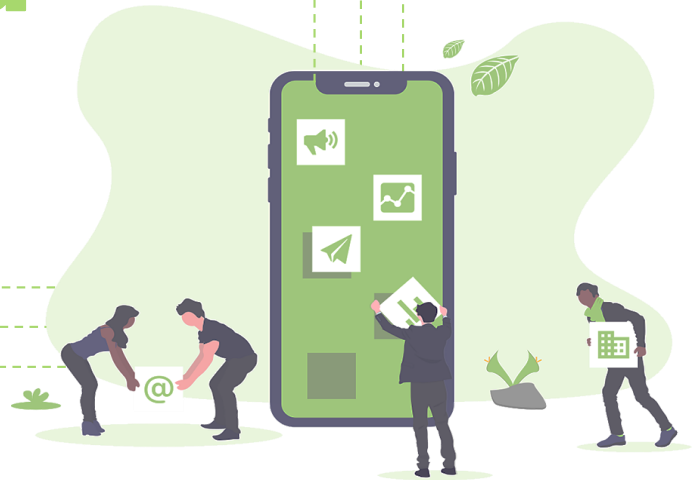


# The Data Science Track



Prepared By: R. Daynalo

1

1


## DATA SCIENCE Specialization

- an introduction to the key ideas behind working with data in a scientific way that will produce new and reproducible insight
- an introduction to the tools that will allow you to execute on a data analytic strategy, from raw data in a database to a completed report with interactive graphics
- on giving you plenty of hands and time on practice so you can learn the techniques for yourself


Prepared By: R. Daynalo

2


2




## Why do Data Science?



"It is not the critic who counts: not the man who points out how the strong man stumbles or where the doer of deeds could have done better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood, who strives valiantly, who errs and comes up short again and again, because there is no effort without error or shortcoming, but who knows the great enthusiasms, the great devotions, who spends himself for a worthy cause; who, at the best, knows, in the end, the triumph of high achievement, and who, at the worst, if he fails, at least he fails while daring greatly, so that his place shall never be with those cold and timid souls who knew neither victory nor defeat."




*Theodore Roosevelt, 26th President of the United States*




Prepared By: R. Daynola

3


3




## The Key Challenge in Data Science



"Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew all of the given information in advance? Where you didn't have a surplus information and have to filter it out, or you didn't have insufficient information to go find some?"



[Dan Myer, Mathematics Educator](#)





Prepared By: R. Daynola

4

4


# Why Data Science?

**The data deluge**

Technology

Businesses, governments and society are only starting to tap its vast potential



Print edition | Leaders >  
Feb 25th 2010

<https://www.economist.com/leaders/2010/02/25/the-data-deluge>

Prepared By: R. Daynola

5

5

# Why Data Science?

McKinsey Global Institute



May 2011

## Big data: The next frontier for innovation, competition, and productivity

<https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>

Prepared By: R. Daynola

6

6



# Why Statistical Data Science?





The New York Times



---

TECHNOLOGY

## For Today's Graduate, Just One Word: Statistics

By STEVE LOHR AUG. 5, 2009








<https://www.nytimes.com/2009/08/06/technology/06stats.html? r=0>


Prepared By: R. Daynolo


7

7

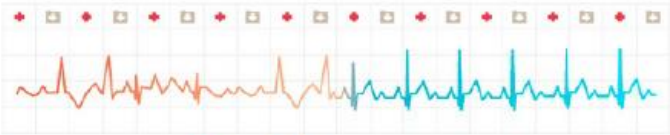


# Why are you lucky?





Information Data Forum Leaderboard



## Improve Healthcare, Win \$3,000,000.


COMPETITION GOAL

Identify patients who will be admitted to a hospital within the next year, using historical claims data.


Prepared By: R. Daynolo


8

8



# Why R?






BUSINESS COMPUTING

## Data Analysts Captivated by R's Power

By ASHLEE VANCE JAN. 6, 2009

<https://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all>

See how this article appeared when it was originally published on NYTimes.com



R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

Left, Stuart Isett for The New York Times; right, Kieran Scott for The New York Times

Prepared By: R. Daynalo

9

9



# Why R?




- It is free
- It has a comprehensive set of packages
  - Data access
  - Data cleaning
  - Analysis
  - Data reporting
- It has one of the best development environments – RStudio
- It has amazing ecosystem of developers
- Packages are easy to install and “play nicely together”

Prepared By: R. Daynalo

10

10




## Who is a Data Scientist?

- Data scientists are a new breed of analytical data expert who have the technical skills to solve complex problems – and the curiosity to explore what problems need to be solved.

Prepared By: R. Daynalo

11

11



## Who is a Data Scientist?

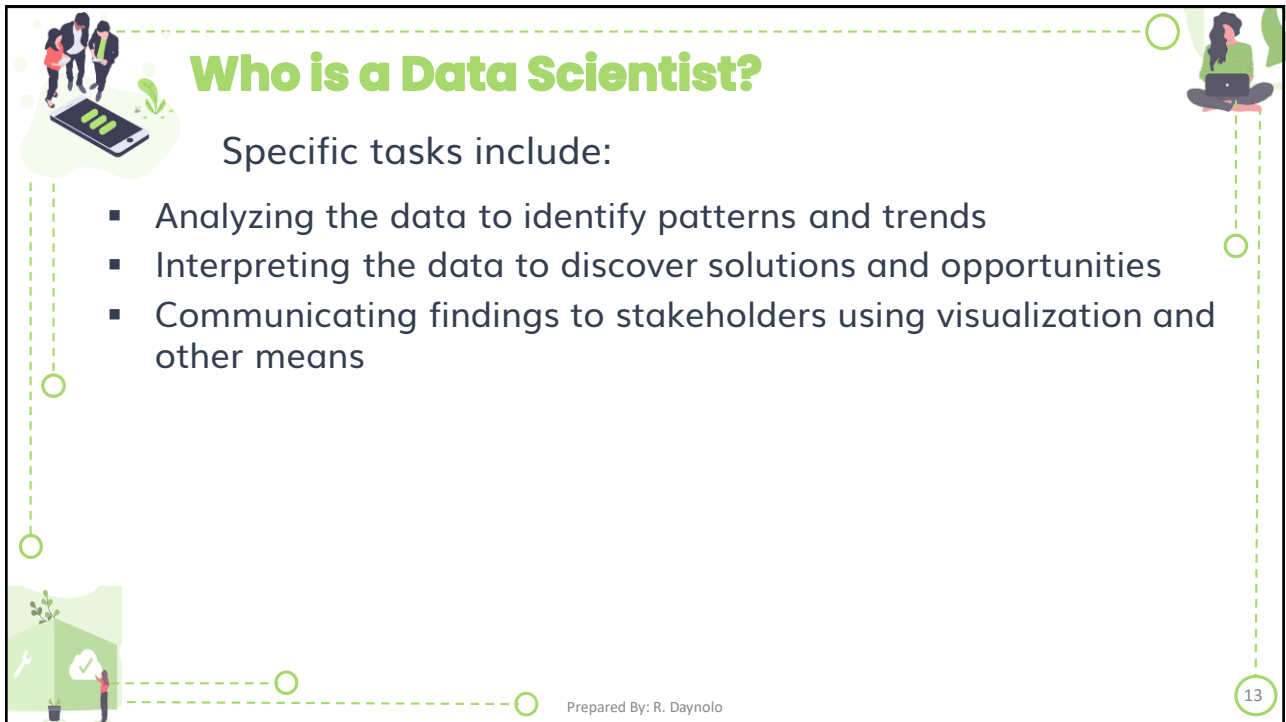
Specific tasks include:

- Identifying the data-analytics problems that offer the greatest opportunities to the organization
- Determining the correct data sets and variables
- Collecting large sets of structured and unstructured data from disparate sources
- Cleaning and validating the data to ensure accuracy, completeness, and uniformity
- Devising and applying models and algorithms to mine the stores of big data

Prepared By: R. Daynalo

12

12



## Who is a Data Scientist?

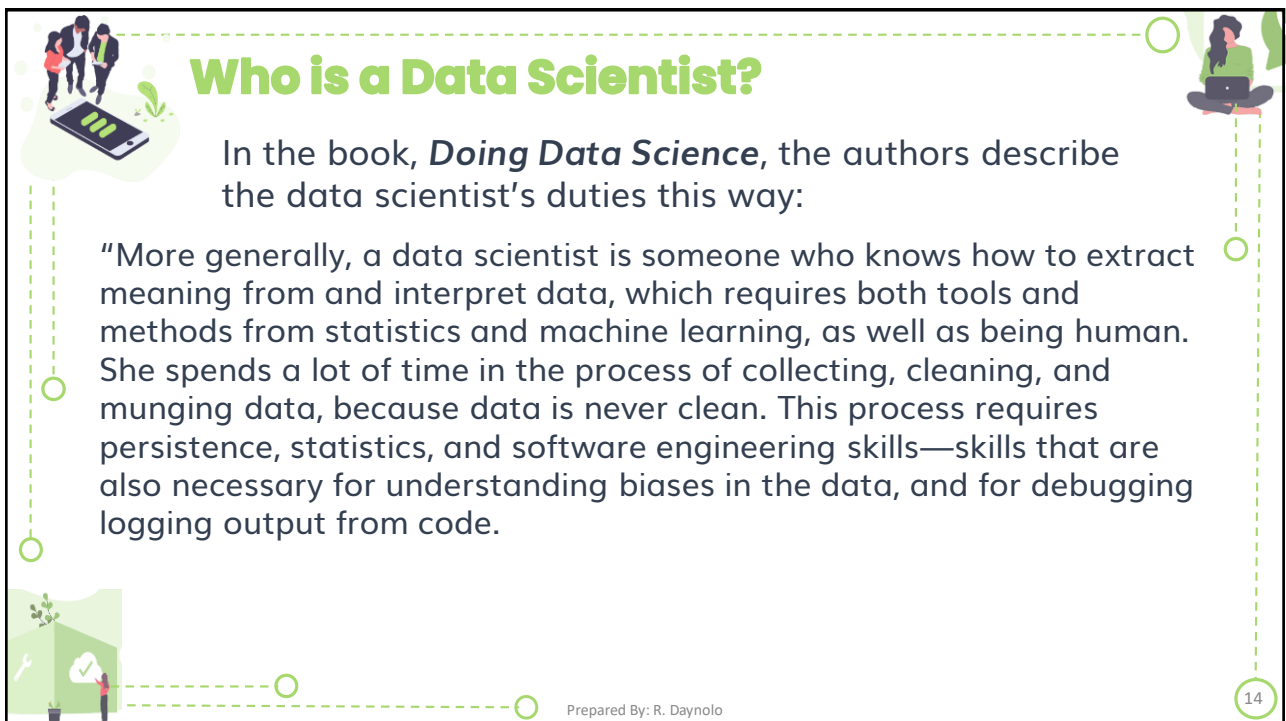
Specific tasks include:

- Analyzing the data to identify patterns and trends
- Interpreting the data to discover solutions and opportunities
- Communicating findings to stakeholders using visualization and other means

Prepared By: R. Daynolo

13

13



## Who is a Data Scientist?

In the book, *Doing Data Science*, the authors describe the data scientist's duties this way:

"More generally, a data scientist is someone who knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning, as well as being human. She spends a lot of time in the process of collecting, cleaning, and munging data, because data is never clean. This process requires persistence, statistics, and software engineering skills—skills that are also necessary for understanding biases in the data, and for debugging logging output from code.

Prepared By: R. Daynolo

14

14



## Who is a Data Scientist?

In the book, *Doing Data Science*, the authors describe the data scientist's duties this way:

Once she gets the data into shape, a crucial part is exploratory data analysis, which combines visualization and data sense. She'll find patterns, build models, and algorithms—some with the intention of understanding product usage and the overall health of the product, and others to serve as prototypes that ultimately get baked back into the product. She may design experiments, and she is a critical part of data-driven decision making. She'll communicate with team members, engineers, and leadership in clear language and with data visualizations so that even if her colleagues are not immersed in the data themselves, they will understand the implications."

Source: O'Neil, C., and Schutt, R. *Doing Data Science*. First edition.

Prepared By: R. Daynolo

15

15

## Who is a Data Scientist?



Nate Silver, Founder and Editor-in-Chief of FiveThirtyEight



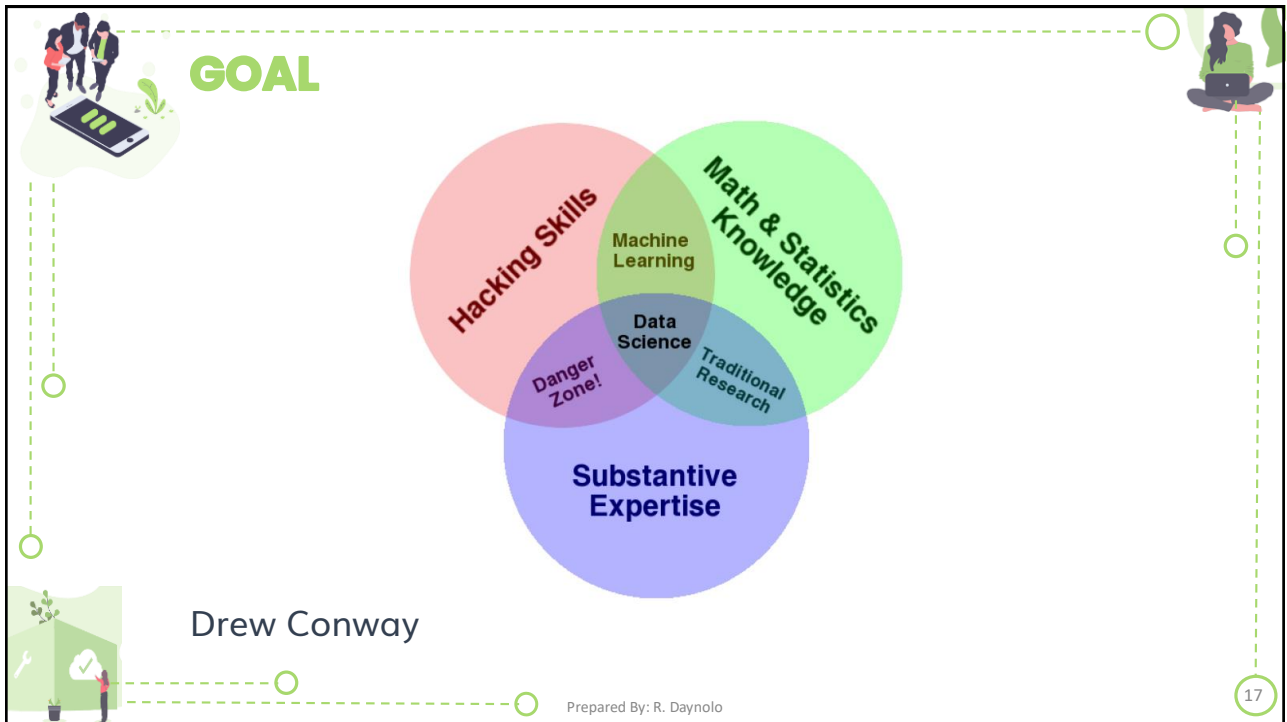
Daphne Koller, CEO of Coursera

Prepared By: R. Daynolo

16

16





17

**JOBS**

21,726 views | Dec 11, 2017, 08:32pm

# LinkedIn's Fastest-Growing Jobs Today Are In Data Science And Machine Learning

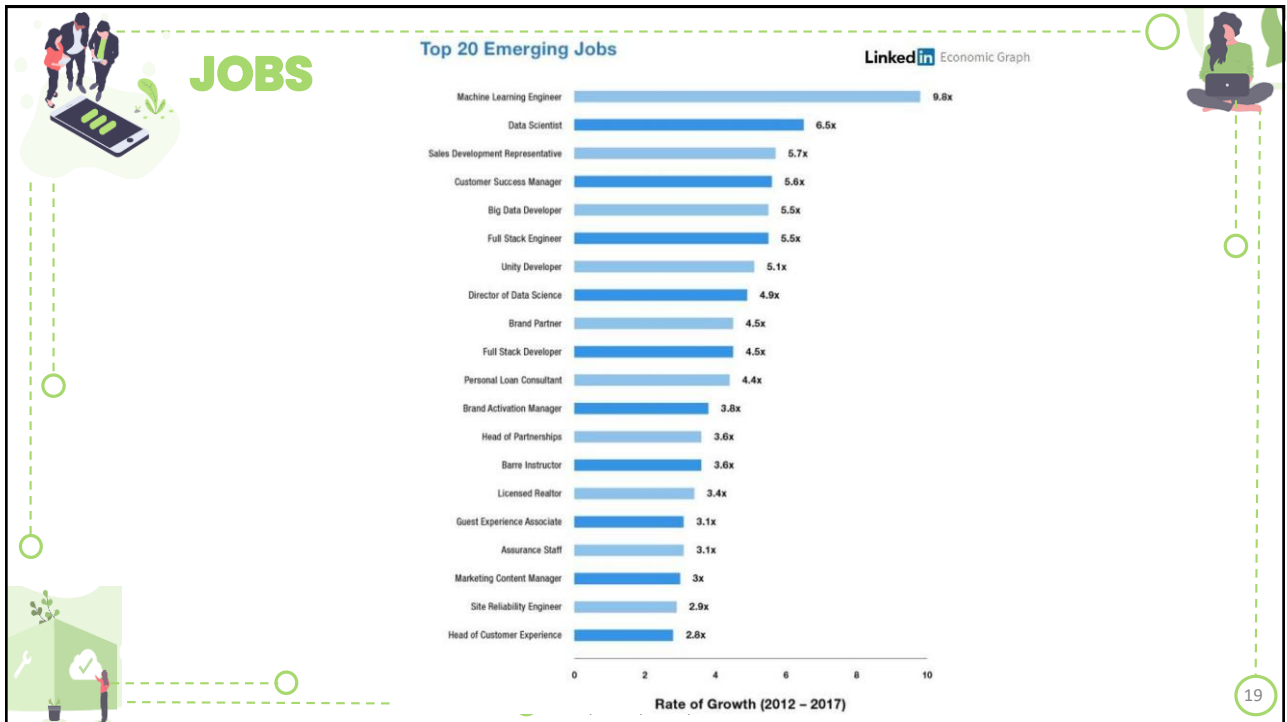
**Louis Columbus** Contributor ⓘ

<https://www.forbes.com/sites/louiscolumnbus/2017/12/11/linkedins-fastest-growing-jobs-today-are-in-data-science-machine-learning/#634af74c51bd>

Prepared By: R. Daynola

18



18



19

# 1. The Data Scientists' Toolbox

20



## What do Data Scientists do?

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code
- Distribute results to other people

Prepared By: R. Daynola

21

21

## The main workhorse of Data Science

### The R Project for Statistical Computing

#### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

#### News

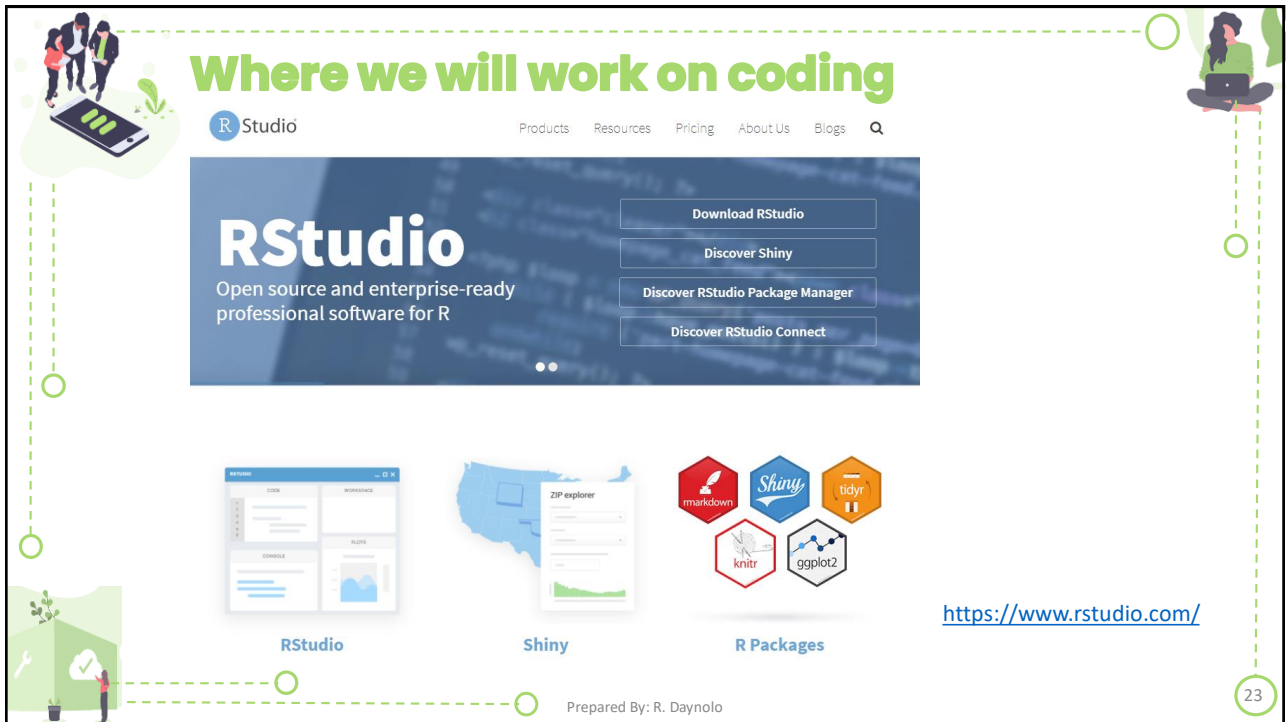
- [R version 3.5.2 \(Eggshell Igloo\)](#) has been released on 2018-12-20.
- The R Foundation Conference Committee has released a [call for proposals](#) to host useR! 2020 in North America.
- You can now support the R Foundation with a renewable subscription as a [supporting member](#)
- The R Foundation has been awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers.

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

Prepared By: R. Daynola

22

22



**Where we will work on coding**

RStudio

Products Resources Pricing About Us Blogs

**RStudio**

Open source and enterprise-ready professional software for R

Download RStudio

Discover Shiny

Discover RStudio Package Manager

Discover RStudio Connect

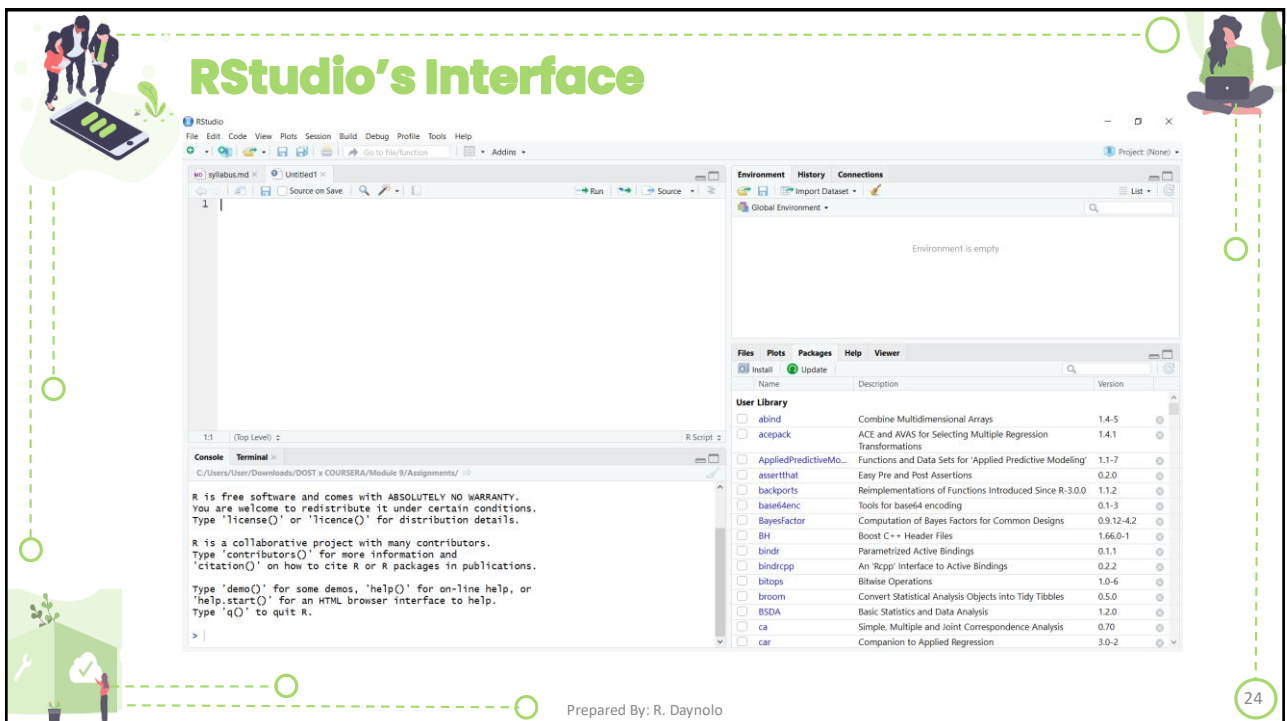
RStudio Shiny R Packages

<https://www.rstudio.com/>

Prepared By: R. Daynalo

23

23



**RStudio's Interface**

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

1:1 (Top Level) z

Console Terminal

C:\Users\User\Downloads\DOST > COURSE/Module 9/Assignments/

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.

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.

Environment History Connections

Global Environment

Environment is empty

Files Plots Packages Help Viewer

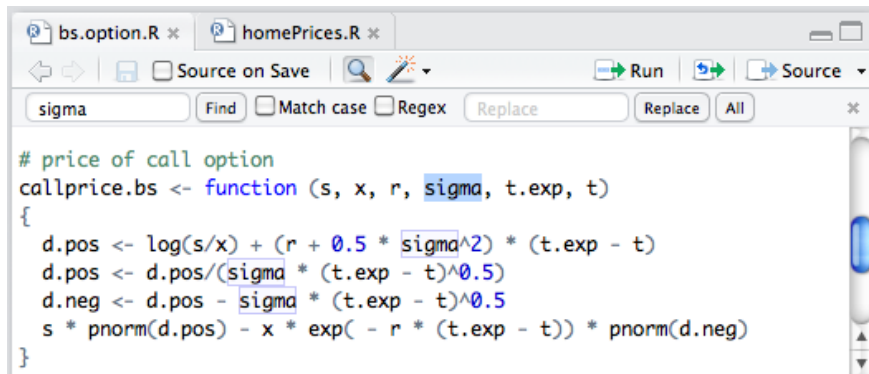
Name	Description	Version
abind	Combine Multidimensional Arrays	1.4-5
acepack	ACE and AVAS for Selecting Multiple Regression	1.4-1
AppliedPredictiveMo...	Functions and Data Sets for 'Applied Predictive Modeling'	1.1-7
assertthat	Easy Pre and Post Assertions	0.2.0
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.2
base64enc	Tools for base64 encoding	0.1-3
Bayesfactor	Computation of Bayes Factors for Common Designs	0.9.12-4.2
BH	Boost C++ Header Files	1.66.0-1
bindr	Parameterized Active Bindings	0.1.1
bindrcpp	An 'Rcpp' interface to Active Bindings	0.2.2
bitops	Bitwise Operations	1.0-6
broom	Convert Statistical Analysis Objects into Tidy Tibbles	0.5.0
BSDA	Basic Statistics and Data Analysis	1.2.0
ca	Simple, Multiple and Joint Correspondence Analysis	0.7.0
car	Companion to Applied Regression	3.0-2

Prepared By: R. Daynalo

24

24

## Primary File Types – R Script



```
bs.option.R * homePrices.R *
Source on Save Find Match case Regex Replace Replace All
sigma
# price of call option
callprice.bs <- function (s, x, r, sigma, t.exp, t)
{
  d.pos <- log(s/x) + (r + 0.5 * sigma^2) * (t.exp - t)
  d.pos <- d.pos/(sigma * (t.exp - t)^0.5)
  d.neg <- d.pos - sigma * (t.exp - t)^0.5
  s * pnorm(d.pos) - x * exp(- r * (t.exp - t)) * pnorm(d.neg)
}
```

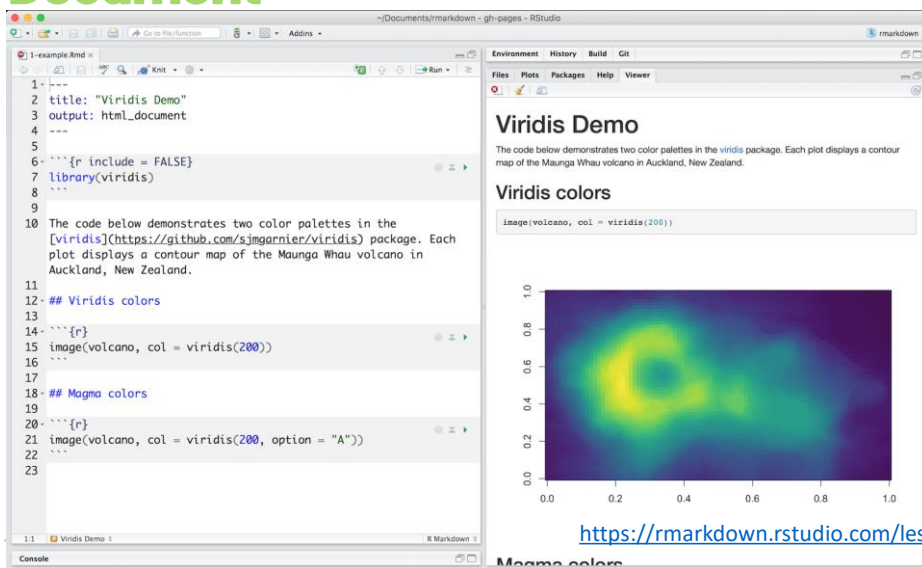
<https://support.rstudio.com/hc/en-us/articles/200484448-Editing-and-Executing-Code>

Prepared By: R. Daynola

25

25

## Primary File Types – R Markdown Document



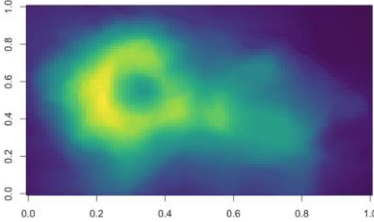
```
1- example.Rmd
2 title: "Viridis Demo"
3 output: html_document
4 ---
5
6 {r include = FALSE}
7 library(viridis)
8
9
10 The code below demonstrates two color palettes in the
11 [viridis](https://github.com/sjmgarnier/viridis) package. Each
12 plot displays a contour map of the Maunga Whau volcano in
13 Auckland, New Zealand.
14
15 ## Viridis colors
16 {r}
17 image(volcano, col = viridis(200))
18
19 ## Magma colors
20 {r}
21 image(volcano, col = viridis(200, option = "A"))
22
23
```

**Viridis Demo**

The code below demonstrates two color palettes in the `viridis` package. Each plot displays a contour map of the Maunga Whau volcano in Auckland, New Zealand.

**Viridis colors**

```
image(volcano, col = viridis(200))
```



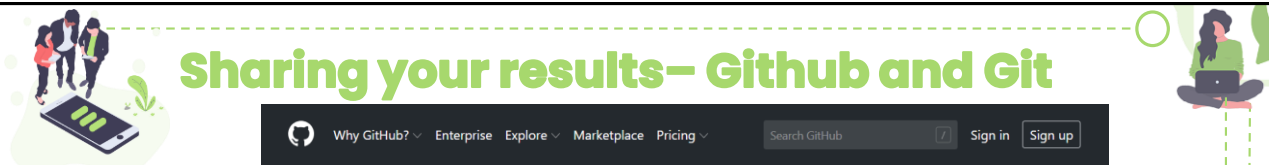
**Magma colors**

<https://rmarkdown.rstudio.com/lesson-2.html>

26

26

# Sharing your results– Github and Git



<https://github.com/>

**Built for developers**

GitHub is a development platform inspired by the way you work. From **open source** to **business**, you can host and review code, manage projects, and build software alongside 31 million developers.

Username  
Pick a username

Email  
you@example.com

Password  
Create a password

Make sure it's more than 15 characters OR at least 8 characters including a number and a lowercase letter.  
Read our documentation on [safer password practices](#).

**Sign up for GitHub**

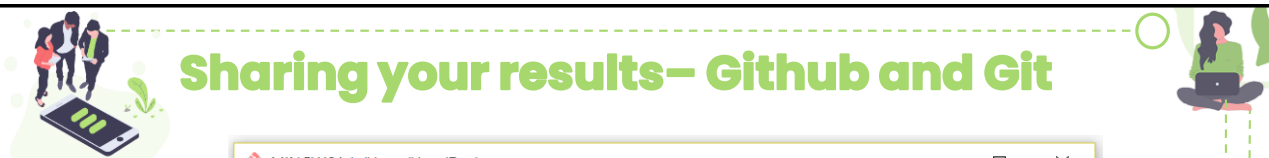
By clicking "Sign up for GitHub", you agree to our [terms of service](#) and [privacy statement](#). We'll occasionally send you account related emails.

Prepared By: R. Daynolo

27

27

# Sharing your results– Github and Git



MINGW64; c:/Users/User/Desktop

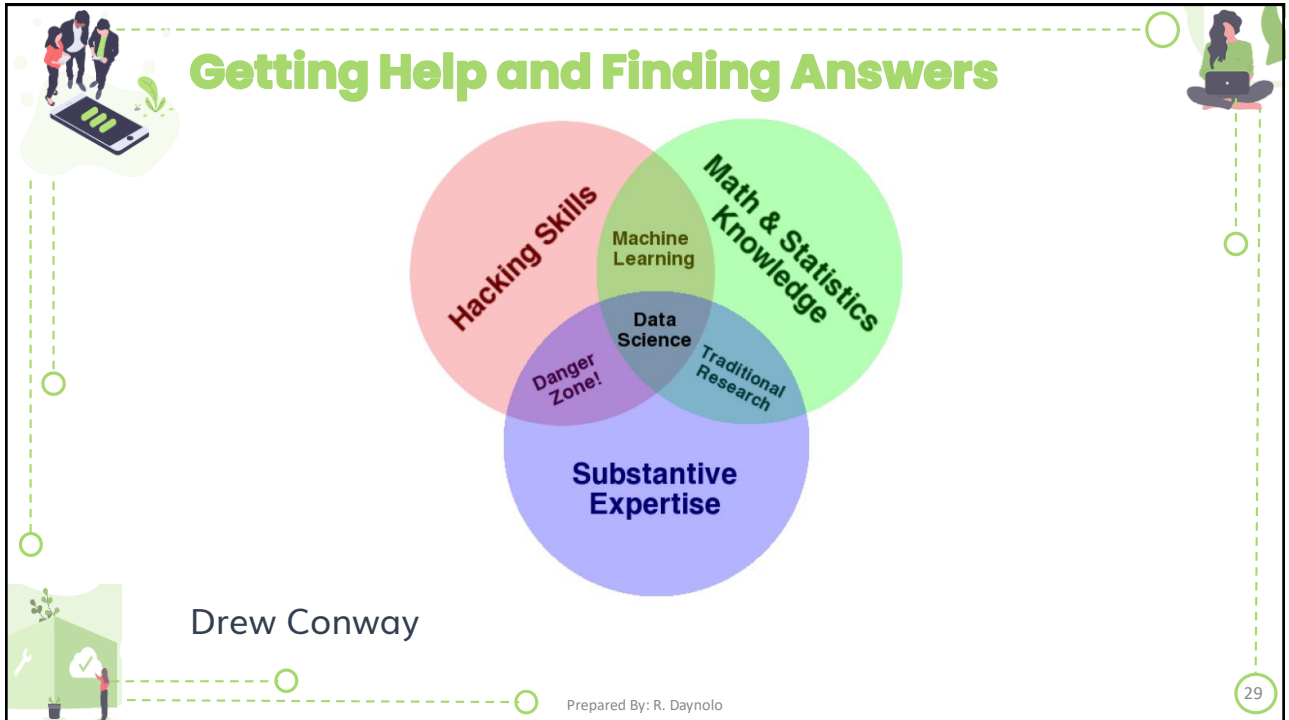
Raquel A. Daynolo@RAQUEL MINGW64 ~/Desktop

\$ |

Prepared By: R. Daynolo

28

28



29

## Getting Help and Finding Answers

Key Characteristics of Hackers:

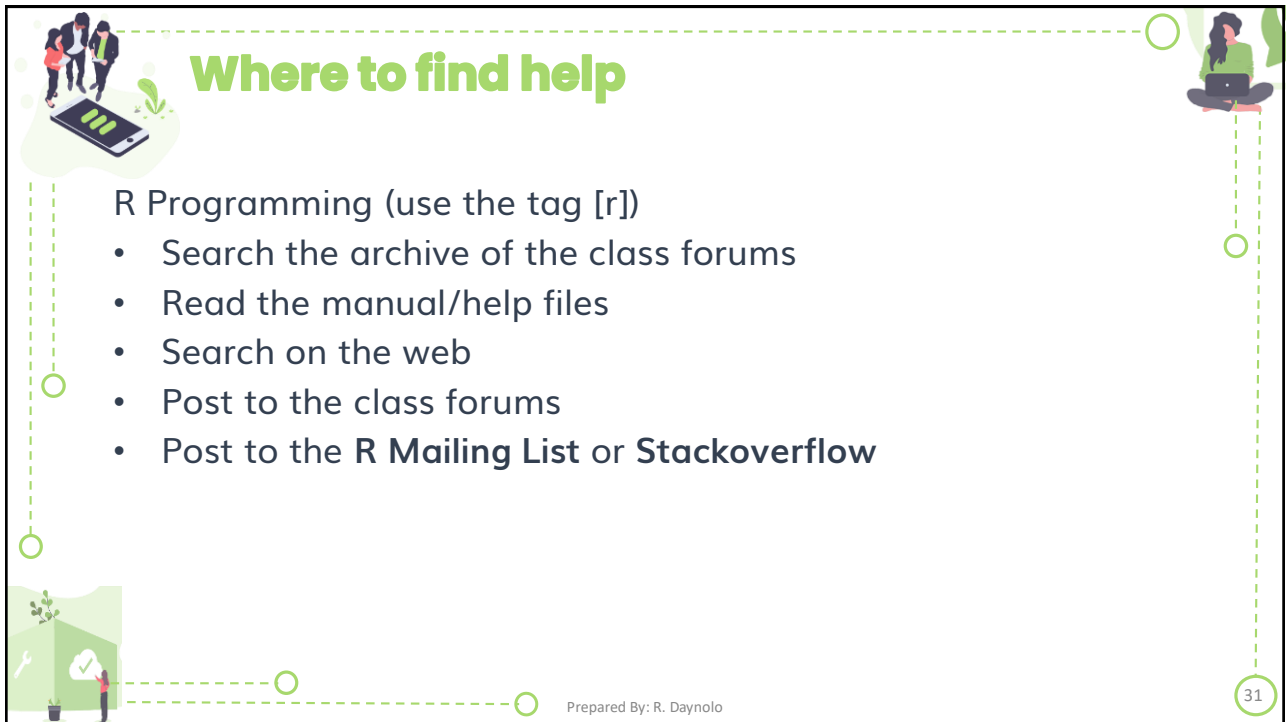
- Willing to find answer on their own
- Knowledgeable about where to find answers on their own
- Unintimidated by new data types or packages
- Unafraid to say they don't know the answer
- *Polite but relentless/persistent*

Prepared By: R. Daynolo

30

30





## Where to find help

R Programming (use the tag [r])

- Search the archive of the class forums
- Read the manual/help files
- Search on the web
- Post to the class forums
- Post to the **R Mailing List** or **Stackoverflow**

Prepared By: R. Daynola

31