Microsoft, Open Source, R:

You Gotta be Kidding Me!

Bio - Niels Berglund

- **Software Specialist - Derivco**
  - lots of production dev. plus figuring out ways to "use and abuse" existing and new technologies
- **Author - "First Look at SQL Server 2005 for Developers"**
- **Researcher / Instructor - DevelopMentor**
- **Speaker - TechEd, DevWeek, SQL Pass, etc.**
- **Longtime user of SQL Server**

- **www.derivco.com**
- **niels.it.berglund@gmail.com**
- **@nielsberglund**
- **http://nielsberglund.com**

Derivco

- **World's leading development house for online gaming software; Casino, Poker, Bingo etc.**
- **Offices in Durban, Cape Town, Pretoria**
  - Estonia, Hong Kong, Sweden, UK
- **Technology company**
  - one of the world's largest install base of SQL Server's
  - SQL Server 2014, 2016
  - .NET 4.5
  - Hadoop, Windows Azure
  - stream processing, Complex Event Processing
  - data science R, Azure ML, etc.
  - RabbitMQ, CouchBase, in-memory databases, etc.

We Are Hiring

Agenda

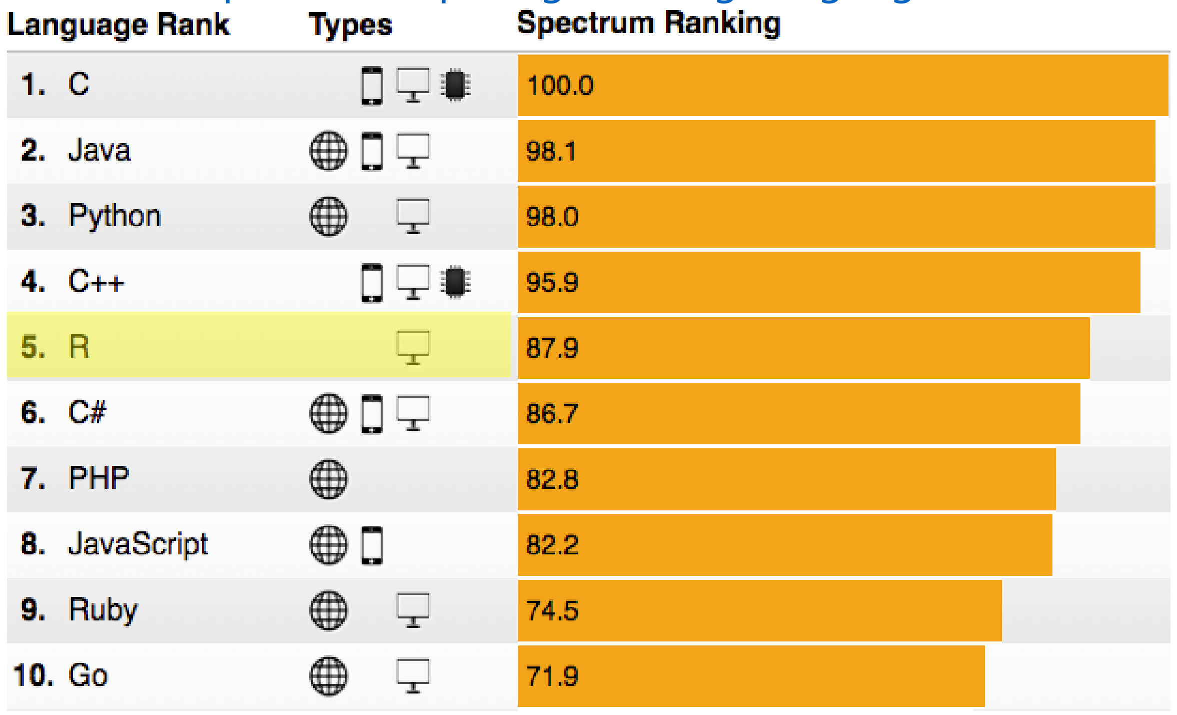
- R?
- Microsoft R Services
- R in SQL Server

R

- **Interpreted open source language for statistical computing**
- **Probably the most popular language for advanced analytics**

Language Popularity

[IEEE Spectrum Top Programming Languages](), 2016



Language Rank	Types	Spectrum Ranking
1. C	📱 🖥 ▪	100.0
2. Java	🌐 📱 🖥	98.1
3. Python	🌐 🖥	98.0
4. C++	📱 🖥 ▪	95.9
5. R	🖥	87.9
6. C#	🌐 📱 🖥	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐 📱	82.2
9. Ruby	🌐 🖥	74.5
10. Go	🌐 🖥	71.9

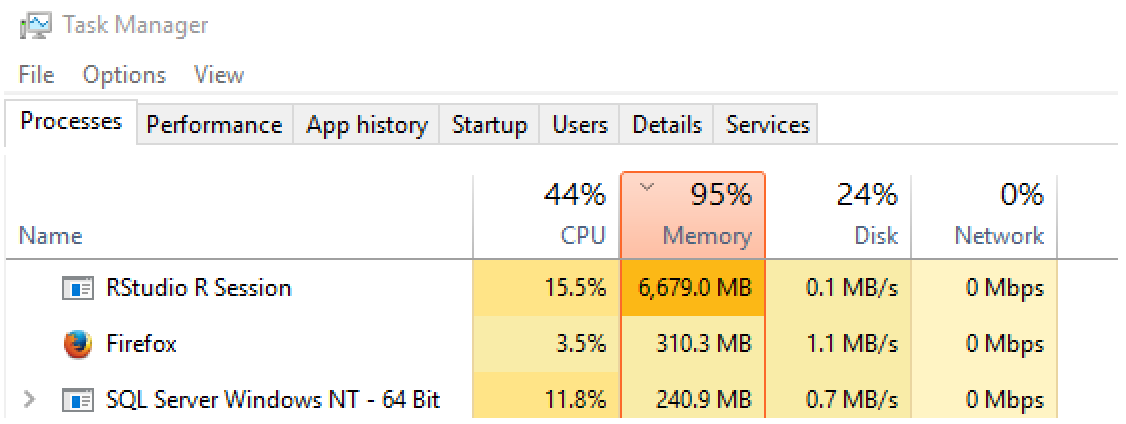
R challenges

- **Data movement**
  - data has to be moved from source to R
- **Operationalization**
  - you now have a model, how is it being called from your app??
- **Scale / performance**
  - R single threaded
  - datasets need to fit in memory

R - I

```
library(RODBC)

conn <- odbcDriverConnect(connection = "Driver={SQL Server native Client 11.0};
server=win10-dev;database=MortgageDb;uid=sa;pwd=sapwd")

mydata <- sqlQuery(conn, "SELECT CreditScore, HouseAge, YearsEmp, CreditCardDebt,
Year, DidDefault FROM MortgageData")

mydata$HouseAge <- factor(mydata$HouseAge)
mydata$Year <- factor(mydata$Year)

logit <- glm(DidDefault ~ HouseAge + Year + CreditScore + YearsEmp + CreditCardDebt,
data = mydata, family = "binominal")
```

R - II



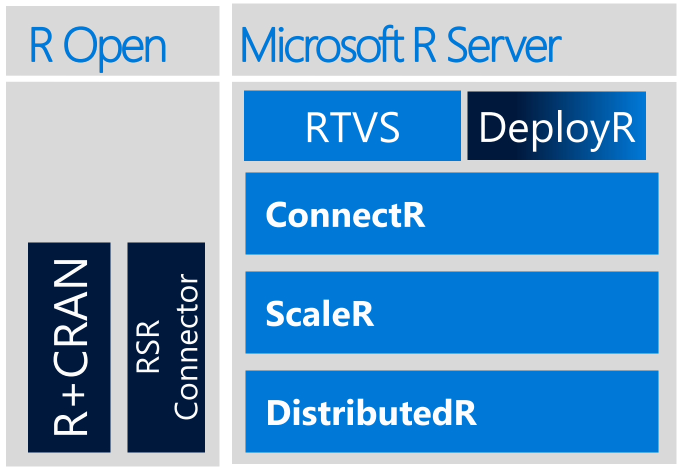
Task Manager

File Options View

Processes | Performance | App history | Startup | Users | Details | Services

Name	44% CPU	95% Memory	24% Disk	0% Network
RStudio R Session	15.5%	6,679.0 MB	0.1 MB/s	0 Mbps
Firefox	3.5%	310.3 MB	1.1 MB/s	0 Mbps
SQL Server Windows NT - 64 Bit	11.8%	240.9 MB	0.7 MB/s	0 Mbps

Microsoft R Server



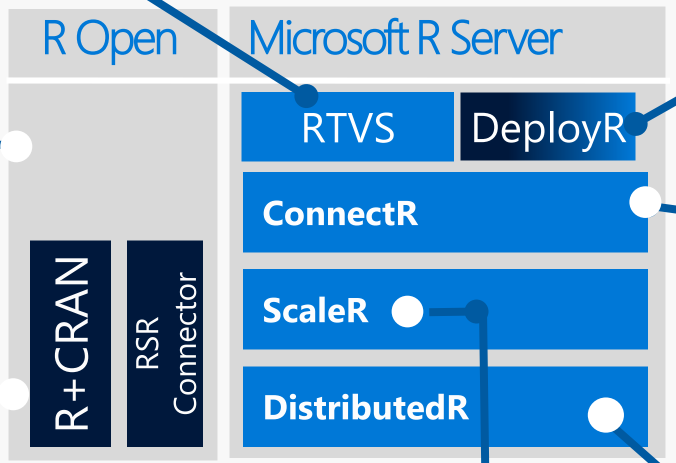
- **Enterprise class R**
  - Revolution Analytics (RevoScaleR package)
- **Works with open source R**
- **Enterprise scale and performance**
- **Secure, scalable R deployment / operationalization**
- **Write once deploy anywhere for multiple platforms**
  - RDBMS: SQL Server & TeraData
  - Windows, Linux: RedHat & SUSE
  - Hadoop: HortonWorks, Cloudera, MapR
  - Cloud: AzureVMs, Azure HDInsight
- **R tools for Visual Studio**

Microsoft R Server: Key Components

R+CRAN
- Open source R interpreter
  - R 3.1.2
- Freely-available huge range of R algorithms
- Algorithms callable by RevoR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages

Microsoft R Open
- Based on open source R
- High-performance math library to speed up linear algebra functions
- Checkpoint package to easily share R code and replicate results using specific R package versions

R Tools for Visual Studio
- State of the art, R Tools for Visual Studio IDE



ScaleR
- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Range of predictive functions
- User tools for distributing customized R algorithms across nodes
- Wide data sets supported – thousands of variables

DeployR
- RESTful APIs for easy integration from Java, JavaScript, .NET
- Enterprise authentication & security
- Horizontal scaling

ConnectR
- High-speed & direct connectors

Available for:
- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS (text & XDF)
- Teradata Database & Aster
- EDWs and ADWs
- ODBC

DistributedR
- Distributed computing framework
- Delivers cross-platform portability

ScaleR

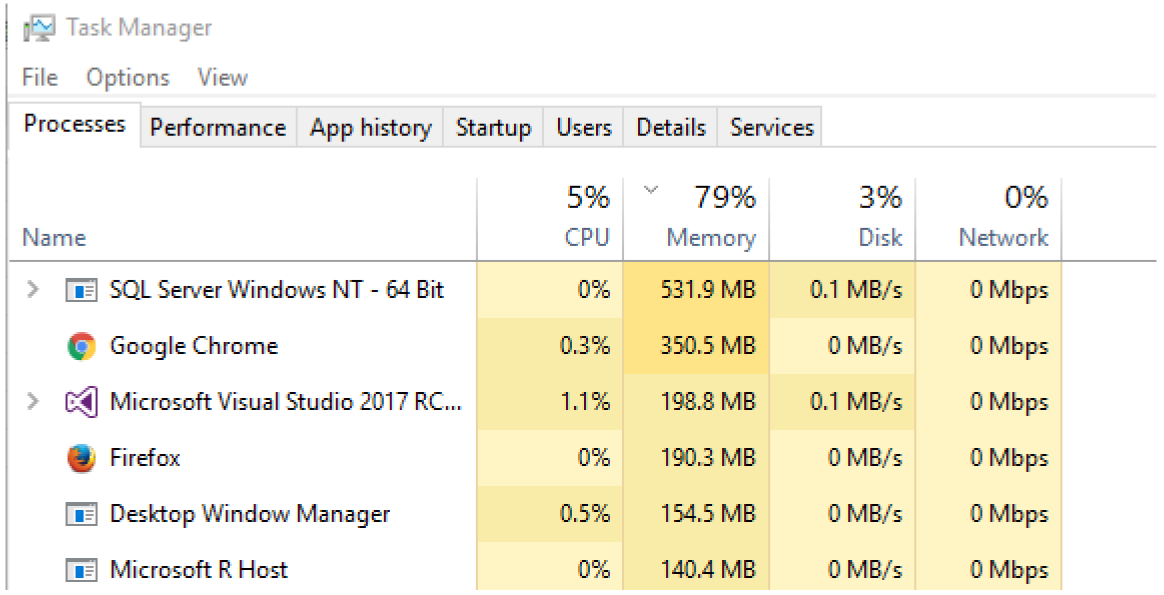
- **R package providing High Performance Computing and High Performance Analytics**
- **Distribute execution across cores and nodes**
- **The package introduces R Open Source equivalent functions (plus more)**
  - name normally starts with rx

```
# set the context, run on local machine or on a server machine
rxSetComputeContext("local")

rxGetComputeContext()
```

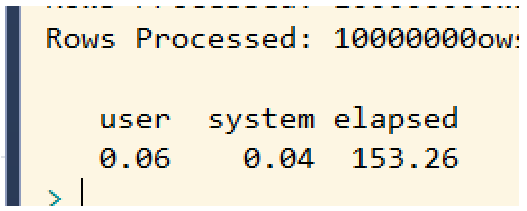
ScaleR - I

```
sqlServerConnString <- "Driver=SQL Server;server=win10-dev; database=MortgageDb;uid=sa;pwd=sapwd"

mydata <- RxSqlServerData(sqlQuery = "SELECT CreditScore, HouseAge, YearsEmp, CreditCardDebt, Year, DidDefault FROM
MortgageData",  connectionString = sqlServerConnString, rowsPerRead = 1000000)

rxHistogram( ~ CreditScore, data = mydata);
rxGetInfo(mydata, numRows = 5);
system.time(print(rxSummary( ~ ., data = mydata, blocksPerRead = 2)));

system.time(
logit <- rxLogit(DidDefault ~ F(HouseAge) + F(Year) + CreditScore + YearsEmp + CreditCardDebt,
        data = mydata, blocksPerRead = 2, reportProgress = 1))
```

ScaleR - II


Task Manager
File Options View
Processes Performance App history Startup Users Details Services

Name	5% CPU	79% Memory	3% Disk	0% Network
SQL Server Windows NT - 64 Bit	0%	531.9 MB	0.1 MB/s	0 Mbps
Google Chrome	0.3%	350.5 MB	0 MB/s	0 Mbps
Microsoft Visual Studio 2017 RC...	1.1%	198.8 MB	0.1 MB/s	0 Mbps
Firefox	0%	190.3 MB	0 MB/s	0 Mbps
Desktop Window Manager	0.5%	154.5 MB	0 MB/s	0 Mbps
Microsoft R Host	0%	140.4 MB	0 MB/s	0 Mbps



```
Rows Processed: 10000000ows

   user  system elapsed
   0.06    0.04  153.26
>
```

SQL Server R services

- New feature in SQL Server 2016
- Starts new workload in SQL Server
- Using R as the language

SQL Server 2016 R vs. R challanges

- Data movement - execute on SQL Server
- Operationalization - use T-SQL stored procedures
- Scale / performance - execute in parallel, leverage in-memory, column store, etc.

R in SQL

- **R engine callable from SQL Server 2016**
- **SQL Server introduces Launchpad**
  - a service to execute external scripts in SQL Server!!!
- **External scripts need to be enabled**
  - server needs to be restarted

```
EXEC sp_configure  'external scripts enabled', 1
RECONFIGURE  WITH  OVERRIDE
```

Execute R in SQL Server 2016 - I

- **R code is executed via the Launchpad service**
- **It is executed as external scripts**
  - sp_execute_external_script
- **Parameters to define external script (language) specific concepts**

Execute R in SQL Server 2016 - II

```
DECLARE @input nvarchar(max) = '"SELECT CreditScore, HouseAge, YearsEmp, CreditCardDebt, Year, DidDefault
                                 FROM MortgageData WHERE DidDefault = -1"'

BEGIN

  DECLARE @model varbinary(max) = (SELECT TOP 1 model  FROM dbo.tb_RModel);
  EXEC sp_execute_external_script @language = N'R',
     @script = N'
       mod <- unserialize(as.raw(model));
       print(summary(mod))
       OutputDataSet<-rxPredict(modelObject = mod,
           data = InputDataSet,
           outData = NULL,
           predVarNames = "DidDefault", type = "response",
           writeModelVars = FALSE, overwrite = TRUE);
       str(OutputDataSet)
       print(OutputDataSet)',
  @input_data_1 = @input,
  @params = N'@model varbinary(max)',
  @model = @model
  WITH RESULT SETS ((Salary bigint));

END
```

Operationalizing, users and key scenarios

- **Data scientist: data exploration, predictive modeling**
  - use R IDE of choice, execute scripts in-database get results back (plots, models, etc.)
  - models can be stored in the db!!
- **Developer: creating applications using the models from the data scientist**
  - execute T-SQL procedures, which run R scripts, getting results back to application(s)
- **DBA: manage the database**
  - manage, secure and control resources for R

Summary

- R is de facto standard for analytics
- There are certain limitations of R
- Microsoft R Server enterprise class R implementation
- R scripts can be run inside SQL Server

Questions???

niels.it.berglund@gmail.com
@nielsberglund
http://nielsberglund.com