

Analyzing Measurement and Representational Errors in Smart Survey Datasets used for Machine Learning

Machine Learning and Survey Methodology

Chris Lam, Marco Puts, Jonas Klingwort, Tim de Jong, Vera Toepoel

September 18, 2024

Introduction

Chris Lam

- Background in Computer Science
- Work for Statistics Netherlands (Dutch abbr. CBS)
- Methodologist / Data Scientist
 - At the Department of Methodology



Introduction

- Machine Learning (ML) in (Smart) Surveys
- ML in Survey Methodology
 - How is ML applied in surveys?
 - What effects does ML have on the statistics?



Machine Learning in Surveys

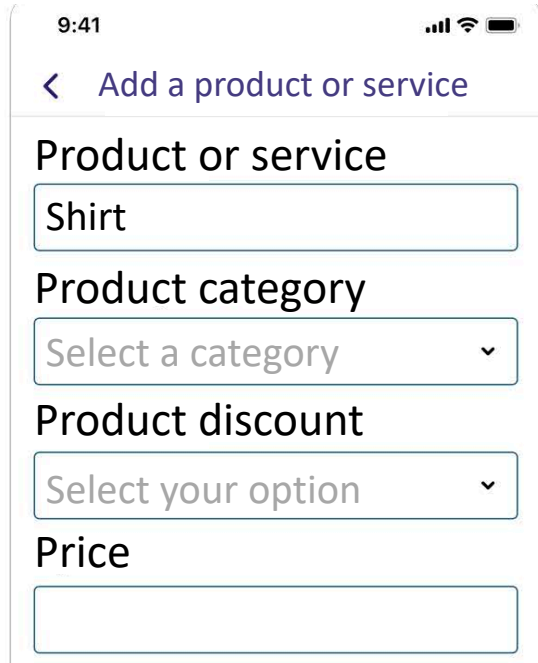


Machine Learning in Surveys: Household budget Survey

In Smart Surveys, ML can automate parts of the response process -> Lowering Response burden



Machine Learning in Surveys: Household budget Survey



9:41

< Add a product or service

Product or service

Shirt

Product category

Select a category ▾

Product discount

Select your option ▾

Price

In Smart Surveys, ML can automate parts of the response process -> Lowering Response burden



Machine Learning in Surveys: Household budget Survey

9:41

< Add a product or service

Product or service

Shirt

Product category

Select a category ▾

Product discount

Select your option ▾

Price



In Smart Surveys, ML can automate parts of the response process -> Lowering Response burden



Machine Learning in Surveys: Household budget Survey



15:52

Add Receipt Ready

Store
Walmart
Wholefoods

Date
Today

Total Price
\$6.00

Product Category
Foods

14:27

Add spending Ready

Store
Walmart

Date
Today

Others
Abroad Online

1x	Eggs	1.95
1.95	Milk, dairy, eggs	1.95
1x	Bread	2.70
2.7	Bread and bakery ...	2.70

In Smart Surveys, ML can be used to automate parts of the response process -> Lowering Response burden



Machine Learning in Surveys: Household budget Survey



```
( 330 ) 339 - 3991
MANAGER DIANA EARNEST
231 BLUEBELL DR SW
NEW PHILADELPHIA OH 44663
ST# 02115 OP# 009044 TE# 44 TR# 01301
PET TOY 004747571658 1.97 X
FLOPPY PUPPY 004747514846 1.97 X
SSSUPREME S 070060332153 4.97 X
2.5 SQUEAK 084699803238 5.92 X
MUNCHY DMBEL 068113108796 3.77 X
DOG TREAT 007119013654 2.92 X
PED PCH 1 002310011802 0.50 X
PED PCH 1 002310011802 0.50 X
COUPON 23100 052310037000 1.00-O
HNYMD SMORES 088491226837 F 3.98 O
FRENCH DRNG 004132100655 F 1.98 O
3 ORANGES 001466835001 F 5.47 N
BABY CARROTS 003338366602 I 1.48 N
COLLARDS 000000004614KI 1.24 N
CALZONE 005208362080 F 2.50 O
MM RVW MNT 003399105848 19.77 X
STKOBRLPLABL 001558679414 1.97 X
STKOBRLPLABL 001558679414 1.97 X
STKO SUNFLWR 001558679410 0.97 X
STKO SUNFLWR 001558679410 0.97 X
STKO SUNFLWR 001558679410 0.97 X
STKO SUNFLWR 001558679410 0.97 X
BLING BEADS 076594060699 0.97 X
GREAT VALUE 007874203191 F 9.97 O
L.TPTON 001200011224 F 4.48 X
```

Smart Survey Mode: Receipt Text digitization
& automatic product categorization



Machine Learning in Surveys: Household budget Survey

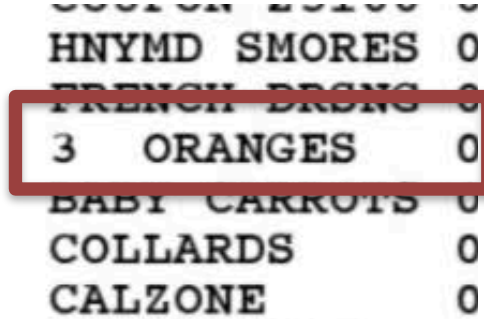


(330) 339 - 3991
MANAGER DIANA EARNEST
231 BLUEBELL DR SW
NEW PHILADELPHIA OH 44663

ST# 02115	OP# 009044	TE# 44	TR# 01301
PET TOY	004747571658		1.97 X
FLOPPY PUPPY	004747514846		1.97 X
SSSUPREME S	070060332153		4.97 X
2.5 SQUEAK	084699803238		5.92 X
MUNCHY DMBEL	068113108796		3.77 X
DOG TREAT	007119013654		2.92 X
PED PCH 1	002310011802		0.50 X
PED PCH 1	002310011802		0.50 X
COUPON 23100	052310037000		1.00-0
HNYMD SMORES	088491226837	F	3.98 0
FRENCH DRNG	004132100655	F	1.98 0
3 ORANGES	001466835001	F	5.47 N
BABY CARROTS	003338366602	I	1.48 N
COLLARDS	000000004614KI		1.24 N
CALZONE	005208362080	F	2.50 0
MM RVW MNT	003399105848		19.77 X
STKOBRLPLABL	001558679414		1.97 X
STKOBRLPLABL	001558679414		1.97 X
STKO SUNFLWR	001558679410		0.97 X
STKO SUNFLWR	001558679410		0.97 X
STKO SUNFLWR	001558679410		0.97 X
STKO SUNFLWR	001558679410		0.97 X
BLING BEADS	076594060699		0.97 X
GREAT VALUE	007874203191	F	9.97 0
L.T.PTON	001200011224	F	4.48 X



Text
recognition



Smart Survey Mode: Receipt Text digitization
& automatic product categorization



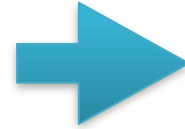
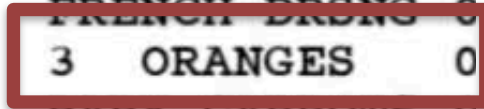
Machine Learning in Surveys: Household budget Survey



(330) 339 - 3991
MANAGER DIANA EARNEST
231 BLUEBELL DR SW
NEW PHILADELPHIA OH 44663
ST# 02115 OP# 009044 TE# 44 TR# 01301
PET TOY 004747571658 1.97 X
FLOPPY PUPPY 004747514846 1.97 X
SSSUPREME S 070060332153 4.97 X
2.5 SQUEAK 084699803238 5.92 X
MUNCHY DMBEL 068113108796 3.77 X
DOG TREAT 007119013654 2.92 X
PED PCH 1 002310011802 0.50 X
PED PCH 1 002310011802 0.50 X
COUPON 23100 052310037000 1.00-0
HNYMD SMORES 088491226837 F 3.98 0
FRENCH DRNG 004132100655 F 1.98 0
3 ORANGES 001466835001 F 5.47 N
BABY CARROTS 003338366602 I 1.48 N
COLLARDS 000000004614KI 1.24 N
CALZONE 005208362080 F 2.50 0
MM RVW MNT 003399105848 19.77 X
STKOBRLPLABL 001558679414 1.97 X
STKOBRLPLABL 001558679414 1.97 X
STKO SUNFLWR 001558679410 0.97 X
STKO SUNFLWR 001558679410 0.97 X
STKO SUNFLWR 001558679410 0.97 X
STKO SUNFLWR 001558679410 0.97 X
BLING BEADS 076594060699 0.97 X
GREAT VALUE 007874203191 F 9.97 0
L.T.PTON 001200011224 F 4.48 X



Text
recognition



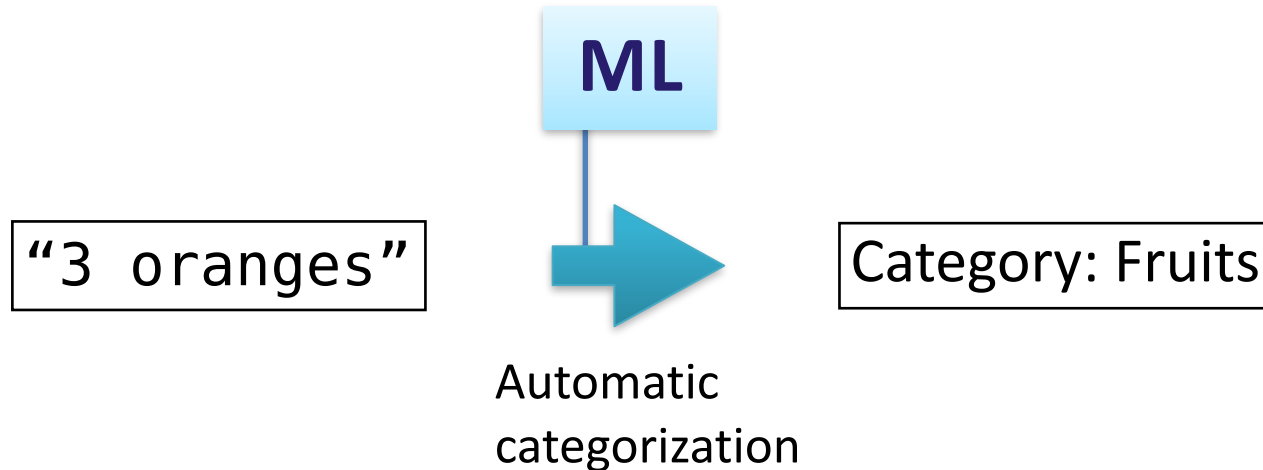
Text
Digitization

"3 oranges"

Smart Survey Mode: Receipt Text digitization
& automatic product categorization



Machine Learning in Surveys: Response modes



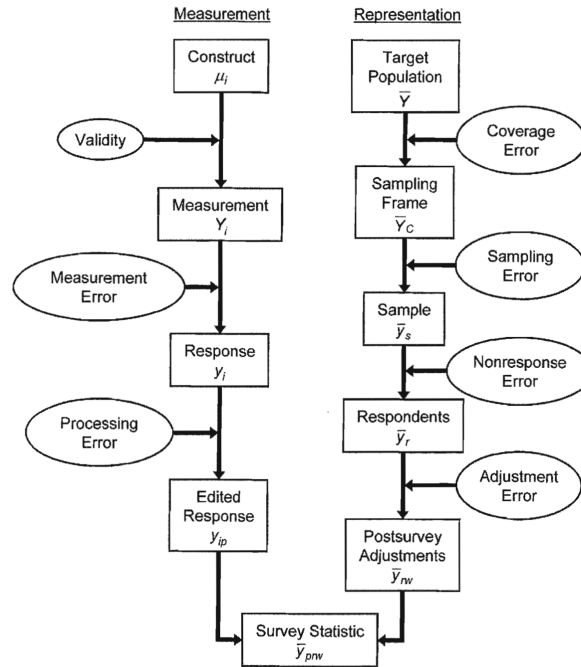
Smart Survey Mode: Receipt Text digitization
& automatic product categorization



Machine Learning in Survey Methodology



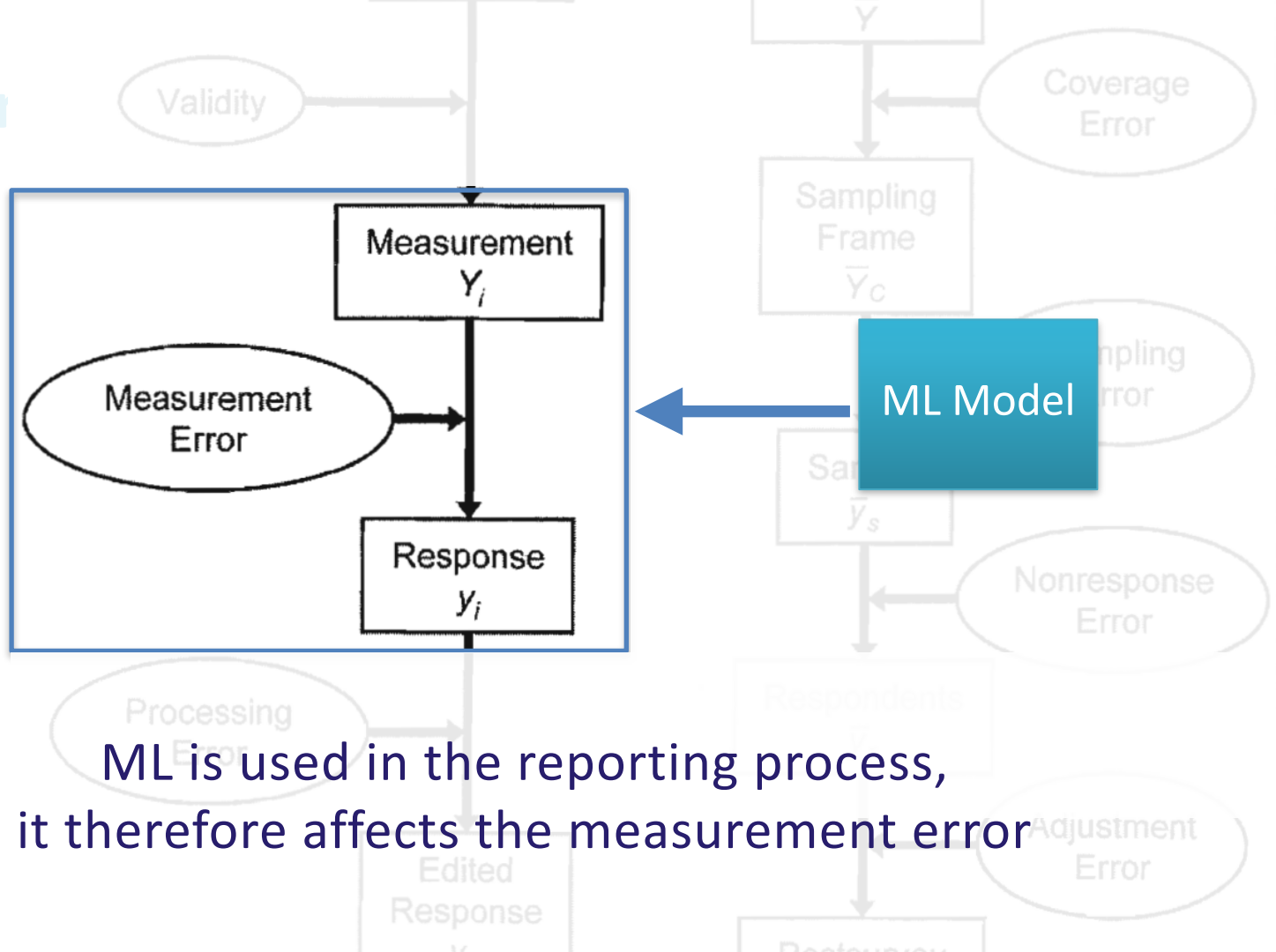
Machine Learning in Survey Methodology



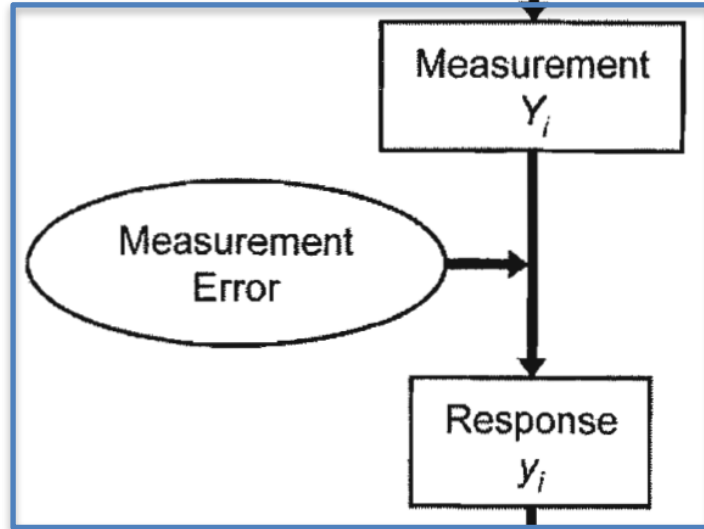
Groves et al. (2004)



How does ML affect the Total Survey Error framework?



ML is used in the reporting process,
it therefore affects the measurement error



Is there a way to evaluate the errors from the ML model?

ML Model

ML is used in the reporting process, it therefore affects the measurement error

Errors in Machine Learning



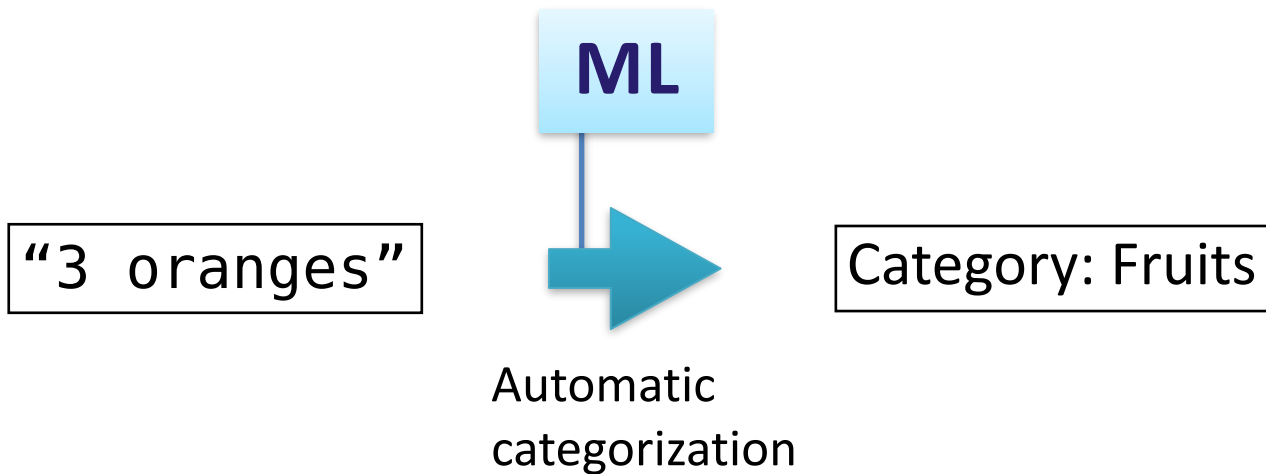
Errors in Machine Learning

- Let's look at an example:
 - Develop an ML-model that can determine the category of a product using ML



Errors in Machine Learning

- Let's look at an example:
 - Develop an ML-model that can determine the category of a product using ML



Errors in Machine Learning

Training

Application



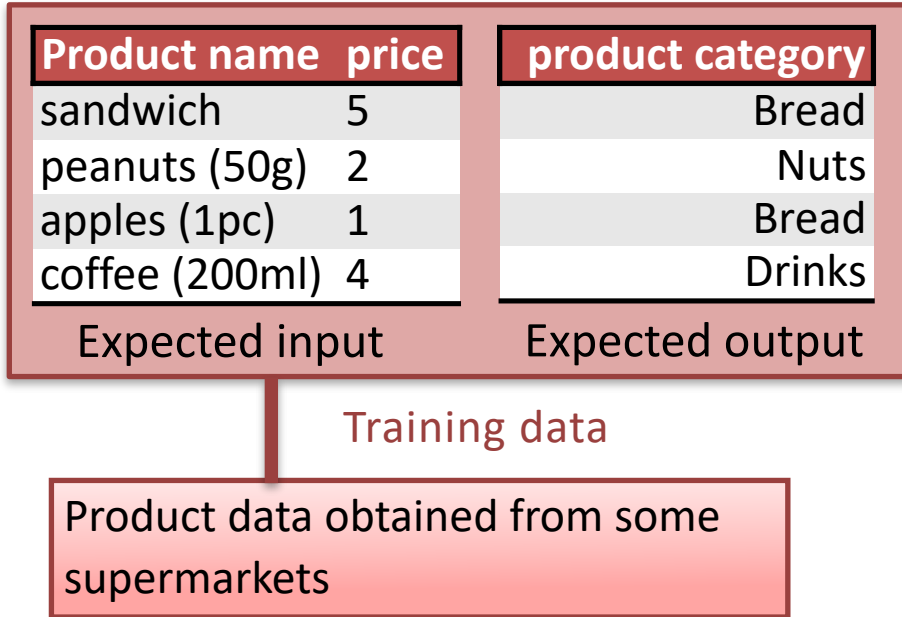
Errors in Machine Learning

Product name	price	product category
sandwich	5	Bread
peanuts (50g)	2	Nuts
apples (1pc)	1	Bread
coffee (200ml)	4	Drinks

Expected input Expected output

Training data

Errors in Machine Learning



Errors in Machine Learning

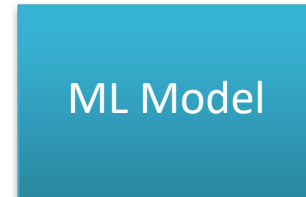
Product name	price	product category
sandwich	5	Bread
peanuts (50g)	2	Nuts
apples (1pc)	1	Bread
coffee (200ml)	4	Drinks

Expected input Expected output

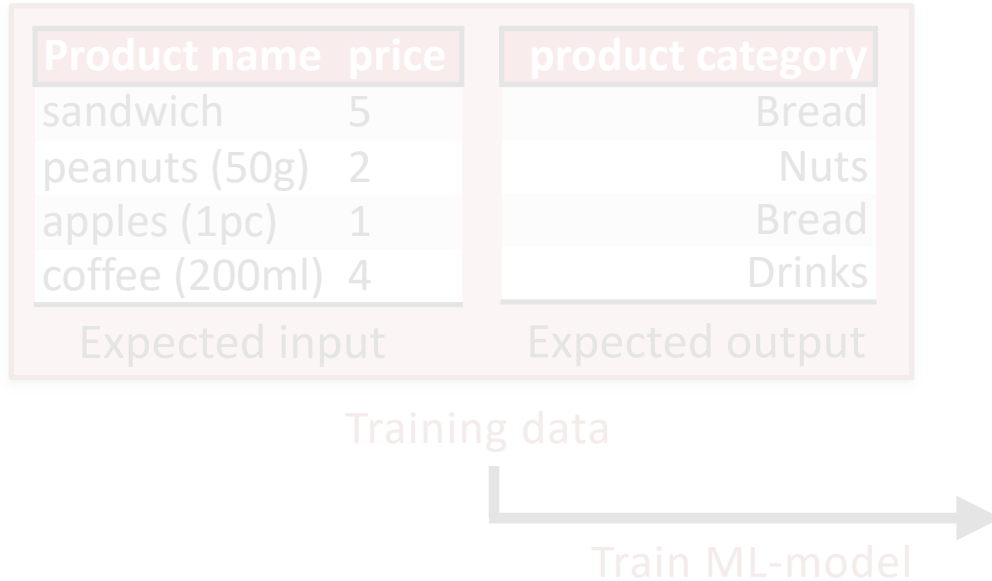
Training data



Train ML-model

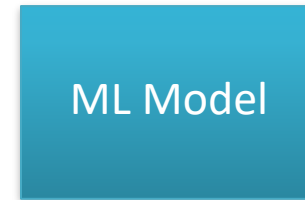


Errors in Machine Learning



Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

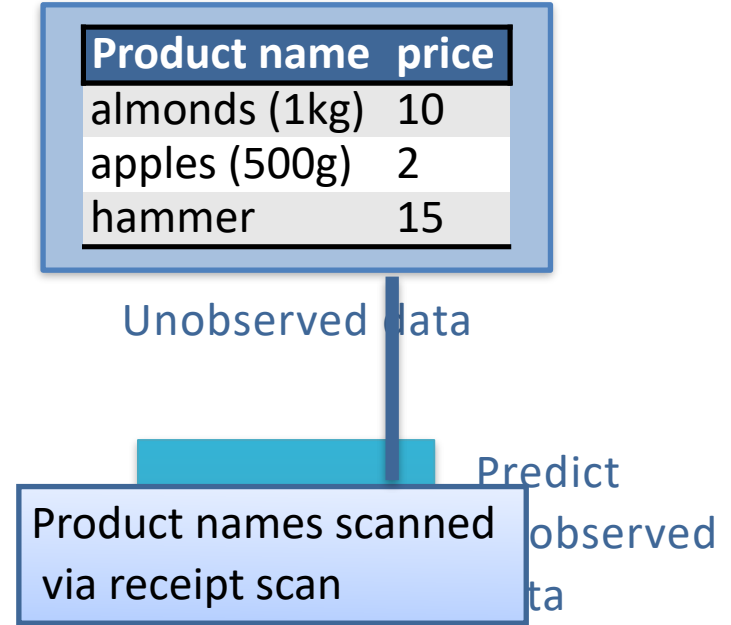
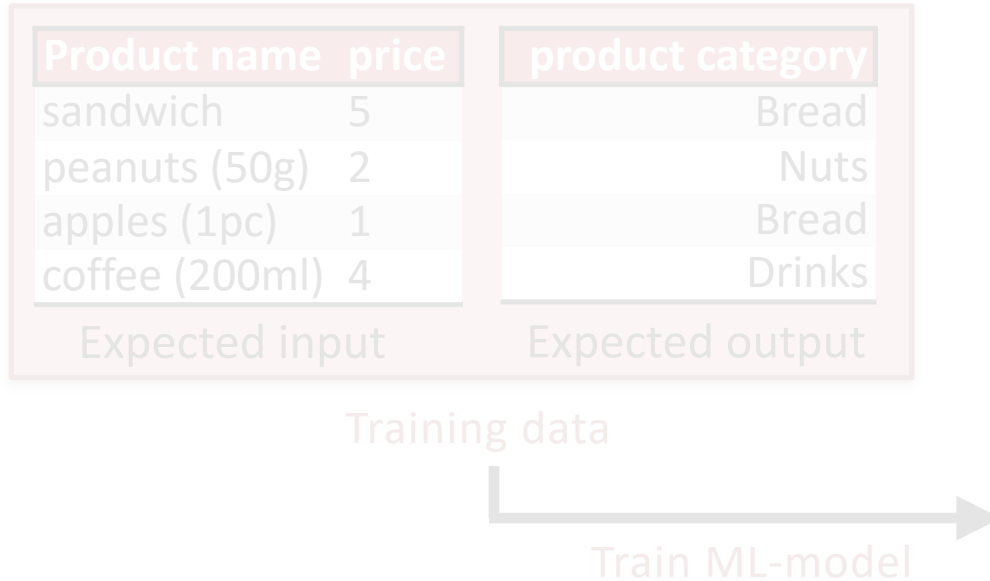
Unobserved data



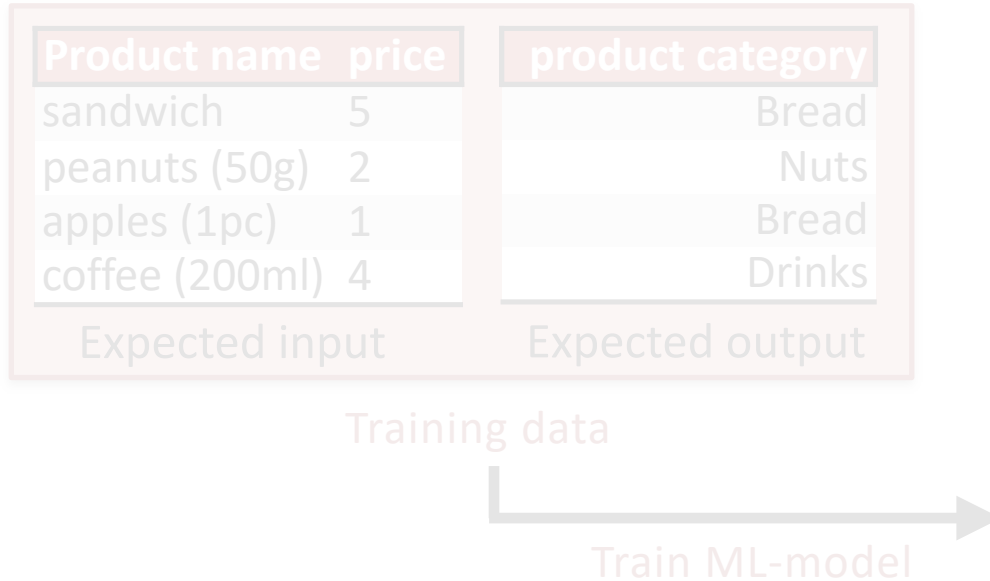
Predict unobserved data



Errors in Machine Learning



Errors in Machine Learning



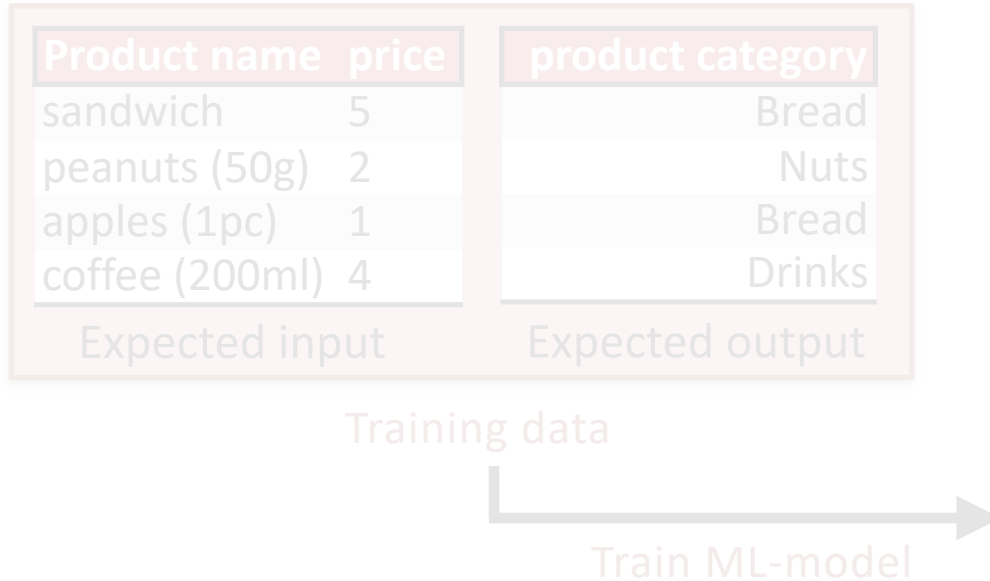
Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

Unobserved data



Predict unobserved data

Errors in Machine Learning



Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

Unobserved data



Predict unobserved data



Product category
Nuts
Bread
Fish



Errors in Machine Learning

Product name	price	product category
sandwich	5	Bread
peanuts (50g)	2	Nuts
apples (1pc)	1	Bread
coffee (200ml)	4	Drinks

Expected input Expected output

Training data

Train ML-model

Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

Unobserved data



Predict unobserved data

Product category
Nuts
Bread
Fish



Errors in Machine Learning

Product name	price	product category
sandwich	5	Bread
peanuts (50g)	2	Nuts
apples (1pc)	1	Bread
coffee (200ml)	4	Drinks

Expected input Expected output

Training data

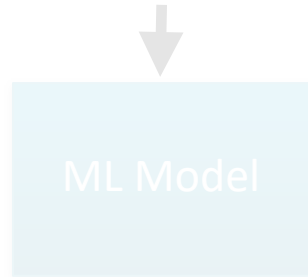
model

Can we assume that the training data:

- Has no errors in the data? (measurement error)
 - e.g. apples = bread?

Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

Unobserved data



Predict unobserved data

Product category
Nuts
Bread
Fish



Errors in Machine Learning

Product name	price	product category
sandwich	5	Bread
peanuts (50g)	2	Nuts
apples (1pc)	1	Bread
coffee (200ml)	4	Drinks

Expected input Expected output

Training data

model

Can we assume that the training data:

- Has no errors in the data? (measurement error)
 - e.g. apples = bread?
- Is representative? (coverage)

Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

Unobserved data



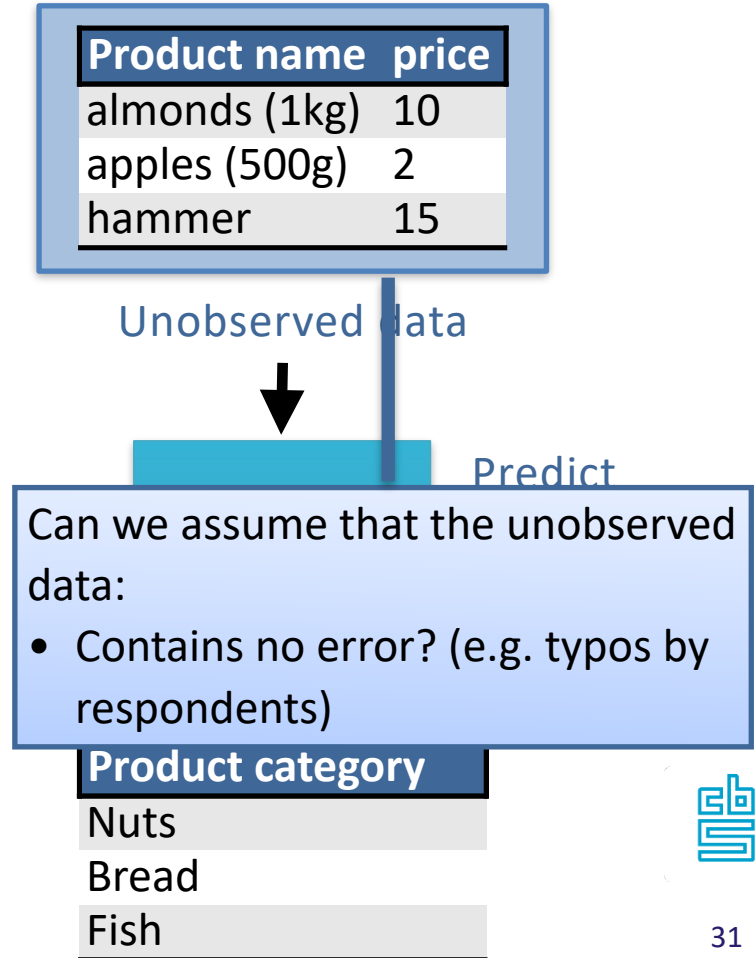
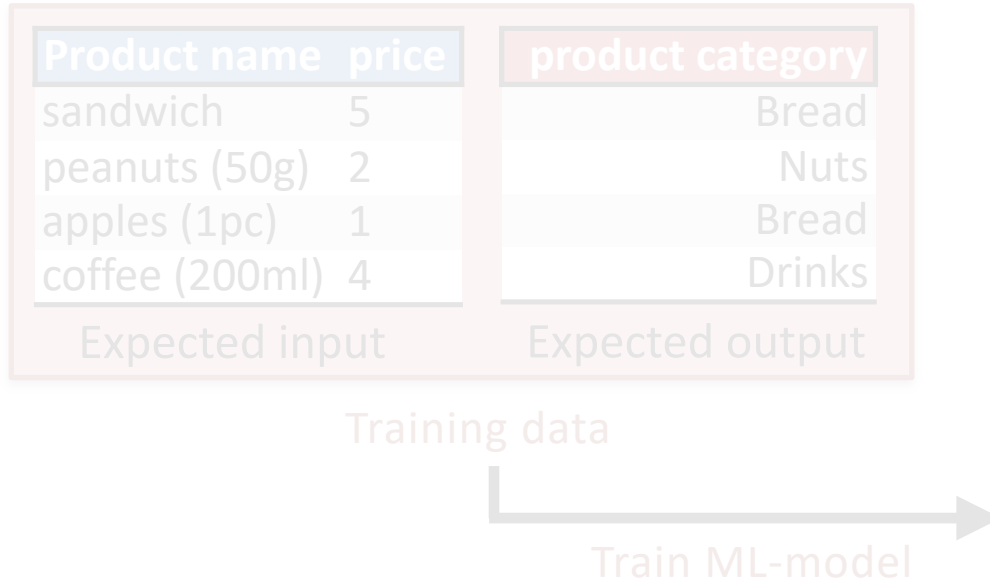
Predict unobserved data



Product category
Nuts
Bread
Fish



Errors in Machine Learning



Errors in Machine Learning

Product name	price	product category
sandwich	5	Bread
peanuts (50g)	2	Nuts
apples (1pc)	1	Bread
coffee (200ml)	4	

Expected input

Product name	price
almonds (1kg)	10
apples (500g)	2
hammer	15

Unobserved data

Aspects from TSE (Groves et al., 2004) can be used for the training and application of ML models!

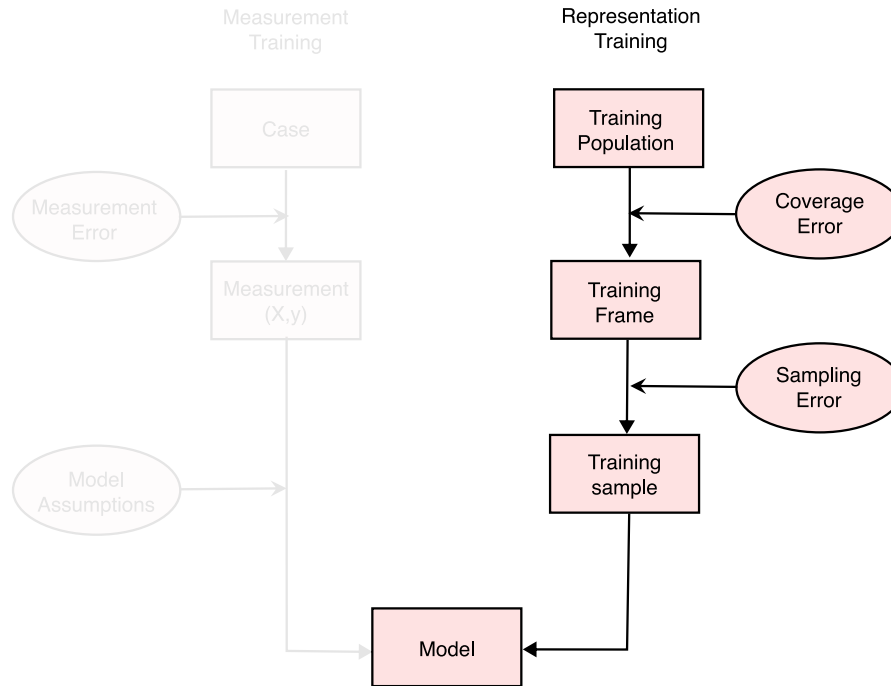
Train

Train ML-model

Predict unobserved data

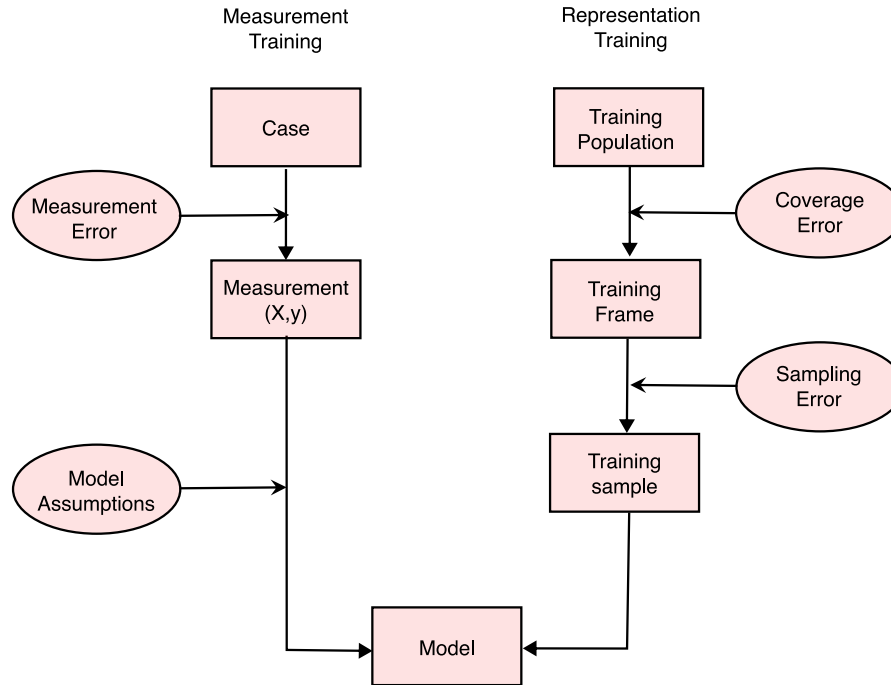
Product category
Nuts
Bread
Fish

Errors in Machine Learning



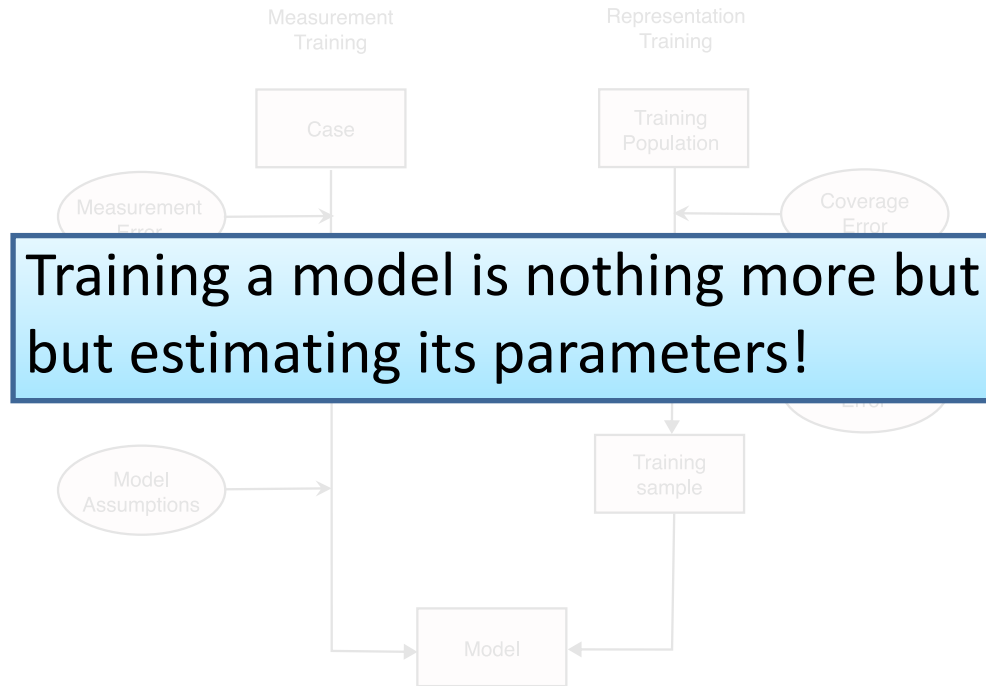
Total Machine Learning Error Framework (TMLE): Training Phase
(Puts, Salgado & Daas, 2024)

Errors in Machine Learning



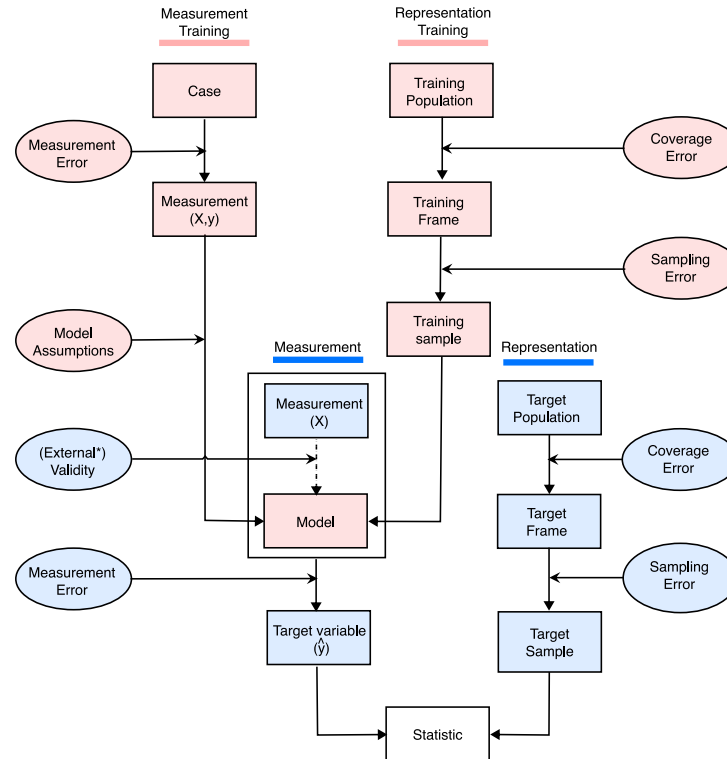
Total Machine Learning Error Framework (TMLE): Training Phase
(Puts, Salgado & Daas, 2024)

Errors in Machine Learning

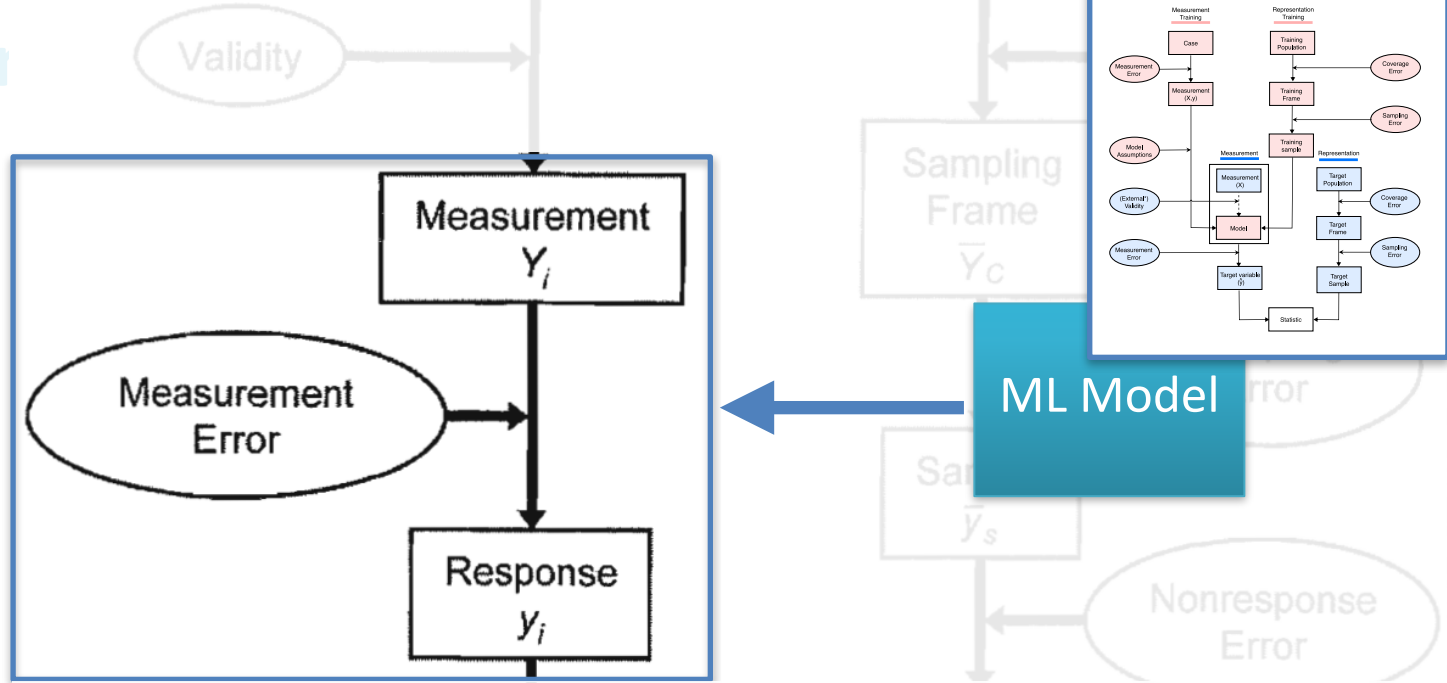


Total Machine Learning Error Framework (TMLE): Training Phase
(Puts, Salgado & Daas, 2024)

Errors in Machine Learning



Total Machine Learning Error Framework (TMLE): Training Phase
(Puts, Salgado & Daas, 2024)



ML is used in the reporting process, it therefore affects the measurement error

Applying the Total Machine Learning Framework

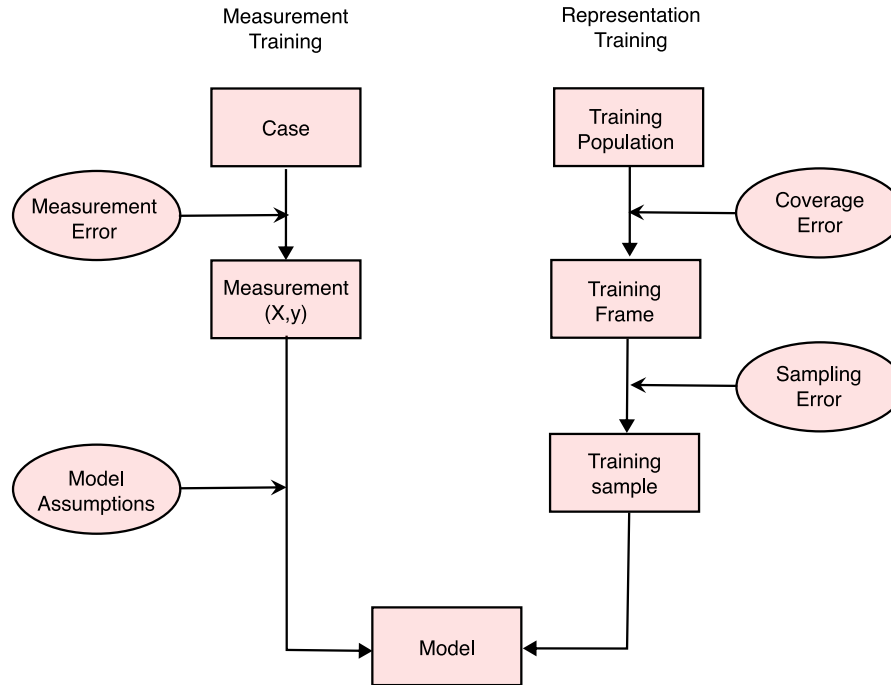


Applying TMLE on ML for Product Categorization

The TMLE applied on the ML for product categorization



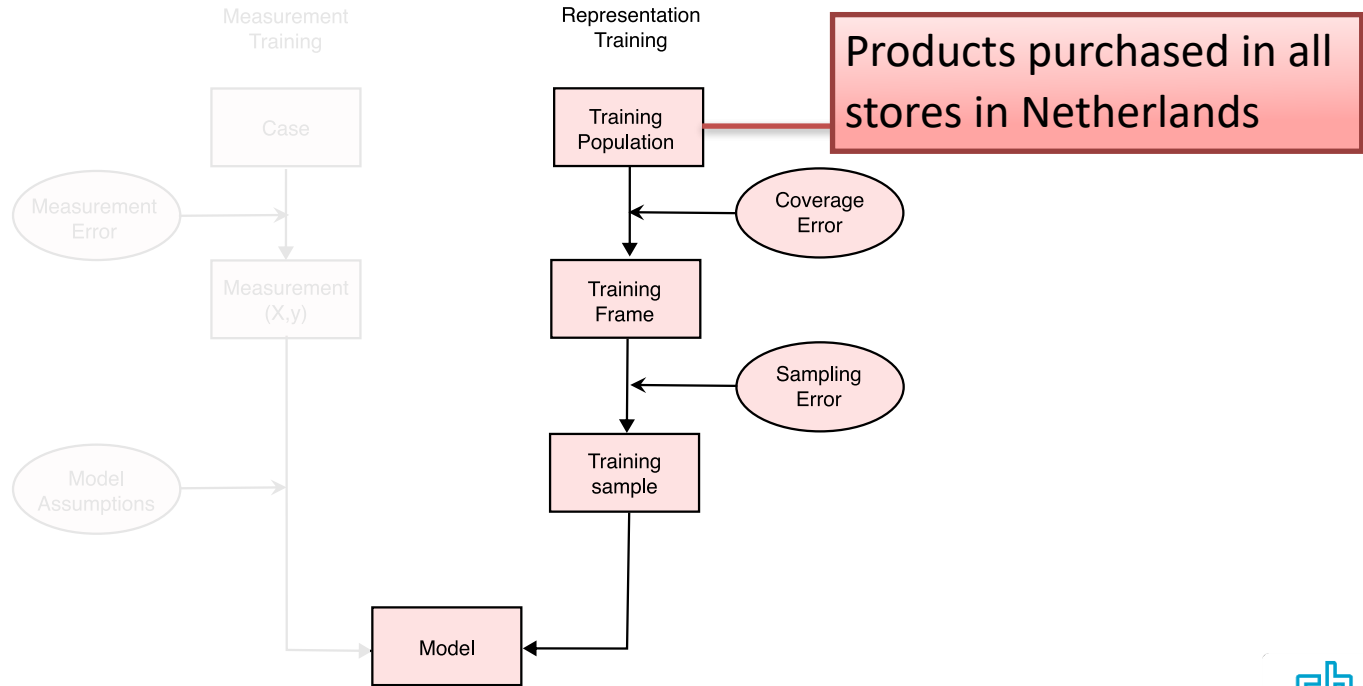
Applying TMLE on ML for Product Categorization



The TMLE applied on the ML for product categorization



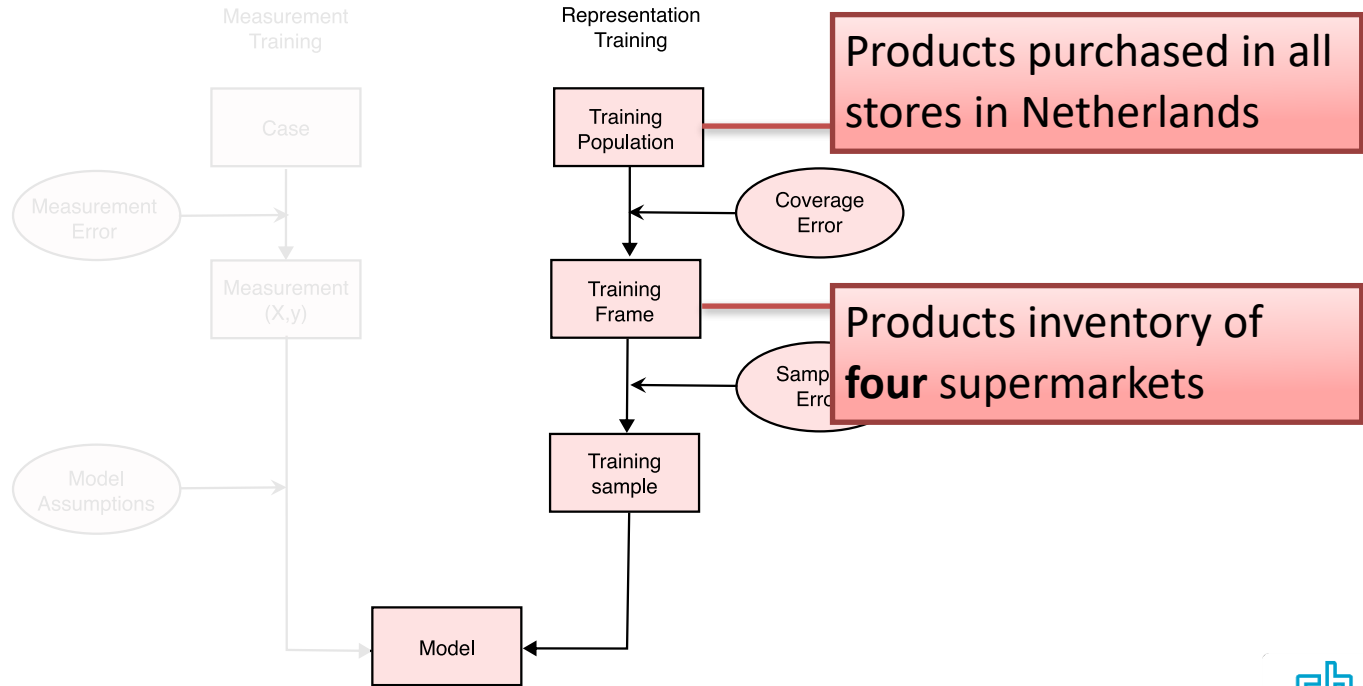
Applying TMLE on ML for Product Categorization



The TMLE applied on the ML for product categorization



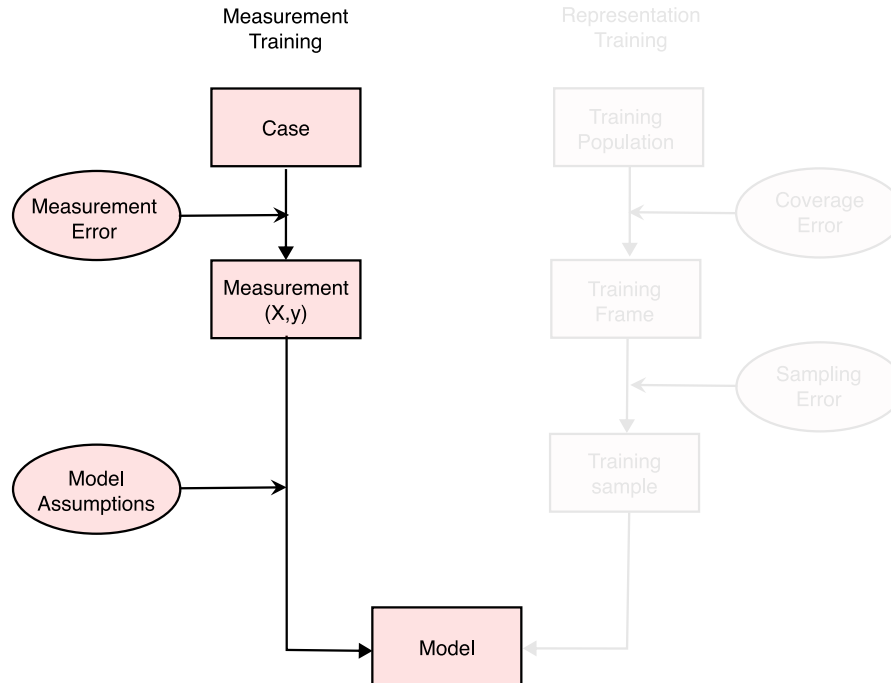
Applying TMLE on ML for Product Categorization



The TMLE applied on the ML for product categorization

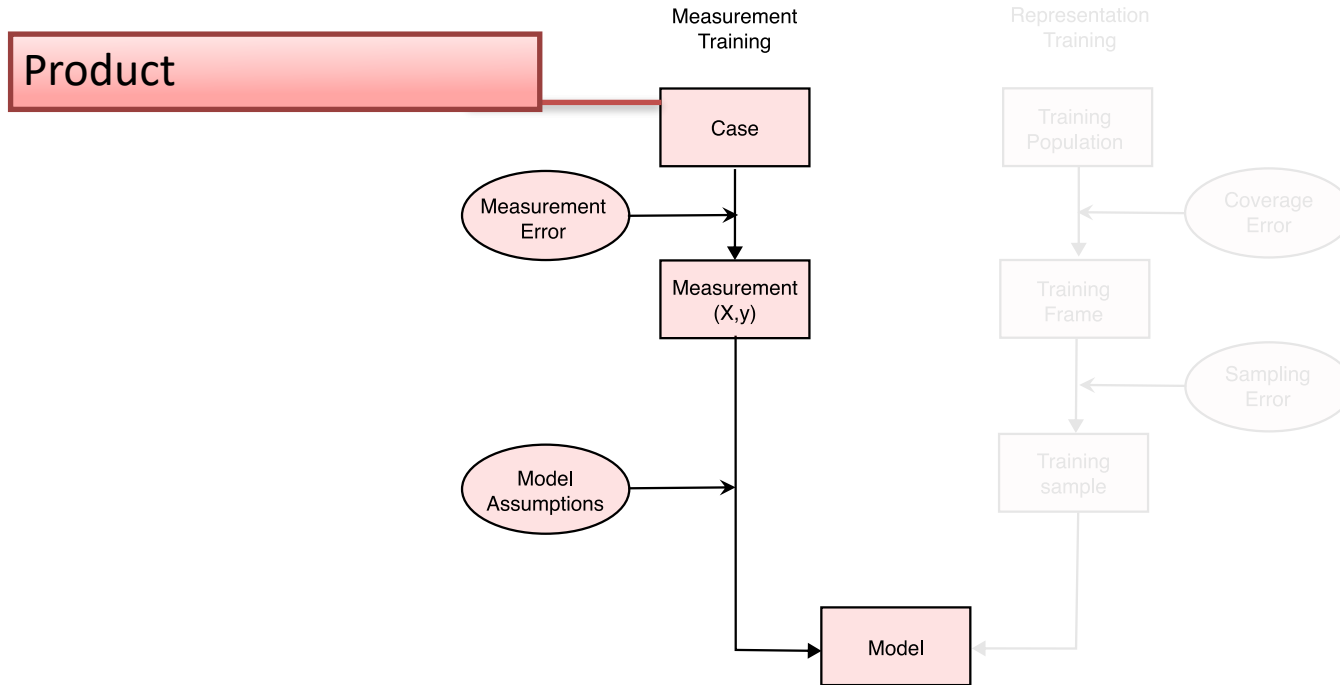


Applying TMLE on ML for Product Categorization



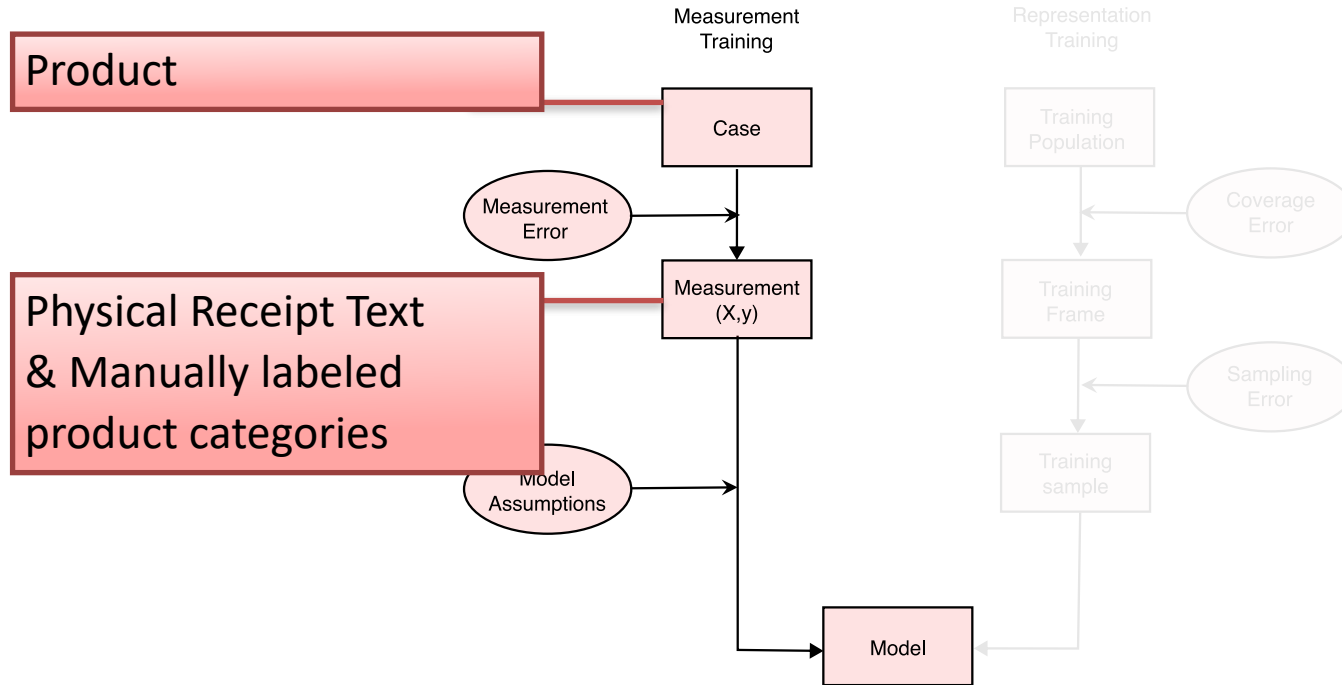
The TMLE applied on the ML for product categorization

Applying TMLE on ML for Product Categorization



The TMLE applied on the ML for product categorization

Applying TMLE on ML for Product Categorization



The TMLE applied on the ML for product categorization

Conclusions



Conclusions

- ML can be used to lower response burden
 - Affects the measurement error of the survey (TSE)
- Aspects from TSE (Groves et al., 2004) can be used to assess the quality of a ML model
 - *Total Machine Learning Error Framework* (Puts, Salgado & Daas, 2024)
- For example in ML for product categorization:
 - coverage error: four stores vs. all stores
 - measurement error in training data: wrong product categories
- We have done similar work for the mobility survey (Smart survey, ML)



Future work

- Expand the types of errors in TMLE
 - e.g. under- and over-coverage
- Quantification of errors in the TMLE

Special thanks to

Co-authors

Marco Puts

Jonas Klingwort

Vera Toepoel

Tim de Jong

Acknowledgements

Yvonne Gootzen

Piet Daas

Contact

Chris Lam

c.lam@cbs.nl

+ 31 6 25 48 41 62



References

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. John Wiley & Sons.

Puts, M., Salgado, D., & Daas, P. (2024). Leveraging Machine Learning for Official Statistics: A Statistical Manifesto. *arXiv preprint arXiv:2409.04365*. <https://doi.org/10.48550/arXiv.2409.04365>

