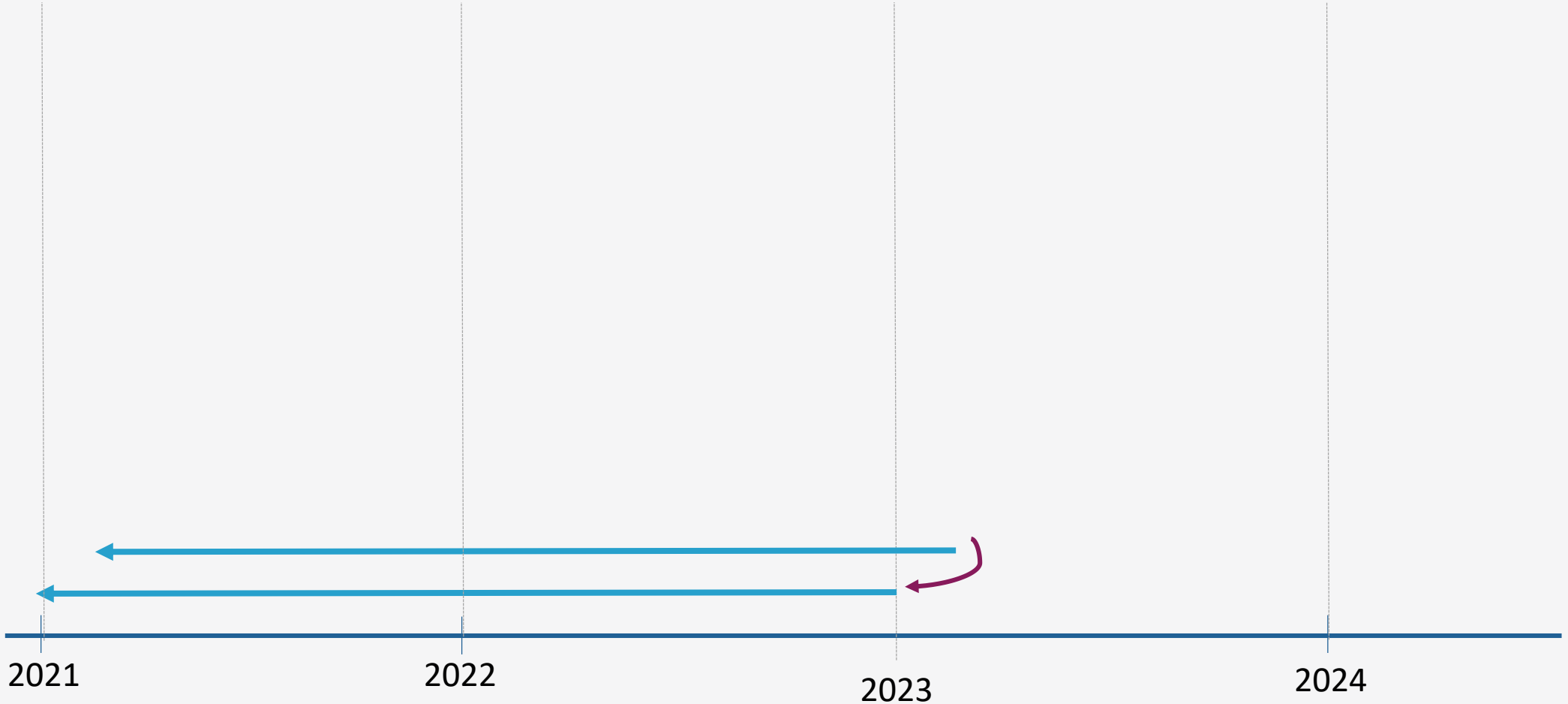
How does a traditional basket update work within a 25-month window multilateral method framework?

1. Clean and classify 25 months historic data and calculate GEKS across full window



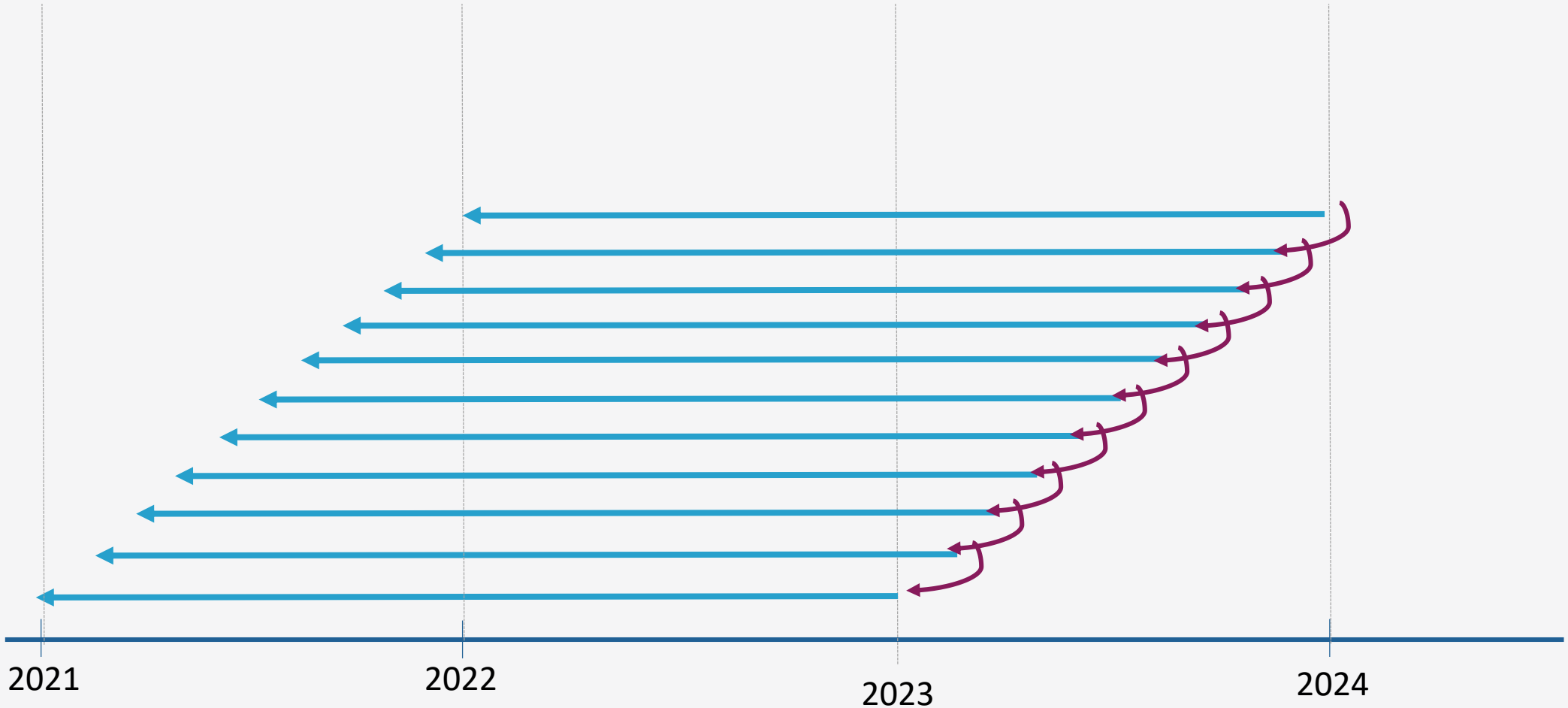
Office for National Statistics

2. Append month and calculate new 25 month window, mean splice onto existing series, rereference to base_month(y)=100



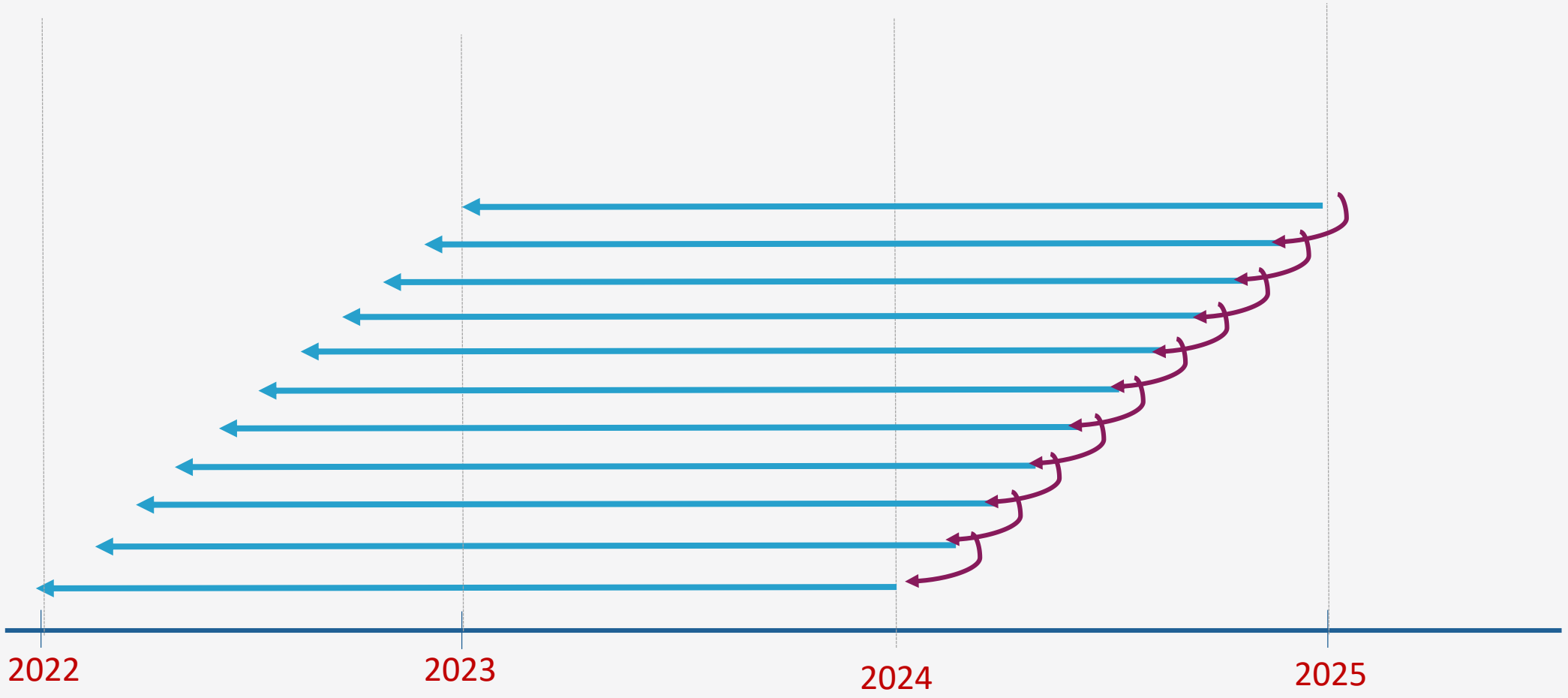
Office for National Statistics

3. Repeat monthly until new base month



Office for National Statistics

4. Repeat process for each consecutive year with updated consumption segments



Office for National Statistics

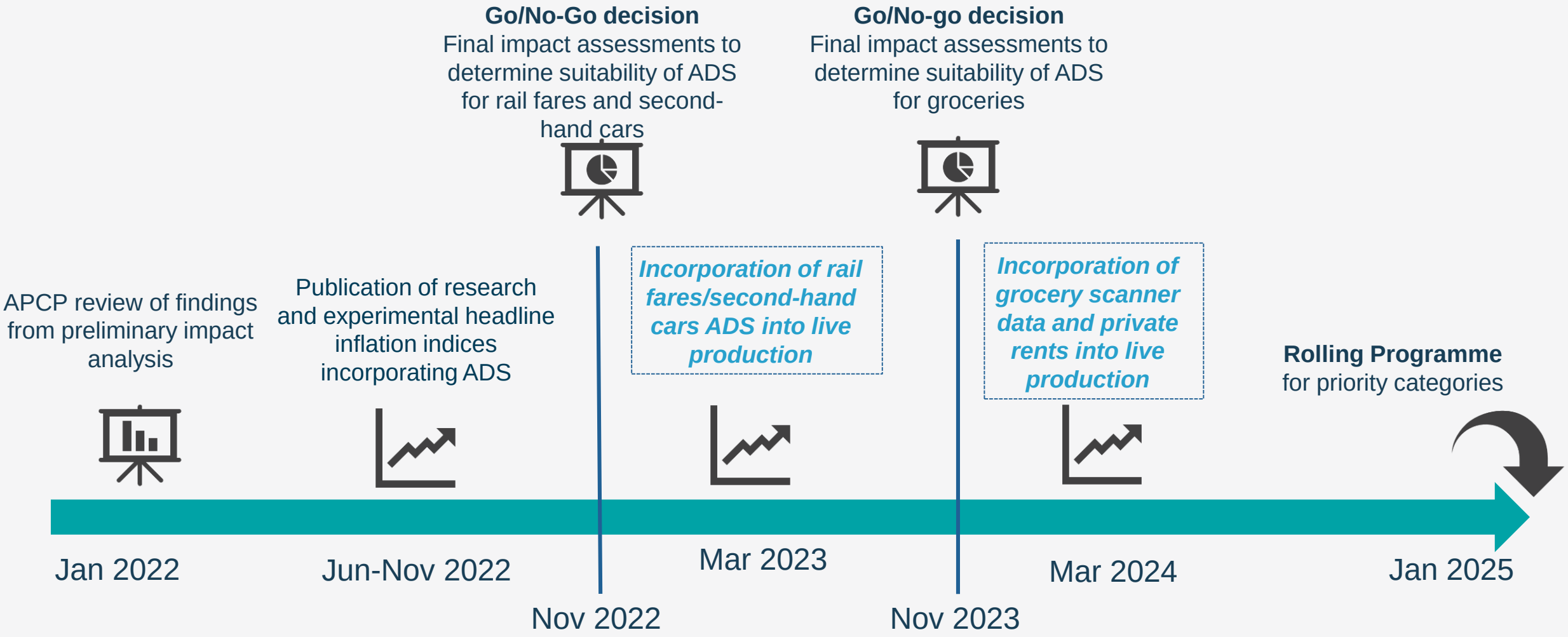
Ongoing work and next steps

The puzzle still isn’t complete!

Ongoing work

- Data cleaning methods
- Continued exploration of multilateral index methods
- Interaction between multilateral methods and higher level chaining
- Exploration into use of web-scraped data, including continued work on ML-based classification, product grouping and expenditure proxies
- Building end-end system

Timeline


Go/No-Go decision
Final impact assessments to determine suitability of ADS for rail fares and second-hand cars
Go/No-go decision
Final impact assessments to determine suitability of ADS for groceries
APCP review of findings from preliminary impact analysis
Publication of research and experimental headline inflation indices incorporating ADS
Incorporation of rail fares/second-hand cars ADS into live production
Incorporation of grocery scanner data and private rents into live production
Rolling Programme
for priority categories
Jan 2022
Jun-Nov 2022
Nov 2022
Mar 2023
Nov 2023
Mar 2024
Jan 2025

Questions?

Helen.Sands@ons.gov.uk