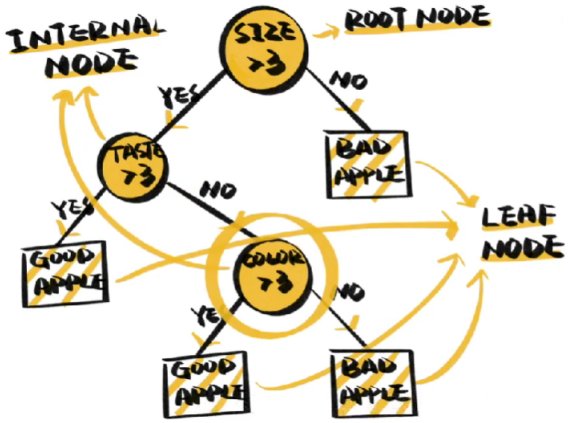
什么是决策树

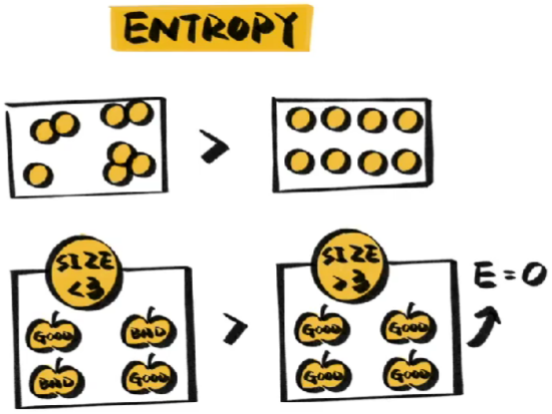
在游戏中遇到敌人是选择攻击还是逃跑？如果选择攻击，是选择普通的物理攻击还是魔法攻击？为达到目标根据一定的条件进行选择的过程，就是决策树(DT Tree)。

决策树模型非常经典，在机器学习中常被用于分类，构成它的元素是节点和边，节点会根据样本的特征做出判断，最初的分支点被称为**根节点**，其余的被称为**子节点**，不再有分支的节点则被称为**叶子节点**，他们代表样本的分类结果，边则指示着方向。

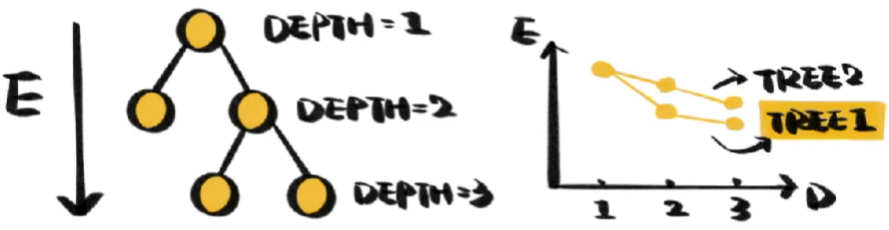


特征这么多，谁来做根结点?接下来的节点又该是什么?

为了构建决策树，人们找到了一个衡量标准，在热力学中，熵被用来描述一个系统内在的混乱程度，在决策树中，熵代表的是分支下样本种类的丰富性，样本种类越多越混乱，熵就越大。如果分支下的样本完全属于同一类，熵就等于0。



构造树的基本思路是随着树的深度，也就是层数的增加，让熵快速降低，熵降低的速度越快，代表决策树分类效率越高。



决策树最大的优点是天然的可解释性，苹果之所以是好苹果，是因为它又大又红又甜，它的缺点也很明显。

但是数据都是有特例的，如果一棵树能将训练样本完美分类，那它一定是过拟合的。

解决方法很简单，去掉一些分支，**剪枝**(Pruning)有两种

1. 预剪枝是在训练开始前规定条件，比如树达到某一深度就停止训练。
2. 后剪枝则是先找到树，再依据一定条件，如限制叶子结点的个数，去掉一部分分支。