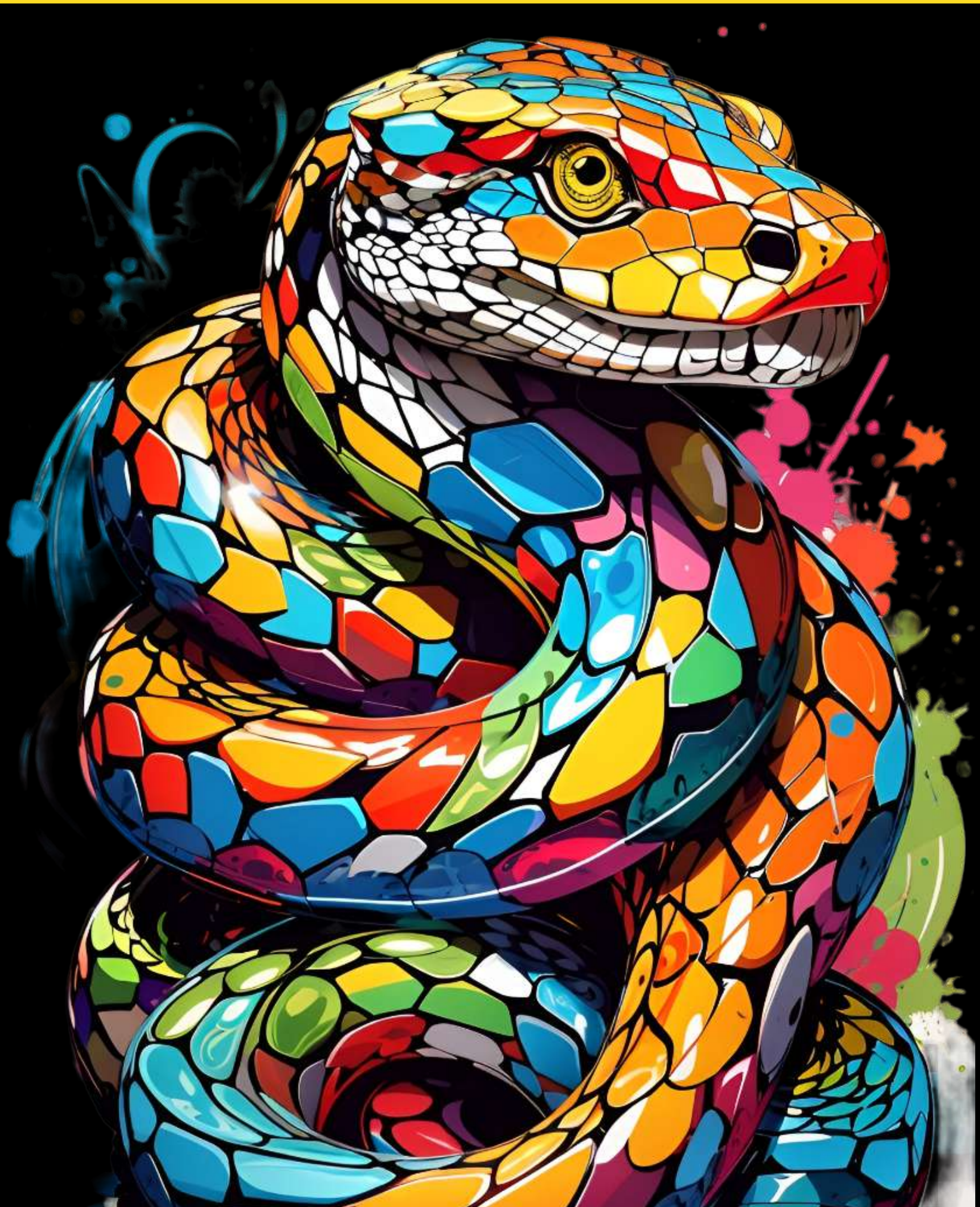
Top Python Libraries for Data Science



hossam-eldin

@codewello

Intro

Python offers a diverse range of data science libraries that continue to grow.

This Carousel focuses on five main areas: Essential Data Science tools, Machine Learning, AutoML, Deep Learning, and Natural Language Processing.

hossam-eldin

@codewello

Staple Python Libraries for Data Science

hossam-eldin

@codewello

NumPy

NumPy is a popular Python library for scientific calculations. It helps with fast math operations on large sets of data, especially matrices. People often choose NumPy Arrays instead of lists because they use less memory and are more convenient and efficient.

⭐ **GitHub Stars: 25.1K**

hossam-eldin

@codewello

Pandas

Pandas is a free tool widely used in data science for tasks like analyzing, manipulating, and cleaning data. It makes data modeling and analysis easier with its simple code. Described as fast, powerful, flexible, Pandas is an open-source tool for data analysis and manipulation.

⭐ **GitHub Stars: 40.4K**

hossam-eldin

@codewello

Matplotlib

Matplotlib is a big library in Python for making different types of visualizations, like fixed, interactive, and animated ones. Many other packages add to what Matplotlib can do, including some easier-to-use plotting tools like Seaborn, HoloViews, ggplot, and more.

⭐ **GitHub Stars: 18.4K**

hossam-eldin

@codewello

Seaborn

Seaborn is a Python data visualization tool built on Matplotlib.

It makes it easy to create attractive statistical visuals for studying data. Seaborn works well with NumPy and pandas.

⭐ **GitHub Stars: 11.4K**

hossam-eldin

@codewello

Plotly

Plotly is a widely-used tool for making interactive data visualizations. It's based on the Plotly JavaScript library (plotly.js) and allows you to create web-friendly visualizations.

⭐ **GitHub Stars: 14.5K**

hossam-eldin

@codewello

Scikit-Learn

Machine learning and scikit-learn go hand in hand. Scikit-learn is a popular Python library for machine learning. It's built on NumPy, SciPy, and Matplotlib, and it's open-source. It's a straightforward and effective tool for predictive data analysis.

⭐ **GitHub Stars: 56.5K**

hossam-eldin

@codewello

Machine Learning Python Libraries

hossam-eldin

@codewello

LightGBM

LightGBM is a widely used tool for boosting in machine learning. It's great because:

1. Trains fast and works efficiently
2. Uses less memory
3. Gives better accuracy
4. Supports parallel, distributed, and GPU learning
5. Handles big data well.

hossam-eldin

@codewello

XGBoost

It uses gradient boosting to solve data science problems quickly.

- Large and growing user base.
- Versatile applications in regression, classification, and ranking.
- Compatible with OS X, Windows, and Linux.
- Easily integrated with cloud services.
- Trusted and used in production by many organizations.

hossam-eldin

@codewello

CatBoost

Catboost is a fast, scalable, high-performance gradient boosting on decision trees library used for ranking, classification, regression, and other machine learning tasks for Python, R, Java, and C++. It supports computation on CPU and GPU.

hossam-eldin

@codewello

Statsmodels

Statsmodels provides classes and functions that allow users to estimate various statistical models, conduct statistical tests, and do statistical data exploration. A comprehensive list of result statistics is then provided for each estimator. The accuray of results can then be tested against existing statistical packages.

hossam-eldin

@codewello

RAPIDS.AI cuDF and cuML

RAPIDS is a set of free software tools that run data science and analytics pipelines on GPUs. It works smoothly on GPU workstations, multi-GPU servers, and clusters with Dask. NVIDIA backs the project, and it uses Numba, Apache Arrow, and other open-source projects.

hossam-eldin

@codewello

RAPIDS.AI cuDF and cuML

cuDF – cuDF is a GPU DataFrame library used to load, join, aggregate, filter, and otherwise manipulate data.

cuML is a set of libraries that applies machine learning algorithms and mathematical functions. It shares compatible APIs with other RAPIDS projects.

hossam-eldin

@codewello

Optuna

This open-source hyperparameter optimization framework is used primarily to automate hyperparameter searches.

hossam-eldin

@codewello

Automated Machine Learning (AutoML) Python Libraries

hossam-eldin

@codewello

PyCaret

This popular machine learning library in Python, known for its open-source nature, simplifies machine learning with minimal code. It's a comprehensive tool for managing models and speeding up the experiment process.

hossam-eldin

@codewello

H2O

H2O is a platform for making smart predictions using big data. It helps build and use machine learning models easily in a business setting. H2O's main code is in Java, and it uses Java's Fork/Join framework for speed and H2O's Map/Reduce framework for distributed computing.

hossam-eldin

@codewello

TPOT

TPOT is a smart tool for making machine learning easier. It works with scikit-learn and uses Genetic Programming to find the best way to analyze data. With its special genetic programming, TPOT can automatically make the data better and create the best machine learning model for accurate results.

hossam-eldin

@codewello

Auto-sklearn

Auto-sklearn is a smart tool for machine learning that's like a replacement for scikit-learn. It automatically fine-tunes settings and picks the best algorithm, saving a lot of time for machine learning users. Its design incorporates recent improvements in meta-learning, ensemble building, and Bayesian optimization.

hossam-eldin

@codewello

Deep Learning Python Libraries

hossam-eldin

@codewello

TensorFlow

TensorFlow is a popular open-source library for high-performance numerical computation developed by the Google Brain team at Google, and a mainstay in the field of deep learning research.

hossam-eldin

@codewello

PyTorch

PyTorch is a powerful tool for machine learning, making it quicker to go from testing ideas to putting them into action. It's a specialized library for deep learning on GPUs and CPUs, serving as an alternative to TensorFlow.

hossam-eldin

@codewello

FastAI

FastAI is a user-friendly deep learning library that effortlessly produces top-notch results. It provides easy-to-use high-level components and interchangeable low-level components for developing new approaches, all without sacrificing usability, flexibility, or performance.

hossam-eldin

@codewello

Keras

Keras is a user-friendly deep learning tool made for people, not machines. It's designed to be easy to use, with simple and consistent interfaces, fewer steps for common tasks, and clear error messages.

TensorFlow liked it so much that they made Keras their main API in the TF 2.0 version.

hossam-eldin

@codewello

PyTorch Lightning

PyTorch Lightning makes using PyTorch easier. It's a fast and lightweight framework that simplifies deep learning experiments. It separates research from engineering, making it easy to understand and reproduce code.

hossam-eldin

@codewello

Python Libraries for Natural Language Processing

hossam-eldin

@codewello

NLTK

NLTK is a key tool for building Python programs that analyze human language. It offers simple interfaces to over 50 datasets and word resources, such as WordNet. NLTK also includes text processing libraries and supports powerful NLP libraries.

hossam-eldin

@codewello

spaCy

spaCy is a powerful open-source library for natural language processing in Python. It's great for big tasks like pulling information from large amounts of text, and it's built for efficiency using Cython. If your application deals with huge web data, spaCy is the perfect choice.

hossam-eldin

@codewello

Gensim

Gensim is a Python library for topic modeling, document indexing, and similarity retrieval with large corpora. Its principle usership is in the information retrieval and natural language processing communities.

hossam-eldin

@codewello

Hugging Face Transformers

Hugging Face Transformers, an open-source library, lets you effortlessly access and train advanced pre-trained models through APIs. This helps cut down on computing expenses, lowers carbon footprint, and saves time compared to training a model from the beginning.

hossam-eldin

@codewello

Thanks !!

if you enjoyed this, please:

- **Follow me** for more badass content
- **Share** this post so others can learn as well
- **Hit the notification** so you don’t miss a thing

And most importantly...

Keep on learning

Happy Coding