Transformation of the Canadian Consumer Price Index (CPI)

Data Science Leaders Network Sprint

January 24, 2024

Serge Goussev and Wesley Yung

Outline

- Background
- Business drivers for transformation
- Phases of the transformation
- Limitations and challenges
- How we are overcoming those challenges:
  - Principles
  - Focusing on holistic solution for all alternative data
  - MLOps
- Lessons learned and summary

Background 1/3

- The Canadian CPI is an indicator of the change in consumer prices
- It uses a fixed basket, so changes should reflect only price changes
- Important uses include measuring inflation and contract and pension increases
- The CPI uses a sampling approach to cover the many different products and the geography of Canada
  - Products are placed into product classes
  - Canada is divided into geographical collection areas

Background 2/3

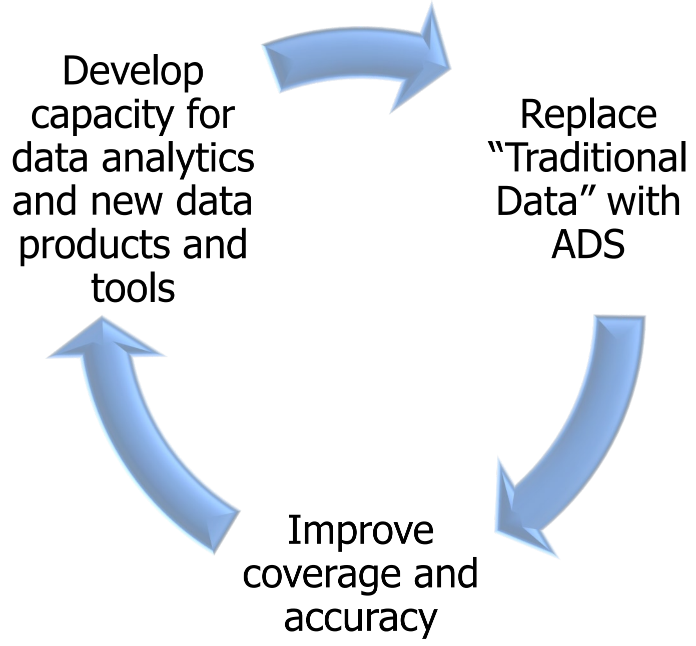
- Geographical areas are selected to ensure representation of all of Canada
- Within the selected areas, a sample of outlets are selected
- Prices for representative products for product classes are then collected from the selected outlets
- Price index theory then use these collected prices to produce the CPI
  - For more details see The Canadian Consumer Price Index Reference Paper (statcan.gc.ca)

Background 3/3

- For many years, prices were essentially collected ‘in-person’
- In 2015, StatCan started to receive scanner data from one retailer after much negotiation
  - Win/win situation for both StatCan and retailer
  - Weekly file (very large) which contained prices, quantity sold, SKU, UPC, store identification and description of the item
- Benefits of scanner data
  - Cost Savings: Ability to decrease field collection earlier
  - Reduced response burden: No collection burden on this retailer’s outlets
  - Accuracy: More prices are used
  - Representativeness: Ability to use quantity data to initiate the product sample as well as for substitution product selection

Business drivers for transformation

- ➢ Increasingly adopting alternative data sources (ADS) for a more accurate and relevant Canadian CPI
- ➢ Utilize Machine Learning (ML) and advanced price index methods to process near universe set of products consumed in Canada
- ➢ Develop dynamic processing systems to be more adaptable, scalable and easier to use.
- ➢ Produce experimental series and alternative data products to support insight on price trends in Canada.

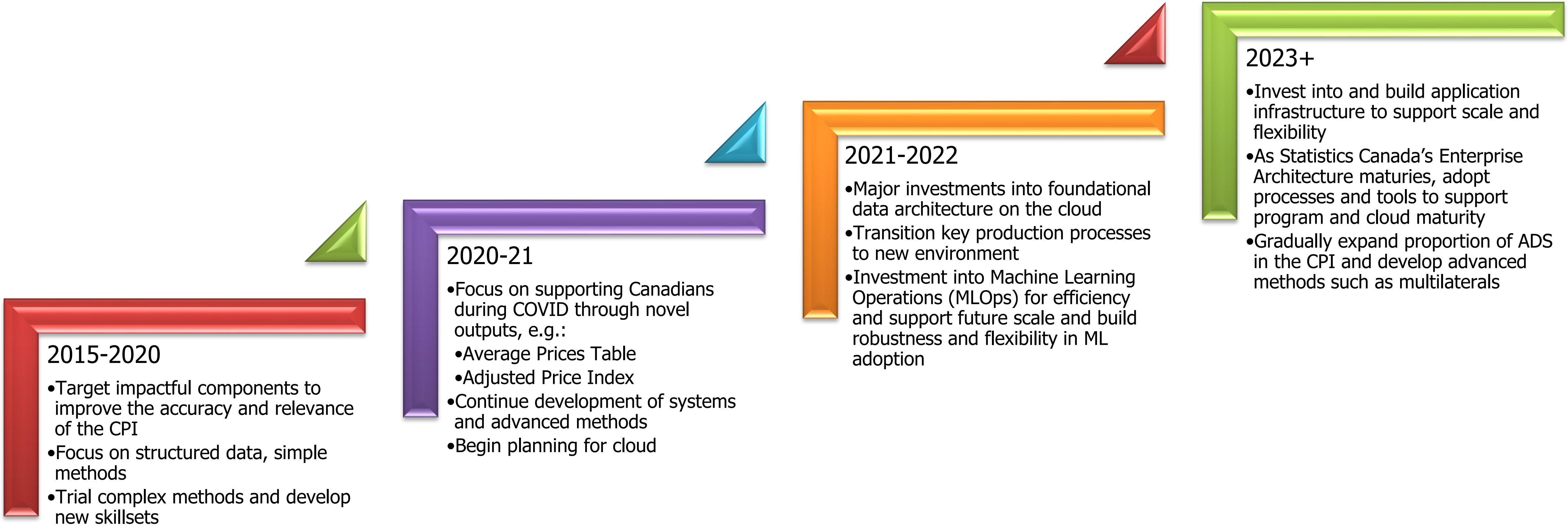


Develop capacity for data analytics and new data products and tools

Replace "Traditional Data" with ADS

Improve coverage and accuracy

Phases of the transformation



2015-2020

- Target impactful components to improve the accuracy and relevance of the CPI
- Focus on structured data, simple methods
- Trial complex methods and develop new skillsets

2020-21

- Focus on supporting Canadians during COVID through novel outputs, e.g.:
  - Average Prices Table
  - Adjusted Price Index
- Continue development of systems and advanced methods
- Begin planning for cloud

2021-2022

- Major investments into foundational data architecture on the cloud
- Transition key production processes to new environment
- Investment into Machine Learning Operations (MLOps) for efficiency and support future scale and build robustness and flexibility in ML adoption

2023+

- Invest into and build application infrastructure to support scale and flexibility
- As Statistics Canada's Enterprise Architecture maturies, adopt processes and tools to support program and cloud maturity
- Gradually expand proportion of ADS in the CPI and develop advanced methods such as multilaterals

Challenges faced during initial phases

Technical and business

- Scale of the data considerable (billions of rows, millions of unique products, dozens of terabytes). Acquisition leads to exponential increase of data volume
- Machine Learning at scale brings its own challenges:
  - Most processes including ML model production needs to be automated
  - Need high level of transparency, governance and management
- Data acquisition via third parties and segregated processing leads to challenges of lineage tracking, transparency and data quality
- Introduction of new models require back-testing experiments and parallel (shadow) deployments
- Infrastructure to enable horizontal access to data at scale

Organizational

- Investing and upskilling staff, increasing technical skills
- Change management as data scale and approaches require adoption of new processes and tools
- Coordination within the program and agency for effective use of data
- Data governance framework to ensure accessibility control

Principles to mitigate the challenges

Transparency in production and R&D processes

Focus on development of reproducible pipelines for production or R&D

Ability to register models and datasets (including metadata for discovery and interoperability)

Version control of code and orchestration pipelines

Horizontal access to the data for R&D and analytics

'Break down the data silos'

Provide analytical insight from all data sources

High processing capacity

Ability to process large data at scale, and scale down upon run completion

Adoption of appropriate tools, open standard and solutions

Access modern tools for R&D or production (critical for Data Science work)

Security

Maintain access control for datasets throughout all environments and their entire lifecycle, not just at source

Auditability of access

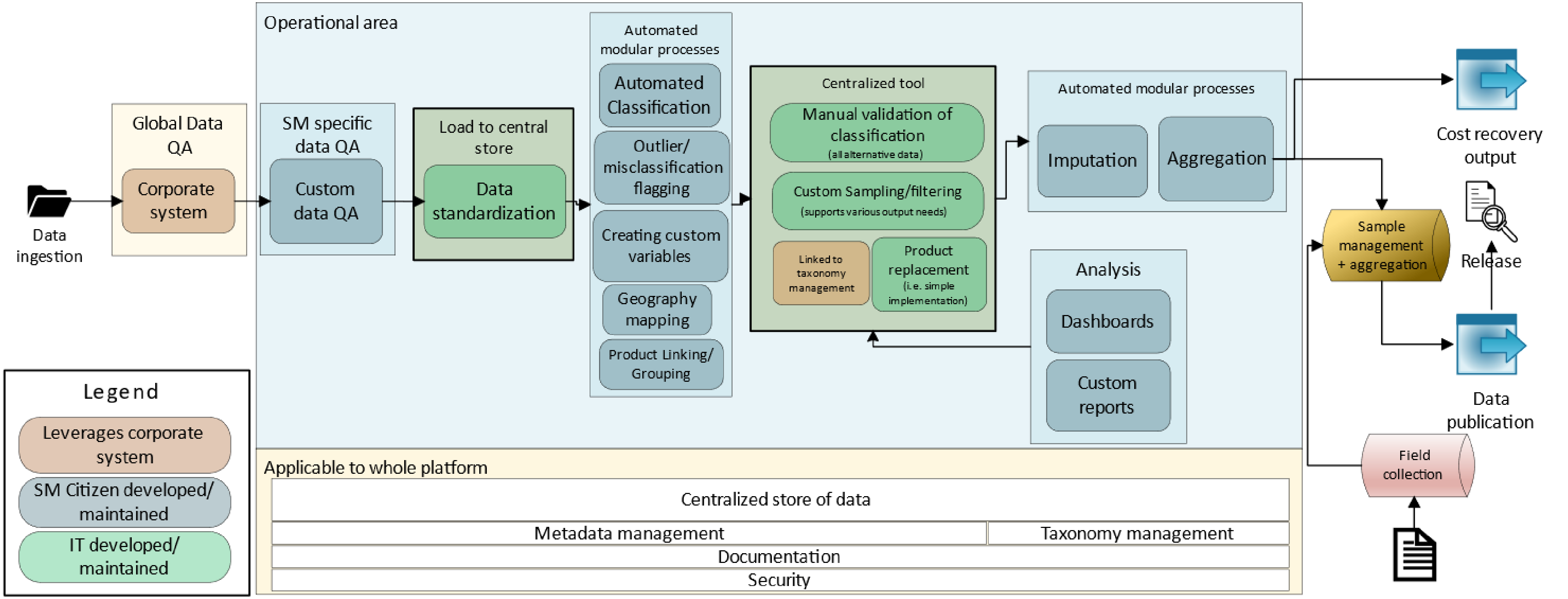
Cost-effectiveness

Elastic processing capacity

Cost transparency

Designing through the lens of standard capabilities

Capabilities for CPI production, focus on alternative data

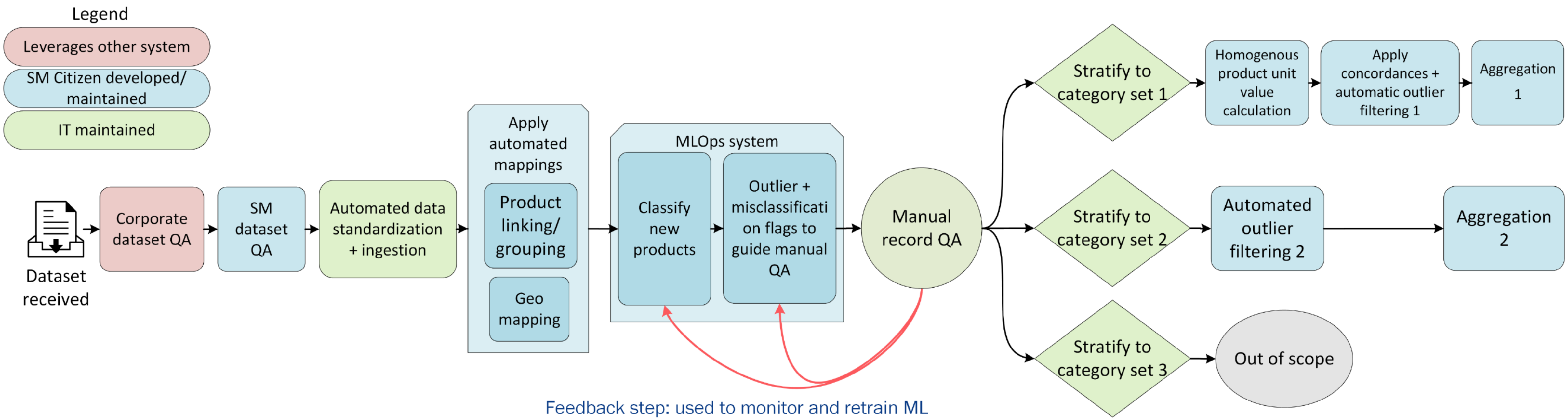


Key takeaways:

- Alternative data requires a separate pipeline from traditional field collection
- Capabilities necessary for multiple alternative data sources are quite consistent – capabilities are currently being drafted by the UN Task Team on Scanner data
- Implementation/development can be done by different groups working on components in a modular fashion

Applying this to an example

End to end example of 3 production pipelines leveraging one retailer dataset


Legend
Leverages other system
SM Citizen developed/ maintained
IT maintained
Dataset received
Corporate dataset QA
SM dataset QA
Automated data standardization + ingestion
Apply automated mappings
Product linking/ grouping
Geo mapping
MLOps system
Classify new products
Outlier + misclassificati on flags to guide manual QA
Manual record QA
Stratify to category set 1
Homogenous product unit value calculation
Apply concordances + automatic outlier filtering 1
Aggregation 1
Stratify to category set 2
Automated outlier filtering 2
Aggregation 2
Stratify to category set 3
Out of scope

Feedback step: used to monitor and retrain ML Models + tune manual outlier/misclassification flags

Key takeaways:

- Not all capabilities needed all at once – each retailer will need a different set of production pipelines to produce several outputs.
- Development of capabilities in a modular way allows interchangeability as methods need to evolve or to incorporate improvement in technology or tools
- Transparent development enables trust and partnership between statistical programs in the agency, allowing robust integration of one data source for multiple statistical outputs

MLOps in focus

- What is MLOps:
  - Automated effective, efficient, *transparent*, iterative delivery of ML models for production while also focusing on business and regulatory requirement

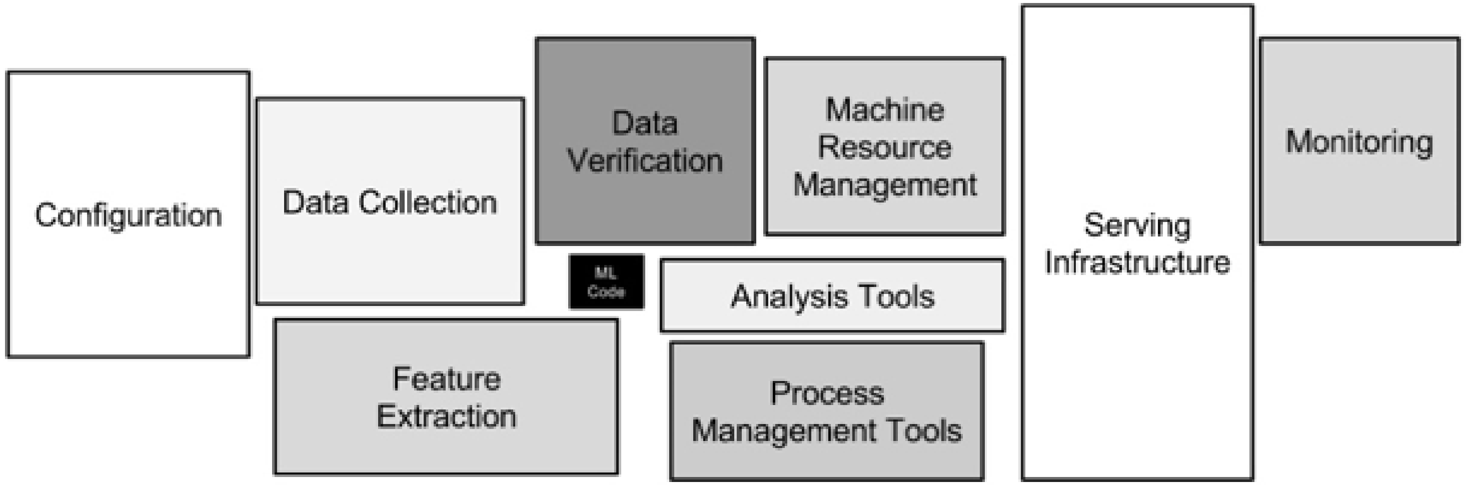


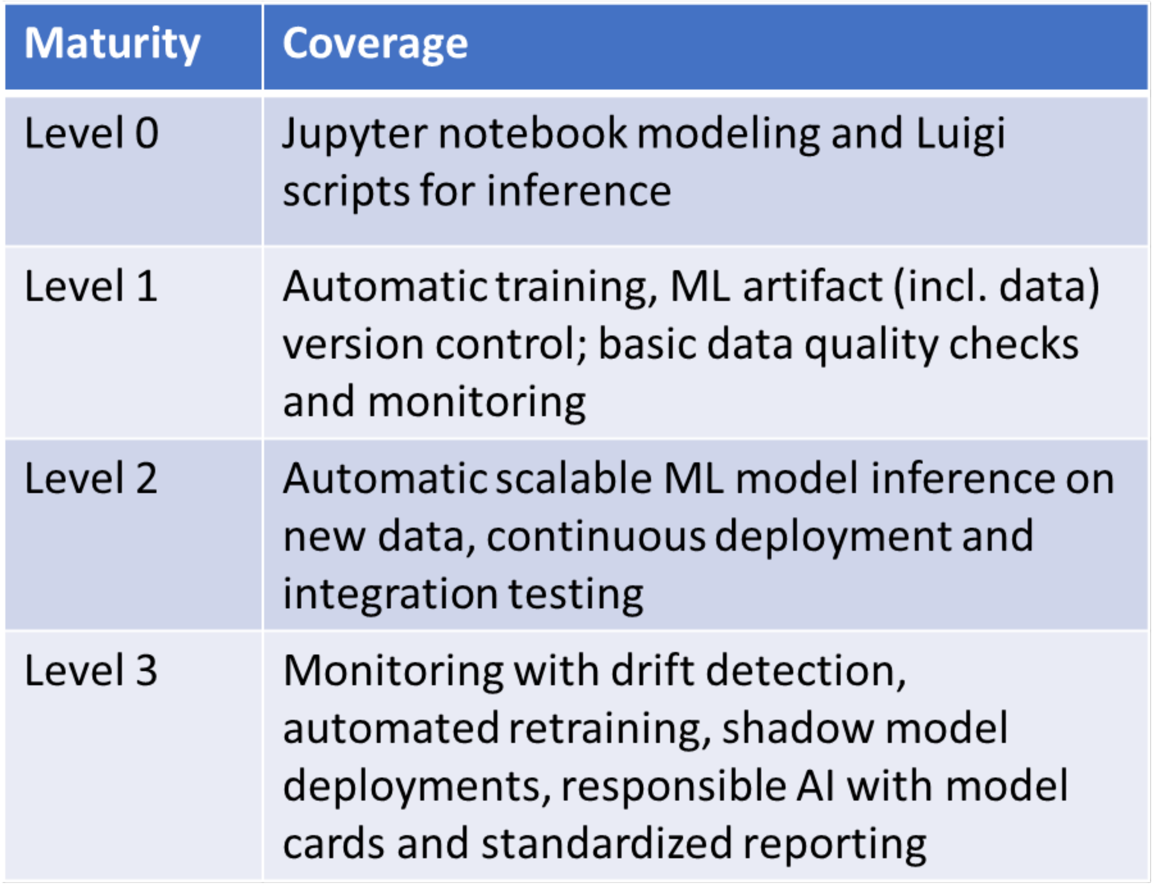
Figure 1: Only a small fraction of real-world ML systems is composed of ML code. The surrounding infrastructure is vast and complex. Reference: "Hidden Technical Debt in Machine Learning Systems", Sculley 2015

MLOps components

MLOps System Component	Description
Pipeline Orchestration & Engines	Execute ML training and model serving via pipelines on compute infrastructure
Version Control	Version control ML artifacts along the ML lifecycle
Data Quality & Drift	fine-grained data schema validation
Model Deployment & Serving	Enable batch, streaming or real-time inference with ML models
Monitoring	Track infrastructure, software, data and model quality, KPIs
Governance and Risk Mitigation	Tracking and Accountability
Responsible AI	Focus on ethical sound and unbiased application

MLOps maturity model for adoption

- Implementation of MLOps System in multiple iterations – separated into MLOps maturity levels
- Each maturity level brings benefits aligns with business requirements



Maturity	Coverage
Level 0	Jupyter notebook modeling and Luigi scripts for inference
Level 1	Automatic training, ML artifact (incl. data) version control; basic data quality checks and monitoring
Level 2	Automatic scalable ML model inference on new data, continuous deployment and integration testing
Level 3	Monitoring with drift detection, automated retraining, shadow model deployments, responsible AI with model cards and standardized reporting

Progress to date

- MLOPs for the CPI – path to adoption:
  - Statistics Canada initially applied ML in production to support the CPI at MLOps maturity level 0:
    - Weekly retailer files preprocessed, with new unique products packaged for an ML model Classification model trained on a manually versioned Jupyter Notebook
    - All new unique products quality assured, and used to retrain the model when necessary
    - ML microservice - classification model and outlier/misclassification detection rules - orchestrated as a Luigi pipeline run on Windows desktop
  - Since 2021, the CPI program has invested into foundational data and application architecture on the cloud to develop a modern, modular, and scalable production processing platform
  - Work continues to develop the first MLOps-focused production process at Statistics Canada, successfully addressed business needs

Lessons learned

- Lessons learned/challenges:
  - Investment in data science skills to design processes were needed, task was achievable but not trivial;
  - Infrastructure costs need to be managed as best practices for keeping costs low important;
  - MLOps Engineering role needs organizational change and inclusion of all parties, close work with IT.
  - Biggest road blockers are related to accessibility and the complexity of cloud networks

Summary

- CPI enhancements will continue as scale of alternate data adoption increases
- Work is ongoing to enable full use of alternative data to support the CPI and related price statistics needs, including through the use of multilateral price index methods
- The scale and robustness necessary to process alternative data is not possible without data science methods and tools

Stay connected!

StatsCAN app

Eh Sayers podcast

StatsCAN Plus

The Daily

Website

Surveys and statistical programs

Data service centres

My StatCan

Questions? Contact us: infostats@statcan.gc.ca