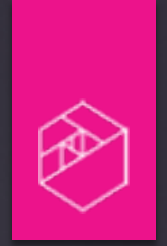
Intro to Classification & *k*-Nearest Neighbors

REVIEW:
What is Supervised Learning?

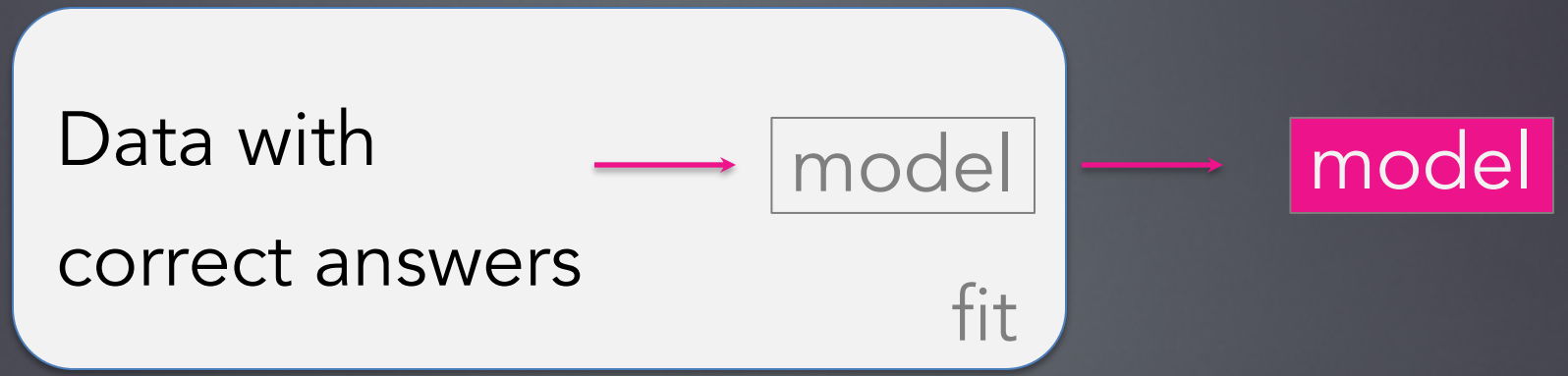
Supervised Learning



- [x] learn model from training data
- [x] make predictions about unseen/future data
- [x] supervised = labeled data available
- [x] e.g. price of house, emails spam/not spam, etc.

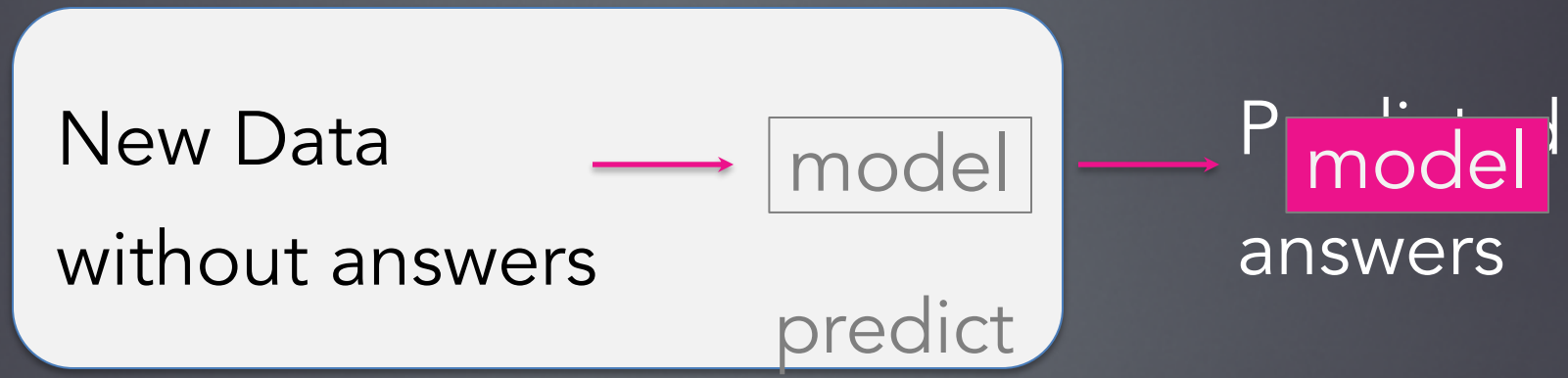
Supervised Learning

Use the labeled data to fit a machine learning model



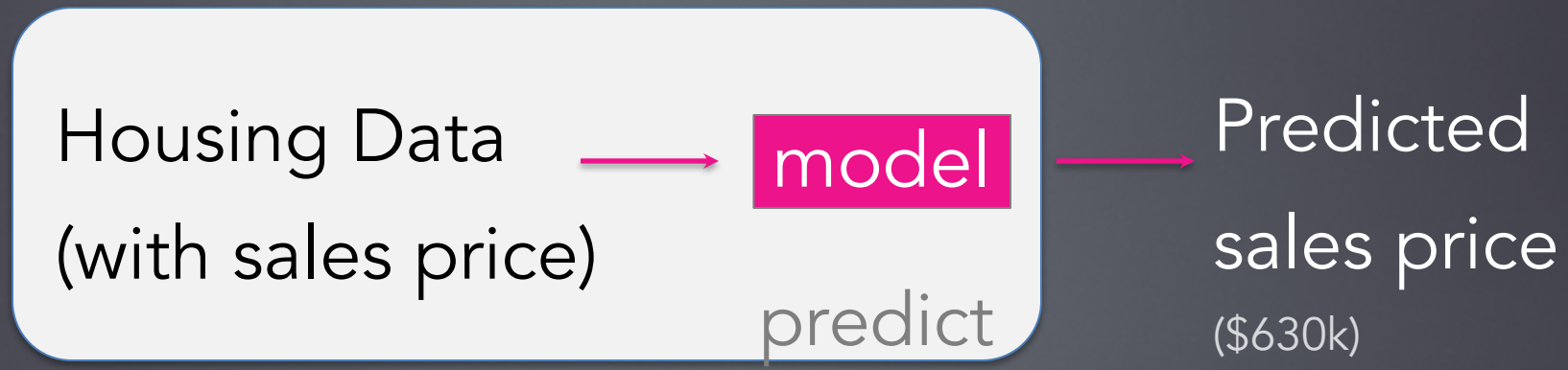
Supervised Learning

Once our model has been trained (and tested!), we can use it on unlabeled data to make new predictions.



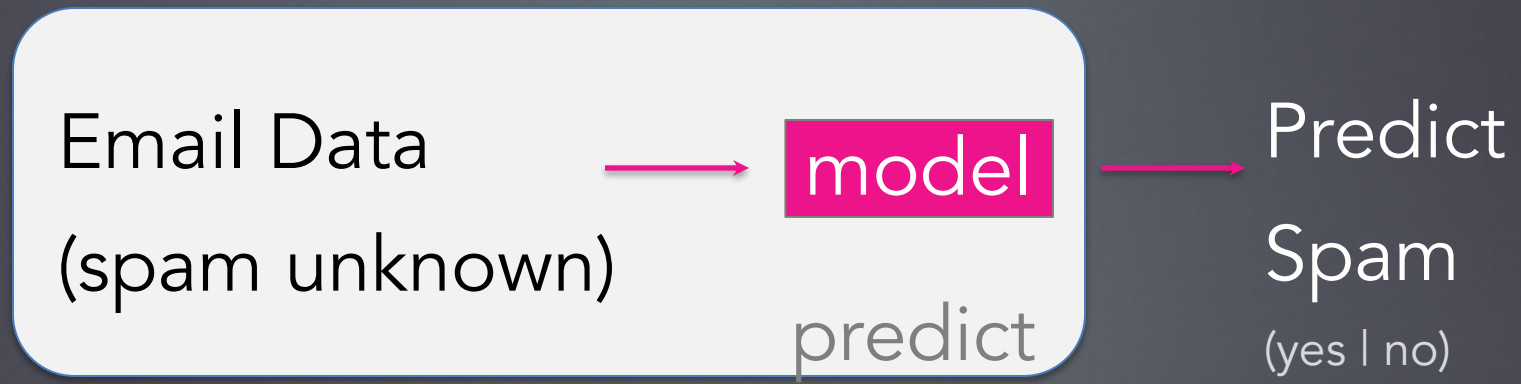
Supervised Learning

Regression: "Answers" from model are numeric

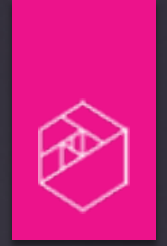


Supervised Learning

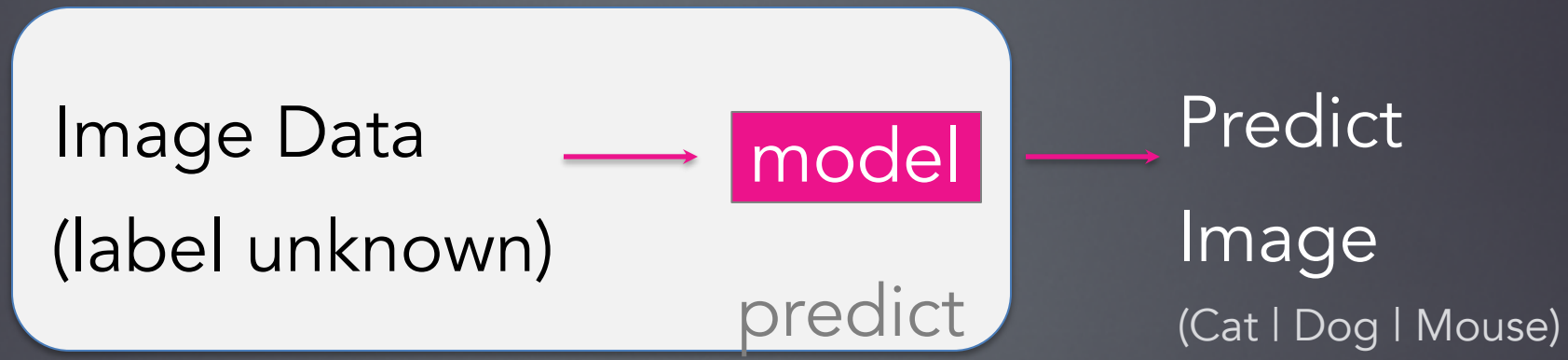
Classification: “Answers” from model are categories



Supervised Learning

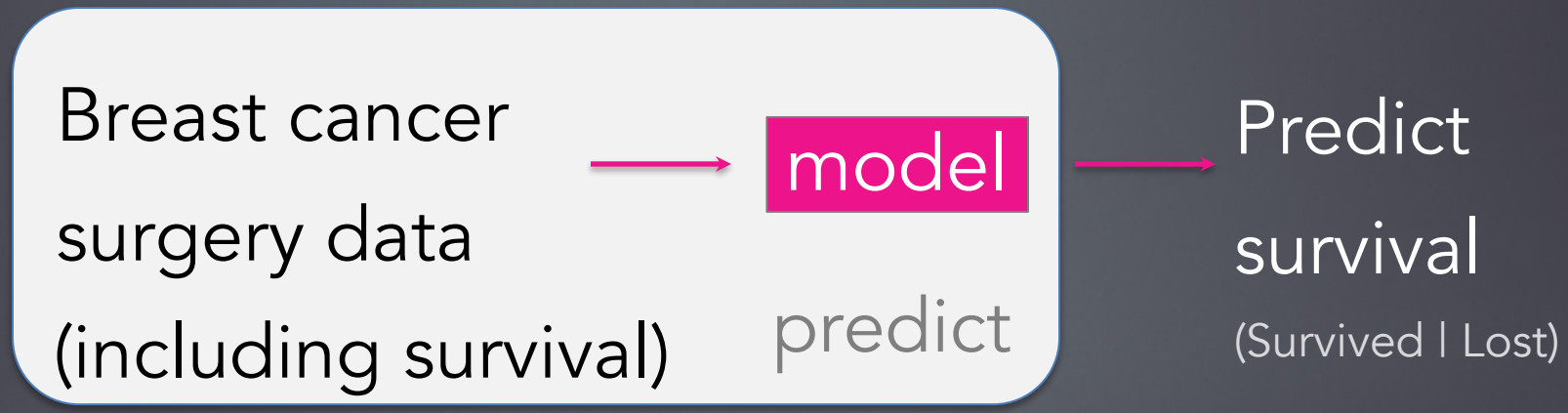


Classification: “Answers” from model are categories

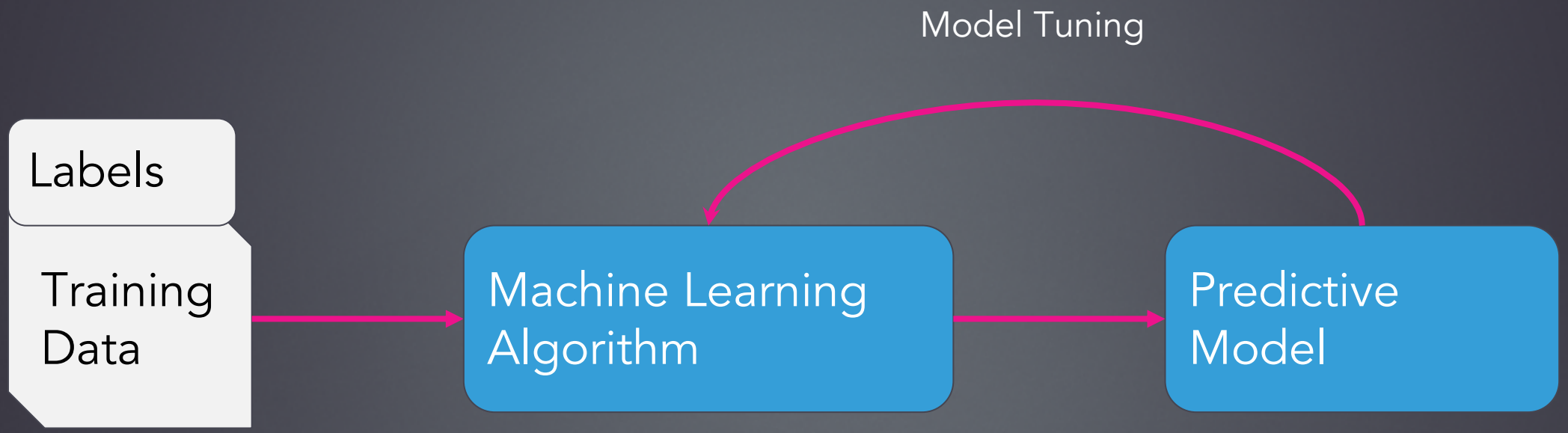


Supervised Learning

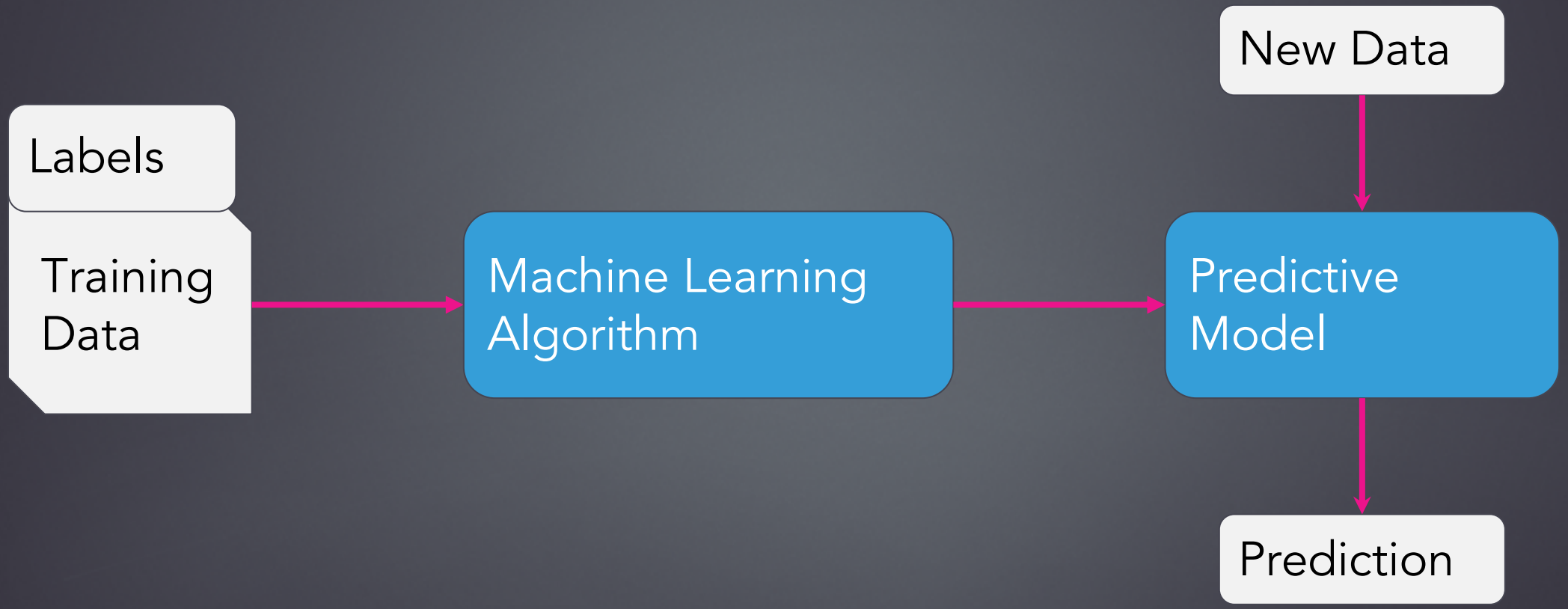
Classification: “Answers” from model are categories



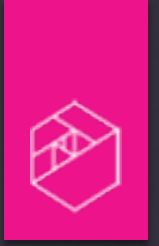
Supervised Learning - Train



Supervised Learning - Predict



Supervised Learning - Vernacular

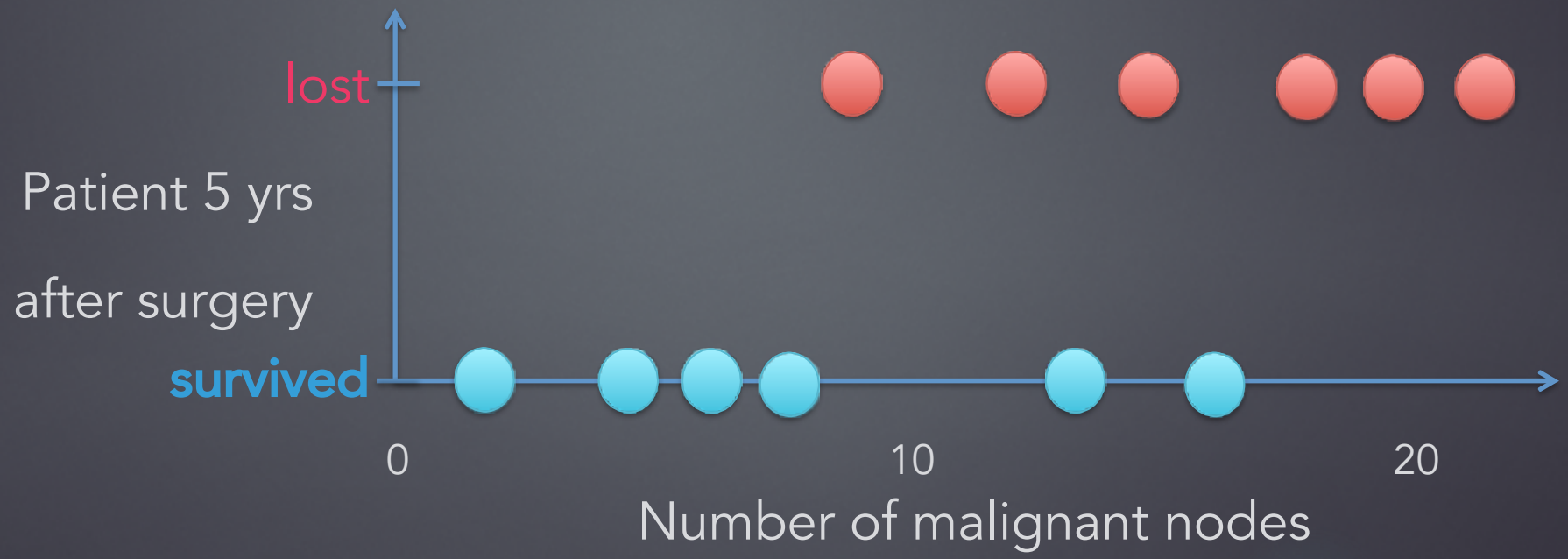


Observation	each data point	one row
Target	predicted property	column to predict
Label	target/category of observation	value of target column
Feature	property of observation, used for prediction	non-target columns

Supervised Learning - Vernacular

1 Feature. Number of malignant nodes

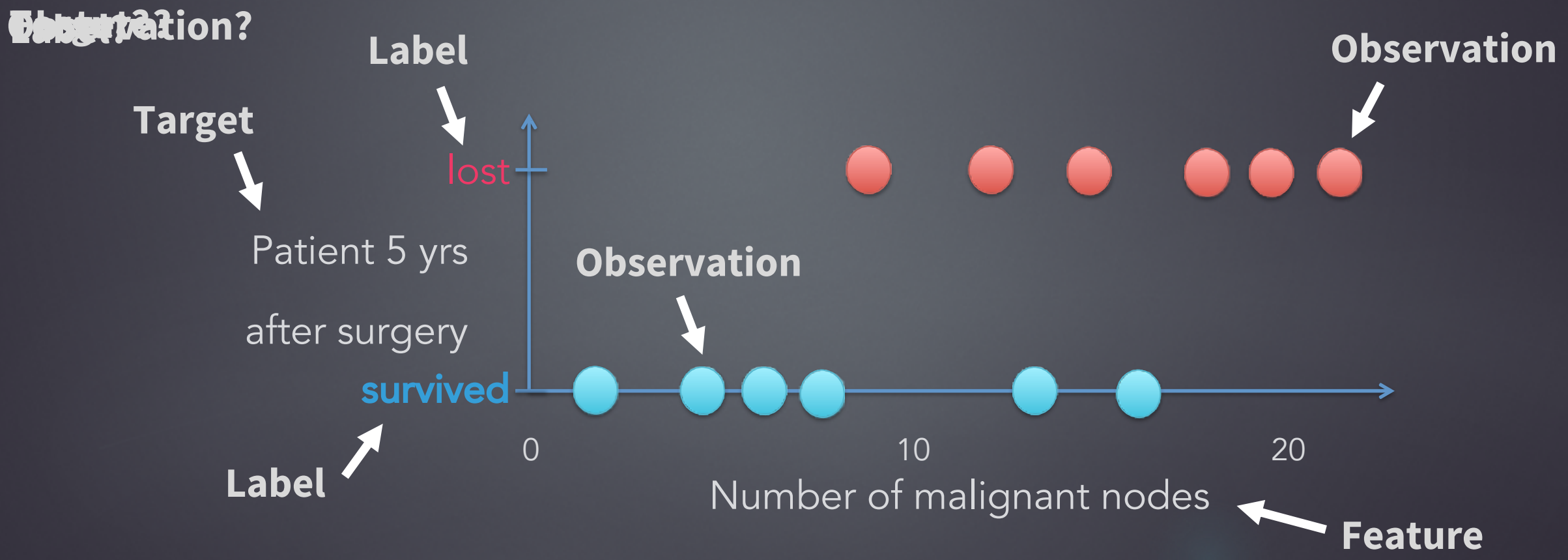
2 Labels. Survived / Lost



Supervised Learning - Vernacular

1 Feature. Number of malignant nodes

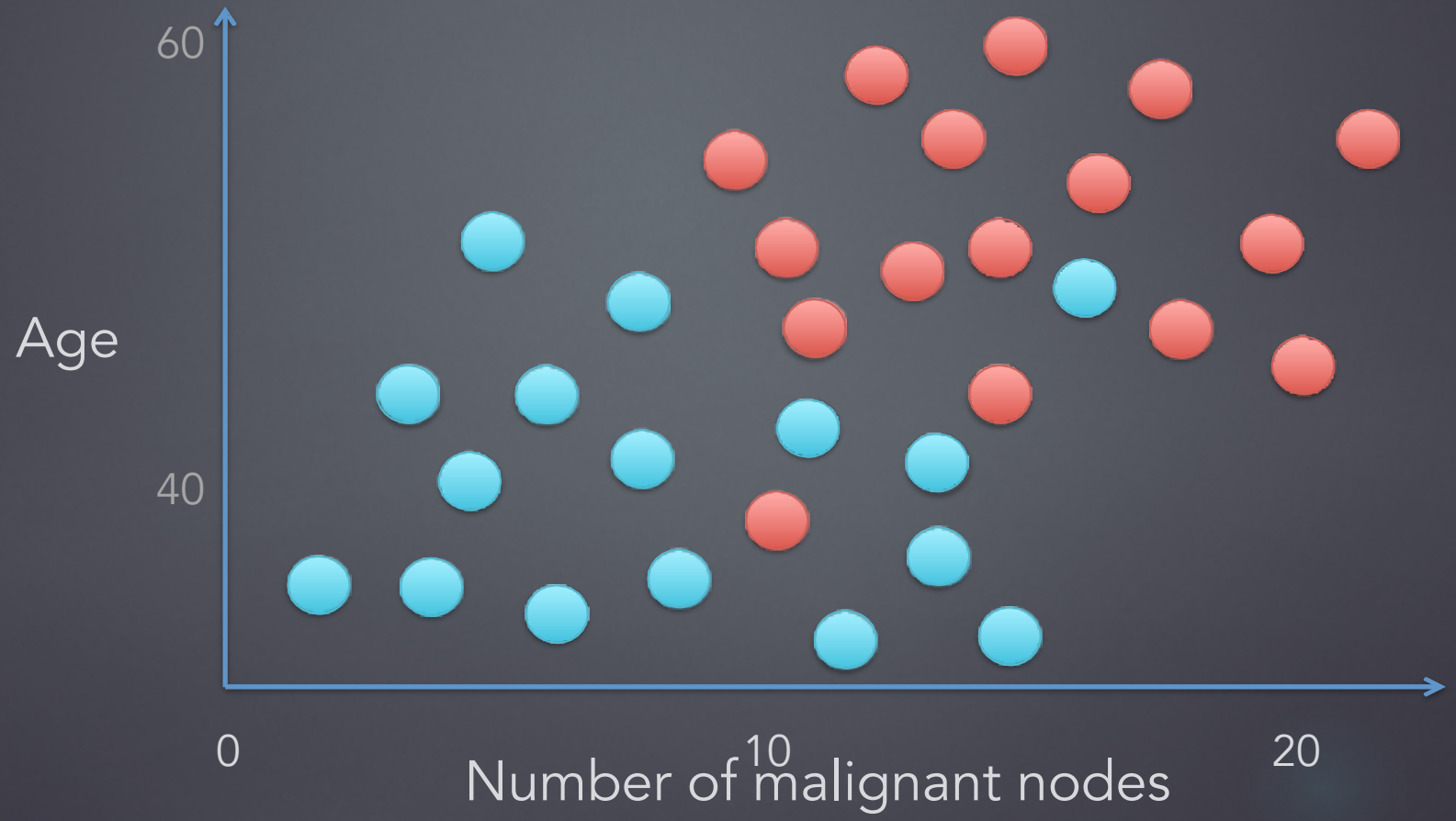
2 Labels. Survived / Lost



Supervised Learning - Example

2 Features. No of malignant nodes / Age

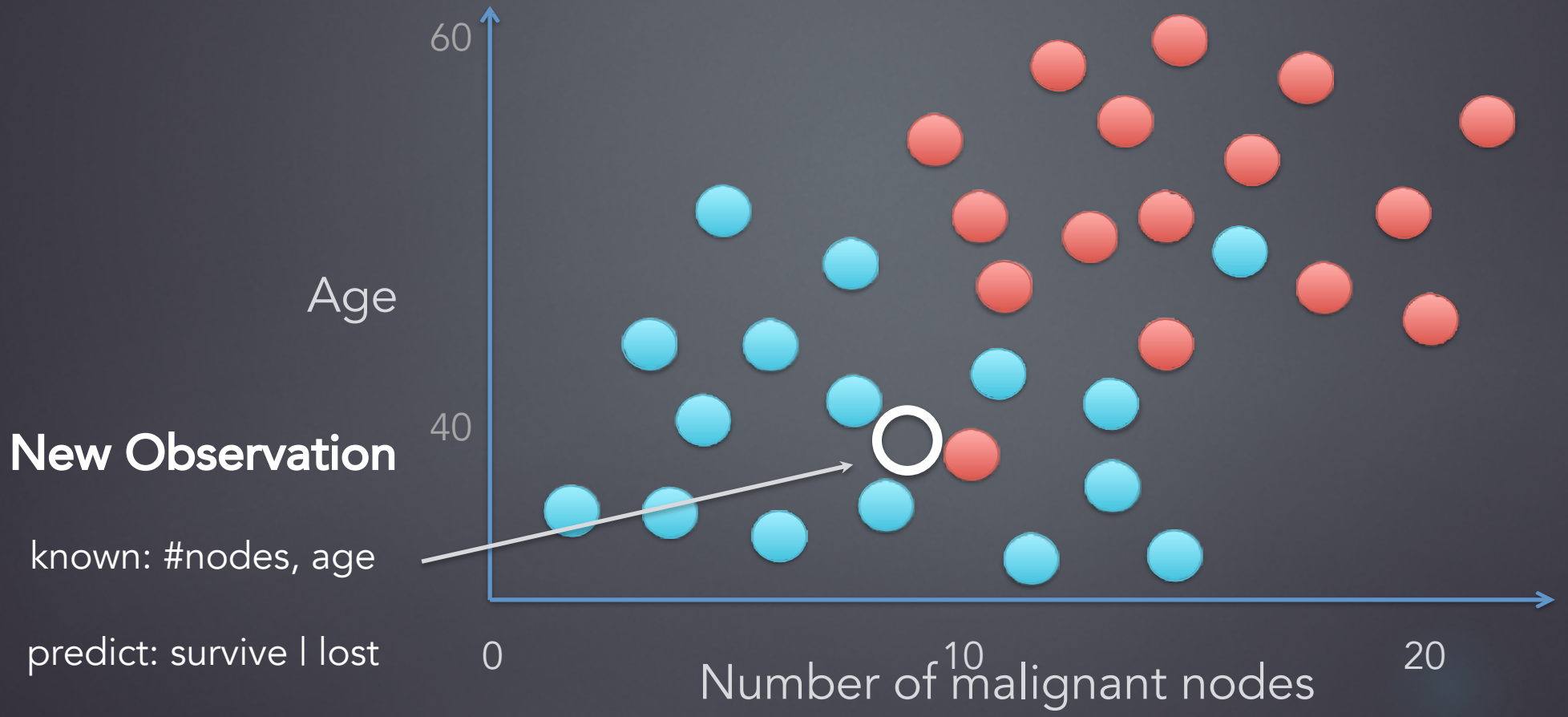
2 Labels. Survived / Lost



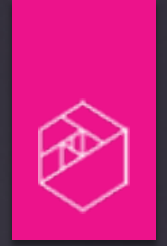
Supervised Learning - Example

2 Features. No of malignant nodes / Age

2 Labels. Survived / Lost



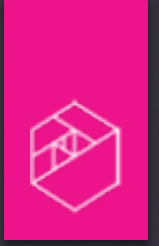
k-Nearest Neighbors Algorithm



“Tell me who you hang out with and I’ll tell you who you are.”

– Everyone’s parents … Also, KNN

KNN - Overview



The KNN Algorithm is fairly straightforward and can be summarized by the following three steps:

STEP 1: Choose *k* and distance metric *(commonly Euclidean)*

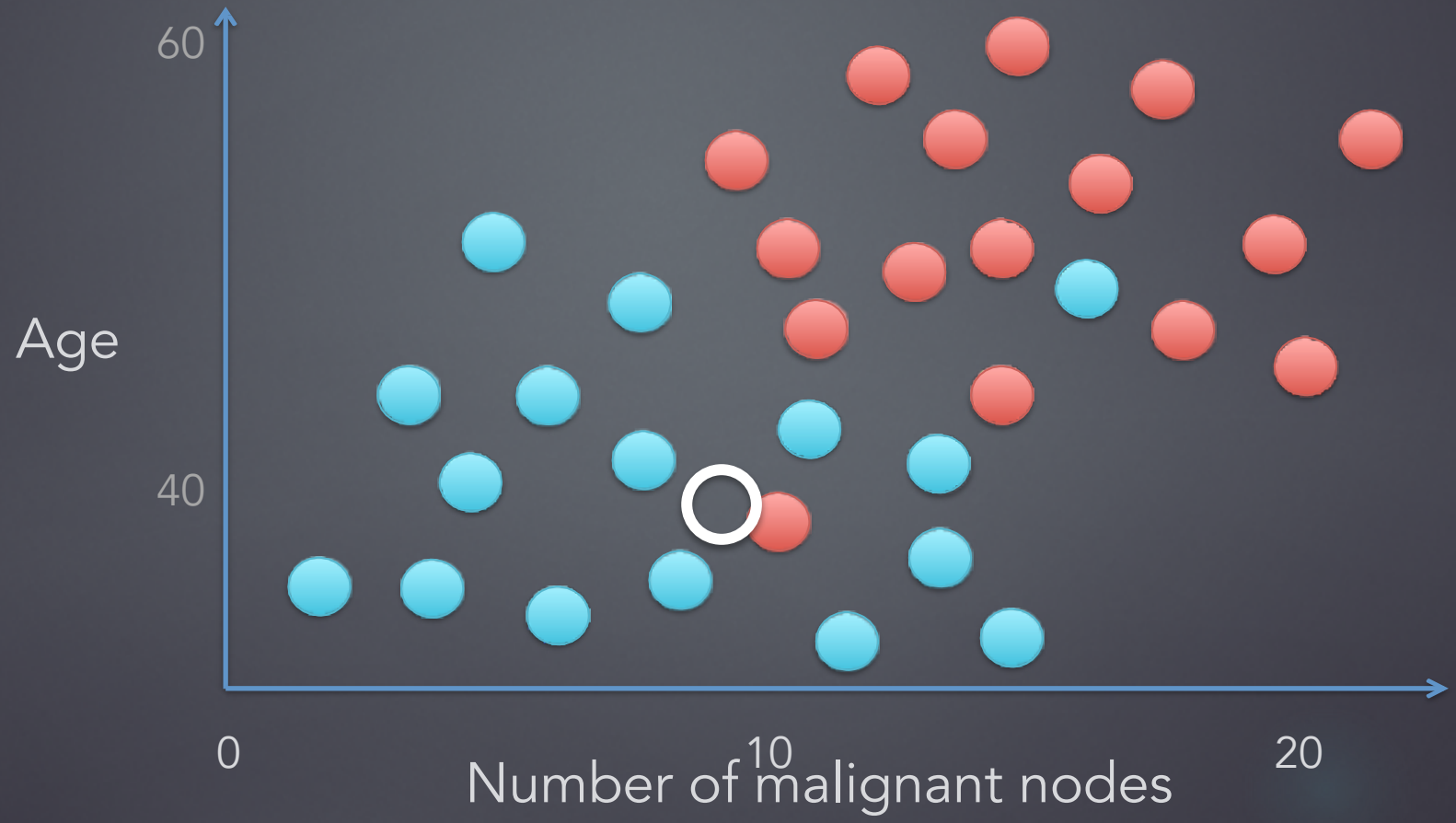
STEP 2: Find *k* nearest neighbors of observation to be classified

STEP 3: Assign class label by majority vote

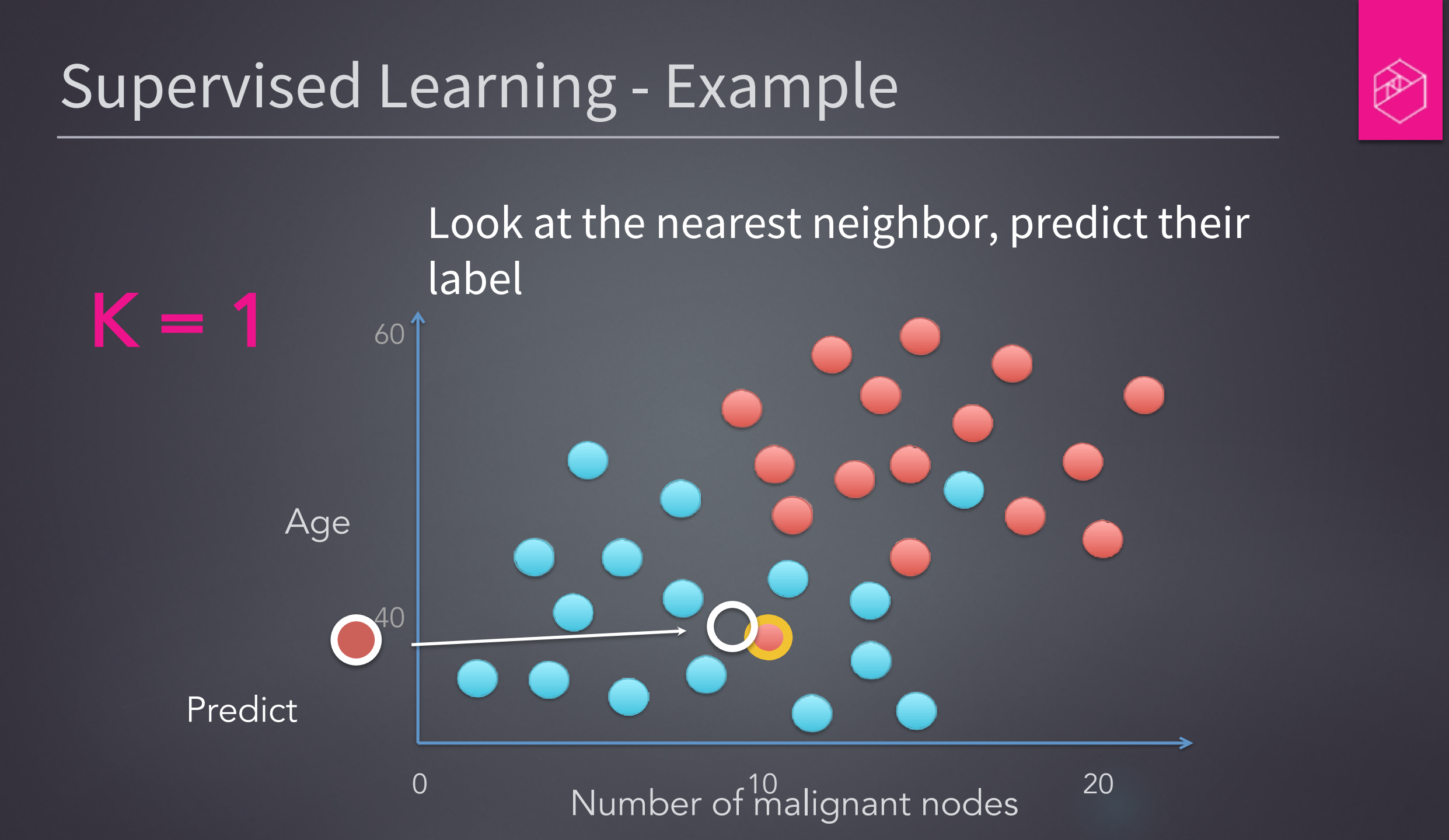
Supervised Learning - Example

Look at the nearest neighbor, predict their label

K = 1



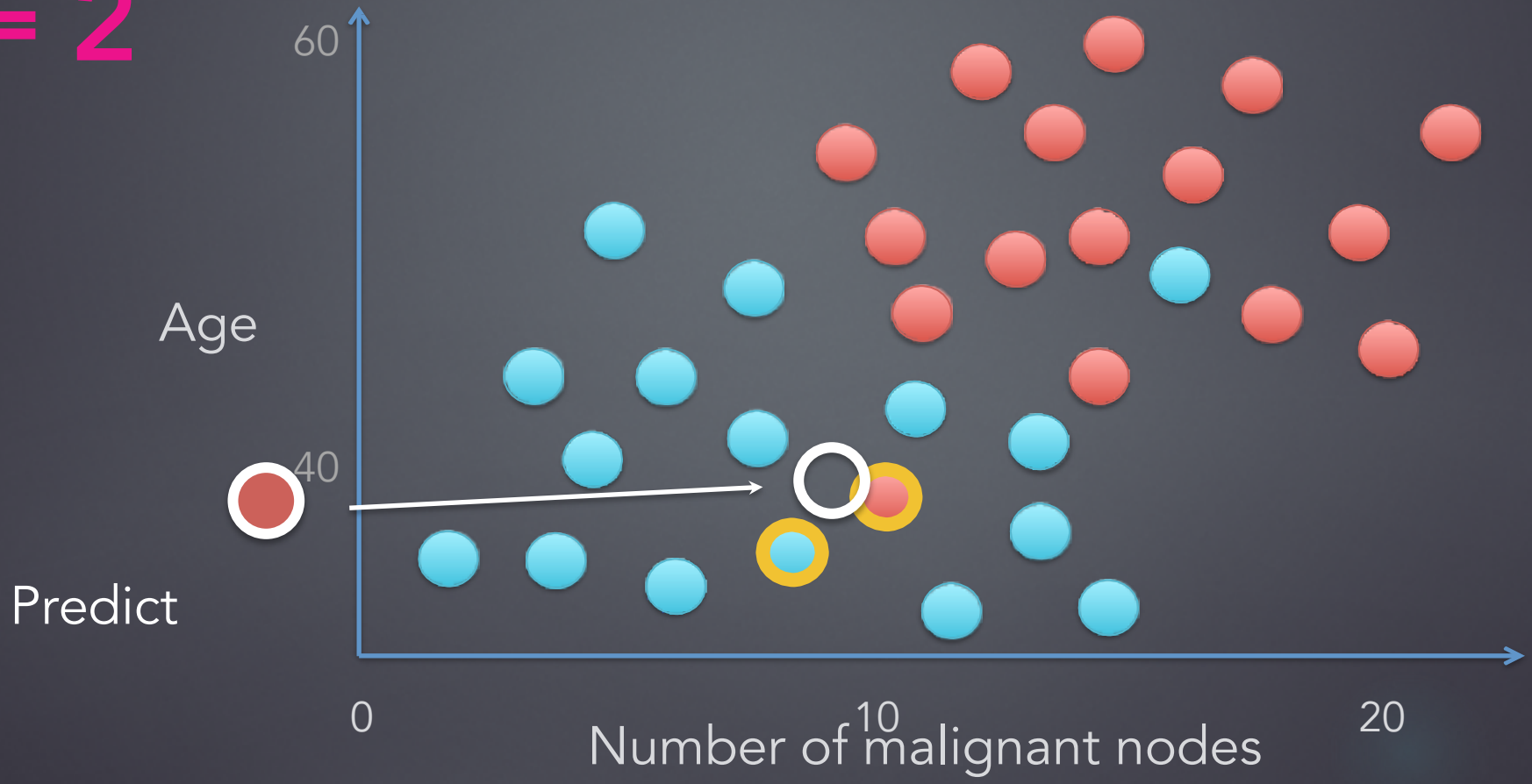
Supervised Learning - Example



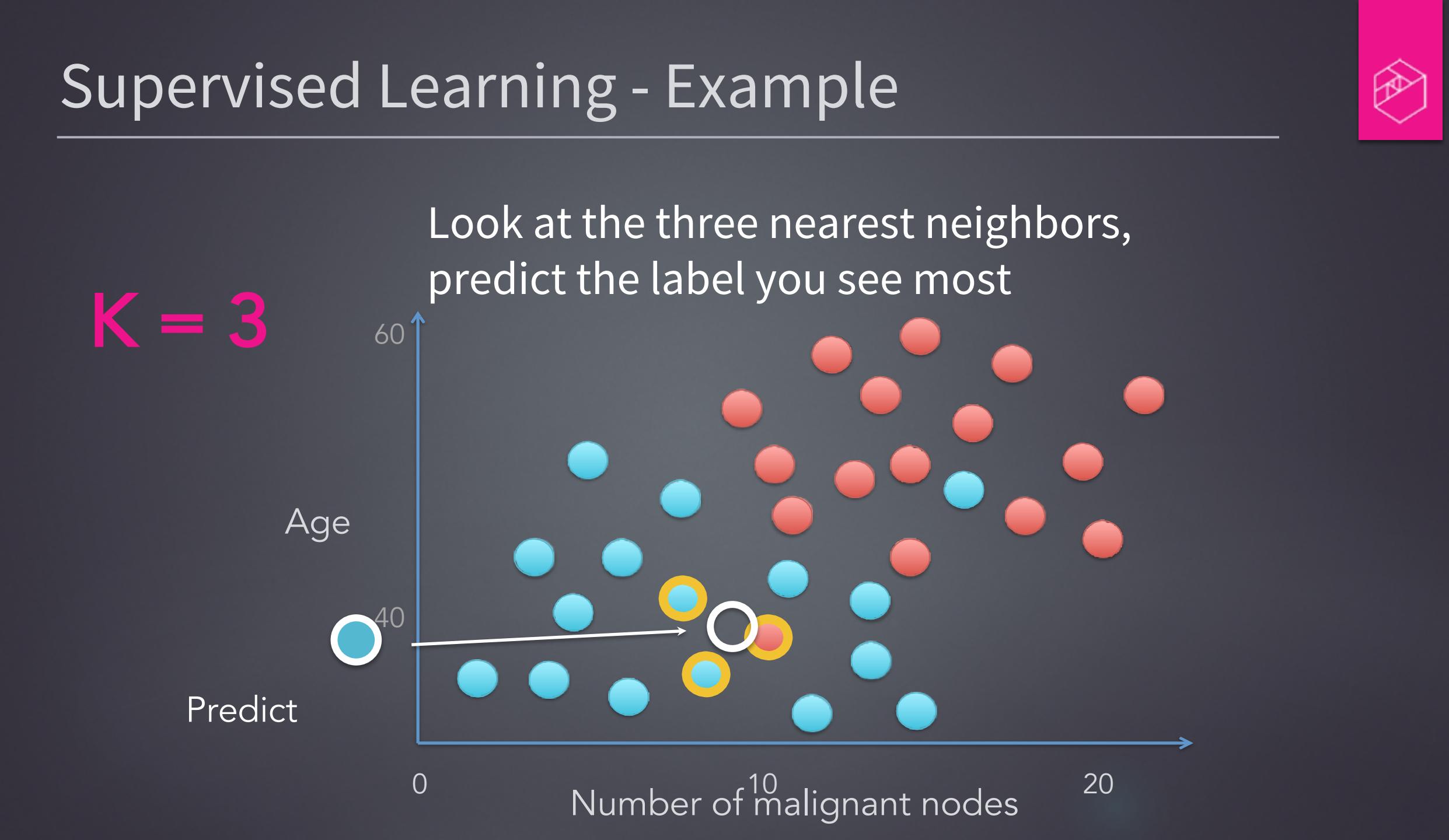
Supervised Learning - Example

Look at the two nearest neighbors, predict the label you see most

K = 2



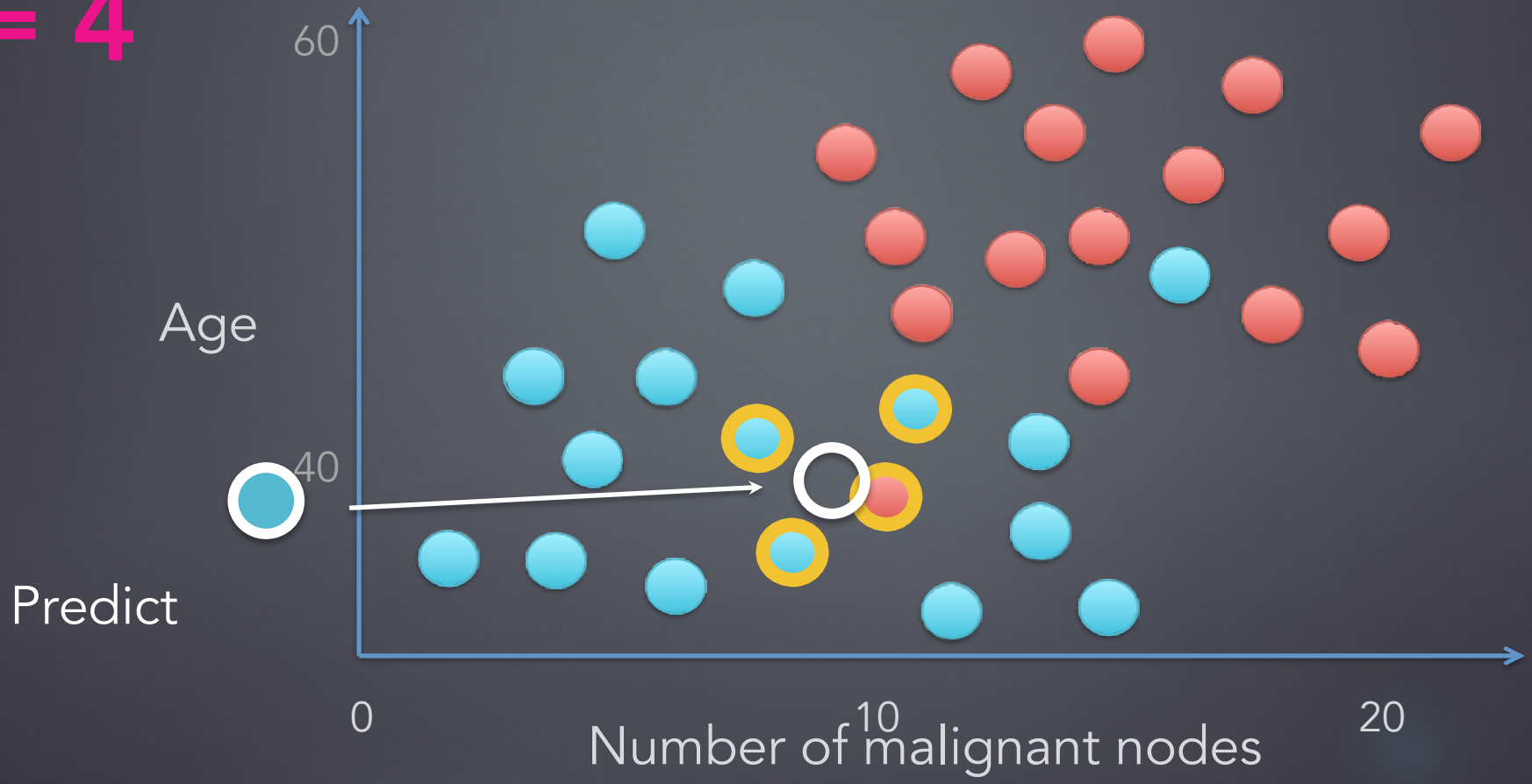
Supervised Learning - Example



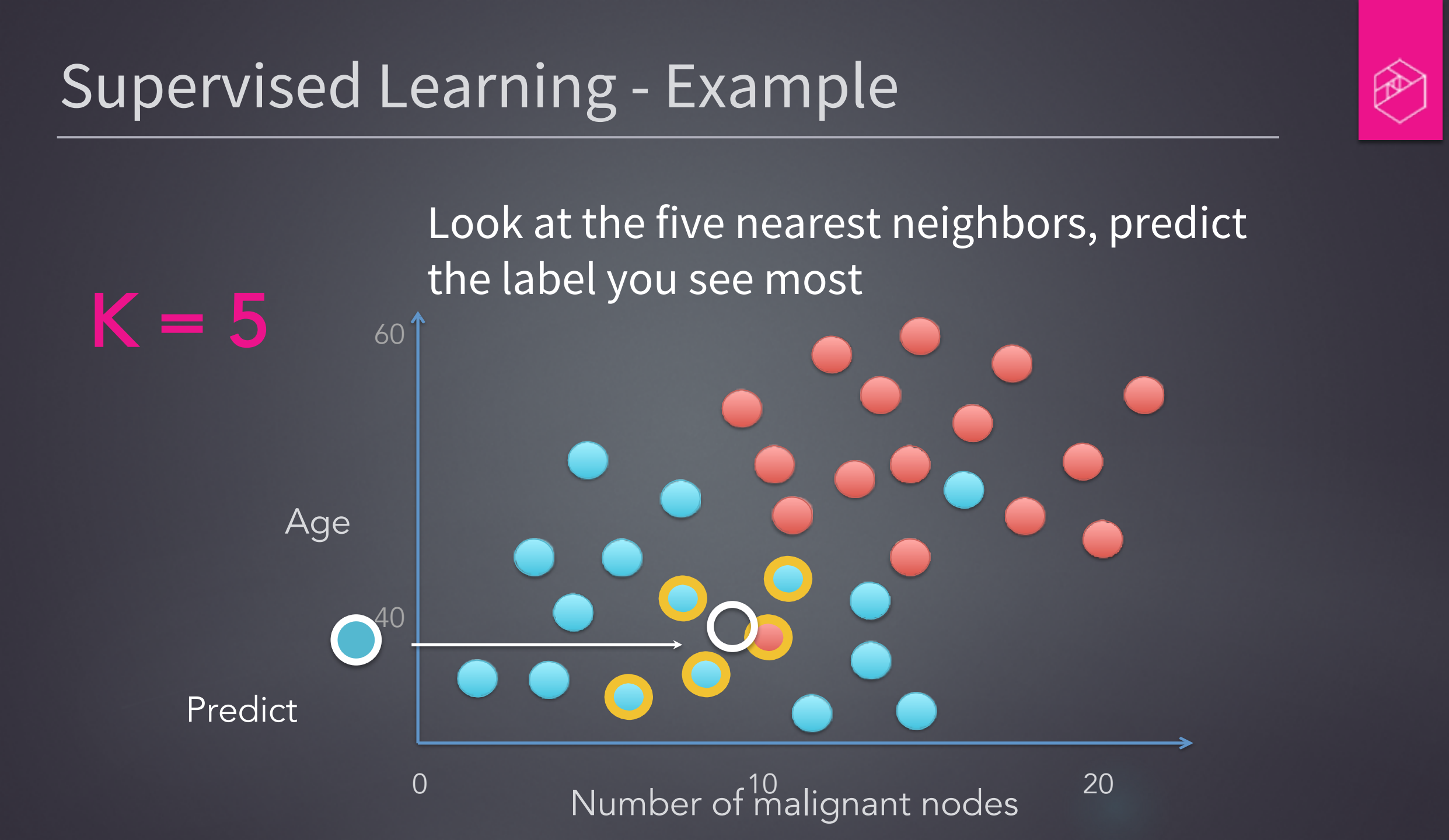
Supervised Learning - Example

Look at the four nearest neighbors, predict the label you see most

K = 4



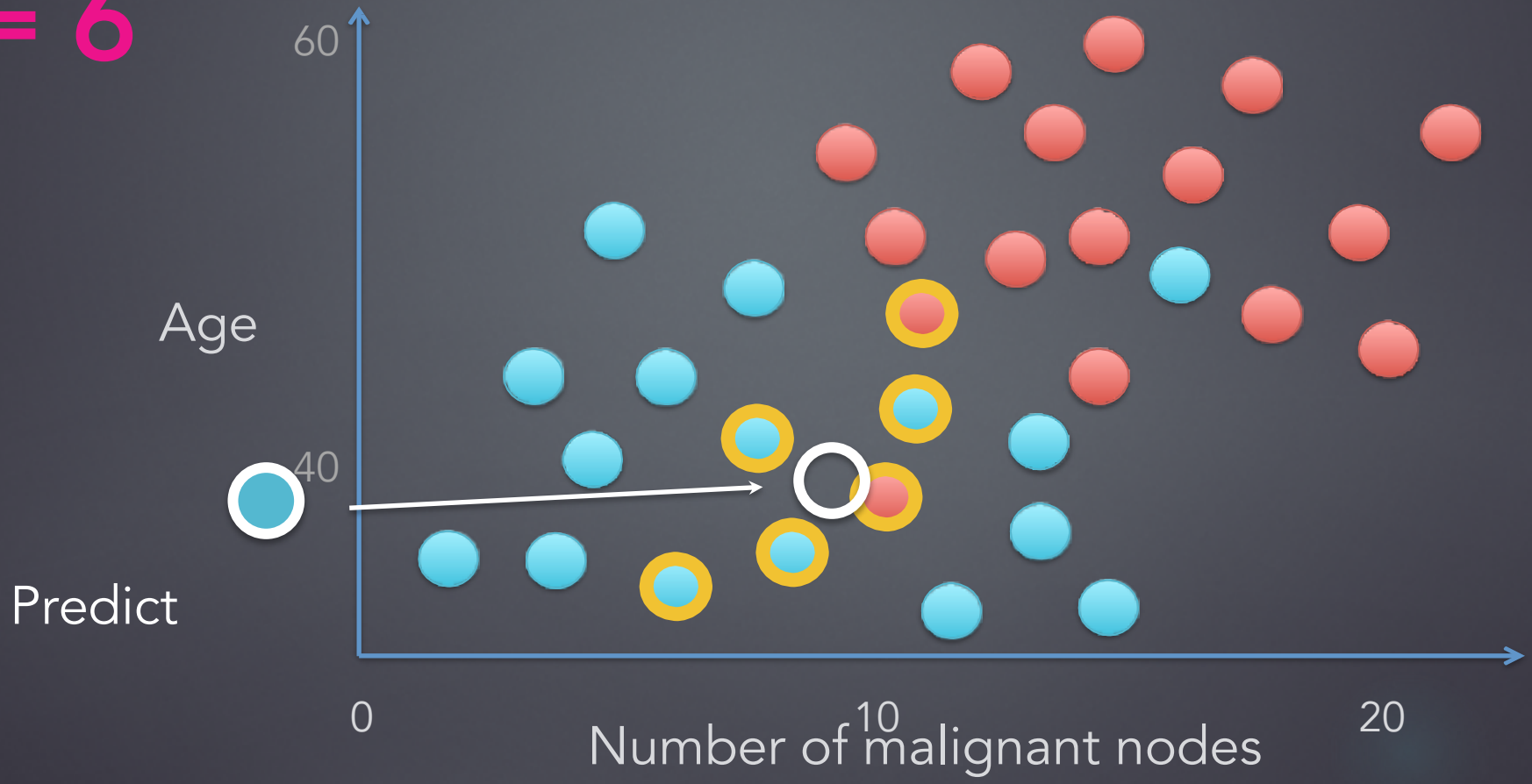
Supervised Learning - Example



Supervised Learning - Example

Look at the six nearest neighbors, predict the label you see most

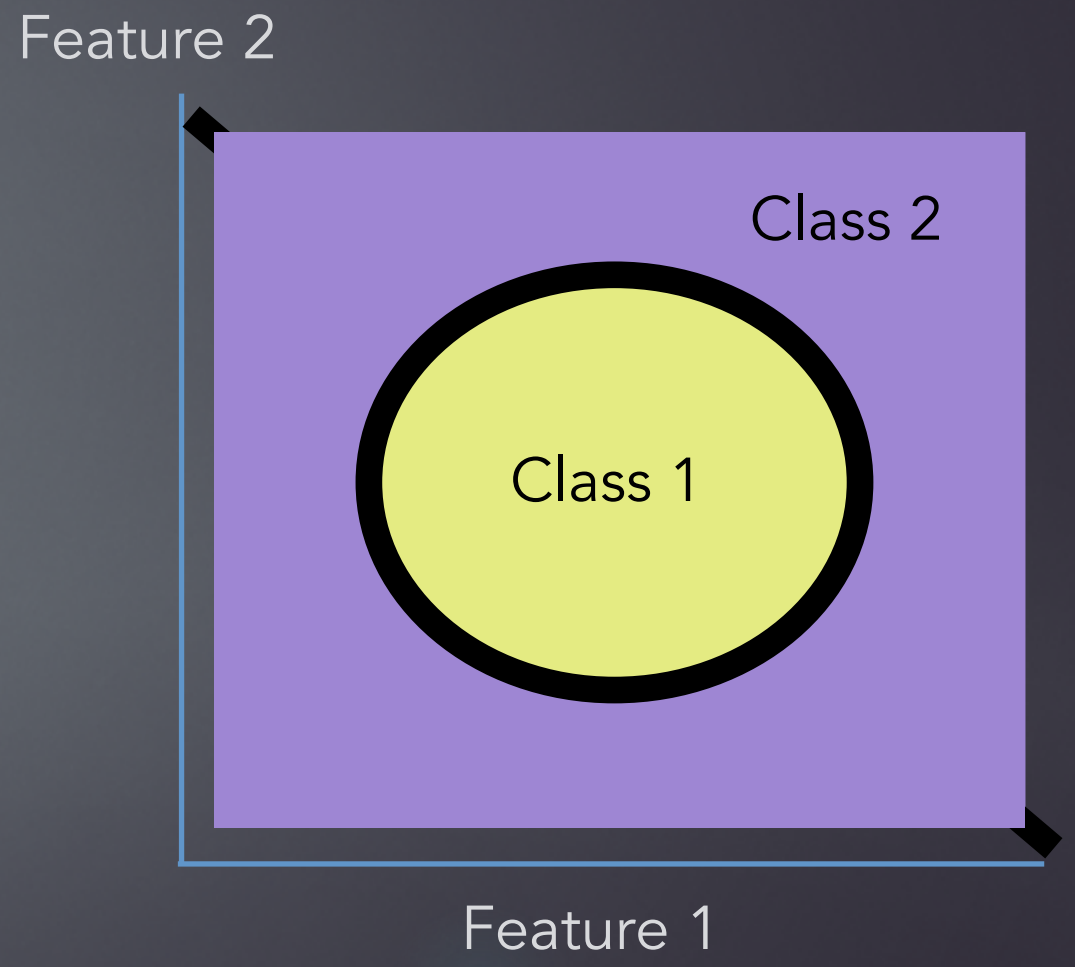
K = 6



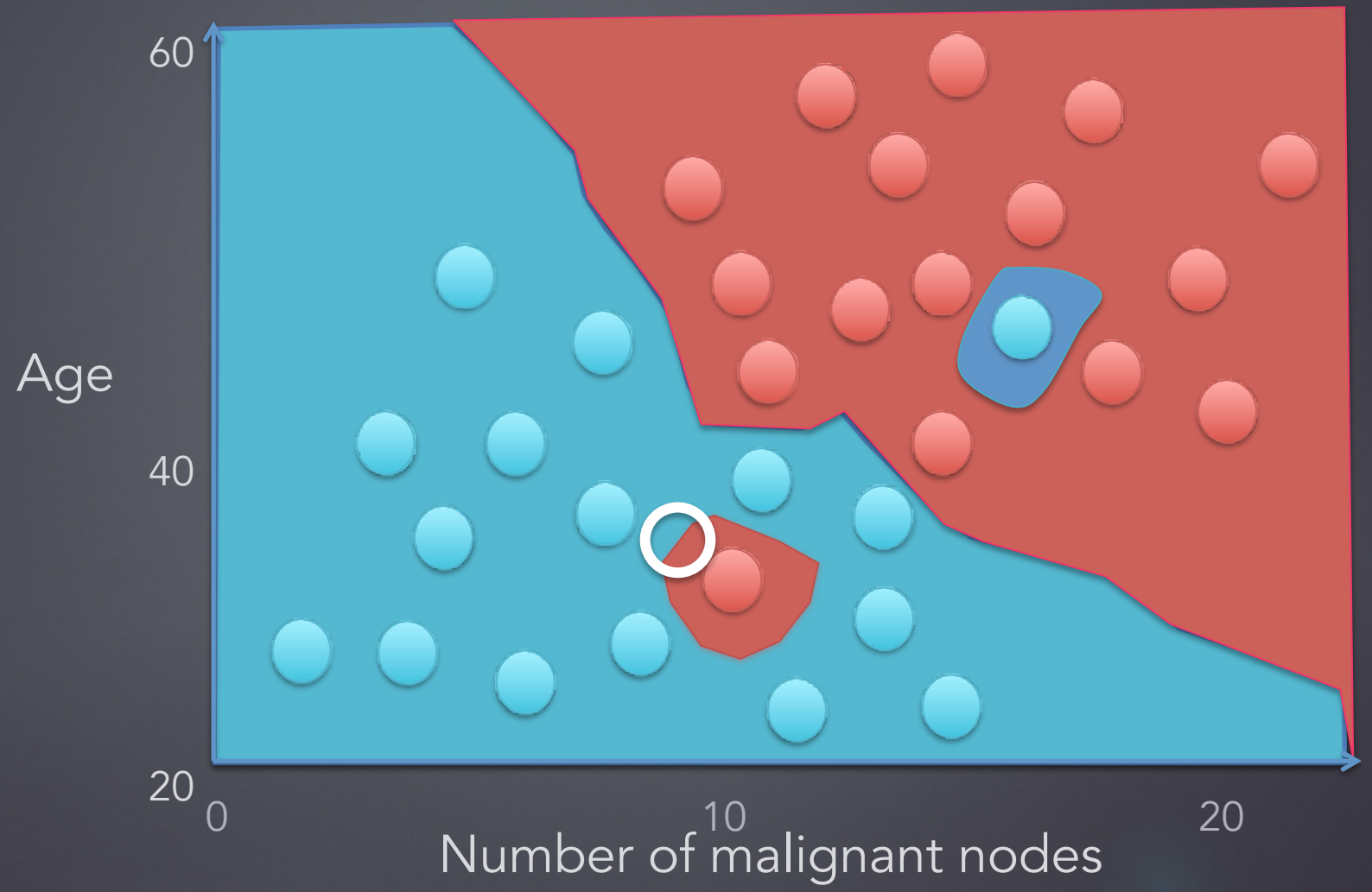
Decision Regions

Decision Regions - What are they?

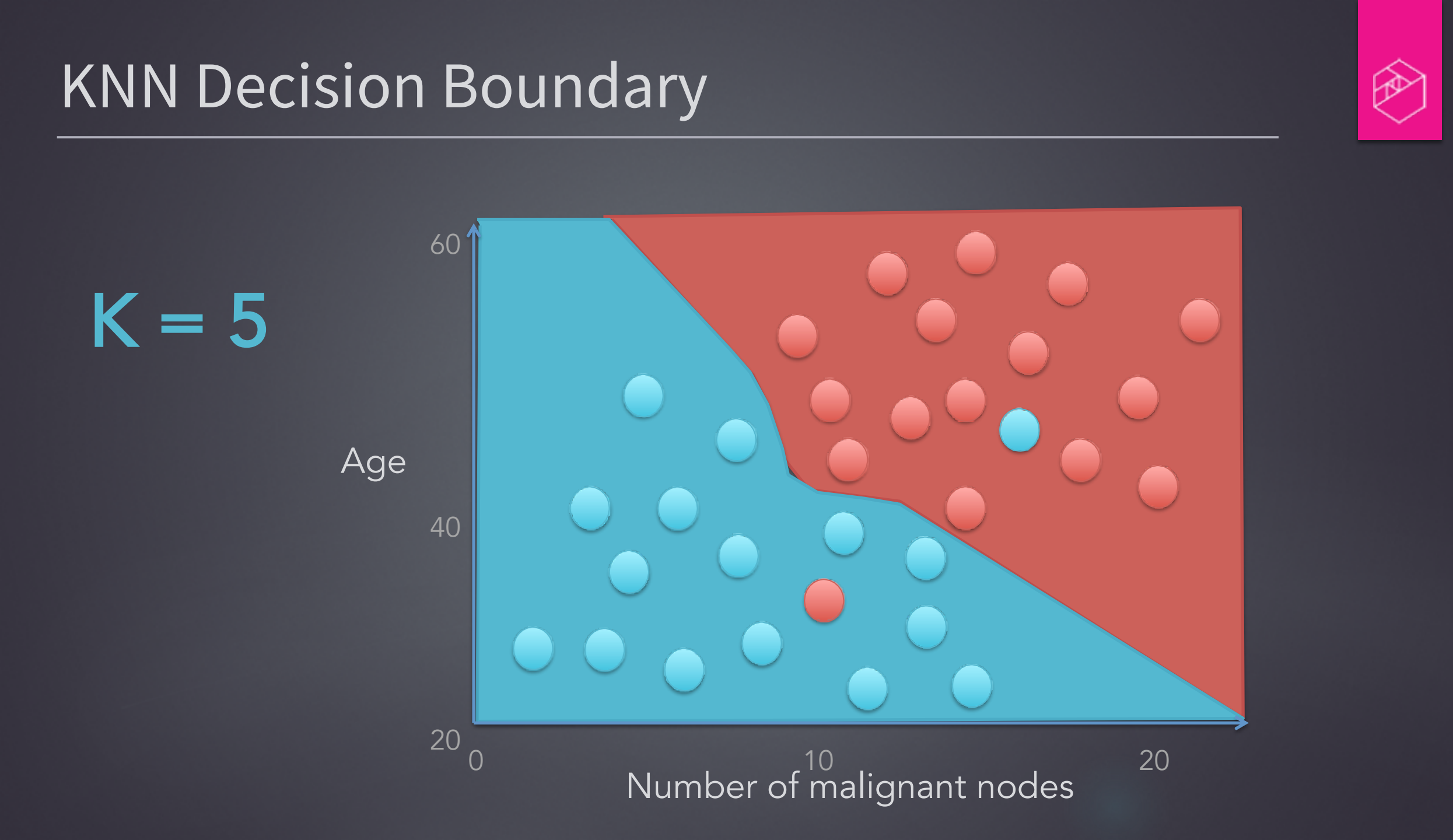
- In general, a machine learning classifier partitions the feature space into volumes called **decision regions**.
- All observations **inside** a decision region are assigned to the same category
- The decision regions are separated by surfaces called **decision boundaries**. These boundaries represent points where there are ties between two or more categories



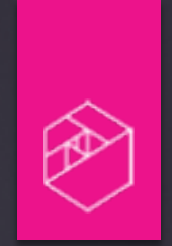
KNN Decision Boundary



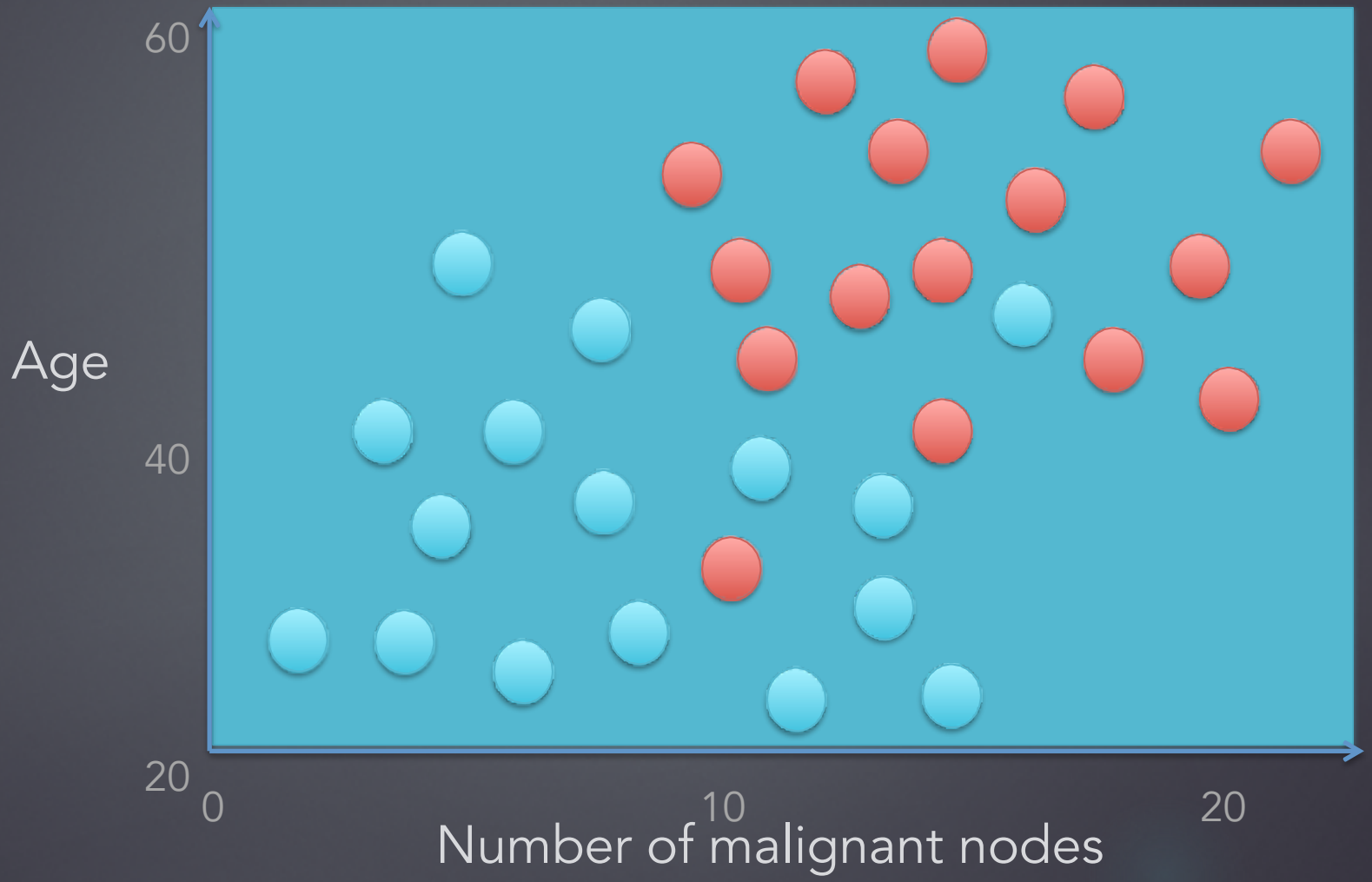
KNN Decision Boundary



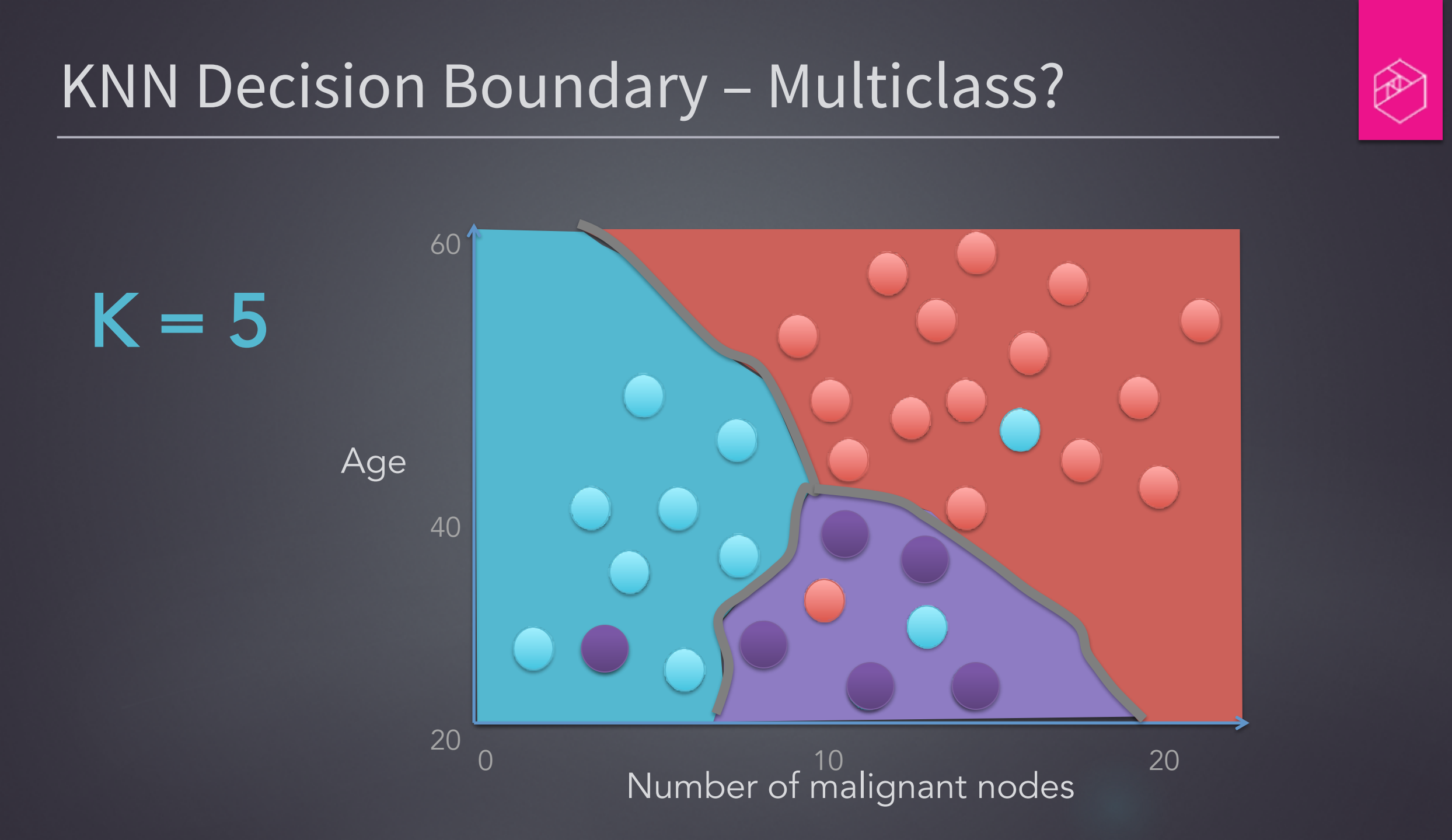
KNN Decision Boundary



K = 34

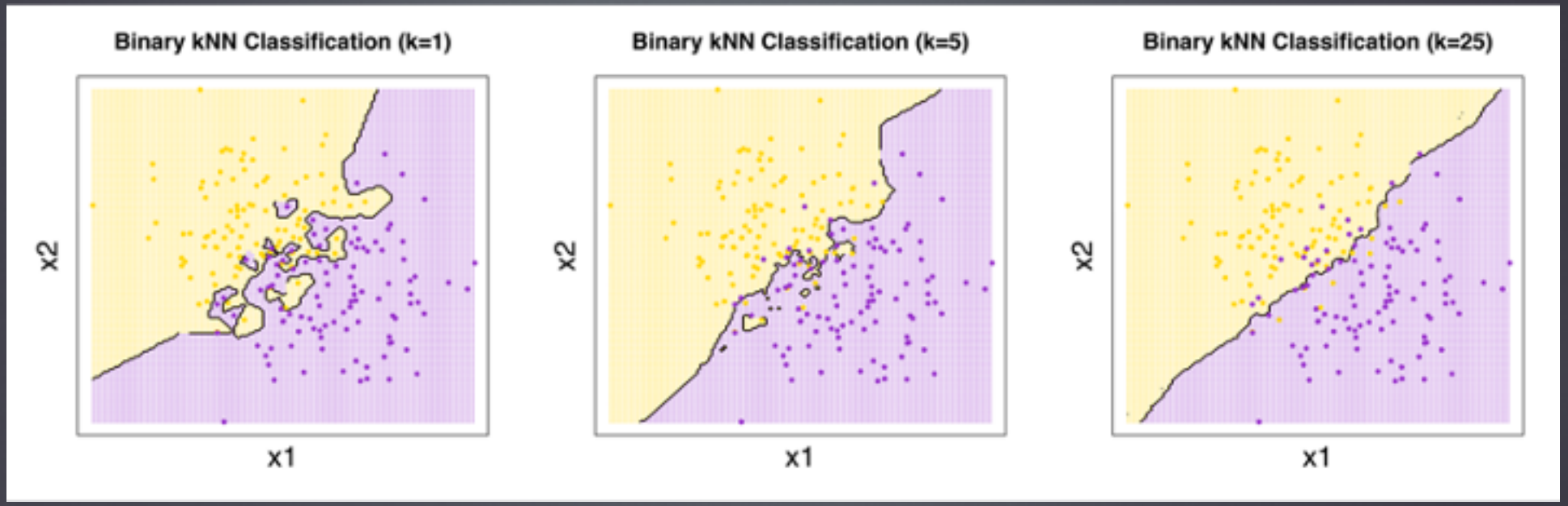


KNN Decision Boundary – Multiclass?

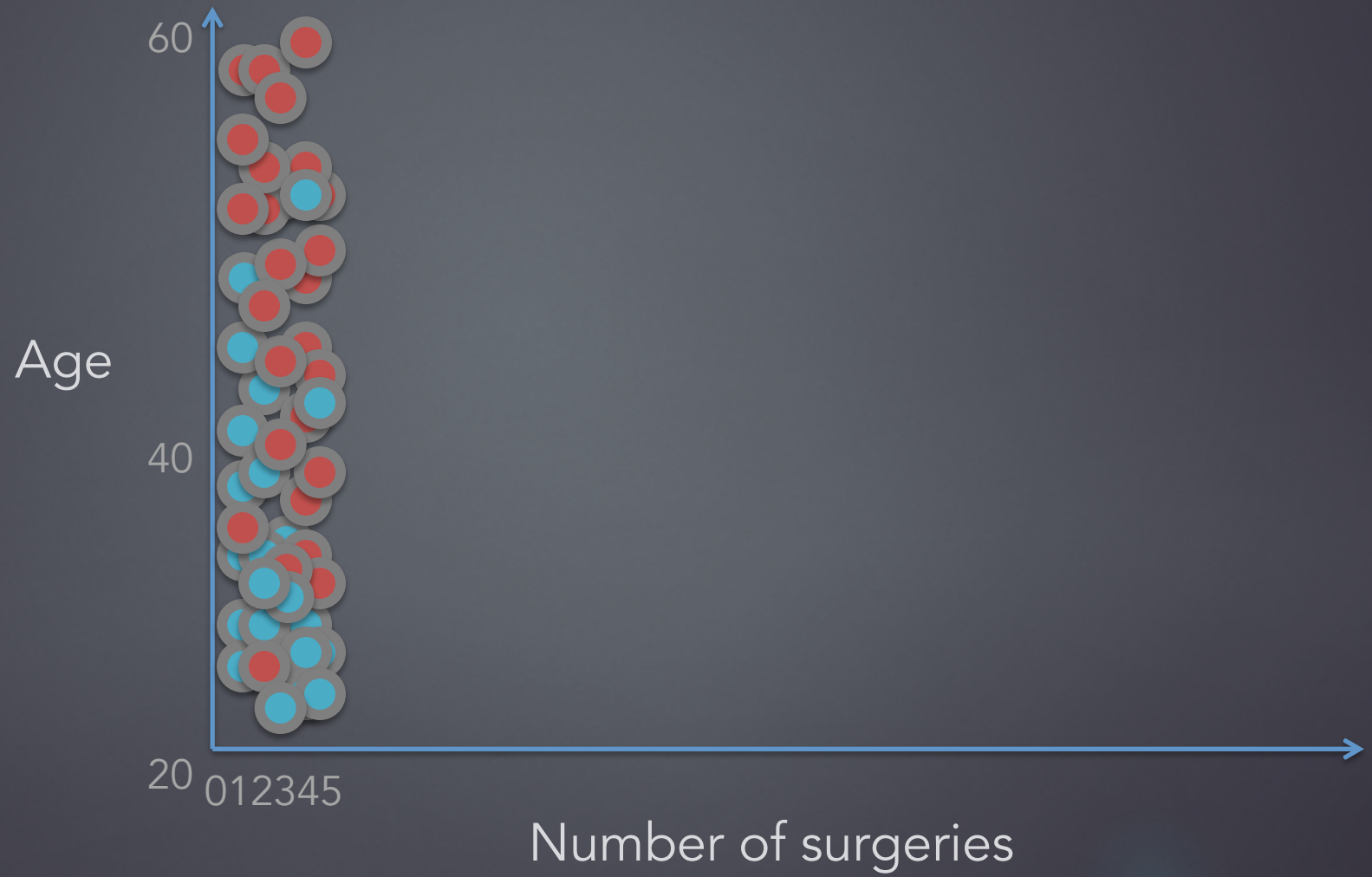


K = 5

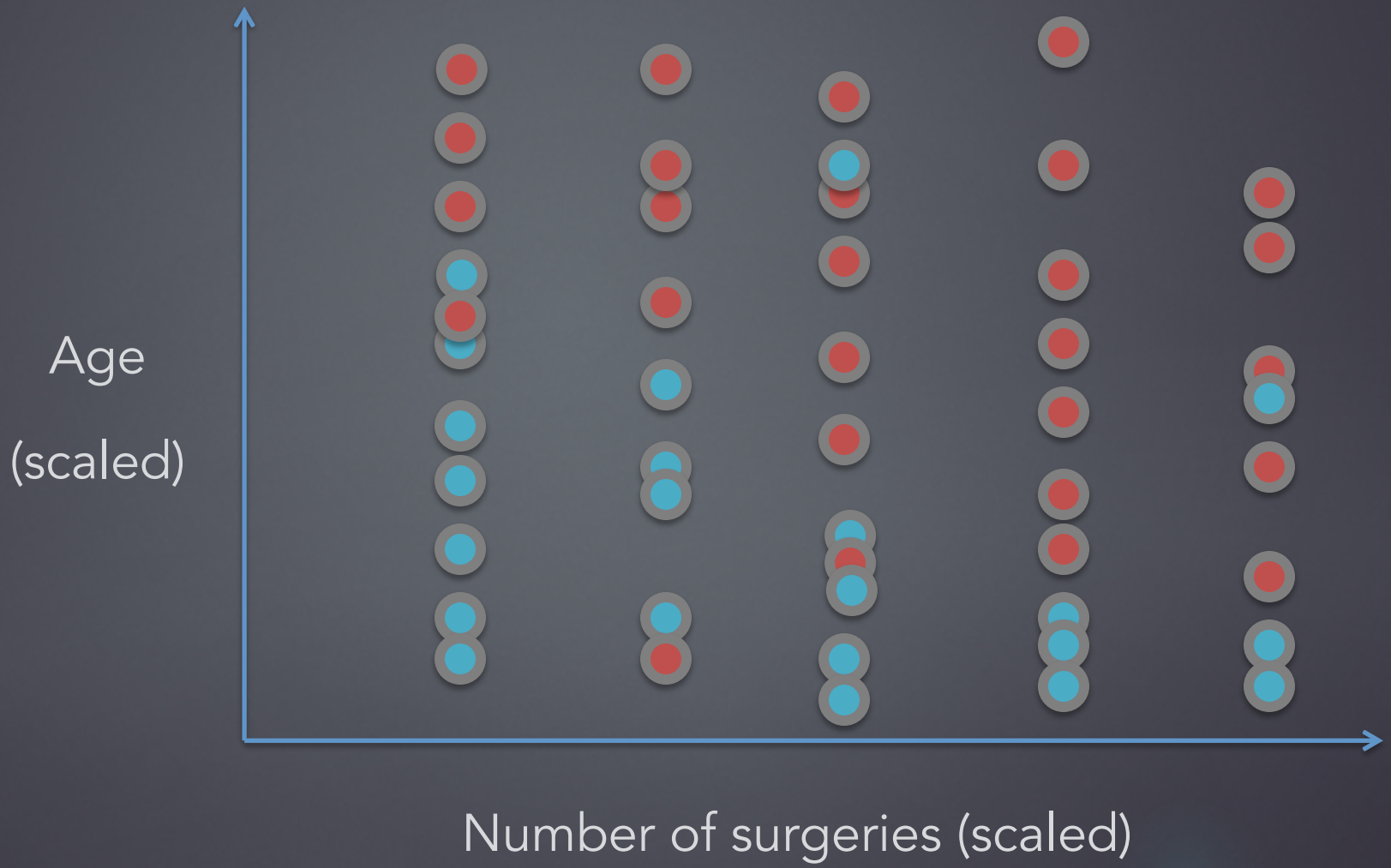
KNN Decision Boundary



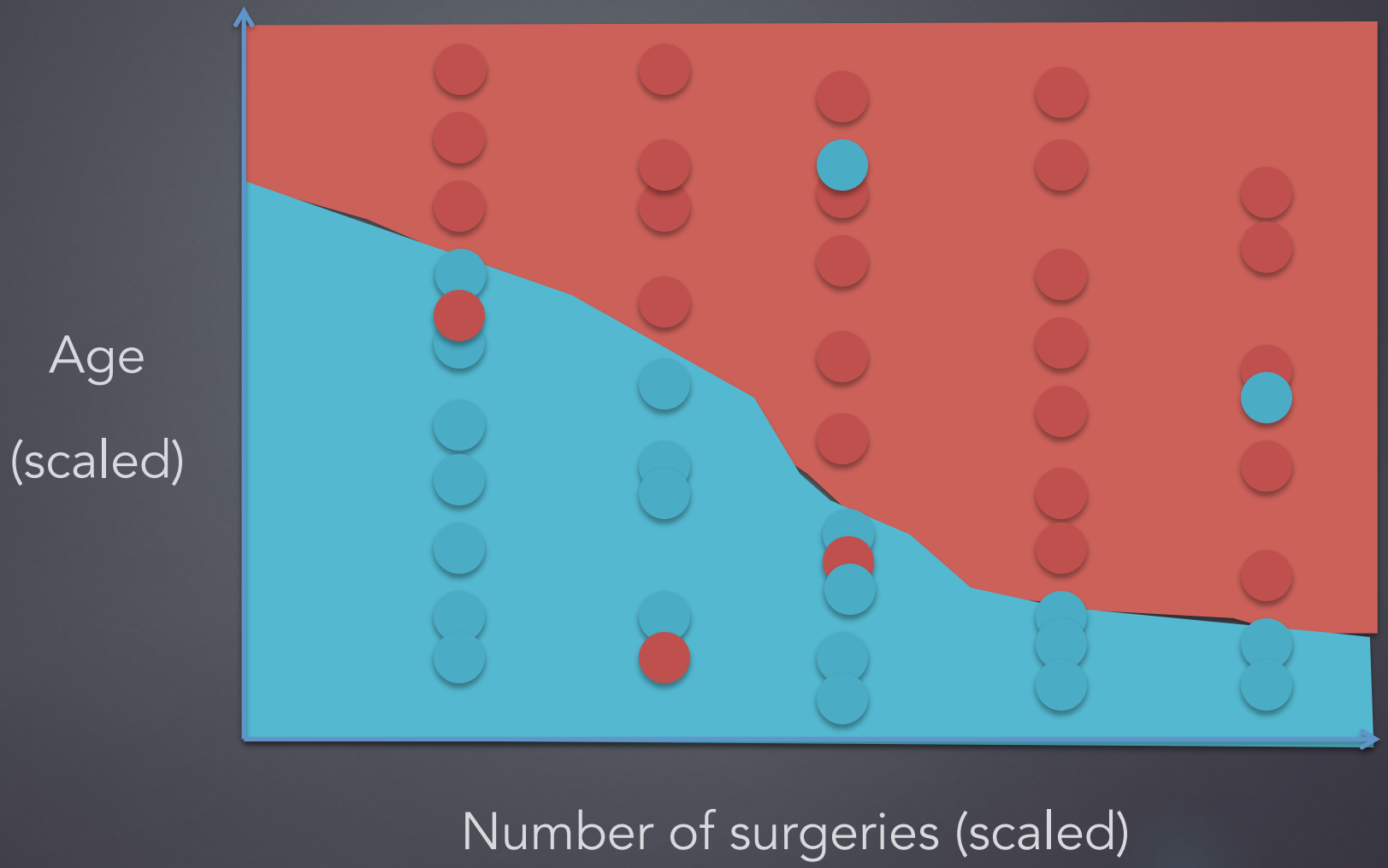
Remember to Scale!



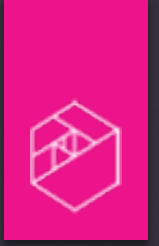
Remember to Scale!



Remember to Scale!



KNN - Overview



Pros

- Easy to implement
- Interpretable
- Classifier adapts as new training data is collected

Cons

- “Lazy” - KNN doesn’t have any “training time”, instead memorizes the training data
- Memory intensive
- Prediction time scales linearly as more training data is introduced