visualizing data

🔗 bit.ly/dataviz-ncpma

dr. mine çetinkaya-rundel
duke university

why visualize

> *"The simple graph has brought more information to the data analyst's mind than any other device."*
>
> John Tukey

We visualize data to ...

- discover patterns that may not be obvious from numerical summaries

We have 13 datasets, each with 142 observations. For each observation we have values on two variables recorded: an X and a Y.

Summary statistics for these two variables for each of the datasets is given on the right.

How, if at all, are these 13 datasets different from each other?

dataset	n	Average x	Average y
Dataset 1	142	54.3	47.8
Dataset 2	142	54.3	47.8
Dataset 3	142	54.3	47.8
Dataset 4	142	54.3	47.8
Dataset 5	142	54.3	47.8
Dataset 6	142	54.3	47.8
Dataset 7	142	54.3	47.8
Dataset 8	142	54.3	47.8
Dataset 9	142	54.3	47.8
Dataset 10	142	54.3	47.8
Dataset 11	142	54.3	47.8
Dataset 12	142	54.3	47.8
Dataset 13	142	54.3	47.8

Some more summary statistics…

How, if at all, are these 13 datasets different from each other?

dataset	n	Average x	Average y	St Dev x	St Dev y
Dataset 1	142	54.3	47.8	16.8	26.9
Dataset 2	142	54.3	47.8	16.8	26.9
Dataset 3	142	54.3	47.8	16.8	26.9
Dataset 4	142	54.3	47.8	16.8	26.9
Dataset 5	142	54.3	47.8	16.8	26.9
Dataset 6	142	54.3	47.8	16.8	26.9
Dataset 7	142	54.3	47.8	16.8	26.9
Dataset 8	142	54.3	47.8	16.8	26.9
Dataset 9	142	54.3	47.8	16.8	26.9
Dataset 10	142	54.3	47.8	16.8	26.9
Dataset 11	142	54.3	47.8	16.8	26.9
Dataset 12	142	54.3	47.8	16.8	26.9
Dataset 13	142	54.3	47.8	16.8	26.9

And some more summary statistics...

How, if at all, are these 13 datasets different from each other?

dataset	n	Average x	Average y	St Dev x	St Dev y	Correlation
Dataset 1	142	54.3	47.8	16.8	26.9	-0.1
Dataset 2	142	54.3	47.8	16.8	26.9	-0.1
Dataset 3	142	54.3	47.8	16.8	26.9	-0.1
Dataset 4	142	54.3	47.8	16.8	26.9	-0.1
Dataset 5	142	54.3	47.8	16.8	26.9	-0.1
Dataset 6	142	54.3	47.8	16.8	26.9	-0.1
Dataset 7	142	54.3	47.8	16.8	26.9	-0.1
Dataset 8	142	54.3	47.8	16.8	26.9	-0.1
Dataset 9	142	54.3	47.8	16.8	26.9	-0.1
Dataset 10	142	54.3	47.8	16.8	26.9	-0.1
Dataset 11	142	54.3	47.8	16.8	26.9	-0.1
Dataset 12	142	54.3	47.8	16.8	26.9	-0.1
Dataset 13	142	54.3	47.8	16.8	26.9	-0.1

And finally a visualization!

How, if at all, are these 13 datasets different from each other?

We visualize data to ...

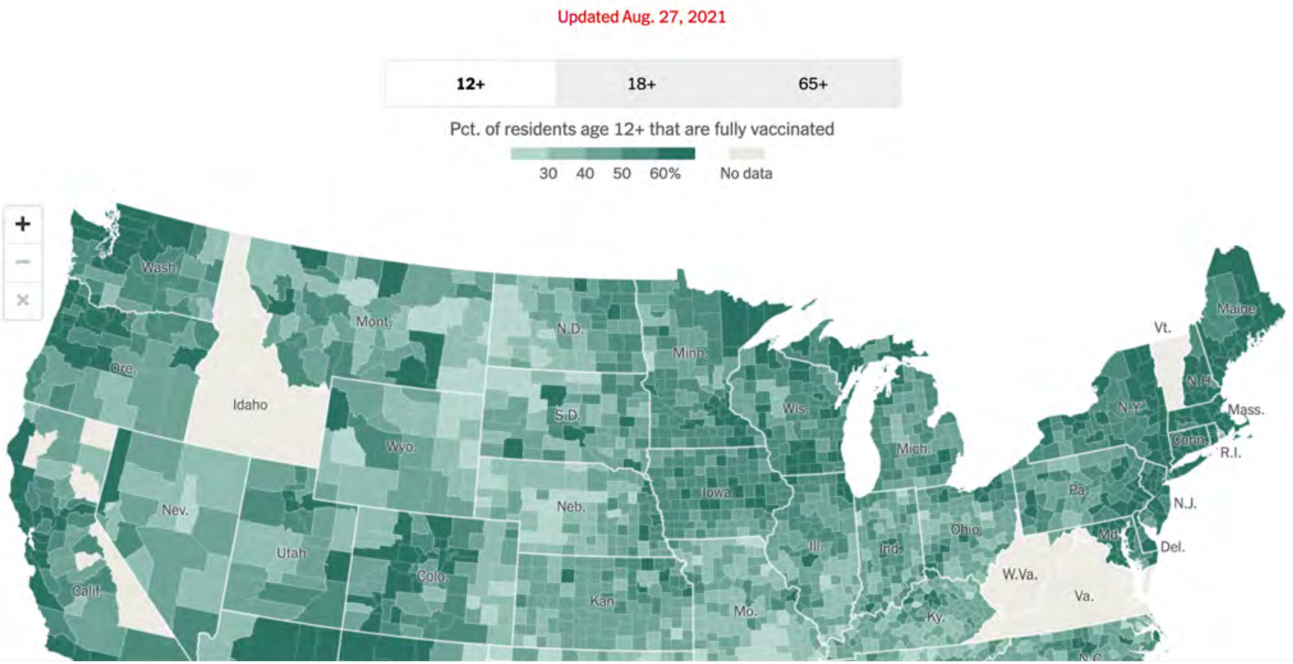
- discover patterns that may not be obvious from numerical summaries
- convey information in a way that is otherwise difficult/impossible to convey

Describe, in words, what this visualization shows.

Source: Financial Times, 27 Aug 2021.

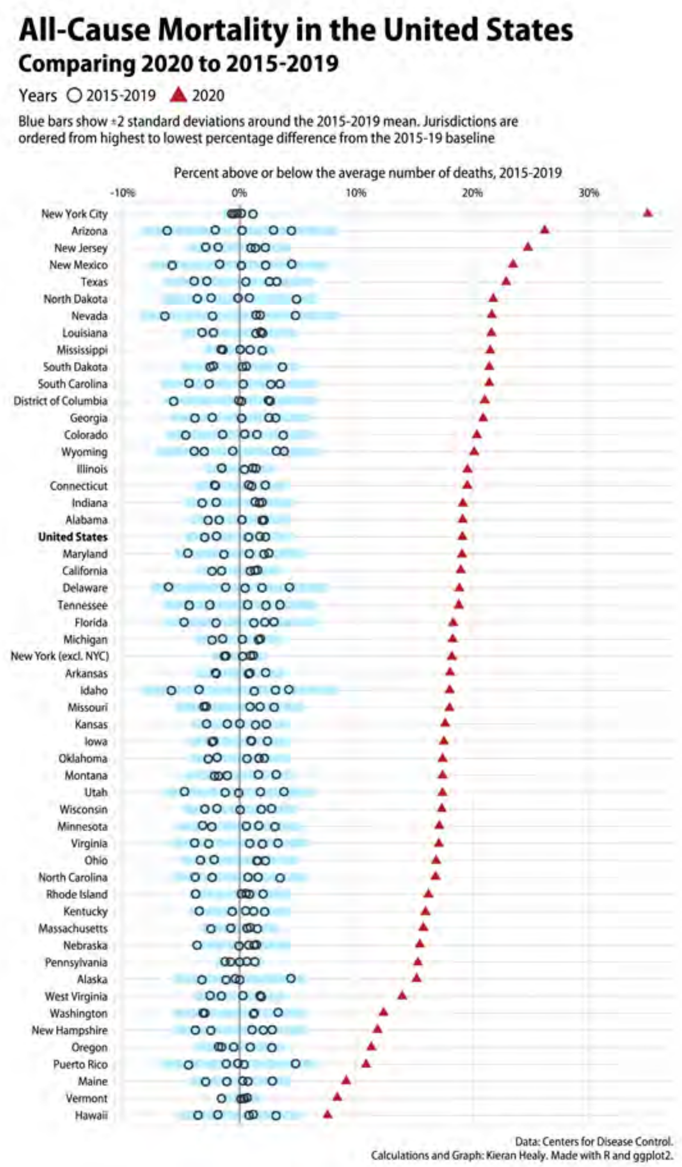
Describe, in words, what this visualization shows.

Source: New York Times, 27 Aug 2021.



Describe, in words, what this visualization shows.

Source: Kieran Healy - Excess Deaths in 2020, 21 Oct 2021.



how visualize

and how not

Case study 1:

Trends instructional staff employees in universities

The American Association of University Professors (AAUP) is a nonprofit membership association of faculty and other academic professionals. This report by the AAUP shows trends in instructional staff employees between 1975 and 2011, and contains the following image.

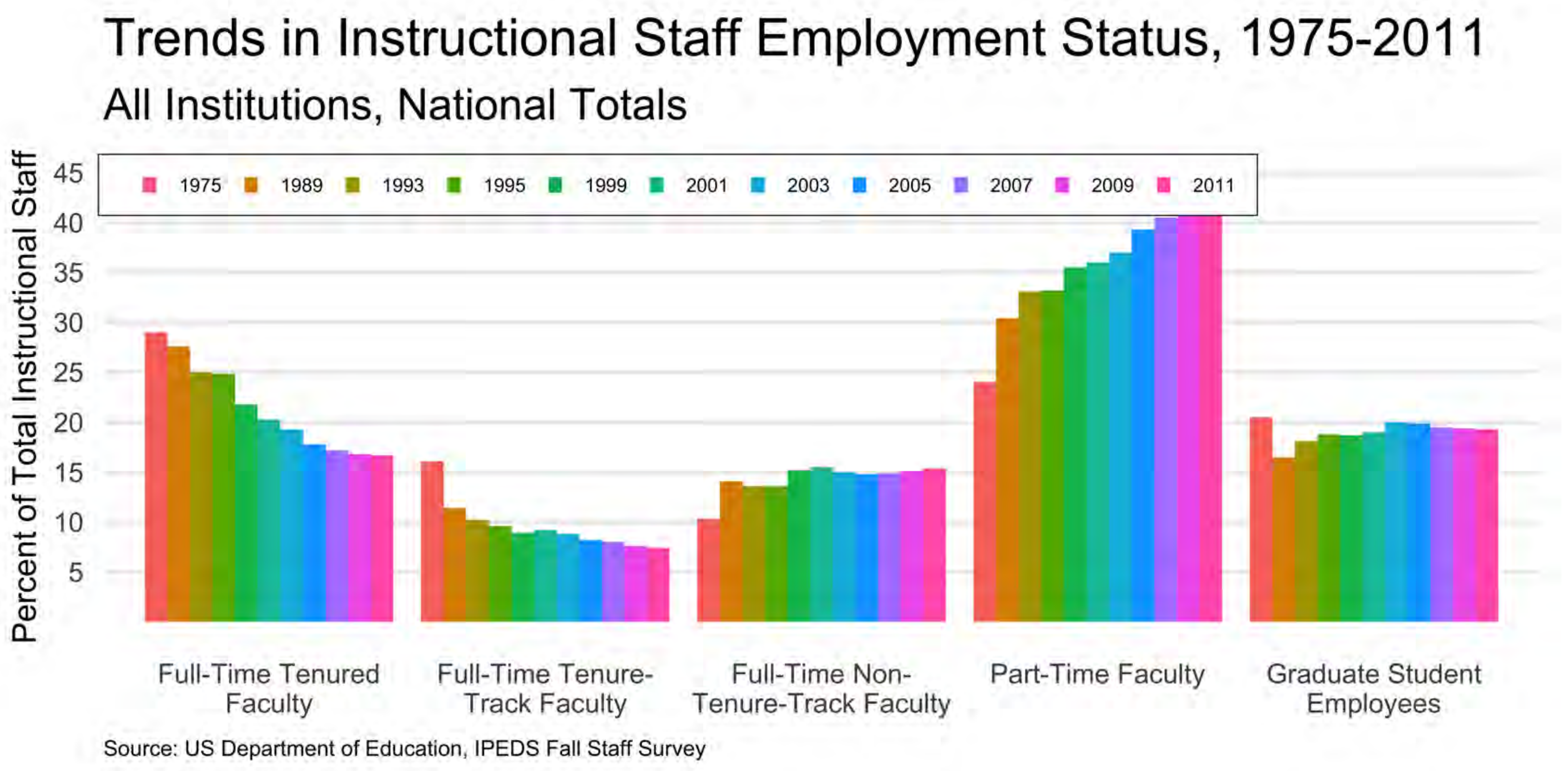
> What trends are apparent in the visualization on the right?

Data

Each row in this dataset represents a faculty type, and the columns are the years for which we have data. The values are percentage of hires of that type of faculty for each year.

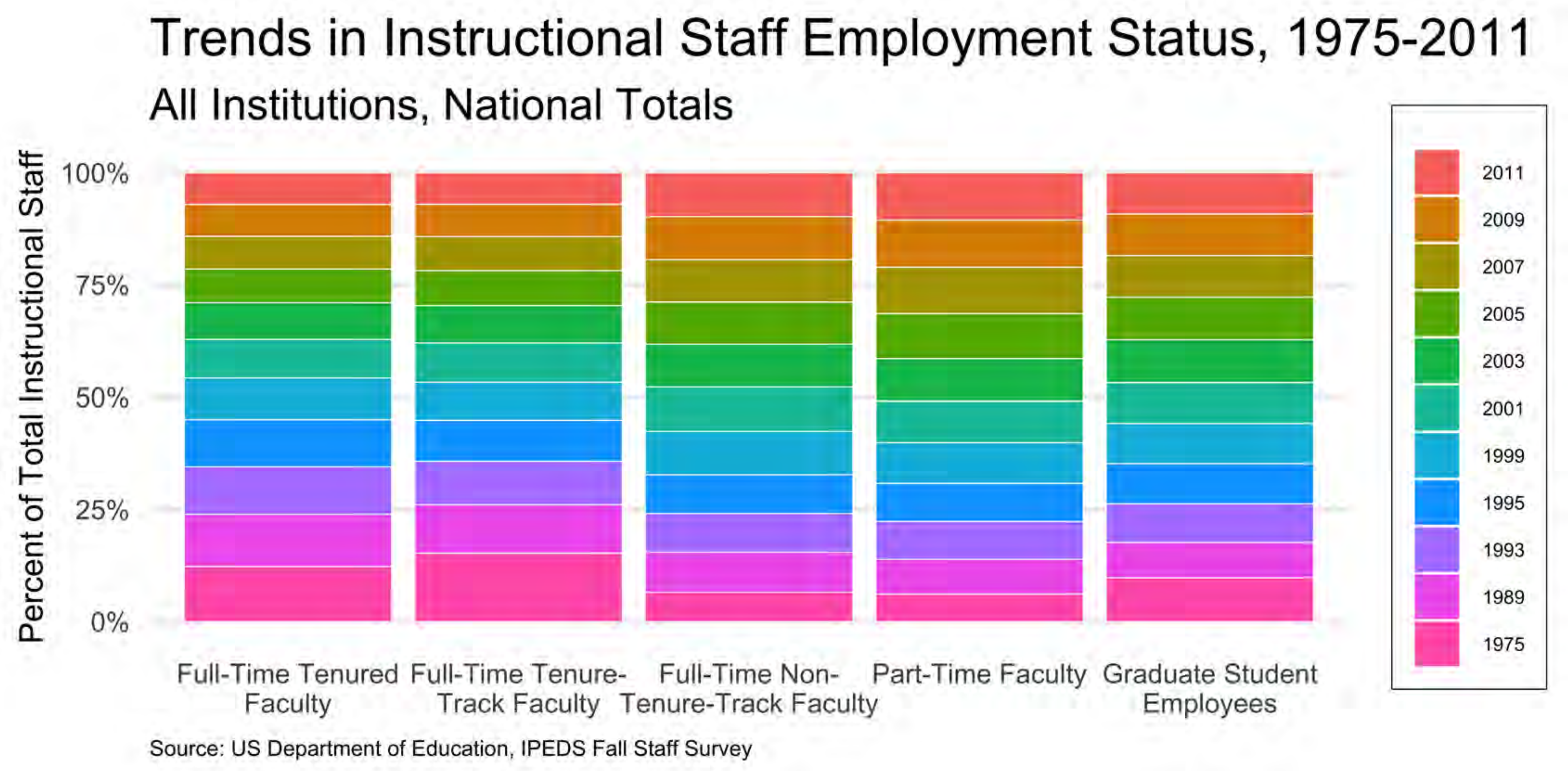
Faculty type	1975	1989	1993	1995	1999	2001	2003	2005	2007	2009	2011
Full-Time Tenured Faculty	29.0	27.6	25.0	24.8	21.8	20.3	19.3	17.8	17.2	16.8	16.7
Full-Time Tenure-Track Faculty	16.1	11.4	10.2	9.6	8.9	9.2	8.8	8.2	8.0	7.6	7.4
Full-Time Non-Tenure-Track Faculty	10.3	14.1	13.6	13.6	15.2	15.5	15.0	14.8	14.9	15.1	15.4
Part-Time Faculty	24.0	30.4	33.1	33.2	35.5	36.0	37.0	39.3	40.5	41.1	41.3
Graduate Student Employees	20.5	16.5	18.1	18.8	18.7	19.0	20.0	19.9	19.5	19.4	19.3

Recreate

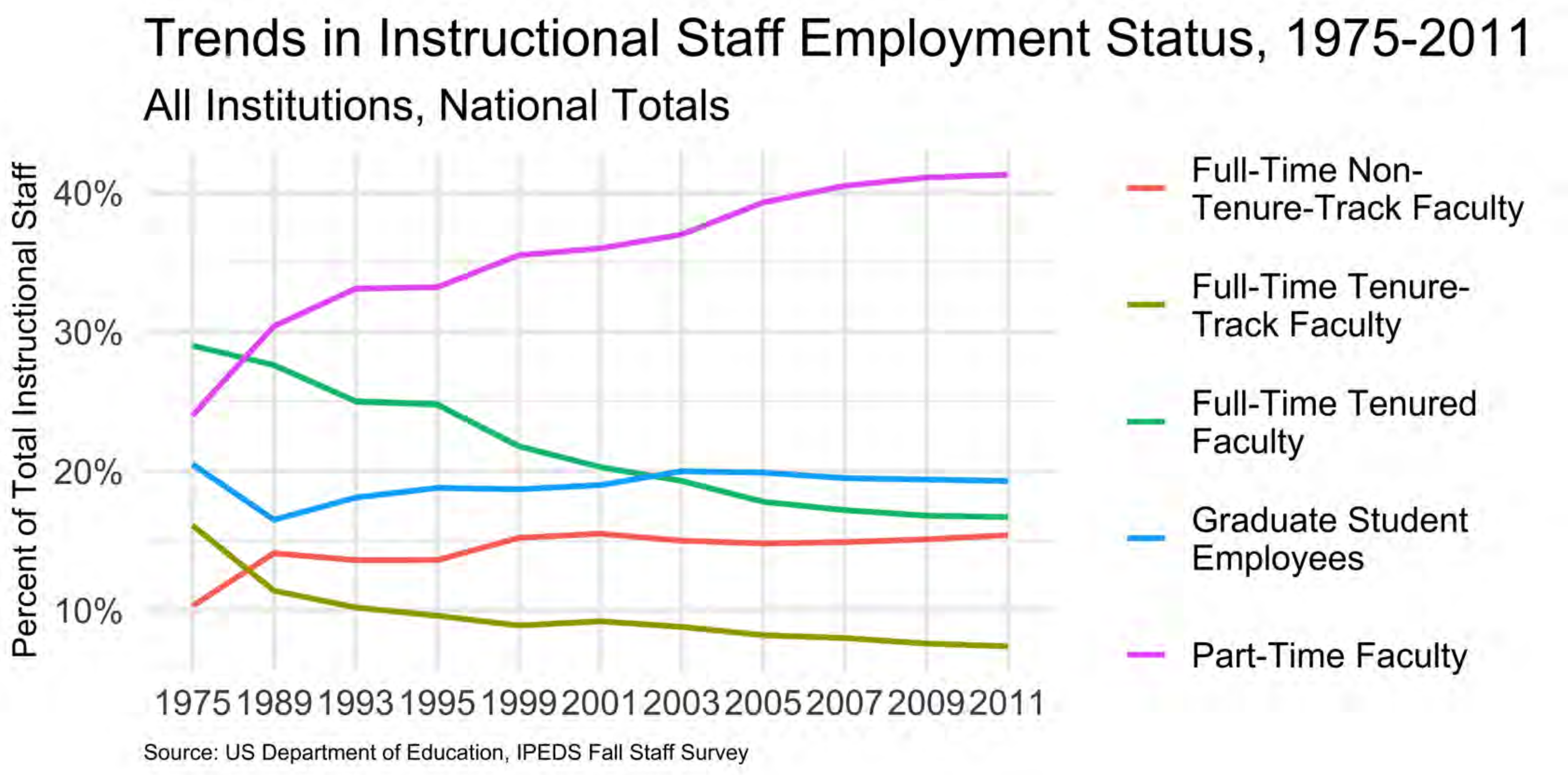


Trends in Instructional Staff Employment Status, 1975-2011
All Institutions, National Totals
Percent of Total Instructional Staff
45
40
35
30
25
20
15
10
5
1975
1989
1993
1995
1999
2001
2003
2005
2007
2009
2011
Full-Time Tenured Faculty
Full-Time Tenure-Track Faculty
Full-Time Non-Tenure-Track Faculty
Part-Time Faculty
Graduate Student Employees
Source: US Department of Education, IPEDS Fall Staff Survey

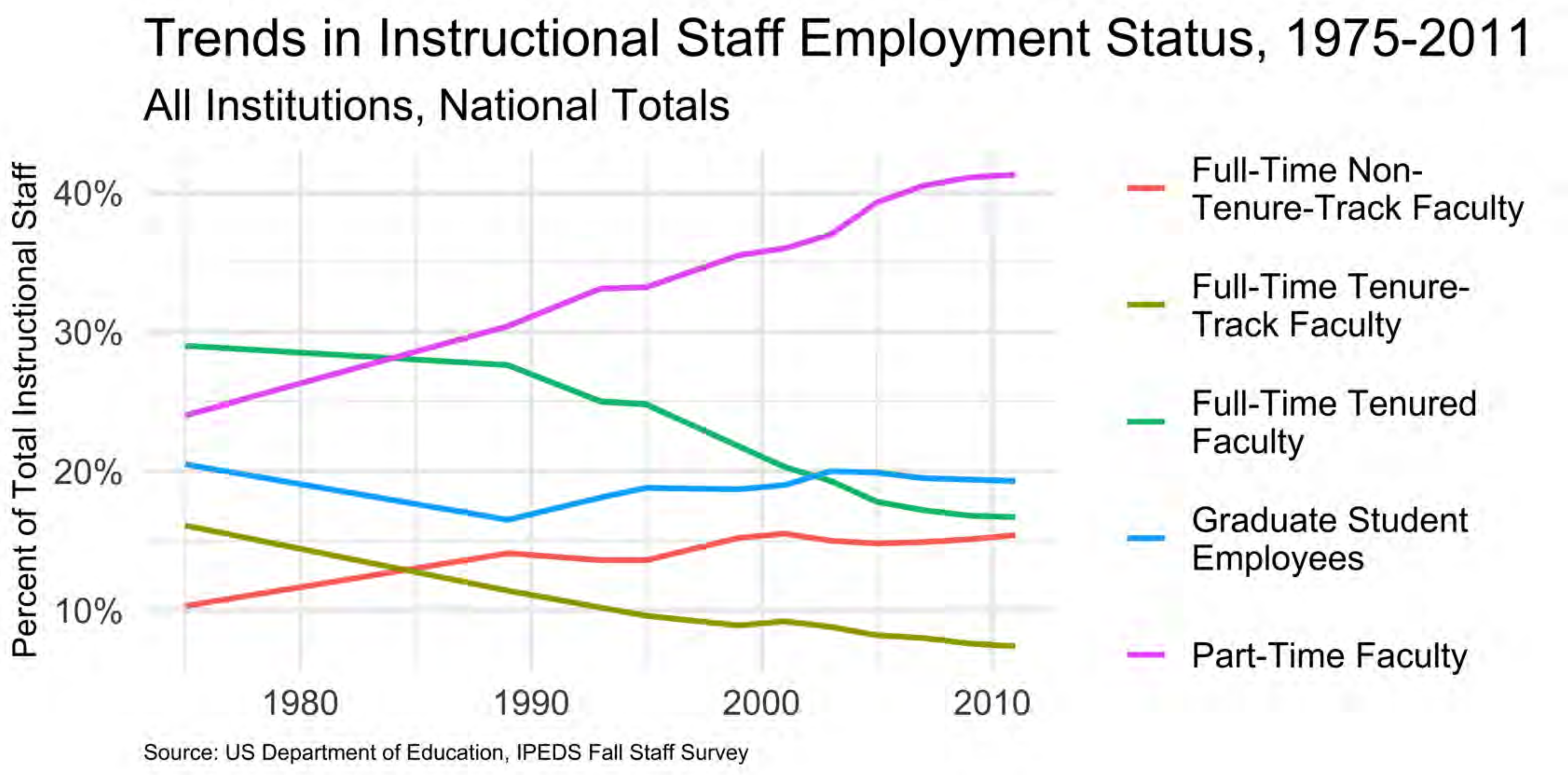
Represent percentages as parts of a whole



Place time on x-axis

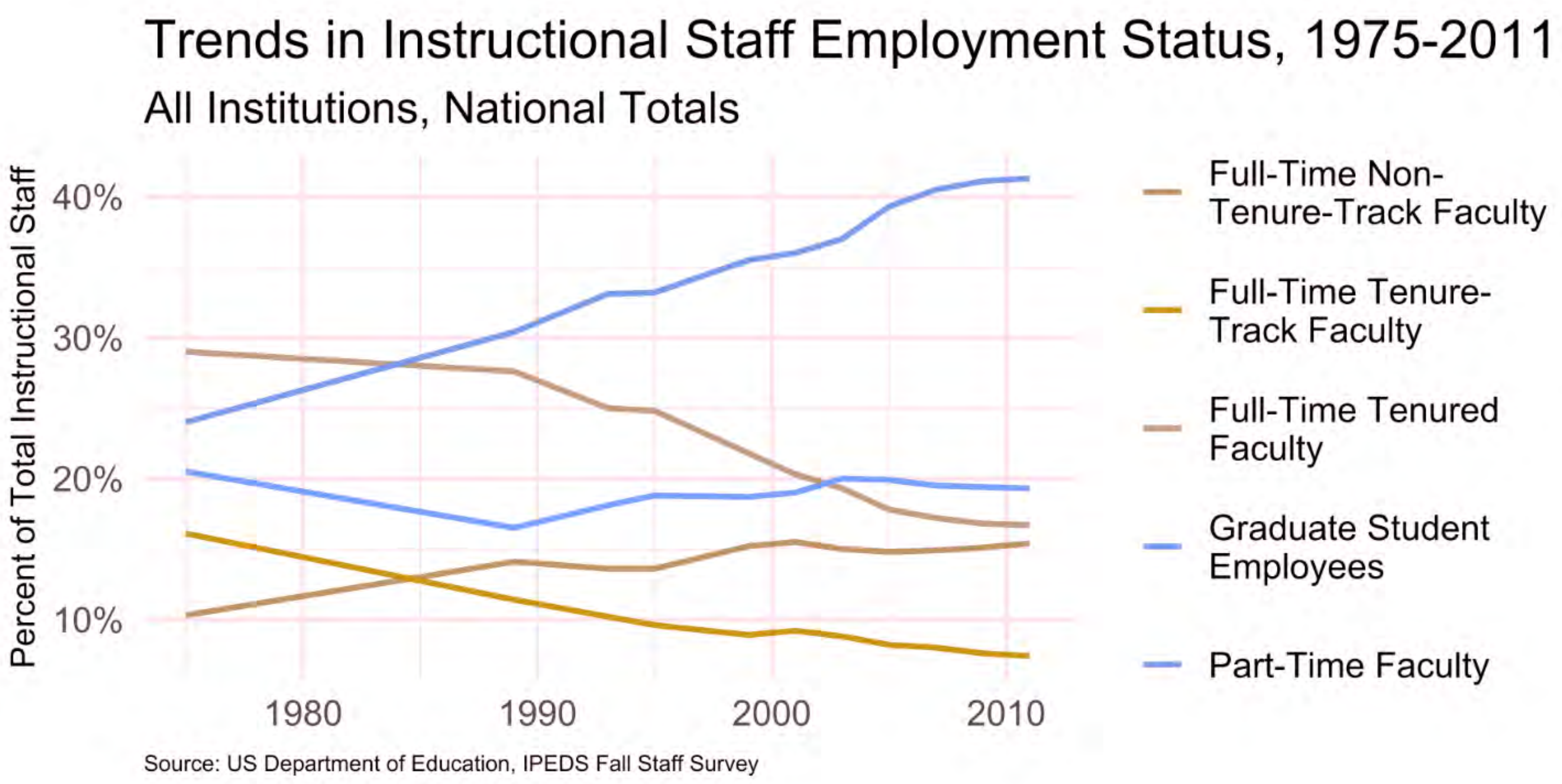


Represent time as time



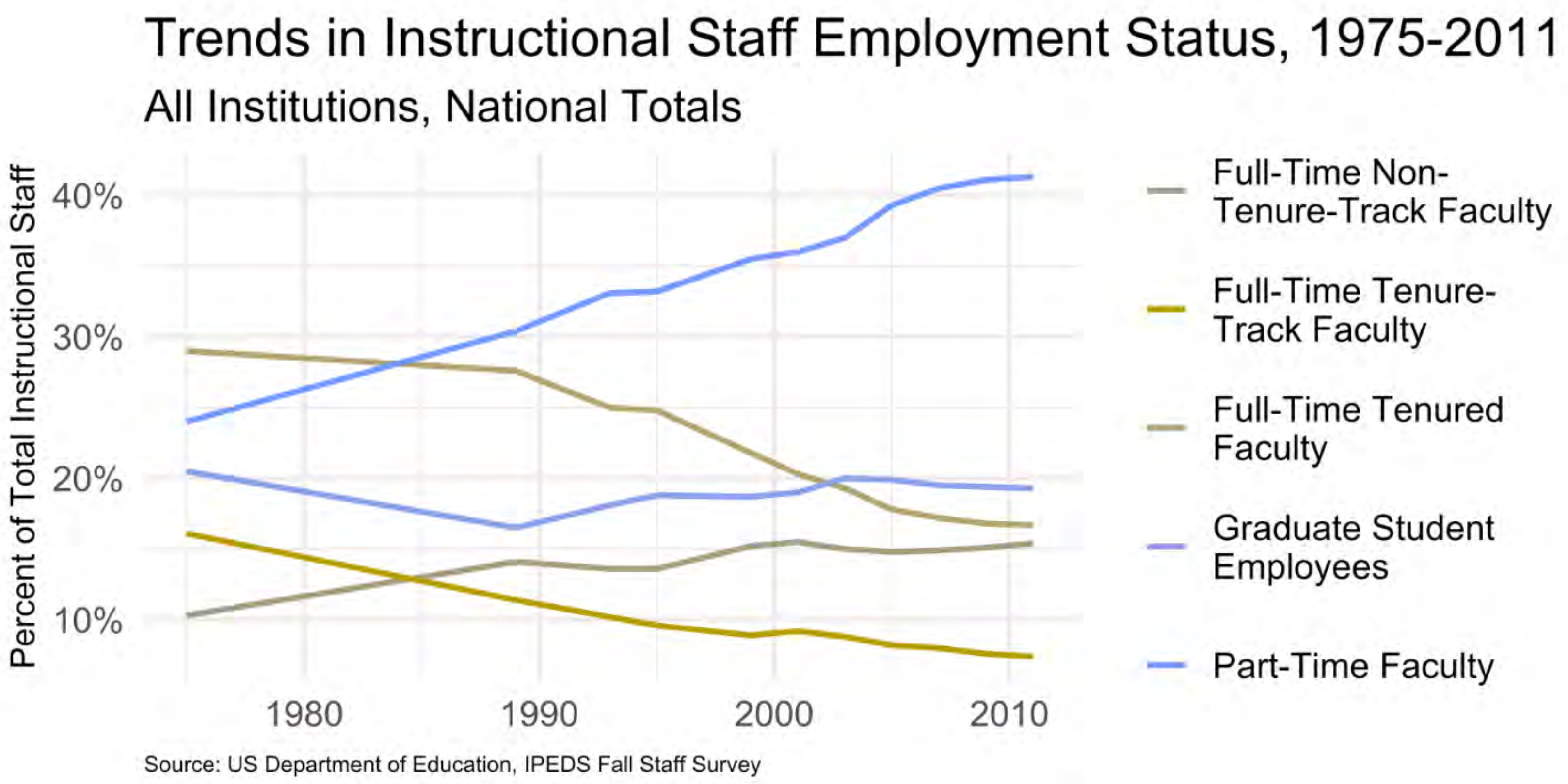
Use an accessible color scale

This is how the previous plot might look like to someone with Deuteranopia (a type of red-green confusion)

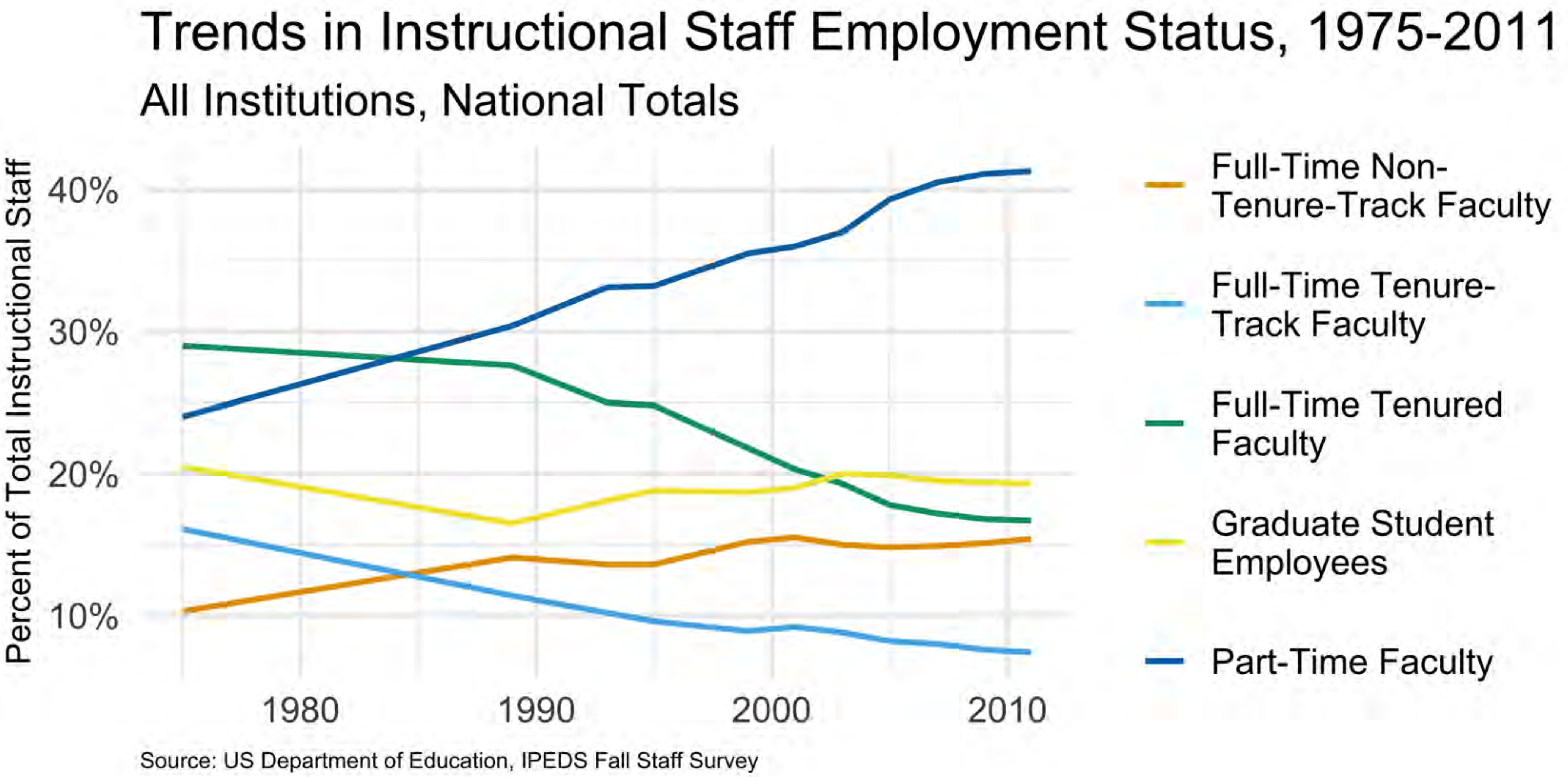


Use an accessible color scale

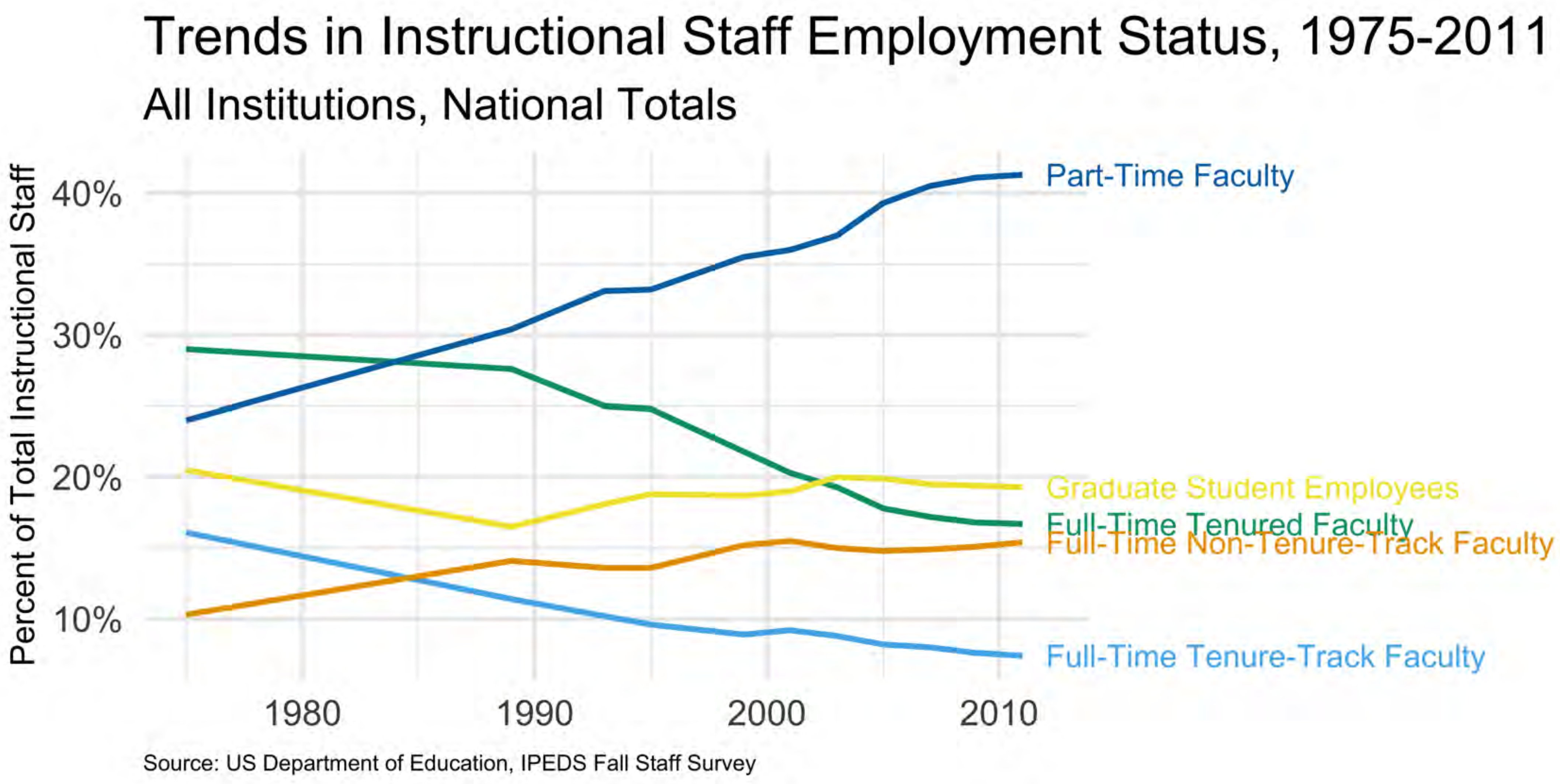
This is it might look like to someone with Protanopia (also a type of red-green confusion)



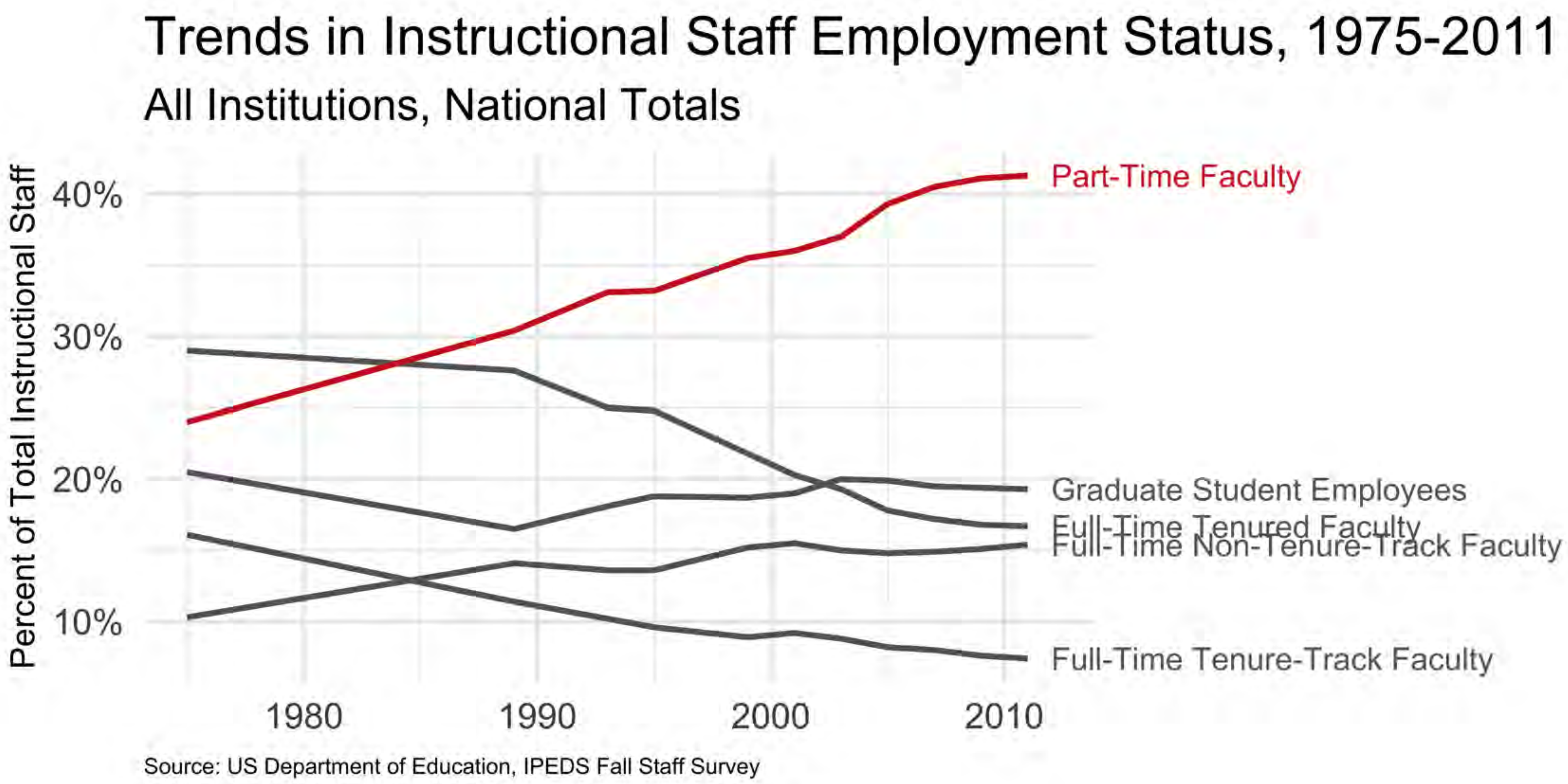
Use an accessible color scale



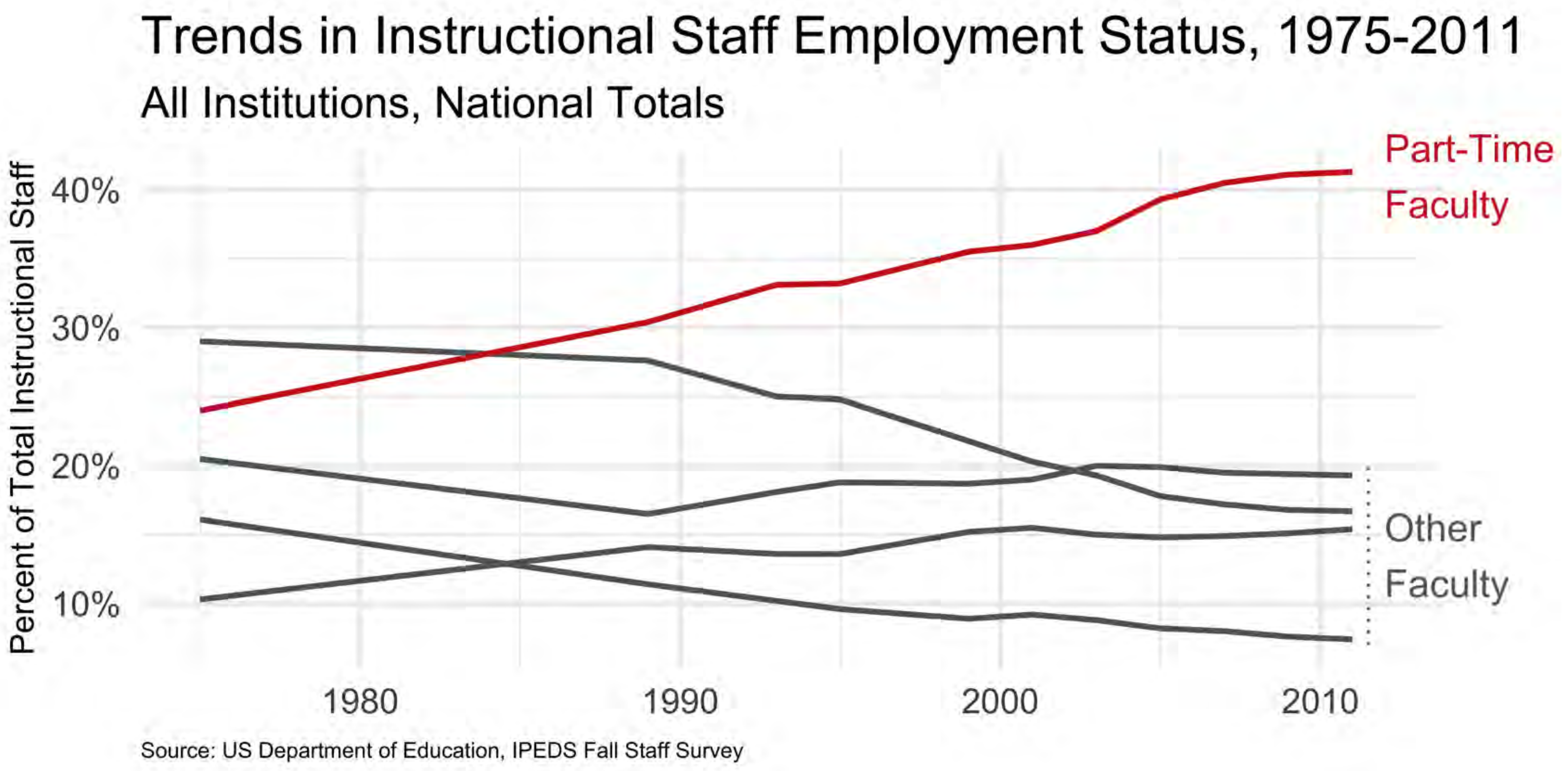
Use direct labeling



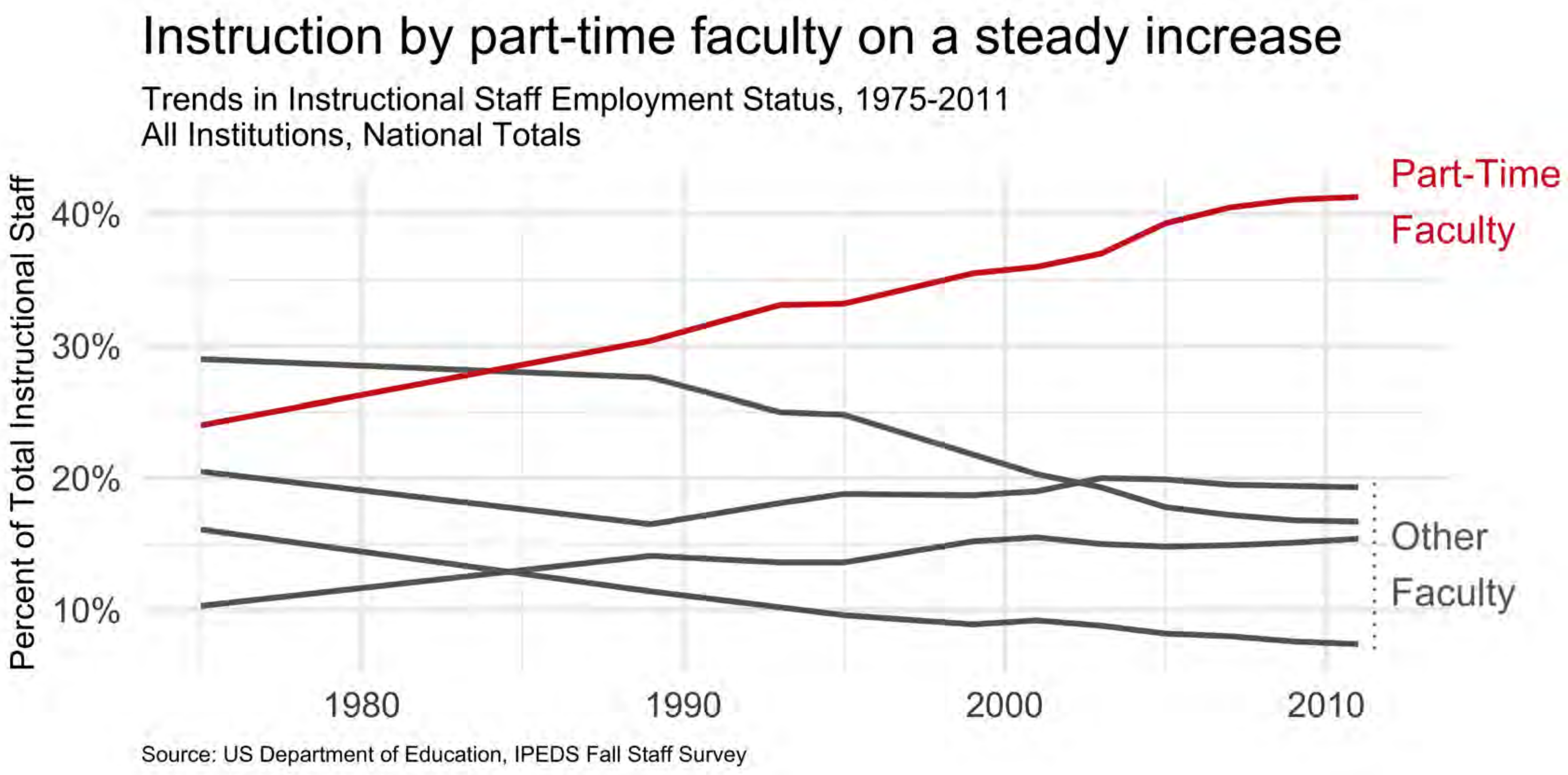
Use color to draw attention



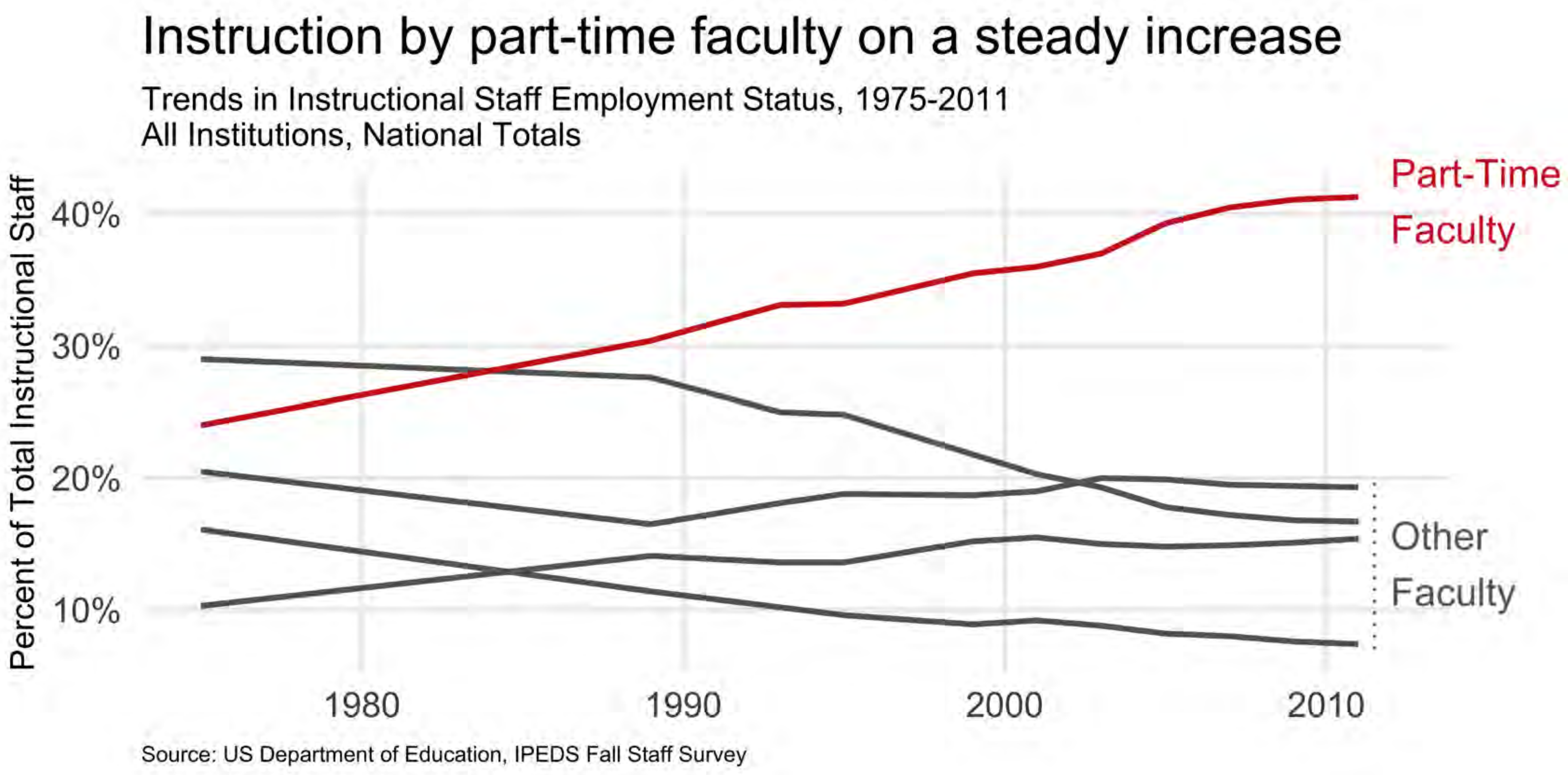
Pick a purpose



Use labels to communicate the message



Simplify


Instruction by part-time faculty on a steady increase
Trends in Instructional Staff Employment Status, 1975-2011
All Institutions, National Totals
Percent of Total Instructional Staff
40%
30%
20%
10%
1980
1990
2000
2010
Part-Time Faculty
Other Faculty
Source: US Department of Education, IPEDS Fall Staff Survey

Summary

- Represent percentages as parts of a whole
- Place variables representing time on the x-axis when possible
- Pay attention to data types, e.g., represent time as time on a continuous scale, not years as levels of a categorical variable
- Prefer direct labeling over legends
- Use accessible colors
- Use color to draw attention
- Pick a purpose and label, color, annotate for that purpose
- Communicate your main message directly in the plot labels
- Simplify before you call it done (a.k.a. "Before you leave the house, look in the mirror and take one thing off")

Case study 2:
Bachelor's degrees

Data

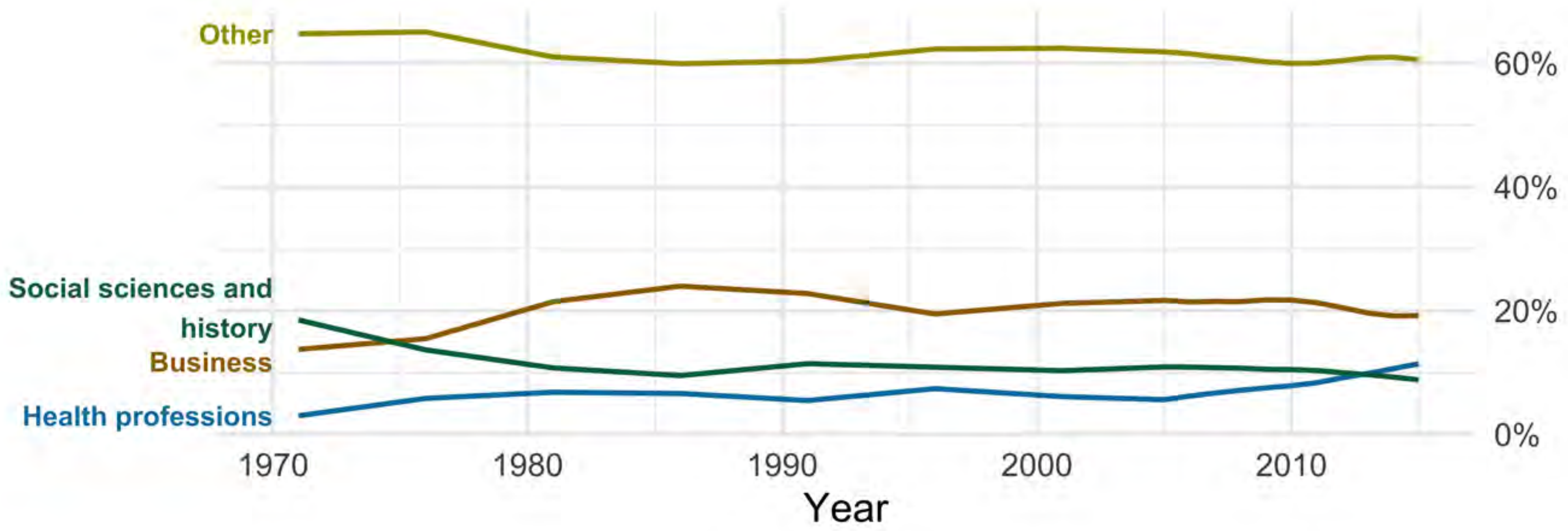
Each row in this dataset represents a field / year combination. For each combination we know the number and the percentage of graduates. Only the most popular three fields are identified, the remaining fields are lumped into "Other".

year	field	perc
1971	Business	0.1374204
1971	Health professions	0.0300370
1971	Social sciences and history	0.1849690
1971	Other	0.6475736
1976	Business	0.1546547
1976	Health professions	0.0582071
1976	Social sciences and history	0.1365342
1976	Other	0.6506039
1981	Business	0.2144289

Should these data be displayed in a table or a plot?

Popular Bachelor's degrees over the years

Field	1971	1976	1981	1986	1991	1996	2001	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Business	14%	15%	21%	24%	23%	19%	21%	22%	21%	21%	21%	22%	22%	21%	20%	20%	19%	19%
Health professions	3%	6%	7%	7%	5%	7%	6%	6%	6%	7%	7%	8%	8%	8%	9%	10%	11%	11%
Social sciences and history	18%	14%	11%	9%	11%	11%	10%	11%	11%	11%	11%	11%	10%	10%	10%	10%	9%	9%
Other	65%	65%	61%	60%	60%	62%	62%	62%	62%	61%	61%	60%	60%	60%	60%	61%	61%	61%



Tables vs. plots

Tables:

- To look up or compare individual values
- To display precise values
- To include detail and summary values
- To display quantitative values including more than one unit of measure

Plots:

- To reveal relationships among whole sets of values
- To display a message that is contained in the shape of the values (e.g., patterns, trends, exceptions)

Add visualizations to your table

Popular Bachelor's degrees over the years																			
Field	Trend	1971	1976	1981	1986	1991	1996	2001	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Business		14%	15%	21%	24%	23%	19%	21%	22%	21%	21%	21%	22%	22%	21%	20%	20%	19%	19%
Health professions		3%	6%	7%	7%	5%	7%	6%	6%	6%	7%	7%	8%	8%	8%	9%	10%	11%	11%
Social sciences and history		18%	14%	11%	9%	11%	11%	10%	11%	11%	11%	11%	11%	10%	10%	10%	10%	9%	9%
Other		65%	65%	61%	60%	60%	62%	62%	62%	62%	61%	61%	60%	60%	60%	60%	61%	61%	61%

Summary

- A table may be preferable to a visualization
- A table can be enhanced with a visualization

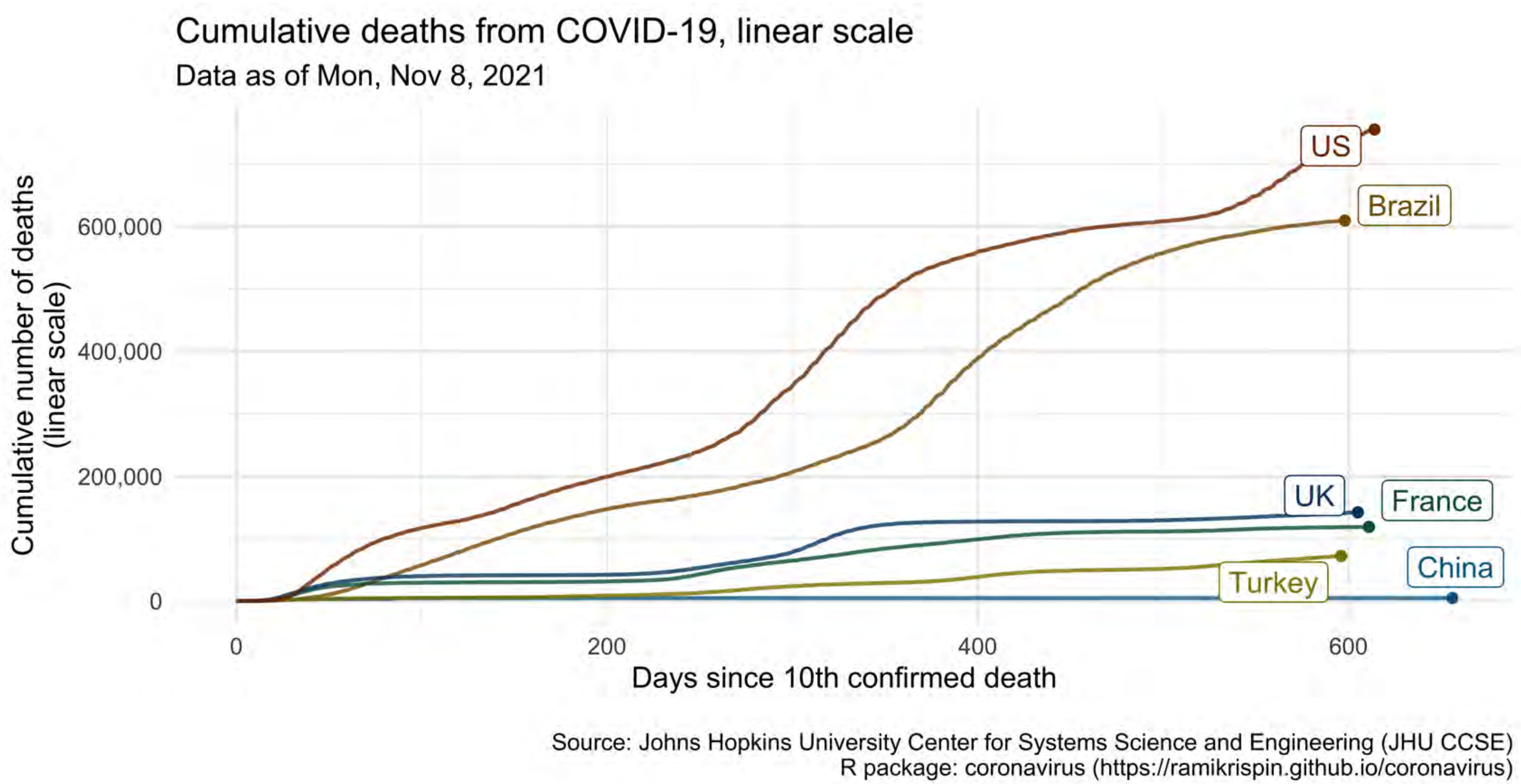
Case study 3:
COVID-19 Deaths

Data

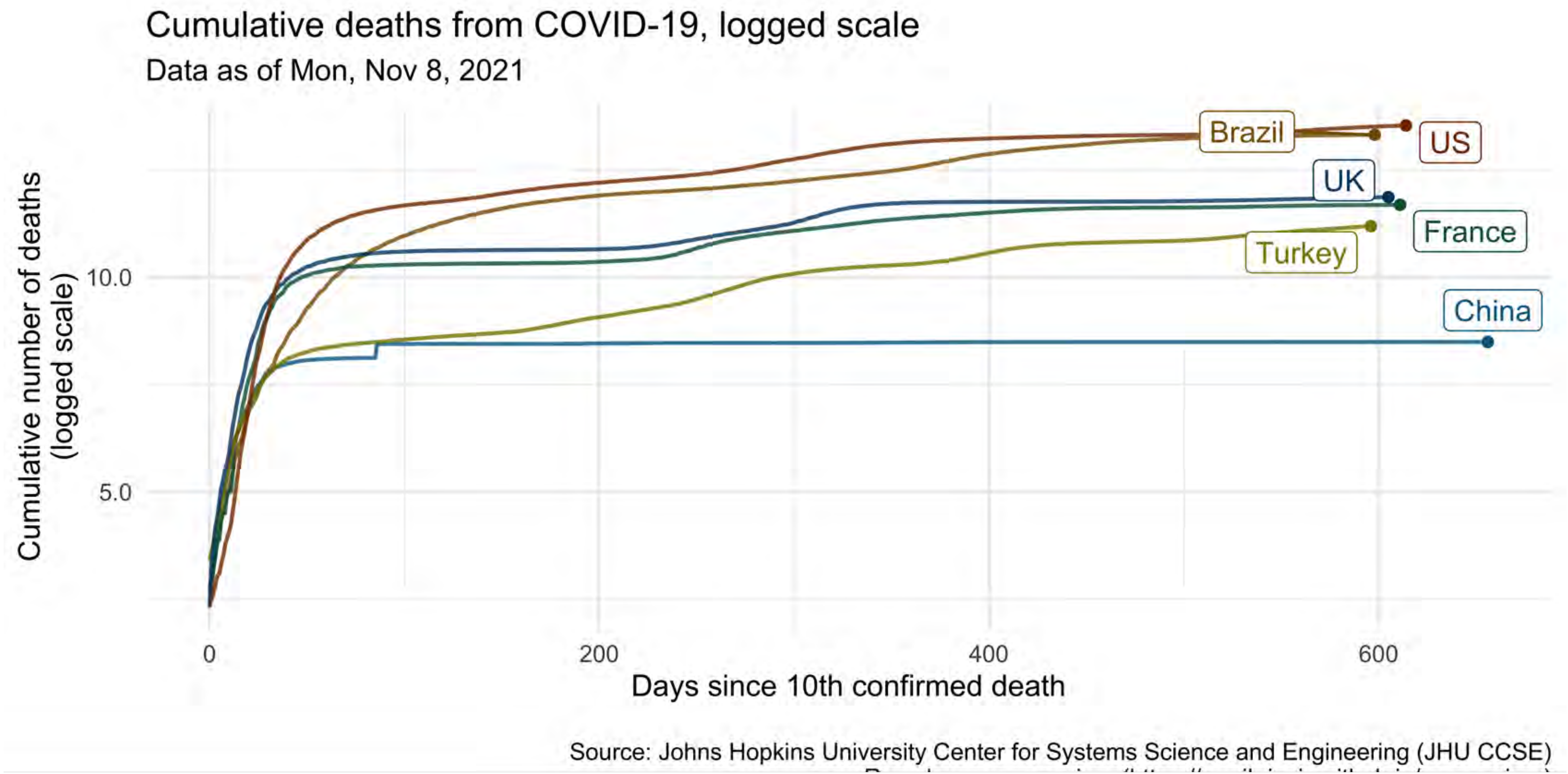
Each row represents a country date combination. For each combination we have the total number of cases, the cumulative cases, and the days elapsed since 10th confirmed COVID-19 case in that country.

country	date	tot_cases	cumulative_cases	days_elapsed
China	2020-01-22	17	17	0
China	2020-01-23	1	18	1
China	2020-01-24	8	26	2
China	2020-01-25	16	42	3
China	2020-01-26	14	56	4
China	2020-01-27	26	82	5
China	2020-01-28	49	131	6
China	2020-01-29	2	133	7
China	2020-01-30	38	171	8

Plot 1: Linear scale



Plot 2: Logged scale



Which plot do you prefer, and why?

Zoom in to the first 25 days: Which plot do you prefer, and why?

Summary

- Your data scale matters!
- Keep in mind not just best practices, but also your audience and the amount of supplementary information you can provide

visualizion **resources**

Resources

- Books:
  - Data Visualization: A practical introduction by Kieran Healy
  - Fundamentals of Data Visualization by Claus O. Wilke
  - How charts lie by Alberto Cairo
  - Presenting Data Effectively by Stephanie Evergreen
  - Datavision by David McCandless
- Community: Data Visualization Society
- Tools: All visualizations presented have been created with R and ggplot2
  - For those who are interested, the source code can be found here