

PSYC 259: Principles of Data Science

Week 1: Course Goals

Today

1. Introductions
2. Course overview
3. Workflow examples
4. Logistics
5. 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
 - Time spent working on methods > results

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 + SPSS until a 4 years ago

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 economics³

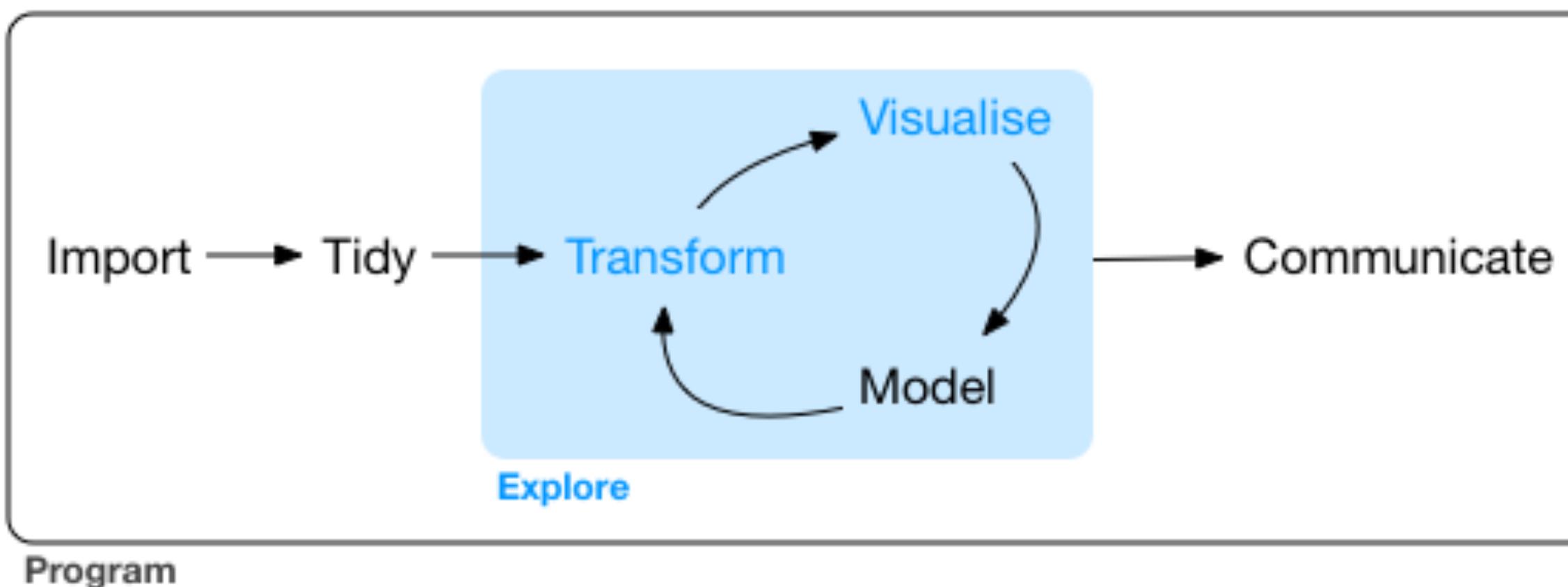
Why this class?

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

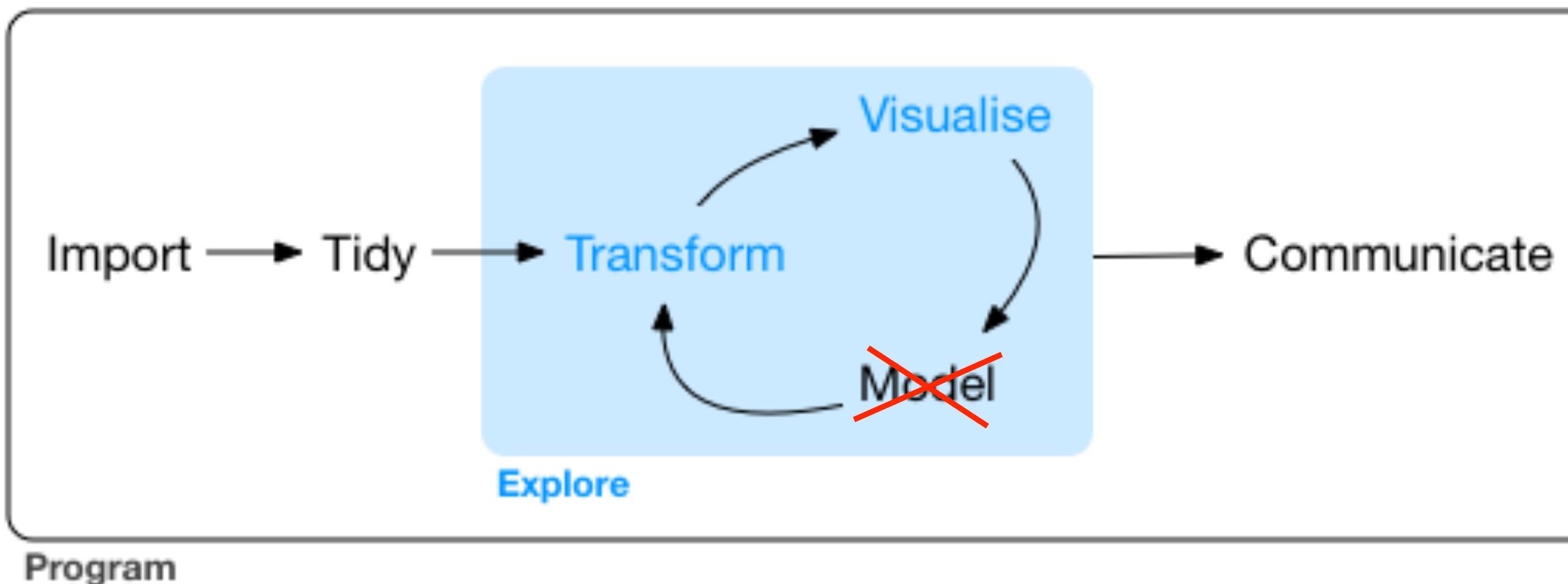
Why this class?

- 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 means mastering the art of data *processing* and data *exploration*

Course focus: Programming for data exploration & communication



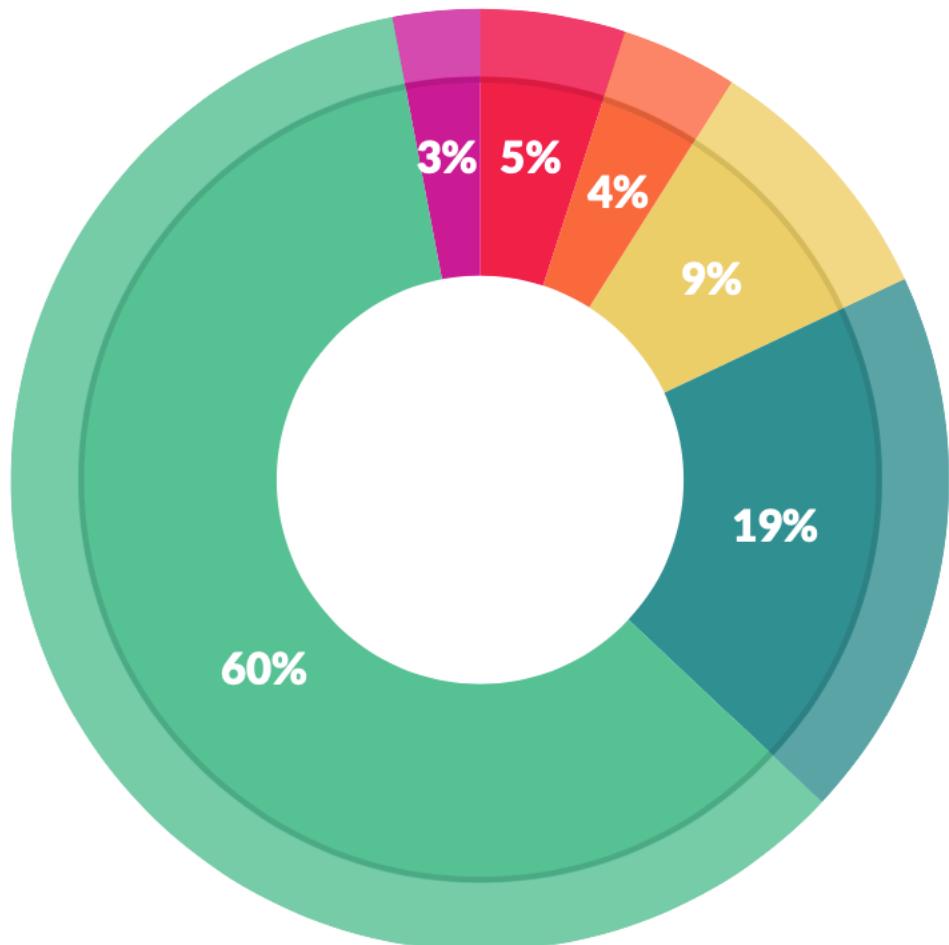
Course focus: Programming for data exploration & communication



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

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%*

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_handles', 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

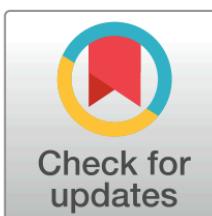
* m.ziemann@deakin.edu.au

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.



OPEN ACCESS

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>

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

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	Exp1data	6191.txt	Participant 6192
BigWalkerSpeed08-16-07.doc	Exp1documents	6192.txt	Age 21
SpatialWalkerDynamics	experiment2	6193.txt	Sex male
WalkingPerception	experiment3	6194.txt	Order 2
	experiment4	6195.txt	First Speed 2.4
	experiment5position	output	6/19/2006 12:00:26 PM
	perspective	perspective.sav	*****
			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	
			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

▶ Exp1data
▶ Exp1documents
▶ experiment2
▶ experiment3
▶ experiment4
▶ experiment5position
▶ perspective
▶ 6191.txt
▶ 6192.txt
▶ 6193.txt
▶ 6194.txt
▶ 6195.txt
▶ output
▶ perspective.sav

▶ 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
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

a 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	16	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

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	plc ond	
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	
					70.00	2.40	far	

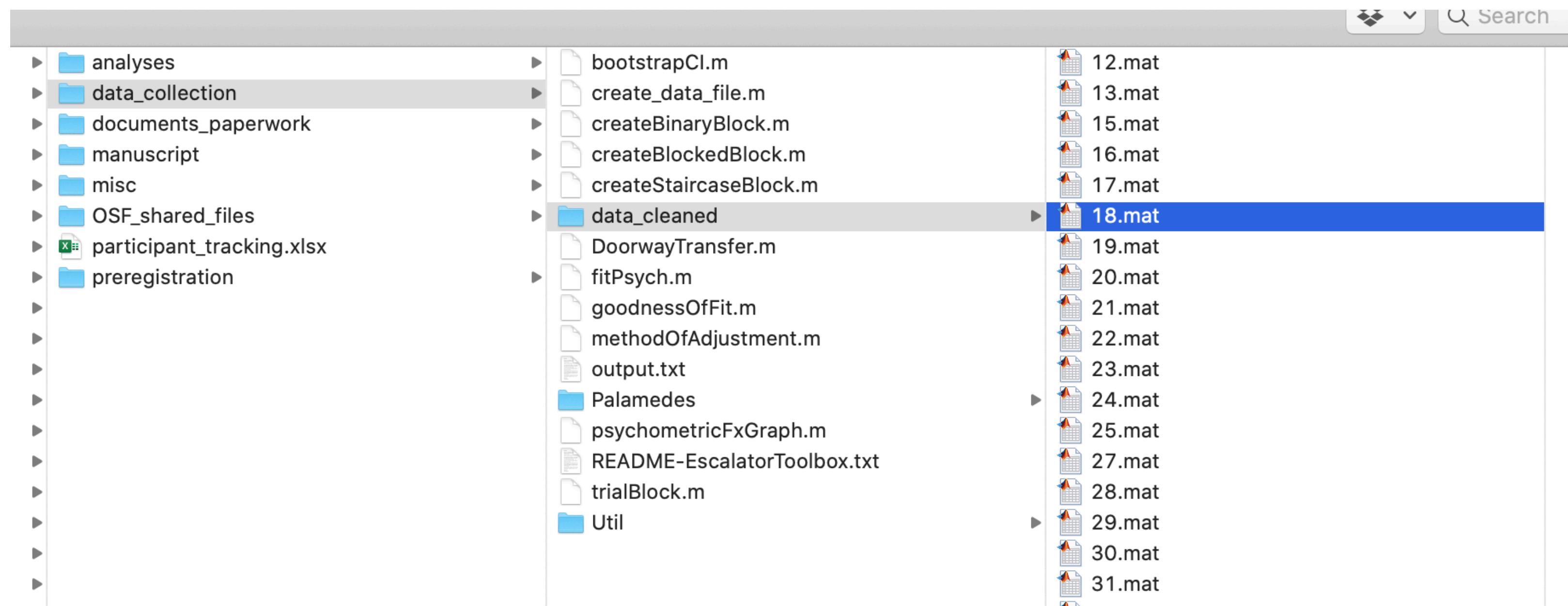
List of crimes against data science

- 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



Command Window

```

>> sex

sex =
    1x1 cell array

    {'m'}

>> id

id =
    1x1 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: [1x30 double]
        trial_unit_fit: [1x30 double]
        trial_resp: [1x30 double]
        trial_resp_fit: [1x30 double]
        trial_mode: [1x30 double]
        trial_subblock: [1x30 double]
        stim_levels: [1x107 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
<input checked="" type="checkbox"/> save_figs	1
save_file	'C:\Users\labuse...
sex	1x1 cell
stim_levels	1x107 double
weight1	1x1 cell
weight2	1x1 cell

```

id,sex,judgecond,practcond,height,weight,doormeasure,aff_s,span_s,aff_f,span_f,pre1,pre2,pre3,pre4,pre5,pre6,pst1,pst2,
pst3,pst4,pst5,pst6
100,f,S,S,60,415,47,3,25,3,19,-0,5,41,8,40,15,44,2,43,1,44,35,41,9,43,5,36,05,31,6,30,55,27,75,30,25,25
101,m,F,F,67,65,64,7,24,7,27,6,1,45,1,3,49,5,46,5,40,45,44,2,43,35,43,05,48,45,47,9,47,25,49,3,47,75,6
102,f,S,S,62,55,42,9,22,55,25,1,0,5,41,1,12,44,95,41,2,42,42,2,42,75,42,55,25,45,25,7,23,35,24,4,24,25,23,85
103,m,F,F,71,68,28,15,29,4,1,44,9,45,95,46,85,44,7,46,6,44,15,43,6,47,1,44,7,44,65,44,5,43,65,45,1
104,m,S,F,70,117,6,34,9,36,6,2,54,6,6,38,65,37,85,41,85,42,55,37,8,43,7,38,65,35,45,37,55,38,35,37,75,37,8
105,f,S,S,57,9,42,9,23,95,24,2,3,5,45,2,5,35,65,32,8,33,45,35,25,37,9,33,8,25,65,24,7,27,45,27,5,23,6,25,7
106,f,F,S,65,5,81,7,31,6,31,1,0,5,47,6,7,50,7,51,35,55,2,51,25,52,8,48,4,48,7,48,05,51,8,48,15,49,6
107,f,F,F,62,05,48,1,25,4,30,7,1,5,42,2,1,44,8,44,65,44,75,43,85,42,95,42,3,43,95,42,5,43,75,44,65,46,4,46,2
108,f,S,S,71,1,120,6,37,35,28,1,1,5,57,6,6,5,54,3,53,75,51,15,49,6,52,2,54,35,57,8,56,45,53,6,56,3,53,05,57,3
109,m,F,F,70,6,75,5,39,2,31,0,45,5,1,5,34,95,37,3,38,1,35,95,38,3,35,7,40,35,42,5,43,9,39,15,41,8,42,75
nboot = 112,F,S,F,71,75,110,8,41,05,37,5,5,61,5,2,46,4,43,65,48,35,43,75,40,8,42,2,44,7,49,5,39,9,40,55,43,55,37,9
113,M,F,F,71,79,27,65,24,5,0,45,4,50,6,48,75,58,7,49,1,49,4,49,35,51,7,48,7,49,05,48,2,48,4,48,5
for i =
if
115,m,S,F,73,35,79,25,27,15,33,1,0,5,43,6,2,36,5,35,35,1,44,44,3,43,6,45,4,39,55,41,45,40,15,40,6,42,1
116,m,S,F,73,94,7,27,1,25,5,2,5,46,5,5,32,15,28,85,32,55,31,6,29,9,29,9,36,15,32,4,30,25,32,05,31,9,30,95
12,f,F,S,61,54,2,25,15,28,6,0,5,43,1,4,5,50,54,4,52,5,53,55,53,9,54,55,45,5,47,2,45,8,45,85,47,8,47,45
13,f,F,S,62,5,49,6,25,85,23,8,1,5,39,8,3,5,44,1,44,4,44,45,2,42,7,43,65,37,05,39,2,40,15,40,4,38,8,38,65
15,m,F,F,67,95,55,1,26,05,27,8,0,45,8,5,45,47,05,49,8,51,5,55,51,8,47,7,47,7,51,55,50,2,51,9,51,5
16,m,S,F,NaN,NaN,NaN,33,2,3,5,47,7,9,5,35,95,36,75,33,55,37,25,36,4,39,95,43,9,46,15,46,43,9,45,75,45,4
17,f,F,S,66,9,48,6,36,2,28,2,3,39,7,3,42,35,43,35,41,7,39,35,34,9,36,95,48,48,7,49,35,49,6,47,95,47
18,m,S,S,67,6,63,9,38,15,33,7,9,42,2,2,53,45,53,2,58,4,54,9,53,8,51,7,48,7,40,1,39,75,40,4,39,25
19,f,F,F,60,4,84,5,34,05,36,1,1,5,56,6,4,56,75,52,15,53,9,55,15,55,95,57,95,49,2,46,85,47,1,49,15,50,7,49,85
20,m,S,F,67,5,86,4,27,9,25,3,2,5,47,8,5,43,95,44,55,40,55,41,75,40,5,40,1,34,75,36,7,34,05,35,5,33,6,33,45
21,f,F,S,59,25,45,8,25,75,26,6,1,5,39,1,6,41,55,38,6,39,2,38,9,41,15,40,7,42,8,38,4,39,35,39,85,38,7,40,7
22,m,S,S,66,5,75,7,28,05,30,6,-0,5,44,6,1,38,75,38,45,39,7,36,2,38,3,39,8,40,1,40,35,41,9,44,45,43,5,44,3
23,f,F,F,65,45,76,8,32,55,29,6,4,53,6,4,5,45,85,47,9,50,2,49,75,48,75,51,55,54,35,53,15,58,8,51,95,52,65,53,75
24,f,S,F,62,75,52,9,22,95,25,1,43,3,5,45,95,39,38,85,37,75,37,8,36,55,36,35,35,9,35,3,36,25,36,15,33
25,m,F,S,66,2,91,8,42,7,37,4,49,5,-0,5,57,65,59,2,56,75,58,2,59,45,63,5,58,75,58,8,58,9,58,8,58,75,57,85
27,m,F,F,71,5,66,7,36,45,28,1,43,10,49,3,54,5,58,1,51,45,53,3,56,85,54,25,54,95,52,95,55,35,54,05,56,95
28,m,S,S,73,4,74,6,25,9,26,9,1,42,4,5,5,33,65,33,1,32,95,34,05,36,55,39,6,29,25,30,15,28,55,29,85,27,9,27,15
29,f,S,S,61,3,59,2,27,55,25,6,0,42,1,6,43,5,43,05,43,5,45,9,47,95,47,5,28,75,29,45,32,7,38,5,32,1,28,15
30,f,F,F,60,85,62,25,85,28,7,0,5,45,7,12,5,48,35,46,35,47,75,49,45,49,2,51,15,47,9,50,44,25,49,95,48,2,48,5
31,m,S,F,71,6,76,1,25,95,28,1,1,5,42,6,4,41,15,42,85,45,75,42,2,43,7,39,3,36,25,39,42,25,38,7,39,9,40,25
32,m,F,S,71,6,77,27,65,25,6,0,42,1,3,5,51,65,49,35,49,2,46,65,45,1,43,5,47,05,47,5,44,45,46,55,45,15,45,55
33,m,S,S,72,5,83,3,27,15,29,3,6,42,8,3,46,35,43,05,41,1,41,37,55,40,5,30,95,31,1,31,2,29,8,31,25,28,25

```

- “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

The screenshot shows the RStudio interface with several panes:

- Top Left (Code):** The script editor pane displays the R code for "analyses.R". The code reads a CSV file ("output.txt"), creates factor variables for "id", "practice", "judge", and "match", and calculates error terms for each judge category.
- Top Right (Environment):** The environment pane shows the global environment with objects: "ds" (100 obs. of 26 variables), "my_theme" (List of 11), and "pd" (Environment). It also lists values: "cbp1" (a character vector of length 8).
- Bottom Left (Console):** The console pane shows the current working directory as "/Dropbox/past_projects/study_doorwaytransfer/analyses/" and a "Column specification" section.
- Bottom Right (Files):** The files pane shows a directory structure under "analyses" in "study_doorwaytransfer". The contents are:

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.tex

ION1.pdf

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

Markup More...

Typeset LaTeX Macros Tags Labels Templates

replicability \citep{AsendorptConner2013, 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 (\url{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 conservative. The resulting effect size ($\eta^2 = .02$) resulted from a technical error in the pre-registration. Although postural sway measurements from the accelerometer data were measured relative to trials and not absolute, the pre-registration we recorded the absolute error. In this paper, we refer to this as "mismatching"; here we report the absolute error.

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...

Absolute error (cm)

Pretest Phase Posttest

Practiced Action
Congruent (Orange circle)
Incongruent (Blue circle)
Judged Action
Squeezing (Orange square)
Fitting (Blue square)

12 Quarterly Journal of Experimental Psychology XX(X)

104
105 \subsection{Participants and Materials}

106
107 The final sample included 54 participants (46 female, 8 male) assigned to one of two conditions in a 2 Judged Action (congruent, incongruent) \times 2 Practiced Action (squeezing-congruent, squeezing-incongruent) \times 2 Judged Phase (male), \textit{squeezing-male}, \textit{squeezing-incongruent-male}, \textit{fitting-congruent-male}, \textit{fitting-incongruent-male}, and \textit{fitting-incongruent-female} ($n = 11$ male). Judgment phase was a within-subjects factor. Two additional participants were excluded for failure to follow instructions and one participant was replaced for failure to follow repeated requests from the experimenter.

Absolute error (cm)

Pretest Phase Posttest

Practiced Action
Congruent (Orange circle)
Incongruent (Blue circle)
Judged Action
Squeezing (Orange square)
Fitting (Blue square)

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 \times phase, practiced action \times phase) and three-way (judged action \times practiced action \times phase) interactions. Including or excluding outliers did not affect the significance of any of the effects in the model.

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?

What should be better?

- 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 leave plenty of 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

Logistics

Schedule

Week 01, 01/03: Course Goals and Setup

- Goals of the class
- Logistics
- Workflow examples
- Software setup
- Introduction to Github

Week 02, 01/10: File Organization and Workflow

- File organization
- Version control
- *SKILLS*: R language basics (variables, assignment, comments, calling functions port/export)

Week 03, 01/17: NO CLASS - MLK DAY

Week 04, 01/24: Data Transformations

- Basic data types (numbers, strings)
- Variable naming
- *SKILLS*: dplyr data wrangling, logical statements, strings

Week 05, 01/31: Data Structure

- Tidy data (observations/cases)
- More advanced data types (factors, dates, vectors, lists)
- *SKILLS*: More data wrangling, factors, dates

Week 06, 02/07: Automation: Functions and Iteration

- Finding new packages/APIs vs. writing your own functions
- Automating your analyses
- Writing more efficient code
- *SKILLS*: Writing custom functions, iteration (map, for loops)

Week 07, 02/14: Exploratory Data Analysis

- **WORKFLOW PRESENTATIONS**
- Data validation
- Automating visualizations
- *SKILLS*: Basic visualizations with ggplot2

Week 08, 02/21: NO CLASS: PRESIDENTS' DAY

Week 09, 02/28: Data Sharing and Reproducibility

- Reuse-minded project management
- Reproducible reports
- Preserving programming environment and analyses
- *SKILLS*: R Markdown, package control

Week 10, 03/07: Communication

- Communicating through graphical styles
- Interactive plots for data exploration
- *SKILLS*: ggplot and extensions

Readings

- *R for Data Science*
 - Available free online
 - Chapters correspond to practical skills that I will cover in lectures
- Other readings
 - Conceptual articles that we will discuss as a group
 - Be sure to read before class so that you can participate in the discussion

Class time

1. Lecture/tutorial
2. Some weeks
 - a. Article discussions ("Other readings")
 - b. Student presentations
3. Hands-on time to work on coding assignments w/ instructor/TA help

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 (Jake's or mine)

Two bigger assignments

- Workflow self-critique
 - 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 an online repository
 - Will include a reproducible report (R Markdown) that communicates exploratory or confirmatory analyses with visualizations in R

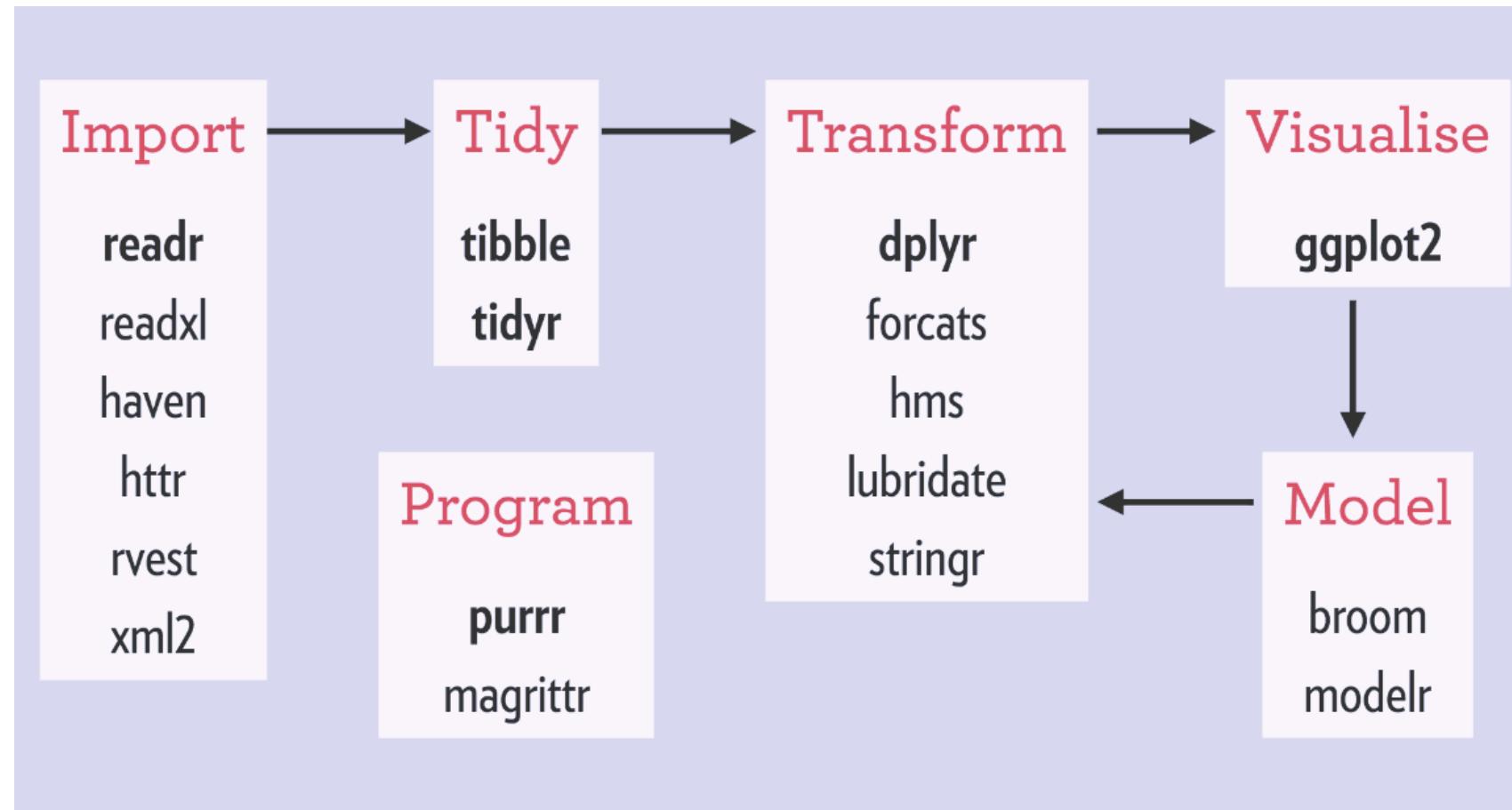
Class Github page

- Contains syllabus and readings
- Will contain (and preserve) all lecture examples

Why R (for this class)?

- Free, open-source, statistics-focused
- Powerful, extensible
- End-to-end solutions
- RStudio IDE
- ggplot2
- R Markdown for academic reports

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 Teator
- Use whatever works for you!

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/psych-259-data-science-2022/259-install-packages>

PSYC-259-Data-Science / 259-files-import

forked from [JohnFranchak/259-files-import](#)

[Watch](#) 0

[Star](#) 0

[Fork](#) 1

[Code](#)

[Pull requests](#)

[Actions](#)

[Projects](#)

[Wiki](#)

[Security](#)

[Insights](#)

[Settings](#)

[master](#) ▾

1 branch

0 tags

[Go to file](#)

[Add file](#) ▾

[Code](#) ▾

About



No description or website provided.

This branch is even with JohnFranchak:master.

[Pull request](#) [Compare](#)

[import](#) [week2](#)



JohnFranchak edited examples

8d496c7 17 hours ago 15 commits

data_cleaned	added RData example	3 days ago
data_raw	reorganized files, added comments	4 days ago
eda	added ggsave example	3 days ago
.DS_Store	added ggsave example	3 days ago
.gitignore	added two data import examples	4 days ago
01_vocab.R	cleaned up vocab	17 hours ago
02_eye_tracking_basic.R	edited examples	17 hours ago
03_eye_tracking_advanced.R	edited examples	17 hours ago
259-files-import.Rproj	added vocab example	5 days ago

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

R 100.0%

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

[Add a README](#)

Fork a repository to make your own copy of it

- Log in to github.com with your username
- Go to <https://github.com/psych-259-data-science-2022/259-install-packages>
- Now it won't be PSYC-259-Data-Science's anymore, it will be yours
- Let's get it off the Github website and onto your local computer by **cloning it**

PSYC-259-Data-Science / 259-files-import

forked from [JohnFranchak/259-files-import](#)

[Watch](#) 0

[Star](#) 0

[Fork](#) 1

[Code](#)

[Pull requests](#)

[Actions](#)

[Projects](#)

[Wiki](#)

[Security](#)

[Insights](#)

[Settings](#)

[master](#) ▾

1 branch

0 tags

[Go to file](#)

[Add file](#) ▾

[Code](#) ▾

This branch is even with JohnFranchak:master.



JohnFranchak edited examples

data_cleaned	added RData example	
data_raw	reorganized files, added com	
eda	added ggsave example	
.DS_Store	added ggsave example	3 days ago
.gitignore	added two data import examples	4 days ago
01_vocab.R	cleaned up vocab	21 hours ago
02_eye_tracking_basic.R	edited examples	21 hours ago
03_eye_tracking_advanced.R	edited examples	21 hours ago
259-files-import.Rproj	added vocab example	5 days ago

[Clone](#)

[HTTPS](#) [SSH](#) [GitHub CLI](#)

<https://github.com/PSYC-259-Data-Sc>

?

Use Git or checkout with SVN using the web URL.

[Open with GitHub Desktop](#)

[Download ZIP](#)

3 days ago

4 days ago

21 hours ago

21 hours ago

21 hours ago

5 days ago

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

[Add a README](#)

About

No description or website provided.

[import](#) [week2](#)

Releases

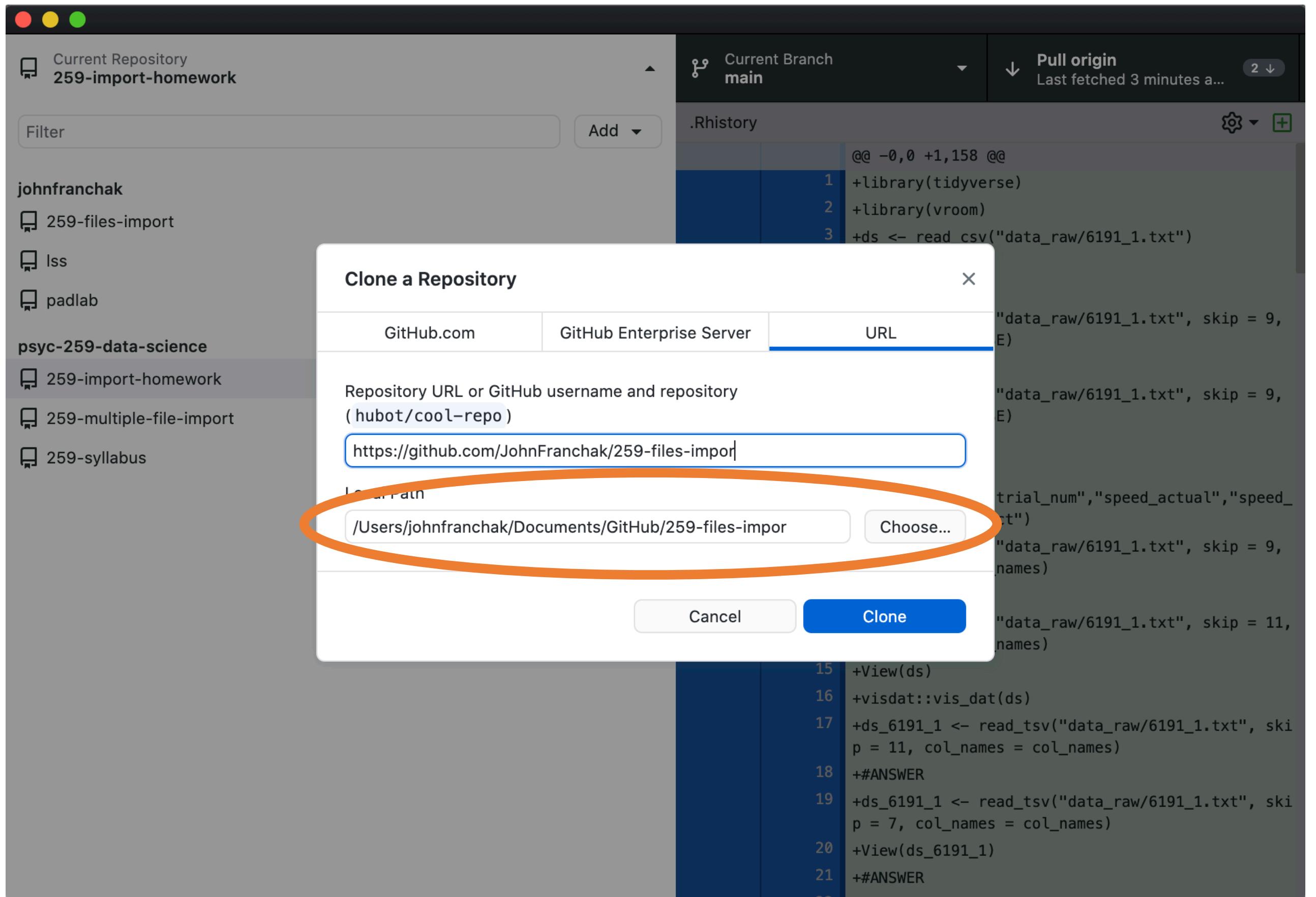
No releases published
[Create a new release](#)

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