EthicalML / **awesome-production-machine-learning** Public

A curated list of awesome open source libraries to deploy, monitor, version and scale your machine learning

🔗 ethical.institute/principles.html

⚖ MIT License

⭐ **11.3k** stars   🍴 **1.6k** forks

☆ Star ▾   👁 Watch ▾

‹› Code   ⊙ Issues **16**   ⑂ Pull requests **2**   ▷ Actions   ▦ Projects   📖 Wiki   ⊘ Secu

⑂ master ▾   ⋯

👤 **axsaucedo** Merge pull request #245 from aspfohl/patch-1   ⋯   ✓ 14 hours ago   ⏱ **449**

View code

| awesome | Maintained? YES | Release PROD | Languages MULTI | License MIT | 🐦 Follow 2.5k |

Awesome production machine learning

This repository contains a curated list of awesome open source libraries that will help you deploy, monitor, version, scale, and secure your production machine learning 🚀

Quick links to sections in this page

🔍 Explaining predictions & models	🔒 Privacy preserving ML	📃 Model & data versioning
🏁 Model Training Orchestration	💪 Model Serving and Monitoring	🤖 Neural Architecture Search

|		
📔 Reproducible Notebooks	📊 Visualisation frameworks	🔤 Industry-strength NLP
🧵 Data pipelines & ETL	🏷️ Data Labelling	📅 Metadata Management
📡 Functions as a service	🗺️ Computation distribution	🛠️ Model serialisation
🧮 Optimized computation frameworks	🔖 Data Stream Processing	🔴 Outlier and Anomaly Detection
🌀 Feature engineering	🎁 Feature Stores	⚔️ Adversarial Robustness
💰 Commercial Platforms	💾 Data Storage Optimization	

10 Min Video Overview

This 10 minute video provides an overview of the motivations for machine learning operations as well as a high level overview on some of the tools in this repo.
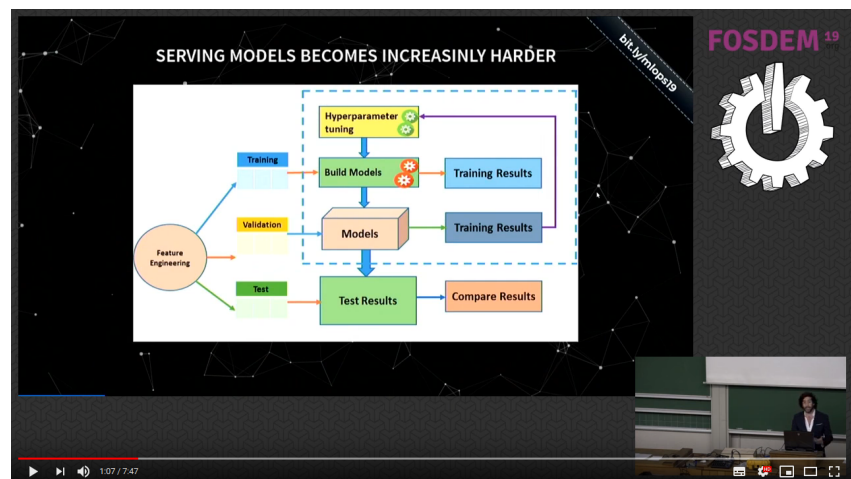

Want to receive recurrent updates on this repo and other advancements?

You can join the Machine Learning Engineer newsletter. You will receive updates on open source frameworks,

tutorials and articles curated by machine learning professionals.
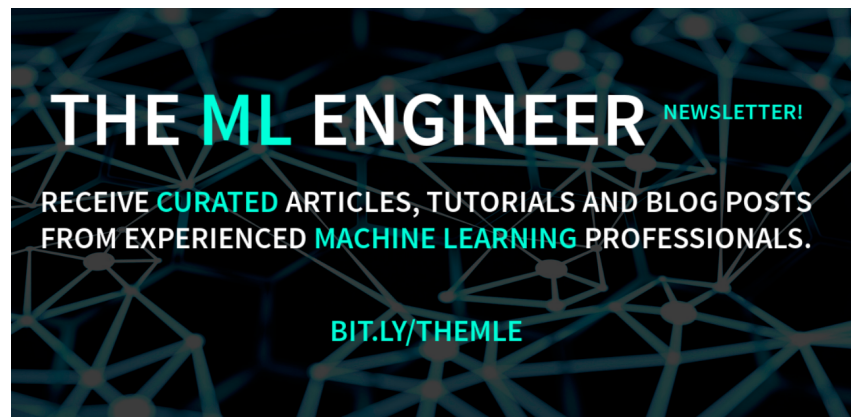
Also check out the [Awesome Artificial Intelligence Guidelines](#) List, where we aim to map the landscape of "Frameworks", "Codes of Ethics", "Guidelines", "Regulations", etc related to Artificial Intelligence.
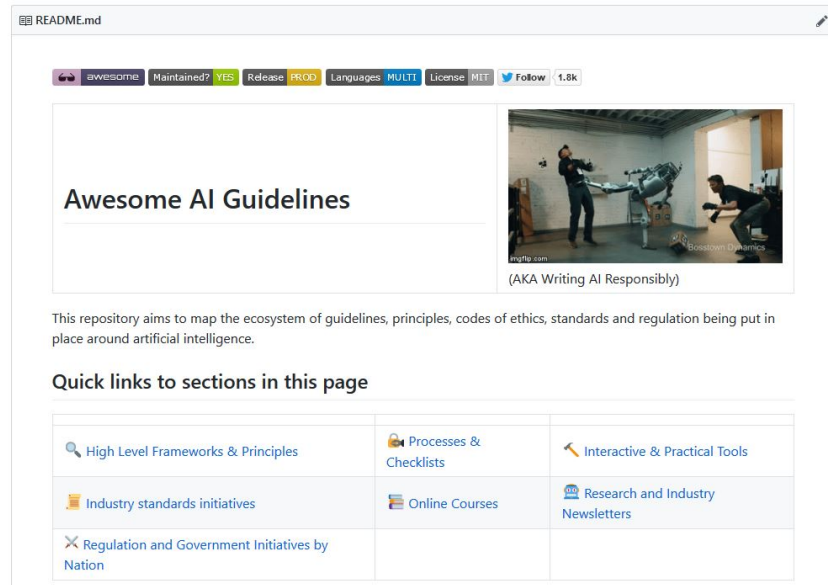
Main Content

Explaining Black Box Models and Datasets

- [Aequitas](#) **Stars** `468` - An open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive risk-assessment tools.
- [Alibi](#) **Stars** `1.6k` - Alibi is an open source Python library aimed at machine learning model inspection and interpretation. The initial focus on the library is on black-box, instance based model explanations.
- [anchor](#) **Stars** `698` - Code for the paper ["High precision model agnostic explanations"](#), a model-agnostic system that explains the behaviour of complex models with high-precision rules called anchors.

- captum `Stars 3.1k` - model interpretability and understanding library for PyTorch developed by Facebook. It contains general purpose implementations of integrated gradients, saliency maps, smoothgrad, vargrad and others for PyTorch models.
- casme `Stars 72` - Example of using classifier-agnostic saliency map extraction on ImageNet presented on the paper "Classifier-agnostic saliency map extraction".
- ContrastiveExplanation (Foil Trees) `Stars 41` - Python script for model agnostic contrastive/counterfactual explanations for machine learning. Accompanying code for the paper "Contrastive Explanations with Local Foil Trees".
- DeepLIFT `Stars 622` - Codebase that contains the methods in the paper "Learning important features through propagating activation differences". Here is the slides and the video of the 15 minute talk given at ICML.
- DeepVis Toolbox `Stars 3.8k` - This is the code required to run the Deep Visualization Toolbox, as well as to generate the neuron-by-neuron visualizations using regularized optimization. The toolbox and methods are described casually here and more formally in this paper.
- ELI5 `Stars 2.5k` - "Explain Like I'm 5" is a Python package which helps to debug machine learning classifiers and explain their predictions.
- FACETS - Facets contains two robust visualizations to aid in understanding and analyzing machine learning datasets. Get a sense of the shape of each feature of your dataset using Facets Overview, or explore individual observations using Facets Dive.
- Fairlearn `Stars 1.2k` - Fairlearn is a python toolkit to assess and mitigate unfairness in machine learning models.
- FairML `Stars 325` - FairML is a python toolbox auditing the machine learning models for bias.
- fairness `Stars 135` - This repository is meant to facilitate the benchmarking of fairness aware machine learning algorithms based on this paper.
- GEBI - Global Explanations for Bias Identification `Stars 16` - An attention-based summarized post-hoc explanations for detection and identification of bias in data. We propose a global explanation and introduce a step-by-step framework on how to detect and test bias. Python package for image data.
- IBM AI Explainability 360 `Stars 1.1k` - Interpretability and explainability of data and machine learning models including a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics.
- IBM AI Fairness 360 `Stars 1.7k` - A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

- iNNvestigate `Stars` `968` - An open-source library for analyzing Keras models visually by methods such as DeepTaylor-Decomposition, PatternNet, Saliency Maps, and Integrated Gradients.
- Integrated-Gradients `Stars` `459` - This repository provides code for implementing integrated gradients for networks with image inputs.
- InterpretML `Stars` `4.7k` - InterpretML is an open-source package for training interpretable models and explaining blackbox systems.
- keras-vis `Stars` `2.9k` - keras-vis is a high-level toolkit for visualizing and debugging your trained keras neural net models. Currently supported visualizations include: Activation maximization, Saliency maps, Class activation maps.
- L2X `Stars` `102` - Code for replicating the experiments in the paper "Learning to Explain: An Information-Theoretic Perspective on Model Interpretation" at ICML 2018
- Lightly `Stars` `1.5k` - A python framework for self-supervised learning on images. The learned representations can be used to analyze the distribution in unlabeled data and rebalance datasets.
- Lightwood `Stars` `253` - A Pytorch based framework that breaks down machine learning problems into smaller blocks that can be glued together seamlessly with an objective to build predictive models with one line of code.
- LIME `Stars` `9.7k` - Local Interpretable Model-agnostic Explanations for machine learning models.
- LOFO Importance `Stars` `455` - LOFO (Leave One Feature Out) Importance calculates the importances of a set of features based on a metric of choice, for a model of choice, by iteratively removing each feature from the set, and evaluating the performance of the model, with a validation scheme of choice, based on the chosen metric.
- MindsDB `Stars` `6.5k` - MindsDB is an Explainable AutoML framework for developers. With MindsDB you can build, train and use state of the art ML models in as simple as one line of code.
- mljar-supervised `Stars` `1.9k` - An Automated Machine Learning (AutoML) python package for tabular data. It can handle: Binary Classification, MultiClass Classification and Regression. It provides feature engineering, explanations and markdown reports.
- NETRON `Stars` `18k` - Viewer for neural network, deep learning and machine learning models.
- pyBreakDown `Stars` `41` - A model agnostic tool for decomposition of predictions from black boxes. Break Down Table shows contributions of every variable to a final prediction.

- rationale `Stars` `347` - Code to implement learning rationales behind predictions with code for paper "Rationalizing Neural Predictions"
- responsibly `Stars` `65` - Toolkit for auditing and mitigating bias and fairness of machine learning systems
- SHAP `Stars` `16k` - SHapley Additive exPlanations is a unified approach to explain the output of any machine learning model.
- SHAPash `Stars` `1.7k` - Shapash is a Python library that provides several types of visualization that display explicit labels that everyone can understand.
- Skater `Stars` `1k` - Skater is a unified framework to enable Model Interpretation for all forms of model to help one build an Interpretable machine learning system often needed for real world use-cases
- Tensorboard's Tensorboard WhatIf `Stars` `5.8k` - Tensorboard screen to analyse the interactions between inference results and data inputs.
- Tensorflow's cleverhans `Stars` `5.5k` - An adversarial example library for constructing attacks, building defenses, and benchmarking both. A python library to benchmark system's vulnerability to adversarial examples
- tensorflow's lucid `Stars` `4.4k` - Lucid is a collection of infrastructure and tools for research in neural network interpretability.
- tensorflow's Model Analysis `Stars` `1.1k` - TensorFlow Model Analysis (TFMA) is a library for evaluating TensorFlow models. It allows users to evaluate their models on large amounts of data in a distributed manner, using the same metrics defined in their trainer.
- themis-ml `Stars` `93` - themis-ml is a Python library built on top of pandas and sklearn that implements fairness-aware machine learning algorithms.
- Themis `Stars` `86` - Themis is a testing-based approach for measuring discrimination in a software system.
- TreeInterpreter `Stars` `701` - Package for interpreting scikit-learn's decision tree and random forest predictions. Allows decomposing each prediction into bias and feature contribution components as described in http://blog.datadive.net/interpreting-random-forests/.
- woe `Stars` `236` - Tools for WoE Transformation mostly used in ScoreCard Model for credit rating
- XAI - eXplainableAI `Stars` `806` - An eXplainability toolbox for machine learning.

Privacy Preserving Machine Learning

- Flower ⓞ **Stars** ‹ 952 › - Flower is a Federated Learning Framework with a unified approach. It enables the federation of any ML workload, with any ML framework, and any programming language.
- Google's Differential Privacy ⓞ **Stars** ‹ 2.5k › - This is a C++ library of ε-differentially private algorithms, which can be used to produce aggregate statistics over numeric data sets containing private or sensitive information.
- Intel Homomorphic Encryption Backend ⓞ **Stars** ‹ 208 › - The Intel HE transformer for nGraph is a Homomorphic Encryption (HE) backend to the Intel nGraph Compiler, Intel's graph compiler for Artificial Neural Networks.
- Microsoft SEAL ⓞ **Stars** ‹ 2.5k › - Microsoft SEAL is an easy-to-use open-source (MIT licensed) homomorphic encryption library developed by the Cryptography Research group at Microsoft.
- PySyft ⓞ **Stars** ‹ 8k › - A Python library for secure, private Deep Learning. PySyft decouples private data from model training, using Multi-Party Computation (MPC) within PyTorch.
- Rosetta ⓞ **Stars** ‹ 445 › - A privacy-preserving framework based on TensorFlow with customized backend Operations using Multi-Party Computation (MPC). Rosetta reuses the APIs of TensorFlow and allows to transfer original TensorFlow codes into a privacy-preserving manner with minimal changes.
- Substra ⓞ **Stars** ‹ 139 › - Substra is an open-source framework for privacy-preserving, traceable and collaborative Machine Learning.
- Tensorflow Privacy ⓞ **Stars** ‹ 1.6k › - A Python library that includes implementations of TensorFlow optimizers for training machine learning models with differential privacy.
- TF Encrypted ⓞ **Stars** ‹ 1k › - A Framework for Confidential Machine Learning on Encrypted Data in TensorFlow.
- Uber SQL Differencial Privacy ⓞ **Stars** ‹ 375 › - Uber's open source framework that enforces differential privacy for general-purpose SQL queries.

Model and Data Versioning

- Aim ⓞ **Stars** ‹ 2.3k › - A super-easy way to record, search and compare AI experiments.
- Apache Marvin ⓞ **Stars** ‹ 94 › is a platform for model deployment and versioning that hides all complexity under the hood: data scientists just need to set up the server and write their code in an extended jupyter notebook.
- Catalyst ⓞ **Stars** ‹ 2.9k › - High-level utils for PyTorch DL & RL research. It was developed with a focus on reproducibility, fast experimentation and code/ideas reusing.

- ClearML `Stars` `3.1k` - Auto-Magical Experiment Manager & Version Control for AI (previously Trains).
- D6tflow `Stars` `939` - A python library that allows for building complex data science workflows on Python.
- Data Version Control (DVC) `Stars` `9.6k` - A git fork that allows for version management of models.
- Dolt `Stars` `10k` - Dolt is a SQL database that you can fork, clone, branch, merge, push and pull just like a git repository.
- FGLab `Stars` `220` - Machine learning dashboard, designed to make prototyping experiments easier.
- Flor `Stars` `112` - Easy to use logger and automatic version controller made for data scientists who write ML code
- GUILD AI `Stars` `684` - Open source toolkit that automates and optimizes machine learning experiments.
- Hub `Stars` `4.4k` - Store, access & manage datasets with version-control for PyTorch/TensorFlow locally or on any cloud with scalable data pipelines.
- Hangar `Stars` `184` - Version control for tensor data, git-like semantics on numerical data with high speed and efficiency.
- Keepsake `Stars` `1.6k` - Version control for machine learning.
- lakeFS `Stars` `2.4k` - Repeatable, atomic and versioned data lake on top of object storage.
- MLflow `Stars` `12k` - Open source platform to manage the ML lifecycle, including experimentation, reproducibility and deployment.
- MLWatcher `Stars` `90` - MLWatcher is a python agent that records a large variety of time-serie metrics of your running ML classification algorithm. It enables you to monitor in real time.
- ModelChimp `Stars` `124` - Framework to track and compare all the results and parameters from machine learning models (Video)
- ModelDB `Stars` `1.4k` - An open-source system to version machine learning models including their ingredients code, data, config, and environment and to track ML metadata across the model lifecycle.
- ModelStore `Stars` `159` - An open-source Python library that allows you to version, export, and save a machine learning model to your cloud storage provider.
- Pachyderm `Stars` `5.4k` - Open source distributed processing framework build on Kubernetes focused mainly on dynamic building of production machine learning pipelines - (Video)

- Polyaxon ⭘ Stars ⟨ 3.1k ⟩ - A platform for reproducible and scalable machine learning and deep learning on kubernetes. - (Video)
- PredictionIO ⭘ Stars ⟨ 13k ⟩ - An open source Machine Learning Server built on top of a state-of-the-art open source stack for developers and data scientists to create predictive engines for any machine learning task
- Quilt Data ⭘ Stars ⟨ 1.1k ⟩ - Versioning, reproducibility and deployment of data and models.
- Sacred ⭘ Stars ⟨ 3.8k ⟩ - Tool to help you configure, organize, log and reproduce machine learning experiments.
- steppy ⭘ Stars ⟨ 132 ⟩ - Lightweight, Python3 library for fast and reproducible machine learning experimentation. Introduces simple interface that enables clean machine learning pipeline design.
- Studio.ML ⭘ Stars ⟨ 374 ⟩ - Model management framework which minimizes the overhead involved with scheduling, running, monitoring and managing artifacts of your machine learning experiments.
- TerminusDB ⭘ Stars ⟨ 1.7k ⟩ - A graph database management system that stores data like git.

Model Training Orchestration

- CML ⭘ Stars ⟨ 3.1k ⟩ - Continuous Machine Learning (CML) is an open-source library for implementing continuous integration & delivery (CI/CD) in machine learning projects.
- Determined ⭘ Stars ⟨ 1.7k ⟩ - Deep learning training platform with integrated support for distributed training, hyperparameter tuning, and model management (supports Tensorflow and Pytorch).
- Flyte ⭘ Stars ⟨ 2.2k ⟩ - Lyft's Cloud Native Machine Learning and Data Processing Platform. (Demo)
- Hopsworks ⭘ Stars ⟨ 698 ⟩ - Hopsworks is a data-intensive platform for the design and operation of machine learning pipelines that includes a Feature Store. (Video).
- Kubeflow ⭘ Stars ⟨ 11k ⟩ - A cloud native platform for machine learning based on Google's internal machine learning pipelines.
- MLeap ⭘ Stars ⟨ 1.4k ⟩ - Standardisation of pipeline and model serialization for Spark, Tensorflow and sklearn
- NVIDIA TensorRT ⭘ Stars ⟨ 5.1k ⟩ - TensorRT is a C++ library for high performance inference on NVIDIA GPUs and deep learning accelerators.
- Onepanel ⭘ Stars ⟨ 588 ⟩ - Production scale vision AI platform, with fully integrated components for model building, automated labeling, data processing and model

training pipelines.

- Open Platform for AI `Stars` `2.3k` - Platform that provides complete AI model training and resource management capabilities.
- Planout `Stars` `1.6k` - PlanOut is a multi-platform framework and programming language for online field experimentation. PlanOut was created to make it easy to run and iterate on sophisticated experiments, while satisfying the constraints of deployed Internet services with many users.
- PyCaret `Stars` `5.5k`) - low-code library for training and deploying models (scikit-learn, XGBoost, LightGBM, spaCy)
- Redis-ML `Stars` `305` - Module available from unstable branch that supports a subset of ML models as Redis data types. (Replaced by Redis AI)
- Skaffold `Stars` `13k` - Skaffold is a command line tool that facilitates continuous development for Kubernetes applications. You can iterate on your application source code locally then deploy to local or remote Kubernetes clusters.
- Tensorflow Extended (TFX) `Stars` `1.7k` - Production oriented configuration framework for ML based on TensorFlow, incl. monitoring and model version management.
- TonY `Stars` `672` - TonY is a framework to natively run deep learning jobs on Apache Hadoop. It currently supports TensorFlow, PyTorch, MXNet and Horovod.
- ZenML `Stars` `1.9k` - ZenML is an extensible, open-source MLOps framework to create reproducible ML pipelines with a focus on automated metadata tracking, caching, and many integrations to other tools.

Model Serving and Monitoring

- Backprop `Stars` `224` - Backprop makes it simple to use, finetune, and deploy state-of-the-art ML models.
- BentoML `Stars` `3.4k` - BentoML is an open source framework for high performance ML model serving
- Cortex `Stars` `7.7k` - Cortex is an open source platform for deploying machine learning models—trained with any framework—as production web services. No DevOps required.
- Deepchecks `Stars` `1.4k` - Deepchecks is an open source package for comprehensively validating your machine learning models and data with minimal effort during development, deployment or in production.
- DeepDetect `Stars` `10` - Machine Learning production server for TensorFlow, XGBoost and Cafe models written in C++ and maintained by Jolibrain

- Evidently `⚙ Stars 2.3k` - Evidently helps analyze machine learning models during development, validation, or production monitoring. The tool generates interactive reports from pandas DataFrame.
- ForestFlow `⚙ Stars 57` - Cloud-native machine learning model server.
- Jina `⚙ Stars 14k` - Cloud native search framework that supports to use deep learning/state of the art AI models for search.
- KFServing `⚙ Stars 1.4k` - Serverless framework to deploy and monitor machine learning models in Kubernetes - (Video)
- m2cgen `⚙ Stars 2.1k` - A lightweight library which allows to transpile trained classic machine learning models into a native code of C, Java, Go, R, PHP, Dart, Haskell, Rust and many other programming languages.
- MLServer `⚙ Stars 207` - An inference server for your machine learning models, including support for multiple frameworks, multi-model serving and more
- mltrace `⚙ Stars 375` - a lightweight, open-source Python tool to get "bolt-on" observability in ML pipelines.
- Model Server for Apache MXNet (MMS) `⚙ Stars 862` - A model server for Apache MXNet from Amazon Web Services that is able to run MXNet models as well as Gluon models (Amazon's SageMaker runs a custom version of MMS under the hood)
- Mosec `⚙ Stars 69` - A rust-powered and multi-stage pipelined model server which offers dynamic batching and more. Super easy to implement and deploy as micro-services.
- OpenScoring `⚙ Stars 561` - REST web service for scoring PMML models built and maintained by OpenScoring.io
- Pandas Profiling - Creates HTML profiling reports from pandas DataFrame objects. It extends the pandas DataFrame with df.profile_report() for quick data analysis.
- Redis-AI `⚙ Stars 674` - A Redis module for serving tensors and executing deep learning models. Expect changes in the API and internals.
- Seldon Core `⚙ Stars 3.1k` - Open source platform for deploying and monitoring machine learning models in kubernetes - (Video)
- Tempo `⚙ Stars 85` - Open source SDK that provides a unified interface to multiple MLOps projects that enable data scientists to deploy and productionise machine learning systems.
- Tensorflow Serving `⚙ Stars 5.4k` - High-performant framework to serve Tensorflow models via grpc protocol able to handle 100k requests per second per core
- TorchServe `⚙ Stars 2.5k` - TorchServe is a flexible and easy to use tool for serving PyTorch models.

- [Transformer-deploy](#) `⌬ Stars` `834` - Transformer-deploy is an efficient, scalable and enterprise-grade CPU/GPU inference server for Hugging Face transformer models.
- [Triton Inference Server](#) `⌬ Stars` `3.5k` - Triton is a high performance open source serving software to deploy AI models from any framework on GPU & CPU while maximizing utilization.
- [WhyLogs](#) `⌬ Stars` `933` - Lightweight solution for profiling and monitoring your ML data pipeline end-to-end

Adversarial Robustness Libraries

- [AdvBox](#) `⌬ Stars` `1.2k` - generate adversarial examples from the command line with 0 coding using PaddlePaddle, PyTorch, Caffe2, MxNet, Keras, and TensorFlow. Includes 10 attacks and also 6 defenses. Used to implement [StealthTshirt](#) at DEFCON!
- [Adversarial DNN Playground](#) `⌬ Stars` `120` - think [TensorFlow Playground](#), but for Adversarial Examples! A visualization tool designed for learning and teaching - the attack library is limited in size, but it has a nice front-end to it with buttons you can press!
- [AdverTorch](#) `⌬ Stars` `1k` - library for adversarial attacks / defenses specifically for PyTorch.
- [Alibi Detect](#) `⌬ Stars` `1.3k` - alibi-detect is a Python package focused on outlier, adversarial and concept drift detection. The package aims to cover both online and offline detectors for tabular data, text, images and time series. The outlier detection methods should allow the user to identify global, contextual and collective outliers.
- [Artificial Adversary](#) `⌬ Stars` `363` AirBnB's library to generate text that reads the same to a human but passes adversarial classifiers.
- [CleverHans](#) `⌬ Stars` `5.5k` - library for testing adversarial attacks / defenses maintained by some of the most important names in adversarial ML, namely Ian Goodfellow (ex-Google Brain, now Apple) and Nicolas Papernot (Google Brain). Comes with some nice tutorials!
- [Counterfit](#) `⌬ Stars` `502` - Counterfit is a command-line tool and generic automation layer for assessing the security of machine learning systems.
- [DEEPSEC](#) `⌬ Stars` `168` - another systematic tool for attacking and defending deep learning models.
- [EvadeML](#) `⌬ Stars` `152` - benchmarking and visualization tool for adversarial ML maintained by Weilin Xu, a PhD at University of Virginia, working with David Evans. Has a tutorial on re-implementation of one of the most important adversarial defense papers - [feature squeezing](#) (same team).

- **Foolbox** `Stars` `2.2k` – second biggest adversarial library. Has an even longer list of attacks - but no defenses or evaluation metrics. Geared more towards computer vision. Code easier to understand / modify than ART - also better for exploring blackbox attacks on surrogate models.
- **IBM Adversarial Robustness 360 Toolbox (ART)** `Stars` `2.9k` – at the time of writing this is the most complete off-the-shelf resource for testing adversarial attacks and defenses. It includes a library of 15 attacks, 10 empirical defenses, and some nice evaluation metrics. Neural networks only.
- **MIA** `Stars` `96` – A library for running membership inference attacks (MIA) against machine learning models.
- **Nicolas Carlini's Adversarial ML reading list** - not a library, but a curated list of the most important adversarial papers by one of the leading minds in Adversarial ML, Nicholas Carlini. If you want to discover the 10 papers that matter the most - I would start here.
- **Robust ML** - another robustness resource maintained by some of the leading names in adversarial ML. They specifically focus on defenses, and ones that have published code available next to papers. Practical and useful.
- **TextFool** `Stars` `91` – plausible looking adversarial examples for text generation.
- **Trickster** `Stars` `43` – Library and experiments for attacking machine learning in discrete domains using graph search.

Neural Architecture Search

- **Autokeras** `Stars` `8.4k` – AutoML library for Keras based on "Auto-Keras: Efficient Neural Architecture Search with Network Morphism".
- **ENAS via Parameter Sharing** - Efficient Neural Architecture Search via Parameter Sharing by authors of paper.
- **ENAS-PyTorch** `Stars` `2.5k` – Efficient Neural Architecture Search (ENAS) in PyTorch based on this paper.
- **ENAS-Tensorflow** `Stars` `110` – Efficient Neural Architecture search via parameter sharing(ENAS) micro search Tensorflow code for windows user.
- **Katib** `Stars` `1.2k` – A Kubernetes-based system for Hyperparameter Tuning and Neural Architecture Search.
- **Maggy** `Stars` `80` – Asynchronous, directed Hyperparameter search and parallel ablation studies on Apache Spark (Video).
- **Neural Architecture Search with Controller RNN** `Stars` `393` – Basic implementation of Controller RNN from Neural Architecture Search with Reinforcement Learning and

Learning Transferable Architectures for Scalable Image Recognition.
- Neural Network Intelligence ⚫ Stars ‹ 11k - NNI (Neural Network Intelligence) is a toolkit to help users run automated machine learning (AutoML) experiments.

Data Science Notebook Frameworks

- Apache Zeppelin ⚫ Stars ‹ 5.6k - Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and more.
- Binder - Binder hosts notebooks in an executable environment (for free).
- H2O Flow - Jupyter notebook-like interface for H2O to create, save and re-use "flows"
- Hydrogen - A plugin for ATOM that enables it to become a jupyter-notebook-like interface that prints the outputs directly in the editor.
- Jupyter Notebooks ⚫ Stars ‹ 9k - Web interface python sandbox environments for reproducible development
- ML Workspace ⚫ Stars ‹ 2.5k - All-in-one web IDE for machine learning and data science. Combines Jupyter, VS Code, Tensorflow, and many other tools/libraries into one Docker image.
- .NET Interactive ⚫ Stars ‹ 1.7k - .NET Interactive takes the power of .NET and embeds it into your interactive experiences.
- Papermill ⚫ Stars ‹ 4.6k - Papermill is a library for parameterizing notebooks and executing them like Python scripts.
- Ploomber ⚫ Stars ‹ 2.3k - Ploomber allows you to develop workflows in Jupyter and execute them in a distributed environment without code changes. It supports Kubernetes, AWS Batch, and Airflow.
- Polynote ⚫ Stars ‹ 630 - Polynote is an experimental polyglot notebook environment. Currently, it supports Scala and Python (with or without Spark), SQL, and Vega.
- RMarkdown ⚫ Stars ‹ 2.4k - The rmarkdown package is a next generation implementation of R Markdown based on Pandoc.
- Stencila ⚫ Stars ‹ 630 - Stencila is a platform for creating, collaborating on, and sharing data driven content. Content that is transparent and reproducible.
- Voilà ⚫ Stars ‹ 4k - Voilà turns Jupyter notebooks into standalone web applications that can e.g. be used as dashboards.

Industrial Strength Visualisation libraries

- Apache ECharts - Apache ECharts is a powerful, interactive charting and data visualization library for browser.

- Bokeh ⬤ Stars ⟨16k⟩ - Bokeh is an interactive visualization library for Python that enables beautiful and meaningful visual presentation of data in modern web browsers.
- Geoplotlib ⬤ Stars ⟨963⟩ - geoplotlib is a python toolbox for visualizing geographical data and making maps
- ggplot2 ⬤ Stars ⟨5.4k⟩ - An implementation of the grammar of graphics for R.
- gradio ⬤ Stars ⟨5.8k⟩ - Quickly create and share demos of models - by only writing Python. Debug models interactively in your browser, get feedback from collaborators, and generate public links without deploying anything.
- matplotlib ⬤ Stars ⟨15k⟩ - A Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.
- Missingno ⬤ Stars ⟨3.1k⟩ - missingno provides a small toolset of flexible and easy-to-use missing data visualizations and utilities that allows you to get a quick visual summary of the completeness (or lack thereof) of your dataset.
- PDPBox ⬤ Stars ⟨672⟩ - This repository is inspired by ICEbox. The goal is to visualize the impact of certain features towards model prediction for any supervised learning algorithm. (now support all scikit-learn algorithms)
- Perspective ⬤ Stars ⟨4.3k⟩ Streaming pivot visualization via WebAssembly https://perspective.finos.org/
- Pixiedust ⬤ Stars ⟨1k⟩ - PixieDust is a productivity tool for Python or Scala notebooks, which lets a developer encapsulate business logic into something easy for your customers to consume.
- Plotly Dash ⬤ Stars ⟨16k⟩ - Dash is a Python framework for building analytical web applications without the need to write javascript.
- Plotly.py ⬤ Stars ⟨11k⟩ - An interactive, open source, and browser-based graphing library for Python.
- Plotly.NET ⬤ Stars ⟨279⟩ - Plotly.NET provides functions for generating and rendering plotly.js charts in .NET programming languages.
- PyCEbox ⬤ Stars ⟨126⟩ - Python Individual Conditional Expectation Plot Toolbox
- pygal ⬤ Stars ⟨2.5k⟩ - pygal is a dynamic SVG charting library written in python
- Redash ⬤ Stars ⟨21k⟩ - Redash is anopen source visualisation framework that is built to allow easy access to big datasets leveraging multiple backends.
- seaborn ⬤ Stars ⟨9.3k⟩ - Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.
- Streamlit ⬤ Stars ⟨19k⟩ - Streamlit lets you create apps for your machine learning projects with deceptively simple Python scripts. It supports hot-reloading, so your app

updates live as you edit and save your file

- Superset `Stars` `46k` - A modern, enterprise-ready business intelligence web application.
- XKCD-style plots - An XKCD theme for matblotlib visualisations
- yellowbrick `Stars` `3.5k` - yellowbrick is a matplotlib-based model evaluation plots for scikit-learn and other machine learning libraries.

Industrial Strength NLP

- AdaptNLP `Stars` `392` - Built atop Zalando Research's Flair and Hugging Face's Transformers library, AdaptNLP provides Machine Learning Researchers and Scientists a modular and adaptive approach to a variety of NLP tasks with an Easy API for training, inference, and deploying NLP-based microservices.
- Blackstone `Stars` `535` - Blackstone is a spaCy model and library for processing long-form, unstructured legal text. Blackstone is an experimental research project from the Incorporated Council of Law Reporting for England and Wales' research lab, ICLR&D.
- CTRL `Stars` `1.6k` - A Conditional Transformer Language Model for Controllable Generation released by SalesForce
- Facebook's XLM `Stars` `2.6k` - PyTorch original implementation of Cross-lingual Language Model Pretraining which includes BERT, XLM, NMT, XNLI, PKM, etc.
- Flair `Stars` `11k` - Simple framework for state-of-the-art NLP developed by Zalando which builds directly on PyTorch.
- Github's Semantic `Stars` `8.3k` - Github's text library for parsing, analyzing, and comparing source code across many languages .
- GluonNLP `Stars` `2.4k` - GluonNLP is a toolkit that enables easy text preprocessing, datasets loading and neural models building to help you speed up your Natural Language Processing (NLP) research.
- GNES `Stars` `1.2k` - Generic Neural Elastic Search is a cloud-native semantic search system based on deep neural networks.
- Grover `Stars` `830` - Grover is a model for Neural Fake News -- both generation and detection. However, it probably can also be used for other generation tasks.
- Kashgari `Stars` `2.3k` - Kashgari is a simple and powerful NLP Transfer learning framework, build a state-of-art model in 5 minutes for named entity recognition (NER), part-of-speech tagging (PoS), and text classification tasks.
- OpenAI GPT-2 `Stars` `15k` - OpenAI's code from their paper "Language Models are Unsupervised Multitask Learners".

- sense2vec ⓞ **Stars** `1.4k` - A Pytorch library that allows for training and using sense2vec models, which are models that leverage the same approach than word2vec, but also leverage part-of-speech attributes for each token, which allows it to be "meaning-aware"
- Snorkel ⓞ **Stars** `5.1k` - Snorkel is a system for quickly generating training data with weak supervision https://snorkel.org.
- SpaCy ⓞ **Stars** `23k` - Industrial-strength natural language processing library built with python and cython by the explosion.ai team.
- Stable Baselines ⓞ **Stars** `3.5k` - A fork of OpenAI Baselines, implementations of reinforcement learning algorithms http://stable-baselines.readthedocs.io/.
- Tensorflow Lingvo ⓞ **Stars** `2.4k` - A framework for building neural networks in Tensorflow, particularly sequence models. Lingvo: A TensorFlow Framework for Sequence Modeling.
- Tensorflow Text ⓞ **Stars** `926` - TensorFlow Text provides a collection of text related classes and ops ready to use with TensorFlow 2.0.
- Wav2Letter++ - A speech to text system developed by Facebook's FAIR teams.
- YouTokenToMe ⓞ **Stars** `800` - YouTokenToMe is an unsupervised text tokenizer focused on computational efficiency. It currently implements fast Byte Pair Encoding (BPE) [Sennrich et al.].
- Transformers ⓞ **Stars** `61k` - Huggingface's library of state-of-the-art pretrained models for Natural Language Processing (NLP).

Data Pipeline ETL Frameworks

- Apache Airflow ⓞ **Stars** `25k` - Data Pipeline framework built in Python, including scheduler, DAG definition and a UI for visualisation
- Apache Nifi ⓞ **Stars** `3.1k` - Apache NiFi was made for dataflow. It supports highly configurable directed graphs of data routing, transformation, and system mediation logic.
- Argo Workflows ⓞ **Stars** `11k` - Argo Workflows is an open source container-native workflow engine for orchestrating parallel jobs on Kubernetes. Argo Workflows is implemented as a Kubernetes CRD (Custom Resource Definition).
- Azkaban ⓞ **Stars** `4k` - Azkaban is a batch workflow job scheduler created at LinkedIn to run Hadoop jobs. Azkaban resolves the ordering through job dependencies and provides an easy to use web user interface to maintain and track your workflows.
- Basin ⓞ **Stars** `25` - Visual programming editor for building Spark and PySpark pipelines

- Bonobo `Stars` `1.5k` - ETL framework for Python 3.5+ with focus on simple atomic operations working concurrently on rows of data
- Chronos `Stars` `4.3k` - More of a job scheduler for Mesos than ETL pipeline. [OUTDATED]
- Couler `Stars` `698` - Unified interface for constructing and managing machine learning workflows on different workflow engines, such as Argo Workflows, Tekton Pipelines, and Apache Airflow.
- Dagster `Stars` `4.6k` - A data orchestrator for machine learning, analytics, and ETL.
- DBT `Stars` `4.5k` - ETL tool for running transformations inside data warehouses
- Flyte `Stars` `2.2k` - Lyft's Cloud Native Machine Learning and Data Processing Platform. (Demo)
- Genie `Stars` `1.6k` - Job orchestration engine to interface and trigger the execution of jobs from Hadoop-based systems
- Gokart `Stars` `246` - Wrapper of the data pipeline Luigi
- Kedro `Stars` `6.8k` - Kedro is a workflow development tool that helps you build data pipelines that are robust, scalable, deployable, reproducible and versioned. Visualization of the kedro workflows can be done by `kedro-viz`
- Luigi `Stars` `16k` - Luigi is a Python module that helps you build complex pipelines of batch jobs, handling dependency resolution, workflow management, visualisation, etc
- Metaflow `Stars` `5.5k` - A framework for data scientists to easily build and manage real-life data science projects.
- Neuraxle `Stars` `512` - A framework for building neat pipelines, providing the right abstractions to chain your data transformation and prediction steps with data streaming, as well as doing hyperparameter searches (AutoML).
- Oozie - Workflow scheduler for Hadoop jobs
- PipelineX `Stars` `182` - Based on Kedro and MLflow. Full comparison given at https://github.com/Minyus/Python_Packages_for_Pipeline_Workflow
- Prefect Core `Stars` `8.7k` - Workflow management system that makes it easy to take your data pipelines and add semantics like retries, logging, dynamic mapping, caching, failure notifications, and more.
- SETL `Stars` `138` - A simple Spark-powered ETL framework that helps you structure your ETL projects, modularize your data transformation logic and speed up your development.
- Snakemake `Stars` `1.3k` - Workflow management system for reproducible and scalable data analyses.

- Towhee `⚙ Stars` `384` - General-purpose machine learning pipeline for generating embedding vectors using one or many ML models.

Data Labelling Tools and Frameworks

- brat rapid annotation tool `⚙ Stars` `1.5k` - Web-based text annotation tool for Named-Entity-Recogntion task.
- COCO Annotator `⚙ Stars` `1.5k` - Web-based image segmentation tool for object detection, localization and keypoints
- Computer Vision Annotation Tool (CVAT) `⚙ Stars` `7.2k` - OpenCV's web-based annotation tool for both VIDEOS and images for computer algorithms.
- Doccano `⚙ Stars` `6k` - Open source text annotation tools for humans, providing functionality for sentiment analysis, named entity recognition, and machine translation.
- ImageTagger `⚙ Stars` `210` - Image labelling tool with support for collaboration, supporting bounding box, polygon, line, point labelling, label export, etc.
- ImgLab `⚙ Stars` `821` - Image annotation tool for bounding boxes with auto-suggestion and extensibility for plugins.
- Label Studio `⚙ Stars` `8.1k` - Multi-domain data labeling and annotation tool with standardized output format
- Labelimg `⚙ Stars` `17k` - Open source graphical image annotation tool writen in Python using QT for graphical interface focusing primarily on bounding boxes.
- makesense.ai `⚙ Stars` `2.2k` - Free to use online tool for labelling photos. Prepared labels can be downloaded in one of multiple supported formats.
- MedTagger `⚙ Stars` `83` - A collaborative framework for annotating medical datasets using crowdsourcing.
- OpenLabeling `⚙ Stars` `802` - Open source tool for labelling images with support for labels, edges, as well as image resizing and zooming in.
- PixelAnnotationTool `⚙ Stars` `1.2k` - Image annotation tool with ability to "colour" on the images to select labels for segmentation. Process is semi-automated with the watershed marked algorithm of OpenCV
- Rubrix `⚙ Stars` `982` - Open-source tool for tracking, exploring, and labeling data for AI projects.
- Semantic Segmentation Editor `⚙ Stars` `1.2k` - Hitachi's Open source tool for labelling camera and LIDAR data.

- Superintendent `Stars` `160` - superintendent provides an ipywidget-based interactive labelling tool for your data.
- VGG Image Annotator (VIA) - A simple and standalone manual annotation software for image, audio and video. VIA runs in a web browser and does not require any installation or setup.
- Visual Object Tagging Tool (VOTT) `Stars` `3.8k` - Microsoft's Open Source electron app for labelling videos and images for object detection models (with active learning functionality)

Metadata Management

- Amundsen `Stars` `3.2k` - Amundsen is a metadata driven application for improving the productivity of data analysts, data scientists and engineers when interacting with data.
- Apache Atlas `Stars` `1.2k` - Apache Atlas framework is an extensible set of core foundational governance services – enabling enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem.
- DataHub `Stars` `5.3k` - DataHub is LinkedIn's generalized metadata search & discovery tool.
- Marquez `Stars` `1k` - Marquez is an open source metadata service for the collection, aggregation, and visualization of a data ecosystem's metadata.
- Metacat `Stars` `1.2k` - Metacat is a unified metadata exploration API service. Metacat focusses on solving these three problems: 1) Federate views of metadata systems. 2) Allow arbitrary metadata storage about data sets. 3) Metadata discovery.
- ML Metadata `Stars` `454` - a library for recording and retrieving metadata associated with ML developer and data scientist workflows. Also TensorFlow ML Metadata.
- Model Card Toolkit `Stars` `266` - streamlines and automates generation of Model Cards

Data Storage Optimisation

- Alluxio `Stars` `5.6k` - A virtual distributed storage system that bridges the gab between computation frameworks and storage systems.
- Apache Arrow - In-memory columnar representation of data compatible with Pandas, Hadoop-based systems, etc

- Apache Druid ![Stars] `12k` - A high performance real-time analytics database.
  https://druid.apache.org/. An introduction to Druid, your Interactive Analytics at (big)
  Scale.
- Apache Ignite ![Stars] `4.1k` - A memory-centric distributed database, caching, and

  Apache Ignite
- Apache Kafka - Distributed streaming platform framework
- Apache Parquet - On-disk columnar representation of data compatible with Pandas,
  Hadoop-based systems, etc
- Apache Pinot ![Stars] `3.9k` - A realtime distributed OLAP datastore
  https://pinot.apache.org. Comparison of the Open Source OLAP Systems for Big Data:
  ClickHouse, Druid, and Pinot.
- BayesDB - Database that allows for built-in non-parametric Bayesian model discovery
  and queryingi for data on a database-like interface - (Video)
- ClickHouse ![Stars] `23k` - ClickHouse is an open source column oriented database
  management system supported by Yandex - [(Video)](https://
- Delta Lake ![Stars] `4.3k` - Delta Lake is a storage layer that brings scalable, ACID
  transactions to Apache Spark and other big-data engines.
- EdgeDB - NoSQL interface for Postgres that allows for object interaction to data
  stored
- HopsFS ![Stars] `255` - HDFS-compatible file system with scale-out strongly consistent
  metadata.
- InfluxDB ![Stars] `23k` Scalable datastore for metrics, events, and real-time analytics.
- Milvus ![Stars] `10k` Milvus is a cloud-native, open-source vector database built to
  manage embedding vectors generated by machine learning models and neural
  networks.
- TimescaleDB ![Stars] `13k` An open-source time-series SQL database optimized for
  fast ingest and complex queries. Packaged as a PostgreSQL extension. Time-series ML
  in TimescaleDB www.youtube.com/watch?v=zbjub8BQPyE)
- Weaviate ![Stars] `2.4k` - A low-latency vector search engine (GraphQL, RESTful) with
  out-of-the-box support for different media types. Modules include Semantic Search,
  Q&A, Classification, Customizable Models (PyTorch/TensorFlow/Keras), and more.
- Zarr ![Stars] `885` - Python implementation of chunked, compressed, N-dimensional
  arrays designed for use in parallel computing.

Function as a Service Frameworks

- [Apache OpenWhisk](#) `Stars` `5.6k` - Open source, distributed serverless platform that executes functions in response to events at any scale.
- [Fission](#) `Stars` `6.9k` - (Early Alpha) Serverless functions as a service framework on Kubernetes
- [Hydrosphere Mist](#) `Stars` `317` - Serverless proxy for Apache Spark clusters
- [Hydrosphere ML Lambda](#) `Stars` `252` - Open source model management cluster for deploying, serving and monitoring machine learning models and ad-hoc algorithms with a FaaS architecture
- [KNative Serving](#) `Stars` `4.4k` - Kubernetes based serverless microservices with "scale-to-zero" functionality.
- [Nuclio](#) `Stars` `4.3k` - A high-performance serverless framework focused on data, I/O, and compute intensive workloads. It is well integrated with popular data science tools, such as Jupyter and Kubeflow; supports a variety of data and streaming sources; and supports execution over CPUs and GPUs
- [OpenFaaS](#) `Stars` `21k` - Serverless functions framework with RESTful API on Kubernetes

Computation load distribution frameworks

- [Analytics Zoo](#) `Stars` `2.5k` - A unified Data Analytics and AI platform for distributed TensorFlow, Keras and PyTorch on Apache Spark/Flink & Ray
- [Apache Spark MLlib](#) - Apache Spark's scalable machine learning library in Java, Scala, Python and R
- [Bagua](#) `Stars` `711` - Bagua is a performant and flexible distributed training framework for PyTorch, providing a faster alternative to PyTorch DDP and Horovod. It supports advanced distributed training algorithms such as quantization and decentralization.
- [Beam](#) `Stars` `5.4k` Apache Beam is a unified programming model for Batch and Streaming https://beam.apache.org/
- [BigDL](#) `Stars` `3.9k` - Deep learning framework on top of Spark/Hadoop to distribute data and computations across a HDFS system
- [Dask](#) `Stars` `9.8k` - Distributed parallel processing framework for Pandas and NumPy computations - [(Video)](#)
- [DEAP](#) `Stars` `4.7k` - A novel evolutionary computation framework for rapid prototyping and testing of ideas. It seeks to make algorithms explicit and data

structures transparent. It works in perfect harmony with parallelisation mechanisms such as multiprocessing and SCOOP.

- DeepSpeed `Stars` `6.6k` - A deep learning optimization library (lightweight PyTorch wrapper) that makes distributed training easy, efficient, and effective.
- Fiber `Stars` `963` - Distributed computing library for modern computer clusters from Uber.
- Hadoop Open Platform-as-a-service (HOPS) - A multi-tenancy open source framework with RESTful API for data science on Hadoop which enables for Spark, Tensorflow/Keras, it is Python-first, and provides a lot of features
- Hivemind `Stars` `1k` - Decentralized deep learning in PyTorch
- Horovod `Stars` `12k` - Uber's distributed training framework for TensorFlow, Keras, and PyTorch
- NumPyWren `Stars` `84` - Scientific computing framework build on top of pywren to enable numpy-like distributed computations
- PyWren - Answer the question of the "cloud button" for python function execution. It's a framework that abstracts AWS Lambda to enable data scientists to execute any Python function - (Video)
- PyTorch Lightning `Stars` `18k` - Lightweight PyTorch research framework that allows you to easily scale your models to GPUs and TPUs and use all the latest best practices, without the engineering boilerplate - (Video)
- Ray `Stars` `20k` - Ray is a flexible, high-performance distributed execution framework for machine learning (VIDEO)
- TensorFlowOnSpark `Stars` `3.8k` - TensorFlowOnSpark brings TensorFlow programs to Apache Spark clusters.
- Vespa `Stars` `3.9k` Vespa is an engine for low-latency computation over large data sets. https://vespa.ai

Model serialisation formats

- Java PMML API - Java libraries for consuming and producing PMML files containing models from different frameworks, including:
  - pyspark2pmml
  - r2pmml
  - sklearn2pmml
  - sparklyr2pmml
- MMdnn `Stars` `5.6k` - Cross-framework solution to convert, visualize and diagnose deep neural network models.

- [Neural Network Exchange Format (NNEF)](#) - A standardard format to store models across Torch, Caffe, TensorFlow, Theano, Chainer, Caffe2, PyTorch, and MXNet
- [ONNX](#) ⚙ Stars ‹ 12k › - Open Neural Network Exchange Format
- [PFA](#) - Created by the same organisation as PMML, the Predicted Format for Analytics is an emerging standard for statistical models and data transformation engines.
- [PMML](#) - The Predictive Model Markup Language standard in XML - ([Video](#))_

Optimized computation frameworks

- [CuDF](#) ⚙ Stars ‹ 4.6k › - Built based on the Apache Arrow columnar memory format, cuDF is a GPU DataFrame library for loading, joining, aggregating, filtering, and otherwise manipulating data.
- [CuML](#) ⚙ Stars ‹ 2.7k › - cuML is a suite of libraries that implement machine learning algorithms and mathematical primitives functions that share compatible APIs with other RAPIDS projects.
- [CuPy](#) ⚙ Stars ‹ 5.9k › - An implementation of NumPy-compatible multi-dimensional array on CUDA. CuPy consists of the core multi-dimensional array class, cupy.ndarray, and many functions on it.
- [H2O-3](#) ⚙ Stars ‹ 5.8k › - Fast scalable Machine Learning platform for smarter applications: Deep Learning, Gradient Boosting & XGBoost, Random Forest, Generalized Linear Modeling (Logistic Regression, Elastic Net), K-Means, PCA, Stacked Ensembles, Automatic Machine Learning (AutoML), etc.
- [Jax](#) ⚙ Stars ‹ 17k › - Composable transformations of Python+NumPy programs: differentiate, vectorize, JIT to GPU/TPU, and more
- [Modin](#) ⚙ Stars ‹ 7.1k › - Speed up your Pandas workflows by changing a single line of code
- [Numba](#) ⚙ Stars ‹ 7.4k › - A compiler for Python array and numerical functions
- [NumpyGroupies](#) ⚙ Stars ‹ 149 › Optimised tools for group-indexing operations: aggregated sum and more
- [OpenVINO™ integration with TensorFlow](#) ⚙ Stars ‹ 134 › - Highly optimized Neural Network inference with Tensorflow on Intel platforms by adding a single line of code
- [Vaex](#) ⚙ Stars ‹ 7k › Vaex is a high performance Python library for lazy Out-of-Core DataFrames (similar to Pandas), to visualize and explore big tabular datasets. Vaex uses memory mapping, zero memory copy policy and lazy computations for best performance (no memory wasted).
- [Vulkan Kompute](#) ⚙ Stars ‹ 843 › - Blazing fast, lightweight and mobile phone-enabled Vulkan compute framework optimized for advanced GPU data processing usecases.

- Weld `Stars` `2.8k` High-performance runtime for data analytics applications, Interview with Weld's main contributor

Data Stream Processing

- Apache Flink `Stars` `19k` - Open source stream processing framework with powerful stream and batch processing capabilities.
- Apache Samza `Stars` `735` - Distributed stream processing framework. It uses Apache Kafka for messaging, and Apache Hadoop YARN to provide fault tolerance, processor isolation, security, and resource management.
- Brooklin `Stars` `761` - Distributed stream processing framework. It uses Apache Kafka for messaging, and Apache Hadoop YARN to provide fault tolerance, processor isolation, security, and resource management.
- Faust `Stars` `6.1k` - Streaming library built on top of Python's Asyncio library using the async kafka client inspired by the kafka streaming library.
- Kafka Streams `Stars` `22k` - Kafka client library for buliding applications and microservices where the input and output are stored in kafka clusters
- Spark Streaming `Stars` `33k` - Micro-batch processing for streams using the apache spark framework as a backend supporting stateful exactly-once semantics

Outlier and Anomaly Detection

- adtk `Stars` `799` - A Python toolkit for rule-based/unsupervised anomaly detection in time series.
- Alibi-Detect `Stars` `1.3k` - Algorithms for outlier and adversarial instance detection, concept drift and metrics.
- dBoost `Stars` `12` - Outlier detection in heterogeneous datasets using automatic tuple expansion. Paper.
- Deequ `Stars` `2.2k` - A library built on top of Apache Spark for defining "unit tests for data", which measure data quality in large datasets.
- Deep Anomaly Detection with Outlier Exposure `Stars` `427` - Outlier Exposure (OE) is a method for improving anomaly detection performance in deep learning models. Paper
- Open Distro Anomaly Detection `Stars` `75` - The Open Distro (Amazon) for the Elasticsearch Anomaly Detection plugin.
- PyOD `Stars` `5.5k` - A Python Toolbox for Scalable Outlier Detection (Anomaly Detection).

- SUOD (Scalable Unsupervised Outlier Detection) `Stars` `316` - An Acceleration System for Large-scale Outlier Detection (Anomaly Detection)
- Tensorflow Data Validation (TFDV) `Stars` `624` - Library for exploring and validating machine learning data.

Feature Engineering Automation

- auto-sklearn `Stars` `6.2k` - Framework to automate algorithm and hyperparameter tuning for sklearn
- AutoGluon `Stars` `4.4k` - Automated feature, model, and hyperparameter selection for tabular, image, and text data on top of popular machine learning libraries (Scikit-Learn, LightGBM, CatBoost, PyTorch, MXNet)
- AutoML-GS `Stars` `6.3k` - Automatic feature and model search with code generation in Python, on top of common data science libraries (tensorflow, sklearn, etc)
- automl `Stars` `1.6k` - Automated feature engineering, feature/model selection, hyperparam. optimisation
- Colombus - A scalable framework to perform exploratory feature selection implemented in R
- Feature Engine `Stars` `17` - Feature-engine is a Python library that contains several transformers to engineer features for use in machine learning models.
- Featuretools - An open source framework for automated feature engineering
- go-featureprocessing `Stars` `74` - A feature pre-processing framework in Go that matches functionality of sklearn
- keras-tuner `Stars` `2.5k` - Keras Tuner is an easy-to-use, distributable hyperparameter optimization framework that solves the pain points of performing a hyperparameter search. Keras Tuner makes it easy to define a search space and leverage included algorithms to find the best hyperparameter values.
- sklearn-deap `Stars` `688` Use evolutionary algorithms instead of gridsearch in scikit-learn.
- TPOT `Stars` `8.5k` - Automation of sklearn pipeline creation (including feature selection, pre-processor, etc)
- tsfresh `Stars` `6.3k` - Automatic extraction of relevant features from time series
- mljar-supervised `Stars` `1.9k` - An Automated Machine Learning (AutoML) python package for tabular data. It can handle: Binary Classification, MultiClass Classification and Regression. It provides feature engineering, explanations and markdown reports.

Feature Stores

- Butterfree `Stars` `207` - A tool for building feature stores which allows you to transform your raw data into beautiful features.
- Feature Store for Machine Learning (FEAST) `Stars` `3.1k` - Feast (Feature Store) is a tool for managing and serving machine learning features. Feast is the bridge between models and data.
- Hopsworks Feature Store `Stars` `698` - Offline/Online Feature Store for ML (Video).
- Ivory `Stars` `7` - ivory defines a specification for how to store feature data and provides a set of tools for querying it. It does not provide any tooling for producing feature data in the first place. All ivory commands run as MapReduce jobs so it assumed that feature data is maintained on HDFS.
- Veri `Stars` `39` - Veri is a Feature Label Store. Feature Label store allows storing features as keys and labels as values. Querying values is only possible with knn using features. Veri also supports creating sub sample spaces of data by default.

Commercial Platforms

- Algorithmia - Cloud platform to build, deploy and serve machine learning models (Video)
- allegro ai Enterprise - Automagical open-source ML & DL experiment manager and ML-Ops solution.
- Amazon SageMaker - End-to-end machine learning development and deployment interface where you are able to build notebooks that use EC2 instances as backend, and then can host models exposed on an API
- Apheris - A platform for federated and privacy-preserving data science that lets you securely collaborate on AI with partners without sharing any data.
- Arize AI - ML observability and automated model monitoring to help ML practitioners understand how their models perform in production, troubleshoot issues, and improve model performance. ML teams can upload offline (training or validation) baselines into an evaluation/inference store alongside online production data for model validation, drift detection, data quality checks, and model performance management.
- Backprop - Serverless hosting of ML models.
- bigml - E2E machine learning platform.
- Censius - Censius is an AI Observability Platform that assists enterprises in continuously monitoring, analyzing, and explaining their production models. It combines monitoring, accountability, and explainability into one Observability Platform.

- **cnvrg.io** - An end-to-end platform to manage, build and automate machine learning
- **Comet.ml** - Machine learning experiment management. Free for open source and students (Video)
- **Cubonacci** - The Cubonacci platform manages deployment, versioning, infrastructure, monitoring and lineage for you, eliminating risk and minimizing time-to-market.
- **D2iQ KUDO for Kubeflow** - Enterprise machine learning platform that runs in the cloud, on premises (incl. air-gapped), in hybrid environments, or on the edge; based on Kubeflow and open-source Kubernetes Universal Declarative Operators (KUDO).
- **DAGsHub** - Community platform for Open Source ML – Manage experiments, data & models and create collaborative ML projects easily.
- **Dataiku** - Collaborative data science platform powering both self-service analytics and the operationalization of machine learning models in production.
- **DataRobot** - Automated machine learning platform which enables users to build and deploy machine learning models.
- **Datatron** - Machine Learning Model Governance Platform for all your AI models in production for large Enterprises.
- **deepsense AIOps** - Enhances multi-cloud & data center IT Operations via traffic analysis, risk analysis, anomaly detection, predictive maintenance, root cause analysis, service ticket analysis and event consolidation.
- **Deep Cognition Deep Learning Studio** - E2E platform for deep learning.
- **deepsense Safety** - AI-driven solution to increase worksite safety via safety procedure check, thread detection and hazardous zones monitoring.
- **deepsense Quality** - Automating laborious quality control tasks.
- **Diffgram** - Training Data First platform. Database & Training Data Pipelines for Supervised AI. Integrated with GCP, AWS, Azure and top Annotation Supervision UIs (or use built-in Diffgram UI, or build your own). Plus a growing list of integrated service providers! For Computer Vision, NLP, and Supervised Deep Learning / Machine Learning.
- **Google Cloud Machine Learning Engine** - Managed service that enables developers and data scientists to build and bring machine learning models to production.
- **Graphsignal** - Machine learning profiler that helps make model training and inference faster and more efficient.
- **H2O Driverless AI** - Automates key machine learning tasks, delivering automatic feature engineering, model validation, model tuning, model selection and deployment, machine learning interpretability, bring your own recipe, time-series and automatic pipeline generation for model scoring. (Video)
- **Hypervector** - API-driven test data fixtures for machine learning and data science features, making it easy to build automated continuous integration steps for your

data-driven components

- **IBM Watson Machine Learning** - Create, train, and deploy self-learning models using an automated, collaborative workflow.
- **Iguazio Data Science Platform** - Bring your Data Science to life by automating MLOps with end-to-end machine learning pipelines, transforming AI projects into real-world business outcomes, and supporting real-time performance at enterprise scale.
- **Labelbox** - Image labelling service with support for semantic segmentation (brush & superpixels), bounding boxes and nested classifications.
- **Logical Clocks Hopsworks** - Enterprise version of Hopsworks with a Feature Store and scale-out ML pipeline design and operation.
- **MCenter** - MLOps platform automates the deployment, ongoing optimization, and governance of machine learning applications in production.
- **Microsoft Azure Machine Learning service** - Build, train, and deploy models from the cloud to the edge.
- **MLJAR** - Platform for rapid prototyping, developing and deploying machine learning models.
- **Neptune.ai** - community-friendly platform supporting data scientists in creating and sharing machine learning models. Neptune facilitates teamwork, infrastructure management, models comparison and reproducibility.
- **Nimblebox** - A full-stack MLOps platform designed to help data scientists and machine learning practitioners around the world discover, create, and launch multi-cloud apps from their web browser.
- **Prodigy** - Active learning-based data annotation. Allows to train a model and pick most 'uncertain' samples for labeling from an unlabeled pool.
- **Scribble Enrich** - Customizable, auditable, privacy-aware feature store. It is designed to help mid-sized data teams gain trust in the data that they use for training and analysis, and support emerging needs such drift computation and bias assessment.
- **SKIL** - Software distribution designed to help enterprise IT teams manage, deploy, and retrain machine learning models at scale.
- **Skytree 16.0** - End to end machine learning platform (Video)
- **Spell** - Flexible end-to-end MLOps / Machine Learning Platform. (Video)
- **SuperAnnotate** - A complete set of solutions for image and video annotation and an annotation service with integrated tooling, on-demand narrow expertise in various fields, and a custom neural network, automation, and training models powered by AI.
- **Superb AI** - ML DataOps platform providing various tools to build, label, manage and iterate on training data.
- **Syndicai** - Easy-to-use cloud agnostic platform that deploys, manages, and scales any trained AI model in minutes with no configuration & infrastructure setup.

- Talend Studio
- Valohai - Machine orchestration, version control and pipeline management for deep learning.
- Weights & Biases - Machine learning experiment tracking, dataset versioning, hyperparameter search, visualization, and collaboration

Releases 5

🏷 **11-2020 Release** (Latest)
  on Nov 25, 2021

+ 4 releases

Packages

No packages published

Contributors 133

+ 122 contributors

Environments 1

🚀 github-pages (Active)