

如何入门人工智能？

“人工智能”是一个充满憧憬的词，但目前电影里神一般的智能还是不存在的，现有的人工智能可以把一些要人来做的工作变成机器操作从而提升效率，不要幻想通过时间的学习就能变得无所不能。并且，我们这里只是说如何入门，并不是说如何成为大神，所以我们的目标是短期内能自己构造出一个模型来解决一些问题。

因此，我们要带着问题来学习。

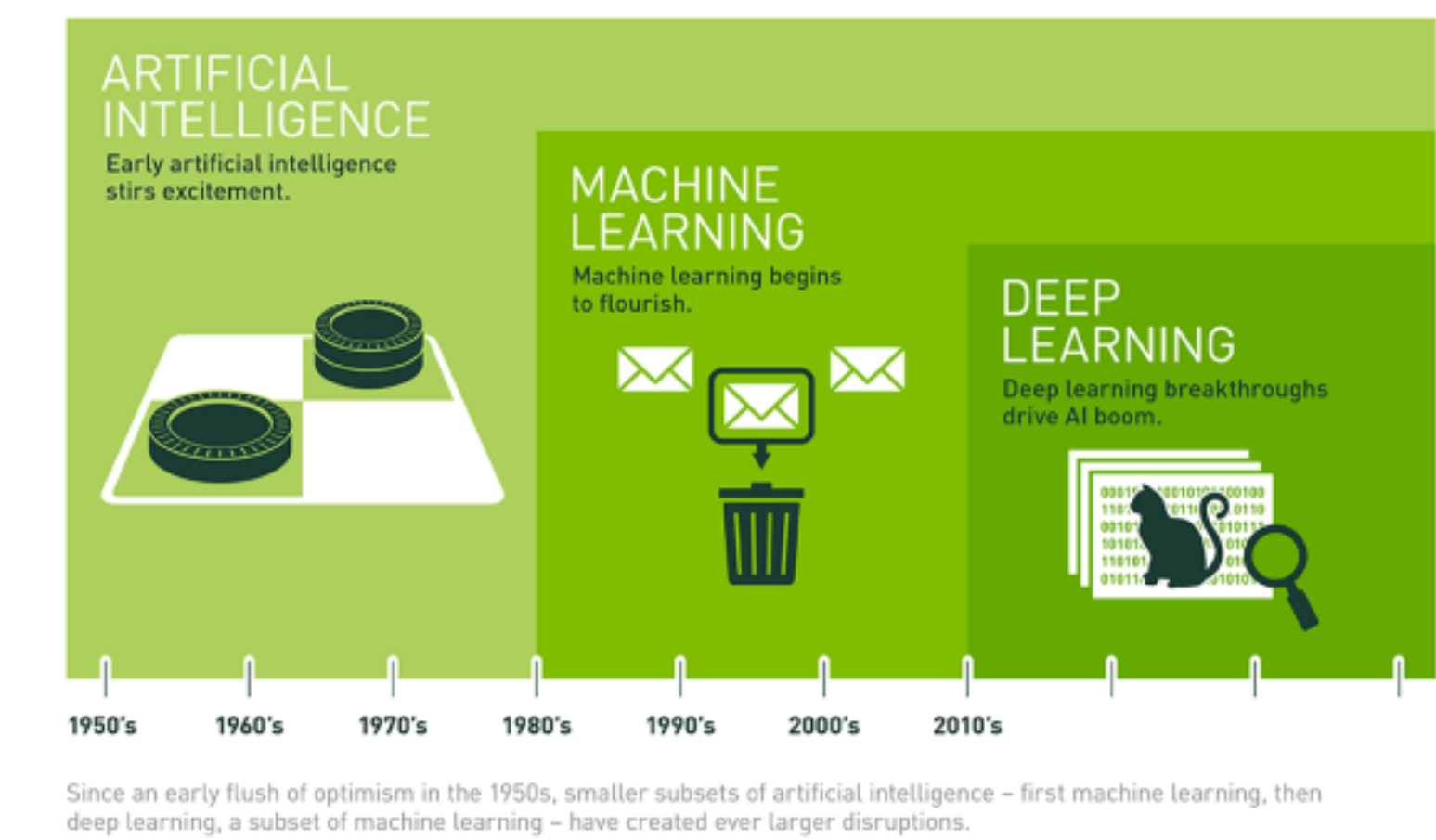
首先，我们要了解解决一个实际问题需要哪些技术。通常，人工智能技术由四个部分组成：认知、预测、决策和集成解决方案。

- **认知**：是指收集信息和解析信息来感知世界，比如图片识别、语音识别、自然语言处理等；
- **预测**：是指通过计算，来预测行为和结果。比如广告推荐，歌曲推荐等；
- **决策**：是指确定实现的方式和路径，比如移动路线规划、自动买卖股票等；
- **集成解决方案**：是指人工智能和其他技术结合时，产生的多种集成解决方案，比如和汽车结合就是无人驾驶，和医疗器械结合就是手术机器人。

目前商业化比较普遍的，是认知和预测领域的应用。

我们这里选“图片识别”和“验证码识别”两个实际问题来解决，在解决这些问题的过程中，逐步地掌握相应的知识。通过项目来学，以代码实现为主，中间穿插理论知识。在项目做完的同时，也就掌握了解决一类问题的思路、基本工具和相应的理论知识。

人工智能、机器学习、深度学习，这三者是什么关系？



机器学习有哪些分类？

机器学习的范围很广，主要分为三大类：监督学习（Supervised Learning）、非监督学习（Unsupervised Learning）和强化学习（Reinforcement Learning）。

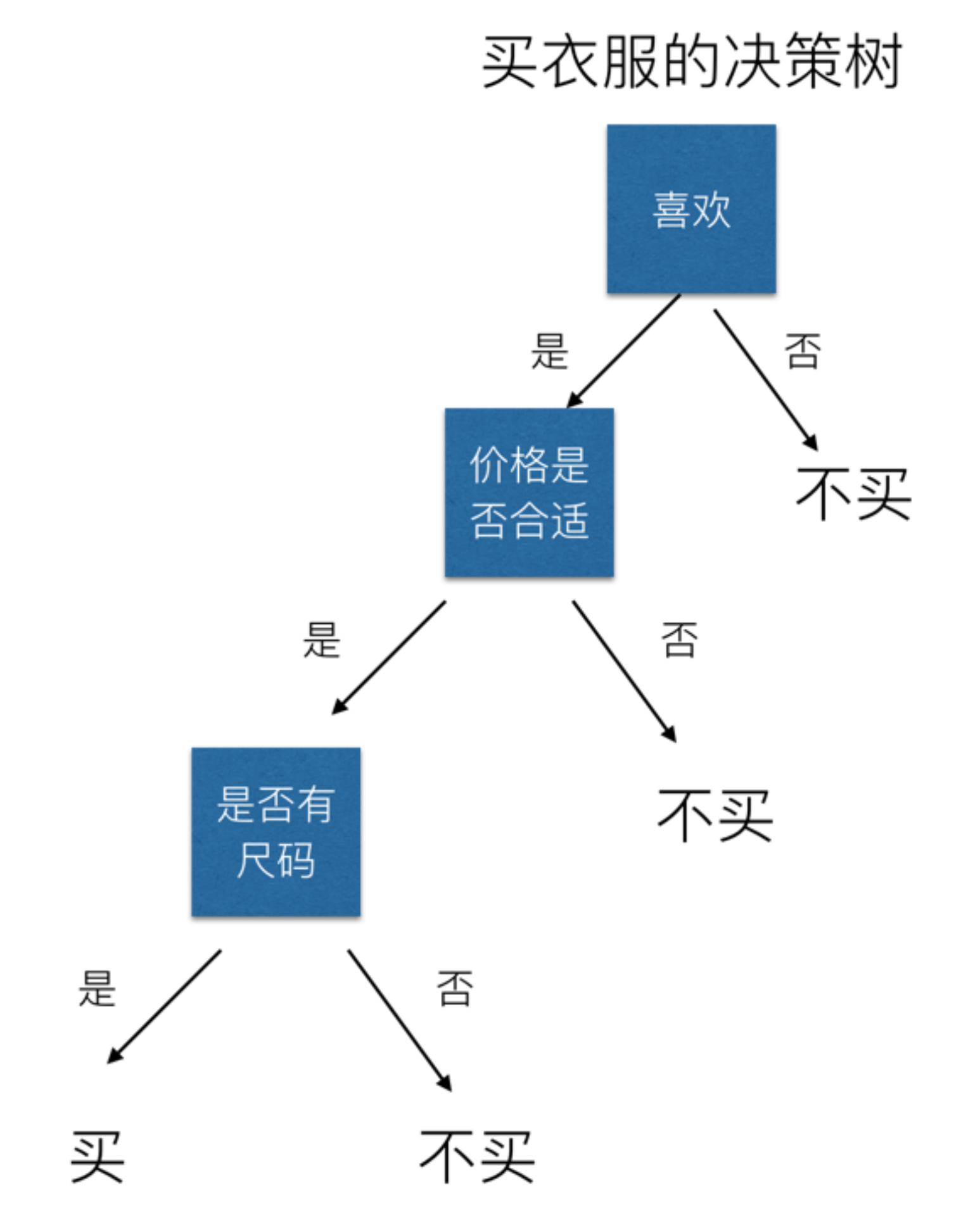
- **监督学习**：监督学习学的是带标准答案的样本。拿猫和狗的识别来举例子。算法看一张图就告诉它，这是猫；再看一张图片，告诉它这也是猫，再看一张图，告诉它这是狗，如此往复。当它看了几十万张猫和狗的图片后，你再给它一张陌生的猫或者狗的图片，就基本能“认”出来，这是哪一种。这样的学习方法很有可能造成模型把所有答案都记了下来，但碰到新的题目又不会了的情况，这种情况叫做“过拟合”。
- **非监督学习**：非监督学习学的是没有标准答案的样本。同样拿猫和狗的识别举例。算法要自己去寻找这些图片的不同特征，然后把这些图片分为两类。它实际上不知道这两类是什么，但它知道这两类各有什么特征，当再出现符合这些特征的图片时它能识别出来，这是第一类图片，那是第二类图片。
- **强化学习**：强化学习的学习方式是通过不断做出决策并获得结果反馈后，学会自动进行决策，得到最优结果。我们小时候，看到马戏团的猴子居然会做算术题，感到很惊讶，这是怎么做到的呢？其实就是每次拿对了数字的时候，训练人员就给它一些食物作为奖励，这些奖励让他“知道”，这么做是“对的”，如果拿错了，可能就会有惩罚，这些惩罚就是要让它“知道”，这样做是“错的”。如此一来，经过不断的训练，猴子就“会”做算术题了。

机器学习有哪些常见算法呢？

这里我们对常见的机器学习算法进行简单的介绍，不求精准，只求容易理解。

1，决策树

决策树是一种用于对数据进行分类的树形结构。比如我们要买一件衣服，如果喜欢则进入下一步，看价格，如过价格合适，就看是否有合适的尺码，有合适的尺码，就下单购买。用决策树表示如下图：



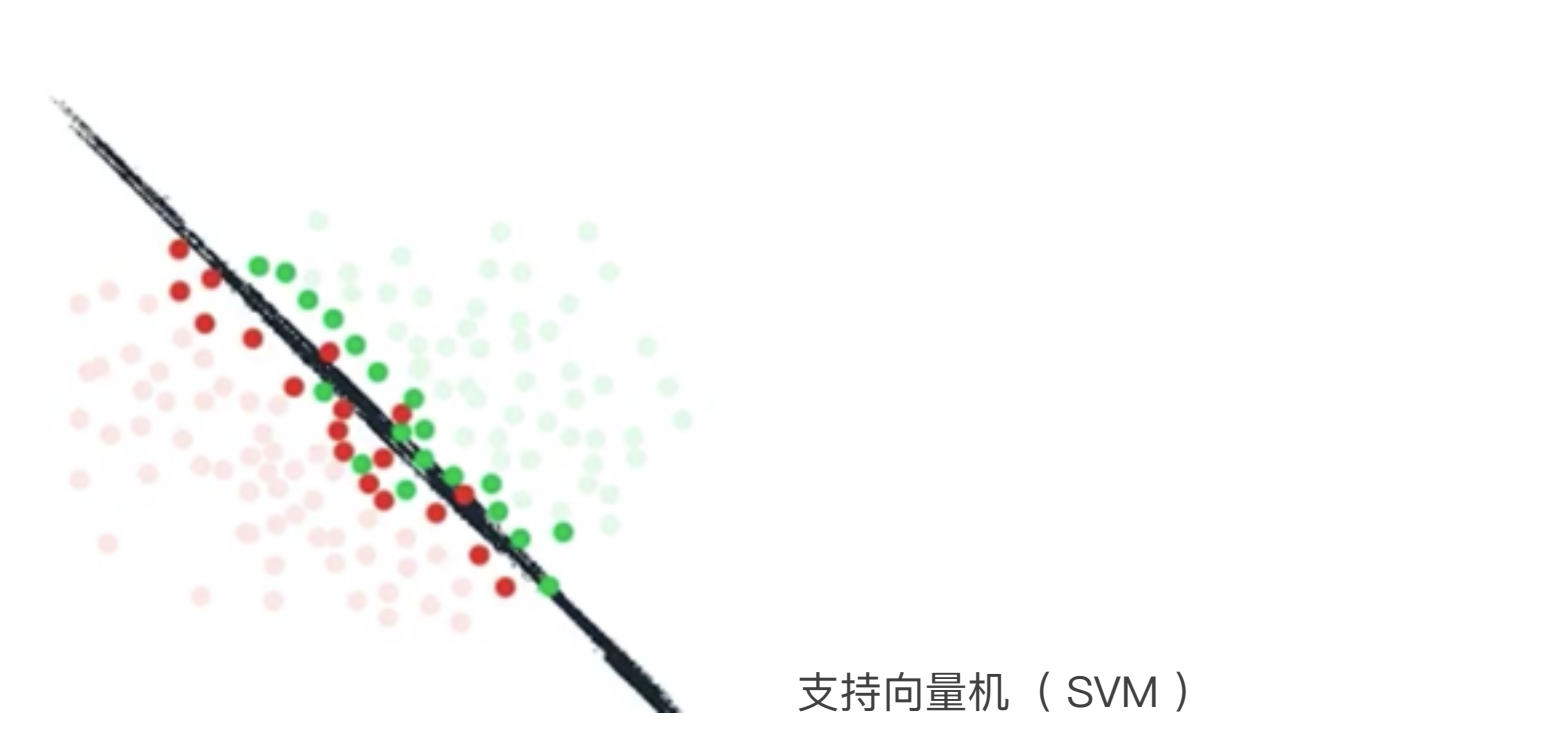
买衣服的决策树

2，线性回归

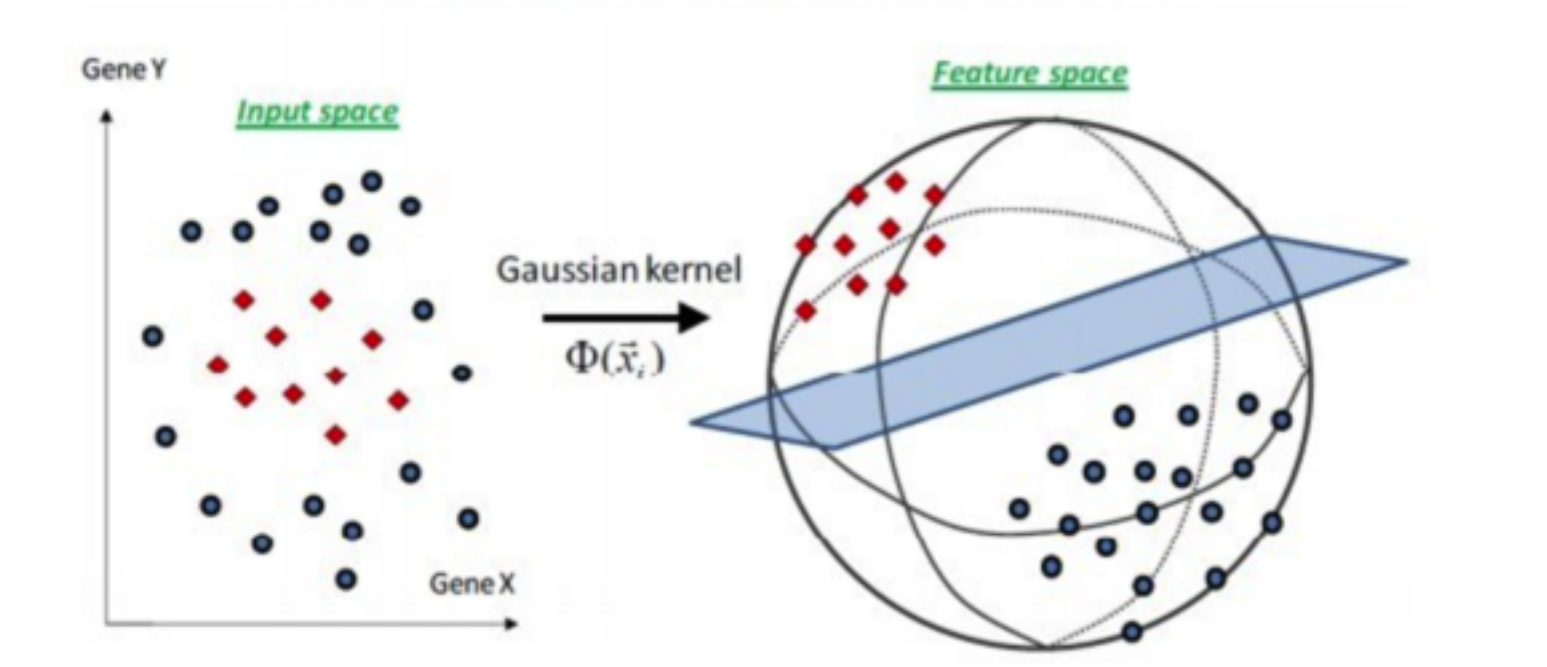
试想，在纸上有很多的点，我们打算画一条直线，让这些点到这条直线的距离之和最短，怎么找到这条直线呢？这个方法就是线性回归。画一条线，让样本以及后面预测的点都尽量在这条线附近。

3，支持向量机和核函数

支持向量机是一种分类方法，力求在样本中划出一道线，让线距离两边样本的距离最大。它在文本分类、图像分类有较多应用。如果桌子上有红豆和绿豆，我们可以把SVM想象成一个忍者，他画了一条线，把红豆和绿豆区分开来。



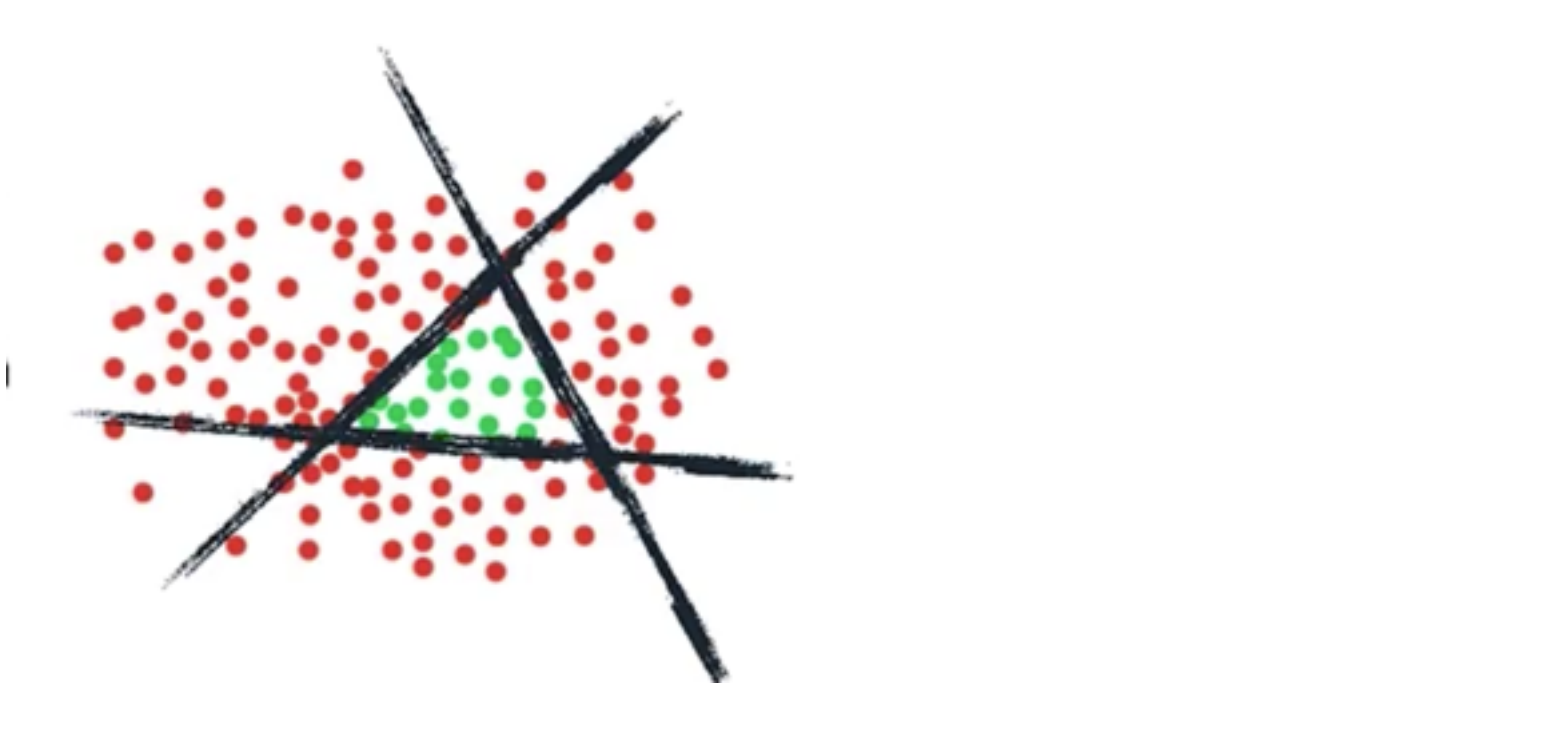
但有的时候豆子掺和在了一起，怎么办呢？我们可以针对红豆和绿豆的不同特性，把这些豆子都通过核函数进行计算，把他们映射到高维空间中去，这样豆子自然就被分开了。



加入核函数后的SVM

4，神经网络

神经网络也是一种分类器。它是由很多个虚拟的神经元组成的一个网络，我们可以把一个神经元看做一个分类器，那很多个神经元组成的网络就能对样本进行很多次分类。还是拿忍者 and 豆子区分举例。一个神经元，相当于忍者可以划一刀，多个神经元就可以划多刀，划的越多，自然分的越细。这里只是做简单的介绍，大家有概念即可，更详细的在后面会更新。



神经网络

5，朴素贝叶斯分类器

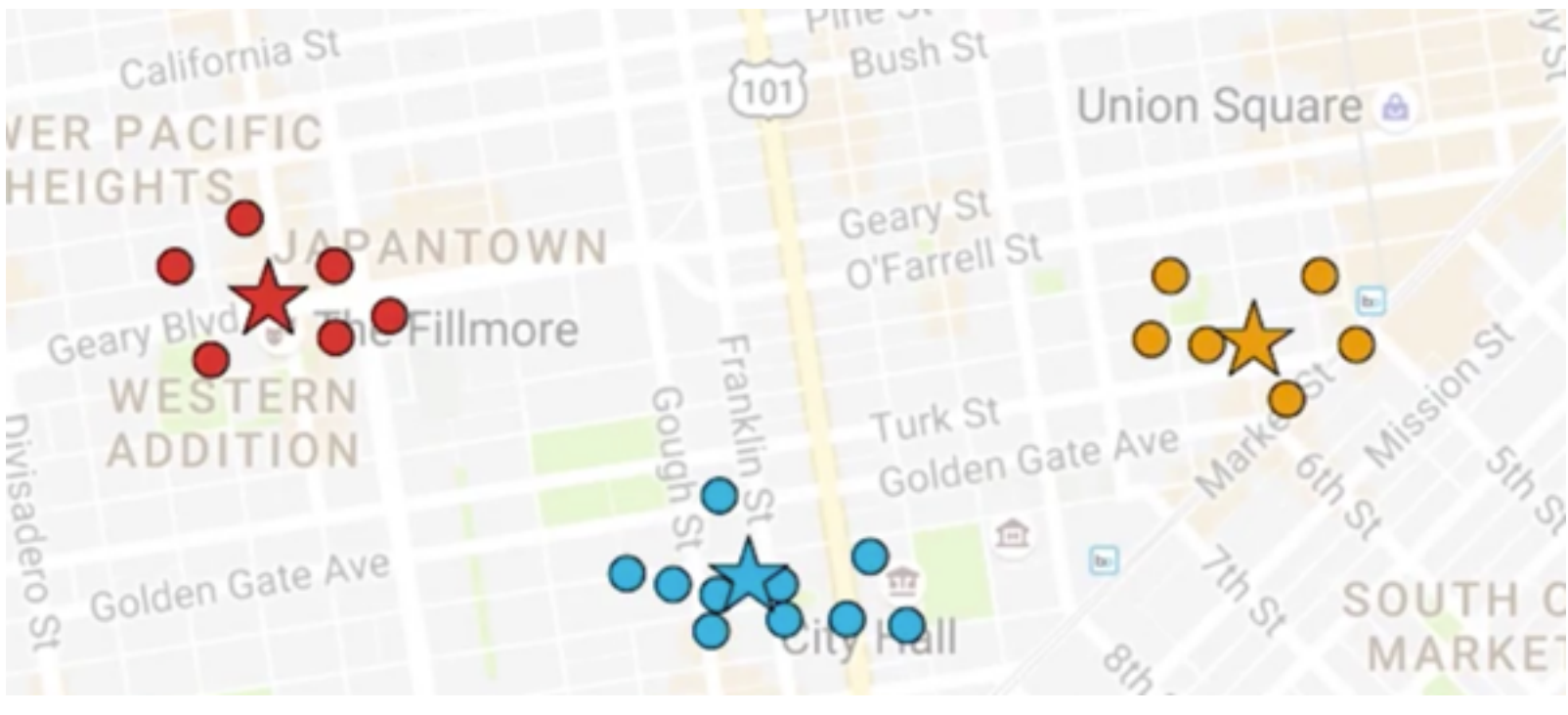
贝叶斯是一个定理，它的意思是：当你不能准确知悉一个事物的本质时，你可以依靠与事物特定本质相关的事件出现的多少去判断其本质属性的概率。

比如说，我们要识别一封邮件是不是垃圾邮件。我们随机挑选出100封垃圾邮件，分析它的特征，我们发现“便宜”这个词出现的频率很高，100封垃圾邮件里，有40封出现了这个词。那我们就以这个认知为依据，得出结论：如果出现了“便宜”，那这封邮件有40%的概率是垃圾邮件。

当我们找到若干个这样的特征，然后用这些特征进行组合后，可以对某些邮件进行判断，它是垃圾邮件的概率超过了我们设定的阈值，我们就自动把这些邮件过滤掉，减少用户受到的打扰。这就是大部分垃圾邮件过滤的原理。

6，聚类

聚类是一种非监督学习的方式。简单的说，就是通过不断的迭代计算，把数据分成若干个组，使得这个组里的都是类似的数据，而不同组之间的数据是不相似的。



聚类

7，强化学习

在没有给出任何答案的情况下，先进行一些尝试，通过尝试所得到的回报，来确定这个尝试是否正确，由一系列的尝试来不断调整和优化算法，最后算法知道在某种情况下，采取何种动作可以得到最好的结果。他的本质是解决“决策问题”，就是通过不断做出决策并获得结果反馈后，学会自动进行决策，得到最优结果。比如上面说过的猴子“学会”做算术题的过程。

8，集成学习

我们在做机器学习的时候，希望能做出各个方面表现都比较好的模型。但常常现实是我们的模型是有偏好的，可能只对某一些情况效果比较好，这个时候我们就希望把若干个这样的模型组合起来，得到一个更好更全面的模型，这种方法，就叫做集成学习。