

causal inference



PRINCIPLES OF STUDY DESIGN

introduction



introduction



- My name is Professor Baiocchi.
- I'm a statistician in Epidemiology, Statistics, Bioinformatics, and SPRC.
- I specialize in causality, mostly in health interventions.
- I do research in: [cardiothoracic interventions](#), [neonates](#), [educational interventions](#), [criminology](#), [anti-labor trafficking](#), and [sexual assault prevention](#).
- You can reach me at: baiocchi@stanford.edu

why do statistics?



three types of analyses



- **Descriptive**
 - What has been happening/what is going on right now?
 - Are things happening in a way that we anticipate?
 - What theories might we come up with to explain what we're seeing?
- **Predictive (or correlational)**
 - Given everything we know, can we do a good job of predicting what might come next?
 - Are there groups that seem to consistently be different than other groups?
- **Causal**

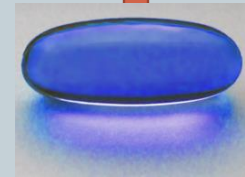
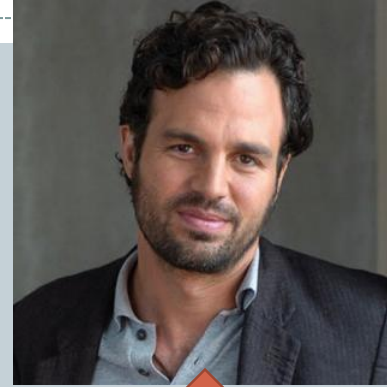
causal inference



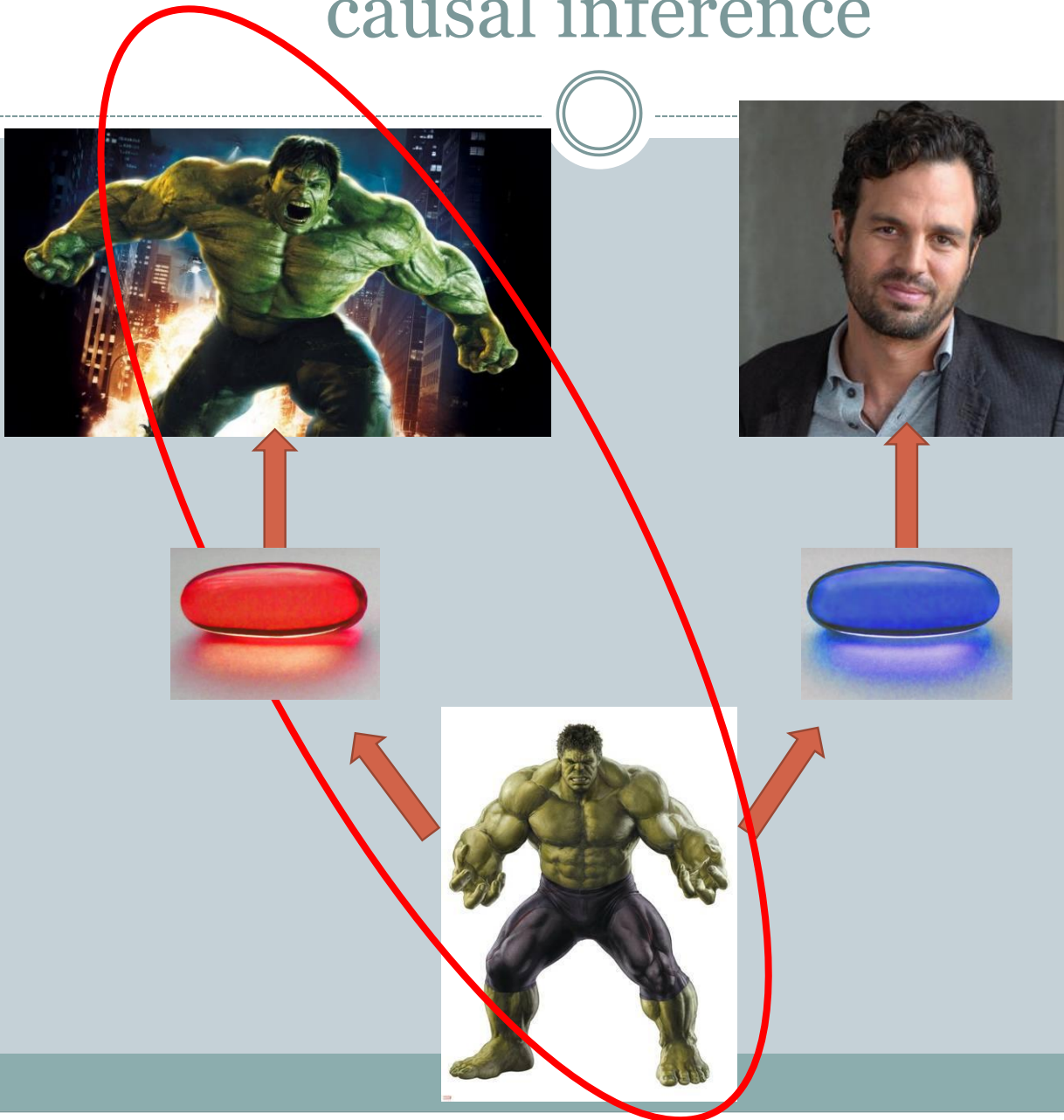
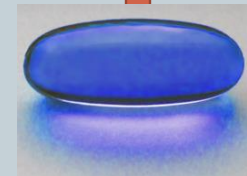
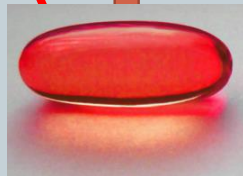
- The goal is to figure out what the change in the outcome will be for a person if we change from the control to the treatment:

$$Y_i(t = 1) - Y_i(t = 0) = \delta_i$$

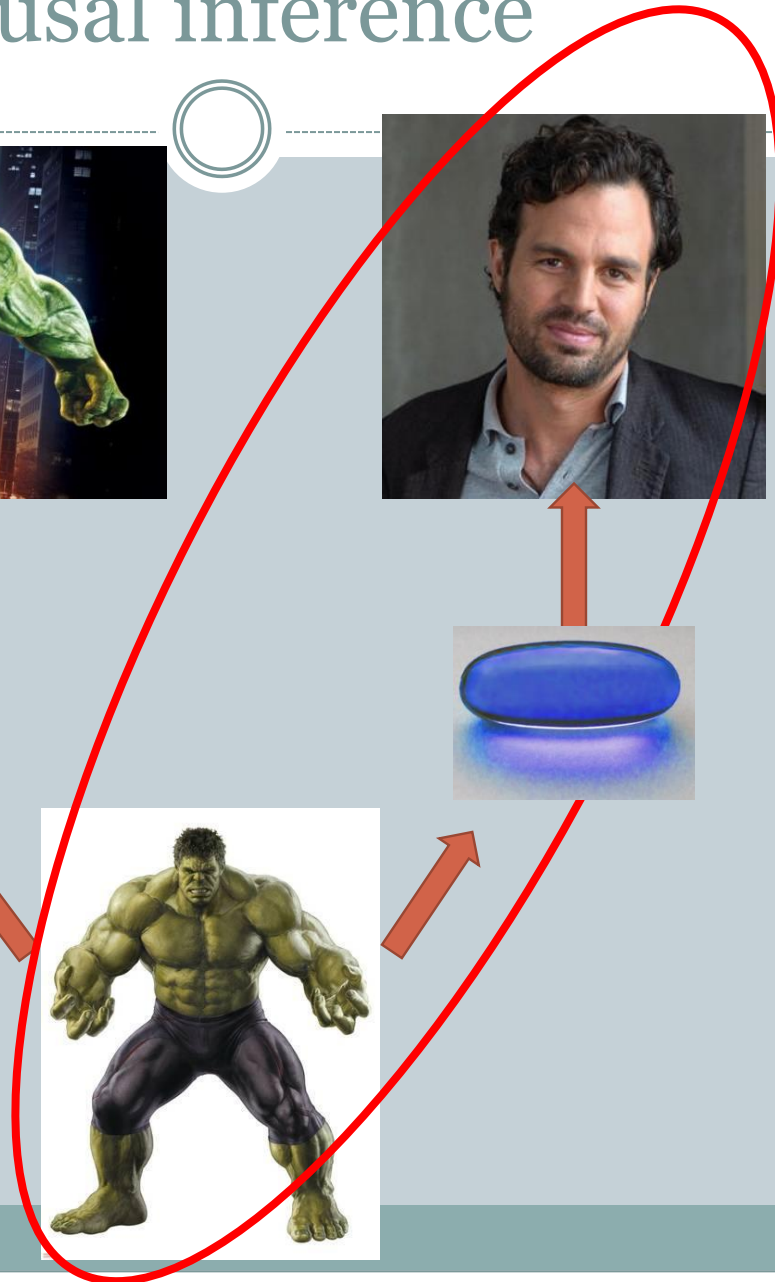
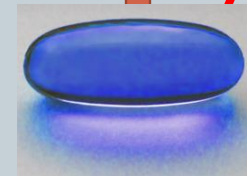
causal inference



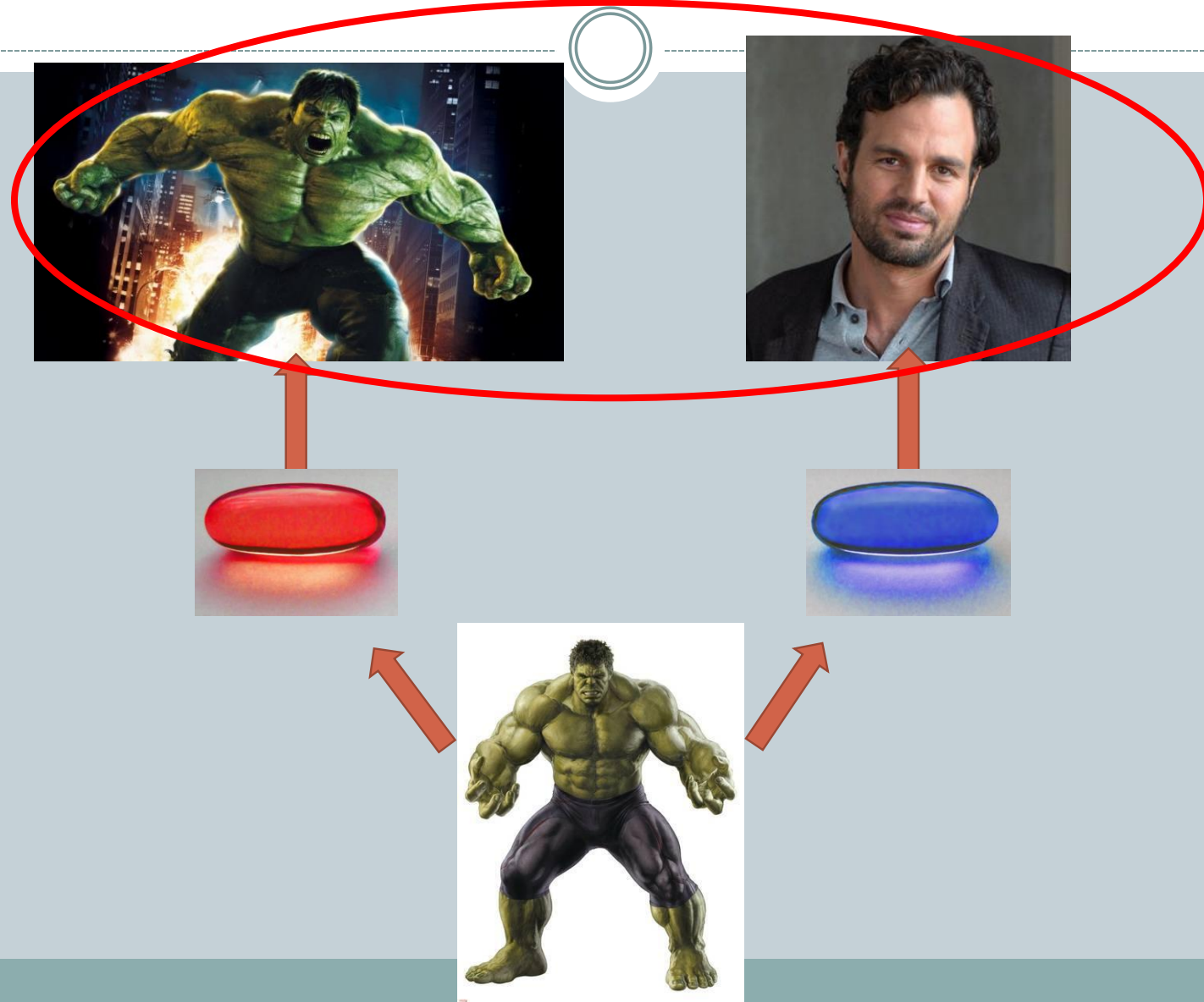
causal inference



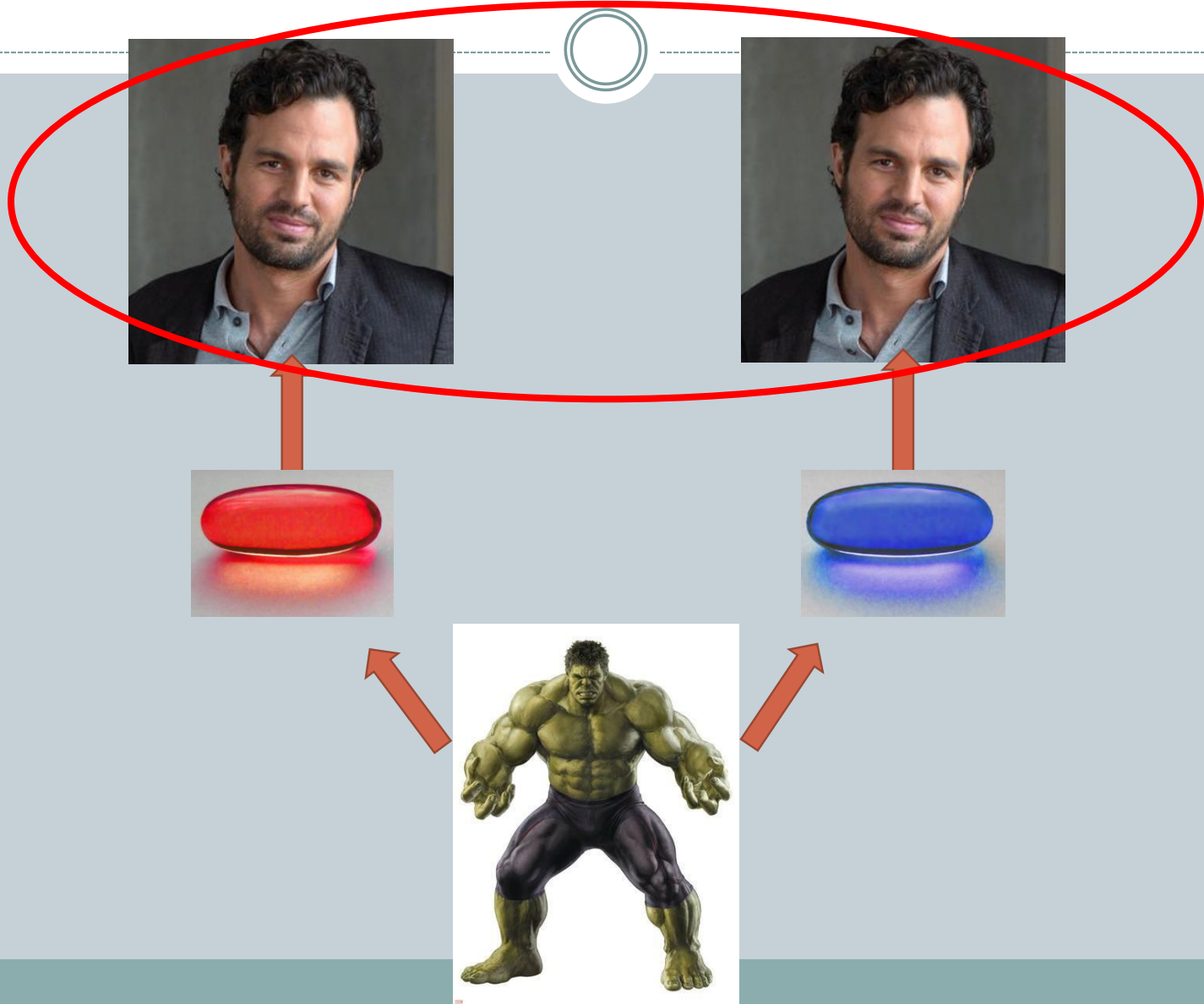
causal inference



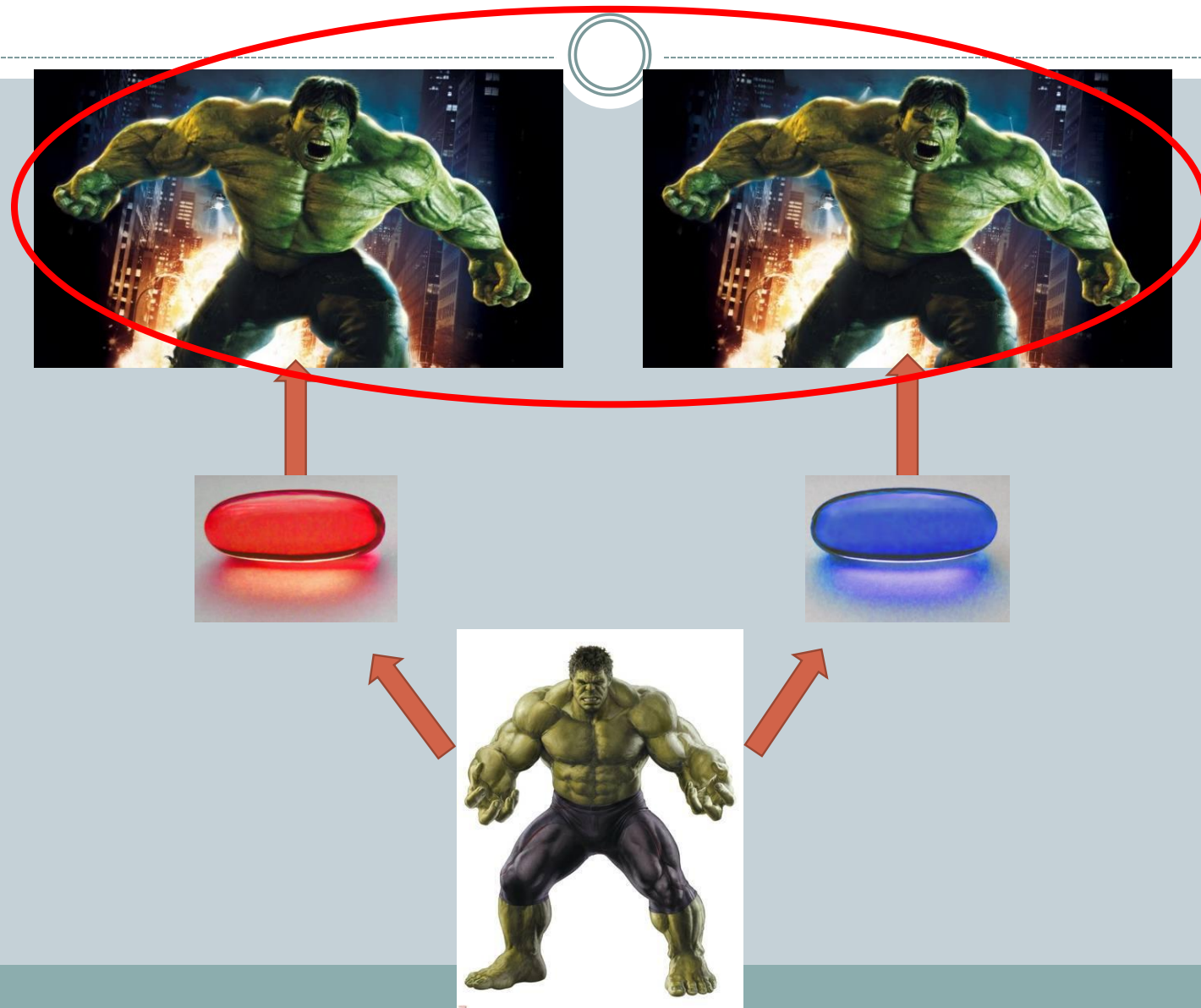
causal inference



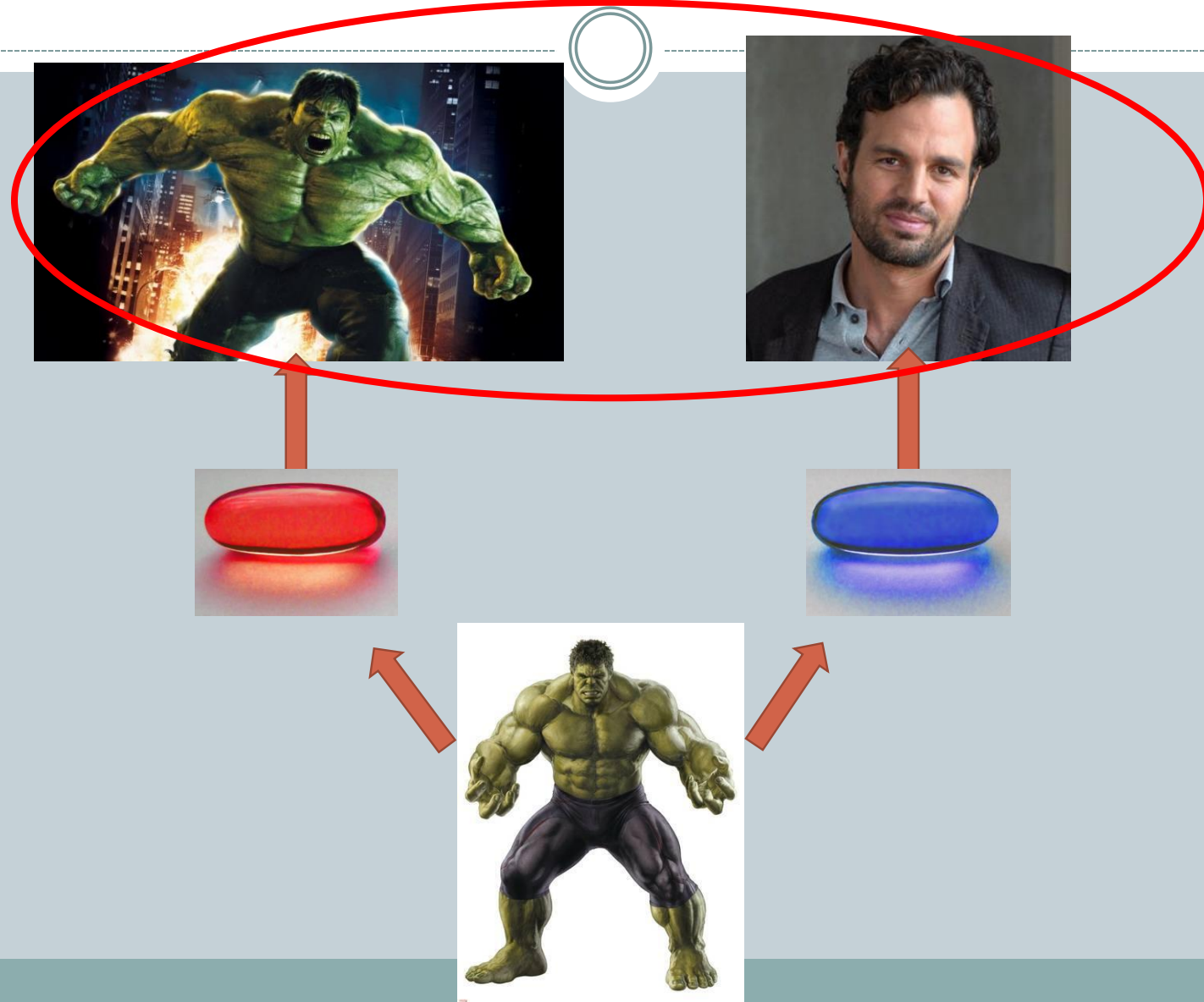
causal inference



causal inference



causal inference



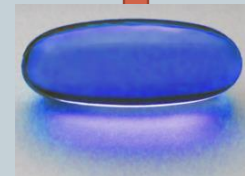
fundamental problem of causal inference



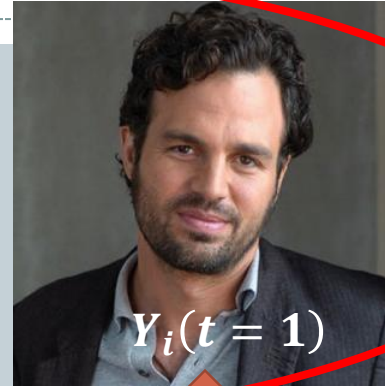
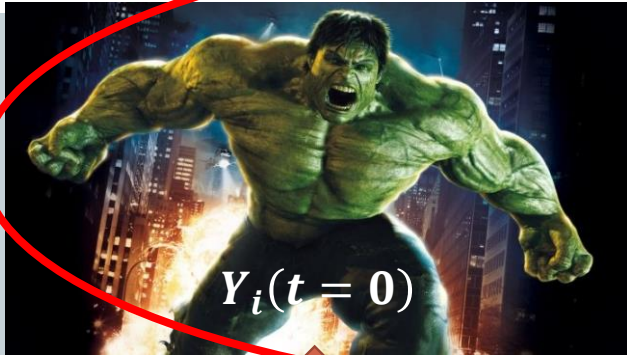
fundamental problem



fundamental problem



fundamental problem



Fundamental problem of causality:
Can't observe $Y_i(t = 1)$ and $Y_i(t = 0)$ at the same time.



so what now?



- Move from individuals to groups.

identification



identification



- How can we express quantities that involve **counterfactuals** using the data that we **observe**?
 - **Expression involving counterfactuals:** What is the expected difference in outcomes if we had, possibly counter to fact, intervened and given treatment, versus if we had given control?

$$E[Y_i(t = 1)] - E[Y_i(t = 0)]$$

- **Expression involving observations:** What is the expected difference in outcomes given receipt of treatment versus receipt of control?

$$E[Y | t = 1] - E[Y | t = 0]$$

identification



- **Statistical inference**

- Observe data.
- Compute some function of the data (estimation).
- Say something about a property in the world and assign some uncertainty in what we say (inference).

identification



- **Causal inference**

- Specify some causal quantity of interest and “tie” that causal quantity to a property in the world. (identification).
- Observe data.
- Compute some function of the data (estimation).
- Say something about a property in the world and assign some uncertainty in what we say (inference).

$$E[Y \mid t = 1] - E[Y \mid t = 0] \stackrel{?}{=} E[Y_i(t = 1)] - E[Y_i(t = 0)]$$

identification



- Identification requires assumptions.
- Sometimes, we have control over how we collect the data so that these assumptions are true by construction (more on this with experiments).
- Other times, we just have to believe them (more on this later, with observational studies).

terminology



**UNIT OF OBSERVATION
COVARIATE**

terminology



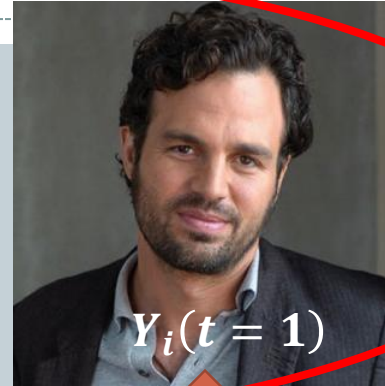
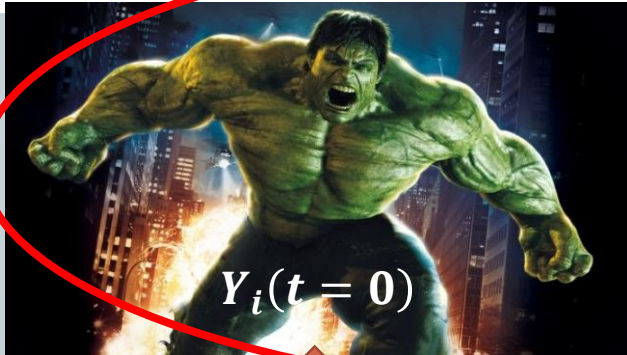
- **Unit of observation** – the element in the study for which the intervention can be applied to or withheld from.
 - In our working example: people (superheroes?).
 - Can have different levels of aggregation being “units of observations”: doctors who treat patients, clinics, municipalities, etc.

terminology



- **Covariate** – a variable, distinct from the intervention and outcome, that can change from unit of observation to unit of observation
 - In our working example: anger, baseline weight, gender, BMI, age, hair color, favorite color...
 - Not all covariates are equally “important.” We’ll revisit this notion when we discuss the concept of *confounding*.

fundamental problem



Fundamental problem of causality:
Can't observe $Y_i(t = 1)$ and $Y_i(t = 0)$ at the same time.



two approaches

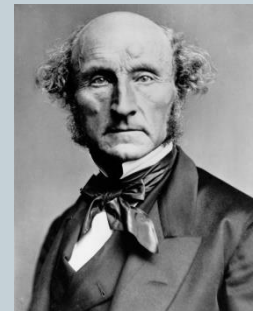


two approaches



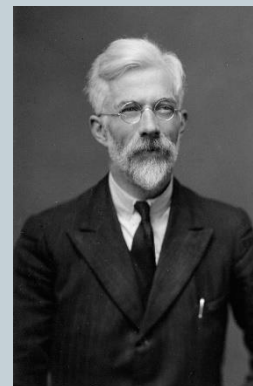
- **John Stuart Mill**

- Philosopher, economist, early feminist and civil servant.
- Estimate effect through “method of differences.”



- **Sir Ronald Fisher**

- Statistician and biologist.
- Estimate effect through “a controlled & random process.”



method of difference



MILL

method of difference

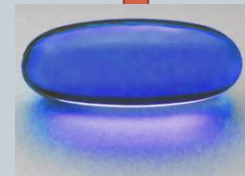


- In 1864, in his *System of Logic: Principles of Evidence and Methods of Scientific Investigation*, Mill proposed four methods of experimental inquiry, including the “method of difference:”

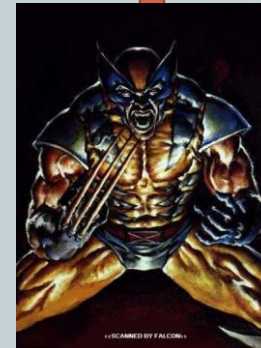
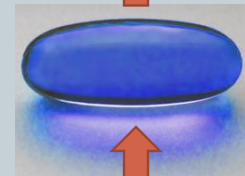
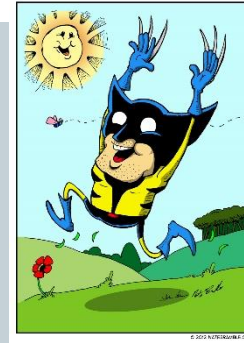
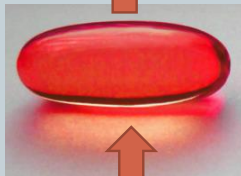
If an instance in which the phenomenon ... occurs and an instance in which it does not ... have every circumstance save one in common ... [then] the circumstance [in] which alone the two instances differ is the ... cause or a necessary part of the cause (III, sec. 8)

- For Mill, homogeneity and sound causal inference were closely linked: he wanted “two instances ... exactly similar in all circumstances except the one” under study.

causal inference



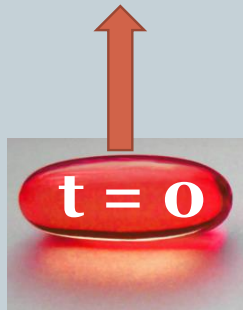
causal inference



causal inference

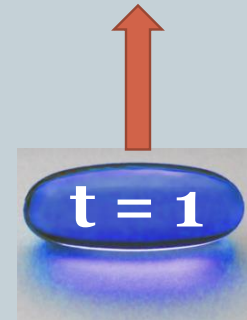


$Y(t=0)$



baseline input

$Y(t=1)$

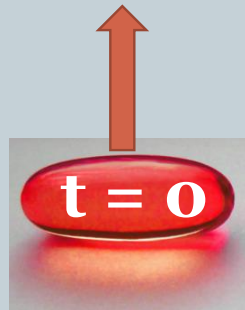


baseline input

causal inference

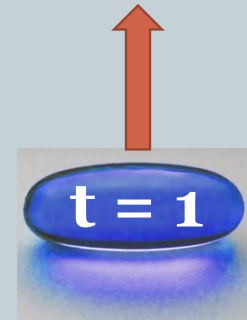


$$Y(t=0) = f(t = 0, X = x)$$



x

$$Y(t=1) = f(t = 1, X = x')$$

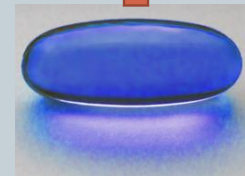


x'

causal inference



method of difference



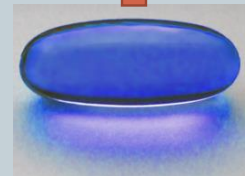
$$Y(t=0) = f(t = 0, X = x)$$



$$Y(t=1) = f(t = 1, X = x')$$



$x =$



$x' =$



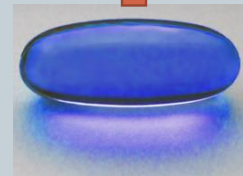
$$Y(t=0) = f(t = 0, X = x)$$



$$Y(t=1) = f(t = 1, X = \textcolor{red}{x})$$



$x =$



$\textcolor{red}{x} =$



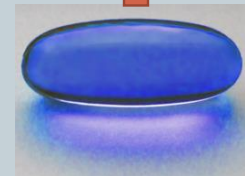
The only difference

$$Y(t=0) = f(t=0, X = \mathbf{x})$$

$$Y(t=1) = f(t=1, X = \mathbf{x})$$



$\mathbf{x} =$



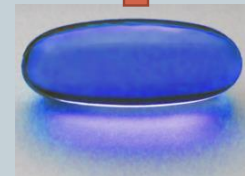
$\mathbf{x} =$

$$Y(t=0) = f(t = 0, X = x)$$

$$Y(t=1) = f(t = 1, X = x')$$



$x =$



$x' =$

terminology



CONFOUNDING

terminology



- **Confounding** – when something (usually a covariate or a set of covariates) makes your estimate of the causal effect biased (*loosely speaking*: makes your estimate – on average – report a number different than the number it should be).
 - In our working example: baseline weight, gender, BMI, age, ~~hair color~~, ~~favorite color~~...
 - Confounding is usually thought of as covariates that cause variation in the outcome as well as the treatment.
 - If the treatment group would have been different than the control group, even if we never applied any form of intervention, then we're almost surely going to experience confounding.

randomization



FISHER

fisher: a deep insight



- Fisher's randomization
- If we control the randomization process then we can describe, with mathematical certainty, how the data will behave.
- Armed with this understanding of data's behavior we can then make statements, with varying levels of certainty, about the state of the world.

fisher: a lady tasting tea



- In his 1935 groundbreaking book, *Design of Experiments*, he discusses an (apocryphal?) encounter he had with a lady at a gathering.
- She contended she could taste the difference between tea which had had its milk poured in first versus tea which had had milk poured in after the tea.
- Fisher thought this was hogwash and proceeded to develop a “test” of her claim.
- Interestingly, he discusses some of the reasoning that led him to this particular test.

fisher: a lady tasting tea



- In in Chapter 2 (p. 18) he wrote:

It is not sufficient remedy to insist that “all the cups must be exactly alike” in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation ... These are only examples of the differences probably present; it would be impossible to present an exhaustive list of such possible differences ... because [they] ... are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labor and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment ...

confounding



- Confounding comes in two flavors:
 - (i) observed confounding (the covariates are in your data set and we can probably do something), and
 - (ii) unobserved confounding (you're going to have a *really* rough time...)
- Assume you have unobserved confounding, until proven otherwise.

fisher: a lady tasting tea



fisher: a lady tasting tea



- His point: You will always have confounding. It will be annoying. Let's move past that.
- His proposal?
- Propose a treatment assignment process that is well-described mathematically and random
- Propose a hypothesis that will explain how the data should look in general
 - This is really important, the theory here should contain information about how the intervention interacts with the outcome,
 - The theory should guide you in which confounders are most impactful, and how to measure the outcome(s).
- Run the experiment and compare the observed data to the actual way the world worked

fisher (and others!): null distributions



- Compare what we anticipated with what the world actually produced.
- This is kind of magical... under this theory, we can investigate counterfactuals – other ways the world could have been.

Wait..... are standard errors important?

Fisher's sharp null



the meaning of “no effect”



- There are at least two definitions of “no effect” floating around
- In this class, we’ll almost exclusively use what is referred to as Fisher’s Sharp Null:

$$H_0: Y_i(0) = Y_i(1)$$

for unit of observation i , the response under control is identical to the response if the unit were to have received treatment.

- In most introductory statistics classes

$$H_0: \mu_C = \mu_T$$

the population C has the same mean as the population T .

a few words about “no effect”



- Fisher’s Sharp Null is quite “sharp,” meaning that it points down to the individual level.
- This grew out of the RCT framework.
- Unsurprisingly, the “no difference in population means” came out of the sampling framework.
- There was no obvious emphasis on the individual level in sampling.
- If Fisher’s Sharp Null is true then the “no difference in population means” is true.
- The converse does not hold.

the permutation test



intuition



- If we assume $H_0: Y_i(0) = Y_i(1)$ then we get a very powerful way of dealing with the fundamental problem of causality

intuition



person i	t_i	Y_i
1	1	3
2	0	2
3	0	6
4	1	7
5	1	3
6	1	2
7	0	5
8	1	2
9	1	1
10	0	4
11	1	7
12	0	7

t_i : Intervention assigned to person i

Y_i : Observed outcome for person i

intuition



person i	$Y_i(0)$	$Y_i(1)$	t_i
1			1
2			0
3			0
4			1
5			1
6			1
7			0
8			1
9			1
10			0
11			1
12			0

t_i : Intervention assigned to person i

Y_i : Observed outcome for person i

$Y_i(t=0)$: Outcome of person i if they were to receive intervention $t=0$

$Y_i(t=1)$: Outcome of person i if they were to receive intervention $t=1$

intuition



person i	$Y_i(0)$	$Y_i(1)$	t_i
1		-15	1
2	-8		0
3	6		0
4		15	1
5		-23	1
6		11	1
7	3		0
8		-3	1
9		-3	1
10	-6		0
11		13	1
12	-5		0

t_i : Intervention assigned to person i

Y_i : Observed outcome for person i

$Y_i(t=0)$: Outcome of person i if they were to receive intervention $t=0$

$Y_i(t=1)$: Outcome of person i if they were to receive intervention $t=1$

intuition




person i	$Y_i(0)$	$Y_i(1)$	t_i	Y_i
1		-15	1	-15
2	-8		0	-8
3	6		0	6
4		15	1	15
5		-23	1	-23
6		11	1	11
7	3		0	3
8		-3	1	-3
9		-3	1	-3
10	-6		0	-6
11		13	1	13
12	-5		0	-5

-2 -0.71



1.29

intuition



person i	$Y_i(0)$	$Y_i(1)$	t_i	Y_i
1		-15	1	-15
2	-8		0	-8
3	6		0	6
4		15	1	15
5		-23	1	-23
6		11	1	11
7	3		0	3
8		-3	1	-3
9		-3	1	-3
10	-6		0	-6
11		13	1	13
12	-5		0	-5

-2 -0.71



1.29

intuition



person i	$Y_i(0)$	$Y_i(1)$	t_i	Y_i
1	-15	-15	0	-15
2	-8		0	-8
3	6		0	6
4		15	1	15
5		-23	1	-23
6		11	1	11
7	3		0	3
8		-3	1	-3
9		-3	1	-3
10	-6		0	-6
11		13	1	13
12	-5		0	-5

intuition



person i	$Y_i(0)$	$Y_i(1)$	t_i	Y_i
1	-15	-15	0	-15
2	-8		0	-8
3	6		0	6
4		15	1	15
5		-23	1	-23
6		11	1	11
7	3		0	3
8		-3	1	-3
9		-3	1	-3
10	-6		0	-6
11		13	1	13
12	-5		0	-5

-4.17 1.67

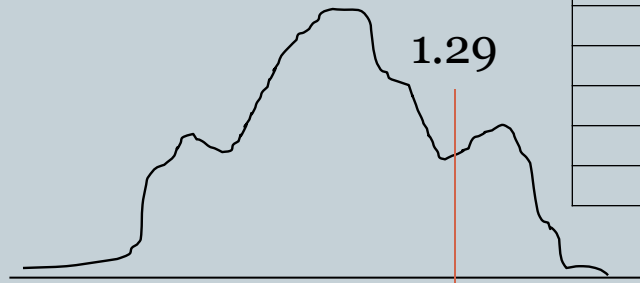


5.83

intuition



person i	$Y_i(0)$	$Y_i(1)$	t_i	Y_i
1	-15	-15	0	-15
2	-8		0	-8
3	6		0	6
4		15	1	15
5		-23	1	-23
6		11	1	11
7	3		0	3
8		-3	1	-3
9		-3	1	-3
10	-6		0	-6
11		13	1	13
12	-5		0	-5



-4.17 1.67



5.83

Repeat these flips over
and over again. Build up
the null distribution.

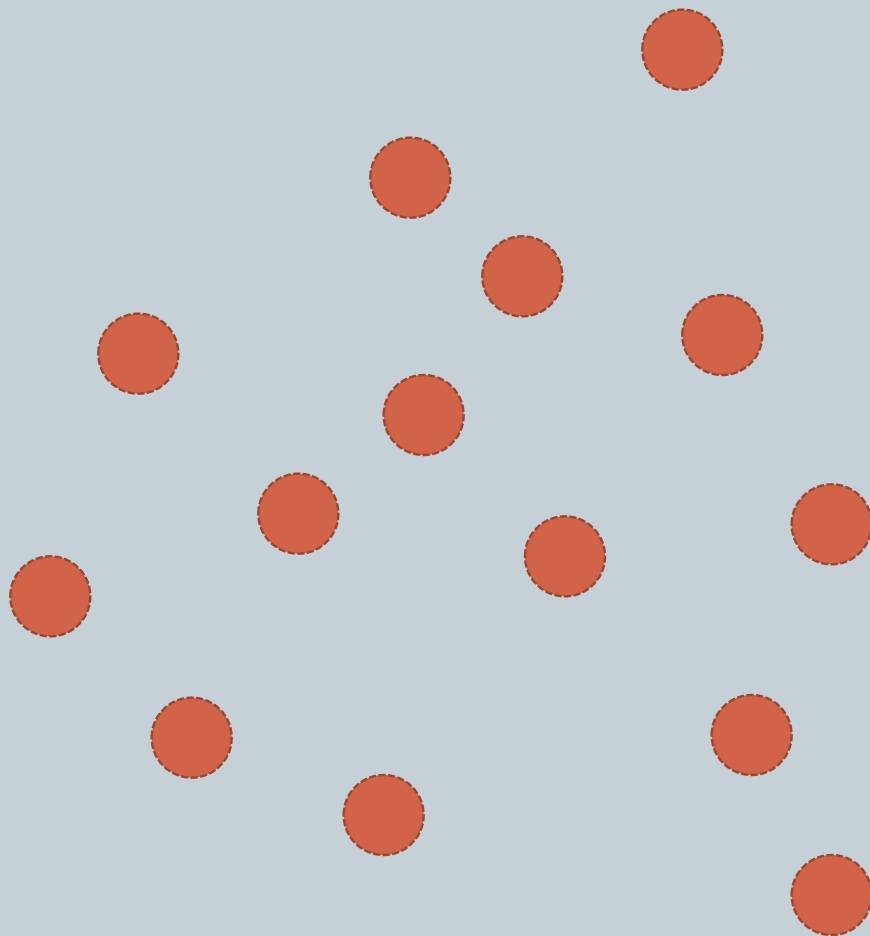
designing RCTs





treatment

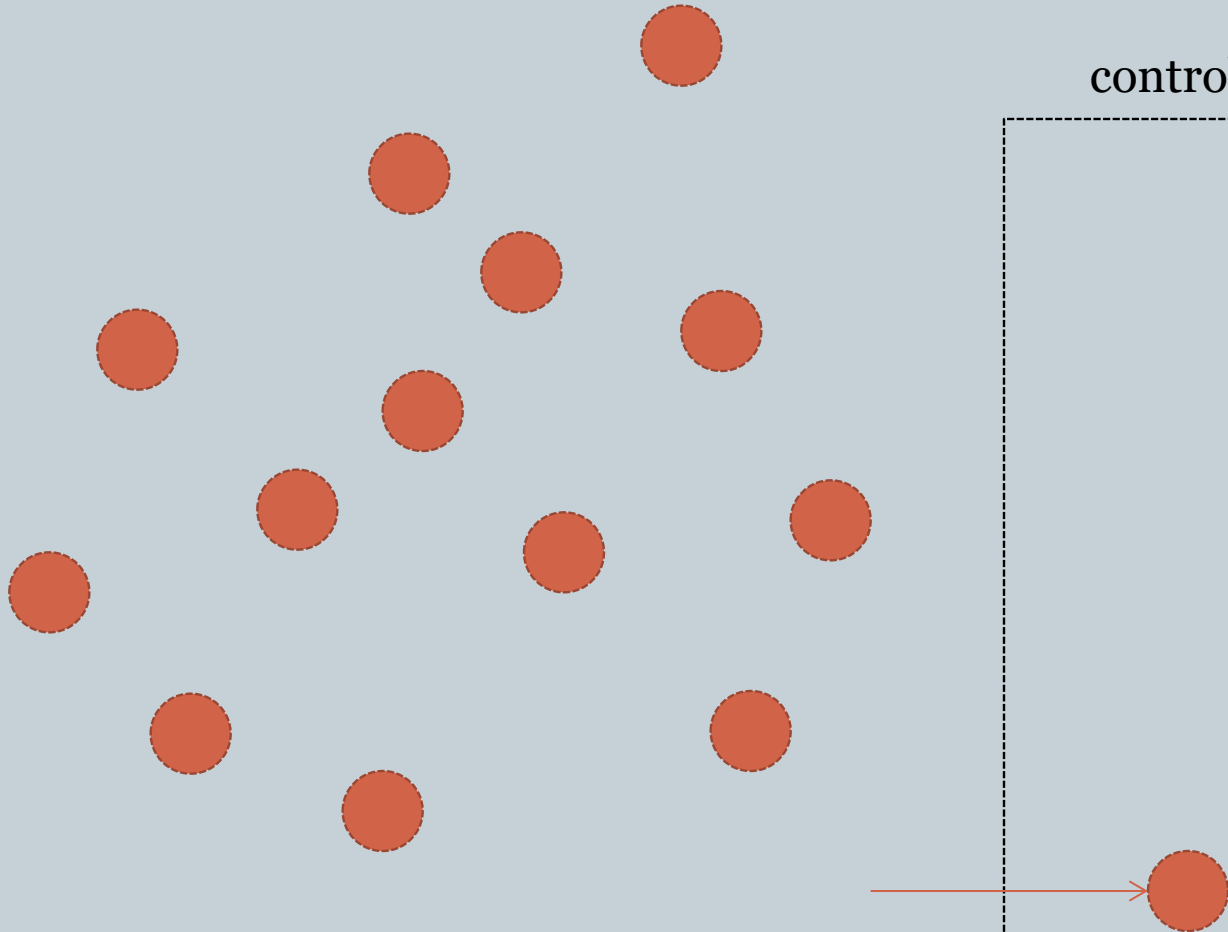
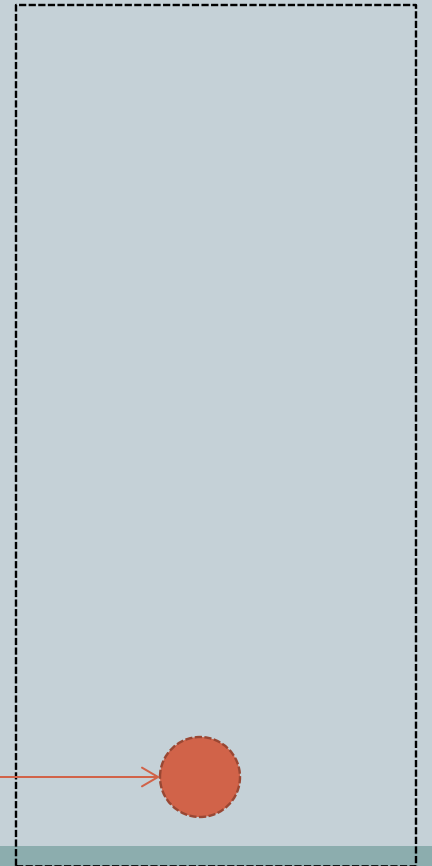
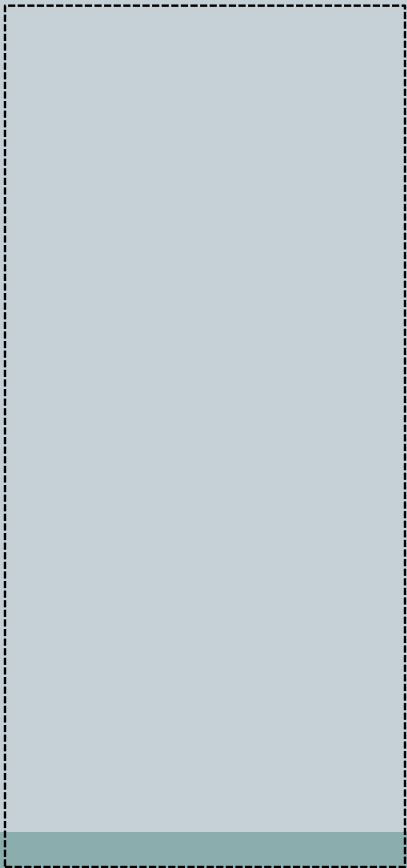
control





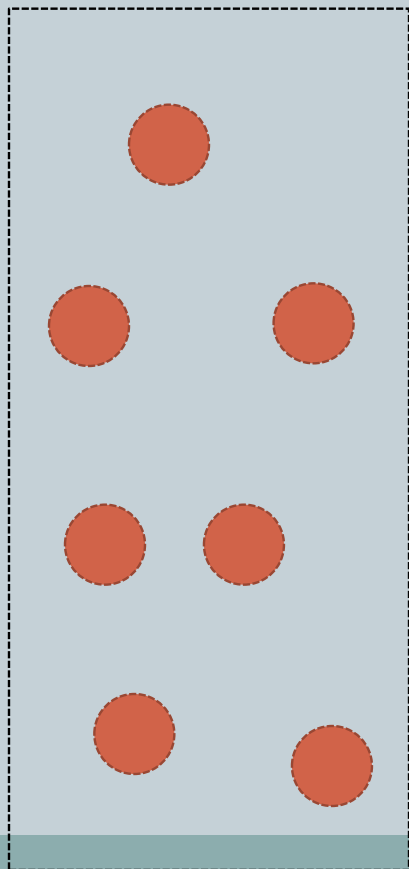
treatment

control

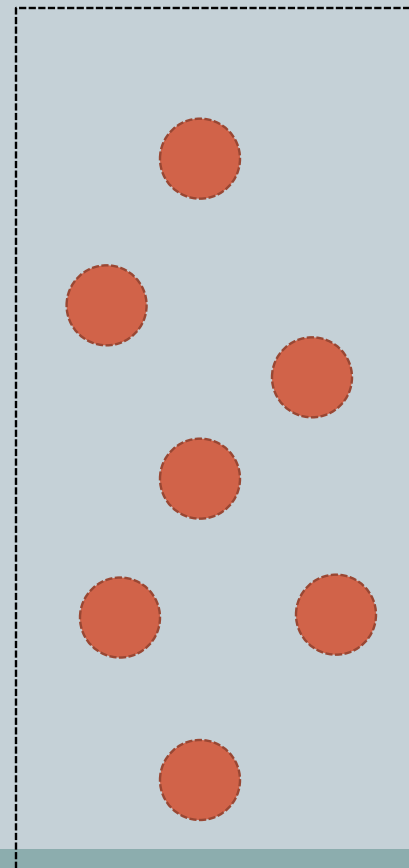




treatment

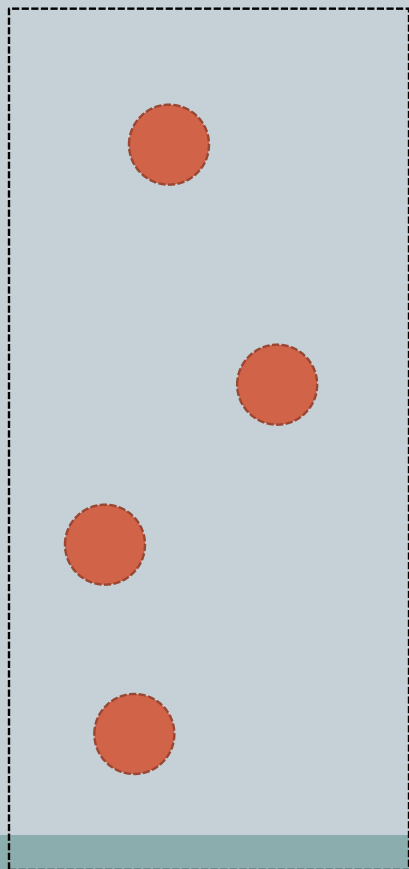


control

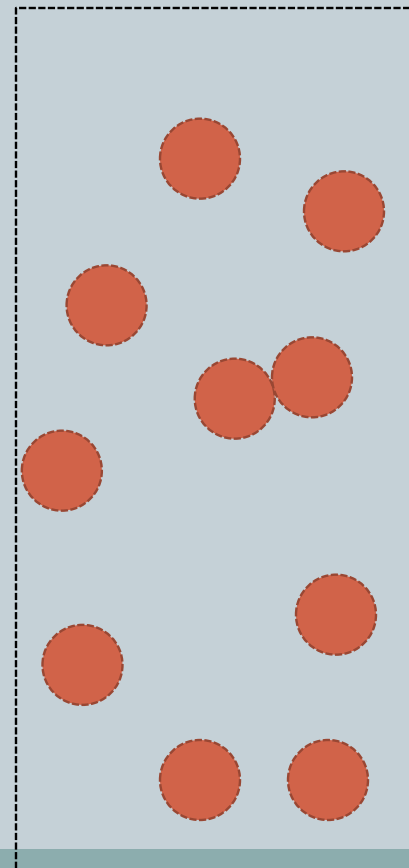




treatment



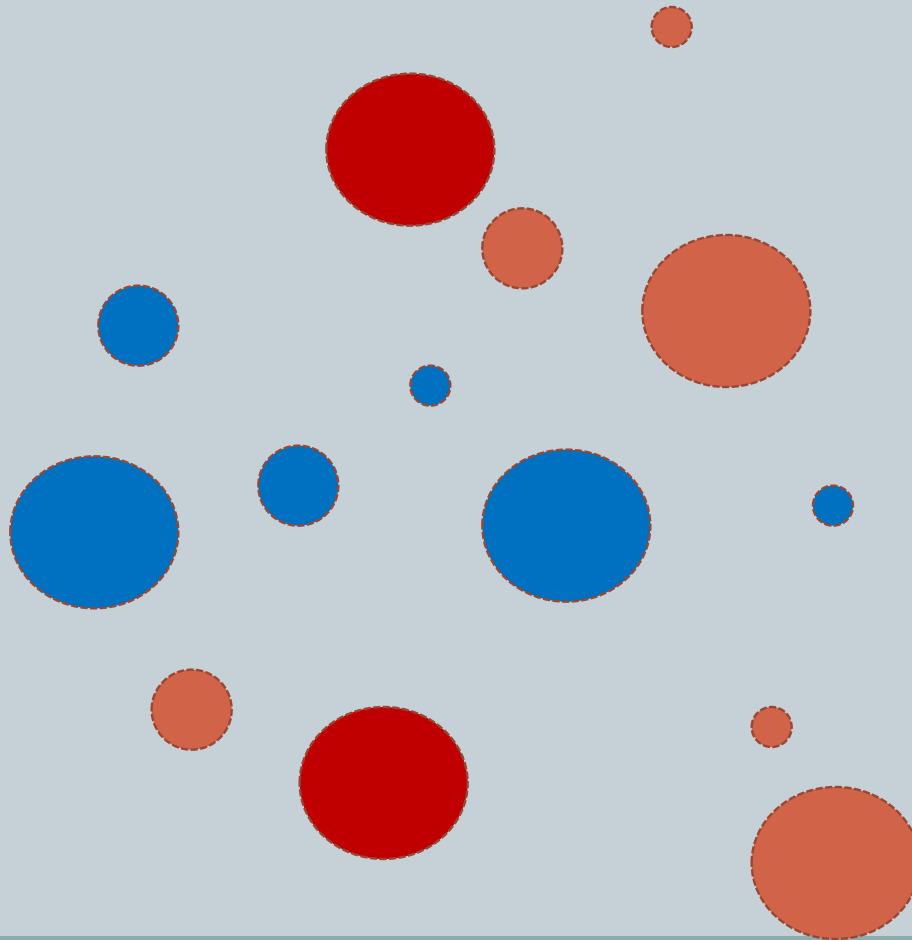
control





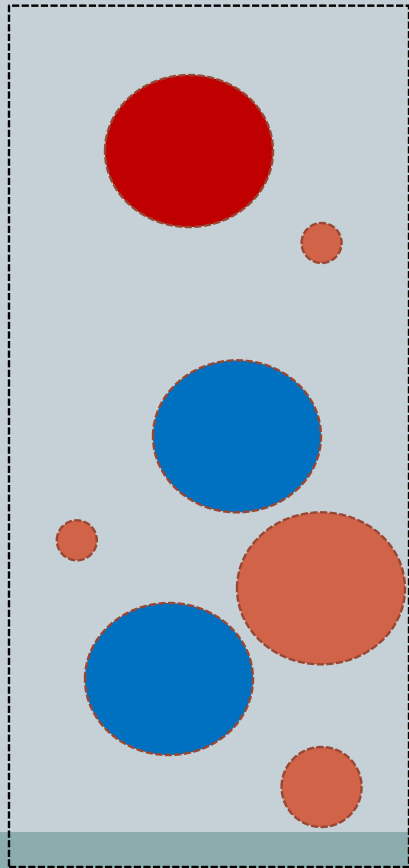
treatment

control

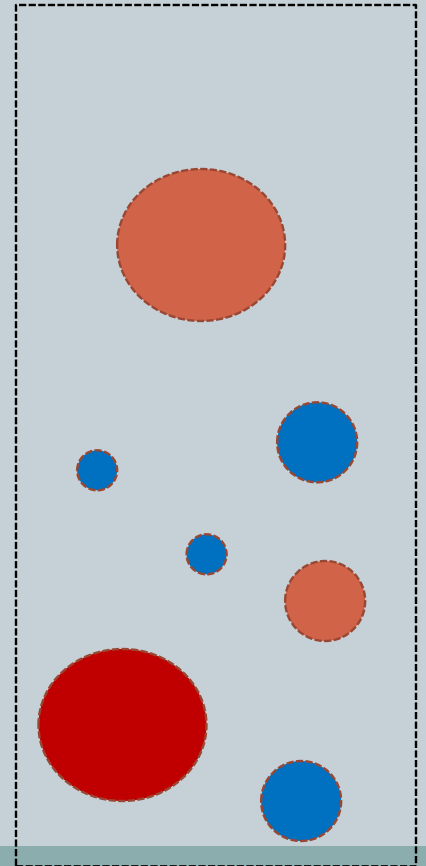




treatment

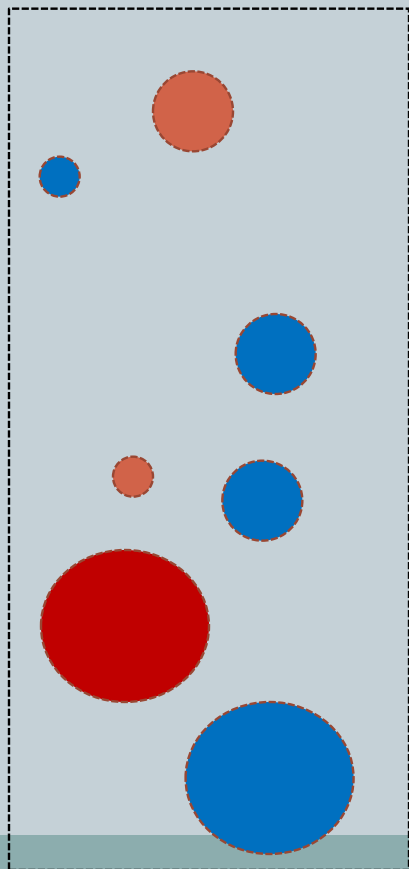


control

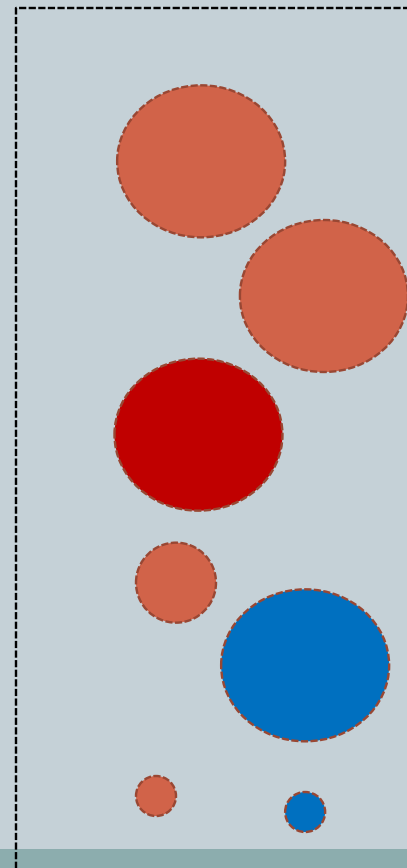




treatment

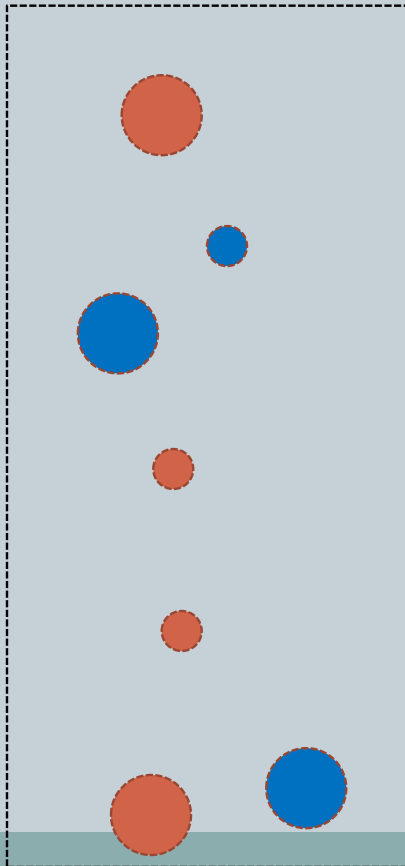


control



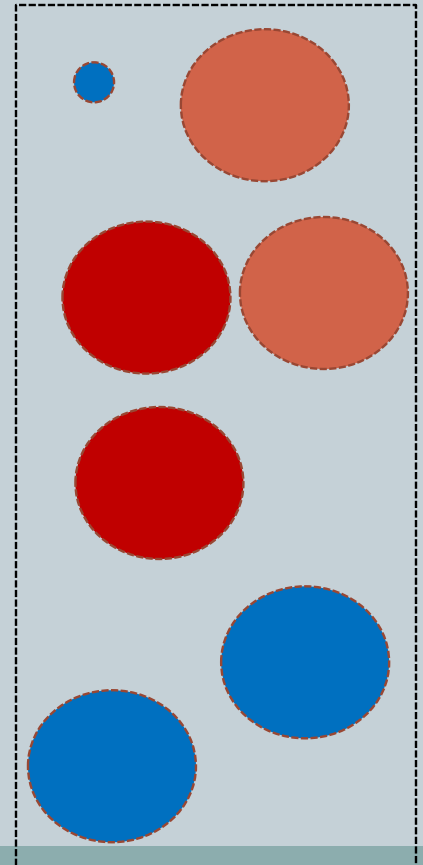


treatment



Even though we randomized
this is absolutely a garbage
study. Throw this away and
start over.

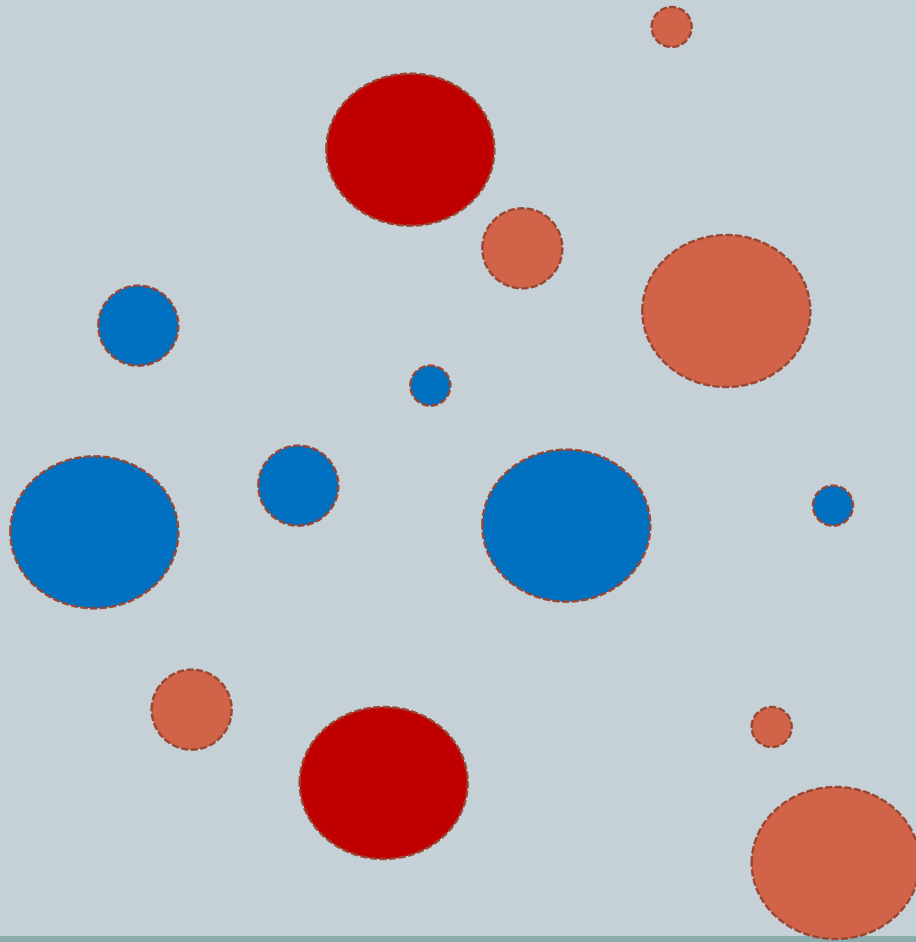
control





treatment

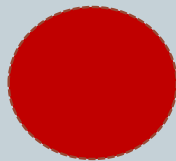
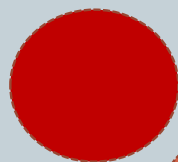
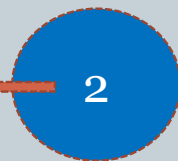
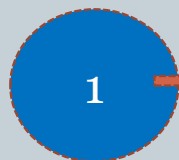
control





treatment

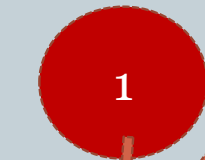
control





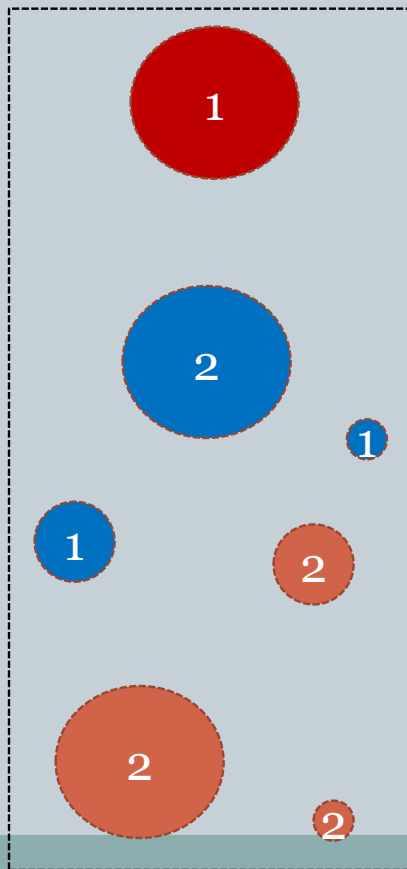
treatment

control

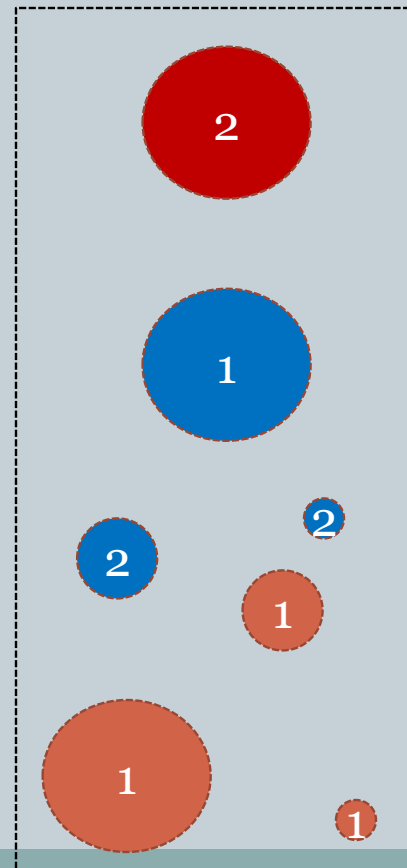




treatment

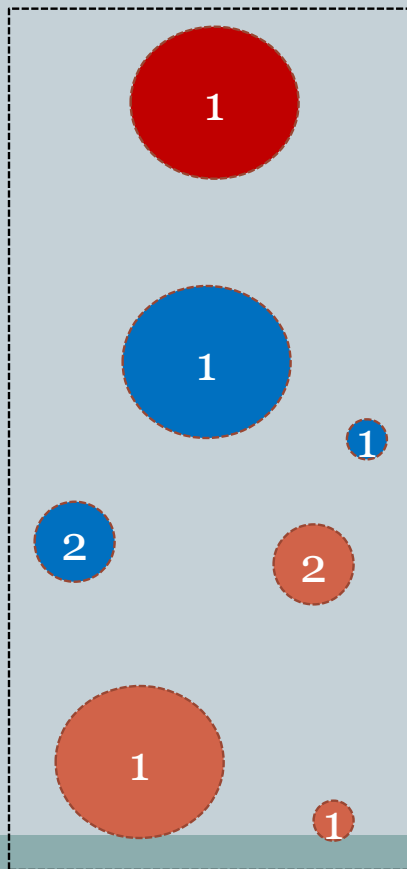


control

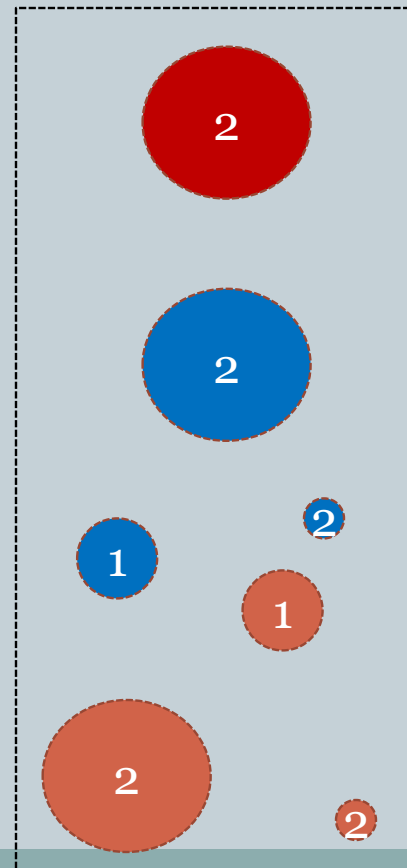




treatment



control



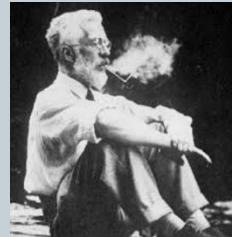
beyond RCTs



beyond RCTs



- Randomized controlled trials (RCTs) are excellent
 - The “controlled” part addresses Mill’s ideas of minimizing differences at baseline
 - The “randomized” part addresses Fisher’s ideas of understanding what-else-could-have-happened
- Unfortunately, RCTs can’t be implemented in all situations:
 - Does smoking cause cancer?
 - Are higher level NICUs better?
 - Expensive? Feasibility? Useful?
- There are study designs that were created to work “in the real world,” and they follow many of these ideas...



beyond RCTs



- The world of “observational studies” is kind of hard to get into because it grew up in several distinct, but overlapping, disciplines:
 - Epidemiology
 - Demography
 - Economics (econometrics)
 - Political Science
 - Sociology
 - Biostatistics
 - Statistics
 - Psychology (psychometrics)
 - Computer Science

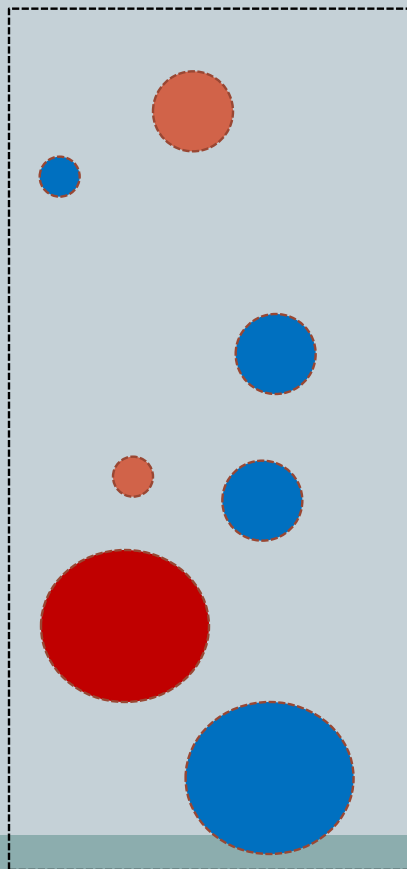
balancing observations studies



TWO APPROACHES & SOME INTUITION



Actual Treated



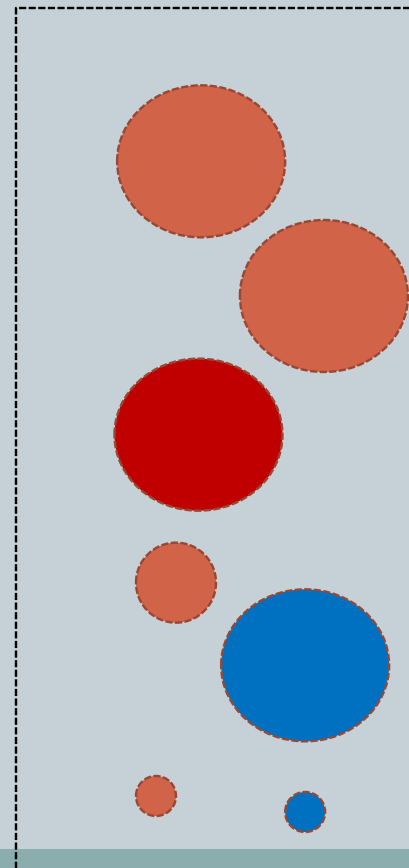
Synthetic Treated



Synthetic Control

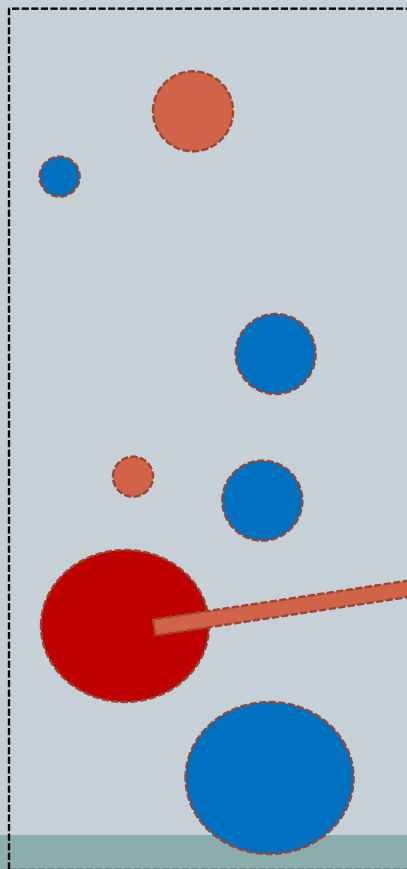


Actual Control

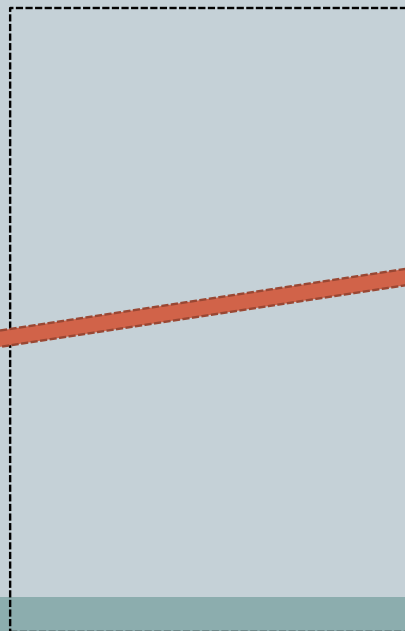




Actual Treated



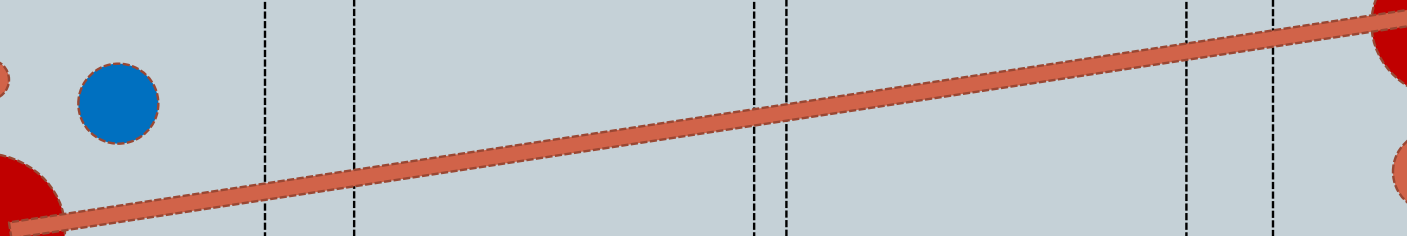
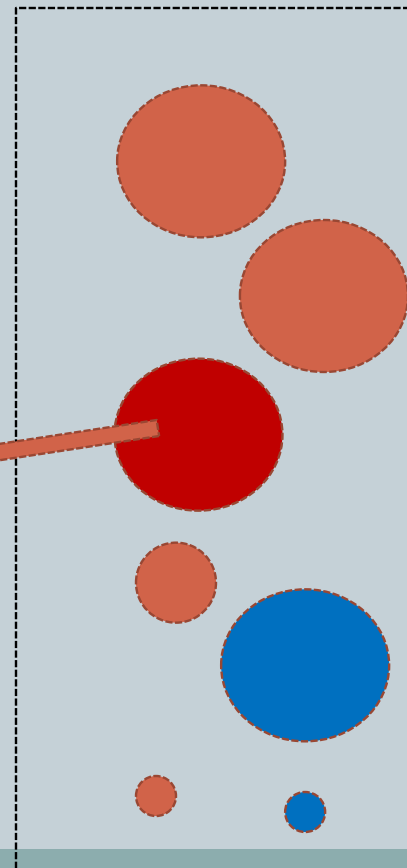
Synthetic Treated



Synthetic Control

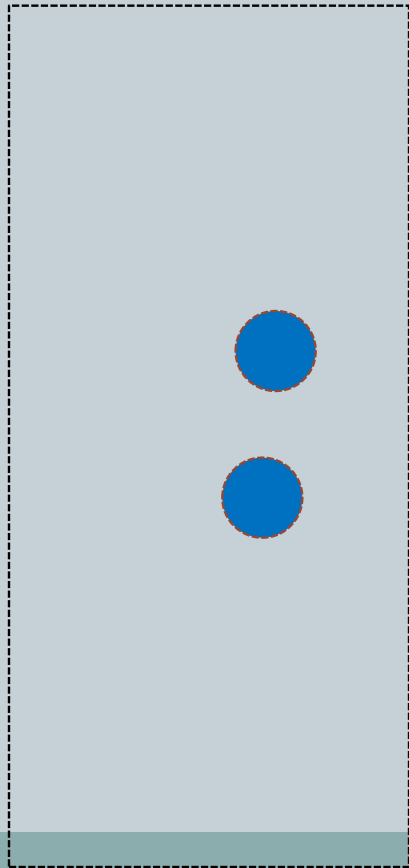


Actual Control

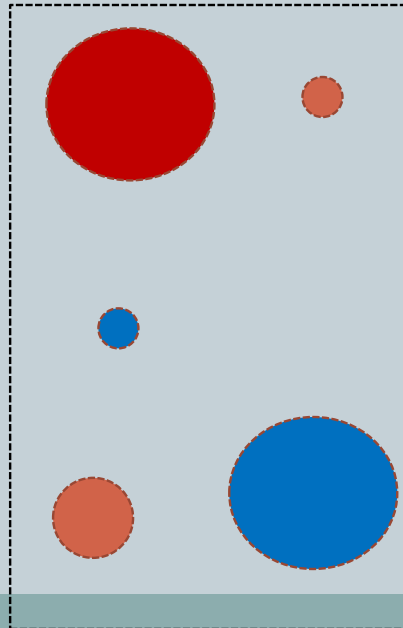




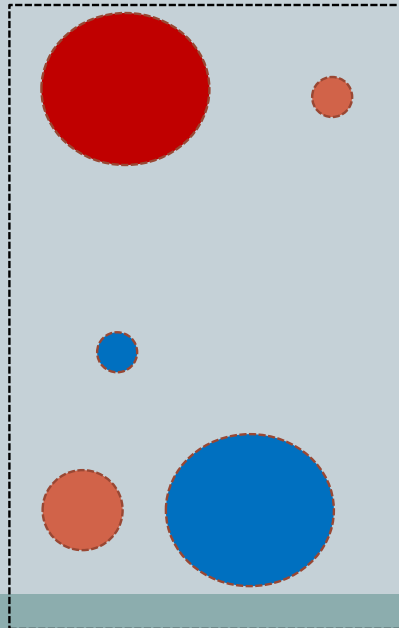
Actual Treated



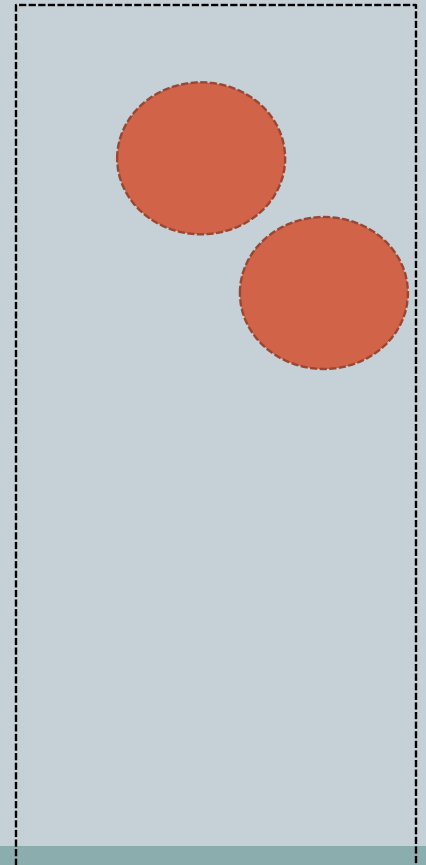
Synthetic Treated



Synthetic Control



Actual Control



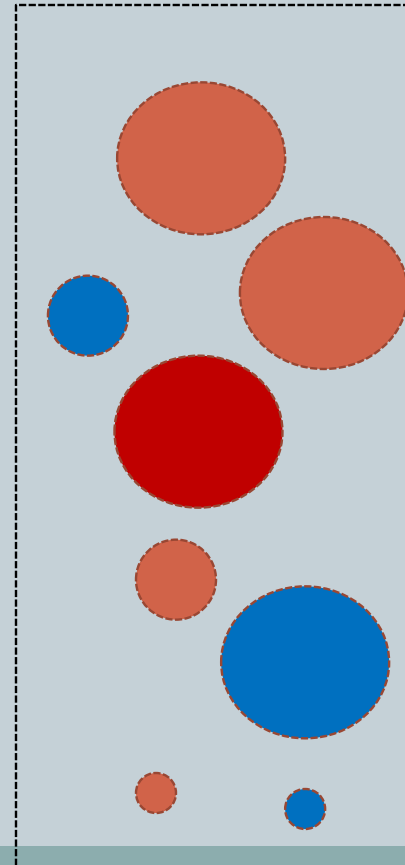
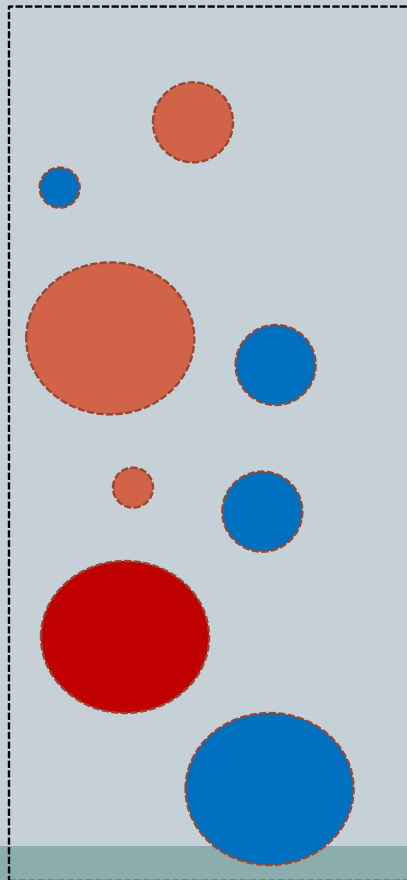


Actual Treated

Synthetic Treated

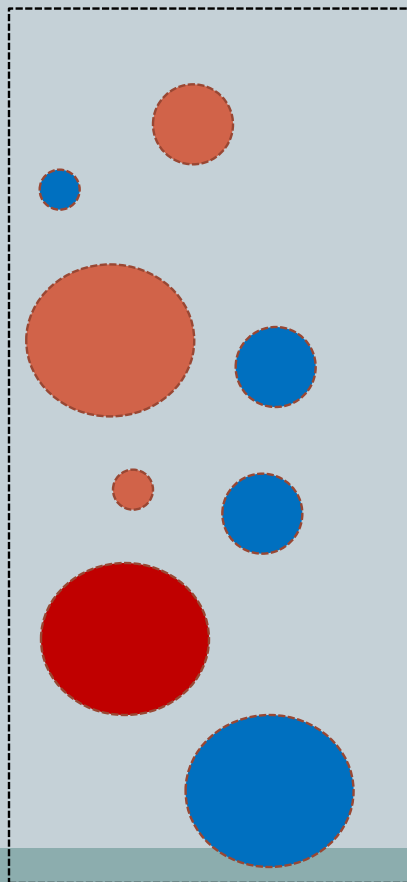
Synthetic Control

Actual Control

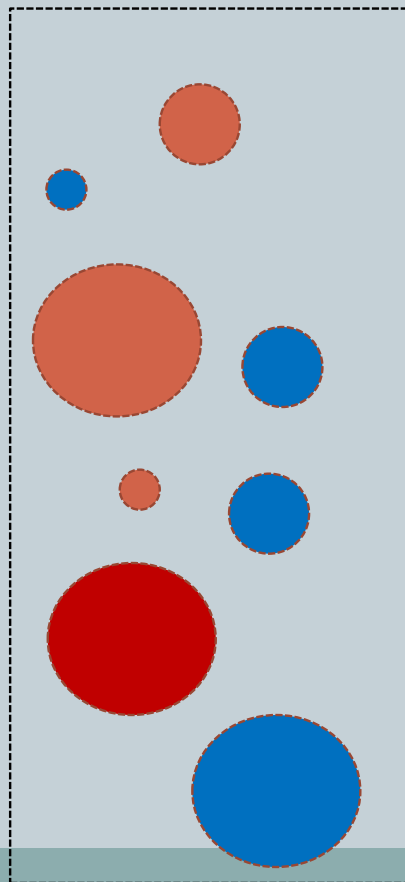




Actual Treated



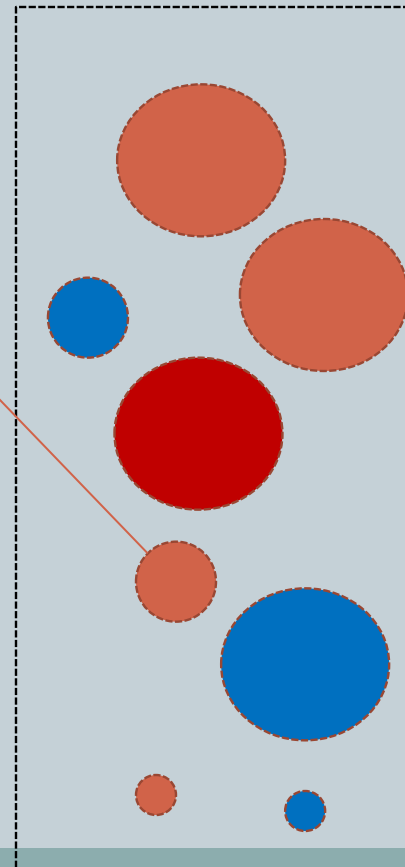
Synthetic Treated



Synthetic Control

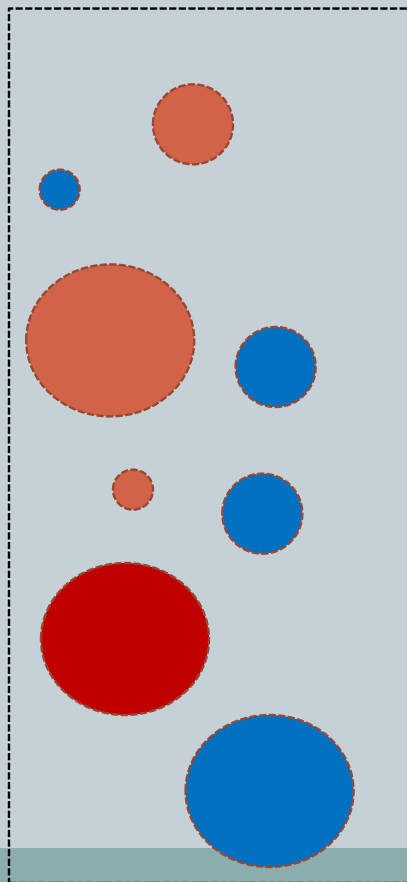


Actual Control

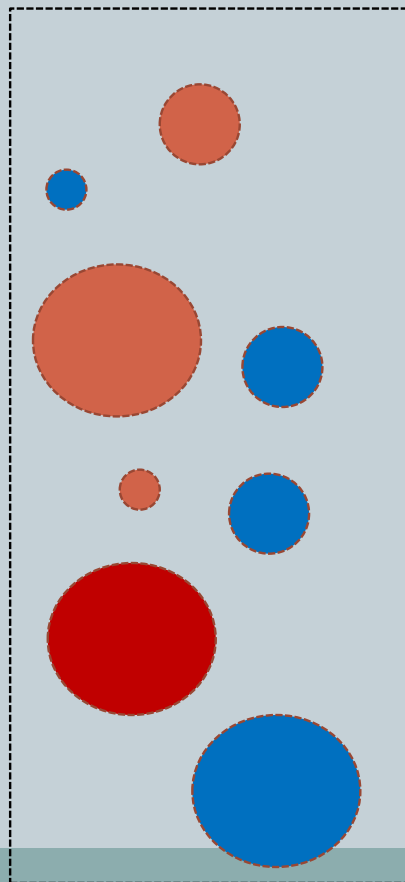




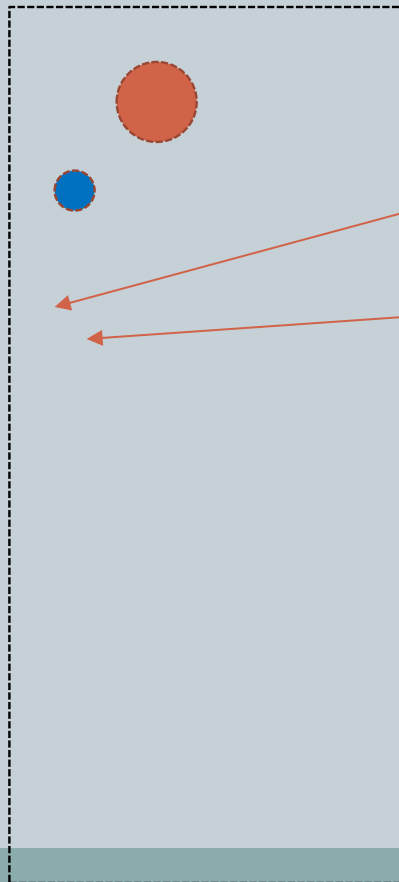
Actual Treated



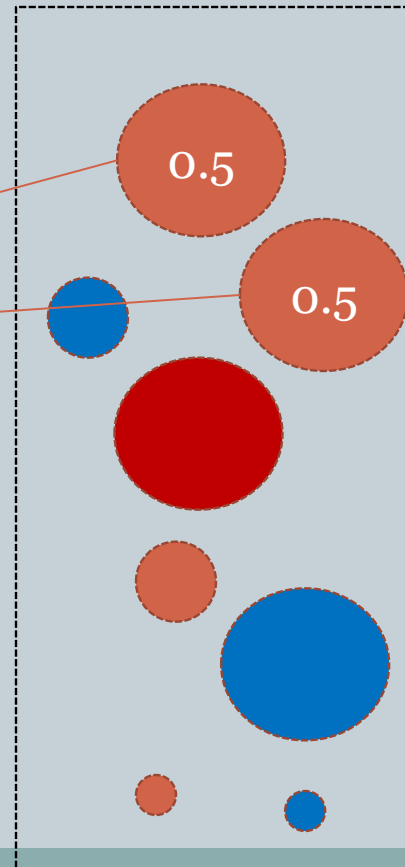
Synthetic Treated



Synthetic Control

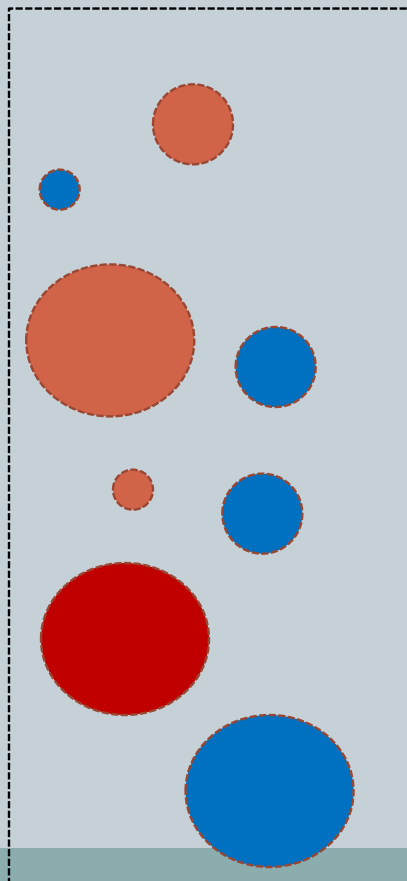


Actual Control

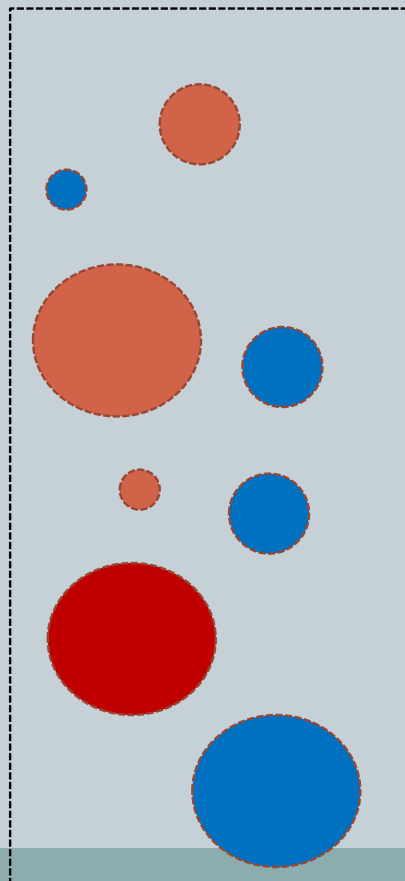




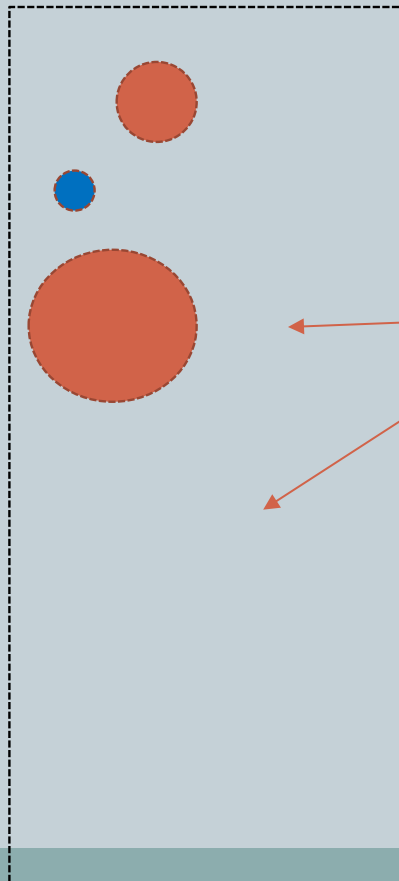
Actual Treated



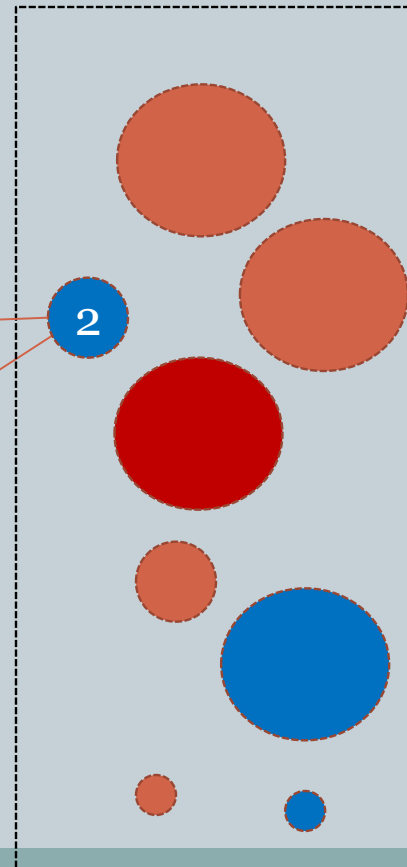
Synthetic Treated



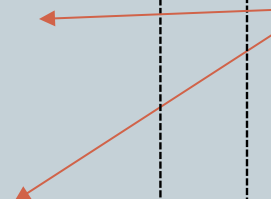
Synthetic Control



Actual Control

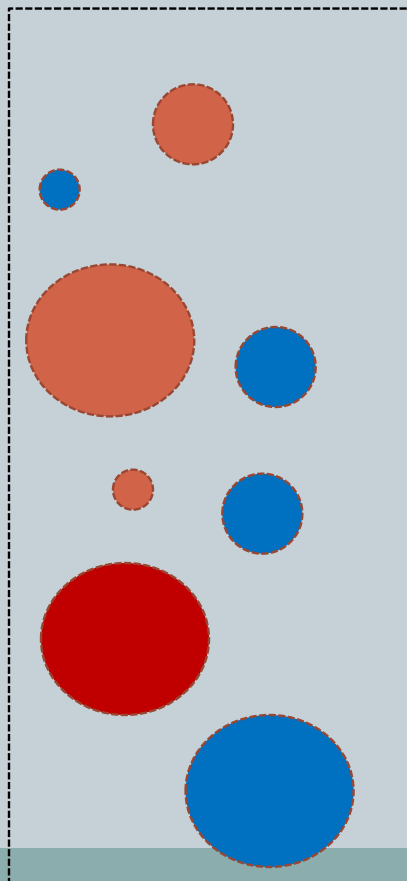


2

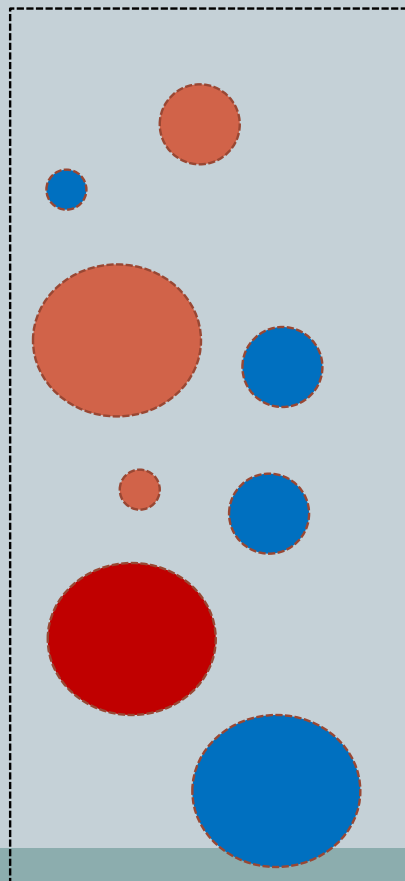




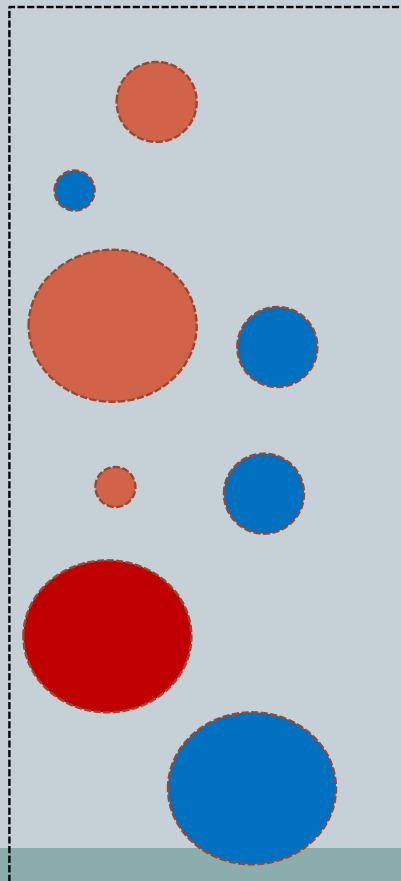
Actual Treated



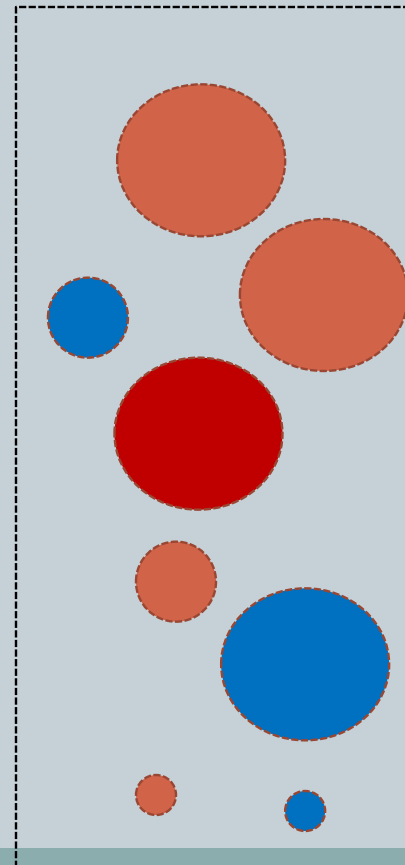
Synthetic Treated



Synthetic Control

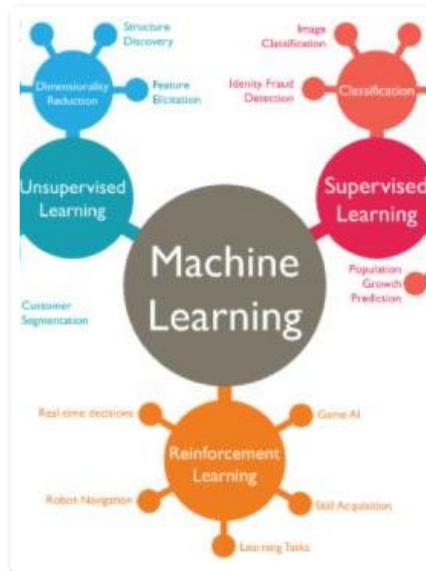


Actual Control



the common task framework

WTF is “the CTF”



the common task framework



- Much of the rapid success in prediction has come from the innovative infrastructure provided by the common task framework (CTF). Big introduction of CTF to the statistical community comes from David Donoho's "50 Years of Data Science." Mark Liberman has thought extensively about the CTF and provided us with a wealth of information.
- The "Netflix Prize" is the most famous example of the CTF.
- The Common Task Framework:
 - Starts with a curated data set
 - Available to everybody
 - A task is described (e.g., given X_i for some new i predict Y_i)
 - A metric for performance is given (e.g., $MSE(y_i, \hat{y}_i)$)
 - A hold out data set is reserved for evaluating performance
 - At the end of the CTF we just "line up" the candidate algos.



the common task framework

- The CTF is fast because it is low friction:
 - No link to sampling,
 - No link to a data generating function,
 - Ground truth is directly observable in the evaluation data,
 - Relative performance is assessed by a targeted metric,
 - A computer does a lot of the work.
- This has allowed us to innovate faster and with fewer restrictions. (Check out Tukey's Sunset Salvo and Breiman's Two Cultures.)
- The CTF has led to wildly complex algorithms which do not rely upon mathematical descriptions of how the algorithms take in variation from covariates and makes use of that variation to map to variation in the prediction space. (“**black box algorithms**”)

the common task framework

- The CTF is also kind of broken.
- The CTF was developed for HLT in the 1980s.
- The CTF was developed in the context of overcoming technical challenges with algorithms.
- It was **not** developed with the intent to determine if a particular algorithm would be “well behaved in” or “felicitous to” the real world.
- People have started working on patching the CTF:
 - ML as alchemy (e.g., [Ali Rahimi](#))
 - Fair AI (e.g., Kristian Lum, Sharad Goel)
 - High Stakes ML (e.g., Cynthia Rudin)

outcome reasoning: a superset of the CTF

- We need not confine ourselves to the CTF when “fixing” ML...
- Arvind Narayanan et al have been by doing a taxonomy of ML studies and then inductively looking for rules of success/failure:
<https://predictive-optimization.cs.princeton.edu/>

Against Predictive Optimization:
On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy

Angelina Wang, Sayash Kapoor, Solon Barocas, Arvind Narayanan.
[Read our draft paper](#)

Our argument

Predictive optimization is a distinct type of automated decision making that has **proliferated widely**. It is sold as accurate, fair, and efficient.

We identify a **recurring set of flaws** that apply broadly to predictive optimization, are hard to fix technologically, and negate its claimed benefits.

Any application of predictive optimization should be considered **illegitimate** by default unless the developer justifies how it avoids these flaws.

What is predictive optimization?

We coin the term predictive optimization to refer to automated decision-making systems where machine learning is used to make predictions about some future outcome pertaining to individuals, and those predictions are used to make decisions about them.

Automating existing rules
e.g., welfare allocation
Automated decision-making about people

Automating judgment
e.g., automated essay grading
Predictive optimization
e.g., recidivism prediction

Simulation
e.g., weather forecasting
Prediction and forecasting

ML applications
e.g., email spam filtering
e.g., traffic forecasting

outcome reasoning: a superset of the CTF

- We need not confine ourselves to the CTF when “fixing” ML...
- We worked out a more general framework we call “outcome reasoning” – we propose modifications for using outcome reasoning in deployment (rather than in development).
<https://muse.jhu.edu/article/883478/summary>
- Outcome reasoning focuses on evaluating performance purely in the space of the outcome, bypassing detailed evaluation of the mapping of the covariates to the outcome.
- A friendly introduction to this framework, that highlights the differences between traditional styles of reasoning used in causal inference (i.e., warranted and model reasoning) and the CTF:
<https://muse.jhu.edu/article/799741>

resources at Stanford



resources at Stanford



- Short term help (pretty much free):
 - Statistics department's [Statistical Consulting](#) hours
 - [SPECTRUM](#)

resources at Stanford



- [Data Studio](#) (free):
 - Bunch of really smart statisticians listen to your proposal and come up with amazing ideas about how to make your research awesome.
 - Sometimes they'll offer to help (longshot).
 - SPECTRUM is a gatekeeper and your project needs to be interesting.

resources at Stanford



- Classes (not exhaustive)
- **Applied (bio)statistics**
 - **EPI 261: Intermediate Biostatistics: Discrete Data** (Mike Baiocchi, Winter)
 - **EPI 262: Intermediate Biostatistics: Regression** (Mike Baiocchi, Spring)
 - **EPI 225: Intro to Epidemiologic and Clinical Research Methods** (Rita Popat, Fall)
- **Causal inference for obs. studies**
 - **ECON 272: Intermediate Econometrics III** (Guido Imbens, Spring)
 - **EPI 227: Advanced Epidemiologic Methods** (Michelle Odden, Fall)
 - **EPI 292: Advanced Methods for Obs. Studies** (Mike Baiocchi, Online)
- **Causal inference with experiments**
 - **STATS 363: Design of Experiments** (Art Owen, Winter)
- **Both obs. and exp. causal inference**
 - **BIODS 251: Causal Inf. in Trials and Obs. Studies** (Lu Tian, Ying Lu, Winter)
 - **STATS 361: Causal Inference** (Stefan Wager, Winter, *heavy on theory*)