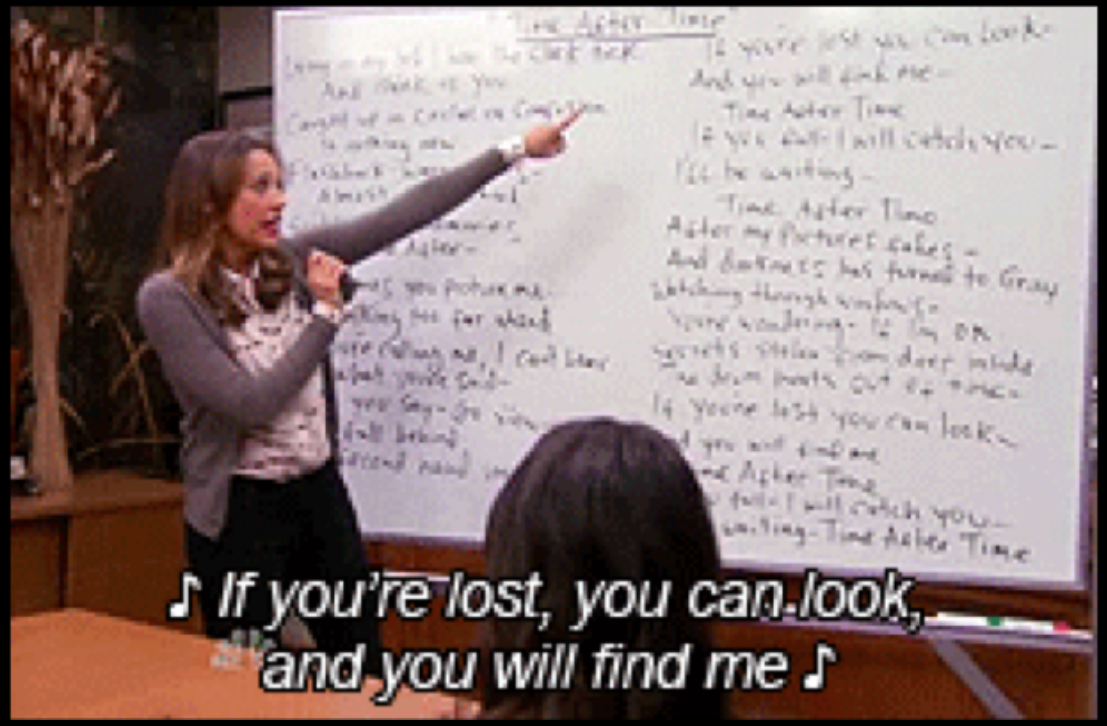
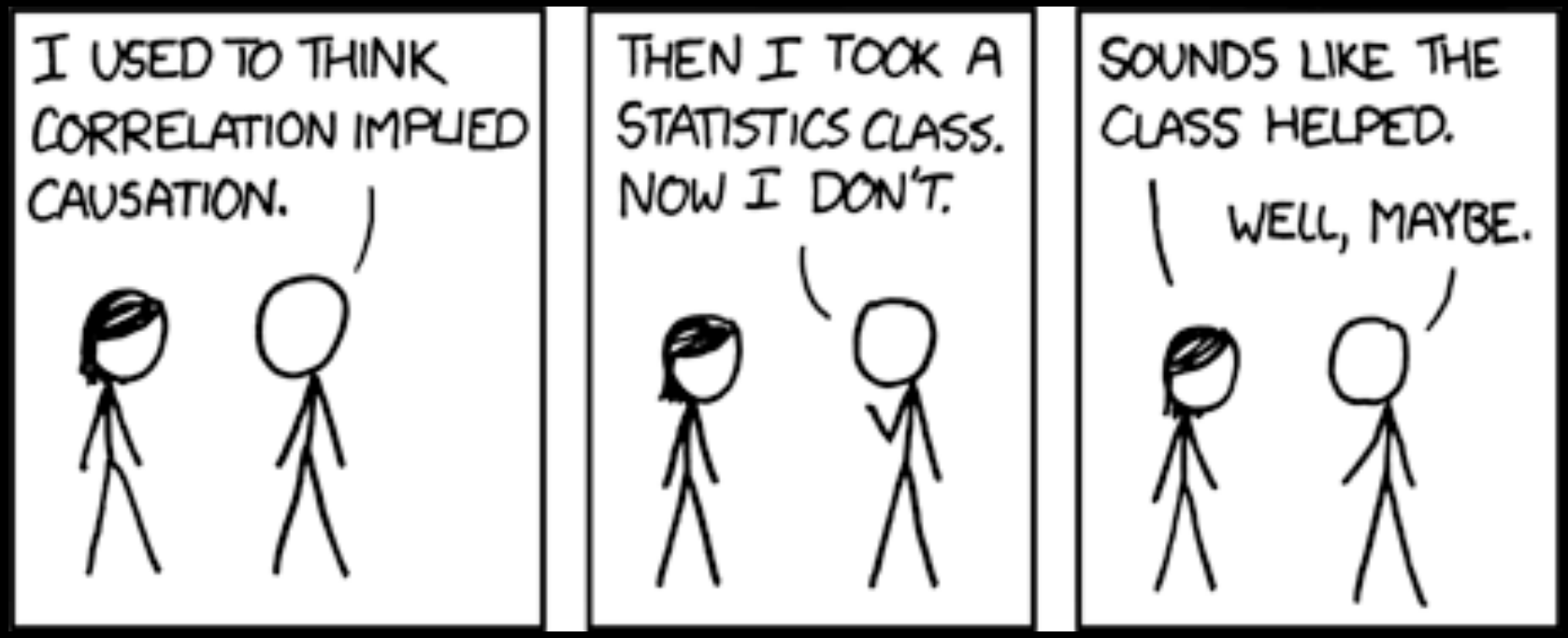
Making Causal Claims as a Data Scientist: Tips and Tricks Using R

Lucy D'Agostino McGowan

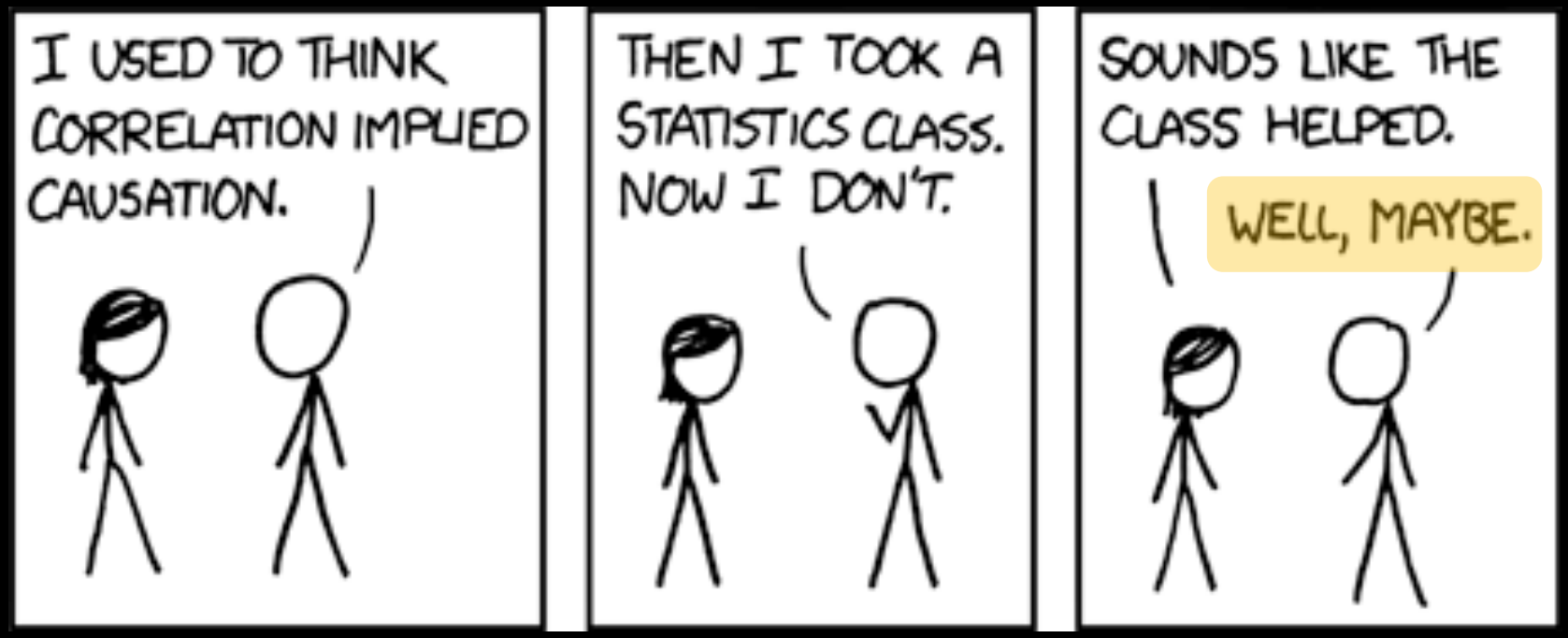
Data Day Texas 2018

http://bit.ly/LucyStatsDDTX18



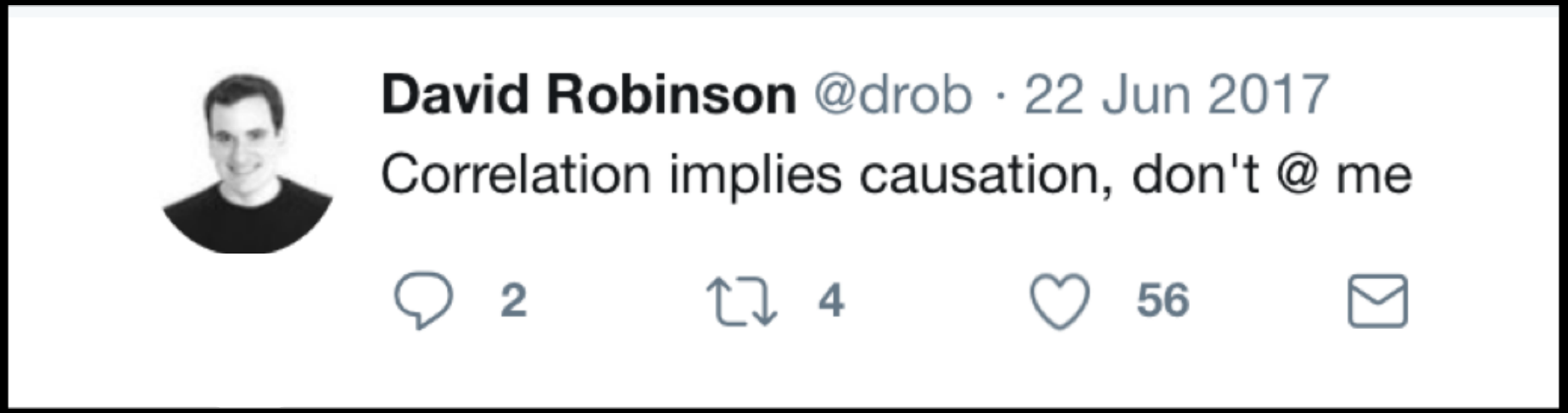


https://xkcd.com/552/



https://xkcd.com/552/


Lucy @LucyStats · 11 Oct 2017
Kaplan claims: As statisticians we are too focused on "abstinence only" method, we need to teach **safe** causality #SSI2017 #ASASymposium2017
GIF
Lucy @LucyStats
Daniel Kaplan makes an exciting claim: correlation *is* causation! (cc: @theeffortreport) #SSI2017 #ASASymposium2017
Show this thread
13
45


David Robinson @drob · 22 Jun 2017
Correlation implies causation, don't @ me
2
4
56

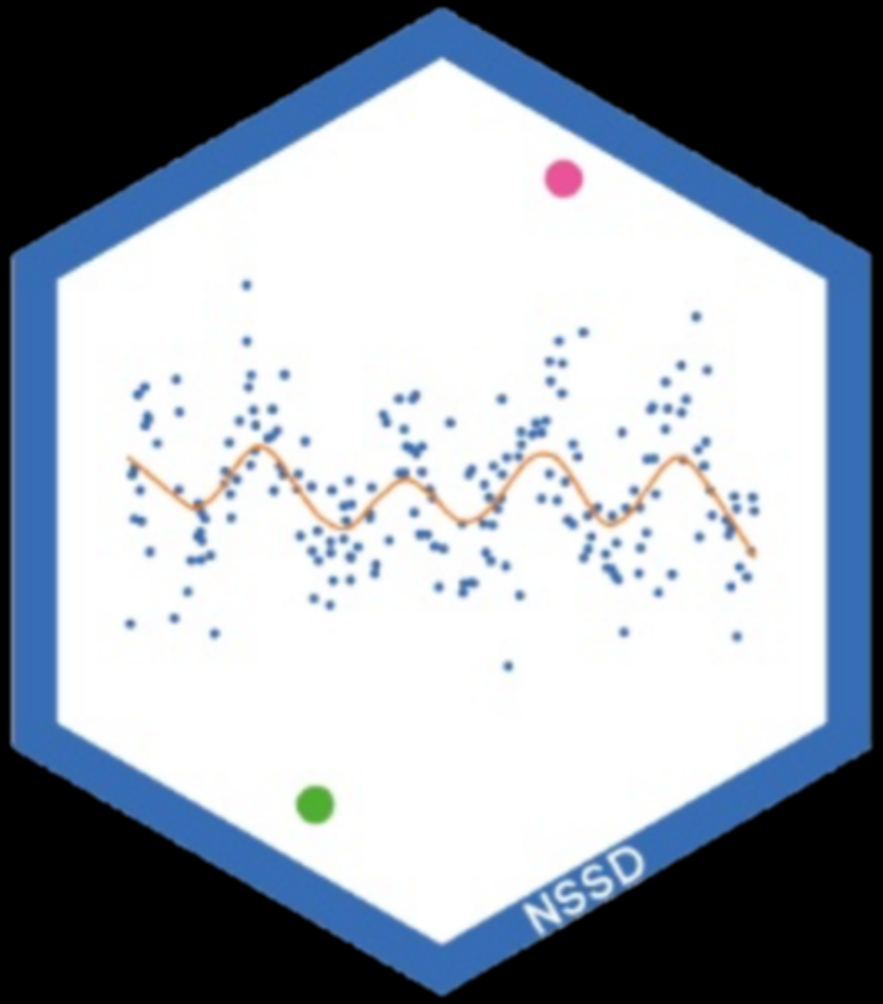
✌ parts

1. Discuss some ways to strengthen a causal argument: **Hill's criteria**

2. Discuss a specific causal inference method: **propensity scores + sensitivity analyses**

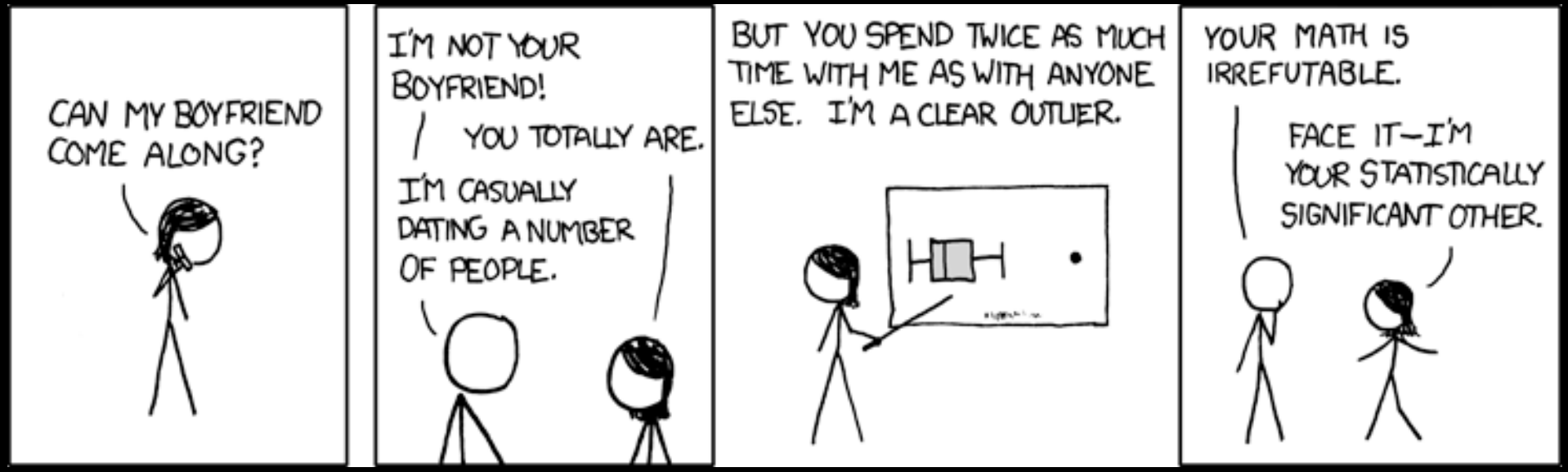
Hill's Criteria

https://commons.wikimedia.org/wiki/File:Austin_Bradford_Hill.jpg



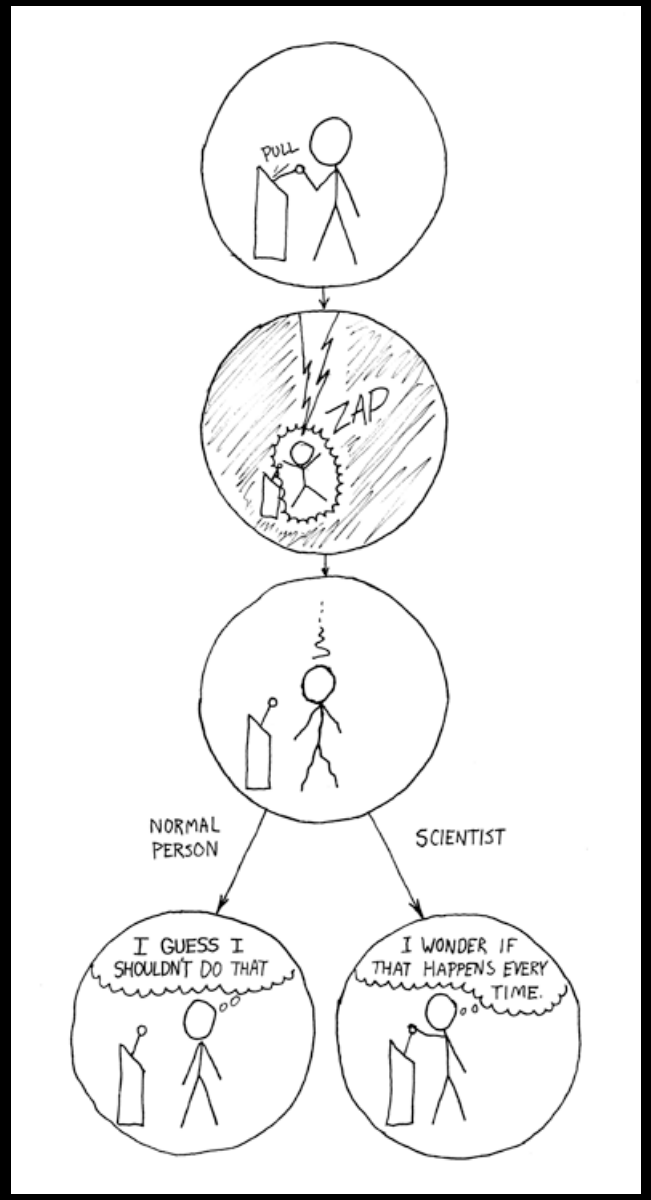
http://nssdeviations.com

http://livefreeordichotomize.com/2016/12/15/hill-for-the-data-scientist-an-xkcd-story/



https://xkcd.com/539/

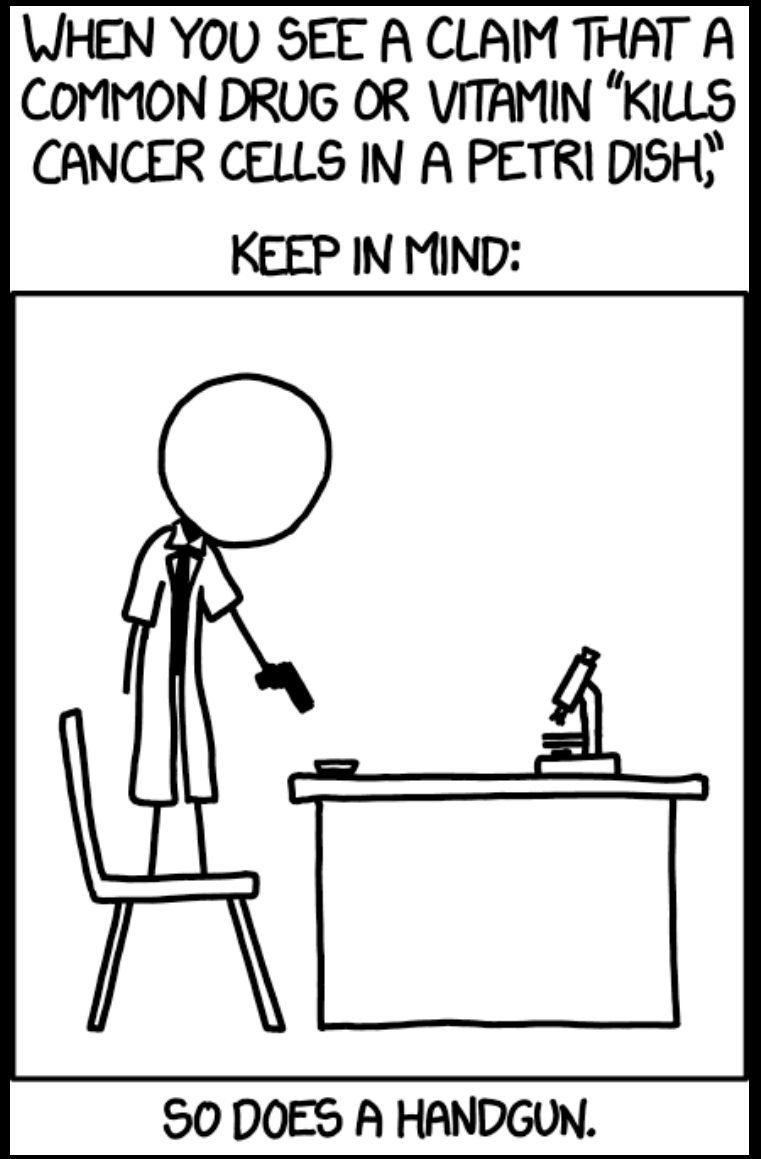
Strength



https://xkcd.com/242/

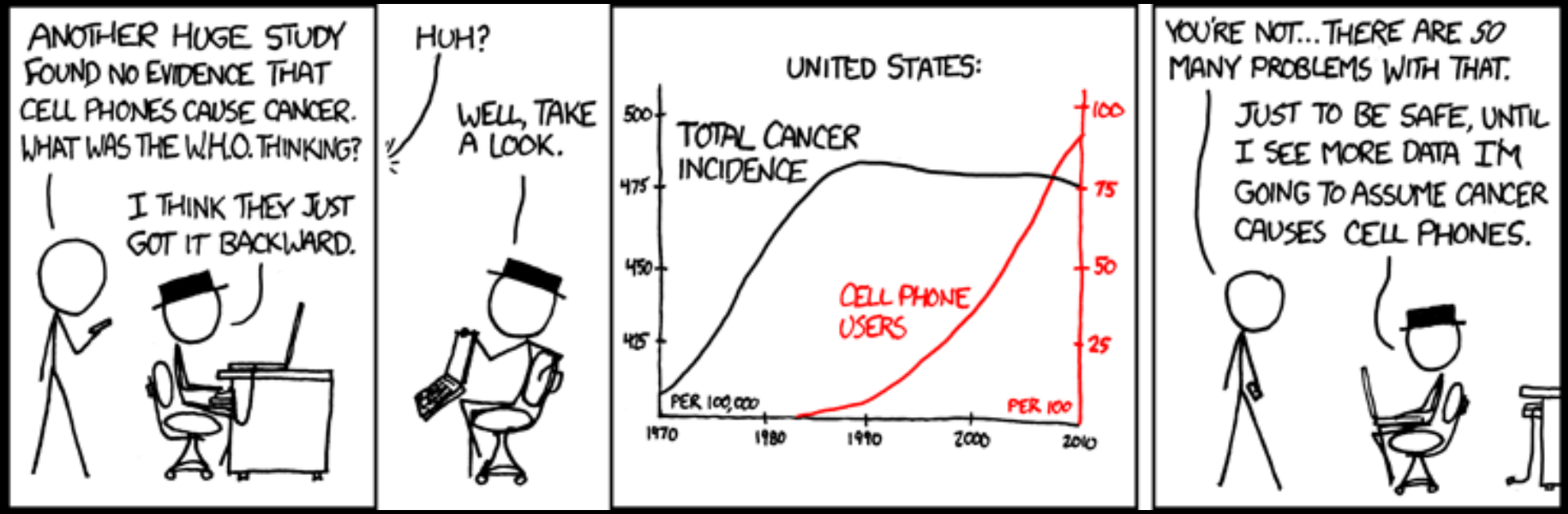
Consistency

Specificity



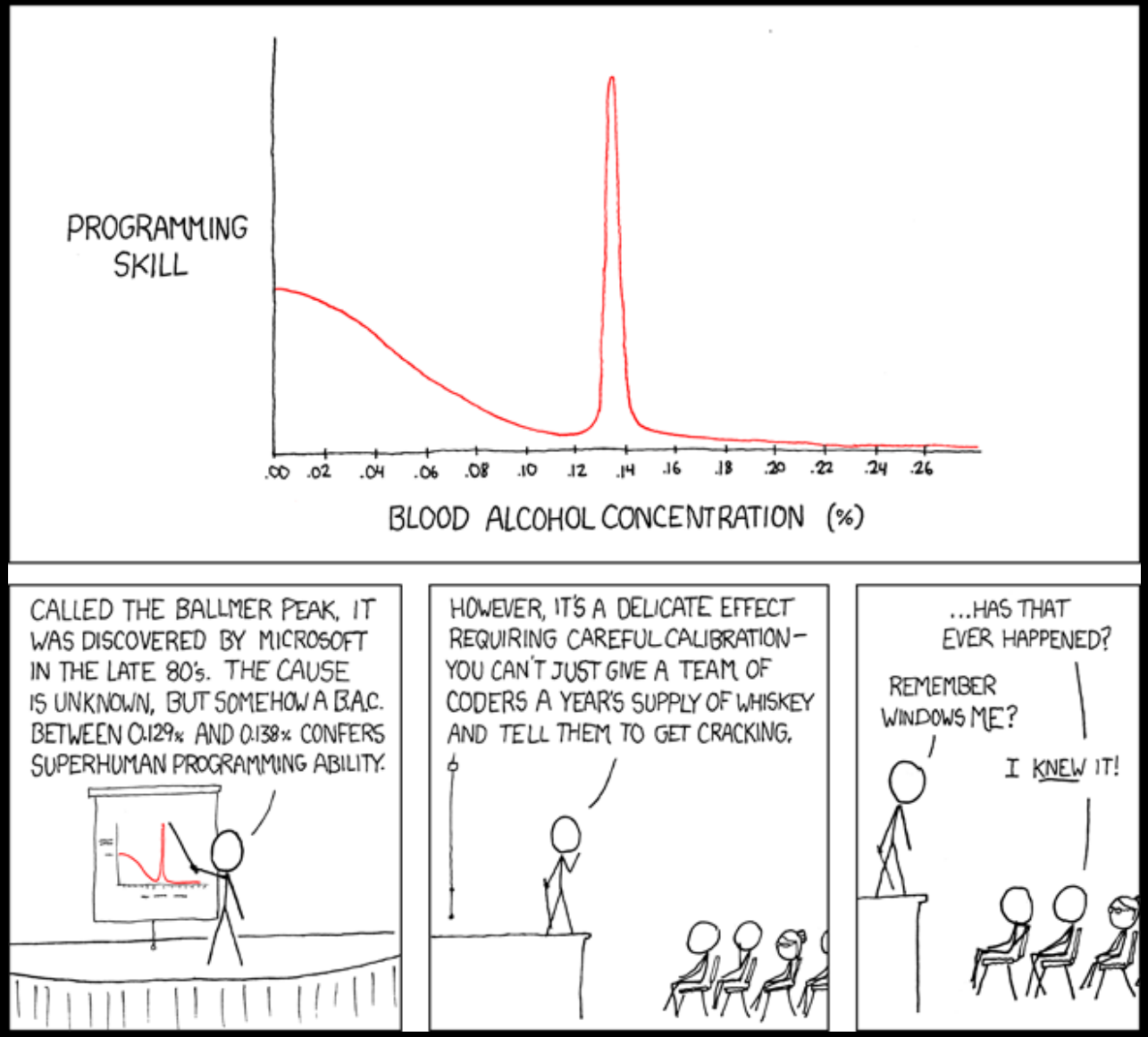
https://xkcd.com/1217/

Temporality



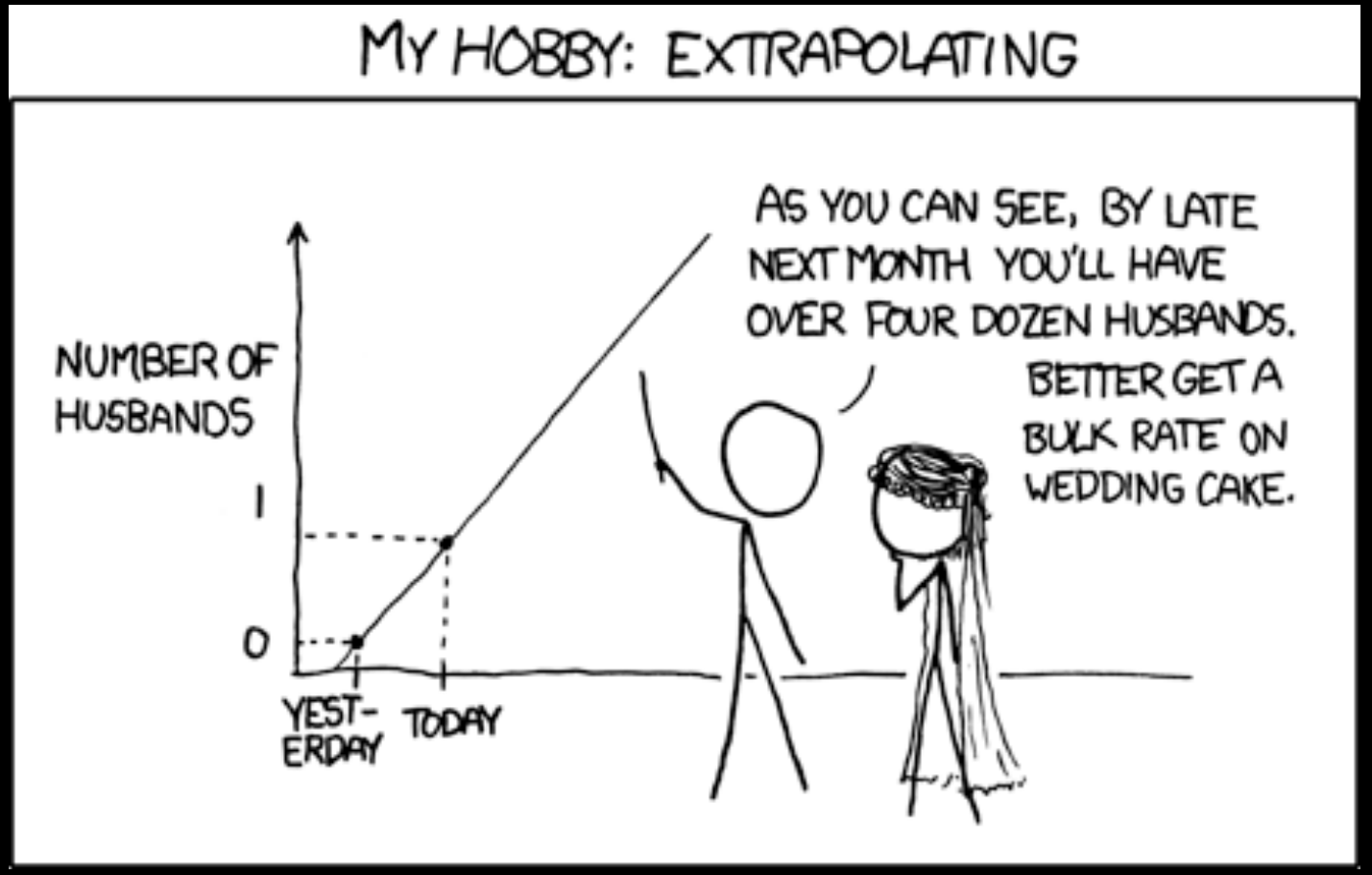
https://xkcd.com/925/

Biological gradient

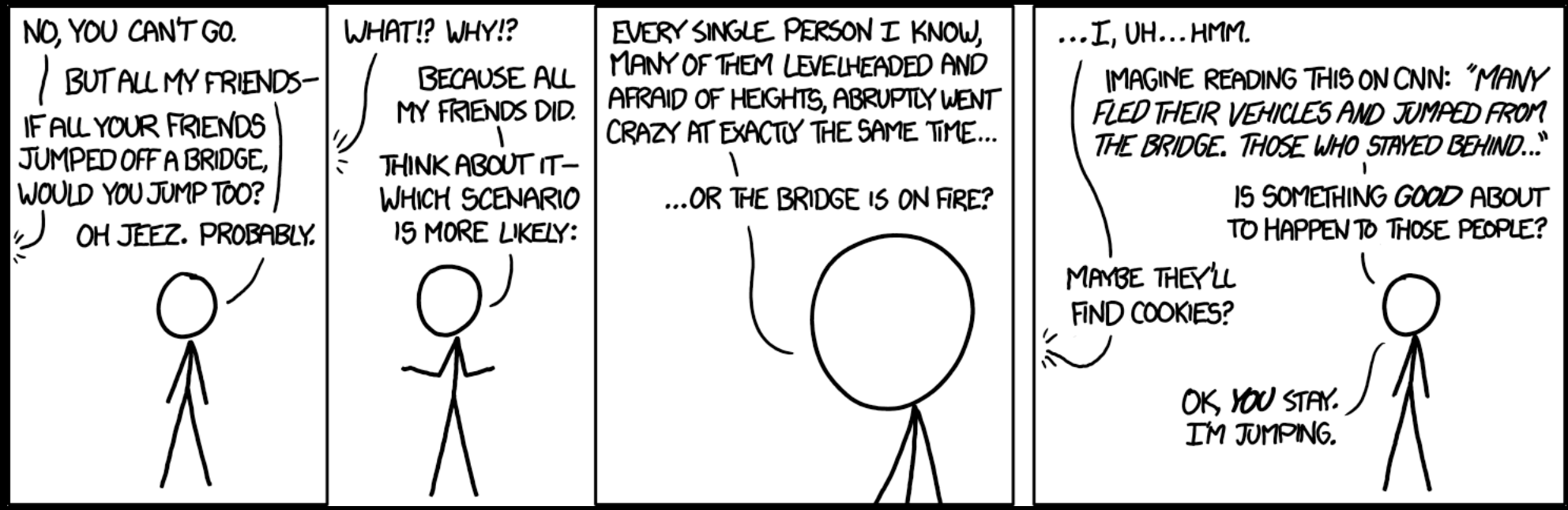


https://xkcd.com/323/

Plausibility



https://xkcd.com/605/

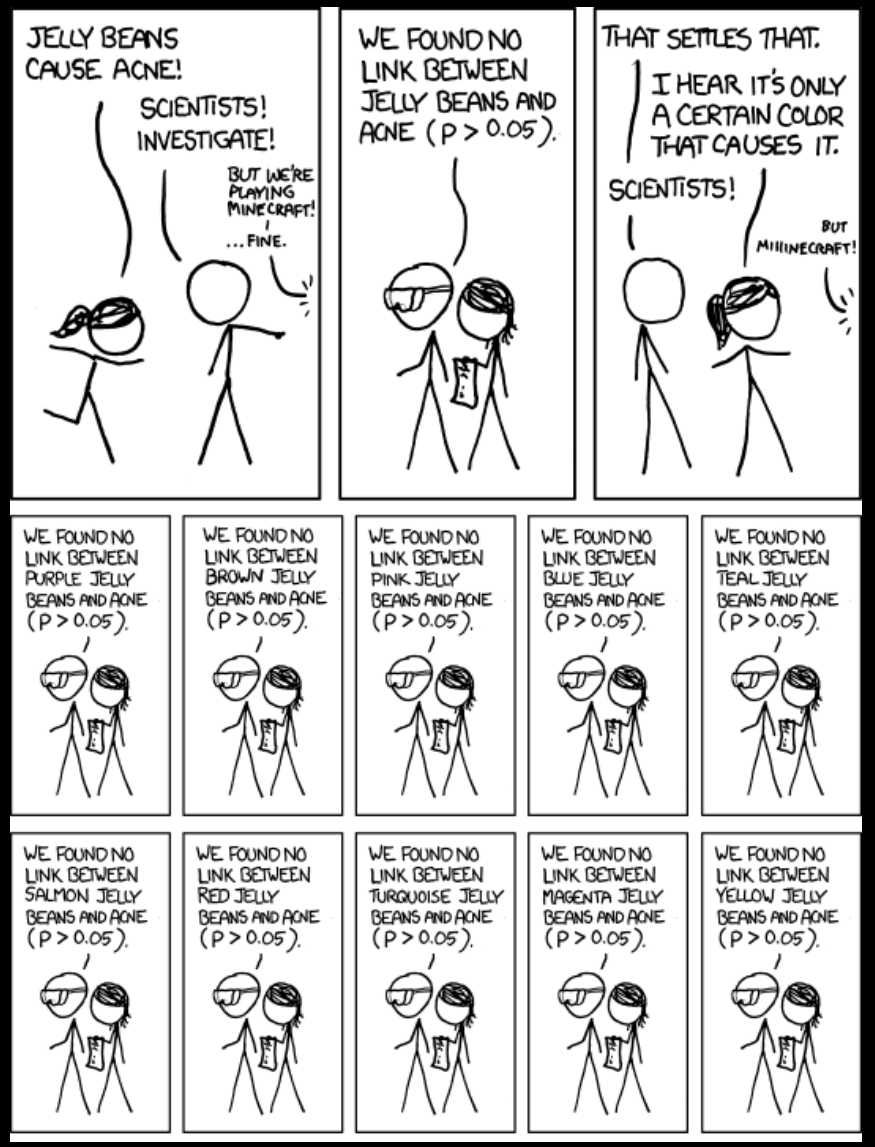


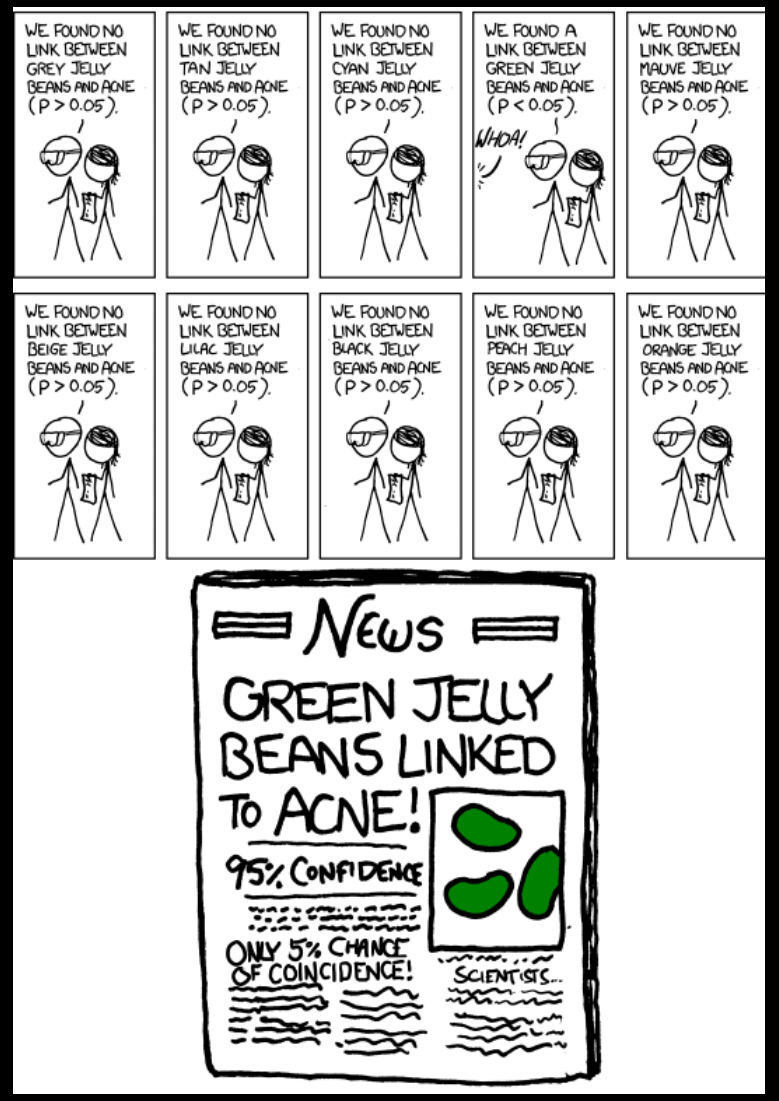
https://xkcd.com/1170/

Coherence

DATA DAY TEXAS 2018

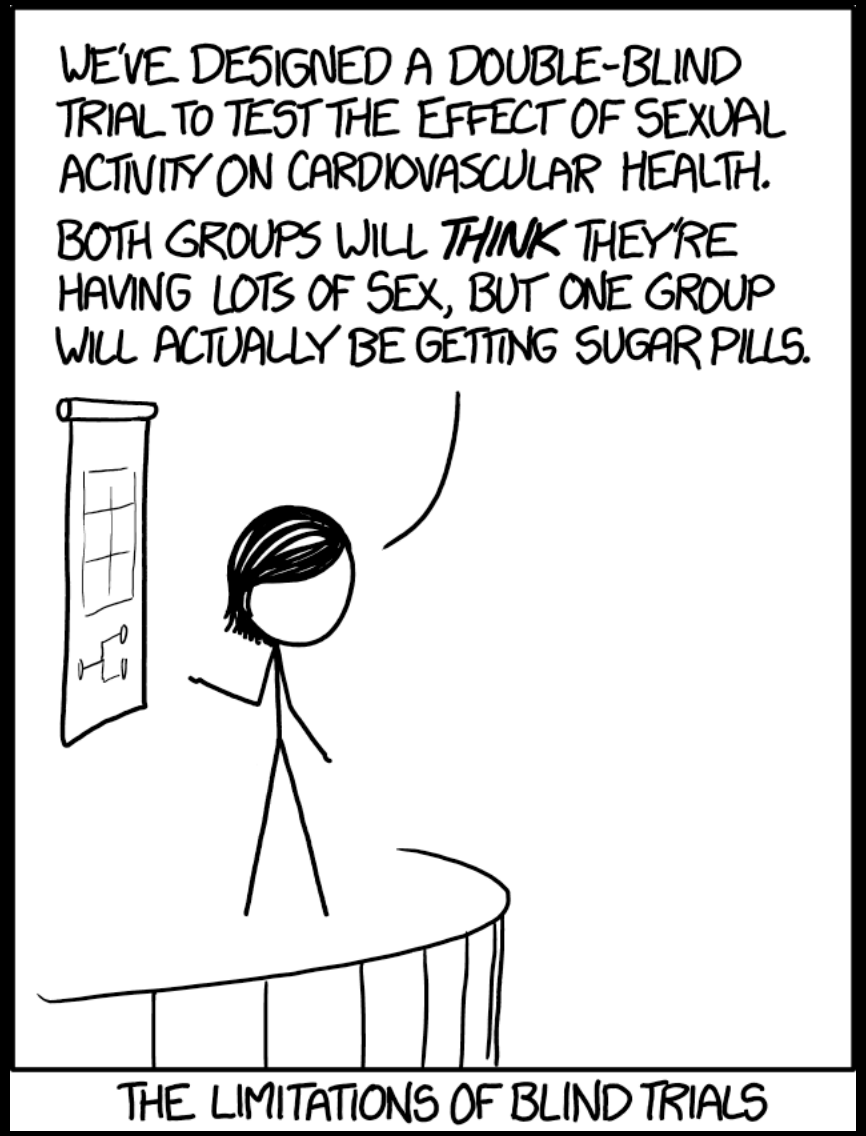
Analogy





https://xkcd.com/882/

Experiment

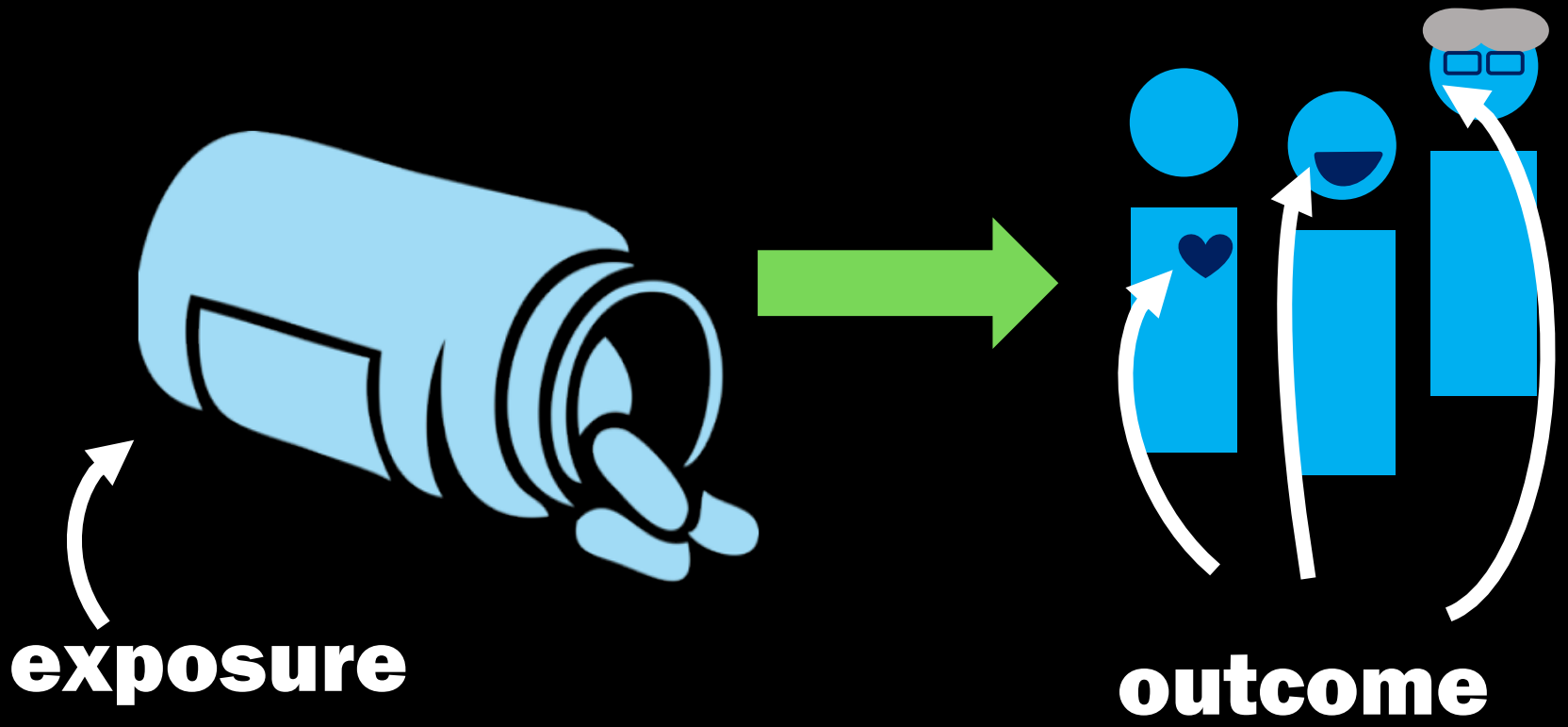


https://xkcd.com/1462/

What if you can't do a controlled experiment?

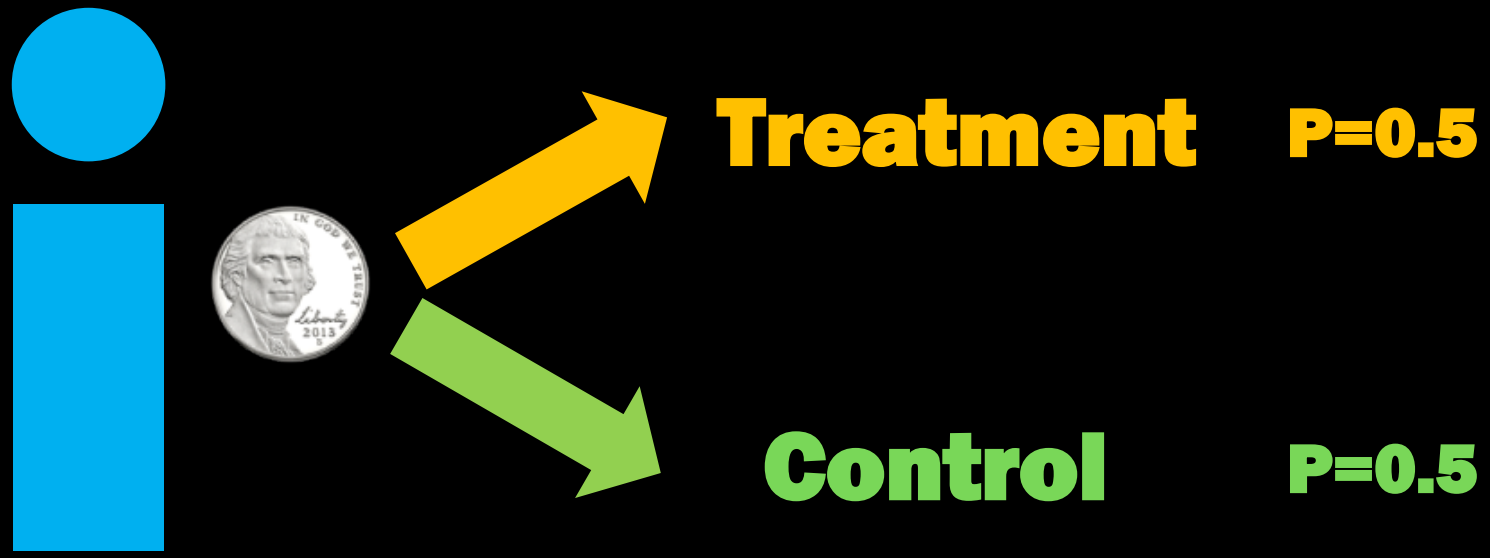
Observational Studies

Goal: Answer a research question



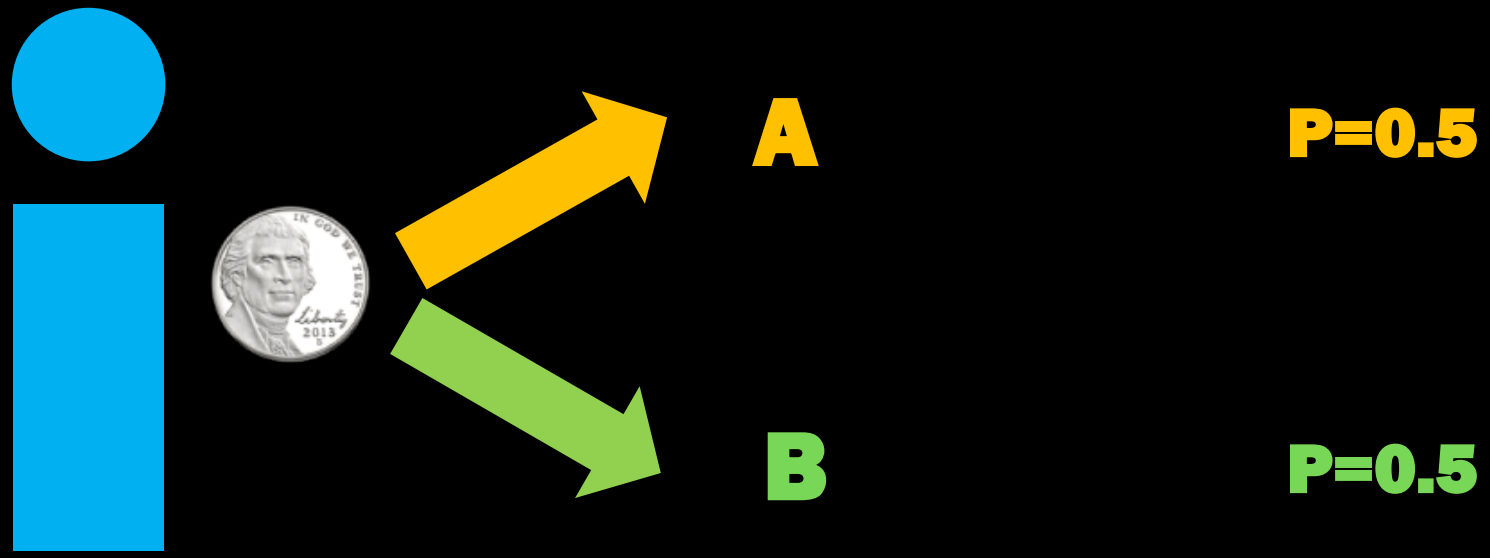
~~Observational Studies~~

"Gold Standard" – Randomized Controlled Trials



~~Observational Studies~~

"Gold Standard" – Randomized Controlled Trials

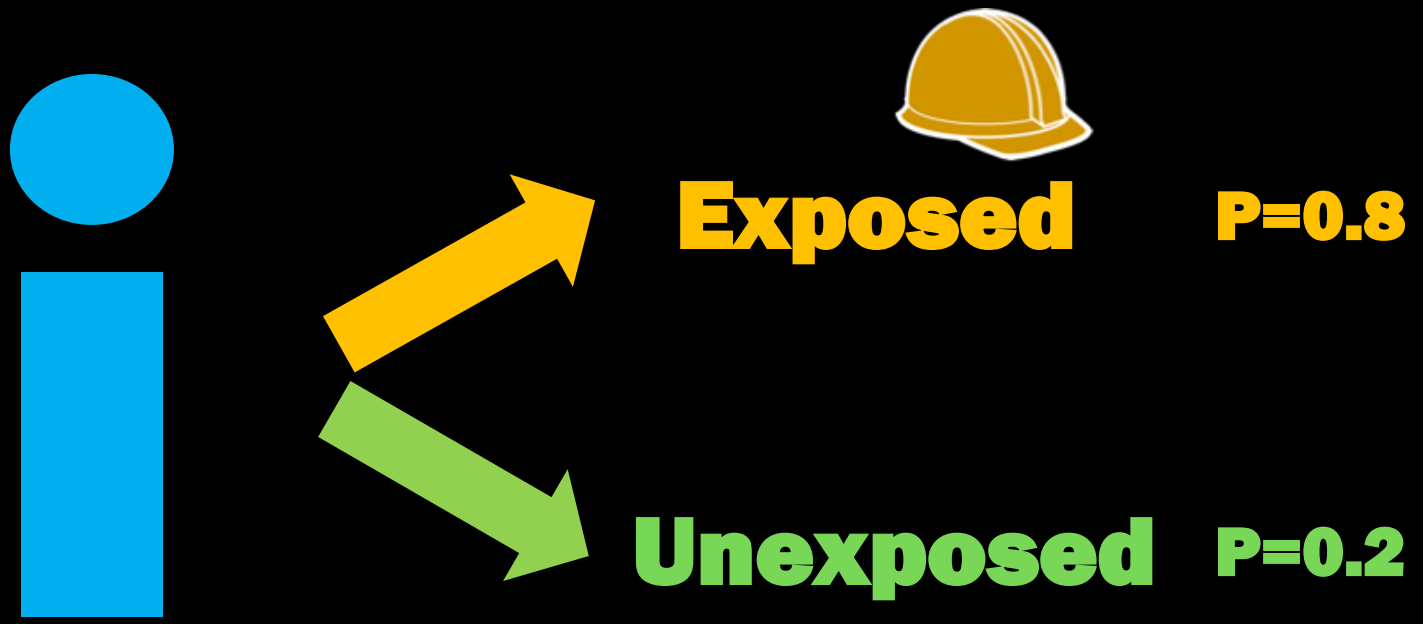


Observational Studies

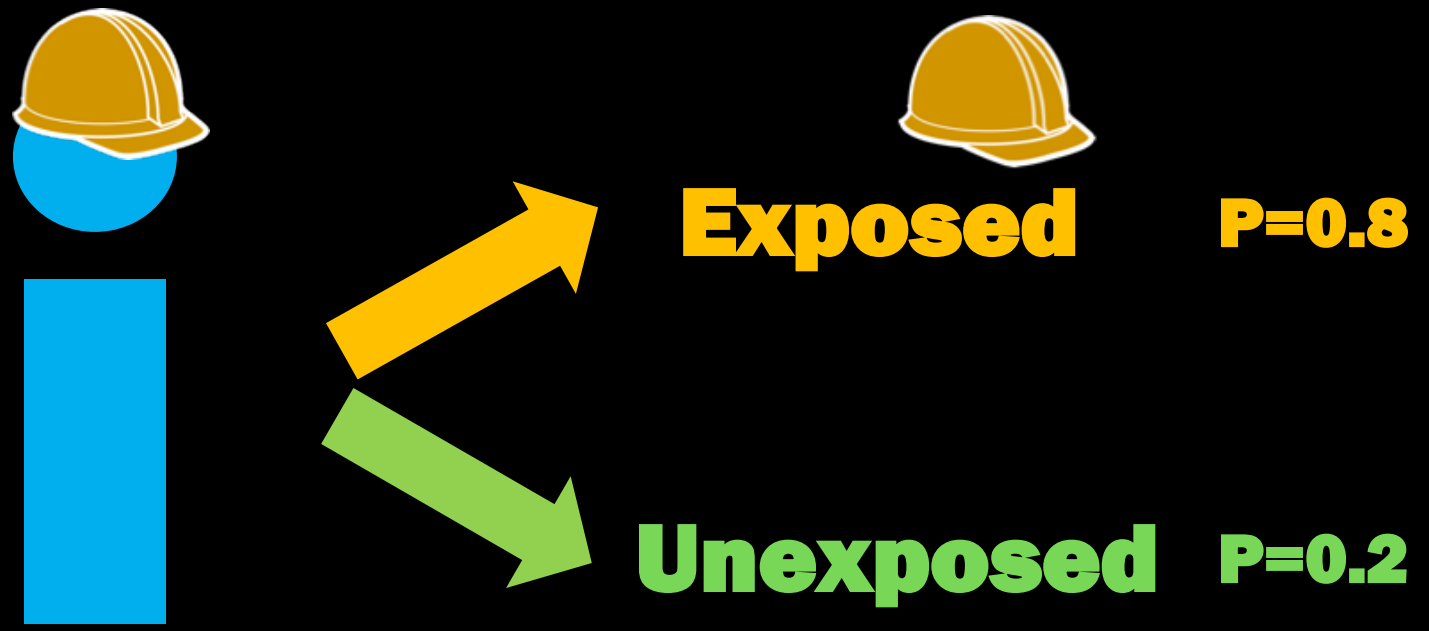
Treatment P=0.9

Control P=0.1

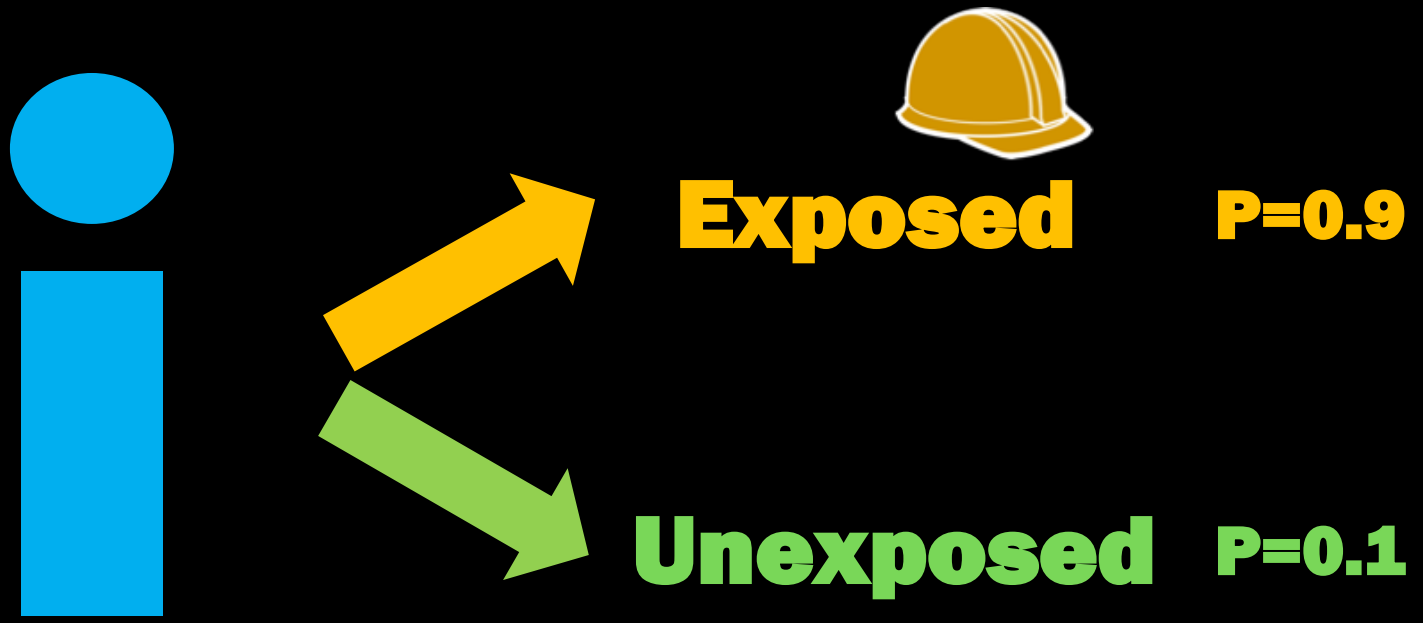
Observational Studies



Observational Studies



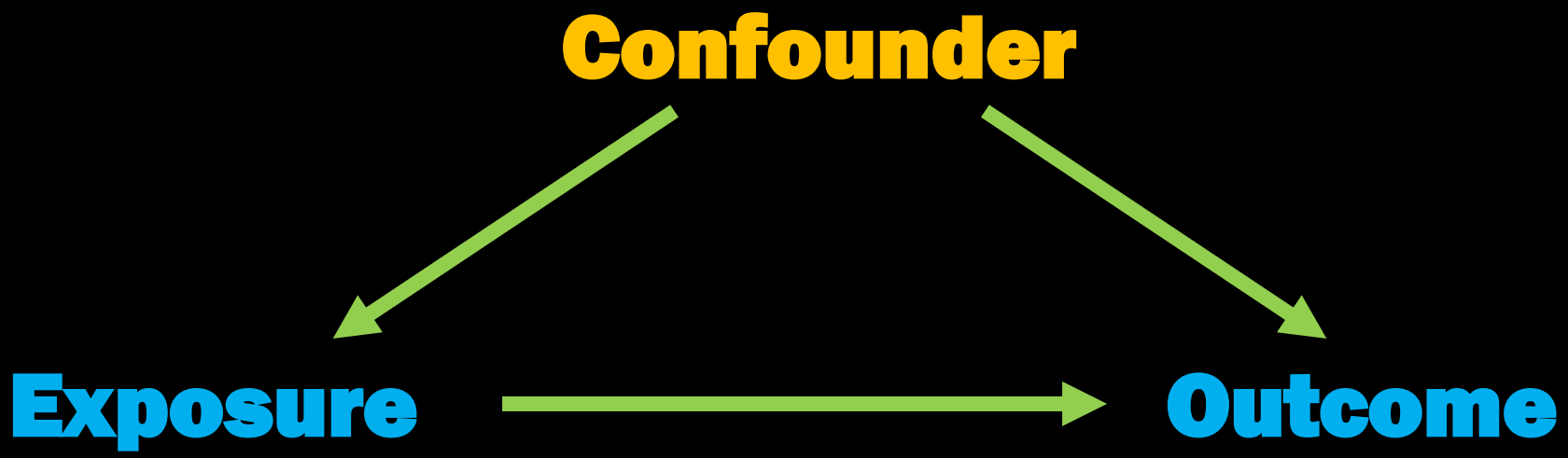
Observational Studies



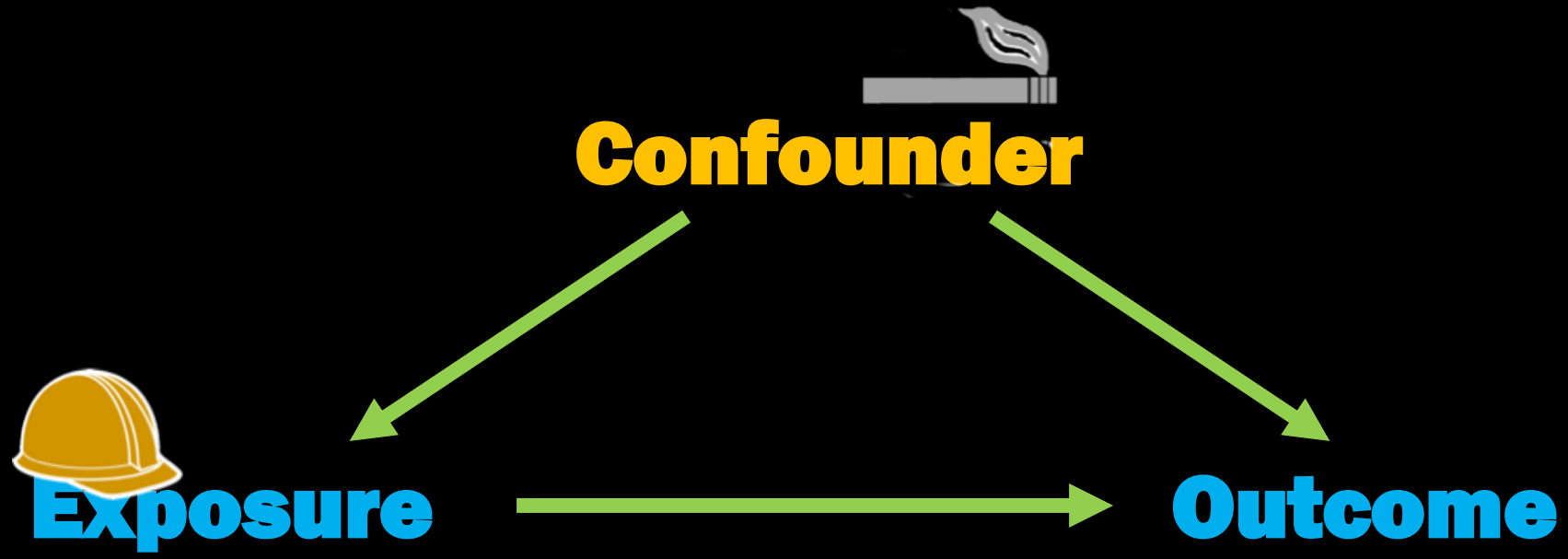
DATA DAY TEXAS 2018

DATA DAY TEXAS 2018

Confounding



Confounding

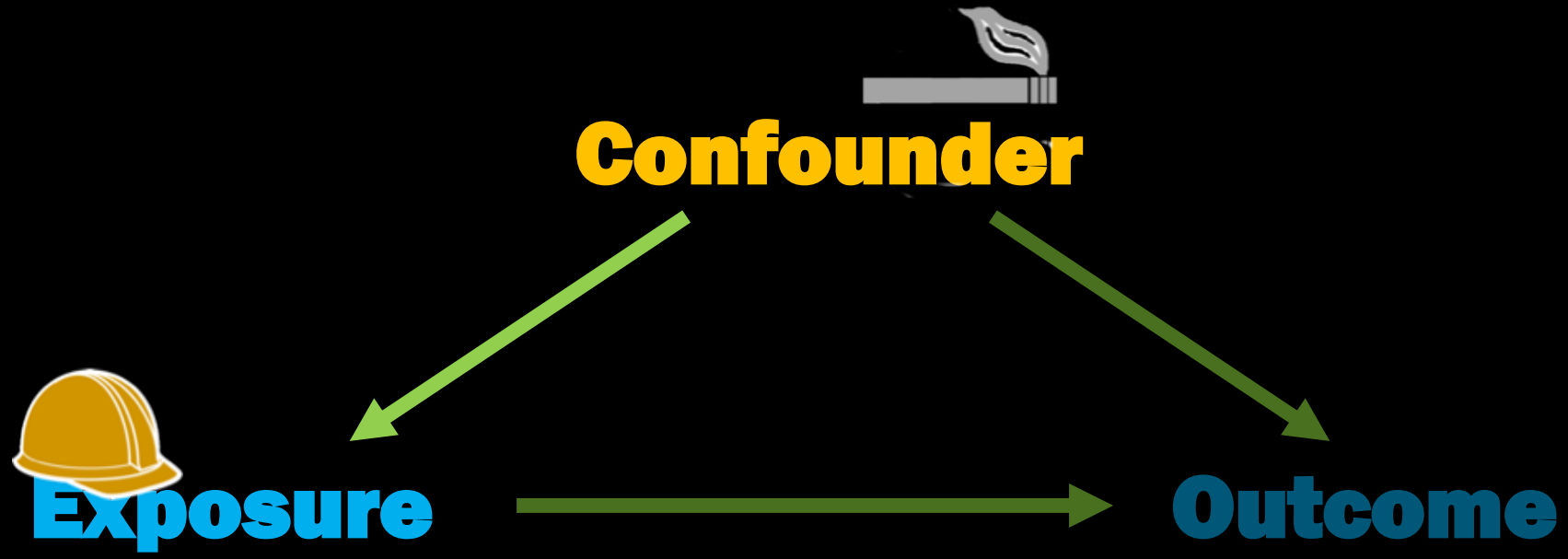


Meaningful confounders

1. How **imbalanced** is the confounder between the exposure groups?

2. How **predictive** is the confounder of the outcome?

Confounding

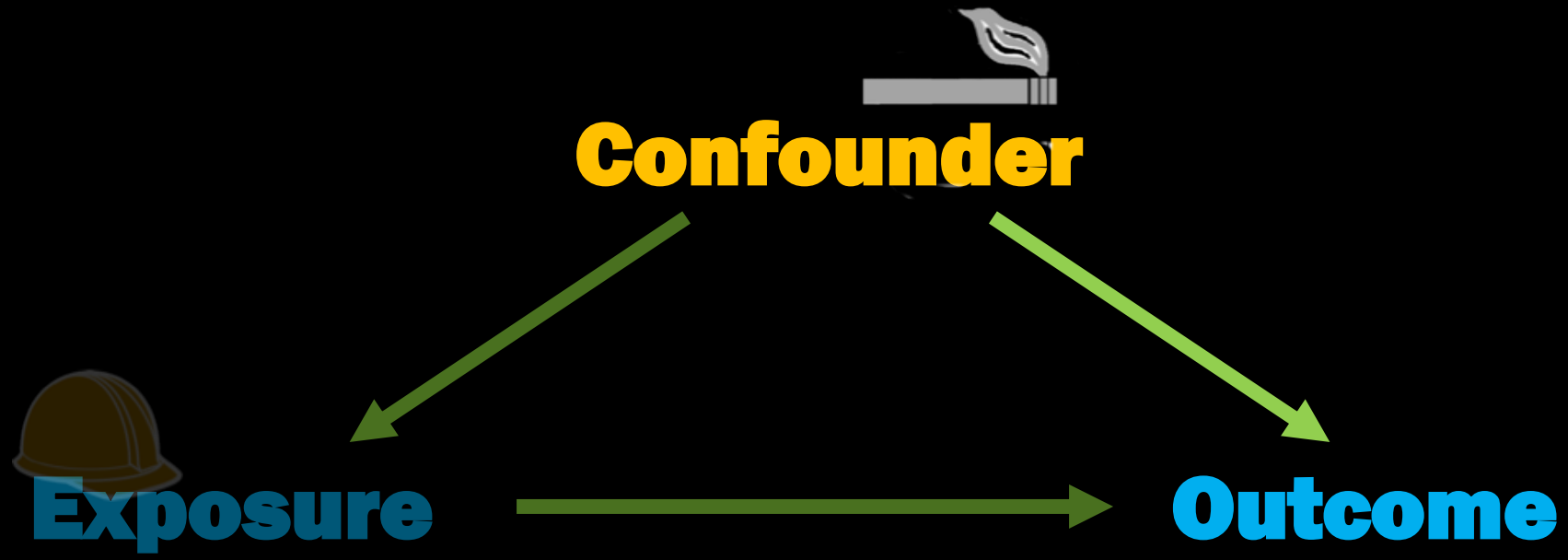


Meaningful confounders

1. How **imbalanced** is the confounder between the exposure groups?

2. How **predictive** is the confounder of the outcome?

Confounding



DATA DAY TEXAS 2018

Propensity scores

Rosenbaum and Rubin showed in observational studies, conditioning on **propensity scores** can lead to unbiased estimates of the exposure effect

1. There are no unmeasured confounders

2. Every subject has a nonzero probability of receiving either exposure

Propensity scores

Rosenbaum and Rubin showed in observational studies, conditioning on **propensity scores** can lead to unbiased estimates of the exposure effect

1. There are no unmeasured confounders

2. Every subject has a nonzero probability of receiving either exposure

Propensity scores

- Fit a **logistic regression** predicting exposure using known covariates

$$\Pr(\text{exposure} = 1) = \frac{1}{1 + \exp(-\mathrm{X}\beta)}$$

- Each individuals' predicted values are the **propensity scores**

Propensity scores

```
glm(exposure ~ smoker + cowboy_hat + sick . . .,
    data = df,
    family = binomial)
```

Propensity scores

```
glm(exposure ~ smoker + cowboy_hat + sick . . .,
    data = df,
    family = binomial) %>%
predict(type = “response”)
```

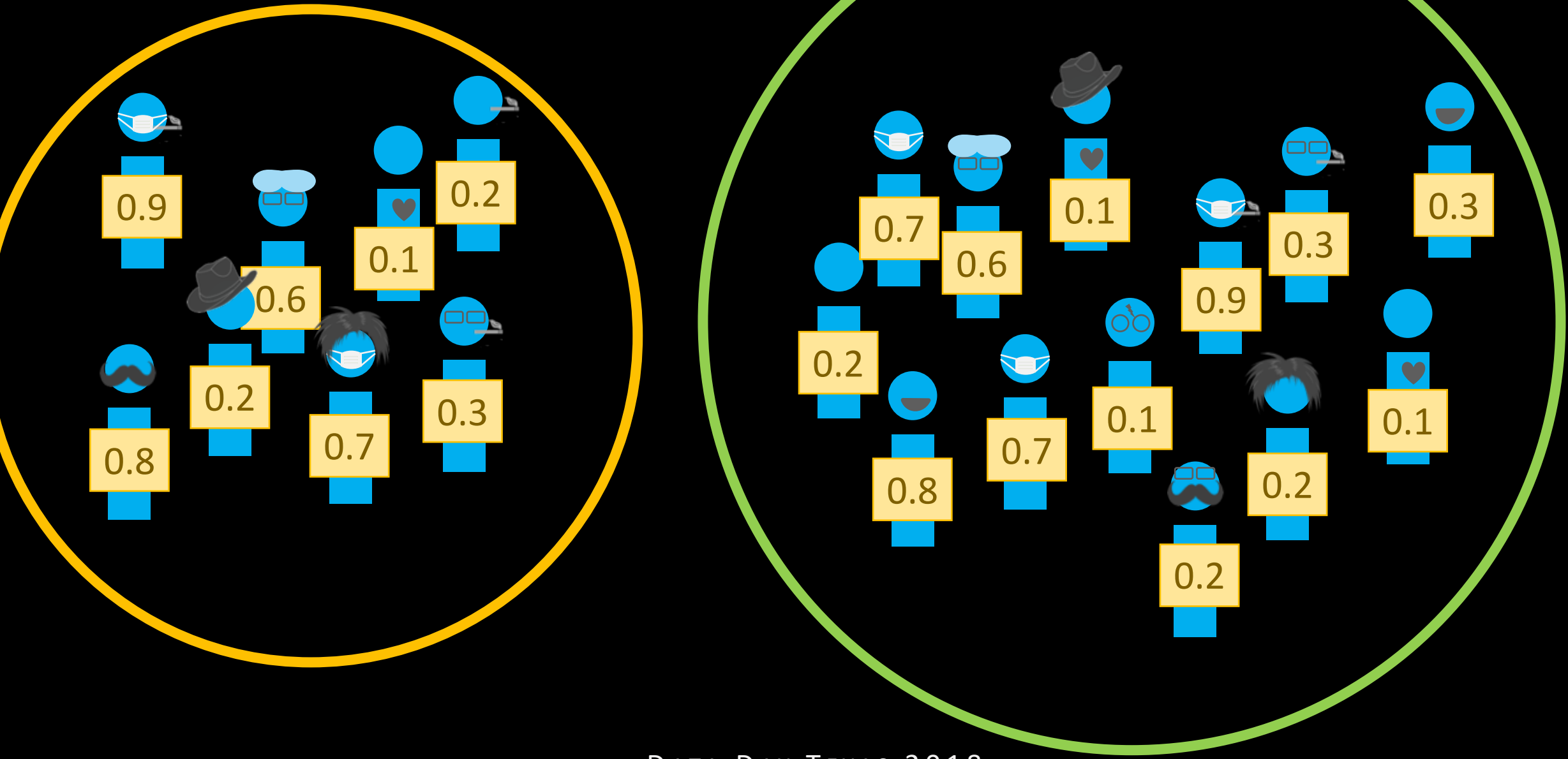
Propensity scores

```
glm(exposure ~ smoker + cowboy_hat + sick . . .,
    data = df,
    family = binomial) %>%
predict(type = “response”)
#> 1         2         3         4         5
#> 0.6967404 0.6370019 0.7195434 0.9079320 0.5905542
```

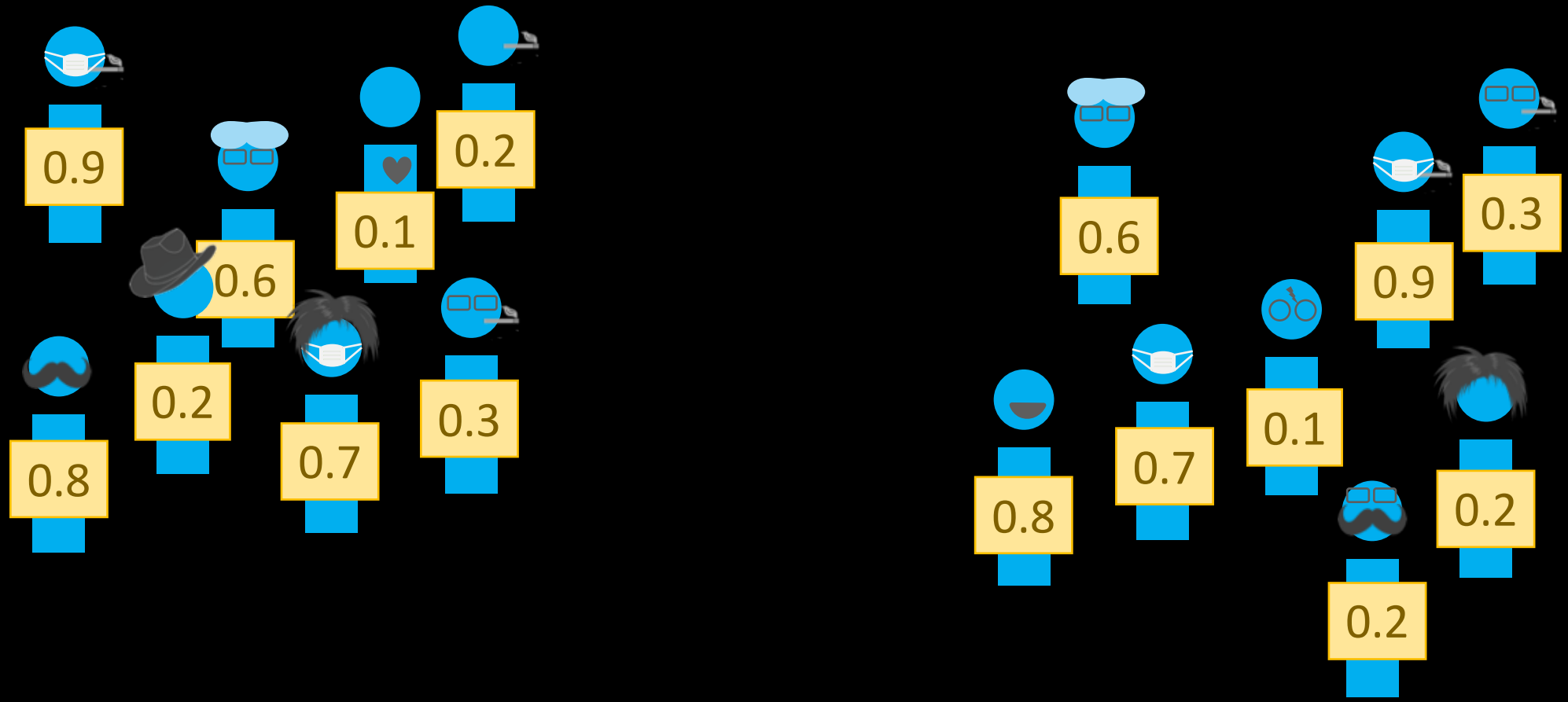
Propensity scores

```
p <- glm(exposure ~ smoker + cowboy_hat + sick . . .,
         data = df,
         family = binomial) %>%
     predict(type = “response”)
```

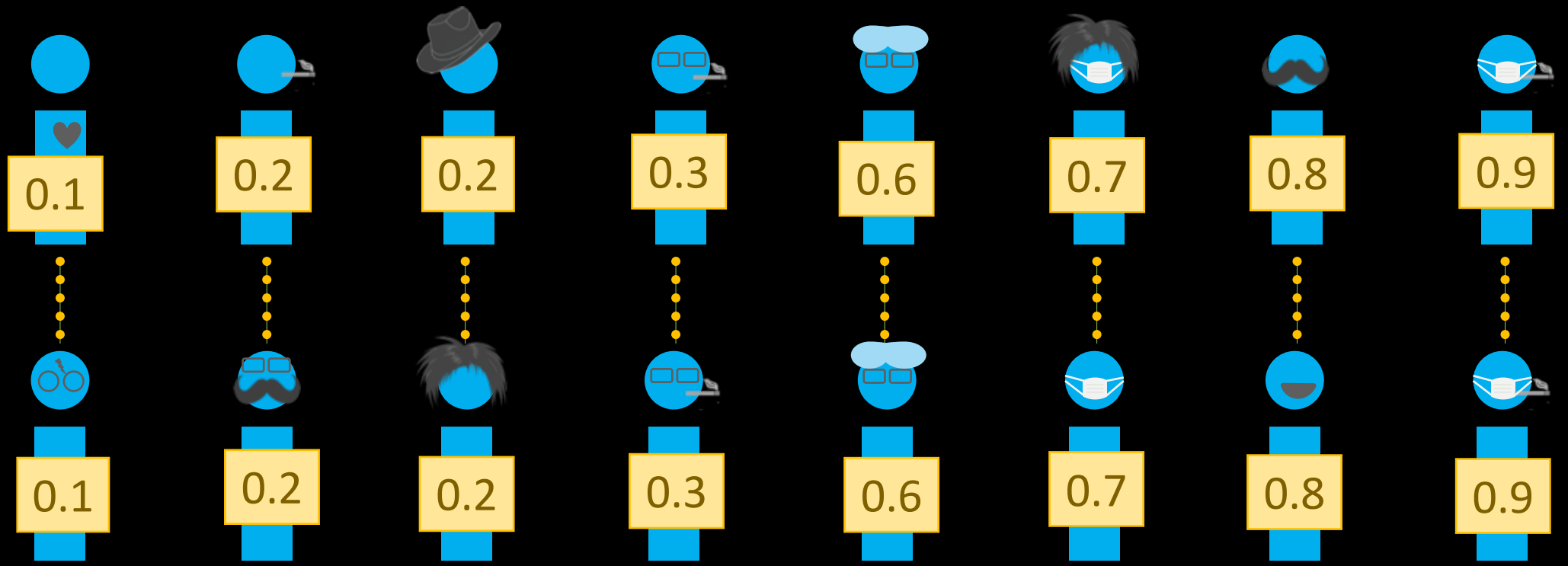
Propensity scores



Propensity scores



Propensity scores



Propensity Scores

- Weighting
- Matching
- Stratification
- Direct Adjustment
- …

Propensity Scores

- **Weighting**
- Matching
- Stratification
- Direct Adjustment
- …

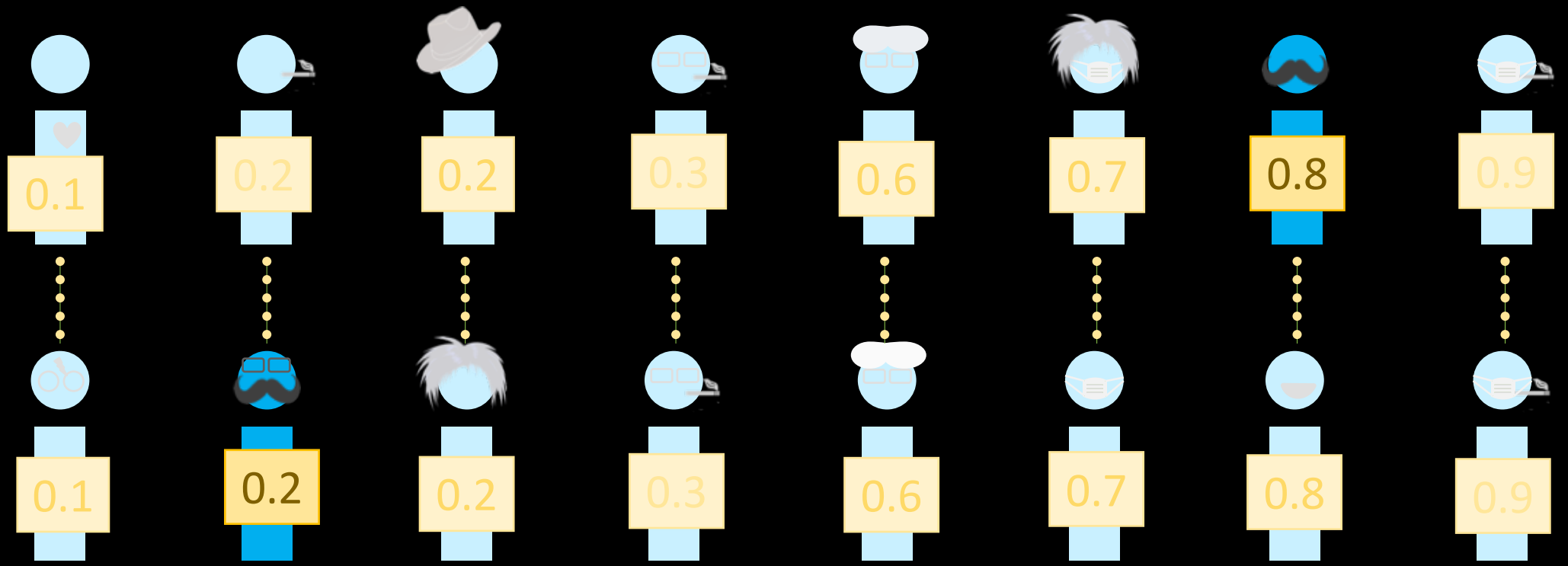
Weighting

$$w_{ATM} = \frac{\min\{p_i, 1 - p_i\}}{Z_i p_i + (1 - Z_i)(1 - p_i)}$$

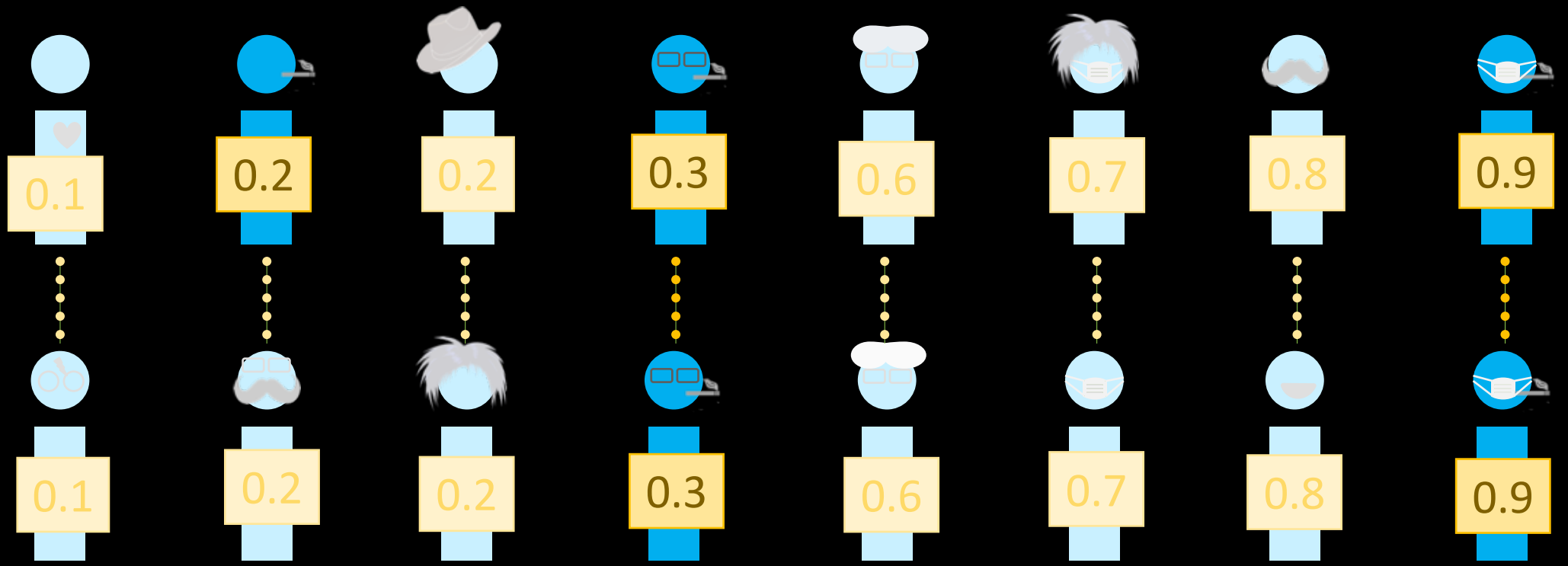
Weighting

```
p_0 <- 1 - p
df$weight <- pmin(p, p_0) /
             ifelse(exposure == 1, p, p_0)
```

Imbalance between exposures



Imbalance between exposures



$$d = \frac{\bar{x}_{exposed} - \bar{x}_{unexposed}}{\sqrt{\frac{s^2_{exposed} + s^2_{unexposed}}{2}}}$$

Standardized mean difference

Standardized Mean Difference

```
library(survey)

svy_des <- svydesign(
  ids = ~ 1,
  data = df,
  weights = ~ weight)
```

Standardized Mean Difference

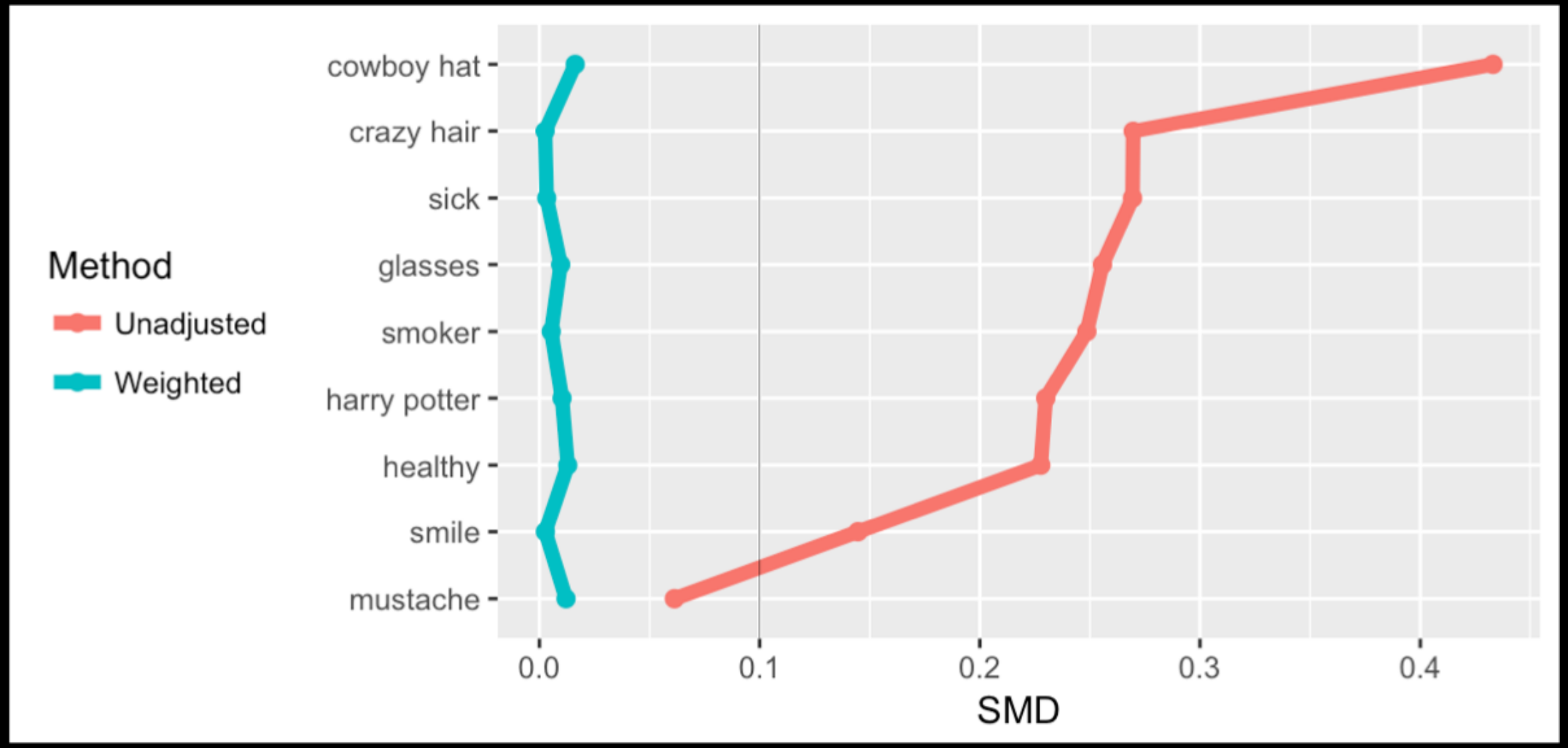
```
library(tableone)

smd_table <- svyCreateTableOne(
  vars = c("smoker", "cowboy_hat", "sick", . . .),
  strata = "exposure",
  data = svy_des,
  test = FALSE)

print(smd_table, smd = TRUE)
```

Standardized Mean Difference

	Exposed	Unexposed	SMD
n	1897.07	1900.24	
mustache	60.71 (16.95)	60.90 (15.61)	0.012
cowboy hat	204.59 (101.45)	202.90 (106.44)	0.016
glasses	70.77 (27.00)	70.51 (27.50)	0.01
smoker	37.46 (10.93)	37.40 (11.28)	0.005
crazy hair	2.27 (2.46)	2.27 (1.82)	0.003
smile	2.49 (5.48)	2.47 (4.80)	0.003
sick	30.92 (8.23)	30.95 (7.54)	0.003
harry potter	3.04 (0.69)	3.03 (0.93)	0.01
healthy	145.4 (7.7)	152.2 (8.0)	0.013

Standardized Mean Difference

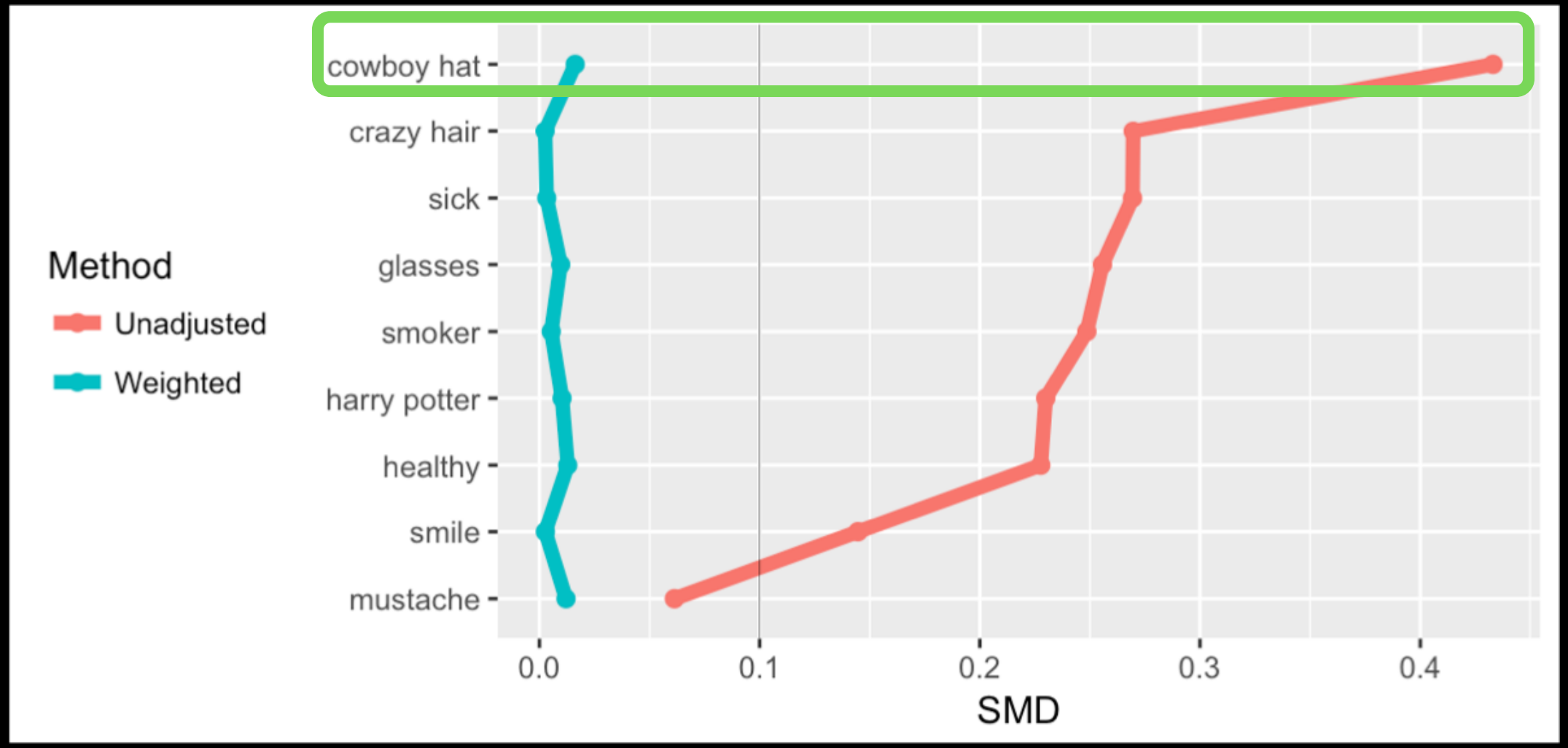
```
library(tableone)

smd_table_unweighted <- CreateTableOne(
  vars = c("smoker", "cowboy_hat", "sick", . . .),
  strata = "exposure",
  data = df,
  test = FALSE)

print(smd_table_unweighted, smd = TRUE)
```



Love Plots

DATA DAY TEXAS 2018

Imbalance between exposures



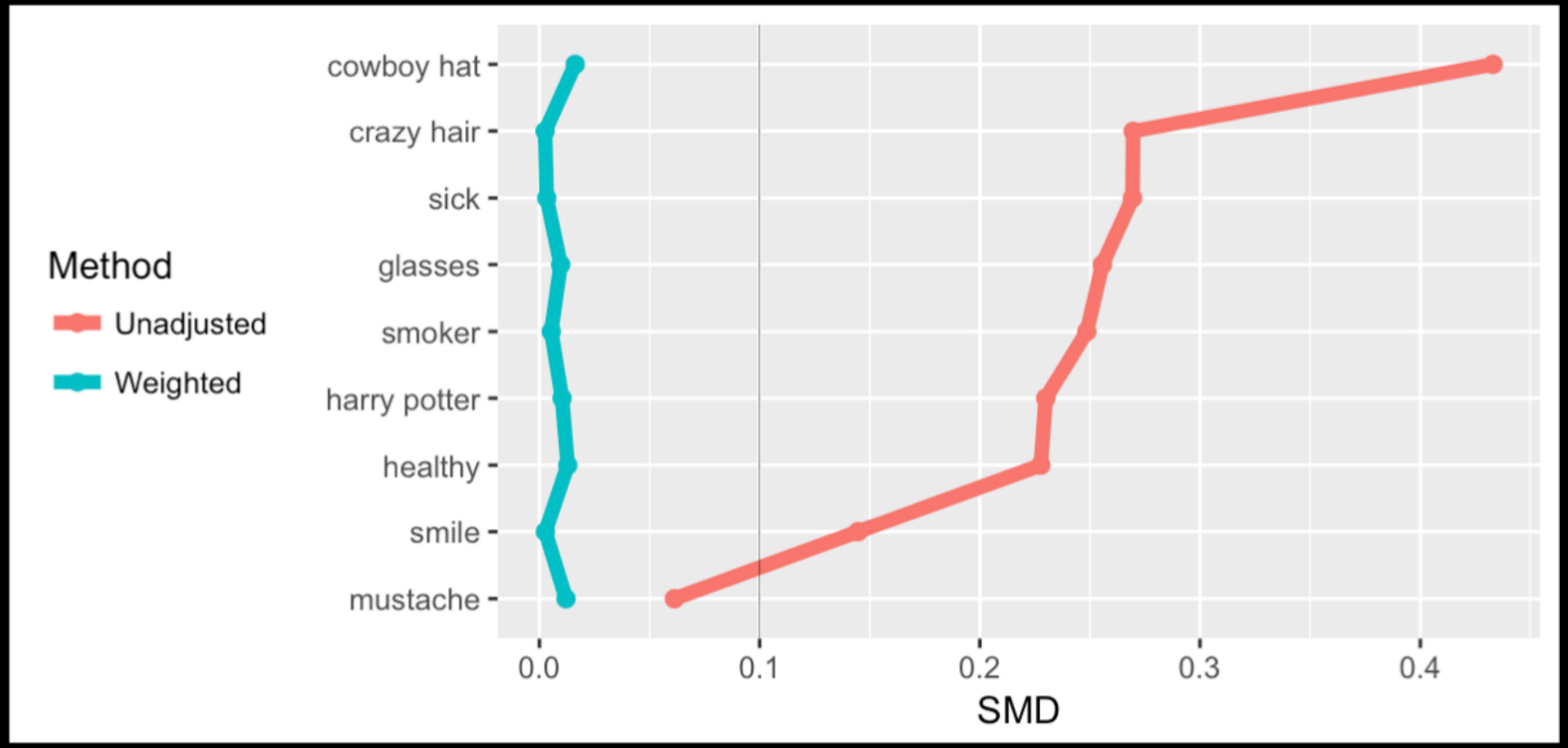
Love Plots

Standardized Mean Difference

```
library(tidyr)

plot_df <- data.frame(
  var = names(ExtractSmd(smd_table)),
  Unadjusted = ExtractSmd(smd_table_unweighted ),
  Weighted  = ExtractSmd(smd_table)) %>%
  gather("Method", "SMD", -var)
```

Standardized Mean Difference

```
library(ggplot2)

ggplot(
  data = plot_df,
  mapping = aes(x = var, y = SMD, group = Method, color = Method)
  ) +
  geom_line() +
  geom_point() +
  geom_hline(yintercept = 0.1, color = "black", size = 0.1) +
  coord_flip()
```



Love Plots

DATA DAY TEXAS 2018

Outcome model

```
svyglm(
  outcome ~ exposure,
  data = svy_des,
  family = binomial)
```

Outcome model

```
svyglm(
  outcome ~ exposure,
  data = svy_des,
  family = binomial) %>%
  confint(parm = “exposure”)
```

Outcome model

```
svyglm(
  outcome ~ exposure,
  data = svy_des,
  family = binomial) %>%
  confint(parm = “exposure”) %>%
  exp()
#>              2.5%   97.5%
#> exposure  1.51   2.01
```

unmeasured confounding

All you need

- ✓ Exposure-outcome effect
- ✓ Exposure-unmeasured confounder effect
- ✓ Outcome-unmeasured confounder effect

All you need

- ✓ ~~Exposure-outcome effect~~ generally estimated from a model, for e o, or risk ratio
- ✓ ~~Exposure-u~~
- ✓ ~~Outcome-u~~

Odds Ratio
Confidence Interval:
(1.5, 2)

All you need

- ✓ ~~Exposure-outcome effect~~ generally estimated from a model, for example an odds ratio, hazard ratio, or risk ratio
- ✓ **Exposure-unmeasured confounder effect**
- ✓ **Outcome-unmeasured confounder effect**

Tipping point analyses

what will tip our confidence bound to cross 1

DATA DAY TEXAS 2018

All you need

- ✓ ~~Exposure-outcome effect~~ generally estimated from a model, for example an odds ratio, hazard ratio, or risk ratio
- ✓ ~~Exposure-unmeasured confounder effect~~ can use my observed variables to estimate this

- ✓ Outcome-unmeasured confounder effect

All you need

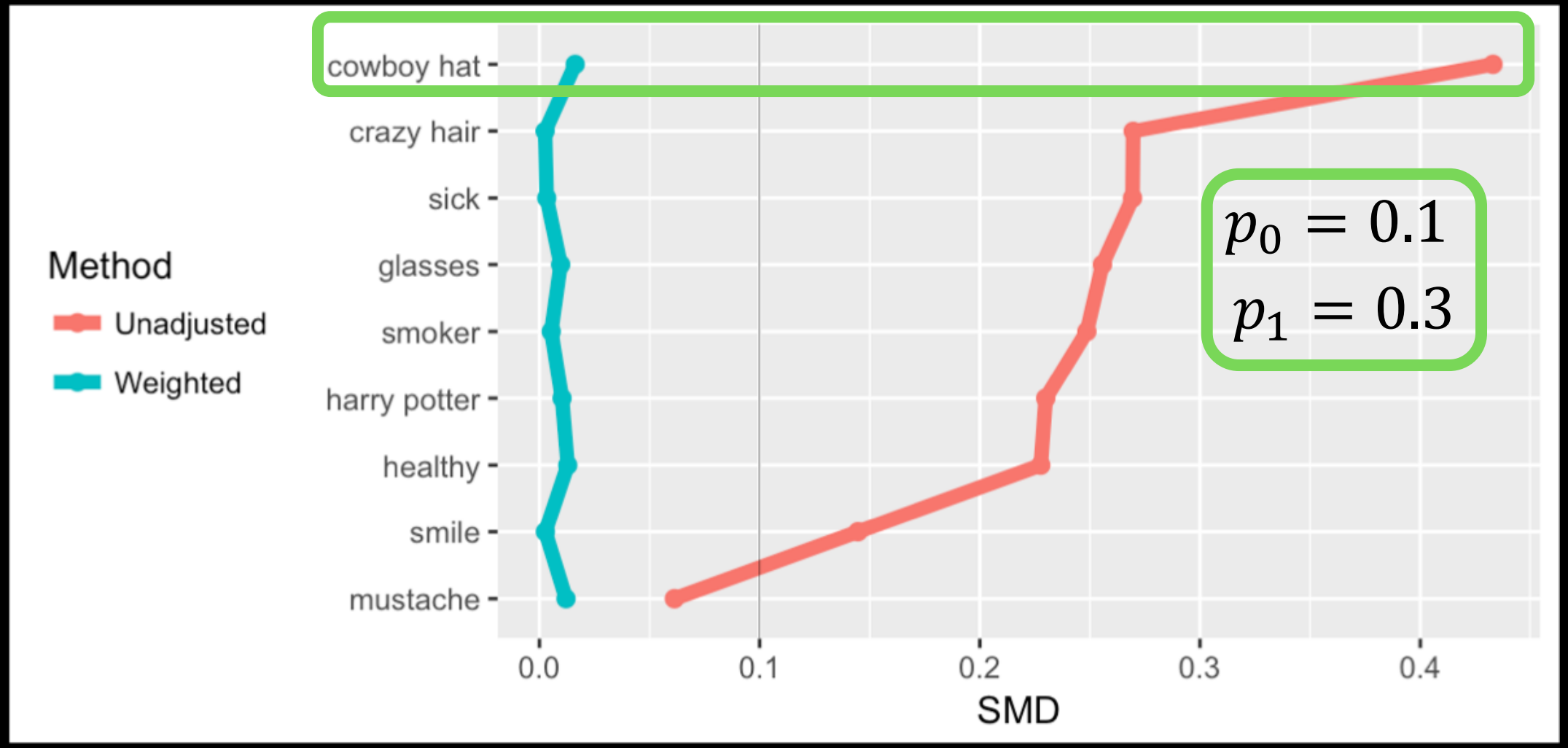
- ✓ ~~**Exposure-outcome effect**~~ generally estimated from a model, for example an odds ratio, hazard ratio, or risk ratio
- ✓ **Exposure-unmeasured confounder effect** can use my observed variables to estimate this

- ✓ **Outcome-unmeasured confounder effect**

Exposure-unmeasured confounder effect

- Prevalence of the unmeasured confounder among the **exposed**

- Prevalence of the unmeasured confounder among the **unexposed**

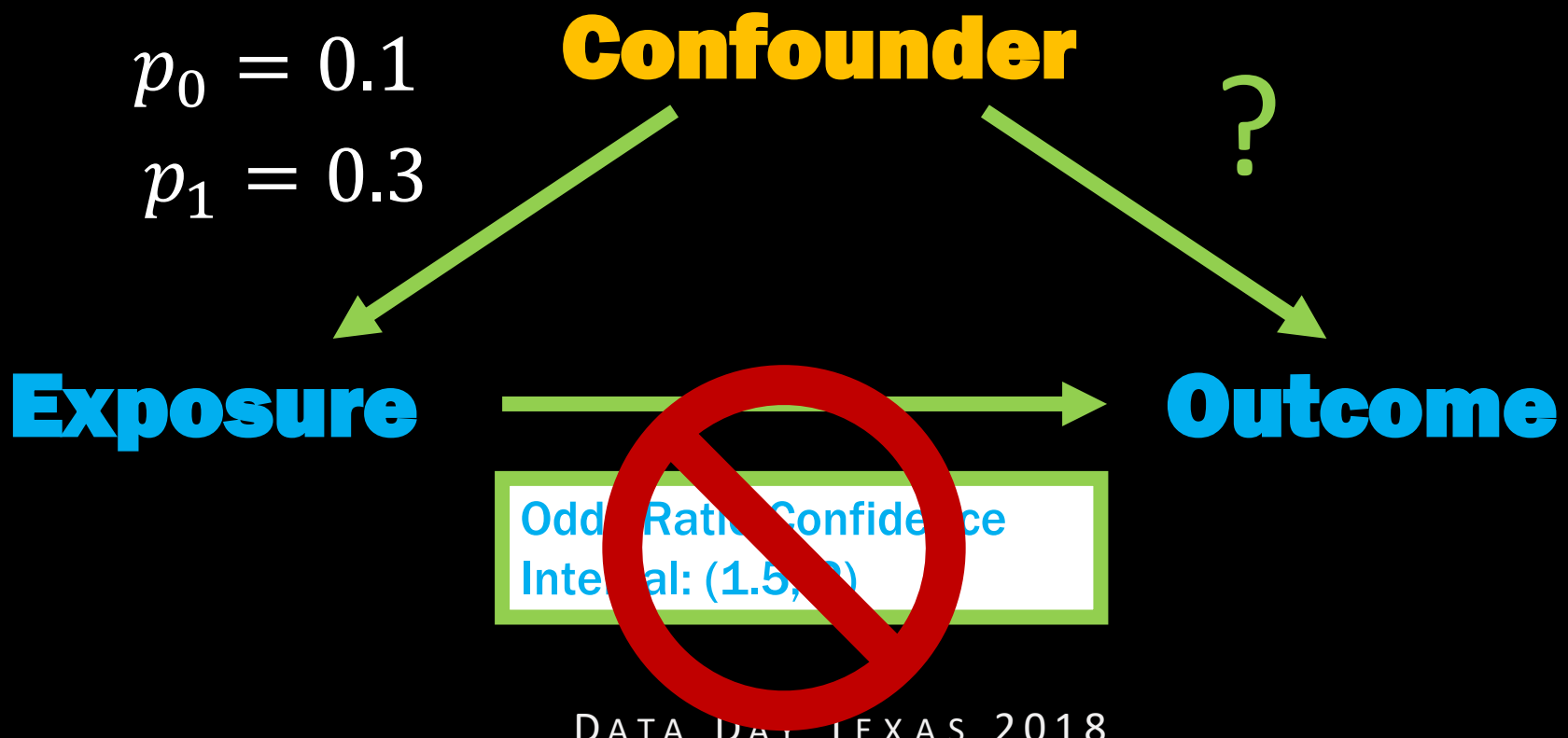

cowboy hat
crazy hair
sick
glasses
smoker
harry potter
healthy
smile
mustache
Method
Unadjusted
Weighted
$p_0 = 0.1$
$p_1 = 0.3$
0.0
0.1
0.2
0.3
0.4
SMD

Love Plots

Confounding

$p_0 = 0.1$

$p_1 = 0.3$



Standardized Mean Difference

```
library(tipr)

tip_with_binary(
  p0 = 0.1,
  p1 = 0.3,
  lb = 1.5,
  ub = 2)
```

Standardized Mean Difference

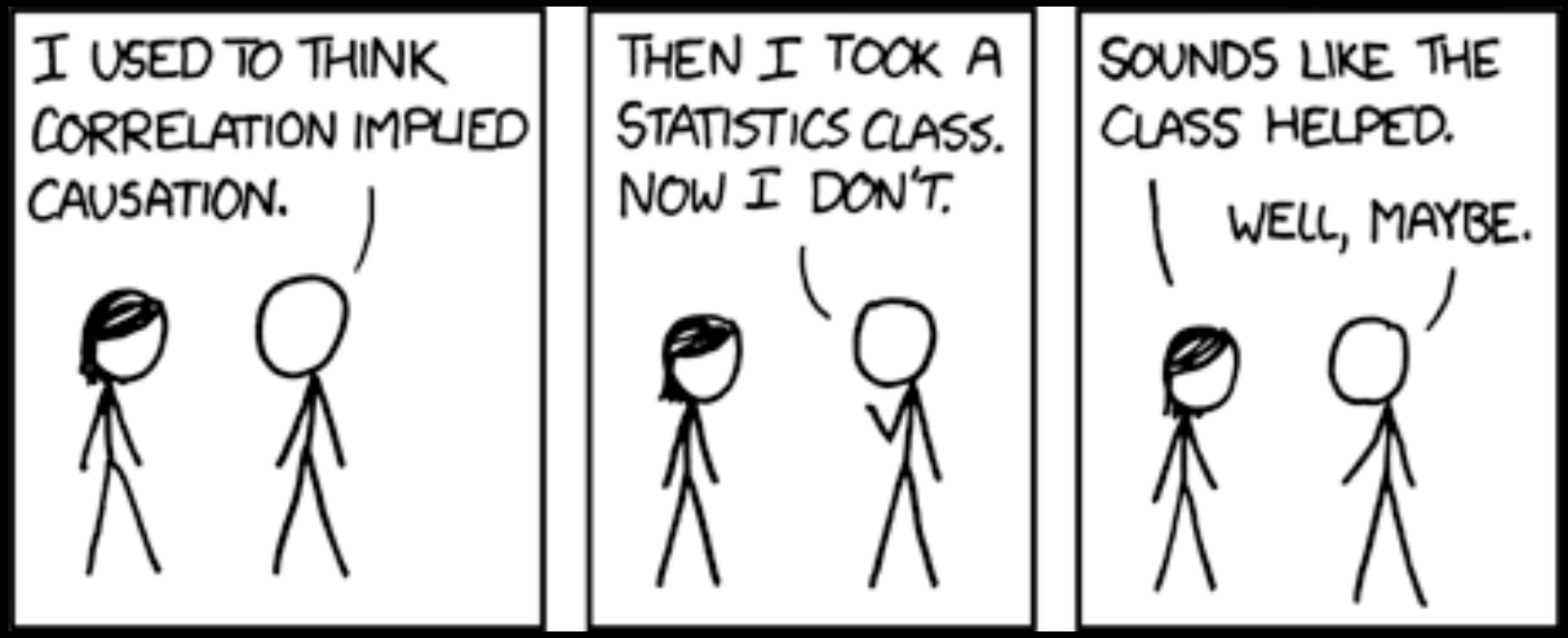
```
library(tipr)

tip_with_binary(
  p0 = 0.1,
  p1 = 0.3,
  lb = 1.5,
  ub = 2)
#> [1] 4.333333
```

> A hypothetical unobserved binary confounder that is prevalent in 30% of the exposed population and 10% of the unexposed population would need to have an association with Y of 4.33 to tip this analysis at the 5% level, rendering it inconclusive.

✌ parts

1. Discuss some ways to strengthen a causal argument: **Hill's criteria**

2. Discuss a specific causal inference method: **propensity scores + sensitivity analyses**



https://xkcd.com/552/

R 

- survey
- tableone
- tidyr
- ggplot2
- tipr

References

1. Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (2009). Smoking and lung cancer: recent evidence and a discussion of some questions. 1959. *International journal of epidemiology* (Vol. 38, pp. 1175–1191). Oxford University Press. http://doi.org/10.1093/ije/dyp289
2. Schlesselman, J. J. (1978). Assessing effects of confounding variables. *American Journal of Epidemiology*, *108*(1), 3–8.
3. Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41. http://doi.org/10.2307/2335942
4. Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, *54*(3), 948–963.
5. https://cran.r-project.org/web/packages/tableone/vignettes/smd.html

Thank you!

@LucyStats

http://bit.ly/LucyStatsDDTX18

lucymcgowan.com