K NN
K-Nearest Neighbors

KNN

- While KNN can be used for regression tasks, its performance can be quite poor **and** less efficient than other algorithms, so we've decided not to exhibit its use for regression.
- However if you do want to use it for regression it is very easy to swap in the KNNRegressor model with scikit-learn.

KNN – Vs K mean

- You may have also heard of K means algorithm.
- K means is unrelated to KNN, be careful not to confuse the two due to their similar sounding names!

KNN

- ISLR Relevant Reading
  - Chapter 2
  - Formula 2.12 starts discussion on KNN for classification.

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

KNN

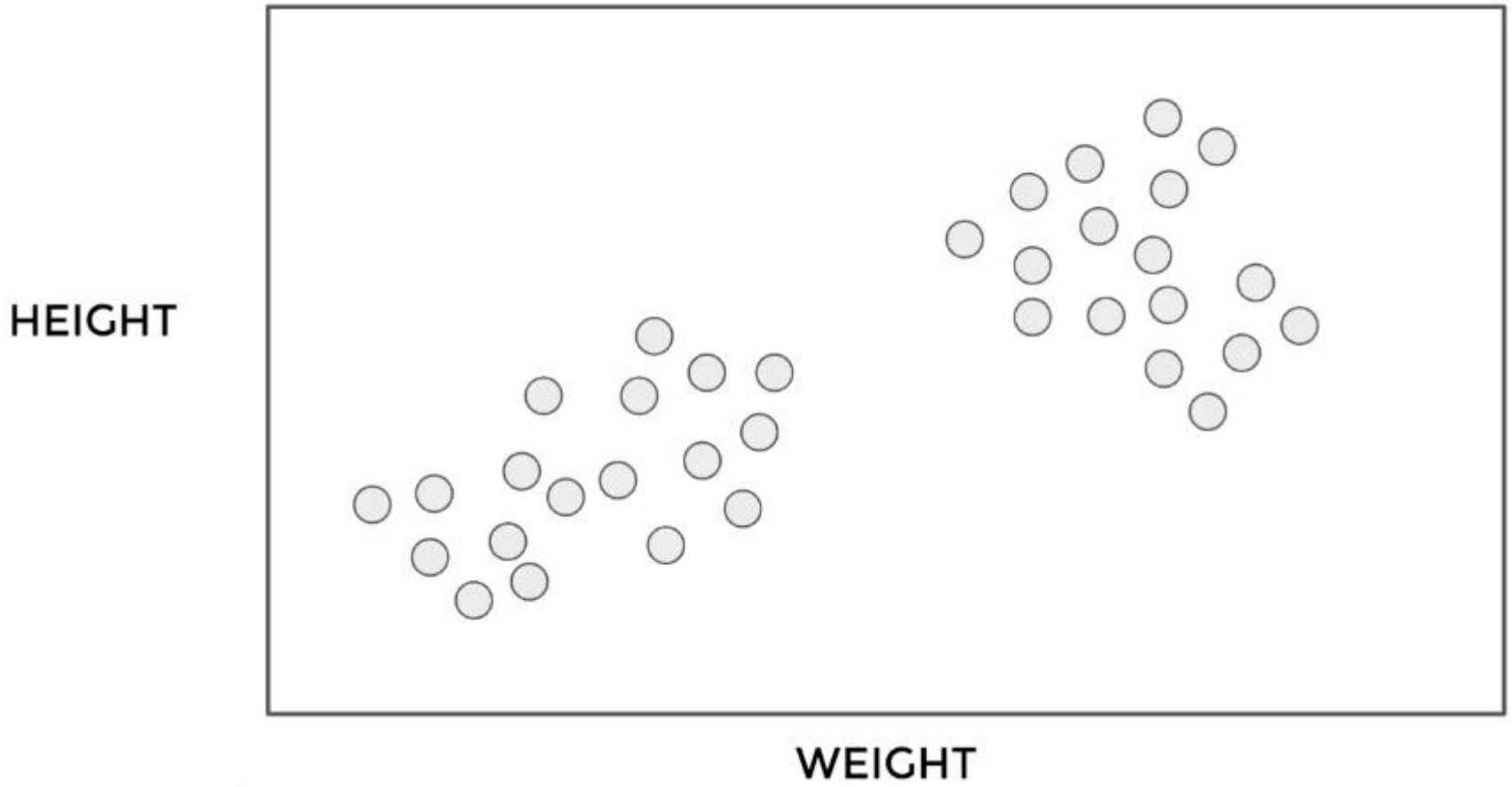
- K nearest neighbors is one of the simplest machine learning algorithms.
- It simply assigns a label to new data based on the **distance** between the old data and new data.
- Let's go through the intuition with an example use case...

KNN – by example

- Sexing chicks is still a very manual process:
  - en.wikipedia.org/wiki/Chick_sexing
  - Let's imagine we gathered a dataset of baby chick heights and weights.
  - How could we train an algorithm to identify the sex of a new baby chick based on historical features?

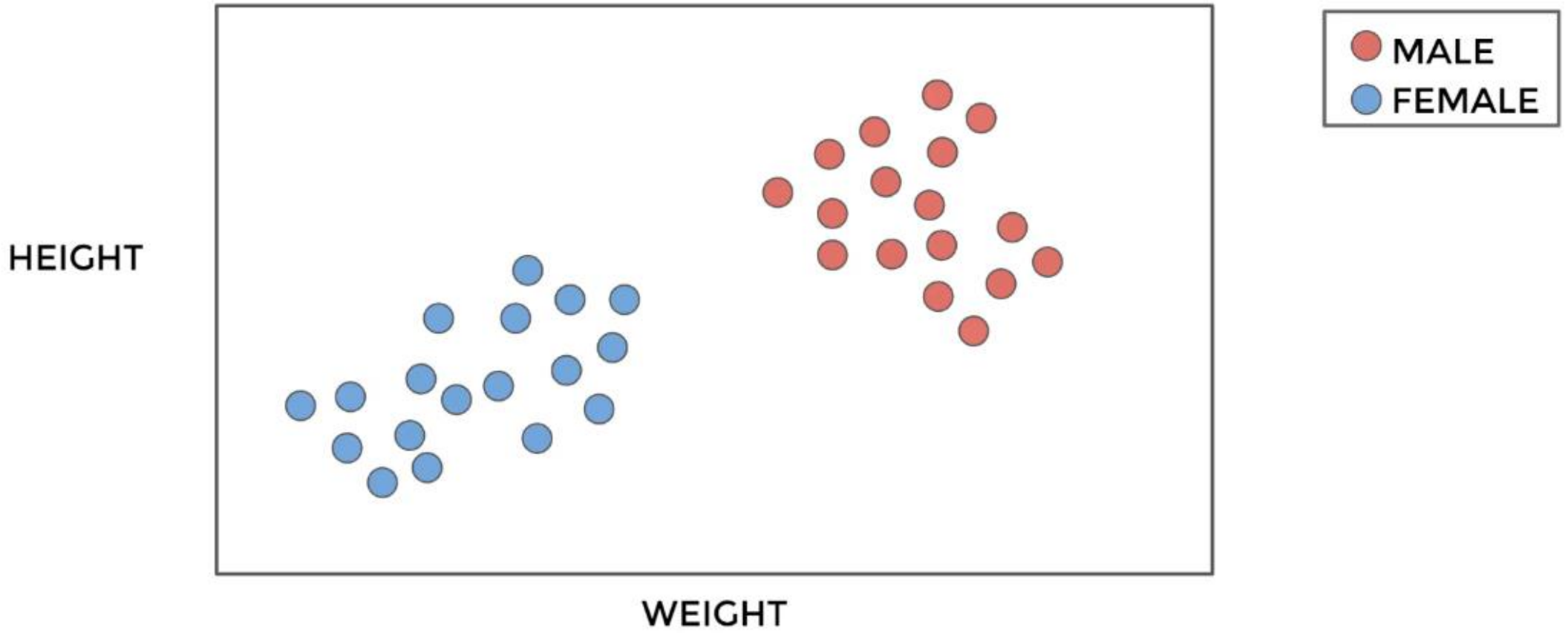
KNN – by example

- Imagine a height and weight data set



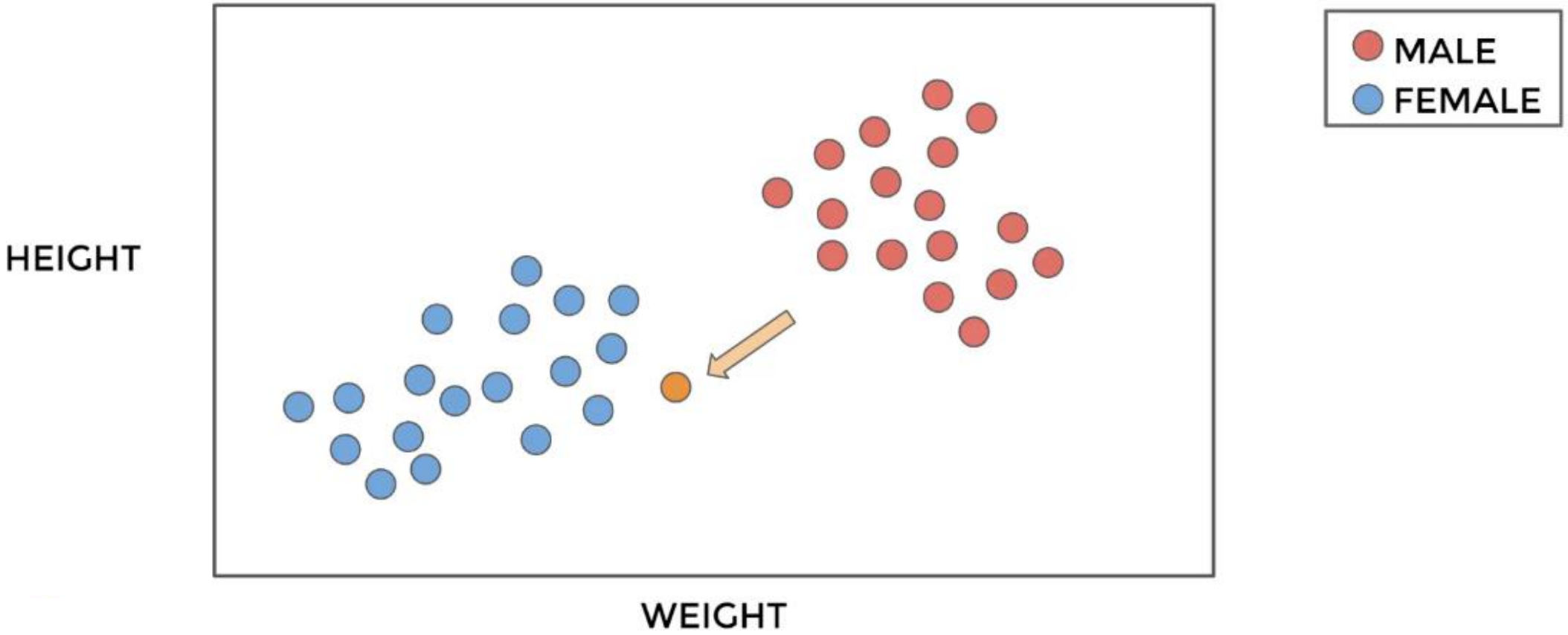
KNN – by example

- We historically know the sex of the chicks:



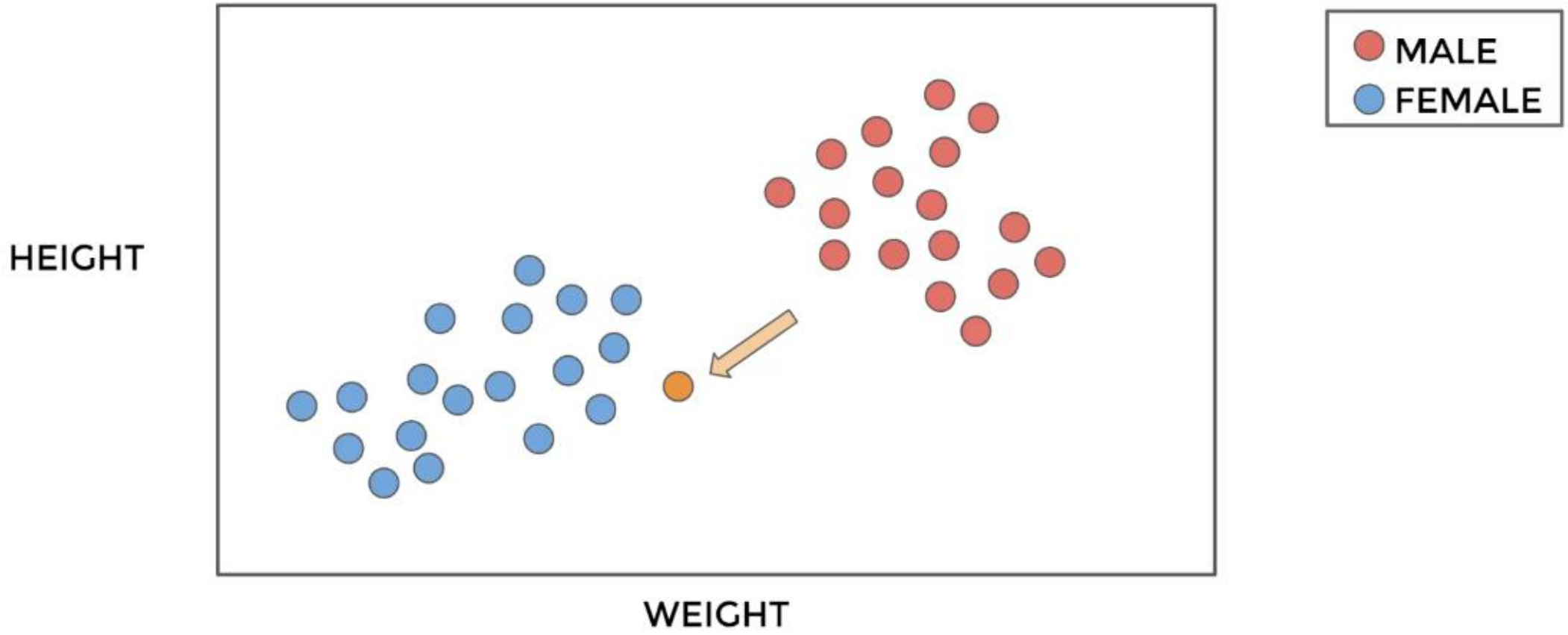
KNN – by example

- How would we assign sex to a new point?



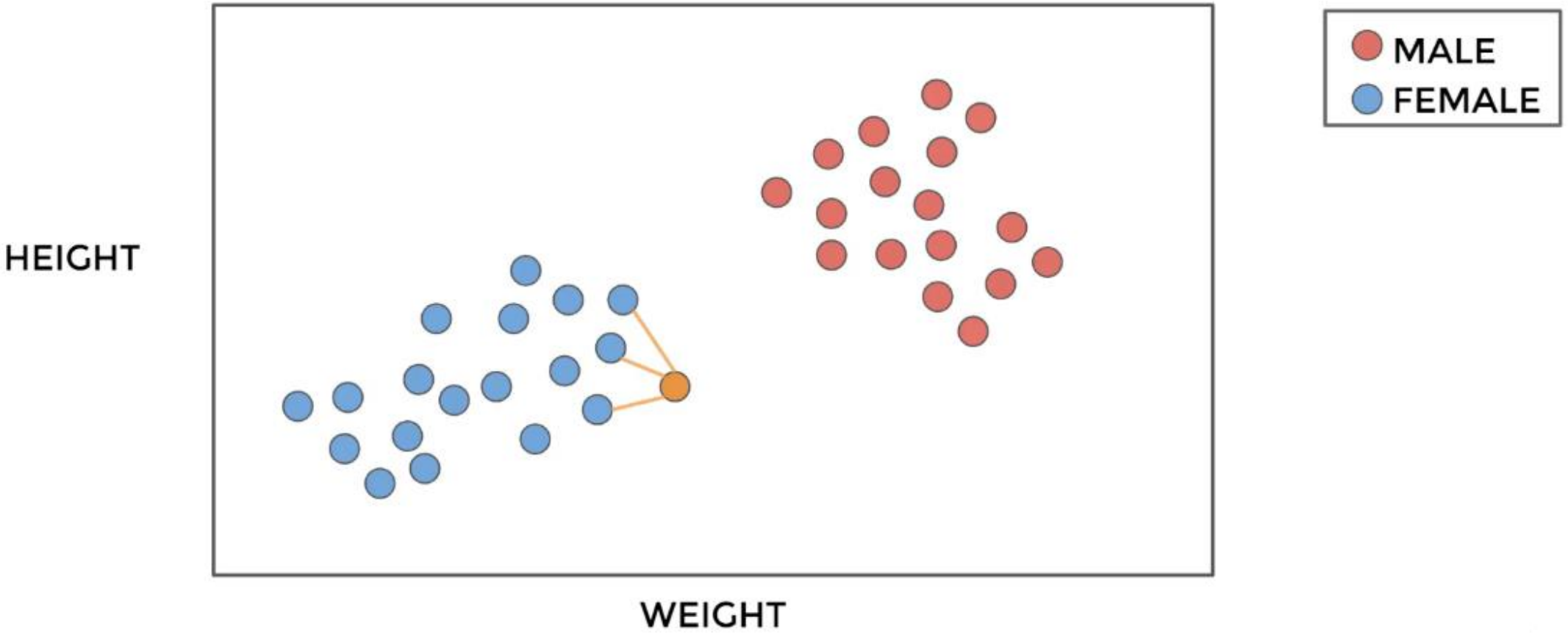
KNN – by example

- We intuitively “know” this is likely female.



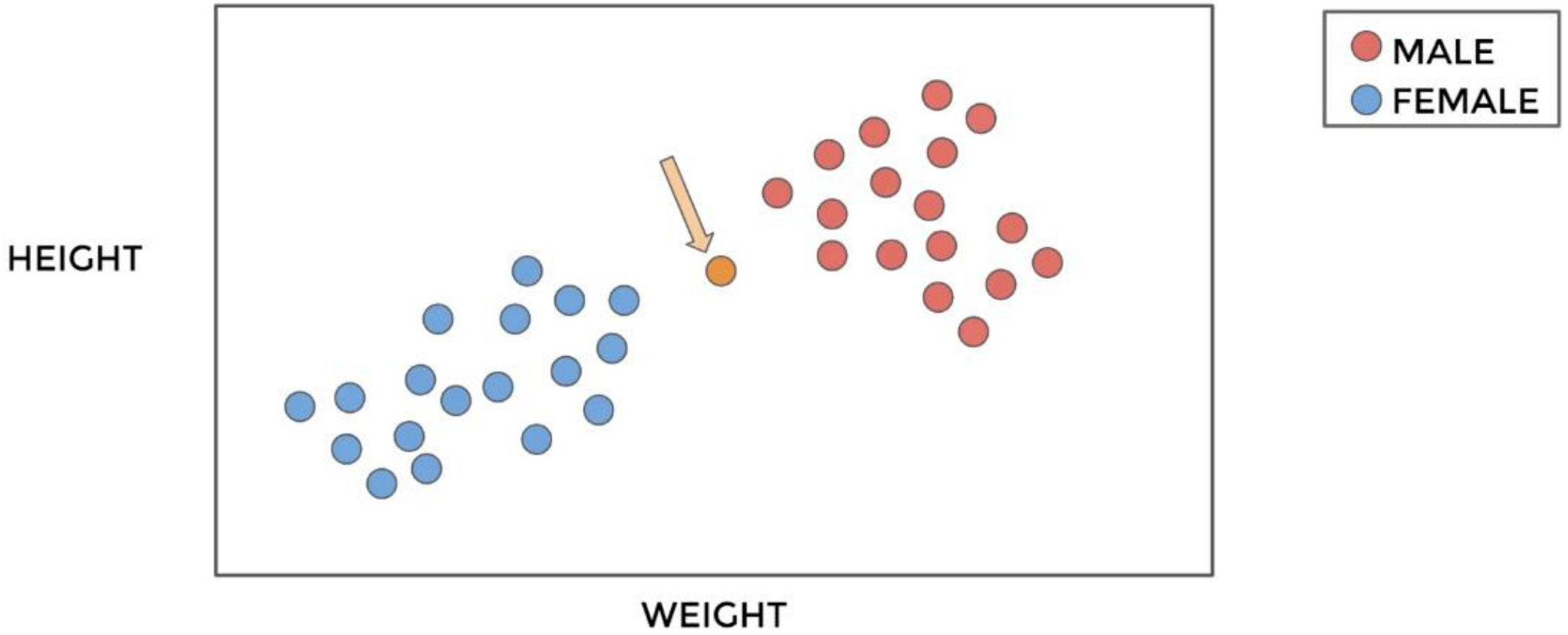
KNN – by example

- Intuition comes from **distance** to points!



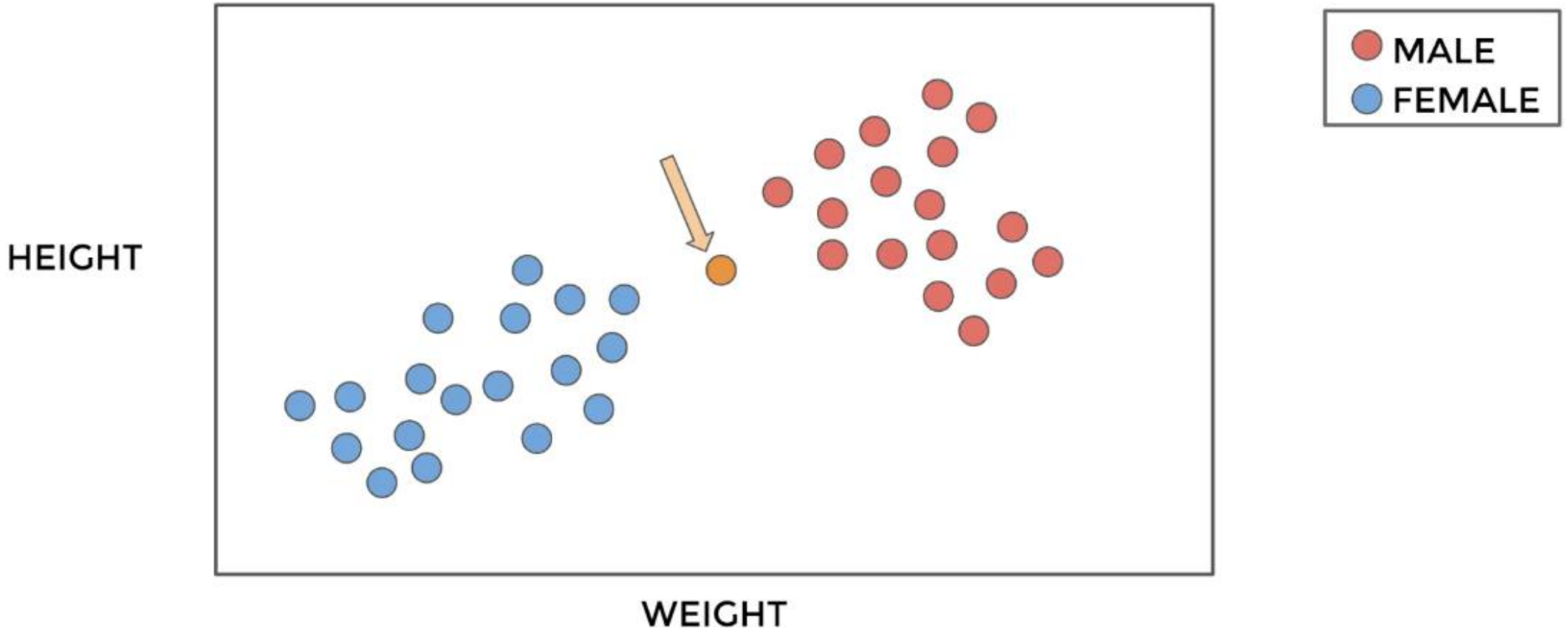
KNN – by example

- What about a less obvious point?



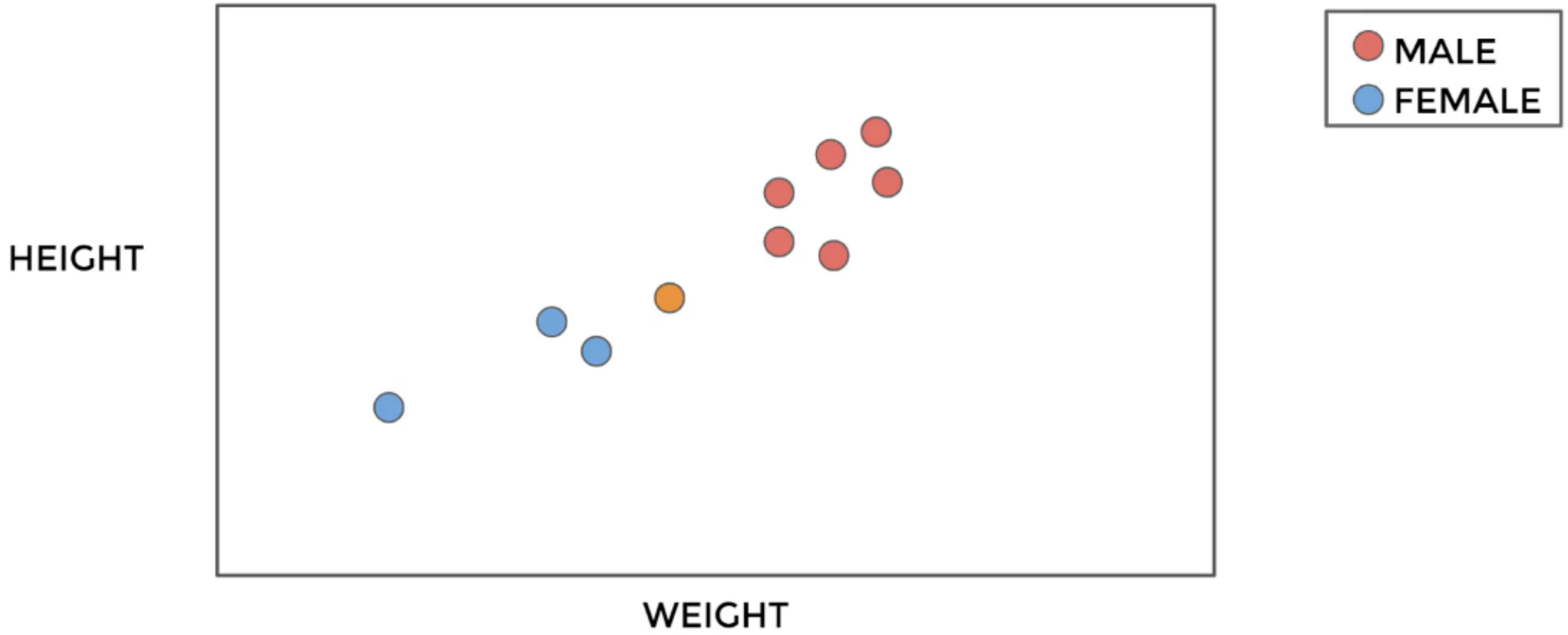
KNN – by example

- How many points to we consider?



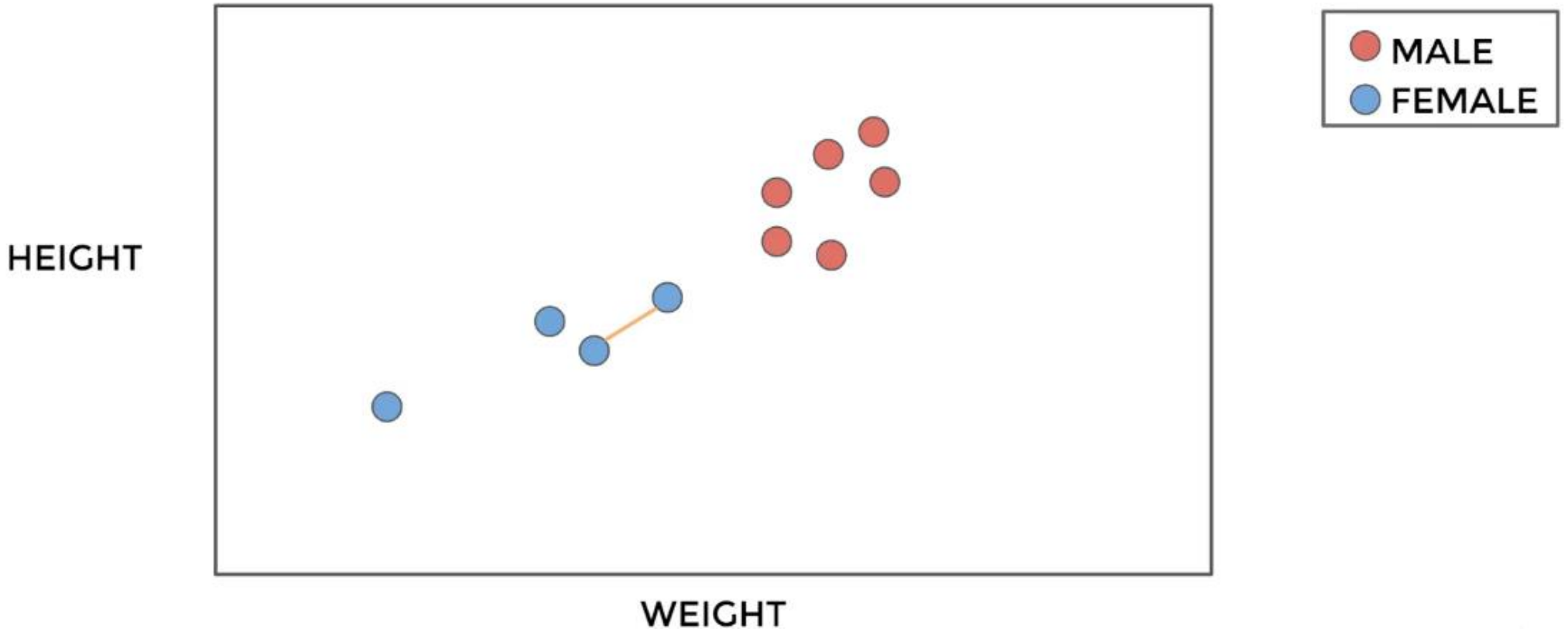
KNN – by example

- Let's imagine a situation like this:



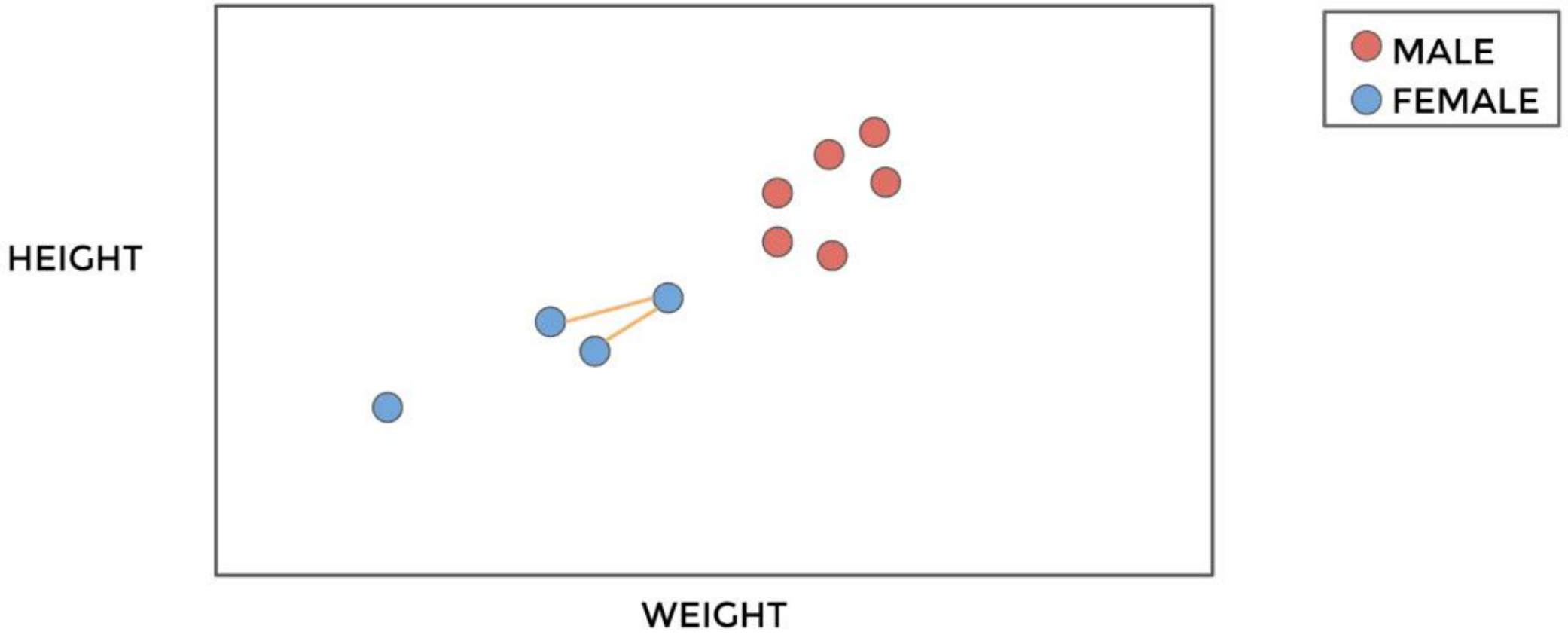
KNN – by example

- K=1



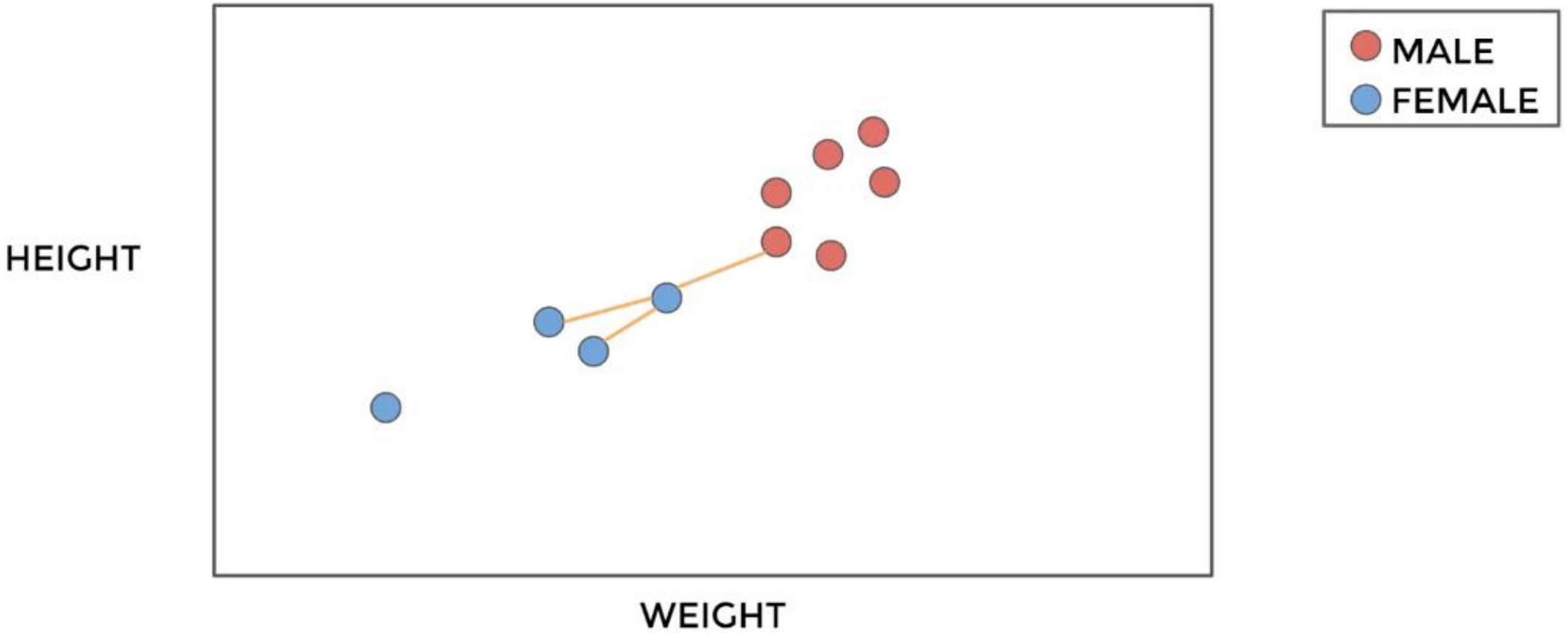
KNN – by example

- K=2



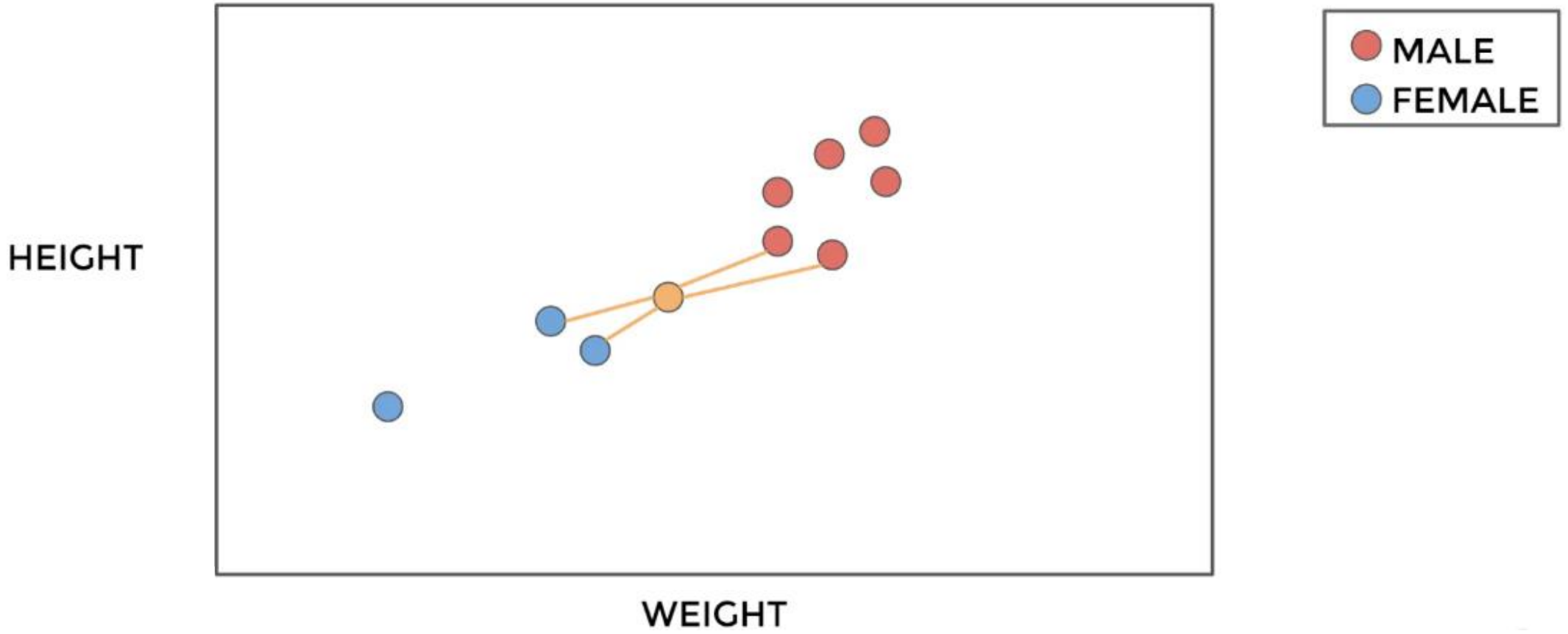
KNN – by example

- K=3



KNN – by example

- K=4 leads to a tie!



KNN – by example

- Tie considerations and options:
  - Always choose an odd K.
  - In case of tie,simply reduce K by 1 until tie is broken.
  - Randomly break tie.
  - Choose nearest class point.

KNN – by example

- **What does Scikit-Learn do in case of tie?**
  - *Warning: Regarding the Nearest Neighbors algorithms, if it is found that two neighbors, neighbor k+1 and k, have identical distances but different labels, the results will depend on the ordering of the training data.*

KNN – by example

- **What does Scikit-Learn do in case of tie?**
  - *In the case of ties, the answer will be the class that happens to appear first in the set of neighbors.*
  - *Results are ordered by distance, so it chooses the class of the closest point.*

KNN – by example

Notes

See Nearest Neighbors in the online documentation for a discussion of the choice of `algorithm` and `leaf_size`.

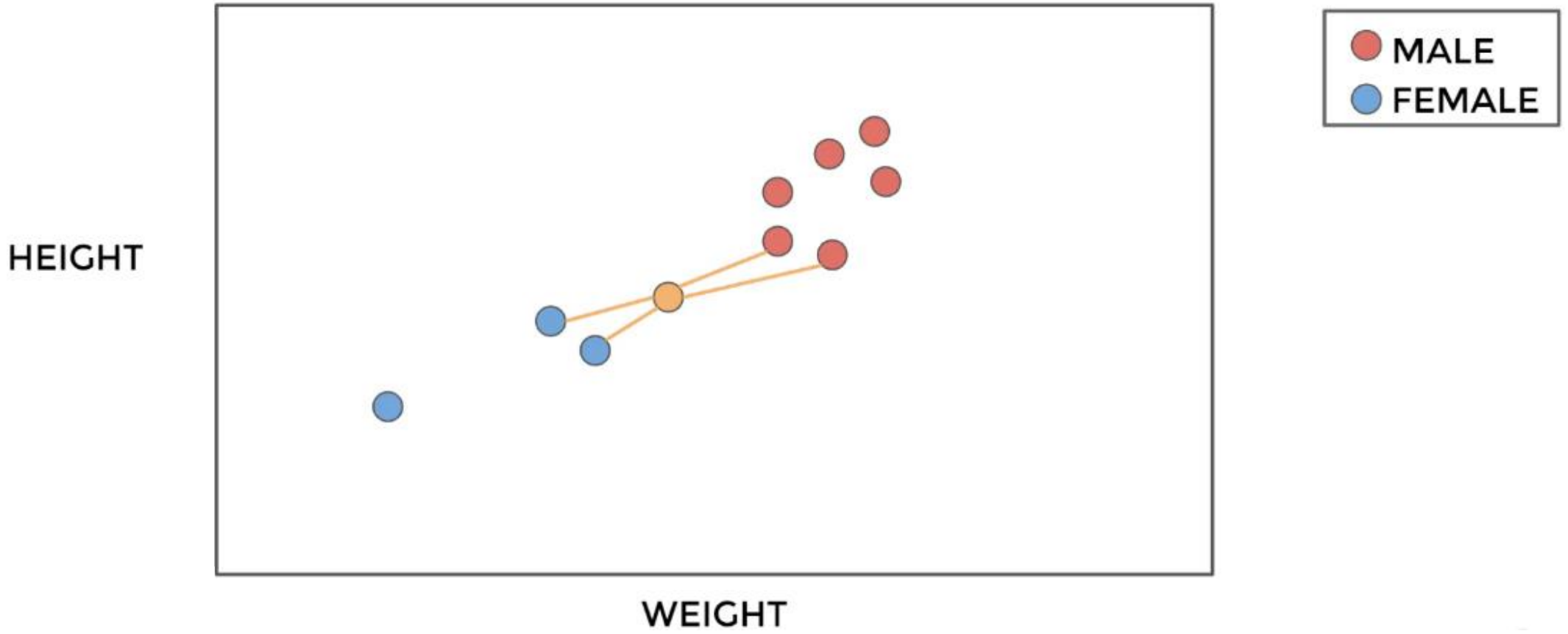
Warning: Regarding the Nearest Neighbors algorithms, if it is found that two neighbors, neighbor `k+1` and `k`, have identical distances but different labels, the results will depend on the ordering of the training data.

https://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

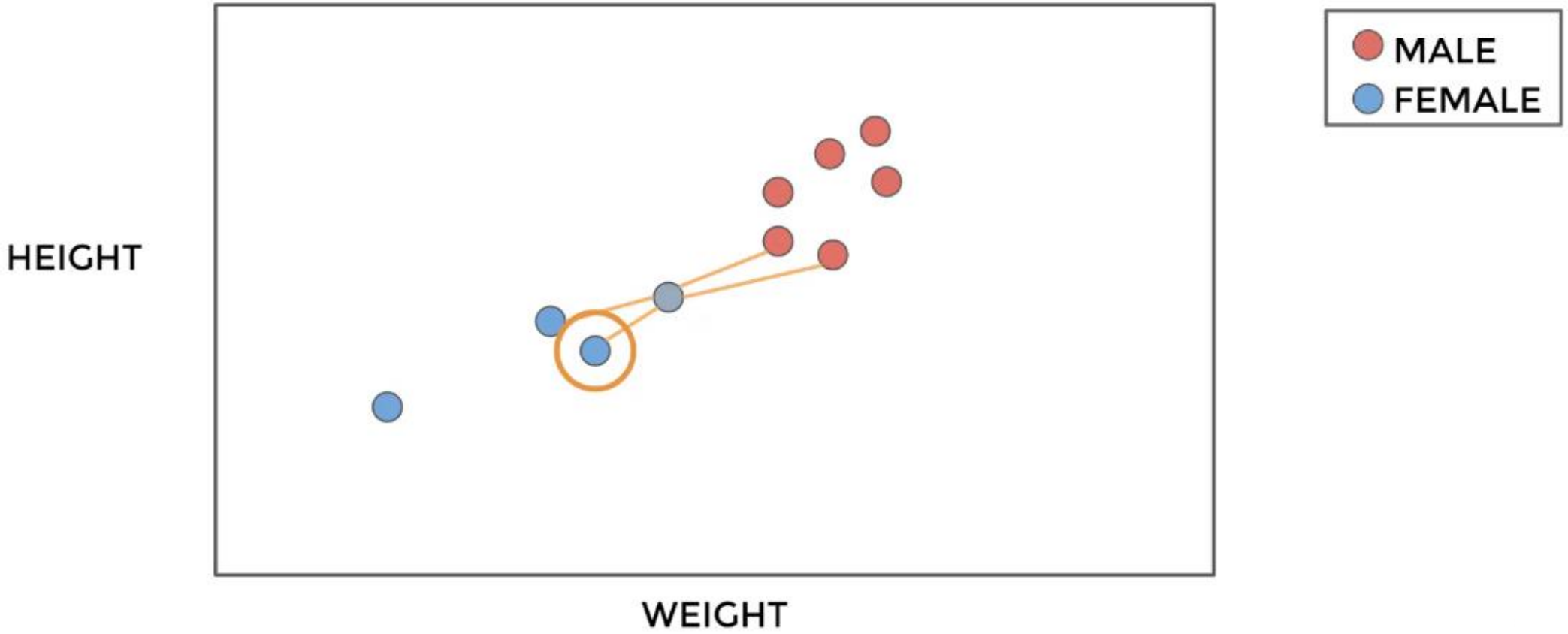
KNN – by example

- K=4 leads to a tie!



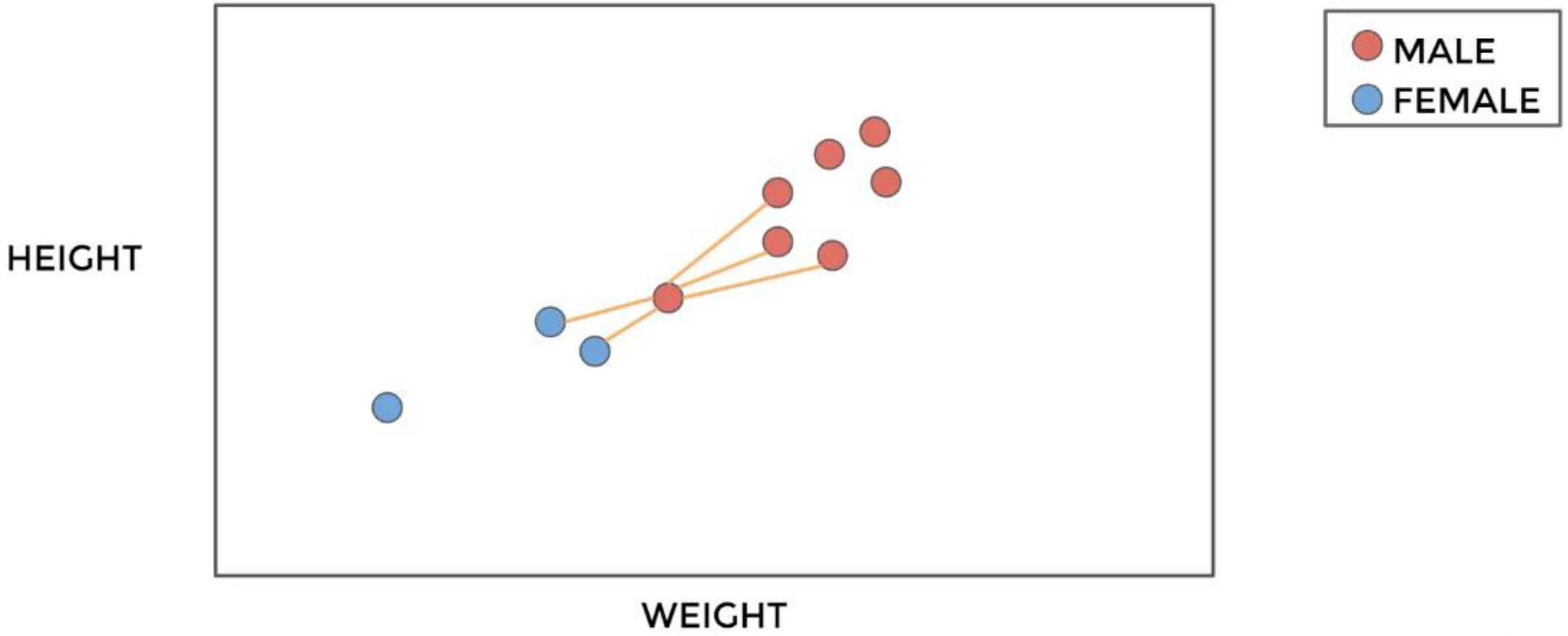
KNN – by example

- Choose closest K



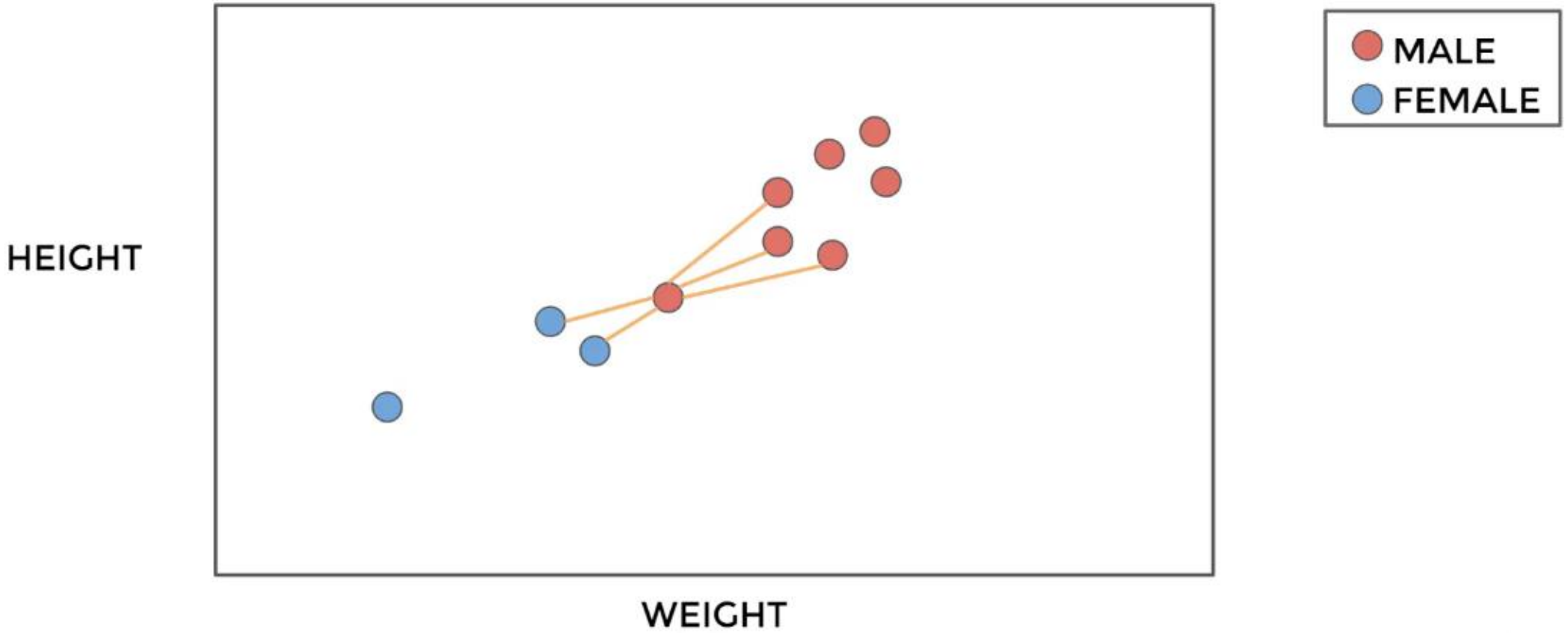
KNN – by example

- K=5 causes a switch from previous K values.



KNN – by example

- K=5 causes a switch from previous K values.



KNN – by example

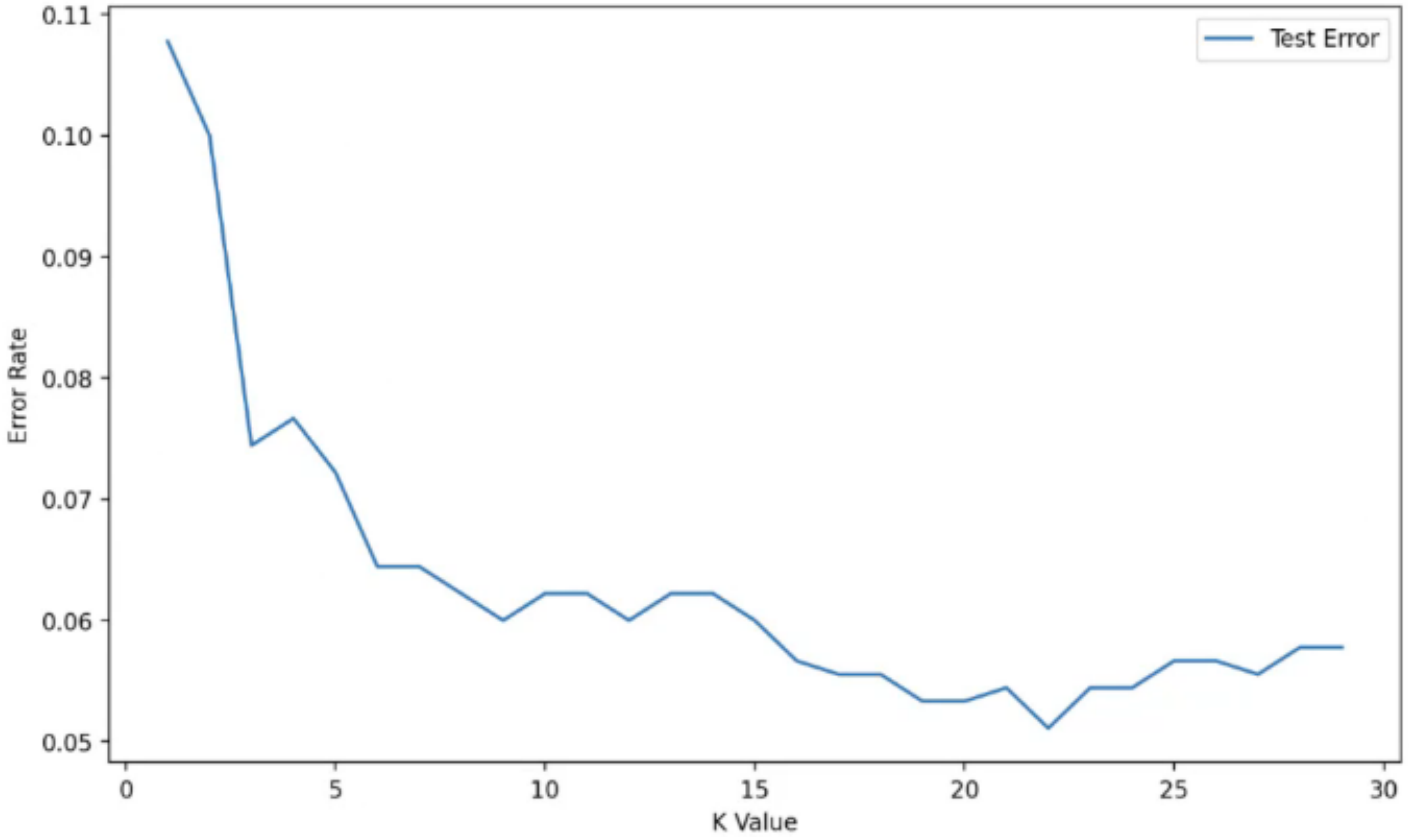
- How to choose best K value?

KNN – by example

- We want a K value that **minimizes** error:
  - Error = 1 - Accuracy
- Two methods:
  - Elbow method.
  - Cross validate a grid search of multiple K values and choose K that results in lowest error or highest accuracy.

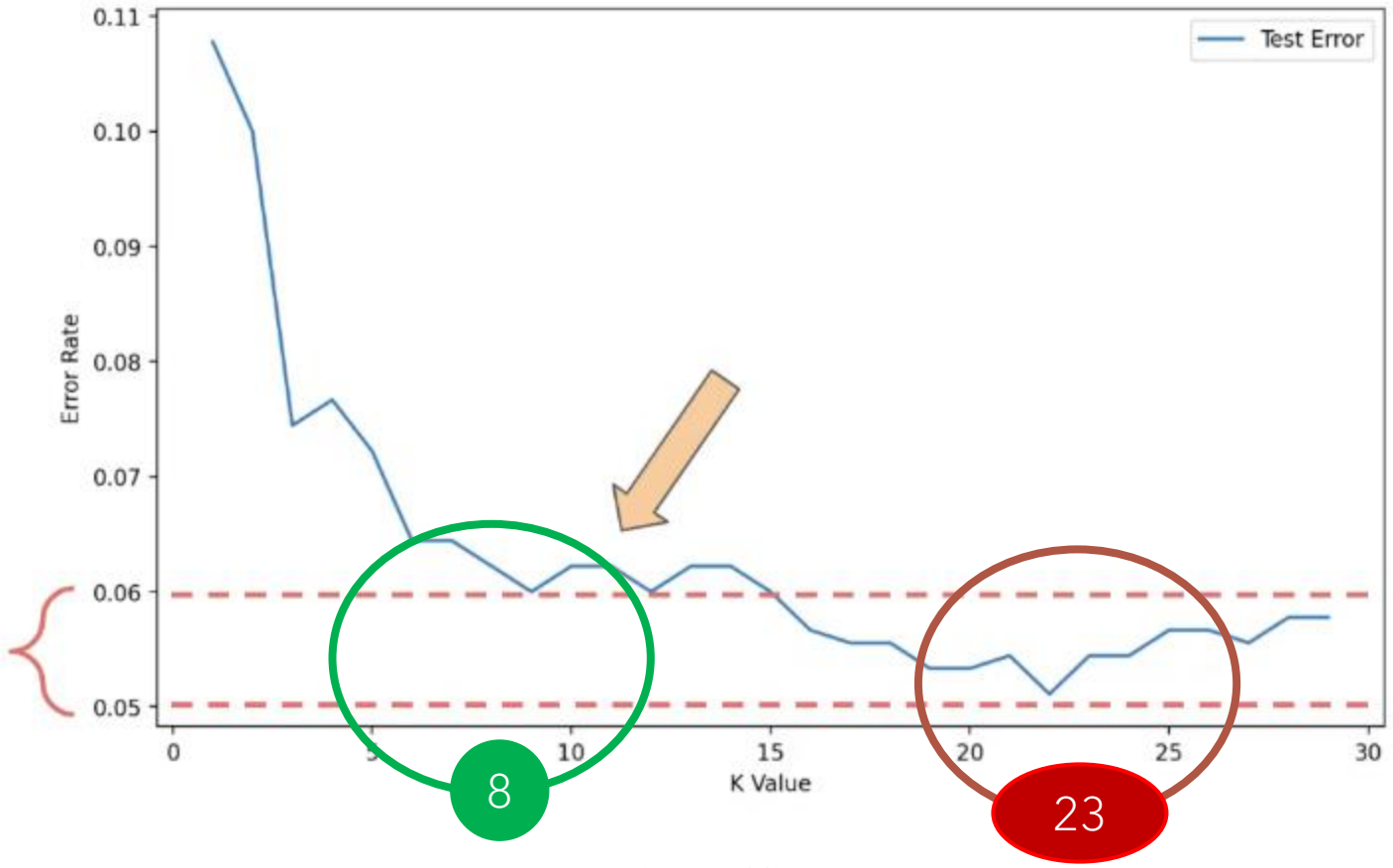
KNN – Elbow Method

- Elbow method:



KNN – Elbow Method

- Elbow method:



KNN – Elbow Method

- Cross validation only takes into account the K value with the lowest error rate across multiple folds.
- This could result in a more complex model (higher value of K).
- Consider the context of the problem to decide if larger K values are an issue.

KNN

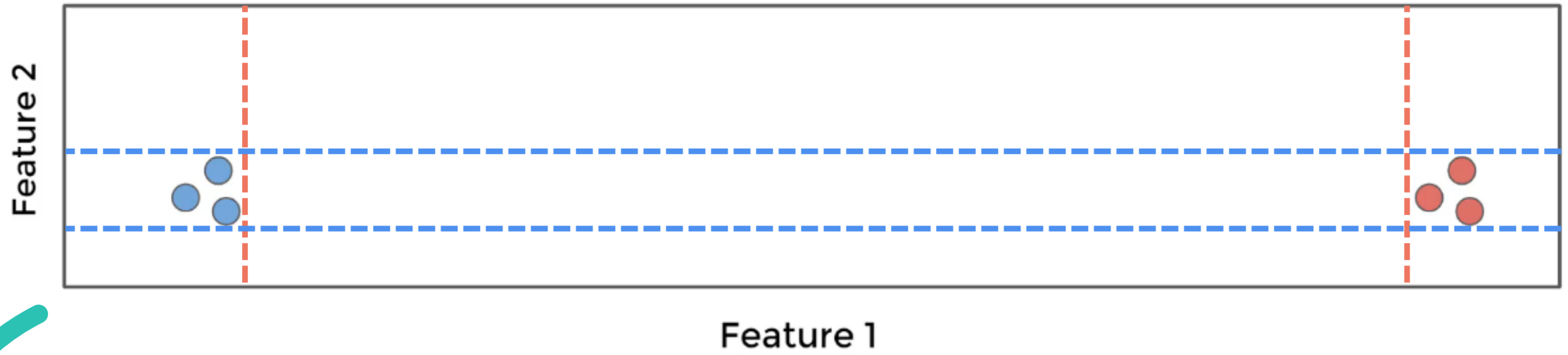
- KNN Algorithm
  - Choose K value.
  - Sort feature vectors (N dimensional space) by distance metric.
  - Choose class based on K nearest feature vectors.

KNN – Distance Metrics

- KNN Considerations:
  - Distance Metric
    - Many ways to measure distance:
      - Minkowski
      - Euclidean
      - Manhattan
      - Chebyshev

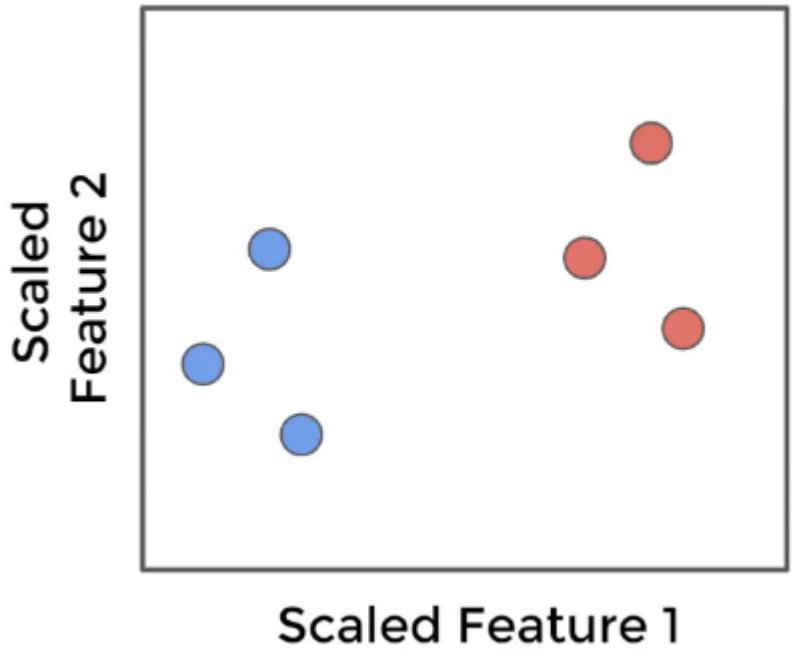
KNN – Scaling Distance

- KNN Considerations:
  - Scaling for Distance
    - Features could have vastly different value ranges!



KNN – Scaling Distance

- KNN Considerations:
  - Scaling is necessary for KNN.



KNN – Scaling Distance

- While the KNN Algorithm is relatively simple, keep in mind the following considerations:
  - Choosing the optimal K value.
  - Scaling features.
  - Let's continue to explore how to perform KNN for classification!

KNN

- Let’s See how we create the data model , LoR model and calculate the metrics using SKLearn