# Top Machine learning languages

* Python
* Java
* R
* C++
* C
* JavaScript
* Scala
* Julia

# R and Python

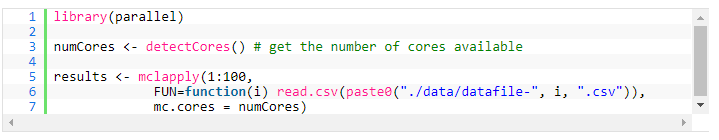
R is a workhorse for statistical analysis and machine learning. R is a software platform use for data understanding and exploration using statistical methods and graphs. It has many machine learning algorithms. The [most popular platform for machine learning competitors such as Kaggle](http://blog.kaggle.com/2011/11/27/kagglers-favorite-tools/). Python is a popular scientific and machine learning language.

R and Python provide tools that allow data scientists to leverage large-scale architectures to collect, write, wrangle, and manipulate data, as well as train and validate models on multicore architectures. This paper will cover the parallel package, data.table, caret, and multidplyr in R and the paratext, joblib, and scikit-learn package in Python.

## Parallel package

Parallel package provides drop-in parallel replacements for most of the functionality of apply, with integrated handling of random-number generation. This package is principally concerned with coarse-grained parallelization. The crucial point is that these chunks of computation are unrelated and do not need to communicate in any way.

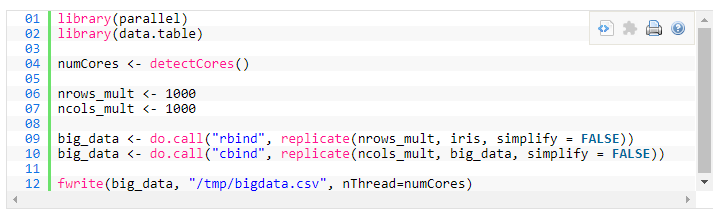
Read a directory of 100 csv files from a folder in parallel leveraging multicore architectures to parse and consume data on disk more quickly.



## Data.table

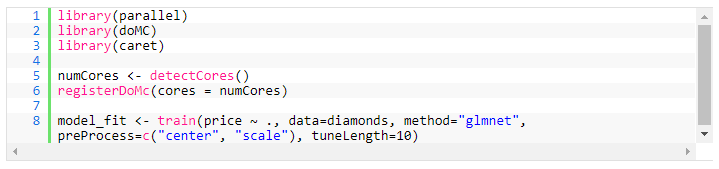
Data.table is a powerful package. It is a high-performance implementation of R’s data frame. It provides a highly optimized tabular data structure for most common analytical operations.

The data.table fwrite package is used to write a large CSV leveraging multicore architectures.



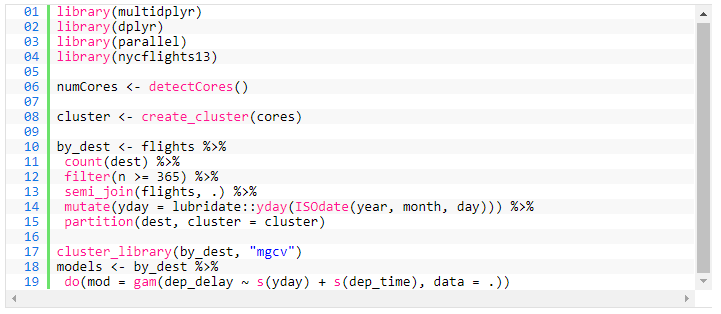
## Caret

The caret package is a set of functions that streamline the process for creating predictive models. This package contains tools for data splitting, preprocessing, feature selection, model tuning using resampling, variable importance estimation, and other functionality. The caret package leverages multicore functionality by taking advantage of the fact that many of the operations in model training are parallel.



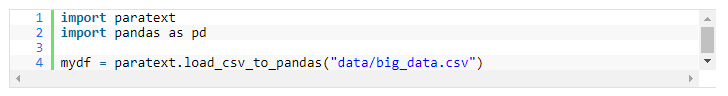
## Multidplyr

Multidplyr partitions a data frame across multiple cores. Multidplyr is being used to train many GAM models on a dataset.



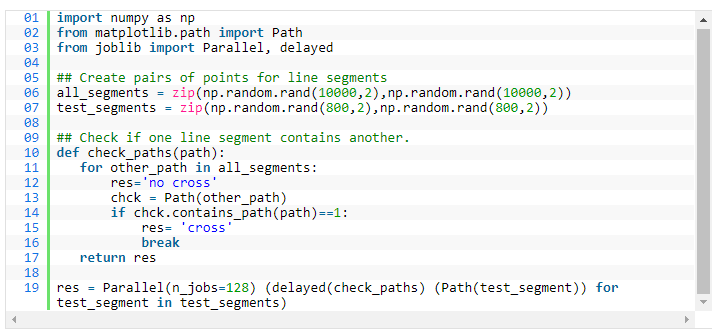
## Paratext

The paratext library delivers mutlicore processing to CSV reading and parsing. ParaText is a C++ library used to read text files in parallel on multi-core machines.



## Joblib

Joblib is a set of tools for lightweight pipelining in Python. Joblib is optimized to be fast and robust on large data. It has been optimizations for numpy arrays. Joblib offers distributed and multicore processing tasks.



## Scikit-learn

Scikit-learn supports various classification, regression, and clustering algorithms including support for vector machines, random forests, gradient boosting, and k-mean. It is designed to work with numerical and scientific libraries NumPy and SciPy. It leverage large multicore servers by the simple usage of the n\_jobs parameter.

The RandomForestClassifier is used to predict the species in the iris dataset leveraging multiple cores in parallel.

