Visualising Data Wrangling Operations: Joins and Reshaping

CHARCO HUI

SUPERVISORS: ANNA FERGUSSON, CHRIS WILD

Question

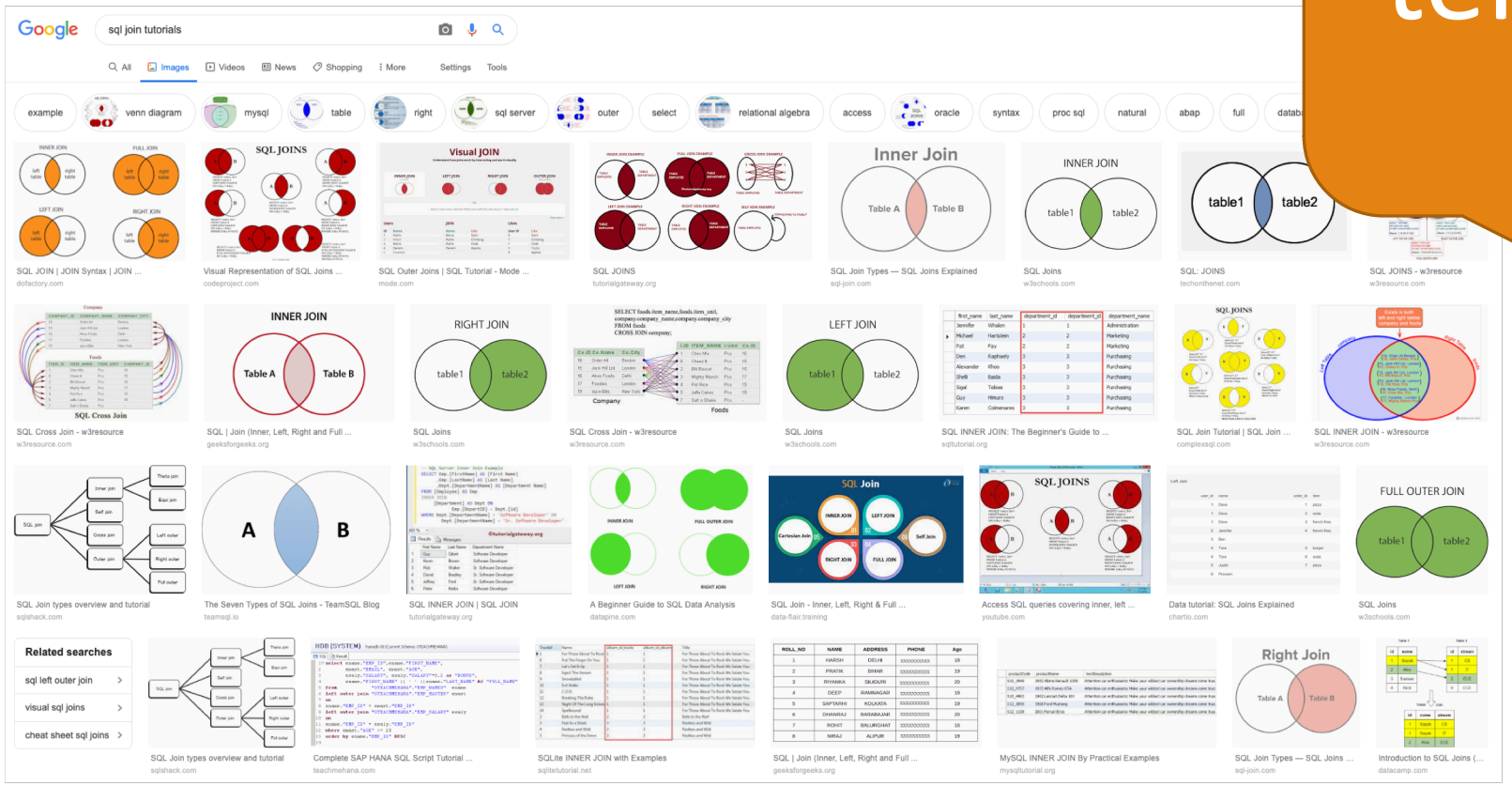
What is Left Join?

What joins are about

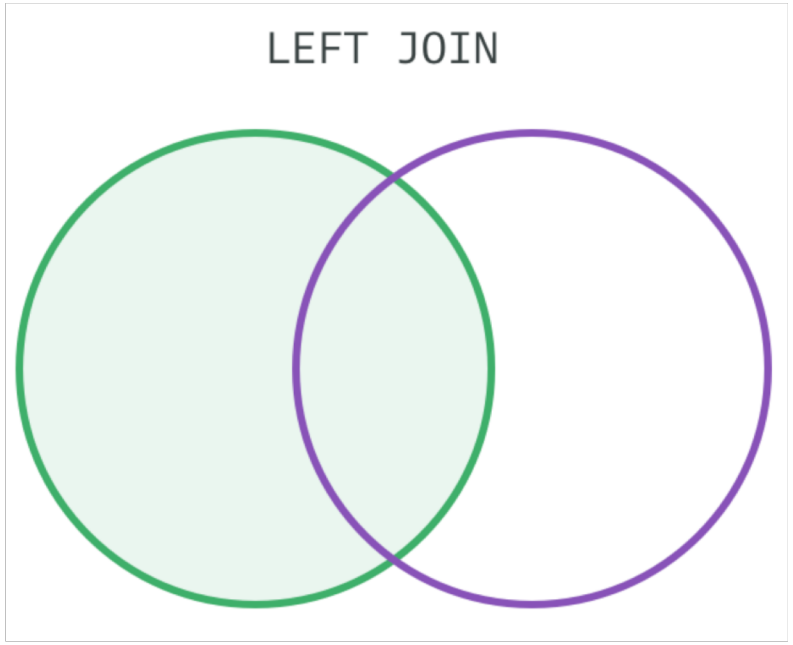
- Combining information from different data tables by using key columns to match records.
- Different type of joins
  - Left Join
  - Right Join
  - Inner Join
  - Complete Join

Most Popular Visualisatio

What do these tell you about a join?



Current Approach 1



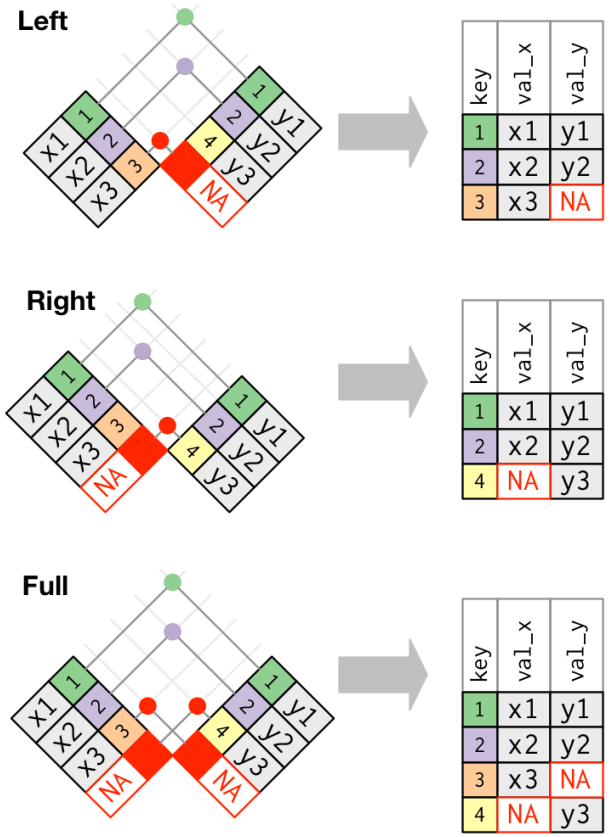
Limitations:

- Don’t look like data tables.
- Vague on how the joins work.
- Does not explain unmatched rows or multiple matches.

Useful Features:

- Useful reminder about what rows get used.

Current Approach 2



R for Data Science Approach
(Wickham and Grolemund, 2017)

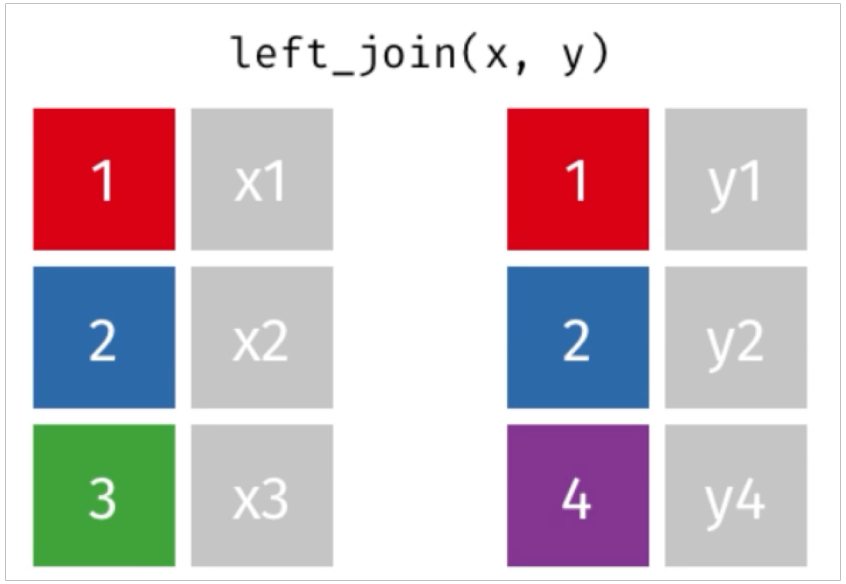
Advantages compared to Venn Diagrams:

- Each table and rows are shown nicely.
- Line linking rows that are matched together.
- Joined table shown.

What is missing:

- No verbal descriptions.
- Does not explain multiple matches.
- Would be messy for larger tables.

Current Approach 3



Left join Animation – **tidyexplain** (Aden-Buie, 2018)

Limitations:

- Lack explanations of how unmatched and duplicated rows are treated.
- No explanations of how rows are joined together.

Advantages:

- Animation catches attention.

Improvements

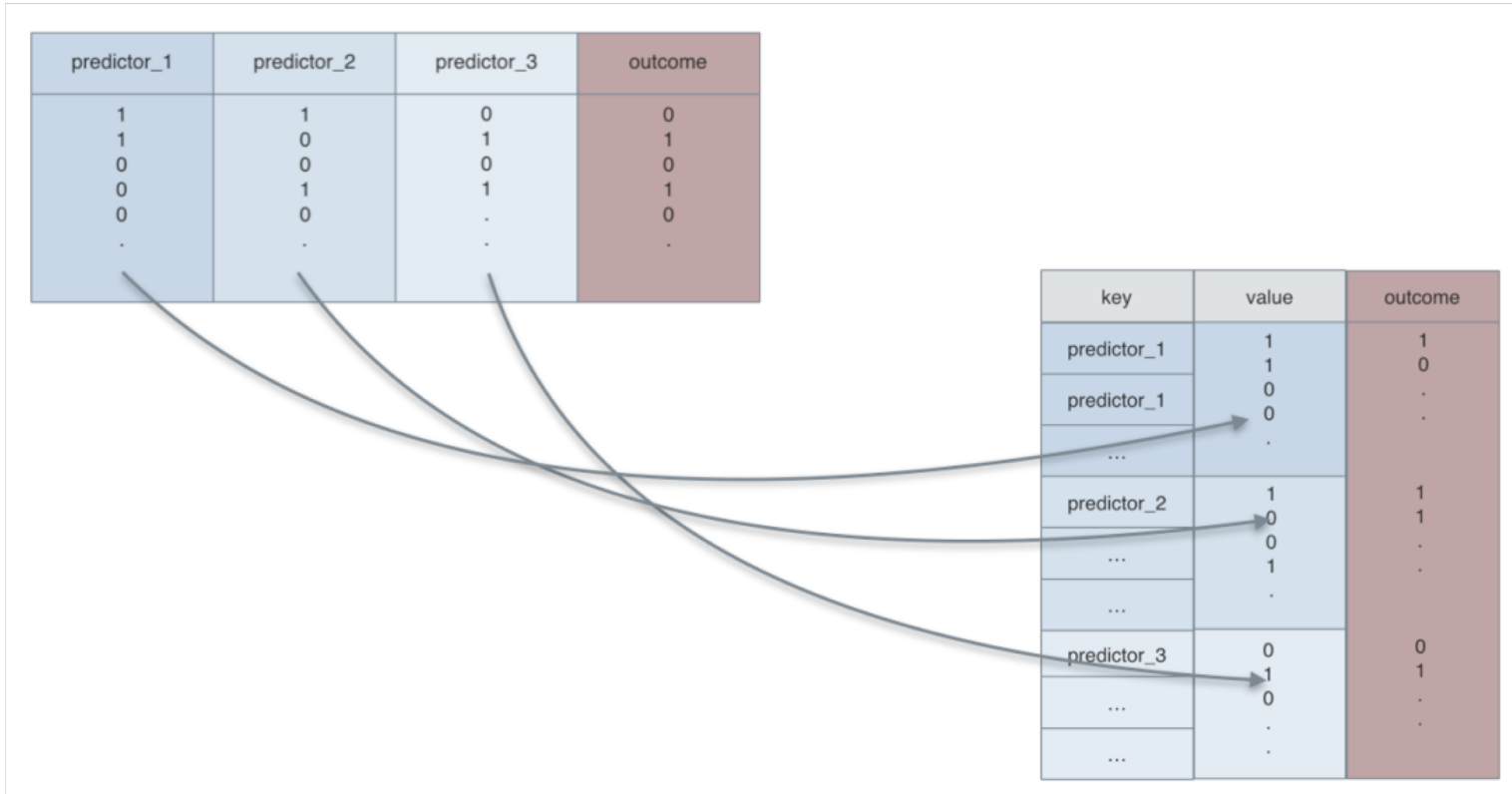
- Show why or how rows are joined together.
- Show how duplicated, unmatched and multiple matching rows are joined together.
- Convey the idea of key columns.
- Displaying messages to show the logic behind each step of the joins.
- Allow users to input their own data to produce the animations.
- Control playback speed.
- Ability to save an animation for use in webpages and presentations.

dataAnim Approach

Name	Gender	Height
Alex	M	140
Ben	M	150
Sam	F	160
Sarah	M	170
Emily	F	180
Hannah	F	190

Age	Name	Weight
30	Ben	40
17	Alex	50
20	Tony	60
21	Ben	70
45	Sarah	80
32	Sean	90

Data Transformation



Data transformation (Sauser, 2016)

dataAnim Approach
Reshaping

Make a new column Subject to contain the old column names

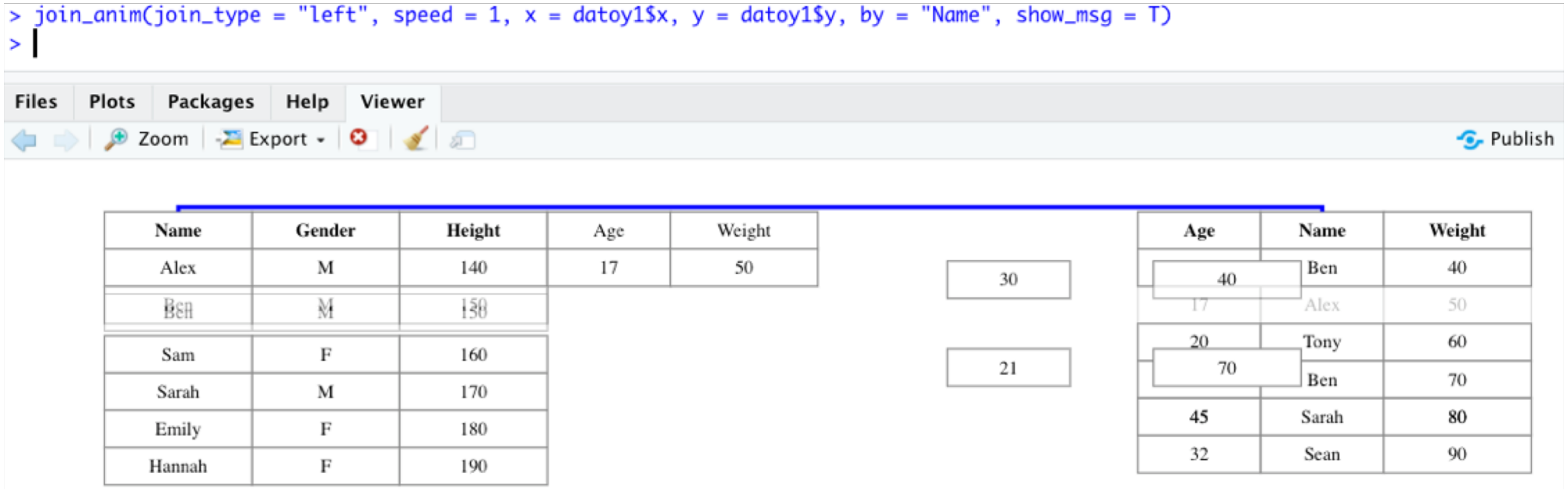
Name	English	Maths
Alex	40.3	39.3
Ben	46.6	68.7
Sam	28.1	43.4

dataAnim Package

Rstudio

Installing dataAnim:

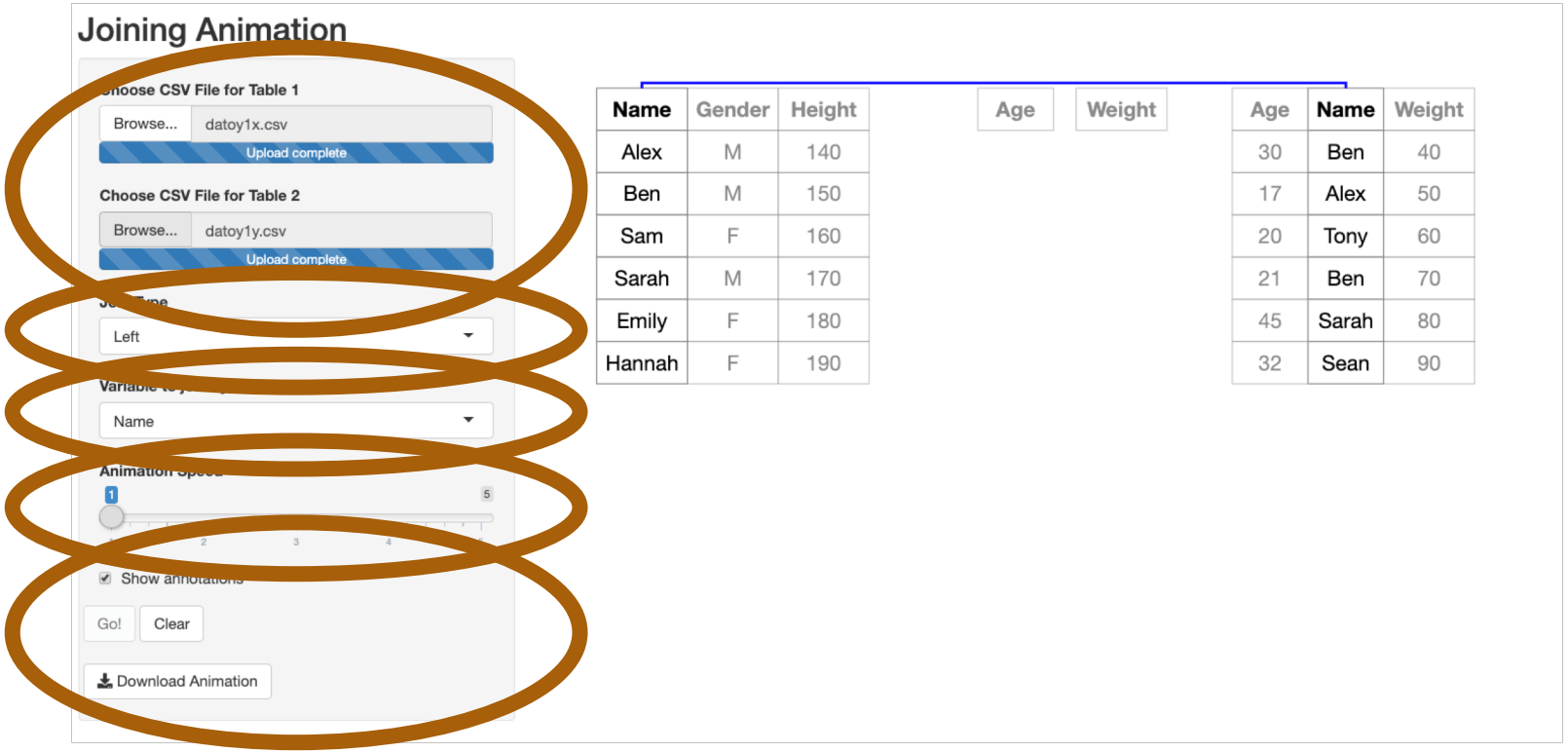
devtools::install_github("chrk623/dataAnim")



dataAnim Package

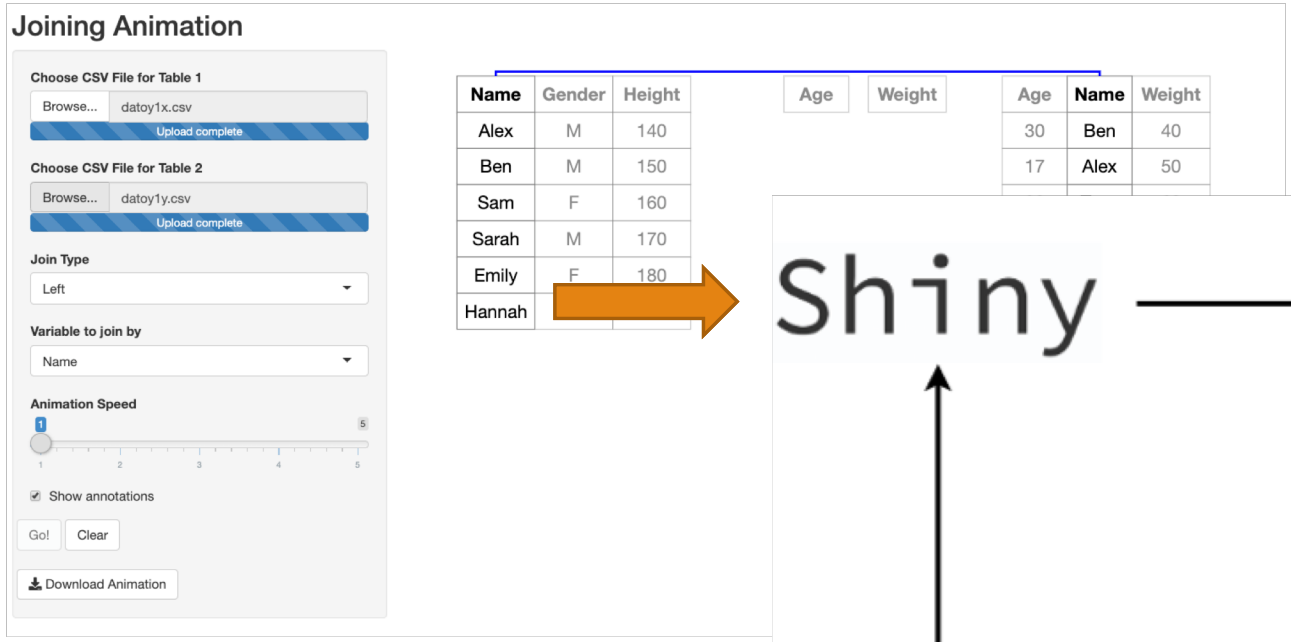
Shiny App

Shiny App: https://chrk623.shinyapp.io/dataAnim_shiny



dataAnim Package

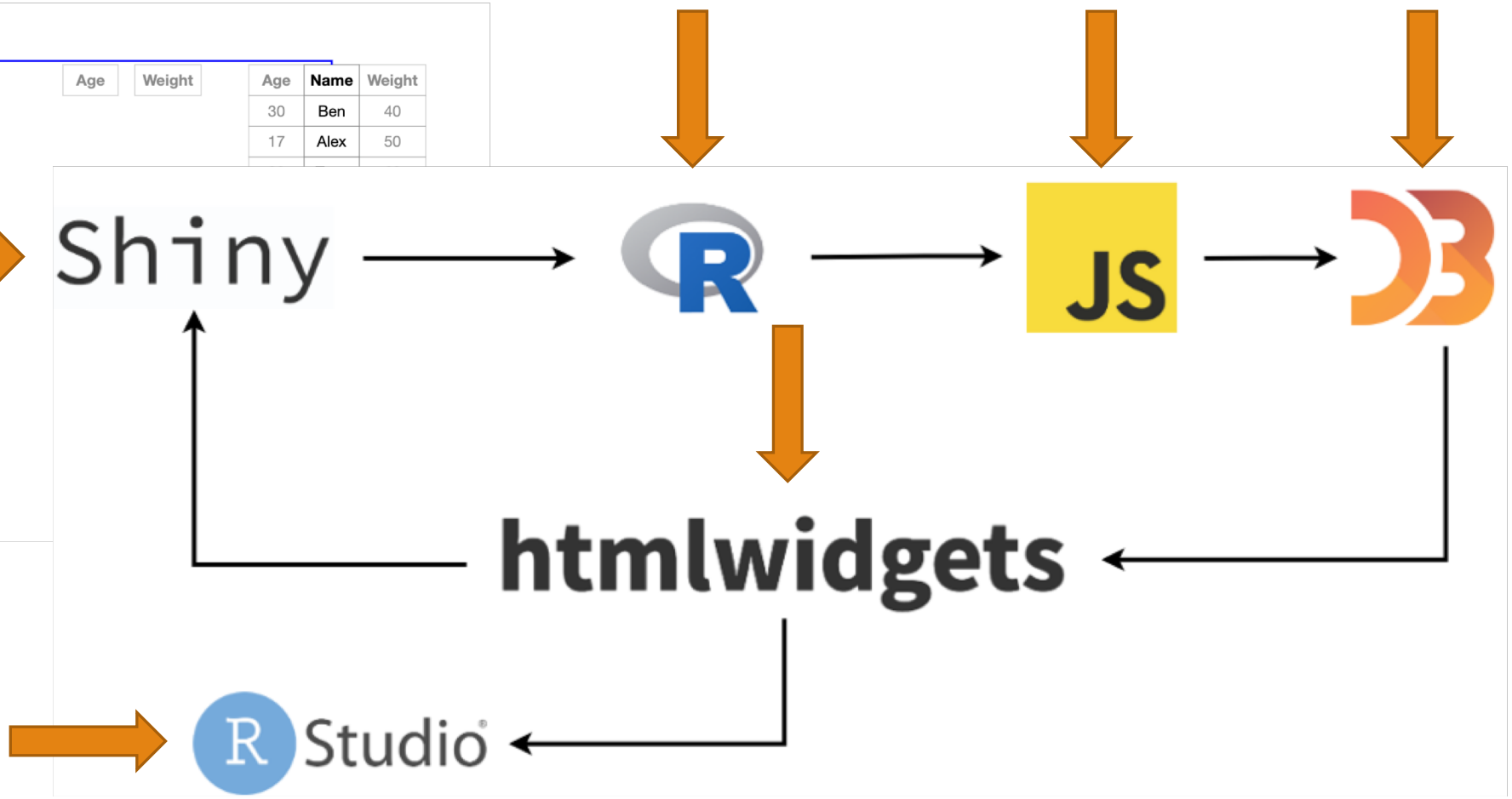
Software Structure



Name	Gender	Height
Alex	M	140
Ben	M	150
Sam	F	160
Sarah	M	170
Emily	F	180
Hannah		

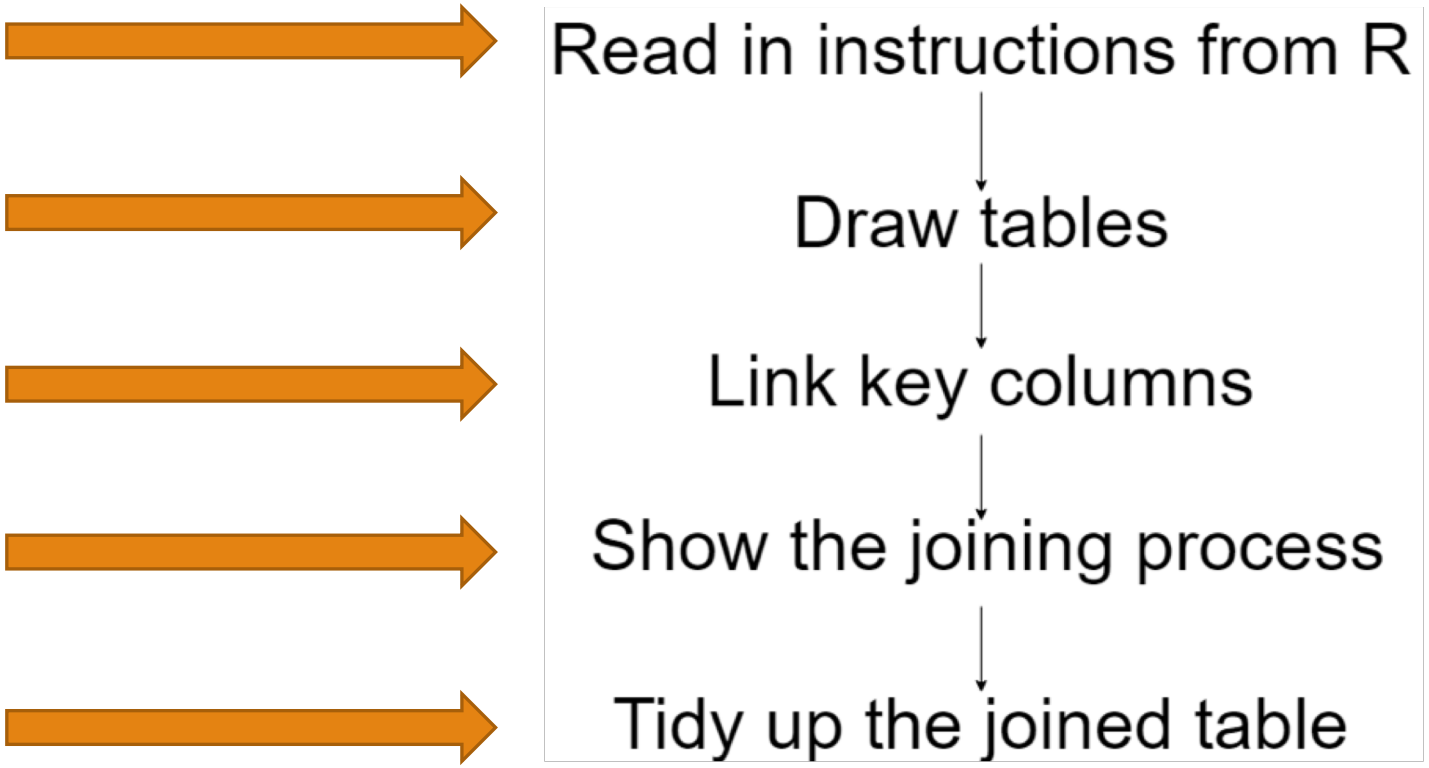
Age	Weight

Age	Name	Weight
30	Ben	40
17	Alex	50



dataAnim Package

JavaScript Program Flow



Conclusion

dataAnim Package:

- Different techniques used to catch users attention.
- Messages to explain steps of different data wrangling operations.

Shiny App:

- Extension to a shiny app for ease of use.
- Used in the help page of iNZight.

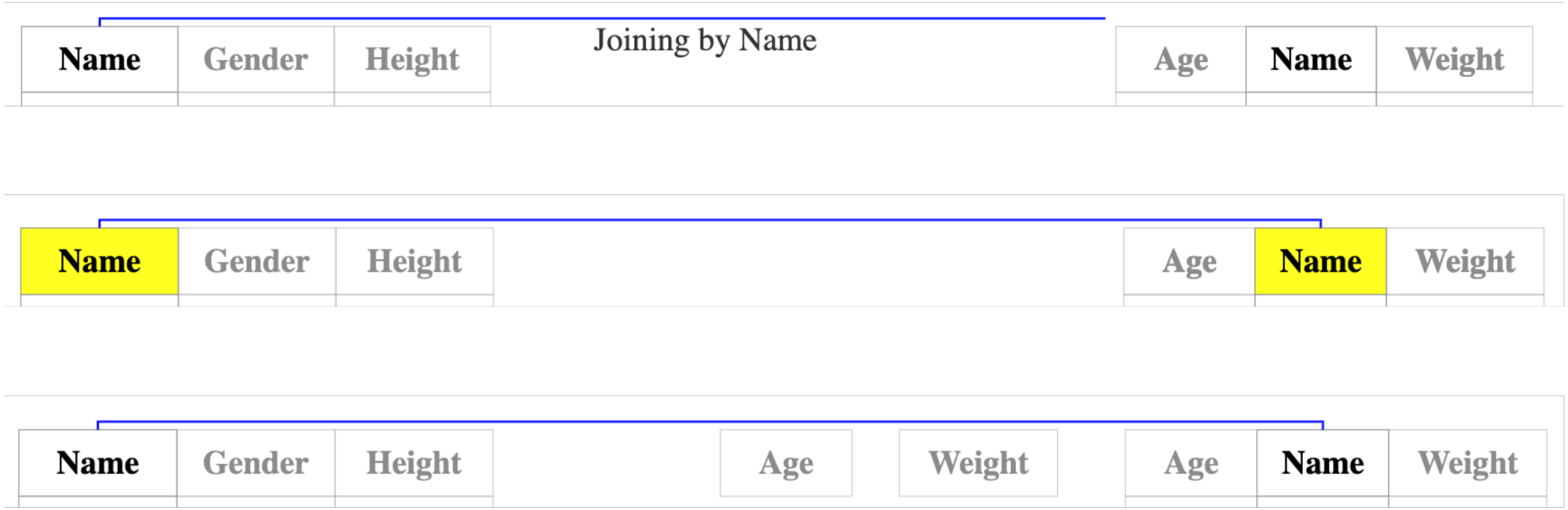
Future work:

- Testing on students/teachers.
- Allow for more complicated situations in the software.
- Export as movie.
- SQL module.

Thank You.

dataAnim Left Join

Key column



dataAnim Left Join

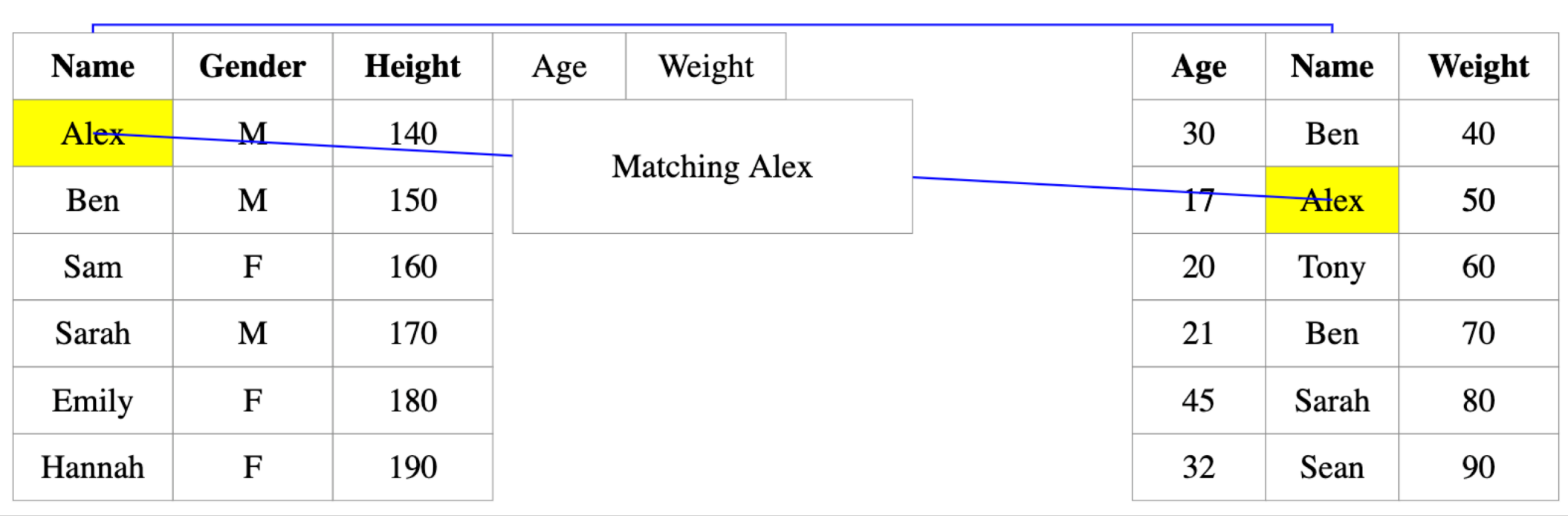
Single match – Step 1

Name	Gender	Height	Age	Weight
Alex	M	140		
Ben	M	150		
Sam	F	160		
Sarah	M	170		
Emily	F	180		
Hannah	F	190		

Age	Name	Weight
30	Ben	40
17	Alex	50
20	Tony	60
21	Ben	70
45	Sarah	80
32	Sean	90

dataAnim Left Join

Single Match – Step 2



dataAnim Left Join

Single match – Step 3

Name	Gender	Height	Age	Weight
Alex	M	140		
Ben	M	150		
Sam	F	160		
Sarah	M	170		
Emily	F	180		
Hannah	F	190		

17	50

Age	Name	Weight
30	Ben	40
17	Alex	50
20	Tony	60
21	Ben	70
45	Sarah	80
32	Sean	90

dataAnim Left Join

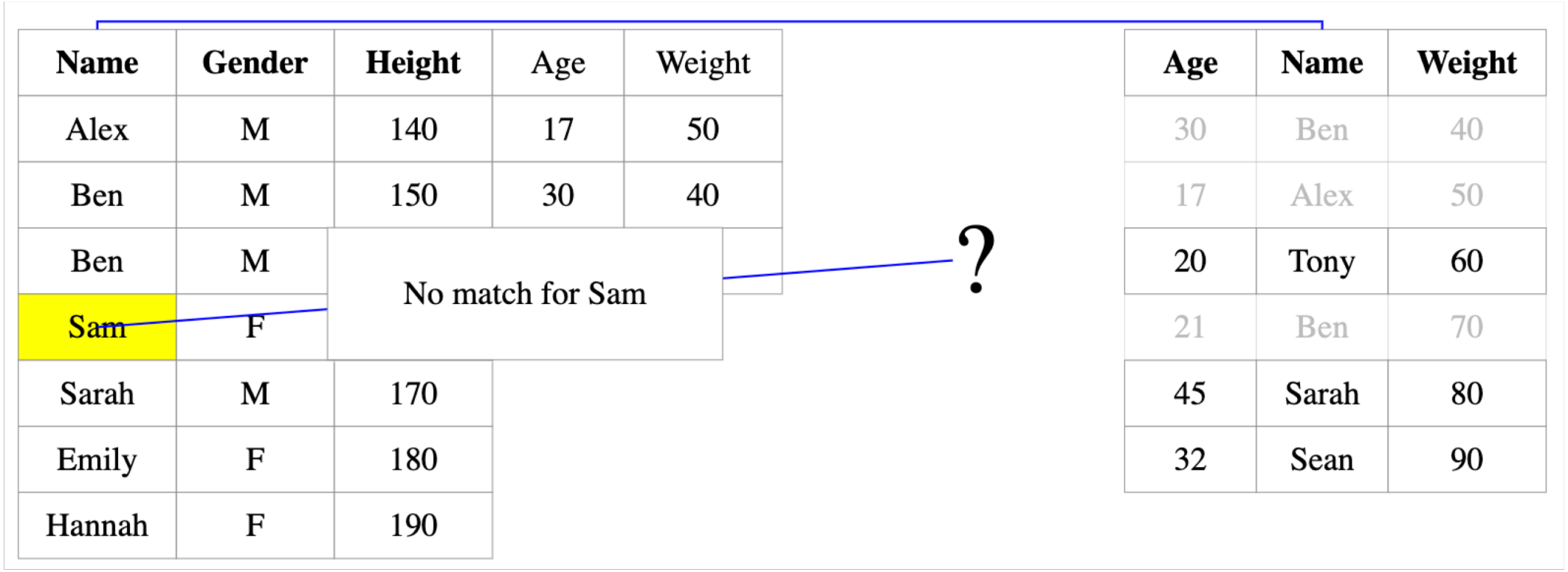
Single Match – Step 4

Name	Gender	Height	Age	Weight
Alex	M	140	17	50
Ben	M	150		
Sam	F	160		
Sarah	M	170		
Emily	F	180		
Hannah	F	190		

Age	Name	Weight
30	Ben	40
17	Alex	50
20	Tony	60
21	Ben	70
45	Sarah	80
32	Sean	90

dataAnim Left Join

No Match



dataAnim Left Join

No Match – Step 2

Name	Gender	Height	Age	Weight
Alex	M	140	17	50
Ben	M	150	30	40
Ben	M	150	21	70
Sam	F	160	NA	NA
Sarah	M	170		
Emily	F	180		
Hannah	F	190		

Age	Name	Weight
30	Ben	40
17	Alex	50
20	Tony	60
21	Ben	70
45	Sarah	80
32	Sean	90

dataAnim Left Join

Multiple Match – Step 1

dataAnim Left Join

Multiple Match – Step 2

Name	Gender	Height	Age	Weight
Alex	M	140	17	50
Ben	M	150		
Sam	F	160		
Sarah	M	170		
Emily	F	180		
Hannah	F	190		

30

21

Age	Name	Weight
30	40	40
17	Alex	50
20	Tony	60
21	70	70
45	Sarah	80
32	Sean	90

dataAnim Left Join

End of joining

Name	Gender	Height	Age	Weight
Alex	M	140	17	50
Ben	M	150	30	40
Ben	M	150	21	70
Sam	F	160	NA	NA
Sarah	M	170	45	80
Emily	F	180	NA	NA
Hannah	F	190	NA	NA

Age	Name	Weight
30	Ben	40
17	Alex	50
20	Tony	60
21	Ben	70
45	Sarah	80
32	Sean	90

dataAnim Left Join

Tidy up

Name	Gender	Height	Age	Weight
Alex	M	140	17	50
Ben	M	150	30	40
Ben	M	150	21	70
Sam	F	160	NA	NA
Sarah	M	170	45	80
Emily	F	180	NA	NA
Hannah	F	190	NA	NA