



An overview of R programming language

By : Naeem Khoshnevis
Center for Earthquake Research and Information
University of Memphis

May 2017

R (Programming language)

- R is an open source programming language
- A software environment for statistical computing
- It is widely used among statisticians and data miners for developing statistical software and data analysis.
- The source code is written primarily in C, Fortran, and R.
- R is freely available under the GNU General Public License.
- Runs on almost any standard computing platform/OS (even on the PlayStation 3)

R language	
Paradigms	Multi-paradigm: Array, object-oriented, imperative, functional, procedural, reflective
Designed by	Ross Ihaka and Robert Gentleman
Developer	R Core Team ^[1]
First appeared	August 1993; 23 years ago ^[2]
Stable release	3.3.3 / March 6, 2017; 47 days ago
Typing discipline	Dynamic
License	GNU GPL v2 ^[3]
Filename extensions	.r, .R, .RData, .rds, .rda
Website	www.r-project.org

R (Programming language)

- R is an open source programming language
- A software environment for statistical computing
- It is widely used among statisticians and data miners for developing statistical software and data analysis.
- The source code is written primarily in C, Fortran, and R.
- R is freely available under the GNU General Public License.
- Runs on almost any standard computing platform/OS (even on the PlayStation 3)

R language	
	
Paradigms	Multi-paradigm: Array, object-oriented, imperative, functional, procedural, reflective
Designed by	Ross Ihaka and Robert Gentleman
Developer	R Core Team ^[1]
First appeared	August 1993; 23 years ago ^[2]
Stable release	3.3.3 / March 6, 2017; 47 days ago
Typing discipline	Dynamic
License	GNU GPL v2 ^[3]
Filename extensions	.r, .R, .RData, .rds, .rda
Website	www.r-project.org 

R (Programming language)

- R is an open source programming language
- A software environment for statistical computing
- It is widely used among statisticians and data miners for developing statistical software and data analysis.
- The source code is written primarily in C, Fortran, and R.
- R is freely available under the GNU General Public License.
- Runs on almost any standard computing platform/OS (even on the PlayStation 3)

R language	
	
Paradigms	Multi-paradigm: Array, object-oriented, imperative, functional, procedural, reflective
Designed by	Ross Ihaka and Robert Gentleman
Developer	R Core Team ^[1]
First appeared	August 1993; 23 years ago ^[2]
Stable release	3.3.3 / March 6, 2017; 47 days ago
Typing discipline	Dynamic
License	GNU GPL v2 ^[3]
Filename extensions	.r, .R, .RData, .rds, .rda
Website	www.r-project.org 

Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. (<https://www.r-project.org/>)

History of R

- 1991 Created in New Zealand by Ross Ihaka and Robert Gentleman.
- 1993 First announcement of R to the public.
- 1995 Martin Mächler convinces Ross and Robert to use the GNU General Public License to make R free software.
- 1996 A public mailing list is created (R-help and R-devel)
- 1997 The R Core Group is formed (containing some people associated with S-PLUS). The core group controls the source code for R.
- 2000 R version 1.0.0 is released.
- 2013 R version 3.0.2 is released on December 2013.
- 2017 R version 3.4.0 is released on April 2017.



Quick Facts about R

R is the highest paid IT skill
[\(Dice.com survey, January 2014\)](#)

R most-used data science language after SQL
[\(O'Reilly survey, January 2014\)](#)

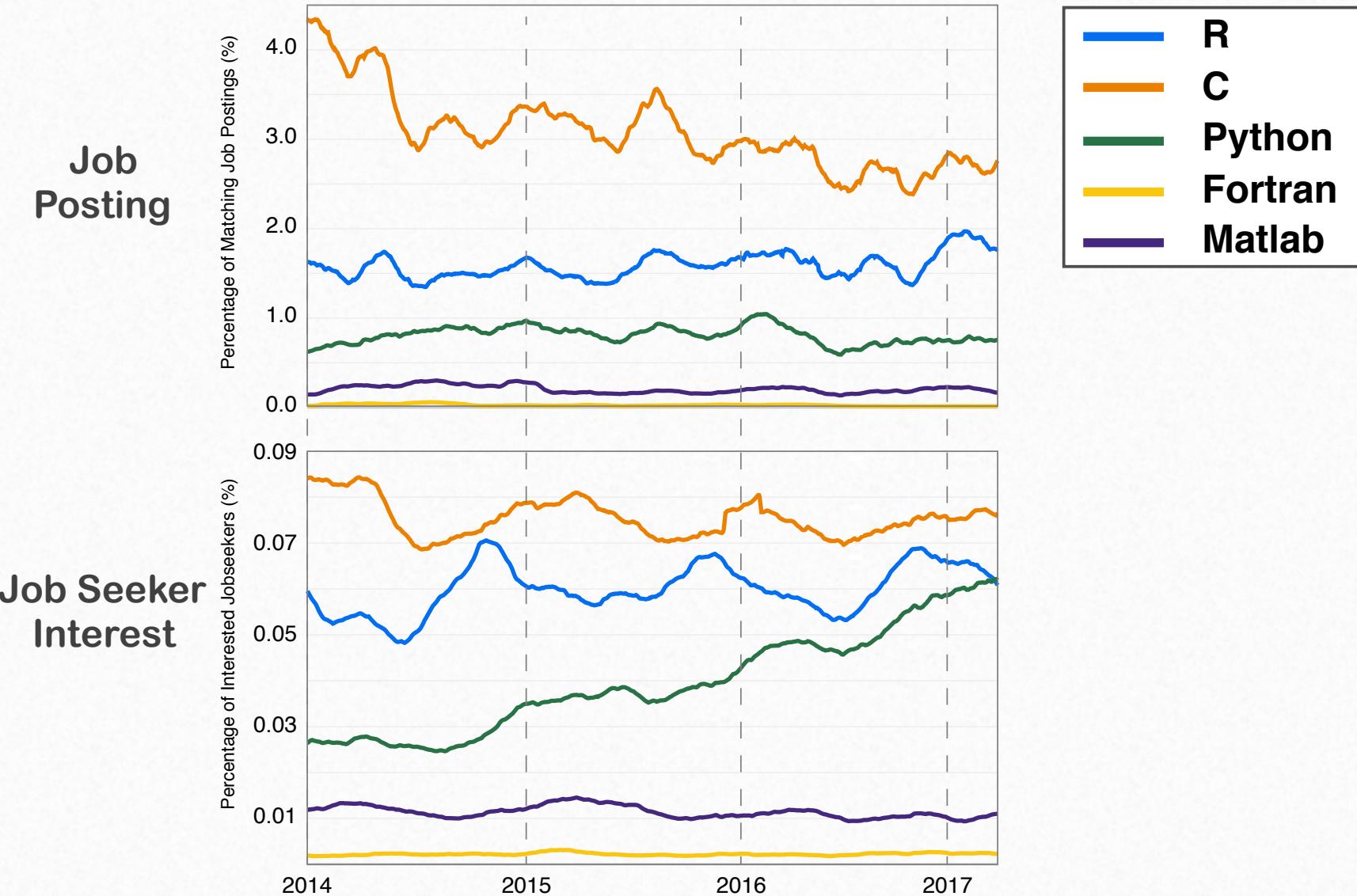
R is used by 70% of data miners
[\(Rexer survey, October 2013\)](#)

R growing faster than any other data science language
[\(KD Nuggets survey, August 2013\)](#)

R is the #1 Google Search for Advanced Analytics software
[\(Google Trends, March 2014\)](#)

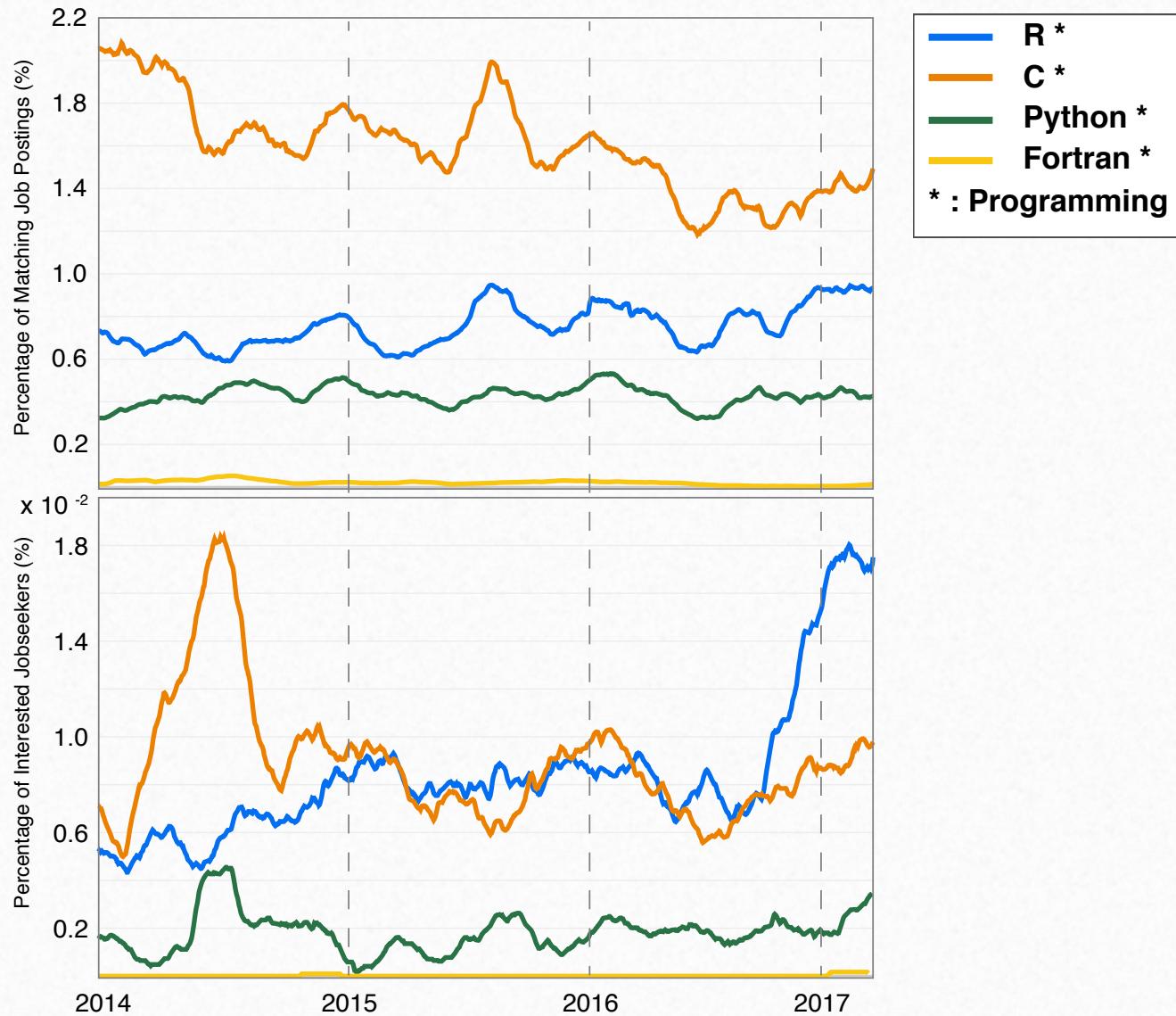
R has more than 2 million users worldwide
[\(Oracle estimate, February 2012\)](#)

R from job perspective



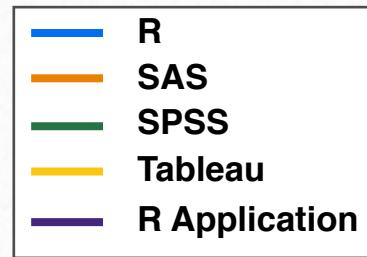
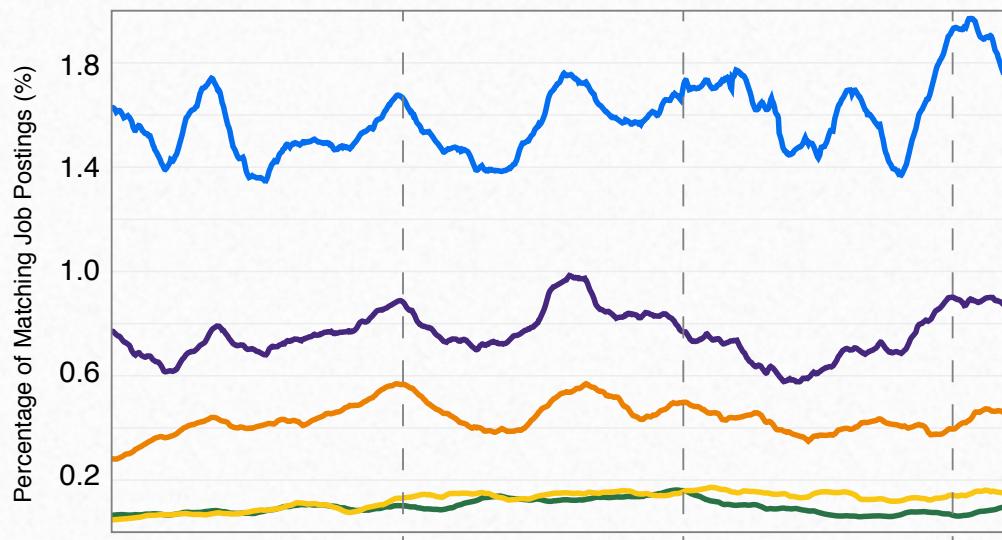
R from job perspective

Job
Posting

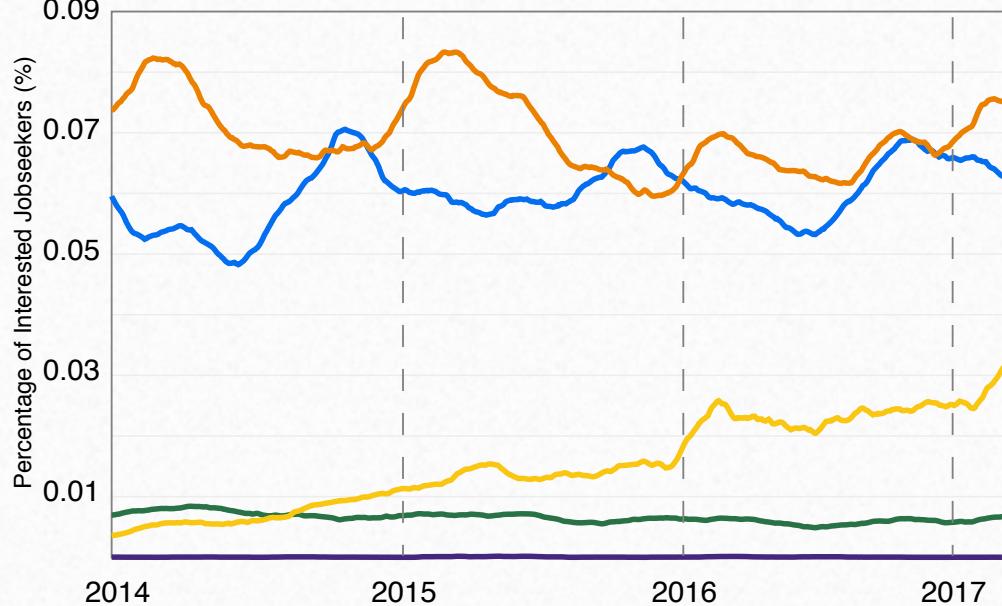


R from job perspective

Job
Posting



Job
Seeker
Interest



R – R studio Environment

Download R from:

<https://www.r-project.org>

Download RStudio from:

<https://www.rstudio.com>

The screenshot shows the official RStudio website. At the top, there's a navigation bar with links for rstudio::conf, Products, Resources, Pricing, About Us, Blogs, and a search icon. Below the navigation is a large banner with the RStudio logo and the text "Open source and enterprise-ready professional software for R". To the right of the banner is a vertical sidebar with buttons for "Download RStudio", "Discover Shiny", "shinyapps.io Login", and "Discover RStudio Connect". Below the banner, there are three main sections: "RStudio" (showing a screenshot of the RStudio interface), "Shiny" (showing a screenshot of a Shiny app interface), and "R Packages" (showing icons for various R packages like rmarkdown, Shiny, tidyverse, knitr, and ggplot2). Each section has a "Learn More" button.

R Studio

rstudio::conf Products Resources Pricing About Us Blogs

RStudio
Open source and enterprise-ready professional software for R

Download RStudio

Discover Shiny

shinyapps.io Login

Discover RStudio Connect

RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

[Download](#) [Learn More](#)

Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

[Learn More](#)

R Packages

Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

[Learn More](#)

R – R studio Environment

Download R from:

<https://www.r-project.org>

Download RStudio from:

<https://www.rstudio.com>

R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.68 (7288) x86_64-apple-darwin13.4.0]

Warning: namespace 'swirl' is not available and has been replaced
by .GlobalEnv when processing object 'pathToFile'
[Workspace restored from /Users/naeem/.RData]
[History restored from /Users/naeem/.Rapp.history]

>

1: ---
2: output: html_document
3: ---
4: ## An Overview of R Programming
5: ##### by: Naeem Khoshnevis
6:
7: ###### Summary
8: In this section I present
9: the Iris dataset.
10: ``{r}
11: data("iris") # load data
12: n_rows <- nrow(iris) # number of rows of data
13:
14: ---
1:1 Title

Console ~ / ↵
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Warning: namespace 'swirl' is not available and has been replaced
by .GlobalEnv when processing object 'pathToFile'
[Workspace loaded from ~/.RData]

> |

Global Environment
pathToFile /Library/Frameworks/R.framework/Versions/3.3/Resources/library/MASS
rng num [1:2] 3 40.1
site0 chr [1:33] "1.5" "1.12" ...
site1 chr [1:18] "1.5" "1.12" ...
wcol num [1:5] 3 4 5 11 13
x0sub num [1:122] NA NA 4.9 4.8...
x1sub num [1:30] 5.5 7.5 8.75 N...

Environment History
Import Dataset List
Global Environment
scale_manual (ggplot2) R Documentation
Create your own discrete scale Find In Topic
scale_manual (ggplot2) R Documentation

Create your own discrete scale

Description

This allows you to specify your own set of mappings from levels in the data to aesthetic values.

Data Types

Variables provide a means of accessing the data stored in memory.

R does not provide direct access to the computer's memory but rather provides a number of specialized data structures (objects)

Data structure:

- Vector : (1d) (Homogenous)
- List : (1d)(Heterogeneous)
- Matrix: (2d) (Homogenous)
- Data frame (2d)(Heterogeneous)
- Array(nD)

Programming in R

```
My_fun <- function(arg1,...){  
  statements  
  return(object)  
}
```

```
while (test_expression)  
{  
  statement  
}  
  
if (condition) {  
  # do something  
} else {  
  # do something else  
}
```

```
for (i in 1:n)  
{  
  statements  
}
```

Apply functions (R base package)

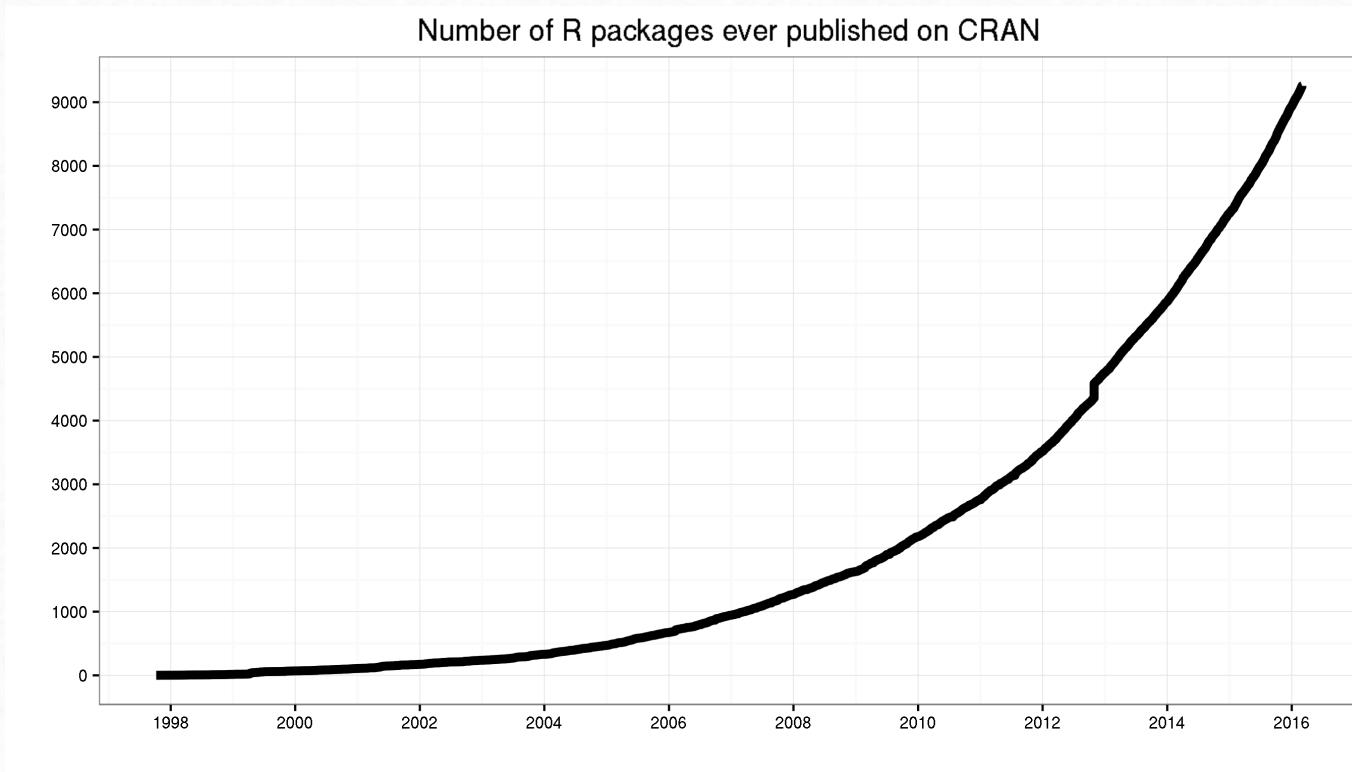
- keep you from having loop
- Apply a function on every row, column, or element
 - ✓ apply
 - ✓ lapply
 - ✓ sapply

`apply(data, margin, function)`

What is R's strength?

What is R's strength?

R packages



R Features

R can do regular expressions/text processing ([tm](#), [quanteda](#))

R can get data out of a database ([RMySQL](#), [mongolite](#))

R can process nasty data ([plyr](#), [reshape2](#))

R can process images ([jpeg](#))

R can handle different data formats([XML](#), [RJSONIO](#) , [xlsx](#))

R can interact with APIs ([rcurl](#), [httr](#), [twitteR](#))

R can build apps/interactive graphics ([shiny](#), [rCharts](#))

R can create dynamic documents ([knitr](#))

R can play with Hadoop ([rhadoop](#))

R can create interactive teaching modules ([swirl](#))

R interfaces very nicely with C if you need to be more active ([Rcpp](#), [rPython](#))

Comprehensive R Archive Network (CRAN)

<https://cran.r-project.org>

Thousands of packages are submitted.

Thousands of packages are being downloaded on a daily basis.

Here of some of packages categories:

- Bayesian
- Economics
- Finance
- Genetics
- Graphics
- High Performance Computing
- Machine Learning
- Natural Language Processing
- Numerical Mathematics
- Optimization
- Social science
- Time Series
- ...

Comprehensive R Archive Network (CRAN)

<https://cran.r-project.org>

Thousands of packages are submitted.

Thousands of packages are being downloaded on a daily basis.

Here of some of packages categories:

- Bayesian
- Economics
- Finance
- Genetics
- Graphics
- High Performance Computing
- Machine Learning
- Natural Language Processing
- Numerical Mathematics
- Optimization
- Social science
- Time Series
- ...

```
> install.packages("packagename")
```

Data Visualization (ggplot)

ggplot2 is a plotting system for R, based on the grammar of graphics.

Developed by Hadley Wickham (Iowa State University)

You are limited only by your imagination

- Object initialization

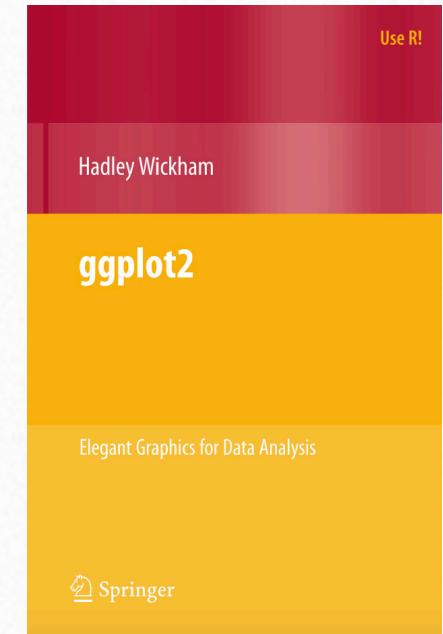
```
ggplot(data, mapping=aes())
```

- Aesthetic mapping (aes)

Describes how variables in the data are mapped to visual properties.

- Geometric objects

```
geom_point, geom_smooth, geom_boxplot, geom_jitter, ...
```

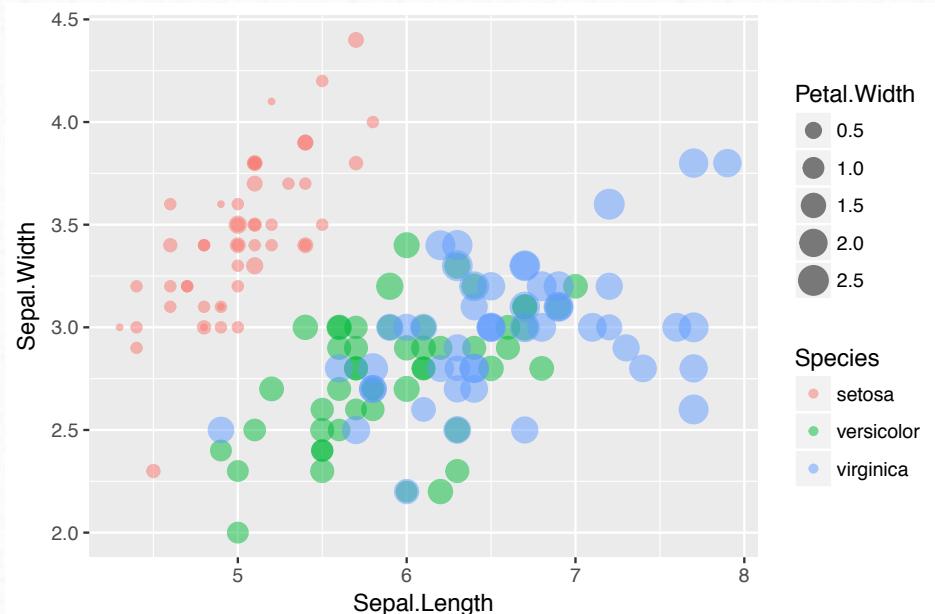


Data Visualization (ggplot)

```
> library(ggplot2)
> data("iris")
> Sepal.Length Sepal.Width Petal.Length Petal.Width Species
  5.1          3.5          1.4          0.2      setosa
> g <- ggplot(iris)
```

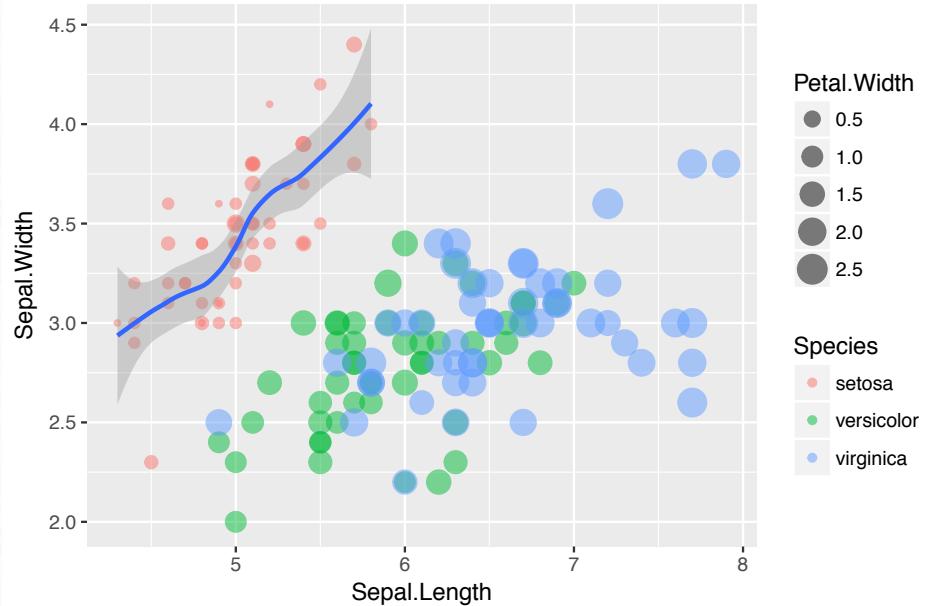
Data Visualization (ggplot)

```
> library(ggplot2)
> data("iris")
> Sepal.Length Sepal.Width Petal.Length Petal.Width Species
  5.1          3.5          1.4          0.2      setosa
> g <- ggplot(iris)
> g <- g + geom_point(aes(x=Sepal.Length,y=Sepal.Width, color=Species, size =Petal.Width), alpha=0.5)
```



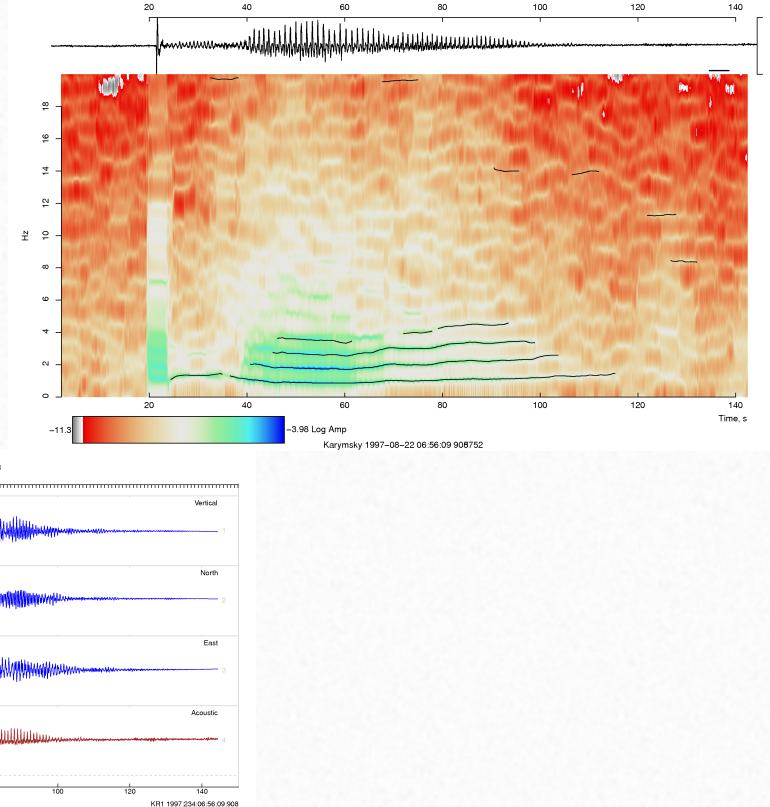
Data Visualization (ggplot)

```
> library(ggplot2)
> data("iris")
> Sepal.Length Sepal.Width Petal.Length Petal.Width Species
      5.1          3.5          1.4          0.2      setosa
> g <- ggplot(iris)
> g <- g + geom_point(aes(x=Sepal.Length,y=Sepal.Width, color=Species, size =Petal.Width), alpha=0.5)
> g <- g + geom_smooth(data = iris[iris$Species == 'setosa'], aes(x=Sepal.Length, y=Sepal.Width))
```



R packages for Earth Science (RSEIS)

- Seismic and Time Series Analysis
- Reads seismic data SAC, SEGY, AH, ASCII
- Spectrograms (MTM, AR)
- Data-Base Extraction of seismic traces
- Filtering
- Deconvolution
- Hodograms (Particle Motion)
- Predict Arrival times
- Moveout Displays
- Event Location
- Corner Frequency
- Attenuation

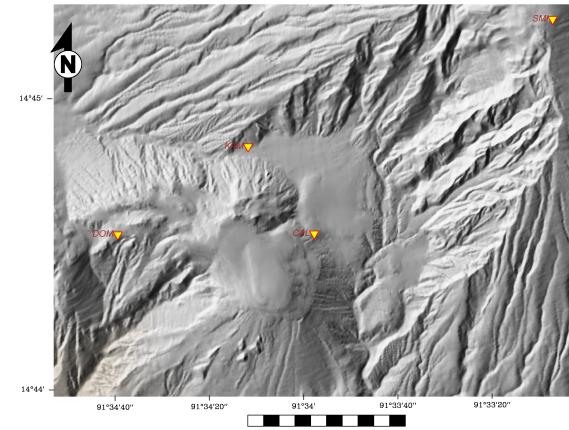
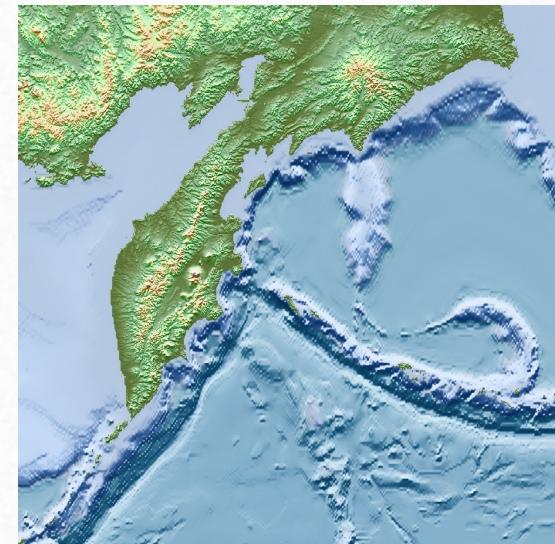


[Download from CRAN Website](#)

Slide from Jonathan Lees

R packages for Earth Science (GEOmap)

- Topographic/Geologic Maps
- Interactive Map
- Geographic Projection
- Geologic Symbols
- Polygon/Area Analysis
- Great Circles
- Cross Sections



[Download from CRAN Website](#)

Slide from Jonathan Lees

Reproducible Research

Deception at Duke

- A group of researchers published a paper at Nature Medicine in 2006, however, other researcher could not redo the process to get the same results. The paper retracted and several law suits filed.
- The idea of reproducible research more and more highlighted and different algorithm and procedures has been defined.
- The ultimate standard for strengthening scientific evidence is replication of findings and conducting studies with independent
 - ✓ Investigators
 - ✓ Data
 - ✓ Analytical methods
 - ✓ Laboratories
 - ✓ Instruments

<https://www.youtube.com/watch?v=W5sZTNPMQRM>

<https://www.coursera.org/learn/reproducible-research>



Reproducible Research (Knitr)

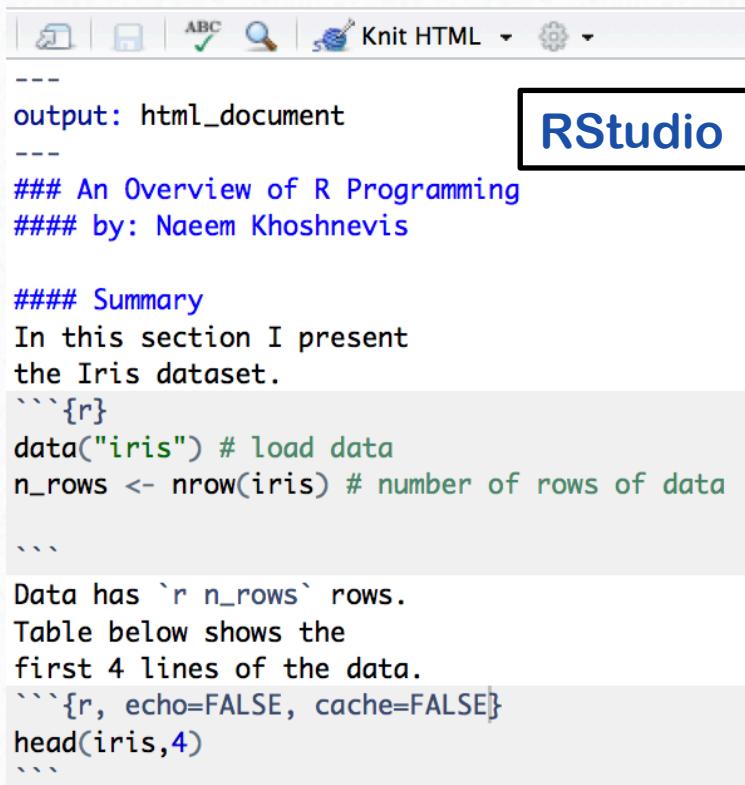
- Knitr is a package to literate (statistical) programming
- Knitr uses R as the programming language (although others are allowed) and variety of documentation languages:

LaTeX, Markdown, HTML

Reproducible Research (Knitr)

- Knitr is a package to literate (statistical) programming
- Knitr uses R as the programming language (although others are allowed) and variety of documentation languages:

LaTeX, Markdown, HTML



The screenshot shows the RStudio interface with a Knit HTML session open. The title bar includes icons for file operations, ABC, search, and Knit HTML. The main pane displays R code and its output. A large blue box highlights the word "RStudio" in the title bar.

```
output: html_document
---
### An Overview of R Programming
#### by: Naeem Khoshnevis

##### Summary
In this section I present
the Iris dataset.

```{r}
data("iris") # load data
n_rows <- nrow(iris) # number of rows of data

```
Data has `r n_rows` rows.
Table below shows the
first 4 lines of the data.

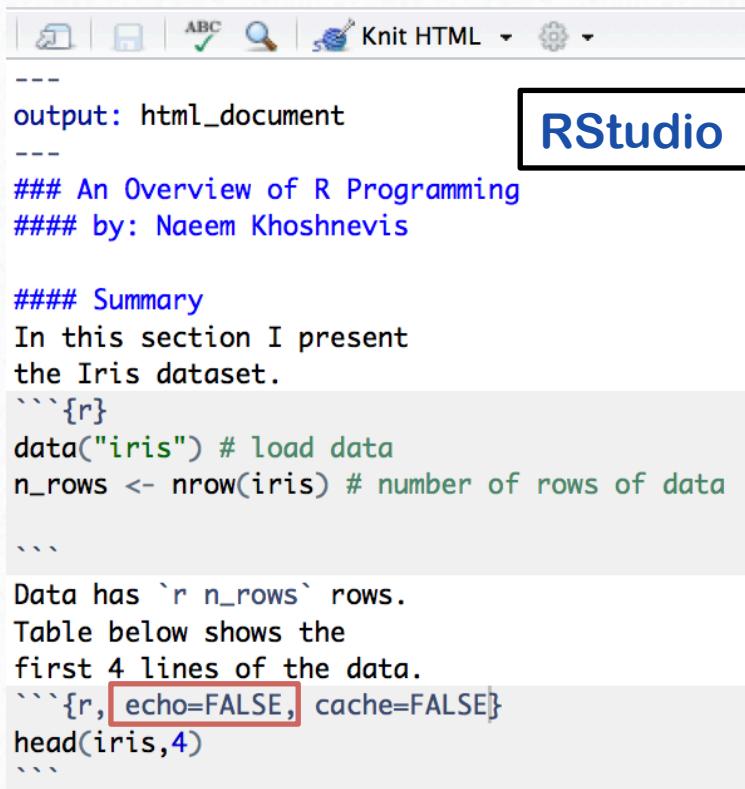
```{r, echo=FALSE, cache=FALSE}
head(iris,4)
```

```

Reproducible Research (Knitr)

- Knitr is a package to literate (statistical) programming
- Knitr uses R as the programming language (although others are allowed) and variety of documentation languages:

LaTeX, Markdown, HTML



The screenshot shows the RStudio interface. The top bar includes icons for file operations, ABC, search, and Knit HTML. The main area displays R code:

```
output: html_document
---
### An Overview of R Programming
#### by: Naeem Khoshnevis

#### Summary
In this section I present
the Iris dataset.

```{r}
data("iris") # load data
n_rows <- nrow(iris) # number of rows of data
```
Data has `r n_rows` rows.
Table below shows the
first 4 lines of the data.
```{r, echo=FALSE, cache=FALSE}
head(iris,4)
```
```

An Overview of R Programming

by: Naeem Khoshnevis

HTML

Summary

In this section I present the Iris dataset.

```
data("iris") # load data
n_rows <- nrow(iris) # number of rows of data
```

Data has 150 rows. Table below shows the first 4 lines of the data.

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|------|--------------|-------------|--------------|-------------|---------|
| ## 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| ## 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| ## 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| ## 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |

Machine Learning with R (CARET)

Classification And REgression Training

CARET is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

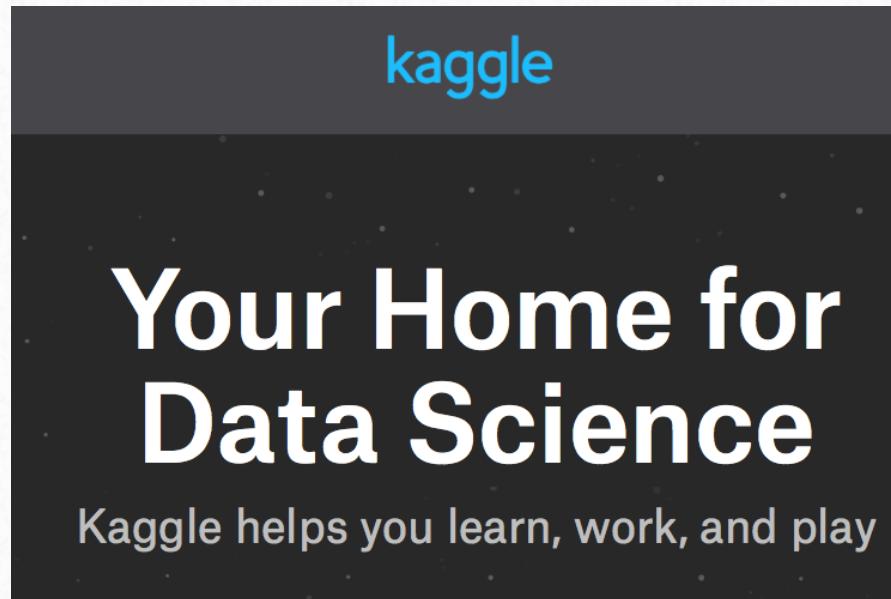
- data splitting
- pre-processing
- feature selection
- model tuning using resampling
- variable importance estimation

the caret package



AdaBoost, Bagged AdaBoost, Bayesian Additive Regression Trees, C4.5-like Trees, Generalized Linear Model, k-Nearest Neighbors, Linear Regression, Multi-Layer Perceptron, Naive Bayes, Neural Network, Random Forest, ...

Machine Learning with R (Kaggle)



Machine Learning with R (Kaggle)

| 9 active competitions | | | |
|---|--|-------------------------|----------------|
| | | Sort by | Prize |
| Active | All | Entered | All Categories |
|  | Intel & MobileODT Cervical Cancer Screening
Which cancer treatment will be most effective?
· 2 months to
Featured go | \$100,000
483 teams | |
|  | Google Cloud & YouTube-8M Video Understanding Challenge
Can you produce the best video tag predictions?
· a month to
Featured go | \$100,000
509 teams | |
|  | Planet: Understanding the Amazon from Space
Use satellite data to track the human footprint in the Amazon rainforest
· 3 months to
Featured go | \$60,000
219 teams | |
|  | Sberbank Russian Housing Market
Can you predict realty price fluctuations in Russia's volatile economy?
· 2 months to
Featured go | \$25,000
431 teams | |
|  | NOAA Fisheries Steller Sea Lion Population Count
How many sea lions do you see?
· 2 months to
Featured go | \$25,000
121 teams | |
|  | Quora Question Pairs
Can you identify question pairs that have the same intent?
· a month to
Featured go | \$25,000
2,259 teams | |



The image shows the top portion of the Kaggle homepage. It features the Kaggle logo in blue at the top right. Below it is a large white banner with the text "Your Home for Data Science" in bold black letters. Underneath that, a smaller white banner says "Kaggle helps you learn, work, and play".

Machine Learning with R (Kaggle)

Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 6,908 teams · 3 years to go

Overview Data Kernels Discussion Leaderboard More Submit Predictions

Naeem Khoshnevis
Tuned cforest (Public score: 0.80861)

last run 4 months ago · R script · 112 views
using data from [Titanic: Machine Learning from Disaster](#) · Public

[Code](#) [Output \(1\)](#) [Comments \(0\)](#) [Log](#) [Versions \(7\)](#)

Submission
✓ Ran successfully
Submitted by Naeem 4 months ago

Public Score
0.80861

[Titanic: Machine Learning ...](#) 3 years to go · Top 8% **512th** of 6670

<https://www.kaggle.com/naeemkh/titanic/tuned-cforest-public-score-0-80861>

Machine Learning with R (H2O)

Open Source AI Platform

- H2O is the world's leading open source deep learning platform.
- H2O is used by over 80,000 data scientists and more than 9,000 organizations around the world.



H2O.ai brings AI to enterprise

H2O

The #1 open source platform for AI.

Deep Water

World's best frameworks and enterprise GPU support for TensorFlow, MxNet, Caffe and H2O.

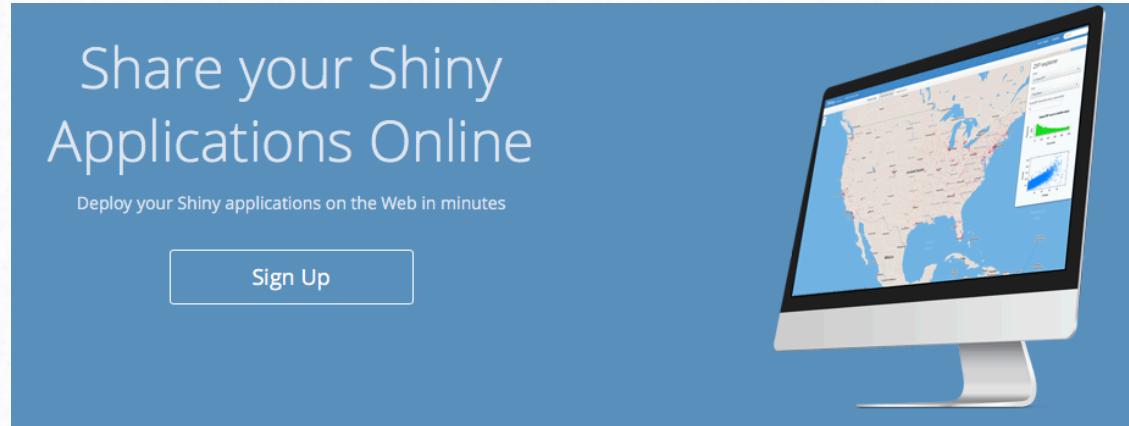
Sparkling Water

Enterprise grade machine learning pipelines.

Steam

Operationalize data science and dev ops with data products. Design and deploy AI apps quickly and easily.

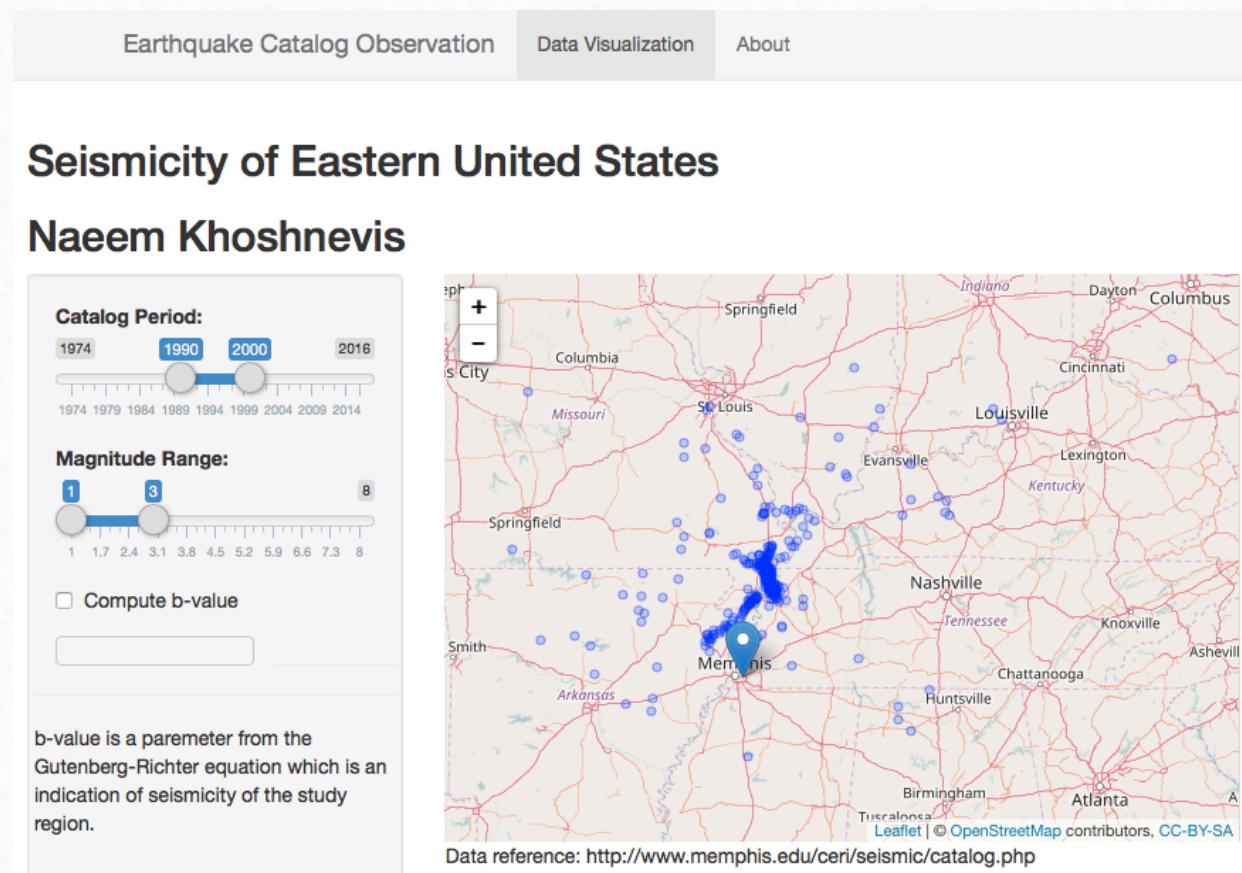
Developing Data Products (Shiny App)



- Shiny is a web application framework for R.
- Shiny allows you to create a graphical interface so that users can interact with your visualizations, models, and algorithms without needing to know R themselves.
- Shiny is developed at R Studio.
- It has two main files (Server.R and Ui.R)

<https://www.coursera.org/learn/data-products>

Developing Data Products (Shiny + Leaflet)



an open-source JavaScript library
for mobile-friendly interactive maps

https://naeem.shinyapps.io/shiny_app/

<https://github.com/Naeemkh/DDP>

Natural Language Processing (NLP)

Numerous number of NLP packages are submitted to CRAN.

(quanteda + shiny app) → Easier Typing application.

Developing steps:

- Preprocessing:
 - ✓ removing punctuation
 - ✓ removing numbers
 - ✓ removing URLs
 - ✓ removing profanity words
- Tokenizing
- Creating n-grams
- Developing the Shiny Application

Natural Language Processing (ShinyApp)

n-grams example:

Abraham Lincoln was an American politician and lawyer.

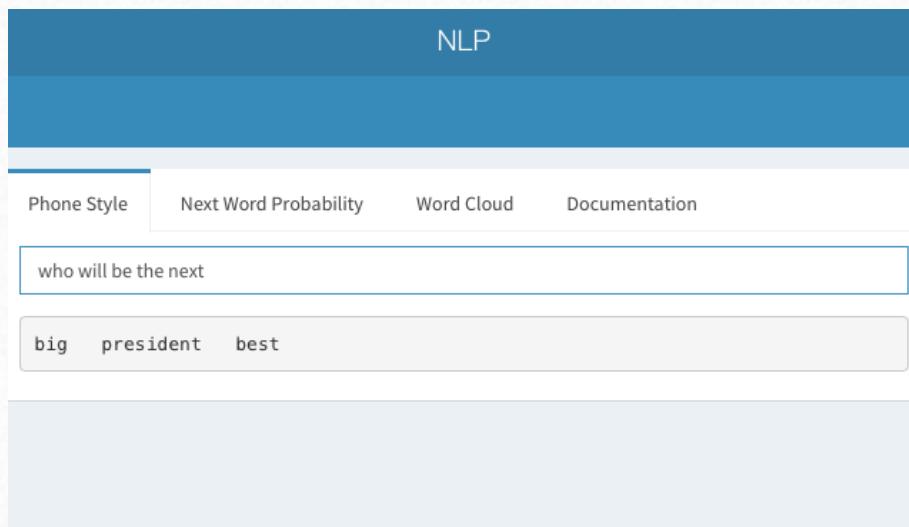
1-grams -> Abraham, Lincoln, was, ...

2-grams -> Abraham_Lincoln, Lincoln_was, was_an, ...

3-grams -> Abraham_Lincoln_was, Lincoln_was_an, ...

4-grams -> Abraham_Lincoln_was_an, ...

5-grams -> Abraham_Lincoln_was_an_American, ...



| 3-grams | frequency |
|--------------------|-----------|
| thank_for_the | 204746 |
| looking_forward_to | 77827 |
| thank_you_for | 75281 |
| i_love_you | 73548 |
| going_to_be | 73002 |

<https://naeem.shinyapps.io/shinyapp-NLP/>

<https://github.com/Naeemkh/DataScienceCapstone>

Revolution Analytics What is R?





An overview of R programming language

By : Naeem Khoshnevis
Center for Earthquake Research and Information
University of Memphis

May 2017