



Weekly Discussion Sections & Readings

Teaching Fellows (TA)

Name	Office	Email
Mengting Gu	Bass 437	mengting.gu (at) yale.edu
Paul Muir	Bass 437	Paul.muir (at) yale.edu

Please E-mail cbb752@gersteinlab.org so both TAs will be informed

DS Format

- Group 1: Friday 1:00 - 2:00 PM, BASS 405
 - Group 2: Tuesday 6:00 - 7:00 PM, BASS 405
1. We will discuss two selected articles from the primary literature.
 2. Write-up approx. a half page (2-3 paragraph) summaries of each paper.
 3. Each student will give approx. 20 min presentations about the paper.

Write-ups and Presentations

- For write-ups and presentation, think about the following:
 - What was missing in the field? (introduction/background)
 - What were the questions the paper aim to address? (hypothesis)
 - What they did and what was the result? (method/results)
 - Conclusion and future direction (discussion/conclusion)
- During each presentation, student audiences will ask questions and discuss the topic presented.
- Please sign up for the presentation using this [spreadsheet](#). Also available on class website: cbb752b17.gersteinlab.org/materials



Introduction to R, Python & GitHub

Slides adapted from last year

Programming Assignments (Req'd for CBB and CS students)

- For the programming assignments, you can use either **R** or **Python**. However, if you would like to use other programming languages, please contact the TAs and request for a permission.
- There will be **THREE** homework assignments. We will try to promote the idea of reproducible research and using version control system, specifically **GitHub**, in facilitating the process of homework submission.



Introduction to R

What is R

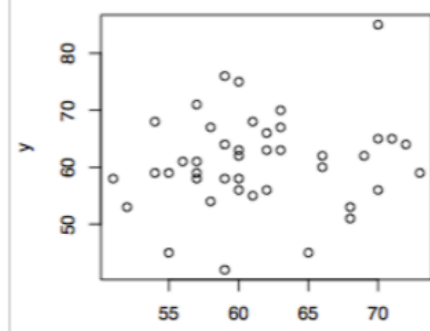
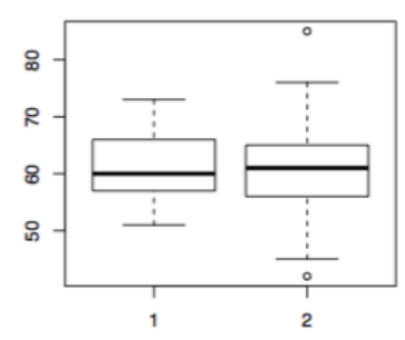
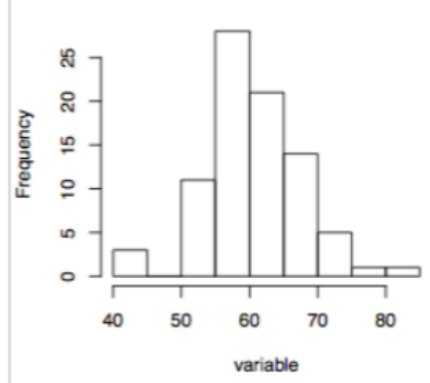
- A well-developed, simple and effective programming language;
- Statistical packages:
 - Statistical computing and graphics (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...);
- Command-line interface;
- Matrix-based programming language;
- Free.


```
> 3+5
[1] 8
> 2*7/9+2/(3+10)
[1] 1.709402
> a<-matrix(1:9,3,3)
> a
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> t(a)
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
> b<-c(1,2,3)
> a%*%b
      [,1]
[1,]   30
[2,]   36
[3,]   42
```

R examples

Mathematical calculations

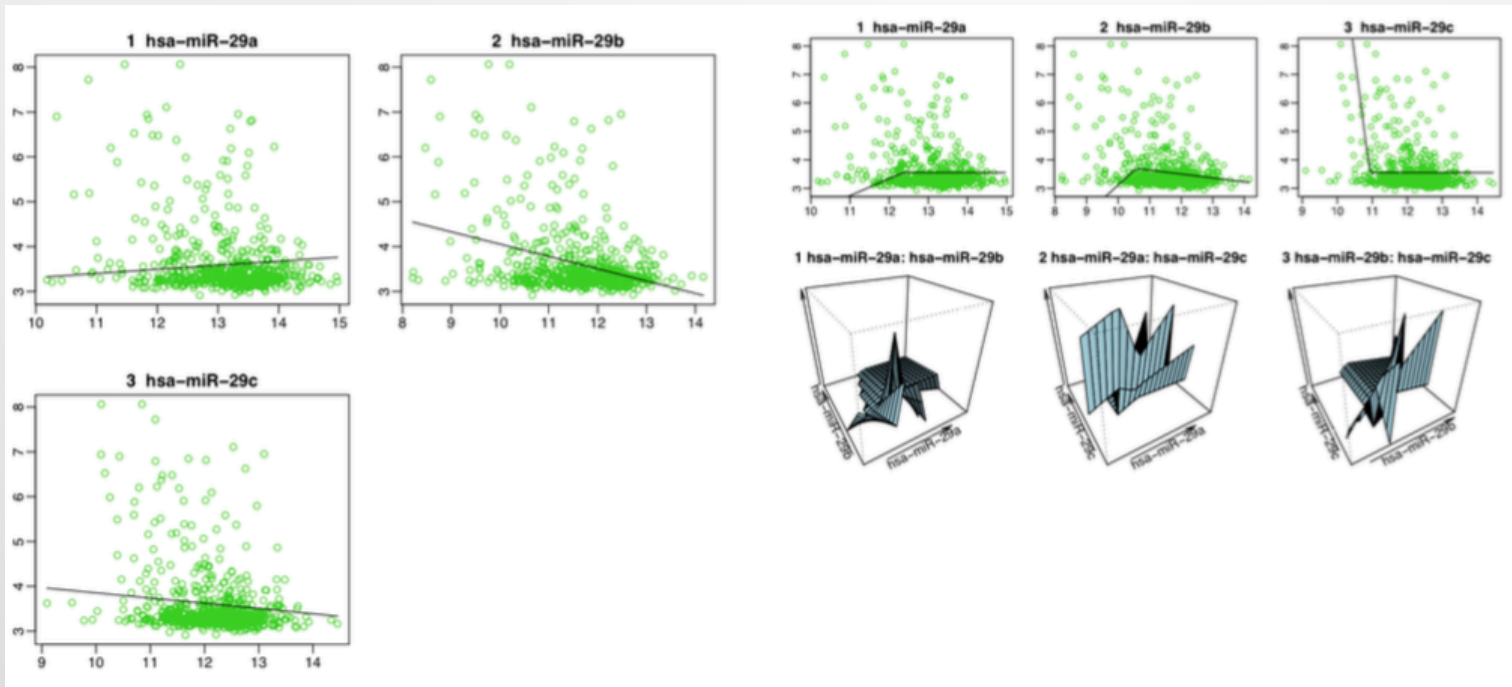
histogram



```
> summary(dat)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 42.00  57.00   60.50   61.24  65.25   85.00
```

R examples

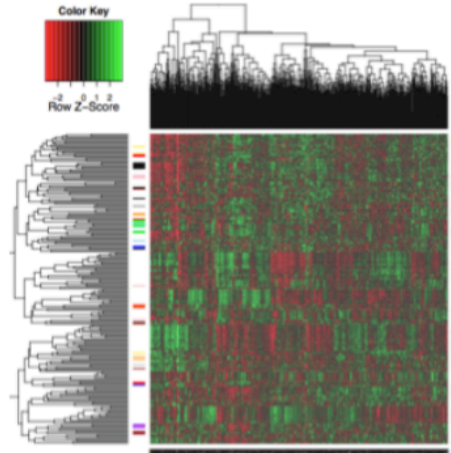
Data Summary



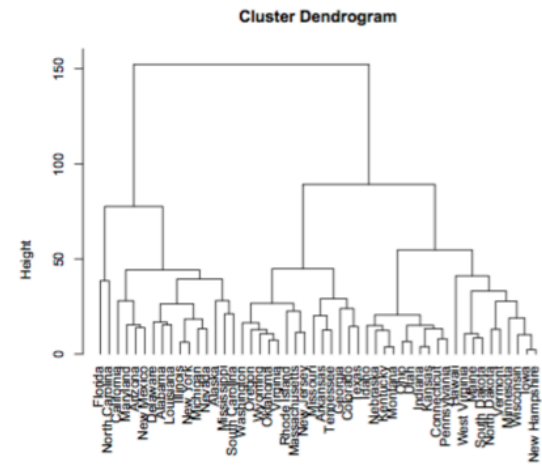
R examples

Data modeling

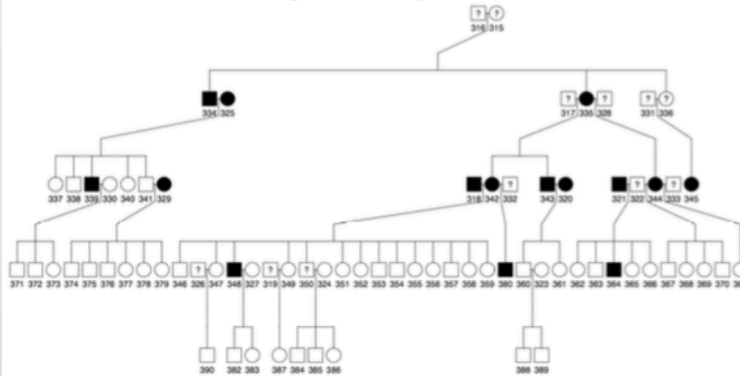
Heatmap



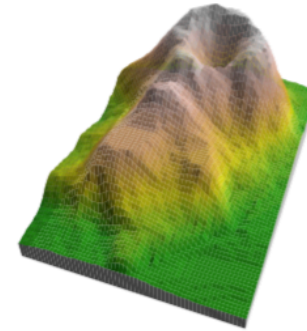
Clustering



Pedigree analysis



Fancier graphs

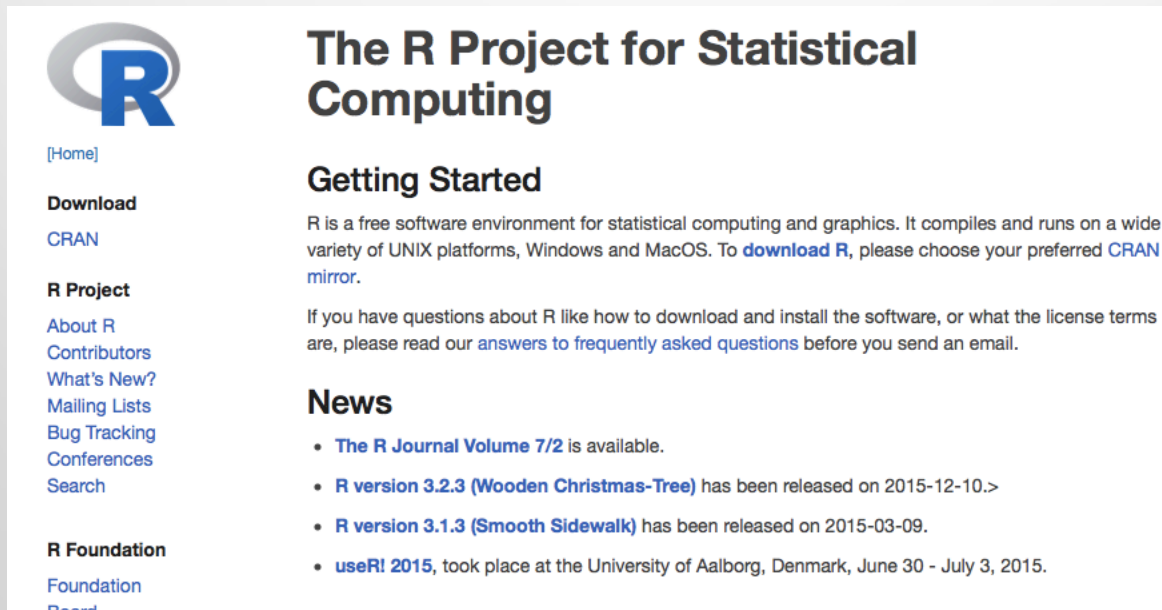


R examples


Data Visualization

Installing R

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



The screenshot shows the homepage of the R Project. On the left is a navigation menu with links for [Home], Download (CRAN), R Project (About R, Contributors, What's New?, Mailing Lists, Bug Tracking, Conferences, Search), and R Foundation (Foundation, Board). The main content area features the R logo, the title 'The R Project for Statistical Computing', and a 'Getting Started' section. The 'Getting Started' section explains that R is a free software environment for statistical computing and graphics, and provides instructions on how to download R from a CRAN mirror. Below this is a 'News' section with three bullet points: 'The R Journal Volume 7/2 is available.', 'R version 3.2.3 (Wooden Christmas-Tree) has been released on 2015-12-10.>', and 'R version 3.1.3 (Smooth Sidewalk) has been released on 2015-03-09.'. The final bullet point states 'useR! 2015, took place at the University of Aalborg, Denmark, June 30 - July 3, 2015.'



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](#).

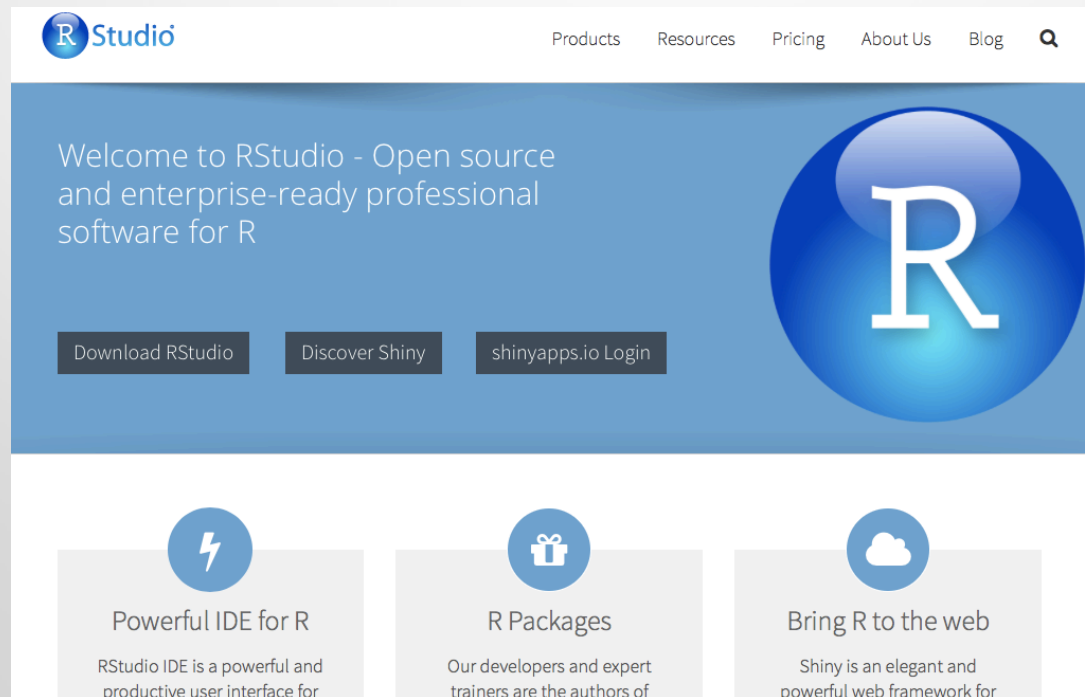
If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

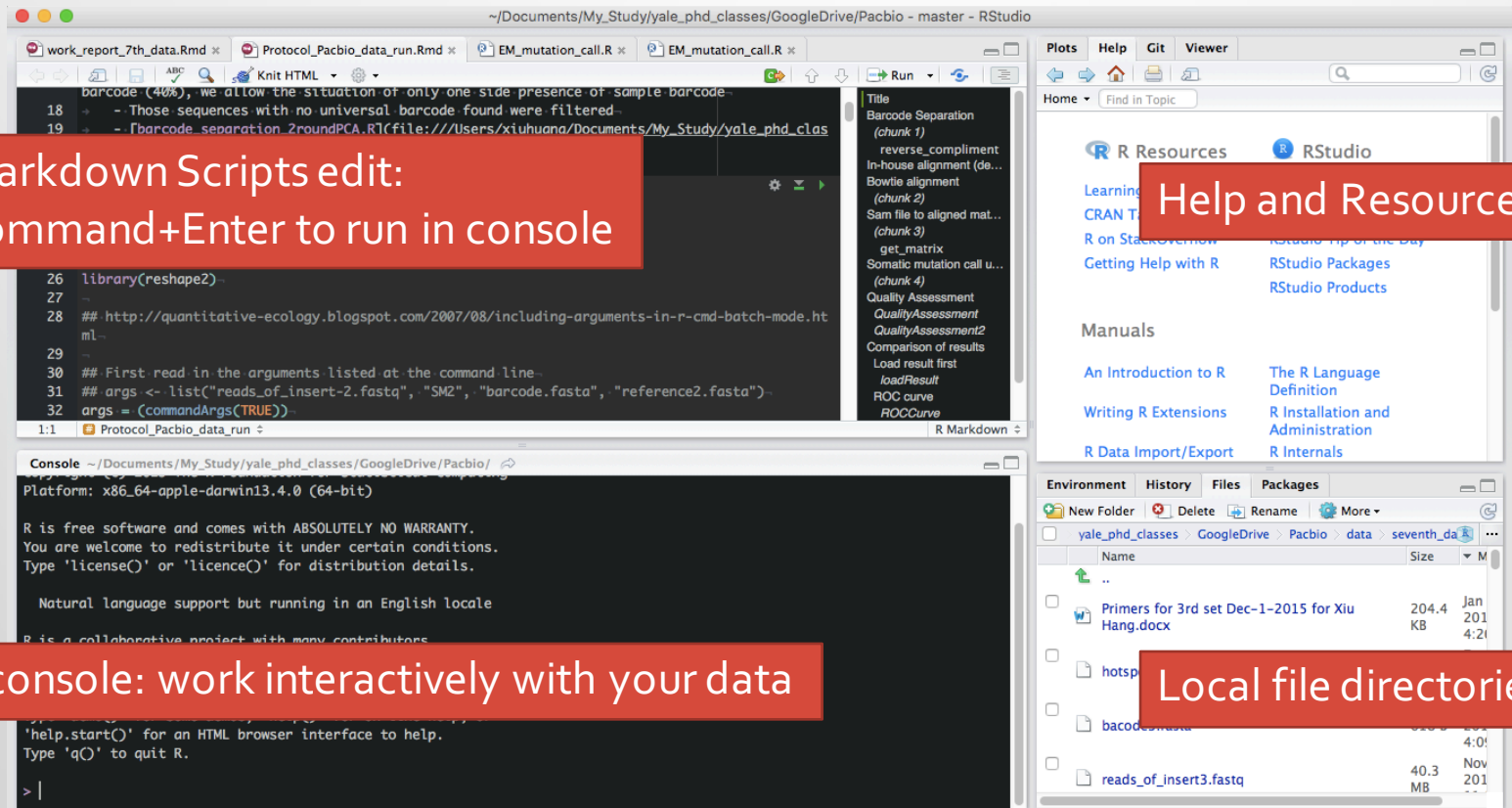
- [The R Journal Volume 7/2](#) is available.
- [R version 3.2.3 \(Wooden Christmas-Tree\)](#) has been released on 2015-12-10.>
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.
- [useR! 2015](#), took place at the University of Aalborg, Denmark, June 30 - July 3, 2015.

Installing RStudio

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



(my) RStudio Interface



R/Rmarkdown Scripts edit:
Hit Command+Enter to run in console

Help and Resources

R console: work interactively with your data

Local file directories

Basic Manipulation

- Create objects with `<-` or `=`;
- Remove objects by `rm()`;
- Show current objects by `ls()`;
- Names of objects and functions are case sensitive;
- Indices begin with `1`, not `0`;
- Comments: `#`;
- Math operators, vector, list, matrix, data frame, logical. (See code.)

Data Visualization

- Histogram;
 - `hist()`
- Scatter plot;
 - `plot()`
- Pairwise scatter plot;
 - `pairs()`
- Boxplot;
 - `boxplot()`
- Some other plots;

Export Graphs

- Export to PDF
 - `pdf("iris.pdf")`
- Plot multiple figures in one page;
 - `par(mfrow=c(2,2))`
- Shut down graph exporting;
 - `dev.off()`

Statistics in R

- Statistical tests (t test, chi square test);
 - `t.test(iris[1:50,1], iris[51:100,1])`
- Linear regression;
 - `lm1 <- lm(Petal.Length~Petal.Width, data=iris)`

Other Functions

- For loops;
- Conditional statements;
- Your own functions;

Input and Output Data from/to Files

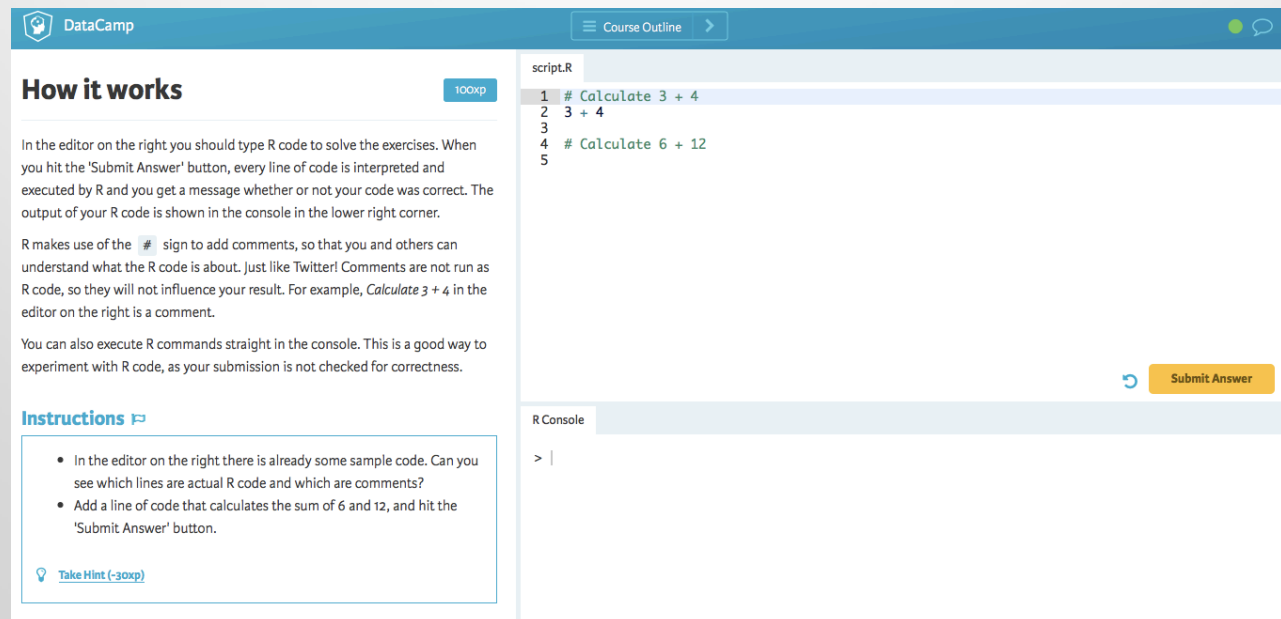
- Input data from files;
 - `x <- read.table("iris.txt", header=T, as.is=T)`
- Output data to files;
 - `write.table(x, "iris2.txt", sep="\t", quote=F, col.names=T, row.names=F)`

Getting Help

- `help(function);`
- `?function;`
- `help.search(function);`
- Open source – type function name without ();

Resources to Learn More about R

- Great Resources to Learn Basics of R **INTERACTIVELY**:
 - DataCamp: <https://www.datacamp.com/home>



The screenshot shows the DataCamp interface for an interactive R exercise. The top navigation bar includes the DataCamp logo, a 'Course Outline' menu, and a chat icon. The main content area is titled 'How it works' with a '100xp' badge. It contains three paragraphs of text explaining the workflow: typing R code in the editor, hitting 'Submit Answer' to execute code and see console output, and using the '#' symbol for comments. Below this is an 'Instructions' section with a list of two tasks: identifying R code vs. comments and adding a new calculation line. A 'Take Hint (-30xp)' button is at the bottom left of the instructions. On the right, the 'script.R' editor shows five lines of code: two comment lines and two calculation lines. Below the editor is the 'R Console' area with a prompt '> |' and a 'Submit Answer' button.

How it works 100xp

In the editor on the right you should type R code to solve the exercises. When you hit the 'Submit Answer' button, every line of code is interpreted and executed by R and you get a message whether or not your code was correct. The output of your R code is shown in the console in the lower right corner.

R makes use of the `#` sign to add comments, so that you and others can understand what the R code is about. Just like Twitter! Comments are not run as R code, so they will not influence your result. For example, *Calculate 3 + 4* in the editor on the right is a comment.

You can also execute R commands straight in the console. This is a good way to experiment with R code, as your submission is not checked for correctness.

Instructions

- In the editor on the right there is already some sample code. Can you see which lines are actual R code and which are comments?
- Add a line of code that calculates the sum of 6 and 12, and hit the 'Submit Answer' button.

Take Hint (-30xp)

```
script.R
1 # Calculate 3 + 4
2 3 + 4
3
4 # Calculate 6 + 12
5
```

R Console

> |

Submit Answer

Resources to Learn More about R

- Great Resources to Learn Basics of R **INTERACTIVELY** :
 - TryR: <http://tryr.codeschool.com>

The screenshot shows the TryR interface. On the left is a navigation menu with a table of contents:

1. Using R	
Expressions	0 of 3
Logical Values	0 of 3
Variables	0 of 5
Functions	0 of 3
Help	0 of 3
Files	0 of 3
2. Vectors	
3. Matrices	
4. Summary Statistics	
5. Factors	
6. Data Frames	
7. Real-World Data	
8. What's Next	

The main content area features an O'Reilly logo and a progress indicator for '1 Complete to Unlock'. The title 'Expressions' is displayed with a sub-section '1.1'. The text reads: 'Type anything at the prompt, and R will evaluate it and print the answer. Let's try some simple math. Type the below command. [Or, if you prefer, click on the command and it will be typed into the console for you!]' Below this is a button with the text '1 + 1' and a dark console area with a prompt character '>'.

Resources to Learn More about R

- Great Resources to Learn Basics of R **INTERACTIVELY** :
 - Swirl: Learn R, in R!

```
| To begin, you must install a course. I can install a course for you from the
| internet, or I can send you to a web page
| (https://github.com/swirldev/swirl_courses) which will provide course options
| and directions for installing courses yourself. (If you are not connected to
| the internet, type 0 to exit.)

1: R Programming: The basics of programming in R
2: Regression Models: The basics of regression modeling in R
3: Don't install anything for me. I'll do it myself.

Selection: 1

| Course installed successfully!

| Please choose a course, or type 0 to exit swirl.

1: R Programming
2: Take me to the swirl course repository!

Selection: 1

| Please choose a lesson, or type 0 to return to course menu.

1: Basic Building Blocks      2: Workspace and Files
3: Sequences of Numbers      4: Vectors
5: Missing Values            6: Subsetting Vectors
7: Matrices and Data Frames  8: Logic
9: Functions                  10: lapply and sapply
11: vapply and tapply        12: Looking at Data
13: Simulation                14: Dates and Times
15: Base Graphics

Selection: 4

|
```

Home Learn Teach Contribute Blog FAQ Help

{swirl}

Learn R, in R.

swirl teaches you R programming and data science interactively, at your own pace, and right in the R console!

Follow @swirlstats

Got questions? Join our [discussion group!](#)

Resources to Learn More about R


- More Resources to read about R:
 - The beginner's guide to R web posts:
<http://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html>

Reproducible Research in R

- The reason we choose RStudio is that it offers a convenient interface to carry out reproducible research with RMarkdown.
- Advantages of RMarkdown:
 - Easy syntax to generate reports that can easily transformed to Doc, Latex, PDF and all the other different formats;
 - Imbedded R code chunks with beautiful syntax highlight, running automatically, caching automatically, and generating results in a report kind of way. Neat!

Go through this 5-minute RMarkdown Guide and You are a Pro already

- <https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>

Syntax	Becomes
Plain text	Plain text
End a line with two spaces to start a new paragraph.	End a line with two spaces to start a new paragraph.
<code>*italics*</code> and <code>_italics_</code>	<i>italics</i> and <i>italics</i>
<code>**bold**</code> and <code>__bold__</code>	bold and bold
<code>superscript^2^</code>	superscript ²
<code>~~strikethrough~~</code>	strikethrough
<code>[link](www.rstudio.com)</code>	link
<code># Header 1</code>	Header 1
<code>## Header 2</code>	Header 2
<code>### Header 3</code>	Header 3
<code>#### Header 4</code>	Header 4
<code>##### Header 5</code>	Header 5
<code>##### Header 6</code>	Header 6
<code>endash: --</code>	endash: –
<code>emdash: ---</code>	emdash: —
<code>ellipsis: ...</code>	ellipsis: ...
<code>inline equation: \$A = \pi * r^{2}\$</code>	inline equation: $A = \pi * r^2$
<code>image: </code>	image: 
horizontal rule (or slide break):	horizontal rule (or slide break):



Introduction to GitHub

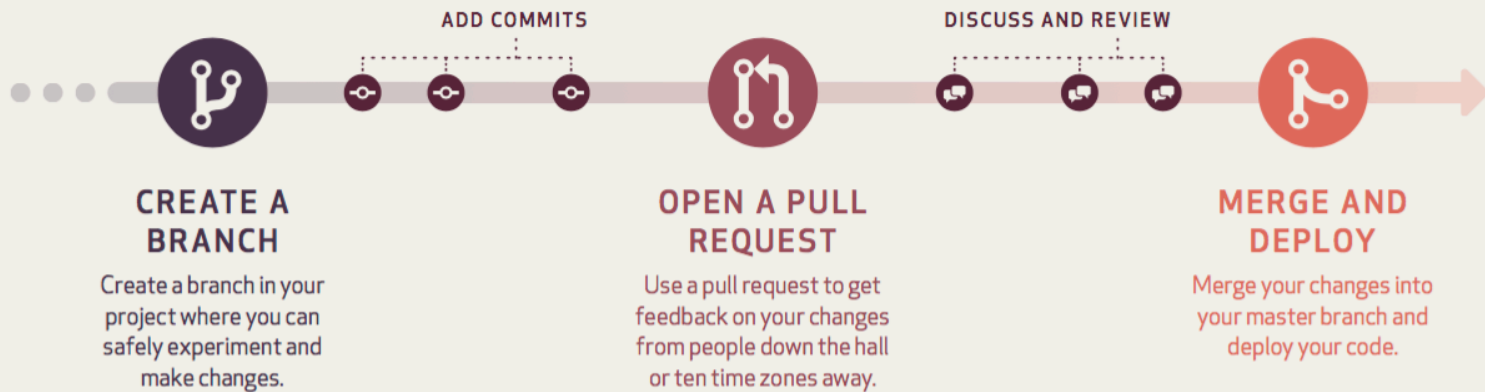
Why GitHub?

- Version control: keep track of your changes
- Easy interface: no command line needed at all!
- Easy collaboration: work together on the same project

The GitHub Concept

WORK FAST WORK SMART THE GITHUB FLOW

The GitHub Flow is a lightweight, branch-based workflow that's great for teams and projects with regular deployments. Find this and other guides at <http://guides.github.com/>.



GitHub App

The screenshot displays the GitHub App interface for a repository named 'hooyden0329/Bioinformatics'. The top bar shows 'No Uncommitted Changes' and 'History'. The main area is divided into a left sidebar with repository filters and a main content area. The main content area shows a merge pull request #2 from 'hoody/master' by 'Donghoon' (hooyden0329) 2 months ago. The pull request details include the file 'README.md' and a diff view. The diff shows changes to the README.md file, with a red background for deletions and a green background for additions. The content of the README.md file is as follows:

```
@@ -99,10 +99,13 @@ These go beyond Basic Math (calculus), Biology,
Chemistry & Physics taught in pr
- Synteny
- Function Classification & Orthologs
- Genome Annotation
- Gene Prediction
- Regulatory site and network prediction
- miRNA prediction and targeting site prediction
- Pseudogene prediction and functional prediction
+ Coding
+ - Gene Prediction
+ - Functional Annotation
+ Non-coding
+ - Regulatory site and network prediction
+ - miRNA prediction and targeting site prediction
+ - Pseudogene prediction and functional prediction

#### Next-Gen Sequencing Data Processing
```


GitHub Web UI

The screenshot shows the GitHub web interface for a repository named "Bioinformatics" by user "hoyden0329". The repository has 8 commits, 2 branches, 0 releases, and 2 contributors. The main content area displays a commit history table with two entries: "Update README.md" (2 months ago) and "venn diagram plot" (3 months ago). Below the commit history, the README file is open, showing the title "Categories of Knowledge for Bioinformatics Education" and a legend for the categories: U = Undergrad. level, G = Grad. level, C = CS, S = Stats/Math, B = Bio/Chem/Phys, I = Intro. Bioinformatics Topic, and A = Advanced Bioinformatics Topic.

GitHub repository page for **hoyden0329 / Bioinformatics**. The page shows the repository name, user profile, and navigation options (Code, Issues, Pull requests, Wiki, Pulse, Graphs, Settings). The repository has 8 commits, 2 branches, 0 releases, and 2 contributors. The main content area displays the commit history for the `master` branch, showing two recent commits: "Update README.md" (2 months ago) and "venn diagram plot" (3 months ago). Below the commit history, the README file is open, showing the title "Categories of Knowledge for Bioinformatics Education" and a legend for the categories: U = Undergrad. level, G = Grad. level, C = CS, S = Stats/Math, B = Bio/Chem/Phys, I = Intro. Bioinformatics Topic, and A = Advanced Bioinformatics Topic.

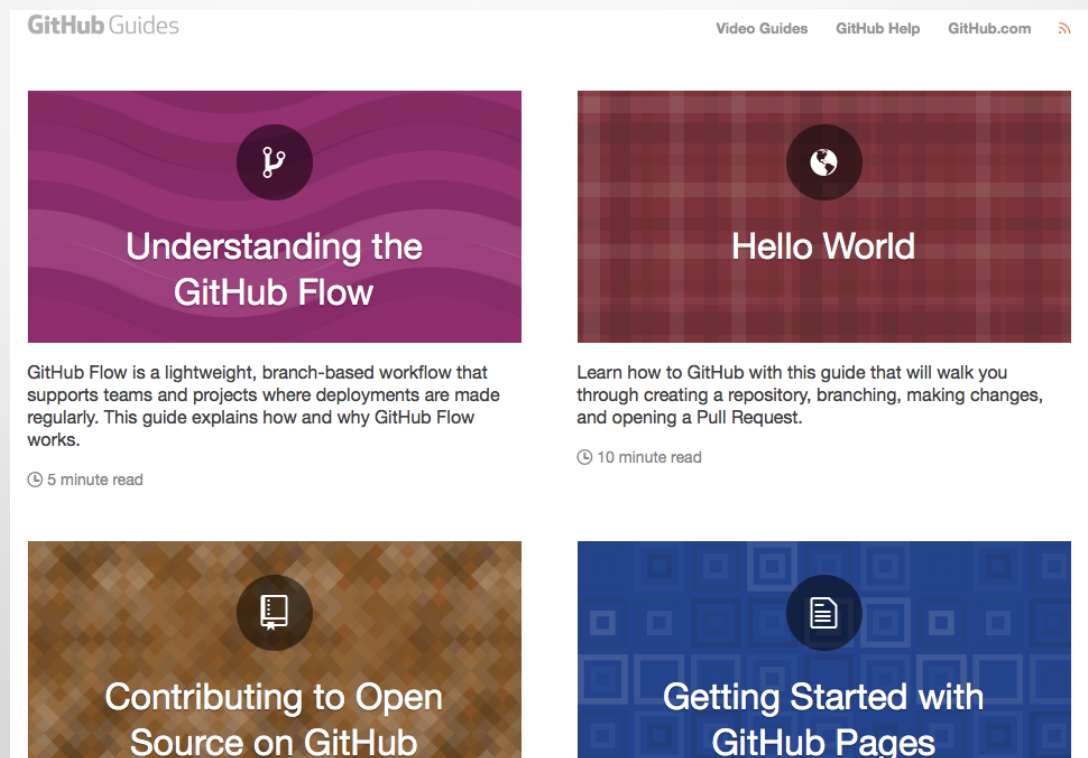
File	Commit Message	Time
README.md	Update README.md	2 months ago
venn.png	venn diagram plot	3 months ago

Categories of Knowledge for Bioinformatics Education

U = Undergrad. level G = Grad. level
C = CS S = Stats/Math B = Bio/Chem/Phys
I = Intro. Bioinformatics Topic (i.e., in a class like CBB752) A = Advanced Bioinformatics Topic (i.e., maybe beyond CBB752)

More GitHub Learning Resources

- <https://guides.github.com>



The screenshot displays the GitHub Guides website with a white background and a blue header. The header includes the text "GitHub Guides" on the left and "Video Guides", "GitHub Help", and "GitHub.com" on the right. Below the header, there are four guide cards arranged in a 2x2 grid. Each card has a colored background with a circular icon and a title. The first card is purple with a GitHub logo icon and the title "Understanding the GitHub Flow". The second card is red with a globe icon and the title "Hello World". The third card is brown with a document icon and the title "Contributing to Open Source on GitHub". The fourth card is blue with a document icon and the title "Getting Started with GitHub Pages". Below each card is a short description and a "minute read" indicator.

GitHub Guides Video Guides GitHub Help GitHub.com

Understanding the GitHub Flow

GitHub Flow is a lightweight, branch-based workflow that supports teams and projects where deployments are made regularly. This guide explains how and why GitHub Flow works.

🕒 5 minute read

Hello World

Learn how to GitHub with this guide that will walk you through creating a repository, branching, making changes, and opening a Pull Request.

🕒 10 minute read

Contributing to Open Source on GitHub

Getting Started with GitHub Pages



Introduction to Python

What is Python?



- Interpreted (no need for compile)
- Human readable clear syntax (fewer lines of code than C++ or Java)
- Libraries like NumPy, SciPy, BioPython allow the effective use of Python in scientific computing

Installation

- We recommend Python 2.7 and up
- OS X (Mac), Linux and Unix: You already have one!
- Windows: <https://www.python.org/downloads/windows/>
- For more information:
<https://wiki.python.org/moin/BeginnersGuide/Download>
- For writing python, you can use any text editor. I personally use Emacs.
- If you are looking for a Python IDE, there are IDLE, Eclipse (PyDev), PyCharm, etc

Python Basics

- Python uses white-space indentation rather than `}`
- Use tab or spaces, but do not mix them
- No need for `;` at the end of line
- In Python `==` compares by value
- Python uses ***and***, ***or***, ***not*** for its boolean operators rather than `&&`, `||`, `!`
- Lists are mutable and written as `[1, 2, 3]`
- Tuples are immutable and written as `(1, 2, 3)`
- Use curly braces for dictionary (key-value pair, not ordered) e.g., `{'key1': 1.0, 'key2': False}`

Example

```
test.py
1 print "hello world!"
2
3 x = 1
4
5 if x == 1:
6     print "x is 1"
7 else:
8     print "x is not 1"
9
10 for y in range(1,10):
11     print y
12
13 def dist(x1, y1, x2, y2):
14     return ((x2-x1)**2 + (y2-y1)**2)**0.5
15
16 print dist(0,0,1,1)
17
18 addressbook = {'Donghoon':'donghoon.lee@yale.edu', 'Xiu':'xiu.huang@yale.edu'}
19 print addressbook['Donghoon']
```

```
[Donghoons15rMBP:test Donghoon$ ll
total 8
-rw-r--r--  1 Donghoon  staff  309 Jan 18 16:29 test.py
[Donghoons15rMBP:test Donghoon$ python test.py
hello world!
x is 1
1
2
3
4
5
6
7
8
9
1.41421356237
donghoon.lee@yale.edu
Donghoons15rMBP:test Donghoon$
```

Resources

- [Python Cookbook, 3rd Edition: Recipes for Mastering Python 3 by David Beazley, Brian K. Jones](#) (Publisher: O'Reilly Media). You can view [Online copy at Yale library](#) and this is [Github source code from book](#).
- 13 hr self-taught crash course on python, <https://www.codecademy.com/learn/python>
- And of course, Google and Stack Overflow are your best programming friend.