

Observational studies and experiments

INTRODUCTION TO DATA IN R

Mine Cetinkaya-Rundel

Associate Professor at Duke University &
Data Scientist and Professional Educator
at RStudio



Types of studies

- **Observational study:**
 - Collect data in a way that does not directly interfere with how the data arise
 - Only correlation can be inferred
- **Experiment:**
 - Randomly assign subjects to various treatments
 - Causation can be inferred

Design a study

Screens at bedtime and attention span

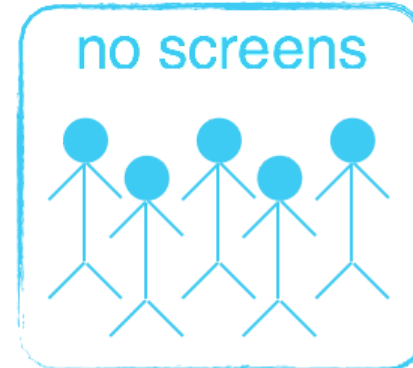
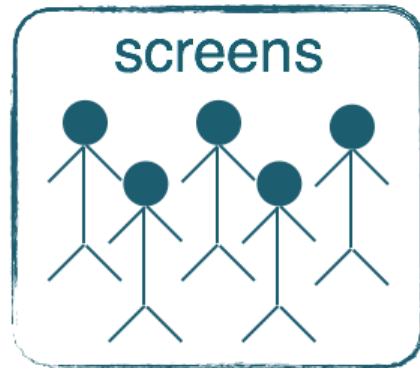
observational
study

experiment

Design a study

Screens at bedtime and attention span

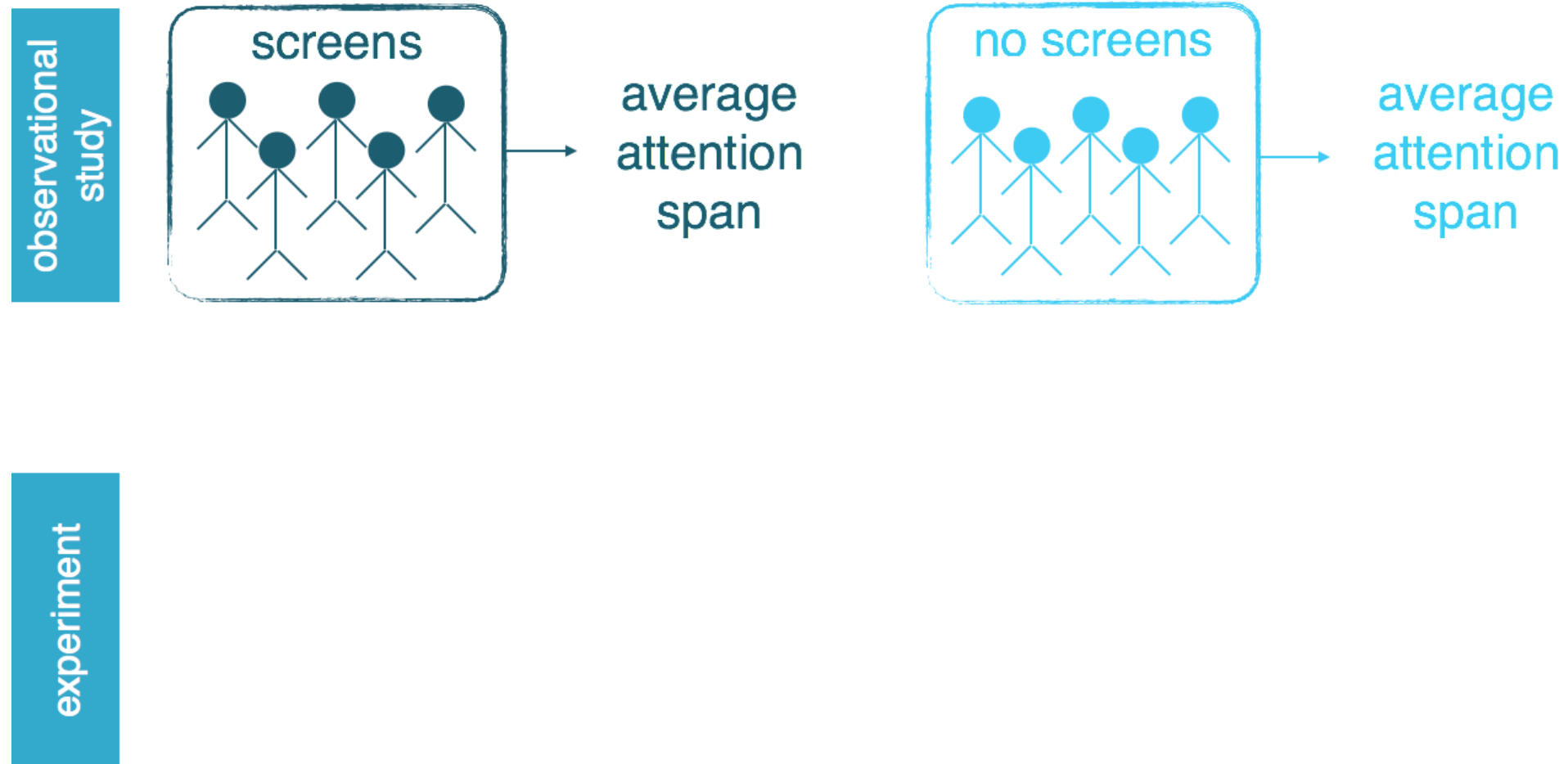
observational
study



experiment

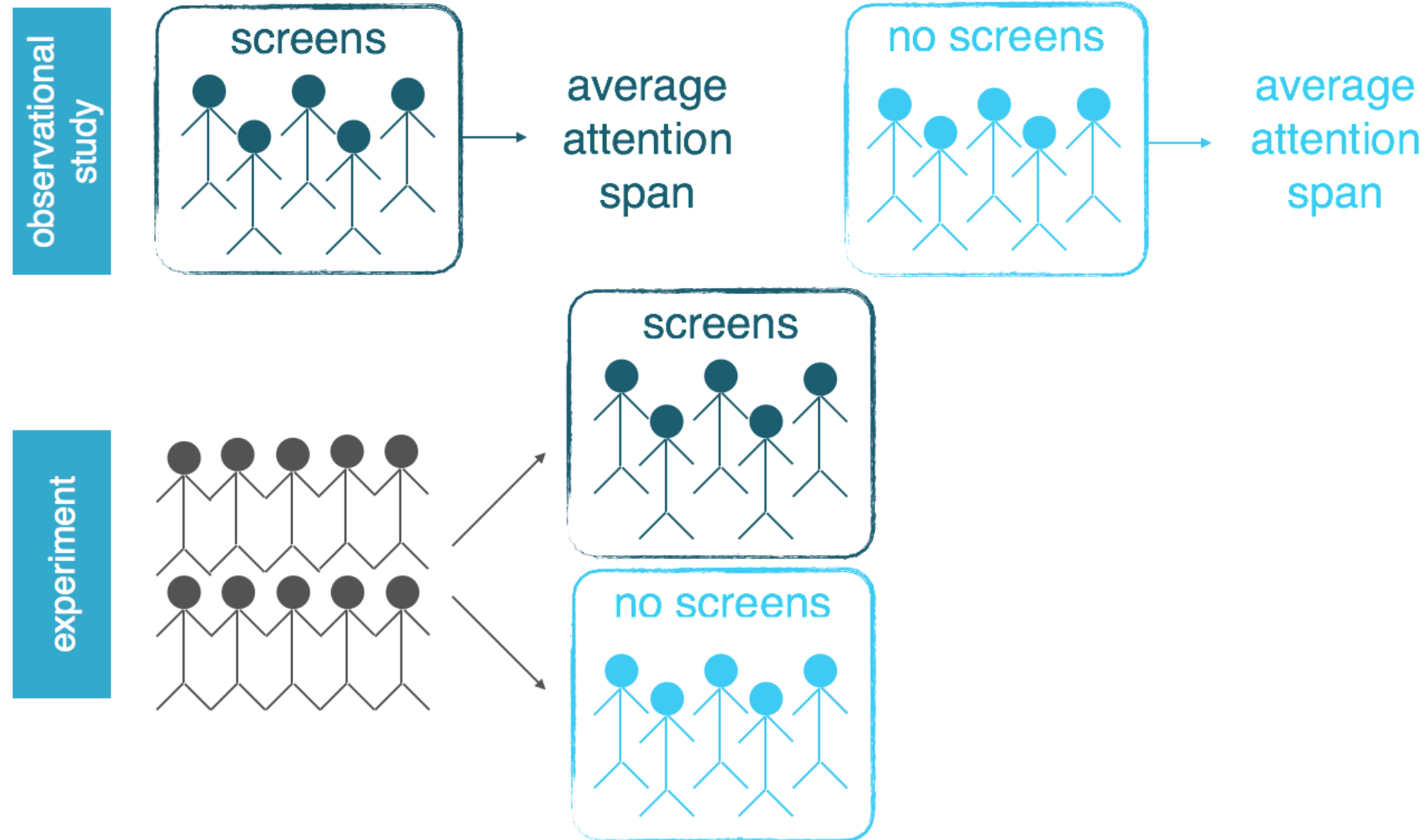
Design a study

Screens at bedtime and attention span



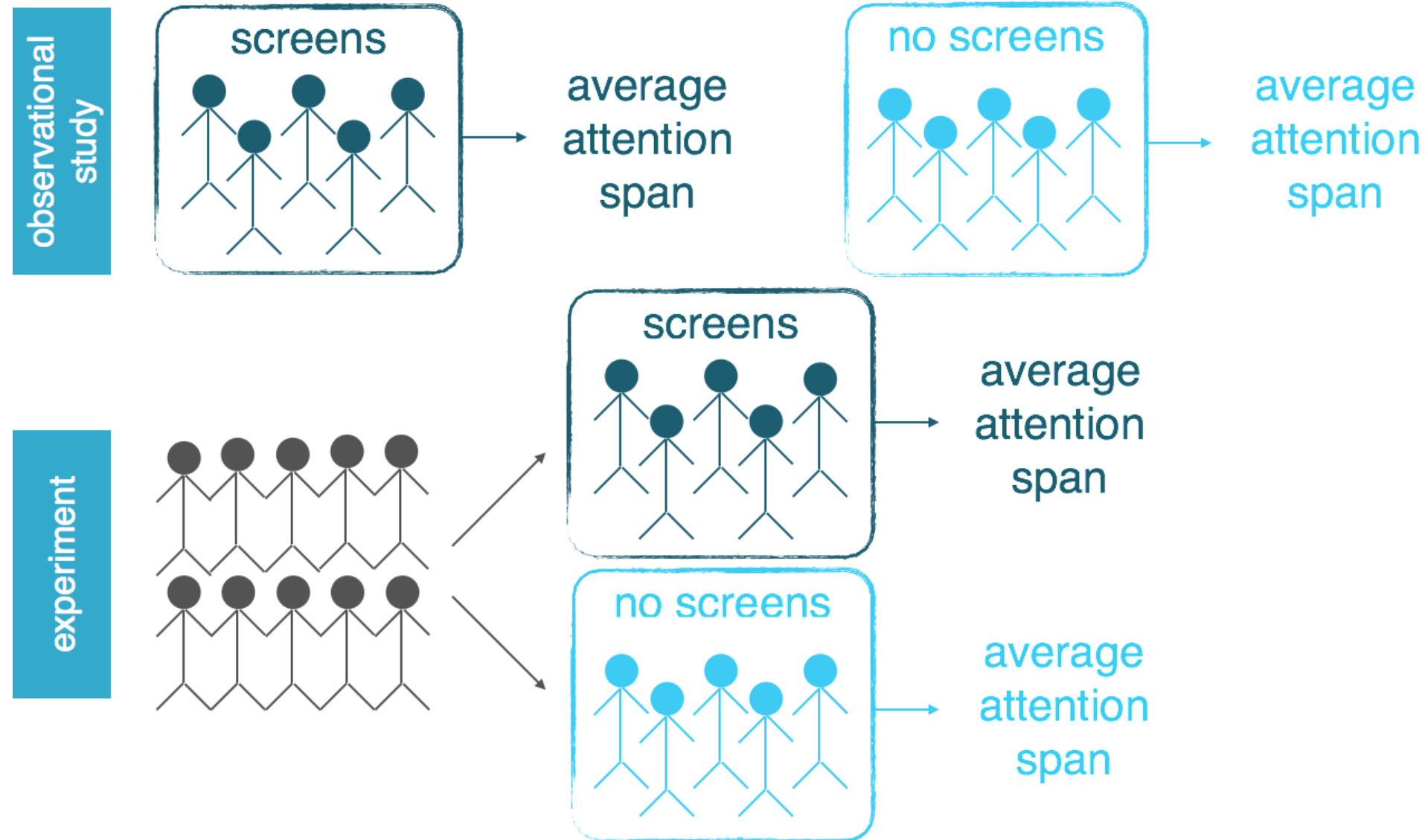
Design a study

Screens at bedtime and attention span



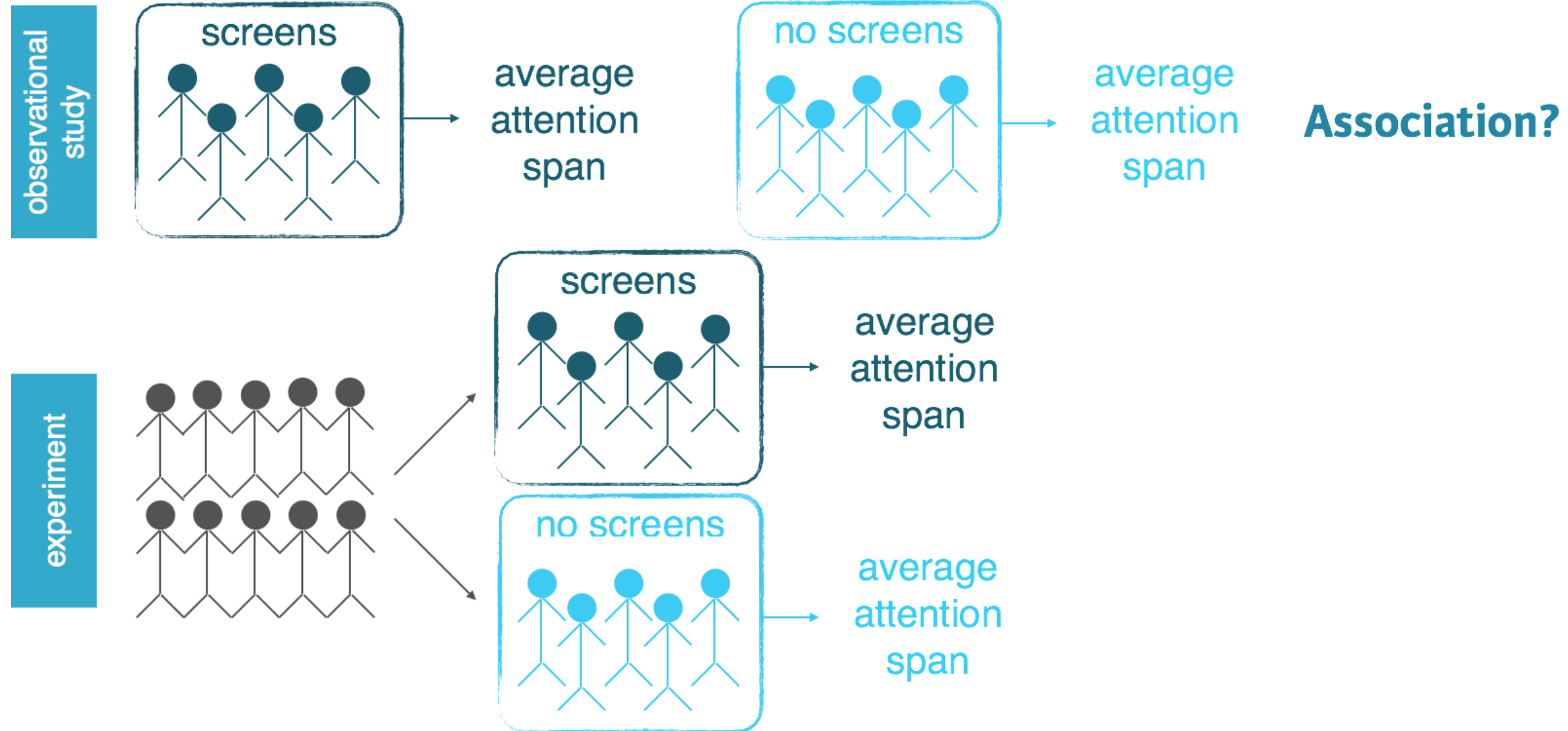
Design a study

Screens at bedtime and attention span



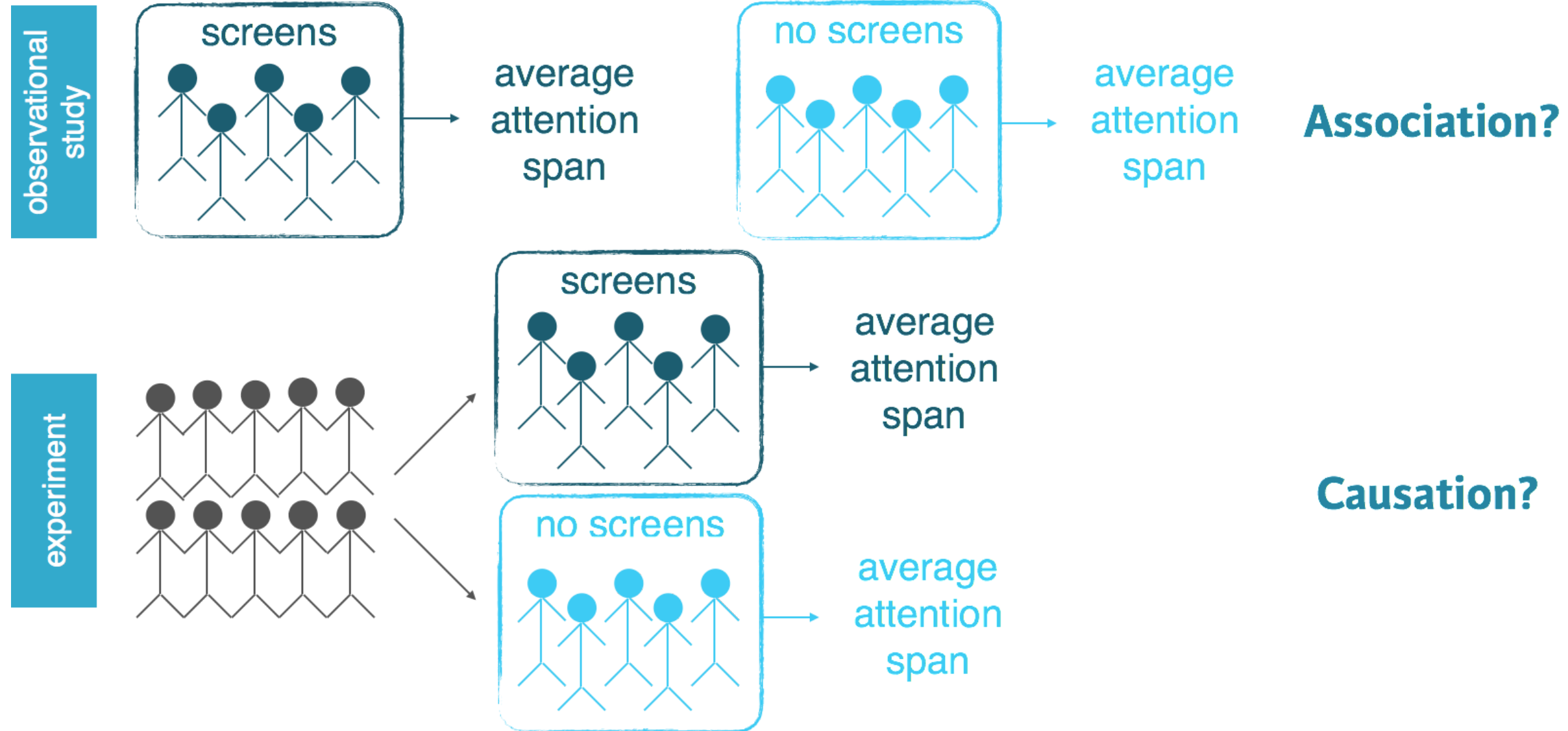
Design a study

Screens at bedtime and attention span



Design a study

Screens at bedtime and attention span



Let's practice!
INTRODUCTION TO DATA IN R

Random sampling and random assignment

INTRODUCTION TO DATA IN R

Mine Cetinkaya-Rundel

Associate Professor at Duke University &
Data Scientist and Professional Educator
at RStudio



Random...

- **Random sampling:**
 - At selection of subjects from population
 - Helps generalizability of results
- **Random assignment:**
 - At selection of subjects from population
 - Helps infer causation from results

Scope of inference

	Random assignment	No random assignment	
Random sampling	Causal and generalizable	Not causal, but generalizable	Generalizable
No random sampling	Causal, but not generalizable	Neither causal nor generalizable	Not generalizable
	Causal	Not causal	

Let's practice!
INTRODUCTION TO DATA IN R

Simpson's paradox


INTRODUCTION TO DATA IN R



Mine Cetinkaya-Rundel

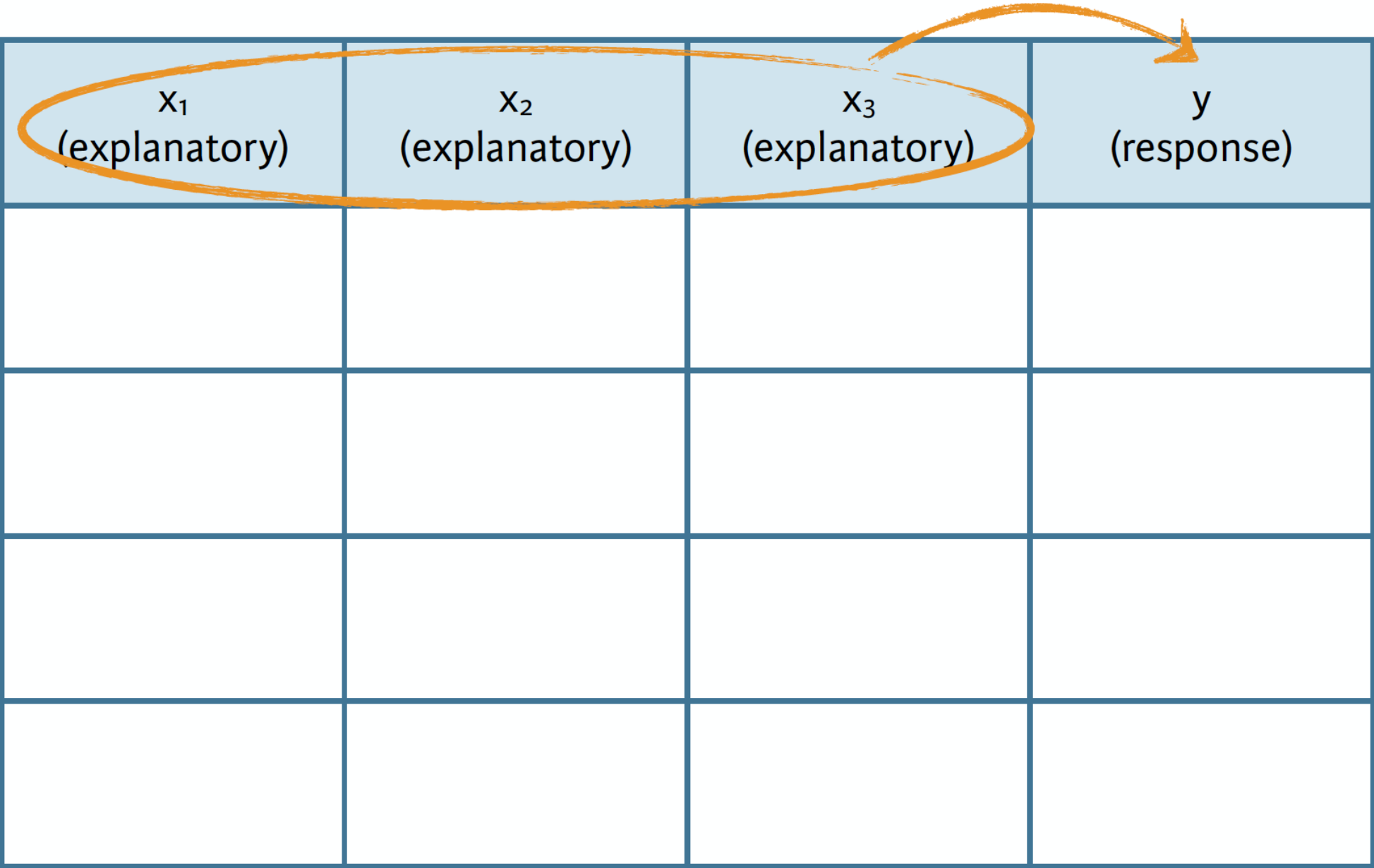
Associate Professor at Duke University &
Data Scientist and Professional Educator
at RStudio

Explanatory and response



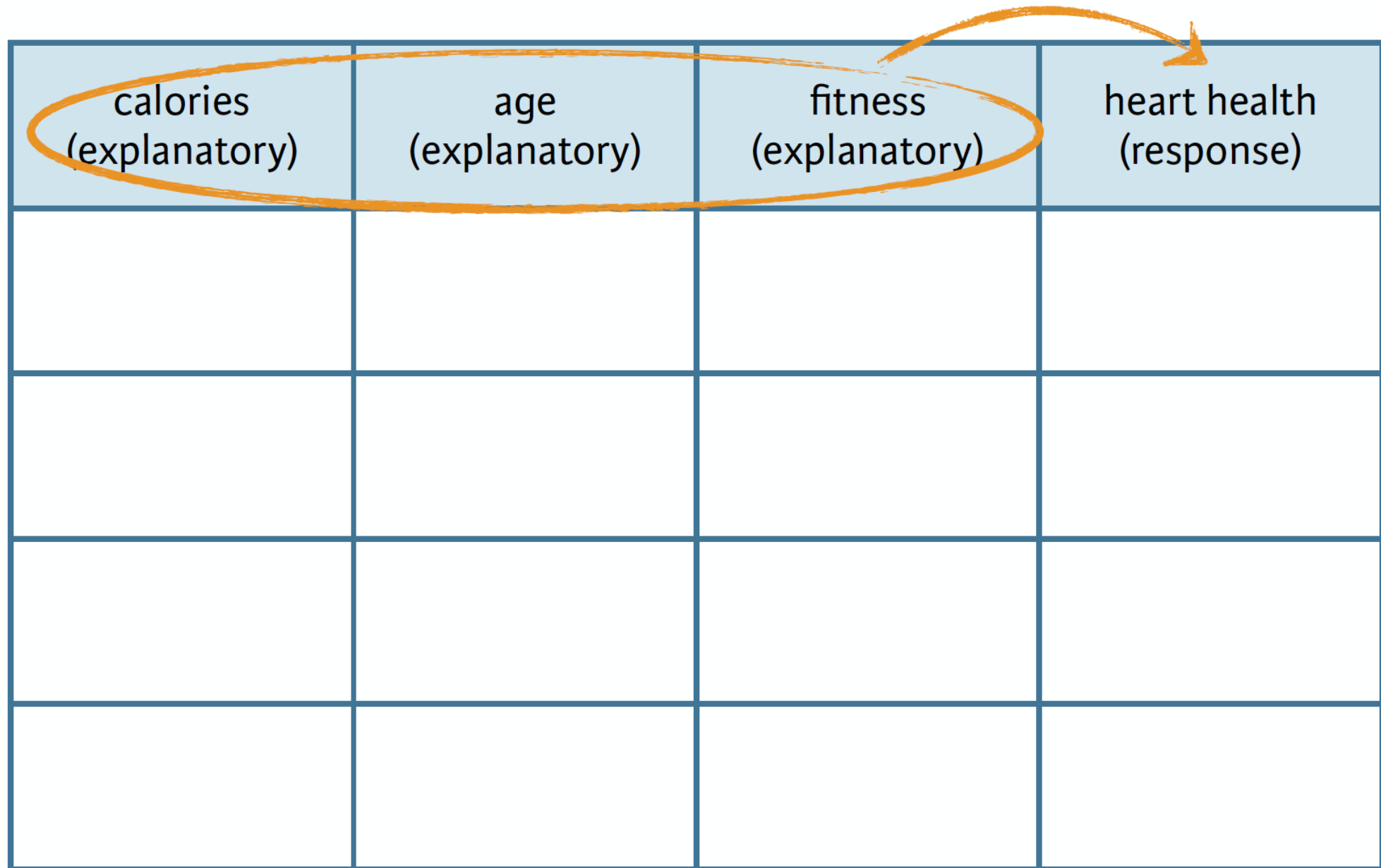
x (explanatory)	y (response)

Multivariate relationships



x_1 (explanatory)	x_2 (explanatory)	x_3 (explanatory)	y (response)

Multivariate relationships



A diagram illustrating multivariate relationships. It features a table with four columns: 'calories (explanatory)', 'age (explanatory)', 'fitness (explanatory)', and 'heart health (response)'. The first three columns are grouped by an orange oval, and an orange arrow points from this group to the fourth column. The table has five rows, with the first row containing the column headers and the remaining four rows being empty.

calories (explanatory)	age (explanatory)	fitness (explanatory)	heart health (response)

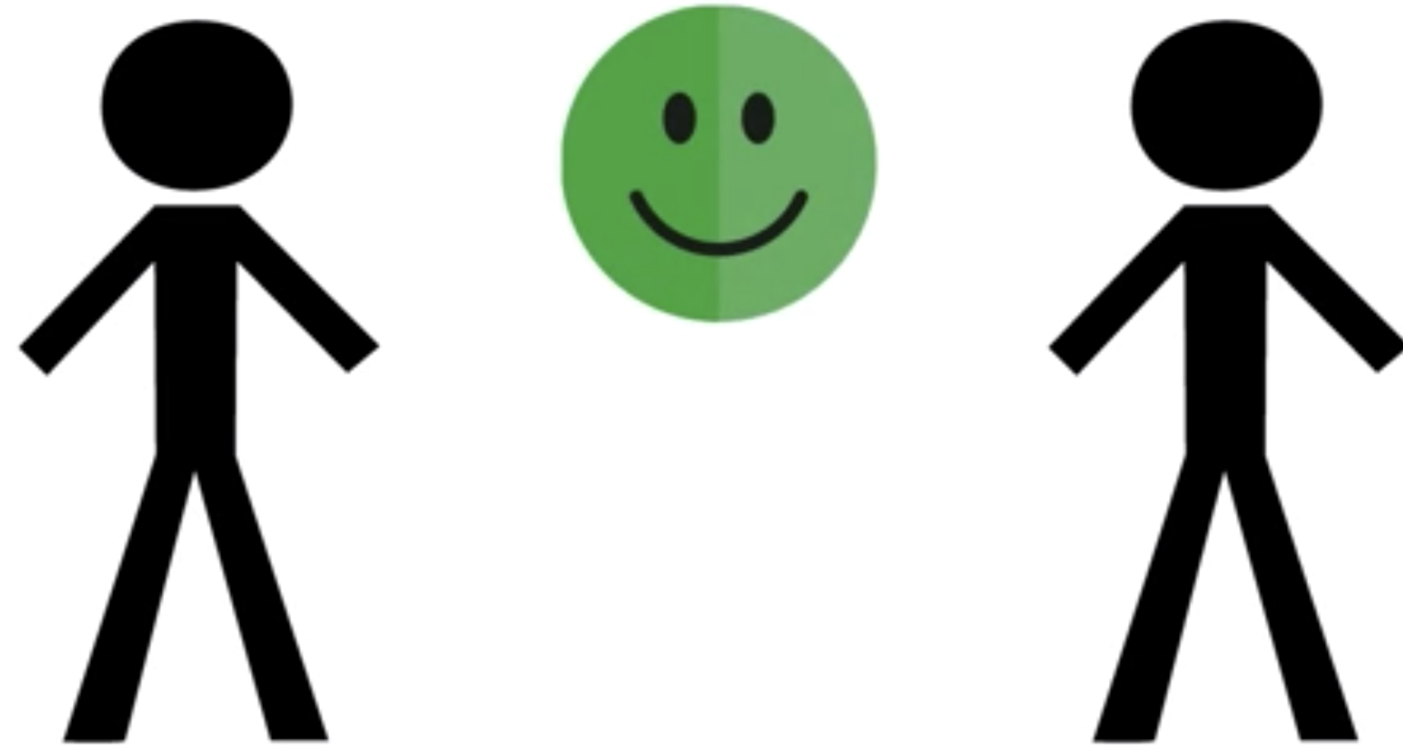
Simpson's paradox



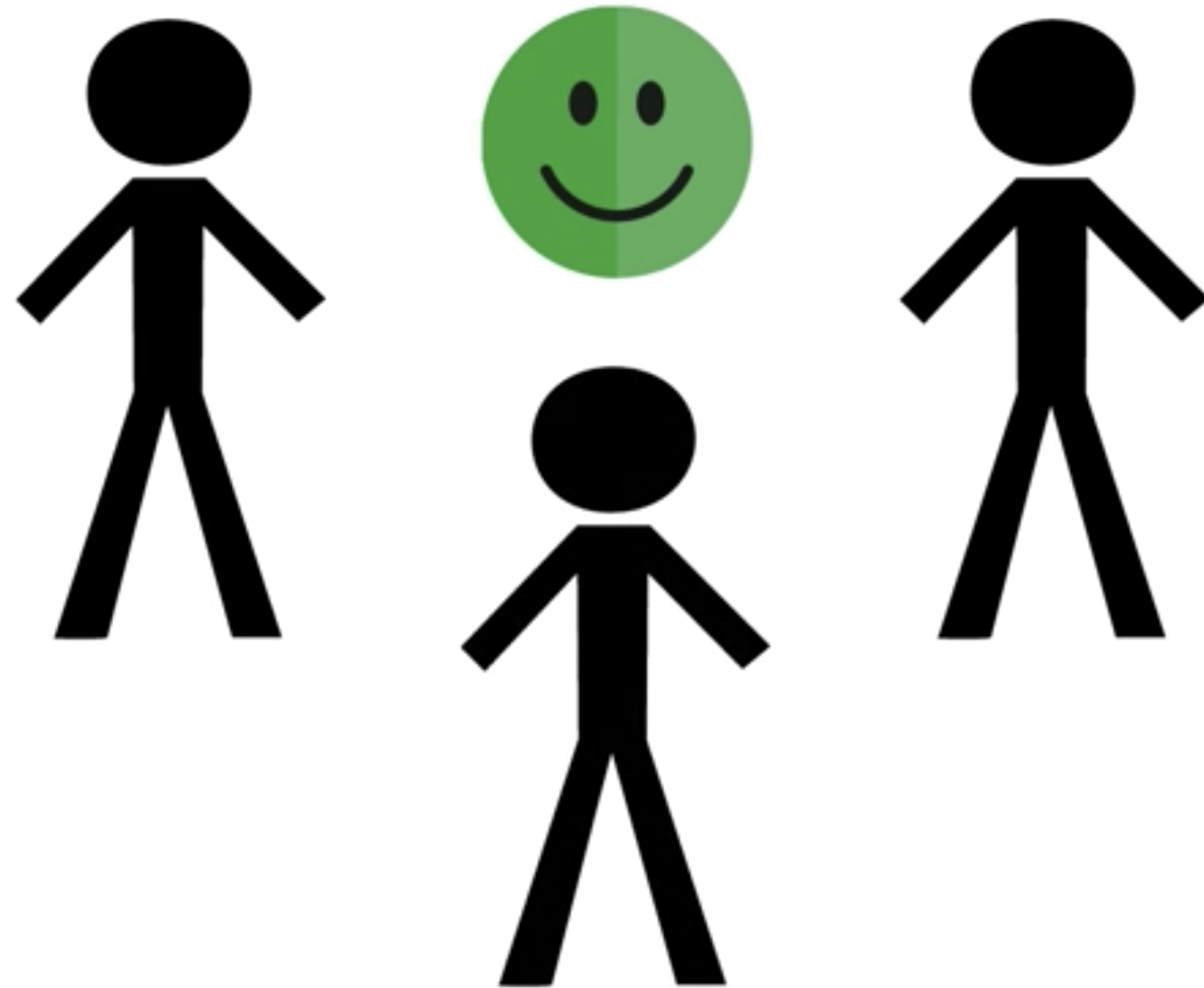
Simpson's paradox



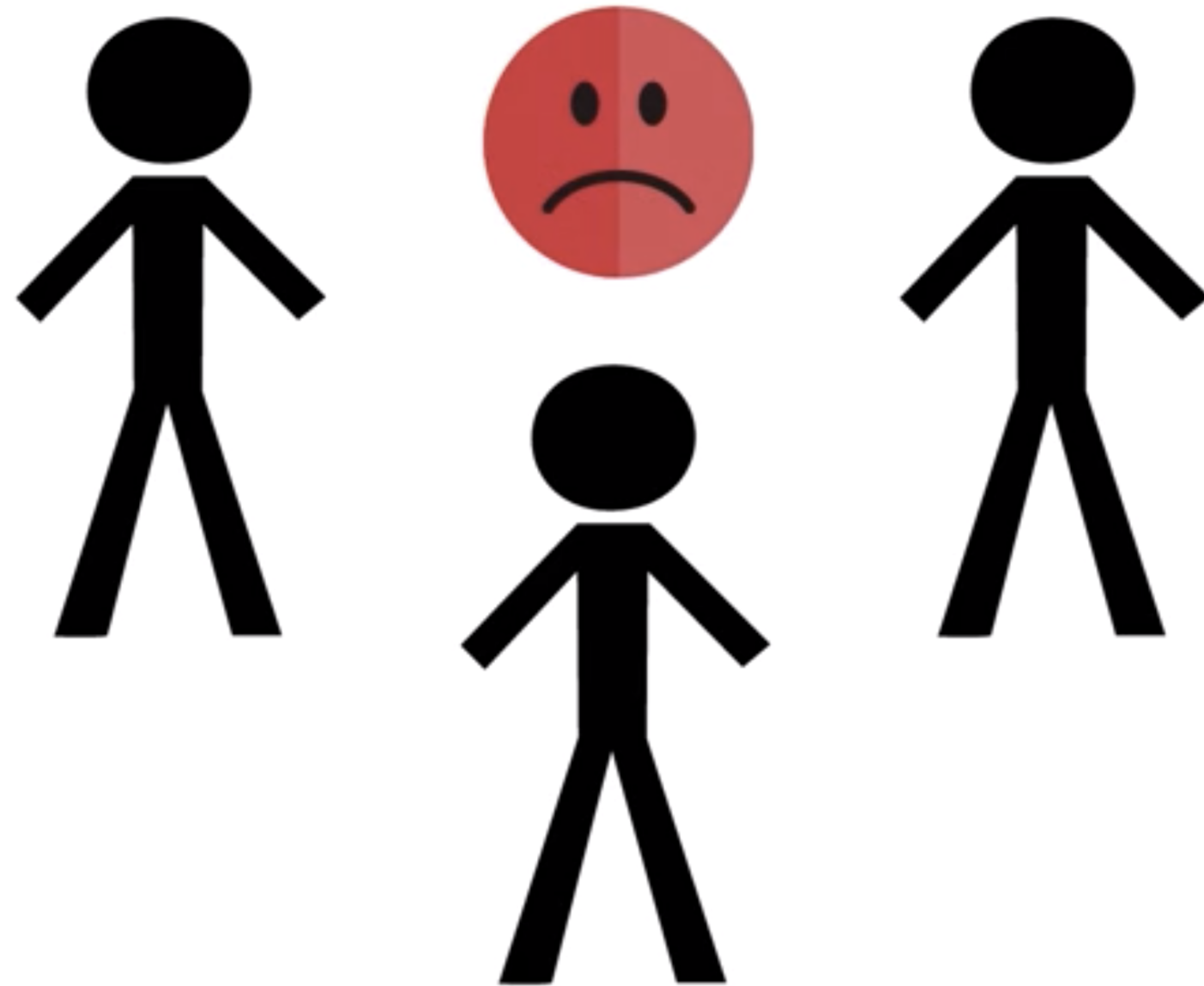
Simpson's paradox



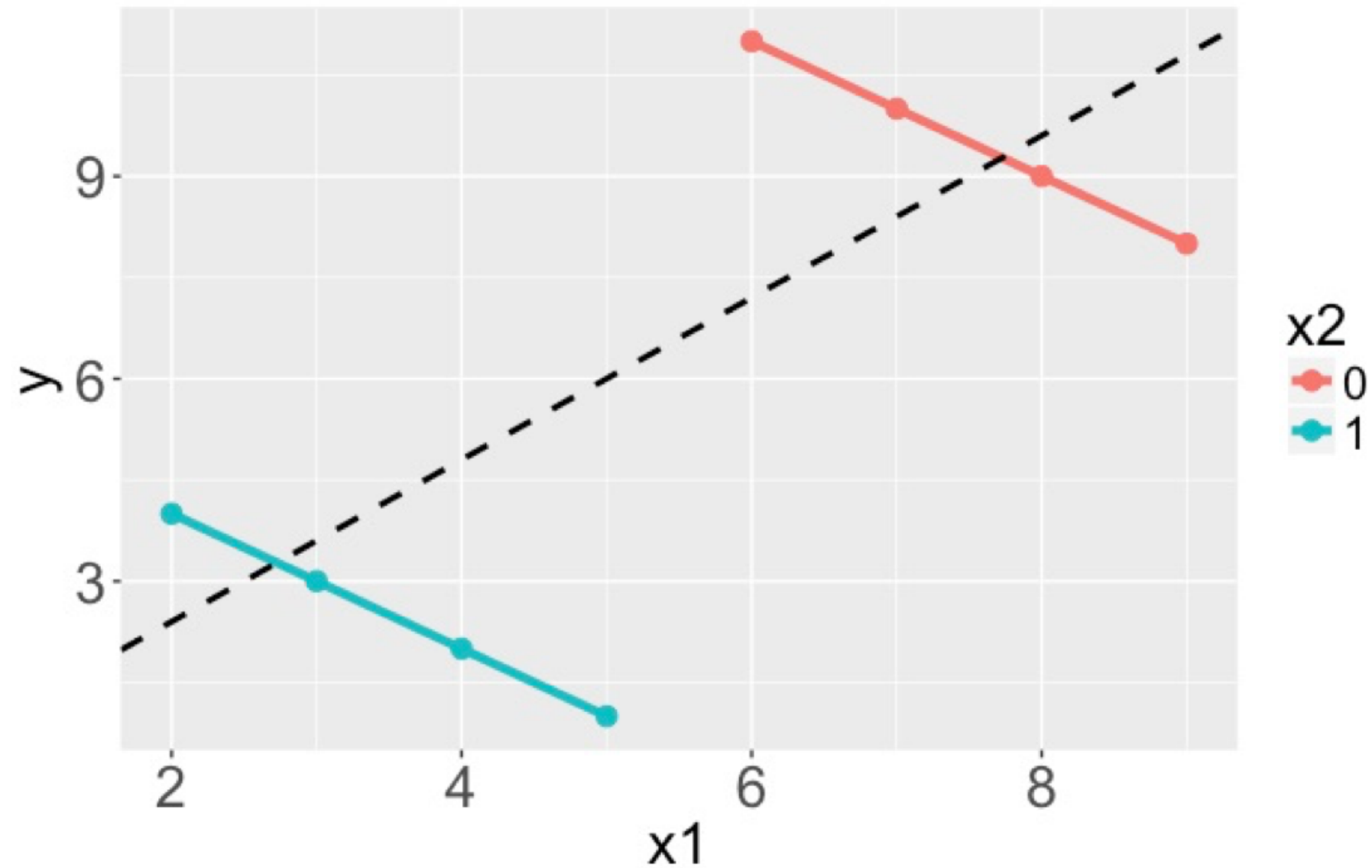
Simpson's paradox



Simpson's paradox



Simpson's paradox



Berkeley admission data

	Admitted	Rejected
Male	1198	1493
Female	557	1278

Let's get started!

INTRODUCTION TO DATA IN R

Recap: Simpson's paradox

INTRODUCTION TO DATA IN R



Mine Cetinkaya-Rundel

Associate Professor at Duke University &
Data Scientist and Professional Educator
at RStudio

Simpson's paradox

- Overall: males more likely to be admitted
- Within most departments: females more likely
- When controlling for department, relationship between gender and admission status is reversed
- Potential reason:
 - Women tended to apply to competitive departments with low admission rates
 - Men tended to apply to less competitive departments with high admission rates

Let's practice!
INTRODUCTION TO DATA IN R