# Classifiers

# Readings for today

- Chapter 4: Classification. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer.

# Topics

1. Logistic regression

2. Linear & quadratic discriminant analysis

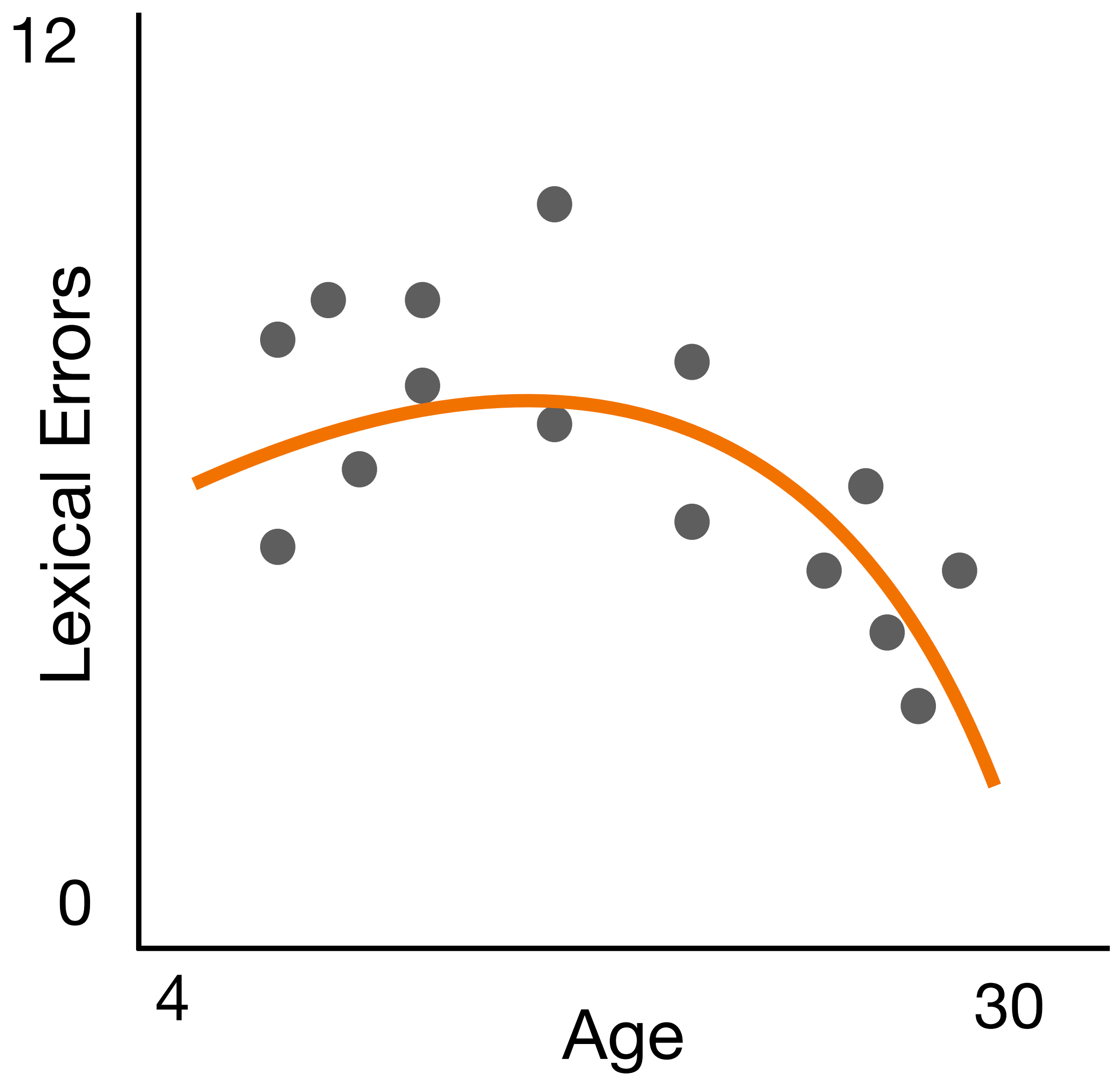
# Logistic regression

# Classes of models

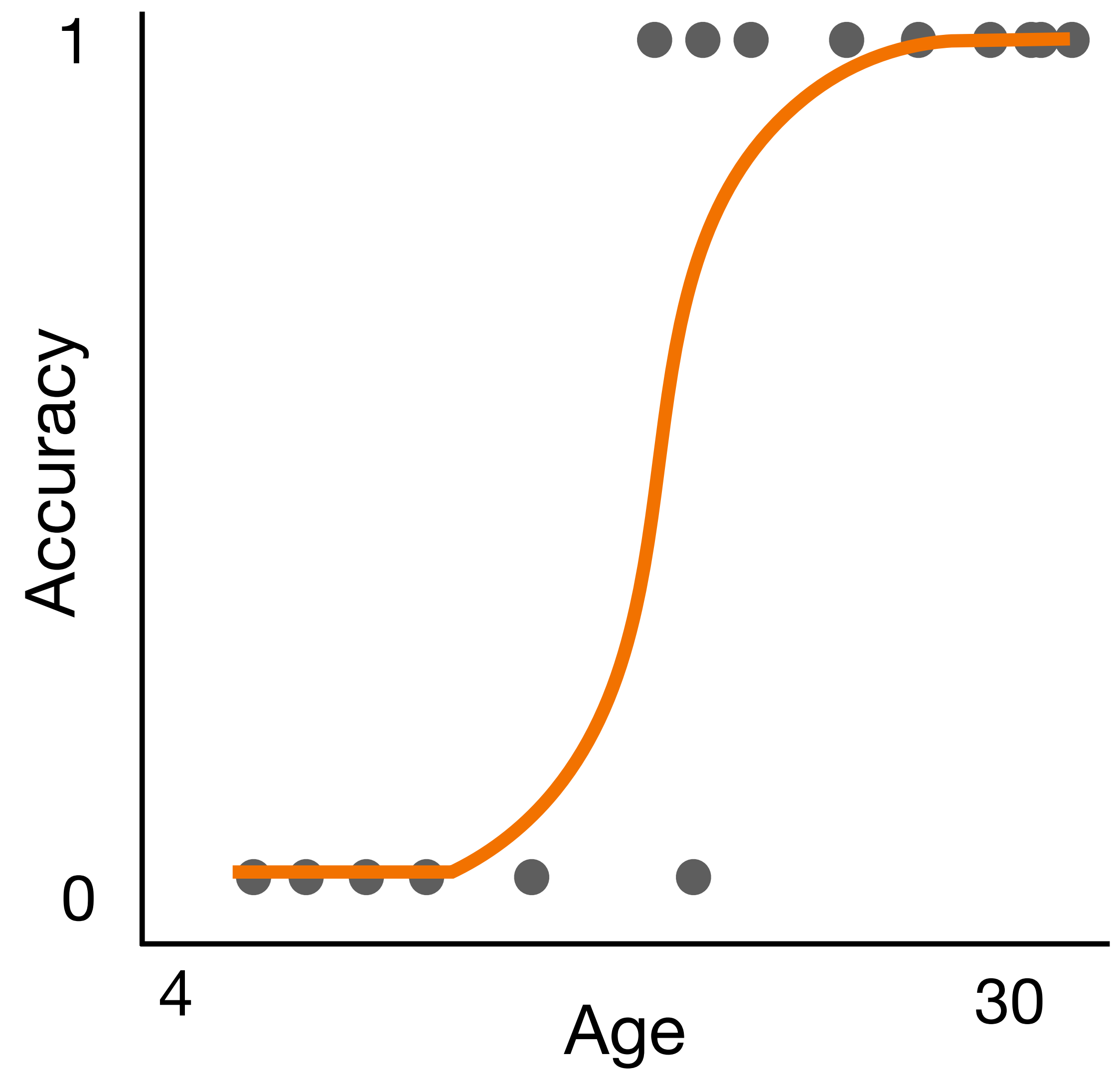
$Y$

| | Quantitative | Qualitative |
|---|---|---|
| **Regression** | ✓ | |
| **Classification** | | ✓ |

**Regression**



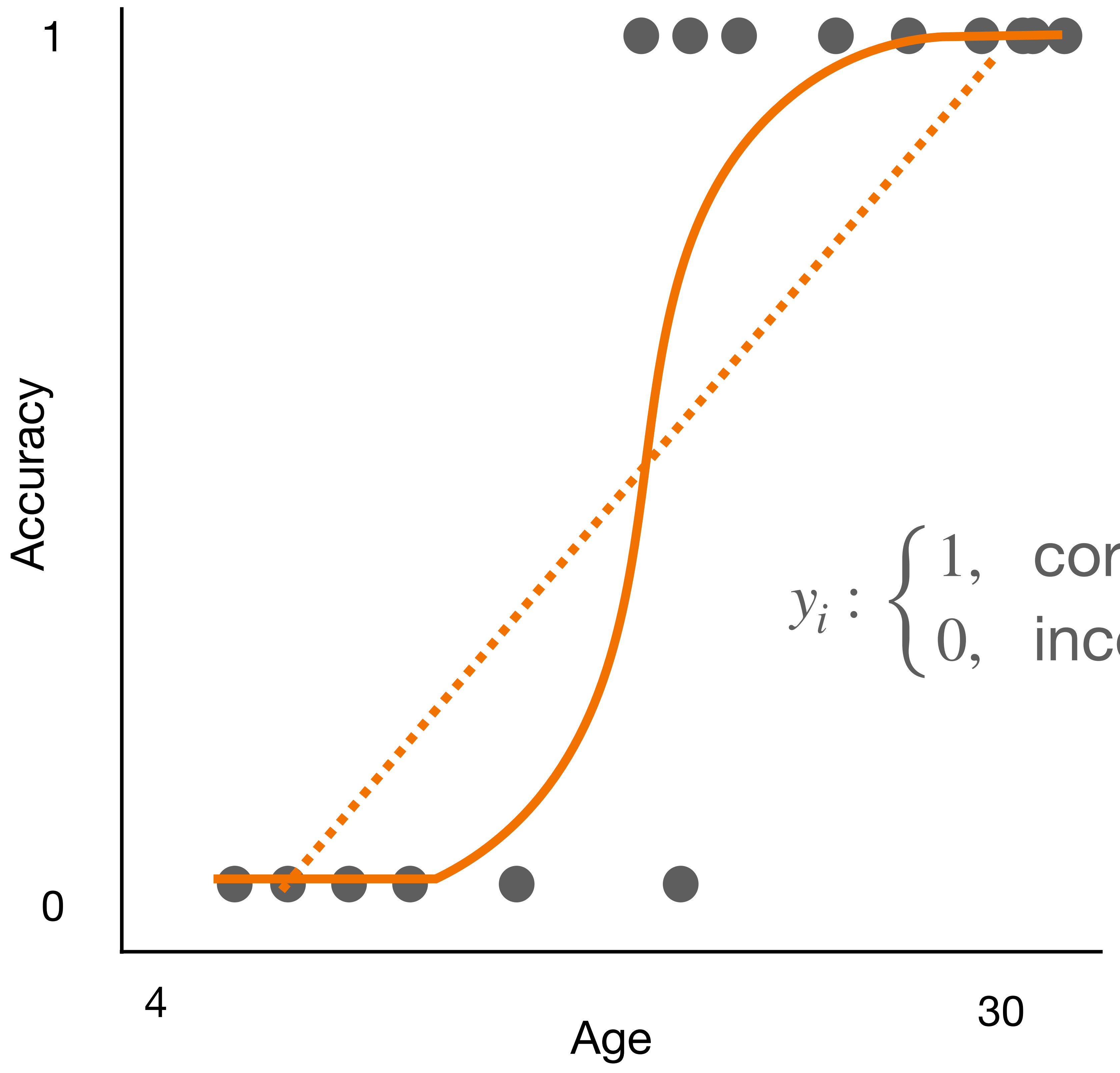

**Classification**






(James et al. 2013)

# Logistic regression




$$y_i : \begin{cases} 1, & \text{correct} \\ 0, & \text{incorrect} \end{cases}$$

**Linear regression**

$$f(y_i \mid x_i, \beta_1, \beta_0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

Residual errors are normally distributed

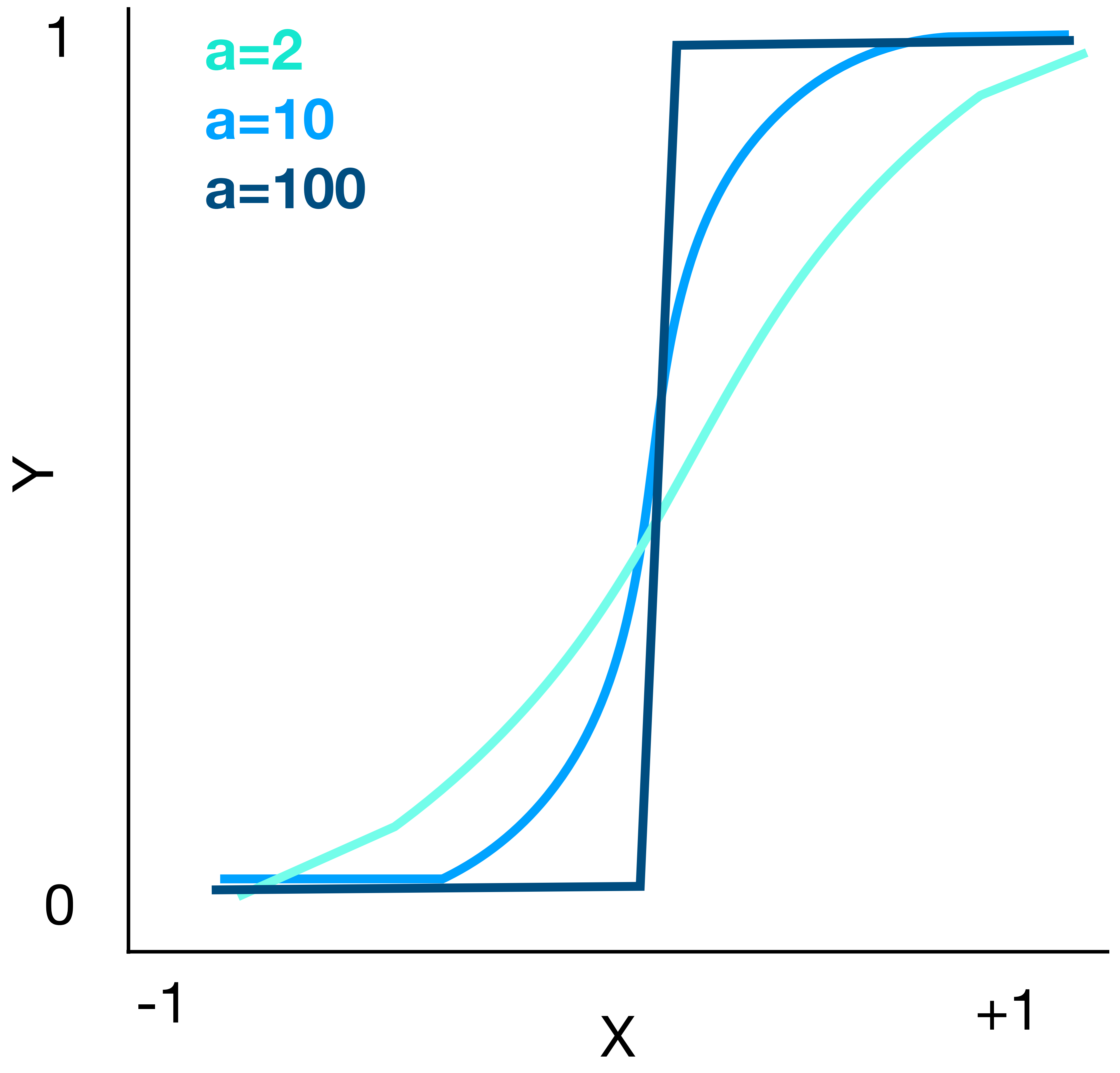
**Logistic regression**

$$p(y = 1 \mid x) = f(y_i \mid x_i, \beta_1, \beta_0, \sigma) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

Residual errors are binomial (i.e., exist in one of 2 states)

# The standard logistic function

**Standard logistic function:** $p(y = 1 \mid x) = \dfrac{e^{ax}}{1 + e^{ax}}$




- Lower *growth rate* parameter ($a$), means slower transition between states of $y$ (i.e., greater uncertainty).
- In a regression context better separation of $y$ as $x$ changes leads to larger $a$ (i.e., $a = \sum_{j=1}^{p} \hat{\beta}_j X_j + \hat{\beta}_0$)

**Odds**

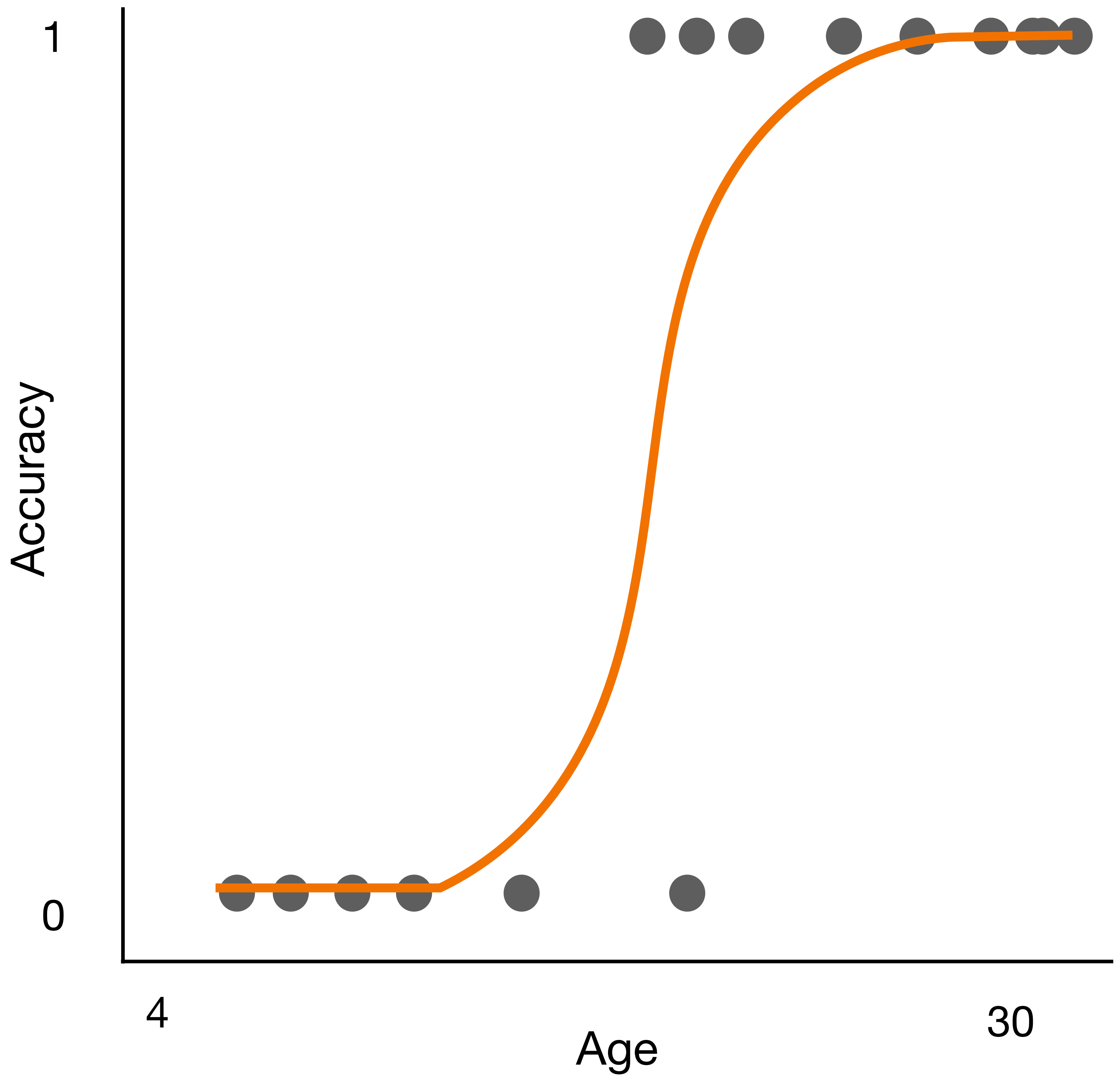
$$odds(y = 1) = \frac{p(y = 1)}{1 + p(y = 1)} = e^{ax}$$

$$ln(odds(y = 1)) = ln(\frac{p(y = 1)}{1 + p(y = 1)}) = ax = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$$



(James et al. 2013)

# Logistic regression




**Assumptions:**

1. $f(X)$ describes a linear relationship between $X$ and $Y$.
2. $Y$ is binary/dichotomous.
3. There is no collinearity between features in $X$.
4. $f(X)$ is stationary.



(James et al. 2013)

# Interpretation

**Linear regression**

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} = 1 + 0.5x_i$$

Interpretation

For a 1 unit change in $x$, $y$ changes 0.5, added to a baseline of 1.

**Logistic regression**

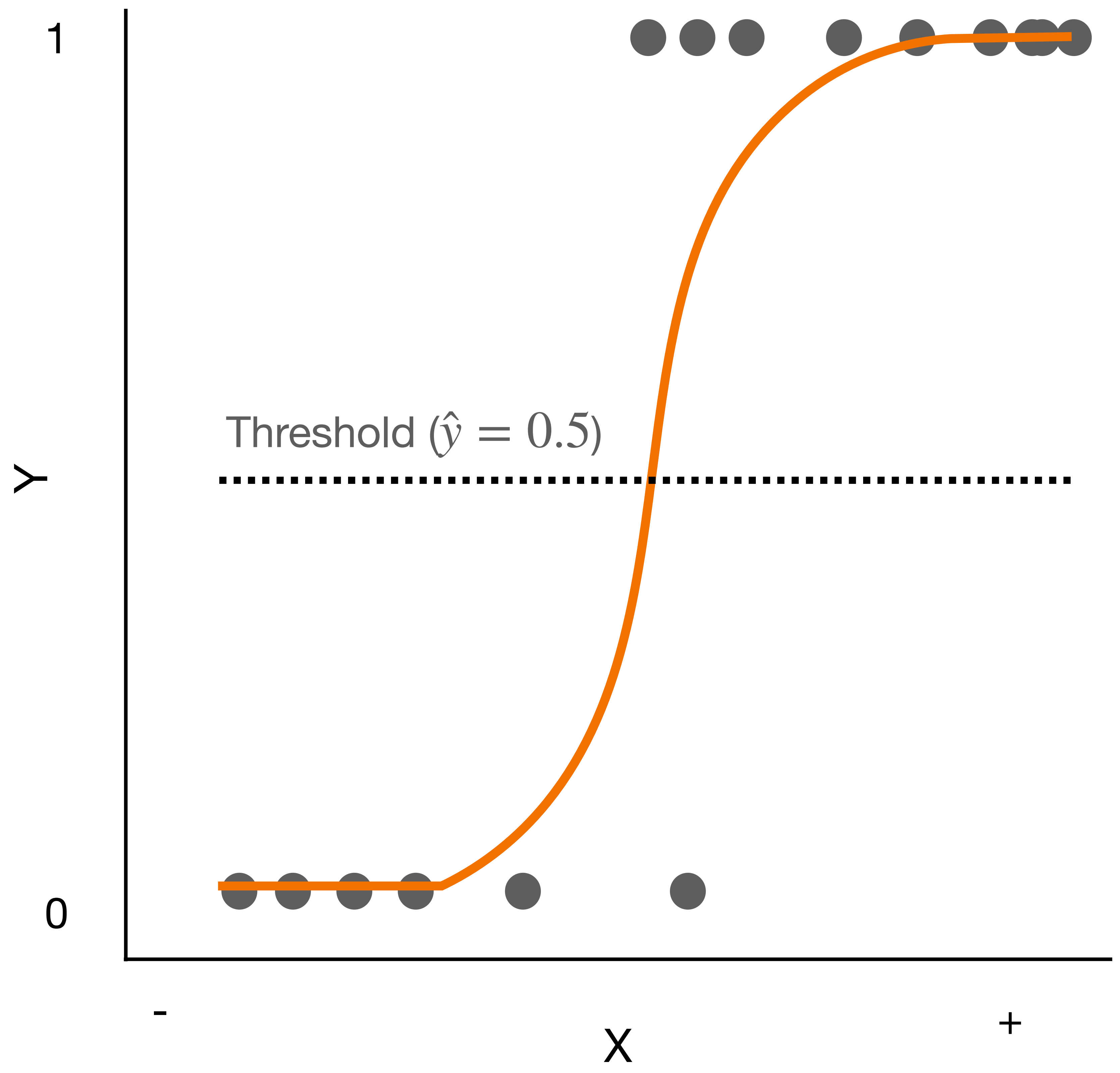
$$\hat{y}_i = e^{\hat{\beta}_0 + \sum_{j=1}^{p} \beta_j x_{i,j}} = e^{1+0.5x_i}$$

For a 1 unit change in $x$, the log odds of $y = 1$ changes by 0.5, multiplied by a baseline odds of $e^1$



(James et al. 2013)

# Prediction




- $\hat{f}(X)$ is still a continuous function.
- In order to predict $y$, an *a priori* threshold needs to be determined to force a category for every $\hat{f}(x_i) = \hat{y}_i$.
- Position of this threshold determines the bias of your predictions.

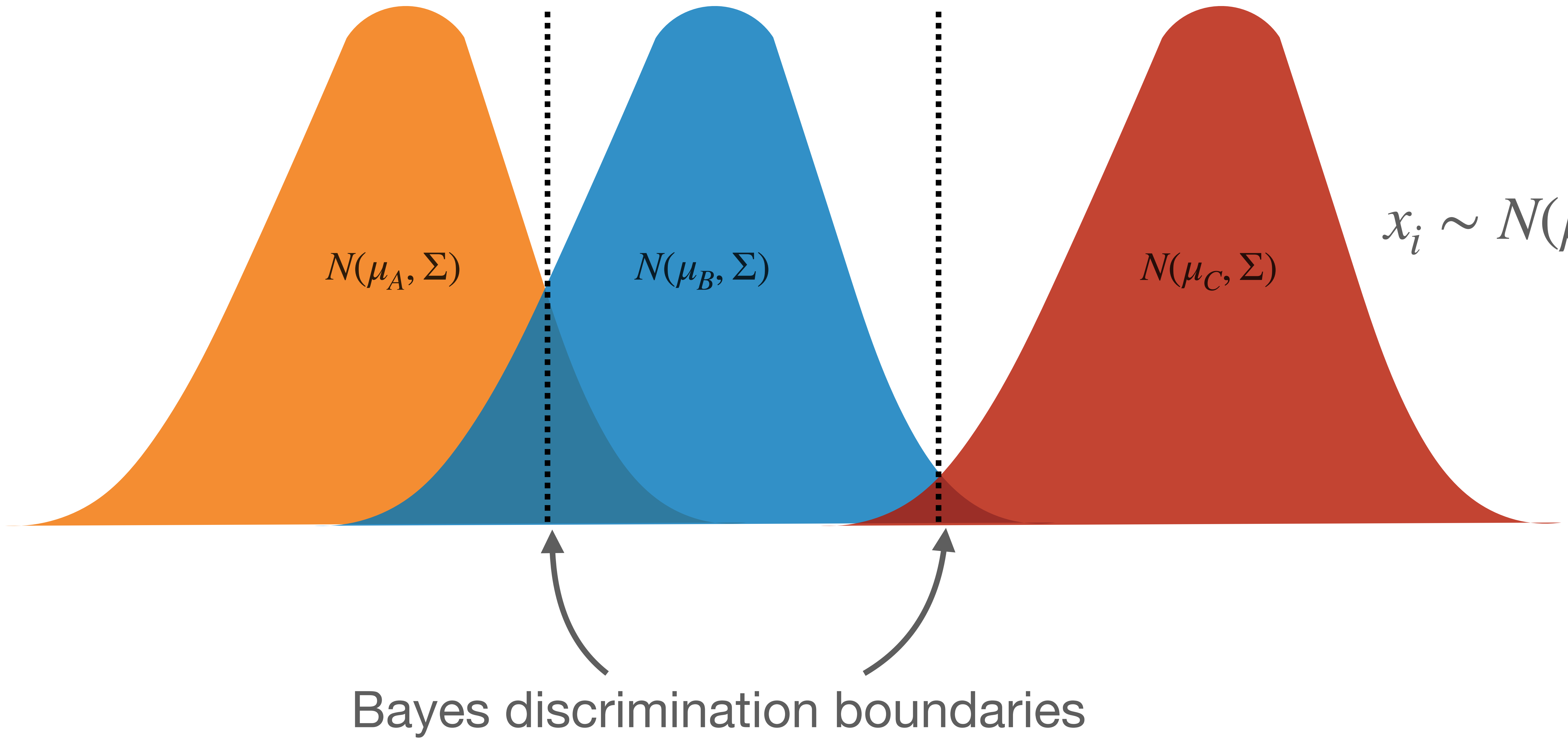


(James et al. 2013)

# Linear & Quadratic Discriminant Analysis

# Discriminant analysis

Q: Which distribution does $x_i$ come from?



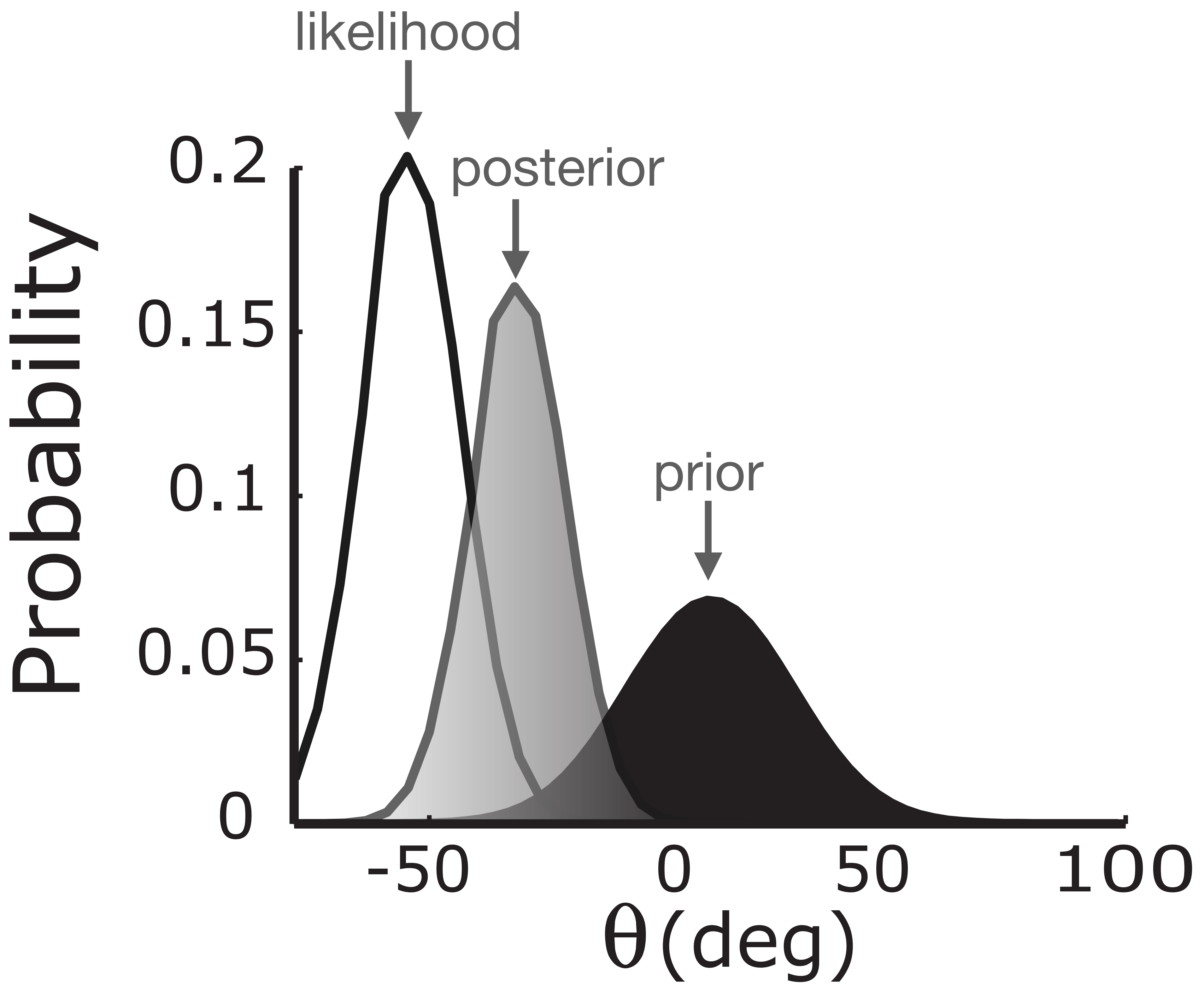

$x_i \sim N(\mu_k, \Sigma)$



(James et al. 2013)

# Bayes Theorem

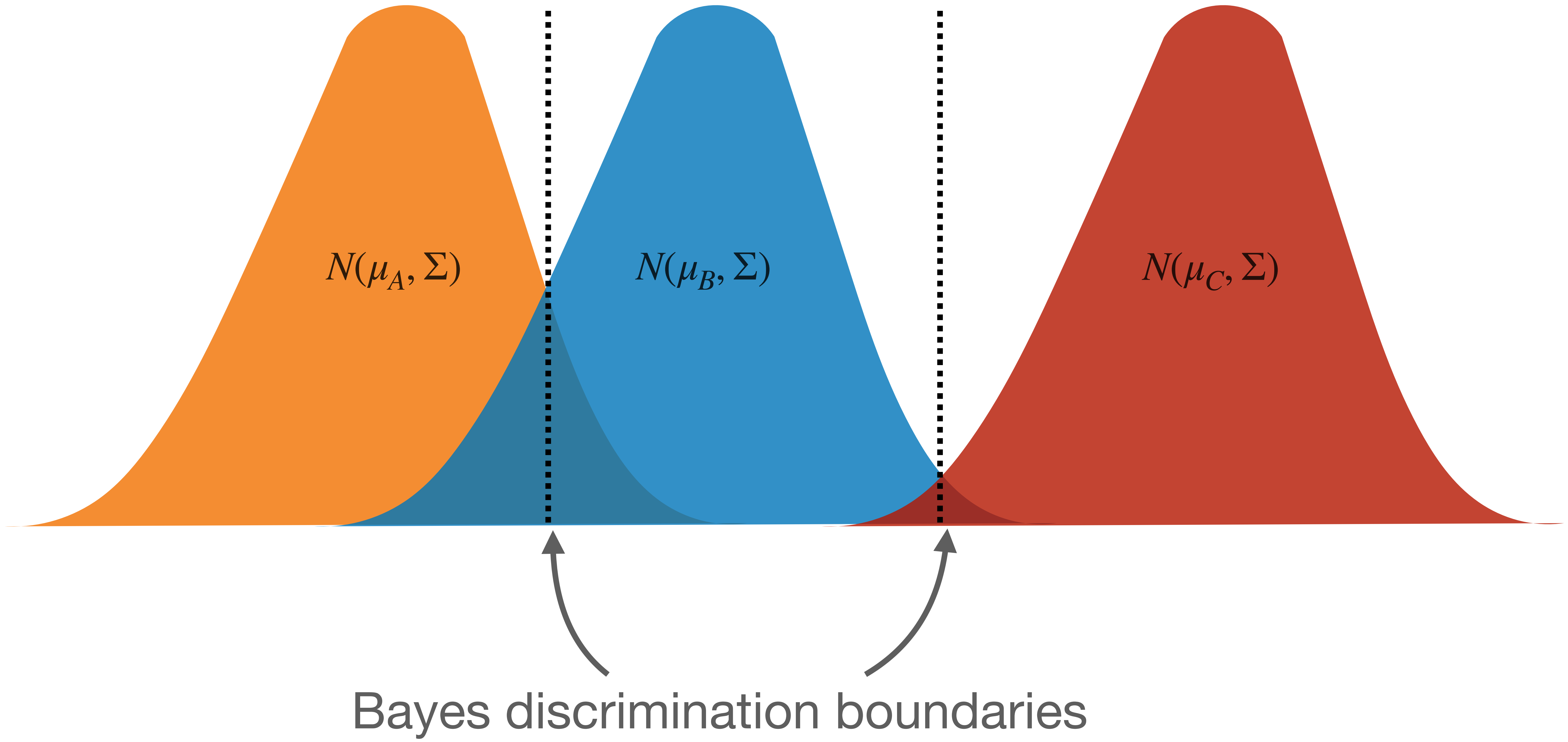
$$\underbrace{P(A\mid B)}_{\text{posterior}} = \frac{\overbrace{P(B\mid A)}^{\text{likelihood}}\ \overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{marginal}}}$$






(James et al. 2013; Verstynen & Sabes 2011)

# Bayesian classifier




$$P(Y = k \mid x_i) = \frac{P(x_i \mid Y = k)P(Y = k)}{P(x_i)}$$

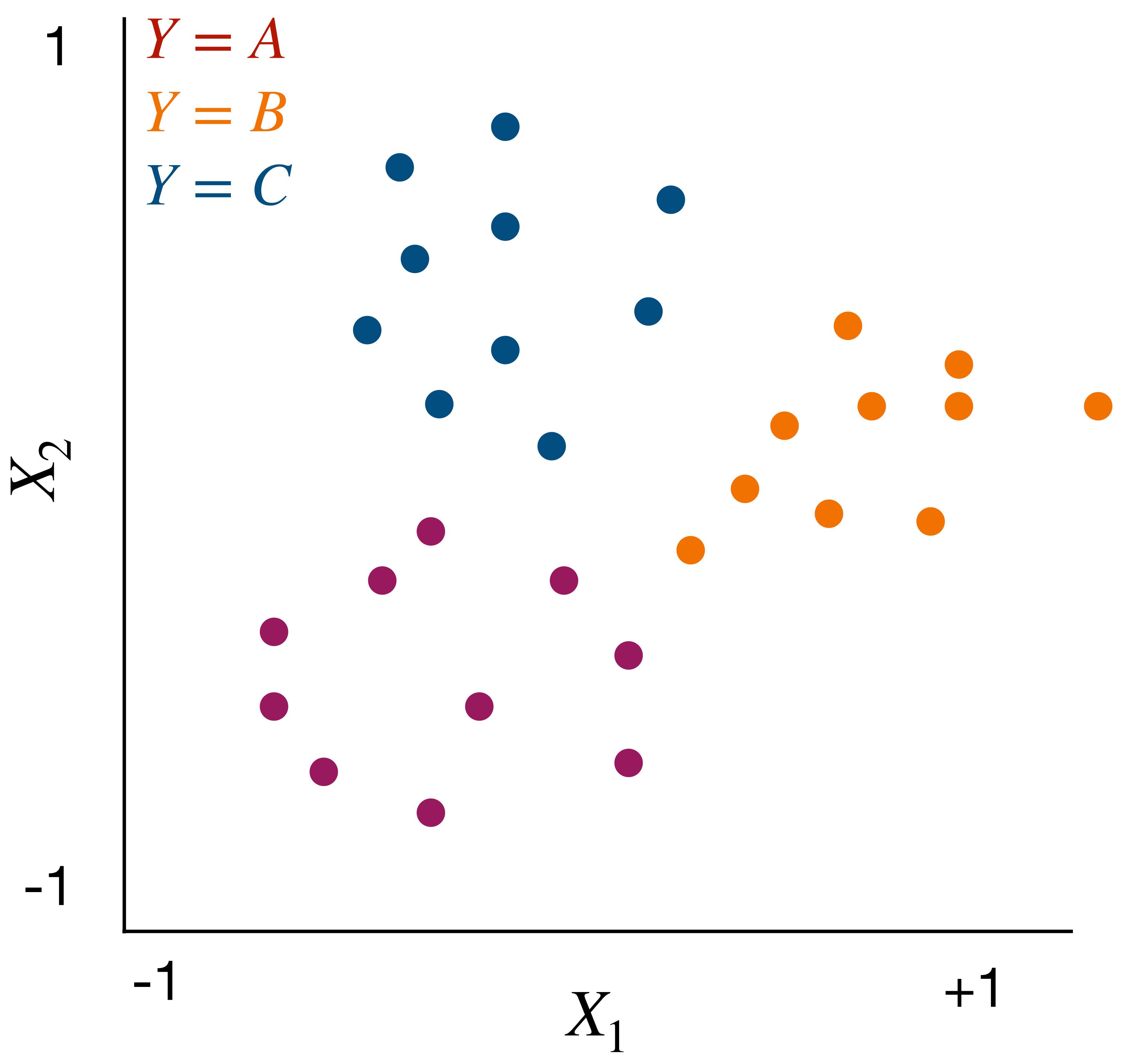
$$\propto P(x_i \mid Y = k)P(Y = k)$$

Use probability theory to infer the separation of the underlying generative distributions.



(James et al. 2013)

# Linear Discriminant Analysis (LDA)



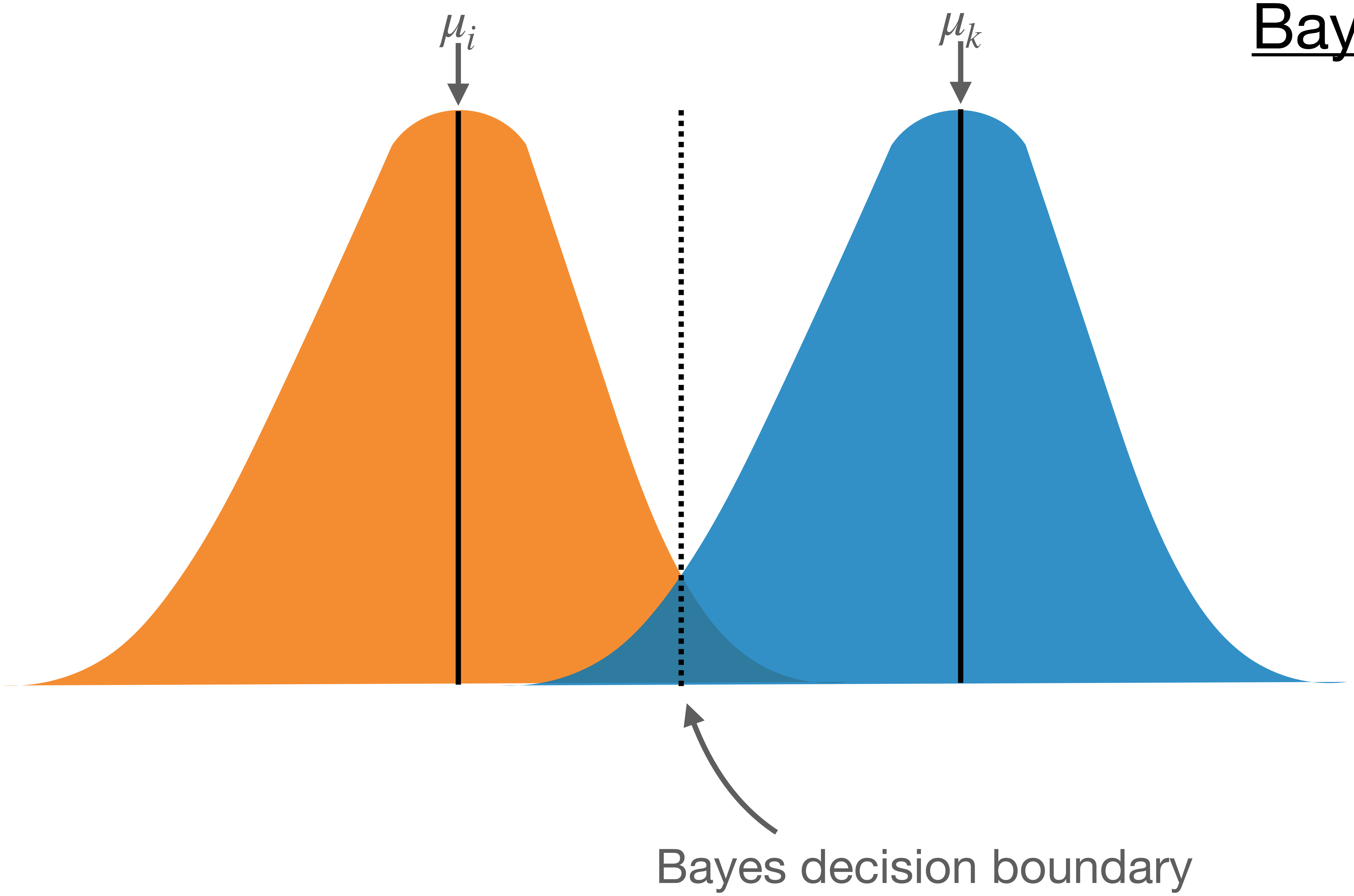

LDA

$\pi_i \sim$ prior for distribution $i$

$$P(Y = k \mid X = x_i) = P_k(x_i) = \frac{\pi_k f_k(x_i)}{\sum_{l=1}^{p} \pi_l f_l(x_i)}$$

$$P_k(x_i) = \frac{\pi_k \frac{1}{\Sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu_k)^2}{2\Sigma^2}}}{\sum_{l=1}^{p} \pi_l \frac{1}{\Sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu_l)^2}{2\Sigma^2}}}$$

# Optimal separation of Gaussians

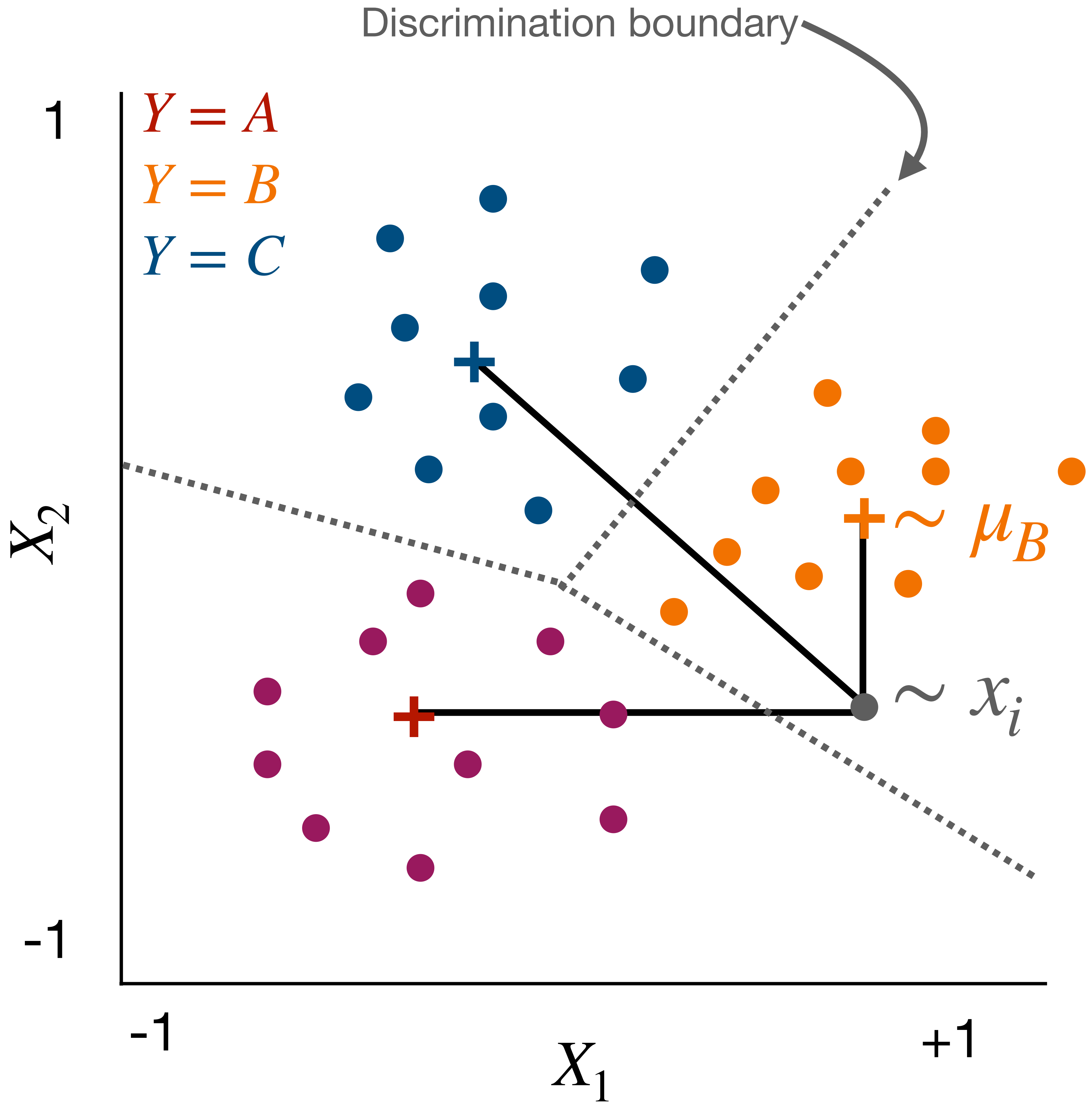



<u>Bayes Decision Boundary</u>

$$\frac{\mu_i^2 - \mu_k^2}{2(\mu_i - \mu_k)} = \frac{\mu_i + \mu_k}{2}$$

Observations on one side of the boundary are group $i$, observations on the other are $k$.

# Group discriminations




<u>Discriminant function</u>

$$\pi_k = \frac{n_k}{n_{all}}$$

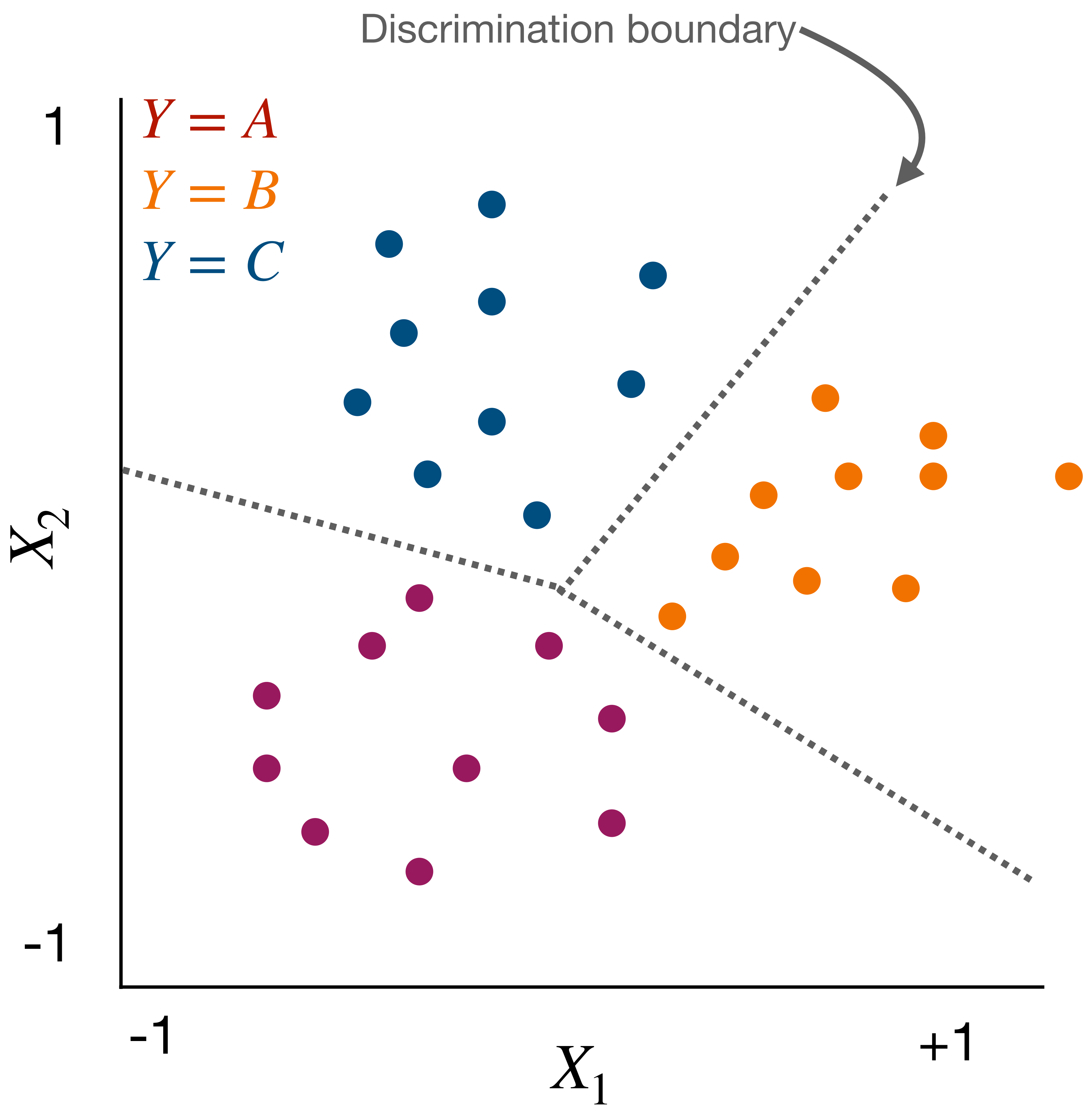
$$\delta_k(x_i) = x_i \frac{\mu_k}{\Sigma^2} - \frac{\mu_k^2}{2\Sigma^2} - ln(\pi_k)$$

$\uparrow \delta_k$ means $x_i$ more likely arises from distribution $k$.



(James et al. 2013)

# Linear Discriminant Analysis (LDA)



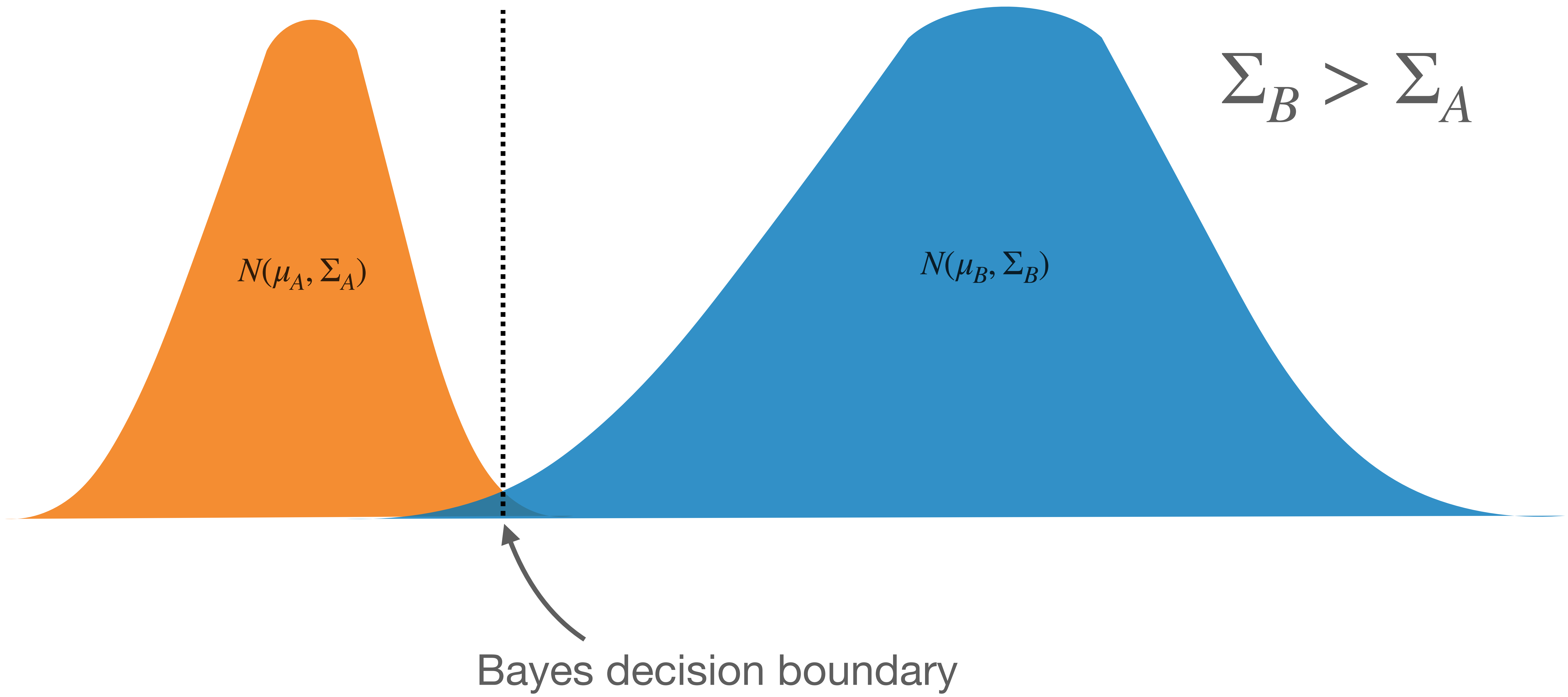

**Assumptions:**

1. $x_i$ is a multivariate normal random variable.
2. All generative distributions have the same variance $\Sigma^2$.
3. There is no collinearity between groups.
4. Errors are independent.



(James et al. 2013)

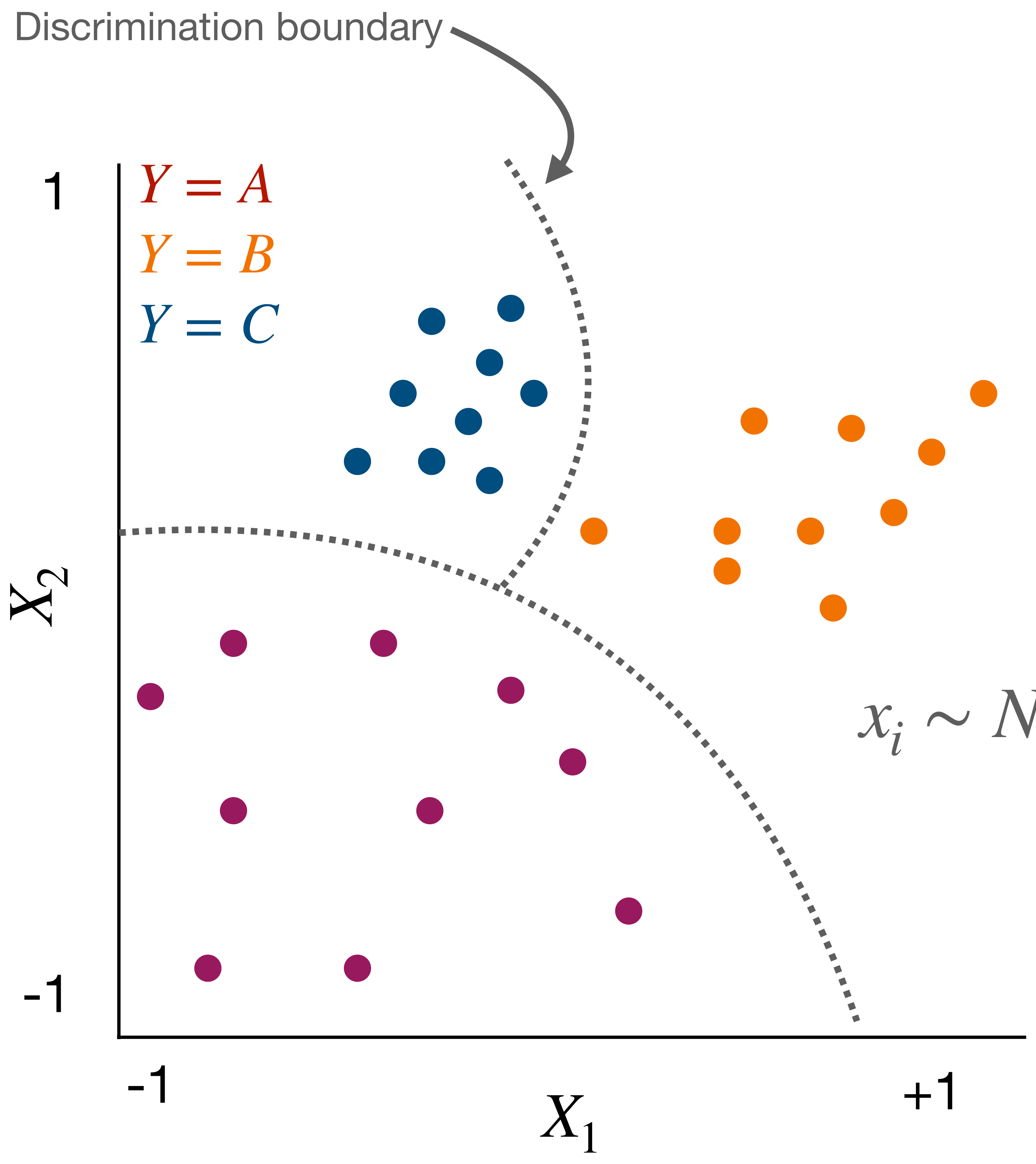
# What if variances are not the same?






(James et al. 2013)

# Quadratic Discriminant Analysis (QDA)




$x_i \sim N(\mu_k, \Sigma_K)$

Remove the assumption of equal variances and the decision boundaries become curves, instead of straight lines.

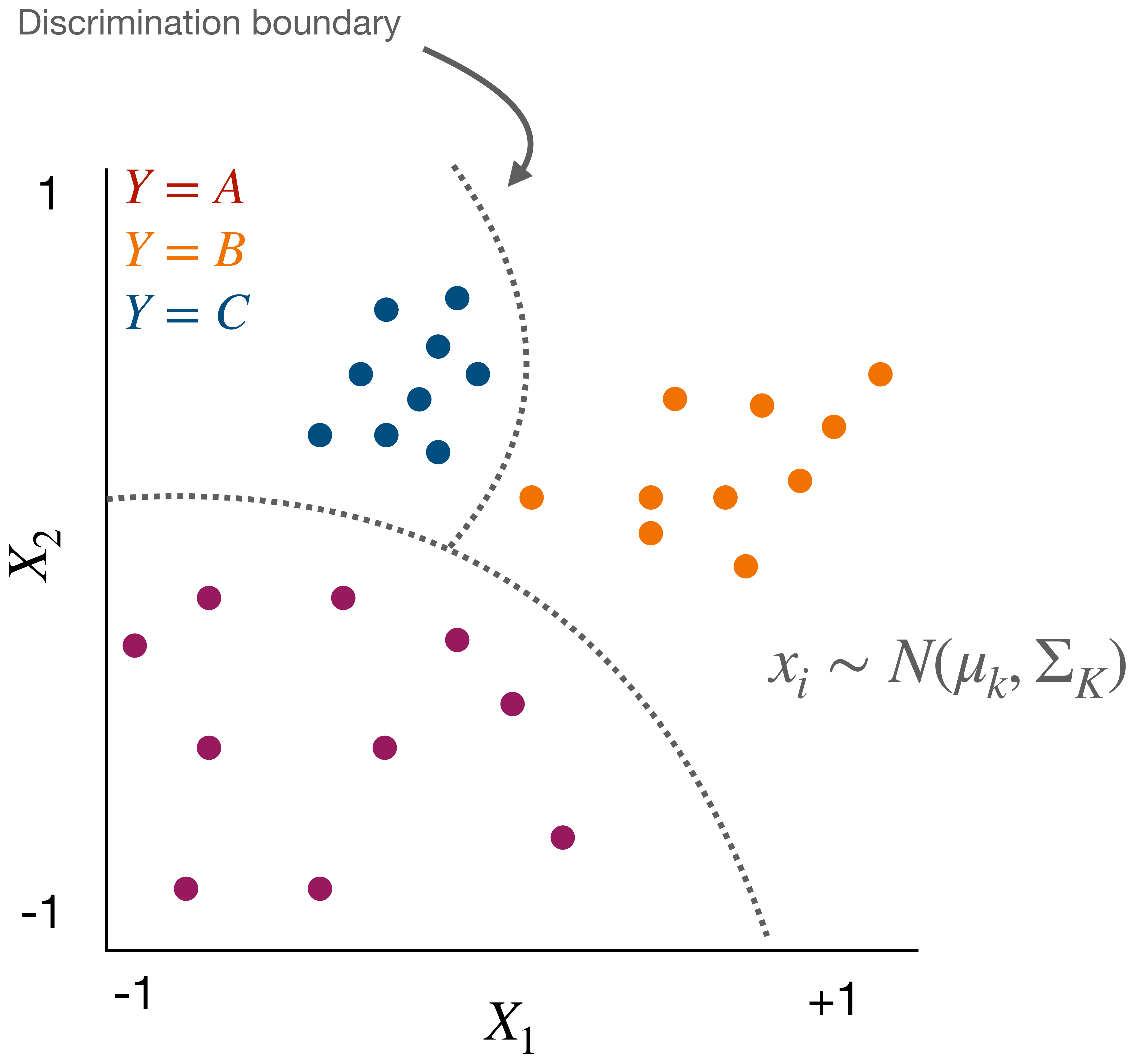
## Discriminant function

$$\delta_k(x_i) = -\frac{(x_i - \mu_k)}{2\Sigma_k(x_i - \mu_k)} - \frac{1}{2} ln(\Sigma_k) + ln(\pi_k)$$



(James et al. 2013)

# Quadratic Discriminant Analysis (QDA)




## Assumptions:

1. $x_i$ is a multivariate normal random variable.
2. There is no collinearity between groups.
3. Errors are independent.



(James et al. 2013)

# Take home message

- While logistic regression offers an intuitive extension to linear regression for classification problems, discriminant analyses are more flexible for classification in cases where your data consists of more than 2 categories.