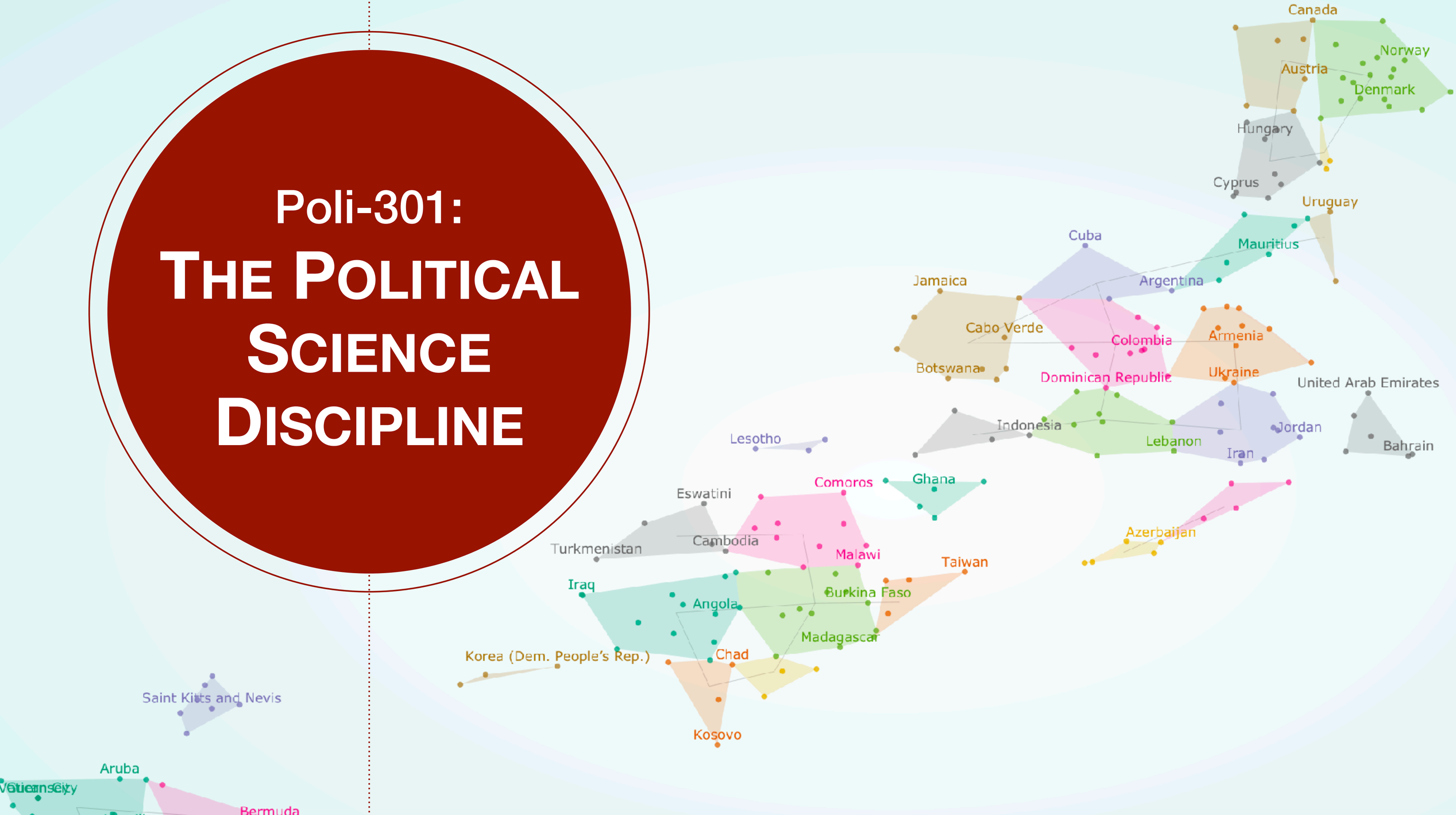
Poli-301:
The Political Science Discipline



TODAY'S AGENDA

1. Wrap-up fixed effects
2. Matching (Coarsened Exact Matching)

Last week

We want to close backdoors from X to Y by controlling Z

We can't measure Z, or there might be many Z's

We can use fixed effects to control for differences *across units,* even if we don't know Z

Caveat: only controls for Z's that are *constant within the unit*

Example

Can reading fiction (for fun) help students get better grades?

name	sex	race	income	grade	reading(min)	bullied	age	weather	masters	last_teach	tutoring
Boise	Female	White	1.83	88.8	35	no	12	80.2	yes	Kujuan	no
Boise	Female	White	1.83	85.3	41	no	12	80.9	no	Calii	no
Boise	Female	White	1.83	85.5	52	no	14	83.2	yes	Wilhem	no
Charilyn	Male	Hispanic	2.06	83.9	28	no	10	80.6	no	Jenevy	yes
Charilyn	Male	Hispanic	2.06	89.7	37	yes	10	79.5	yes	Loralea	yes
Charilyn	Male	Hispanic	2.06	88.5	41	no	13	86.9	yes	Robret	yes

What should the DAG look like?

Varies within vs. varies across

			Constant						Varies	Varies	
name	sex	race	income	grade	reading(min)	bullied	age	weather	masters	last_teach)	tutoring
Boise	Female	White	1.83	88.8	35	no	12	80.2	yes	Kujuan	no
Boise	Female	White	1.83	85.3	41	no	12	80.9	no	Calii	no
Boise	Female	White	1.83	85.5	52	no	14	83.2	yes	Wilhem	no
Charilyn	Male	Hispanic	2.06	83.9	28	no	10	80.6	no	Jenevy	yes
Charilyn	Male	Hispanic	2.06	89.7	37	yes	10	79.5	yes	Loralea	yes
Charilyn	Male	Hispanic	2.06	88.5	41	no	13	86.9	yes	Robret	yes
	Constant	Constant				Varies	Varies	Varies			Constant

Remember: we don't want to control for all variables; just the backdoors!

What does the FE do?

Fixed effect on student accounts for all those variables that are constant within student

E.g., even if we don't know someone's family income, we would control for it with the student FE…

…so long as it is *constant within students*.

Even for variables that vary within students (bullied), the FE will still control some of its effect

Today: matching

Matching is a method commonly used to evaluate policies

E.g.: some counties rolled out GOTV shuttle services; did it work?

The situation: some *treatment* has been applied to some people/states/groups/etc., but not others

Our goal is to compare the treated and untreated in a way that makes sense

Example: LRDP

Problem: many in Colombian countryside own land *informally*

Bought land, have lived on it for decades, but have no formal proof of ownership

Creates potential for conflict, abuse, under-investment

Colombia + USAID create program (LRDP) to increase formal ownership

Did the program: reduce conflict? Improve investment? Increase trust in local govt?

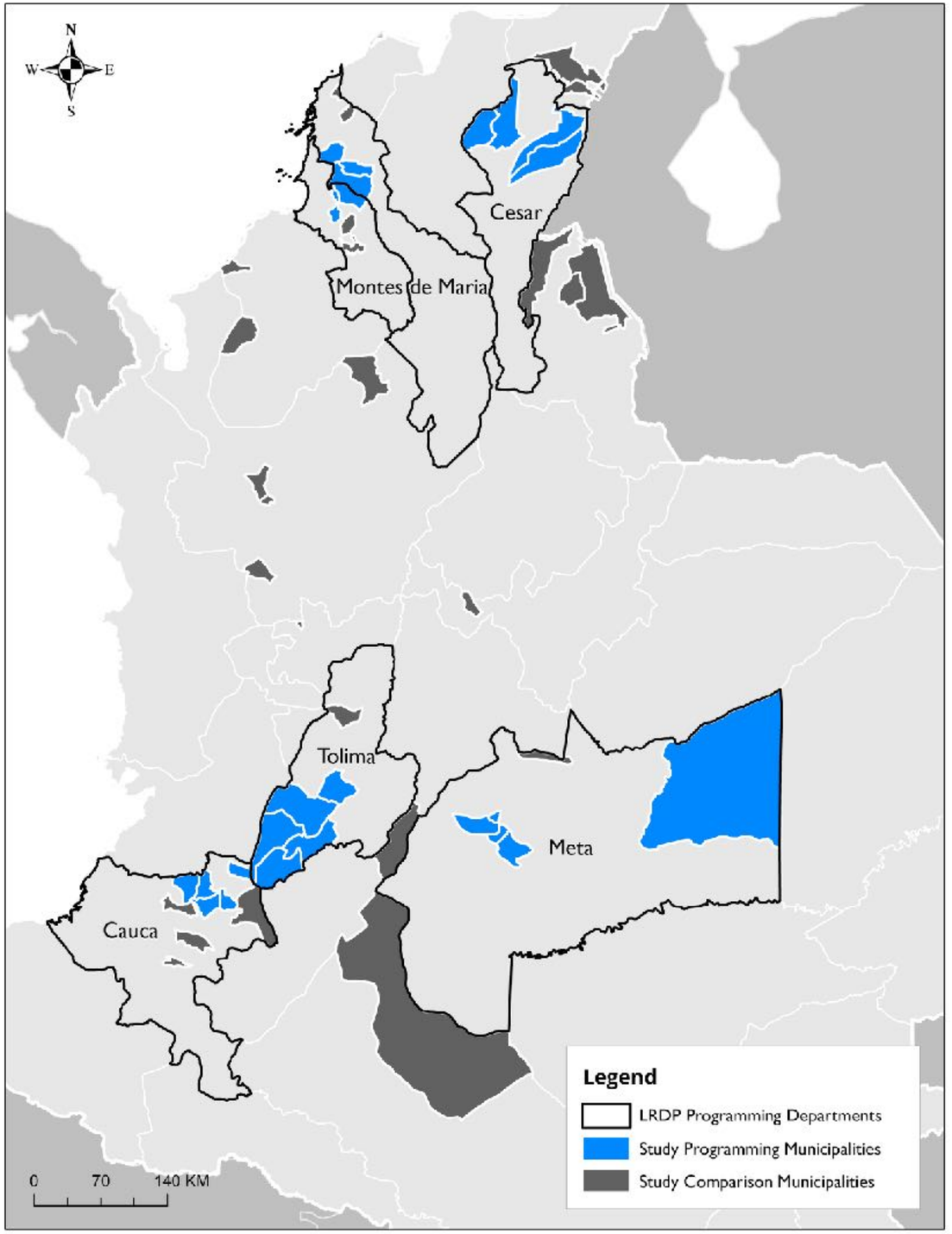
If LRDP had been randomly assigned we could just compare conflict levels in treated vs non-treated towns

But LRDP was not randomly assigned!

Implementers chose places with history of conflict, poverty; other ad-hoc decisions

What to compare the towns that got LRDP programming to?

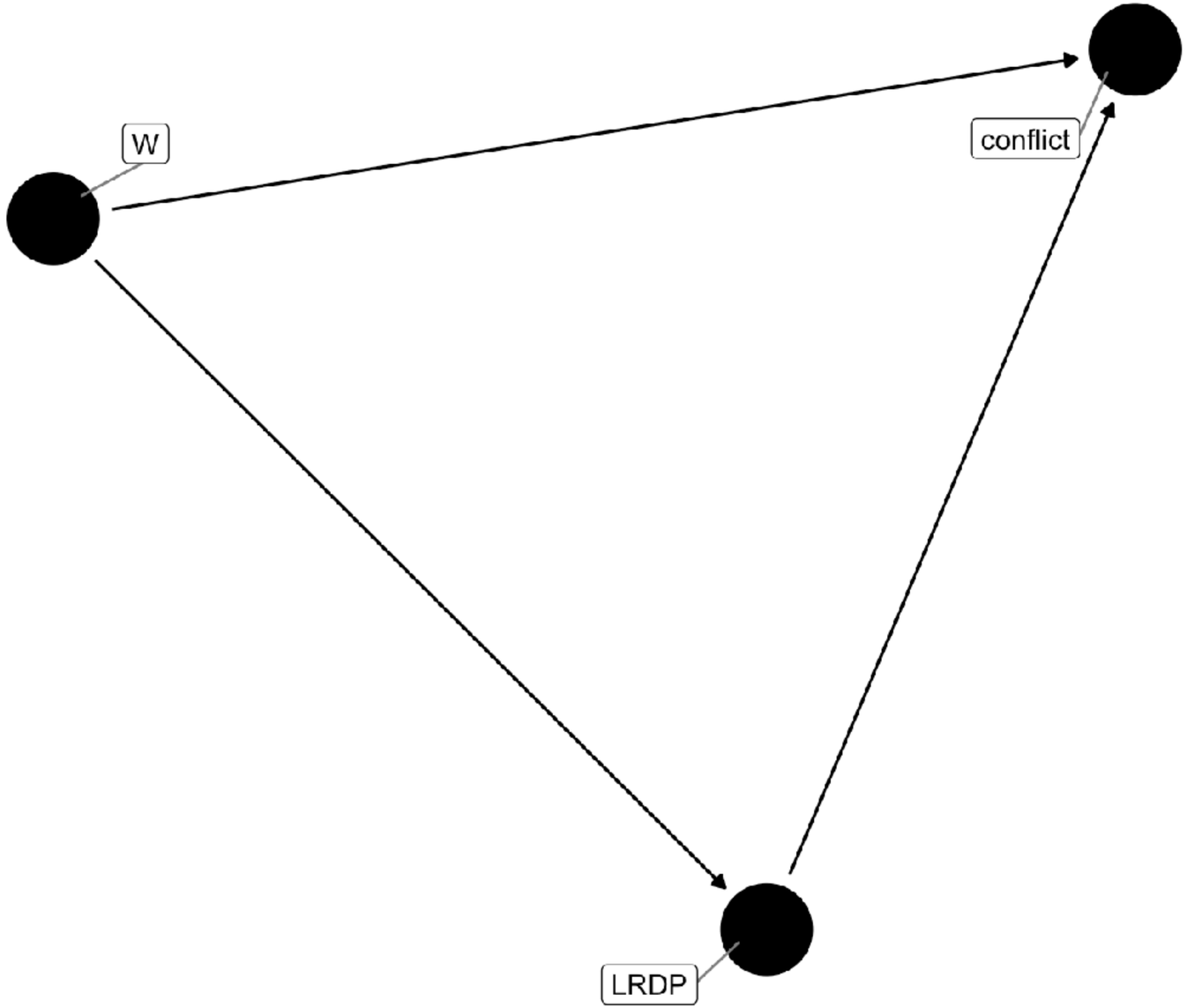
We used matching to identify comparable towns



The DAG

The worry here is there is a backdoor from LRDP to conflict

“W” might be:
poverty,
Prior conflict,
State presence,
Road quality,
Etc.



Call and response

“We find a negative correlation between being in LRDP and conflict”

“Well, that’s only because the places that got LRDP were already likely to have conflict. That’s why they were chosen!”

“Well, even if we ONLY look at towns with similar levels of past conflict, we STILL see a negative correlation”

Call and response II

“We find a positive correlation between having kids and engaging in extramarital affairs”

“Well, that’s only because people who have kids tend to be older, and older people tend to have more affairs!”

“Well, if we ONLY look at people of similar ages, we see that having children reduces likelihood of affair”

Apples to apples

Matching: pick towns that are very similar to one another *except some got treatment and others didn't*

Approach: look at our treated towns, look at their background characteristics (W), and pick non-treated towns that have similar levels of W

Why do we want background characteristics to be similar?

Remember the rats

One of the advantages of experiments is that we *know* there are no backdoors to X

One of the consequences of randomization is that other characteristics (not X or Y) are, on average, *balanced*

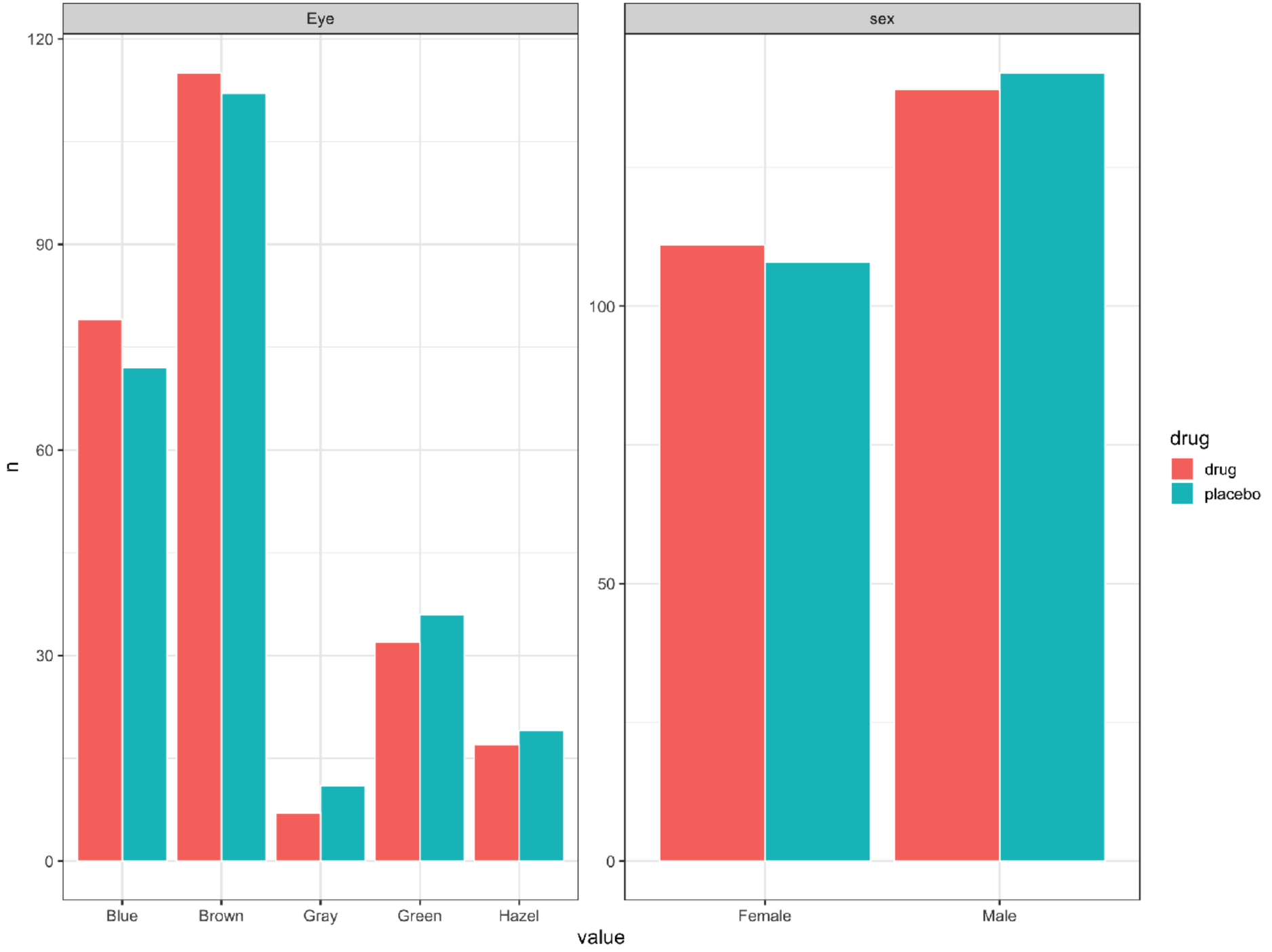
That is, apart from what drug they got and blood sugar levels, rats in treatment and control group look pretty similar

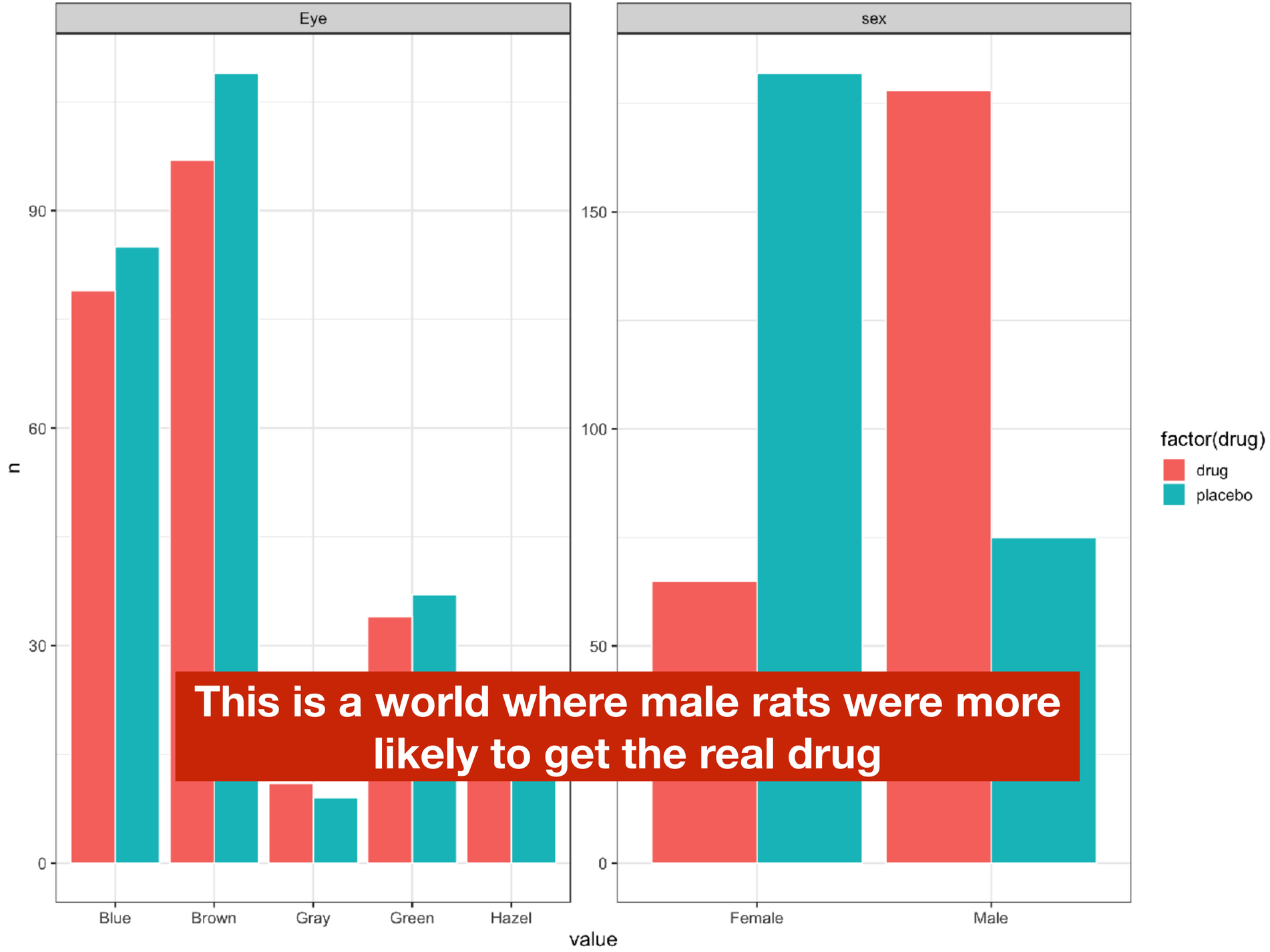
```
A tibble: 100 x 3
    rat drug    blood_sugar
  <int> <fct>         <I<dbl>>
1     1 drug          -2.97
2     2 placebo       -0.823
3     3 placebo        1.26
4     4 placebo       -1.22
5     5 drug          -2.53
6     6 drug          -1.94
7     7 drug          -0.934
8     8 drug          -1.83
9     9 placebo        0.785
0    10 drug          -1.91
```

Rats again

Table 1

rat	blood_sugar	drug	Height	Age	sex	Eye
Jenin	0.00268	drug	76	40	Female	Blue
Reigna	-0.418	placebo	68	83	Male	Brown
Madaline	-1.11	drug	71	68	Female	Blue
Jace	-0.833	placebo	71	62	Male	Blue
Lorali	-1.35	drug	67	80	Male	Blue
Leith	0.685	placebo	69	37	Female	Green


Eye
sex
120
90
60
30
0
n
100
50
0
Blue
Brown
Gray
Green
Hazel
Female
Male
value
drug
drug
placebo



This is a world where male rats were more likely to get the real drug

The search for doppelgängers

The basic premise of matching is that there are people/towns out there that didn't get treatment, but are otherwise nearly identical (*doppelgangers*)

If we can identify *otherwise similar* units that didn't get treatment, this is almost as good as having done an experiment

How to match

A ton of approaches to matching

We'll cover basic concept

And then do one version of it (Coarsened Exact Matching)

The steps

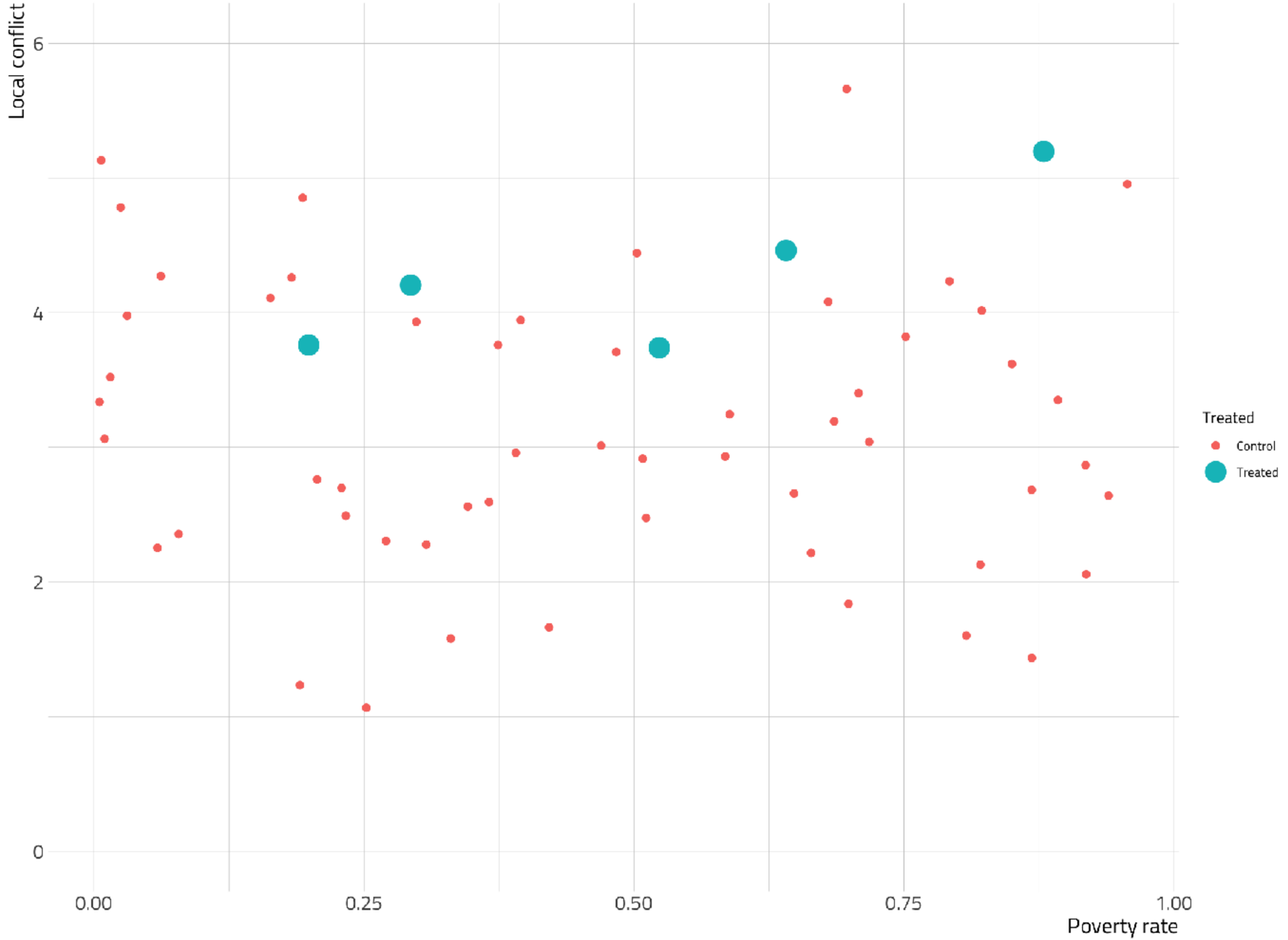
Pick set of variables to match on

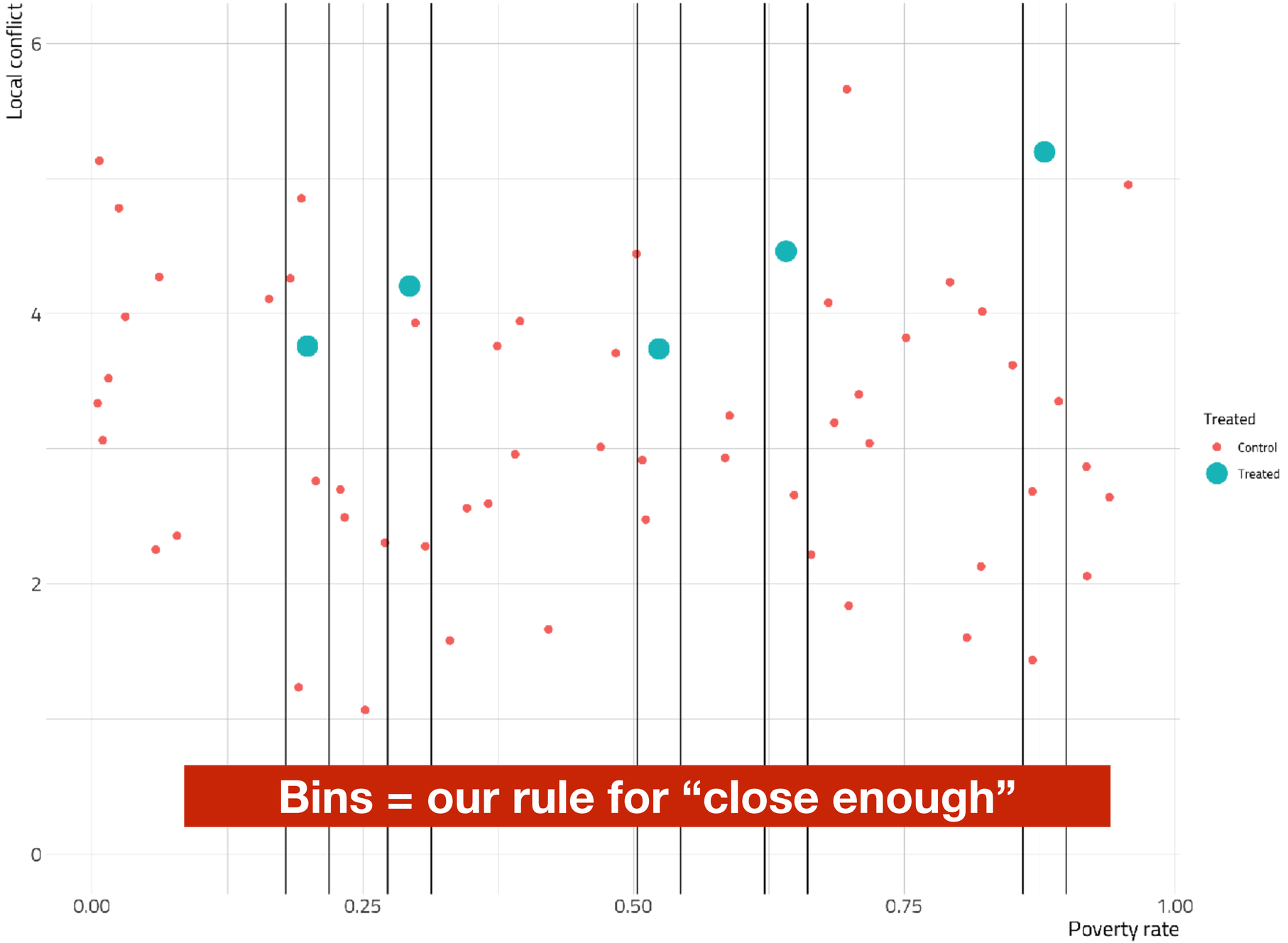
Separate out treated and untreated units

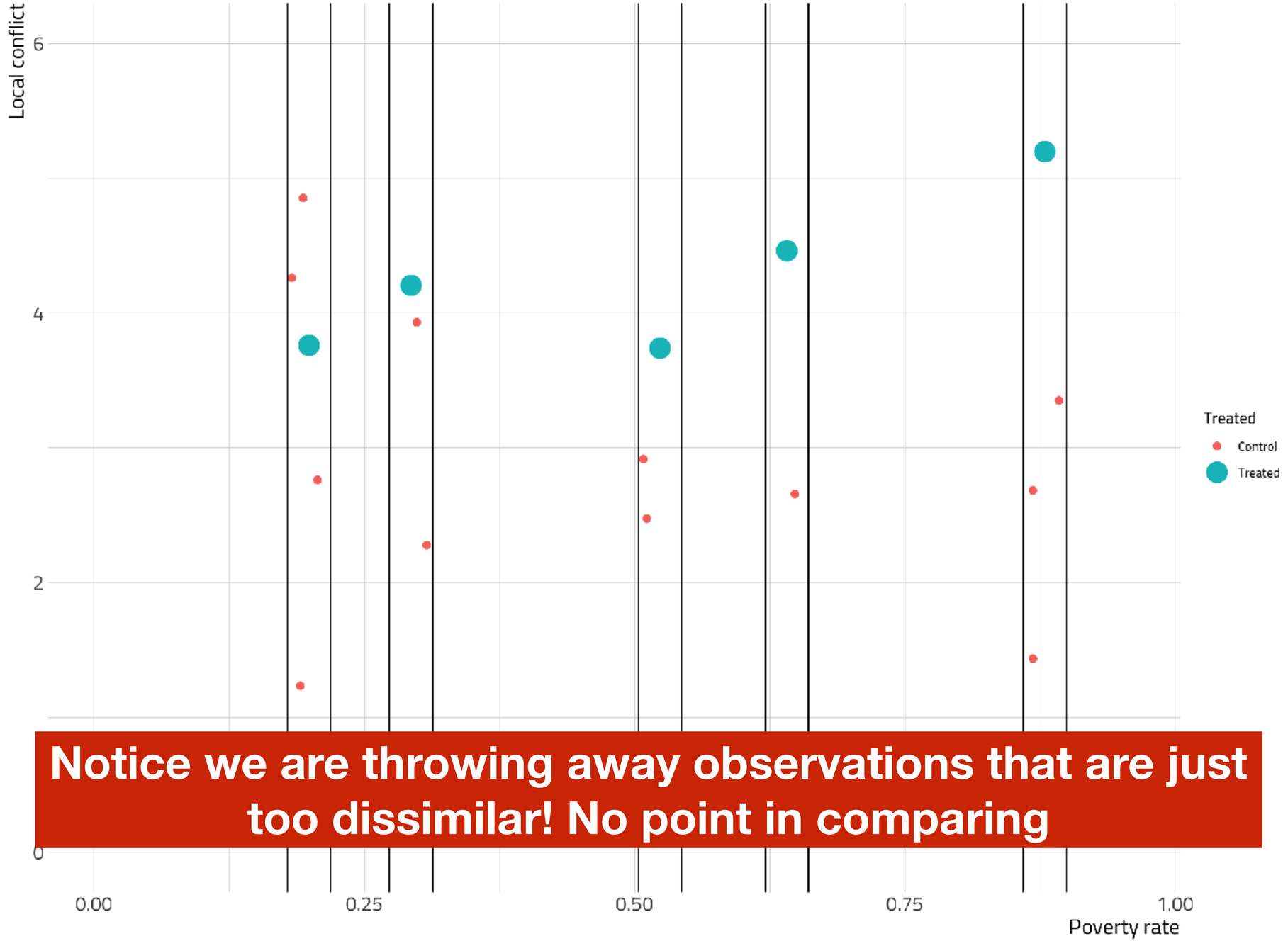
For each treated unit, check how similar each untreated unit is

Decide what counts as “similar”

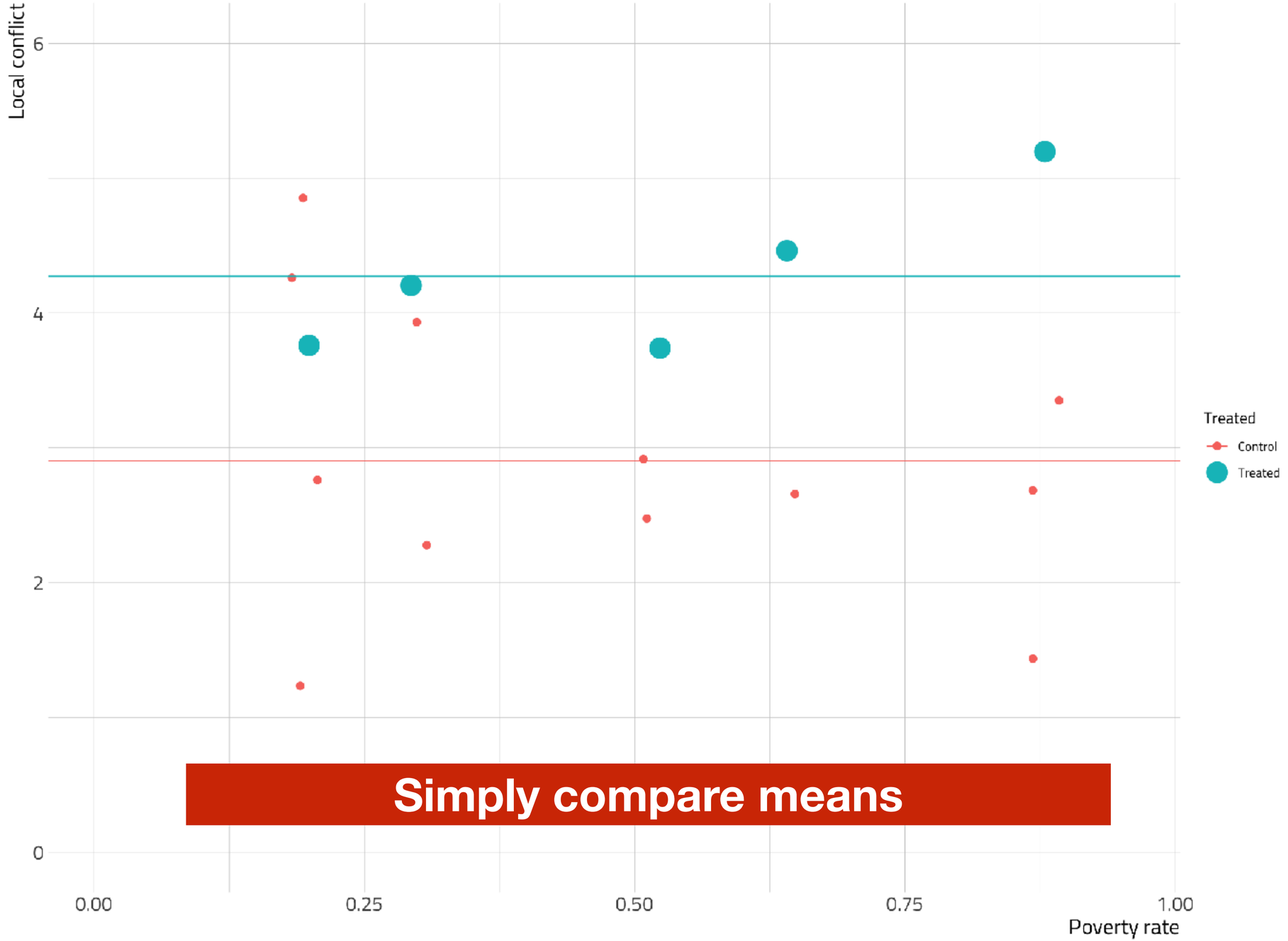
Compare the average treated Y vs. the average untreated Y, *counting observations more heavily the closer they are*


Local conflict
6
4
2
0
0.00
0.25
0.50
0.75
1.00
Poverty rate
Treated
Control
Treated


Local conflict
6
4
2
0
0.00
0.25
0.50
0.75
1.00
Poverty rate
Treated
Control
Treated
Bins = our rule for “close enough”



Notice we are throwing away observations that are just too dissimilar! No point in comparing


Local conflict
6
4
2
0
0.00
0.25
0.50
0.75
1.00
Poverty rate
Treated
Control
Treated
Simply compare means

Matching

Trivial to do in this example

But imagine doing this with multiple variables at the same time

Think too about drawing those “bins”, and how drawing them for one variable would impact a different variable

This is where matching really shines

Coarsened Exact Matching

Very popular approach to matching

Only count someone as a match if they are exactly the same on every matching variable

This is not possible with continuous variables, so you cut these variables into categories first ("coarsen" them)

Weight untreated observations so each treated observation is matched to the same number of treated observations

Follow along in R

Affairs example again

New libraries: Matchit
(and you will need to install cem as well)

```
cem_match = matchit(treatment ~ w1 + w2
 + w3 + w4, data = df, method = “cem”)
```

```
Note: treatment must be 0/1
```

```
summary(cem_match)
```

```
matched_data = match.data(cem_math)
```

```
M1 = lm(outcome ~ treatment, data =
  matched_data, weights = weights)
```

Assessing balance

```
# step 3: evaluate balance
summary(match_model)
```

```
Summary of balance for all data:
              Means Treated Means Control
distance             0.8224        0.4467
yearsmarried        10.1887        3.1209
religiousness        3.2116        2.8772


Summary of balance for matched data:
              Means Treated Means Control
distance             0.7885        0.7878
yearsmarried         9.2287        9.2193
religiousness        3.1561        3.1561
```

```
Sample sizes:
          Control Treated
All           171     430
Matched       171     346
Unmatched       0      84
Discarded       0       0
```

Matching affairs

Point of matching is to find untreated observations to serve as control group

We can’t know what would have happened had all those people who didn’t have children had instead had children

But if we can pick the most comparable group that’s as good as it gets!

Why this vs. controls?

Matching is just one form of closing backdoors

Like controlling, with matching we need to identify all backdoors

Why match? matching has some nice statistical properties and is easier to interpret

Why this vs. controls?

“We find a negative correlation between being in LRDP and conflict”

“Well, that’s only because the places that got LRDP were already likely to have conflict. That’s why they were chosen!”

Using controls:
“Well, we subtracted out the parts of being in LRDP and conflict explained by past conflict, and the effect remains”

What does this mean? And did it “work”?

Using matching:
“Well, even if we only look at places with similar past levels of violence the effect remains”

More intuitive and can check balance

Practice

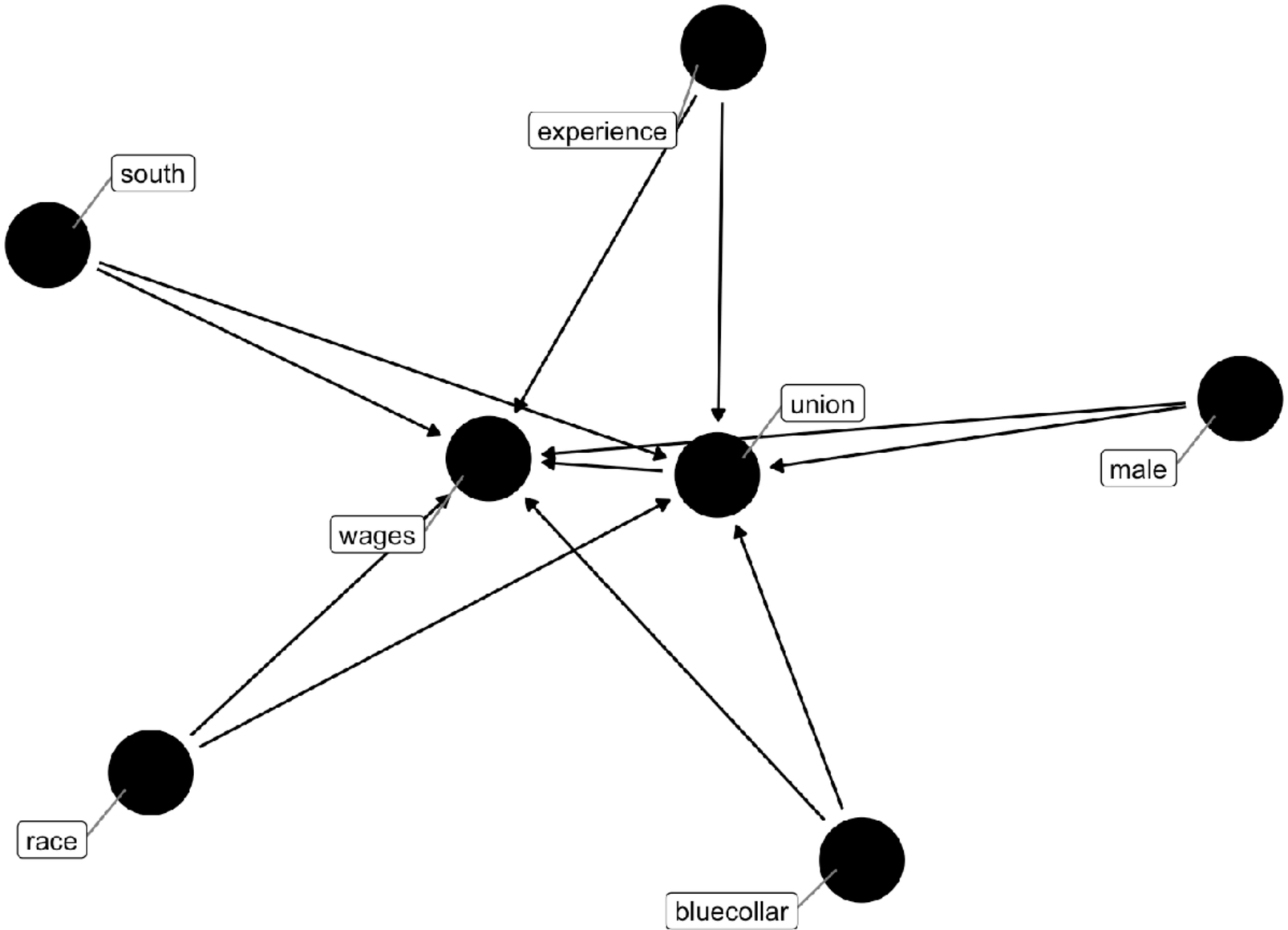
Does being in a union pay off?

Survey workers, find out if in union, how much they make, other characteristics

```
# practice: Does being in a unio
library(Ecdat)
data("Wages")
Wages %>%
  sample_n(5)
```

lwage	union	black	bluecol	south	exp	sex
6.91	no	no	no	no	9	male
6.77	yes	no	yes	no	38	male
6.07	yes	no	yes	no	10	male
6.4	no	no	yes	yes	21	male
7.09	no	no	no	yes	18	male

Practice



Practice

Match on all backdoors

Examine balance

Estimate effects of union on wages