



Machine Learning Frameworks

And other resources to get you started.

BIG DATA & AI LANDSCAPE 2018

INFRASTRUCTURE

The collage is organized into three main sections, each with a title and a collection of logos:

- HADOOP ON-PREMISE:** Includes logos for Cloudera, Hortonworks, MapR, Pivotal, IBM InfoSphere, and BlueData. Jethro is also listed below the main group.
- HADOOP IN THE CLOUD:** Includes logos for AWS, Microsoft Azure, Google Cloud, IBM InfoSphere BigInsights, Treasure Data, Databricks, Altilscale, CAZENAX, and CenturyLink.
- STREAMING / IN-MEMORY:** Includes logos for AWS, Databricks, StreamSets, Confluent, GridGain, Oracle, DataArtisans, Hazelcast, Terracotta, Jax, and FastData.

Below the logos, there is a row of social media icons for Facebook, Twitter, LinkedIn, and YouTube.

The image displays a collection of logos for various data and analytics companies, organized into four distinct categories. Each category is represented by a colored header bar at the top of its respective column.

- DATA TRANSFORMATION** (Orange header): Includes logos for **talend**, **pentaho**, **alteryx**, **TRIFRATA**, **lumen**, **Panaya**, **StreamSets**, and **UNIFI**.
- DATA INTEGRATION** (Yellow header): Includes logos for **SAP Data Services**, **Informatica**, **TEALUM**, **mapinfo**, **Segment**, **enigma**, **redbus**, **alooma**, **axenty**, **ZALONI**, **Stitch**, **import.io**, **InfoWorks**, and **ATTUNITY**.
- DATA GOVERNANCE** (Green header): Includes logos for **Informatica**, **SailPoint**, **IBM**, **McAfee Stylus Secure Shield**, **colibra**, **Alation**, **Veracode**, **Alkerm**, and **PIXERA**.
- MGMT / MONITORING** (Blue header): Includes logos for **aws**, **New Relic**, **active**, **APPOYNICS**, **nubik**, **WAVEFORM**, **dynatrace**, **splunk**, **Signifix**, **GraphIQ**, **unwired**, **pagerduty**, and **Humany**.

The collage features logos for the following companies:

- Storage:** AWS, Microsoft Azure, Google Cloud, IBM Storage, Pure Storage, Alluxio, Membrane.io, Datomic, Carbonix, Cohesity.
- Cluster Svcs:** AWS, Kubernetes, Docker, Mesosphere, CoreOS, CoreOS, OpenShift, CaaS.
- App Dev:** Logentries, Keen IO, Rainforest, CaaS.
- Crowd-Sourcing:** Amazon Mechanical Turk, Upwork, oDesk, Figure Eight, Scale, Hivive.
- Hardware:** Google TPU, ARM, Intel, NVIDIA, Mythic, Movidius, Wavve, Hailo.
- GPU DBs:** Kinetica, Axiom, Mythic, Blazegraph, Bryte.io, PGStats.

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS 1010DATA vmware TIBCO TERADATA ORACLE NetApp syncsort MAPR cloudera

ANALYTICS

DATA ANALYST PLATFORMS

- Microsoft
- pentaho
- alteryx
- guavus
- AYASDI
- ATTIVO
- Datameer
- Quid
- incorta
- inter ana
- ClearStory
- Origami

DATA SCIENCE PLATFORMS

- IBM
- KNIME
- dataiku
- DOMINO
- rapidminer
- CONTINUUM ANALYTICS
- ALGORITHMIA
- DATAWATCH
- SAS

BI PLATFORMS

- Microsoft
- aws
- Wave Analytics
- loder
- INDUSTRIO
- ATSCALE
- Information Builders
- GoogleData
- birst
- MicroStrategy

VISUALIZATION

- + Tableau
- SAP
- Google Cloud
- Celonis
- Aliko
- Parasite Data
- ZEPL
- Xenova
- ISI
- Ploidy
- CHARTIO
- FEDWAG TECH

MACHINE LEARNING

- aws
- Google Cloud
- H2O
- DataRobot
- gamalan
- ELEMENT
- VIZENSE
- VERSIVE
- Bonsai

COMPUTER VISION

- Microsoft Azure
- Amazon Rekognition
- Clarifai
- Cloud Vision API
- Ever AI
- Deepomatic
- Twintyten
- Neuro

HORIZONTAL AI

- Wolfram
- Sentient
- Affective
- Numenta
- Blue Vision
- Cortana
- Voyager Labs
- Prophesee
- Petuum
- OSARO
- IBM Watson
- Veritas
- OpenScale

SPEECH & NLP

- Google Cloud
- Amazon Alexa
- Semantic Machines
- Mobvoi
- SoundHound Inc.
- Verbena
- Twitter
- Narrative Science
- Efer Technologies
- PRIMER
- Snips
- Yocypop

The diagram illustrates various analytics tools organized into four categories:

- SEARCH ANALYTICS**: Includes Elasticsearch, Oracle Enterprise, Kogniton, Lucidworks, Swiftype, Alpinelabs, Omnisus, Coveo, ATTIVO!, Algolia, MAANA, SINEQUA.
- LOG ANALYTICS & SPUNK**: Includes Sumologic, LOGGY, TIMBUKtu, kibana, logz.io.
- SOCIAL ANALYTICS**: Includes Hootsuite Sprinklr, NETBASE, Synthesio, SimpleReach, bitly, predata, SimilarWeb.
- WEB / MOBILE / COMMERCE ANALYTICS**: Includes Google Analytics, Mixpanel, Amplitude, SUMALL, Airtable, RESCII, SIGOPT, granify, custora.

OPEN SOURCE

The banner displays a wide array of tools used in data science and machine learning, organized into eleven categories:

- FRAMEWORK**: Includes TensorFlow, PyTorch, Keras, JAX, PySpark, Flink, YARN, Hadoop, and Spark.
- QUERY / DATA FLOW**: Includes Spark SQL, Presto, SLAM DATA, Google Cloud Dataflow, and Drift.
- DATA ACCESS**: Includes Cassandra, MongoDB, CouchDB, Redis, and others.
- COORDINATION**: Includes Talend, Apache Airflow, and others.
- STREAMING**: Includes Spark Streaming, Flink, Kafka, and others.
- STAT TOOLS**: Includes Jupyter, Scalalab, and others.
- AI / MACHINE LEARNING / DEEP LEARNING**: Includes TensorFlow, Theano, Caffe, PyTorch, and others.
- SEARCH**: Includes Elasticsearch and Solr.
- LOGGING & MONITORING**: Includes Elasticsearch, Kibana, and others.
- VISUALIZATION**: Includes Tableau, Rodeo, and others.
- COLLABORATION**: Includes Anaconda and others.
- SECURITY**: Includes Apache Ranger and others.

DATA SOURCES & APIs

HEALTH

Apple, VALIDIC, practicefusion, fitbit, GARMIN, myantapp, kinsa

IOT

GE Digital, lifepointe, thingiverse, helium, samvera, bluewin, iotforall

FINANCIAL & ECONOMIC DATA

Bloomberg, THOMSON REUTERS, DOW JONES, SP, CAPITALIQ, CB, xignite, Quandl, INVESTMENT THOUGHTS, PREMISE, estimate, FINANCIAL MARKETS, Eagle Alpha, StockTwits, PLAD, Thomson, Earnest

DATA SCIENCE & AI

AIR / SPACE / SEA: Orbital Insights, planet, Alphasense, AIRBETICS, spire, keepy, AEROSPACE, WINDWARD, teslu, DroneDeploy, Bluewin

PEOPLE / ENTITIES

axion, experian, EPILION, InsideView, Citicore Hexion, Quantcast, BASIS, SAFEGRAPH

LOCATION INTELLIGENCE

FOURSQARE, location, mapbox, senselog, athenaflow, REASON, PlaceIQ, esri, factual, HERE, Mapillary, HERE, cuebio, A. Radu, HERE

OTHER

qualtrics, DATA GOV, enigma, CRUX, HERE, HERE

APPLICATIONS – ENTERPRISE

[illegible]

APPLICATIONS – INDUSTRY

[illegible]

DATA RESOURCES

DATA SERVICES

- Palantir
- UCS
- OPERA
- DATA SCIENCE
- fractal
- kaggle
- DataKind
- EXL
- interplay

INCUBATORS & SCHOOLS

- PLURAL EIGHT
- DA
- galvanize
- DataCamp
- DataElite
- INSIGHT
- The Data Incubator
- netix

RESEARCH

- OpenAI
- facebook research
- MIRI
- VECTOR INSTITUTE
- AI2
- ALL TOP INSTITUTE FOR ARTIFICIAL INTELLIGENCE

STAT TOOLS



ScalaLab



AI / MACHINE LEARNING / DEEP LEARNING



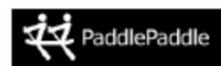
theano



Caffe



OpenAI



FeatureFu



VELES



neon™



DIMSUM

Aerosolve

DSSTNE



We'll focus on Python here.



Best for Classical ML Algorithms



- Python library built on Numpy, SciPy, Matplotlib
- Supports most of the classical supervised and unsupervised algorithms: *linear and logistic regressions, SVM, Naive Bayes, etc.*
- Easy to read *[arguably]*
- Limited support for Neural Networks; i.e. can't be used for Deep Learning
- Has datasets ready to play around with; e.g. MNIST, Iris flower, housing prices
- Huge community

XGBoost, LightGBM and CatBoost: packages for gradient boosting over decision trees algorithms



The most popular Machine Learning and Deep
Learning Library



Popular for Neural Networks and Deep Neural Networks

Offers a tool to visualize your ML models: TensorBoard

A bit of a learning curve...

No pretrained models

Must define the entire computational graph of your model before running

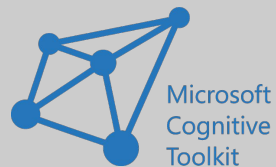
Can be deployed on multiple CPUs and GPUs



Keras



theano

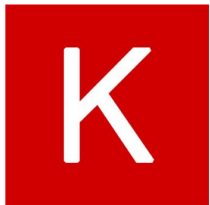


CUDA/cuDNN

BLAS, Eigen

GPU

CPU



Keras

- Adds a lot of convenience to for example tensorflow
- Build up your neural network in building blocks
- You have been seen it a bunch during this week

theano

TensorFlow's less popular cousin...

theano

At its core it is an auto-differentiation framework.

Simpler to use and faster than TensorFlow when it comes to usability and speed

Raw theano is somewhat low-level

Error messages are not too helpful

Can only be deployed on a single GPU

Not fully maintained anymore (basically only maintained to the level that PyMC needs it). Use at your own risk.



Another Deep Learning Library

'the new kid on the block'



More Pythonic and flexible than TensorFlow

Less 'mature' than TensorFlow, but the support and community is growing

Also an auto-differentiation framework at its core

Supports CPU and GPU - accelerated computations

Very popular in research today

Ok, so...which do I use?

Comparisons:

Software	Open Source	Platform	Interface	OpenMP Support	Parallel	Used for
Keras	Yes	Linux, macOS, Windows	Python, R	If using Theano backend		DL
Matlab + NN Toolbox	No	Linux, macOS, Windows	MATLAB	No		DL
Microsoft CNTK	Yes	Linux, Windows	Python, C++, CLI	Yes		DL
PyTorch	Yes	Linux, Windows, macOS	Python	Yes		DL
ONNX	Yes					DL
Scikit Learn	Yes	Linux, Windows, macOS	Python [Library]			ML
Tensorflow	Yes	Linux, Windows, macOS, Android	Python, C++, Java, Go, R, Julia, Swift, Javascript	No		DL

Want to learn more?

Tutorials and Blogs

Coursera & Other Online Courses

Stack Overflow and Google in general ;)

Kaggle

Textbooks on statistics and machine learning.

Suggested Resources:

[Scikit-learn tutorials](#)

An Introduction Book to ML with Jupyter Notebooks: '[Hands-On Machine Learning with Scikit Learn & TensorFlow](#)'

[Keras Tutorial](#): The Ultimate Beginner's Guide to Deep Learning in Python

Stanford's CS231N: <http://cs231n.stanford.edu/>, including [pdf slide sets](#) for more on Convolutional Neural Network

Some Book Suggestions: 'The Elements of Statistical Learning', [Deep Learning Textbook](#)....

Python Data Science Handbook

- Tutorial style book to learn how to properly use Python and its data science tools
- Available for free online in its entirety in the form of Jupyter notebooks

<https://github.com/jakevdp/PythonDataScienceHandbook>

