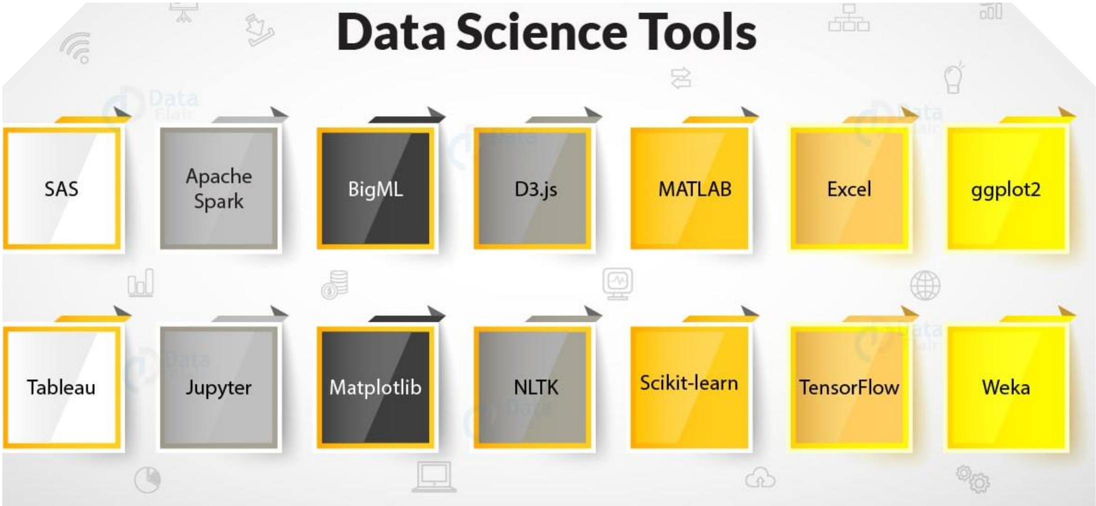
Introduction to Data Science

Data Science has emerged out as one of the most popular fields of 21st Century. Companies employ Data Scientists to help them gain insights about the market and to better their products.

Data Scientists work as decision-makers and are largely responsible for analyzing and handling a large amount of unstructured and structured data.

In order to do so, he requires various tools and **programming languages for Data Science to mend the day** in the way he wants. We will go through some of these data science tools utilizes to analyze and generate predictions.



Top Data Science Tools

Here is the list of 14 best data science tools that most of the data scientists used.

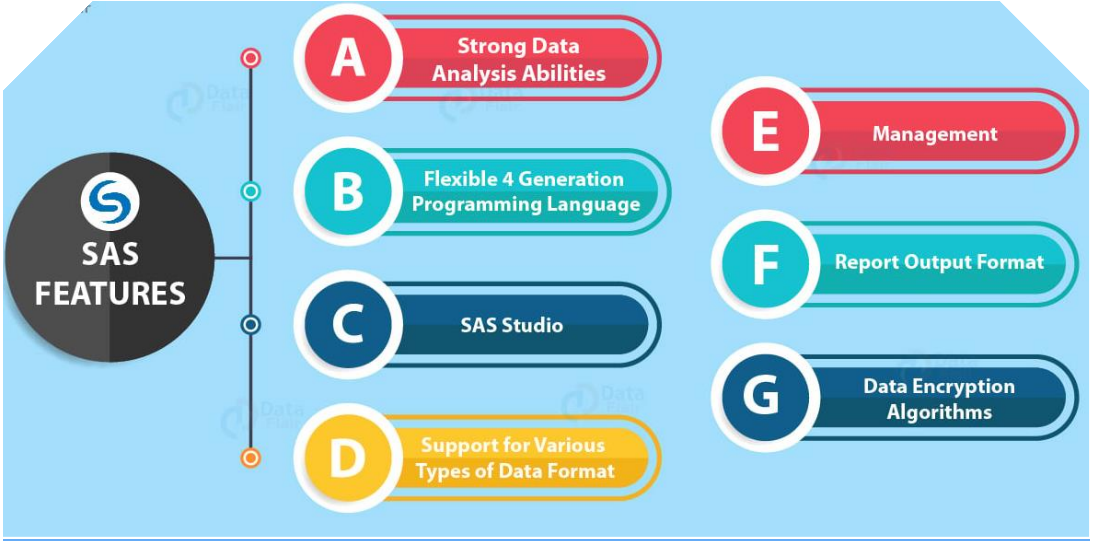
1. SAS

It is one of those data science tools which are specifically designed for statistical operations. **SAS is a closed source proprietary software** that is used by large organizations to analyze data. SAS uses base SAS programming language which for performing statistical modeling.

It is widely used by professionals and companies working on reliable commercial software. **SAS offers numerous statistical libraries** and tools that you as a Data Scientist can use for modeling and organizing their data.

While SAS is highly reliable and has strong support from the company, it is highly expensive and is only used by larger industries. Also, SAS pales in comparison with some of the more modern tools which are open-source.

Furthermore, there are several libraries and packages in SAS that are not available in the base pack and can require an expensive upgradation.

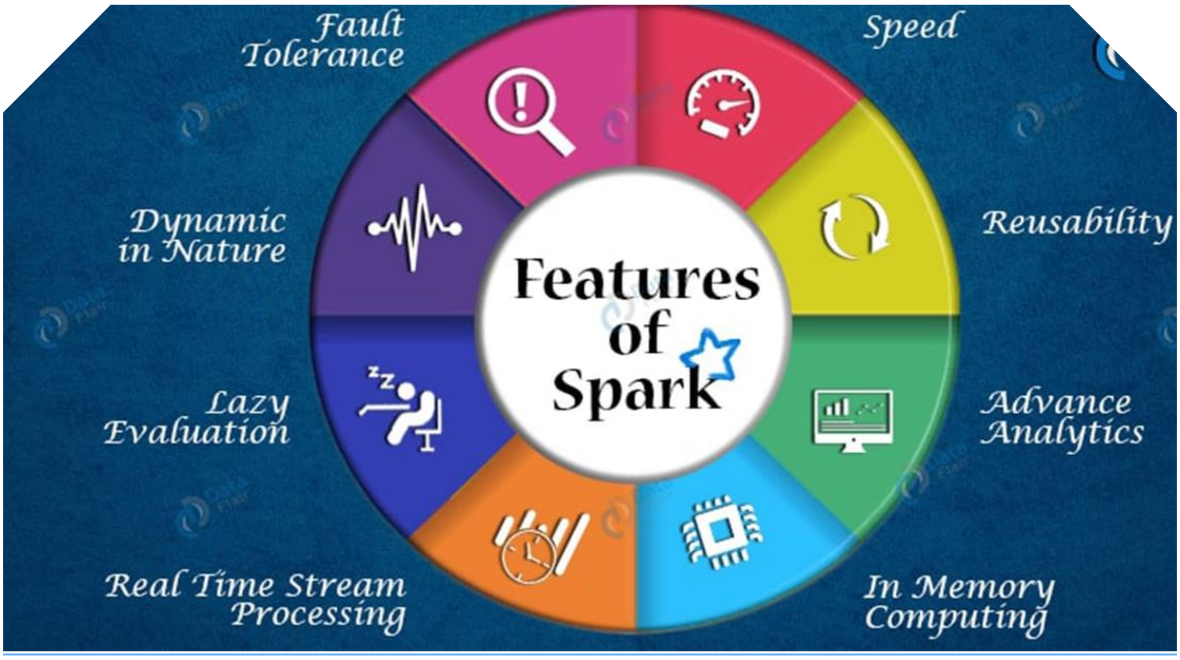

SAS FEATURES
A
Strong Data Analysis Abilities
B
Flexible 4 Generation Programming Language
C
SAS Studio
D
Support for Various Types of Data Format
E
Management
F
Report Output Format
G
Data Encryption Algorithms

2. Apache Spark

Apache Spark or simply Spark is an all-powerful analytics engine and it is the most used Data Science tool. Spark is specifically designed to handle batch processing and **Stream Processing.** It is covered in all data science course.

It comes with many APIs that facilitate Data Scientists to make repeated access to data for Machine Learning, Storage in SQL, etc. It is an improvement over Hadoop and can perform 100 times faster than MapReduce.

Spark has many Machine Learning APIs that can help Data Scientists to make powerful predictions with the given data.



Spark does better than other Big Data Platforms in its ability to handle streaming data. This means that **Spark can process real-time data** as compared to other analytical tools that process only historical data in batches.

Spark offers various APIs that are programmable in Python, Java, and R. But the most powerful conjunction of Spark is with Scala programming language which is based on **Java Virtual Machine** and is cross-platform in nature.

Spark is highly efficient in cluster management which makes it much better than Hadoop as the latter is only used for storage. It is this cluster management system that allows Spark to process applications at a high speed.

3. BigML

BigML, it is another widely used Data Science Tool. It provides a fully interactable, cloud-based GUI environment that you can use for processing **Machine Learning Algorithms**. BigML provides standardized software using cloud computing for industry requirements.

Through it, companies can use Machine Learning algorithms across various parts of their company. For example, it can use this one software across for sales forecasting, risk analytics, and product innovation.

BigML specializes in predictive modeling. It uses a wide variety of Machine Learning algorithms like clustering, classification, time-series forecasting, etc.

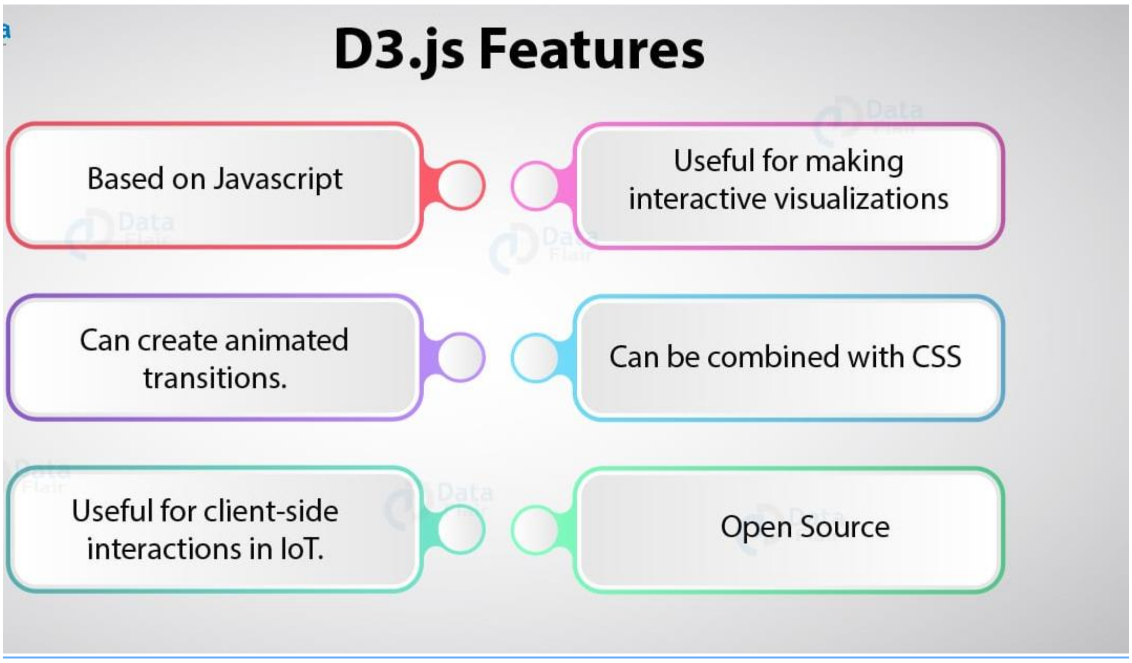
BigML provides an easy to use web-interface using Rest APIs and you can create a free account or a premium account based on your data needs. It allows interactive visualizations of data and provides you with the ability to export visual charts on your mobile or IOT devices.

Furthermore, BigML comes with various automation methods that can help you to automate the tuning of hyperparameter models and even automate the workflow of reusable scripts.

4. D3.js

Javascript is mainly used as a client-side scripting language. D3.js, a Javascript library allows you to make interactive visualizations on your web-browser. With several APIs of D3.js, you can use several functions to create dynamic visualization and analysis of data in your browser.

Another powerful feature of D3.js is the usage of animated transitions. D3.js makes documents dynamic by allowing updates on the client side and actively using the change in data to reflect visualizations on the browser.


D3.js Features
Based on Javascript
Useful for making interactive visualizations
Can create animated transitions.
Can be combined with CSS
Useful for client-side interactions in IoT.
Open Source

You can combine this with CSS to create illustrious and transitory visualizations that will help you to implement customized graphs on web-pages.

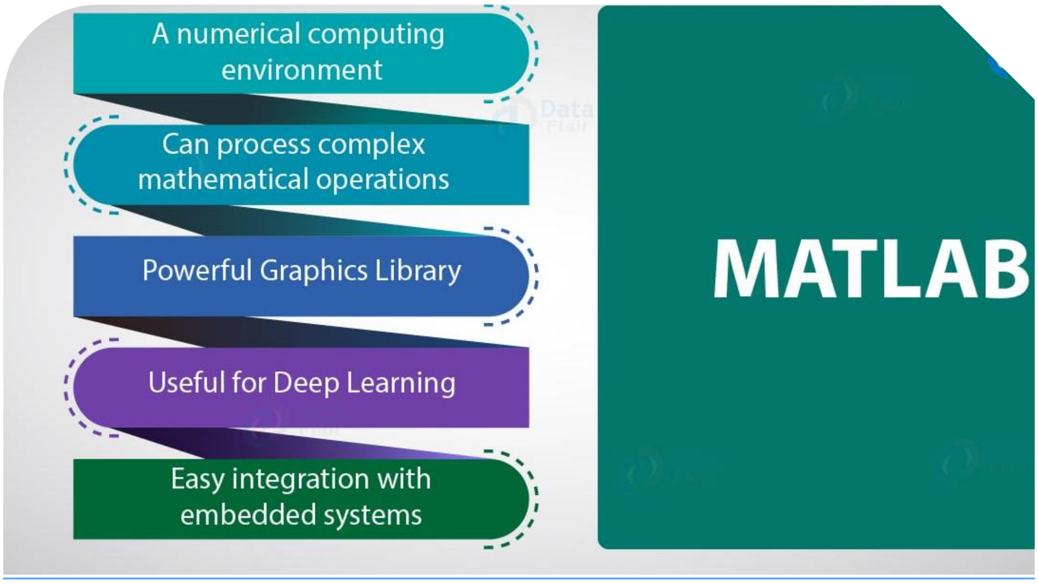
D3.js is a popular option in many fields where data transmission and exploration are crucial due to its versatility and capacity to produce aesthetically attractive and interactive data visualisations. Developers may create robust and dynamic data-driven apps for a variety of uses because of its interaction with web technologies like HTML, SVG, and CSS.

5. MATLAB

MATLAB is a multi-paradigm numerical computing environment for processing mathematical information. It is a closed-source software that facilitates matrix functions, algorithmic implementation and statistical modeling of data. MATLAB is most widely used in several scientific disciplines.

In Data Science, MATLAB is used for simulating **neural networks** and fuzzy logic. Using the MATLAB graphics library, you can create powerful visualizations. MATLAB is also used in image and signal processing.

This makes it a very versatile tool for Data Scientists as they can tackle all the problems, from data cleaning and analysis to more advanced **Deep Learning** algorithms.


A numerical computing environment
Can process complex mathematical operations
Powerful Graphics Library
Useful for Deep Learning
Easy integration with embedded systems
MATLAB

Furthermore, MATLAB’s easy integration for enterprise applications and embedded systems make it an ideal Data Science tool.

It also helps in automating various tasks ranging from the extraction of data to re-use of scripts for decision making. However, it suffers from the limitation of being a closed-source proprietary software.

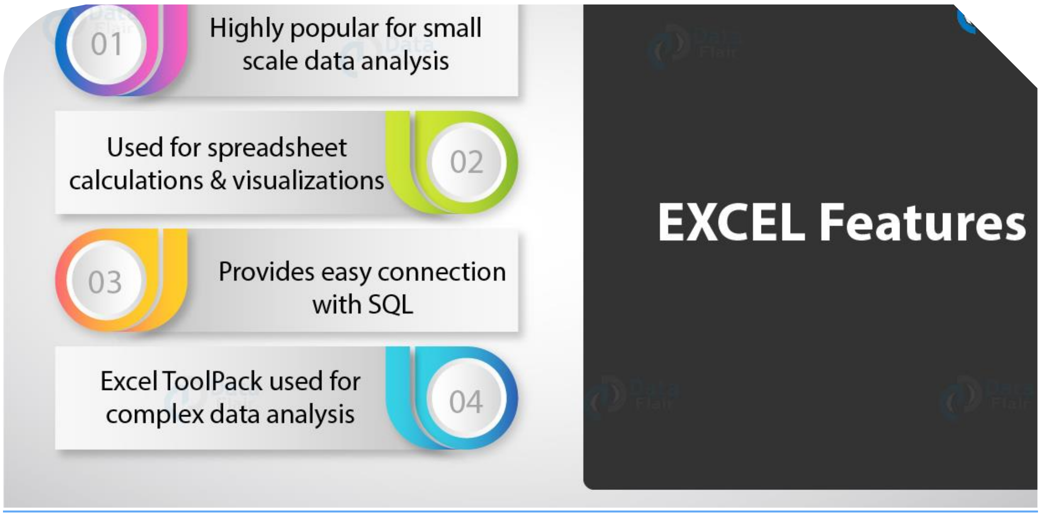
6. Excel

Probably the most widely used Data Analysis tool. Microsoft developed Excel mostly for spreadsheet calculations and today, it is widely used for data processing, visualization, and complex calculations.

Excel is a powerful **analytical tool for Data Science**. While it has been the traditional tool for data analysis, Excel still packs a punch.

Excel comes with various formulae, tables, filters, slicers, etc. You can also create your own custom functions and formulae using Excel. While Excel is not for calculating the huge amount of Data, it is still an ideal choice for creating powerful data visualizations and spreadsheets.

You can also connect SQL with Excel and can use it to manipulate and analyze data. A lot of Data Scientists use Excel for data cleaning as it provides an interactable GUI environment to pre-process information easily.


01
Highly popular for small scale data analysis
Used for spreadsheet calculations & visualizations
02
03
Provides easy connection with SQL
Excel ToolPack used for complex data analysis
04
EXCEL Features

With the release of ToolPak for Microsoft Excel, it is now much easier to compute complex analyzations. However, it still pales in comparison with much more advanced Data Science tools like SAS. Overall, on a small and non-enterprise level, Excel is an ideal tool for data analysis.

7. ggplot2

ggplot2 is an advanced data visualization package for the **R programming language**. The developers created this tool to replace the native graphics package of R and it uses powerful commands to create illustrious visualizations.

It is the most widely used library that Data Scientists use for creating visualizations from analyzed data.Ggplot2 is part of tidyverse, a package in R that is designed for Data Science.

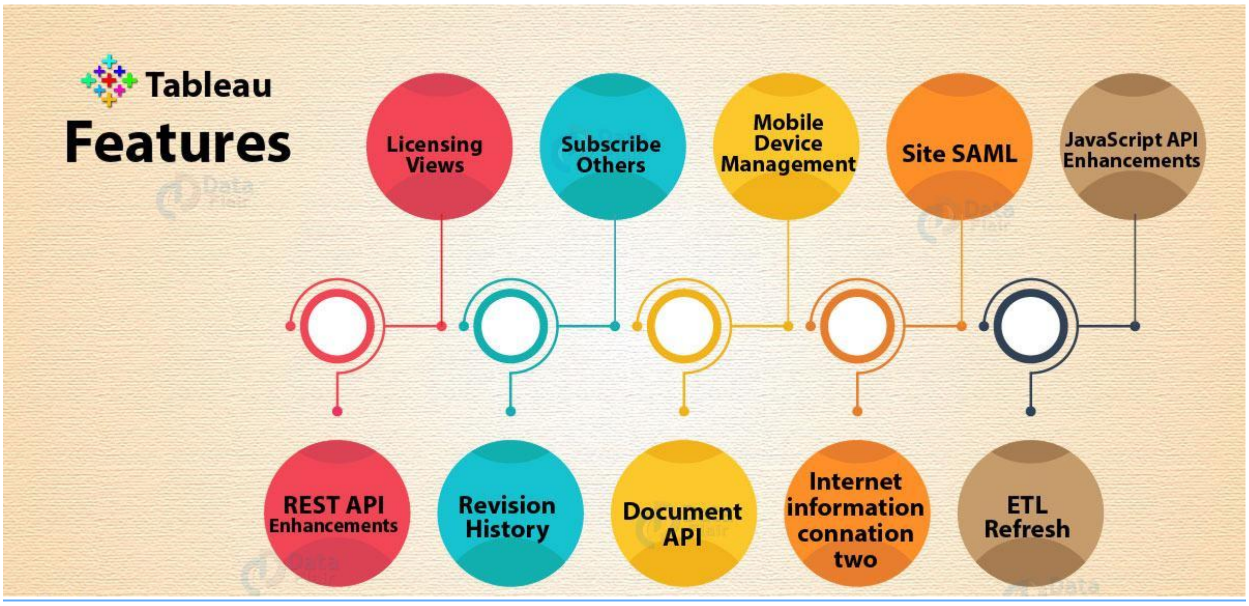
One way in which ggplot2 is much better than the rest of the data visualizations is aesthetics. With ggplot2, Data Scientists can create customized visualizations in order to engage in enhanced storytelling.

Using ggplot2, you can annotate your data in visualizations, add text labels to data points and boost intractability of your graphs. You can also create various styles of maps such as choropleths, cartograms, hexbins, etc. It is the most used data science tool.

8. Tableau

Tableau **is a Data Visualization software** that is packed with powerful graphics to make interactive visualizations. It is focused on industries working in the field of business intelligence.

The most important aspect of Tableau is its ability to interface with databases, spreadsheets, OLAP (Online Analytical Processing) cubes, etc. Along with these features, Tableau has the ability to visualize geographical data and for plotting longitudes and latitudes in maps.



Along with visualizations, you can also use its analytics tool to analyze data. Tableau comes with an active community and you can share your findings on the online platform. While Tableau is enterprise software, it comes with a free version called Tableau Public.

Users may construct dashboards and visualisations that continually update in real-time and display live data thanks to Tableau's real-time data connectivity features.

Tableau is a useful tool in a variety of businesses and areas due to its adaptability and simplicity of usage. It is a popular option for data-driven decision-making and storytelling because of its capacity to translate complicated data into useful insights through interactive visualisations.

9. Jupyter

Project **Jupyter** is an open-source tool based on IPython for helping developers in making open-source software and experiences interactive computing. Jupyter supports multiple languages like Julia, **Python**, and R.

It is a web-application tool used for writing live code, visualizations, and presentations. Jupyter is a widely popular tool that is designed to address the requirements of Data Science.

It is an interactable environment through which Data Scientists can perform all of their responsibilities. It is also a powerful tool for storytelling as various presentation features are present in it.

Using Jupyter Notebooks, one can perform data cleaning, statistical computation, visualization and create predictive **machine learning models**. It is 100% open-source and is, therefore, free of cost.

There is an online Jupyter environment called Collaboratory which runs on the cloud and stores the data in Google Drive.

10. Matplotlib

Matplotlib is a plotting and visualization library developed for Python. It is the most popular tool for generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc.

Matplotlib has several essential modules. One of the most widely used modules is pyplot. It offers a MATLAB like an interface. Pyplot is also an open-source alternative to MATLAB's graphic modules.

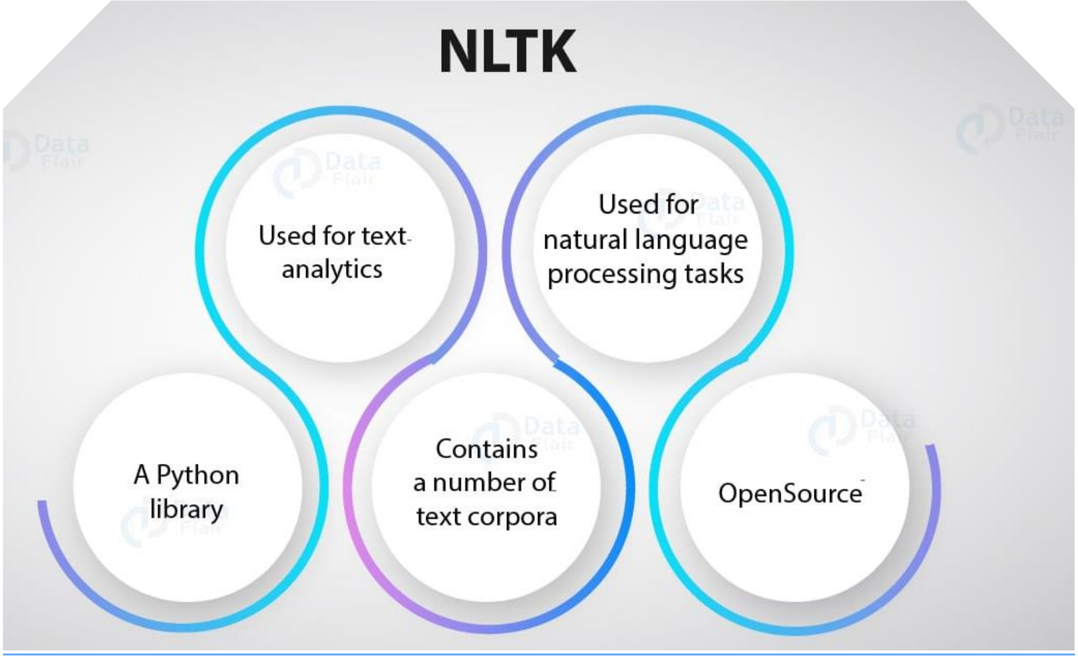
Matplotlib is a preferred tool for data visualizations and is used by Data Scientists over other contemporary tools.

As a matter of fact, NASA used Matplotlib for illustrating data visualizations during the landing of Phoenix Spacecraft. It is also an ideal tool for beginners in learning data visualization with Python.

11. NLTK

Natural Language Processing has emerged as the most popular field in Data Science. It deals with the development of statistical models that help computers understand human language.

These statistical models are part of Machine Learning and through several of its algorithms, are able to assist computers in understanding natural language. Python language comes with a collection of libraries called **Natural Language Toolkit (NLTK)** developed for this particular purpose only.


NLTK
Used for text analytics
Used for natural language processing tasks
A Python library
Contains a number of text corpora
OpenSource

NLTK is widely used for various language processing techniques like tokenization, stemming, tagging, parsing and machine learning. It consists of over 100 corpora which are a collection of data for building machine learning models.

It has a variety of applications such as Parts of Speech Tagging, Word Segmentation, Machine Translation, Text to Speech Speech Recognition, etc.

12. Scikit-learn

Scikit-learn is a library-based in Python that is used for implementing Machine Learning Algorithms. It is simple and easy to implement a tool that is widely used for analysis and data science.

It supports a variety of features in Machine Learning such as data preprocessing, classification, regression, clustering, dimensionality reduction, etc

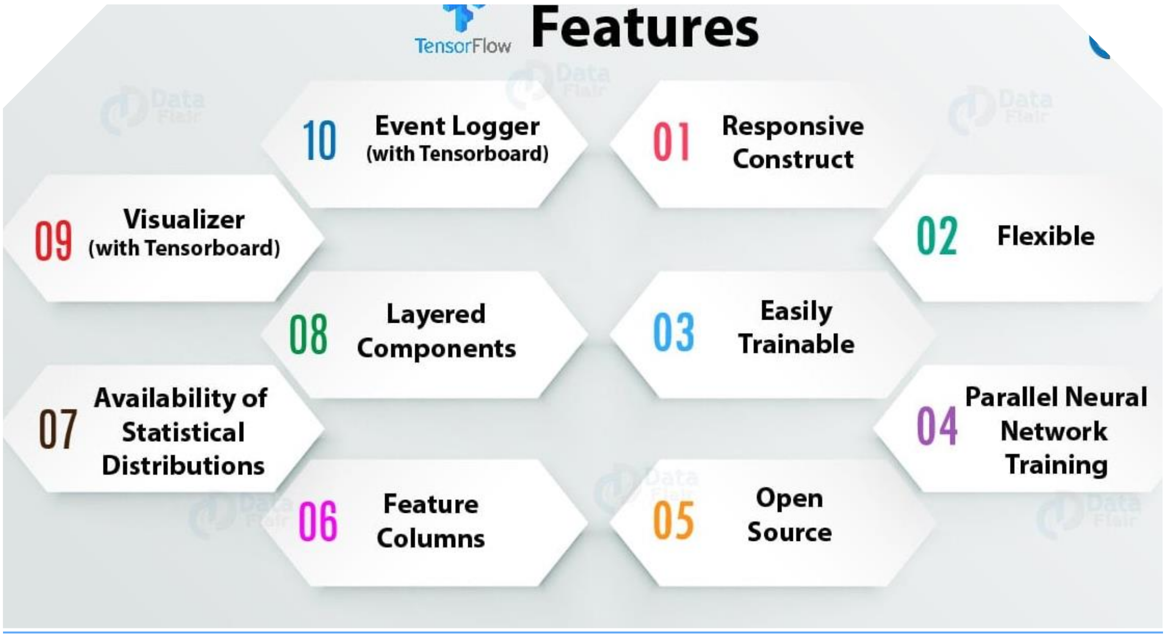
Scikit-learn makes it easy to use complex machine learning algorithms. It is therefore in situations that require rapid prototyping and is also an ideal platform to perform research requiring basic Machine Learning. It makes use of several underlying **libraries of Python** such as SciPy, Numpy, Matplotlib, etc.

13. TensorFlow

TensorFlow **has become a standard tool** for Machine Learning. It is widely used for advanced machine learning algorithms like Deep Learning. Developers named TensorFlow after Tensors which are multidimensional arrays.

It is an open-source and ever-evolving toolkit which is known for its performance and high computational abilities. TensorFlow can run on both CPUs and GPUs and has recently emerged on more powerful TPU platforms.

This gives it an unprecedented edge in terms of the processing power of advanced machine learning algorithms.


TensorFlow
Features
01 Responsive Construct
02 Flexible
03 Easily Trainable
04 Parallel Neural Network Training
05 Open Source
06 Feature Columns
07 Availability of Statistical Distributions
08 Layered Components
09 Visualizer (with Tensorboard)
10 Event Logger (with Tensorboard)
Data Flair

Due to its high processing ability, **Tensorflow has a variety of applications** such as speech recognition, image classification, drug discovery, image and language generation, etc. For Data Scientists specializing in Machine Learning, Tensorflow is a must-know tool.

14. Weka

Weka or **W**aikato **E**nvironment for **K**nowledge **A**nalysis is a machine learning software written in Java. It is a collection of various Machine Learning algorithms for data mining. Weka consists of various machine learning tools like classification, clustering, regression, visualization and data preparation.

It is an open-source GUI software that allows easier implementation of machine learning algorithms through an interactable platform.

You can understand the functioning of Machine Learning on the data without having to write a line of code. It is ideal for Data Scientists who are beginners in Machine Learning.

Summary

We saw how data science requires a vast array of tools. The tools for data science are for analyzing data, creating aesthetic and interactive visualizations and creating powerful predictive models using machine learning algorithms.

Most of the data science tools deliver complex data science operations in one place. This makes it easier for the user to implement functionalities of data science without having to write their code from scratch. Also, there are several other tools that cater to the application domains of data science.