Data Science Campus

Machine Learning for Social Surveys LCF/SLC/HFS

Claus Sthamer
Technical Project Lead

26th January 2022

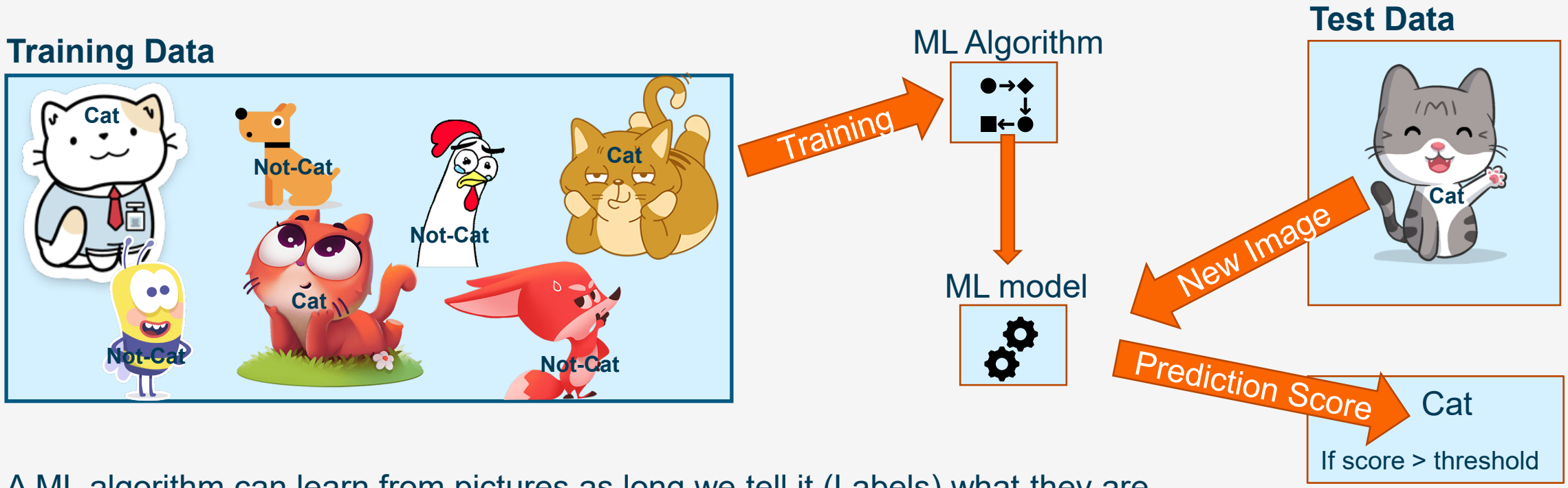
Data Science Campus

What is Machine Learning?

Machine Learning is used to identify rules and patterns in data humans can not.

Example: How to recognise a cat

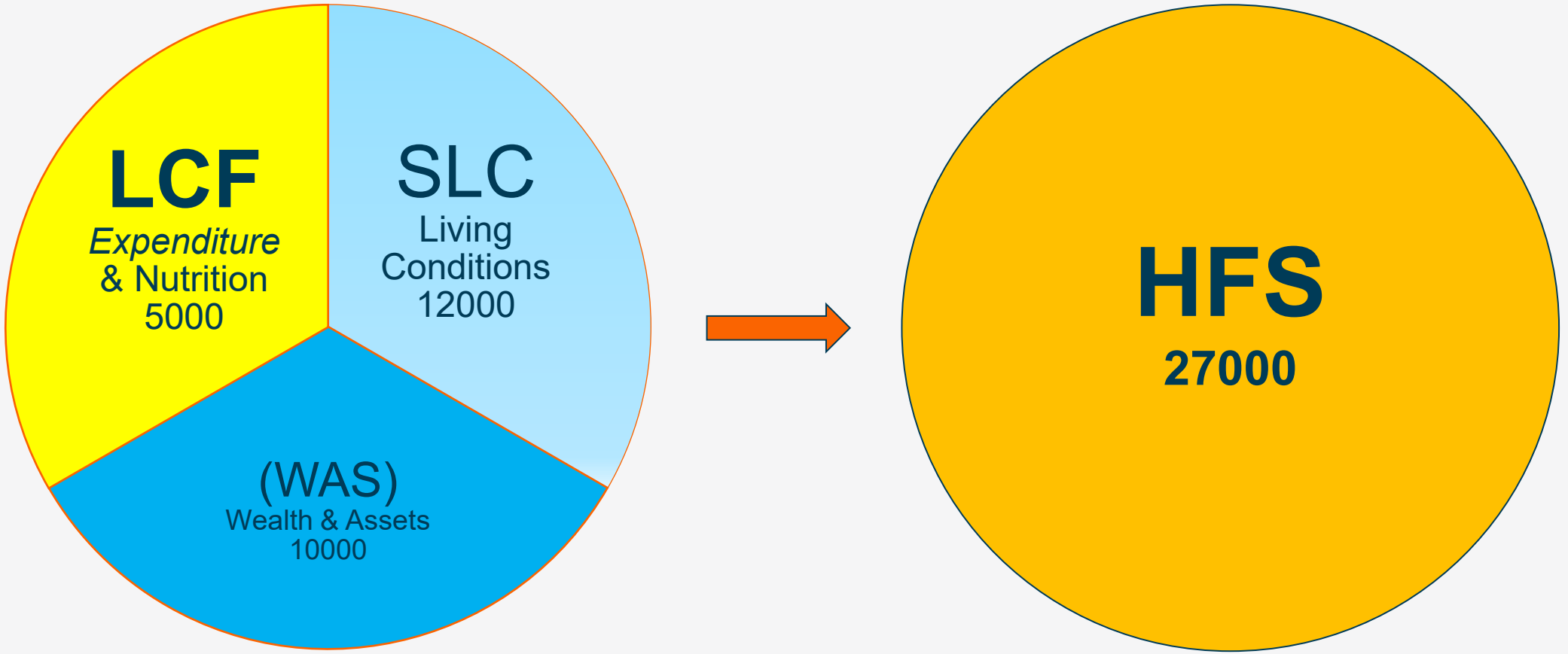
But not just one cat but all cats. What are the rules?



A ML algorithm can learn from pictures as long we tell it (Labels) what they are

ML can make an inference of the class of new pictures, it gives a score for the most likely class

The HFS and it's component Surveys



Number of co-operating Households each year

Survey Specific Editing (Currently in Production)

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

- 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 50% (?) 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

Editing of LCF income data with ML

LCF Editing Instructions for Income → extensive manual Editing → Ground Truth

5 most often changed Independent Variables:

- Net income (after deductions)
- Income Tax
- Gross pay (before deductions)
- National Insurance paid
- Deduction for pension

Any change of more than 10% is counted as a change → Label/Target/Dependant

Two Class Classification Problem:

- No-Change
- Change

Supervised Learning

Training data → Train the algorithm
Test data → Test the trained model

Random Forest ML algorithm

Results of 2912 LCF Test Cases – What is good enough?

Battle between Recall and Precision, they can't both be 100%

Prediction Threshold	20%	25%	30%	35%	40%	45%	50%
Recall	95.7%	94.0%	91.8%	88.9%	84.2%	81.0%	77.4%
Precision	37.0%	41.3%	47.5%	55.1%	61.8%	69.8%	77.4%
F1-Score	53.3%	57.4%	62.6%	68.0%	71.3%	75.0%	77.4%
TP	352	346	338	**327**	310	298	285
FP	600	491	374	**267**	192	129	83

Exp v62.1

327/368 = 88.9%

If prediction score > Threshold → Case belongs to the Change Class

Objective:

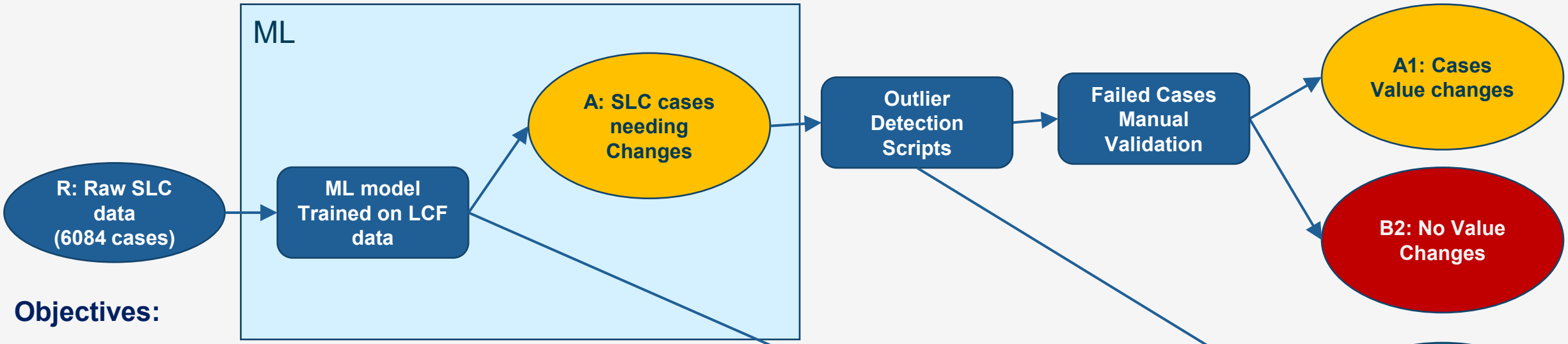
1. Find as many True Positive (TP) as possible with small number of False Positives (FP) - 88.9% ☑
2. Find all cases with large value changes > 500% - 100% ☑
3. Reduce the number of cases to be manually analysed (from 3000 to about 600) - 80% ☑

The same team could check 5 x the number of cases for inconsistent data → HFS

Then we looked at SLC survey data

- SLC also collects Income data, but uses a deterministic editing process
- No SLC Ground Truth, but we know which cases have been changed (Labels)
- Transfer Learning – Can we use the trained LCF ML model?
- Objective:
  - Reduce: number of cases flagged by the deterministic SLC process, but do not receive a value change
  - Keep all cases with Changes
  - Keep all cases with large value changes

Proposed SLC Editing Pipeline



Objectives:

- Retain the SLC Editing method
- Add ML as a filter to remove cases that:
  Are flagged by the scripts, but do not need any changes → 'Wasted Effort'
- Reducing this Burden is key to use the SLC Editing method for HFS
- Retain cases that
  - need changes
  - need large value changes
- A longer term objective is to have only one Editing process

Results of 6084 SLC Cases – What is good enough?

Battle between Recall and Precision, they can't both be 100%

SLC Prediction Results:

Prediction Threshold	20%	25%	30%	35%	40%	45%	50%
Recall	99.7%	99.7%	99.3%	98.4%	97.4%	95.2%	92.6%
Precision	34.5%	35.8%	38.7%	43.7%	52.2%	60.9%	69.4%
F1-Score	51.2%	52.7%	55.7%	60.6%	67.9%	74.4%	79.3%
TP	753	753	750	743	734	**719**	699
FP	1432	1351	1188	956	674	**462**	308
Burden cases filtered out				**76%**		**82.6%**	**85.3%**
				736 of 968		**800 of 968**	**826 of 968**

Exp v65.1

If prediction score > Threshold → Case belongs to the Change Class

Objective:

1. Find as many True Positive (TP) as possible with small number of False Positives (FP) - 95.2% ☑
2. Find all cases with large value changes >300% - 100% ☑
3. Reduce the number of cases manually checked, without receiving value changes (240 hours annually) - 82.6% ☑

The same team could check 5 x the number of cases for inconsistent data → HFS

Challenges

Results show that ML can make data Editing more efficient
Questions:

- What is good enough?
- What is it we want to achieve?
- How do we put this into production?
- How will ML fit into the data pipeline?

What about:

- Training Data
- Model/Data Drift
- Ethics?
- Explainability?

The Research and Implementation ML Workshop Part 1 will look at these in more detail

Technical ML sessions will look in more details how ML works and how to do it.

Thank you