Visualizing Proportions and Enrichments

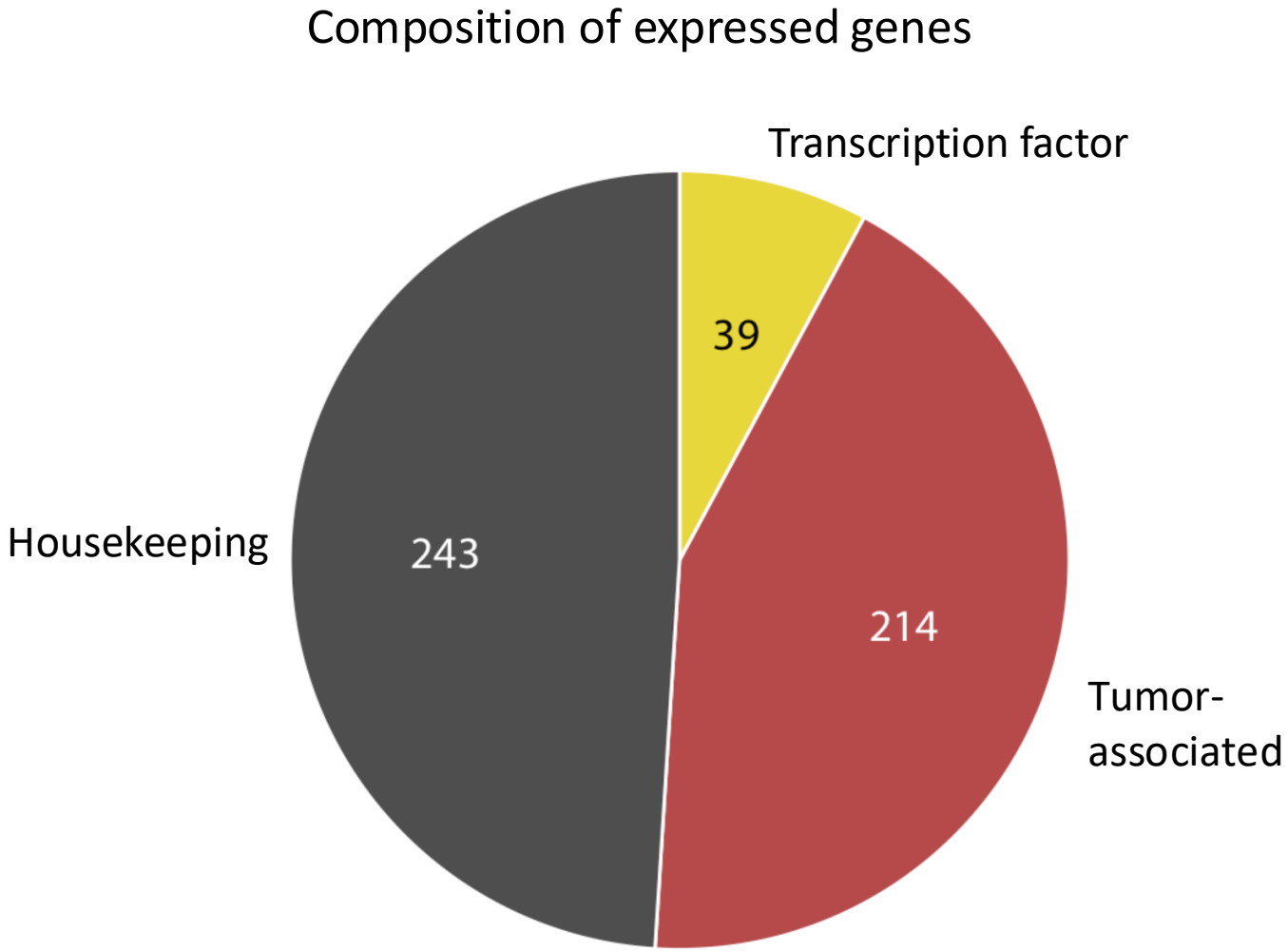
HUGEN 2073

Genomic Data Visualization and Integration

Slides borrowed/modified from H. J. Park with permission

Pie charts to show proportions

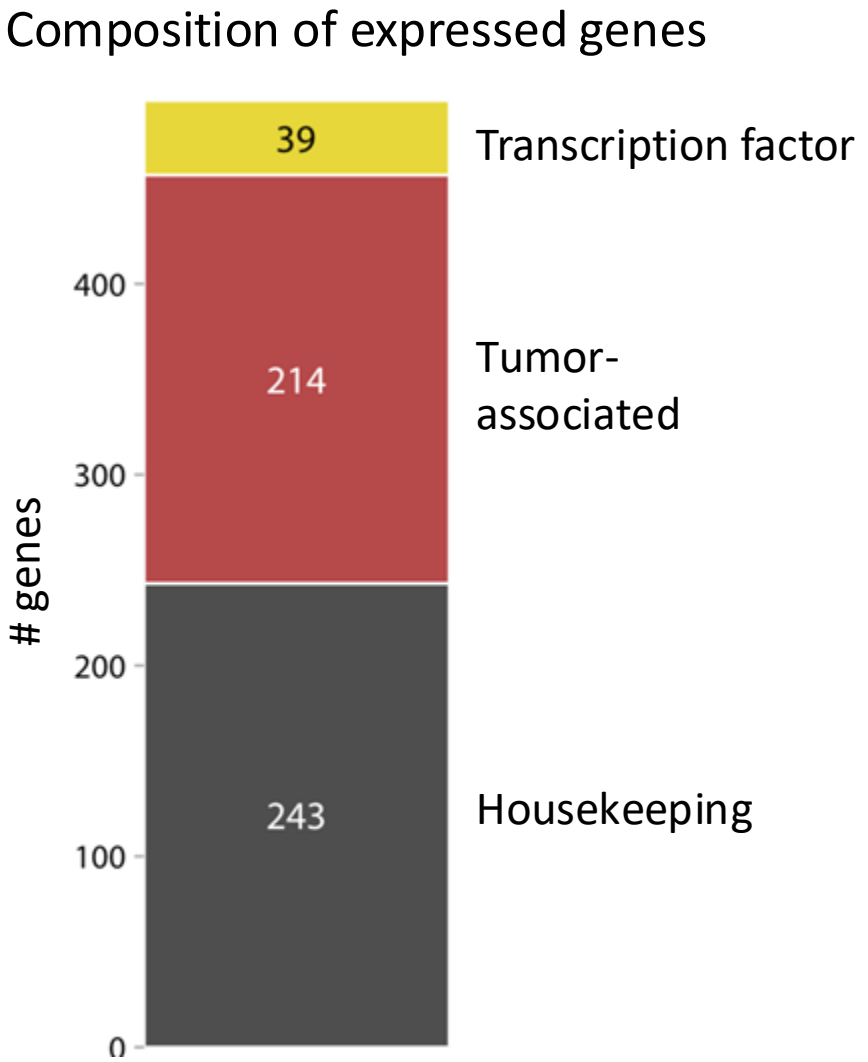
- Pros:
  - can show how some entity breaks down into pieces (e. g., n=496)
  - easy to read when the data amounts to an entirety
- Cons:
  - Difficult to see the total number



Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Stacked bar to show proportions

- ~~Pros: can show how some entity breaks down into pieces~~
- Pros: easy to see the total number
- Cons: less sense of totality (there might be other classes)

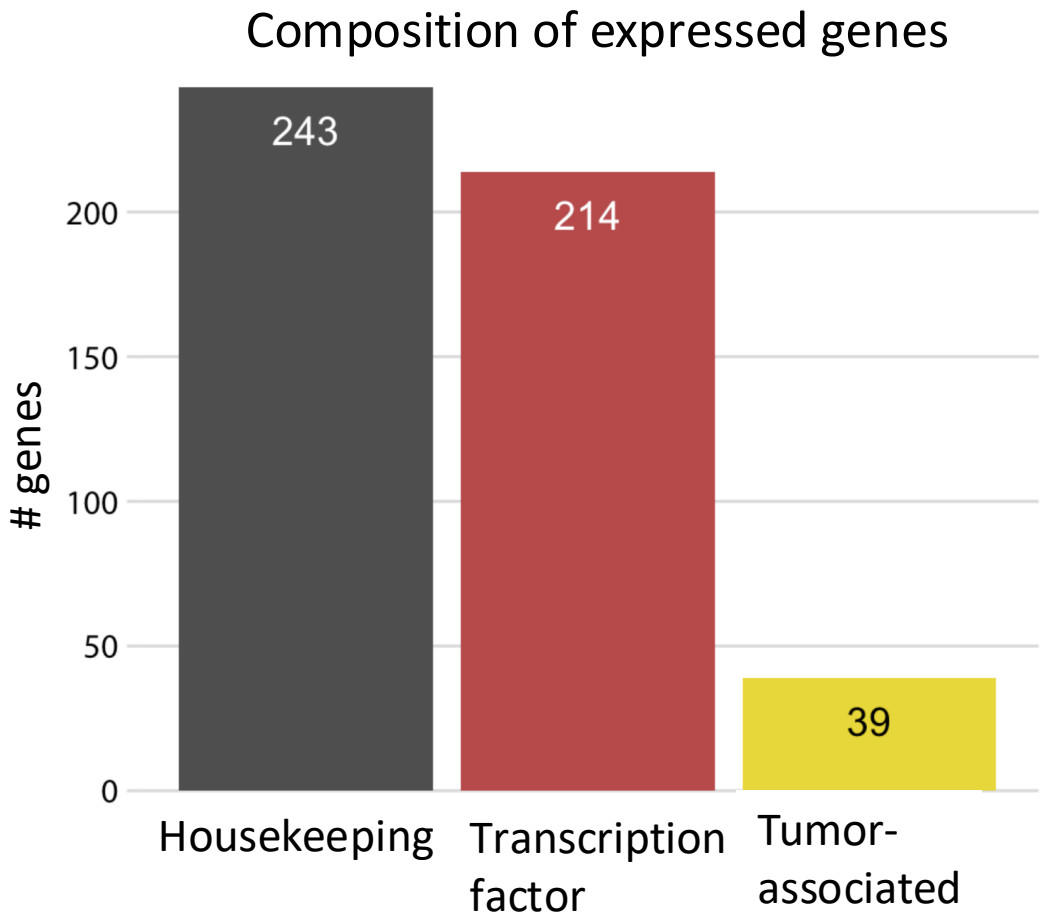


Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Side-by-side bar to show proportions

- Pros: better to compare
- Cons: the relationship of each to the total not obvious

➡ can annotate percentage values

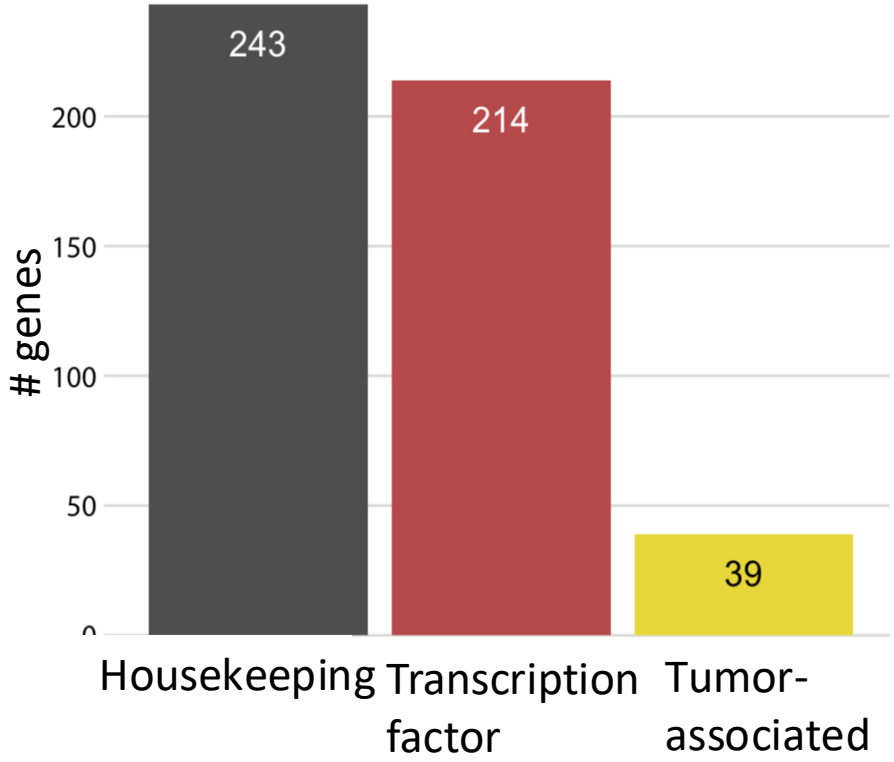


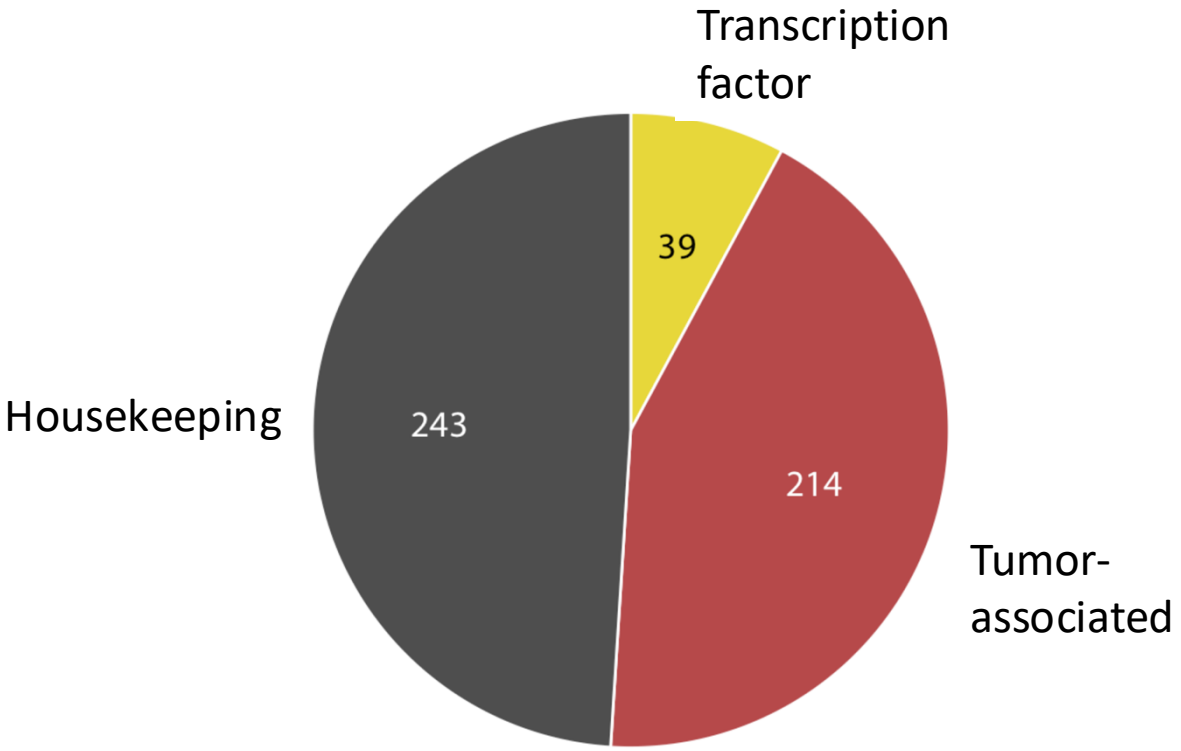
Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Pie chart and barchart make different points

Q: Which is better to show when your message is that…

- the combination of TF and TA jointly had a small majority over the HK or
- HK is big compared to each of the others ?



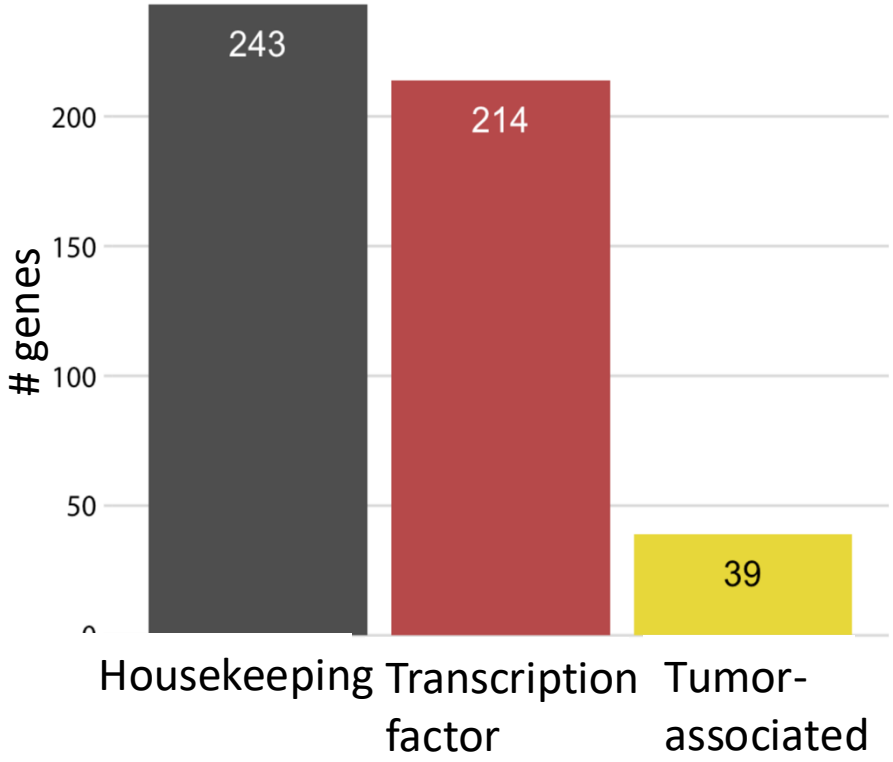


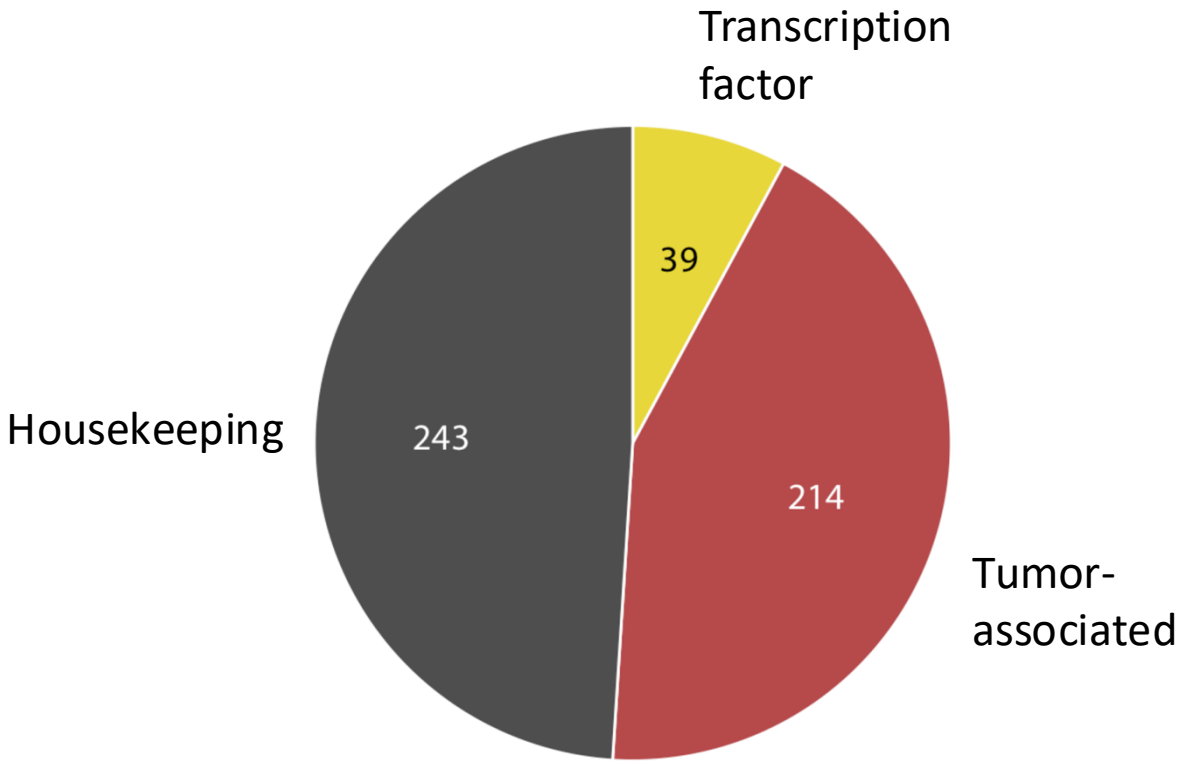
Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Pie chart and barchart make different points

Q: Which is better to use when there are

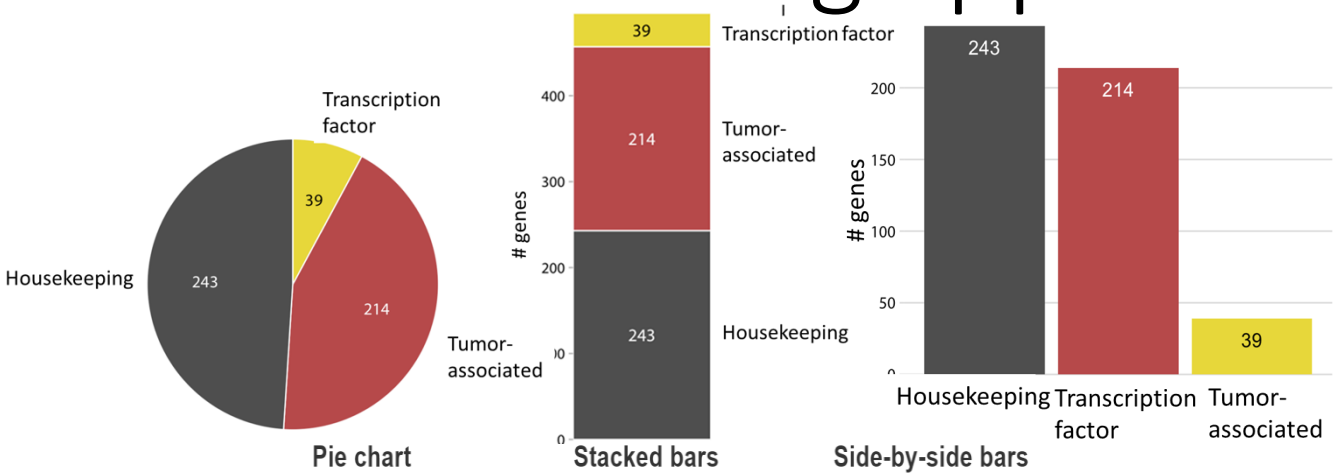
- only those three gene groups or
- other gene groups?





Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Pros and cons of the visualizing approaches for proportions



Clearly visualizes the data as proportions of a whole	✓	✓	×
Allows easy visual comparison of the relative proportions	×	×	✓
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✓	×	×
Looks visually appealing even for very small datasets	✓	×	✓
Works well when the whole is broken into many pieces	×	×	✓
Works well for the visualization of many sets of proportions or time series of proportions	×	✓	×

By "closing" the plots

By putting them in such a way that facilitates comparison
Using the behavior of the human visual system
Yes for pie chart, not sure for side-by-side bars

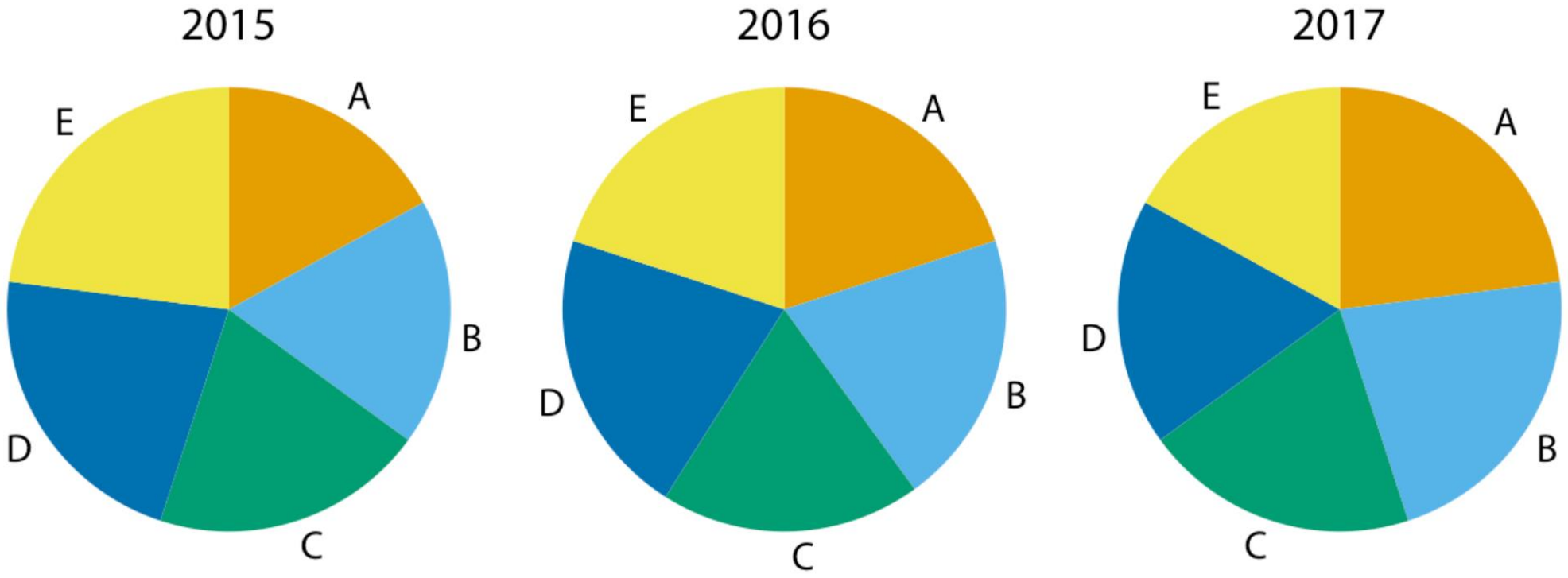
By having them all start at zero baseline

By putting multiple proportions in a bar

Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

A case for side-by-side bars

Proportions of five gene groups across time points, A–E

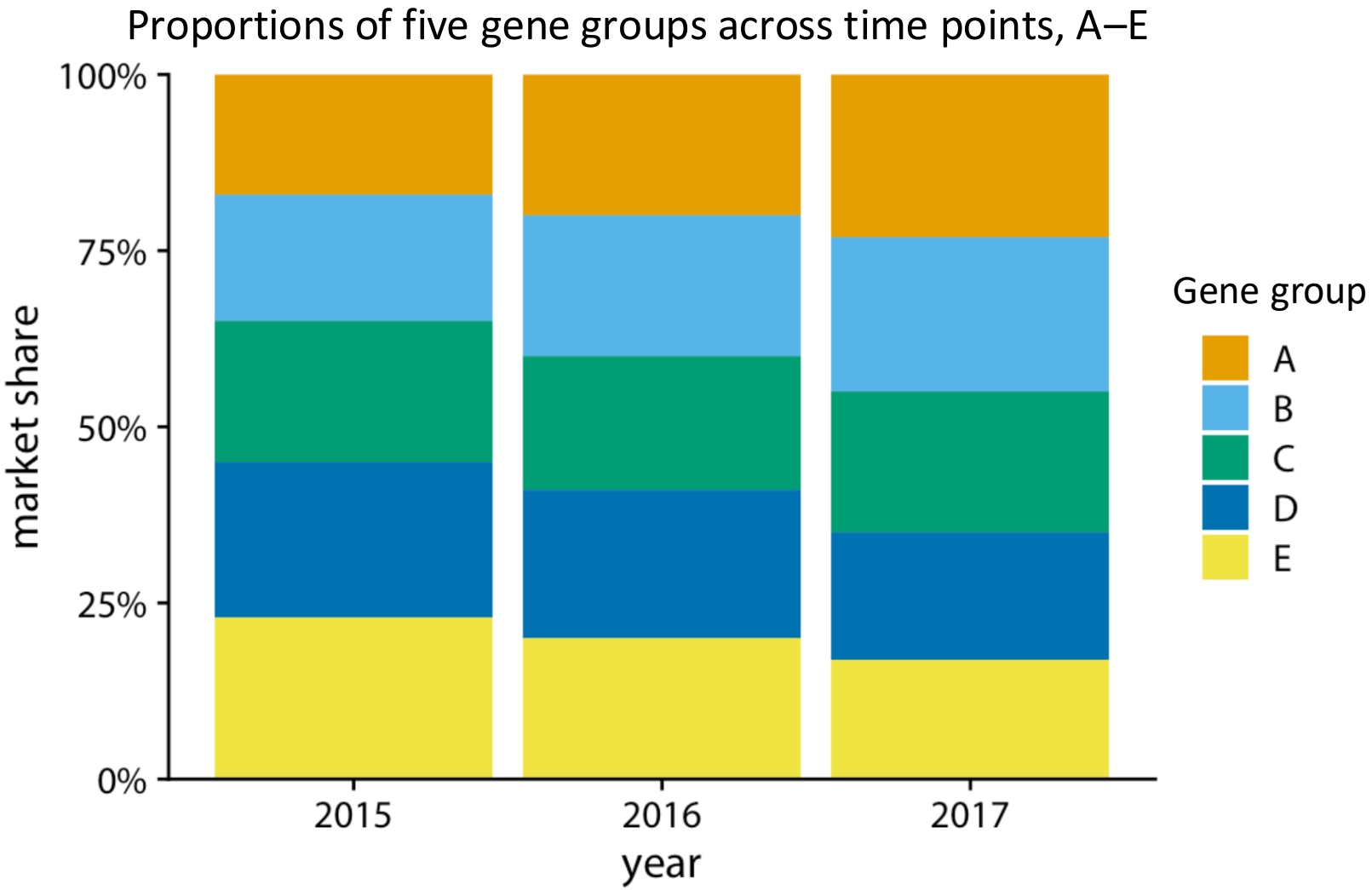


Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

- Cons: unclear for small changes

A case for side-by-side bars

- Pros: clearer for big changes, e. g., gene group A and E
- Cons:
  - Unclear for small changes especially placed in the middle
  - Difficult to compare within each year



Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

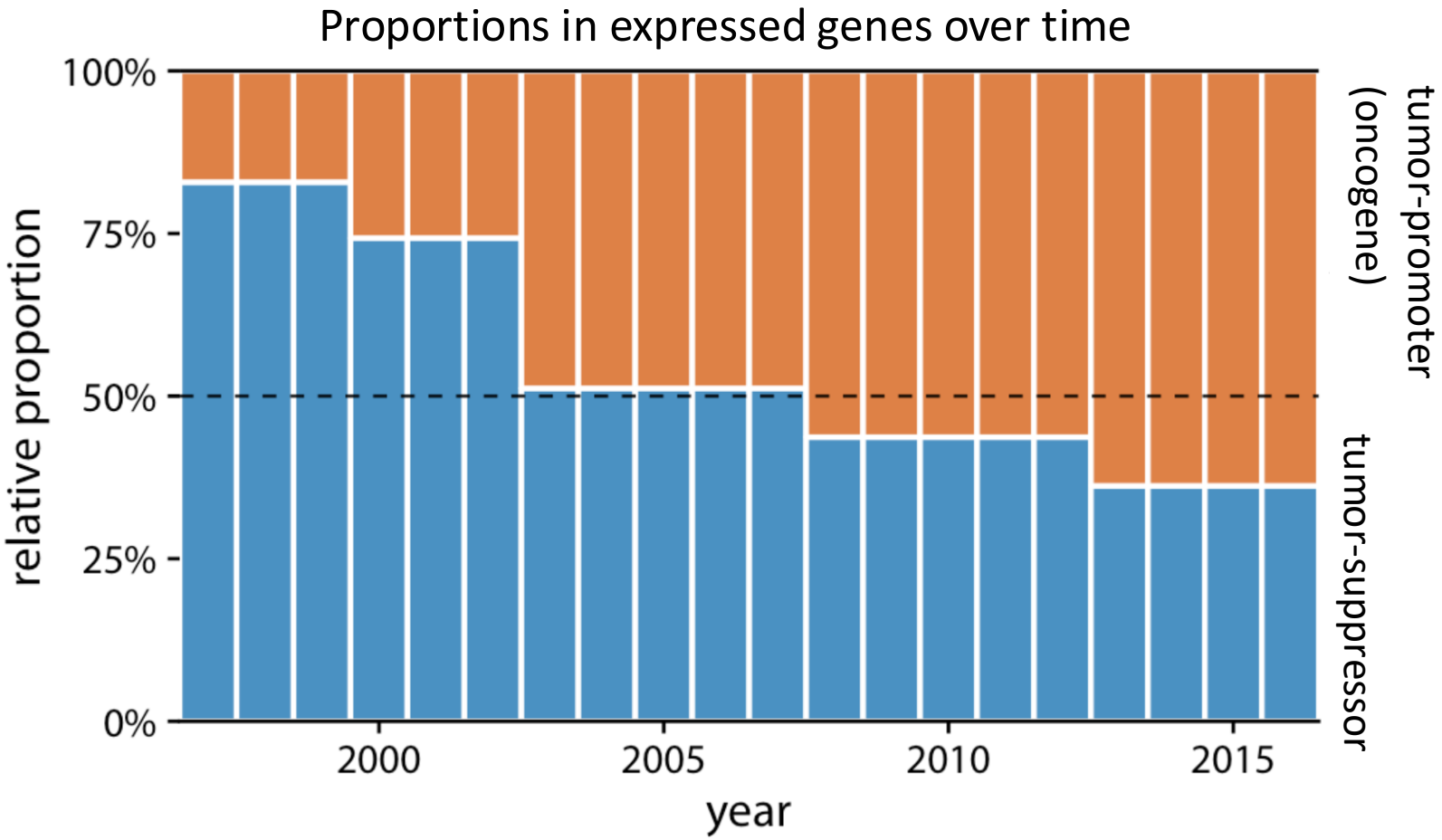
A case for side-by-side bars

- Pros: clear for all changes regardless of their relative positions, e.g. gene group B
- Cons: difficult to read if too many bars

Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

A case for stacked bars

- Pros: work well for two-group data
- A good use of horizontal line at 50%



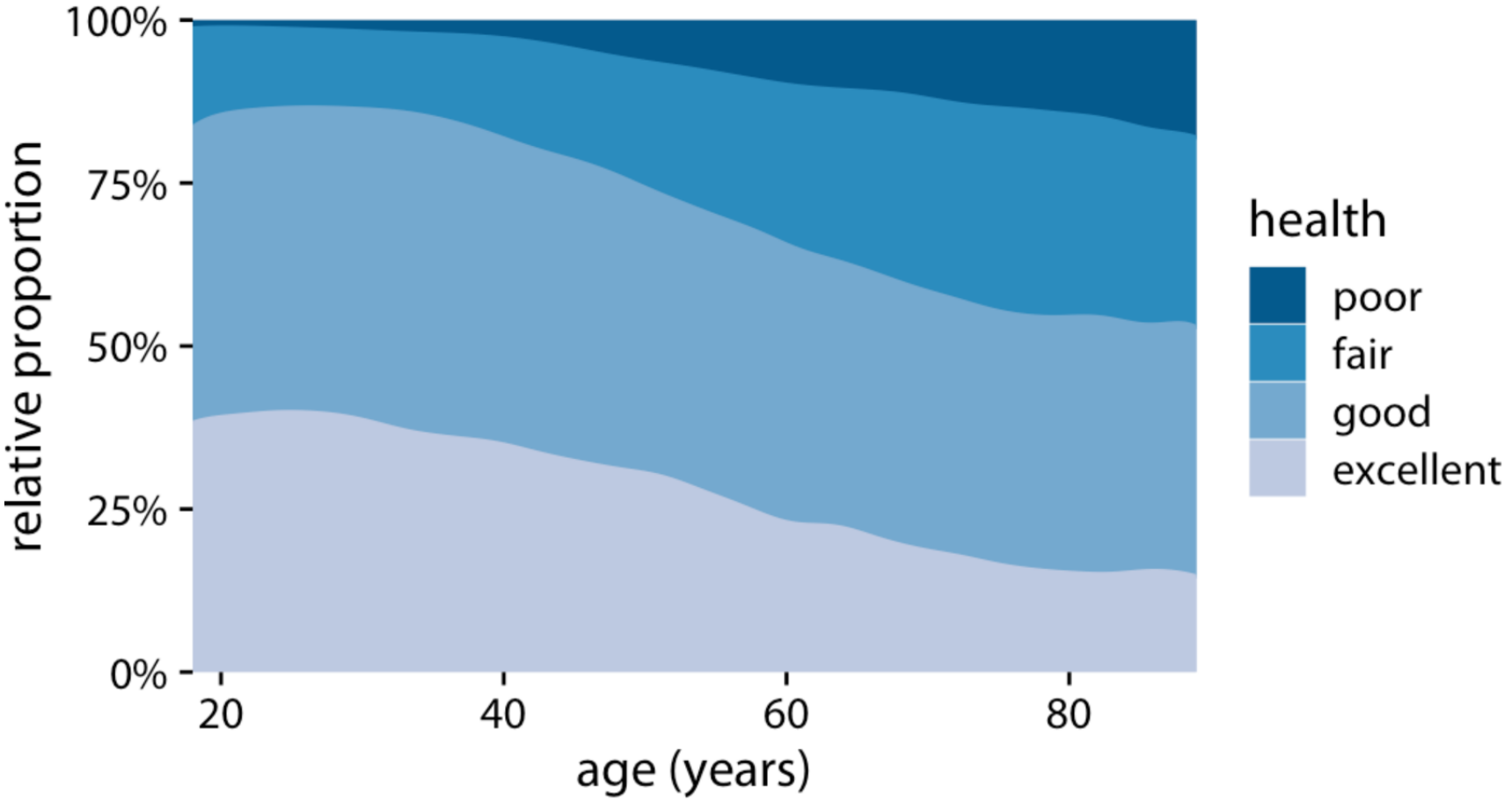
Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Stacked densities show how proportions change in response to a continuous variable

Cons:

- The bars can't be compared easily because of different baseline
- Can't indicate the size of different age groups, e. g., young people vs. old people

Health status by age, as reported by the general social survey (GSS)

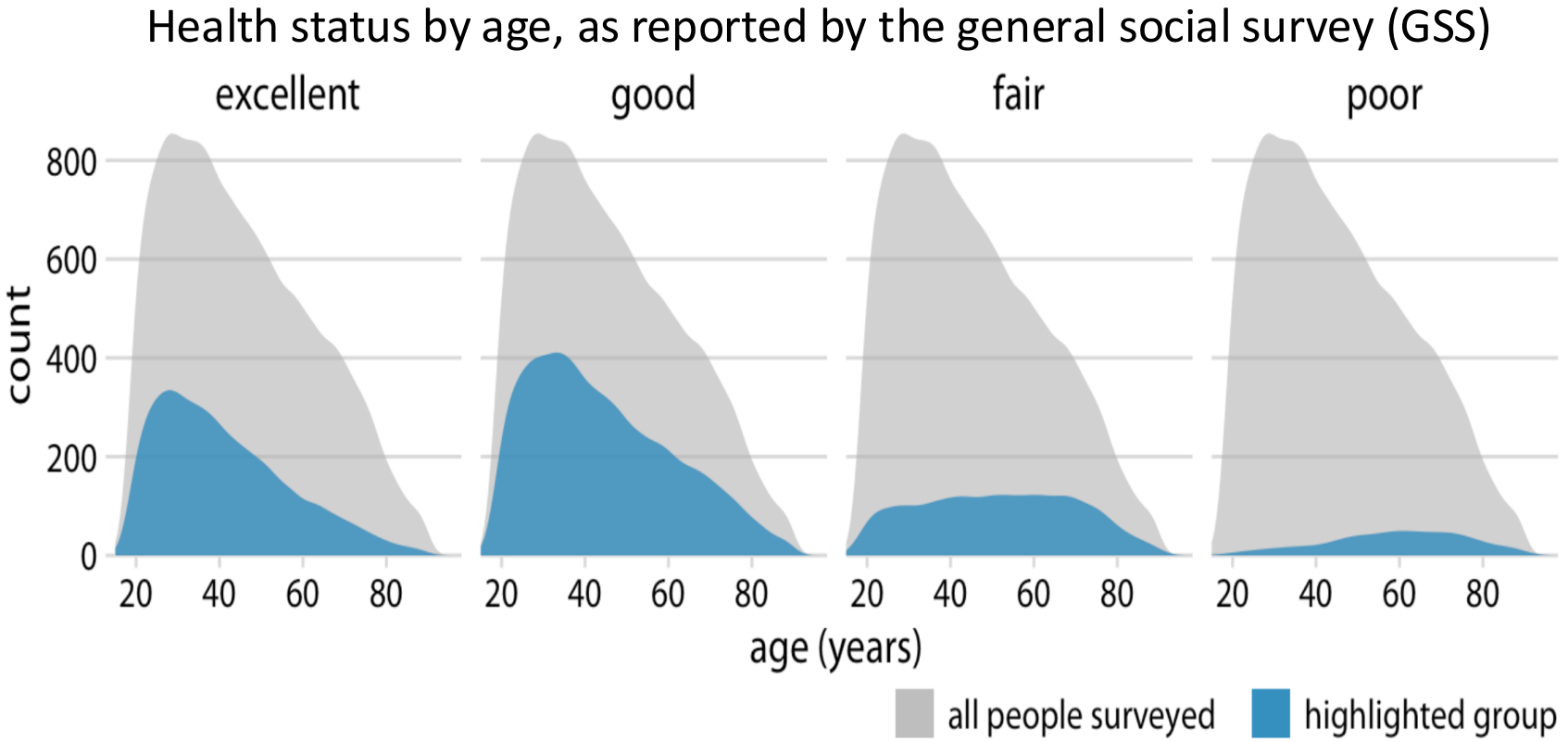


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Side-by-side stacked densities showing numbers as parts of the total

Solution:

- Making a separate plot for each category
- Plotting the size of different age groups

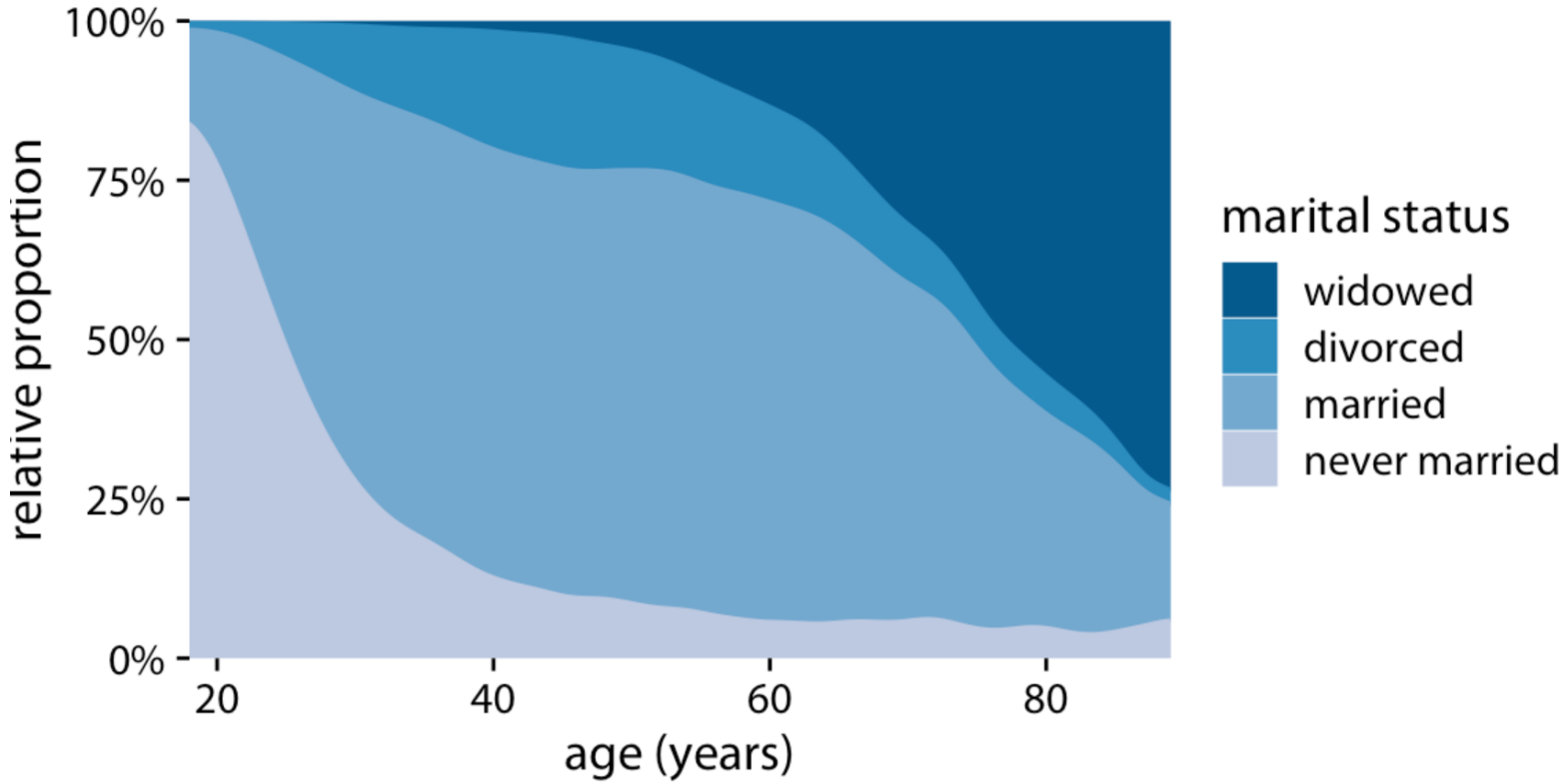


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Stacked densities overwhelmed by prevailing signals

e. g., drastic changes in widowed and never married distort for divorced and married

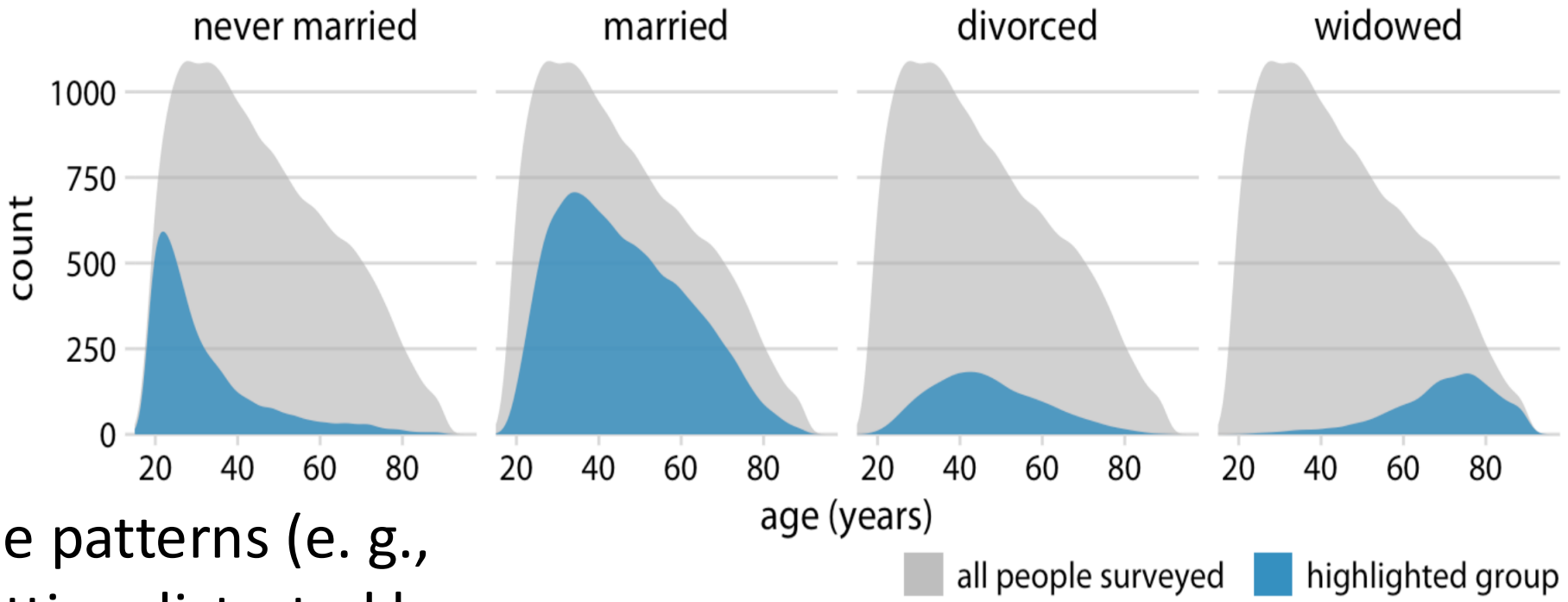
Marital status by age, as reported by the general social survey (GSS)



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Side-by-side stacked densities showing numbers as parts of the total

Marital status by age, as reported by the general social survey (GSS)

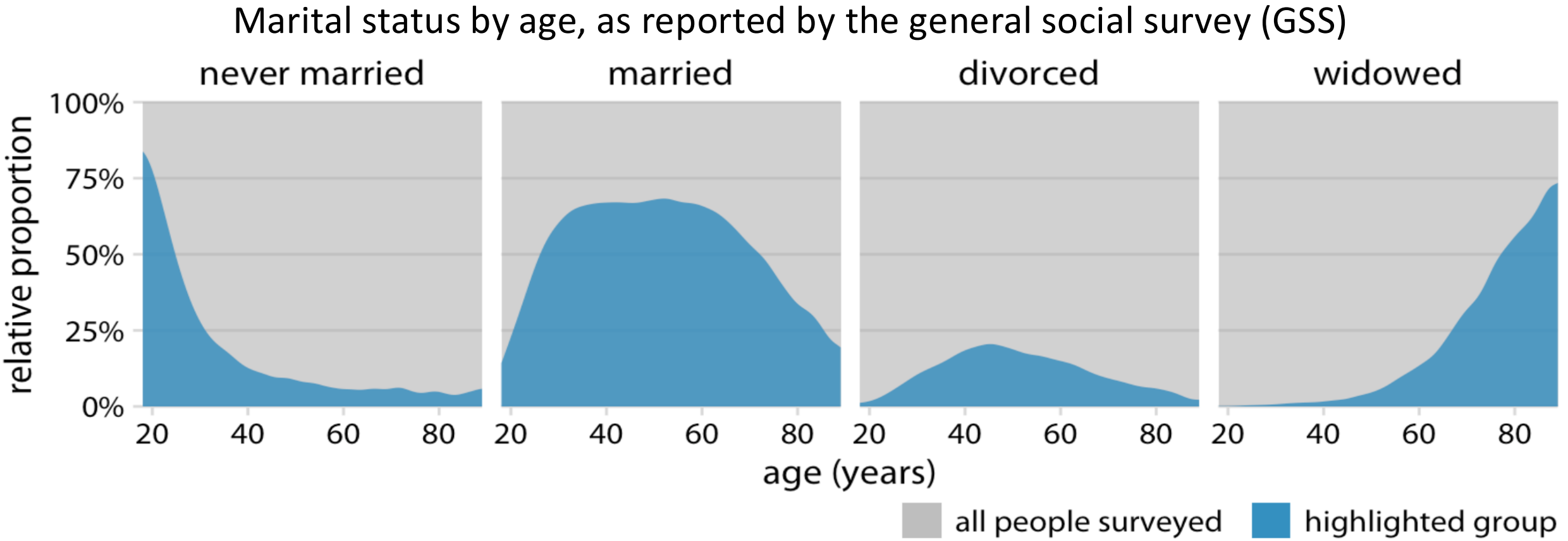


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

- Pros: easily see the patterns (e. g., peaks) without getting distorted by other populations

What if we're interested in determining relative proportions, e. g., at what age > 50% of people married?

Proportions separately as parts of the total

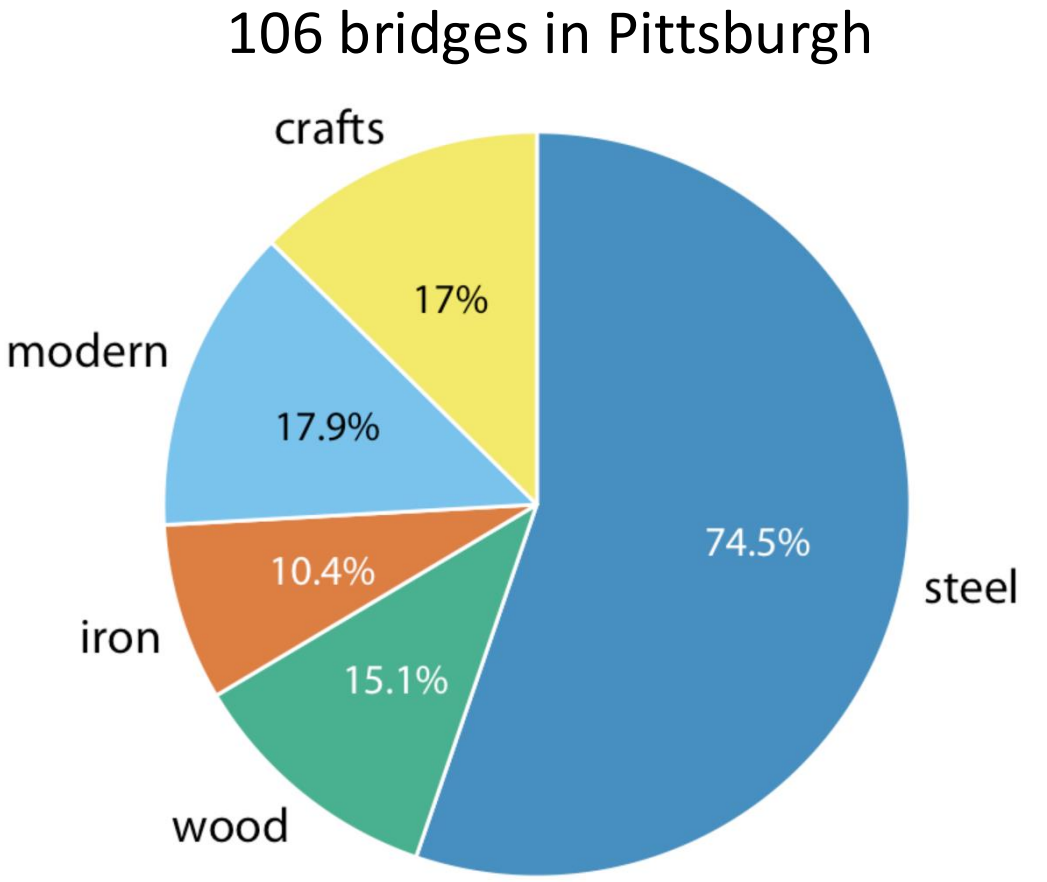


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Nested proportions in pie chart to show proportions in two categorical variables

Nested proportions for two categories : e. g.,

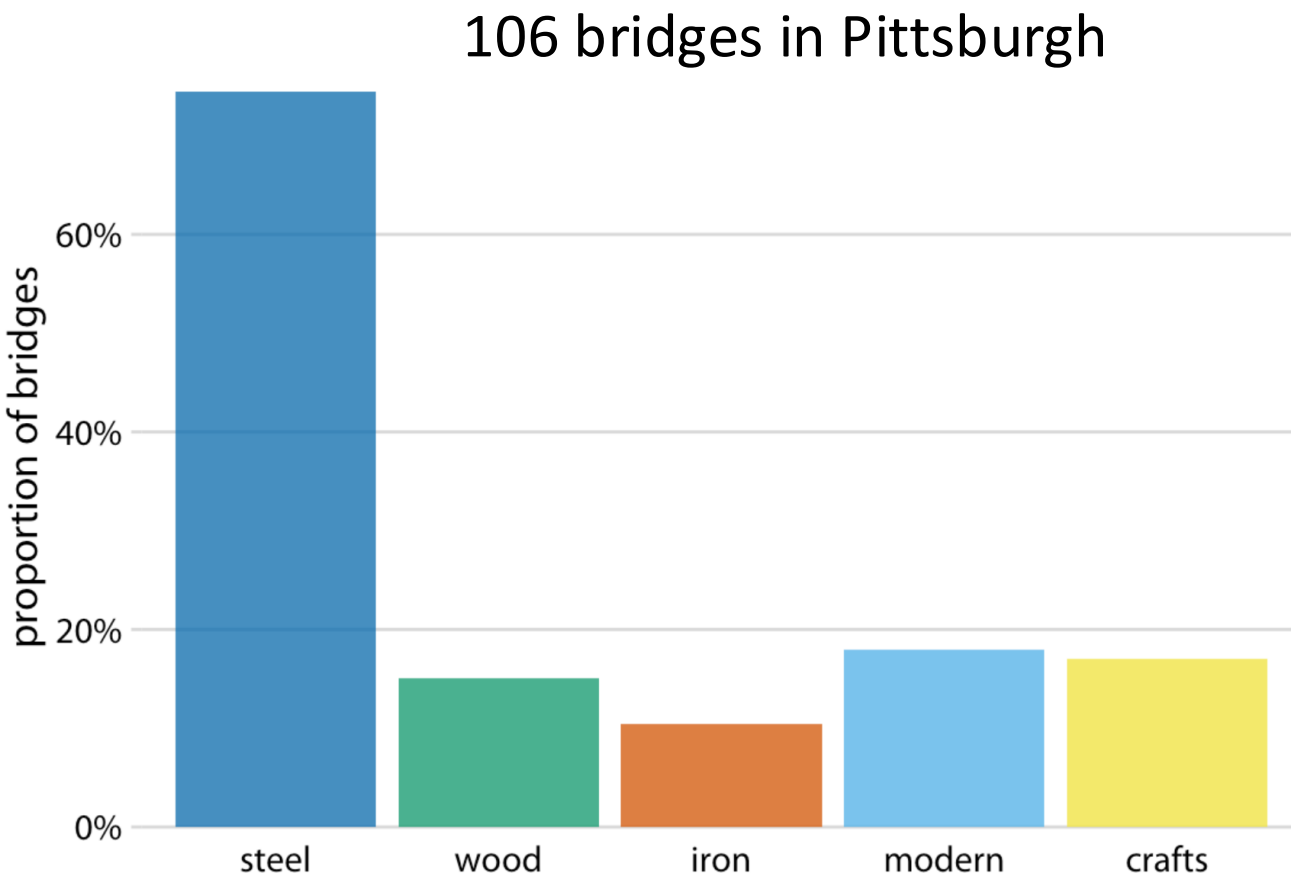
- Health and martial status or
- Genes up- or down-regulated (case vs. control) and whether tumor suppressors or tumor promoters
- Bridges by construction material (steel, wood, iron) and by date of construction (crafts, before 1870, and modern, after 1940)



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Nested proportions in pie chart to show proportions in two categorical variables

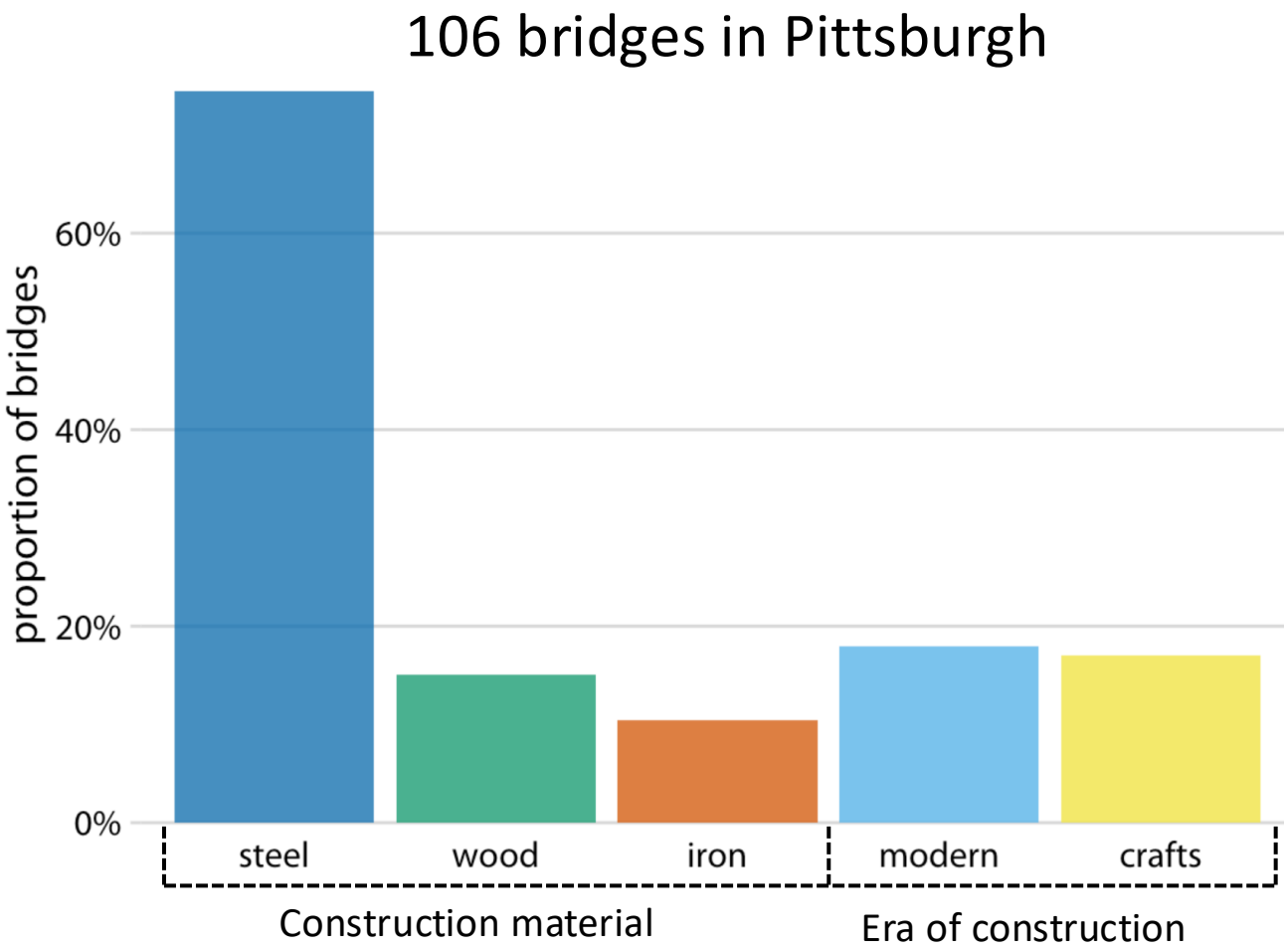
- What's shown: by construction material (steel, wood, iron) and by date of construction (crafts, before 1870, and modern, after 1940)
- It's OK they don't add up to 100 (%) (?)



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Nested proportions in pie chart to show proportions in two categorical variables

- E. g., by construction material (steel, wood, iron) and by date of construction (crafts, before 1870, and modern, after 1940)
- It's OK they don't add up to 100 (%)
- Better to indicate the separation

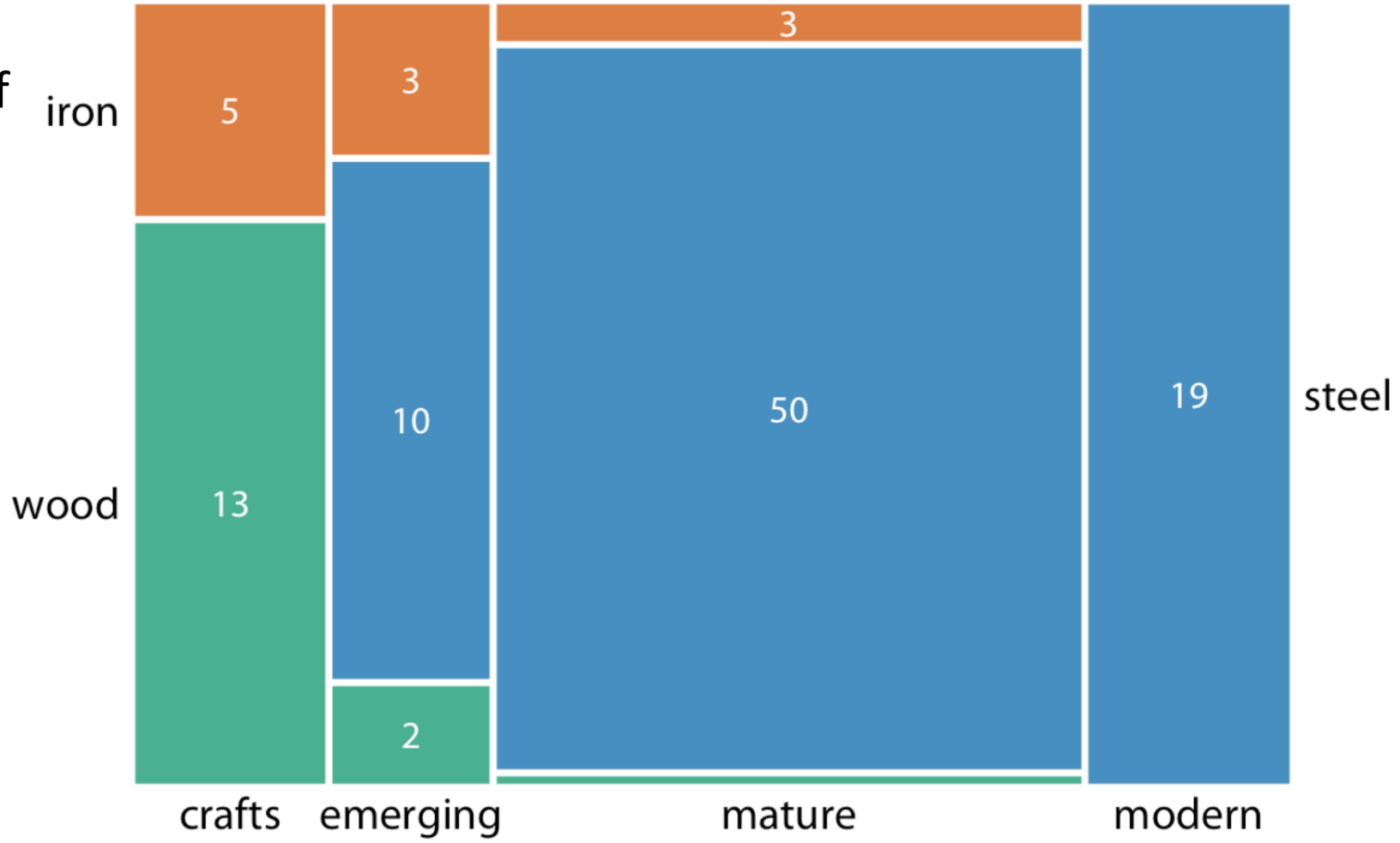


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Nested proportions in mosaic plot

106 bridges in Pittsburgh

- Like stacked barplots: bar heights relative proportions of the y variable
- Unlike stacked barplots: bar width relative proportion of the x variable
- Resulting in the rectangles proportional to the number of cases for each combination

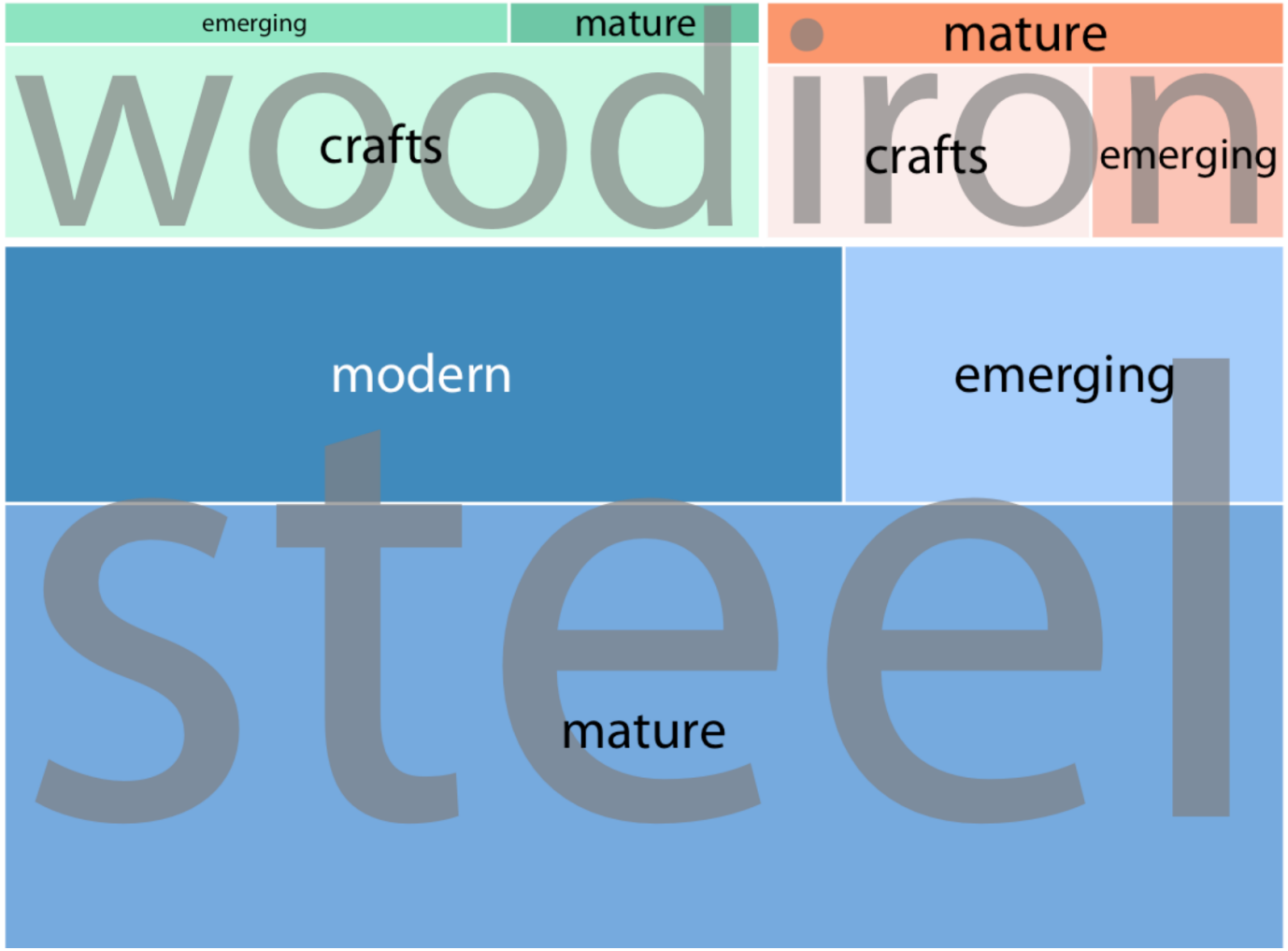


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Nested proportions in treemap

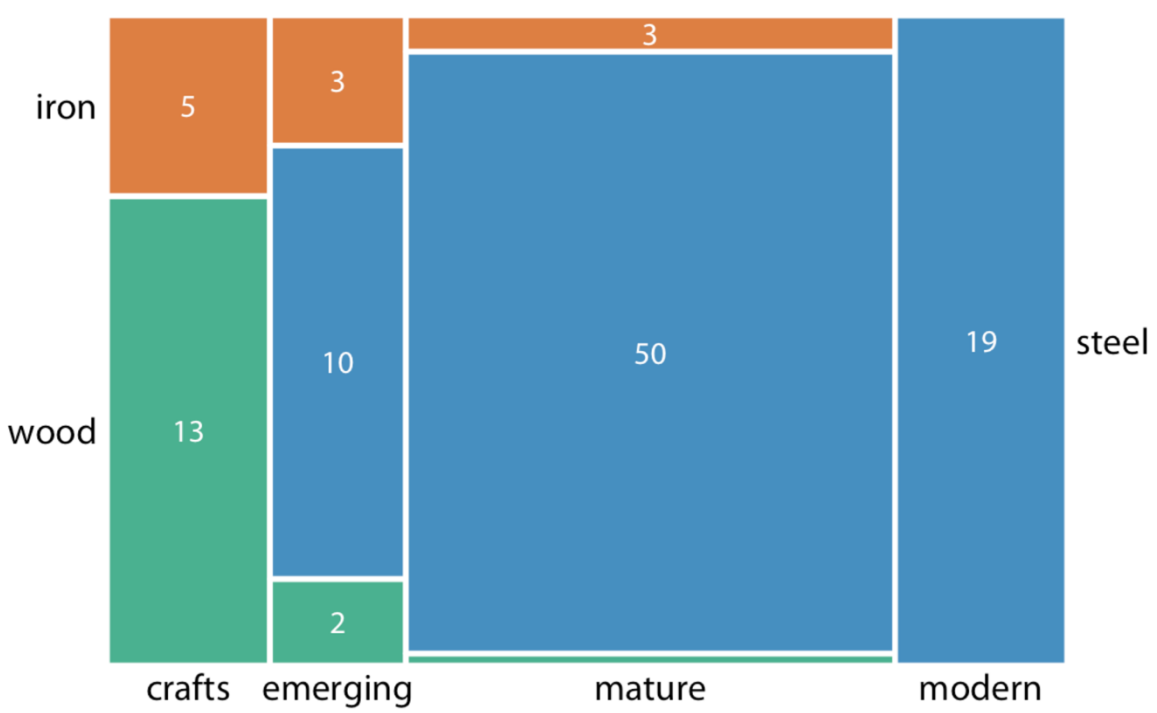
- Like mosaic plot: subdividing enclosing rectangles into smaller areas to represent the proportions
- Unlike mosaic plot: recursively nest rectangles inside each other

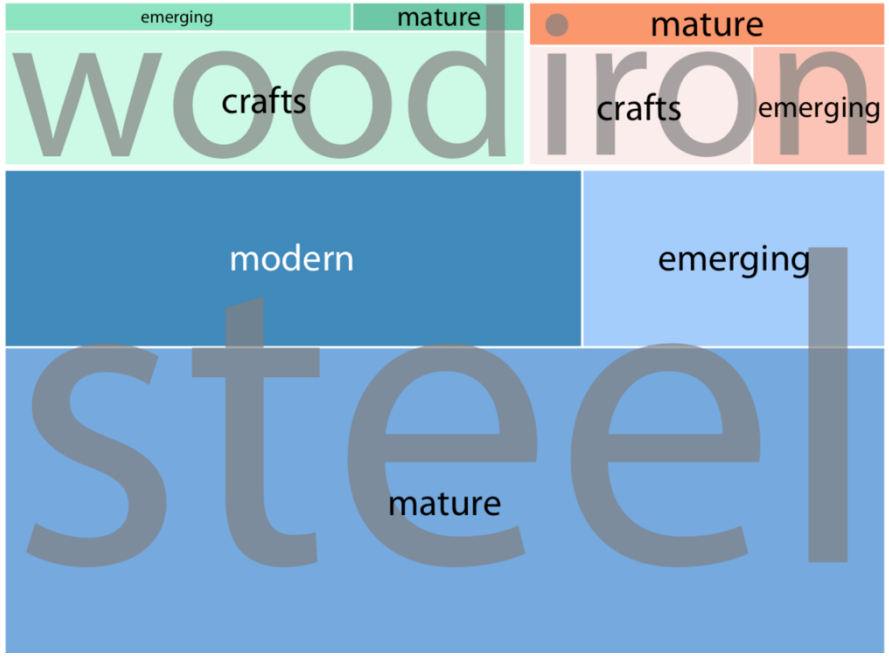
106 bridges in Pittsburgh



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Mosaic plot vs. treemap



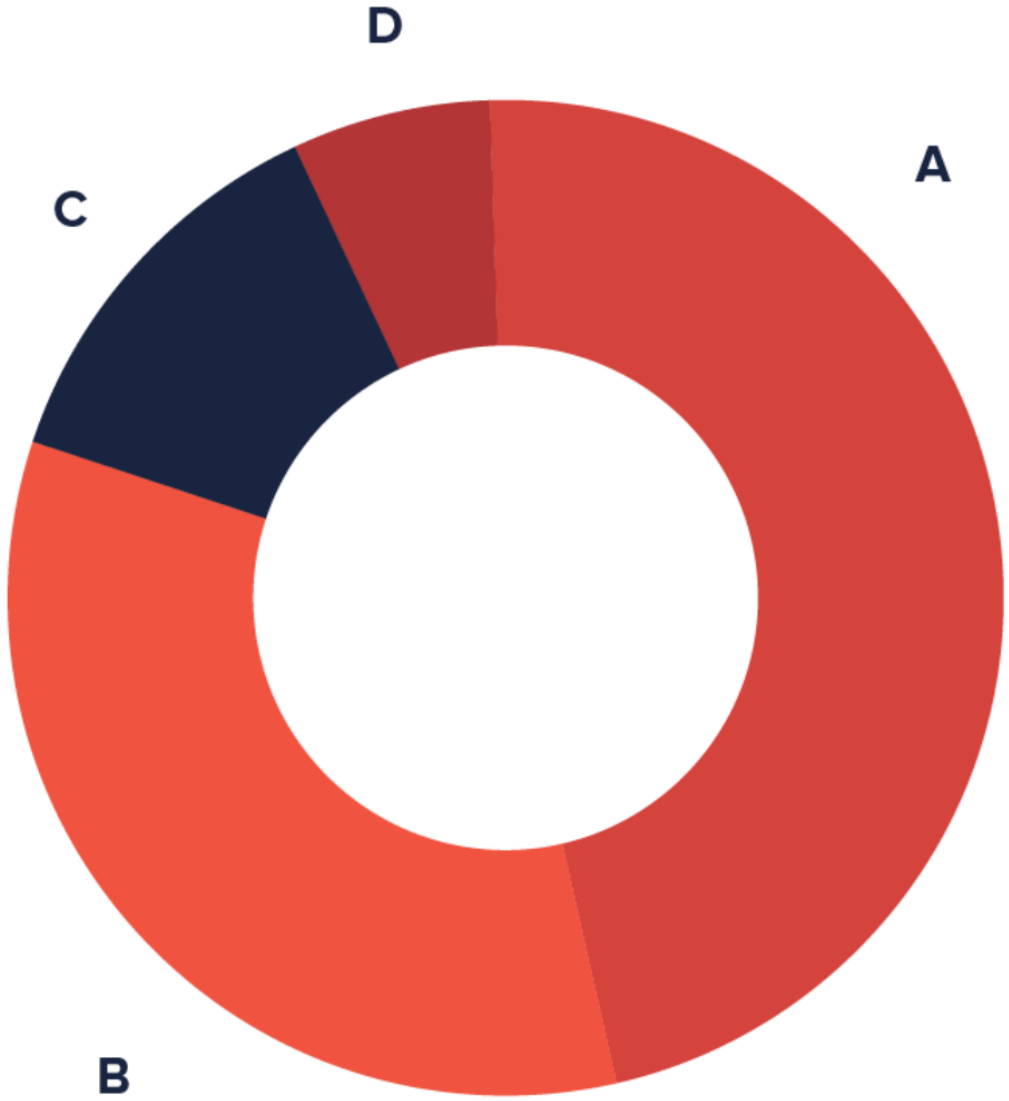


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

- Mosaic plots assume that the orthogonal variables identify all of the proportions
- Treemaps do not require that

Donut chart equivalent to piechart

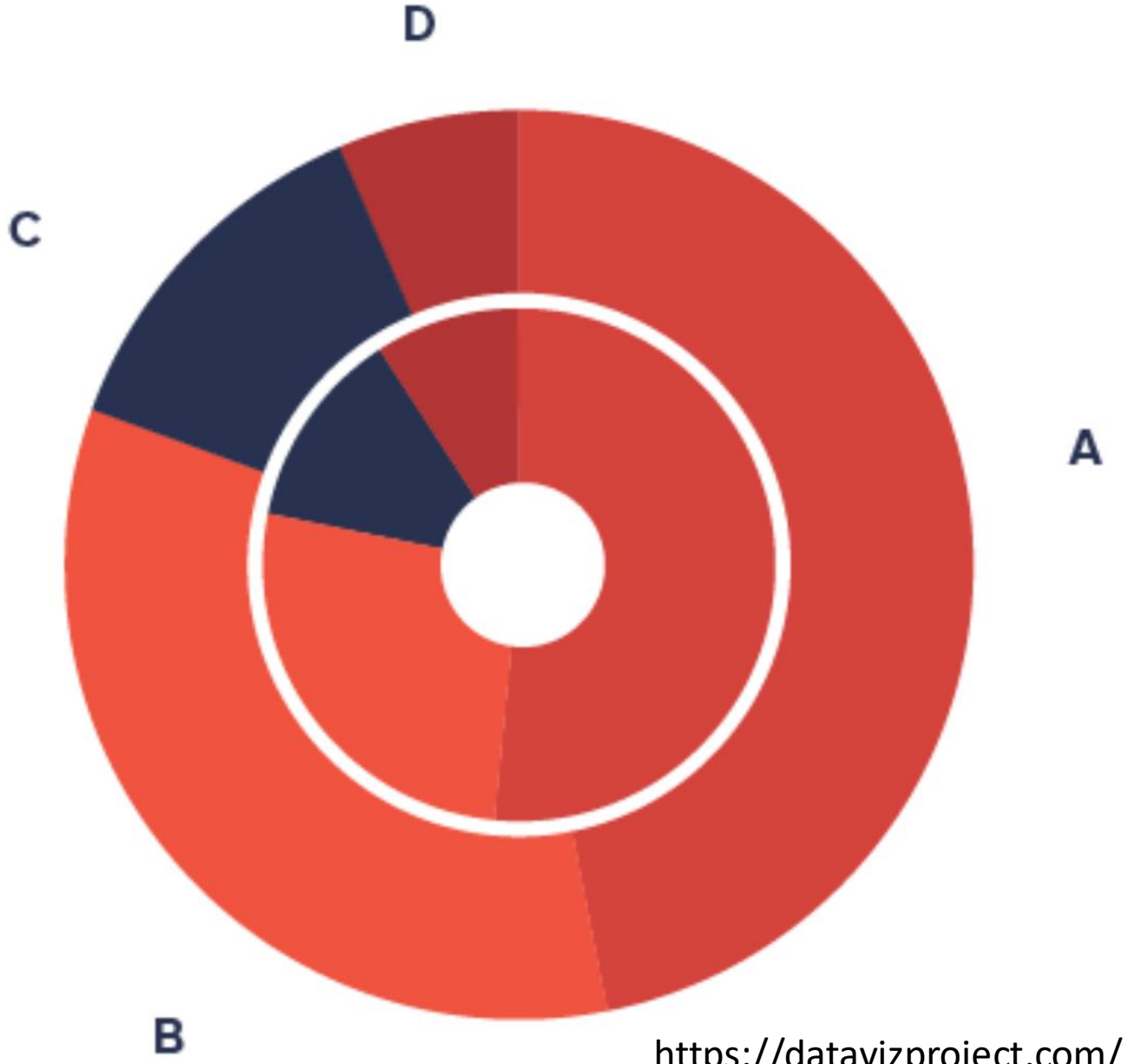
- Identical function to a pie chart
- Provide a better data intensity ratio because of the blank center?



https://datavizproject.com/

Multi-layer donut chart for multiple proportion

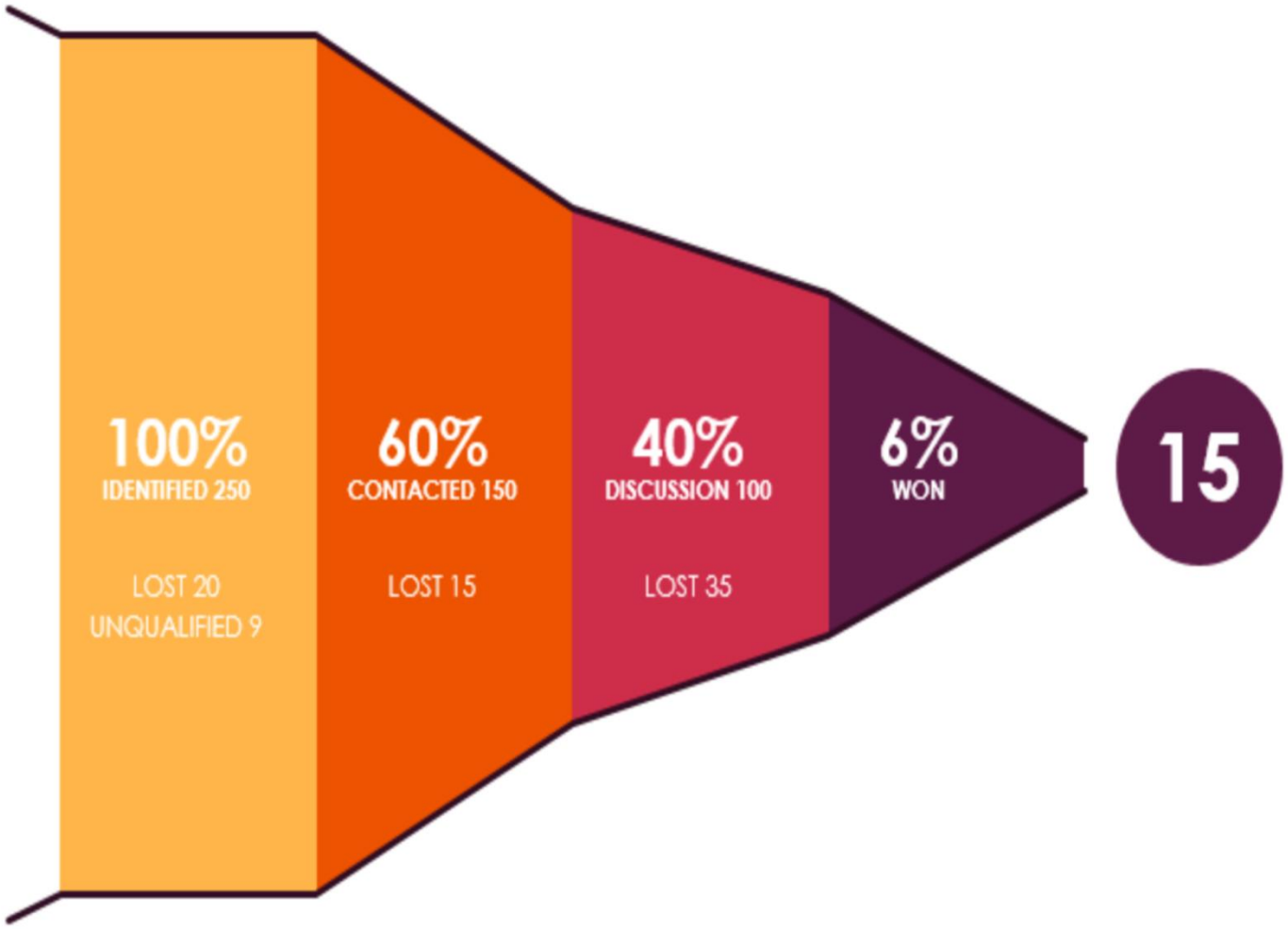
- Concentric circles each representing proportion
- Good for comparing across the circles?
- Other limitations?



https://datavizproject.com/

Funnel chart for show proportion during a process

- Each slice represents a process filtering out data



https://exceldashboardschool.com/sales-funnel-chart/

Enrichment can be estimated by overlap

Is your gene set enriched in a predefined gene set, e. g., apoptosis?

➡ Does your gene set overlap with the gene set?

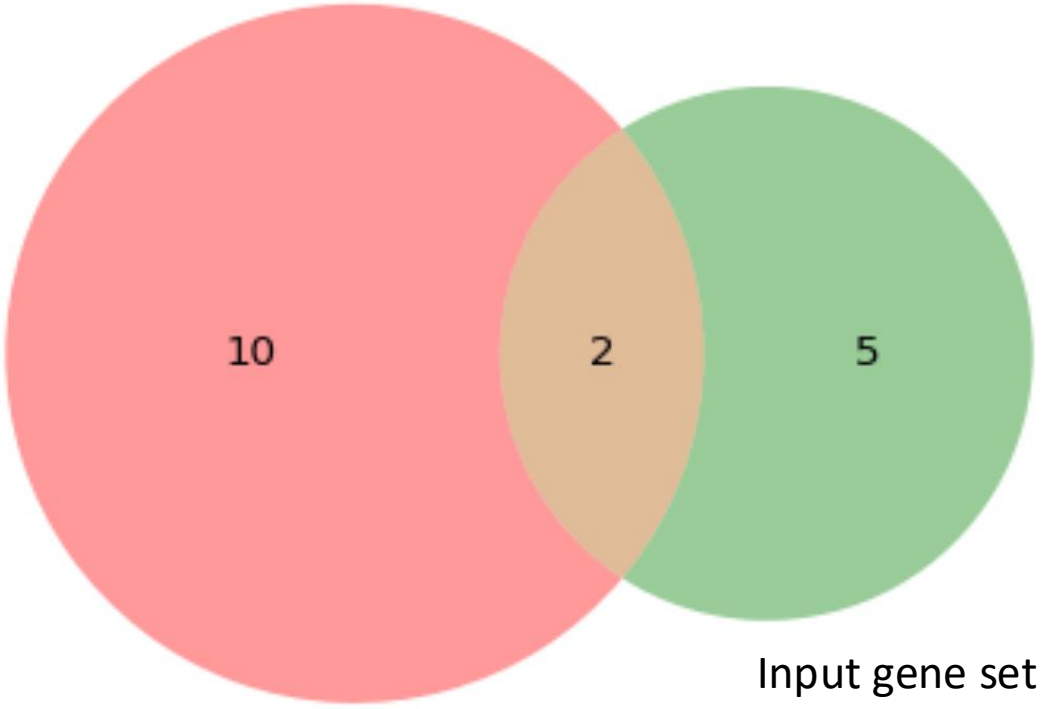
How enriched is your gene set in the gene set?

➡ How significant is the overlap?

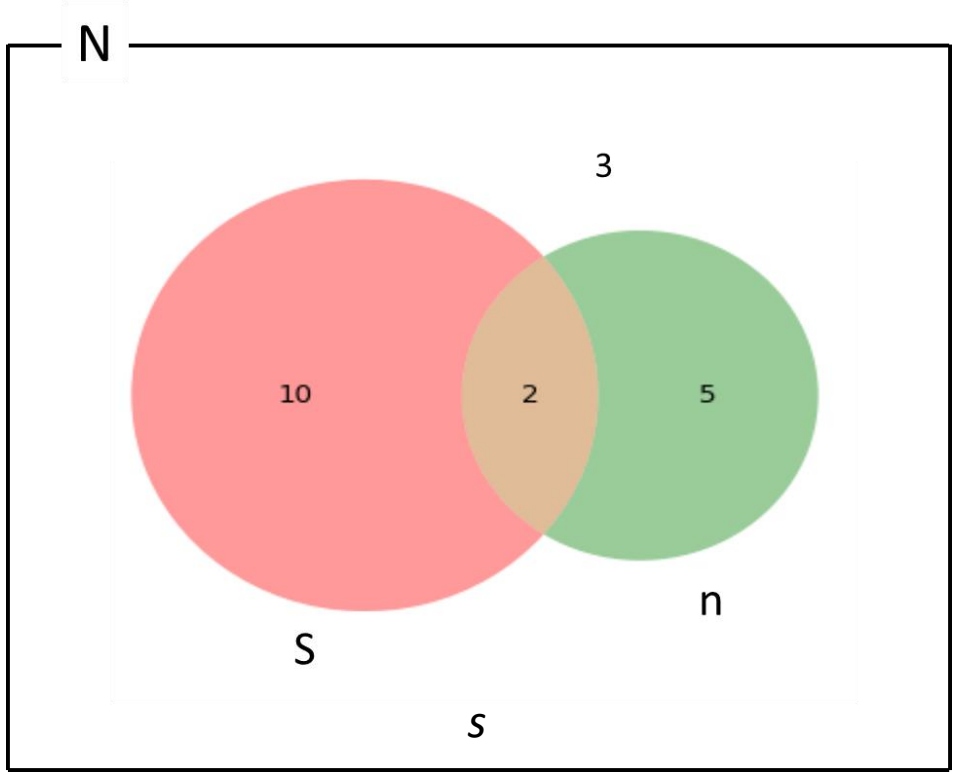
Note that enrichment can consider more than overlap to make more informative decisions (e.g., gene set enrichment analysis, GSEA)

Enrichment often presented by overlap

Venn diagram to show enrichment



Hypergeometric test to estimate significance of overlap



$$P(s) = \frac{{}_{S}C_{s} * {}_{N-S}C_{N-s}}{{}_{N}C_{n}}$$

when

S = successes from population

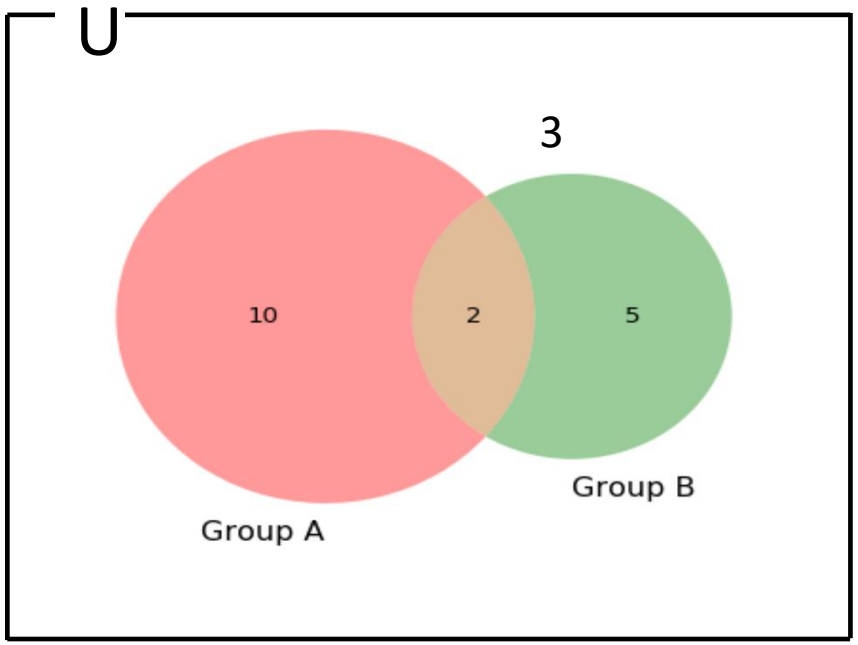
s = successes from sample

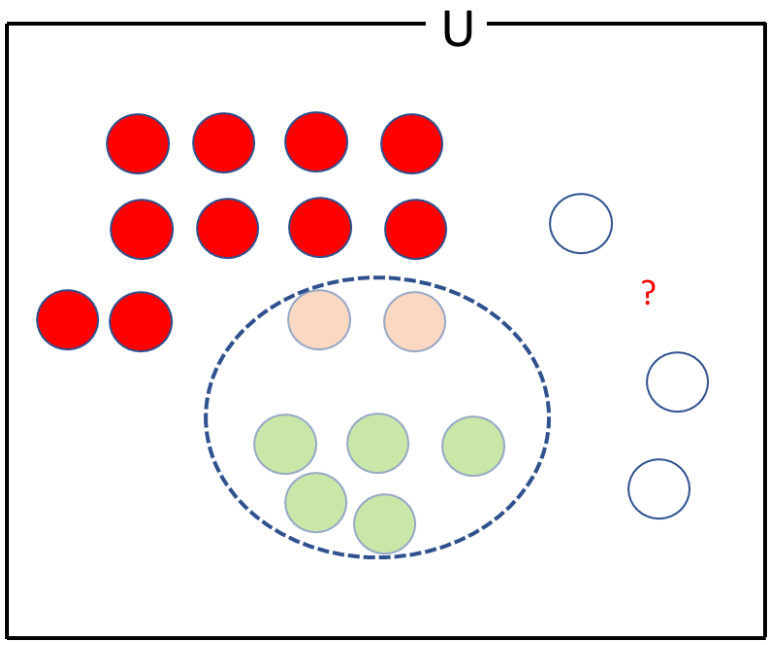
N = population size

n = sample size

- Experiment
  - Draw n observations without replacement from a total population of size N
  - S out of N are "successes"
  - What's the probability of drawing exactly s of the S "successes" in your sample of size n?
  - Or, what's the probability of exactly this much overlap between your sample and the "success" set, S ⊆ N?
- To estimate significance of the overlap, we need to determine the size of the universe (N)!

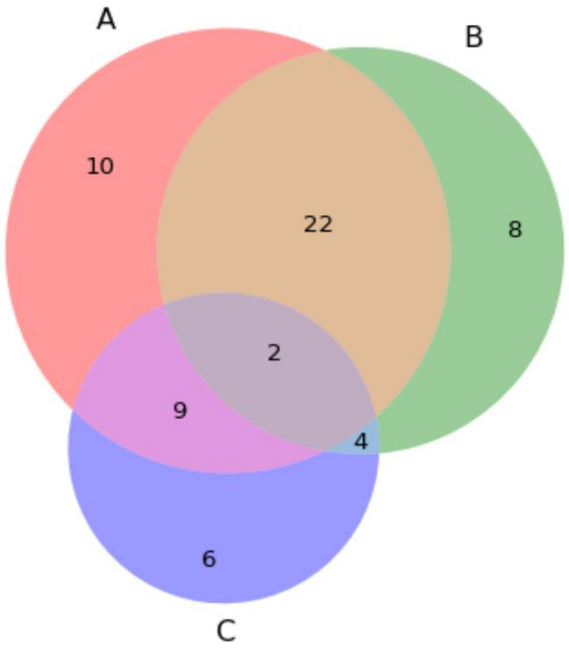
Hypergeometric test to estimate significance of overlap

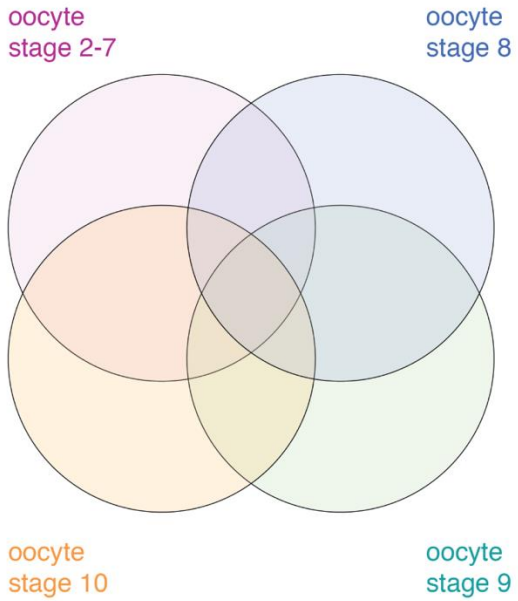




- To estimate significance of the overlap, we need to determine the size of the “universe”
- You can play with the size of the universe to make a favorable significance

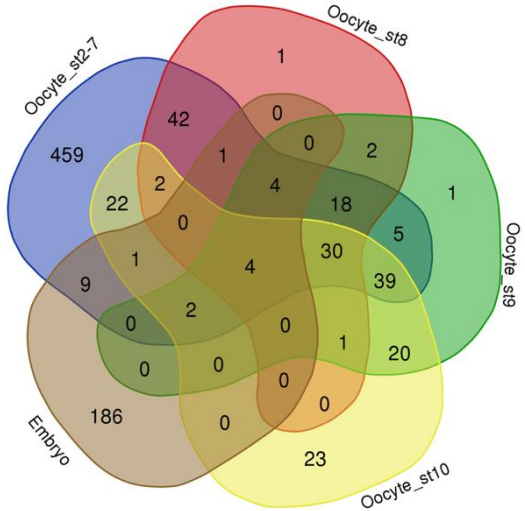
Venn diagram to show overlap for 3 or fewer sets

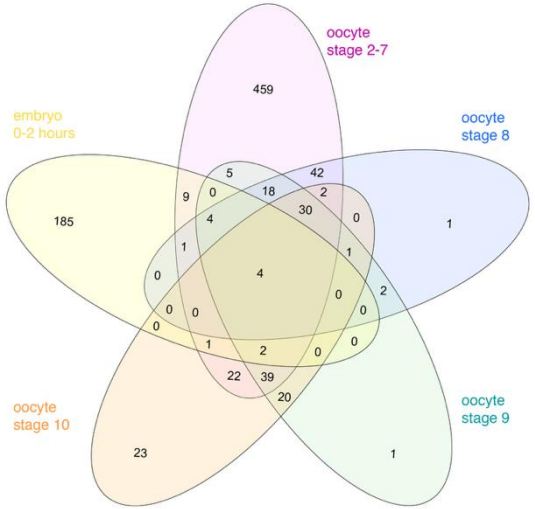


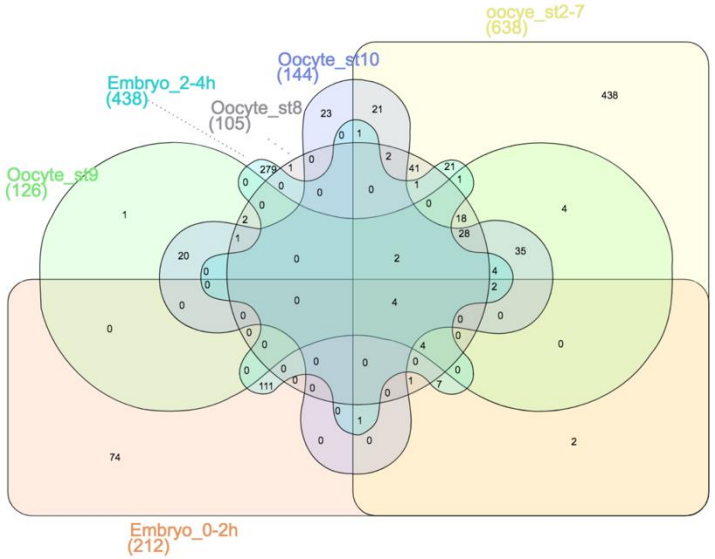


- Difficult to focus
- Not OK to show an interesting subset of overlaps
- Encouraged to use only for exploration
- What's "wrong" with the figure on the right?

Venn diagram to show overlap for ≥ three sets

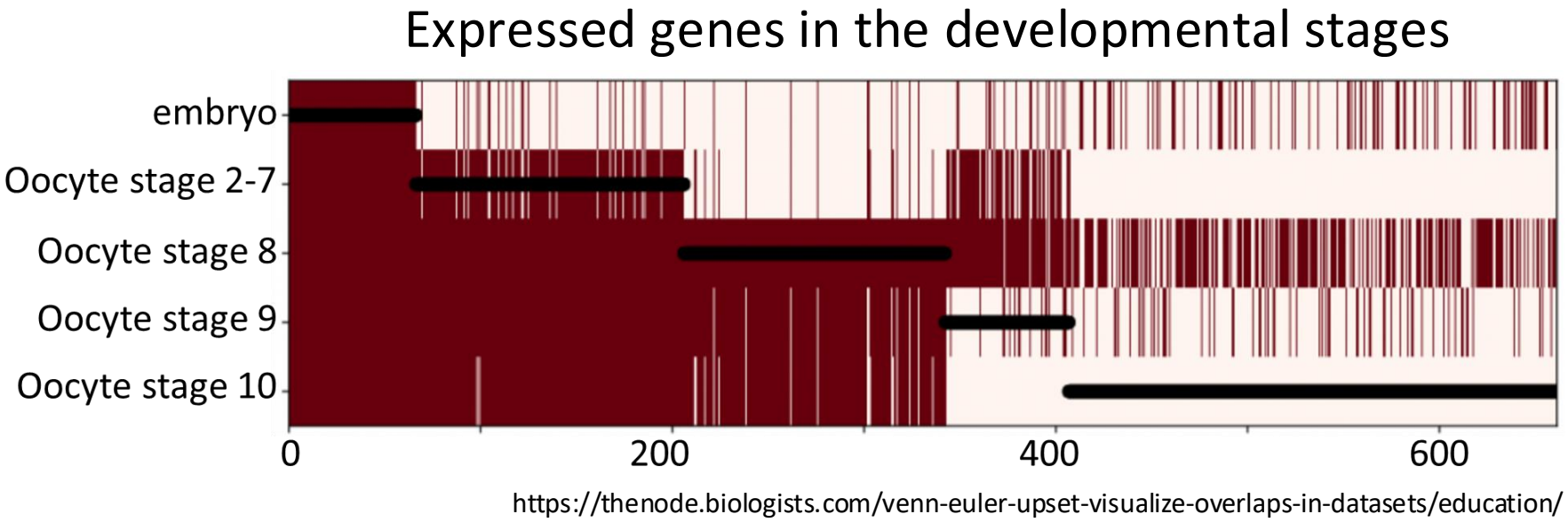






- Web-based: Draw Venn (Yves Vandepeer, Univ of Gent) and InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams (Heberle et al. BMC Bioinformatics, 2015)
- R: nVennR: (Victor Quesada) – seems to be gone from CRAN
- Try plotVenn

Heatmaps to draw overlap among > three sets

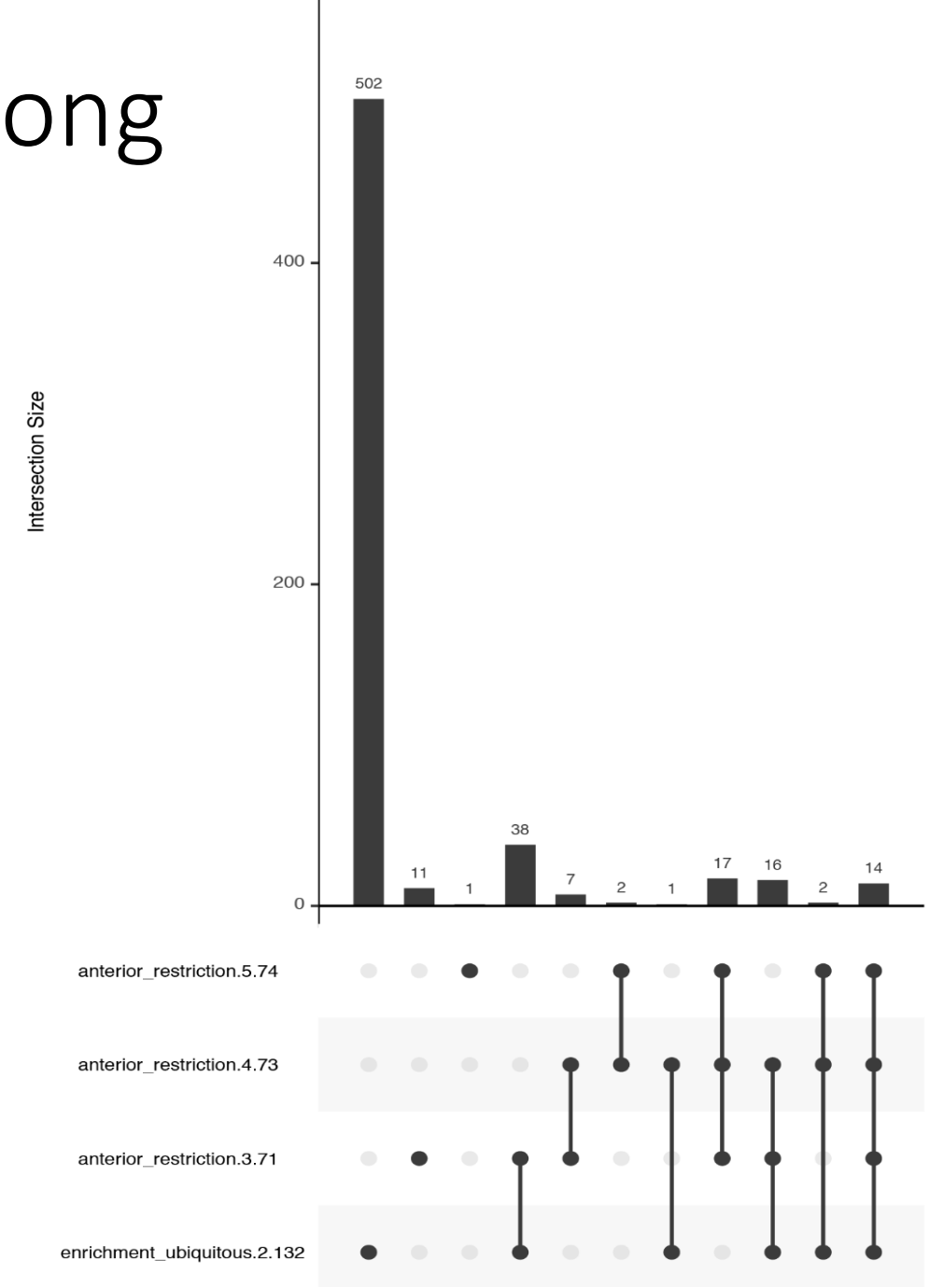


https://thenode.biologists.com/venn-euler-upset-visualize-overlaps-in-datasets/education/

- Pros: Easy to check the overall pattern
- Cons: Difficult to see amount of the overlaps

Upset plots to show overlaps among > three sets

- Presence of dataset elements in a given intersection is shown in a table
- The size of the intersection is represented with a bar chart
- Available in R: UpSetR (Conway et al. Bioinformatics, 2017)

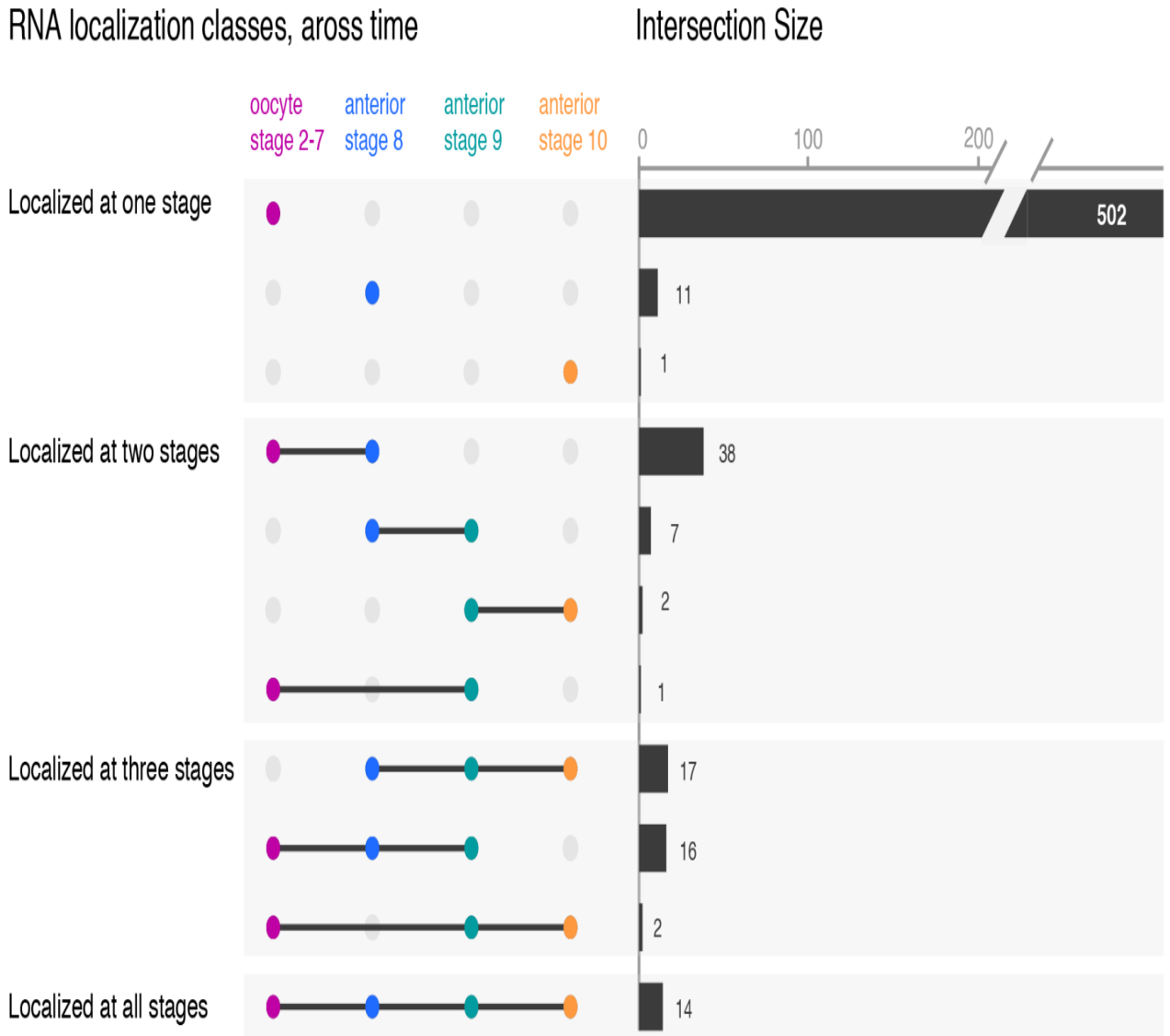


https://thenode.biologists.com/venn-euler-upset-visualize-overlaps-in-datasets/education/

Upset plots to show overlaps among > three sets

Possible improvements:

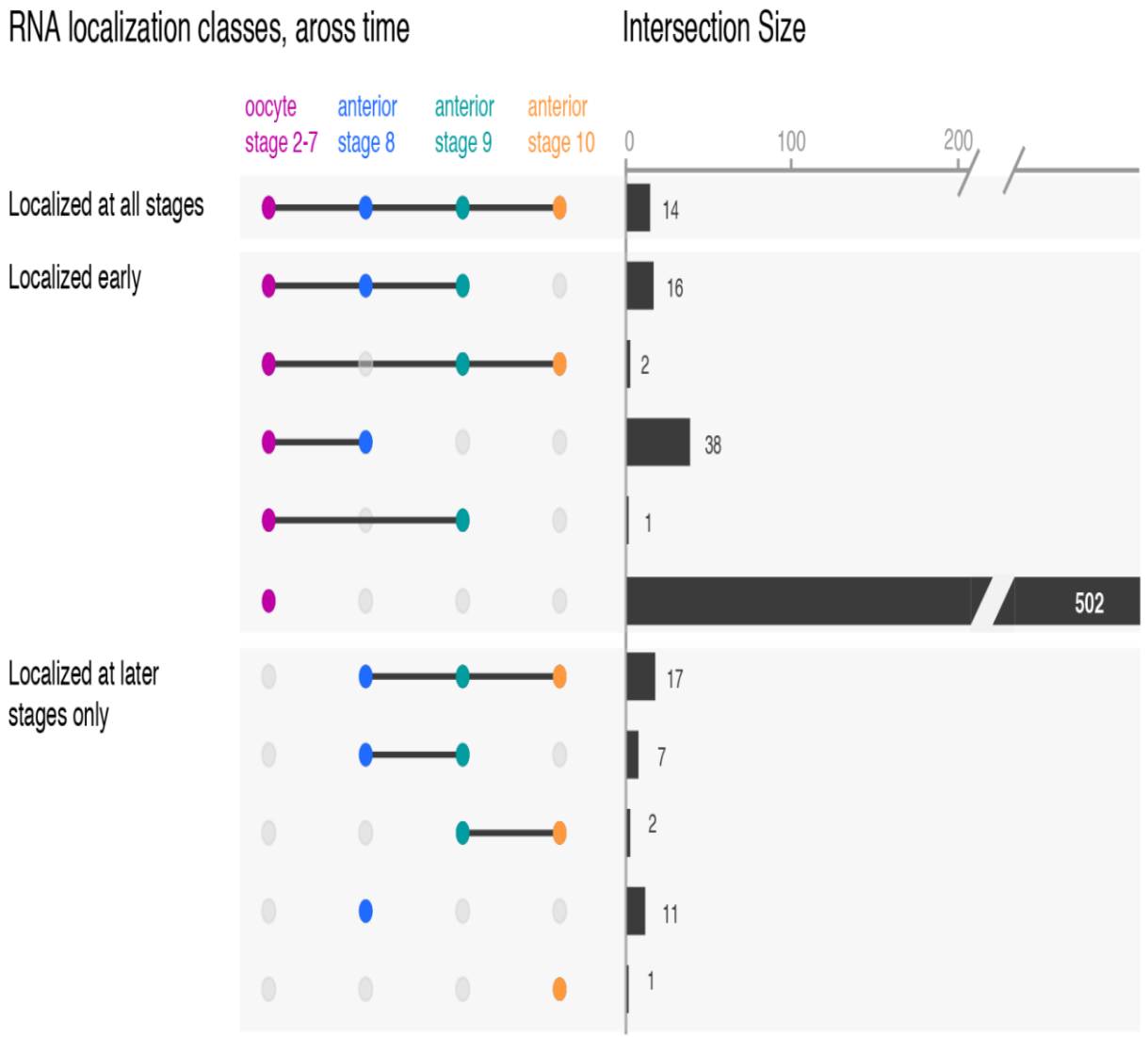
1. Cutting the y-axis
2. Tilting the plot.
3. Group by the number of intersection
4. Color-code



https://thenode.biologists.com/venn-euler-upset-visualize-overlaps-in-datasets/education/

Upset plots to show overlaps among > three sets

- The sorting order will reflect your point
- e.g. RNAs localized early vs. late stages only
- More suitable for exploration



https://thenode.biologists.com/venn-euler-upset-visualize-overlaps-in-datasets/education/