今天，构造 AI 或机器学习系统比以往任何时候都更加容易。我们有许多开源的最前沿的工具，如 TesorFlow，Torch，Spark 等，也有 AWS、Google Cloud 以及其他云服务提供商提供的大量计算力，这意味着你可以悠哉地一边喝着咖啡一边用 laptop 训练模型。

虽然不算人工智能这列火车的车头，但 AI 革命的幕后英雄是数据——得益于各大研究机构和企业的辛苦工作，我们有机会获取大量有标签和注释的数据。这些研究机构和企业也认识到，数据的民主化是加速 AI 发展的一个必要步骤。

然而，大多数包含机器学习或 AI 的产品都严重依赖非公开的专有数据集。因此，很难判定哪些公开数据集有用。

重要的是，在数据集上表现得性能良好并不能保证机器学习系统在真实的产品场景中表现良好。许多搞 AI 的人忘记了构建新 AI 解决方案或开发产品的最难的部分不是 AI 本身或者算法，而是数据的收集和标记。标准数据集可以用于验证模型，或作为构建更加定制化的解决方案的一个好的起点。

以下是我们精心收集的一些非常好的开放数据集，也是做 AI 研究不容错过的数据集。

标签解释



【经典】这些是在 AI 领域中非常著名、众所周知的数据集。很少有研究者或工程师没有听说过它们。

【有用】这些是更加接近现实世界的、精心设计的数据集。而且，这些数据集通常在产品和研发两方面都有用。

【学术】这些是在机器学习和 AI 的学术研究中通常作为基准或基线使用的数据集。无论好坏，研究人员都使用这些数据集来验证算法。

【陈旧】这些数据集，无论是否实用，已经有相当长历史了。

计算机视觉

【学术、经典、陈旧】**MNIST**：最常用的完整性检查数据集，图像大小为25x25的B＆W手写数字，但在 MNIST 上性能良好，并不意味着模型本身很好。

地址：http://pjreddie.com/projects/mnist-in-csv/

【经典、陈旧】**CIFAR 10 & CIFAR 100**：32x32的彩色图像数据集，虽然已经不常用，但也可以用作完整性检查。

地址：https://www.cs.toronto.edu/~kriz/cifar.html

【有用、学术、经典】**ImageNet**：新算法实际上使用的图像数据集，很多图像 API 公司从其 REST 接口获取标签，这些标签被怀疑与 ImageNet 的下一级 WordNet 的 1000 个类很相似。

地址：http://image-net.org/

**LSUN**：用于场景理解和多任务辅助（房间布局估计，显着性预测等）。

地址：http://lsun.cs.princeton.edu/2016/

【学术】**PASCAL VOC**：一个通用的图像分割/分类数据集，对构建真实图像的注释用处不是特别大，但对于基线很有用。

地址：http://host.robots.ox.ac.uk/pascal/VOC/

【学术】**SVHN**：数据来源于 Google 街景视图中的房屋数量，可以用作野外的周期性 MNIST。

地址：http://ufldl.stanford.edu/housenumbers/

**MS COCO**：一个通用的图像理解/字幕数据集。

地址：http://mscoco.org/

【有用】**Visual Genome**：非常详细的视觉知识数据集，包含约100K图像的深字母。

地址：http://visualgenome.org/

【有用、学术、经典、陈旧】**Labeled Faces in the Wild**：使用名称标识符标记的面部区域数据集，常用于训练面部识别系统。

地址：http://vis-www.cs.umass.edu/lfw/

自然语言处理

【有用、学术】**Text Classification Datasets**：一个文本分类数据集，包含8个可用于文本分类的子数据集，样本大小从120K到3.6M，问题范围从2级到14级，数据来源于 DBPedia、Amazon、Yelp、Yahoo!、Sogou 和 AG。

地址：http://t.cn/RJDVxr4

【有用、学术】**WikiText**：由 Salesforce MetaMind 设计的大型语言建模语料库，来源于维基百科文章。

地址：http://t.cn/RJDVSRy/

【有用】**Question Pairs**：第一个来源于 Quora 的包含重复/语义相似性标签的数据集。

地址：https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

【有用、学术】**SQuAD**：斯坦福大学的问答数据集，广泛用于问题回答和阅读理解，其中每个问题和答案都是文本片段的形式。

地址：https://rajpurkar.github.io/SQuAD-explorer/

**CMU Q/A Dataset**：人工生成的问题/答案对，难度评级来自维基百科文章。

地址：http://www.cs.cmu.edu/~ark/QA-data/

【有用】**Maluuba Datasets**：用于状态性的自然语言理解研究的人工制作的精细数据集。

地址：https://datasets.maluuba.com/

【有用、学术】**Billion Words**：一个大型、通用的语言建模数据集，常用于如 word2vec 或 Glove 的分布式词语表征。

地址：http://www.statmt.org/lm-benchmark/

【有用、学术】**Common Crawl**：Petabyte 级规模的网络爬行数据集，常用于学习词嵌入。

地址：http://commoncrawl.org/the-data/

【学术、经典】**bAbi**：来自 FAIR 的阅读理解和问答应答数据集。

地址：https://research.fb.com/projects/babi/

【学术】**The Children’s Book Test**：从古登堡计划的童书中提取的（问题+上下文，答案）的基线，该数据集对问题回答、阅读理解和模拟陈述有用。

地址：https://research.fb.com/projects/babi/

【学术、经典、陈旧】**Stanford Sentiment Treebank**：一个标准情感数据集，数据集中每个句子解析树的每个节点都有精细的情感注释。

地址：http://nlp.stanford.edu/sentiment/code.html

【经典、陈旧】**20 Newsgroups**：一个文本分类的经典数据集，通常用于纯分类或作为任何 IR／索引算法的基准。

地址：http://qwone.com/~jason/20Newsgroups/

【经典、陈旧】**Reuters**：一个较旧，完全基于分类的新闻文本数据集，常用于教程。

地址：http://t.cn/RJDfi7T

【经典、陈旧】**IMDB**：一个比较旧，规模也相对较小的二院情感分类数据集。

地址：http://ai.stanford.edu/~amaas/data/sentiment/

【经典、陈旧】**UCI’s Spambase**：这是一个年代较久远的、经典的垃圾电子邮件数据集，来源是著名的 UCI 机器学习库。由于该数据集在设计细节上的独特之处，可以用作学习个性化垃圾邮件过滤的一个有趣的基线。

地址：https://archive.ics.uci.edu/ml/datasets/Spambase

语音



大多数语音识别数据集是专有的，因为这些数据对于创建该数据集的公司来说具有很大价值。因此，这部分的可用公开数据集多数比较陈旧。

【学术、陈旧】**2000 HUB5 English**：仅包含英语的语音数据集，百度最近的论文《深度语音：扩展端对端语音识别》使用的是这个数据集。

地址：https://catalog.ldc.upenn.edu/LDC2002T43

【学术】**LibriSpeech**：包含文本和语音的有声读物数据集，由近500小时的多人朗读的清晰音频组成，且包含书籍的章节结构。

地址：http://www.openslr.org/12/

【有用、学术】**VoxForge**：带口音的语音清洁数据集，对测试模型在不同重音或语调下的鲁棒性非常有用。

地址：http://www.voxforge.org/

【学术、经典、陈旧】**TIMIT**：英文语音识别数据集。

地址：https://catalog.ldc.upenn.edu/LDC93S1

【有用】**CHIME**：包含环境噪音的语音识别挑战赛数据集。该数据集包含真实、模拟和清洁的语音录音，具体来说，包括4个扬声器在4个有噪音环境下进行的将近9000次录音，模拟数据是将多个环境组合及在无噪音环境下记录的数据。

地址：http://spandh.dcs.shef.ac.uk/chime\_challenge/data.html

**TED-LIUM**：TED Talk 的音频数据集，包含1495个TED演讲的录音及全文的文字稿。

地址：http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus

推荐和排序系统

【经典、陈旧】**Netflix Challenge**：第一个主要的 Kaggle 挑战赛数据集，但由于隐私问题，只有非正式的数据集提供。

地址：http://www.netflixprize.com/

【有用、学术、经典】**MovieLens**：多种大小的电影评论数据，通常用于基线协同过滤。

地址：https://grouplens.org/datasets/movielens/

**Million Song Dataset**：Kaggle 上的大型、元数据丰富的开源数据集，对混合推荐系统有用。

地址：https://www.kaggle.com/c/msdchallenge

【有用】**Last.fm**：可访问底层社交网络及其他元数据的音乐推荐数据集，这些元数据对混合系统很有用。

地址：http://grouplens.org/datasets/hetrec-2011/

网络和图表



【学术】**Amazon Co-Purchasing and Amazon Reviews**：亚马逊网站的“买了该产品的用户也买了......”板块的数据，以及相关产品的亚马逊评论数据。适合用于推荐系统。

地址：http://snap.stanford.edu/data/amazon-meta.html

**Friendster Social Network Dataset**：包含103,750,348个 Friendster 用户的好友列表的匿名数据集。

地址：https://archive.org/details/friendster-dataset-201107

地理空间数据

【有用、经典】**OpenStreetMap**：免费许可的全球矢量数据集，包含美国人口普查局的 TIGER数据。

地址：http://wiki.openstreetmap.org/wiki/Planet.osm

【有用】**Landsat8**：卫星拍摄的地球表面照片数据，每隔几周更新一次。

地址：https://landsat.usgs.gov/landsat-8

【有用】**NEXRAD**：多普勒雷达扫描的美国大气环境数据。

地址：https://www.ncdc.noaa.gov/data-access/radar-data/nexrad

结语：

人们常常以为在一个数据集上解决了问题就等同于得到好的产品了。但在使用这些数据集作为验证或概念证明时，不要忘记用更新、更接近现实的数据来测试产品的功能，从而能够作出改进。一个成功的以数据作为驱动力的企业通常能够从他们收集新的、专有的数据的能力中获益，进而提升竞争力。