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)

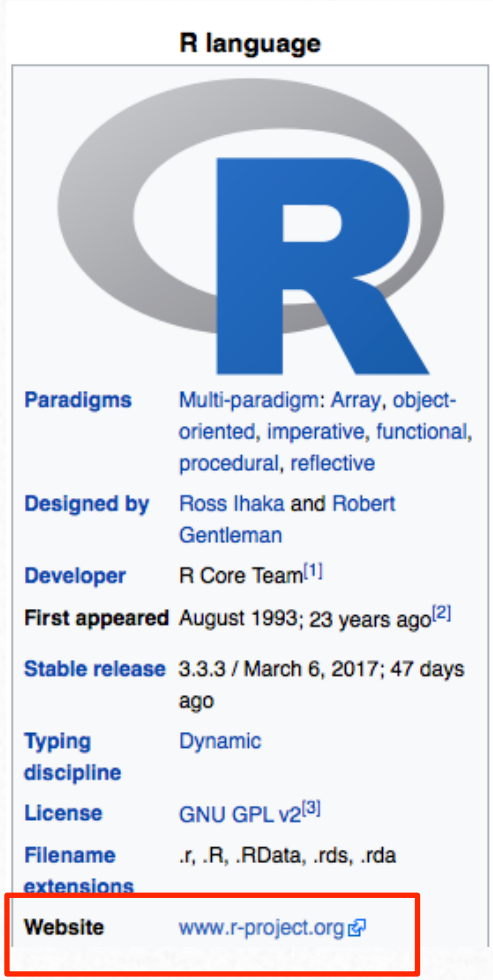
R language

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

https://en.wikipedia.org/wiki/R_(programming_language)

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)



https://en.wikipedia.org/wiki/R_(programming_language)

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/)

https://en.wikipedia.org/wiki/R_(programming_language)

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.

https://www.r-project.org, https://www.coursera.org/learn/r-programming

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
(KDNuggets survey, August 2013)

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

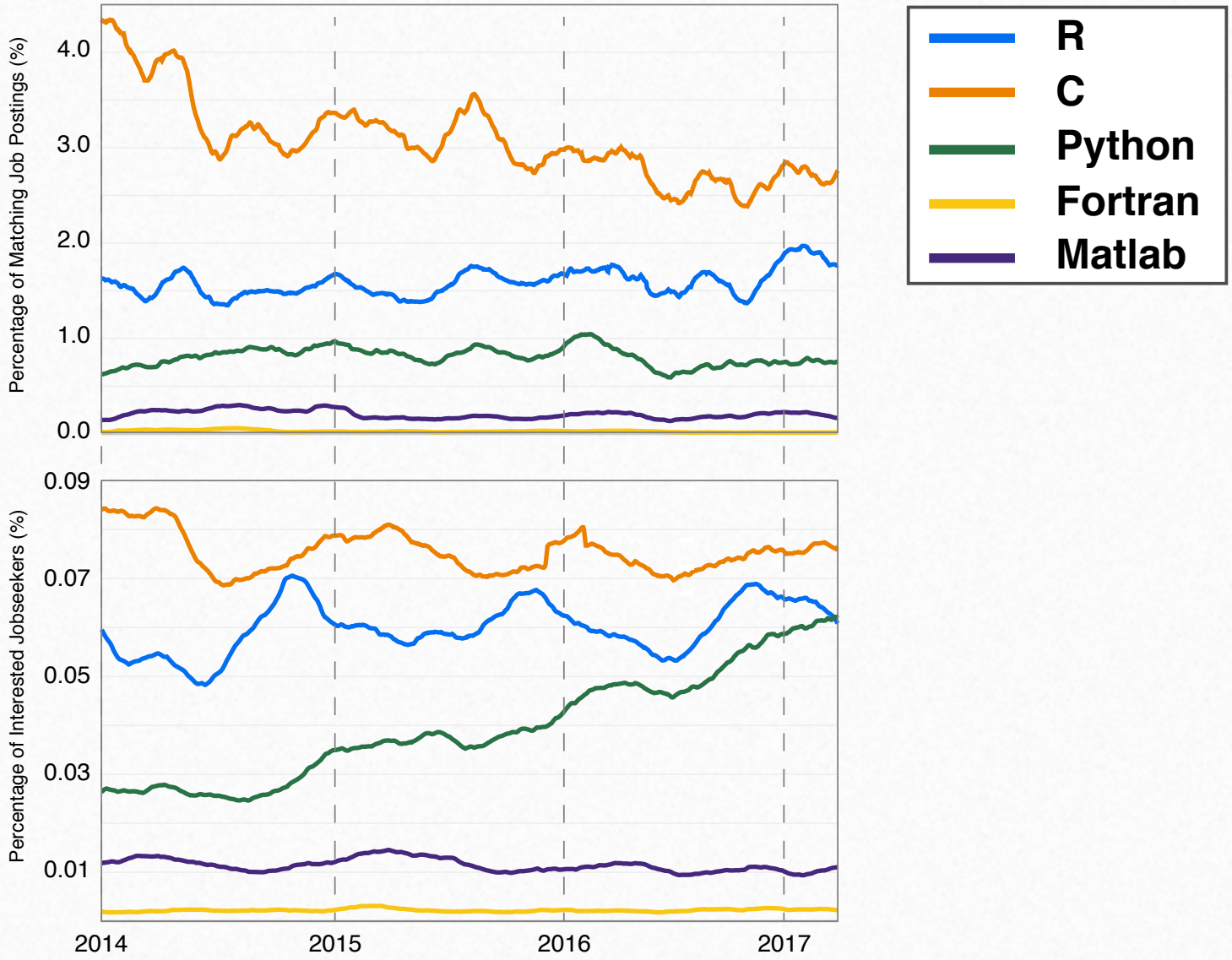
R has more than 2 million users worldwide
(Oracle estimate, February 2012)

http://blog.revolutionanalytics.com/2014/04/seven-quick-facts-about-r.html

R from job perspective

Job Posting

Job Seeker Interest

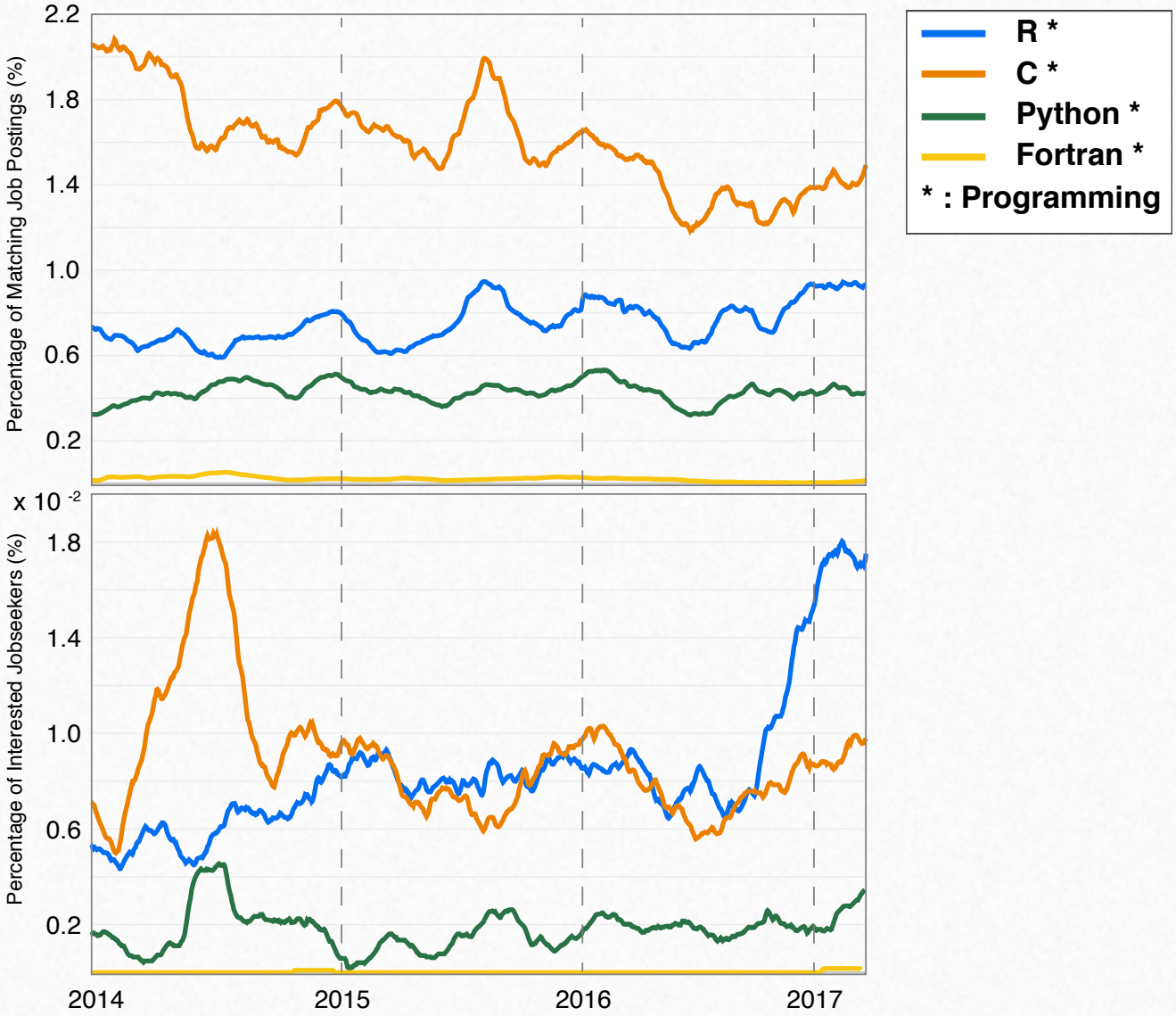


https://www.indeed.com/ (access 04/22/2017)

R from job perspective

Job Posting

Job Seeker Interest

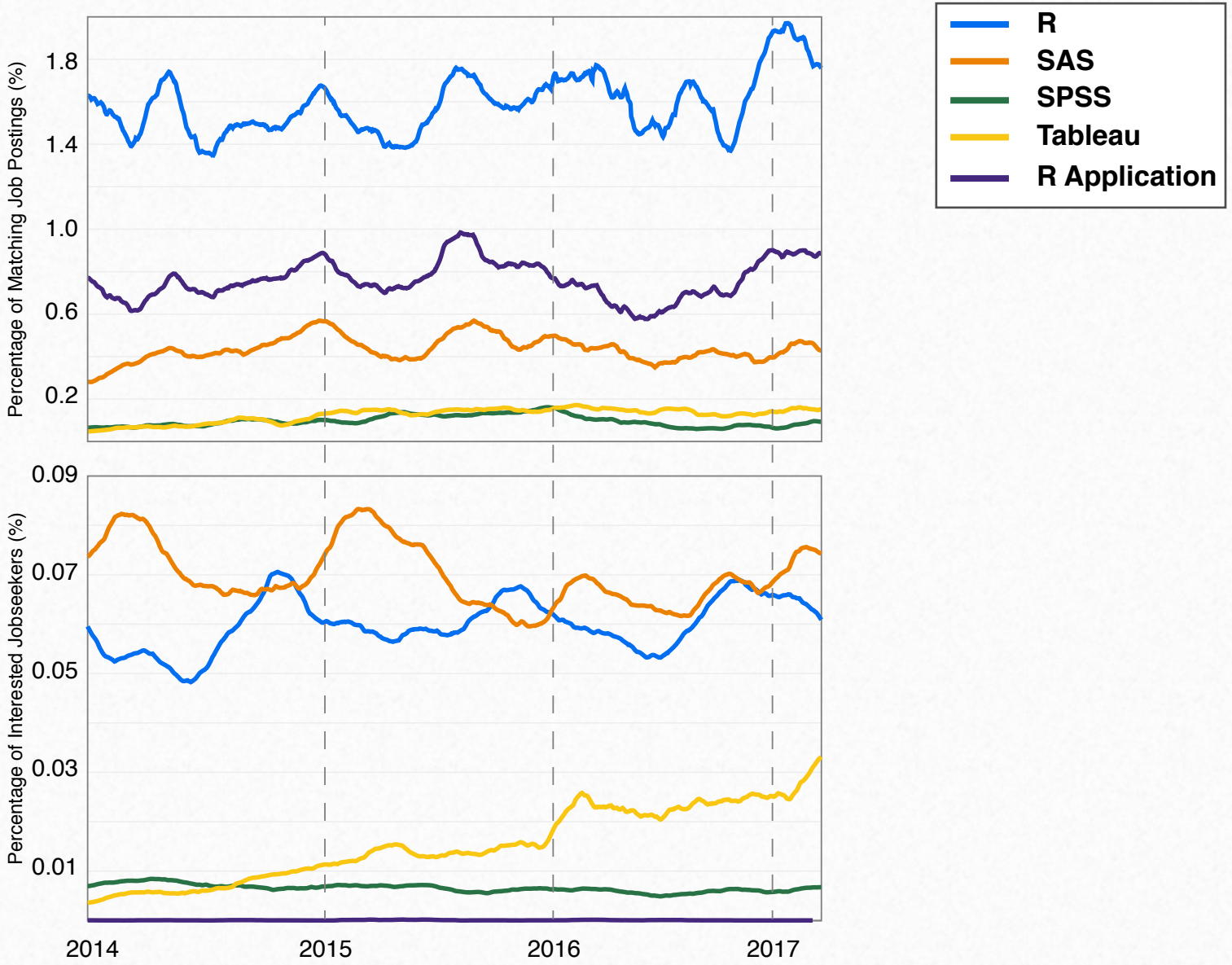


https://www.indeed.com/ (access 04/22/2017)

R from job perspective

Job Posting

Job Seeker Interest

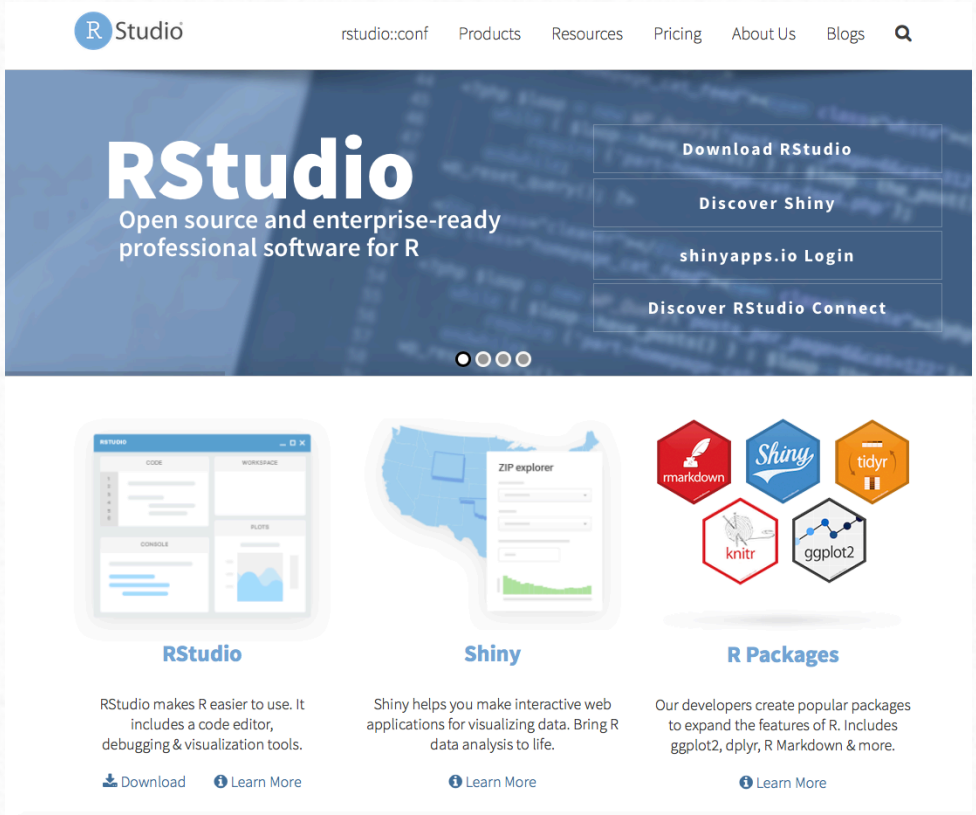


https://www.indeed.com/ (access 04/22/2017)

R – R studio Environment

Download R from: https://www.r-project.org

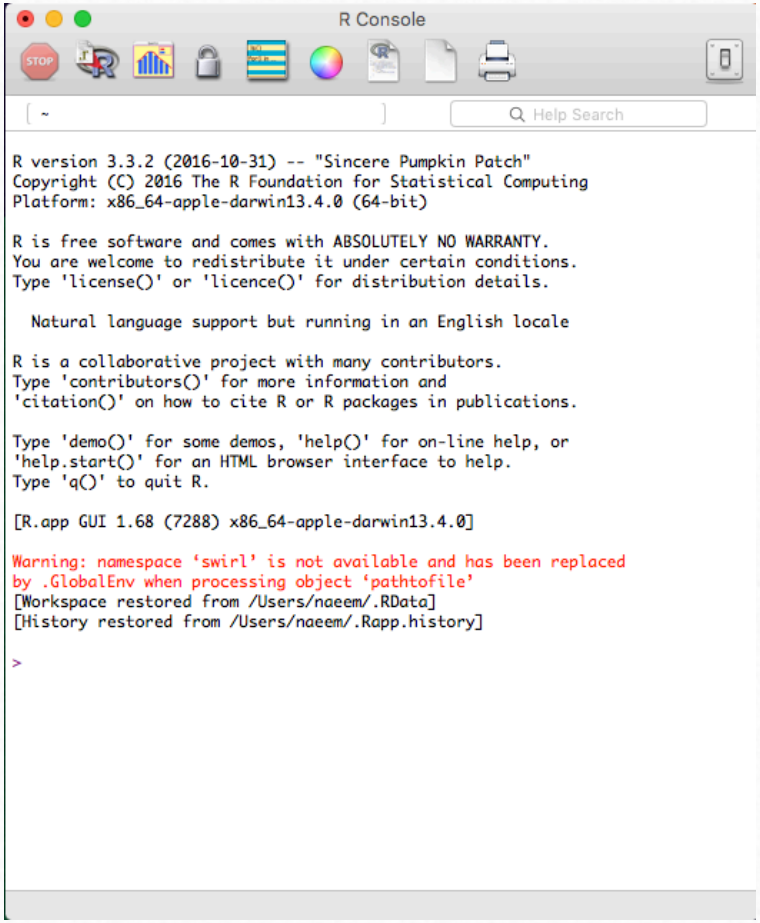
Download RStudio from: https://www.rstudio.com

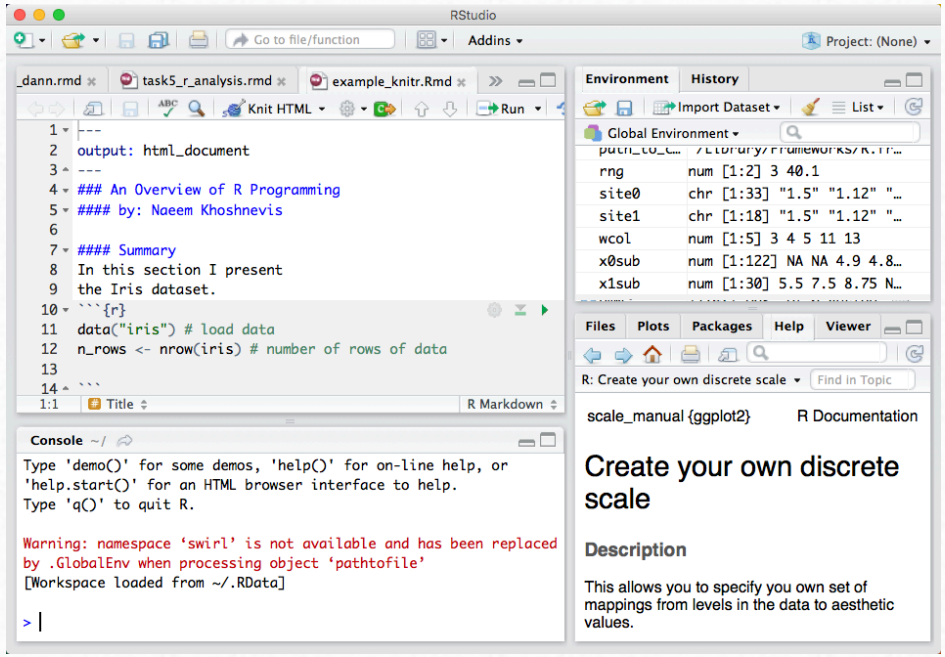


R – R studio Environment

Download R from: https://www.r-project.org

Download RStudio from: https://www.rstudio.com





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

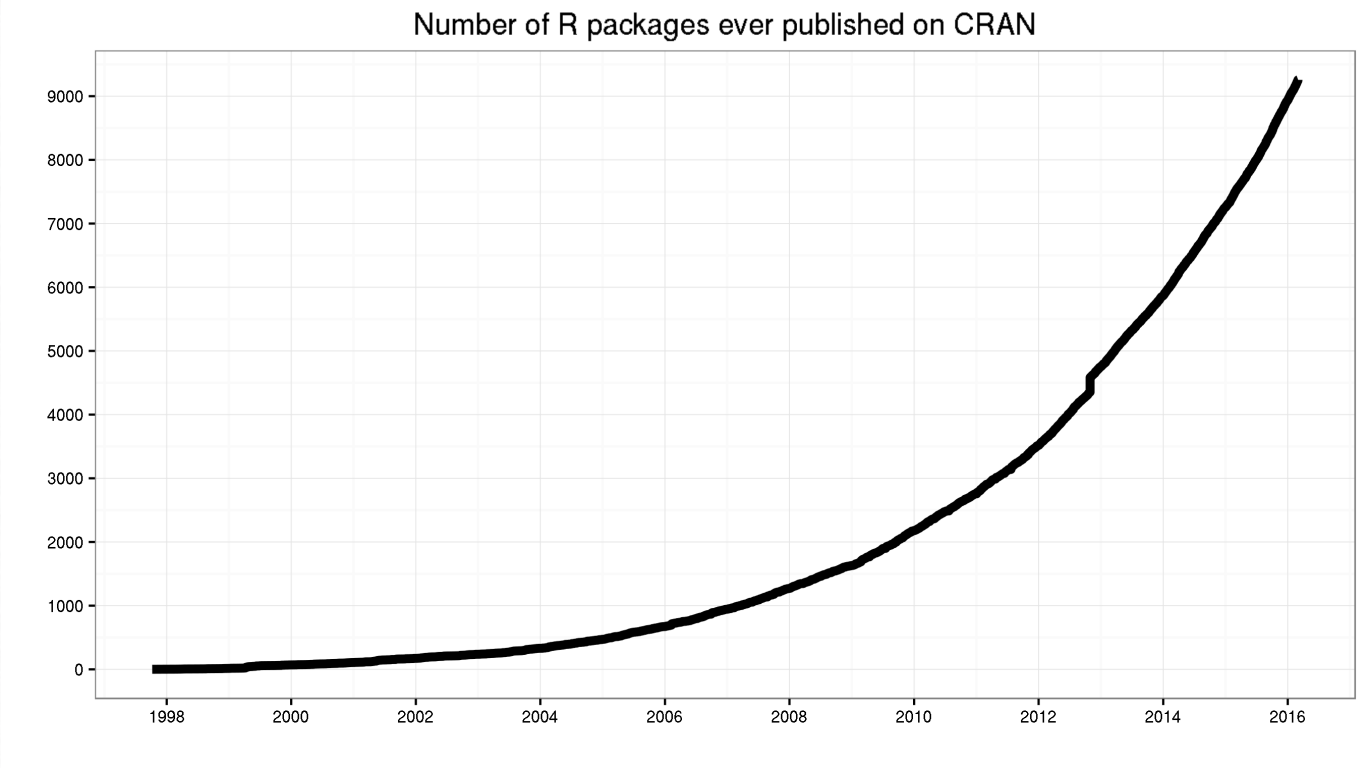
apply(data, margin, function)

What is R's strength?

https://gist.github.com/daroczig/3cf06d6db4be2bbe3368

What is R's strength?

R packages



https://gist.github.com/daroczig/3cf06d6db4be2bbe3368

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)

http://simplystatistics.org/2013/12/30/some-things-r-can-do-you-might-not-be-aware-of/

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.

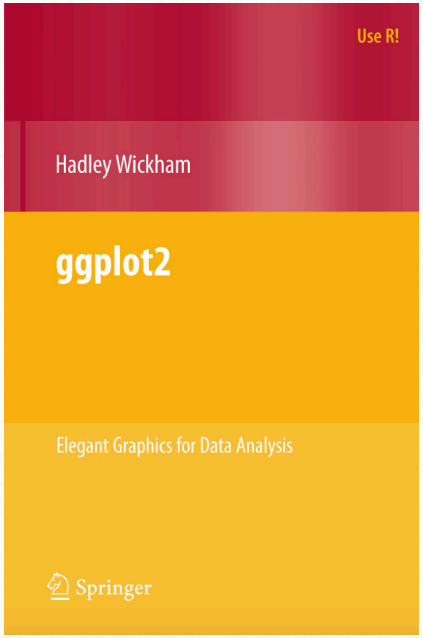
Thousands of packages are being downloaded on a daily basis.

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 varibale in the data are mapped to visual properties.

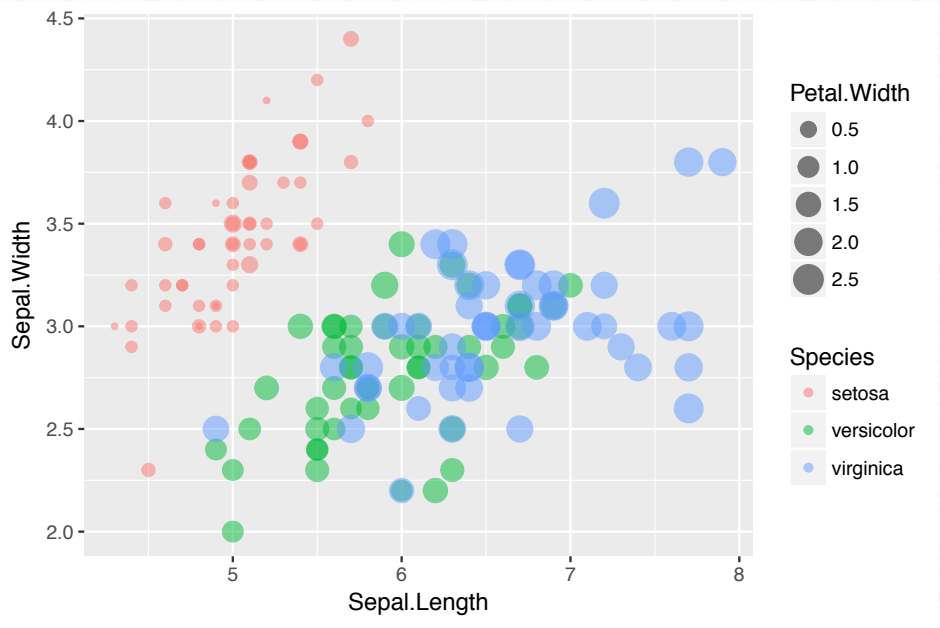
- Geometric objects
  geom_point, geom_smooth,geom_boxplot,geom_jitter, …

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)
```

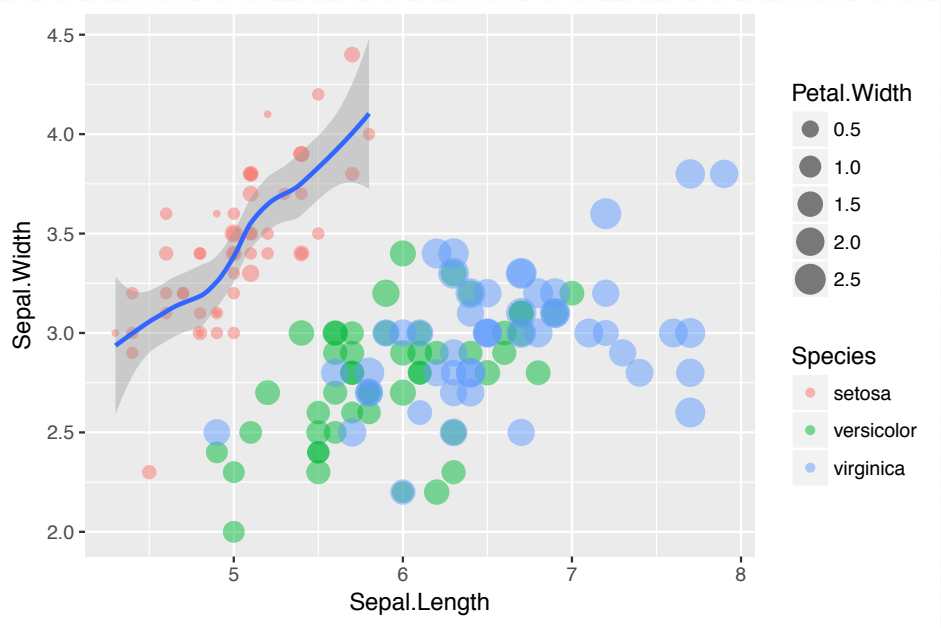
Data Visualization (ggplot)

```
> 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)
```



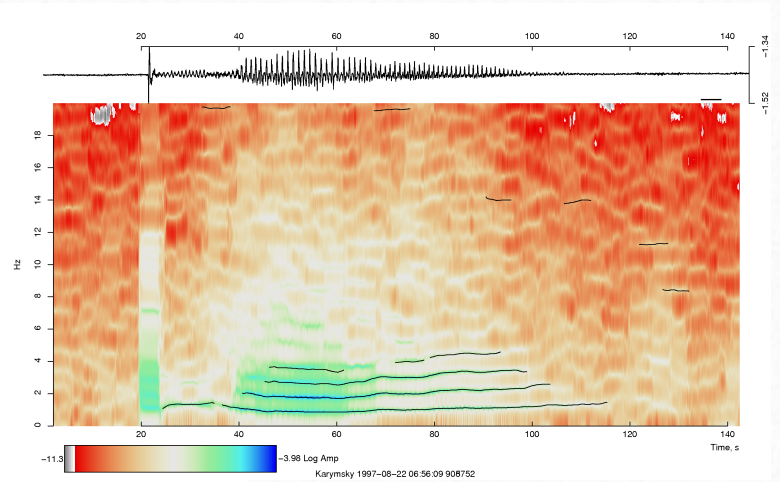
Data Visualization (ggplot)

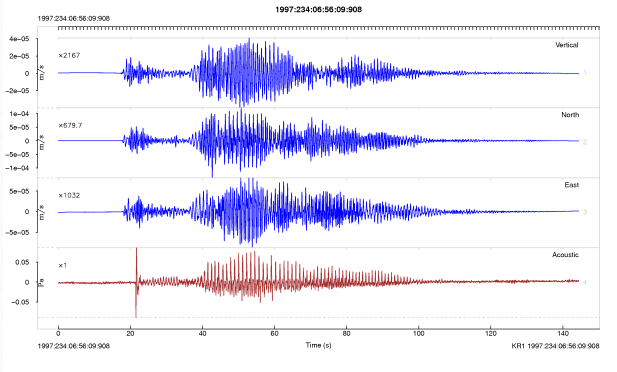
```
> library(ggplot2)
> data("iris")
> Sepal.Length Sepal.Width Petal.Length Petal.Width Species
          5.1           3.5          1.4           0.2       setosa
> g <- ggplot(iris)
> g <- g + geom_point(aes(x=Sepal.Length,y=Sepal.Width, color=Species, size
=Petal.Width), alpha=0.5)
> g <- g + geom_smooth(data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



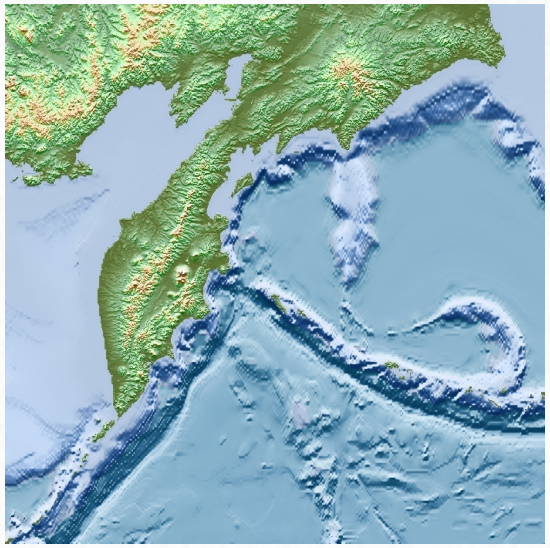


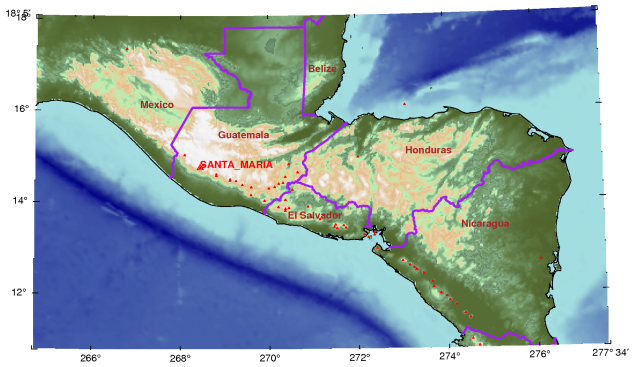
[Download from CRAN Website]

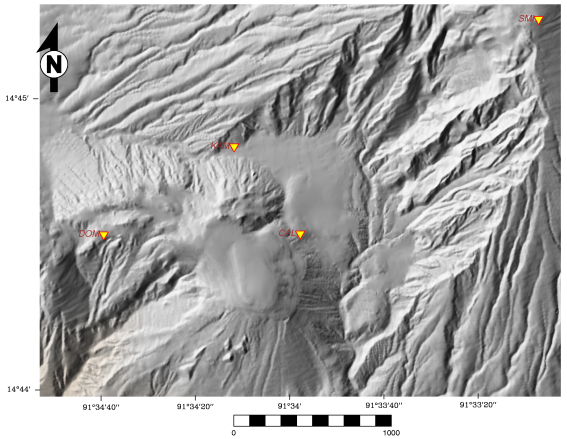
Slide from Jonathan Lees

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)

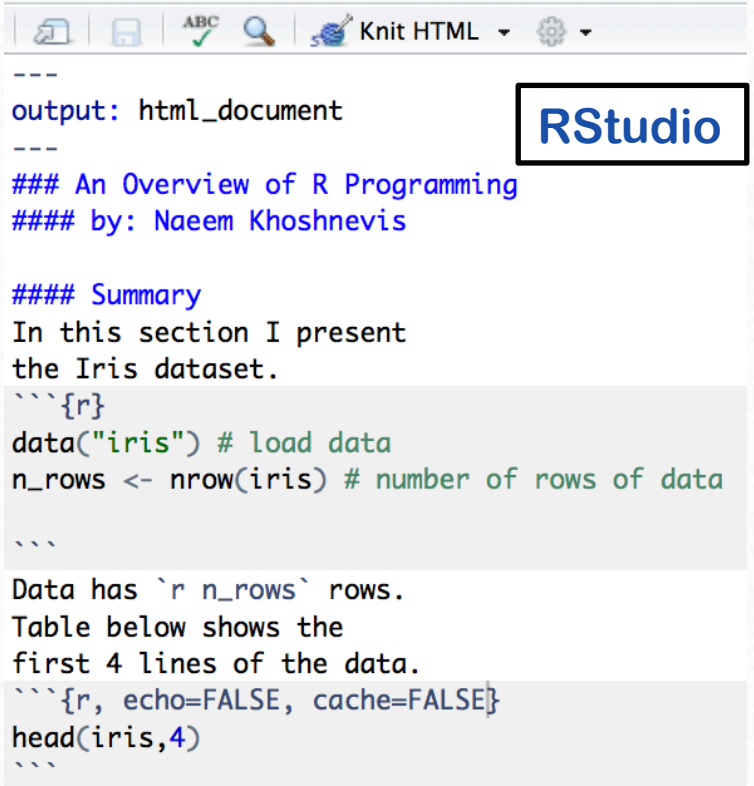
- Knitr is a package to literate (statistical) programming
- Knitr uses R as the programming language (although others are allowed) and variety of documentation languages:

  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




```
---
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

Knit HTML

RStudio

```
---
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)
```
```

HTML

An Overview of R Programming

by: Naeem Khoshnevis

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

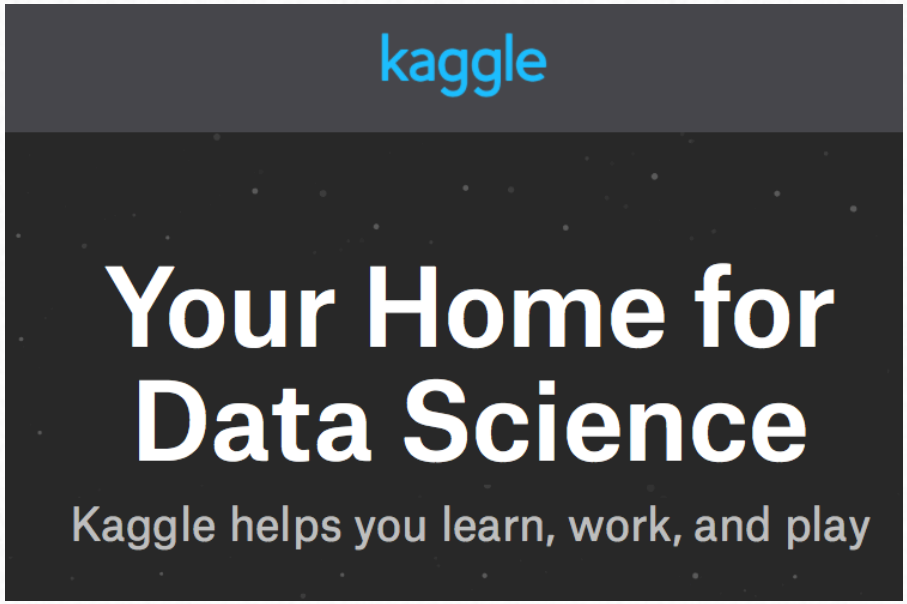
CARET is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

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

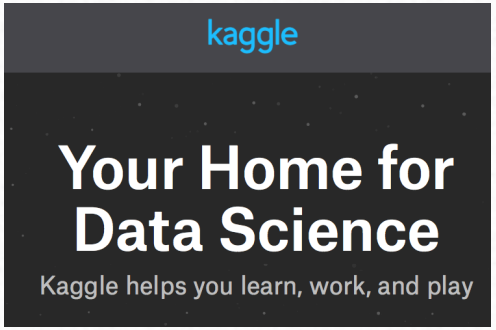
AdaBoost, Bagged AdaBoost, Bayesian Additive Regression Trees, C4.5-like Trees, Generalized Linear Model, k-Nearest Neighbors, Linear Regression, Multi-Layer Perceptron, Naive Bayes, Neural Network, Random Forest, …

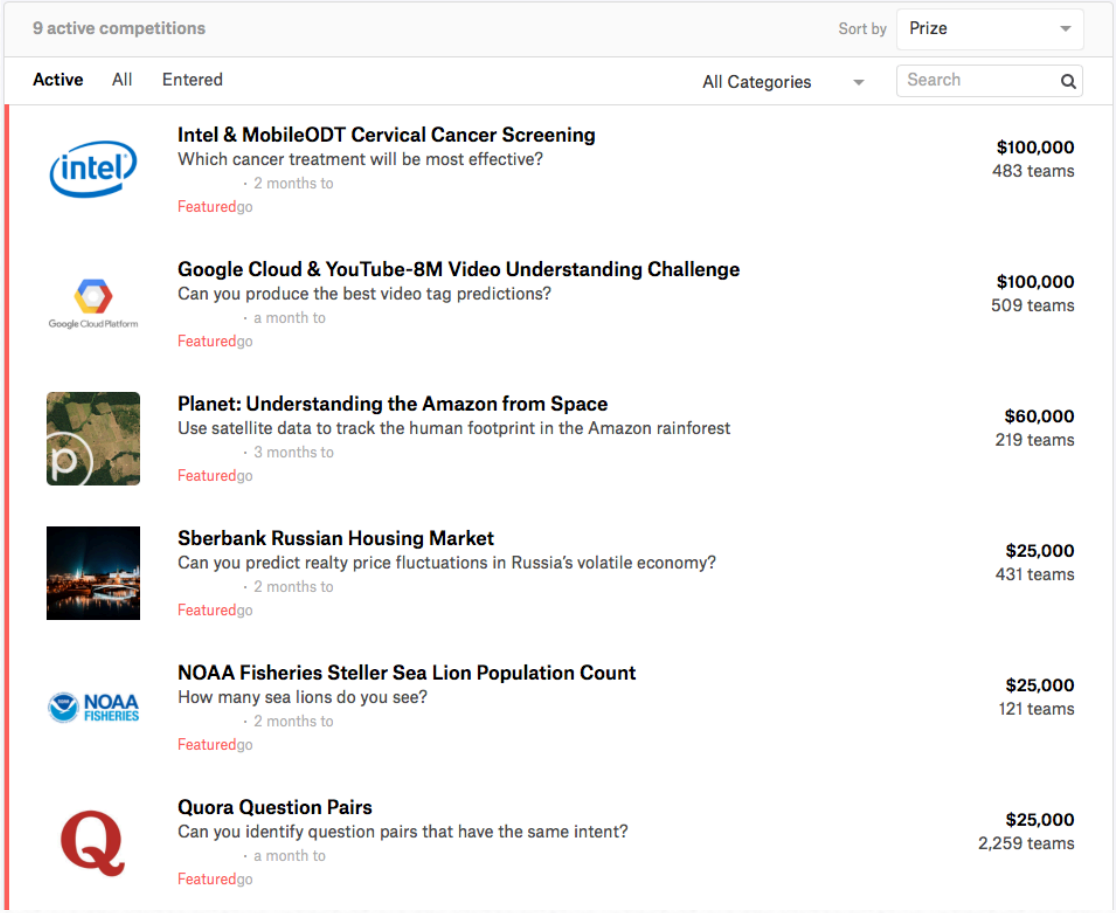
http://topepo.github.io/caret/index.html

Machine Learning with R (Kaggle)

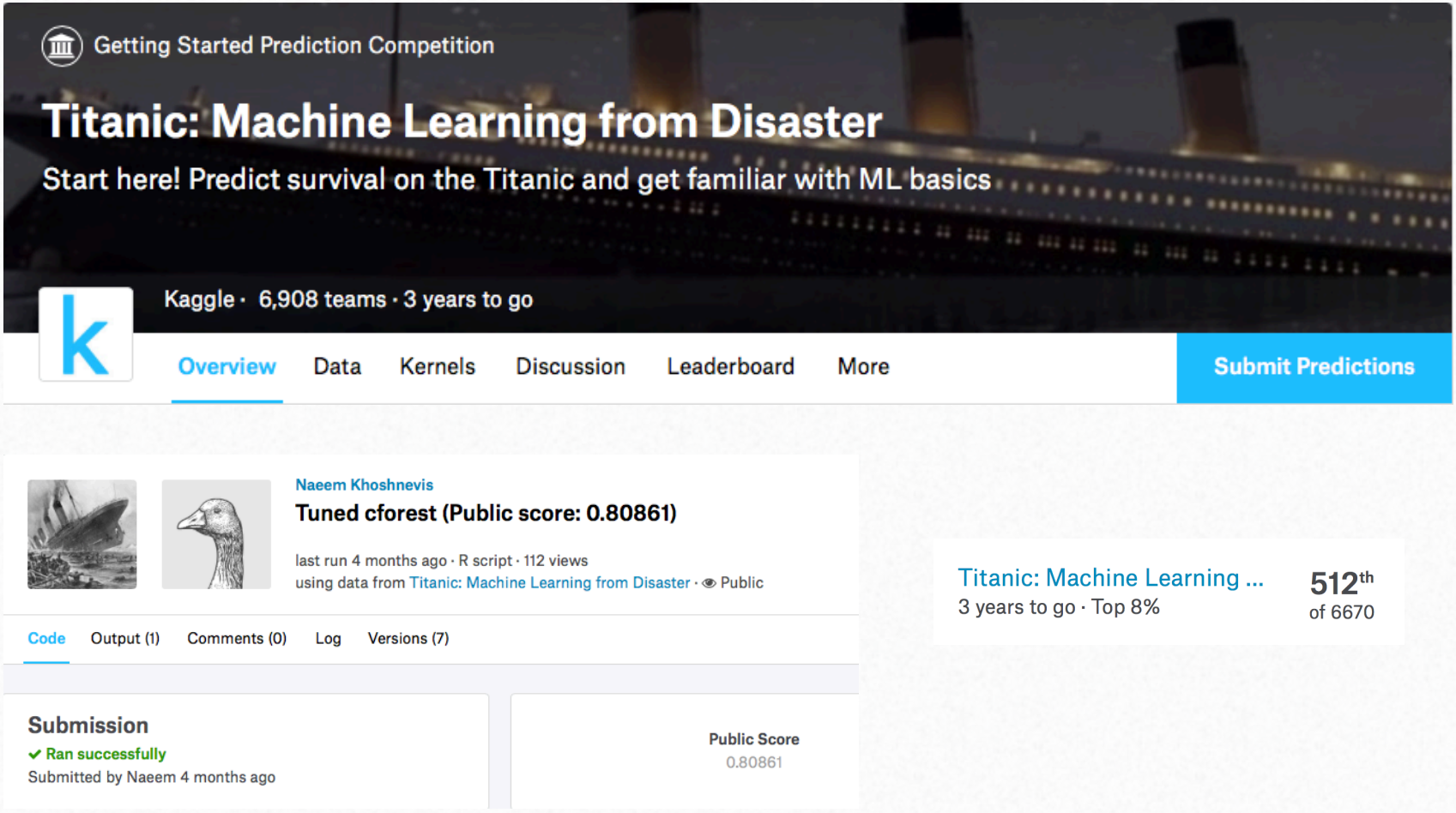


Machine Learning with R (Kaggle)





Machine Learning with R (Kaggle)



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

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.

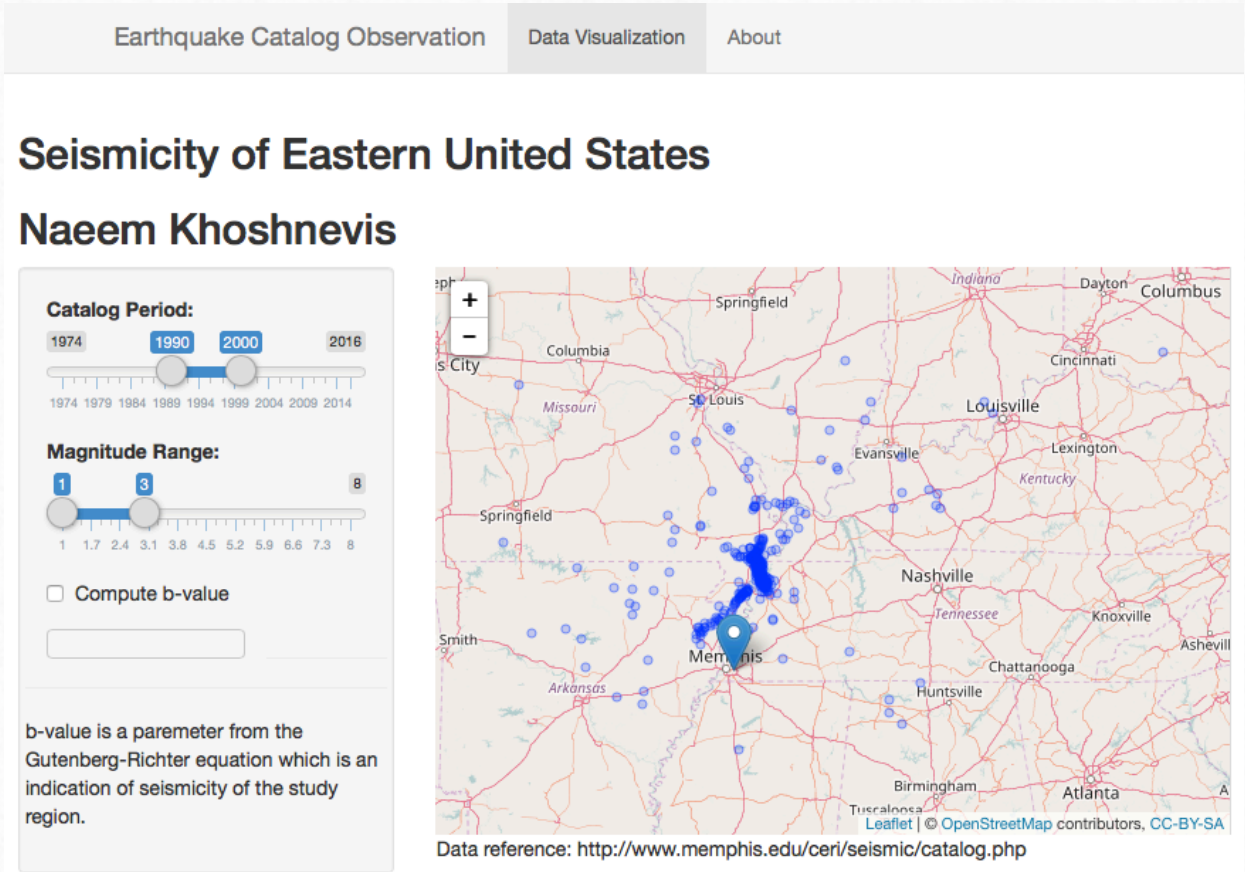
https://www.h2o.ai

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 needed to know R themselves.
- Shiny is developed at R Studio.
- It has two main file (Server.R and Ui.R)

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

Developing Data Products (Shiny + Leaflet)



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.

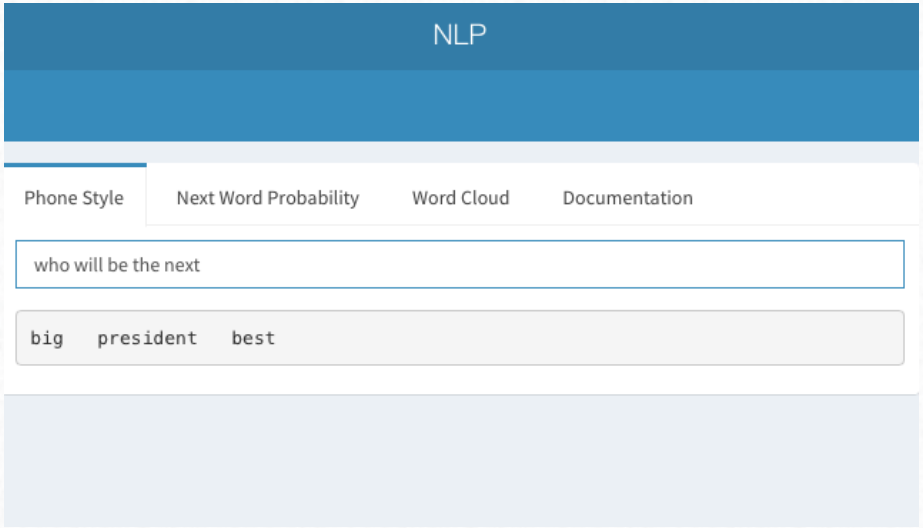
Developing steps:

- 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