

Sharpening the Tools in Your Data Science Toolbox



Jessica Minnier, PhD

OHSU-PSU School of Public Health

Knight Cancer Institute Biostatistics Shared Resource

Oregon Health & Science University

Joint Statistical Meetings, July 29, 2019

 bit.ly/jmin-jsm19
 [datapointier](#)

Who am I?

- Use R every day, git often, unix sometimes, python rarely
- Teach intro to R, R markdown, intro to stats for data science, R Shiny (and Math/Stats)
- Collaborative statistician (mostly applied work, no PhD students, no Postdocs, some MS students and staff)
- Involved in: organizing Cascadia R conf, PDX R Meetup, R Ladies, BioData Club

Statistics is changing

A non-random sample of buzzwords:

- Big Data
- Data Science
- Machine Learning
- Reproducible Coding
- Version Control
- Interactive Dashboards

ASA Statement on the Role of Statistics in Data Science

1 OCTOBER 2015

13,976 VIEWS

13 COMMENTS

Sustained and substantial collaborative effort with researchers with expertise in data organization and in the flow and distribution of computation. Statisticians must engage them, learn from them, teach them, and work with them.

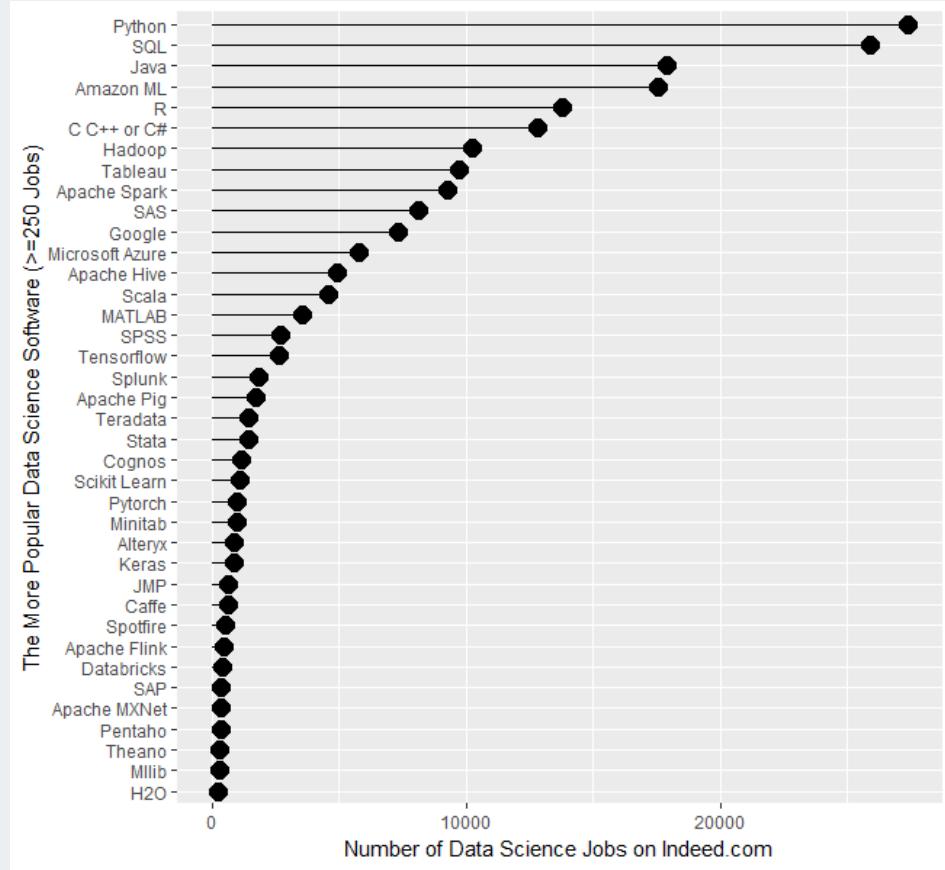
Statistical education and training must continue to evolve—the next generation of statistical professionals needs a broader skill set and must be more able to engage with database and distributed systems experts.

The next generation must include more researchers with skills that cross the traditional boundaries of statistics, databases, and distributed systems; there will be an ever-increasing demand for such “multi-lingual” experts.

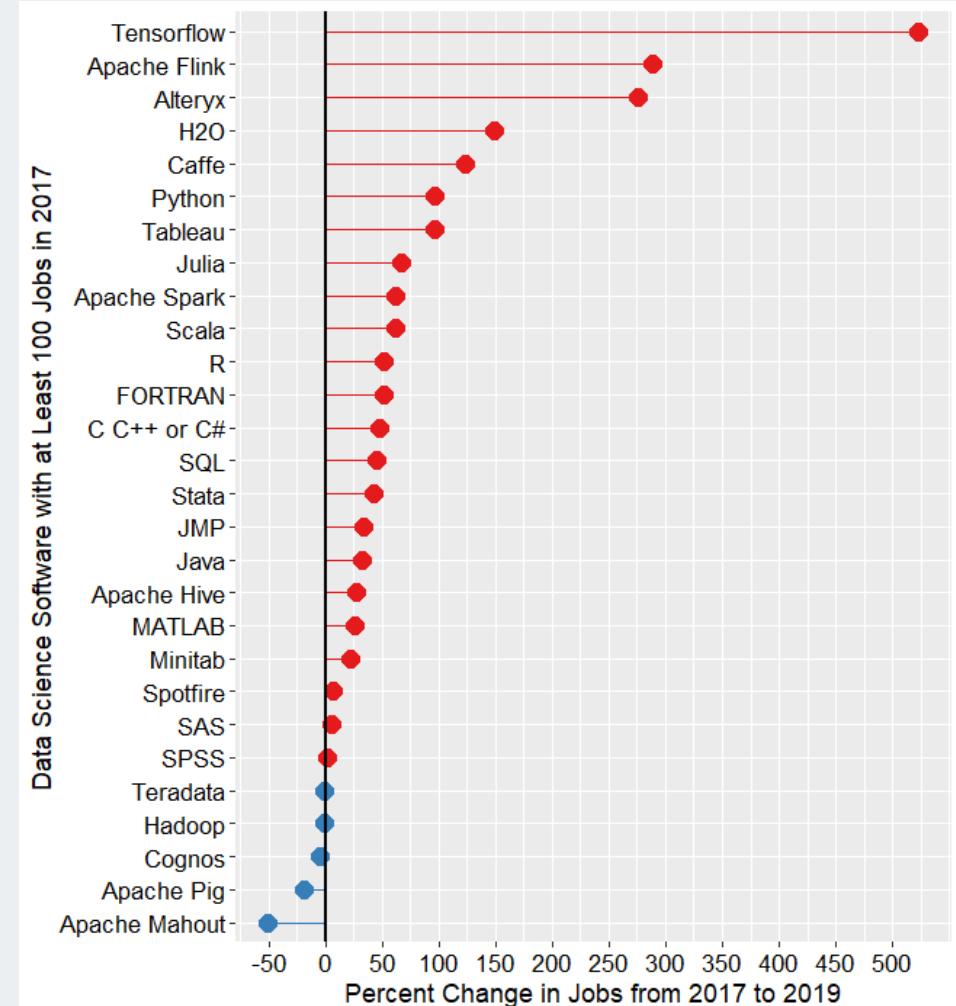
Programming itself is changing

- Data Science is programming intensive
 - Big data and complex models (machine learning beyond regression)
- Beyond SAS
 - R is easier to learn than ever before (see: tidyverse)
 - R packages are easier to make and distribute
 - R is not "slow" anymore
 - FDA allows R for clinical trials
- Python is most common in data science
 - but not statistics!

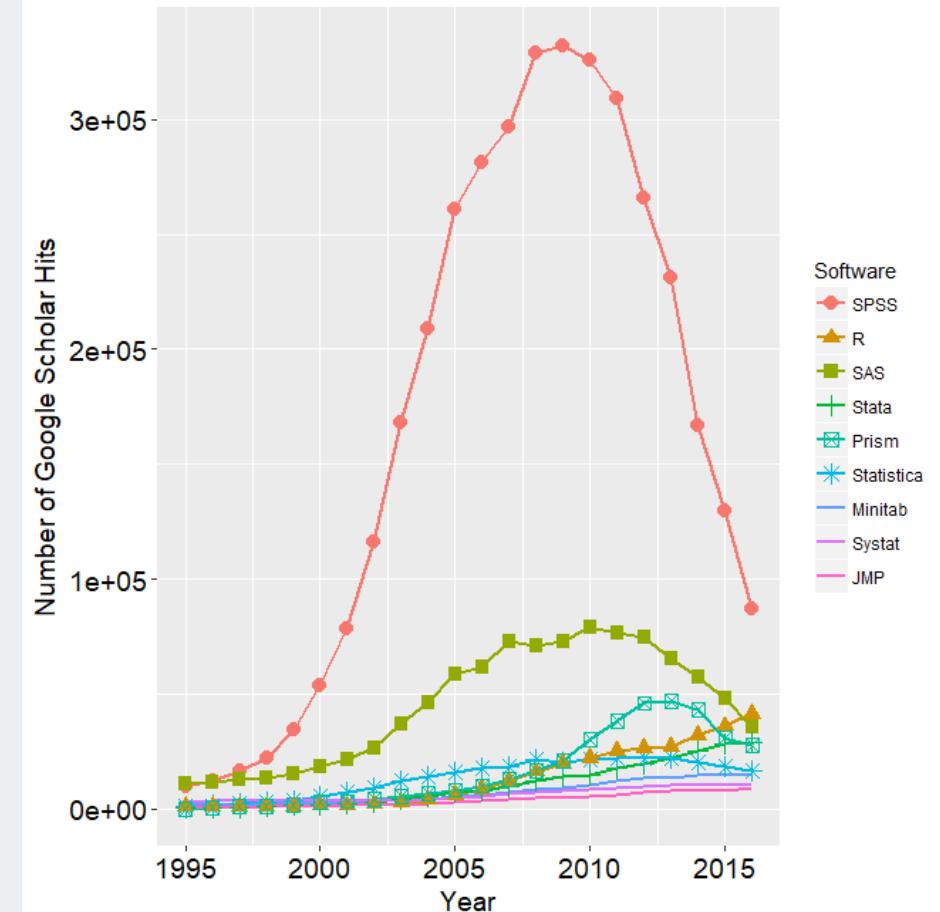
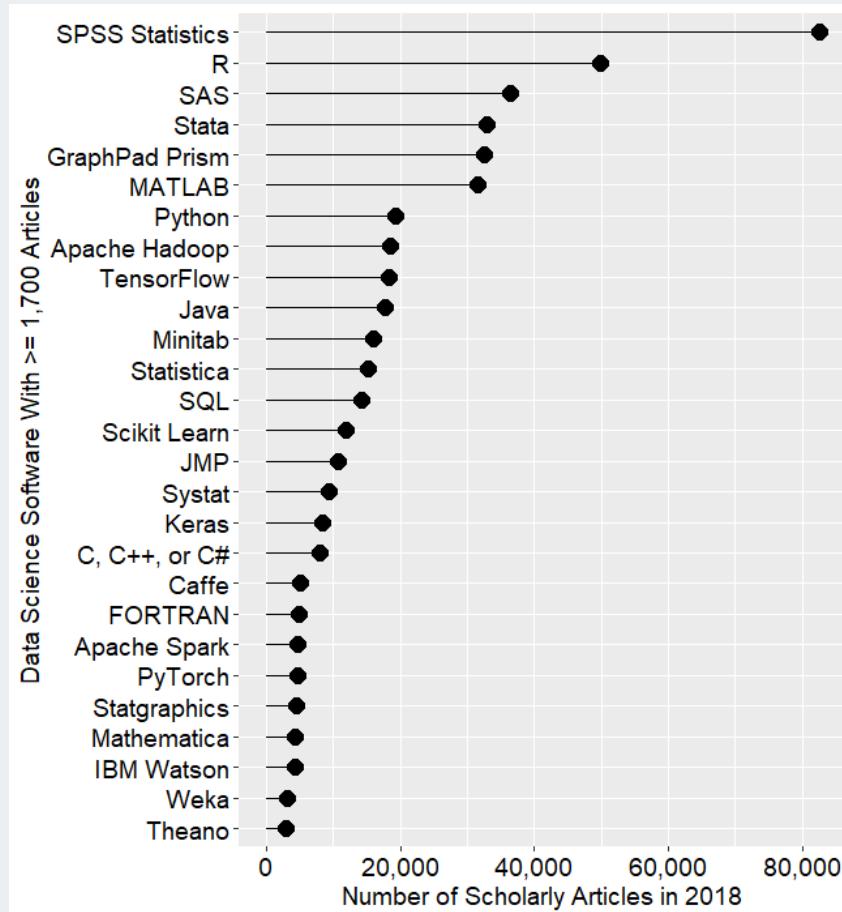
Data Science Jobs: Python, SQL, R



R4stats - Robert A. Muenchen



By contrast, scholarly articles: SPSS, R, SAS



R4stats - Robert A. Muenchen

How to keep up?

Learning takes time,
dedication, courage.

How to draw an Owl.

"A fun and creative guide for beginners"

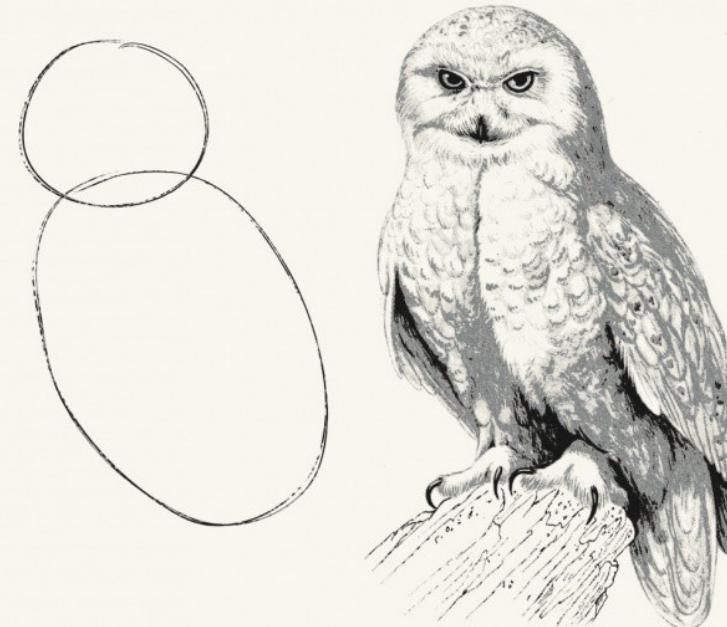


Fig 1. Draw two circles

Fig 2. Draw the rest of the Owl

Highlights of my education in programming

CS courses in college (C, Java):

Name	Date Modified	Size	Kind
AlgorithmBook	Sep 20, 2017, 12:05 AM	384 KB	Folder
Dots	Sep 19, 2017, 9:31 PM	17 KB	Folder
Dots.class	Apr 16, 2007, 4:58 PM	4 KB	Java class file
Dots.java	Feb 26, 2006, 2:55 PM	5 KB	Java Source
DotsFrame.class	Apr 16, 2007, 4:58 PM	826 bytes	Java class file
DotsFrame.java	Jan 30, 2006, 8:00 PM	537 bytes	Java Source
DotsListener.class	Apr 16, 2007, 4:58 PM	1 KB	Java class file
DotsListener.java	Feb 26, 2006, 2:52 PM	1 KB	Java Source
DotsPanel.class	Apr 16, 2007, 4:58 PM	2 KB	Java class file
DotsPanel.java	Jan 30, 2006, 8:00 PM	1 KB	Java Source
Final Project RosenKonig	Today, 1:11 PM	122 KB	Folder
Go Moku	Sep 19, 2017, 9:38 PM	17 KB	Folder
HWK 1 Ch 7	Sep 19, 2017, 9:58 PM	20 KB	Folder
Die2.class	Apr 16, 2007, 4:58 PM	2 KB	Java class file
Die2.java	Jan 23, 2006, 7:29 PM	1 KB	Java Source
LinkedList2.class	Apr 16, 2007, 4:58 PM	3 KB	Java class file
LinkedList2.java	Jan 23, 2006, 8:22 PM	2 KB	Java Source

My PhD advisor wrote a ton of code

Postdoc (learned git/github):



First year Assistant Professor (R
Markdown, Shiny were all new &
exciting!)



Learning while teaching

START App

R Shiny Transcriptome Analysis Resource Tool

Jessica Minnier

email: minnier@ohsu.edu

Wednesday, December 7, 2016

<https://github.com/jminnier/STARTapp>

Slides available at <http://bit.ly/rmeetup-start>

Building Shiny Apps: With Great Power Comes Great Responsibility

CSP 2018

Jessica Minnier, PhD

Oregon Health & Science University

Twitter: [@datapointier](#)

Slides available at <http://bit.ly/shiny-csp18>

February 16, 2018

emrselect

Automated Feature Selection of Predictors in Electronic Medical Records Data

Jessica Minnier; minnier@ohsu.edu

Women Who Code + PDX R User Meetup

Tuesday, January 10, 2017

<https://github.com/jminnier/emrselect>

Slides available at <http://bit.ly/wwwc-emrselect>

Shiny Apps in Genomics and Trials

R/Pharma 2018

Jessica Minnier, PhD

Oregon Health & Science University

 @datapointier

August 16, 2018

Slides available at <http://bit.ly/rpharma-minnier>

Building Shiny Apps

Challenges and Responsibilities

Jessica Minnier

email: minnier@ohsu.edu

@datapointier

Saturday, January 27, 2018, Data/R Day Texas

Slides available at <http://bit.ly/shiny-ddtx>

github.com/jminnier/talks_etc

Learning while teaching with new tools

(also: find collaborators better than you at all this - thanks Ted Laderas!)

The screenshot shows two separate shiny applications side-by-side, both titled "Continuous Data".

Left Shiny App: This app is titled "The effect of Age on the Data". It displays a histogram of deaths by age group. The x-axis is labeled "age" and ranges from 19 to 83. The y-axis is labeled "count" and ranges from 0.00 to 1.00. A legend indicates three categories: "alive" (light blue), "dead" (dark red), and "not known" (grey). The histogram shows a significant peak for the "alive" category across most ages, with a smaller peak for the "dead" category.

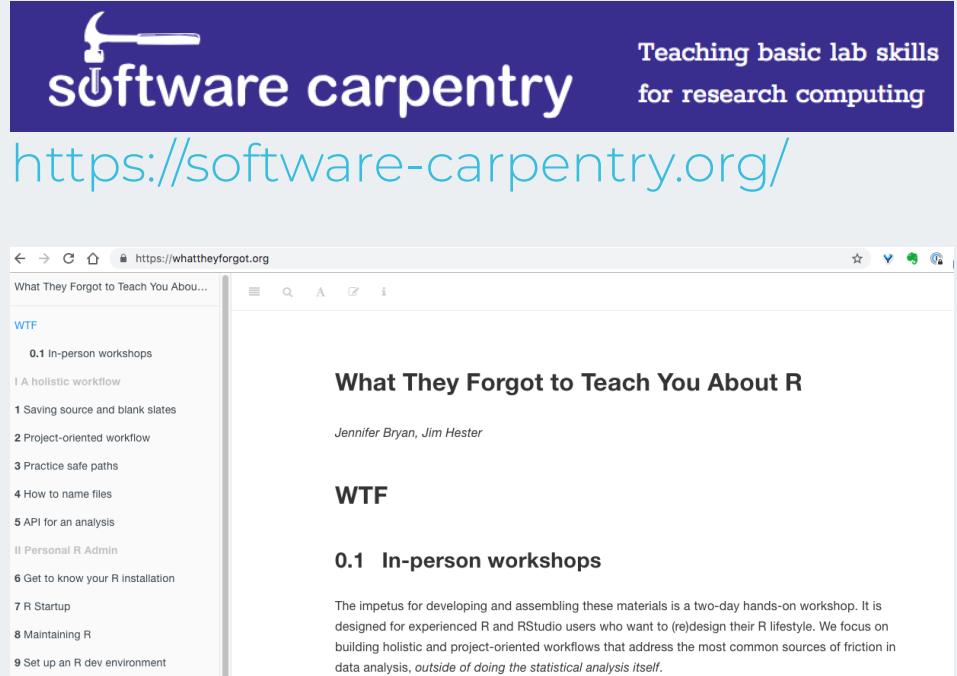
Right Shiny App: This app is titled "Histogram and density plots". It displays a histogram of BMI with a superimposed density curve. The x-axis is labeled "BMI" and ranges from 0 to 80. The y-axis is labeled "count" and ranges from 0 to 1500. The distribution is skewed to the right, with a large peak around 25-30 BMI and a long tail extending towards higher values.

Common Features: Both apps have a header with the R logo and "Bootcamp". They include sections for "Learning Objectives for this Session" and "What is Exploratory Data Analysis?". There are also sections for "Continuous Data" and "EDA with continuous variables". The right app includes a code snippet for generating the histogram and density plot.

r-bootcamp.netlify.com; minnier.shinyapps.io/ODSI_categoricalData/; minnier.shinyapps.io/ODSI_continuousData

Software development

- Unix/shell (i.e. for cluster computing)
- Programming skills
- Version control (i.e. git/github/gitlab, bitbucket)
- Unit Testing (does your code work?)
- Collaborative coding
- Using databases (i.e. SQL)
- Automation (i.e. make)



The screenshot shows the Software Carpentry website. At the top, there's a purple header with the Software Carpentry logo (a hammer icon) and the text "Teaching basic lab skills for research computing". Below the header is a teal URL: <https://software-carpentry.org/>. The main content area has a white background. On the left, a sidebar lists "What They Forgot to Teach You About R" under the heading "WTF". The sidebar items include:

- 0.1 In-person workshops
- I A holistic workflow
- 1 Saving source and blank slates
- 2 Project-oriented workflow
- 3 Practice safe paths
- 4 How to name files
- 5 API for an analysis
- II Personal R Admin
- 6 Get to know your R installation
- 7 R Startup
- 8 Maintaining R
- 9 Set up an R dev environment

On the right, the main content area is titled "What They Forgot to Teach You About R" and credits "Jennifer Bryan, Jim Hester". It then lists "0.1 In-person workshops" and provides a detailed description of the workshop's purpose and target audience.

<https://whattheyforgot.org/>

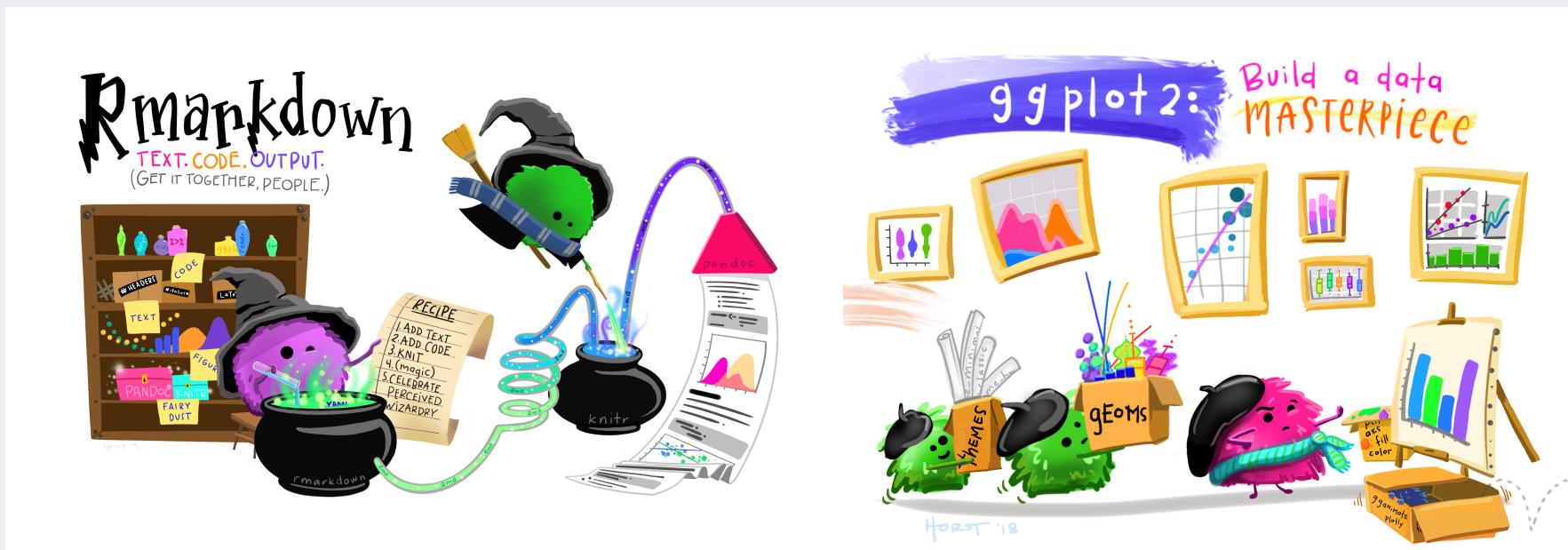
Improve your coding skills

Similar to learning a new language

- Difficult to do when busy/sporadically, unless you set aside time (make a deadline) or find an immersive experience
- Learn more R? SAS/STATA? Find workshops! (JSM, CSP, SDSS, Rstudio::conf, UseR!, local R meetups and conferences)
- Learn git/github/software development? Host a [Software Carpentry](#) workshop

Coding in R is more fun these days

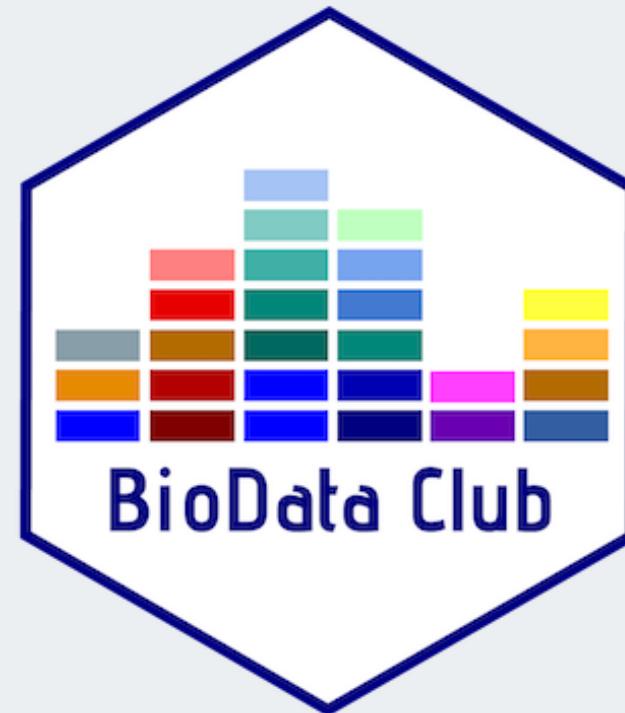
- R Markdown → reproducible documents
 - transformative for collaborative statistical analyses
 - these slides are made with knitr/r markdown
 - make textbooks, webpages, much more
- Shiny → webpages to show off methodology/analyses



Find a community & learn together

We all have something to learn!

- Meetup groups (R Users, R Ladies, Women Who Code)
- Hackweeks
- Data Science Workshops
- Data competitions (DataFest)
- Workshops at conferences
- Online: [#rstats Twitter](#), [Rstudio community](#), [R 4 Data Science Learning community](#), [Tidy Tuesday](#)
- Start your own club!



<https://biodata-club.github.io/>

It takes time to learn how to learn

R specific learning checks from Alison Hill's [Big Magic with R: Creative Learning Beyond Fear](#) - watch the [video!](#)



Alison Hill, Associate Professor of Pediatrics & Assistant Director of OHSU's Center for Spoken Language Understanding

[Twitter: @apreshill](#) [GitHub: @apreshill](#) [Blog: https://alison.rbind.io/](#)
[Slides: http://bit.ly/cascadiarconf-magic](#)

All-purpose courage

- Package [reference](#) docs
(usage / arguments / examples)
- Package [vignette](#)
- GitHub [README](#)
- Is there an [RStudio Cheatsheet](#)?
(<https://www.rstudio.com/resources/cheatsheets/>)
- Are there Stat545 materials by [Jenny Bryan](#)?
(<http://stat545.com>)
- Is there a roundup by [Mara Averick](#)?
(<https://maraaverick.rbind.io/tags/roundups/>)
- [RWeekly](#) "R Tutorials": <https://rweekly.org/#Tutorials>

Practice, practice, practice

- Learning a tool and then forgetting to use it is the best way to unlearn it
- Write down what you've learned (make a list of commands, write a blog post)

Give talks

- Sign up to give talks at local meetups
- Teach workshops to students/fellow faculty/coworkers
- Journal/book clubs with your co-workers/team

Resources →

- Interactive lessons
- Online courses
- Blog posts

Ask for help

- Stack Overflow
- Rstudio Community

The screenshot shows a GitHub repository page for 'jminnier / awesome-rstats'. The file 'learn-r.md' is displayed. The content of the file is a list of resources for learning R/Rstudio, categorized into several main sections:

- Learning R/Rstudio for beginners
 - Interactive lessons
 - Slides/videos/online courses
 - Textbooks (online, free)
 - Blog posts
 - Resources
 - \$\$ Options
 - Blogs
- Statistics in R
 - Interactive lessons
- Intermediate/Advanced/More R
 - Textbooks/tutorials

Below the file content, there are two sections with links:

- Learning R/Rstudio for beginners**
 - Interactive lessons**
 - R bootcamp - interactive lessons in tidyverse/broom/stats, [Ted Laderas](#) and Jessica Minnier (me)
 - Rstudio Cloud primers - Interactive lessons (using learnr) on the basics of R, visualization, tidyverse. Requires a free Rstudio Cloud account.
 - Swirl
 - Slides/videos/online courses**
 - OHSU OCTRI BERD Workshops, with [audio recordings](#) - workshops taught by Jessica Minnier and Meike Niederhausen
 - Introduction to R and Rstudio [slides](#)
 - Data Wrangling in R with the Tidyverse [part 1 slides](#), [part 2 slides](#)

github.com/jminnier

We don't want to be left behind

- Work on projects outside of your comfort zone
- Grow as statisticians, grow as a field
- Be open to other fields collaborating on "data science"

Don't forget to have fun!

Thank you!

Contact me: [✉ minnier-\[at\]-ohsu.edu](mailto:minnier-[at]-ohsu.edu), [@datapointier](https://twitter.com/datapointier), [jminnier](https://github.com/jminnier)

Slides available: bit.ly/jmin-jsm19

Slide code and files available at:
github.com/jminnier/talks-etc

Slides created via the R package [xaringan](#) by [Yihui Xie](#) with the [xaringanthemer](#) package by [Garrick Aden-Buie](#), inspired by slides and slide formatting by [Alison Hill](#), [Jennifer Thompson](#), and [Chester Ismay](#)



Allison Horst