Zen And The aRt Of Workflow Maintenance

Jennifer Bryan
RStudio, University of British Columbia
@JennyBryan @jennybc

rstd.io/jenny-latinr

links to stuff in this talk!!

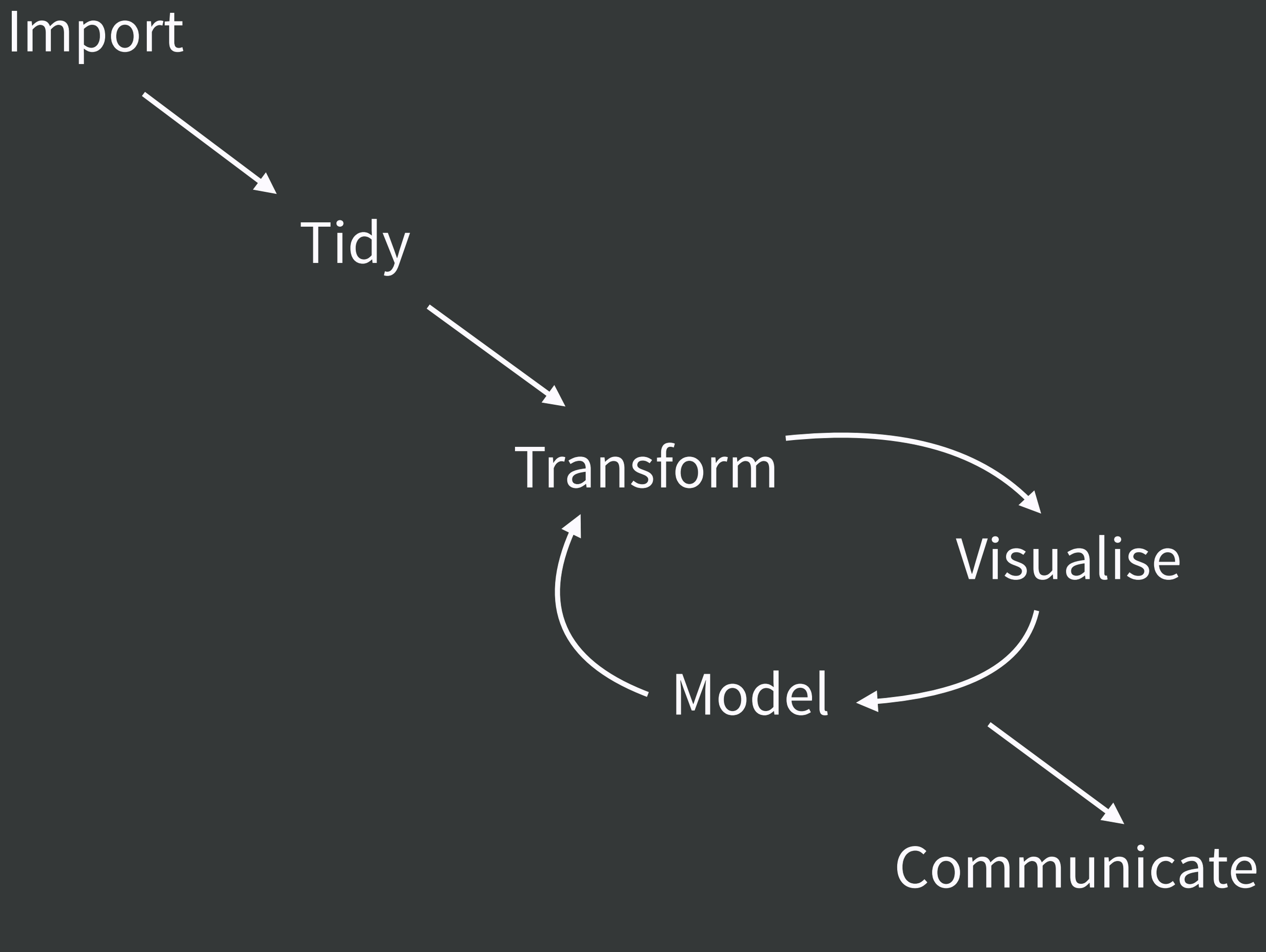
Is data science just a trendy term for statistics?

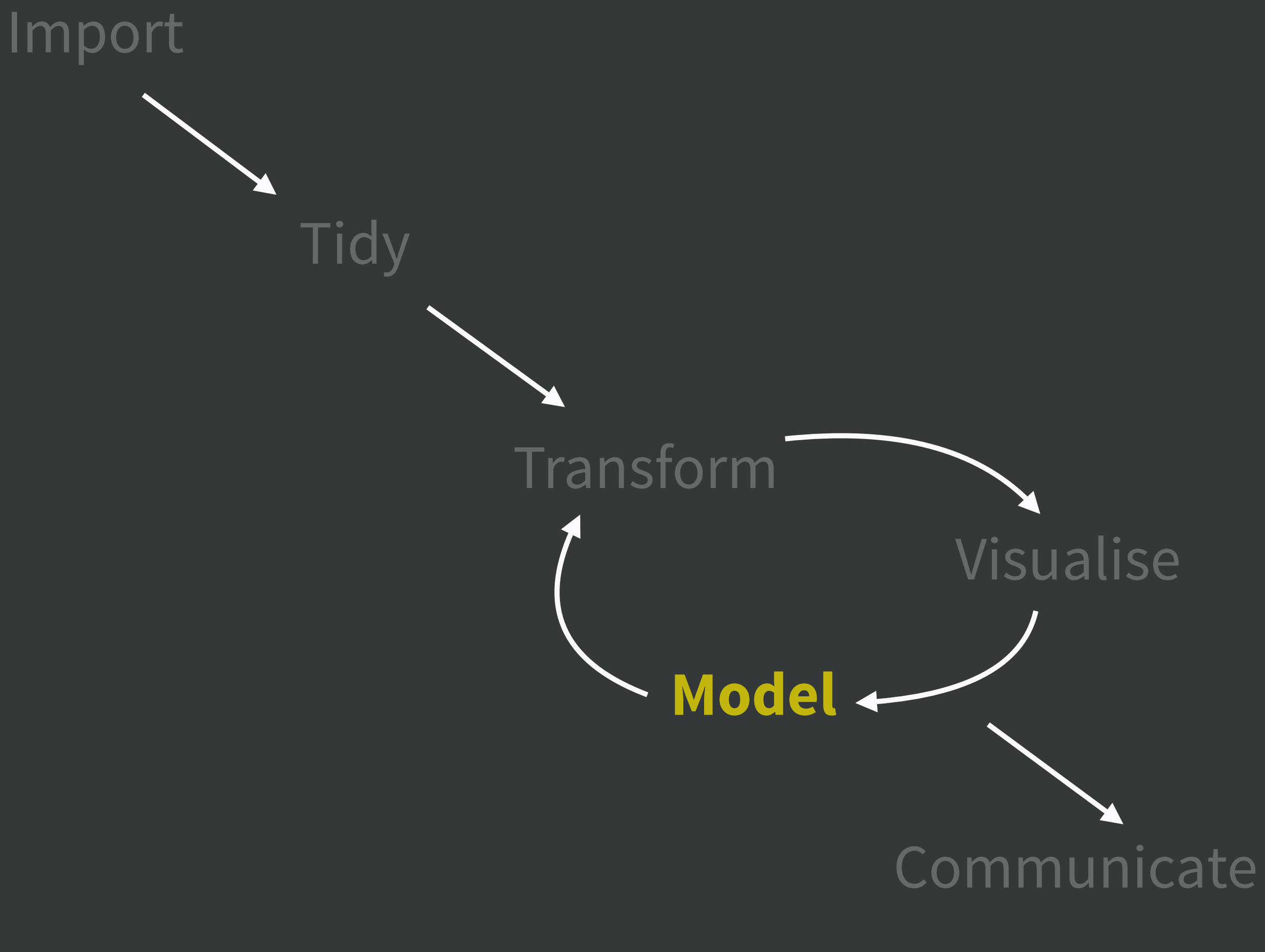
7 Life-Changing Workflow Tips Every useR Should Know

Is data science just a trendy term for statistics?

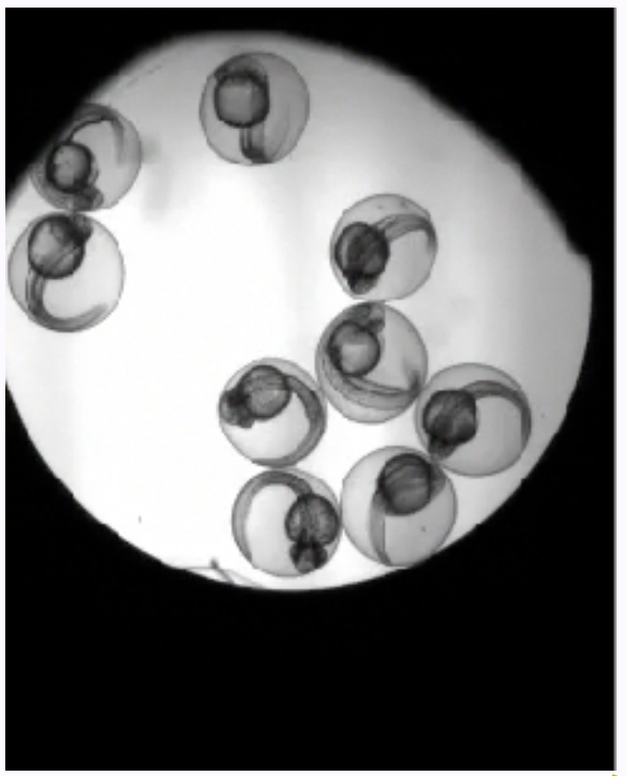
Is data science just a trendy term for statistics?

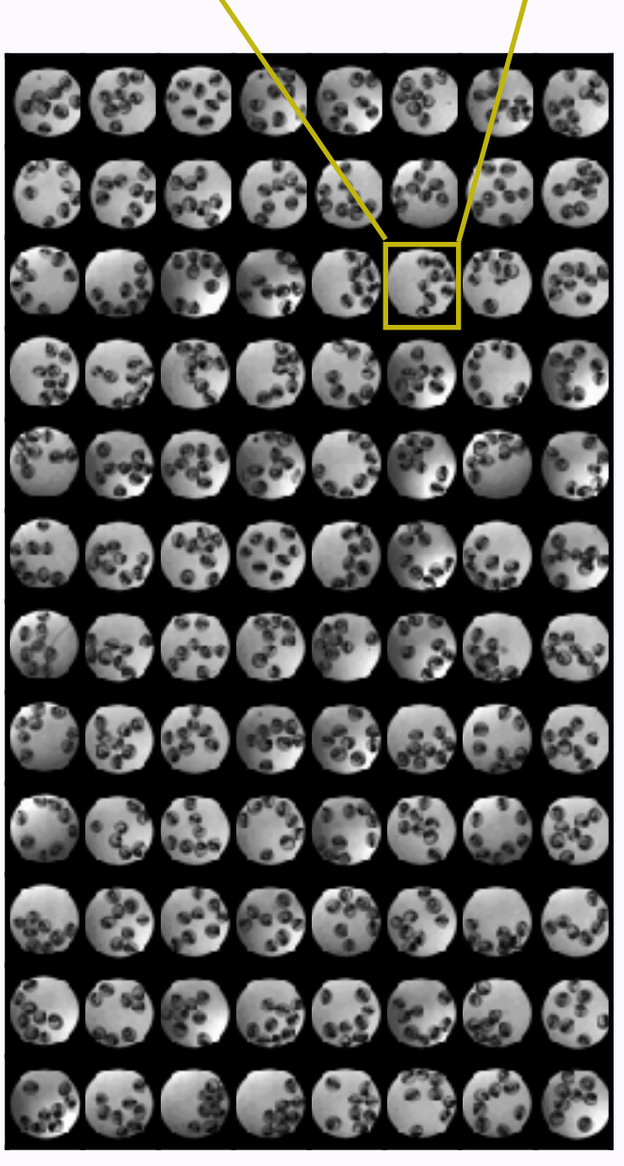
No.

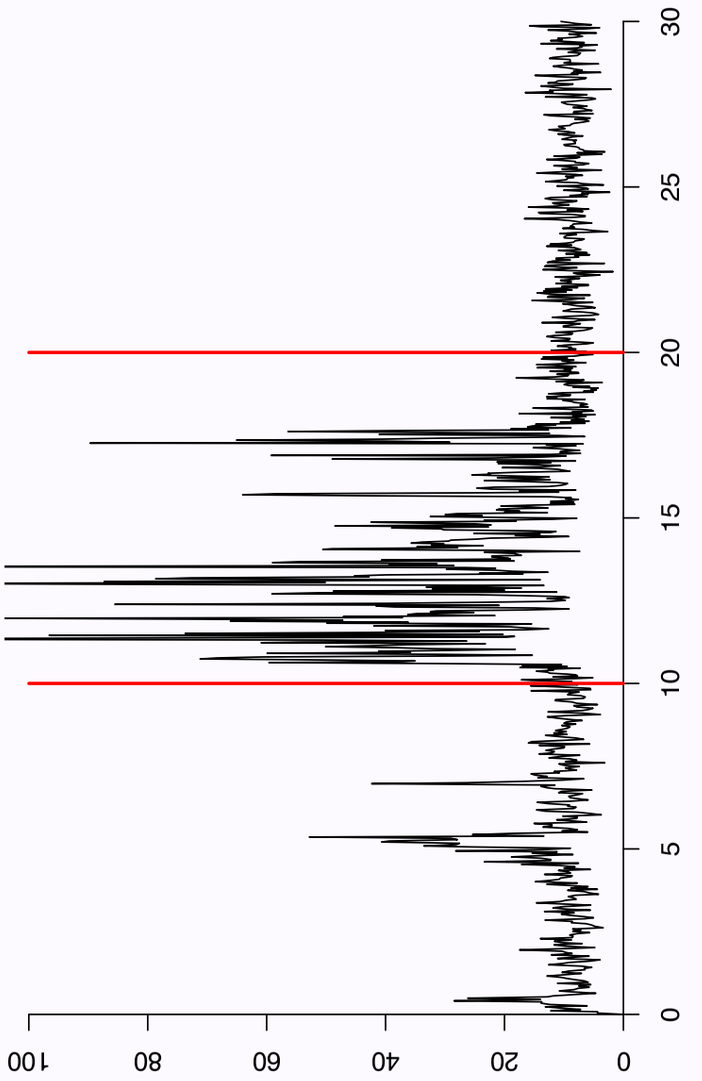

Import
Tidy
Transform
Visualise
Model
Communicate

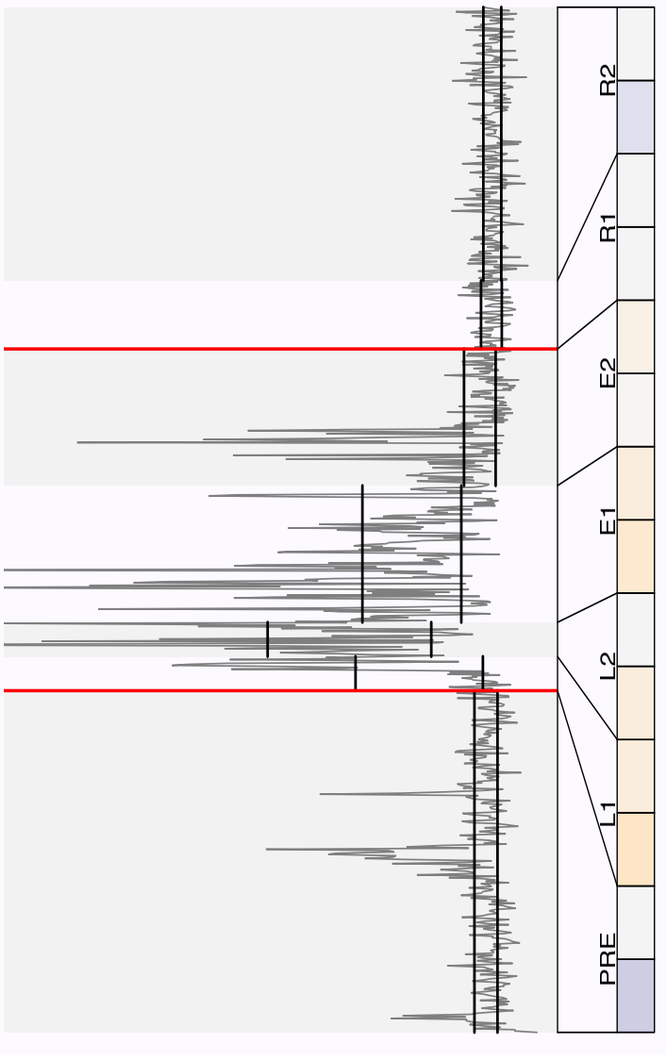

Import
Tidy
Transform
Visualise
Model
Communicate

Quandary of the applied statistician







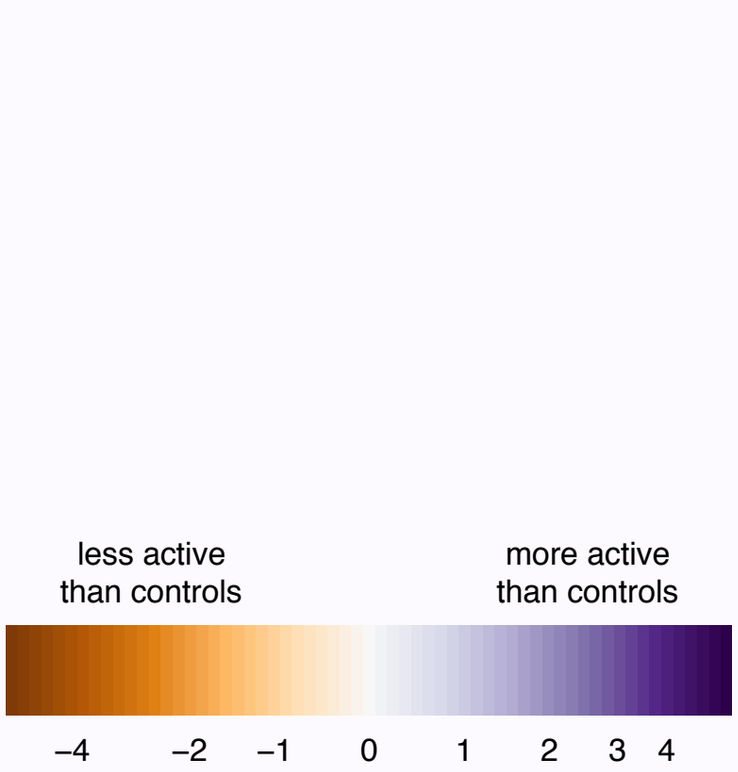


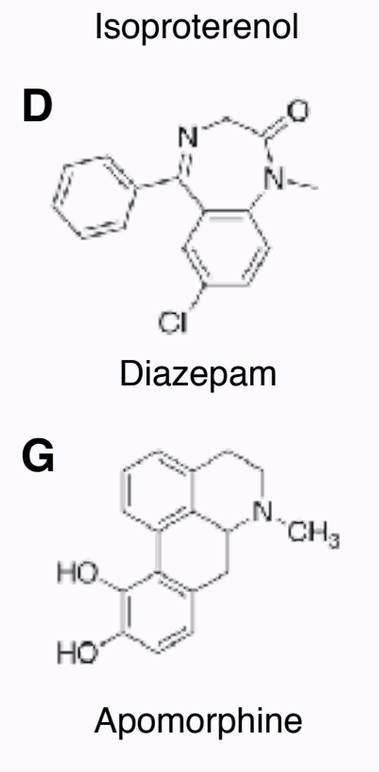
Motion Index

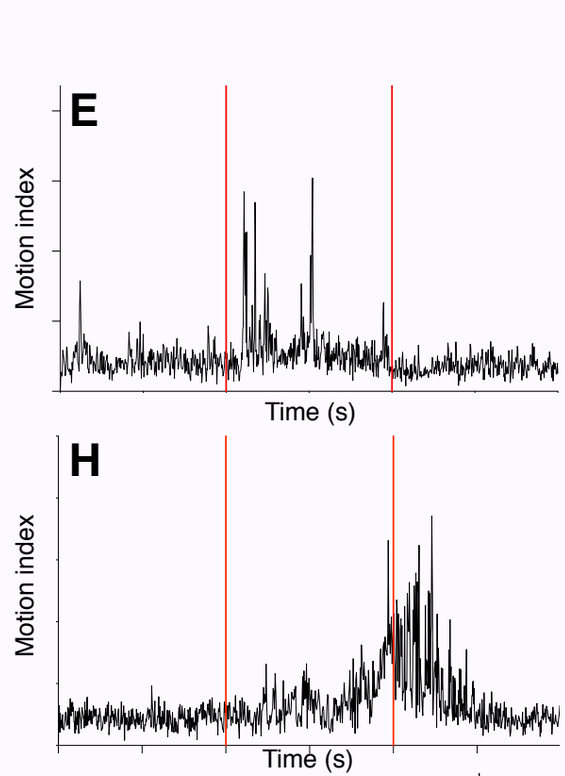
Rapid behavior-based identification of neuroactive small molecules in the zebrafish

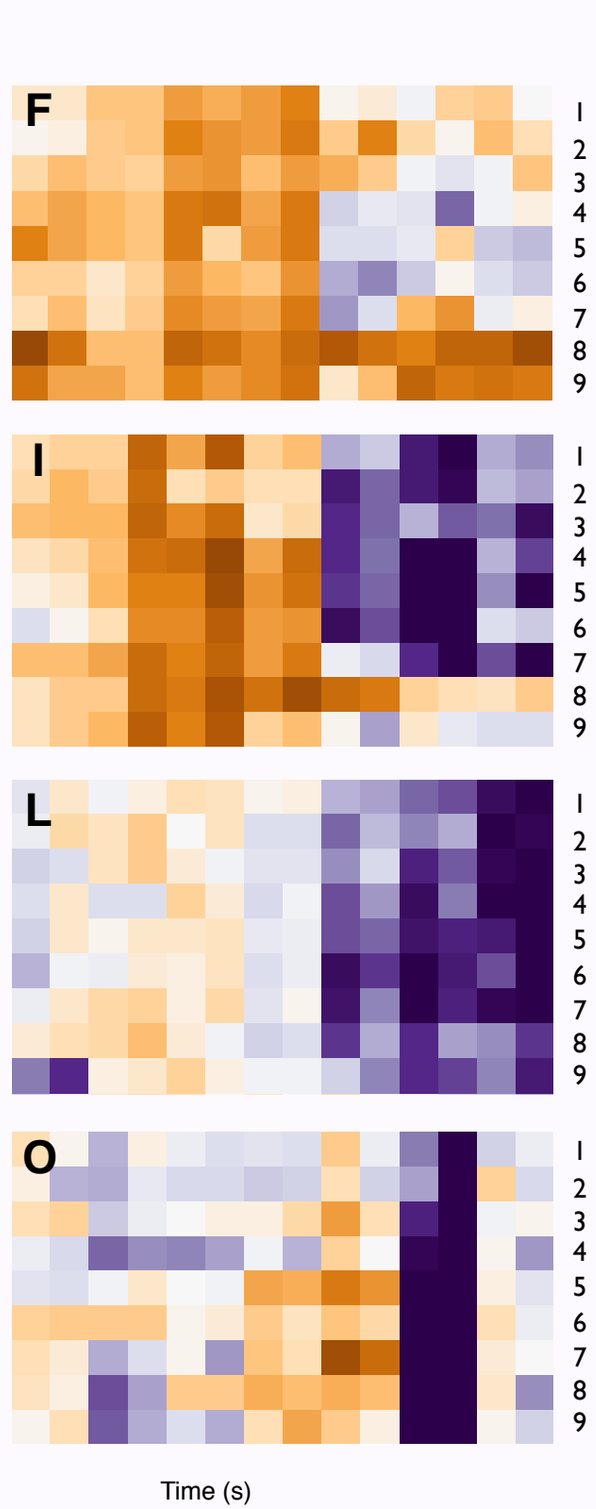
David Kokel[1,2*], Jennifer Bryan[3,4], Christian Laggner[5], Rick White[3], Chung Yan J Cheung[1,2], Rita Mateus[1,2], David Healey[1,2], Sonia Kim[1,2], Andreas A Werdich[1], Stephen J Haggarty[2,6,7], Calum A MacRae[1], Brian Shoichet[5] & Randall T Peterson[1,2*]

NATURE CHEMICAL BIOLOGY | VOL 6 | MARCH 2010

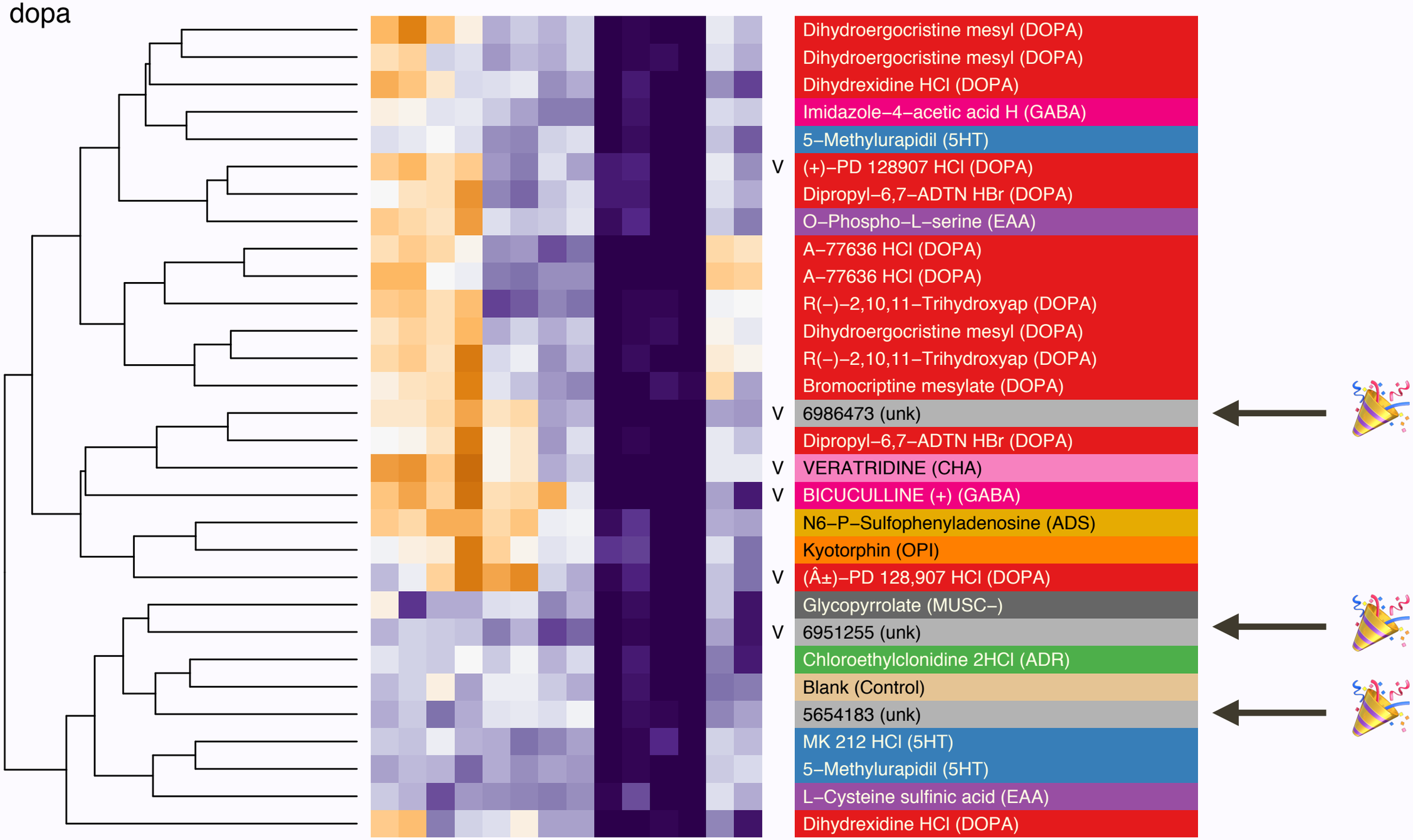








dopa


Dihydroergocristine mesyl (DOPA)
Dihydroergocristine mesyl (DOPA)
Dihydrexidine HCl (DOPA)
Imidazole–4–acetic acid H (GABA)
5–Methylurapidil (5HT)
V (+)–PD 128907 HCl (DOPA)
Dipropyl–6,7–ADTN HBr (DOPA)
O–Phospho–L–serine (EAA)
A–77636 HCl (DOPA)
A–77636 HCl (DOPA)
R(–)–2,10,11–Trihydroxyap (DOPA)
Dihydroergocristine mesyl (DOPA)
R(–)–2,10,11–Trihydroxyap (DOPA)
Bromocriptine mesylate (DOPA)
V 6986473 (unk)
Dipropyl–6,7–ADTN HBr (DOPA)
V VERATRIDINE (CHA)
V BICUCULLINE (+) (GABA)
N6–P–Sulfophenyladenosine (ADS)
Kyotorphin (OPI)
V (Â±)–PD 128,907 HCl (DOPA)
Glycopyrrolate (MUSC–)
V 6951255 (unk)
Chloroethylclonidine 2HCl (ADR)
Blank (Control)
5654183 (unk)
MK 212 HCl (5HT)
5–Methylurapidil (5HT)
L–Cysteine sulfinic acid (EAA)
Dihydrexidine HCl (DOPA)

STAT 545

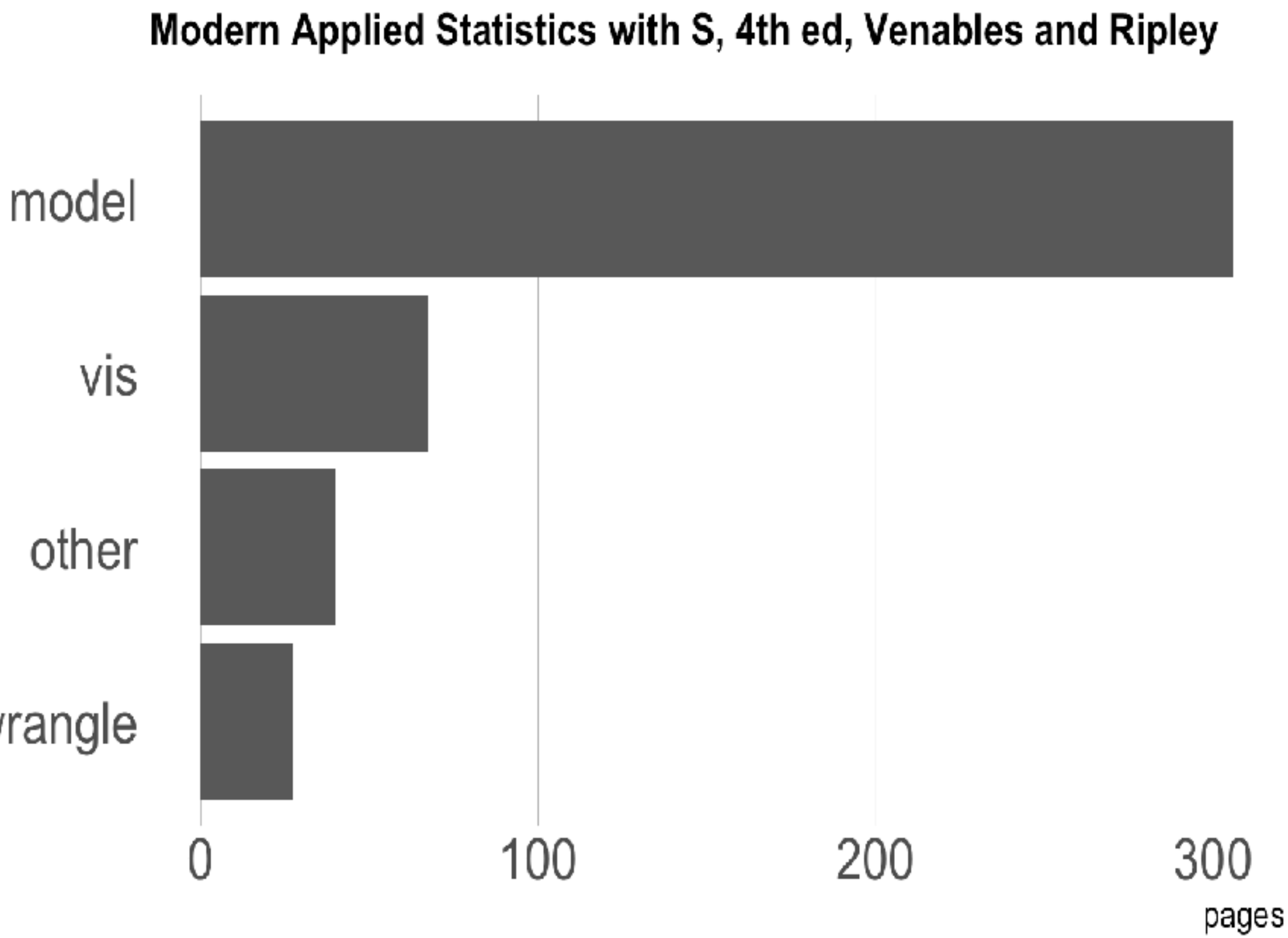
http://stat545.com

Statistics and Computing

W.N. Venables
B.D. Ripley

Modern Applied Statistics with S-Plus

Springer Science+
Business Media, LLC


Modern Applied Statistics with S, 4th ed, Venables and Ripley
model
vis
other
wrangle
0
100
200
300
pages

How STAT 545 projects went sideways:
An Incomplete List

inability to
  … scrape data off the web
  … request data from an API
  … parse JSON or XML
utter defeat by date times
text encoding fiascos
ineptitude with regular expressions
R scripts that consume infinite time and RAM
software installation gong shows

What if …
I actually taught that?!

Step 1:
get better at it myself!

UBC MASTER OF DATA SCIENCE

Professional Masters degree
10-months full-time
24 1-credit course modules
6-credit Capstone Project
Collaborative effort by STAT & CS (Faculty of Science)

14/30 credits

STAT	Descriptive Statistics and Probability
	Statistical Inference and Computation I
	Statistical Inference and Computation II
	Regression I
	Regression II
	Spatial and Temporal Models
	Experimentation and Causal Inference
CS	Algorithms and Data Structures
	Databases and Data Retrieval
STAT/CS	Supervised Learning I
	Supervised Learning II
	Unsupervised Learning
	Feature and Model Selection
	Advanced Machine Learning

16/30 credits

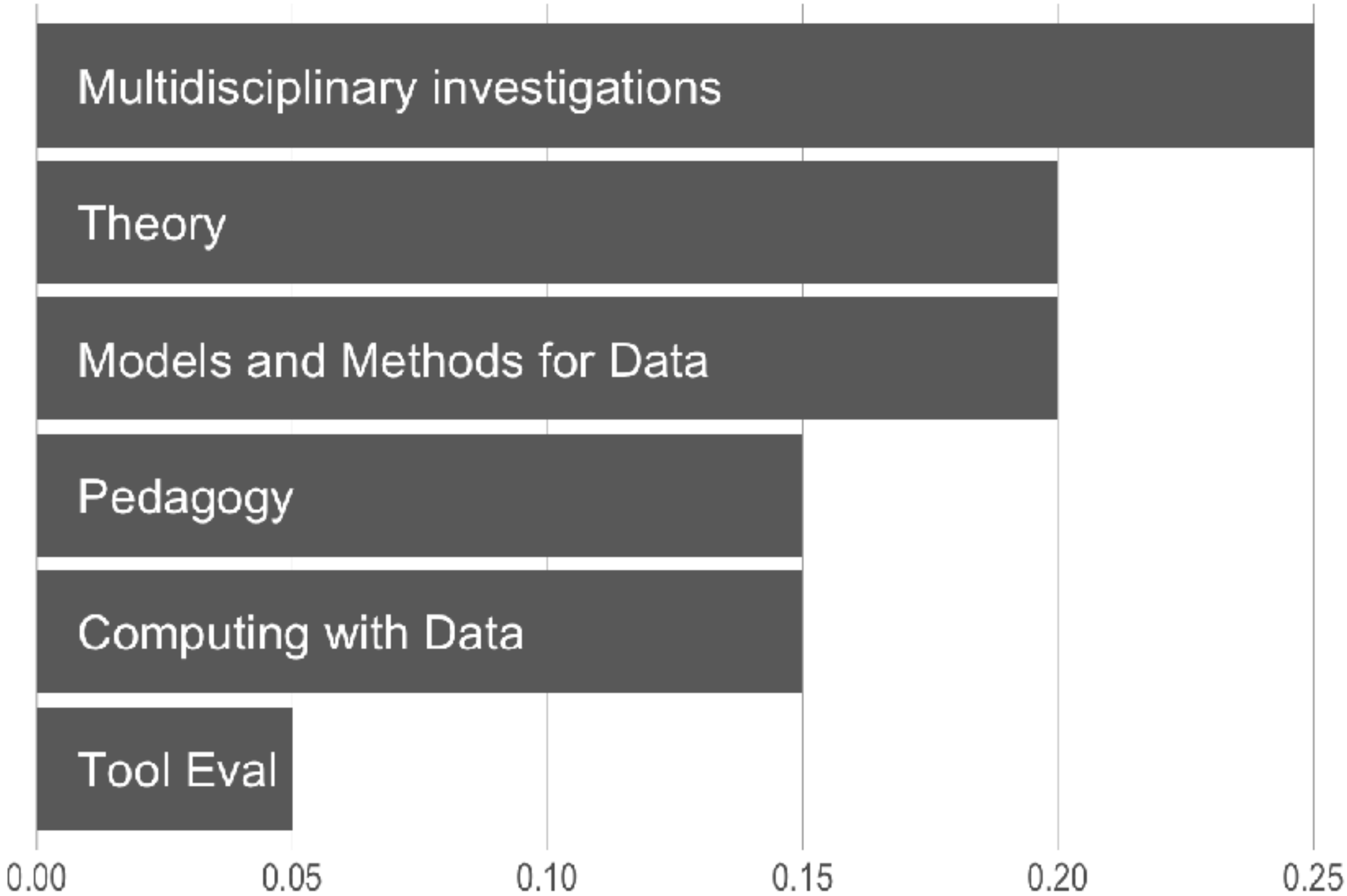
~~STAT~~

~~CS~~

DS

Programming for Data Science
Computing Platforms for Data Science
Data Science Workflows
Collaborative Software Development
Web and Cloud Computing
Data Wrangling
Data Visualization I
Data Visualization II
Privacy, Ethics, and Security
Communication and Argumentation
Capstone Project

Cleveland: proposed resource allocation



via Donoho's 50 years of Data Science

"We don't have to teach data science, it's just a fancy word for statistics"

"Why would we teach programming? This is a statistics course"

Tweet by David Robinson @drob

pick **one**:

☐ data science is ‘just’ statistics

☐ data wrangling is not statistics

programming

version control

visualization

testing

web apps

...

pick, at most, **one**:

☐ data science is ‘just’ statistics

☐ data wrangling is not statistics

programming

version control

visualization

testing

web apps

...

We can say
"data science is just statistics"
if and only if

we broaden the
definition of "statistics".

Journal
Journal of Computational and Graphical Statistics ›
Volume 26, 2017 - Issue 4

Free access

Discussion

Data Science: A Three Ring Circus or a Big Tent?

Jennifer Bryan & Hadley Wickham

Pages 784-785 | Published online: 19 Dec 2017

https://doi.org/10.1080/10618600.2017.1389743

https://doi.org/10.1080/10618600.2017.1389743

Is data science just a trendy term for statistics?

No.

Users Become Developers

- Good Programming Practice, by Martin Mächler
- Language Interfaces (.Call & .External), by Peter Dalgaard
- Packaging, Documentation, Testing, by Kurt Hornik

This talk is . . .

- *not* a one or two days' course (from Insightful or . . .)
- *not* systematic and comprehensive like a *book* such as
  Chambers "Programming with Data" (1998),
  Venables + Ripley "S Programming" (2000),
  Uwe Ligges "R Programmierung" (2004) [in German]
- *not* for complete newbies
- *not* really for experts either
- *not* about C (or Fortran or C++ . . .) programming
- *not* always entirely serious ☺

from Mächler's talk

R has changed since you graduated

http://kbroman.org/hipsteR/

hipsteR re-educating people who learned R before it was cool

… my knowledge of R seems stuck in 2001. I keep finding out about “new” R functions (like replicate, which was new in 2003).

This is a tutorial for people like me, or people who were taught by people like me.

Import

readr
readxl
haven
xml2

Tidy

tibble
tidyr

Transform

dplyr
forcats
hms
lubridate
stringr

Visualise

ggplot2

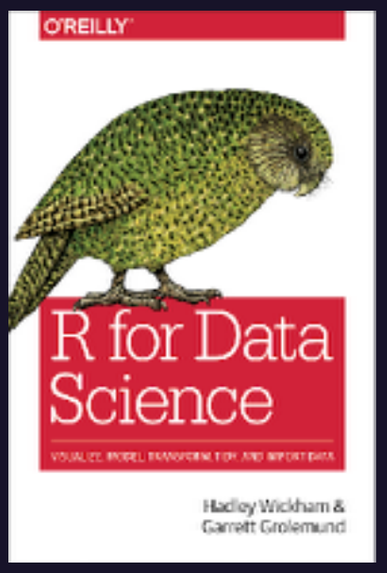
Model

rsample
tidyposterior
yardstick
broom
modelr
recipes

purrr
magrittr

Program

tidyverse.org



r4ds.had.co.nz

Use an IDE

Integrated Development Environment

RStudio

Emacs + ESS

vim + Nvim-R

…

source is real

“The source code is real. The objects are realizations of the source code. Source for EVERY user modified object is placed in a particular directory or directories, for later editing and retrieval.”

-- from the ESS manual

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 fraiche
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 (50-25 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 30ml 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.

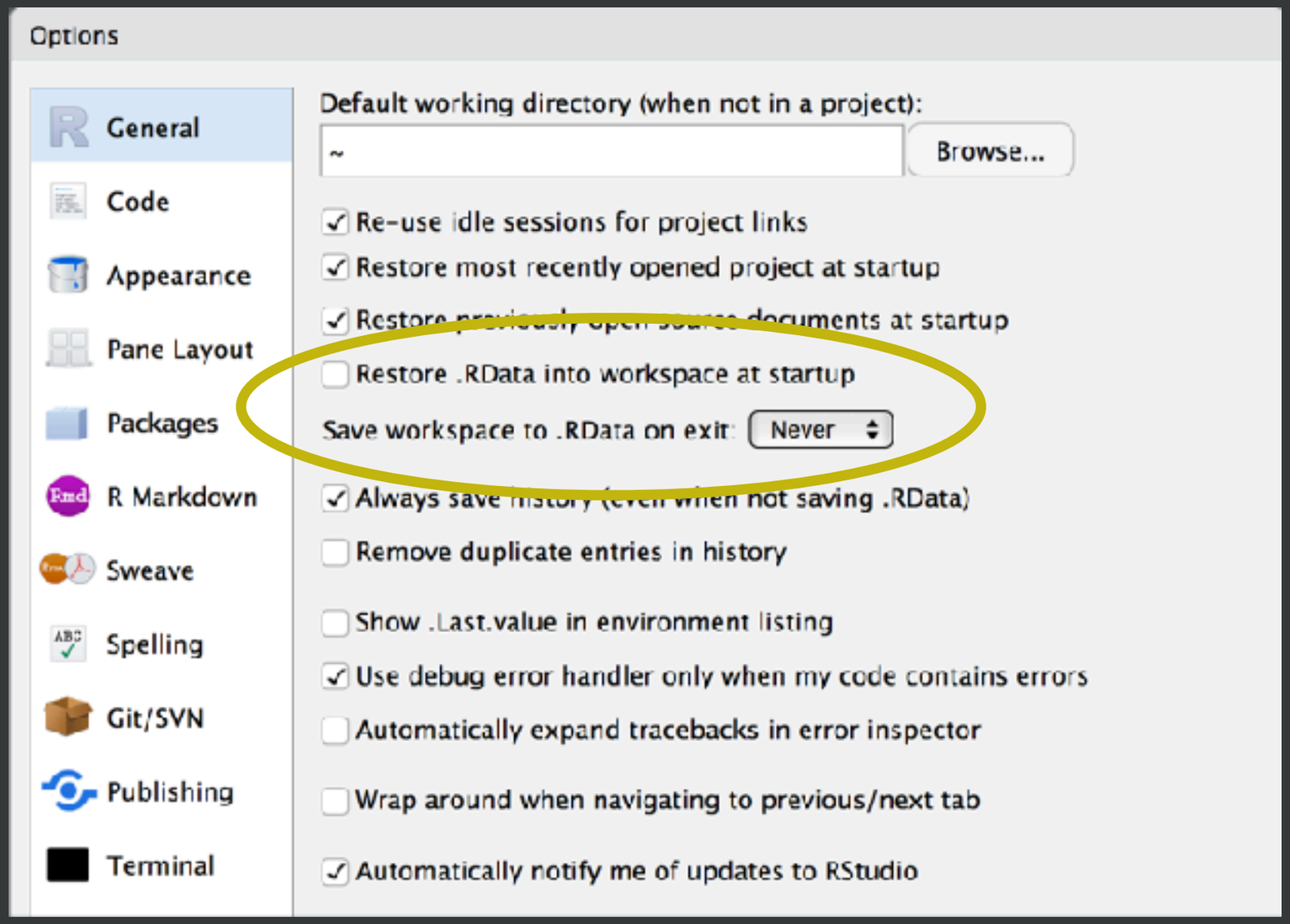
4. Meanwhile, combine crème fraiche, milk, rind, icing sugar and remaining vanilla seeds in a bowl, adding extra milk if necessary until spreadable. Spread one-third over set jelly, top with a sponge round, fill any gaps with trimmings, drizzle with 125ml crème de mûre. Scatter over remaining berries, 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.

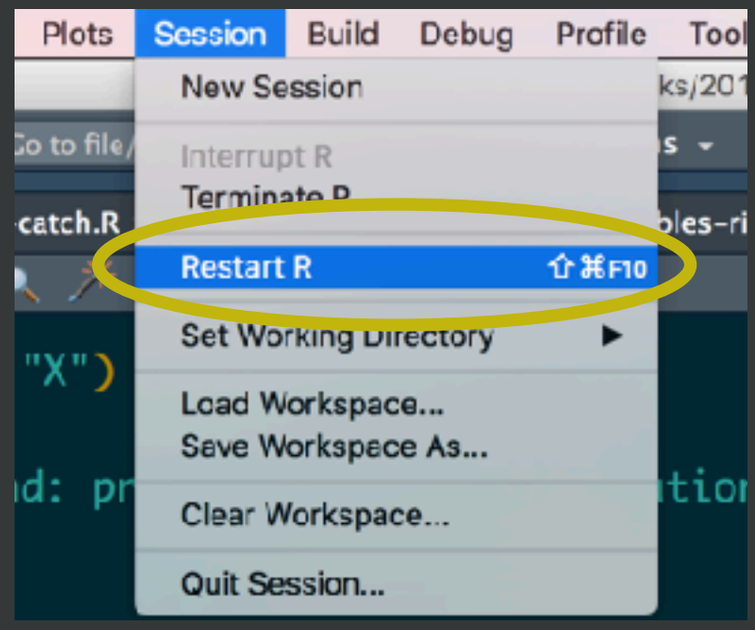
If the first line of your R script is

```
rm(list = ls())
```

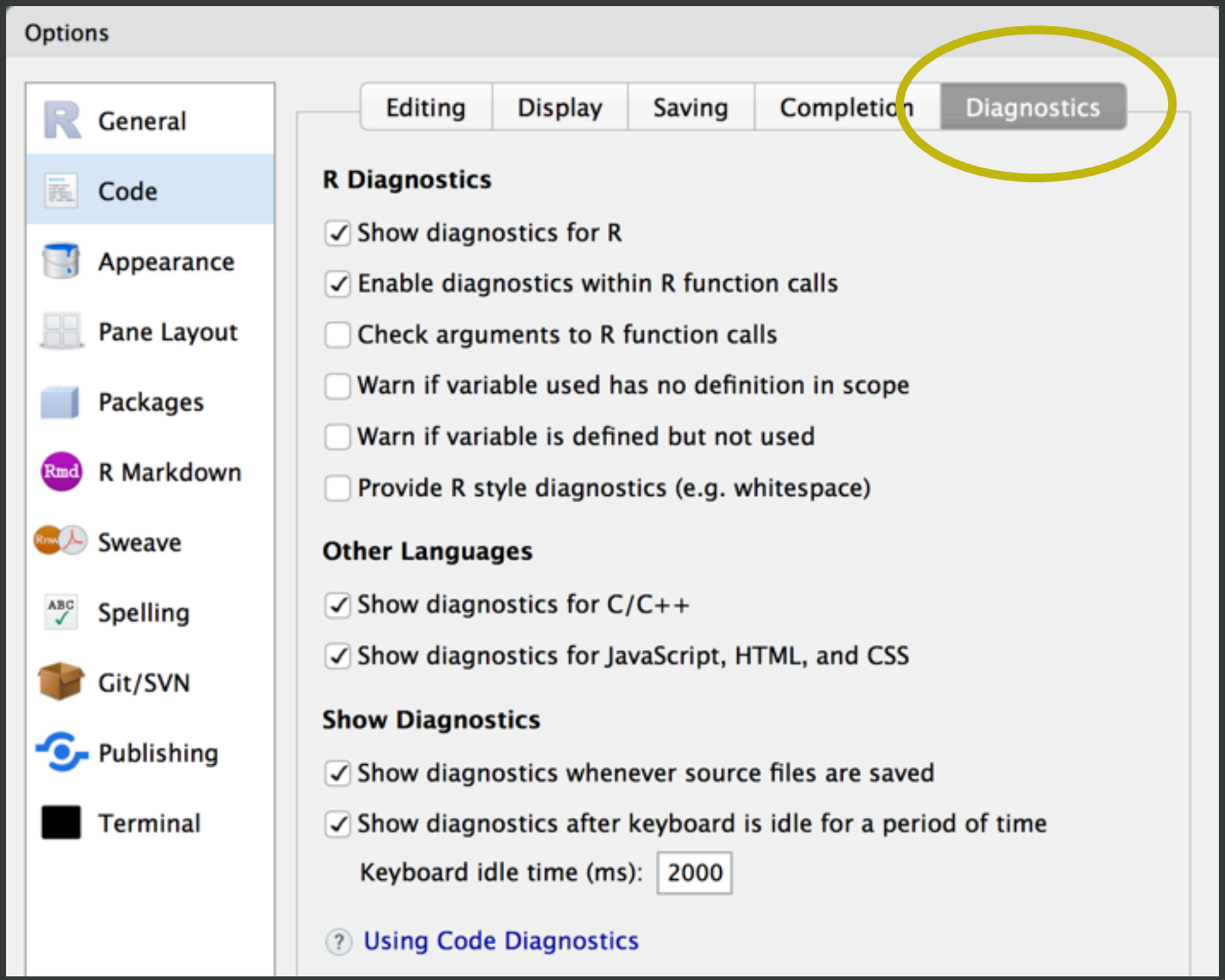
I will come into your office and

SET YOUR COMPUTER ON FIRE 🔥.





Restart R with a clean slate OFTEN,
e.g., multiple times per day



Accept help re: missing)'s or errant ,s

use [Pp]rojects

Emily Riederer
@EmilyRiederer

What #rstats tricks did it take you way too long to learn?

Jesse Maegan @kierisi · Aug 19

I cannot retweet this enough - it took me ages to use Projects and now I can't believe I did it any other way!

> **Emily Riederer** @EmilyRiederer
>
> Replying to @EmilyRiederer
>
> And using Projects in RStudio

Jenny Bryan
@JennyBryan

Replying to @kierisi

what made you reluctant? were you worried it would be complicated? curious because I see the reluctance, then joy/relief, often in @STAT545

Jesse Maegan
@kierisi

Follow

Replying to @JennyBryan @STAT545

I couldn't find a good explanation on what Projects *did* & why it was better than manually saving all my scripts in the same folder.


rexcel
blarg
jenny
eigencoder
foofy
STAT545-
happy-git-with-

One folder per project

That folder is an

- RStudio Project (package? website? whatever)
- Or similar implementation in your IDE of choice
- Git repo, with associated GitHub remote

Work on multiple projects at once w/ multiple instances of RStudio (or other IDE)

- Each gets own child R process
- R & file browser have sane working directory

use portable

file paths

If the first line of your R script is

```
setwd("C:\Users\jenny\path\that\only\I\have")
```

I* will come into your office and
SET YOUR COMPUTER ON FIRE 🔥.

* or maybe Timothée Poisot will

Blog post: Project-oriented workflow

Don’t rush into a complicated folder hierarchy

Build paths relative to project’s top-level folder

But once you need sub-folders …

Use the **here** package to build paths

```
install.packages("here")
```

```
ggsave(here("figs", “cleveland-alloc.png”))
```

Works on my machine, works on yours!

Works even if working directory is in a sub-folder

Works for RStudio projects, Git repos, R packages, …

Works with knitr / rmarkdown

expect to

iterate

Trevor A. Branch
@TrevorABranch

Follow

My rule of thumb: every analysis you do on a dataset will have to be redone 10–15 times before publication. Plan accordingly. #Rstats

Reason to iterate #1:

Get it right!

New data?

New understanding of data?

Reason to iterate #2:

Refine and Extend

Make your code more

Readable

Efficient

Resilient

General

beware of monoliths

break logic & output into pieces

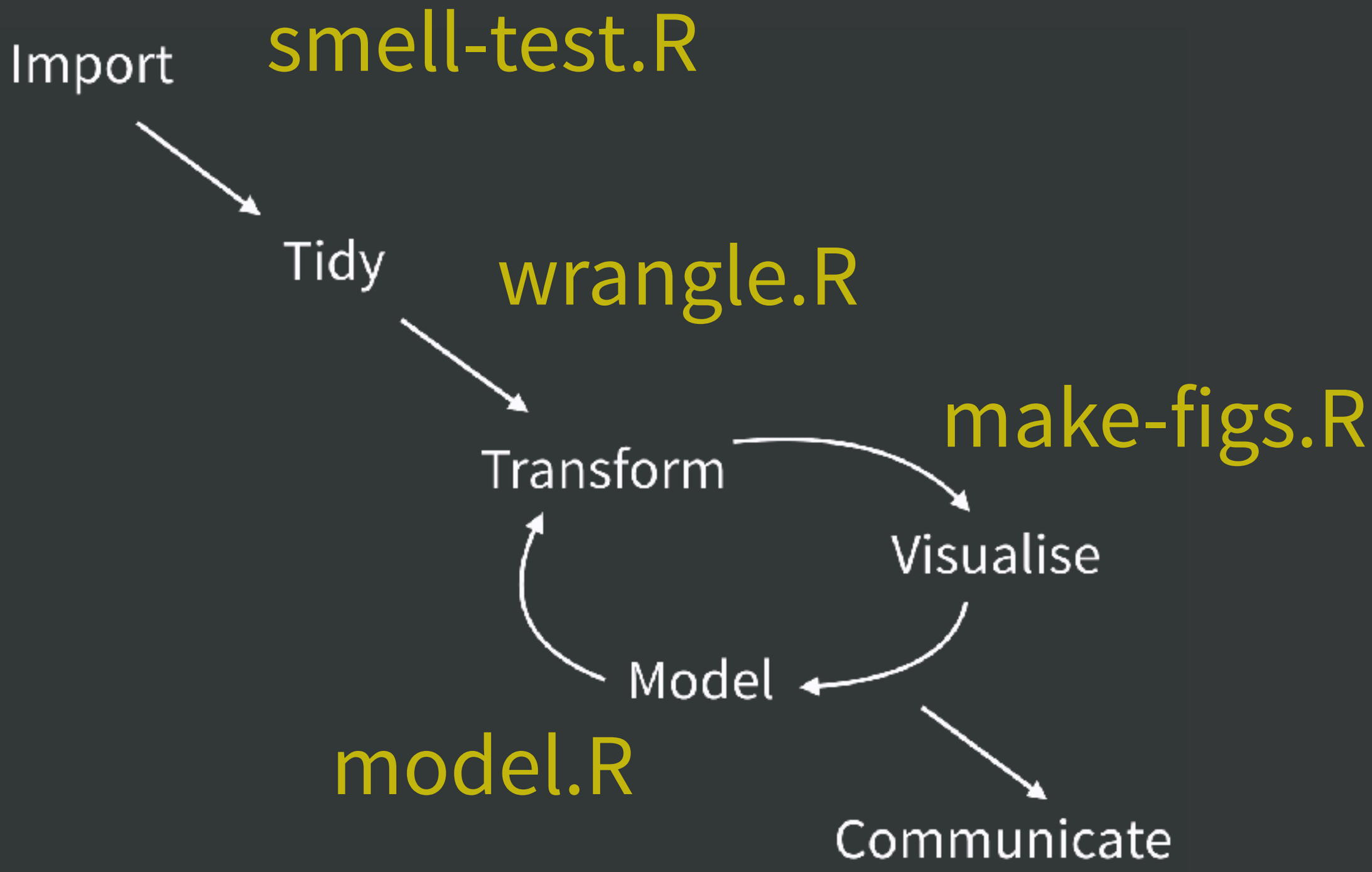
smell-test.R

wrangle.R

model.R >>> everything.R

make-figs.R

report.Rmd


Import
smell-test.R
Tidy
wrangle.R
Transform
make-figs.R
Visualise
Model
model.R
Communicate
report.Rmd

raw-data.xlsx

data.csv

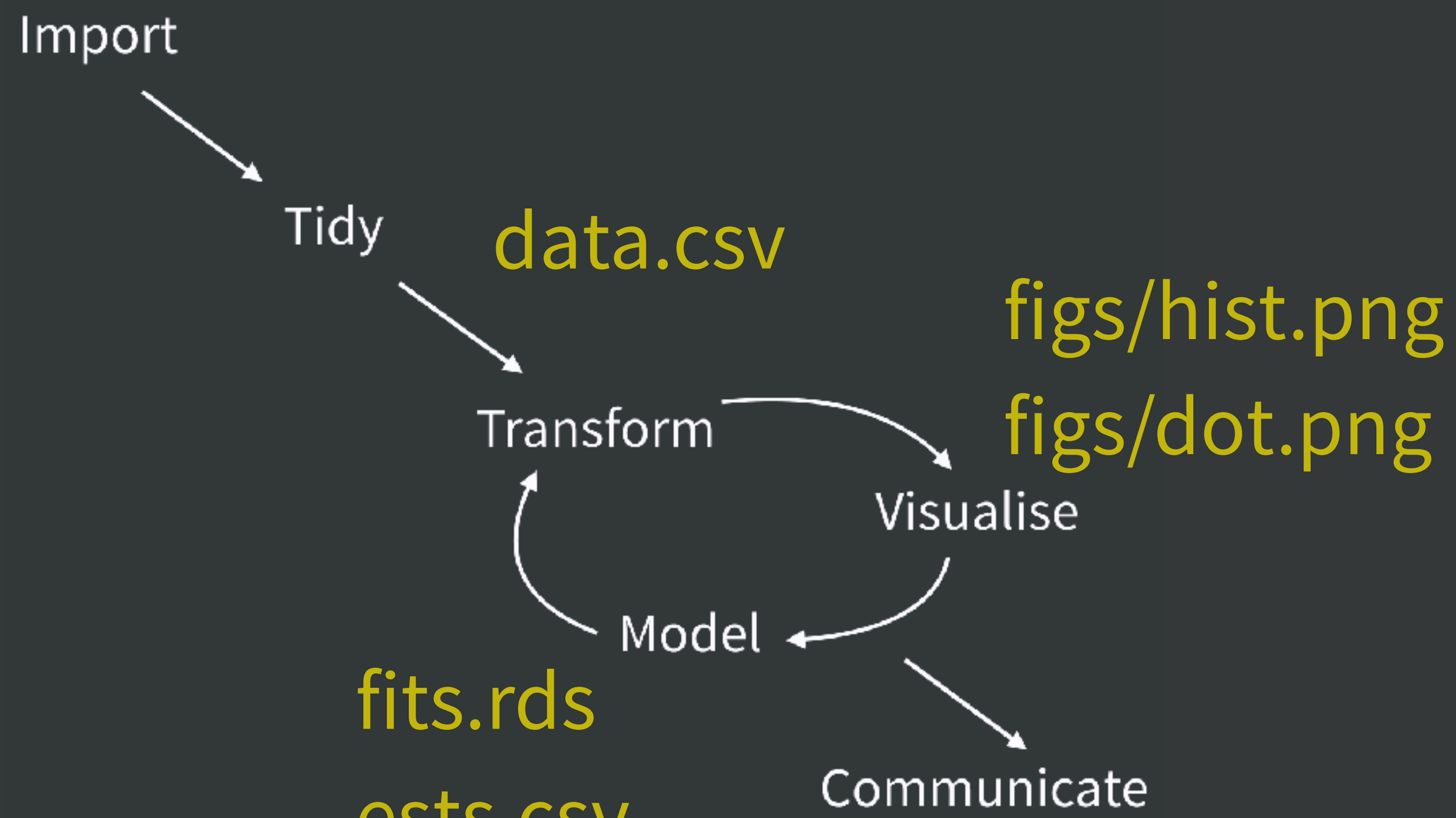
fits.rds

ests.csv

>>>

.Rdata

raw-data.xlsx



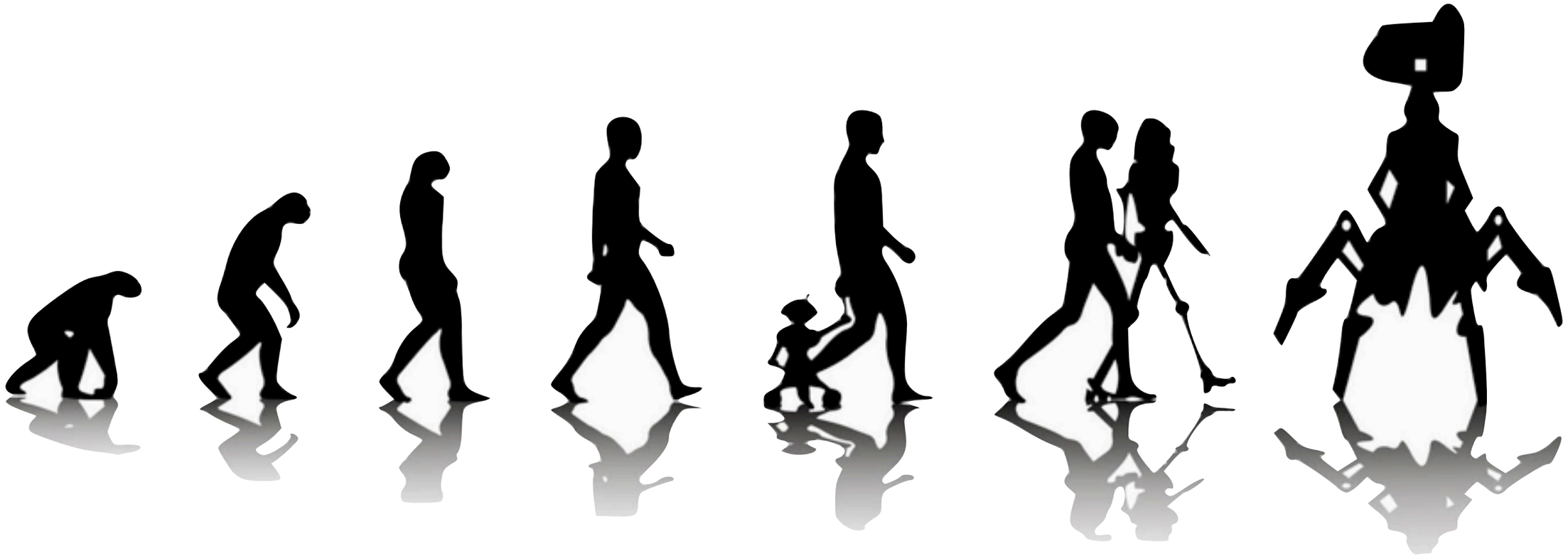
Input	Code	Output
raw data	`smell-test.R`	*wisdom*
raw data	`wrangle.R`	`data.csv`
`data.csv`	`model.R`	`fits.rds` `ests.csv`

`data.csv`
`fits.rds`
`ests.csv`
	`make-figs.R`	`figs/*`

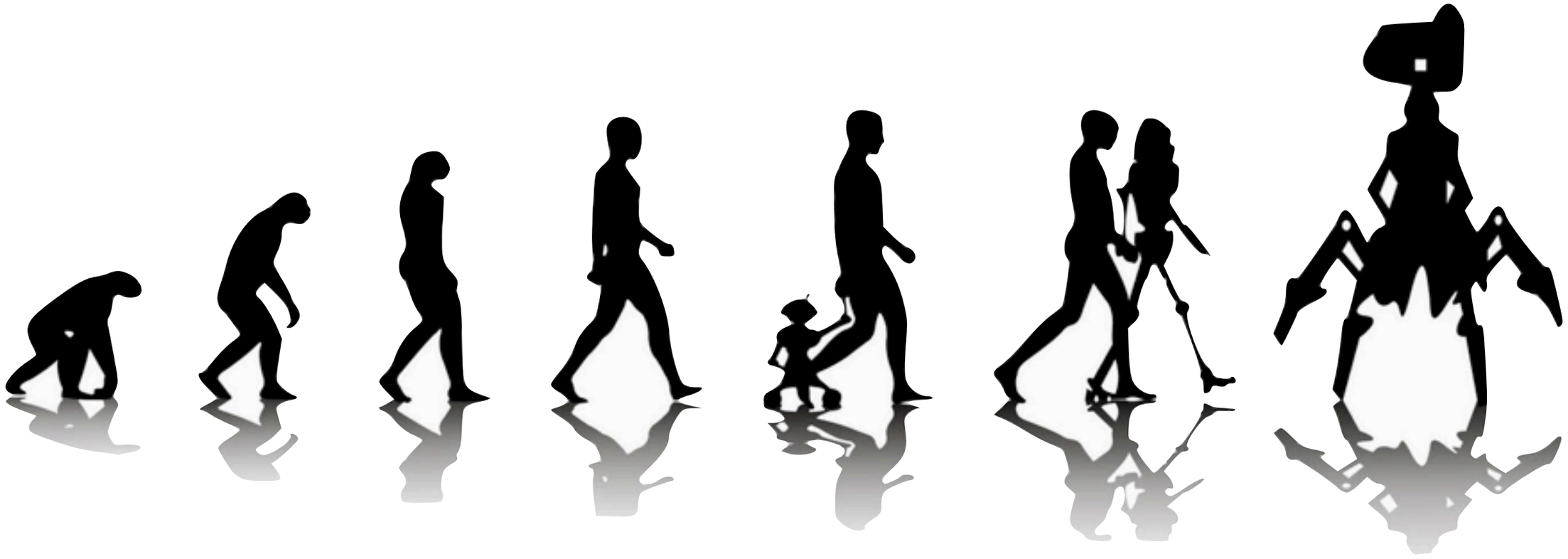
`figs/*`
`ests.csv`
	`report.Rmd`	`report.html`
		`report.docx`
		`report.pdf`

a humane API

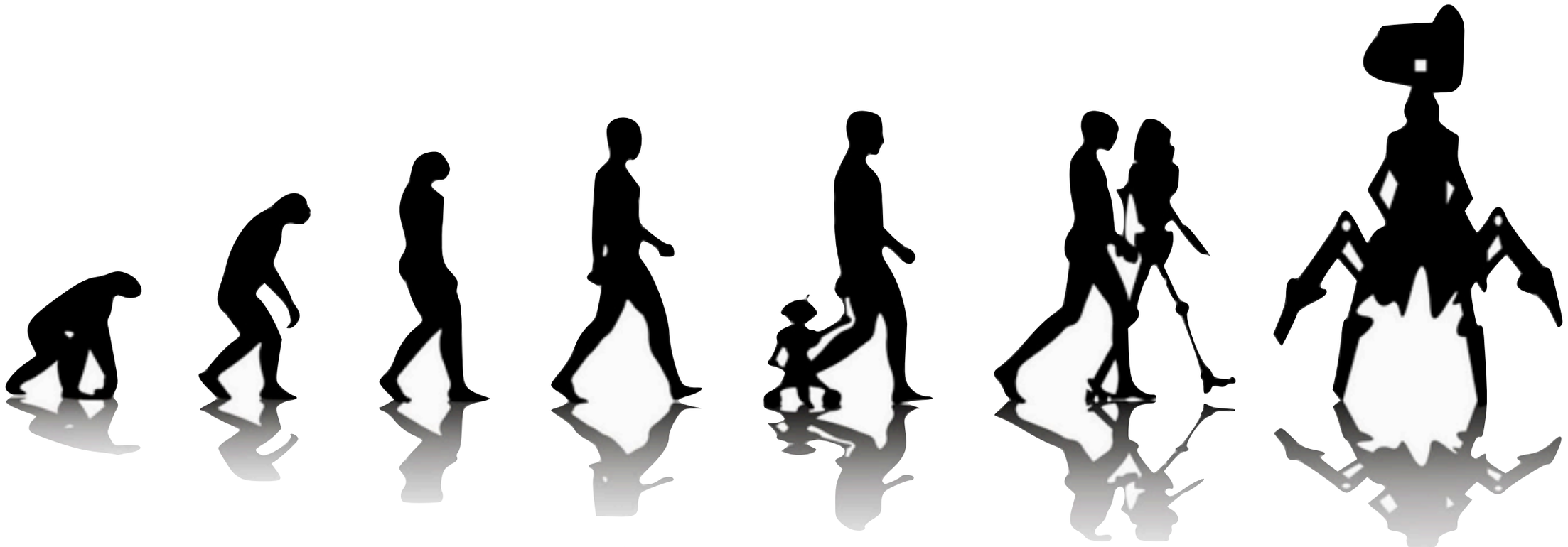
for your analysis



consider version control



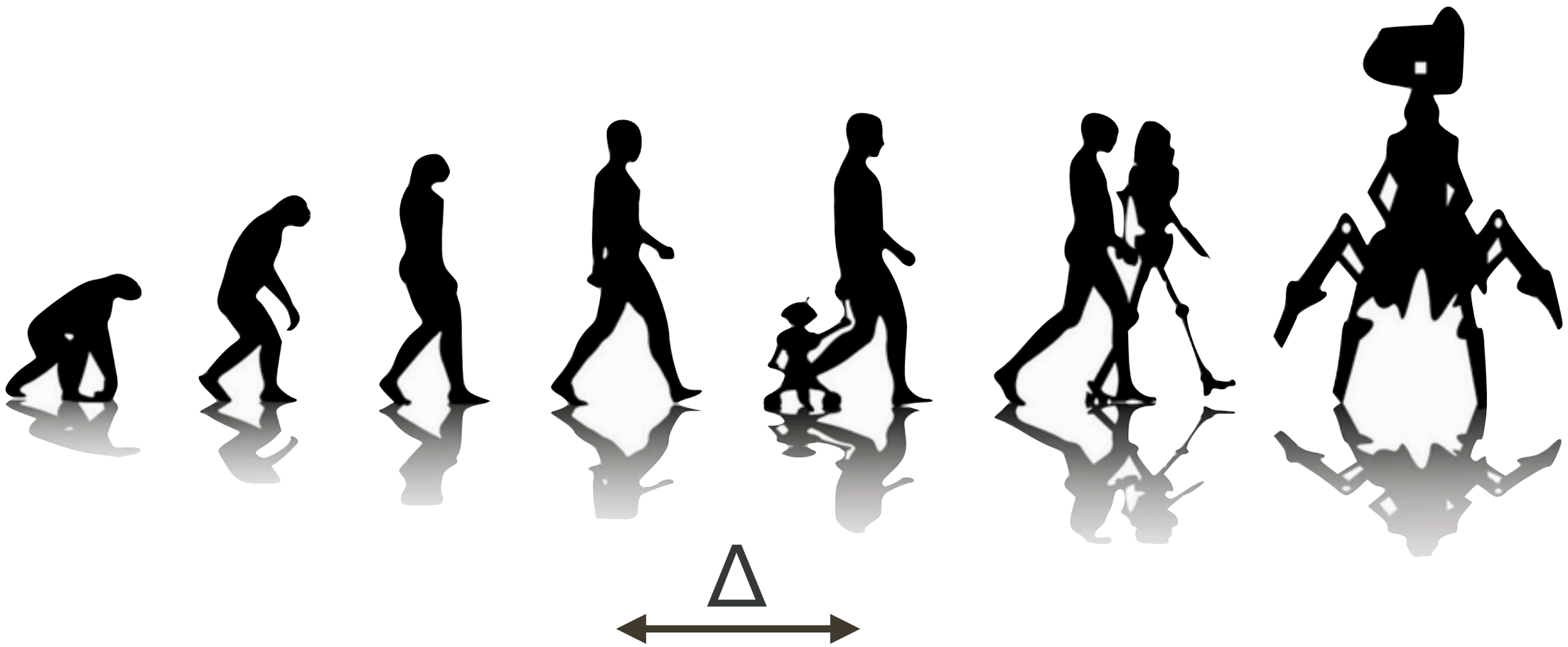
I use Git + GitHub



“commit”

a file or project state that is **meaningful to you**

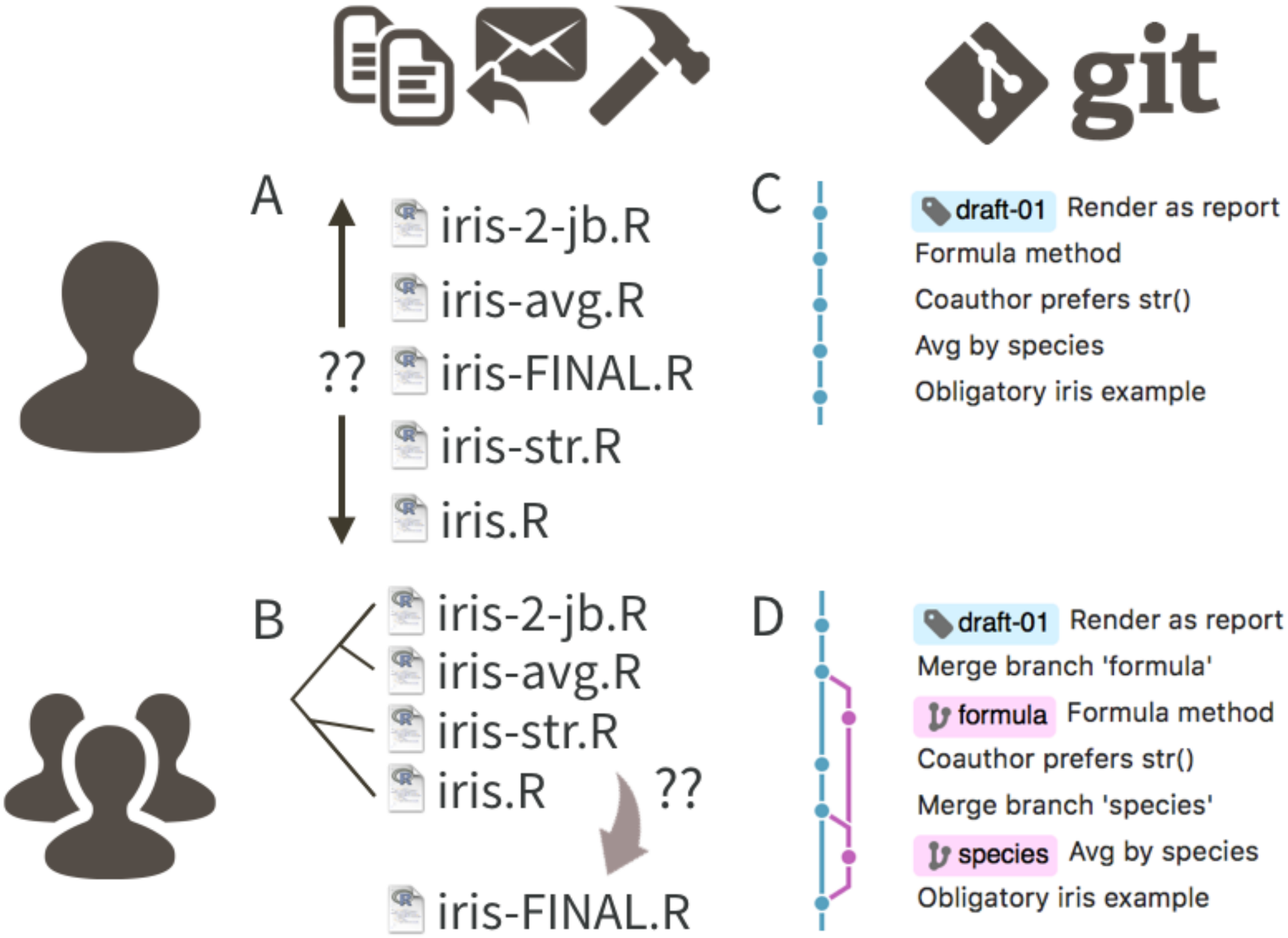
for inspection, comparison, restoration

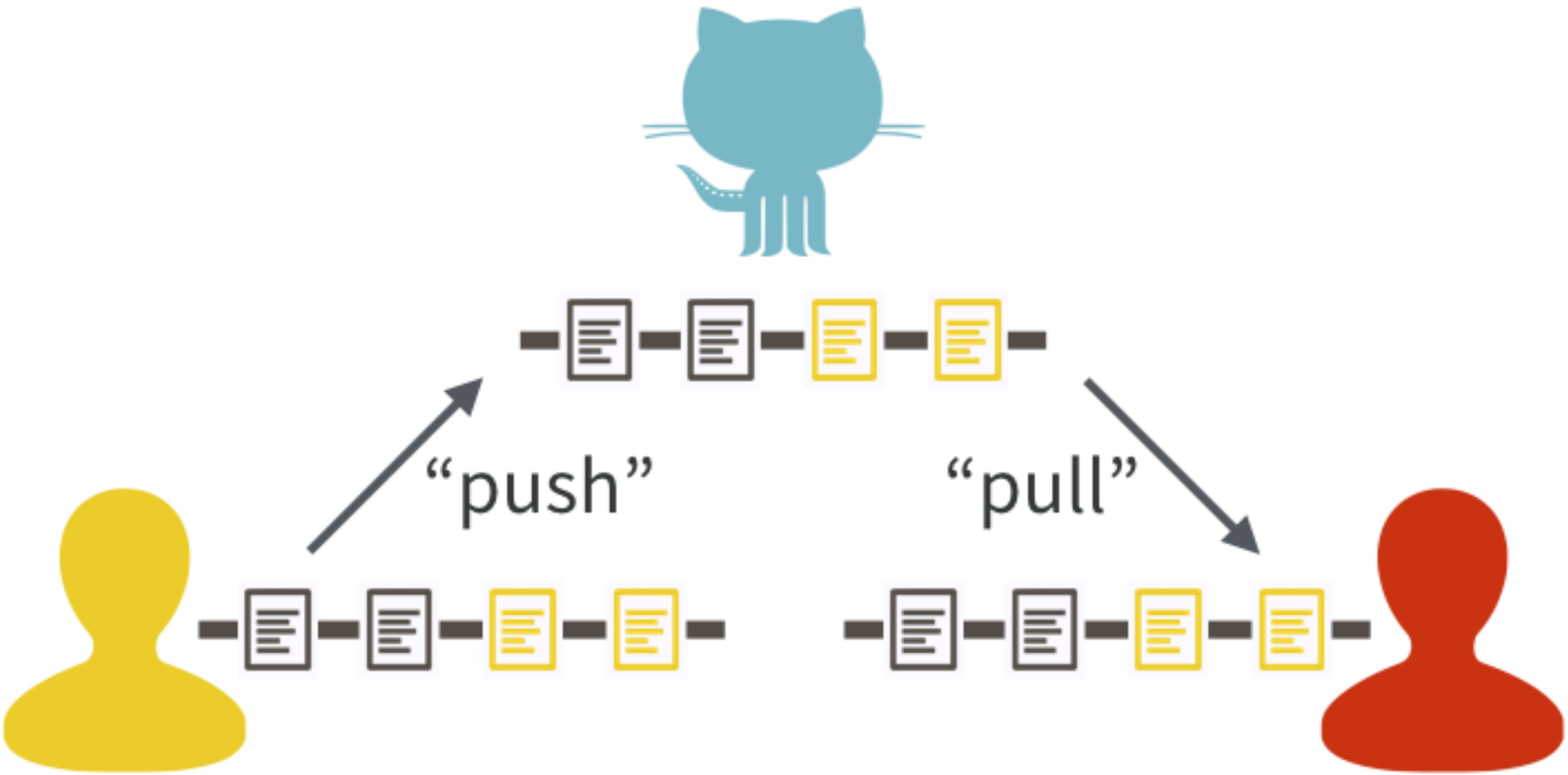


“diff”

What changed here?

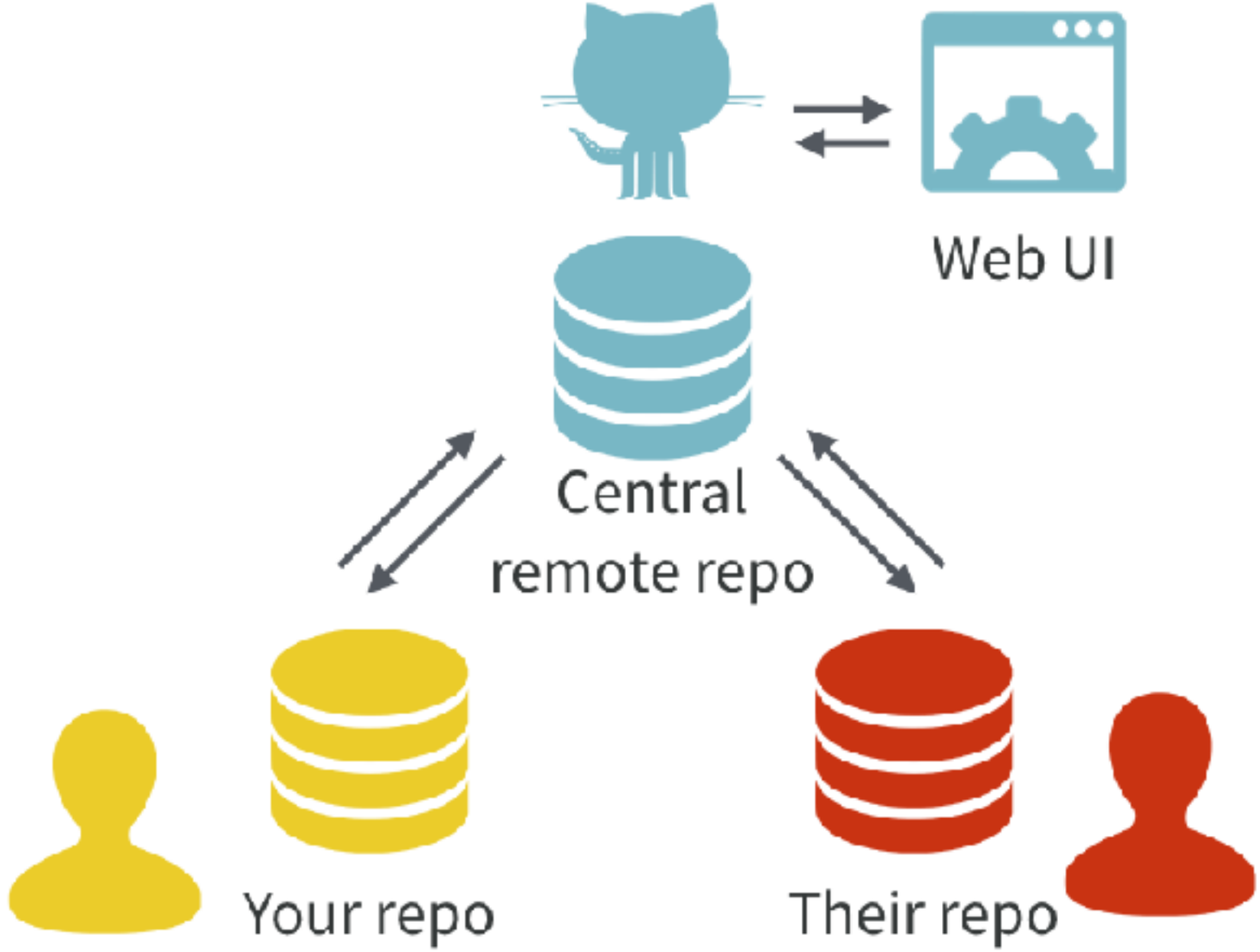
Why?


A
iris-2-jb.R
iris-avg.R
??
iris-FINAL.R
iris-str.R
iris.R
B
iris-2-jb.R
iris-avg.R
iris-str.R
iris.R
??
iris-FINAL.R
C
draft-01 Render as report
Formula method
Coauthor prefers str()
Avg by species
Obligatory iris example
D
draft-01 Render as report
Merge branch 'formula'
formula Formula method
Coauthor prefers str()
Merge branch 'species'
species Avg by species
Obligatory iris example


“push”
“pull”

Excuse me, do you have a moment to talk about version control?

https://doi.org/10.7287/peerj.preprints.3159v2



happygitwithr.com

Practical Data Science for Stats

http://bit.ly/practical-data-sci

Good enough practices in scientific computing

Wilson, Bryan, Cranston, Kitzes, Nederbragt, Teal

https://doi.org/10.1371/journal.pcbi.1005510

http://bit.ly/good-enuff

@JennyBryan
@jennybc

Thanks:

Mara Averick

Matthew Lincoln

Hadley Wickham

STAT 545 TAs

UBC MDS Fellows & Faculty