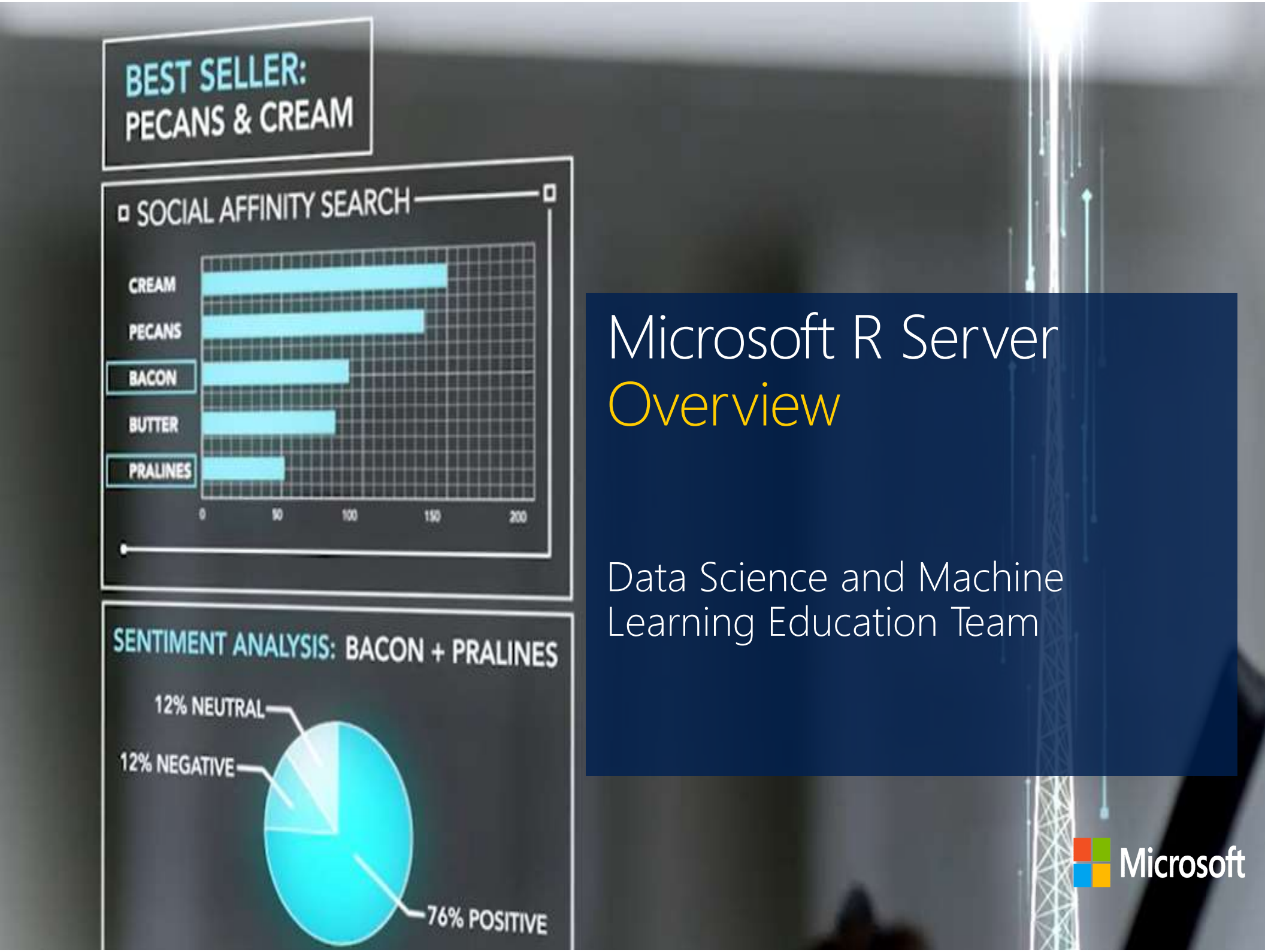


Microsoft R Server
Overview

Data Science and Machine Learning Education Team

Microsoft

What is R?

Open-source programming language for statistical computing

- Free (cran.r-project.org)
- Highly extensible
- Focused on statistics and machine learning
- Transparent and reproducible

- Single-threaded
- Data stored in memory

A brief history of R

1993
Research project in Auckland, NZ

1995
Open source

Photo credit: Robert Gentleman

1997
R-core

2000
R-1.0.0

2003
R Foundation

2004
First UseR!

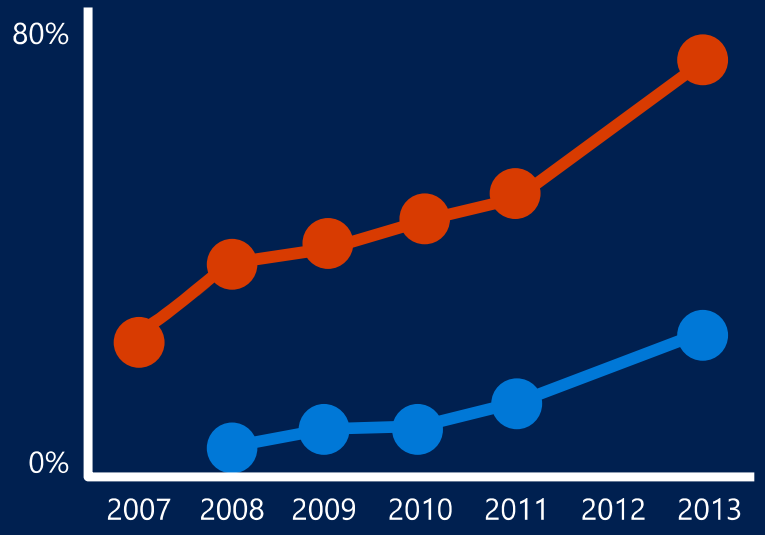
2009
New York Times

2015
R-3.2.0
R Consortium

R's popularity is growing rapidly

R Usage Growth

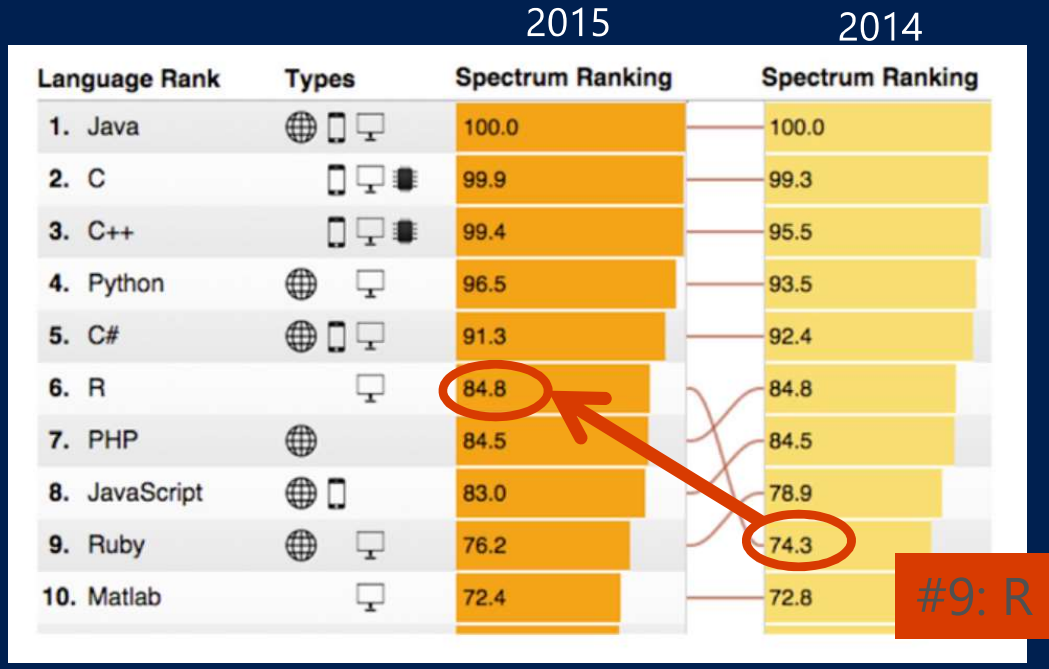
Rexer Data Miner Survey, 2007-2013



Rexer Data Miner Survey

Language Popularity

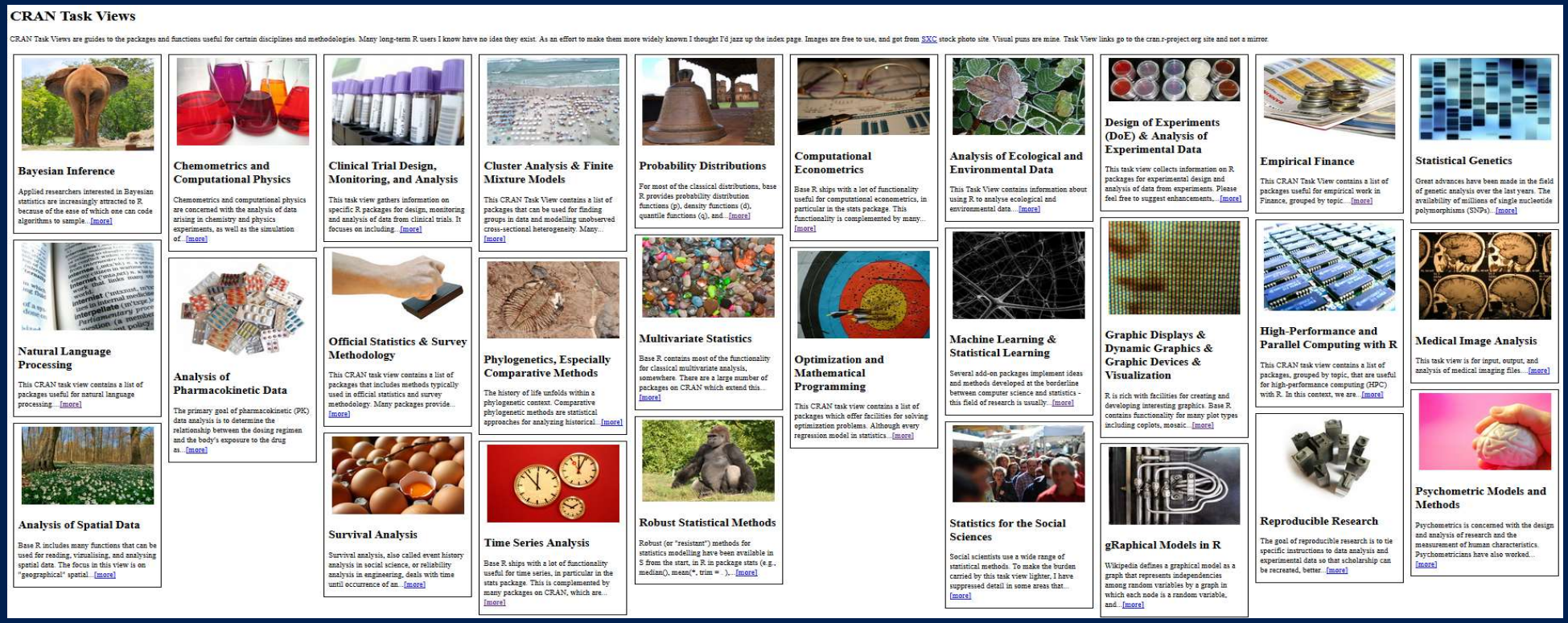
IEEE Spectrum Top Programming Languages



IEEE Spectrum July 2015

CRAN

The Comprehensive R Archive Network



CRAN Task Views

CRAN Task Views are guides to the packages and functions useful for certain disciplines and methodologies. Many long-term R users I know have no idea they exist. As an effort to make them more widely known I thought I'd jazz up the index page. Images are free to use, and got from SXC stock photo site. Visual puns are mine. Task View links go to the cran.r-project.org site and not a mirror.

Bayesian Inference

Applied researchers interested in Bayesian statistics are increasingly attracted to R because of the ease of which one can code algorithms to sample.. [more]

Chemometrics and Computational Physics

Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of.. [more]

Clinical Trial Design, Monitoring, and Analysis

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including.. [more]

Cluster Analysis & Finite Mixture Models

This CRAN Task View contains a list of packages that can be used for finding groups in data and modelling unobserved cross-sectional heterogeneity. Many... [more]

Probability Distributions

For most of the classical distributions, base R provides probability distribution functions (p), density functions (d), quantile functions (q), and... [more]

Computational Econometrics

Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many... [more]

Analysis of Ecological and Environmental Data

This Task View contains information about using R to analyse ecological and environmental data... [more]

Design of Experiments (DoE) & Analysis of Experimental Data

This task view collects information on R packages for experimental design and analysis of data from experiments. Please feel free to suggest enhancements,.. [more]

Empirical Finance

This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic... [more]

Statistical Genetics

Great advances have been made in the field of genetic analysis over the last years. The availability of millions of single nucleotide polymorphisms (SNPs).. [more]

Natural Language Processing

This CRAN task view contains a list of packages useful for natural language processing... [more]

Analysis of Pharmacokinetic Data

The primary goal of pharmacokinetic (PK) data analysis is to determine the relationship between the dosing regimen and the body's exposure to the drug as... [more]

Official Statistics & Survey Methodology

This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide... [more]

Phylogenetics, Especially Comparative Methods

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical... [more]

Multivariate Statistics

Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this... [more]

Optimization and Mathematical Programming

This CRAN task view contains a list of packages which offer facilities for solving optimization problems. Although every regression model in statistics.. [more]

Machine Learning & Statistical Learning

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually.. [more]

Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic.. [more]

High-Performance and Parallel Computing with R

This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are.. [more]

Medical Image Analysis

This task view is for input, output, and analysis of medical imaging files... [more]

Analysis of Spatial Data

Base R includes many functions that can be used for reading, visualising, and analysing spatial data. The focus in this view is on "geographical" spatial.. [more]

Survival Analysis

Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an.. [more]

Time Series Analysis

Base R ships with a lot of functionality useful for time series, in particular in the stats package. This is complemented by many packages on CRAN, which are... [more]

Robust Statistical Methods

Robust (or "resistant") methods for statistics modelling have been available in S from the start, in R in package stats (e.g., median(), mean(*, trim = .)... [more]

Statistics for the Social Sciences

Social scientists use a wide range of statistical methods. To make the burden carried by this task view lighter, I have suppressed detail in some areas that... [more]

gRaphical Models in R

Wikipedia defines a graphical model as a graph that represents independencies among random variables by a graph in which each node is a random variable, and... [more]

Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better.. [more]

Psychometric Models and Methods

Psychometrics is concerned with the design and analysis of research and the measurement of human characteristics. Psychometricians have also worked... [more]

In addition to CRAN, Bioconductor, GitHub, others distribute R packages

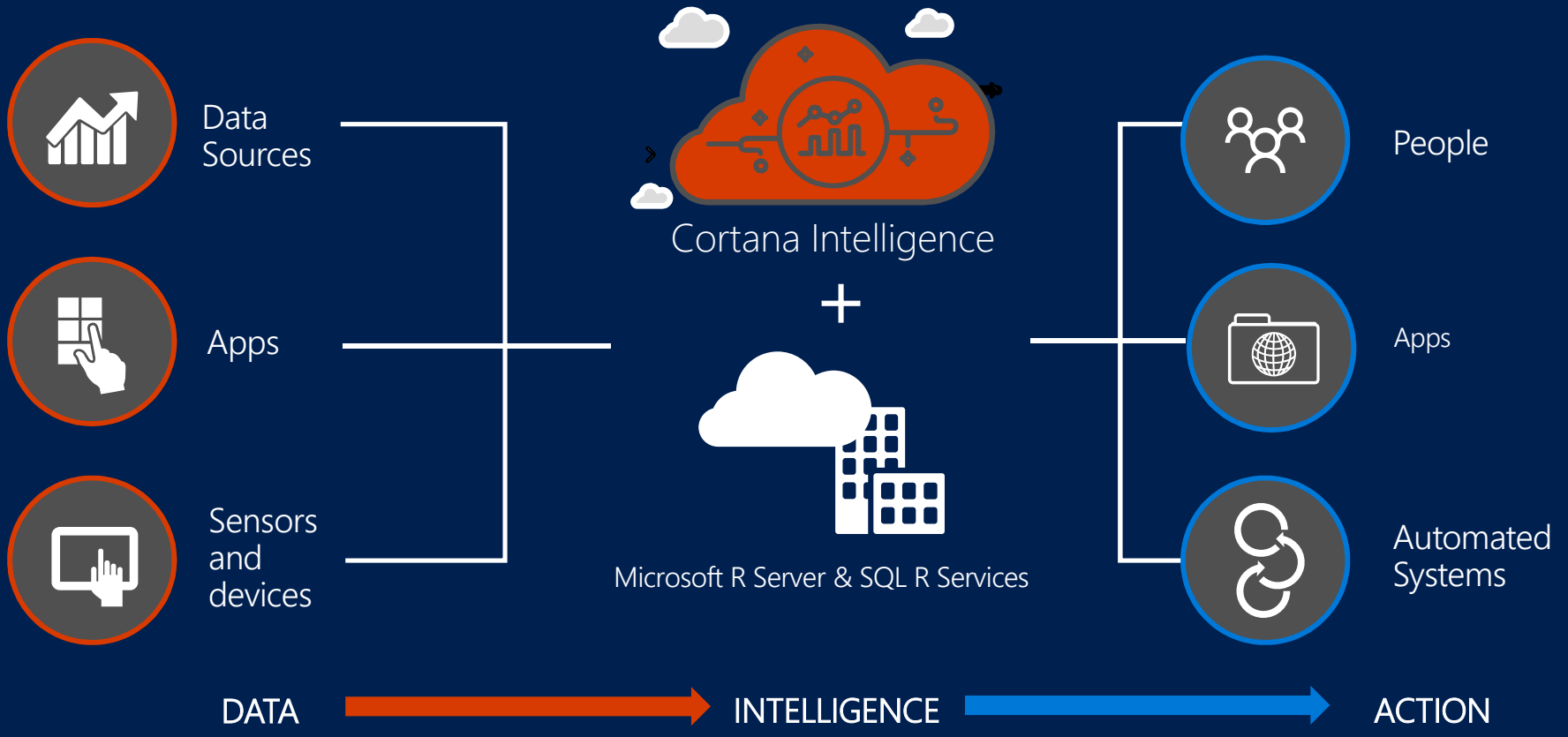
Microsoft R Server

MRS extends open-source R to allow:

- Multi-threading
  - Matrix operations, linear algebra, and many other math operations run on all available cores
- Parallel processing
  - ScaleR functions utilize all available resources, local or distributed
- On-disk data storage
  - RAM limitations lifted – Break Through Your Memory Barrier!

Microsoft R Server family

From Data To Action On Premises



What is

Language Platform

- A statistics programming language
- A data visualization tool
- Open source

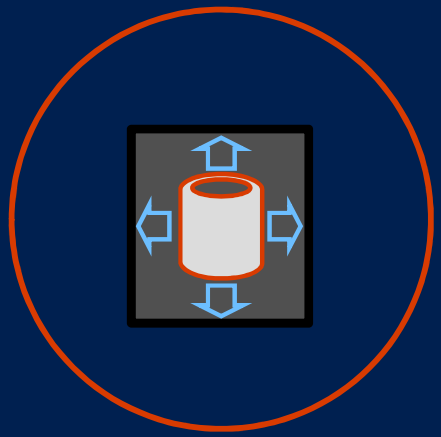
Community

- 2.5+M users
- Taught in most universities
- New and recent grad's use it
- Thriving user groups worldwide

Ecosystem

- 10,000+ free algorithms in CRAN
- Scalable to big data
- Rich application & platform integration

Challenges posed by open source R



Limited Data Scale

Inadequate Modeling Performance

Lack of Commercial Support

Complex Deployment Processes

R from Microsoft brings

Peace of mind

Efficiency

Speed and scalability

Flexibility and agility

Microsoft R portfolio

Microsoft R portfolio

Community

Microsoft
R Open

Commercial

SQL Server
R Services

Windows

Microsoft
R Server

Linux

Hadoop | Teradata

Spark

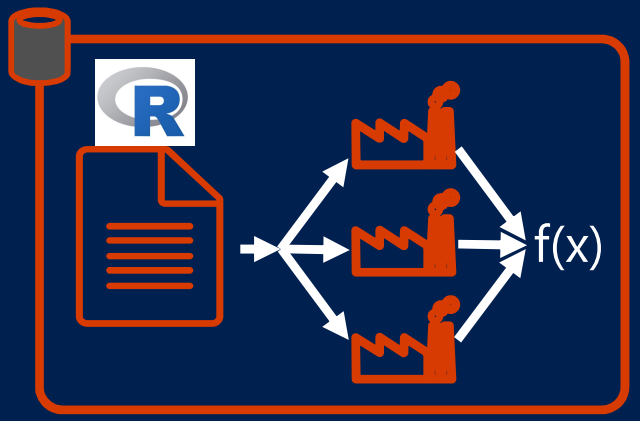
Microsoft R Scales to Big Data for Enterprises

Escapes R's traditional memory limits

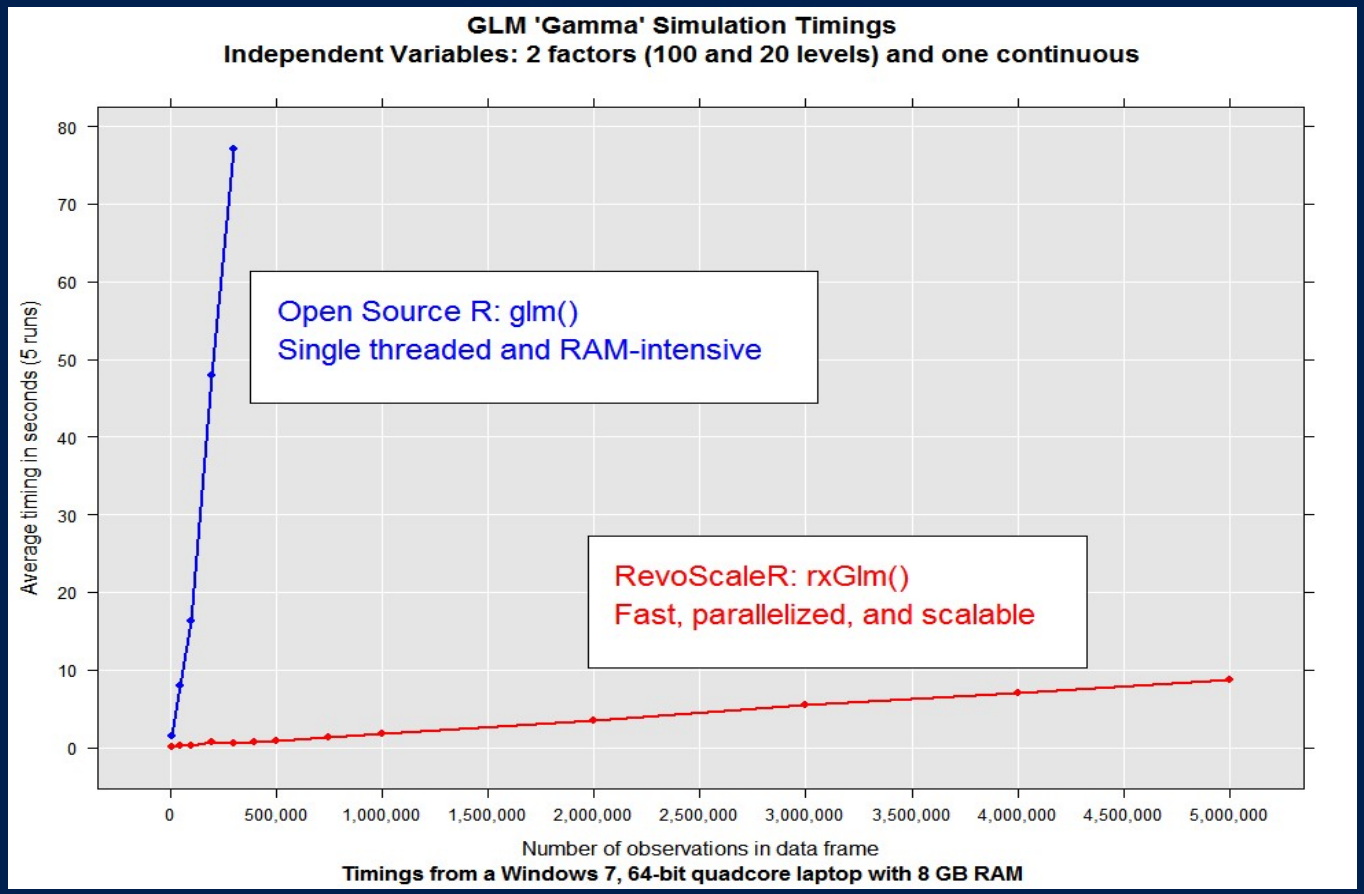
Scales predictive modeling using parallelization

Distributes computation cores & nodes

Minimizes data movement using in-database, in-MapReduce and in-Apache Spark execution



Scalable algorithms

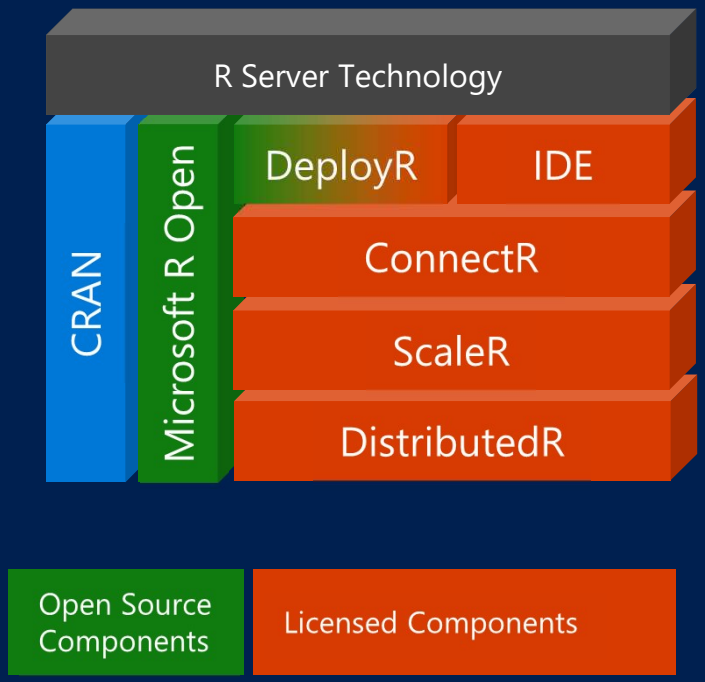


Introducing Microsoft R Server

Linux, Windows, Hadoop & Teradata

High-performance, Scalable R

100% open source R
CRAN, Bioconductor, MRAN, GitHub compatibility
Big-data connectivity
Scalable analytics
Multi-platform
In-database, in-cluster scalability
Choice of IDE



Introducing SQL Server 2016 R services

Enterprise speed and scale

Near-DB analytics

Parallel threading and processing

Reuse SQL skills for data engineering

In-database deployment

Memory and disk scalability

No R memory limits

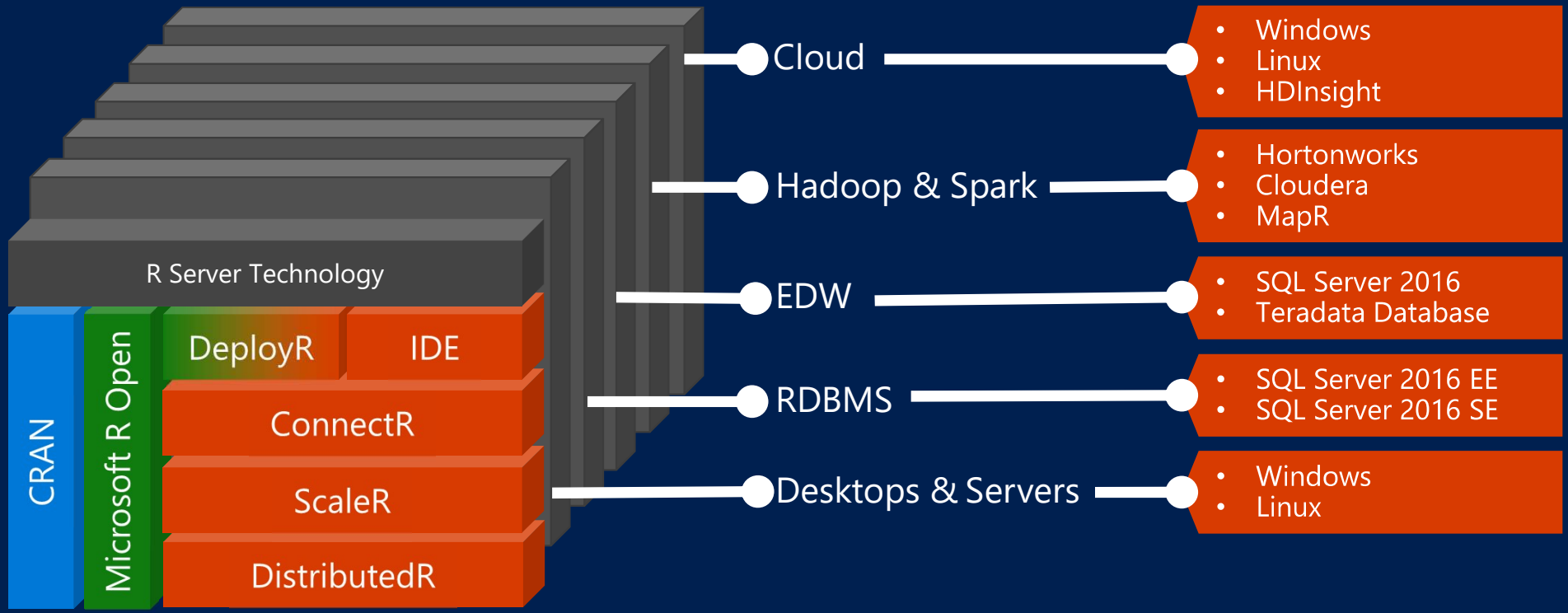
Write once, deploy anywhere

Included in SQL Server 2016

Reuse and optimize existing R code

Eliminate data movement

Portability & investment assurance



Write Once – Deploy Anywhere

MRS in Different Contexts

- On a workstation, that means:
  - All available cores will be used for math operations and parallel processes
  - Hard drive capacity sets limit for data size, not RAM

- On a cluster:
  - Parallel utilization of all available nodes
  - Distributed file systems like HDFS greatly expand possible data sizes

MRS in Different Contexts

Code written on a workstation will run on a cluster by tweaking a single function call:

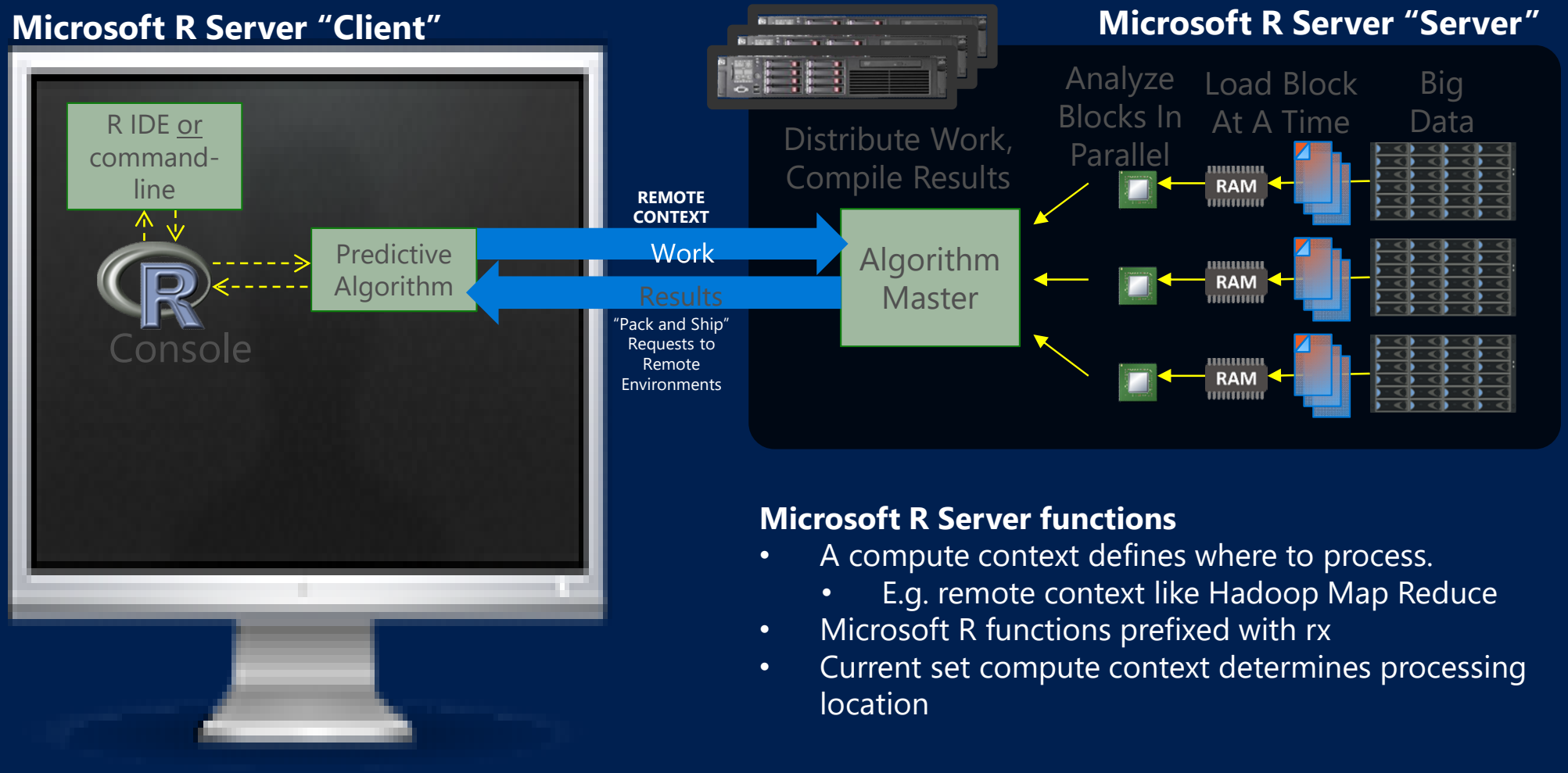
```
# Use your local computer:
rxSetComputeContext( LocalParallel() )

# Switch to your cluster:
rxSetComputeContext( RxSpark(...) )
```

How MRS Works

Parallel External Memory Algorithms (PEMAs)

1. A chunk/subset of data is extracted from the main dataset
2. An intermediate result is calculated from that chunk of data
3. The intermediate results are combined into a final dataset

Distributed R - How Does Remote Compute Context ?



Microsoft R Server functions

- A compute context defines where to process.
  - E.g. remote context like Hadoop Map Reduce
- Microsoft R functions prefixed with rx
- Current set compute context determines processing location

Available Algorithms

- Linear regression (rxLinMod)
- Generalized linear models (rxLogit, rxGLM)
- Decision trees (rxDTree)
- Gradient boosted decision trees (rxBTree)
- Decision forests (rxDForest)
- K-means (rxKmeans)
- Naïve Bayes (rxNaiveBayes)

Note: models available in open-source R packages won't be made parallel automatically

Parallelized, Remote Execution Algorithms

Data Step

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling

- Subsample (observations & variables)
- Random Sampling

Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Quantiles (approx.)
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

Variable Selection

- Stepwise Regression

Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

Cluster Analysis

- K-Means

Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes

Combination

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms

Best Uses of MRS

- Working with data too big to fit into memory
- Building models that take too long to run
- Working with clusters and distributed file systems

MRS's Native Data Format: The XDF File

- Chunk-oriented
  - Easy to distribute to nodes
  - Fast to append
- Column-oriented
  - Fast retrieval of variables
- Pre-computed metadata

Moving Data to Disk

- Text files, binary files, databases…
  - MRS can work directly with many of these formats
  - Advantages and disadvantages to each

- The eXternal Data Frame file (XDF)

Importing to XDF

- `rxImport`
  - `inData`
  - `outFile`
  - `varsToKeep, varsToDrop`
  - `numRows`
  - `rowSelection`
  - `overwrite`
  - `append`

Input and Output

- `inData`
  - CSV, SAS, SPSS, an ODBC connection...

- `outFile`
  - An XDF file; returns a data frame if left blank

Subset of Variables

- `varsToKeep`
- `varsToDrop`

Subset of Rows

- `numRows`
- `rowSelection`

Data Sources

- Data sources are wrappers that help MRS work with different kinds of data
- Often implicit, more powerful when explicit
  - Specify data types, a query to use over ODBC, rows per read, etc.

Data Sources

- Text files (delimited, fixed-width, etc)
- SAS, SPSS
- Teradata
- HDFS
- Databases via ODBC
- Runs in-database in SQL Server 2016

Importing from Databases

- Set up ODBC first
- Each data source (RxOdbcData) represents one query (not one database)
- SQL Server 2016 can run MRS internally; no ODBC required!!!

rxDataStep

- Subset rows with criteria (`rowSelection`)
- Select columns by name (`varsToKeep`, `varsToDrop`)
- Create and modify variables (`transforms`)
- Pull data into an in-memory `data.frame`

Subsetting Rows

- `rowSelection` takes a logical vector, just like `subset()`
- Chain multiple criteria together with `&` and `|`
- (Use `numRows = N` to get the first N rows of a dataset)

Selecting Columns

- `varsToKeep, varsToDrop`

- One quirk: can’t keep/drop when inData == outFile

`transforms`

- Create new variables
- Modify existing variables
- Change variable types
- Takes a list of named elements – each a new variable

"Complex" Transformations

- Simple transformations depend on a single row of data
- Complex transforms depend on multiple rows and/or objects
- In a distributed context, that means moving results between nodes

Managing Factors

- Factors count as "complex" because levels, level order, and level encodings can vary across chunks
- Use rxFactors to create and modify factors
- The F() shortcut

How Algorithms Work in Microsoft R Server: Chunk by Chunk

(aka Parallel External Memory Algorithms/PEMAs)

- Data just needs to fit on disk
- Chunks of data distributed to all available cores/nodes
- Intermediate results calculated in-memory for each chunk
- Final results assembled in-memory

PEMAs in Context

On a laptop:

- Chunks pulled from local disk
- All cores process chunks in parallel

Computing cluster

- Chunks partitioned across nodes
- All cores on nodes process local chunks in parallel

Analyzing Data with MRS

Pre-computed metadata

- rxGetInfo, rxGetVarInfo

Summary statistics

- rxSummary, rxQuantile
- rxCrossTabs, rxCube

Predictive modeling

- Regressions: rxLinMod, rxLogit, rxGLM
- Decision trees and forests: rxDTree, rxBTree, rxDForest
- K-means and Naive Bayes: rxKmeans, rxNaiveBayes

Metadata Retrieval

- rxGetInfo, rxGetVarInfo, rxGetVarNames
- All calculated on import, and retrieved from the XDF file header

Numeric Variables: rxSummary

- Standard summary stats: mean, standard deviation, minimum, maximum, number missing
- Works for groupwise summaries, too
- Formula interface

Using Formula Syntax in rxSummary

- One variable:

```
rxSummary( ~ arr_delay,
           data = flightsXdf)
```

- Two variables:

```
rxSummary( ~ arr_delay + dep_delay,
           data = myXdf)
```

- Groupwise:

```
rxSummary(arr_delay ~ dayOfWeek_F,
          data = myXdf)
```

Numeric Variables: rxQuantile

- Calculates quantiles
- … just one variable at a time

Categorical Variables

- rxCrossTabs for frequency tables
- rxCube for "long" tables
- Requires factor inputs
- Formula interface:

```
rxCrossTabs( ~ origin_F : dest_F,
             data = flightsXdf)
```

Modeling Workflow in MRS

- Load data (rxImport)
- Exploratory analysis (rxGetInfo, rxSummary, rxCube)
- Clean data (rxDataStep, rxFactors)
- Build a model – or several! (rxLinMod, rxGLM, etc)
- Evaluate and Predict (rxPredict)

Modeling Algorithms

- Linear regression (rxLinMod)
- Generalized linear models (rxLogit, rxGLM)
- Decision trees (rxDTree)
- Gradient boosted decision trees (rxBTree)
- Decision forests (rxDForest)
- K-means (rxKmeans)
- Naïve Bayes (rxNaiveBayes)

Using Formula Syntax in Models

- One predictor:
  ```
  rxLinMod(y ~ x, data = myXdf)
  ```
- Two predictors:
  ```
  rxLinMod(y ~ x + z, data = myXdf)
  ```
- Two predictors with interaction term:
  ```
  rxLinMod(y ~ x * z, data = myXdf)
  ```

Some Simple Examples

```
rxLinMod(mpg ~ hp + wt,
         data = mtcars)

rxLogit(delayed ~ dep_time + dayOfWeek_F,
        data = flightsXdf)

rxNaiveBayes(Species ~ Sepal.Length + Sepal.Width,
             data = iris)
```

Modeling Workflow in MRS

- Load data (rxImport)
- Exploratory analysis (rxGetInfo, rxSummary, rxCube)
- Clean data (rxDataStep, rxFactors)
- Build a model – or several! (rxLinMod, rxGLM, etc)
- Evaluate and Predict (rxPredict)

Model Evaluation and Prediction

- MRS models don't include fitted values or residuals by default
- Generate fitted values, residuals, and predictions with rxPredict:

```
# Fitted values: data used to fit model
rxPredict(modelObject = delayMod,
          data = flightsXdf,
          outData = flightsXdf)
```

Model Evaluation and Prediction

- Other options
  - Residuals: `computeResiduals = TRUE`
  - Standard Errors: `computeStdErrors = TRUE`
  - Confidence intervals: `interval = “confidence”`
  - Prediction intervals: `interval = “prediction”`
- For binary classifiers: rxRocCurve
  - Compares actual values to one or more predictions generated by rxPredict

Modeling Workflow in MRS

- Load data (rxImport)
- Exploratory analysis (rxGetInfo, rxSummary, rxCube)
- Clean data (rxDataStep, rxFactors)
- Build a model – or several! (rxLinMod, rxGLM, etc)
- Evaluate and Predict (rxPredict)