Urząd Statystyczny
w Opolu

ZASTOSOWANIE KLASTRA W OBLICZANIU WSKAŹNIKÓW CEN TOWARÓW I USŁUG KONSUMPCYJNYCH DLA DANYCH POZYSKIWANYCH METODĄ WEBSCRAPINGU

Związek Marcin, Widera Katarzyna

Urząd Statystyczny w Opolu

Plan prezentacji

1. Webscraping - pozyskiwanie danych
2. Wykorzystanie uczenia maszynowego
3. Budowa reprezentanta jako klastra
4. Metody obliczania wskaźników cen

Dobór próby do badania

Produkty (towary/usługi)

Źródła danych

Klasyfikacja produktów

~30 000 produktów

Przyczyny podjęcia wyzwania badawczego

Dostęp do dużej ilości danych

Zmienność produktów w czasie

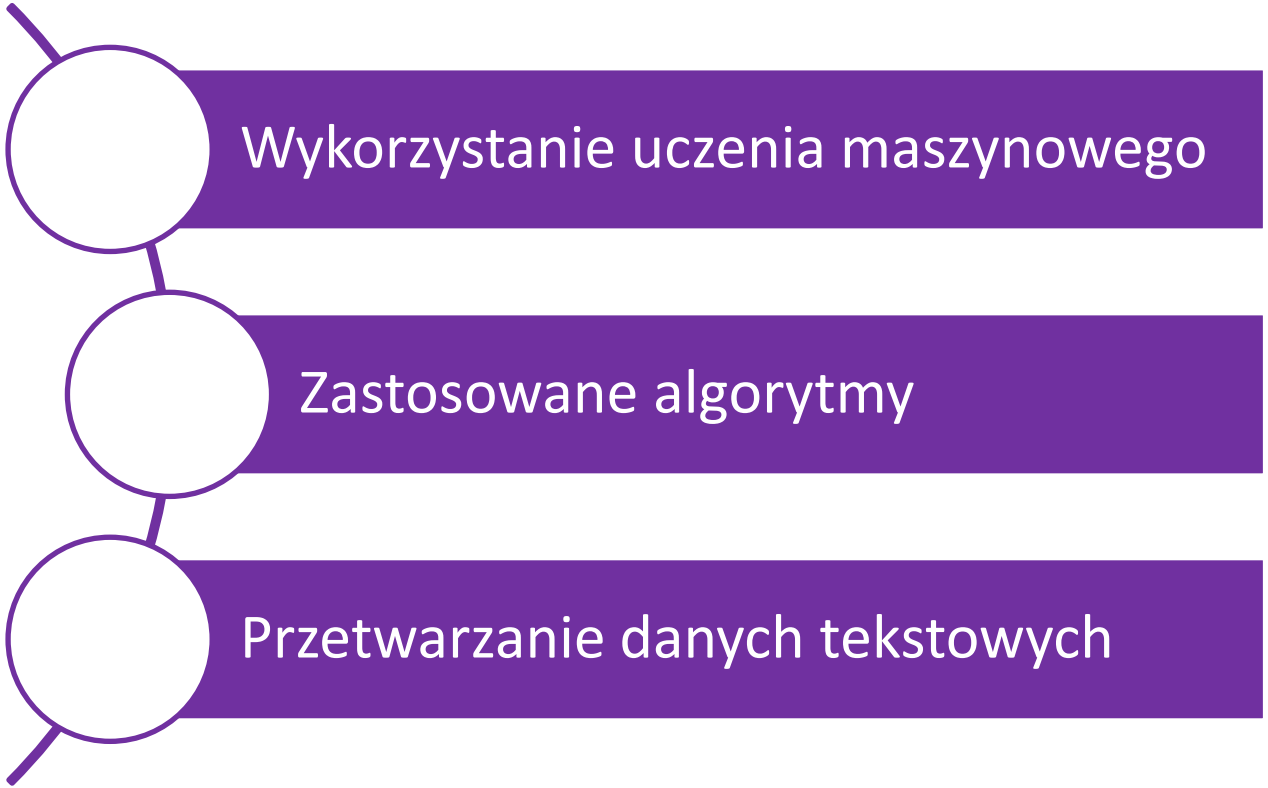
Brak ciągłości notowań produktów

Nowa metoda pozyskiwania danych
- udoskonalenie wskaźnika cen

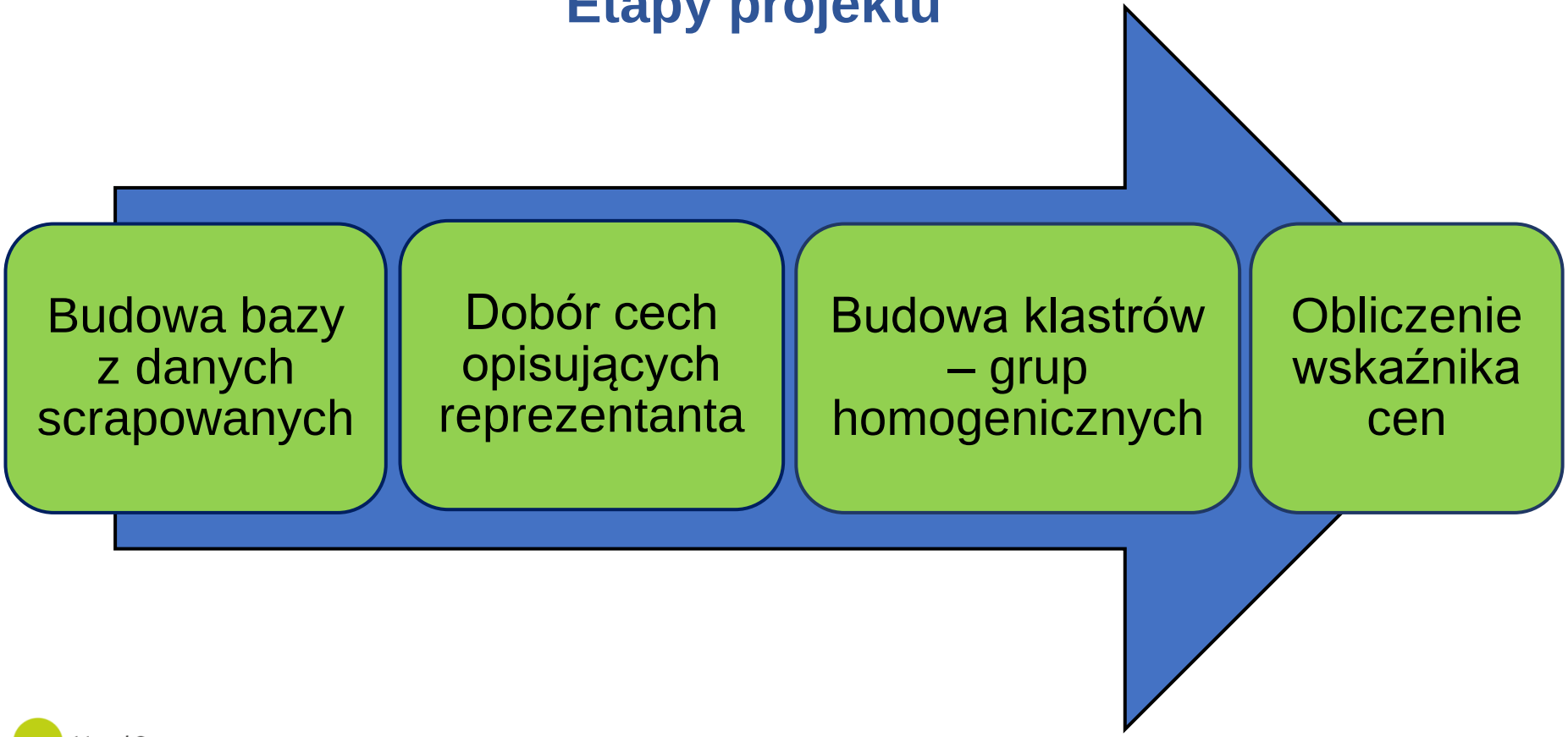
Webscraping

- Czyszczenie zbiorów, jakość danych, walidacja
- Format danych
- Częstotliwość, sposób scrapowania
- Zalety i wady wykorzystania webscrapingu

Uczenie Maszynowe



Etapy projektu



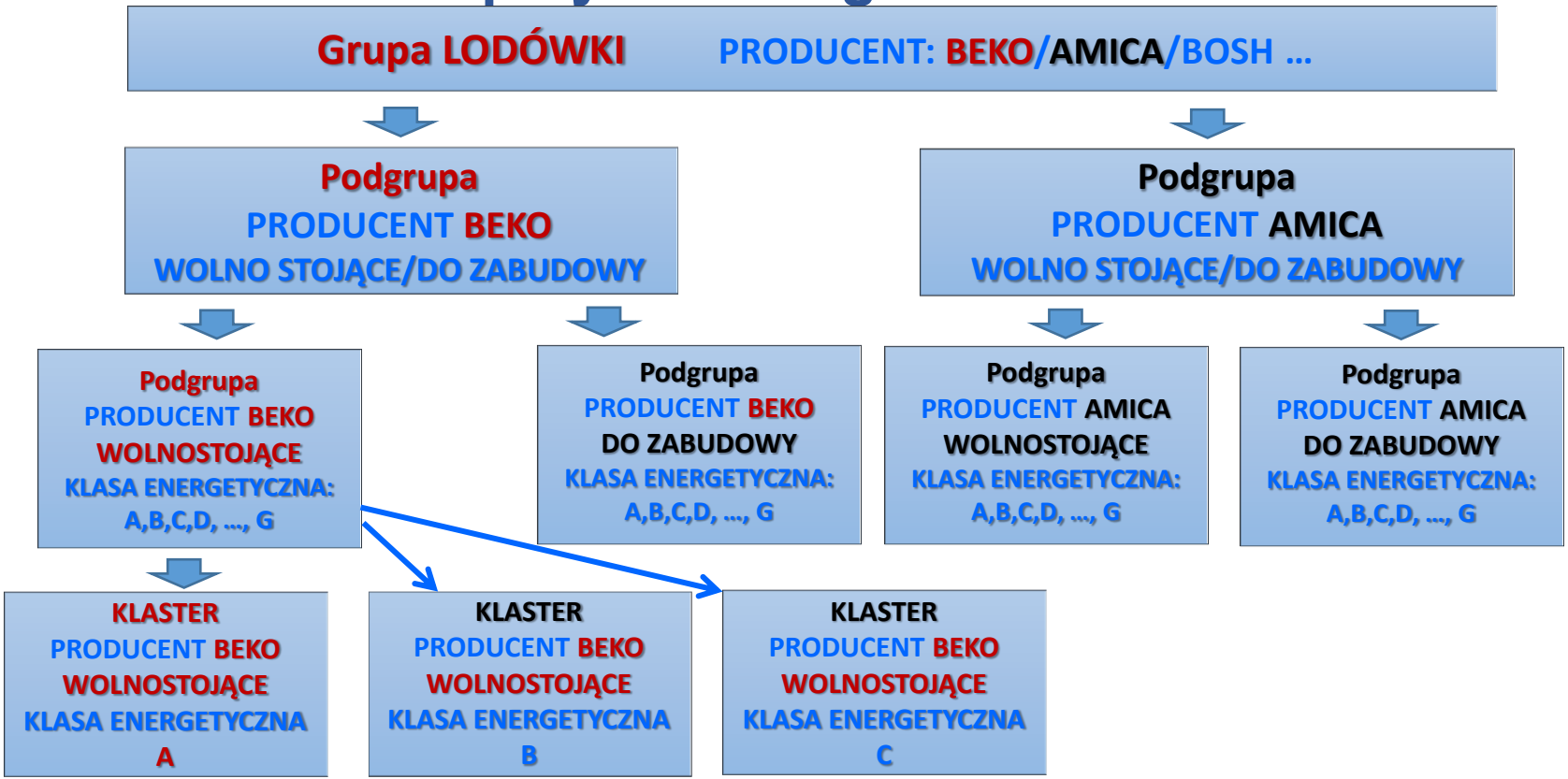
Dobór cech REPREZENTANTA

Budowa klastra - etapy klasyfikacji produktu

KRYTERIA DOBORU:

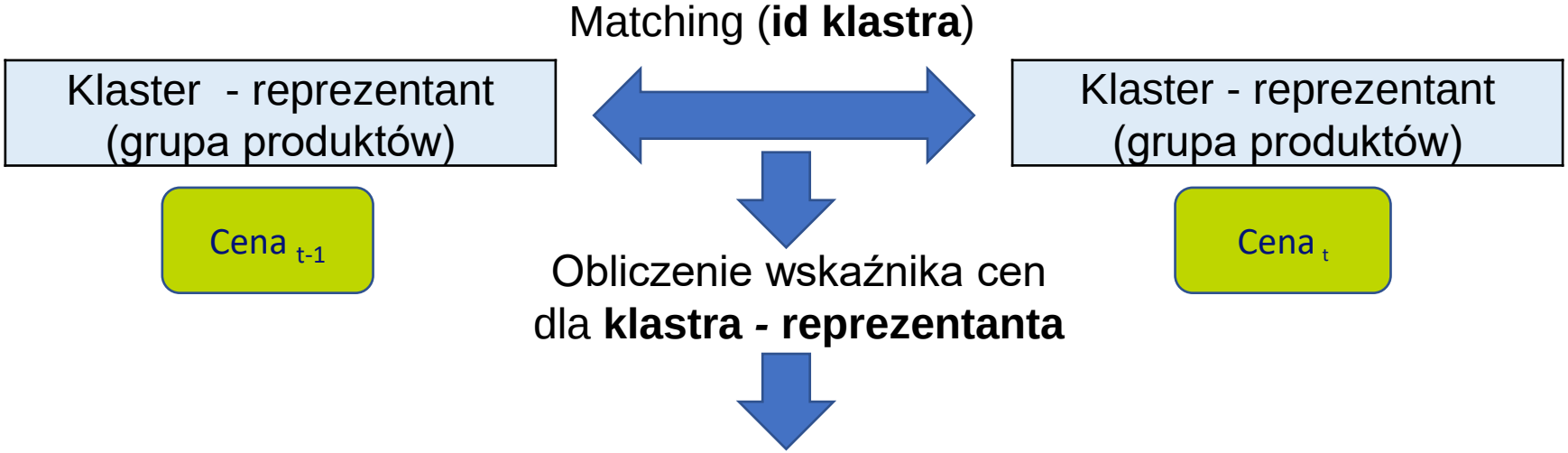
- wybór konsumenta
- zróżnicowanie poziomu cen przy podziale na kategorie (cechy)

Budowa przykładowego KLASTRA



Obliczanie wskaźnika cen

Obliczenie ceny **reprezentanta** - średnia arytmetyczna cen jednostkowych produktów w klastrze - cena dla każdego klastra



Obliczenie wskaźnika cen dla danej grupy **produktu –** (towaru/usługi)

Kwestie do rozwiązania

1. Wybór cech do budowy klastra – reprezentanta.
2. Liczebność klastra – filtr ilościowy.
3. Tzw. wagi - q (quantity):

- do zastosowania indeksów multilateralnych,
- by zmniejszyć fluktuację danych (przy zwiększonej częstotliwości ich pozyskiwania).

Podsumowanie

Webscraping jako źródło danych o cenach produktów pozwala na:

- pozyskanie dużej ilości danych bez obciążenia ankieterów statystycznych,
- zmniejszenie kosztu badań statystycznych,
- zastosowanie metod multilateralnych do obliczania wskaźników cen.

Klaster jako reprezentant pozwala na:

- obliczenie wskaźnika cen towarów w przypadku utraty ciągłości notowań dla pojedynczego produktu.

Literatura:

Metcalfe E., Flower T., Lewis T., Mayhew M., Rowland E. (2016): *Research indices using web scraped price data: clustering large datasets into price indices (CLIP)* Office for National Statistic of UK https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip

Mayhew M. (2017): *ONS methodology working paper series number 12 – a comparison of index number methodology used on UK web scraped price data* Office for National Statistics https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber12acomparisonofindexnumbermethodologyusedonukwebscrapedpricedata

Ayoubkhani D., Thomas H. (2022) Estimating Weights for Web-Scraped Data in Consumer Price Indices Journal of Official Statistics, Vol. 38, No. 1, 2022, pp. 5–21, http://dx.doi.org/10.2478/JOS-2022-0002

Guide on Multilateral Methods in the Harmonised Index of Consumer Prices (2022) Manuals and Guidelines, Eurostat

Dziękujemy za uwagę