Microsoft

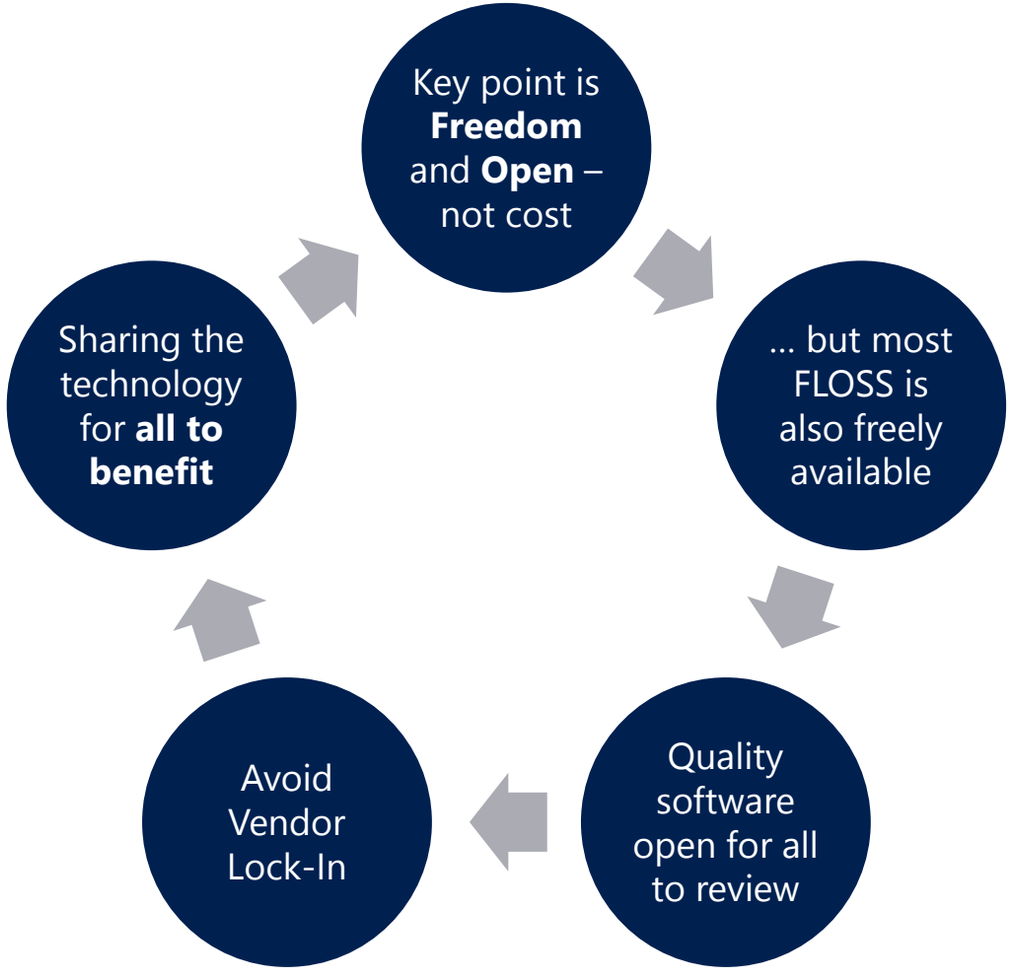
Microsoft R and Beyond

Mithun Prasad, PhD
miprasad@Microsoft.com

What you will be able to do after this training

- Code in R
- Develop an understanding of MRS capabilities
- How to manage data with dplyr
- Use RevoScaleR package to develop models and score
- Use SQL Server to develop in-database applications

Free and Open Source Software



FREE SOFTWARE FOUNDATION

open source initiative

FOSS

FLOSS

Libre Software

python

Spark

julia

hadoop

Data-Driven Documents

Microsoft

R Statistical Software
The Rich, The Powerful, The Ugly

A platform for best of breed open and closed software

Empowering today's developers to build intelligent applications

Microsoft

What is

Language Platform

- The most popular statistical 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

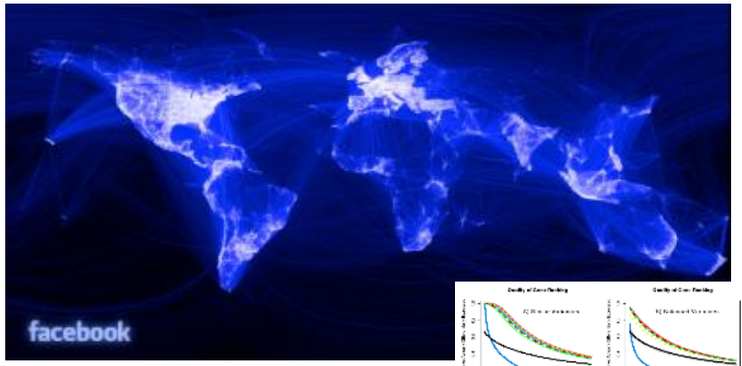
Ecosystem

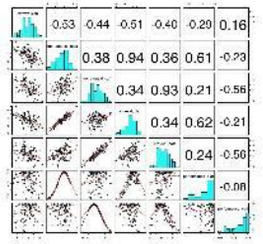
- 8000+ contributed packages
- Rich application & platform integration

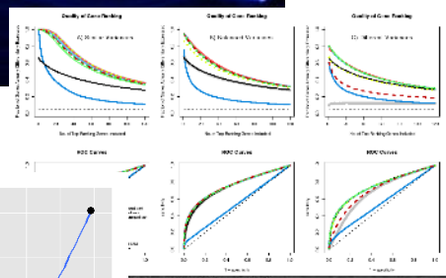
Microsoft

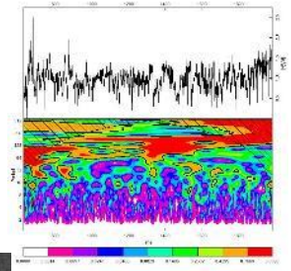
Open source R (CRAN R)

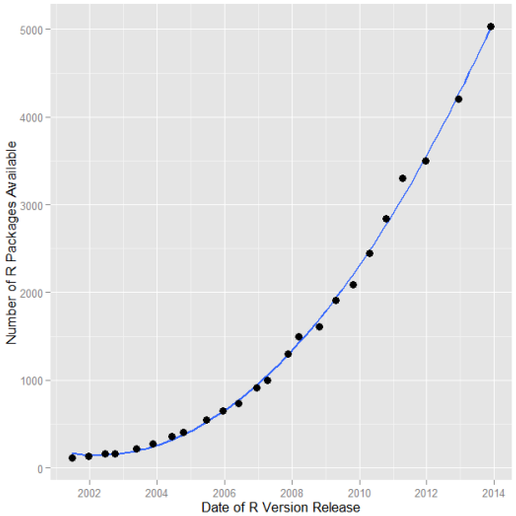
- Developed by Robert Gentleman & Ross Ihaka in 1993 from S+
- Version 1.0 open source in 2000
- 3.0+ Million Global Users
- 8000+ “Packages” – huge range of data manipulation, descriptive, predictive and visualisation capability
- R in universities provides new talent pools in large numbers
- Open Source means there is access to innovation at a pace that no commercial company can keep up with
- Very flexible and extensible programming language – much faster to programme than legacy alternatives

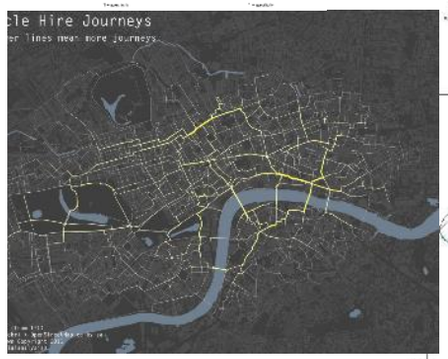


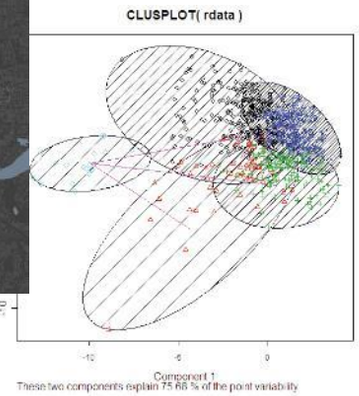






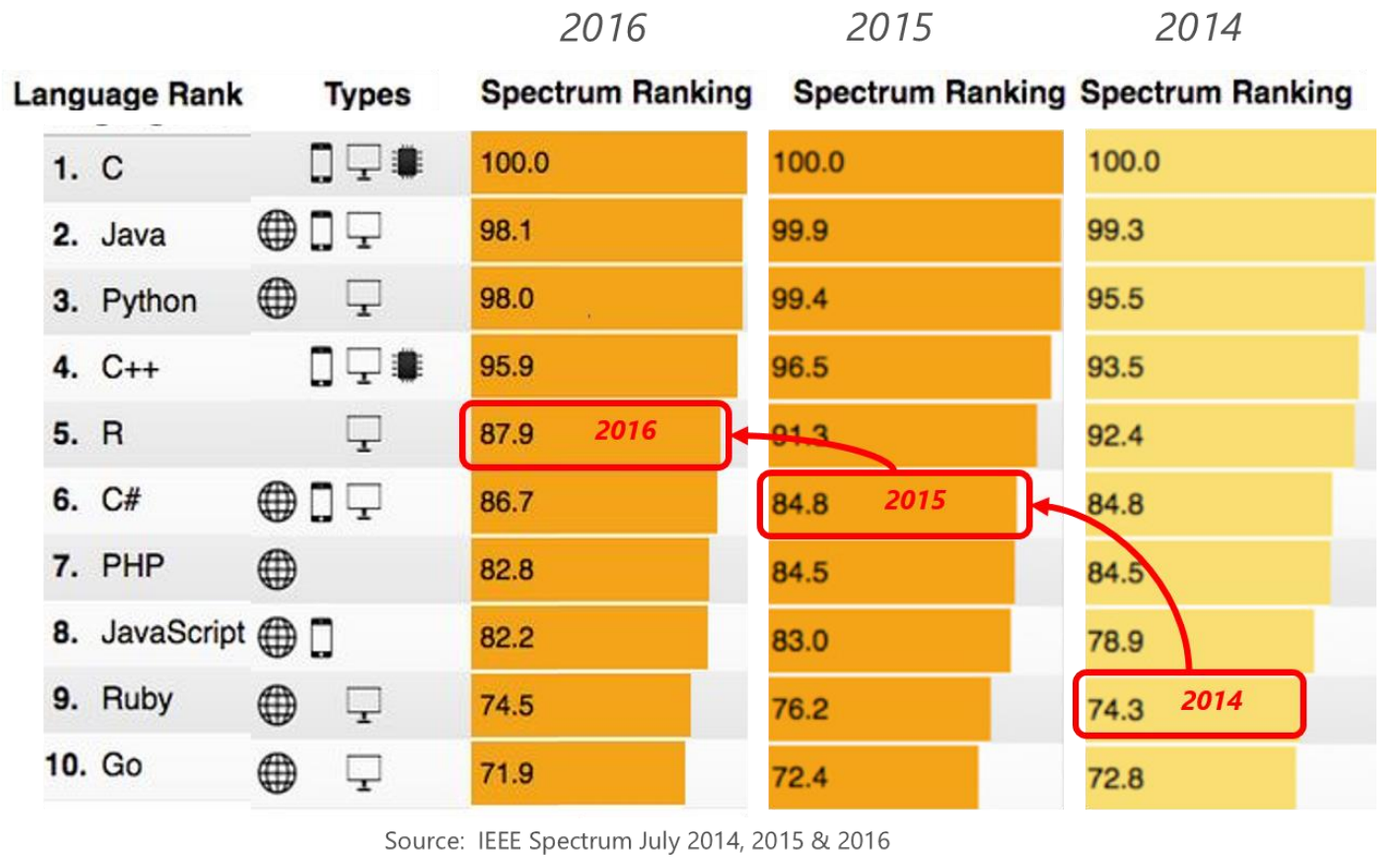






Microsoft

R Adoption is Growing Significantly
How to Support R as Enterprise Class



Source: IEEE Spectrum July 2014, 2015 & 2016

Data Flows and Open Source R

- In-Memory Operation
- Lack of Parallelism
- Data Movement & Duplication

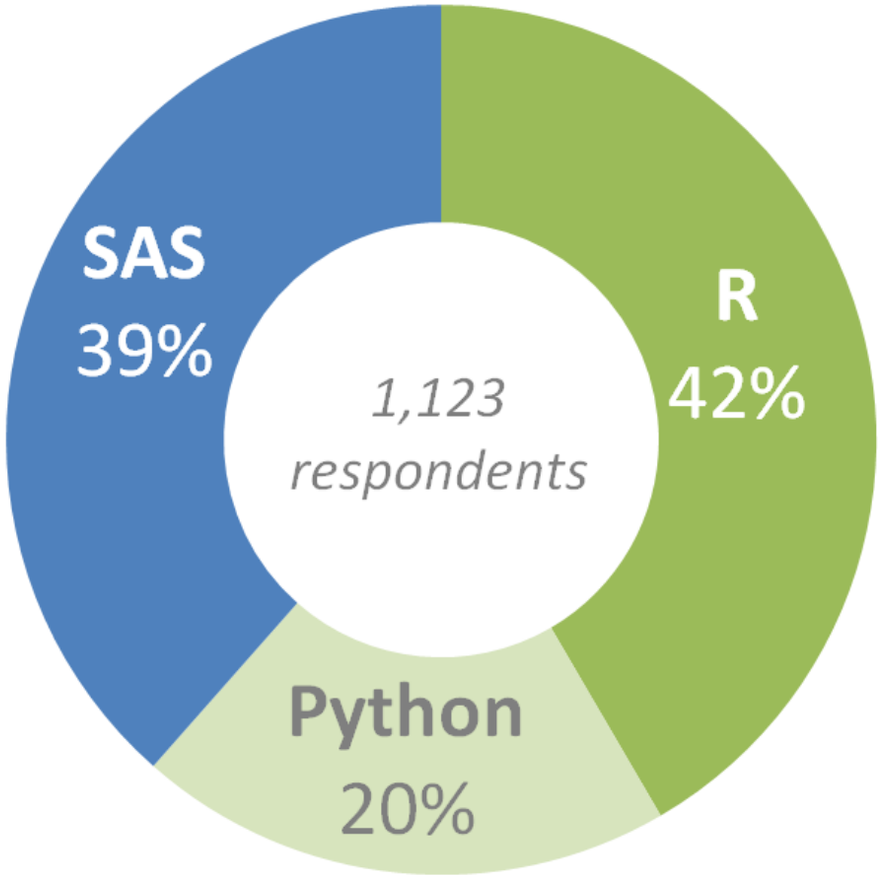
Enterprise Ready

- Gaps in Community Support
- Lack of Support Timeliness
- No SLAs or Support models

Microsoft

Preferred language by Analytics Pros

Which do you prefer to use: SAS, R, or Python?



Microsoft

Source: SAS, R or Python Survey 2016, by Burtch Works

Power of R: Language + Packages
CRAN: 8000+ Add-on packages for R

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]

Natural Language Processing

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

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]

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]

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]

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]

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]

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]

Phylogenetics, Especially Comparative Methods

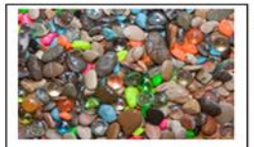
The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical... [more]

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]

Probability Distributions

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

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]

Computational Econometrics

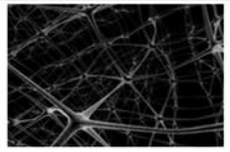
Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many... [more]

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]

Analysis of Ecological and Environmental Data

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

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]

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]

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]

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]

Empirical Finance

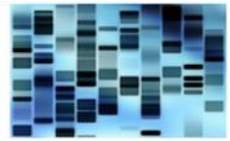
This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic... [more]

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]

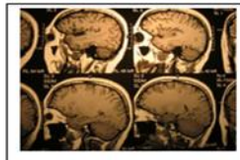
Reproducible Research

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



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]



Medical Image Analysis

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

Psychometric Models and Methods

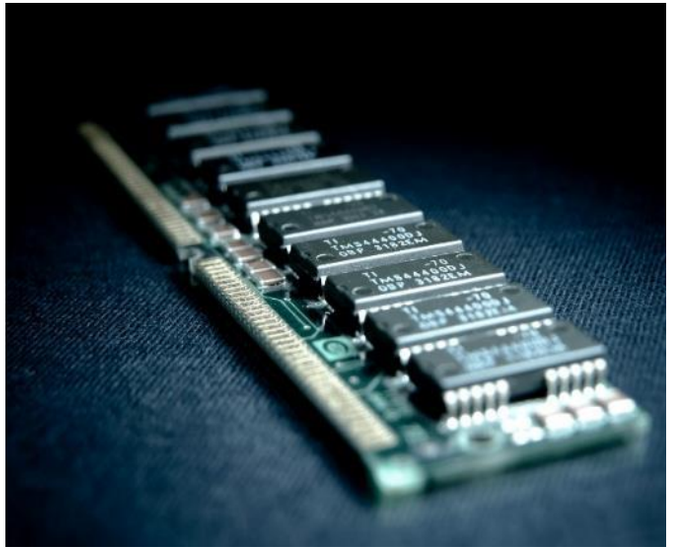
Psychometrics is concerned with the design and analysis of research and the measurement of human characteristics. Psychometricians have also worked... [more]

CRAN Task View by Barry Rowlingson: http://www.maths.lancs.ac.uk/~rowlings/R/TaskViews/

More packages on Github and BioConductor project

Microsoft

Enterprise use of open source R



R needs data in memory to start a computation*

R is single threaded*

R requires skilled resource to scale out computations across a cluster and needs re-coding for R map-reduce in Hadoop

Open source R is supported by the community

Microsoft R Server solves these problems!

*Open source R work-arounds are available for some of these problems but do not work in all cases

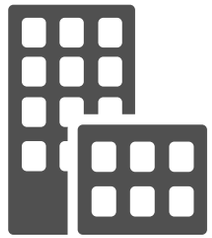
Microsoft

R Server: Scale-Out R, Enterprise Class!

- 100% compatible with open source R
  - Any code/package that works today with R will work in R Server.
- Ability to parallelize any R function
  - Ideal for parameter sweeps, simulation, scoring.
- Wide range of scalable and distributed "***rx***" pre-fixed functions in "RevoScaleR" package.
  - *Transformations*: rxDataStep()
  - *Statistics*: rxSummary(), rxQuantile(), rxChiSquaredTest(), rxCrossTabs()…
  - *Algorithms*: rxLinMod(), rxLogit(), rxKmeans(), rxBTrees(), rxDForest()…
  - *Parallelism*: rxSetComputeContext()

Microsoft

Microsoft R portfolio



Microsoft R portfolio

Community

Open Source R

R Open

R Client

Commercial

SQL Server R Services

R Server

Windows

Linux

Teradata

Hadoop

Spark

Microsoft

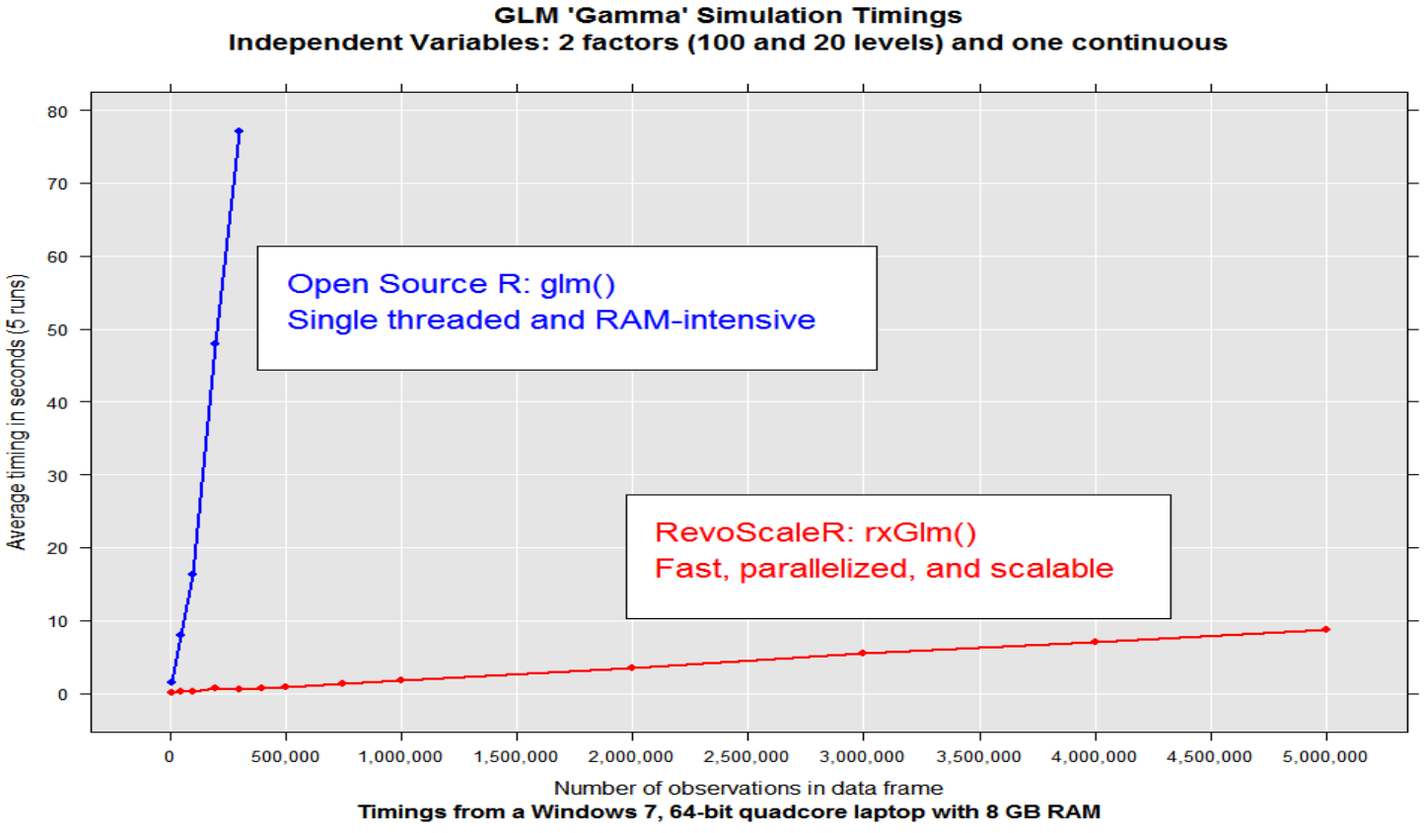
CRAN, MRO, MRS Comparison

		**MRO** 	**MRS** 
Datasize	In-memory	In-memory	In-Memory or **Disk Based**
Speed of Analysis	Single threaded	**Multi-threaded**	Multi-threaded, **parallel processing 1:N servers**
Support	Community	Community	Community **+ Commercial**
Analytic Breadth & Depth	7500+ innovative analytic packages	7500+ innovative analytic packages	7500+ innovative packages **+ commercial parallel high-speed functions**
Licence	Open Source	Open Source	Commercial license.

Microsoft

ScaleR - Performance comparison

Microsoft R Server has no data size limits in relation to size of available RAM. When open source R operates on data sets that exceed RAM it will fail. In contrast Microsoft R Server scales linearly well beyond RAM limits and parallel algorithms are much faster.



File Name	Compressed File Size (MB)	No. Rows	Open Source R (secs)	Revolution R (secs)
Tiny	0.3	1,235	0.00	0.05
V. Small	0.4	12,353	0.21	0.05
Small	1.3	123,534	0.03	0.03
Medium	10.7	1,235,349	1.94	0.08
Large	104.5	12,353,496	60.69	0.42
Big (full)	12,960.0	123,534,969	Memory!	4.89
V.Big	25,919.7	247,069,938	Memory!	9.49
Huge	51,840.2	494,139,876	Memory!	18.92

- US flight data for 20 years
- Linear Regression on Arrival Delay
- Run on 4 core laptop, 16GB RAM and 500GB SSD

Microsoft

MRS

Microsoft

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

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

Metadata Retrieval

- All calculated on import, and retrieved from the XDF file header.

- rxGetInfo, rxGetVarInfo, rxGetVarNames

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

- The eXternal Data Frame (XDF)

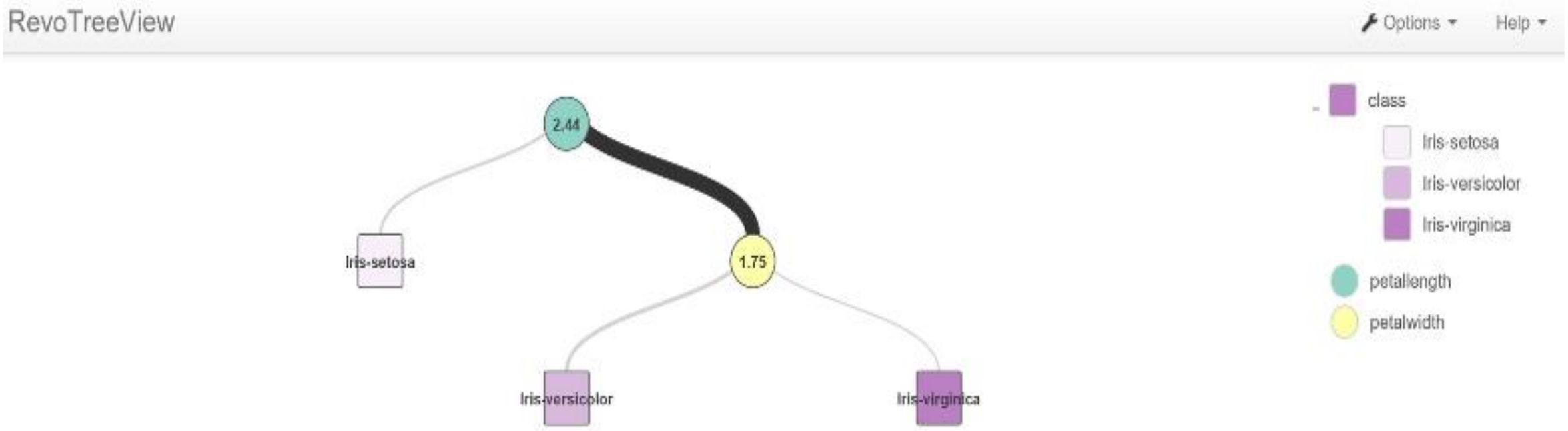
Modeling Algorithms

- 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)

Trees

root 150 100 Iris-setosa (0.33333333 0.33333333 0.33333333)
  petallength< 2.445 50 0 Iris-setosa (1.00000000 0.00000000 0.00000000) *
  petallength>=2.445 100 50 Iris-versicolor (0.00000000 0.50000000 0.50000000)
    petalwidth< 1.7495 54 5 Iris-versicolor (0.00000000 0.90740741 0.09259259) *
    petalwidth>=1.7495 46 1 Iris-virginica (0.00000000 0.02173913 0.97826087) *

Visualization



Microsoft

Write Once Deliver Anywhere

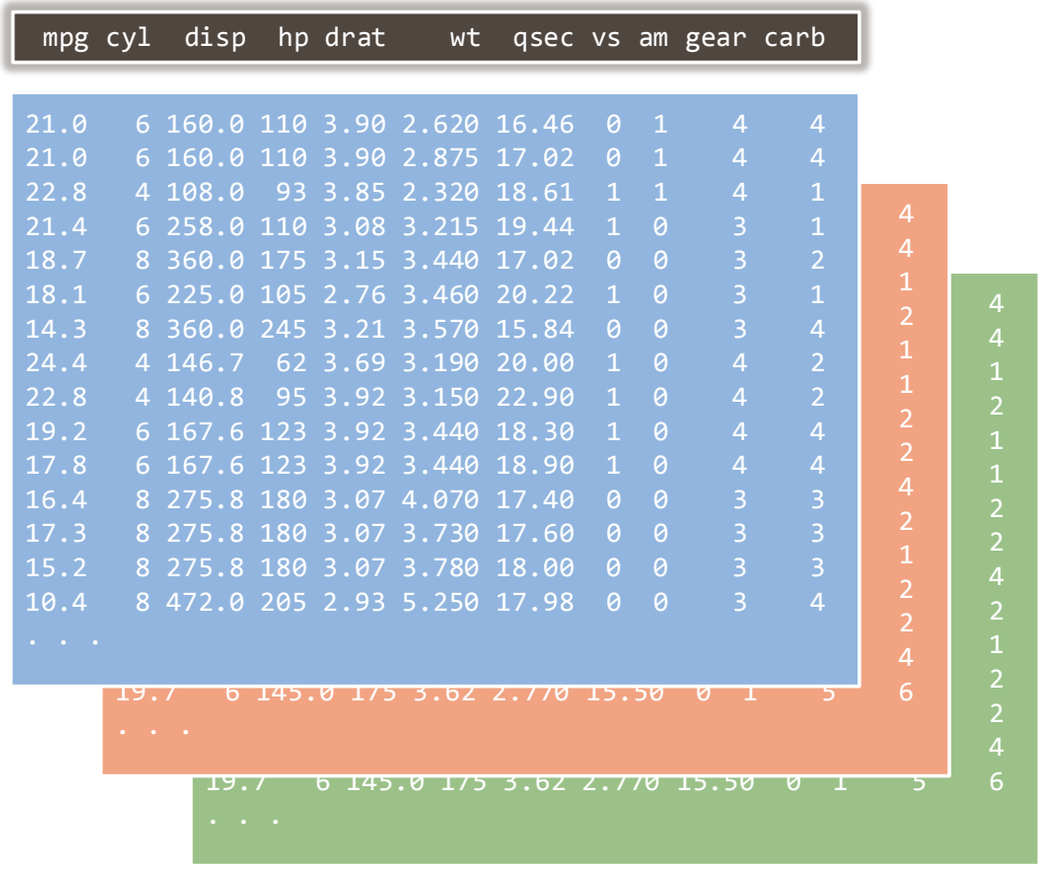
Delivering analytic models faster.

Building analytic models over big data.

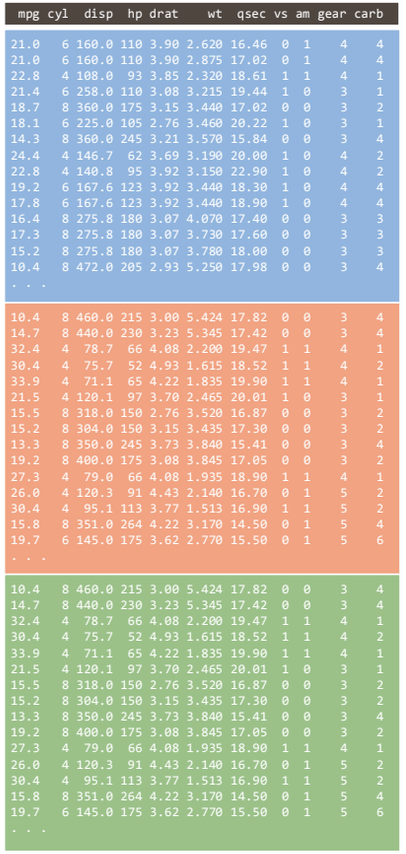
Microsoft

ScaleR – Parallel + Big Data

- Partition Datasets on Disk



In an Xdf file (local)



Microsoft

Local compute context

```
### LOCAL COMPUTE CONTEXT ###
  rxSetComputeContext("local")

### CREATE DIRECTORY AND FILE OBJECTS ###
    AirlineDatabase <-file.path("datasets","AirlineDemoSmall")
    AirlineDataSet <- RxXdfData(file.path(AirlineDatabase,"AirlineDemoSmall.xdf"))
```

```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
  rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
  rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
  arrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
  plot(arrLateLinMod$coefficients)
```

Microsoft

Remote compute: Teradata

```
### SETUP TERADATA ENVIRONMENT VARIABLES ###
dbConnStr <- "Driver=Teradata; Server=dbHostName; Database=RevoDb; Uid=xxxx; pwd=xxxx"
myTeradataCC <- RxInTeradata(connectionString = dbConnStr, shareDir = "/tmp",
    remoteShareDir = "/tmp/revoJobs", revoPath = "/usr/lib64/Revo-7.0/R-3.0.2/lib64/R")

### TERADATA COMPUTE CONTEXT ###
rxSetComputeContext(myTeradataCC)

### CREATE TERADATA DATA SOURCE ###
AirlineDemoQuery <- "SELECT * FROM AirlineDemoSmall;"
AirlineDataSet <- RxTeradata(connectionString = dbConnStr, sqlQuery = AirlineDemoQuery)
```

```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
  rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
  rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
  arrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
  plot(arrLateLinMod$coefficients)
```

Microsoft

Remote compute: Hadoop

```
### SETUP HADOOP ENVIRONMENT VARIABLES ###
myNameNode <- "master"
myUser <- "root"
myPort <- 8020
myHadoopCluster <- RxHadoopMR(sshUsername = myUser, sshHostname = myNameNode, port = myPort)

### HADOOP COMPUTE CONTEXT USING HDFS ###
rxSetComputeContext(myHadoopCluster)

### CREATE HDFS, DIRECTORY AND FILE OBJECTS ###
hdfsFS <- RxHdfsFileSystem(hostName=myNameNode, port=myPort)
AirlineDatabase <-file.path("datasets","AirlineDemoSmall")
AirlineDataSet <- RxXdfData(file.path(AirlineDatabase,"AirlineDemoSmall.xdf"), fileSystem = hdfsFS)
```

```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
  rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
  rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
  arrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
  plot(arrLateLinMod$coefficients)
```

Microsoft

Remote compute: SQL Server

```
### SETUP SQL SERVER ENVIRONMENT VARIABLES ###
dbConnStr <- "Driver=SQL Server; Server=dbHostName; Database=RevoDb; Uid=xxxx; pwd=xxxx"
mySqlServerCC <- RxInSqlServer(connectionString = dbConnStr, consoleOutput = TRUE)

### SQL SERVER COMPUTE CONTEXT ###
rxSetComputeContext(mySqlServerCC)

### CREATE SQL SERVER DATA SOURCE ###
AirlineDemoQuery <- "SELECT * FROM AirlineDemoSmall;"
AirlineDataSet <- RxSqlServer(connectionString = dbConnStr, sqlQuery = AirlineDemoQuery)
```

```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
  rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
  rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
  arrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
  plot(arrLateLinMod$coefficients)
```

Microsoft

The Data Science Virtual Machine

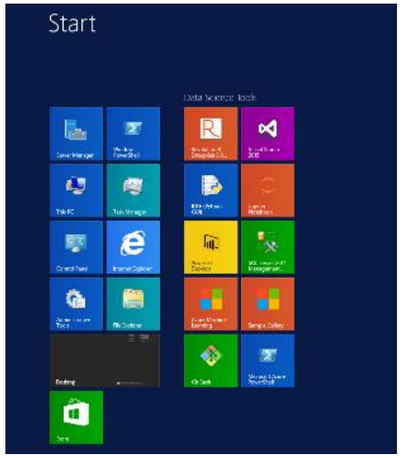
An Azure VM for best of Open Source Data Science quickly.

The Data Science Super Computer when we need it.

Microsoft

The Data Science Super Computer

- Specialized VM image on Azure.
- Data Science and Azure tools and SDKs.
- Pre-configured and ready to use.
- Pay for cloud hardware usage only.
- No separate software charges!
- Windows and **Linux** Versions.
- Up and running quickly – 5 minutes.

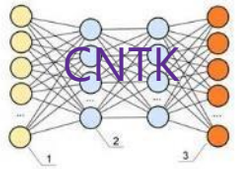


Microsoft

What's Included?

xgboost

Vowpal Wabbit



Rattle

Microsoft

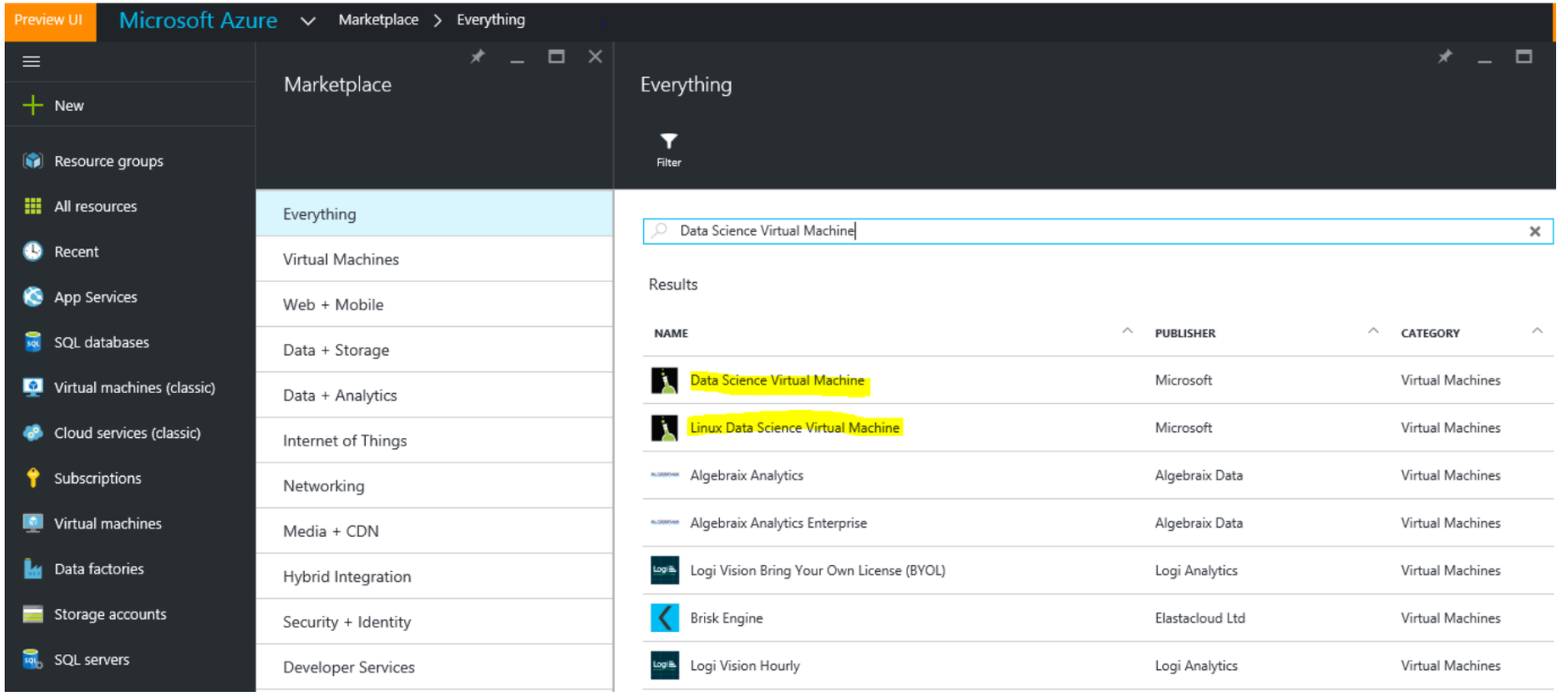
VM Versions comparison

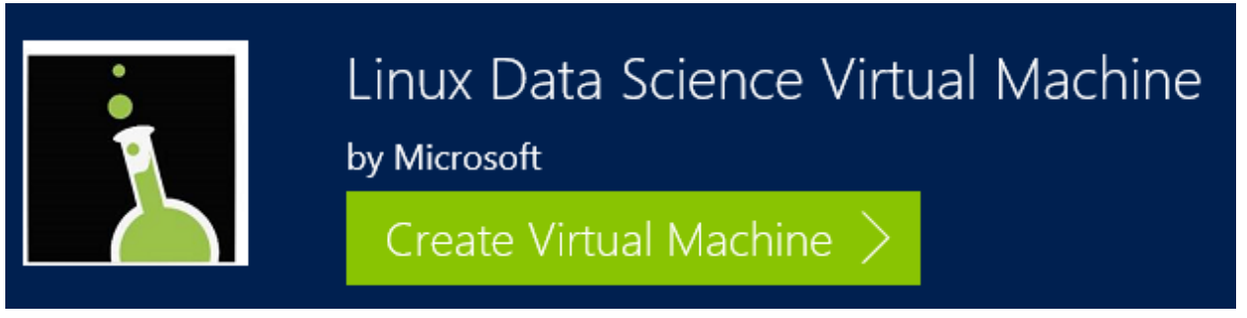
- Windows Version
  - Microsoft R Server (Enterprise R, R Open, MKL)
  - Anaconda Python 2.7, 3.5
  - Jupyter Notebooks (R, Python)
  - SQL Server 2014 Express
  - Visual Studio Community Edition 2015
    - Azure SDKs, HDInsight Tools, Data lake Tools
    - Python and R Tools or Visual Studio (IDE)
  - Power BI Desktop
  - ML Tools
    - Integrations to Azure Machine Learning
    - CNTK (Deep Learning)
    - Xgboost (Popular tool in data science competitions)
    - Vowpal Wabbit (Fast Online Learner)
    - Rattle (Visual quick start data analytics tool)
  - APIs to access Azure and Cortana Intelligence Suite services
  - Tools for data transfer to and from accessing Azure and Big Data storage technologies (Azure Storage Explorer, Powershell)
  - Git
  - Linux/Unix utilities through Git-Bash and Windows Command Prompt

- Linux Version
  - Microsoft R Open (Open Source R + MKL)
  - Anaconda Python 2.7, 3.5
  - Jupyter Notebooks (R, Python)
  - Postgres, SQuirreL SQL (Database tool), SQL Server Drivers and Command Line (bcp, sqlcmd)
  - Eclipse with Azure toolkit plugin
  - Emacs (with ESS, auctex)
  - ML Tools
    - Integrations to Azure Machine Learning
    - CNTK (Deep Learning)
    - Xgboost (Popular tool in data science competitions)
    - Vowpal Wabbit (Fast Online Learner)
    - Rattle (Visual quick start data analytics tool)
  - APIs to access Azure and Cortana Intelligence Suite services
  - Azure Command Line for administration
  - Azure Storage Explorer
  - Git

Microsoft

Creating a DSVM



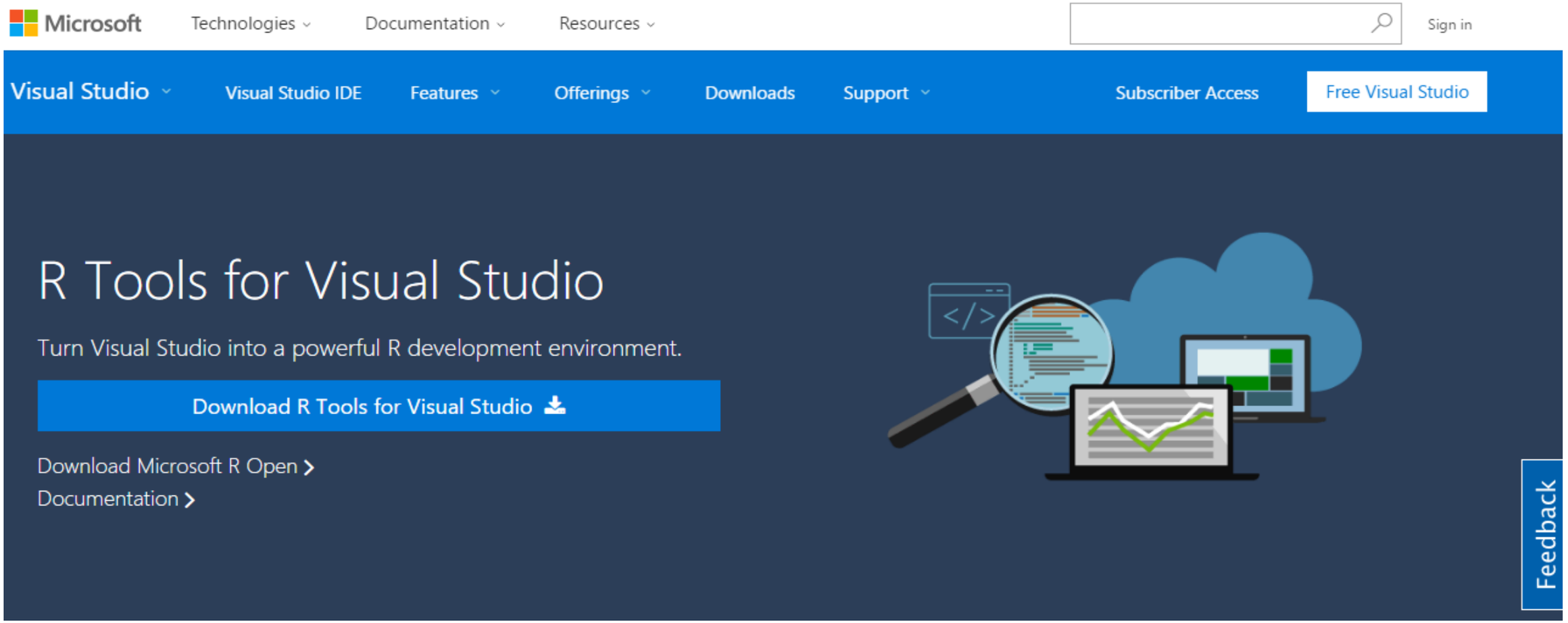


R Tools for Visual Studio

Microsoft

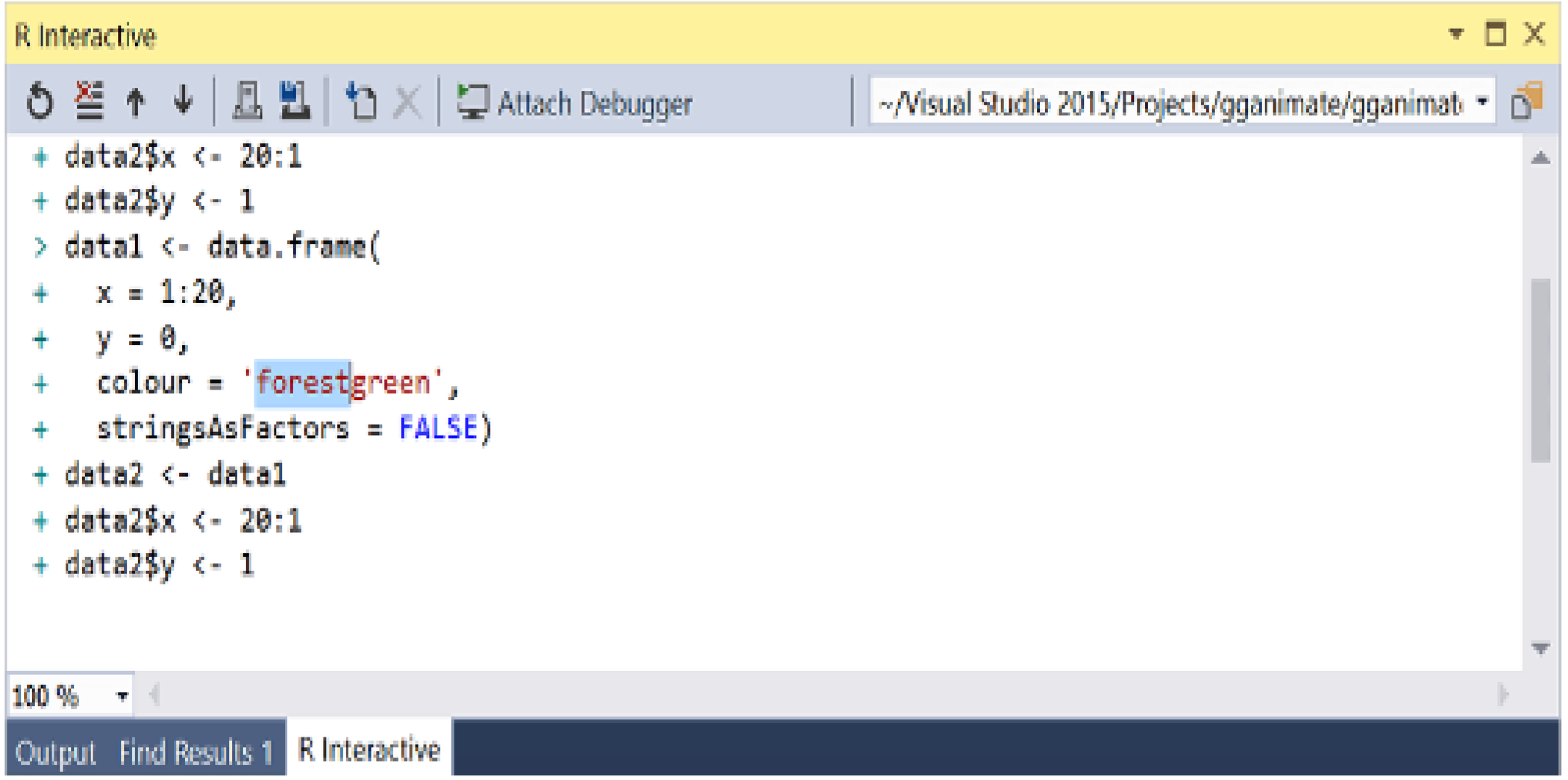
Visual Studio

https://www.visualstudio.com/vs/rtvs/



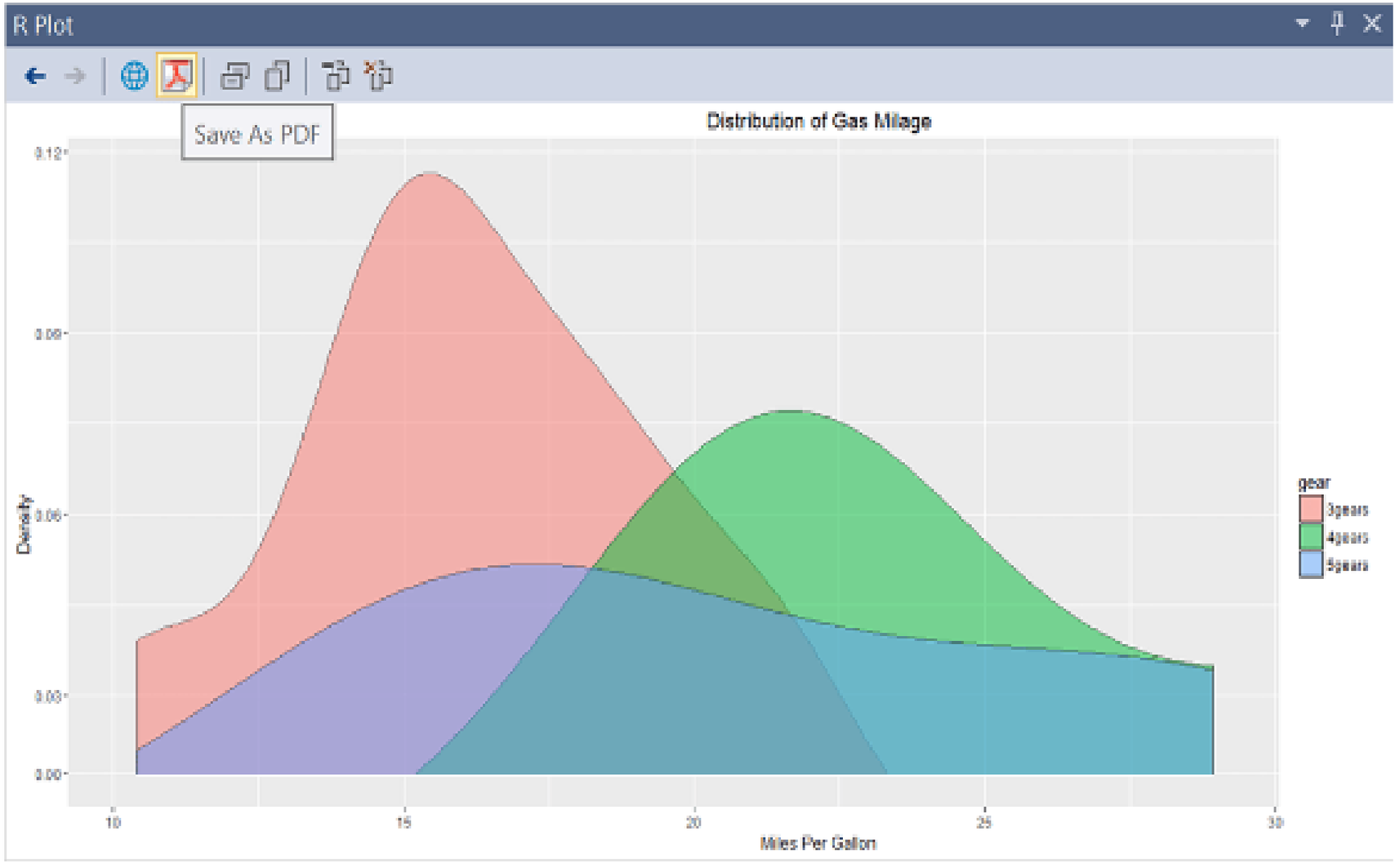
Microsoft

R Interactive window



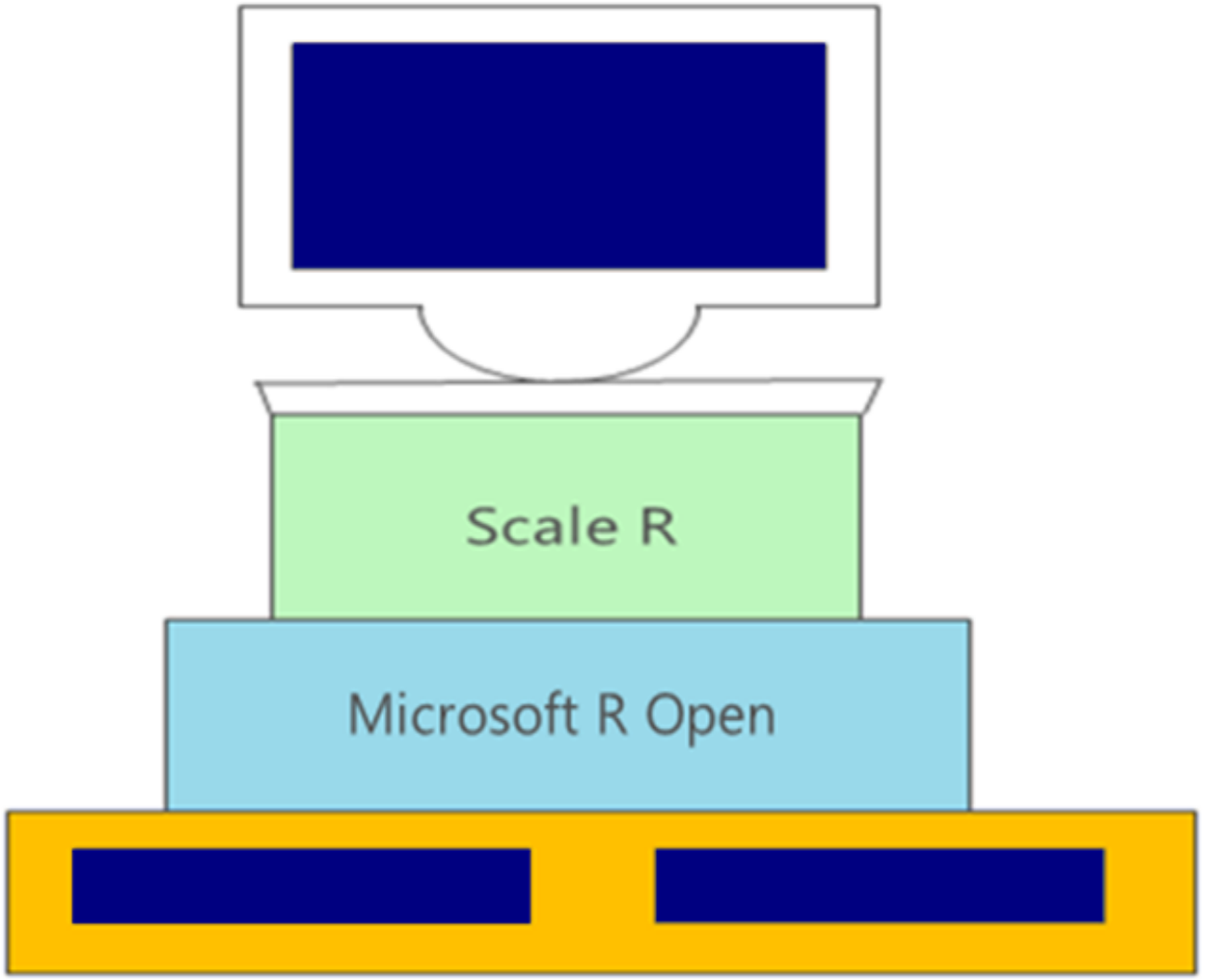
Microsoft

Plots



Microsoft

R Client



Microsoft

Labs

IRIS Dataset

Features

Sepallength
Sepalwidth

Petallength
Petalwidth

Classes

Iris-setosa

Iris-versicolor

Iris-virginica

Microsoft

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)

Importing to XDF

rxImport

- InData
- OutFile
- VarsTokeep, varsToDrop
- numRows
- rowsPerRead
- rowSelection
- Overwrite
- append

Microsoft

Formulas in RevoScaleR

- RevoScaleR uses a variant of the Wilkinson-Rogers formula notation (Wilkinson & Rogers, 1973)
- Similar to standard R modeling functions
- The dependent variables are separated from the predictor, or independent, variables by a tilde (~)
- Independent variables (predictors) are separated by plus signs (+)
- Interaction terms are joined with a colon (:)

Using Formula Syntax in Models

- One predictor:

  rxLinMod(y ~ x, data = myXdf)

- Two predictors:

  rxLinMod(y ~ x + z, data = myXdf)

- Two predictos with interaction term:

  rxLinMod(y ~ x * z, data = myXdf)