



Anaconda's Guide to Open-Source

Tools and Libraries for Enterprise Data Science
and Machine Learning

What's Inside

- 3**.....Introduction
- 4**.....Fundamental Data Science Tools and Libraries
- 7**.....Machine Learning
- 10**.....Data Visualization
- 13**.....Image Processing
- 16**.....Scalable Computing
- 19**.....Data Preparation / ETL
- 21**.....Natural Language Processing (NLP)
- 24**.....Looking Ahead: AI Frontiers
- 28**.....How Can I Manage Open Source Enterprise?
- 29**.....About Anaconda Enterprise





Open-source collaboration has led to some of the most innovative and advanced technologies of our time. These are data science and machine learning tools and libraries that equip data scientists in every industry, including engineering, manufacturing, cybersecurity, medicine, genetics, and astronomy. Open-source technologies empower organizations to do breakthrough data science and create differentiating AI and machine learning technologies.

Python is the most commonly used and most recommended language for data science and machine learning, which is why many of the open-source tools and libraries are built for Python. It is also growing in popularity among developers -- it is currently the second most popular language on GitHub. As Python becomes a common language between developers and data scientists, getting machine learning models and applications through production becomes more efficient. All of the tools listed in this guide are compatible with Python.

There are thousands of open-source data science and machine learning packages. This guide focuses on a common set of tools that cover most fundamental tasks in the realm of data science and machine learning. We also touch on a few tools to take ML and data science to the next level as well as cutting-edge tools that are at the forefront of solving the next great challenges in AI.

Fundamental Data Science Tools and Libraries

This collection of open-source Python tools and libraries consists of very popular packages that are frequently used together to do data science. The fundamental tools are not only essential and powerful for individual practitioners, but they are also essential for doing enterprise data science with Python. Many other tools and libraries in the Python data science and ML ecosystem are dependent upon these fundamental packages.



WHAT IT IS:

Jupyter is an open-source project created to support interactive data science and scientific computing across programming languages. Jupyter offers a web-based environment for working with notebooks containing code, data, and text. Jupyter notebooks are the standard workspace for most Python data scientists.

WHAT IT'S USED FOR:

Jupyter notebooks are used to create and share live code, equations, visualizations and text. It has become the tool of choice for presenting data science projects.

PROJECTS:

Jupyter is used by Google, Microsoft, IBM, Bloomberg, NASA, and many other companies and universities. It is safe to say that if an organization has data scientists working in Python, they use Jupyter notebooks.

MORE INFORMATION:

jupyter.org



WHAT IT IS:

A library for tabular data structures, data analysis, and data modeling tools, including built-in plotting using Matplotlib.

WHAT IT'S USED FOR:

Data manipulation and indexing, reshaping and pivoting of data sets, label-based slicing and alignment, high-performance merging and joining of data sets, and time series data analysis. Pandas includes efficient methods for reading and writing a wide variety of data, including CSV files, Excel sheets, and SQL queries.

PROJECTS:

Many companies have found that pandas is easy to use across teams and boosts productivity for data analysis. For example, Appnexus uses pandas across their engineering, mathematician, and analyst teams. Datadog uses pandas to process time series data on their production servers. It's safe to say, if a company is doing data science, they are using Pandas.

LEARN MORE:

<https://pandas.pydata.org/>



WHAT IT IS:

The SciPy library consists of a specific set of fundamental scientific and numerical tools for Python that data scientists use to build their own tools and programs, not to be confused with the SciPy community and the SciPy conference, which include anyone working on scientific computing with Python.

WHAT IT'S USED FOR:

Routines for numerical integration, interpolation, linear algebra, and statistics.

PROJECTS:

SciPy is used by Instacart, WalMart, and Vital Labs, among others. Vital Labs uses SciPy to power their analytics tools.

LEARN MORE:

<https://www.scipy.org/about.html>



WHAT IT IS:

A core package for scientific computing with Python. Numpy enables array formation and basic operations with arrays.

WHAT IT'S USED FOR:

Numpy is used for indexing and sorting but can also be used for linear algebra and other operations. SciPy is more fully featured when it comes to algebra modules and numerical algorithms. Many other data-science libraries for Python are built on NumPy internally, including Pandas and SciPy.

PROJECTS:

Numpy is used by Instacart, Walmart, and Vital Labs for data analysis. It is also used as a foundation in most other Python data-science packages.

LEARN MORE:

<https://numpy.org/>

Machine Learning

Machine learning (ML) is a discipline within AI that involves developing and studying algorithms and models machines use to learn and perform tasks without being explicitly programmed to do so. Deep learning is a subfield of ML that involves processing with neural networks and high-performance computing. These are three of the most popular open-source machine learning technologies.



WHAT IT IS:

TensorFlow is an open-source deep learning platform from Google that includes an ecosystem of tools and libraries that enable the building and deployment of AI and deep learning applications. Keras is a high-level API used to build and train deep learning models, originally as a separate library but now included with TensorFlow.

WHAT IT'S USED FOR:

TensorFlow and Keras are used together to efficiently build, train, and deploy deep learning models, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs). If you download the latest version of TensorFlow, Keras is included.

PROJECTS:

Airbnb uses Tensorflow to classify images and detect objects at scale. Airbus uses TensorFlow to extract information from satellite images, and Twitter used Tensorflow to create their ranked timeline, which shows users the most important tweets first.

LEARN MORE:

<https://www.tensorflow.org/>

WHAT IT IS:

An open-source deep learning framework that consists of fundamental tools and libraries for Python AI and machine learning development.

WHAT IT'S USED FOR:

To build and train deep learning models, such as CNNs and GANs. A rich ecosystem of libraries extends the capabilities of PyTorch for natural language processing and computer vision.

PROJECTS:

Salesforce, among many others, uses PyTorch for natural language processing and multi-task learning.

LEARN MORE:

<https://pytorch.org/>



**WHAT IT IS:**

A powerful and versatile machine learning library for machine learning basics like classification, regression, and clustering. It includes both supervised and unsupervised ML algorithms with important functions like cross-validation and feature extraction. Scikit-learn is the most frequently downloaded machine learning library.

WHAT IT'S USED FOR:

Efficient for predictive analytics and building machine learning models with Python. It also includes tools that make it easy to include deep-learning models in a scikit-learn pipeline.

PROJECTS:

Booking.com and Spotify use scikit-learn for their recommendation engines. Spotify has said scikit-learn is the "most well-designed ML package we've seen so far." J.P. Morgan uses it for predictive analytics, and MARS for supply chain management.

LEARN MORE:

<https://scikit-learn.org/stable/>



Data Visualization

Data visualization is essential to data exploration, analysis, and communication, allowing data scientists to understand their data and share that understanding with others. Python has many, many viz tools available (see pyviz.org/tools.html for a complete list), but we will highlight a few here.



WHAT IT IS:

Matplotlib is the most well-established Python data visualization tool, focusing primarily on two-dimensional plots (line charts, bar charts, scatter plots, histograms, and many others). It works with many GUI interfaces and file formats, but has relatively limited interactive support in web browsers.

WHAT IT'S USED FOR:

Matplotlib is used to analyze, explore, and show relationships between data.

PROJECTS:

Nearly every company with data scientists is using Matplotlib somewhere, whether directly, or often via Pandas or the high-level interfaces made for data scientists like Seaborn, HoloViews, or plotnine. Matplotlib and other open-source Python tools were used to create the first image of a black hole in the Event Horizon Telescope project.

LEARN MORE:

<https://matplotlib.org>



Bokeh & Plotly

WHAT THEY ARE:

Popular and powerful browser-based visualization libraries that let you create interactive, JavaScript-based plots from Python.

WHAT THEY ARE USED FOR:

Bokeh and Plotly create not just static plots, but interactive visualizations with panning, zooming, linking between plots, and other features that let you work in Python but use the power of modern web technologies to share your results widely.

PROJECTS:

Thousands of web sites are built on these tools, either directly or using the higher level interfaces hvPlot, HoloViews, or Chartify (for Bokeh) or Cufflinks and plotly_express (for Plotly).

LEARN MORE:

<https://bokeh.org> and <https://plot.ly/python>



Panel / Voila / Streamlit / Dash

WHAT THEY ARE:

Python frameworks for building custom visualization-rich apps and dashboards for the web.

WHAT THEY ARE USED FOR:

Using Python to create custom applications with live plots, widgets, and other controls to share running applications on the web, backed with the power of Python. Each toolkit has its own focus and strengths: Panel (simple, Pythonic code, easily transitioning from Jupyter to standalone servers), Voila (directly serving Jupyter notebooks), Streamlit (apps from Python scripts), Dash (direct control over HTML/CSS styling, stateless deployment).

PROJECTS:

The best way to see what projects are possible with these tools is to see the examples at awesome-panel.org, voila-gallery.org, awesome-streamlit.org, and dash-gallery.plotly.host

LEARN MORE:

panel.holoviz.org, voila.readthedocs.io, www.streamlit.io, and plot.ly/dash

WHAT IT IS:

HoloViz is an Anaconda project to simplify and improve Python-based visualization by adding high-performance server-side rendering (Datashader), simple plug-in replacement for static visualizations with interactive Bokeh-based plots (hvPlot), and declarative high-level interfaces for building large and complex systems (HoloViews and Param).

WHAT IT'S USED FOR:

The HoloViz project provides extensive free tutorials showing how to use these tools for working with billions of data points interactively, for constructing plots and dashboards from a few lines of Python code, and for working with streaming, geographic, network, or other more complex types of data.

PROJECTS:

See demos and tutorials for the many types of visualizations possible with HoloViz at <http://holoviews.org/gallery/index.html>.

LEARN MORE:

holoviz.org

PIL/Pillow

WHAT IT IS:

Pillow (a “friendly fork” of the older PIL library) is a Python imaging library and a general image processing tool with support for opening, manipulating, and saving images in many different file formats.

WHAT IT'S USED FOR:

Data preparation for image training and basic image manipulation.

PROJECTS:

Data scientists, analysts, and others in banking, finance and health care industries have used Pillow for image manipulation.

LEARN MORE:

<https://pillow.readthedocs.io/>



WHAT IT IS:

Scikit-Image is an open-source Python package containing a collection of image-processing algorithms, including segmentation, geometric transformations, color space manipulation, and feature detection. It uses NumPy arrays as image objects.

WHAT IT'S USED FOR:

scikit-image is used for processing large volumes of images, and it is commonly used for scientific applications ranging from biomedical imaging to astronomy.

PROJECTS:

INRIA has used scikit-image for neuroimaging and computer vision to support leading-edge research.

LEARN MORE:

<https://scikit-image.org>



WHAT IT IS:

An open-source library of programming functions for real-time computer vision with C++, Java, Python and MATLAB interfaces.

WHAT IT'S USED FOR:

OpenCV is the most commonly used library for robotics. It's also used for face tracking and detection and image processing and recognition. OpenCV has been used to build intrusion detection and monitoring tools and to help robots navigate and identify objects.

PROJECTS:

OpenCV is used by Google, Yahoo, Microsoft, Intel, Honda, Toyota for computer vision. Famous projects based on OpenCV include the Robot Operating System and Integrating Vision Toolkit.

LEARN MORE:

<https://opencv.org/>

Scalable Computing

Scalable computing, including distributed and parallel computing, speeds up analysis, model training and performance. It enables multiple tasks and calculations to be performed simultaneously across computers or processors. These packages can be used as boosters for many Python data science and machine learning tasks.



WHAT IT IS:

Numba is a high-performance Python compiler. It makes Python faster and optimizes the performance of Numpy arrays, reaching the speed of FORTRAN and C without a compiler.

WHAT IT'S USED FOR:

Accelerating Python functions and parallelizing algorithms for GPUs and CPUs, such as in Datashader.

PROJECTS:

Datashader, a data visualization tool, uses Numba for acceleration. Fortune 100 finance firms have used it for financial modeling, and it is also commonly used for building simulations. Numba was also used, among other tools, in the Xenon1T experiment to detect dark matter.

LEARN MORE:

<http://numba.pydata.org/>



WHAT IT IS:

Dask is a Python package used to scale NumPy workflows with parallel processing to enable multi-dimensional data analysis, enabling users to store and process data larger than their computer's RAM. Dask can scale out to clusters, or scale down to a single computer. Dask mimics the pandas and NumPy API, making it more intuitive for Python data scientists than Apache Spark.

WHAT IT'S USED FOR:

Dask is used to accelerate processing in a variety of fields, including research in Earth science, satellite imagery, and genomics. It is also used in business and engineering. For example, it is used to increase efficiency in cashflow model management systems and civic modeling.

PROJECTS:

With implementations of Dask, Capital One reduced model training times by 91%. Other organizations have used Dask for genome sequencing, cashflow modeling systems, satellite imagery processing.

LEARN MORE:

<https://stories.dask.org/>

RAPIDS

WHAT IT IS:

RAPIDS is basically a tool for running Pandas, Scikit-Learn, and NetworkX (graph analytics library) on GPUs. It also integrates with some deep learning libraries

WHAT IT'S USED FOR:

Accelerating data science and analytics pipelines by utilizing GPUs.

PROJECTS:

Capital One uses Rapids in conjunction with Dask to speed up their data science workflows and scale on GPUs. They also find that former SAS users and other data scientists because they do not have to learn Spark or Java to be effective.

LEARN MORE:

<https://rapids.ai/about.html>



WHAT IT IS:

A fault-tolerant cluster computing framework and interface for programming clusters launched by UC Berkeley. Developed for Java/Hadoop ecosystem but with support for Python. PySpark is the Python API for Spark.

WHAT IT'S USED FOR:

Spark is a multi-purpose tool that can be used for data preparation and processing as well as training ML algorithms. Spark is great for managing data streams in real time and interactive analytics through interactive queries.

PROJECTS:

Spark is used by a wide variety of companies. eBay uses Apache Spark for log transaction aggregation and analytics. MyFitnessPal uses Spark to clean up users' data and to build recommendation engines for foods and recipes.

LEARN MORE:

<https://spark.apache.org/>

Data Preparation / ETL

Data preparation is a prerequisite to doing data analysis, data science and machine learning, and it can also be the most rigorous and time-consuming part of the whole process. Most data-science workflows initially use custom Pandas and other data-manipulation code, but these data preparation / ETL (extract, transform, and load) tools help automate the process to make data preparation more efficient in production for companies and large organizations.

**WHAT IT IS:**

An open-source workflow automation tool by Apache for creating data workflows, scheduling tasks and monitoring results. It integrates with multiple cloud providers, including AWS, Azure, and Google Cloud.

WHAT IT'S USED FOR:

Airflow is used to manage and automate data pipelines for use in data analysis and machine learning models.

PROJECTS:

Airflow was created by developers from Airbnb for managing big data pipelines from multiple sources. Currently used for data pipeline management by Airbnb, Slack, Walmart, Lyft and Hello Fresh among others.

LEARN MORE:

<https://airflow.apache.org/>

**WHAT IT IS:**

A data ingest/loading library for a wide variety of file formats and data services, with hierarchical cataloguing, searching, and interactivity with remote storage platforms under a single interface.

WHAT IT'S USED FOR:

Intake lets an organization catalog data of all types, including fitted model descriptions, images, and unstructured log entries, so Python data scientists can then focus on their analyses rather than boilerplate I/O code. Catalogs are text files that can easily be shared with others and reused between projects.

PROJECTS:

Intake is currently used by Zillow, NASA, and USGS to catalog data of many types for use in Python.

LEARN MORE:

<https://intake.readthedocs.io>

Natural Language Processing (NLP)

Natural Language Processing (NLP) involves programming machines to parse and understand human language and to interact with humans through both written and spoken language. The field of NLP includes speech recognition, language generation, document analysis, and information retrieval.

NLTK

WHAT IT IS:

An open-source Python natural language toolkit for symbolic and statistical NLP. It includes a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning in multiple languages.

WHAT IT'S USED FOR:

NLTK is used to process human language via tokenization, parsing, classification, and semantic reasoning.

PROJECTS:

NLTK has been used to analyze large bodies of text in academic research projects in a variety of fields, including software engineering studies, cinematics, communications, and sociology.

LEARN MORE:

<https://www.nltk.org/>

gensim

WHAT IT IS:

A Python library for topic modeling, document indexing, and similarity retrieval for large bodies of text with efficient multicore implementations of NLP algorithms.

WHAT IT'S USED FOR:

Gensim is great for the efficient analysis of large bodies of text and extraction of semantic topics.

PROJECTS:

Companies use Gensim to search for relevant information and themes in large bodies of text. For example, DynAdmic, an online video advertising company, uses Gensim to curate digital video ads. Tailwind, an app for scheduling Pinterest and Instagram posts, uses it to help customers post relevant content. Sports Authority uses this tool to analyze text fields from customer surveys and social media commentary.

LEARN MORE:

<https://pypi.org/project/gensim/>



spaCy

WHAT IT IS:

spaCy is an open-source Python library for NLP and one of the fastest, if not the fastest, syntactic parser.

WHAT IT'S USED FOR:

spaCy is used for a wide variety of NLP tasks, especially for large-scale information extraction tasks. It is also used to prepare text for deep learning and is interoperable with TensorFlow, PyTorch, and scikit-learn.

PROJECTS:

spaCy is used by Airbnb, Uber, Stitch Fix, Quora, and many other organizations. Quill used spaCy to develop a free online tool that helps students improve their grammar and writing. It has also been used for quote extraction and attribution.

LEARN MORE:

<https://spacy.io/>





WHAT IT IS:

An open neural network exchange making machine learning models portable between frameworks and platforms. Microsoft and Facebook started this community in 2017 to create an open ecosystem for interchangeable models.

WHAT IT'S USED FOR:

Interoperability and portability. The exchange enables data scientists and developers to move AI models between tools and platforms, which saves a significant amount of time and headaches in the process of operationalizing models. It is also commonly used for serving models.

PROJECTS:

ONNX is used and supported by AMD, AWS, HP, IBM, Intel, NVIDIA, and other companies on the cutting edge of AI/ML.

LEARN MORE:

<https://onnx.ai/>

FairLearn

WHAT IT IS:

A burgeoning project by open-source developers at Microsoft. FairLearn is a Python package for assessing fairness and mitigating unfairness in ML models and AI systems.

WHAT IT'S USED FOR:

Evaluating fairness of AI/ML models and training data and for mitigating bias in models determined to be unfair.

PROJECTS:

Because this project is fairly new, not many companies have published case studies or overviews of their use of the tool. One example project provided is the mitigation of racial disparities in ranking of law school applicants.

LEARN MORE:

<https://github.com/fairlearn/fairlearn/blob/master/README.md>

AI Fairness 360 (AIF360)

WHAT IT IS:

A comprehensive open-source Python toolkit of metrics that checks for and measures bias in datasets and ML models. It also included algorithms to mitigate bias. This toolkit was developed by IBM's open-source team.

WHAT IT'S USED FOR:

Similar to FairLearn, it's used for evaluating fairness of AI/ML models and training data and mitigating bias in current models.

PROJECTS:

|AI Fairness 360 has been used to detect bias in credit scoring algorithms and to mitigate racial bias in healthcare utilization scoring.

LEARN MORE:

<https://aif360.mybluemix.net/>

InterpretML

WHAT IT IS:

An open-source Python package that makes it easy to compare algorithms for interpretability. It provides a "scikit-learn style uniform API" and includes an interactive visualization platform and dashboard so data scientists can compare algorithms with ease.

WHAT IT'S USED FOR:

InterpretML is used to explain any existing "black box" model (models with means of making decisions that are incomprehensible to humans), and it can also be used to train new models that are designed to be interpretable, "glass box" models (models explainable to humans).

PROJECTS:

InterpretML was started by open-source developers at Microsoft, and it has been used to make credit fraud, churn, and medical prediction models more interpretable.

LEARN MORE:

<https://github.com/interpretml/interpret>

LIME

WHAT IT IS:

LIME is a PyPI package and a model-agnostic interpretability tool. LIME explains individual predictions for text classifiers that act on tables or images. Support for scikit-learn classifiers is built into the tool.

WHAT IT'S USED FOR:

Lime is used to help data scientists understand and explain the decisions of black box models with two or more classes. It works by perturbing data samples to understand how this changes the model's predictions, narrowing down the logic that was used to make a particular decision.

PROJECTS:

LIME has been used by both academic and corporate data scientists to understand model decision-making. The main contributor to LIME is a researcher at Microsoft.

LEARN MORE:

<https://github.com/marcotcr/lime>

HOW DO I START USING ALL THESE TOOLS?

All of these libraries and packages can be downloaded individually with pip, but more than 250 of the most commonly used open-source data science and machine learning packages are automatically installed when you download the Anaconda Distribution, and many others can be installed by simply typing `conda install [package-name]`. Anaconda Distribution is an installer and package management system and the easiest and most efficient way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. It updates packages and their dependencies and also creates, saves, loads, and switches between environments on your computer.

 **ANACONDA.DISTRIBUTION**

Learn more

<https://www.anaconda.com/distribution/>

How Can I Manage Open Source in the Enterprise?



While Anaconda Distribution is perfect for individual practitioners, it is not well-equipped for package management or collaboration at the enterprise level. With [Anaconda Team Edition](#), companies can mirror Anaconda's powerful repository onto corporate infrastructure for control over availability, reporting on common vulnerabilities and exposures (CVEs), user access control, license type control, and private and shared channels for package management. Know who's using what packages in which models and blacklist or whitelist packages as needed.



Another option is to manage open-source packages with our end-to-end machine learning platform. [Anaconda Enterprise](#) combines package management, collaboration on projects via Jupyter notebooks, governance and one-click deployment for a full-featured data science and machine learning platform that meets enterprise requirements.

About Anaconda

With more than 20 million users, Anaconda is the world's most popular data science platform and the foundation of modern machine learning. We pioneered the use of Python for data science, champion its vibrant community, and continue to steward open-source projects that make tomorrow's innovations possible. Our enterprise-grade solutions enable corporate, research, and academic institutions around the world to harness the power of open-source for competitive advantage, groundbreaking research, and a better world.

Visit <https://www.anaconda.com> to learn more.

