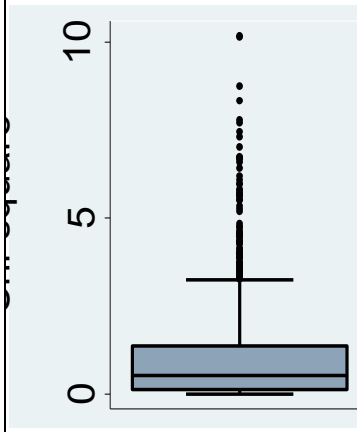
Biostat 202: Opportunities and Challenges of "Big Data" - Causal Inference in 'Big Data'

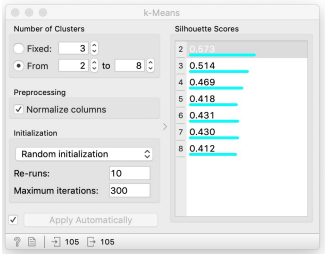
Aaron Wolfe Scheffler,
Division of Biostatistics,
Department of Epidemiology and Biostatistics

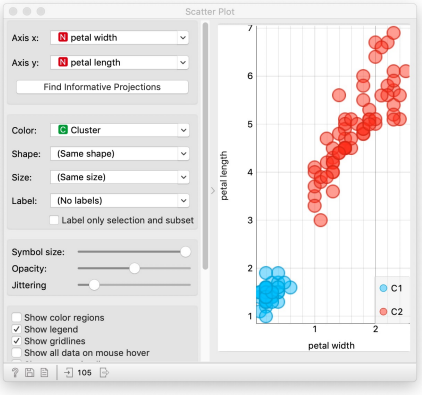


Review

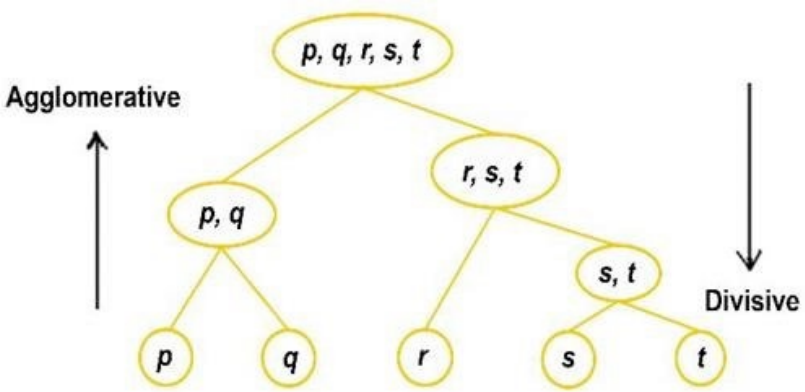
Clustering

1. k-means



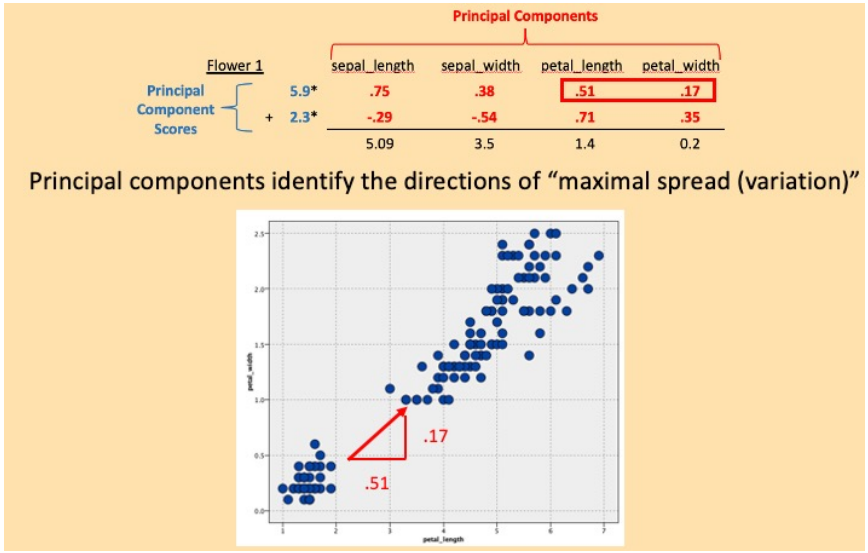


2. heirarchical

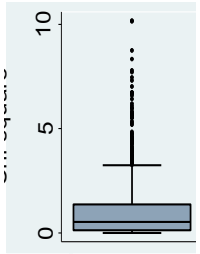


Dimension Reduction

3. PCA



4. t-SNE



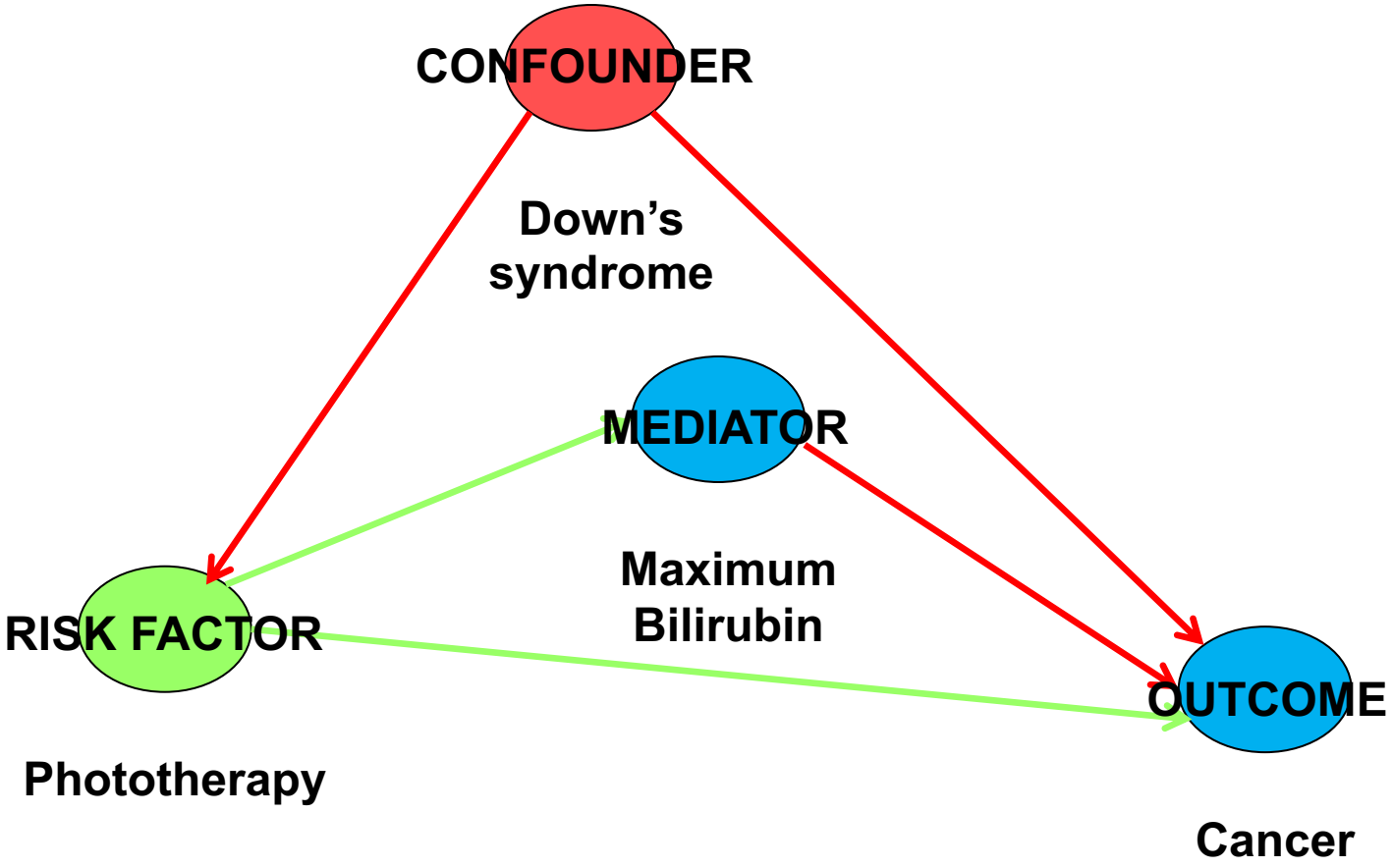
Outline

- Introduction: prediction versus causation
- Causal methods in (very) brief
- Some uses of big data and machine learning in causal inference
- Summary

Prediction versus causality

- Data-mining and machine learning algorithms are only based on associations.
- Often interested in causation, i.e., understand the cause of illness or develop interventions that cause improvement.
- Correlation does not imply causation
  - https://www.tylervigen.com/spurious-correlations
- Often senior researchers have lapses and over-interpret correlations causally
- The (social) media frequently confuses causation and correlation

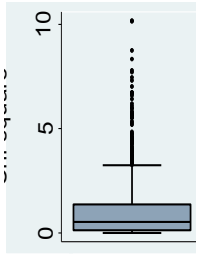
Causal inference – the 5¢ tour



Adapted from Newman, et al Pediatrics, 2016

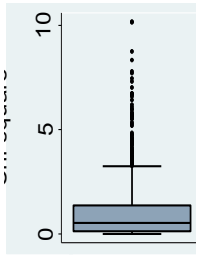
Biostat 215 covers these topics in detail

Which of these are prediction versus causation questions?



1. Effect of Abaloparatide on fractures in postmenopausal women.
2. Estimating fetal risk of a disease using genetic screening.
3. Effect of cerebral protection device on brain lesions following aortic valve implantation.
4. Estimation of temporal trends in preterm birth rates and their association with obstetric interventions.
5. Survival for children with trisomy 13 and 18.

Which of these are prediction versus causation questions?



1. Effect of Abaloparatide on fractures in postmenopausal women.
2. Estimating fetal risk of a disease using genetic screening
3. Effect of cerebral protection device on brain lesions following aortic valve implantation.
4. Estimation of temporal trends in preterm birth rates and their association with obstetric interventions .
5. Estimation of survival for children with trisomy 13 and 18.

Potential outcomes

- Imagine a hypothetical experiment in which you get to observe each participant under both the treatment and control conditions holding all else the same, like a perfect cross-over experiment.

$$Y_{trt}, Y_{ctl}$$

- Often, we only get to observe one of Y_{trt} or Y_{ctl}, depending on whether the participant is in the treatment or control condition.
- Sometimes described as the **"fundamental problem in causal inference."**

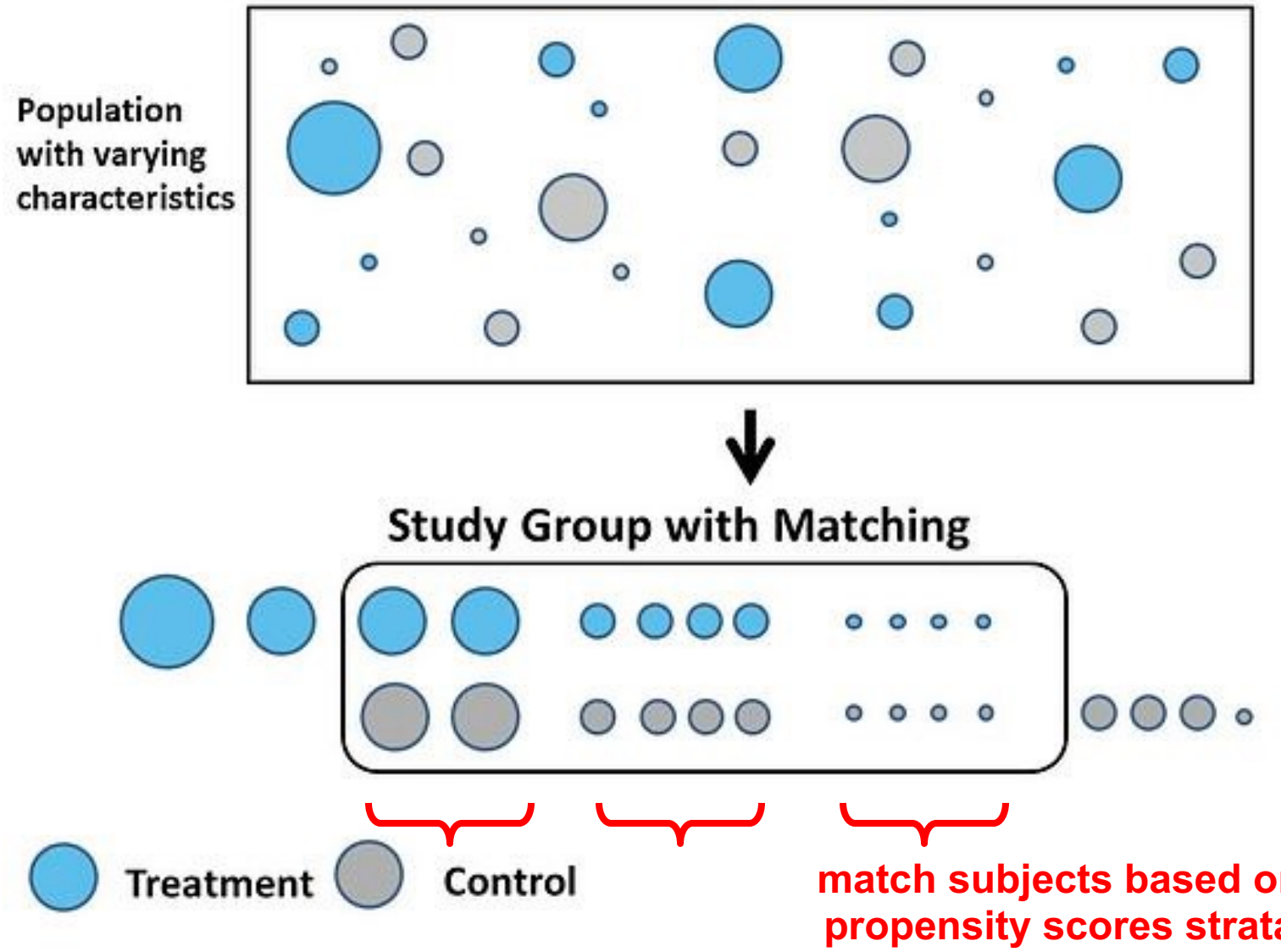
Potential Outcomes – Average Causal Effect

- Often, a reasonable quantity to estimate is the average causal effect (ACE): the average of the individual causal effects across the entire population.

$$\text{ACE} = \frac{1}{n}\sum_{i=1}^{n} Y_{trt\ i} \_ Y_{ctrl\ i}$$

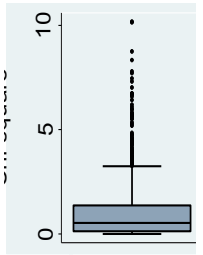
- Nice in theory but you cannot directly observe the individual causal effects! E.g. you cannot observe someone smoke and not smoke over their lifetime.
- So to estimate causal effects, we'll need to make some assumptions

E.g. Propensity score matching



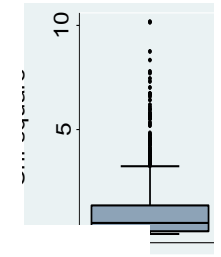
https://www.summitllc.us/propensity-score-matching

Propensity score estimation using Orange

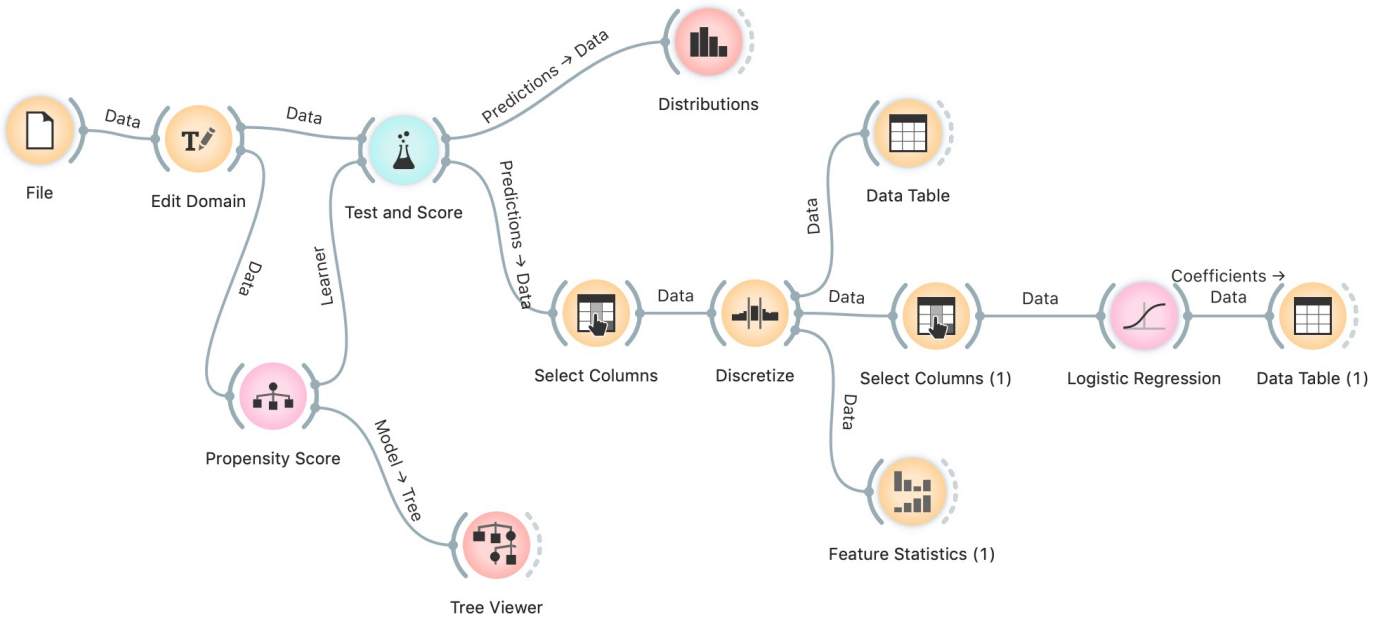


- Often divide propensity scores into quintiles in order to adjust.
- For example, bin the propensity scores into five equal categories using a Bin Node (bin method = Tiles (equal count), quintiles) and then use a logistic regression node (for a binary outcome) or a linear regression model node (for a numeric outcome).
- The only predictors are the propensity score bins (as a nominal variable) and treatment.

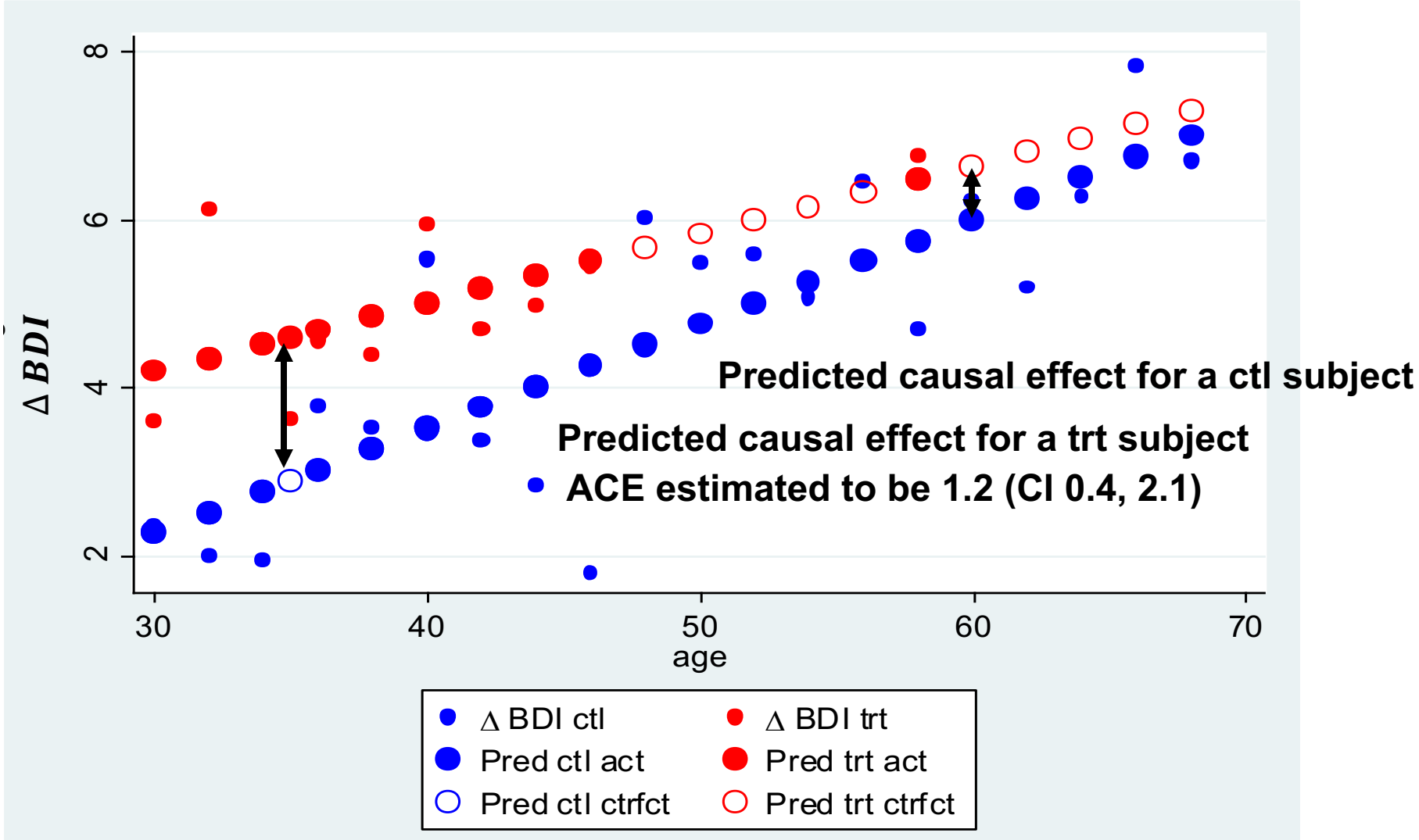
Propensity score example



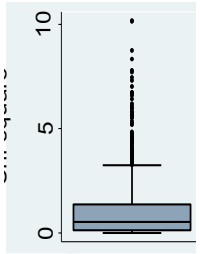
I have posted an example Orange workflow example and dataset on the website.



Potential outcomes estimation



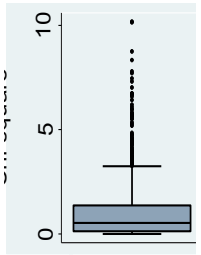
Potential outcomes estimation using Orange



- Create duplicate rows of data with the treatment assignment switched and missing outcome data:

Patient_ID	Sex	Treatment	Systolic_BP	…other predictors
1	F	1	120	
2	M	1	135	
3	F	0	142	
…				
1	F	0	missing	
2	M	0	missing	
3	F	1	missing	

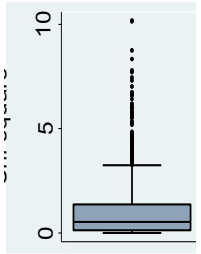
Potential outcomes estimation using Orange



- Want to end up with a predicted outcome under treatment and under control:

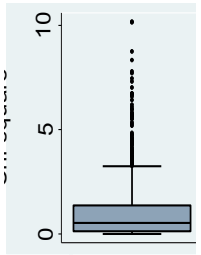
Patient_ID	Sex	Treatment	Systolic_BP	Pred_under_Trt	Pred_und_Ctl
1	F	1	120	122	135
2	M	1	135	137	142
3	F	0	142	137	140

Potential outcomes estimation using Orange



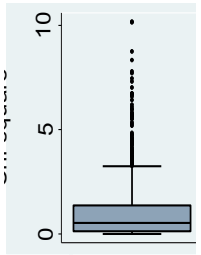
- Read the data using a File widget
- Read in a duplicate version of the data using another File widget. In this part of the stream, switch the coding of treatment and control and recode the outcome variable to missing.
- *Concatenate* the duplicate data to the original data. So now the dataset is twice as large as the original and has a copy of every person under control and under treatment (though only one of them has the outcome – the other is missing).

Potential outcomes estimation using Orange



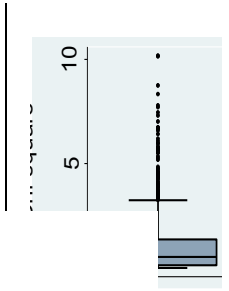
- Separate the data into data for the treatment group and control groups.
- Fit your favorite machine learning method to predict the outcome in the treatment group and do the same for the control group. This will give predictions for the data that were missing previously.
- *Merge* the datasets together so everyone has a predicted outcome under the treatment and under the control conditions.

Potential outcomes estimation using Orange

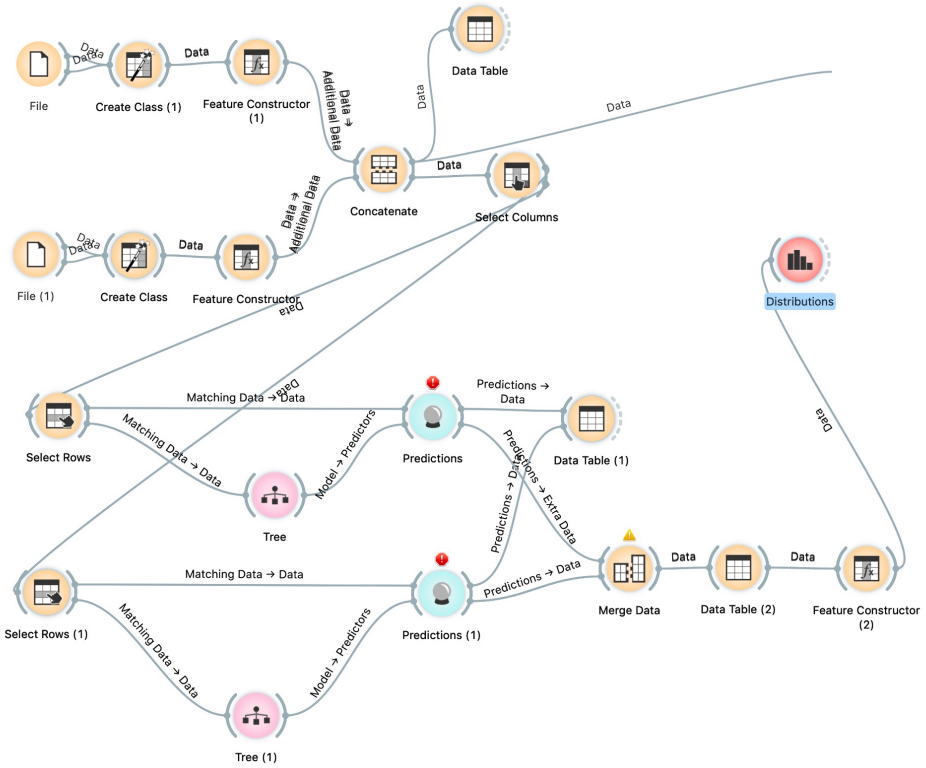


- Average all the differences between the under-treatment and under-control conditions to estimate the average causal effect (ACE).
- Or the ACE in different subgroups.

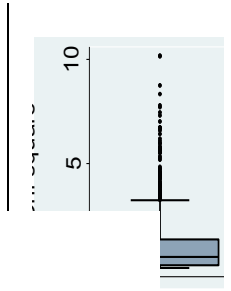
Potential outcomes estimation



I have posted an example Orange workflow and dataset on the website illustrating the idea for this and the next section (subgroup analysis).

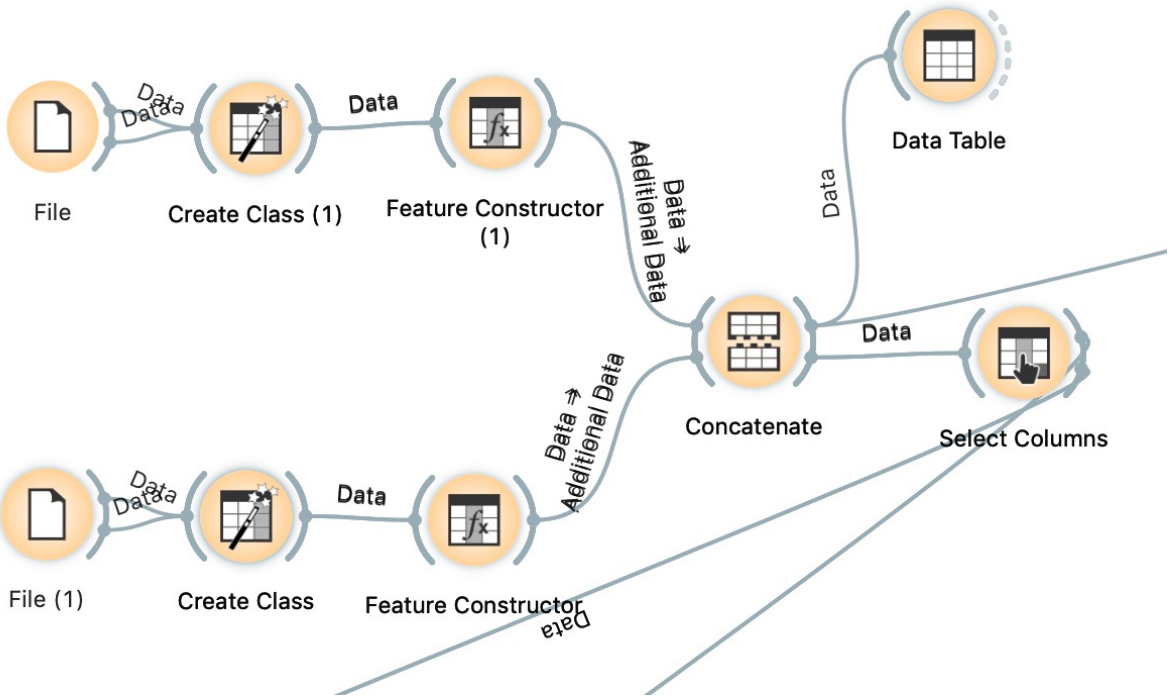


Potential outcomes estimation

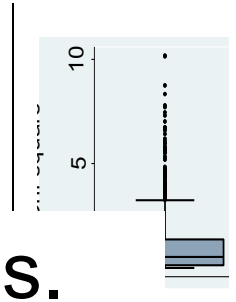


Looks complicated, but let's break it down.

1. Create duplicate records

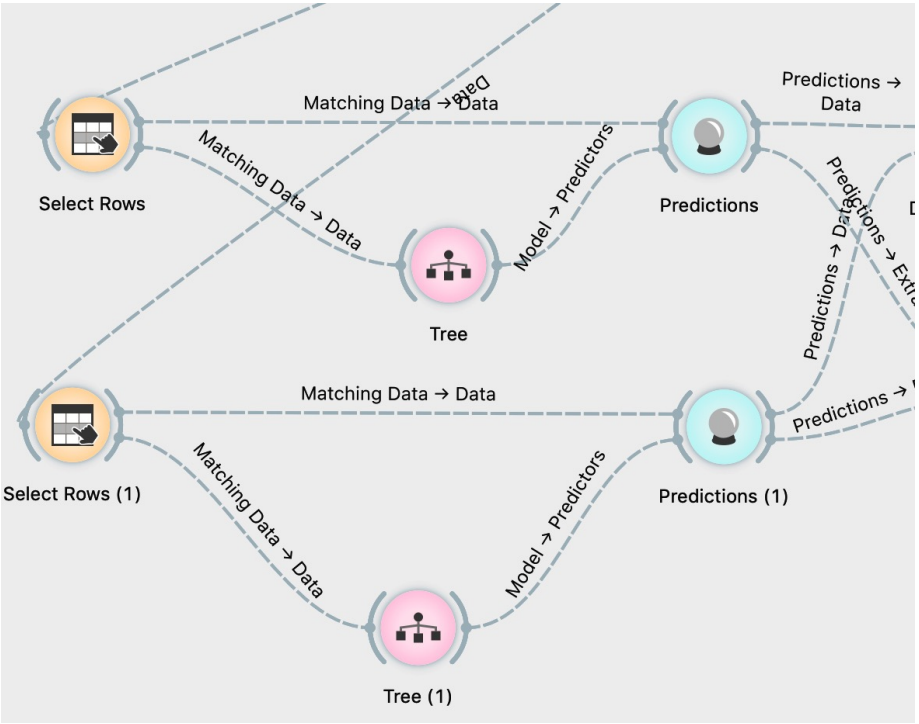


Potential outcomes estimation



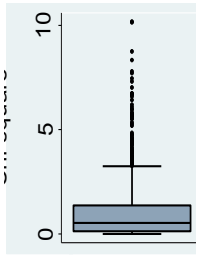
Looks complicated, but let's look at the steps.

2. Estimate effects in treatment and control



3. (Next) Use potential outcomes to get ACE

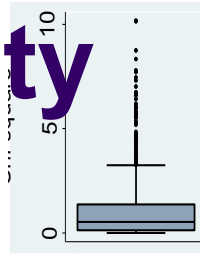
Potential outcomes estimation using Orange



Issues

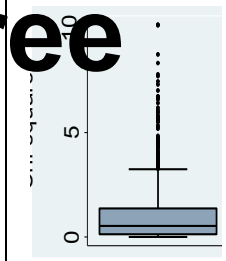
- (+) Can use flexible models in each condition to help avoid issues with incorrectly specifying the model.
- (-) Requires large sample sizes in treatment and in control groups.
- (0) Does not solve the problem of "unmeasured confounders". Not surprisingly, you can only adjust for measured quantities.

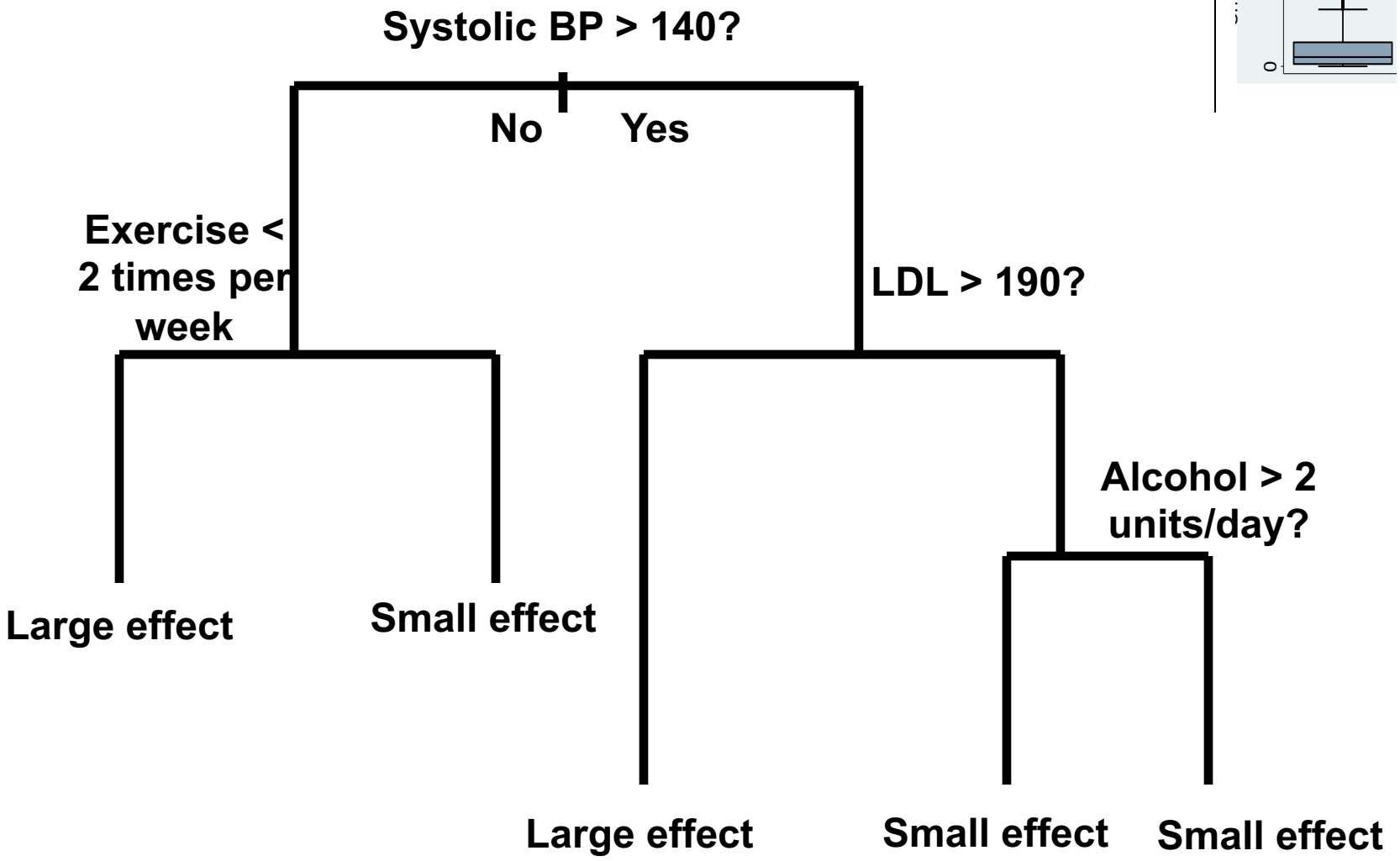
Potential outcomes vs propensity score adjustment



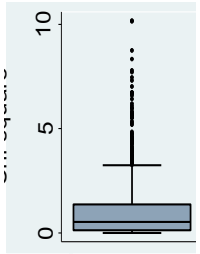
- Potential outcomes: try to **estimate** the potential outcome using supervised learning
- Propensity score adjustment: try to construct **balanced** groups
- Both rely on models!!! (not a silver bullet)
- Can be used alone or in tandem ("doubly robust")
- Neither helps us extrapolate to samples that do not represent the training data

Subgroup analysis: decision tree





Summary



- Distinguish questions by whether they can be answered by prediction only or require more detailed analysis.
  - effect, intervention = causal,
  - estimate, association, correlation = prediction
- Machine learning methods are not designed for drawing causal conclusions.
- However, *aspects of* causal analyses may be prediction problems for which we can use the methods in the class
  - Estimating propensity scores and predicting potential outcomes (e.g. regression estimation)
  - Identifying subgroups on the basis of different estimated treatment effects