POLI 7002/7450: Computational Methods for Comparative Politics

Prof. Frederick Solt

Introduction to the Course

CMCR

CMCR

- Reproducibility

CMCR

- Reproducibility
- Data Science

CMCR

- Reproducibility
- Data Science
  - Visualization

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming
- Computational Methods

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming
- Computational Methods
  - Crowdsourcing Data

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming
- Computational Methods
  - Crowdsourcing Data
  - Applied Bayesian Analysis

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming
- Computational Methods
  - Crowdsourcing Data
  - Applied Bayesian Analysis
    - Multilevel Models

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming
- Computational Methods
  - Crowdsourcing Data
  - Applied Bayesian Analysis
    - Multilevel Models
    - Cross-National Latent Variables

CMCR

- Reproducibility
- Data Science
  - Visualization
  - Data Wrangling
  - Programming
- Computational Methods
  - Crowdsourcing Data
  - Applied Bayesian Analysis
    - Multilevel Models
    - Cross-National Latent Variables
    - Comparative Dynamic Public Opinion

Reproducibility

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors

Transparency requires making visible both the empirical foundation and the logic of inquiry of research. We agree that by January 15, 2016 we will:

- Require authors to ensure that cited data are available at the time of publication through a trusted digital repository. Journals may specify which trusted digital repository shall be used
- Require authors to delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to all relevant analytic materials. If such materials

http://dx.doi.org/10.1017/psrm.2015.44

Reproducibility

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors

Transparency requires making visible both the empirical foundation and the logic of inquiry of research. We agree that by January 15, 2016 we will:

- Require authors to ensure that cited data are available at the time of publication through a trusted digital repository. Journals may specify which trusted digital repository shall be used
- Require authors to delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to all relevant analytic materials. If such materials

AJPS Replication and Verification Policy

The corresponding author of a manuscript that is accepted for publication in the *American Journal of Political Science* must provide replication materials that are sufficient to enable interested researchers to reproduce all of the analytic results that are reported in the text and supporting materials.

http://dx.doi.org/10.1017/psrm.2015.44
https://ajps.org/ajps-replication-policy/

Data Science

Too often, students are either expected to figure out how to program on their own during their econometrics classes, or they are offered short, graduate-student-led workshops to introduce basic skills. Coding is now too central to our discipline to be given this second-tier status in our curriculum. If we are going to expect our students to engage in computational research, it is our obligation to equip them with the tools they need…

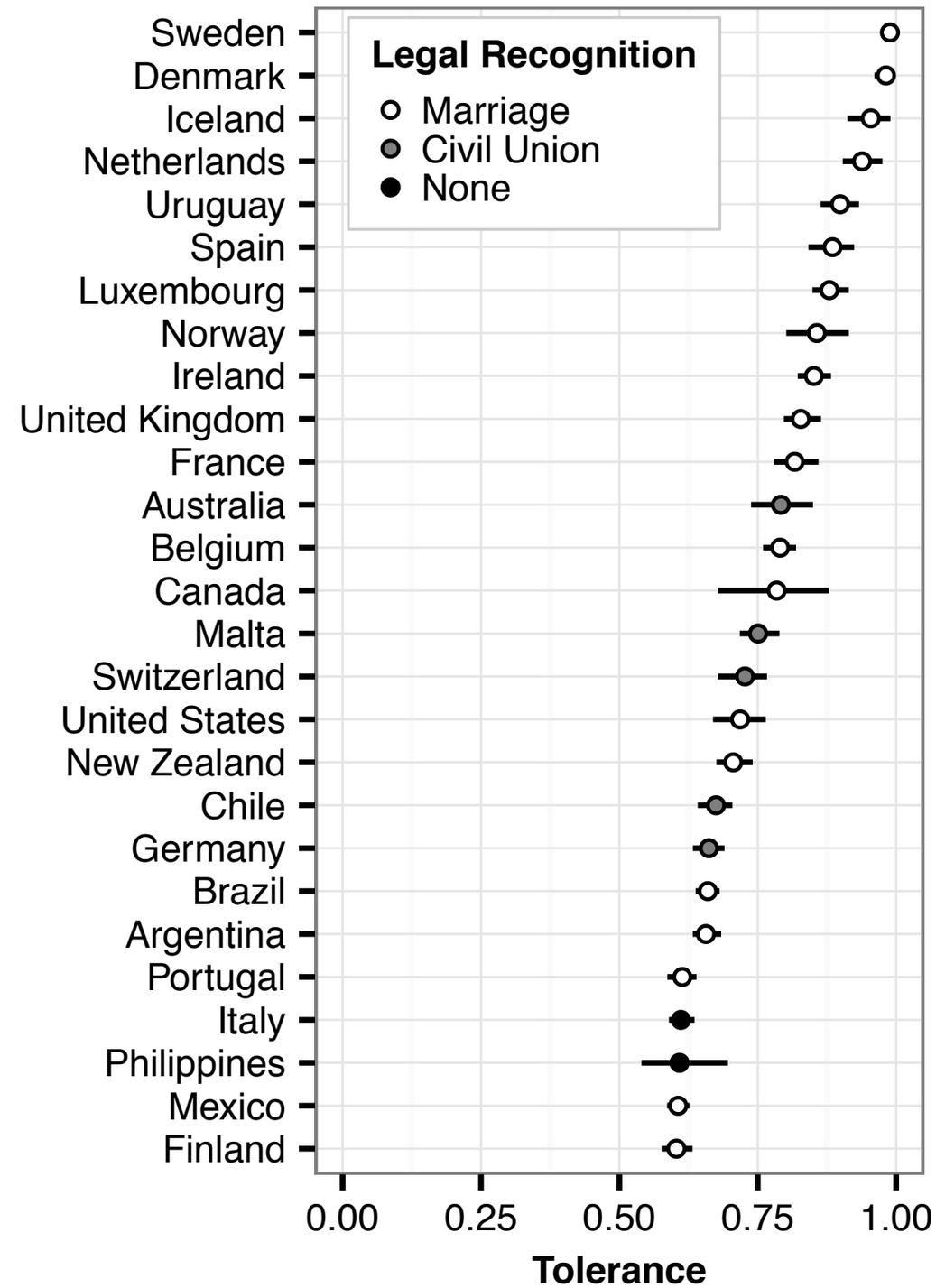
Nicholas Eubank
Stanford University

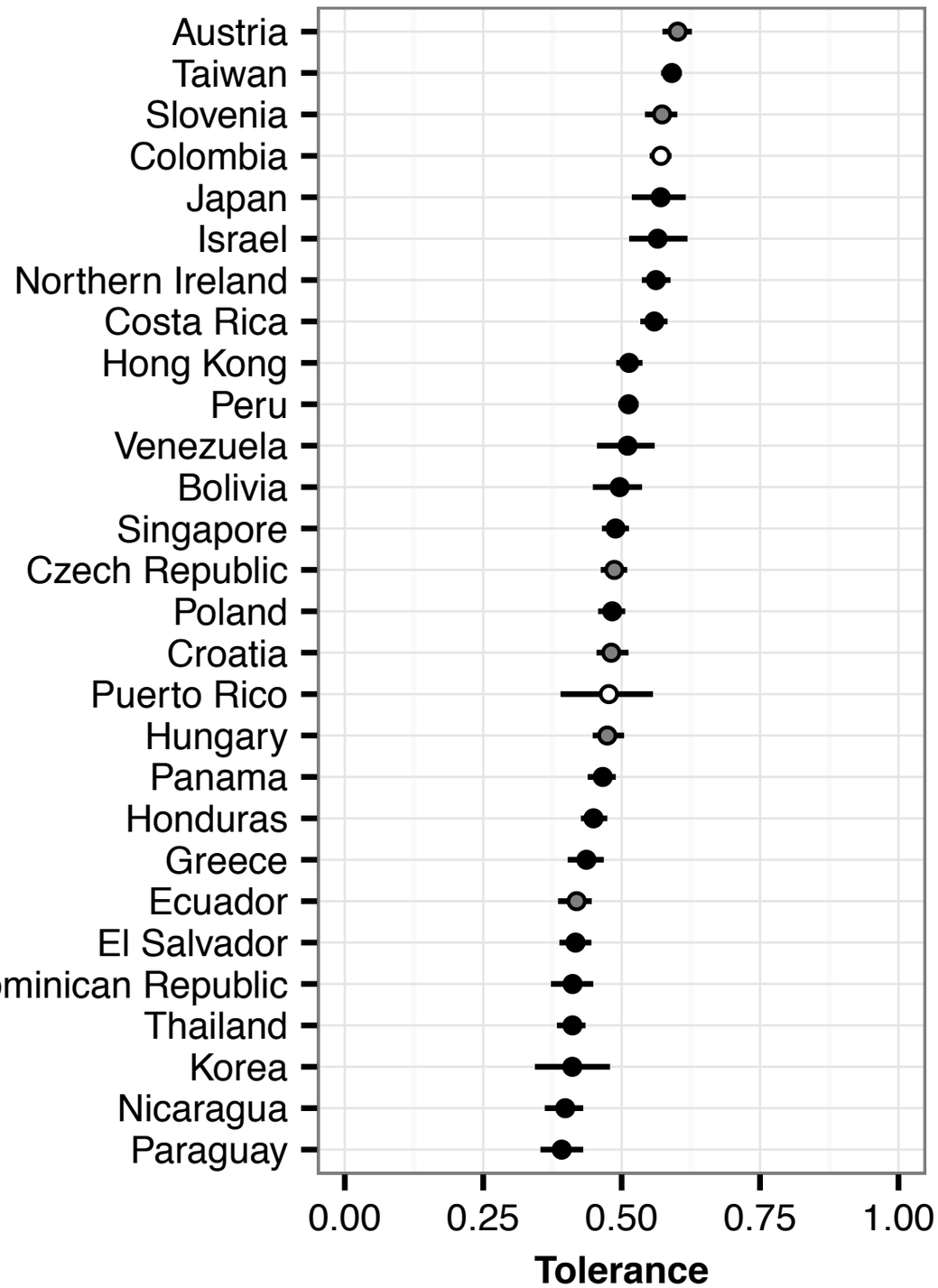
The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 23, Number 2, Spring 2016

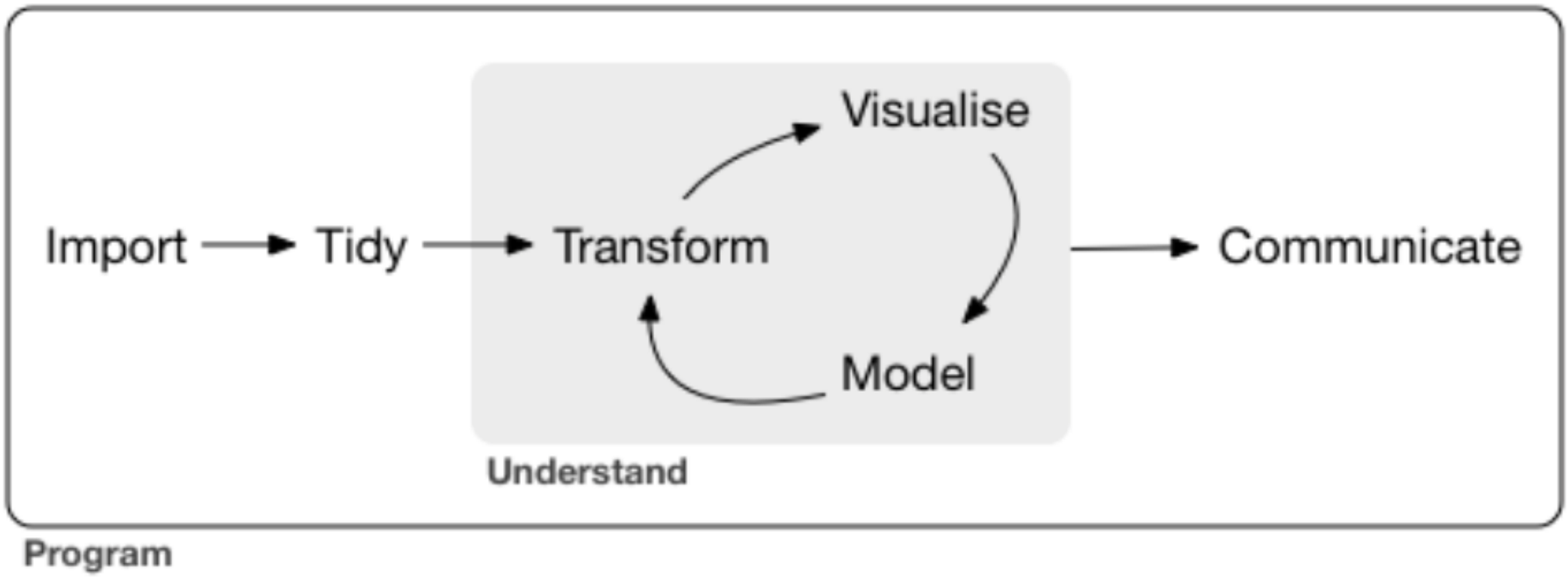
https://thepoliticalmethodologist.files.wordpress.com/2016/08/tpm_v23_n2.pdf

Computational Methods



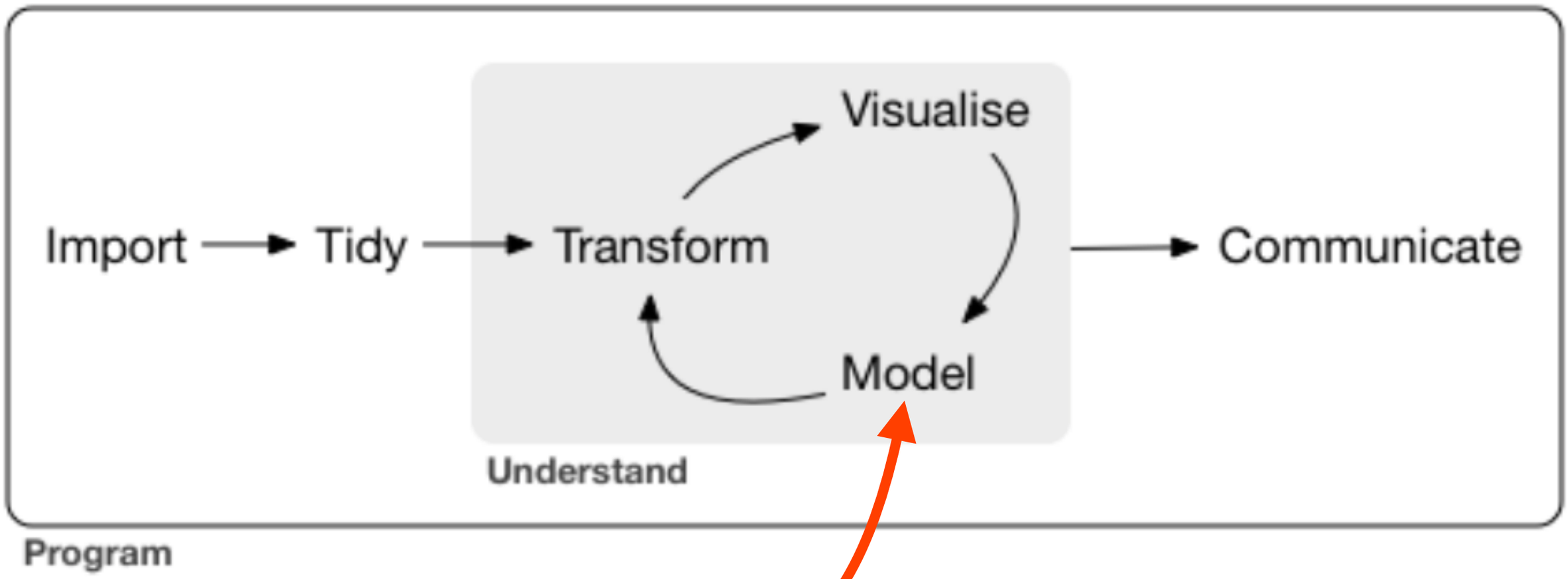


Data Science



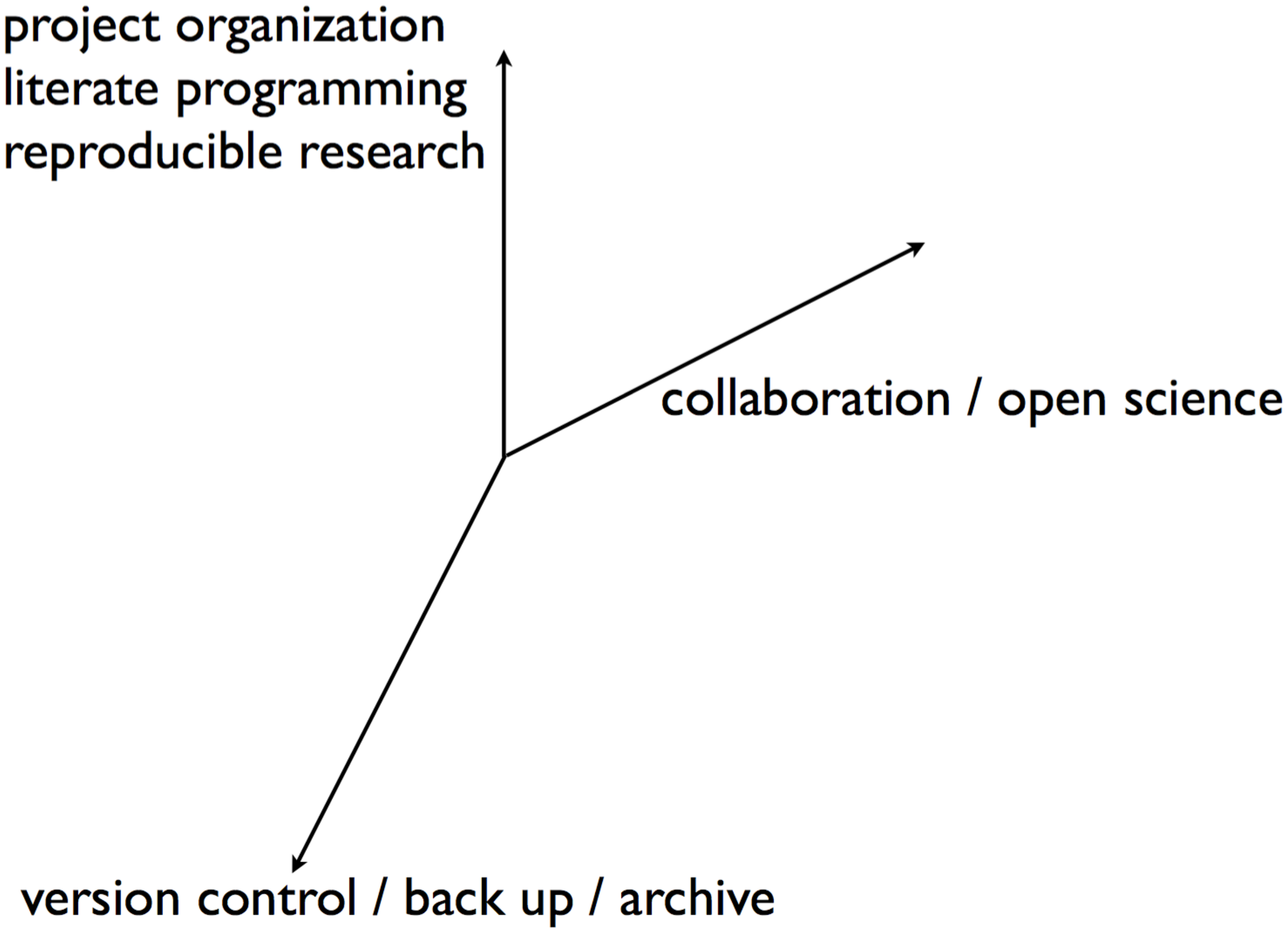
Grolemund and Wickham,
R for Data Science

Data Science

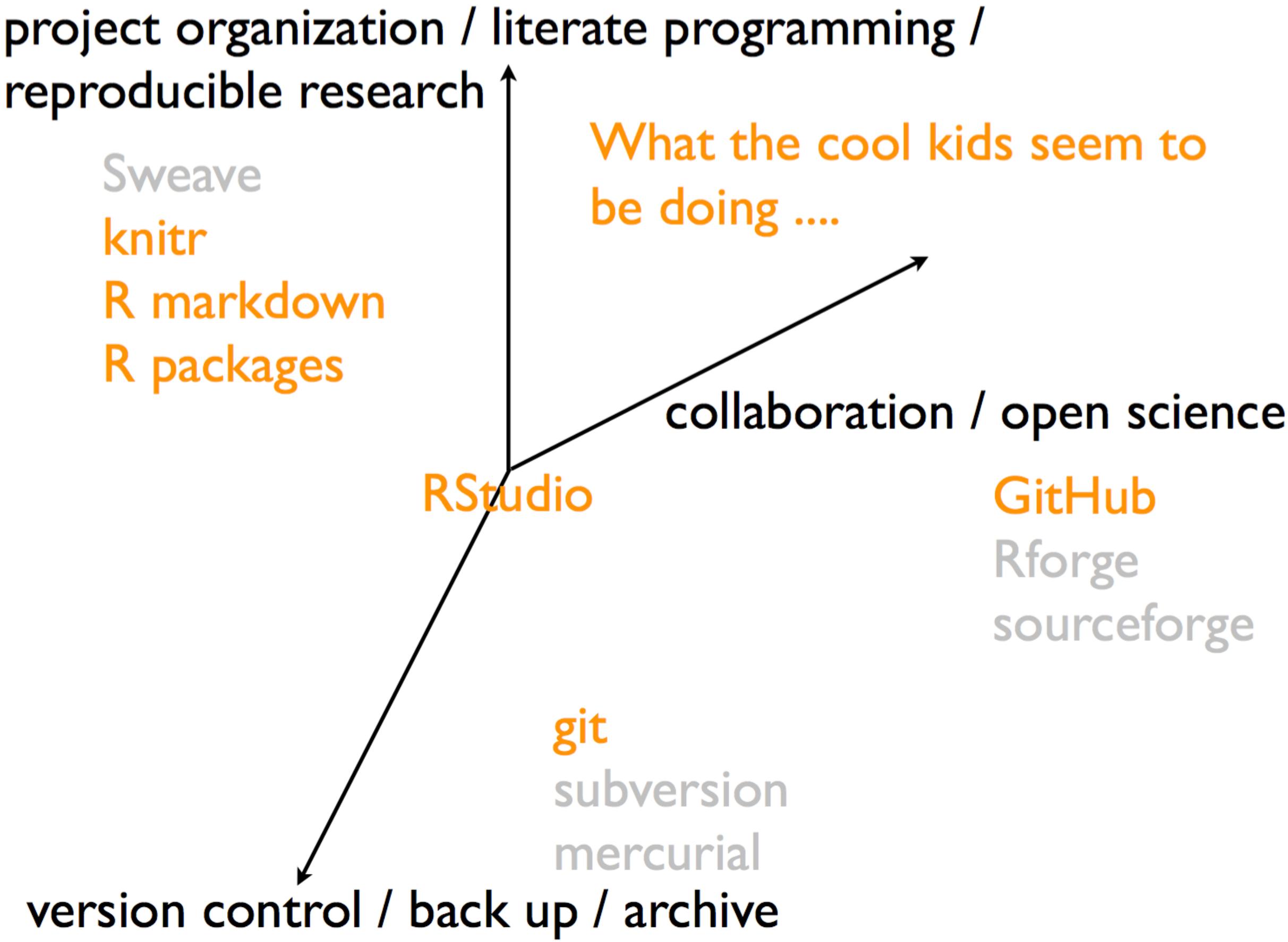


Our courses typically focus only on this…

Grolemund and Wickham,
R for Data Science

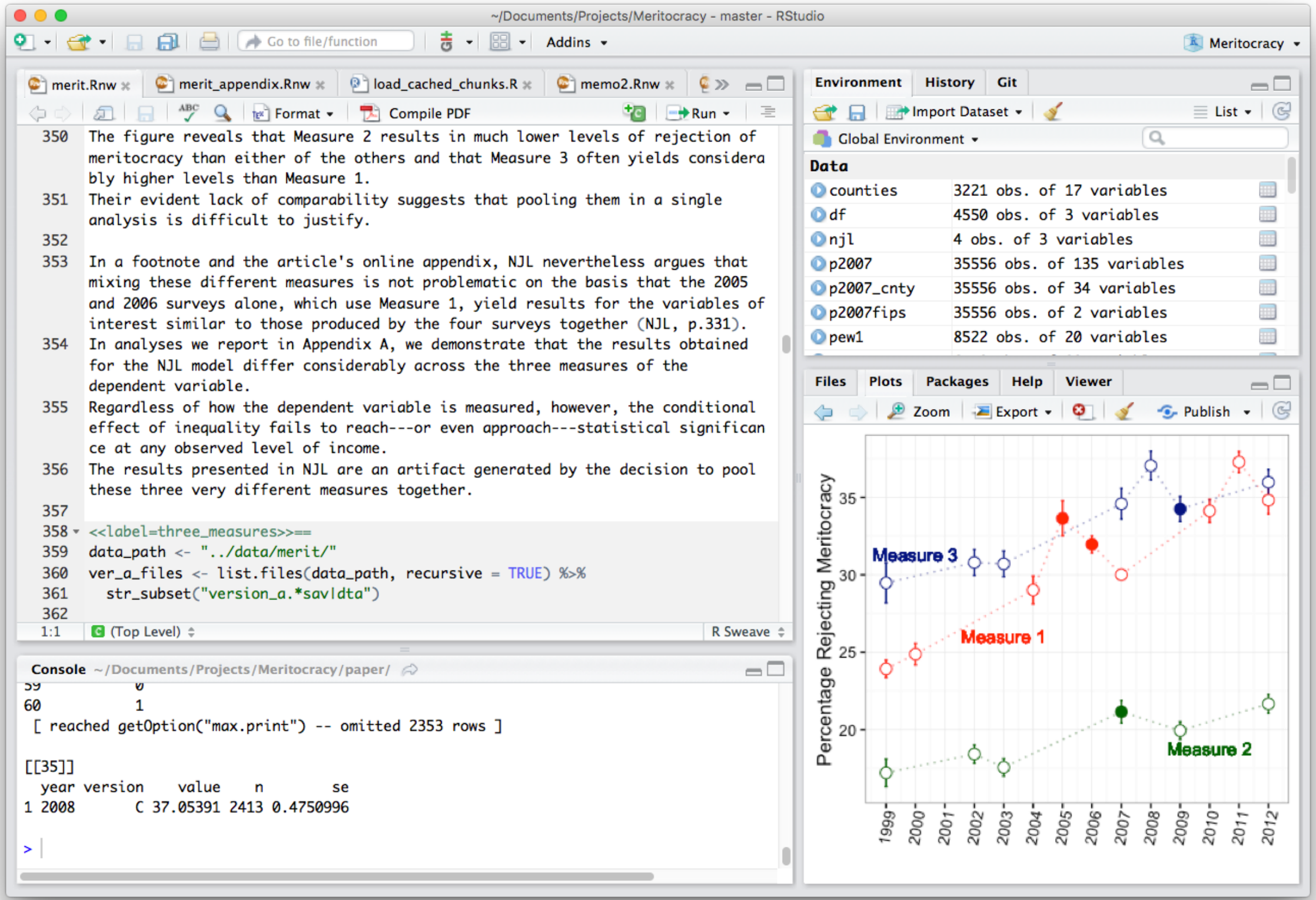


Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course



Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

RStudio


~/Documents/Projects/Meritocracy - master - RStudio
Go to file/function
Addins
Meritocracy
merit.Rnw
merit_appendix.Rnw
load_cached_chunks.R
memo2.Rnw
Format
Compile PDF
Run
350 The figure reveals that Measure 2 results in much lower levels of rejection of meritocracy than either of the others and that Measure 3 often yields considera bly higher levels than Measure 1.
351 Their evident lack of comparability suggests that pooling them in a single analysis is difficult to justify.
352
353 In a footnote and the article's online appendix, NJL nevertheless argues that mixing these different measures is not problematic on the basis that the 2005 and 2006 surveys alone, which use Measure 1, yield results for the variables of interest similar to those produced by the four surveys together (NJL, p.331).
354 In analyses we report in Appendix A, we demonstrate that the results obtained for the NJL model differ considerably across the three measures of the dependent variable.
355 Regardless of how the dependent variable is measured, however, the conditional effect of inequality fails to reach---or even approach---statistical significan ce at any observed level of income.
356 The results presented in NJL are an artifact generated by the decision to pool these three very different measures together.
357
358 <<label=three_measures>>==
359 data_path <- "../data/merit/"
360 ver_a_files <- list.files(data_path, recursive = TRUE) %>%
361 str_subset("version_a.*sav|dta")
362
1:1 (Top Level) R Sweave
Console ~/Documents/Projects/Meritocracy/paper/
59 0
60 1
[reached getOption("max.print") -- omitted 2353 rows]
[[35]]
year version value n se
1 2008 C 37.05391 2413 0.4750996
>
Environment History Git
Import Dataset List
Global Environment
Data
counties 3221 obs. of 17 variables
df 4550 obs. of 3 variables
njl 4 obs. of 3 variables
p2007 35556 obs. of 135 variables
p2007_cnty 35556 obs. of 34 variables
p2007fips 35556 obs. of 2 variables
pew1 8522 obs. of 20 variables
Files Plots Packages Help Viewer
Zoom Export Publish
Percentage Rejecting Meritocracy
35
30
25
20
Measure 3
Measure 1
Measure 2
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012

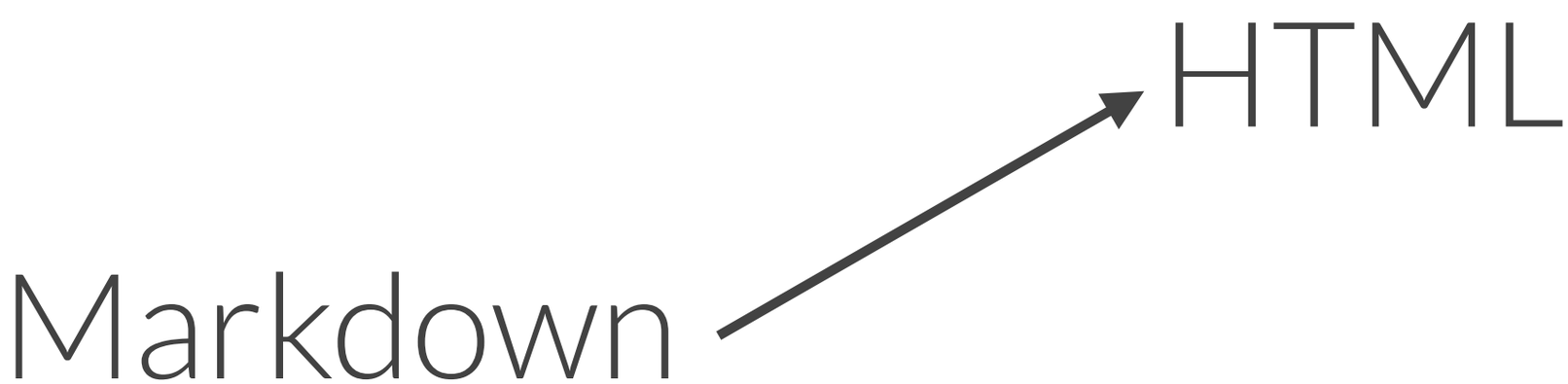
RStudio

- RStudio ≠ R
- It's an IDE—a way to interact with R
- It's made R much easier to use for reproducible research and collaboration
  - R Markdown (& **`knitr`**)
  - R + Git(Hub)

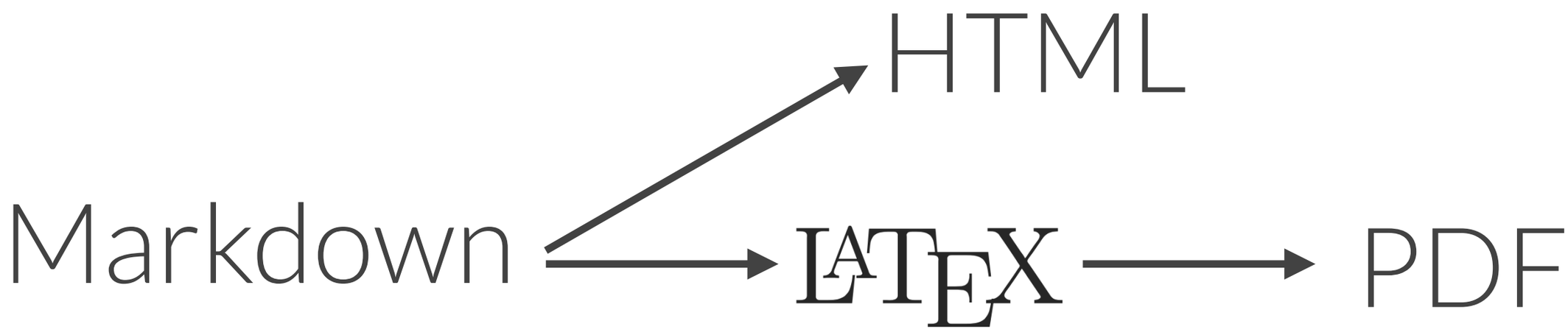
R Markdown

Markdown

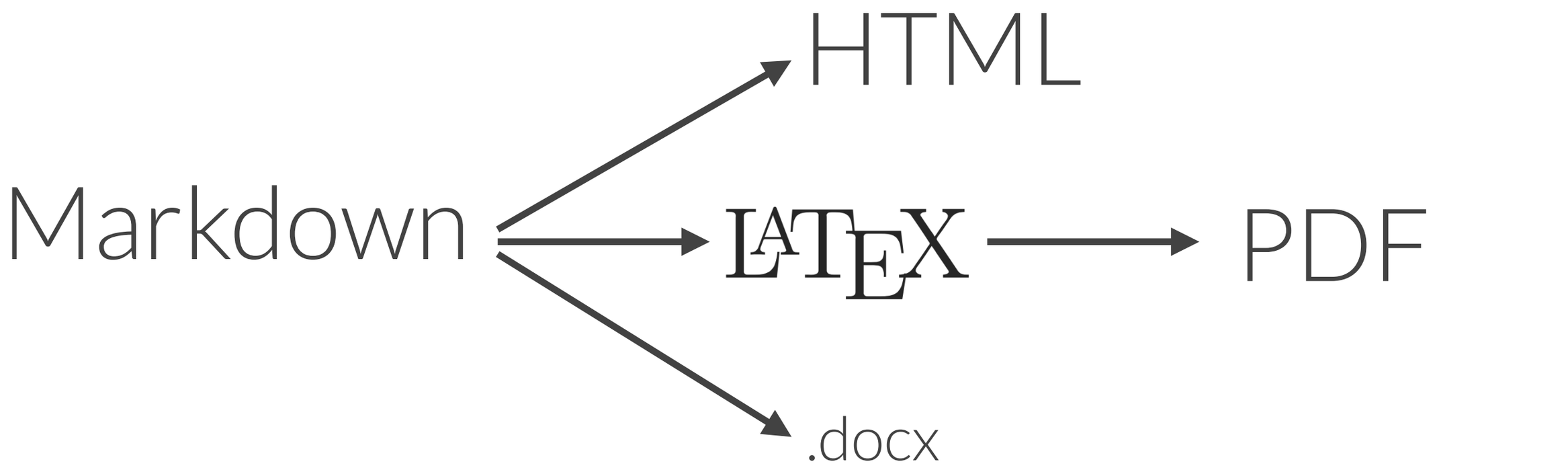
R Markdown



R Markdown



R Markdown



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
```

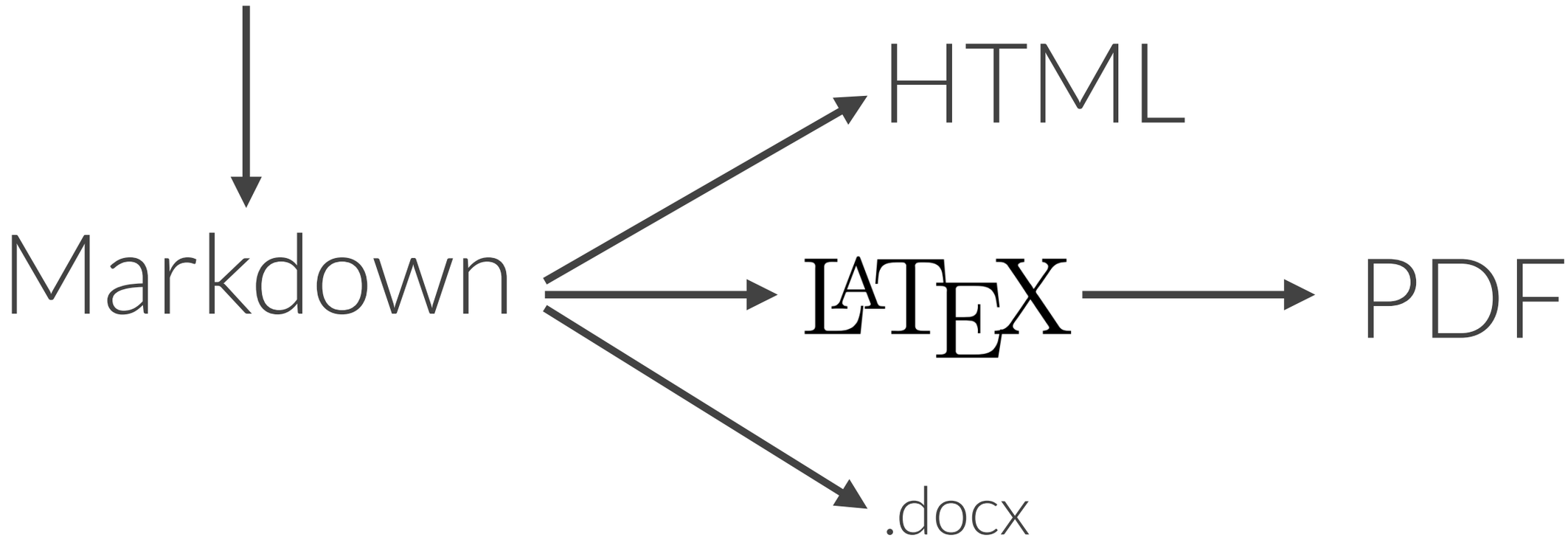
I can haz equations. Inline equations, such as … the average is computed as $\frac{1}{n}\sum_{i=1}^{n} x_i$. Or display equations like this:

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

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

R Markdown

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.

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

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

...
````

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

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)

...
```

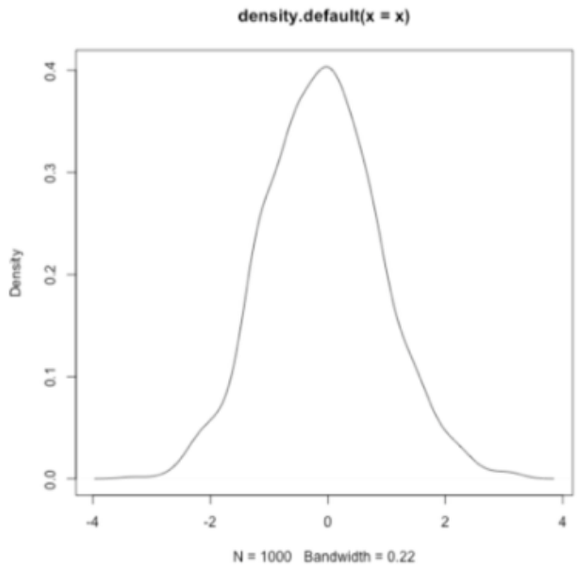
R Markdown rocks

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.



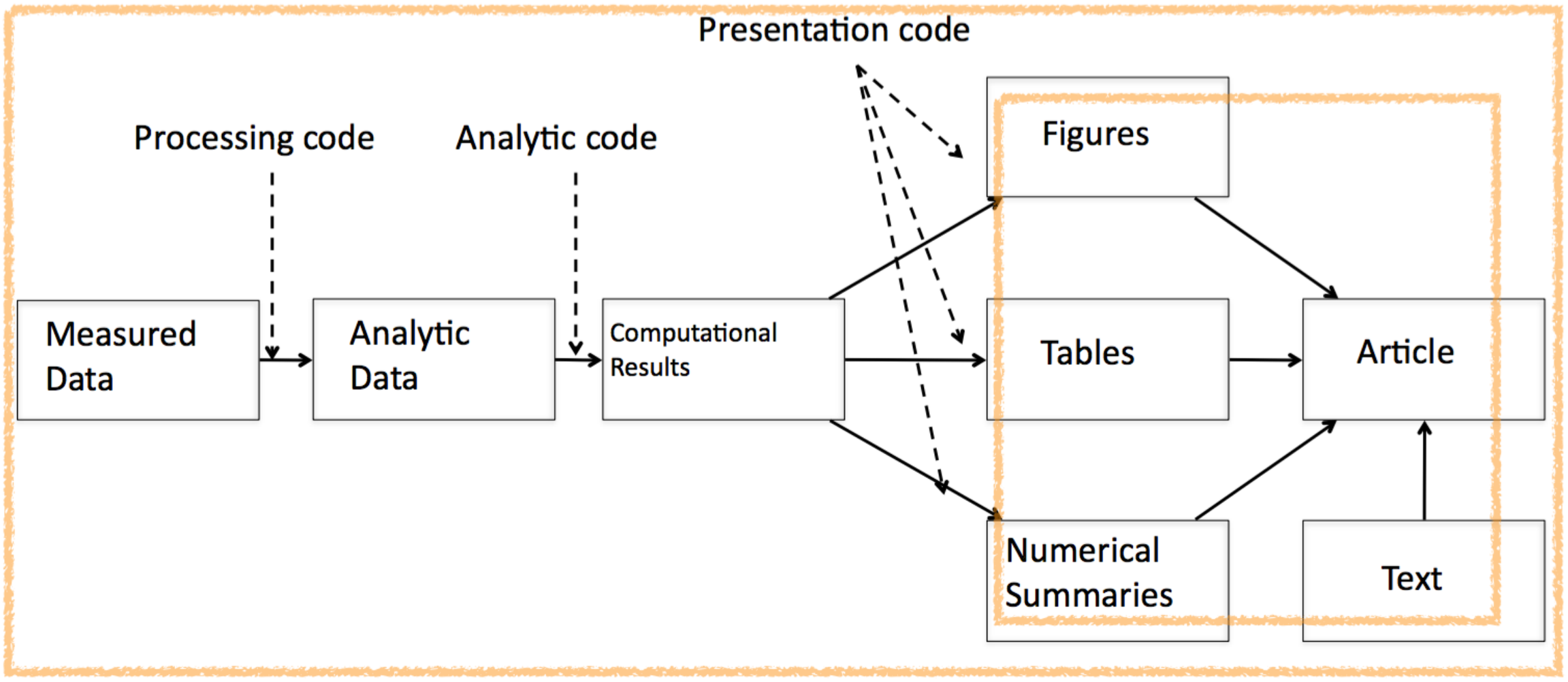
Note that all the previously demonstrated math typesetting still works. You don't have to choose between having math cred and being web-friendly!

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

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

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

R Markdown



Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

R Markdown

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
300 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 fraîche, 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 fraîche mixture, then remaining sponge. Drizzle with remaining crème de mûre, top with remaining crème fraîche mixture. Cover, refrigerate overnight. Serve scattered with extra berries and

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

Git & GitHub

Git & GitHub

- But how do I make my source code available to the world?

Git & GitHub

- But how do I make my source code available to the world?
- Or just share it with collaborators?

Git & GitHub

- But how do I make my source code available to the world?

- Or just share it with collaborators?

Advice to preserve sanity:
Stop doing this via email, attachments, and tracking changes in Word. Get that stuff into plain text, put it under version control and get it out on the web.

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

Git

http://www.phdcomics.com/comics/archive.php?comicid=1531

Git

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

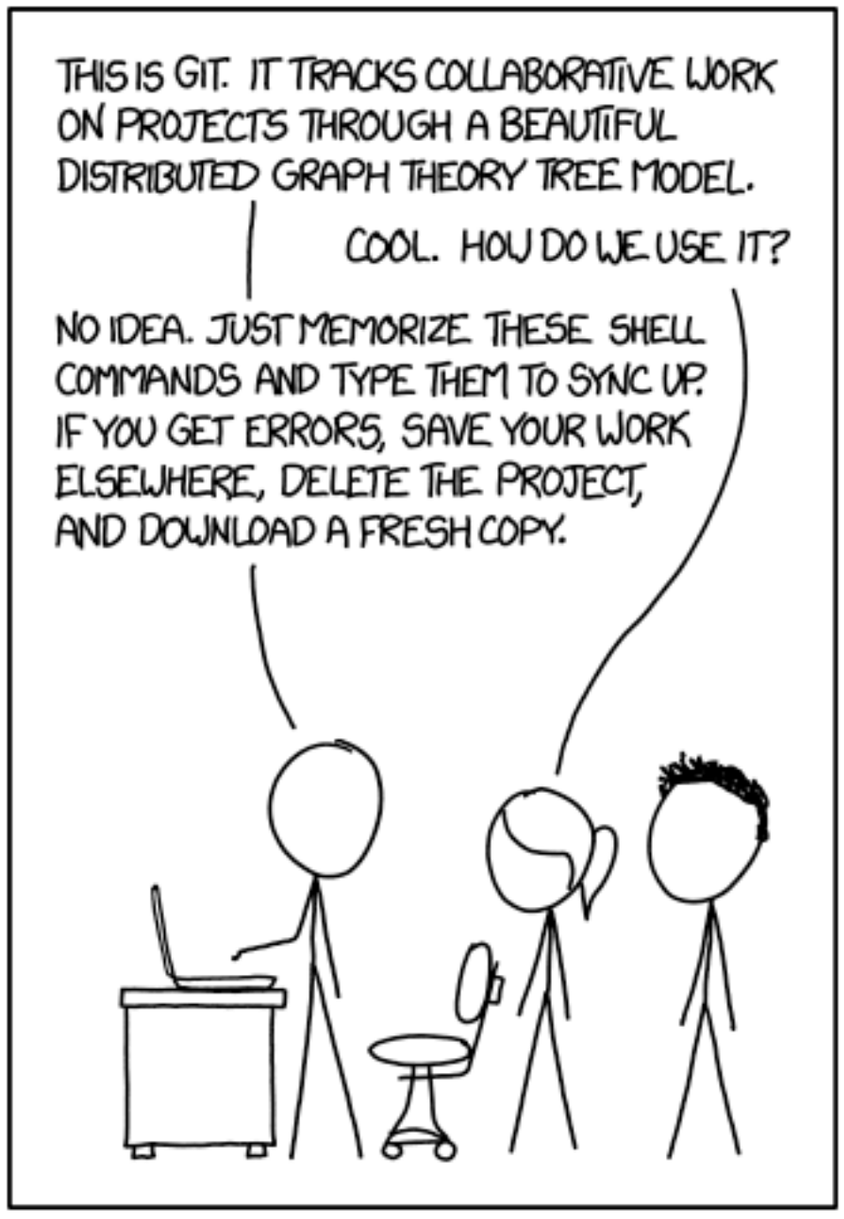
repo = repository

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

Git

collaboration = the "killer app" of version control

Learning Git has been -- and continues to be -- painful.



http://xkcd.com/1597

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

Git

collaboration = the "killer app" of version control

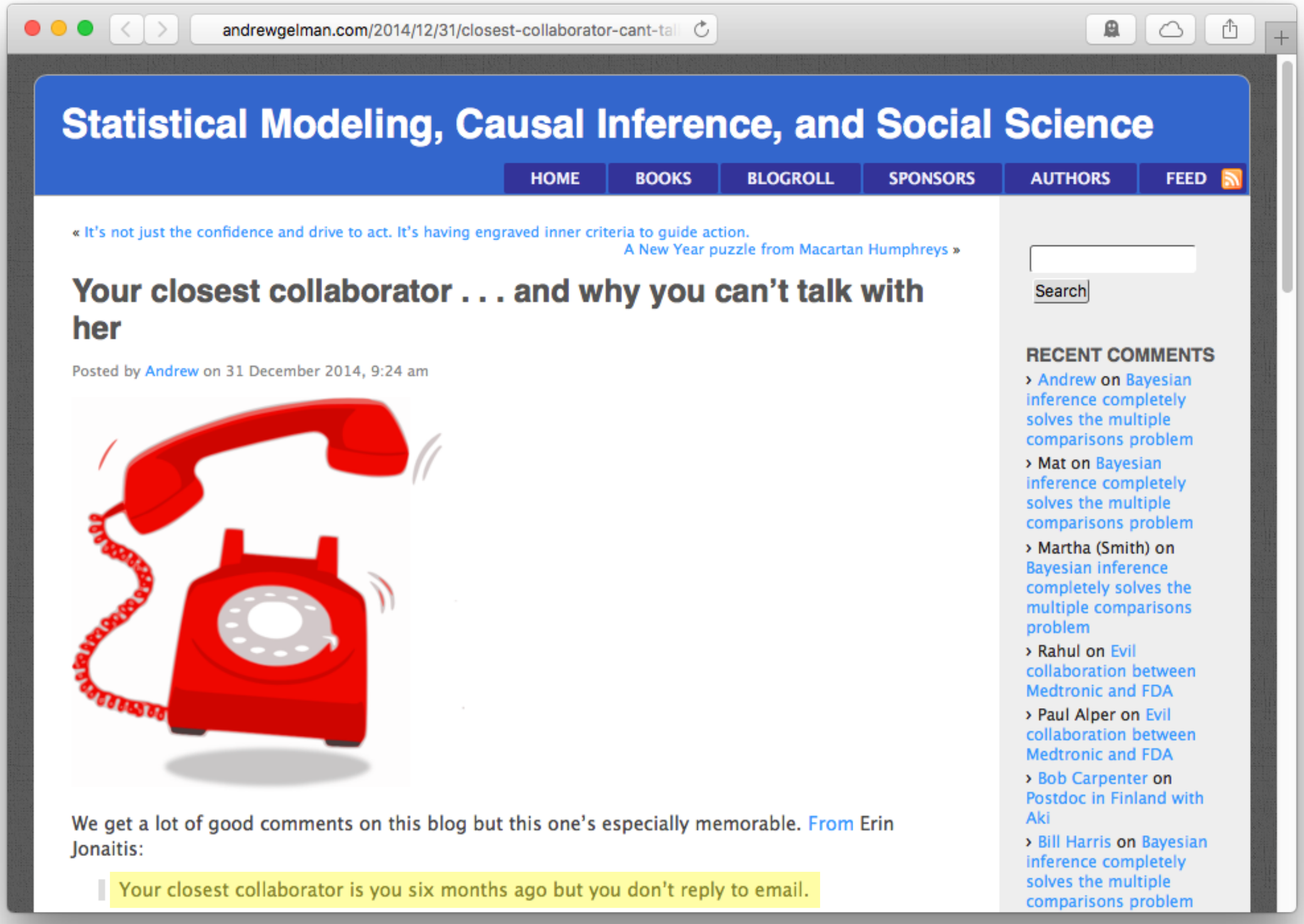
Learning Git has been -- and continues to be -- painful. But not nearly as crazy-making as the alternatives:

- 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

Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

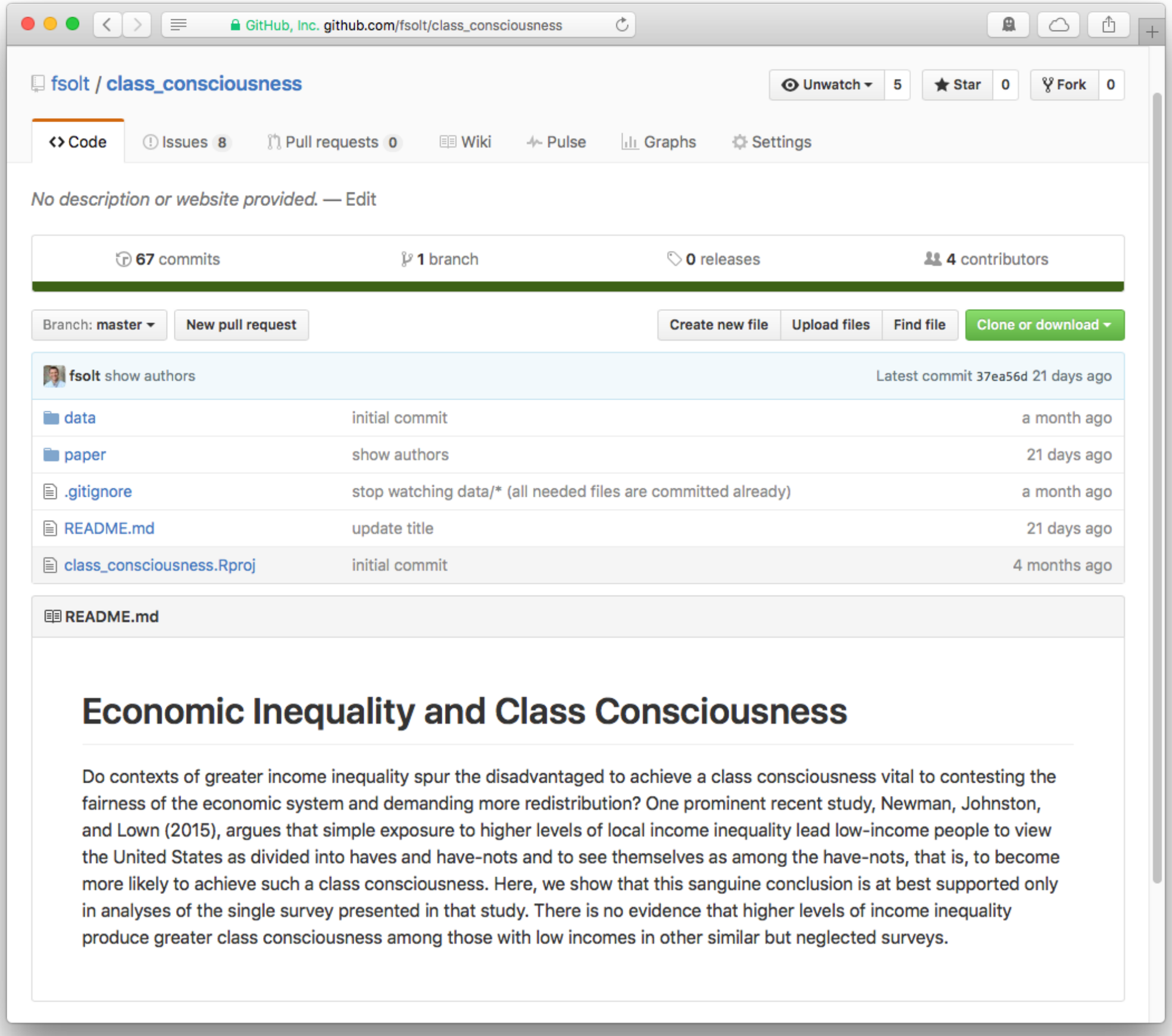
Git

and remember:

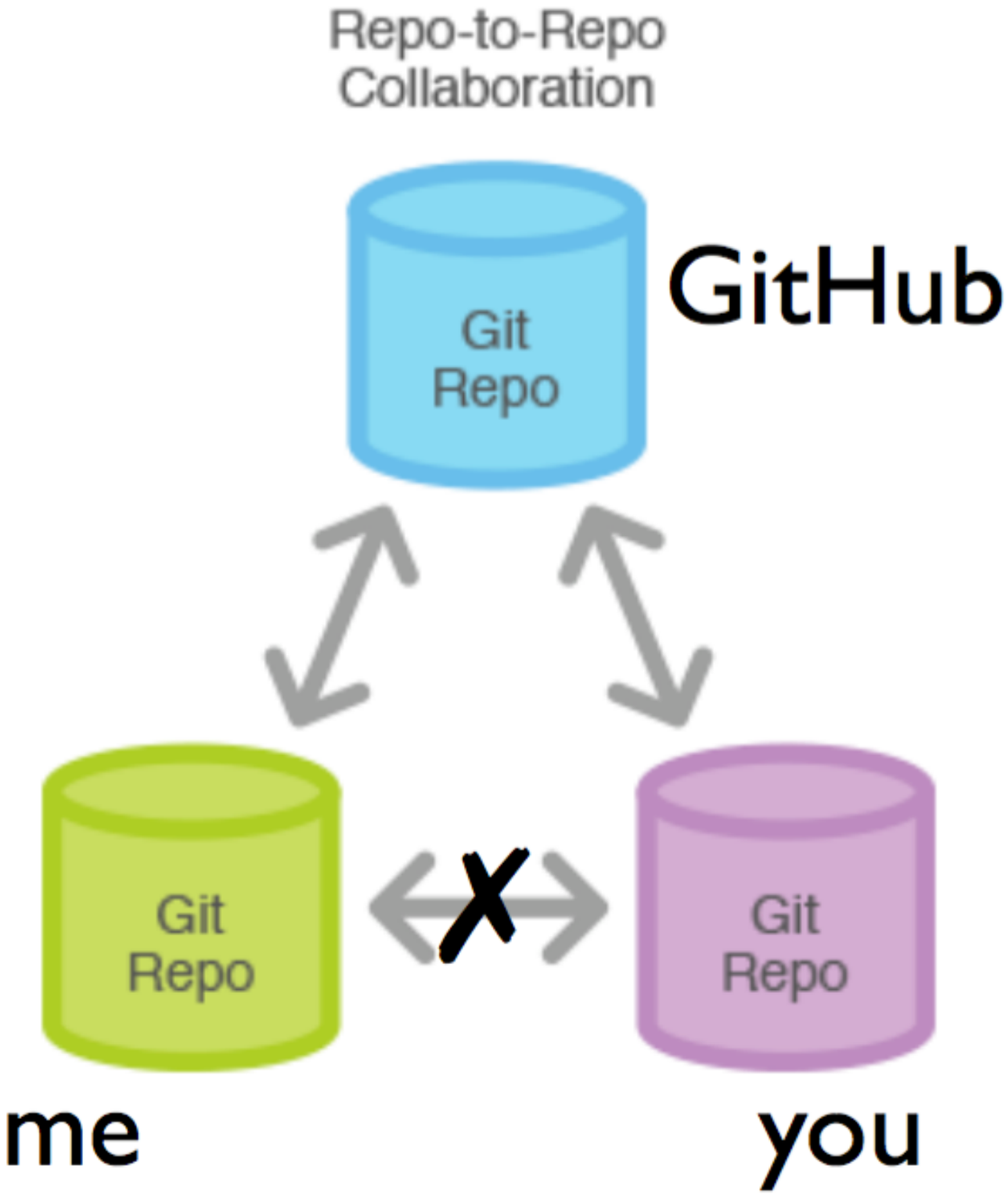


http://andrewgelman.com/2014/12/31/closest-collaborator-cant-talk

GitHub


GitHub, Inc. github.com/fsolt/class_consciousness
fsolt / class_consciousness
Unwatch 5
Star 0
Fork 0
Code
Issues 8
Pull requests 0
Wiki
Pulse
Graphs
Settings
No description or website provided. — Edit
67 commits
1 branch
0 releases
4 contributors
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
fsolt show authors
Latest commit 37ea56d 21 days ago
data
initial commit
a month ago
paper
show authors
21 days ago
.gitignore
stop watching data/* (all needed files are committed already)
a month ago
README.md
update title
21 days ago
class_consciousness.Rproj
initial commit
4 months ago
README.md
Economic Inequality and Class Consciousness
Do contexts of greater income inequality spur the disadvantaged to achieve a class consciousness vital to contesting the fairness of the economic system and demanding more redistribution? One prominent recent study, Newman, Johnston, and Lown (2015), argues that simple exposure to higher levels of local income inequality lead low-income people to view the United States as divided into haves and have-nots and to see themselves as among the have-nots, that is, to become more likely to achieve such a class consciousness. Here, we show that this sanguine conclusion is at best supported only in analyses of the single survey presented in that study. There is no evidence that higher levels of income inequality produce greater class consciousness among those with low incomes in other similar but neglected surveys.

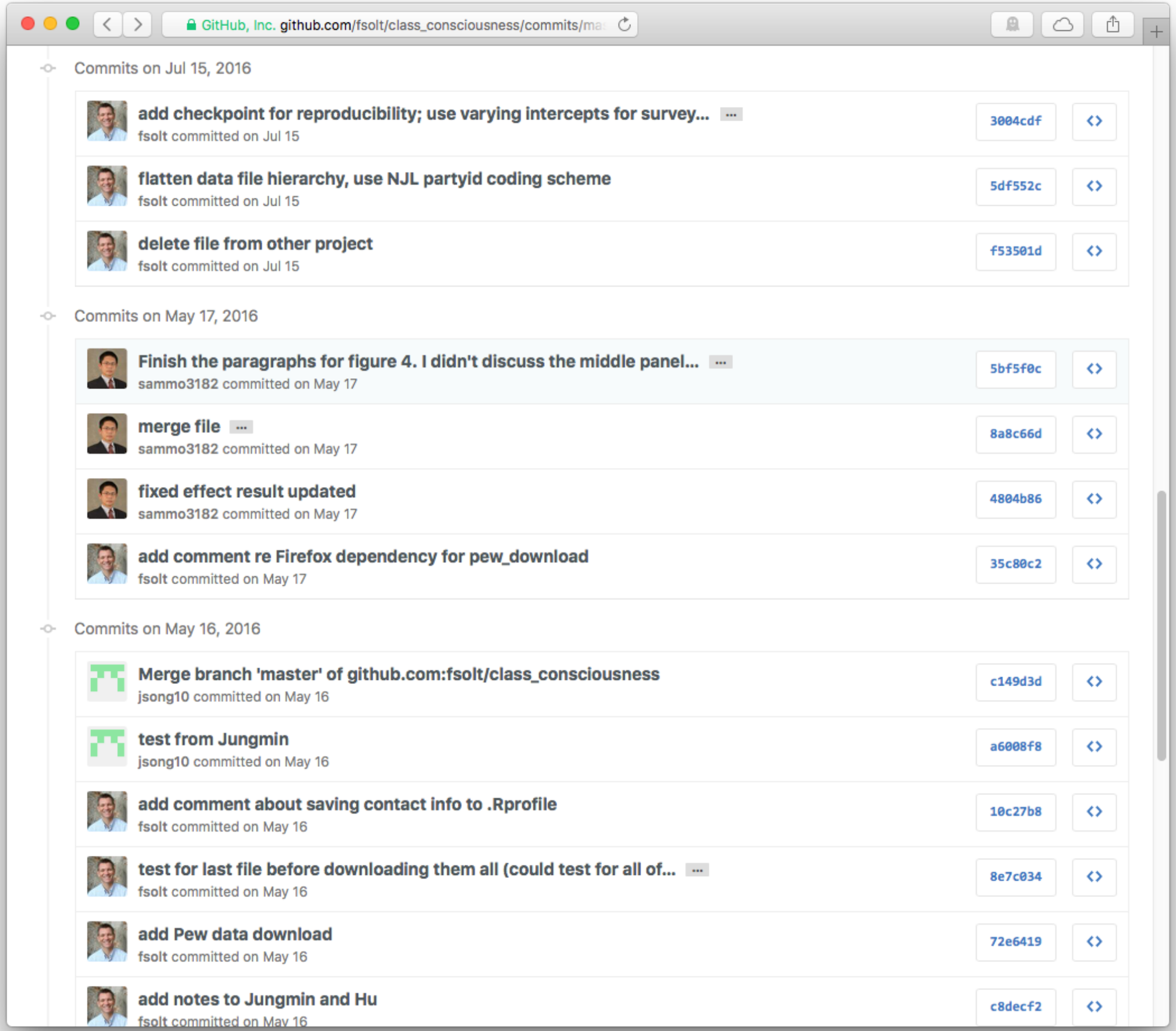
GitHub



Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course

GitHub

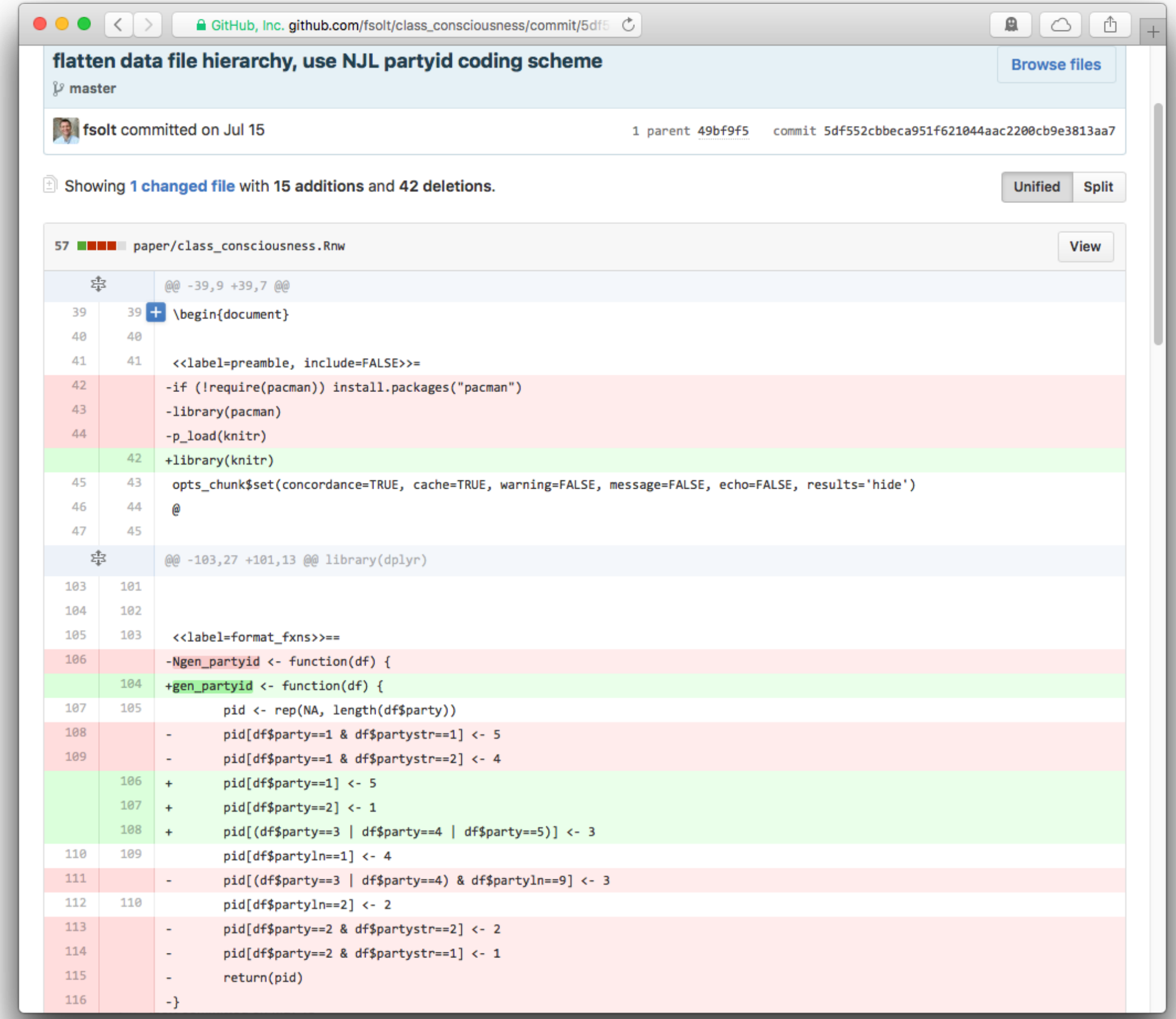
History of *all* changes

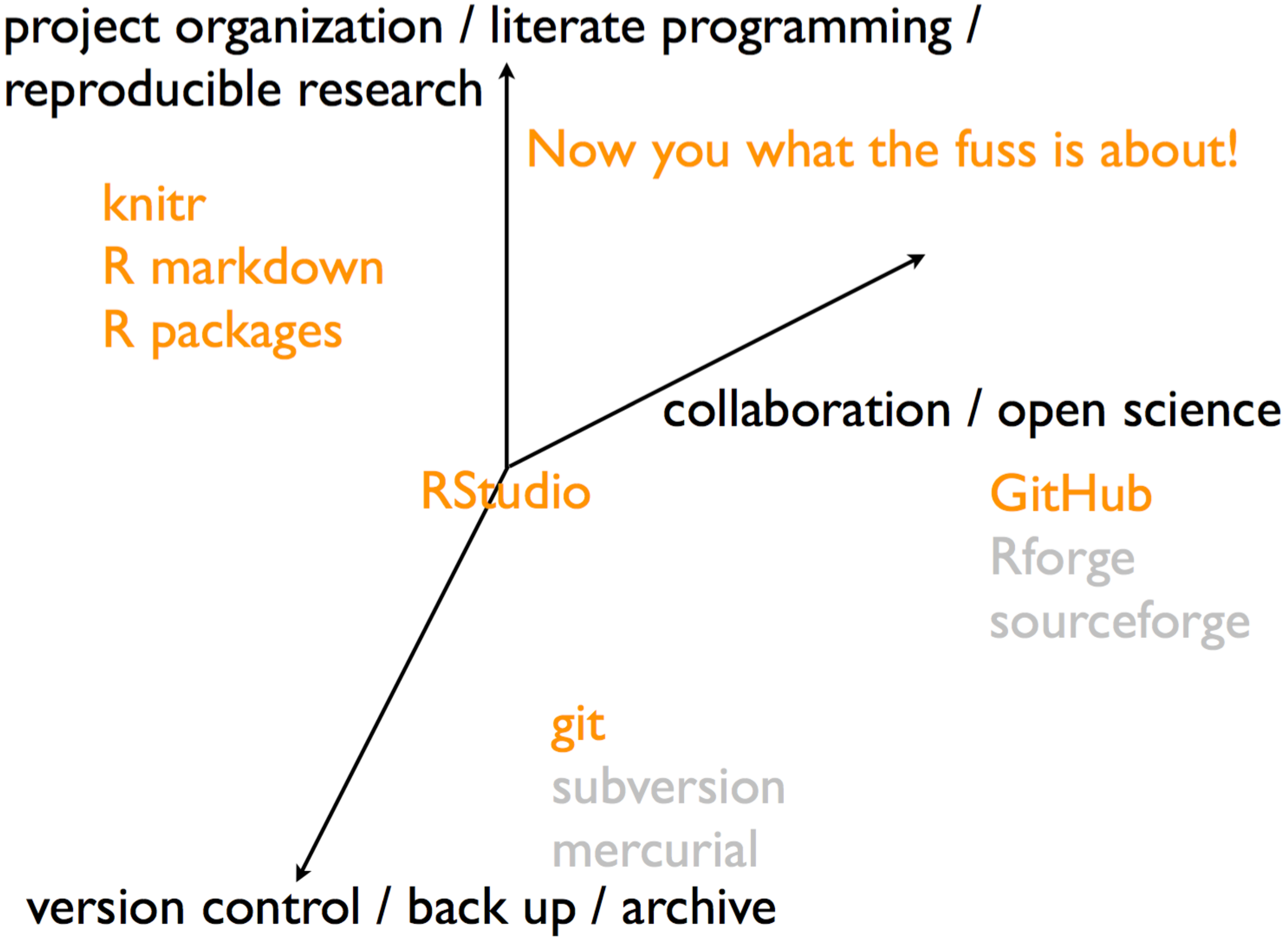


GitHub

History of *all* changes

‘diff’ within documents





Jennifer Bryan, https://speakerdeck.com/jennybc/ubc-stat545-2015-cm001-intro-to-course