Causal Inference Libraries: What They Do and What I'd Like Them to Do

Kevin Klein, QuantCo

Agenda

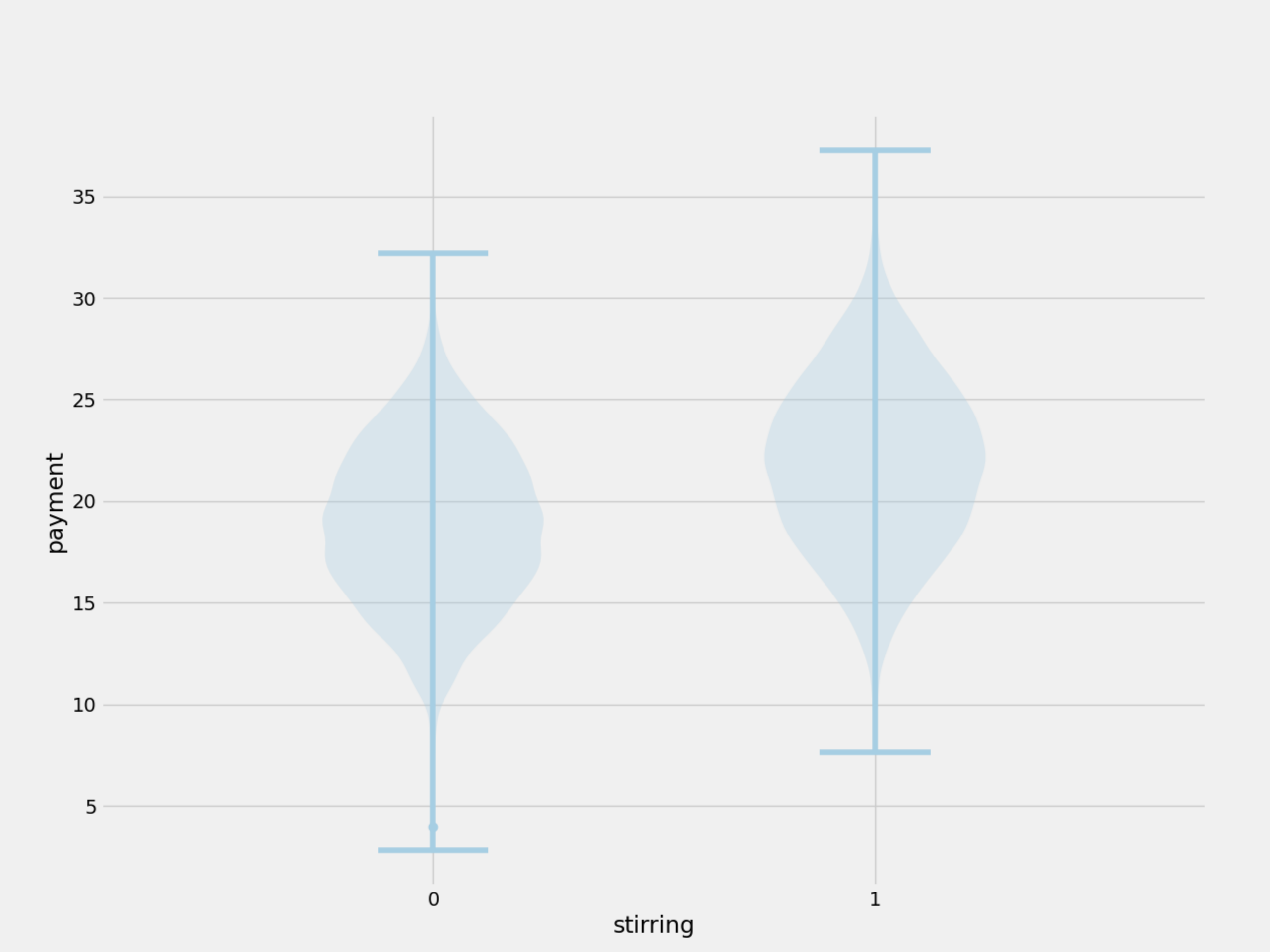
1. Why Causal Inference and why heterogeneity?
2. How can we estimate heterogeneous treatment effects on paper?
3. How can we estimate heterogeneous treatment effects in practice?
4. What are we missing from EconML and CausalML?

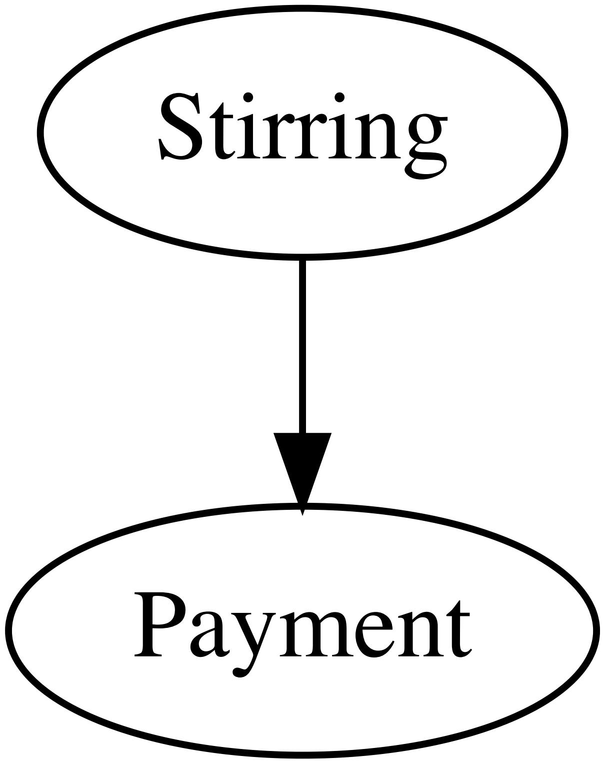
Risotto

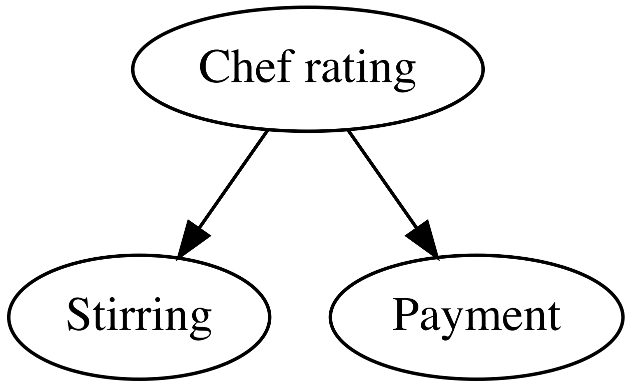
- Risotto can either be prepared
    - in a laborous and delicate fashion, involving a lot of **stirring** or
    - in a cut-throat, cantine style fashion, **not** involving a lot of **stirring**
- Consumers of risotto are **free to decide how much they pay** for their risotto.
- Naturally we wonder: should we be stirring?

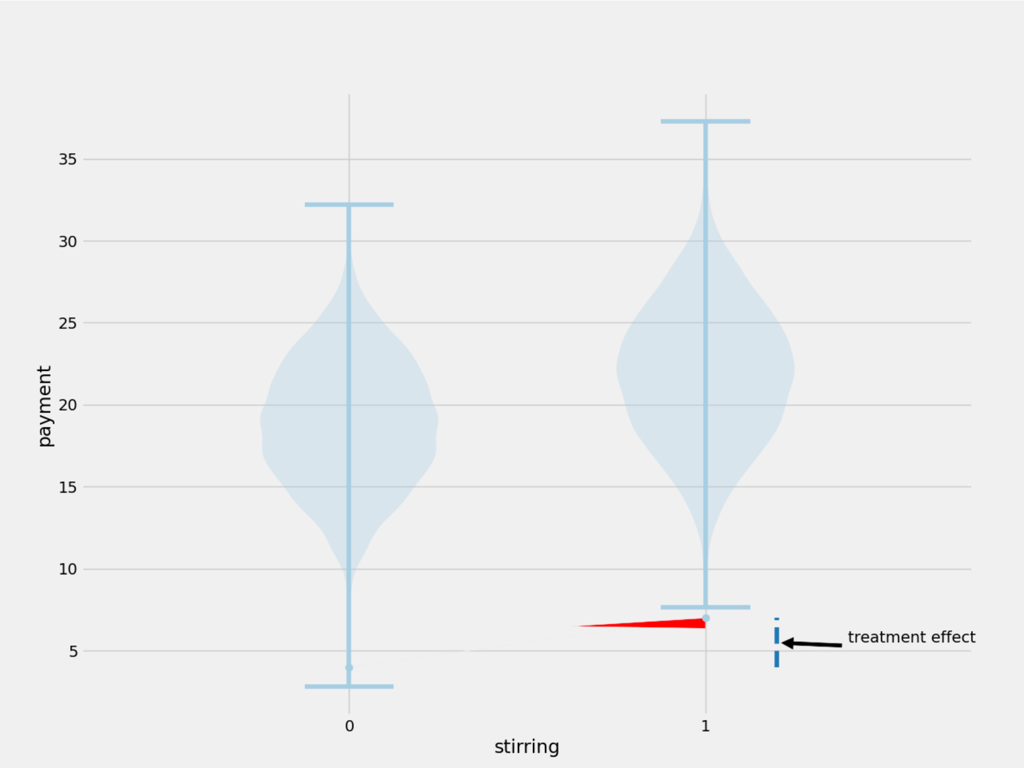
When prediction fails

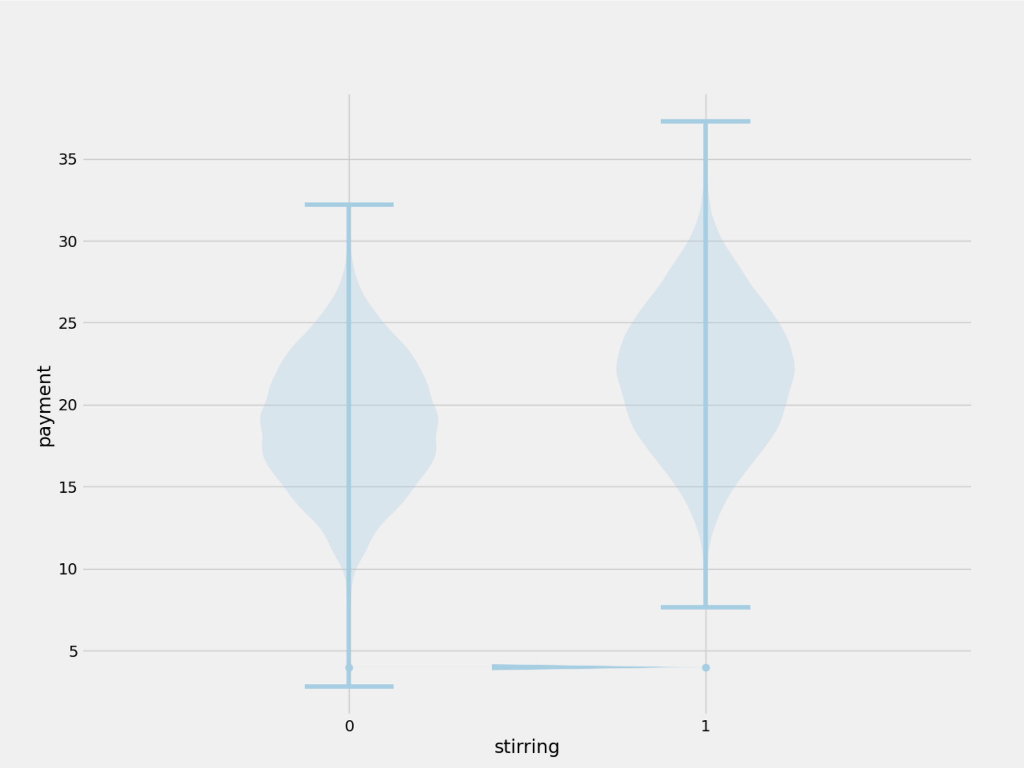
What happens when we intervene on a data point from the left, i.e. `stirring = 0`, and now - keeping everything else unchanged - make sure that a gas stove is used, i.e. `stirring = 1`?



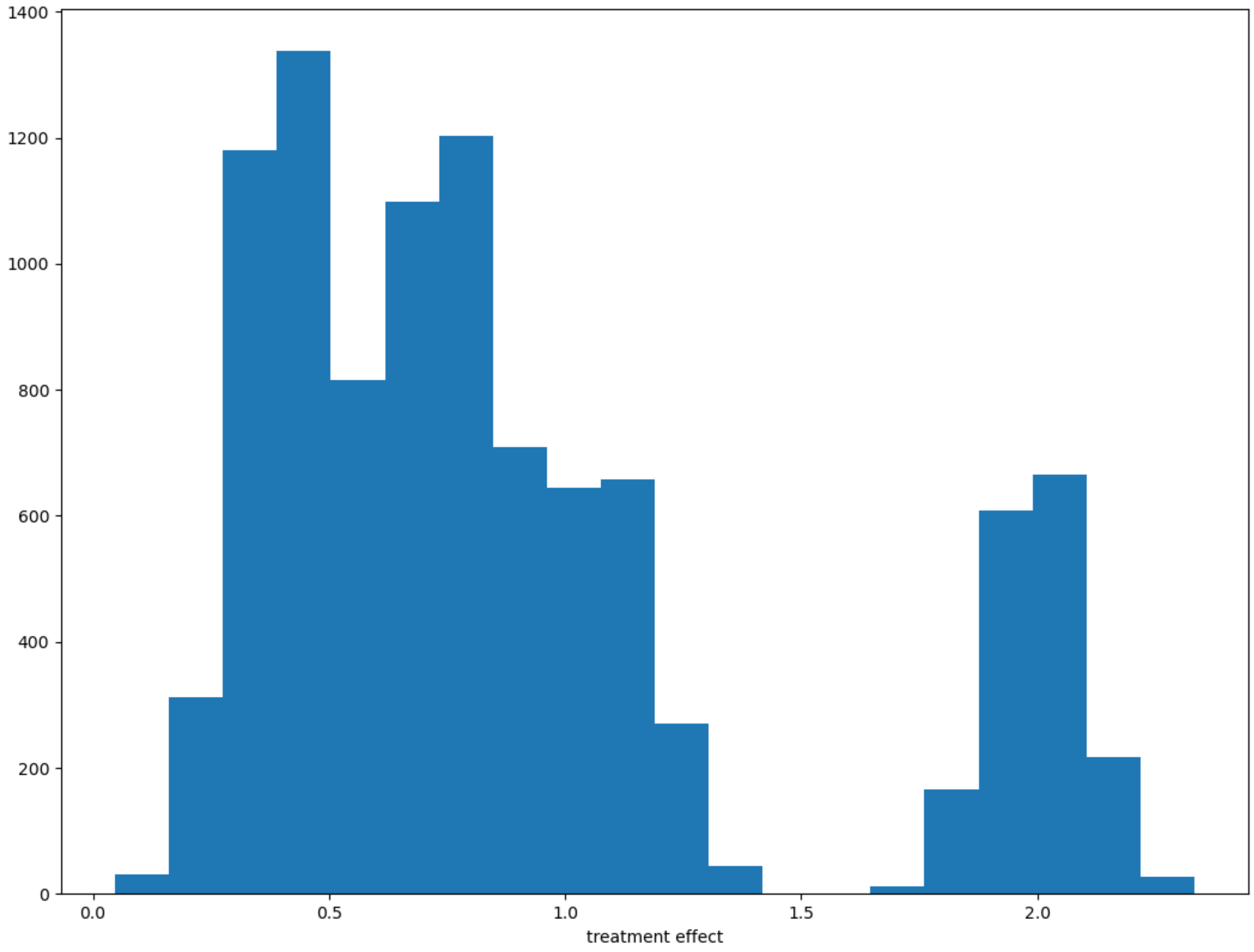

Stirring
Payment


Chef rating
Stirring
Payment


35
30
25
20
15
10
5
payment
0
1
stirring
treatment effect


35
30
25
20
15
10
5
payment
0
1
stirring

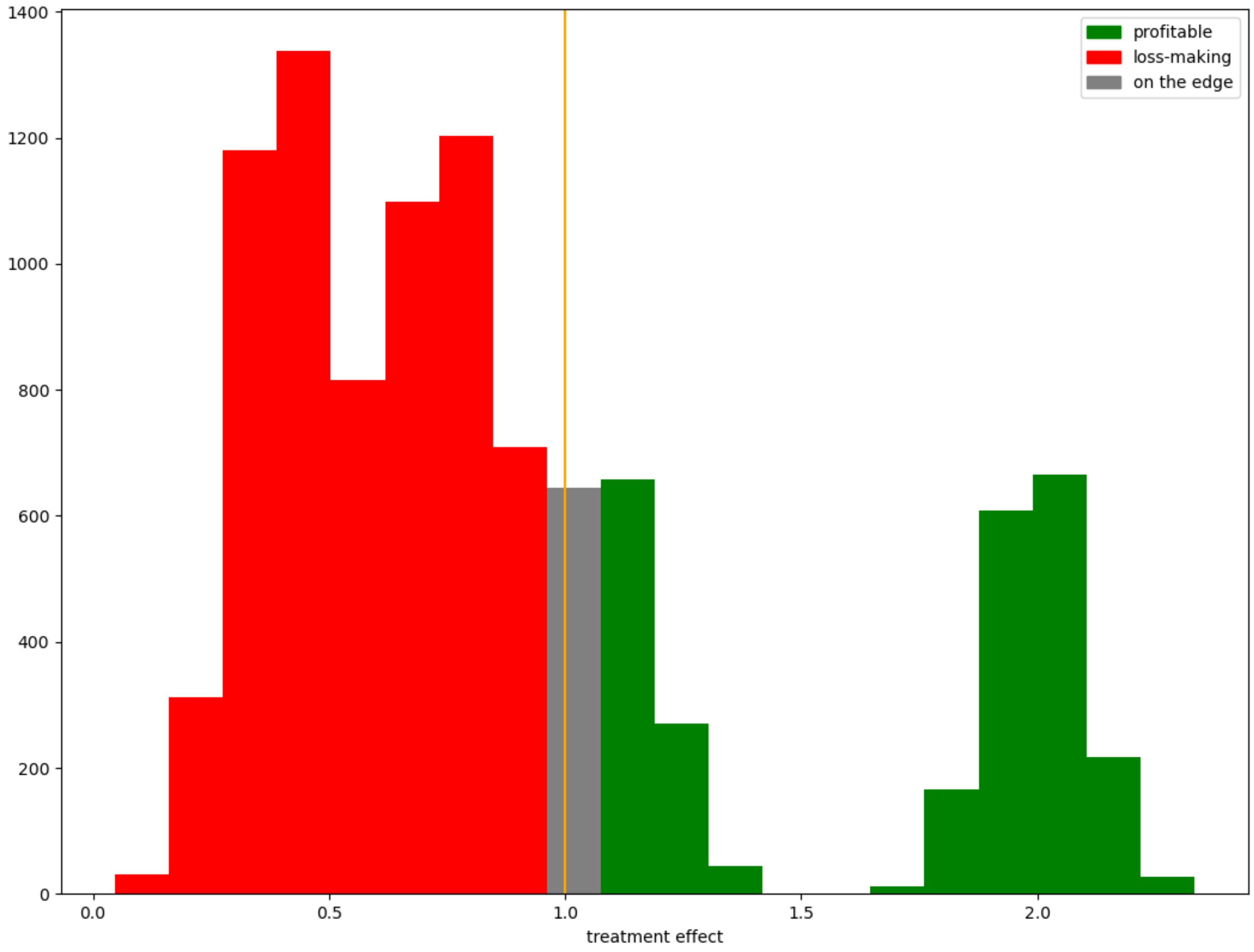
Why heterogeneity



Kevin Klein, @kevkle

To stir or not to stir, the maths

- Assume that the cost of stirring amounts to 1$ per unit.
- Also assume that the overall revenue when never stirring is R.
- Then, the overall revenue when **always stirring** is $R - n \cdot 1 + \delta_1$
  - The plot from the previous slide tells us that $n \cdot 1 > \delta$.
- The overrall revenue of **stirring when we expect it to pay off**: $R - k \cdot 1 + \delta_\pi$
  - We can condition on certain 'covariates'/features to decide for whom it pays off.
  - When doing this 'right', we get that $\delta_\pi > k \cdot 1$.



profitable
loss-making
on the edge
1400
1200
1000
800
600
400
200
0
0.0
0.5
1.0
1.5
2.0
treatment effect

Estimating heterogeneity on paper

The fundamental problem of Causal Inference

Desire

Consumer	Age of consumer	Non-stirred outcome/payment	Stirred outcome/payment	Individual treatment effect
Bob	28	21	21.8	.8
Anne	10	12	12	0

The fundamental problem of Causal Inference

Reality

Consumer	Age of consumer	Non-stirred outcome/payment	Stirred outcome/payment	Individual treatment effect
Bob	28	21	?	?
Anne	10	?	12	?

What now?

- We can't know the Individual Treatment Effect (ITE).
- Yet, we can define an estimand, the Conditional Average Treatment Effect (CATE), which we can actually estimate:
  $\tau(X) := \mathbb{E}[\text{payment}|X, \text{stirring}] - \mathbb{E}[\text{payment}|X, \text{no stirring}]$

Conventional assumptions for estimating heterogeneous treatment effects

- Positivity/overlap
- Conditional ignorability/unconfoundedness
- Stable Unit Treatment Value (SUTVA)

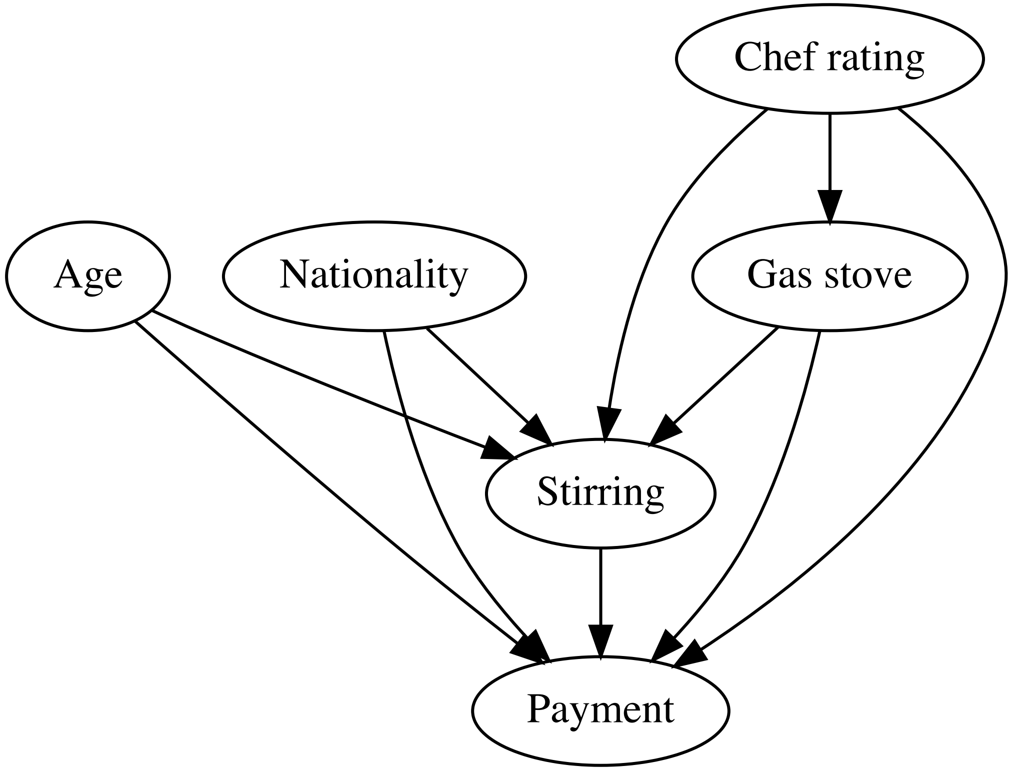
A randomized control trial usually gives us the first two for free.

MetaLearners

- MetaLearners are CATE models which rely on typical, arbitrary machine learning estimators (classifiers or regressors) as components. Their output is an estimate of the heterogeneous treatment effect.
- An simple, intuitive, yet often somewhat disappointing MetaLearner is the T-Learner: $\hat{\tau}(X) = \mu_1(X) - \mu_0(X)$
- Other examples are: S-Learner, F-Learner, X-Learner, R-Learner, M-Learner, DR-Learner

Estimating heterogeneity in practice

	EconML	**CausalML**
Developed by	MSR/py-why	Uber
License	MIT	Apache 2.0
Features	asdf	asdf
	asdf	asdf
	asdf	asdf
	asdf	asdf
MetaLearner API	sklearn	sklearn



Risotto consumption: a simulation

Risotto consumption: a simulation

age	nationality	chef_rating	gas_stove	μ	T	τ	Y
50.77	Indonesia	0.53	1	20.73	1	0.34	21.08
59.48	Iraq	0.46	0	20.46	0	0.76	20.46
47.25	India	0.46	0	24.29	0	0.19	24.29
22.21	Italy	0.58	0	15.90	1	0.88	16.79

$\mu(X) \equiv$ the 'base outcome', i.e. outcome/payment without stirring

$T \equiv$ the treatment, whether the risotto has been stirred or not

$\tau(X) \equiv$ the heterogeneous treatment effect

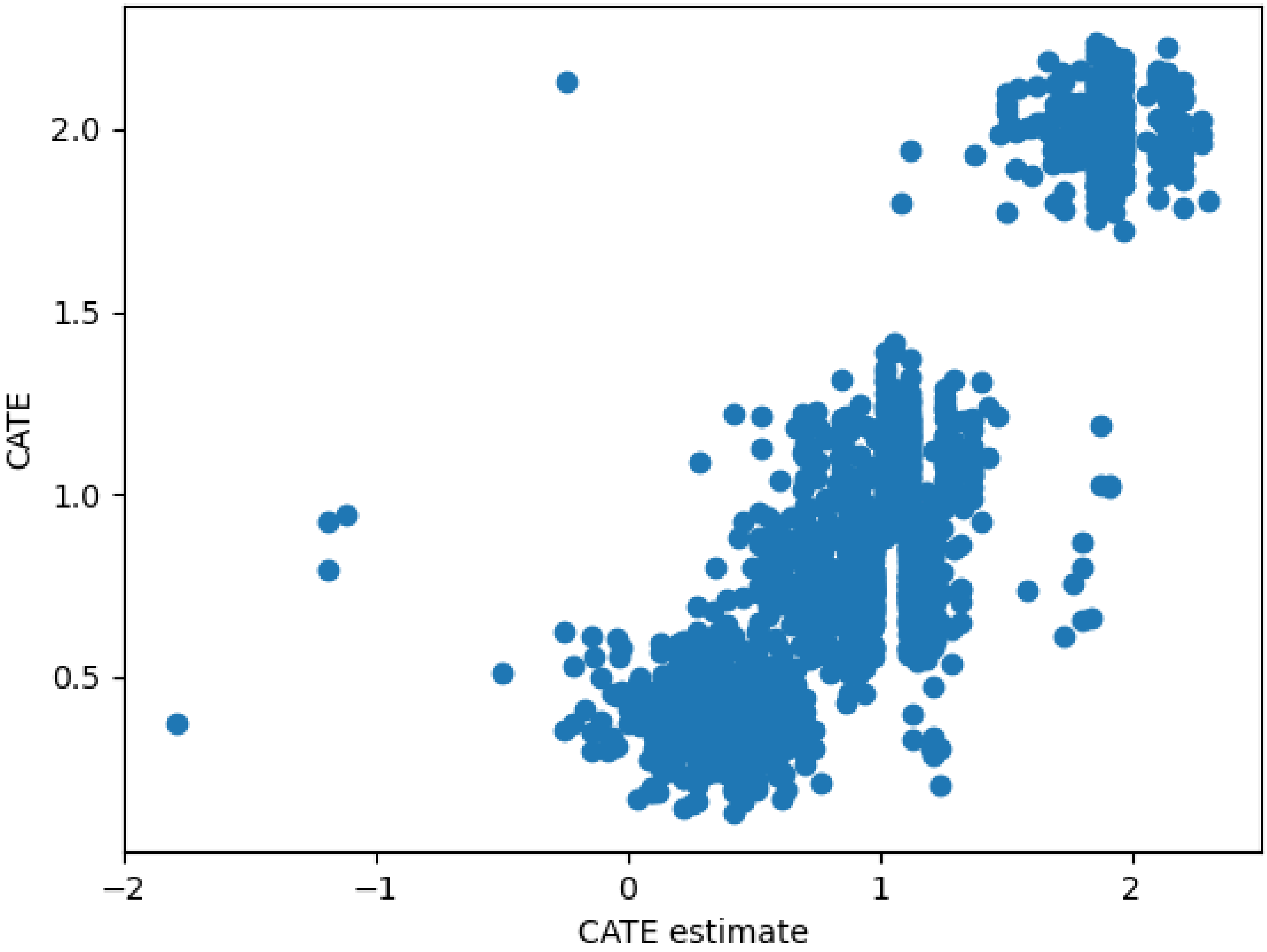
$Y \equiv$ the outcome, the final payment

$$Y = \mu(X) + T \cdot \tau(X)$$

```python
# One-hot encoding
X = pd.concat([
    df[numerical_covariates],
    pd.get_dummies(df["nationality"])
], axis=1)

# Model definition
reg = lgbm.LGBMRegressor()
clf = lgbm.LGBMClassifier()
model = causalml.BaseRRegressor(
    outcome_learner=reg,
    effect_learner=reg,
    propensity_learner=clf,
)

# Model training and prediction
model.fit(X=X, treatment=df[treatment], y=df[outcome])
cate_estimates = model.predict(X)
```

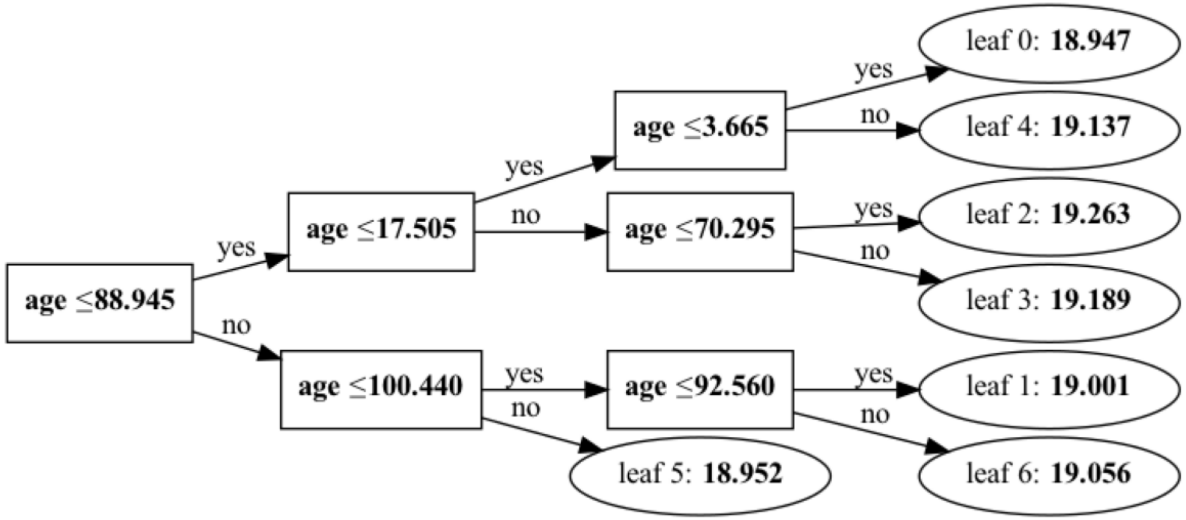


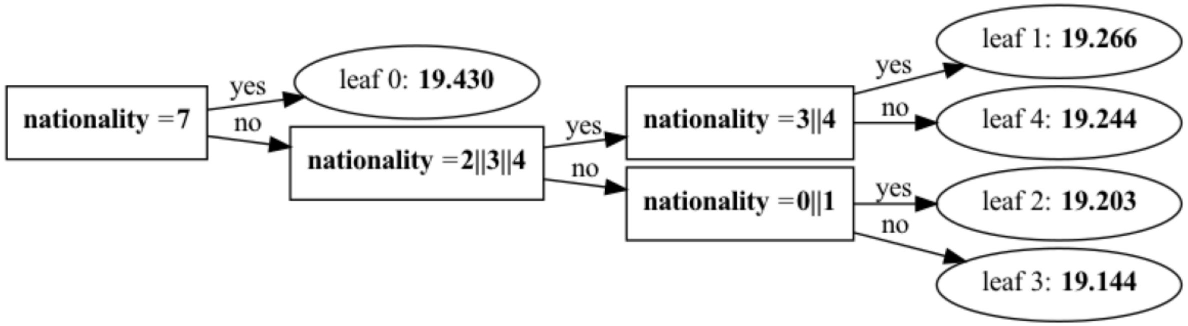
CATE
2.0
1.5
1.0
0.5
−2
−1
0
1
2
CATE estimate

Pains and problems in practice (P^3)

P^3#1: Categorical features

- `lightgbm` is a very popular choice for prediction on tabular datasets.
- One of is nice features is that it works natively with categorical features.
  E.g. instead of having to one-hot encode categoricals, one can 'tell' `lightgbm` that a single columns is to be treated as a categorical.


age ≤88.945
yes
no
age ≤17.505
age ≤100.440
age ≤3.665
age ≤70.295
age ≤92.560
leaf 0: 18.947
leaf 4: 19.137
leaf 2: 19.263
leaf 3: 19.189
leaf 1: 19.001
leaf 6: 19.056
leaf 5: 18.952


nationality =7
leaf 0: 19.430
nationality =2||3||4
nationality =3||4
nationality =0||1
leaf 1: 19.266
leaf 4: 19.244
leaf 2: 19.203
leaf 3: 19.144
yes
no

- Option 1: Use `pandas` `category` dtype

```
df["nationality"] = df["nationality"].astype("category")
model = lgbm.LGBMRegressor()
model.fit(df[["nationality"]], df["payment"])
```

- Option 2: Explicitly set `categorical_indices`

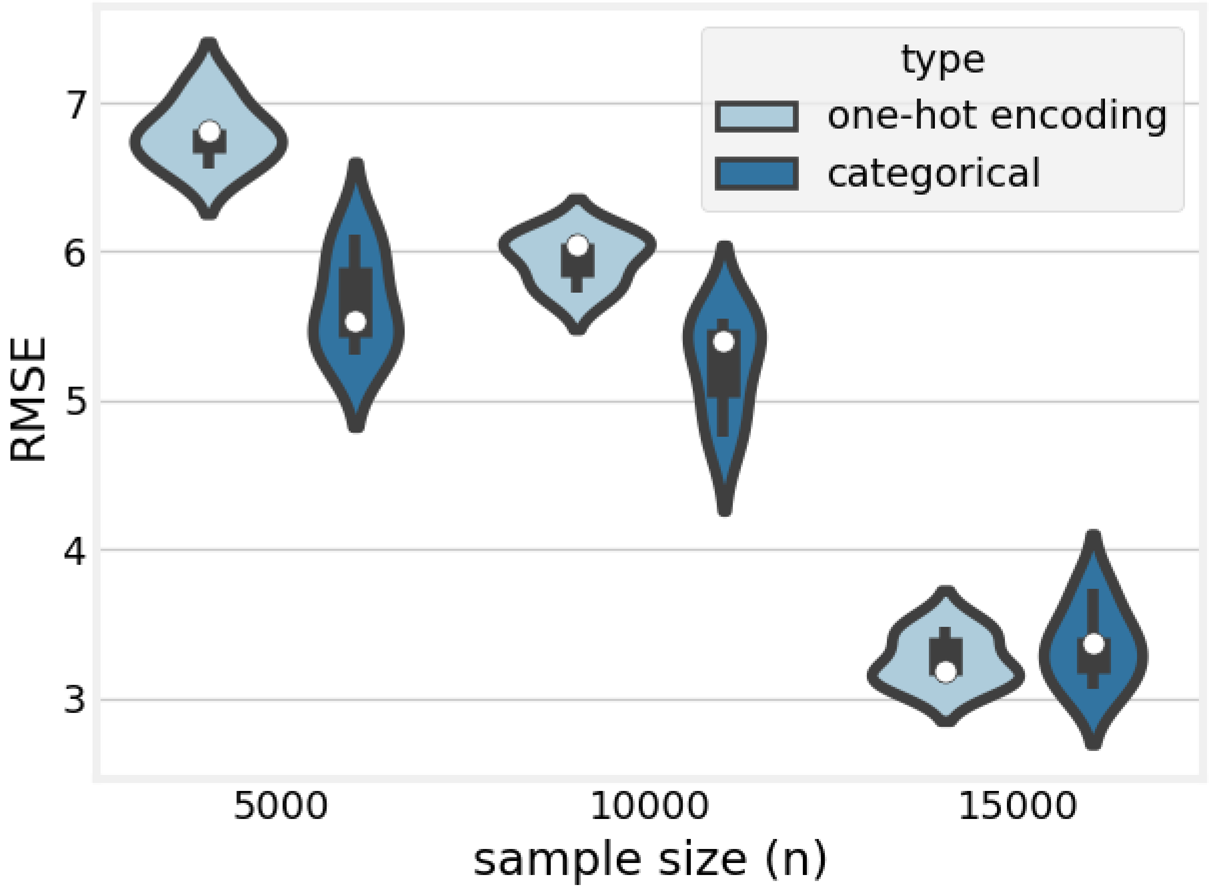
```
df["nationality"] = df["nationality"].astype("category").cat.codes
model = lgbm.LGBMRegressor(categorical_feature=[0])
model.fit(df[["nationality"]], df["payment"])
```

- Unfortunately, both options don't work with `causalml` and `econml`.
- Option 1 is not possible since both convert `pandas` input to `numpy` objects in a 'validation' step.
  - `X, treatment, y = convert_pd_to_np(X, treatment, y)`
  - https://github.com/uber/causalml/blob/3b3daaa3cd2ef1960028908c152cfd242b37712c/causalml/inference/meta/rlearner.py#L100
- Option 2 is not possible since constructor parameters can't be passed.

- A hack is - of course - possible to indirectly use option 2:

```
from functools import partialmethod
from lightgbm import LGBMRegressor
LGBMRegressor.fit = partialmethod(
  LGBMRegressor.fit,
  categorical_feature=[0],
)
```

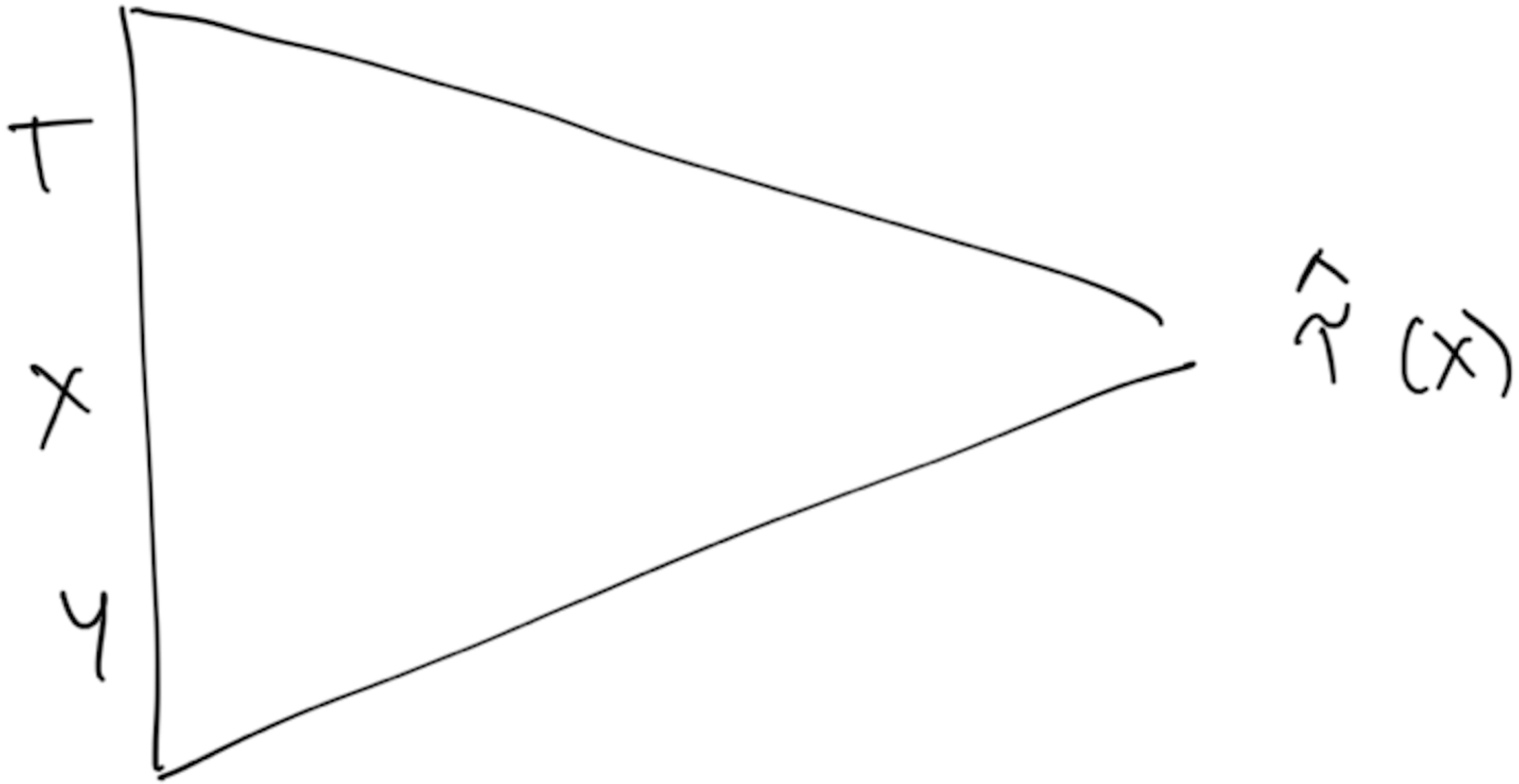
Tying back to our example: what's the difference?



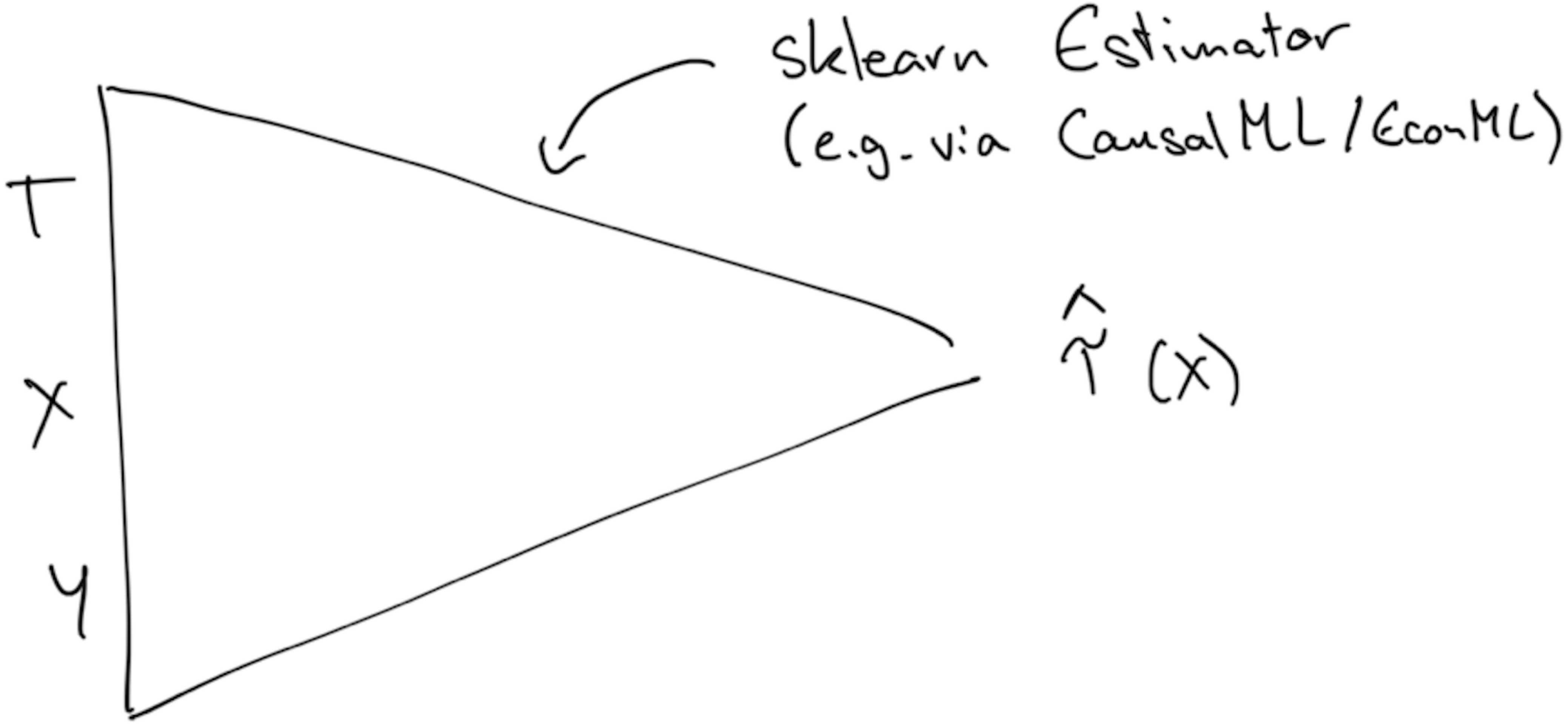
P^3 #2: Reusing component models

- There are many free parameters when training meta learners:
   i. Choose which meta learner (e.g. R-Learner)
      - This is hard, see Curth, van der Schar (2021)
   ii. Per estimand, choose an estimator (e.g. boosted trees)
   iii. Per estimator
      - per hyperparameter (e.g. depth), choose a value (e.g.12)
- In practice, this often boils down to 'trying out' different constellations, e.g. via random search or grid search.
- See EconML issue 646.

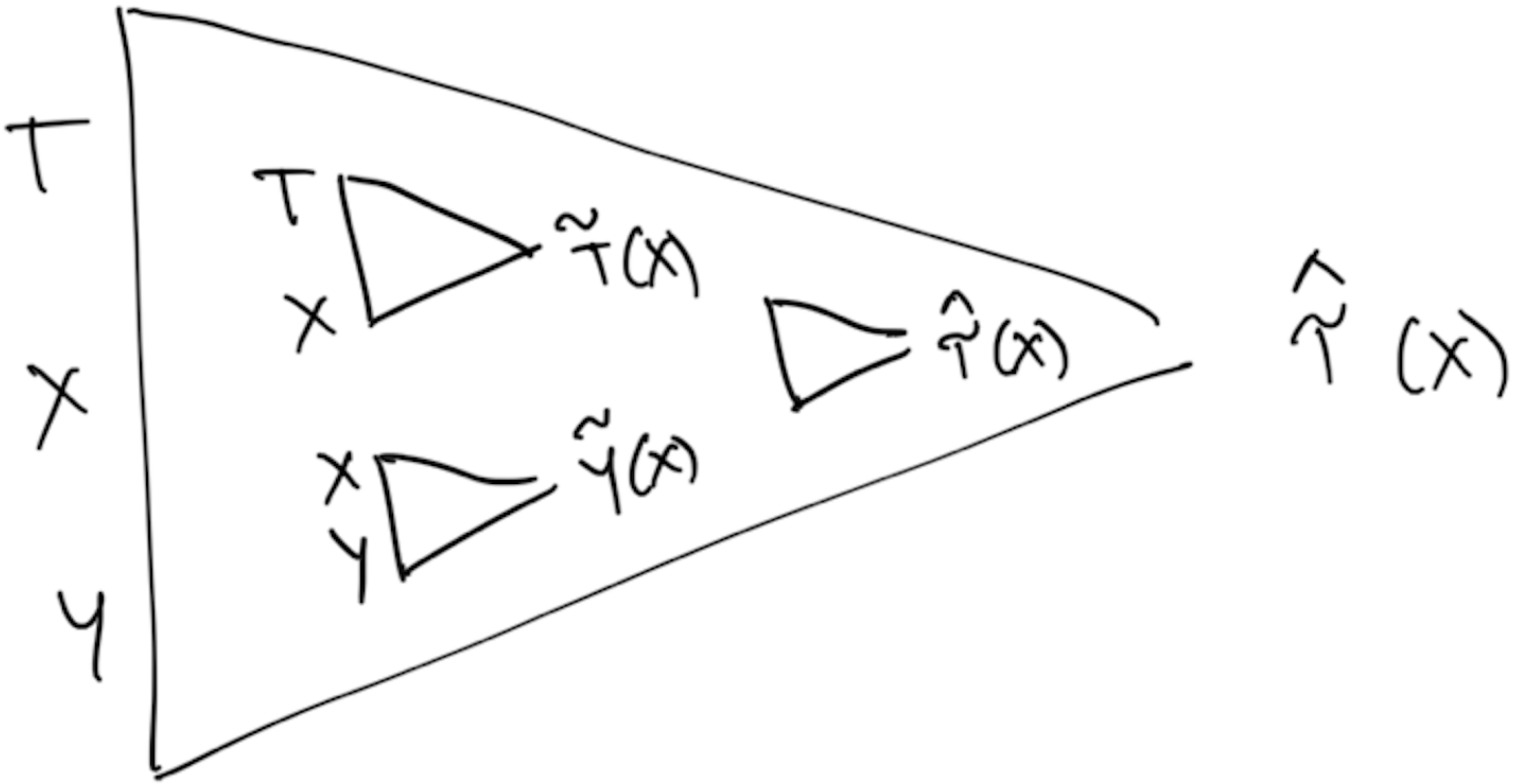
A MetaLearner



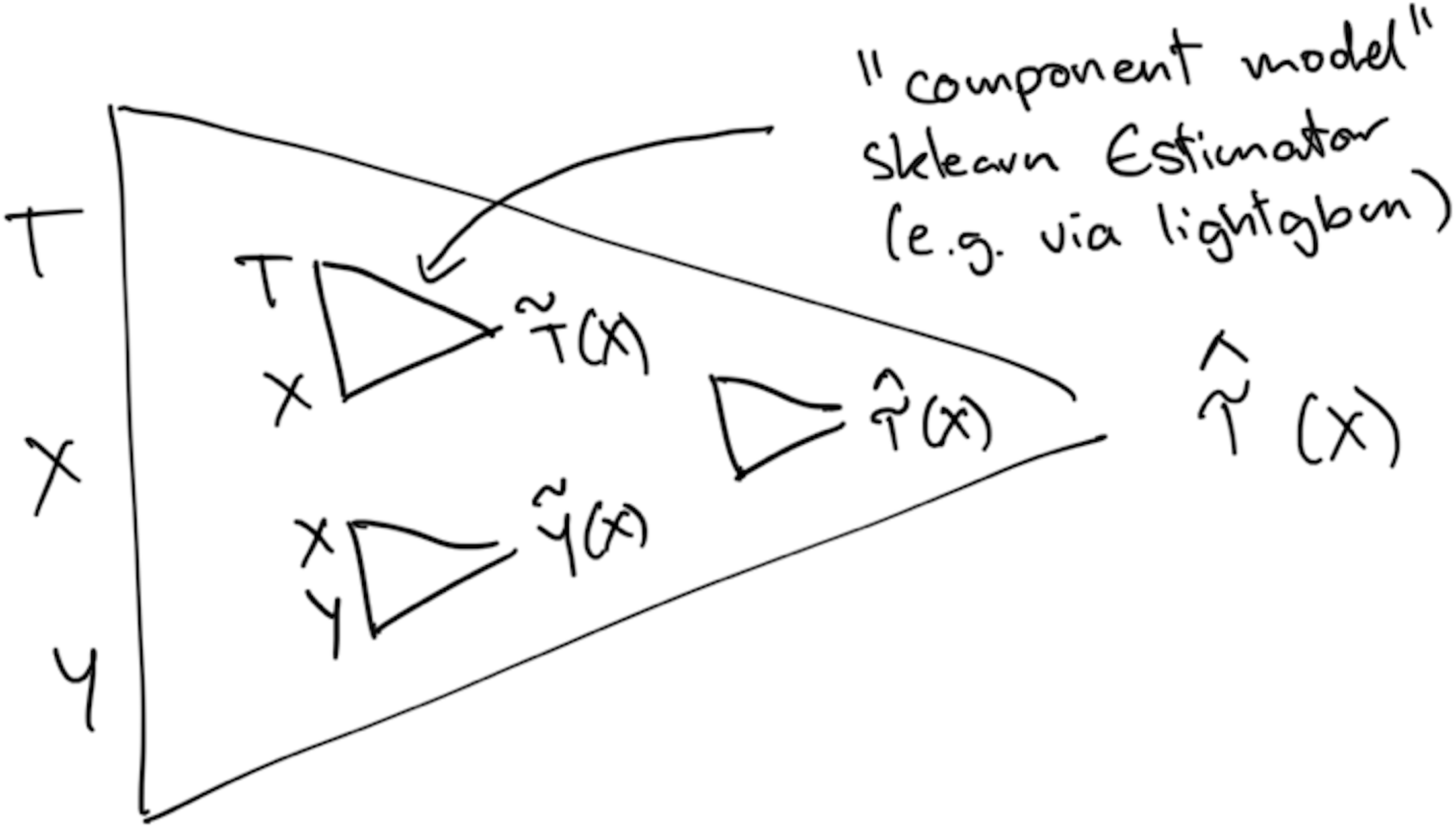
A MetaLearner



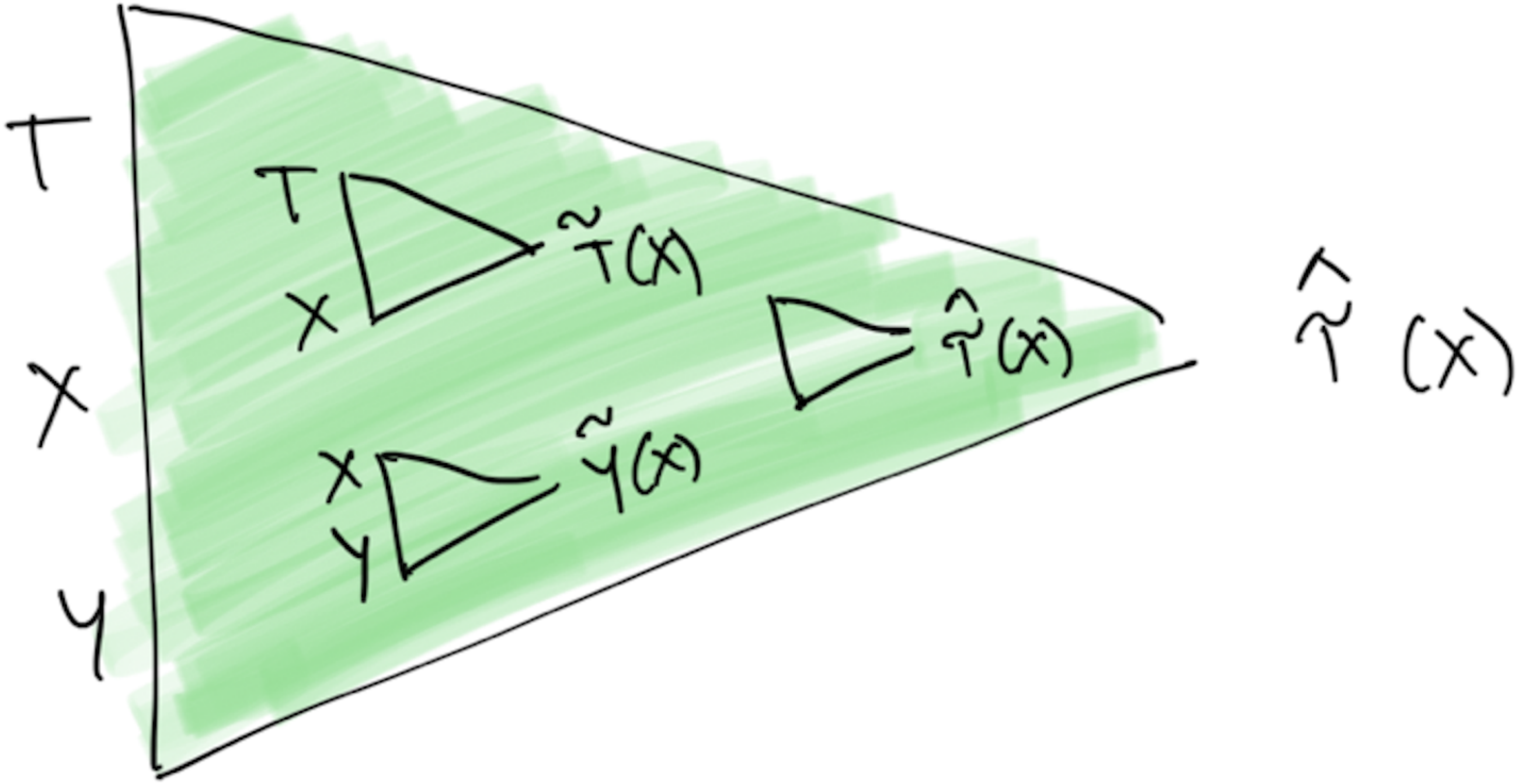
The R-Learner



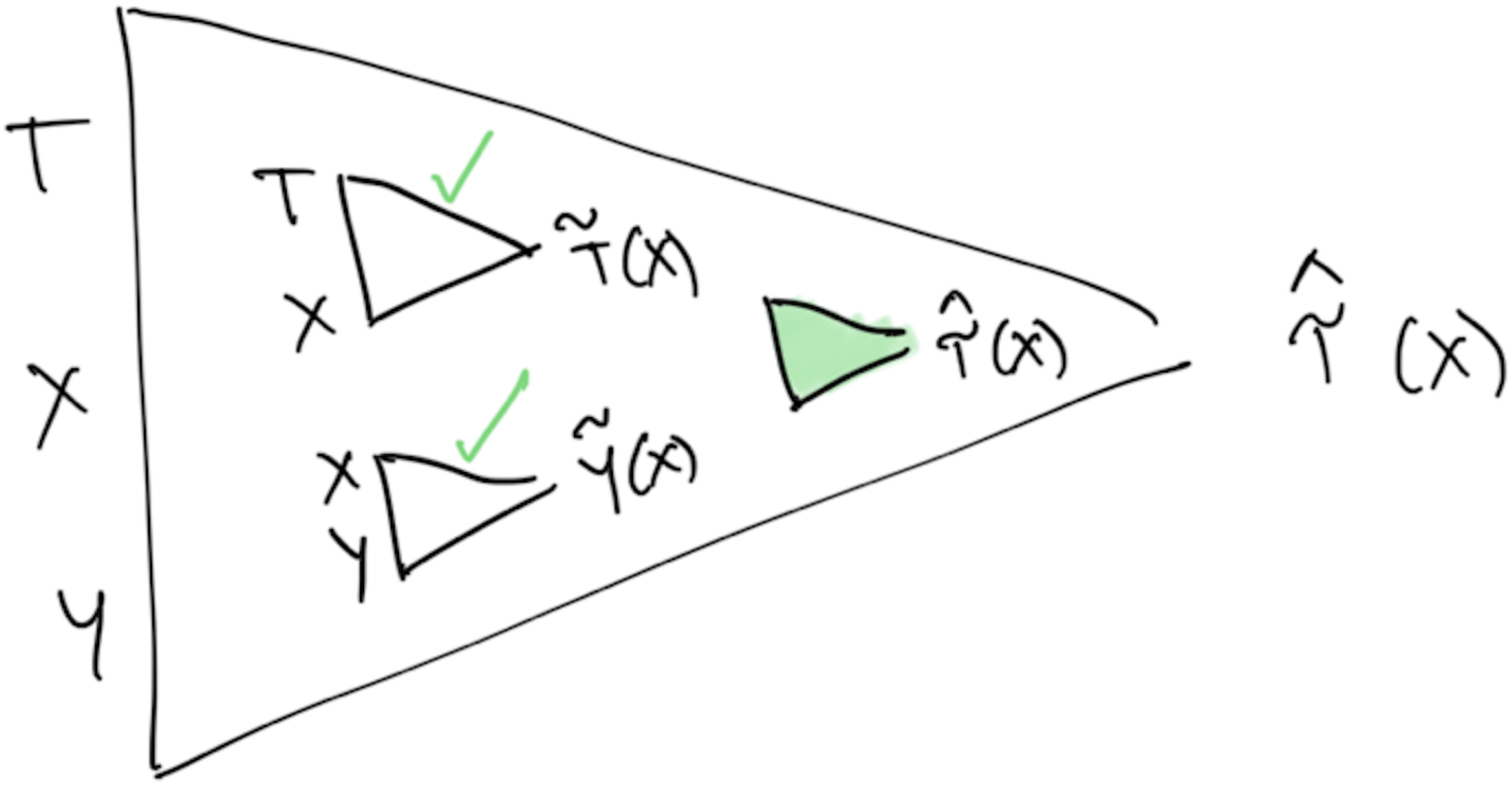
The R-Learner

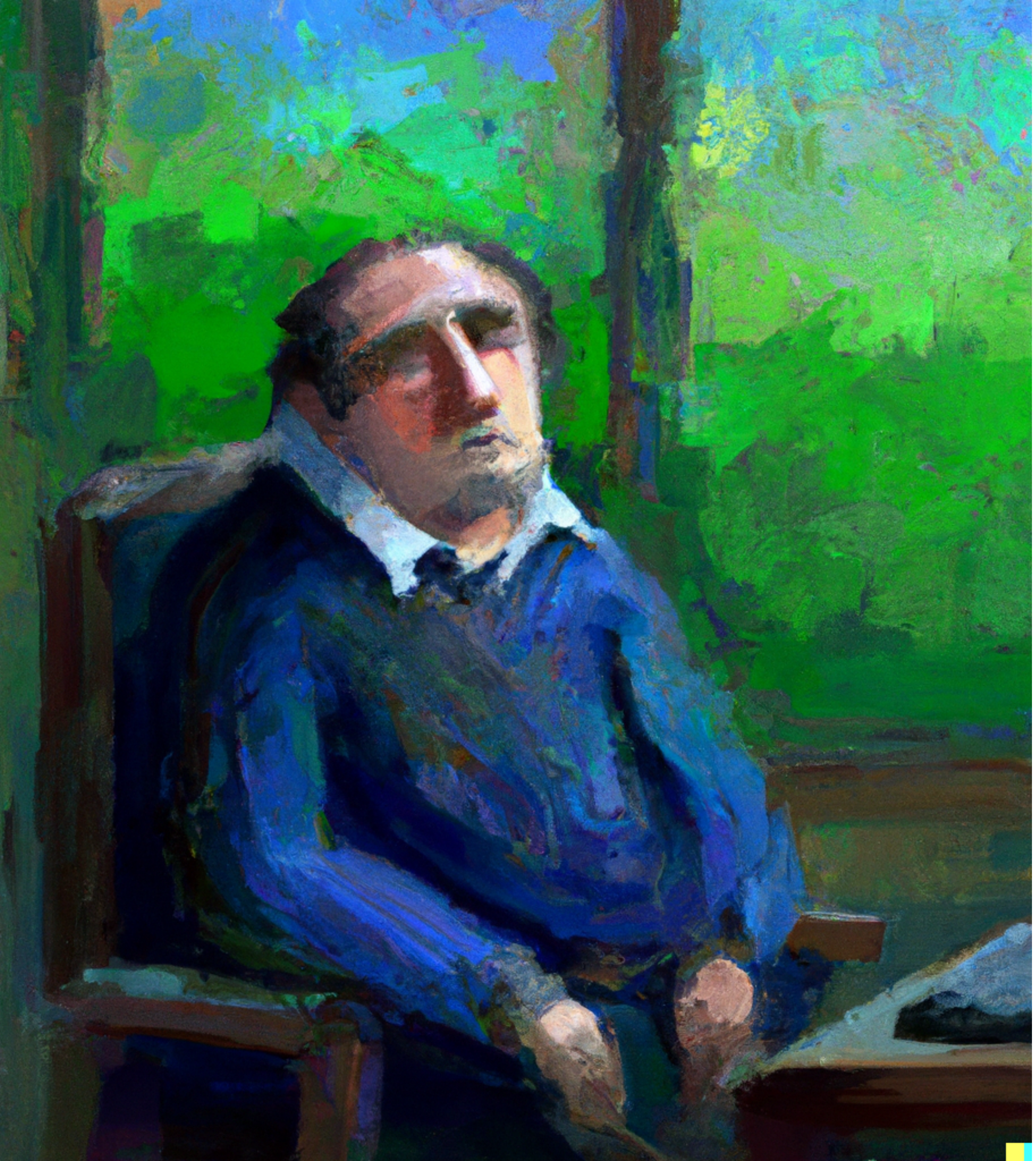


The R-Learner: Hyperparameter tuning



The R-Learner: Hyperparameter tuning





The R-Learner: Hyperparameter tuning

- We can expect roughly a tripleing of runtime due to not being able to train and reuse component models in isolation.
- This is even amplified when trying to use a particular component model for other MetaLearners, too.

P^3 #3: Distinct covariate sets

- Different covariates for different components
  - E.g. we know that the treatment effect is only a function of stirring while the base outcome is a function of many features.
- Different covariates for different treatments
  - E.g. assume we have the following treatment variants:

  i. No stirring

  ii. Stirring for 20'

  iii. Stirring for 40'

  - where the second and third variant have the additional covariate of the spoon type. We would like to use that covariant to capture heterogeneity, but can't specify that we only have it for specific variants.
- These features are not at all supported by EconML and CausalML.

And more...

- DoubleML: Biased final stage
- Tricky to combine cross-fitting with further cross-splitting (e.g. super learning or splits) -> also an engineering problem (e.g. multiprocessing)
- Read out treatment effects of categoricals when using DML

quantco

Do you also prefer Causal Inference and critizing libraries over doing pretty slides?

Join us :)

https://www.quantco.com/

ENGINEERING

Cloud Site Reliability Engineer
EUROPE ENGINEERING HYBRID
APPLY

Deep Learning Engineer
EUROPE ENGINEERING FULL-TIME HYBRID
APPLY

Developer Productivity Engineer
EUROPE ENGINEERING FULL-TIME HYBRID
APPLY

Software Engineer
EUROPE ENGINEERING FULL-TIME HYBRID
APPLY

Software Engineering Intern
EUROPE ENGINEERING INTERN ON-SITE
APPLY

DATA SCIENCE

Data Science Intern (Causal Inference/Machine Learning)
EUROPE DATA SCIENCE INTERN ON-SITE
APPLY

Data Scientist (Causal Inference/Machine Learning)
EUROPE DATA SCIENCE FULL-TIME HYBRID
APPLY

Data Scientist | Commercial Insurance Pricing
COLOGNE, NORTH RHINE-WESTPHALIA DATA SCIENCE FULL-TIME HYBRID
APPLY

Medical Physicist - Virdx
ESSEN, NORTH RHINE–WESTPHALIA DATA SCIENCE FULL-TIME HYBRID
APPLY

Quantitative Researcher
GLOBAL DATA SCIENCE FULL-TIME HYBRID
APPLY

Acknowledgements

- Matheus Facure: Causal Inference for the Brave and True
- Matthias Lux, Norbert Stoop