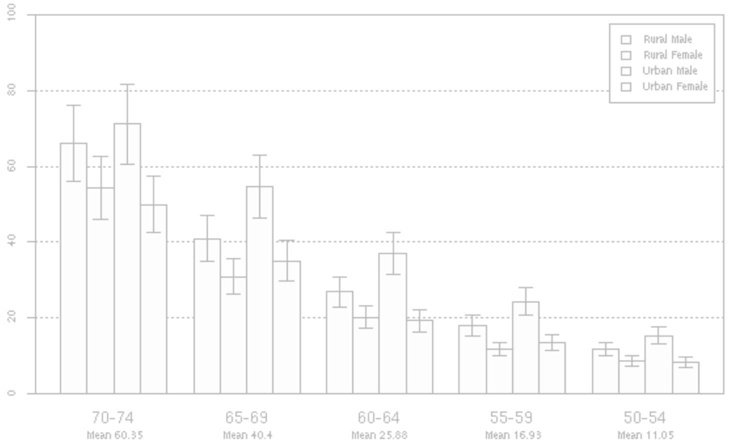
REVOLUTION ANALYTICS

Revolution Analytics

R

and

Data Science

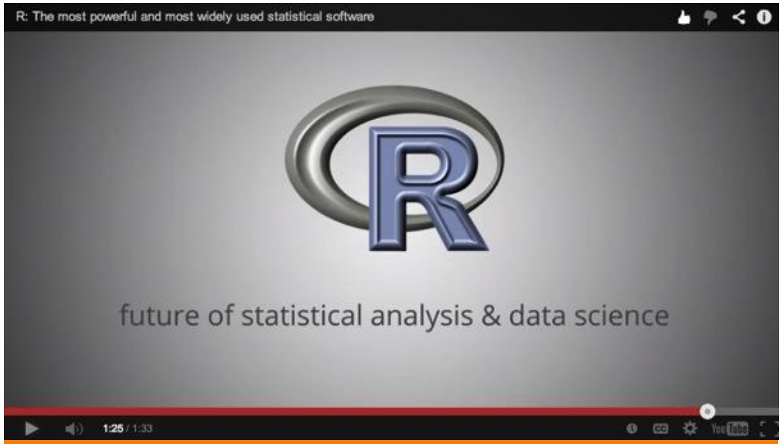


Joseph B Rickert

September 25, 2014

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
- Platform for beautiful and unique data visualizations
  - As seen in New York Times, Twitter and Flowing Data
- Thriving open-source community
  - Leading edge of analytics research


R: The most powerful and most widely used statistical software
future of statistical analysis & data science
1:25 / 1:33

www.revolutionanalytics.com/what-r

REVOLUTION ANALYTICS

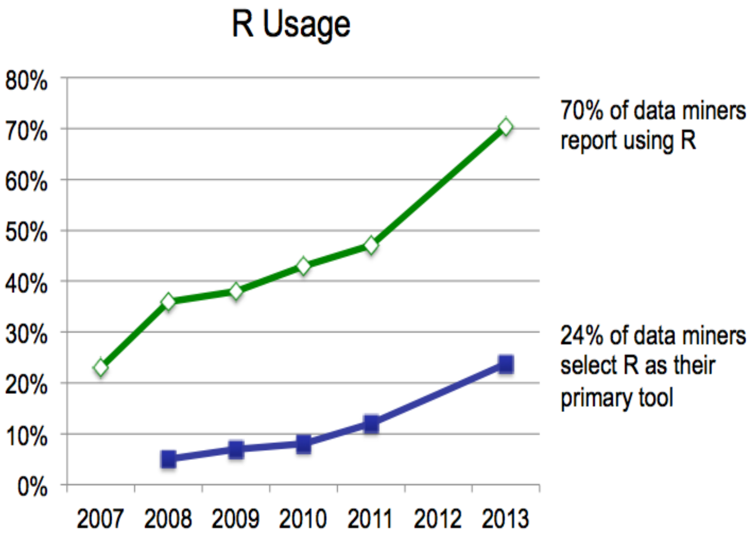
OPEN SOURCE R

UNIVERSIDAD DE CASTILLA - LA MANCHA
PABELLON DE GOBIERNO
The R User Conference 2013

R’s popularity is growing rapidly

R Usage Growth

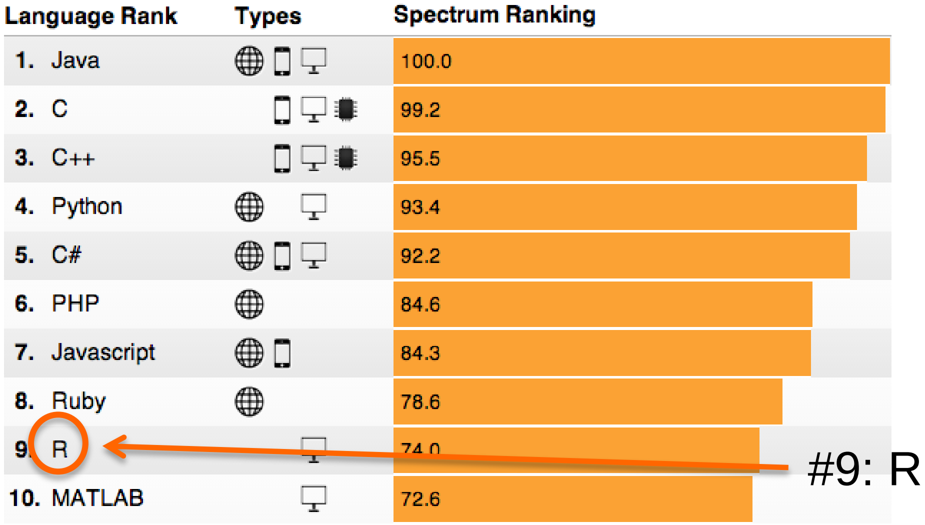
Rexer Data Miner Survey, 2007-2013



- [Rexer Data Miner Survey]

Language Popularity

IEEE Spectrum Top Programming Languages



Language Rank		Types	Spectrum Ranking
1.	Java	🌐 📱 🖥	100.0
2.	C	📱 🖥 ▣	99.2
3.	C++	📱 🖥 ▣	95.5
4.	Python	🌐 🖥	93.4
5.	C#	🌐 📱 🖥	92.2
6.	PHP	🌐	84.6
7.	Javascript	🌐 📱	84.3
8.	Ruby	🌐	78.6
9.	R	🖥	74.0
10.	MATLAB	🖥	72.6

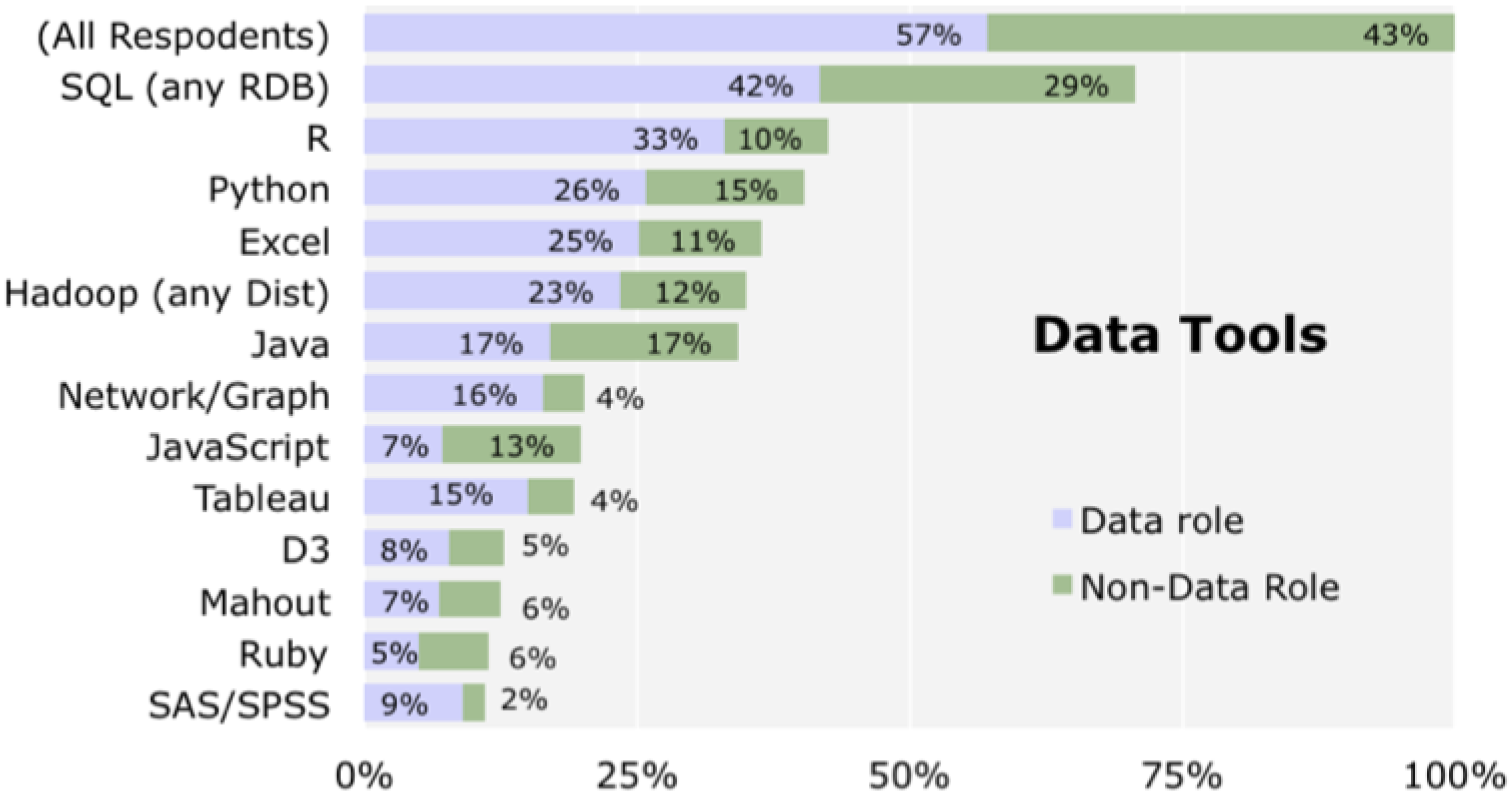
#9: R

- [IEEE Spectrum, July 2014]

Poll Question #1

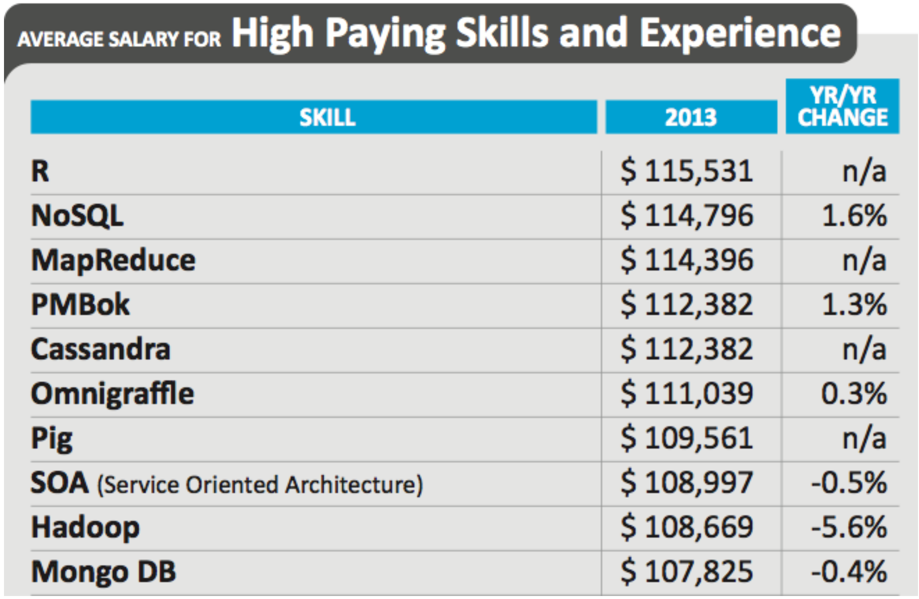
- What are the statistical programming languages/platforms you are most familiar with? (choose all that apply)
  - A) R
  - B) SAS
  - C) SPSS
  - D) KXEN
  - E) Statistica

Tools for Data Science



Source: O'Reilly Data Science Survey

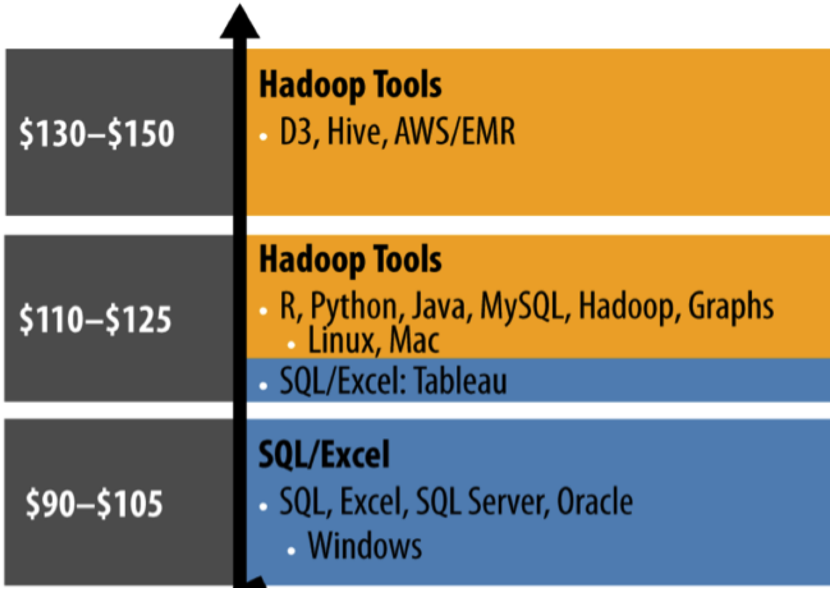
R is among the highest-paid IT skills in the US



AVERAGE SALARY FOR **High Paying Skills and Experience**

SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%

Dice Tech Salary Survey, January 2014



$130–$150

Hadoop Tools

- D3, Hive, AWS/EMR

$110–$125

Hadoop Tools

- R, Python, Java, MySQL, Hadoop, Graphs
  - Linux, Mac
- SQL/Excel: Tableau

$90–$105

SQL/Excel

- SQL, Excel, SQL Server, Oracle
  - Windows

O'Reilly Strata 2013 Data Science Salary Survey

Photo by Ksayer1 on flickr.

Why R for Data Science?

Algorithms

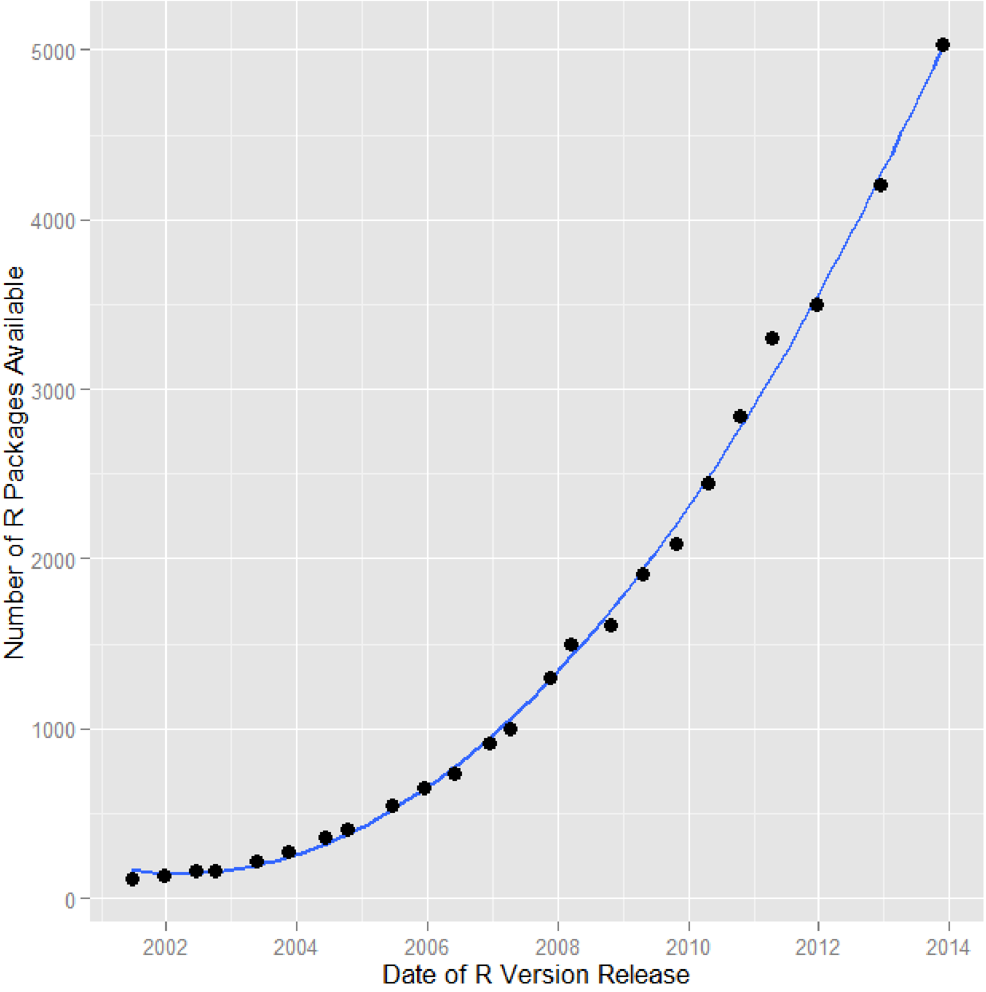
```
X <- if (!is.empty.model(mt))
        model.matrix(mt, mf, contrasts)
    else matrix(, NROW(Y), 0L)
    weights <- as.vector(model.weights(mf))
    if (!is.null(weights) && !is.numeric(weights))
        stop("'weights' must be a numeric vector")
    if (!is.null(weights) && any(weights < 0))
        stop("negative weights not allowed")
    offset <- as.vector(model.offset(mf))
    if (!is.null(offset)) {
        if (length(offset) != NROW(Y))
            stop(gettextf("number of offsets is %d should equal %d (number of observations)",
                length(offset), NROW(Y)), domain = NA)
    }
    mustart <- model.extract(mf, "mustart")
    etastart <- model.extract(mf, "etastart")
    fit <- eval(call(if (is.function(method)) "method" else method,
        x = X, y = Y, weights = weights, start = start, etastart = etastart,
        mustart = mustart, offset = offset, family = family,
        control = control, intercept = attr(mt, "intercept") >
            0L))
    if (length(offset) && attr(mt, "intercept") > 0L) {
        fit2 <- eval(call(if (is.function(method)) "method" else method,
            x = X[, "(Intercept)", drop = FALSE], y = Y, weights = weights,
            offset = offset, family = family, control = control,
            intercept = TRUE))
        if (!fit2$converged)
            warning("fitting to calculate the null deviance did not converge -- increase 'maxit'?")
        fit$null.deviance <- fit2$deviance
    }
    if (model)
        fit$model <- mf
    fit$na.action <- attr(mf, "na.action")
    if (x)
        fit$x <- X
    if (!y)
        fit$y <- NULL
    fit <- c(fit, list(call = call, formula = formula, terms = mt,
        data = data, offset = offset, control = control, method = method,
        contrasts = attr(X, "contrasts"), xlevels = .getXlevels(mt,
            mf)))
    class(fit) <- c(fit$class, c("glm", "lm"))
    fit
```

Task Views

R Growth

Put this astonishing growth in perspective:

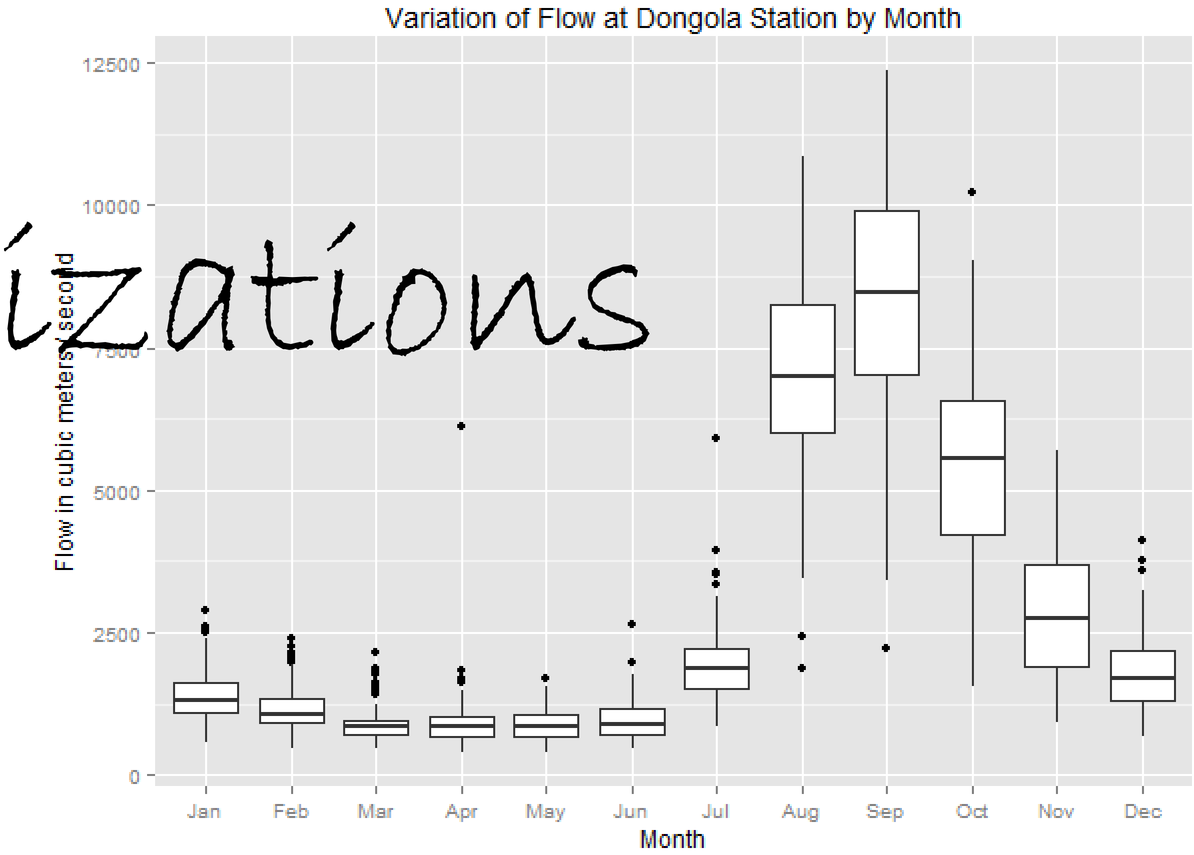
- SAS.V 9.3S contains ~ 1,200 commands that are roughly equivalent to R functions
- R packages contain a median of 5 functions
- Therefore R has ~ 36,820 functions
- *During 2013 alone, R added more functions than SAS Institute has written in its entire history!*

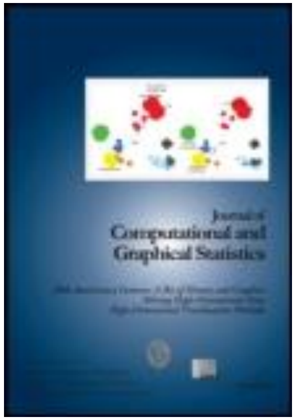


Bob Muenchen

Why R for Data Science?

Visualizations





Journal of Computational and Graphical Statistics

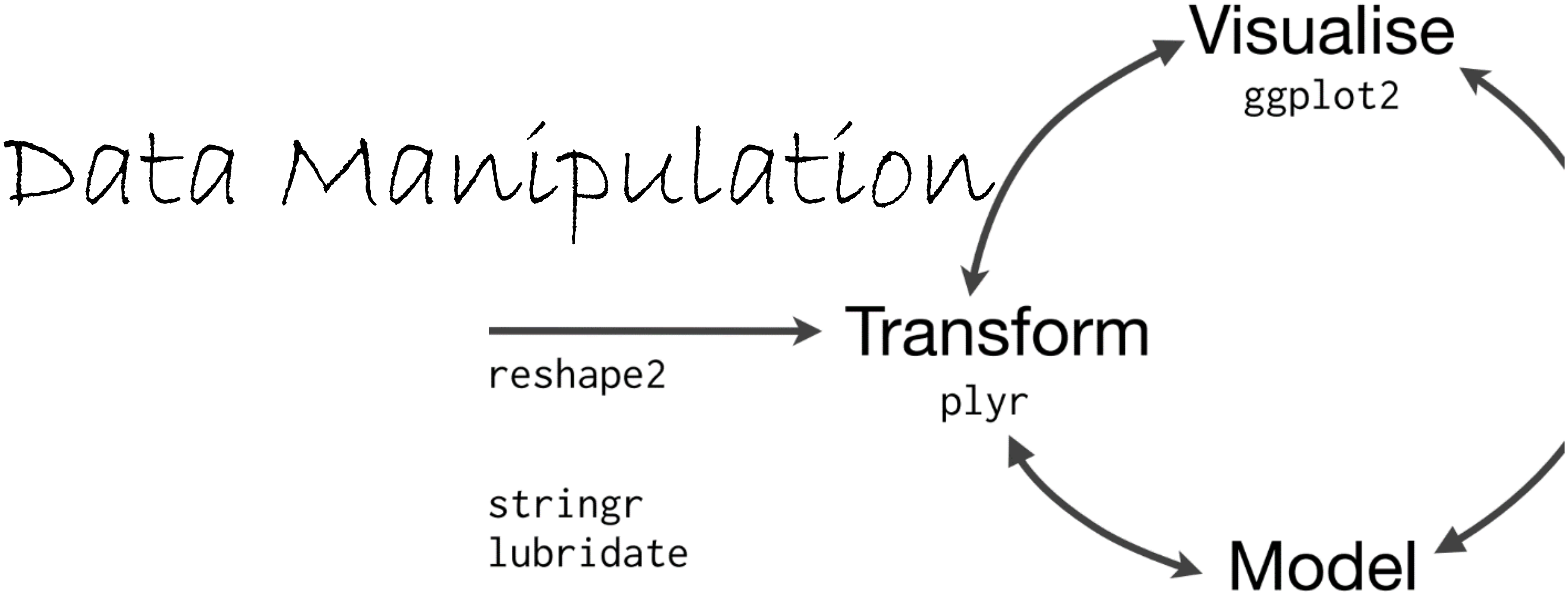
Why R for Data Science?

Programming

- Scripting
- Functional programming
- Parallel programming
- Data structures
- Objects
- Data Types
- Regular expressions
- Data connections
- Interfaces to other languages

REVOLUTION ANALYTICS

Why R for Data Science?



Data Manipulation

Visualise

ggplot2

reshape2

Transform

plyr

stringr
lubridate

Model

“It's often said that 80% of the effort of analysis is spent just getting the data ready to analyse, the process of data cleaning. Data cleaning is not only a vital first step, but it is often repeated multiple times over the course of an analysis as new problems come to light.” Hadley Wickham Tidy Data

Why R for Data Science?

R Integrates

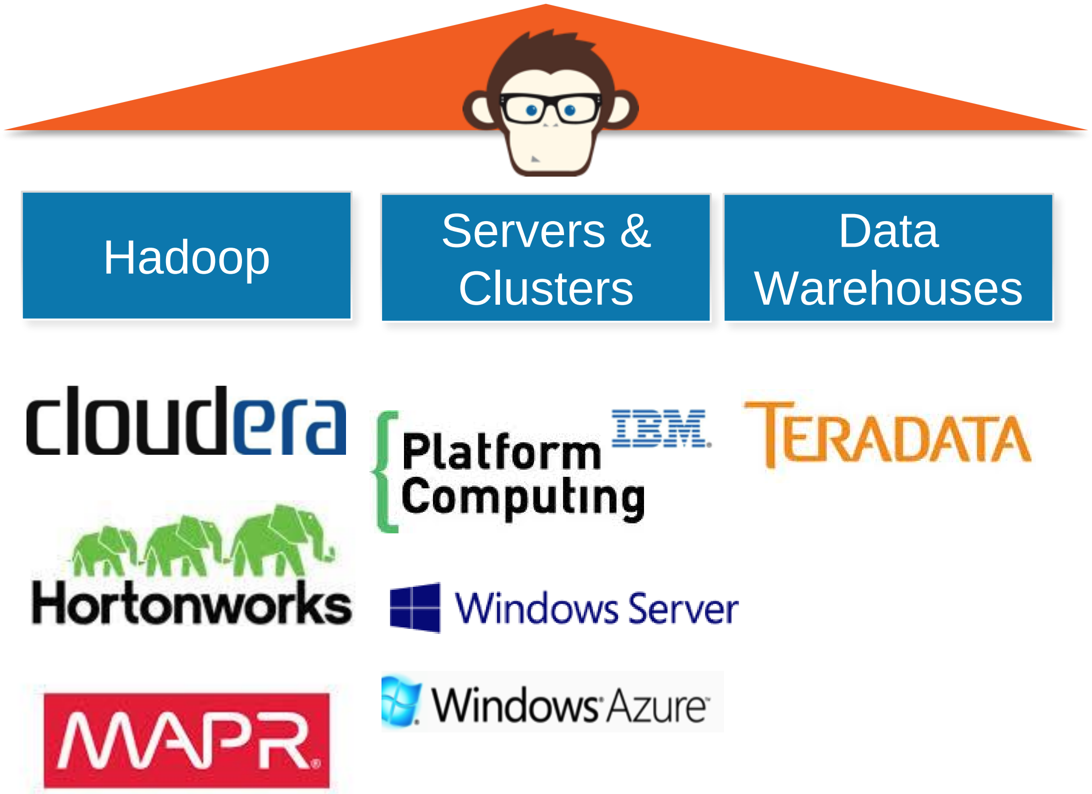
- Web applications
- Internet graphics
  - D3
  - Potly
- Other Languages
  - C, C++
  - Java
- BI Tools
- Data bases
  - SQL
  - MongoDB

Poll Question #2

- What are the data platforms that you are connecting to regularly? (choose all that apply)
  - A) Hadoop
  - B) Spark
  - C) Cloud-based (Azure/AWS/Google)
  - D) Data Warehouses
  - E) Servers (Grid or Cluster)

Why R for Data Science

R Scales


Hadoop

Servers & Clusters

Data Warehouses

Poll Question #3

- What are the types of models that you are working with most? (choose all that apply)
  - A) Linear models / Regression / GLM
  - B) Decision Trees / Random Forests
  - C) Survival Models
  - D) GBM
  - E) Time Series models

Let’s look at some code.

www.revolutionanalytics.com
1.855.GET.REVO
Twitter: @RevolutionR

Why is R Right for Data Science?

- R is open source
- R is a powerful language
  - Data Manipulation
  - Computational Statistics
  - Machine Learning
- R is an innovation engine
- R has a rich and expanding ecosystem

Q&A / Resources

R Code and Markdown Files
https://github.com/joseph-rickert/DataScienceRWebinar

What is R?
`revolutionanalytics.com/what-is-r`

Companies using R
`revolutionanalytics.com/companies-using-r`

AcademyR training
`revolutionanalytics.com/AcademyR`

AcademyR Certification
`revolutionanalytics.com/AcademyR-certification`

Contact Revolution Analytics
`revolutionanalytics.com/contact-us`

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com, 1.855.GET.REVO, Twitter: @RevolutionR