半监督学习（**Semi-supervised Learning**）：**训练集同时包含有标记样本数据和未标记样本数据**，**不需要人工干预**，让学习器不依赖外界交互、自动地利用未标记样本来提升学习性能，就是半监督学习。

半监督学习为什么可行？这个就要了解一下半监督学习中的理论前提，即模型假设，当模型假设正确时，无标签的样例能够帮助改进学习性能。而理论前提，就是我们之前在提到过的相似性原理，相似性聚簇，**在现实世界中普遍存在，在晚上，在中国大片土地上用灯光的聚簇就可以定位大城市、特大城市。**

**实现半监督算法的概率视角**

半监督生成模型是以概率视角求解半监督算法的。该方法假设所有数据（无论是否有标记）都是由同一潜在的模型生成的，分布类型相同。通过这个假设潜在模型的参数将未标记数据与学习目标联系起来，而未标记数据的标记可作为模型的缺失参数，通常基于EM算法进行极大似然估计求解。生成式方法的重点是对于生成式模型的假设，不同的模型假设将产生不同的方法。

这个方法的关键也就是这个模型假设必须准确，即假设的生成式模型必须与真实数据分布吻合；否则利用未标记数据反倒会降低泛化性能。 这个不好控制，也是好多学习机器学习的人比较讨厌的，但这是错误的！模型的构建过程允许试错，也就是当结果不是很理想，可以反思先验是否准确，继而，修正先验，继续迭代模型。