Getting started with R

Essential R for data manipulation

JGI Data Science Training
Wyatt Hartman
Feb 14, 2018

Slides and R code notebook available at

https://github.com/JGI-Bioinformatics/JGI-Data-Science/Getting-started-with-R/

What is R ?

Open source statistical programming language

Forked from “S” language developed by Bell Labs (1997)

Used by about 50% of data scientists

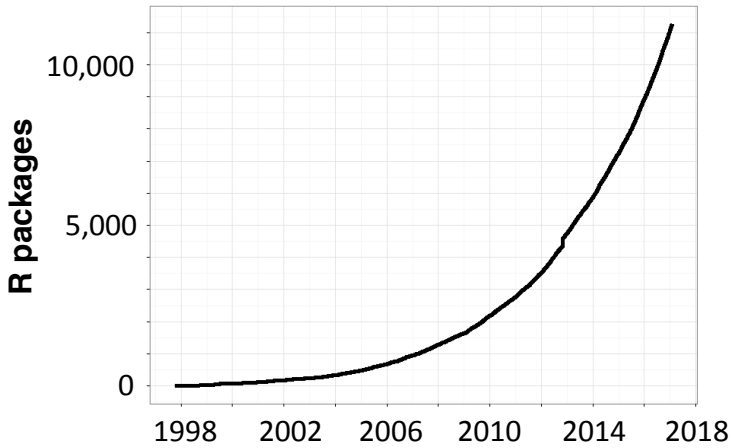
https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html/

“Least disliked” programming language

https://stackoverflow.blog/2017/10/31/disliked-programming-languages/

“Academic Language”

- New methods often deployed first
- Lots of learning resources
- vs. Python…



Why use R ?

Stop clicking, start coding

Eliminate repetition, make reproducible workflows

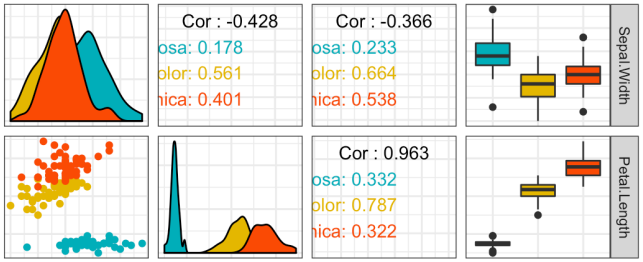
Build on top of work you've already done

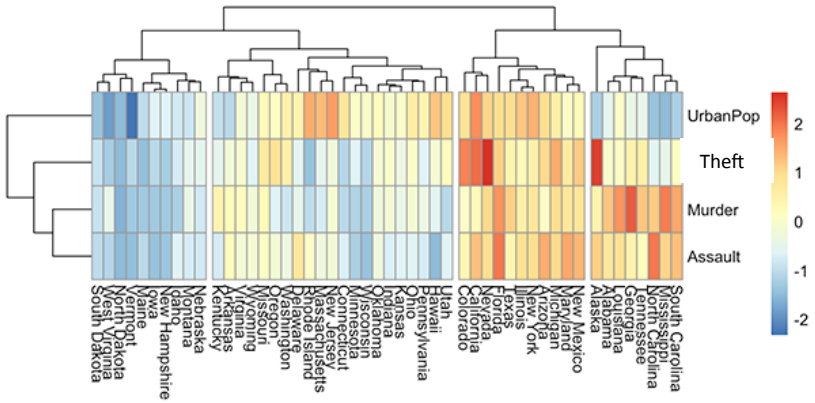
Stand on the shoulders of those who made code for you

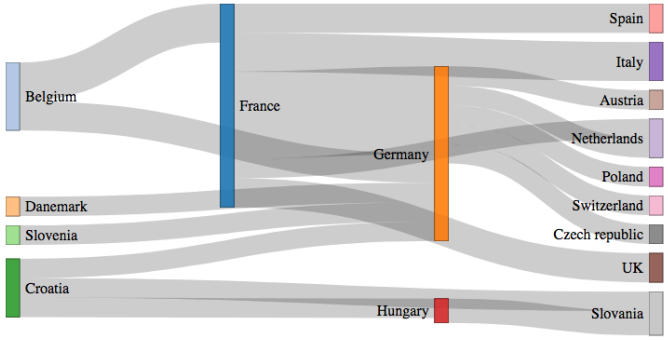
http://www.computergeekblog.com/7-excel-tips-tricks-impress-boss/

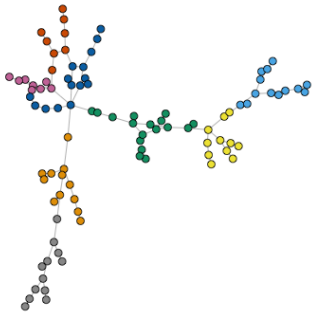
Why use R ?

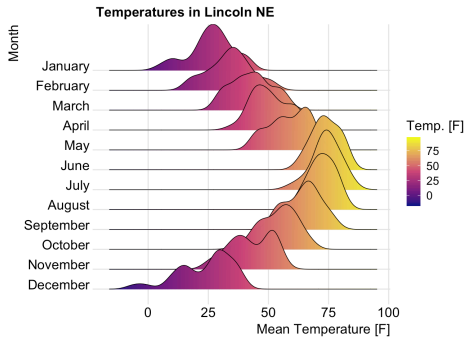
Striking data visualizations





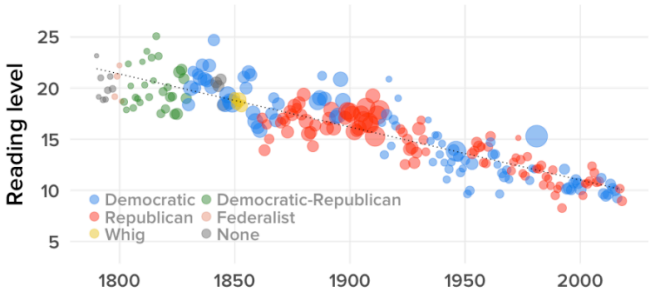


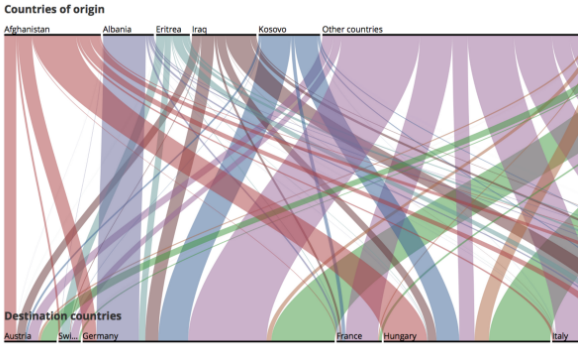


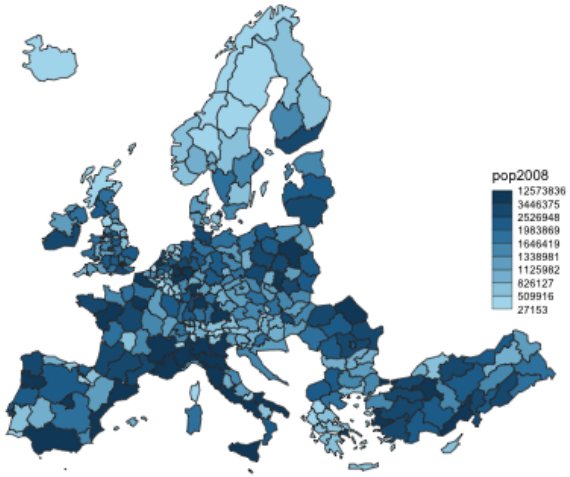


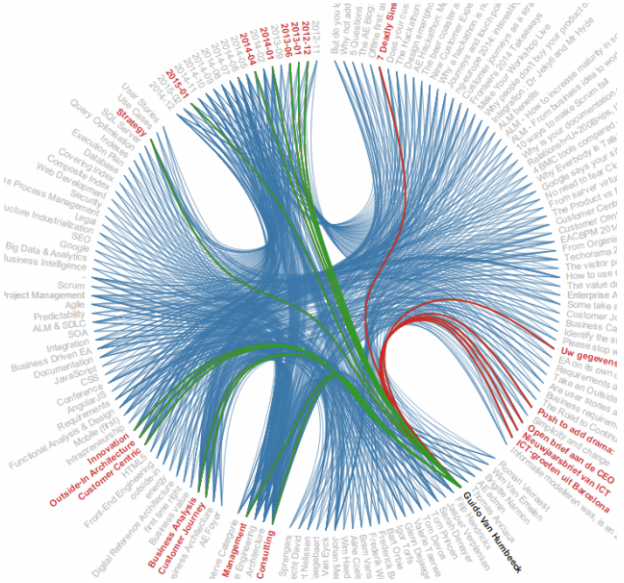
Why use R ?

Polished, publication quality graphics







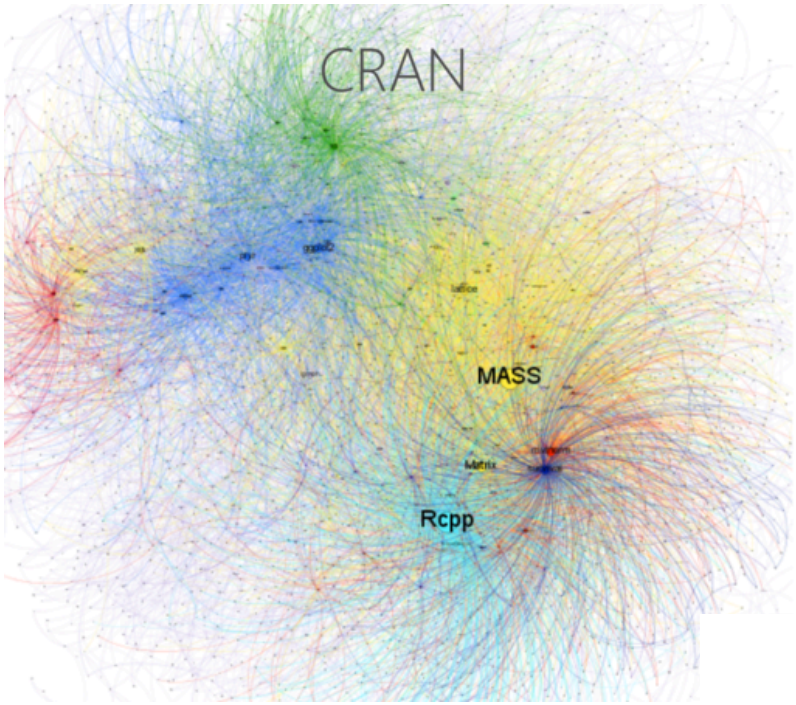


Why use R ?

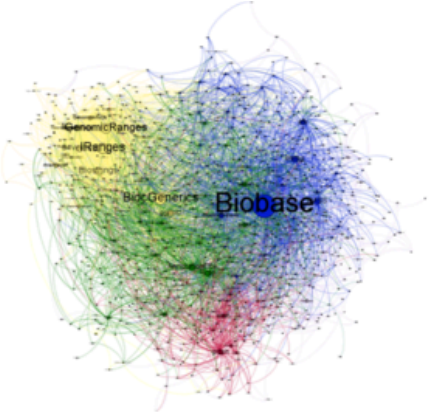
Extensive and powerful statistical libraries

> 11,000 R packages in CRAN
Comprehensive R Archive Network

> 6,000 Bioconductor packages
for analysis of genomic data



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



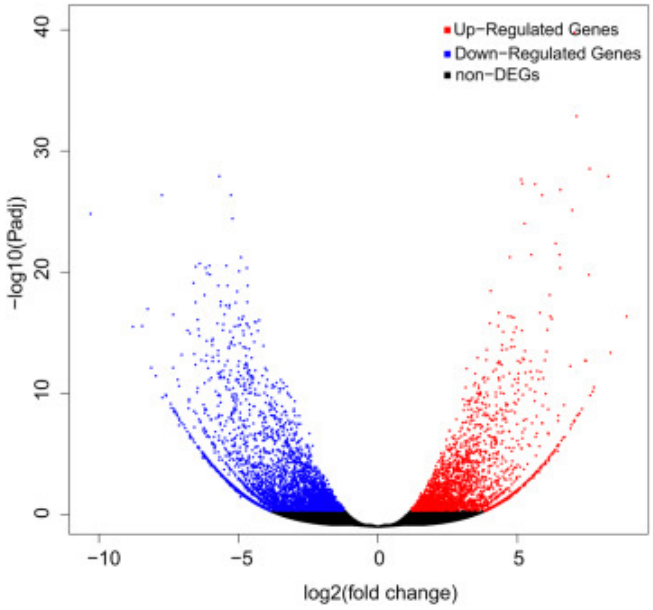
http://blog.revolutionanalytics.com/2015/08/differences-in-the-network-structure-of-cran-and-bioconductor.html

Why use R ?

Bioconductor: Popular sequence data workflows

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

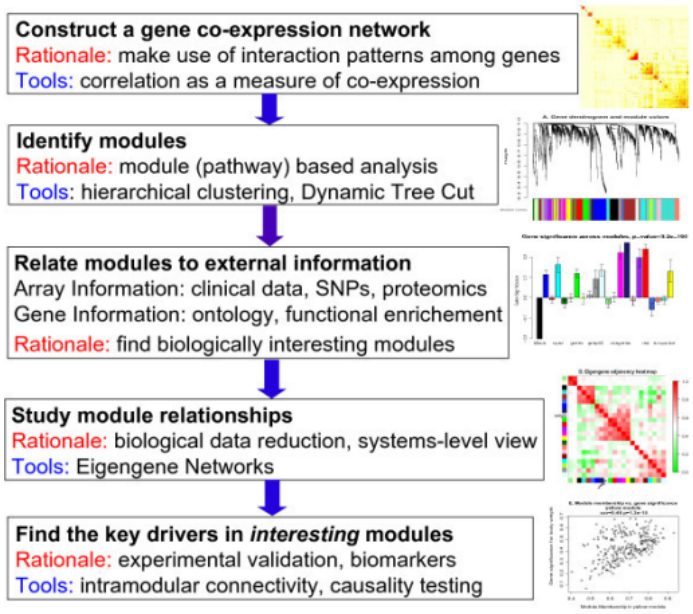
Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]



(2014) Genome Biol. 15: 550

WGCNA: an R package for weighted correlation network analysis

Peter Langfelder[1] and Steve Horvath*[2]

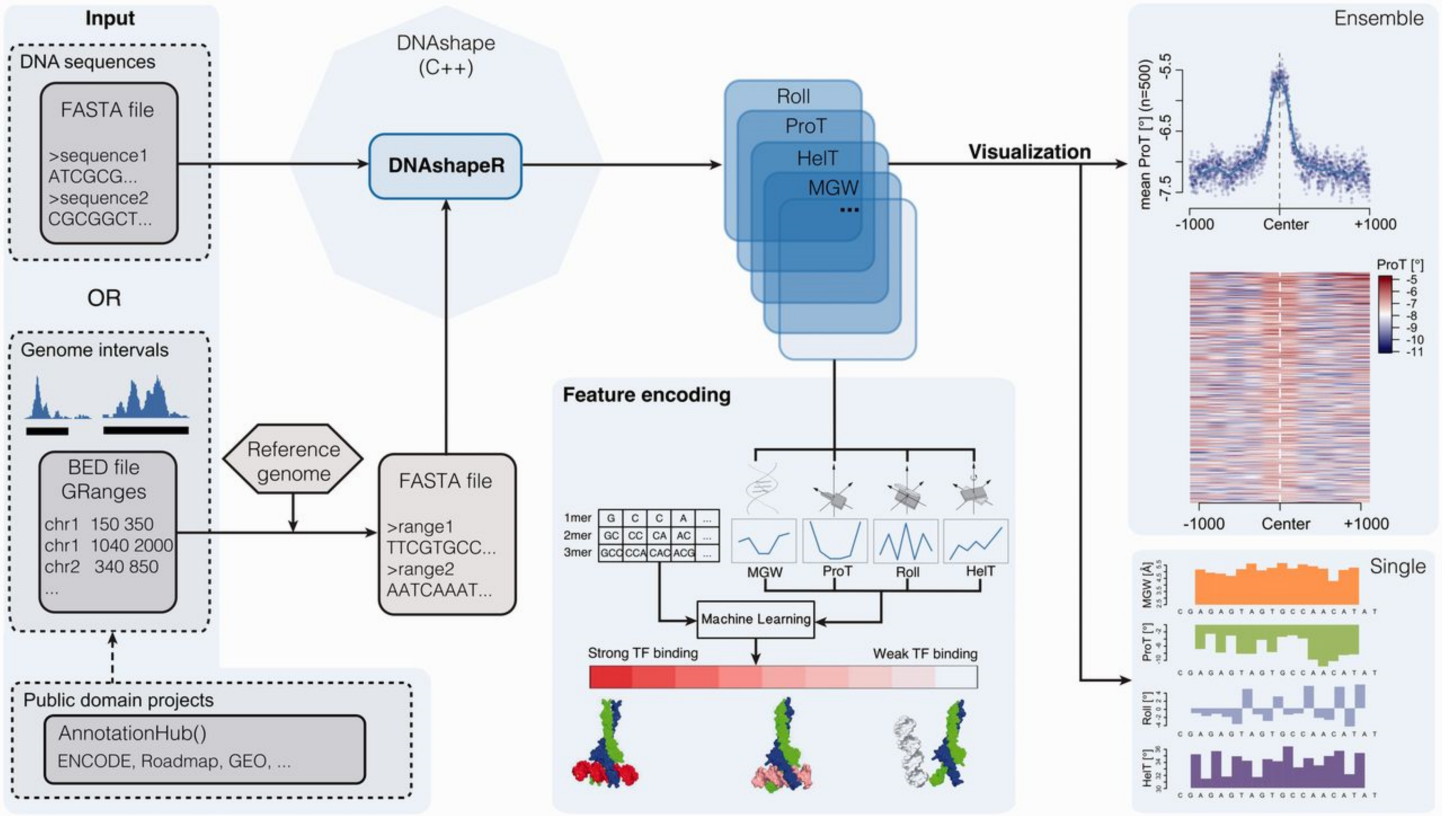


(2008) BMC Bioinf. 9: 559

Why use R ?

Intensive genomics analyses in Bioconductor

DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding



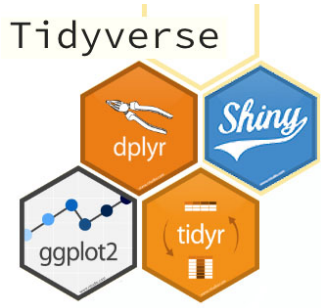
Bioinformatics. 2015;32(8):1211-1213. doi:10.1093/bioinformatics/btv735

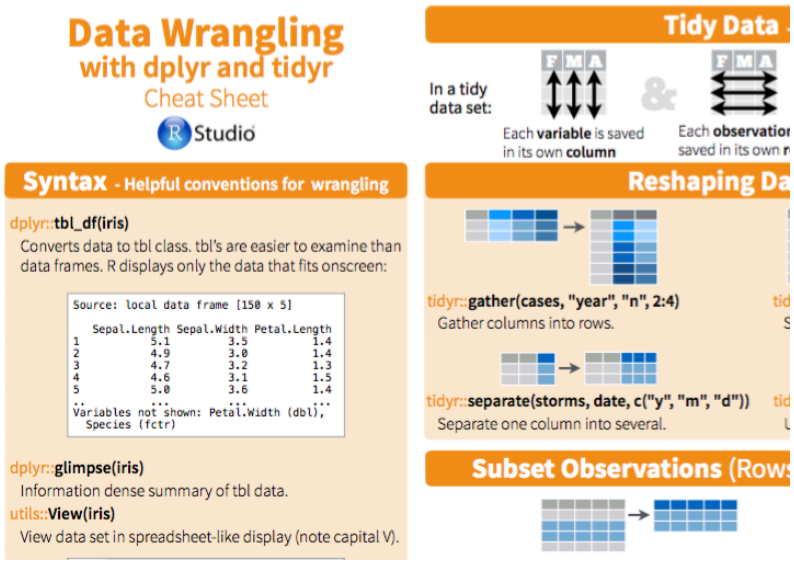
Why use R ?

Data manipulation and graphing libraries

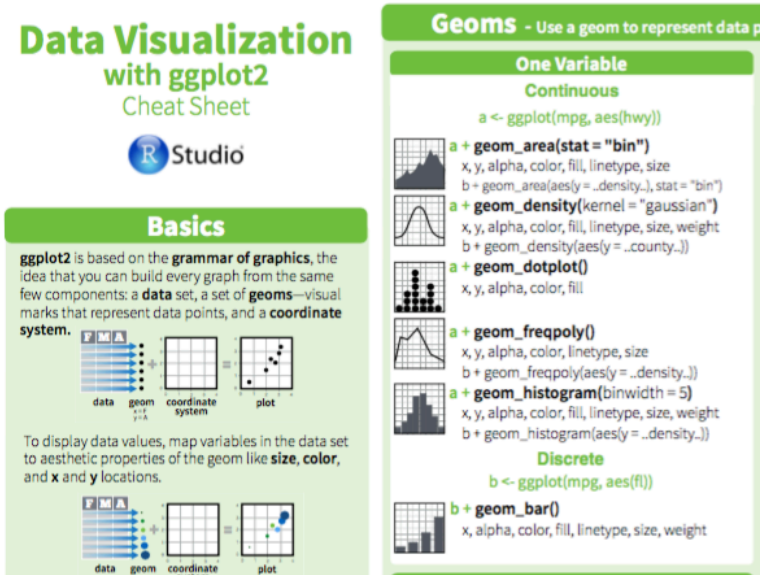
Modern, well documented libraries

cover the basics, cleanly





https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

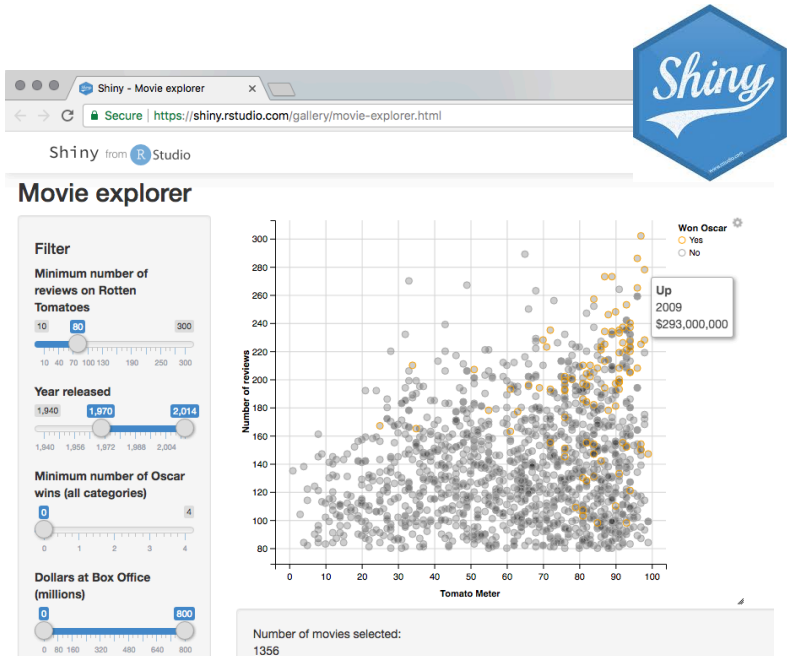


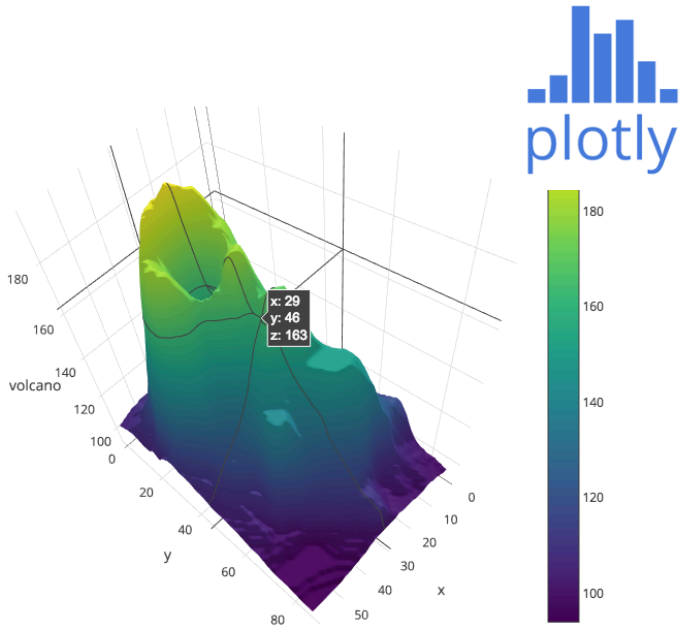
https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

Why use R ?

Interactive online graphics

Build dashboards and live graphics on the web





https://moderndata.plot.ly/interactive-r-visualizations-with-d3-ggplot2-rstudio/

http://shiny.rstudio.com/tutorial/

https://help.plot.ly/tutorials/

How do I use R ?

Rstudio is easiest to install yourself

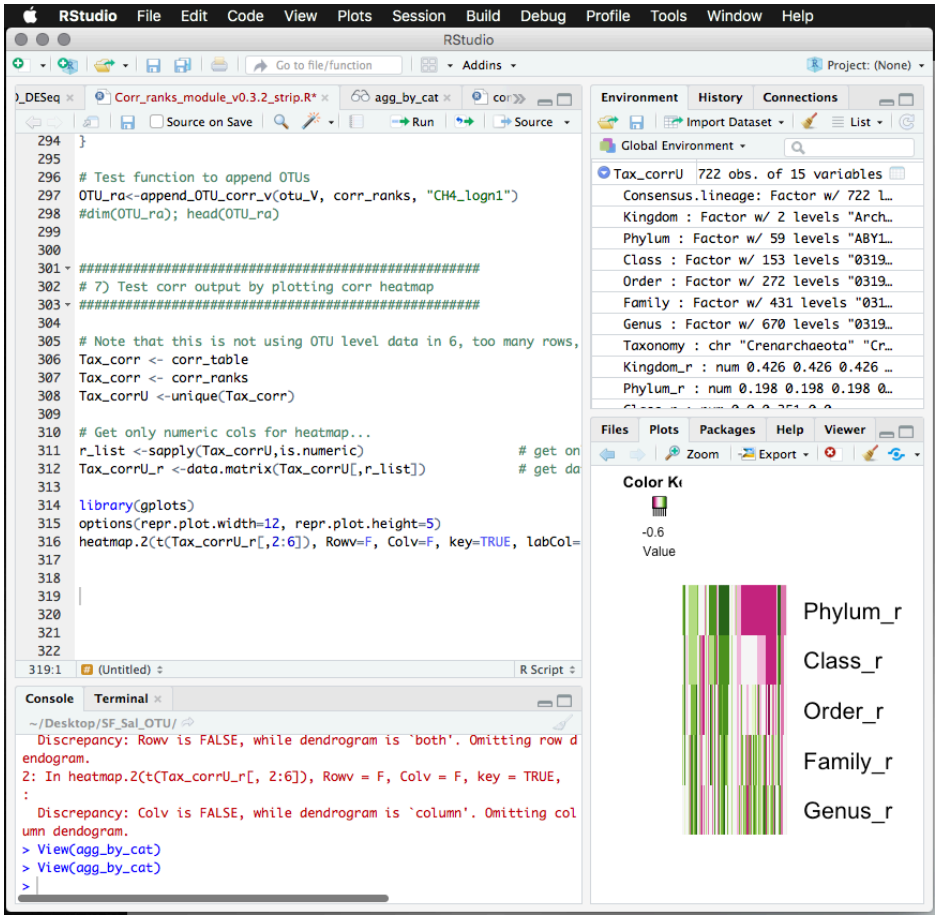
Easy to install software
Mac – Win - Linux

Helps track code, bugs, objects and graphics

Use Rstudio on NERSC:
https://rstudio.nersc.gov/

Install R studio:

https://www.rstudio.com/products/rstudio/download/



How do I use R ?

Jupyter notebooks: Portable R code

jupyter

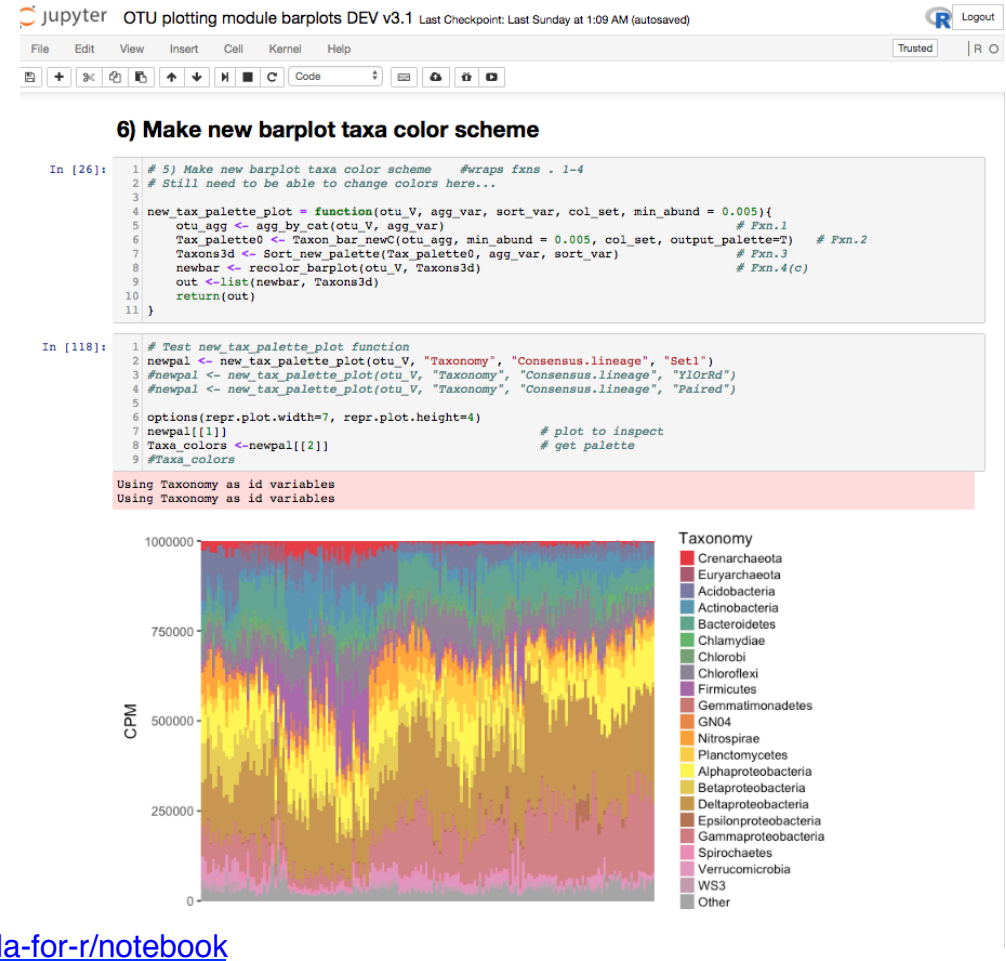
Runs in web browser,
Github compatibility

Simplified interface,
helps organize code

R in Jupyter on NERSC:

https://jupyter.nersc.gov/

Installation more complicated,
requires anaconda Python



https://anaconda.org/chdoig/jupyter-and-conda-for-r/notebook
https://www.datacamp.com/community/blog/jupyter-notebook-r

Getting started with R

Installing packages

> install.packages("ggplot2")

R studio:

Tools -> Install Packages

Bioconductor:

https://www.bioconductor.org/install/

> source('http://bioconductor.org/biocLite.R')
> biocLite('phyloseq')

Getting help

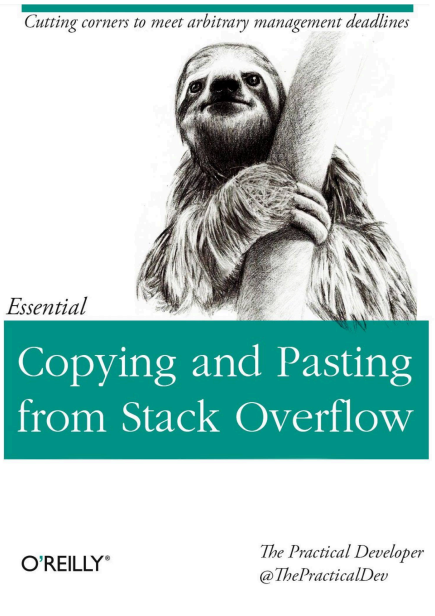
? Before a function or package gets help

> ?read.table > ?ggplot2

- Search for tutorials, else a Google rabbit hole…
  - Look for package **vignettes**

- Excellent tutorials for beginners here:

http://www.sthda.com/english/



Data lab objectives

Our R notebook exercise will help you learn to:

- Install and get help with R packages
- Get your data into (and out of) R
- Access and manipulate data
- Make simple data summaries and plots
- Massage data for plotting

Our next workshop covers graphics with the ggplot2 library

- This one gives you the tools you need to be ready, and a few examples