Responsible ML in Official Stats: Explainability & Uncertainty

Saeid Molladavoudi

Senior Data Science Advisor

Statistics Canada

(Joint work with Wesley Yung)

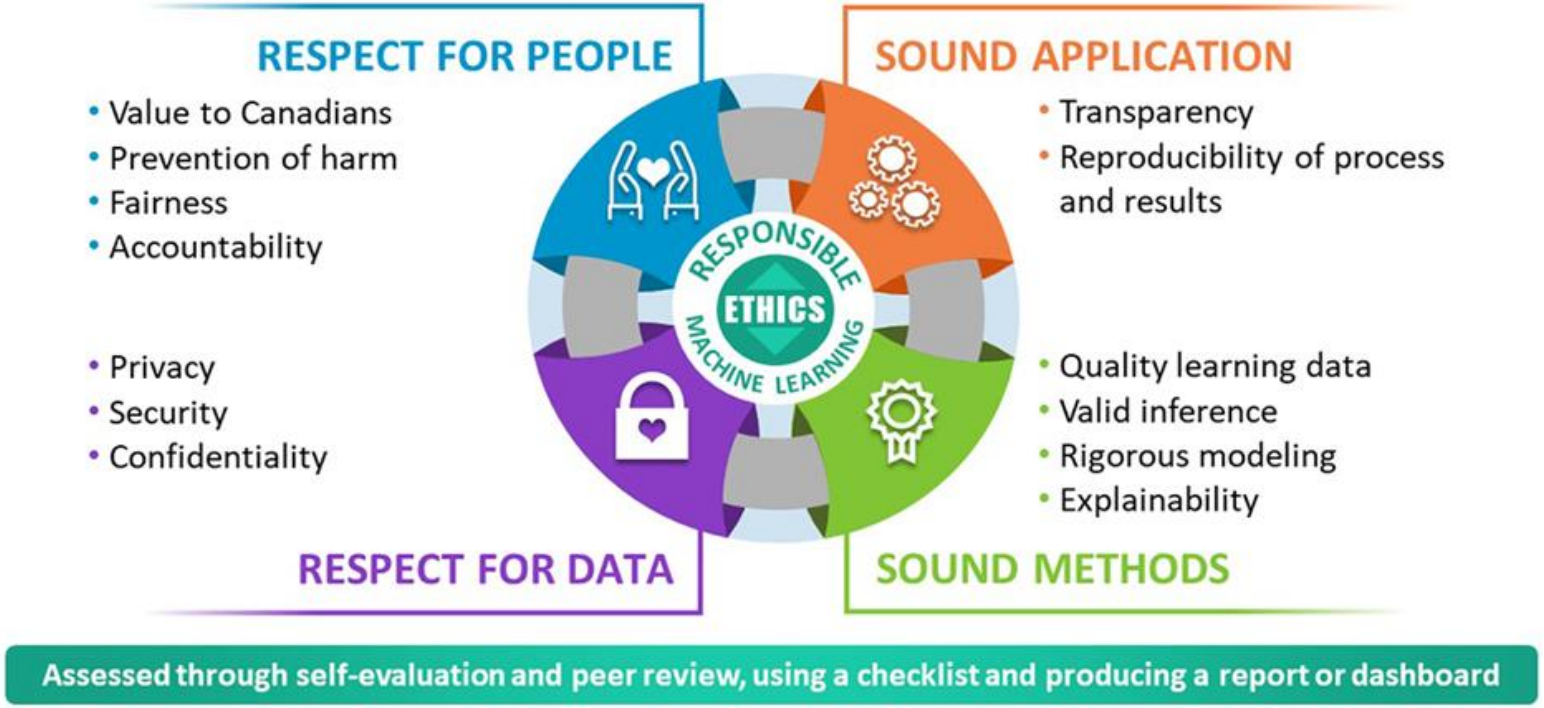
Delivering insight through data for a better Canada

Background

- Existing quality assurance frameworks developed before ML.
- Statistics Canada's quality guidelines define:
  - Accuracy, relevance, timeliness, accessibility, interpretability, and coherence (Statistics Canada, 2019).
- **QF4SA (2022)** proposed complementary quality dimensions:
  - Accuracy, explainability, reproducibility, timeliness, and cost-effectiveness
- **Responsible ML** covers some of these, and much more, e.g., fairness, ethics, accountability, robustness, privacy, etc.

Responsible ML for Official Statistics

- Statistics Canada's Framework for Responsible ML:



- International Framework on the Responsible AI for Official Statistics (HLG-MOS).

Transparency, Explainability & Interpretability

- **Transparency**: model, design, and algorithms (inductive biases),
- **Interpretability**: conformity of the 'knowledge' encoded in the model with human domain experts.
- **Explainability**: faithful secondary interpretable algorithms to extract insight about what a black box model has learned.
- **PDR Framework (Murdoch et al, 2019)**:
  - **Predictive accuracy**: model selection to address the problem at-hand,
  - **Descriptive accuracy**: description of the process to produce outcomes,
  - **Relevance**: judged relative to a human domain expert.

Review of Explainable ML Methods

- **Categories**: Global vs local, model-specific or model-agnostic methods
- **Local Interpretable Model-Agnostic Explanations (LIME):**

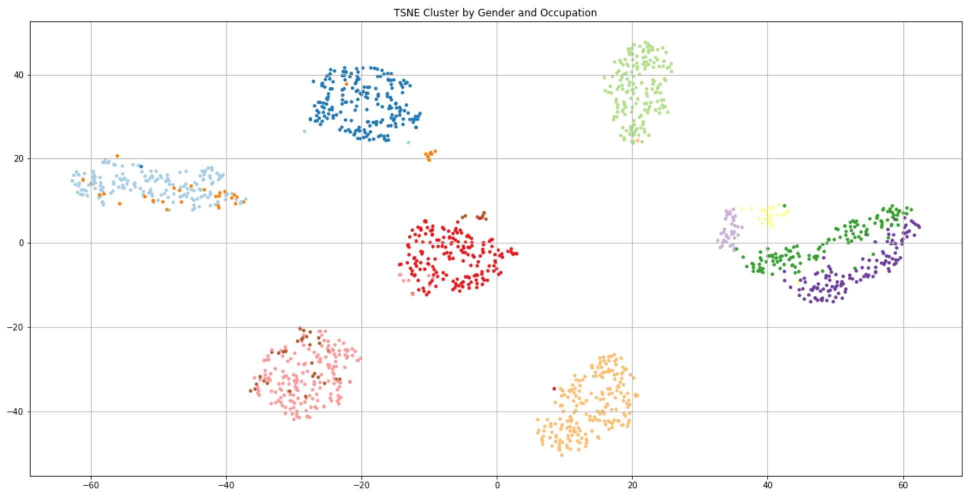
  Generates perturbed samples from the original dataset near the decision boundary.
- **Shapley Values and SHapley Additive exPlanations (SHAP):**

  Use cooperative game theory to explain feature importance (features represent 'players'!).

  SHAP: the Shapley values of a conditional expectation function of the original model.
- **Counterfactual explanations:** What would the adjustments in the feature values be in order to shift the prediction to a desired outcome?
- **Anchors:** generate local perturbations of instances with user-friendly if-then rules.

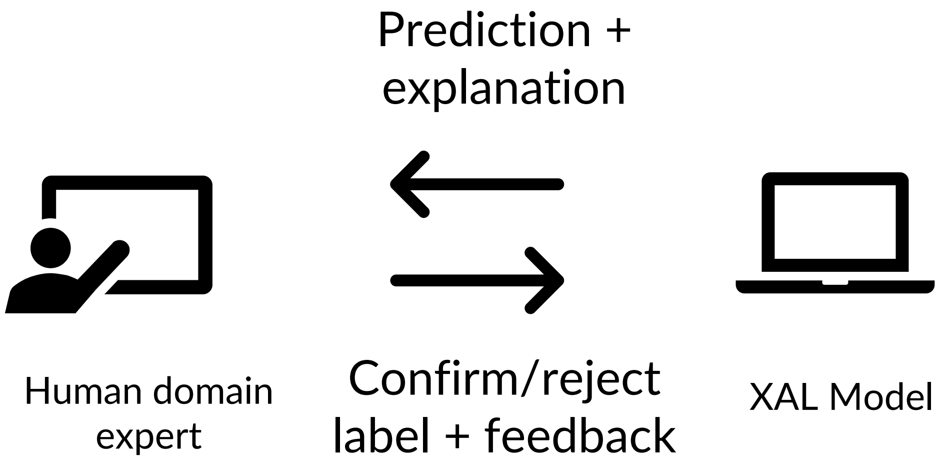
Applications of Explainable ML

(a) Understanding non-response mechanisms and sub-structures



‘Black box’ model + Local explainable ML + Visualization

(b) Continuous model monitoring (Explainable Active Learning - XAL)



Uncertainty in ML

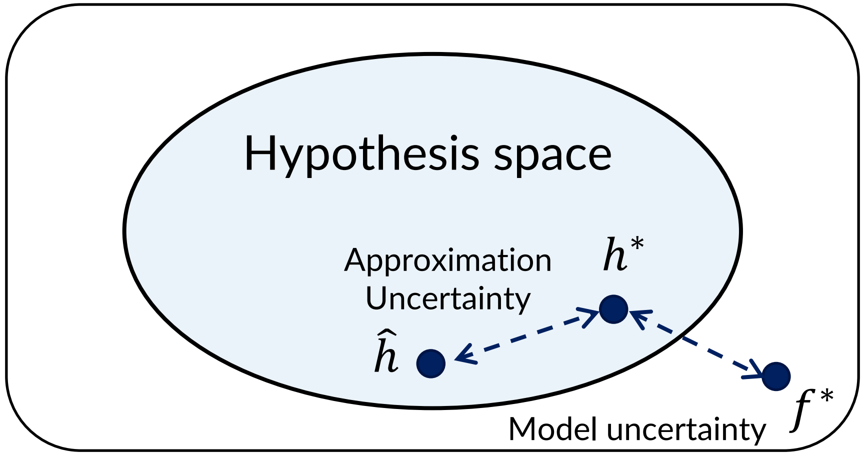
- Types of uncertainty in **statistical learning theory**: *aleatoric* vs *epistemic*

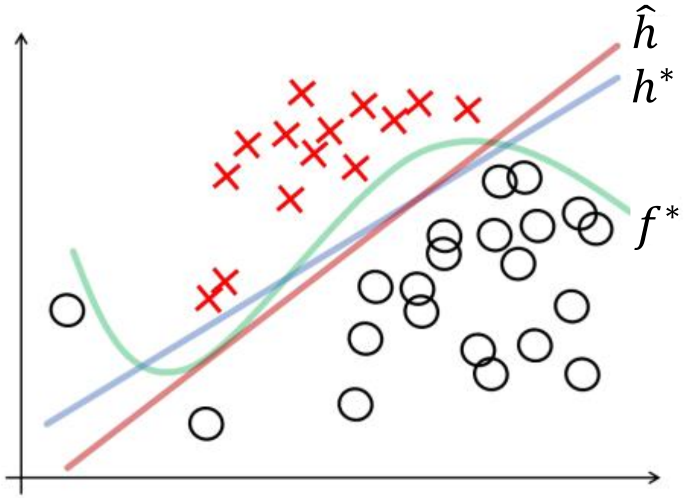
Training data (D) → Background Knowledge →

Model Induction
$\hat{h} = Ind(D)$

→ **Model $\hat{h}$**
Prediction: $\hat{y} = \hat{h}(\boldsymbol{x})$

← Test data ($\boldsymbol{x}$)





Quality Indicators

- Existing quality indicators, e.g., CV in a survey-based framework
- Current uncertainty quantification methods in ML (e.g. supervised learning):
  - **Bayesian methods** to approximate posterior distribution over model parameters $P(\boldsymbol{\theta}|D)$ and use for inference ($\boldsymbol{x}$):

$$P(y\,|\boldsymbol{x}\,,D) = \int P(y\,|\boldsymbol{x},\boldsymbol{\theta})\,P(\boldsymbol{\theta}|\,D)\,d\boldsymbol{\theta}$$

  - **Conformal prediction:** distribution-free prediction sets around any model type. It provides coverage guarantee and is based on data exchangeability. For a non-conformity score function, e.g., $r_i = |y_i - f(\boldsymbol{x}_i)|$, with $i \in$ hold-out dataset, threshold τ, and error rate α,

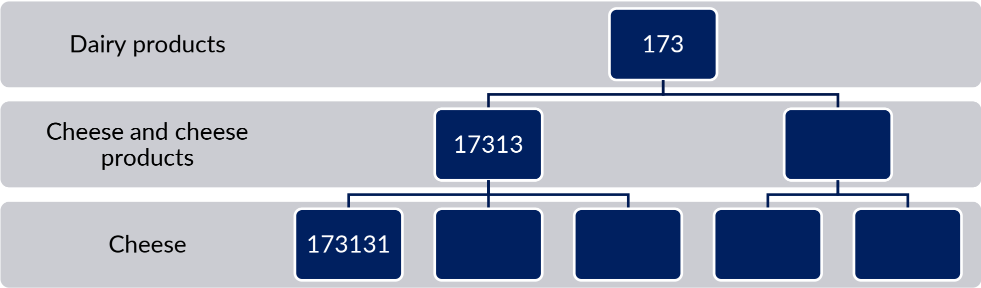
$$C(\boldsymbol{x}_{n+1}) = \{y\,|\,r_{n+1} \leq \tau\}, \qquad P\big(y_{n+1} \in C(\boldsymbol{x}_{n+1})\big) \geq 1 - \alpha$$

  - **Other methods:** Ensemble method, selective abstention, confidence calibration, etc.

Applications of Uncertainty Quantification

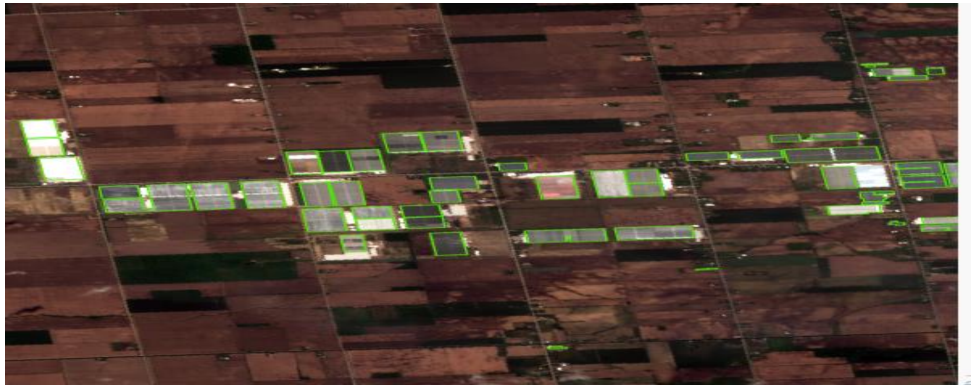
(a) Hierarchical text classification

- Industrial auto-coding is wide-spread.
- *Conformal risk control*, based on a geometric non-conformity cost function, i.e., costs based on semantic distance.
- Coarser/finer prediction sets, w.r.t. leaves.



(b) Image segmentation

- Pixel-wise classification (e.g., crop types).
- False negative rate, as the cost function to be controlled, at a user-specific rate.
- Provides distribution-free and finite-sample guarantees (data exchangeability).



Applications of Uncertainty Quantification

- **Prediction-powered ML (Jordan et al, 2023): Model-assisted survey estimation**
- Use a model $f: X \to Y$ to estimate population mean $\hat{\mu}_y$ of the response $y \in Y$ (Model Assisted Estimator - MAE):

$$\hat{\mu}_y = \frac{1}{N}\sum_{i \in U} f(\boldsymbol{x}_i) + \frac{1}{n}\sum_{i \in S} \frac{y_i - f(\boldsymbol{x}_i)}{\pi_i}$$

- Write as a constrained convex optimization $\mu_y = arg\, min_{\mu'} E[(y - \mu')^2]$.
- Form confidence intervals that covers the true value of μ_y, while making the interval tighter than the classical interval.
- This works well in the regime $n \ll N$, with provable asymptotic properties.

Conclusions

- There is more work to reconcile ML-based quality control with the existing quality assurance frameworks (e.g., QF4SA's complementary criteria?).
- There are interesting applications to be explored further with respect to **explainable ML** and **uncertainty quantification**, e.g., **(1) continuous model monitoring**, **(2) explainable active learning**, **(3) hierarchical text classification**, **(4) image segmentation**, and **(5) model-assisted survey estimation**.
- There are more reasons to consider these dimensions, such as upcoming regulations: **EU AI Act**, **Digital Services Act**, **AI and Data Act**, etc.
- We have a session in the **ISI WSC 2023:** RML in the context of Official Stats.

Let’s continue exploring the quality dimensions!

Thank you/Merci!

saeid.molladavoudi@statcan.gc.ca

wesley.yung@statcan.gc.ca

Le contenu de cette présentation représente la position de l’auteur, mais pas nécessairement celle de Statistique Canada.
The content of this presentation represents the position of the author and may not necessarily represent that of Statistics Canada.