Tools for Data Science

Glossary

Term	Definition
Apache MLlib	Language that makes machine learning scalable
Apache Spark	A general-purpose cluster-computing framework allowing you to process data using compute clusters
API	Application programming interface allows communication between two pieces of software
Caffe	A deep learning algorithm repository built with C++ with Python and Matlab bindings
CDLA	Community Data License Agreement
Classification models	Are used to predict whether some information or data belongs to a category (or “class”)
CLI	Command line interface
C++	A general-purpose programming language. It is an extension of the C programming language or C with Classes
Data set	A structured collection of data
Deeplearning4	Language for deep learning
Deep learning	A specialized type of machine learning. It refers to a general set of models and techniques that loosely emulate the way the human brain solves a wide range of problems
ELT	Extract, Load, Transform
ETL	Extract, Transform, and Load
FSF	Free Software Foundation
ggplot2	A popular library for data visualization in R
GPU	Graphics processing units
Git	De facto standard for code asset management, also known as version management or version control. Around Git emerged several services, GitHub, and GitLab
Hadoop	Application of Java which manages data processing and storage for big data applications running in clustered systems
Java	Object-oriented programming language
Java-ML	Language for machine learning
JVM	Java Virtual Machine
JavaScript	A general-purpose language that extended beyond the

	browser with the creation of Node.js and other server-side approaches
Julia	A language for high-performance numerical analysis and computational science
Jupyter Notebook	A browser-based application that allows you to create and share documents containing code, equations, visualizations, narrative text links, and more
Jupyter Lab	A browser-based application that allows you to access multiple Jupyter Notebook files, other code, and data files
Kernel	An execution environment for the different programming languages
Lattice	It is a high-level data visualization library that can handle graphics without customizations
Library	A collection of functions and methods that allow you to perform many actions without writing the code
Leaflet	Used for creating interactive plots
ML	Machine learning uses algorithms – also known as “models” - to identify patterns in the data
Matplotlib	package for data visualization
Model training	The process by which the model learns patterns from data
MNIST	Modified National Institute of Standards and Technology
MongoDB	A NoSQL database for big data management that was built with C++
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NumPy	Libraries are based on arrays and matrices, allowing you to apply mathematical functions to the arrays
OSI	Open-Source Initiative
PaaS	Platform as a service
Pandas	A library that offers data structures and tools for effective data cleaning, manipulation, and analysis
Plotly	Used for web-based data visualizations that can be displayed or saved as individual HTML files
PMML	Predictive Model Markup Language
Python	A high-level, general-purpose programming language. It has a large, standard library that provides tools suited to many different tasks, including Databases,

	Automation, Web scraping, Text processing, Image processing, Machine learning, and Data analytics
R	A statistical computing language
Regression models	Are used to predict a numeric (or “real”) value
Reinforcement Learning	Loosely based on the way human beings and other organisms learn.
REST	RE stands for Representational; the S stands for State, and the T stands for Transfer
RStudio	Unifies programming, execution, debugging, remote data access, data exploration, and visualization into one tool
SaaS	Software as a service
Scala	Is a combination of scalable and language. A general-purpose programming language that provides support for functional programming and is a strong static type system
Spyder	Integrates code, documentation, and visualizations, among others, into a single canvas
SQL	Structured Query Language that is non-procedural, used for querying and managing data
Supervised Learning	A learning in which a human provides input data and correct outputs
TensorFlow	Deep Learning library for dataflow that was built with C++
Unsupervised Learning	The data is not labeled by a human. Examples are Clustering models used to divide each record of a dataset into one of a similar group
Watson Studio	A fully integrated development environment for data scientists
Weka	Language for data mining