Introduction to Data Analysis in R (Day 2)

Dr. Behnam Yousefi
Institute of medical systems biology, UKE
March 2024

Hypothesis Testing

T-test

Fisher's exact test

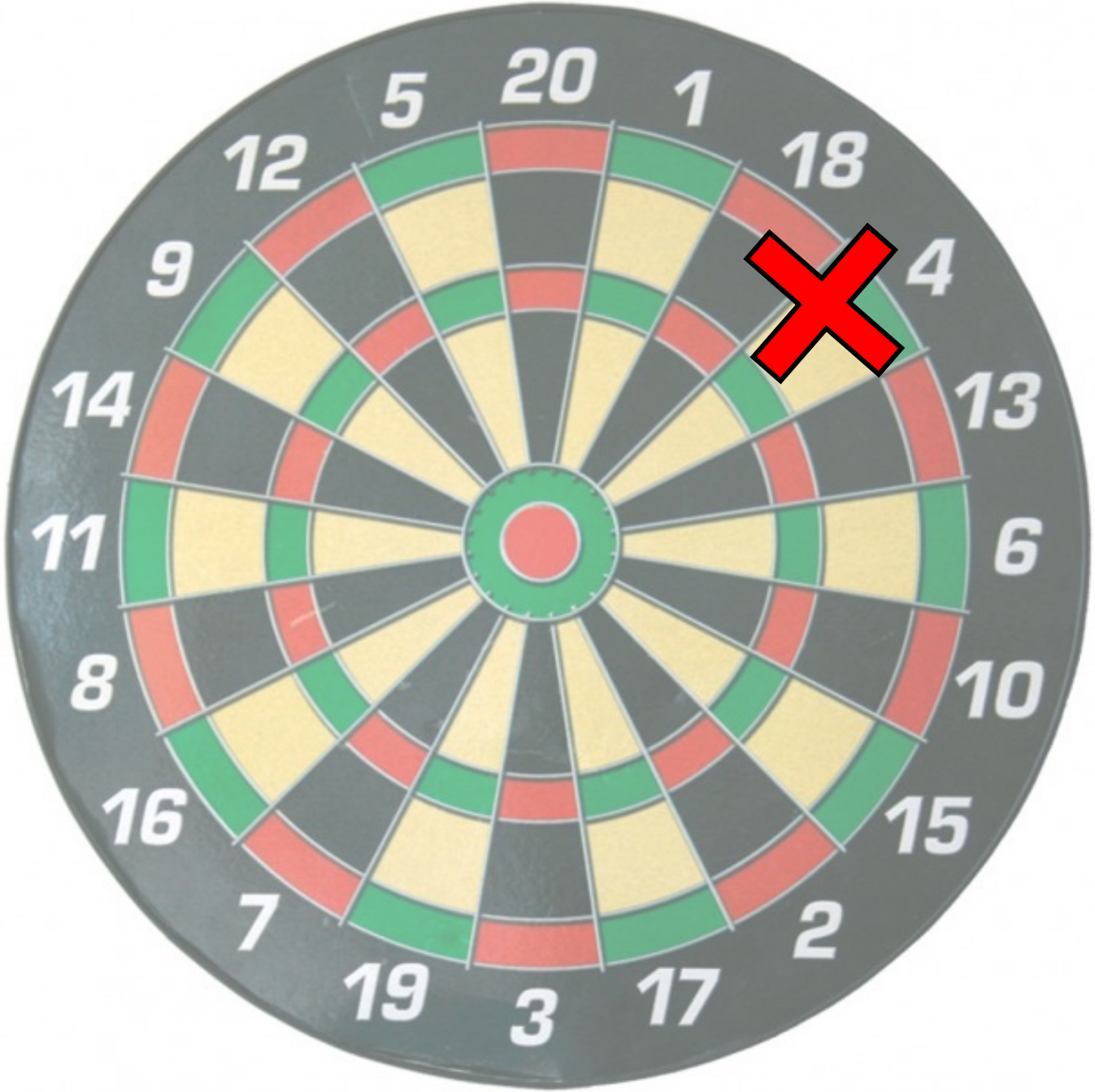
Wilcoxson rank sum test

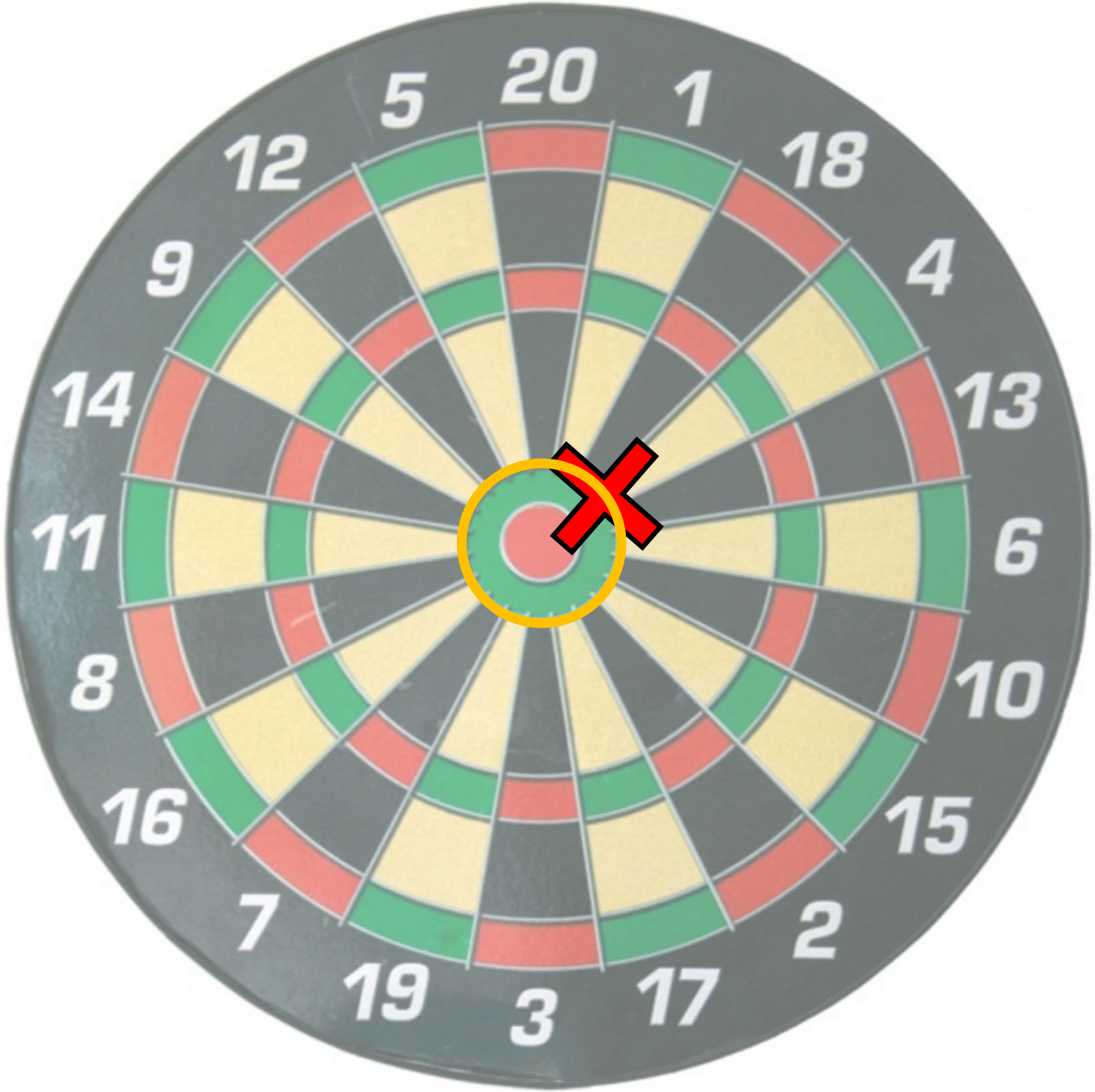
Correlation test

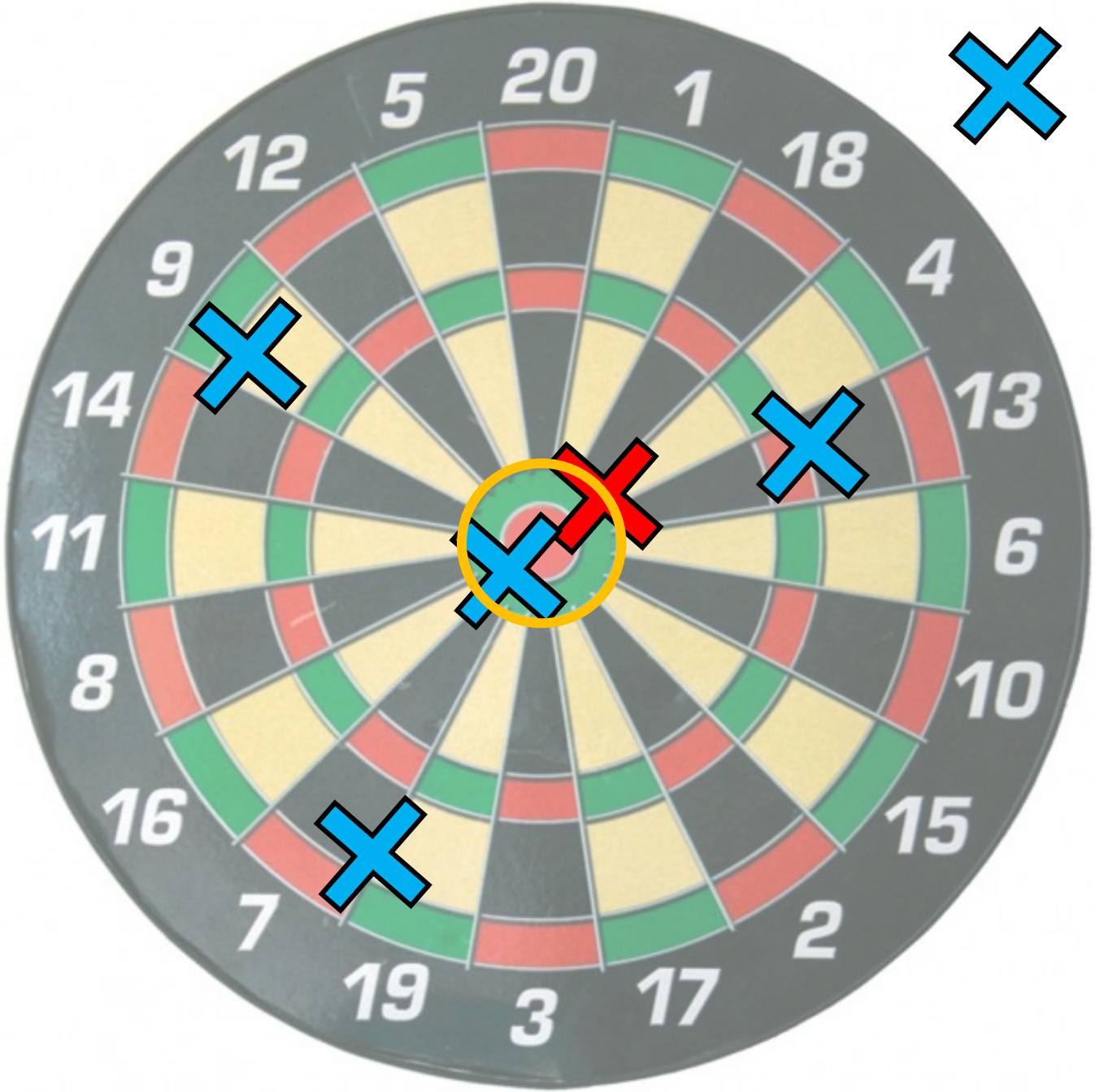
Hypergeometric test

Chi squared test


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

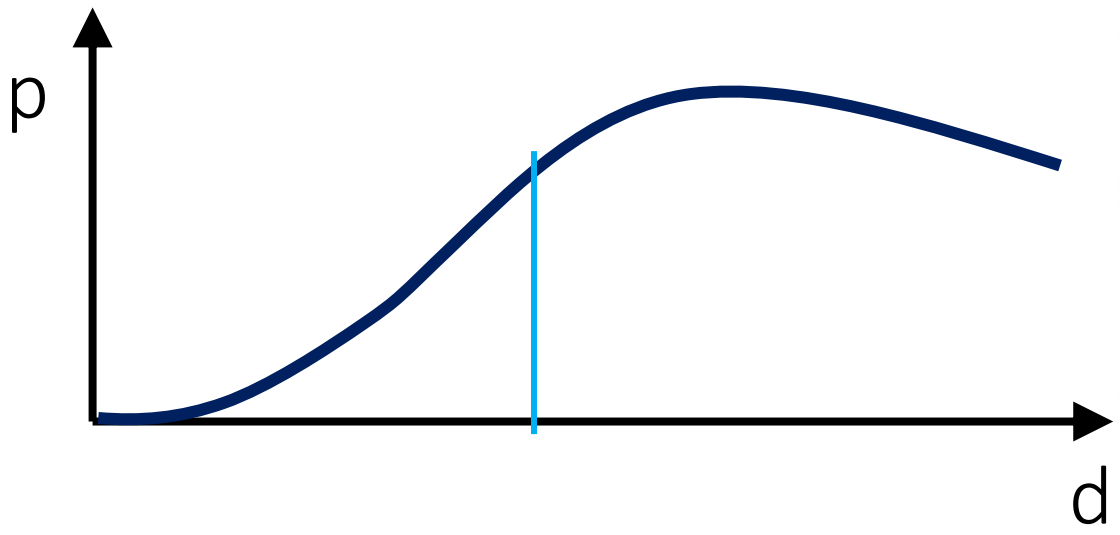

20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

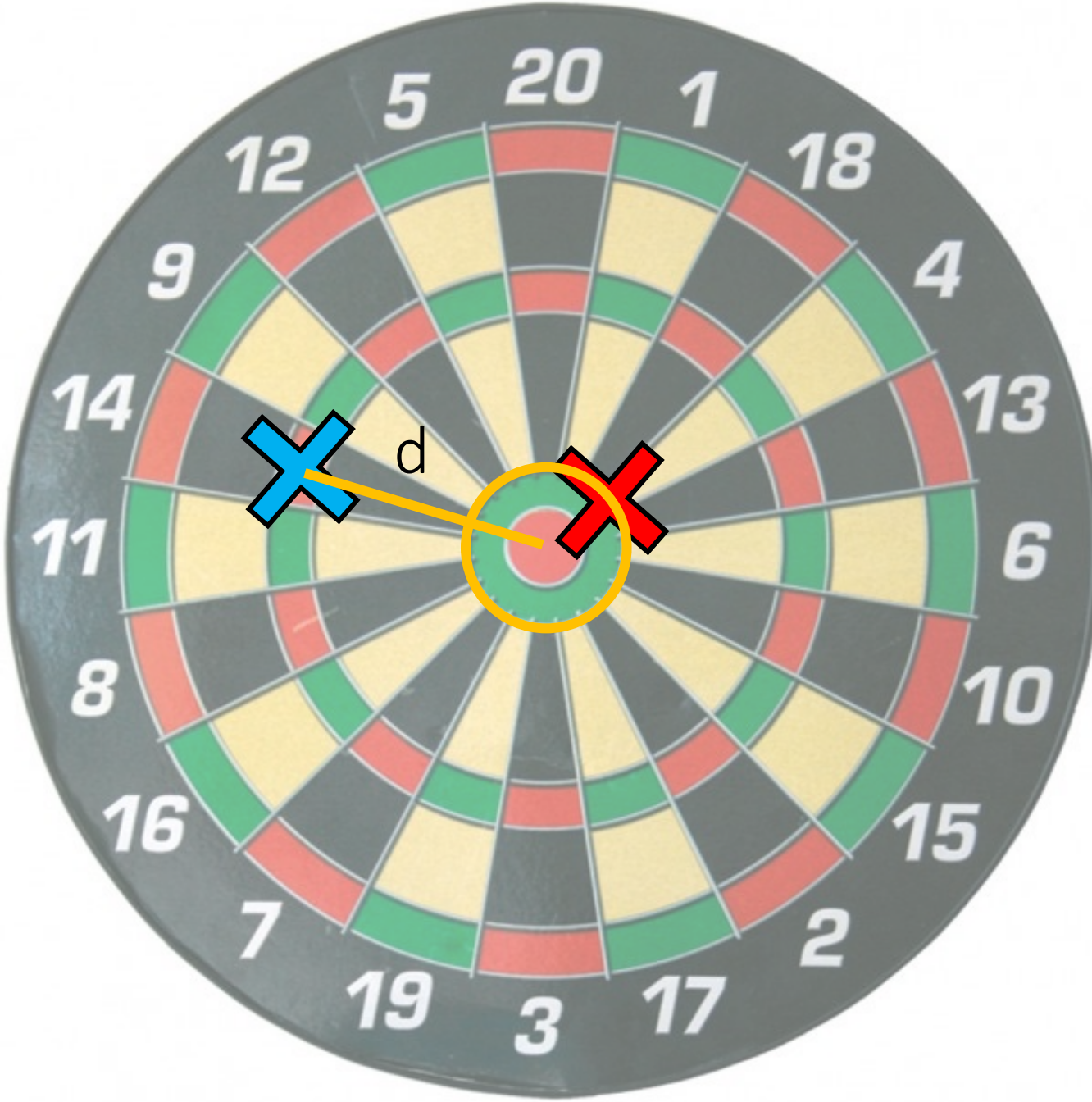

20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

1. Experiment
2. Being pessimistic:

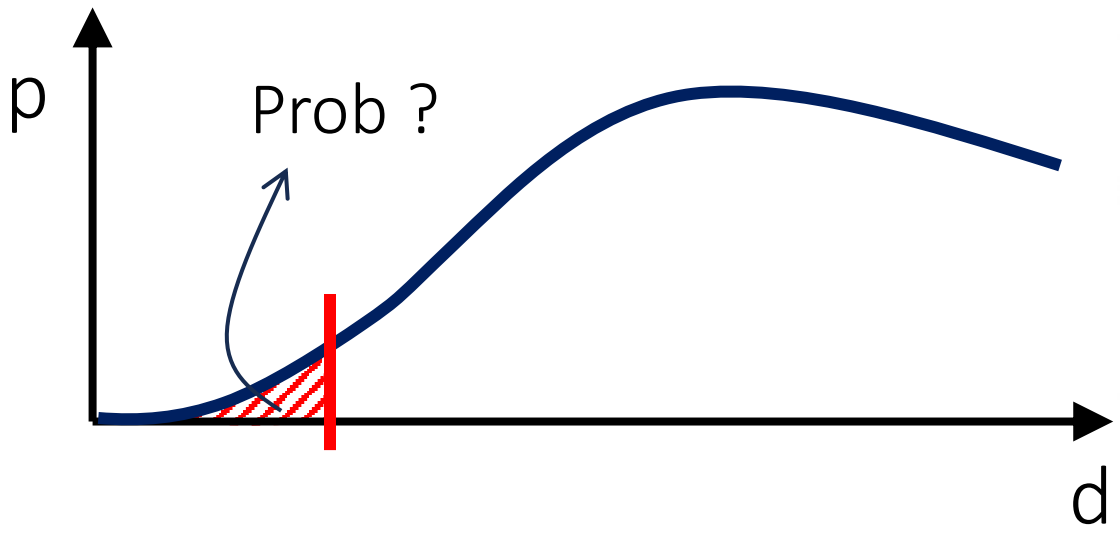
   You got there only by chance (null hypothesis)

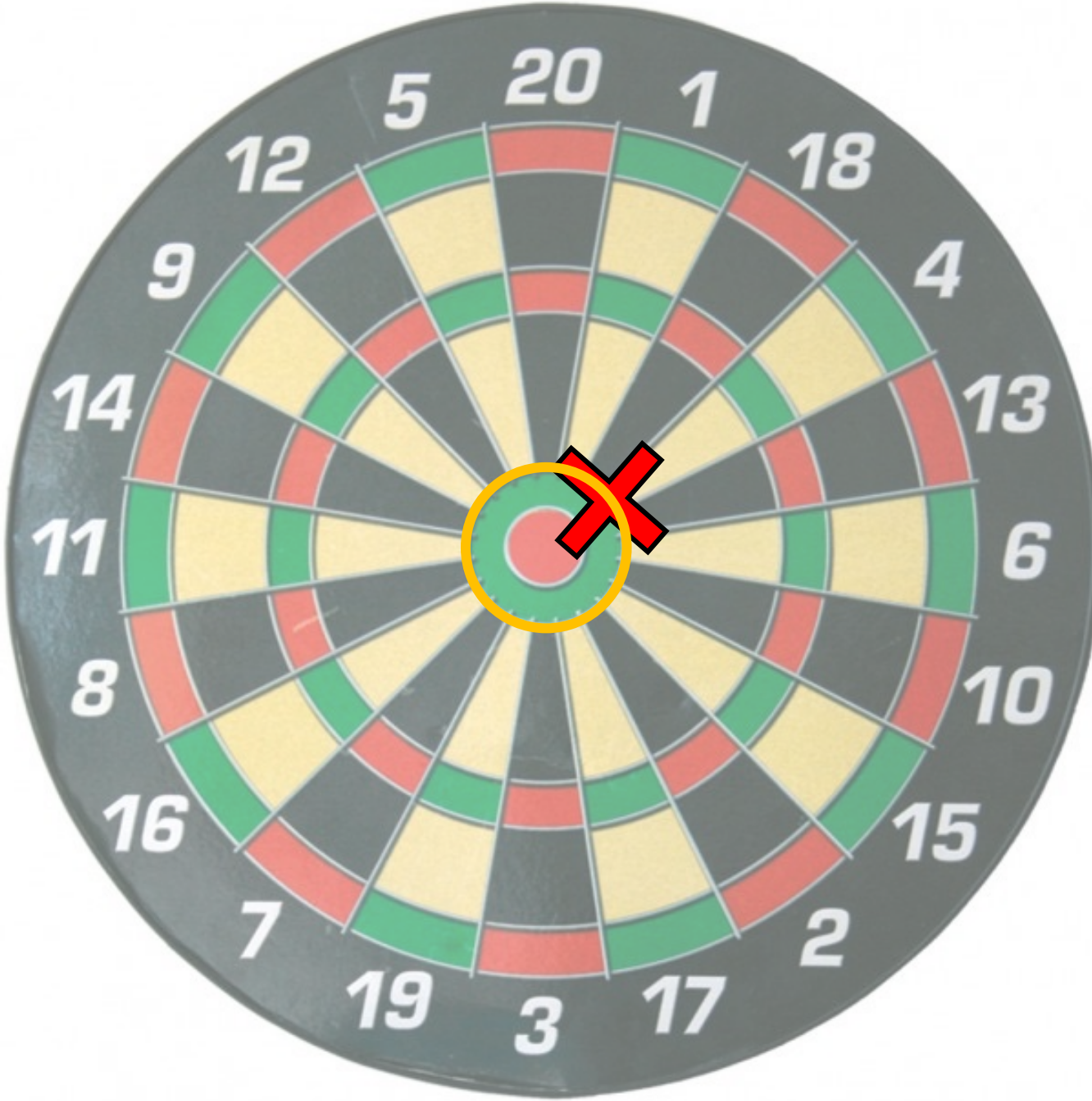
3. Obtain the null distribution





1. Experiment
2. Being pessimistic:
   You got there only by chance
   (null hypothesis)
3. Obtain the null distribution
4. Hypothesis testing





Hypothesis testing

- **What does P value Mean?**

  The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.

Dahiru et al. 2008

Hypothesis testing

$$p - value < 0.05 \quad (\alpha)$$

We can **reject** the null hypothesis and accept the **alternative hypothesis**

$$p - value > 0.05 \quad (\alpha)$$

We do not have **enough evidenced to reject** the null hypothesis.

Hypothesis Testing

- Null Hypothesis: H_0

- Alternative Hypothesis: H_1

- Significance level ($\boldsymbol{\alpha}$)

T-test: one sample test

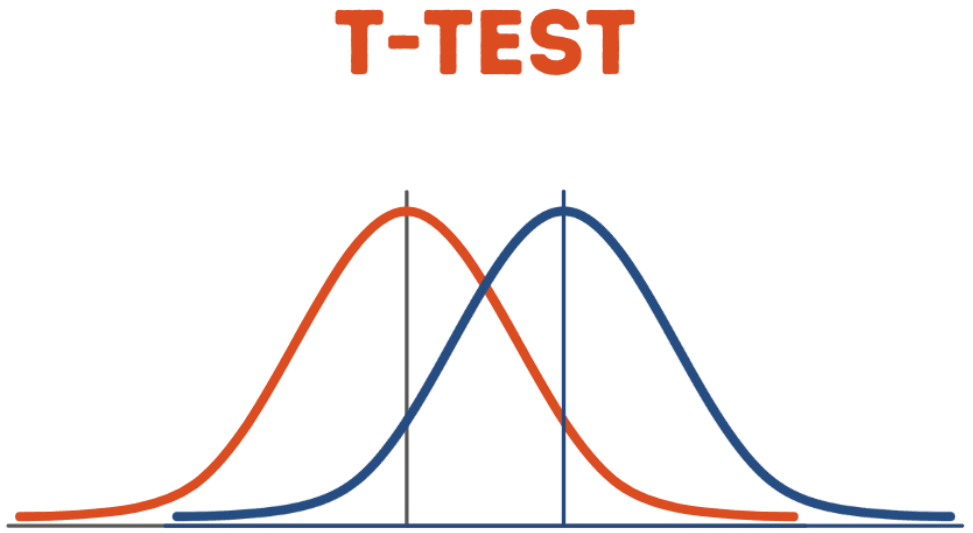
- Null Hypothesis (H_0): $\mu_1 = 0$

- Alternative Hypothesis (H_1) $\mu_1 \neq 0$

T-test: two sample test

- Null Hypothesis (H_0): $\mu_1 = \mu_2$

- Alternative Hypothesis (H_1)



For Loops

```
> for (i in 1:10){
>     # your code
> }
```

```
> set = c(1, 10, 20)
> for (i in set){
>     # your code
> }
```

Adjusting p-value

- Bonferroni correction
- Benjamini-Hochberg correction

Fisher’s Exact Test

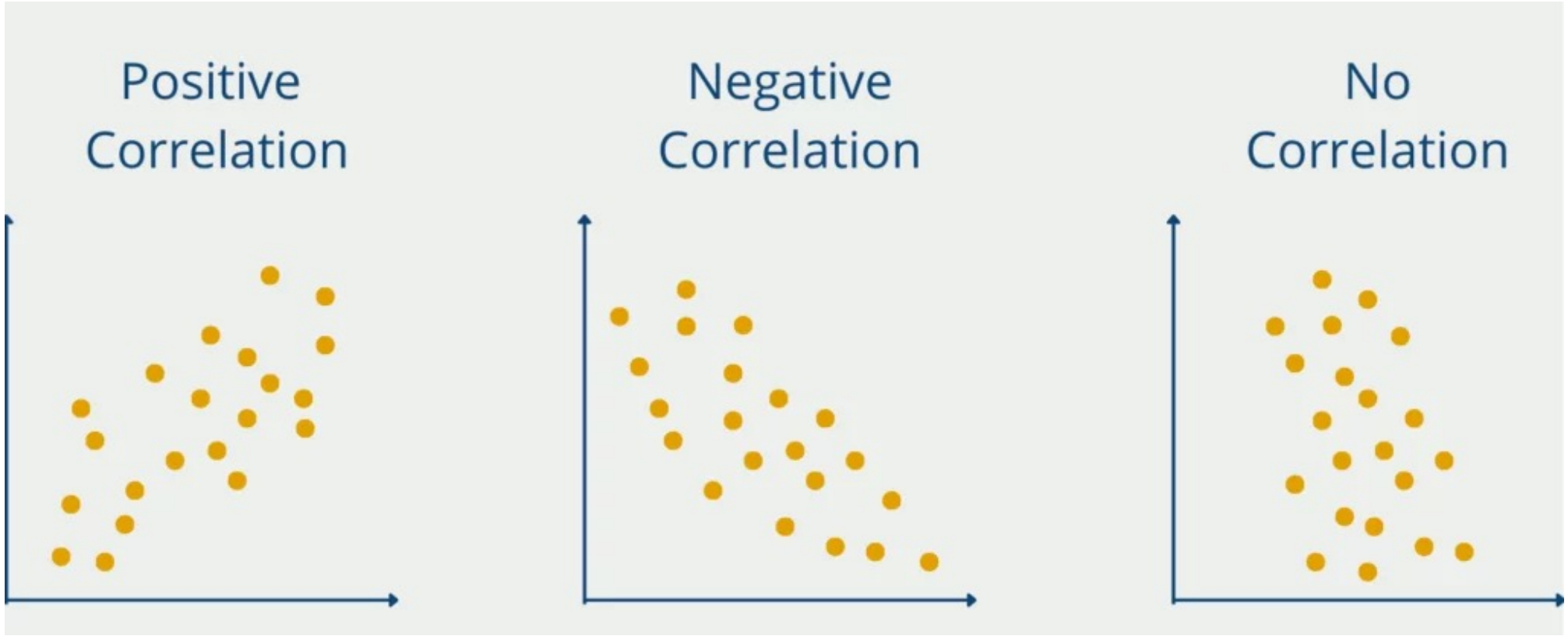
Fisher's Exact Test

	Group 1	Group 2
Condition 1	a	b
Condition 2	c	d

$$H_0: \frac{a}{b} = \frac{c}{d}$$

Covariance and Correlation

Covariance and Correlation



Covariance and Correlation

Variance

$$var(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Covariance

$$cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Pearson Correlation

$$cor(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

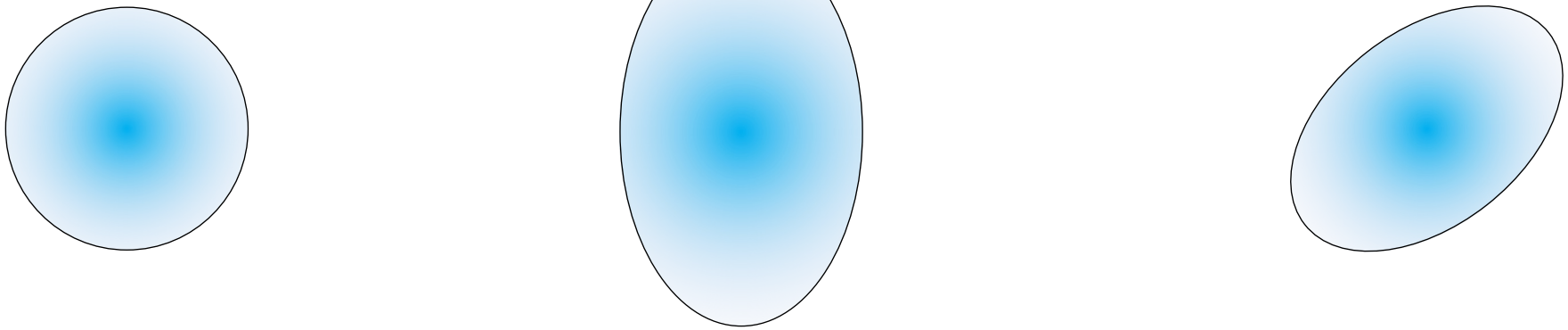
Covariance Matrix

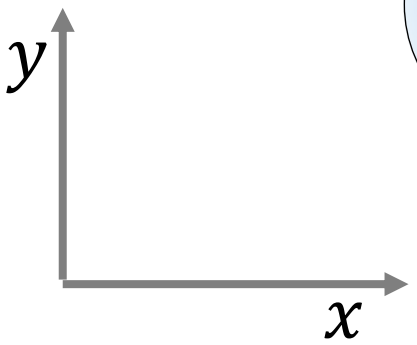
$$C_{xy} = \begin{bmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{bmatrix}$$

Covariance Matrix

$$C_{xy} = \begin{bmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$





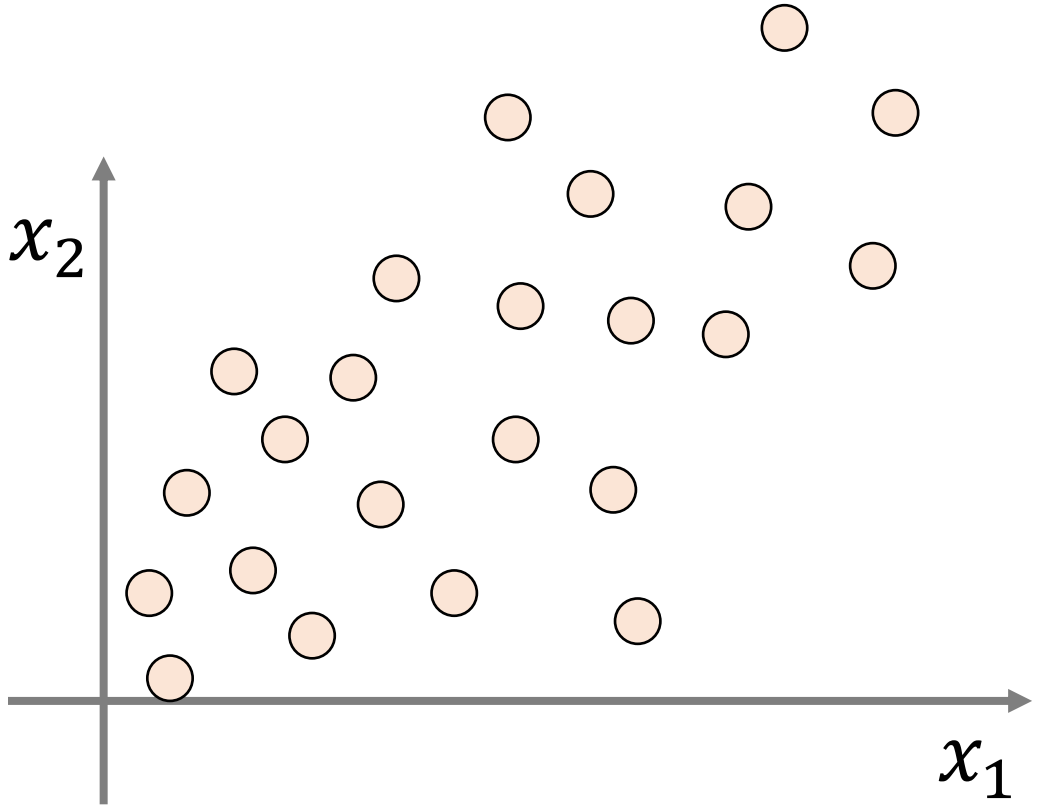
https://distill.pub/2019/visual-exploration-gaussian-processes/

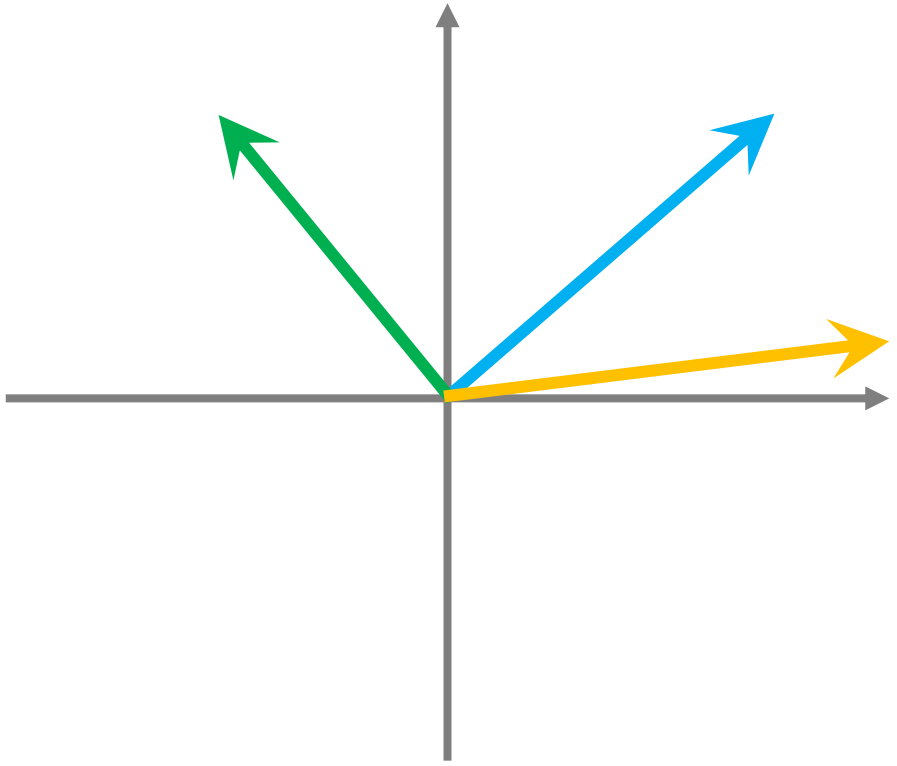
Principle Component Analysis

Principle Component Analysis

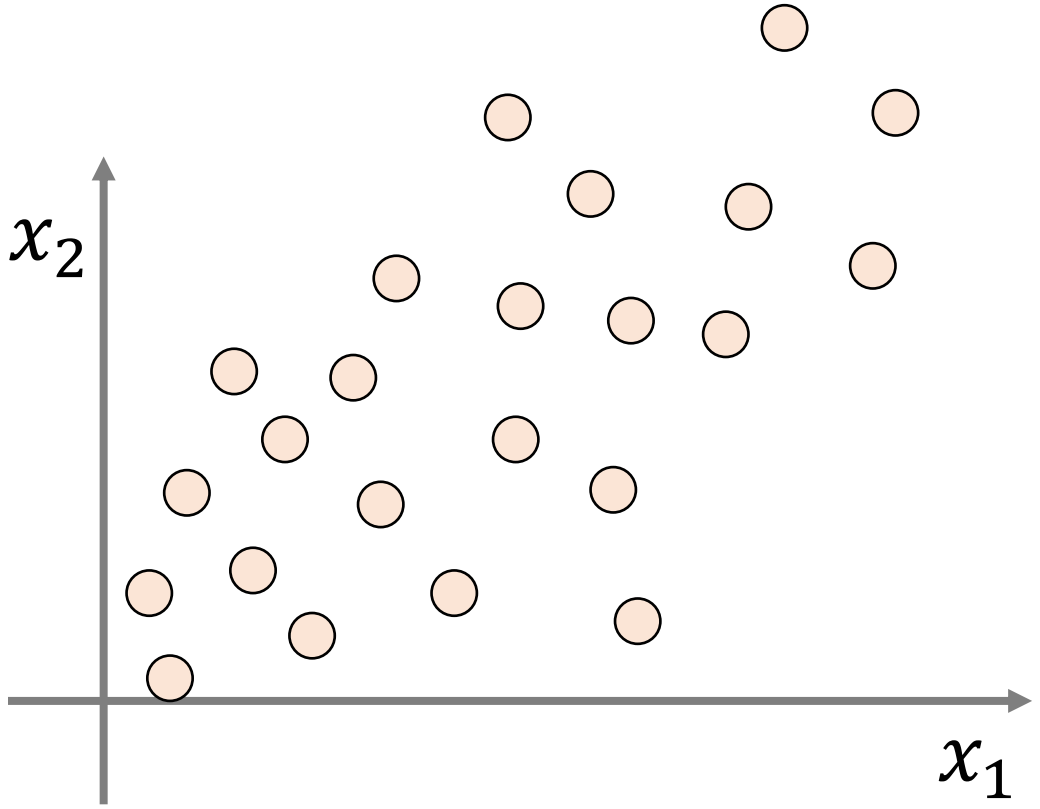
What animal does this shadow belong to?

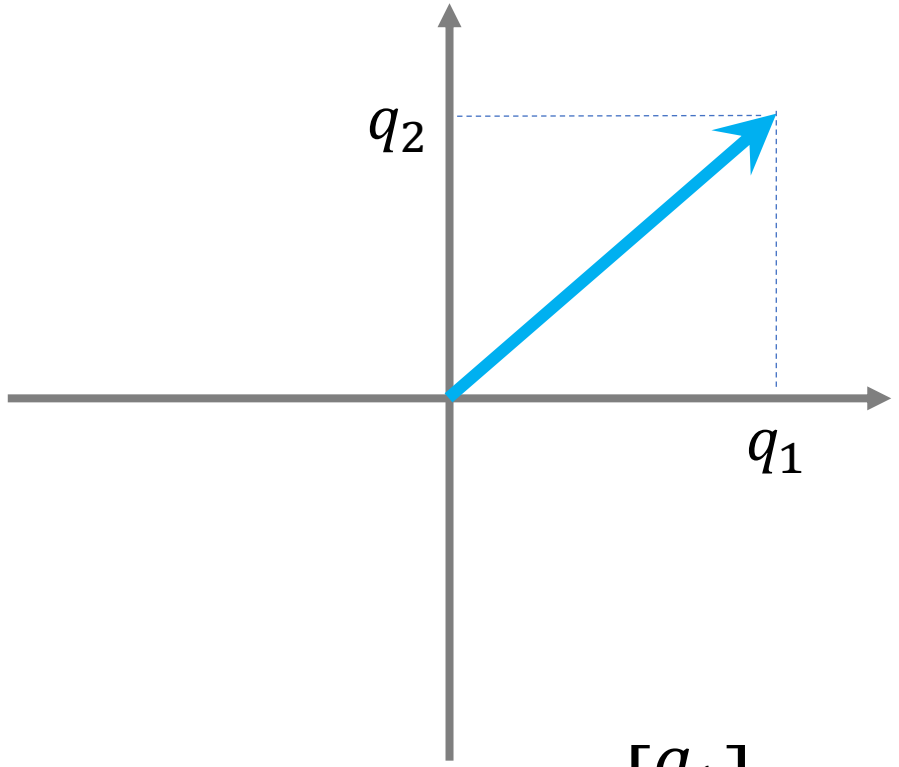
Principle Component Analysis





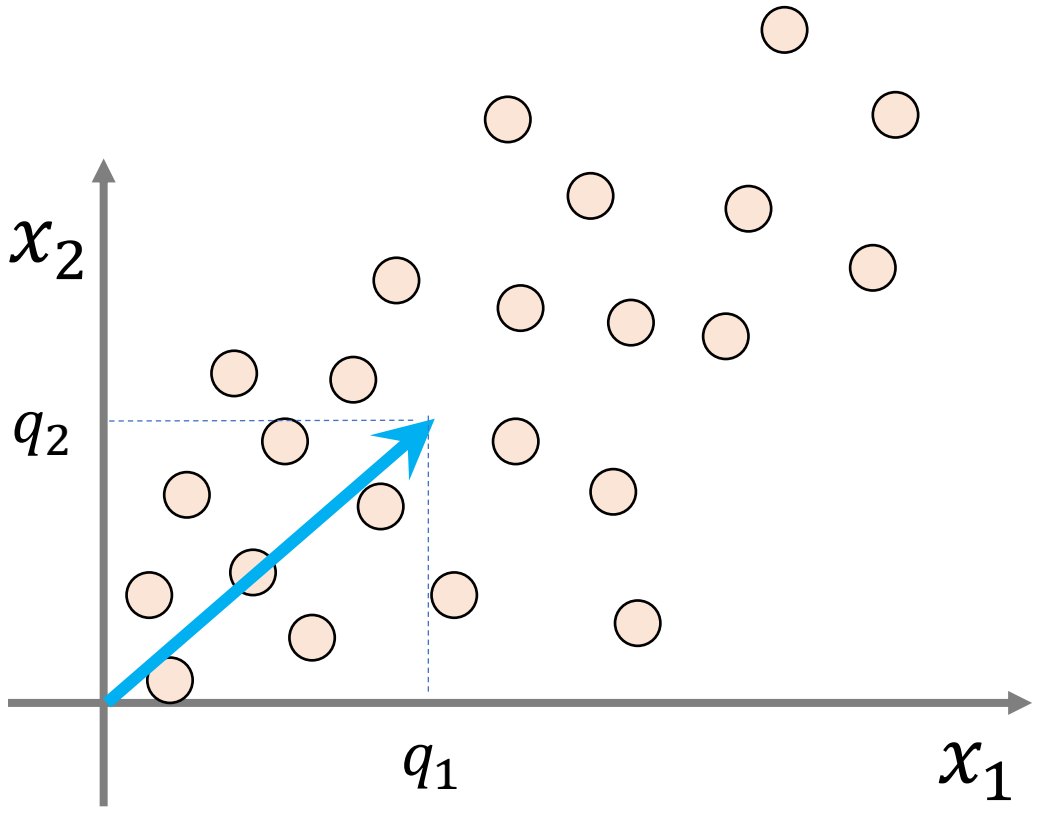
Principle Component Analysis





$$\boldsymbol{q} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$$

Principle Component Analysis



$\boldsymbol{q} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$ is an **eigen-vector** of

$C = \begin{bmatrix} var(x_1) & cov(x_1, x_2) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix}$

Eigen Analysis

Eigen-vector Eigen-value

$$C\boldsymbol{q} = \boldsymbol{q}\lambda$$

$$C\boldsymbol{q}_1 = \boldsymbol{q}_1\lambda_1 \qquad \lambda_1$$

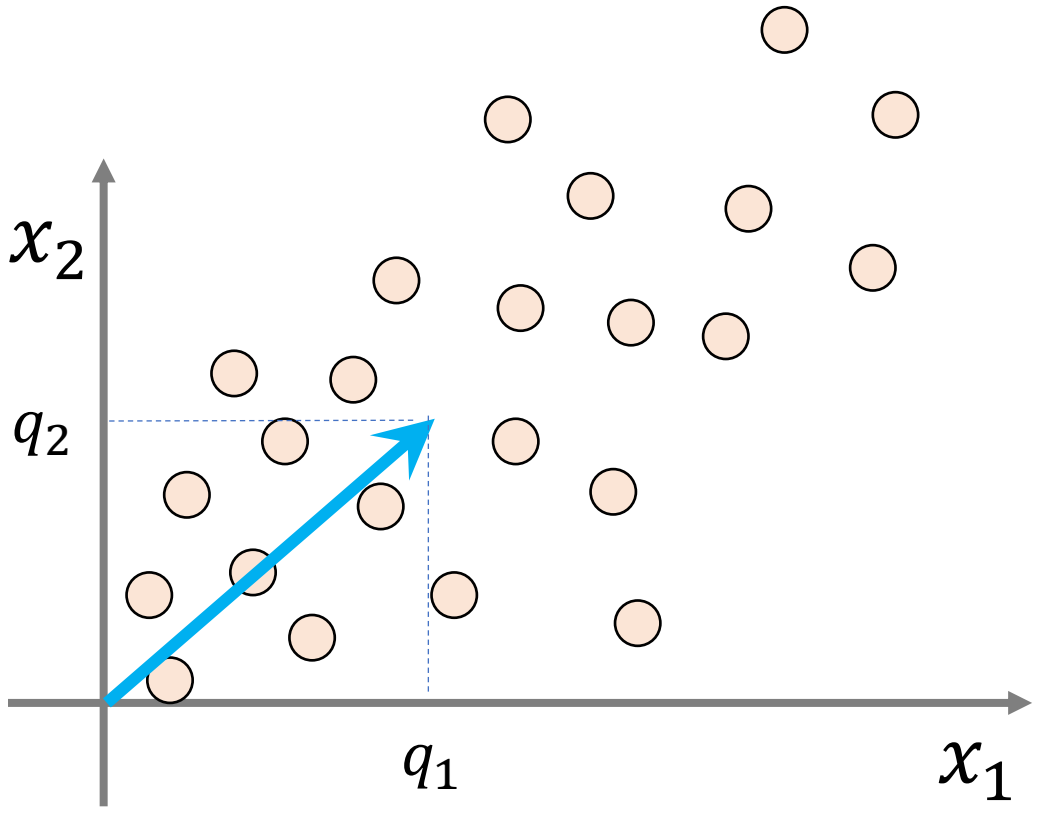
$$\vee$$

$$C\boldsymbol{q}_2 = \boldsymbol{q}_2\lambda_2 \qquad \lambda_2$$

$$\vee$$

$$\ldots \qquad \ldots$$

Principle Component Analysis

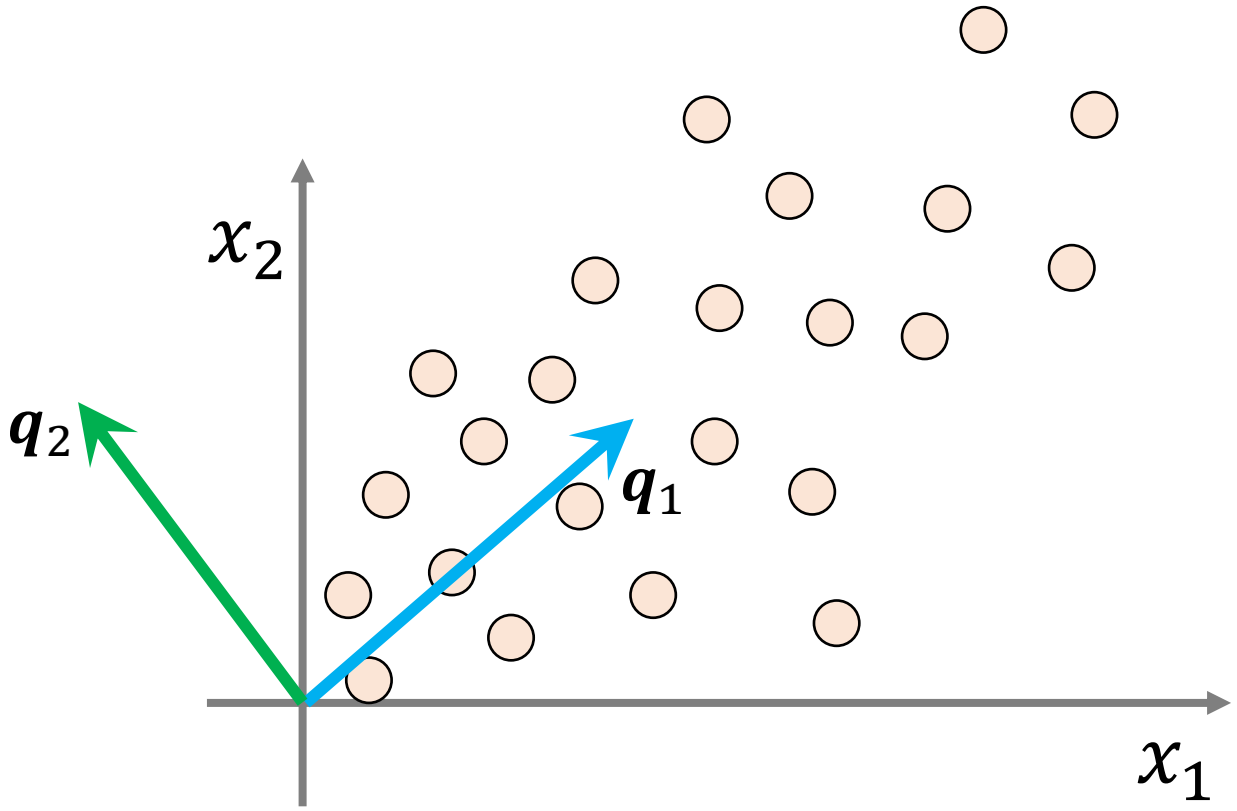


$\boldsymbol{q} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$ is the **eigen-vector** of

$$C = \begin{bmatrix} var(x_1) & cov(x_1, x_2) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix}$$

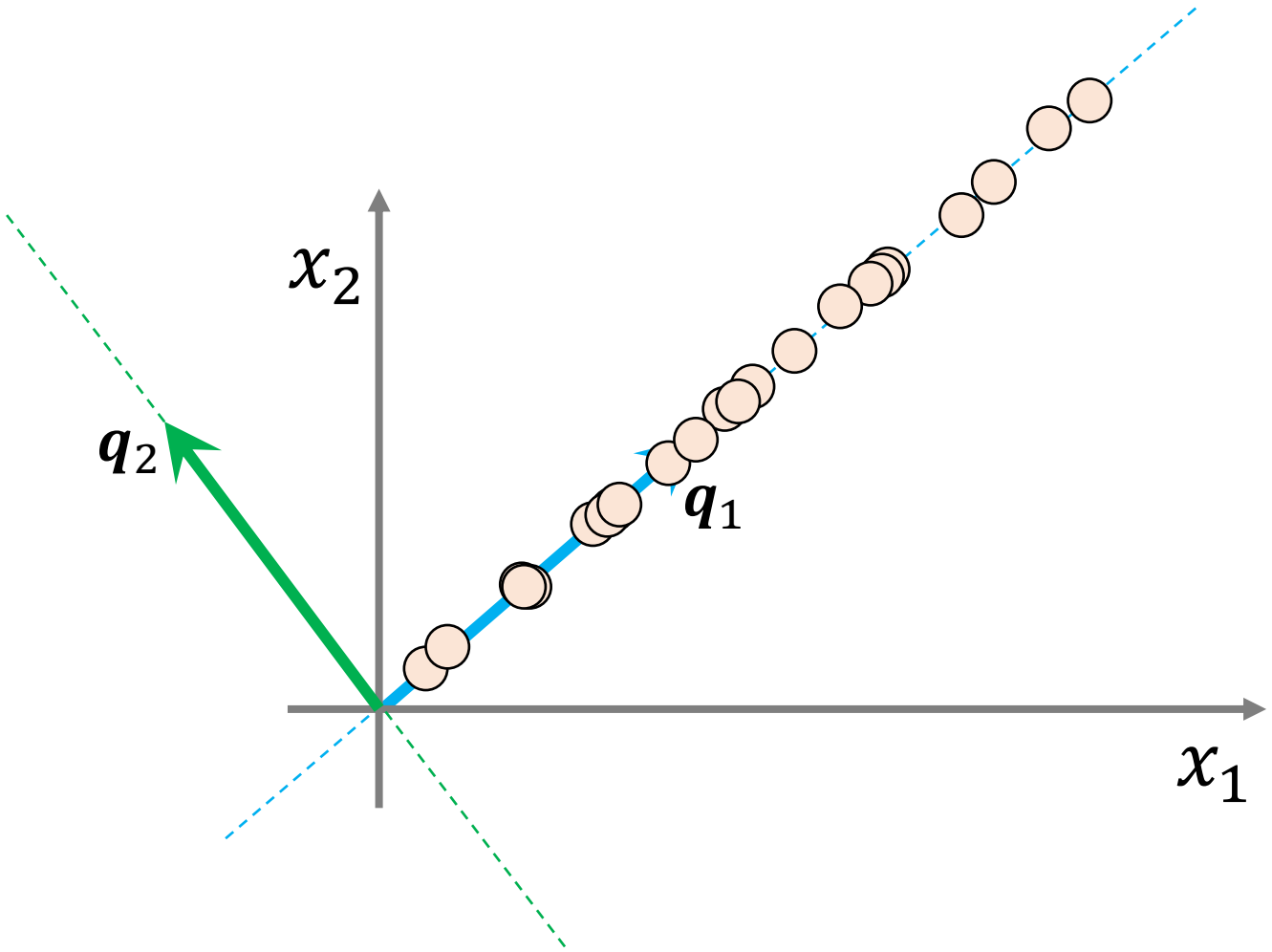
that corresponds to the largest **eigen-value**

Principle Component Analysis



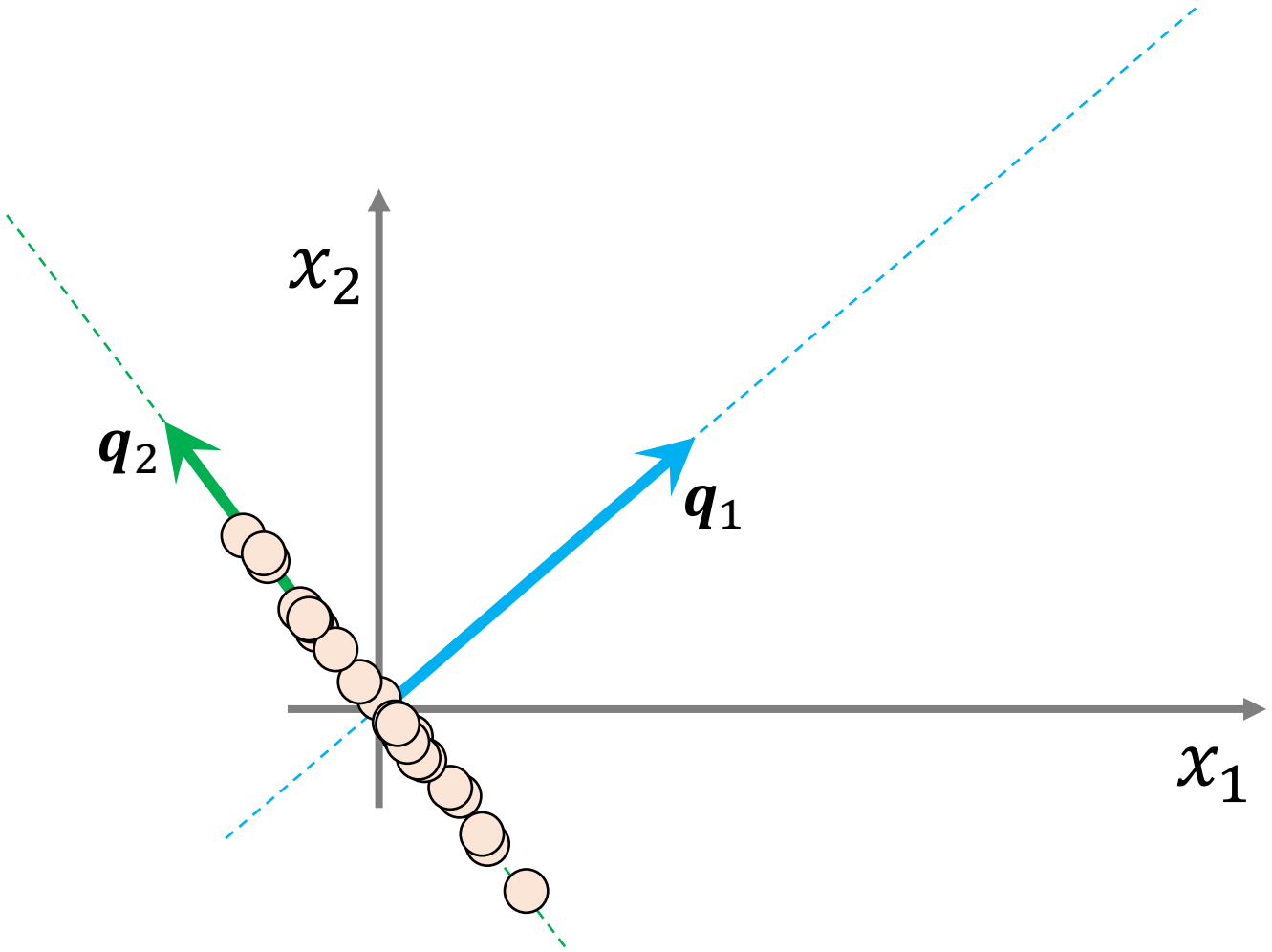
$$q_1 \perp q_2 \perp \cdots$$

Principle Component Analysis



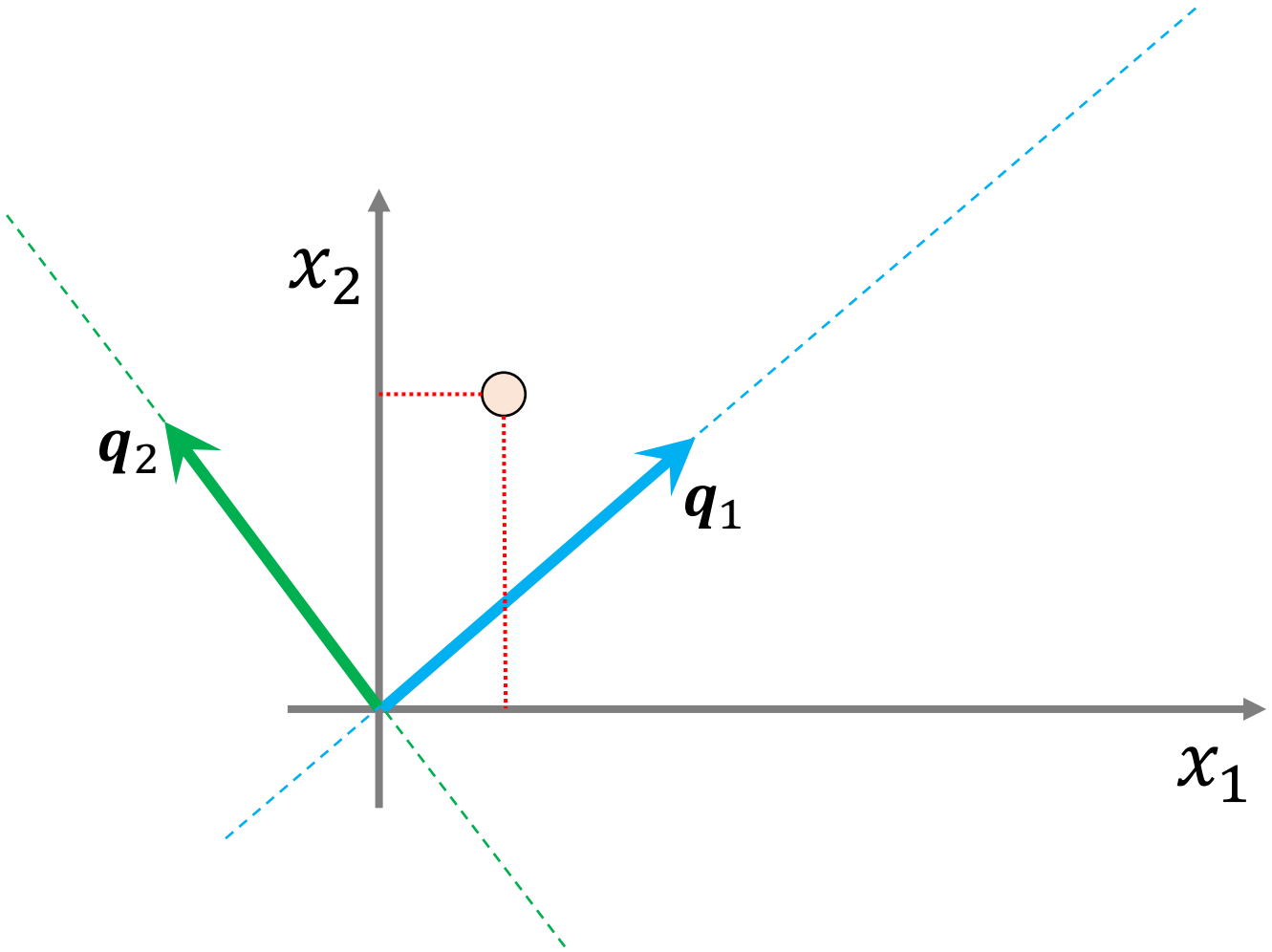
$$\boldsymbol{pc}_1 = X\boldsymbol{q}_1$$

Principle Component Analysis

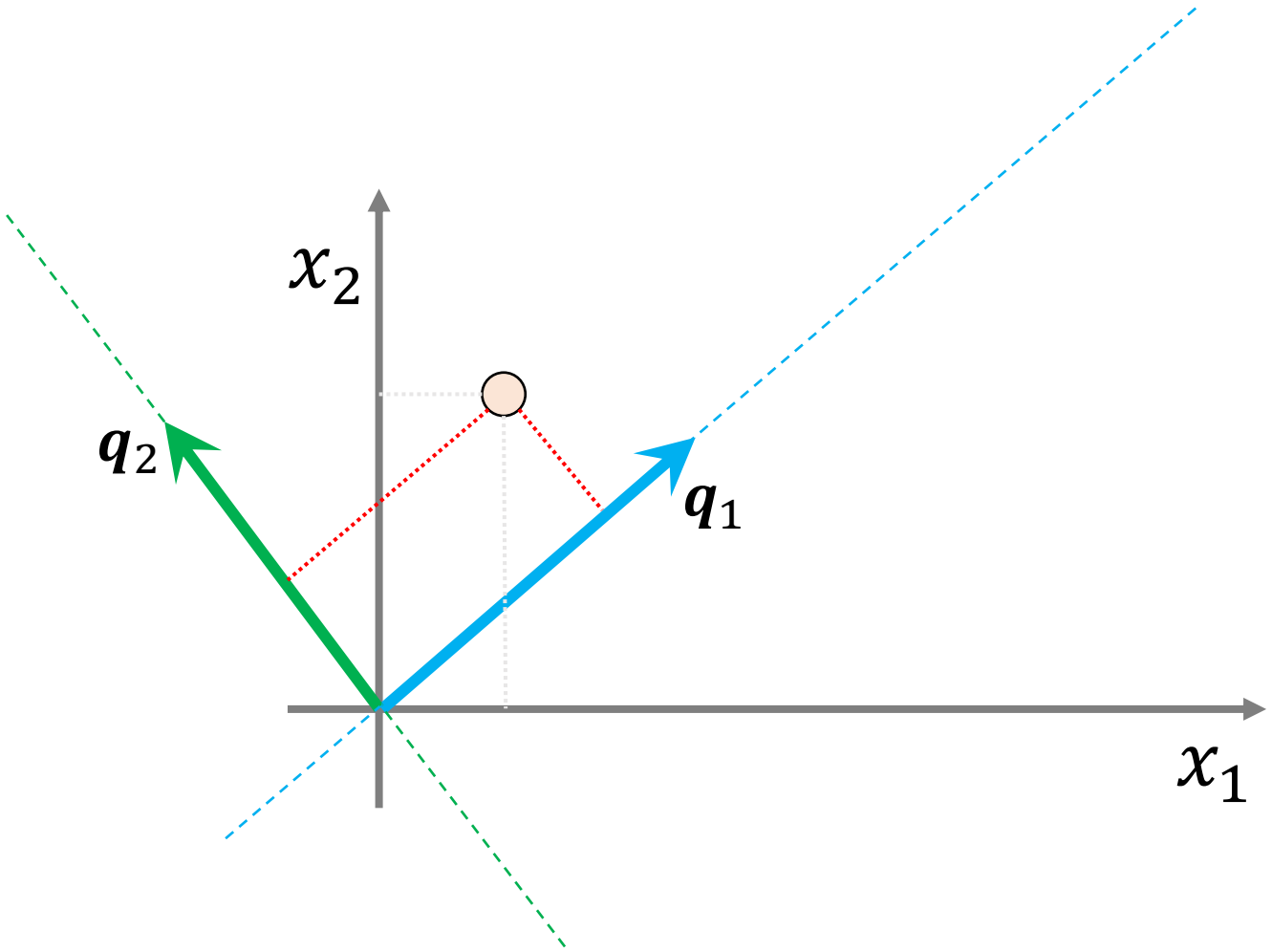


$$\boldsymbol{pc}_2 = X\boldsymbol{q}_2$$

Principle Component Analysis



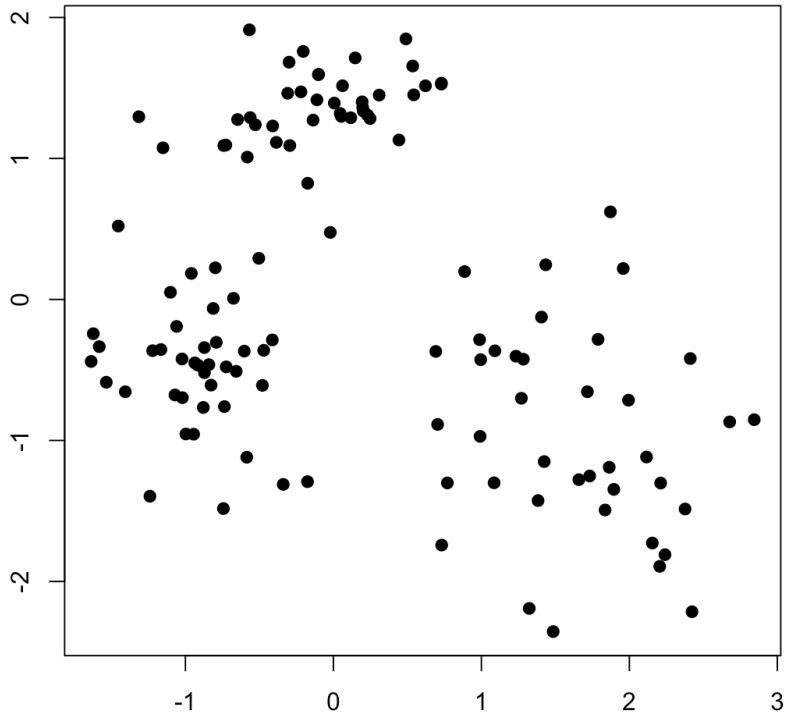
Principle Component Analysis

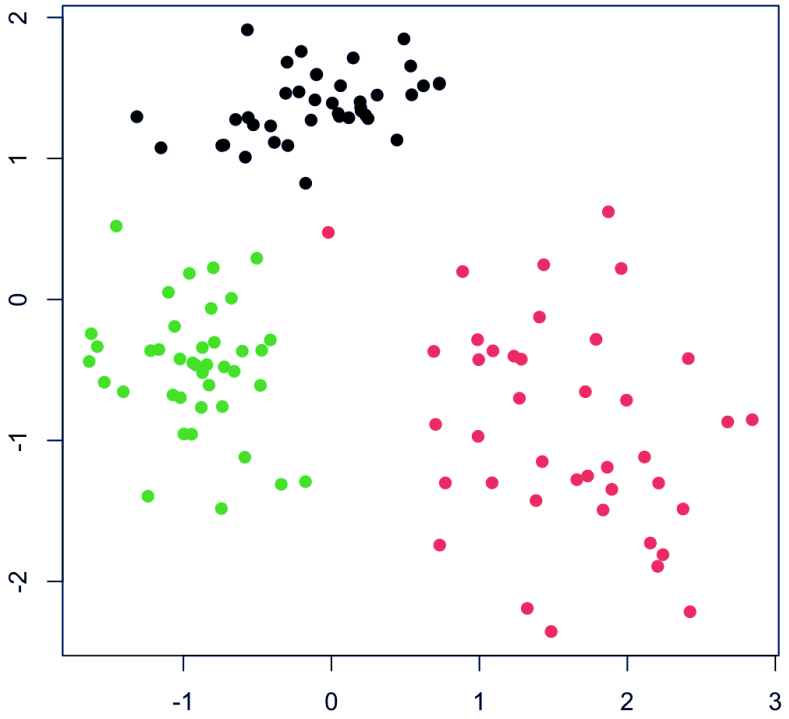


Clustering

Clustering

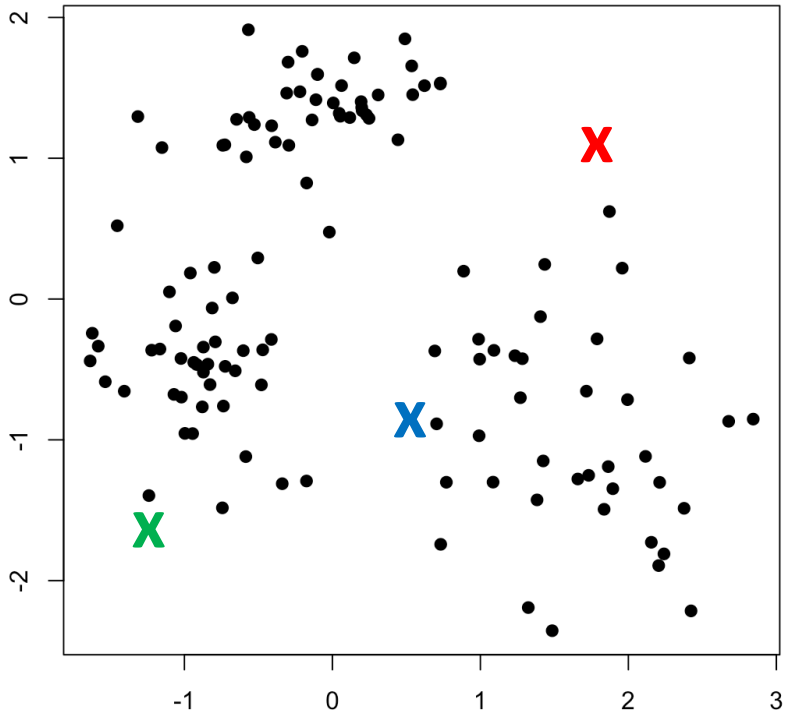
Identify groups of similar data points in an unsupervised manner.

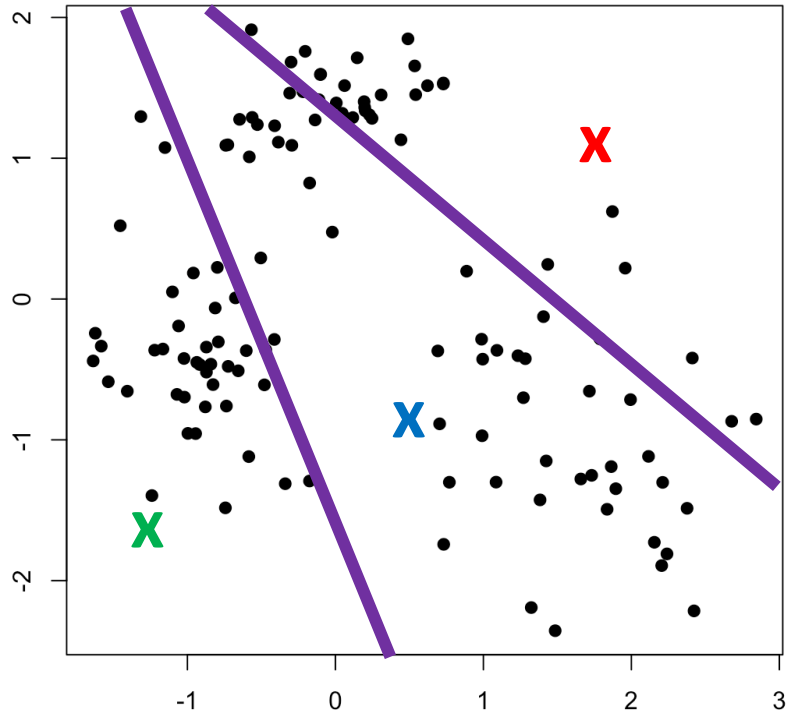


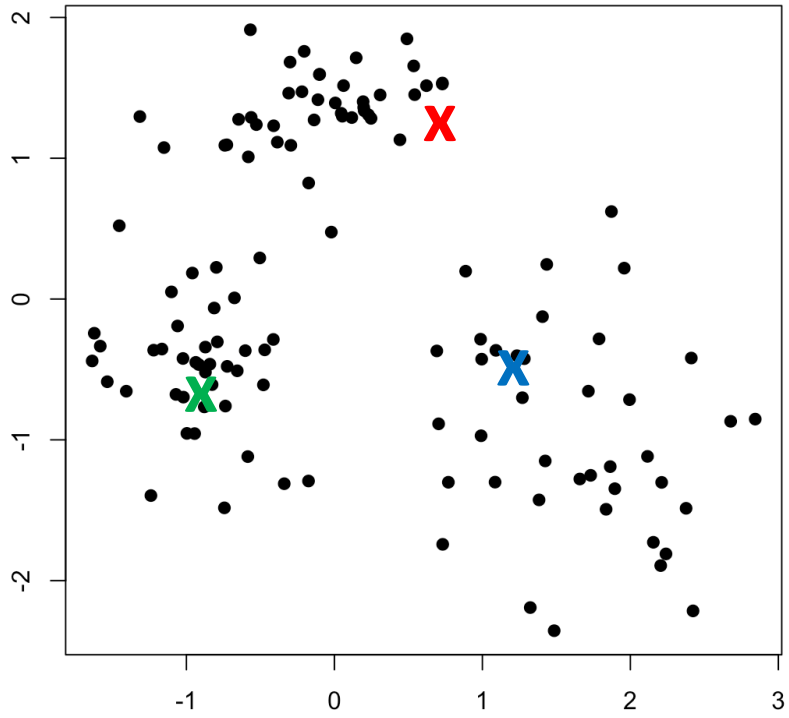


Clustering

K-means clustering

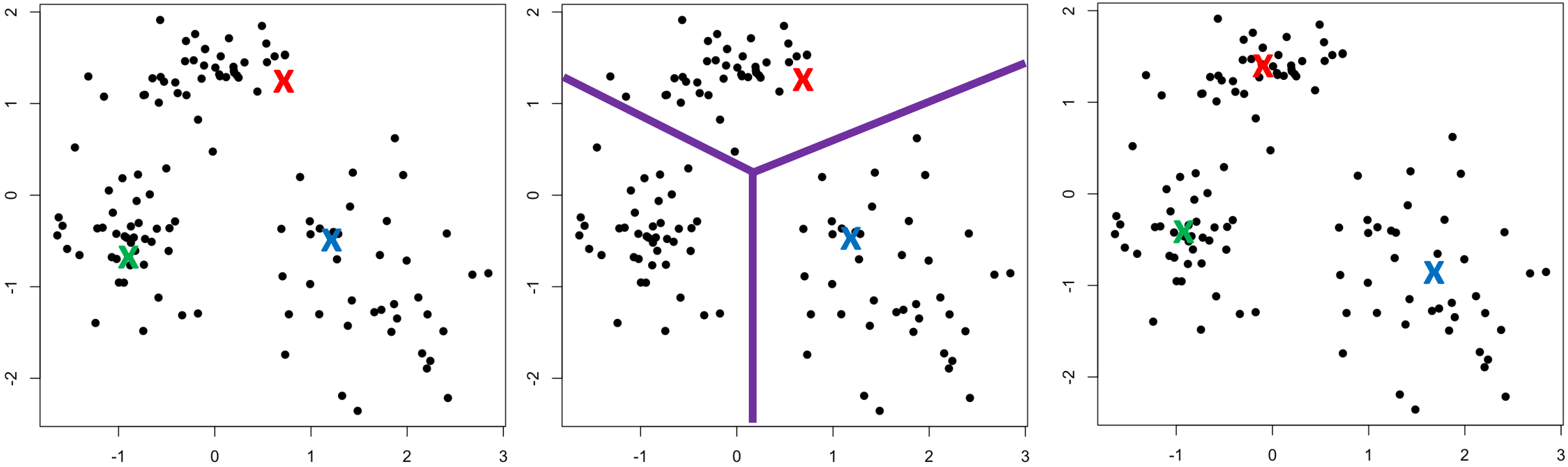






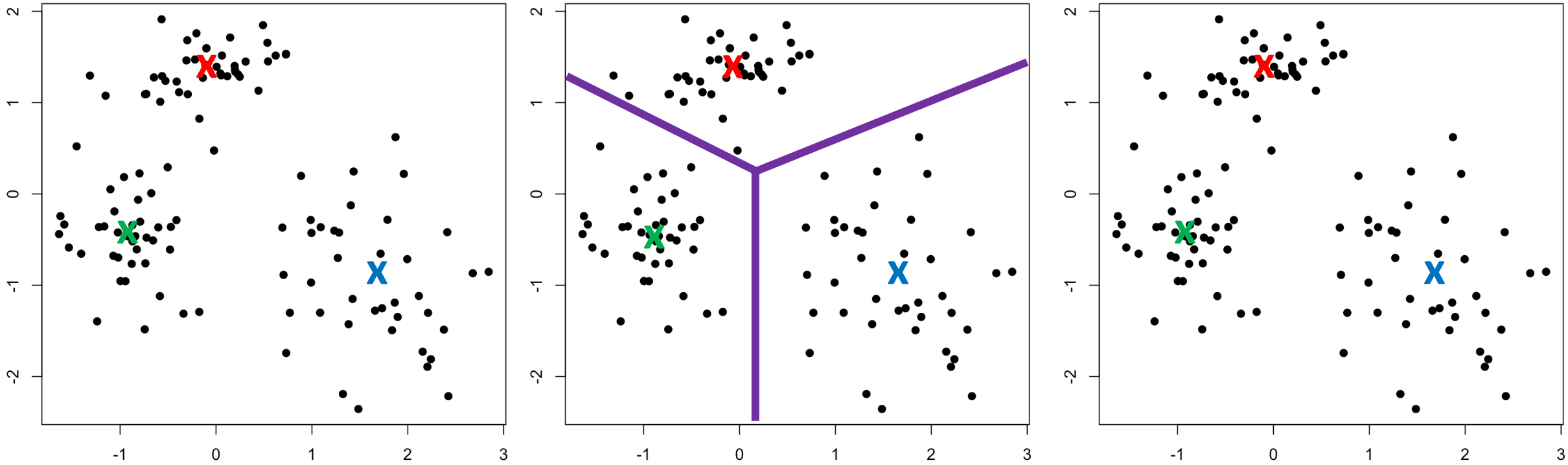
Clustering

K-means clustering



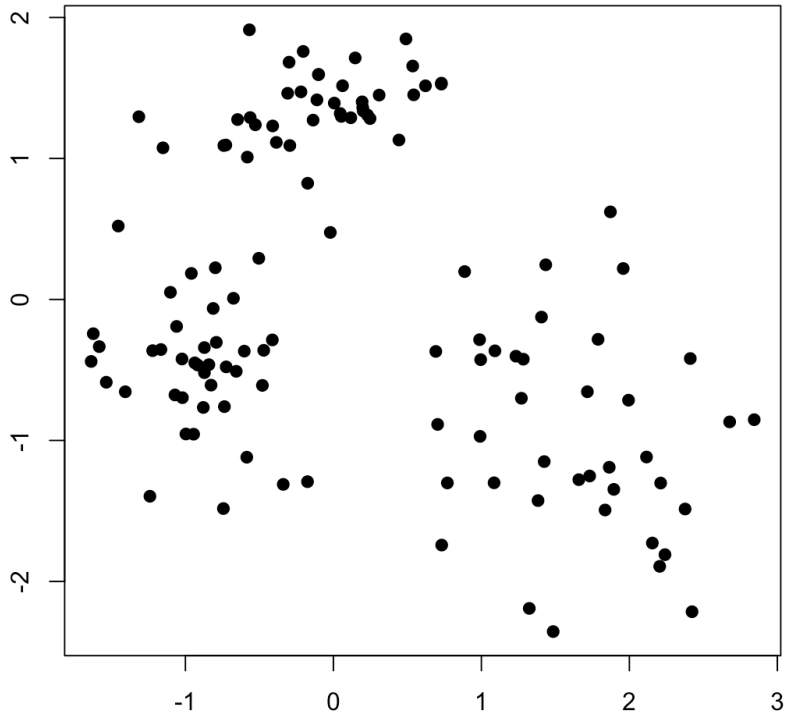
Clustering

K-means clustering

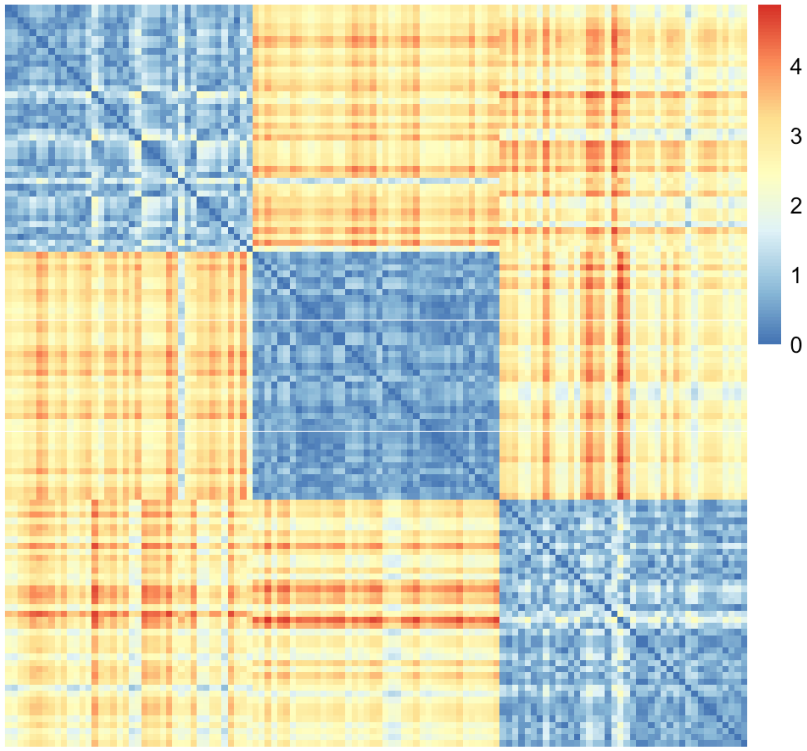


Clustering

Hierarchical clustering

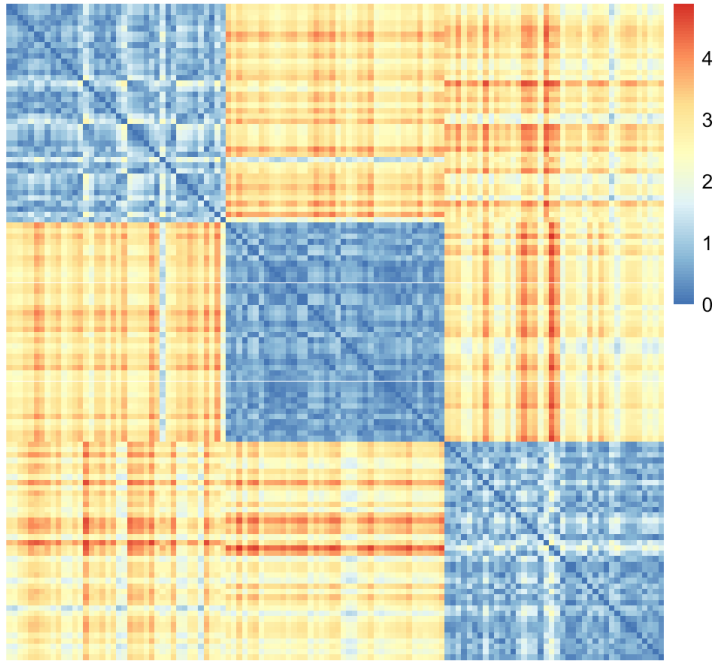


distances

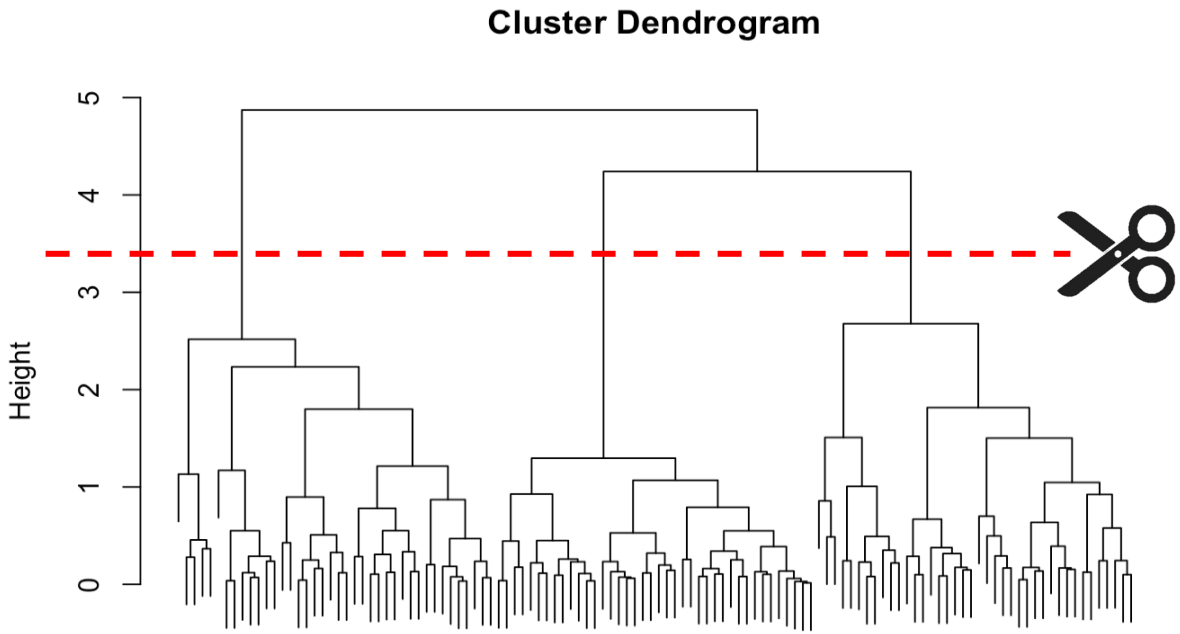


Clustering

Hierarchical clustering



Make tree



Thank You :)

Please give us feedback!