Sources of data sets

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Data are defined by how they are collected

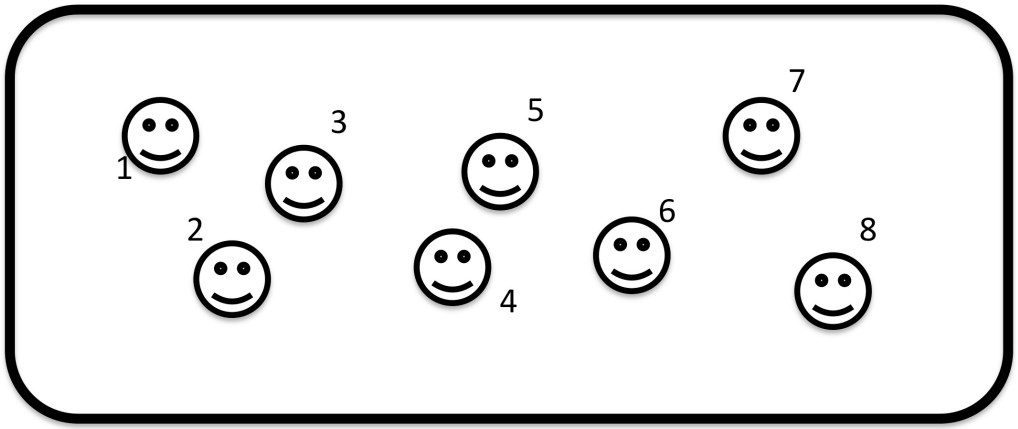
Main types

- Census (descriptive)
- Observational study (inferential)
- Convenience sample (all types - may be biased)
- Randomized trial (causal)

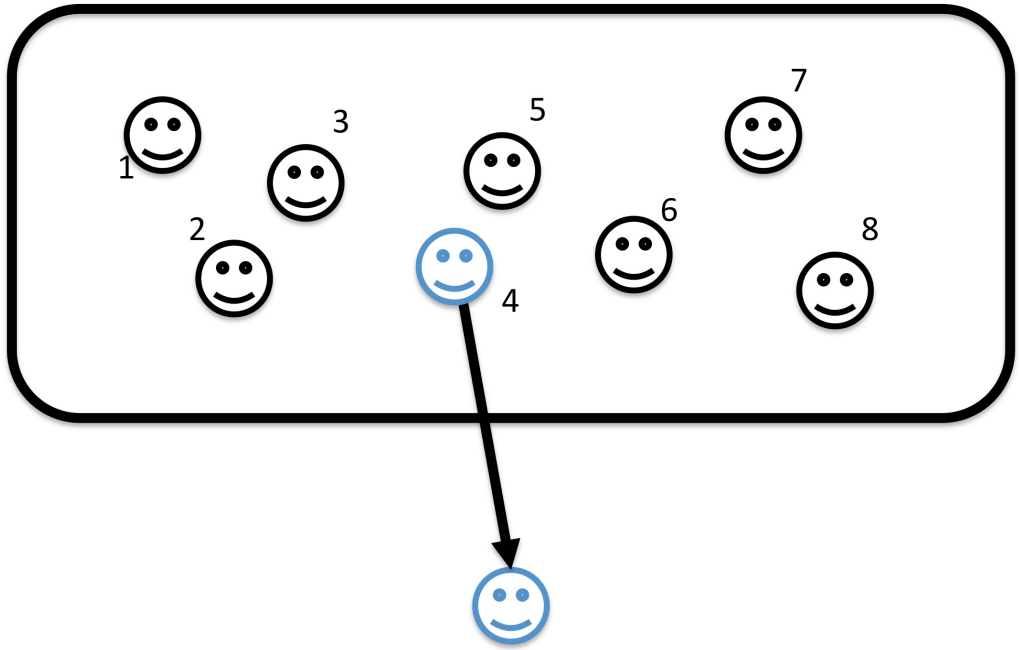
Other types

- Prediction study (prediction)
- Studies over time
    - Cross sectional (inferential)
    - Longitudinal (inferential, predictive)
- Retrospective (inferential)

A population

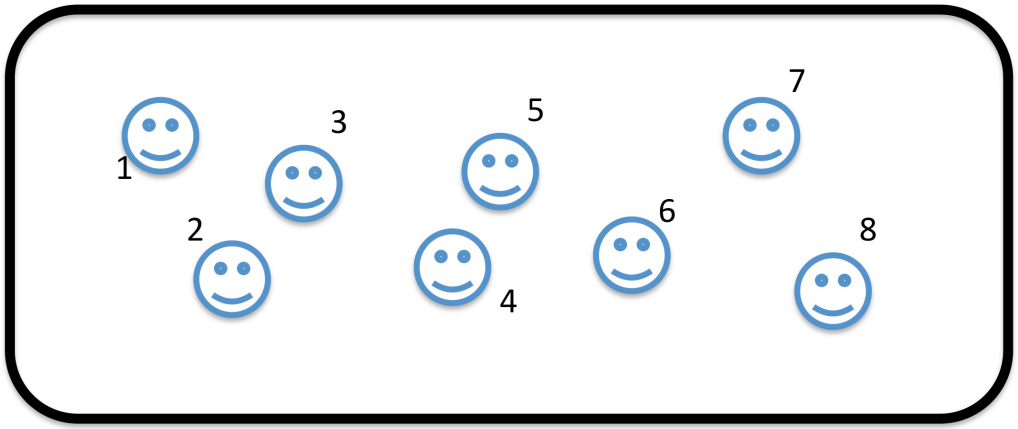


Pick a person and measure



Name	Cancer	Smoker	Baltimore	Exercise
Sarah	Yes	No	No	Yes

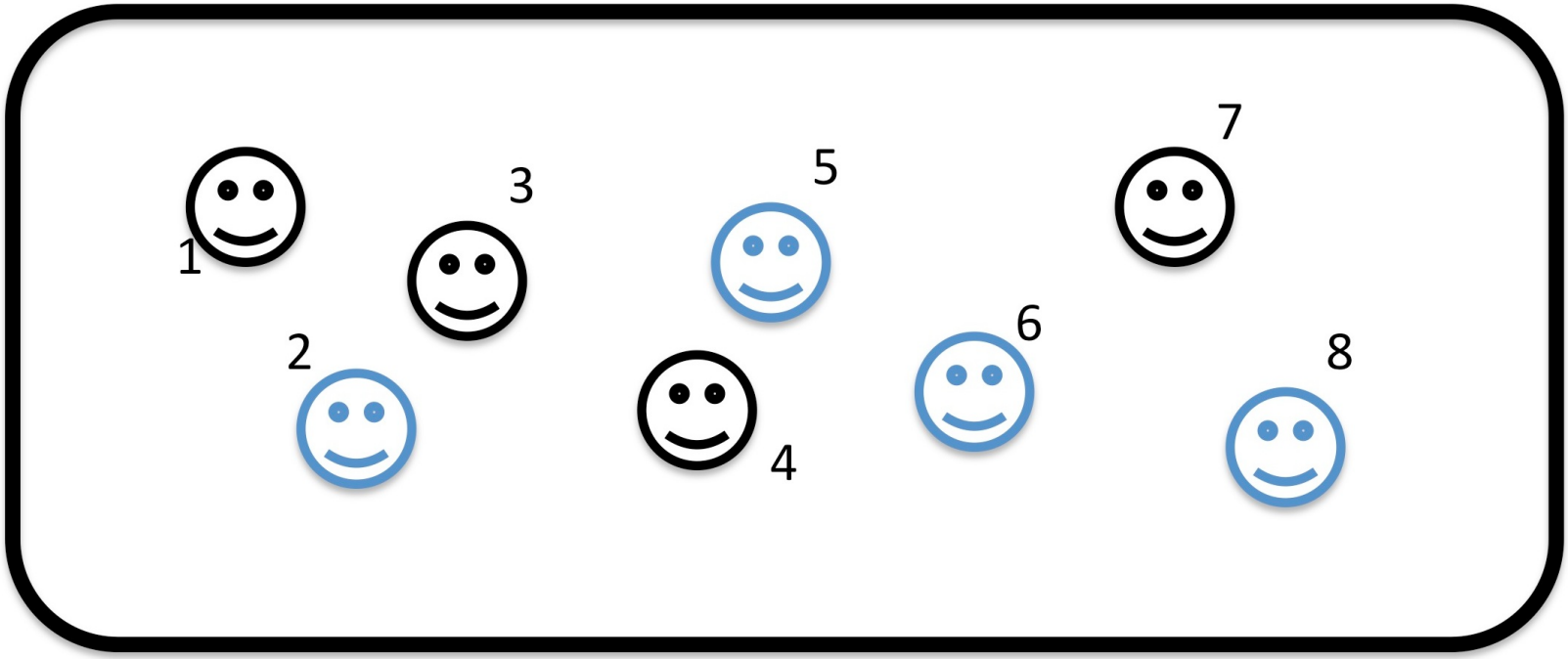
Census



Observational study

```
set.seed(5)
sample(1:8,size=4,replace=FALSE)
```

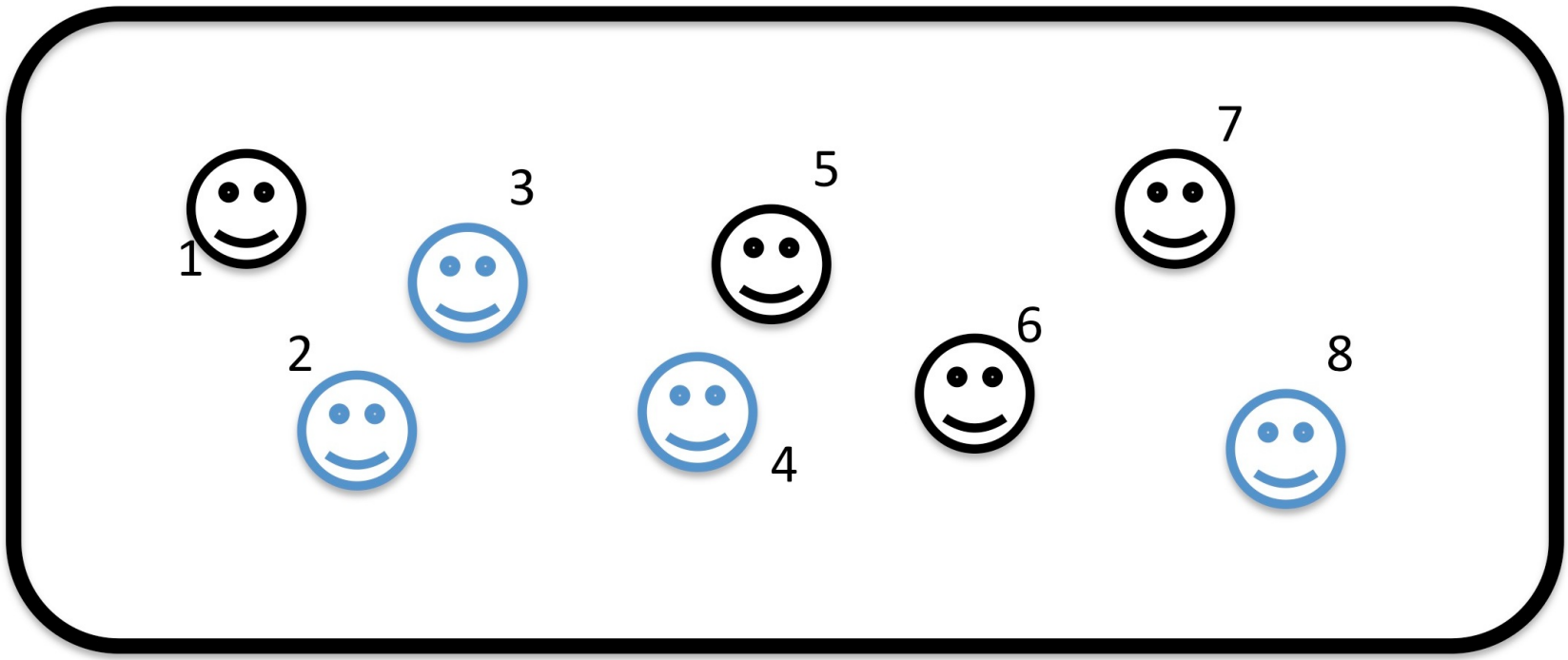
```
[1] 2 5 6 8
```



Convenience sample

```
probs = c(5,5,5,5,1,1,1,1)/16
sample(1:8,size=4,replace=FALSE,prob=probs)
```

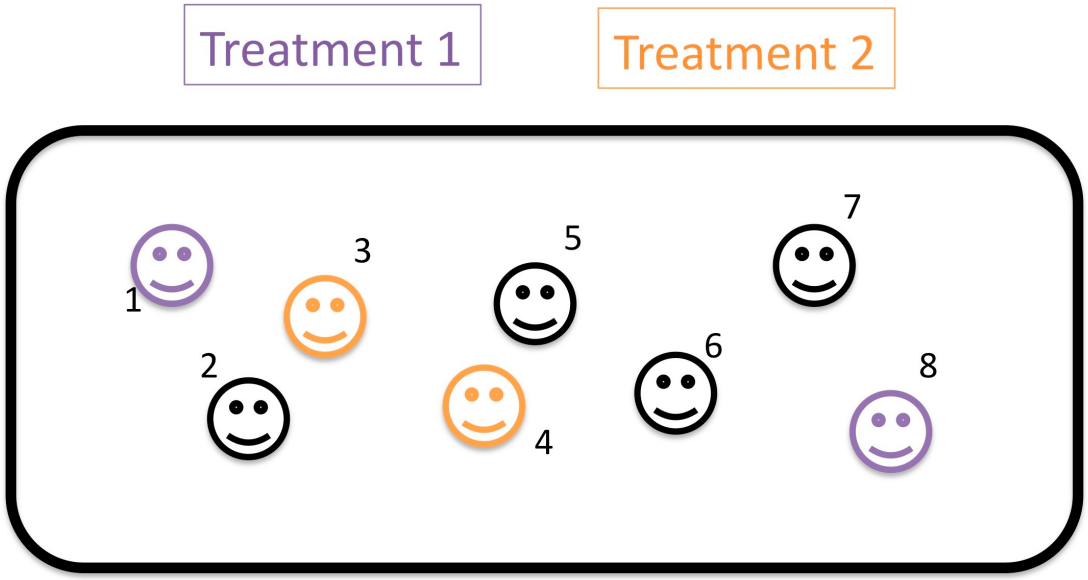
```
[1] 4 8 3 2
```



Randomized trial

```
treat1 = sample(1:8,size=2,replace=FALSE); treat2 = sample(2:7,size=2,replace=FALSE)
c(treat1,treat2)
```

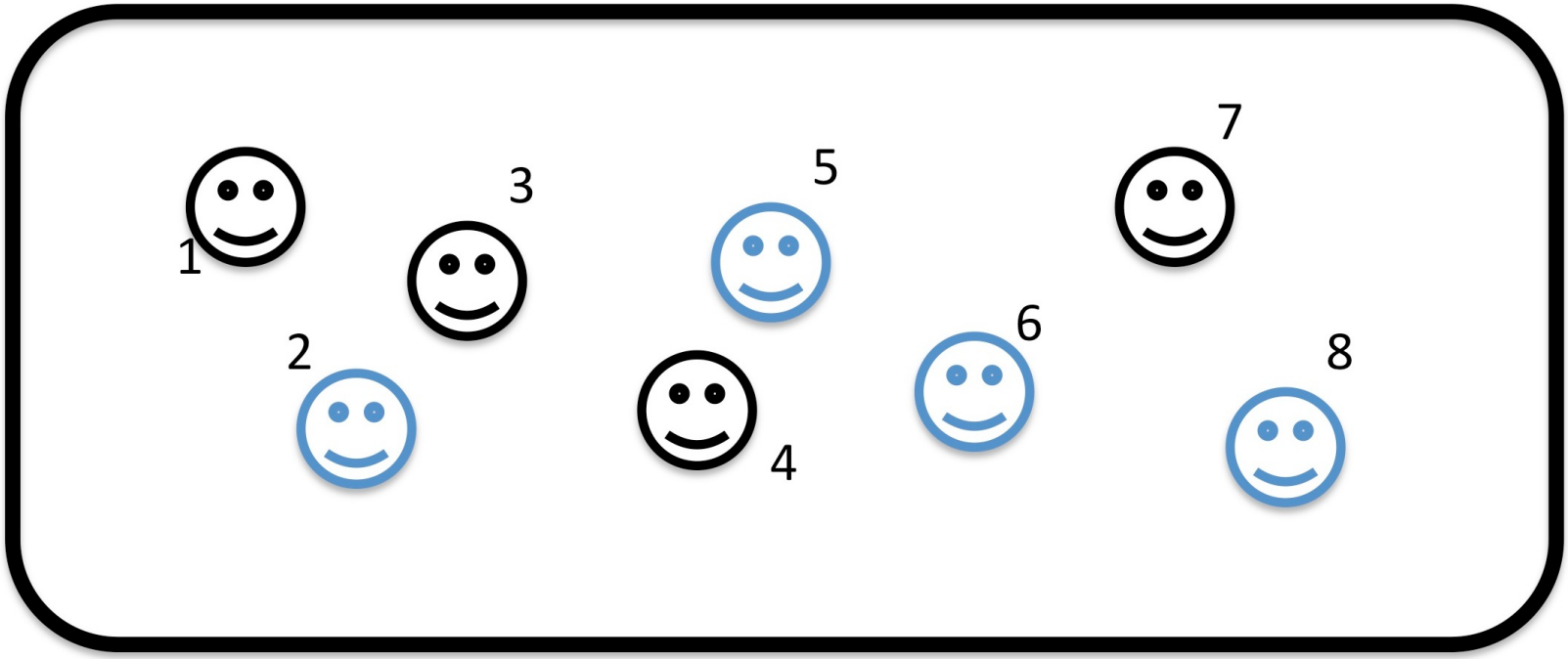
```
[1] 8 1 3 4
```



Prediction study: train

```
set.seed(5)
sample(1:8,size=4,replace=FALSE)
```

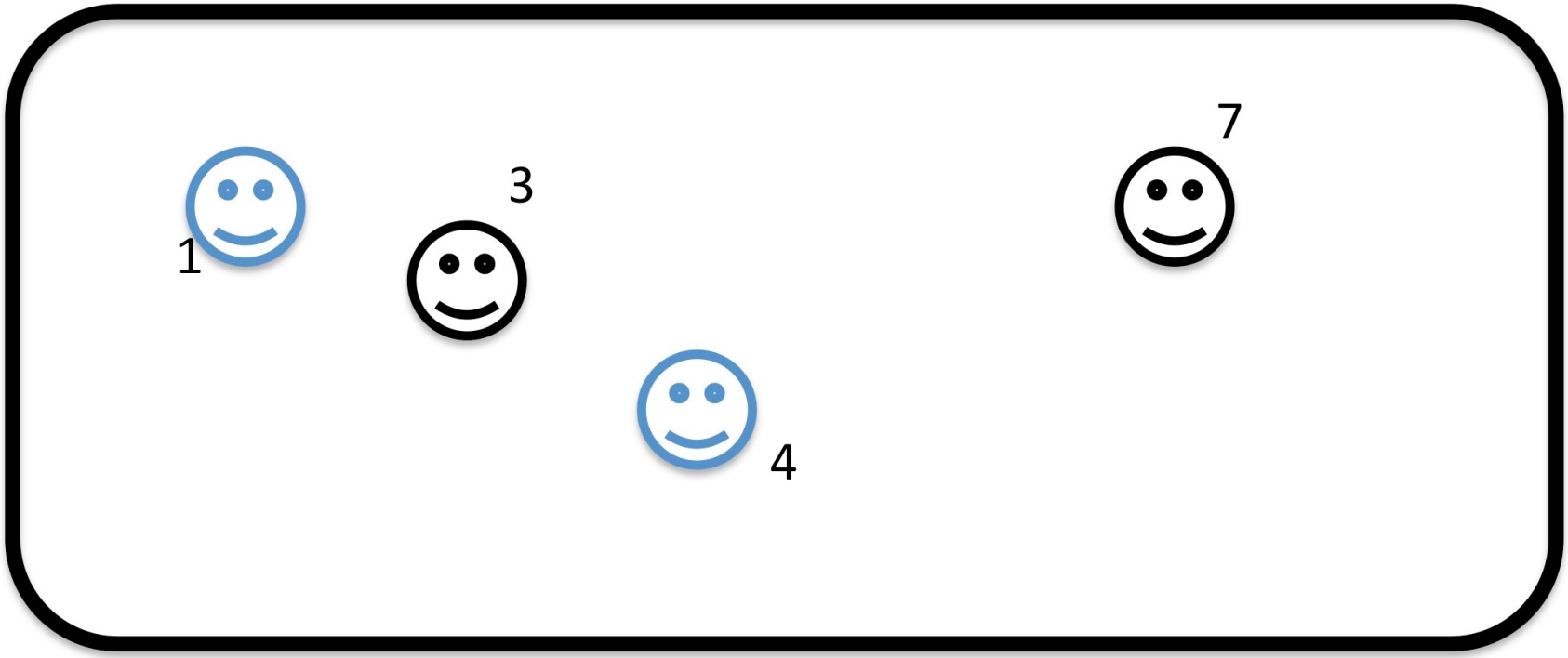
```
[1] 2 5 6 8
```



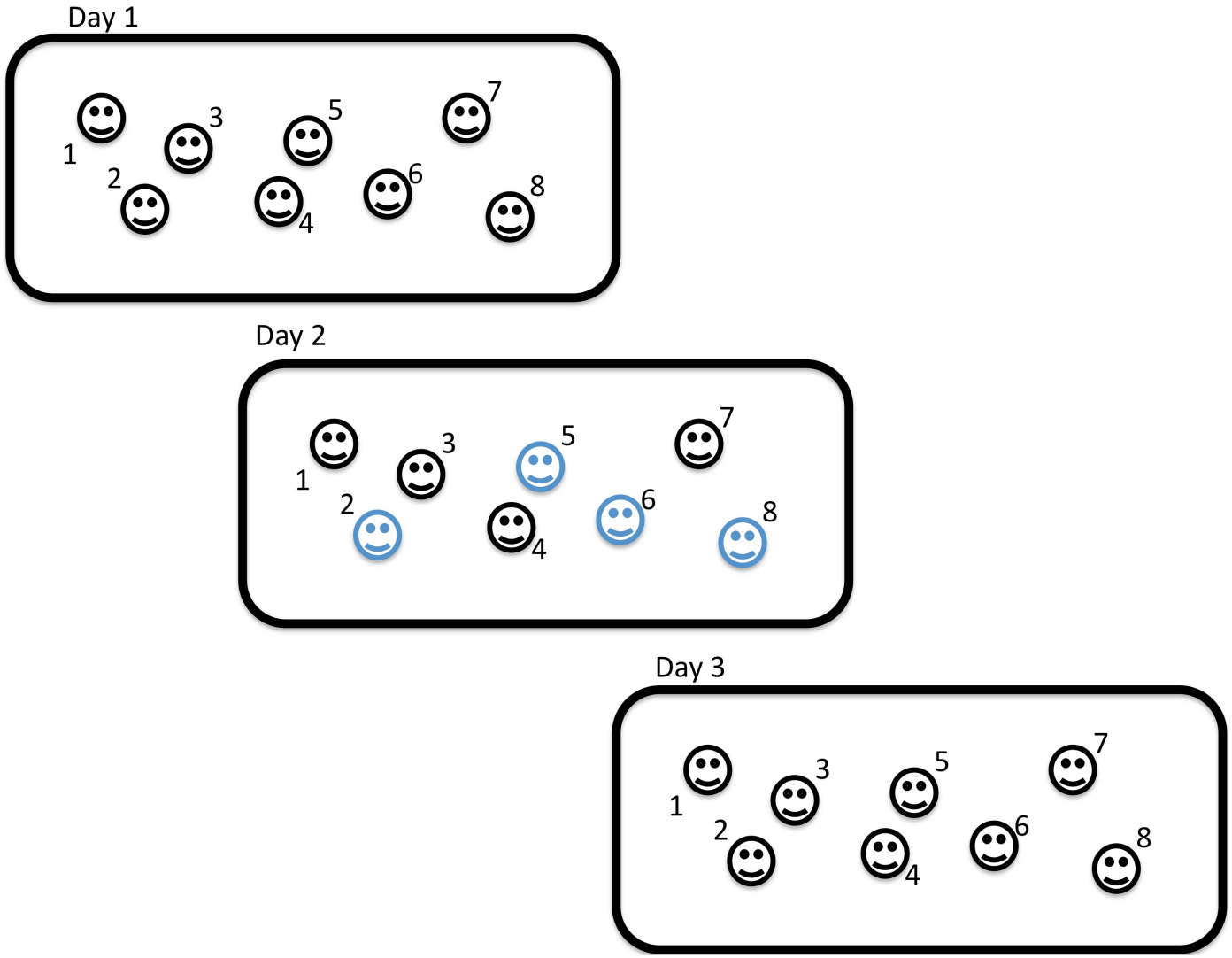
Prediction study: test

```
sample(c(1,3,4,7),size=2,replace=FALSE)
```

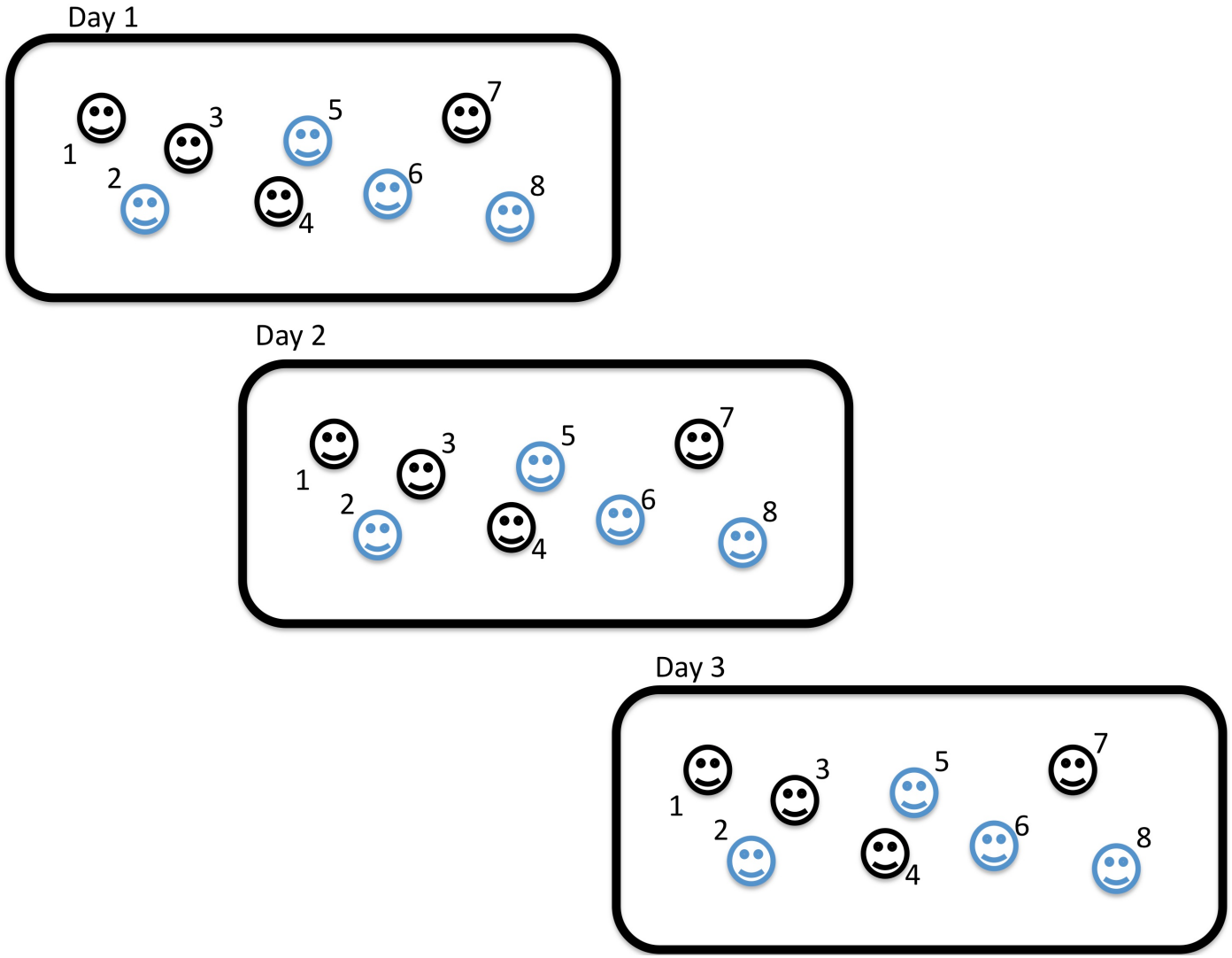
```
[1] 1 4
```



Study over time: cross-sectional



Study over time: longitudinal



Study over time: retrospective