Causal Machine Learning for Counterfactual Evaluations

Marcel Tkacik

PhD student @ VSE, Visiting Researcher @ WU (till September),
Visiting Researcher at IDC, Herzliya (September – October)

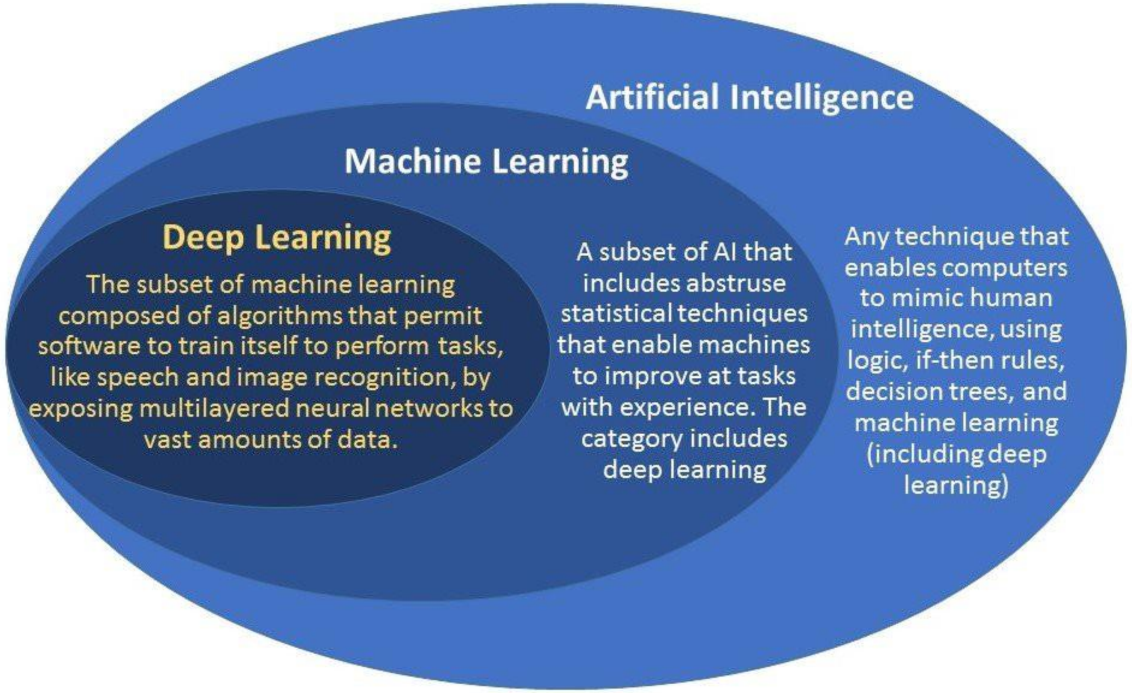
Outline

1. What is Machine Learning (ML)
2. Why ML has superior predictive performance
3. Why is ML not ready-made for causal inference
4. Causal ML models: double/debiased machine learning, causal forest

What is Machine Learning (ML)

- **Economics definition**
  - In very simple terms: fitting function between dep. var. and indep. Var.
  - What’s new: the parameters of the function are optimized to increase the out-of-sample predictive performance (and not in-sample)
  - The process of estimating the parameters is called learning
  - ML field has its own jargon, but most of the terms are analogous to the terms we already use!
- **Computer science definition**
  - Computer algorithms that improve automatically through experience and by the use of data

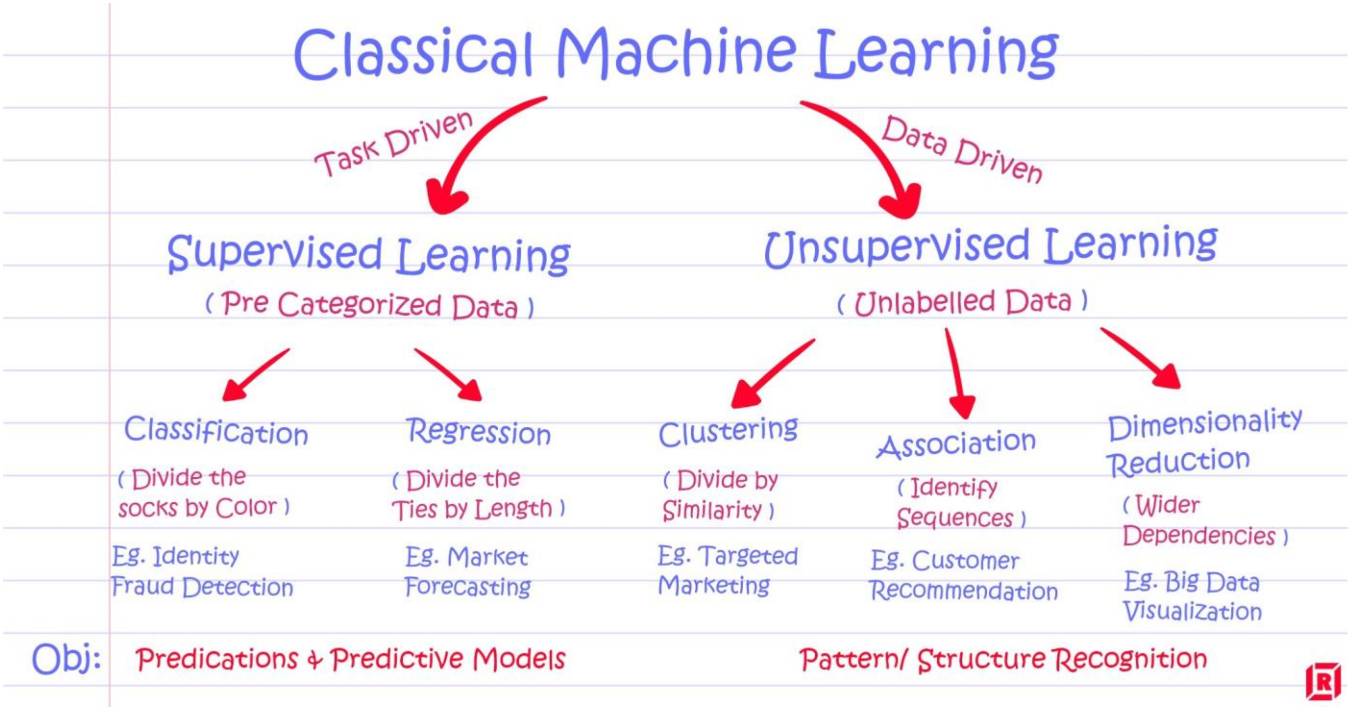
Where did this ML thing come from



Artificial Intelligence
Machine Learning
Deep Learning
The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.
A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning
Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)

Pic credit and nice explanation: https://medium.com/@harish_6956/what-is-machine-learning-deep-learning-7788604004da

The world of ML

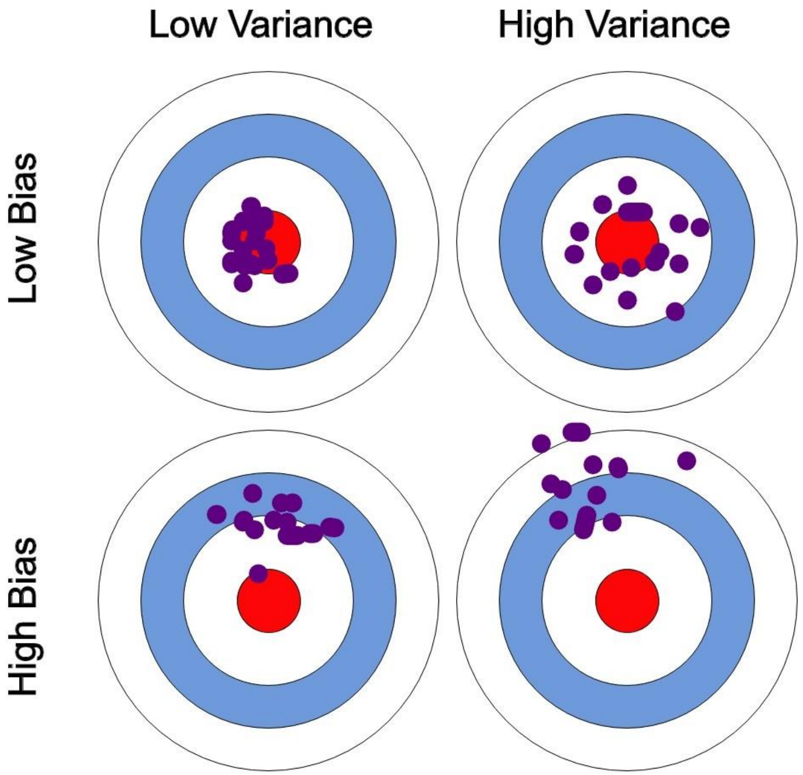


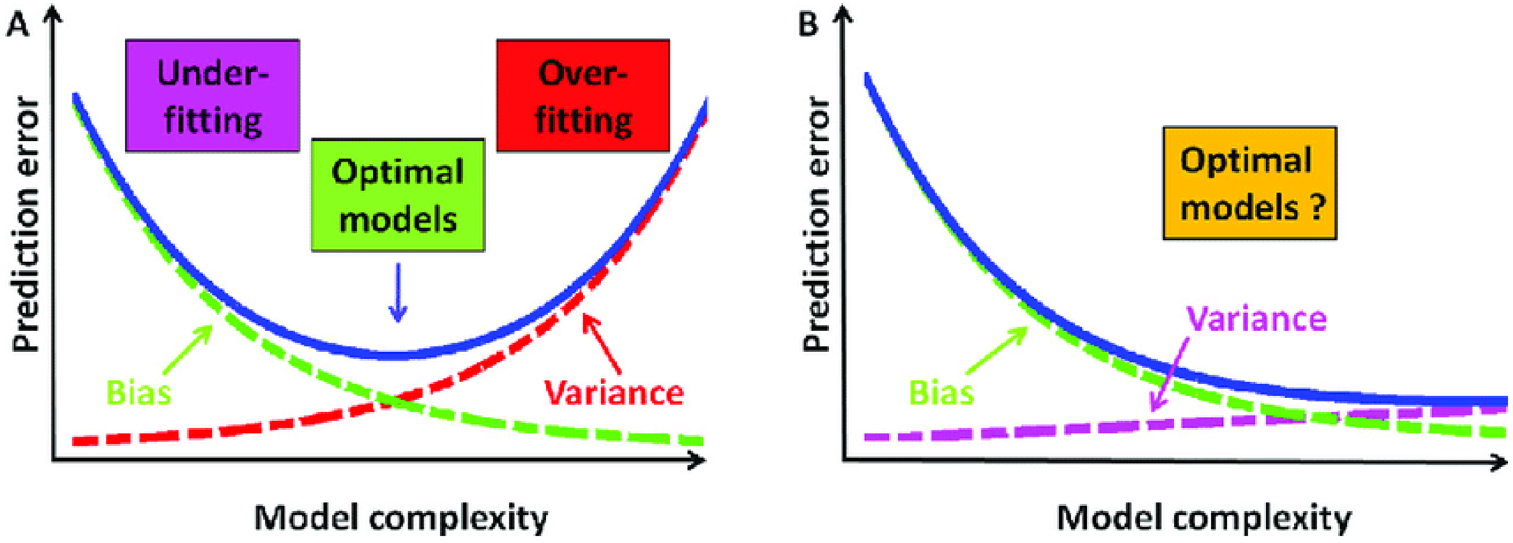
Pic credit and nice explanation: https://medium.com/@recrosoft.io/supervised-vs-unsupervised-learning-key-differences-cdd46206cdcb

Why are ML predictions better?

- Jargon: target = dep. var, features =indep.vars, learning=estimating
- Most used ML algorithm is.. A regression model
- Contrary to common practice – playing with model selection by testing p-values of different variables in the reg. model, ML always takes all available variables
- ML enhances OLS with two add-ons:
  - **Train-test sample split:** the model is trained on one part of the total data (usually 70%) and the parameters of the model are adjusted to increase the predictive performance of target variable in the test set (30%)
  - **Penalization of high variance** in the model by regularization methods (next slide): performs automatic variable selection (LASSO).

Regularization – trading off the variance and bias





Typical regularization methods: LASSO (L1), Ridge (L2), ElasticNet (L1+L2)

- Note that both train-test sample split and regularization by definition increase the external validity of your results! See Varian (2014) here: https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

That sounds good.. Where is the problem?

- This is all very good for business applications where you only care about predictions and not the causal structure
- BUT when we ask causal questions, we are usually interested in one parameter of interest – the treatment effect!
- Introducing the bias also biases the estimate of the treatment effect, the TE can be already biased
- This means that the estimate of TE is not “true”, not even statistical inference works (the CIs are not valid, p-values are off, bootstrapping won’t consistently solve this)
- Hence, for a long time the proper identification of causal effect (TE) was deemed unsolvable issue under regularization
- However, three years ago a solution was found!

Double/debiased ML

- Chernozhukov et al. (2018): **Double/debiased machine learning for treatment and structural parameters**, https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097
- See a popular article with nice explanation: https://medium.com/data-science-at-microsoft/causal-inference-part-2-of-3-selecting-algorithms-a966f8228a2d
- Double machine learning is a method for estimating treatment effects when all potential confounders are observed
  - BUT are either too many for classical statistical approaches to be applicable (curse of dimensionality, some vars are going to be significant only by accident)
  - OR their effect on the treatment and outcome cannot be satisfactorily modeled by parametric functions (i.e. the unconfoundedness assumption doesn't hold) => the treatment selection is not random (which is more like a norm in observational studies than an exception, exclusion restriction usually only holds in real RCTs)
- What it does is that in the first stage you run regularized regression and in the second stage you debias the estimate of the treatment effect by predicting the treatment from the controls (because you estimate twice, that's why it's called double ML). This debiases the TE both from bias induced by the regularization and the confoundedness of non-random treatment assignment
- This approach has been proved to provide valid ATE estimates, CIs and p-values

Causal Forest

- Wager & Athey (2017): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- The name of the method is IMHO little bit misleading
- What it does is that it allows us to study the heterogeneity of treatment effect
- In most cases, in observational studies we are interested in ATE, or ATT, but the heterogeneity of TE is rather a norm than the exception
- Studying the heterogeneity of TE is important – it shows us how different subgroups respond to the treatment

Heterogeneity of TE

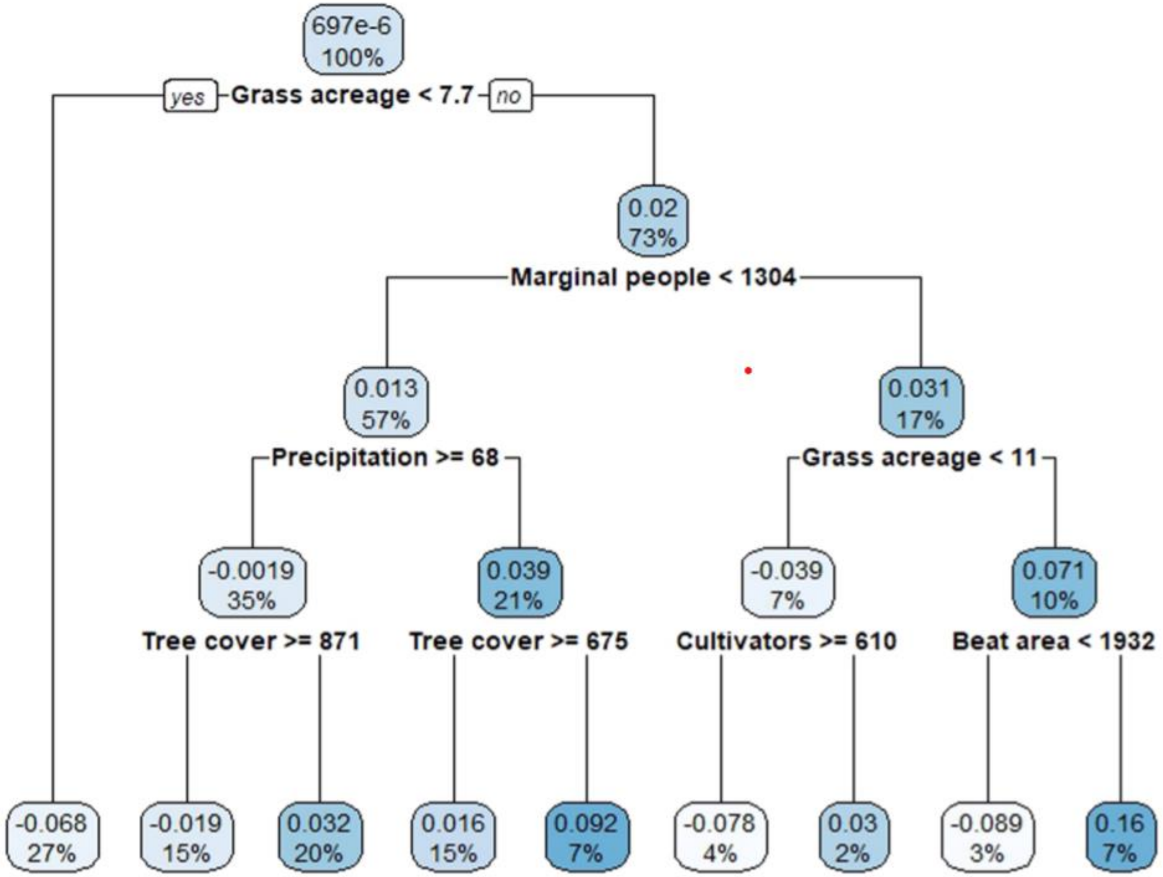
- Before causal forest, you would estimate local ATE by estimating conditional ATE (ATE conditioned on a membership in subgroup)
- A challenge with this approach is that you have to define the subgroups and each subgroup may have substantially less data than the study as a whole, so if the study has been powered to detect the main effects without subgroup analysis, there may not be enough data to properly judge the effects on subgroups.
- Causal forest overcomes this by fitting a random forest algorithm (ensemble of decision trees) – this defines the subgroups automatically

Causal Forest – an application

- Case study: microcredit in Morrocco
- Who benefits from microcredit the most? The highest local ATE is found on young households with children aged 6 – 16, https://towardsdatascience.com/causal-machine-learning-for-econometrics-causal-forests-5ab3aec825a7
- **Case study: Machine learning to analyze the social-ecological impacts of natural resource policy: Insights from community forest management in the Indian Himalaya**
- Dep. Var: NDVI (normalized difference vegetation index)
- Treatment: community forest management
- The authors found no average TE of community forest management on the NDVI
- HOWEVER they found significant local ATEs, hence for policy it has important implications – target the treatment

NDVI – which areas responded the most to the treatment

a. CPM Causal Tree



Deep Instrumental Variables

- Aim: Observational studies with confounders
- IV = tackle causality by identifying the sources of treatment randomization that are conditionally independent from the outcomes (this source is called an instrument)
- Deep IV is a method that trains deep networks to minimize the counterfactual prediction error and validate the resulting models on held-out-data
- Hartford et al (2017): Deep IV: A Flexible Approach for Counterfactual Prediction, http://proceedings.mlr.press/v70/hartford17a.html
- Still waiting for additional validation results

How to use all these methods

- **Python:** Microsoft implements all these methods in their new Causal Inferene and EconML Python libraries: https://github.com/Microsoft/EconML, https://github.com/microsoft/dowhy
- **R:** Causal Forest is implemented in grf package (https://cran.r-project.org/web/packages/grf/grf.pdf), DML is implemented in DoubleML package (https://github.com/DoubleML/doubleml-for-r)

Thank you for your attention!

This presentation is available at my website:
http://marceltkacik.me/research