Statistics Finland

Application of Alternative Data Sources for Price Statistics: Finland

Kristiina Nieminen
1st July 2025
UN Webinar

Kristiina Nieminen

- Master of Business administration and IT
- 25 years experience in statistics, of which 11 years in CPI production and 5 in IT department
- international cooperation with Mauritius, Kosovo, Bosnia-Herzegovina, North-Macedonia, Cyprus…
- focus areas: developing and improving CPI processes, delivering information services, and providing training on CPI fundamentals to both national and international colleagues.

/knieminen

+358 50 5925519

Kristiina.nieminen@stat.fi

Content of the presentation

Alternative data sources in CPI
Benefits and challenges of different datasources
Use of scanner-data for other purposes

Organisation of data acquisition
CPI monthly process
Components in CPI production system

Some ideas for making the process more efficient

Alternative data sources for CPI

Dimension	Traditional data collection by price interviewers	Direct data collection by CPI-team	Scanner-data (15 datasets)	Web-scraping / API (3 scrapers)	Administrative data/-registers (several)
Coverage	Sample of products and outlets	Sample of products and outlets	All products sold. Sample of companies, coverage >90%	Sample or most of the sold products. Sample of companies	Population
Number of observations/ month	13 700	100	Millions of observations	Thousands of observations	Hundreds of thousands of observations
Content	Prices	Prices, sometimes also number of sales	Sales value and number of sold products	Prices and product characteristics	Prices and quantities
Target COICOP-subclass	e.g. clothes, shoes, furniture, toys and games,	e.g. tobacco, water charge, new cars, car rental, refuse collection, passport	e.g. Daily products, pet products Alcoholic beverages, Pharmaceutical products Household appliances, Apartments, townhouses and detached houses, rents …	Prices for foreign flights Prices for holiday cottages Fuels	Rents
Data provider	Companies websites	Companies, government and municipal websites, web-questionnaire for restaurants	Grocery trade, Retailers Pharmaca Health Intelligence Federation of Real Estate Agency Popular online market place	Companies websites Service provider API	Tax Authority Social Insurance Institutions' register

Benefits and challenges of alternative data sources

Data source	Benefit	Challenges
Traditionally collected data	Robust procedures for product replacements	Prices are not always collected according to CPI instructions
Directly collected data	Robust procedures for product replacements	Static price catalogs are seldom updated
Scanner-data	Actual sales transaction All products sold – Large number of observations Complete data, including quantities sold	All products need to be classified according to COICOP Deployment requires methodological development Managing product substitution is more challenging Changes in data transfer and data structure
Web-scraped/API data	Large number of observations	Deployment requires methodological development Constantly changing web-sites
Administrative data	Large number of observations	Delay in data delivery

Use of new data sources for other purposes

In CPI production

- Derivation of the CPI weights at lower level of COICOP hierachy
- Derivation of retailer specific weights.
- Update of outlet and product sample.
- Selection of items/varieties to price collection

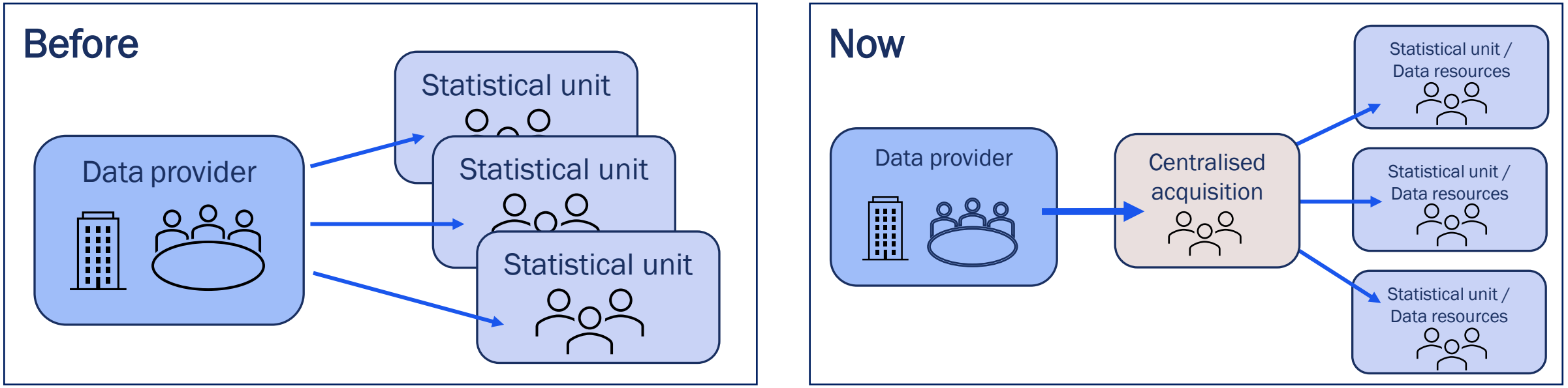
In information services

- Customer-specific feedback
- Chargeable information services such as ”Sales values of grocery products by product group”

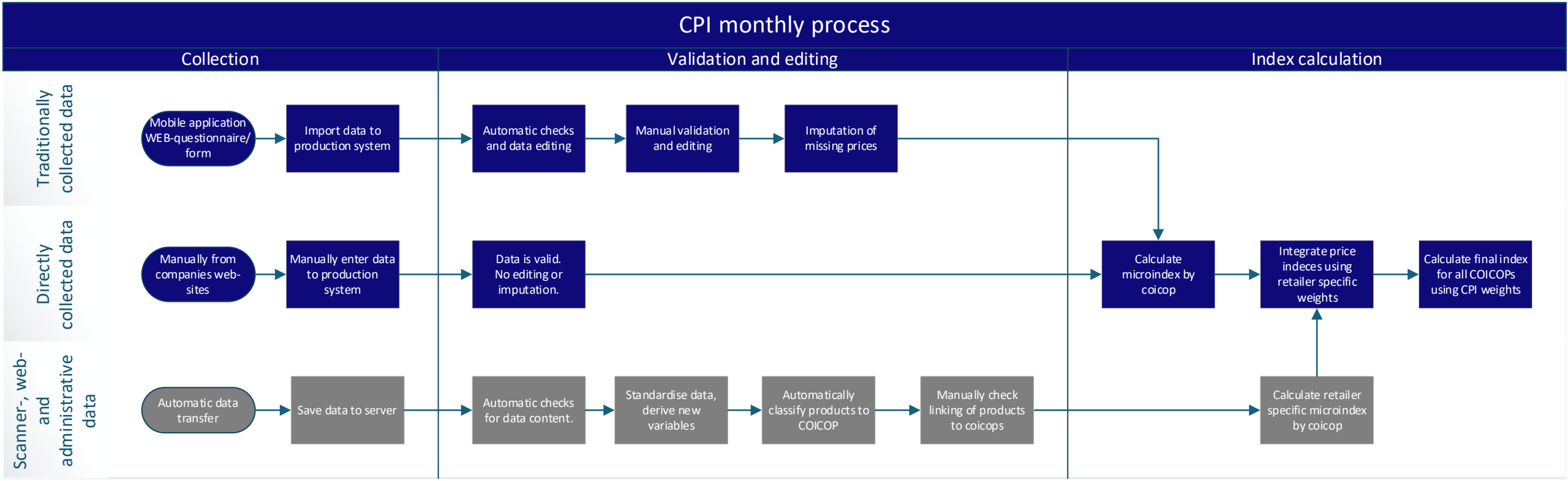
Other

- Data source for ICP (International comparison program)
- Data source for National Accounts /private consumption expenditures

Organisation of data acquisition

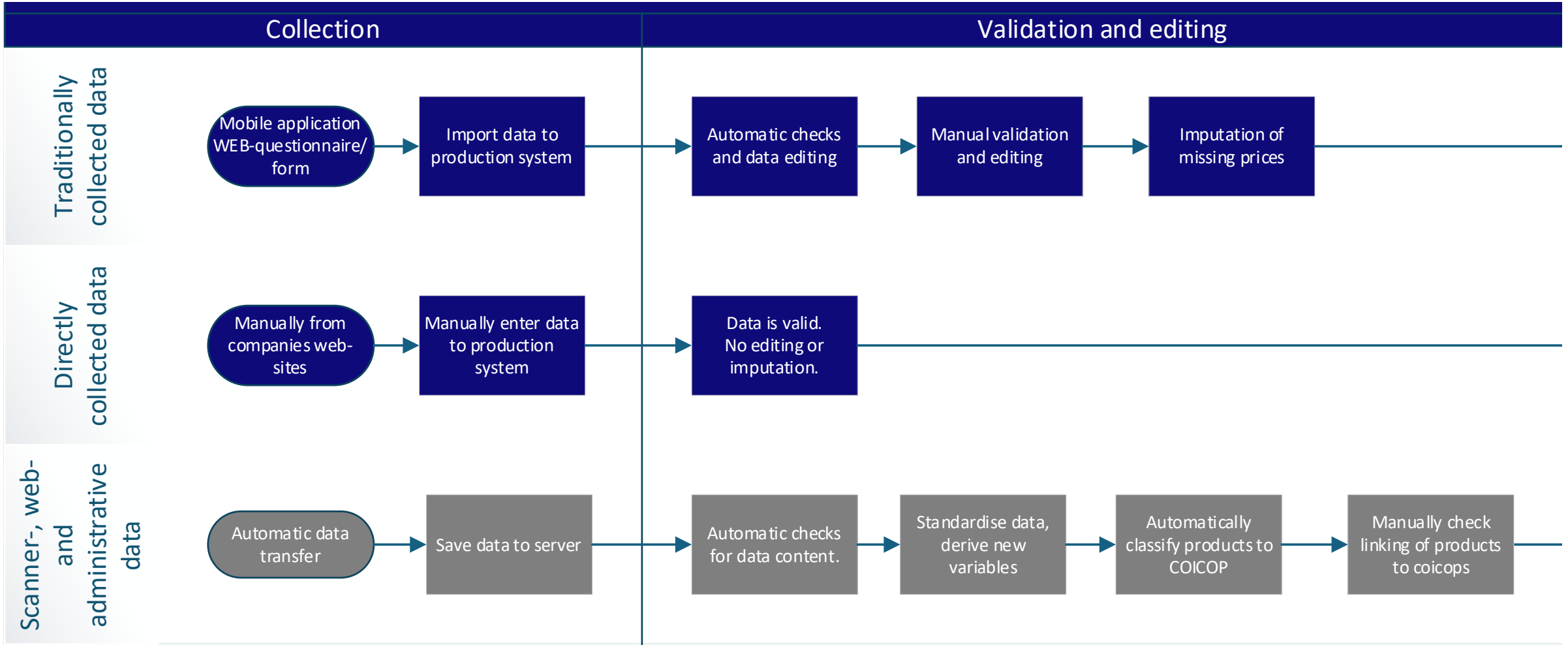


- **Previously**, data providers had to deliver the same data separately to multiple statistical units within Statistics Finland, which limited the broader utilization of the data and placed a significant burden on the providers.
- **Now,** data is delivered through a centralized acquisition process, enabling more versatile data dissemination and improved utilization.



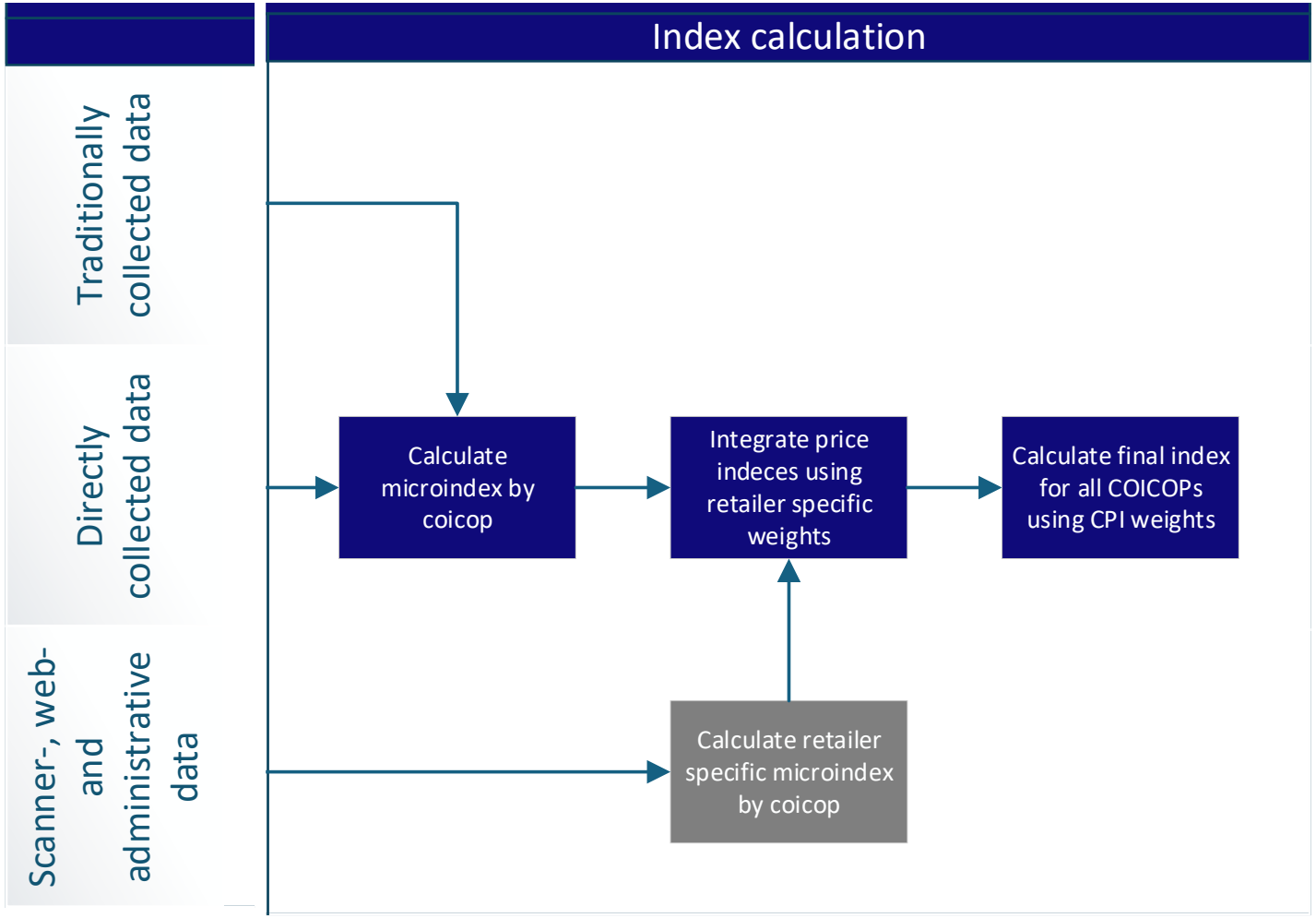
- The chart illustrates the monthly Consumer Price Index (CPI) production process in Finland. Although different terminology is used, the phases align with the steps of the Generic Statistical Business Process Model (GSBPM).

CPI monthly process: collection and validation ...



An independent production process was established parallel to the existing CPI-production (in blue) to accommodate the processing of new data sources

CPI monthly process: ...Index calculation



Before combining the different data sources, separate micro-indices are calculated for each dataset. These micro-indices are then aggregated using commodity- and retailer-specific weights.

Components in CPI production system

Traditionally and directly collected data

- Data collection mode:
  - Mobile phone application (in-house built)
  - Web-questionnaire (in-house built)
  - Excel
- Data storage: MS SQL server/database
- Data processing: In-house-built .NET application
- Data analysis: In-house-built .NET application

Scanner-, web- and administrative data

- Data collection mode:
  - Transmission codes (python) for automatically transferred csv-files
  - Python code for extracting data from API-layer or web-scraping internet pages
- Data storage: SAS server
- Data processing: SAS software
- Data analysis:SAS software

- Index calculation: SAS software
- Generating of dissemination tables: SAS software

Some ideas for making the process more efficient

- Traditional data collection (including direct collection) is planned to be replaced with scanner-data when scanner-data is available
  - Amendment of EU-legislation 223/2009– giving us power to ask for new data sources, including big data
  - Automatic classifiers and current index calculation methods need to be improved to introduce new products and service in the CPI
- Introduce web-based form for price interviewers
  - Replaces current mobile application
  - Used for product groups having fragmented markets such as hairdresser's charge, canteens, jewellery and clocks, ….
- Abandoning the regional index-calculation in 2026
  - Limit the number of price observations and minimize manual data validation and editing
  - Re-sample regional outlets without loss of coverage
- Introduce new production system within 2027-2029
  - Moving from databases to cloud-based storage and services
  - Replacing current in-house built user interface with simplified R Shiny applications.
  - Introduce PriceIndices-package. Introduce Power BI or R/Python in data analysis.

Thank you for your attention.

- Kristiina.nieminen@stat.fi