food	Link	Purpose
20 Newsgroups	http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html	*The text from 20000 messages taken from 20 Usenet newsgroups for text analysis, classification, etc.*
Amazon Reviews	http://jmcauley.ucsd.edu/data/amazon/	*Over 142 million product reviews for sentiment analysis, recommender systems, and more.*
Football Strategy	http://jmcauley.ucsd.edu/data/amazon/	*Thousands of scenarios to make the best coaching decisions.*
Horses for Courses	http://jmcauley.ucsd.edu/data/amazon/	
Human Activity Recognition with Smartphones	http://jmcauley.ucsd.edu/data/amazon/	*Sensor data for recognizing the human activity - walking, sitting, etc.*
Labeled Faces in the Wild	csv	*13,000 named faces for facial recognition. Multiple training and test sets*
National Survey on Drug Use and Health	http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34933	
NORB 3D Object Recognition	http://www.cs.nyu.edu/~ylclab/data/norb-v1.0/	*Binocular images of 50 toy figurines for 3D object recognition from image.*
One Million Songs	http://labrosa.ee.columbia.edu/millionsong/	*Audio features and metadata for a subset (10,000) of the one million popular songs dataset for recognition/classification.*
SMS Spam Collection	http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/	*A collection of 5,574 SMS (text) messages, some spam, some normal, for spam filtering.*
Hate Speech Identification	https://www.crowdflower.com/wp-content/uploads/2016/03/twitter-hate-speech-classifier-DFE-a845520.csv	*A sampling of Twitter posts that have been judged based on whether they are offensive or contain hate speech, as a training set for text analysis.*
Hidden Beauty of Flickr Pictures	http://www.di.unito.it/~schifane/dataset/beauty-icwsm15/	*15,000 Flickr photo IDs that have received ratings based on aesthetics, for image analysis.*
Yahoo Instant Messenger Friends Connectivity Graph	http://webscope.sandbox.yahoo.com/catalog.php?datatype=g	*Connections between Yahoo users who communicate with each other using Yahoo messenger, can be used to identify key social contacts/influencers. Add dataset to cart to access.*
Record of Heart Sound	http://mldata.org/repository/data/viewslug/record-of-heart-sound/	*Recordings of normal and abnormal heartbeats, used to recognize heart murmur, etc.*
mcdonalds logo scene image dataset	http://www.cancerimagingarchive.net/	*Tumor and nontumor samples, used to recognize prostate cancer.*
Wine Quality	http://archive.ics.uci.edu/ml/datasets/Wine+Quality	*Chemical properties of red and white wines (separately) and quality, for classification.*
Mushroom Identification	http://archive.ics.uci.edu/ml/datasets/Mushroom	*For hypothetically classifying mushrooms as edible or poisonous based on its characteristics.*
UFO Reports	https://github.com/planetsig/ufo-reports	*80,000 historic reports for classification or regression. This dataset has been standardized from the source data at nuforc.org.*
Militarized Interstate Disputes	http://www.correlatesofwar.org/data-sets/MIDs	*Nearly 200 years of international threats, conflicts, etc. for modelling or prediction. Includes action taken, level of hostility, fatalities, and outcomes.*
NBA & MLB Stats	http://www.dougstats.com/	*Current and past season stats for teams and players for fantasy sports predictions.*
Sign Language	http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php	
MusicNet	http://homes.cs.washington.edu/~thickstn/musicnet.html	*MusicNet is a collection of 330 freely-licensed classical music recordings, together with over 1 million annotated labels indicating the precise time of each note every recording, the instrument that plays each note, and the note's position in the metrical structure of the composition. The labels are acquired from musical scores aligned to recordings by dynamic time warping. The labels are verified by trained musicians; we estimate a labeling error rate of 4%. We offer the MusicNet labels to the machine learning and music communities as a resource for training models and a common benchmark for comparing results.*
ProductHunt	https://data.world/producthunt/product-hunt-research	
Reddit	https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/	*1.7 billion Reddit comments*
VQA2	https://arxiv.org/pdf/1612.00837.pdf	*visual question answering dataset, now 2X larger*
UCI ML Repo	https://archive.ics.uci.edu/ml/datasets.html	*351 datasets*
Hacker News	http://aaron-hoffman.blogspot.com/2016/10/hacker-news-dataset-october-2016.html	*the comment dump for HN*
FIRE	http://www.ics.forth.gr/cvrl/fire/	Fundus Image Registration Dataset
LASIESTA	http://www.gti.ssr.upm.es/data/LASIESTA	*Labeled and Annotated Sequences for Integral Evaluation of SegmenTation Algorithms*
LAKH MIDI Dataset	http://colinraffel.com/projects/lmd/	*Its goal is to facilitate large-scale music information retrieval, both symbolic (using the MIDI files alone) and audio content-based (using information extracted from the MIDI files as annotations for the matched audio files).*
Lamem	http://memorability.csail.mit.edu/	*Large-scale Image Memorability*
Pratheepan dataset	http://cs-chan.com/project1.htm	*Human Skin Detection dataset*
COCO-Stuff dataset	http://calvin.inf.ed.ac.uk/datasets/coco-stuff	*COCO-Stuff semantic segmentation dataset*
NewsQA	http://datasets.maluuba.com/NewsQA	*Maluuba's News QA is a new **machine reading comprehension dataset** for developing algorithms capable of answering questions requiring human-level comprehension and reasoning skills.* ***This dataset of CNN news articles has over 110K Q&A pairs.*** *Questions are written by humans in natural language. Questions may not have answers and answers may be multiword passages.*
Awesome Public Datasets	eminem still we dance you dance	*A massive Github repo of accessible, public datasets. The datasets are not, by nature, completely clean and purpose-built for ML.*

food	Link	Purpose
ImageNet	http://image-net.org/download.php	*The ImageNet project is a large visual database designed for use in visual object recognition software research*
Element List Scientific Data Directory	http://www.elementlist.com/scientific_data/	*An online repository of links to free, publicly available scientific datasets, mostly from university, industry, and government research programs.*
IMDB dataset	ftp://ftp.fu-berlin.de/pub/misc/movies/database/	
MSCOCO	http://mscoco.org/	*Image segmentation and object recognition*
Google Books Ngrams	**https://aws.amazon.com/datasets/google-books-ngrams/**	
OpenML repository	http://www.openml.org/search?type=data	*Almost 20k datasets*
Enron Email Corpus	https://en.wikipedia.org/wiki/Enron_Corpus	*The Enron Corpus is a large database of over 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse.*
German Traffic Signs	http://benchmark.ini.rub.de/	*German Traffic Sign Detection Benchmark (GTSDB). The first was used in a competition at IJCNN 2011.*
SYNTHIA	http://www.synthia-dataset.net	*500.000 frames of annotated vídeo from a virtualcity. labels for stereo, optical flow, semántica segmentación, odometry...*
Elektra	http://adas.cvc.uab.es/elektra	*over 20 different autonomous driving datasets: pedestrians, semantic segmentation, stereo...*
Cornell Movie--Dialogs Corpus	http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html	This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts:
Virtual KITTI	http://www.xrce.xerox.com/Our-Research/Computer-Vision/Proxy-Virtual-Worlds	Large photo-realistic synthetic video understanding dataset (high res. videos @30FPS generated with the Unity Game Engine). Automatically, exactly, and fully annotated for all 2D and 3D ground truths at the pixel level (object detection & tracking, segmentation, optical flow, depth, structure from motion, ...).
Bureau of Labor Statistics	http://www.bls.gov/data/	*Dozens of longitudinal datasets provided by the US Department of Labor (CPI, PPI, employment, population, pay, etc.)*
KITTI Vision Benchmark Suite	http://www.cvlibs.net/datasets/kitti/	*Computer vision benchmarks: stereo, flow, odometry, object detection or tracking*
Allen Institute for Artificial Intelligence Datasets	http://allenai.org/data.html	*Datasets for computer vision, reasoning and inference, question answering, and natural language understanding*
Numenta Anomaly Benchmark (NAB)	https://github.com/numenta/NAB	*This repository contains the data and scripts comprising the Numenta Anomaly Benchmark (NAB). NAB is a novel benchmark for evaluating algorithms for anomaly detection in streaming, real-time applications.*
Cityscapes Dataset	https://www.cityscapes-dataset.com/	*Targets semantic understanding of urban street scenes. Great for visual perception applications in automotive industry (ADAS, self-driving).*
MS MARCO (machine reading comprehension & question answering dataset)	http://www.msmarco.org	*A dataset with 100K questions from real users, passages from web pages that could answer the question, and human generated natural language answers*
UCF101 dataset	http://crcv.ucf.edu/data/UCF101.php	*UCF101 a trimmed video datasets for human action recognition, 13k videos*
HMDB51 dataset	http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/	*HMDB51 a large human motion database, 5,6k videos*
Stanford Drone Dataset	http://cvgl.stanford.edu/projects/uav_data/	*When humans navigate a crowed space such as a university campus or the sidewalks of a busy street, they follow common sense rules based on social etiquette. In order to enable the design of new algorithms that can fully take advantage of these rules to better solve tasks such as target tracking or trajectory forecasting, we need to have access to better data. To that end, we contribute the very first large scale dataset (to the best of our knowledge) that collects images and videos of various types of agents (not just pedestrians, but also bicyclists, skateboarders, cars, buses, and golf carts) that navigate in a real world outdoor environment such as a university campus. In the above images, pedestrians are labeled in pink, bicyclists in red, skateboarders in orange, and cars in green.*
High-Resolution Settlement Layer	https://ciesin.columbia.edu/data/hrsl/	*The High Resolution Settlement Layer (HRSL) provides estimates of human population distribution at a resolution of 1 arc-second (approximately 30m) for the year 2015*
	http://cse.iitkgp.ac.in/~abhijnan/	
Oxford Robotcar Dataset	http://robotcar-dataset.robots.ox.ac.uk/	*1 year and approximately 1000km of recorded driving with over 20 million images collected from 6 cameras mounted to the vehicle, along with LIDAR, GPS and INS ground truth. Data was collected in all weather conditions.*
Kepler Data Products	http://archive.stsci.edu/kepler/data_products.html	*https://arxiv.org/pdf/1408.1496.pdf*
Broad Bioimage Benchmark Collection (BBBC)	https://data.broadinstitute.org/bbbc/	*Collection of freely downloadable microscopy image sets. In addition to the images themselves, each set includes a description of the biological application and some type of "ground truth" (expected results).*
Trump Data	https://github.com/brandtg/trump-data	*Collection of data from Donald Trump's 2016 presidential campaign*
Caltech Pedestrian Detection Benchmark	https://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/	*The Caltech Pedestrian Dataset consists of approximately 10 hours of 640x480 30Hz video taken from a vehicle driving through regular traffic in an urban environment. About 250,000 frames (in 137 approximately minute long segments) with a total of 350,000 bounding boxes and 2300 unique pedestrians were annotated. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels. More information can be found in our PAMI 2012 and CVPR 2009 benchmarking papers.*
PoseNet	http://mi.eng.cam.ac.uk/projects/relocalisation/#dataset	*PoseNet was trained with the Cambridge Landmarks Dataset. This is a large urban relocalisation dataset with 6 scenes from around Cambridge University containing over 12,000 images labelled with their full 6-DOF camera pose.*
Scrape Cars	https://www.youtube.com/watch?v=xhp47v5OBXQ	*Building a car image dataset from scraping.*
Swedish Military	http://labs.europeana.eu/data/swedish-military-aviation-in-historical-images	*Over 13,000 photographs, postcards, posters, floor plans of Swedish Air Force*
Volcanoes on Venus	http://kdd.ics.uci.edu/databases/volcanoes/volcanoes.html	*Images of small volcanoes in the large set of Venus collected by the Magellan spacecraft from 1990 to 1994.*
Online News Popularity	http://archive.ics.uci.edu/ml/machine-learning-databases/00332/	*Statistics associated with articles published by Mushable*
Wind	http://lib.stat.cmu.edu/datasets/wind.data	*Daily average wind speeds for 1961-1978 at 12 synoptic meteorological stations in the Republic of Ireland*

food	Link	Purpose
Geographical Analysis Spatial Data	http://lib.stat.cmu.edu/datasets/space_ga	*Contains 3,107 observations on U.S. county votes cast in the 1980 presidential election.*
Air Quality	https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r	*Air Quality in New York City*
Endangered Species Act Critical Habitat	http://www.nmfs.noaa.gov/gis/data/critical.htm	*Fisheries Data: Critical Habtat for each species.*
Residential Fire Fatalities in the News	https://apps.usfa.fema.gov/civilian-fatalities/incident/reportList	*Between January 1, 2016 and December 20, 2016 2158 civilian home fire fatalities were reported by U.S. news media*
Tropical Cyclone Information System	ftp://mwsci.jpl.nasa.gov/outgoing/	*It contains satellite depictions of hurricanes over the globe from 1999-2010.*
North American Bat Ranges	https://catalog.data.gov/dataset/north-american-bat-ranges-direct-download	*Our current understanding of the distributions of United States and Canadian bat species during the past 100-150 years*
Frames	https://datasets.maluuba.com/Frames	*Maluuba's Frames is a new human-generated dataset consisting of consists of 19,986 turns that can be used to help train deep-learning algorithms on natural conversations. These text-based conversations were recorded between two humans, simulating the conversation between a vacation seeker and a travel agent*
4D Light Field Dataset (HCI Heidelberg & CVIA Konstanz)	http://hci-lightfield.iwr.uni-heidelberg.de/	*A synthetic light field dataset with 24 scenes.* *Data provided for each scene:* *- 9x9x512x512x3 light fields as individual PNGs* *- config files with camera settings and disparity ranges*
CASIA WebFace	http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html	*494414 "in the wild" facial images from 10575 labelled subjects. Institutional access only.*
VGG Face Dataset	http://www.robots.ox.ac.uk/~vgg/data/vgg_face/	*~2.6 million "in the wild" facial images from ~2600 labelled subjects. Only URLs to publicly available images and face bounding boxes provided.*
Youtube Faces	http://www.cs.tau.ac.il/~wolf/ytfaces/	*Large dataset of facial images cropped from youtube videos, labelled by subject.*
LibriSpeech ASR corpus	http://www.openslr.org/12	*LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.*
TED-LIUM	http://www.openslr.org/7/	*English speech recognition training corpus from TED talks, created by Laboratoire d'Informatique de l'Université du Maine (LIUM)*
EveryPolitician	http://everypolitician.org	*The world's richest open dataset on politicians*
SceneNet RGB-D	robotvault.bitbucket.org/scenenet-rgbd.html	*5M Photo-realistic synthetic images for indoor scenes*
NYU Depth Dataset V2	http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html	*Indoor Segmentation and Support Inference from RGBD Images* *ECCV 2012*
Dataset of Object Scans	http://redwood-data.org/3dscan/index.html	*Over 10,000 objects densely scanned and reconstructed. Data captured from the real world by non-technical operators.*
CelebFaces	http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	
YouTube Bounding Boxes	https://research.googleblog.com/2017/02/advancing-research-on-video.html	*Today, in order to facilitate progress in video understanding research, we are introducing YouTube-BoundingBoxes, a dataset consisting of 5 million bounding boxes spanning 23 object categories, densely labeling segments from 210,000 YouTube videos. To date, this is the largest manually annotated video dataset containing bounding boxes, which track objects in temporally contiguous frames. The dataset is designed to be large enough to train large-scale models, and be representative of videos captured in natural settings. Importantly, the human-labelled annotations contain objects as they appear in the real world with partial occlusions, motion blur and natural lighting.*