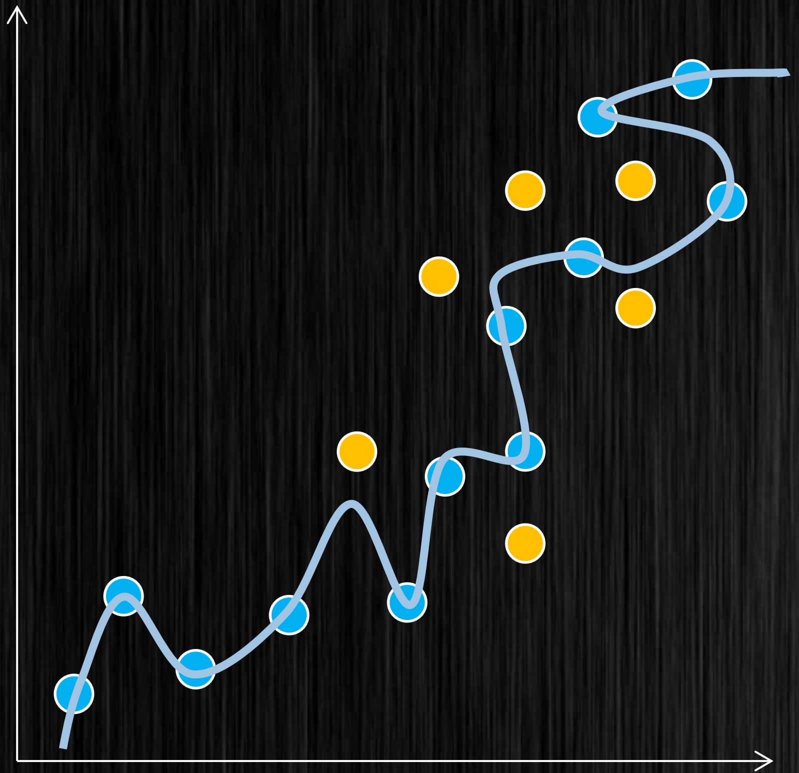
Variance

Variance

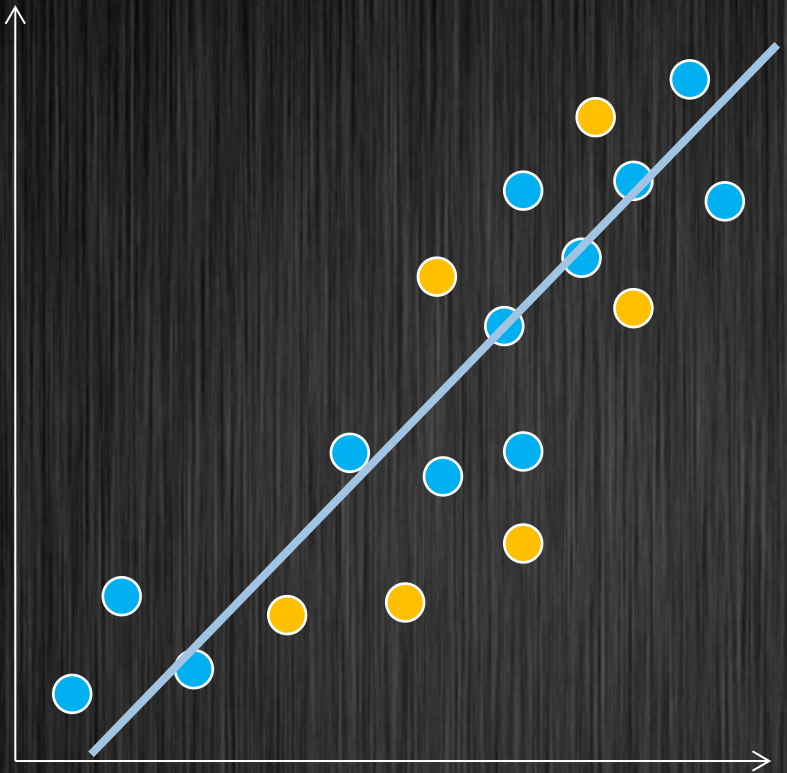
Chirag Rathi
A023119820021

Definition:

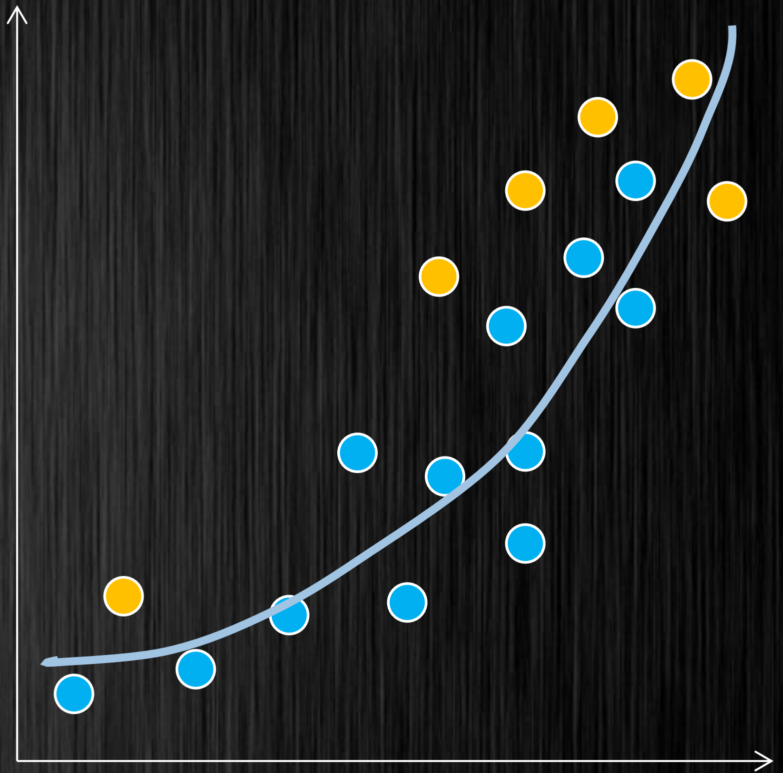
Variance refers to **the changes in the model when using different portions of the training data set**. Simply stated, variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set.



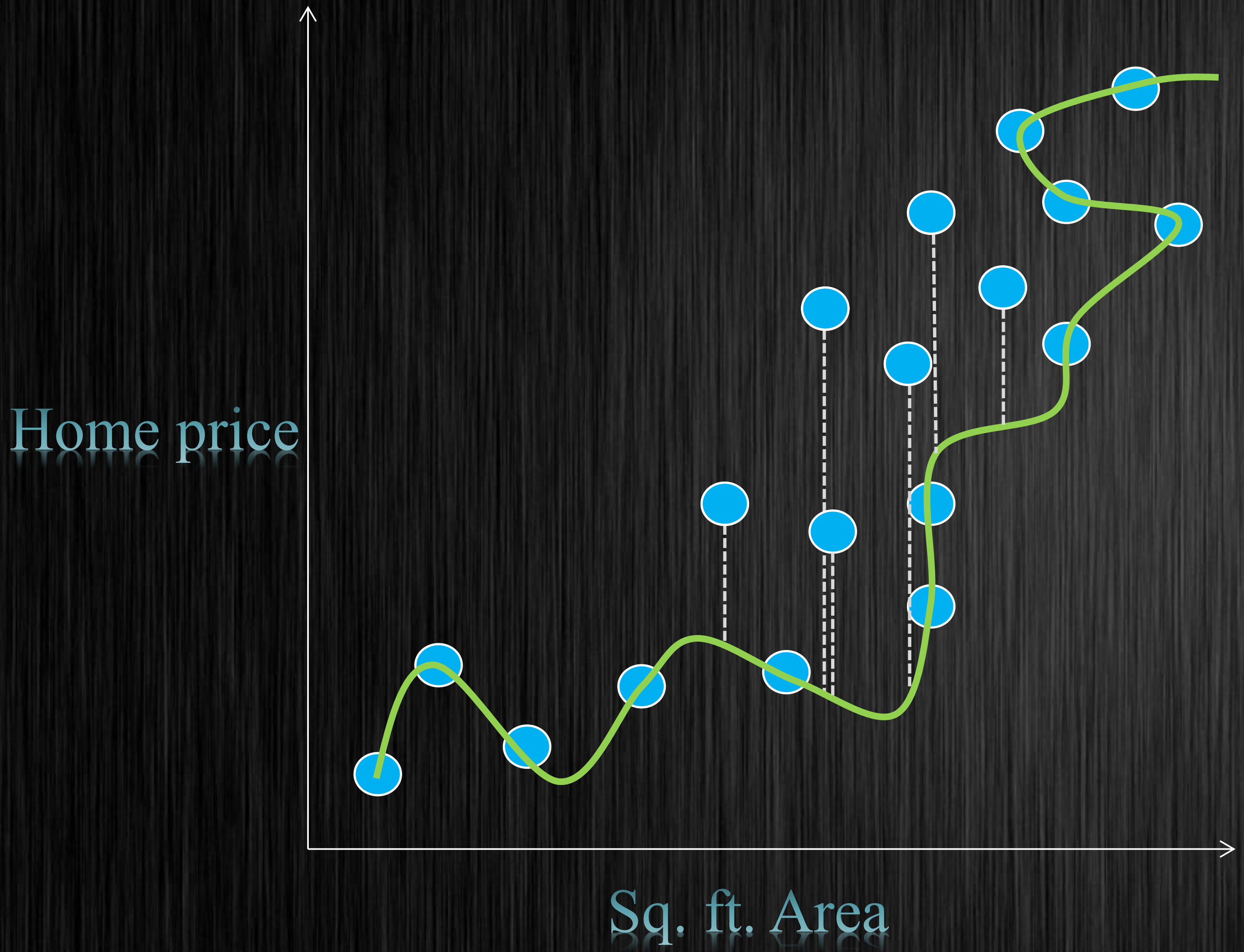
overfit



underfit

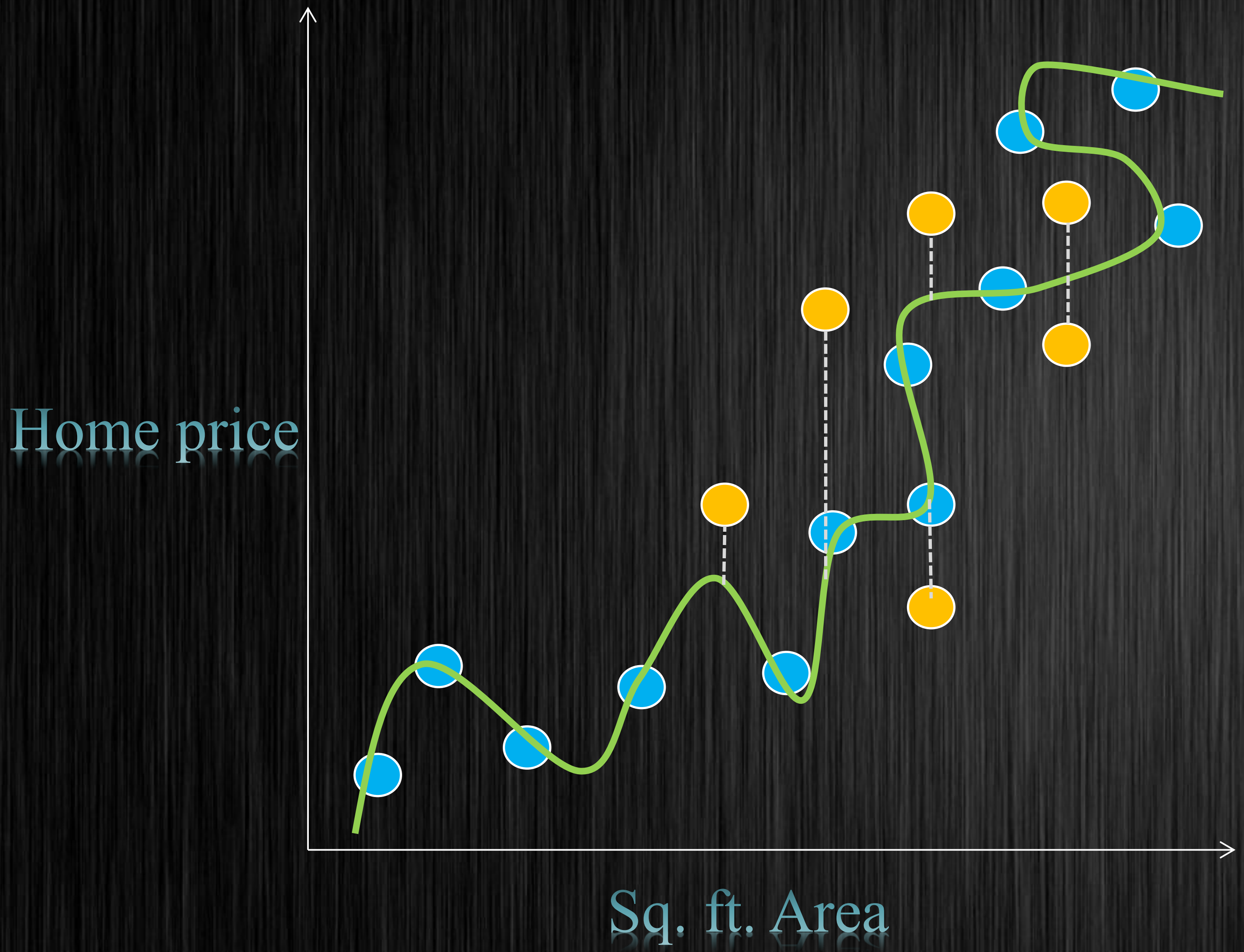


Balanced fit



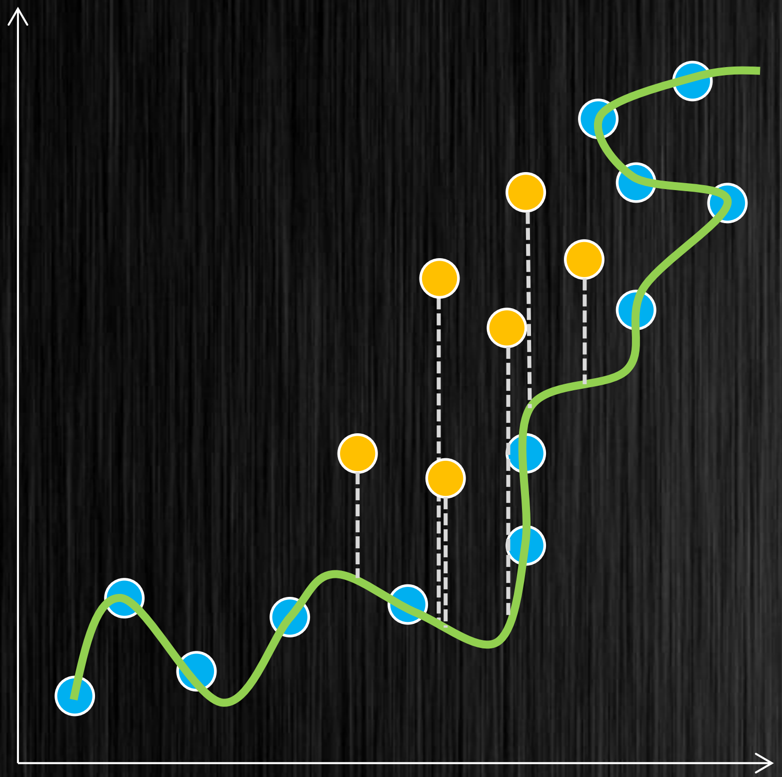
Training dataset error = 0

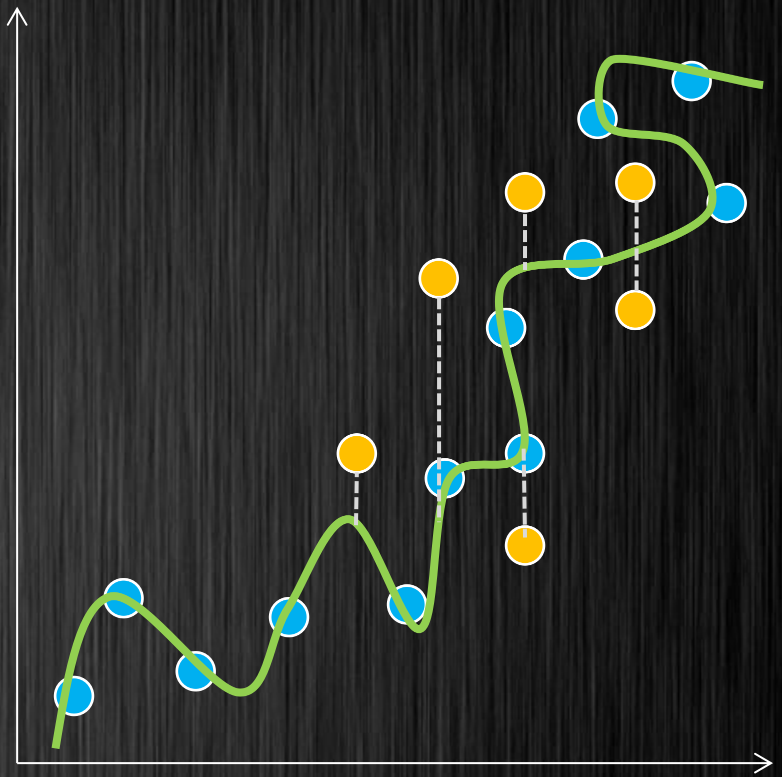
Test dataset error = 100



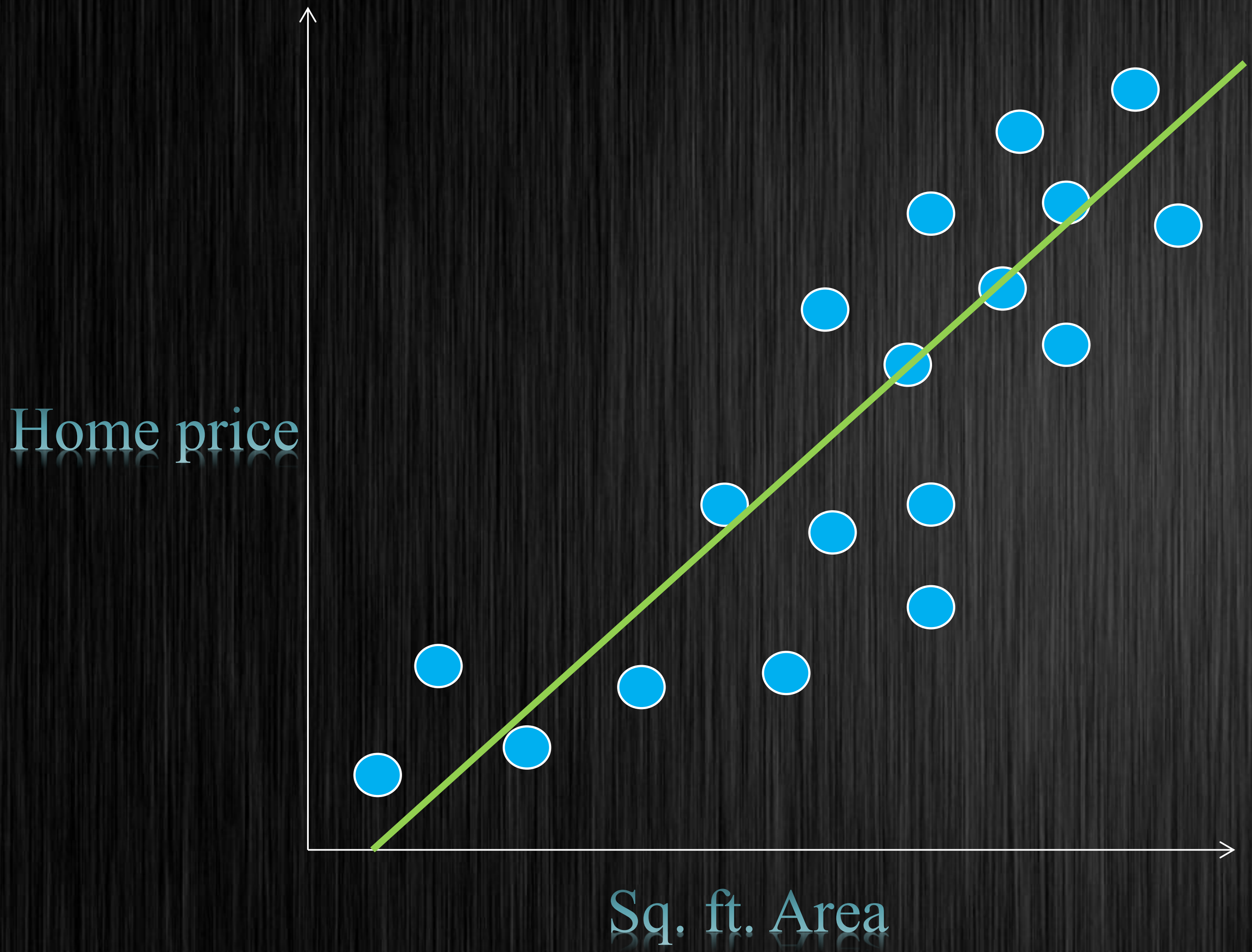
Training dataset error = 0

Test dataset error = 27



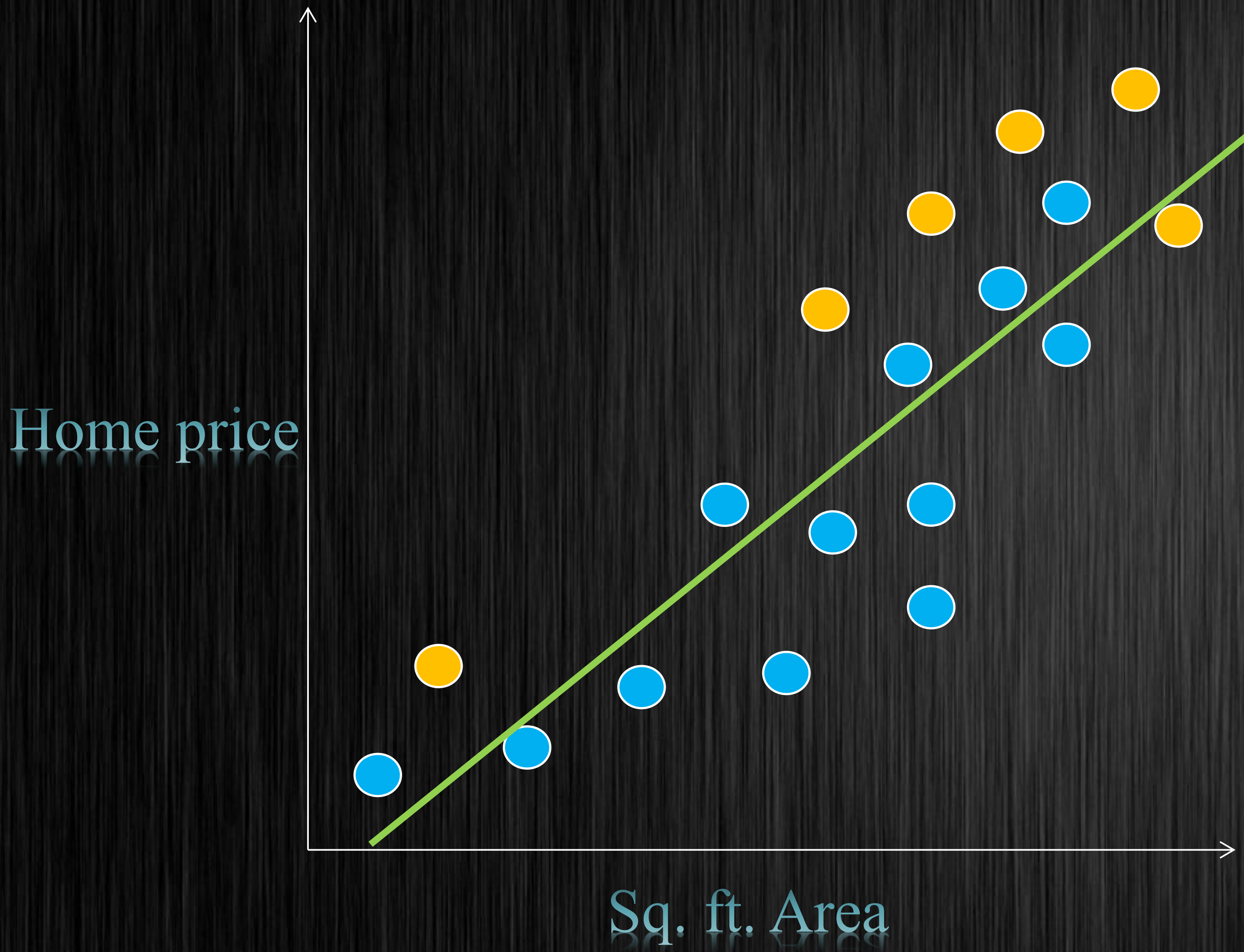


Test Error = 100 ← **High variance** → Test Error = 27



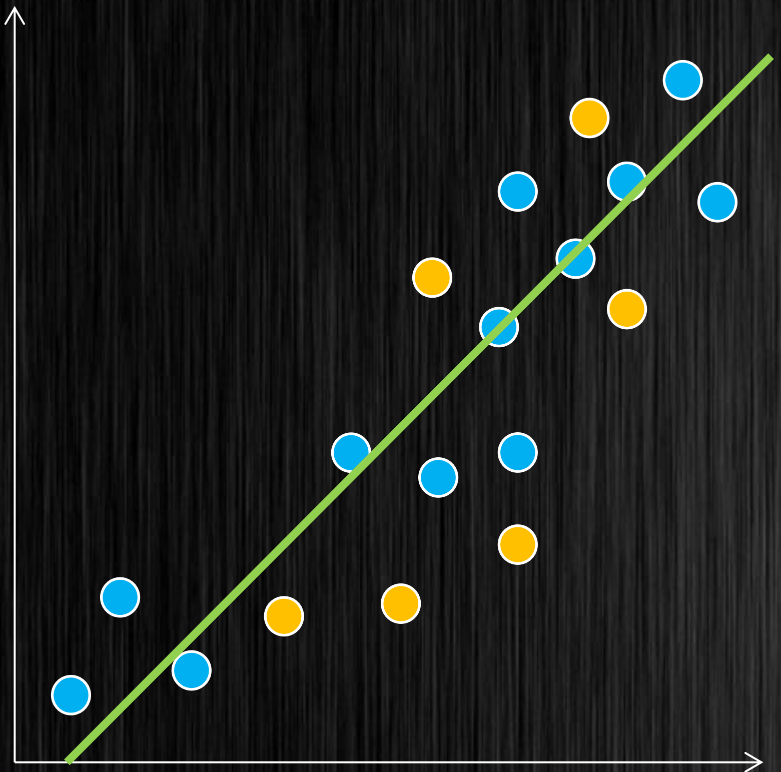
Training dataset error = 43

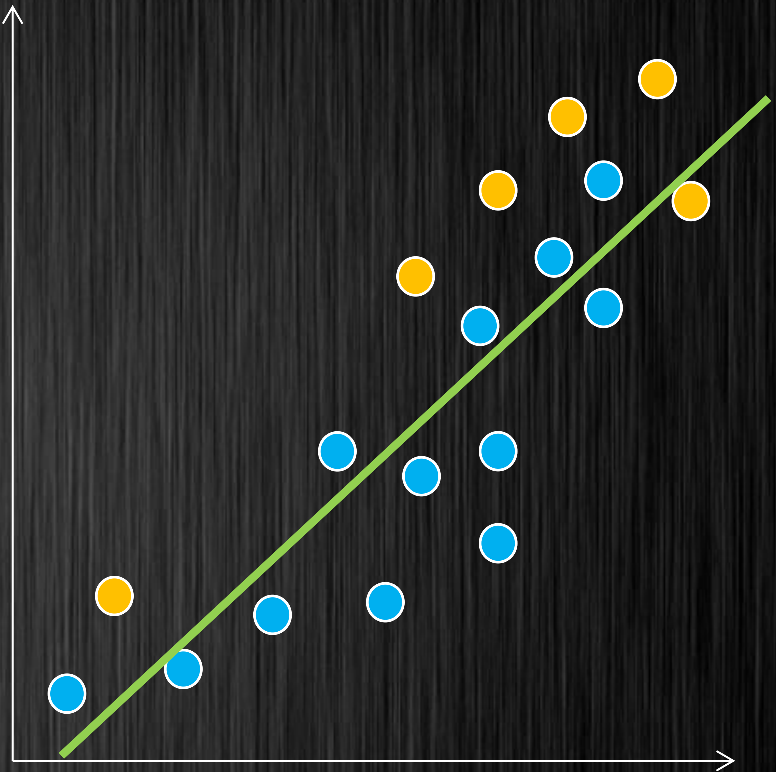
Test dataset error = 47



Training dataset error = 41

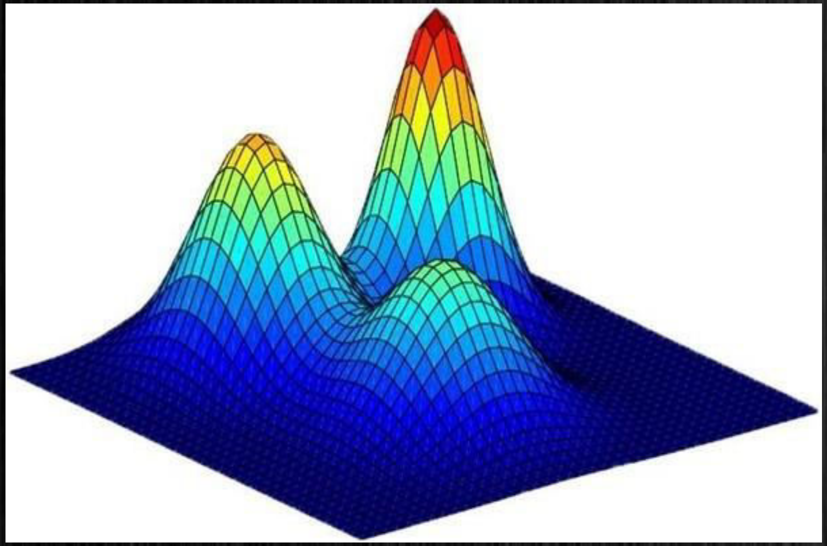
Test dataset error = 37

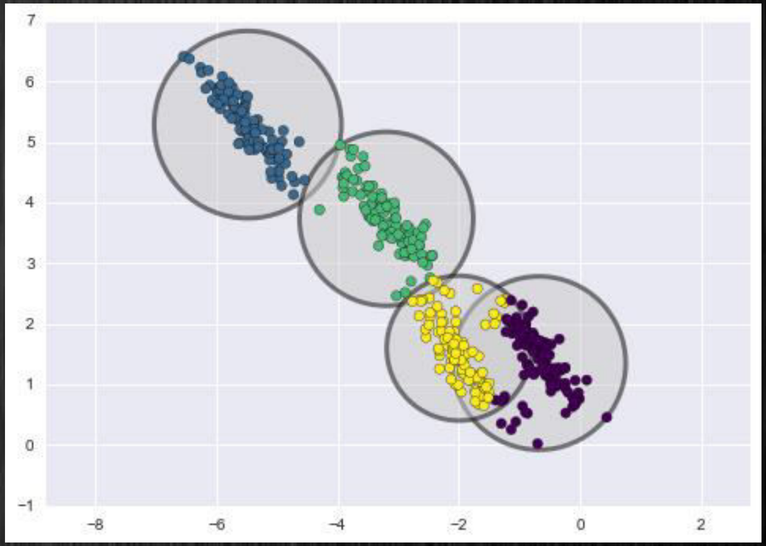




Test Error = 47 ← **Low variance** → Test Error = 37

Model – Based Clustering





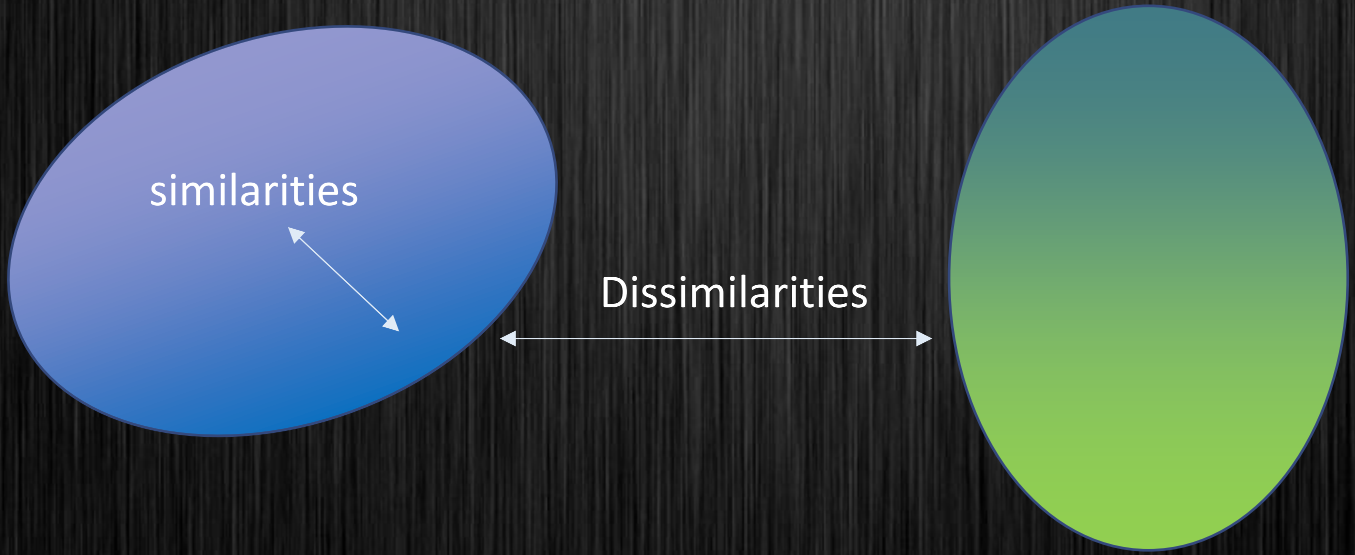
Gaussian Mixture Models

Definition:

Model-based clustering is **a statistical approach to data clustering**. The observed (multivariate) data is assumed to have been generated from a finite mixture of component models. Each component model is a probability distribution, typically a parametric multivariate distribution.

What is Clustering

The procedure of portioning a set of observations into a set of meaningful subclasses



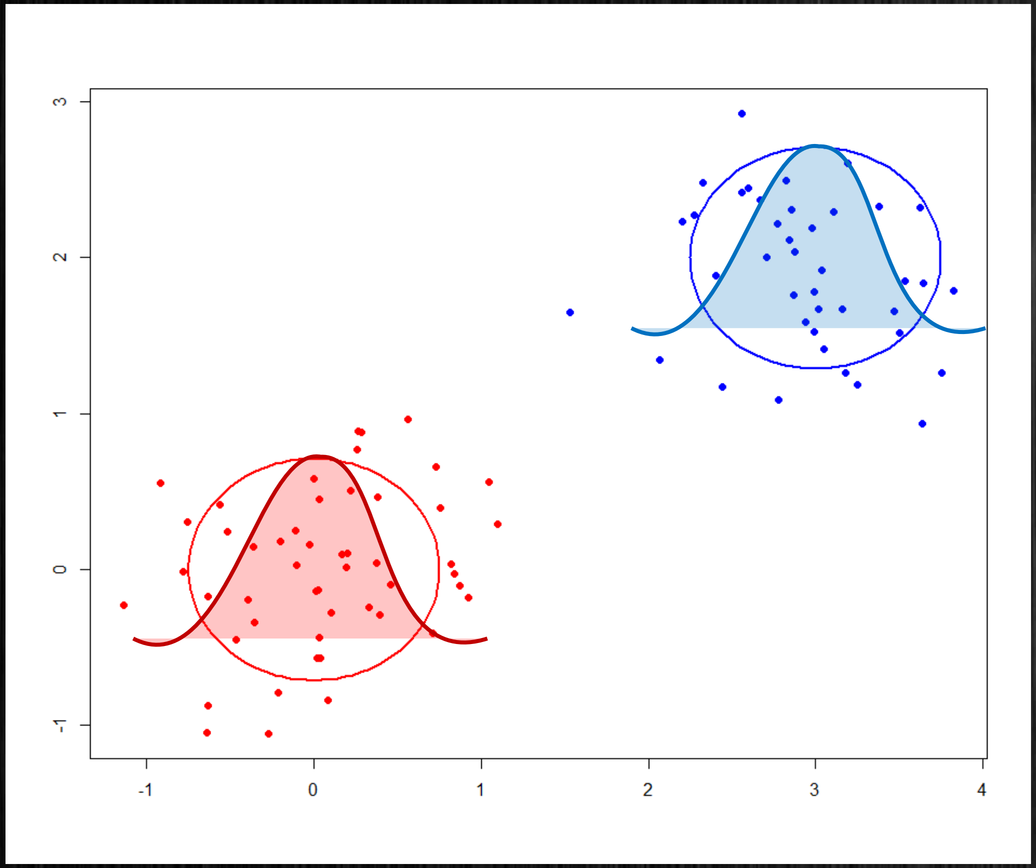
Applications of Clustering

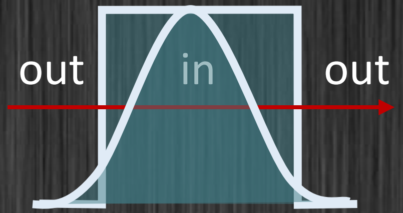
- Medicine
  - Ex. In medical imaging to distinguish between different types of tissues

- Business
  - Ex. To discover distinctive group of customers to develop targeted marketing programs

- Social Sciences
  - Ex. To identify zones in a city by the type of committed crimes to manage law enforcement resources more efficiently

How do we “see” clusters





This is just one example of Model – Based Clustering

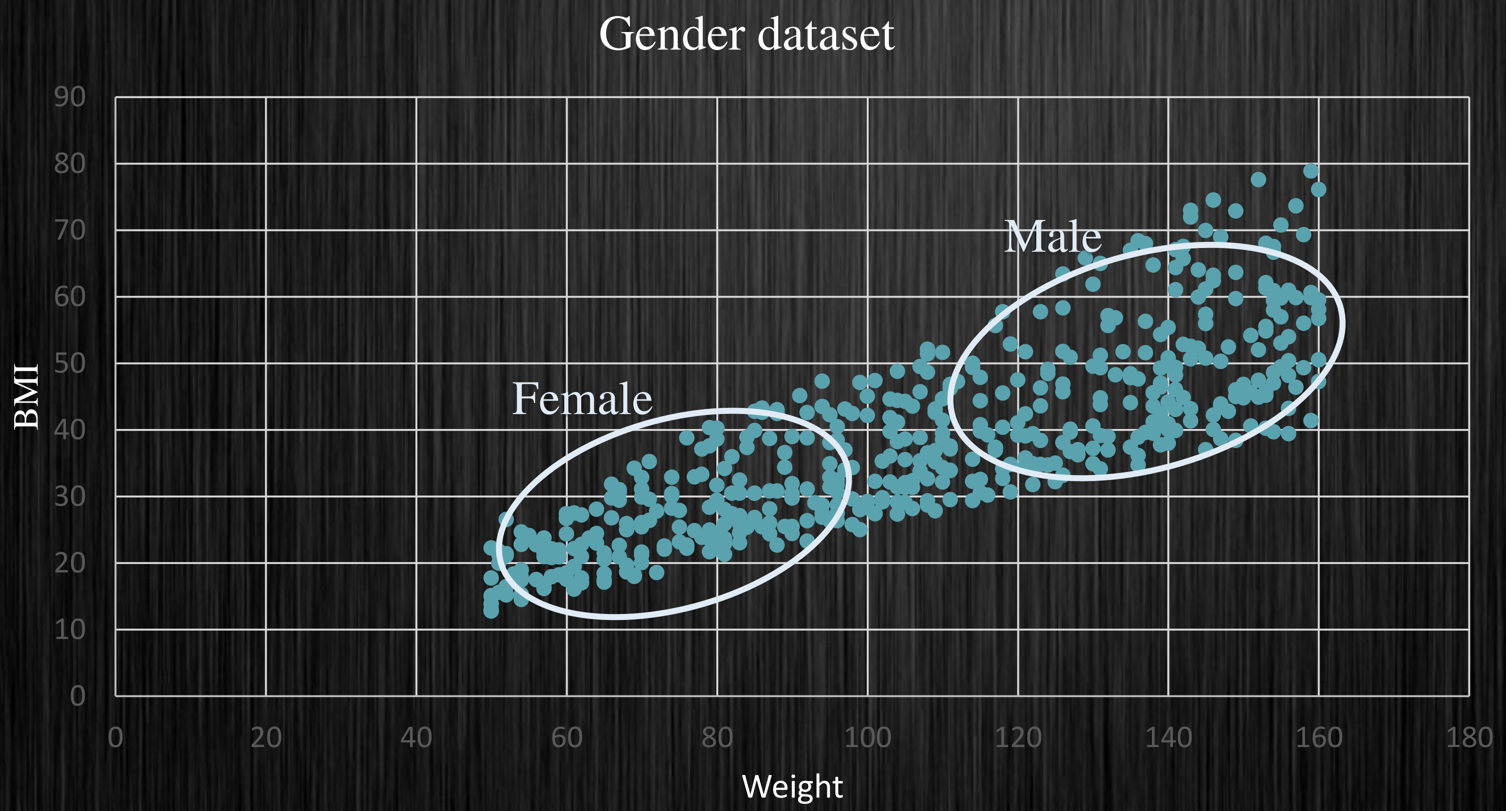
Gaussian Mixture Model (GMM)

Model Based Clustering aims to find:

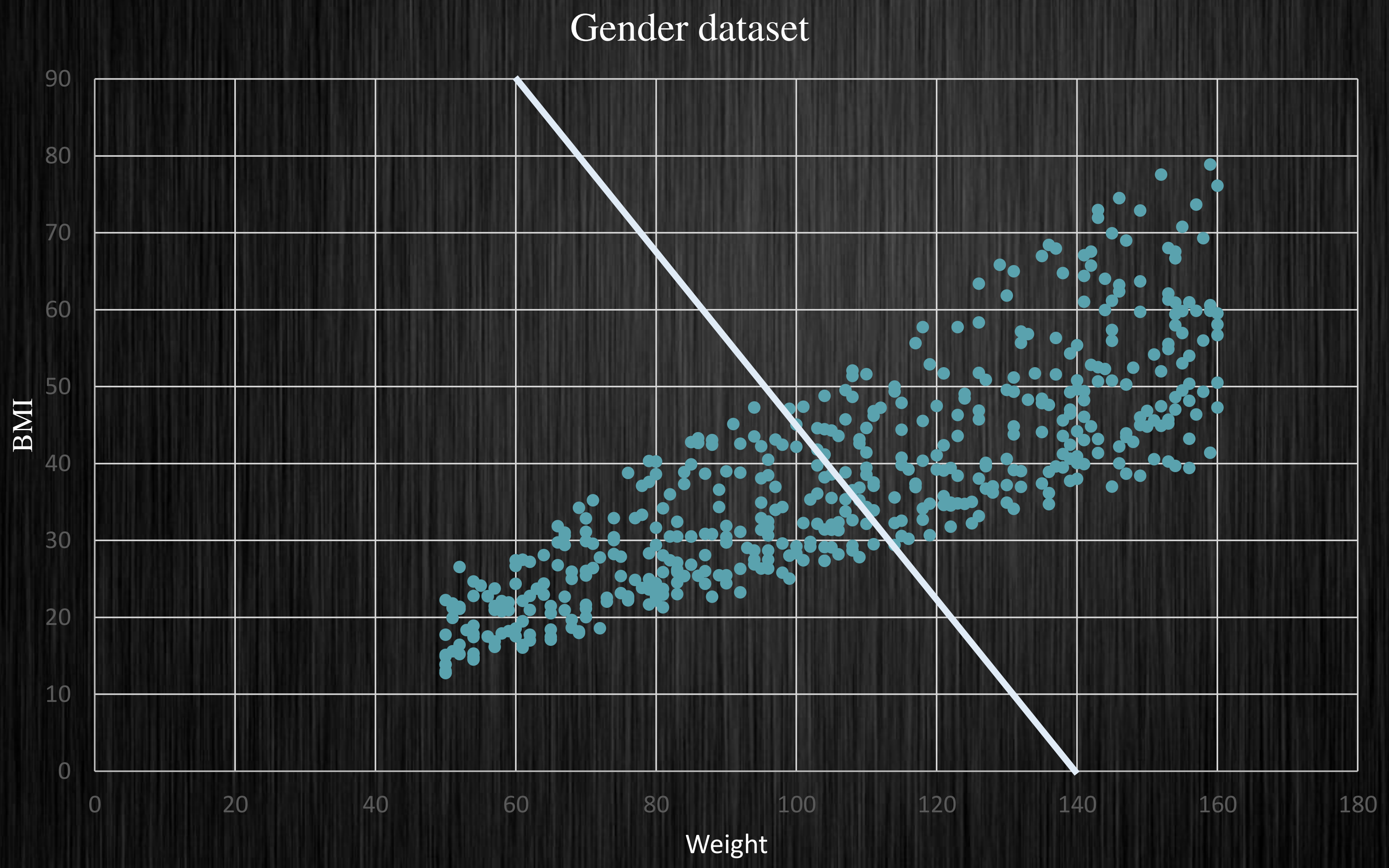
1) **The number of gaussians**
2) **Their locations**
3) **Their (CO)variance(“width”)**

Gender Dataset: Can you guess the gender?

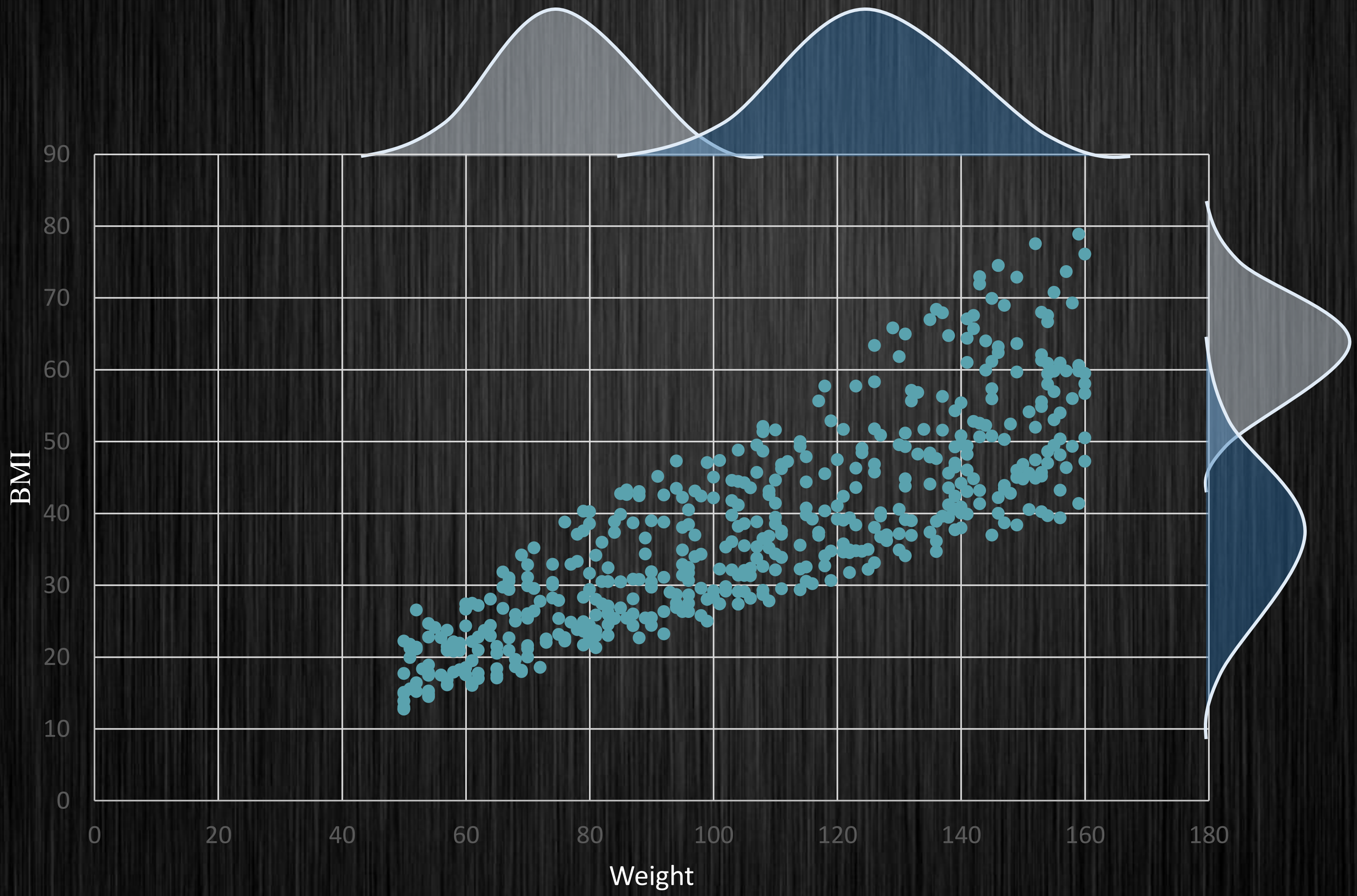
```
library(ggplot2)
ggplot(gender, aes(x = weight, y = BMI)) + geom_points()
```



Under traditional cluster approaches



Model – based Clustering


BMI
Weight