STAT 545A: Class Meeting 001

Course Intro + prompts to install and sign up for lots of stuff!

Web companion: STAT 545 web home > syllabus > cm001

Tuesday, September 5, 2017

Who am I?

Dr. Vincenzo Coia
Department of Statistics
University of British Columbia

vincen.coia@stat.ubc.ca
https://github.com/vincenzocoia
https://twitter.com/VincenzoCoia

- Background:
  - PhD in Statistics at UBC
  - MSc in Mathematical Statistics at Brock U
  - BSc in Math + Stats at Brock U
  - BSc in Biology w/ Earth Science at Brock U

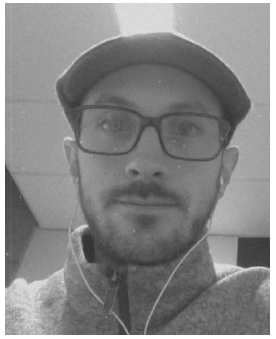
Meet the Teaching Team!

Vincenzo Coia, Instructor

Katharine Sedivy-Haley, TA

Pedro Gonzalez-Espinosa, TA

Derek Cho, TA



Giulio Valentino Dalla Riva, Instructor

Joey Bernhardt, TA

Ke Dai, TA

Instructor Schedule

STAT 545A

Week	cm	Date	Instructor
1	cm001	Sept. 05	Vincenzo
	cm002	Sept. 07	Vincenzo
2	cm003	Sept. 12	Vincenzo
	cm004	Sept. 14	Vincenzo
3	cm005	Sept. 19	Vincenzo
	cm006	Sept. 21	Vincenzo
4	cm007	Sept. 26	Vincenzo
	cm008	Sept. 28	Giulio
5	cm009	Oct. 03	Giulio
	cm010	Oct. 05	Giulio
6	cm011	Oct. 10	Vincenzo
	cm012	Oct. 12	Vincenzo
7	cm013	Oct. 17	Vincenzo
	cm014	Oct. 19	Tamara Munzner

STAT 547M

Week	cm	Date	Instructor
1 (8)	cm101	Oct. 24	Giulio
	cm102	Oct. 26	Giulio
2 (9)	cm103	Oct. 31	Giulio
	cm104	Nov. 02	Giulio
3 (10)	cm105	Nov. 07	Giulio
	cm106	Nov. 09	Giulio
4 (11)	cm107	Nov. 14	Giulio
	cm108	Nov. 16	Giulio
5 (12)	cm109	Nov. 21	Vincenzo
	cm110	Nov. 23	Vincenzo
6 (13)	cm111	Nov. 28	Giulio
	cm112	Nov. 30	Giulio

Culture of the class

- Teaching you to fish (vs. giving you a fish)
  - It's amazing what a determined individual can learn from documentation, small learning examples, and ... <gasp> Googling. And also stackoverflow.
- Rewarding engagement, intellectual generosity and curiosity
  - Speaking up, sharing success OR failure, showing some interest in something will earn marks.
- Zero tolerance of plagiarism
  - Generating your own approach, writing some code, and describing the process is the whole point. Process is generally more important than product.

THE UNIVERSITY OF BRITISH COLUMBIA

UBC BROADCAST EMAIL

If you are a manager of staff whose work is not computer-based, please print this email and display it in a common work area for them to review.

To: Faculty, staff and students in Vancouver and the Okanagan

Hate is antithetical to UBC's commitment to diversity and inclusion

In recent months we have all sadly witnessed incidents of hate, racism and xenophobia across the globe.

Tragedies such as today's attack in Barcelona, and the unrest in Charlottesville, only serve to remind us that we cannot *and must not* tolerate hatred and racism.

These incidents are antithetical to the very fabric of a just society. In particular, they are antithetical to UBC's commitment to diversity and inclusion and to the values we hold and champion as an academic institution.

We must remember that Canada is not immune to similar challenges and we must work together to build a just and inclusive society.

As we welcome faculty, staff and especially new and returning students to UBC at the start of the coming academic year, let us work to ensure that *everybody* feels welcome and safe at our university.

Professor Santa J. Ono,
President & Vice-Chancellor

Where marks will come from

- Weekly homework; marked coarsely (think check, check minus, check plus), with peer evaluation
- Eventually, flexibility to work with a dataset you choose or to spin the problem a certain way
  - Think about datasets you'd like to prepare and analyze!
- Adjust the difficulty level relative to where you are now (and where you need/want to be!)
- Peer review; marked coarsely (good review vs. "needs more")
- Engagement and participation: in class, in our GitHub world

stat545-ubc.github.io

Home FAQ Syllabus Topics People

Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

Selected topics

- Introduction to R and the RStudio IDE; scripts, the workspace, RStudio Projects
- Generate reports from R scripts and R Markdown
- Care and feeding of data in R
- Data aggregation; "apply" functions, `plyr`, `dplyr`
- Data visualization with `ggplot2`
- Graphs and descriptive stats for quantitative and categorical variables
- Writing R functions
- Coding style and project organization
- Version control with Git; collaboration via GitHub
- Character data; regular expressions
- Interactive pages, apps, and graphics with Shiny and `ggvis`
- Get data off the web and expose data, code, results on the web
- Distribute data and code via an R package
- Automate an analytical pipeline, e.g. via `Make`

More info?

Use the navigation bar above!

@STAT545 on Twitter

The course organization on GitHub: https://github.com/STAT545-UBC

Repo that creates this website: https://github.com/STAT545-UBC/STAT545-UBC.github.io

http://stat545-ubc.github.io

Home

Data wrangling

UBC STAT 545A and 5

Learn how to

- explore, groom, visualize, and analyze
- make all of that reproducible, reusabl
- using R

Selected topics

- Introduction to R and the RStudio IDE
- Generate reports from R scripts and R
- Care and feeding of data in R
- Data aggregation; "apply" functions,
- Data visualization with `ggplot2`
- Graphs and descriptive stats for quant
- Writing R functions
- Coding style and project organization
- Version control with Git; collaboratio
- Character data; regular expressions
- Interactive pages, apps, and graphics
- Get data off the web and expose data
- Distribute data and code via an R pac
- Automate an analytical pipeline, e.g. v

More info?

Use the navigation bar above!

@STAT545 on Twitter

The course organization on GitHub: https://github.com/STAT545-UBC

Repo that creates this website: https://github.com/STAT545-UBC/STAT545-UBC.github.io

STAT545-UBC/STAT545-UBC.github.io

This repository Search Explore Gist Blog Help jennybc

STAT545-UBC / STAT545-UBC.github.io

Unwatch 3 Star 1 Fork 1

Main repository for STAT 545 @ University of British Columbia, a course in data wrangling, exploration, and analysis with R. http://stat545-ubc.github.io — Edit

31 commits 1 branch 0 releases 5 contributors

branch: master STAT545-UBC.github.io / +

Compile html version for @jooolia

BernhardKonrad authored 17 minutes ago latest commit 09d7176cca

course-admin	Generate link-y course chronology	12 hours ago
include	Mirror gh-pages content in master	a day ago
libs	Mirror gh-pages content in master	a day ago
.gitignore	Mirror gh-pages content in master	a day ago
README.md	Create README.md	8 hours ago
_output.yaml	Mirror gh-pages content in master	a day ago
block000_r-rstudio-install.html	R and RStudio install instructions	4 hours ago
block000_r-rstudio-install.md	R and RStudio install instructions	4 hours ago
block001_git-install.html	Git install instructions	2 hours ago
block001_git-install.md	Git install instructions	2 hours ago
cm001_course-intro-sw-install-acc...	Update/correct Git client info	2 hours ago
cm001_course-intro-sw-install-acc...	Update/correct Git client info	2 hours ago
faq.Rmd	Words re marks	13 hours ago
faq.html	Generate link-y course chronology	12 hours ago

Code Issues 1 Pull Requests 0 Pulse Graphs Settings

HTTPS clone URL https://github.com/S

You can clone with HTTPS, SSH, or Subversion.

Clone in Desktop Download ZIP

stat545-ubc.github.io

Home FAQ Syllabus Topics People

Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

Selected topics

- Introduction to R and the RStudio IDE; scripts, the workspace, RS
- Generate reports from R scripts and R Markdown
- Care and feeding of data in R
- Data aggregation; "apply" functions, `plyr`, `dplyr`
- Data visualization with `ggplot2`
- Graphs and descriptive stats for quantitative and categorical var
- Writing R functions
- Coding style and project organization
- Version control with Git; collaboration via GitHub
- Character data; regular expressions
- Interactive pages, apps, and graphics with Shiny and `ggvis`
- Get data off the web and expose data, code, results on the web
- Distribute data and code via an R package
- Automate an analytical pipeline, e.g. via `Make`

More info?

Use the navigation bar above!

@STAT545 on Twitter

The course organization on GitHub: https://github.com/STAT545-UBC

Repo that creates this website: https://github.com/STAT545-UBC/STAT54

Home FAQ Syllabus

Syllabus

Until we truly get rolling, you can also see a chronology of the course from 2013.

STAT 545A

date	notes
sep-01 mon	no class; Labor Day
sep-03 wed	cm001: Intro to course; S/W install; acct sign-ups
sep-08 mon	
sep-10 wed	

stat545-ubc.github.io/cm001_course-intro-sw

Home FAQ Syllabus

- cm001 2014-09-03 Wednesday overview
  - Git
  - R and RStudio
  - Twitter
  - GitHub
  - Git(Hub) client
  - RPubs

cm001 2014-09-03 Wednesday overview

- Introduction to the course *slides to be posted after class*

web companion: STAT 545 web home > Syllabus > cm001

stat545-ubc.github.io/cm001_course-intro-sw-install-account-signup.html

Home FAQ Syllabus Topics People

- cm001 2014-09-03 Wednesday overview
  - Git
  - R and RStudio
  - Twitter
  - GitHub
  - Git(Hub) client
  - RPubs

cm001 2014-09-03 Wednesday overview

- Introduction to the course *slides to be posted after class*
- Students to do for Monday 2014-09-08 (see below)
  - install software
  - sign up for accounts
- Don't panic if there are glitches. That's normal.
- What's coming? We'll help you begin to use all your new toys next week!

Homework!

Git

Follow these instructions.

R and RStudio

Follow these instructions. Stick with it to the bitter end, where you try to get Git and RStudio talking to each other.

If you have some ancient version or R and/or RStudio, just go ahead and update! It often requires more technical skill to function with old software (and hardware).

Twitter

https://twitter.com

I will use the @STAT545 Twitter account to make micro-announcements, share interesting links, and facilitate a conversation amongst ourselves in public.

Follow these instructions. Stick with it to the bitter end, where you try to get Git and RStudio talking to each other.

If you have some ancient version or R and/or RStudio, just go ahead and update! It often requires more technical skill to function with old software (and hardware).

Twitter

https://twitter.com

I will use the @STAT545 Twitter account to make micro-announcements, share interesting links, and facilitate a conversation amongst ourselves **in public**.

In class, we'll talk about Twitter, its scholarly use, and privacy. Some relevant links:

- Disposable Twitter Accounts for Classroom Use from ProfHacker on the Chronicle of Higher Education
  - Good quote: "I strongly encourage you to create a disposable account if for any reason you prefer not to share your personal account for classroom activities."
- Resources for exploring digital identity, privacy and authenticity from Catherine Cronin

Twitter, GitHub ... are PUBLIC

cultivate your professional / scholarly profile with intention

if you join, make sure @STAT545 follows you back

CAREERS

NATURAL SOLUTIONS Seeking cancer cures in native flora **p.615**

WORDS COUNT Men drop female physicians' titles in talks **p.615**

STEP UP A course to quell harrassment on sight **p.615**

TOBY KEANE/ALAN TURING INSTITUTE

The Alan Turing Institute in London is an interdisciplinary hub for the growing field of data science.

INFORMATION MANAGEMENT

Data domination

Software programming, algorithm development and other technological skills can give scientists an edge in their fields.

BY GAIA DONATI AND CHRIS WOOLSTON

Karthik Ram had to reinvent himself in 2009, as have many other scientists in this data-driven age. When he started his postdoctoral work on how climate change affects elk in Yellowstone National Park in Wyoming, he thought of himself as an ecologist. But interpreting data from satellites and the tracking collars used to follow the animals pushed him to expand that mindset.

To make sense of the shifting ecosystem, he had to hone his programming and learn how to manage mountains of information — skills that have changed the way he views himself and his career. "I use the term 'ecologist' less and less often," he says. "Now, I mainly call myself a data scientist."

Data science was a young field in 2009, but it has quickly matured, and now intersects with many disciplines. Although its definition varies, data science generally involves using computing tools to manage and interpret large data sets.

Ram, now at the Berkeley Institute for Data Science at the University of California, Berkeley, works with former neuroscientists, social scientists and biologists who have also moved into the world of data. "Everyone at the institute is like me," he says. "We have computational skills and statistical skills that we can bring to bear on our particular fields."

The demand for data scientists has expanded beyond academia to industry, health care, government and any institution that generates complex information. IBM projects that there could be more than 2.7 million US jobs in data science and analytics by 2020, a 15% increase from 2015. Numbers are similar in Europe, according to the European Data Science Academy, a training and education group that identifies and collects job advertisements in Europe seeking data-science skills. The academy has identified more than 3 million such ads since 2015, including 290,000 posted during a 3-month period this year.

For those seeking a data-scientist role, the challenge isn't so much finding a job, but finding the best position for their aptitudes and interests (see 'Dig into the data world'). Identifying "the right fit can be tricky", says Amelia Taylor, a former tenure-track mathematician at Colorado College in Colorado Springs and now a data scientist for Zymergen, a company based in Emeryville, California, that is developing new uses for genetically engineered microbes. "Data science can look very different at different places. There are so many companies out there, it's hard to know which ones to look at."

Too many options — when many PhD holders in other fields are facing too few — is a good problem to have. Scientists who develop the right skills and understand their opportunities can expect a rewarding, data-driven future.

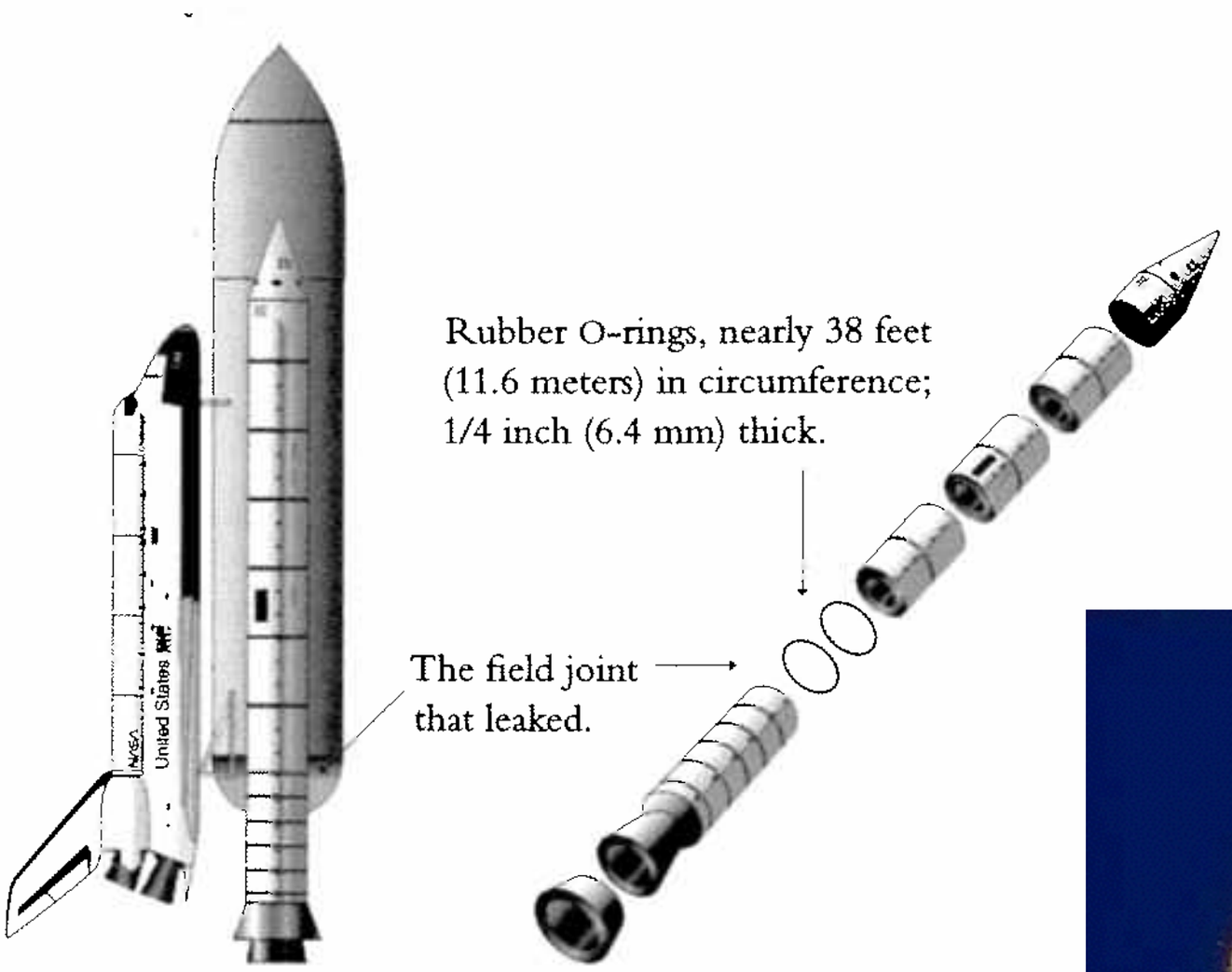
A MULTITUDE OF ROLES

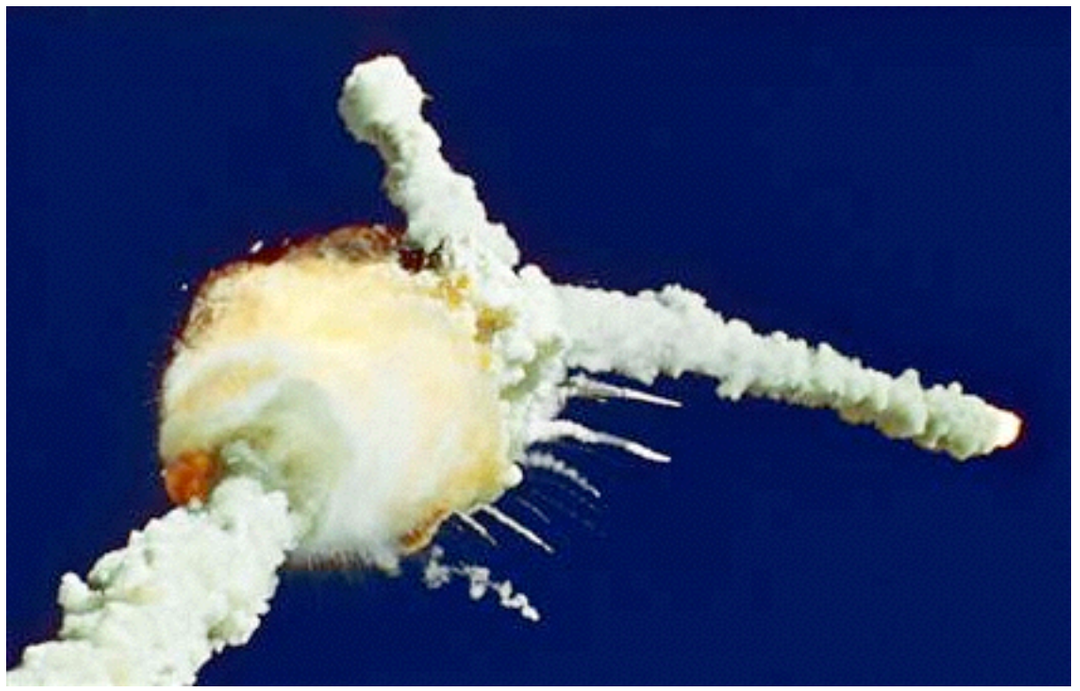
The rising tide of data science has lifted many boats. Along with a surge in 'data-scientist' searches, 'data engineer' and 'data analyst' are also popular terms on job-search boards. The differences in these roles are subtle but important. "The core skill of a data engineer is building robust systems that won't fail," explains Marc Warner, chief executive at ASI Data Science, a London-based firm that offers consulting services and a data-science fellowship programme with industry placements.

One key difference between data scientists and analysts, he says, is that scientists tend to follow data where they lead them — a 'data-first' approach — whereas analysts generally use numbers to test an established hypothesis.

“A picture is worth a thousand words”

1986 Challenger space shuttle disaster
Favorite example of Edward Tufte





http://msnbcmedia1.msn.com/j/msnbc/Components/Photos/050709/050609_columbia_hmed_6p.hmedium.jpg

TEMPERATURE CONCERN ON SRM JOINTS

27 JAN 1986

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°--66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

BLOW BY HISTORY

SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

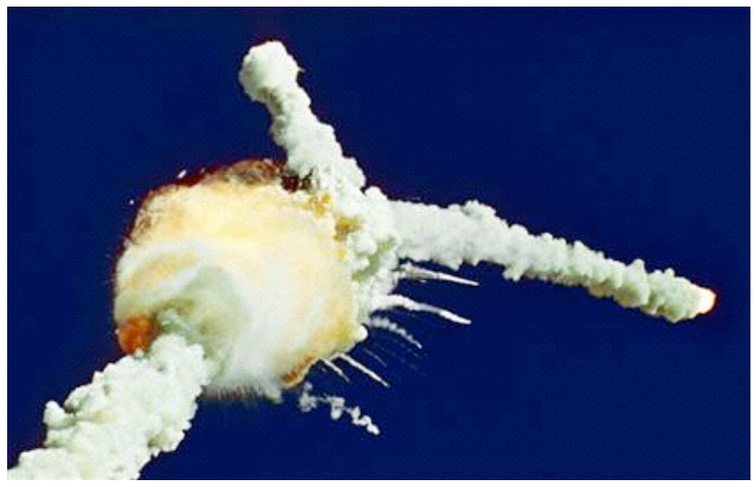
SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

MOTOR	O-RING
DM-1	47
DM-2	52
QM-3	48
QM-4	51
SRM-15	53
SRM-22	75
SRM-25	29 27

“A picture is worth a thousand words”

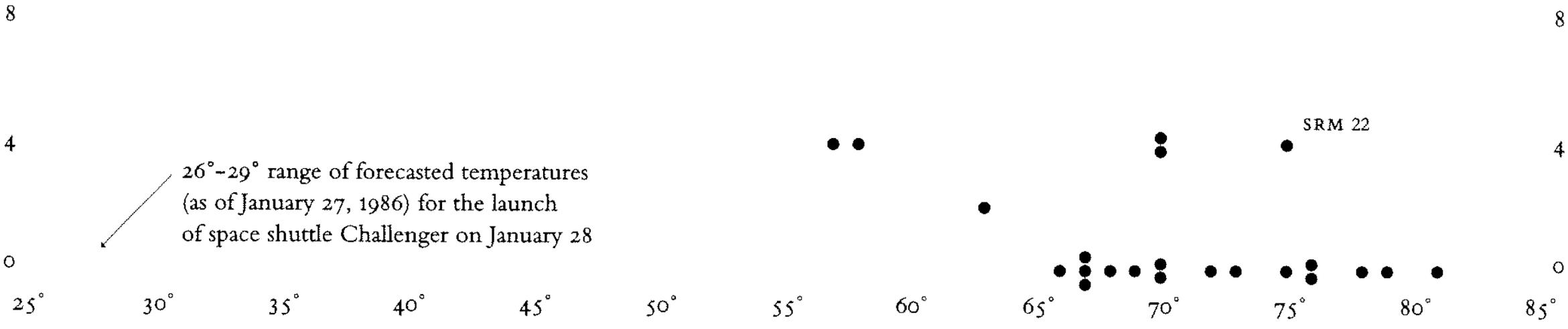


O-ring damage
index, each launch

12

12

SRM 15



Temperature (°F) of field joints at time of launch

“A picture is worth a thousand words”

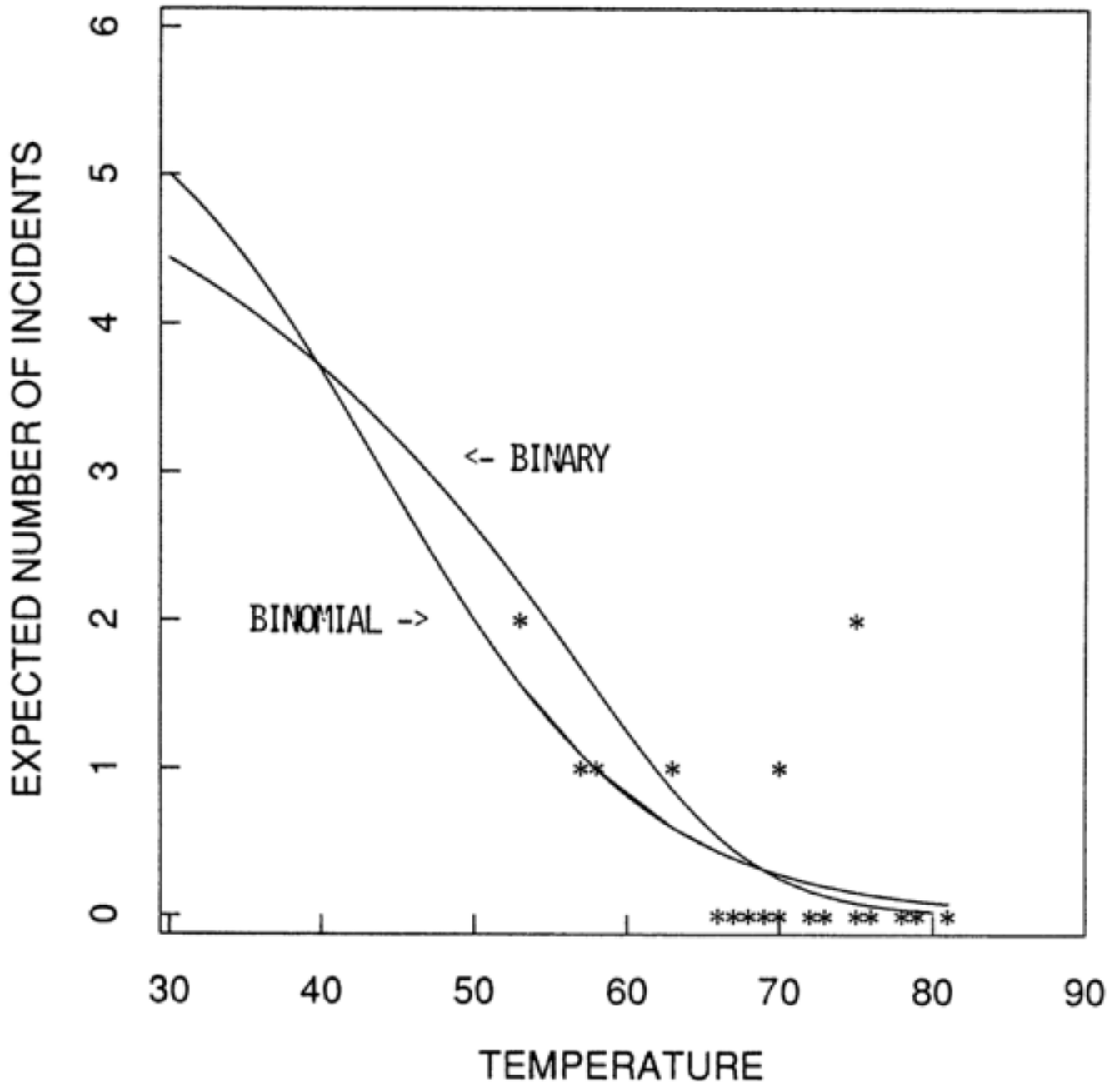
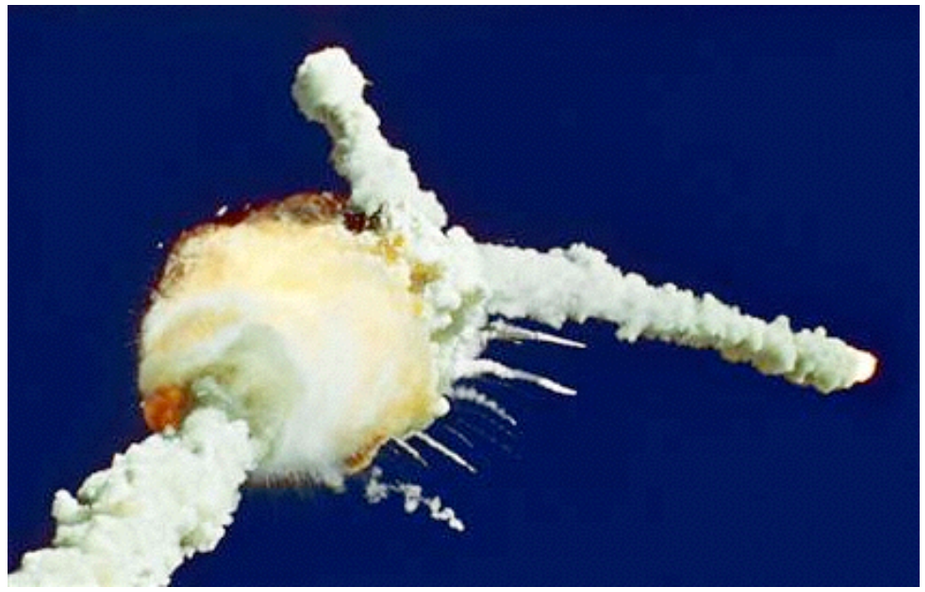


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.



Siddhartha R. Dalal; Edward B. Fowlkes; Bruce Hoadley. Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. JASA, Vol. 84, No. 408 (Dec., 1989), pp. 945-957. Access via JSTOR.

Edward Tufte
http://www.edwardtufte.com

BOOK:
Visual Explanations: Images and Quantities, Evidence and Narrative

Ch. 5 deals with the Challenger disaster
That chapter is available for $7 as a downloadable booklet:
http://www.edwardtufte.com/tufte/books_textb

“A picture is worth a thousand words”

Always, always, always plot the data.

Replace (or complement) ‘typical’ tables of data or statistical results with figures that are more compelling and accessible.

Whenever possible, generate figures that overlay / juxtapose observed data and analytical results, e.g. the ‘fit’.

“A picture is worth a thousand words”

Why?

- find bizarre data and results when it is least embarrassing and painful
- facilitate comparisons and reveal trends

Recommended reference: Gelman A, Pasarica C, Dodhia R. “Let's Practice What We Preach: Turning Tables into Graphs”. The American Statistician, Volume 56, Number 2, 1 May 2002 , pp. 121-130(10). via JSTOR

Statistical Computing and Graphics

Let’s Practice What We Preach: Turning Tables into Graphs

Andrew GELMAN, Cristian PASARICA, and Rahul DODHIA

Statisticians recommend graphical displays but often use tables to present their own research results. Could graphs do better? We study the question by going through the tables in a recent issue of the *Journal of the American Statistical Association*. We show how it is possible to improve the presentations using graphs that actually take up less space than the original tables. We find a particularly effective tool to be multiple repeated line plots,

plays. Our advice follows well-known principles of data display (see, e.g., Tufte 1983; Cleveland 1985) but applied to the presentation or research results as well as raw data.

2. DISPLAYING NUMERICAL RESULTS

Statistical research requires the display of many different kinds of numerical results, including raw numbers, data reductions, inferences, and—for research in theory and

source is real

PERFECT MATCH: TRIFLE WITH MOSCATO

AT A GLANCE

SERVES 20 PEOPLE

1 HR PREPARATION
50 MIN COOKING (PLUS COOLING, SETTING)

You'll need

1.5 kg	blackberries or mulberries, plus extra to serve (see note)
300 gm	caster sugar
2	vanilla beans, split and seeds scraped
10	gelatine leaves (titanium strength), softened in cold water for 5 minutes
800 ml	pink moscato
1	lemon, juice only
330 ml	crème de mûre (see note)
1.25 kg	crème fraîche
150 ml	milk, or enough to thin
2	lemons, finely grated rind only
40 gm	(¼ cup) pure icing sugar, sifted
	Sponge
8	eggs, at room temperature
250 gm	raw caster sugar
250 gm	plain flour, sieved
50 gm	butter, melted and cooled

Method

1. For sponge, preheat oven to 175C. Whisk eggs and sugar in an electric mixer until tripled in volume (7 minutes). Fold through flour in batches, fold in butter, pour into a 28cm-square cake tin lined with baking paper. Bake until golden and centre springs back when pressed (20-25 minutes). Cool in tin, turn out, halve sponge horizontally, trim each half to fit a 6 litre-capacity glass bowl, then remove from bowl and set aside, reserving trimmings.

2. Meanwhile, combine 1kg berries, sugar, 1 vanilla bean and seeds and 1.1 litres water in a large saucepan, simmer over low heat until infused (50 minutes). Strain through a fine sieve (discard solids), transfer 1 litre hot liquid to a bowl (reserve remainder). Squeeze excess water from gelatine, add to bowl, stir to dissolve. Add moscato, lemon juice and 80ml crème de mûre. Strain half into trifle bowl, scatter over 250gm berries and refrigerate until set (2-2½ hours). Chill remaining berry jelly, removing from refrigerator if it starts to set.

3. Reduce 250ml remaining liquid (discard excess) over high heat to 50ml or until syrupy (10-15 minutes), refrigerate until required.

4. Meanwhile, combine crème fraiche, milk, rind, icing sugar and remaining vanilla seeds in a bowl, adding extra milk if necessary until spreadable. Spread one-third over set jelly, top with a sponge round, fill any gaps with trimmings, drizzle with 125ml crème de mûre. Scatter over remaining berries, pour over remaining jelly (mixture should be starting to set). Refrigerate until set (2-2½ hours). Top with half the remaining crème fraiche mixture, then remaining sponge. Drizzle with remaining crème de mûre, top with remaining crème fraiche mixture. Cover, refrigerate overnight. Serve scattered with extra berries and drizzled with blackberry syrup.

Data Science Tools – Get to know your Neighbour!

Turn to a neighbour you don't know. Ask:

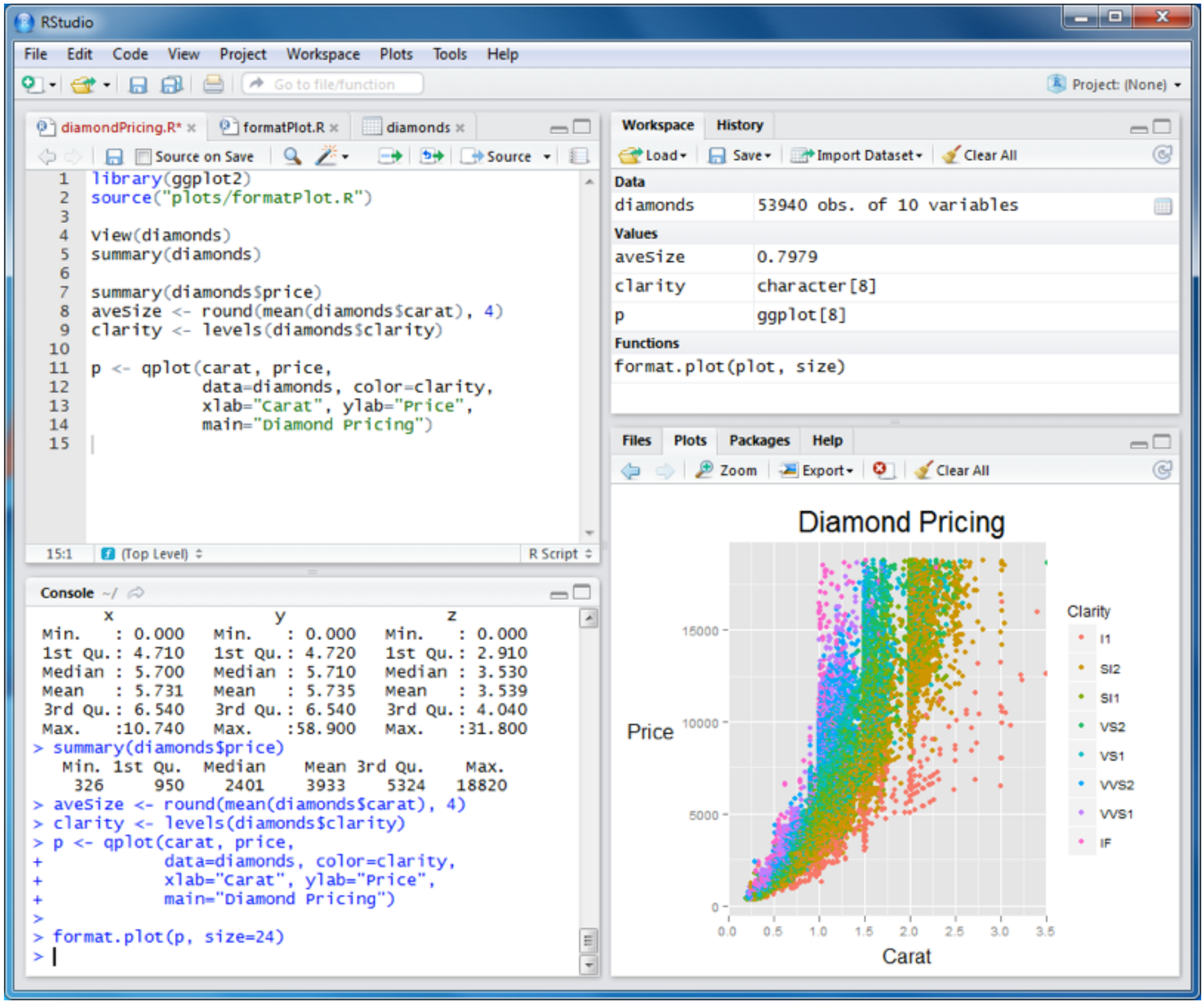
1. Their name
2. Their program
3. What data science tools they use (Excel? R? …)

Volunteer to share what you find out!

RStudio®

DATA SCIENCE TOOL:
R AND RSTUDIO

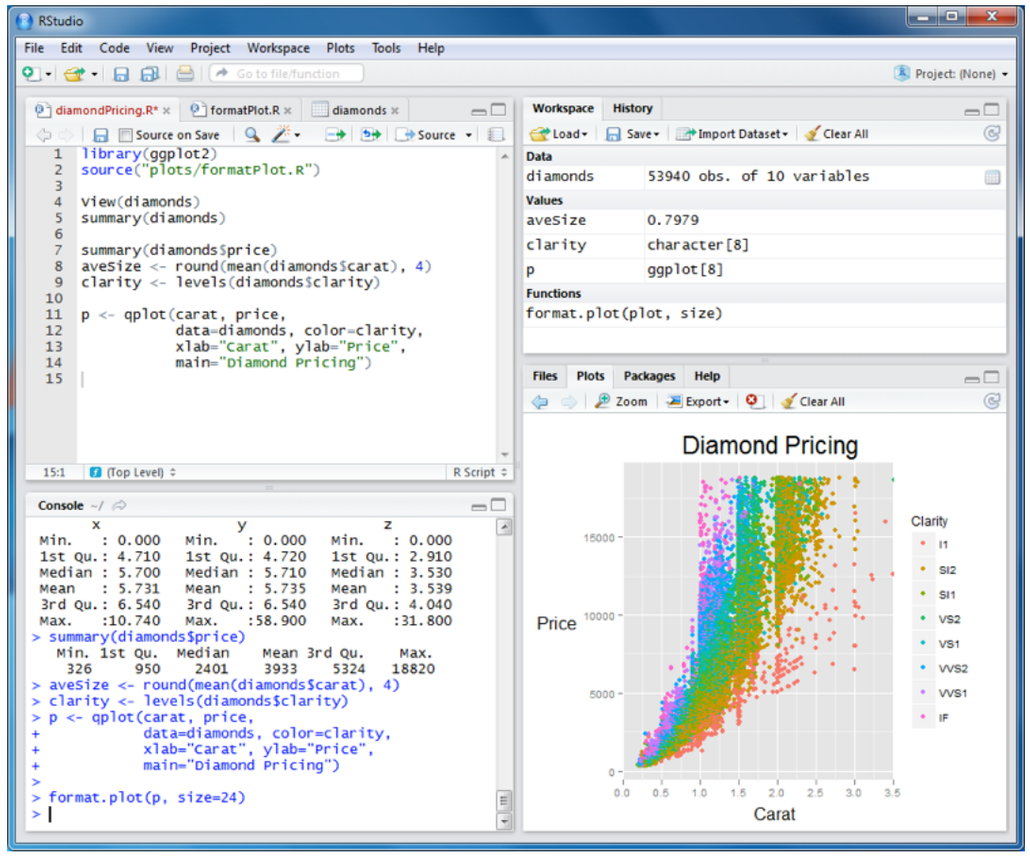
RStudio is an integrated development environment (IDE) for R

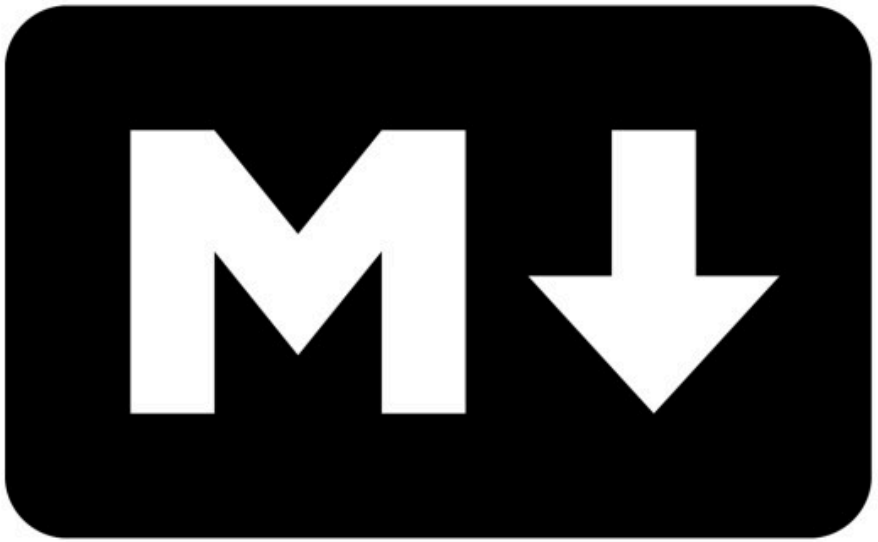


R ≠ RStudio

RStudio mediates your interaction with R; it would replace Emacs + ESS or Tinn-R, but not R itself

Rstudio is a product of -- actually, more a driver of -- the emergence of R Markdown, `knitr`, R + Git(Hub)





DATA SCIENCE TOOL:
(R) MARKDOWN AND HTML (OR PDF)

Markdown ⟶ HTML

foo.md ⟶ foo.html

easy to write
(and read!)

easy to publish
easy to read in
browser

Markdown → HTML

```
Title (header 1, actually)
==========================

This is a Markdown document.

## Medium header (header 2, actually)

It's easy to do *italics* or __make things bold__.

> All models are wrong, but some are useful. An
approximate answer to the right problem is worth a
good deal more than an exact answer to an
approximate problem. Absolute certainty is a
privilege of uneducated minds-and fanatics. It is,
for scientific folk, an unattainable ideal. What
you do every day matters more than what you do
once in a while. We cannot expect anyone to know
anything we didn't teach them ourselves.
Enthusiasm is a form of social courage.


Code block below. Just affects formatting here but
we'll get to R Markdown for the real fun soon!

```
x <- 3 * 4
```

I can haz equations. Inline equations, such as ...
the average is computed as $\frac{1}{n} \sum_{i=1}
^{n} x_{i}$. Or display equations like this:

$$
\begin{equation*}
|x|=
\begin{cases} x & \text{if $x≥0$,} \\\\
-x &\text{if $x\le 0$.}
\end{cases}
\end{equation*}
$$
```

Title (header 1, actually)

This is a Markdown document.

Medium header (header 2, actually)

It's easy to do *italics* or **make things bold**.

> All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. What you do every day matters more than what you do once in a while. We cannot expect anyone to know anything we didn't teach them ourselves. Enthusiasm is a form of social courage.

Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!

```
x <- 3 * 4
```

I can haz equations. Inline equations, such as … the average is computed as $\frac{1}{n}\sum_{i=1}^{n} x_i$. Or display equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

R markdown

R Markdown → Markdown

````
R Markdown rocks
================================

This is an R Markdown document.

```{r}
x <- rnorm(1000)
head(x)
```

See how the R code gets executed and a
representation thereof appears in the document?
`knitr` gives you control over how to represent all
conceivable types of output. In case you care, then
average of the `r length(x)` random normal variates
we just generated is `r round(mean(x), 3)`. Those
numbers are NOT hard-wired but are computed on-the-
fly. As is this figure. No more copy-paste ... copy-
paste ... oops forgot to copy-paste.

```{r}
plot(density(x))
```

Note that all the previously demonstrated math
typesetting still works. You don't have to choose
between having math cred and being web-friendly!

Inline equations, such as ... the average is
computed as $\frac{1}{n} \sum_{i=1}^{n} x_{i}$. Or
display equations like this:

$$
\begin{equation*}
|x|=
\begin{cases} x & \text{if $x≥0$,} \\\\
-x &\text{if $x\le 0$.}
\end{cases}
\end{equation*}
$$
````

````
R Markdown rocks
================================

This is an R Markdown document.


```
r
x <- rnorm(1000)
head(x)
```


```
## [1] -1.3007  0.7715  0.5585 -1.2854  1.1973
2.4157
```


See how the R code gets executed and a
representation thereof appears in the document?
`knitr` gives you control over how to represent all
conceivable types of output. In case you care, then
average of the 1000 random normal variates we just
generated is -0.081. Those numbers are NOT hard-
wired but are computed on-the-fly. As is this
figure. No more copy-paste ... copy-paste ... oops
forgot to copy-paste.


```
plot(density(x))
```


![plot of chunk unnamed-chunk-2](figure/unnamed-
chunk-2.png)

...
````

Markdown ⟶ HTML

````
R Markdown rocks
================================

This is an R Markdown document.


```r
x <- rnorm(1000)
head(x)
```


```
## [1] -1.3007  0.7715  0.5585 -1.2854  1.1973
2.4157
```



See how the R code gets executed and a
representation thereof appears in the document?
`knitr` gives you control over how to represent all
conceivable types of output. In case you care, then
average of the 1000 random normal variates we just
generated is -0.081. Those numbers are NOT hard-
wired but are computed on-the-fly. As is this
figure. No more copy-paste ... copy-paste ... oops
forgot to copy-paste.


```r
plot(density(x))
```

![plot of chunk unnamed-chunk-2](figure/unnamed-
chunk-2.png)

...
````

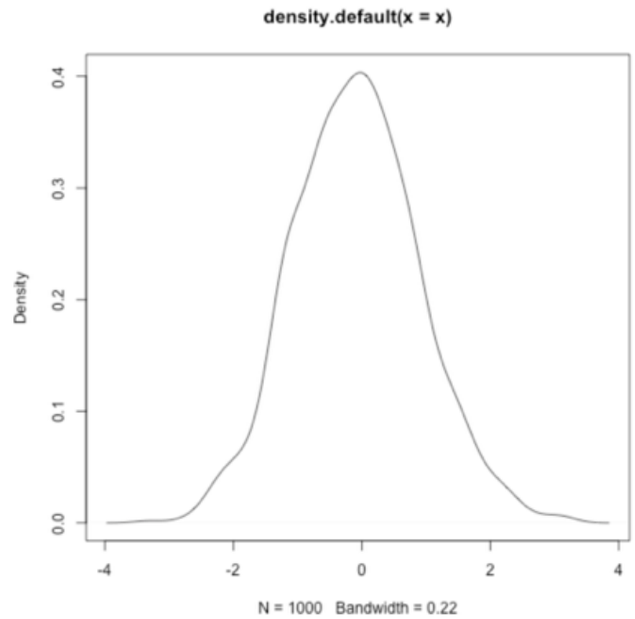
R Markdown rocks

This is an R Markdown document.

```r
x <- rnorm(1000)
head(x)
```

```
## [1] -1.3007  0.7715  0.5585 -1.2854  1.1973  2.4157
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the 1000 random normal variates we just generated is -0.081. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste … copy-paste … oops forgot to copy-paste.



Note that all the previously demonstrated math typesetting still works. You don't have to choose between having math cred and being web-friendly!

Inline equations, such as … the average is computed as $\frac{1}{n}\sum_{i=1}^{n} x_i$. Or display equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

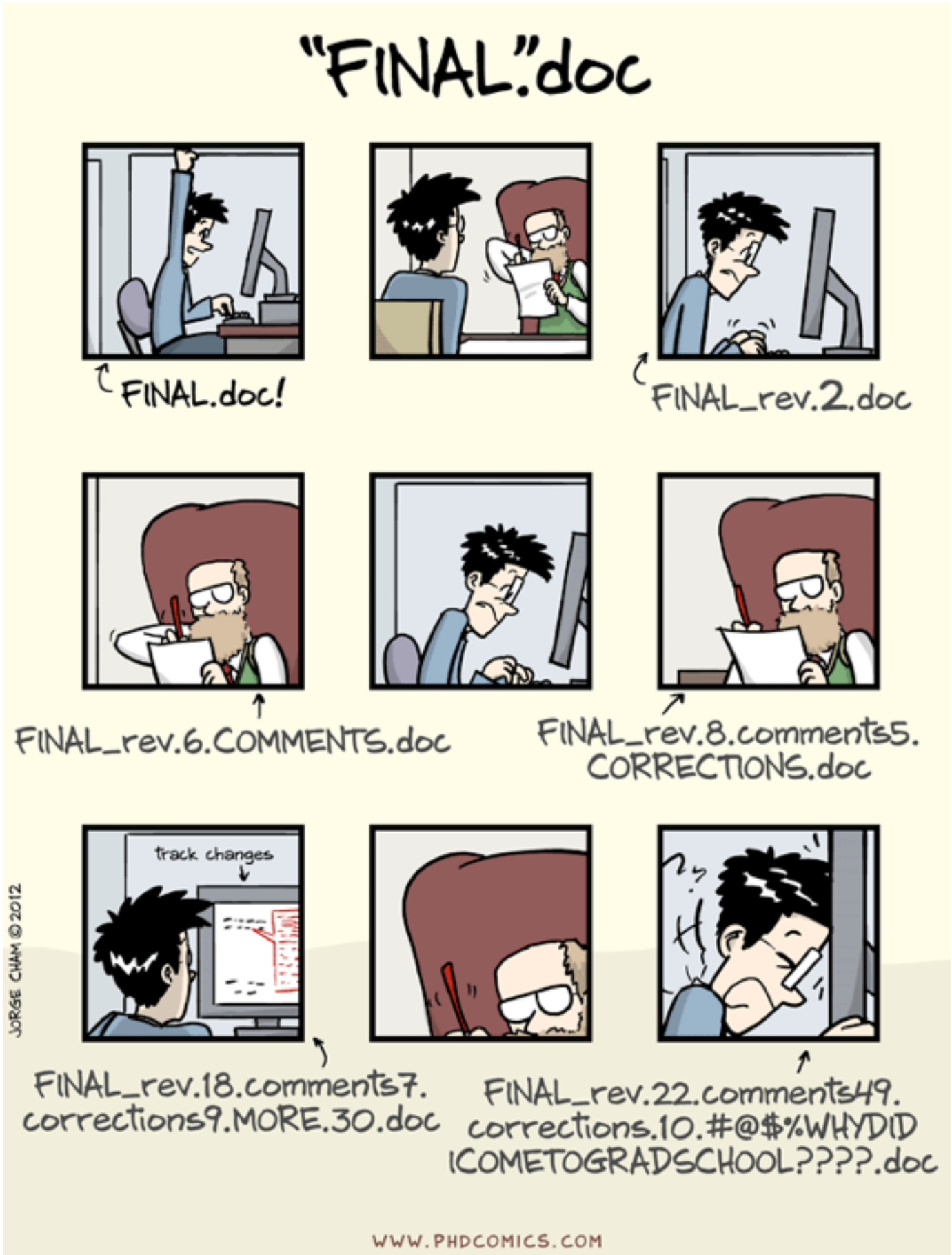
R Markdown → Markdown → HTML

foo.rmd → foo.md → foo.html

easy to write
(and read!)

easy to publish
easy to read in
browser

DATA SCIENCE TOOL:
GIT AND GITHUB



http://www.phdcomics.com/comics/archive.php?comicid=1531 via Ram, 2013 doi:10.1186/1751-0473-8-7

Version control systems (VCS) were created to help groups of people develop software

Git, in particular, is being "repurposed" for activities other than pure software development ... like the messy hybrid of writing, coding and data wrangling

Git repository = a bunch of files you want to manage in a sane way

repo = repository

Git repository = a bunch of files you want to manage in a sane way

repo = repository

you can set up repo ... then start your work

or you can make a set of existing files and make them into a repo

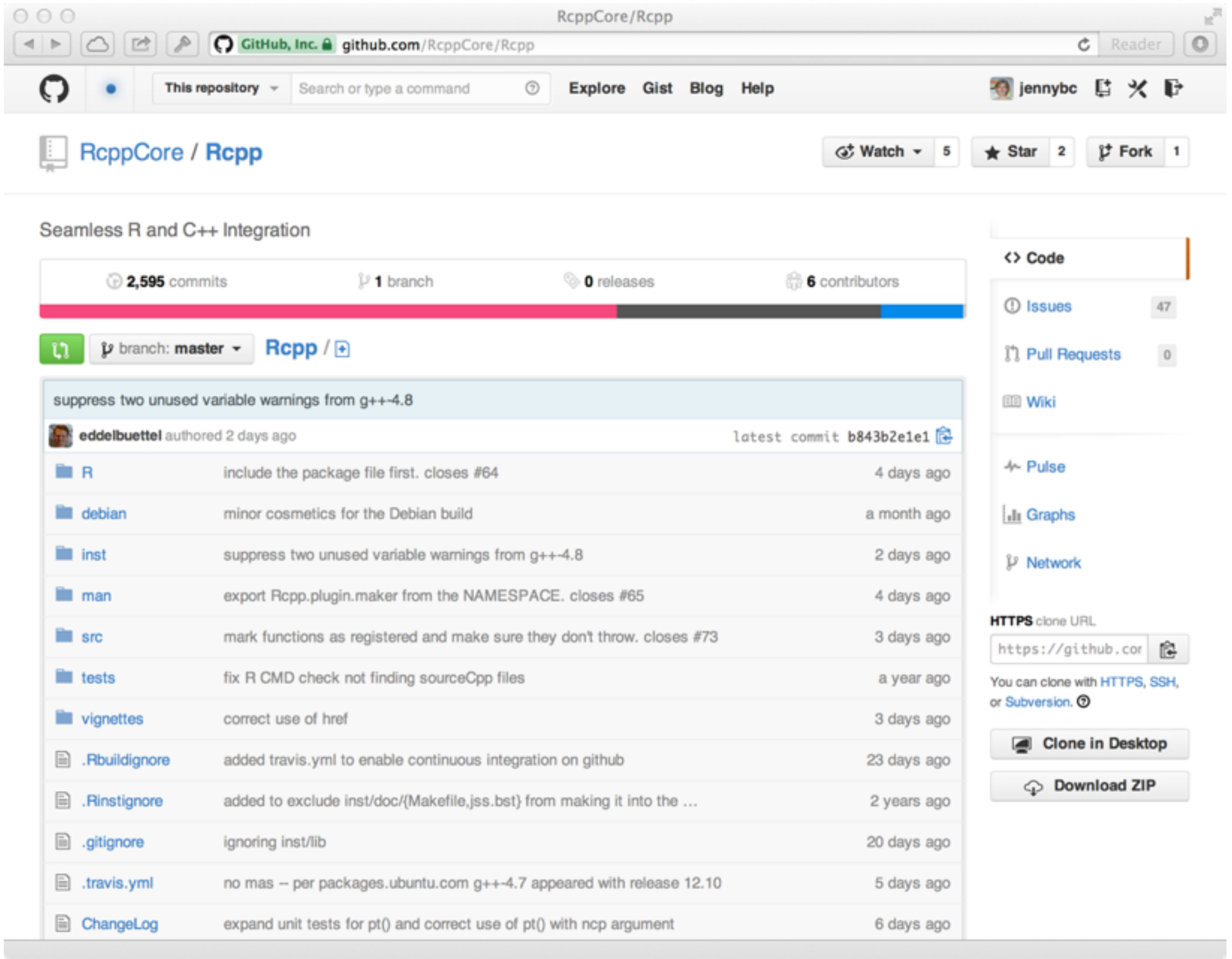
collaboration = the “killer app” of version control

Learning Git has been -- and continues to be -- painful. But not nearly as crazy-making as the alternatives:

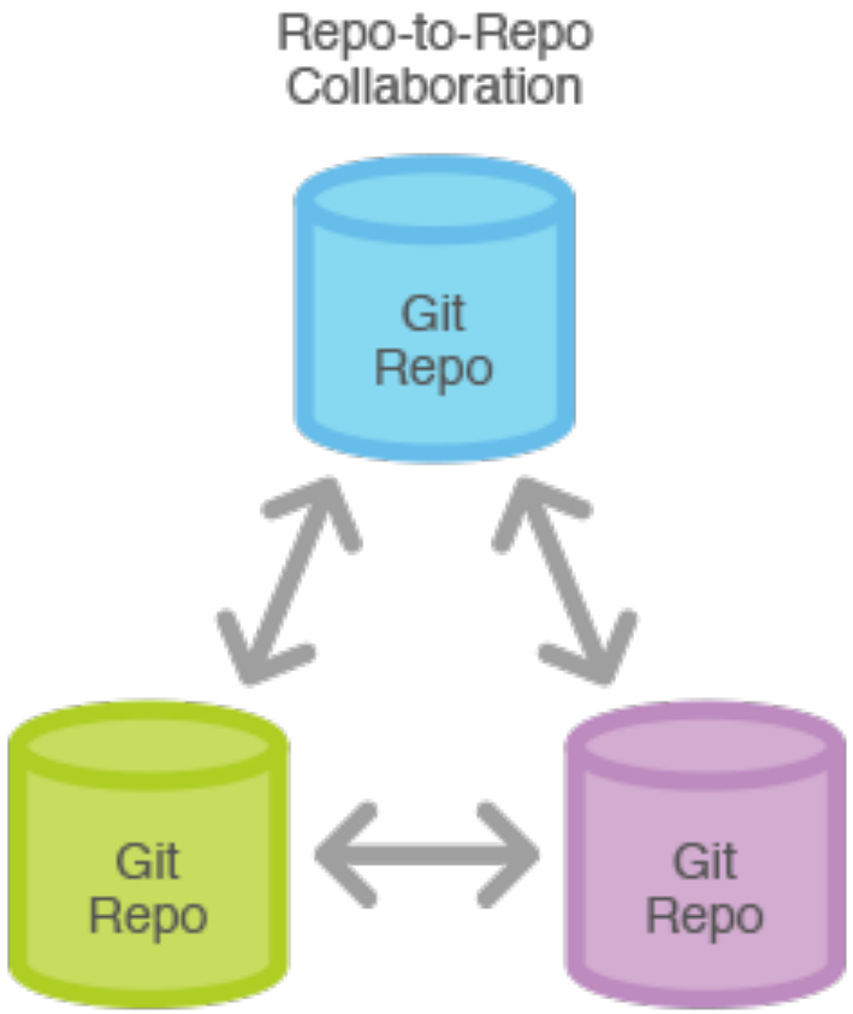
- documents as email attachments
- uncertainty about which version is “master”
- am I working with the most recent data?
- archaelogical “digs” on old email threads
- uncertainty about how/if certain changes have been made or issues solved
- hair-raising ZIP archives containing file salad

GitHub = a place to host Git repositories on the web

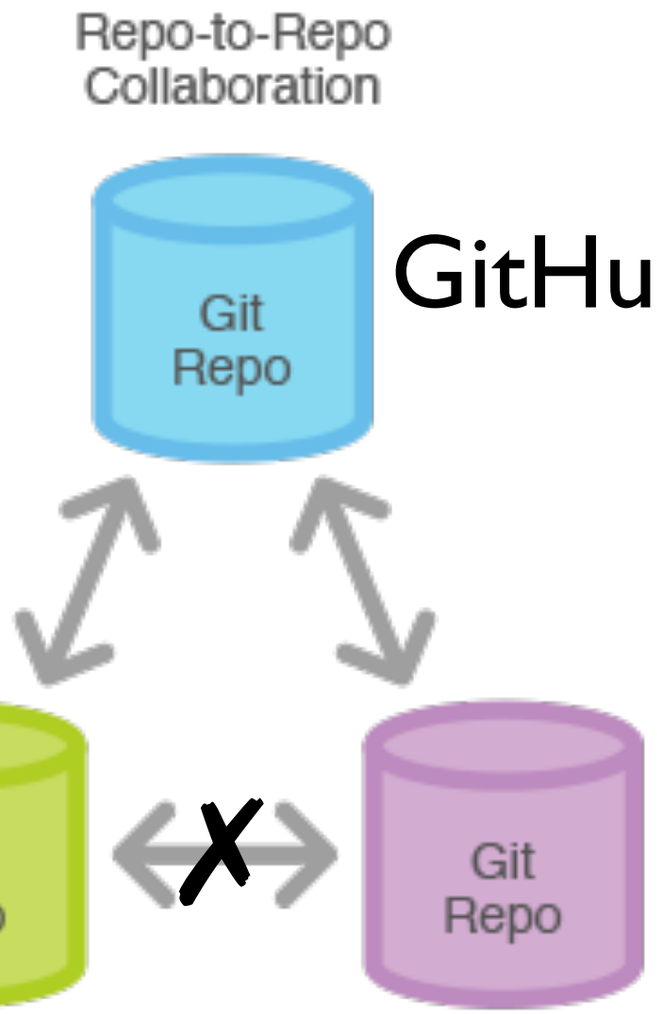
GitHub ≠ Git



possible, in theory



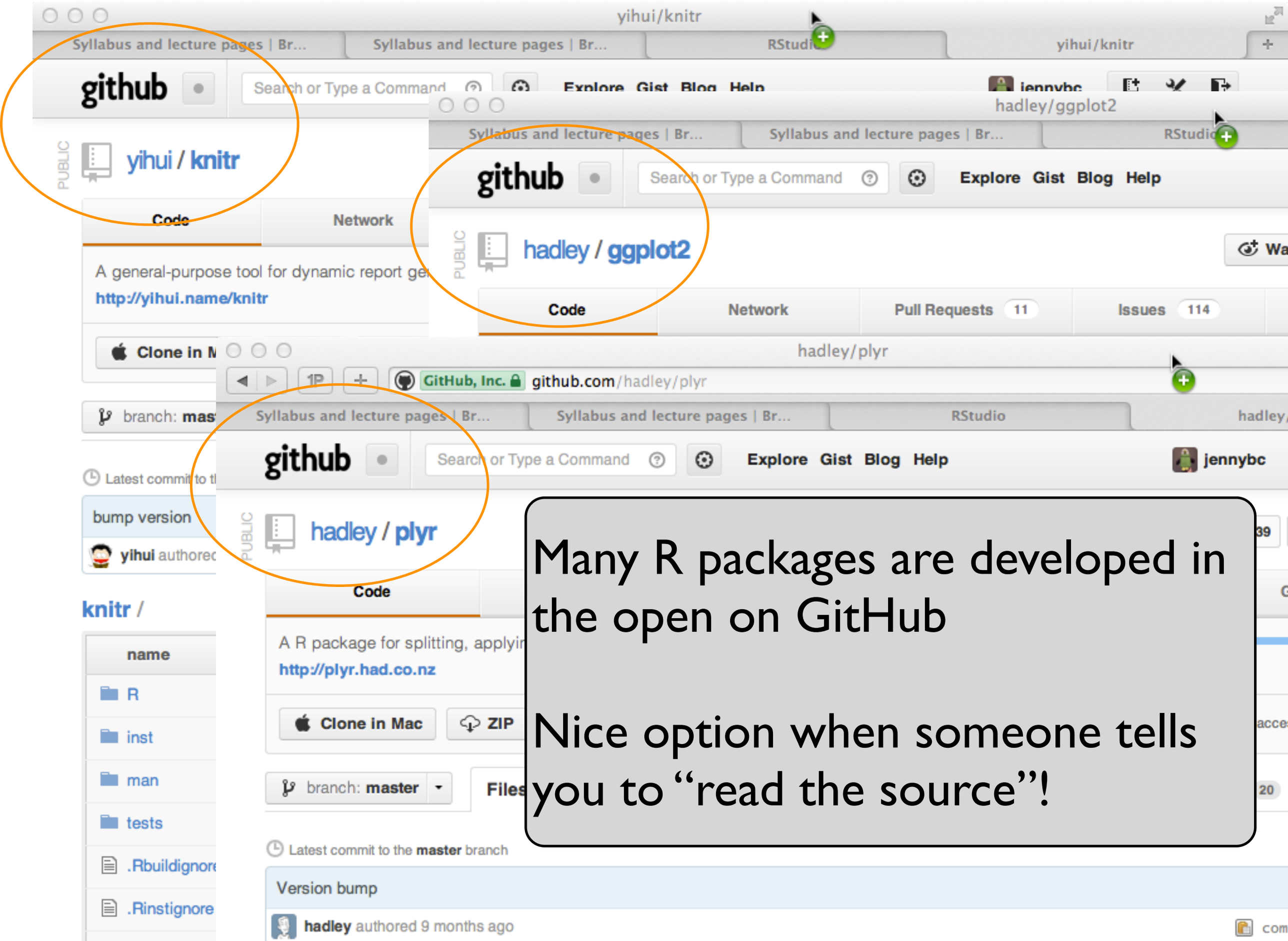
more typical



GitHub

me you

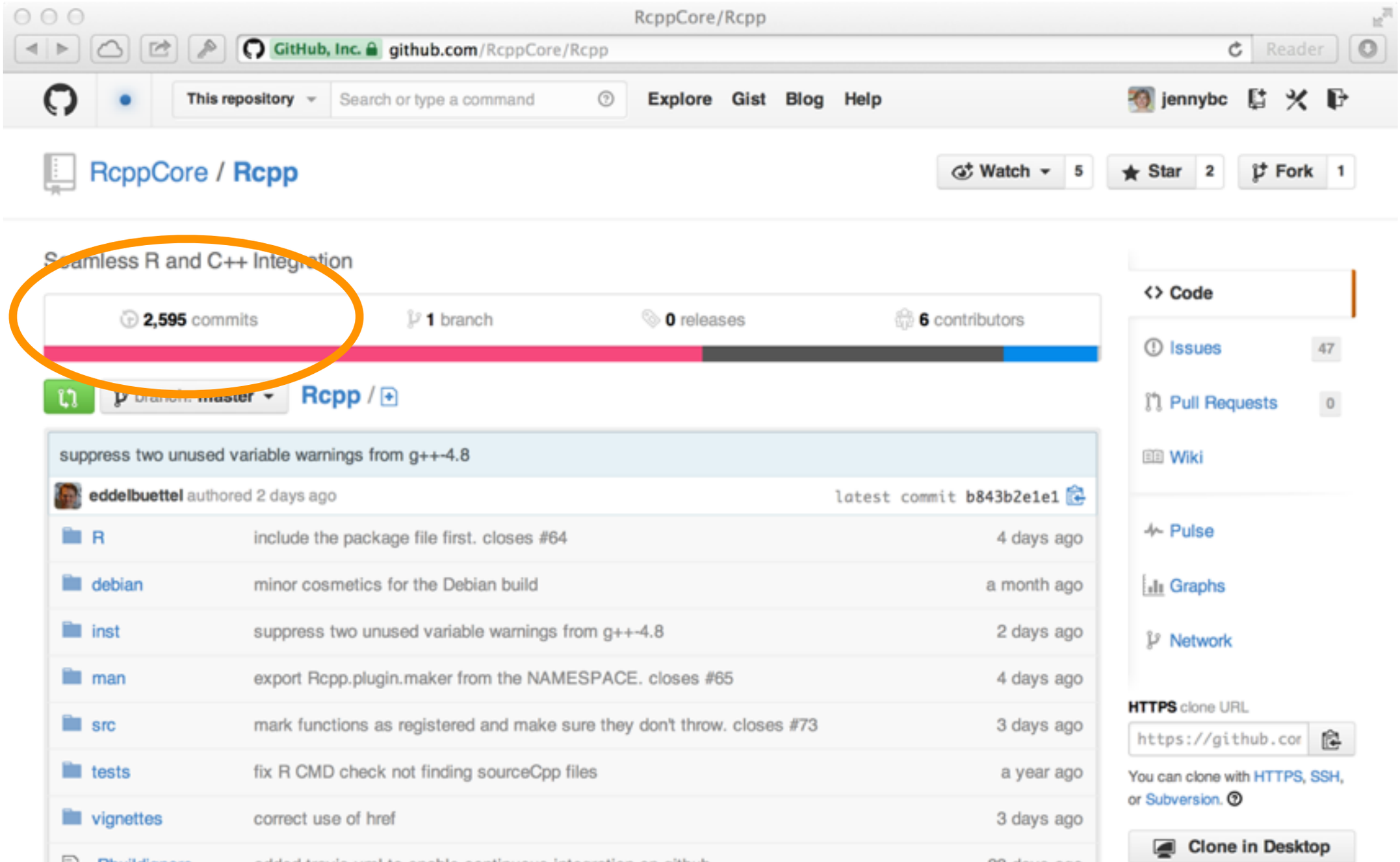
Image from https://www.atlassian.com/git/tutorial/git-basics#!clone



Many R packages are developed in the open on GitHub

Nice option when someone tells you to "read the source"!

You can see exactly how files have changed, when, and by whom. If commit message is good, you'll see why.
Commit = a formal "checkpoint" or snapshot of the state of the repository



GitHub provides a fantastic visual "diff" view of *exactly* what changed. Incredibly useful.

mark functions as registered and make sure they don't throw. closes #73 Browse code

Loading branch information...

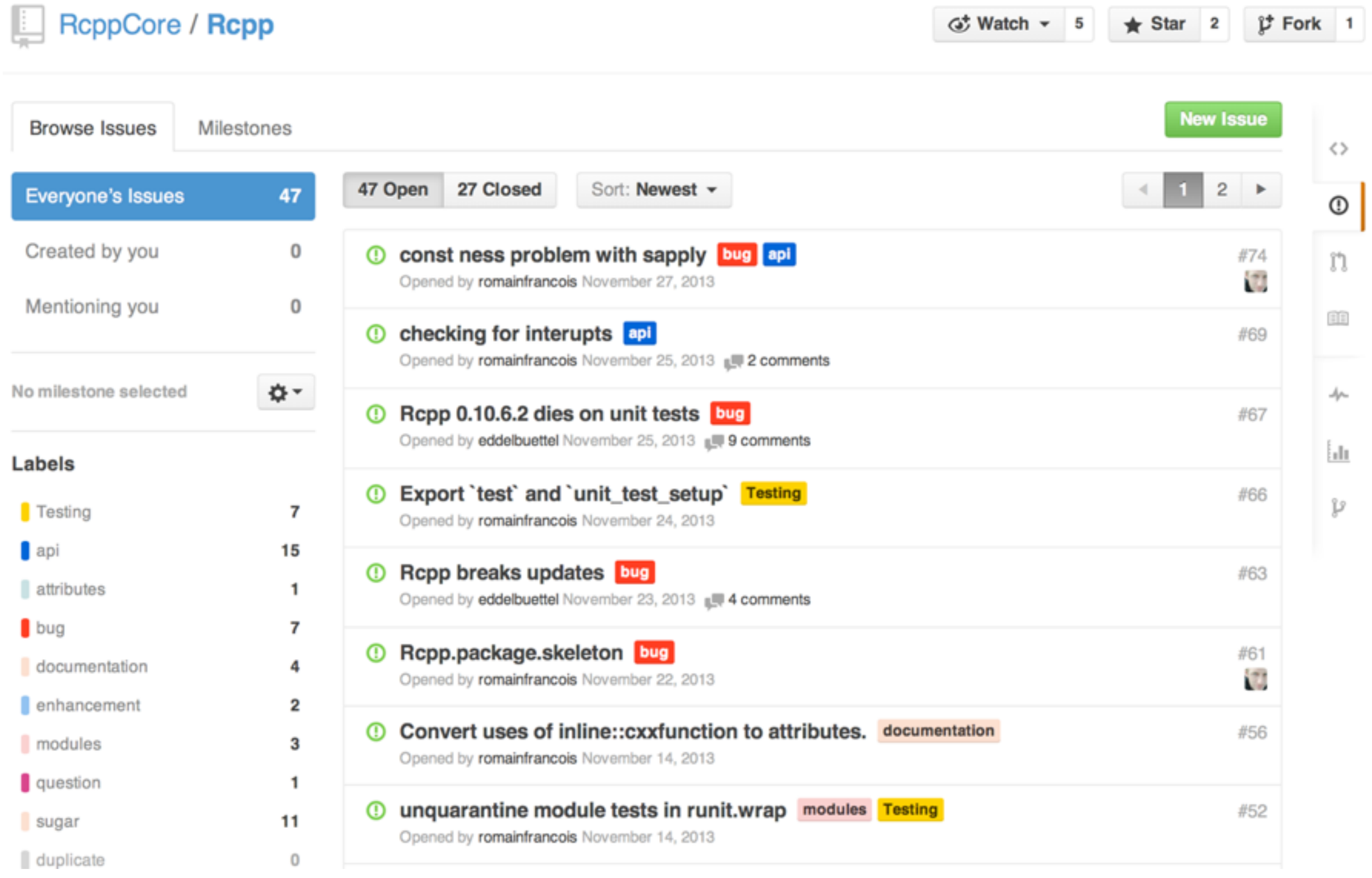
romainfrancois authored November 26, 2013 1 parent d1bc779 commit 3331e7b48934936aa8c4782ba6ec06714da7f670

Showing **2 changed files** with **29 additions** and **11 deletions**. Show Diff Stats

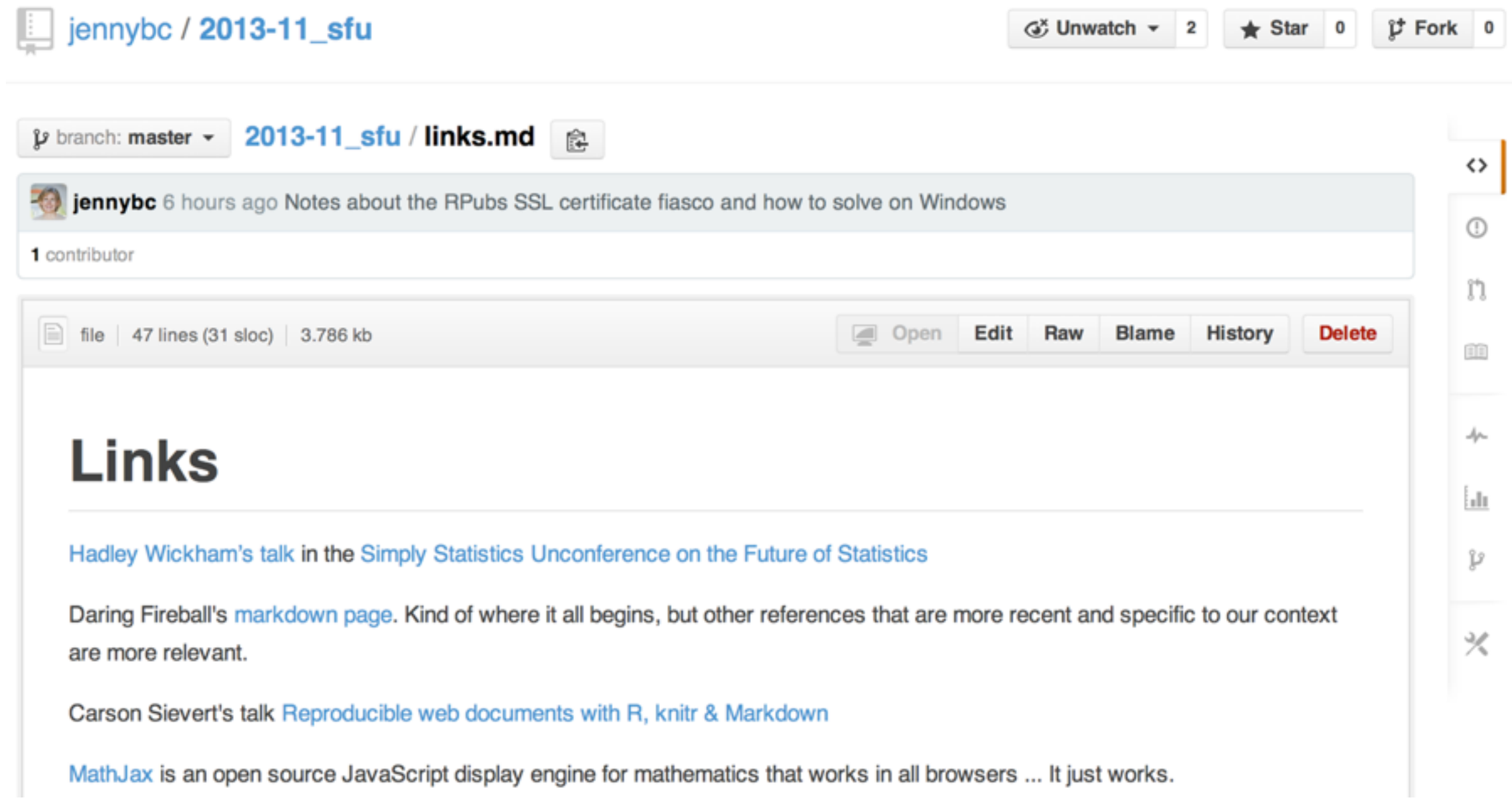
19 src/api.cpp View file @ 3331e7b

```
...  ...  @@ -227,10 +227,11 @@ SEXP stack_trace( const char* file, int line ){
227  227      #if defined(__GNUC__)
228  228          #if defined(WIN32) || defined(__FreeBSD__) || defined(__NetBSD__) || defined(__OpenBSD__) || defined(__CYGWIN_
229  229              // Simpler version for Windows and *BSD
230       -            Rcpp::List trace = Rcpp::List::create(
231       -                Rcpp::Named( "file"  ) = file,
232       -                Rcpp::Named( "line"  ) = line,
233       -                Rcpp::Named( "stack" ) = "C++ stack not available on this system" ) ;
     230  +            List trace = List::create(
     231  +                _[ "file"  ] = file,
     232  +                _[ "line"  ] = line,
     233  +                _[ "stack" ] = "C++ stack not available on this system"
     234  +            ) ;
```

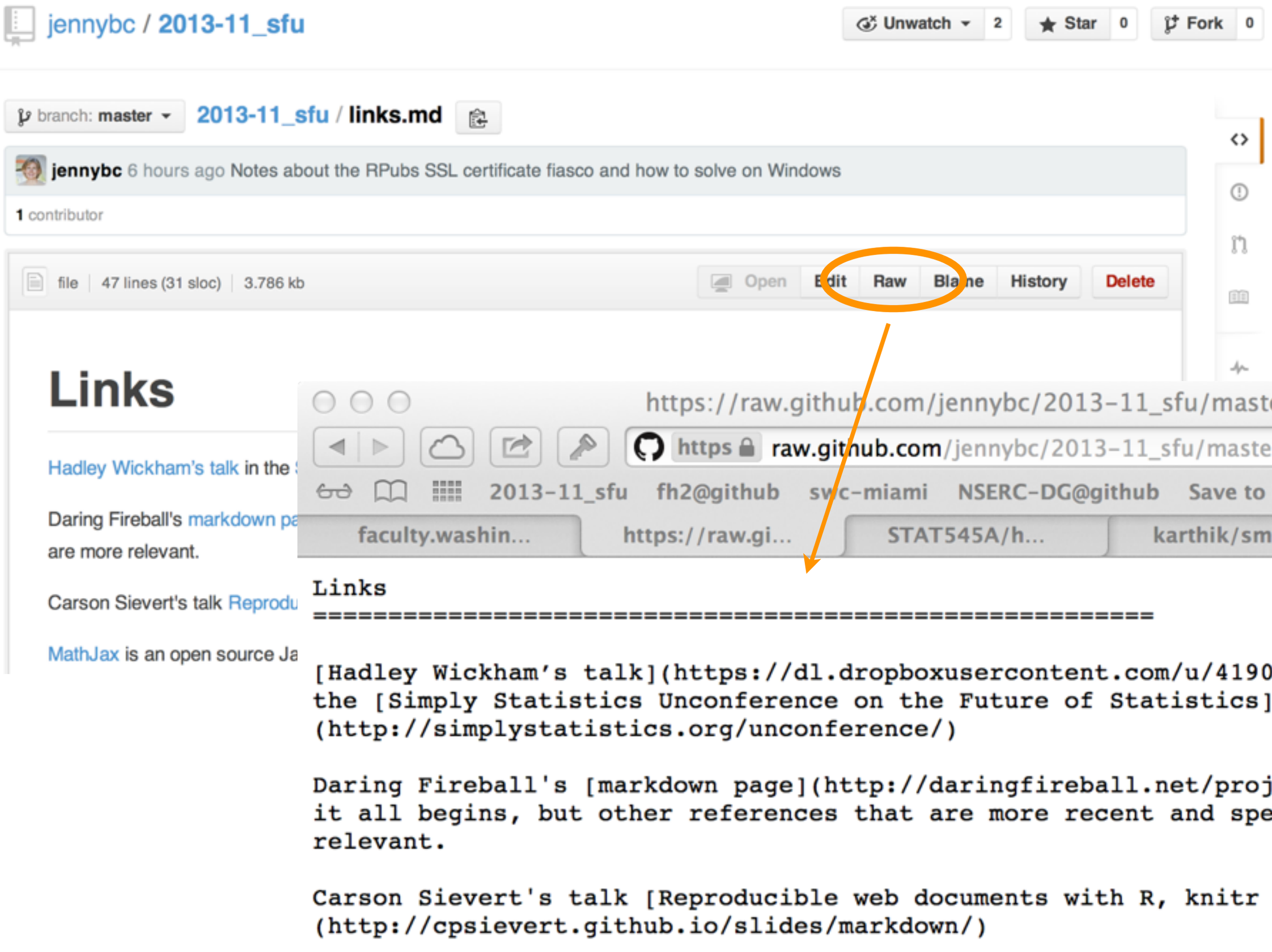
GitHub *issues*: think "bug tracker", "to do list".



GitHub renders Markdown files nicely
Example: `links.md` in workshop repo of mine



You can see the raw Markdown too!



GitHub renders comma (.csv) and tab (.tsv) delimited files nicely
Example: Lord of the Rings data I found for STAT 545A

jennybc / lotr

Unwatch 1 | Star 0 | Fork 1

branch: master | lotr / lotr_clean.tsv

jennybc 2 months ago Add early exploration/cleaning

1 contributor

file | 684 lines (683 sloc) | 42.64 kb | Open | Edit | Raw | Blame | History | Delete

Search this file...

	Film	Chapter	Character	Race	Words
2	The Fellowship Of The Ring	01: Prologue	Bilbo	Hobbit	4
3	The Fellowship Of The Ring	01: Prologue	Elrond	Elf	5
4	The Fellowship Of The Ring	01: Prologue	Galadriel	Elf	460
5	The Fellowship Of The Ring	02: Concerning Hobbits	Bilbo	Hobbit	214
6	The Fellowship Of The Ring	03: The Shire	Bilbo	Hobbit	70
7	The Fellowship Of The Ring	03: The Shire	Frodo	Hobbit	128
8	The Fellowship Of The Ring	03: The Shire	Gandalf	Wizard	197
9	The Fellowship Of The Ring	03: The Shire	Hobbit Kids	Hobbit	10

GNU Operating System

Sponsored by the Free Software Foundation

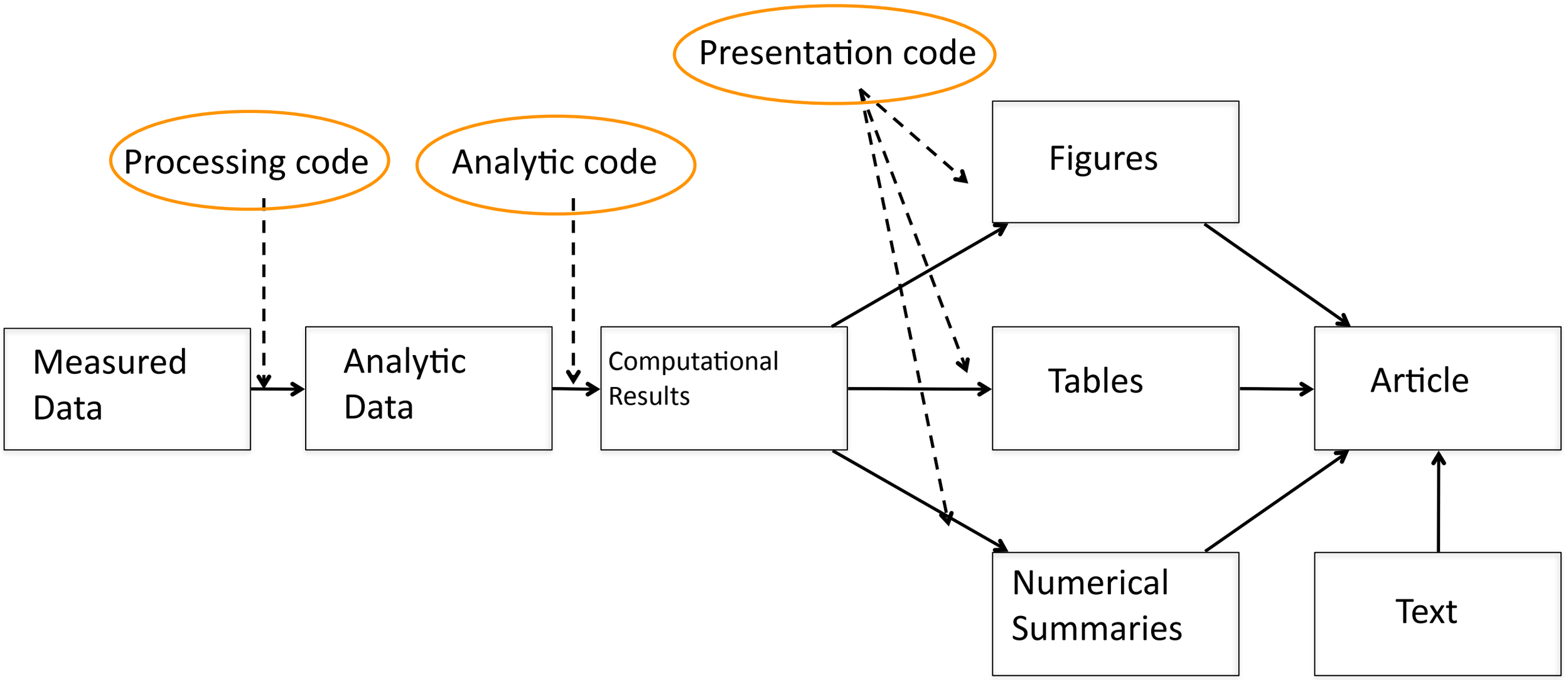
ABOUT GNU PHILOSOPHY

GNU Make

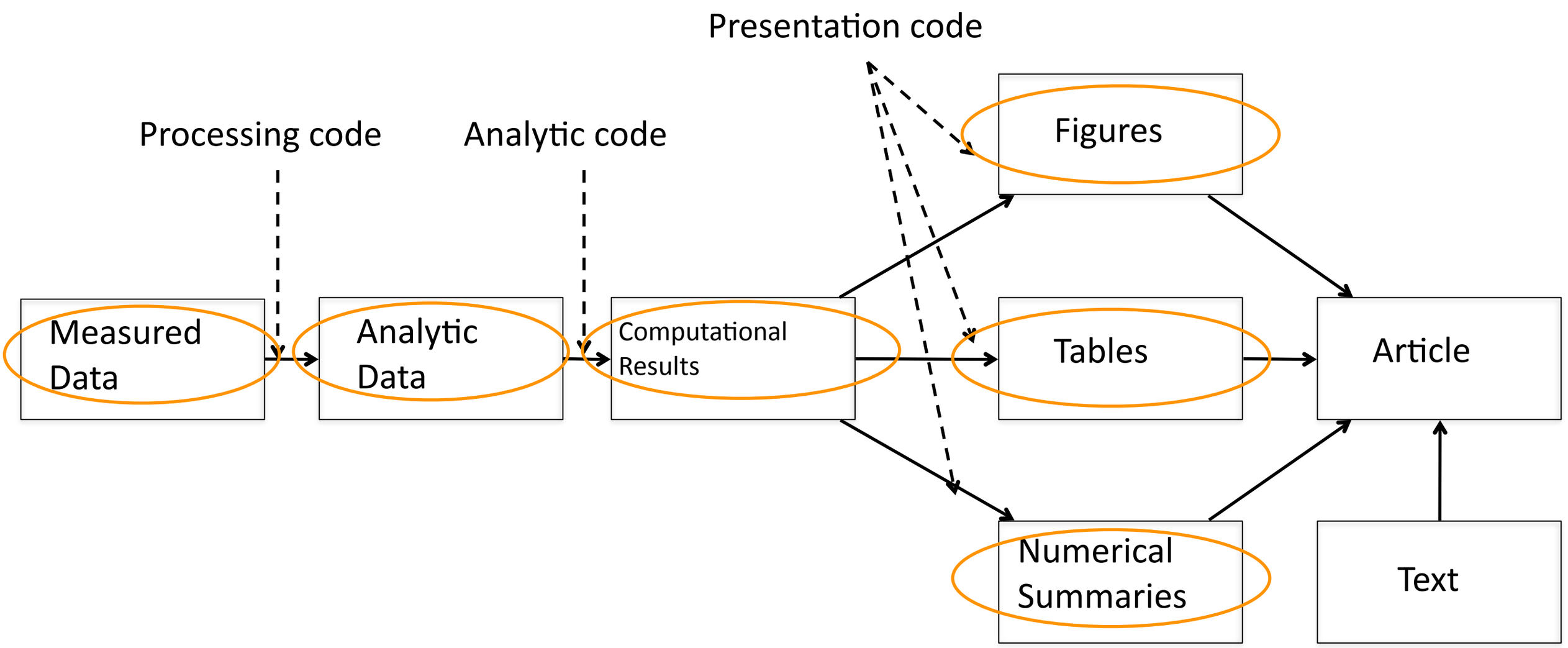
DATA SCIENCE TOOL: MAKE

slide modified from Roger Peng http://www.biostat.jhsph.edu/~rpeng/

R (or Python) scripts



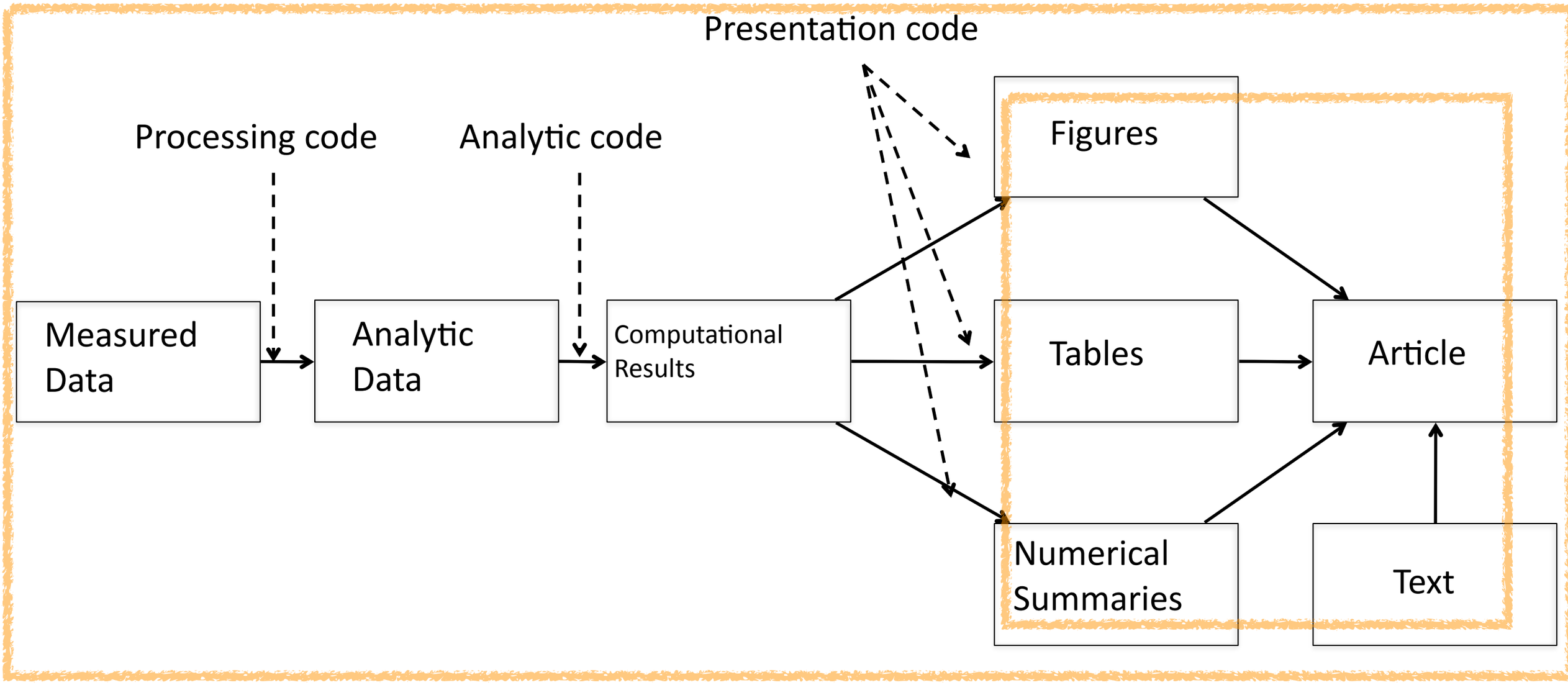
slide modified from Roger Peng http://www.biostat.jhsph.edu/~rpeng/



delimited or other structured, agnostic files

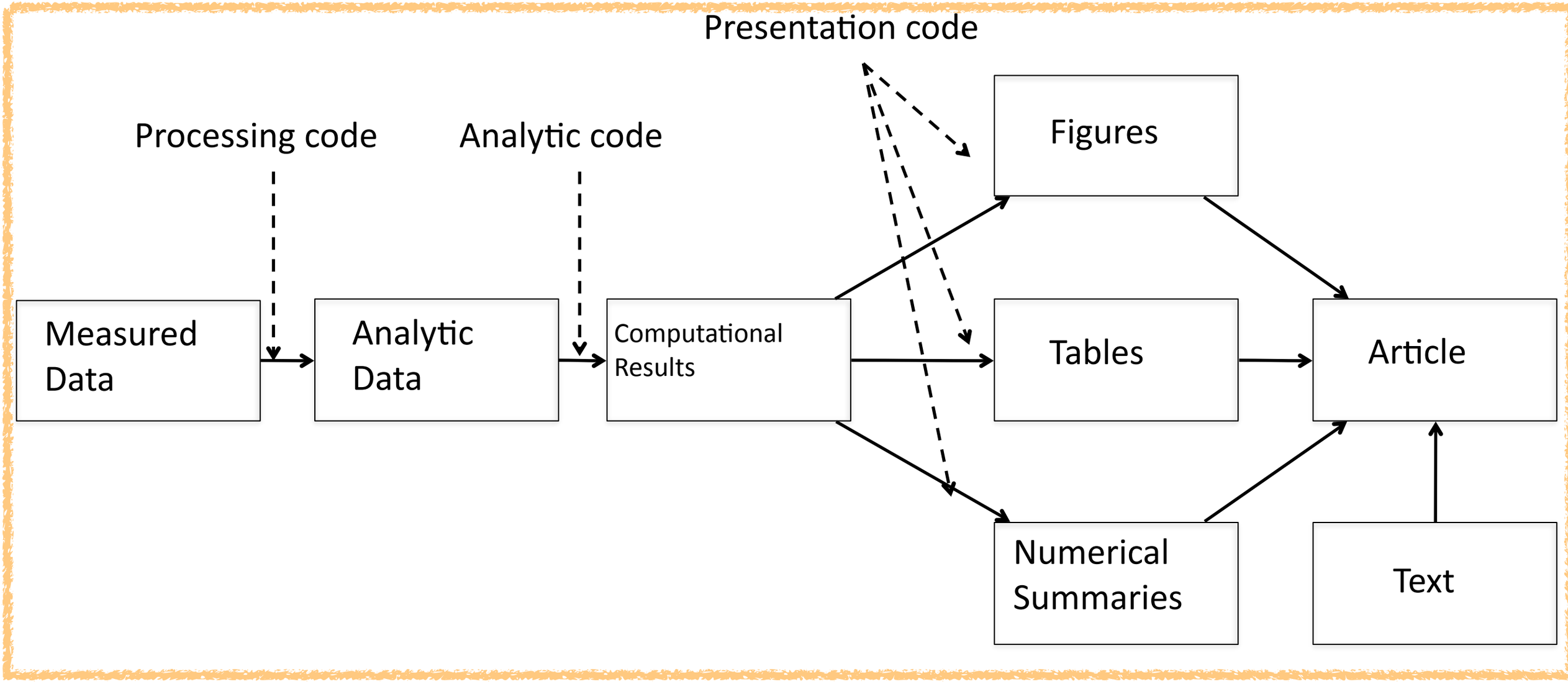
slide modified from Roger Peng http://www.biostat.jhsph.edu/~rpeng/

How to make end products more integrated and more reproducible?



slide modified from Roger Peng http://www.biostat.jhsph.edu/~rpeng/

How to keep everything up-to-date?
If the data changes, how do we remember to re-make the figures 2B and 4?



Like Git,
GNU Make is another old school tool that is being repurposed to meet a need in data-intensive workflows.

Originally intended to orchestrate compiling complicated software, it's now used to express what depends on what and keep everything "in sync".

Makefile

```
all: data model paper
data: raw.csv
model: model.Rout
paper: plot.Rout paper.pdf

raw.csv: get_data.py
    python get_data.py

clean.csv: clean.sh raw.csv
    source clean.sh

model.Rout: model.R clean.csv
    R CMD BATCH model.R

plot.Rout: plot.R model.Rout
    R CMD BATCH plot.R

paper.pdf: paper.tex
    $(TEXCMD) $<
    $(TEXCMD) $<
    bibtex *.aux
    $(TEXCMD) $<
```

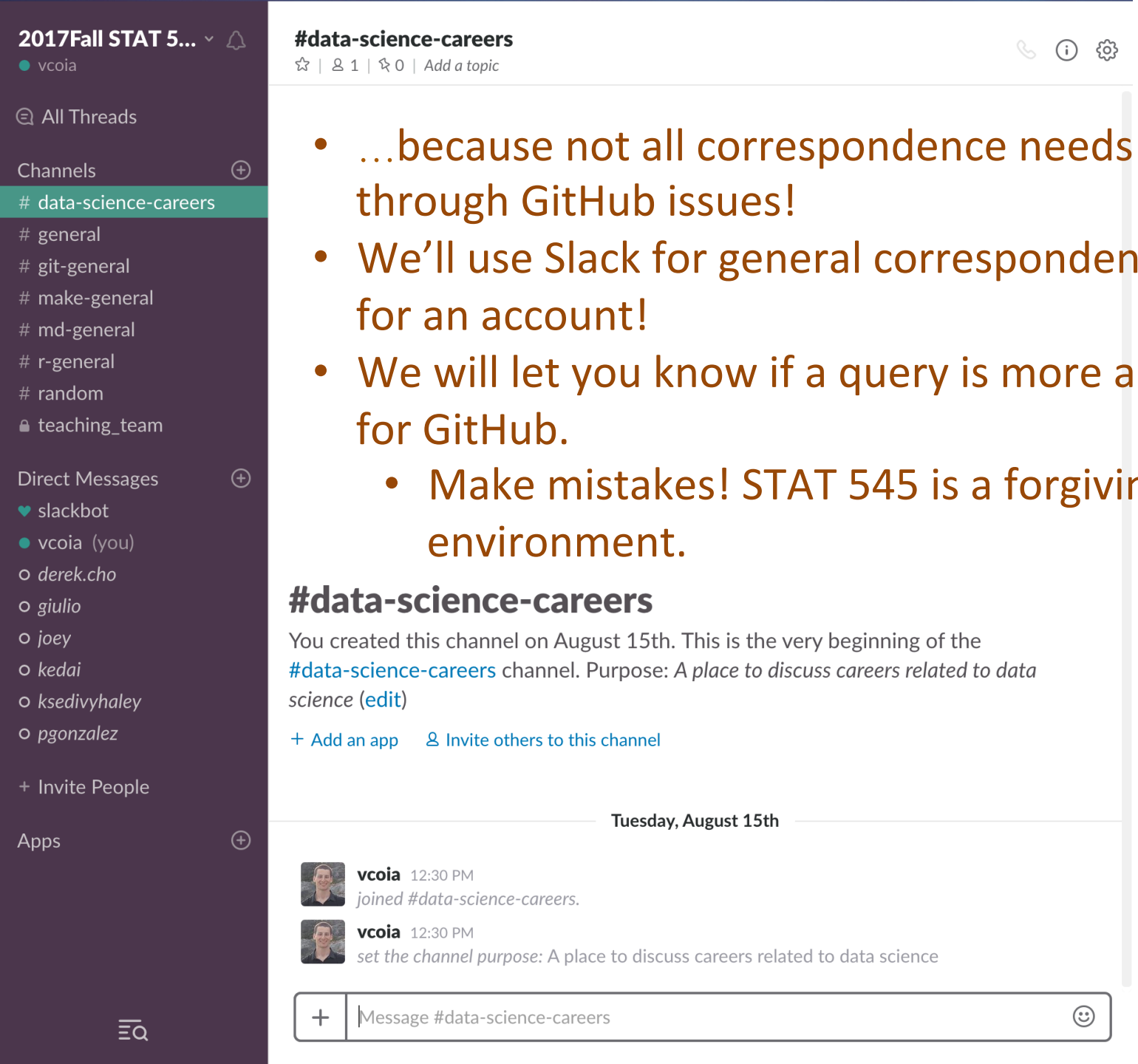
http://zmjones.com/make.html

course stuff

COURSE TOOL: STAT545.COM

COURSE TOOL: SLACK

stat545-2017.slack.com


2017Fall STAT 5...
vcoia
All Threads
Channels
data-science-careers
general
git-general
make-general
md-general
r-general
random
teaching_team
Direct Messages
slackbot
vcoia (you)
derek.cho
giulio
joey
kedai
ksedivyhaley
pgonzalez
Invite People
Apps
#data-science-careers
1 | 0 | Add a topic
…because not all correspondence needs to be through GitHub issues!
We'll use Slack for general correspondence. Sign up for an account!
We will let you know if a query is more appropriate for GitHub.
Make mistakes! STAT 545 is a forgiving environment.
#data-science-careers
You created this channel on August 15th. This is the very beginning of the #data-science-careers channel. Purpose: A place to discuss careers related to data science (edit)
Add an app
Invite others to this channel
Tuesday, August 15th
vcoia 12:30 PM
joined #data-science-careers.
vcoia 12:30 PM
set the channel purpose: A place to discuss careers related to data science
Message #data-science-careers

how to get help

Office hours Tues/Thurs after class (or Wed this week only)

Open an issue on a GitHub repository and tag one or more instructors

Tweet to @STAT545

Not for 2017 -- use Slack.

Direct twitter message to @STAT545

Email to an instructor

http://stat545-ubc.github.io/help-STAT545.html

respond to our prompt to figure out who you are!

we want to match up various info with what UBC provides (e.g. Twitter handle, Github username)

what class meetings will look like … sort of?

9:30 - 9:50	“lecture”
9:50 - 10:35	hands-on work!
10:35 - 10:55	“lecture”

rhythm of each week

submit a unit of work → tues class → thurs class

work on your own
work in class
consult peers and instructors in class
office hrs
online interaction via GitHub, Twitter

the end