



\*PASS

# BUSINESS ANALYTICS DAY

JANUARY 11, 2017 CHICAGO, IL

## In-Database Analytics

with R and SQL Server 2016

**David Smith and Seth Mottaghinejad**

Microsoft

alteryx



ATTUNITY



SolidQ



PLEASE SILENCE  
CELL PHONES

# Agenda

- Introduction to Data Science and R
- Workshop: In-database analytics with R and SQL Server 2016
  - How to build and store and retrieve an R analytics model and use it to score new data
  - How to use R to produce data visualizations and manage them for retrieval by SQL Reporting Services and Power BI
  - Examples of everyday, practical data-wrangling operations where R can outshine SQL
  - Using Microsoft R's RevoScaleR package to scale analytics on very large datasets

**Session Prerequisites:** Laptop with one of

- **SQL Server 2016** (Enterprise or Express), including “R Services” component
- Azure account, and **Azure Data Science Virtual Machine**



# Your Presenters



DAVID SMITH

R Community Lead at Revolution Analytics, a Microsoft Company



David Smith is the R Community Lead at Microsoft. With a background in data science, he writes daily about applications of predictive analytics at the Revolutions blog ([blog.revolutionanalytics.com](http://blog.revolutionanalytics.com)), and is a co-author of "Introduction to R", the R manual. Follow David on Twitter as [@revodavid](https://twitter.com/revodavid).



SETH MOTTAGHINEJAD

Data Scientist, Microsoft

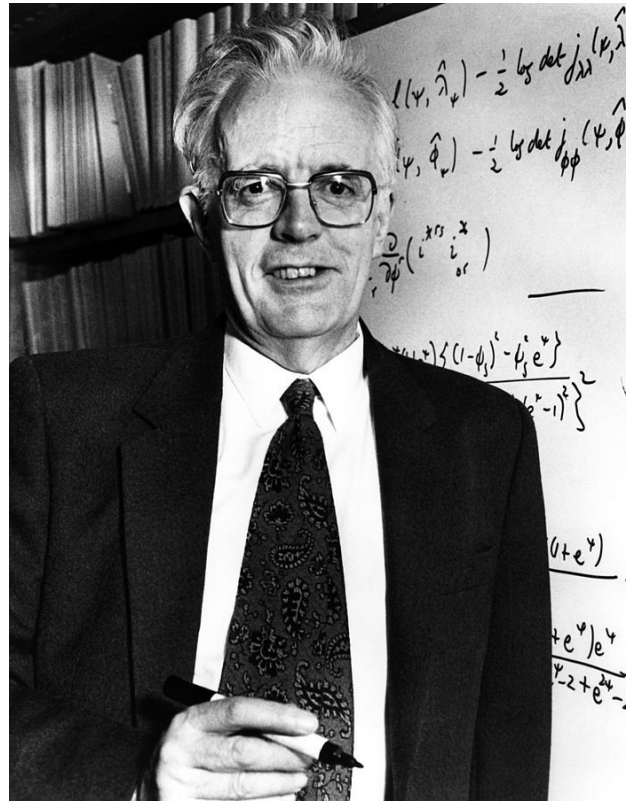


Seth is a data scientist at Microsoft who specializes in training and consulting for clients using Microsoft R Server (MRS). His past work includes training teams of data scientists to use R and MRS, showing how MRS fits in the big-data architecture, and helping with migration from tools such as SAS to R and MRS, and optimizing R performance. Before joining Microsoft, Seth worked as an analytics consultant at Revolution Analytics, the R-based big data and analytics company that was acquired by Microsoft in May 2015. Seth also has experience in marketing and customer analytics from prior jobs at American Express and Saks Fifth Avenue. He is a passionate "R-vangelist", an avid outdoorsman (who moved to Seattle to be close to lakes and mountains), and an amateur globetrotter.

An aerial photograph of a city skyline, likely Chicago, with a dense cluster of skyscrapers. The image is overlaid with a semi-transparent purple filter. The text "Introduction to R and Data Science" is centered in white. The background shows a body of water to the left and a cloudy sky above the city.

# Introduction to R and Data Science

# What's a data scientist, anyway?



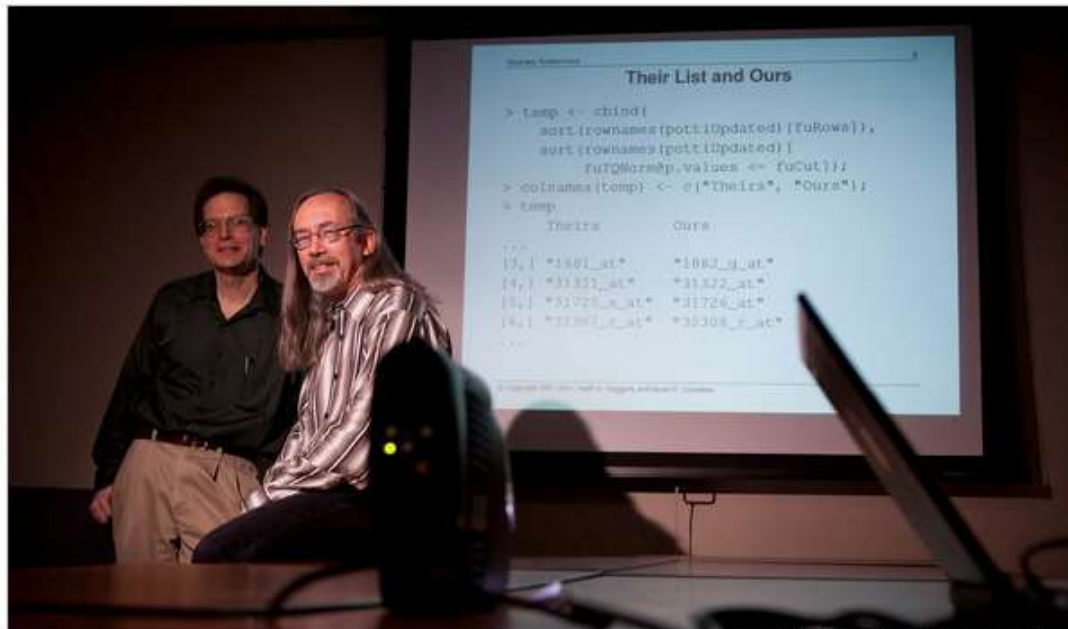


# What's a data scientist, anyway?



# What's a data scientist, anyway?

## How Bright Promise in Cancer Testing Fell Apart



Michael Stravato for The New York Times

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

By GINA KOLATA

Published: July 7, 2011



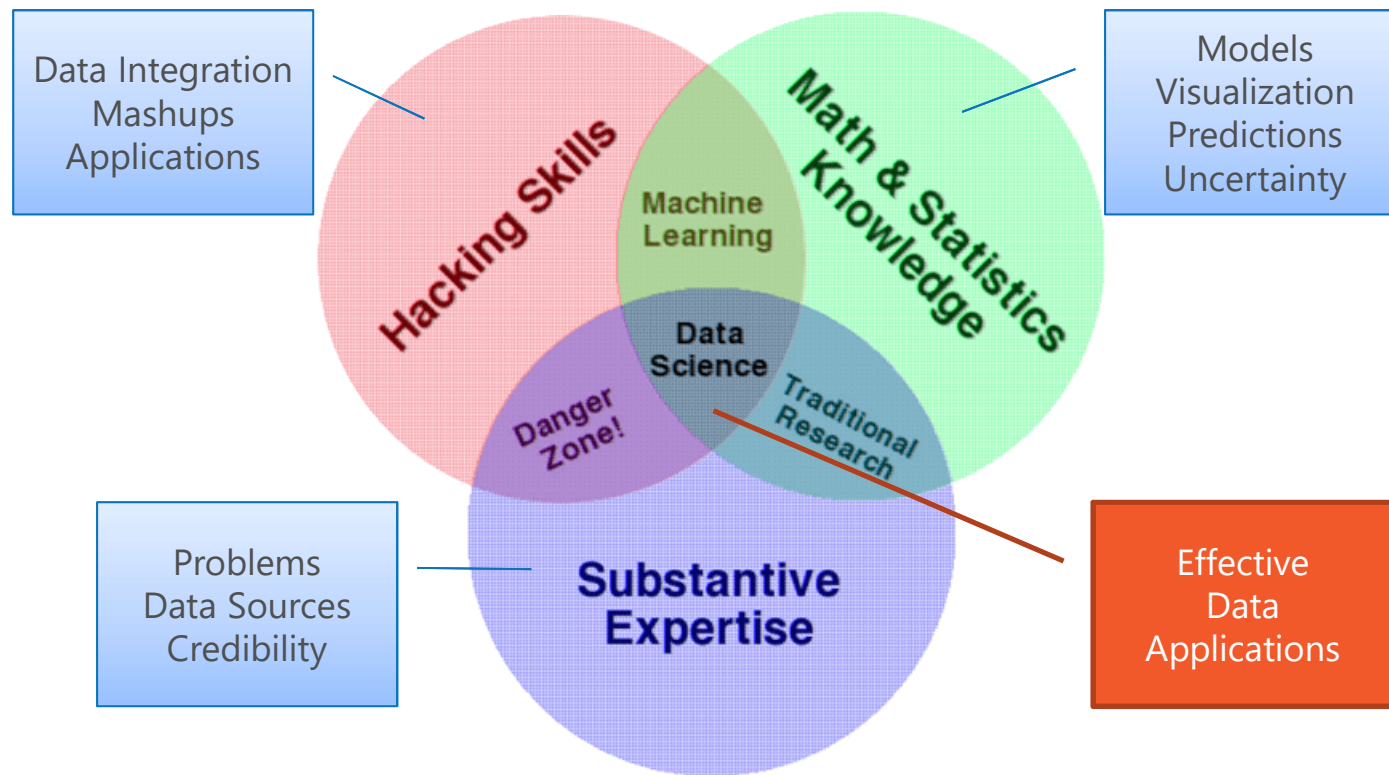
**D.J. Patil**

First U.S. Chief Data Scientist





# Three Essential Skills of Data Scientists





“It resonated with many people. It's not just a pretty picture, it's a reaffirmation of the impact we have in connecting people, even across oceans and borders.” — **Paul Butler**, data scientist, Facebook

# The New York Times

## What Happens After the I.P.O.?

Since 1980, there have been about 2,400 technology, Internet and telecom I.P.O.'s. On the first day of trading, the average stock rose 32 percent above its offer price.

But in the three years after that, most companies had negative returns, according to statistics compiled by Jay Ritter, a professor of finance at the University of Florida. Companies with higher values compared with their revenue before the I.P.O. have fared especially poorly.

### CHART KEY

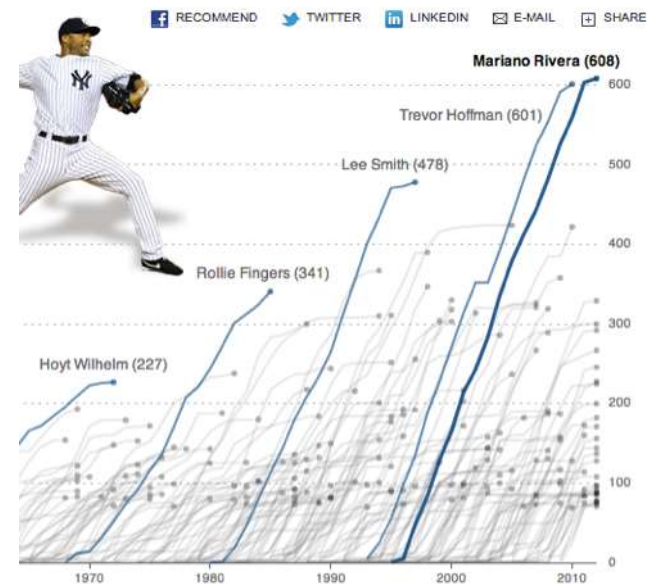
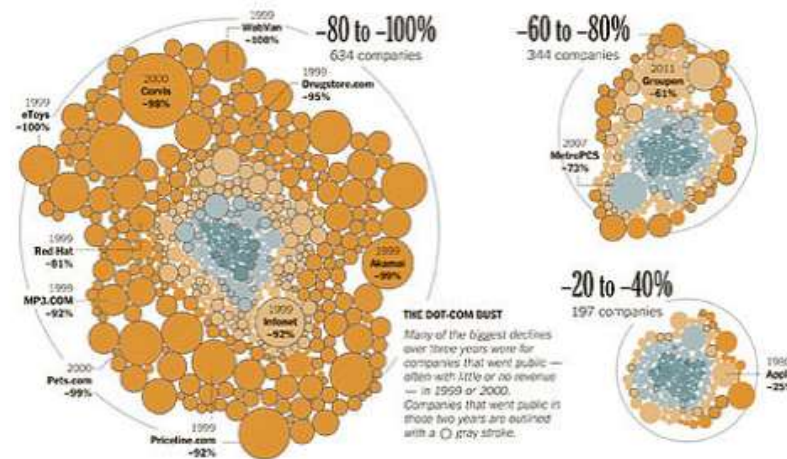
Circles are sized by value at the end of the first trading day, in today's dollars.



Colors show the ratio of the company's value to its revenue in the 12 months before the I.P.O.

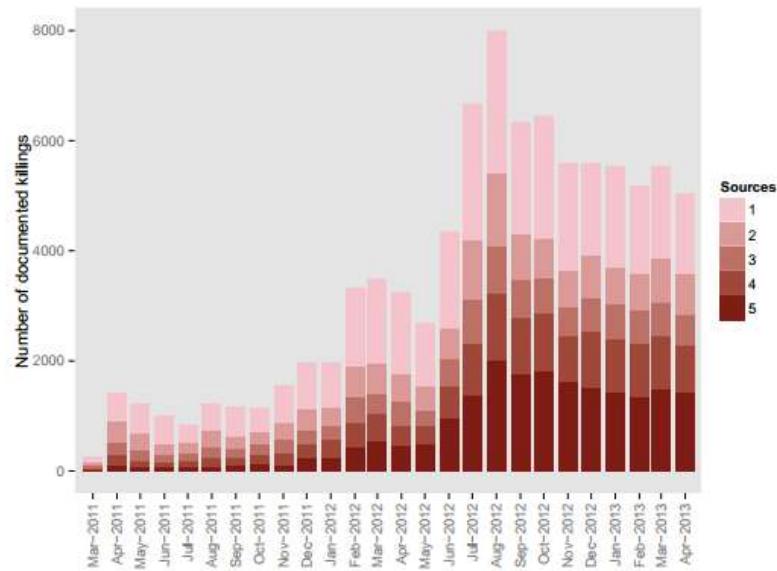


### Return three years after the I.P.O.: The decliners ...



- [Facebook IPO](#)
- [Baseball legends](#)

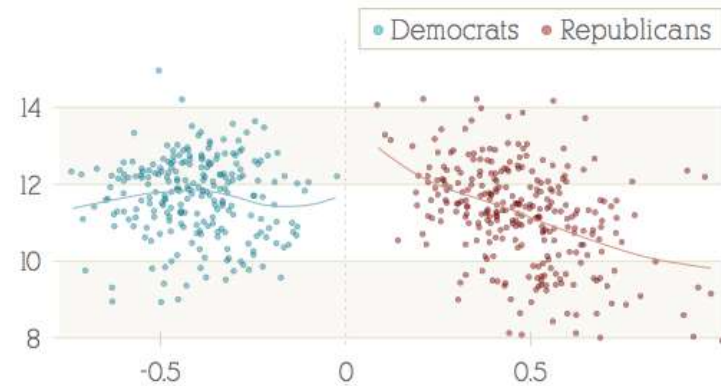




- [Casualty estimation in Warzones](#)

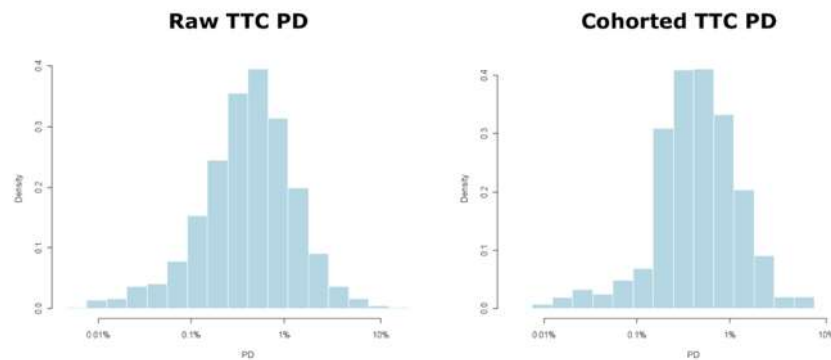


### Ideology and Grade Level of Congressional Record Speeches, Current Members

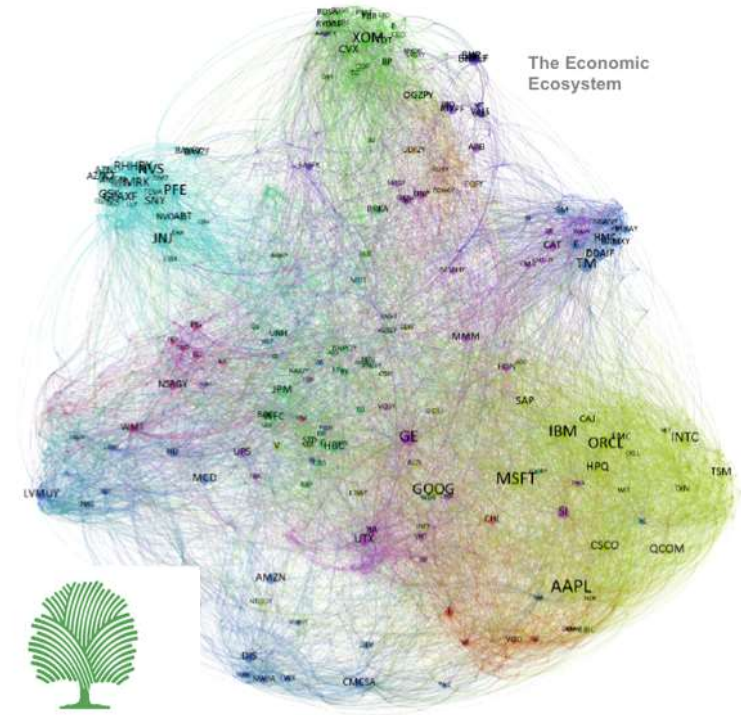


Liberal-Conservative Voting Score, 112th Congress  
-1 (most liberal) to 1 (most conservative)

- [Political Analysis](#)



- [Credit Risk Analysis](#)



- [Financial Networks](#)

# What is R?

Most widely used data analysis software

- Used by 2M+ data scientists, statisticians and analysts

Most powerful statistical programming language

- Flexible, extensible and comprehensive for productivity

Create beautiful and unique data visualizations

- As seen in New York Times, The Economist and FlowingData

Fills the Data Science talent gap

- New graduates prefer R

Thriving open-source community

- Leading edge of analytics research





# A brief history of R

1993 Research project in Auckland, NZ

- Ross Ihaka and Robert Gentleman

1995 Released as open-source software

- Generally compatible with the "S" language

1997 R core group formed

2003 R Foundation formed in Austria

2007 Revolution Analytics founded

2014 Revolution R Open launched

2015 R Consortium founded

2015 Microsoft acquires Revolution Analytics

2016 SQL Server 2016 R Services released



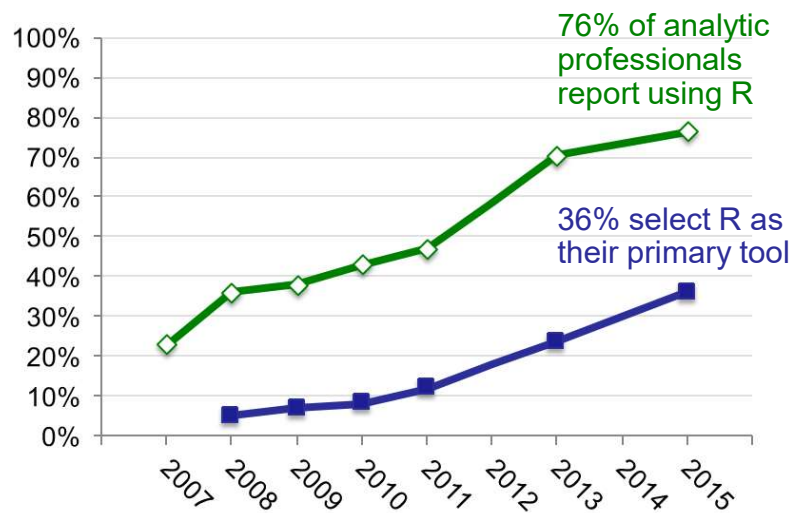
Photo credit: Robert Gentleman

# R: The #1 software for Data Science

... and #5 amongst general-purpose programming languages

## R Usage Growth

Rexer Data Miner Survey, 2007-2015



## 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
















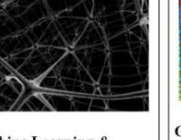
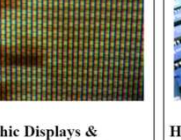
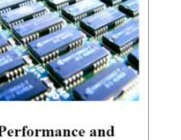

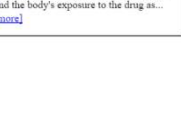





# 200 Local R User Groups Worldwide



[blog.revolutionanalytics.com/local-r-groups.html](http://blog.revolutionanalytics.com/local-r-groups.html)



# CRAN: 9000+ add-on packages for R

 <p><b>Bayesian Inference</b></p> <p>Applied researchers interested in Bayesian statistics are increasingly attracted to R because of the ease of which one can code algorithms to sample... <a href="#">[more]</a></p>	 <p><b>Chemometrics and Computational Physics</b></p> <p>Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of... <a href="#">[more]</a></p>	 <p><b>Clinical Trial Design, Monitoring, and Analysis</b></p> <p>This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including... <a href="#">[more]</a></p>	 <p><b>Cluster Analysis &amp; Finite Mixture Models</b></p> <p>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... <a href="#">[more]</a></p>	 <p><b>Probability Distributions</b></p> <p>For most of the classical distributions, base R provides probability distribution functions (p), density functions (d), quantile functions (q), and... <a href="#">[more]</a></p>	 <p><b>Computational Econometrics</b></p> <p>Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many... <a href="#">[more]</a></p>	 <p><b>Analysis of Ecological and Environmental Data</b></p> <p>This Task View contains information about using R to analyse ecological and environmental data... <a href="#">[more]</a></p>	 <p><b>Design of Experiments (DoE) &amp; Analysis of Experimental Data</b></p> <p>This task view collects information on R packages for experimental design and analysis of data from experiments. Please feel free to suggest enhancements... <a href="#">[more]</a></p>	 <p><b>Empirical Finance</b></p> <p>This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic... <a href="#">[more]</a></p>
 <p><b>Natural Language Processing</b></p> <p>This CRAN task view contains a list of packages useful for natural language processing... <a href="#">[more]</a></p>	 <p><b>Analysis of Pharmacokinetic Data</b></p> <p>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... <a href="#">[more]</a></p>	 <p><b>Optimization and Mathematical Programming</b></p> <p>This CRAN task view contains a list of packages which offer facilities for solving optimization problems. Although every regression model in statistics... <a href="#">[more]</a></p>	 <p><b>Phylogenetics, Especially Comparative Methods</b></p> <p>The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical... <a href="#">[more]</a></p>	 <p><b>Multivariate Statistics</b></p> <p>Base R contains most of the functionality for classical multivariate analysis, somewhere. There are a large number of packages on CRAN which extend this... <a href="#">[more]</a></p>	 <p><b>Official Statistics &amp; Survey Methodology</b></p> <p>This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide... <a href="#">[more]</a></p>	 <p><b>Machine Learning &amp; Statistical Learning</b></p> <p>Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually... <a href="#">[more]</a></p>	 <p><b>Graphic Displays &amp; Dynamic Graphics &amp; Graphic Devices &amp; Visualization</b></p> <p>R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic... <a href="#">[more]</a></p>	 <p><b>High-Performance and Parallel Computing with R</b></p> <p>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... <a href="#">[more]</a></p>
 <p><b>Analysis of Spatial Data</b></p> <p>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... <a href="#">[more]</a></p>	 <p><b>Time Series Analysis</b></p> <p>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... <a href="#">[more]</a></p>	 <p><b>Robust Statistical Methods</b></p> <p>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 = ...). <a href="#">[more]</a></p>	 <p><b>Survival Analysis</b></p> <p>Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an... <a href="#">[more]</a></p>	 <p><b>Statistics for the Social Sciences</b></p> <p>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... <a href="#">[more]</a></p>	 <p><b>gRaphical Models in R</b></p> <p>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... <a href="#">[more]</a></p>	 <p><b>Reproducible Research</b></p> <p>The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better... <a href="#">[more]</a></p>		


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

# Introduction to R

www.datacamp.com

The screenshot shows the DataCamp website homepage. The browser address bar displays 'https://www.datacamp.com'. The navigation bar includes links for 'Courses', 'Pricing', 'Business', 'Community', 'Sign in', and a 'Create Free Account' button. The main heading reads 'THE EASIEST WAY TO Learn Data Science Online'. Below this, a paragraph states: 'Master data analysis from the comfort of your browser, at your own pace, tailored to your needs and expertise. Whether you want to learn R, Python or Data Visualization, we want to help!'. Two buttons are visible: 'Start Learning R' (highlighted with a red circle) and 'Start Learning Python'. On the right, a 'Create Your Free Account' form includes social login options for LinkedIn, Facebook, and Google+, and input fields for 'Email' and 'Password', followed by a 'Get Started' button.

← → ↻ https://www.datacamp.com ☆

 **DataCamp**  
We're hiring!

[Courses](#) [Pricing](#) [Business](#) [Community](#) | [Sign in](#) [Create Free Account](#)

THE EASIEST WAY TO

# Learn Data Science Online


Master data analysis from the comfort of your browser, at your own pace, tailored to your needs and expertise. Whether you want to learn R, Python or Data Visualization, we want to help!


[Start Learning R](#)

[Start Learning Python](#)

### Create Your Free Account

[in](#) [f](#) [G+](#)





[Get Started](#)






An aerial photograph of a city skyline, likely Chicago, with numerous skyscrapers and a body of water visible in the background. The entire image is overlaid with a semi-transparent blue filter. The text 'Workshop' is centered in the upper half of the image.

Workshop

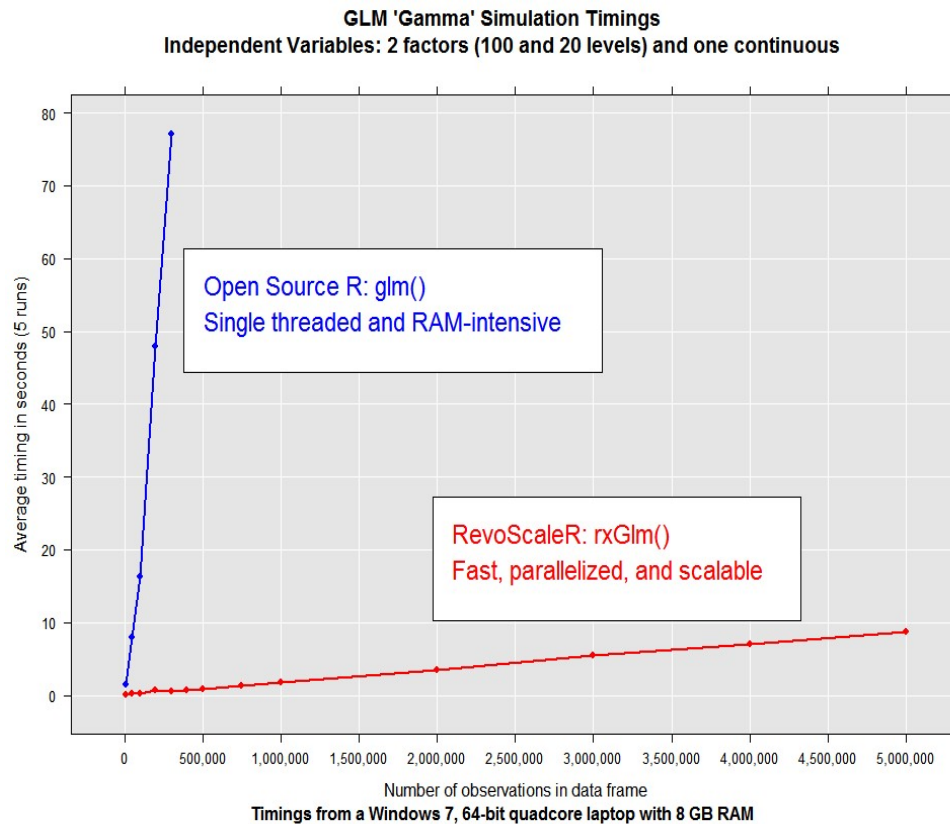
# Benefits of Microsoft R Server



# CRAN, MRO, MRS Comparison

		<b>R Open</b> 	<b>R Server</b> 
<b>Datasize</b>	In-memory	In-memory	In-Memory or Disk Based
<b>Speed of Analysis</b>	Single threaded	Multi-threaded	Multi-threaded, parallel processing 1:N servers
<b>Support</b>	Community	Community	Community + Commercial
<b>Analytic Breadth &amp; Depth</b>	7500+ innovative analytic packages	7500+ innovative analytic packages	7500+ innovative packages + commercial parallel high-speed functions
<b>Licence</b>	Open Source	Open Source	Commercial license. Supported release with indemnity

# RevoScaleR Overview



- (1) **Performance:** RevoScaleR functions work with data in RAM, but because they are parallel they outperform their open-source counterparts.
- (2) **No memory limit:** Once data size on disk surpasses available RAM, RevoScaleR functions can still be used to process data, chunk-wise, without loading the data into RAM in its entirety.
- (3) **Code portability:** RevoScaleR functions also work with data stored in SQL Server or on HDFS and can be executed in a remote compute context (thereby preserving data locality).

# Code Portability

RevoScaleR models can be deployed in SQL Server with only minor changes to the overall code structure

Compute context R script – sets where the model will run

## Local Parallel processing – Linux or Windows

```
### SETUP LOCAL ENVIRONMENT VARIABLES ###
myLocalCC <- RxLocalParallel()

### LOCAL COMPUTE CONTEXT ###
rxSetComputeContext(myLocalCC)

### POINT TO DATA STORED LOCALLY ###
AirlineDataSet <- RxXdfData("AirlineDemoSmall.xdf")
```

## In – SQL

```
### SETUP SQL COMPUTE CONTEXT ###
sqlCC <- RxInSqlServer(...)

### SET COMPUTE CONTEXT TO SQL SERVER ###
rxSetComputeContext(sqlCC)

### POINT TO DATA IN SQL SERVER ###
AirlineDataSet <- RxSqlServerData(connectionString =
..., table = "AirlineSmall")
```

Functional model R script – does not need to change to run in SQL

```
### 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
hdfsXdfArrLateLinMod <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
plot(hdfsXdfArrLateLinMod$coefficients)
```

An aerial photograph of a city skyline, likely Chicago, with numerous skyscrapers and a body of water visible in the background. The entire image is overlaid with a semi-transparent blue filter. The text is centered over the image.

Workshop

# In-database analytics with R and SQL Server 2016



# SQL Server 2016 R Services

- Bring analytics to data = in-database scalable computing
- Let SQL handle scaling, data governance, security, etc.
- Let R handle advanced analytics
- Develop in R IDE
- Operationalize as SQL Stored Procedures

# Meet our Heroes – The Data Scientist

I love R

Cool prototypes

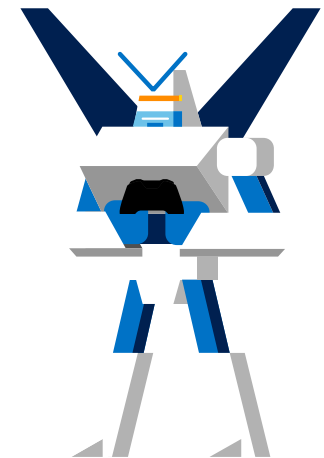
Huge potential

Management excited

Developer turnaround time

Access to production

Scale and performance



# Meet our Heroes – The Developer

I see the potential

Need to integrate with applications

Need to re-write

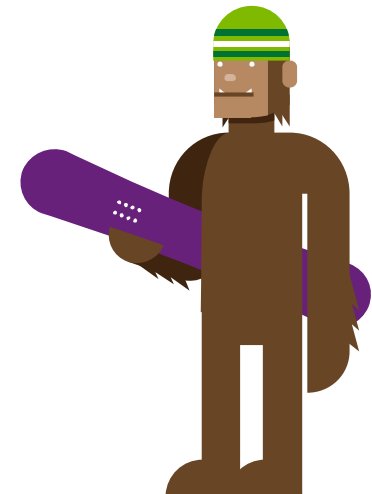
Re-write is difficult

Long turnaround

Error Prone

Scale and performance

Letting everybody down



# Meet our Heroes – The DBA

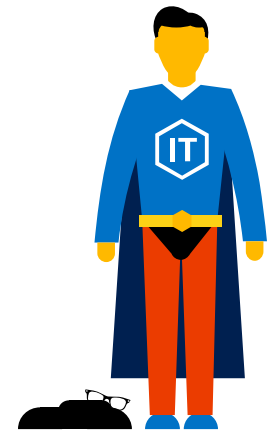
I see the potential

Pulling data out of production databases

Security and privacy

Scale

Data fragmentation



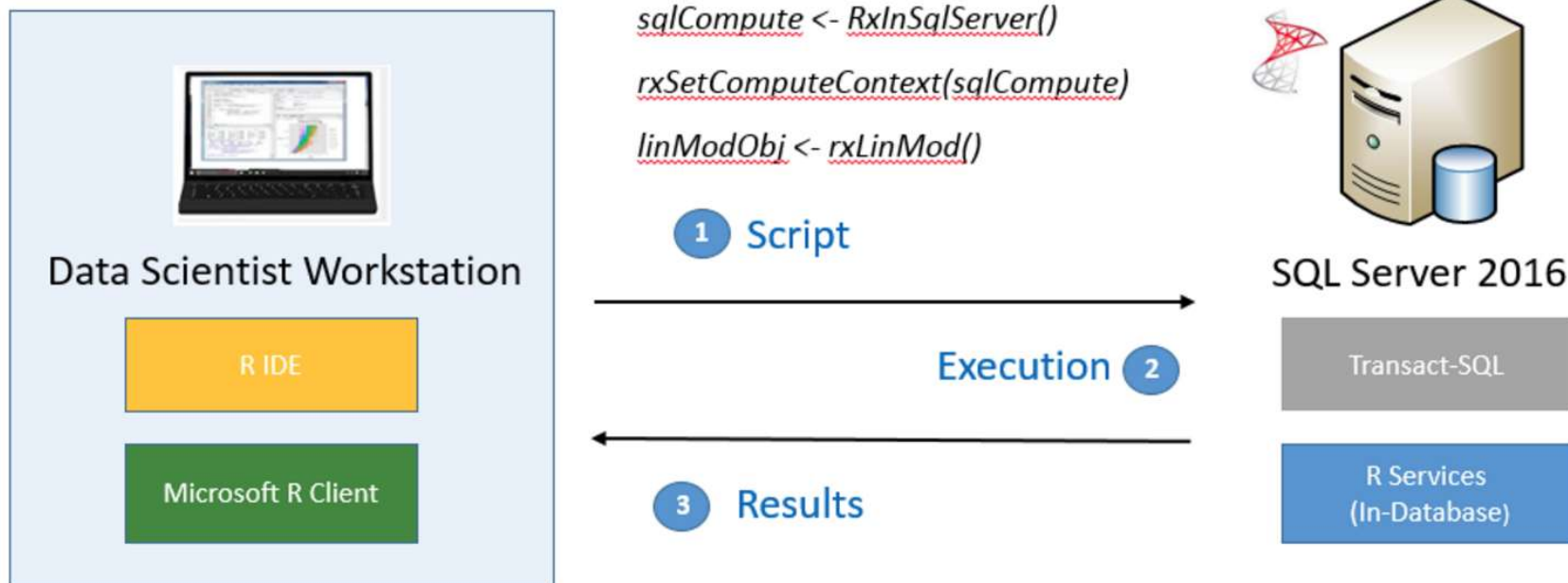


## Two ways to run R

### Two options to execute inside SQL

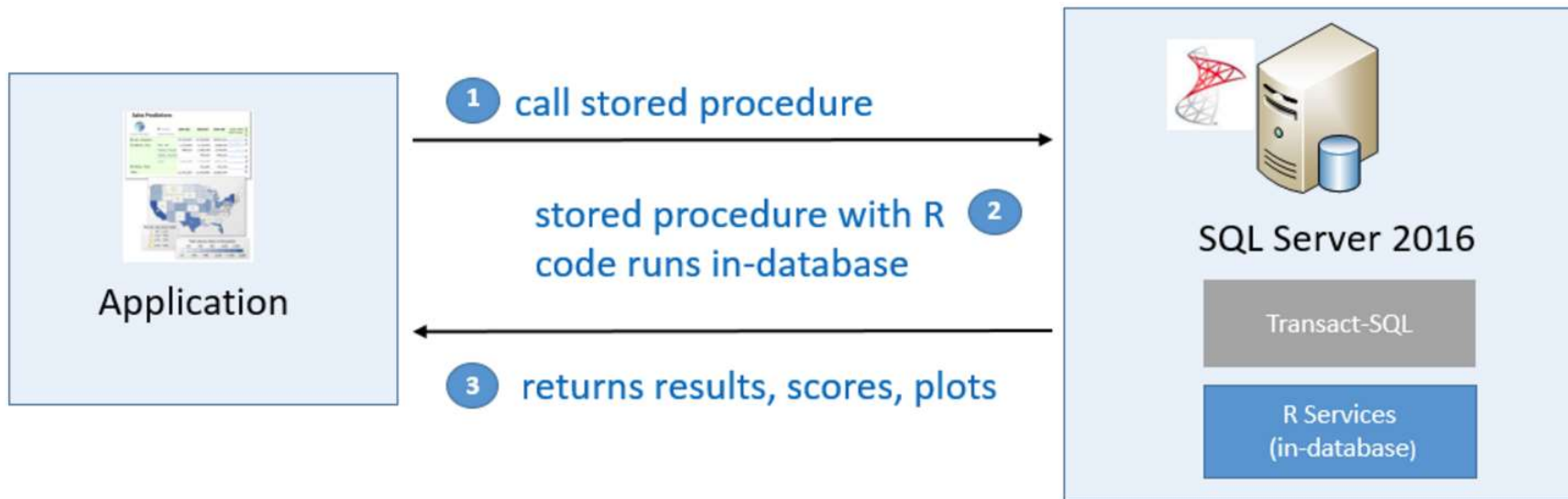
- **Data scientist scenario:** Start from R GUI and execute R code remotely
- **Production scenario:** Start from T-SQL with embedded R script

# Data scientist scenario



<https://msdn.microsoft.com/en-us/library/mt604885.aspx>

# Production scenario



<https://msdn.microsoft.com/en-us/library/mt604885.aspx>

# Lab agenda

Let's learn how we can use R and RevoScaleR to

- Taking data into SQL Server
- Pointing to data on SQL Server and dealing with column types
- Summarizing and visualizing data
- Prepping data for analysis
- Creating a predictive model and storing it in SQL Server
- Scoring on out of sample data in SQL Server



An aerial photograph of a city skyline, likely Chicago, featuring numerous skyscrapers and a body of water in the background. The entire image is overlaid with a semi-transparent purple filter. The text "Q & A" is centered in a white, serif font.

Q & A

An aerial photograph of a city skyline, likely New York City, with numerous skyscrapers. The image is overlaid with a semi-transparent purple filter. A large white diagonal line cuts across the right side of the image.

# Thank You

for attending this session  
and  
PASS Business Analytics Day!

alteryx  ATTUNITY  SolidQ