AIBridge

Lecture 3

Classification!

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

→ Model →

White = 1

Red = 0

- categorical label outputs are named “**classes**”

Classification!

quick review
wine dataset

that's a lot of features!

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

→ Model →

White = 1

Red = 0

- categorical label outputs are named "**classes**"

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

■ Linear models might not be the best in some cases

5 ML Models for Classification

1. Decision Tree
2. Random Forest
3. Support Vector Machine (SVM)
4. Naïve Bayes
5. K Nearest Neighbors

Decision Trees

Decision Trees

Can I afford it?

Decision Trees

Can I afford it?

Is it comfortable?

Decision Trees

Can I afford it?

Is it comfortable?

Is it fashionable?

Decision Trees

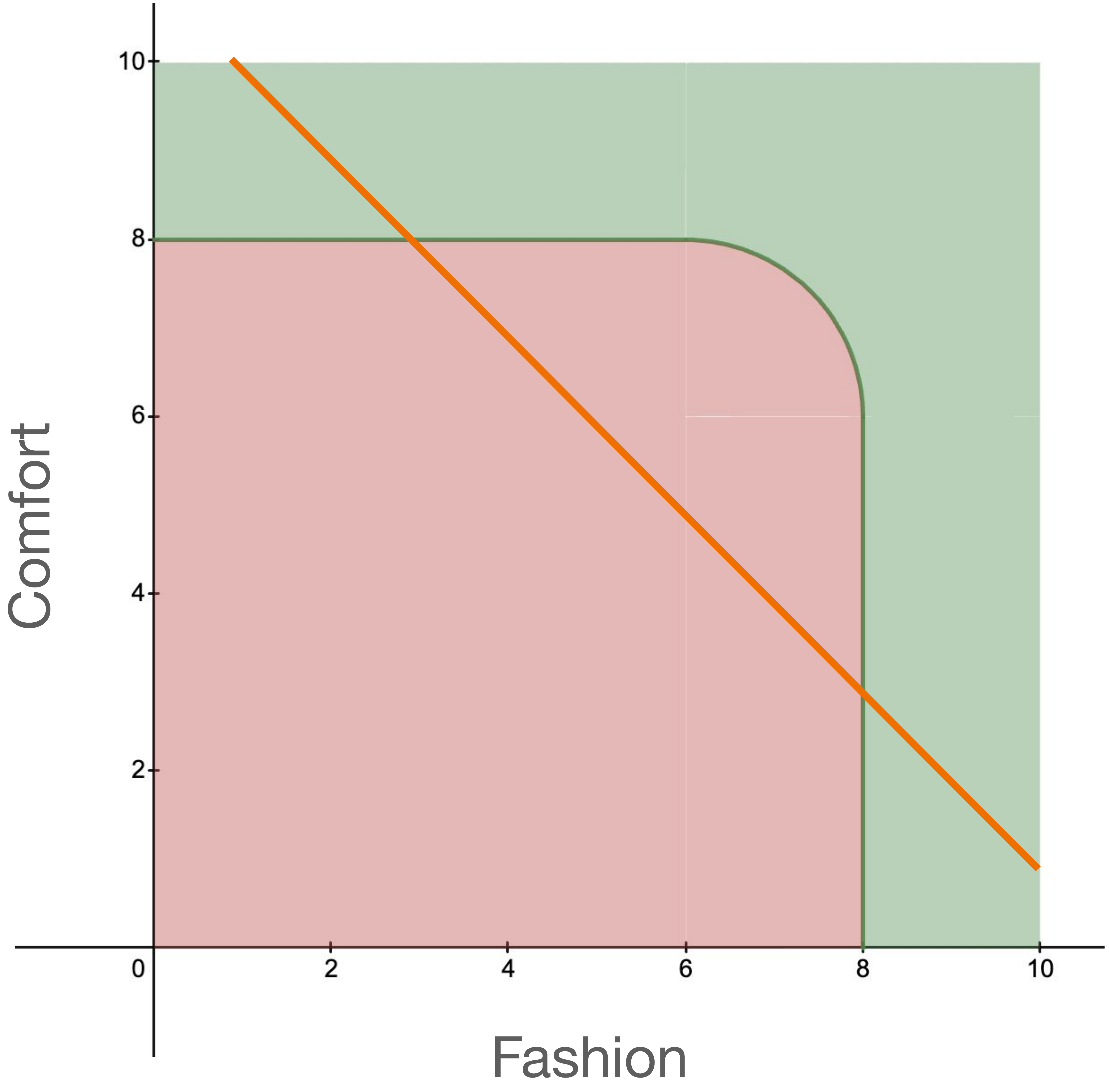
Price	Comfort	Fashion	Purchased?
$70	4	6	No
$120	5	8	No
$20	4	4	No
$60	1	8	Yes
$60	6	3	No
$80	8	8	Yes

Decision Trees

Can I afford it?

Is it comfortable?

Is it fashionable?



Decision Trees

Decision Trees

- flowcharts of decisions can create an explainable and repeatable graph of predictions

Decision Trees

Price	Comfort	Fashion	Purchased?
$70	4	6	No
$120	5	8	No
$20	4	4	No
$60	1	8	Yes
$60	6	3	No
$80	8	8	Yes

Decision Trees

Purchased?

No

No

No

Yes

No

Yes

Decision Trees

No	Yes
No	No
No	Yes

Decision Trees

No	Yes
No	No
No	Yes

Decision Trees

Which one is a better split?

No	Yes
No	No
No	Yes

All no

No	
No	No
No	Yes
Yes	

Mostly no

Decision Trees

Gini impurity

- as a group becomes more **homogeneous**, its **Gini Impurity** decreases.

Decision Trees

Fraction of that one class in group

Fraction of not that one class in the group

$$G = \sum_{i=1}^{C} P(i) \cdot (1 - P(i$$

Add them up for all classes (in one side of the split)

- **Gini impurity** measures the homogeneity in a group

Decision Trees

Purchased?	
No	0
No	
No	0.5
Yes	
No	
Yes	
	0.5

Decision Trees

Purchased?	
No	
No	0.38
No	
Yes	
No	0.5
Yes	
	0.88

Decision Trees

we gotta do better than this, right?

Purchased?	
No	
No	0
No	
Yes	
No	0.44
Yes	
	0.44

Decision Trees

just split again!

Purchased?	
No	
No	0
No	
Yes	
No	0.44
Yes	
	0.44

Decision Trees

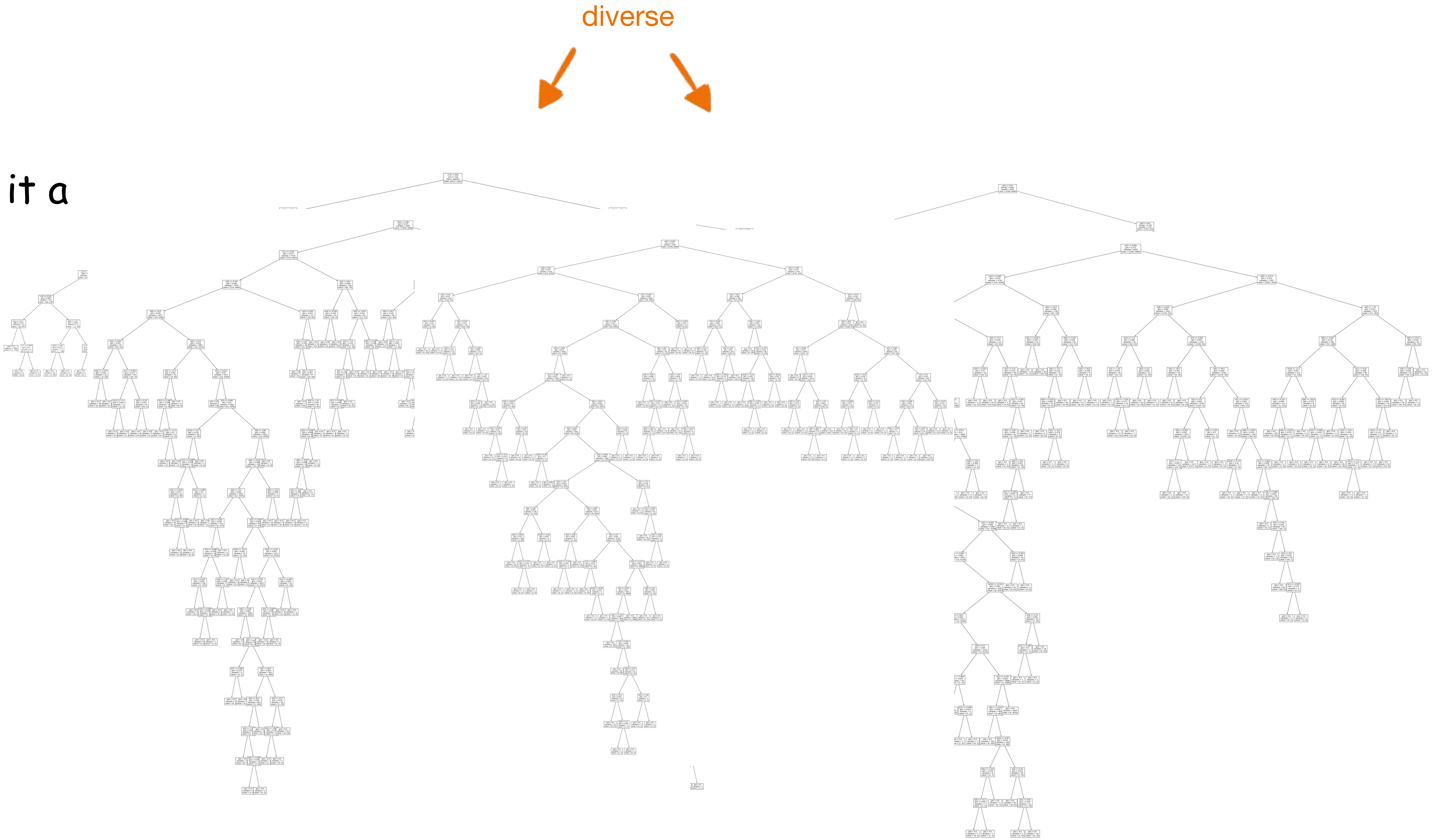
1. Calculates Gini impurities for ALL possible splits
2. Select the split that results in the lowest Gini impurity sum
3. Implements the split
4. **Split again** – repeat steps 1-3 as much as needed

a **hyperparameter**

5 ML Models for Classification

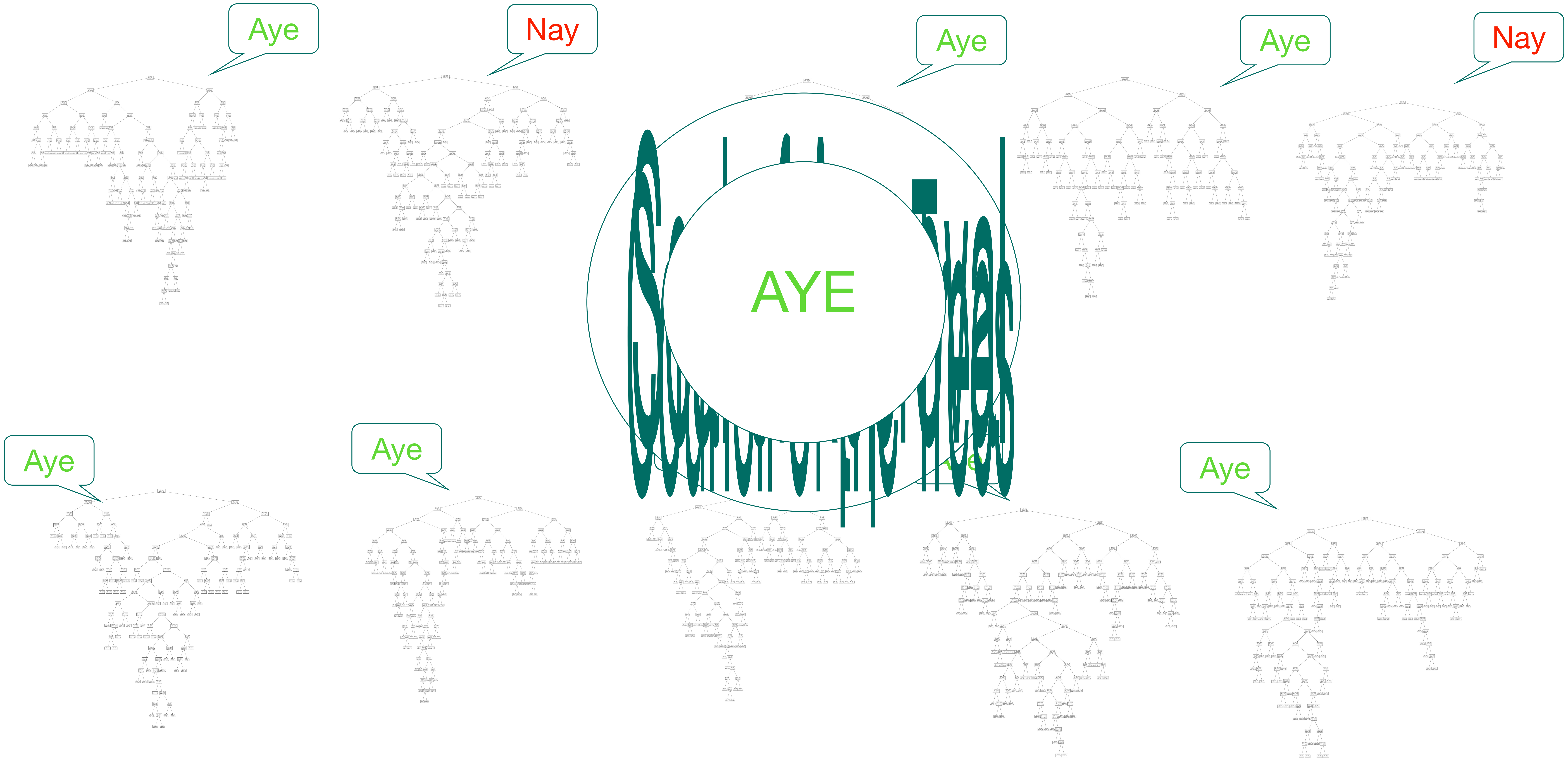
1. Decision Tree
2. Random Forest
3. Support Vector Machine (SVM)
4. Naïve Bayes
5. K Nearest Neighbors

Random Forest



What if we do it a lot?

Random Forest



Random Forest

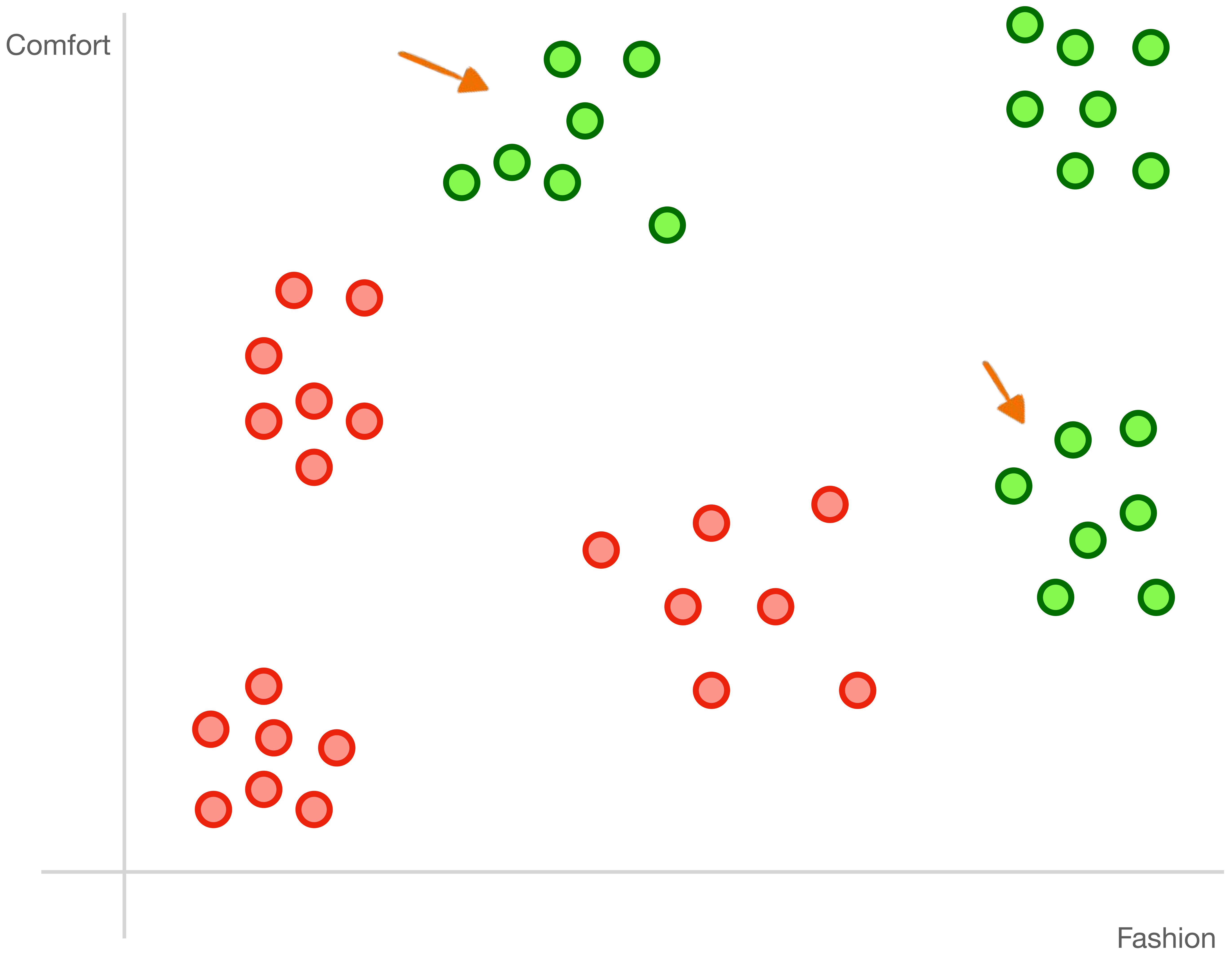
1. Make a lot of decision trees, on different portions of the data
2. For a new sample, run all of them
3. Combine their votes and take the majority

5 ML Models for Classification

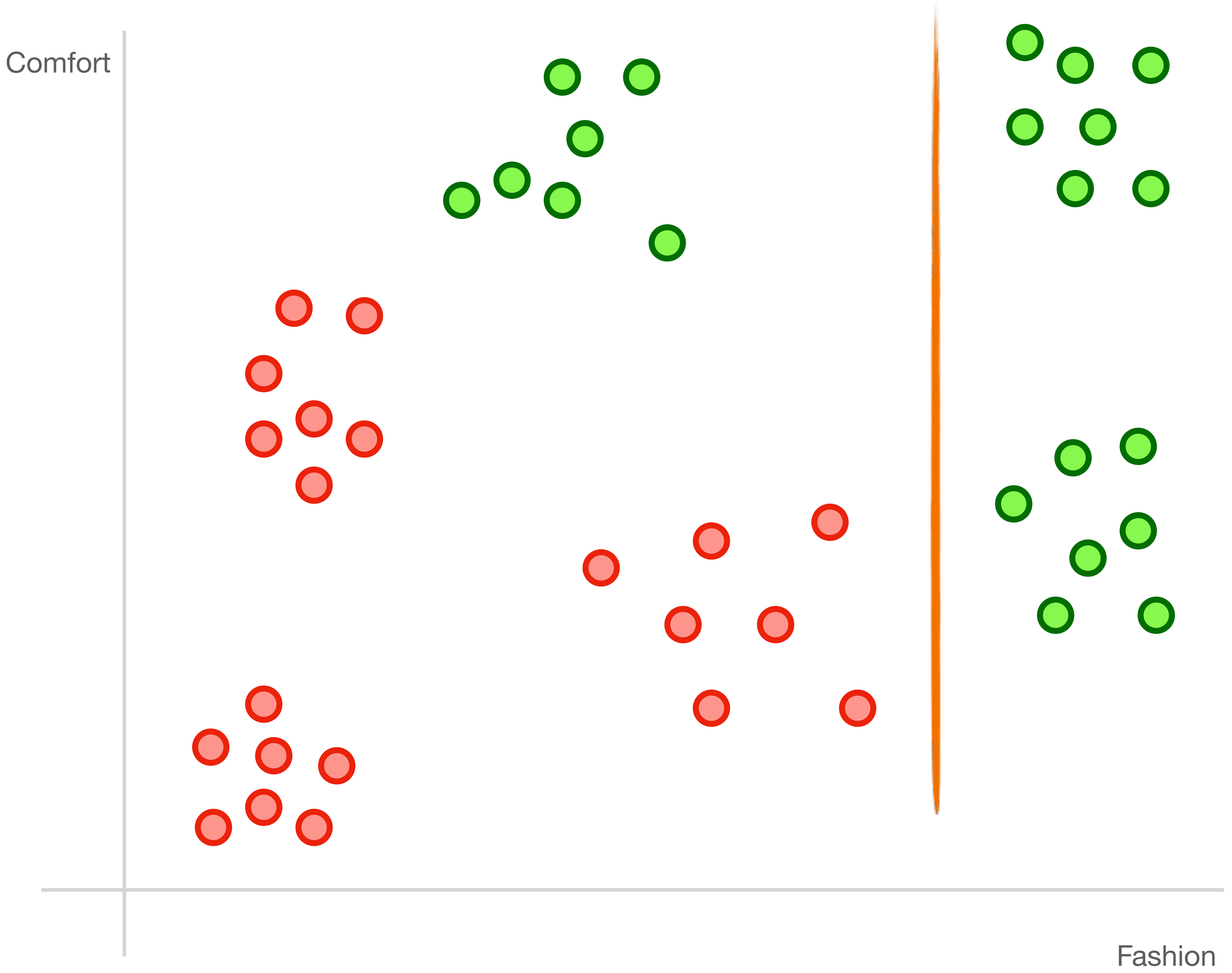
1. Decision Tree
2. Random Forest
3. Support Vector Machine (SVM)
4. Naïve Bayes
5. K Nearest Neighbors

“split”

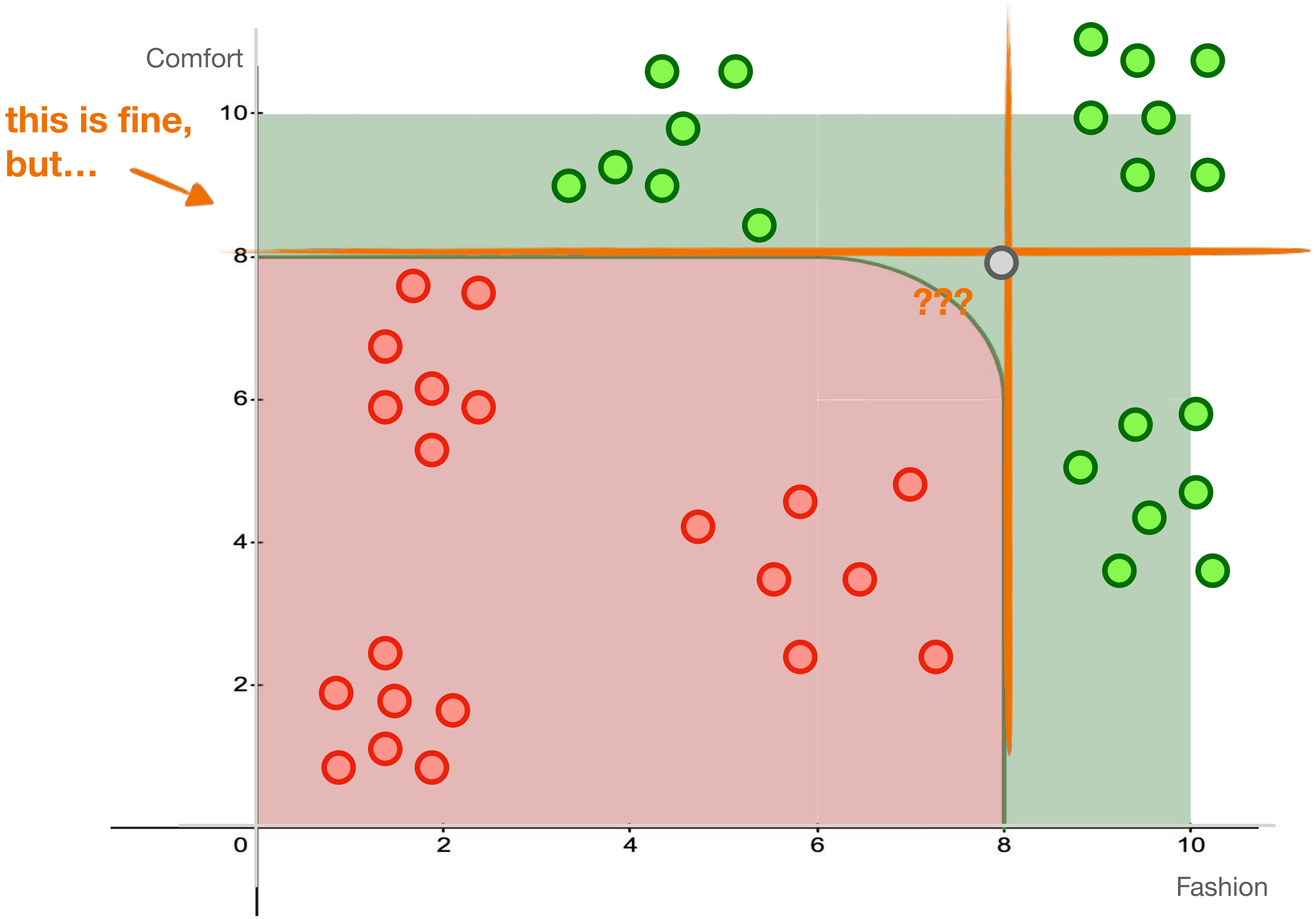
“split”



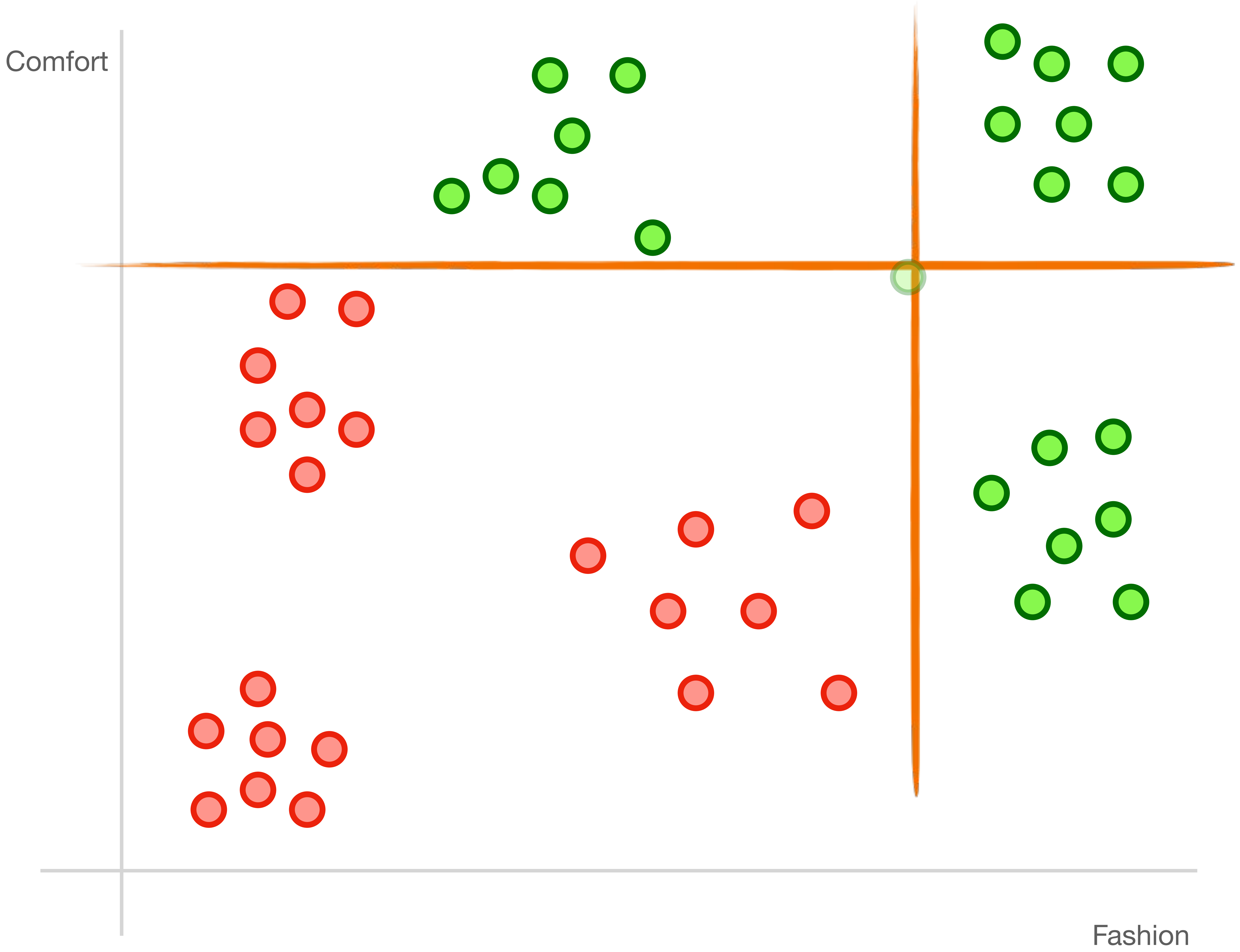
“split”



“split”



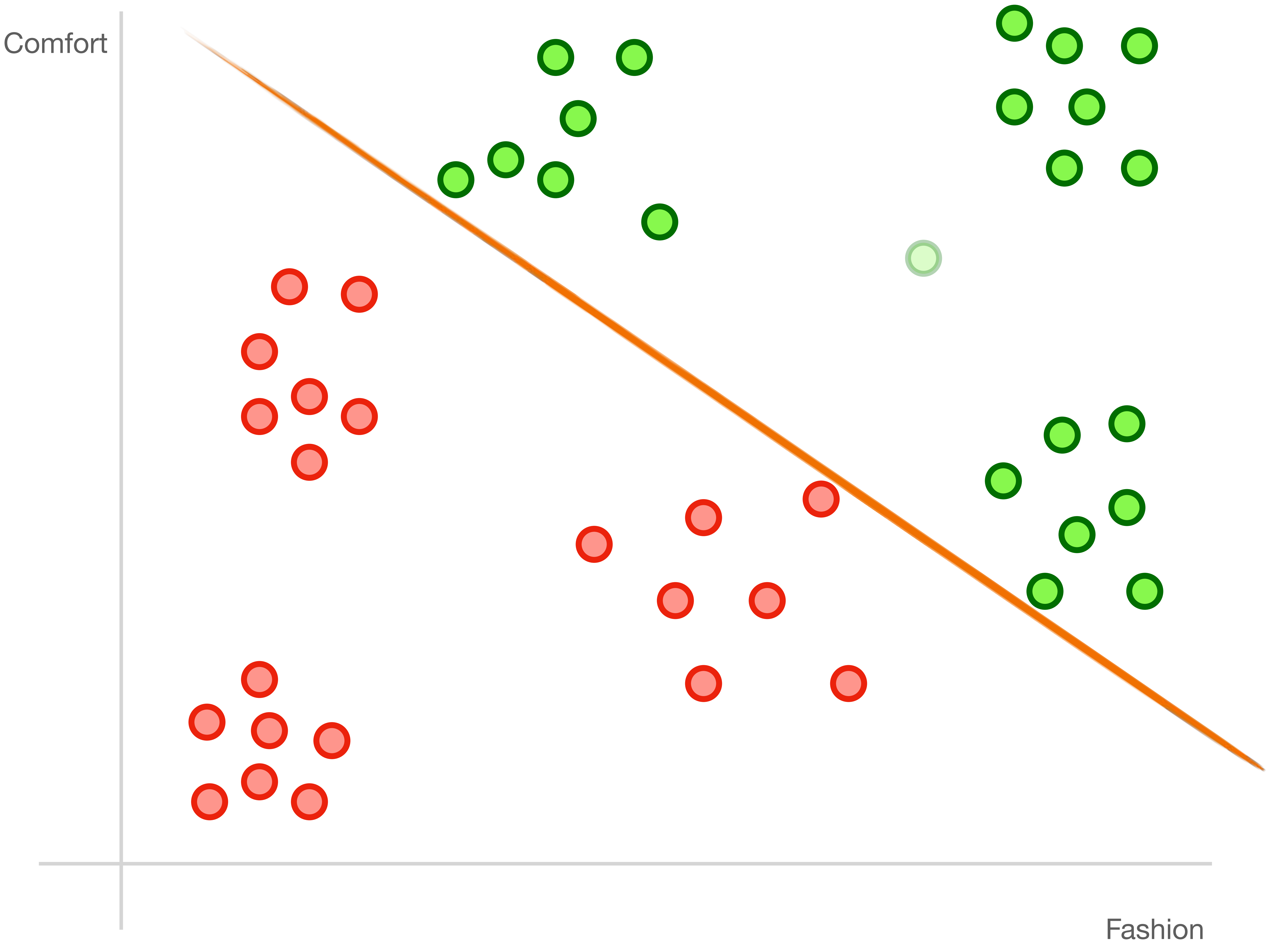
“split”



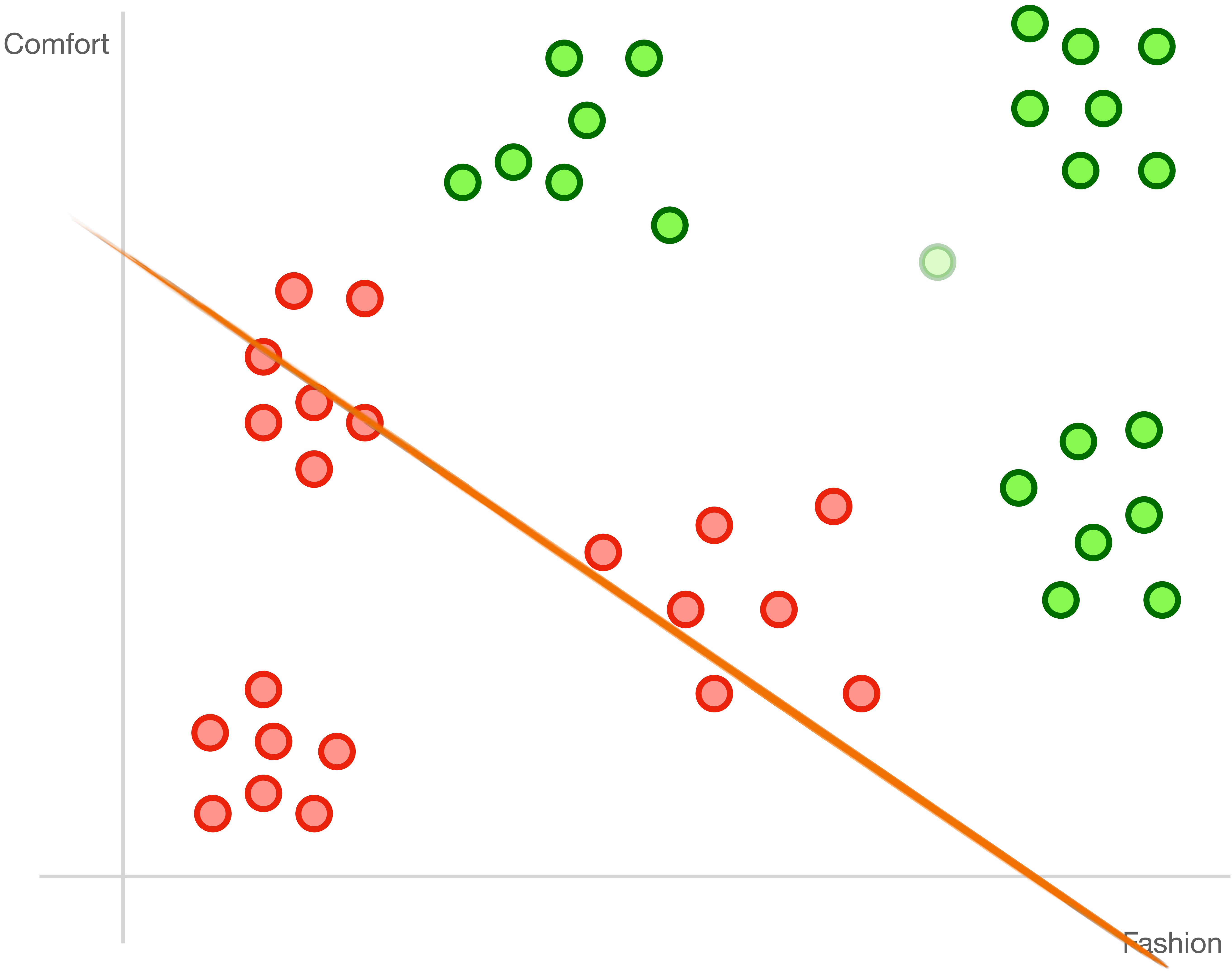
we need a more complex split

Support vector machines!

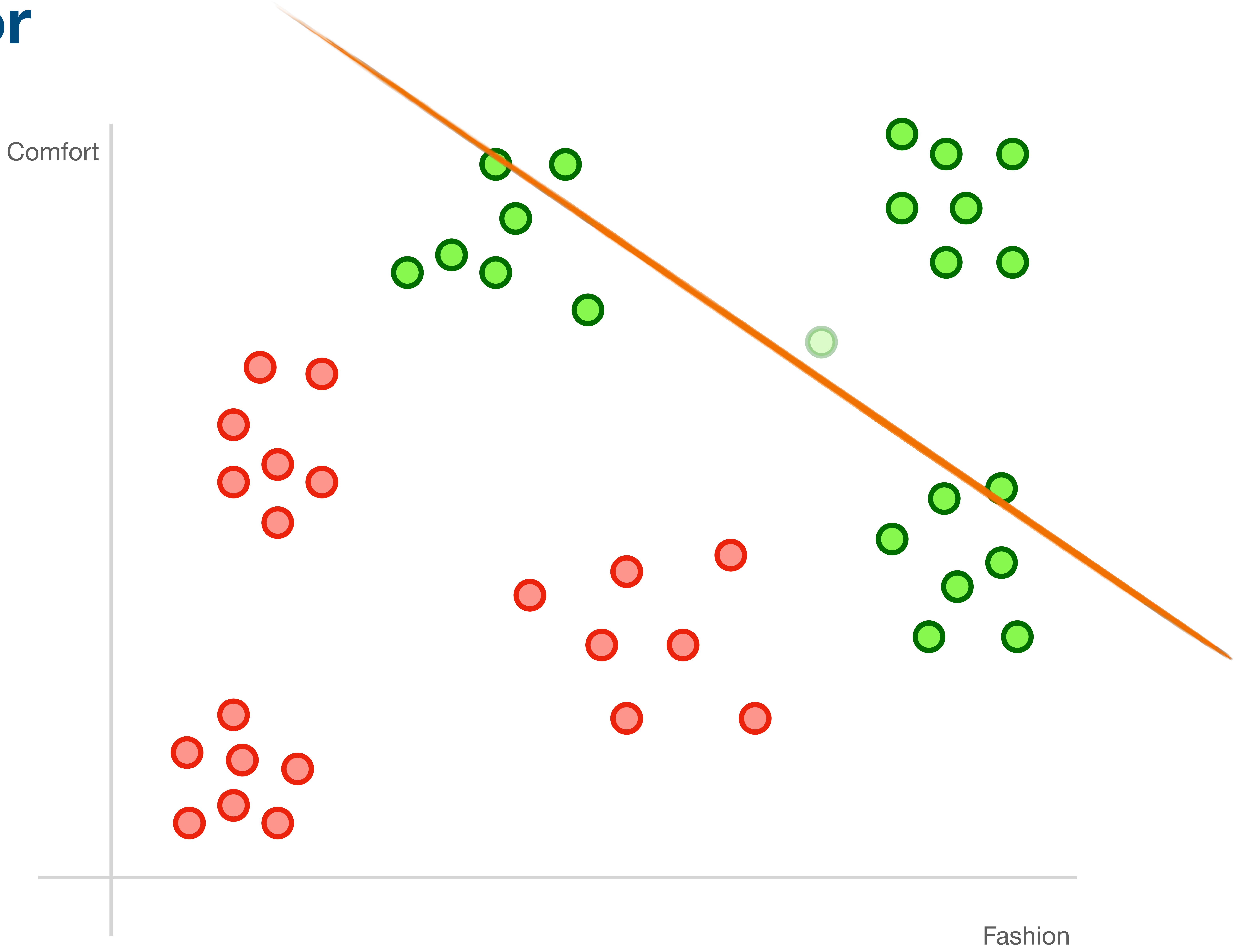
Support Vector Machines



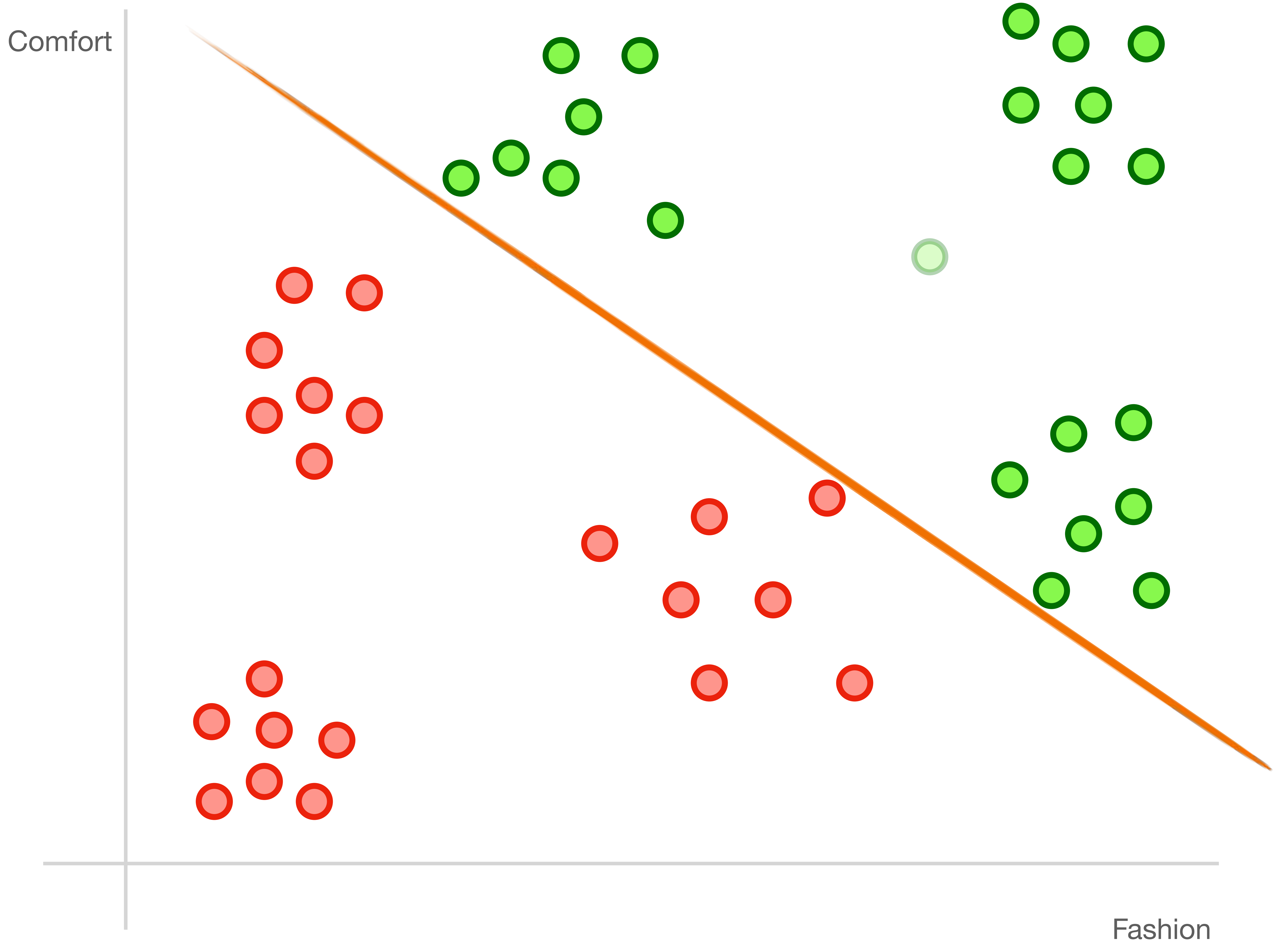
Support Vector Machines

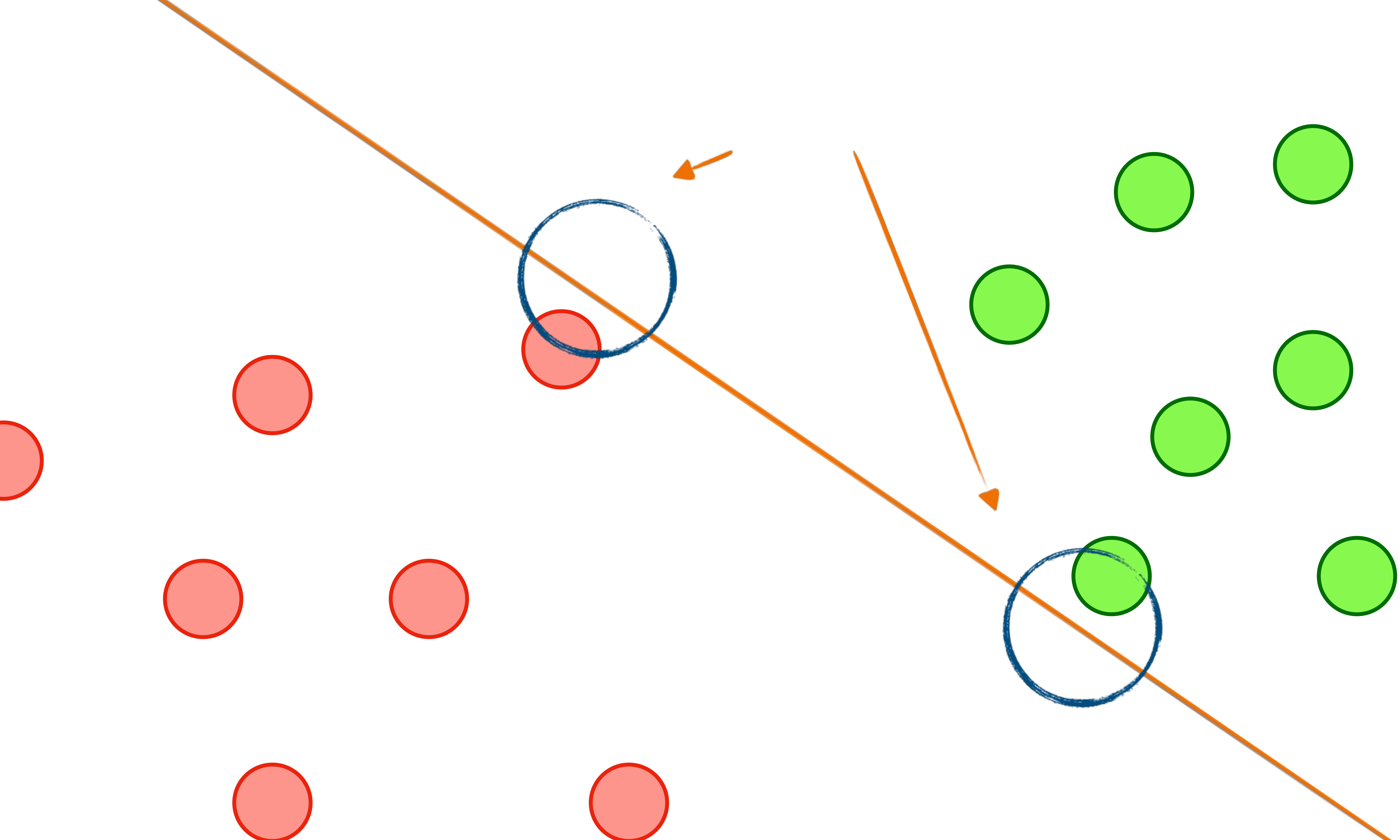


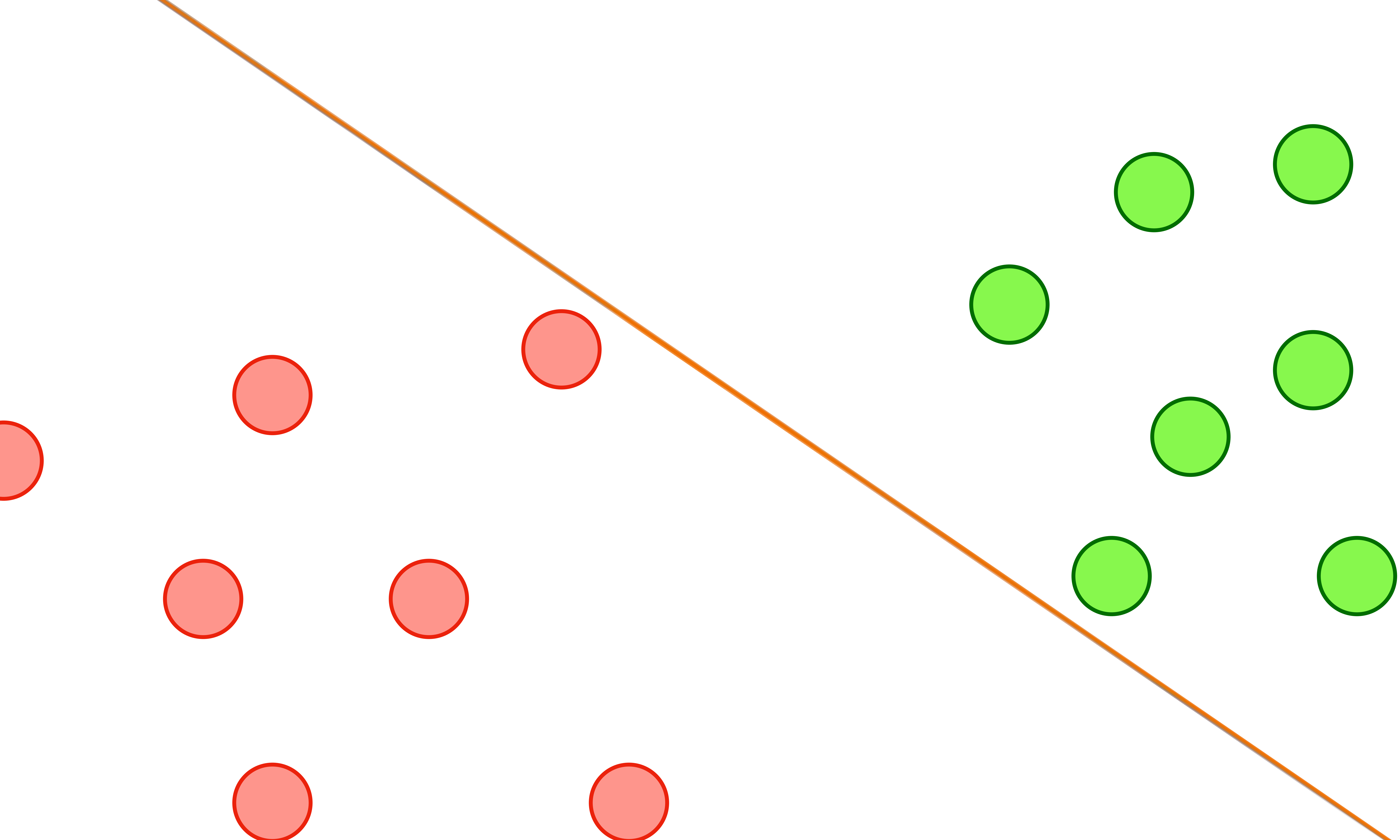
Support Vector Machines

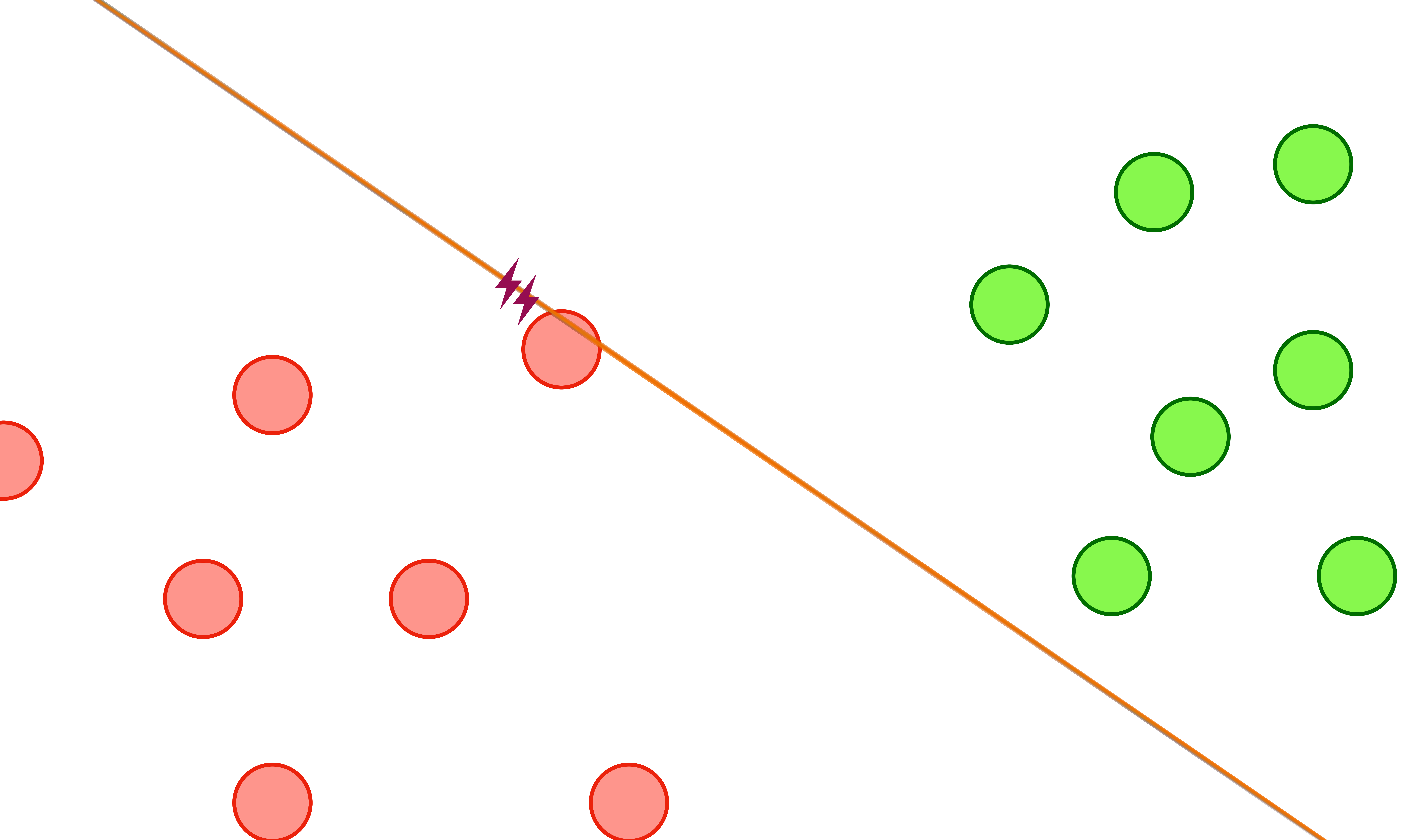


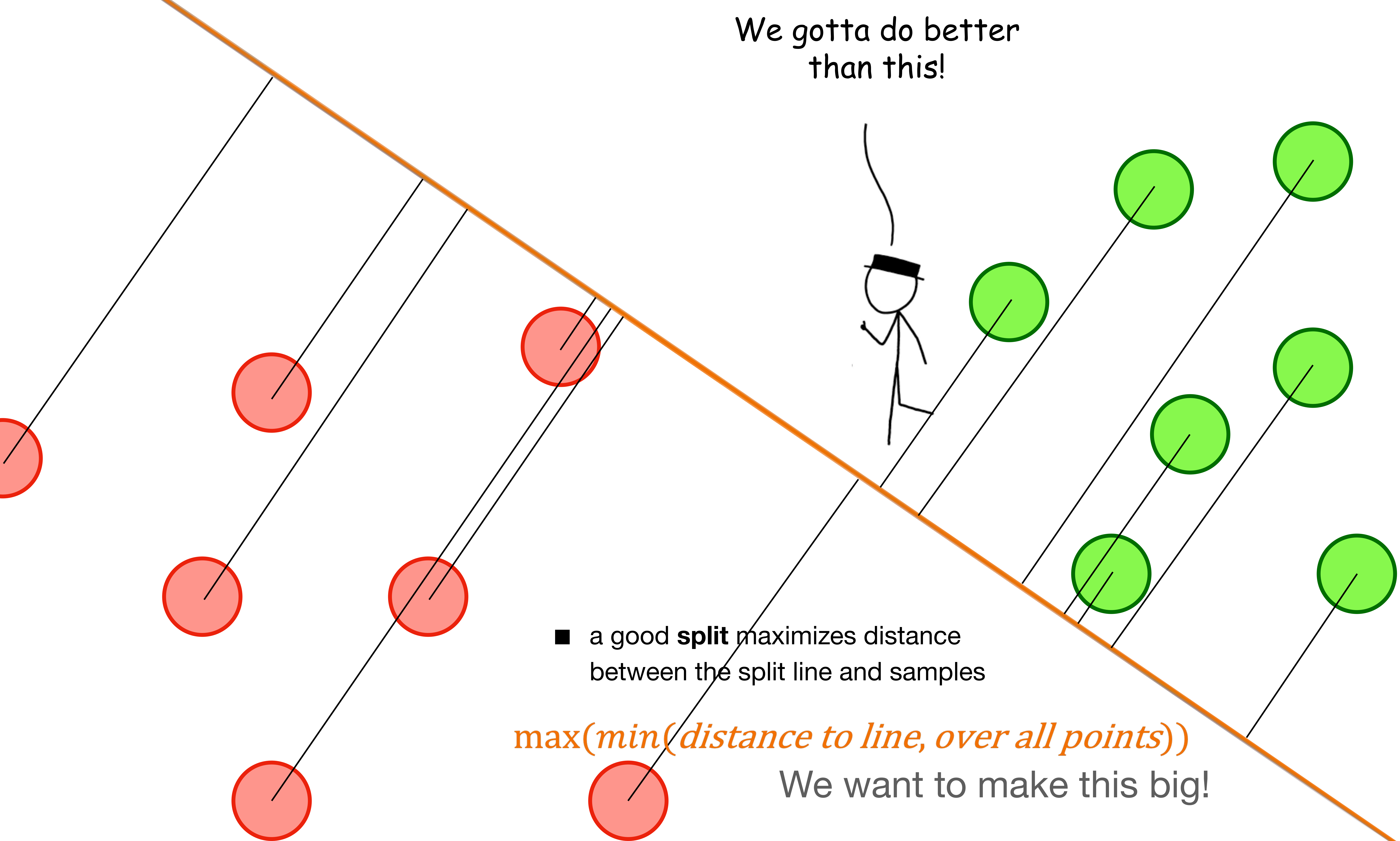
Support Vector Machines











- a good **split** maximizes distance between the split line and samples

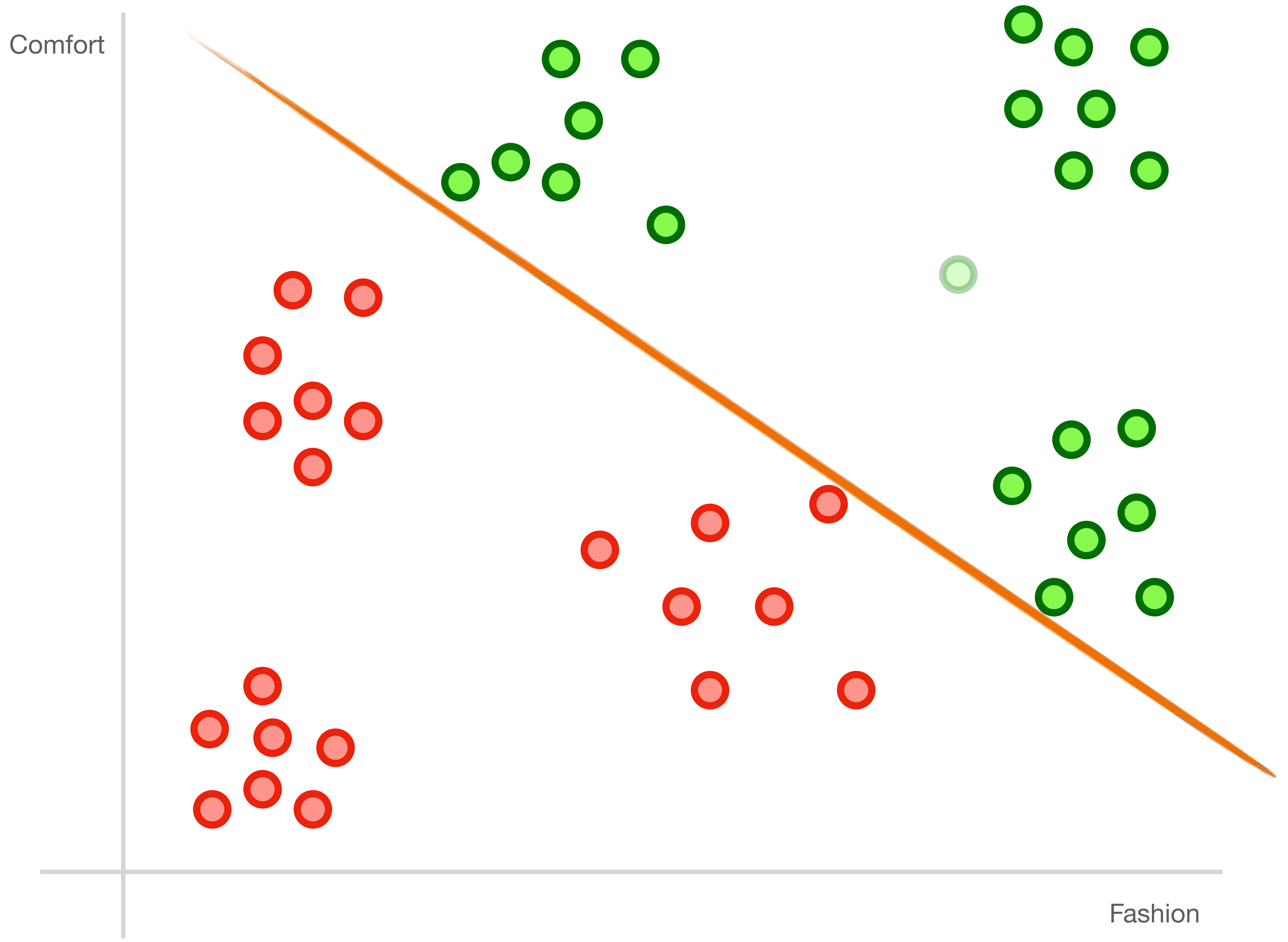
$$\max(min(distance\ to\ line,\ over\ all\ points))$$

We want to make this big!

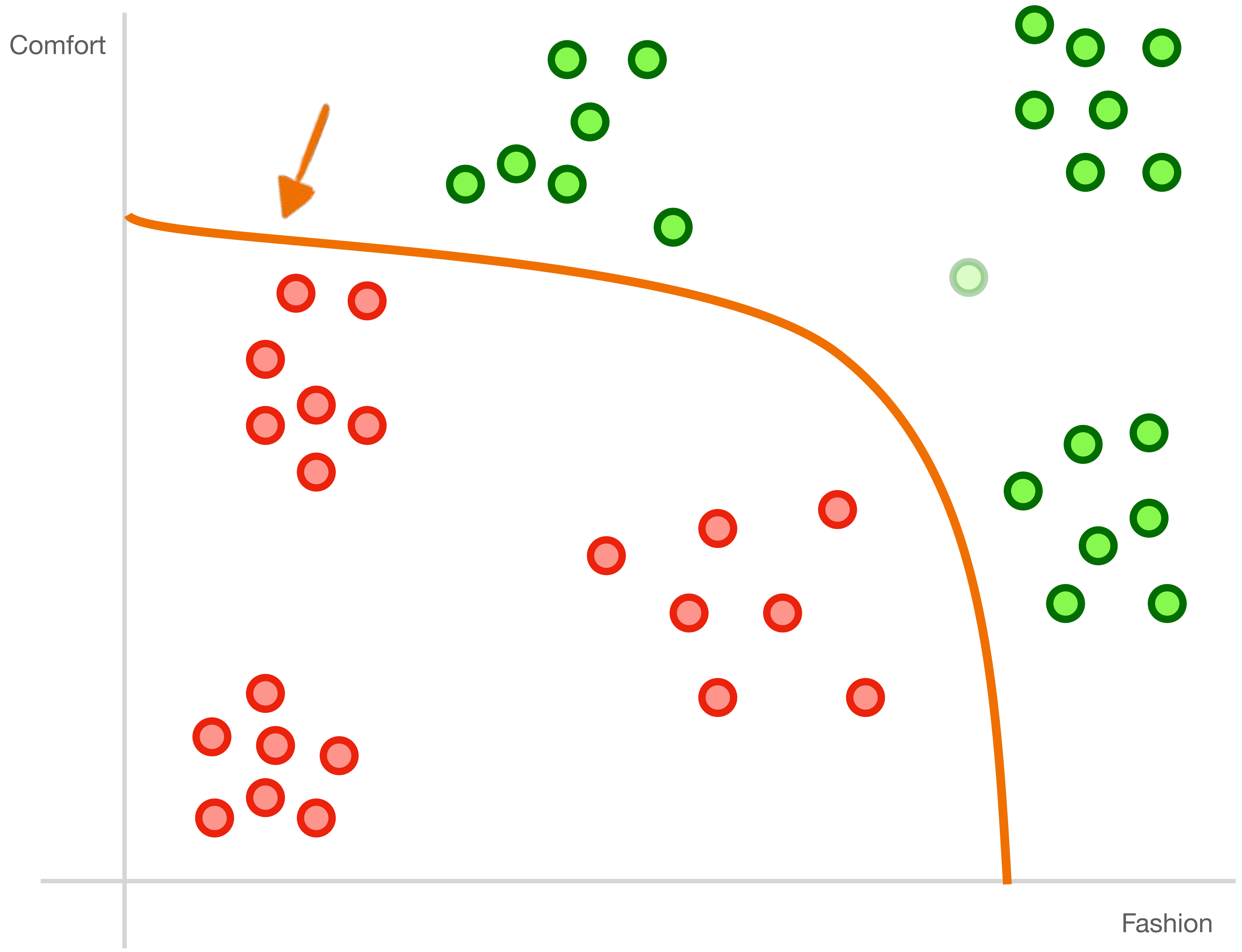
Support Vector Machines

$\min(\textit{distance to line, over all points})$

We want to make this big!



Support Vector Machines



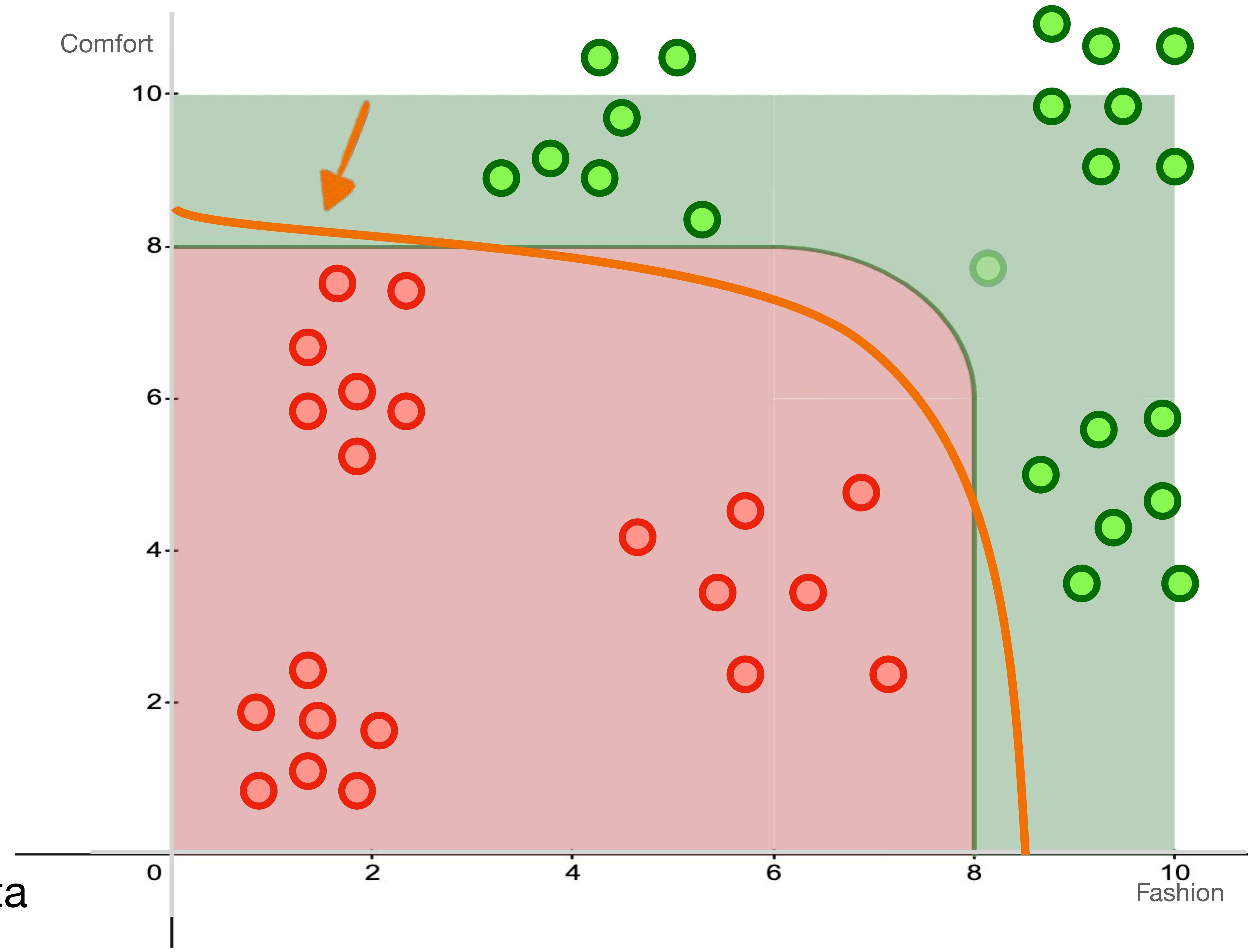
$\min(\textit{distance to line, over all points})$

We want to make this big!

Support Vector Machines

$\min(\textit{distance to line, over all points})$

We want to make this big!



- **support-vector machines** are classifiers that divide data by class, aiming to create a **margin** that's as wide as possible.
- They can use non-linear functions

5 ML Models for Classification

1. Decision Tree
2. Random Forest
3. Support Vector Machine (SVM)
4. Naïve Bayes
5. K Nearest Neighbors

Internal Memo:

146 Hagley Road, Birmingham
Birmingham B3 3PJ

From the Desk of
Mr. Jerry Smith
Date: 13/01/14

Attn: Sir/Madam,

I seize this opportunity to extend my unalloyed compliments of the new season to you and your family hopping that this year will bring more joy, happiness and prosperity into your house hold.

I am certain that by the time you read this letter I might have already gone back to my country **United Kingdom**. I visited South Africa during the New Year period and during my stay, I used the opportunity to send you this letter believing that it will reach you in good state.

Internal Memo:

146 Hagley Road, Birmingham
Birmingham B3 3PJ

From the Desk of
Mr. Jerry Smith
Date: 13/01/14

Attn: Sir/Madam,

I seize this opportunity to extend my unalloyed compliments of the new season to you and your family hopping that this year will bring more joy, happiness and prosperity into your house hold.

I am certain that by the time you read this letter I might have already gone back to my country **United Kingdom**. I visited South Africa during the New Year period and during my stay, I used the opportunity to send you this letter believing that it will reach you in good state.

“unalloyed complements” → Spam

“$100,000 dollars” → Spam

“relative dying of cancer” → Spam

IF we have this

we get this

“unalloyed complements”	→	Spam
“$100,000 dollars”	→	Spam
“relative dying of cancer”	→	Spam

we get this IF we have this

we get this IF we have this

$$A|B$$

we get this IF we have this

$$A \mid B$$

- Is Spam
- "Nigerian Prince"

we get this IF we have this

$spam | nigerianprinc$

we get this IF we have this

$$P(spam|nigerianprince)$$

high? Nigerian prince → spam likely

low? Nigerian prince → not spam

- **conditional probabilities** can be used as a classifier!

Naïve Bayes

% of spam in dataset that relates to Nigerian prince

% of spam in dataset

$$P(spam|nigerianprince) = \frac{P(spam)P(nigerianprince|spa}{P(nigerianprince)}$$

% of Nigerian prince in dataset

Naïve Bayes

Classifier

$$P(spam|nigerianprince, offer) = \frac{P(spam)P(nigerianprince|spam)P(offer|spam)}{P(nigerianprince)P(offer)}$$

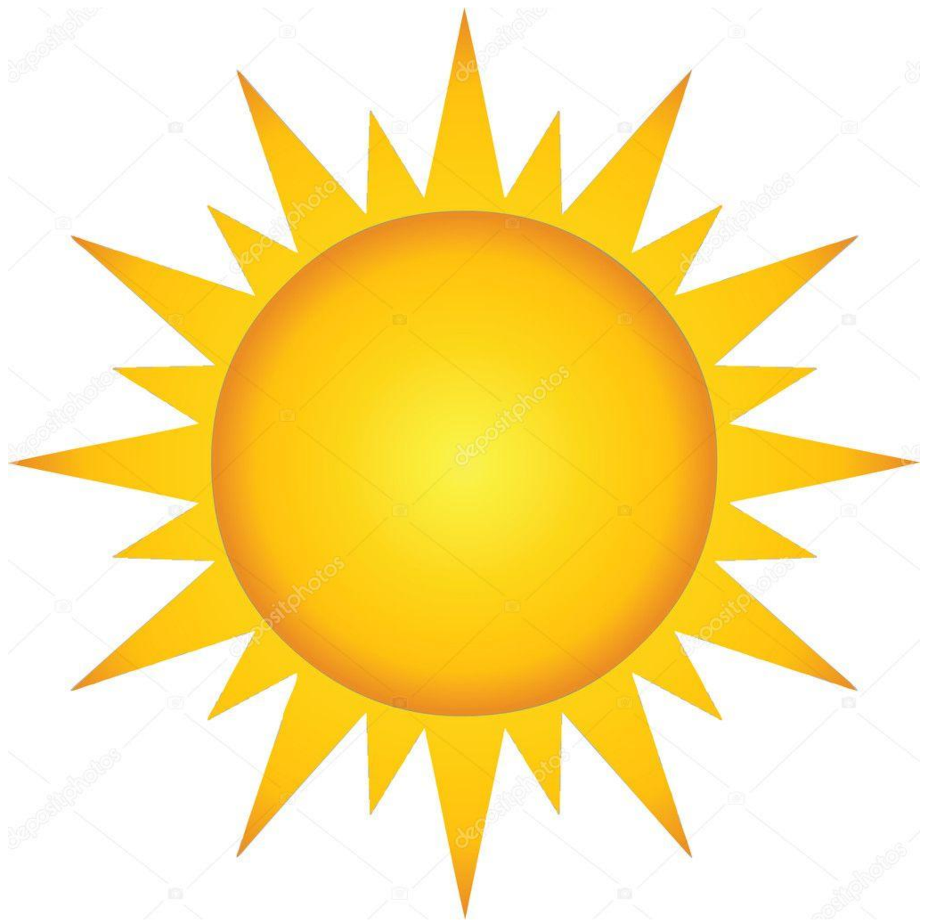
"naïve"

- **conditional probabilities** can be used as a classifier!
- a classifier made this way, however, is "**naïve**" when extended to multiple features

Naïve Bayes

1. For each feature, calculate the probability of this feature given the class and the probability of this feature given the entire dataset
   - For the spam dataset, iterate through each word in the entire dataset and use each word as a feature
2. Given an unlabeled test sample, determine which features it contains
   - For the spam dataset, use all the distinct words in the dataset that were in the training data as features
3. Calculate the probability that this sample belongs to a certain class using the Naïve Bayes classifier
4. Classify the sample to that class if probability is greater a threshold (50%)

Naïve Bayes
Independence



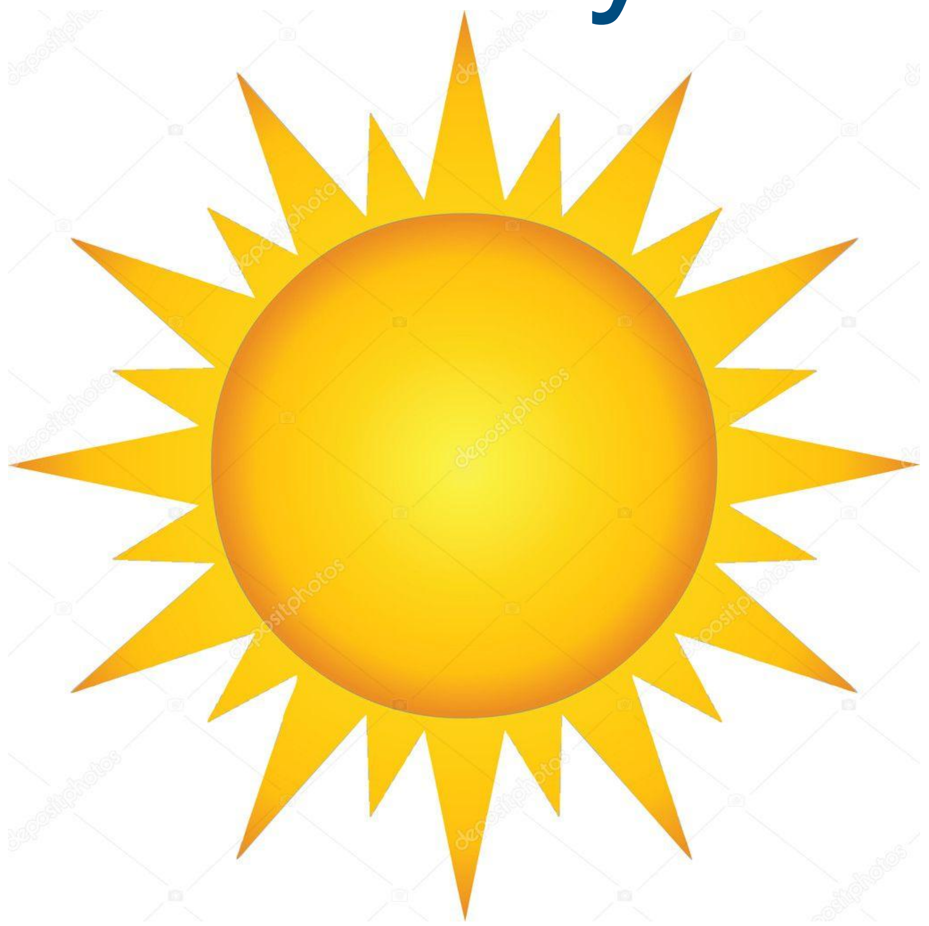
50%

50%

Naïve Bayes

Independence

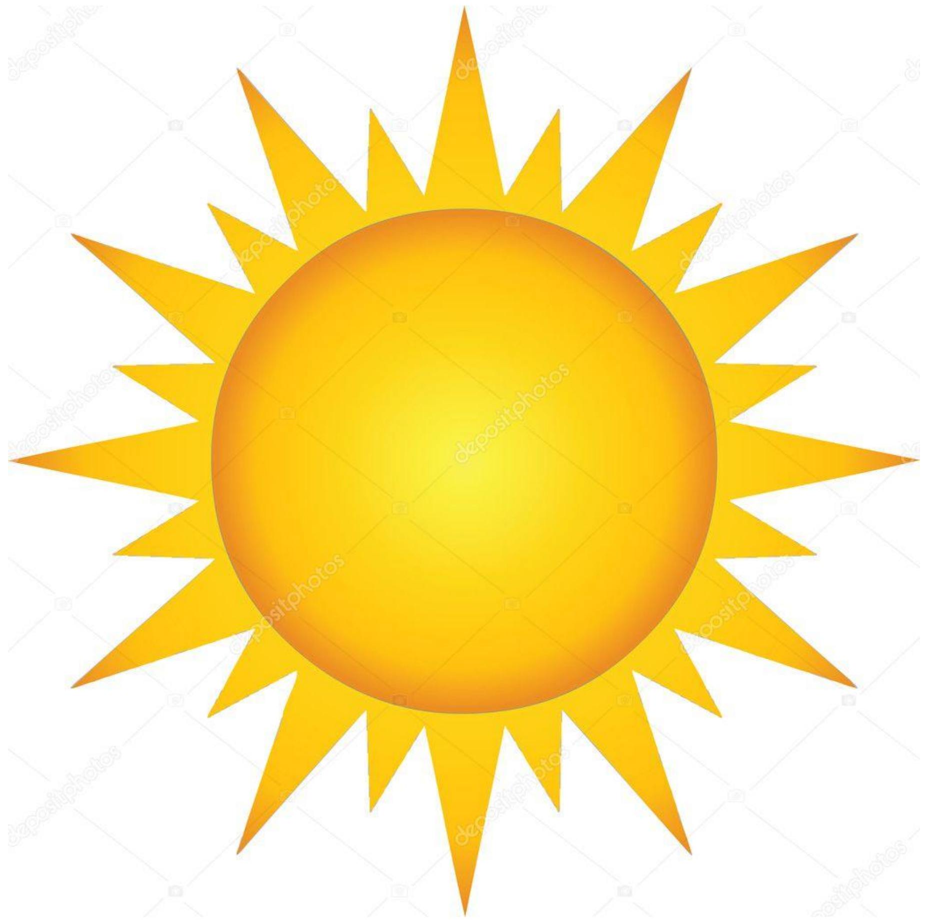
January 1st



50%

50%

January 2nd



50%

50%

Naïve Bayes
Independence

$P(\text{Rain} \mid \text{January } 1^{st})$

$=50\%$

$P(\text{Rain} \mid \text{January } 2^{nd})$

$=50\%$

Naïve Bayes
Independence

P(Rain | January 1st AND Rain | January 2nd) =45%

Is NOT

P(Rain | January 1st) * P(Rain | January 2nd) =25%

Naïve Bayes Classifier

$$P(spam|nigerianprince, offer) = \frac{P(spam)P(nigerianprince|spam)P(offer|spam)}{P(nigerianprince)P(offer)}$$

"naïve"

- **conditional probabilities** can be used as a classifier!
- a classifier made this way, however, is "**naïve**" when extended to multiple features

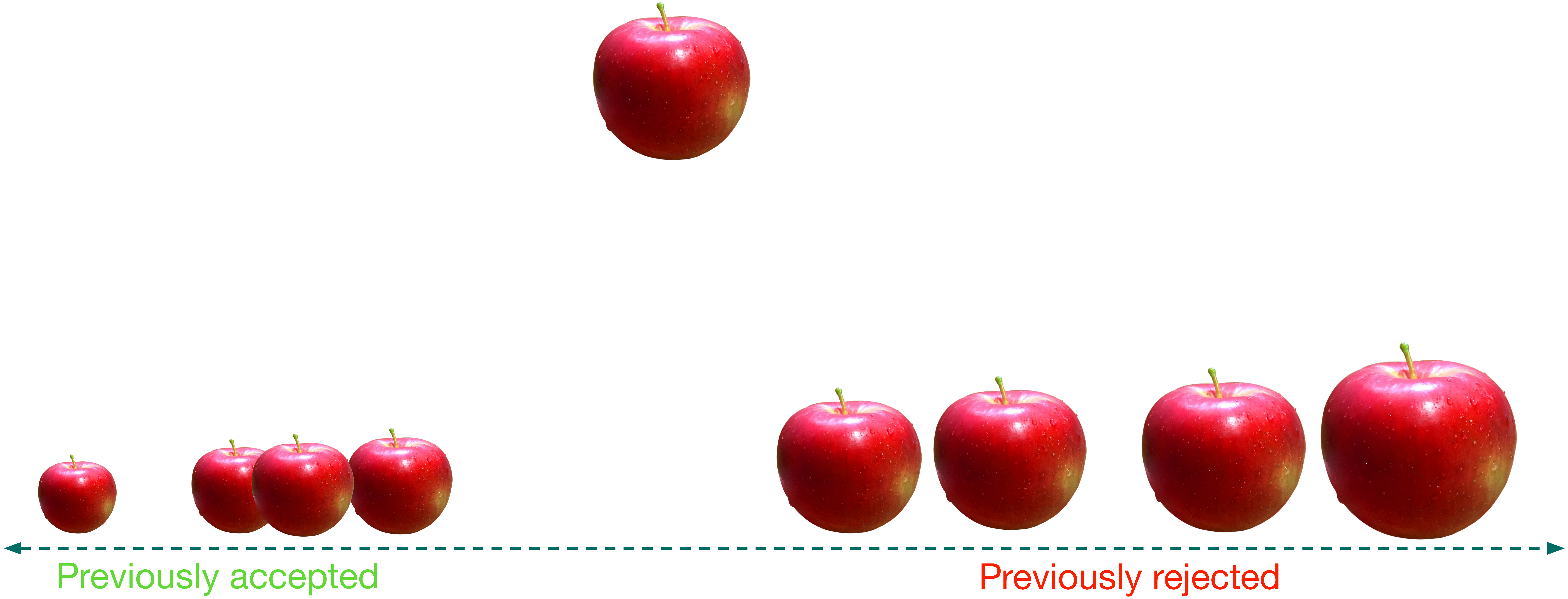
5 ML Models for Classification

1. Decision Tree
2. Random Forest
3. Support Vector Machine (SVM)
4. Naïve Bayes
5. K Nearest Neighbors (KNN)

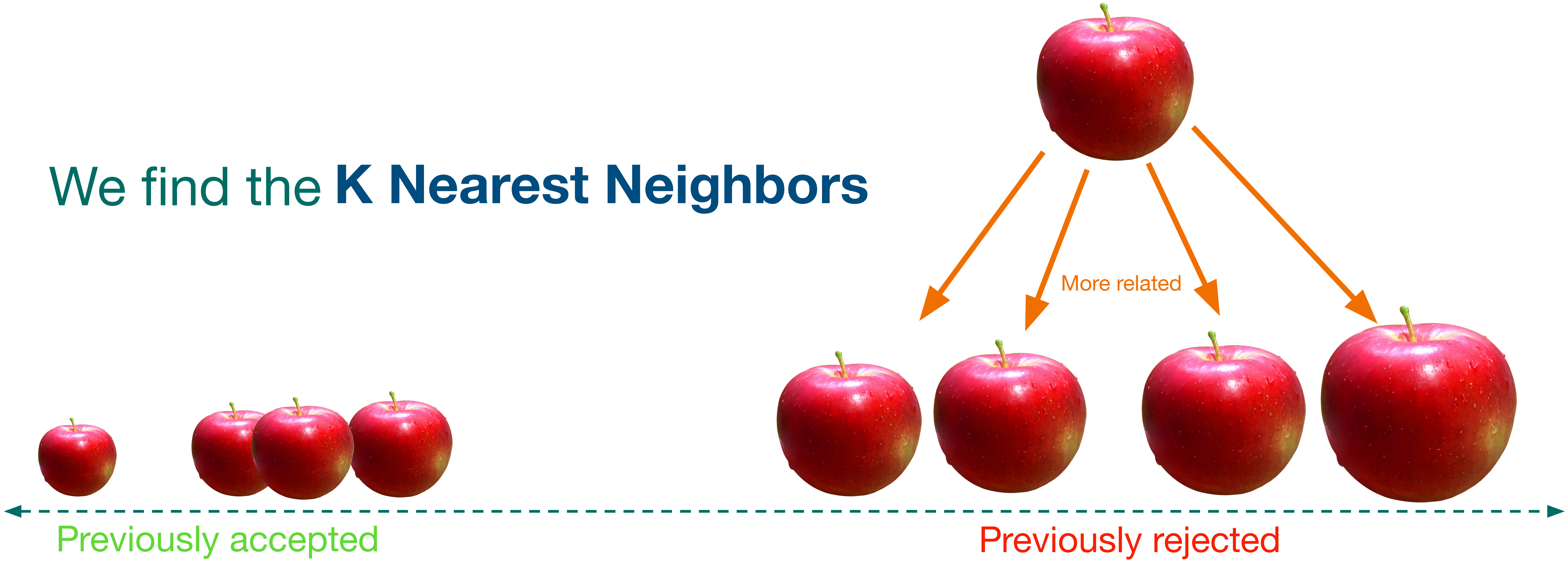
Buy? Don’t buy?

Did Buy

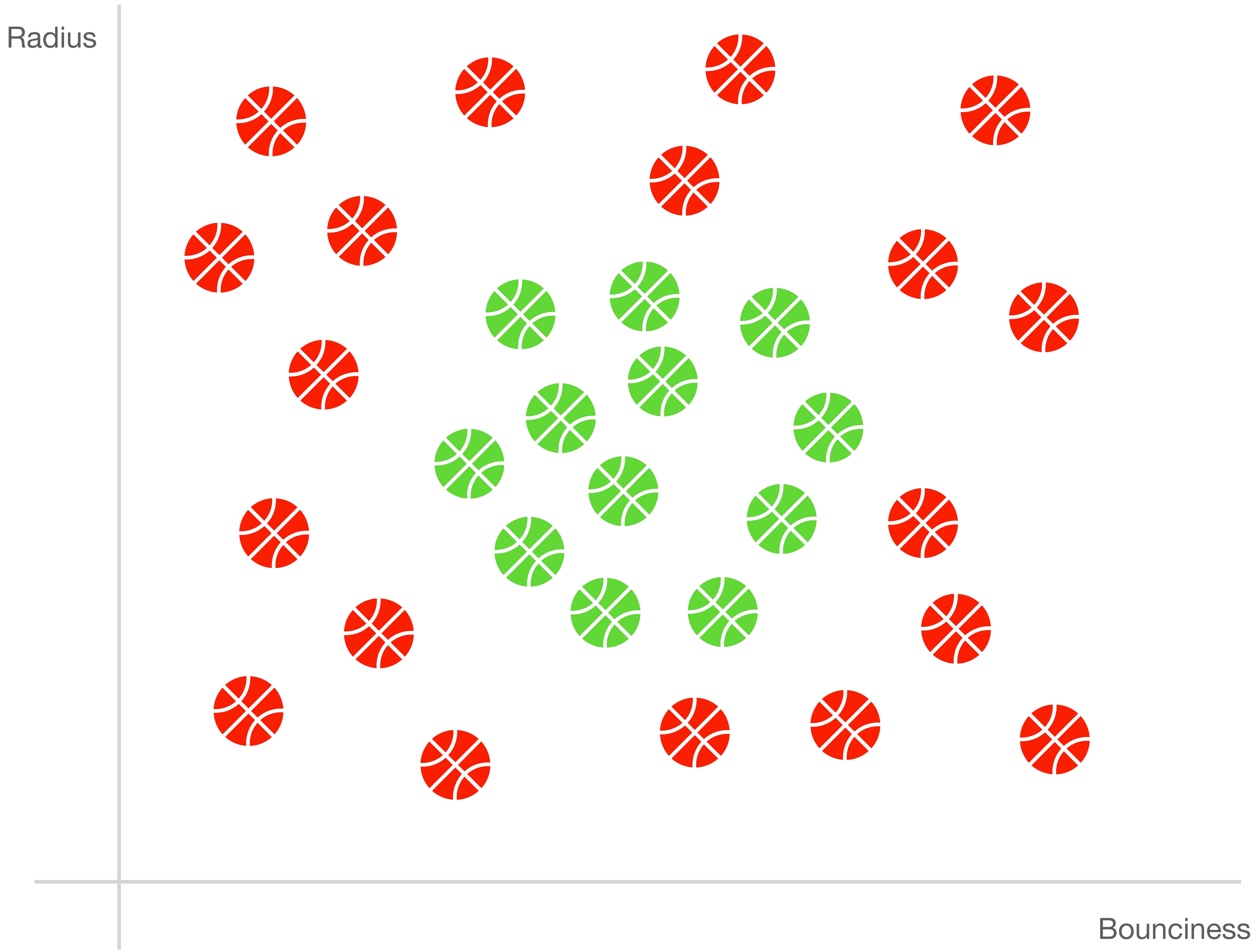
Rejected


Previously accepted
Previously rejected

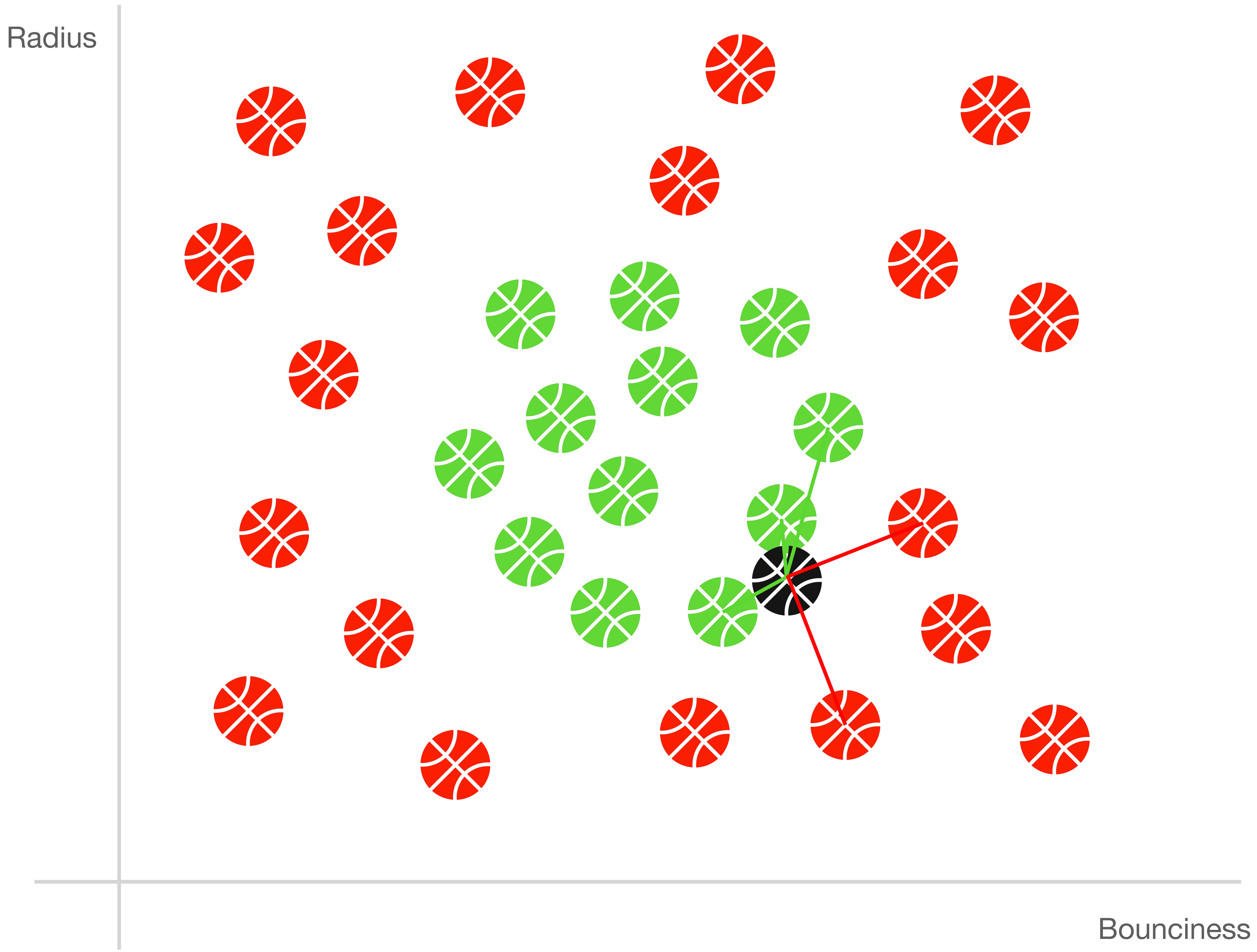
We find the **K Nearest Neighbors**



K Nearest Neighbors



K Nearest Neighbors

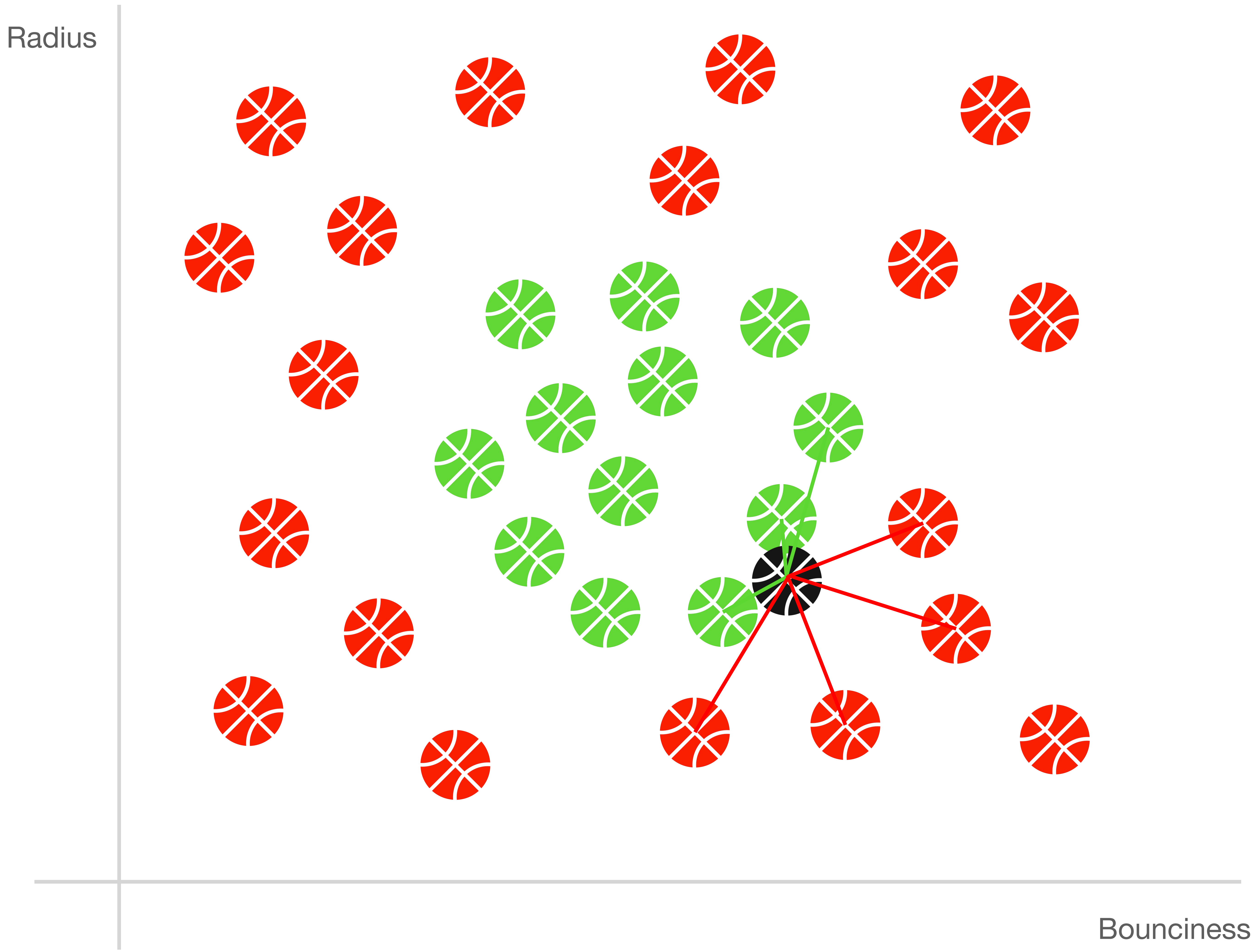


another **hyperparameter**

K=5

Buy

K Nearest Neighbors



another **hyperparameter**

K=7

Don't Buy

Five classifiers! That's a lot.

Let's get to the lab!