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

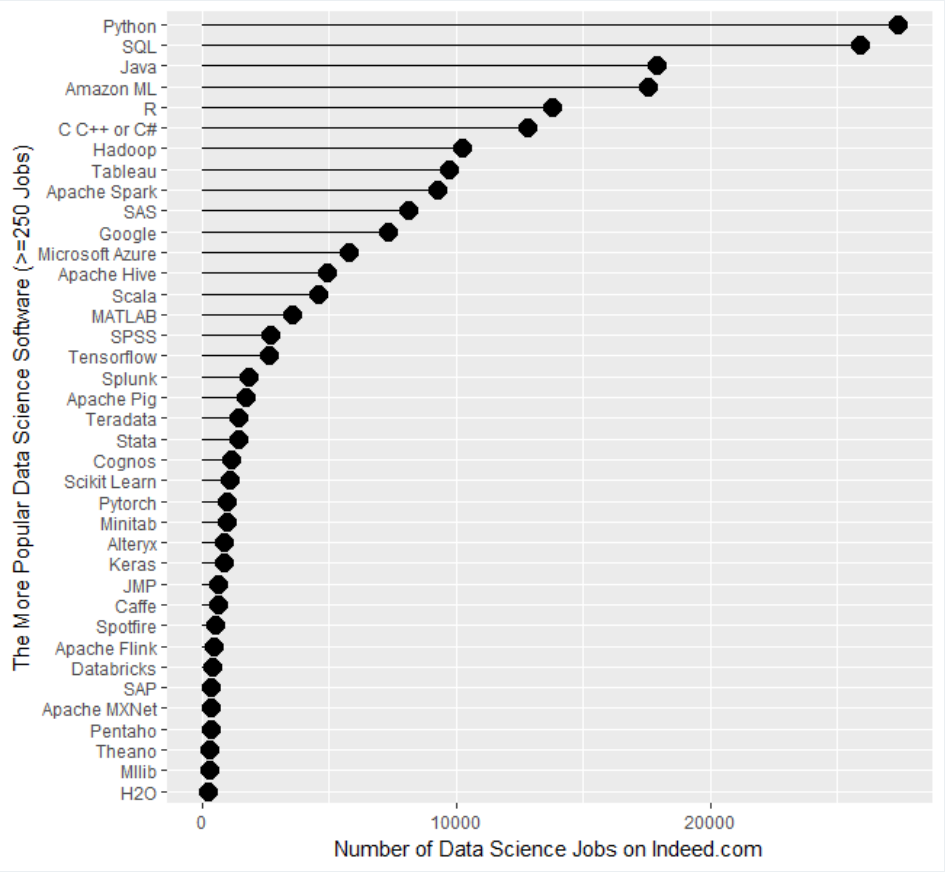
1 OCTOBER 2015 13,976 VIEWS 13 COMMENTS

> 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.

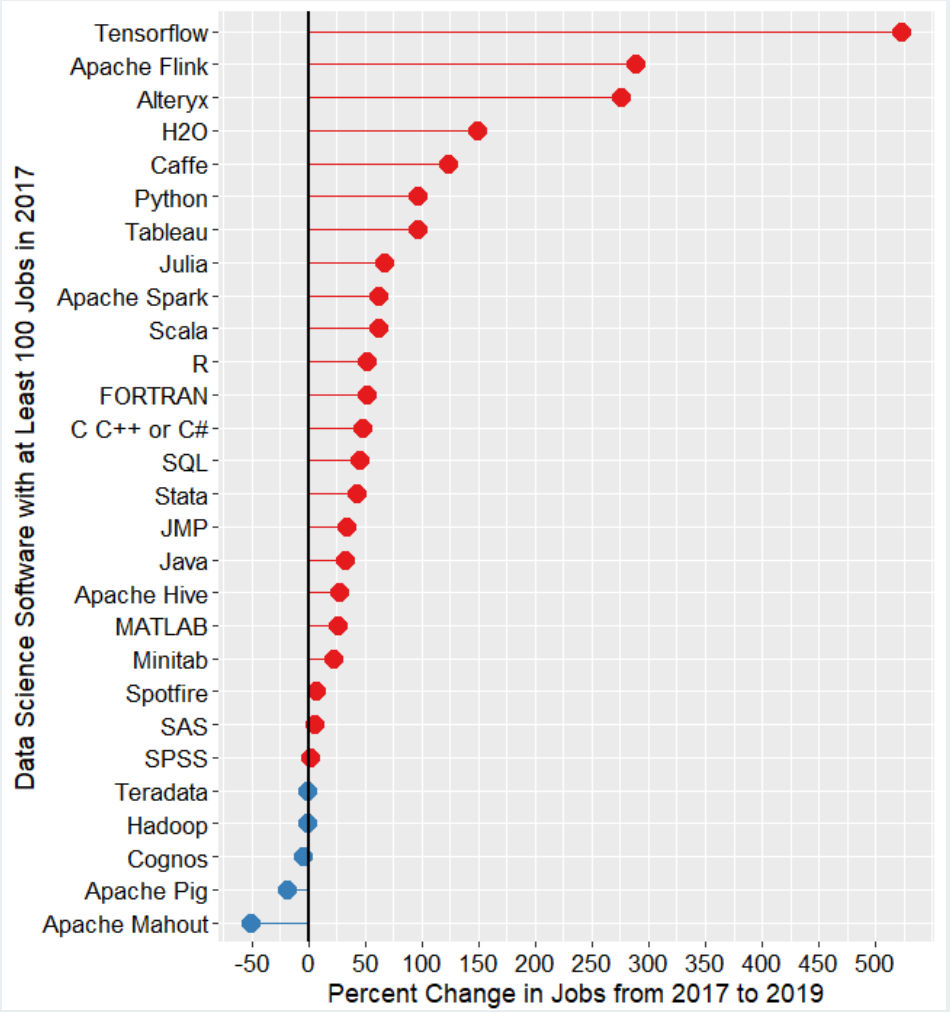
Programming itself is changing

- 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!

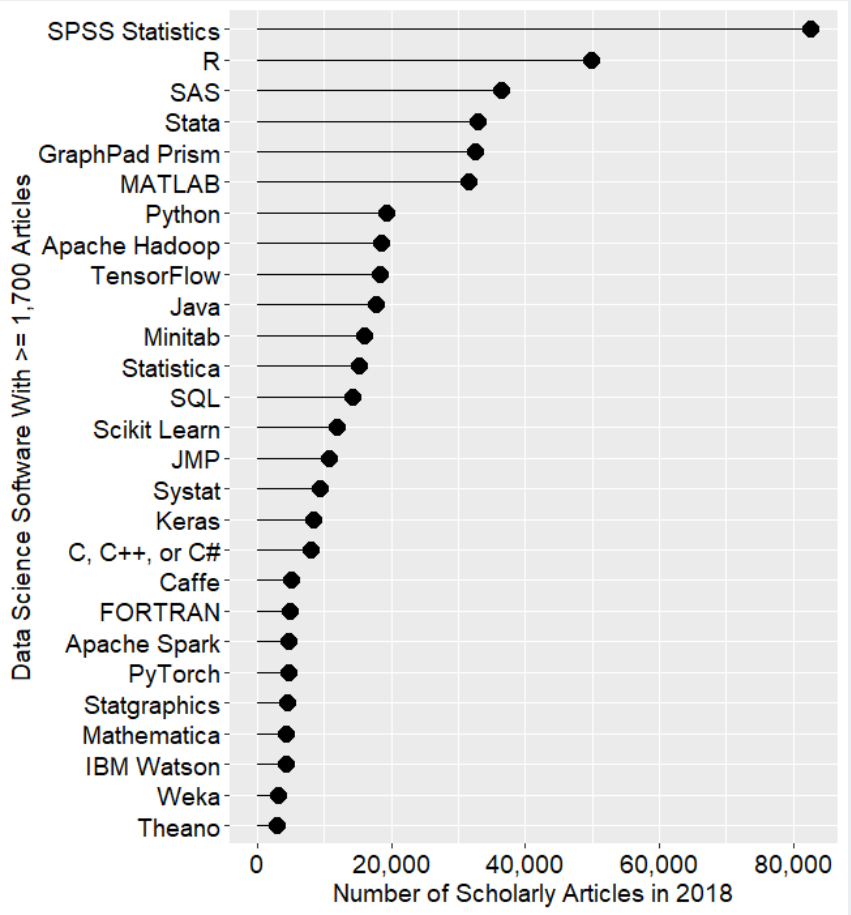
Data Science Jobs: Python, SQL, ... R

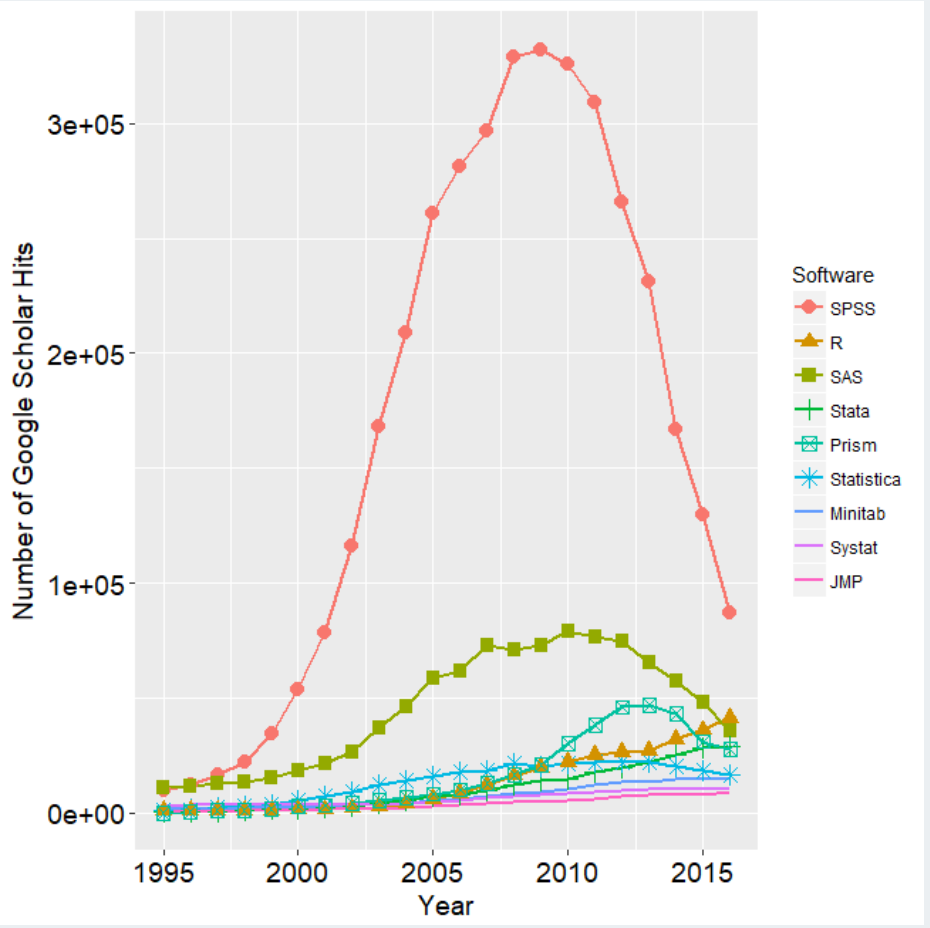


R4stats - Robert A. Muenchen



By contrast, scholarly articles: SPSS, R, SAS

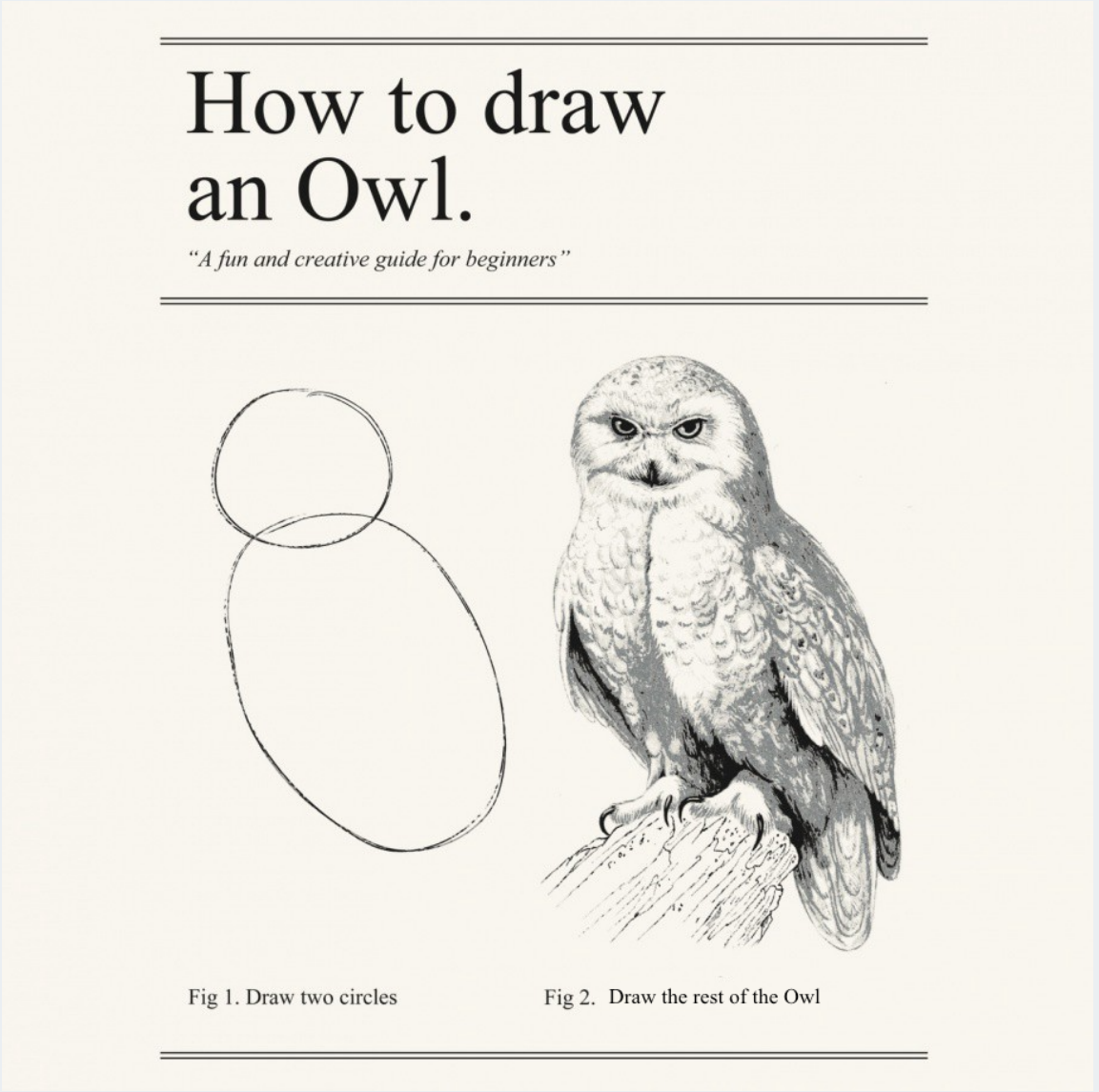




R4stats - Robert A. Muenchen

How to keep up?

Learning takes time,
dedication, courage.



How to draw
an Owl.

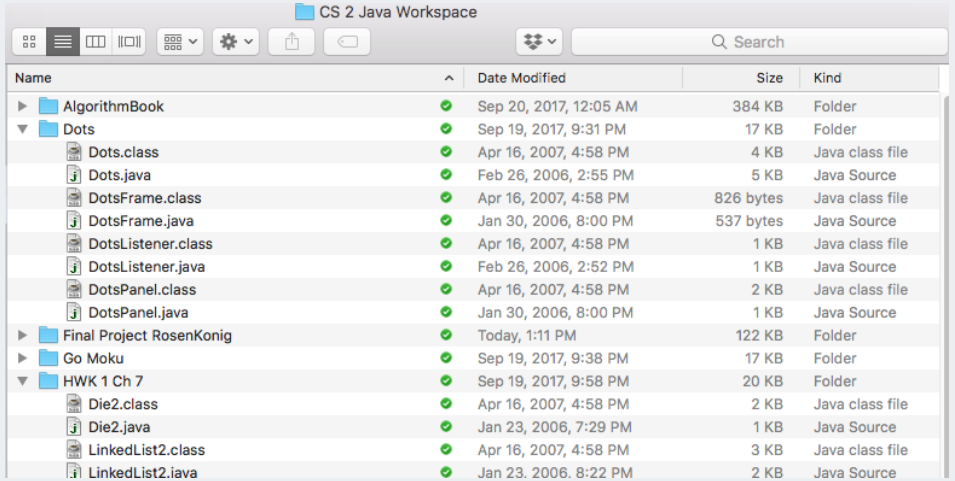
"A fun and creative guide for beginners"

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):

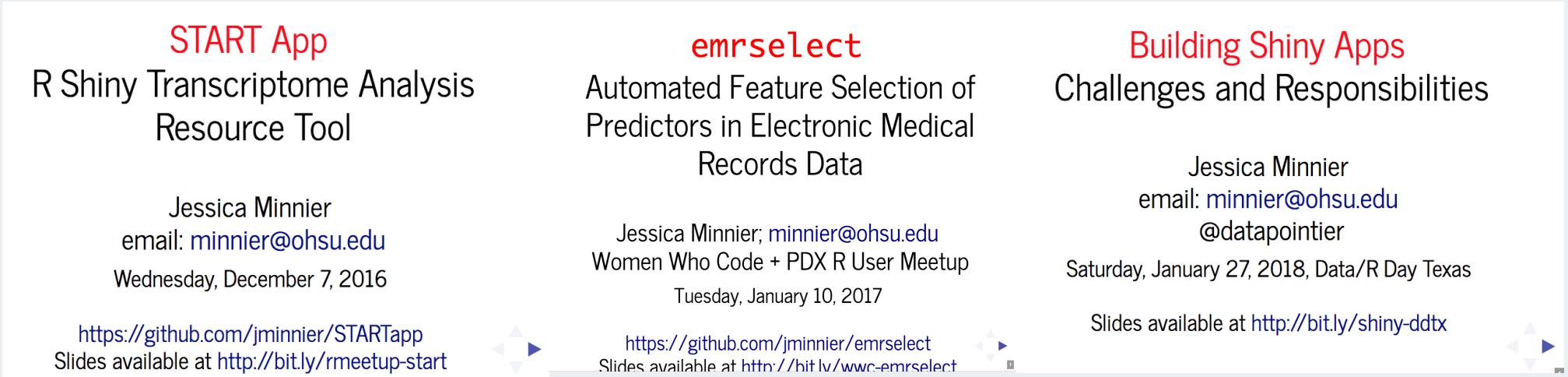


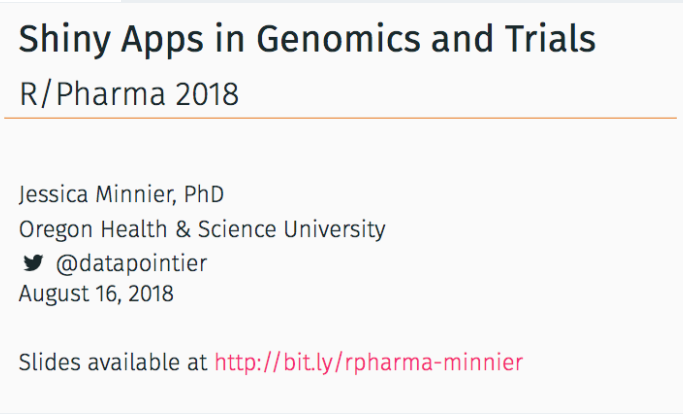
My PhD advisor wrote a *ton* of code

Postdoc (learned git/github):

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

Learning while teaching

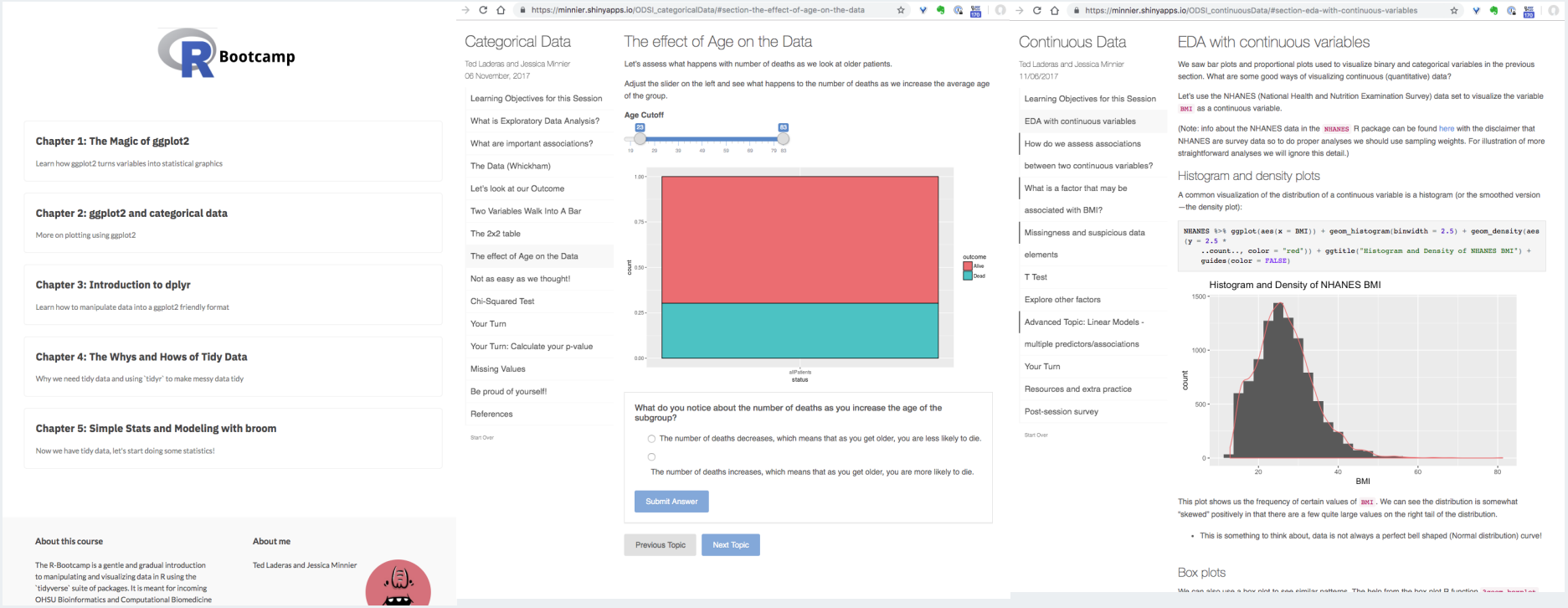




github.com/jminnier/talks_etc

Learning while teaching with new tools

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

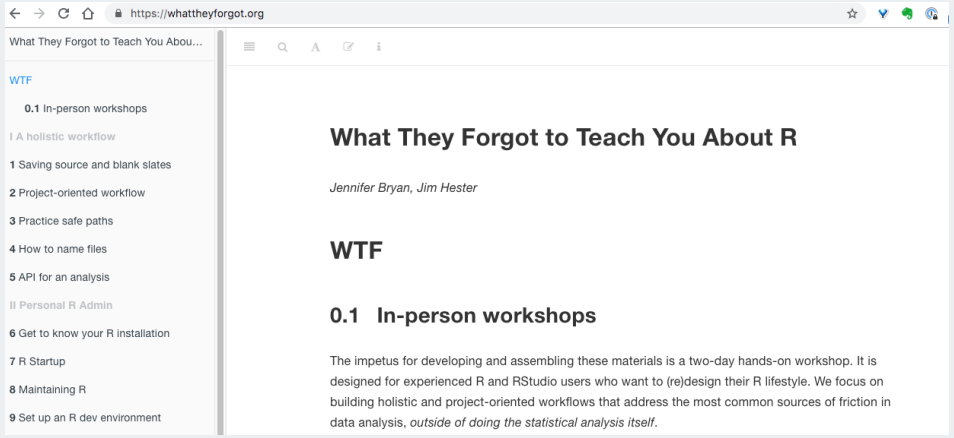


r-bootcamp.netlify.com; minnier.shinyapps.io/ODSI_categoricalData/; minnier.shinyapps.io/ODSI_continuousData

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)

https://software-carpentry.org/



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!

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

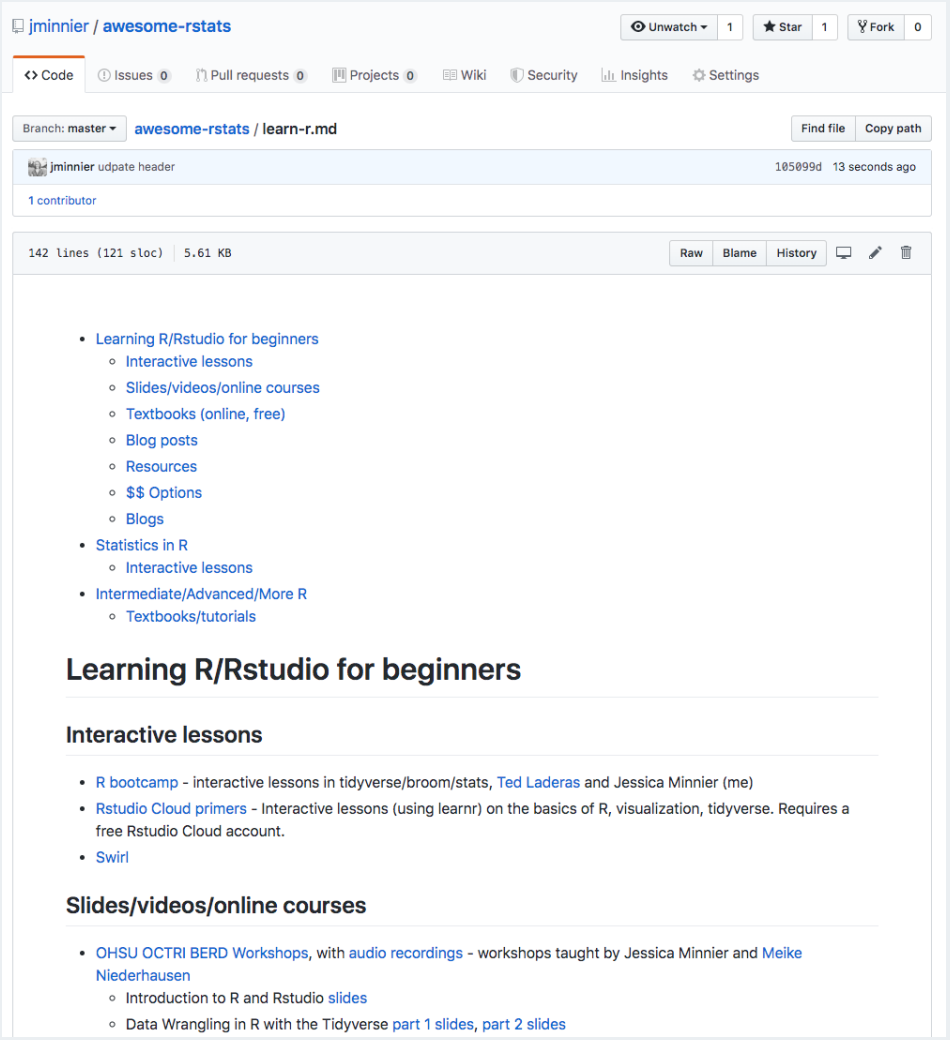
- Sign up to give talks at local meetups
- Teach workshops to students/fellow faculty/coworkers
- Journal/book clubs with your co-workers/team

Resources →

- Interactive lessons
- Online courses
- Blog posts

Ask for help

- Stack Overflow
- Rstudio Community



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, datapointier, 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