Accessing MIDFIELD

A workshop for R beginners

Richard Layton, Russell Long, Susan Lord, Matthew Ohland, Marisa Orr

2019–10–16

FIE Conference, Cincinnati, OH

Accessing MIDFIELD: A workshop for R beginners

Look for the conference wifi

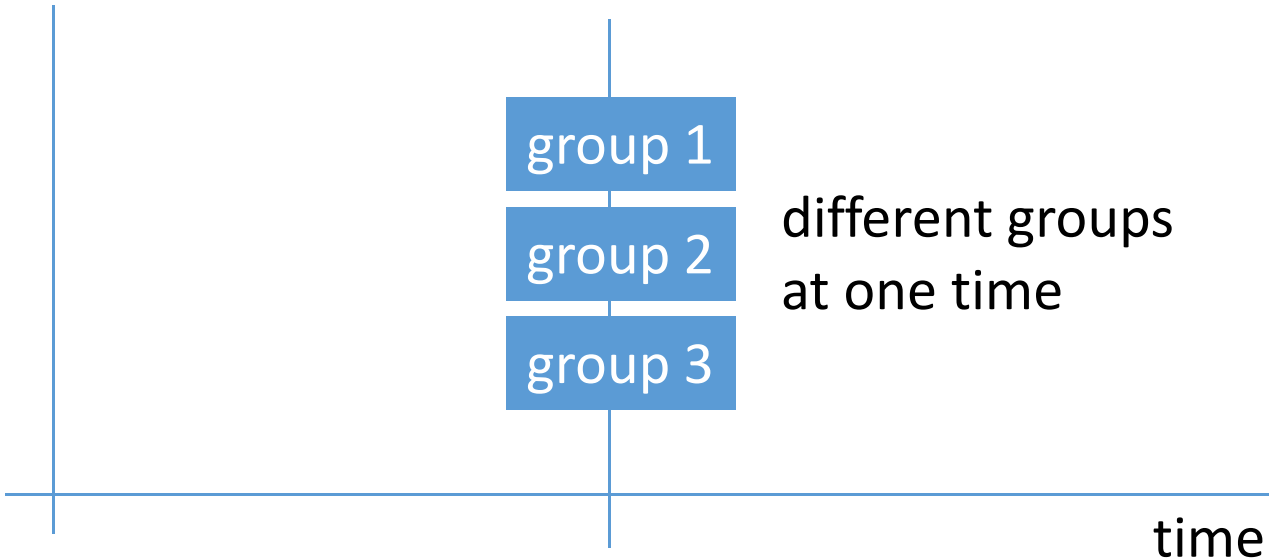
password: OCT2019

R-Bar volunteers can help with software issues

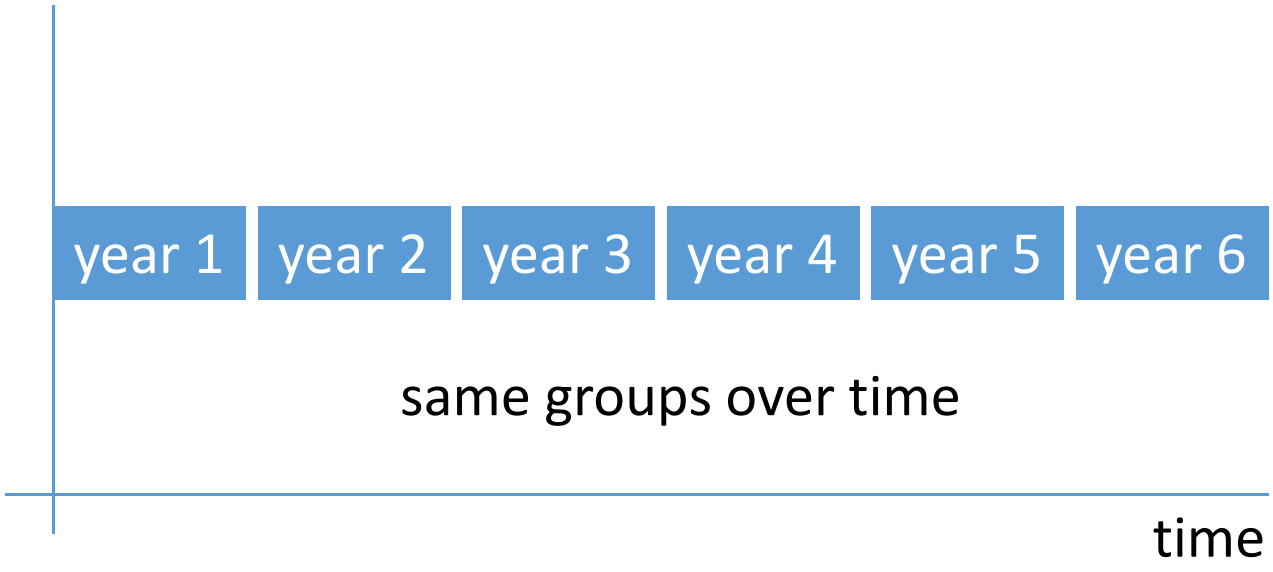
Min	Topic
10	Introductions
5	The MIDFIELD data structures
30	Finding stories in the data
35	Starting with R (tutorial)
15	— break —
20	Elements of effective graphs
5	Introducing midfieldr
50	Starting with midfieldr (tutorial)
10	Next steps

The MIDFIELD data structures

In education, cross-sectional designs are typical



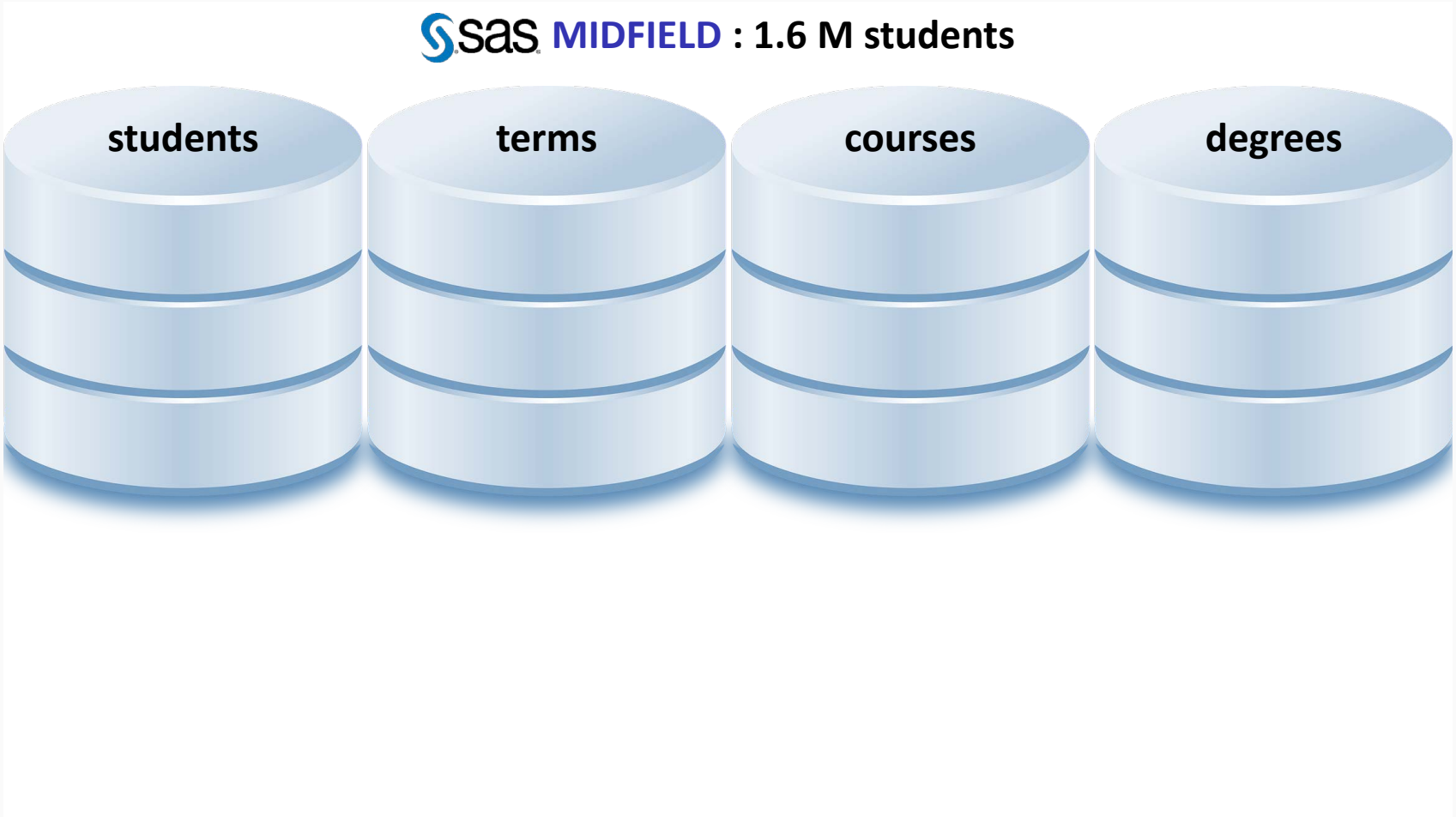
Longitudinal studies offer some advantages



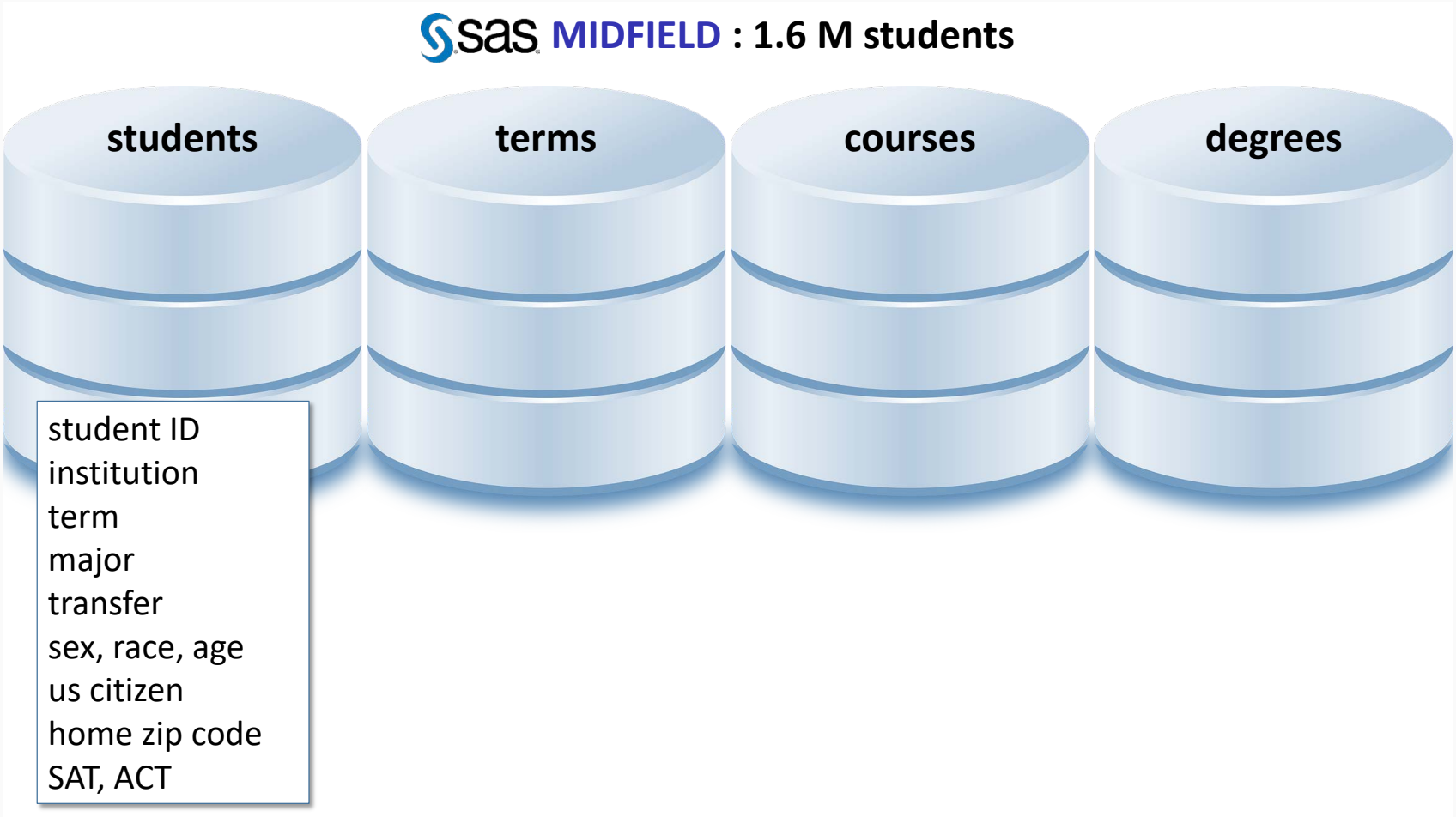
MIDFIELD is a database for longitudinal studies

- 1.6 M undergraduate students at 22 US institutions
- whole-population data from registrars
- 1987–present

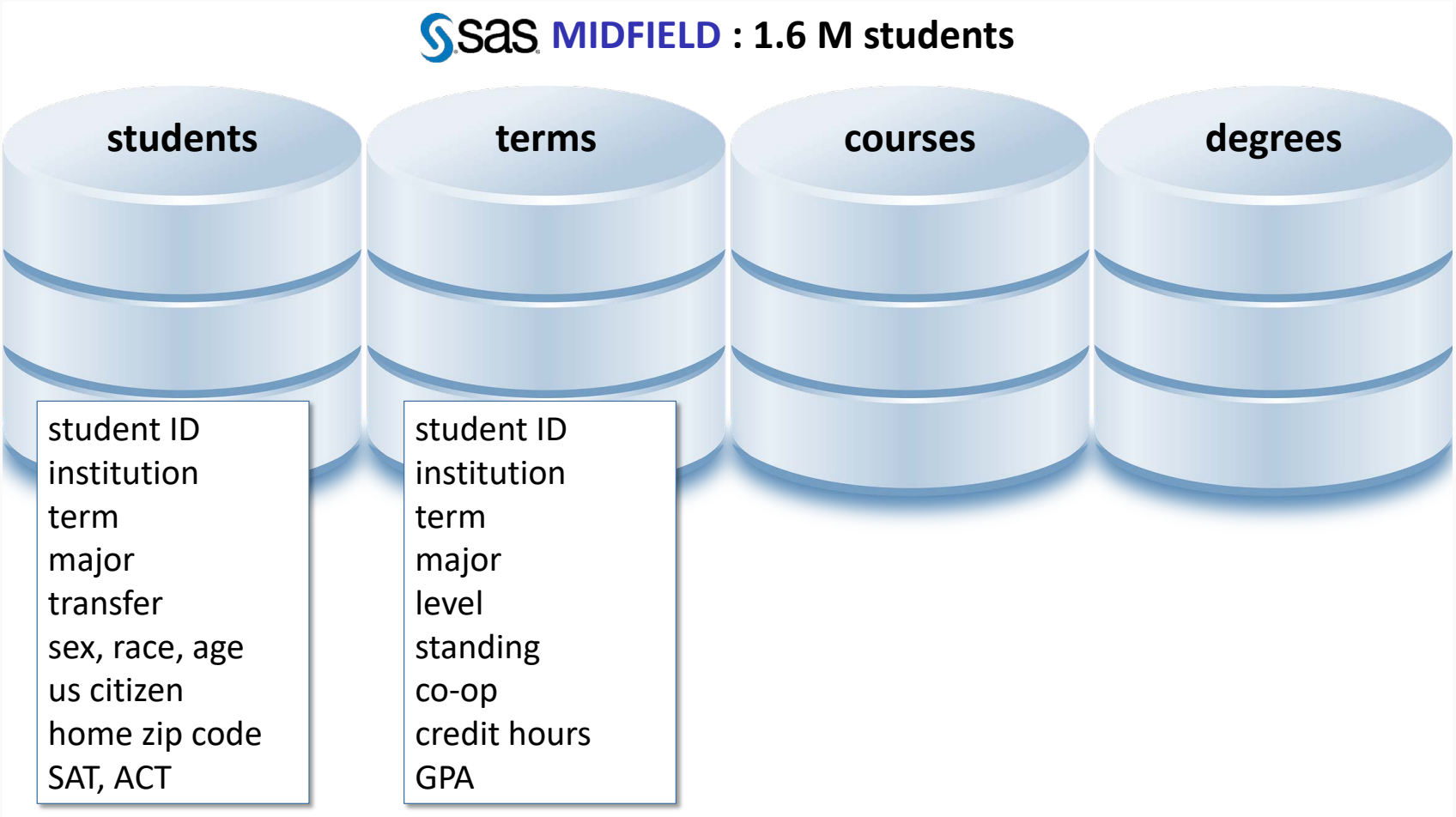
MIDFIELD data are curated in four categories



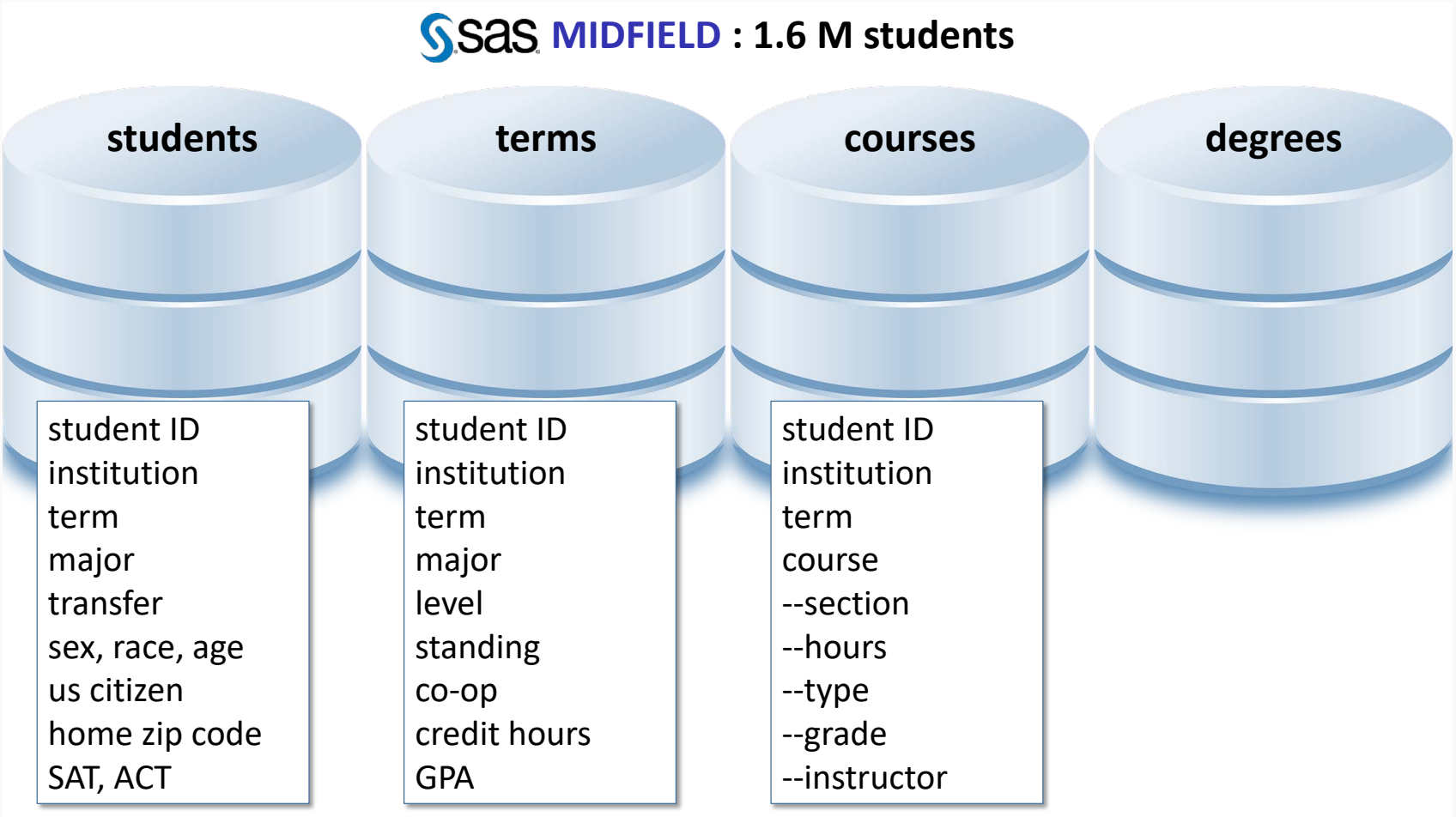
Each observation is a unique student



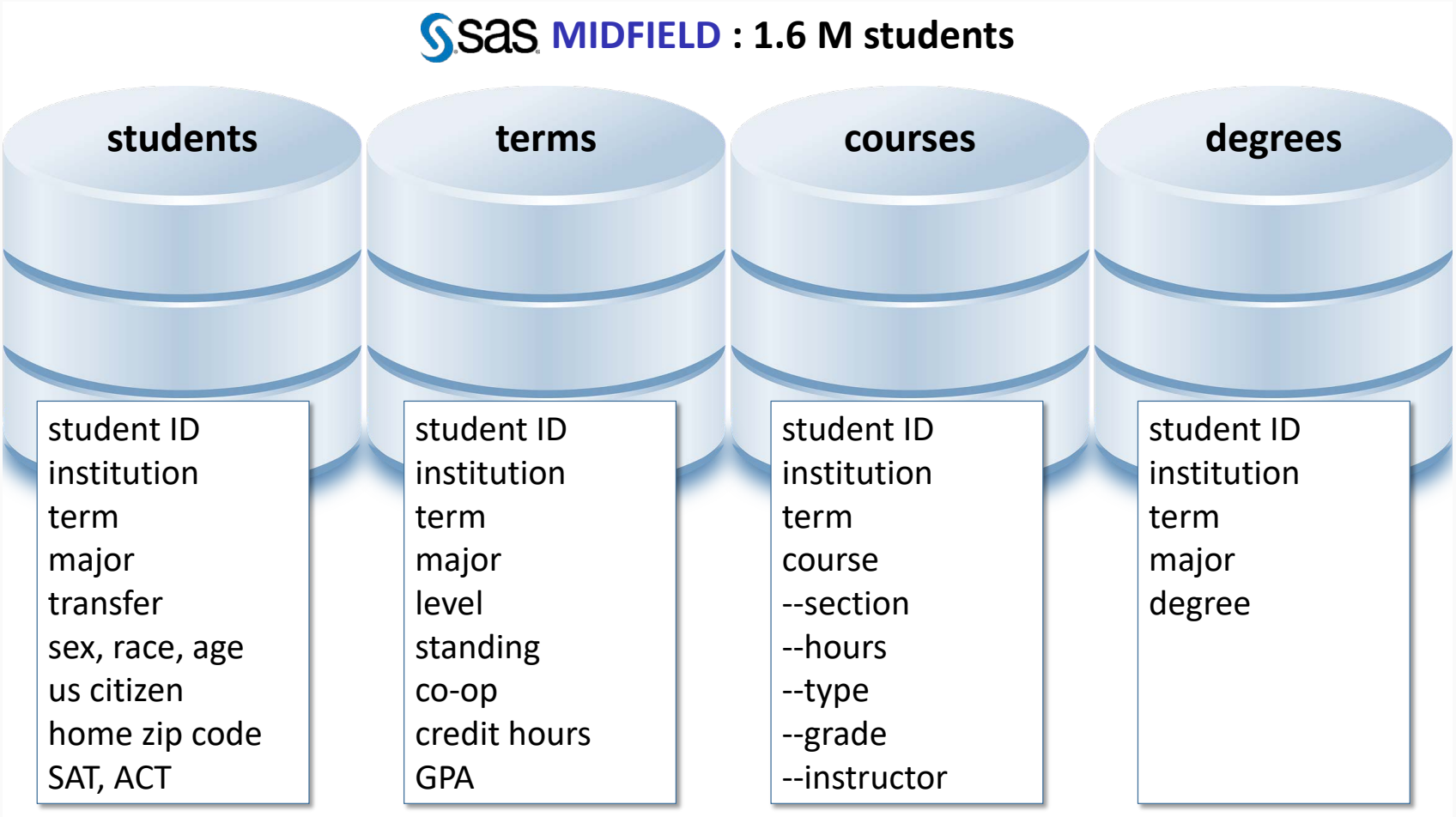
Each observation is one term for one student



Each observation is one course for one student



Each observation is a unique student



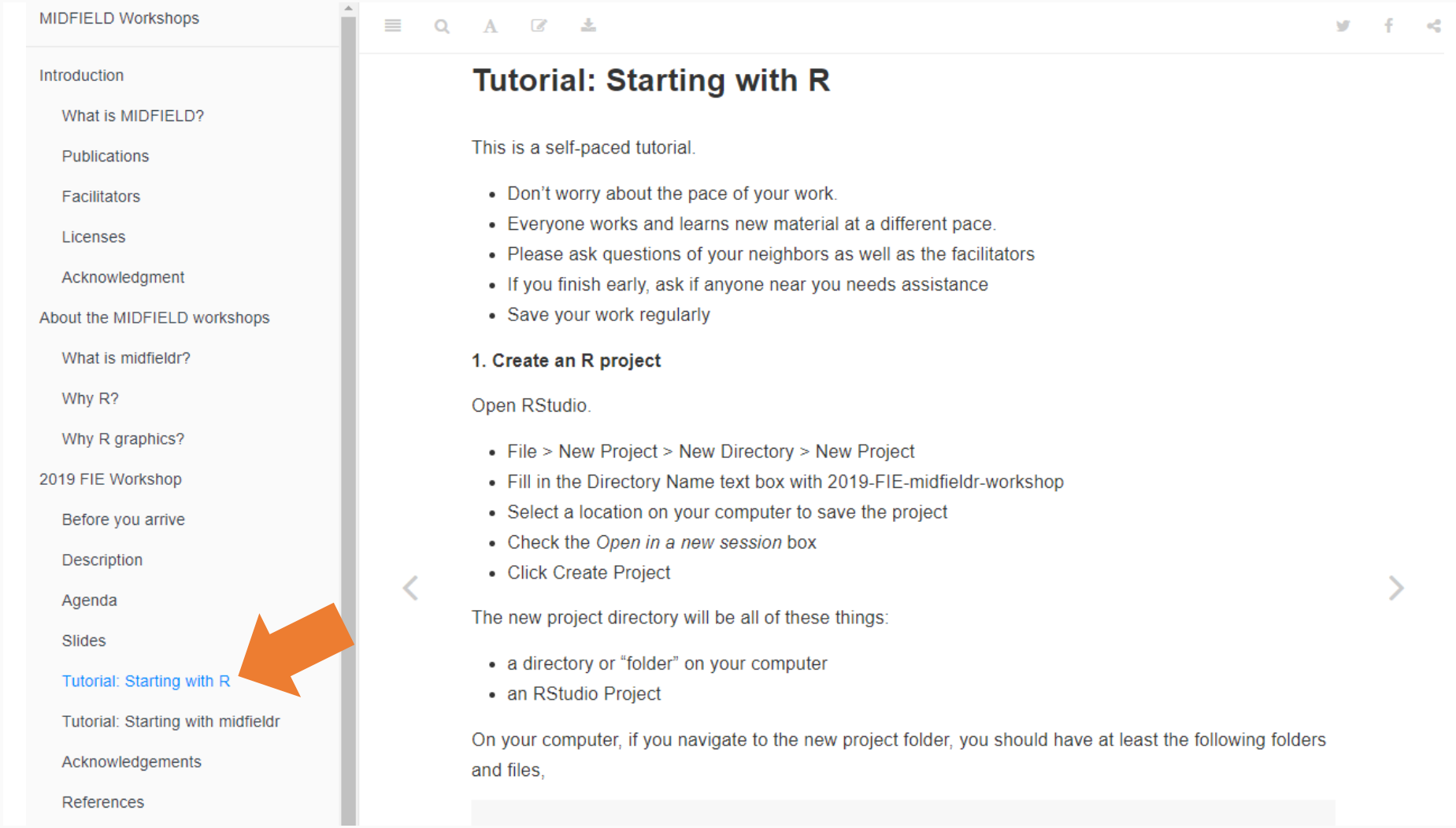
Finding stories in the data

Starting with R (tutorial)

This self-paced tutorial introduces basic R

- Don’t worry about the pace of your work.
- Everyone works and learns new material at a different pace.
- Please ask questions of your neighbors as well as the facilitators
- If you finish early, ask if anyone near you needs assistance
- Save your work regularly

https://midfieldr.github.io/2019-FIE-workshop/

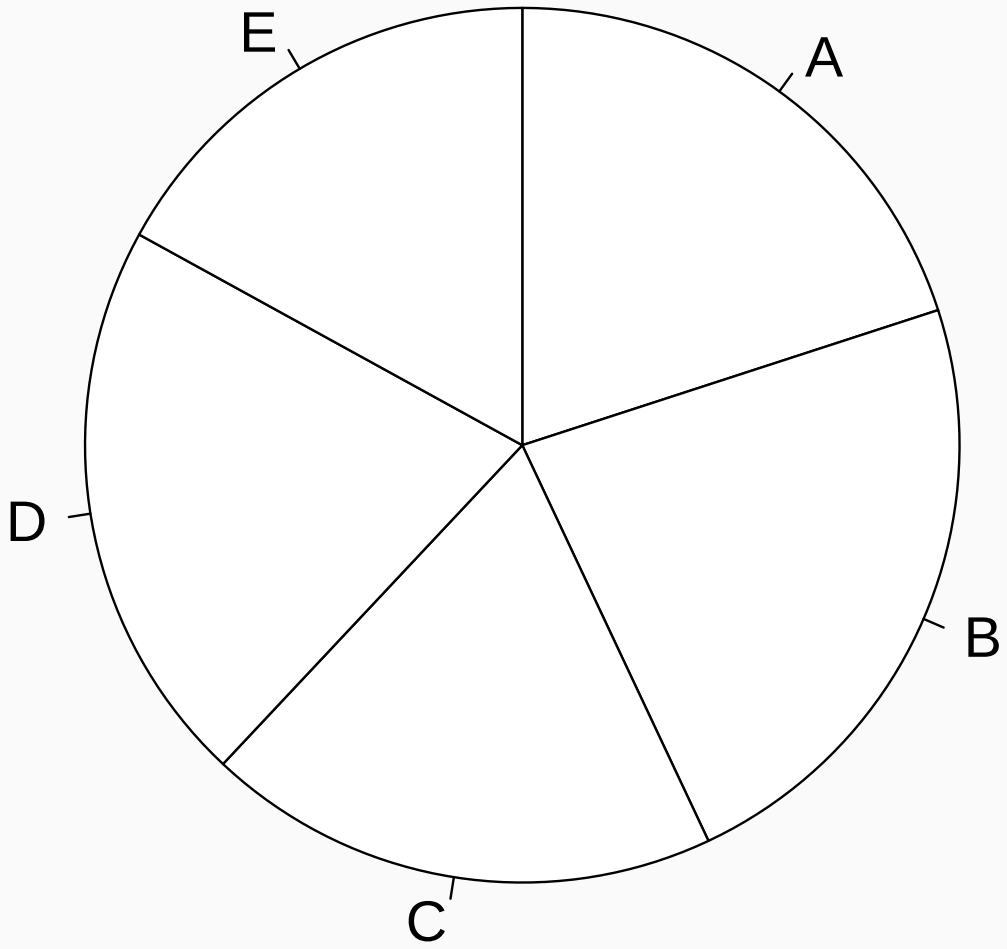


Create an R project, start an R script, add code, rinse, repeat.

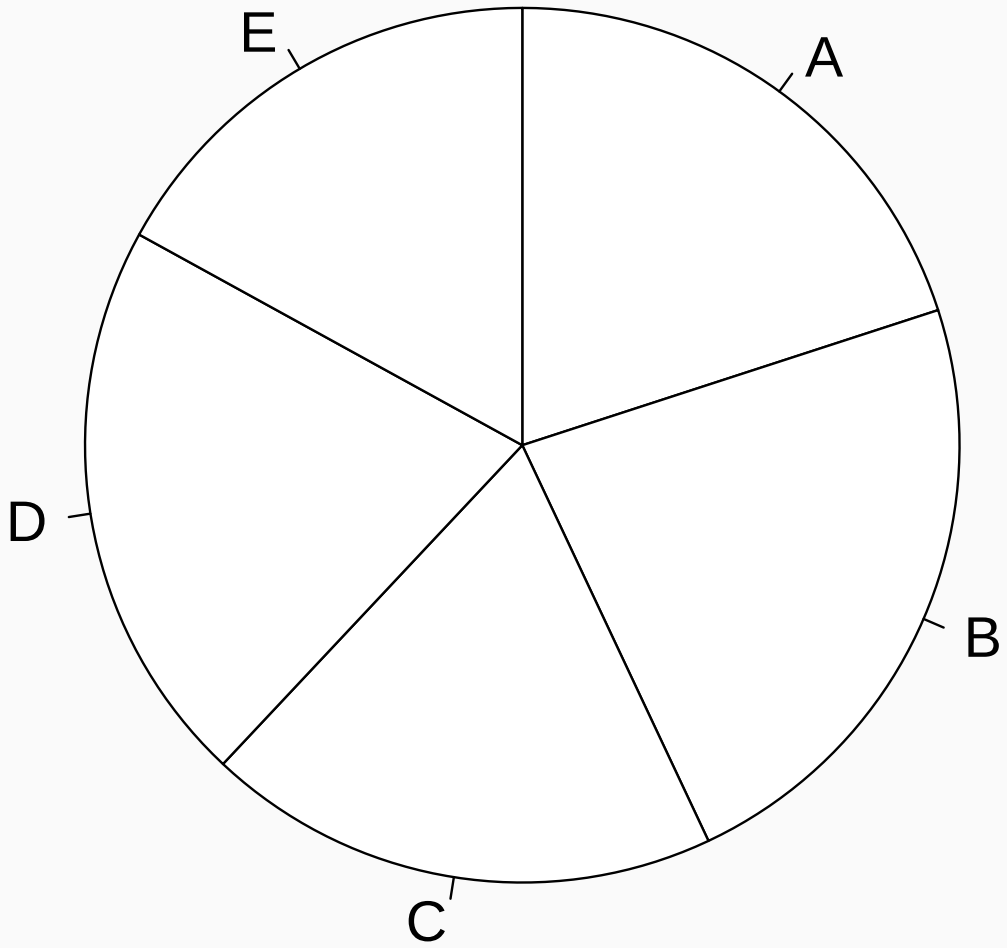
— break —

Elements of effective graphs

In your handout, list the slices A thru E from largest to smallest

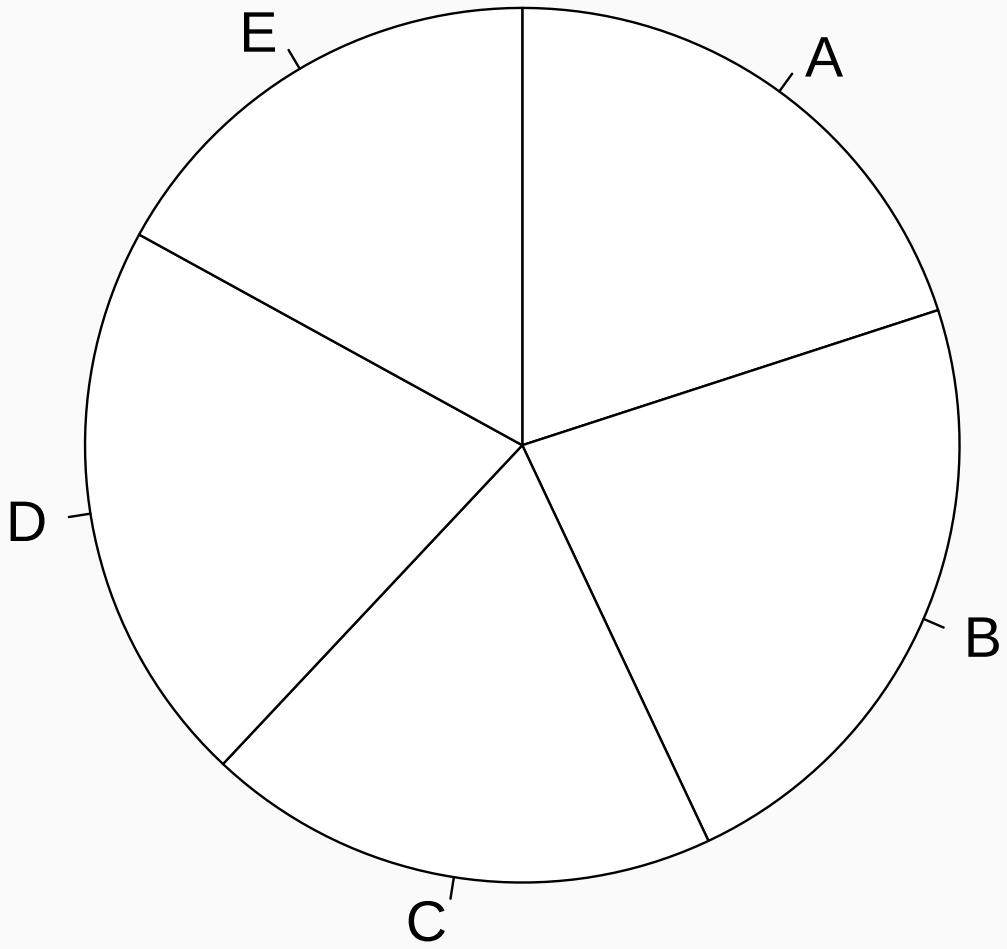


In your handout, list the slices A thru E from largest to smallest



- B (largest)

In your handout, list the slices A thru E from largest to smallest

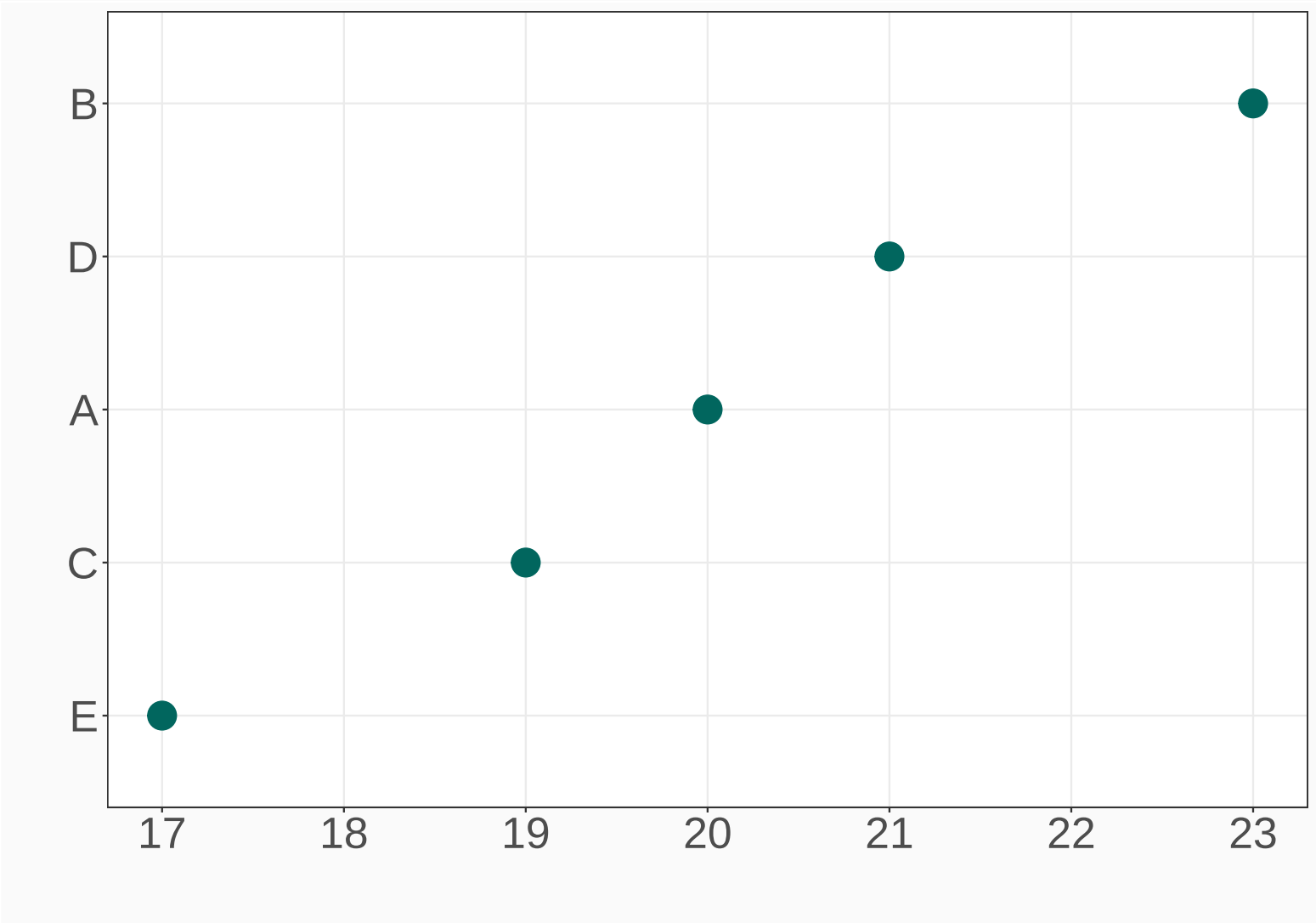


- B (largest)
- D
- A
- C
- E (smallest)

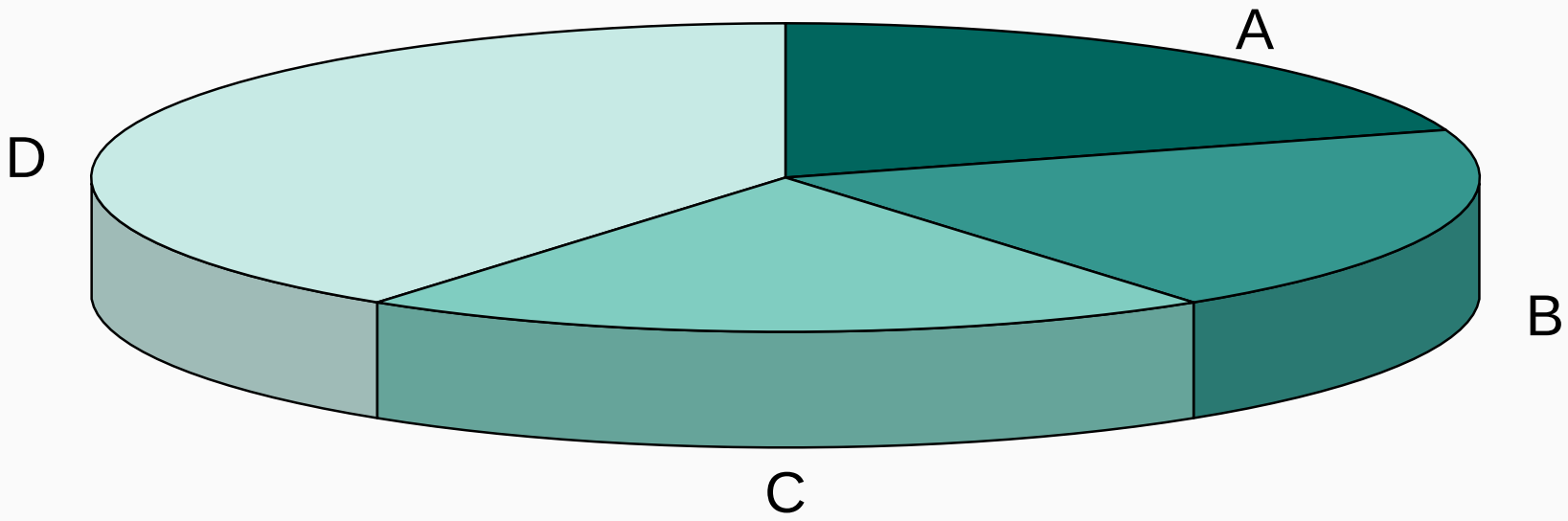
Adapted from (Robbins 2013) Ch. 2

The same data arranged along a common axis

Comparing values along a common axis is a high-accuracy visual task.



Slices are what percentage of the whole?



Fill in the blanks

The total should be 100%

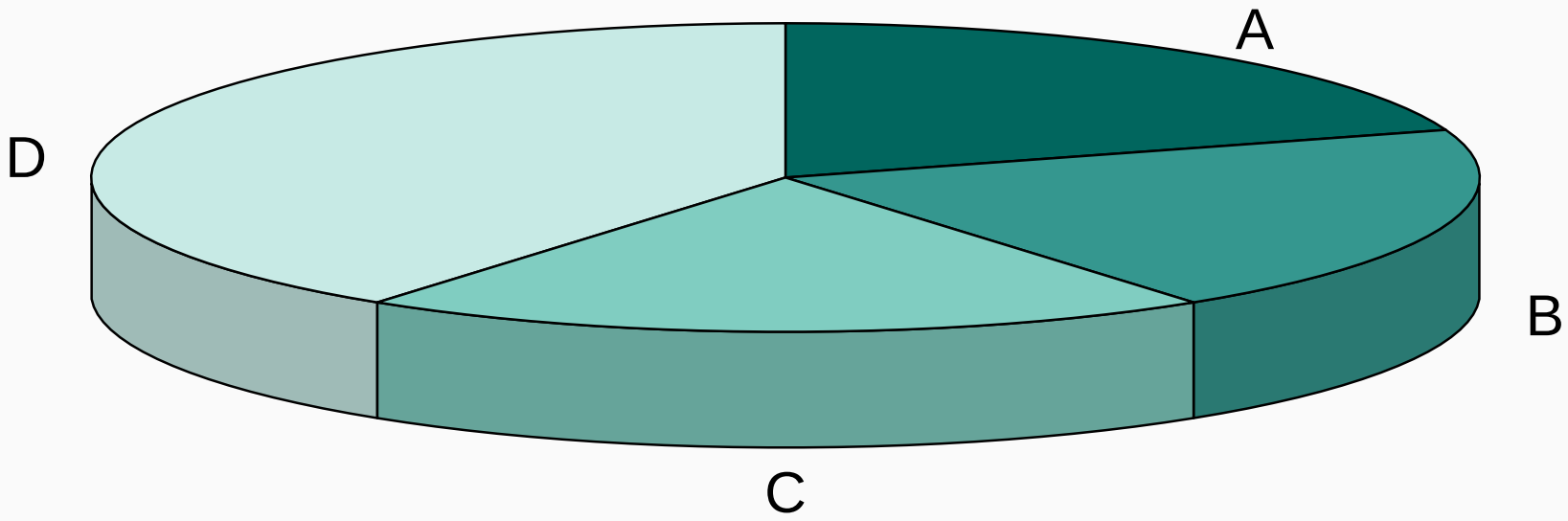
A. _____

B. _____

C. _____

D. _____

3D-effects distort our judgment



Fill in the blanks

The total should be 100%

A. 20%

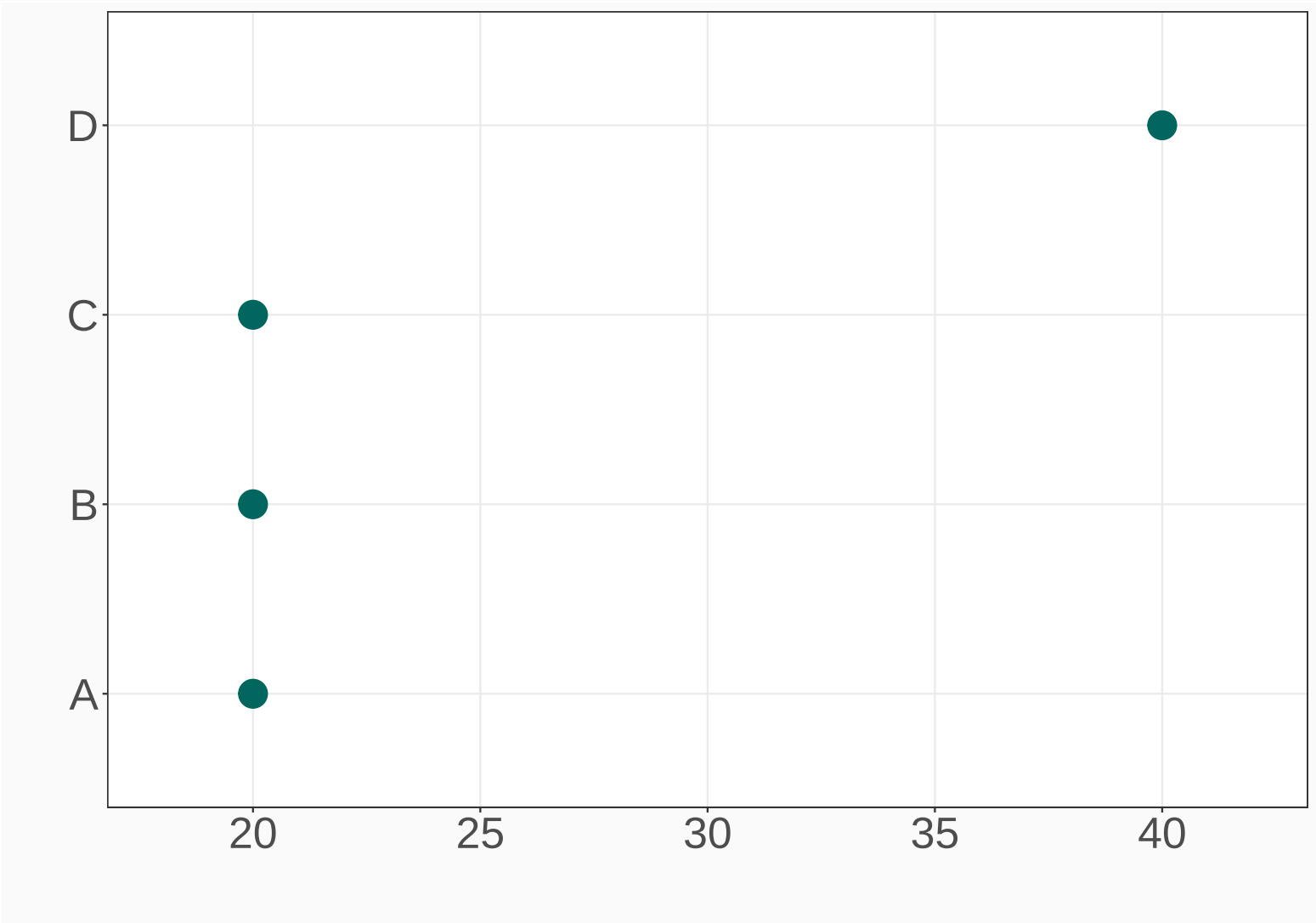
B. 20%

C. 20%

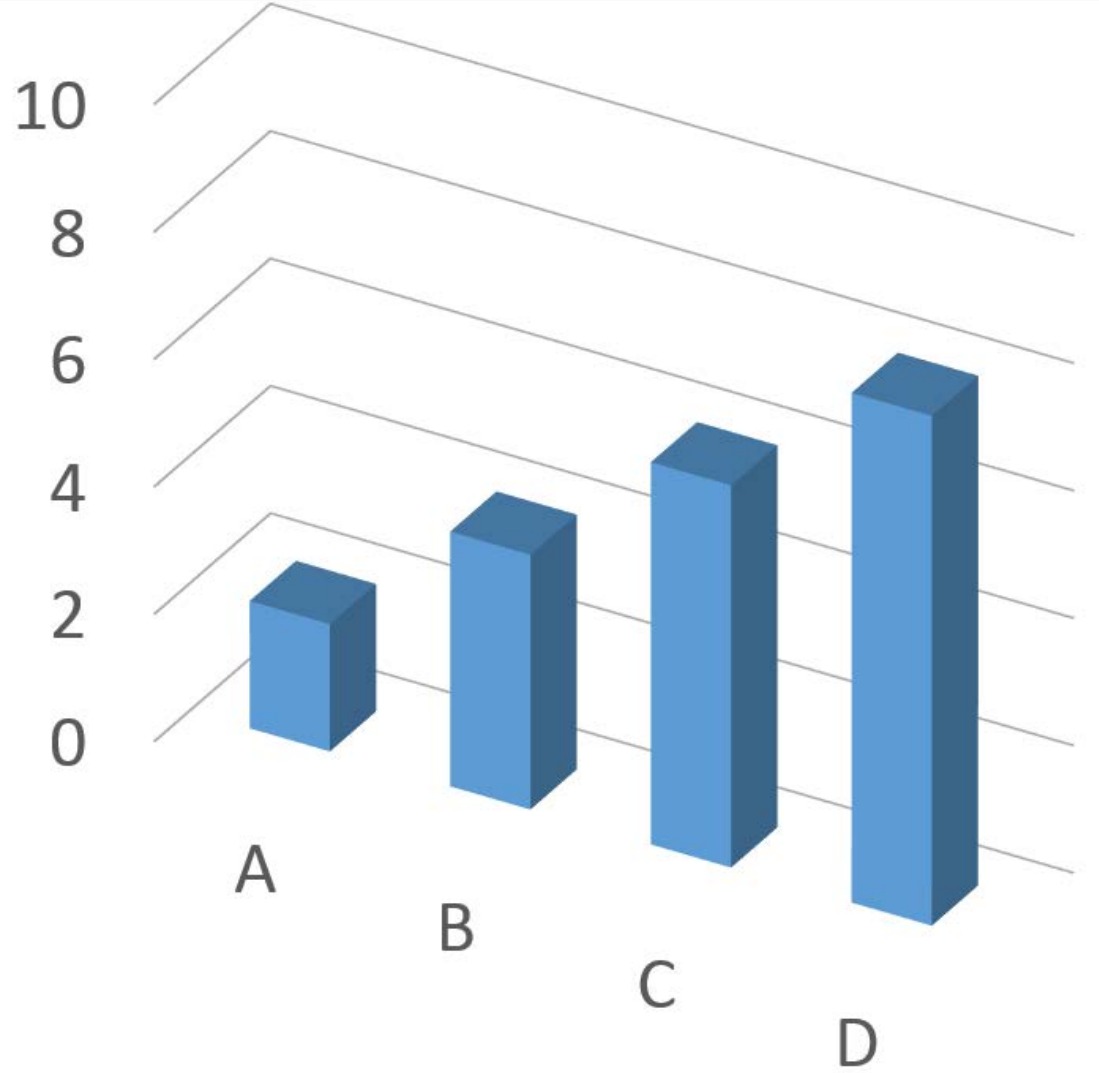
D. 40%

Again, the same data arranged along a common axis

A high-accuracy visual task.



Write down the heights of the bars



This is a visual inspection only.

Fill in the blanks

A. _____

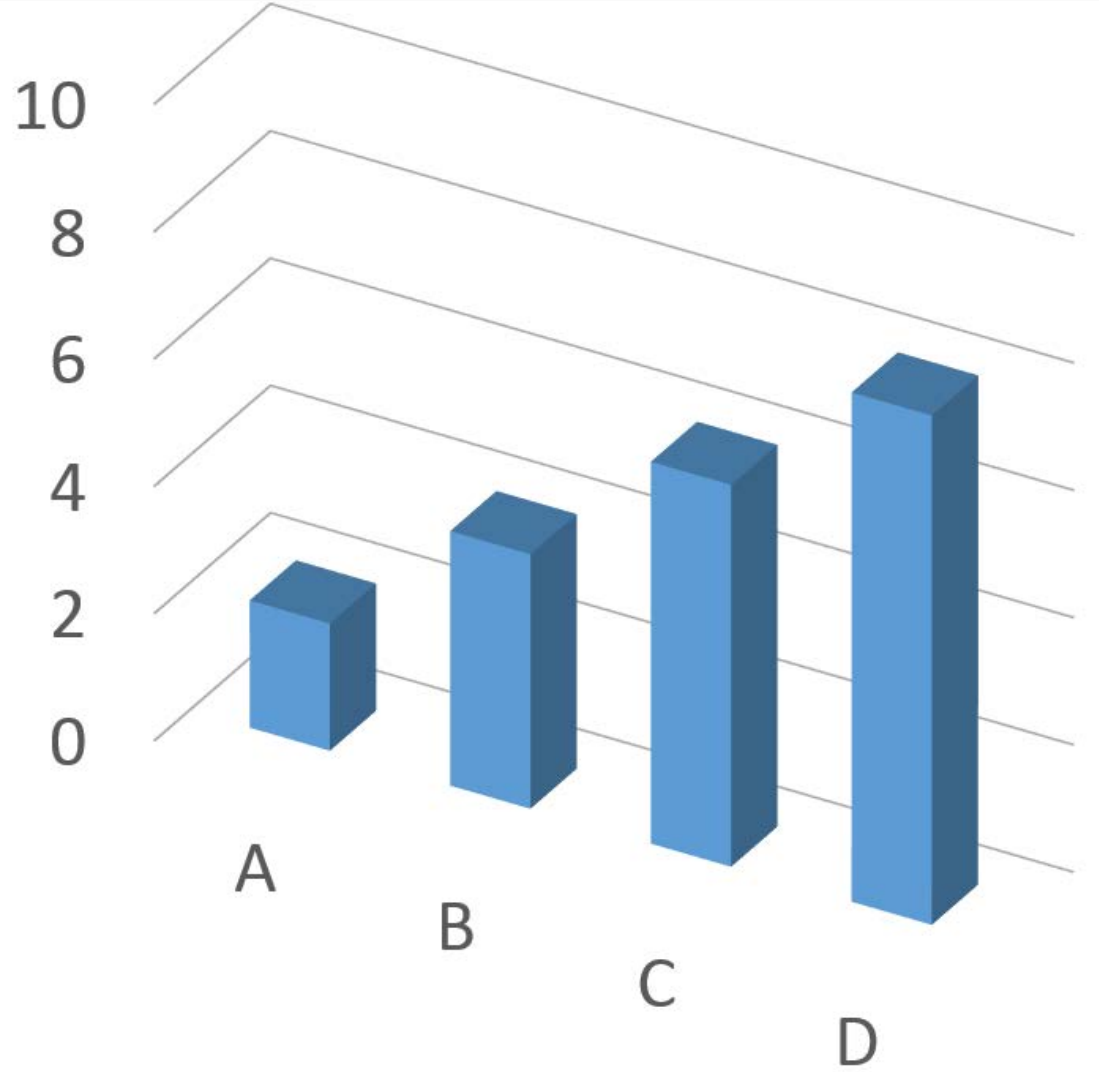
B. _____

C. _____

D. _____

Adapted from (Robbins 2013) p. 22

Again, 3D-effects distort our judgment



This is a visual inspection only.

Fill in the blanks

A. 2

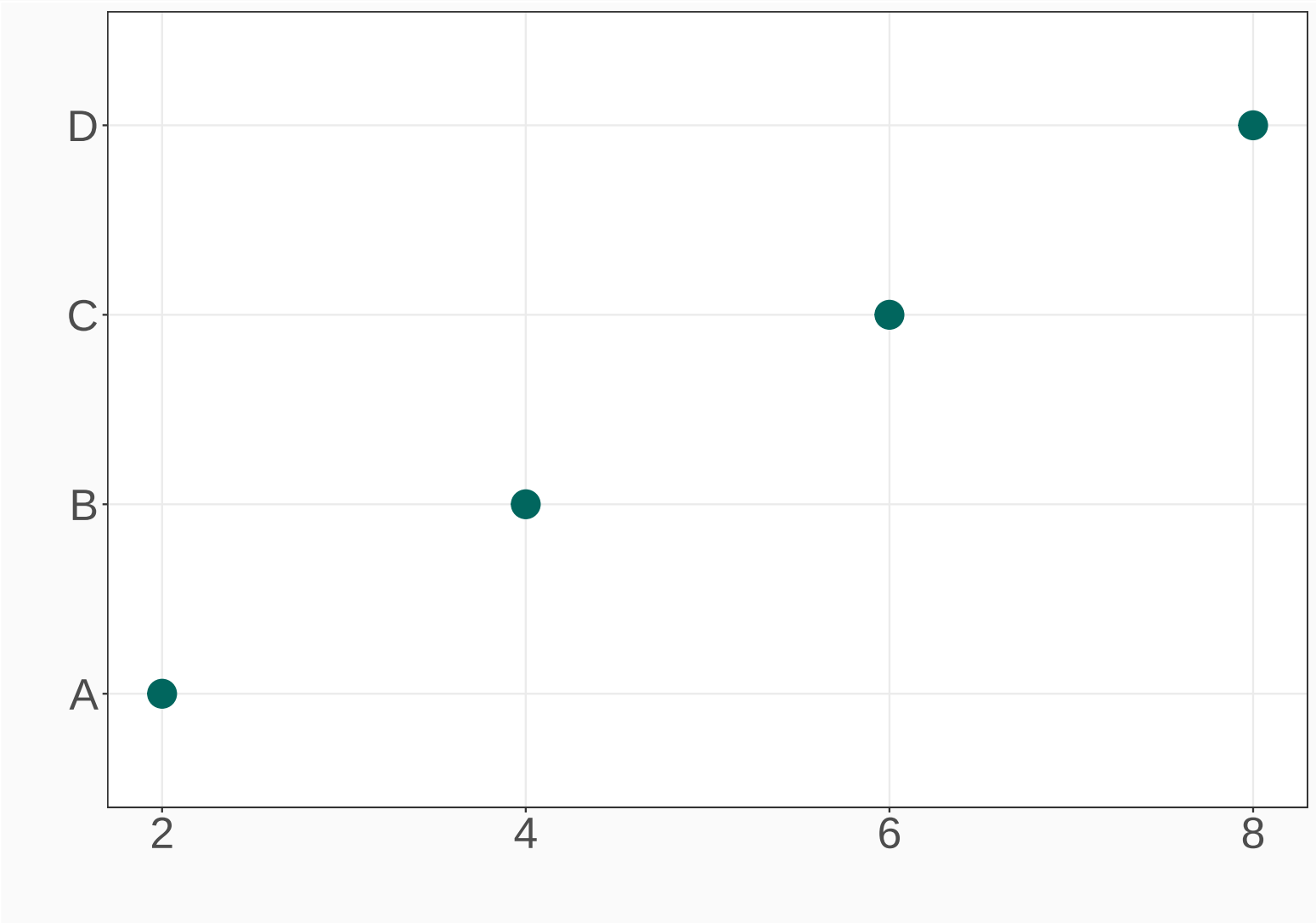
B. 4

C. 6

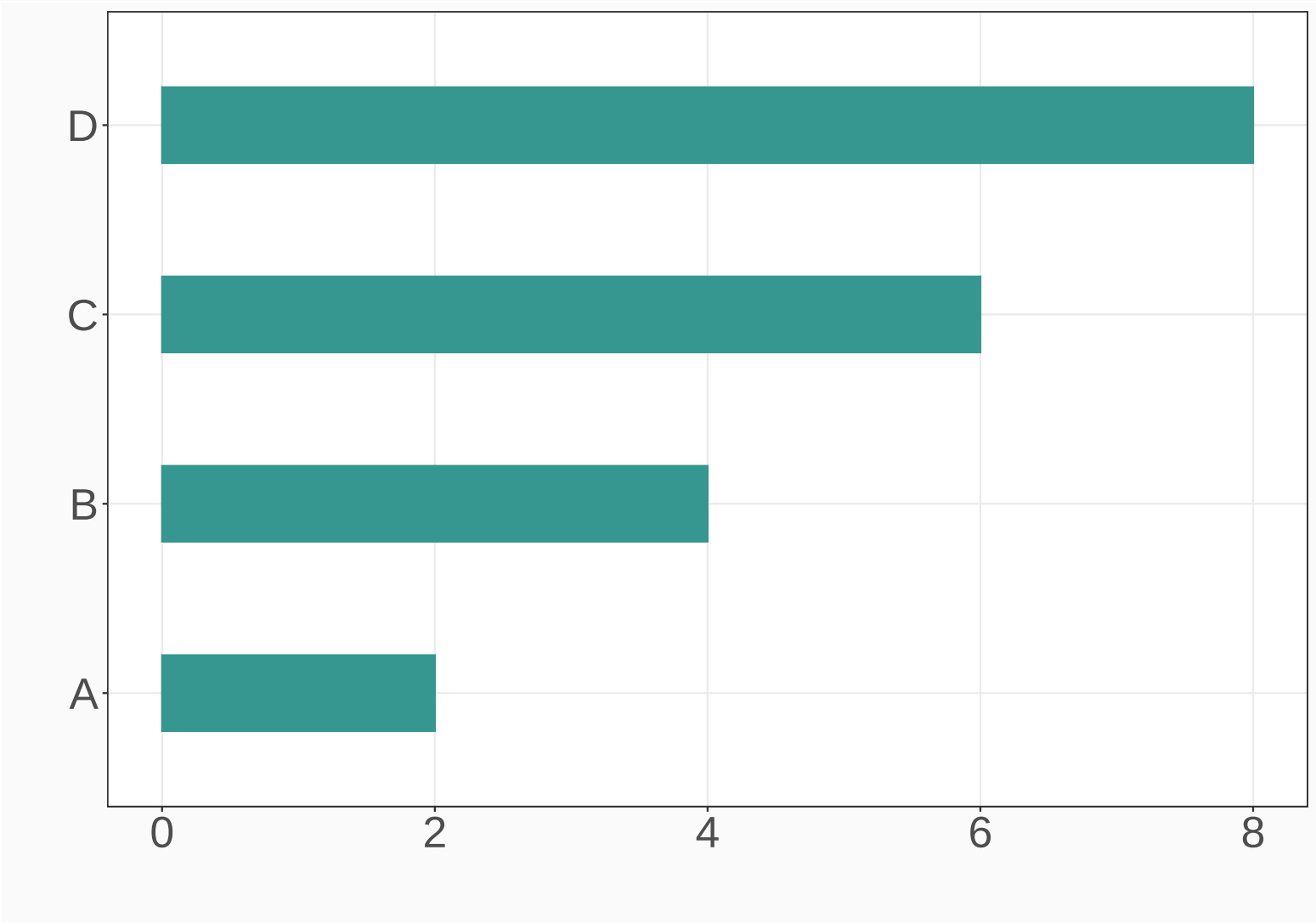
D. 8

Again, the same data arranged along a common axis

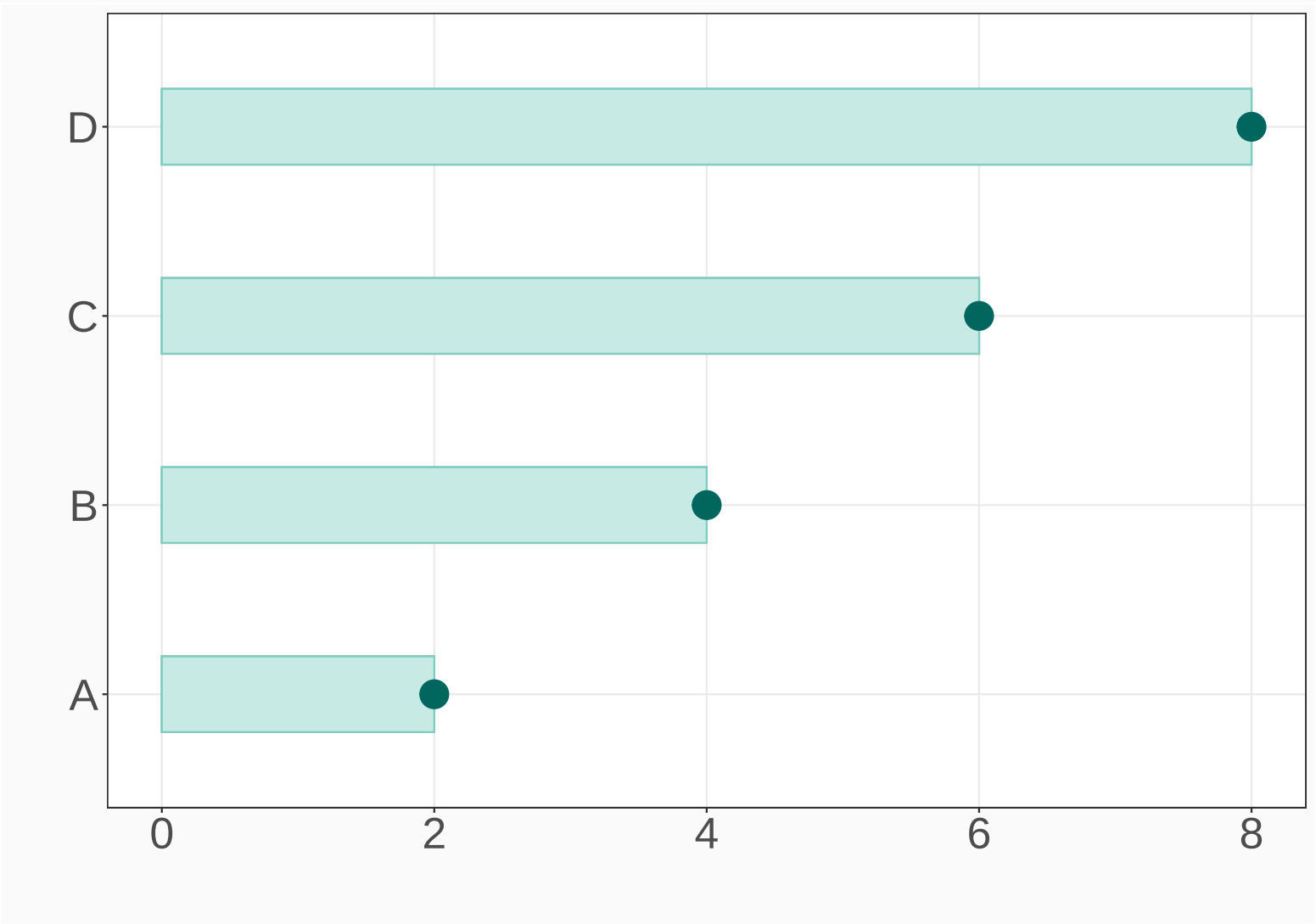
A high-accuracy visual task.



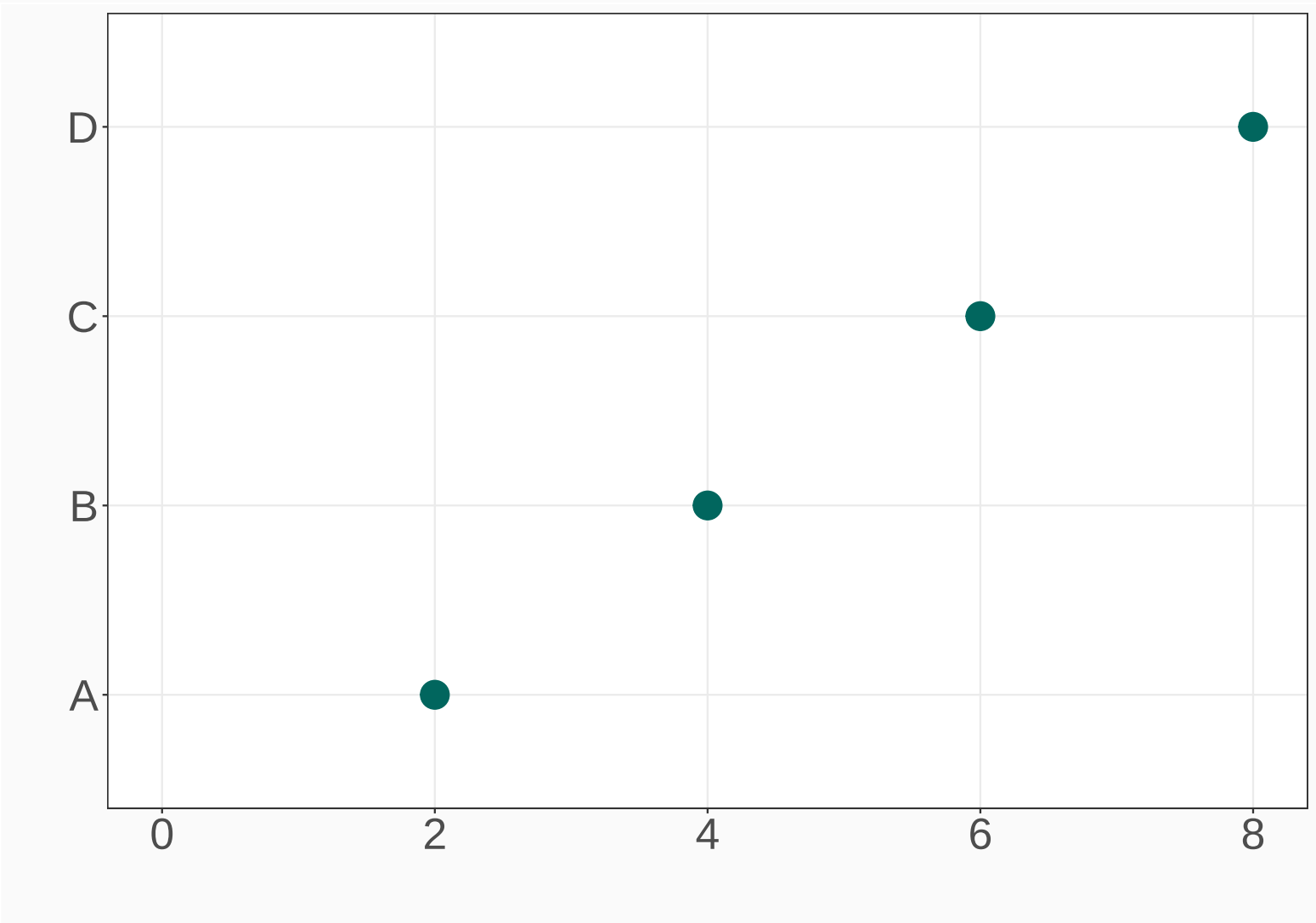
You can use bars, but must include zero



If you mark the endpoints, you can omit the bar



Producing a “dot plot” with rows ordered per the data



Try estimating areas of three states

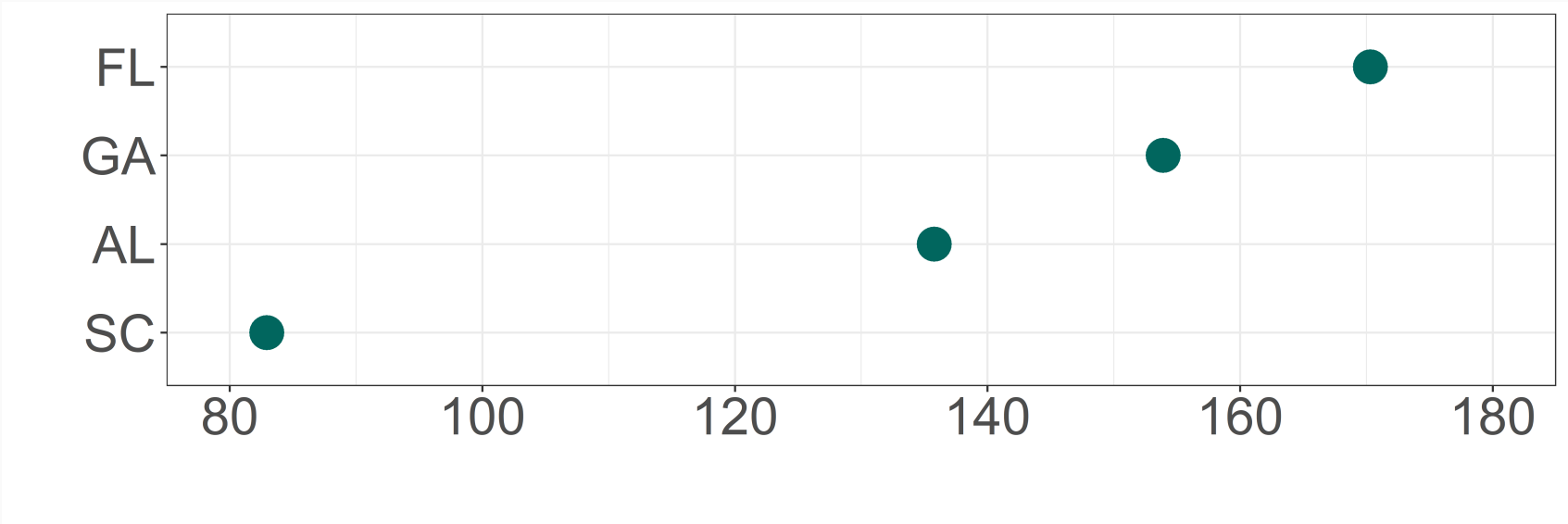
Visual estimation of area is a low-accuracy task.

South Carolina (SC) ≈ 83,000 sq km.

FL	_____	x 1000 sq. km
GA	_____	x 1000 sq. km
AL	_____	x 1000 sq. km
SC	83	x 1000 sq. km

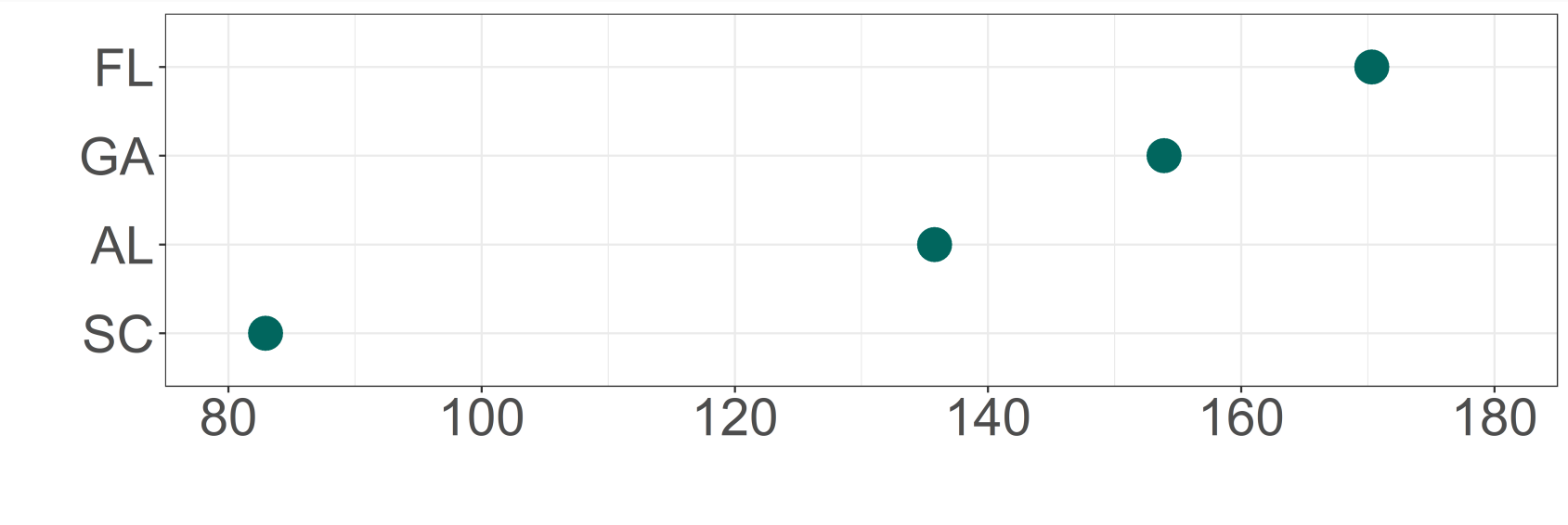
Adapted from (Ihaka 2007)

Again, the same data arranged along a common axis



FL	_____	x 1000 sq. km
GA	_____	x 1000 sq. km
AL	_____	x 1000 sq. km
SC	83	x 1000 sq. km

Your estimates have probably improved



FL **170** x 1000 sq. km

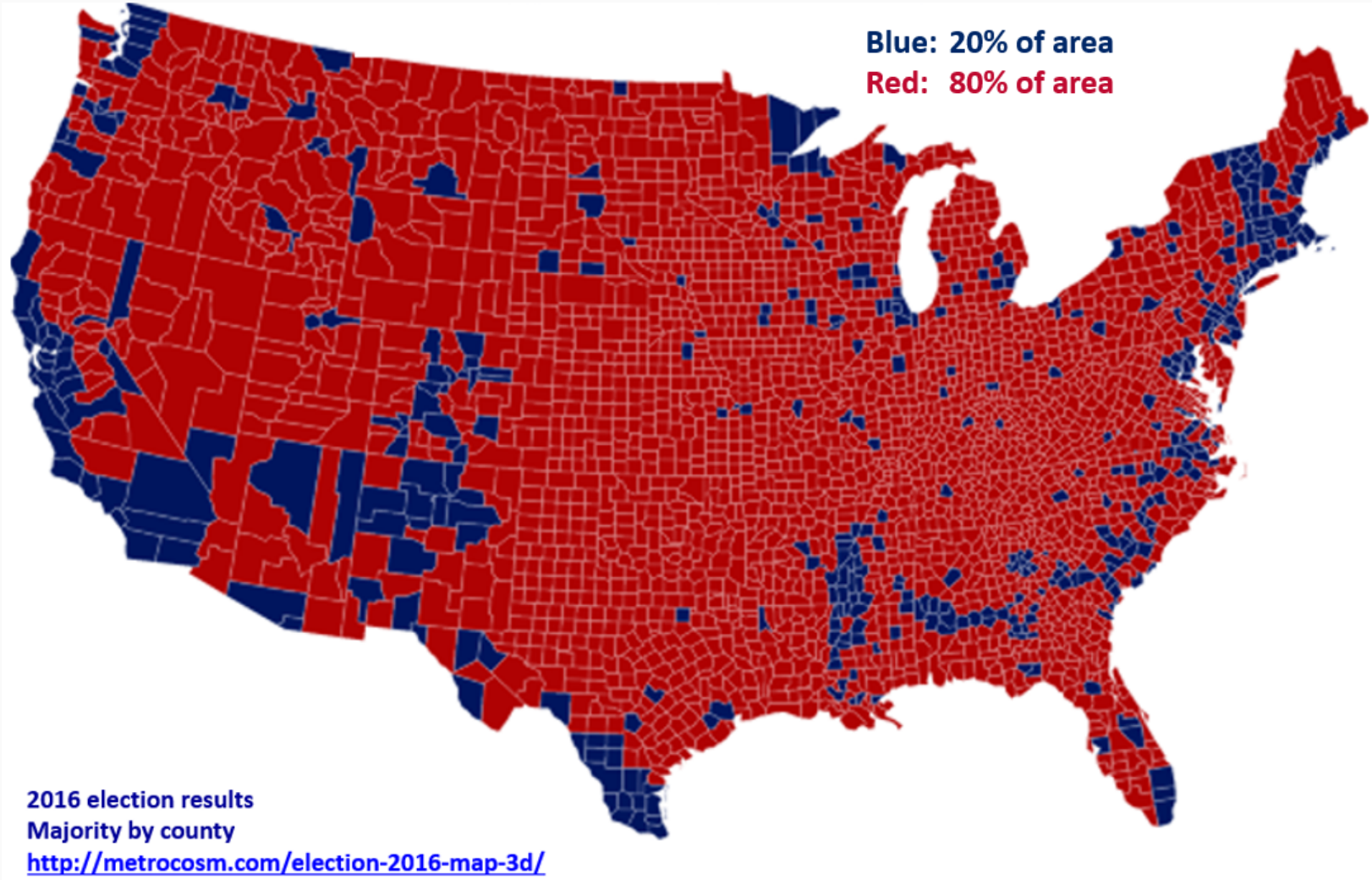
GA **154** x 1000 sq. km

AL **136** x 1000 sq. km

SC **83** x 1000 sq. km

When color represents area, what story emerges?

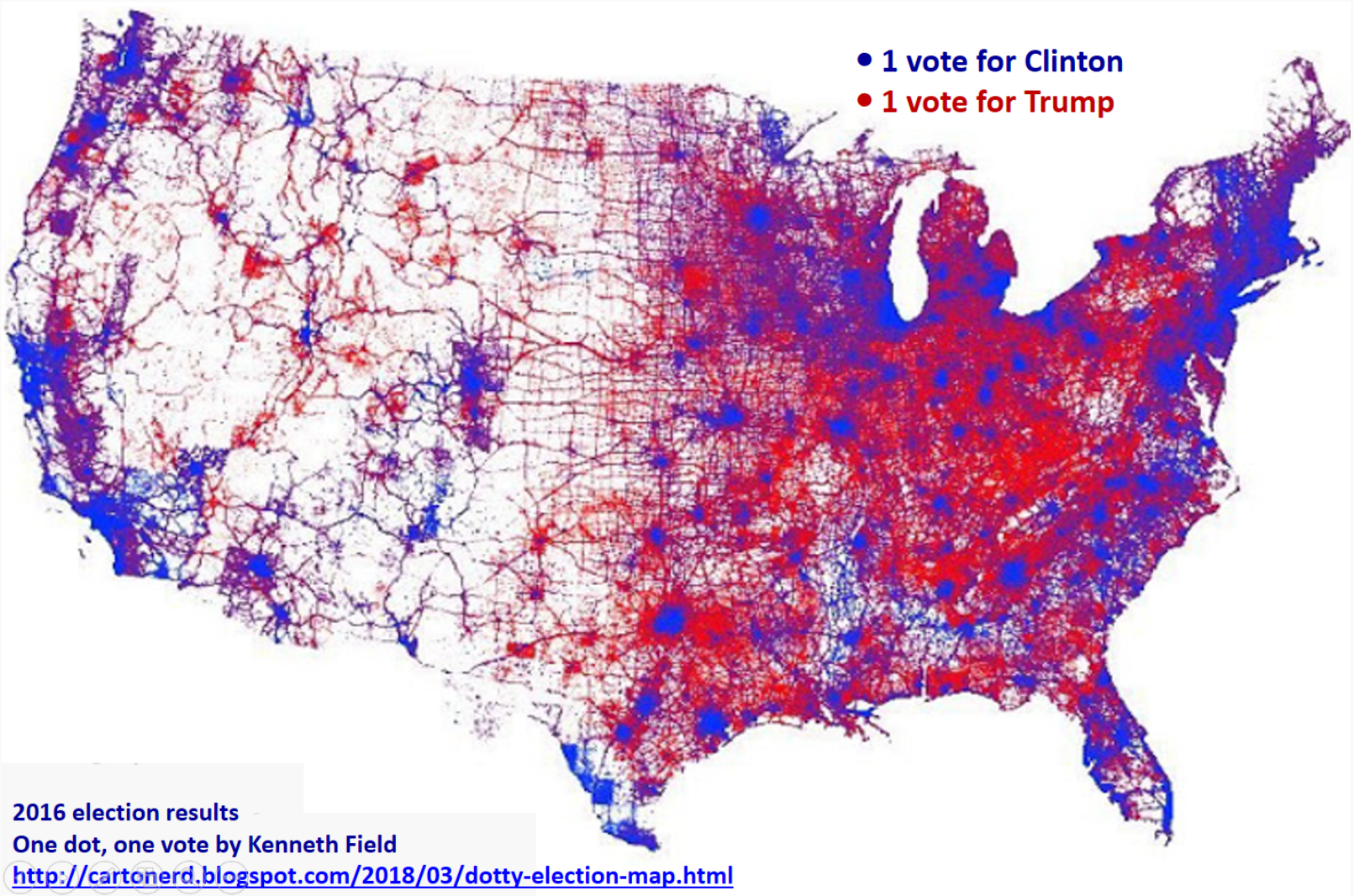
Color used deceptively, 2016 election by county: Clinton, Trump



http://www-personal.umich.edu/~mejn/election/2012/

When color represents voters?

Color used judiciously, each dot 1 votes for: Clinton, Trump



http://coach.weinstein.to/lets-get-specific/election-results/

The experts tell us

(Doumont 2009)

Image from `http://www.principiae.be/pdfs/Principiae-2014.pdf`

Optimal design primarily depends on

- The message to be conveyed
- The variables to be shown

The experts tell us

The task of the designer is to give visual access to the subtle and the difficult — that is, reveal the complex.

(Tufte 1983)

Image from `https://en.wikipedia.org/wiki/Edward_Tufte`

The experts tell us

What's your point?

Seriously, that's the most important question.

(Evergreen 2017)

Image from `https://tei.cgu.edu/people/stephanie-evergreen-phd/`

R is designed with statistical analysis and data graphics in mind

Well-designed data graphics are accessible, even to the beginner

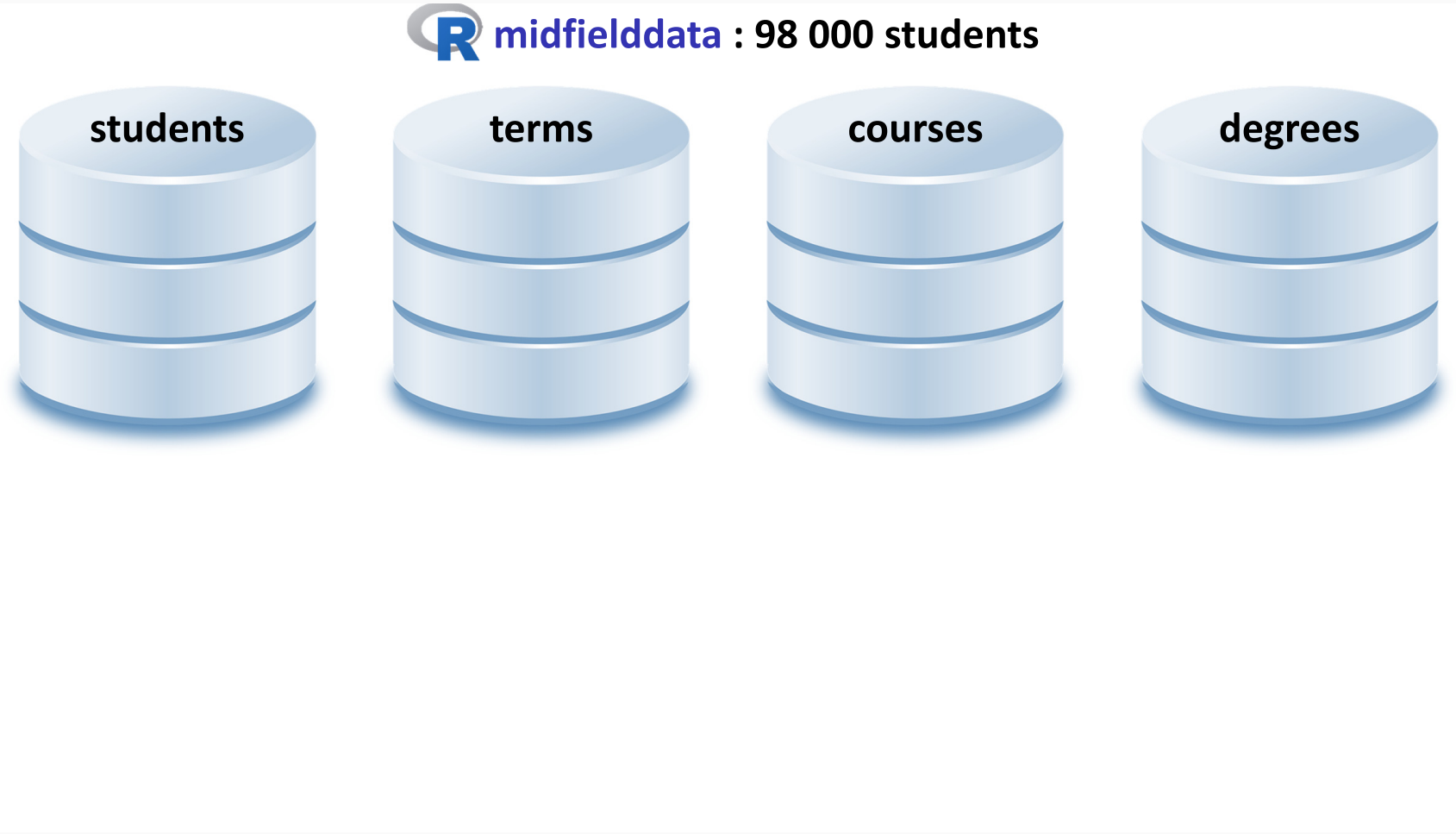
- makes graphical exploration of data accessible to all
- work in progress is easily disseminated via GitHub

And because R is open-source

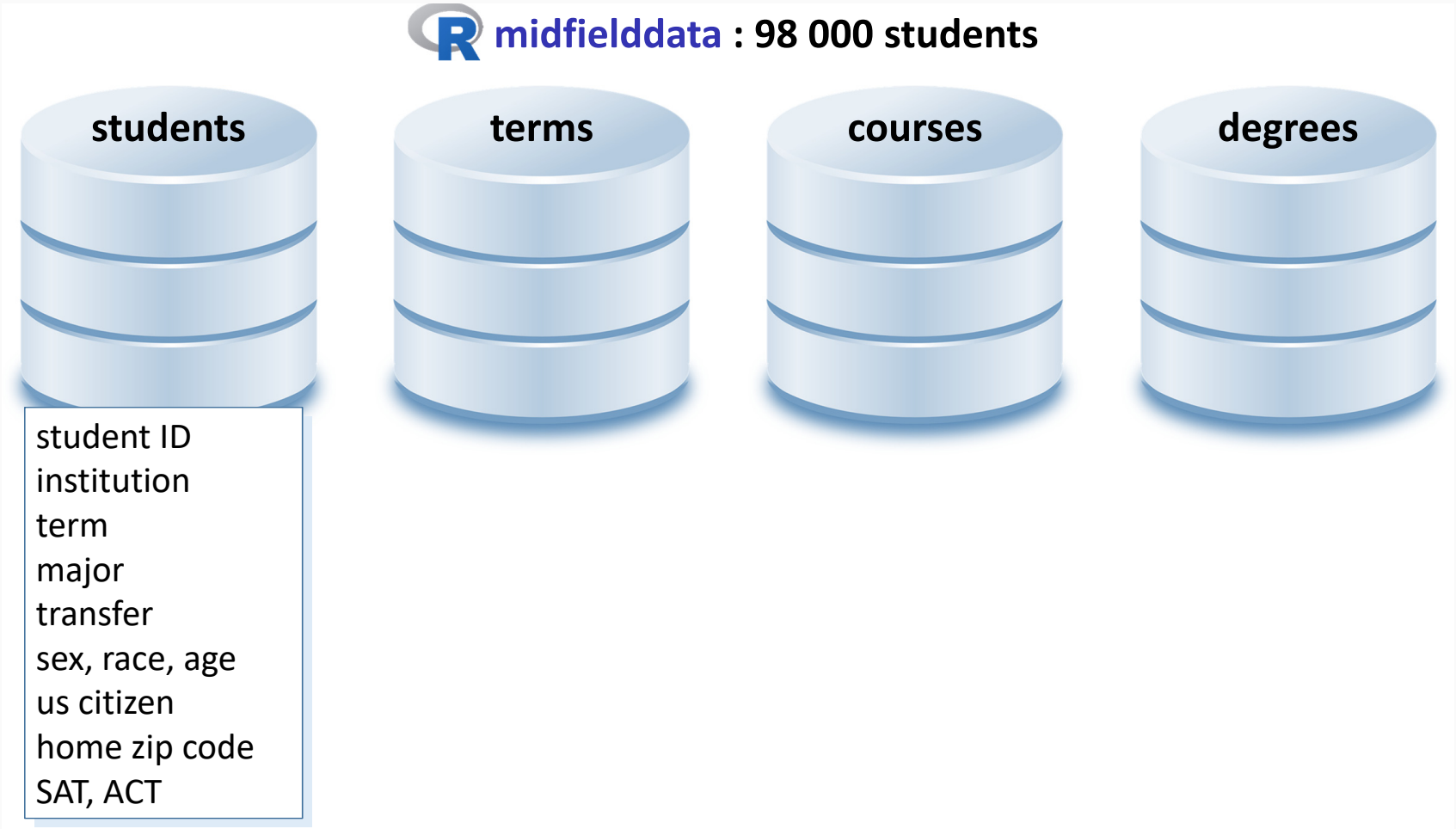
- new packages appear regularly—one might solve *your* problem
- anyone can help us find errors and add features to our packages

Introducing midfieldr

R package midfielddata provides a stratified sample



Each observation is a unique student

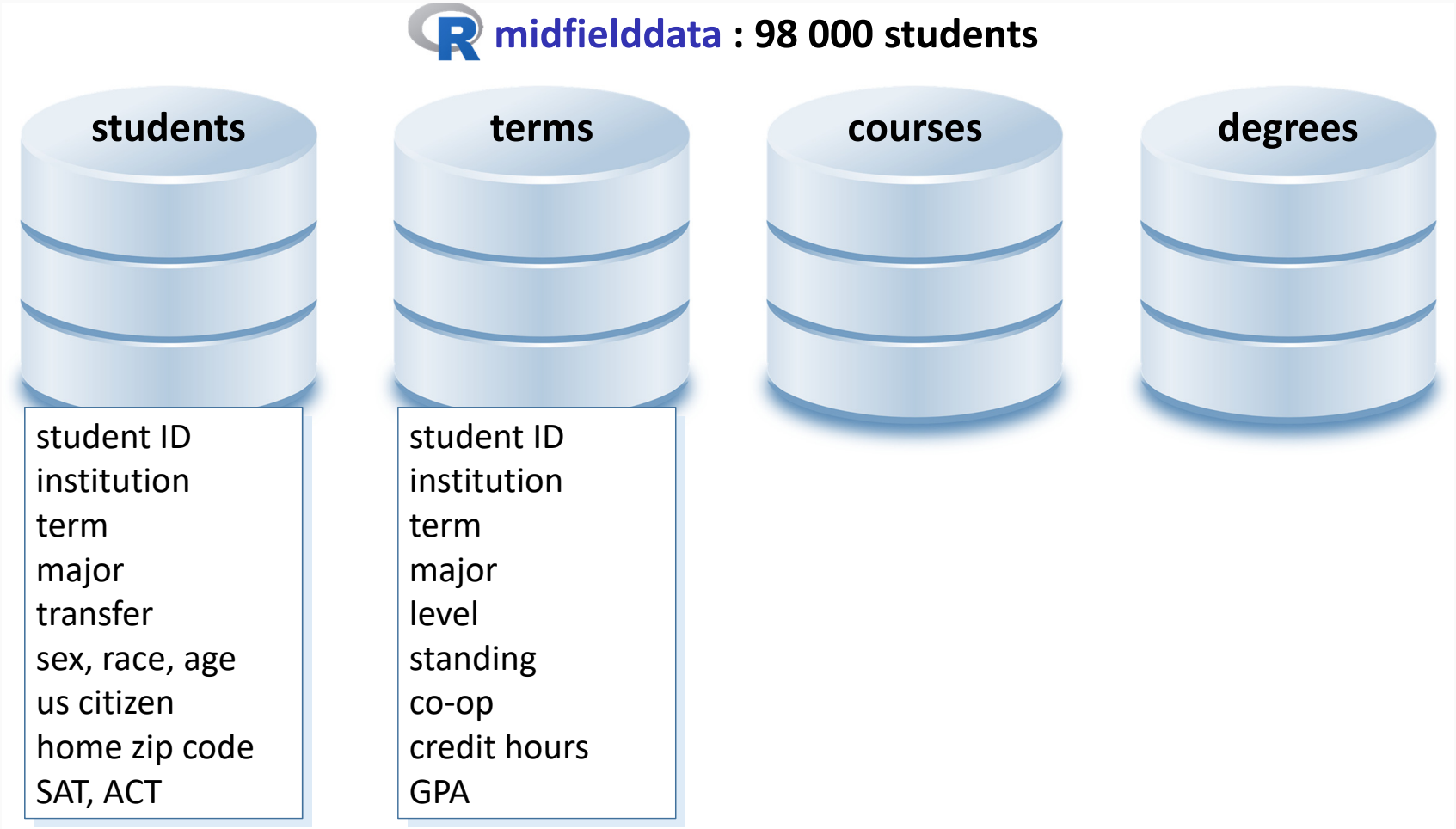


`midfieldstudents`

98,000 observations

19 Mb of memory

Each observation is one term for one student

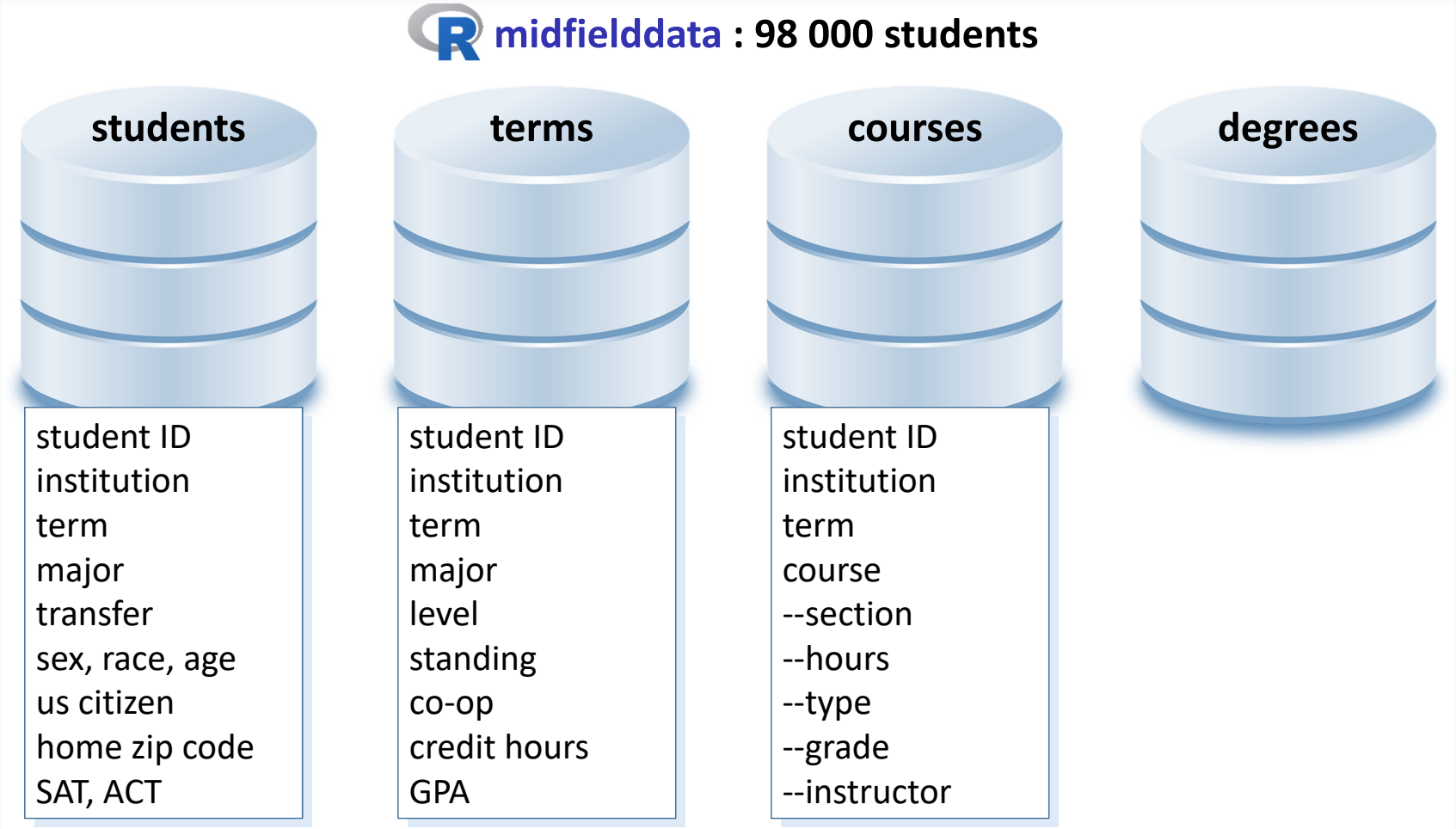


`midfieldterms`

729,000 observations

82 Mb of memory

Each observation is one course for one student

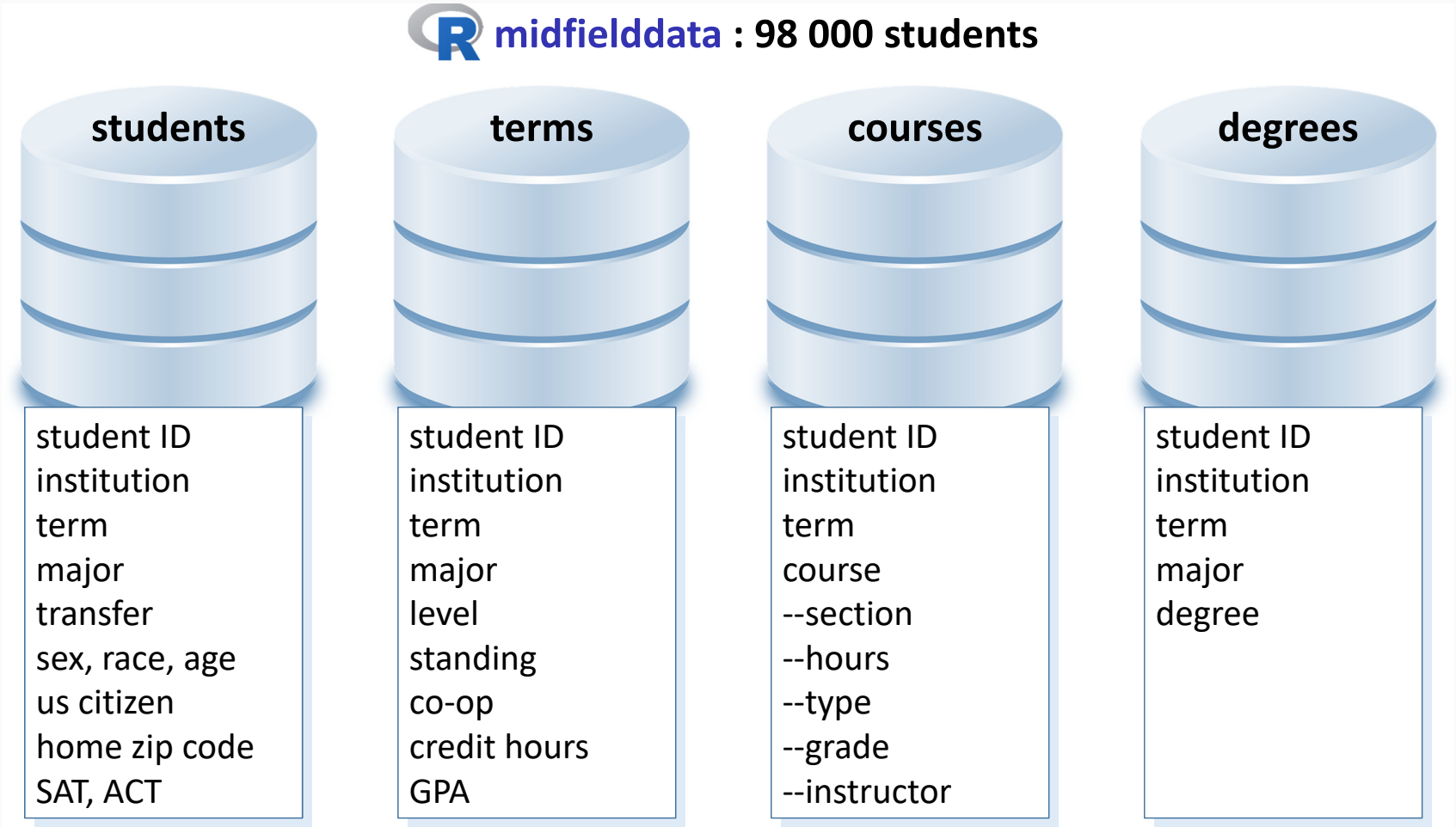


`midfieldcourses`

3.5 M observations

348 Mb of memory

Each observation is a unique student

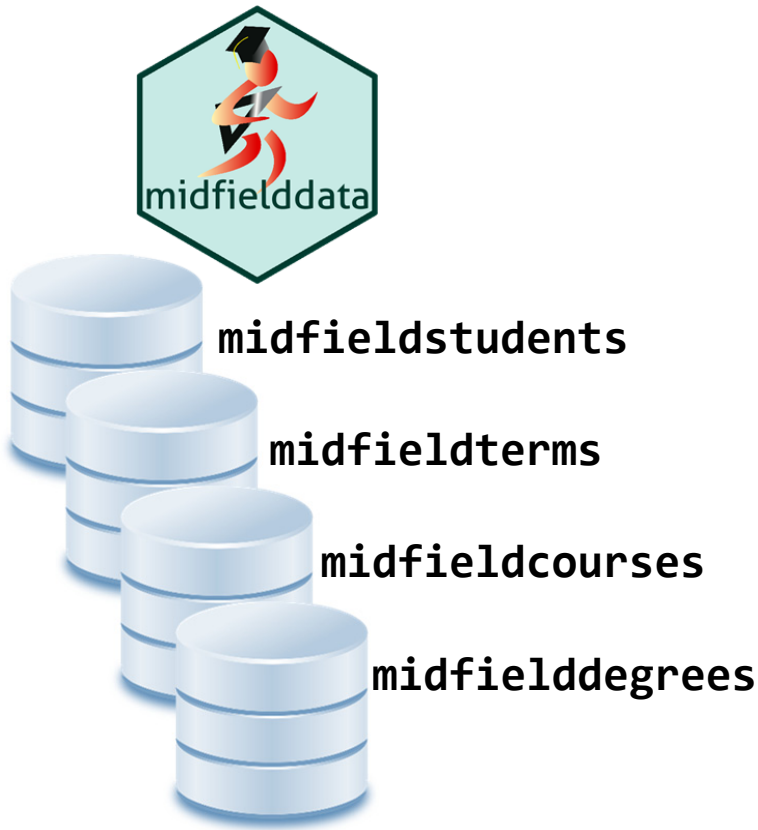


`midfielddegrees`

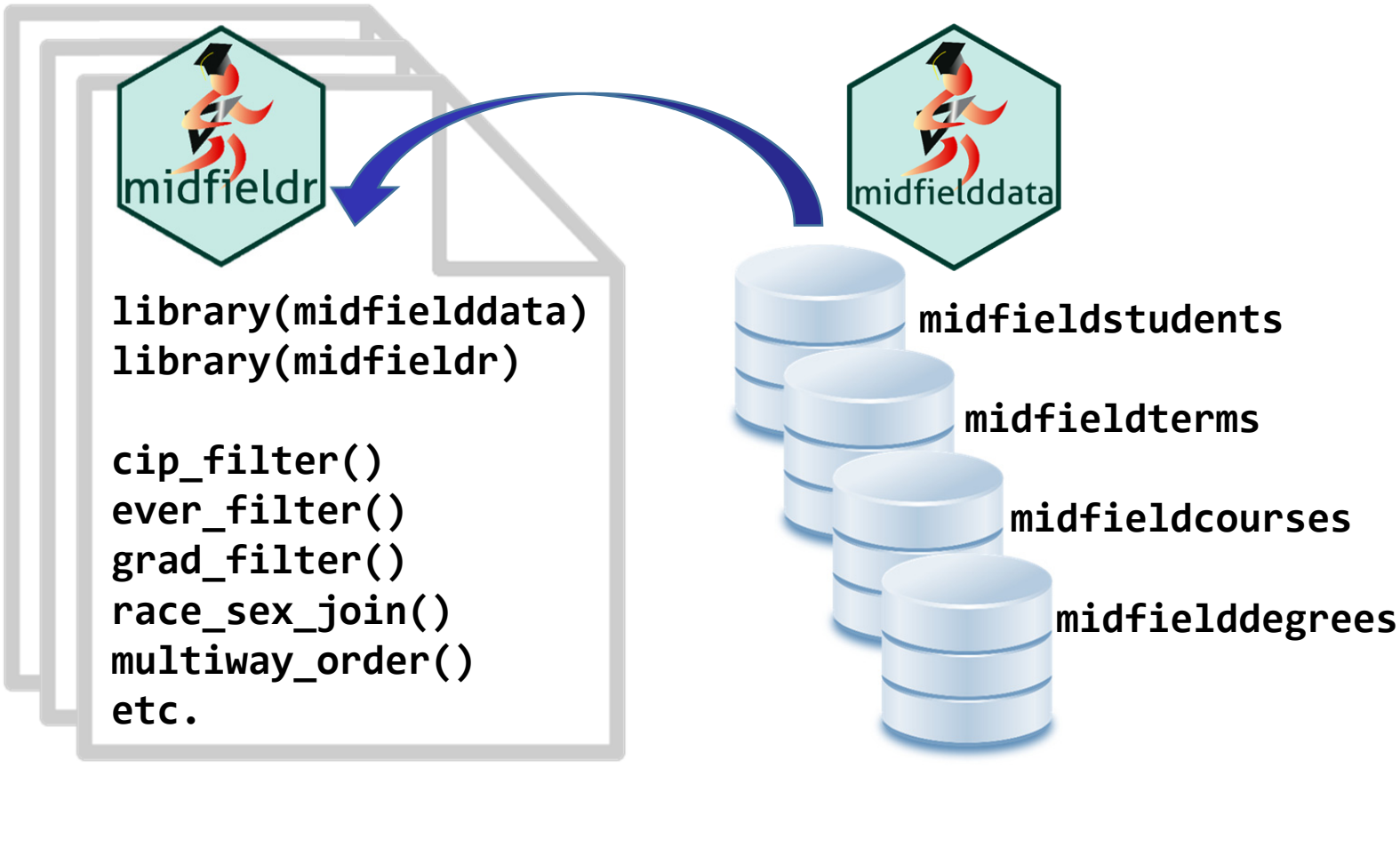
98,000 M observations

10 Mb of memory

midfielddata provides the data



midfieldr provides the tools



Preparing for the workshop, you installed both packages

`https://midfieldr.github.io/midfieldr`

midfieldr

CRAN not published | build passing | coverage 77% | downloads null | License GPL v3

A package for investigating student record data provided by registrars at US universities participating in the MIDFIELD project.

midfielddata

build passing | License CC-0

An R data package containing four data sets:

- `midfieldstudents` Student demographic attributes
- `midfieldcourses` Academic course attributes
- `midfieldterms` Academic term attributes
- `midfielddegrees` Student graduation attributes

midfieldr provides functions for working with midfieldddata

Some of those functions you will use today are:

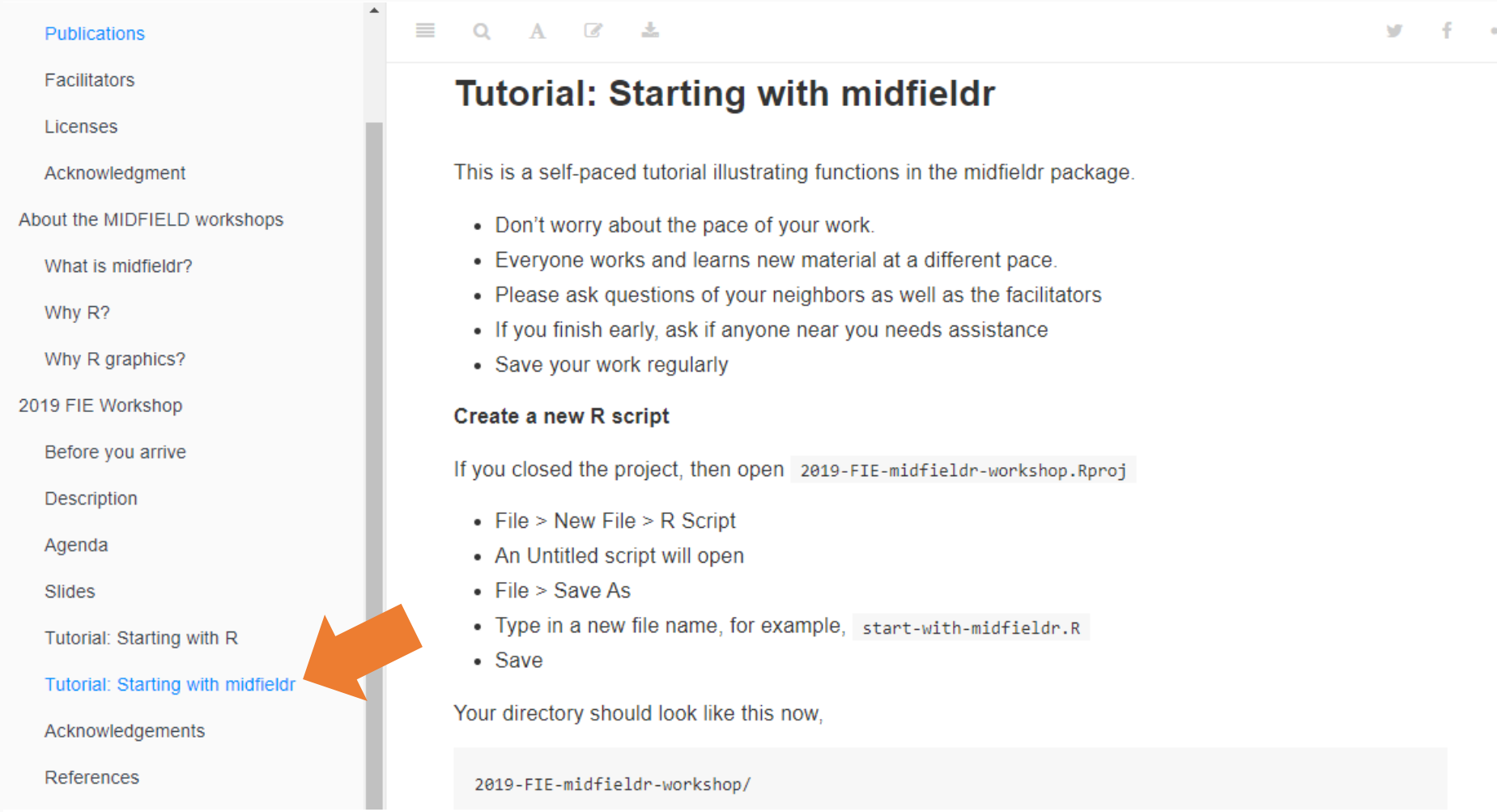
Function	Provides
cip_filter()	Identify programs by CIP code
cip_label()	Label your programs
ever_filter()	Find all students ever enrolled in your programs
grad_filter()	Find all graduates of your programs
race_sex_join()	Join student race/ethnicity and sex to the data
multiway_order()	Order the rows and panels of multiway data

Starting with midfieldr (tutorial)

This self-paced tutorial illustrates midfieldr functions

- Don’t worry about the pace of your work.
- Everyone works and learns new material at a different pace.
- Please ask questions of your neighbors as well as the facilitators
- If you finish early, ask if anyone near you needs assistance
- Save your work regularly

https://midfieldr.github.io/2019-FIE-workshop/



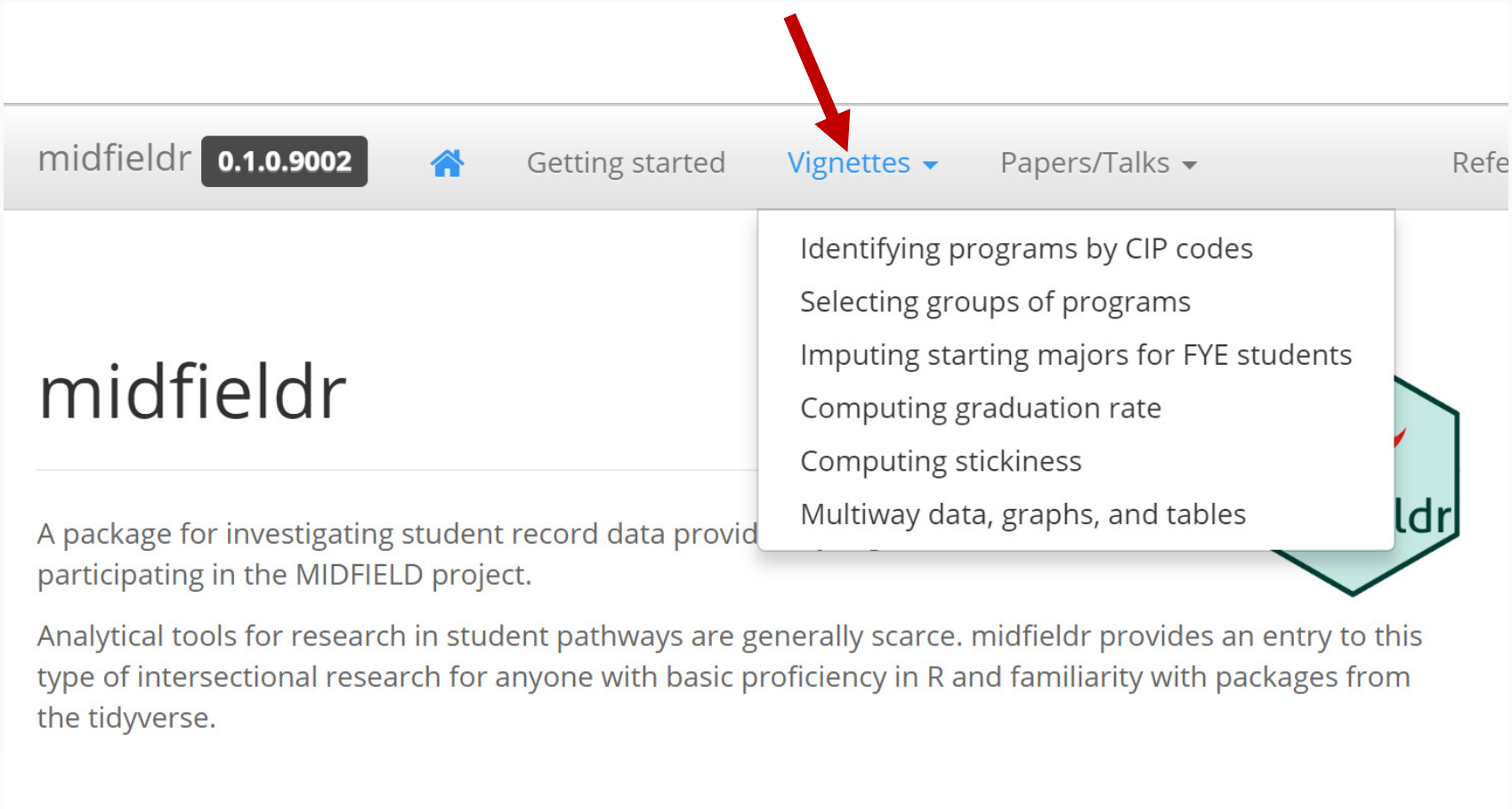
Start a new R script, add a line of code, run it

Examine the result, repeat

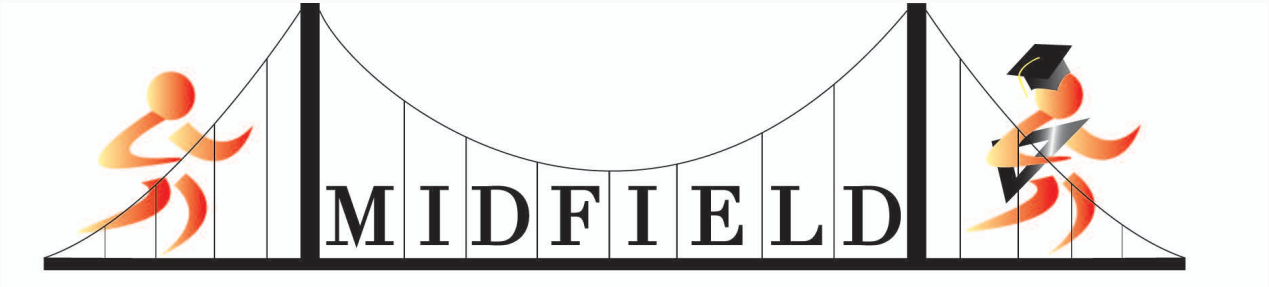
Next steps

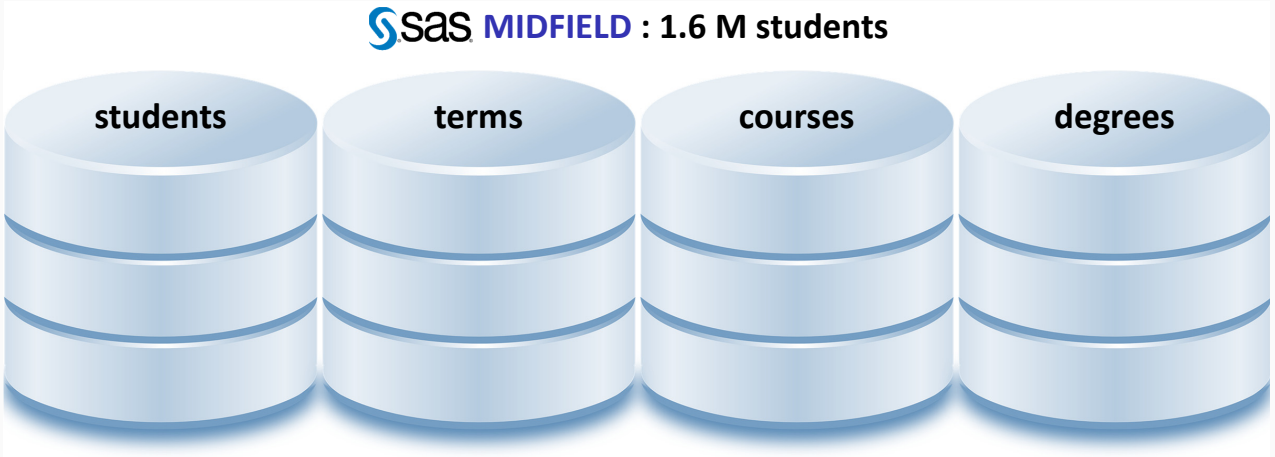
Next steps in learning to use midfieldr

Several more vignettes (tutorials) on the midfieldr website



Next steps if you want more than a MIDFIELD sample

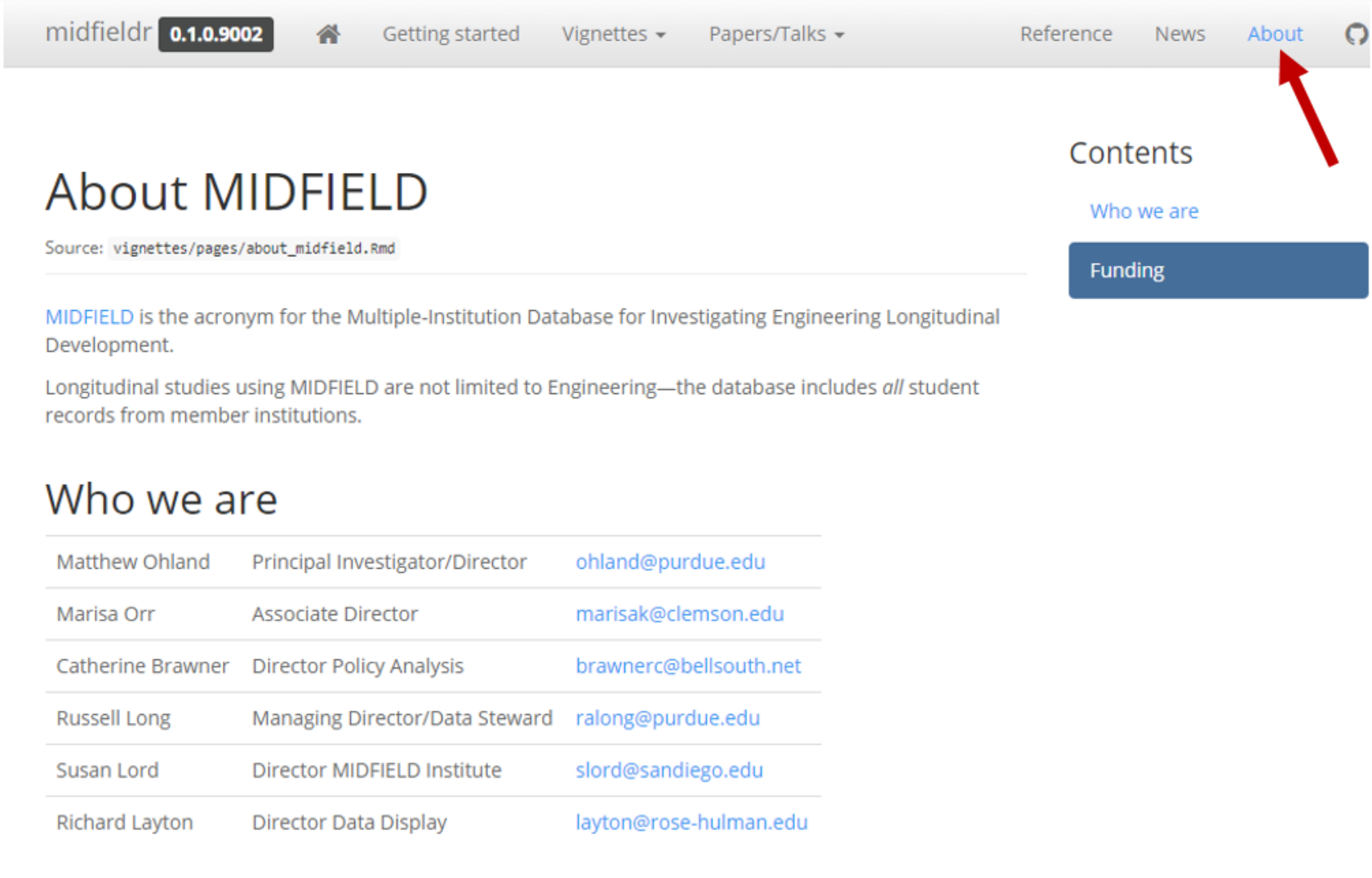




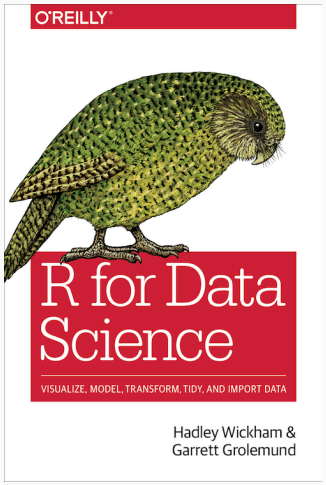
Talk to a member of the MIDFIELD team.

Names and emails on the website.

Talk to a member of the MIDFIELD team

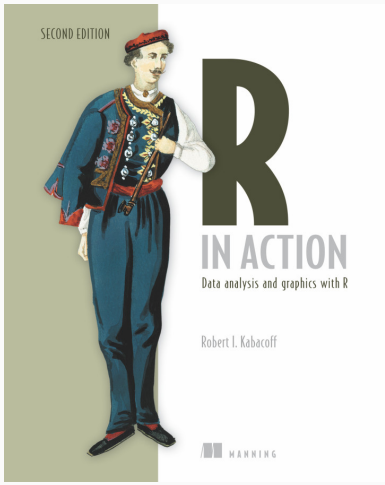


Next steps in learning R

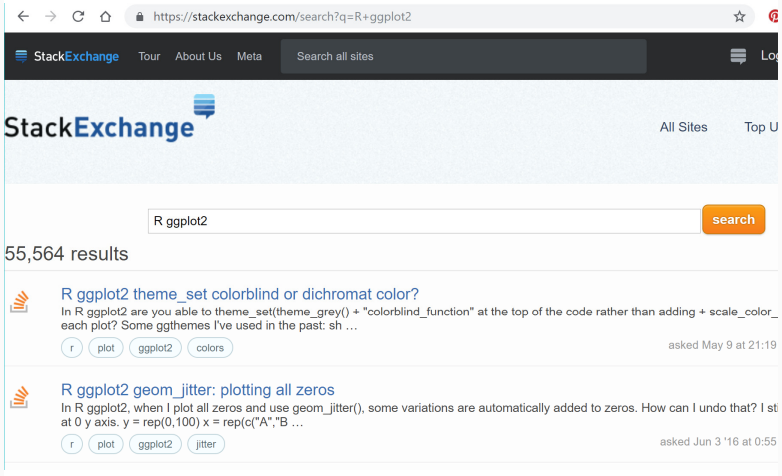


Hadley Wickham

Garrett Grolemund



Robert Kabacoff

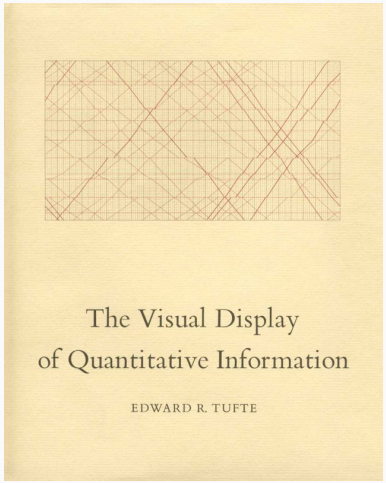


StackExchange.com

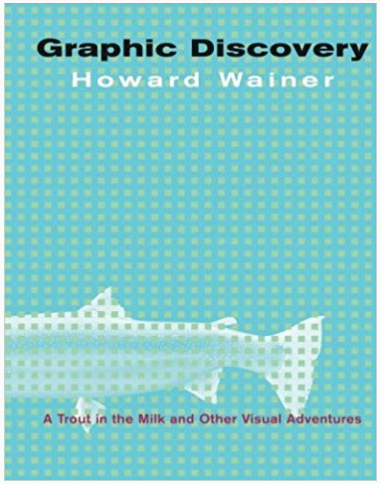
Or just google it

Your problem may already be solved

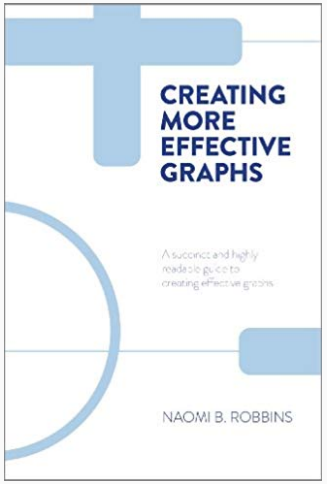
Next steps in learning about graph design



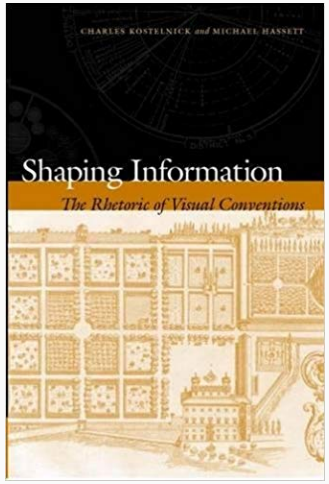
Edward Tufte



Howard Wainer



Naomi Robbins



Charles Kostelnick

Michael Hassett

References

Doumont, Jean-luc. 2009. *Trees, maps, and theorems: Effective communication for rational minds*. 2nd ed. Kraainem, Belgium: Principiae.

Evergreen, Stephanie D. H. 2017. *Effective Data Visualization: The Right Chart for the Right Data*. Sage.

Ihaka, Ross. 2007. "Statistics 787 Lecture slides."

Kabacoff, Robert. 2015. *R in Action: Data Analysis and Graphics with R, 2/e*. Manning Publications Co.

Kostelnick, Charles, and Michael Hassett. 2003. *Shaping Information: The Rhetoric of Visual Conventions*. Southern Illinois University.

Robbins, Naomi. 2013. *Creating More Effective Graphs*. Chart House.

Tufte, Edward. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

Wainer, Howard. 1997. *Visual Revelations: Graphical Tales of Fate and Deception From Napoleon Bonaparte To Ross Perot*. Copernicus.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science*. Sebastopol, CA: O'Reilly Media, Inc.