
Awesome Data Science with Python

A curated list of awesome resources for practicing data science using Python, including not only libraries, but also links to tutorials, code snippets, blog posts and talks.

Core pandas - Data structures built on top of numpy.
scikit-learn - Core ML library, intellex.
matplotlib - Plotting library.
seaborn - Data visualization library based on matplotlib.
ydata-profiling - Descriptive statistics using [ProfileReport](#).
sklearn_pandas - Helpful [DataFrameMapper](#) class.
missingno - Missing data visualization.
rainbow-csv - VSCode plugin to display .csv files with nice colors.

General Python Programming [Python Best Practices Guide](#)

pyenv - Manage multiple Python versions on your system.
poetry - Dependency management.
pyscaffold - Python project template generator.
hydra - Configuration management.
hatch - Python project management.
more_itertools - Extension of itertools.
tqdm - Progress bars for for-loops. Also supports pandas `apply()`.
loguru - Python logging.

Pandas Tricks, Alternatives and Additions [pandasvault](#) - Large collection of pandas tricks.

polars - Multi-threaded alternative to pandas.
xarray - Extends pandas to n-dimensional arrays.
mlx - An array framework for Apple silicon.
pandas_flavor - Write custom accessors like `.str` and `.dt`.
duckdb - Efficiently run SQL queries on pandas DataFrame.
daft - Distributed DataFrame.

Pandas Parallelization [modin](#) - Parallelization library for faster pandas [DataFrame](#).

vaex - Out-of-Core DataFrames.
pandarallel - Parallelize pandas operations.
swifter - Apply any function to a pandas DataFrame faster.

Environment and Jupyter Jupyter Tricks

ipyflow - IPython kernel for Jupyter with additional features.

nteract - Open Jupyter Notebooks with doubleclick.

papermill - Parameterize and execute Jupyter notebooks, tutorial.

nbdime - Diff two notebook files, Alternative GitHub App: ReviewNB.

RISE - Turn Jupyter notebooks into presentations.

qgrid - Pandas [DataFrame](#) sorting.

lux - DataFrame visualization within Jupyter.

pandasgui - GUI for viewing, plotting and analyzing Pandas DataFrames.

dtale - View and analyze Pandas data structures, integrating with Jupyter.

itables - Interactive tables in Jupyter.

handcalcs - More convenient way of writing mathematical equations in Jupyter.

notebooker - Productionize and schedule Jupyter Notebooks.

bamboolib - Intuitive GUI for tables.

voila - Turn Jupyter notebooks into standalone web applications.

voila-gridstack - Voila grid layout.

Extraction textract - Extract text from any document.

Big Data spark - [DataFrame](#) for big data, cheatsheet, tutorial.

dask, dask-ml - Pandas [DataFrame](#) for big data and machine learning library, resources, talk1, talk2, notebooks, videos.

h2o - Helpful [H2OFrame](#) class for out-of-memory dataframes.

datatable - Data Table for big data support.

cuDF - GPU DataFrame Library, Intro.

cupy - NumPy-like API accelerated with CUDA.

ray - Flexible, high-performance distributed execution framework.

bottleneck - Fast NumPy array functions written in C.

petastorm - Data access library for parquet files by Uber.

zarr - Distributed NumPy arrays.

NVTabular - Feature engineering and preprocessing library for tabular data by Nvidia.

tensorstore - Reading and writing large multi-dimensional arrays (Google).

Command line tools, CSV csvkit - Command line tool for CSV files.

csvsort - Sort large csv files.

Classical Statistics

p-values The ASA Statement on p-Values: Context, Process, and Purpose

Greenland - Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations

Blume - Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses

Rubin - Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses

Gigerenzer - Mindless Statistics

Rubin - That's not a two-sided test! It's two one-sided tests!

Correlation Guess the Correlation - Correlation guessing game.

phik - Correlation between categorical, ordinal and interval variables.

hoeffd - Hoeffding's D Statistics, measure of dependence (R package).

Packages statsmodels - Statistical tests.

linearmodels - Instrumental variable and panel data models.

penguin - Statistical tests. Pairwise correlation between columns of pandas DataFrame

scipy.stats - Statistical tests.

scikit-posthocs - Statistical post-hoc tests for pairwise multiple comparisons.

Bland-Altman Plot 1, 2 - Plot for agreement between two methods of measurement.

ANOVA

Effect Size Estimating Effect Sizes From Pretest-Posttest-Control Group Designs - Scott B. Morris, Twitter

Statistical Tests test_proportions_2indep - Proportion test.

G-Test - Alternative to chi-square test, power_divergence.

Comparing Two Populations torch-two-sample - Friedman-Rafsky Test: Compare two population based on a multivariate generalization of the Runstest. Explanation, Application

Power and Sample Size Calculations pwrss - Statistical Power and Sample Size Calculation Tools (R package), Tutorial with t-test

Interim Analyses / Sequential Analysis / Stopping Sequential Analysis - Wikipedia.
sequential - Exact Sequential Analysis for Poisson and Binomial Data (R package).
confseq - Uniform boundaries, confidence sequences, and always-valid p-values.

Visualizations Friends don't let friends make certain types of data visualization
Great Overview over Visualizations
Dependent Probabilities
Null Hypothesis Significance Testing (NHST) and Sample Size Calculation
Correlation
Cohen's d
Confidence Interval
Equivalence, non-inferiority and superiority testing
Bayesian two-sample t test
Distribution of p-values when comparing two groups
Understanding the t-distribution and its normal approximation
Statistical Power and Sample Size Calculation Tools

Talks Inverse Propensity Weighting
Dealing with Selection Bias By Propensity Based Feature Selection

Texts Modes, Medians and Means: A Unifying Perspective
Using Norms to Understand Linear Regression
Verifying the Assumptions of Linear Models
Mediation and Moderation Intro
Montgomery et al. - How conditioning on post-treatment variables can ruin your experiment and what to do about it
Lindeløv - Common statistical tests are linear models
Chatruc - The Central Limit Theorem and its misuse
Al-Saleh - Properties of the Standard Deviation that are Rarely Mentioned in Classrooms
Wainer - The Most Dangerous Equation
Gigerenzer - The Bias Bias in Behavioral Economics
Cook - Estimating the chances of something that hasn't happened yet
Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, Youtube
How large is that number in the Law of Large Numbers?
The Prosecutor's Fallacy
The Dunning-Kruger Effect is Autocorrelation

Rafi, Greenland - Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise

Carlin et al. - On the uses and abuses of regression models: a call for reform of statistical practice and teaching

Evaluation Collins et al. - Evaluation of clinical prediction models (part 1): from development to external validation - Twitter

Epidemiology R Epidemics Consortium - Large tool suite for working with epidemiological data (R packages). Github

incidence2 - Computation, handling, visualisation and simple modelling of incidence (R package).

EpiEstim - Estimate time varying instantaneous reproduction number R during epidemics (R package) paper.

researchpy - Helpful `summary_cont()` function for summary statistics (Table 1).

zEpid - Epidemiology analysis package, Tutorial.

tipr - Sensitivity analyses for unmeasured confounders (R package).

quartets - Anscombe's Quartet, Causal Quartet, Datasaurus Dozen and others (R package).

Exploration and Cleaning Checklist.

pyjanitor - Clean messy column names.

skimpy - Create summary statistics of dataframes. Helpful `clean_columns()` function.

pandera - Data / Schema validation.

impyute - Imputations.

fancyimpute - Matrix completion and imputation algorithms.

imbalanced-learn - Resampling for imbalanced datasets.

tspreprocess - Time series preprocessing: Denoising, Compression, Resampling.

Kaggler - Utility functions (`OneHotEncoder(min_obs=100)`)

Noisy Labels cleanlab - Machine learning with noisy labels, finding mislabelled data, and uncertainty quantification. Also see awesome list below.

doubtlab - Find bad or noisy labels.

Train / Test Split iterative-stratification - Stratification of multilabel data.

Feature Engineering Vincent Warmerdam: Untitled12.ipynb - Using df.pipe()

Vincent Warmerdam: Winning with Simple, even Linear, Models

sklearn - Pipeline, examples.

pdpipe - Pipelines for DataFrames.

scikit-lego - Custom transformers for pipelines.

categorical-encoding - Categorical encoding of variables, vtreat (R package).

dirty_cat - Encoding dirty categorical variables.

patsy - R-like syntax for statistical models.

mlxtend - LDA.

featuretools - Automated feature engineering, example.

tsfresh - Time series feature engineering.

temporian - Time series feature engineering by Google.

pypeln - Concurrent data pipelines.

feature-engine - Encoders, transformers, etc.

Computer Vision Intro to Computer Vision

Feature Selection Overview Paper, Talk, Repo

Blog post series - 1, 2, 3, 4

Tutorials - 1, 2

sklearn - Feature selection.

eli5 - Feature selection using permutation importance.

scikit-feature - Feature selection algorithms.

stability-selection - Stability selection.

scikit-rebate - Relief-based feature selection algorithms.

scikit-genetic - Genetic feature selection.

boruta_py - Feature selection, explanation, example.

Boruta-Shap - Boruta feature selection algorithm + shapley values.

linselect - Feature selection package.

mlxtend - Exhaustive feature selection.

BoostARoota - Xgboost feature selection algorithm.

INVASE - Instance-wise Variable Selection using Neural Networks.

SubTab - Subsetting Features of Tabular Data for Self-Supervised Representation Learning, AstraZeneca.

mrml - Maximum Relevance and Minimum Redundancy Feature Selection, Website.

arfs - All Relevant Feature Selection.

VSURF - Variable Selection Using Random Forests (R package) doc.

FeatureSelectionGA - Feature Selection using Genetic Algorithm.

Subset Selection apricot - Selecting subsets of data sets to train machine learning models quickly.
ducks - Index data for fast lookup by any combination of fields.

Dimensionality Reduction / Representation Learning

Selection Check also the Clustering section and self-supervised learning section for ideas!

Review

PCA - link

Autoencoder - link

Isomaps - link

LLE - link

Force-directed graph drawing - link

MDS - link

Diffusion Maps - link

t-SNE - link

NeRV - link, paper

MDR - link

UMAP - link

Random Projection - link

Ivis - link

SimCLR - link

Neural-network based esvit - Vision Transformers for Representation Learning (Microsoft).

MCML - Semi-supervised dimensionality reduction of Multi-Class, Multi-Label data (sequencing data) paper.

Packages Dangers of PCA (paper).

Phantom oscillations in PCA.

What to use instead of PCA.

Talk, tsne intro. sklearn.manifold and sklearn.decomposition - PCA, t-SNE, MDS, Isomaps and others.

Additional plots for PCA - Factor Loadings, Cumulative Variance Explained, Correlation Circle Plot,

Tweet

sklearn.random_projection - Johnson-Lindenstrauss lemma, Gaussian random projection, Sparse

random projection.

sklearn.cross_decomposition - Partial least squares, supervised estimators for dimensionality reduction and regression.

prince - Dimensionality reduction, factor analysis (PCA, MCA, CA, FAMD).

Faster t-SNE implementations: lvdmaaten, MulticoreTSNE, Fit-SNE umap - Uniform Manifold Approximation and Projection, talk, explorer, explanation, parallel version.

humap - Hierarchical UMAP.

sleepwalk - Explore embeddings, interactive visualization (R package).

somoclu - Self-organizing map.

scikit-tda - Topological Data Analysis, paper, talk, talk, paper.

giotto-tda - Topological Data Analysis.

ivis - Dimensionality reduction using Siamese Networks.

trimap - Dimensionality reduction using triplets.

scanpy - Force-directed graph drawing, Diffusion Maps.

direpack - Projection pursuit, Sufficient dimension reduction, Robust M-estimators.

DBS - DatabionicSwarm (R package).

contrastive - Contrastive PCA.

scPCA - Sparse contrastive PCA (R package).

tmap - Visualization library for large, high-dimensional data sets.

lollipop - Linear Optimal Low Rank Projection.

linearsdr - Linear Sufficient Dimension Reduction (R package).

PHATE - Tool for visualizing high dimensional data.

Visualization All charts, Austrian monuments.

Better heatmaps and correlation plots.

Example notebooks for interactive visualizations(Plotly,Seaborn, Holoviz, Altair)

cufflinks - Dynamic visualization library, wrapper for plotly, medium, example.

physt - Better histograms, talk, notebook.

fast-histogram - Fast histograms.

matplotlib_venn - Venn diagrams, alternative.

joypy - Draw stacked density plots (=ridge plots), Ridge plots in seaborn.

mosaic plots - Categorical variable visualization, example.

scikit-plot - ROC curves and other visualizations for ML models.

yellowbrick - Visualizations for ML models (similar to scikit-plot).

bokeh - Interactive visualization library, Examples, Examples.

lets-plot - Plotting library.

animatplot - Animate plots build on matplotlib.

plotnine - ggplot for Python.
altair - Declarative statistical visualization library.
bqplot - Plotting library for IPython/Jupyter Notebooks.
hvplot - High-level plotting library built on top of holoviews.
dtreeviz - Decision tree visualization and model interpretation.
chartify - Generate charts.
VivaGraphJS - Graph visualization (JS package).
pm - Navigatable 3D graph visualization (JS package).
python-ternary - Triangle plots.
falcon - Interactive visualizations for big data.
hiplot - High dimensional Interactive Plotting.
visdom - Live Visualizations.
mpl-scatter-density - Scatter density plots. Alternative to 2d-histograms.
ComplexHeatmap - Complex heatmaps for multidimensional genomic data (R package).
largeVis - Visualize embeddings (t-SNE etc.) (R package).
proplot - Matplotlib wrapper.
morpheus - Broad Institute tool matrix visualization and analysis software. Source, Tutorial: 1, 2, Code.
jupyter-scatter - Interactive 2D scatter plot widget for Jupyter.

Colors palettable - Color palettes from colorbrewer2.
colorcet - Collection of perceptually uniform colormaps.
Named Colors Wheel - Color wheel for all named HTML colors.

Dashboards py-shiny - Shiny for Python, talk.
superset - Dashboarding solution by Apache.
streamlit - Dashboarding solution. Resources, Gallery Components, bokeh-events.
mercury - Convert Python notebook to web app, Example.
dash - Dashboarding solution by plot.ly. Resources.
visdom - Dashboarding library by Facebook.
panel - Dashboarding solution.
altair example - Video.
voila - Turn Jupyter notebooks into standalone web applications.
voila-gridstack - Voila grid layout.

UI gradio - Create UIs for your machine learning model.

Survey Tools samplics - Sampling techniques for complex survey designs.

Geographical Tools folium - Plot geographical maps using the Leaflet.js library, jupyter plugin.

gmaps - Google Maps for Jupyter notebooks.

stadiamaps - Plot geographical maps.

datashader - Draw millions of points on a map.

sklearn - BallTree.

pynndescent - Nearest neighbor descent for approximate nearest neighbors.

geocoder - Geocoding of addresses, IP addresses.

Conversion of different geo formats: talk, repo

geopandas - Tools for geographic data

Low Level Geospatial Tools (GEOS, GDAL/OGR, PROJ.4)

Vector Data (Shapely, Fiona, Pyproj)

Raster Data (Rasterio)

Plotting (Descartes, Catropy)

Predict economic indicators from Open Street Map.

PySal - Python Spatial Analysis Library.

geography - Extract countries, regions and cities from a URL or text.

cartogram - Distorted maps based on population.

Recommender Systems Examples: 1, 2, 2-ipynb, 3.

surprise - Recommender, talk.

implicit - Fast Collaborative Filtering for Implicit Feedback Datasets.

spotlight - Deep recommender models using PyTorch.

lightfm - Recommendation algorithms for both implicit and explicit feedback.

funk-svd - Fast SVD.

Decision Tree Models Intro to Decision Trees and Random Forests, Intro to Gradient Boosting 1, 2, Decision Tree Visualization

lightgbm - Gradient boosting (GBDT, GBRT, GBM or MART) framework based on decision tree algorithms, doc.

xgboost - Gradient boosting (GBDT, GBRT or GBM) library, doc, Methods for CIs: link1, link2.

catboost - Gradient boosting.

h2o - Gradient boosting and general machine learning framework.

pycaret - Wrapper for xgboost, lightgbm, catboost etc.

forestci - Confidence intervals for random forests.

grf - Generalized random forest.

dtreeviz - Decision tree visualization and model interpretation.
Nuance - Decision tree visualization.
rfpimp - Feature Importance for RandomForests using Permutation Importance.
Why the default feature importance for random forests is wrong: link
bartpy - Bayesian Additive Regression Trees.
merf - Mixed Effects Random Forest for Clustering, video
groot - Robust decision trees.
linear-tree - Trees with linear models at the leaves.

Natural Language Processing (NLP) / Text Processing talk-nb, nb2, talk.

Text classification Intro, Preprocessing blog post.
gensim - NLP, doc2vec, word2vec, text processing, topic modelling (LSA, LDA), Example, Coherence Model for evaluation.
Embeddings - GloVe ([1], [2]), StarSpace, wikipedia2vec, visualization.
magnitude - Vector embedding utility package.
pyldavis - Visualization for topic modelling.
spaCy - NLP.
NLTK - NLP, helpful `KMeansClusterer` with `cosine_distance`.
pytext - NLP from Facebook.
fastText - Efficient text classification and representation learning.
annoy - Approximate nearest neighbor search.
faiss - Approximate nearest neighbor search.
pysparnn - Approximate nearest neighbor search.
infomap - Cluster (word-)vectors to find topics.
datasketch - Probabilistic data structures for large data (MinHash, HyperLogLog).
flair - NLP Framework by Zalando.
stanza - NLP Library.
Chatistics - Turn Messenger, Hangouts, WhatsApp and Telegram chat logs into DataFrames.
textdistance - Collection for comparing distances between two or more sequences.

Bio Image Analysis Awesome Cytodata

Tutorials MIT 7.016 Introductory Biology, Fall 2018 - Videos 27, 28, and 29 talk about staining and imaging.
bioimaging.org - A biologists guide to planning and performing quantitative bioimaging experiments.
Introduction to Bioimage Analysis - Book.

Bio-image Analysis Notebooks - Large collection of image processing workflows, including point-spread-function estimation and deconvolution, 3D cell segmentation, feature extraction using pyclesperanto and others.

python_for_microscopists - Notebooks and associated youtube channel for a variety of image processing tasks.

Datasets jump-cellpainting - Cellpainting dataset.

MedMNIST - Datasets for 2D and 3D Biomedical Image Classification.

CytoImageNet - Huge diverse dataset like ImageNet but for cell images.

Haghighi - Gene Expression and Morphology Profiles.

broadinstitute/lincs-profiling-complementarity - Cellpainting vs. L1000 assay.

Biostatistics / Robust statistics MinCovDet - Robust estimator of covariance, RMPV, Paper, App1, App2.

moderated z-score - Weighted average of z-scores based on Spearman correlation.

winsorize - Simple adjustment of outliers.

High-Content Screening Assay Design Zhang XHD (2008) - Novel analytic criteria and effective plate designs for quality control in genome-wide RNAi screens

Iversen - A Comparison of Assay Performance Measures in Screening Assays, Signal Window, Z' Factor, and Assay Variability Ratio Z-factor - Measure of statistical effect size.

Z'-factor - Measure of statistical effect size.

CV - Coefficient of variation.

SSMD - Strictly standardized mean difference.

Signal Window - Assay quality measurement.

Microscopy + Assay BD Spectrum Viewer - Calculate spectral overlap, bleed through for fluorescence microscopy dyes.

SpectraViewer - Visualize the spectral compatibility of fluorophores (PerkinElmer).

Thermofisher Spectrum Viewer - Thermofisher Spectrum Viewer.

Microscopy Resolution Calculator - Calculate resolution of images (Nikon).

PlateEditor - Drug Layout for plates, app, zip, paper.

Image Formats and Converters OME-Zarr - paper, standard

bioformats2raw - Various formats to zarr.

raw2ometiff - Zarr to tiff.

BatchConvert - Wrapper for bioformats2raw to parallelize conversions with nextflow, video.
REMBI model - Recommended Metadata for Biological Images, BioImage Archive: Study Component Guidance, File List Guide, paper, video, spreadsheet

Matrix Formats anndata - annotated data matrices in memory and on disk, Docs.
muon - Multimodal omics framework.
mudata - Multimodal Data (.h5mu) implementation.
bdz - Zarr-based format for storing quantitative biological dynamics data.

Image Viewers napari - Image viewer and image processing tool.
Fiji - General purpose tool. Image viewer and image processing tool.
vizarr - Browser-based image viewer for zarr format.
avivator - Browser-based image viewer for tiff files.
OMERO - Image viewer for high-content screening. IDR uses OMERO. Intro
fiftyone - Viewer and tool for building high-quality datasets and computer vision models.
Image Data Explorer - Microscopy Image Viewer, Shiny App, Video.
ImSwitch - Microscopy Image Viewer, Doc, Video.
pixmi - Web-based image annotation and classification tool, App.
DeepCell Label - Data labeling tool to segment images, Video.

Napari Plugins napari-sam - Segment Anything Plugin.
napari-chatgpt - ChatGPT Plugin.

Image Restoration and Denoising aydin - Image denoising.
DivNoising - Unsupervised denoising method.
CSBDeep - Content-aware image restoration, Project page.

Illumination correction skimage - Illumination correction (CLAHE).
cidre - Illumination correction method for optical microscopy.
BaSiCPy - Background and Shading Correction of Optical Microscopy Images, BaSiC.

Bleedthrough correction / Spectral Unmixing PICASSO - Blind unmixing without reference spectra measurement, Paper
cytoflow - Flow cytometry. Includes Bleedthrough correction methods.
Linear unmixing in Fiji for Bleedthrough Correction - Youtube.

Bleedthrough Correction using Lumos and Fiji - Link.

AutoUnmix - Link.

Platforms and Pipelines CellProfiler, CellProfilerAnalyst - Create image analysis pipelines.

fractal - Framework to process high-content imaging data from UZH, Github.

atomai - Deep and Machine Learning for Microscopy.

py-clesperanto - Tools for 3D microscopy analysis, deskewing and lots of other tutorials, interacts with napari.

qupath - Image analysis.

Microscopy Pipelines Labsyspharm Stack see below.

BiaPy - Bioimage analysis pipelines.

SCIP - Image processing pipeline on top of Dask.

DeepCell Kiosk - Image analysis platform.

IMCWorkflow - Image analysis pipeline using steinbock, Twitter, Paper, workflow.

Labsyspharm mcmicro - Multiple-choice microscopy pipeline, Website, Paper.

MCQuant - Quantification of cell features.

cylinter - Quality assurance for microscopy images, Website.

ashlar - Whole-slide microscopy image stitching and registration.

scimap - Spatial Single-Cell Analysis Toolkit.

Cell Segmentation microscopy-tree - Review of cell segmentation algorithms, Paper.

Review of organoid pipelines - Paper.

BioImage.IO - BioImage Model Zoo.

MEDIAR - Cell segmentation.

cellpose - Cell segmentation. Paper, Dataset.

stardist - Cell segmentation with Star-convex Shapes.

UnMicst - Identifying Cells and Segmenting Tissue.

ilastik - Segment, classify, track and count cells. ImageJ Plugin.

nnUnet - 3D biomedical image segmentation.

allencell - Tools for 3D segmentation, classical and deep learning methods.

Cell-ACDC - Python GUI for cell segmentation and tracking.

ZeroCostDL4Mic - Deep-Learning in Microscopy.

DL4MicEverywhere - Bringing the ZeroCostDL4Mic experience using Docker.

EmbedSeg - Embedding-based Instance Segmentation.

segment-anything - Segment Anything (SAM) from Facebook.
micro-sam - Segment Anything for Microscopy.
Segment-Everything-Everywhere-All-At-Once - Segment Everything Everywhere All at Once from Microsoft.
deepcell-tf - Cell segmentation, DeepCell.
labkit - Fiji plugin for image segmentation.

Cell Segmentation Datasets cellpose - Cell images.

omnipose - Cell images.
LIVECell - Cell images.
Sartorius - Neurons.
EmbedSeg - 2D + 3D images.
connectomics - Annotation of the EPFL Hippocampus dataset.
ZeroCostDL4Mic - Stardist example training and test dataset.

Evaluation seg-eval - Cell segmentation performance evaluation without Ground Truth labels, Paper.

Feature Engineering Images Computer vision challenges in drug discovery - Maciej Hermanowicz
CellProfiler - Biological image analysis.

scikit-image - Image processing.
scikit-image regionprops - Regionprops: area, eccentricity, extent.
mahotas - Zernike, Haralick, LBP, and TAS features, example.
pyradiomics - Radiomics features from medical imaging.
pyefd - Elliptical feature descriptor, approximating a contour with a Fourier series.
pyvips - Faster image processing operations.

Domain Adaptation / Batch-Effect Correction Tran - A benchmark of batch-effect correction methods for single-cell RNA sequencing data, Code.

R Tutorial on correcting batch effects.
harmonypy - Fuzzy k-means and locally linear adjustments.
pyliger - Batch-effect correction, R package.
nimfa - Nonnegative matrix factorization.
scgen - Batch removal. Doc.
CORAL - Correcting for Batch Effects Using Wasserstein Distance, Code, Paper.

adapt - Awesome Domain Adaptation Python Toolbox.
pytorch-adapt - Various neural network models for domain adaptation.

Sequencing Single cell tutorial.

PyDESeq2 - Analyzing RNA-seq data.

cellxgene - Interactive explorer for single-cell transcriptomics data.

scanpy - Analyze single-cell gene expression data, tutorial.

besca - Beyond single-cell analysis.

janggu - Deep Learning for Genomics.

gdsctools - Drug responses in the context of the Genomics of Drug Sensitivity in Cancer project, ANOVA, IC50, MoBEM, doc.

monkeybread - Analysis of single-cell spatial transcriptomics data.

Drug discovery TDC - Drug Discovery and Development.

DeepPurpose - Deep Learning Based Molecular Modelling and Prediction Toolkit.

Neural Networks Convolutional Neural Networks for Visual Recognition - Stanford CS class.

mit6874 - Computational Systems Biology: Deep Learning in the Life Sciences.

ConvNet Shape Calculator - Calculate output dimensions of Conv2D layer.

Great Gradient Descent Article.

Intro to semi-supervised learning.

Tutorials & Viewer fast.ai course - Practical Deep Learning for Coders.

Tensorflow without a PhD - Neural Network course by Google.

Feature Visualization: Blog, PPT

Tensorflow Playground

Visualization of optimization algorithms, Another visualization

cutouts-explorer - Image Viewer.

Image Related imgaug - More sophisticated image preprocessing.

Augmentor - Image augmentation library.

keras preprocessing - Preprocess images.

albumentations - Wrapper around imgaug and other libraries.

augmix - Image augmentation from Google.

kornia - Image augmentation, feature extraction and loss functions.

augly - Image, audio, text, video augmentation from Facebook.

pyvips - Faster image processing operations.

Lossfunction Related SegLoss - List of loss functions for medical image segmentation.

Activation Functions rational_activations - Rational activation functions.

Text Related ktext - Utilities for pre-processing text for deep learning in Keras.

textgenrnn - Ready-to-use LSTM for text generation.

ctrl - Text generation.

Neural network and deep learning frameworks OpenMMLab - Framework for segmentation, classification and lots of other computer vision tasks.

caffe - Deep learning framework, pretrained models.

mxnet - Deep learning framework, book.

Libs General keras - Neural Networks on top of tensorflow, examples.

keras-contrib - Keras community contributions.

keras-tuner - Hyperparameter tuning for Keras.

hyperas - Keras + Hyperopt: Convenient hyperparameter optimization wrapper.

elephas - Distributed Deep learning with Keras & Spark.

tflearn - Neural Networks on top of TensorFlow.

tensorlayer - Neural Networks on top of TensorFlow, tricks.

tensorforce - TensorFlow for applied reinforcement learning.

autokeras - AutoML for deep learning.

PlotNeuralNet - Plot neural networks.

lucid - Neural network interpretability, Activation Maps.

tcav - Interpretability method.

AdaBound - Optimizer that trains as fast as Adam and as good as SGD, alt.

foolbox - Adversarial examples that fool neural networks.

hiddenlayer - Training metrics.

imgclsmob - Pretrained models.

netron - Visualizer for deep learning and machine learning models.

ffcv - Fast dataloader.

Libs PyTorch Good PyTorch Introduction

skorch - Scikit-learn compatible neural network library that wraps PyTorch, talk, slides.

fastai - Neural Networks in PyTorch.

timm - PyTorch image models.

ignite - Highlevel library for PyTorch.

torchcv - Deep Learning in Computer Vision.

pytorch-optimizer - Collection of optimizers for PyTorch.

pytorch-lightning - Wrapper around PyTorch.

lightly - MoCo, SimCLR, SimSiam, Barlow Twins, BYOL, NNCLR.

MONAI - Deep learning in healthcare imaging.

kornia - Image transformations, epipolar geometry, depth estimation.

torchinfo - Nice model summary.

lovely-tensors - Inspect tensors, mean, std, inf values.

Distributed Libs flexflow - Distributed TensorFlow Keras and PyTorch.

horovod - Distributed training framework for TensorFlow, Keras, PyTorch, and Apache MXNet.

Architecture Visualization Awesome List.

netron - Viewer for neural networks.

visualker - Visualize Keras networks.

Object detection / Instance Segmentation Metrics reloaded: Recommendations for image analysis validation - Guide for choosing correct image analysis metrics, Code, Twitter Thread

Good Yolo Explanation

yolact - Fully convolutional model for real-time instance segmentation.

EfficientDet Pytorch, EfficientDet Keras - Scalable and Efficient Object Detection.

detectron2 - Object Detection (Mask R-CNN) by Facebook.

simplenet - Object Detection and Instance Recognition.

CenterNet - Object detection.

FCOS - Fully Convolutional One-Stage Object Detection.

norfair - Real-time 2D object tracking.

Detic - Detector with image classes that can use image-level labels (facebookresearch).

EasyCV - Image segmentation, classification, metric-learning, object detection, pose estimation.

Image Classification nfnets - Neural network.

efficientnet - Neural network.

pycls - PyTorch image classification networks: ResNet, ResNeXt, EfficientNet, and RegNet (by Facebook).

Applications and Snippets SPADE - Semantic Image Synthesis.

Entity Embeddings of Categorical Variables, code, kaggle

Image Super-Resolution - Super-scaling using a Residual Dense Network.

Cell Segmentation - Talk, Blog Posts: 1, 2

deeplearning-models - Deep learning models.

Variational Autoencoders (VAEs) Variational Autoencoder Explanation Video

disentanglement_lib - BetaVAE, FactorVAE, BetaTCVAE, DIP-VAE.

ladder-vae-pytorch - Ladder Variational Autoencoders (LVAE).

benchmark_VAE - Unifying Generative Autoencoder implementations.

Generative Adversarial Networks (GANs) Awesome GAN Applications

The GAN Zoo - List of Generative Adversarial Networks.

CycleGAN and Pix2pix - Various image-to-image tasks.

TensorFlow GAN implementations

PyTorch GAN implementations

PyTorch GAN implementations

StudioGAN - PyTorch GAN implementations.

Transformers SegFormer - Simple and Efficient Design for Semantic Segmentation with Transformers.

esvit - Efficient self-supervised Vision Transformers.

nystromformer - More efficient transformer because of approximate self-attention.

Deep learning on structured data Great overview for deep learning for tabular data

Graph-Based Neural Networks How to do Deep Learning on Graphs with Graph Convolutional Networks

Introduction To Graph Convolutional Networks

An attempt at demystifying graph deep learning

ogb - Open Graph Benchmark, Benchmark datasets.

networkx - Graph library.

cugraph - RAPIDS, Graph library on the GPU.
pytorch-geometric - Various methods for deep learning on graphs.
dgl - Deep Graph Library.
graph_nets - Build graph networks in TensorFlow, by DeepMind.

Model conversion hummingbird - Compile trained ML models into tensor computations (by Microsoft).

GPU cuML - RAPIDS, Run traditional tabular ML tasks on GPUs, Intro.
thundergbm - GBDTs and Random Forest.
thundersvm - Support Vector Machines.
Legate Numpy - Distributed Numpy array multiple using GPUs by Nvidia (not released yet) video.

Regression Understanding SVM Regression: slides, forum, paper
pyearth - Multivariate Adaptive Regression Splines (MARS), tutorial.
pygam - Generalized Additive Models (GAMs), Explanation.
GLRM - Generalized Low Rank Models.
tweedie - Specialized distribution for zero inflated targets, Talk.
MAPIE - Estimating prediction intervals.
Regressio - Regression and Spline models.

Polynomials orthopy - Orthogonal polynomials in all shapes and sizes.

Classification Talk, Notebook
Blog post: Probability Scoring
All classification metrics
DESLib - Dynamic classifier and ensemble selection.
human-learn - Create and tune classifier based on your rule set.

Metric Learning Contrastive Representation Learning
metric-learn - Supervised and weakly-supervised metric learning algorithms.
pytorch-metric-learning - PyTorch metric learning.
deep_metric_learning - Methods for deep metric learning.
ivis - Metric learning using siamese neural networks.
TensorFlow similarity - Metric learning.

Distance Functions `scipy.spatial` - All kinds of distance metrics.

`pyemd` - Earth Mover's Distance / Wasserstein distance, similarity between histograms. OpenCV implementation, POT implementation

`dcor` - Distance correlation and related Energy statistics.

`GeomLoss` - Kernel norms, Hausdorff divergences, Debiased Sinkhorn divergences (=approximation of Wasserstein distance).

Self-supervised Learning `lightly` - MoCo, SimCLR, SimSiam, Barlow Twins, BYOL, NNCLR.

`vissl` - Self-Supervised Learning with PyTorch: RotNet, Jigsaw, NPID, ClusterFit, PIRL, SimCLR, MoCo, DeepCluster, SwAV.

Clustering Overview of clustering algorithms applied image data (= Deep Clustering).

Clustering with Deep Learning: Taxonomy and New Methods.

Hierarchical Cluster Analysis (R Tutorial) - Dendrogram, Tanglegram

`hdbscan` - Clustering algorithm, talk, blog.

`pyclustering` - All sorts of clustering algorithms.

`FCPS` - Fundamental Clustering Problems Suite (R package).

`GaussianMixture` - Generalized k-means clustering using a mixture of Gaussian distributions, video.

`nmslib` - Similarity search library and toolkit for evaluation of k-NN methods.

`merf` - Mixed Effects Random Forest for Clustering, video

`tree-SNE` - Hierarchical clustering algorithm based on t-SNE.

`MiniSom` - Pure Python implementation of the Self Organizing Maps.

`distribution_clustering`, paper, related paper, alt.

`phenograph` - Clustering by community detection.

`FastPG` - Clustering of single cell data (RNA). Improvement of phenograph, Paper.

`HypHC` - Hyperbolic Hierarchical Clustering.

`BanditPAM` - Improved k-Medoids Clustering.

`dendextend` - Comparing dendrograms (R package).

`DeepDPM` - Deep Clustering With An Unknown Number of Clusters.

Clustering Evalutation Wagner, Wagner - Comparing Clusterings - An Overview * Adjusted Rand Index * Normalized Mutual Information * Adjusted Mutual Information * Fowlkes-Mallows Score * Silhouette Coefficient * Variation of Information, Julia * Pair Confusion Matrix * Consensus Score - The similarity of two sets of biclusters.

Assessing the quality of a clustering (video)

`fpc` - Various methods for clustering and cluster validation (R package).

* Minimum distance between any two clusters * Distance between centroids * p-separation index: Like minimum distance. Look at the average distance to nearest point in different cluster for $p=10\%$ “border” points in any cluster. Measuring density, measuring mountains vs valleys * Estimate density by weighted count of close points

Other measures: * Within-cluster average distance * Mean of within-cluster average distance over nearest-cluster average distance (silhouette score) * Within-cluster similarity measure to normal/uniform * Within-cluster (squared) distance to centroid (this is the k-Means loss function) * Correlation coefficient between distance we originally had to the distance the are induced by the clustering (Huberts Gamma) * Entropy of cluster sizes * Average largest within-cluster gap * Variation of clusterings on bootstrapped data

Multi-label classification `scikit-multilearn` - Multi-label classification, talk.

Signal Processing and Filtering Stanford Lecture Series on Fourier Transformation, Youtube, Lecture Notes.

Visual Fourier explanation.

The Scientist & Engineer’s Guide to Digital Signal Processing (1999) - Chapter 3 has good introduction to Bessel, Butterworth and Chebyshev filters.

Kalman Filter article.

Kalman Filter book - Focuses on intuition using Jupyter Notebooks. Includes Bayesian and various Kalman filters.

Interactive Tool for FIR and IIR filters, Examples.

`filterpy` - Kalman filtering and optimal estimation library.

Filtering in Python `scipy.signal` * Butterworth low-pass filter example * Savitzky–Golay filter, `W pandas.Series.rolling` - Choose appropriate `win_type`.

Geometry `geomstats` - Computations and statistics on manifolds with geometric structures.

Time Series `statsmodels` - Time series analysis, seasonal decompose example, SARIMA, granger causality.

`kats` - Time series prediction library by Facebook.

`prophet` - Time series prediction library by Facebook.

`neural_prophet` - Time series prediction built on PyTorch.

`pyramid`, `pmdarima` - Wrapper for (Auto-) ARIMA.

modeltime - Time series forecasting framework (R package).
pyflux - Time series prediction algorithms (ARIMA, GARCH, GAS, Bayesian).
atspy - Automated Time Series Models.
pm-prophet - Time series prediction and decomposition library.
htsprophet - Hierarchical Time Series Forecasting using Prophet.
nupic - Hierarchical Temporal Memory (HTM) for Time Series Prediction and Anomaly Detection.
tensorflow - LSTM and others, examples: [link](#), [link](#), [seq2seq: 1](#), [2](#), [3](#), [4](#)
tspreprocess - Preprocessing: Denoising, Compression, Resampling.
tsfresh - Time series feature engineering.
tsfel - Time series feature extraction.
thunder - Data structures and algorithms for loading, processing, and analyzing time series data.
gatspy - General tools for Astronomical Time Series, [talk](#).
gendis - shapelets, [example](#).
tslearn - Time series clustering and classification, [TimeSeriesKMeans](#), [TimeSeriesKMeans](#).
pastas - Analysis of Groundwater Time Series.
fastdtw - Dynamic Time Warp Distance.
fable - Time Series Forecasting (R package).
pydlm - Bayesian time series modelling (R package, [Blog post](#))
PyAF - Automatic Time Series Forecasting.
luminol - Anomaly Detection and Correlation library from LinkedIn.
matrixprofile-ts - Detecting patterns and anomalies, [website](#), [ppt](#), [alternative](#).
stumpy - Another matrix profile library.
obspy - Seismology package. Useful [classic_sta_lta](#) function.
RobustSTL - Robust Seasonal-Trend Decomposition.
seglearn - Time Series library.
pyts - Time series transformation and classification, [Imaging time series](#).
Turn time series into images and use Neural Nets: [example](#), [example](#).
sktime, sktime-dl - Toolbox for (deep) learning with time series.
adtk - Time Series Anomaly Detection.
rocket - Time Series classification using random convolutional kernels.
luminaire - Anomaly Detection for time series.
etna - Time Series library.
Chaos Genius - ML powered analytics engine for outlier/anomaly detection and root cause analysis.

Time Series Evaluation TimeSeriesSplit - Sklearn time series split.
tscv - Evaluation with gap.

Financial Data and Trading Tutorial on using cvxpy: 1, 2

pandas-datareader - Read stock data.

yfinance - Read stock data from Yahoo Finance.

findatapy - Read stock data from various sources.

ta - Technical analysis library.

backtrader - Backtesting for trading strategies.

surpriver - Find high moving stocks before they move using anomaly detection and machine learning.

ffn - Financial functions.

bt - Backtesting algorithms.

alpaca-trade-api-python - Commission-free trading through API.

eiten - Eigen portfolios, minimum variance portfolios and other algorithmic investing strategies.

tf-quant-finance - Quantitative finance tools in TensorFlow, by Google.

quantstats - Portfolio management.

Riskfolio-Lib - Portfolio optimization and strategic asset allocation.

OpenBBTerminal - Terminal.

mplfinance - Financial markets data visualization.

Quantopian Stack pyfolio - Portfolio and risk analytics.

zipline - Algorithmic trading.

alphalens - Performance analysis of predictive stock factors.

empyrical - Financial risk metrics.

trading_calendars - Calendars for various securities exchanges.

Survival Analysis Time-dependent Cox Model in R.

lifelines - Survival analysis, Cox PH Regression, talk, talk2.

scikit-survival - Survival analysis.

xgboost - "objective": "survival:cox" NHANES example

survivalstan - Survival analysis, intro.

convoys - Analyze time lagged conversions.

RandomSurvivalForests (R packages: randomForestSRC, ggRandomForests).

pysurvival - Survival analysis.

DeepSurvivalMachines - Fully Parametric Survival Regression.

auton-survival - Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Events.

Outlier Detection & Anomaly Detection sklearn - Isolation Forest and others.

pyod - Outlier Detection / Anomaly Detection.

eif - Extended Isolation Forest.

AnomalyDetection - Anomaly detection (R package).

luminol - Anomaly Detection and Correlation library from LinkedIn.

Distances for comparing histograms and detecting outliers - Talk: Kolmogorov-Smirnov, Wasserstein, Energy Distance (Cramer), Kullback-Leibler divergence.

banpei - Anomaly detection library based on singular spectrum transformation.

telemanom - Detect anomalies in multivariate time series data using LSTMs.

luminaire - Anomaly Detection for time series.

rrcf - Robust Random Cut Forest algorithm for anomaly detection on streams.

Concept Drift & Domain Shift TorchDrift - Drift Detection for PyTorch Models.

alibi-detect - Algorithms for outlier, adversarial and drift detection.

evidently - Evaluate and monitor ML models from validation to production.

Lipton et al. - Detecting and Correcting for Label Shift with Black Box Predictors.

Bu et al. - A pdf-Free Change Detection Test Based on Density Difference Estimation.

Ranking lightning - Large-scale linear classification, regression and ranking.

Causal Inference CS 594 Causal Inference and Learning

Statistical Rethinking - Video Lecture Series, Bayesian Statistics, Causal Models, R, python, numpyro1, numpyro2, tensorflow-probability.

Python Causality Handbook

dowhy - Estimate causal effects.

CausalImpact - Causal Impact Analysis (R package).

causalib - Modular causal inference analysis and model evaluations by IBM, examples.

causalml - Causal inference by Uber.

upliftml - Causal inference by Booking.com.

causality - Causal analysis using observational datasets.

DoubleML - Machine Learning + Causal inference, Tweet, Presentation, Paper.

EconML - Heterogeneous Treatment Effects Estimation by Microsoft.

Papers Bours - Confounding

Bours - Effect Modification and Interaction

Probabilistic Modelling and Bayes Intro, Guide

PyMC3 - Bayesian modelling.

numpyro - Probabilistic programming with numpy, built on pyro.
pomegranate - Probabilistic modelling, talk.
pmlearn - Probabilistic machine learning.
arviz - Exploratory analysis of Bayesian models.
zhusuan - Bayesian deep learning, generative models.
edward - Probabilistic modelling, inference, and criticism, Mixture Density Networks (MNDs), MDN Explanation.
Pyro - Deep Universal Probabilistic Programming.
TensorFlow probability - Deep learning and probabilistic modelling, talk1, notebook talk1, talk2, example.
bambi - High-level Bayesian model-building interface on top of PyMC3.
neural-tangents - Infinite Neural Networks.
bnlearn - Bayesian networks, parameter learning, inference and sampling methods.

Gaussian Processes Visualization, Article

GPyOpt - Gaussian process optimization.
GPflow - Gaussian processes (TensorFlow).
gpytorch - Gaussian processes (PyTorch).

Stacking Models and Ensembles Model Stacking Blog Post

mlxtend - [EnsembleVoteClassifier](#), [StackingRegressor](#), [StackingCVRegressor](#) for model stacking.
vecstack - Stacking ML models.
StackNet - Stacking ML models.
mlens - Ensemble learning.
combo - Combining ML models (stacking, ensembling).

Model Evaluation evaluate - Evaluate machine learning models (huggingface).

pycm - Multi-class confusion matrix.
pandas_ml - Confusion matrix.
Plotting learning curve: [link](#).
yellowbrick - Learning curve.
pyroc - Receiver Operating Characteristic (ROC) curves.

Model Uncertainty awesome-conformal-prediction - Uncertainty quantification.

uncertainty-toolbox - Predictive uncertainty quantification, calibration, metrics, and visualization.

Model Explanation, Interpretability, Feature Importance Princeton - Reproducibility Crisis in ML-based Science

Book, Examples

scikit-learn - Permutation Importance (can be used on any trained classifier) and Partial Dependence

shap - Explain predictions of machine learning models, talk, Good Shap intro.

treeinterpreter - Interpreting scikit-learn's decision tree and random forest predictions.

lime - Explaining the predictions of any machine learning classifier, talk, Warning (Myth 7).

lime_xgboost - Create LIMEs for XGBoost.

eli5 - Inspecting machine learning classifiers and explaining their predictions.

lofo-importance - Leave One Feature Out Importance, talk.

pybreakdown - Generate feature contribution plots.

pycebox - Individual Conditional Expectation Plot Toolbox.

pdpbox - Partial dependence plot toolbox, example.

partial_dependence - Visualize and cluster partial dependence.

contrastive_explanation - Contrastive explanations.

DrWhy - Collection of tools for explainable AI.

lucid - Neural network interpretability.

xai - An eXplainability toolbox for machine learning.

investigate - A toolbox to investigate neural network predictions.

dalex - Explanations for ML models (R package).

interpretml - Fit interpretable models, explain models.

shapash - Model interpretability.

imodels - Interpretable ML package.

captum - Model interpretability and understanding for PyTorch.

Automated Machine Learning AdaNet - Automated machine learning based on TensorFlow.

tpot - Automated machine learning tool, optimizes machine learning pipelines.

autokeras - AutoML for deep learning.

nni - Toolkit for neural architecture search and hyper-parameter tuning by Microsoft.

mljar - Automated machine learning.

automl_zero - Automatically discover computer programs that can solve machine learning tasks from Google.

AlphaPy - Automated Machine Learning using scikit-learn xgboost, LightGBM and others.

Graph Representation Learning Karate Club - Unsupervised learning on graphs.

PyTorch Geometric - Graph representation learning with PyTorch.

DLG - Graph representation learning with TensorFlow.

Convex optimization cvxpy - Modelling language for convex optimization problems. Tutorial: 1, 2

Evolutionary Algorithms & Optimization deap - Evolutionary computation framework (Genetic Algorithm, Evolution strategies).

evol - DSL for composable evolutionary algorithms, talk.

platypus - Multiobjective optimization.

autograd - Efficiently computes derivatives of numpy code.

nevergrad - Derivation-free optimization.

gplearn - Sklearn-like interface for genetic programming.

blackbox - Optimization of expensive black-box functions.

Optometrist algorithm - paper.

DeepSwarm - Neural architecture search.

evotorch - Evolutionary computation library built on Pytorch.

Hyperparameter Tuning sklearn - GridSearchCV, RandomizedSearchCV.

sklearn-deap - Hyperparameter search using genetic algorithms.

hyperopt - Hyperparameter optimization.

hyperopt-sklearn - Hyperopt + sklearn.

optuna - Hyperparameter optimization, Talk.

skopt - [BayesSearchCV](#) for Hyperparameter search.

tune - Hyperparameter search with a focus on deep learning and deep reinforcement learning.

bbopt - Black box hyperparameter optimization.

dragonfly - Scalable Bayesian optimisation.

botorch - Bayesian optimization in PyTorch.

ax - Adaptive Experimentation Platform by Facebook.

lightning-hpo - Hyperparameter optimization based on optuna.

Incremental Learning, Online Learning sklearn - PassiveAggressiveClassifier, PassiveAggressiveRegressor.

river - Online machine learning.

Kaggler - Online Learning algorithms.

Active Learning Talk

modAL - Active learning framework.

Reinforcement Learning YouTube, YouTube

Intro to Monte Carlo Tree Search (MCTS) - 1, 2, 3

AlphaZero methodology - 1, 2, 3, Cheat Sheet

RLLib - Library for reinforcement learning.

Horizon - Facebook RL framework.

Deployment and Lifecycle Management

Workflow Scheduling and Orchestration nextflow - Run scripts and workflow graphs in Docker image using Google Life Sciences, AWS Batch, Website.

airflow - Schedule and monitor workflows.

prefect - Python specific workflow scheduling.

dagster - Development, production and observation of data assets.

ploomber - Workflow orchestration.

kestra - Workflow orchestration.

cml - CI/CD for Machine Learning Projects.

rocketry - Task scheduling.

huey - Task queue.

Containerization and Docker Reduce size of docker images (video)

Optimize Docker Image Size

cog - Facilitates building Docker images.

Data Versioning, Databases, Pipelines and Model Serving dvc - Version control for large files.

kedro - Build data pipelines.

feast - Feature store. Video.

pinecone - Database for vector search applications.

truss - Serve ML models.

milvus - Vector database for similarity search.

mlem - Version and deploy your ML models following GitOps principles.

Data Science Related m2cgen - Transpile trained ML models into other languages.

sklearn-porter - Transpile trained scikit-learn estimators to C, Java, JavaScript and others.

mlflow - Manage the machine learning lifecycle, including experimentation, reproducibility and deployment.

skll - Command-line utilities to make it easier to run machine learning experiments.

BentoML - Package and deploy machine learning models for serving in production.

dagster - Tool with focus on dependency graphs.

knockknock - Be notified when your training ends.

metaflow - Lifecycle Management Tool by Netflix.

cortex - Deploy machine learning models.

Neptune - Experiment tracking and model registry.

clearml - Experiment Manager, MLOps and Data-Management.

polyaxon - MLOps.

sematic - Deploy machine learning models.

zenml - MLOPs.

Math and Background All kinds of math and statistics resources

Gilbert Strang - Linear Algebra

Gilbert Strang - Matrix Methods in Data Analysis, Signal Processing, and Machine Learning

Resources Distill.pub - Blog.

Machine Learning Videos

Data Science Notebooks

Recommender Systems (Microsoft)

Datascience Cheatsheets

Guidelines datasharing - Guide to data sharing.

Books Blum - Foundations of Data Science

Chan - Introduction to Probability for Data Science

Colonescu - Principles of Econometrics with R

Rafael Irizarry - Introduction to Data Science (R Language)

Rafael Irizarry - Advanced Data Science (R Language)

Other Awesome Lists Awesome Adversarial Machine Learning

Awesome AI Bookmarks

Awesome AI on Kubernetes

Awesome Big Data

Awesome Biological Image Analysis

Awesome Business Machine Learning

Awesome Causality

Awesome Community Detection
Awesome CSV
Awesome Cytodata
Awesome Data Science with Ruby
Awesome Dash
Awesome Decision Trees
Awesome Deep Learning
Awesome ETL
Awesome Financial Machine Learning
Awesome Fraud Detection
Awesome GAN Applications
Awesome Graph Classification
Awesome Industry Machine Learning
Awesome Gradient Boosting
Awesome Learning with Label Noise
Awesome Machine Learning
Awesome Machine Learning Books
Awesome Machine Learning Interpretability
Awesome Machine Learning Operations
Awesome Monte Carlo Tree Search
Awesome Neural Network Visualization
Awesome Online Machine Learning
Awesome Pipeline
Awesome Public APIs
Awesome Python
Awesome Python Data Science
Awesome Python Data Science
Awesome Pytorch
Awesome Quantitative Finance
Awesome Recommender Systems
Awesome Satellite Benchmark Datasets
Awesome Satellite Image for Deep Learning
Awesome Single Cell
Awesome Semantic Segmentation
Awesome Sentence Embedding
Awesome Time Series
Awesome Time Series Anomaly Detection
Awesome Visual Attentions

Awesome Visual Transformer

Lectures NYU Deep Learning SP21 - YouTube Playlist.

Things I google a lot Color Codes

Frequency codes for time series

Date parsing codes

Contributing

Do you know a package that should be on this list? Did you spot a package that is no longer maintained and should be removed from this list? Then feel free to read the contribution guidelines and submit your pull request or create a new issue.

License

