PSYC 259: Principles of Data Science

Week 1: Course Goals

Today

1. Introductions
2. Course overview
3. Logistics
4. Next week

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, surveys
  - Methods section > results section

About me

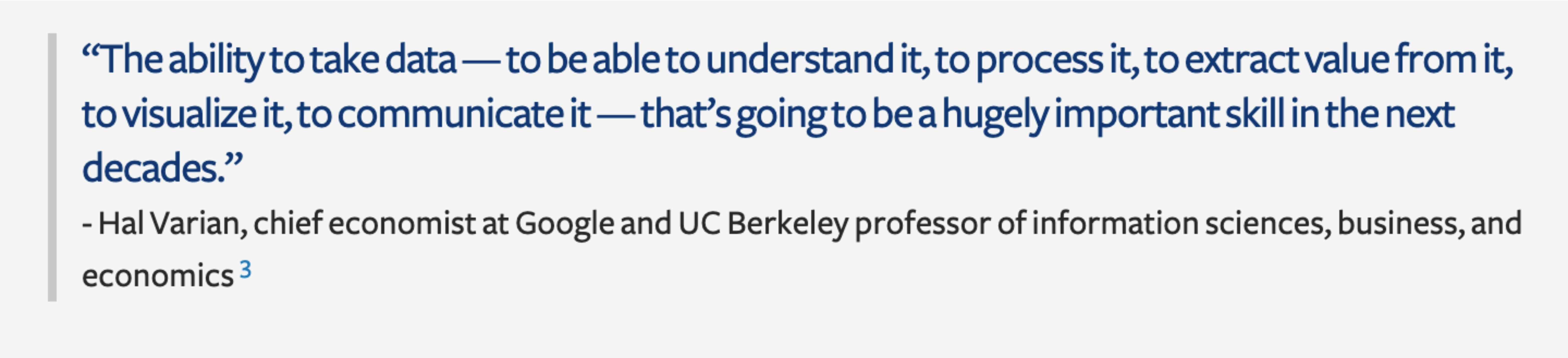
- My programming background
  - First program: Video poker on a TI-83+ graphing calculator
  - C++, visual basic in high school
  - 2 semesters of C++/java in college
  - Learned Matlab on my own as a lab manager, mostly used Matlab + SPSS until a 3-4 years ago

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 economics[3]

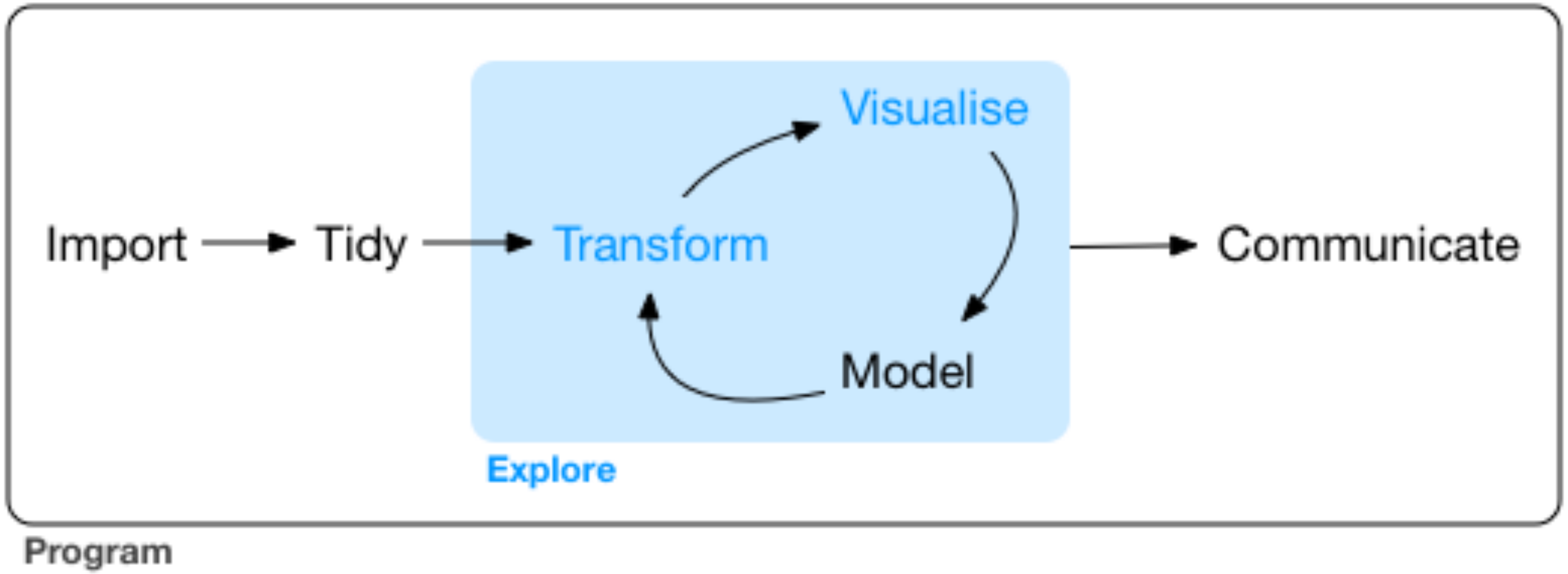
Why this class?

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

Why this class?

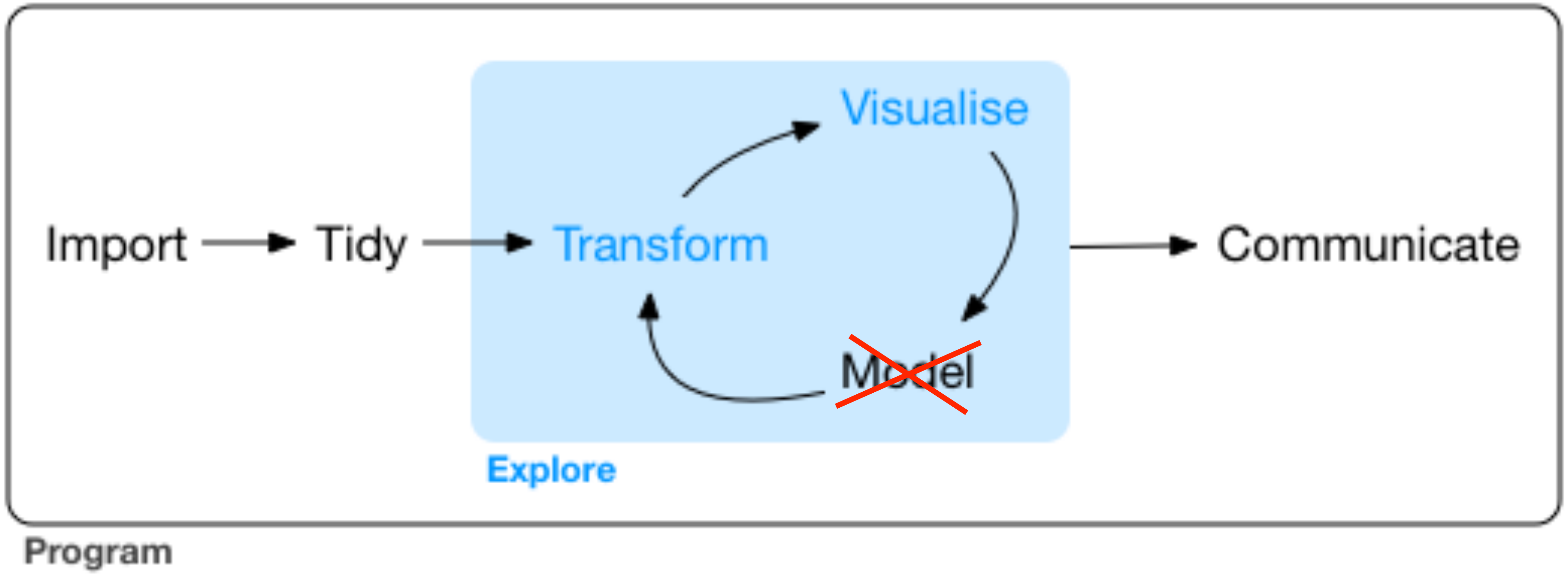
- Don’t we already teach students to understand, extract value, and communicate about data?
- Our statistics classes teach data *analysis* - testing hypotheses, modeling, etc.
- Understanding data means mastering the art of data *processing* and data *exploration*

Course focus: Programming for data exploration & communication



source: ***R for Data Science***

Course focus: Programming for data exploration & communication

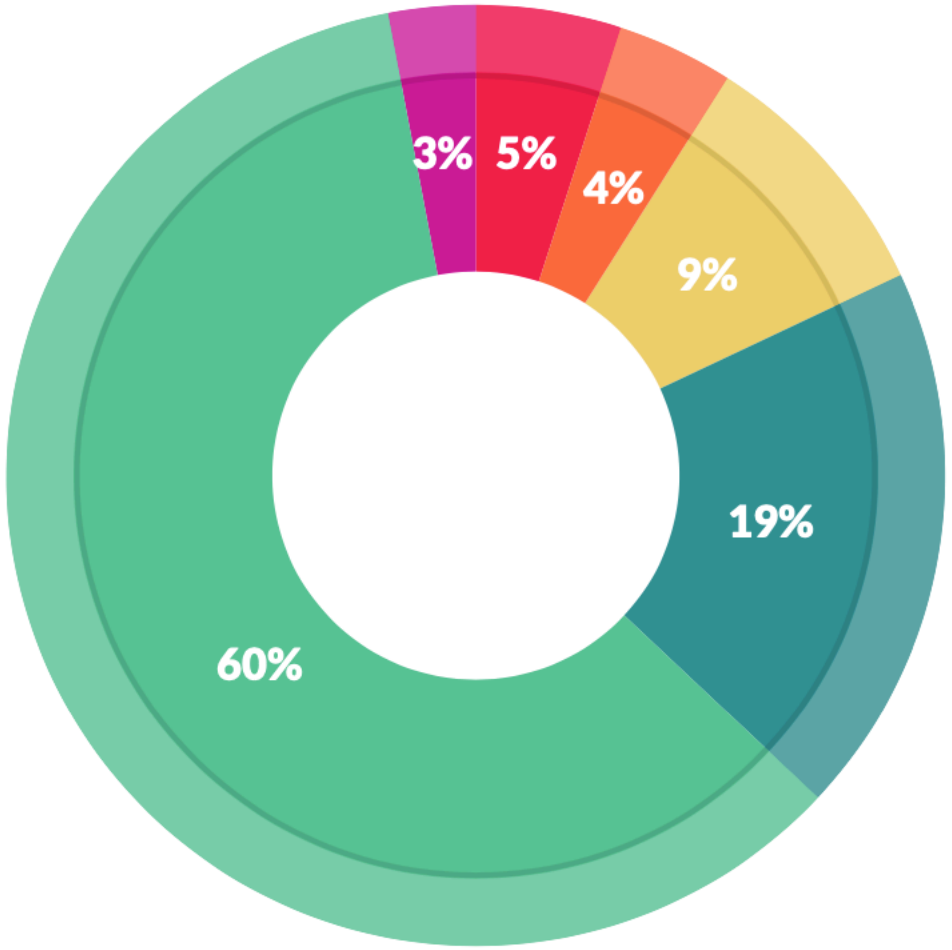


80/20 rule of data science: 80% of the work is getting the data ready to analyze, only 20% of the work is analyzing/reporting

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

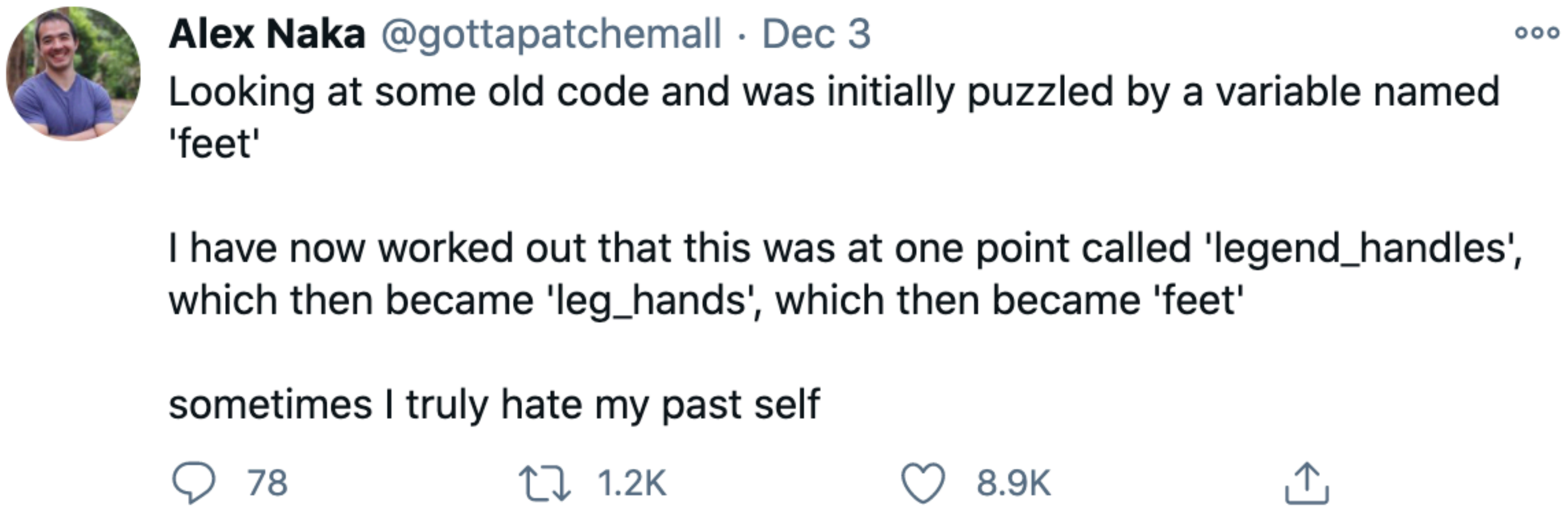
- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

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

Choosing best (*better) practices

- Pipeline A: Experiment software produces an RT measure for each participant in each condition. After running each participant, the RA enters those numbers with the participant # and condition into an Excel file. The grad student imports the data into SPSS to build graphs and run some analyses using the drop-down menus. That output file is saved and used when writing up the paper.

Choosing best (*better) practices

- Pipeline B: Experiment software writes a .txt file with each participant’s RT and metadata. Each .txt file is moved to a central server location. The grad student writes an R script that collects the data from each individual file and checks for errors. Other scripts are written to generate figures and calculate analyses.

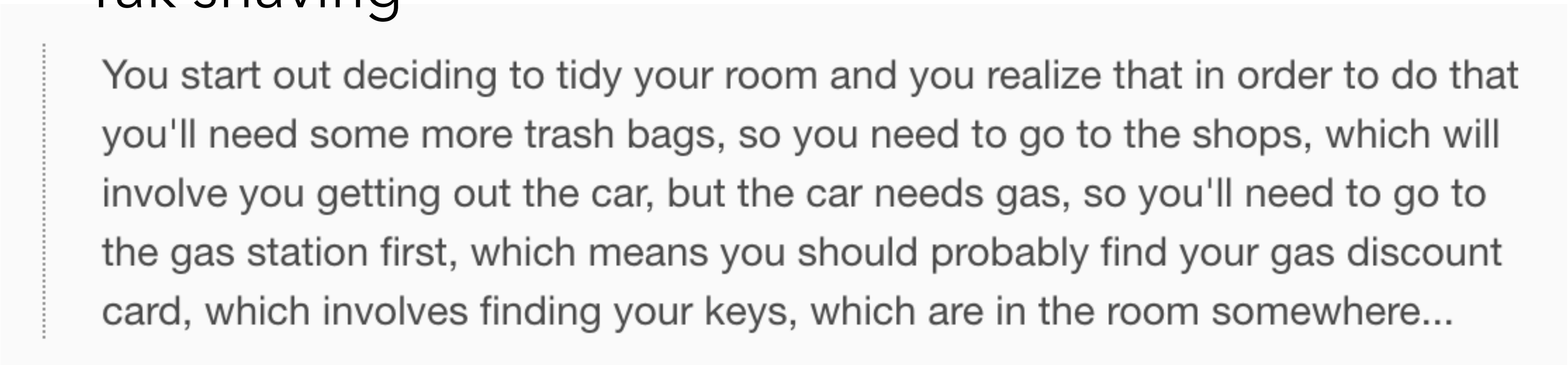
Choosing best (*better) practices

- Scenario #1: How could this RT be 6500 ms?
- Scenario #2: Why don't we have data from participant X? How did it get lost/excluded?
- Scenario #3: The PI/reviewer asks for a completely different analysis/graph
- Scenario #4: Another lab asks you to share the data for a secondary analysis
- Scenario #5: A new grad student wants to know how to run the analyses from that paper

Choosing best (*better) practices

- Difficult problems
  - Will switching to a program make it easier/harder for others on my team to work on the project?
  - I found a better practice, but is there a *best* practice?
  - Is it worth investing the time to change something that already works?

“Yak shaving”



> You start out deciding to tidy your room and you realize that in order to do that you'll need some more trash bags, so you need to go to the shops, which will involve you getting out the car, but the car needs gas, so you'll need to go to the gas station first, which means you should probably find your gas discount card, which involves finding your keys, which are in the room somewhere...

https://joi.ito.com/weblog/2005/03/05/yak-shaving.html

Logistics

Schedule

Week 02, 01/13: Data Workflow

- File organization
- Version control
- Workflow example
- *SKILLS:* Basic I/O functions, getting help, using Github

Week 03, 01/20: Data Structure

- Data types
- Data organization (variables/observations)
- *SKILLS:* Factors/strings, reshaping/aggregating data

Week 04, 01/27: Automation

- Copy/paste is evil (automating data pipeline)
- Drop-down menus are evil (automating your analyses)
- Machine-readable formats
- *SKILLS:* Writing functions, iteration, vectors
- **WORKFLOW PRESENTATIONS GROUP 1**

Week 05, 02/03: Error Checking

- Data validation
- Automating visualizations
- Exploratory data analysis
- *SKILLS:* Basic visualizations, handling missing data
- **WORKFLOW PRESENTATIONS GROUP 2**

Week 06, 02/10: Reuse Part 1: Encapsulation and Writing Functions

- Best option: Use existing APIs and packages
- Next best option: Writing your own general-purpose functions
- Not great option: Writing overly-specific functions that you can never use again
- *SKILLS:* Sourcing functions, parameterization

Week 07, 02/17: Reuse Part 2: Data Sharing, Archiving, and Documentation

- Code as documentation
- Reuse-minded project management
- *SKILLS:* Comments, specifications, licenses

Week 08, 02/24: Communication Part 1: Reproducible Reports

- Reproducible reports
- Preserving programming environment and analyses
- *SKILLS:* R Markdown, papaja, package control

Week 09, 03/03: Communication Part 2: Visualization

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

Week 10, 03/10: Project Presentations

- How have you changed your workflow?
- What have you learned about your data?
- What problems are still unresolved?

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 workflow 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
  - Jake office hours Tuesdays 11am-12pm
  - My office hours by appointment

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?
  - Due week 4; present in class week 4 or 5

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 or papaja) that communicates exploratory or confirmatory analyses with visualizations in R
  - Presentation in week 10 to summarize your changes and describe what problems you solved (and have yet to solve)

RStudio Cloud

- Getting started
  - Use link in syllabus to sign up and create an account
  - Afterwards, join the class workspace
- How we will use it
  - Sharing lecture examples
  - Assignments will be posted here; create a copy to complete the assignment
  - Jake and I can view it to count it as complete/give help as needed

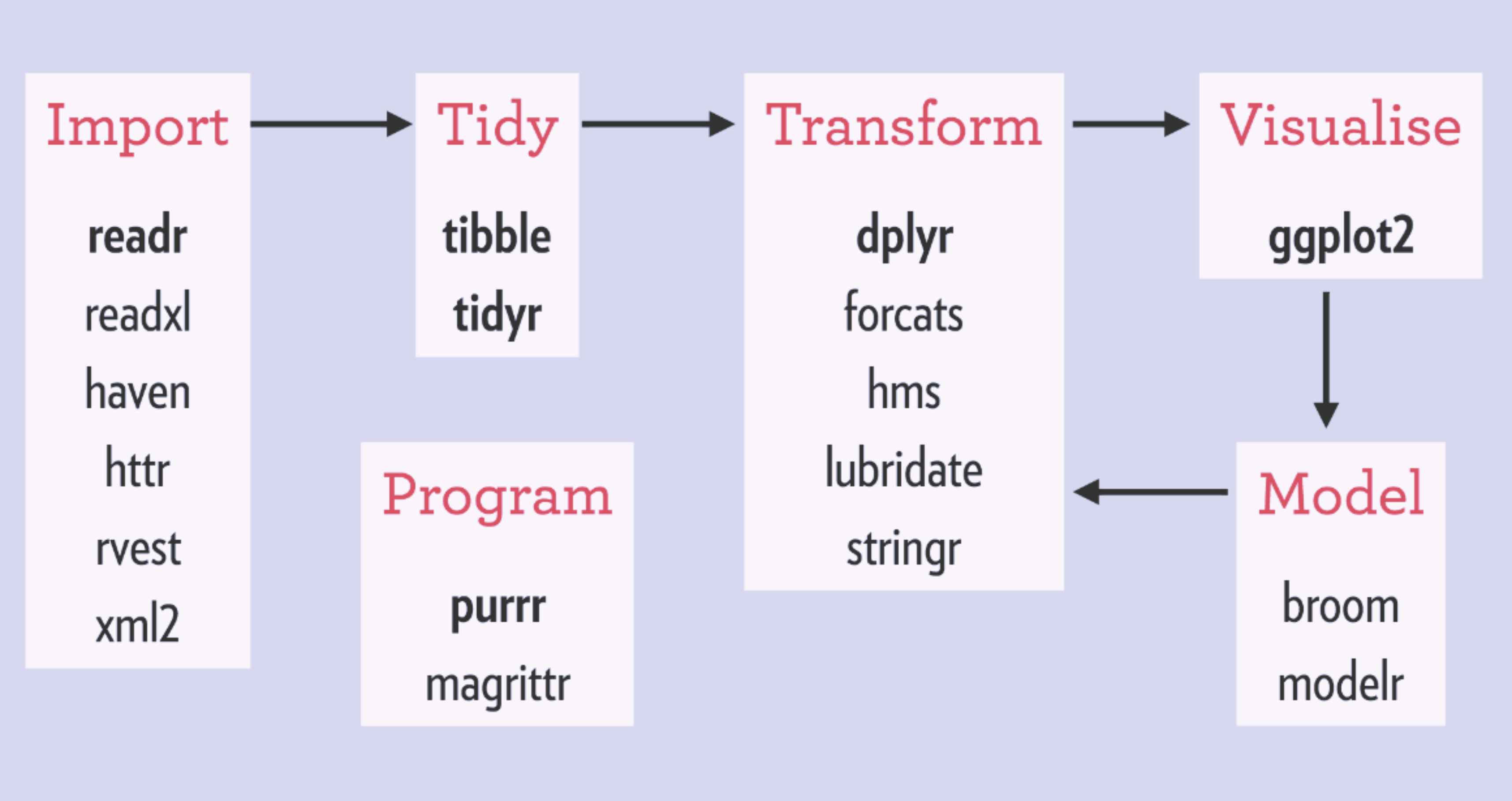
Class Github page

- https://github.com/PSYC-259-Data-Science
- 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 Teetor

- Use whatever works!

For next week

Before next class…

- *R4DS* readings (link in syllabus)
- Rouder article (to discuss, pdf is on Github)
- Make sure you've joined the class RStudio Cloud page
- Finish intro R lessons through swirl() if needed

Before next class…

- Read Github beginner guide (link in syllabus)
- Follow the directions in software_setup.md to get R, RStudio, git, and GitHub set up on your own computer
  - Link is in the updated syllabus

Right now

- Changing class time?
- Stay on if you need help getting onto R Studio Cloud, accessing swirl() lessons, or have other questions