The Data Scientist’s Toolbox

## About us

Data intensive statistics in biology and medicine

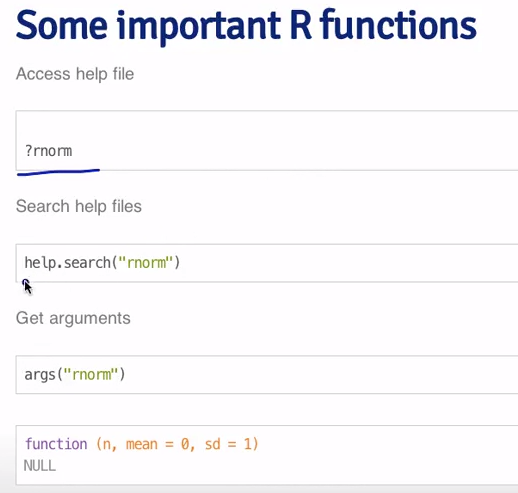
* Brian Caffo
  + Website <http://www.bcaffo.com/>
  + Twitter [@bcaffo](https://twitter.com/bcaffo)
  + Github <https://github.com/bcaffo>
* Jeff Leek
  + Website <http://biostat.jhsph.edu/~jleek/>, <http://simplystatistics.org/>
  + Twitter [@jtleek](https://twitter.com/jtleek)
  + Github <https://github.com/jtleek>
* Roger Peng
  + Website <http://www.biostat.jhsph.edu/~rpeng/>,<http://simplystatistics.org/>
  + Twitter [@rdpeng](https://twitter.com/rdpeng)
  + Github <https://github.com/rdpeng>

RStudio – <http://www.rstudio.com/>

Primary file types – (.R) script – <http://www.rstudio.com/ide/docs/using/source>

Reproducible markdown documents (.Rmd) – <http://www.rstudio.com/ide/docs/authoring/using_markdown>

## Getting help on R





How to get help when using R – video

<https://www.youtube.com/watch?v=ZFaWxxzouCY&list=PLjTlxb-wKvXNSDfcKPFH2gzHGyjpeCZmJ&index=3>

R mailing list

Stackoverflow – with tag “[r]”

CrossValidated

Google – [data type] R package

Machine learning packages

Pamr – yes

Knitr – no

Rpart – yes

Lubridate – no

Gbm – yes

Solidify – no

Caret – yes

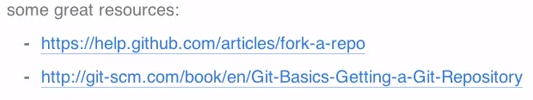
Kernlab – yes

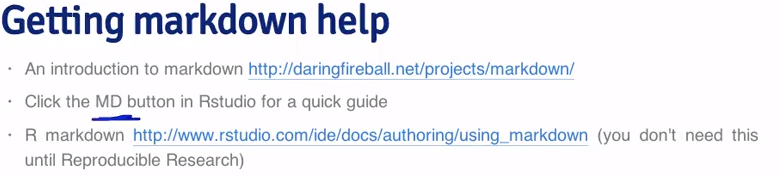
Filehash – no

Shiny – no

Link to install Git

<http://git-scm.com/downloads>





# Types of Data Science Questions (in approximate order of difficulty)

# Descriptive

Just describes a set of data

Commonly applied to census data

Description and interpretation are separate steps

Cannot be generalized

Example – http://www.census.gov/2010census/

# Exploratory

Find relationships you didn’t know about

Good for discovering new connections

Useful for defining future studies(to confirm what you find)

Not usually the final say

Should not be used for generalizing/predicting

Correlation does not imply causation

# Inferential

Use a small sample to say something about a bigger population

Commonly the goal of statistical models

Involves estimating both the quantity you care about and your uncertainty about your estimate

# Predictive

Use the data on some objects to predict values for another object

Accurate prediction depends heavily on measuring right variables

In general, more data and simple model works really well

Prediction is very hard

Samples – predict election outcome, Target predicts girl is pregnant

# Causal

Find what happens to one variable when you change another

Normally applies “on average”

They are the “gold standard” for data analysis

# Mechanistic

Understand exact changes in one variable that leads to exact changes in another

Very hard and rarely done, needs no noise in the data

Example – changes in pavement design

Definition of Data

“set of items” you are measuring is the “population”

Data are the values of the qualitative or quantitative variables for the population

Variables are measurements of characteristics of an item

Government has lots of data - [www.data.gov](http://www.data.gov)

Data sharing plans

<https://github.com/>

<http://figshare.com/>

<https://github.com/jtleek/datasharing> - guide to data sharing