Basic Statistics for Machine Learning

You Could Call It "Statistical Learning"

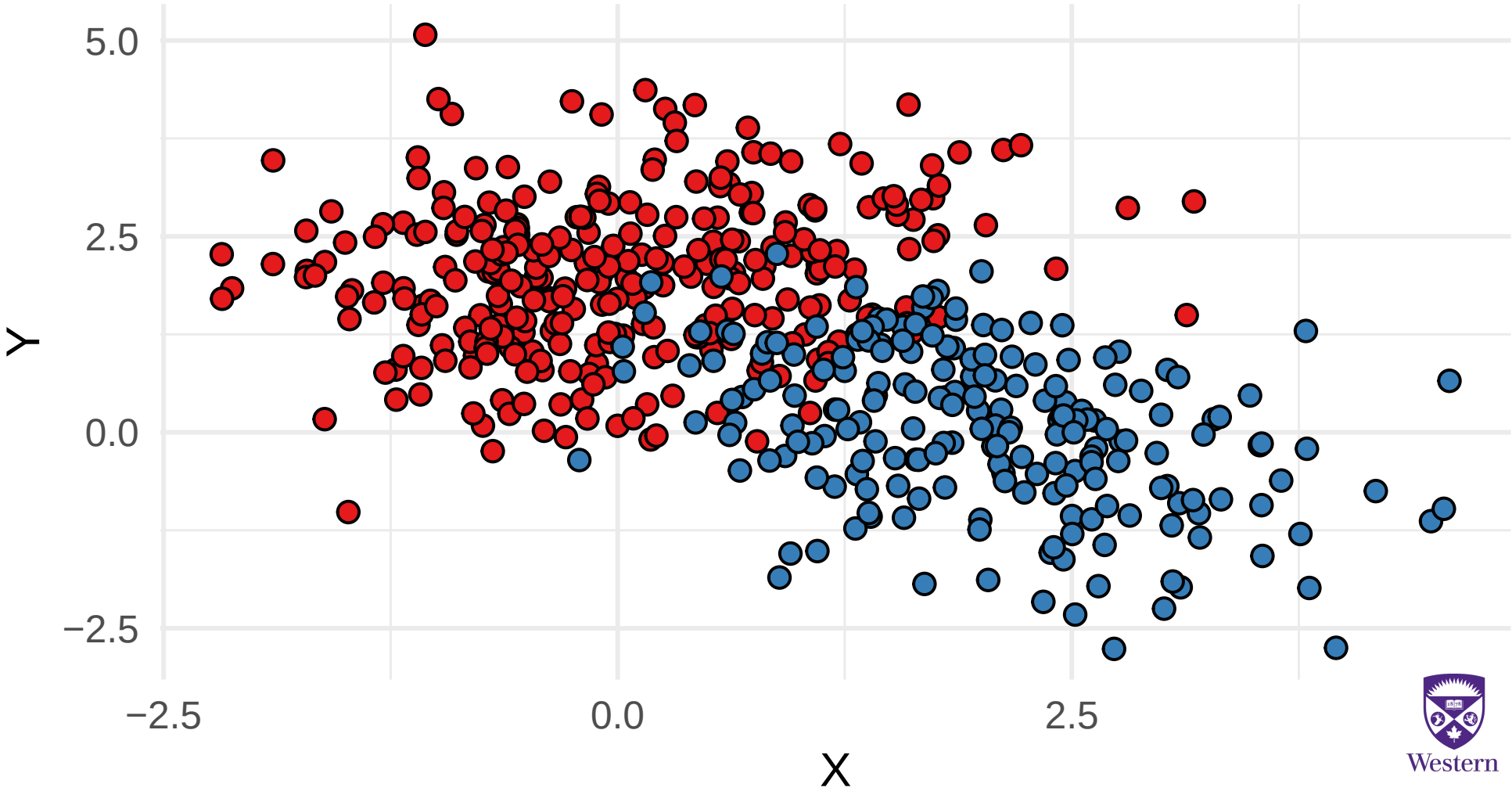
Demetri Pananos

Western

Department of Epidemiology & Biostatistics
Schulich School of Medicine & Dentistry
Western University

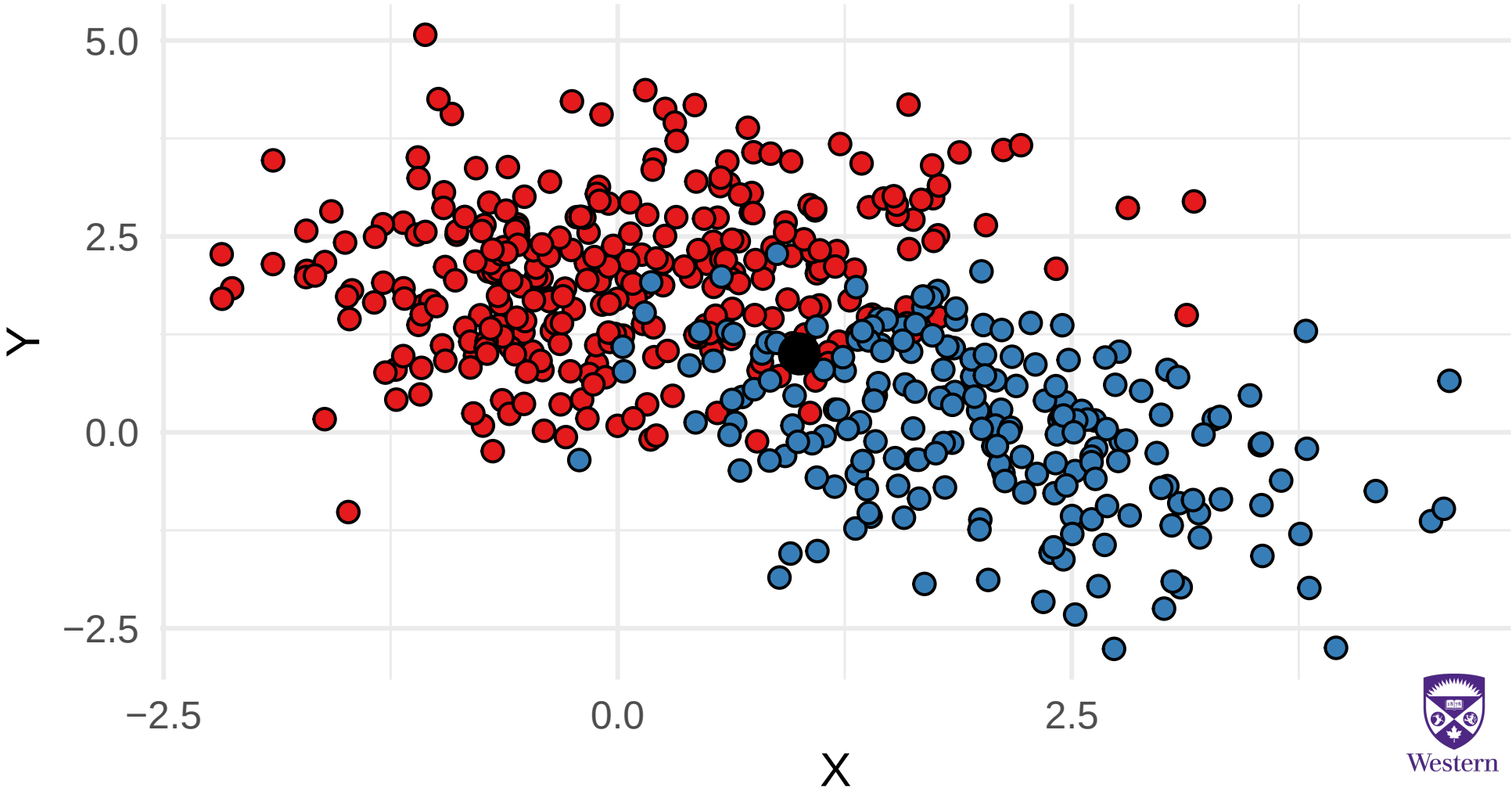
2018-10-31

I Have A Problem



Western

I Have a Problem



Western

How Can I Solve It?

- We can do this in a variety of ways.
- Some are very statistical. Some are very complicated.
- Some are in the middle. I'll call those *statistical learning algorithms*.

Western

Statistical Decision Theory

- The "Theorectically Optimal" way to do this would require knowing $P(\text{Is Red}|X, Y)$.
- Maybe approximating this conditional distribution will be enough.
- But first, a little vocabularty

Western

Vocabulary

Marginal

$P(X)$

Joint

$P(X, Y)$

Conditional

$P(X|Y)$

Probability

Sum Rule

$$P(X) = \sum_{Y} P(X, Y)$$

Product/Chain Rule

$$P(X, Y) = P(Y|X)P(X) \stackrel{or}{=} P(X|Y)P(Y)$$

Probability

We can recover some very powerful rules just from these. For example, Bayes' Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_{Y} P(X|Y)P(Y)}$$

Western

Naive Bayes

Use Bayes' Theorem to compute $P(\text{Is Red}|X, Y)$.

$$P(\text{Is Red}|X, Y) \propto P(\text{Is Red})P(X, Y|\text{Is Red})$$

Naive Bayes

$$P(\text{Is Red}) = \text{The proportion of our sample which belongs to Red}$$

$$P(X, Y|\text{Is Red}) = \text{Distribution of data belonging to Red}$$

Western

The Naive Part

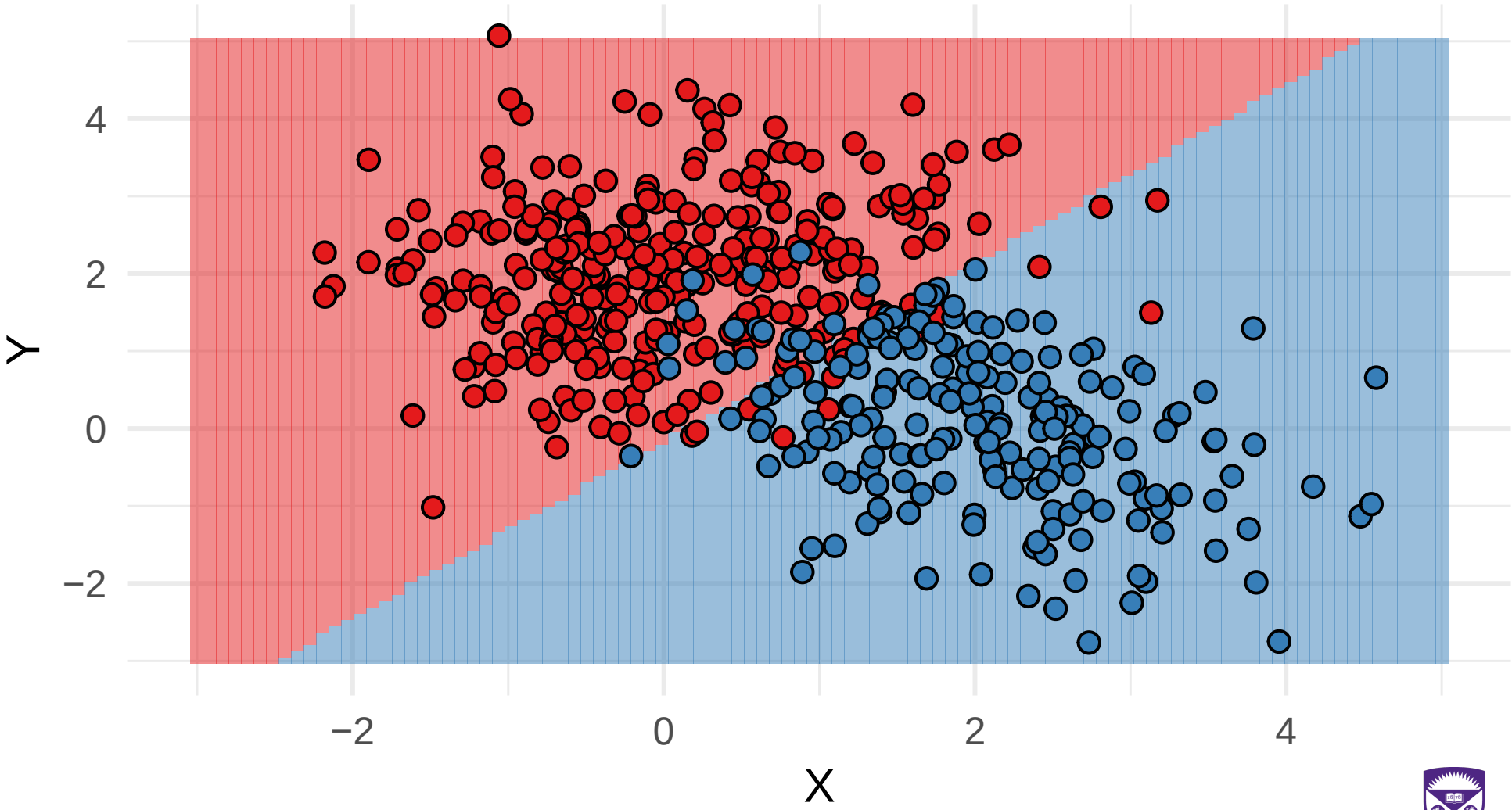
The Naive Part of Naive Bayes

$$P(X, Y|\text{Is Red}) = P(X|\text{Is Red})P(Y|\text{Is Red})$$

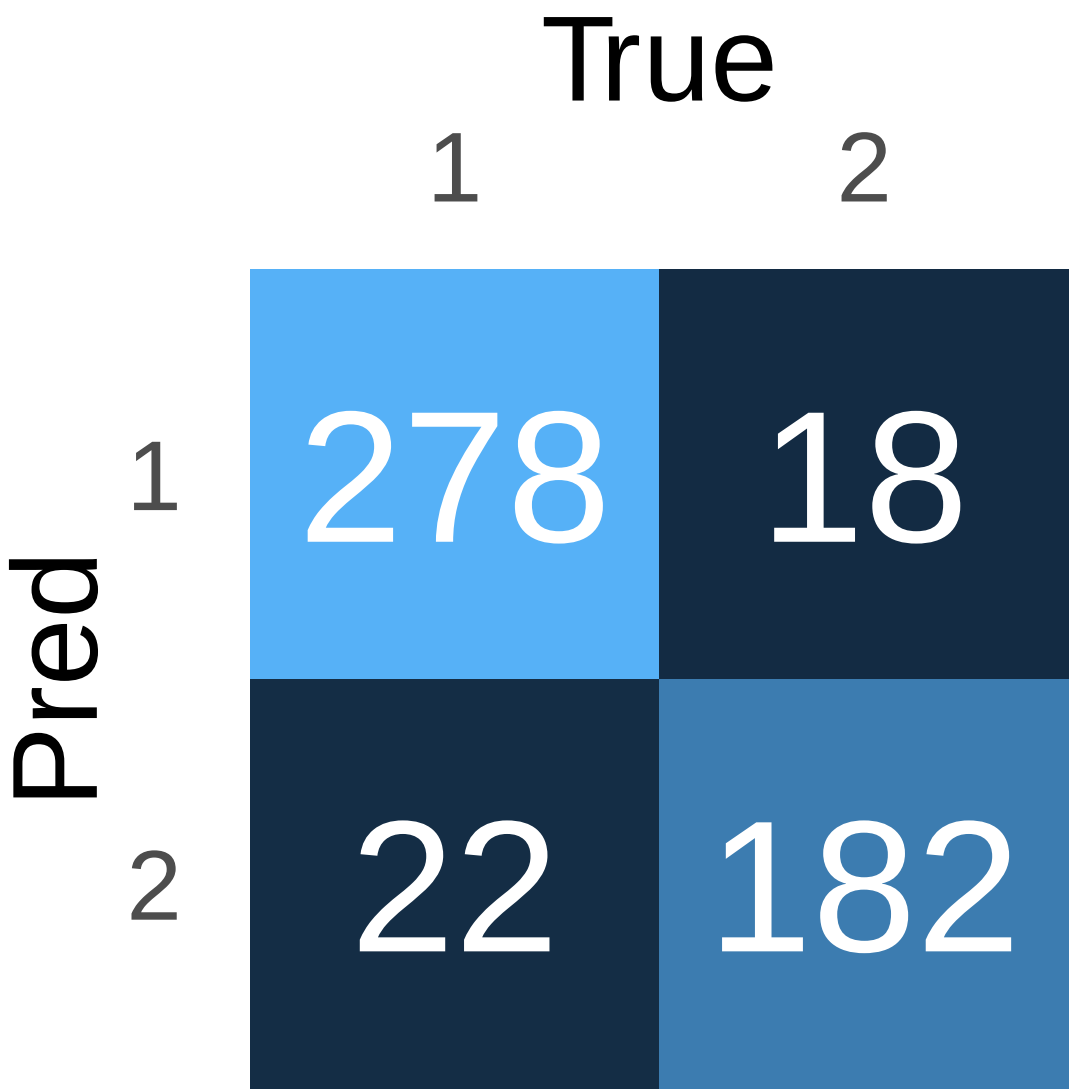
All in All

$$P(\text{Is Red}|X, Y) = P(\text{Is Red})P(X|\text{Is Red})P(Y|\text{Is Red})$$

Make another assumption that $P(X|\text{Is Red})$ is normal.



Western


True
1
2
Pred
1
2
278
18
22
182

Quadratic Discriminant Analysis

Naive Bayes ignores the covariance between data by assuming independence between covariates.
Quadratic Discriminant Analysis will use some information about covariance to help make predictions. In order to fit these models, we need to understand

- Expectation in $\mathbb{R}^n$
- Covariance Matrices

Western

Means and Covariance

- Expectations in multiple dimenions behave like expectations in a single variable.
- If X is an n-dimensional random variable with expectation $\boldsymbol{\mu}$, then

$$E[X] = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \approx \begin{bmatrix} \sum_k \frac{x_{1k}}{N} \\ \vdots \\ \sum_k \frac{x_{nk}}{N} \end{bmatrix}$$

Western

Means and Covariance

- If X is an n-dimensional random variable with covariance matrix Σ, then

$$\Sigma_{ij} = \text{Cov}(X_i, X_j)$$

.

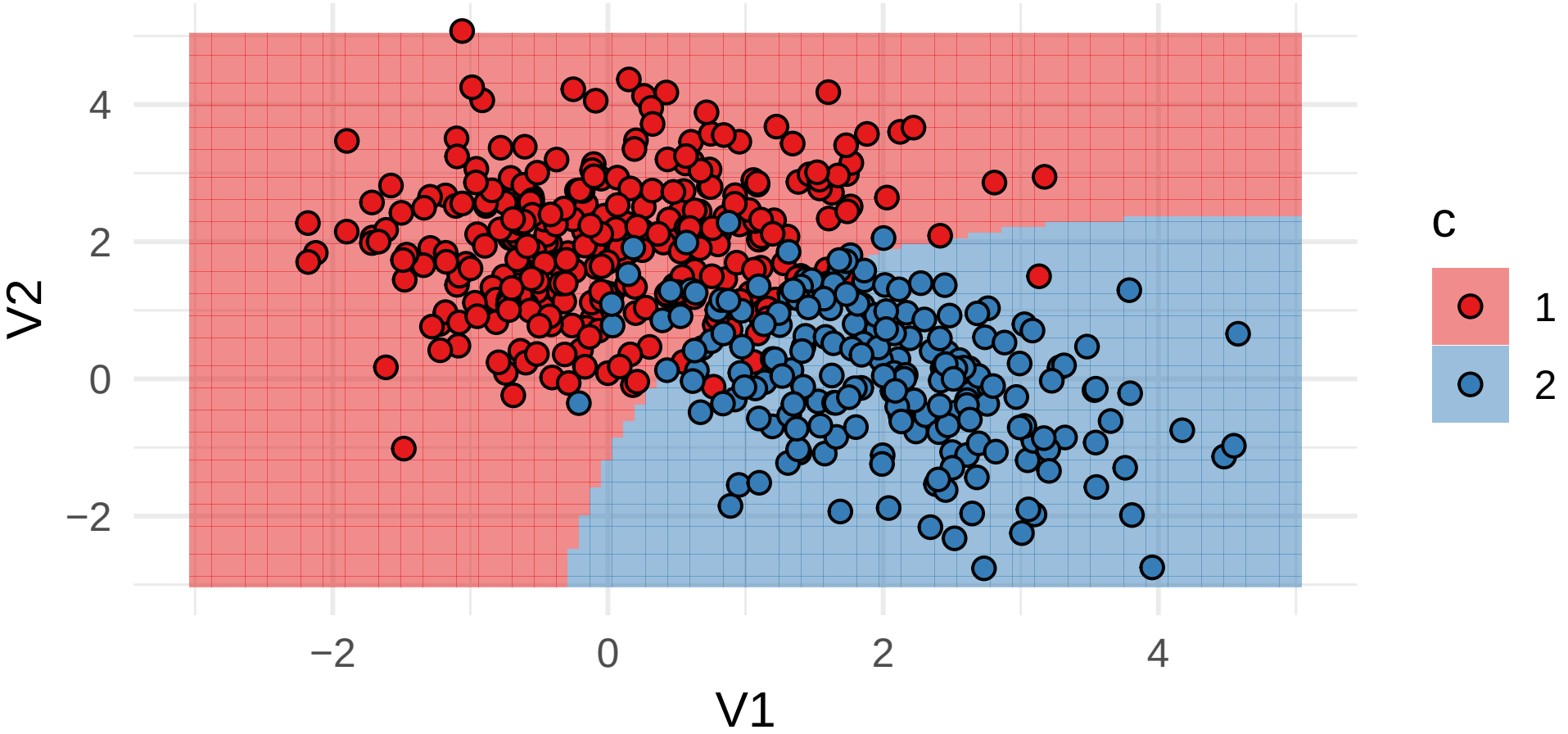
- This implies $\Sigma^T = \Sigma$ since the covariance is a symmetric operator.

Western

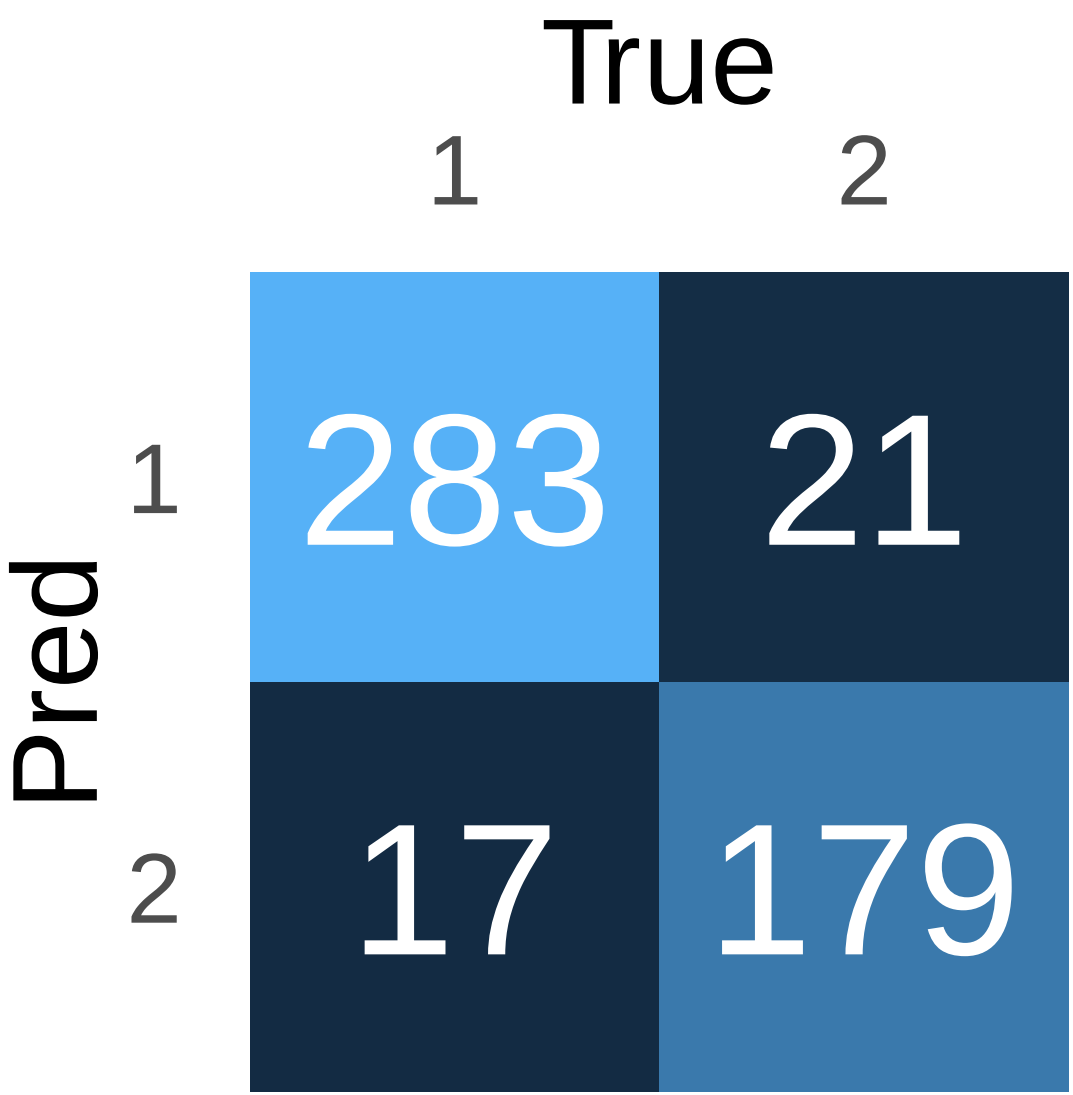
Back To Quadratic Discriminant Analysis

- Assume that $X|C_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$. That is, assume that the two classes have different covariance and different means.
- Compute $P(C = C_k|X)$ using Baye's rule.

Western



Western


True
1
2
Pred
1
2
283
21
17
179

What Have We Seen?

- A little statistics can get you very far.

Western