Sampling strategies

INTRODUCTION TO DATA IN R

Mine Cetinkaya-Rundel

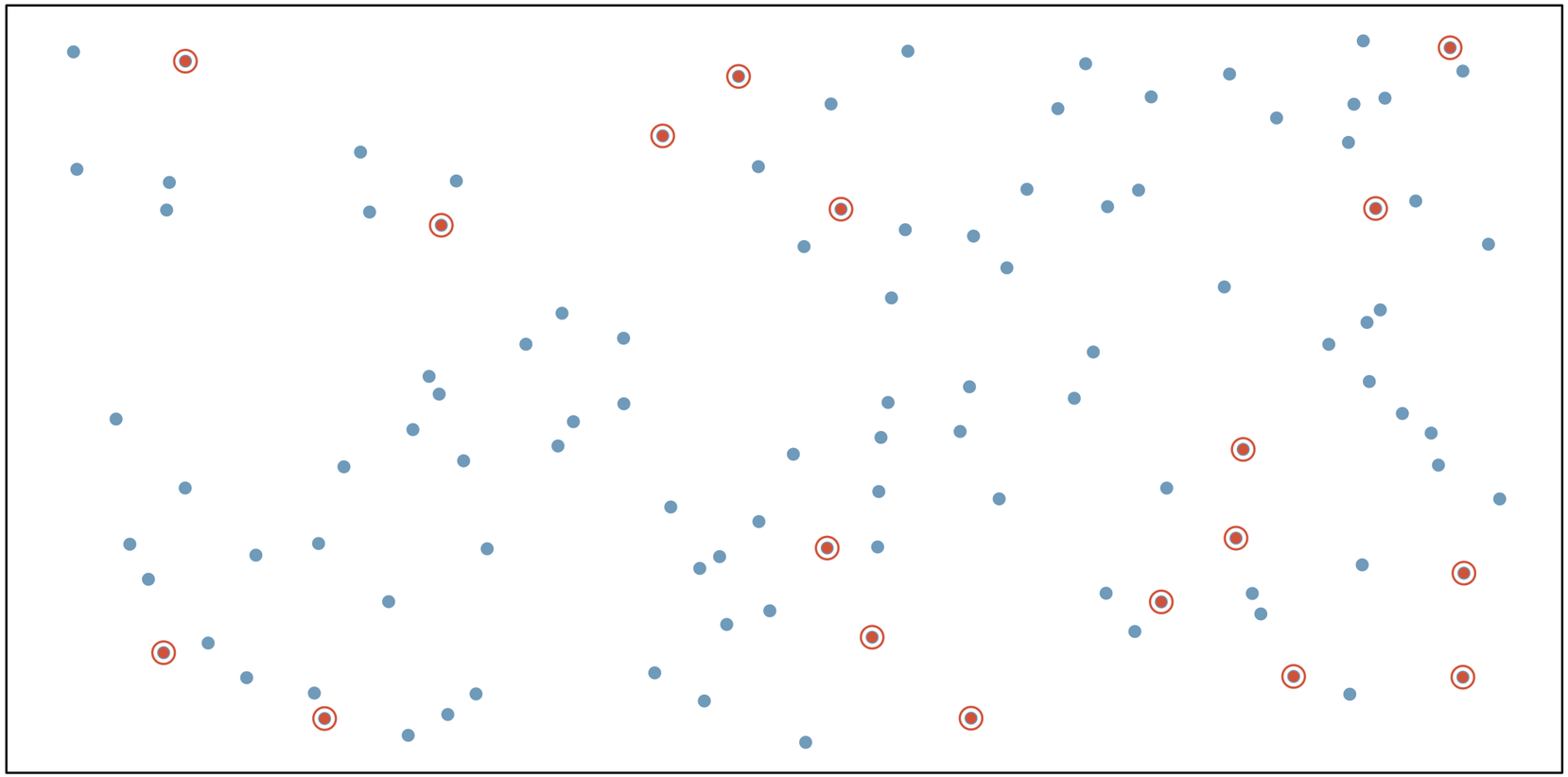
Associate Professor at Duke University &
Data Scientist and Professional Educator
at RStudio

Why not take a census?

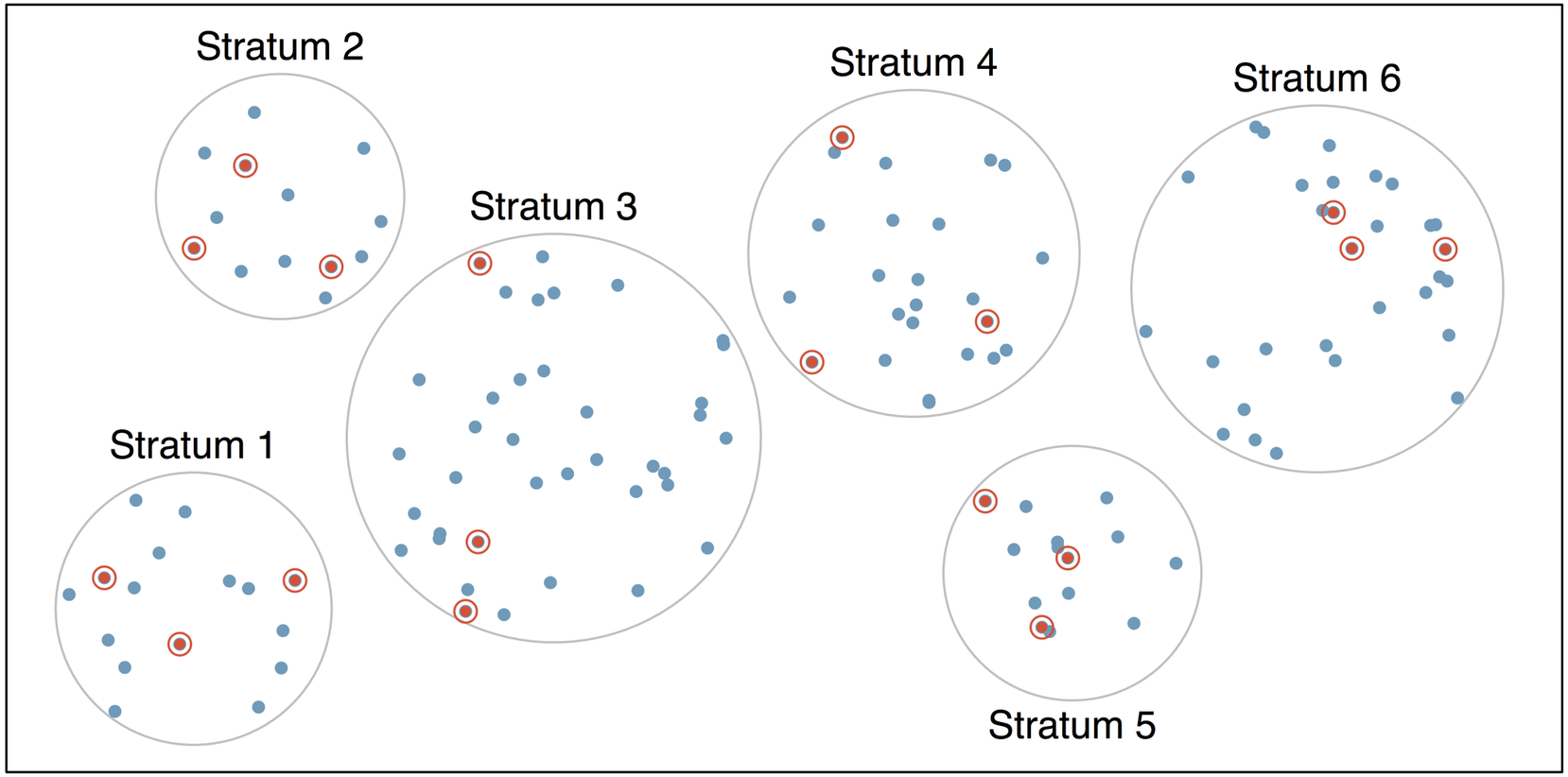
- Conducting a census is very resource intensive
- (Nearly) impossible to collect data from all individuals, hence no guarantee of unbiased results
- Populations constantly change

Sampling is natural

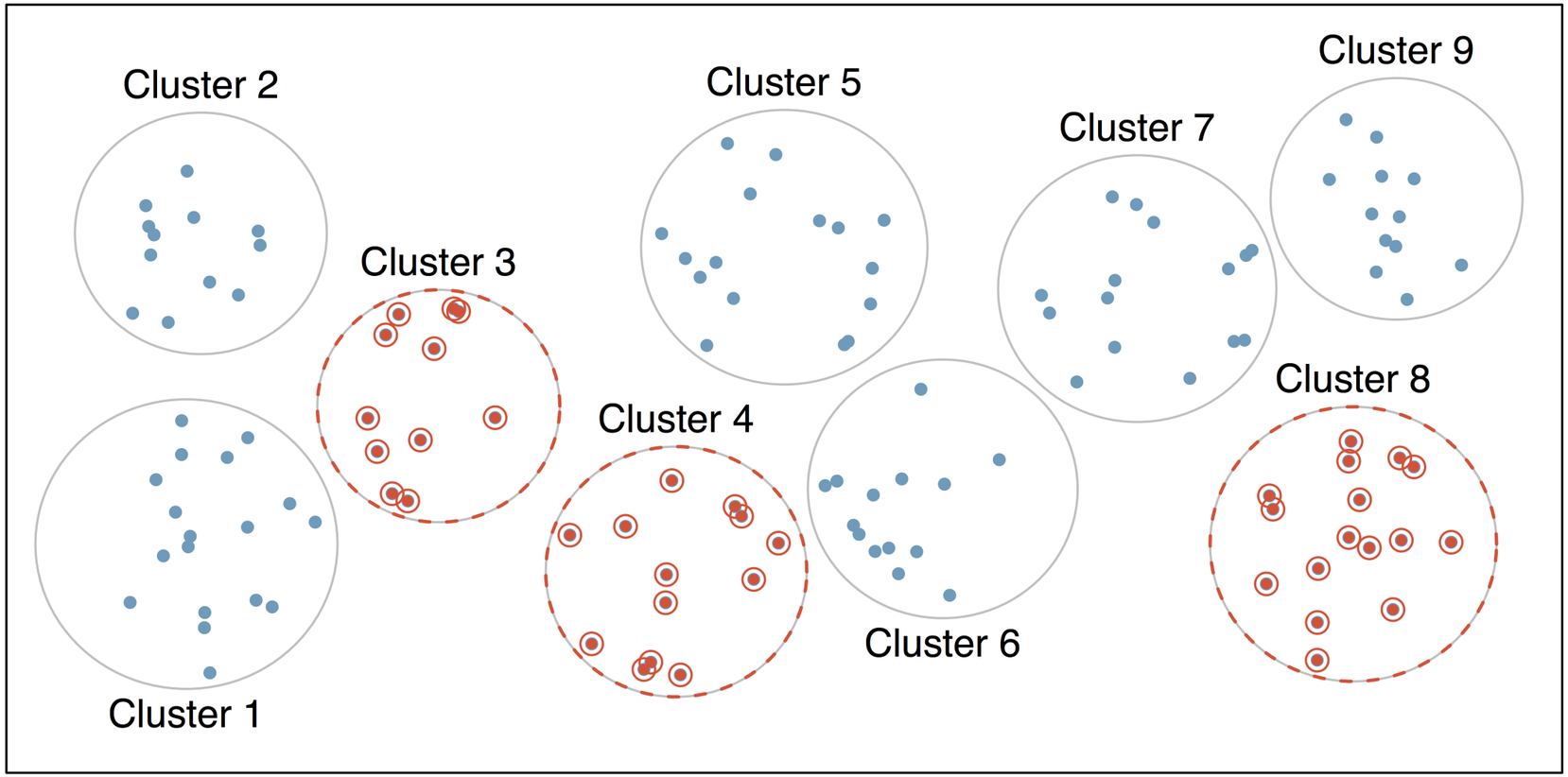
Simple random sample



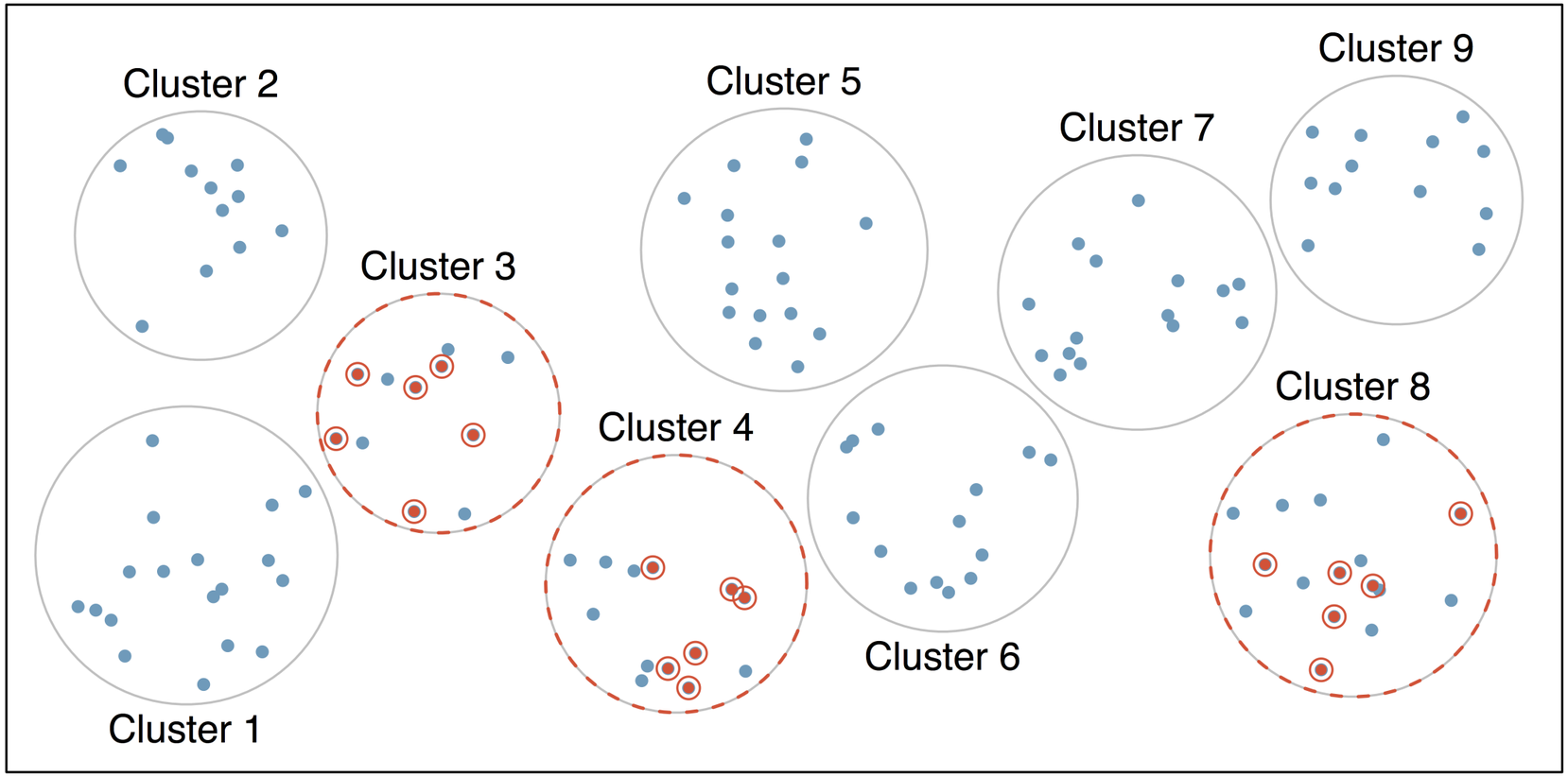
Stratified sample



Cluster sample



Multistage sample



Let's practice!

INTRODUCTION TO DATA IN R

Sampling in R

INTRODUCTION TO DATA IN R

Mine Cetinkaya-Rundel

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

datacamp

Setup

```
# Load packages
library(openintro)
library(dplyr)
# Load county data
data(county)
# Remove DC
county_noDC <- county %>%
    filter(state != "District of Columbia") %>%
    droplevels()
```

Simple random sample

```
# Simple random sample of 150 counties
county_srs <- county_noDC %>%
    sample_n(size = 150)
# Glimpse county_srs
glimpse(county_srs)
```

```
Observations: 150
Variables: 10
$ name <fctr> Clinton County, Muskegon County, D...
$ state <fctr> Ohio, Michigan, Wisconsin, Iowa, U...
$ pop2000 <dbl> 40543, 170200, 43287, 36051, 8238, ...
$ pop2010 <dbl> 42040, 172188, 44159, 35625, 10246,...
$ fed_spend <dbl> 7.444, 7.360, 8.325, 10.616, 7.839,...
$ poverty <dbl> 14.0, 18.0, 12.8, 16.2, 10.5, 17.3,...
$ homeownership <dbl> 70.2, 75.7, 69.8, 76.5, 82.7, 71.4,...
$ multiunit <dbl> 16.7, 14.3, 20.1, 13.9, 7.0, 16.9, ...
$ income <dbl> 22163, 19719, 24552, 22376, 18193, ...
$ med_income <dbl> 46261, 40670, 43127, 40093, 53225, ...
```

SRS state distribution

```
# State distribution of SRS counties
county_srs %>%
    group_by(state) %>%
    count()
```

```
# A tibble: 45 × 2
state n
<fctr> <int>
1 Alabama 2
2 Alaska 1
3 Arizona 1
4 Arkansas 3
5 California 4
6 Colorado 2
# ... with 39 more rows
```

Stratified sample

```
# Stratified sample of 150 counties, each state is a stratum
county_str <- county_noDC %>%
    group_by(state) %>%
    sample_n(size = 3)
# State distribution of stratified sample counties
glimpse(county_str)
```

```
Observations: 150
Variables: 10
$ name <fctr> Bibb County, Washington County, Da...
$ state <fctr> Alabama, Alabama, Alabama, Alaska,...
$ pop2000 <dbl> 20826, 18097, 49129, 13913, 9196, 6...
$ pop2010 <dbl> 22915, 17581, 50251, 13592, 9492, 5...
$ fed_spend <dbl> 7.122, 7.830, 25.775, 12.703, 25.94...
$ poverty <dbl> 12.6, 19.7, 14.8, 10.9, 24.6, 23.6,...
$ homeownership <dbl> 82.9, 83.0, 61.2, 59.2, 56.2, 69.1,...
$ multiunit <dbl> 6.6, 2.6, 13.2, 25.9, 17.4, 2.9, 22...
$ income <dbl> 19918, 18824, 21722, 26413, 20549, ...
$ med_income <dbl> 41770, 36431, 43353, 60776, 53899, ...
```

Let's practice!

INTRODUCTION TO DATA IN R

Principles of experimental design

INTRODUCTION TO DATA IN R

Mine Cetinkaya-Rundel

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

datacamp

Principles of experimental design

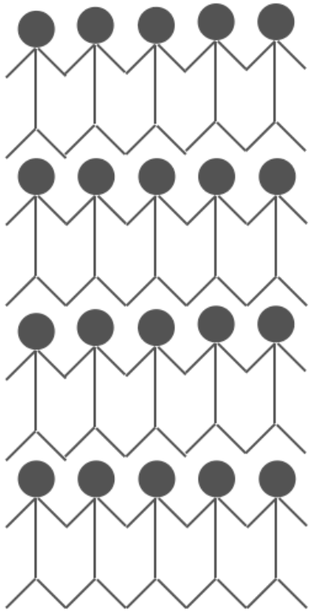
- **Control**: compare treatment of interest to a control group
- **Randomize**: randomly assign subjects to treatments
- **Replicate**: collect a sufficiently large sample within a study, or replicate the entire study
- **Block**: account for the potential effect of confounding variables
  - Group subjects into blocks based on these variables
  - Randomize within each block to treatment groups

Design a study, with blocking

Learning R: lecture or online

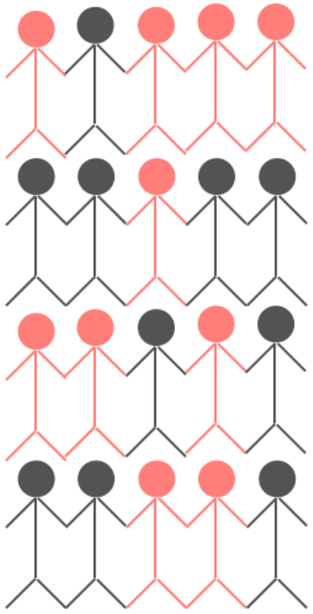
Design a study, with blocking

Learning R: lecture or online

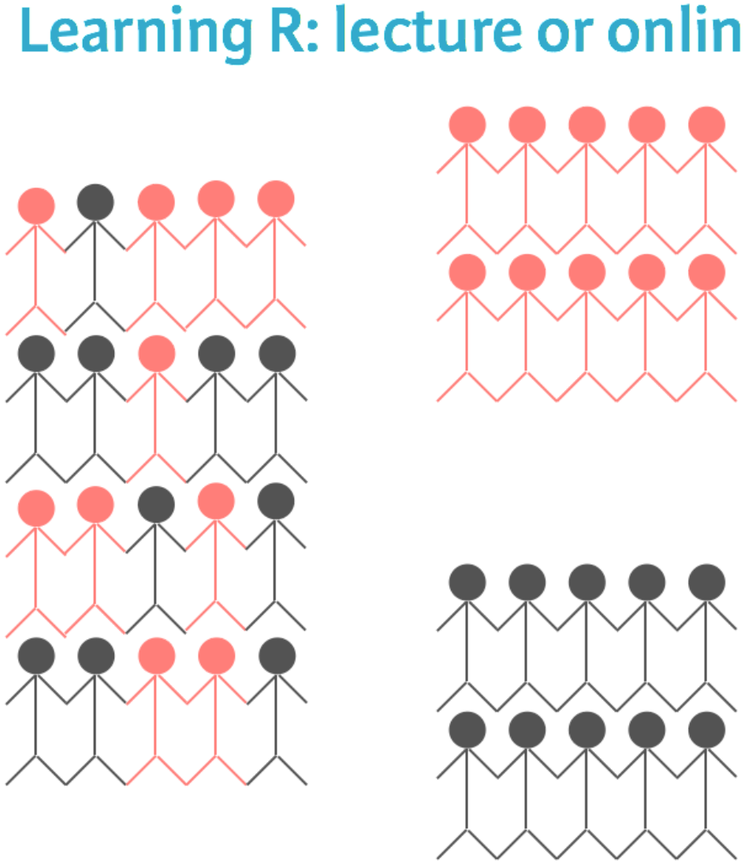


Design a study, with blocking

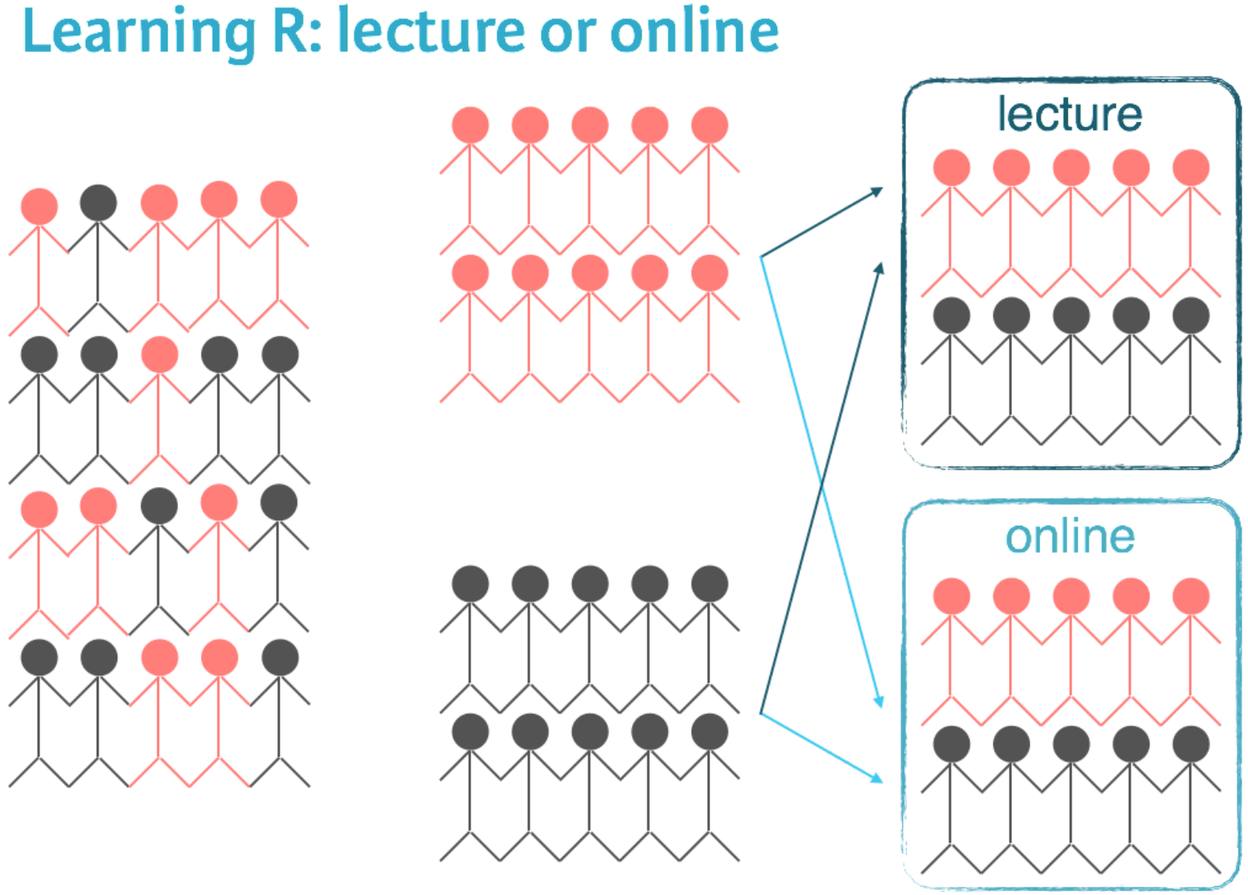
Learning R: lecture or online



Design a study, with blocking



Design a study, with blocking



Let's practice!

INTRODUCTION TO DATA IN R