Data visualization

Tools to analyze data and communicate knowledge

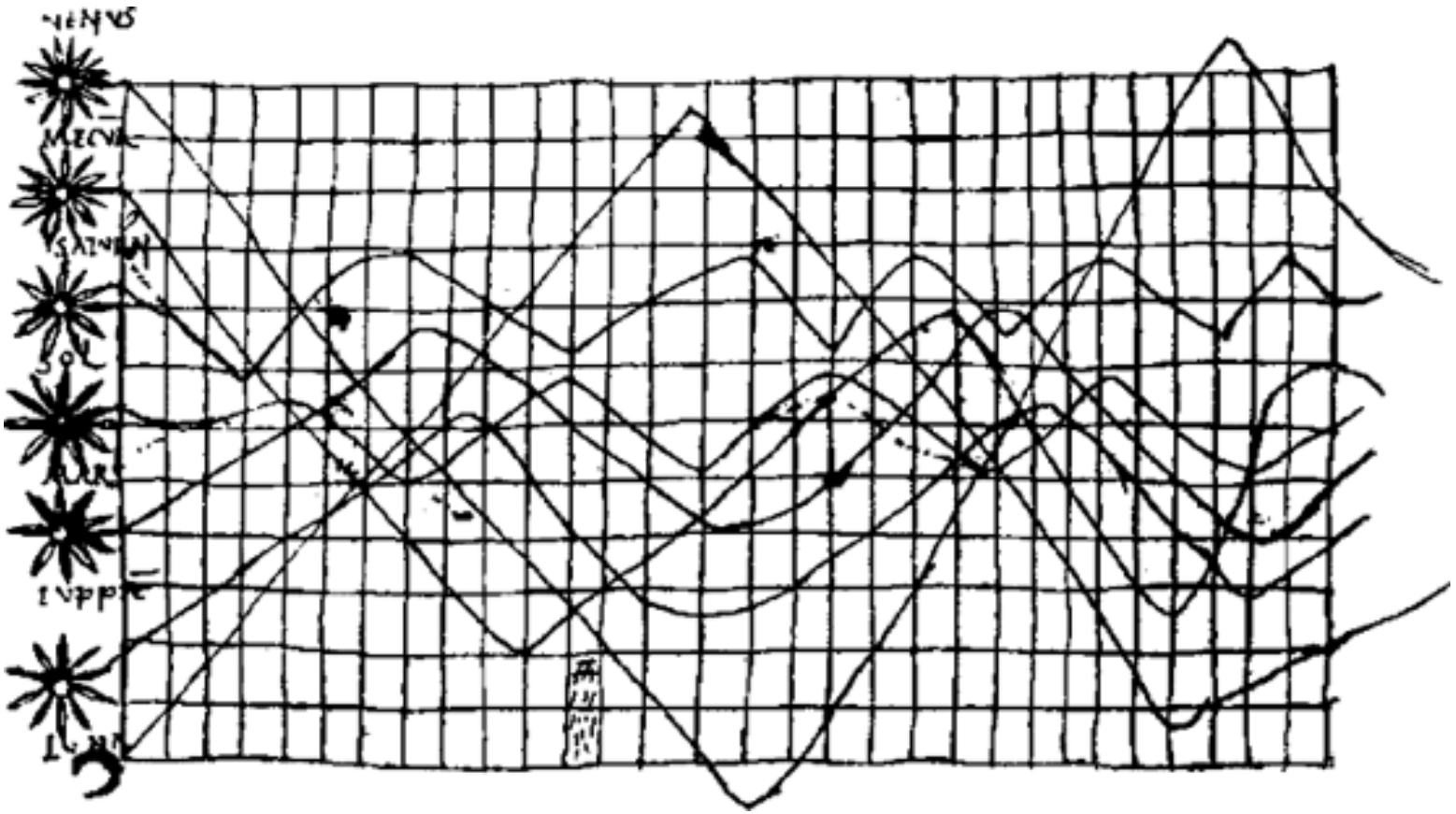
F. Roscioli[1]

[1] *University of Rome Tor Vergata*

29th April 2023

Outline

Historical Background

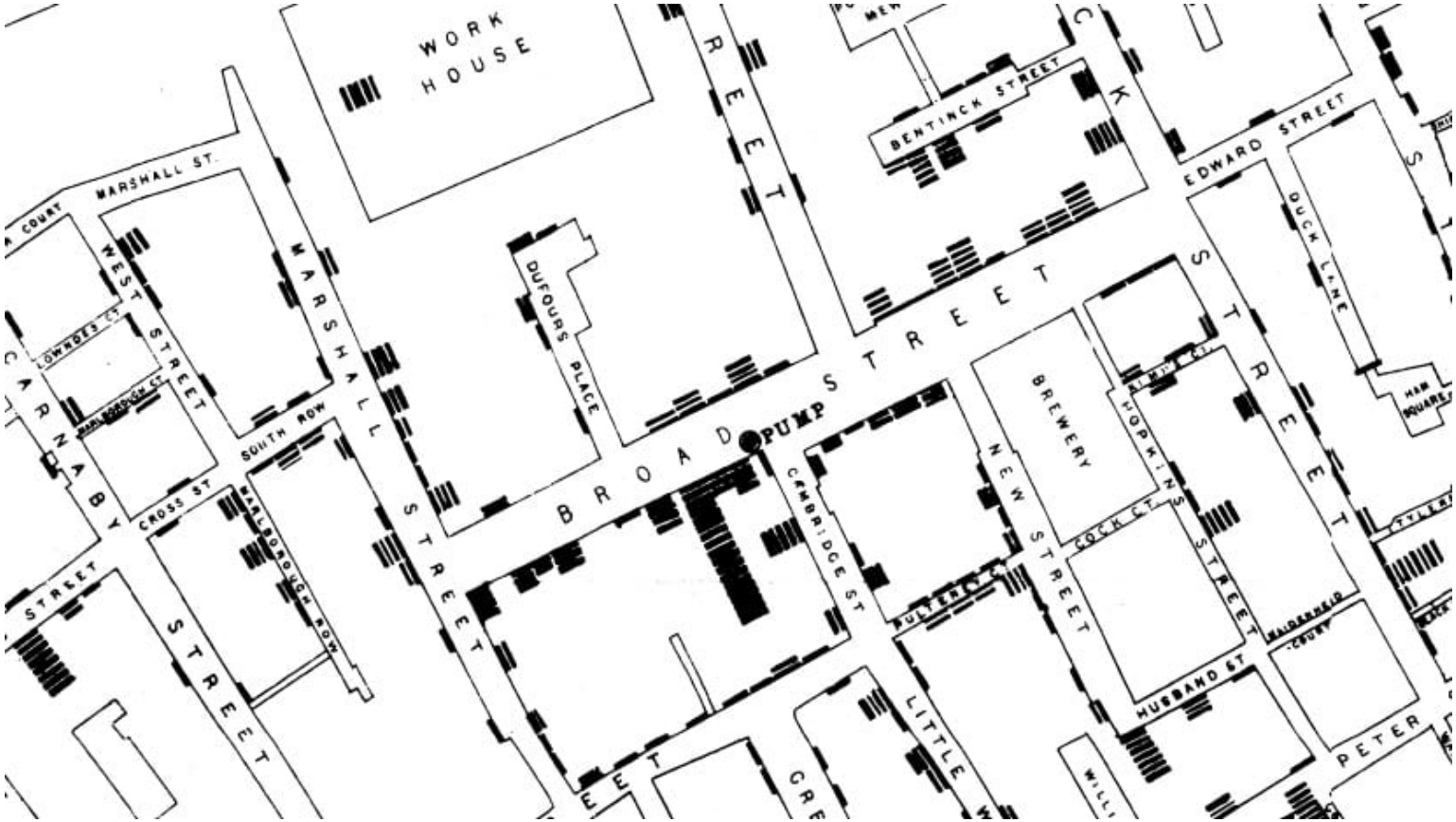


Credit: *Funkhouser (1936, p. 261)*

Figure: Planetary movements shown as cyclic inclinations over time, by an unknown astronomer, appearing in a 10th century appendix to commentaries by A.T. Macrobius on Cicero's In Somnium Sciponis.

(Friendly, 2006)

Historical Background



Credit: *Snow (1855)*

Figure: Deaths from cholera clustered around the Broad Street pump in London.

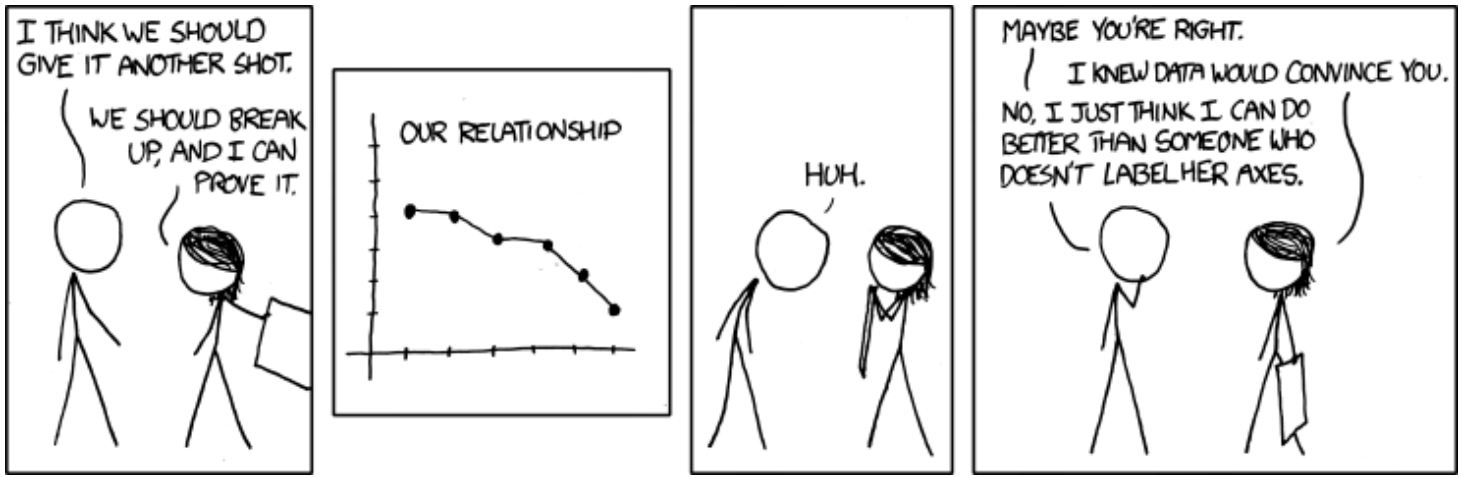
(Friendly, 2006)

Historical Background

- In statistical graphics, all modern forms of data display were invented during the **first half of the 19th century**: bar and pie charts, histograms, line graphs and time series plots, contour plots, scatterplots, etc.
- In October 1831, the first case of Asiatic cholera occurred in Great Britain, and over 52000 people died in the epidemic that ensued over the next 18 months or so (Gilbert, 1958). Subsequent cholera epidemics in 1848-1949 and 1853-1954 produced similarly high death tolls, but the waterborne cause of the disease was unknown until 1855 when **Dr John Snow** produced his famous dot map (Snow, 1855) showing **deaths due to cholera clustered around the Broad Street pump in London**.
- During the last quarter of the twentieth century, data visualization blossomed into a mature, vibrant, and multidisciplinary research area, and software tools for a wide range of visualization methods and data types are available for every computer.

(Friendly, 2006)

Motivation



Credit: *dotTech*

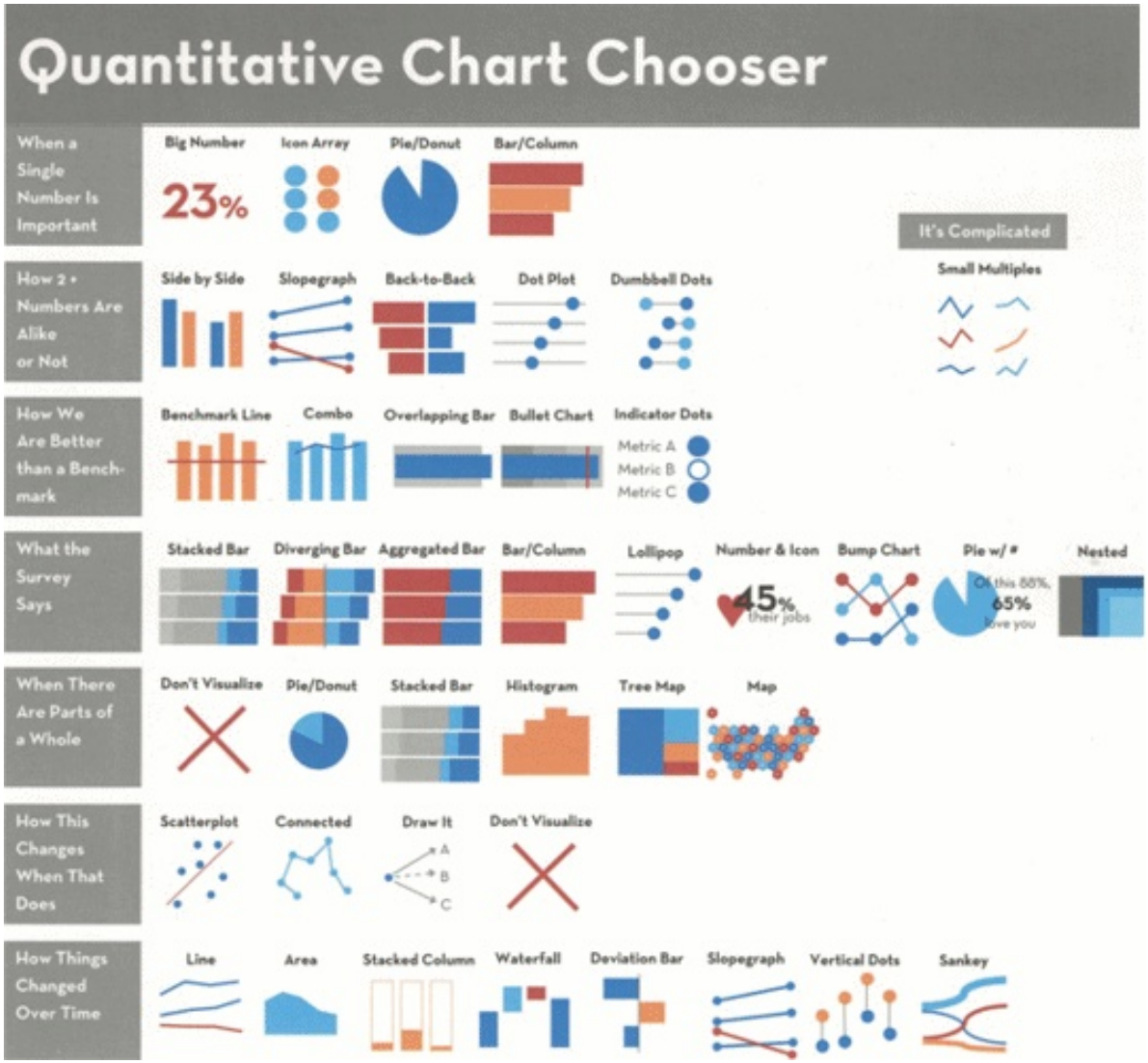
- Research tells us that data are more persuasive when shown in graphs. One factor may be that **we are primarily visual beings and that most of us, most of the time, are skimming the narrative for things that pop-out at us and catch our attention.** Data visualization does just that: it provides pop. Graphs and formulas seem to add credibility to the data, even if they do not contain any new insights beyond what already exists in the narrative.

(Evergreen, 2019)

Key questions

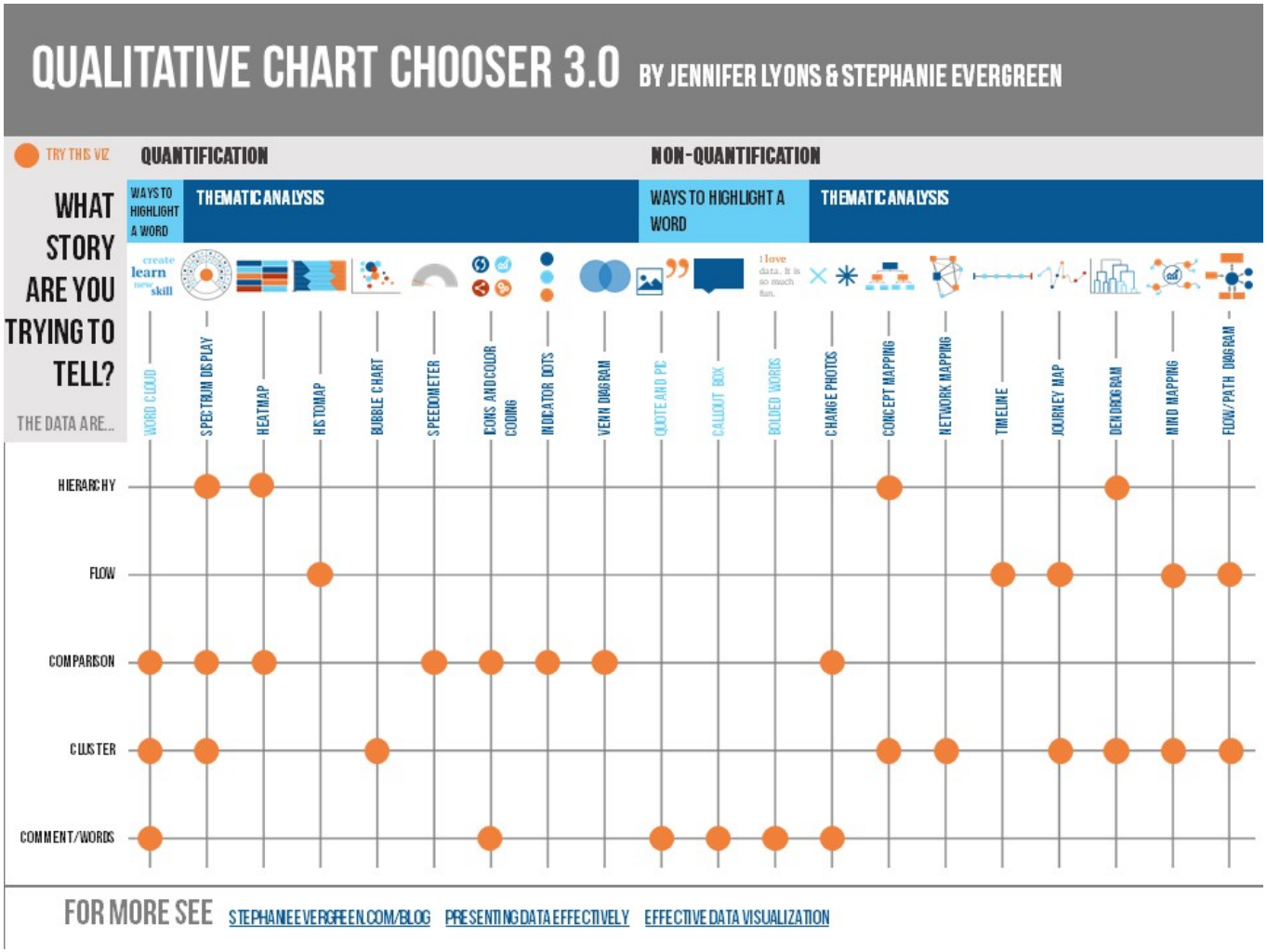
- Before plotting any data, a good researcher must ask herself:
  1. **Why** do I put graphs and tables in my publication?
     To spread the message.
  2. **What** do I want to communicate? What is my point?
     A trend? A comparison? The results of a test?
  3. **Who** is my audience? In presenting the results to others, the presenter assumes that his audience is ignorant. With this said, the audience also depends on the output in which our data will be included.

What is your point?



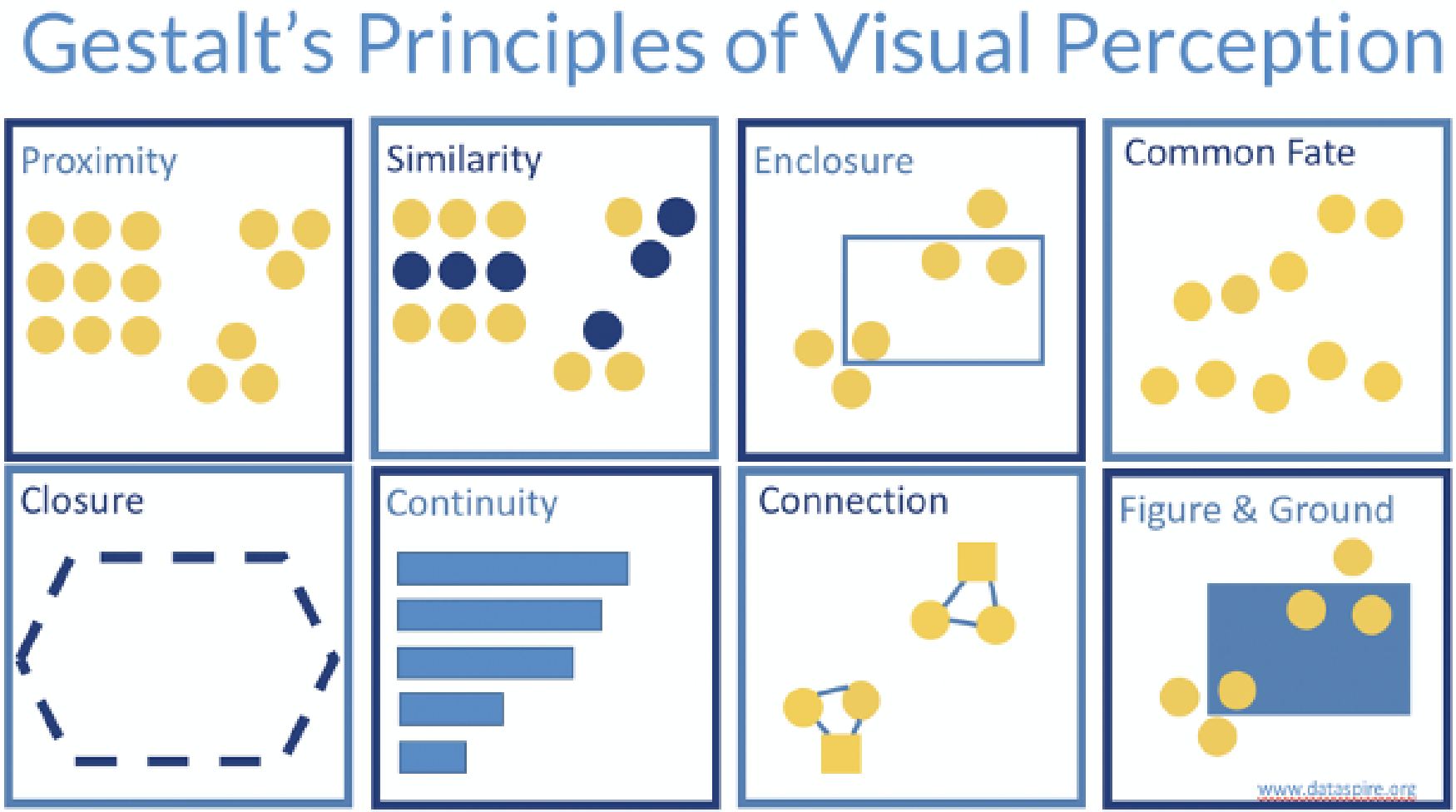
(Evergreen, 2019)

What is your point?



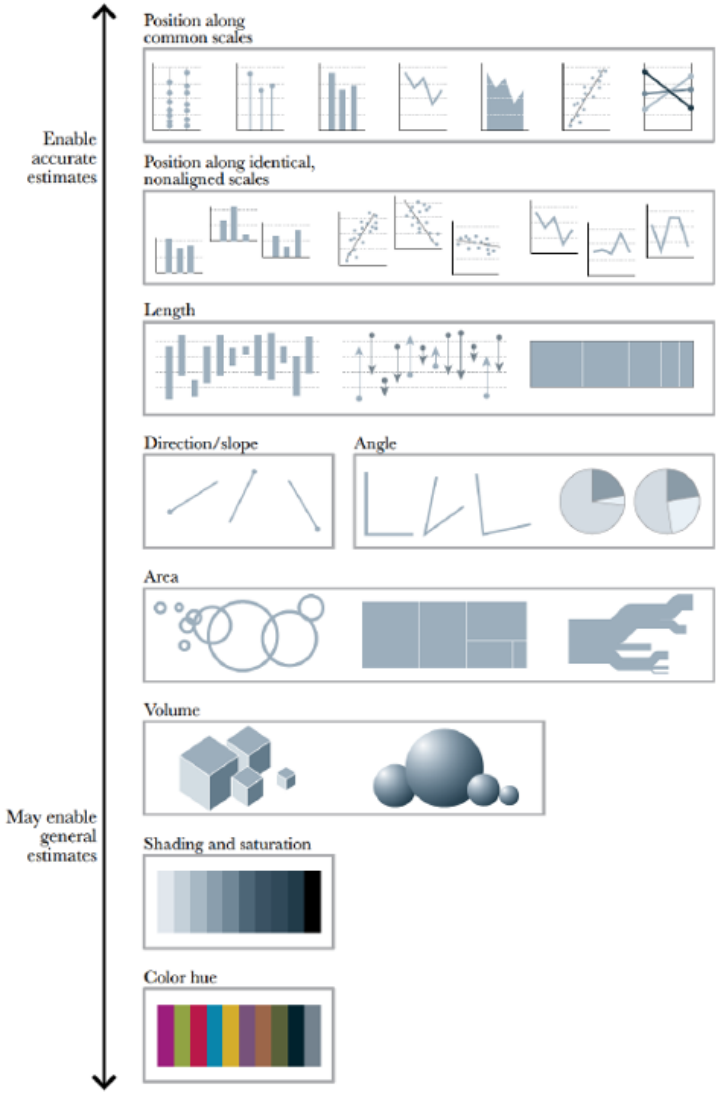
(Evergreen, 2019)

Who is your audience?



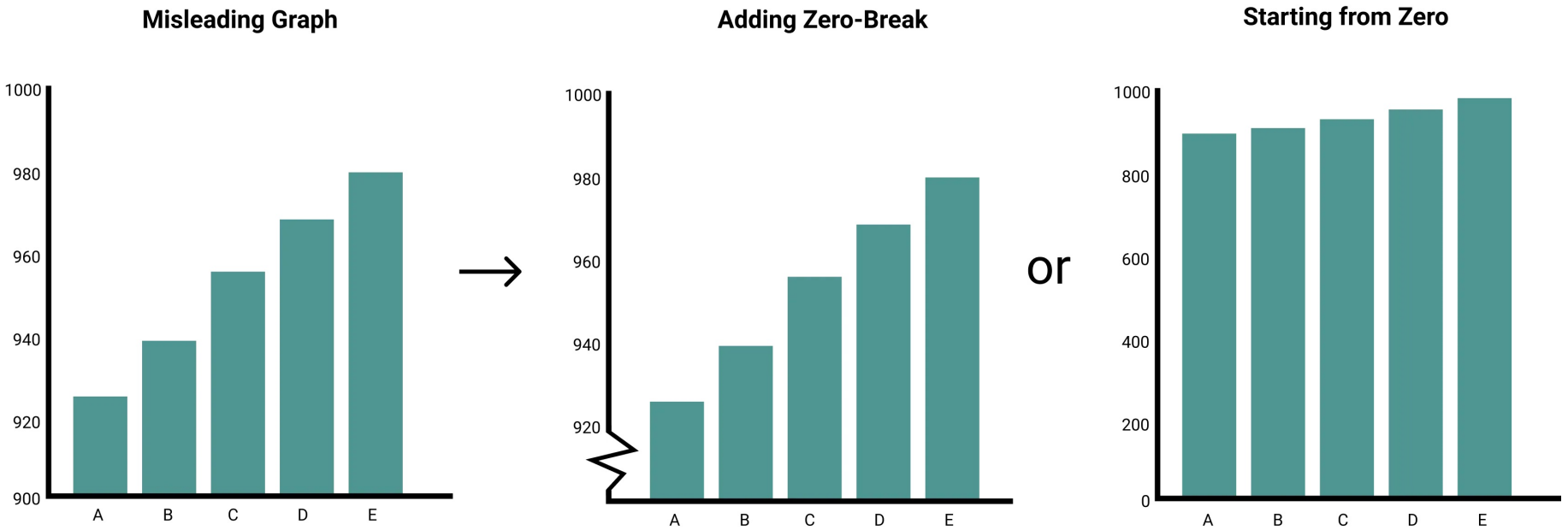
(Todorovic, 2008)

Who is your audience?



(Cleveland & McGill, 198

Who is your audience?



Credit: *Grootendorst*

Who is your audience?

Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Find the highest values.

(Schwabish, 2021)

Who is your audience?

Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	84
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	95	13	64
Jon	83	62	46	48
Karen	30	65	98	82
Ken	38	28	45	71
Lauren	98	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

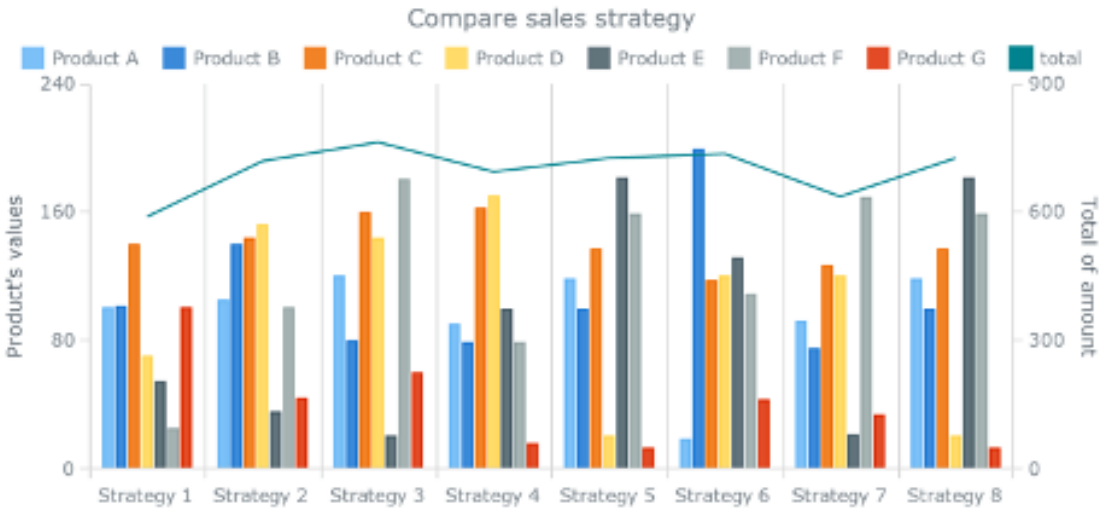
Table 1. Our sales grew to $600 million this year

	Q1	Q2	Q3	Q4
Bob	26	35	72	**84**
Ellie	22	15	61	35
Gerrie	19	20	71	55
Jack	22	**95**	13	64
Jon	83	62	46	48
Karen	30	65	**98**	82
Ken	38	28	45	71
Lauren	**98**	81	41	63
Steve	16	50	23	41
Valerie	46	24	30	57
Total	**$400**	**$475**	**$500**	**$600**

Find the highest values.

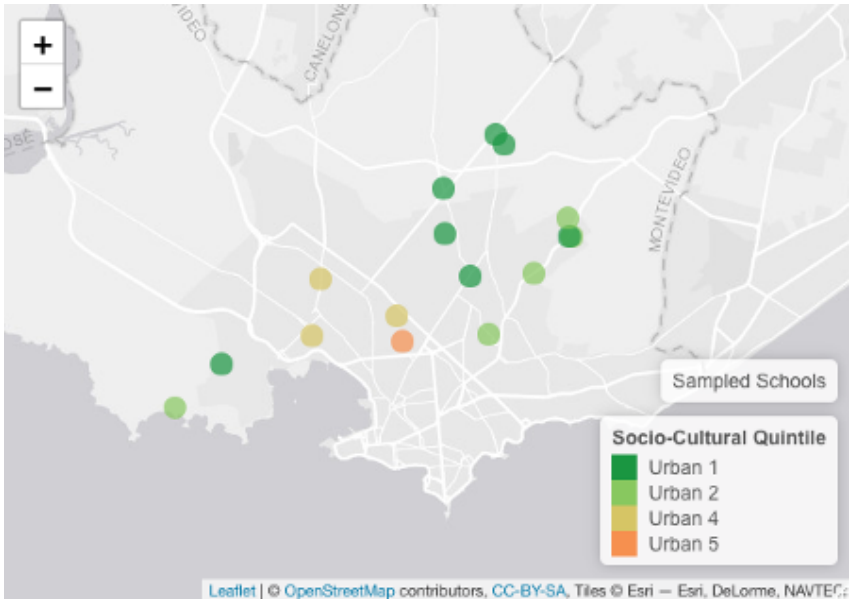
(Schwabish, 2021)

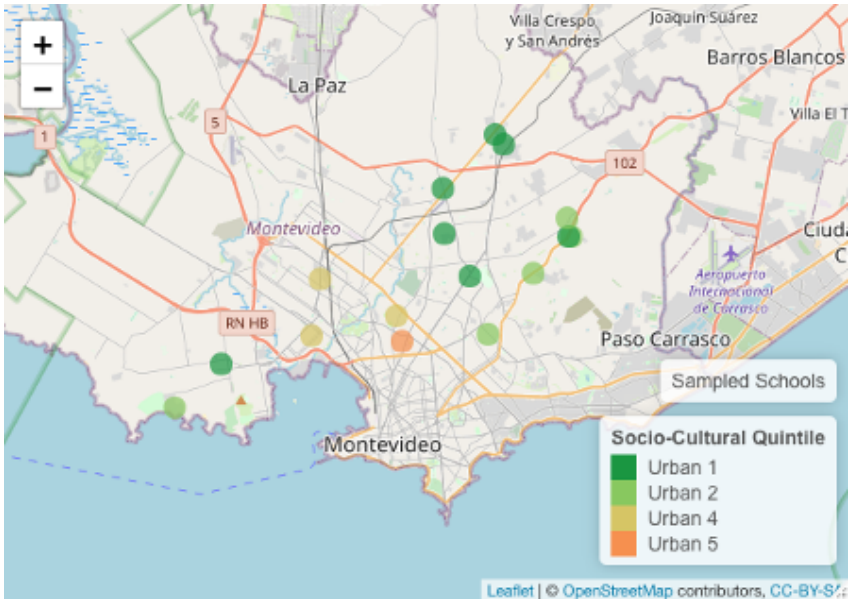
Who is your audience?



- To communicate efficiently and keep the audience interested, we need to use "messages" that are easily understandable. Thus, **visually known** and **easily readable** by the audience.

Who is your audience?





Who is your audience?

	◦Critics 0 Gr=12	◦Critics 1 Gr=13	◦Good atmosphere 0 Gr=14	◦Good atmosphere 1 Gr=14	◦Motivation 0 Gr=11	◦Motivation 1 Gr=18	◦Not curricular 0 Gr=25	◦Not curricular 1 Gr=31
◦Q1 Gr=64	0	0	12	14	2	5	2	0
◦Q2 Gr=64	0	0	0	0	0	0	22	29
◦Q3 Gr=64	2	0	2	0	0	0	1	1
◦Q4 Gr=64	10	13	0	0	9	13	0	1

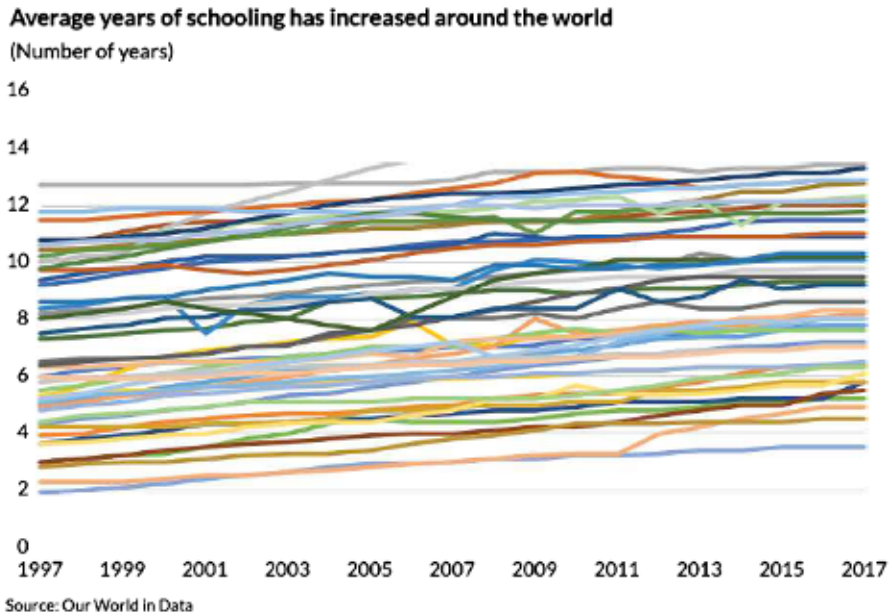
	Bad Atmosphere non PHCE Gr=21	Bad Atmosphere PHCE Gr=11	Behavior non PHCE Gr=72	Behavior PHCE Gr=120	Good Atmosphere non PHCE Gr=24	Good Atmosphere PHCE Gr=51	Orchard Gr=22	Others Gr=1083	Science non PHCE Gr=35	Science PHCE Gr=51	Study non PHCE Gr=181	Study PHCE Gr=191	Work non PHCE Gr=98	Work PHCE Gr=79
Q1 Gr=319	17	4	28	55	20	46	0	287	0	0	43	57	9	3
Q2 Gr=319	2	3	21	32	0	2	6	286	35	51	59	48	3	3
Q3 Gr=319	2	4	23	33	4	3	0	191	0	0	79	86	86	73
Q4 Gr=319	0	0	0	0	0	0	16	319	0	0	0	0	0	0

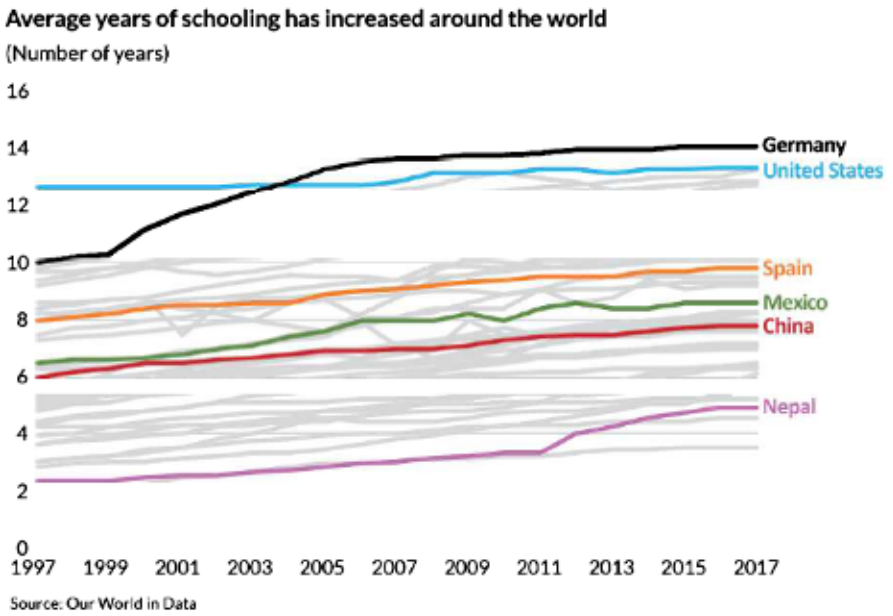
The perfect graph

- Answer What, Why, and Who (see previous section).
- Show the data.
- Reduce disorder.
- Integrate graphic and text.
- Avoid the spaghetti chart.
- Start with grey.

(Schwabish, 2021)

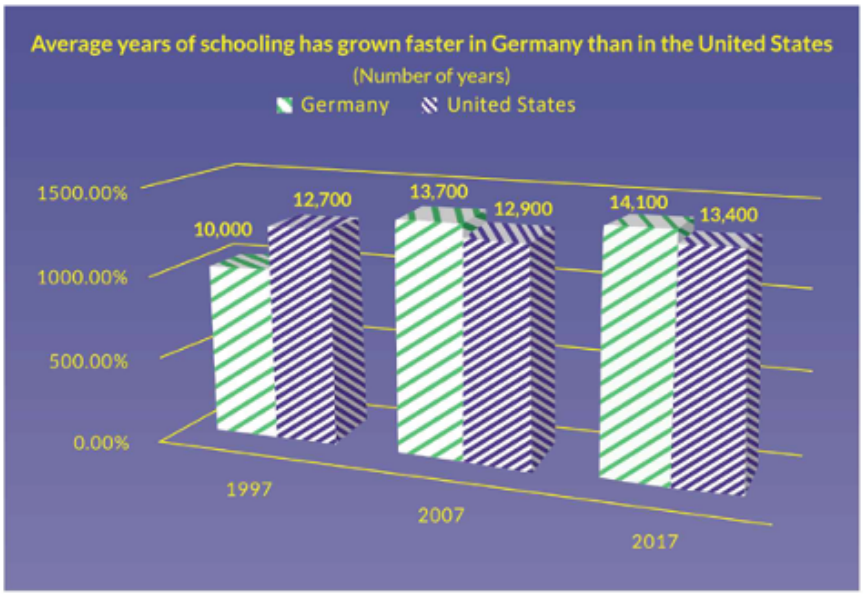
Show the data.

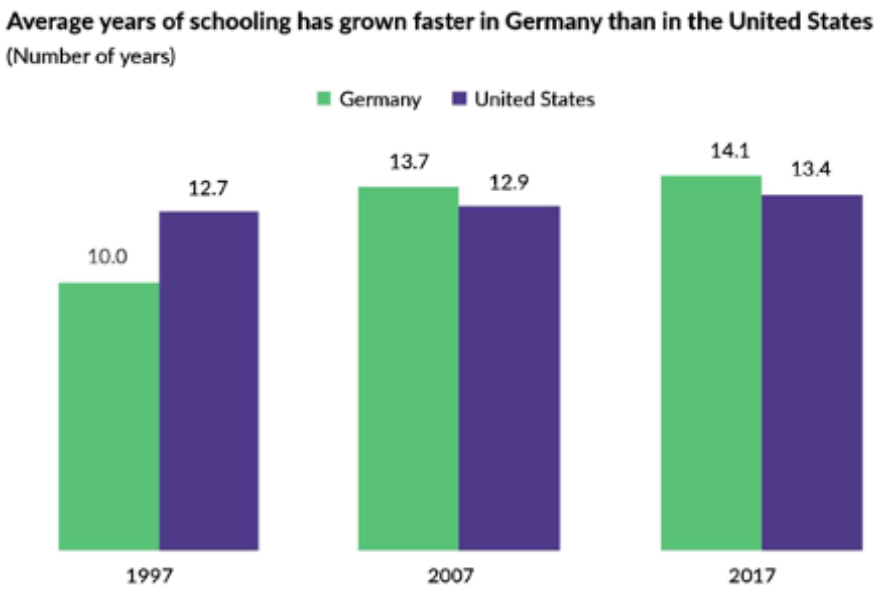




(Schwabish, 2021)

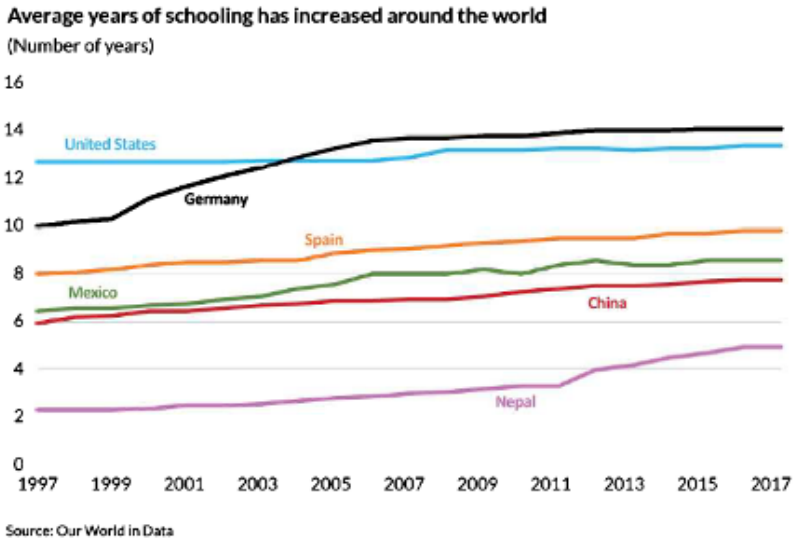
Reduce the disorder.

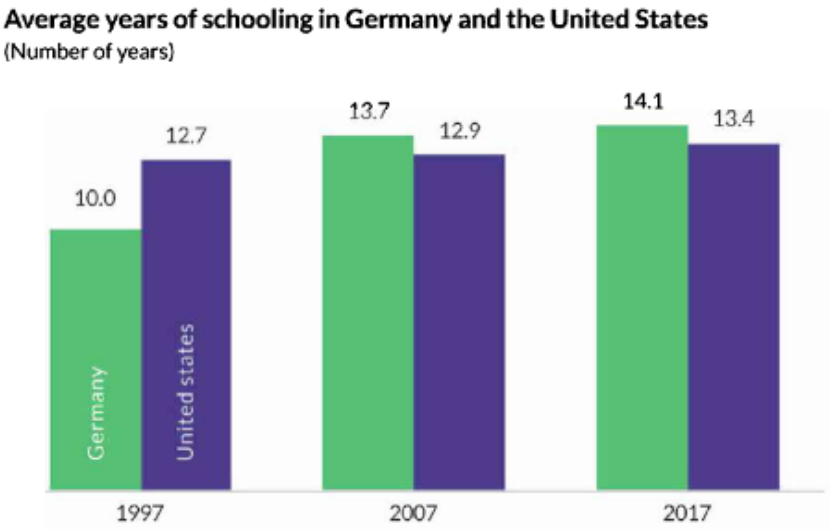




(Schwabish, 2021)

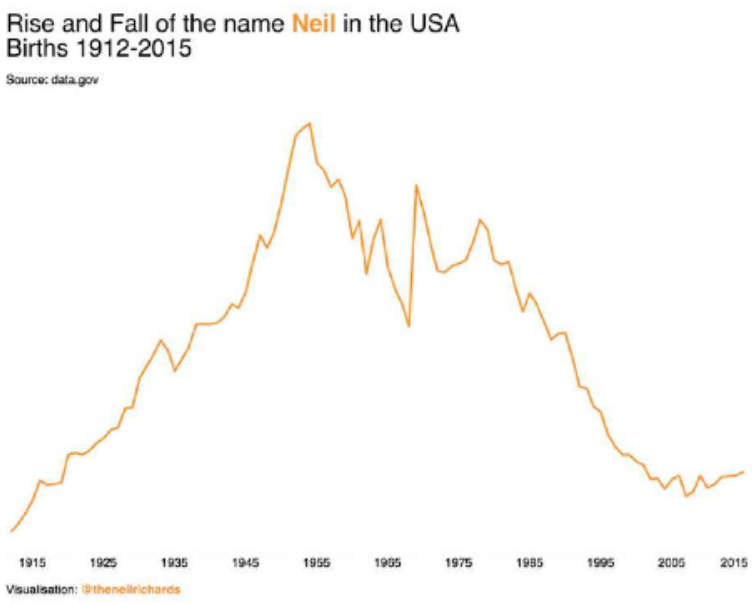
Integrate graphic and text.

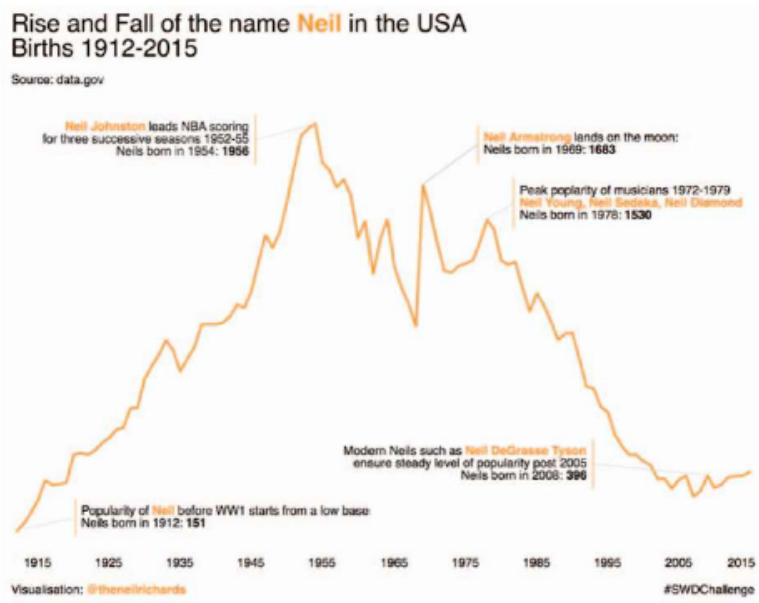




(Schwabish, 2021)

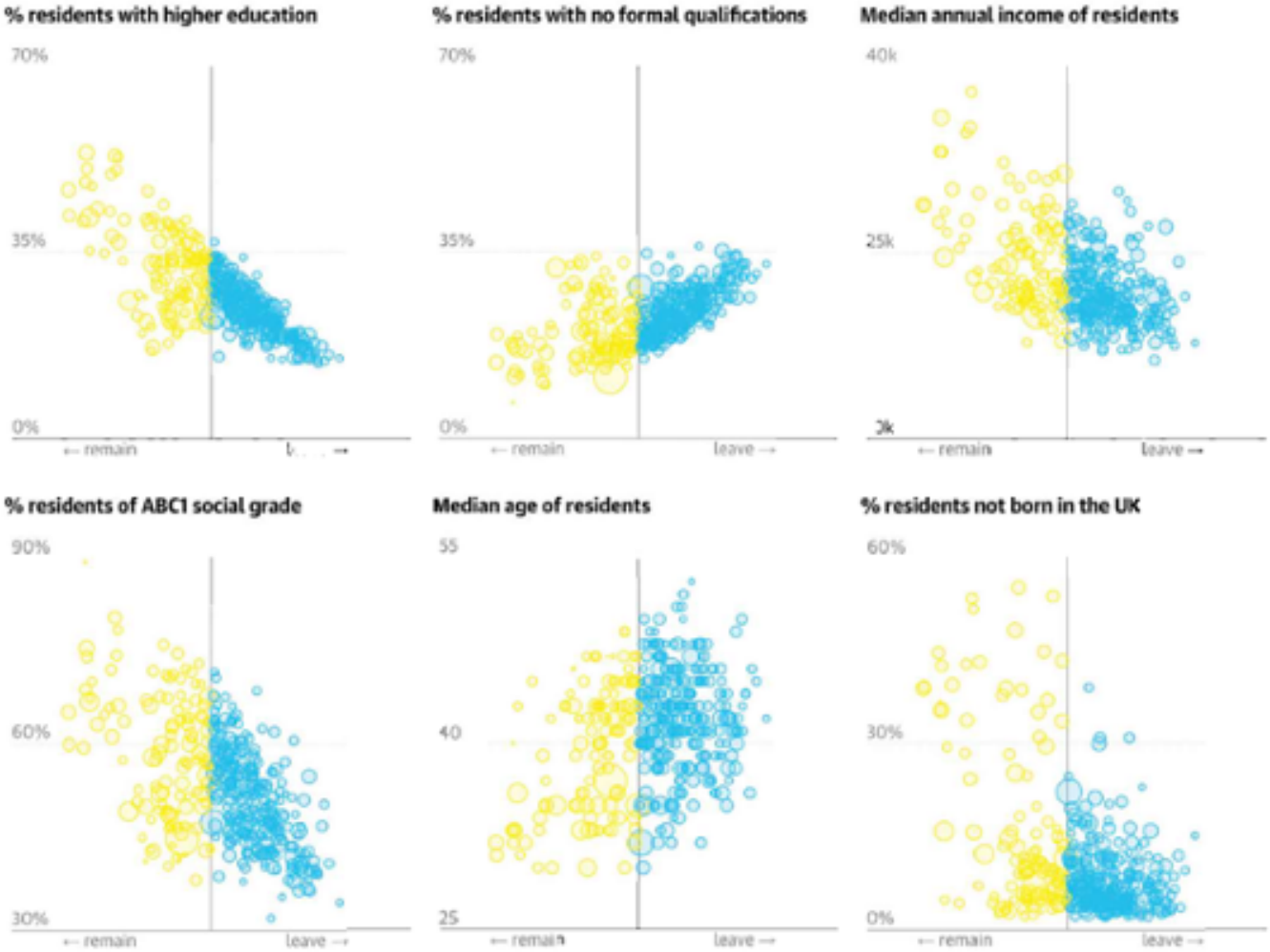
Integrate graphic and text.





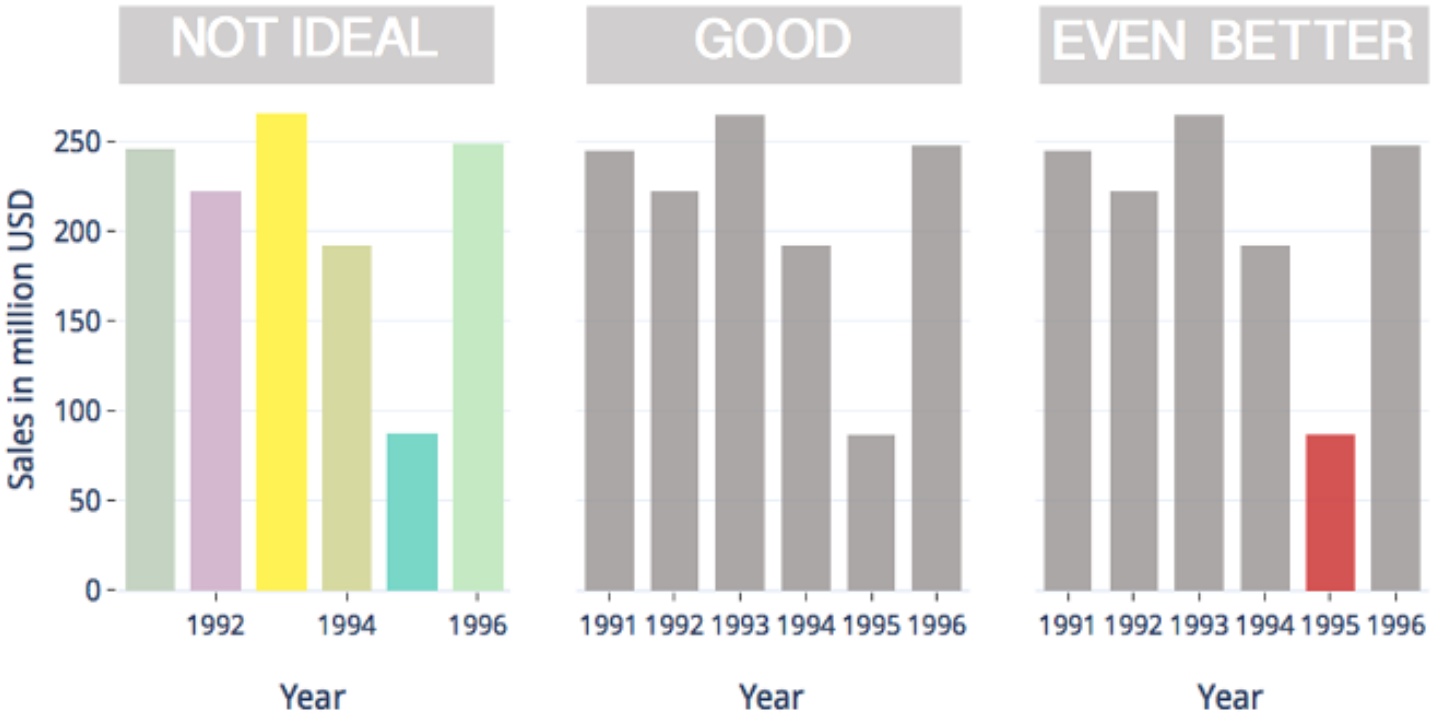
(Schwabish, 2021)

Avoid the spaghetti chart.



(Schwabish, 2021)

Start with grey.



(Schwabish, 2021)

To summarize

- The primary objective in data visualization is to gain insight into an information space. Therefore, **data, information, and message could serve as both the input and output of a visualization process**, raising questions about their exact role in visualization.
  - **Data**: Computerized representations of models and attributes of real or simulated entities.
  - **Information**: Data representing the results of a computational process, such as statistical analysis, for assigning meanings to the data or the transcripts of some meanings assigned by human beings.
  - **Message**: Data that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or transcripts of some knowledge acquired by humans.
- **A visualization process is a search process**. Given a data set, a user first makes decisions about which visualization tools to use to explore the data set. The user then experiments with different controls, such as styles, layout, viewing position, color maps, and transfer functions, until he or she obtains a satisfactory collection of visualization results.

(Chen et al., 2008)

Graphs vs Tables

Credit: *ModernAnalyst.com*

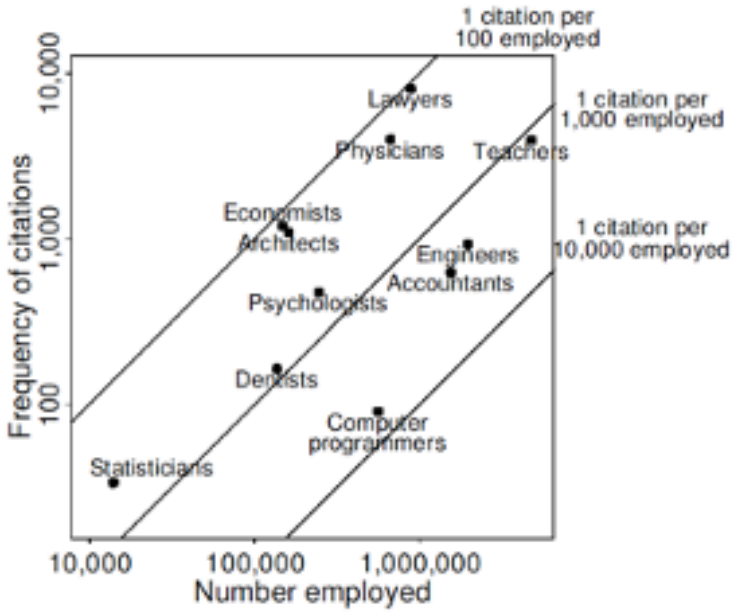
Graphs vs Tables

- General rule:
  - **Tables** are best suited for looking up specific information.
  - **Graphs** are better for perceiving trends, making comparisons, and making predictions.
- **Tables are more effective if the goal is to read off exact numbers.** However, the interest in a statistical paper is typically based on comparisons, not absolute numbers. When dozens of such potential comparisons are possible, they can be seen much more clearly in a well-chosen graphical display than in a table.
- Many people say that graphs make the data easier to see. But this is not always so. Many graphs are not easy to follow. **Graphs usually fail if they do not have a simple story-line to tell.** This restricts them to communicating qualitative aspects of the data, i.e., shapes or orders of magnitude. For this, they can excel. On the contrary, detailed quantitative features are usually difficult to read off of a graph. For that, well-designed numerical tables may be better.

(Chen et al., 2008)

Graphs vs Tables

Profession	*Frequency of recent citations*	*1996 total employed (1,000)*	*Relative frequency*
Lawyers	8101	880	9.2
Economists	1201	148	8.1
Architects	1097	160	6.9
Physicians	3989	667	6.0
Statisticians	34	14	2.4
Psychologists	479	245	2.0
Dentists	165	137	1.2
Teachers (not university)	3938	4724	0.8
Engineers	934	1960	0.5
Accountants	628	1538	0.4
Computer programmers	91	561	0.2
Total	20,657	11,034	1.9



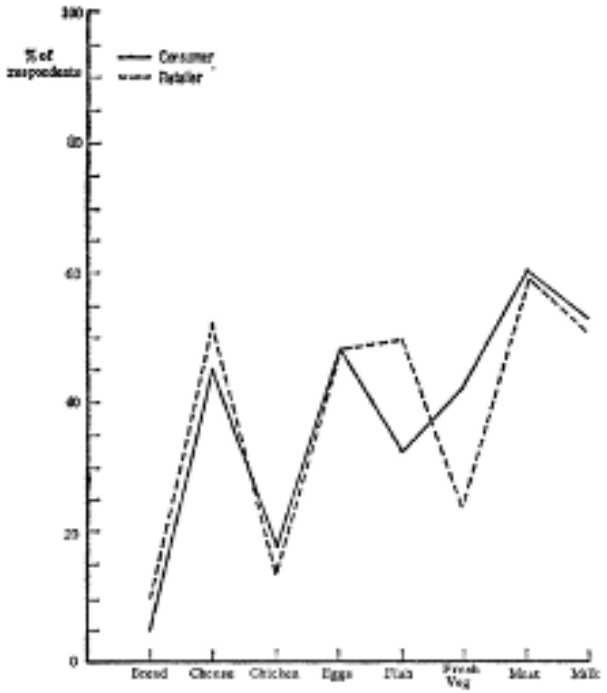
(Ehrenberg, 1978)

Graphs vs Tables

Consumers' (C) and Retailers' (R) ratings of the nutritional and economic values of different foods

Foods	*Nutritional*		*Economic*	
	C	*R*	*C*	*R*
Meat	62	58	14	11
Milk	55	52	44	95
Eggs	49	48	85	61
Cheese	45	52	30	62
Fresh Veg.	42	24	25	18
Fish	33	52	20	10
Chicken	18	13	70	25
Bread	5	11	5	21

*In decreasing order of Consumers' Nutritional Ratings.



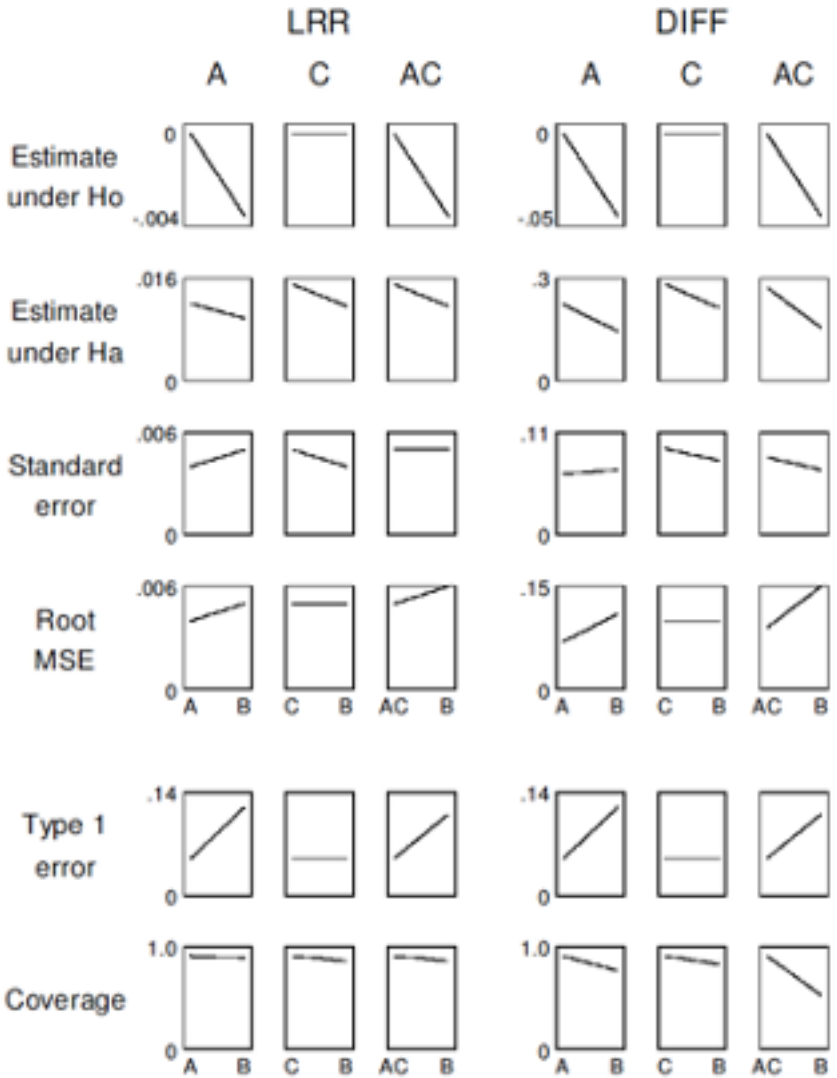
(Ehrenberg, 1978)

Graphs vs Tables

	Estimate under H_0	*Type 1 error*	*Estimate under H_A*	*Standard error*	*$MSE^{1/2}$*	*Coverage*
Distribution A						
DIF^A	.000	.05	.012	.004	.004	.95
DIF^B	−.004	.12	.01	.005	.005	.93
LRR^A	.00	.05	.23	.070	.07	.95
LRR^B	−.05	.12	.15	.070	.110	.80
Distribution C						
DIF^C	.000	.05	.015	.005	.005	.95
DIF^B	.000	.05	.012	.004	.005	.90
LRR^C	.00	.05	.29	.10	.10	.95
LRR^B	.00	.05	.22	.080	.100	.87
Distribution AC						
DIF^{AC}	.000	.05	.015	.005	.005	.95
DIF^B	−.004	.11	.012	.005	.006	.90
LRR^{AC}	.00	.05	.28	.09	.09	.95
LRR^B	−.05	.11	.16	.07	.15	.55

(Ehrenberg, 1978)

Graphs vs Tables



(Ehrenberg, 1978)

The Perfect Table

- **Think about what purpose the table is meant to serve** for the reader. APA says: "Although tables and figures attract attention, they should not be used for mere decoration in an academic paper."
- **Each table must tell a worthwhile and intelligible story**. How does your table support your argument? Do you want your readers to notice a particular pattern in your data? Should they notice an outlier? Or maybe you want to present a list of information, clearly organized so that readers can easily locate the individual fact(s) that they are interested in.
- **Your table should be as simple as possible**, so that it fits nicely onto the page, but also to lessen the strain on the cognitive load of your readers. Include only the data you need to make your argument. Write headings with clear, simple language. Use color and lines/dividers sparingly.
- **Make sure it stands on its own.** The main text should highlight the most important points in the table and give a sense of the significance of the data, but make sure that you do not duplicate the content of the table in text form.
- **Make it accessible.** Your tables must not be images, but rather correctly marked text. This mark-up allows people using screen-reading software to access the content.

(Goforth, 2021)

References

Chen, M., Ebert, D., Hagen, H., Laramee, R. S., Van Liere, R., Ma, K.-L., Ribarsky, W., Scheuermann, G., & Silver, D. (2008). Data, information, and knowledge in visualization. *IEEE computer graphics and applications*, *29*(1), 12–19.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, *79*(387), 531–554.

Ehrenberg, A. (1978). Graphs or tables. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *27*(2), 87–96.

Evergreen, S. D. (2019). Effective data visualization: The right chart for the right data..

Friendly, M. (2006). A brief history of data visualization [(In press)]. In C. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of computational statistics: Data visualization*. Springer-Verlag.

Goforth, J. (2021). *Academic publishing: Creating effective tables.*. http://www.centerforengagedlearning.org/academic-publishing-creating-effective-tables

Schwabish, J. (2021). *Better data visualizations: A guide for scholars, researchers, and wonks.*. Columbia University Press.

Todorovic, D. (2008). Gestalt principles', scholarpedia, 3 (12): 5345.