Next steps

R for Data Science

Basel R Bootcamp

February 2019

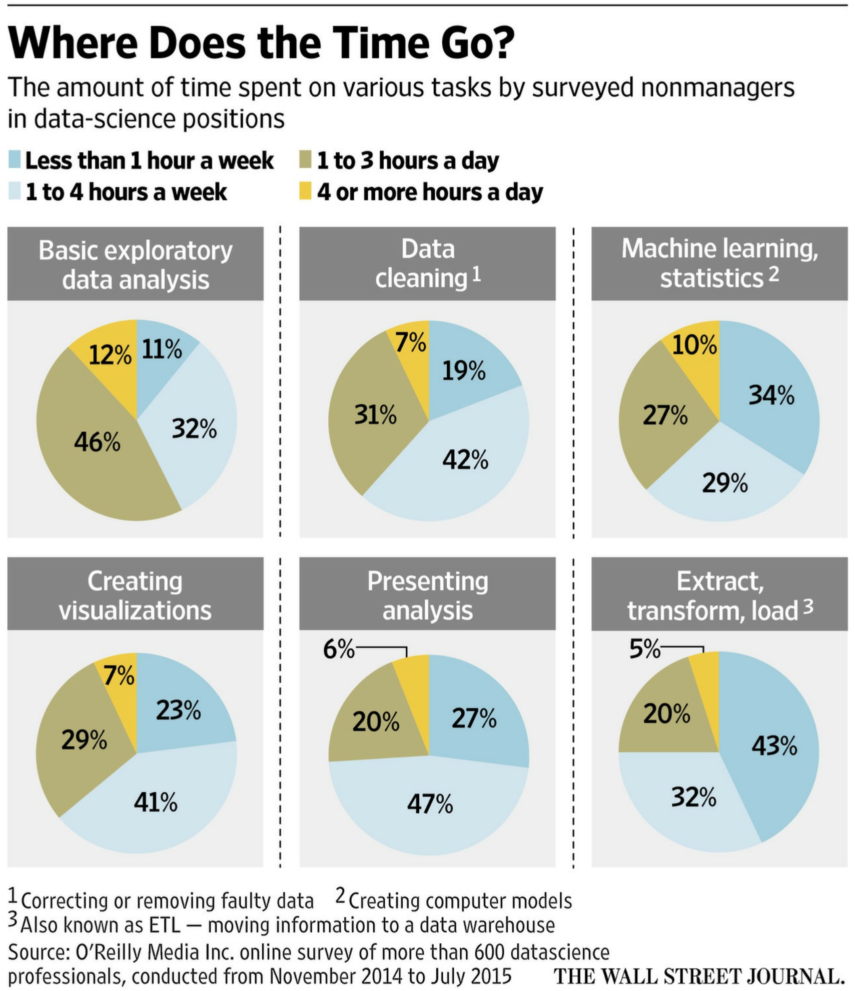
Hello Data Scientist!

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

	Sat, 23 Feb	Sun, 24 Feb
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



Where Does the Time Go?

The amount of time spent on various tasks by surveyed nonmanagers in data-science positions

[1] Correcting or removing faulty data [2] Creating computer models

[3] Also known as ETL — moving information to a data warehouse

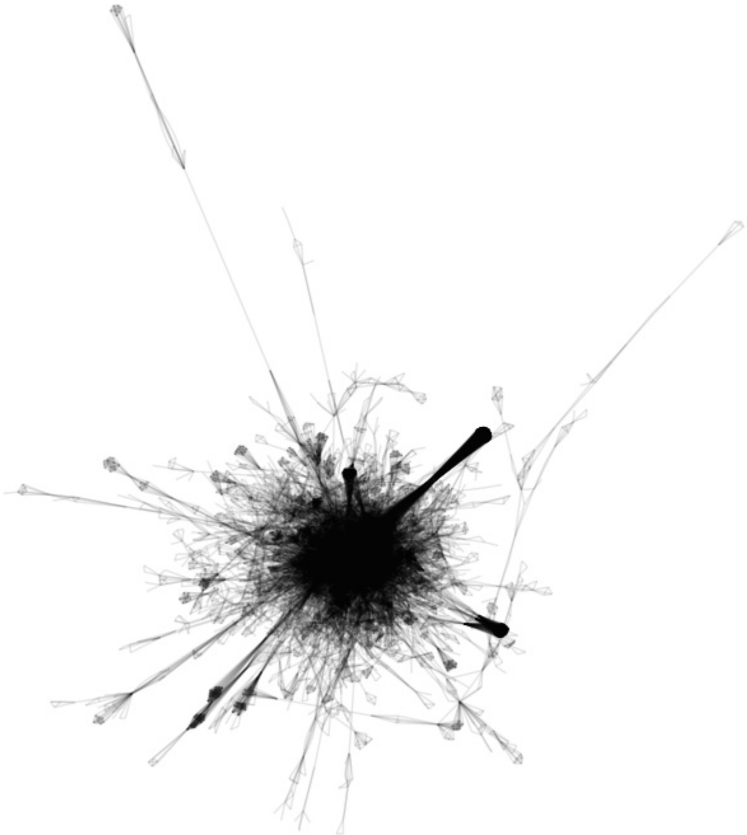
Source: O'Reilly Media Inc. online survey of more than 600 datascience professionals, conducted from November 2014 to July 2015

THE WALL STREET JOURNAL.

Wall Street Journal

Stuff we didn't cover

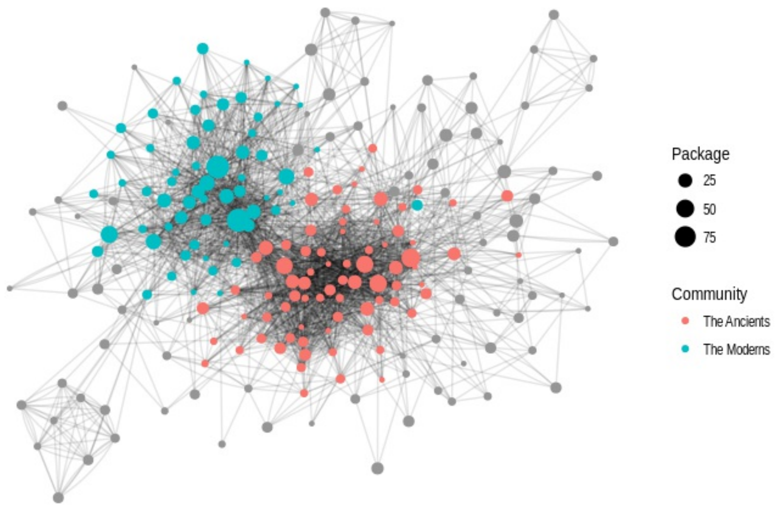
1. Networks
2. Statistics
3. Machine learning
4. Text analysis
5. Rcpp
6. Forms



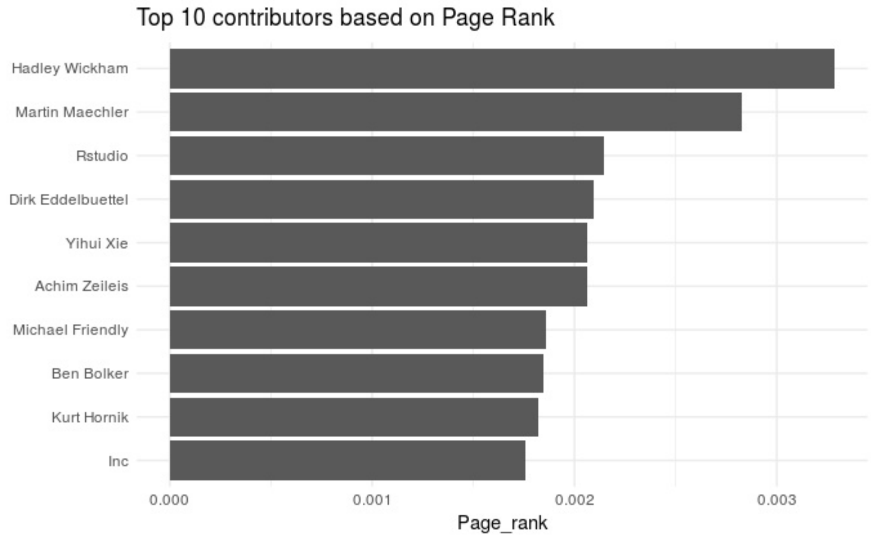
source R-bloggers.com

Networks

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**

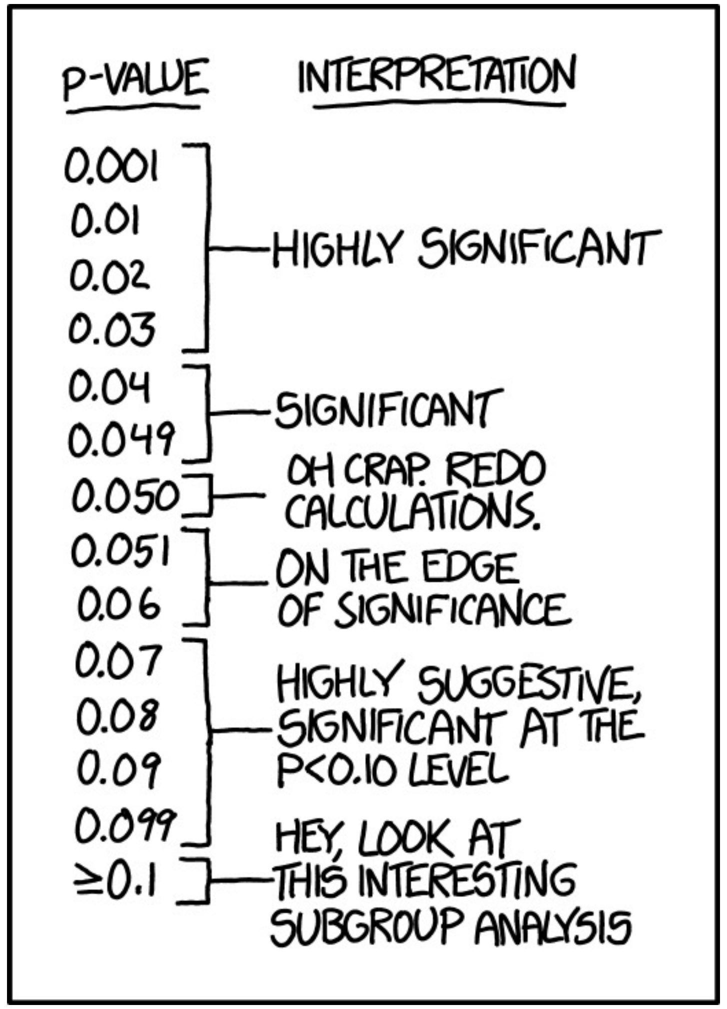


source **R-bloggers.com**

Stats

> "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.

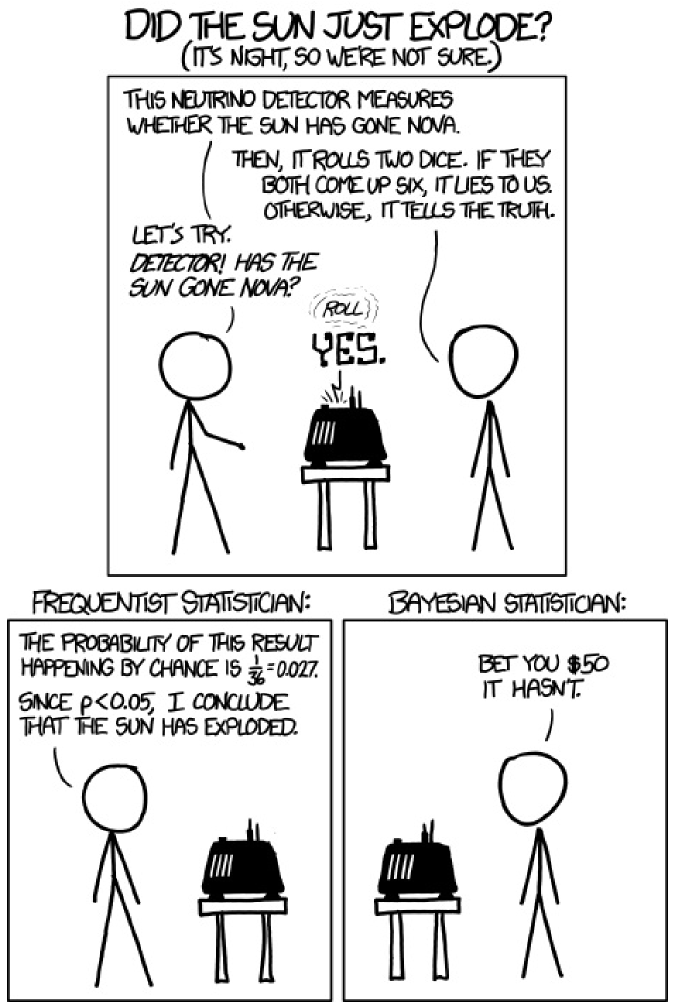

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

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

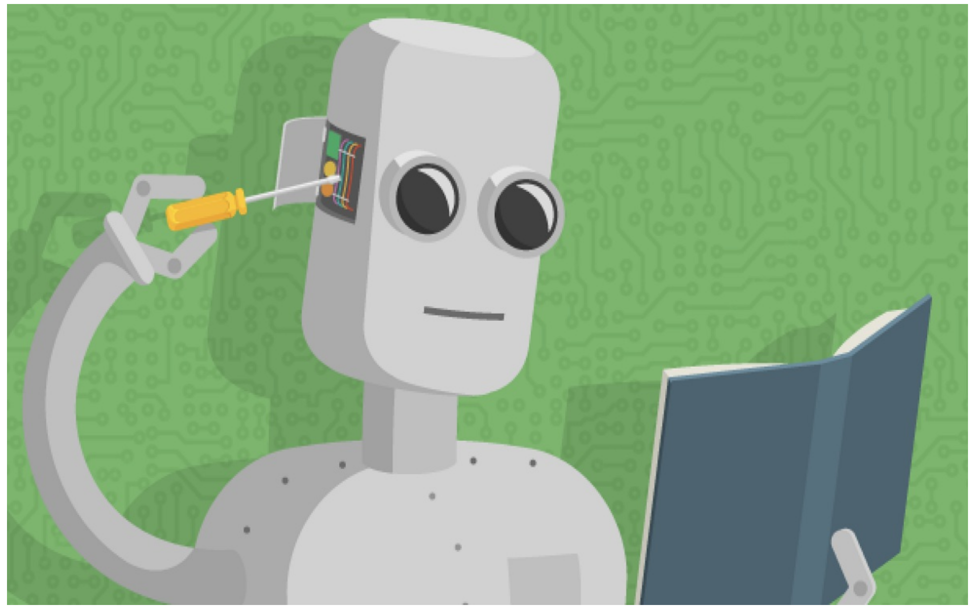
Package	Description
BayesFactor, rstanarm	Bayesian linear models. As easy as non-Bayesian methods.
rjags, rstan	Build flexible, hierarchical Bayesian models.
mcmc	Metropolis algorithms.
bridgesampling	Estimating marginal likelihoods using bridgesampling.


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 $\frac{1}{36}$=0.027. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

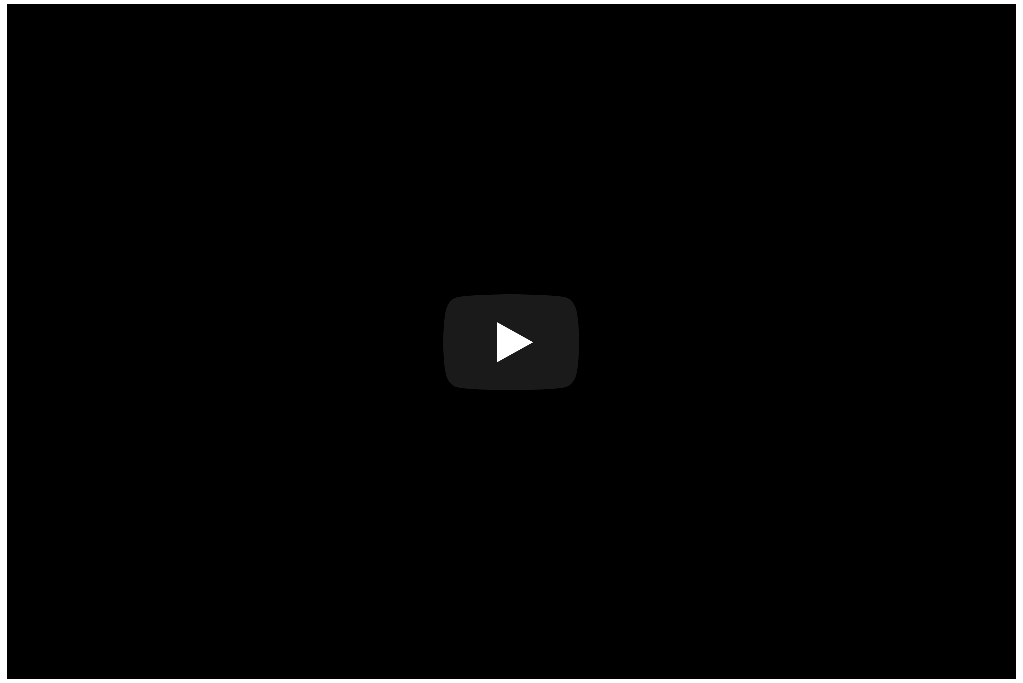
xkcd.com

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`	Deep learning.



Text analysis



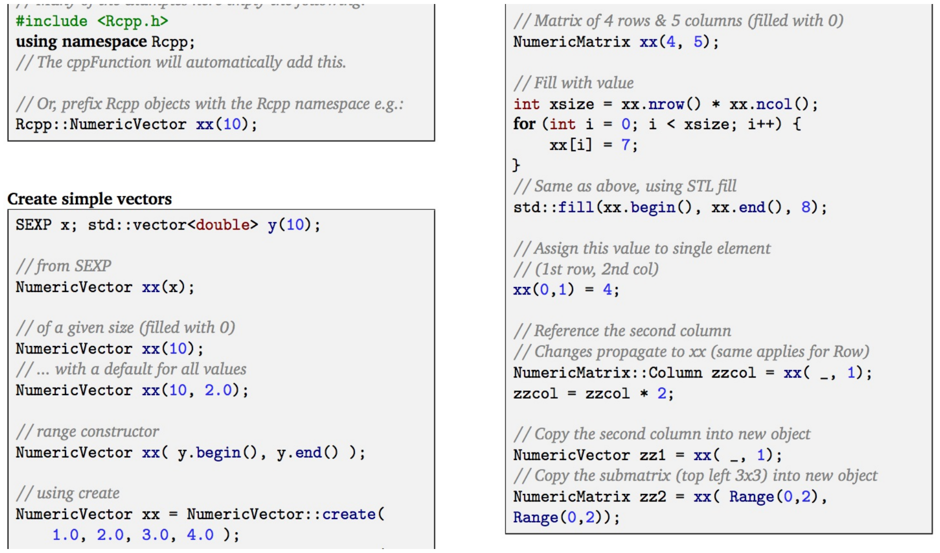
Sundar Pichai @ Google IO, May 2018

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.

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**



```
// Many of the examples here imply the following:
#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 );
```

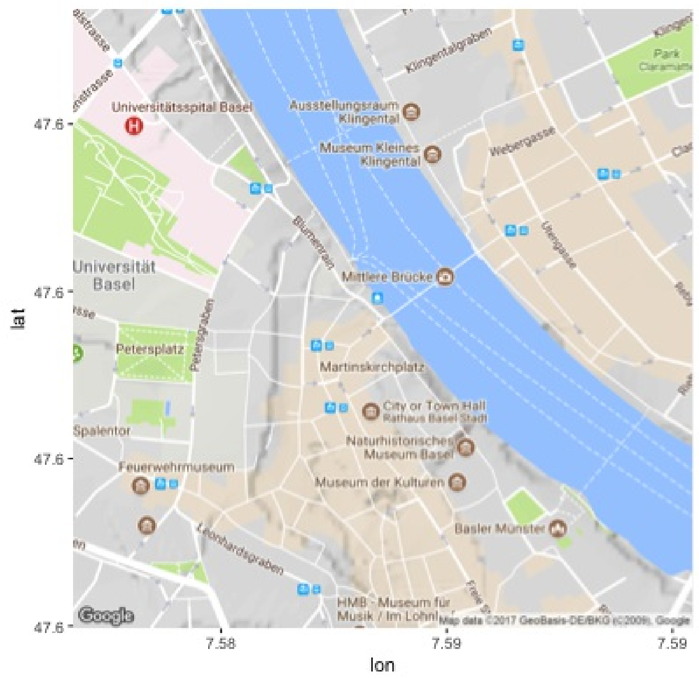
```
// Matrix of 4 rows & 5 columns (filled with 0)
NumericMatrix xx(4, 5);

// 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));
```

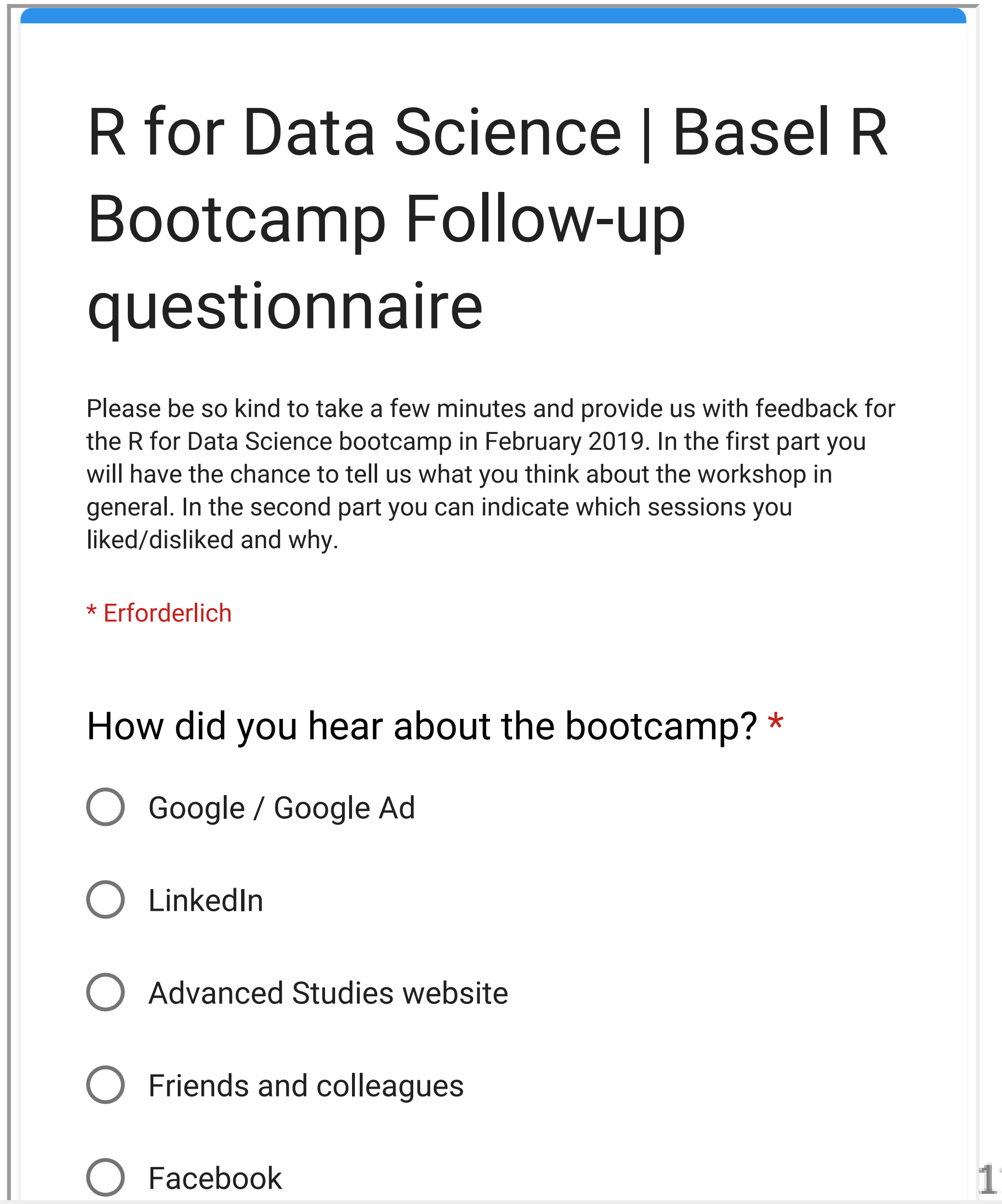
Quick Reference Guide

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))
```





R for Data Science | Basel R Bootcamp Follow-up questionnaire

Please be so kind to take a few minutes and provide us with feedback for the R for Data Science bootcamp in February 2019. 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

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 Pirate's Guide to R
DISCOVERING STATISTICS USING R
ANDY FIELD | JEREMY MILES | ZOË FIELD
O'REILLY R for Data Science
Hadley Wickham & Garrett Grolemund
O'REILLY Hands-On Programming with R
Garrett Grolemund
Hadley Wickham ggplot2 Elegant Graphics for Data Analysis Second Edition
Springer
O'REILLY Text Mining with R
Julia Silge & David Robinson
Max Kuhn · Kjell Johnson Applied Predictive Modeling
Gareth James Daniela Witten Trevor Hastie Robert Tibshirani An Introduction to Statistical Learning with Applications in R
Deep Learning with R
Doing Bayesian Data Analysis
A Tutorial with R, JAGS, and Stan
John K. Kruschke
O'REILLY Efficient R Programming
Colin Gillespie & Robin Lovelace
Dirk Eddelbuettel Seamless R and C++ Integration with Rcpp
THE R BOOK SECOND EDITION
MICHAEL J. CRAWLEY
WILEY
The R Series Advanced R
Hadley Wickham
O'REILLY R Packages
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.

Support & Consulting

Dr. Dirk Wulff

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

Dr. Nathaniel Phillips

nathanieldphillips.com
github.com/ndphillips
yarrr,
FFTrees

Markus Steiner

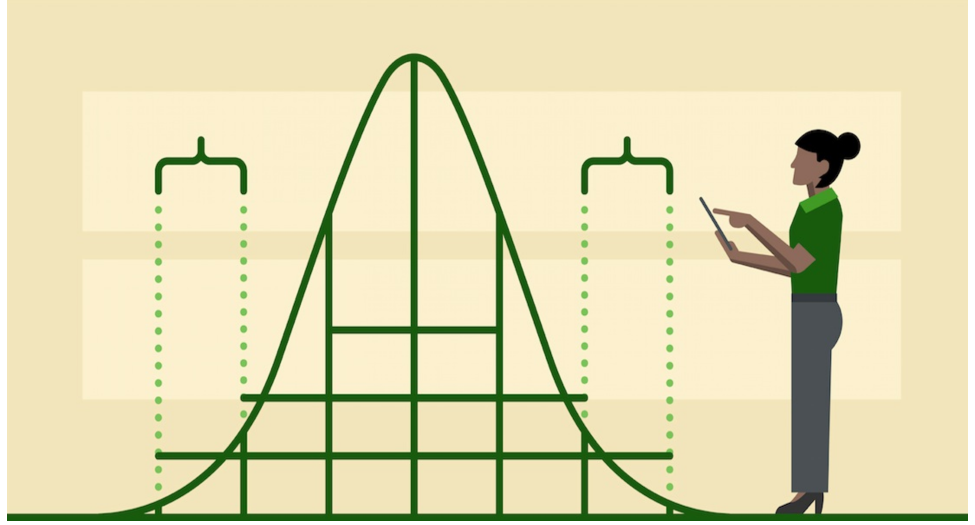
github.com/mdsteiner
ShinyPsych

Dr. Michael Schulte-M.

schulte-mecklenbeck.com
github.com/schultem

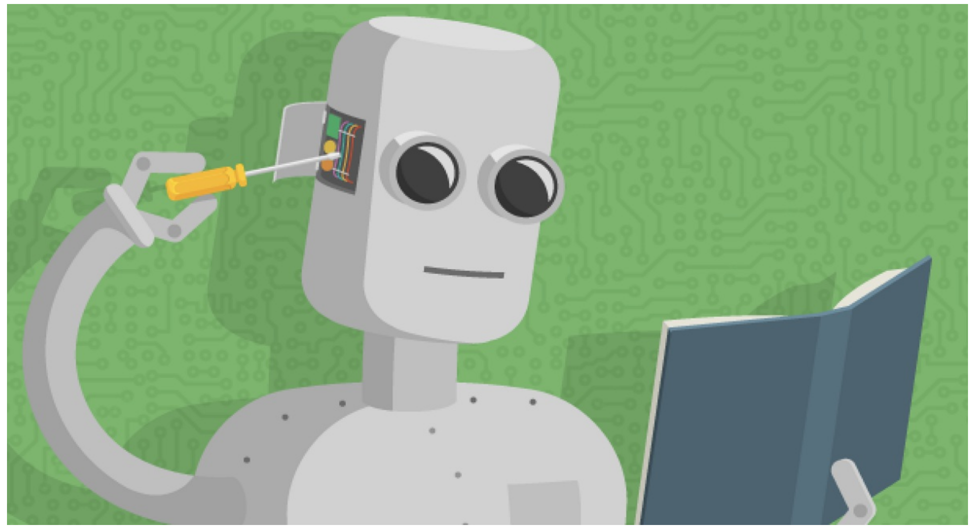
Next Bootcamps

Statistics with R



2 days, April 6-7, 2019

Applied Machine Learning with R



2 days, May 11-12, 2019

Thank you

Here is an R Joke.

And now one more thing...