应用机器学习前，我们建模使用的数据集可以从以下渠道获得

- 从实验室购买收费数据集，一般是卖硬盘，出自各个大数据实验室，大多已经是对方爬好的数据集
- 通过爬虫挖掘关键词数据集，这一步工作需要自行编写爬虫程序
- 下载免费数据集，这些数据集大都是和论文捆绑验证准确率用的，为了争取高额的论文精度奖金，数据集往往被深入提炼，nGB，nT，这些数据集太过于大众和日常，特殊用途有限，多用于验证算法方案。
- 自行使用电子设备拍照制作数据集，这是最简单的定制化建模，一手一脚从 0 开始构建，我以为这是最有实用价值的，你对业务理解越深刻，越容易针对建模，拍照工作并不难做，几千张样本数据集，几天间收集完成。
- 通过业务手段从商业公司获取大数据集，一般是没价值的，要么就是个人名义秘密兜售。
- 非法数据集，黑客，人名义秘密兜售