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), and is a co-author of "Introduction to R", the R manual. Follow David on Twitter as @revodavid.

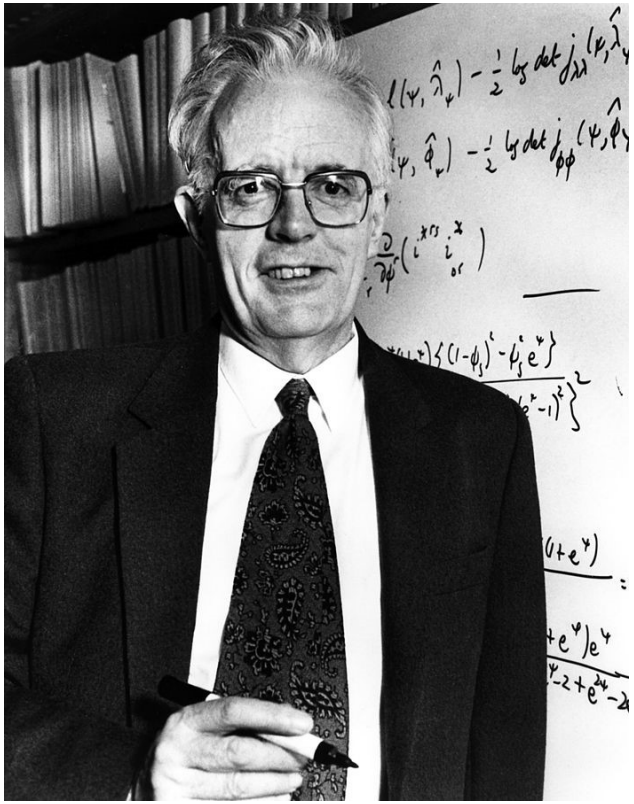
SETH MOTTAGHINEJAD

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.

Introduction to R and Data Science

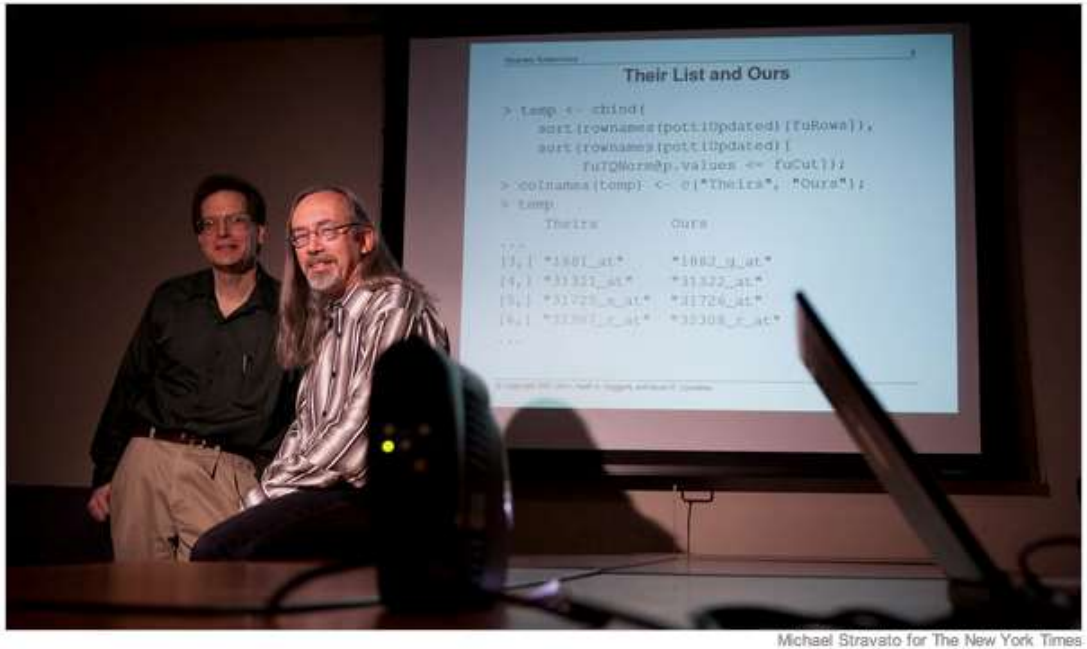
What’s a data scientist, anyway?



What's a data scientist, anyway?

What's a data scientist, anyway?

How Bright Promise in Cancer Testing Fell Apart



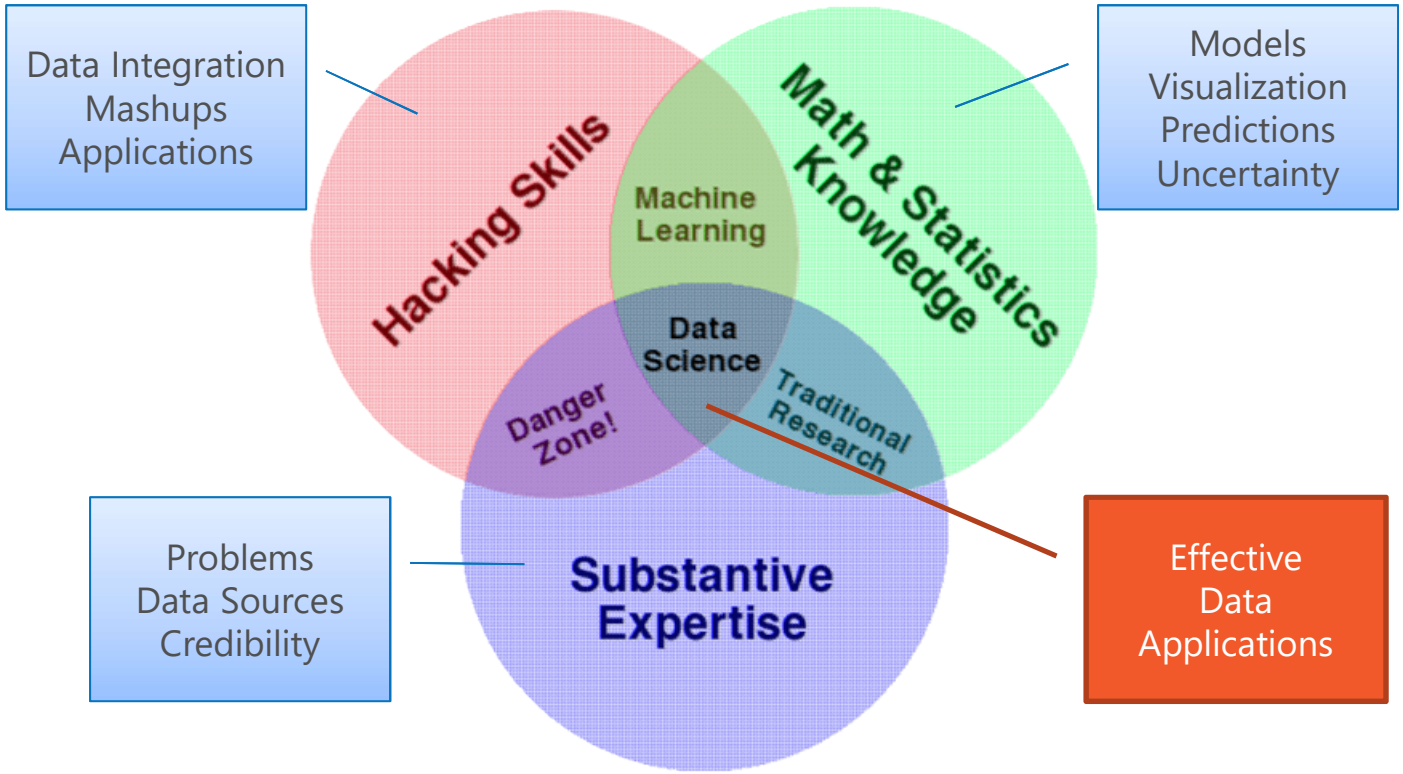
Michael Stravato for The New York Times

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

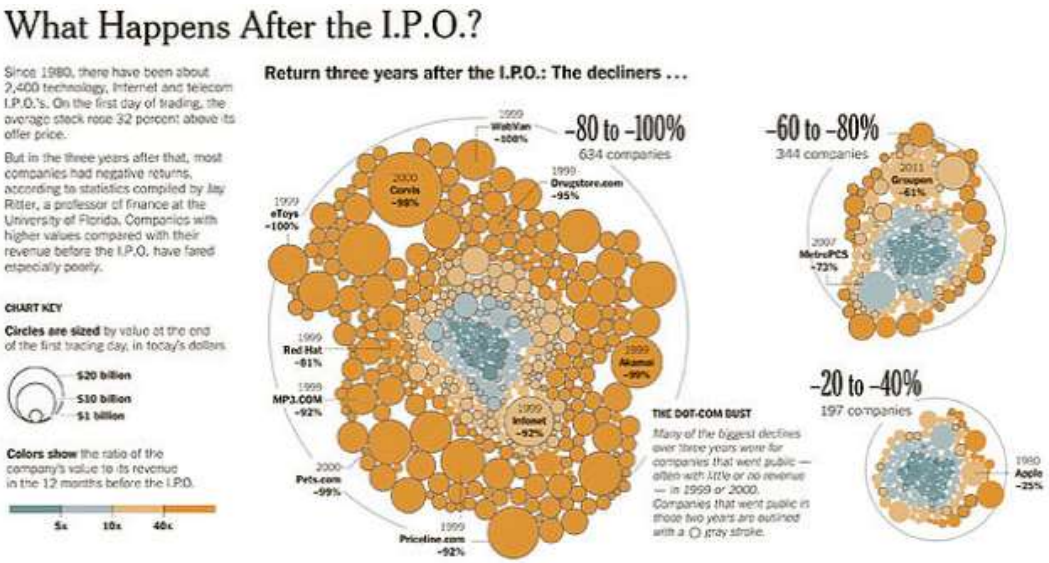
Three Essential Skills of Data Scientists

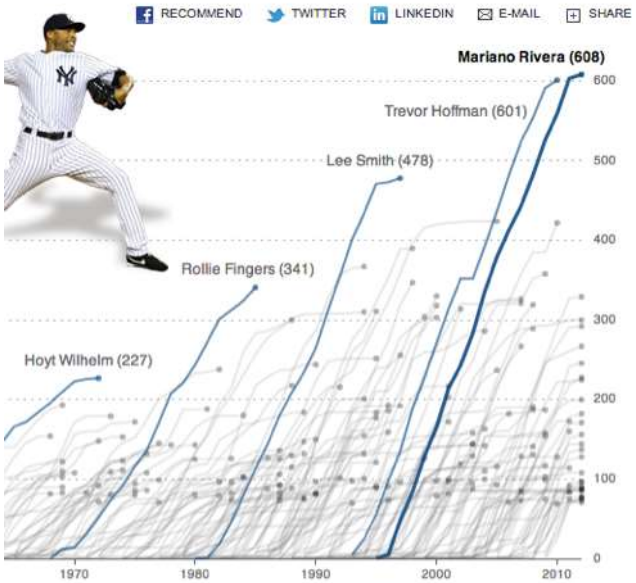


Drew Conway
www.dataists.com/2010/09/the-data-science-venn-diagram/

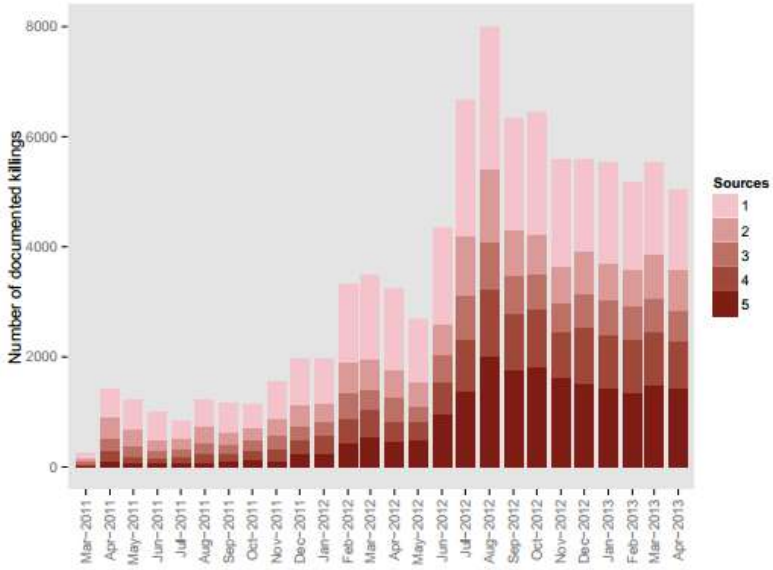
“It resonated with many people. It's not just a pretty picture, it's a reaffirmation of the impact we have in connecting people, even across oceans and borders.” — **Paul Butler, data scientist, Facebook**

The New York Times

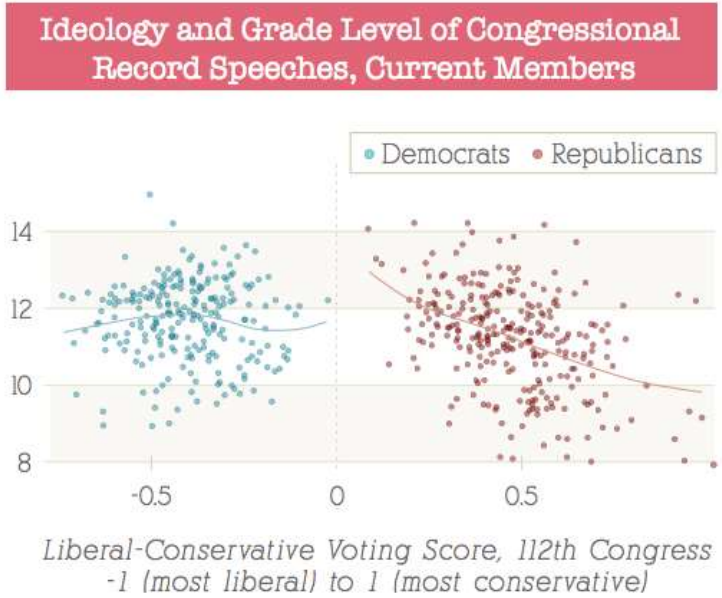




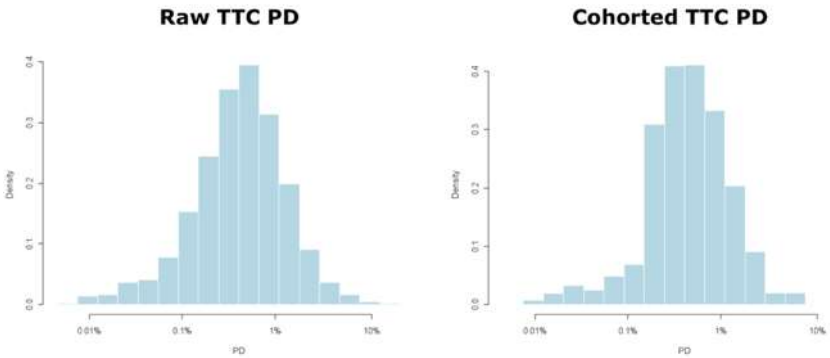
- [Facebook IPO]
- [Baseball legends]



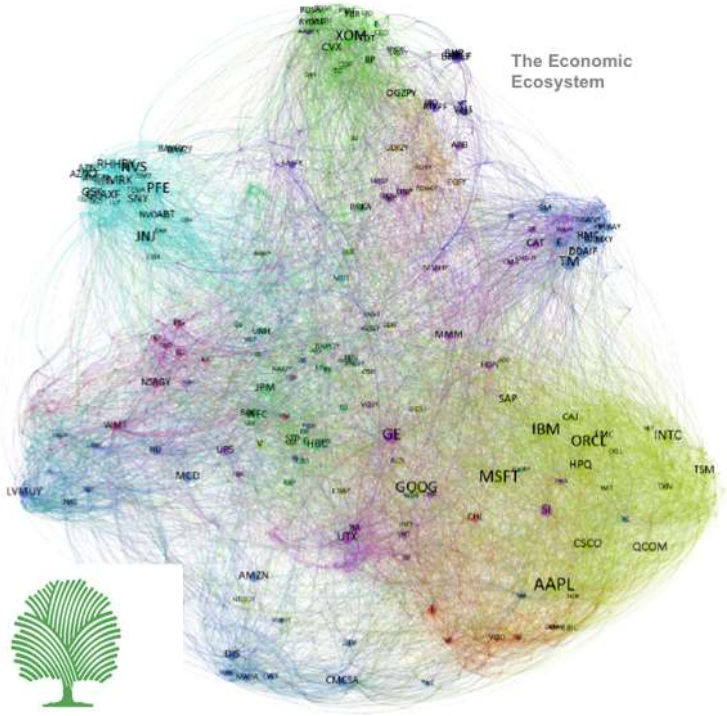
- [Casualty estimation in Warzones]



- [Political Analysis]



- Credit Risk Analysis



AMERICAN CENTURY

- 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 Gentlemen

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

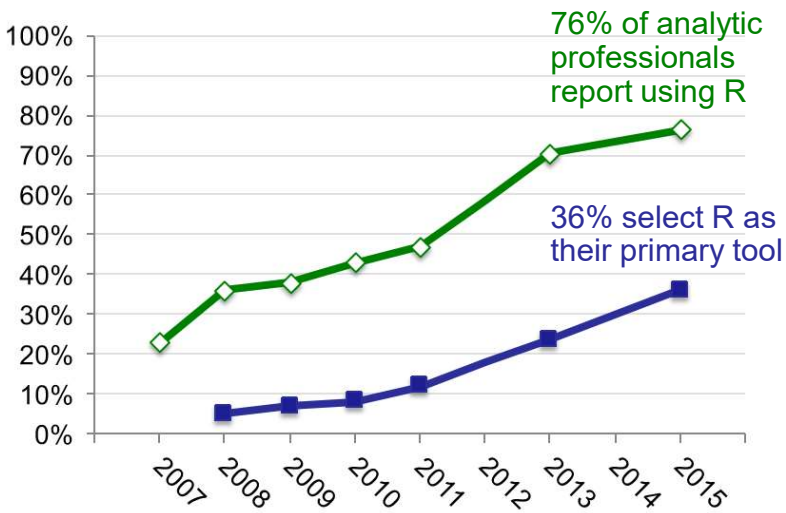
Photo credit: Robert Gentleman

R: The #1 software for Data Science

… and #5 amongst general-purpose programming languages

R Usage Growth

Rexer Data Miner Survey, 2007-2015

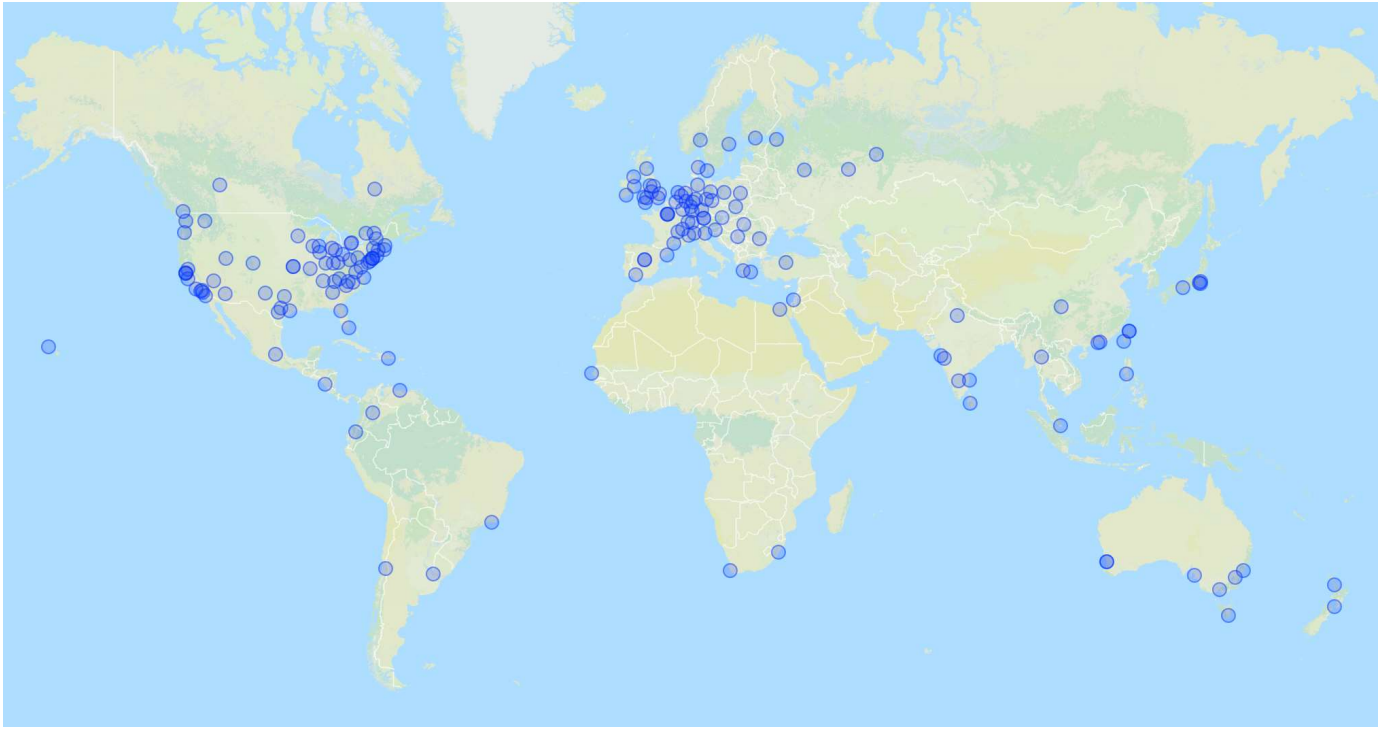


Language Popularity

IEEE Spectrum Top Programming Languages, 2016

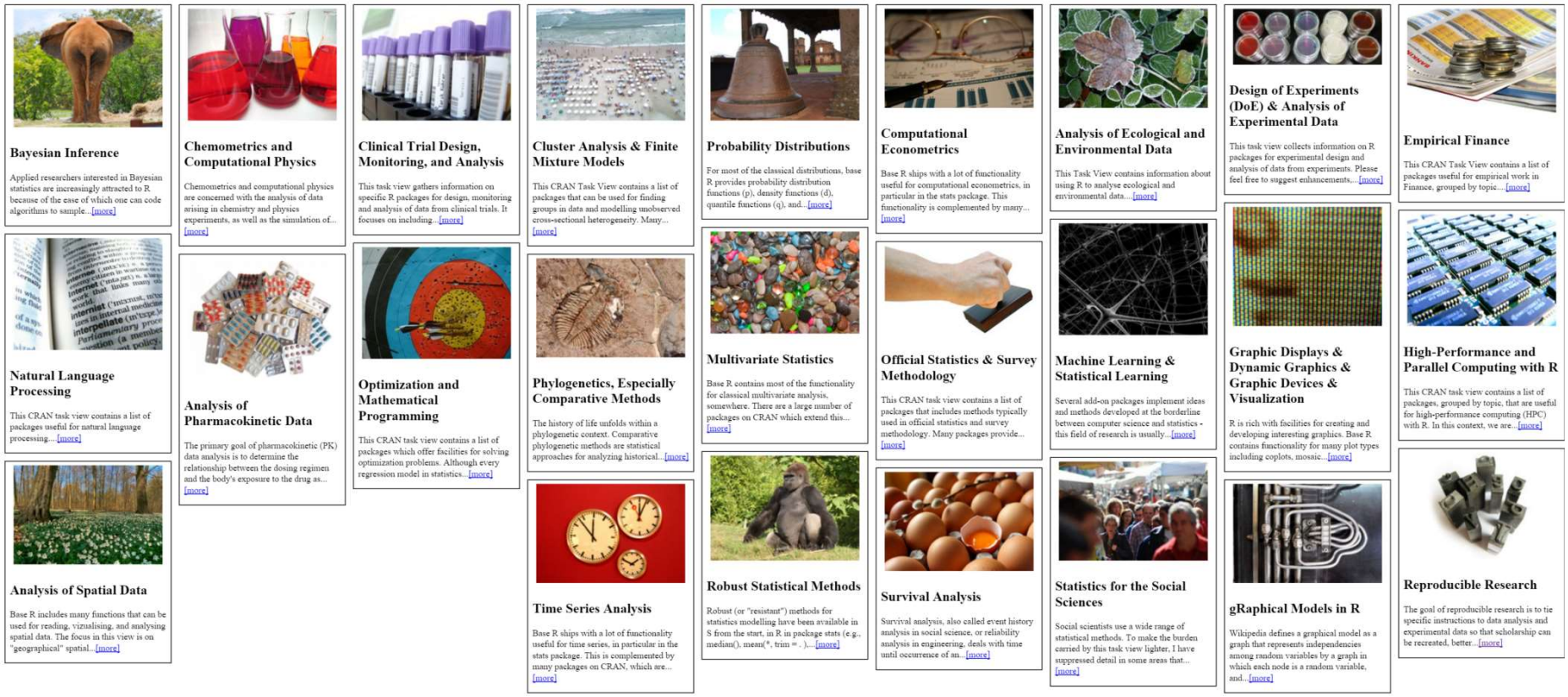
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

200 Local R User Groups Worldwide



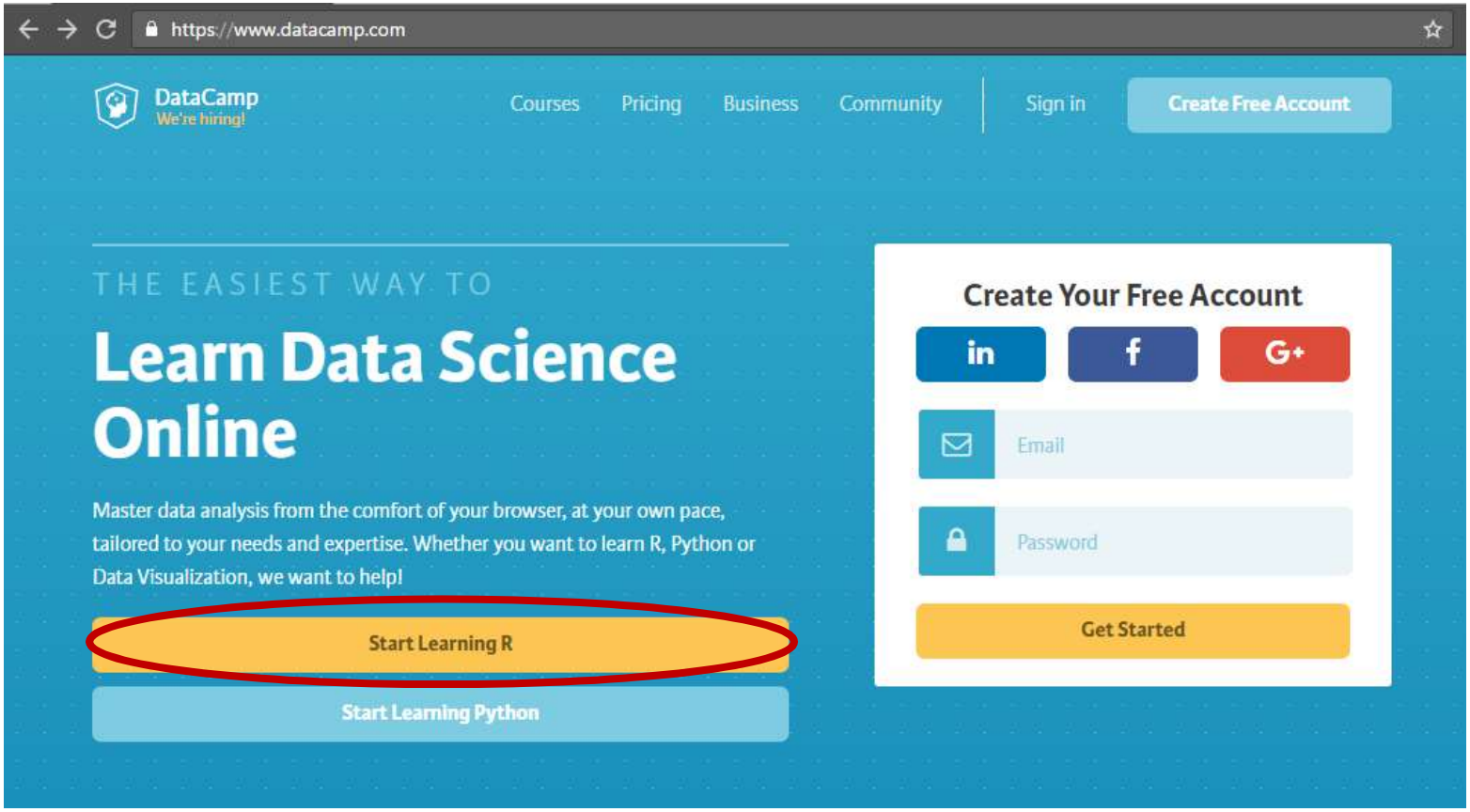
blog.revolutionanalytics.com/local-r-groups.html

CRAN: 9000+ add-on packages for R



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

Introduction to R ww.datacamp.com



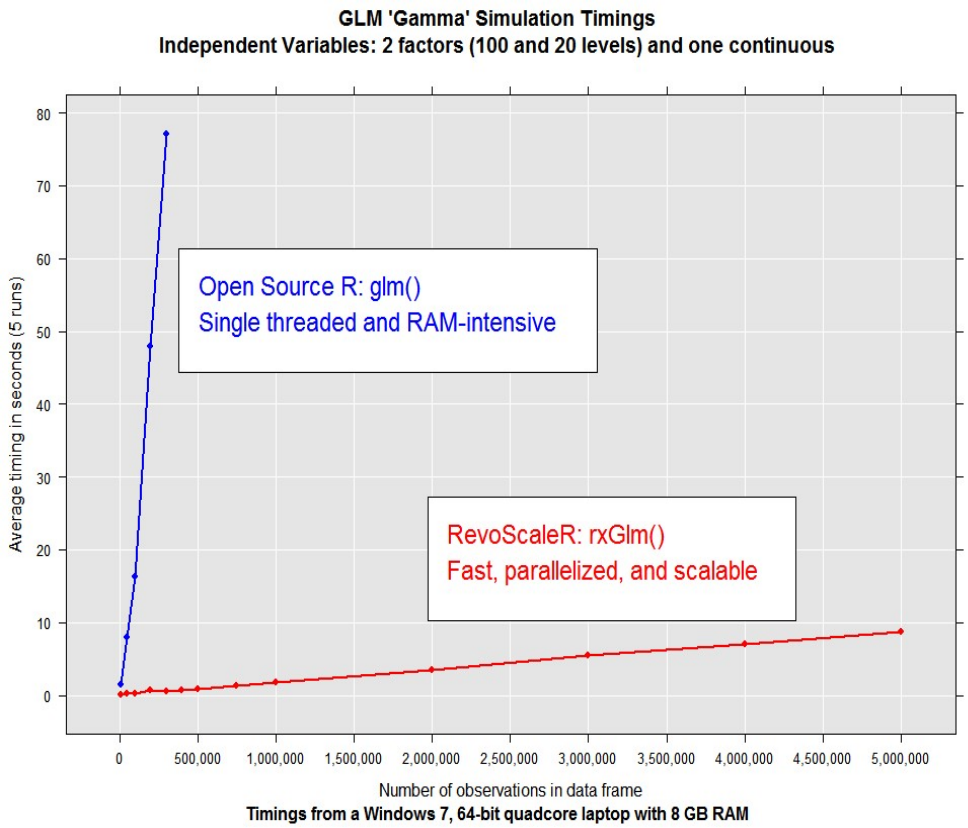
Workshop

Benefits of Microsoft R Server

CRAN, MRO, MRS Comparison

		R Open	R Server
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. 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)
```

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

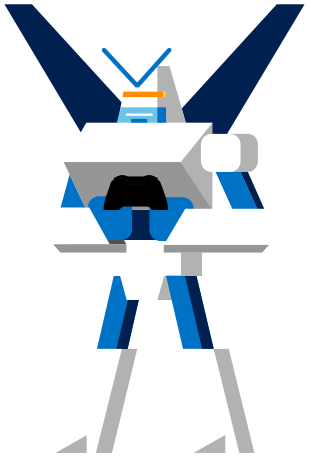
Huge potential

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

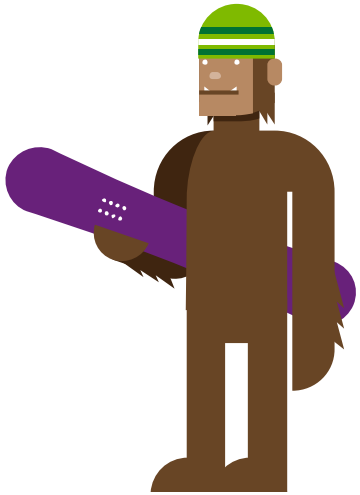
Re-write is difficult

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

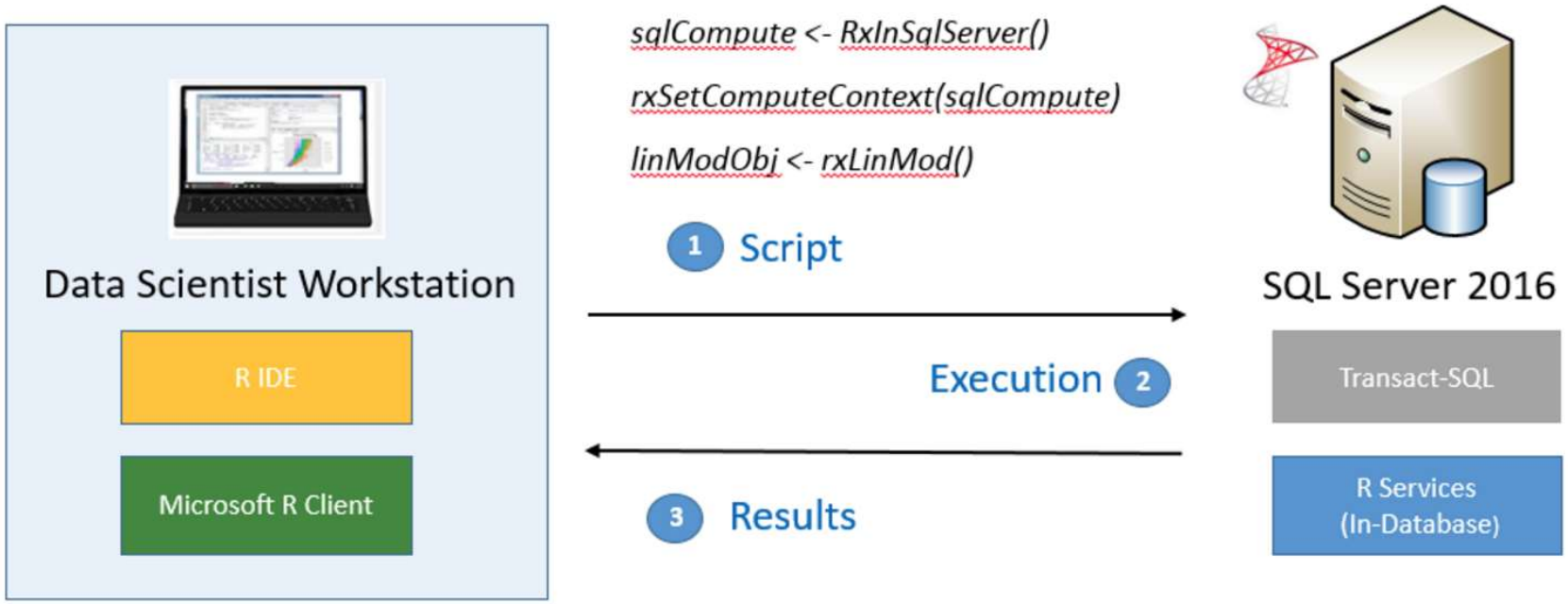
Data fragmentation

Two ways to run R

Two options to execute inside SQL

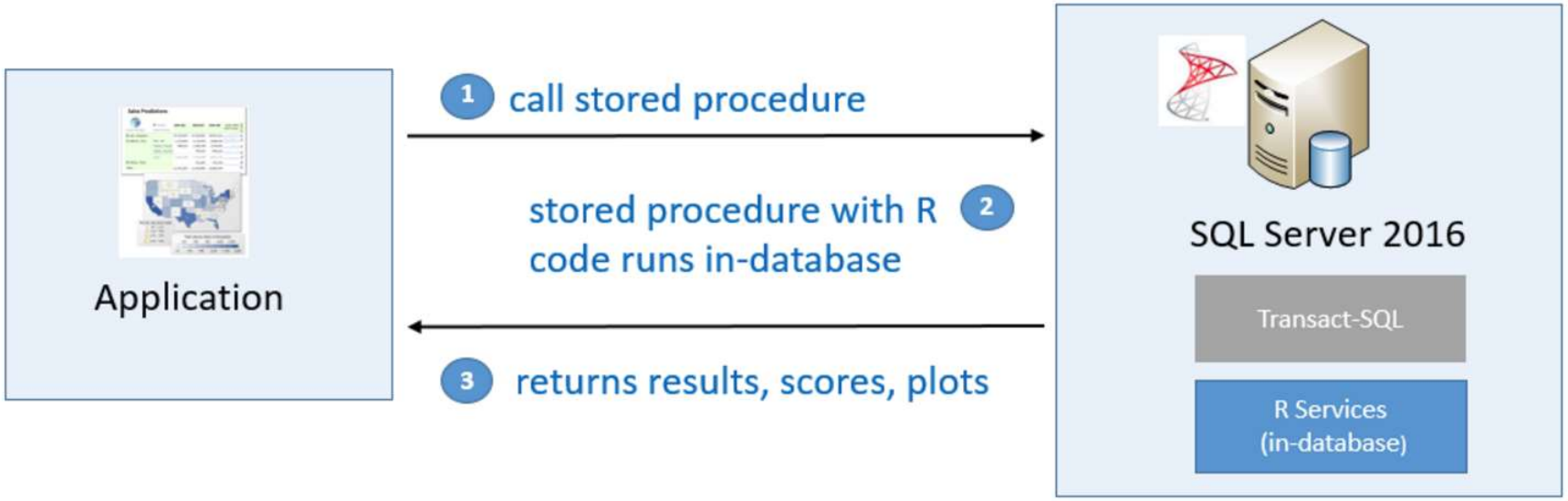
- **Data scientist scenario**: Start from R GUI and execute R code remotely
- **Production scenario**: Start from T-SQL with embedded R script

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


Q & A

Thank You

for attending this session
and
PASS Business Analytics Day!

alteryx ATTUNITY SolidQ