PSYC 259:
Principles of Data Science

Week 1: Course Goals

Today

1. Introductions
2. Course overview
3. Workflow examples
4. Logistics
5. Course focus
6. Software setup

Introductions

About me

- Prof. John Franchak (he/him/his)
  - Please call me John
  - Office hours by appointment (zoom link in syllabus)
- Research:
  - Development of perception and motor control
  - Wide range of data: eye tracking, motion tracking, video/image analysis, behavioral measures, EMA, psychophysics
  - Program research tools: Shiny apps, Matlab region-of-interest software, SMS scheduling

About me

- My programming background
  - First programs: TI-83+ graphing calculator games
  - C++, visual basic in high school
  - 2 semesters of C++/Java in college
  - Learned Matlab on my own as a lab manager, mostly used Matlab and SPSS as a grad student
  - Learned R as an Assistant Prof at UCR
  - Learned Julia as an Associate Prof at UCR

About you

- Name
- Program/lab
- Types of data you work with
- What you're hoping to learn/improve on

Course Overview

What is data science?

> "The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades."
>
> - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics [3]

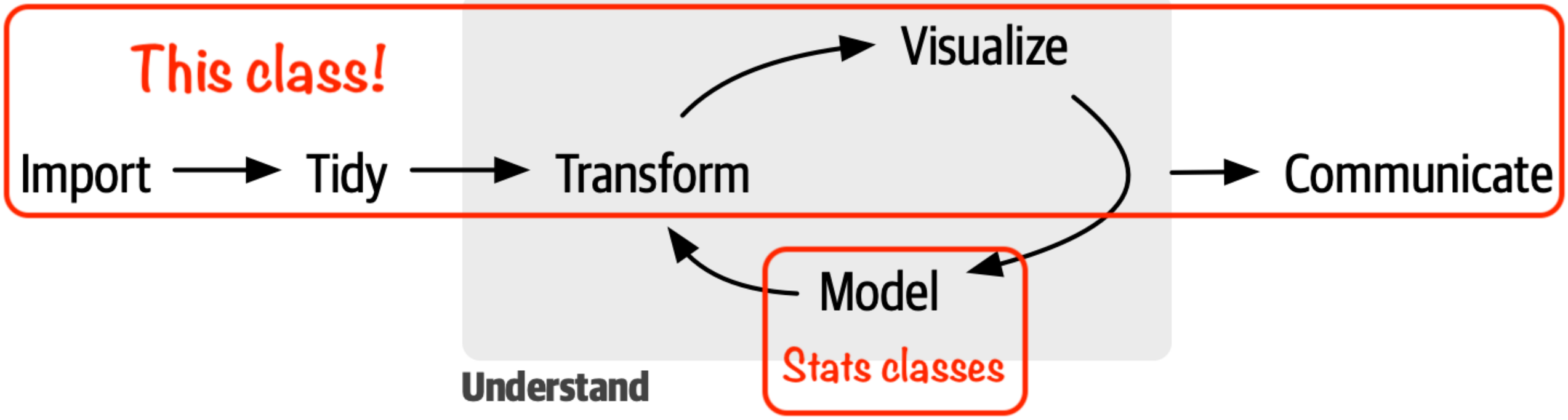
Why this class?

- Don’t we already teach students to understand, extract value, and communicate about data?

Why this class?

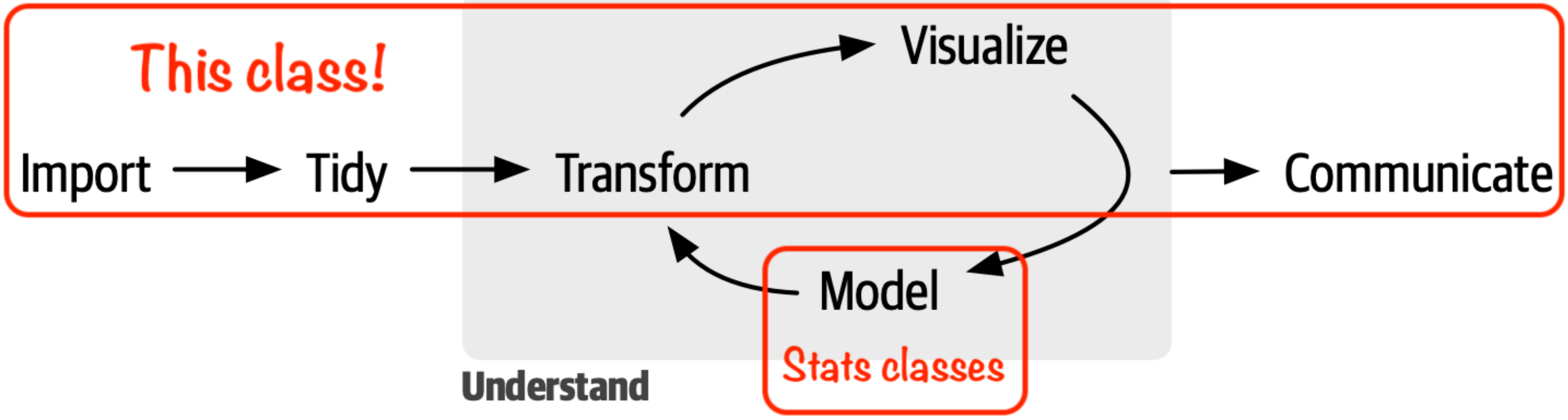
- Don’t we already teach students to understand, extract value, and communicate about data?
- Our statistics classes teach data *analysis* - testing hypotheses, modeling, etc.
- Understanding data requires the art of data *processing* and data *exploration*

Course focus: Programming for data exploration & communication



source: *R for Data Science*

Course focus: Programming for data exploration & communication

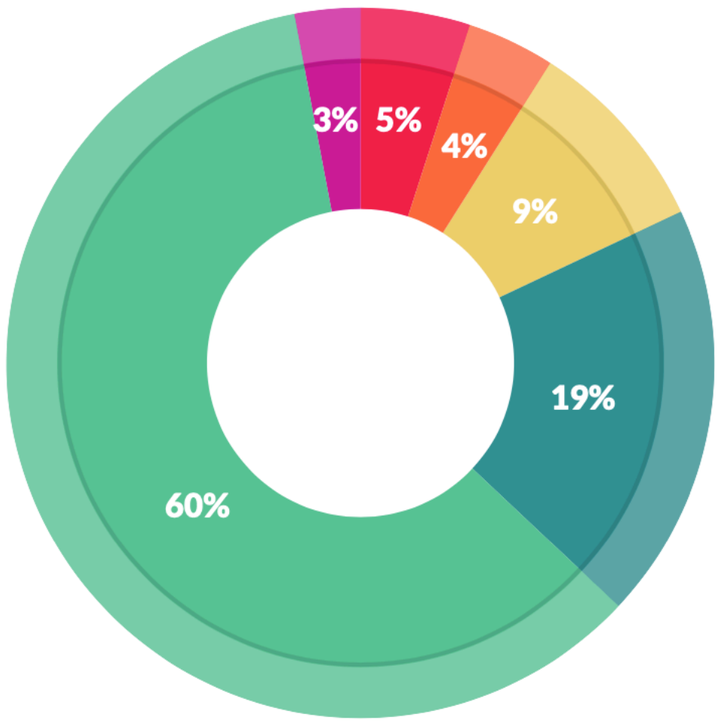


80/20 rule of data science: 80% of the work is getting the data ready to analyze, only 20% of the work is analyzing/reporting

source: *R for Data Science*

How a Data Scientist Spends Their Day

Here's where the popular view of data scientists diverges pretty significantly from reality. Generally, we think of data scientists building algorithms, exploring data, and doing predictive analysis. That's actually not what they spend most of their time doing, however.



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

source: ***Crowdflower Data Science Report (2016)***

You will learn to critically analyze and improve your data analysis workflows. Robust, automated procedures for handling data will:

- Foster open science through increased transparency, reproducibility, and easier data sharing
- Increase the fidelity of your data and analyses by detecting and preventing errors
- Better understand and communicate about your data
- Save time by preventing errors, automating tasks, and reusing code
- Facilitate collaboration with organized and documented workflows

Don’t be “the gift that keeps on giving”

Alex Naka @gottapatchemall · Dec 3

Looking at some old code and was initially puzzled by a variable named 'feet'

I have now worked out that this was at one point called 'legend_handles', which then became 'leg_hands', which then became 'feet'

sometimes I truly hate my past self

78 1.2K 8.9K

Don’t let software make assumptions about your data

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Gene name errors: Lessons not learned

Mandhri Abeysooriya, Megan Soria, Mary Sravya Kasu, Mark Ziemann*

Deakin University, School of Life and Environmental Sciences, Geelong, Australia

* m.ziemann@deakin.edu.au

Abstract

Erroneous conversion of gene names into other dates and other data types has been a frustration for computational biologists for years. We hypothesized that such errors in supplementary files might diminish after a report in 2016 highlighting the extent of the problem. To assess this, we performed a scan of supplementary files published in PubMed Central from 2014 to 2020. Overall, gene name errors continued to accumulate unabated in the period after 2016. An improved scanning software we developed identified gene name errors in 30.9% (3,436/11,117) of articles with supplementary Excel gene lists; a figure significantly higher than previously estimated. This is due to gene names being converted not just to dates and floating-point numbers, but also to internal date format (five-digit numbers). These findings further reinforce that spreadsheets are ill-suited to use with large genomic data.

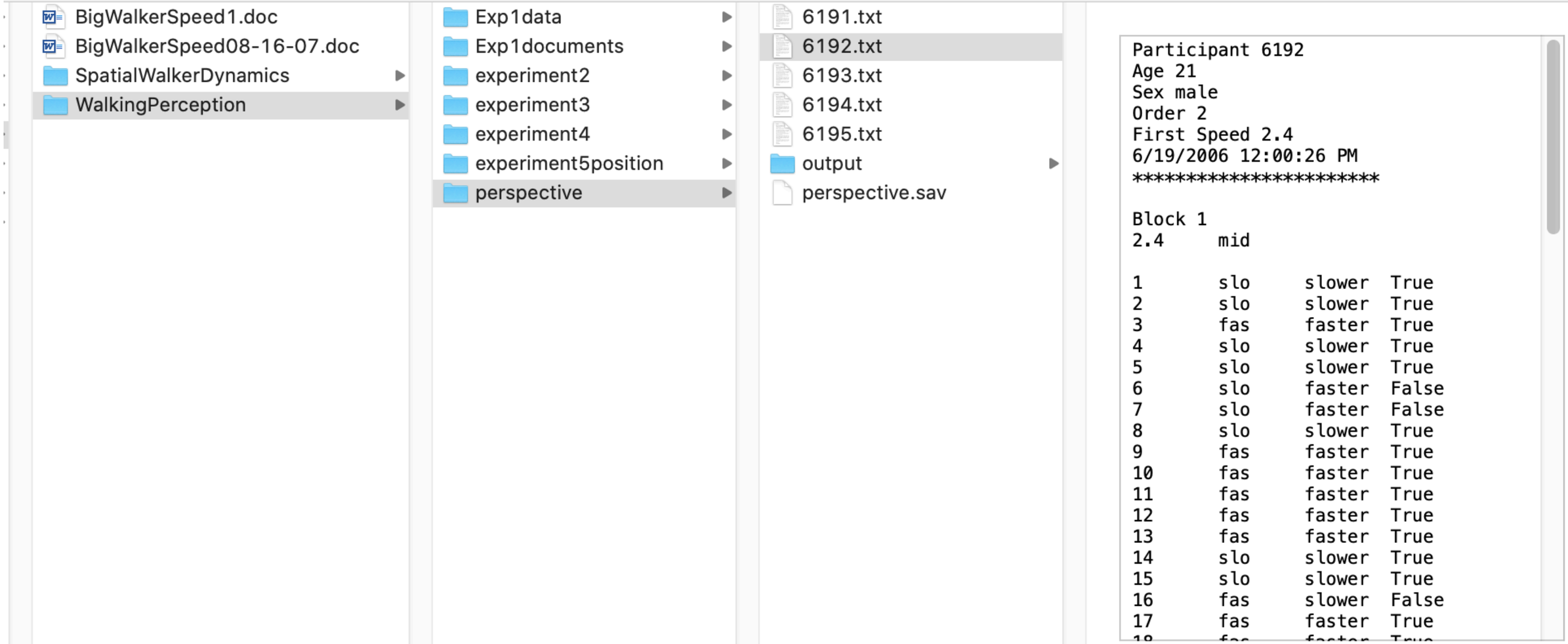
	A	B
1	SEPT8	8-Sep
2	DEC1	1-Dec
3	MARCH3	3-Mar
4		
5		

Citation: Abeysooriya M, Soria M, Kasu MS, Ziemann M (2021) Gene name errors: Lessons not learned. PLoS Comput Biol 17(7): e1008984. https://doi.org/10.1371/journal.pcbi.1008984

Workflow examples

Example 1
It was the year 2006…

- A young lab manager was tasked with testing how people walking on a treadmill perceived their own walking speed compared to life-sized avatars walking alongside them on a projection screen
- Software was written
- Results were analyzed
- A poster was presented
- The files sat dormant in storage for over a decade, never seeing the light of day….until now!


BigWalkerSpeed1.doc
BigWalkerSpeed08-16-07.doc
SpatialWalkerDynamics
WalkingPerception
Exp1data
Exp1documents
experiment2
experiment3
experiment4
experiment5position
perspective
6191.txt
6192.txt
6193.txt
6194.txt
6195.txt
output
perspective.sav
Participant 6192
Age 21
Sex male
Order 2
First Speed 2.4
6/19/2006 12:00:26 PM
****
Block 1
2.4 mid
1 slo slower True
2 slo slower True
3 fas faster True
4 slo slower True
5 slo slower True
6 slo faster False
7 slo faster False
8 slo slower True
9 fas faster True
10 fas faster True
11 fas faster True
12 fas faster True
13 fas faster True
14 slo slower True
15 slo slower True
16 fas slower False
17 fas faster True
6192.txt
Plain Text Document - 3 KB
Information
Created Monday, June 19, 2006 at 10:10 AM
Modified Monday, June 19, 2006 at 10:10 AM


```
Block 1
2.4     mid

1       slo     slower  True
2       slo     slower  True
3       fas     faster  True
4       slo     slower  True
5       slo     slower  True
6       slo     faster  False
7       slo     faster  False
8       slo     slower  True
9       fas     faster  True
10      fas     faster  True
11      fas     faster  True
12      fas     faster  True
13      fas     faster  True
14      slo     slower  True
15      slo     slower  True
16      fas     slower  False
17      fas     faster  True
18      fas     faster  True
19      slo     slower  True
20      fas     faster  True

Correct:        17
Slower:         9

Block 2
2.7     near

1       slo     slower  True
2       slo     slower  True
3       slo     slower  True
4       slo     slower  True
```

id	age	sex	order	firstblock	near24	mid24	far24	near27	mid27	far27	near24s	mid24s	far24s	near27s	mid27s	far27s
6191	25	1	1	1	16	15	16	18	17	15	10	7	12	10	11	13
6192	21	1	2	1	18	17	18	18	19	17	8	9	12	8	11	9
6193	21	1	3	1	17	12	18	18	17	19	7	6	10	8	9	9
6194	22	1	1	2	17	17	17	20	19	17	7	7	11	10	9	11
6195	19	2	2	2	15	17	14	17	17	18	7	9	16	7	7	12
6211	19	1	3	2	20	16	13	17	16	18	10	14	13	13	8	10
6221	19	2	1	1	16	12	18	15	16	12	8	6	12	7	10	10
6222	21	2	2	1	20	13	18	19	18	18	10	13	12	9	8	12
6223	18	2	3	1	17	15	15	18	17	17	7	7	15	8	9	9
6224	18	2	1	2	19	15	17	16	19	16	11	9	13	10	11	8
6231	22	2	2	2	15	15	16	20	16	19	9	9	10	10	8	11
6232	21	1	3	2	14	14	11	14	16	15	4	8	7	6	8	13

id	age	sex	order	firstblock	near24	mid24	far24	near27	mid27	far27	near24s	mid24s	far24s	near27s	mid27s	far27s
6191	25	1	1	1	16	[illegible]	16	18	17	15	10	7	12	10	11	13
6192	21	1	2	1	18	17	18	18	19	17	8	9	12	8	11	9
6193	21	1	3	1	17	[illegible]	18	18	17	19	7	6	10	8	9	9
6194	22	1	1	2	17	17	17	20	19	17	7	7	11	10	9	11
6195	19	2	2	2	15	17	14	17	17	18	7	9	16	7	7	12
6211	19	1	3	2	20	16	13	17	16	18	10	14	13	13	8	10
6221	19	2	1	1	16	12	18	15	16	12	8	6	12	7	10	10
6222	21	2	2	1	20	13	18	19	18	18	10	13	12	9	8	12
6223	18	2	3	1	17	15	15	18	17	17	7	7	15	8	9	9
6224	18	2	1	2	19	15	17	16	19	16	11	9	13	10	11	8
6231	22	2	2	2	15	15	16	20	16	19	9	9	10	10	8	11
6232	21	1	3	2	14	14	11	14	16	15	4	8	7	6	8	13

6192.txt
6193.txt
6194.txt
6195.txt
6211.txt
6221.txt
6222.txt
6223.txt
6224.txt
6231.txt
6232.txt

```
Block 1
2.4     mid

1       slo     slower  True
2       slo     slower  True
3       fas     faster  True
4       slo     slower  True
5       slo     slower  True
6       slo     faster  False
7       slo     faster  False
8       slo     slower  True
9       fas     faster  True
10      fas     faster  True
11      fas     faster  True
12      fas     faster  True
13      fas     faster  True
14      slo     slower  True
15      slo     slower  True
16      fas     slower  False
17      fas     faster  True
18      fas     faster  True
19      slo     slower  True
20      fas     faster  True

Correct:        17
Slower:         9
```

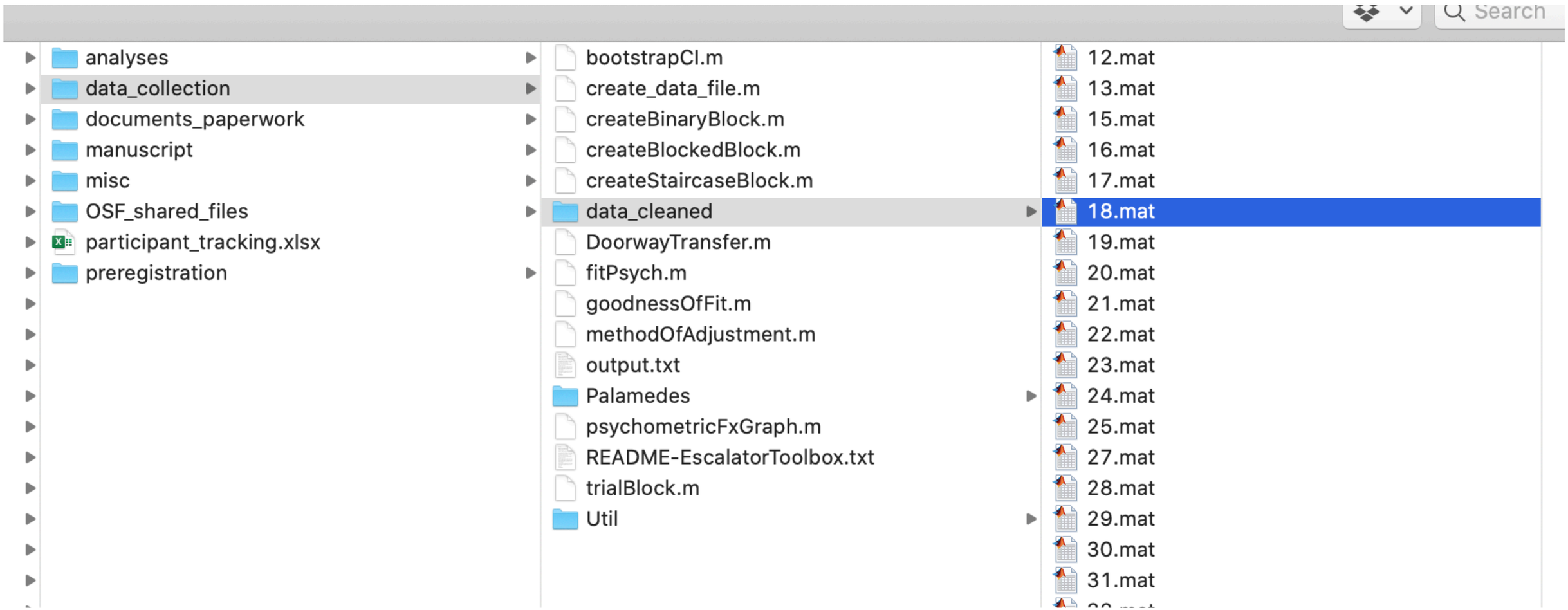
per27n	per27m	per27f	mean24	mean27	data	plspeed	plcond
90.00	85.00	75.00	78.33	83.33	80.00	2.40	near
90.00	95.00	85.00	88.33	90.00	90.00	2.40	near
90.00	85.00	95.00	78.33	90.00	85.00	2.40	near
100.00	95.00	85.00	85.00	93.33	85.00	2.40	near
85.00	85.00	90.00	76.67	86.67	75.00	2.40	near
85.00	80.00	90.00	81.67	85.00	100.00	2.40	near
75.00	80.00	60.00	76.67	71.67	80.00	2.40	near
95.00	90.00	90.00	85.00	91.67	100.00	2.40	near
90.00	85.00	85.00	78.33	86.67	85.00	2.40	near
80.00	95.00	80.00	85.00	85.00	95.00	2.40	near
100.00	80.00	95.00	76.67	91.67	75.00	2.40	near
70.00	80.00	75.00	65.00	75.00	70.00	2.40	near
.	.	.	.	.	75.00	2.40	mid
.	.	.	.	.	85.00	2.40	mid
.	.	.	.	.	60.00	2.40	mid
.	.	.	.	.	85.00	2.40	mid
.	.	.	.	.	85.00	2.40	mid
.	.	.	.	.	80.00	2.40	mid
.	.	.	.	.	60.00	2.40	mid
.	.	.	.	.	65.00	2.40	mid
.	.	.	.	.	75.00	2.40	mid
.	.	.	.	.	75.00	2.40	mid
.	.	.	.	.	75.00	2.40	mid
.	.	.	.	.	70.00	2.40	mid
.	.	.	.	.	80.00	2.40	far
.	.	.	.	.	90.00	2.40	far
.	.	.	.	.	90.00	2.40	far
.	.	.	.	.	85.00	2.40	far

Workflow problems

- Files: unorganized, inconsistently named, used unsupported proprietary formats, no version control, cluttered workspace with intermediary/redundant file
- Automation (lack of)
  - Used copy/paste to transfer data rather than automating input
  - Copied data within SPSS from wide to long
  - Made graphs/analyses from drop-down menus
- No metadata, no documentation

Example 2
The year was 2019…

- An assistant professor had a team of 5 undergrad RAs run participants in a study on perception of fitting through doorways
- The RAs entered the participant and condition info into a Matlab program, which ran the trials and wrote output data for each participant
- The RAs logged metadata into a “participant_tracking.xls” document after running each participant, leaving notes about problems with each participant


Search
analyses
data_collection
documents_paperwork
manuscript
misc
OSF_shared_files
participant_tracking.xlsx
preregistration
bootstrapCI.m
create_data_file.m
createBinaryBlock.m
createBlockedBlock.m
createStaircaseBlock.m
data_cleaned
DoorwayTransfer.m
fitPsych.m
goodnessOfFit.m
methodOfAdjustment.m
output.txt
Palamedes
psychometricFxGraph.m
README-EscalatorToolbox.txt
trialBlock.m
Util
12.mat
13.mat
15.mat
16.mat
17.mat
18.mat
19.mat
20.mat
21.mat
22.mat
23.mat
24.mat
25.mat
27.mat
28.mat
29.mat
30.mat
31.mat

Command Window

```
>> sex

sex =

  1×1 cell array

    {'m'}

>> id

id =

  1×1 cell array

    {'18'}

>> block1

block1 =

  struct with fields:

          mode: 'staircase'
    start_unit: 32
     down_step: 4
       up_step: 3
    num_trials: 16
   use_for_fit: 1

>> output_aff_f

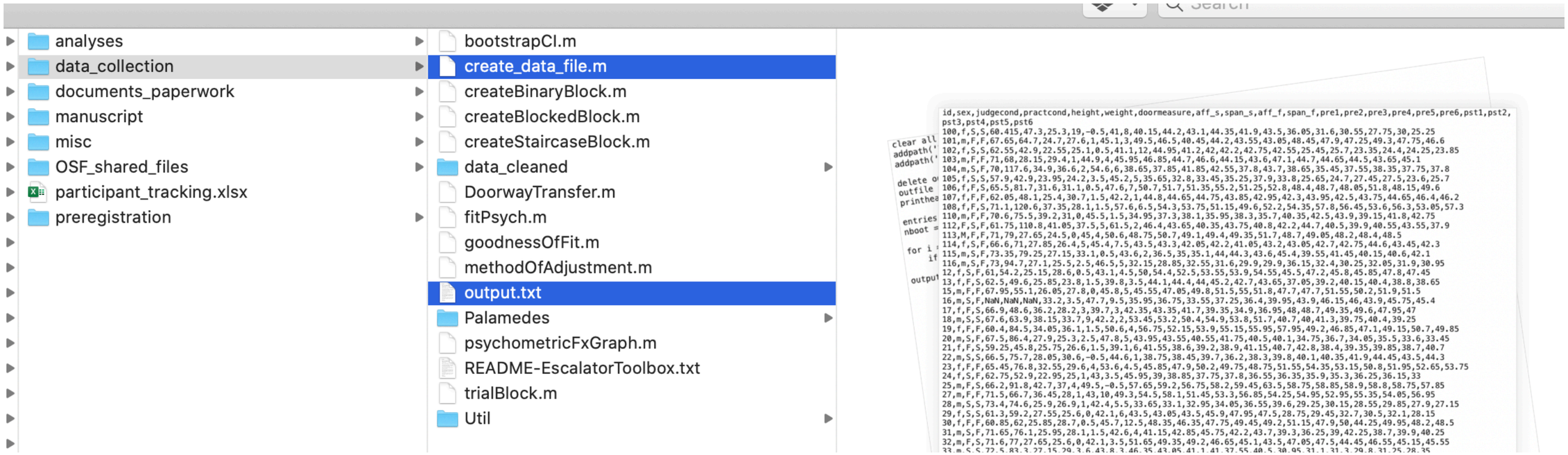
output_aff_f =

  struct with fields:

              id: '18'
       condition: 'aff_f'
       threshold: 40.8259
           slope: 1.0576
      trial_unit: [1×30 double]
  trial_unit_fit: [1×30 double]
      trial_resp: [1×30 double]
  trial_resp_fit: [1×30 double]
      trial_mode: [1×30 double]
  trial_subblock: [1×30 double]
     stim_levels: [1×107 double]
    elapsed_time: 618.8013
```

Workspace

Name ▲	Value
block1	*1x1 struct*
block2	*1x1 struct*
block_fitting	*2x1 cell*
block_squeezing	*2x1 cell*
door_measure1	*1x1 cell*
door_measure2	*1x1 cell*
height1	*1x1 cell*
height2	*1x1 cell*
id	*1x1 cell*
judgment_cond	*1x1 cell*
measurement_...	*1x1 cell*
out_dir	'C:\Users\labuse
output_aff_f	*1x1 struct*
output_aff_s	*1x1 struct*
output_pre_moa	*1x1 struct*
output_pst_moa	*1x1 struct*
practice_cond	*1x1 cell*
save_figs	1
save_file	'C:\Users\labuse
sex	*1x1 cell*
stim_levels	*1x107 double*
weight1	*1x1 cell*
weight2	*1x1 cell*



- “create_data_file.m” loads each individual file in “data_cleaned” and creates rows of data to write to “output.txt”
- “output.txt” gets copied to “analyses” for R analyses

analyses.R

Source on Save | Run | Source

```
30
31  # READ DATA ----
32
33  #Read data and set factor levels
34  ds <- read_csv("output.txt")
35  ds$id <- factor(ds$id)
36  ds$practice <- factor(ds$practcond, levels =c("S","F"),labels = c("Squeezing","Fitting"))
37  ds$judge <- factor(ds$judgecond, levels =c("S","F"),labels = c("Squeezing","Fitting"))
38  ds$match <- ds$practcond == ds$judgecond
39  ds$match <- factor(ds$match, levels = c(T, F), labels = c("Congruent","Incongruent"))
40
41  # Create error terms ====
42  ds$judge_ref[ds$judge == "Fitting"] <- ds$aff_f[ds$judge == "Fitting"]
43  ds$judge_ref[ds$judge == "Squeezing"] <- ds$aff_s[ds$judge == "Squeezing"]
44
45  ds$pre_err1 <- ds$pre1 - ds$judge_ref
46  ds$pre_err2 <- ds$pre2 - ds$judge_ref
47  ds$pre_err3 <- ds$pre3 - ds$judge_ref
48  ds$pre_err4 <- ds$pre4 - ds$judge_ref
49  ds$pre_err5 <- ds$pre5 - ds$judge_ref
50  ds$pre_err6 <- ds$pre6 - ds$judge_ref
51  ds$pst_err1 <- ds$pst1 - ds$judge_ref
52  ds$pst_err2 <- ds$pst2 - ds$judge_ref
53  ds$pst_err3 <- ds$pst3 - ds$judge_ref
54  ds$pst_err4 <- ds$pst4 - ds$judge_ref
```

40:1 | READ DATA | R Script

Console | Terminal | Jobs

~/Dropbox/past_projects/study_doorwaytransfer/analyses/

```
── Column specification ────────────────────────────────────────────────────────
cols(
  .default = col_double(),
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
ds	100 obs. of 26 variables
my_theme	List of 11
pd	Environment
Values	
cbp1	chr [1:8] "#999999" "#E69F00" "#56B4E9" "#009E73" "…

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > Dropbox > past_projects > study_doorwaytransfer > analyses

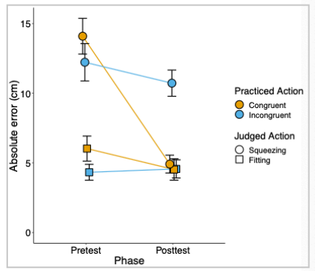
Name	Size	Modified
..		
.RData	136.7 KB	Dec 17, 2019, 3:46 PM
.Rhistory	24.1 KB	Dec 17, 2019, 3:46 PM
age.xlsx	26.1 KB	Jul 23, 2019, 9:46 AM
analyses.R	17.4 KB	Dec 17, 2019, 3:46 PM
analyses.Rproj	204 B	Jan 5, 2021, 1:16 PM
correlations.pdf	9.9 KB	Aug 7, 2019, 3:40 PM
error.pdf	5.4 KB	Aug 7, 2019, 3:42 PM
output.txt	10.4 KB	Jul 17, 2019, 12:22 PM
summarySE.R	1.7 KB	Aug 7, 2018, 4:16 PM

- R reads in output.txt, assigns factors, and calculates DVs
- Creates the analyses and figures for the paper

doorway_transfer_LaTeX_final

Search

apparatus-compressed.pdf
apparatus.pdf
correlations.pdf
DesignFigure.pdf
diff.aux
diff.bbl
diff.blg
diff.log
diff.out
diff.pdf
diff.synctex.gz
diff.tex
doorway_transfer.aux
doorway_transfer.bbl
doorway_transfer.blg
doorway_transfer.log
doorway_transfer.out
doorway_transfer.pdf
doorway_transfer.synctex.gz
doorway_transfer.tex
error.pdf
ExampleTrials.mp4
master.bib
sage_latex_template_4
SageH.bst
sagej.cls
sagej.log
SageV.bst
submission1
submission2
submission3-publication



error.pdf

PDF document - 5 KB

Information Show More

Created Aug 7, 2019 at 3:42 PM
Modified Aug 7, 2019 at 3:42 PM

Tags

Add Tags...

Markup More...

doorway_transfer.tex

Typeset LaTeX Macros Tags Labels Templates

replicability \citep{AsendorpfConner2013, SimmonsNelson2011}, the current study's procedure and analysis plan were pre-registered before data collection began. The pre-registration document was entered on AsPredicted.com ({\tt https://aspredicted.org/s58hb.pdf}). The sample size was determined in advance based on a power analysis: Past work comparing pretest/posttest affordance judgment errors across multiple between-subjects conditions found a large interaction effect size, $\eta^{2} = .152$ \citep{Recal}. However, the transfer effect in the current study might be smaller, so the sample size was conserv
effect size ($\eta^{2} = .0
resulted from a technical
Although postural sway n
accelerometer data were
measurements to trials a
the pre-registration we re
``mismatching"; here we

104
105 \subsection{Participants a
106
107 The final sample included
\textit{SD} = 1.19, 54 fem
conditions in a 2 Judged
(congruent, incongruent)
\textit{squeezing-congrue
male), \textit{squeezing-i
11 male), \textit{fitting-co
male), and \textit{fitting-ir
11 male). Judgment phas
subjects factor. Two addit
replaced for failure to foll
repeated requests from t

doorway_transfer.pdf

Scale 138 Page 12 of 22

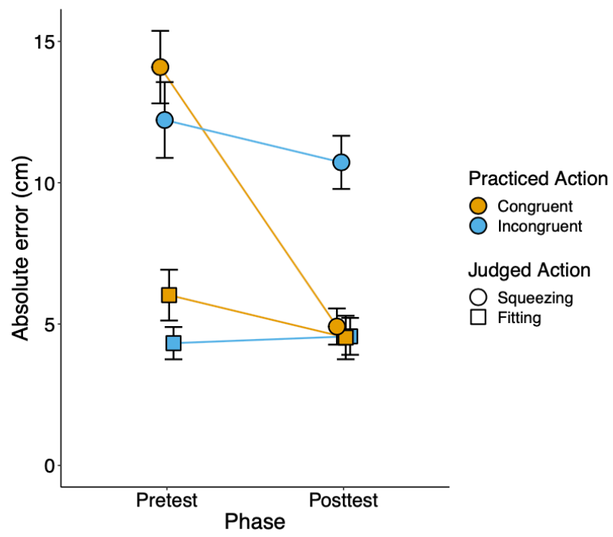


Figure 3. Absolute judgment errors by phase (x-axis: pretest vs posttest), judged action (circles: squeezing; squares: fitting), and practiced action (orange symbols: congruent practice; blue symbols: incongruent practice).

with the pre-registration plan, participants with outlier data (based on inter-quartile range) were excluded and not replaced (3 participants in the FC condition and 1 participant in the FI condition). The influence of outliers on the results is discussed below.

Overall model Table 1 shows results for the overall model predicting absolute judgment errors from judged action, practiced action, and judgment phase (and their interactions) as fixed effects and random intercepts by participant (random slope models failed to converge). As seen in Figure 3, a significant main effect of phase indicates that participants were more accurate in posttest compared with pretest, and a significant main effect of judged action reveals that participants were more accurate when judging fitting compared with squeezing. However, these effects were moderated by significant two-way (judged action×phase, practiced action×phase) and three-way (judged action×practiced action×phase) interactions. Including or excluding outliers did not affect the significance of any of the effects in the

What was good?

- Files: Organized in directories, consistently named, mostly flat text files
- Version control: minimal intermediaries, used GitHub
- Automation: Data collection was automated, cleaning raw data was automated, analyses/figures were automated
- Automation: Minimal copy/paste, minimal user typing
- Metadata saved about each participant to diagnose problems, understand exclusion

What should be better?

Workflow issues

- Using Matlab scripts/data files makes project harder to share, less future-proof; no way to archive the coding environment
- Copy/pasting data/figures between folders, analyses to paper leaves room for error
- Clear set of processes to regenerate from raw data, but not documented well (requires user to run scripts in correct order which only I know)
- No formal checks of data quality
- Lots of repetitive, single-use Matlab and R code

Example 3

- Data and code stored in shared GitHub repository
- R scripts to process raw data .csvs from multiple sources into an analysis file
- Manuscript written in RMarkdown so that analyses and code

main · 2 Branches · 0 Tags · Go to file t · Add file · Code

JohnFranchak got word version working — 027de33 · 2 years ago · 251 Commits

data-cleaned	rewriting pre-processing	3 years ago
data-raw	initial commit	3 years ago
docs	updated slides	3 years ago
manuscript	got word version working	2 years ago
metadata	walkers in table	3 years ago
presentation-conference	walkers in table	3 years ago
presentation-data-report	calculated object holding restraint context	3 years ago
scripts-analysis	small tweaks	3 years ago
scripts-preprocessing	rewriting pre-processing	3 years ago
.gitattributes	Initial commit	3 years ago
.gitignore	trackdown	3 years ago
.nojekyll	no jekyll	3 years ago
LICENSE	Initial commit	3 years ago
walking_ema.Rproj	updated analyses	3 years ago

About

Longitudinal relations between independent walking, body position, and object experiences in home life

johnfranchak.github.io/walking_ema/

MIT license

Activity

0 stars

2 watching

0 forks

Deployments 31

github-pages 3 years ago

+ 30 deployments

Languages

TeX 94.3% **HTML** 3.6% **Rich Text Format** 1.1% **Other** 1.0%

Suggested workflows

Based on your tech stack

Edit | Preview | Spaces | 2 | Soft wrap

```
to reliably categorize when they are held in caregiver's arms or in slings/carriers.

Finally, *object holding* was scored as a 1 if caregivers responded "yes" to question 6a ("Is your child holding an object") and scored a 0 otherwise.

### Responsivity Statistics

To determine if responses reflected truly in-the-moment observations of behavior (rather than caregivers responding to prompts only when convenient), we calculated the *response time* by subtracting the time the response text message was sent from the time the survey was completed by the participant. Figure \@ref(fig:response-time) shows a histogram of all response times in minutes across participants at every age. The two peaks roughly correspond to surveys that ended after the Question 1 ("child sleeping" and "not with child" responses required answering only Question 1) compared with surveys that required caregivers to answer all 6 questions ("child awake"). The maximum possible response time was 15 minutes, after which the survey would automatically close. The right skew of the histogram indicates that most responses were made quickly, with a *median* of `r printnum(median(ds$responsetime))` min. Thus, caregivers' prompt responses to survey requests suggest that observations captured in-the-moment infant behavior.

```{r response-time, fig.cap="Distribution of response time (elapsed time in minutes between receiving the prompt and completing the survey) showing that responses captured infants' behavior at the moment caregivers received survey requests (median response time = 0.5 minutes). For most responses, children were awake with caregivers present to report their behavior (blue bars). Subsequent analyses report experiences as percent of awake samples, excluding times that infants were sleeping or caregivers were not with their child (brown bars).", fig.align='center', fig.height=5, fig.width=7, cache=TRUE}
ds %>% drop_na(mc_presence) %>%
  mutate(mc_presence = factor(mc_presence, labels = c("Child Awake", "Child Sleeping", "Not with Child"))) %>%
ggplot(aes(x = responsetime, fill = mc_presence)) +
  geom_histogram(binwidth = .2) +
  xlab("Response time (min)") +
  ylab("Number of samples") +
  scale_fill_manual(name="",  values = natparks.pals("KingsCanyon", n = 8)[c(6,3,2)])
```


Results
```

Responsivity Statistics. To determine if responses reflected truly in-the-moment observations of behavior (rather than caregivers responding to prompts only when convenient), we calculated the *response time* by subtracting the time the response text message was sent from the time the survey was completed by the participant. Figure 1 shows a histogram of all response times in minutes across participants at every age. The two peaks roughly correspond to surveys that ended after the Question 1 (“child sleeping” and “not with child” responses required answering only Question 1) compared with surveys that required caregivers to answer all 6 questions (“child awake”). The maximum possible response time was 15 minutes, after which the survey would automatically close. The right skew of the histogram indicates that most responses were made quickly, with a *median* of 0.50 min. Thus, caregivers’ prompt responses to survey requests suggest that observations captu

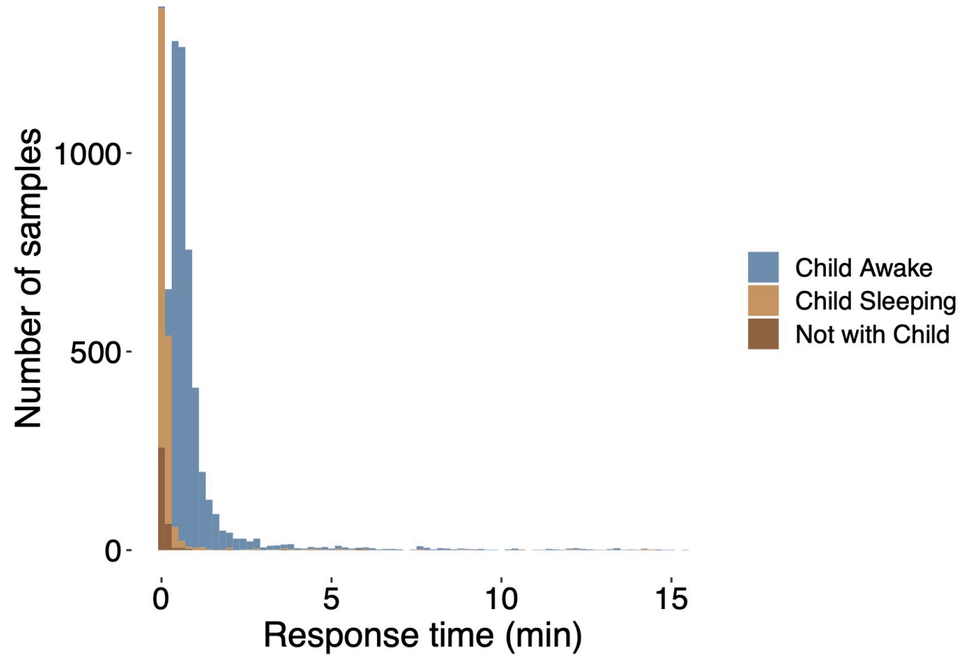


Figure 1. Distribution of response time (elapsed time in minutes between receiving the prompt and completing the survey) showing that responses captured infants’ behavior at

What was good?

- Reproducible manuscript
  - No copying and pasting numbers/figures
  - Self-documenting: I know what code produced every number and figure
  - Anyone else who downloads the code also knows
- Shared data and code with publicly available DOI

What should be better?

What should be better?

- Code might not run the same under a different version of R/packages
- Poor documentation of workflow code -> I'm not sure I know what all the steps are or what the data sources are from
- Messy, manual entry data needed lots of cleanup

Logistics

Schedule, syllabus, and readings on course website

- Let's all go visit:
  - https://2025-psyc-259.github.io/course_site/
- The course site will contain (and preserve) all lecture examples, and link to Github repositories code
- The course site is written in Quarto/R and hosted on GitHub

Class structure

1. Before class - readings from *R for Data Science*
2. During class
    a. Lecture/tutorial on those skills
    b. Hands-on time to work on coding assignments w/ instructor/TA help
3. After class: Finish homework on your own/ in groups to crystallize skills

Weekly assignments

- Designed to give you a chance to practice coding skills covered in lecture
- Group work is OK (but no more than 3 per group)
- Getting help
  - Start assignments in class
  - Office hours

Two bigger assignments

- Workflow self-critique presentation
  - Choose a data analysis project from your lab
  - Describe the end-to-end workflow: What's the raw data? How are data combined/processed? What resources are involved? What does the end product need to look like?
  - Critique the workflow: Where are errors likely to occur? Where could things be automated? How could the workflow be made more transparent and reproducible?

Two bigger assignments

- Final Project
  - Take that workflow and make it better
  - End-to-end workflow (raw data, processing/checking/analysis scripts) shared on a GitHub repository
  - Will include a reproducible report (Quarto) that communicates exploratory or confirmatory analyses with visualizations in R

Course focus:

Why R?

What's the difference between R and RStudio?

R vs. RStudio

- R is a programming language
  - You can run R code from the command line or another IDE (like VSCode)
- RStudio IDE
  - IDE = Integrated Development Environment = software that makes programming easier (panes for scripts, console, graphs, data environment, versioning, etc)
  - You can run Python code or knit Rmarkdown/Quarto files from within RStudio

What's better: Moderate skill in 1 language or novice skill in 2 or 3?

Why R (for this class)?

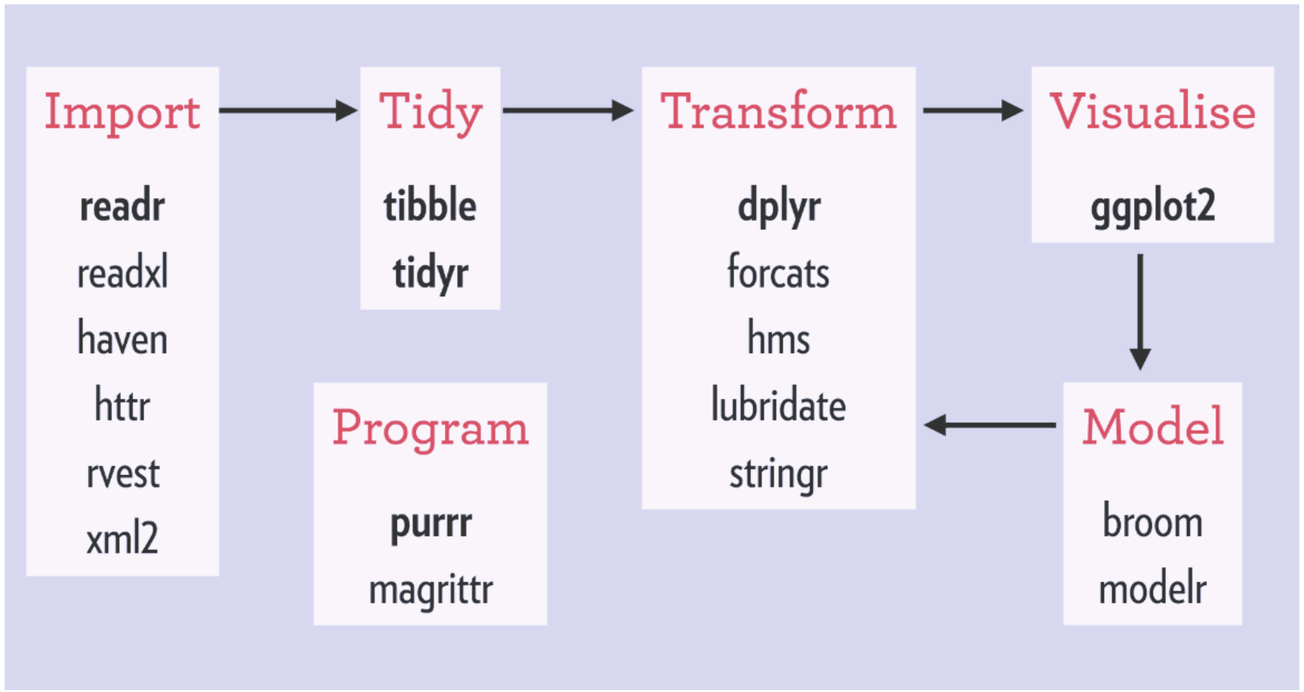
- Build on what you know
- Free, open-source, statistics-focused
- Powerful, extensible
- Reproducible (flat script files, easily shareable)
- RStudio IDE
- ggplot2
- R Markdown/Quarto for academic reports

Why not _____?

- Matlab?
- Python?
- Julia?

- I'm teaching R, but you can use other languages for the big projects

Why focus on the “tidyverse”?



- Well-documented, large user community
- Internally consistent, pieces work well together
- Emphasis on verbs rather than nouns
- Not saying “base R” is bad!

Learn more about base R

- *The Art of R Programming* by Norman Matloff
- *R Cookbook* by Paul Teetor

- Use whatever works for you, but for homework try to do what we're teaching

Software setup help

Fork a repository to make your own copy of it

- Log in to github.com with your username
- Go to https://github.com/2025-PSYC-259/package_install

2025-PSYC-259 / **package_install**

Type / to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

package_install

Public

Edit Pins | Unwatch 1 | Fork 0 | Star 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

JohnFranchak initial commit — ab280d6 · 4 minutes ago — 1 Commit

.Rproj.user	initial commit	4 minutes ago
package_install.R	initial commit	4 minutes ago
package_install.Rproj	initial commit	4 minutes ago

README

Add a README

Help people interested in this repository understand your project by adding a README.

Add a README

About

No description, website, or topics provided.

Activity
Custom properties
0 stars
1 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

R 100.0%

Fork a repository to make your own copy of it

- Log in to github.com with your username
- Go to https://github.com/2025-PSYC-259/package_install
- Now it won't be 2025-PSYC-259's anymore, it will be yours
- Let's get it off the Github website and onto your local computer by **cloning it**

package_install

Public

Edit Pins | Unwatch 1 | Fork 0 | Star 0

main | 1 Branch | 0 Tags

Go to file t | Add file | Code

JohnFranchak initial commit

.Rproj.user	initial commit
package_install.R	initial commit
package_install.Rproj	initial commit

README

Add a README

Help people interested in this repository understand your project by adding a README.

Add a README

Local | Codespaces

Clone

HTTPS SSH GitHub CLI

https://github.com/2025-PSYC-259/package_i

Clone using the web URL.

Open with GitHub Desktop

Download ZIP

About

No description, website, or topics provided.

Activity

Custom properties

0 stars

1 watching

0 forks

Report repository

Releases

No releases published

Create a new release

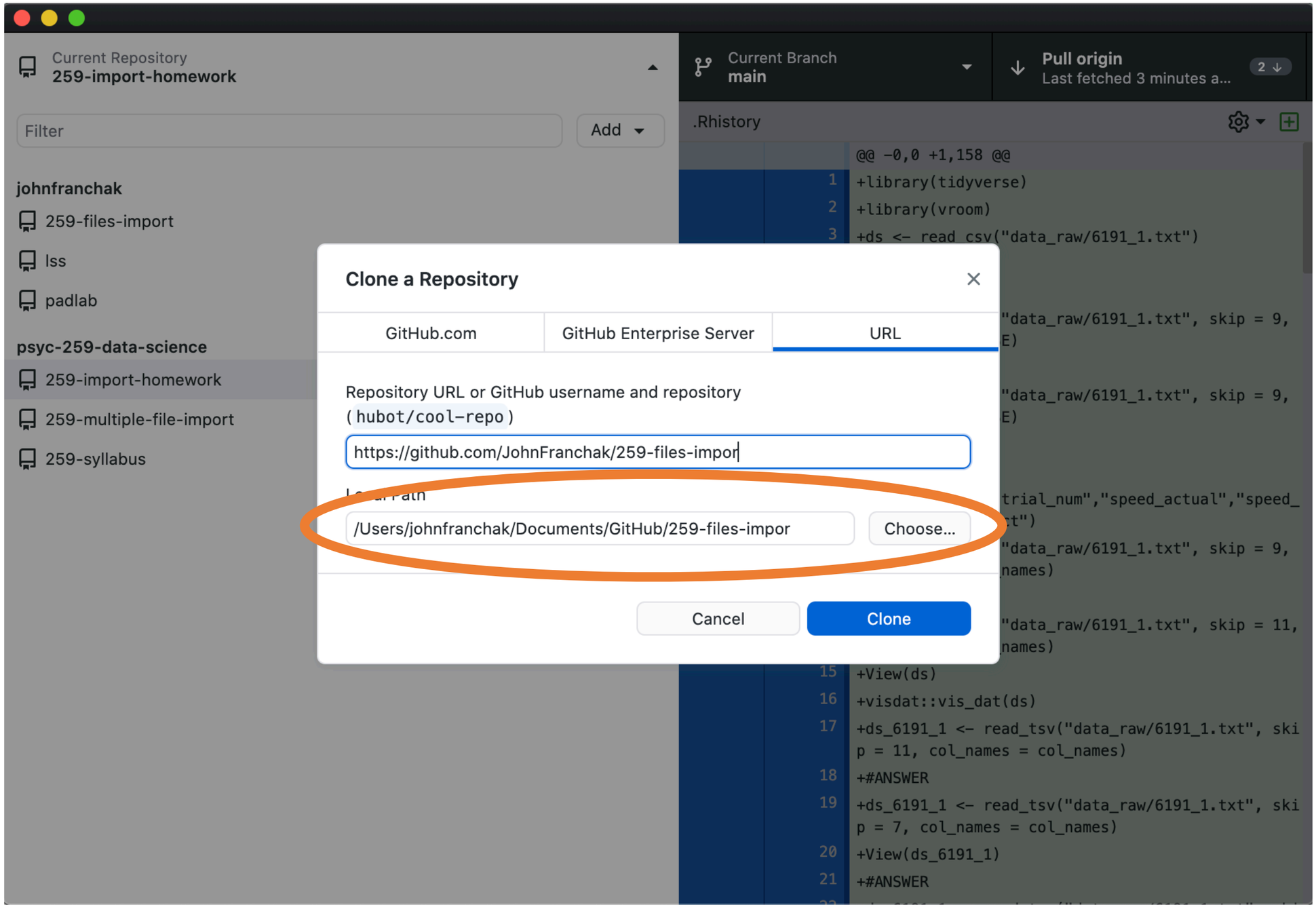
Packages

No packages published

Publish your first package

Languages

R 100.0%



Choose a local destination (not on cloud)

Now we’re set up!

- Before starting work, you should always “fetch/pull origin” to get the most recent version from the origin (GitHub)
- Now make a change to a file in the folder (and save the change)
- Go back to the Github app and see what happens