Data Amnesty

WE ALL COUNT

project for equity in data science

DEMYSTIFY. DEMOCRATIZE. DEMONSTRATE.

Sources of bias can be identified in each step of the data life cycle.

What is data?

Microdata

	A	B	C	D	E
1	**Id Number**	**State**	**Race**	**Gender**	**Income**
2	1001	MA	Asian	F	$28,900
3	1002	MA	White	M	$107,490
4	1003	MA	Black	F	$77,320
5	1004	MA	American Indian	F	$12,302
6	1005	MA	American Indian	F	$54,034
7	1006	MA	Asian	M	$112,560
8	1007	MA	White	M	$245,607
9	1008	MA	Black	M	$103,25[illegible]
10	1009	MA	Black	F	[illegible]
11	1010	MA	Black	F	[illegible]
12	1011	MA	White	F	[illegible]
13	1012	MA	White	M	[illegible]
14	1013	MA	White	M	[illegible]
15	1014	MA	White	M	[illegible]
16	1015	MA	American Indian	M	[illegible]
17	1016	MA	American Indian	M	[illegible]
18	1017	MA	Asian	F	[illegible]
19	1018	MA	White	M	[illegible]
20	1019	MA	Black	M	[illegible]
21	1020	MA	Black	F	[illegible]
22	1021	MA	Black	M	[illegible]
23	1022	MA	White	F	[illegible]
24	1023	MA	Asian	M	[illegible]
25	1024	MA	White	F	[illegible]
26	1025	MA	Black	F	[illegible]
27	1026	MA	Asian	F	$31,84[illegible]
28	1027	MA	Asian	F	$32,845
29	1028	MA	Asian	F	$37,590
30	1029	MA	Black	M	$63,401
31	1030	MA	White	M	$9,045
32	1031	MA	White	M	$11,368
33	1032	MA	Black	F	$37,451
34	1033	MA	Black	F	$42,309
35	1034	MA	White	F	$42,941
36	1035	MA	Asian	M	$73,081
37	1036	MA	Asian	M	$32,058

Aggregate Data

RACE	GENDER	AVERAGE INCOME
American Indian	Male	$71,945
	Female	$33,168
Asian	Male	$64,282
	Female	$45,079
Black	Male	$73,116
	Female	$41,392
White	Male	$91,020
	Female	$47,982

Statistical Information

```
> fit1 <- lme(sumMsrNumDonations ~ gender*race = ~ 1|GWID,na.action=na.omit)
> wald(fit1)
```

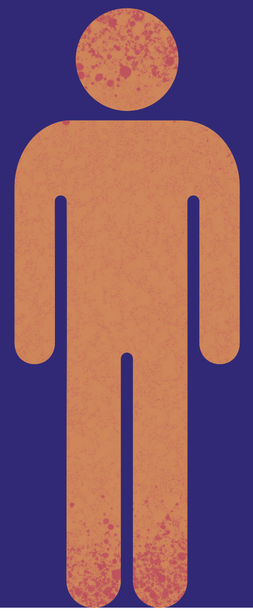
numDF	denDF	F.value	p.value
4	467	108.2909	<.00001

Coefficients	Estimate	Std.Error	DF	t-value	p-value	Lower 0.95	Upper 0.95
(intercept)	1190488.0019	136315.3659	467	8.733337	<.00001	922620.57053	1458355.44333
gender	-20442.3855	13108.4705	467	-1.559479	0.11956	-46201.27404	5316.5030
race	-9902.1188	1599.8718	467	-6.189320	<.00001	-13045.95772	674.8334
gender:race	366.0572	157.1335	467	2.329594	0.02025	57.28096	674.8334

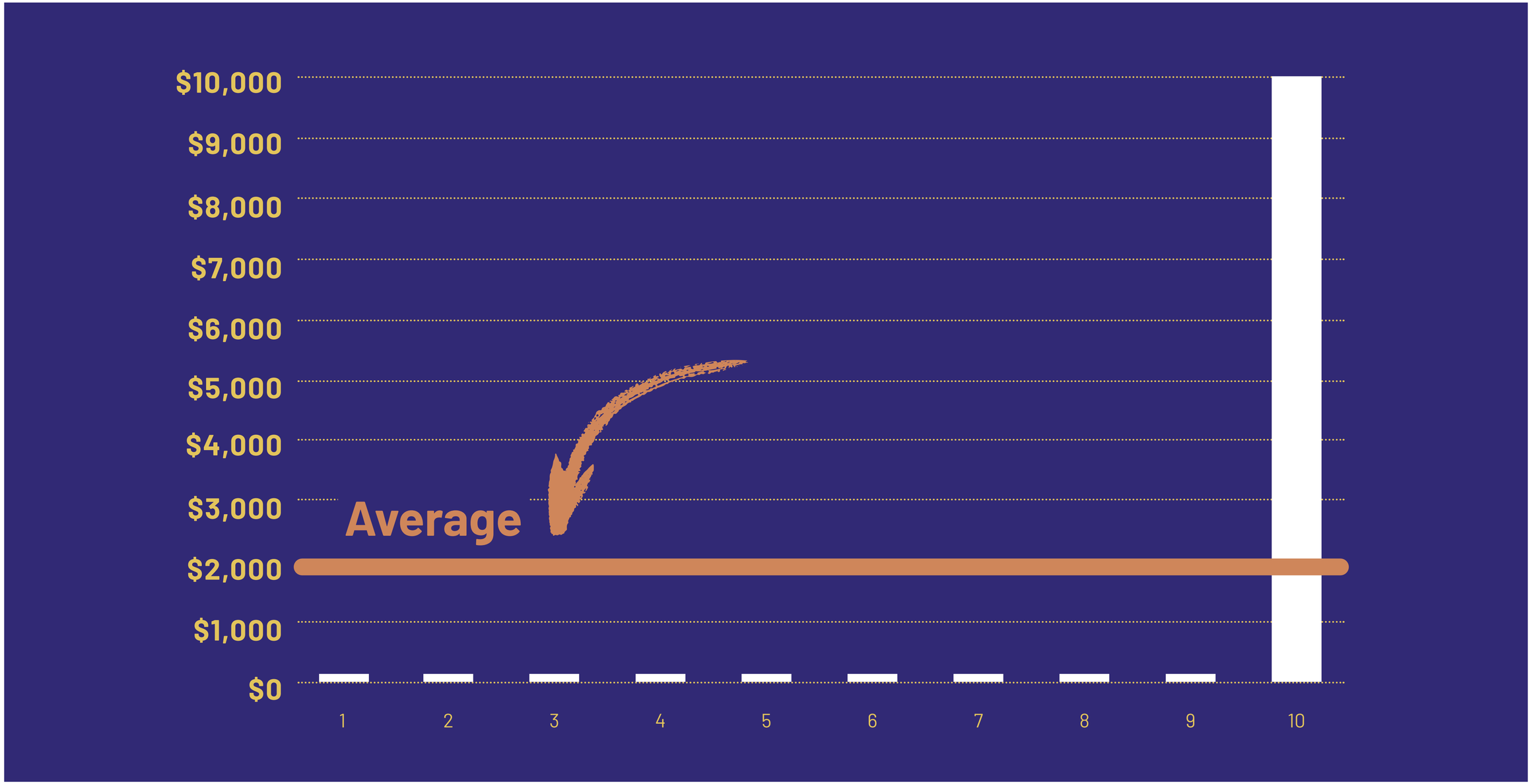
Statistical Information

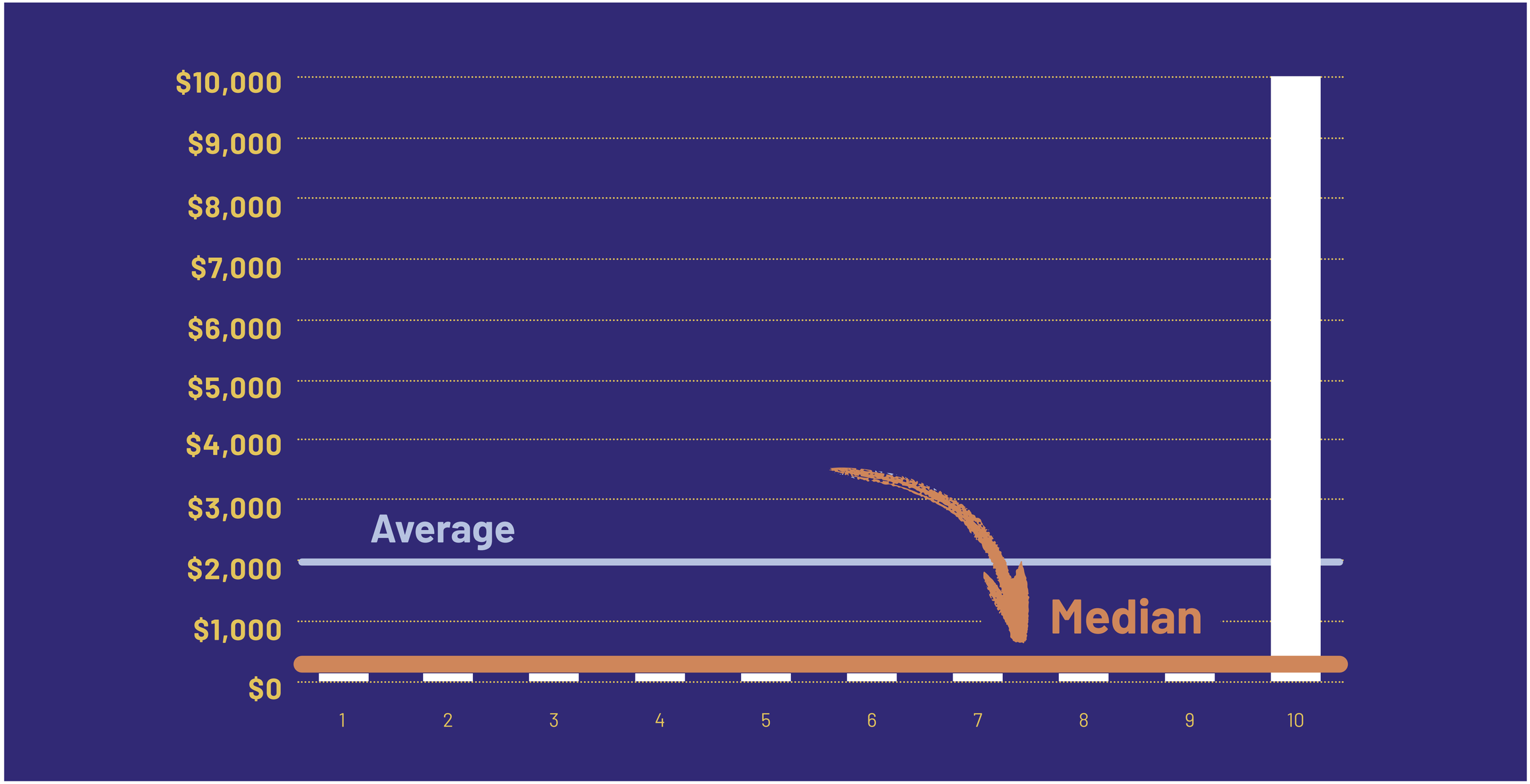
Men in this state tend to earn 84% more than women, on average in this state. Plus or minus 2%.

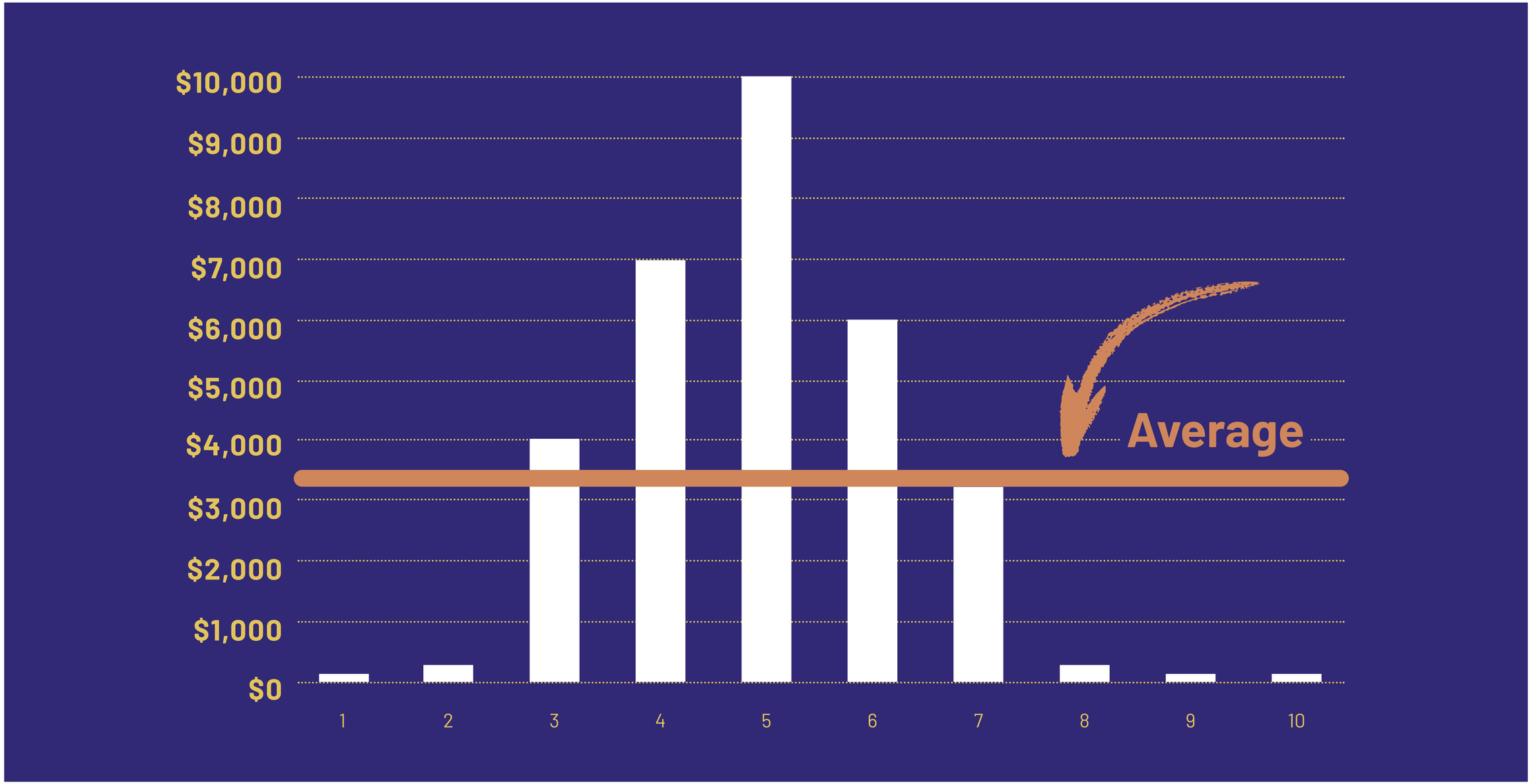
However, if you consider incomes by race, men in this state tend to earn 79% more than women. Plus or minus 4%.

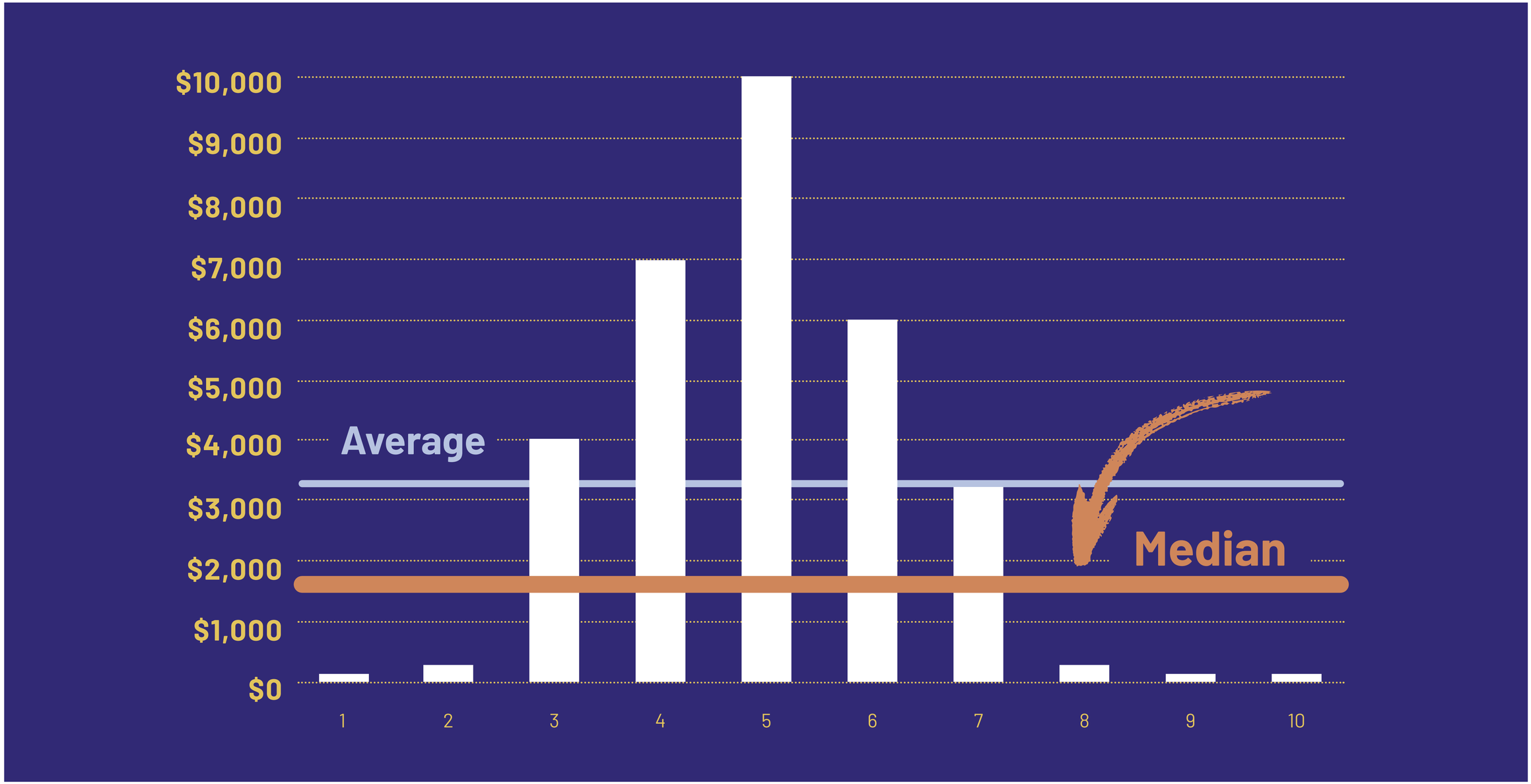


Means, Medians, Measures of Central Tendency


$10,000
$9,000
$8,000
$7,000
$6,000
$5,000
$4,000
$3,000
$2,000
$1,000
$0
Average
1
2
3
4
5
6
7
8
9
10


$10,000
$9,000
$8,000
$7,000
$6,000
$5,000
$4,000
$3,000
$2,000
$1,000
$0
Average
Median
1
2
3
4
5
6
7
8
9
10


$10,000
$9,000
$8,000
$7,000
$6,000
$5,000
$4,000
$3,000
$2,000
$1,000
$0
1
2
3
4
5
6
7
8
9
10
Average


$10,000
$9,000
$8,000
$7,000
$6,000
$5,000
$4,000
$3,000
$2,000
$1,000
$0
Average
Median
1
2
3
4
5
6
7
8
9
10

Let's look at this another way.

Examining the average achievement score of three different schools, we find that each school experienced a 5-point increase:

School A increased their score from 20 to 25 | **School B went from 50 to 55** | **School C rose from 95 to 100**

All three schools experienced the exact same amount of change: 5 points. But because the context varies — each school is working to improve on its unique score — the percentage change is wildly different:

School A improved by 25% | **School B went up by 10%** | **School C saw only a 5% change**

	School A	School B	School C
Performance Score 2015	20	50	95
Performance Score 2016	25	55	100
Amount Increase	5	5	5
Percent Incease	25%	10%	5%

“Census reveals minority population at an all time high”

400 → 500

Smalltown’s minority population grows from 400 to 500

3000 → 4000

Smalltown’s population grows from 3000 to 4000 in the same period

“Census reveals minority population at an all time high”

400 → 500

Smalltown’s minority population grows from 400 to 500

3000 → 4000

Smalltown’s population grows from 3000 to 4000 in the same period

The change in minority population is a drop from **13.3% to 12.5%**

What are the relationships between variables?

5

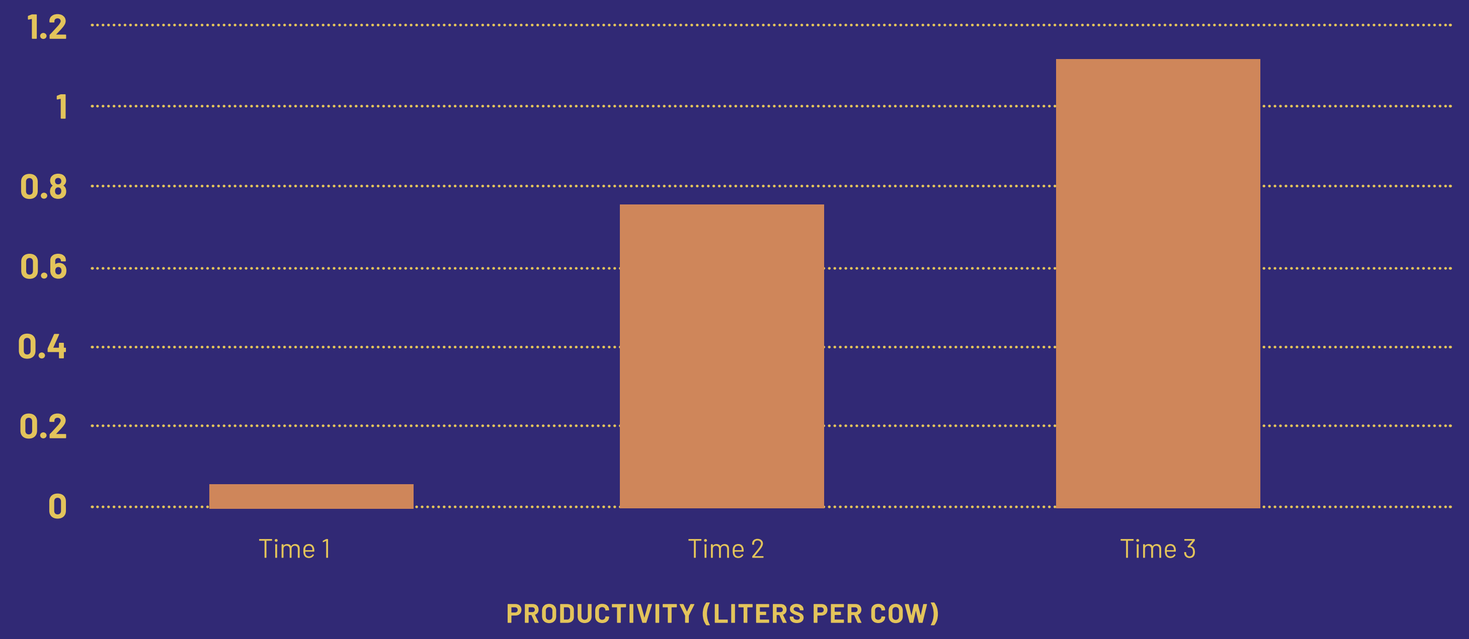
explanations for data relationships:

1. **Direct relationship** *in one direction*
2. **Direct relationship** *in the other direction*
3. **Confounded relationship**
4. **Mediated relationship**
5. **Chance**

Moderators, Mediators, Confounders

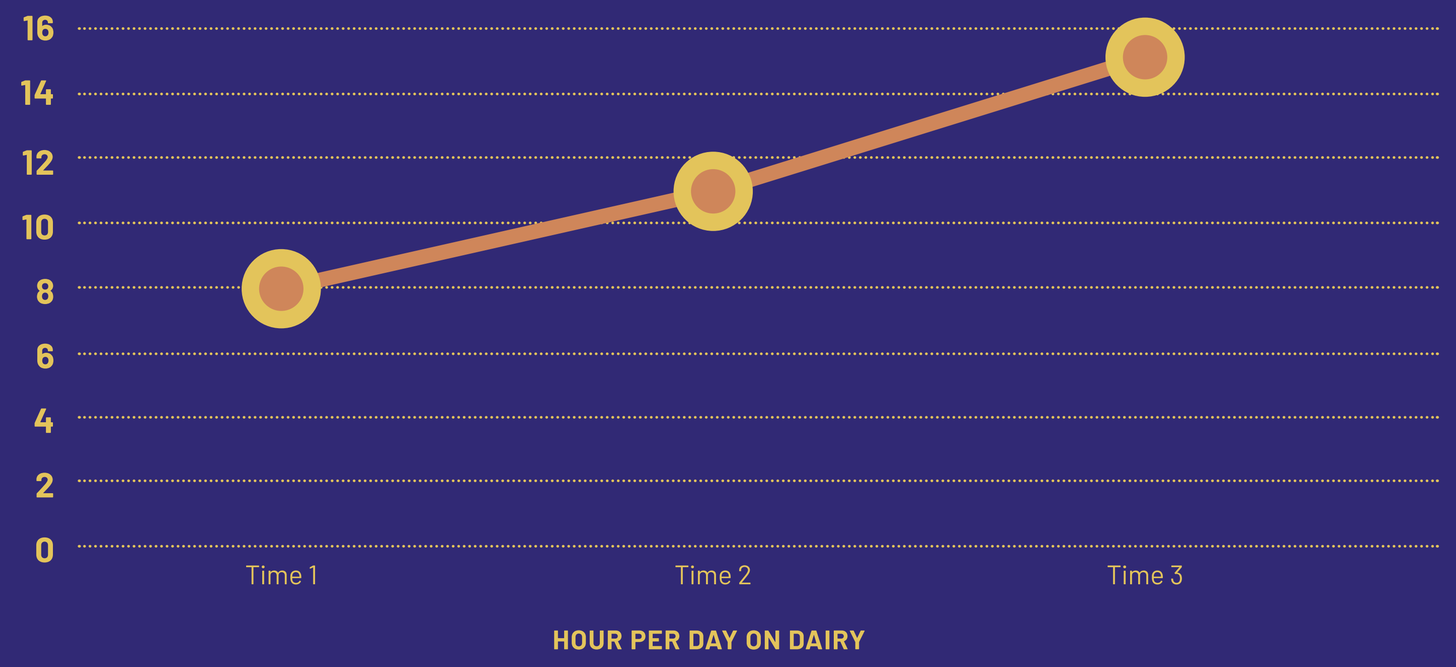
Mediators, Moderators, Confounders

Productivity increases over time



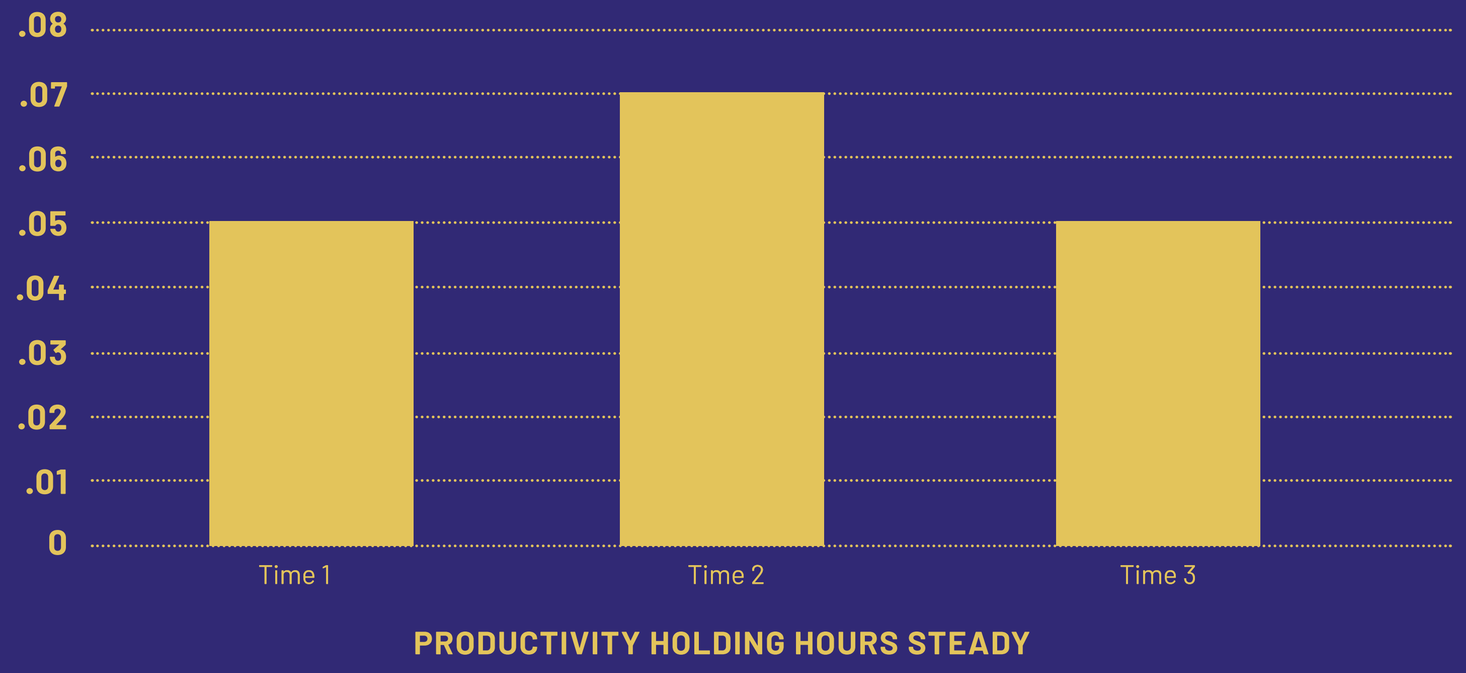
Mediators, Moderators, Confounders

Hours of farm work increases over time



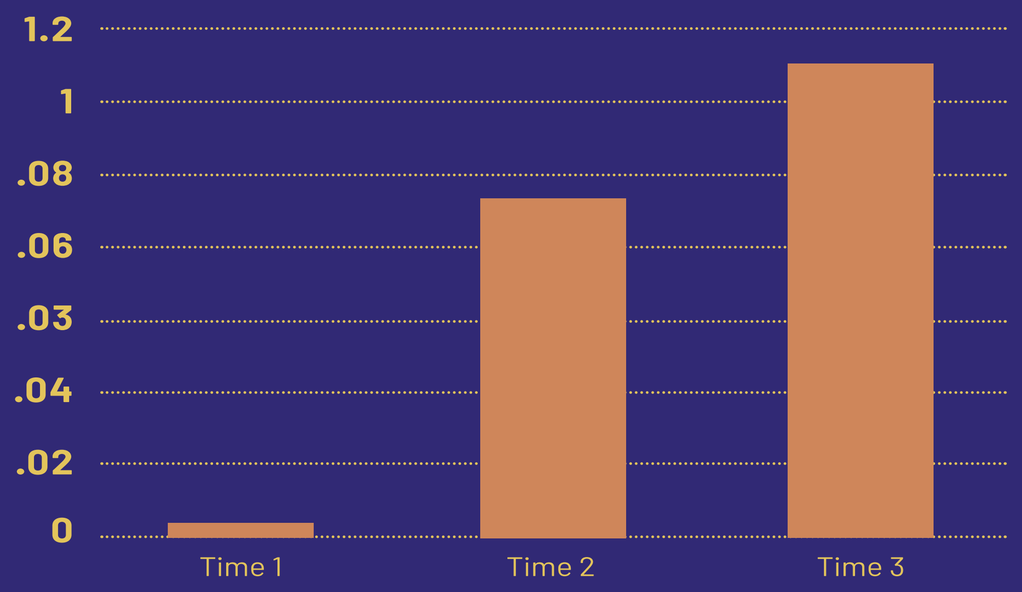
Mediators, Moderators, Confounders

Productivity controlling for increase in work time

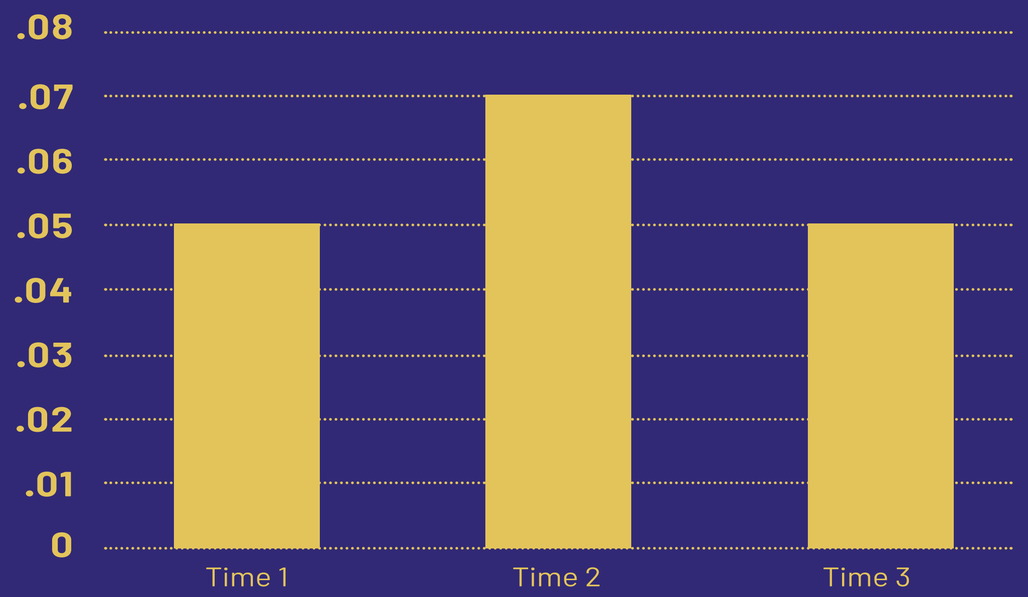


Mediators, Moderators, Confounders

Productivity controlling for increase in work time



PRODUCTIVITY HOLDING HOURS STEADY



PRODUCTIVITY HOLDING HOURS STEADY

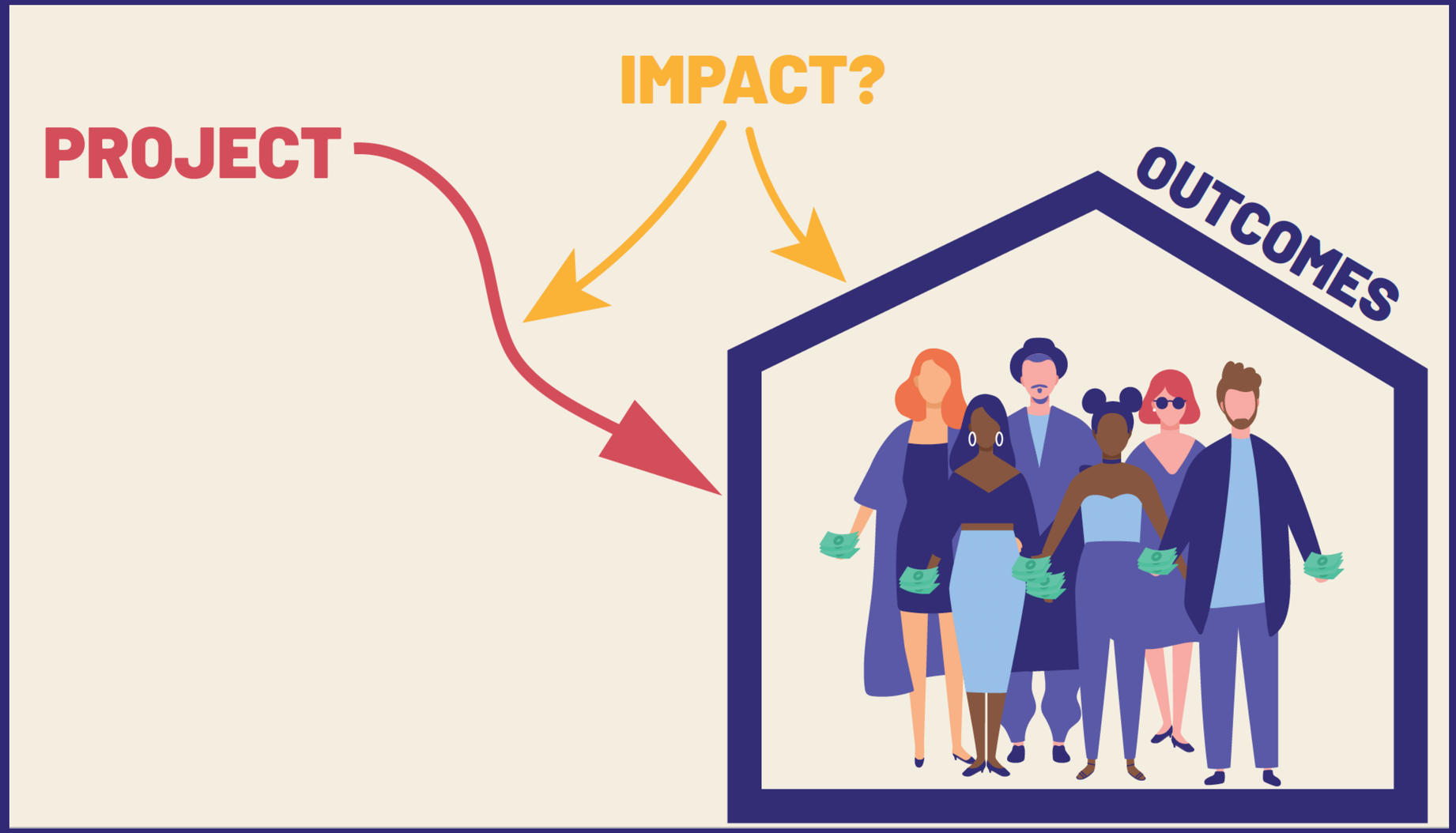
Impact Analysis & RCTs

PROJECT

PROJECT

PROJECT

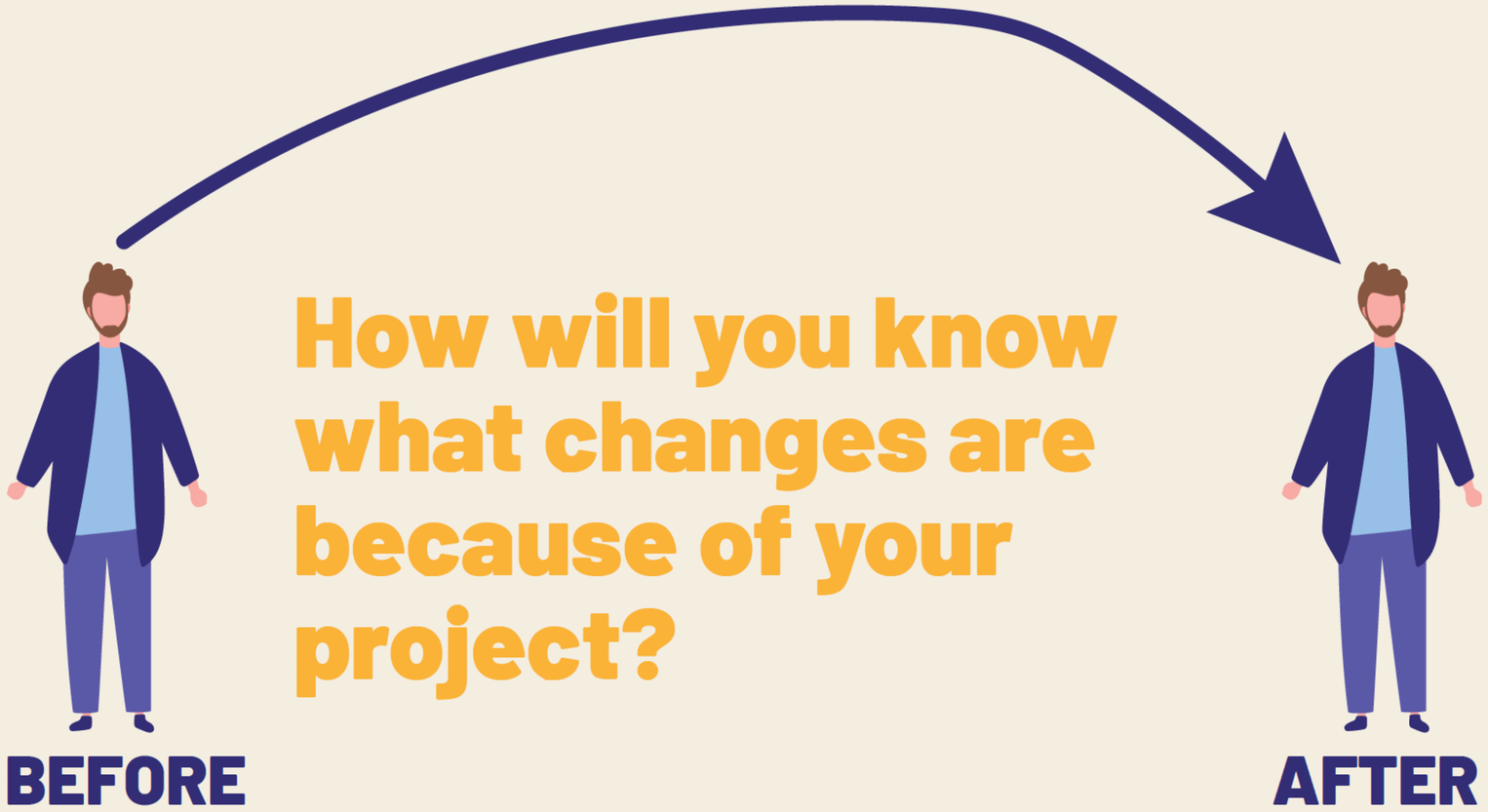
PROJECT


IMPACT?
PROJECT
OUTCOMES

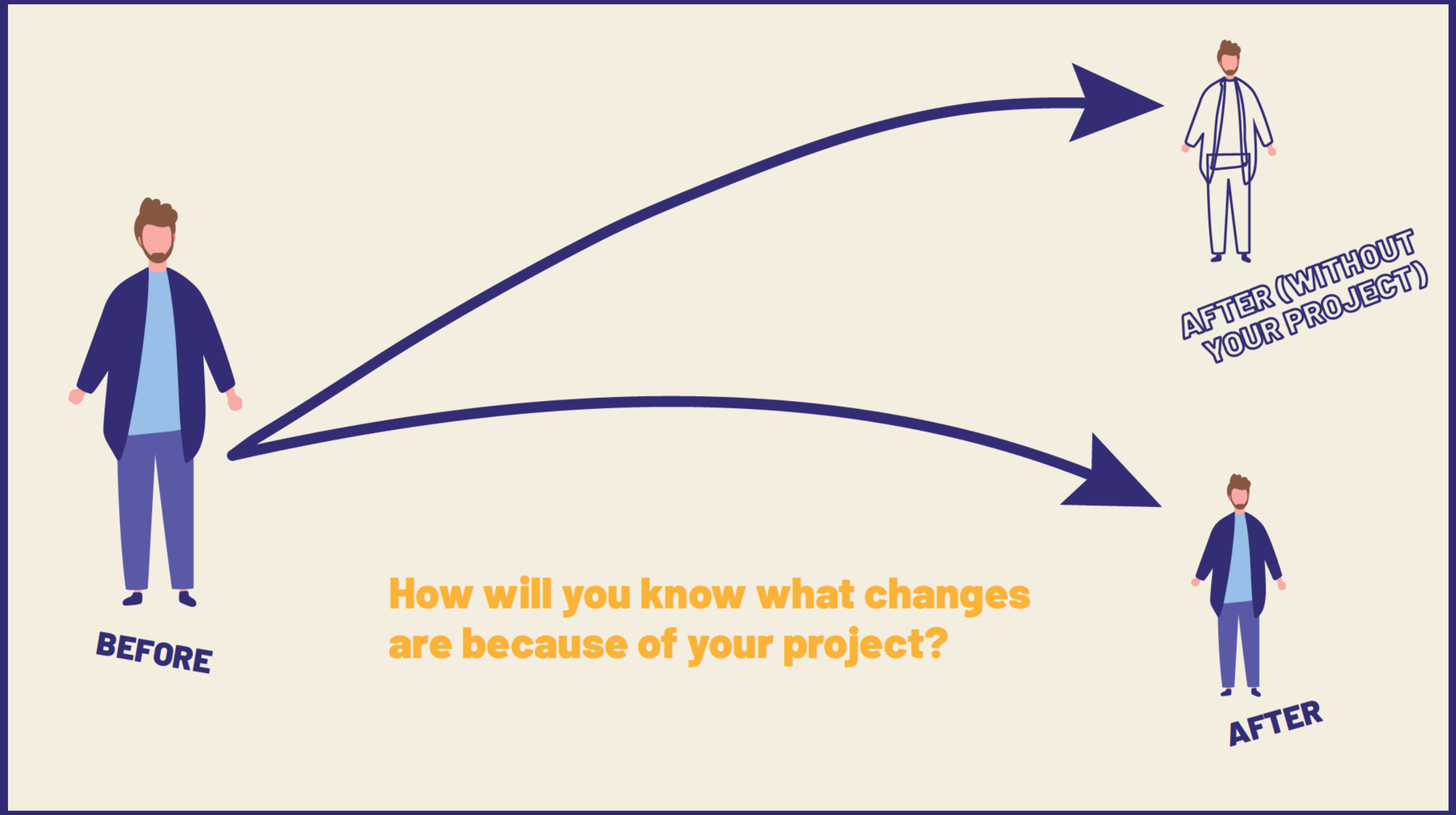
How will you know what changes are because of your project?

BEFORE

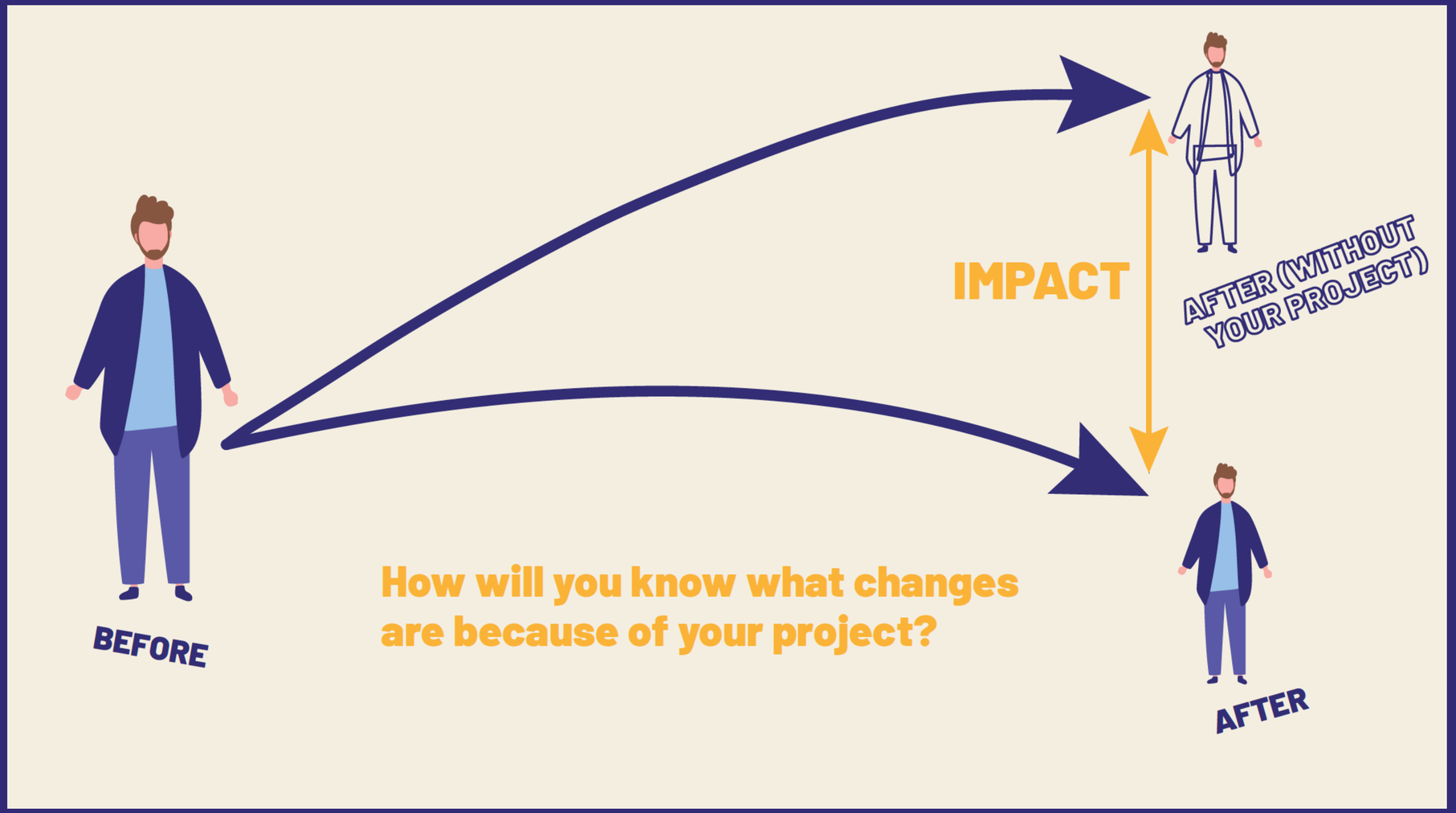
How will you know what changes are because of your project?

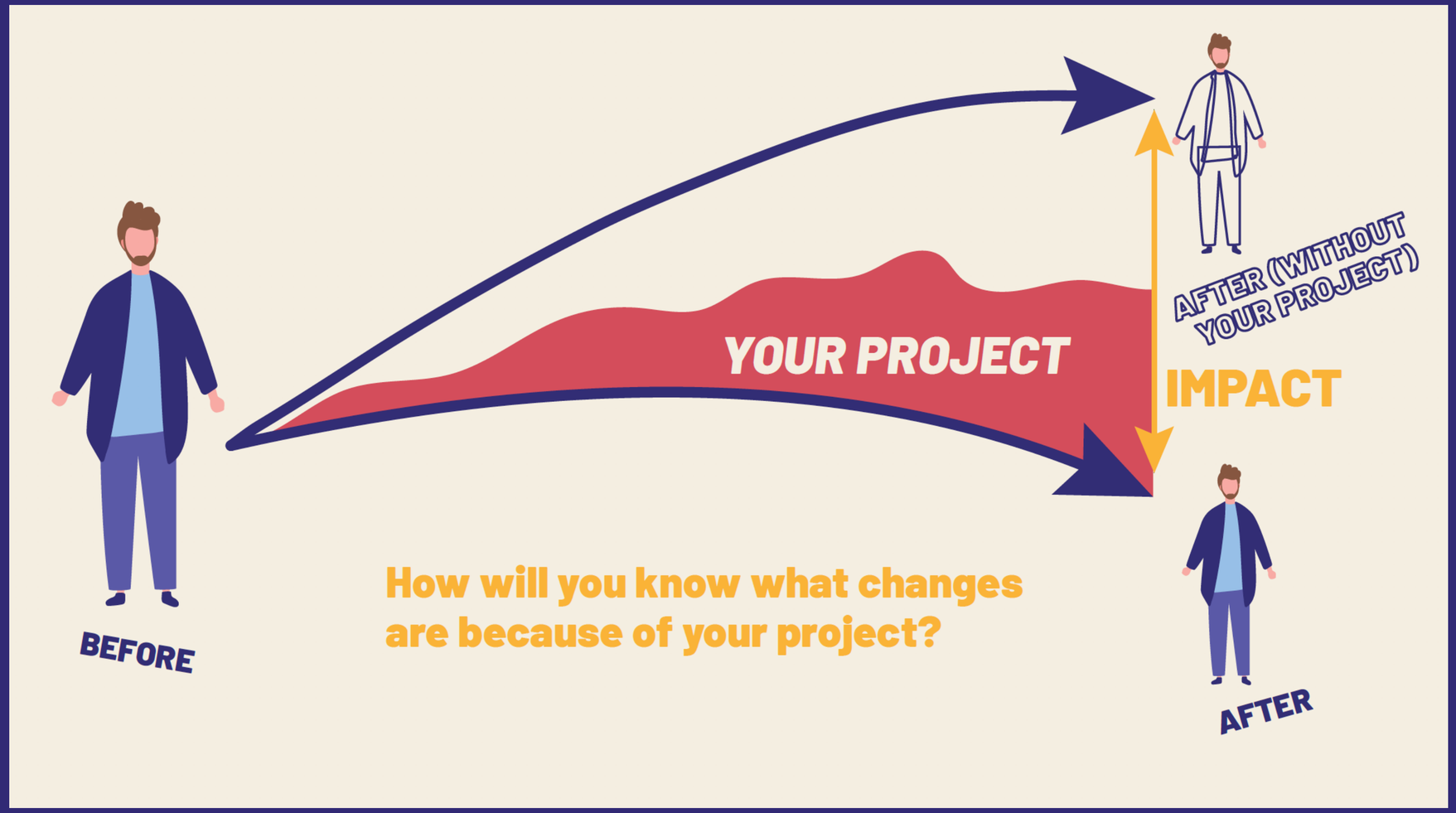


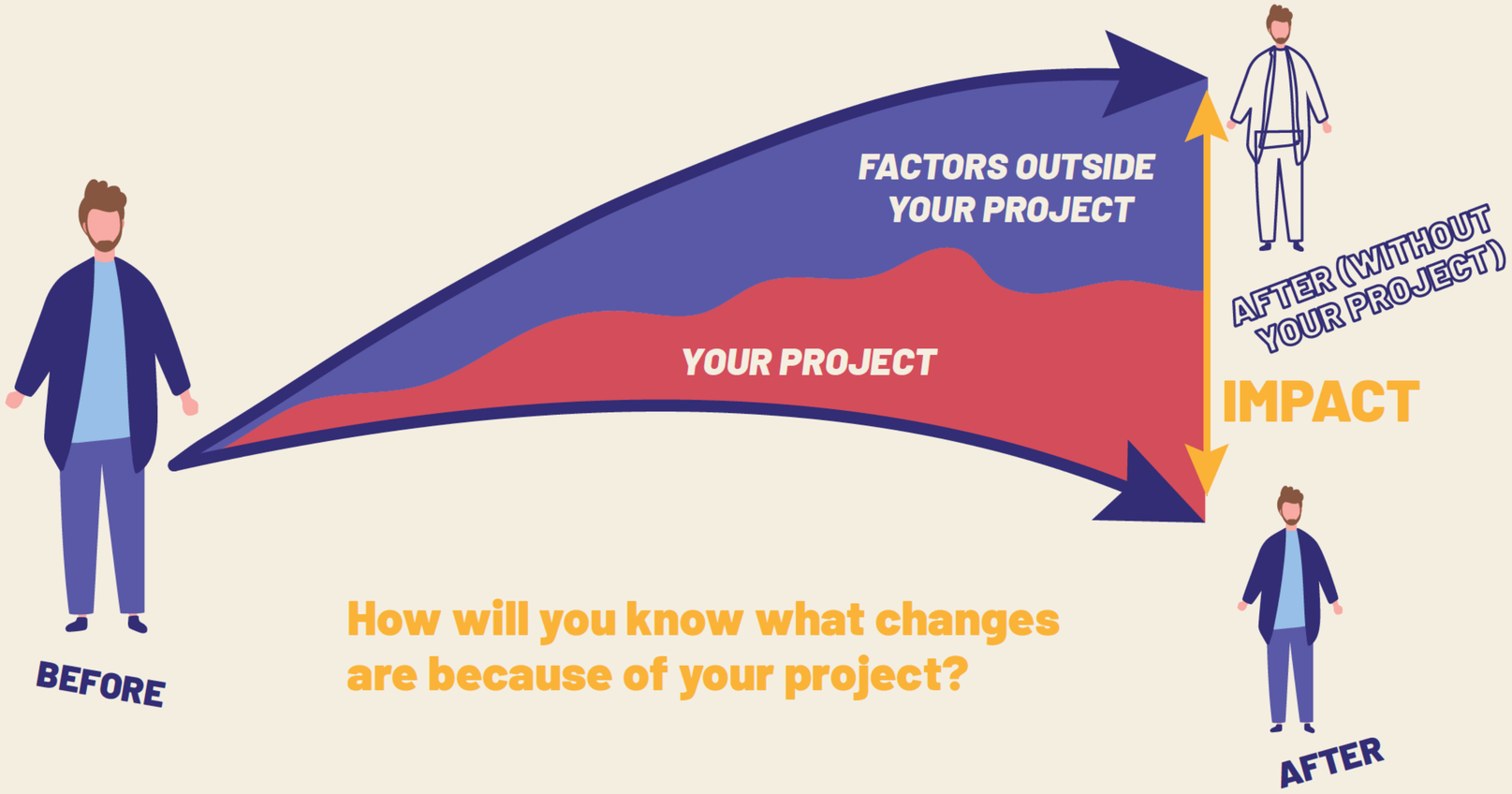
How will you know what changes are because of your project?



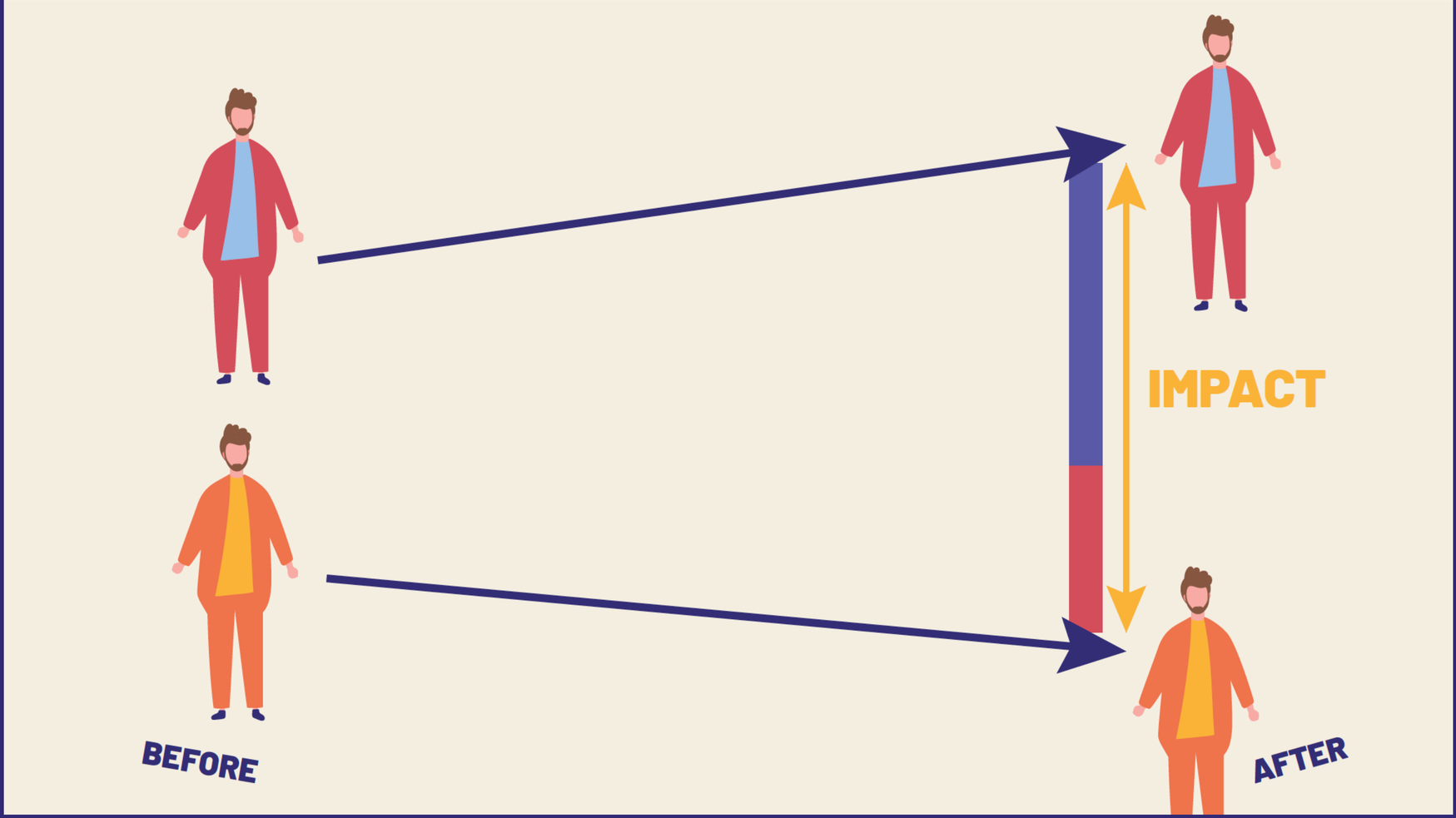
How will you know what changes are because of your project?

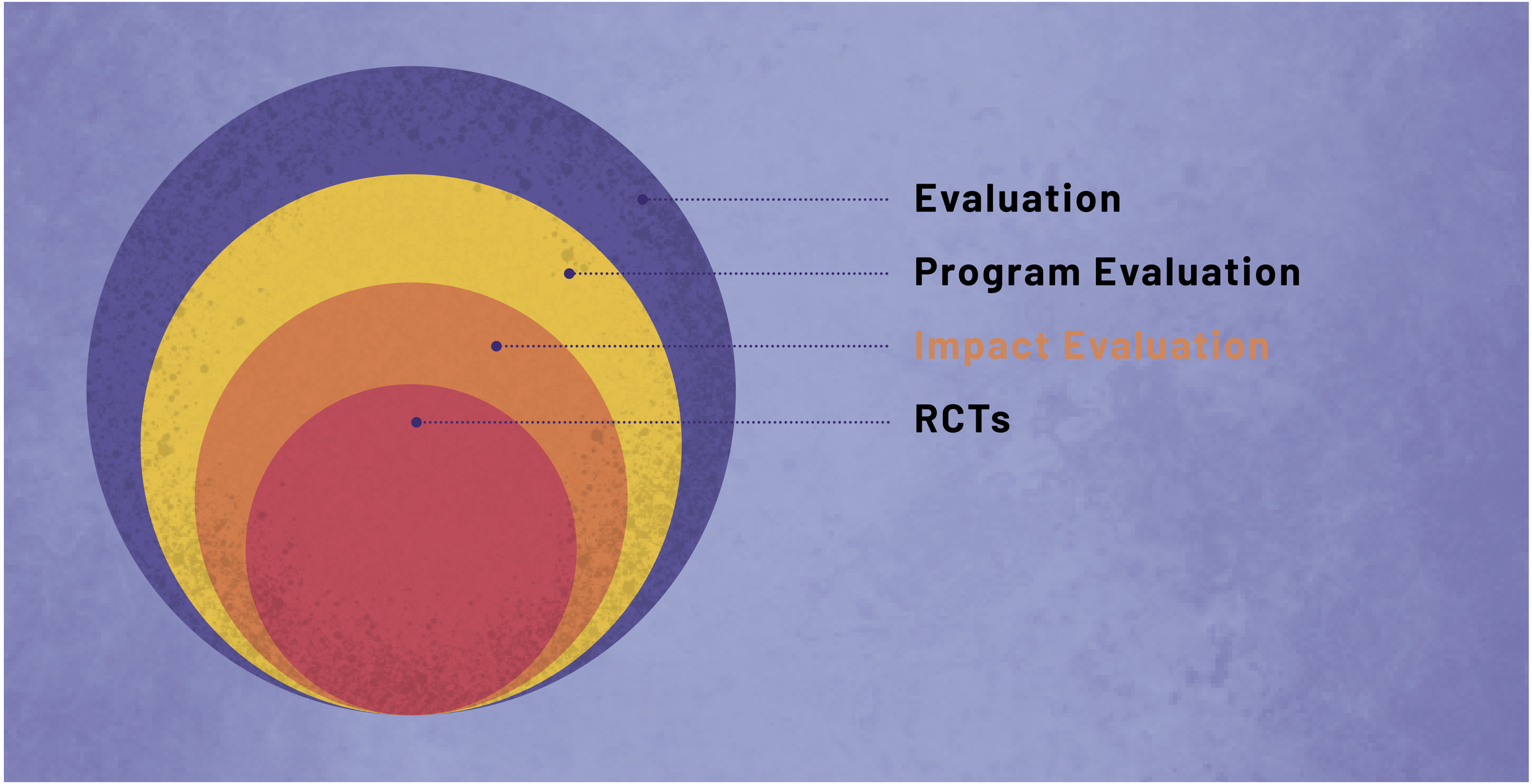

BEFORE
IMPACT
AFTER (WITHOUT YOUR PROJECT)
AFTER
How will you know what changes are because of your project?

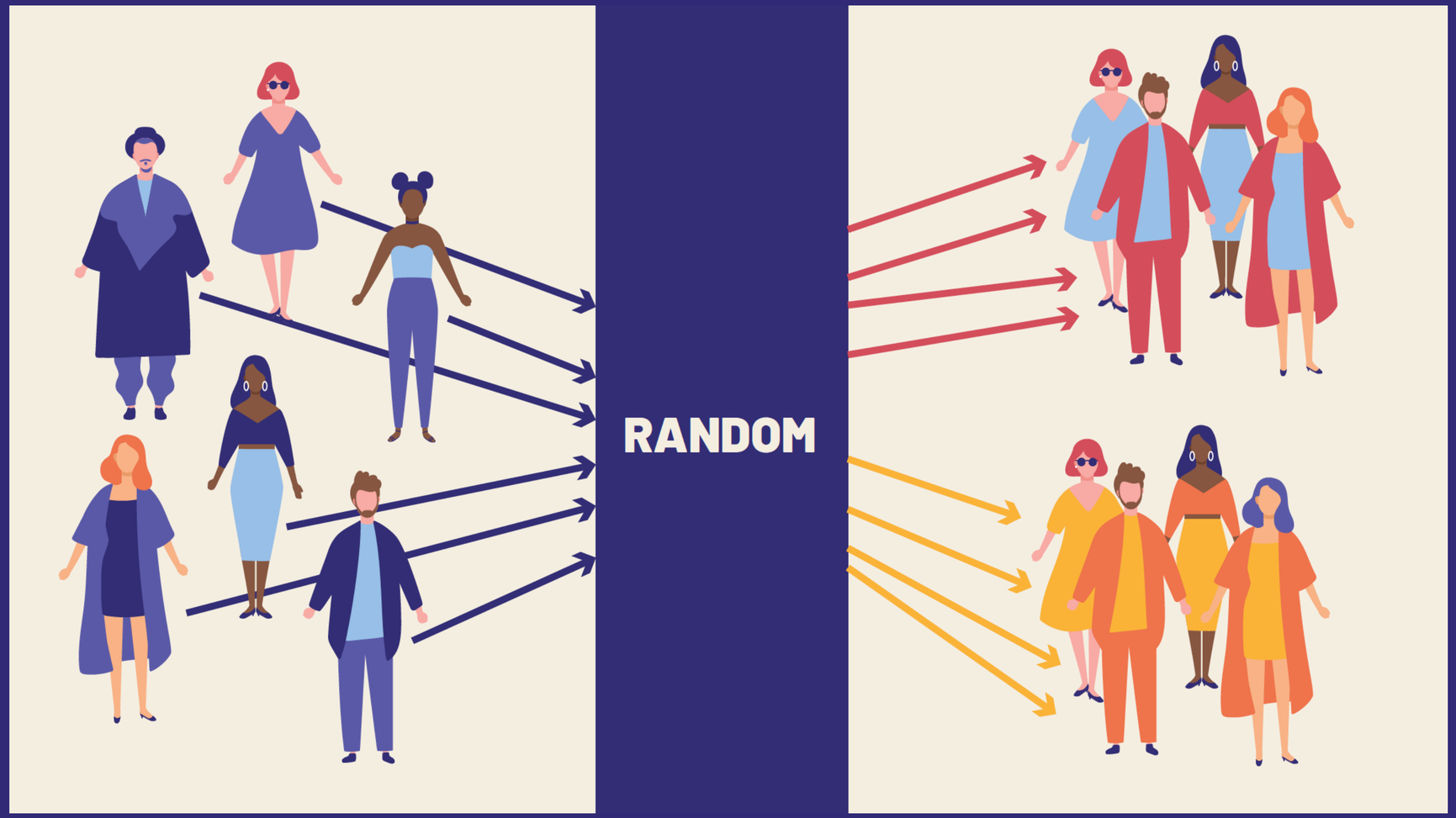

YOUR PROJECT
AFTER (WITHOUT YOUR PROJECT)
IMPACT
BEFORE
AFTER
How will you know what changes are because of your project?



How will you know what changes are because of your project?


IMPACT
BEFORE
AFTER


Evaluation
Program Evaluation
Impact Evaluation
RCTs


RANDOM

RCTs:
The Gold Standard
… of *What?*

Unbiased estimate of average treatment effect of the population.

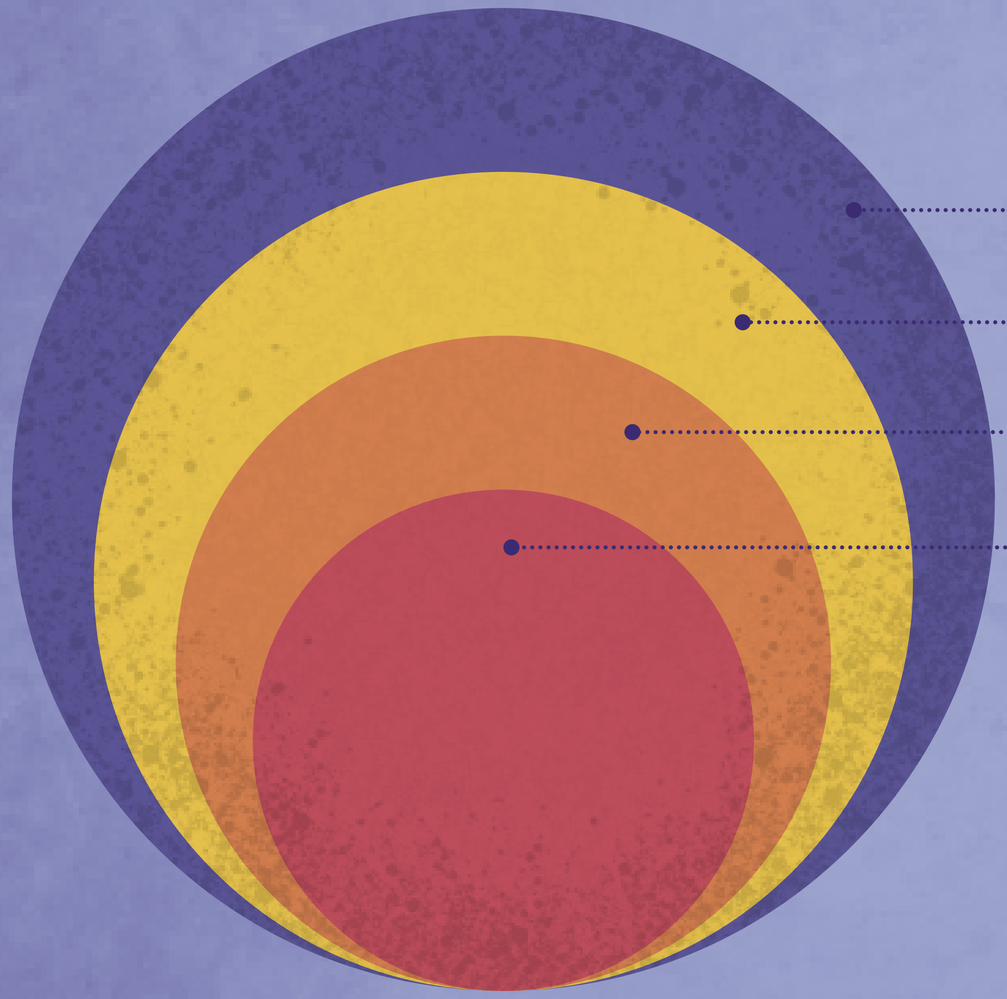
We found that the population average income increased $100*

---

*What we don’t find is that a small number of people had their income increase $1000 and a big group of people had their average income decrease by $100.

Questions that RCTs cannot answer:

- Will this scale?
- Who does this work for?
- Why does this work?
- How long does this work?
- In what context does this work?
- Where does this work?
- Cannot usually answer questions over time.



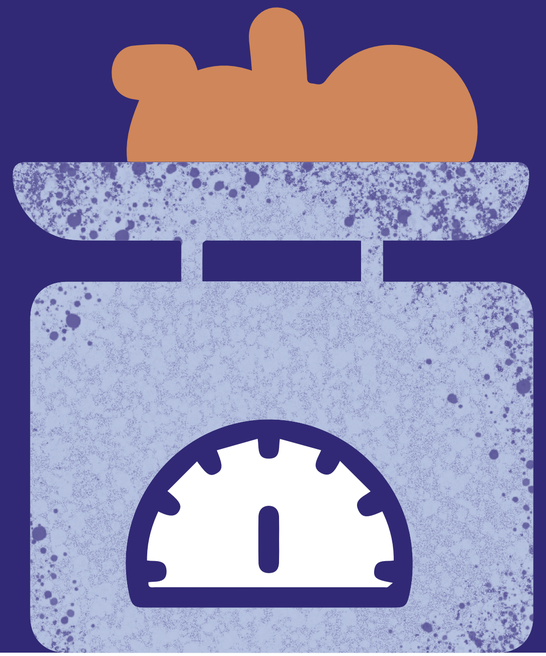
Evaluation

Program Evaluation

Impact Evaluation

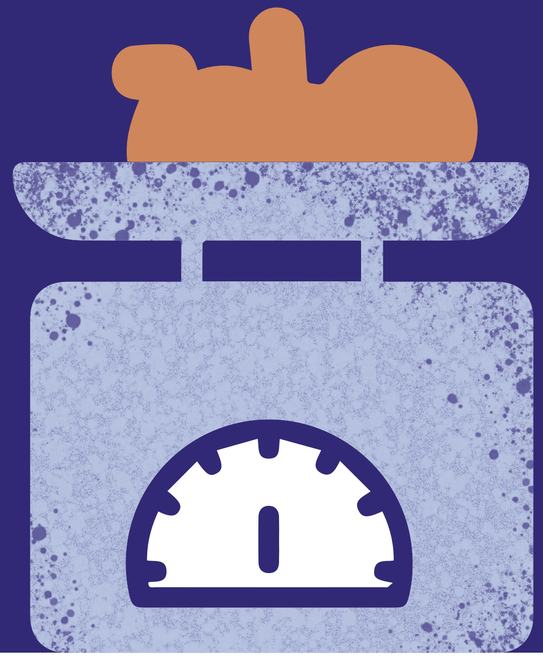
Longitudinal Analysis, DiDs, Matching, RCTs, Network Analysis, KIIs, Focus Groups, etc.

CONTEXTUAL VARIABLES



Chance of having a low birthweight baby is

20%



Chance of having a low birthweight baby

Ethnic Group A

Ethnic Group B

Chance of having a low birthweight baby when taking into account community of residence is between

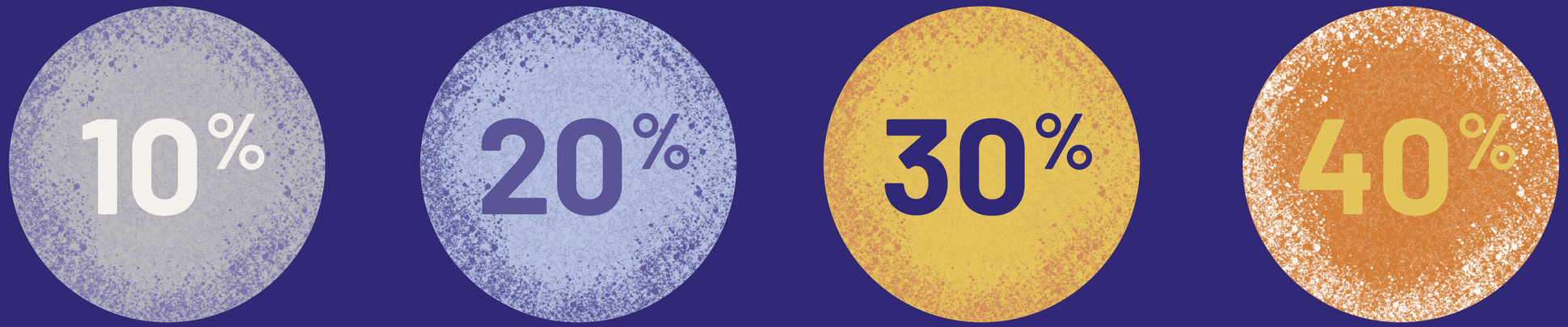
15% & 25%

Chance of having a low birthweight baby when taking into account community of residence and state of residence is between

10% & 30%

CONTEXTUAL VARIABLES

Amount related to:



Chance | Individual | State | Community

Chance of having a low birthweight baby when taking into consideration community and state

Ethnic Group A: 15% & 20%

Ethnic Group B: 15% & 21%

AMOUNT RELATED TO:

Ethnicity	Chance	Individual	State	Community
0%	10%	20%	30%	40%

Contextual Variables

BIRTH WEIGHT =

Person

Person+Ethnicity

Person+Ethnicity+Community

Person+Ethnicity+Community+State

Person+ Ethnicity+Community*State+C(Community)+C(State)

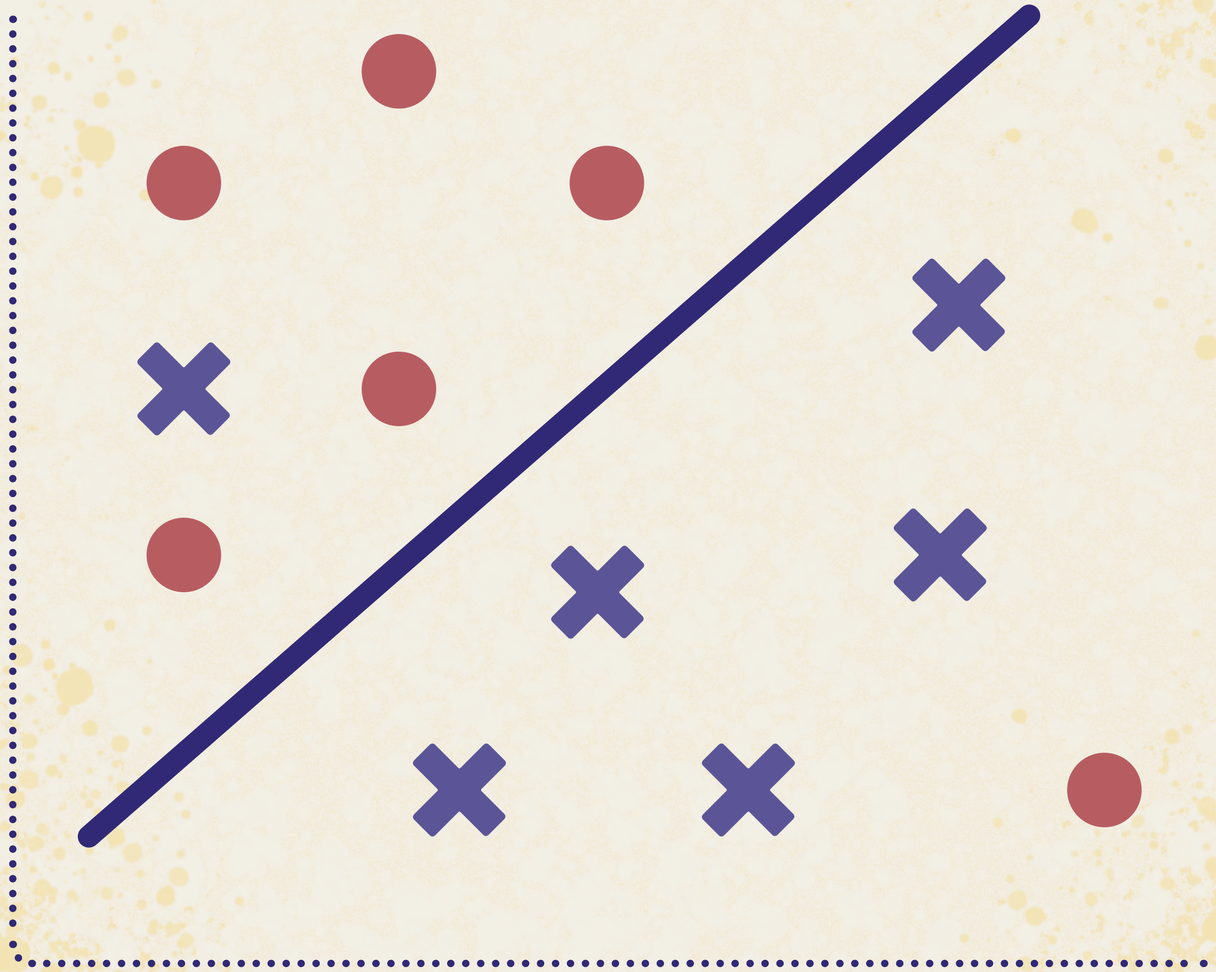
What is Bias?

What is Bias?

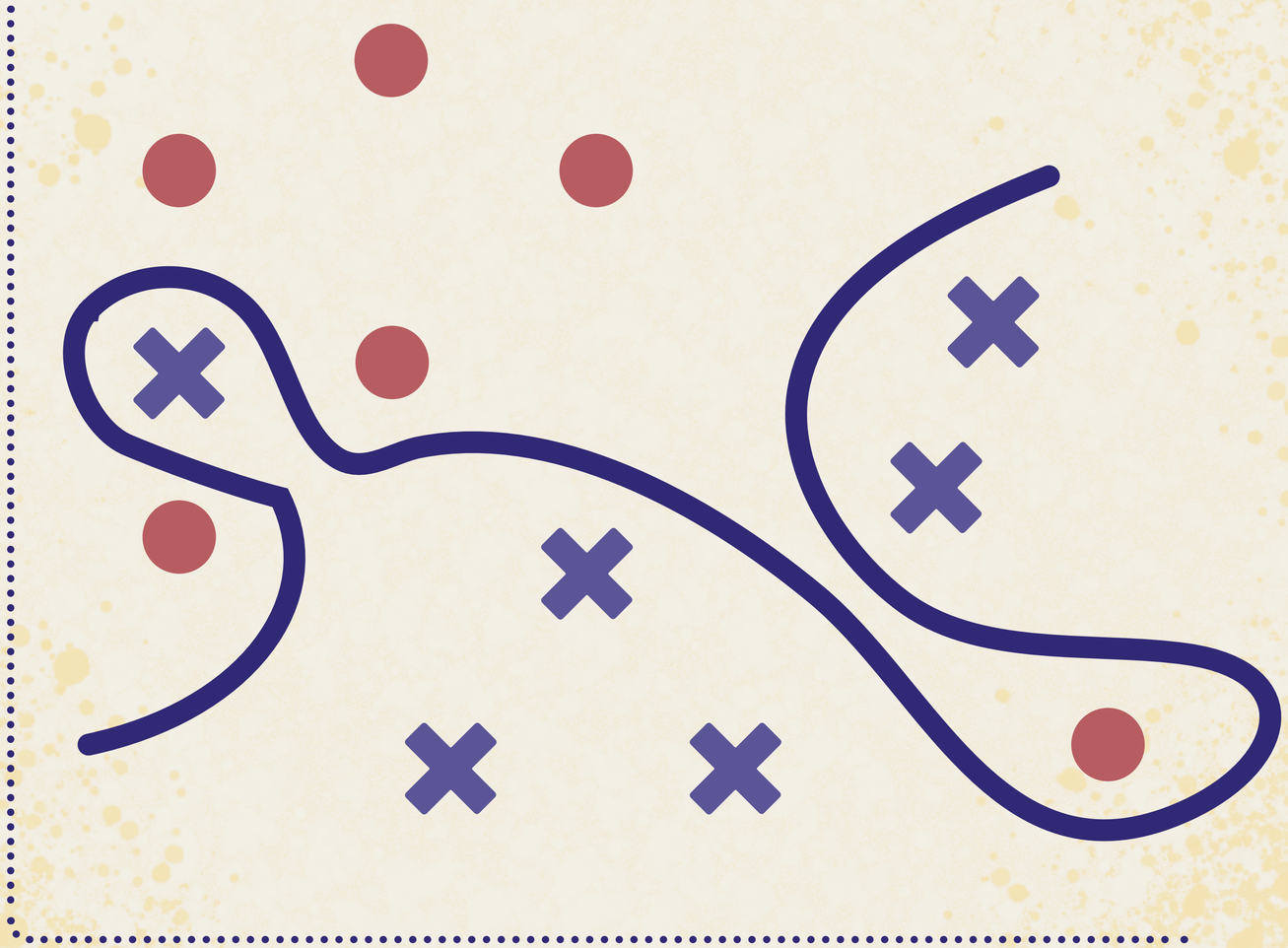
Systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others.

Conscious. Unconscious.
Intentional. Accidental.
Usually a combination.

High Bias



High Variance



What is Metadata?

What is a Data Biography?

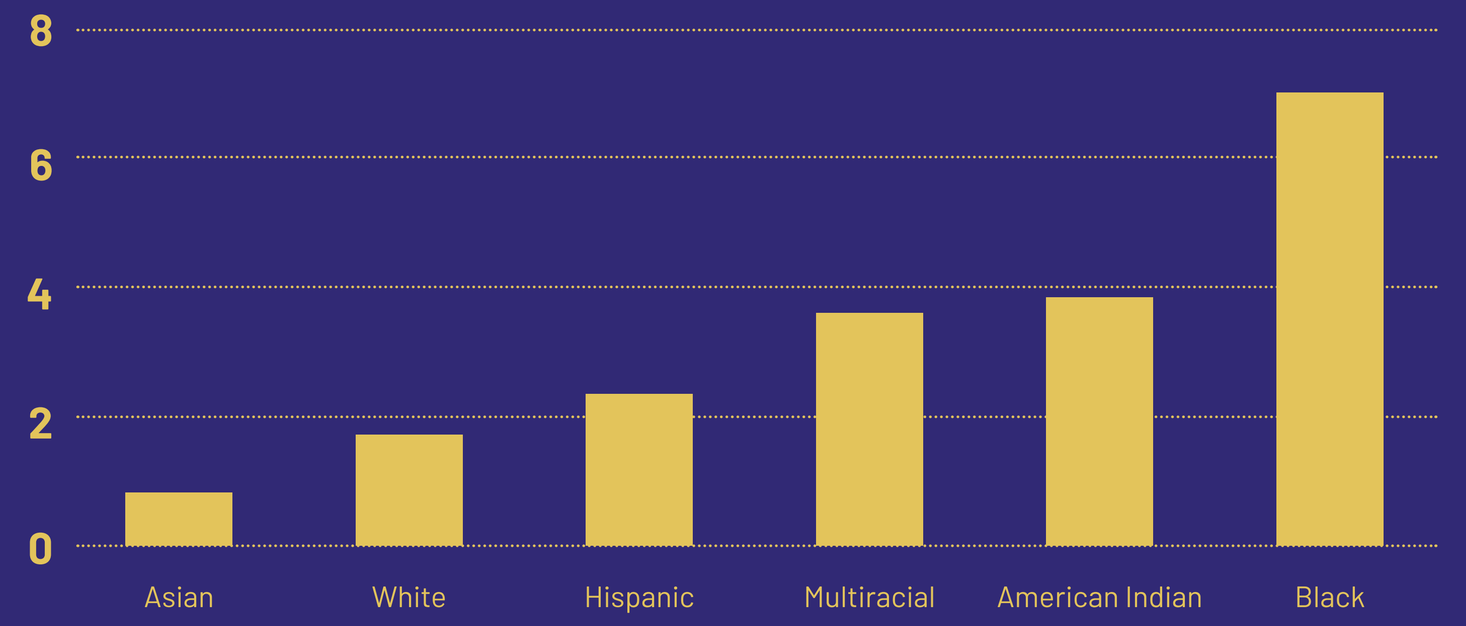
Data Biographies at the bare minimum must accompany each dataset you are using:

When What Who Why How Where

What is Fair?
What is Algorithmic Accountability

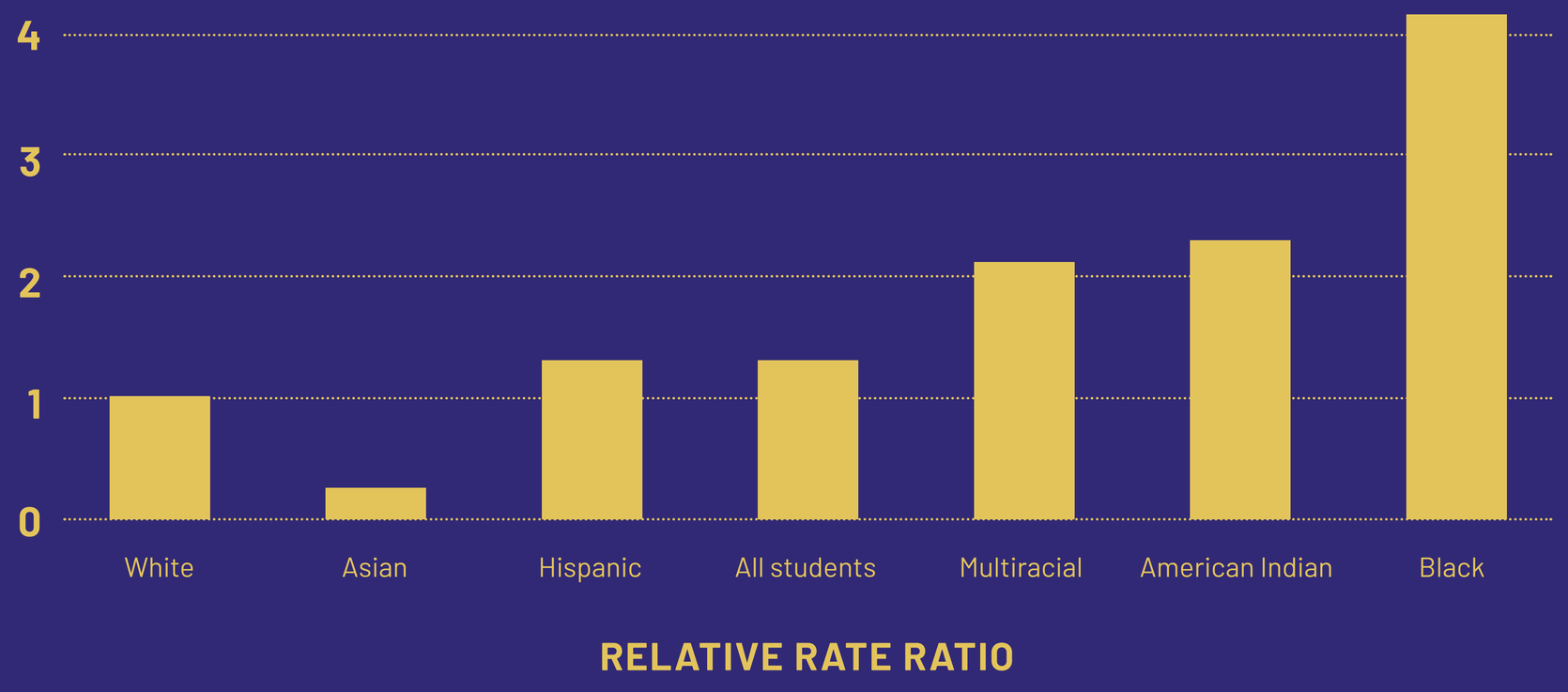
	Student Group 1	Student Group 2	Student Group 3
Number of Students	10	20	30
Count of Students Disciplined	7	5	15
Rate	75/100	25/100	50/100
Rate Relative to Student Group 2	3.0	1.0	2.0
Composition Index	27.3	18.2	54.5
Composition of Enrollment	16.7	33.3	50.0
Difference in Composition (Percentage Points)	10.6	-15.2	4.5
Relative Difference in Composition of Students Disciplines and Enrollment	63.6	-45.5	9.1

Rate of students who experienced one suspension or more, by racial/ethnic group

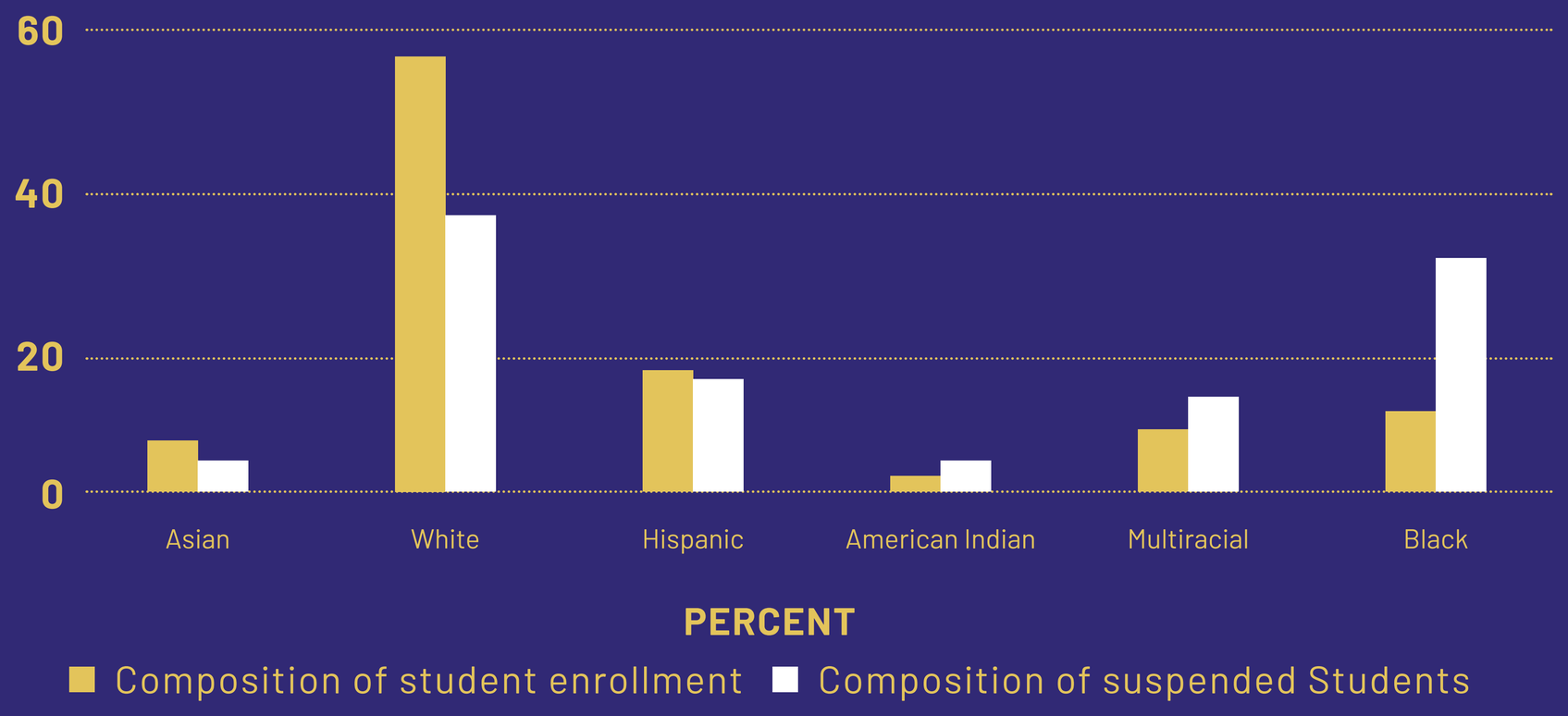


PERCENT OF STUDENTS WHO EXPERIENCED ONE OR MORE SUSPENSIONS

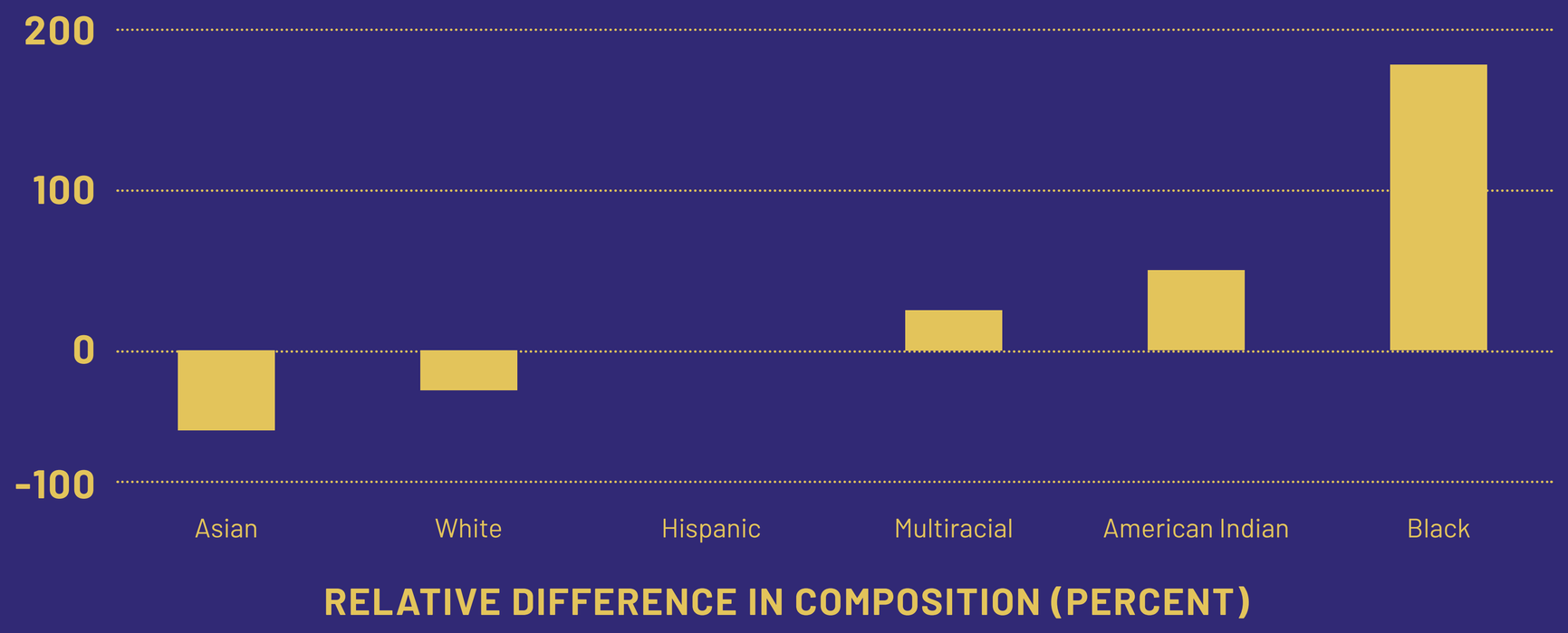
Relative rate ratios comparing the rates of students who experienced one suspension or more in specific racial/ethnic groups with the rate among White students who experienced one suspension or more.

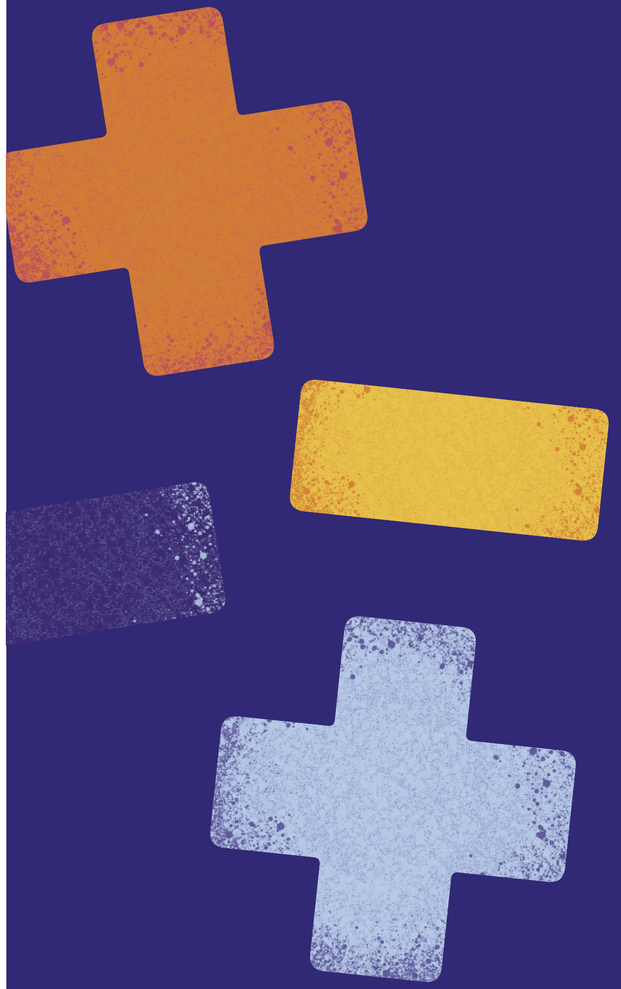


Comparison of two compositions: Proportion of the student group who were suspended and proportion of the group in the student population, by racial/ethnic group.



The relative difference in composition between the proportion of students who experienced suspensions in each racial/ethnic group and proportion of the group in the total student population.





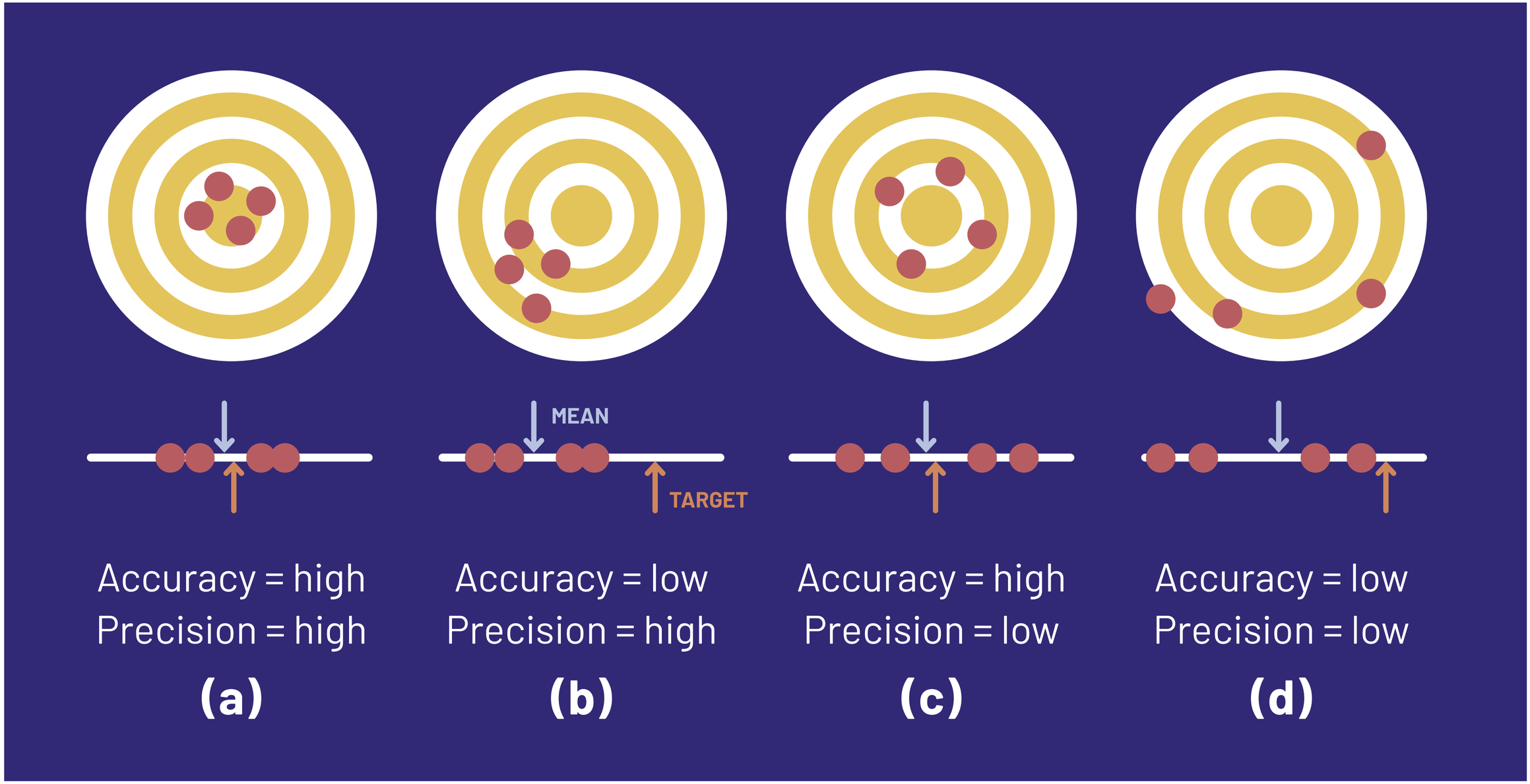
Ways to think about fair

Equal False Negative Rates: the fraction of positives which are marked negative in each group agree.

Equal False Positive Rates: the fraction of negatives which are marked positive in each group agree.

Equal Positive Predictive Values: the fraction of those marked positive which are actually positive in each group agree.

Statistical Parity (equal positive decision rates): the fraction marked positive in each group should agree.


MEAN
TARGET
Accuracy = high
Precision = high
(a)
Accuracy = low
Precision = high
(b)
Accuracy = high
Precision = low
(c)
Accuracy = low
Precision = low
(d)

What is the Margin of Error?

How randomly your survey respondents were chosen

How similar your sample is compared to your population (This one is usually called the "Margin of Error"

Survey Results

How well the results of your survey reflect the ground truth

Ground Truth

How many missing responses there e in your sample

How well your questions are phrased to get accurate answers

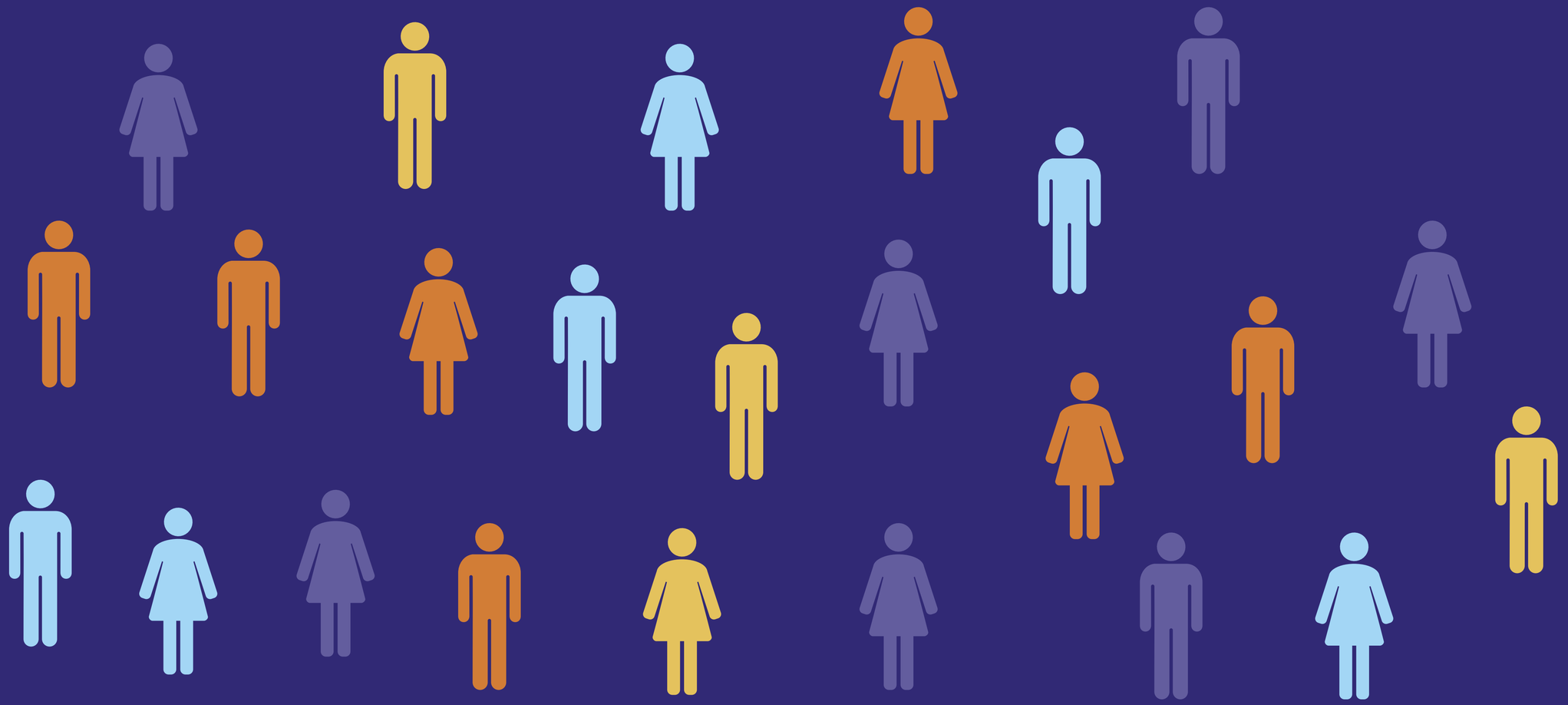
The type of data analysis you used to calculate your results.

How much we can depend on our estimate will depend on:

What question we actually ask.
How we measure.
Who we ask.
How many people we ask.
How good we are at measuring.
How we take into account context.
How good we are at collecting data.
What types of analysis we use.

What is a Sample?

Here is the community you're interested in talking about.

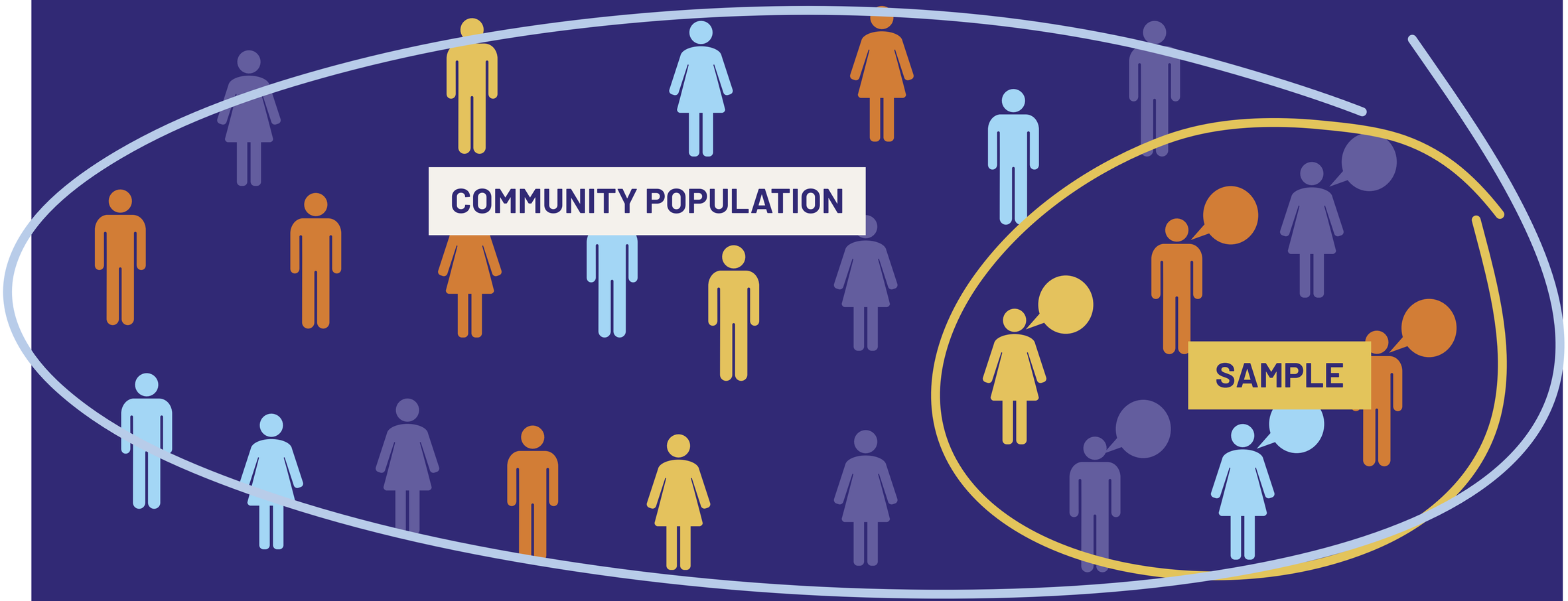


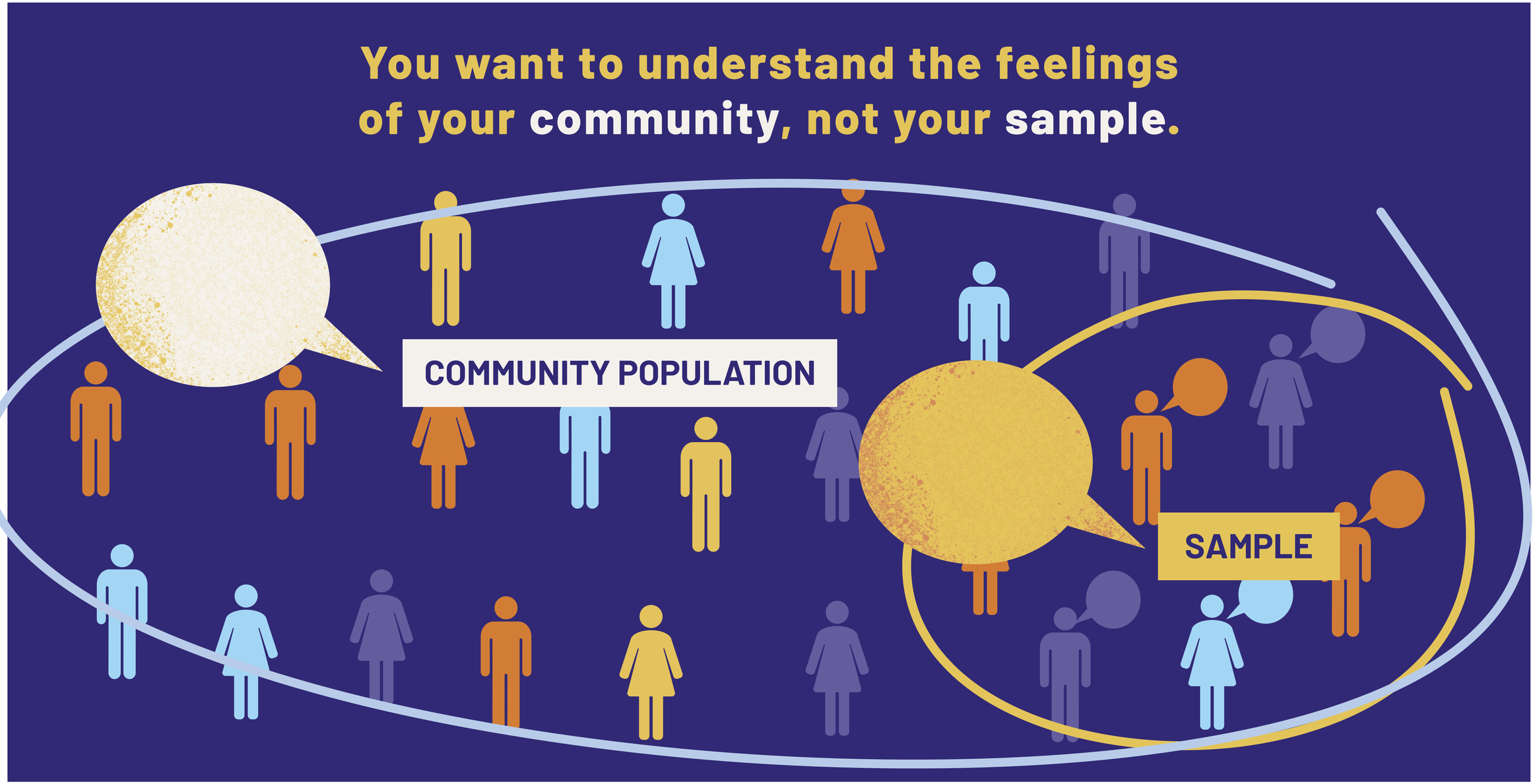
Don’t have the time, money, capacity to ask every single one of these people their opinion.

So you take a survey.

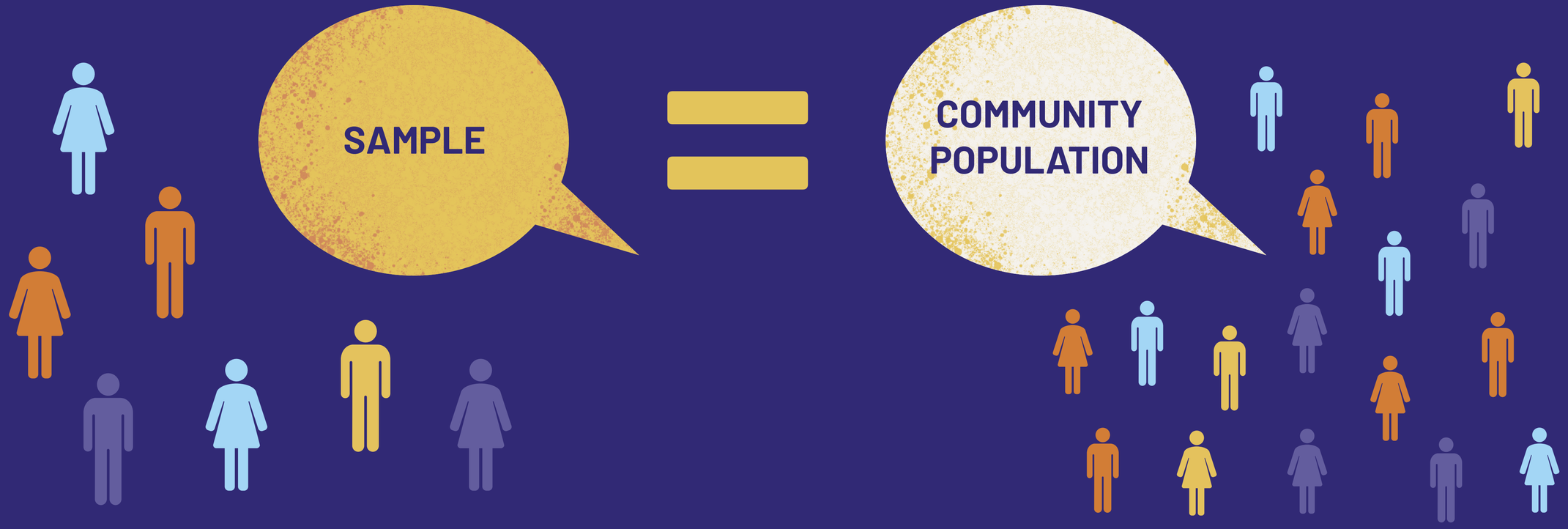
The group of people who answer that survey are called your sample.

The entire group of people is called the community population.

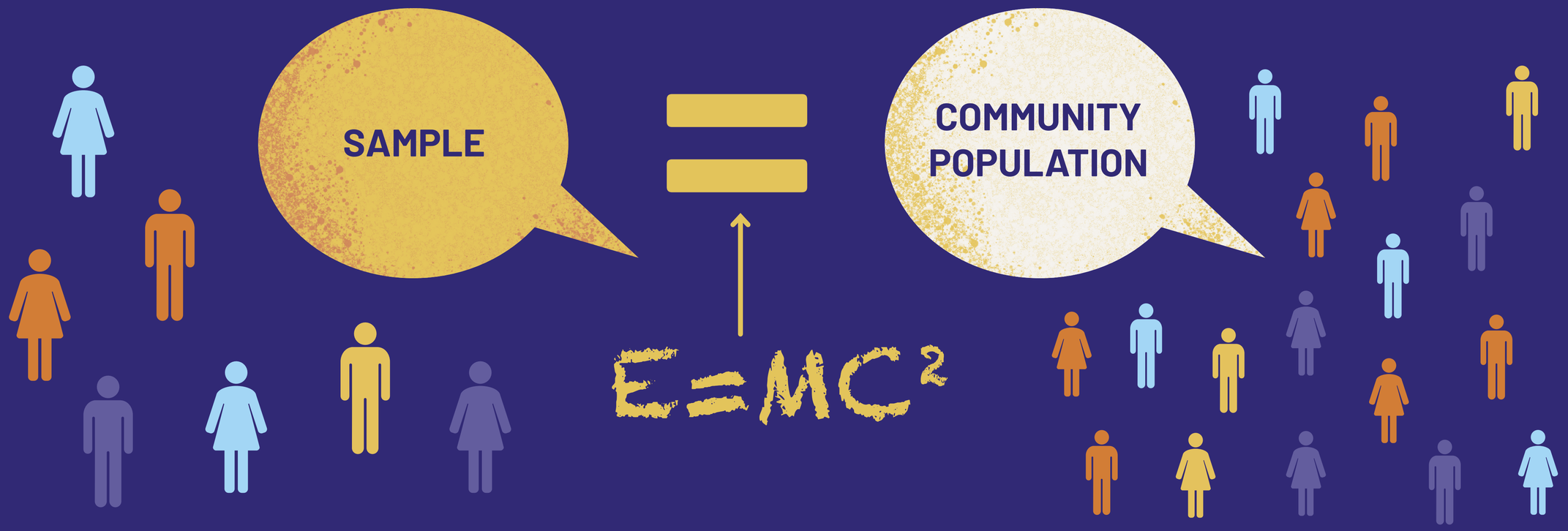



You want to understand the feelings of your community, not your sample.
COMMUNITY POPULATION
SAMPLE

If you've collected your sample in a very fancy scientific way – random sampling – you can assume that the feelings of your sample reflect the feelings of your community.

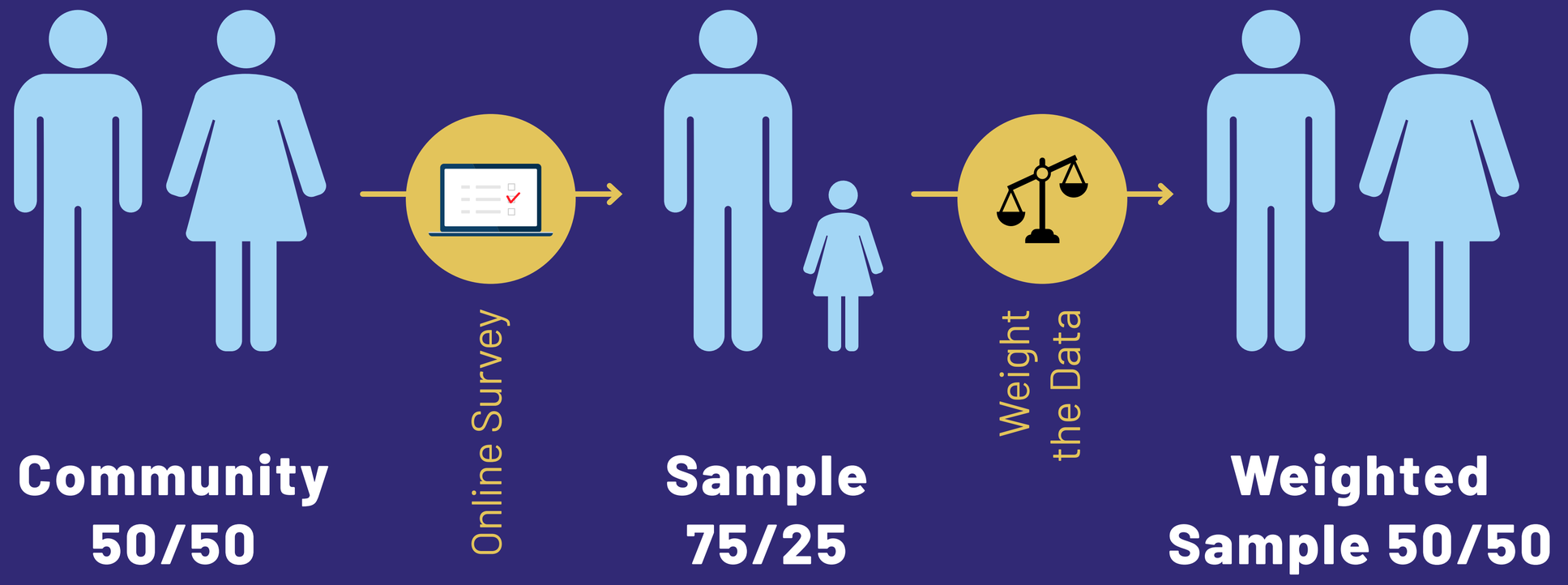


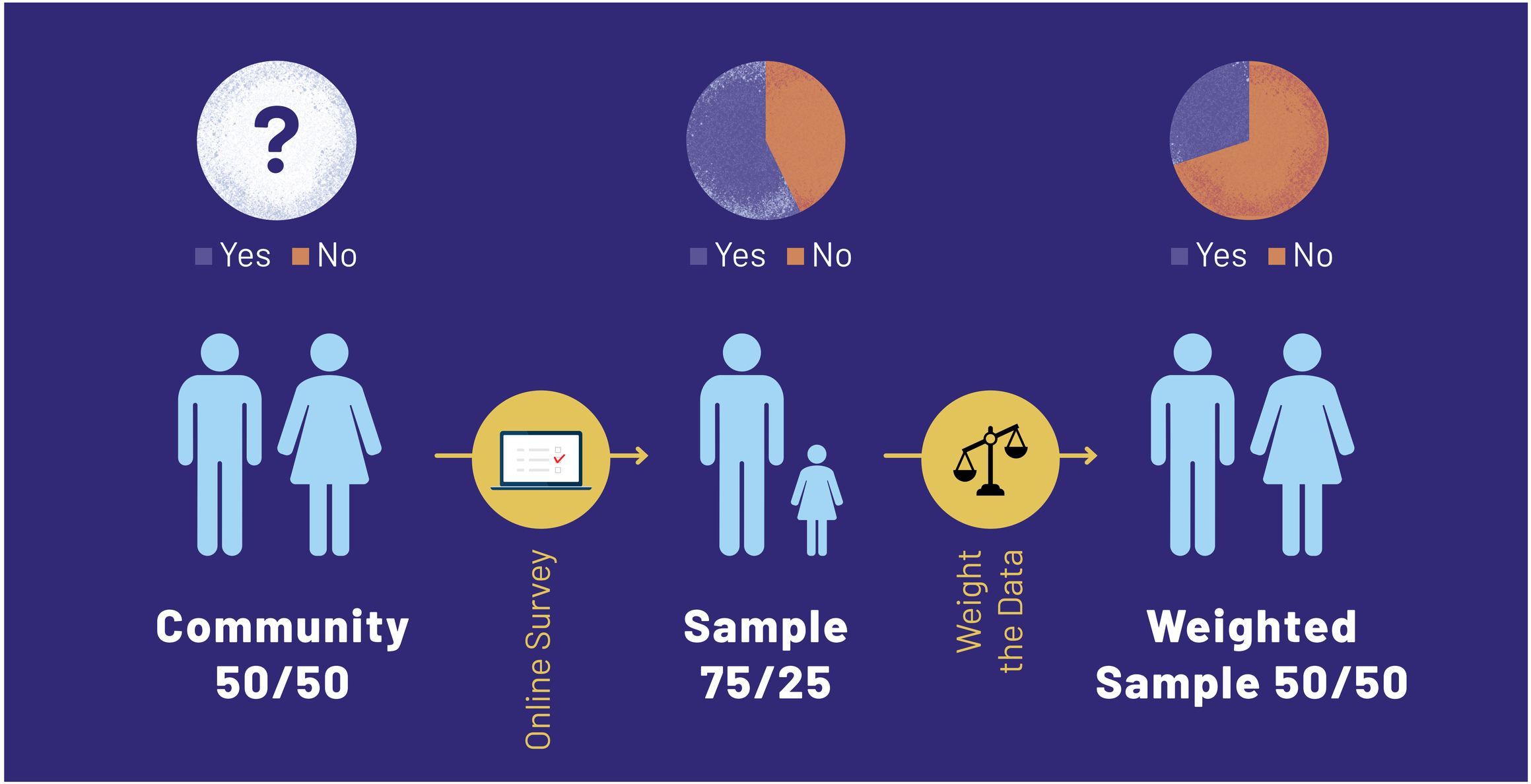
You need to do some fancy math to get the feelings of your sample to more accurately reflect the feelings of your community.

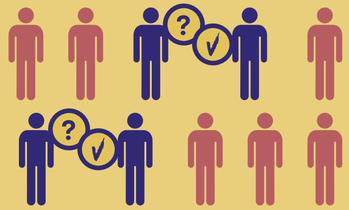


So for example if your community is 50/50 men and women, and your sample is 25% women and 75% men.

So for example if your community is 50/50 men and women, and your sample is 25% women and 75% men.




Yes
No
Yes
No
Yes
No
Community
50/50
Online Survey
Sample
75/25
Weight
the Data
Weighted
Sample 50/50

Random Sampling 	Every member of a population has an equal chance of being selected *E.g. pulling names out of a hat*	For very large samples it provides the best chance of an unbiased representative sample	For large populations it is time-consuming to create a list of every individual
Stratified Sampling 	Dividing the target population into important subcategories Selecting members in proportion that they occur in the population *E.g. 2.5% of British are of Indian origin, so 2.5% of your sample should be of Indian origin...and so on*	A deliberate effort is made to make the sample representative of the target population.	It can be time consuming as the subcategories have to be identified and proportions calculated
Volunteer Sampling 	Individuals who have chosen to be involved in a study, also called self-selecting *E.g. people who have responded to an advert for participants*	Relatively convenient and ethical if it leads to informed consent	Unrepresentative as it leads to bias on the part of the participant *E.g. a daytime TV advert would not attract full-time workers*
Opportunity Sampling 	Simply selecting those people that are avilable at the time. *E.g. going up to people in cafés and asking them to be interviewed*	Quick, convenient and economical. A most common type of sampling in practice.	Very unrepresentative samples and often biased by the researchers who will likely choose people who are 'helpful'

What's an Independent Variable? or a Dependent one?

Independent Variables

Student's Characteristics

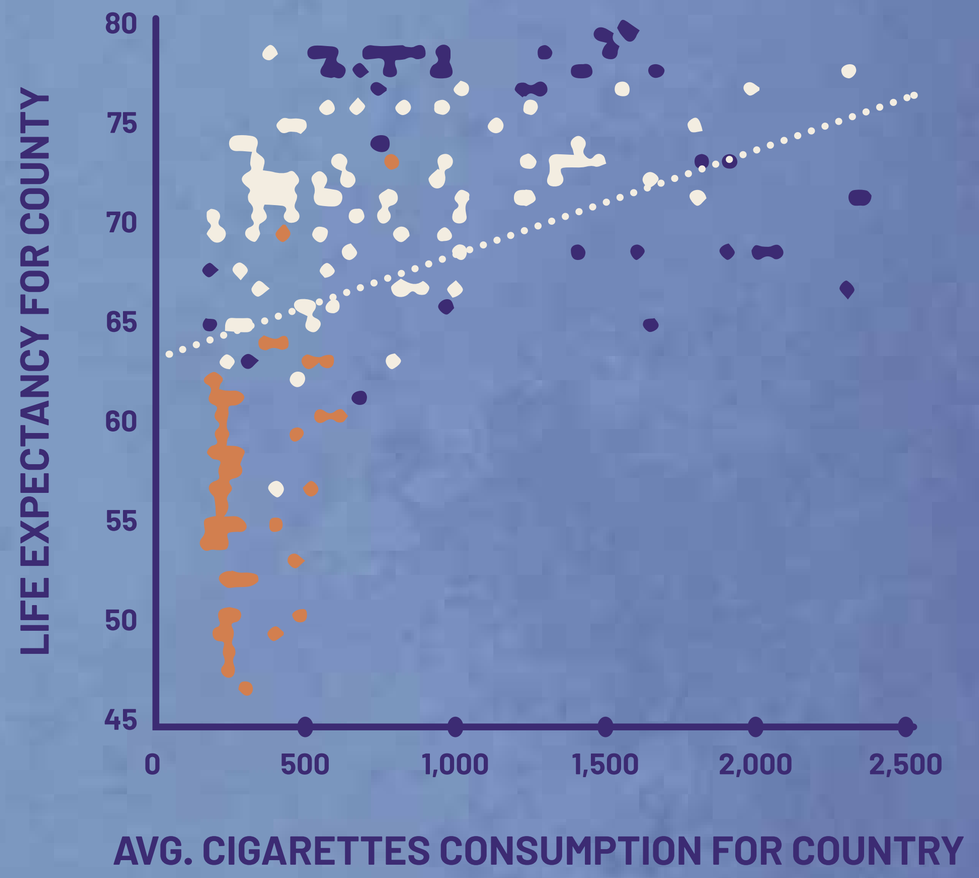
- age
- gender
- relationship with mother
- relationship with father

Dedependent Variables

Total Number of Hours Spent Online

What’s the Ecological Fallacy?

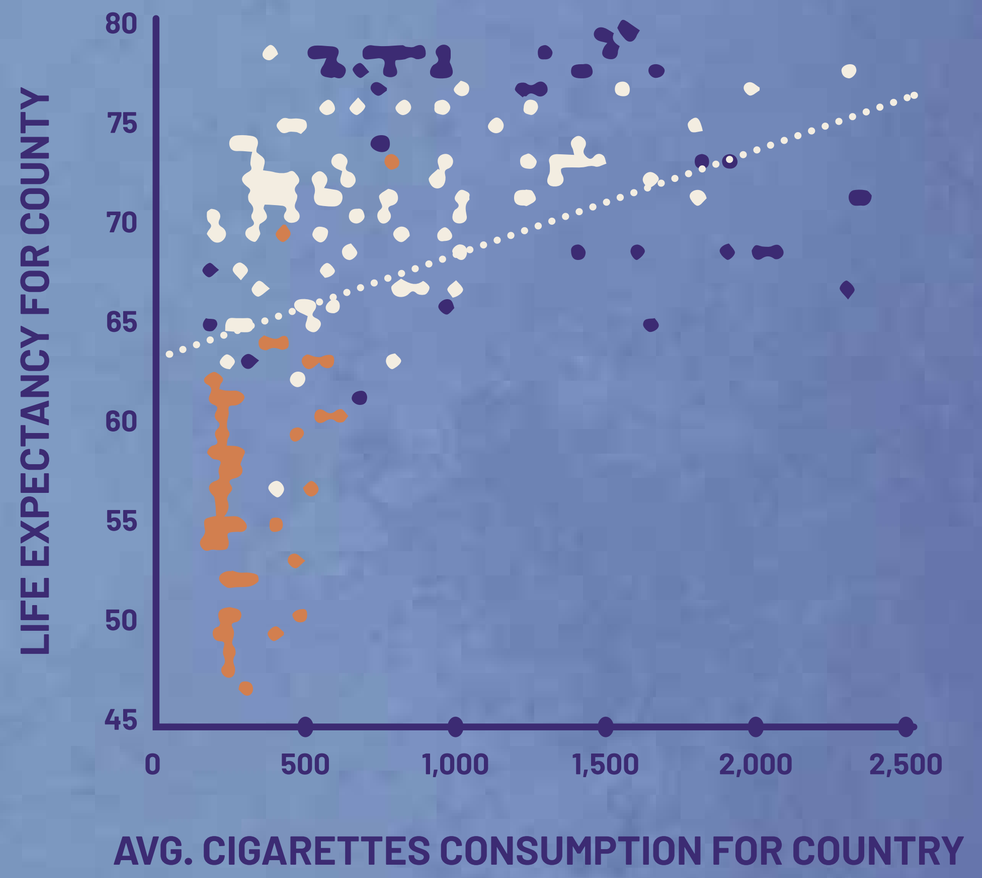
Is smoking cigarettes
good for your health?



Is smoking cigarettes **good for your health?**

STATISTICAL SUMMARY

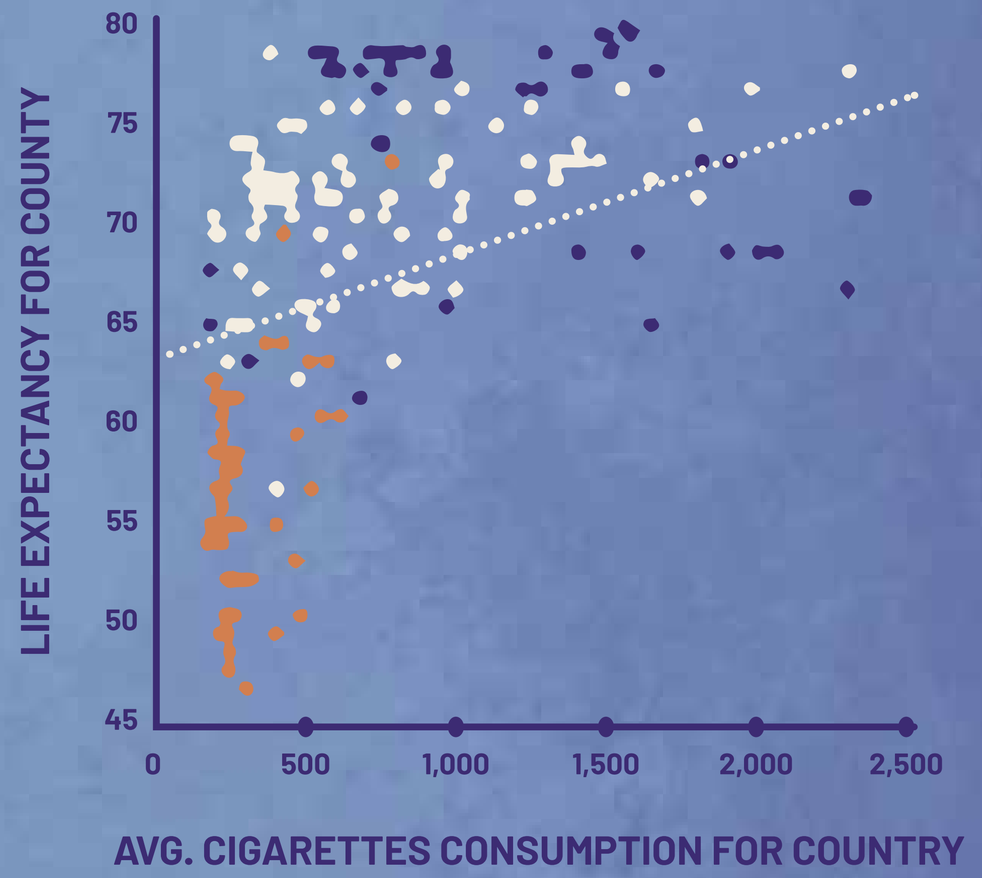
Coefficients	Estimate	Std. Error	DF	t-value	p-value
(Intercept)	65.07	0.85	183	76.02	<.00001
Cigarettes	**0.006**	0.0008	183	8.09	**<.00001**



Is smoking cigarettes **good for your health?**

One extra cigarette per year adds 0.006915 years to your life.

So 4 cigarettes a day will add 10 years to your life.



Is smoking cigarettes **good for your health?**

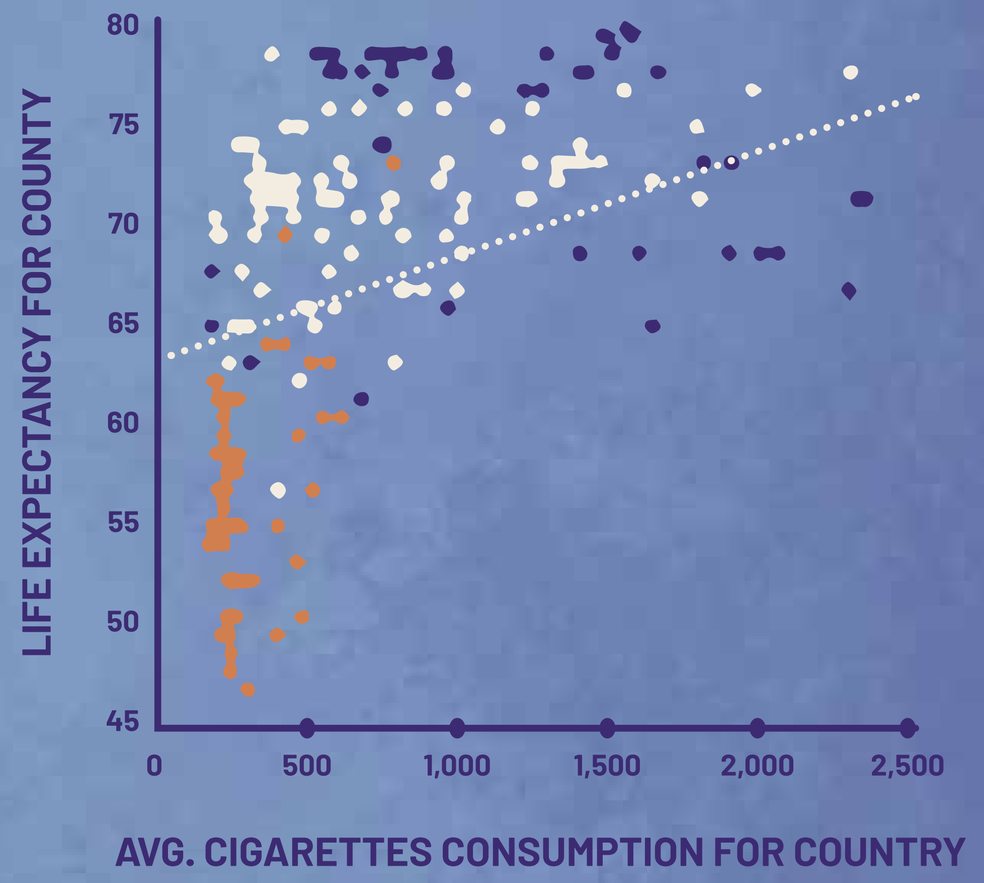
STATISTICAL SUMMARY

Coefficients	Estimate	Std. Error	DF	t-value	p-value
(Intercept)	65.07	0.85	183	76.02	<.00001
Cigarettes	**0.006**	0.0008	183	8.09	**<.00001**

So what's the problem?

No problem in the math or data viz.

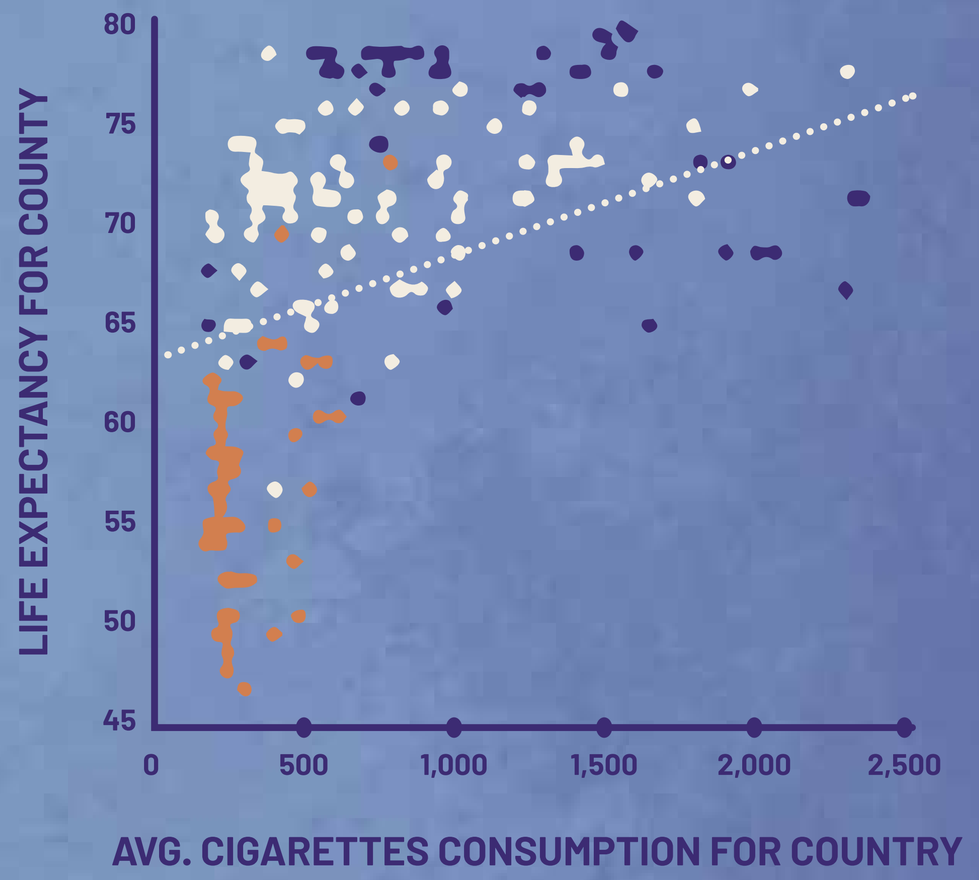
Problem is with the title.



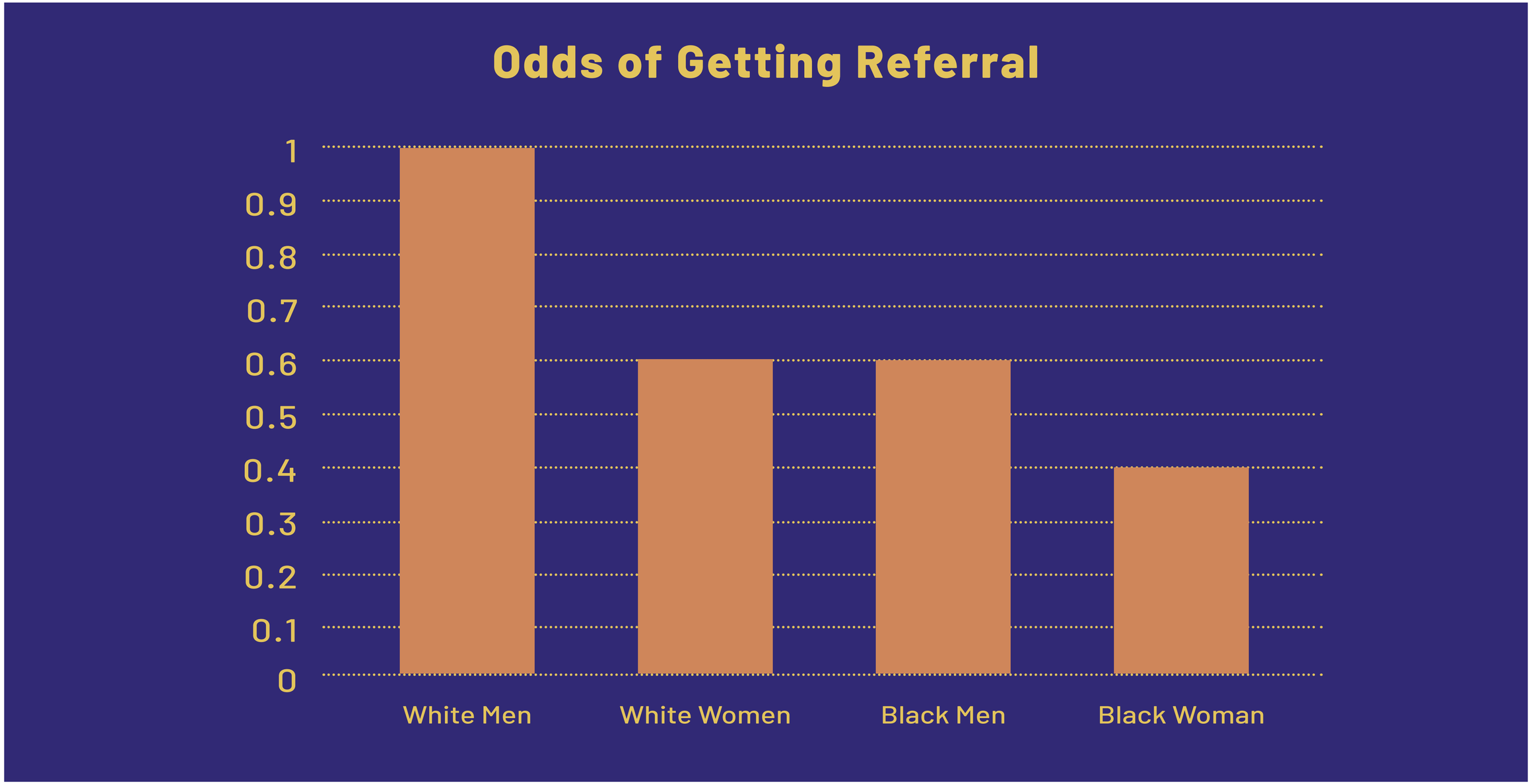
~~Is smoking cigarettes~~
~~**good for your health?**~~

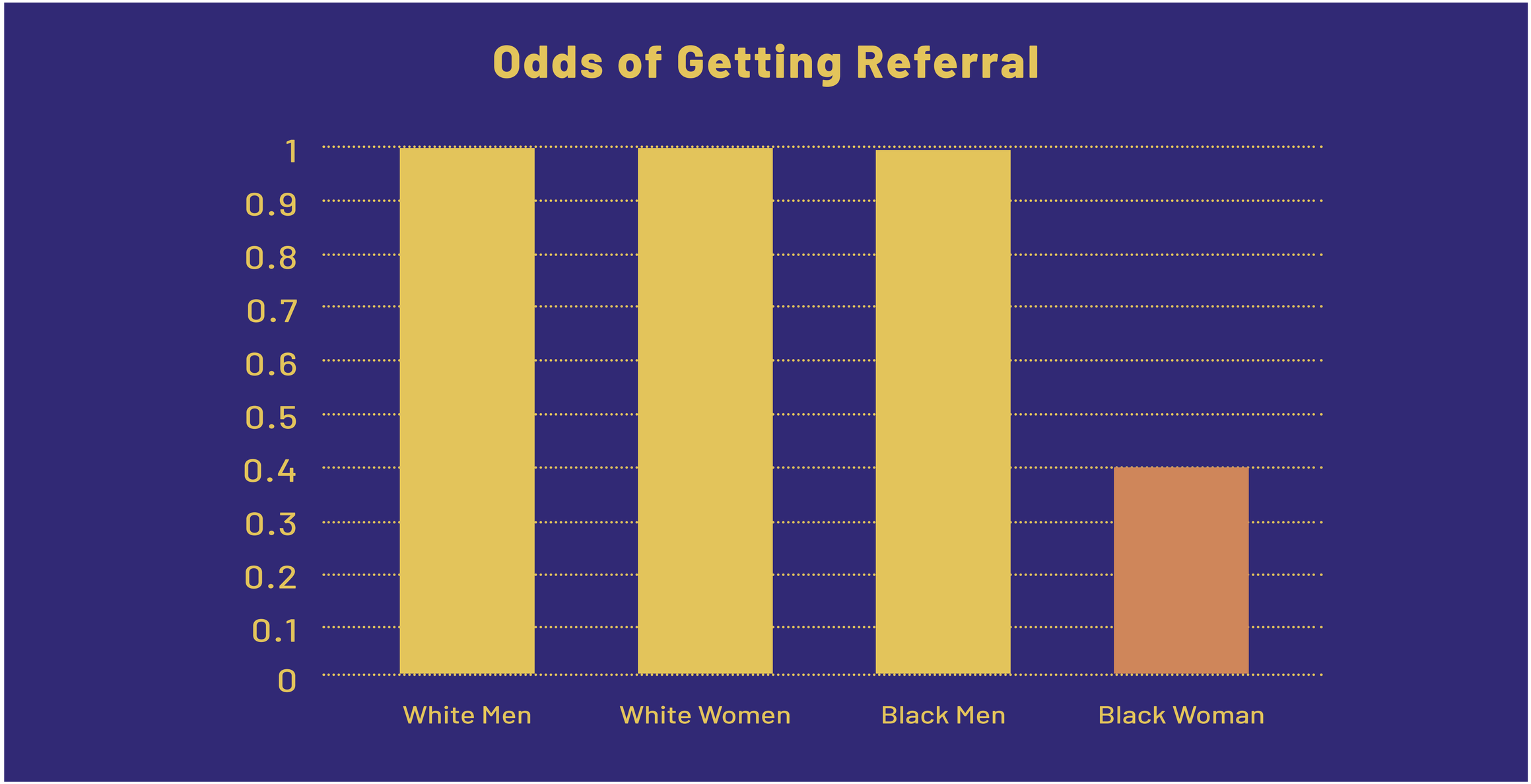
Do countries with higher average cigarette consumption **have longer life expectancies?**

Yes.

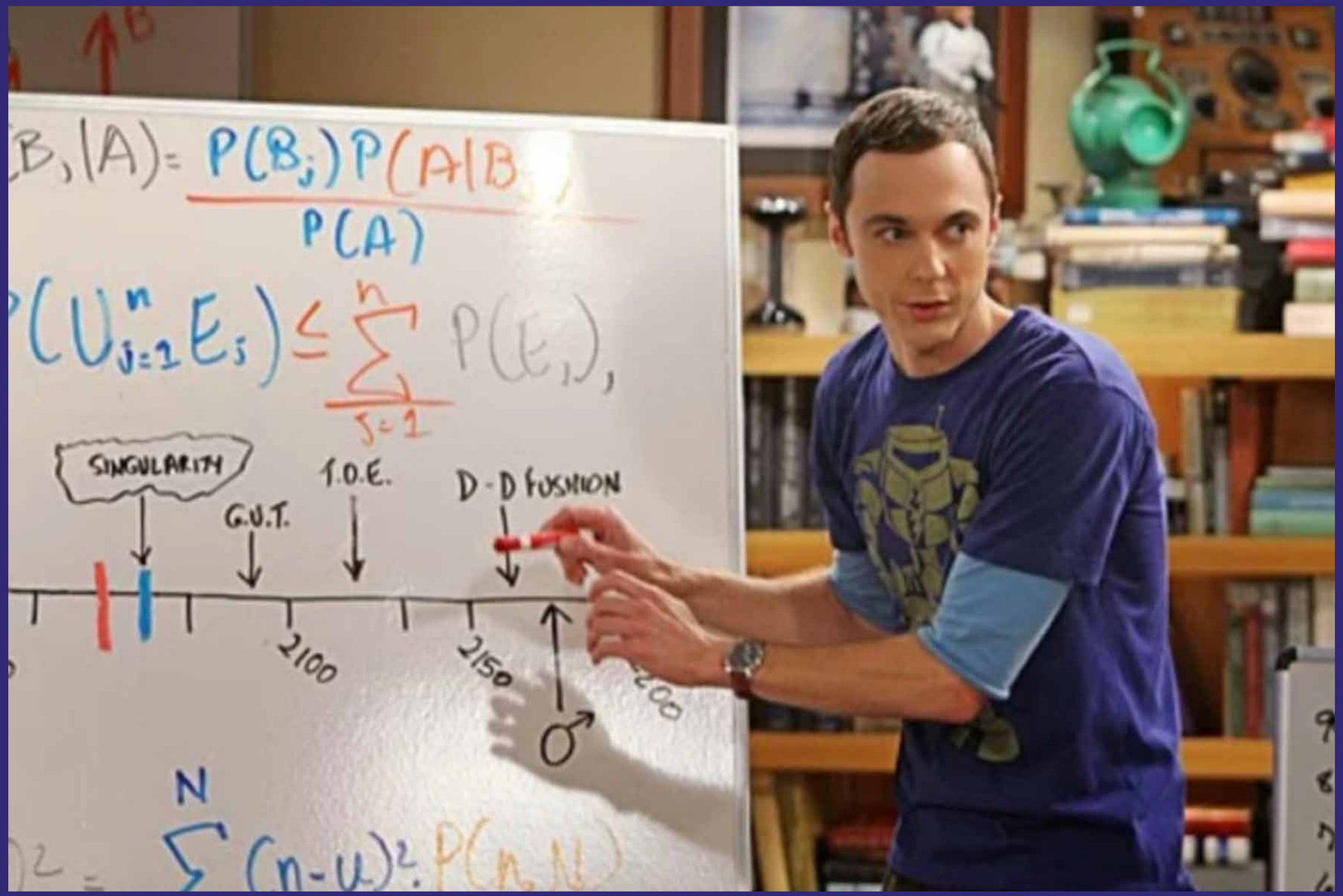


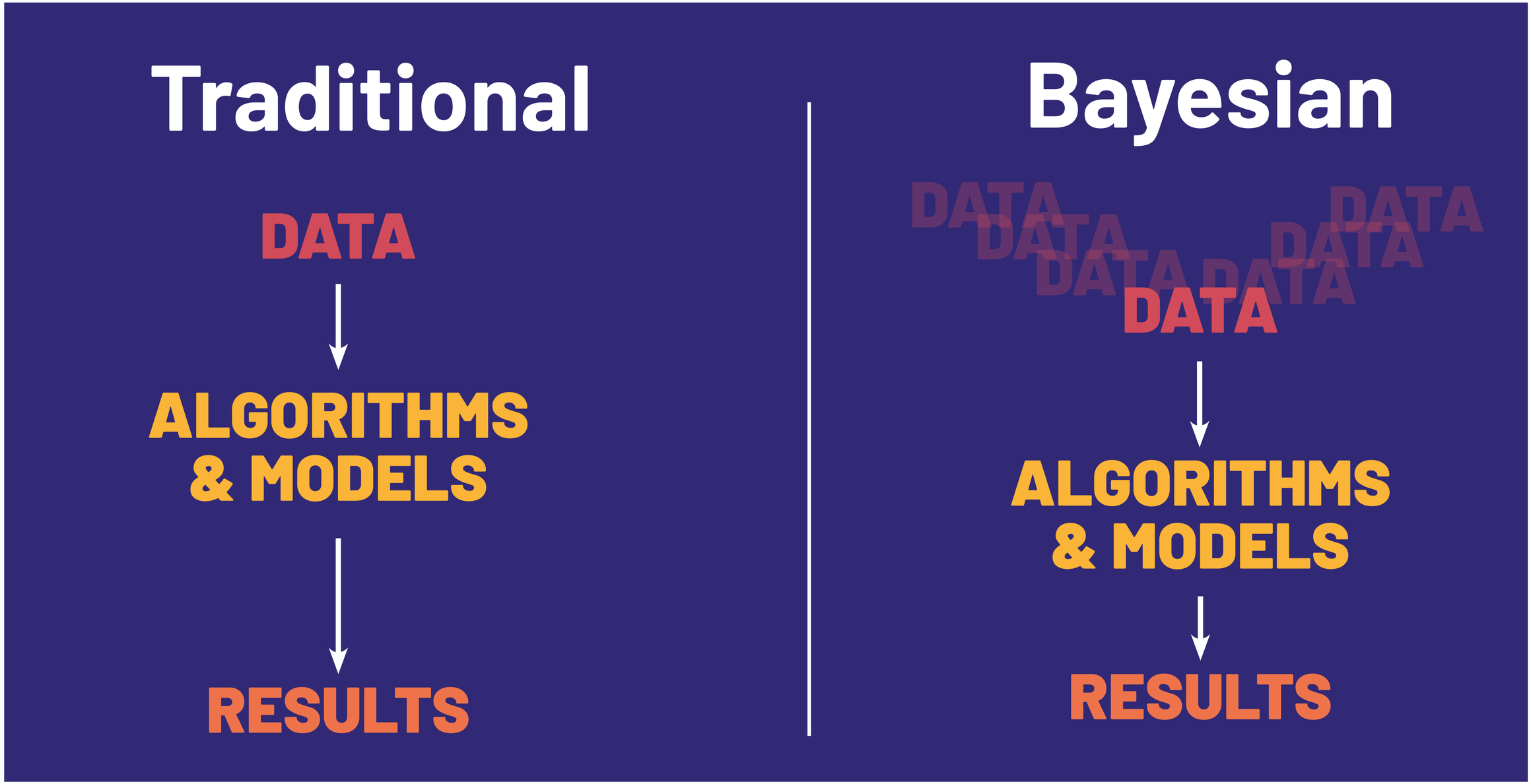
What’s Intersectional Analysis?


Odds of Getting Referral
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
White Men
White Women
Black Men
Black Woman

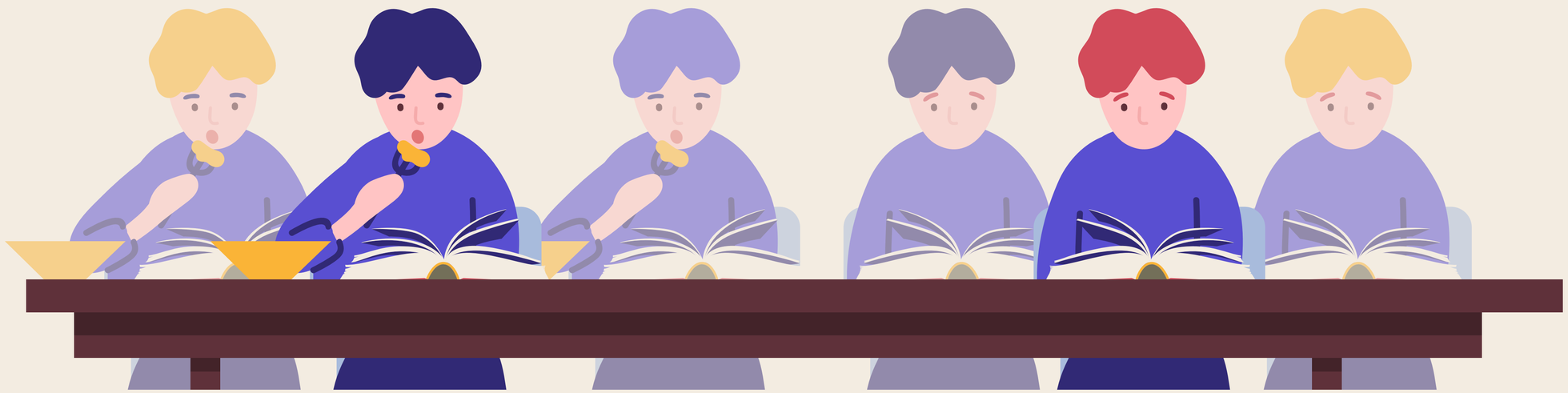

Odds of Getting Referral
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
White Men
White Women
Black Men
Black Woman

What’s Bayesian Analysis?


SINGULARITY
G.U.T.
T.O.E.
D-D FUSHION
2100
2150


Traditional
DATA
ALGORITHMS & MODELS
RESULTS
Bayesian
DATA
DATA
DATA
DATA
DATA
DATA
DATA
ALGORITHMS & MODELS
RESULTS

Traditional analysis will look at this difference over time.



Bayesian analysis will look at the difference over time, as well as incorporating ‘the prior’; results from similar programs and analyses.

The ideal number of cows for optimal productivity is $e^{1.27}$ cows

Thank you.

WeAllCount.com

Heather Krause, PStat
heather@idatassist.com
@datassist