Random Forest and Aggregation

Week 7 - Day 1

Recap on Decision Trees

Decision Tree

=

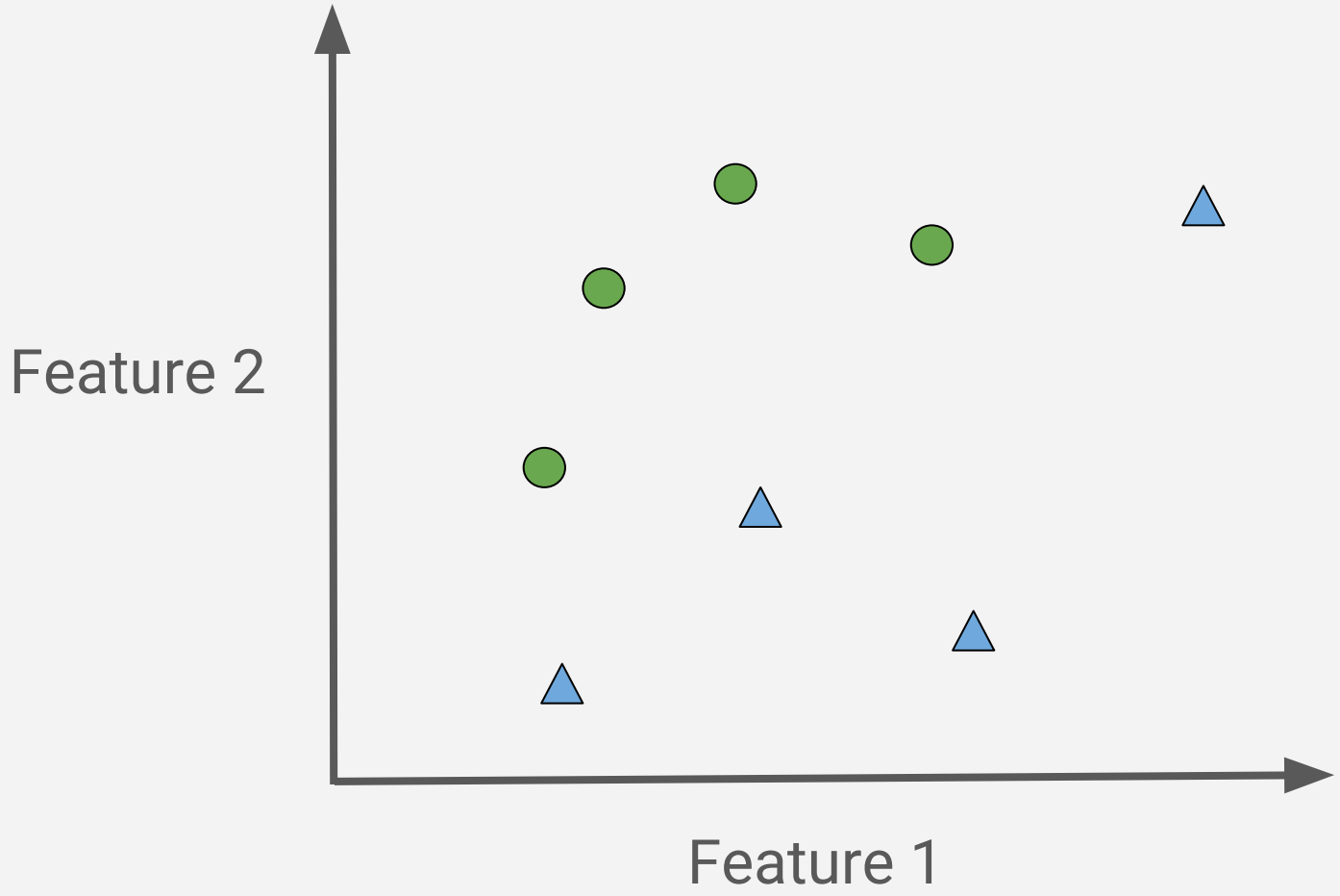
Rules builder

If (weather==sunny) and (temperature<28):

```
If (weather==sunny) and (temperature<28):
        play football
```

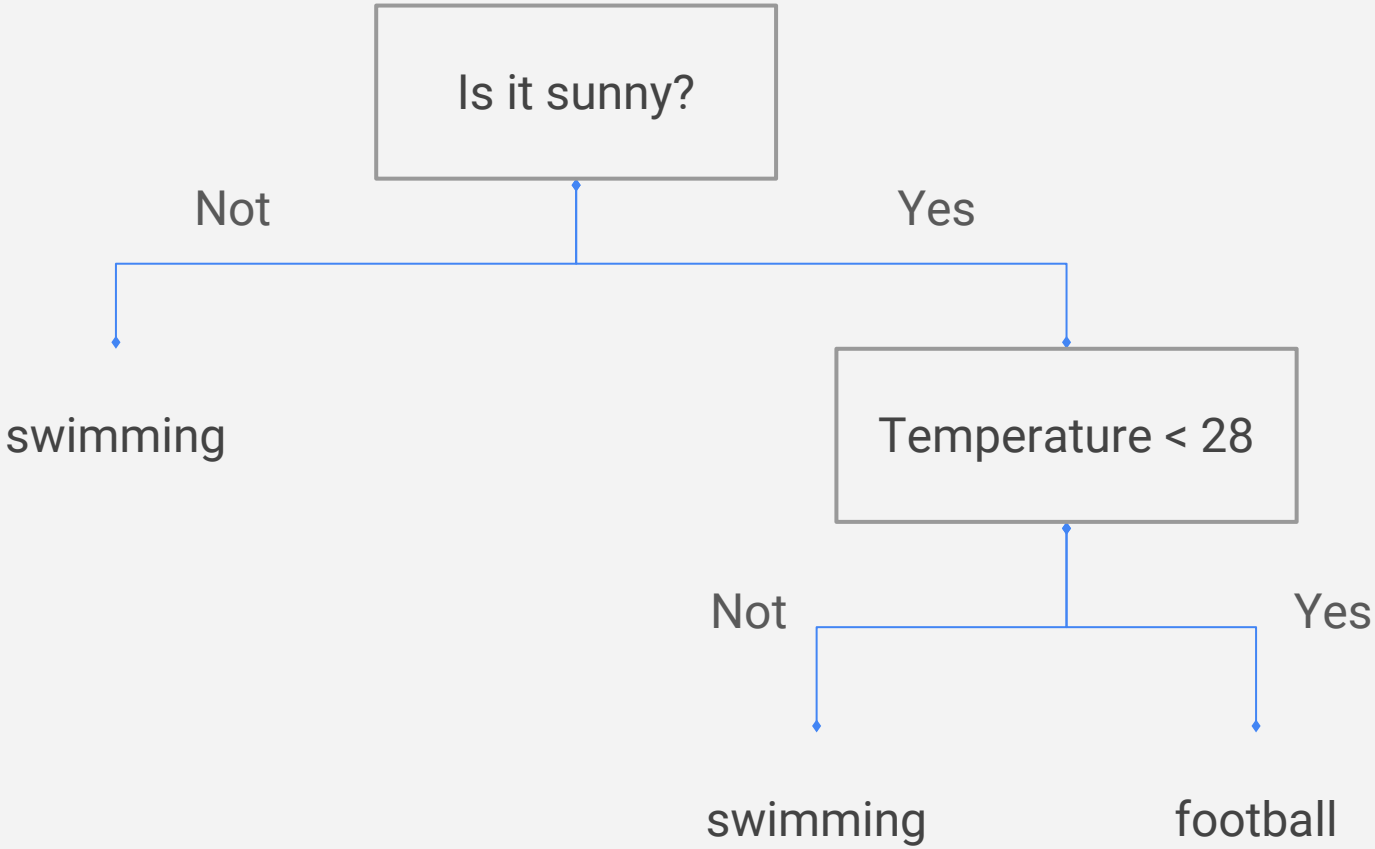
```
If (weather==sunny) and (temperature<28):
        play football
else
        go swimming
```

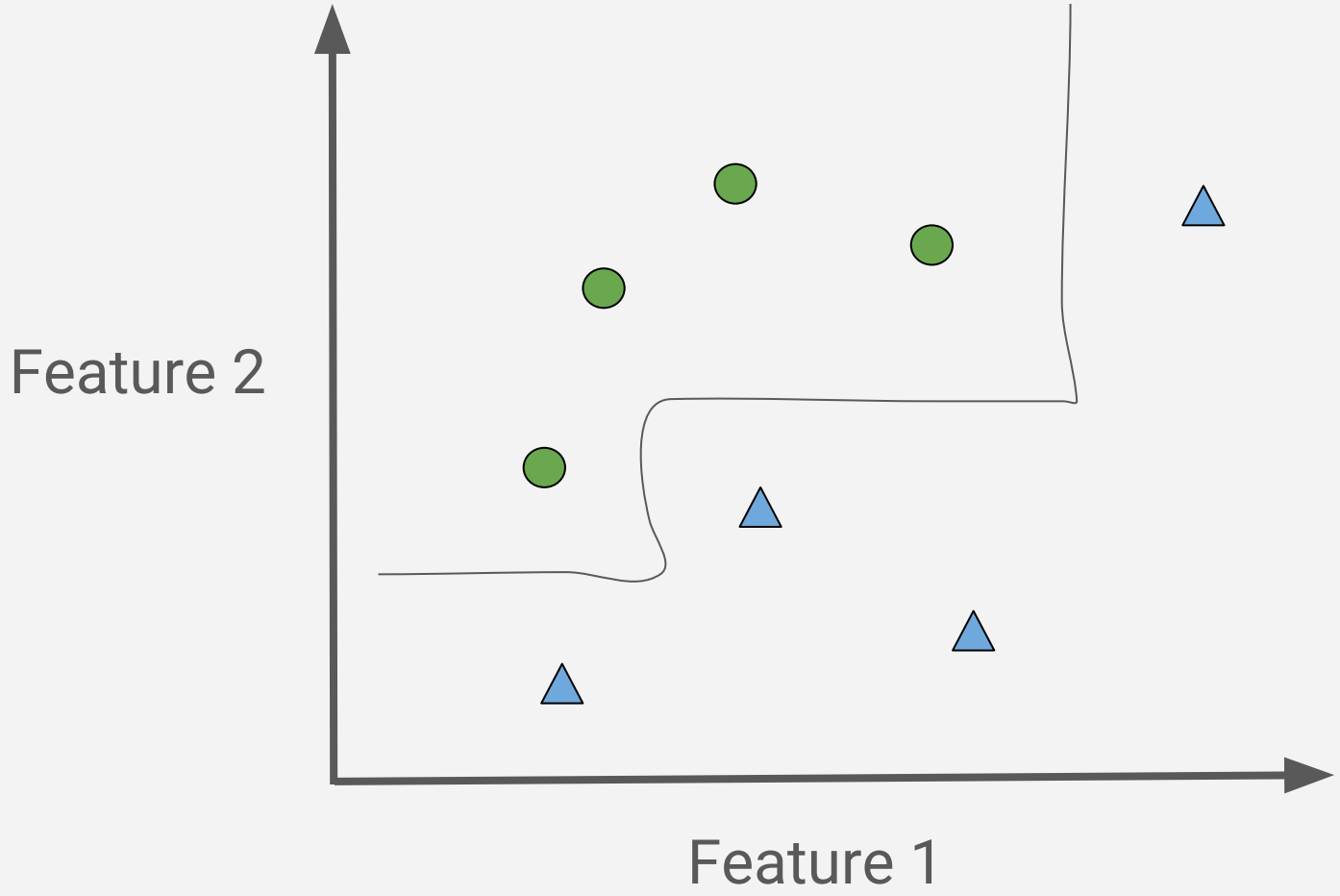
Quiz

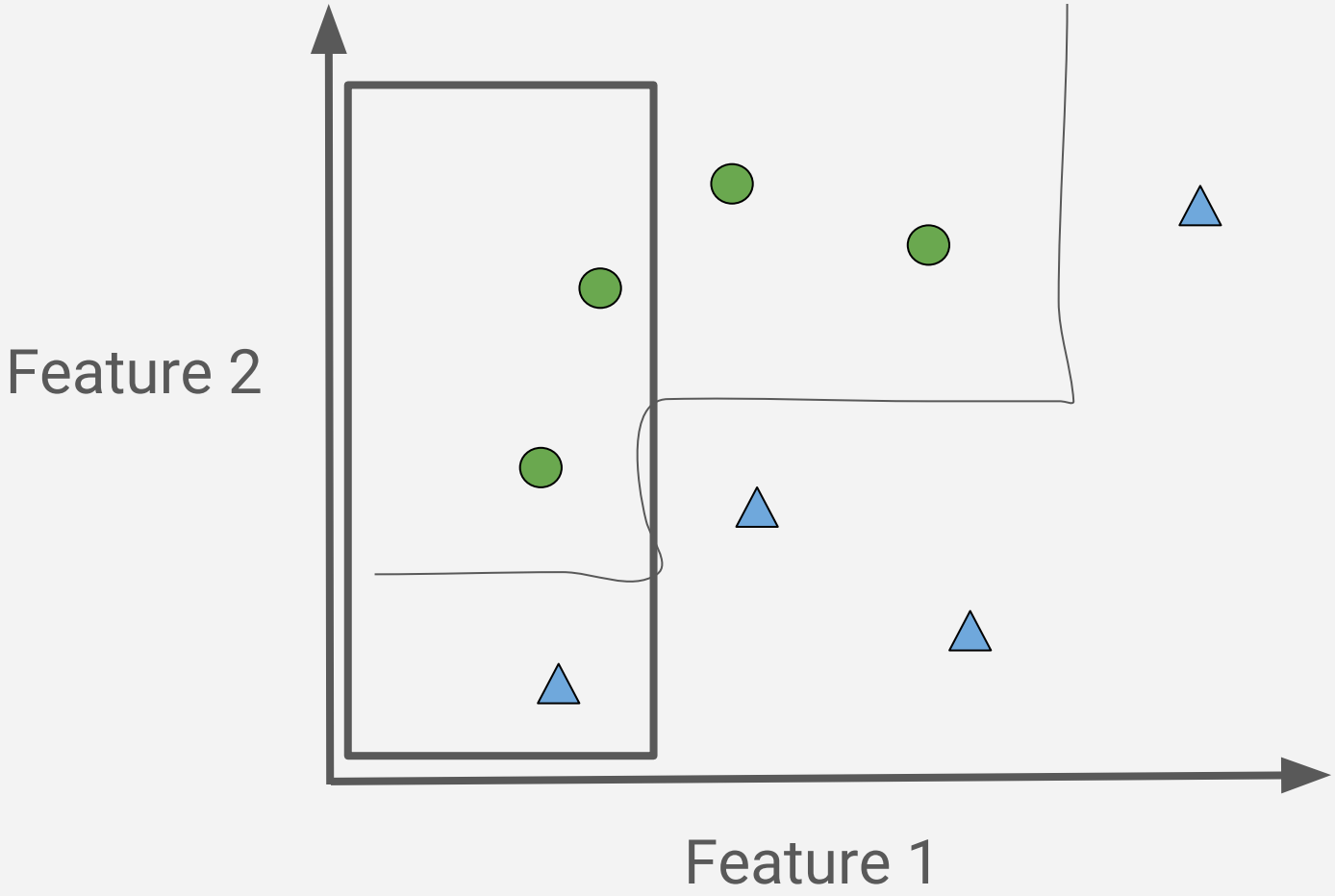

Feature 2
Feature 1

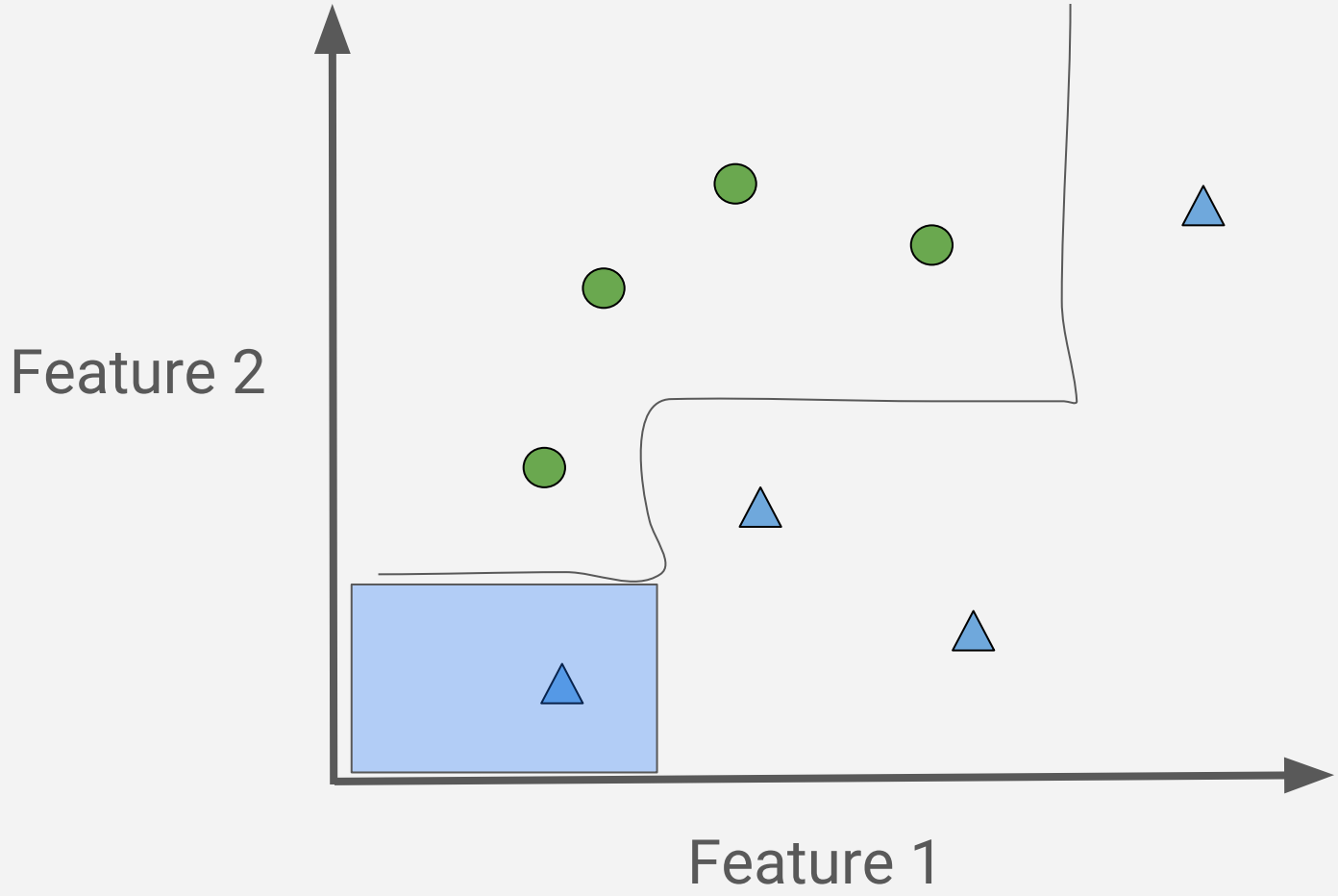

Feature 2
Feature 1

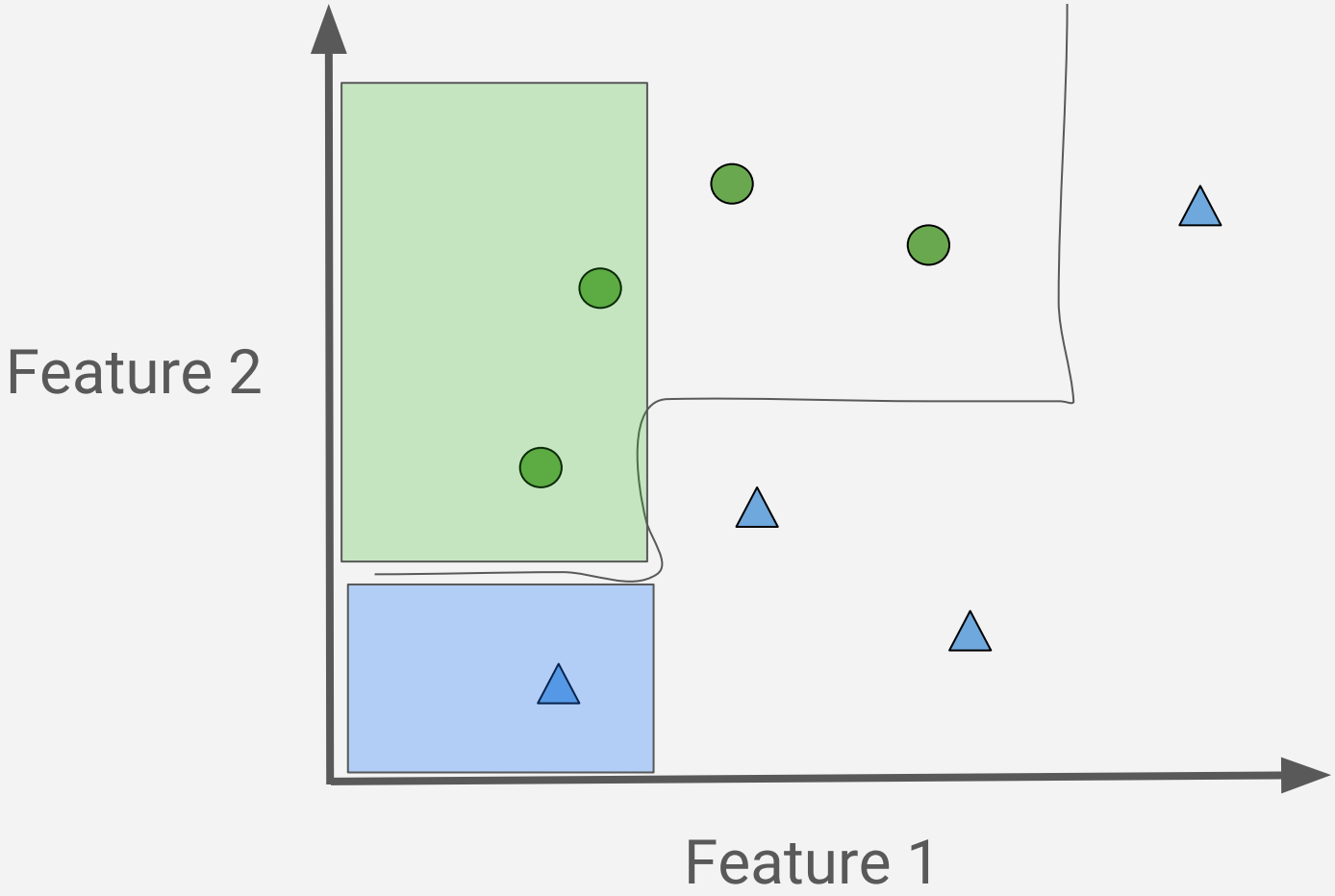
Which one is from a CART?

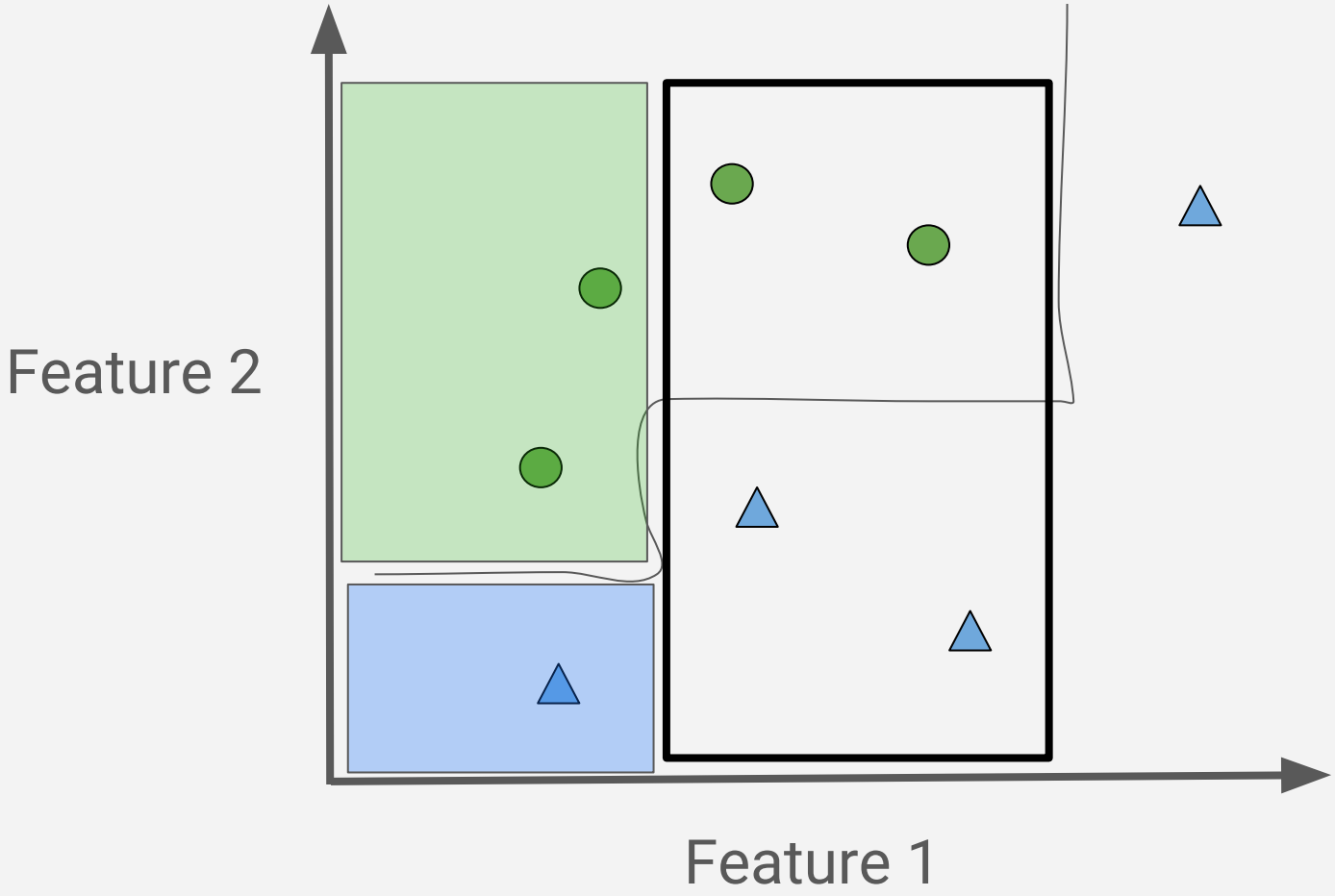

Is it sunny?
Not
Yes
swimming
Temperature < 28
Not
Yes
swimming
football

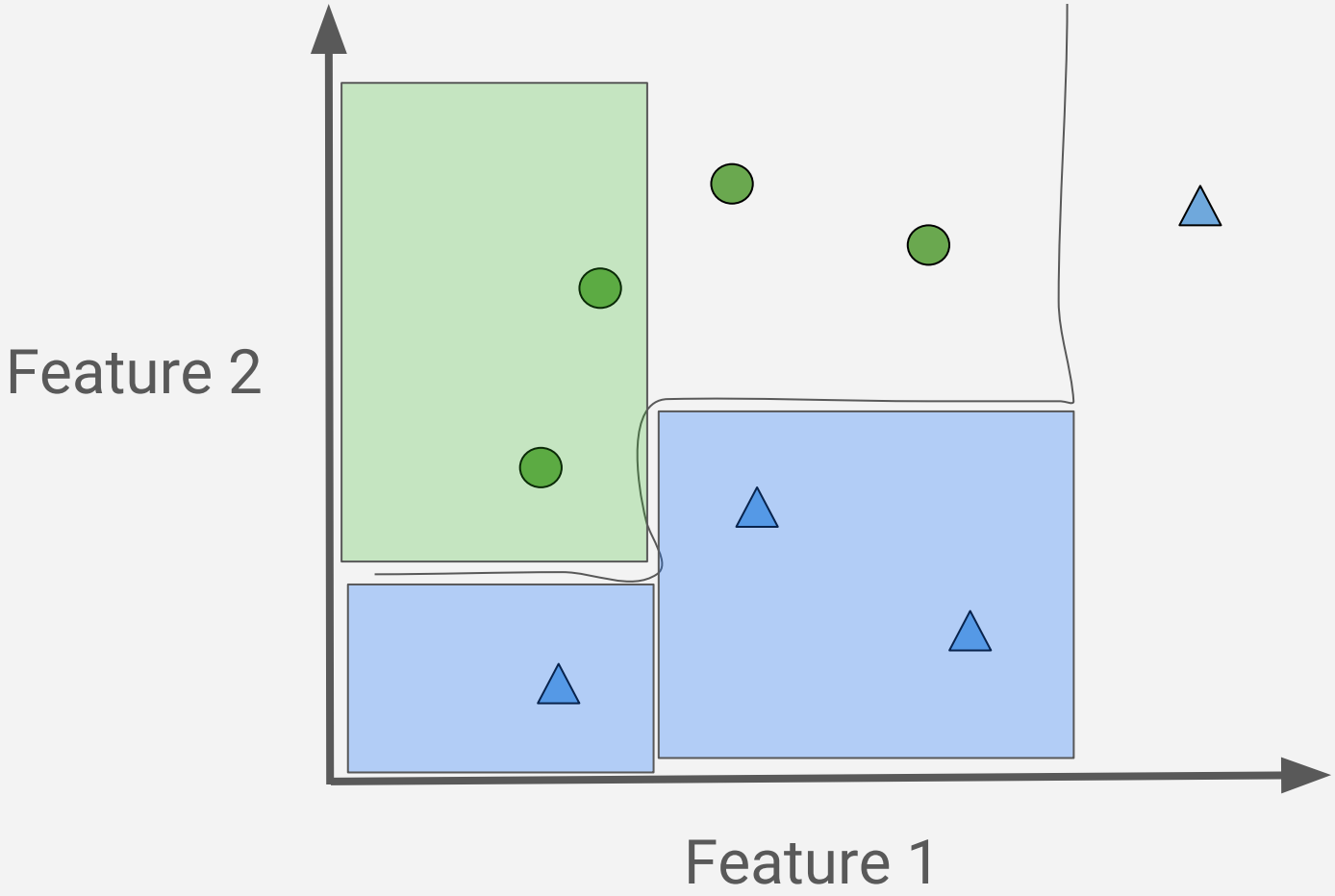

Feature 2
Feature 1


Feature 2
Feature 1


Feature 2
Feature 1


Feature 2
Feature 1


Feature 2
Feature 1


Feature 2
Feature 1

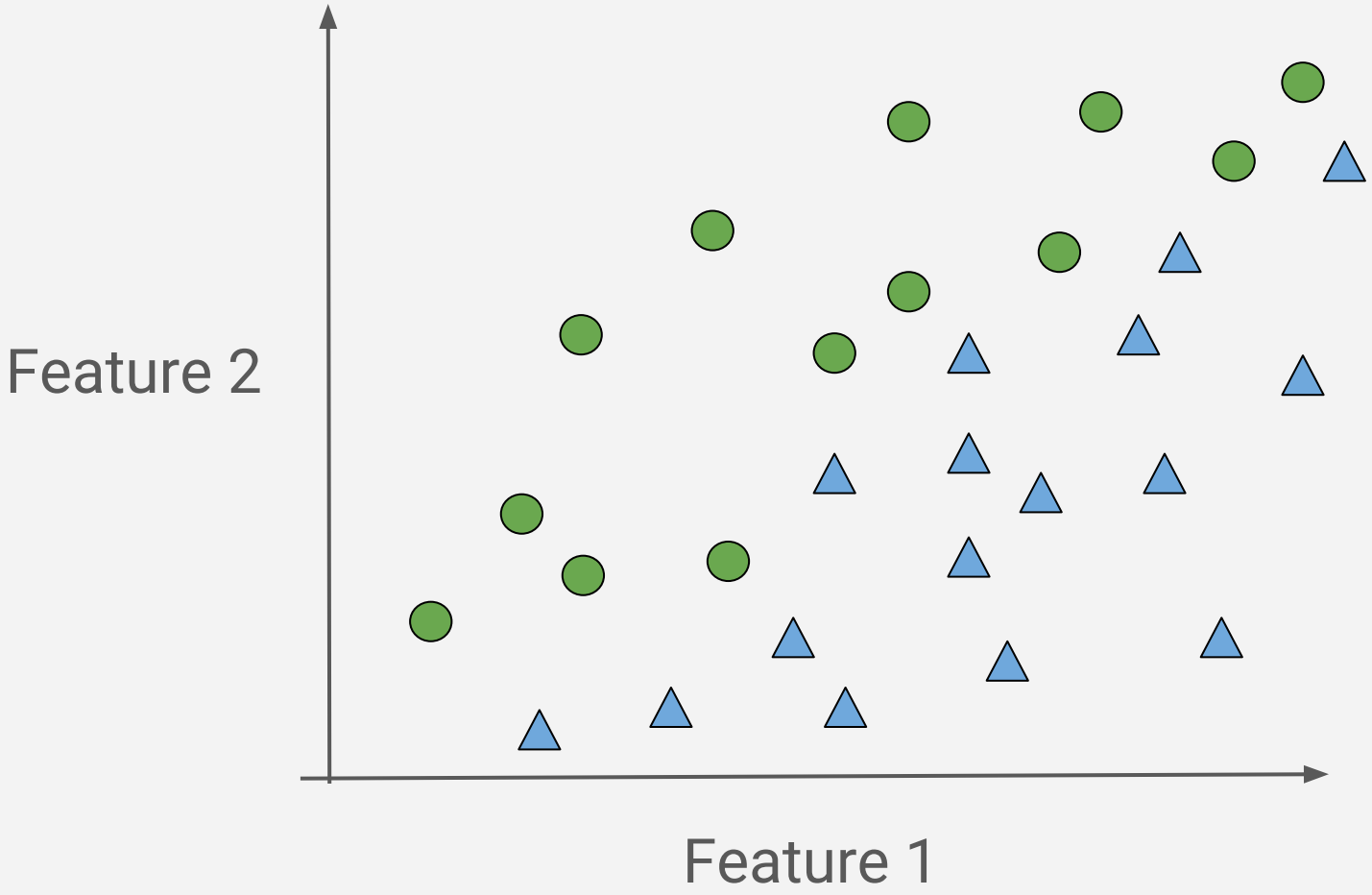

Feature 2
Feature 1

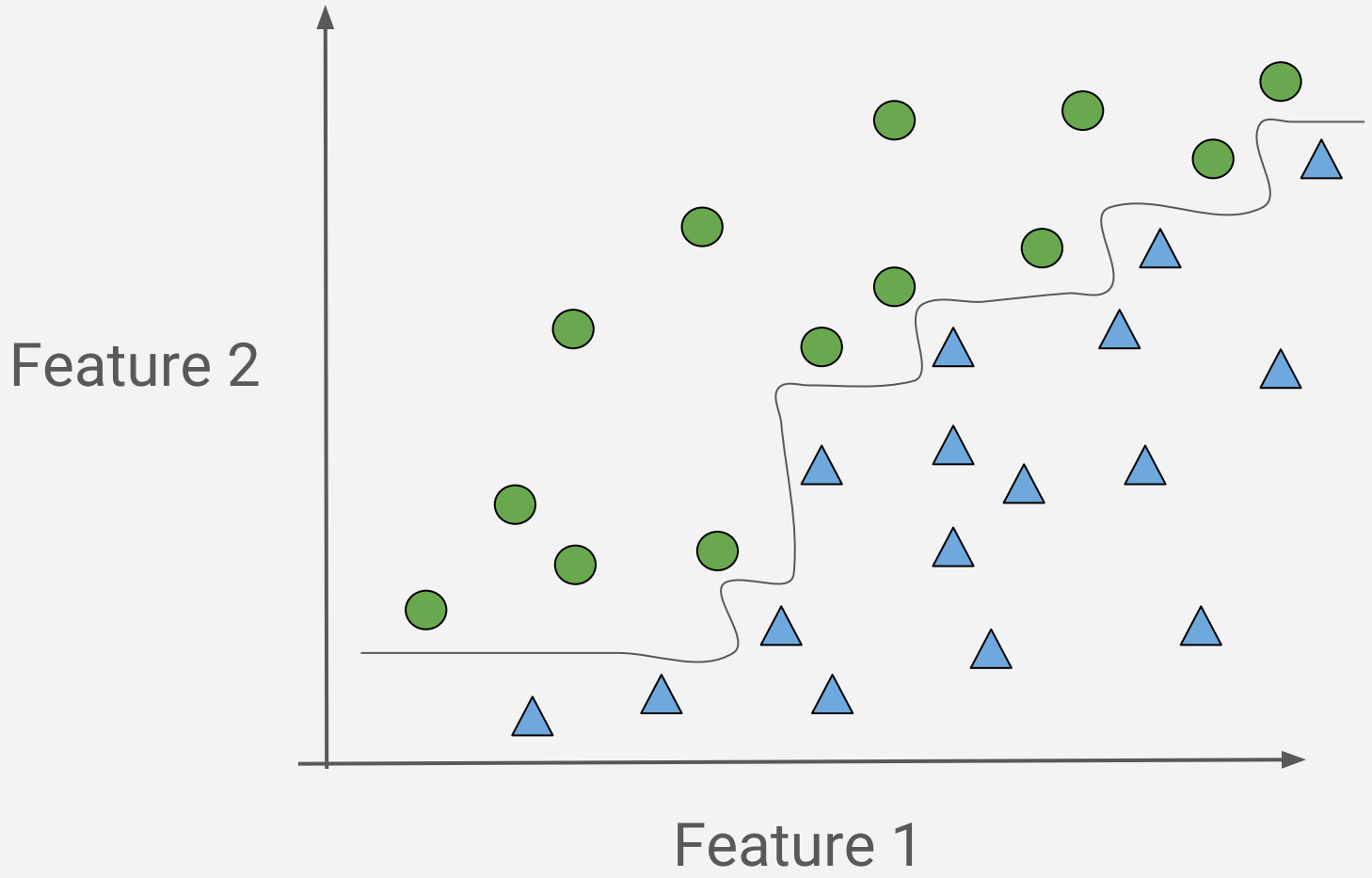

Feature 2
Feature 1

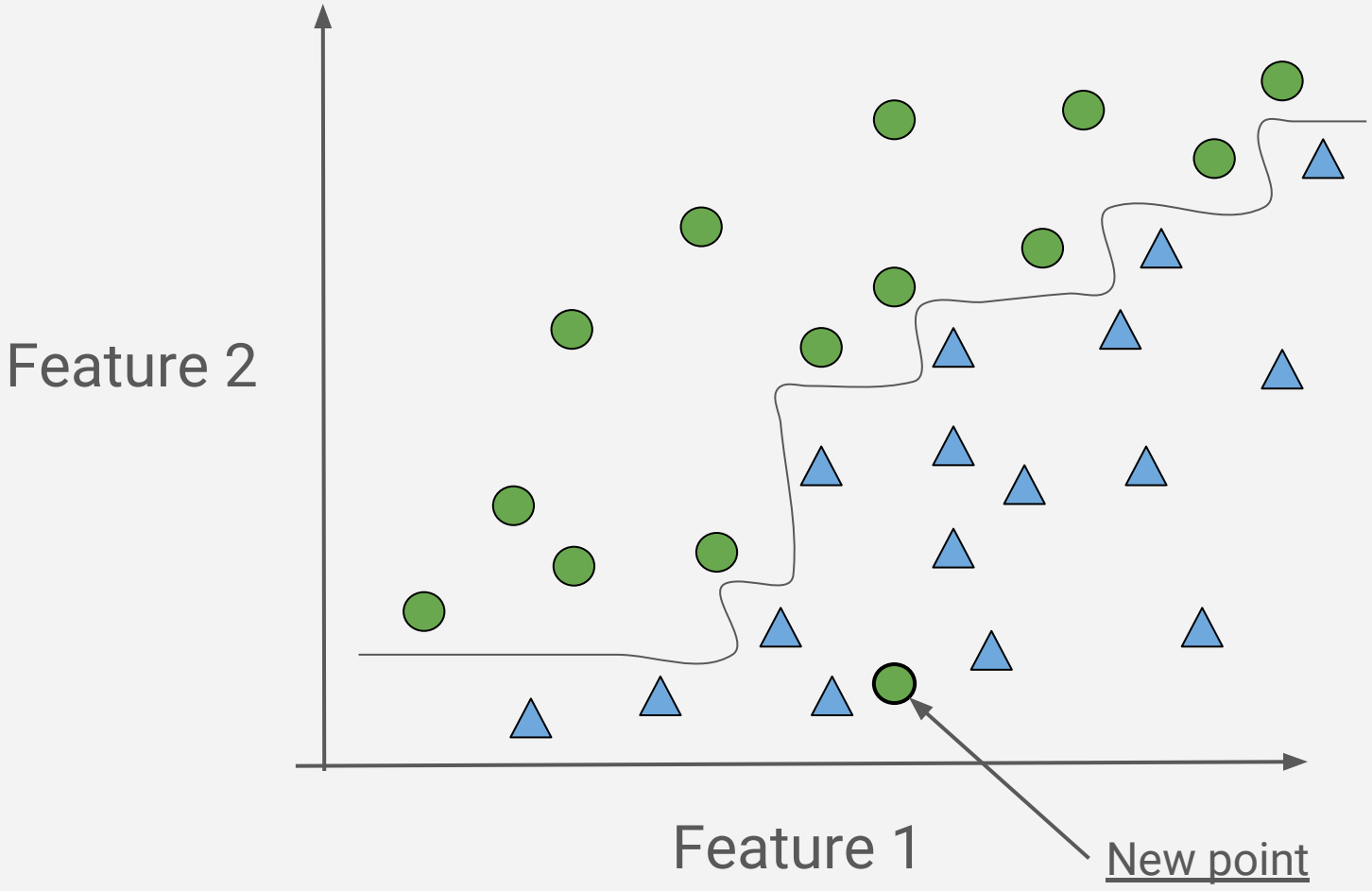
Experiment time!

Guess my height

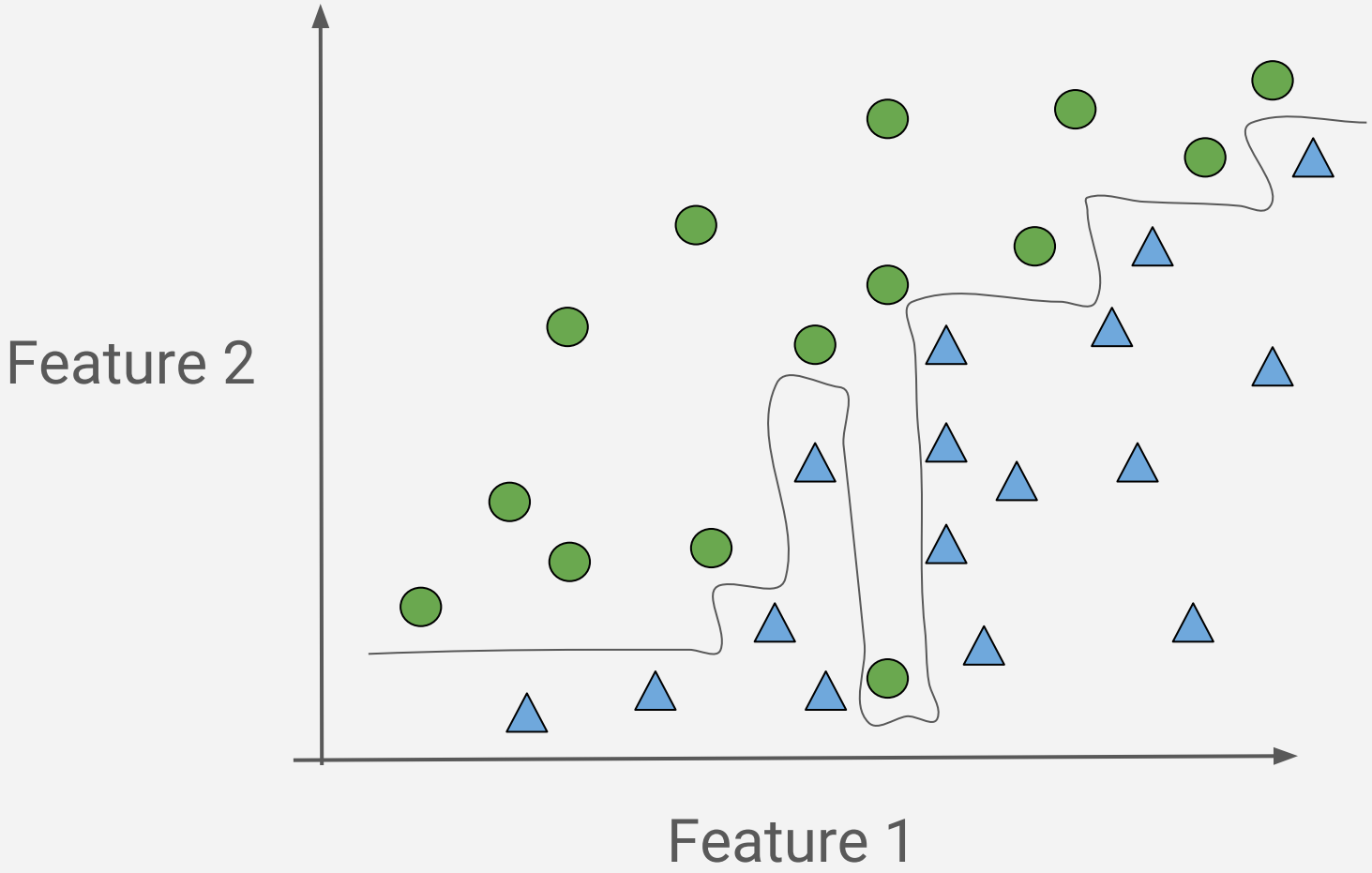
Decision Trees Limits


Feature 2
Feature 1


Feature 2
Feature 1


Feature 2
Feature 1
New point

What happens?


Feature 2
Feature 1

A lot of variance!

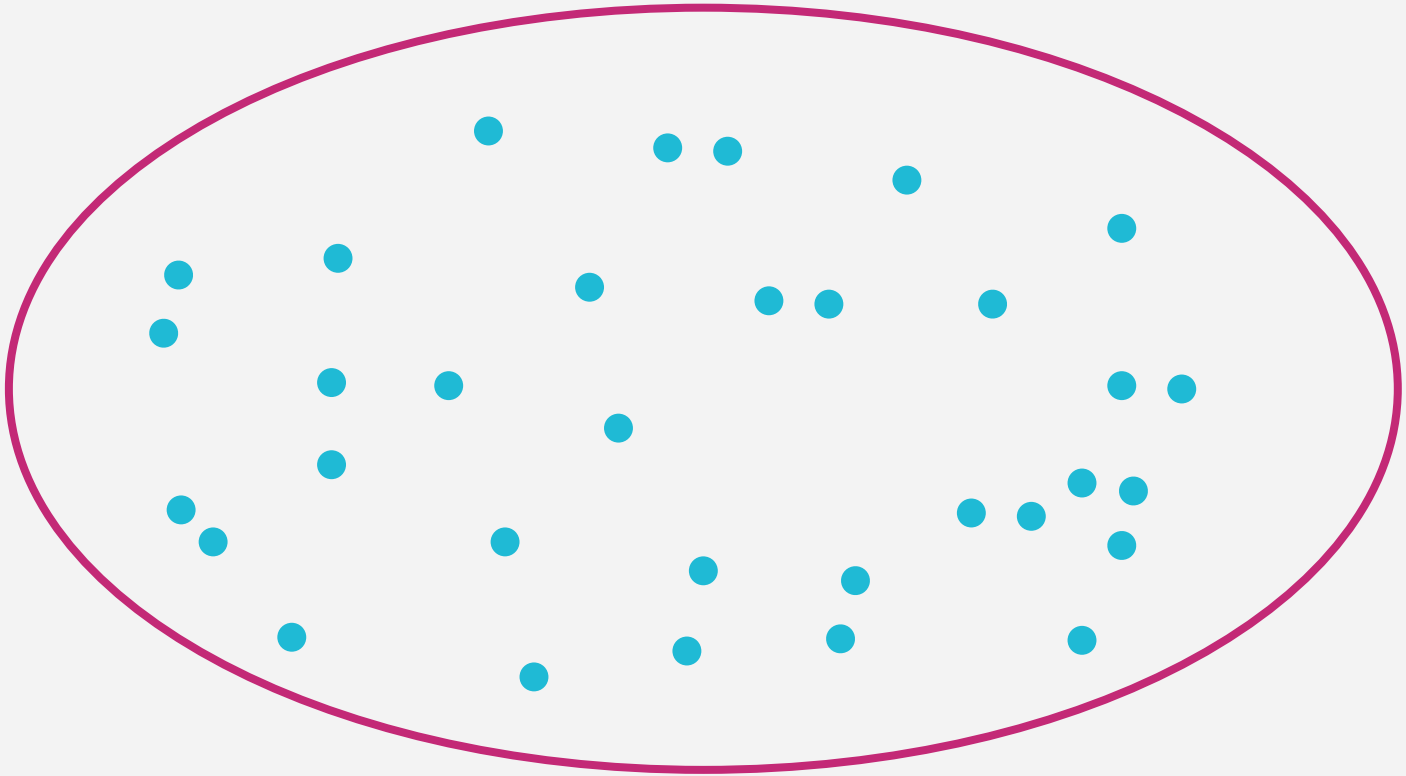
Noise can be a problem

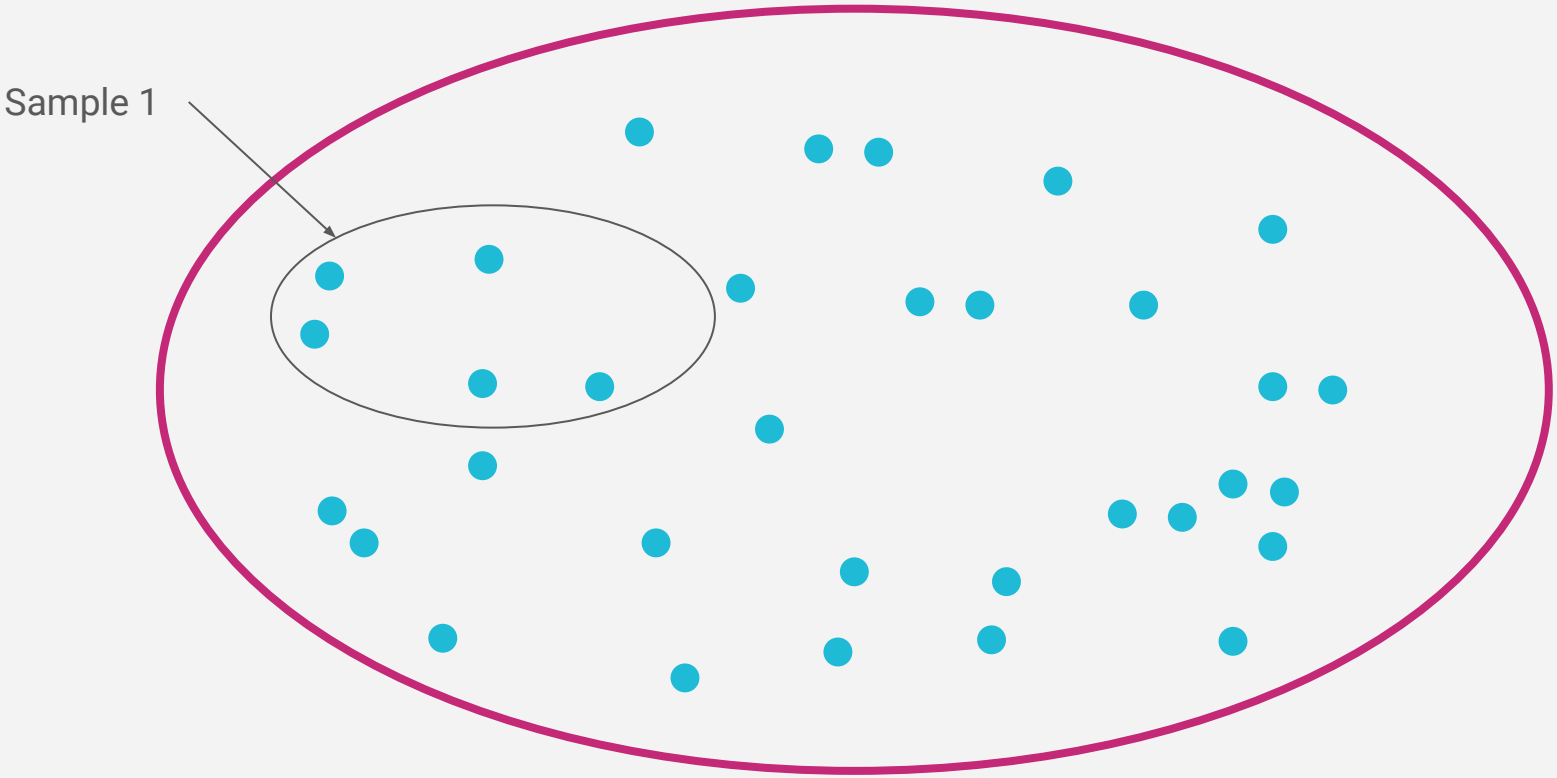
(Pruning can be used, but often it’s not enough)

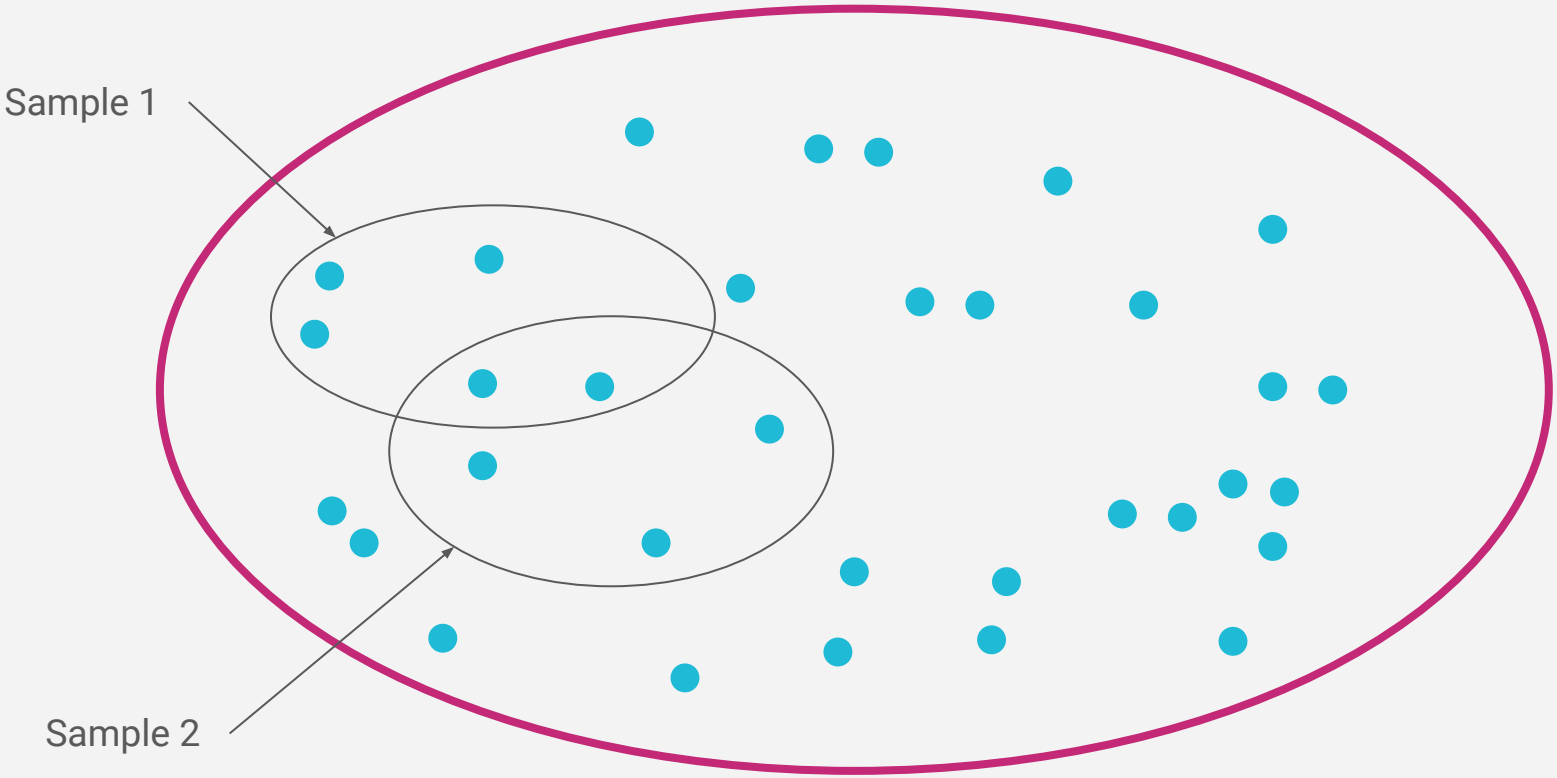
Random Forest

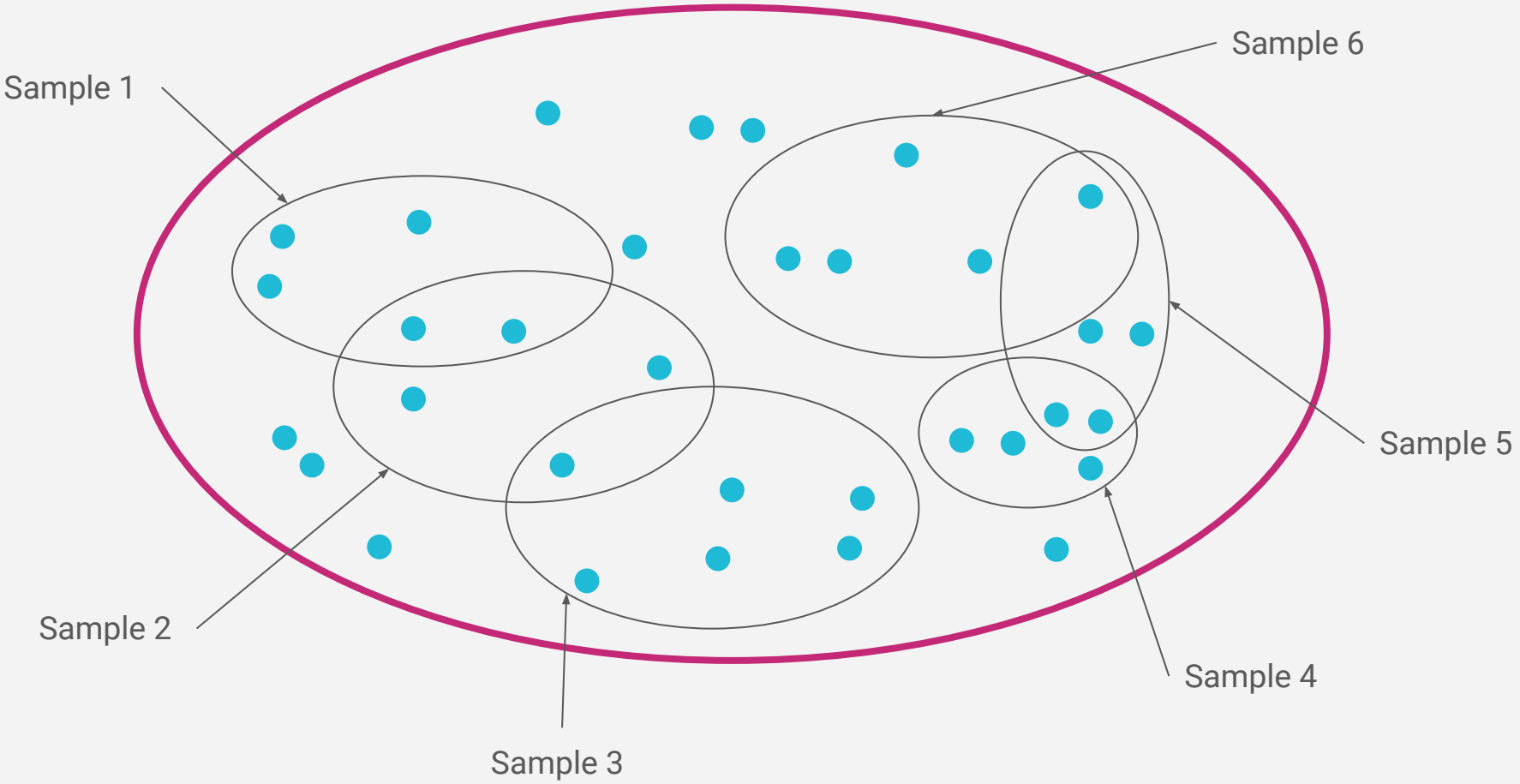
(Aggregation + Bootstrapping)

What’s bootstrapping?

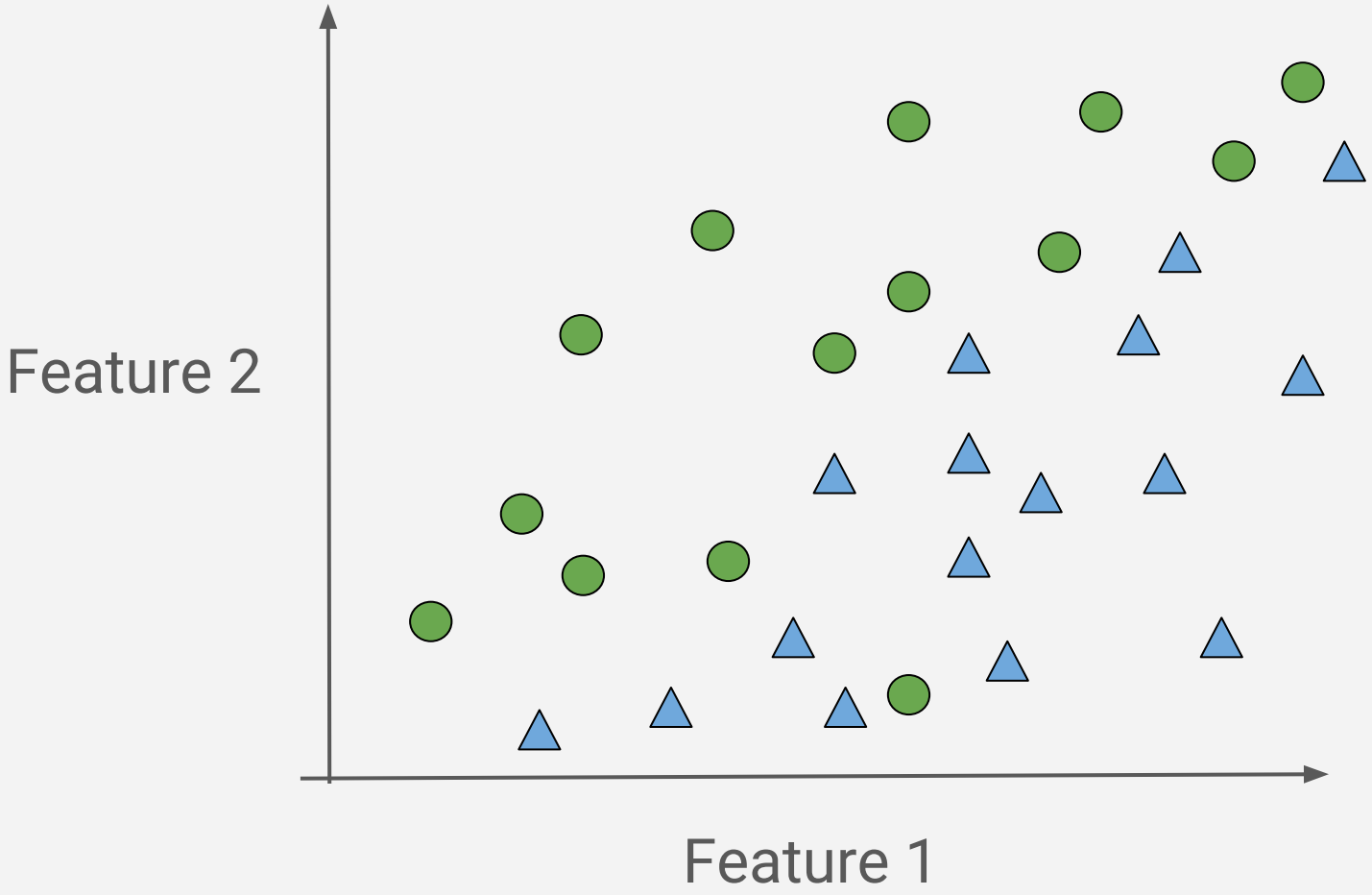


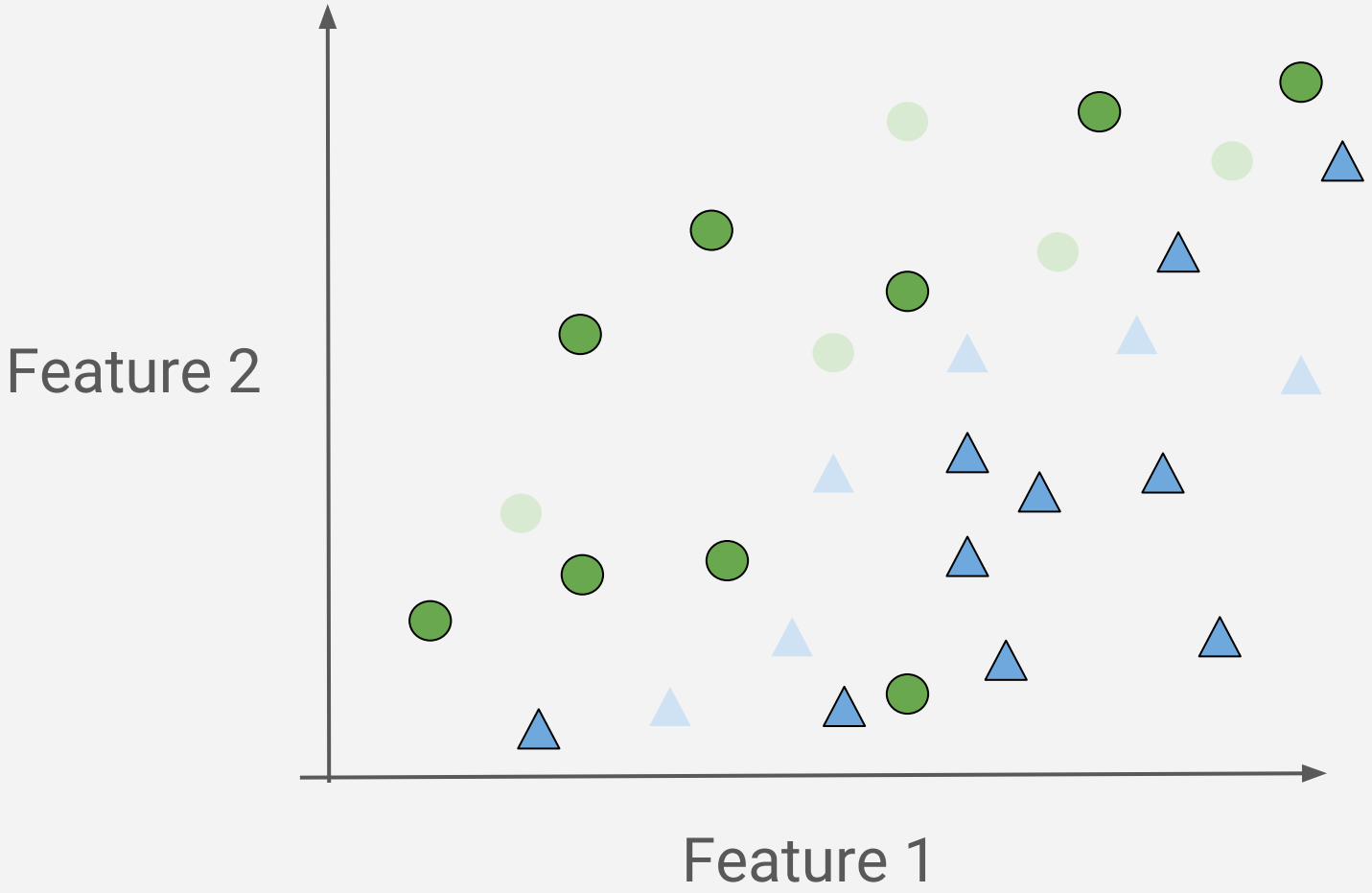

Sample 1

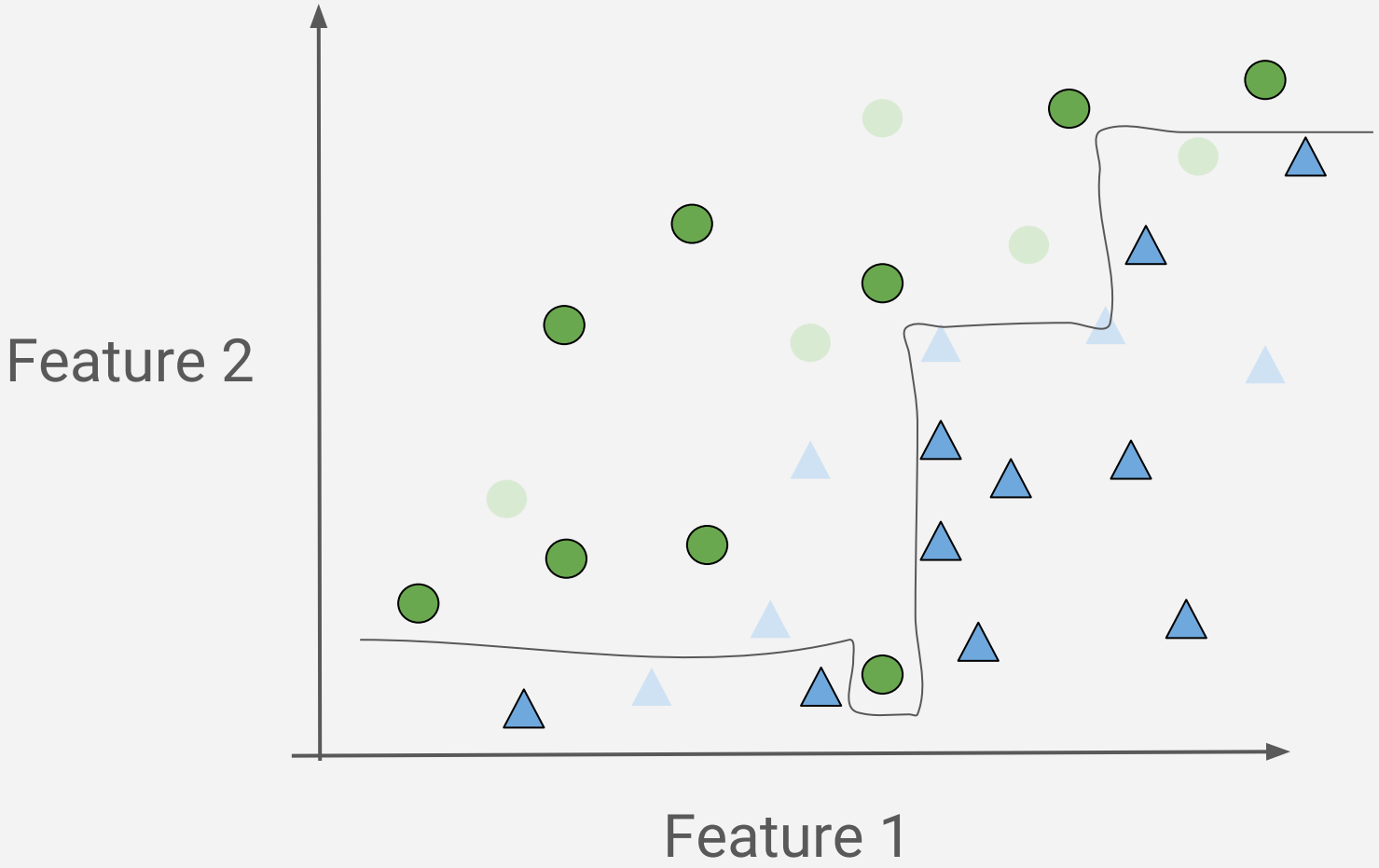

Sample 1
Sample 2

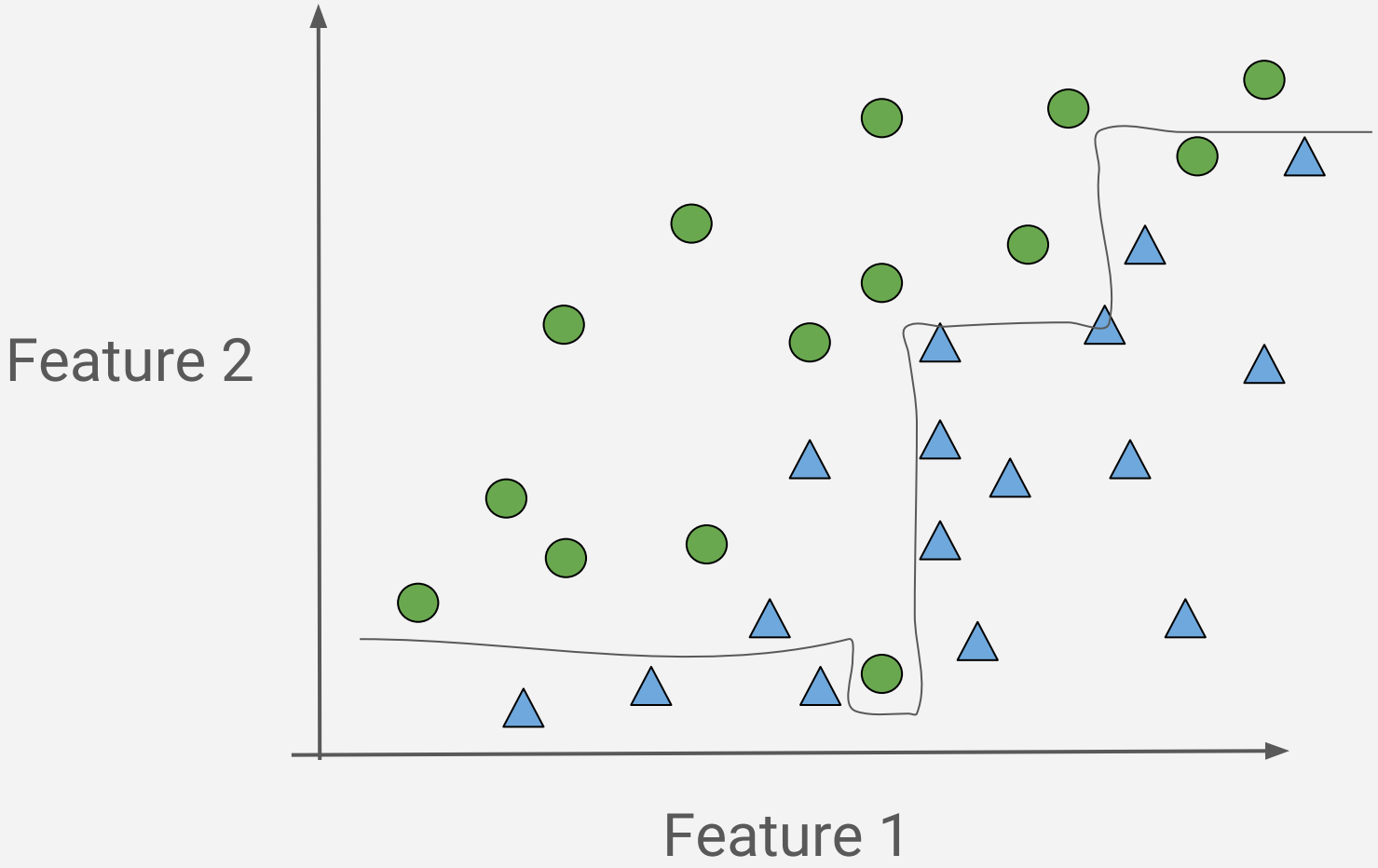

Sample 1
Sample 6
Sample 5
Sample 2
Sample 4
Sample 3

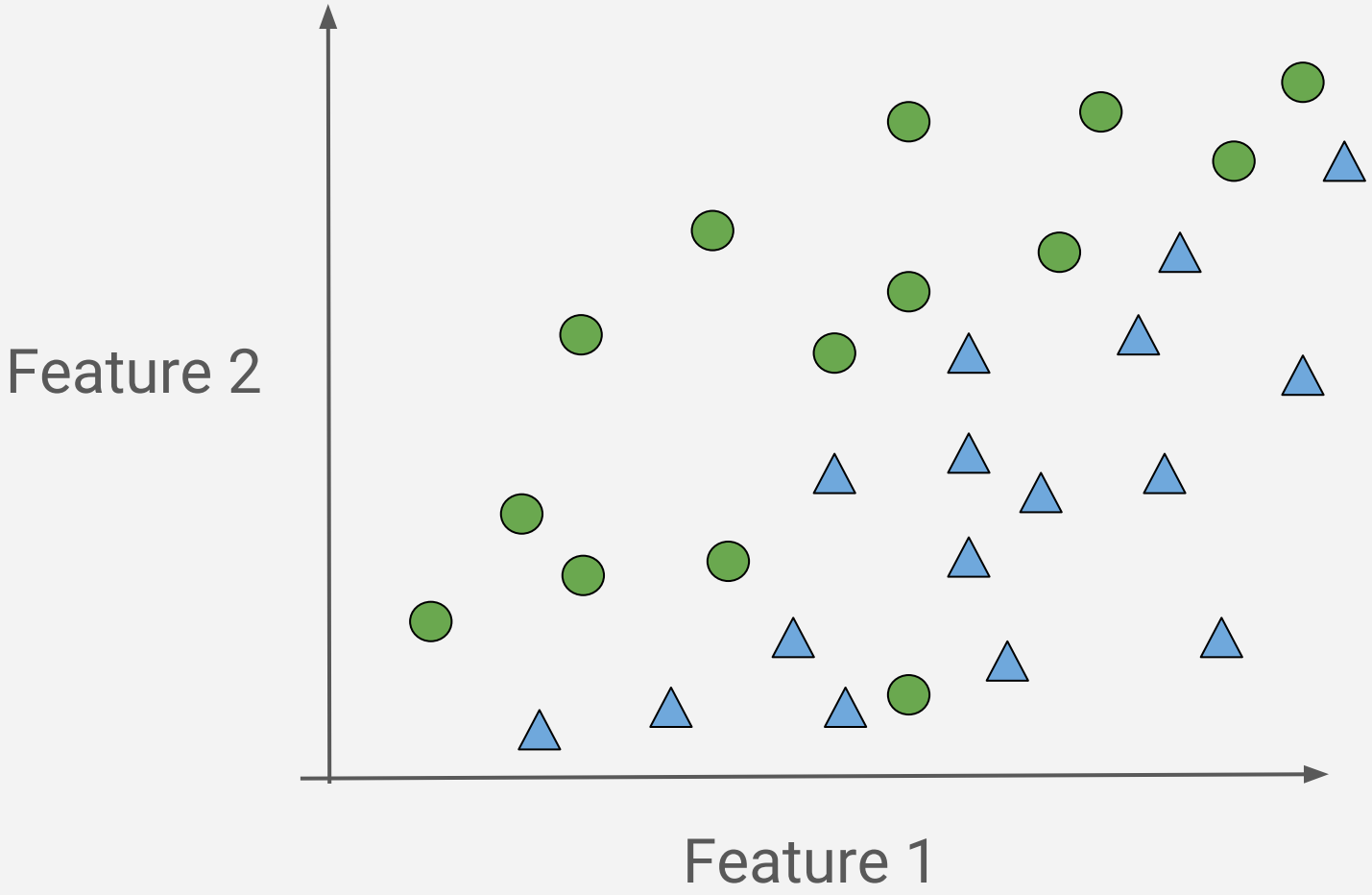
Bootstrapping
+
Decision Trees

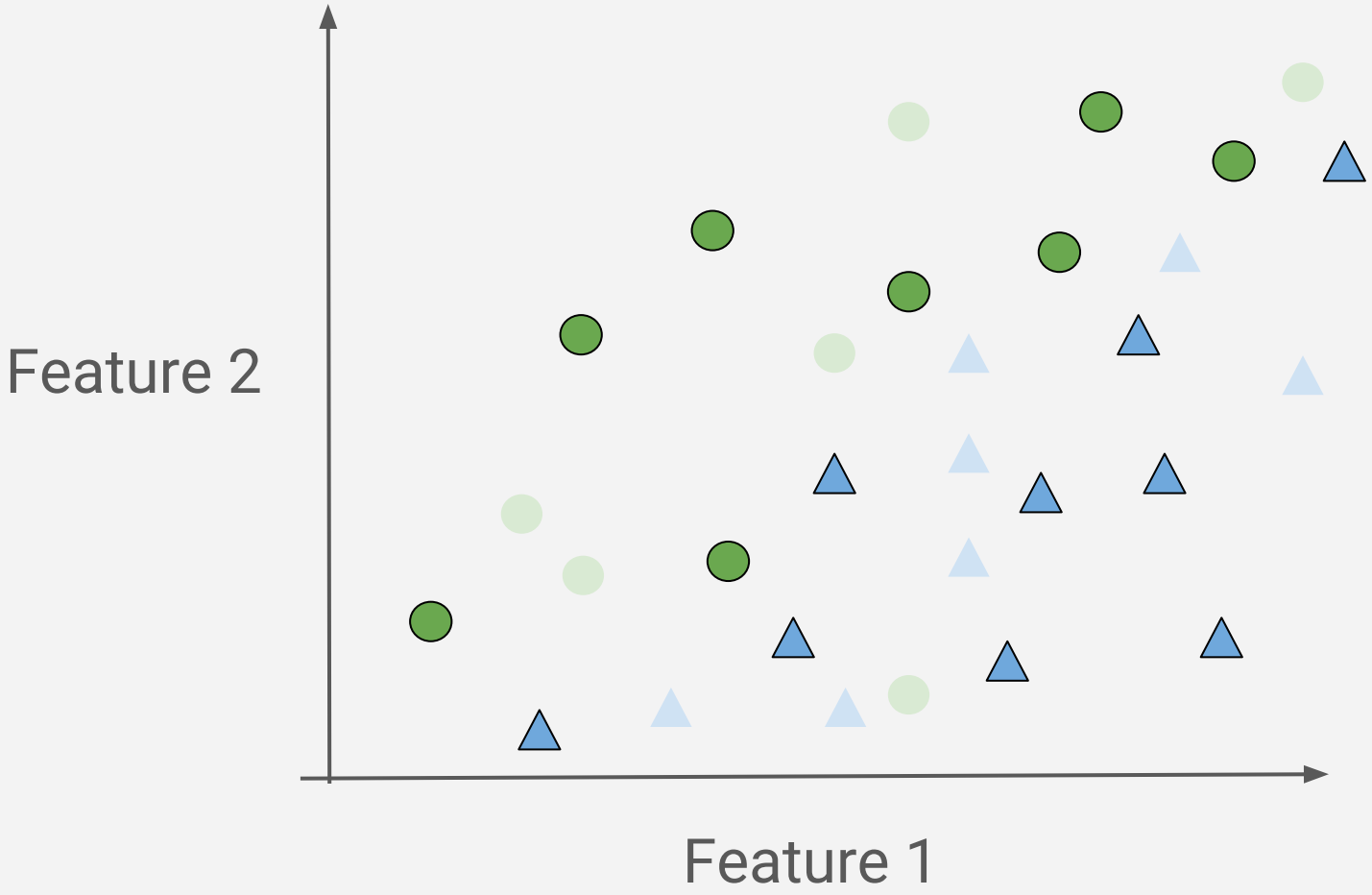

Feature 2
Feature 1

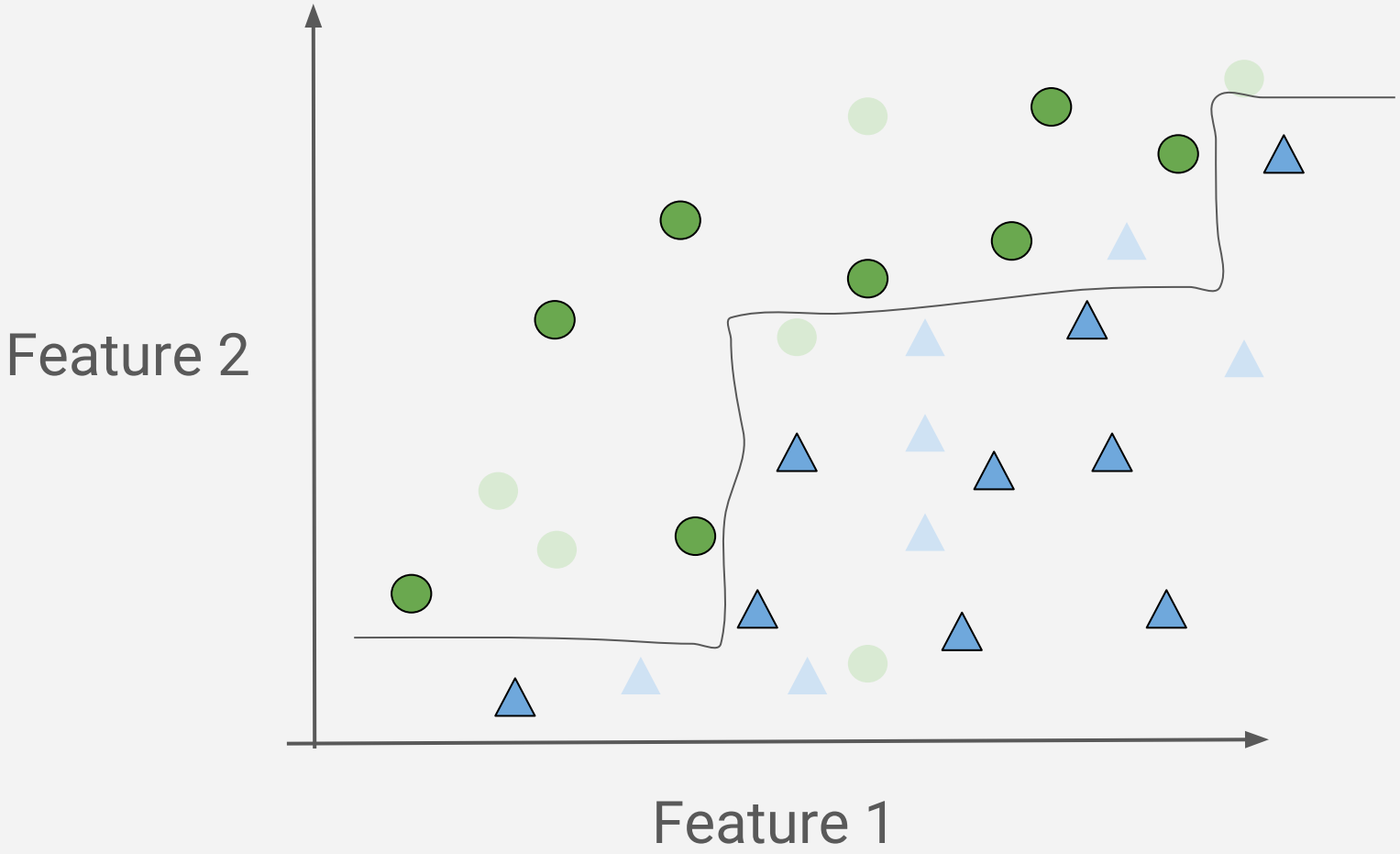

Feature 2
Feature 1

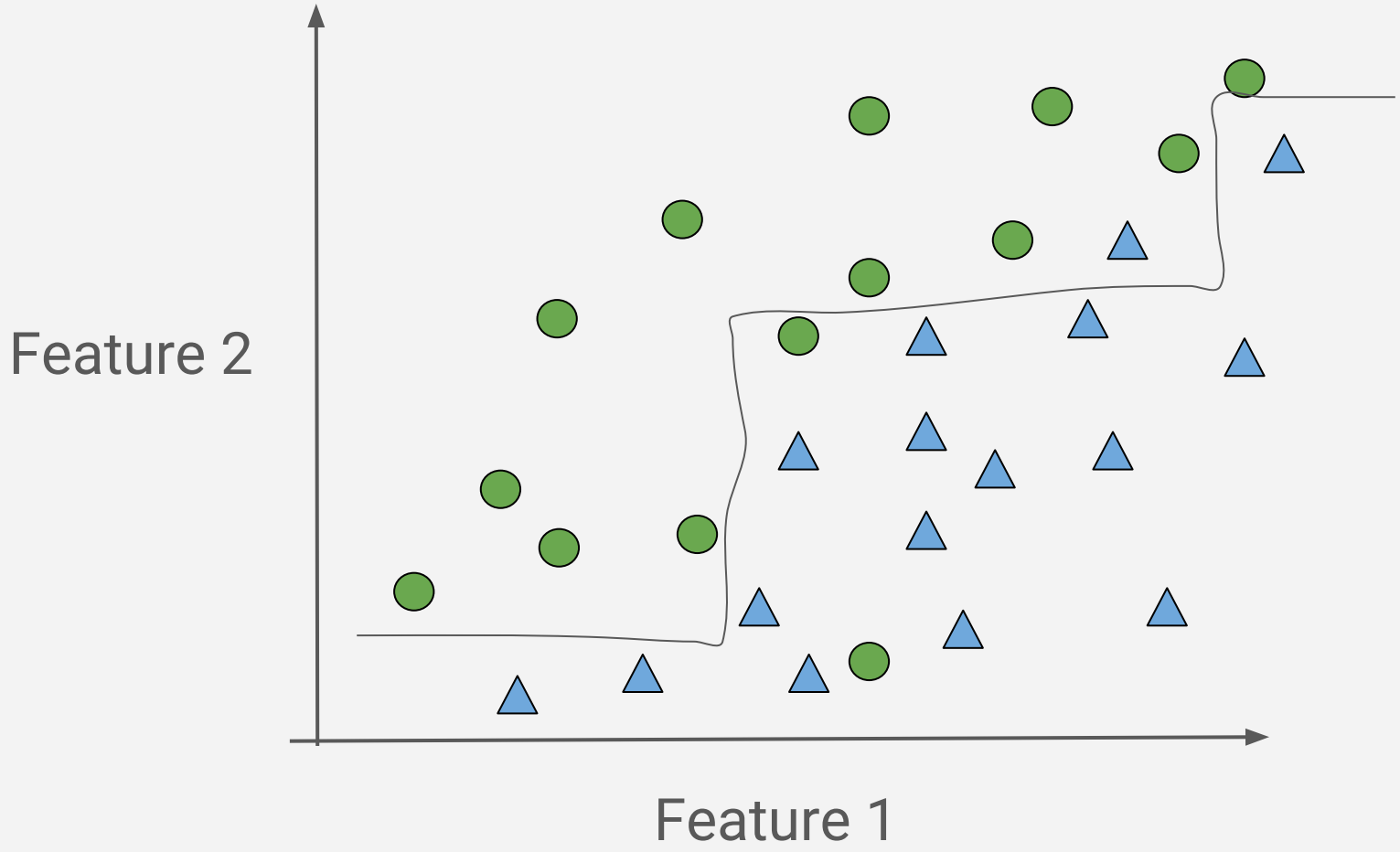

Feature 2
Feature 1

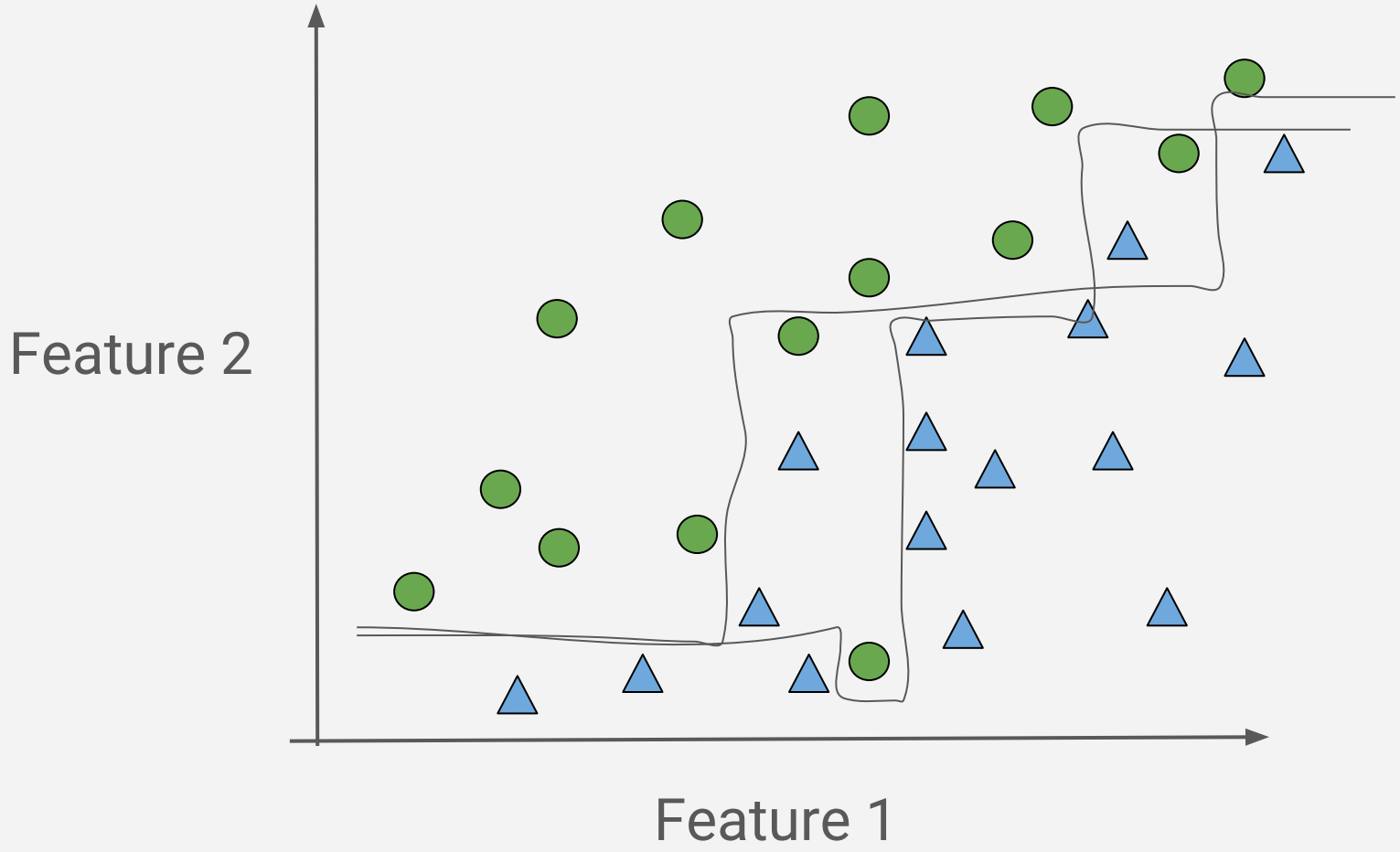

Feature 2
Feature 1

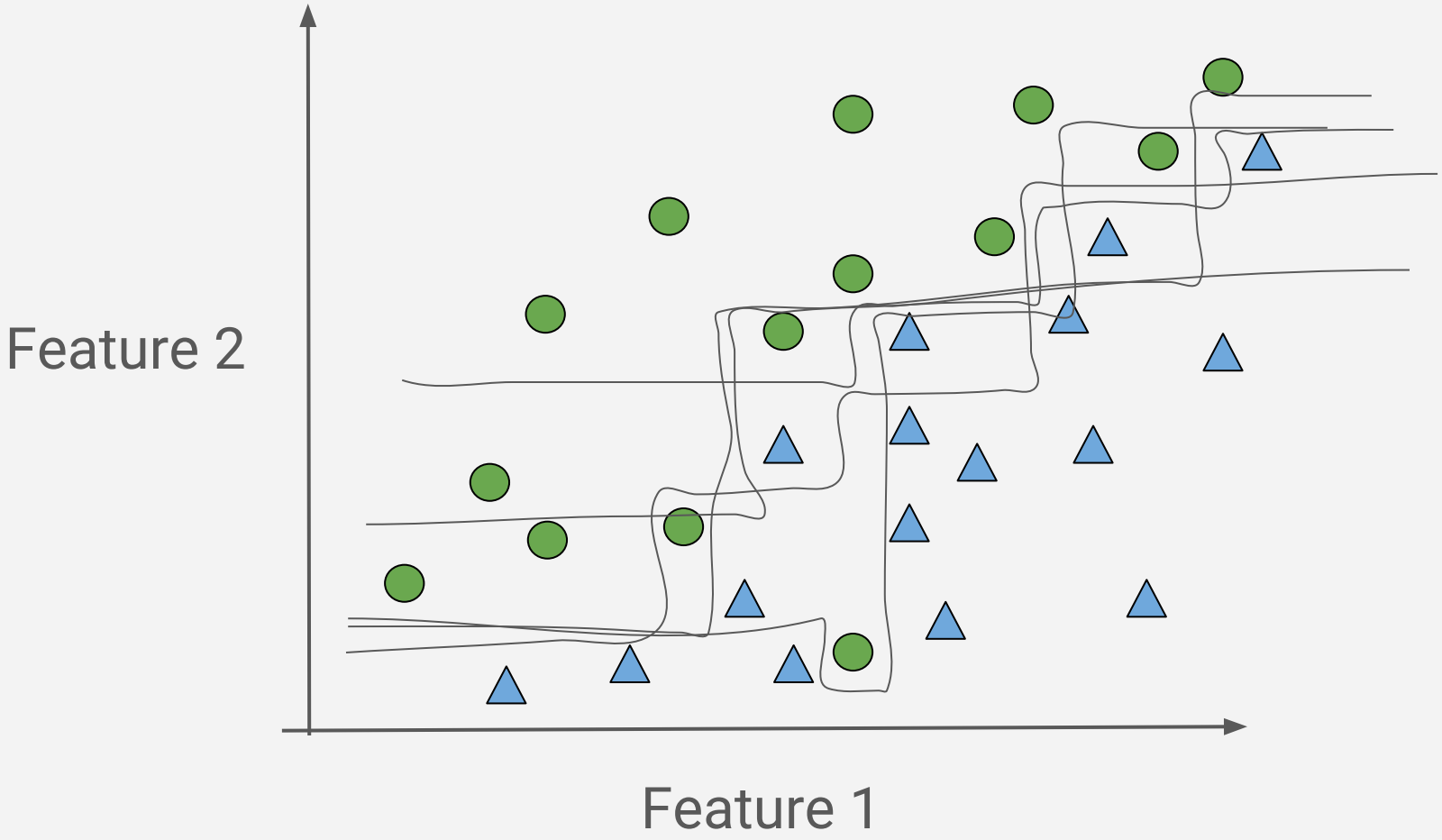

Feature 2
Feature 1


Feature 2
Feature 1

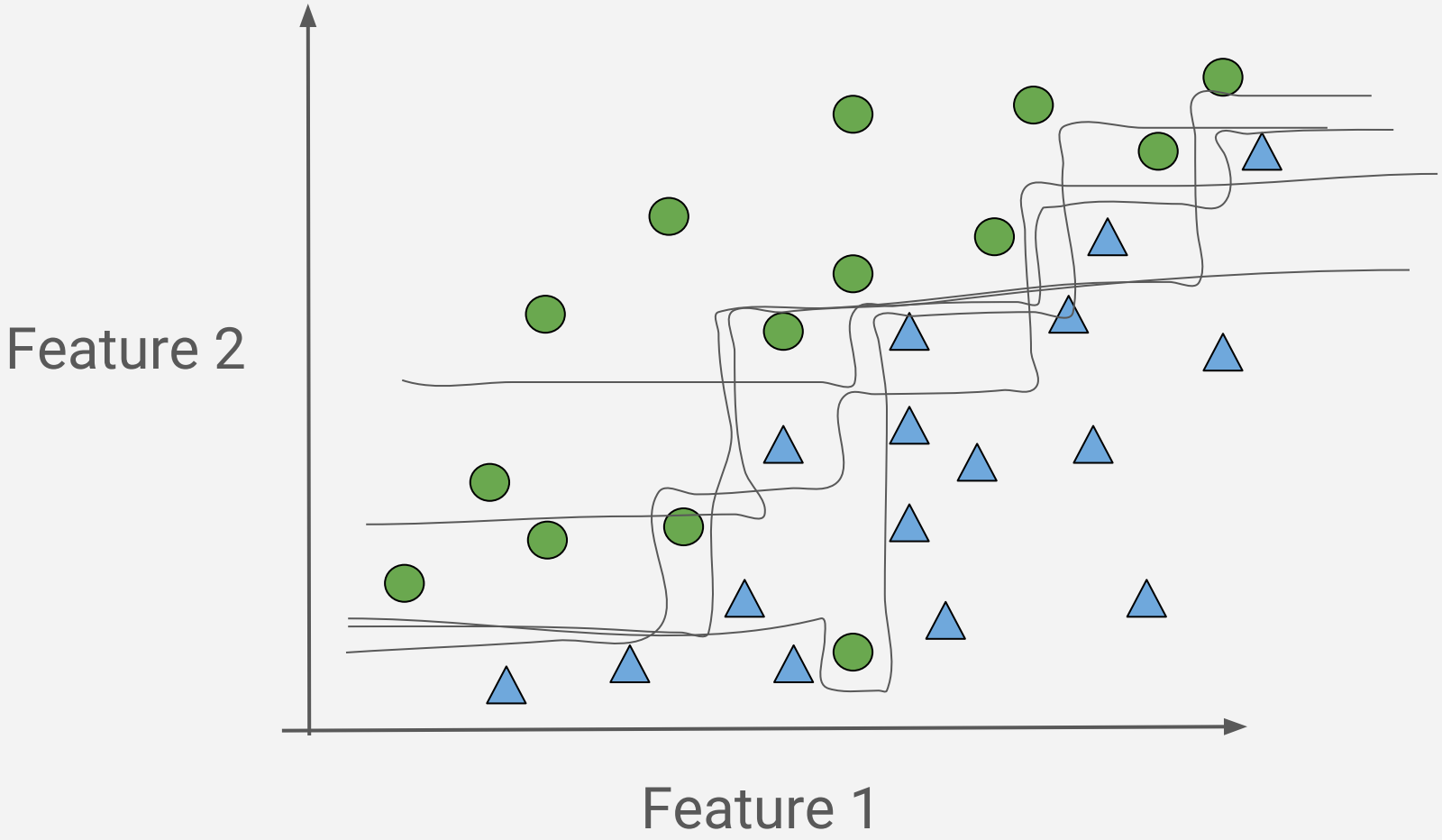

Feature 2
Feature 1

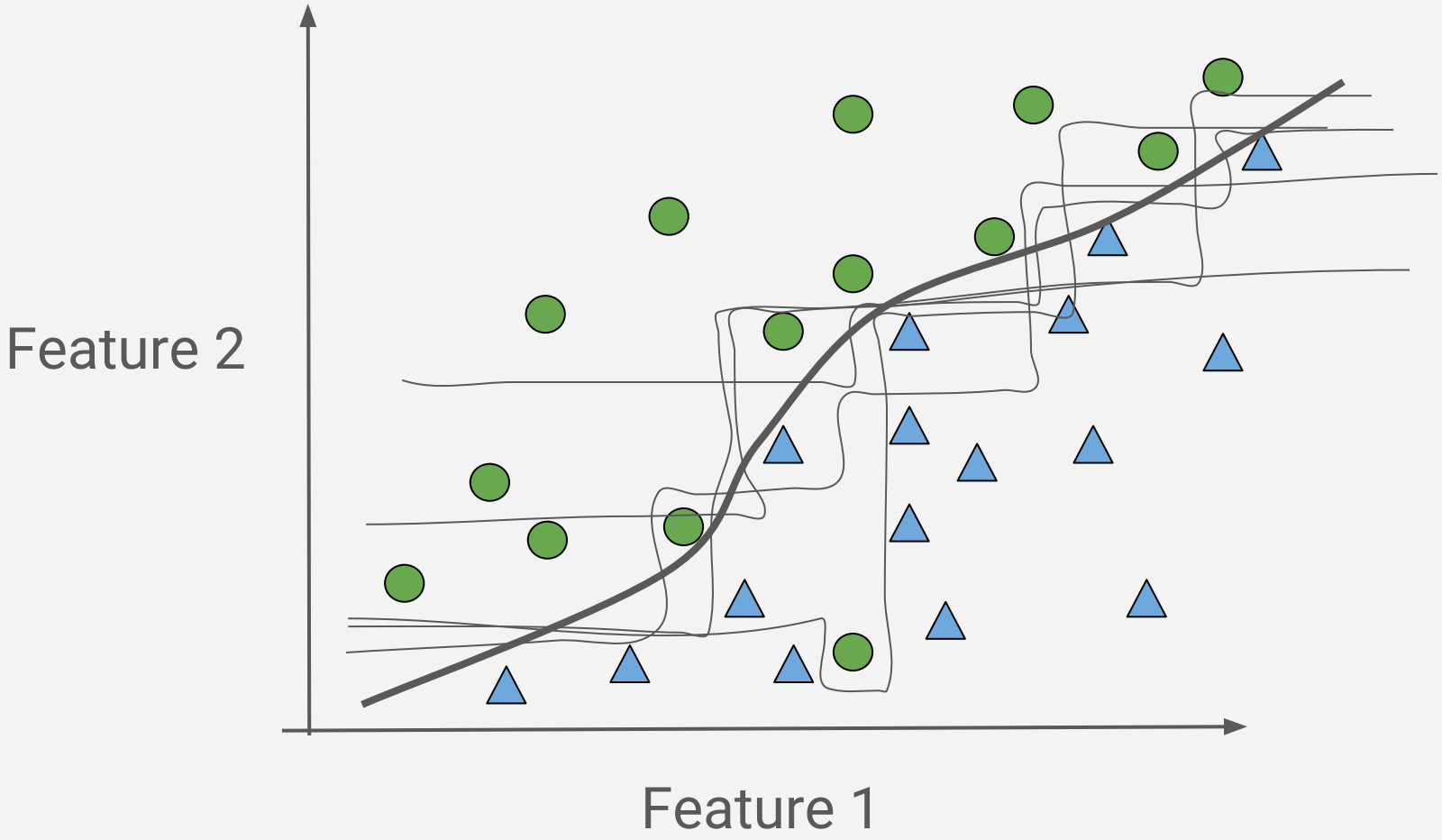

Feature 2
Feature 1


Feature 2
Feature 1


Feature 2
Feature 1

Let’s aggregate!


Feature 2
Feature 1


Feature 2
Feature 1

Bootstrapping a Dataset

Row #	Age	Gender	Weight	Height	Job	Output

Row #	Age	Gender	Weight	Height	Job	Output
1	31	M	87	181	lawyer	**positive**
2	54	M	79	177	developer	**negative**
3	34	F	56	165	lawyer	**positive**
4	25	F	52	161	developer	**negative**

Row #	Age	Gender	Weight	Height	Job	Output
1	31	M	87	181	lawyer	**positive**
2	54	M	79	177	developer	**negative**
3	34	F	56	165	lawyer	**positive**
4	25	F	52	161	developer	**negative**

Row #	Age	Gender	Weight	Height	Job	Output
1	31	M	87	181	lawyer	**positive**
2	54	M	79	177	developer	**negative**
3	34	F	56	165	lawyer	**positive**
4	25	F	52	161	developer	**negative**

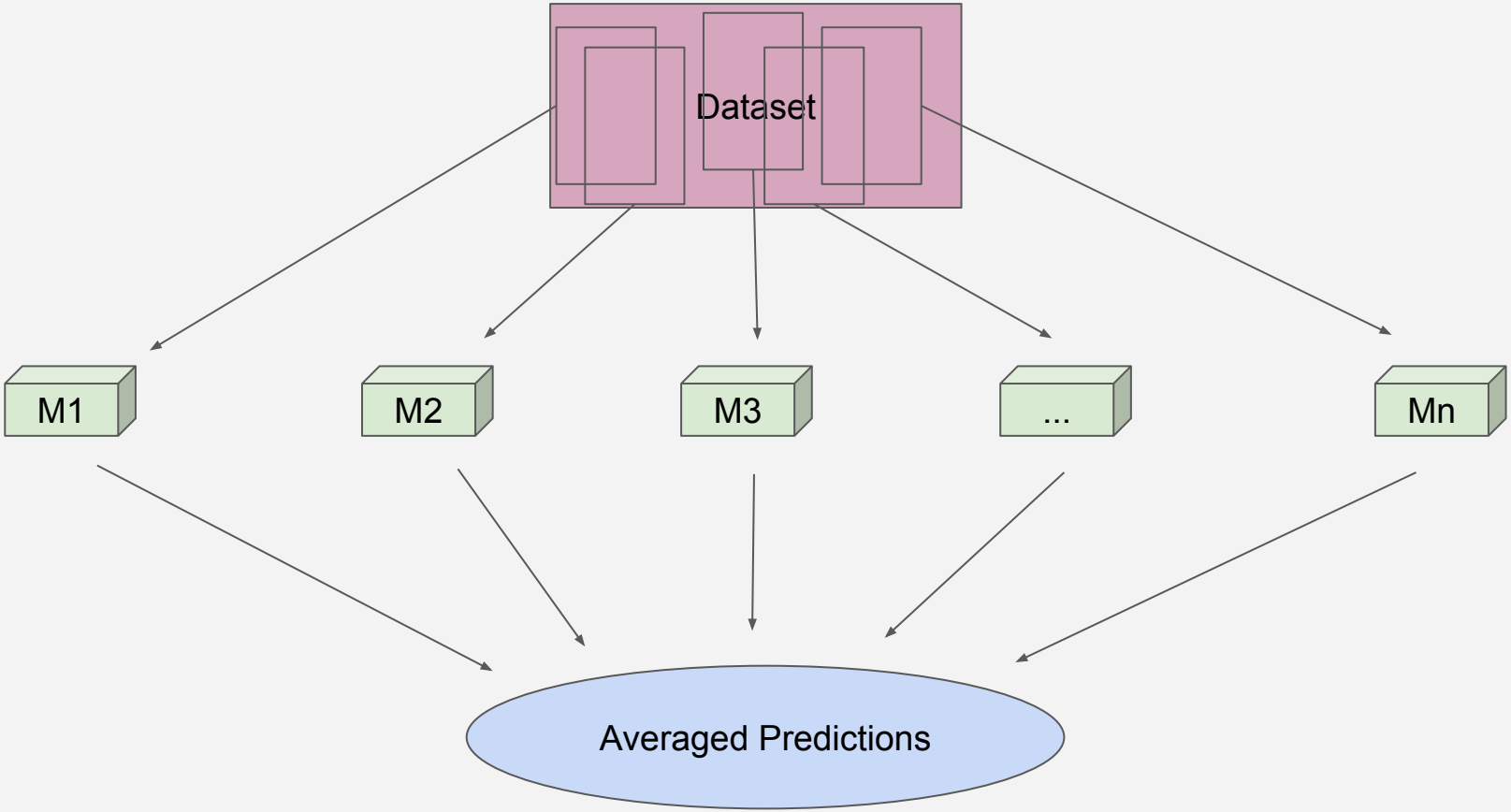
Aggregating:
2 ways

1 . Bagging

Bagging

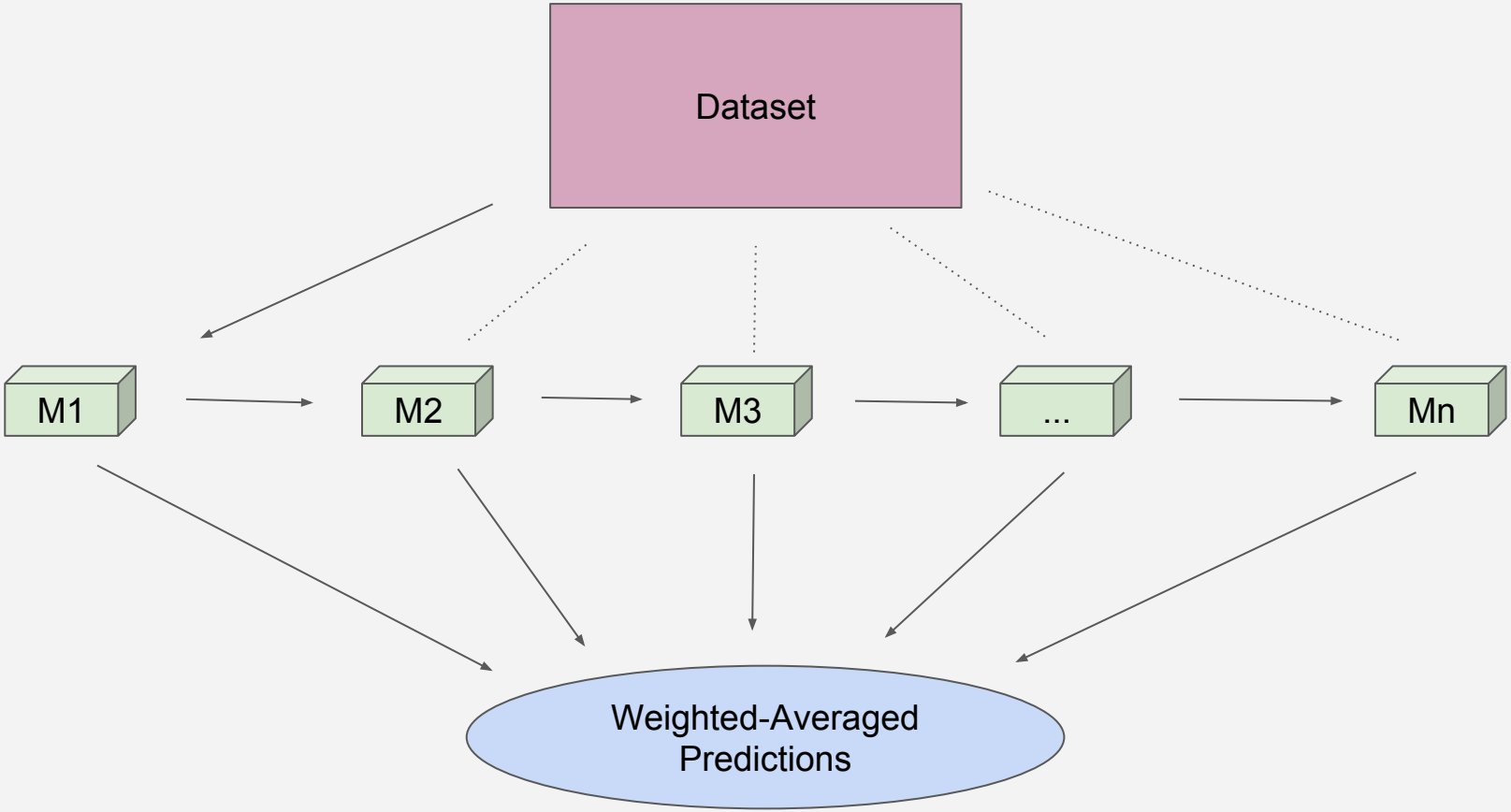
=

Bootstrap + Aggregating


Dataset
M1
M2
M3
...
Mn
Averaged Predictions

E.g. Random Forest

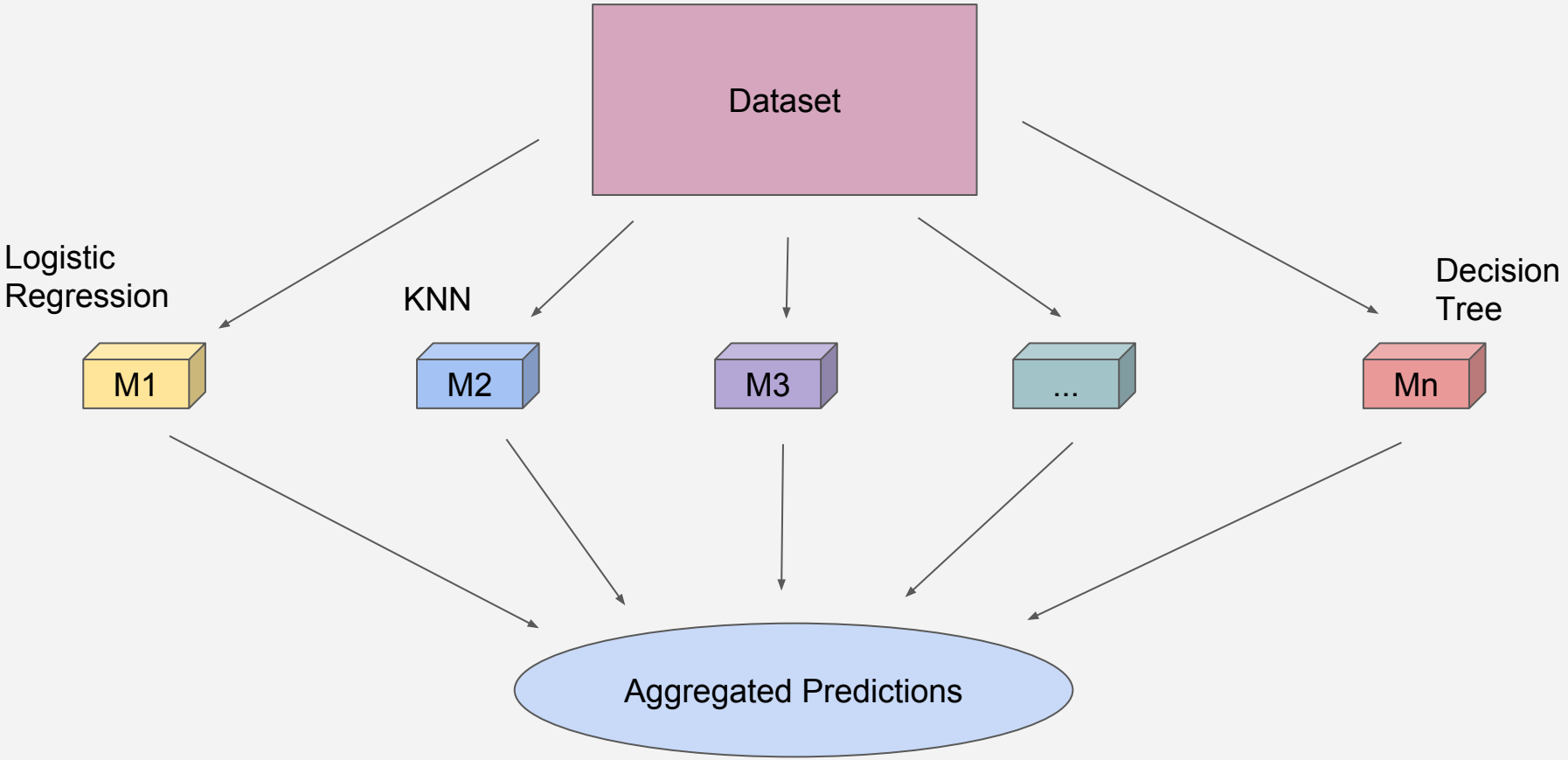
2. Boosting


Dataset
M1
M2
M3
...
Mn
Weighted-Averaged
Predictions

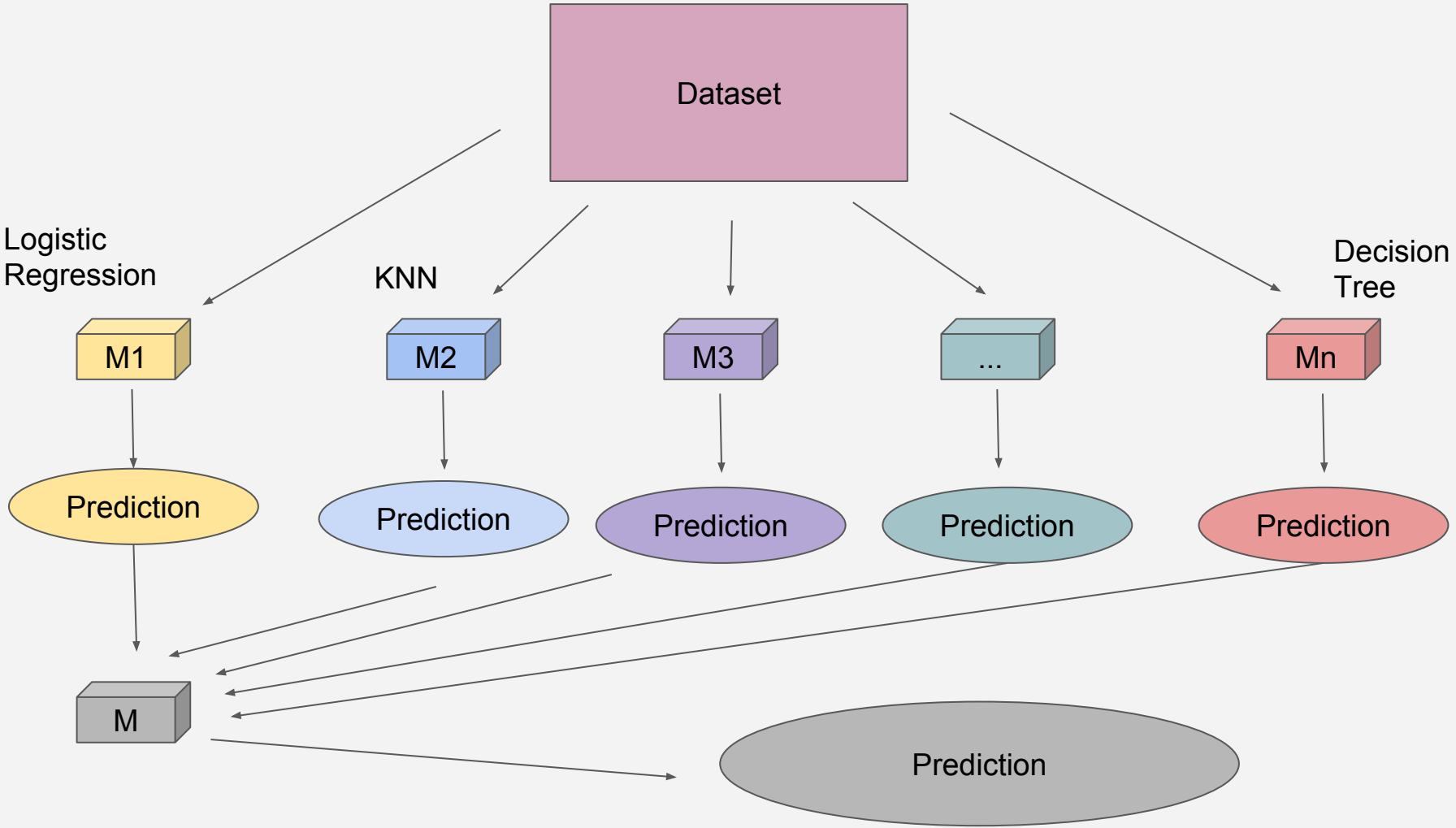
E.g. AdaBoost, Gradient Boost

(lesson tomorrow)

Bonus 1 - Voting


Dataset
Logistic Regression
KNN
Decision Tree
M1
M2
M3
...
Mn
Aggregated Predictions

Bonus 2 - Stacking


Dataset
Logistic Regression
KNN
Decision Tree
M1
M2
M3
...
Mn
Prediction
Prediction
Prediction
Prediction
Prediction
M
Prediction

Takeaways

RF is very powerful

(especially for classification)

RF is very versatile

RF is easy to use

RF is (almost) a black box

Aggregating models produces powerful models