Statistics Canada

CPI Modernization

Presentation to the "Panel on Improving Cost-of-Living Indexes and Consumer Inflation Statistics in the Digital Age" – October 7, 2020

Delivering insight through data for a better Canada

CPI modernization plan

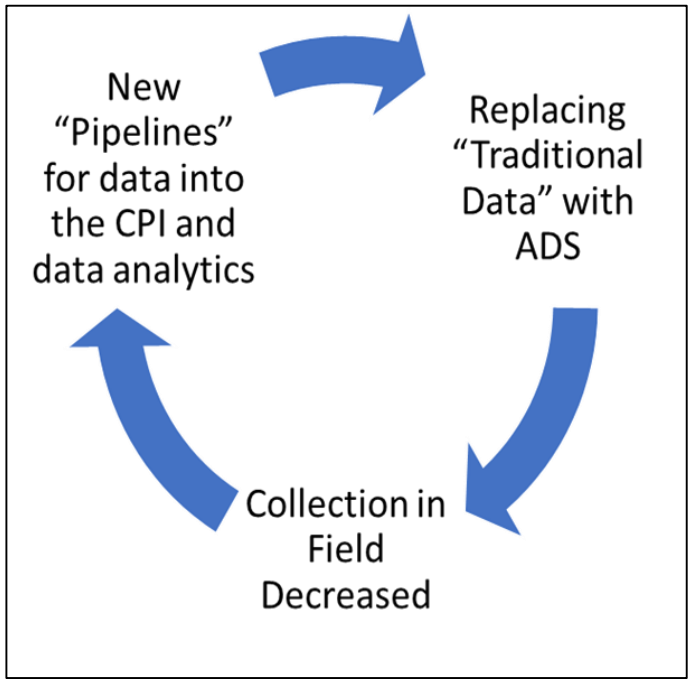
Vision

1. Increase the use of alternative data sources (ADS) for a more efficient and relevant CPI

2. Simplify processing systems to be more adaptable, scalable and easier to use

3. Develop a 'data depot' to increase data accessibility and interpretability and to facilitate analysis

4. Produce experimental series and alternative data products



CPI Modernization plan

Phase 1: Years 1 – 3
Replace *Existing* Prices

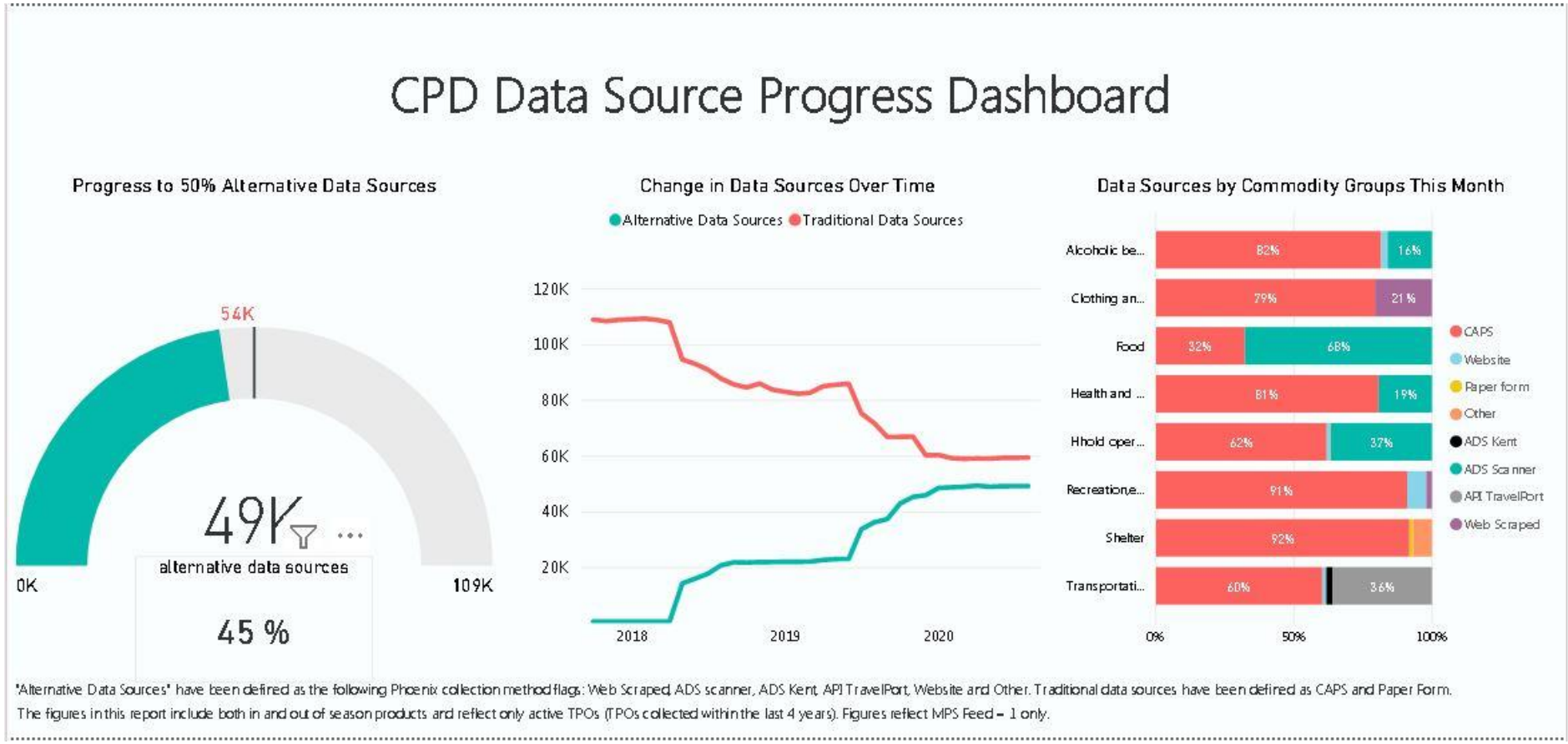
Commodity Class	% of quotes replaced
Food	83%
Household Operations	50%
Health and Personal Care	50%
Transportation	25%
Recreation, education, reading	25%
Alcoholic Beverages and Tobacco	18%
Shelter	10%
Clothing and Footwear	10%
Total Sample	**50%**

By March 2021, 50% of collected prices from alternate data sources. This represents 20% of the CPI Basket Weight.

Phase 2: Years 4 – 5
Use *ALL* the Data

- Research done in parallel to implementing Phase 1 will provide direction for methods and processes for Phase 2.
- By March 2023, 70-80% of collected prices from alternative data sources, representing 55% of basket weight.
- Using ADS, the CPI sample will be augmented through:
  - More products and more outlets
  - Use of quantities sold/revenues

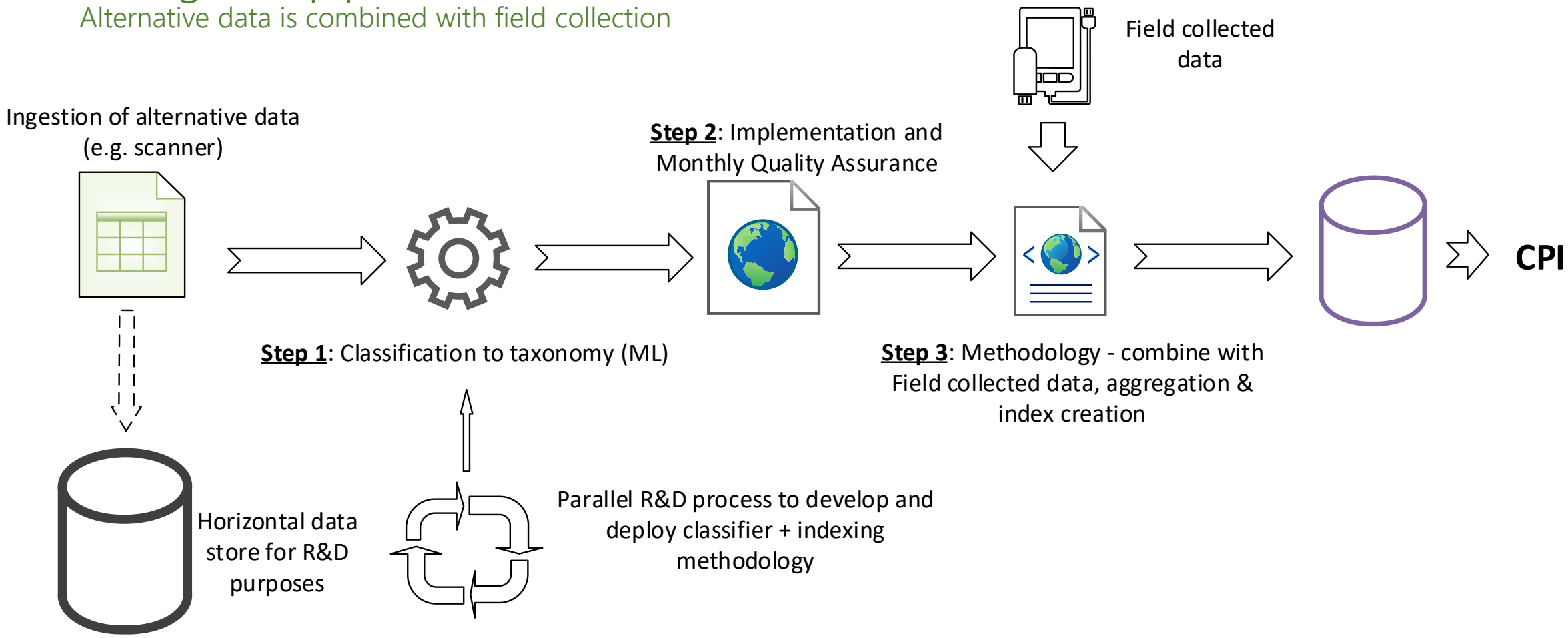
CPI Modernization dashboard



Implementing alternative data in the CPI

Changes to pipeline:

Alternative data is combined with field collection



Step 1: Using Machine Learning (ML) to process and map alternative data

- Classification needed to make sure that the right products in each monthly dataset are assigned automatically to the correct CPI categories, prior to application of various index methods;
- Supervised Machine Learning methods excel at this:
  - Requires a costly process to create the training data, or a subset that has been manually and correctly labelled, from which the classifier 'learns' – for each included retailer;
  - Stringent testing and validation must also be utilized prior to accepting a classifier;
  - Simple methods have proven to be highly effective, as well as incurring less training and infrastructure costs.
- Some error is always present however, hence a quality assurance stage post classification is used to validate monthly sample;

Steps 2 & 3: Implementation & methodology of alternative data

- Three additional applications were created to enable the processing and "pushing" of this type of data into our existing production system

Source	Retailers	Data availability
Scanner Data	Food retailers	Prices, characteristics and quantity
Web-scraping	Canadian clothing retailers	Prices and characteristics
Application Programming Interface (API)	2 aggregators for hotels, flights, car rentals, package holidays, etc.	Prices and characteristics
Other Data	3rd party providers	Prices and some characteristics

ADS implementation – Process lessons learned

- For all sources, key things are
  - Clear processes set out for acquisition and use of data within the department
  - The ability to efficiently classify data (see machine learning)
  - Engage and consult with Production area
    - Ensure they have proper tools to review and understand new type of data
    - Visualizations and well placed reports = key to summarize and support analysis
  - Even simple implementation will result in significant labour costs to maintain
  - Engage and consult with IT and systems support for appropriate tools for ingestion, processing, analysis
- Scanner data
  - Acquisition process is lengthy
  - There is little motivation for retailers to cooperate
- Webscrape data
  - Contracting 3rd party to scrape for us was best option as we did not have expertise internally to fully support
  - Process and expertise needs to be developed internally to make it acceptable for production use
- API data
  - Well structured data and relatively easier to work with

ADS implementation – ML + infrastructure lessons learned

- Homogeneity and churn in the data mean some quality assurance useful:
  - Some categories are naturally less homogeneous than others, meaning that a classifier will struggle with the harder ones. Hence manual sample quality analysis useful where this can impact the final numbers.
- Infrastructure and training:
  - Cloud and open source platforms are very useful, but should be designed with infrastructure that facilitates robust processes (such as replicability and source control). Without a proper setup, manual processes may be highly resource intensive
  - Architecting the ADS approach best done holistically, traditional SQL applications per data source or solutions can lead to a portfolio of siloed applications, making maintenance and adoption slower and harder. Best designed shared UIs and processes.
- Process:
  - Quality assurance is a valuable feedback loop to automatically label more data on a monthly basis.

ADS implementation – Methodology lessons learned

- Challenges
  - Difference between data available and structure/format across retailers
  - Web scrape - no sales information
  - Combine prices from different data sources and use different method
  - Churn aggravates implementation, especially for data sources such as web-scrape clothing
- Considerations
  - Scope of the CPI: domestic vs international retailers
  - Quality of CPI is top priority: cautious and gradual implementation to mitigate risk of error

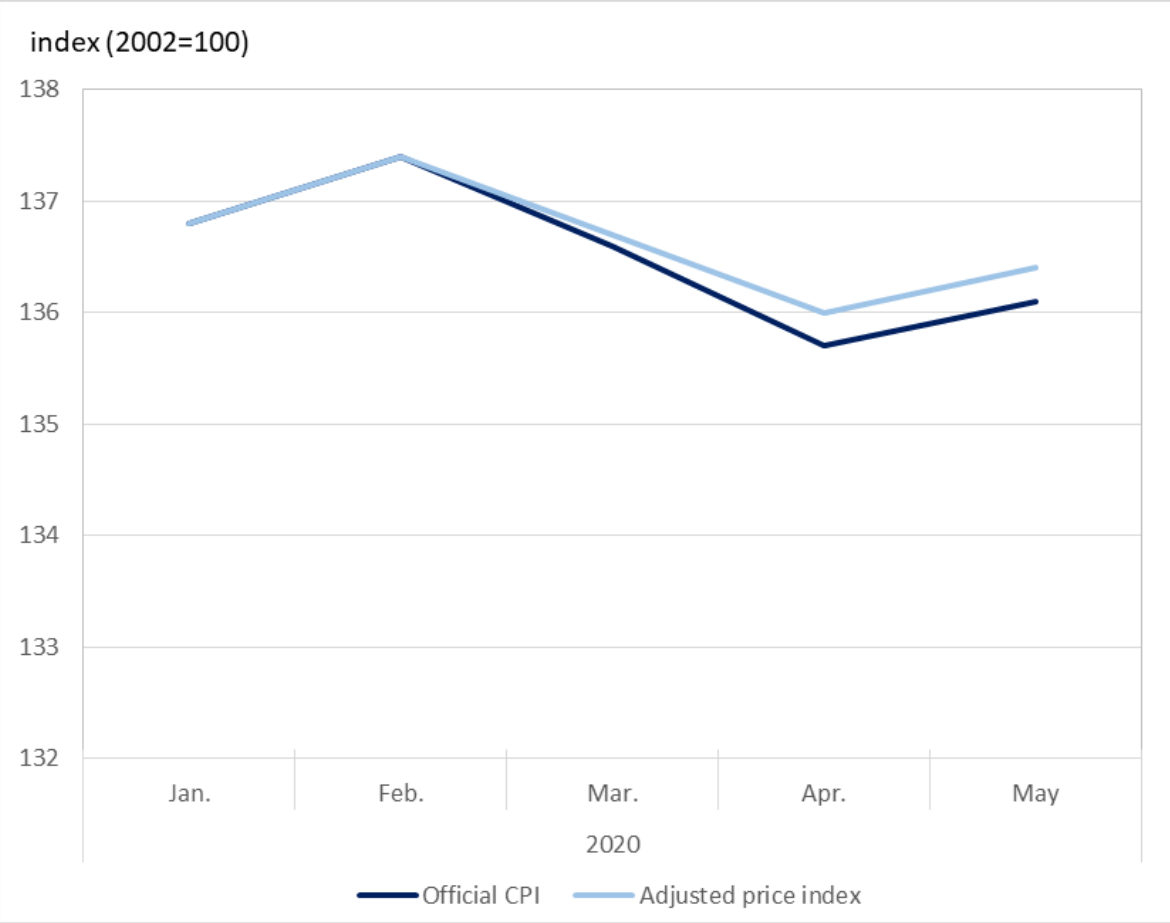
Responding to the COVID-19 pandemic

- experimental work with new sources of expenditure data
- calculating an adjusted 'COVID-CPI'
- informing methods and frequency of future basket updates

CPI basket updates

- Consumer expenditures during COVID-19: An exploratory analysis of the effects of changing consumption patterns on consumer price indexes
- Monthly adjusted price index
- Monthly adjusted consumer expenditure basket weights
- Inter-basket update adjustments

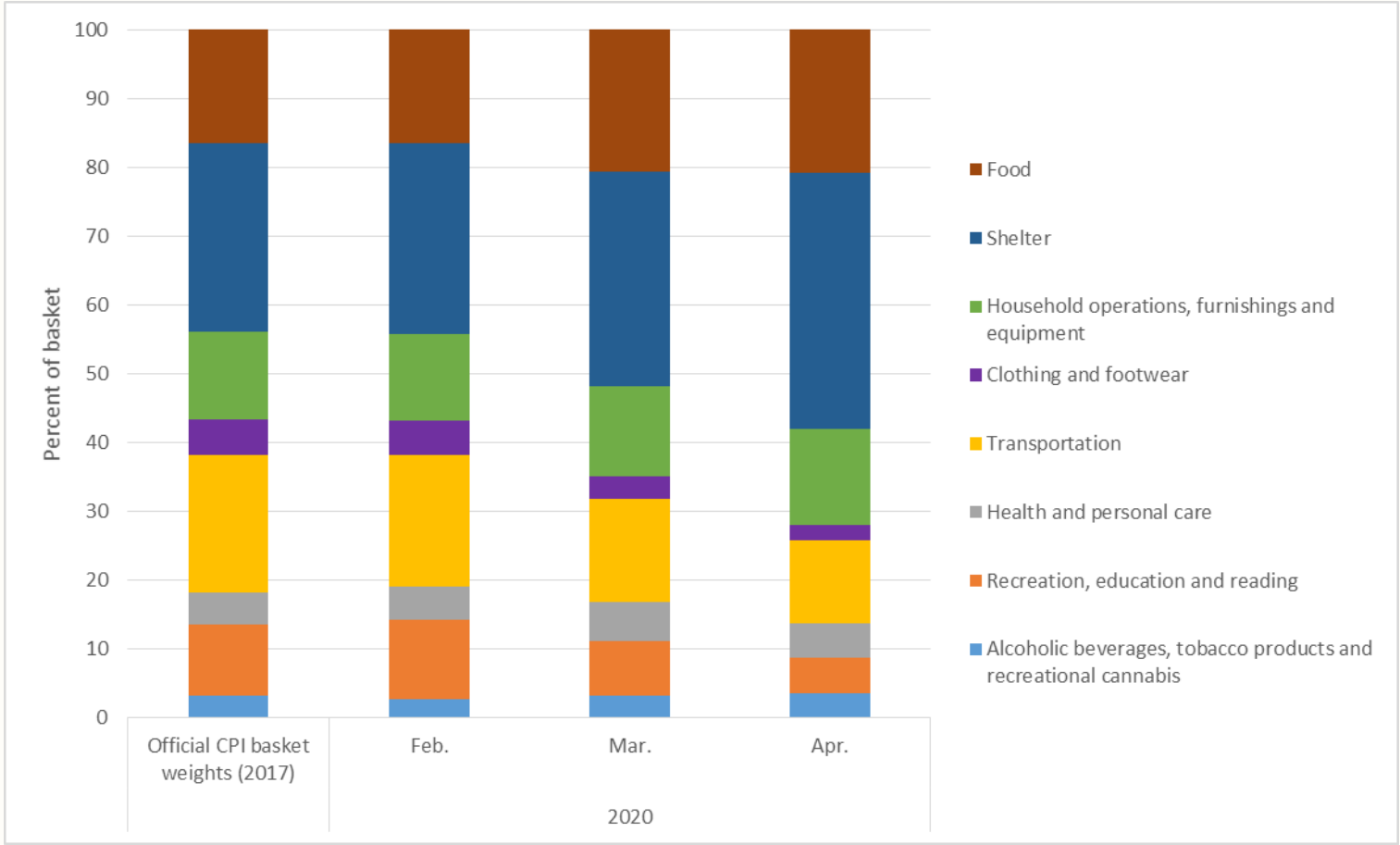
Consumer Price Index (CPI) and adjusted price index, Canada, February 2020 to May 2020 (October 8th release: update to August 2020 with revised data)



Lessons learned & best practices

- Agility over traditional approaches to ‘rigour’
- Perception matters
- Transparency
- Communication and public engagement

Basket weights of major components in the official Consumer Price Index (CPI) and the adjusted price index, Canada, February 2020 to April 2020 (October 8th release: update to July 2020 with revised data)



Services and the digital economy

- development work to enhance representativeness of the CPI basket

- digital products, services, digital platforms, shipping costs, etc.

Services & Digital Economy: Ongoing projects

Service industries

- Services coverage
  - Analyze current sampling of services versus goods, geographical distribution and dispersion, identify problem areas, suggest replacement strategy (admin, online, etc.)
- Improve shipping services methodology
  - Improve current methodology on parcels, conduct research on shipping services in general, and expand the CPI scope
- Free online services
  - Experimental study, plans to design a survey

Digital Economy

Computers and peripherals index

- Used web-scrapped data, created characteristics, input in hedonic and random forest models, create indices

Create a special aggregate for digital economy

- Define the scope of the special aggregate and publish the aggregate in FY2021-22

Increase online price coverage

- Web-scrape certain websites for best-sellers, introduce new products, confront current prices collected, replace certain field price collection

Challenges/Lessons Learned

Data Acquisition

- No official and detailed information about household spending on e-commerce and shipping services
  - Investigating alternative data, leveraging other surveys,
- Sales data for online purchases
  - Use alternative data, such as retail sales, and some information from panel surveys
  - Coming up with methodologies to estimate sales from web-scraped data categories

Methodological challenges

- High product churn in technology intensive products, classification and mapping challenges, creating an API pipeline that fits all (computers versus clothing categories)
  - Heterogeneity of web-scraped data requires multiple approaches to parse efficiently the product characteristics
  - Create Asteroids (in-house system) as temporary measures and future work on a global pipeline
- Survey designs for experimental work; behavioural survey
  - Leverage existing surveys; use alternatives such as Input Output data
- Rapidly evolving industry that requires constant monitoring
  - Challenges in introducing a new products in the system
  - Challenges in making proper quality adjustments

Thank you!

Links to CPI data products, papers and interactive tools:
CPI Portal

heidi.ertl@canada.ca
serge.goussev@canada.ca
chris.li@canada.ca
roobina.keshishbanoosy@canada.ca