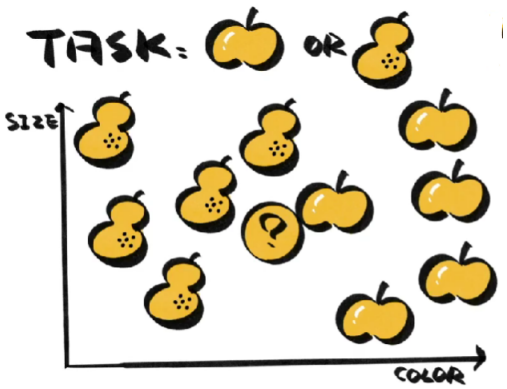
什么是KNN(K近邻算法)

虽然名字中有NN，KNN并不是哪种神经网络，它全名**K-Nearest-Neighbors**：K近邻算法，是机器学习中常用的分类算法。

物以类聚，人以群分。KNN的基础思想很简单，要判断一个新数据的类别，就看它的邻居都是谁。

假设我们的任务是分类水果，虽然不知道新来的水果是梨还是苹果，但通过观察它的大小和颜色，我们找到了它在坐标系中的位置，再看看已经确定的苹果和梨都在哪，如果附近的苹果多，我们就认为它是苹果，反之认为它是梨。



KNN中的**K**指的是**K**个邻居，K=3就是通过距离最近的3个样本，来判断新数据的类别。

大小和颜色是数据的特征，苹果和梨是数据的标签。计算距离时既可以使用两点之间的直线距离，也就是欧式距离，也可以使用坐标轴距离的绝对值的和，也就是曼哈顿距离。

对于KNN来说，K的取值非常重要，如果K的值太小，很容易受个例影响，K的值太大，又会受到距离较远的特殊数据影响。K的取值受问题自身和数据集大小决定，很多时候要靠反复尝试。

KNN算法能做什么?

- 根据花瓣长度、宽度等特征判断植物类别
- 将文本分词、统计词频等处理后判断文章的类型
- 电商、视频网站可以找到与你类似的用户，依据他们的选择推荐你可能感兴趣的商品或内容

简单好用的KNN同样存在一定的缺点，它的流程是先计算新样本和所有样本之间的距离，按由近及远的顺序排序后，再按K值确定分类，因此数据越多，KNN的计算量越大，效率也就越低，很难应用到较大的数据集中 。