


Explanation and Prediction

Double-robust estimators

 Cornell University

arXiv.org > stat > arXiv:1608.00060

Search or Ar
(Help | Advanced)

Statistics > Machine Learning

Double/Debiased Machine Learning for Treatment and Causal Parameters

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins

(Submitted on 30 Jul 2016 (v1), last revised 12 Dec 2017 (this version, v6))

Most modern supervised statistical/machine learning (ML) methods are explicitly designed to solve prediction problems very well. Achieving this goal does not imply that these methods automatically deliver good estimators of causal parameters. Examples of such parameters include individual regression coefficients, average treatment effects, average lifts, and demand or supply elasticities. In fact, estimates of such causal parameters obtained via naively plugging ML estimators into estimating equations for such parameters can behave very poorly due to the regularization bias. Fortunately, this regularization bias can be removed by solving auxiliary prediction problems via ML tools. Specifically, we can form an orthogonal score for the target low-dimensional parameter by combining auxiliary and main ML predictions. The score is then used to build a de-biased estimator of the target parameter which typically will converge at the fastest possible $1/\sqrt{n}$ rate and be approximately unbiased and normal, and from which valid confidence intervals for these parameters of interest may be constructed. The resulting method thus could be called a "double ML" method because it relies on estimating primary and auxiliary predictive models. In order to avoid overfitting, our construction also makes use of the K-fold sample splitting, which we call cross-fitting. This allows us to use a very broad set of ML predictive methods in solving the auxiliary and main prediction problems, such as random forest, lasso, ridge, deep neural nets, boosted trees, as well as various hybrids and aggregators of these methods.

Comments: 71 pages, 2 figures

Subjects: **Machine Learning (stat.ML)**; Econometrics (econ.EM)

MSC classes: 62G

Cite as: arXiv:1608.00060 [stat.ML]
(or arXiv:1608.00060v6 [stat.ML] for this version)

Submission history

From: Christian Hansen [\[view email\]](#)

[\[v1\]](#) Sat, 30 Jul 2016 01:58:04 UTC (58 KB)

[\[v2\]](#) Fri, 5 Aug 2016 05:48:45 UTC (58 KB)

[\[v3\]](#) Thu, 18 Aug 2016 10:45:11 UTC (59 KB)

[\[v4\]](#) Fri, 30 Dec 2016 20:18:36 UTC (65 KB)

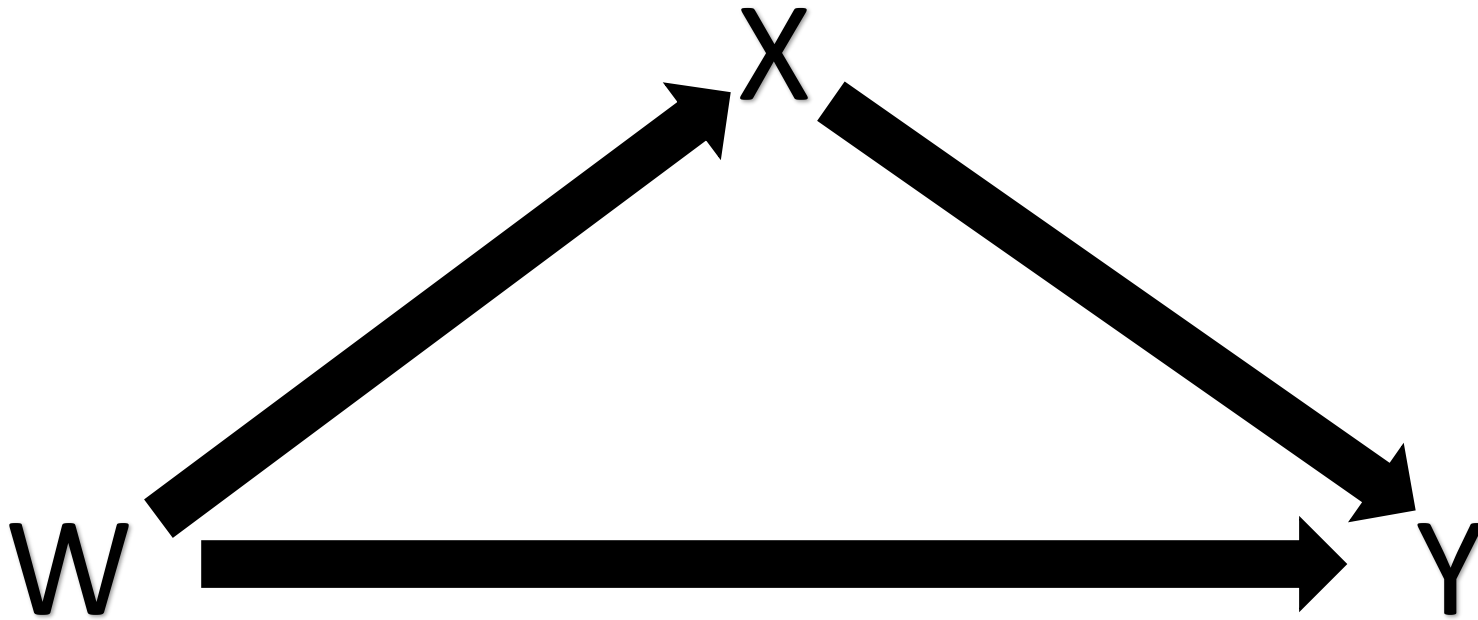
[\[v5\]](#) Wed, 21 Jun 2017 06:27:04 UTC (130 KB)

[\[v6\]](#) Tue, 12 Dec 2017 20:13:33 UTC (130 KB)

Double-robust estimators

- Run a LASSO on $Y \sim X$, call the set of variables selected A
- Run a LASSO on $W \sim X$, call the set of variables selected B
- Run a linear regression $Y \sim (W * (A \cup B))$, the coefficient for W is the ATE
- Called “doubly robust” because bias is a product of bias in two LASSOs (i.e. two numbers between zero and one multiplied together)

Double-robust estimators

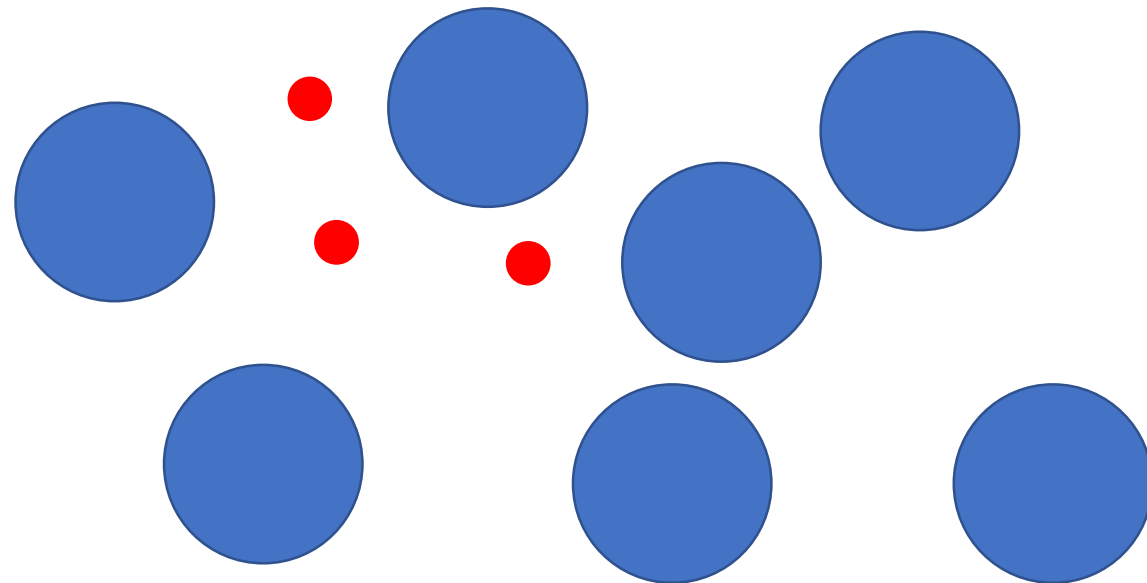
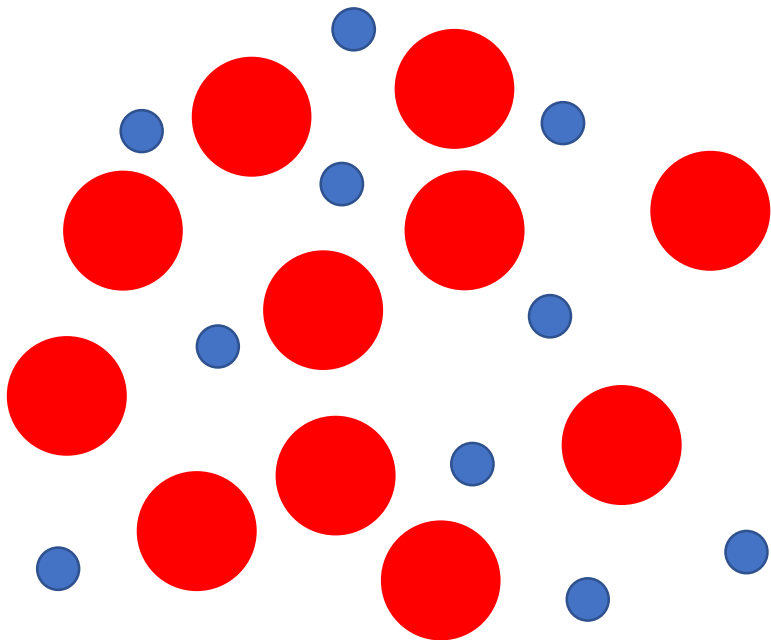
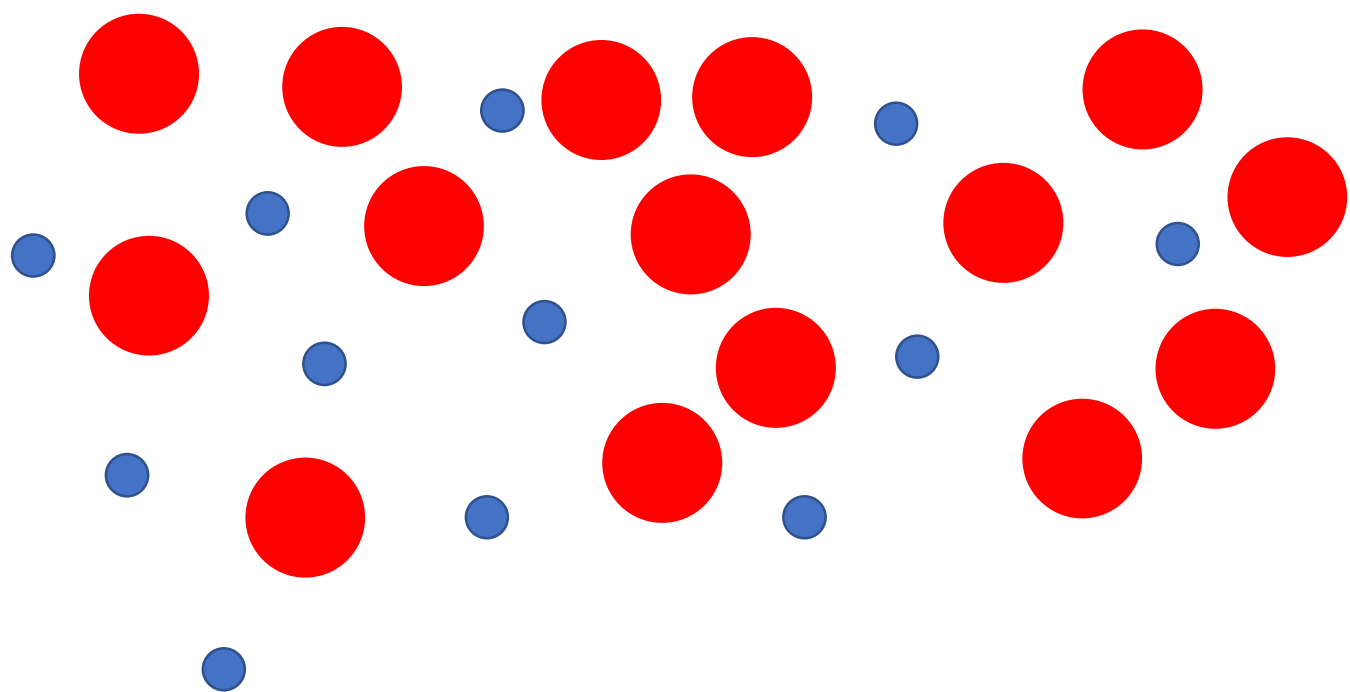
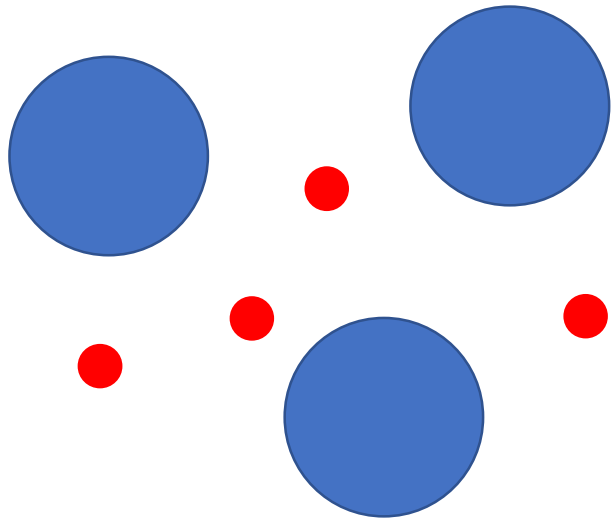


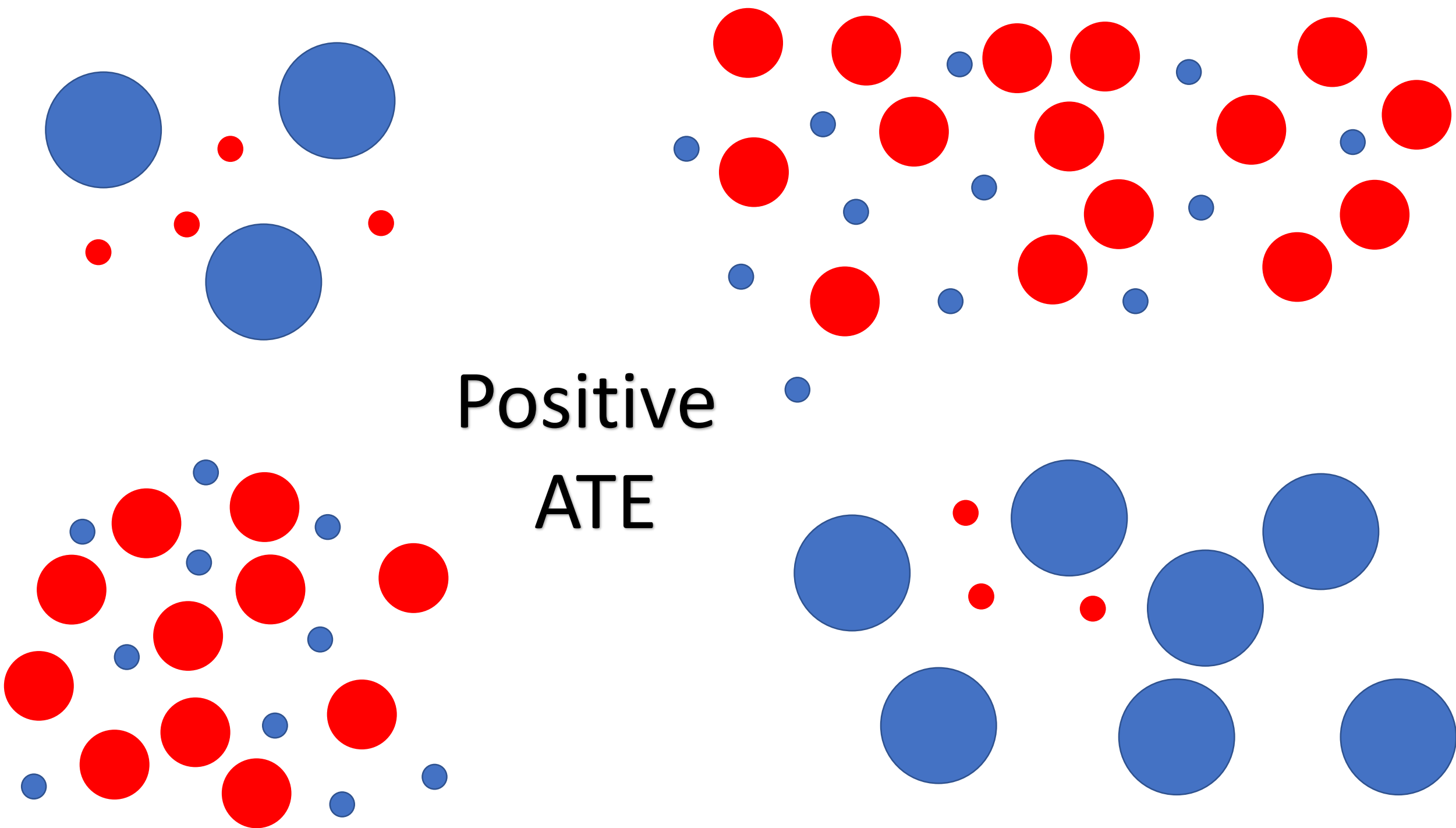
Heterogeneous treatment effects

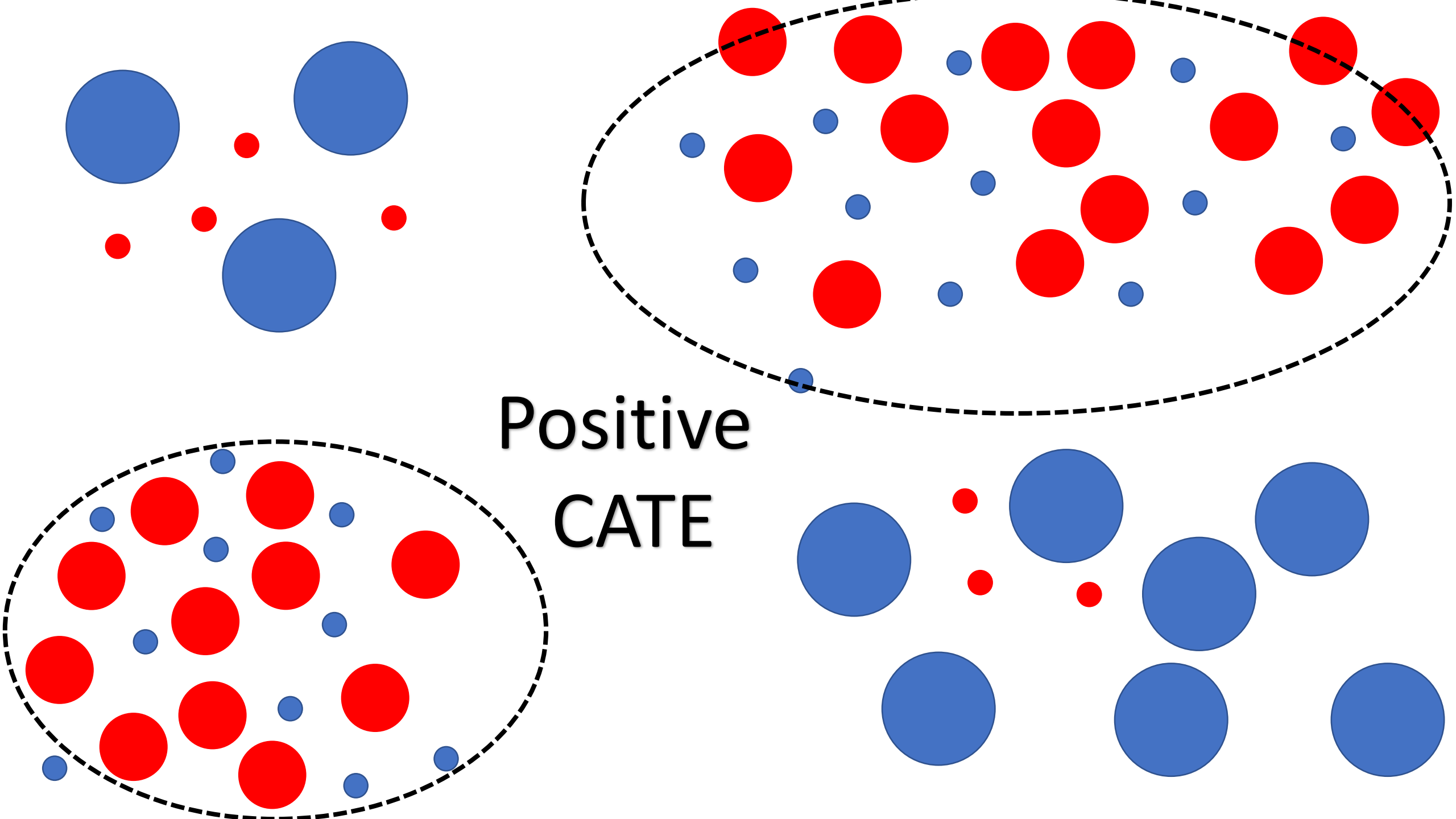
- Sometimes treatments work differently for different folks
- If this is a function of observables, we can estimate
 - Personalized marketing
 - Personalized medicine
 - Policy optimization
- We call the ATE conditional on observables the CATE, “conditional average treatment effect”

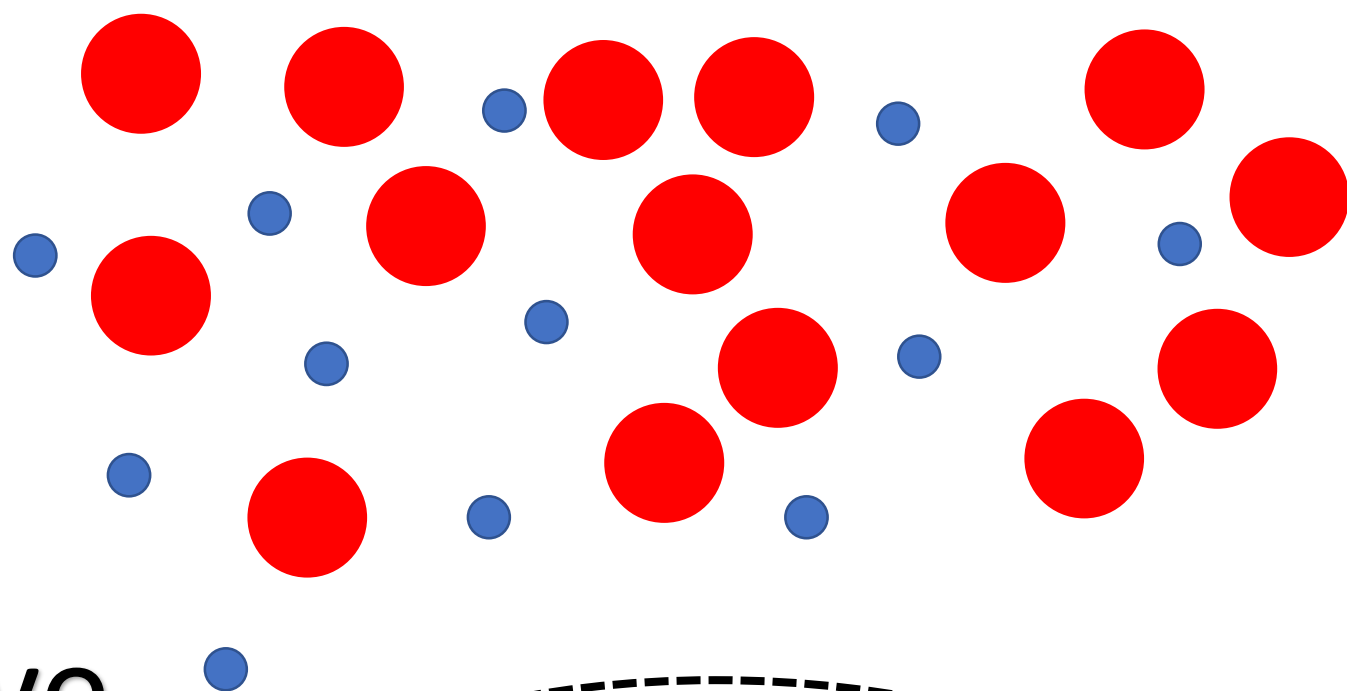
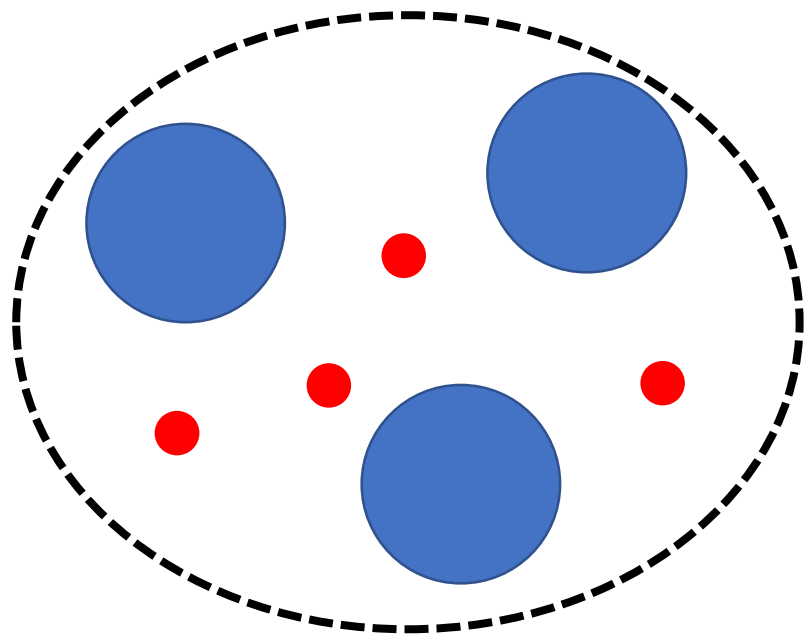
Heterogeneous treatment effects (cont.)

- Hard to estimate in practice
 - Shouldn't just run every possible subset (overfitting!)
 - Hard to know *a priori* which groups will have different CATEs
- How can we use the data to validly find groups with different CATEs?
- More generally, how do we get CATE?
- In general, having the CATE is better than having the ATE

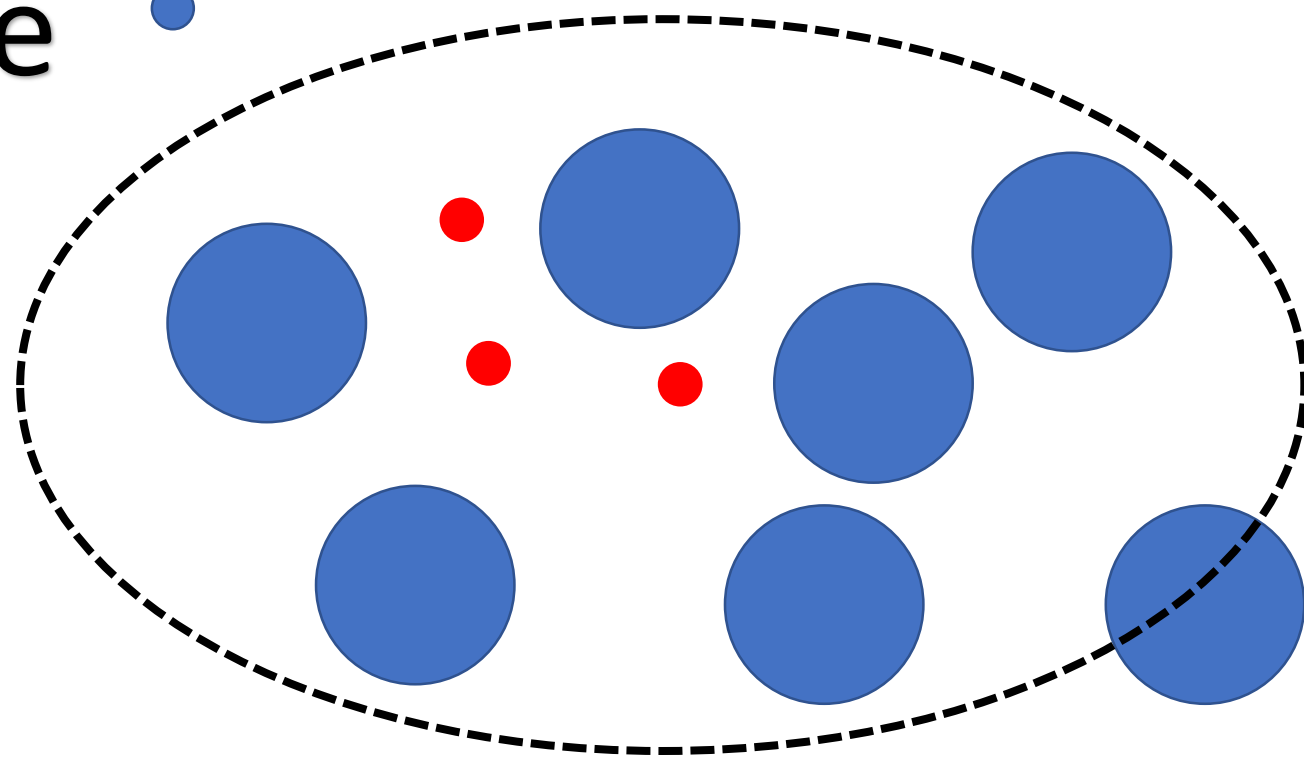
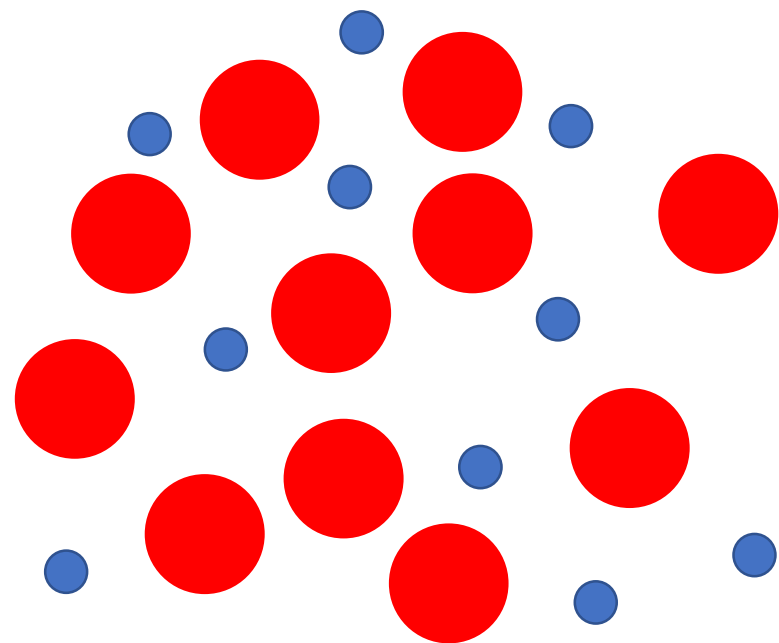


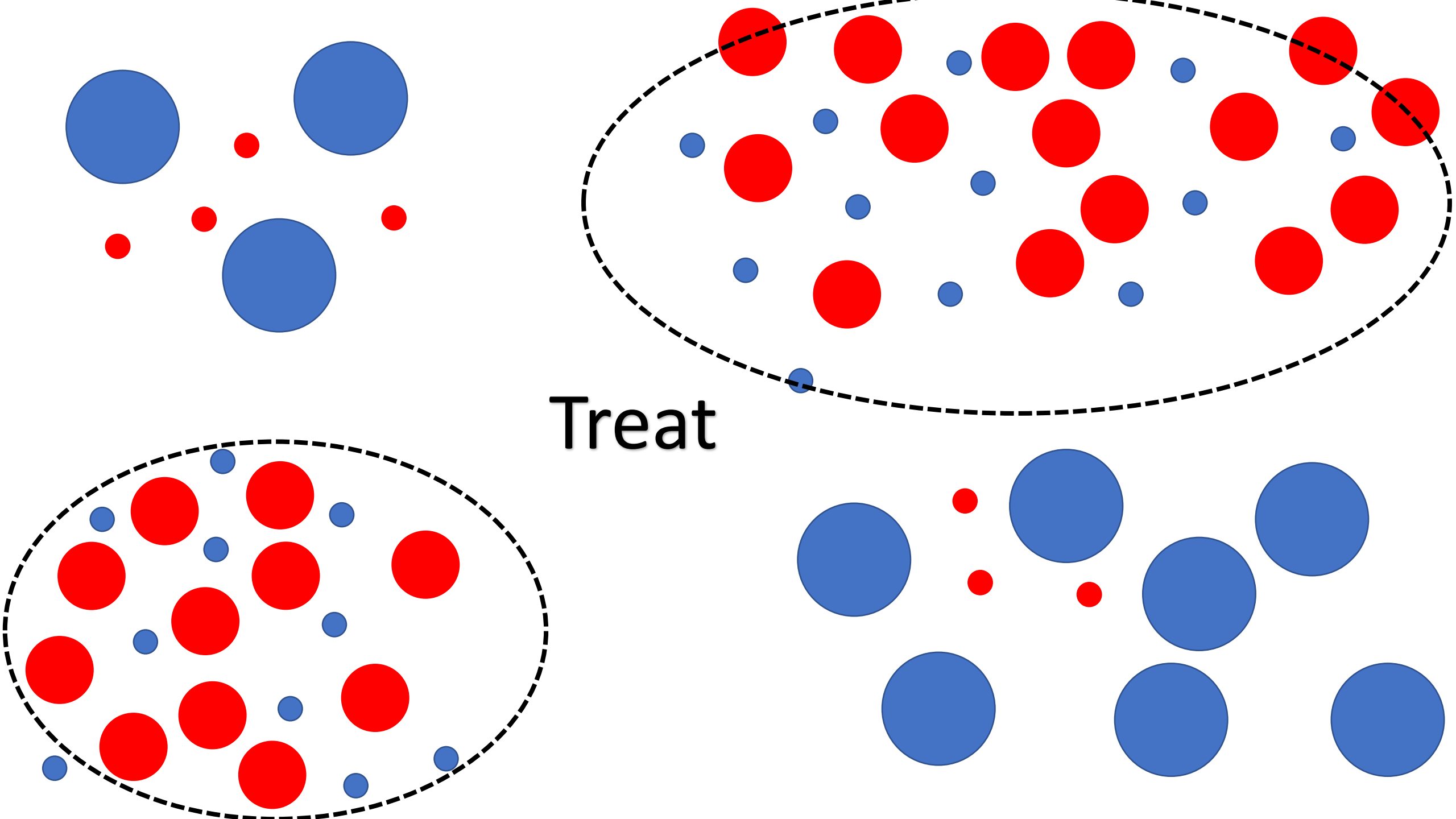


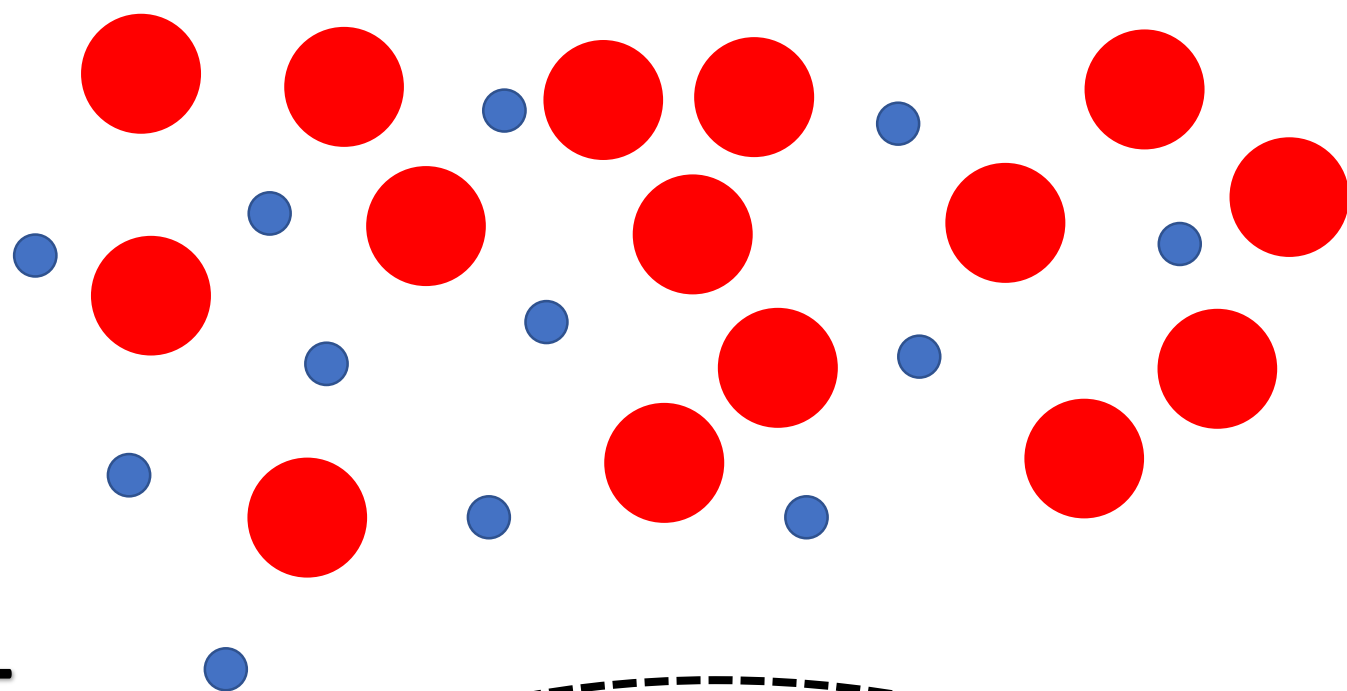
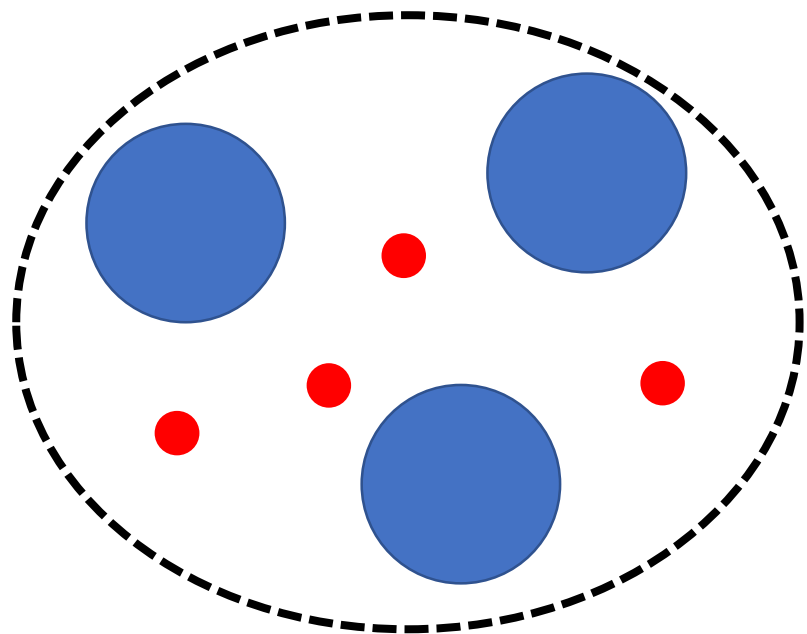




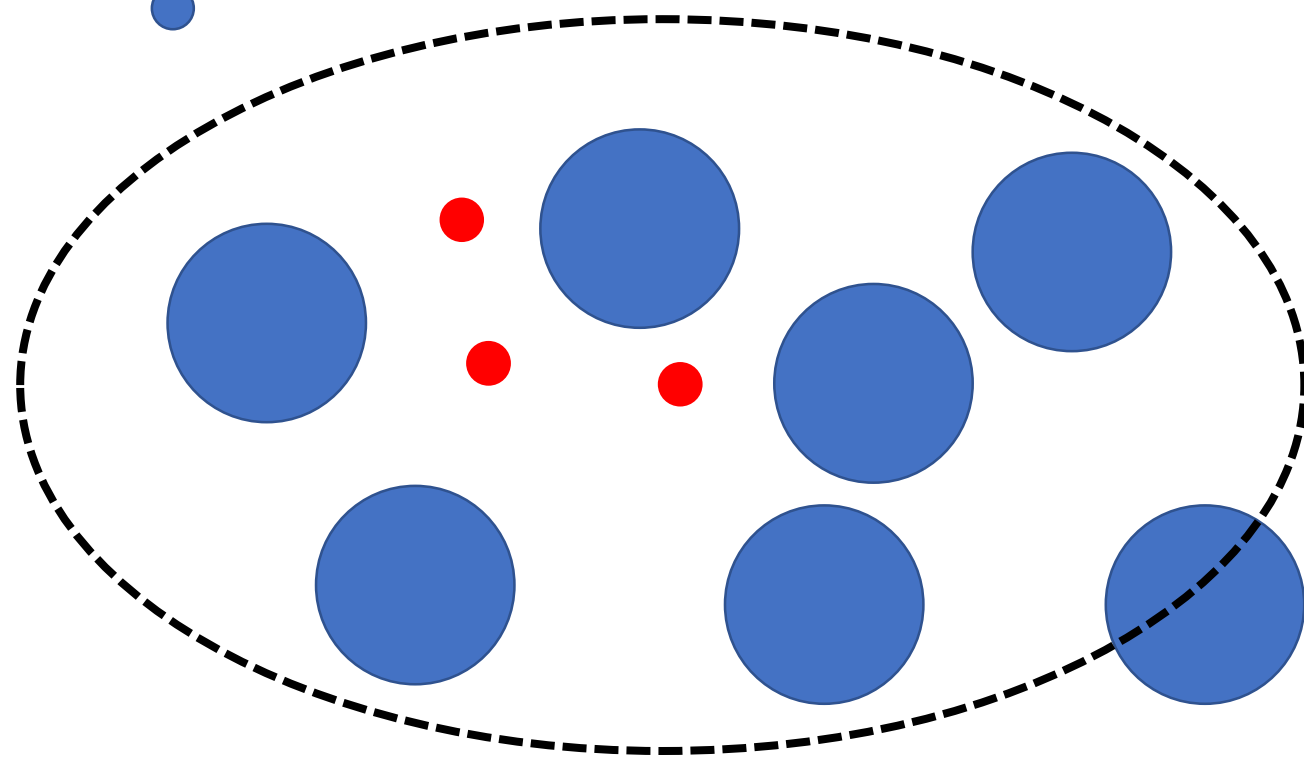
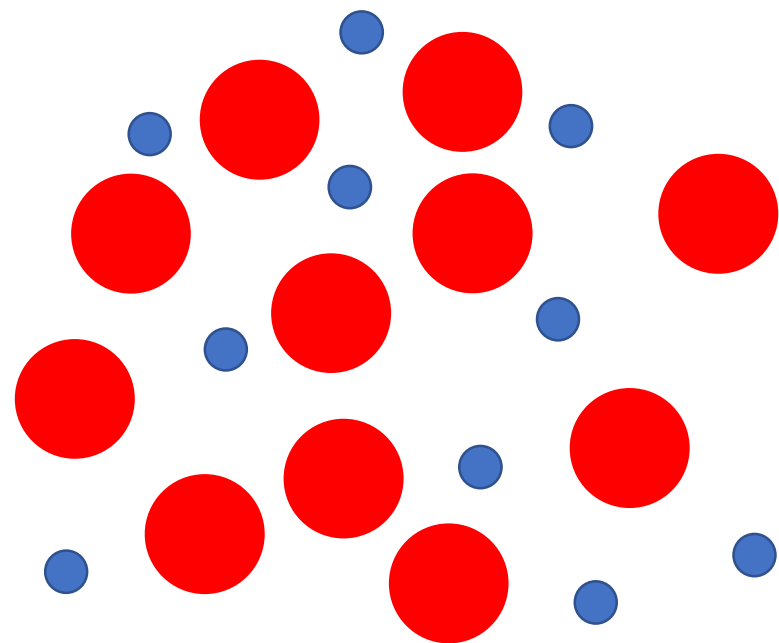
Negative
CATE








Don't
treat



Causal trees

 Cornell University

arXiv.org > stat > arXiv:1504.01132v3

Search or Art
(Help | Advanced s

Statistics > Machine Learning

Recursive Partitioning for Heterogeneous Causal Effects

Susan Athey, Guido Imbens

(Submitted on 5 Apr 2015 (v1), last revised 30 Dec 2015 (this version, v3))

In this paper we study the problems of estimating heterogeneity in causal effects in experimental or observational studies and conducting inference about the magnitude of the differences in treatment effects across subsets of the population. In applications, our method provides a data-driven approach to determine which subpopulations have large or small treatment effects and to test hypotheses about the differences in these effects. For experiments, our method allows researchers to identify heterogeneity in treatment effects that was not specified in a pre-analysis plan, without concern about invalidating inference due to multiple testing. In most of the literature on supervised machine learning (e.g. regression trees, random forests, LASSO, etc.), the goal is to build a model of the relationship between a unit's attributes and an observed outcome. A prominent role in these methods is played by cross-validation which compares predictions to actual outcomes in test samples, in order to select the level of complexity of the model that provides the best predictive power. Our method is closely related, but it differs in that it is tailored for predicting causal effects of a treatment rather than a unit's outcome. The challenge is that the "ground truth" for a causal effect is not observed for any individual unit: we observe the unit with the treatment, or without the treatment, but not both at the same time. Thus, it is not obvious how to use cross-validation to determine whether a causal effect has been accurately predicted. We propose several novel cross-validation criteria for this problem and demonstrate through simulations the conditions under which they perform better than standard methods for the problem of causal effects. We then apply the method to a large-scale field experiment re-ranking results on a search engine.

Subjects: **Machine Learning (stat.ML)**; Econometrics (econ.EM)

Cite as: **arXiv:1504.01132 [stat.ML]**

(or **arXiv:1504.01132v3 [stat.ML]** for this version)

Submission history

From: Susan Athey [[view email](#)]

[v1] Sun, 5 Apr 2015 16:01:44 UTC (168 KB)

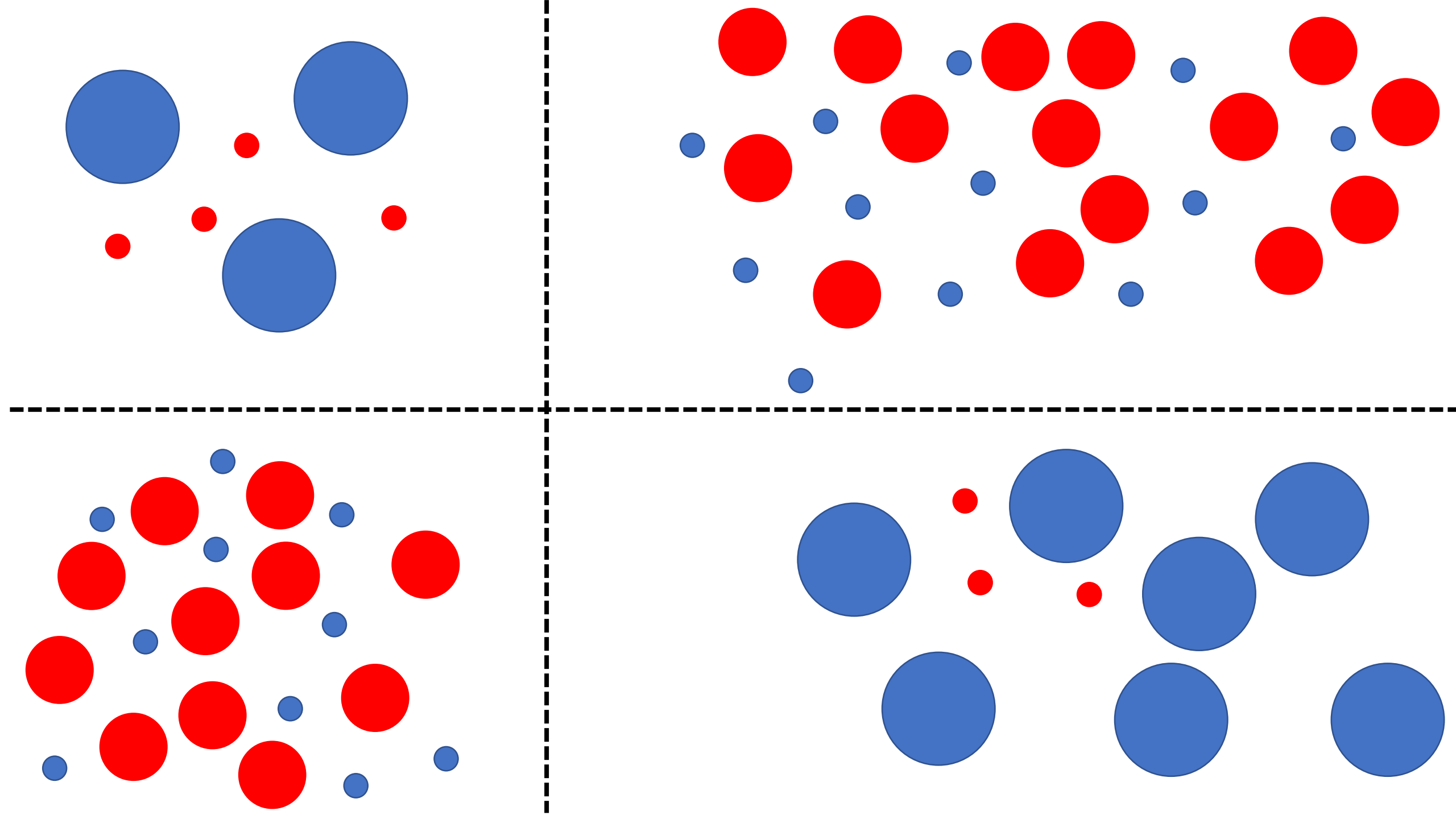
[v2] Mon, 20 Jul 2015 18:24:56 UTC (43 KB)

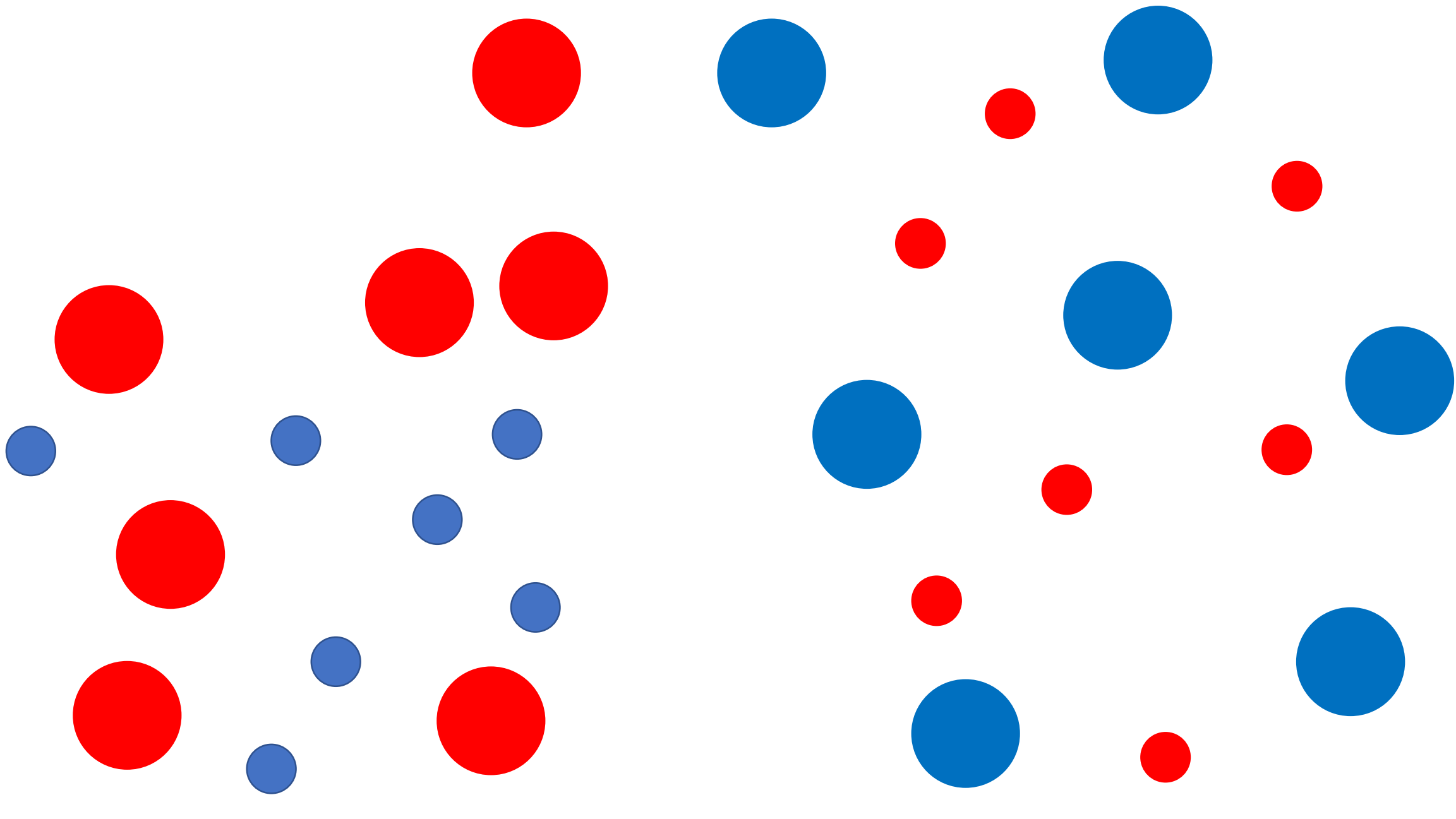
[v3] Wed, 30 Dec 2015 18:01:20 UTC (30 KB)



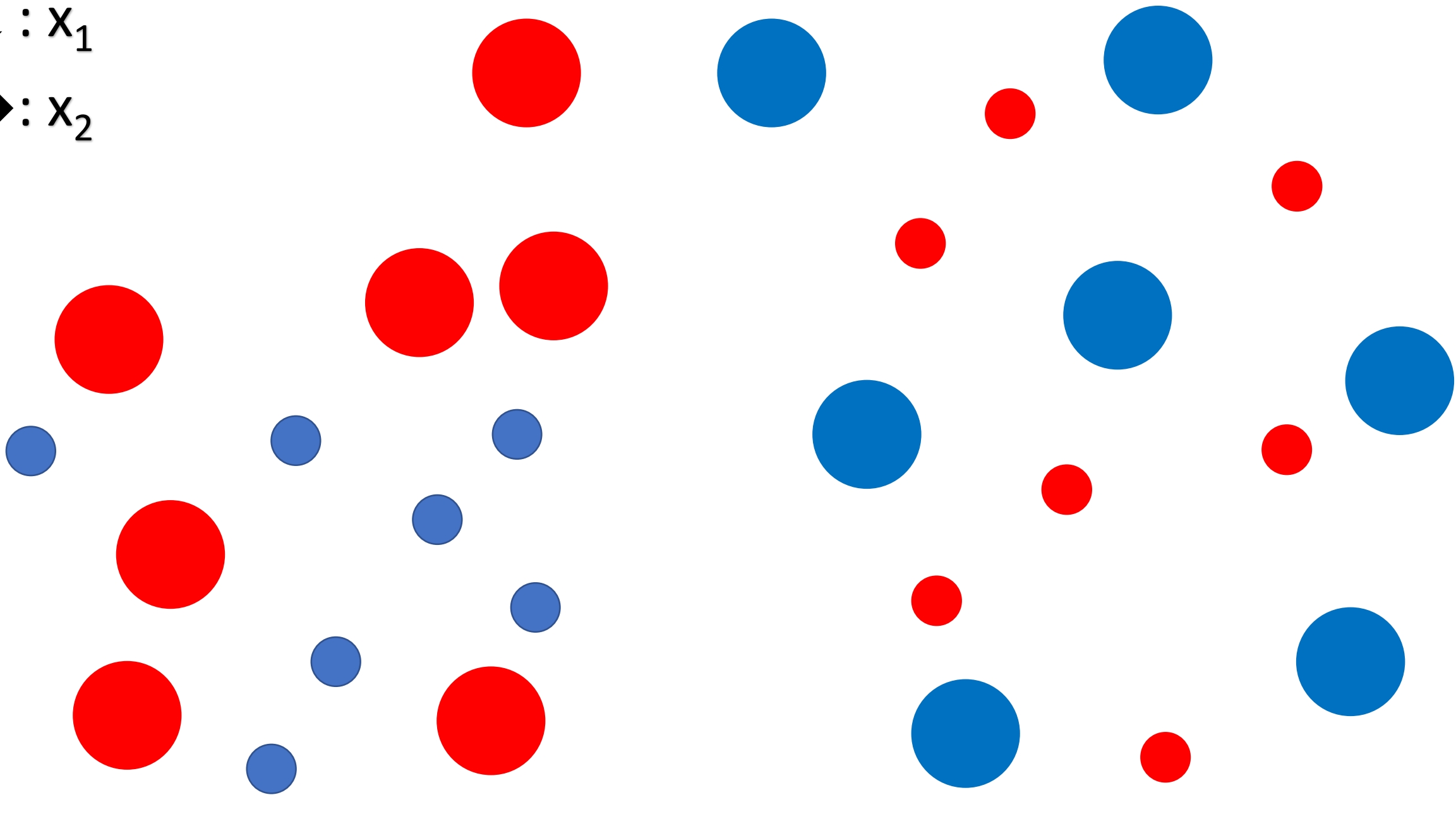
Look for
splits that
Maximize

Difference between
ATE in leaves

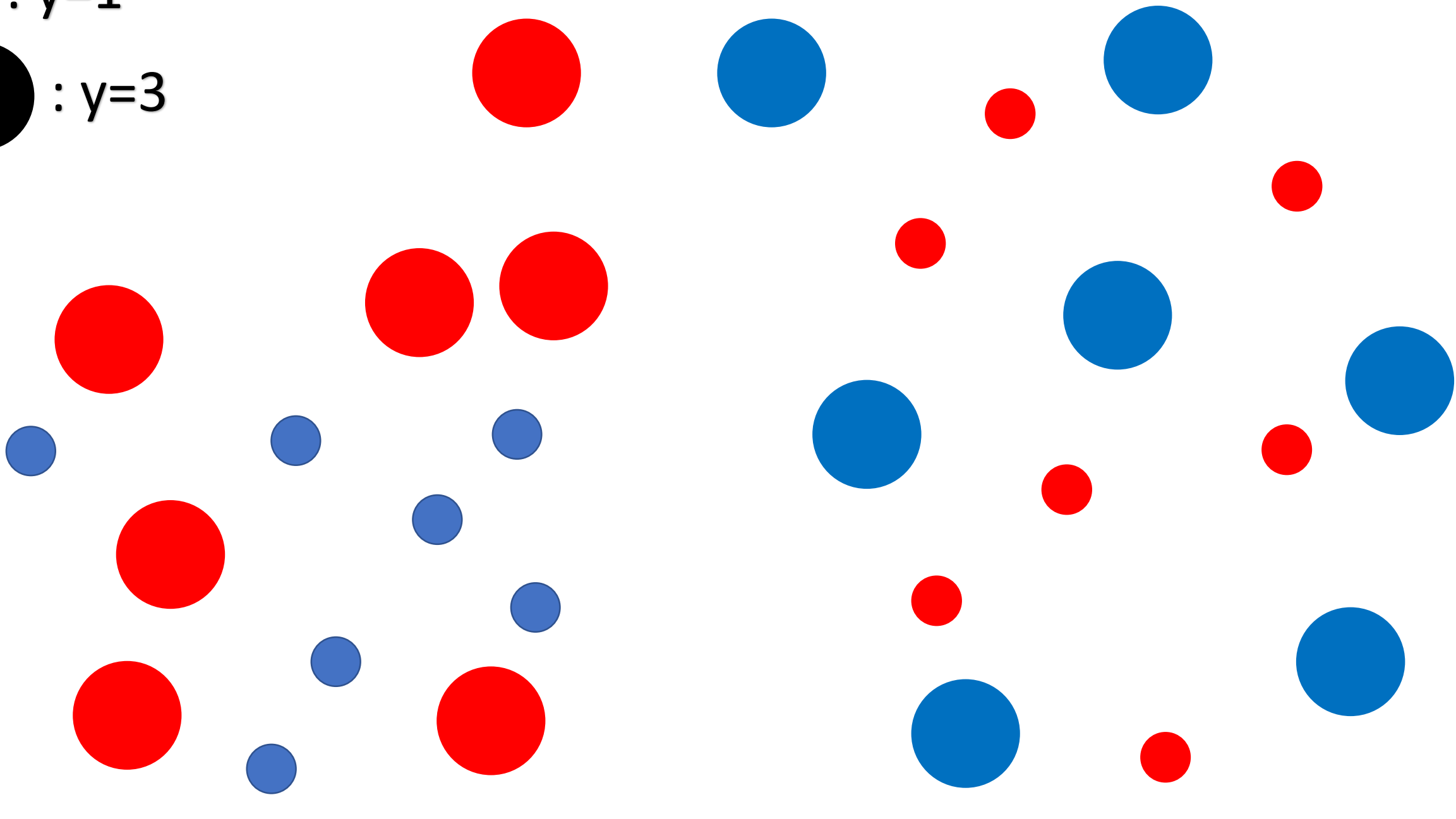




↓ : x_1
→ : x_2

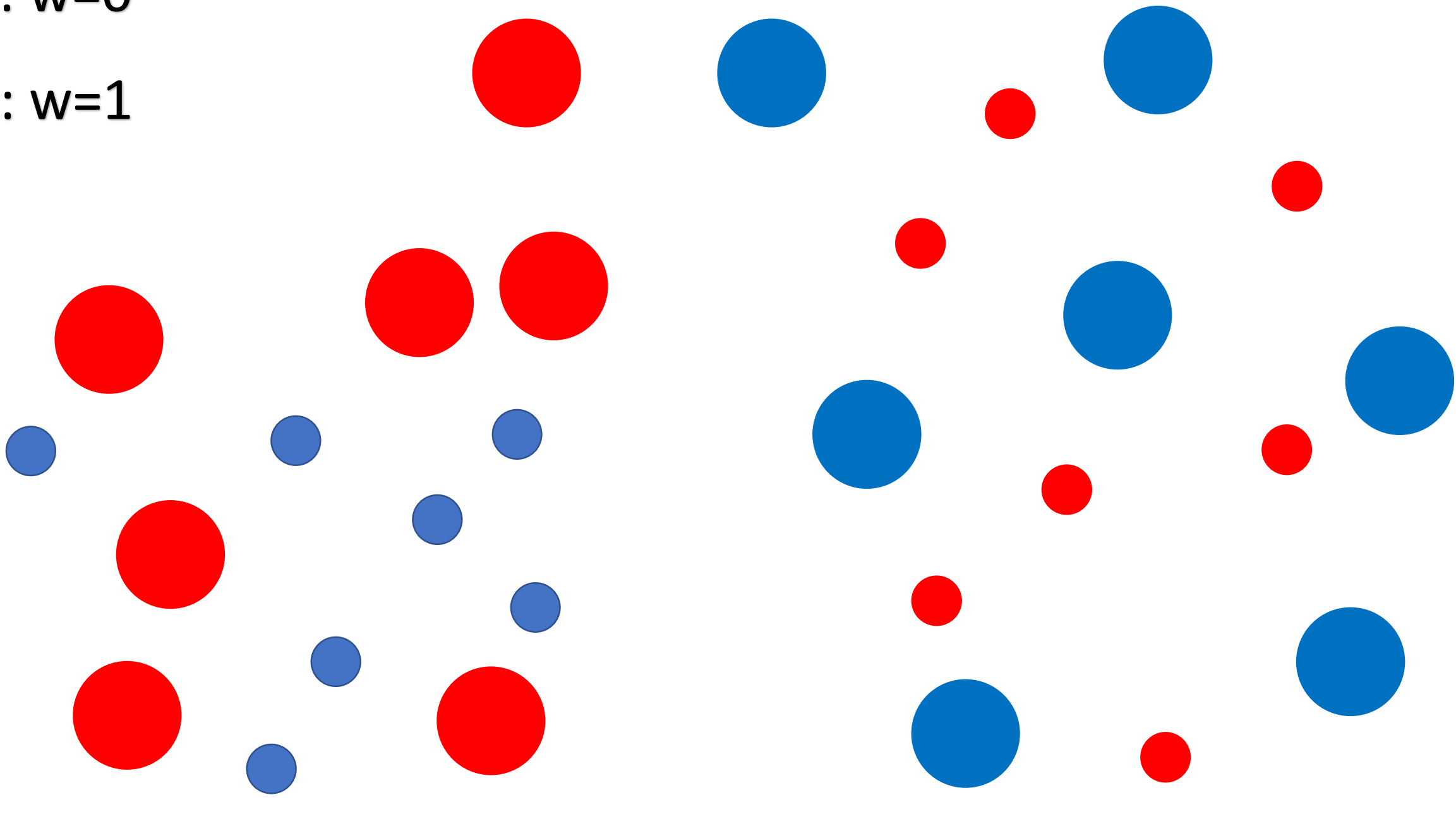


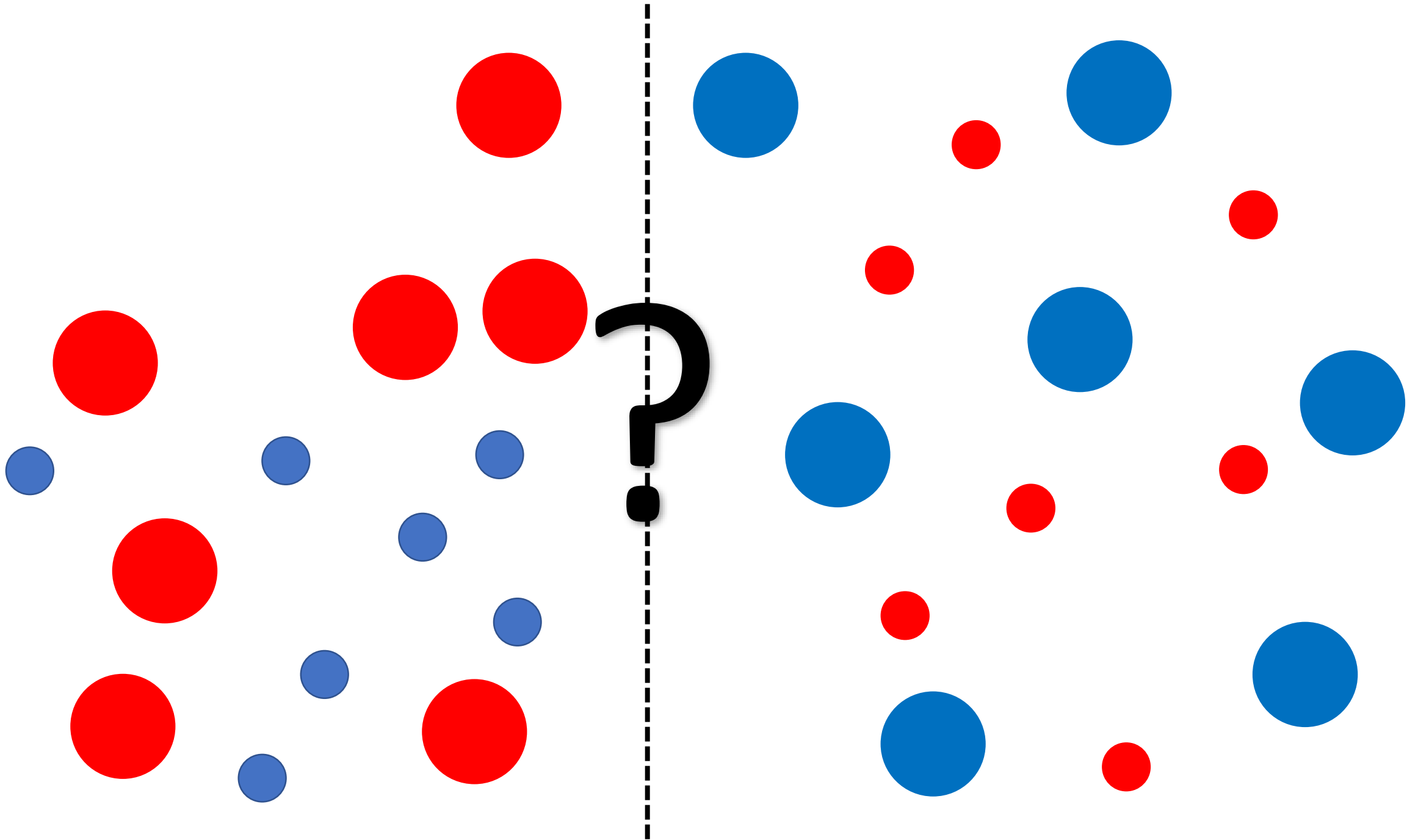
● : $y=1$
● : $y=3$



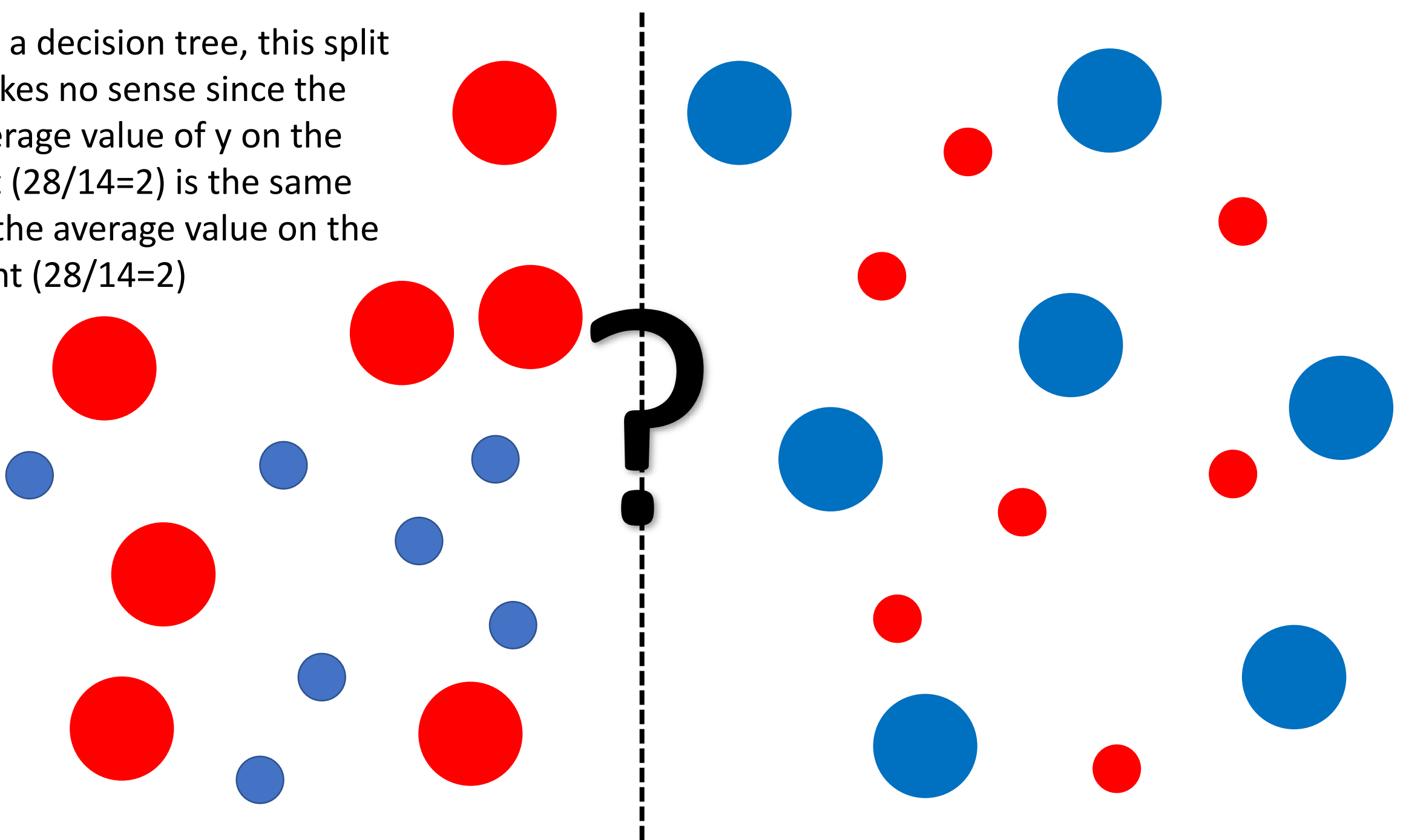
●: $w=0$

●: $w=1$

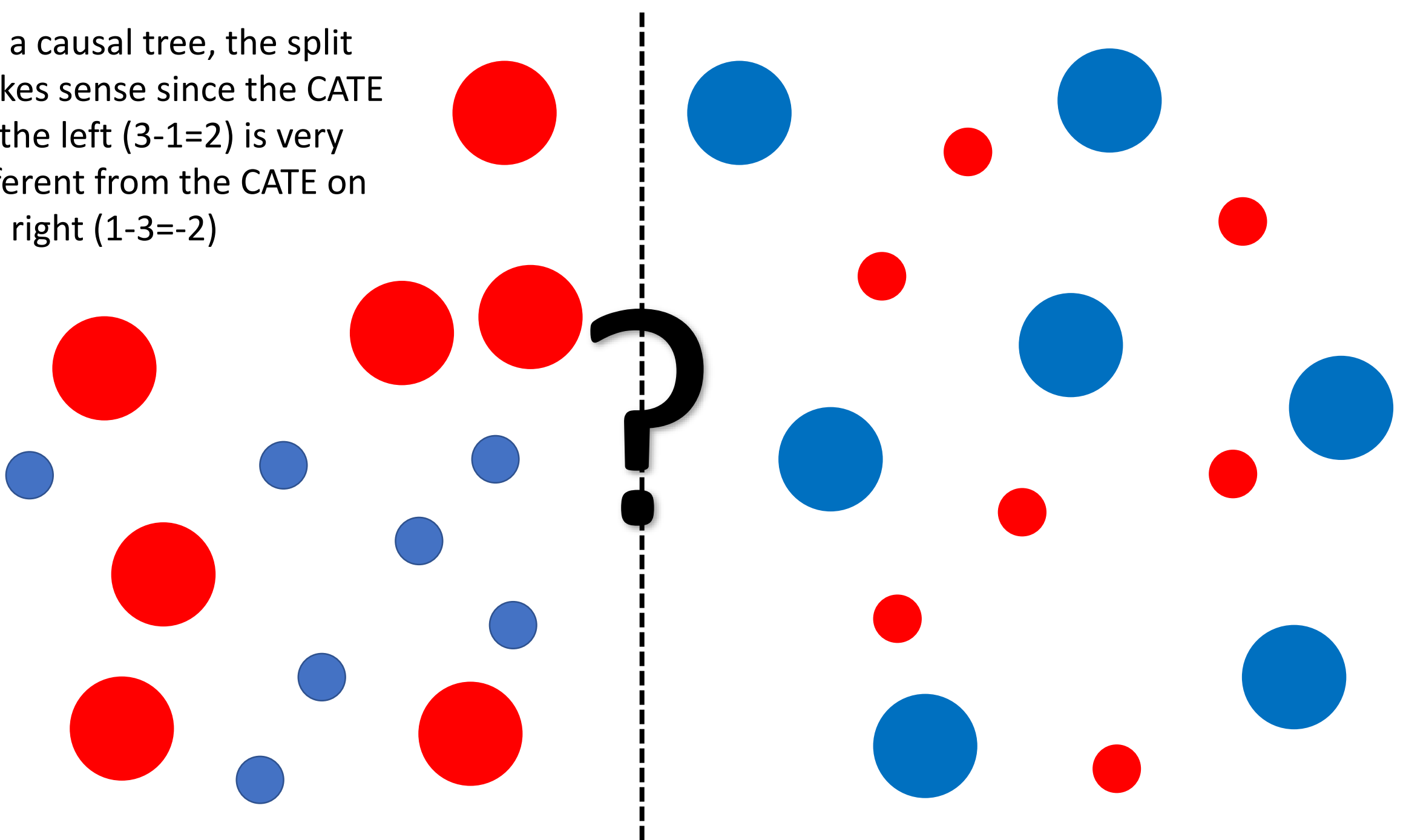




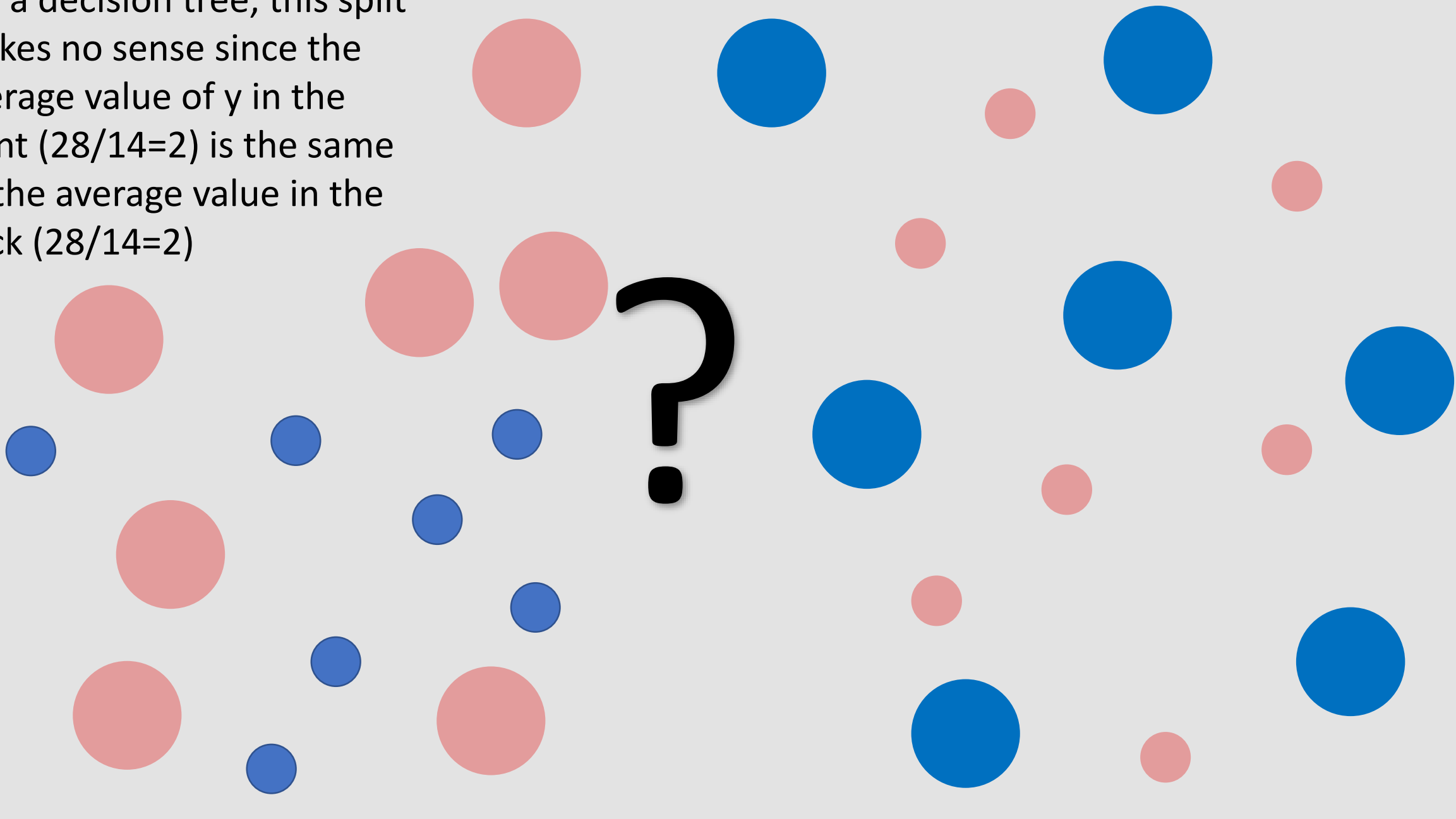
For a decision tree, this split
makes no sense since the
average value of y on the
left ($28/14=2$) is the same
as the average value on the
right ($28/14=2$)

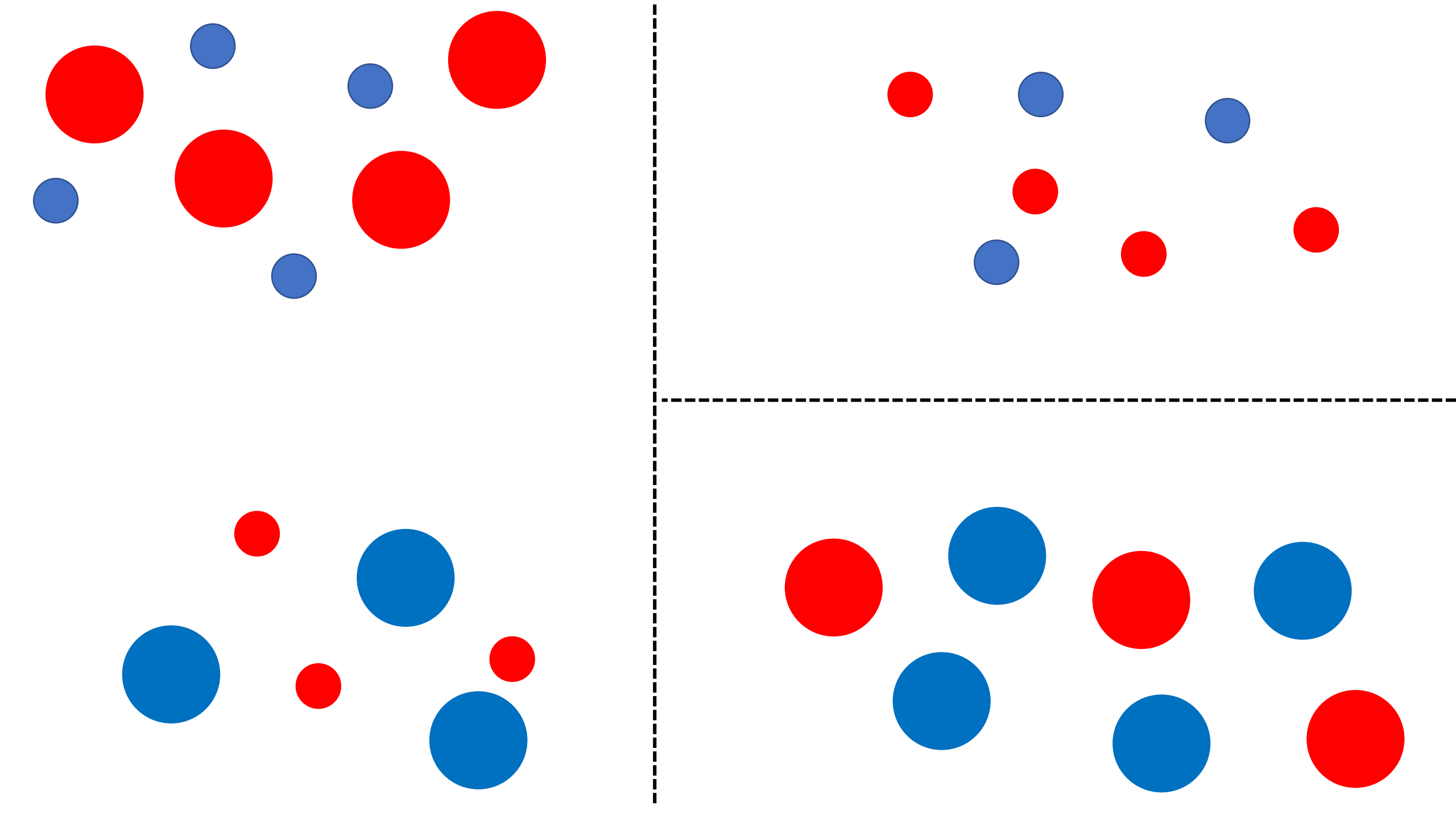


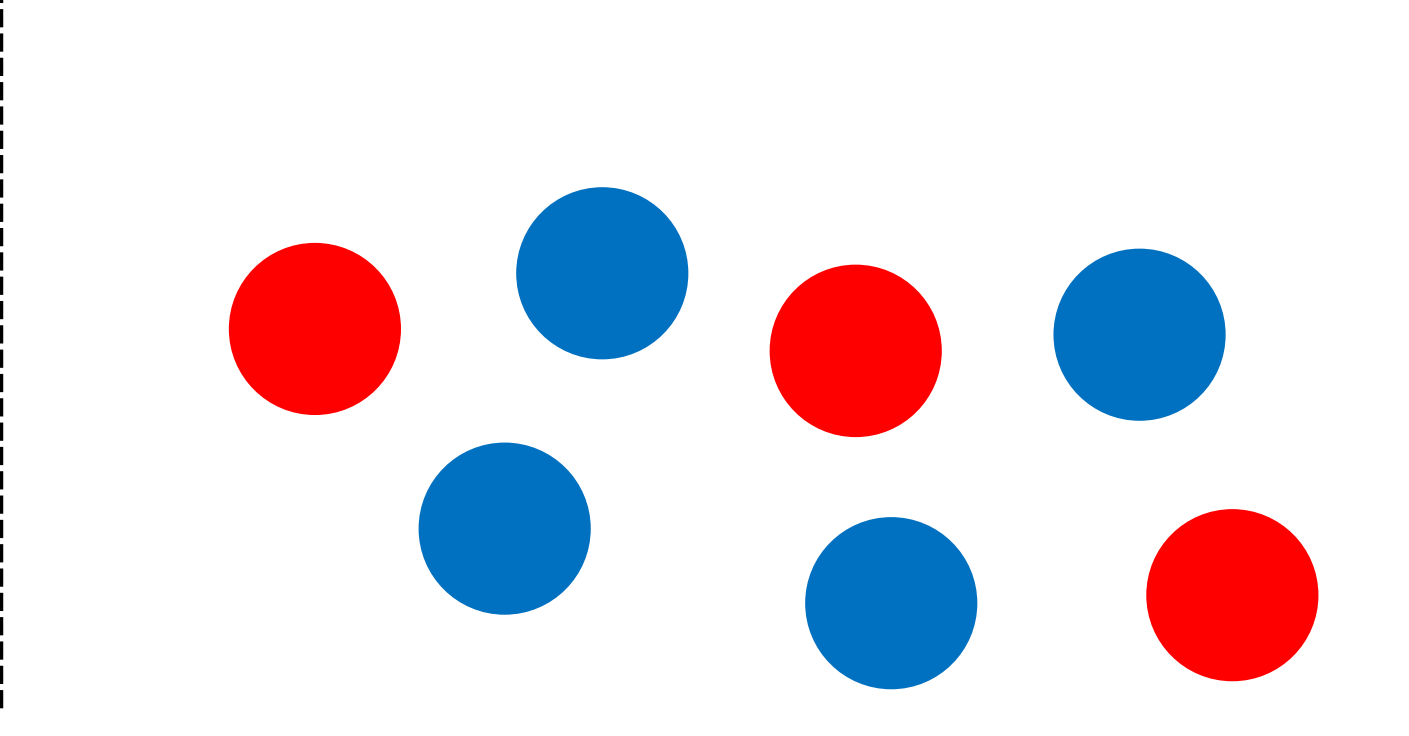
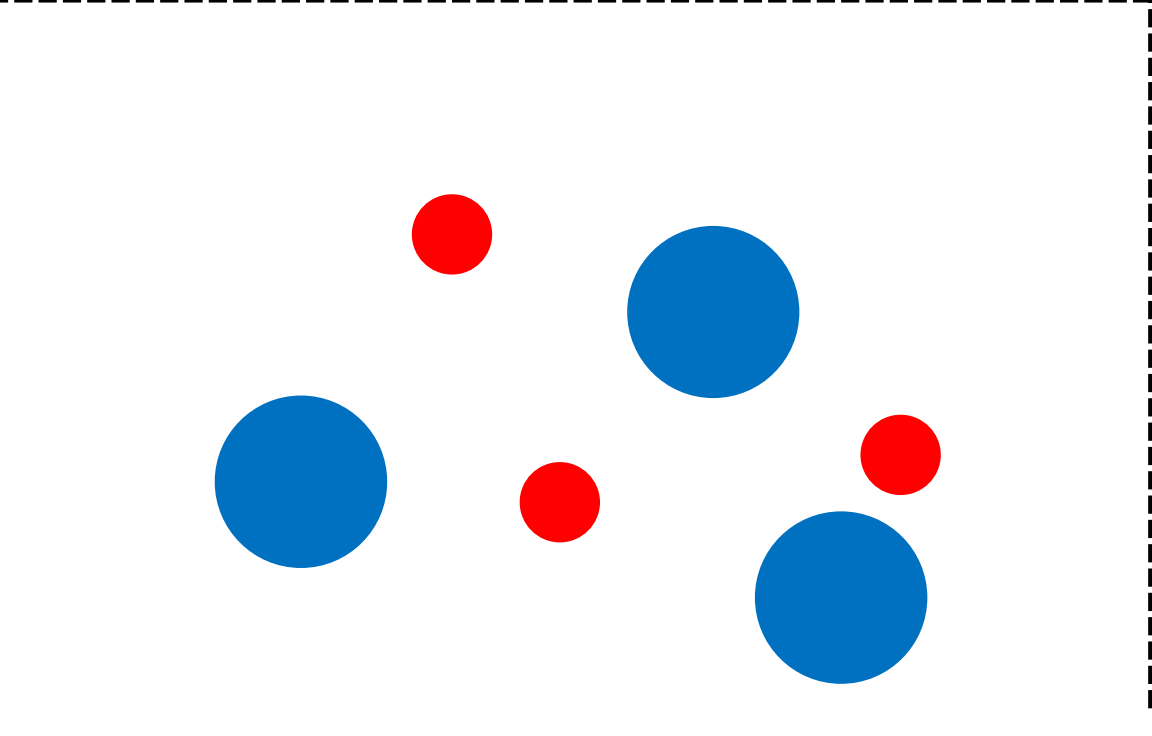
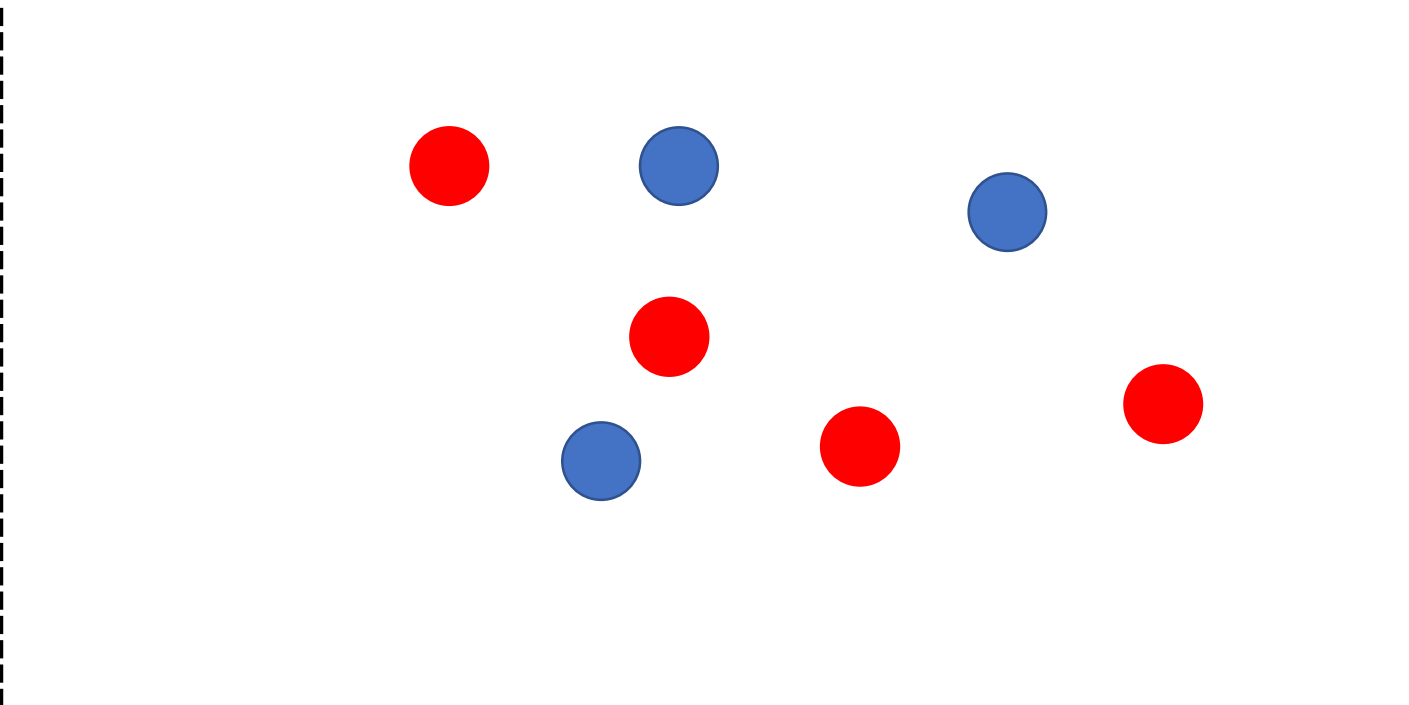
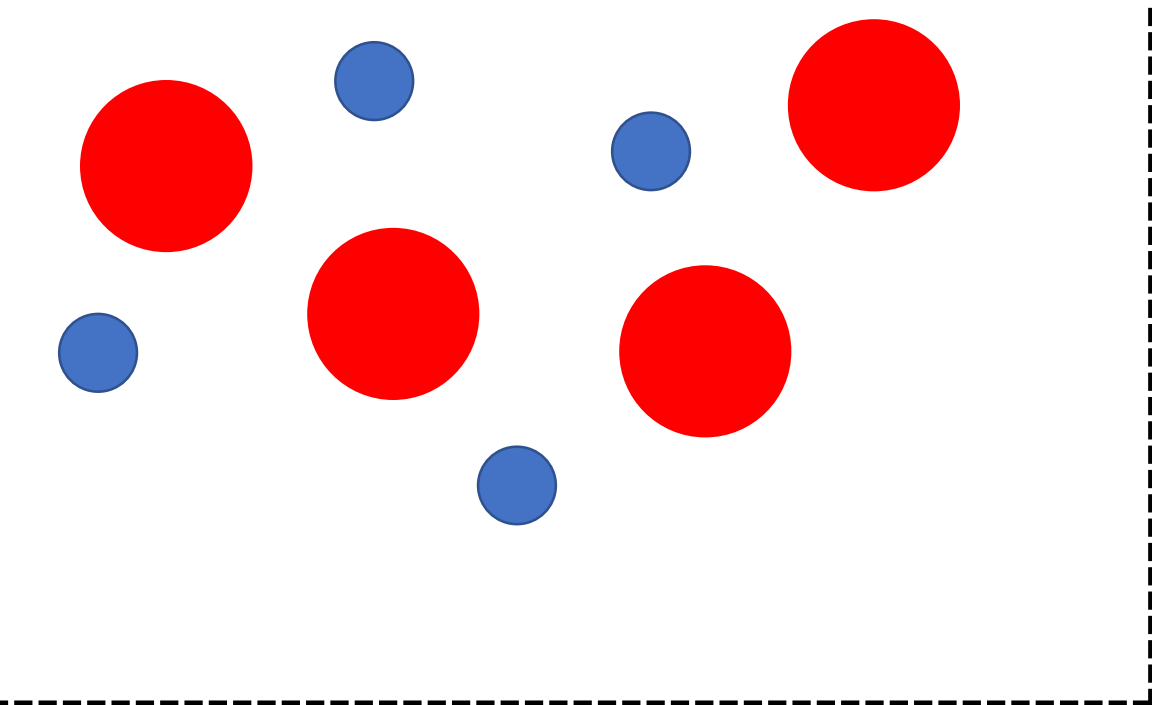
For a causal tree, the split makes sense since the CATE on the left ($3-1=2$) is very different from the CATE on the right ($1-3=-2$)



For a decision tree, this split
makes no sense since the
average value of y in the
front ($28/14=2$) is the same
as the average value in the
back ($28/14=2$)





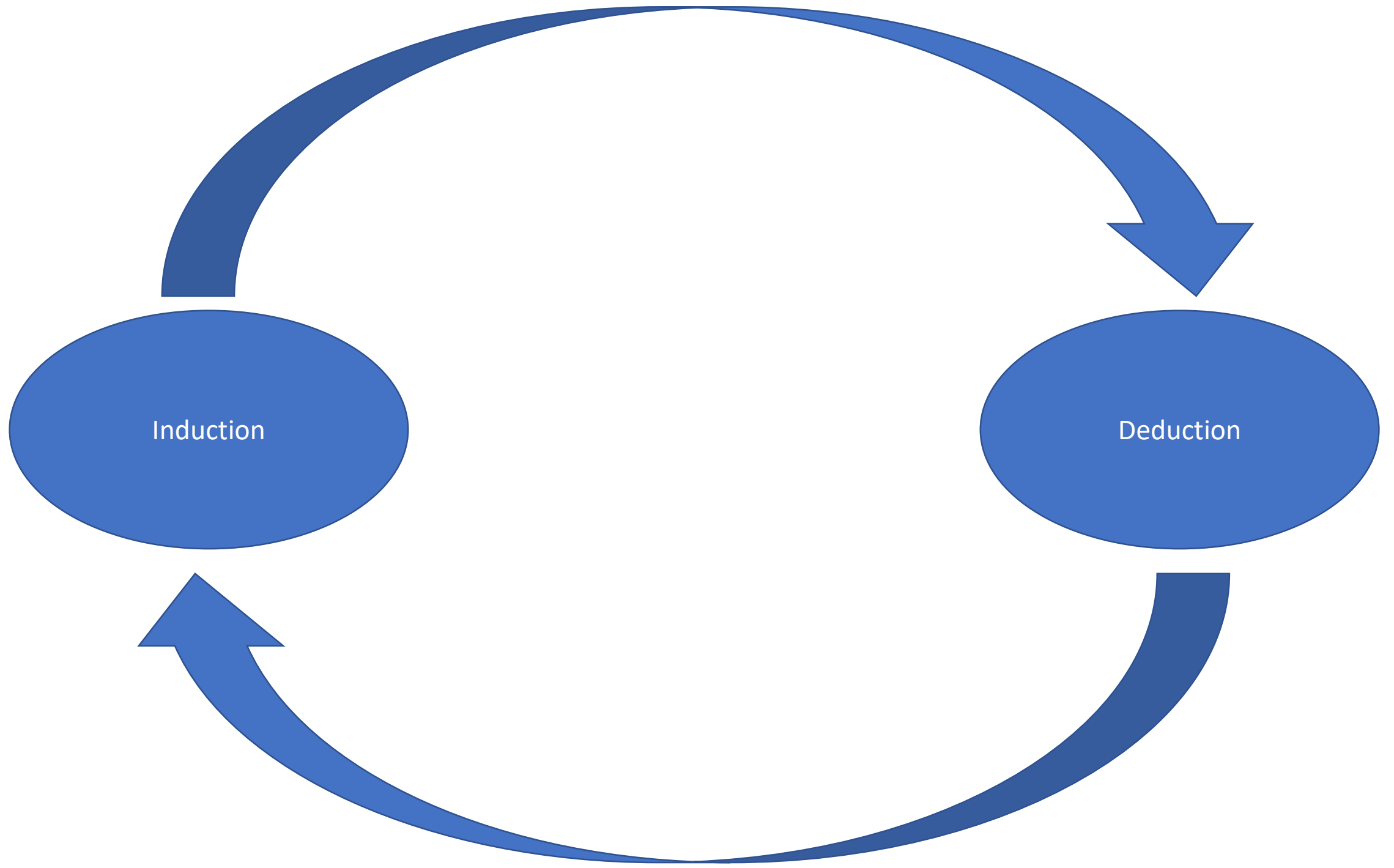


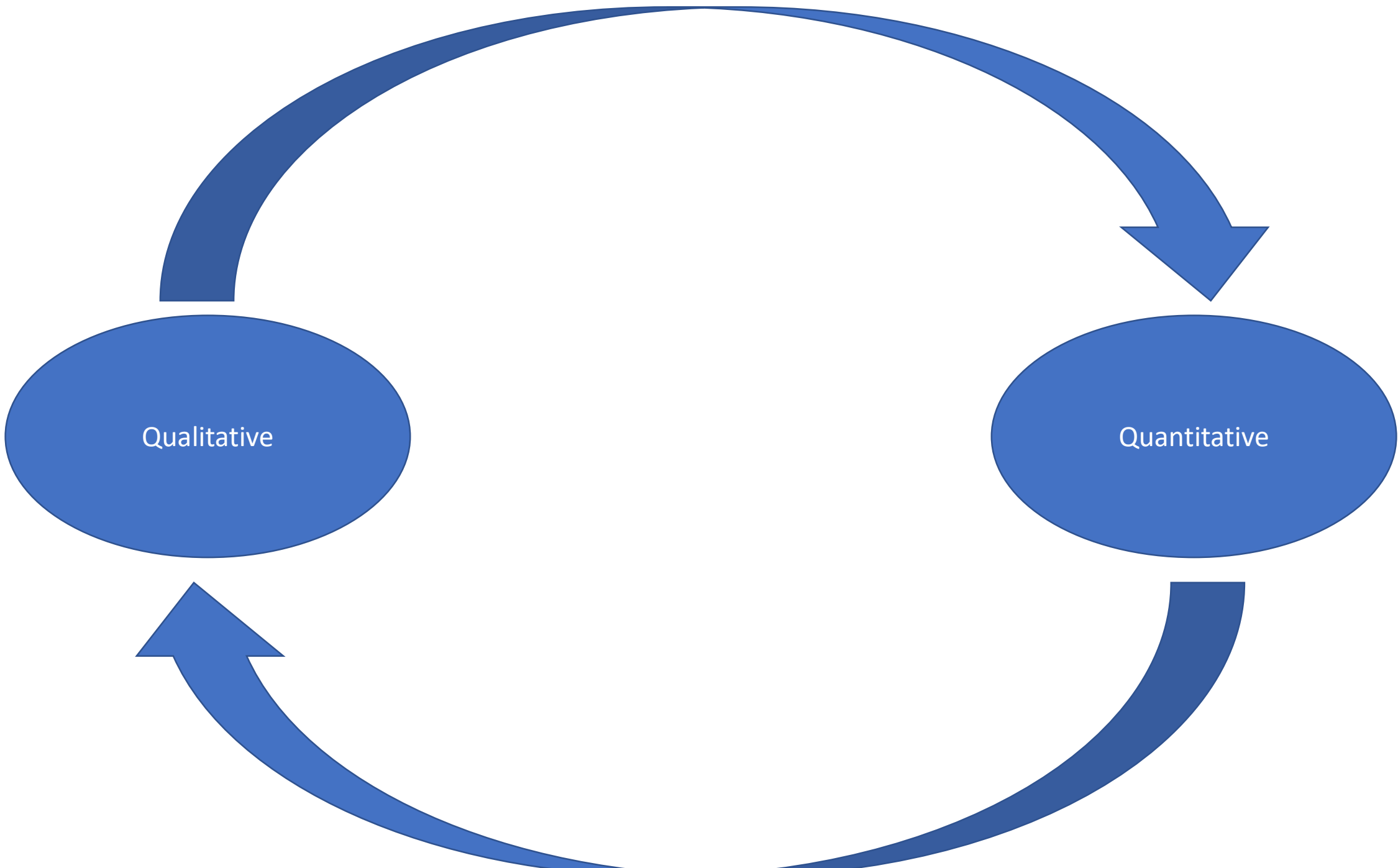
Causal trees

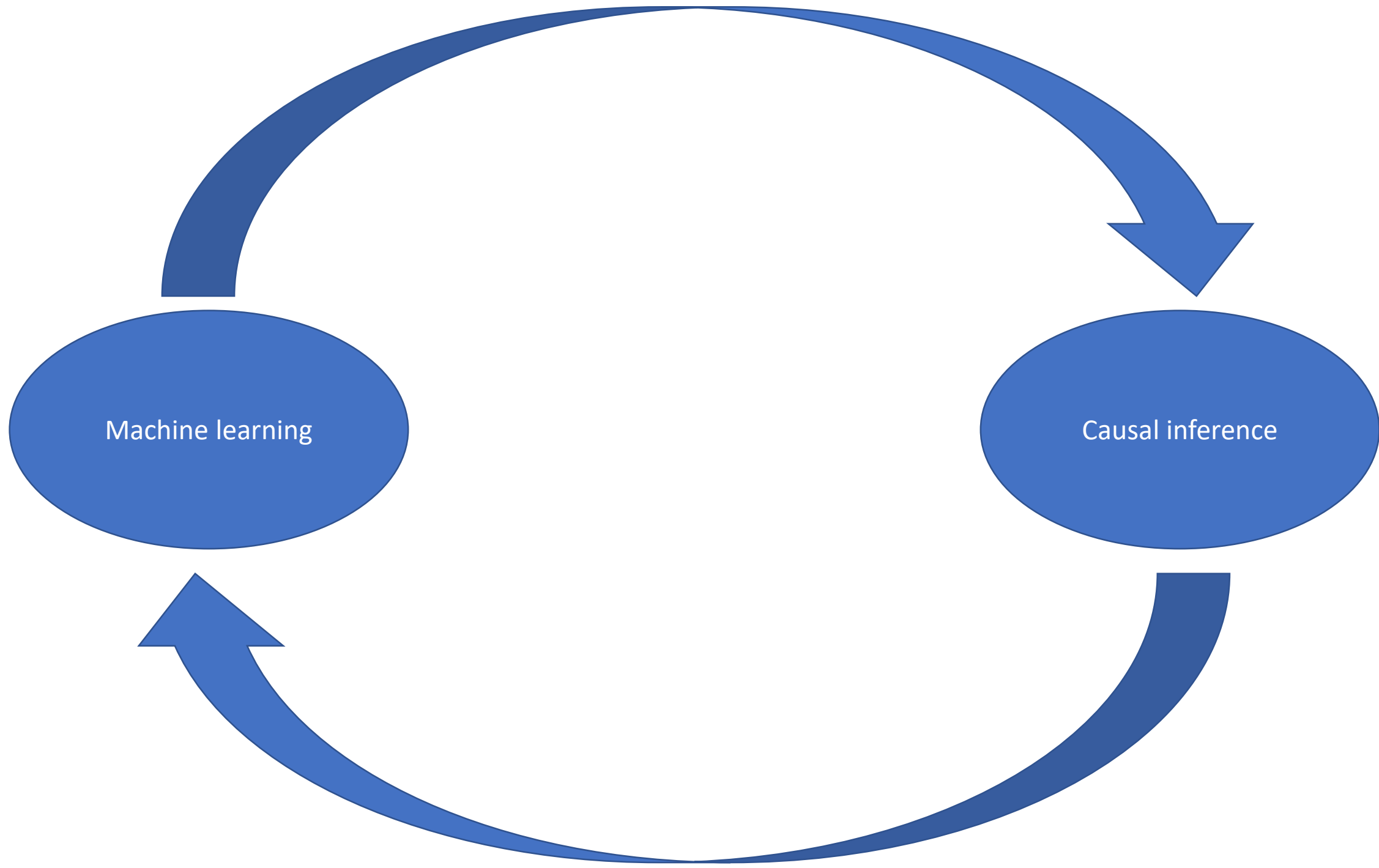
- Look for splits that maximize the difference in means between treated and untreated group
- Provides a data-driven way to split up covariate space for CATE
- Assumes unconfoundedness, but can also weight observations by propensity score to relax assumption of unconfoundedness to selection on observables
- Importantly, introduces idea of “honest estimation”

Casual trees (honest estimation)


- Use one set of data to find the appropriate splits
- Then test these splits on held-out data
- Given this, testing on the held-out data is as if you had a deductive theory of what splits would have different CATEs (can derive SEs and make valid inferences)
- This is an important technique that will be used throughout many different methods you will see
- Importantly, this also allows you to predict ATE (and sometimes can do this more powerfully than other methods)







Causal forests

 Cornell University

arXiv.org > stat > arXiv:1510.04342

Search or Article Number
(Help | Advanced Search)

Statistics > Methodology

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager, Susan Athey

(Submitted on 14 Oct 2015 (v1), last revised 10 Jul 2017 (this version, v4))

Many scientific and engineering challenges -- ranging from personalized medicine to customized marketing recommendations -- require an understanding of treatment effect heterogeneity. In this paper, we develop a non-parametric causal forest for estimating heterogeneous treatment effects that extends Breiman's widely used random forest algorithm. In the potential outcomes framework with unconfoundedness, we show that causal forests are pointwise consistent for the true treatment effect, and have an asymptotically Gaussian and centered sampling distribution. We also discuss a practical method for constructing asymptotic confidence intervals for the true treatment effect that are centered at the causal forest estimates. Our theoretical results rely on a generic Gaussian theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for provably valid statistical inference. In experiments, we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially in the presence of irrelevant covariates.

Comments: To appear in the Journal of the American Statistical Association. Part of the results developed in this paper were made available as an earlier technical report "Asymptotic Theory for Random Forests", available at ([arXiv:1405.0352](https://arxiv.org/abs/1405.0352))

Subjects: **Methodology (stat.ME)**; Statistics Theory (math.ST); Machine Learning (stat.ML)

Cite as: **arXiv:1510.04342 [stat.ME]**
(or **arXiv:1510.04342v4 [stat.ME]** for this version)

Submission history

From: Stefan Wager [[view email](#)]

[v1] Wed, 14 Oct 2015 22:54:59 UTC (346 KB)

[v2] Sat, 21 Nov 2015 00:38:23 UTC (348 KB)


[v3] Sat, 19 Nov 2016 04:08:22 UTC (1,520 KB)

[v4] Mon, 10 Jul 2017 01:15:47 UTC (971 KB)

Topic models

- Assume that a document is made up of a set of topics, and those topics are made up of a “basket of words”
- Assume that the generation of a document is done by selecting how much of each topic will make up a document, and then randomly selecting words within that topic
- With LDA, we can reverse-engineer these topics given a set of documents and know how much of each document is made up of each topic
- NOTE: there are other, better (in my opinion) topic models out there that are very recent, but this is by far the most popular
- Importantly, in this setup we have to decide *a priori* how many topics are present in all the documents
- There’s no perfect way of doing this... What is honestly the accepted way is to try a bunch of values until you find a set of theoretically coherent topics
- But.... Aren’t you at risk of over-fitting then?
- This problem comes in lots of applications where there is some latent dimension of text we are trying to pick up as a variable of interest

Causal inference and text

 Cornell University

arXiv.org > stat > arXiv:1802.02163

Search or Ask Question
(Help | Advanced Search)

Statistics > Machine Learning

How to Make Causal Inferences Using Texts

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, Brandon M. Stewart
(Submitted on 6 Feb 2018)

New text as data techniques offer a great promise: the ability to inductively discover measures that are useful for testing social science theories of interest from large collections of text. We introduce a conceptual framework for making causal inferences with discovered measures as a treatment or outcome. Our framework enables researchers to discover high-dimensional textual interventions and estimate the ways that observed treatments affect text-based outcomes. We argue that nearly all text-based causal inferences depend upon a latent representation of the text and we provide a framework to learn the latent representation. But estimating this latent representation, we show, creates new risks: we may introduce an identification problem or overfit. To address these risks we describe a split-sample framework and apply it to estimate causal effects from an experiment on immigration attitudes and a study on bureaucratic response. Our work provides a rigorous foundation for text-based causal inferences.

Comments: 47 pages
Subjects: **Machine Learning (stat.ML)**; Computation and Language (cs.CL); Methodology (stat.ME)
Cite as: **arXiv:1802.02163 [stat.ML]**
(or **arXiv:1802.02163v1 [stat.ML]** for this version)

Submission history

From: Brandon Stewart [[view email](#)]
[v1] Tue, 6 Feb 2018 19:00:12 UTC (307 KB)

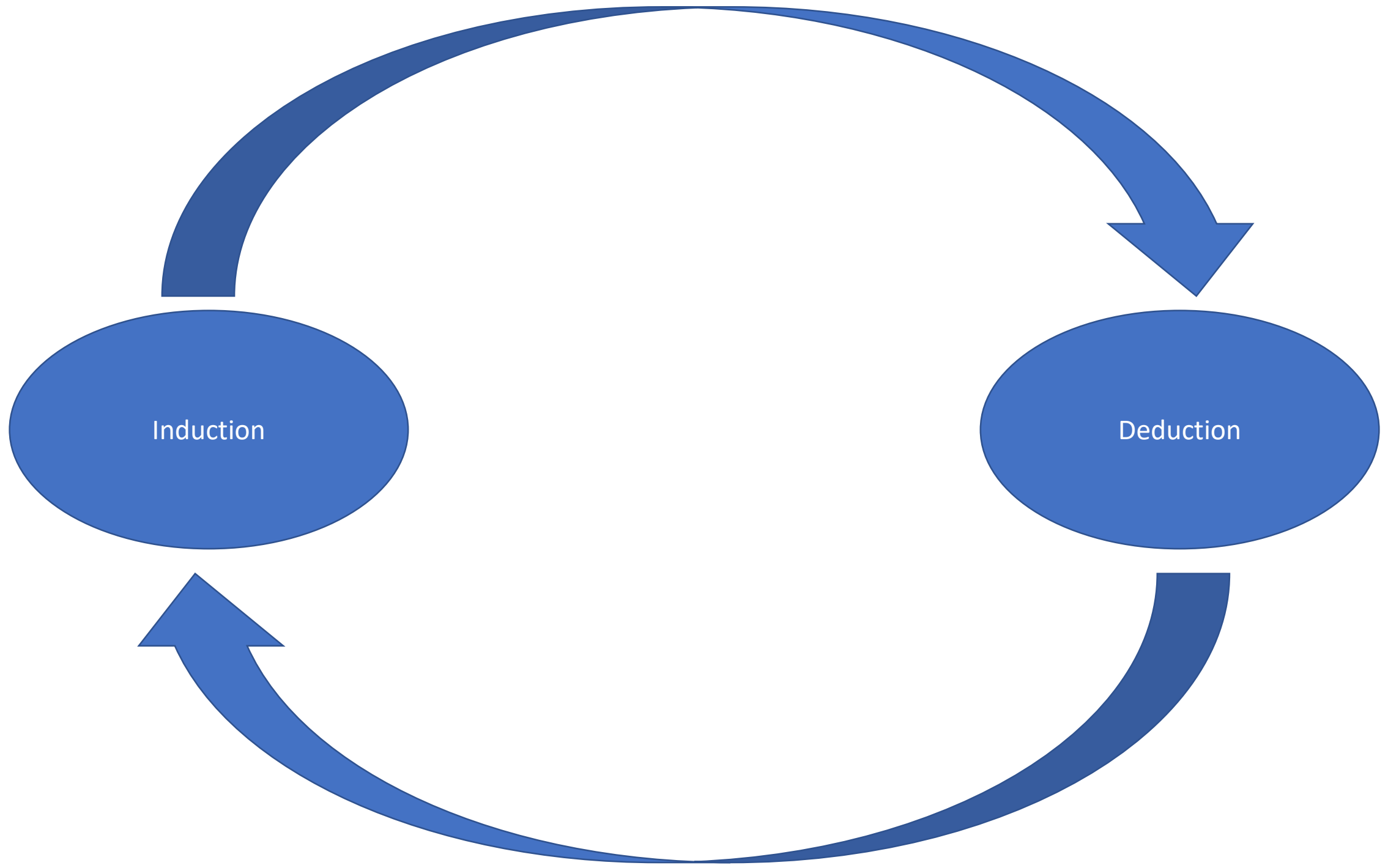
[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))

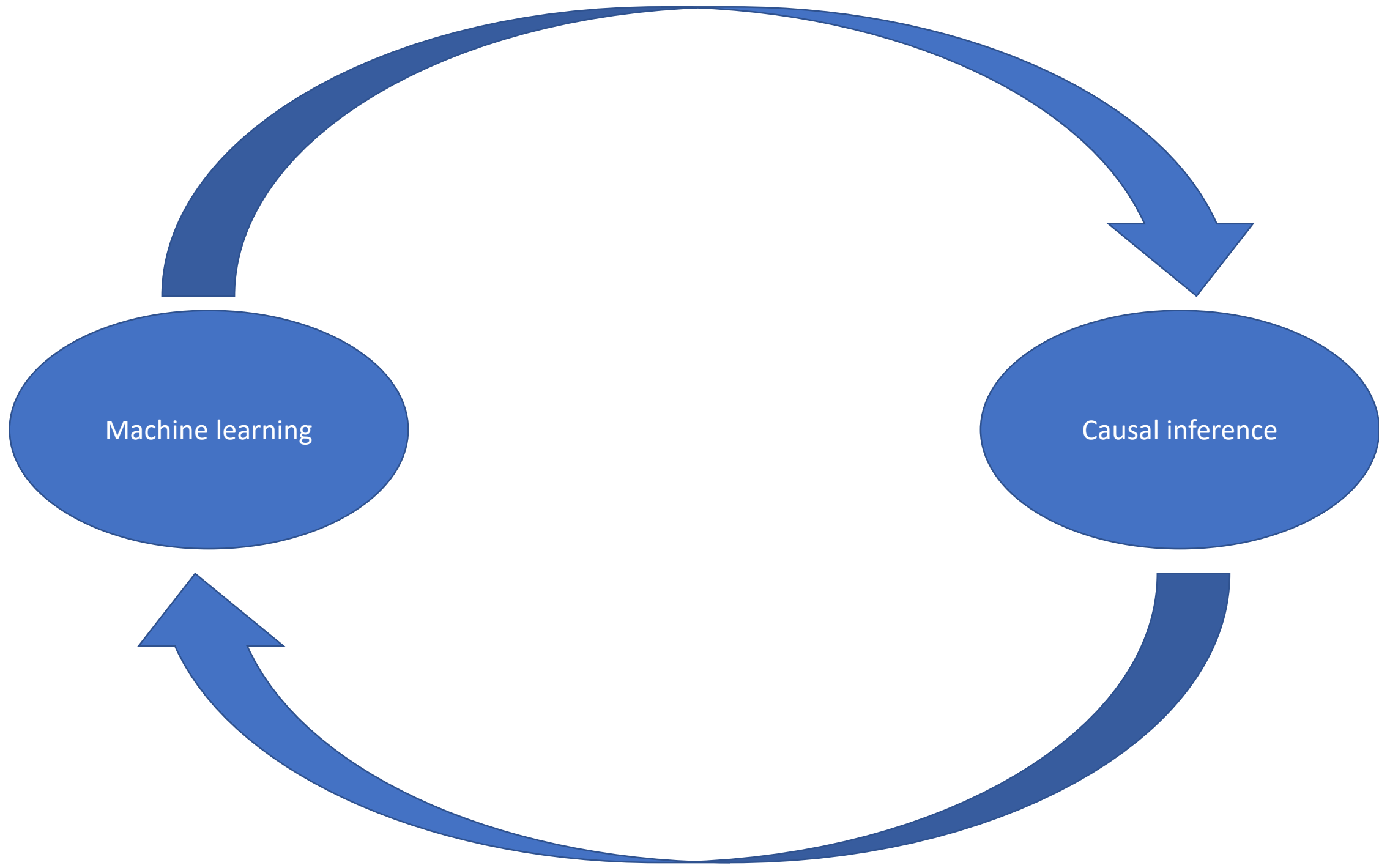
Causal inference and text

- On one set of data, train your topic model until you have theoretically insightful topics
- On the new set of data, simply count the occurrences of each topic in a document and do a t-test, negative binomial regression, or fractional logit (depending on distributions of the outcome and lengths of the document)
- **Text as DV:** how does being exposed to one experimental condition as opposed to another affect the text respondents write to an open-ended question?
- **Text as IV:** at this political rally, folks were exposed to this speech (with certain topics) and at a different rally they were exposed to this speech (with different topics). How did that affect their voting behavior?
- Can also do CATE estimation with text

Causal inference and text (cont.)

- If you only care about doing observational studies, DON'T WORRY ABOUT THIS
- If you already have a pre-defined text variable (e.g. count of words in this pre-defined dictionary), DON'T WORRY ABOUT THIS
- Only need to split when you both (a) want to argue you are measuring a causal effect AND (b) need to derive the latent characteristic from the text
- Just for your information, a lot of “computational social science” now is doing a topic model (or something similar) on a corpus of text and seeing how topic distribution varies with some variable





Fragile Families Challenge



[HOME](#) [OUR BLOG](#) [CONTACT US](#)

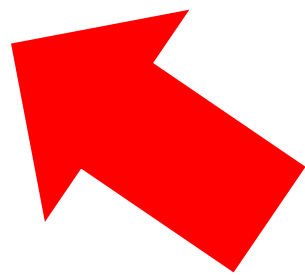


What would happen if hundreds of social scientists and data scientists worked together on a scientific challenge to improve the lives of disadvantaged children in the United States?

Read a quick overview from the Princeton University Office of Communications. More questions? Read blog posts [here](#)

$$\hat{y}_i = \hat{\beta} X_i$$

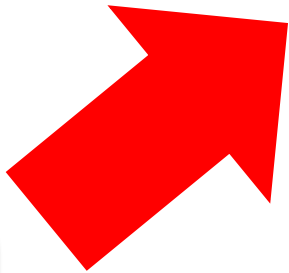
$$\hat{y}_i = \hat{\beta} X_i$$



Prediction

$$\hat{y}_i = \hat{\beta} X_i$$

Explanation



Prediction * Explanation

- **Argument:** if we can't predict an outcome, it's either because we don't have the correct β or because we don't have all the relevant Xs
- Machine learning has gotten very good at estimating y given x
- Missing Xs are “dark matter” in our observations (we know it's there but we can't often observe it)
- It can be very hard to discover Xs with quantitative or computational analysis, but that's the bread-and-butter of qualitative analysis!

Fragile Families Challenge

- Phase I: Get a ton of money
- Phase II: Collect longitudinal data on thousands of American families for 15 years
- Phase III: Offer prizes to individuals who can predict 6 key outcomes (gpa, grit, materialHardship, eviction, layoff, jobTraining) the best
- Phase IV: Combine all the individual submissions using something called an “ensemble method” to build a new model which should out-predict even the best individual model
- Phase V: Identify cases which were not well predicted, go back and interview these individuals to reveal what vital information we’re missing (try to figure out what makes up the “dark matter”)

$$\hat{y}_i = \hat{\beta} X_i$$