Next steps

Introduction to Data Science with R

www.therbootcamp.com

@therbootcamp

October 2018

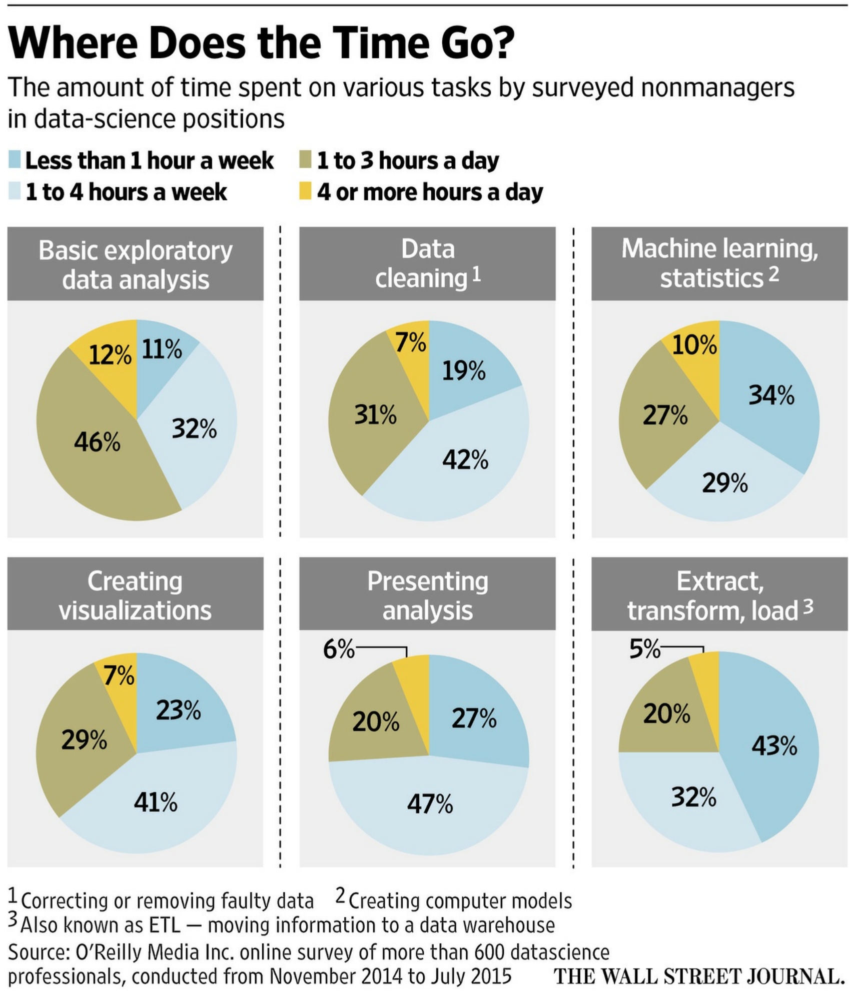
Hello Data Scientist!

In 2 days, 6 sessions, and 16 hours you have come a long way.

	Sat, 6 Oct	Sun, 7 Oct
0900	Welcome	Recap
0930	**Intro to R** +Interactive	**Analysing** +Practical
1200	*Lunch*	*Lunch*
1300	**Data** +Practical	**Plotting** +Practical
1530	**Wrangling** +Practical	**Case studies** +Practical
1630		Next steps
1800	Wrapup	*Apero*

Data Scientist

> Harvard Business Review

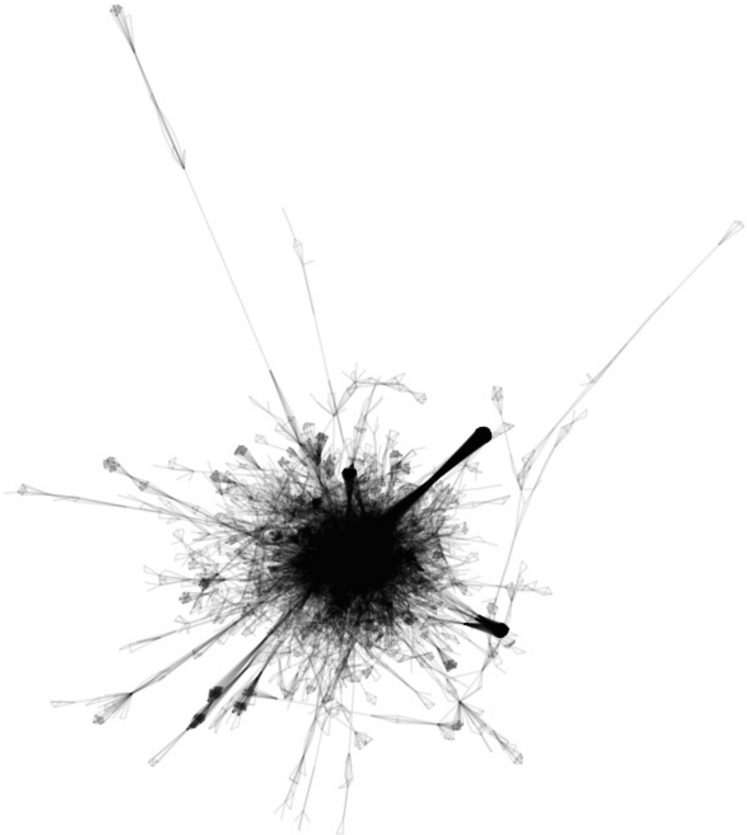


Wall Street Journal

Next steps

1. Networks
2. Statistics
3. Machine learning
4. Text analysis
5. Rcpp
6. Forms
7. R on servers

8. Books
9. Websites R
10. Groups

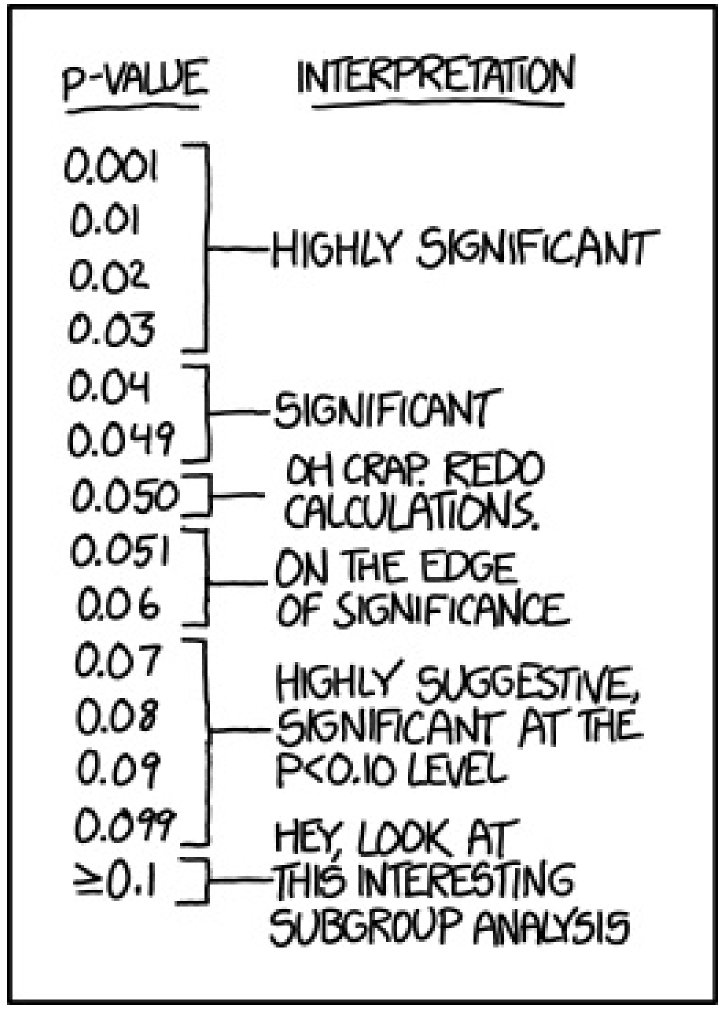


source R-bloggers.com

Stats packages

> "It's easy to lie with statistics; it is easier to lie without them."
>
> Frederick Mosteller

Package	Description
`stats`	Linear, generalized linear models, individual tests, and distributions.
`lme4`, `afex`	Mixed-mode, hierarchical regression.
`sem`, `lavaan`, `OpenMx`	Structural equation modeling.
`survival`	Survival analysis.

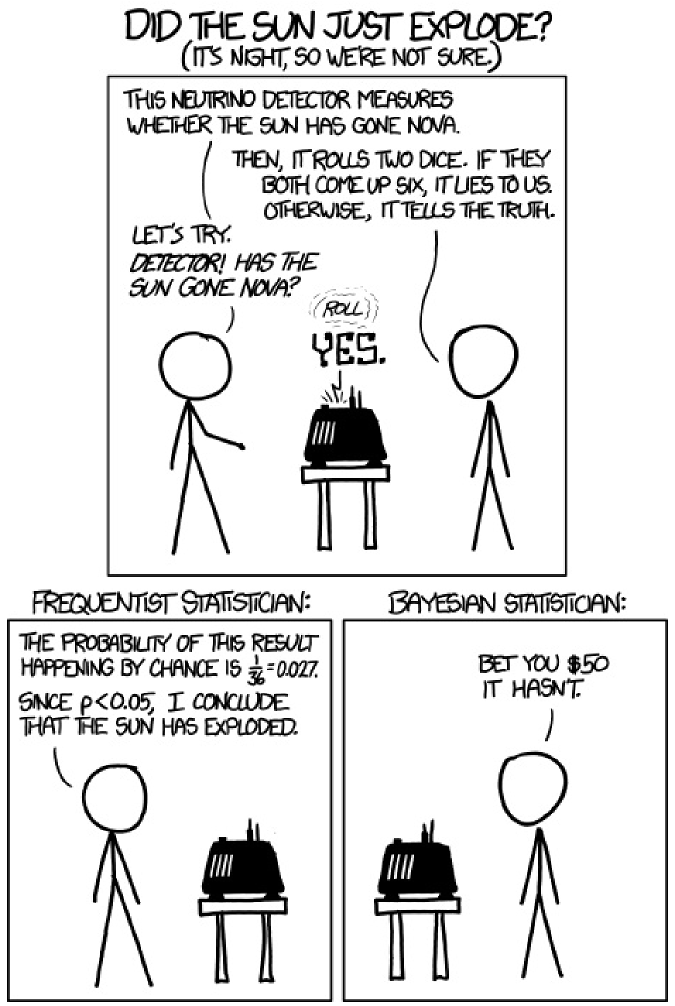

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

xkcd.com

Bayesian statistics

> The subjectivist (i.e. Bayesian) states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science.
>
> I. J. Good

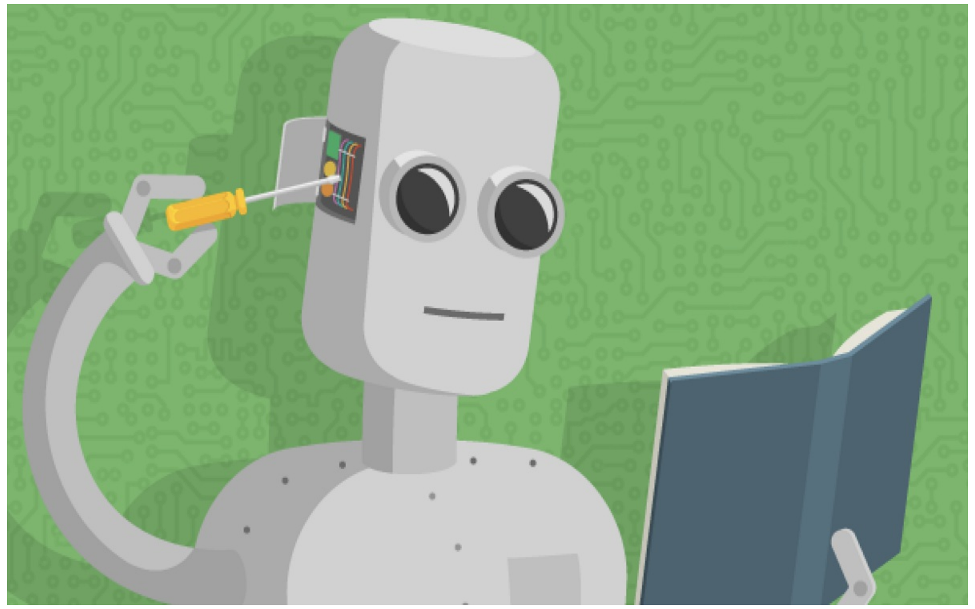
Package	Description
`BayesFactor`, `rstanarm`	Bayesian linear models. As easy as non-Bayesian methods.
`rjags`, `rstan`	Build flexible, hierarchical Bayesian models.
`mcmc`	Metropolis algorithms.
`bridgesampling`	Estimating marginal likelihoods using bridgesampling.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

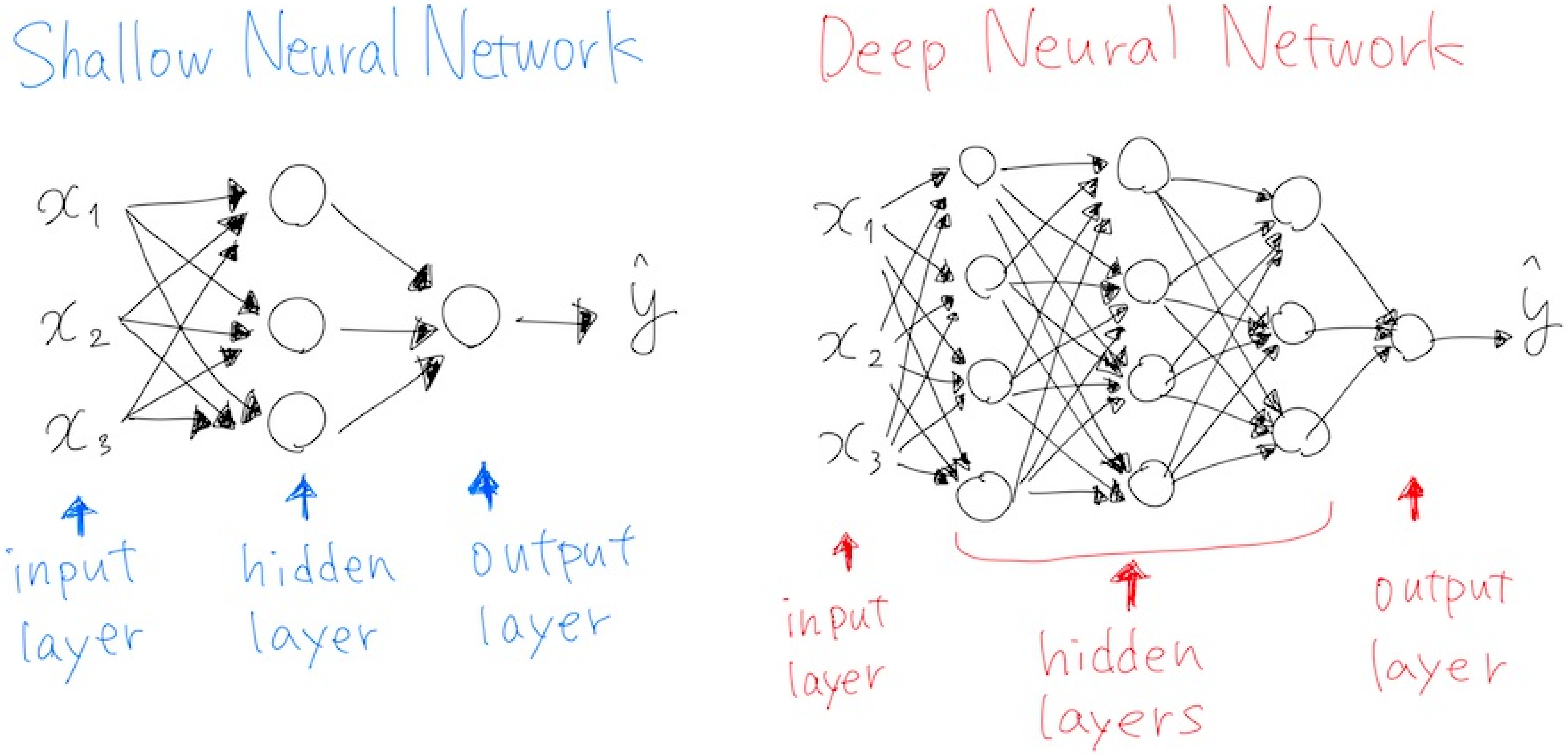
xkcd.com

Machine learning

Package	Description
`caret`	Umbrella package for diverse machine learning algorithms.
`mlr`, `e1071`, etc.	Other umbrella packages.
`randomForest`, `rpart`, `FFTrees`	Decision trees.
`cluster`, `fastcluster`, `cstab`, etc.	Cluster analysis.
`forecast`, `mgm`, `timeSeries`, etc.	Time series models.
`tensorflow`	Time series models.

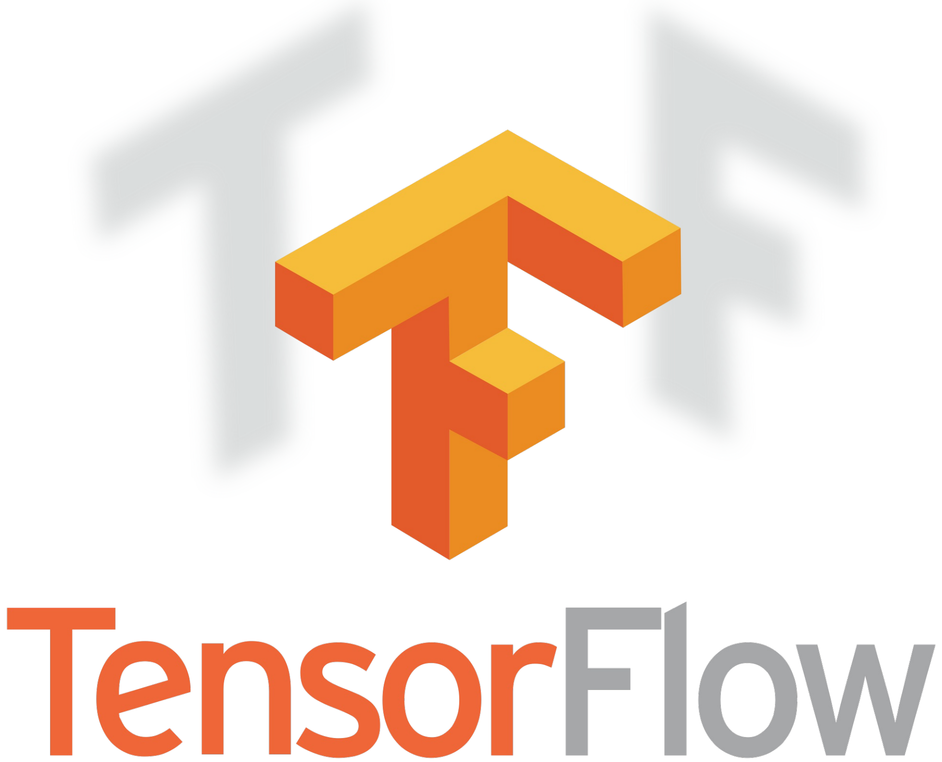


Deep learning

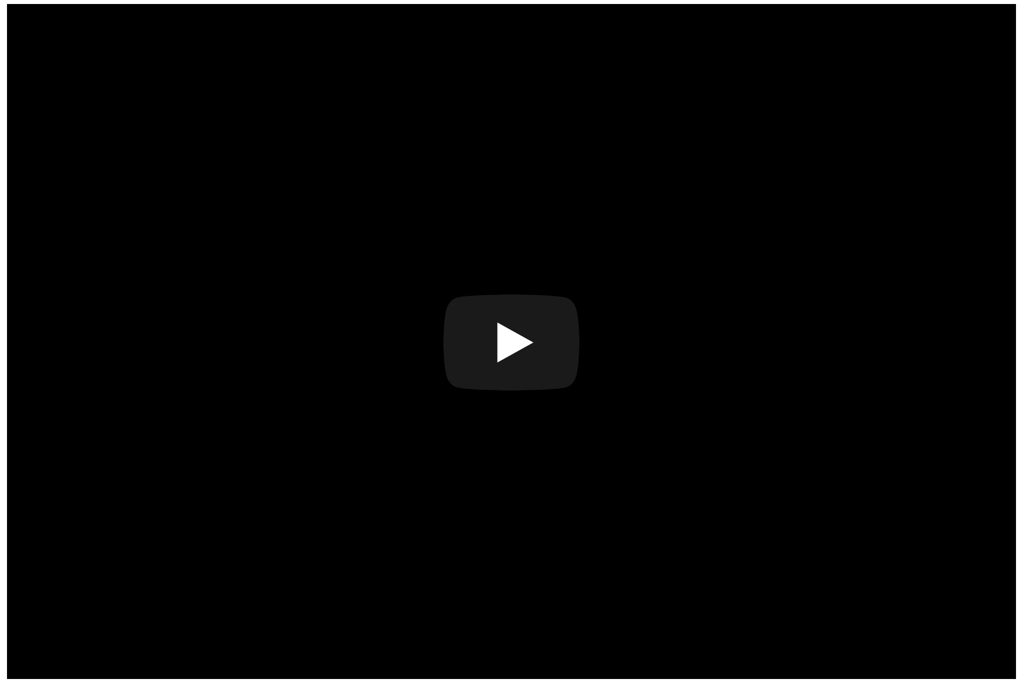


Deep learning: Tensorflow

```r
# Get Keras
install.packages(keras)
library(keras)
install_keras()

# Define model
model <- keras_model_sequential()
model %>%
  layer_dense(units = 256,
              activation = 'relu',
              input_shape = c(784)) %>%
  layer_dropout(rate = 0.4) %>%
  layer_dense(units = 128,
              activation = 'relu') %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 10,
              activation = 'softmax')
```



Text analysis

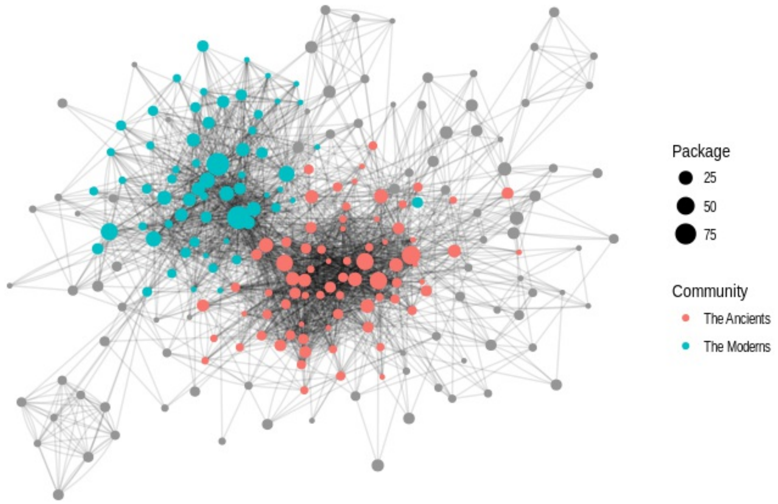


Sundar Pichai @ Google IO, May 2018

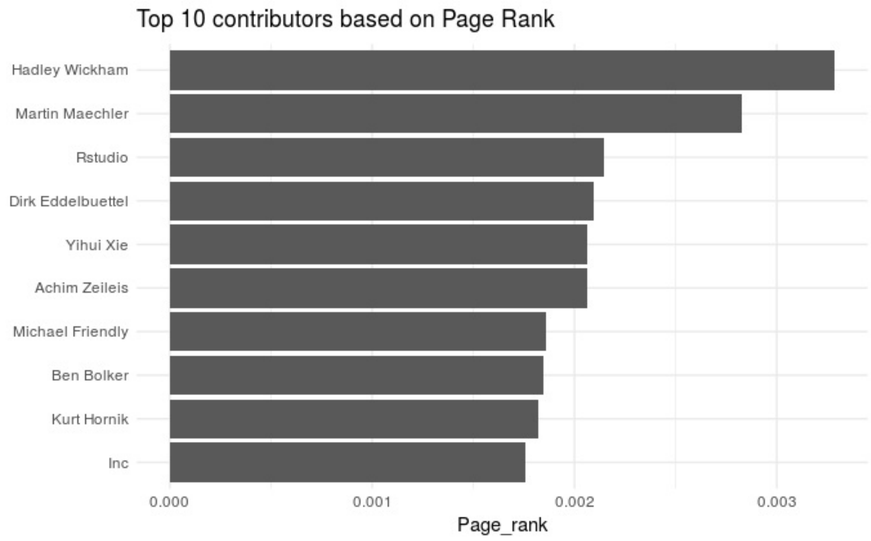
Package	Description
tm, tidytext	General text analysis packages
stringr, stringi	String operations and regular expressions.
rvest, XML	Scraping content of the internet
text2vec	Vector representation of words.
SentimentAnalysis	Sentiment analysis.
twitteR, streamR, jsonlite	Streaming and parsing tweets.
Rfacebook	Access to Facebook API.

Network analysis

A **social graph** of package Co-authors using `tidyverse` plus `ggraph`, an extension for `ggplot2` for graphs (aka networks) and `igraph`, an extremely powerful library for network analysis. Find the code and additional explanations **here**.



source **R-bloggers.com**

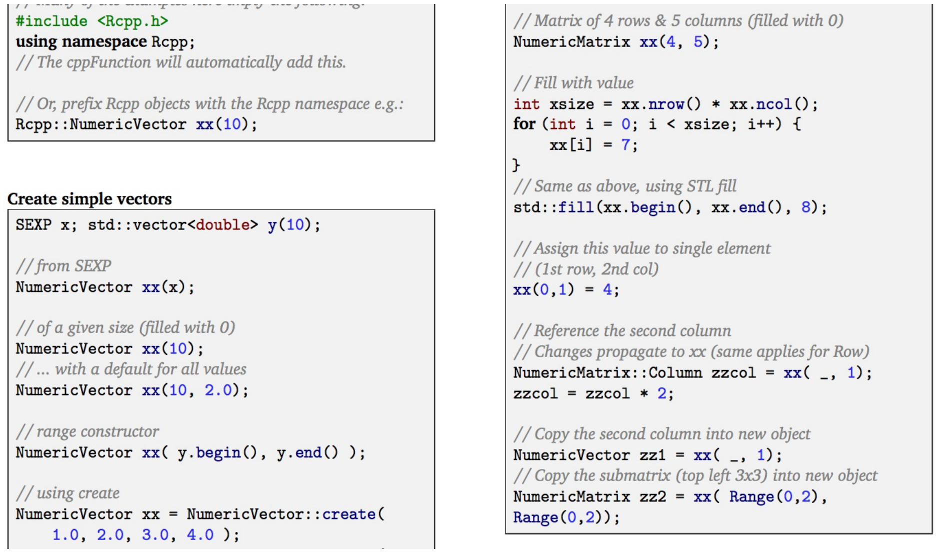


source **R-bloggers.com**

Rcpp

By now one of the most referenced R packages is Rcpp - **R's interface to C++**. With often relatively little effort due to **Rcpp sugar**, Rcpp can provide vast speed improvements, which many packages today rely on Rcpp in the background for **swift code execution**. Rcpp becomes particularly powerful, when supplemented with BH, which makes avaialble a collection **free, peer-reviewed C++ libraries**, and **RcppArmdillo**, which available the high-performance **Armadillo** library for linear algebra methods.

source **classic105.com**



```
#include <Rcpp.h>
using namespace Rcpp;
// The cppFunction will automatically add this.

// Or, prefix Rcpp objects with the Rcpp namespace e.g.:
Rcpp::NumericVector xx(10);
```

Create simple vectors

```
SEXP x; std::vector<double> y(10);

// from SEXP
NumericVector xx(x);

// of a given size (filled with 0)
NumericVector xx(10);
// ... with a default for all values
NumericVector xx(10, 2.0);

// range constructor
NumericVector xx( y.begin(), y.end() );

// using create
NumericVector xx = NumericVector::create(
    1.0, 2.0, 3.0, 4.0 );
```

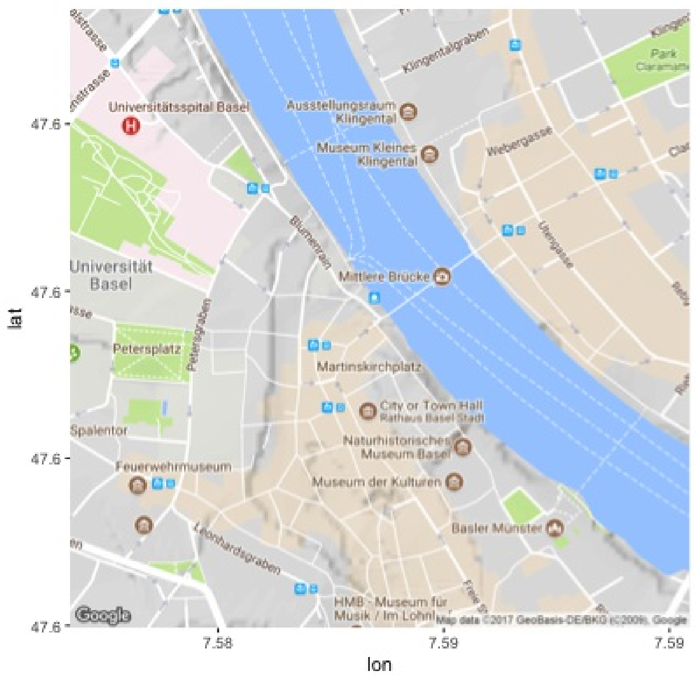
```
// Matrix of 4 rows & 5 columns (filled with 0)
NumericMatrix xx(4, 5);

// Fill with value
int xsize = xx.nrow() * xx.ncol();
for (int i = 0; i < xsize; i++) {
    xx[i] = 7;
}
// Same as above, using STL fill
std::fill(xx.begin(), xx.end(), 8);

// Assign this value to single element
// (1st row, 2nd col)
xx(0,1) = 4;

// Reference the second column
// Changes propagate to xx (same applies for Row)
NumericMatrix::Column zzcol = xx( _, 1);
zzcol = zzcol * 2;

// Copy the second column into new object
NumericVector zz1 = xx( _, 1);
// Copy the submatrix (top left 3x3) into new object
NumericMatrix zz2 = xx( Range(0,2),
Range(0,2));
```

Quick Reference Guide

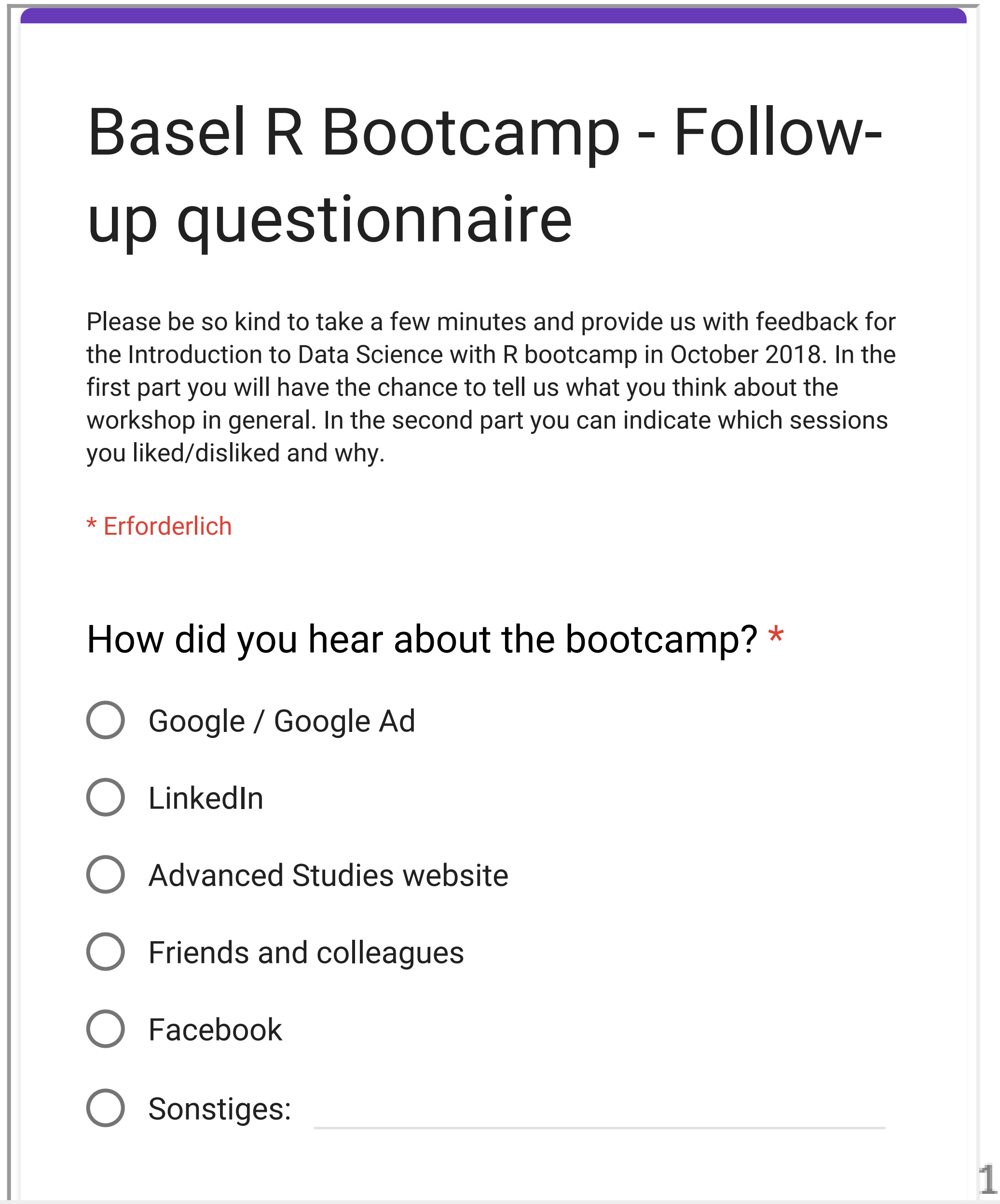
Google Forms & Maps

New packages also allow you to interact with **Google Maps** and **Google Forms**. Use *ggmap* to access Google Maps and `googlesheets` to access Google Forms.

```
library(ggmap)
ggmap(get_map(c(7.588576, 47.559601),zoom=16))
```



lat
lon
47.6
47.6
47.6
47.6
7.58
7.59
7.59
Universitätsspital Basel
Ausstellungsraum Klingental
Museum Kleines Klingental
Universität Basel
Mittlere Brücke
Petersplatz
Martinskirchplatz
City or Town Hall Rathaus Basel Stadt
Spalentor
Naturhistorisches Museum Basel
Feuerwehrmuseum
Museum der Kulturen
Basler Münster
HMB - Museum für Musik / Im Lohnhof
Google
Map data ©2017 GeoBasis-DE/BKG (©2009), Google



Basel R Bootcamp - Follow-up questionnaire

Please be so kind to take a few minutes and provide us with feedback for the Introduction to Data Science with R bootcamp in October 2018. In the first part you will have the chance to tell us what you think about the workshop in general. In the second part you can indicate which sessions you liked/disliked and why.

* Erforderlich

How did you hear about the bootcamp? *

○ Google / Google Ad

○ LinkedIn

○ Advanced Studies website

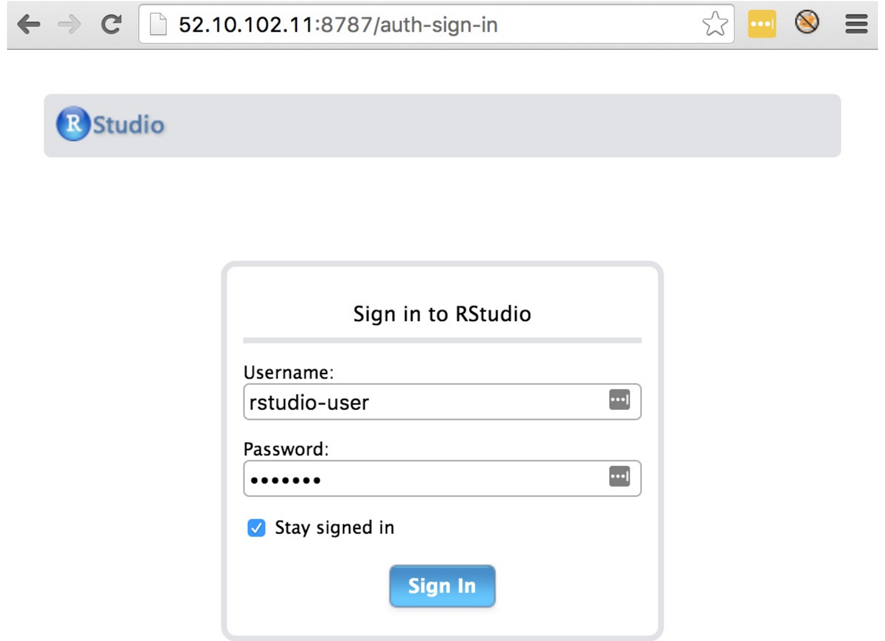
○ Friends and colleagues

○ Facebook

○ Sonstiges: ____________

R on servers

With the need for more computing power for, e.g., machine learning, R is also moving more and more to the **server and cluster sphere**. Rstudio offers a convenient (and free) solution to access **RStudio** sessions on a **remote linux server** via a browser. RStdudio also offers the `sparklyr` package integrating the **Spark** architecture for efficient cluster computing with `dplyr`. Other packages exist to interface with **Hadoop** (RHadoop) and **SQL** (`RSQLite`).



52.10.102.11:8787/auth-sign-in

RStudio

Sign in to RStudio

Username: rstudio-user

Password: •••••••

☑ Stay signed in

Remote databases

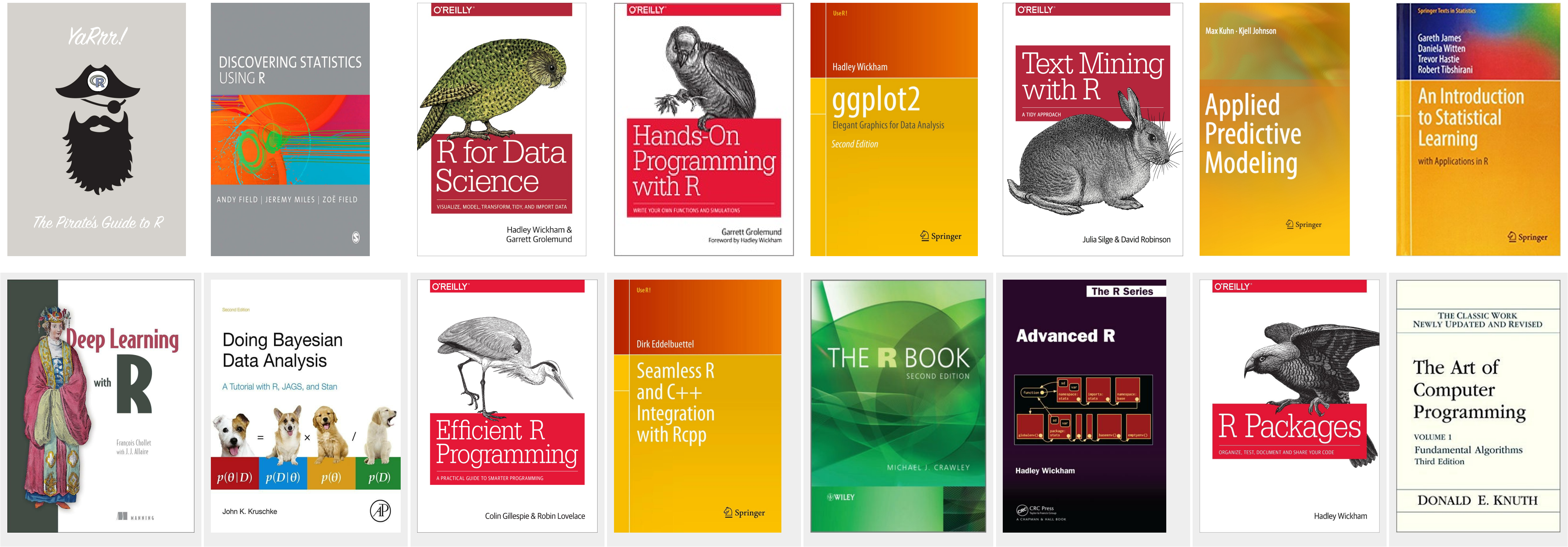
R provides **all necessary tools to pull data from or directly work with** remote databases such as, e.g., a SQL database. Find out more at:

db.rstudio.com

How to continue

Books

Here is a very incomplete series of good books. They are ordered by complexity, beginning with user-friendly books on **learning statistics** in R and ending with books focusing on the more **advanced topics of the R language**.


YaRrr!
The Pirates Guide to R
DISCOVERING STATISTICS USING R
ANDY FIELD | JEREMY MILES | ZOË FIELD
O'REILLY
R for Data Science
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA
Hadley Wickham & Garrett Grolemund
O'REILLY
Hands-On Programming with R
WRITE YOUR OWN FUNCTIONS AND SIMULATIONS
Garrett Grolemund
Foreword by Hadley Wickham
Use R!
Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Second Edition
Springer
O'REILLY
Text Mining with R
A TIDY APPROACH
Julia Silge & David Robinson
Max Kuhn · Kjell Johnson
Applied Predictive Modeling
Springer
Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer
Deep Learning with R
François Chollet
with J. J. Allaire
MANNING
Second Edition
Doing Bayesian Data Analysis
A Tutorial with R, JAGS, and Stan
$p(\theta|D)$ = $p(D|\theta)$ × $p(\theta)$ / $p(D)$
John K. Kruschke
O'REILLY
Efficient R Programming
A PRACTICAL GUIDE TO SMARTER PROGRAMMING
Colin Gillespie & Robin Lovelace
Use R!
Dirk Eddelbuettel
Seamless R and C++ Integration with Rcpp
Springer
THE R BOOK
SECOND EDITION
MICHAEL J. CRAWLEY
WILEY
The R Series
Advanced R
Hadley Wickham
CRC Press
O'REILLY
R Packages
ORGANIZE, TEST, DOCUMENT AND SHARE YOUR CODE
Hadley Wickham
THE CLASSIC WORK
NEWLY UPDATED AND REVISED
The Art of Computer Programming
VOLUME 1
Fundamental Algorithms
Third Edition
DONALD E. KNUTH

Websites

The web is a great place to learn about R.

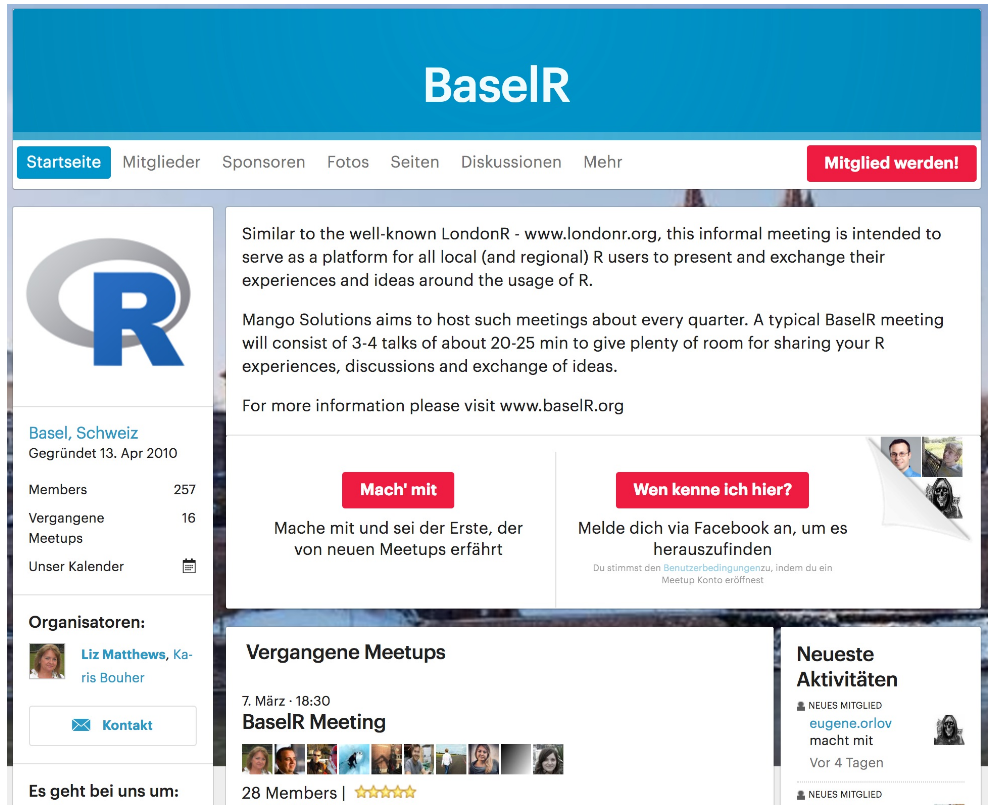
Google or Rseek, which is a wrapper around google to maximize hits related to R. However, most of the time Google works just fine. Just be sure to add to the the search query.

R-bloggers is a website on which R users inform each other on the newest developments. See, e.g., Nathaniel's entry.

Stackoverflow is a website on which R users exchange problems and solutions to problems. Try post something yourself. You will be amazed by the turnaround.

R meetup groups

Great place to meet other **R enthusiast** in and outside of industry and see interesting presentions on R's latest developments.



Similar to the well-known LondonR - www.londonr.org, this informal meeting is intended to serve as a platform for all local (and regional) R users to present and exchange their experiences and ideas around the usage of R.

Mango Solutions aims to host such meetings about every quarter. A typical BaselR meeting will consist of 3-4 talks of about 20-25 min to give plenty of room for sharing your R experiences, discussions and exchange of ideas.

For more information please visit www.baselR.org

Basel, Schweiz
Gegründet 13. Apr 2010

Members	257
Vergangene Meetups	16
Unser Kalender	

Organisatoren:
Liz Matthews, Karis Bouher

Kontakt

Es geht bei uns um:

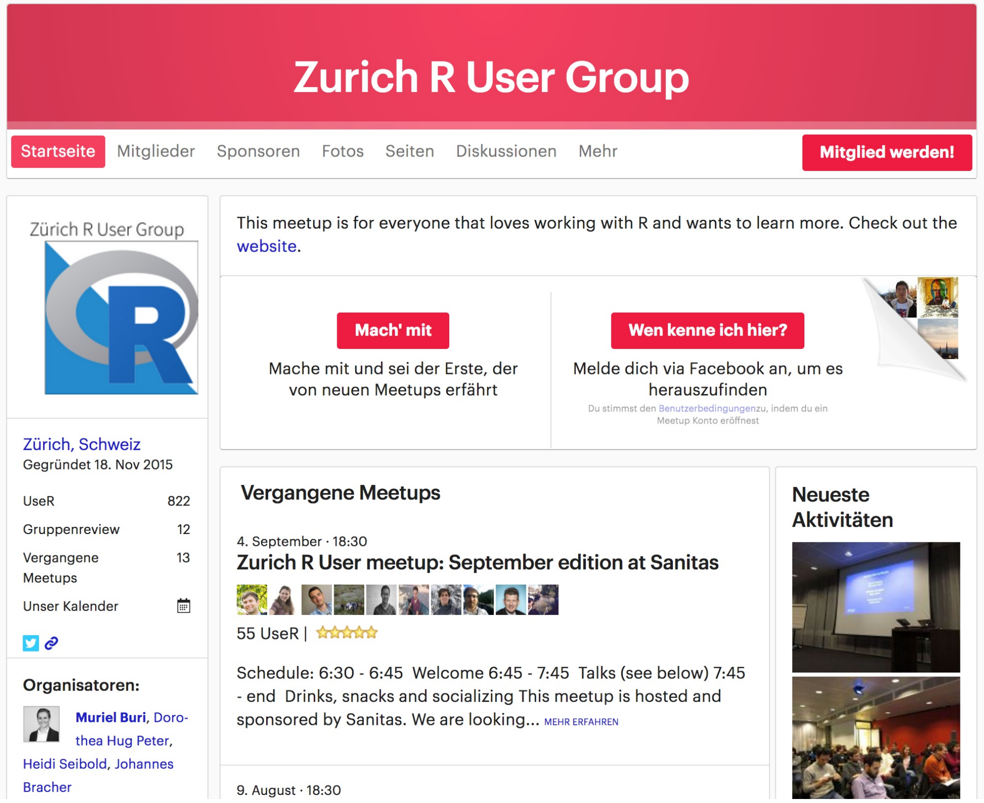
Mach' mit
Mache mit und sei der Erste, der von neuen Meetups erfährt

Wen kenne ich hier?
Melde dich via Facebook an, um es herauszufinden
Du stimmst den Benutzerbedingungenzu, indem du ein Meetup Konto eröffnest

Vergangene Meetups

7. März · 18:30
BaselR Meeting
28 Members | ☆☆☆☆☆

Neueste Aktivitäten
NEUES MITGLIED
eugene.orlov
macht mit
Vor 4 Tagen
NEUES MITGLIED



This meetup is for everyone that loves working with R and wants to learn more. Check out the website.

Zürich, Schweiz
Gegründet 18. Nov 2015

UseR	822
Gruppenreview	12
Vergangene Meetups	13
Unser Kalender	

Organisatoren:
Muriel Buri, Dorothea Hug Peter, Heidi Seibold, Johannes Bracher

Mach' mit
Mache mit und sei der Erste, der von neuen Meetups erfährt

Wen kenne ich hier?
Melde dich via Facebook an, um es herauszufinden
Du stimmst den Benutzerbedingungenzu, indem du ein Meetup Konto eröffnest

Vergangene Meetups

4. September · 18:30
Zurich R User meetup: September edition at Sanitas
55 UseR | ☆☆☆☆☆

Schedule: 6:30 - 6:45 Welcome 6:45 - 7:45 Talks (see below) 7:45 - end Drinks, snacks and socializing This meetup is hosted and sponsored by Sanitas. We are looking... MEHR ERFAHREN

9. August · 18:30

Neueste Aktivitäten

Consulting

Feel free to contact us.

Dr. Dirk Wulff

dirkwulff.org
github.com/dwulff
cstab, mousetrap, choicepp

Dr. Nathaniel Phillips

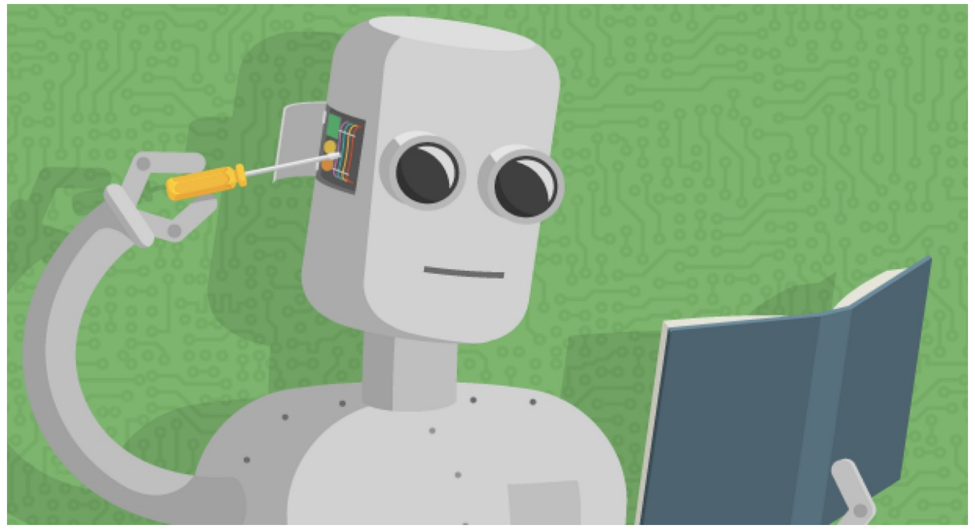
nathanieldphillips.com
github.com/ndphillips
yarrr, FFTrees

Markus Steiner

github.com/mdsteiner
ShinyPsych

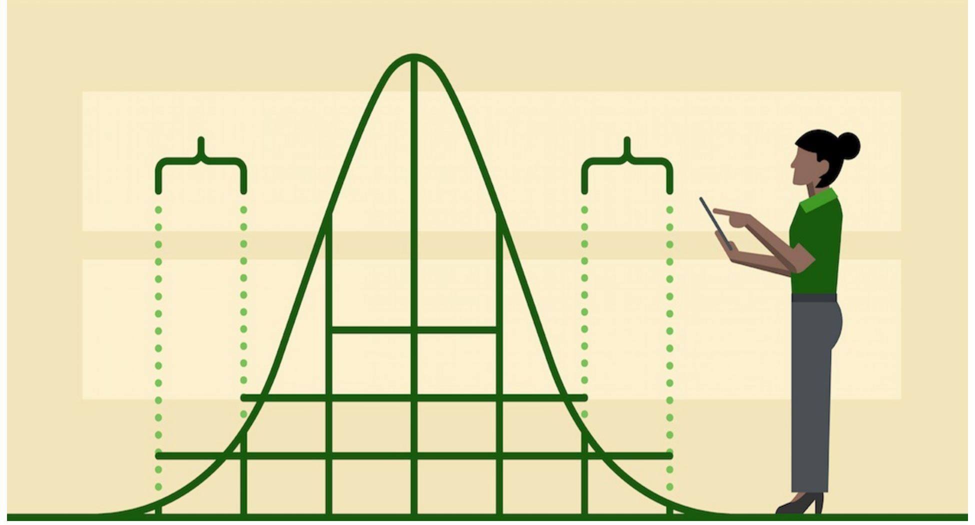
Next courses

Applied Machine Learning with R



2 days, January 19-20, 2019

Statistics with R



2 days, February/March, 2019

Thank you

Here is an R Joke.

And now one more thing...