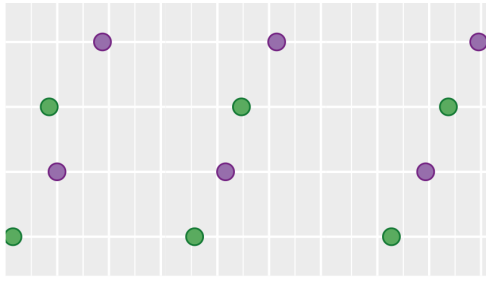
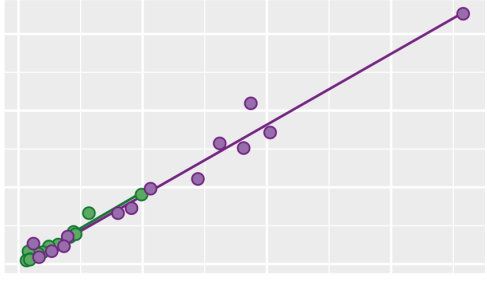
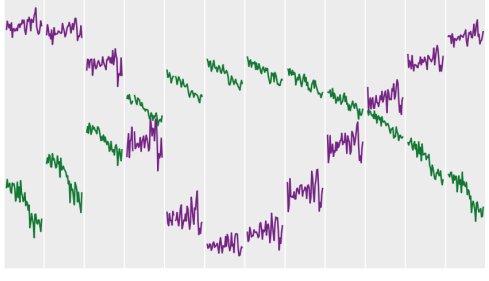
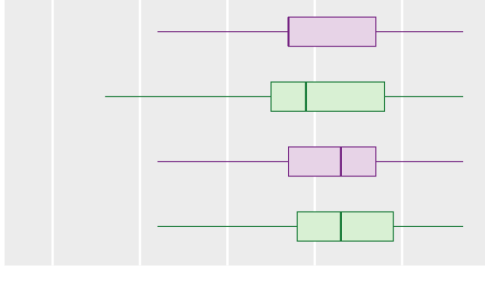








# Expanding your graphical repertoire

## Variables, design, message


2024 MIDFIELD Institute

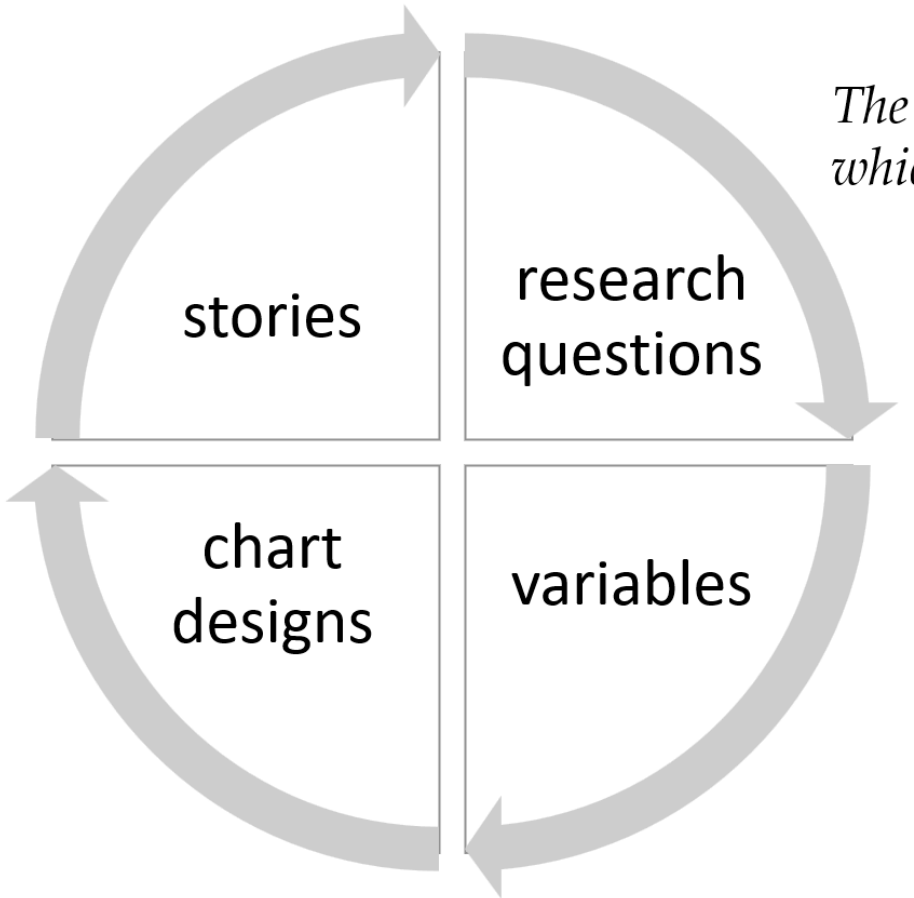
Richard Layton

https://www.graphdoctor.com

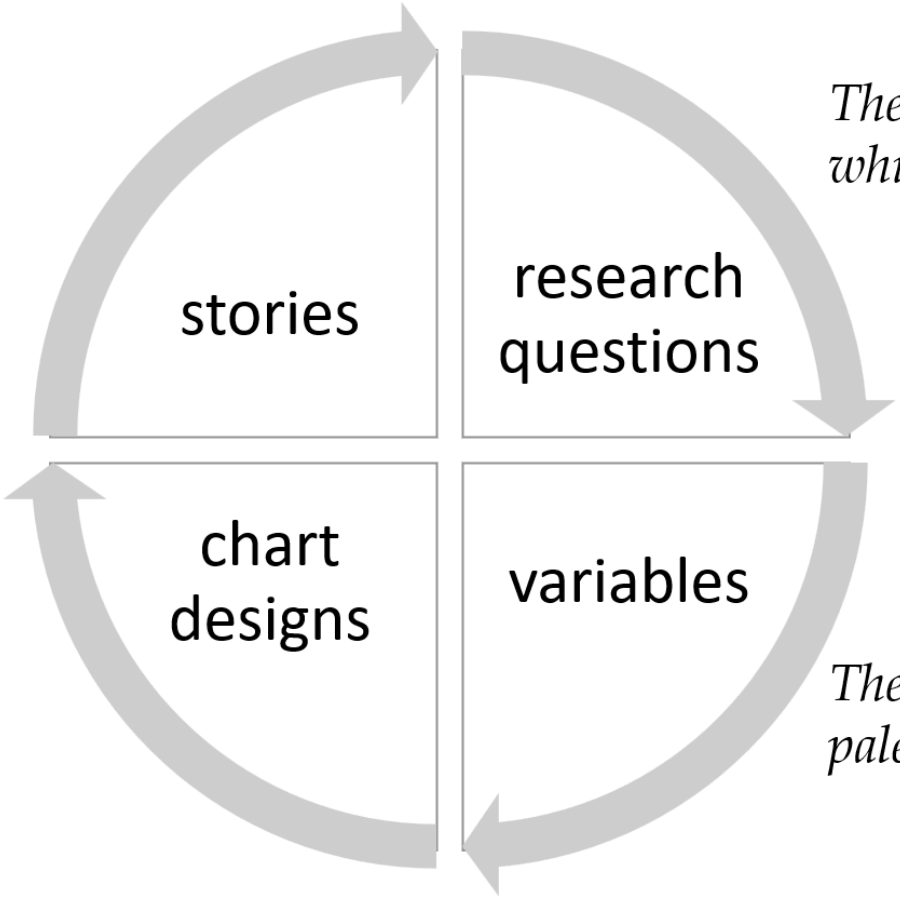
https://github.com/graphdr

graphdoctor@gmail.com

# Why graphical repertoire matters

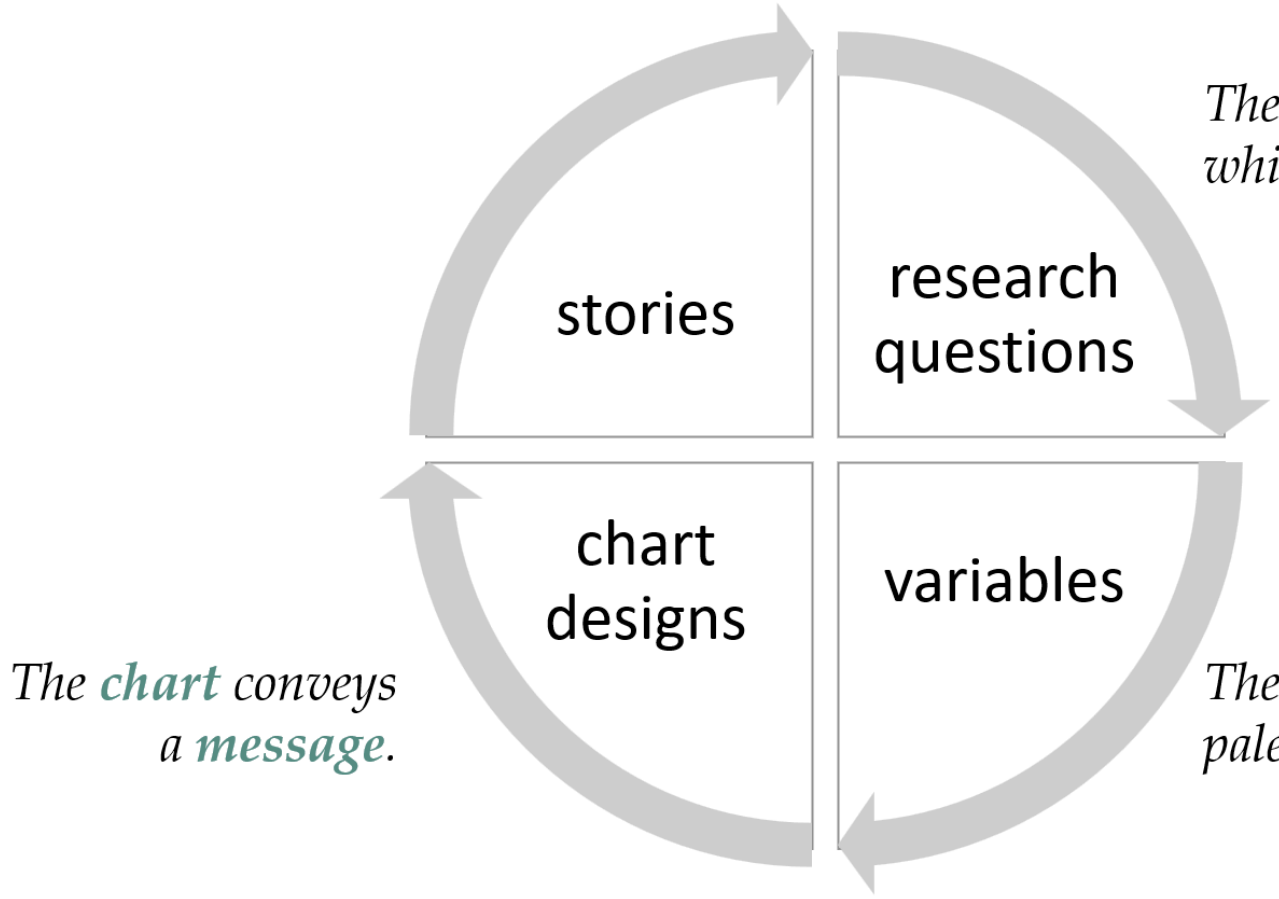



*The **questions** suggest which **data** to collect.*




*The **questions** suggest which **data** to collect.*

*The **data** suggest the palette of **chart** types.*



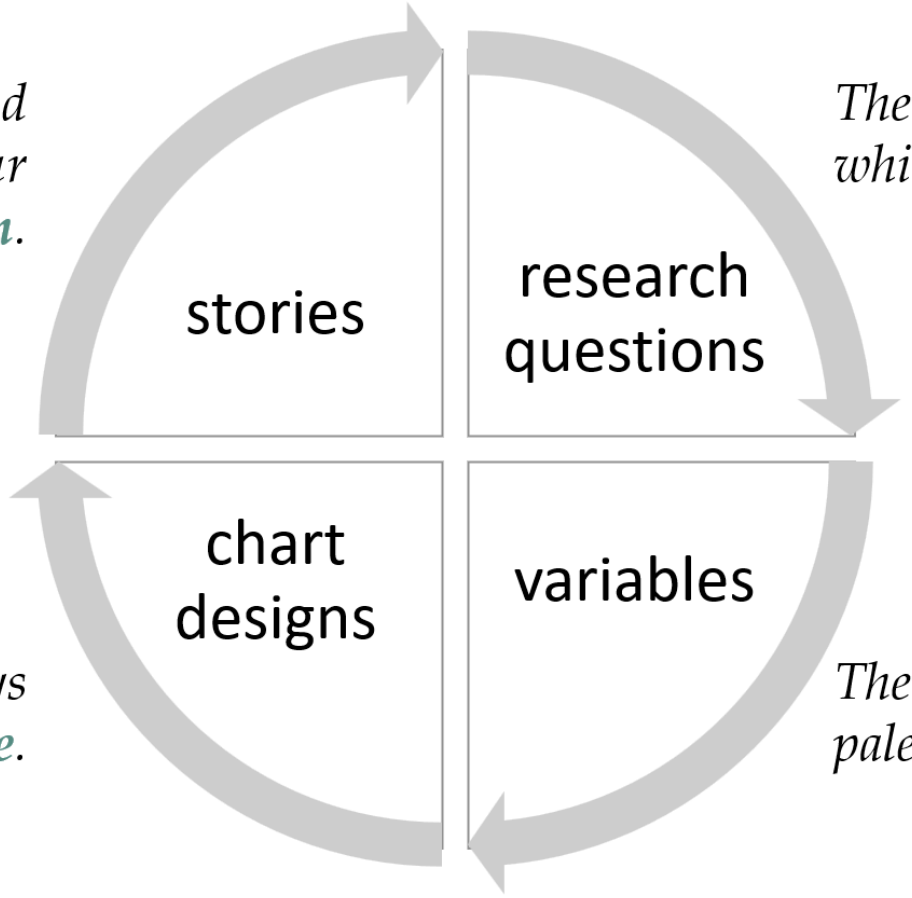

*The **questions** suggest which **data** to collect.*

*The **data** suggest the palette of **chart** types.*

*The **chart** conveys a **message**.*

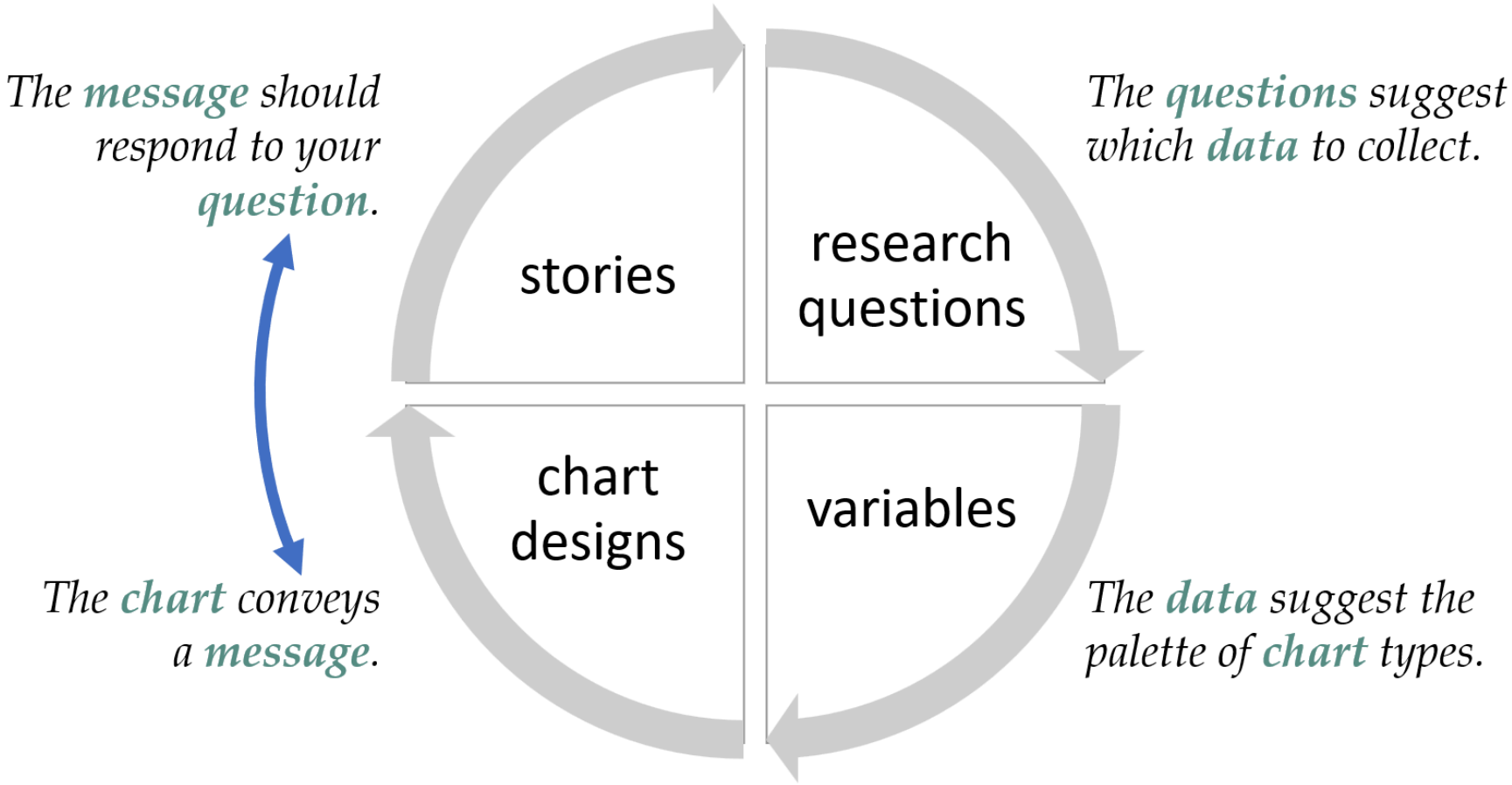
*The **message** should respond to your **question**.*

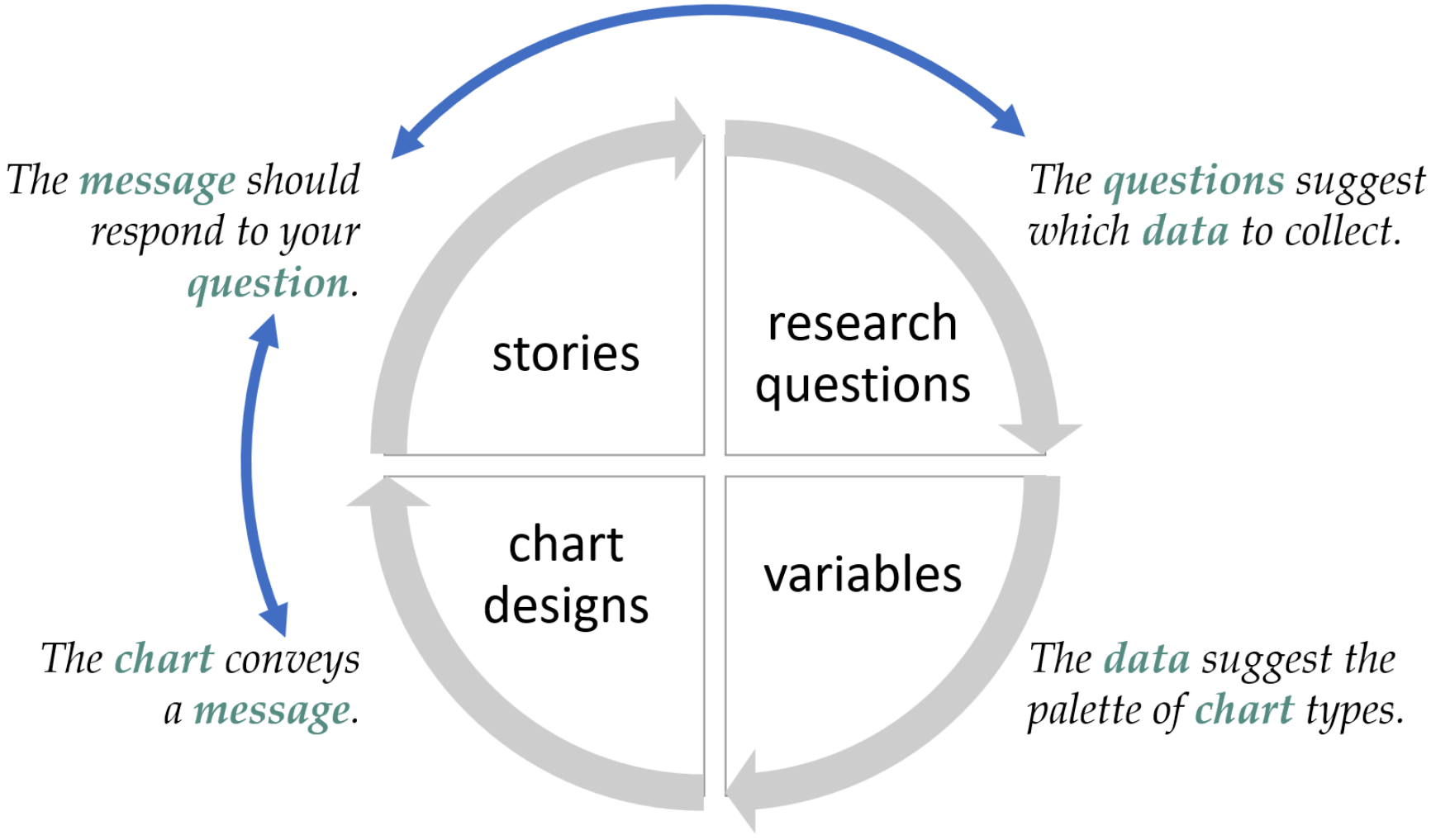
*The **questions** suggest which **data** to collect.*

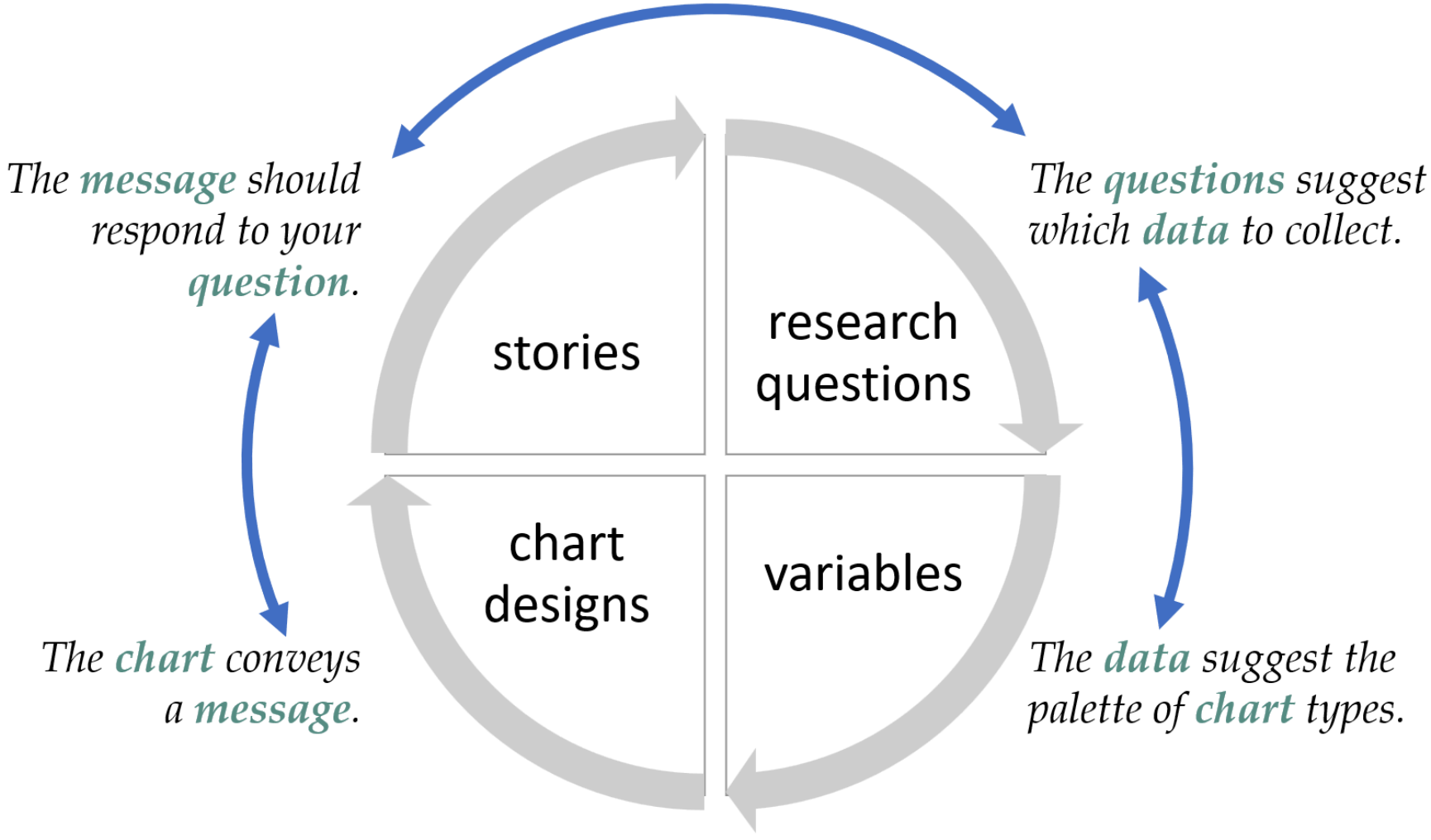


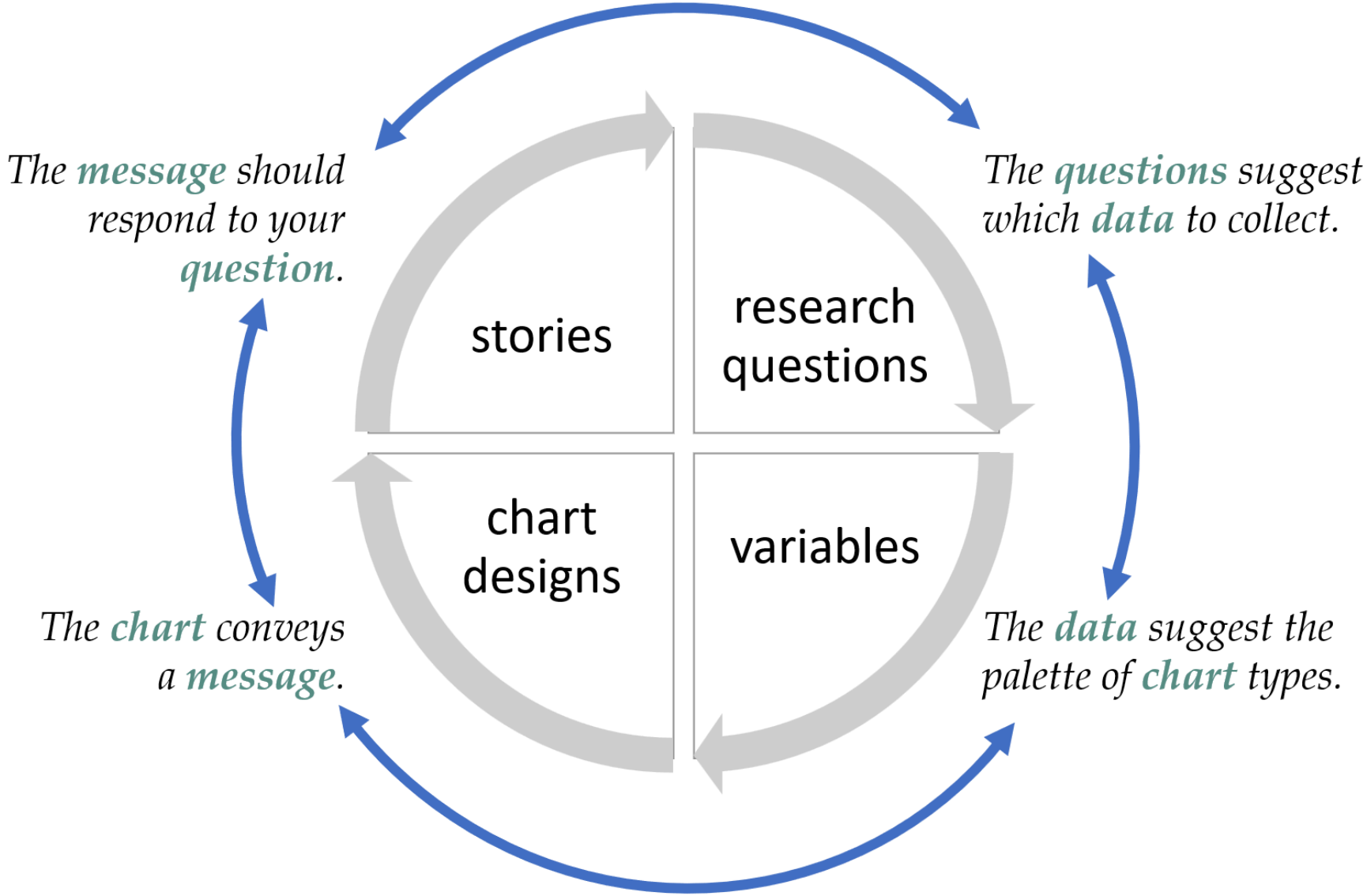

*The **chart** conveys a **message**.*

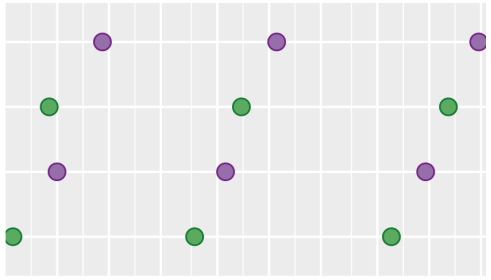
*The **data** suggest the palette of **chart** types.*

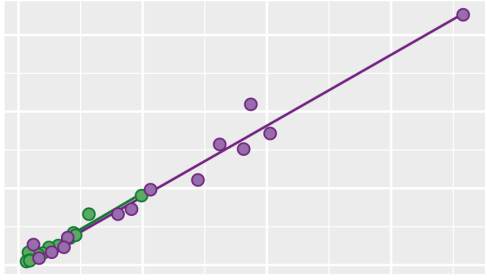

The message should respond to your question.
The questions suggest which data to collect.
stories
research questions
chart designs
variables
The chart conveys a message.
The data suggest the palette of chart types.

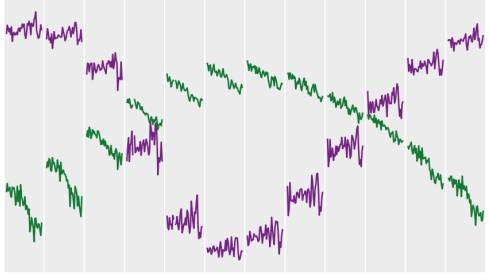

The message should respond to your question.
The questions suggest which data to collect.
stories
research questions
chart designs
variables
The chart conveys a message.
The data suggest the palette of chart types.

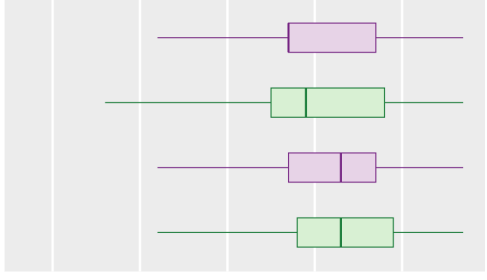

The message should respond to your question.
The questions suggest which data to collect.
stories
research questions
chart designs
variables
The chart conveys a message.
The data suggest the palette of chart types.


The message should respond to your question.
The questions suggest which data to collect.
stories
research questions
chart designs
variables
The chart conveys a message.
The data suggest the palette of chart types.





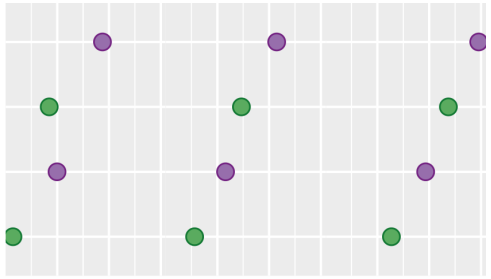




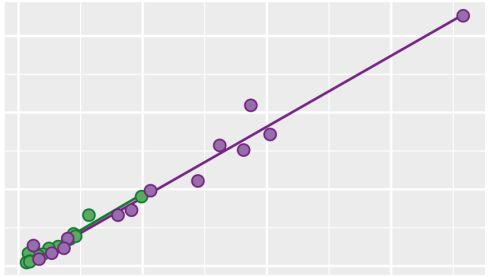
Expanding your repertoire

expands your ability to discover

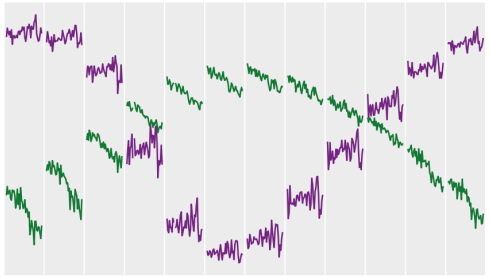
and visualize stories in your data.



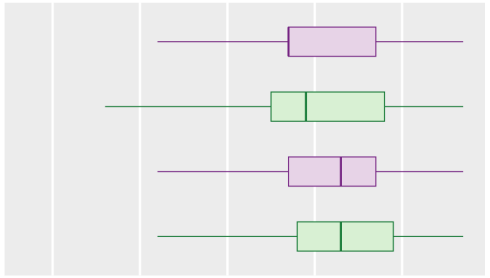
**Comparing quantities**



**Revealing correlations**



**Showing evolution**



**Displaying distributions**



Optimal design categories from: Jean-luc Doumont (2009) *Trees, Maps and Theorems*, Principiae, p. 133.

# Comparing quantities

# Data

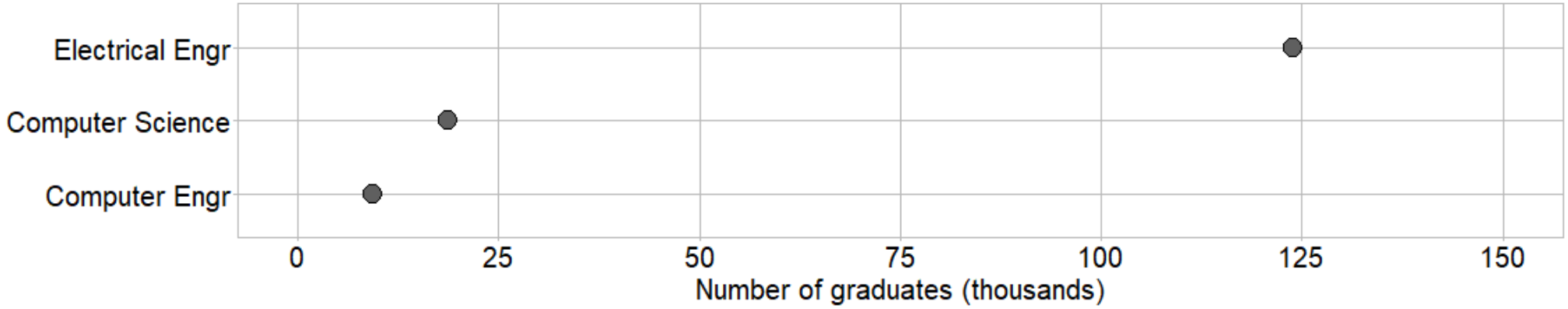
Representation at graduation in 3 engineering programs, 19 US institutions, 1987–2018

| Origin | Sex | Electrical Engr | Computer Engr | Computer Science |
|---|---|---|---|---|
| Domestic | Female | 23426 | 702 | 2923 |
| Domestic | Male | 90150 | 7481 | 13987 |
| International | Female | 1865 | 140 | 365 |
| International | Male | 8530 | 993 | 1442 |



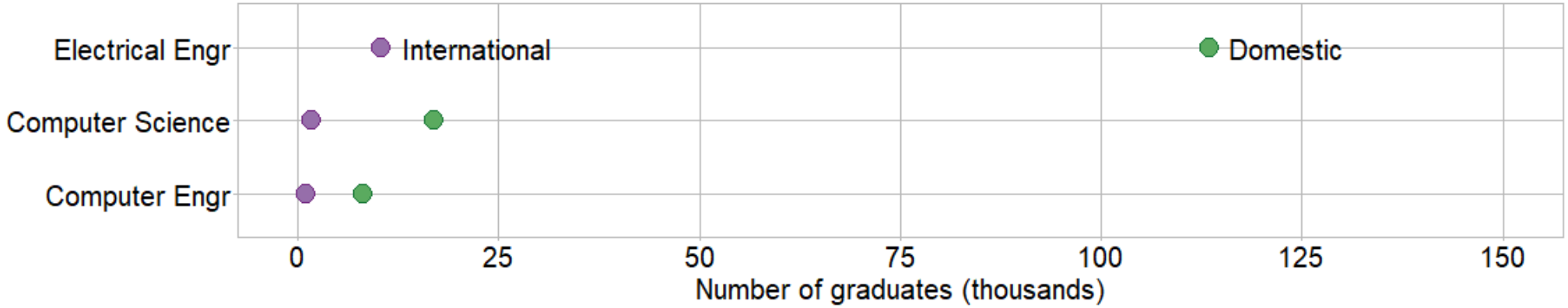
Data source: 2022 MIDFIELD

# Dot chart



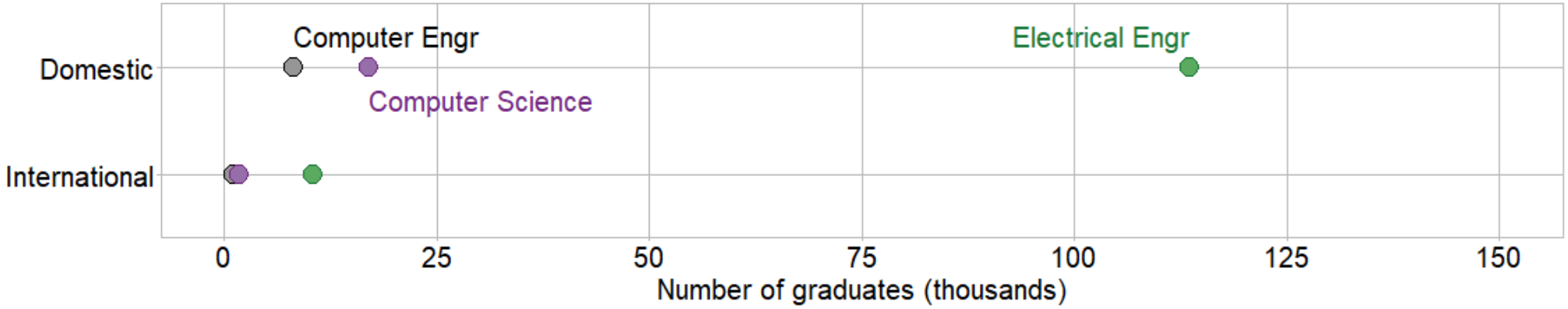

| variable | type |
|---|---|
| program | categorical |
| count of graduates | quantitative |

# Add a second category

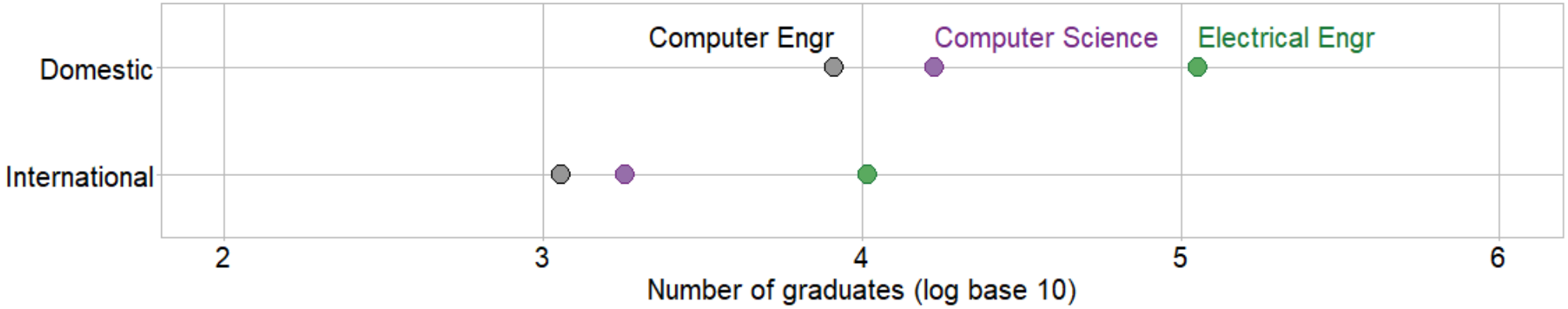



| variable | type |
|---|---|
| program | categorical |
| origin | categorical |
| count of graduates | quantitative |

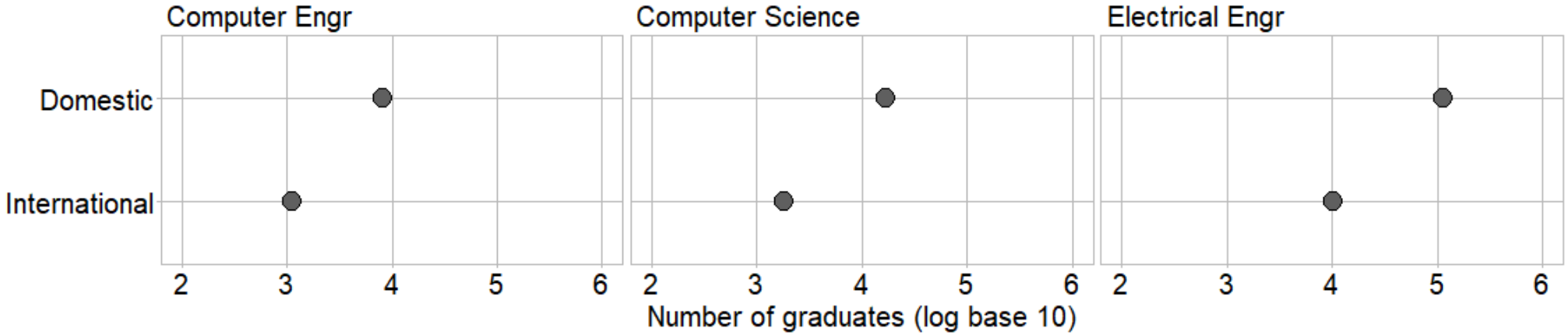
# Exchange mapping of categorical variables



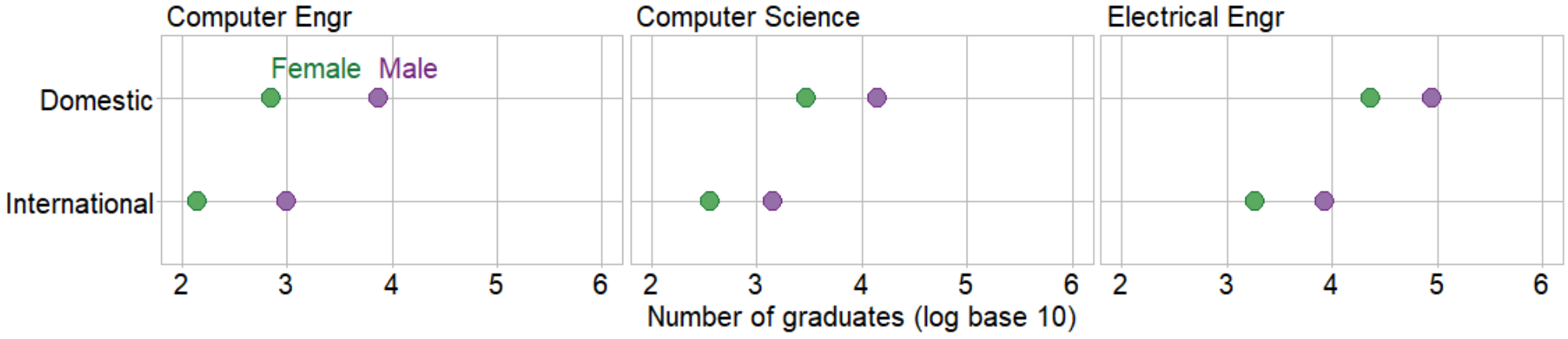
# Logarithmic scale for orders of magnitude differences



# One program per facet

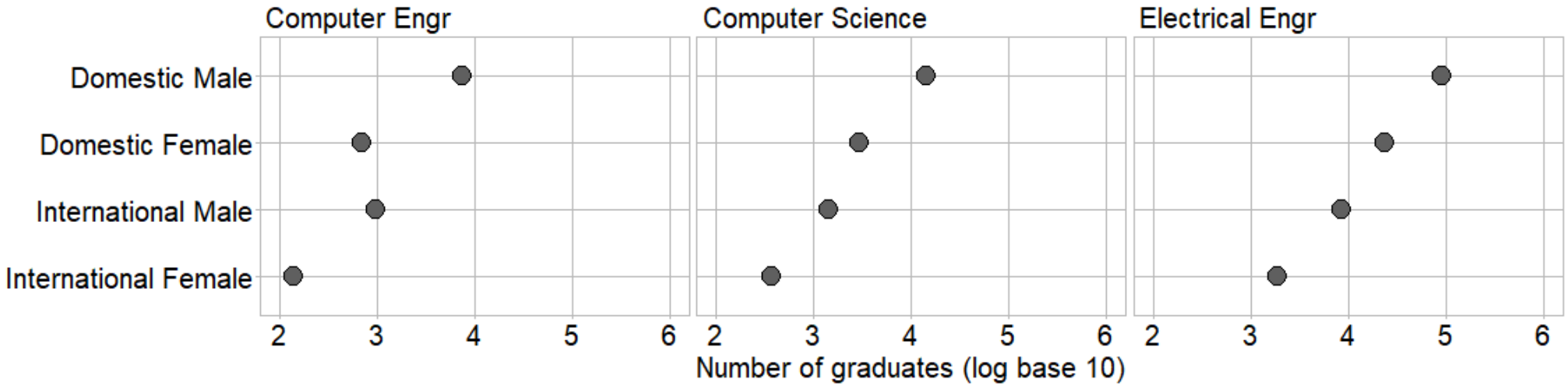


# Add a third category



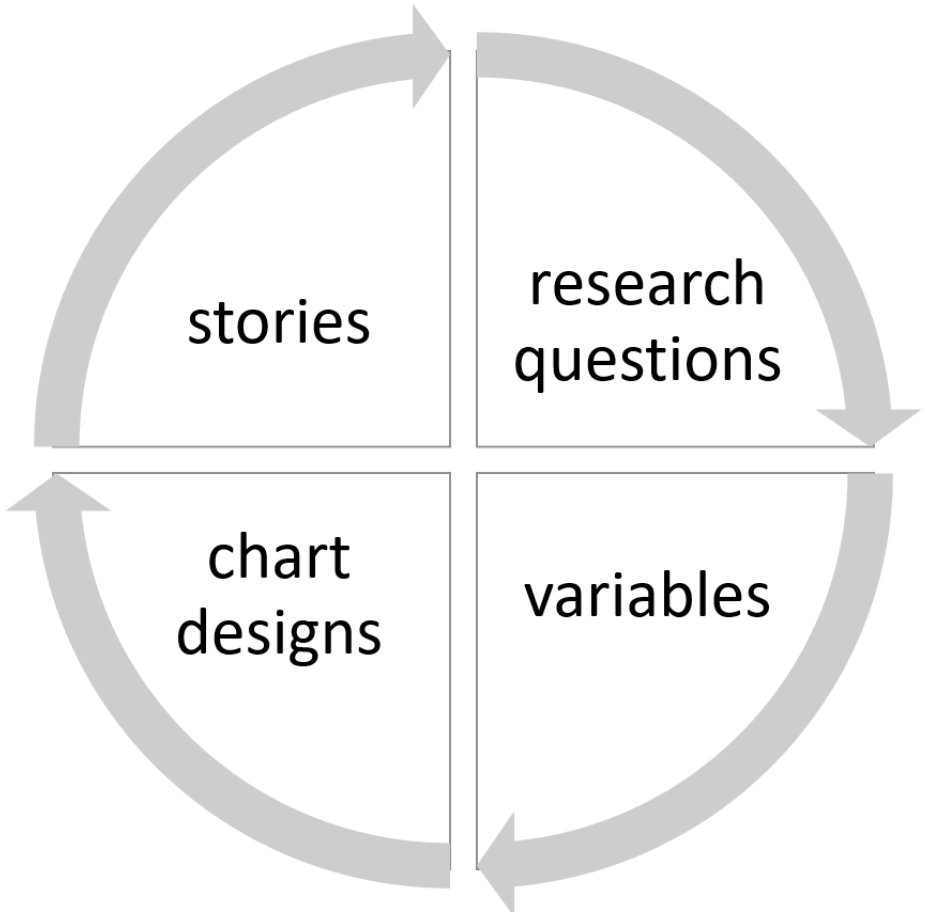

| variable | type |
|---|---|
| program | categorical |
| origin | categorical |
| sex | categorical |
| count of graduates | quantitative |

# Combine categories




| variable | type |
|---|---|
| origin/sex | categorical |
| program | categorical |
| count of graduates | quantitative |

# Discussion




## Comparing quantities

What points seem most important to you so far?

# Revealing correlations

# Data

Engineering students at 14 institutions persisting to year 4 and graduating by year 6, 1987--2019

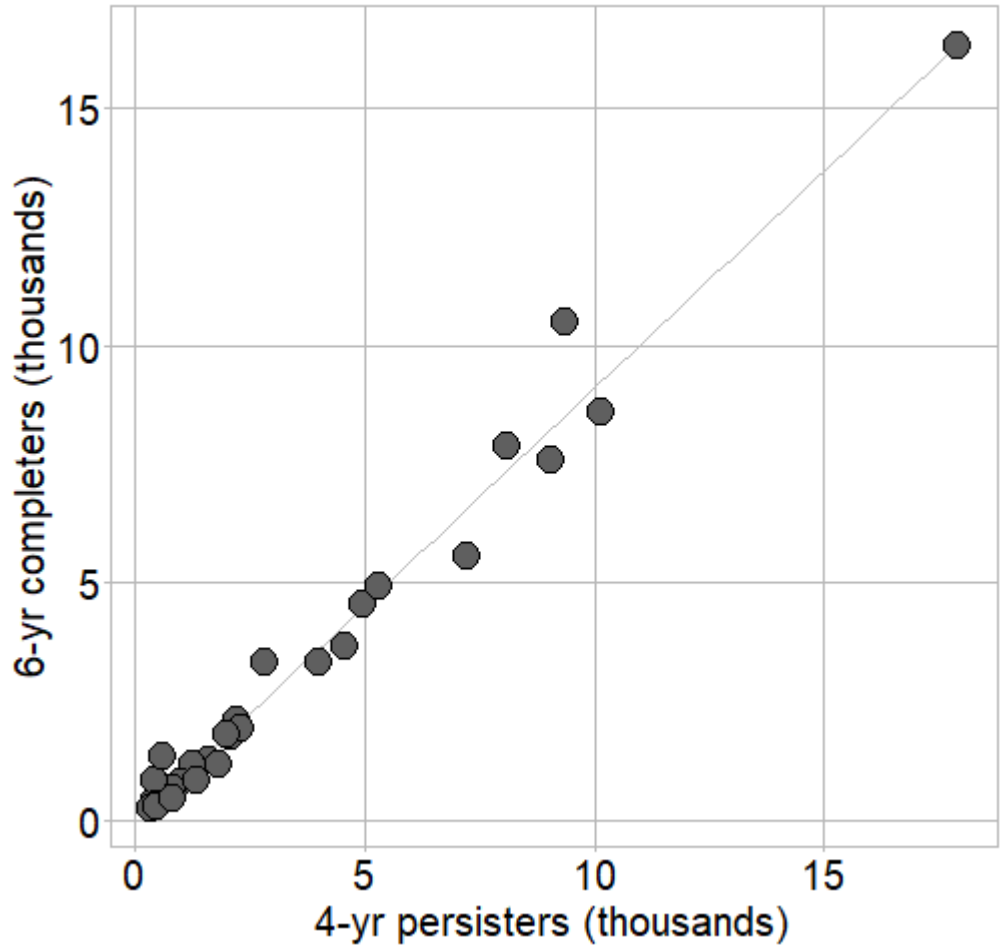
```
    institution    sex    y4    y6
         <char> <char> <int> <int>
 1:           A Female  4953  4525
 2:           A   Male 17897 16312
 3:           B Female  2834  3316
 4:           B   Male  9351 10473
 5:           C Female  2071  1764
 6:           C   Male 10128  8575
 7:           D Female  2217  2096
 8:           D   Male  8099  7863
---
21:           L Female   401   824
22:           L   Male   602  1332
23:           M Female   462   319
24:           M   Male  1829  1160
25:           N Female   322   228
26:           N   Male  1338   838
27:           P Female   457   283
28:           P   Male   827   447
```

| variable | type |
|---|---|
| institution | categorical |
| sex | categorical |
| 4-yr persisters | quantitative |
| 6-yr completers | quantitative |

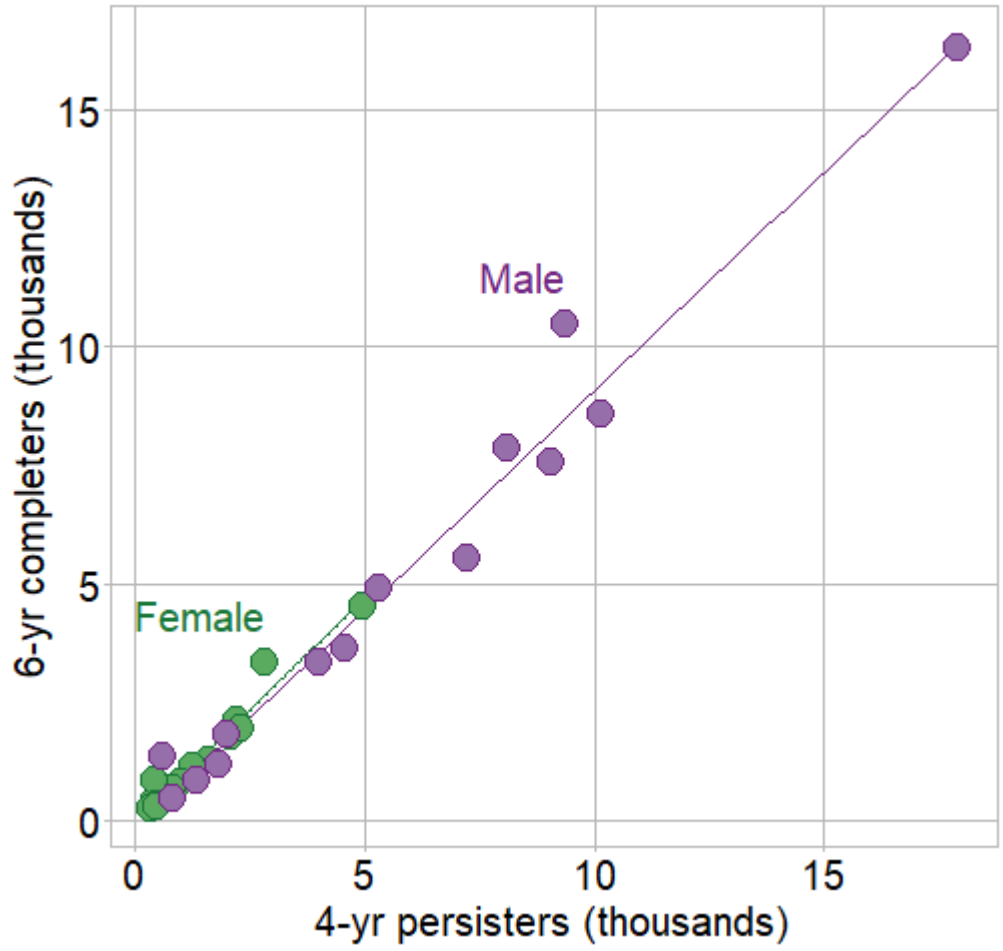


Data source: 2022 MIDFIELD

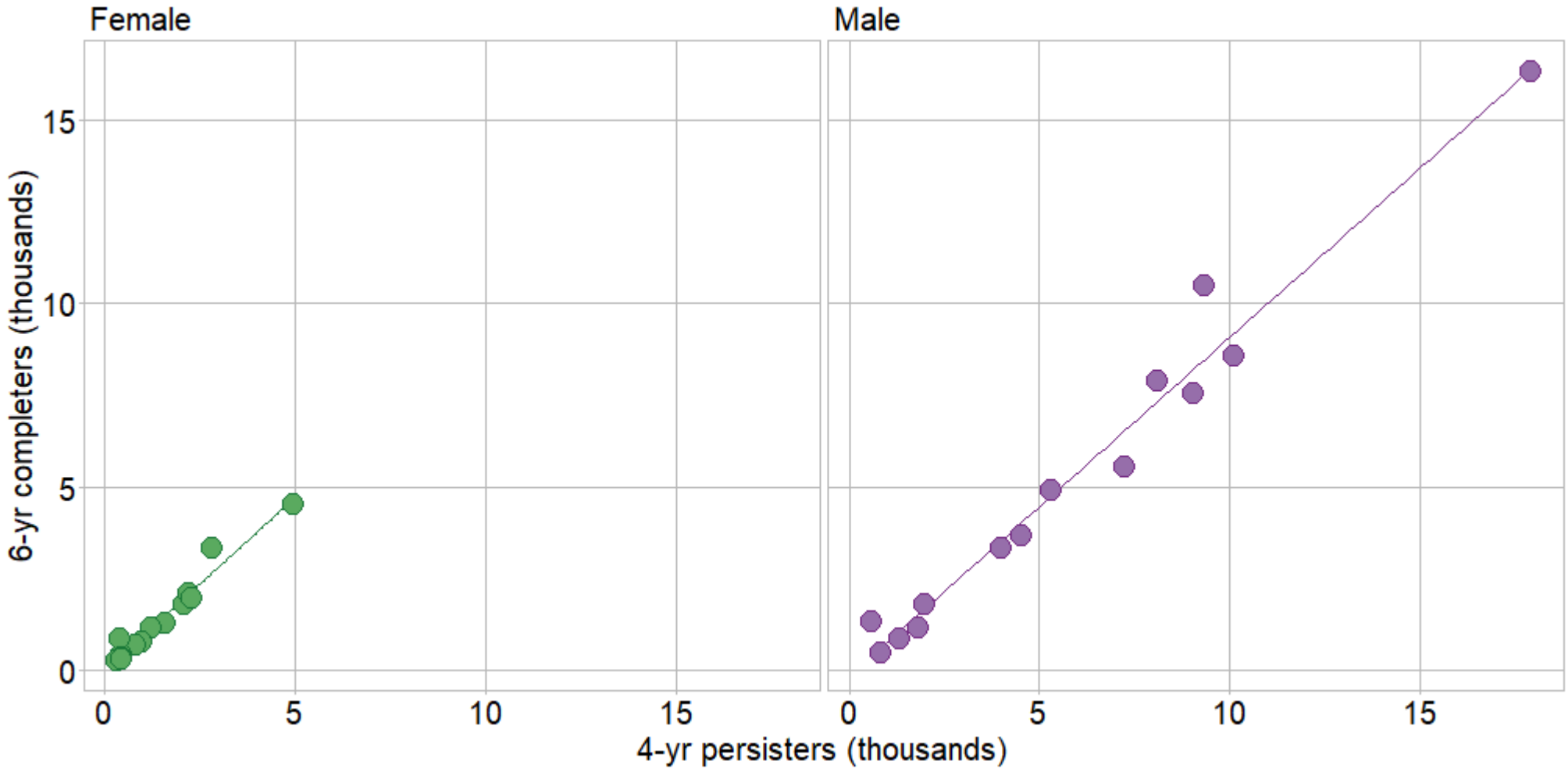
# Scatterplots are designed to reveal correlation



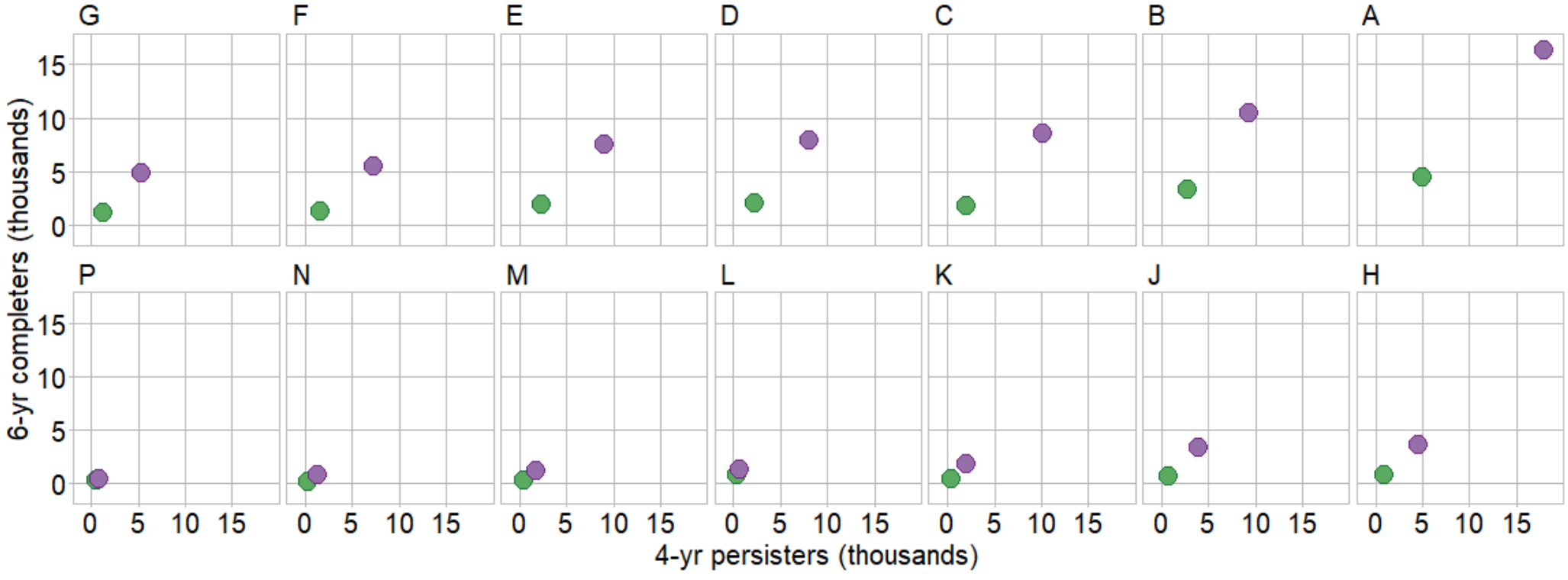
# Add a category



# One facet per sex

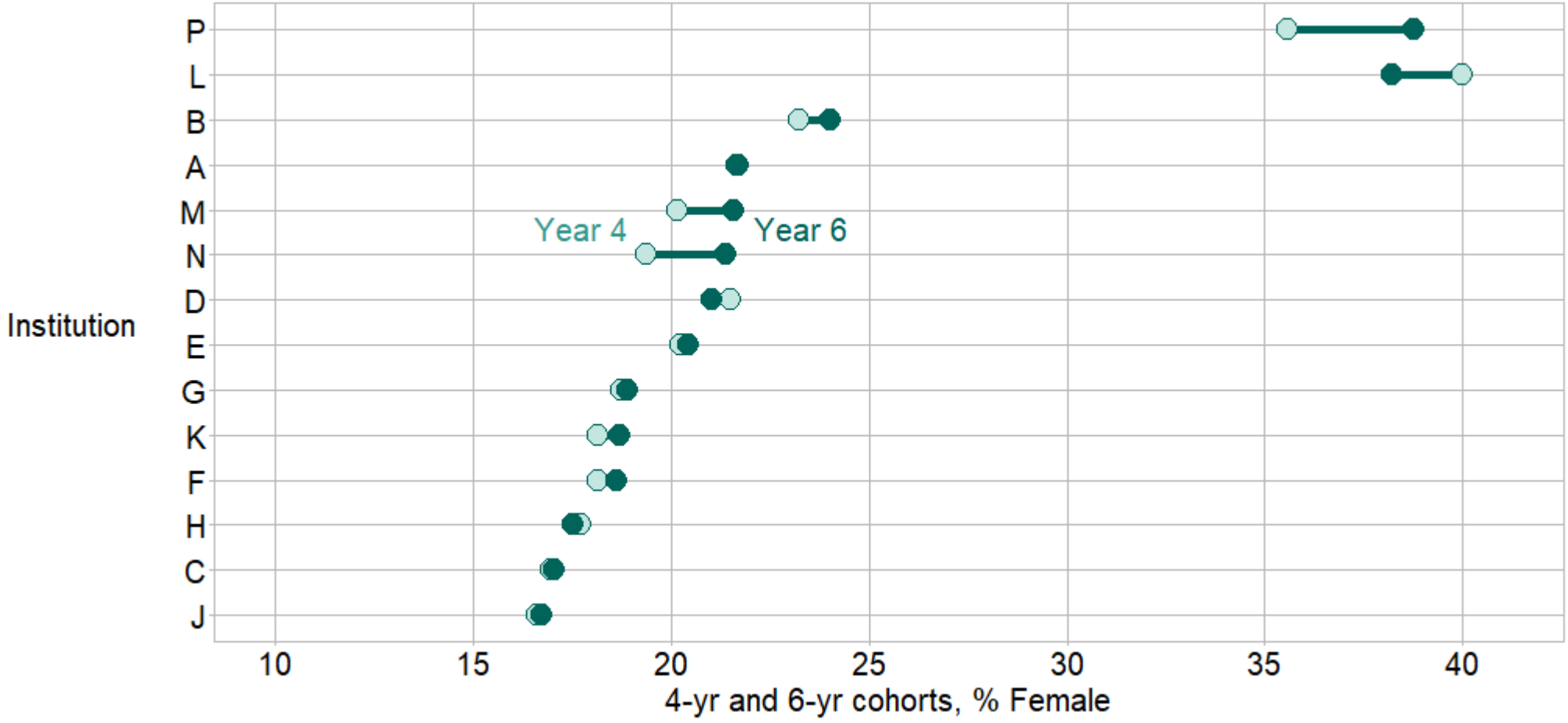


# One facet per institution

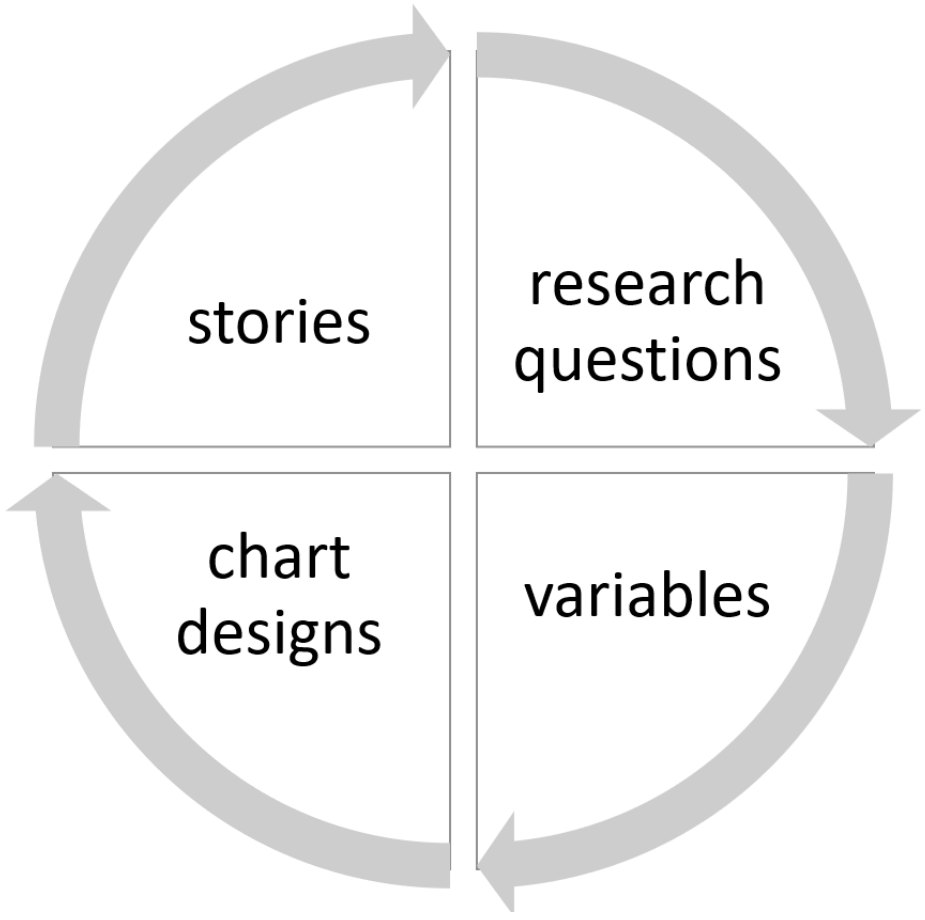


# Change the quantitative variable

Engineering students at 14 institutions persisting to year 4 and graduating by year 6, 1987–2019



# Discussion




## Revealing correlations

We saw a correlation.

We changed the emphasis.

Which chart tells a more compelling story?

# Showing evolution

# Data

University of California: funding and percent White enrollment, 1999–2017

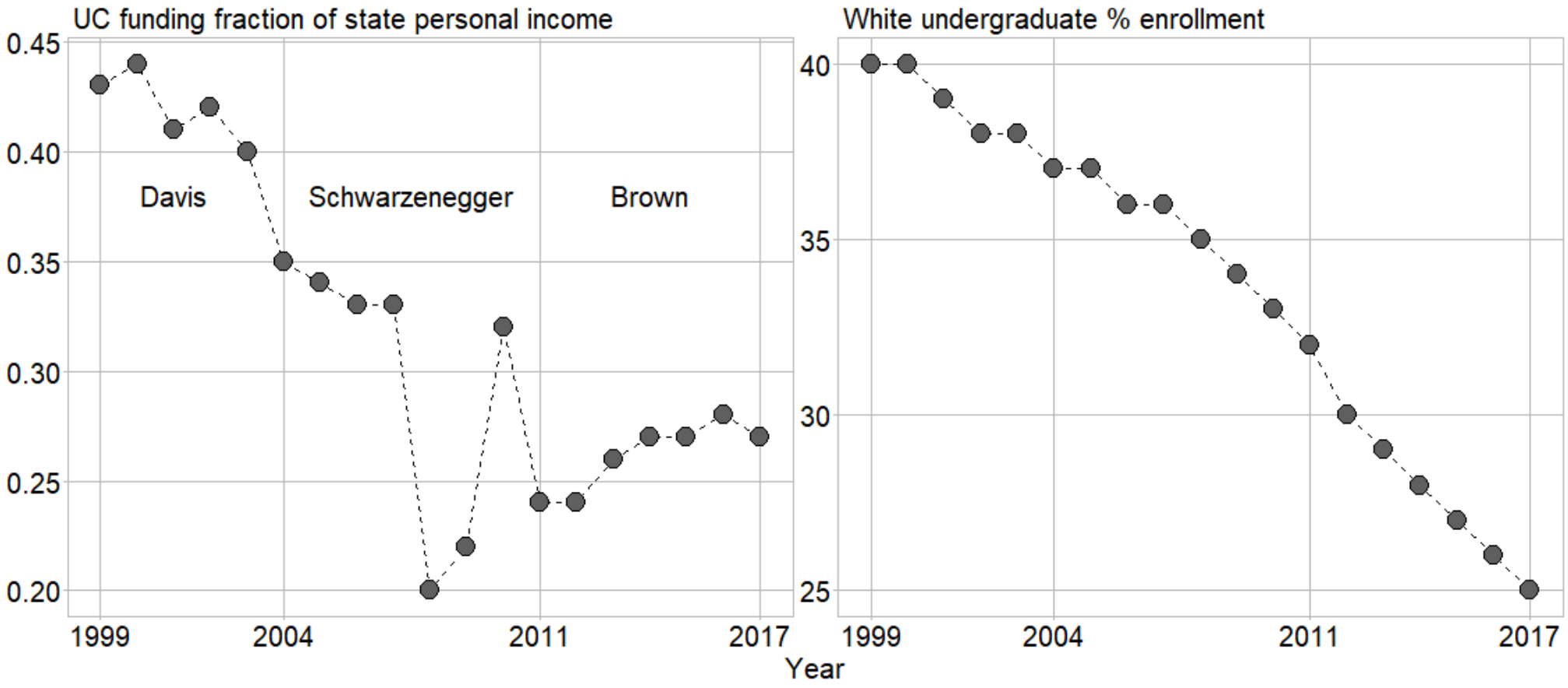
| | year | gov | white_pct | fund_frac |
|---|---|---|---|---|
| | <num> | <char> | <num> | <num> |
| 1: | 1999 | Davis | 40 | 0.43 |
| 2: | 2000 | Davis | 40 | 0.44 |
| 3: | 2001 | Davis | 39 | 0.41 |
| 4: | 2002 | Davis | 38 | 0.42 |
| 5: | 2003 | Davis | 38 | 0.40 |
| 6: | 2004 | Schwarzenegger | 37 | 0.35 |
| 7: | 2005 | Schwarzenegger | 37 | 0.34 |
| 8: | 2006 | Schwarzenegger | 36 | 0.33 |
| 9: | 2007 | Schwarzenegger | 36 | 0.33 |
| 10: | 2008 | Schwarzenegger | 35 | 0.20 |
| 11: | 2009 | Schwarzenegger | 34 | 0.22 |
| 12: | 2010 | Schwarzenegger | 33 | 0.32 |
| 13: | 2011 | Brown | 32 | 0.24 |
| 14: | 2012 | Brown | 30 | 0.24 |
| 15: | 2013 | Brown | 29 | 0.26 |
| 16: | 2014 | Brown | 28 | 0.27 |
| 17: | 2015 | Brown | 27 | 0.27 |
| 18: | 2016 | Brown | 26 | 0.28 |
| 19: | 2017 | Brown | 25 | 0.27 |

| variable | type |
|---|---|
| year | categorical |
| governor | categorical |
| UC funding metric | quantitative |
| White undergraduate % enrollment | quantitative |

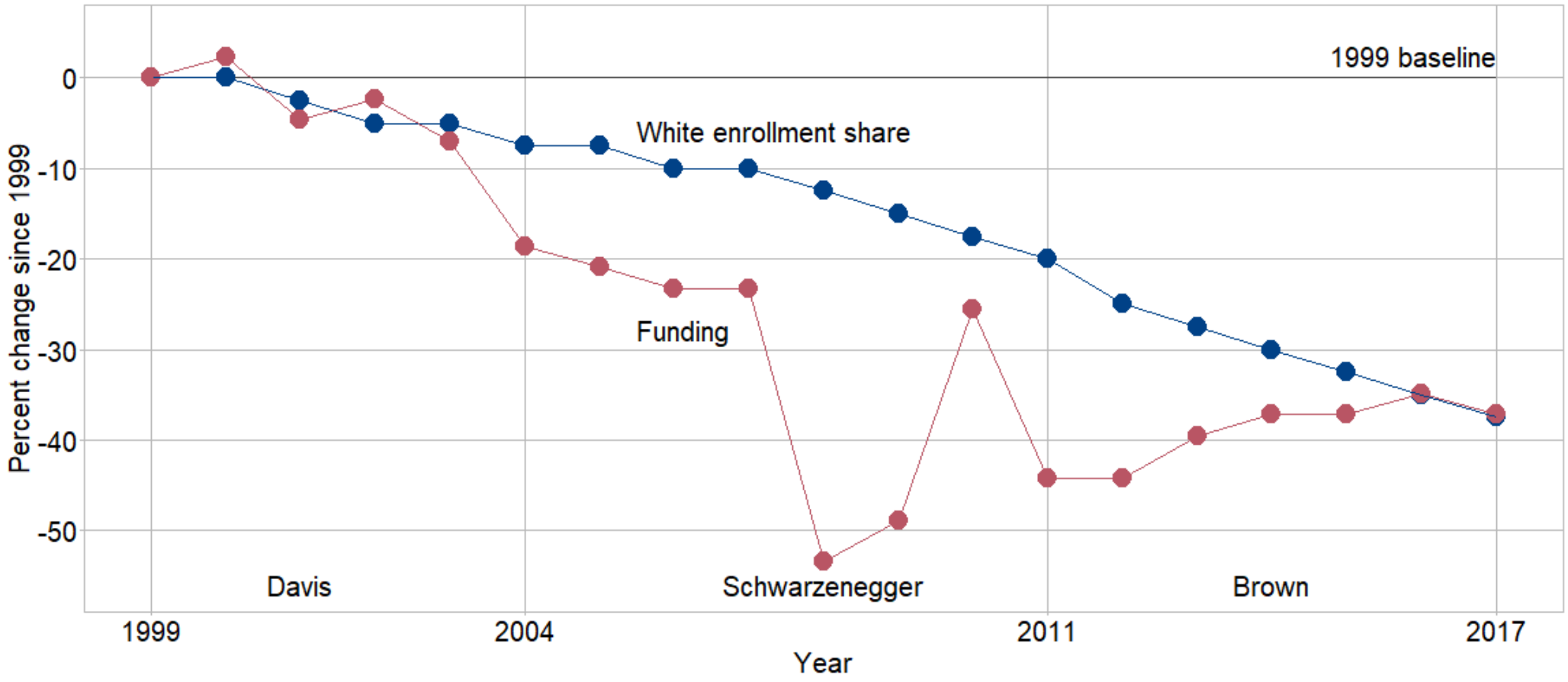


Data source: C. Newfield (2021) Budget justice: Addressing the structural racism of higher education funding, *Academe* 107(2), 57–64.

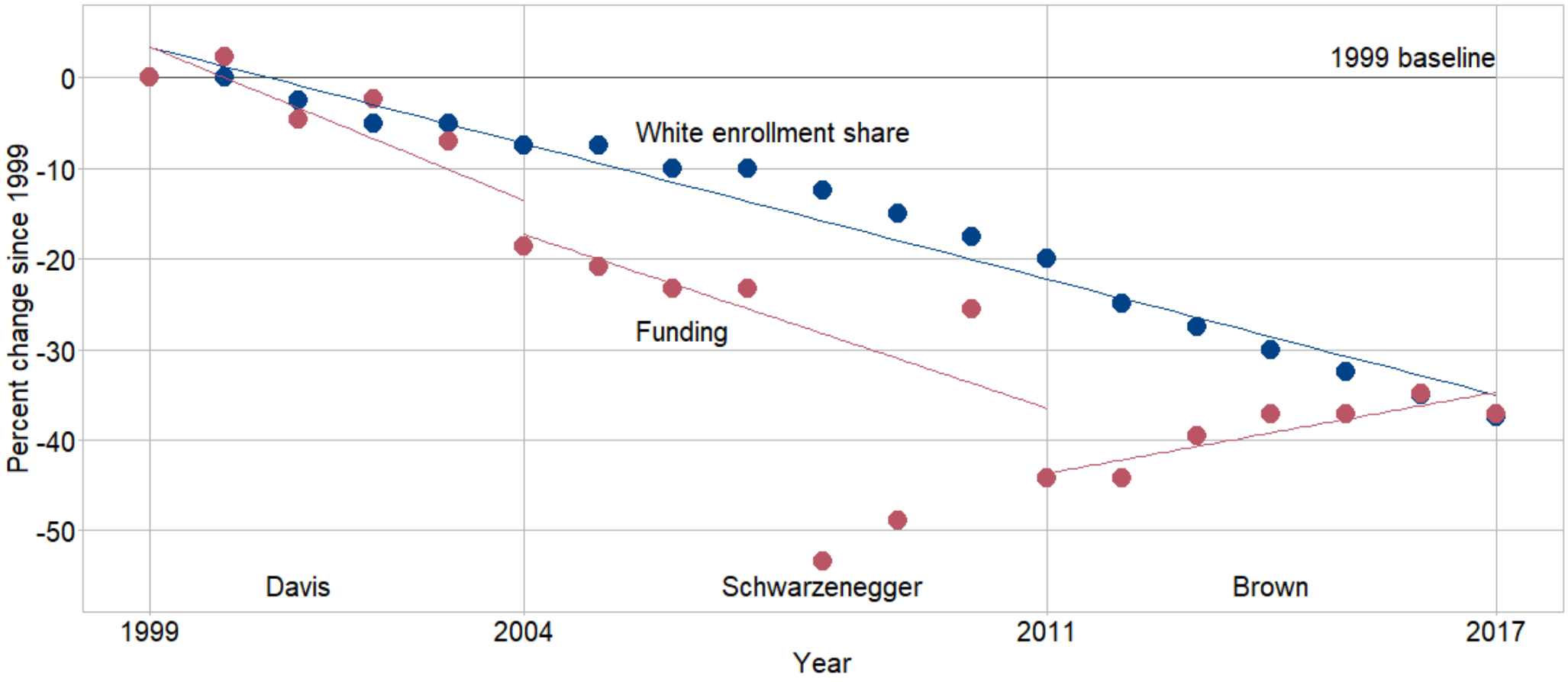
# Two time series



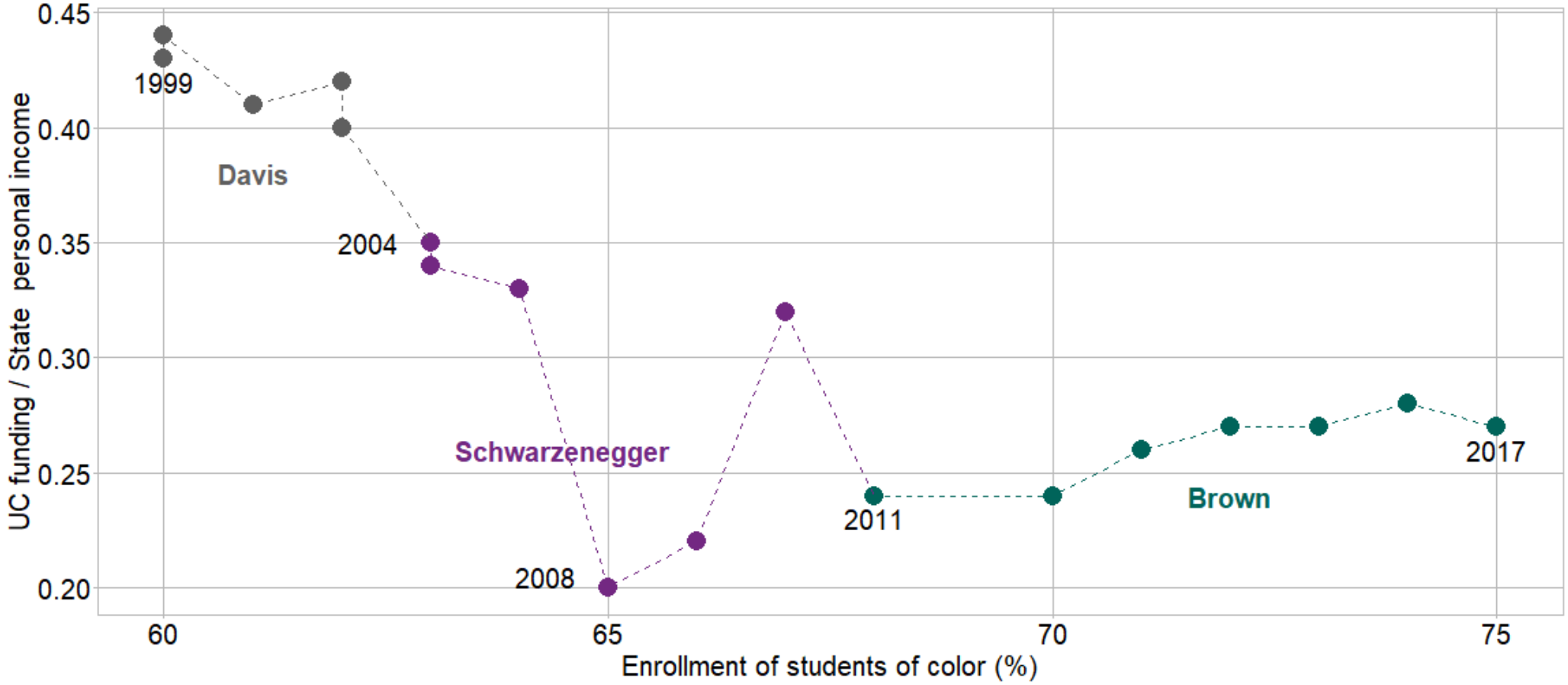
# Indexed time series



# Parallel lines indicate possible correlation



# Connected scatterplot



# Data

Extent of polar ice, 1979–2021

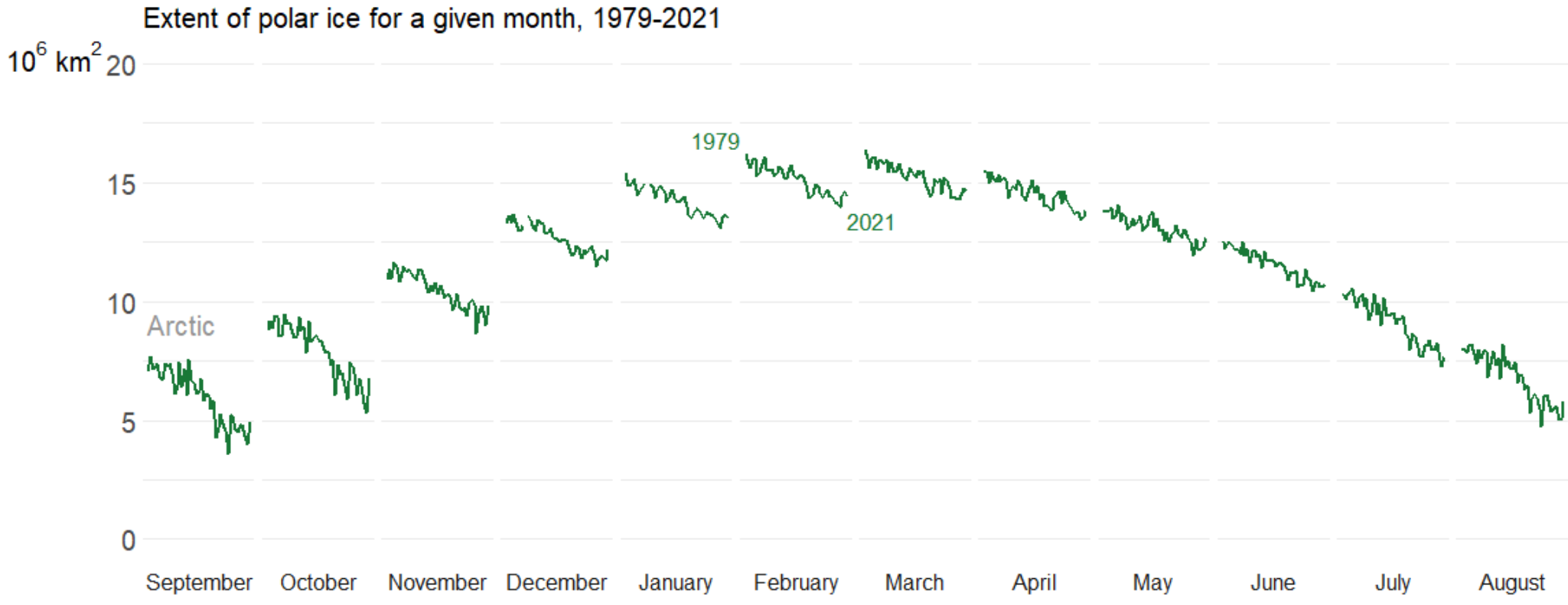
```
         hemis     month  year extent
        <char>    <fctr> <int>  <num>
   1:    Arctic September  1979  7.051
   2:    Arctic September  1980  7.667
   3:    Arctic September  1981  7.138
   4:    Arctic September  1982  7.302
   5:    Arctic September  1983  7.395
   6:    Arctic September  1984  6.805
   7:    Arctic September  1985  6.698
   8:    Arctic September  1986  7.411
   9:    Arctic September  1987  7.279
  10:    Arctic September  1988  7.369
  ---
1023: Antarctic    August  2012 18.097
1024: Antarctic    August  2013 18.664
1025: Antarctic    August  2014 18.908
1026: Antarctic    August  2015 17.749
1027: Antarctic    August  2016 17.892
1028: Antarctic    August  2017 17.219
1029: Antarctic    August  2018 17.417
1030: Antarctic    August  2019 17.478
1031: Antarctic    August  2020 17.758
1032: Antarctic    August  2021 18.131
```

| variable | type |
|---|---|
| hemisphere | categorical |
| month | categorical |
| year | categorical |
| area of polar ice (millions sq km) | quantitative |

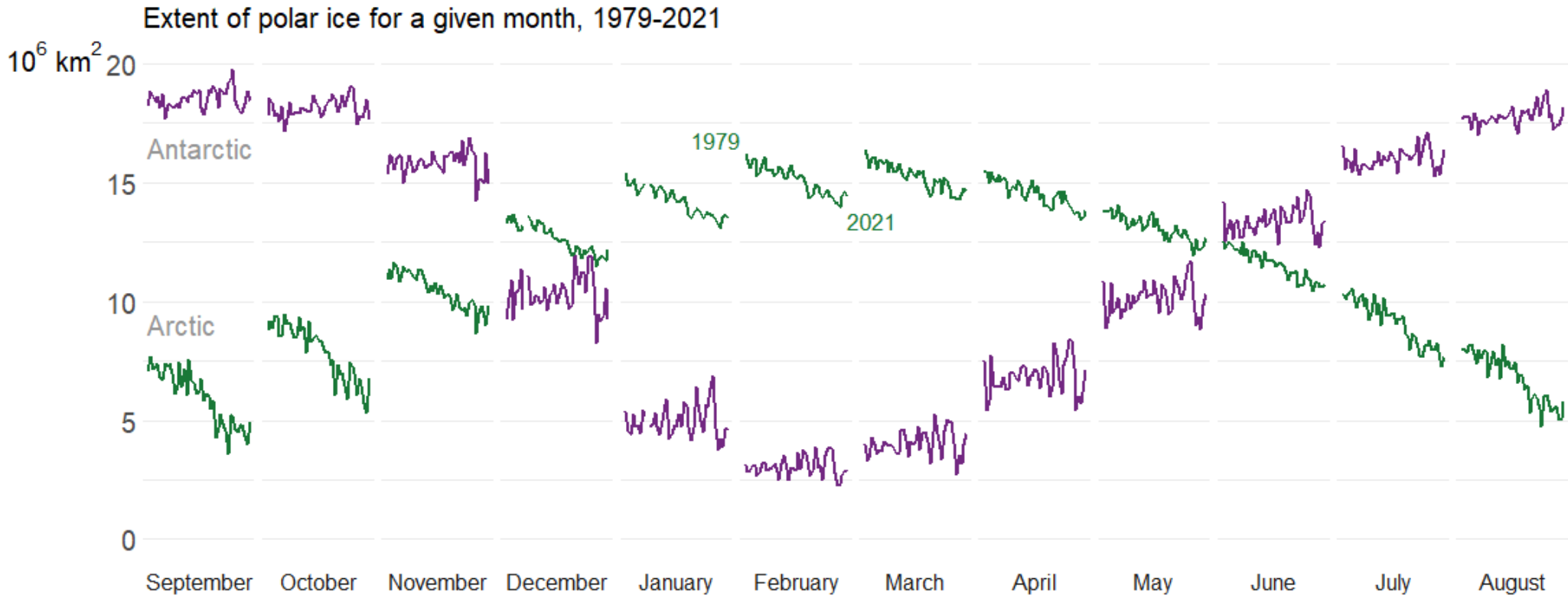


Data source: NASA National Snow and Ice Data Center Sea ice extent and area organized by year

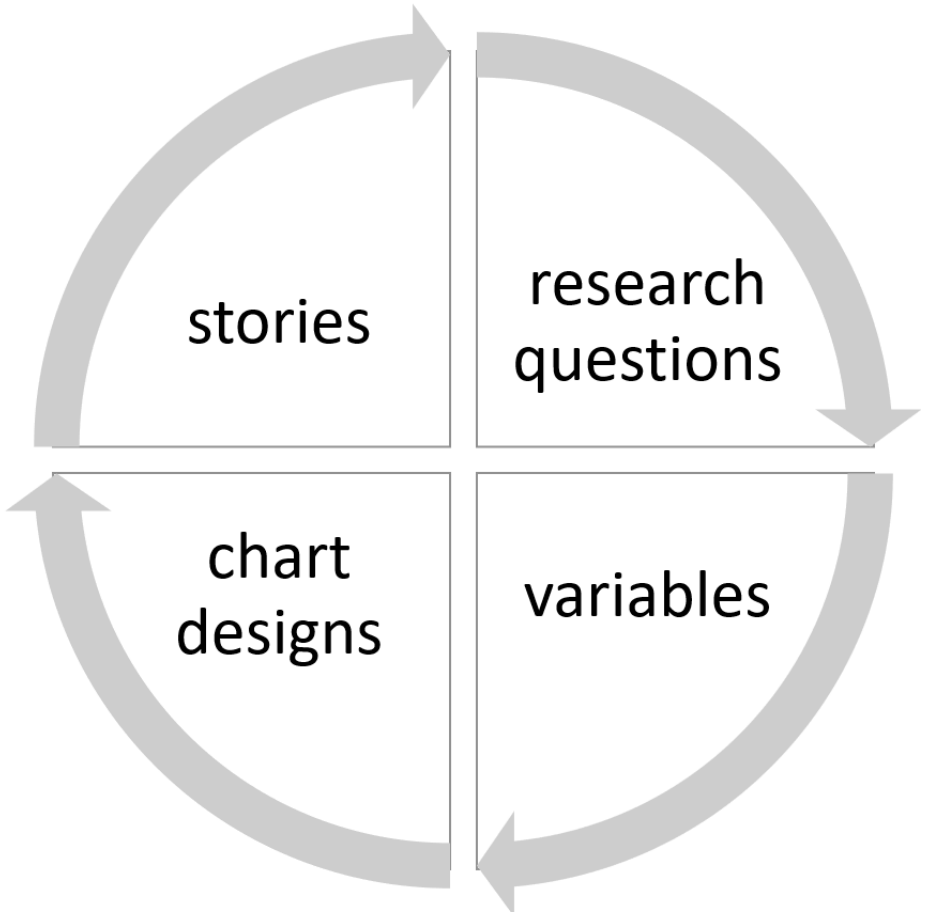
# Cyclic time series



# Add a category



# Discussion




## Showing evolution

Which time series chart design might be used in your own work? Explain.

# Displaying distributions

# Data

World speed skiing competitions, 1953–1995

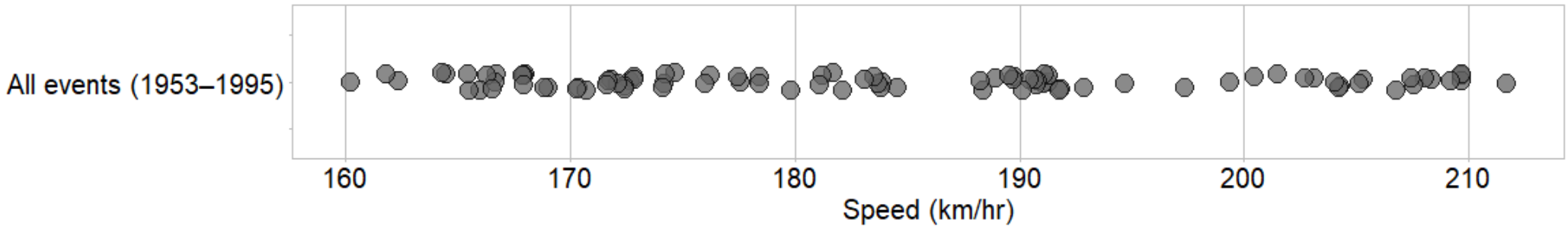
```
Key: <Event, Year, Sex>
             Event  Year    Sex  Speed
            <fctr> <int> <fctr>  <num>
 1: Speed Downhill  1952   Male 167.85
 2: Speed Downhill  1953   Male 168.86
 3: Speed Downhill  1961   Male 165.42
 4: Speed Downhill  1962   Male 172.85
 5: Speed Downhill  1965   Male 189.77
 6: Speed Downhill  1965   Male 172.44
 7: Speed Downhill  1966   Male 176.01
 8: Speed Downhill  1967   Male 188.29
 9: Speed Downhill  1967   Male 172.15
10: Speed Downhill  1969   Male 192.86
---
82:      Speed One  1982   Male 206.80
83:      Speed One  1982   Male 191.29
84:      Speed One  1985 Female 202.70
85:      Speed One  1985   Male 209.69
86:      Speed One  1987   Male 209.70
87:      Speed One  1990 Female 201.51
88:      Speed One  1990 Female 199.35
89:      Speed One  1991   Male 207.59
90:      Speed One  1993   Male 208.33
91:      Speed One  1993   Male 170.30
```

| variable | type |
|---|---|
| event | categorical |
| year | categorical |
| sex | categorical |
| speed (km/hr) | quantitative |



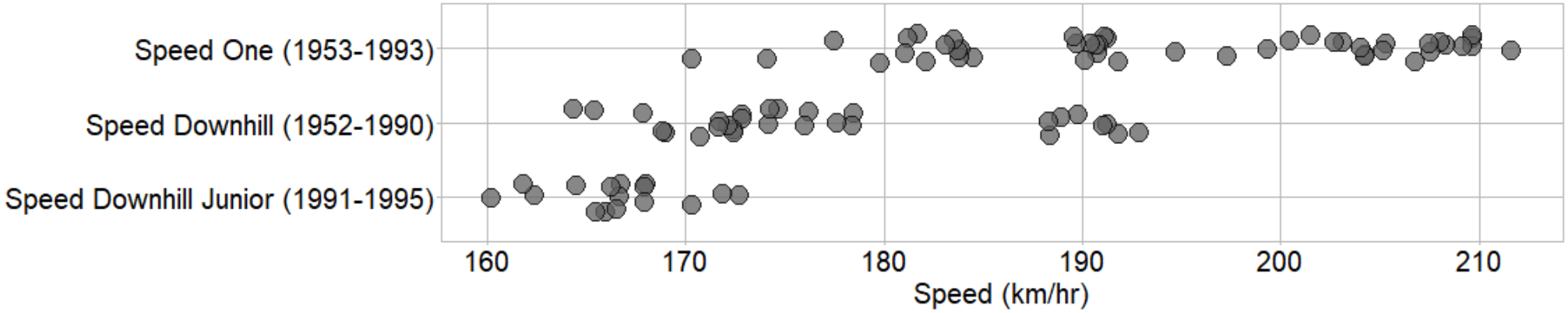
Data source: 2011 World speed skiing competition, in the GDAdata R package

# Strip chart



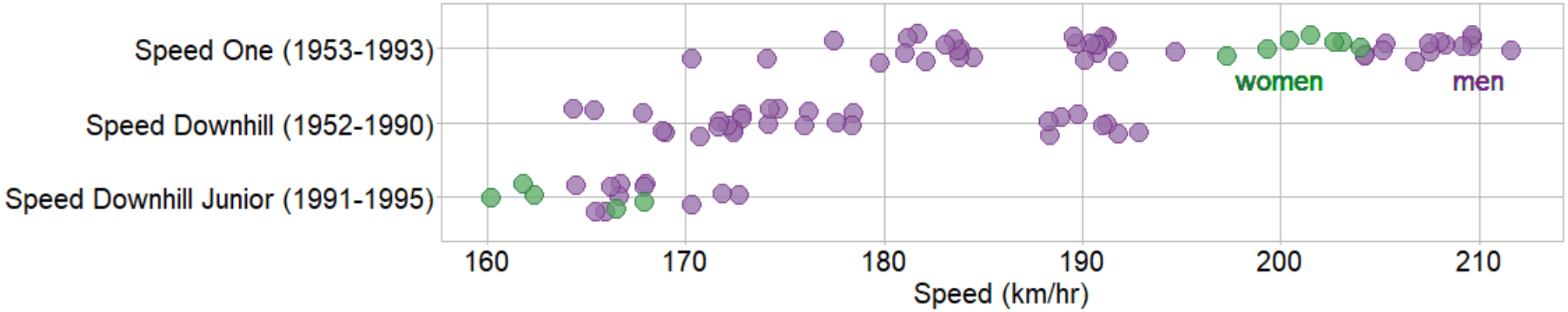

| variable | type |
|---|---|
| speed | quantitative |

# Add a category




| variable | type |
|---|---|
| event | categorical |
| speed | quantitative |

# Add a second category




| variable | type |
|---|---|
| event | categorical |
| sex | categorical |
| speed | quantitative |

# Data

MIDFIELD graduates (N = 270k), enrolled in Engineering, excluding 10th and 90th quantiles

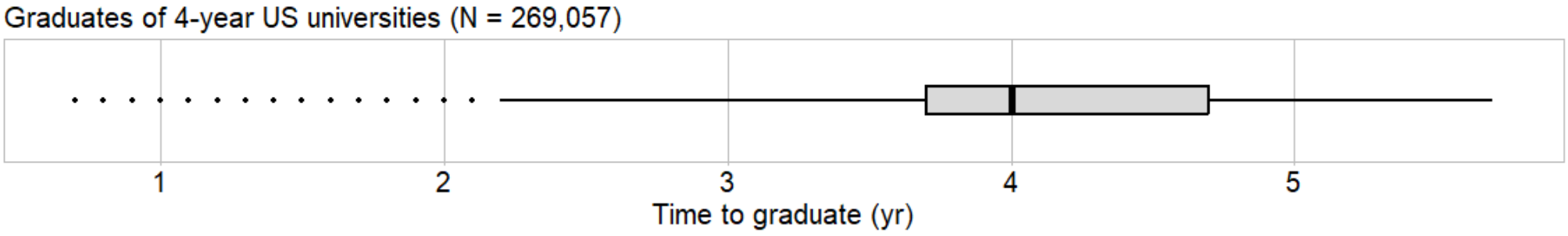
```
Key: <path, sex>
                path    sex years_to_grad
              <char> <char>         <num>
     1: Nontraditional Female           3.9
     2: Nontraditional Female           1.9
     3: Nontraditional Female           3.9
     4: Nontraditional Female           5.3
     5: Nontraditional Female           5.1
     6: Nontraditional Female           3.8
     7: Nontraditional Female           2.7
     8: Nontraditional Female           1.9
     9: Nontraditional Female           2.8
    10: Nontraditional Female           3.9
    ---
269048:    Traditional   Male           5.7
269049:    Traditional   Male           1.7
269050:    Traditional   Male           3.7
269051:    Traditional   Male           4.7
269052:    Traditional   Male           5.7
269053:    Traditional   Male           2.6
269054:    Traditional   Male           1.3
269055:    Traditional   Male           3.0
269056:    Traditional   Male           5.3
269057:    Traditional   Male           0.7
```

| variable | type |
|---|---|
| path | categorical |
| sex | categorical |
| years to graduate | quantitative |



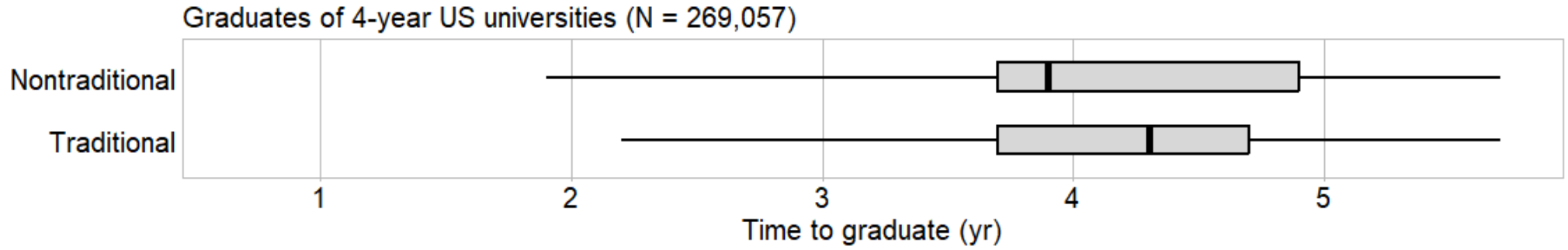
Data source: MIDFIELD subset from the graphclassmate R package

# Box and whisker chart



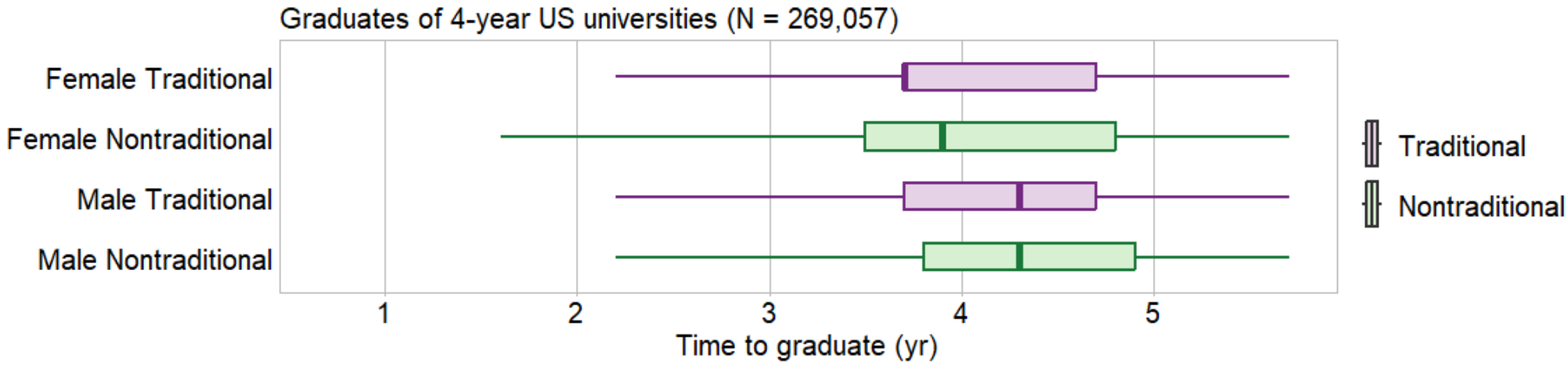

| variable | type |
|---|---|
| years to graduate | quantitative |

# Add a category



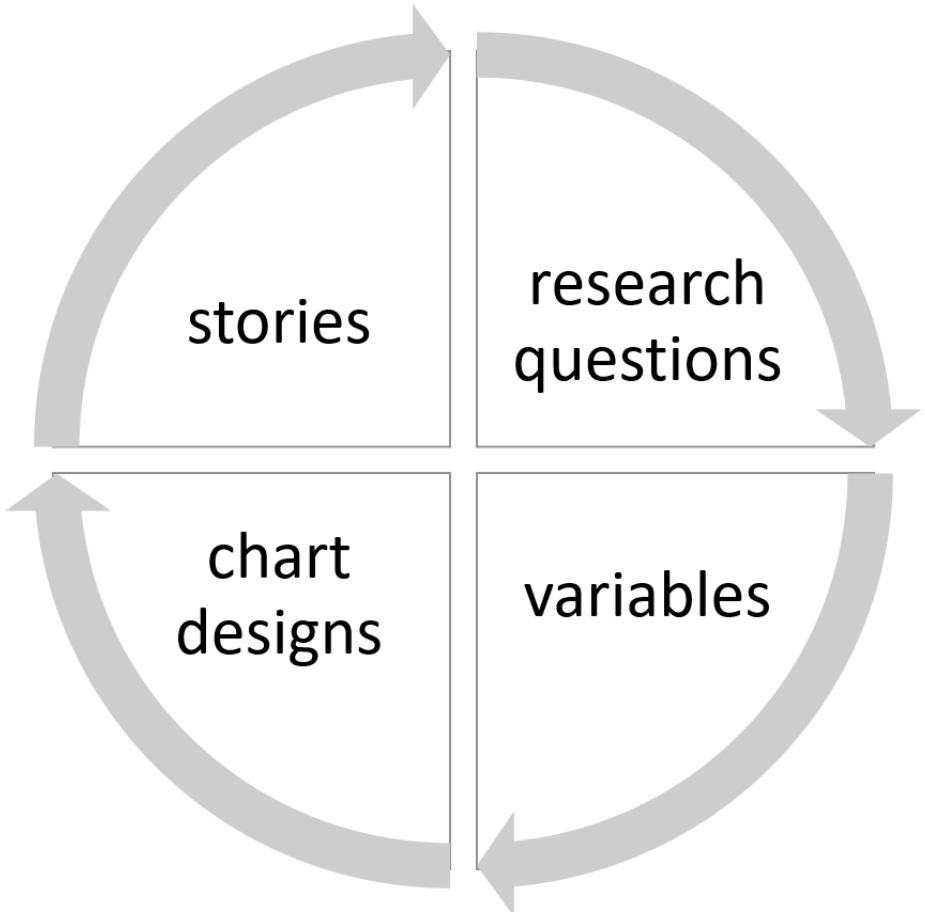

| variable | type |
|---|---|
| path | categorical |
| years to graduate | quantitative |

# Combine a second category




| variable | type |
|---|---|
| sex and path | categorical |
| years to graduate | quantitative |

# Discussion




## Displaying distributions

What MIDFIELD distributions would you like to see:

- what quantitative variable?
- what categorical variables?

# Closing discussion

# Variables, design, message

For you, what was the muddiest point in the session?

Is there a graph design you would have liked to have seen today?

Is there a class of variables you would have liked to have seen today?