Edit and Imputation With Machine Learning from PoC on LCF towards Implementation for SLC/HFS survey data

Claus Sthamer

Data Science Campus

2nd September 2020

Claus.Sthamer@ons.gov.uk

International Definitions:

Edit:
Identifying records that need values changed or missing values inserted

Imputation:
Changing and Inserting missing values

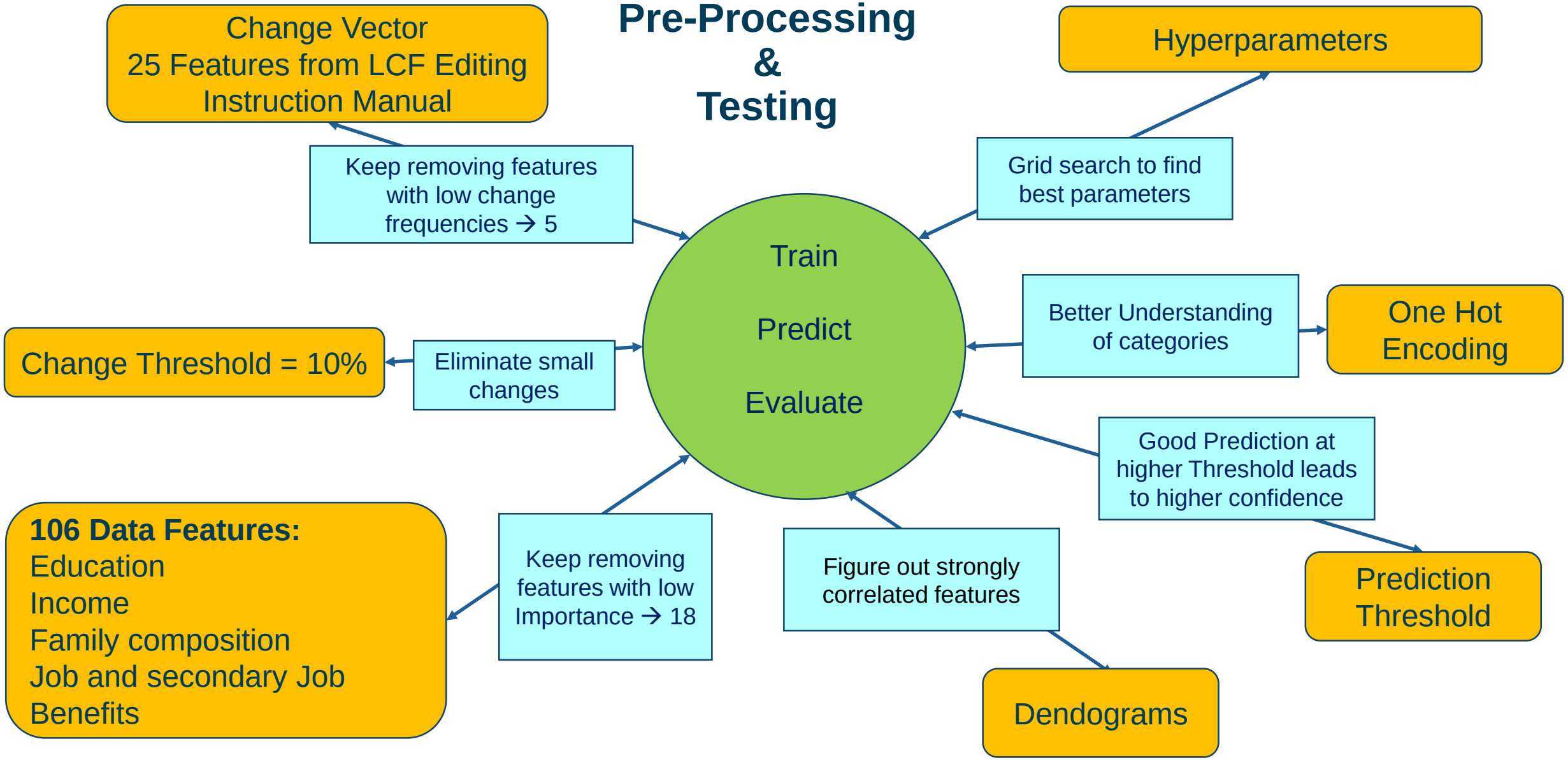
ONS Social Survey Definitions:

Edit:
Correct an incorrect value

Imputation:
Inserting missing values

Claus.Sthamer@ONS.gov.uk

Pre-Processing & Testing



Claus.Sthamer@ONS.gov.uk

Results

Change vector with number of changes:

NetPay	-Last take home pay including	81
IncTax	-Income Tax	235
NIns	-National Insurance Contribution	254
GrossPay	-Last gross pay from main job	336
DedPenAm	-Deduction for pension or superannuation	113

Training Data: 8Q3 3059 Person records, 442 labelled as Change, reduced to 362 with 10% change threshold
Test Data: 8Q2 2912 Person records, 451 labelled as Change , reduced to 361 with 10% change threshold

Features:

106 features reduced to 18 by removing features with low Importance

Prediction Threshold:	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**
Recall	97.2%	94.6%	88.1%	84.0%	78.6	75.2	72.9
Precision	35.7%	40.0%	45.4%	54.5%	67.1	81.3	90.1
F1-Score	52.2%	56.2%	59.1%	66.1%	72.4	78.1	80.6
TP	376	366	341	325	304	291	282
FP	677	550	410	271	149	67	31

Claus.Sthamer@ONS.gov.uk

The Journey so far:

UNECE ML Project
To improve:

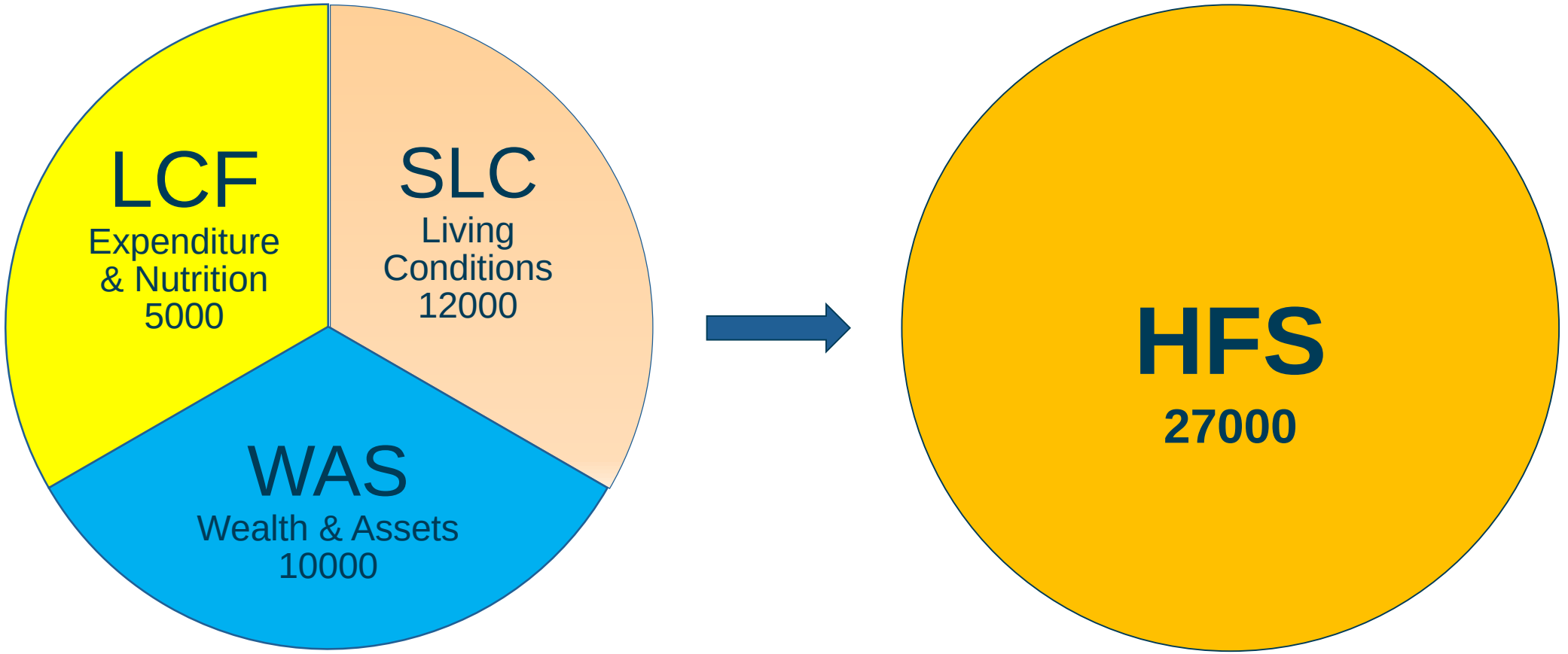
- Time lines
- Accuracy
- Cost

Editing of LCF with ML POC

- ✓ Faster
- ✓
- ✓ 75% Savings

Claus.Sthamer@ONS.gov.uk

The HFS and it's component Surveys



Number of co-operating Households

Claus.Sthamer@ONS.gov.uk

Survey Specific Editing (Currently in Production)

LCF – all cases go through clerical editing

too slow and too labour intensive → **Speed &Cost?** **Over Editing?**

SLC – Scripted outlier detection (range of values)

Only about 10% of changes that are made with the LCF method are made with the SLC method → **Accuracy?**

WAS – Scripted outlier detection (range of values)

Accuracy?

ML

Claus.Sthamer@ONS.gov.uk

Focus has shifted to SLC survey data

- Income Statistics for LCF & SLC differ
- Possible reasons:
  - Different Edit processes?
  - Different Imputation?
  - SLC is Longitudinal – older respondence tend to stick with it

- Which Editing process should be adapted for all 3 surveys?

Claus.Sthamer@ONS.gov.uk

What are the Challenges?

1. Business Knowledge
   - Stakeholders - Who can make decisions?
   - Data usage - Impact on data users
   - Existing Data Pipeline - Integration
2. Knowing the Data
   - Precision ← → Recall : to satisfy business needs
3. Value Proposition, what is the ML value?
   - Speed?
   - Cost?
   - Accuracy?

} What is the driver for all this?

Claus.Sthamer@ONS.gov.uk

Strategic Engagement Outline

- ‘deep dive’ discussions about infrastructure, survey pipelines, etc.

- apply an evaluation framework to a ML PoC designed to understand both its strengths and the challenges it faces in progressing towards operationalisation.

- recommendations for addressing these challenges in the form of short and medium-term roadmaps.

Claus.Sthamer@ONS.gov.uk

Short term roadmap

1. Extend the predictive capability of the model.
2. Relax restrictions on development on legacy infrastructure
3. Develop a plan for collecting baseline data for model evaluation purposes : high workload?
4. Establish thresholds for accuracy-related metrics
5. Test model with WAS and SLC datasets

Claus.Sthamer@ONS.gov.uk

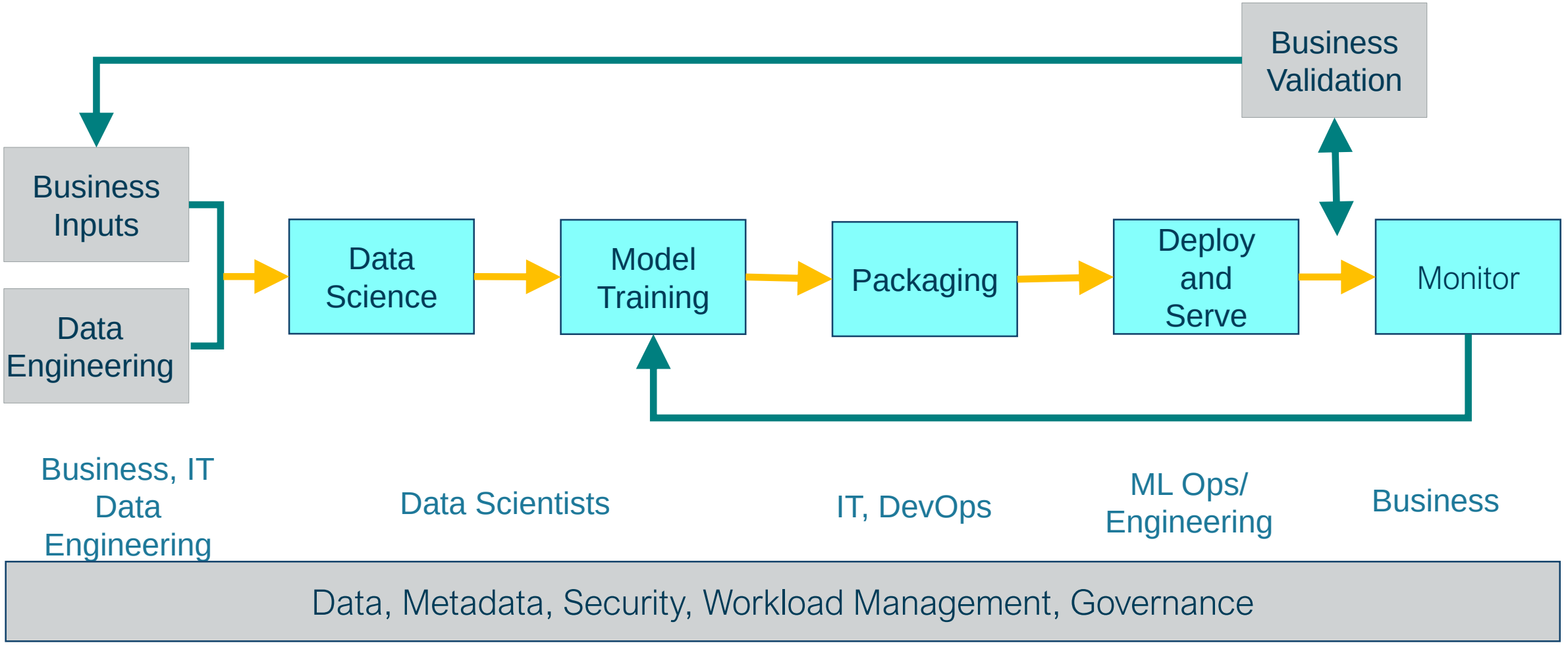
Medium term roadmap

1. Develop protocol for implementing effective ongoing model monitoring and maintenance.
2. Establish protocol for off- and on-lining the editing model when necessary.
3. Incorporate an interpretability framework and explainability approach

Claus.Sthamer@ONS.gov.uk

End-to-end Machine Learning in an Enterprise setting

Production ML is an iterative team sport



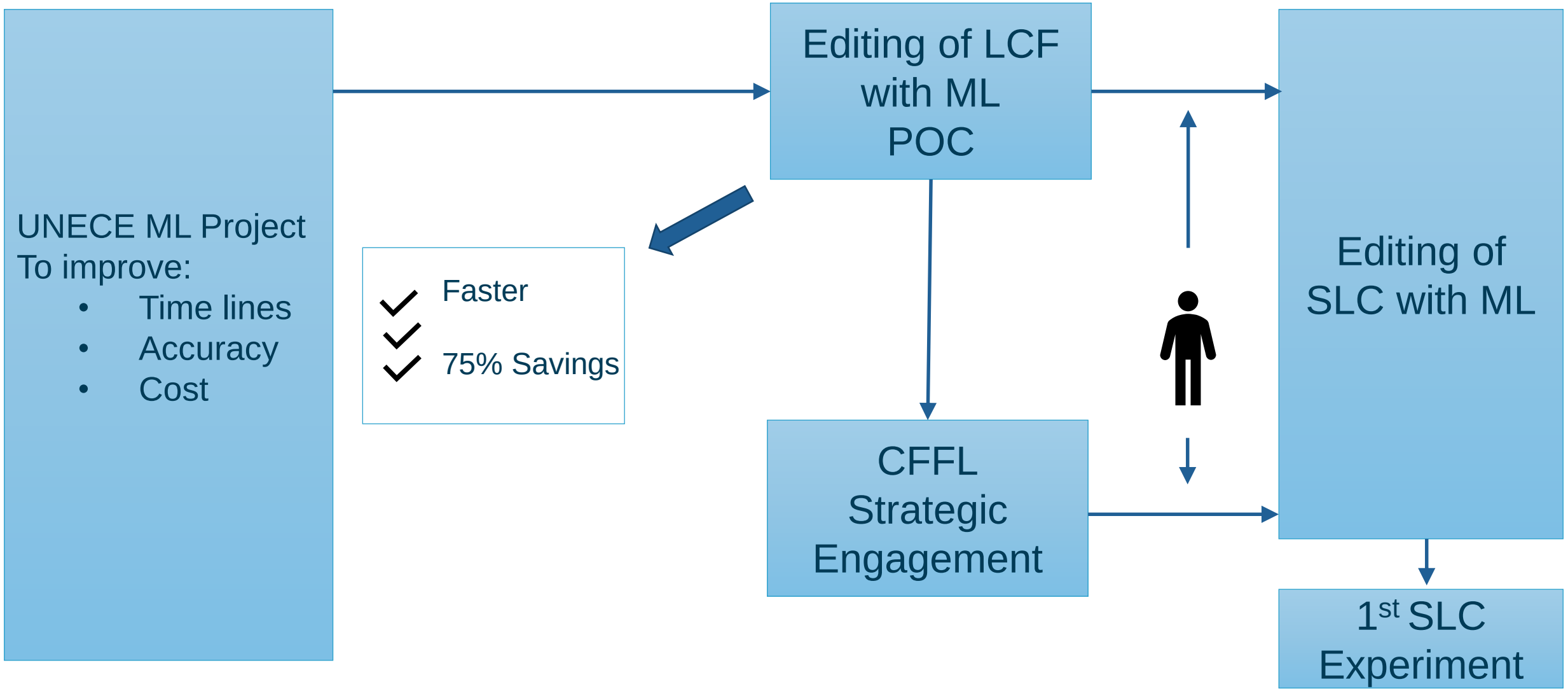
Claus.Sthamer@ONS.gov.uk

Other the Challenges?

- Technology
  - DAP
  - IDP
- Google, AWS, Azure
- Who will do what?
- Ethics & Governance

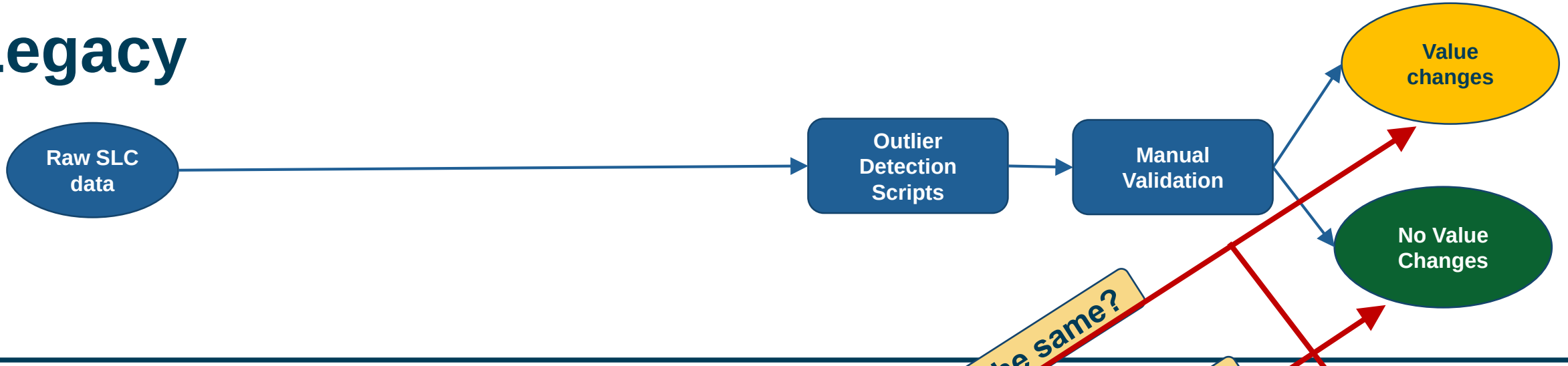
Claus.Sthamer@ONS.gov.uk

The Journey so far:

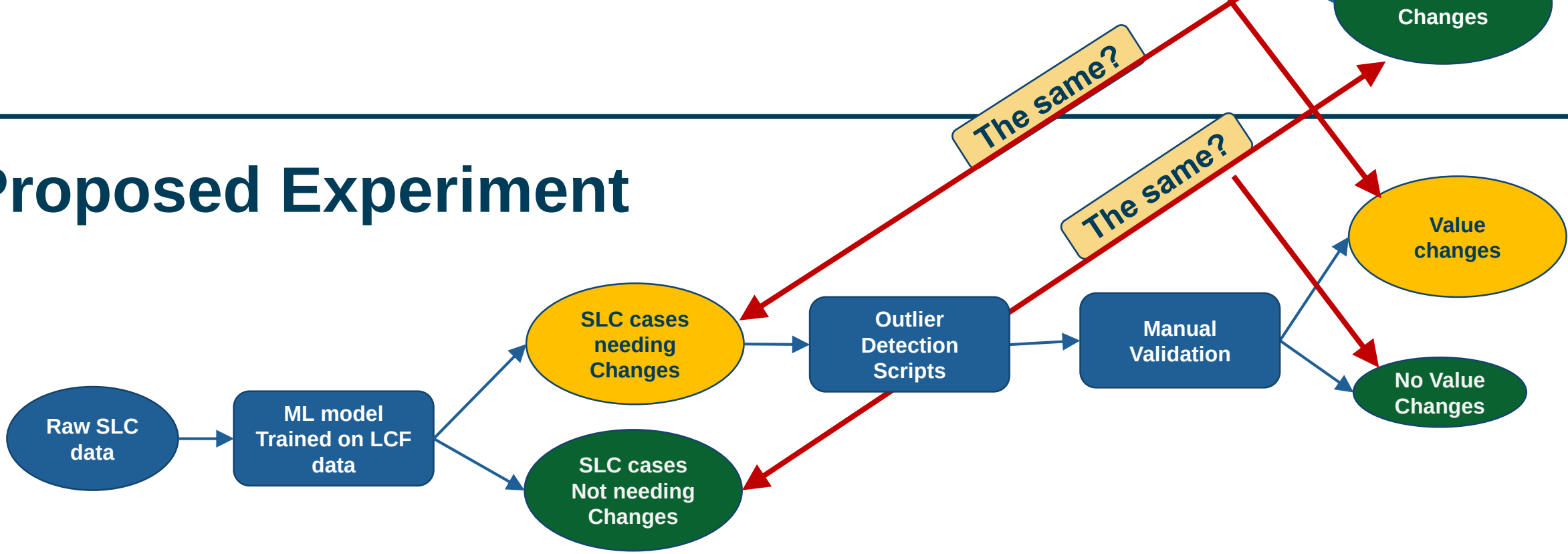


Claus.Sthamer@ONS.gov.uk

Legacy



Proposed Experiment



Claus.Sthamer@ONS.gov.uk

Office for National Statistics

Thank you

Claus Sthamer

Data Science Campus

2nd September 2020

Claus.Sthamer@ons.gov.uk