

What is cluster analysis?

CLUSTER ANALYSIS IN R



Dmitriy (Dima) Gorenshteyn

Lead Data Scientist,

What is clustering?



What is clustering?



























































































































What is clustering?



























































































































What is clustering?

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ♣ | ♥ | ♣ | ♦ | ♠ | ♠ | ♦ | ♦ | ♥ | ♥ |
| ♠ | ♠ | ♥ | ♦ | ♥ | ♦ | ♣ | ♣ | ♣ | ♥ |
| ♣ | ♥ | ♣ | ♦ | ♠ | ♠ | ♦ | ♦ | ♥ | ♥ |
| ♥ | ♥ | ♥ | ♦ | ♠ | ♠ | ♣ | ♣ | ♣ | ♠ |
| ♠ | ♠ | ♥ | ♦ | ♥ | ♦ | ♣ | ♣ | ♣ | ♥ |
| ♣ | ♥ | ♣ | ♦ | ♠ | ♠ | ♦ | ♦ | ♥ | ♥ |

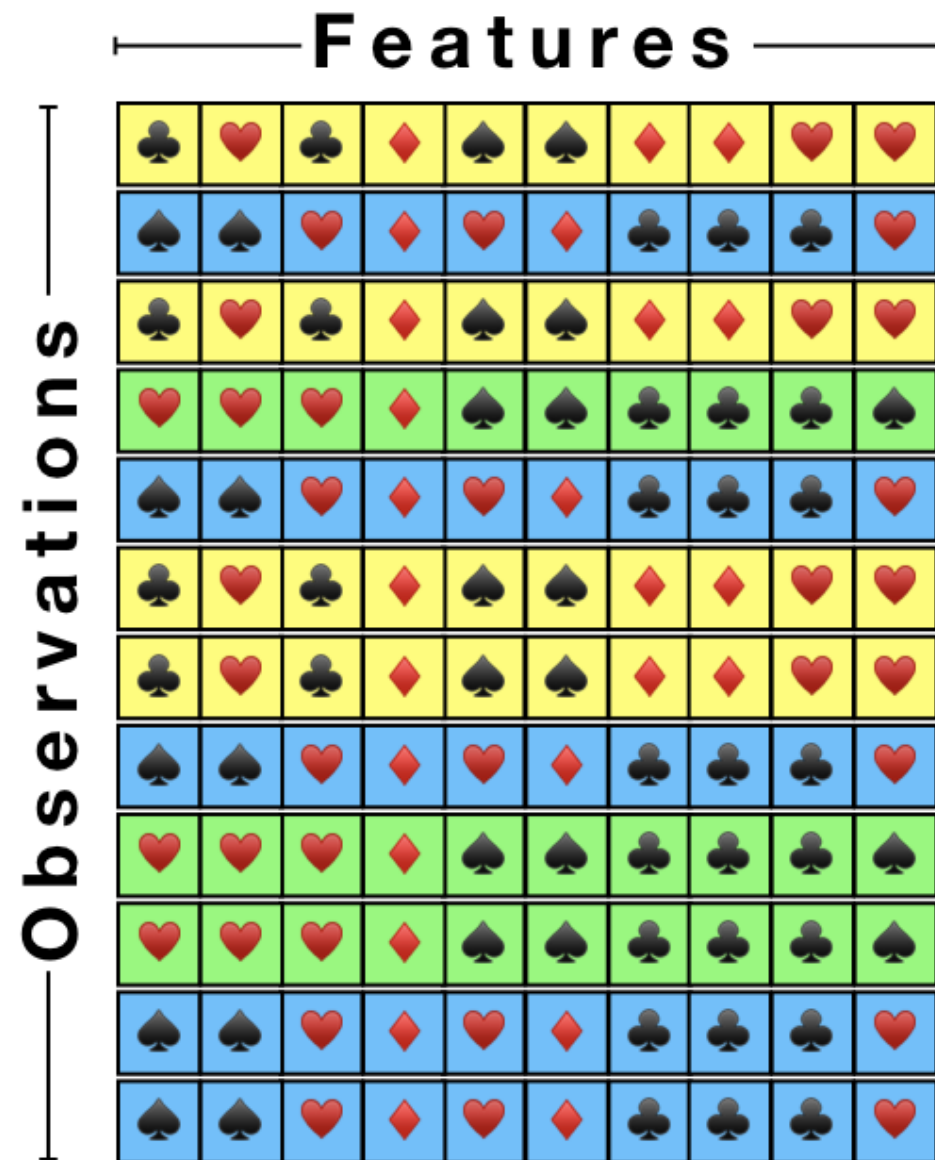
What is clustering?

| | A | B | C | D | E | F | G | H | I | J |
|----|---|---|---|---|---|---|---|---|---|---|
| 1 |  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |  |  |  |  |
| 9 |  |  |  |  |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |  |  |  |  |
| 11 |  |  |  |  |  |  |  |  |  |  |
| 12 |  |  |  |  |  |  |  |  |  |  |

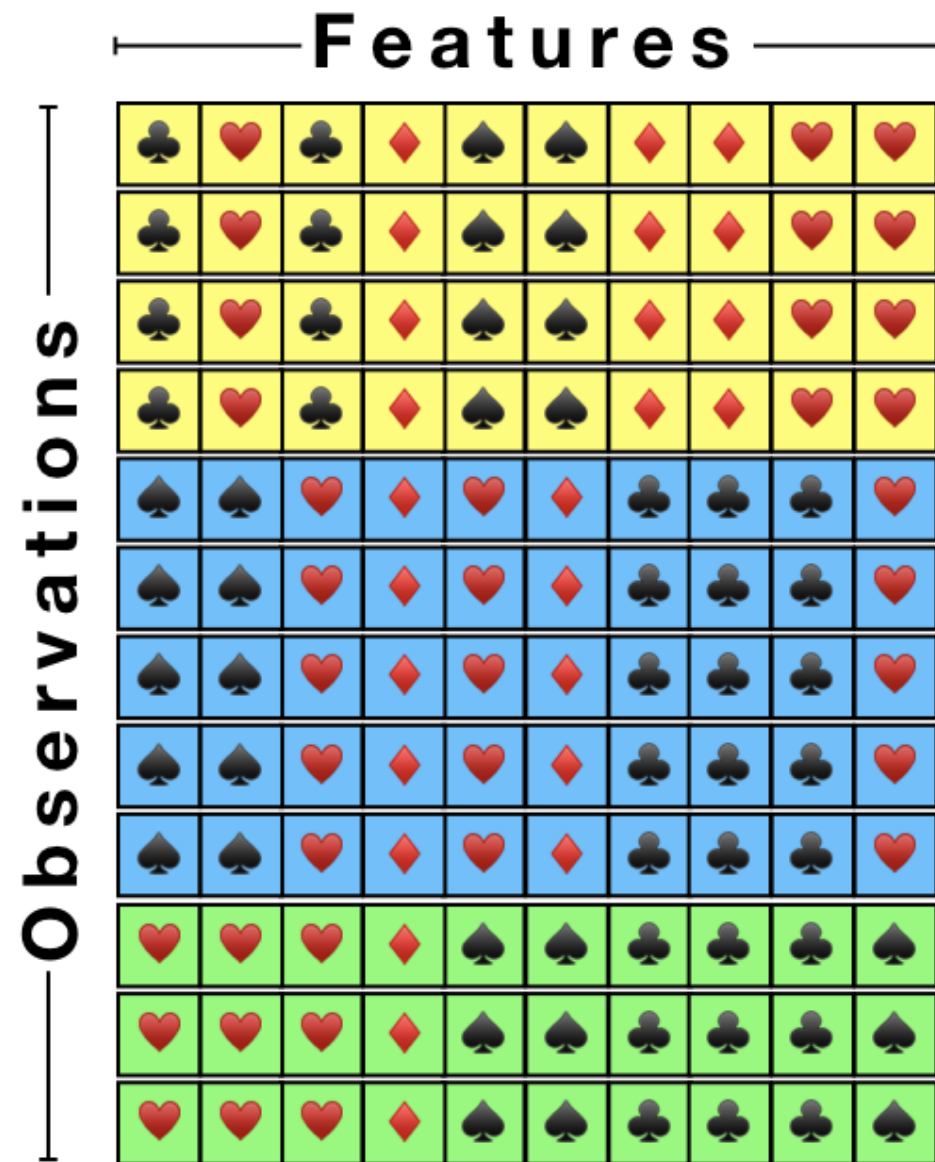
What is clustering?

| Features | | | | | | | | | | |
|--------------|---|---|--|---|---|---|---|---|---|---|
| Observations |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |
| |  |  |  |  |  |  |  |  |  |  |

What is clustering?



What is clustering?



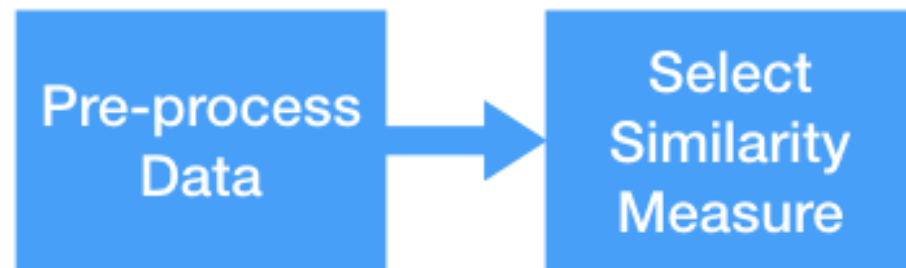
What is clustering?

A form of exploratory data analysis (**EDA**) where **observations** are divided into meaningful groups that share common characteristics (**features**).

The flow of cluster analysis

Pre-process
Data

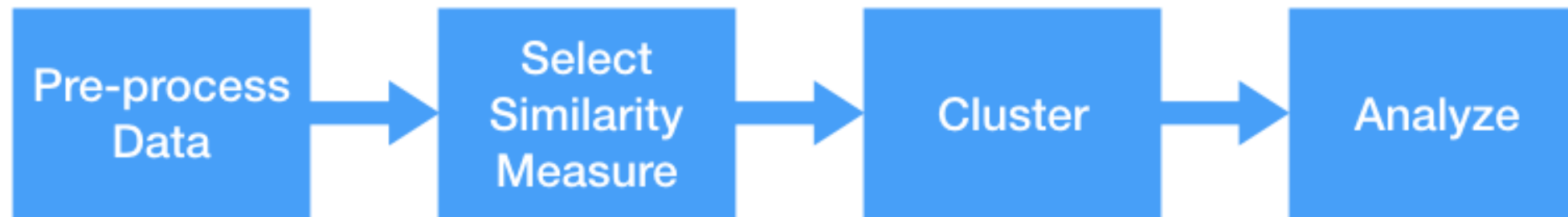
The flow of cluster analysis



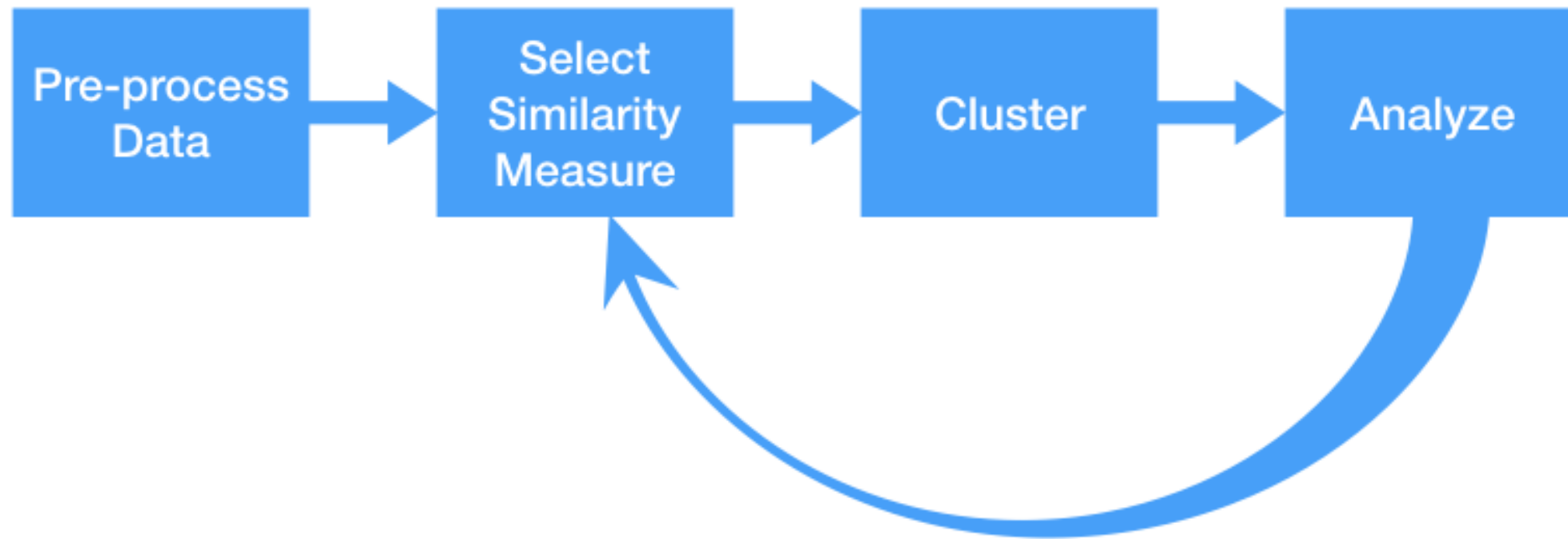
The flow of cluster analysis



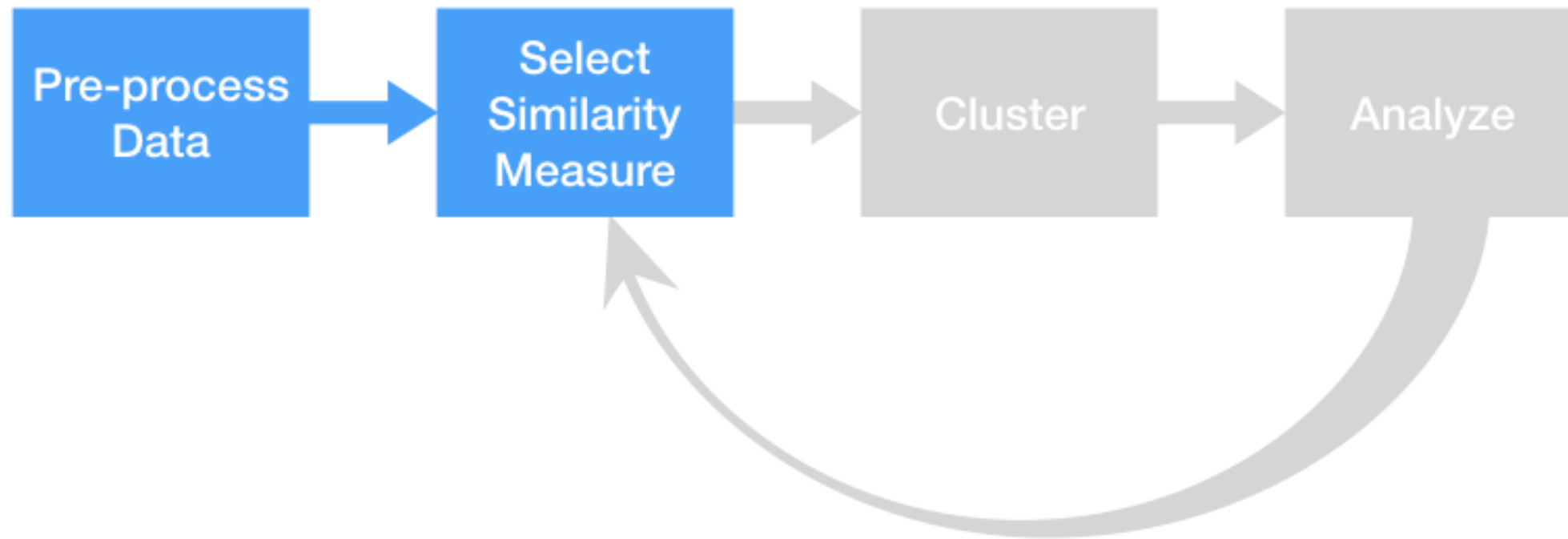
The flow of cluster analysis



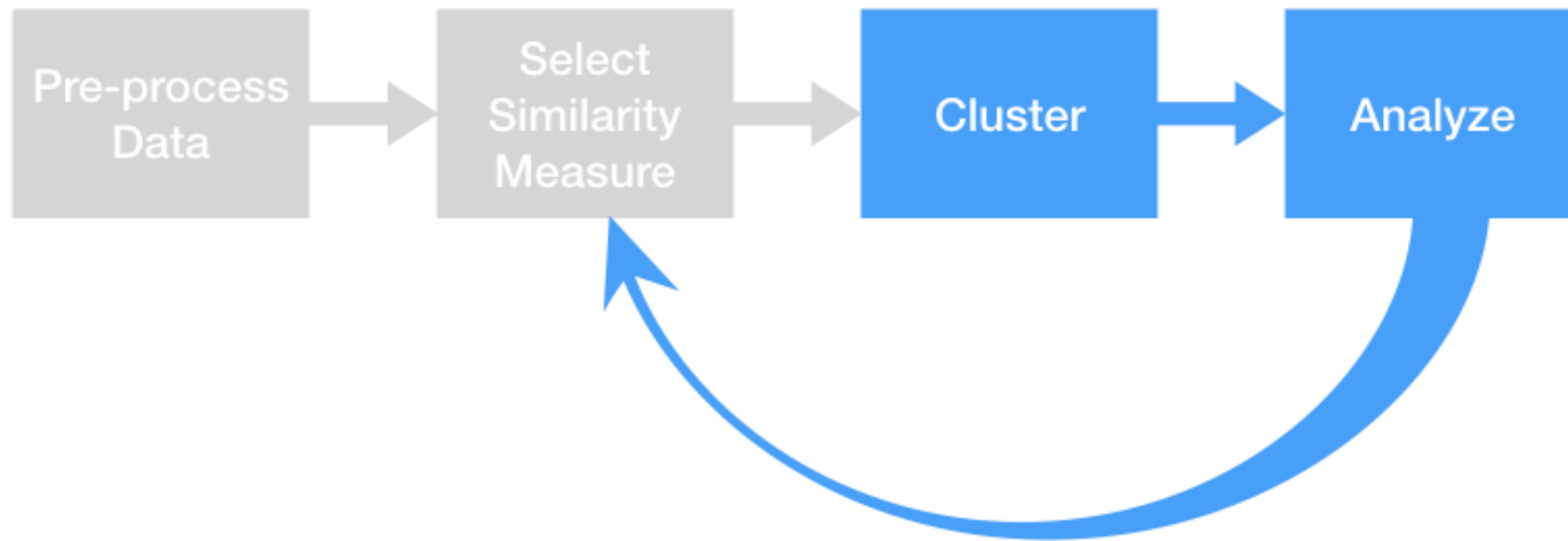
The flow of cluster analysis



Structure of this course



Structure of this course



Let's learn!

CLUSTER ANALYSIS IN R

Distance between two observations

CLUSTER ANALYSIS IN R



Dmitriy (Dima) Gorenshteyn

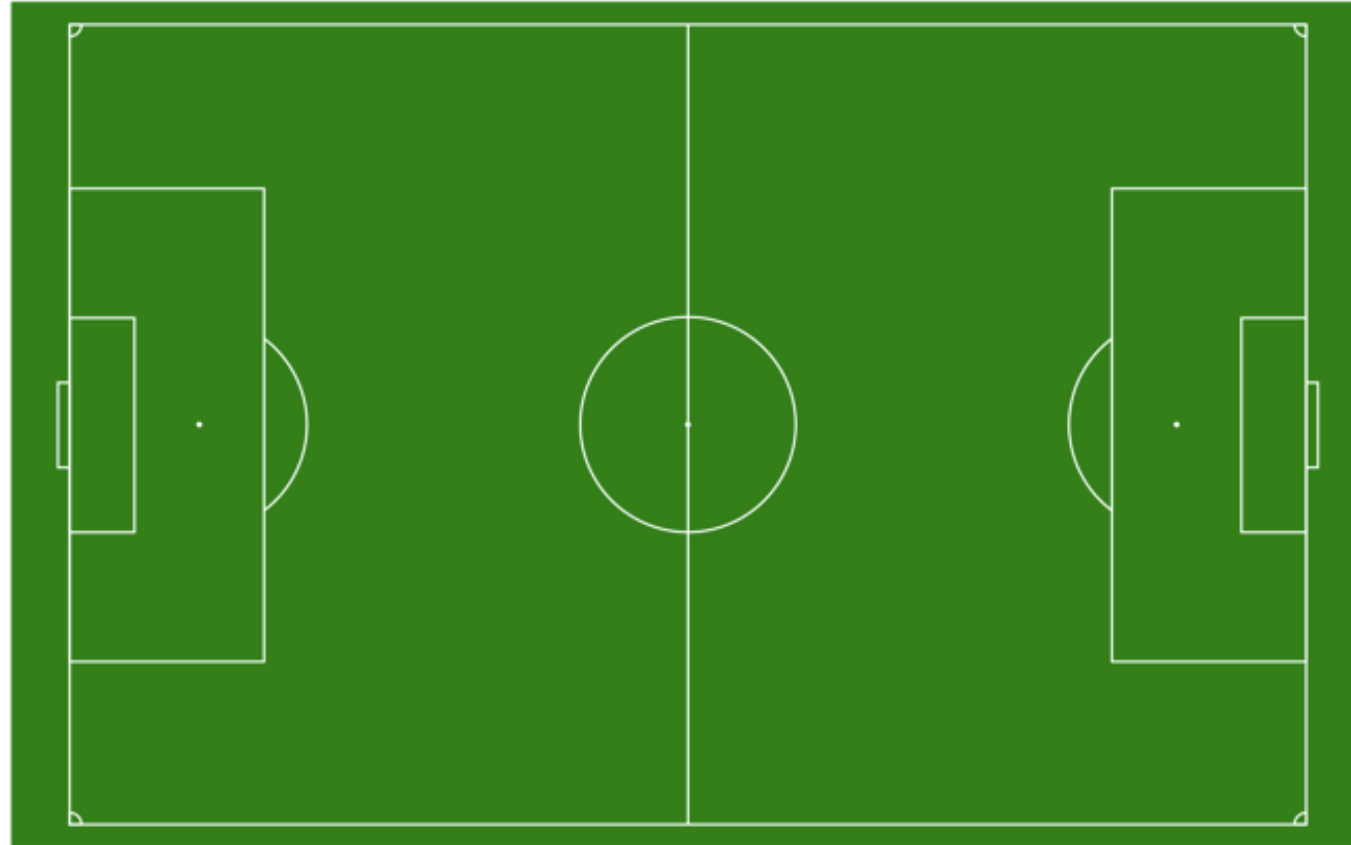
Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

Distance vs Similarity

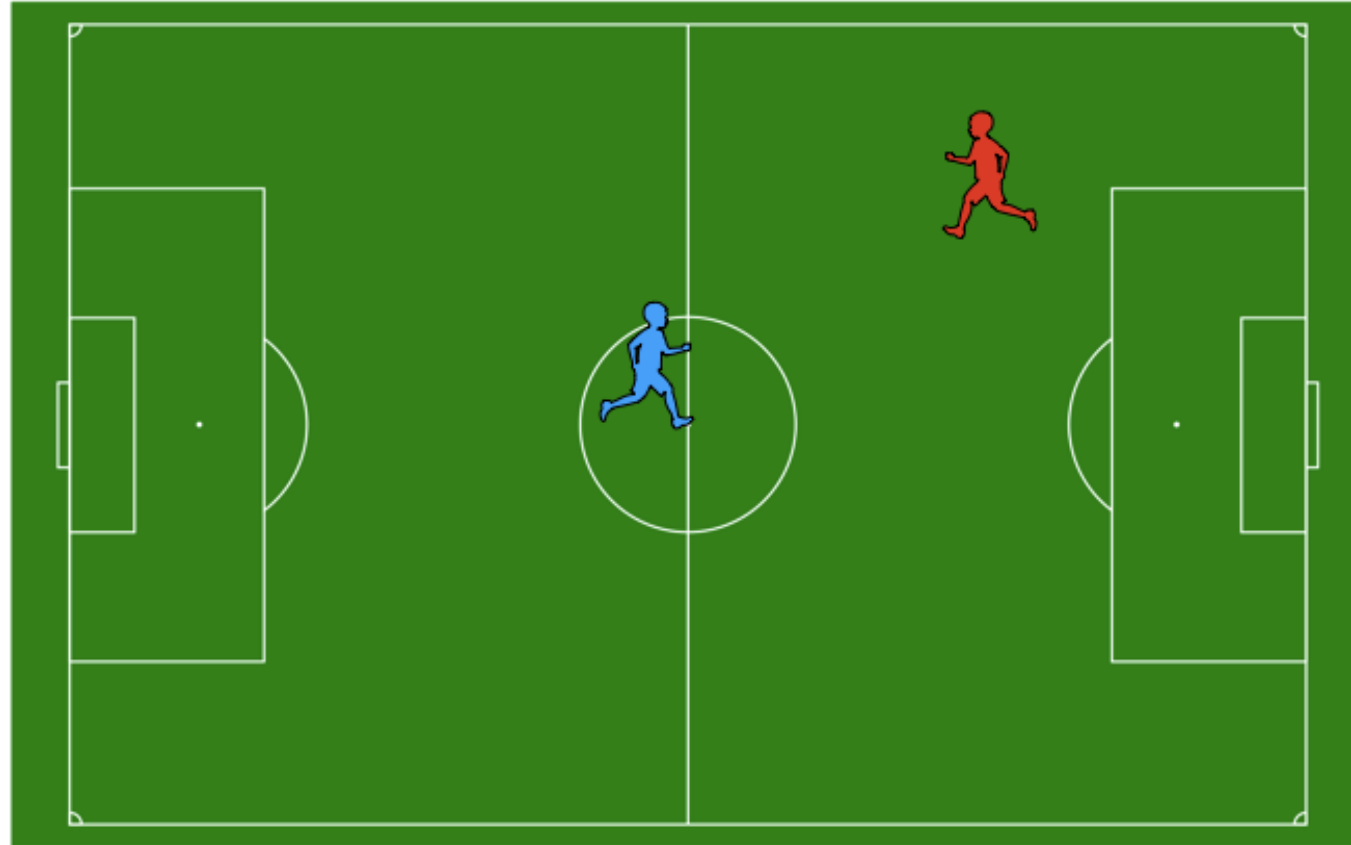
Distance vs Similarity

$$\textit{Distance} = 1 - \textit{Similarity}$$

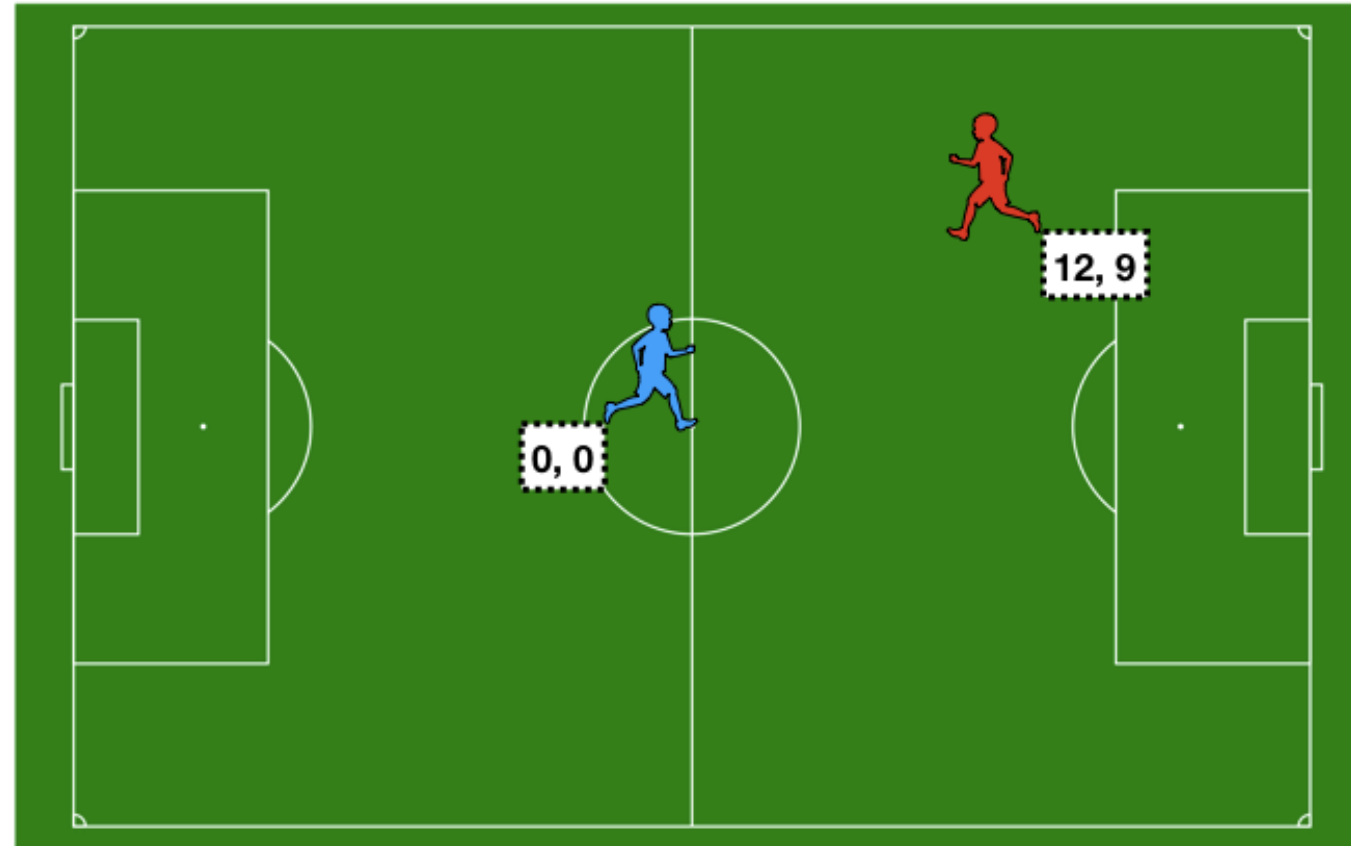
Distance between two players



Distance between two players

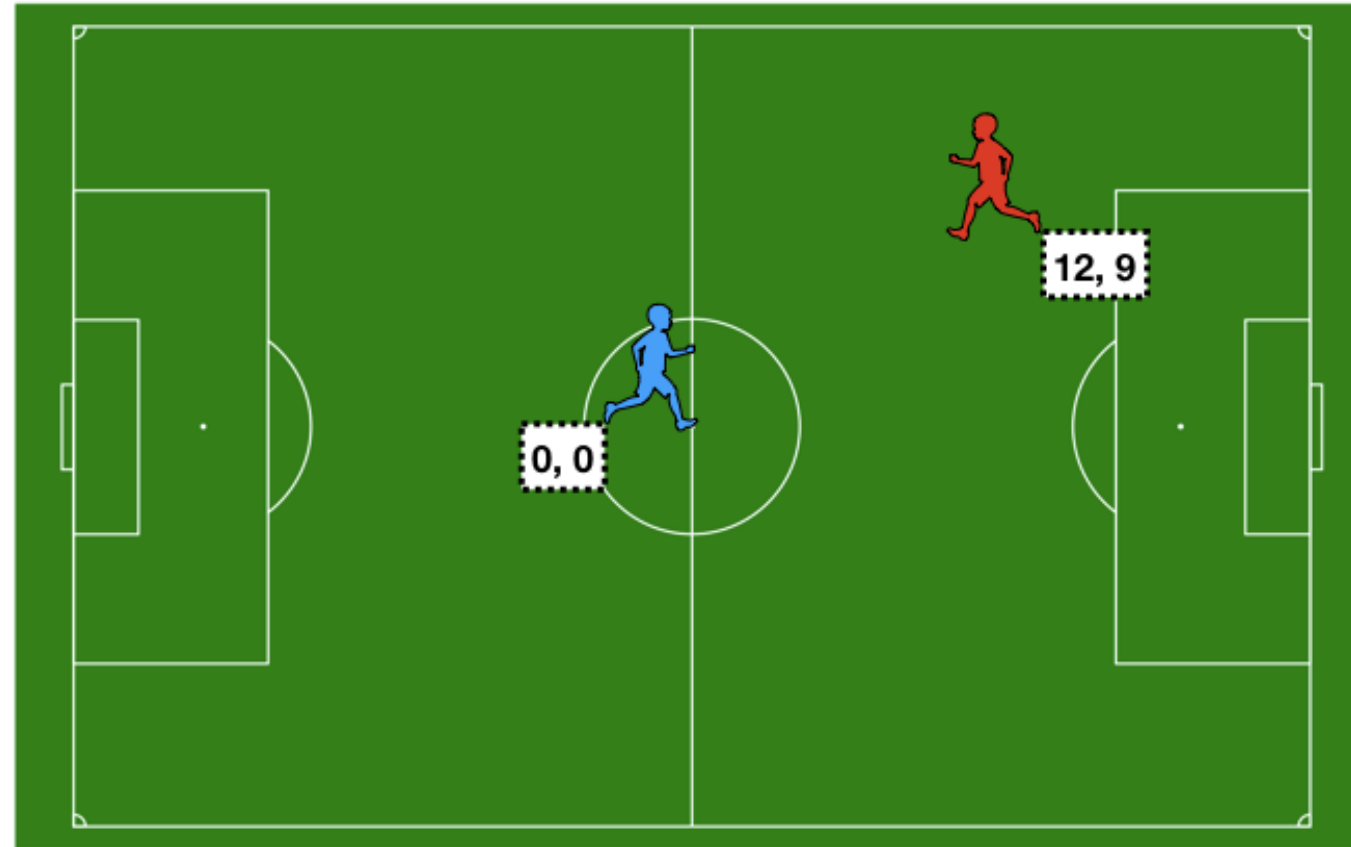


Distance between two players



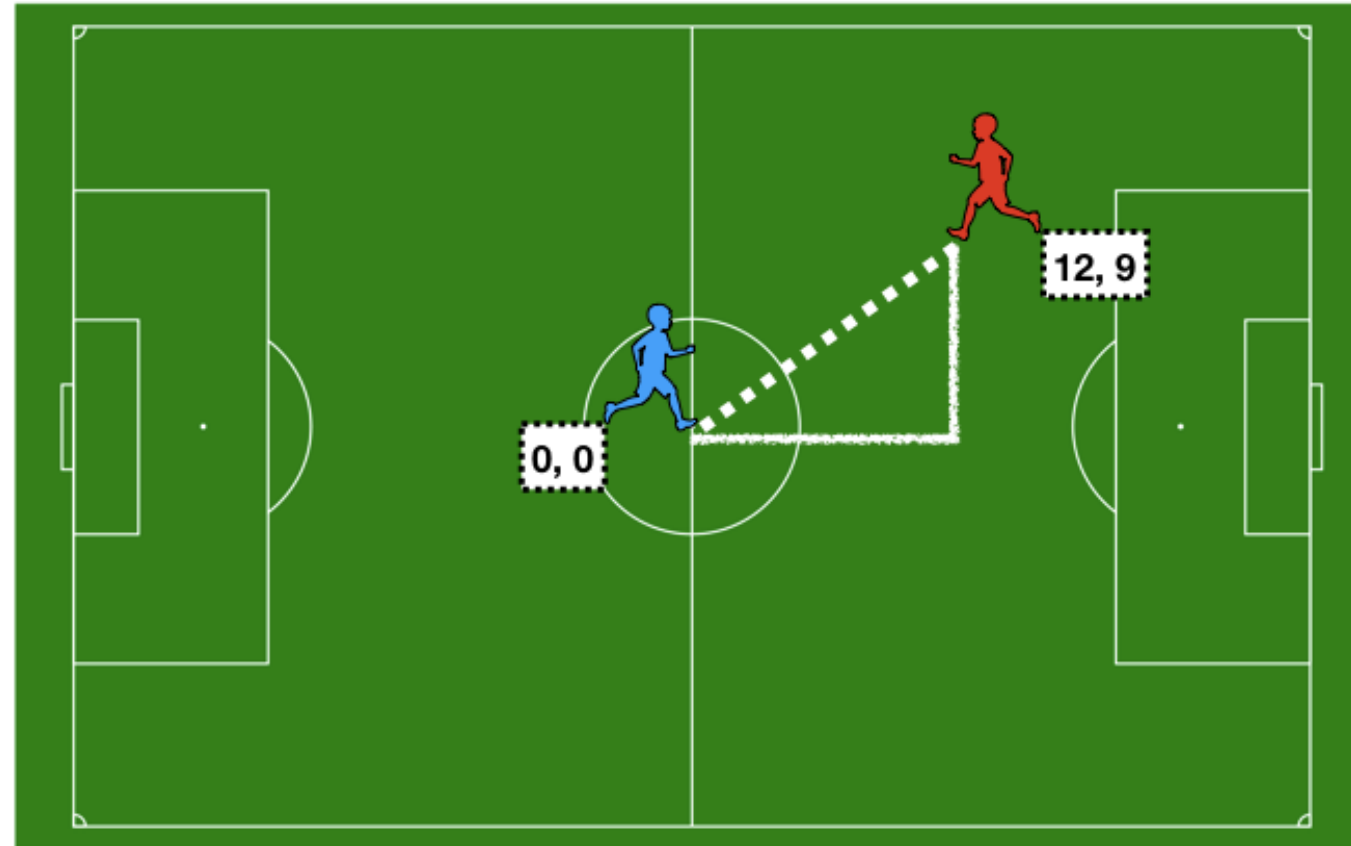
Distance between two players

| | X | Y |
|------|----|---|
| Blue | 0 | 0 |
| Red | 12 | 9 |



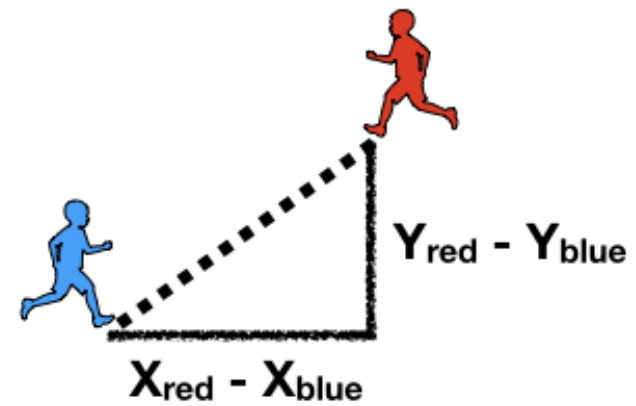
Distance between two players

| | X | Y |
|------|----|---|
| Blue | 0 | 0 |
| Red | 12 | 9 |



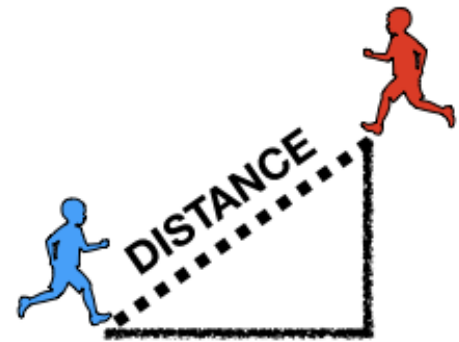
Distance between two players

| | X | Y |
|------|----|---|
| Blue | 0 | 0 |
| Red | 12 | 9 |



Distance between two players

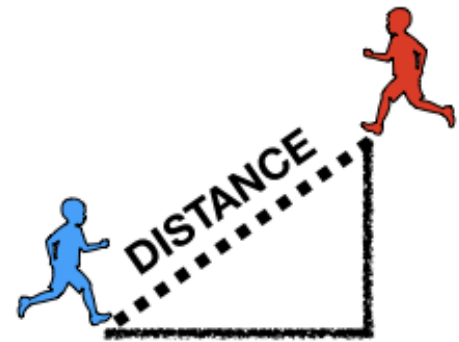
| | X | Y |
|------|----|---|
| Blue | 0 | 0 |
| Red | 12 | 9 |



$$= \sqrt{(X_{\text{red}} - X_{\text{blue}})^2 + (Y_{\text{red}} - Y_{\text{blue}})^2}$$

Distance between two players

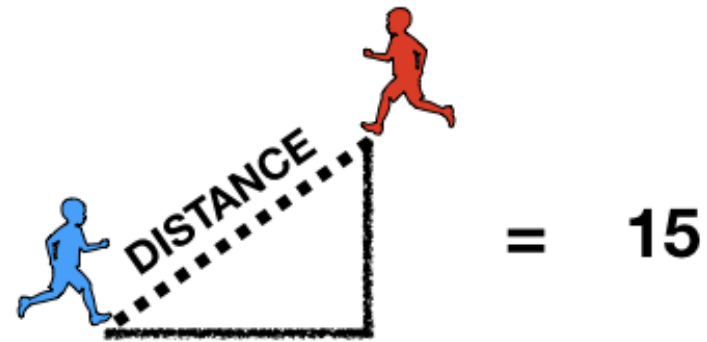
| | X | Y |
|------|----|---|
| Blue | 0 | 0 |
| Red | 12 | 9 |



$$= \sqrt{(12 - 0)^2 + (9 - 0)^2}$$

Distance between two players

| | X | Y |
|------|----|---|
| Blue | 0 | 0 |
| Red | 12 | 9 |



dist() function

```
print(two_players)
```

| | X | Y |
|------|---|----|
| BLUE | 0 | 0 |
| RED | 9 | 12 |

```
dist(two_players, method = 'euclidean')
```

| | BLUE |
|-----|------|
| RED | 15 |

More than 2 observations

```
print(three_players)
```

| | X | Y |
|-------|----|----|
| BLUE | 0 | 0 |
| RED | 9 | 12 |
| GREEN | -2 | 19 |

```
dist(three_players)
```

| | BLUE | RED |
|-------|----------|----------|
| RED | 15.00000 | |
| GREEN | 19.10497 | 13.03840 |

Let's practice!
CLUSTER ANALYSIS IN R

The importance of scale

CLUSTER ANALYSIS IN R



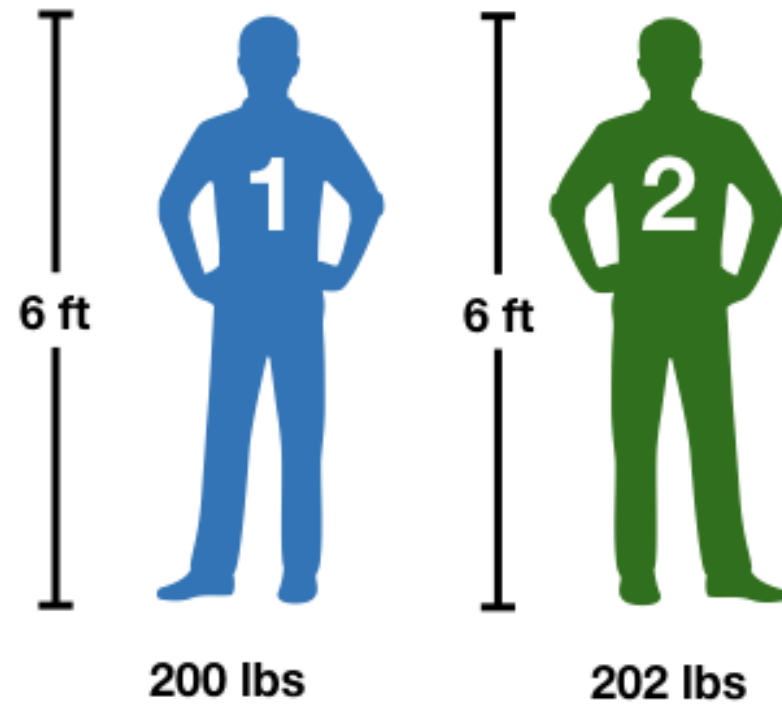
Dmitriy (Dima) Gorenshteyn

Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

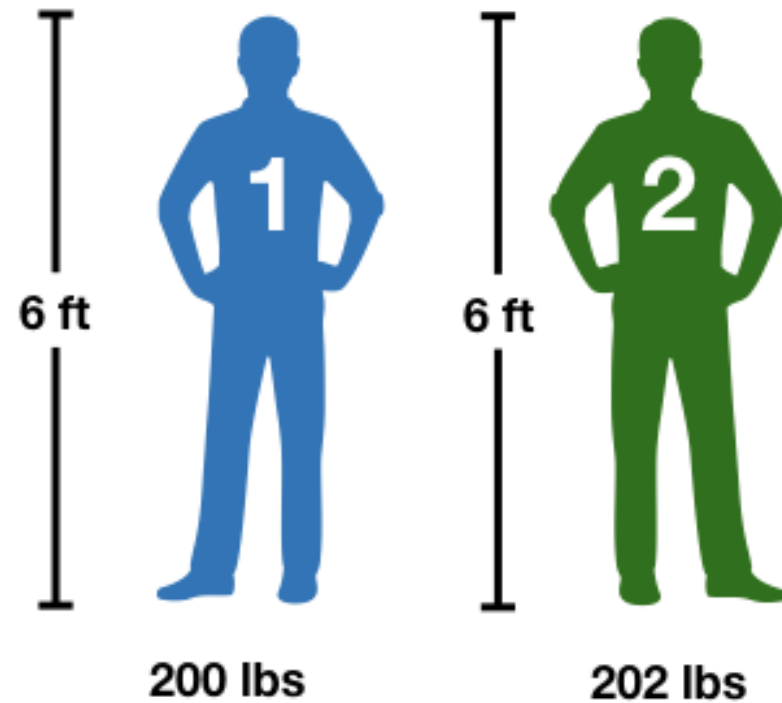
Distance between individuals

| Observation | Height (feet) | Weight (lbs) |
|-------------|---------------|--------------|
| 1 | 6.0 | 200 |
| 2 | 6.0 | 202 |
| 3 | 8.0 | 200 |
| ... | ... | ... |
| ... | ... | ... |

Distance between individuals

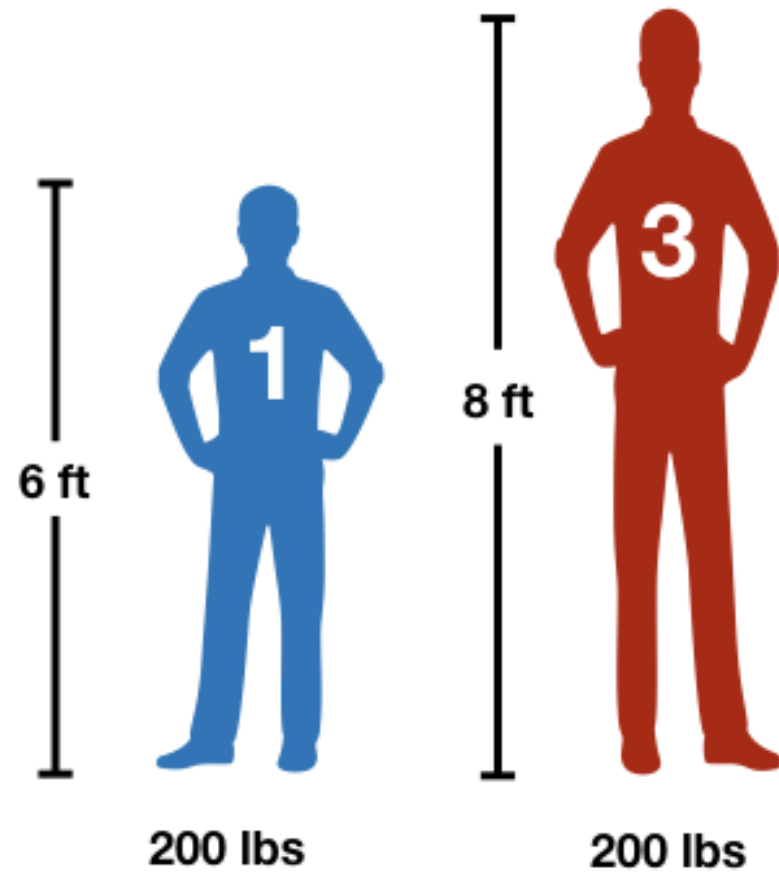


Distance between individuals

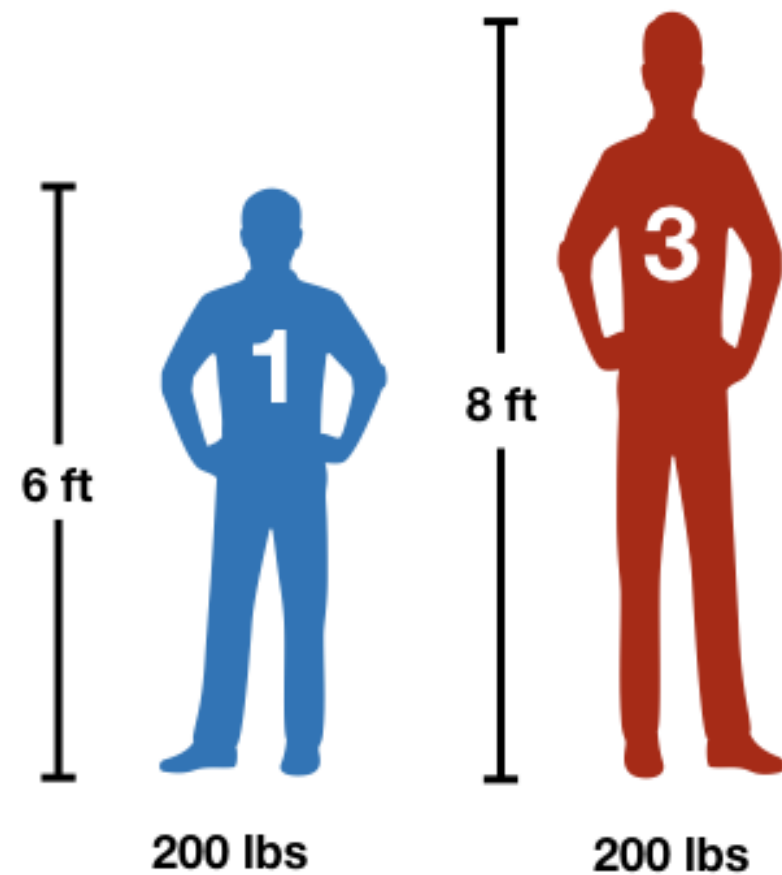


DISTANCE: 2

Distance between individuals

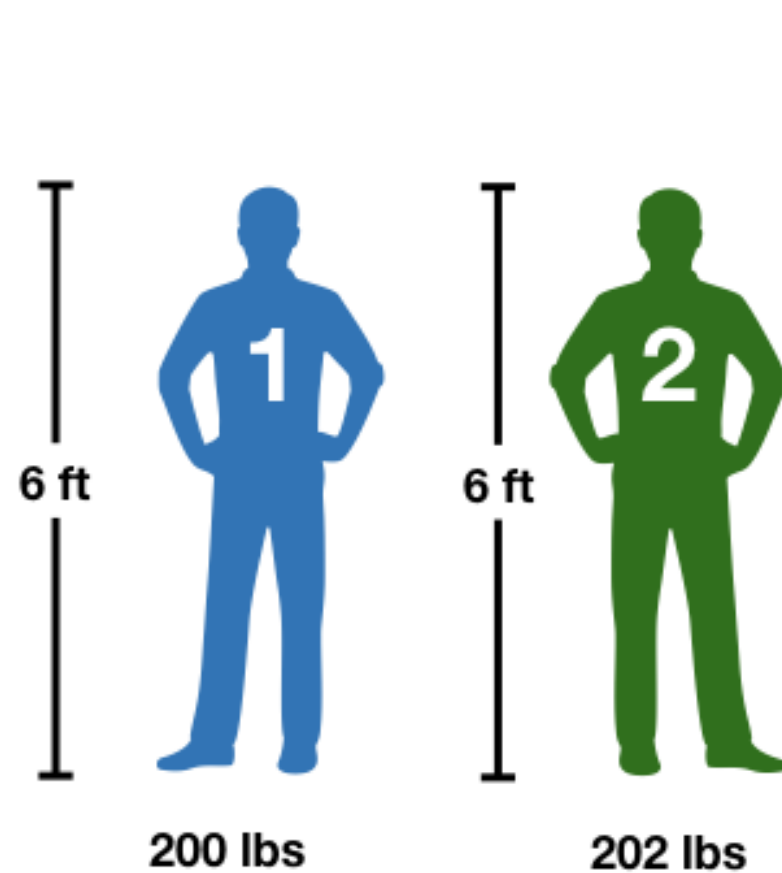


Distance between individuals

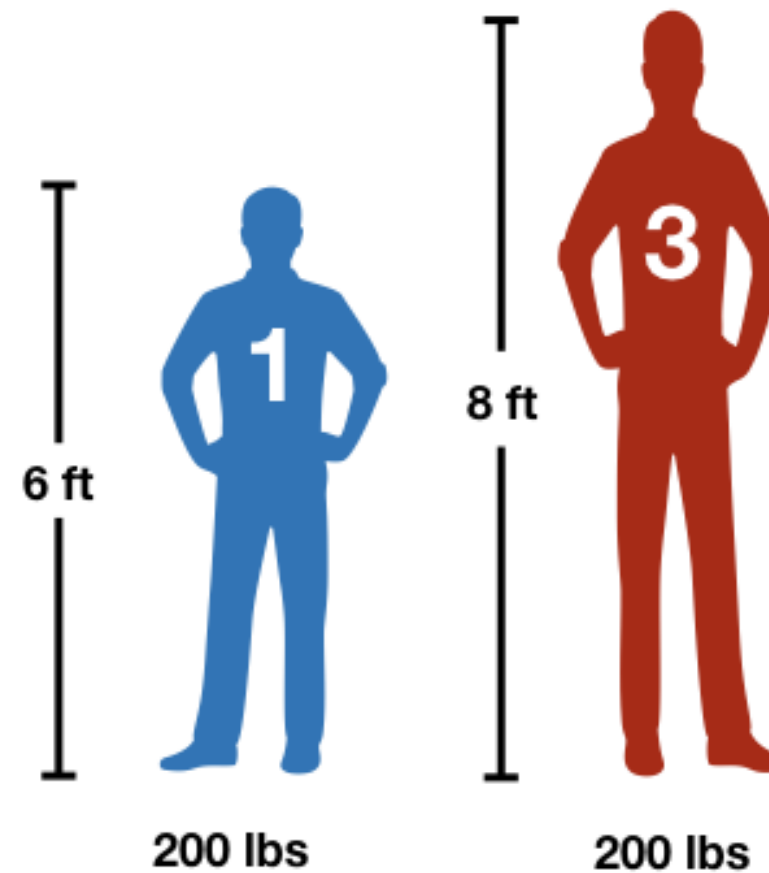


DISTANCE: 2

Distance between individuals



DISTANCE: **2**

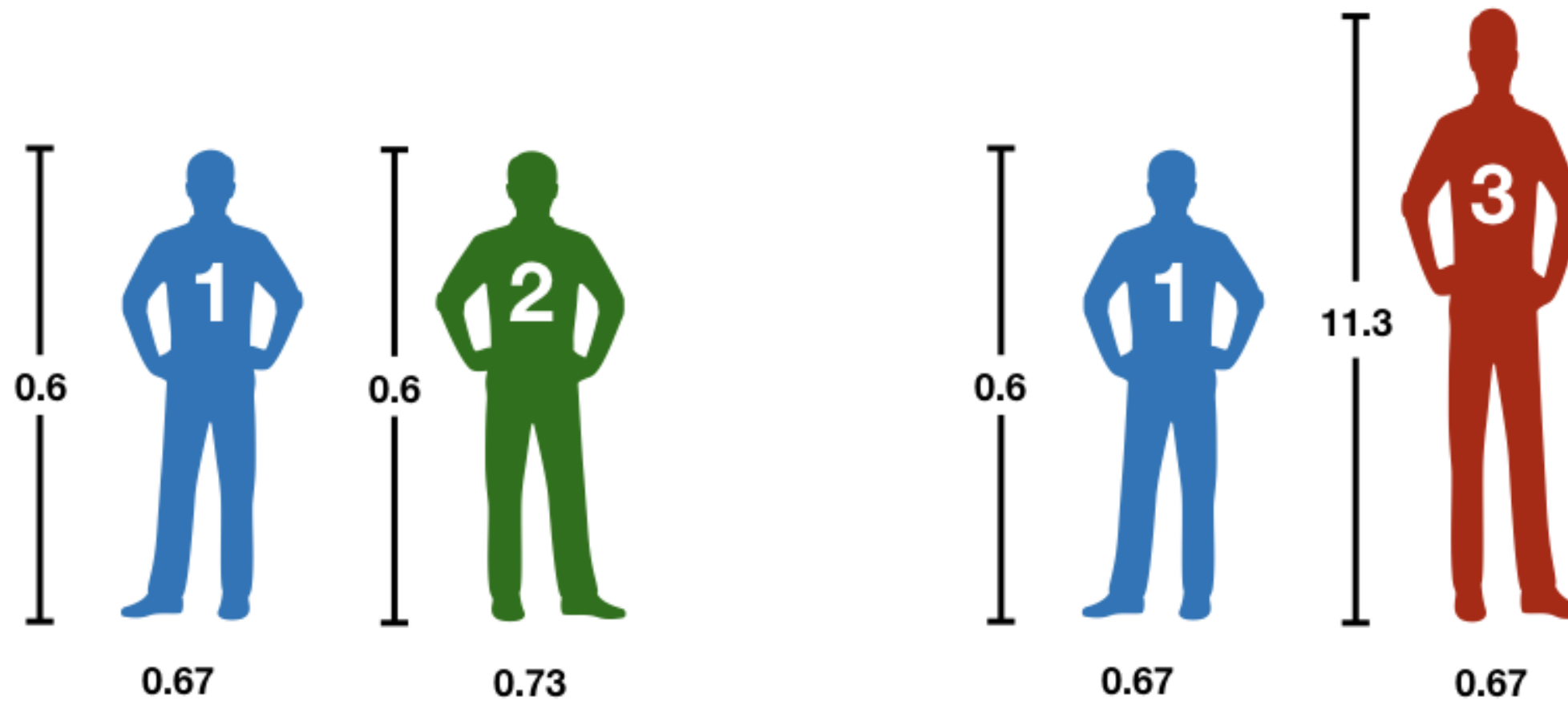


DISTANCE: **2**

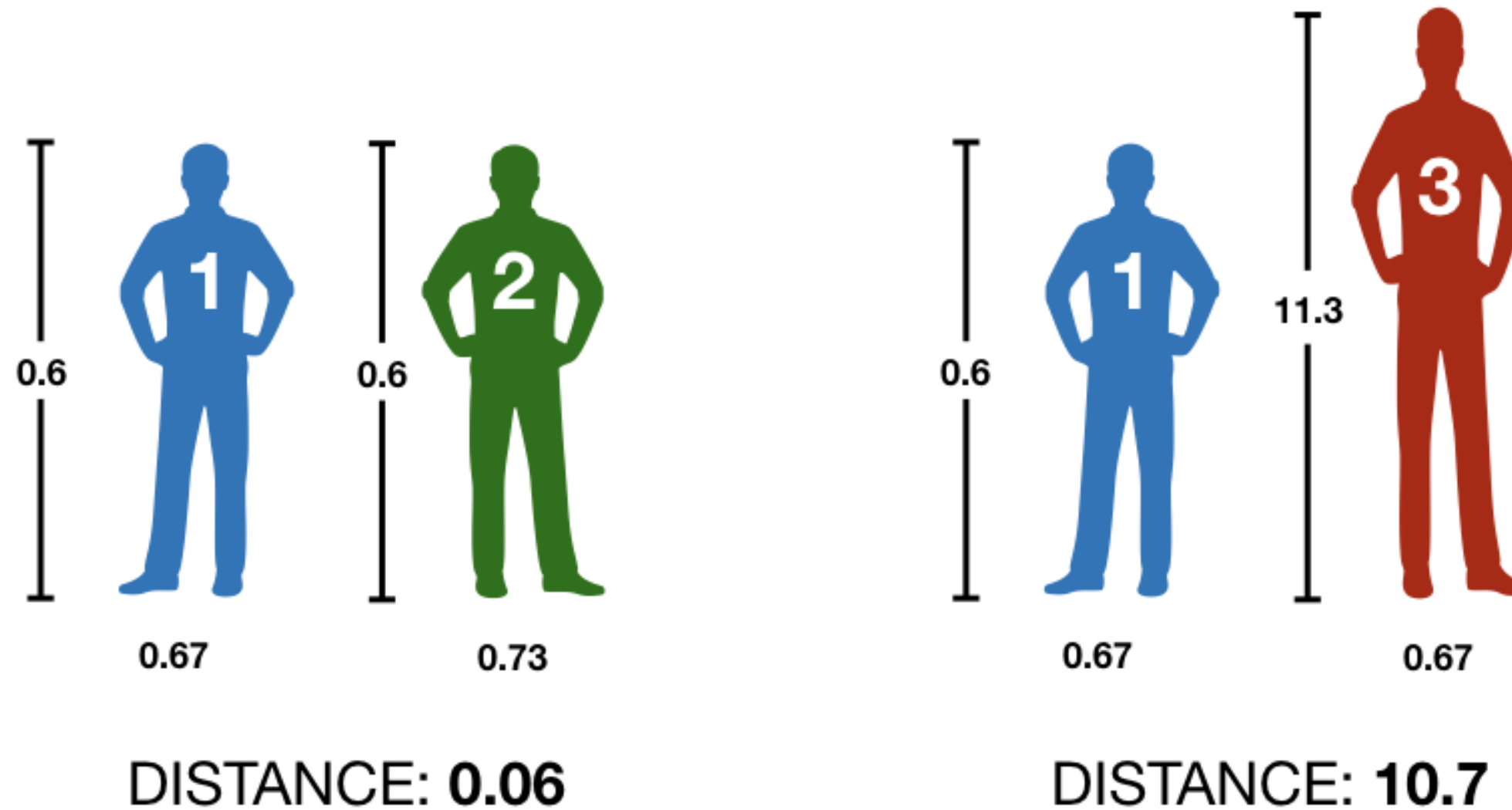
Scaling our features

$$height_{scaled} = \frac{height - \text{mean}(height)}{sd(height)}$$

Distance between individuals



Distance between individuals



scale() function

```
print(height_weight)
```

| | Height | Weight |
|-----|--------|--------|
| 1 | 6 | 200 |
| 2 | 6 | 202 |
| 3 | 8 | 200 |
| ... | ... | ... |

```
scale(height_weight)
```

| | Height | Weight |
|---|--------|--------|
| 1 | 0.60 | 0.67 |
| 2 | 0.60 | 0.73 |
| 3 | 11.3 | 0.67 |

Let's practice!
CLUSTER ANALYSIS IN R

Measuring distance for categorical data

CLUSTER ANALYSIS IN R



Dmitriy (Dima) Gorenshteyn

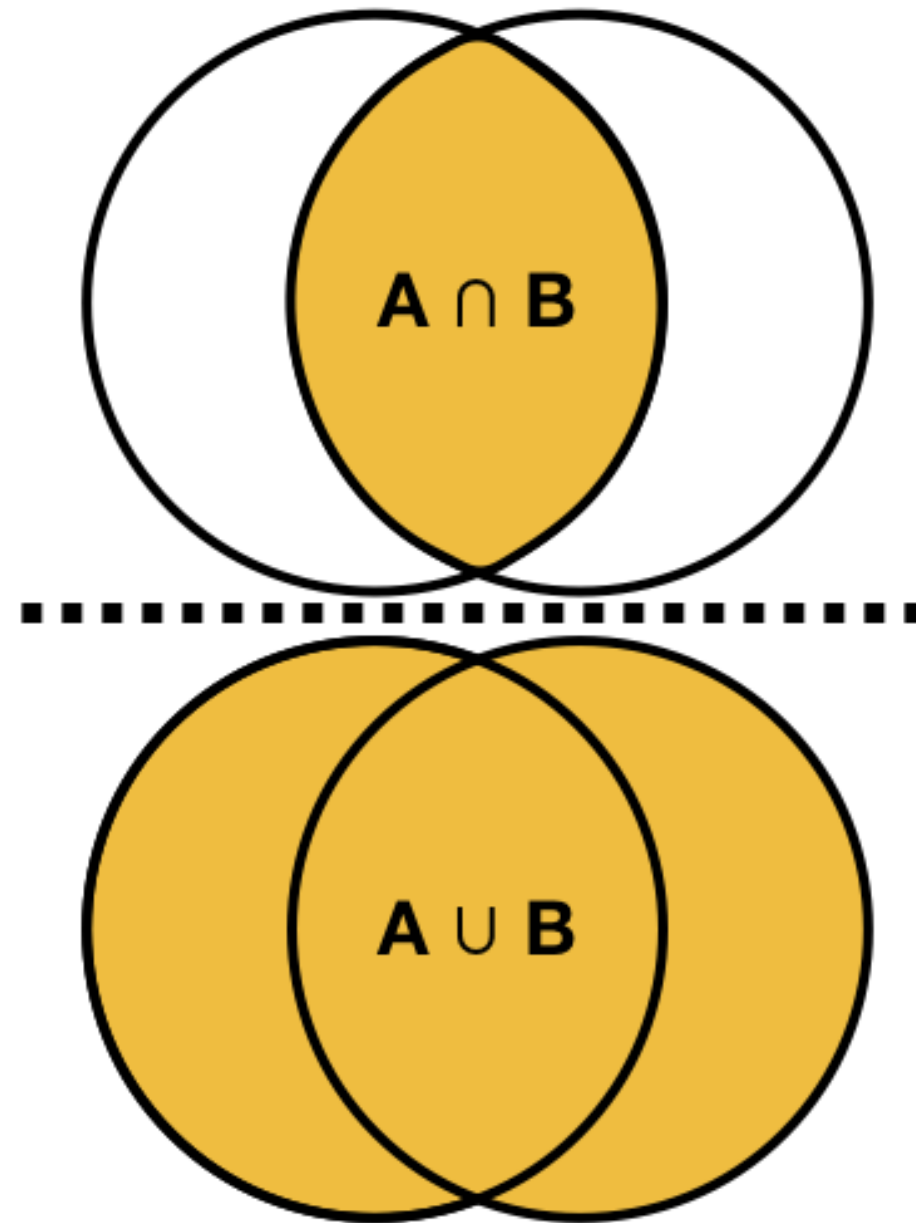
Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

Binary data

| | wine | beer | whiskey | vodka |
|-----|-------|------|---------|-------|
| 1 | TRUE | TRUE | FALSE | FALSE |
| 2 | FALSE | TRUE | TRUE | TRUE |
| ... | ... | ... | ... | ... |

Jaccard index

$$J(A, B) = \frac{A \cap B}{A \cup B}$$



Calculating Jaccard distance

| | wine | beer | whiskey | vodka |
|---|-------|------|---------|-------|
| 1 | TRUE | TRUE | FALSE | FALSE |
| 2 | FALSE | TRUE | TRUE | TRUE |

$$J(1, 2) = \frac{1 \cap 2}{1 \cup 2} = \frac{1}{4} = 0.25$$

$$Distance(1, 2) = 1 - J(1, 2) = 0.75$$

Calculating Jaccard distance in R

```
print(survey_a)

  wine  beer whiskey vodka
<lgl> <lgl>   <lgl> <lgl>
1  TRUE  TRUE    FALSE FALSE
2 FALSE  TRUE    TRUE  TRUE
3  TRUE FALSE    TRUE FALSE
dist(survey_a, method = "binary")

      1      2
2 0.7500000
3 0.6666667 0.7500000
```

More than two categories

| | color | sport |
|---|-------|--------|
| 1 | red | soccer |
| 2 | green | hockey |
| 3 | blue | hockey |
| 4 | blue | soccer |

| | colorblue | colorgreen | colorred | sporthockey | sportsoccer |
|---|-----------|------------|----------|-------------|-------------|
| 1 | 0 | 0 | 1 | 0 | 1 |
| 2 | 0 | 1 | 0 | 1 | 0 |
| 3 | 1 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 0 | 1 |

Dummification in R

```
print(survey_b)
```

```
  color sport  
1  red soccer  
2 green hockey  
3 blue hockey  
4 blue soccer
```

```
library(dummies)
```

```
dummy.data.frame(survey_b)
```

| | colorblue | colorgreen | colorred | sporthockey | sportsoccer |
|---|-----------|------------|----------|-------------|-------------|
| 1 | 0 | 0 | 1 | 0 | 1 |
| 2 | 0 | 1 | 0 | 1 | 0 |
| 3 | 1 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 0 | 1 |

Generalizing categorical distance in R

```
print(survey_b)

  color sport
1  red soccer
2 green hockey
3  blue hockey
4  blue soccer
dummy_survey_b <- dummy.data.frame(survey_b)

dist(dummy_survey_b, method = 'binary')

      1      2      3
2 1.0000000
3 1.0000000 0.6666667
4 0.6666667 1.0000000 0.6666667
```

Let's practice!
CLUSTER ANALYSIS IN R