

Accessing MIDFIELD

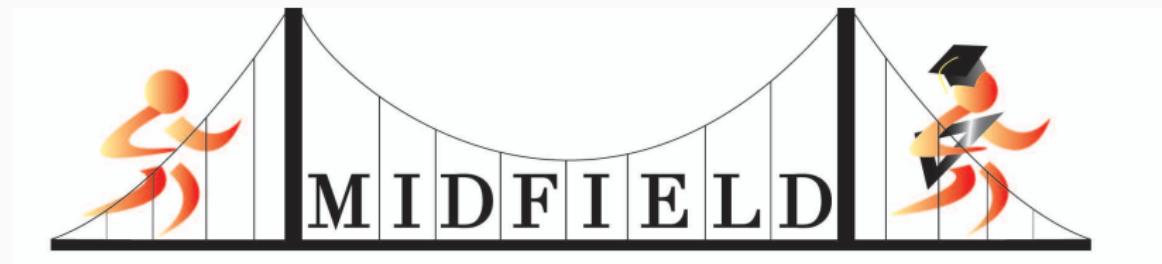
A workshop for R beginners

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

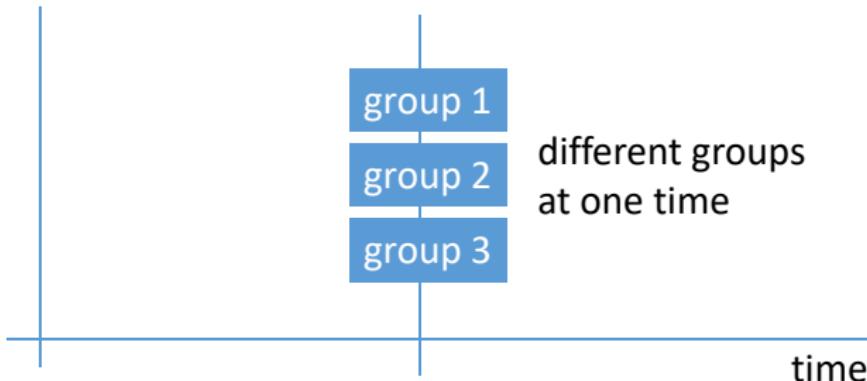
password: OCT2019

R-Bar volunteers can help with software issues

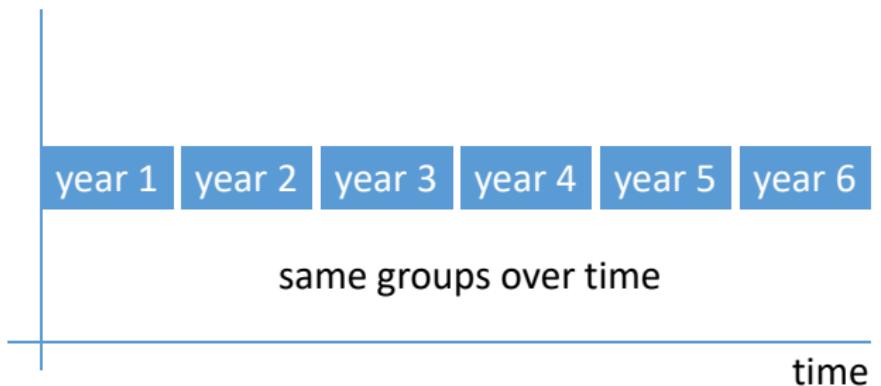
Min	Topic
10	Introductions
5	The MIDFIELD data structures
30	Finding stories in the data
35	Starting with R (tutorial)
15	— break —
20	Elements of effective graphs
5	Introducing midfieldr
50	Starting with midfieldr (tutorial)
10	Next steps

The MIDFIELD data structures

In education, cross-sectional designs are typical



Longitudinal studies offer some advantages



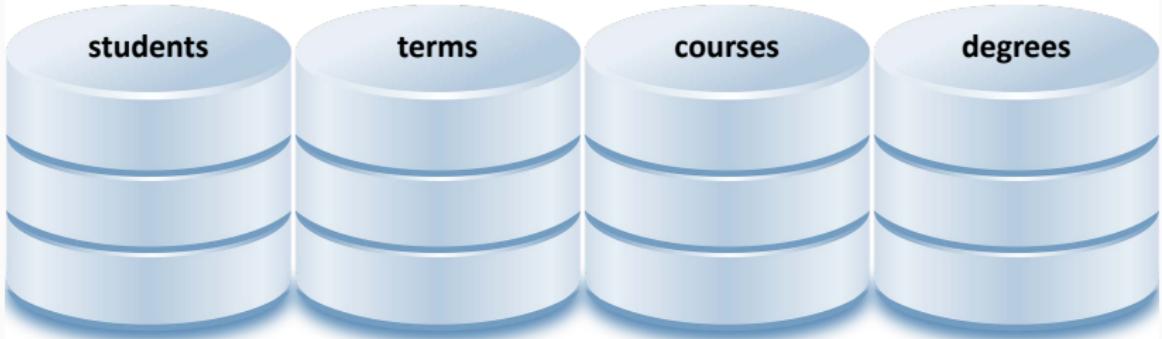
MIDFIELD is a database for longitudinal studies



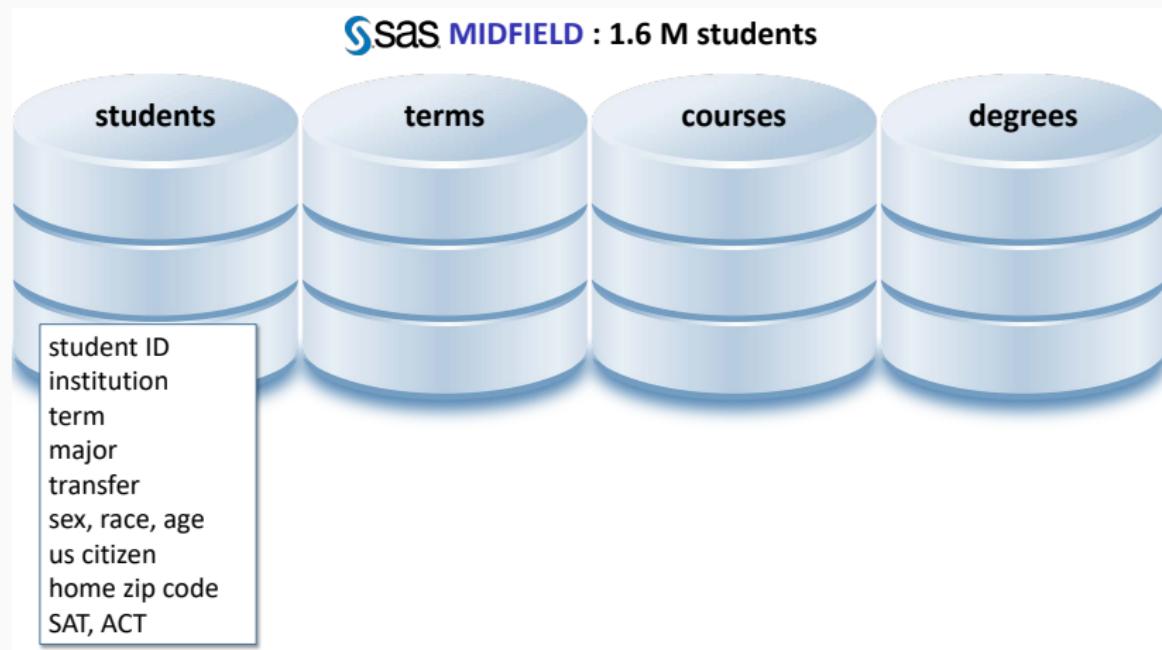
- 1.6 M undergraduate students at 22 US institutions
- whole-population data from registrars
- 1987–present

MIDFIELD data are curated in four categories

Sas MIDFIELD : 1.6 M students

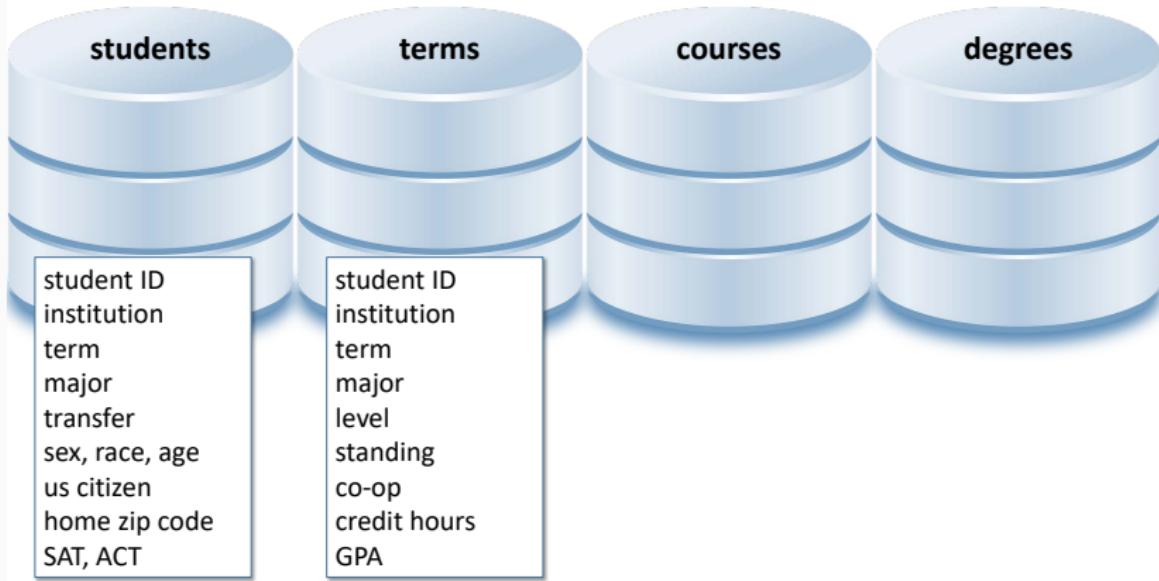


Each observation is a unique student

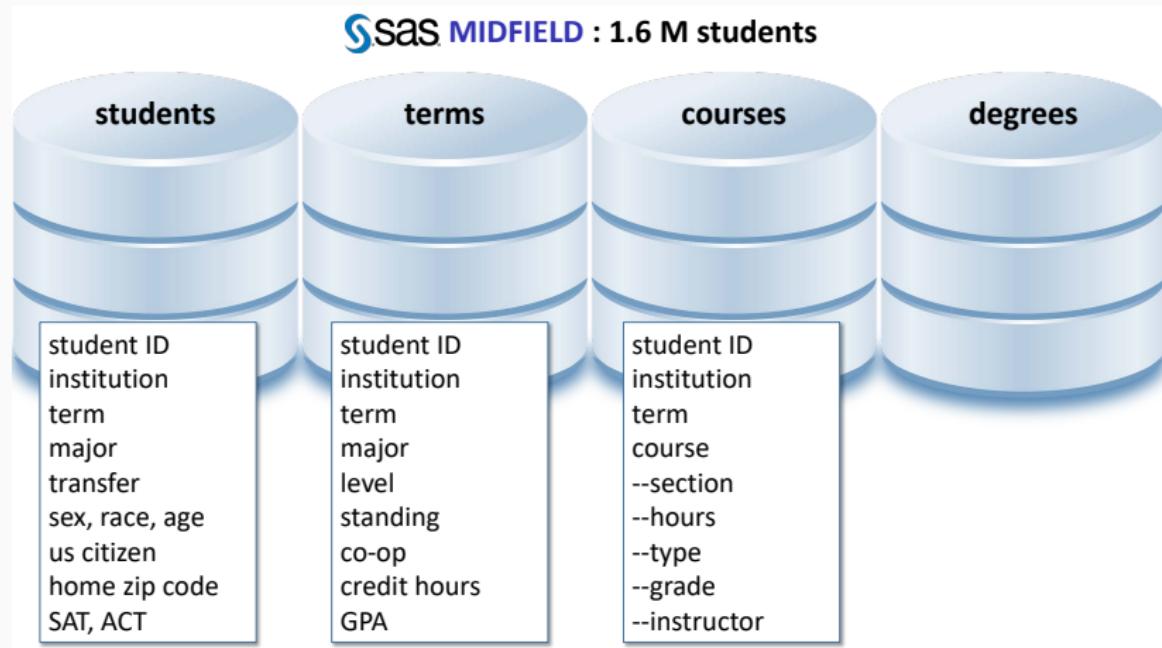


Each observation is one term for one student

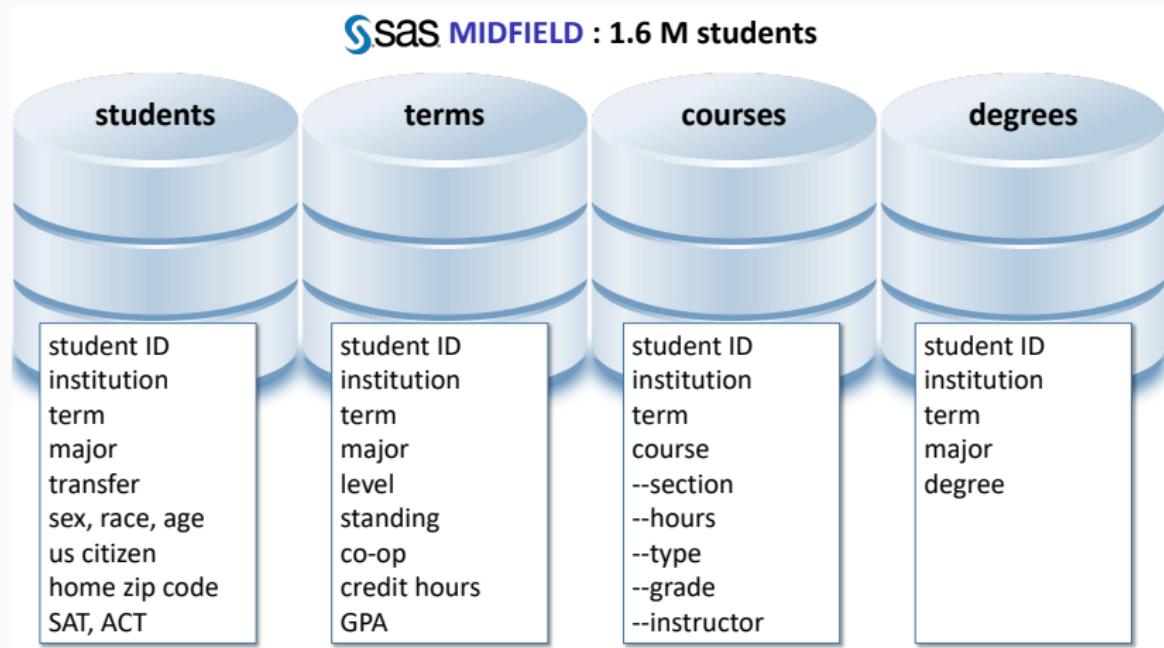
Sas MIDFIELD : 1.6 M students



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



MIDFIELD Workshops

- Introduction
- What is MIDFIELD?
- Publications
- Facilitators
- Licenses
- Acknowledgment
- About the MIDFIELD workshops
- What is midfieldr?
- Why R?
- Why R graphics?
- 2019 FIE Workshop
- Before you arrive
- Description
- Agenda
- Slides
- Tutorial: Starting with R**
- Tutorial: Starting with midfieldr
- Acknowledgements
- References

Tutorial: Starting with R

This is a self-paced tutorial.

- 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

1. Create an R project

Open RStudio.

- File > New Project > New Directory > New Project
- Fill in the Directory Name text box with 2019-FIE-midfieldr-workshop
- Select a location on your computer to save the project
- Check the *Open in a new session* box
- Click Create Project

The new project directory will be all of these things:

- a directory or "folder" on your computer
- an RStudio Project

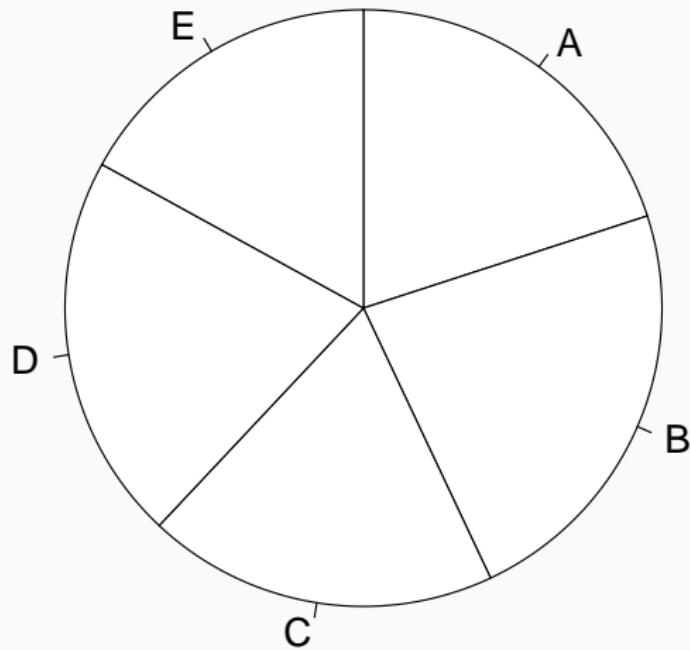
On your computer, if you navigate to the new project folder, you should have at least the following folders and files,

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

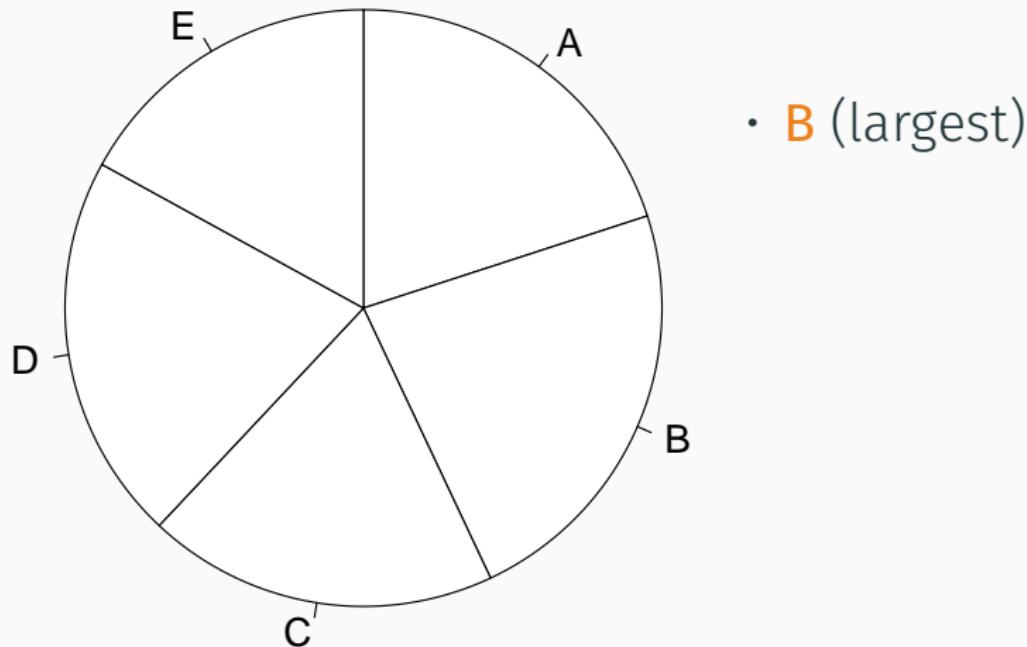
— break —

Elements of effective graphs

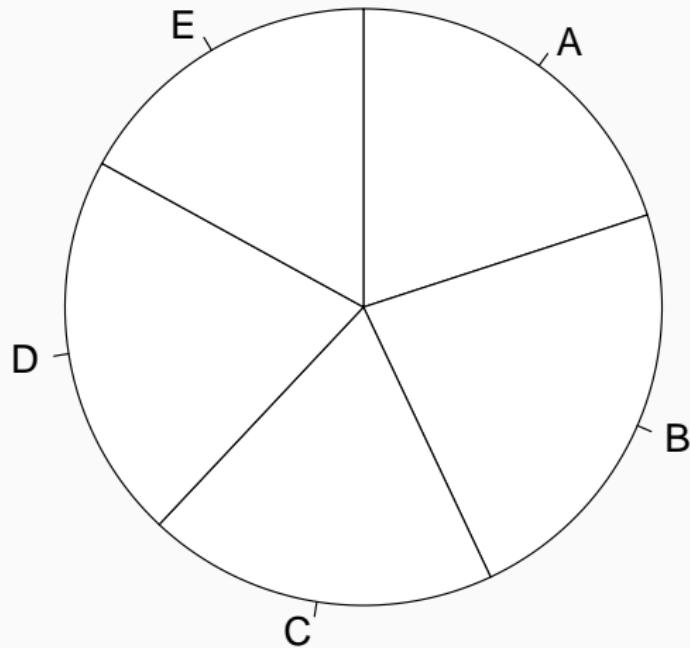
In your handout, list the slices A thru E from largest to smallest



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



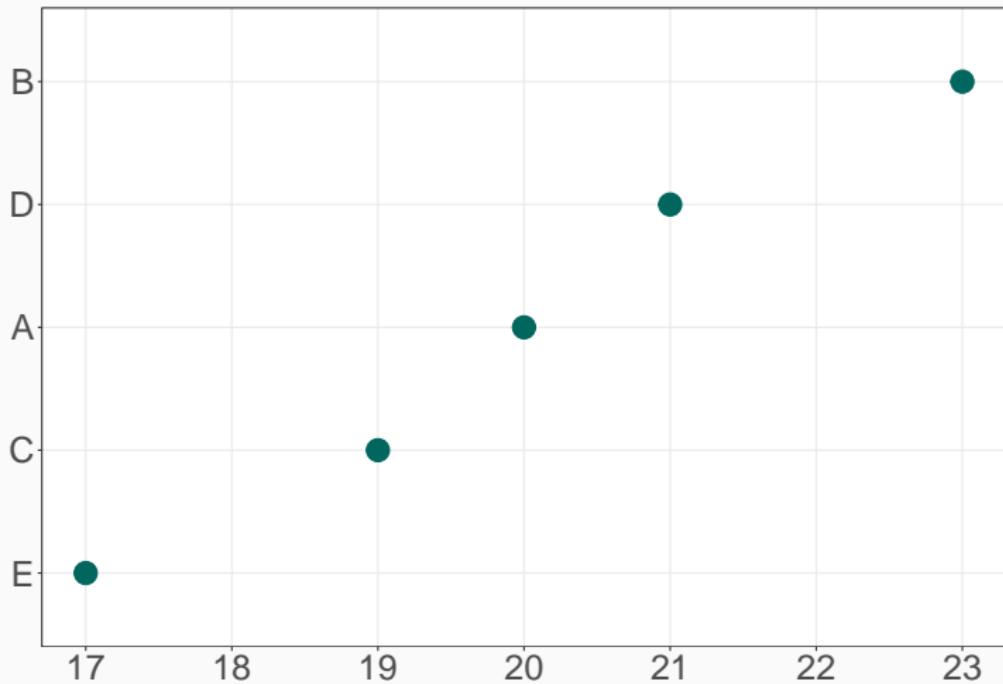
In your handout, list the slices A thru E from largest to smallest



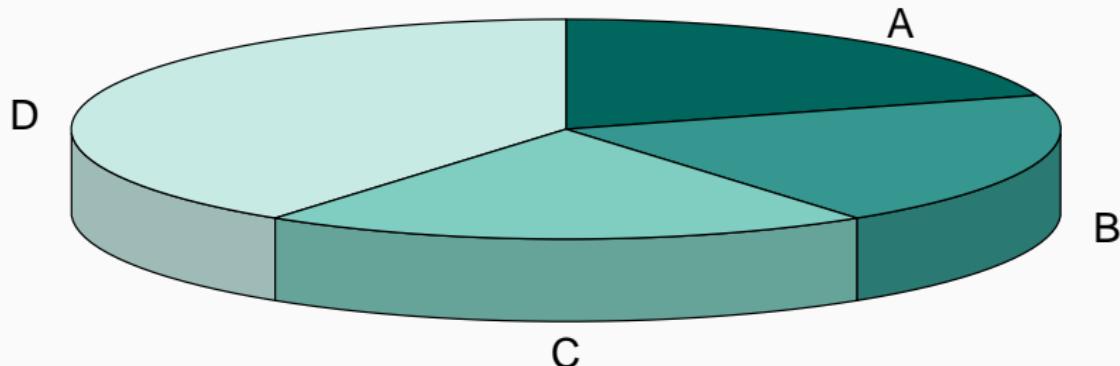
- B (largest)
- D
- A
- C
- E (smallest)

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

A. _____

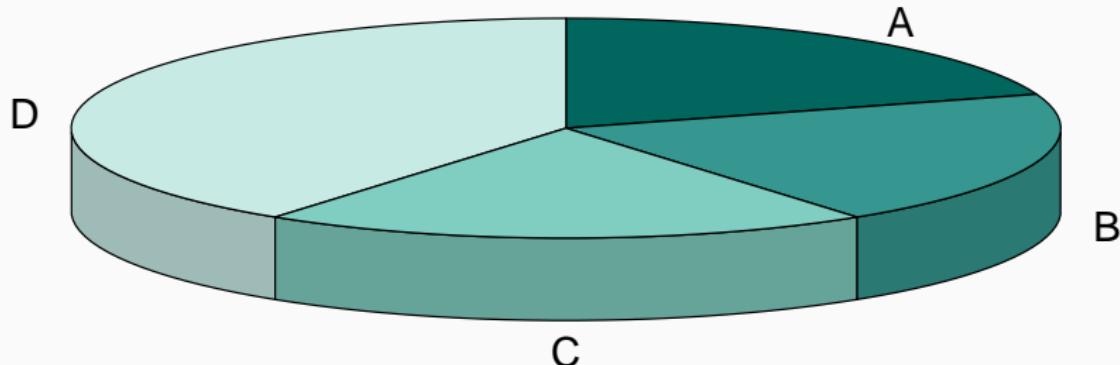
The total should be 100%

B. _____

C. _____

D. _____

3D-effects distort our judgment



Fill in the blanks

A. 20%

The total should be 100%

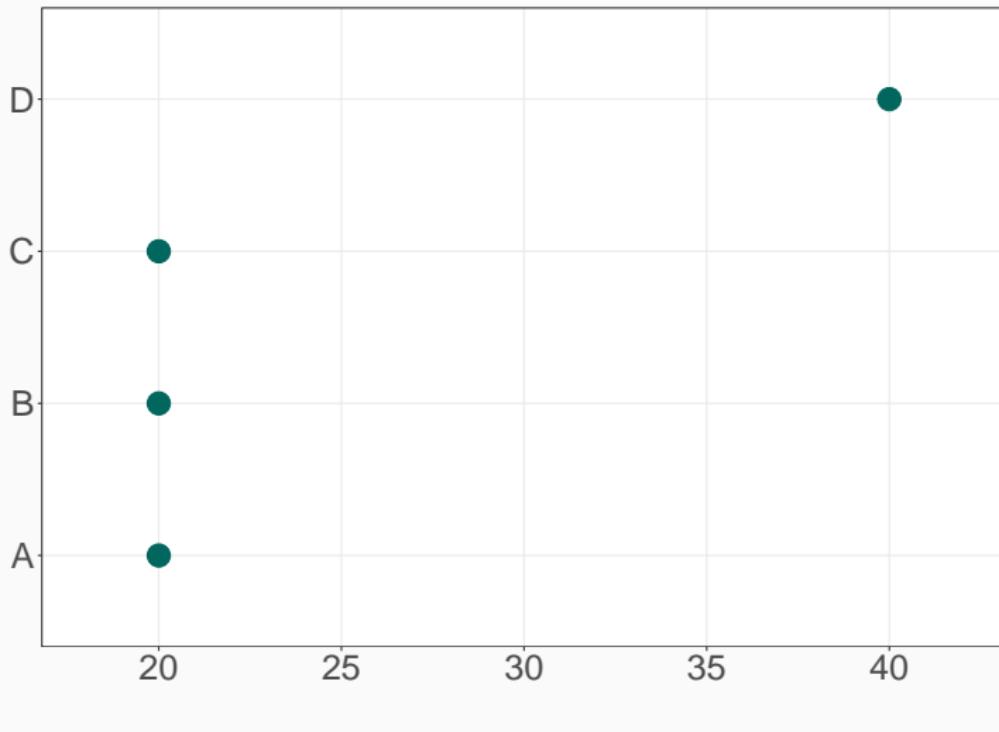
B. 20%

C. 20%

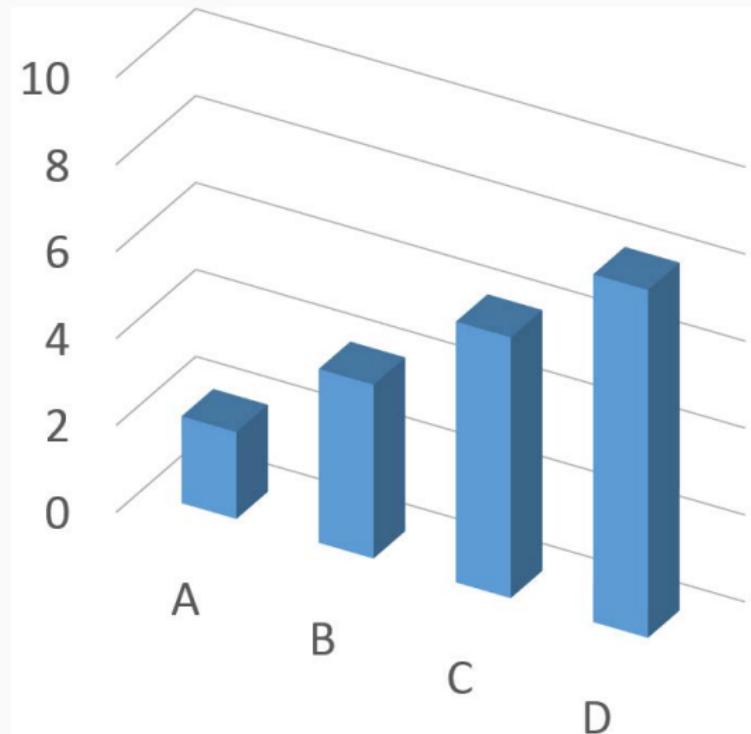
D. 40%

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

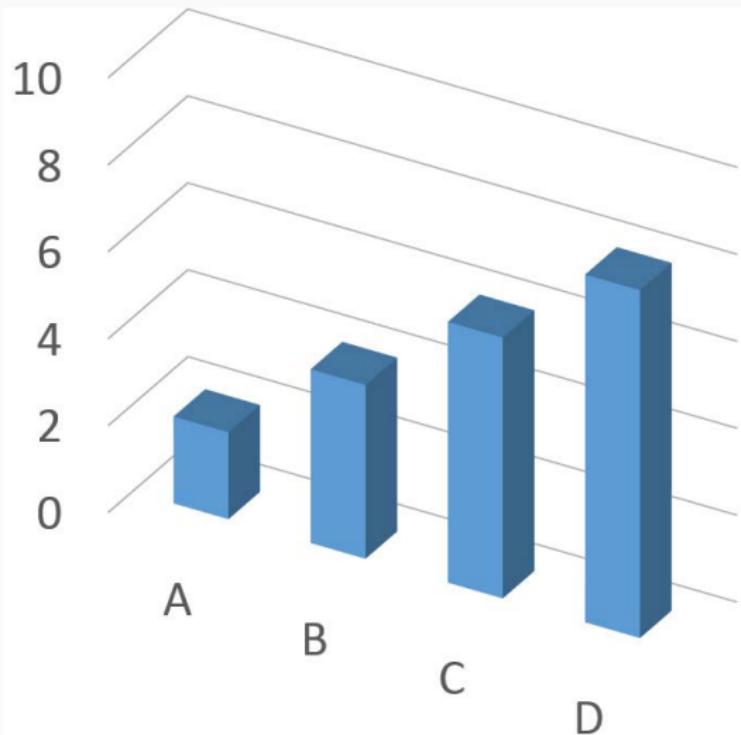
A. _____

B. _____

C. _____

D. _____

Again, 3D-effects distort our judgment



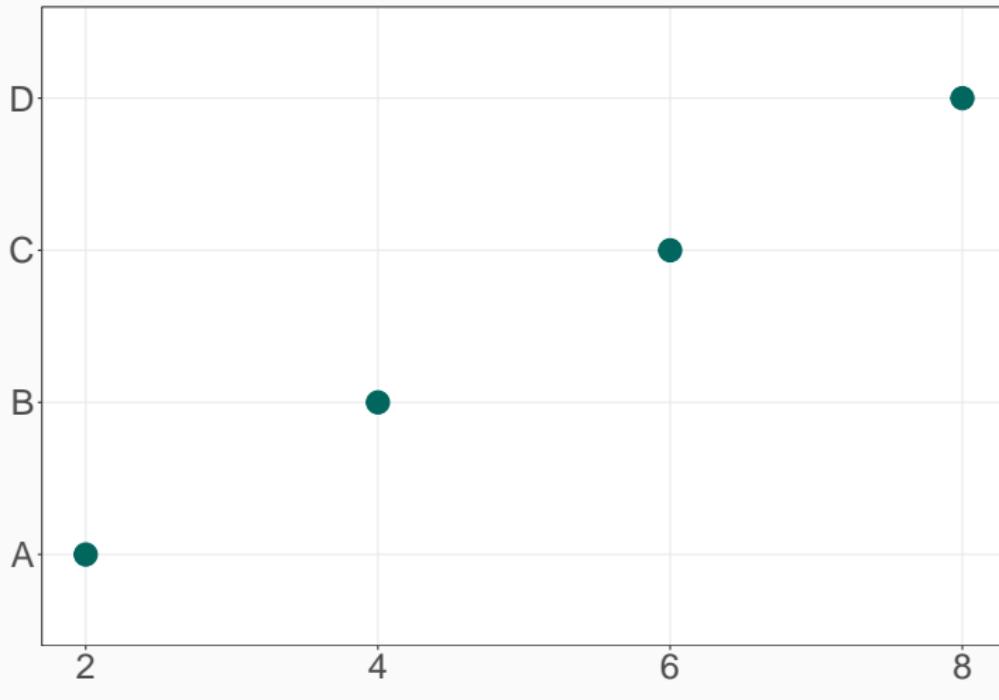
This is a visual inspection only.

Fill in the blanks

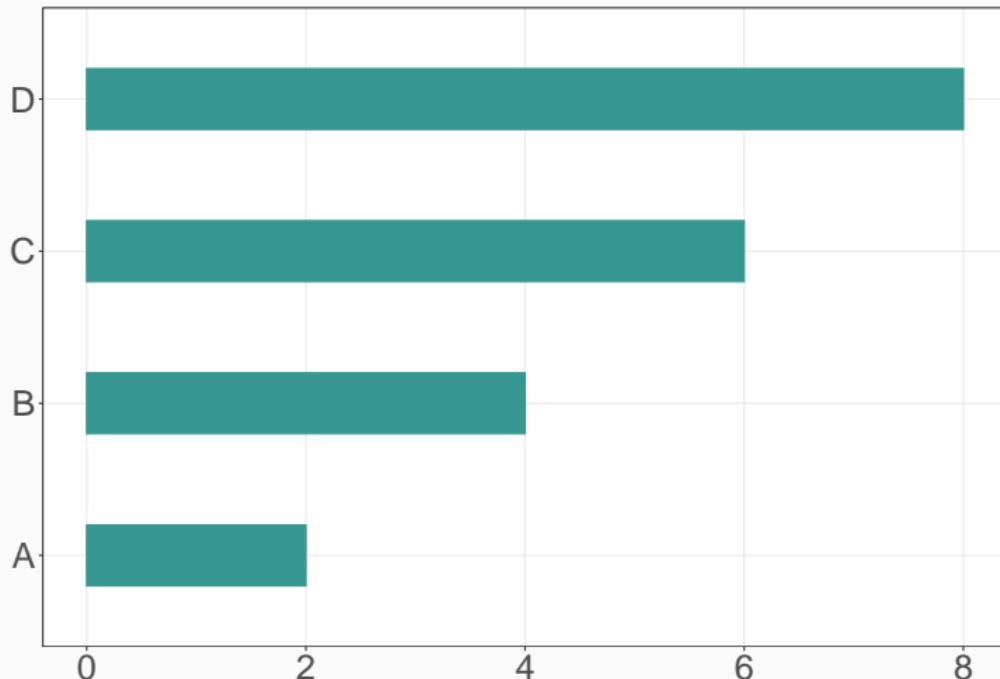
- A. 2
- B. 4
- C. 6
- D. 8

Again, the same data arranged along a common axis

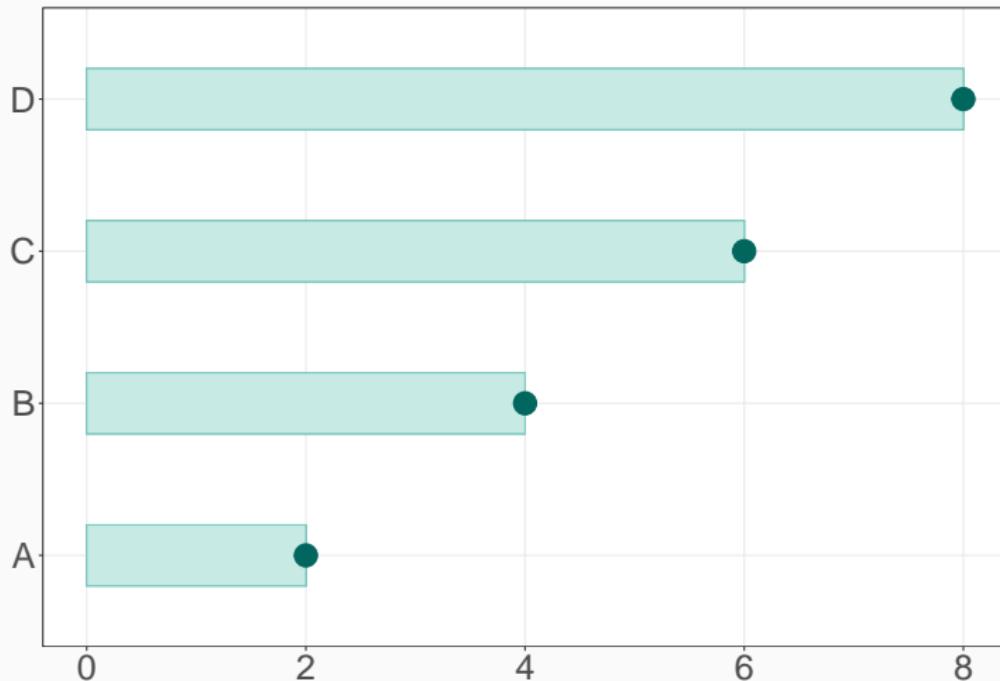
A high-accuracy visual task.



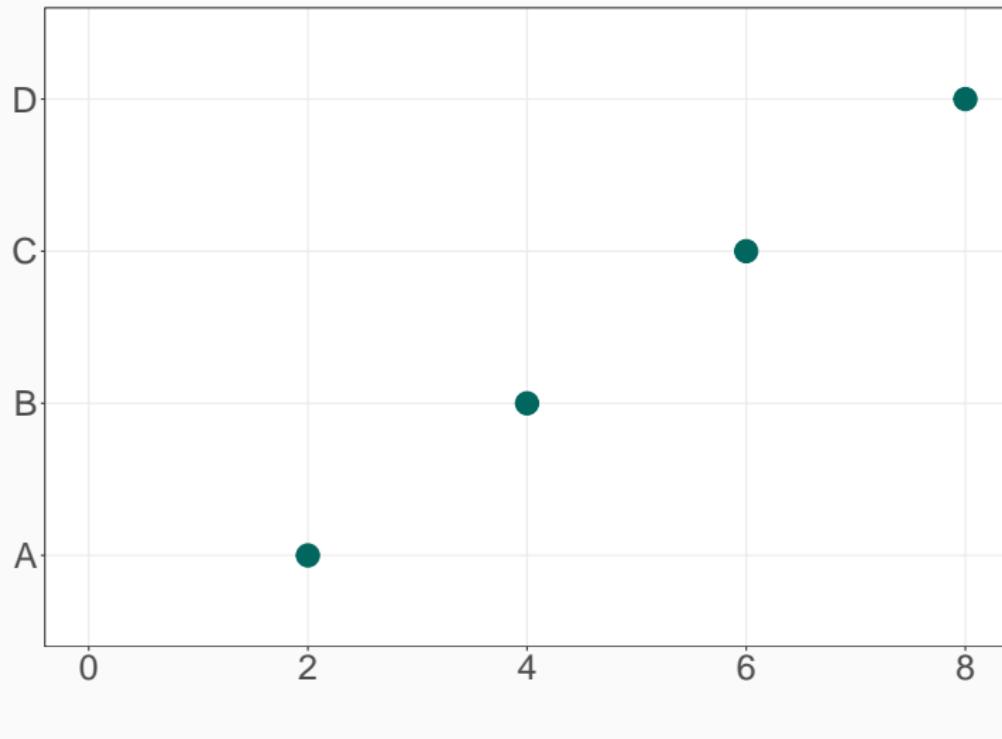
You can use bars, but must include zero



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) \approx 83,000 sq km.

FL _____ \times 1000 sq. km

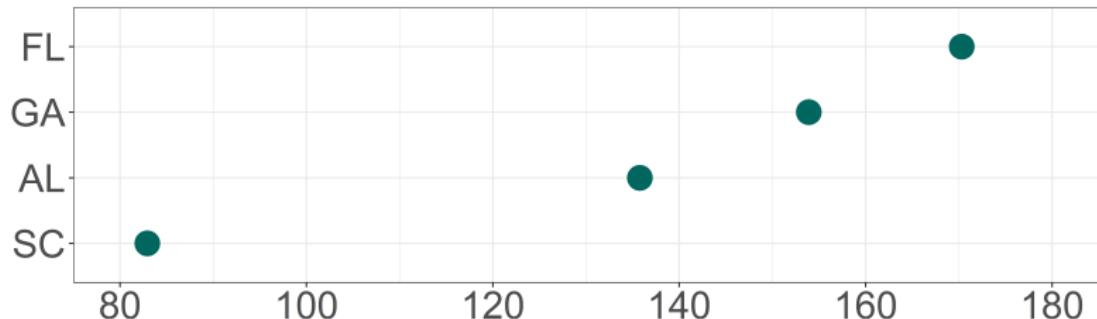
GA _____ \times 1000 sq. km

AL _____ \times 1000 sq. km

SC 83 \times 1000 sq. km

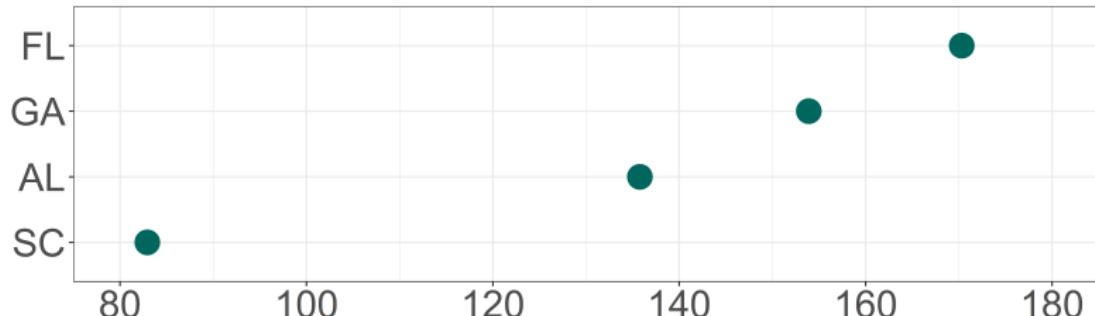
Adapted from (Ihaka 2007)

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

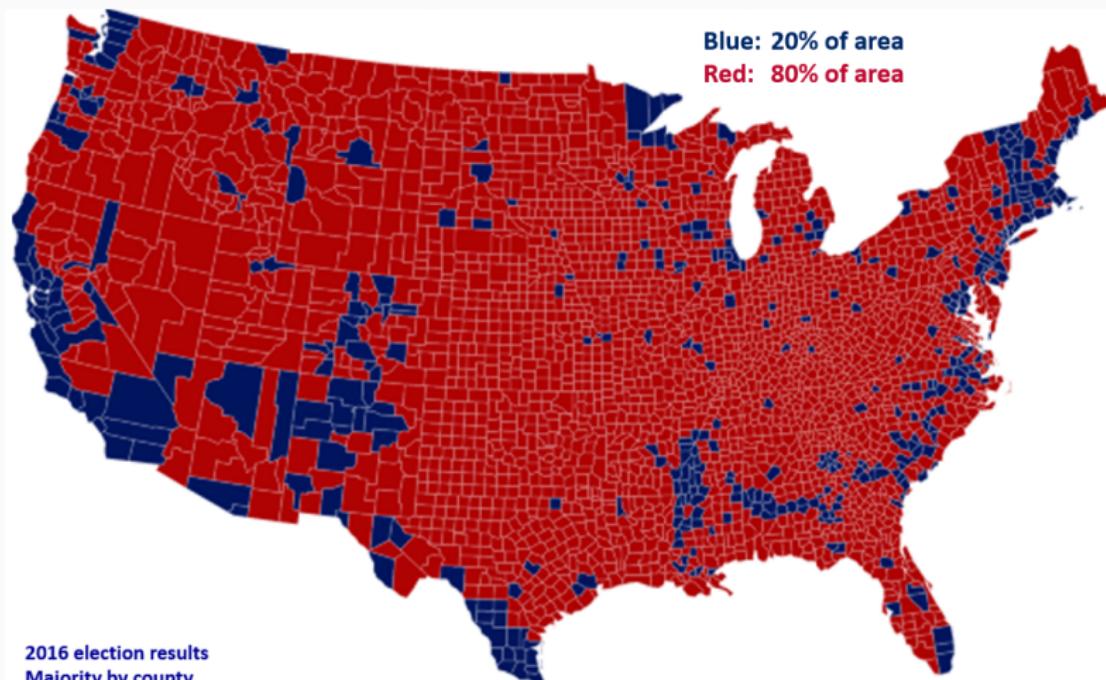
Your estimates have probably improved



- FL **170** x 1000 sq. km
- GA **154** x 1000 sq. km
- AL **136** x 1000 sq. km
- SC **83** x 1000 sq. km

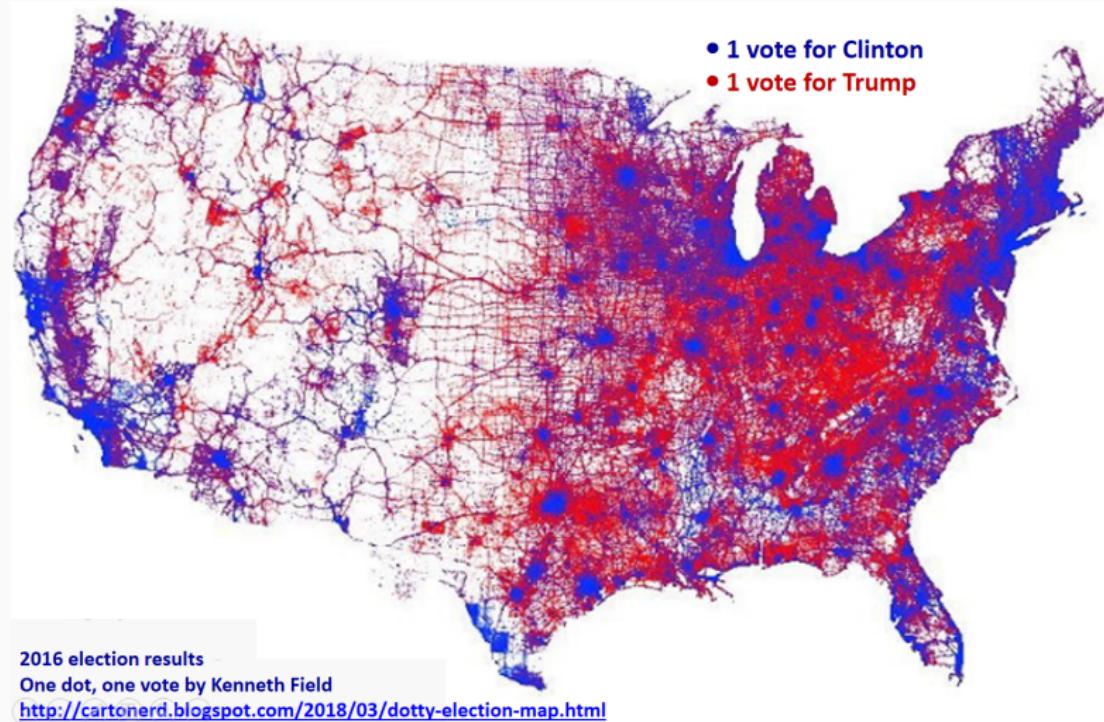
When color represents area, what story emerges?

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



When color represents voters?

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



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

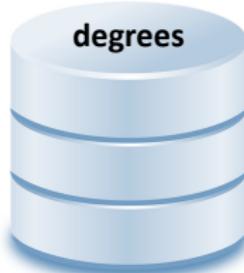
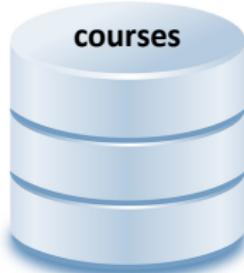
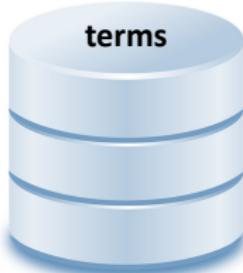
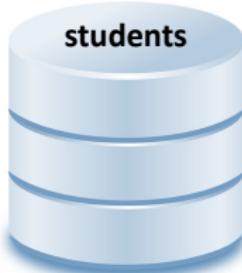
- new packages appear regularly—one might solve *your* problem
- anyone can help us find errors and add features to our packages

Introducing midfieldr

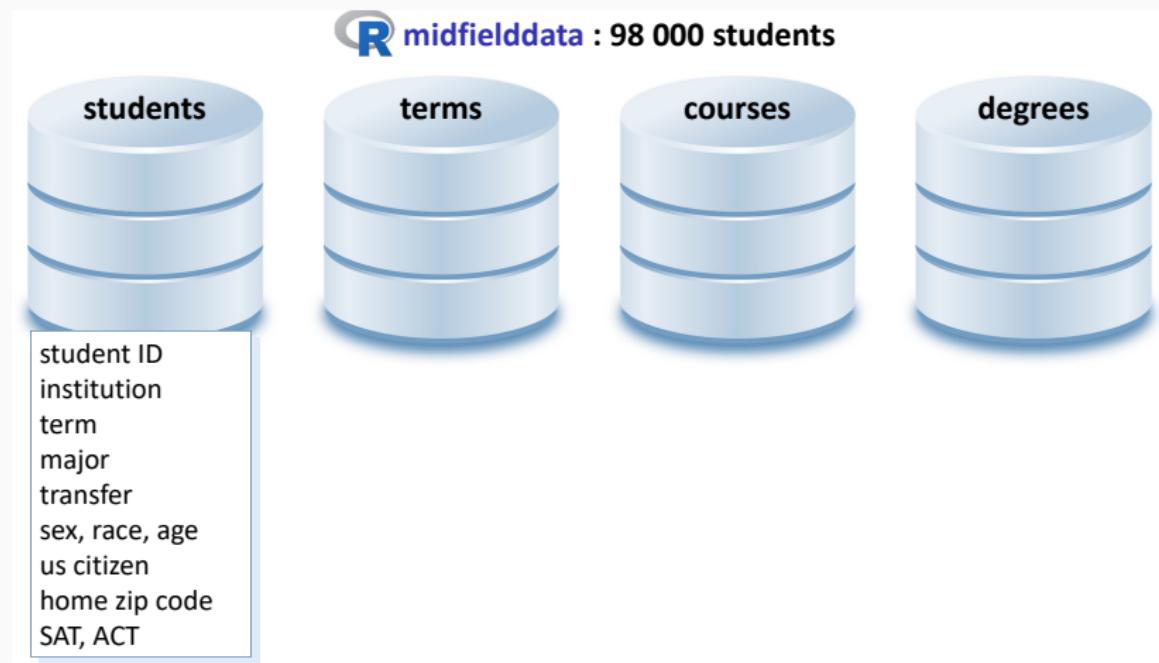
R package `midfielddata` provides a stratified sample



midfielddata : 98 000 students



Each observation is a unique student



midfieldstudents

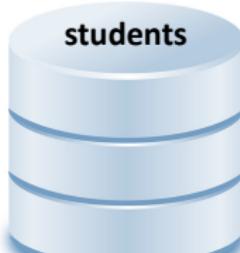
98,000 observations

19 Mb of memory

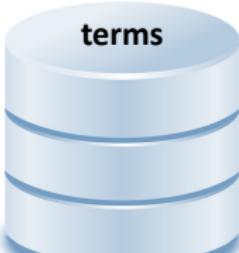
Each observation is one term for one student



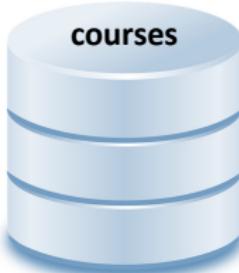
midfielddata : 98 000 students



student ID
institution
term
major
transfer
sex, race, age
us citizen
home zip code
SAT, ACT



student ID
institution
term
major
level
standing
co-op
credit hours
GPA



midfieldterms

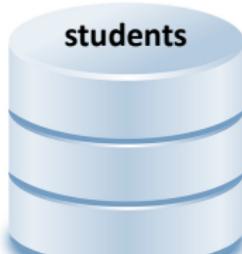
729,000 observations

82 Mb of memory

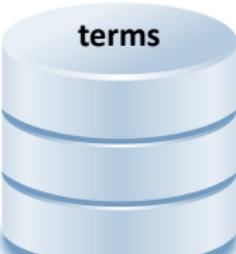
Each observation is one course for one student



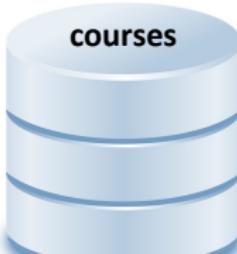
midfielddata : 98 000 students



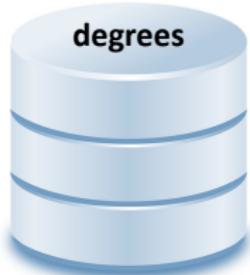
student ID
institution
term
major
transfer
sex, race, age
us citizen
home zip code
SAT, ACT



student ID
institution
term
major
level
standing
co-op
credit hours
GPA



student ID
institution
term
course
--section
--hours
--type
--grade
--instructor



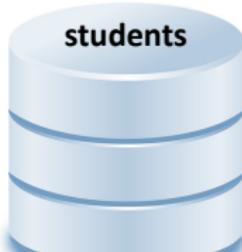
midfieldcourses

3.5 M observations

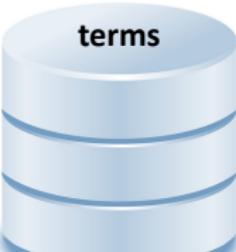
348 Mb of memory

Each observation is a unique student

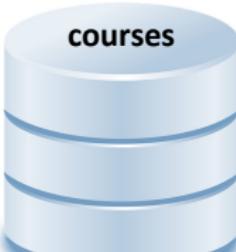
 **midfielddata** : 98 000 students



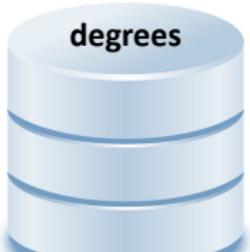
student ID
institution
term
major
transfer
sex, race, age
us citizen
home zip code
SAT, ACT



student ID
institution
term
major
level
standing
co-op
credit hours
GPA



student ID
institution
term
course
--section
--hours
--type
--grade
--instructor



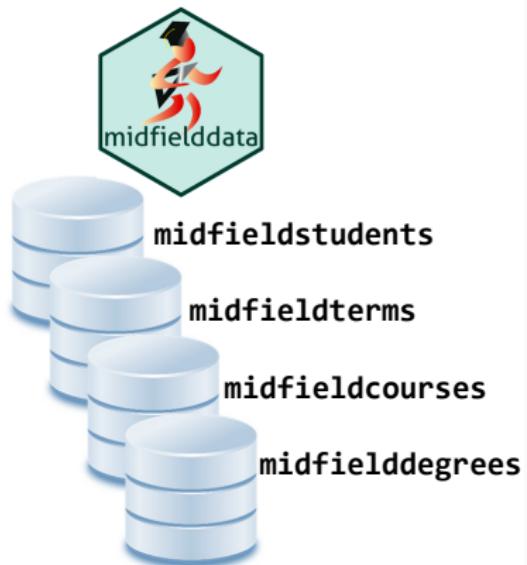
student ID
institution
term
major
degree

midfielddegrees

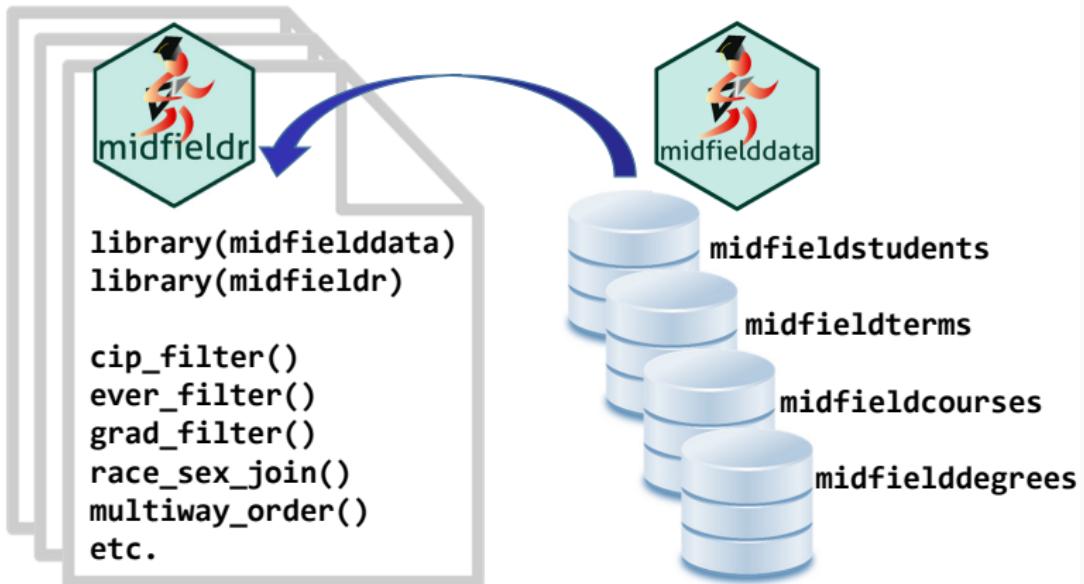
98,000 M observations

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



Publications
Facilitators
Licenses
Acknowledgment
About the MIDFIELD workshops
What is midfieldr?
Why R?
Why R graphics?
2019 FIE Workshop
Before you arrive
Description
Agenda
Slides
Tutorial: Starting with R
Tutorial: Starting with midfieldr
Acknowledgements
References



Tutorial: Starting with midfieldr

This is a self-paced tutorial illustrating functions in the `midfieldr` package.

- 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

Create a new R script

If you closed the project, then open `2019-FIE-midfieldr-workshop.Rproj`

- File > New File > R Script
- An Untitled script will open
- File > Save As
- Type in a new file name, for example, `start-with-midfieldr.R`
- Save

Your directory should look like this now,

```
2019-FIE-midfieldr-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

The screenshot shows the midfieldr website interface. At the top, there is a navigation bar with the text "midfieldr 0.1.0.9002", a blue house icon, "Getting started", "Vignettes" (which is highlighted in blue and has a red arrow pointing to it), "Papers/Talks", and "Ref". Below the navigation bar, the word "midfieldr" is displayed in large, bold, black letters. A horizontal line follows, and then the text "A package for investigating student record data provided by institutions participating in the MIDFIELD project." is shown. Further down, a paragraph explains the package's purpose: "Analytical tools for research in student pathways are generally scarce. midfieldr provides an entry to this type of intersectional research for anyone with basic proficiency in R and familiarity with packages from the tidyverse." To the right of the main content area, a white box contains a list of vignette titles: "Identifying programs by CIP codes", "Selecting groups of programs", "Imputing starting majors for FYE students", "Computing graduation rate", "Computing stickiness", and "Multiway data, graphs, and tables". A green hexagonal logo with the letters "ldr" is partially visible on the right side.

midfieldr 0.1.0.9002

Getting started

Vignettes

Papers/Talks

Ref

midfieldr

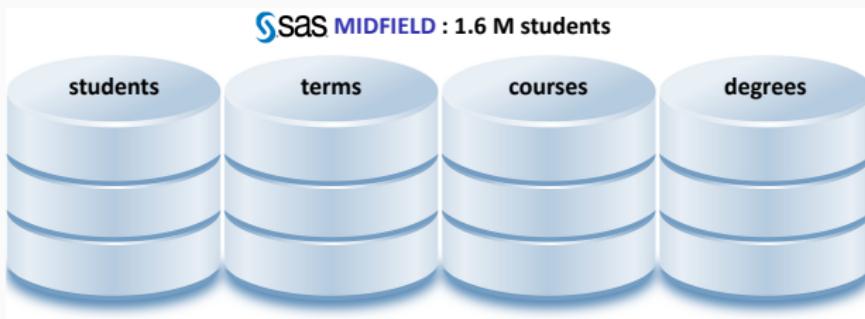
A package for investigating student record data provided by institutions participating in the MIDFIELD project.

Analytical tools for research in student pathways are generally scarce. midfieldr provides an entry to this type of intersectional research for anyone with basic proficiency in R and familiarity with packages from the tidyverse.

- Identifying programs by CIP codes
- Selecting groups of programs
- Imputing starting majors for FYE students
- Computing graduation rate
- Computing stickiness
- Multiway data, graphs, and tables

ldr

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

The screenshot shows the MIDFIELD website's navigation bar at the top, featuring links for "midfieldr 0.1.0.9002", "Getting started", "Vignettes", "Papers/Talks", "Reference", "News", and "About". A red arrow points to the "About" link. Below the navigation bar, the main content area has a title "About MIDFIELD" and a source note "Source: vignettes/pages/about_midfield.Rmd". To the right, a sidebar titled "Contents" lists "Who we are" and "Funding", with "Who we are" being highlighted in blue. The main content area contains two paragraphs about MIDFIELD's purpose and scope.

About MIDFIELD

Source: [vignettes/pages/about_midfield.Rmd](#)

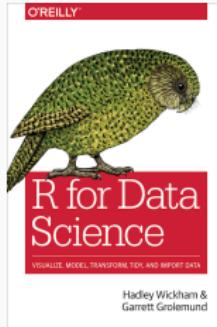
MIDFIELD is the acronym for the Multiple-Institution Database for Investigating Engineering Longitudinal Development.

Longitudinal studies using MIDFIELD are not limited to Engineering—the database includes *all* student records from member institutions.

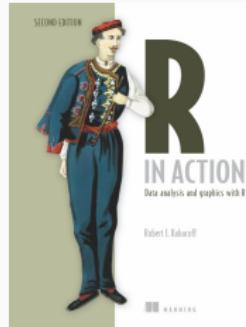
Who we are

Matthew Ohland	Principal Investigator/Director	ohland@purdue.edu
Marisa Orr	Associate Director	marisak@clemson.edu
Catherine Brawner	Director Policy Analysis	brawnerc@bellsouth.net
Russell Long	Managing Director/Data Steward	ralong@purdue.edu
Susan Lord	Director MIDFIELD Institute	slord@sandiego.edu
Richard Layton	Director Data Display	layton@rose-hulman.edu

Next steps in learning R



Hadley Wickham
Garrett Grolemund



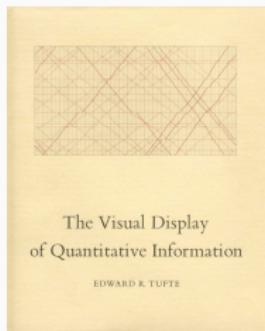
Robert I. Kabacoff

R ggplot2 theme_set colorblind or dichromat color?
In R ggplot2 are you able to theme_set(theme_gray() + "colorblind_function" at the top of the code rather than adding + scale_color_ each plot? Some ggthemes I've used in the past: sh ...
asked May 9 at 21:19

R ggplot2 geom_jitter: plotting all zeros
In R ggplot2, when I plot all zeros and use geom_jitter(), some variations are automatically added to zeros. How can I undo that? I st at 0.0 y axis. $y = rnorm(100)$ $x = rnorm(100)$ "A":B ...
asked Jun 3 '16 at 0:00

StackExchange.com
Or just google it
Your problem may already be solved

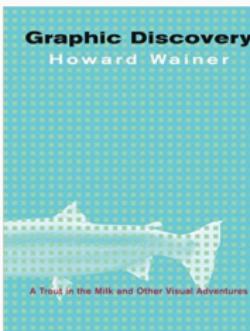
Next steps in learning about graph design



The Visual Display
of Quantitative Information

EDWARD R. TUFTE

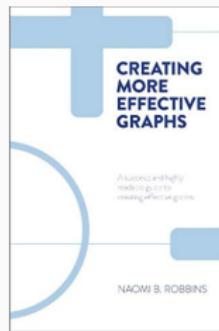
Edward Tufte



Graphic Discovery
Howard Wainer

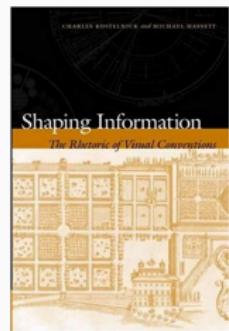
A Trout in the Milk and Other Visual Adventures

Howard Wainer



NAOMI B. ROBBINS

Naomi Robbins



Shaping Information

The Rhetoric of Visual Compositions

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.