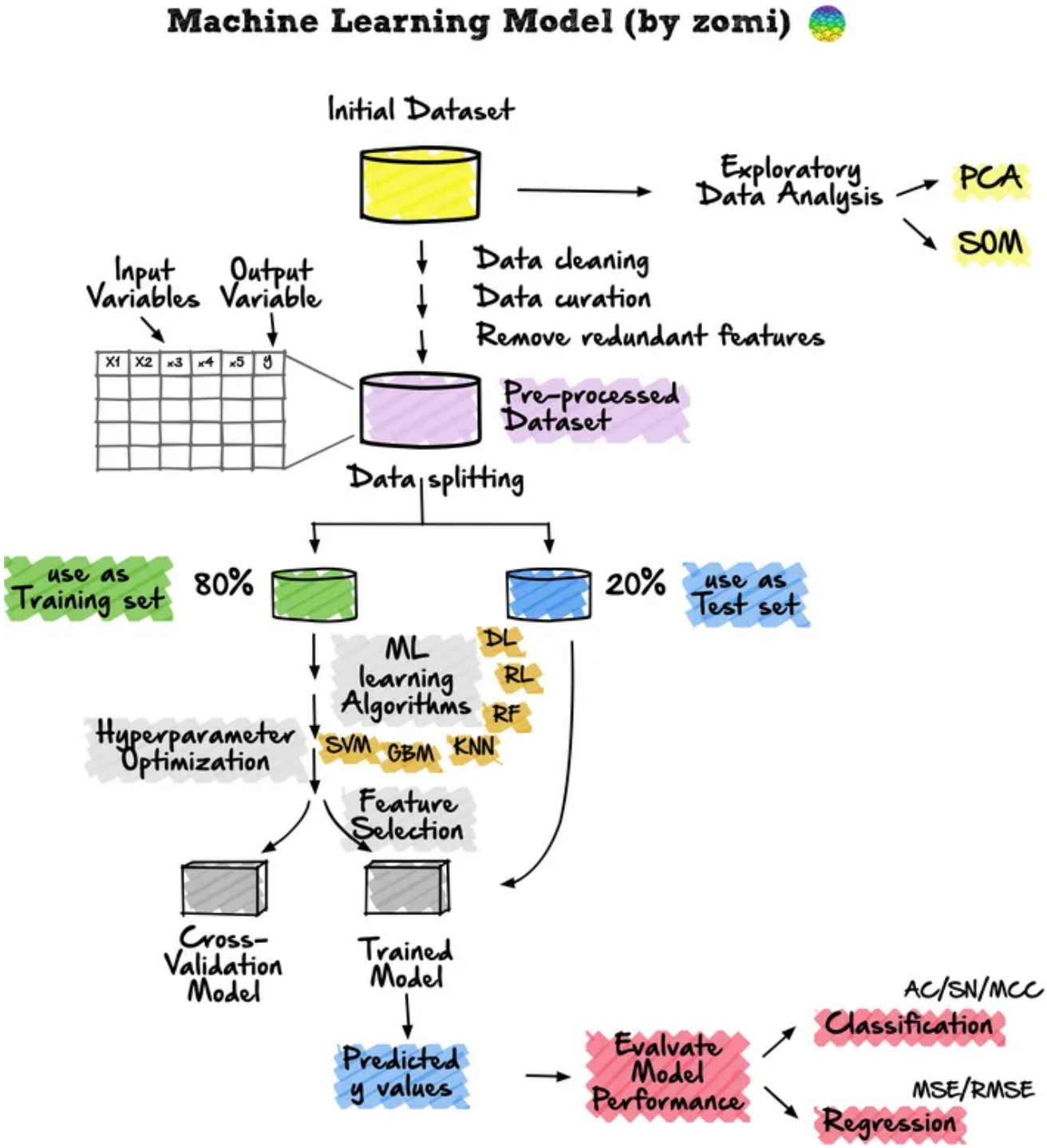
作者：知乎用户

链接：https://www.zhihu.com/question/58339949/answer/2305409944

来源：知乎

1、机器学习的算法流程

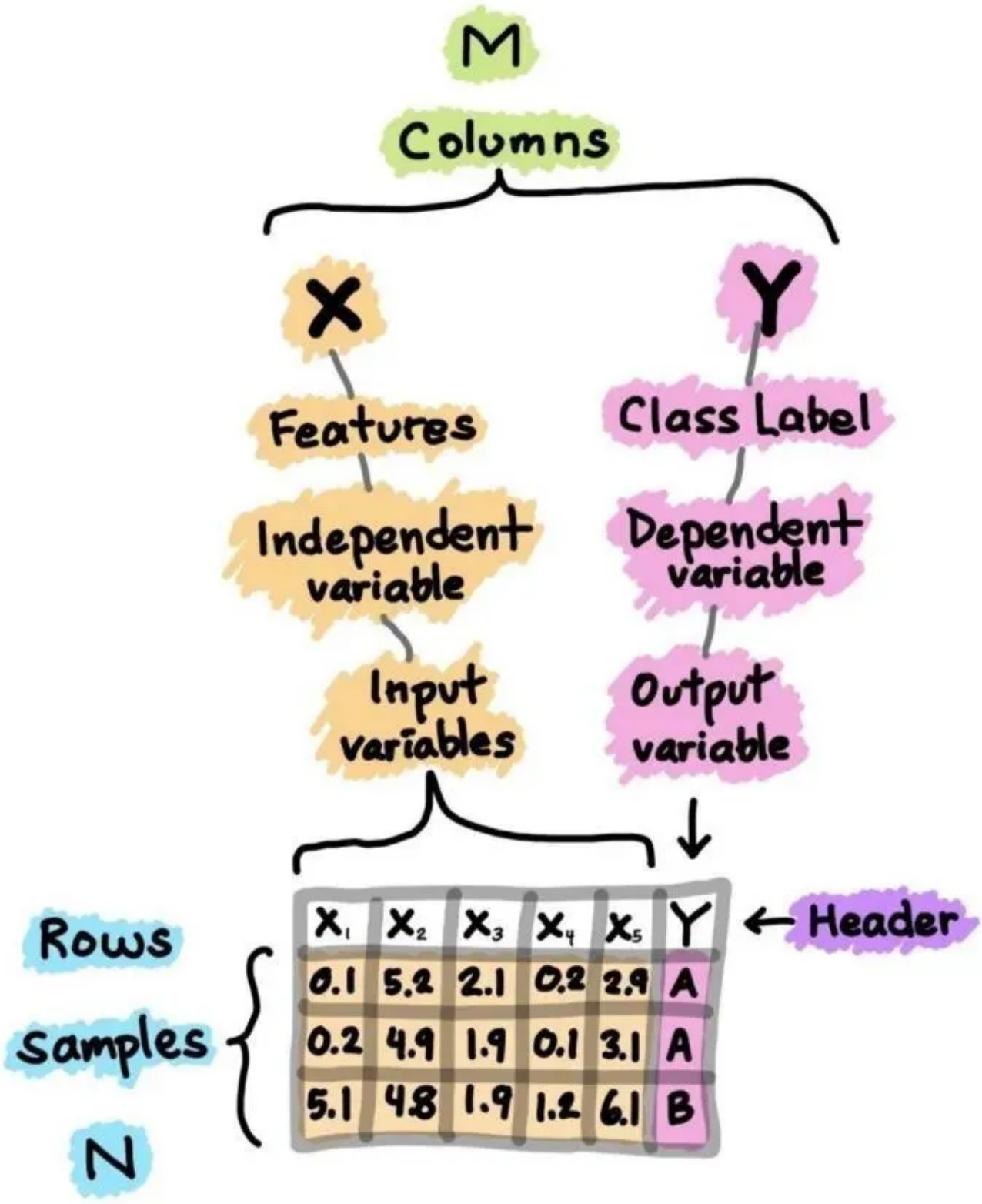
实际上机器学习研究的就是数据科学（听上去有点无聊），下面是机器学习算法的主要流程：主要从1）数据集准备、2）探索性的对数据进行分析、3）数据预处理、4）数据分割、5）机器学习算法建模、6）选择机器学习任务，当然到最后就是评价机器学习算法对实际数据的应用情况如何。



1.1 数据集

首先我们要研究的是数据的问题，数据集是构建机器学习模型流程的起点。简单来说，数据集本质上是一个M×N矩阵，其中M代表列（特征），N代表行（样本）。

列可以分解为X和Y，X是可以指特征、独立变量或者是输入变量。Y也是可以指类别标签、因变量和输出变量。



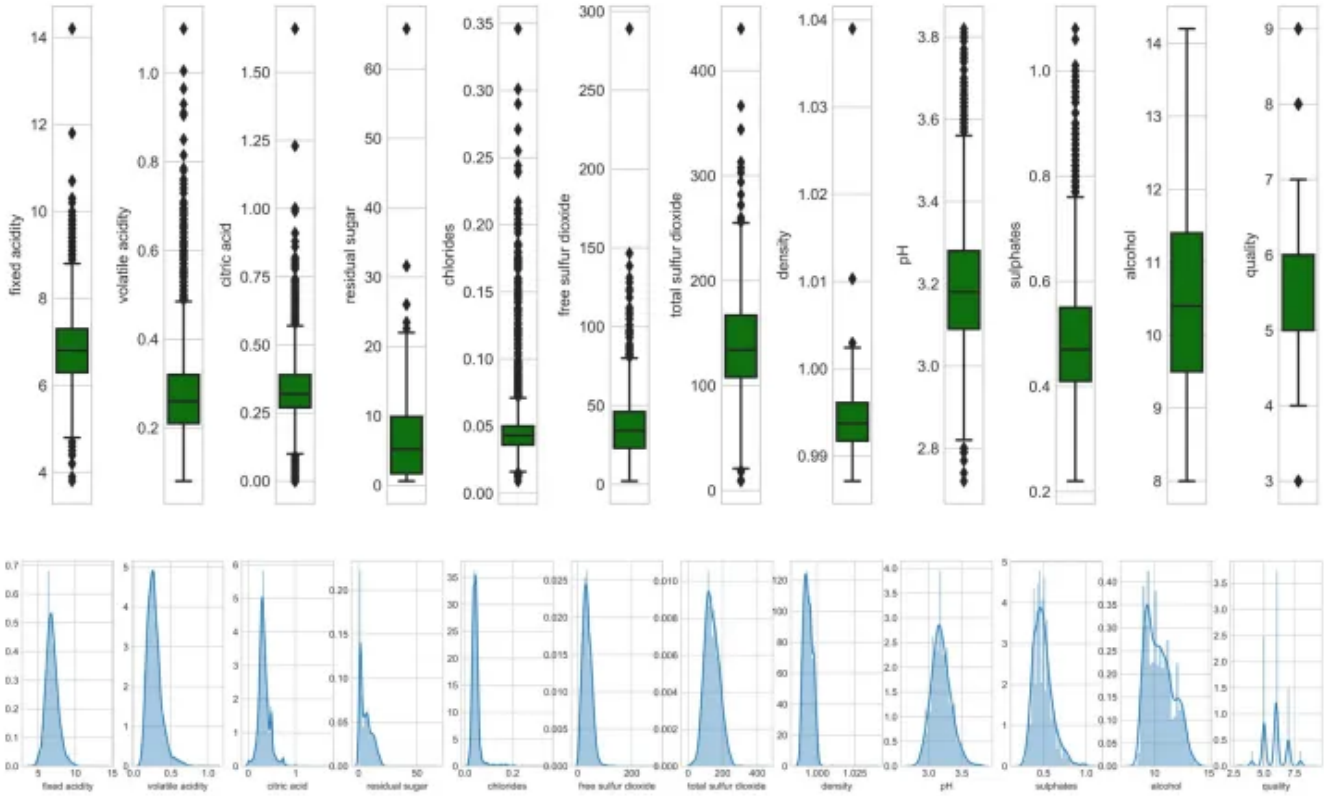
1.2 数据分析

进行探索性数据分析（Exploratory data analysis, EDA）是为了获得对数据的初步了解。EDA主要的工作是：对数据进行清洗，对数据进行描述（描述统计量，图表），查看数据的分布，比较数据之间的关系，培养对数据的直觉，对数据进行总结等。

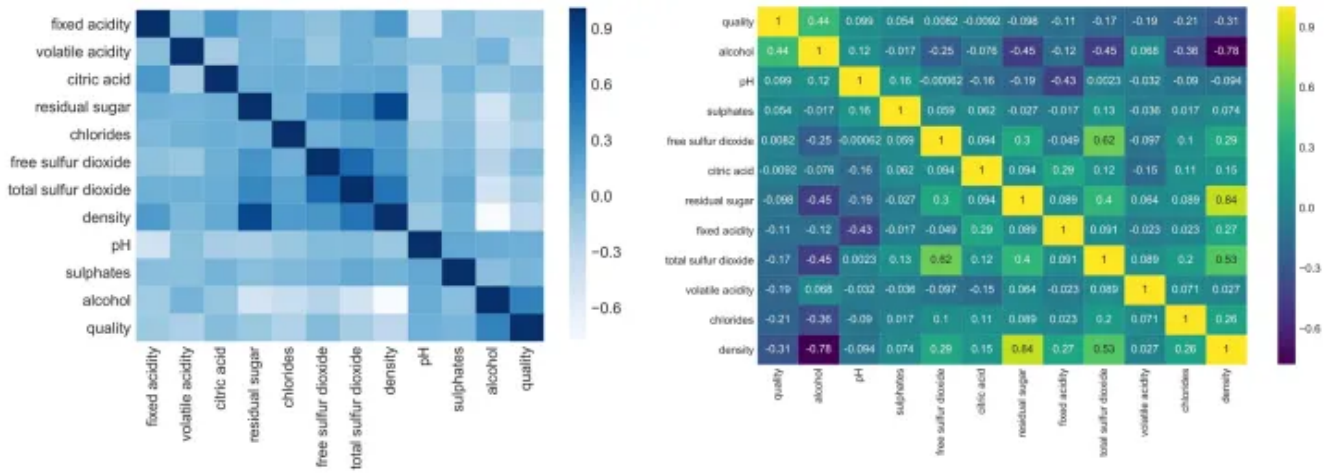
> 探索性数据分析方法简单来说就是去了解数据，分析数据，搞清楚数据的分布。主要注重数据的真实分布，强调数据的可视化，使分析者能一目了然看出数据中隐含的规律，从而得到启发，以此帮助分析者找到适合数据的模型。

在一个典型的机器学习算法流程和数据科学项目里面，我做的第一件事就是通过 "盯住数据"，以便更好地了解数据。个人通常使用的三大EDA方法包括：

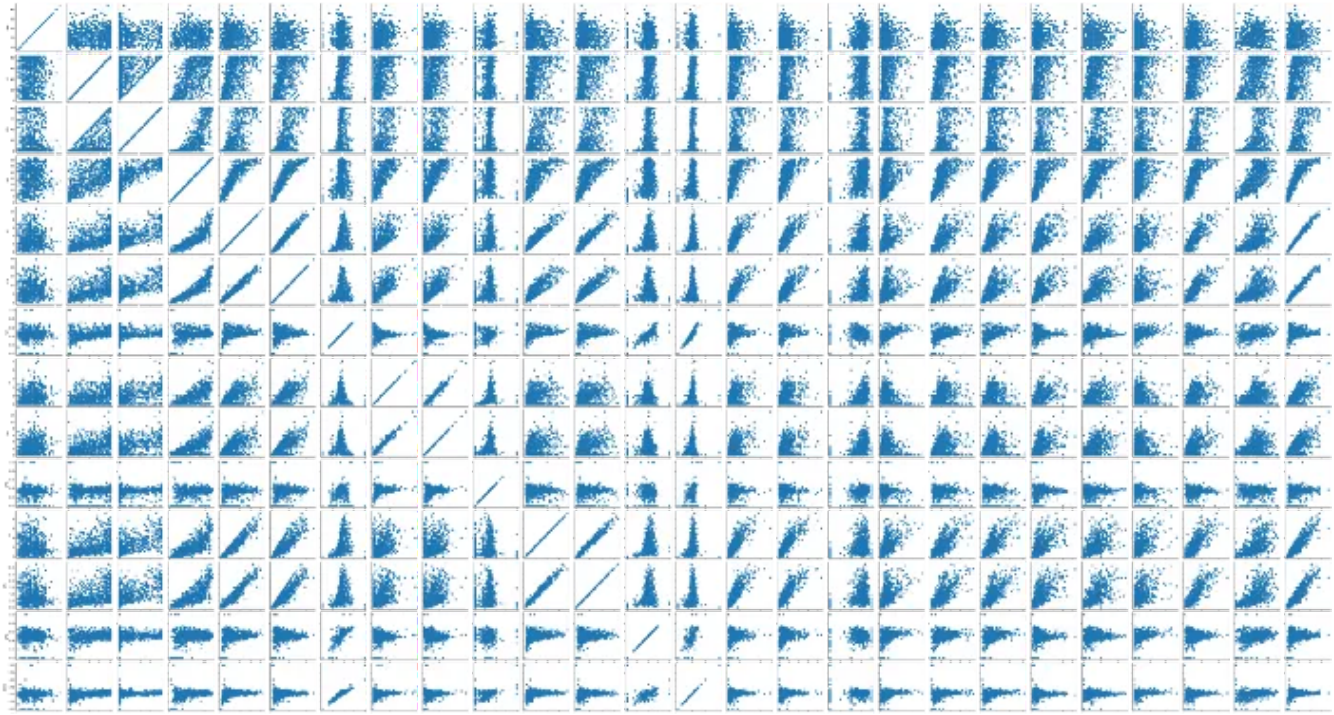
- **描述性统计**：平均数、中位数、模式、标准差。



- **数据可视化**：热力图（辨别特征内部相关性）、箱形图（可视化群体差异）、散点图（可视化特征之间的相关性）、主成分分析（可视化数据集中呈现的聚类分布）等。



- **数据整形**：对数据进行透视、分组、过滤等。



1.3 数据预处理

数据预处理，其实就是对数据进行清理、数据整理或普通数据处理。指对数据进行各种检查和校正过程，以纠正缺失值、拼写错误、使数值正常化/标准化以使其具有可比性、转换数据(如对数转换)等问题。

> 例如对图像进行resize成统一的大小或者分辨率。

数据的质量将对机器学习算法模型的质量好坏产生很大的影响。因此，为了达到最好的机器学习模型质量，传统的机器学习算法流程中，其实很大一部分工作就是在对数据进行分析和处理。

一般来说，数据预处理可以轻松地占到机器学习项目流程中80%的时间，而实际的模型建立阶段和后续的模型分析大概仅占到剩余的20%。

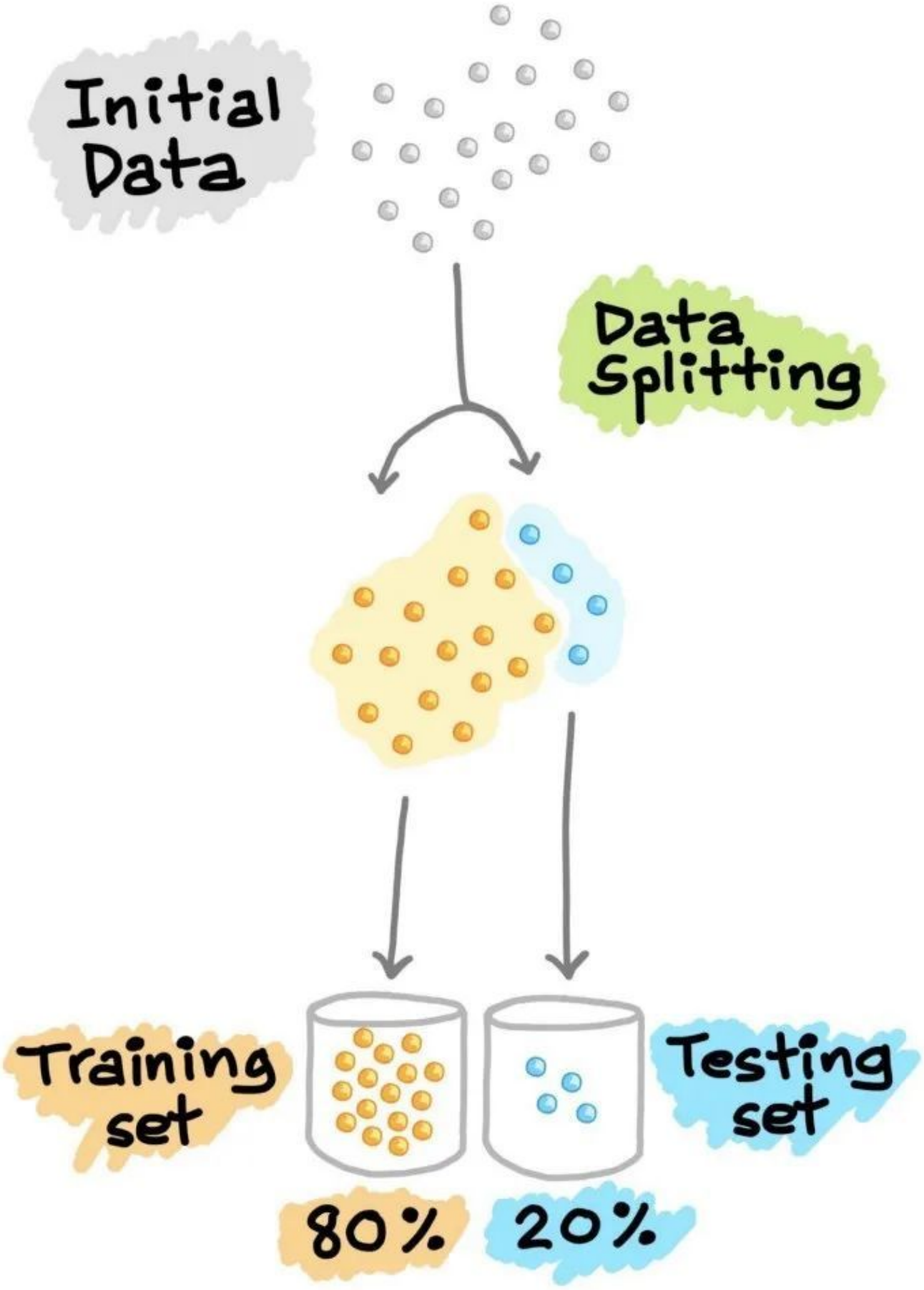
1.4 数据分割

训练集 & 测试集

在机器学习模型的开发流程中，希望训练好的模型能在新的、未见过的数据上表现良好。为了模拟新的、未见过的数据，对可用数据进行数据分割，从而将已经处理好的数据集分割成2部分：训练集合测试集。

第一部分是较大的数据子集，用作训练集（如占原始数据的80%）；第二部分通常是较小的子集，用作测试集（其余20%的数据）。

接下来，利用训练集建立预测模型，然后将这种训练好的模型应用于测试集（即作为新的、未见过的数据）上进行预测。根据模型在测试集上的表现来选择最佳模型，为了获得最佳模型，还可以进行超参数优化。

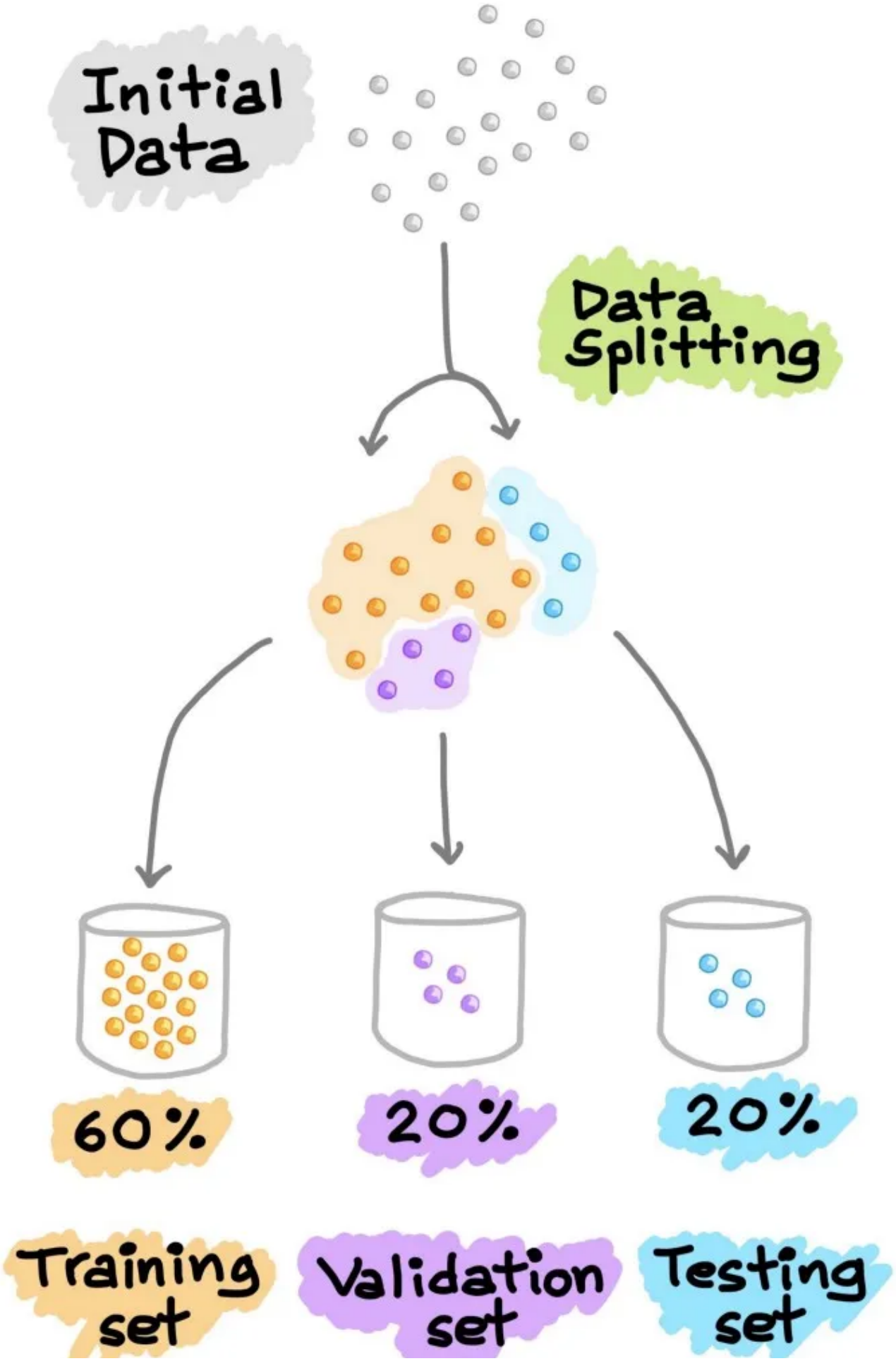


训练集 & 验证集 & 测试集

另一种常见的数据分割方法是将数据分割成3部分：1）训练集，2）验证集和3）测试集。

训练集用于建立预测模型，同时对验证集进行评估，据此进行预测，可以进行模型调优（如超参数优化），并根据验证集的结果选择性能最好的模型。

验证集的操作方式跟训练集类似。不过值得注意的是，测试集不参与机器学习模型的建立和准备，是机器学习模型训练过程中单独留出的样本集，用于调整模型的超参数和对模型的能力进行初步评估。通常边训练边验证，这里的验证就是用验证集来检验模型的初步效果。


Initial Data
Data Splitting
60%
20%
20%
Training set
Validation set
Testing set

交叉验证

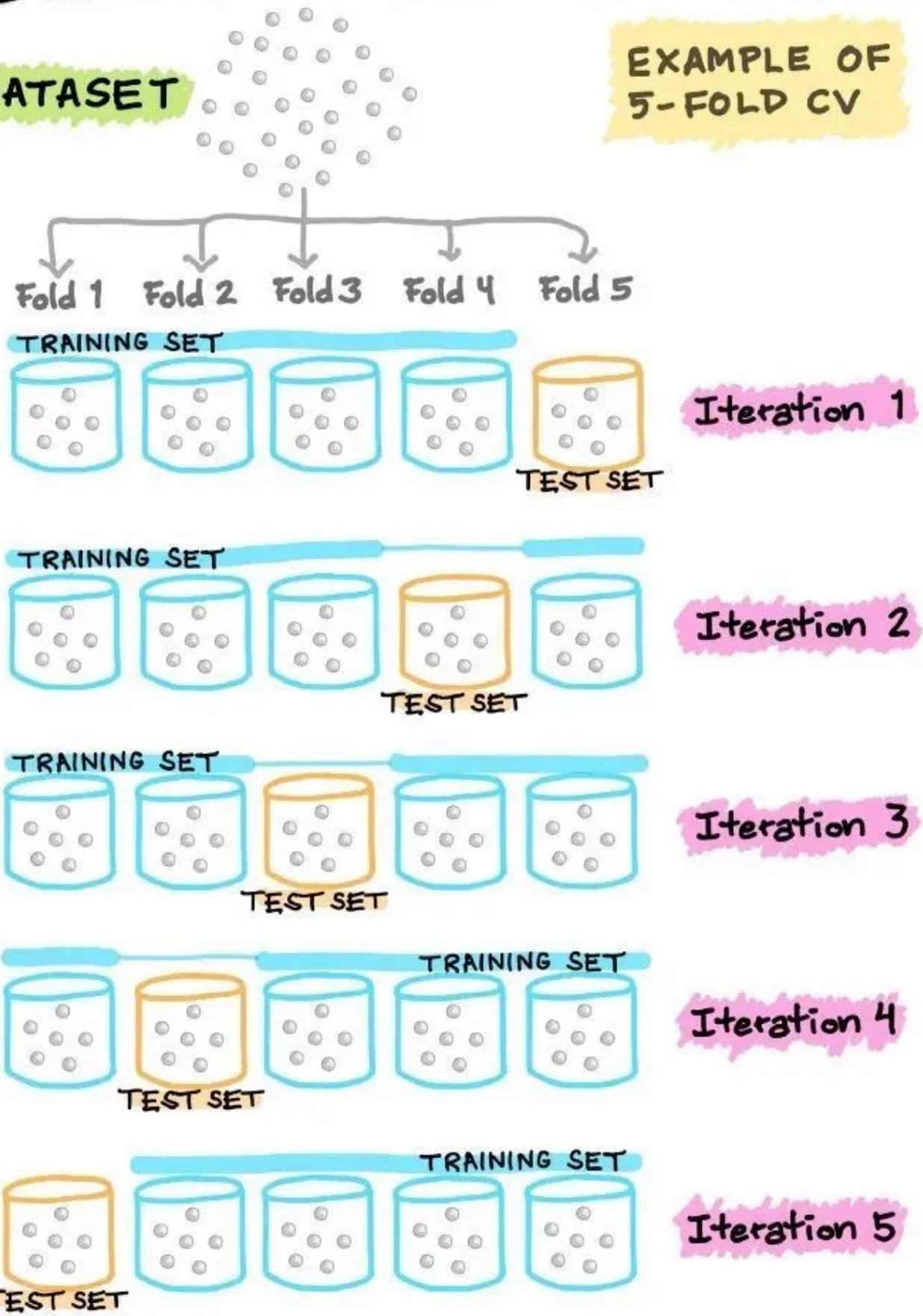
实际上数据是机器学习流程中最宝贵的，为了更加经济地利用现有数据，通常使用N倍交叉验证，将数据集分割成N个。在这样的N倍数据集中，其中一个被留作测试数据，而其余的则被用作建立模型的训练数据。通过反复交叉迭代的方式来对机器学习流程进行验证。

> 这种交叉验证的方法在机器学习流程中被广泛的使用，但是深度学习中使用得比较少哈。

CROSS-VALIDATION

DATASET

EXAMPLE OF 5-FOLD CV



1.5 机器学习算法建模

下面是最有趣的部分啦，数据筛选和处理过程其实都是很枯燥乏味的，现在可以使用精心准备的数据来建模。根据taget变量（通常称为Y变量）的数据类型，可以建立一个分类或回归模型。

机器学习算法

机器学习算法可以大致分为以下三种类型之一：

- **监督学习**：是一种机器学习任务，建立输入X和输出Y变量之间的数学（映射）关系。这样的(X、Y)对构成了用于建立模型的标签数据，以便学习如何从输入中预测输出。
- **无监督学习**：是一种只利用输入X变量的机器学习任务。X变量是未标记的数据，学习算法在建模时使用的是数据的固有结构。
- **强化学习**：是一种决定下一步行动方案的机器学习任务，它通过试错学习（trial and error learning）来实现这一目标，努力使reward回报最大化。

参数调优

传说中的调参侠主要干的就是这个工作啦。超参数本质上是机器学习算法的参数，直接影响学习过程和预测性能。由于没有万能的超参数设置，可以普遍适用于所有数据集，因此需要进行超参数优化。

> 以随机森林为例。在使用randomForest时，通常会对两个常见的超参数进行优化，其中包括mtry和ntree参数。mtry（maxfeatures）代表在每次分裂时作为候选变量随机采样的变量数量，而ntree（nestimators）代表要生长的树的数量。

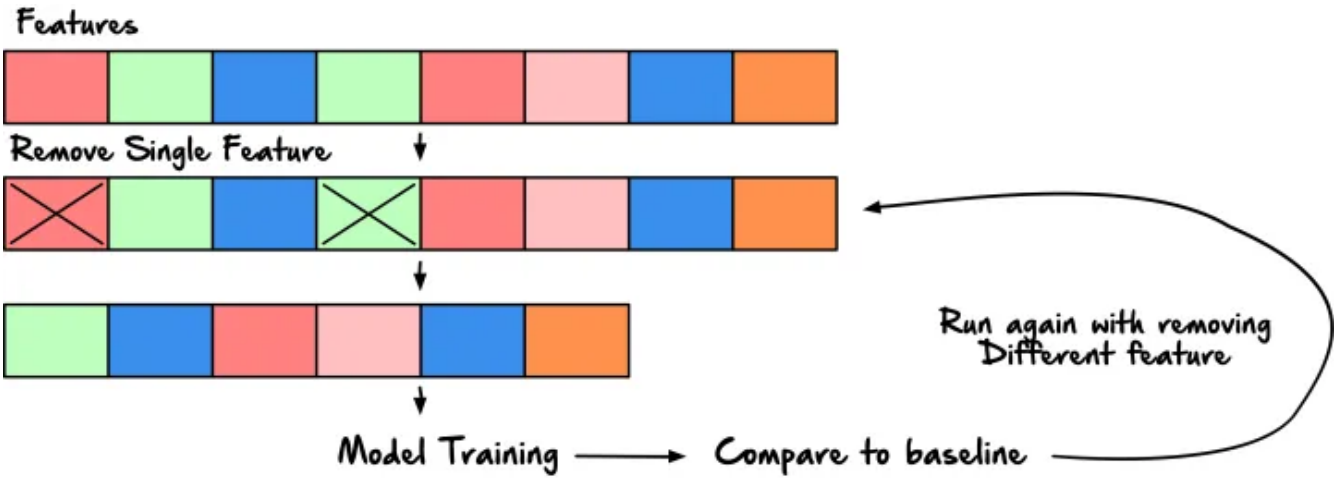
另一种在10年前仍然非常主流的机器学习算法是支持向量机SVM。需要优化的超参数是径向基函数(RBF)内核的C参数和gamma参数。C参数是一个限制过拟合的惩罚项，而gamma参数则控制RBF核的宽度。

调优通常是为了得出超参数的较佳值集，很多时候不要去追求找到超参一个最优值，其实调参侠只是调侃调侃，真正需要理解掌握算法原理，找到适合数据和模型的参数就可以啦。

特征选择

特征选择从字面上看就是从最初的大量特征中选择一个特征子集的过程。除了实现高精度的模型外，机器学习模型构建最重要的一个方面是获得可操作的见解，为了实现这一目标，能够从大量的特征中选择出重要的特征子集非常重要。

特征选择的任务本身就可以构成一个全新的研究领域，在这个领域中，大量的努力都是为了设计新颖的算法和方法。从众多可用的特征选择算法中，一些经典的方法是基于模拟退火和遗传算法。除此之外，还有大量基于进化算法（如粒子群优化、蚁群优化等）和随机方法（如蒙特卡洛）的方法。

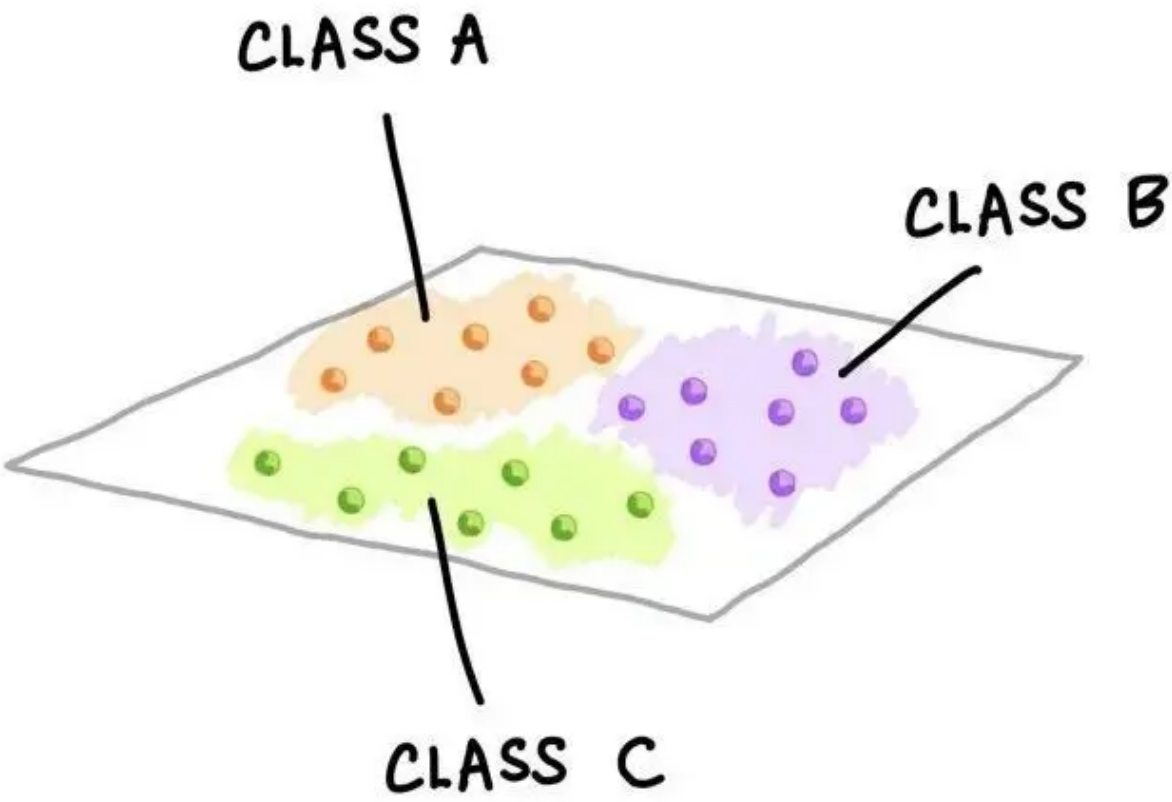


1.6 机器学习任务

在监督学习中，两个常见的机器学习任务包括分类和回归。

分类

一个训练好的分类模型将一组变量作为输入，并预测输出的类标签。下图是由不同颜色和标签表示的三个类。每一个小的彩色球体代表一个数据样本。三类数据样本在二维中的显示，这种可视化图可以通过执行PCA分析并显示前两个主成分（PC）来创建；或者也可以选择两个变量的简单散点图可视化。

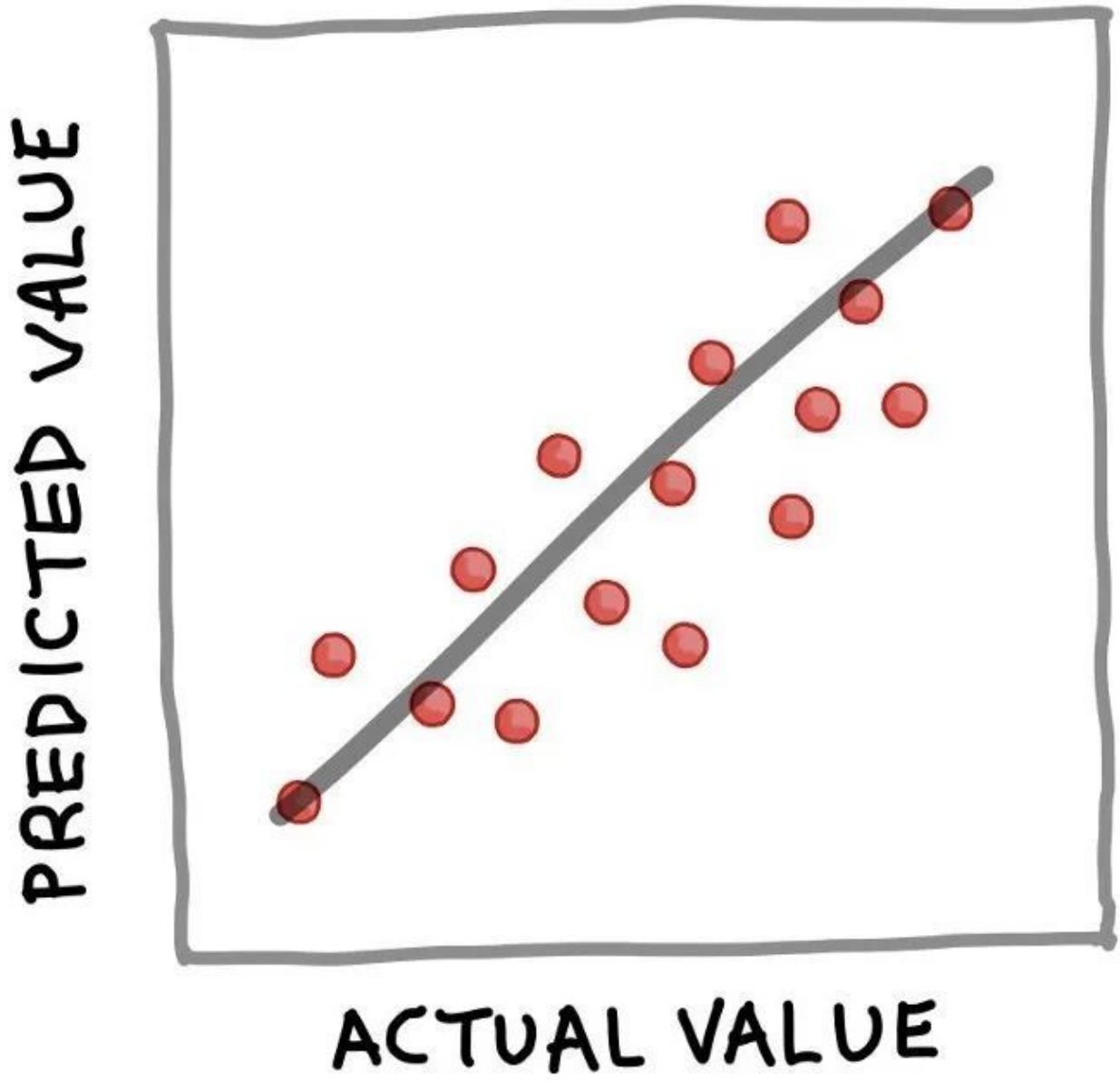


性能指标

如何知道训练出来的机器学习模型表现好或坏？就是使用性能评价指标（metrics），一些常见的评估分类性能的指标包括准确率（AC）、灵敏度（SN）、特异性（SP）和马太相关系数（MCC）。

回归

最简单的回归模式，可以通过以下简单等式很好地总结：Y = f(X)。其中，Y对应量化输出变量，X指输入变量，f指计算输出值作为输入特征的映射函数（从机器学习模型中得到）。上面的回归例子公式的实质是，如果X已知，就可以推导出Y。一旦Y被计算（预测）出来，一个流行的可视化方式是将实际值与预测值做一个简单的散点图，如下图所示。

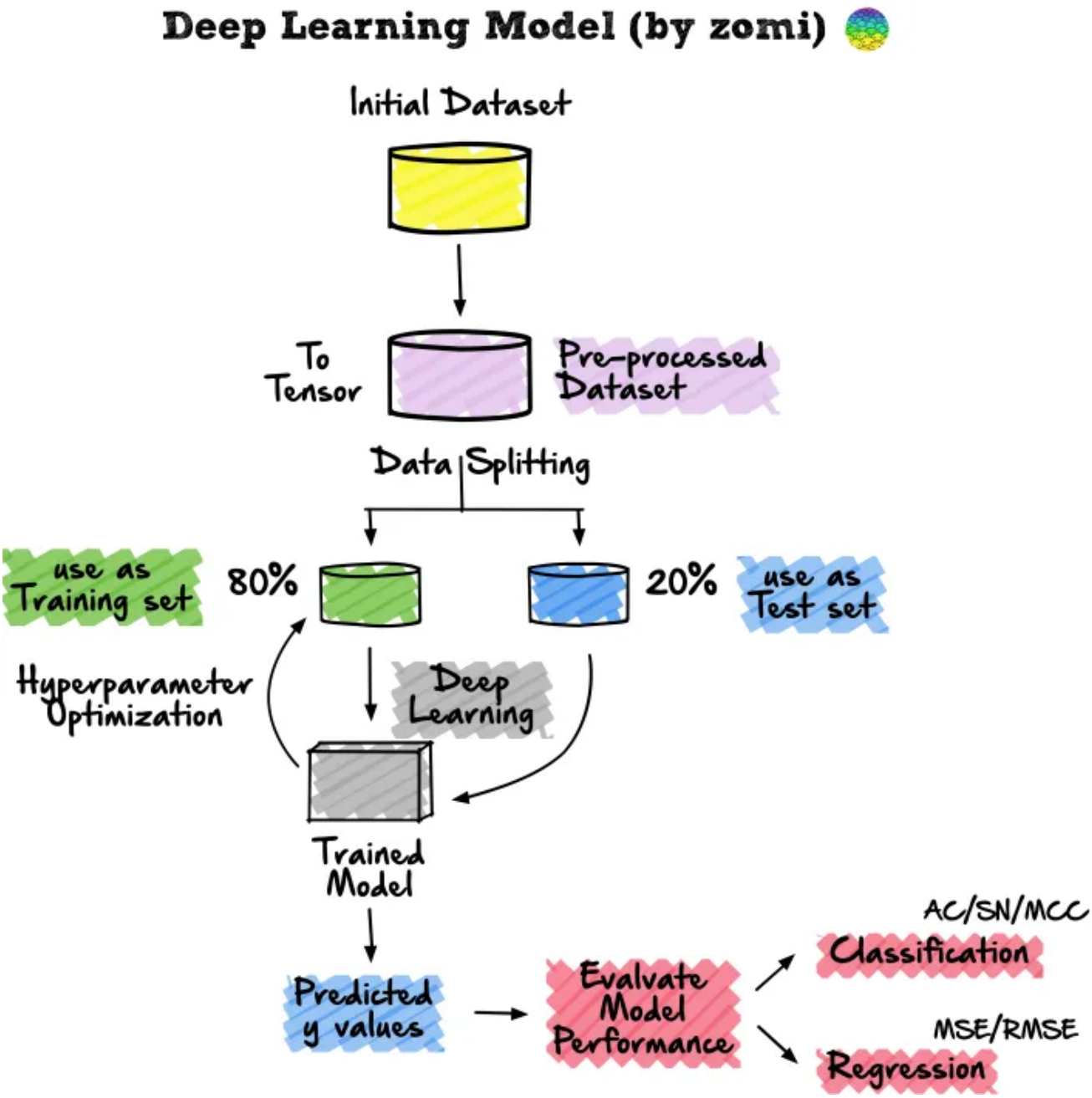


对回归模型的性能进行评估，以评估拟合模型可以准确预测输入数据值的程度。评估回归模型性能的常用指标是确定系数（R^2）。此外，均方误差（MSE）以及均方根误差（RMSE）也是衡量残差或预测误差的常用指标。

2、深度学习算法流程

深度学习实际上是机器学习中的一种范式，所以他们的主要流程是差不多的。深度学习则是优化了数据分析，建模过程的流程也是缩短了，由神经网络统一了原来机器学习中百花齐放的算法。

在深度学习正式大规模使用之前呢，机器学习算法流程中药花费很多时间去收集数据，然后对数据进行筛选，尝试各种不同的特征提取机器学习算法，或者结合多种不同的特征对数据进行分类和回归。



下面是机器学习算法的主要流程：主要从1）数据集准备、2）数据预处理、3）数据分割、4）定义神经网络模型，5）训练网络。

深度学习不需要我们自己去提取特征，而是通过神经网络自动对数据进行高维抽象学习，减少了特征工程的构成，在这方面节约了很多时间。

但是同时因为引入了更加深、更复杂的网络模型结构，所以调参工作变得更加繁重啦。例如：定义神经网络模型结构、确认损失函数、确定优化器，最后就是反复调整模型参数的过程。