

Storytelling with Data

Module 7: The storytelling process, encoding data, visual perception

Scott Spencer
Faculty and Lecturer
Columbia University



Unanswered, or new, questions from discussion?

Agenda

Next deliverable – **draft** infographic

Today's objectives

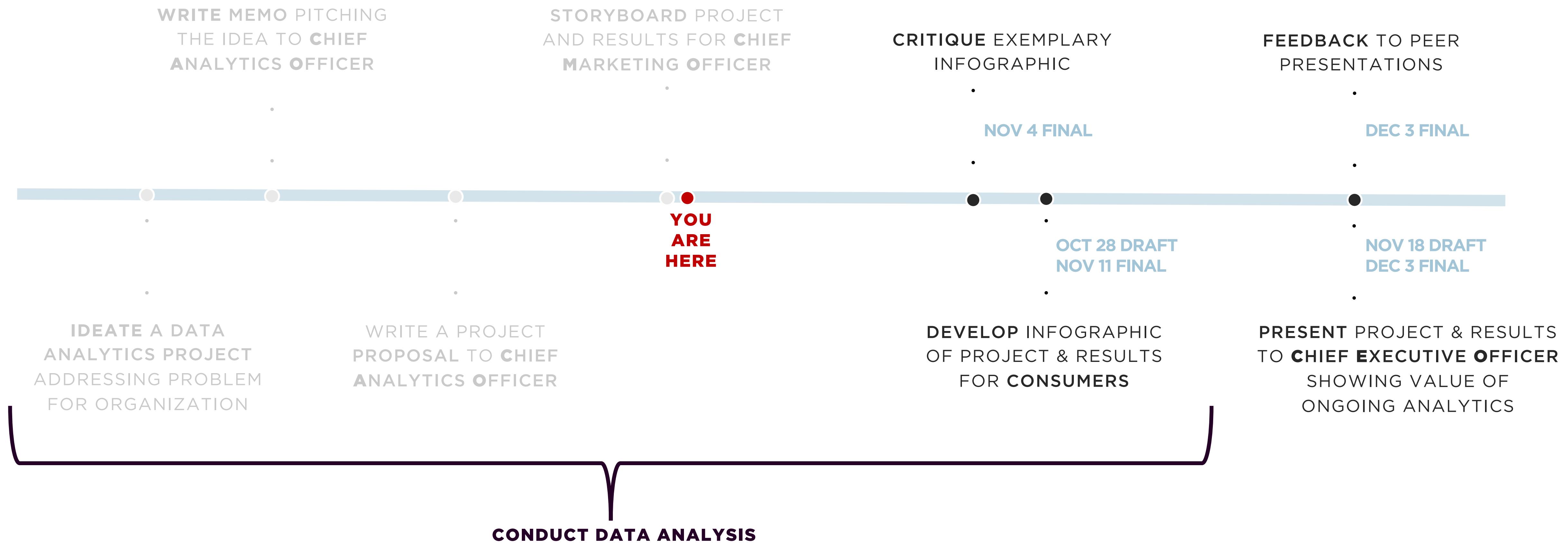
Storytelling process continued

Perception in visual narrative

Our project timeline

Upcoming deliverables

Information graphic – reframe your story, this time building off the messages you built for the marketing team in order to craft an infographic that displays the results of the analytic work in a way that is accessible, engaging, and exciting for a **general or consumer audience**.



Today's Objectives

Objectives

- 1 | Explain importance of effectively framing a story
- 2 | Consider audience perception of visual components of a story

The storytelling process, continued



See, Think, Design, Produce

Corum

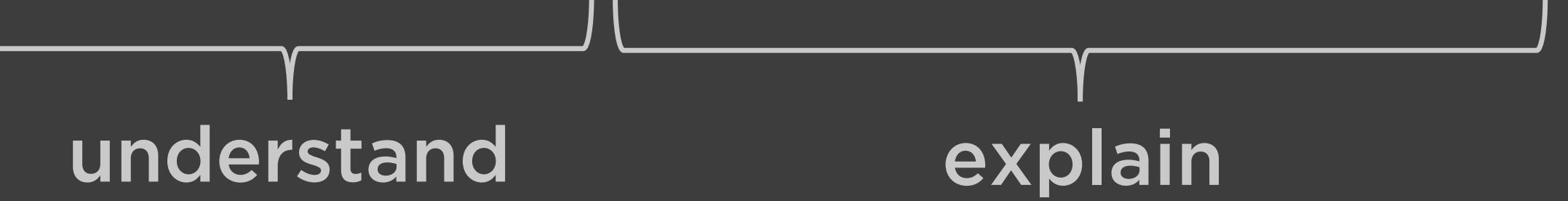
A former student of Edward Tufte at Yale, Jonathan is science graphics editor at The New York Times and has won 28 awards from the Society for News Design and 18 medals from the international Malofiej competition, including Best of Show.

His projects for an audience of NYT readers ...



13pt style.org

See, Think, Design, Produce

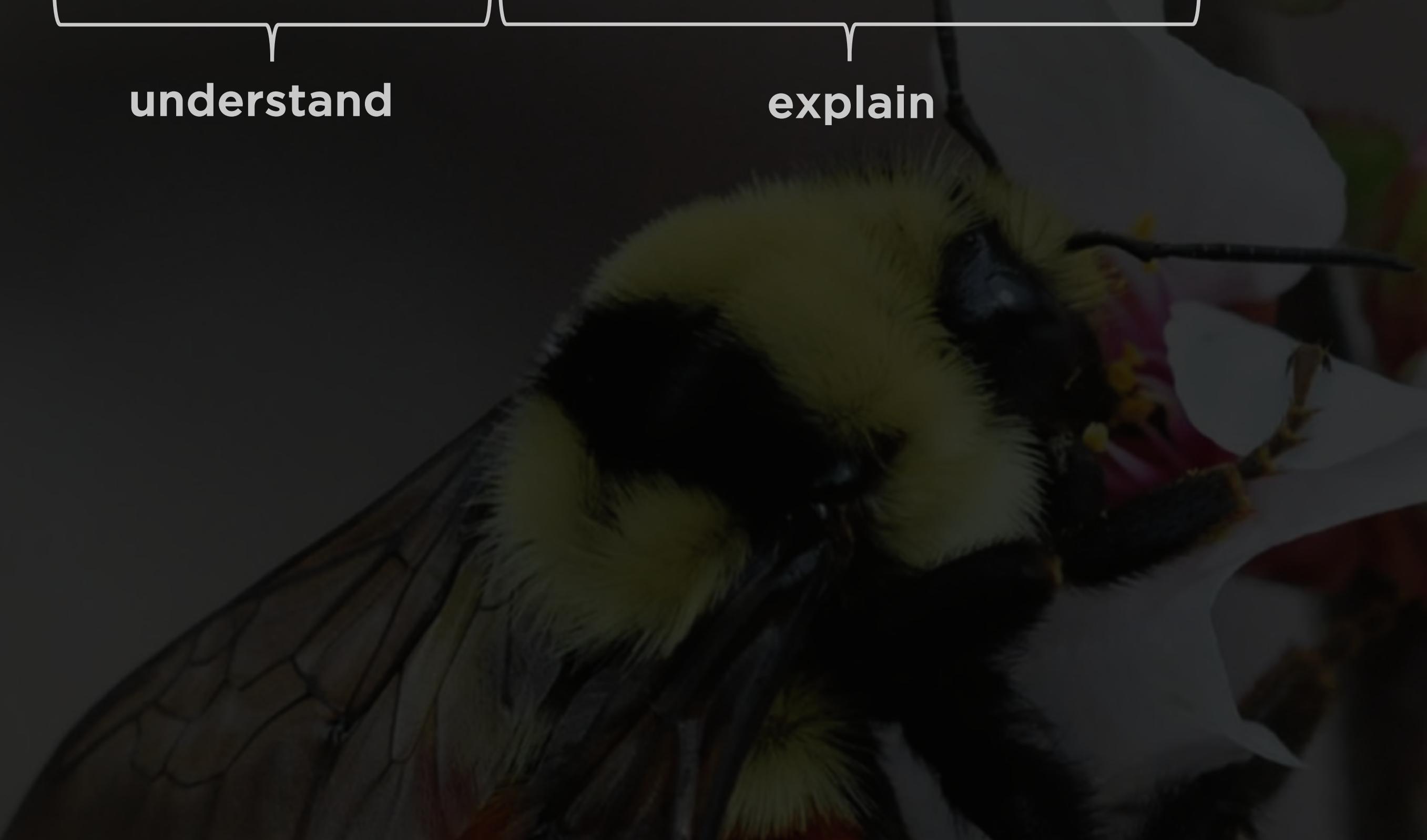


understand explain

See, Think, Design, Produce

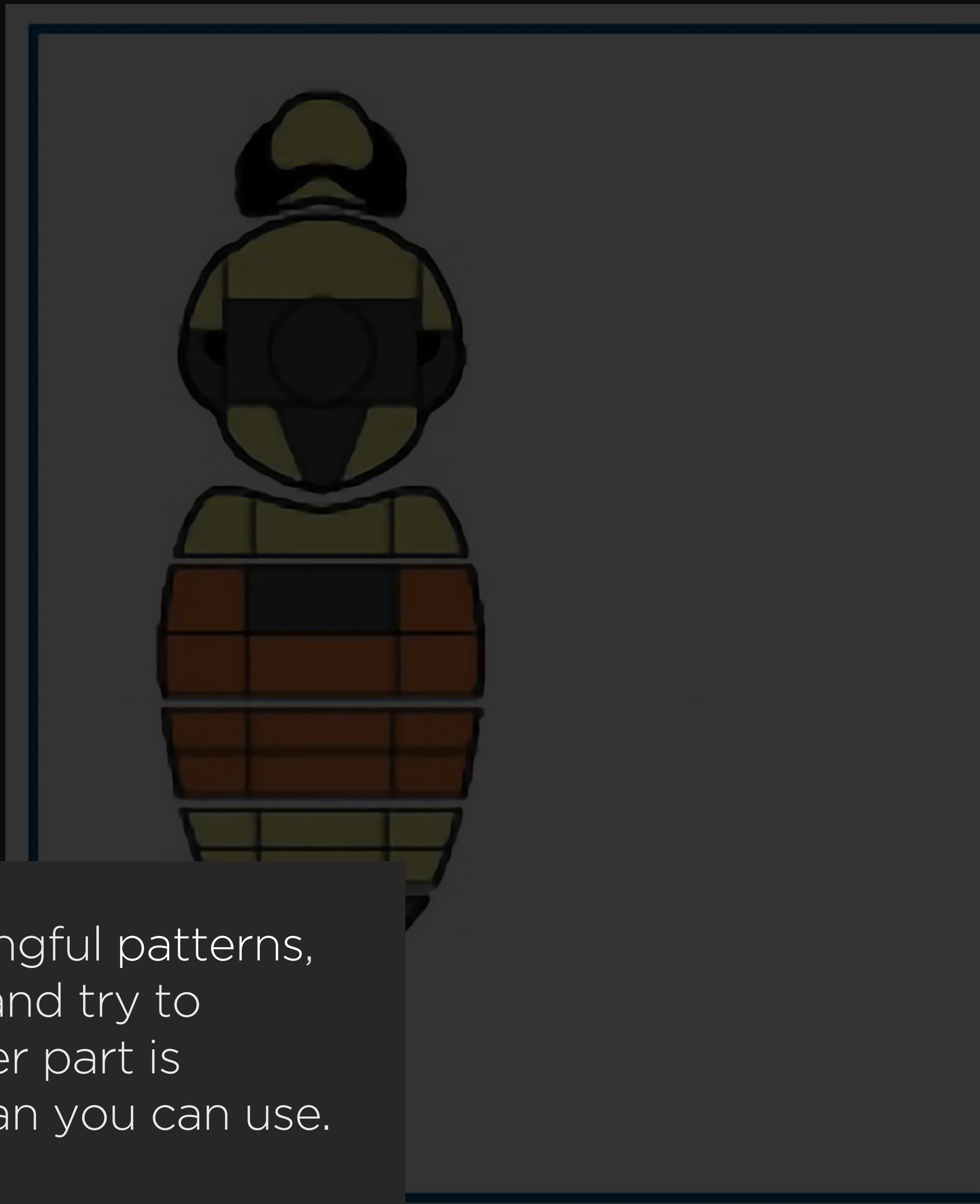
understand

explain



**Search for
patterns
by comparing**

Visualization is not counting. Search for meaningful patterns, try to understand **patterns**, visualize patterns and try to explain them. Part of this is **comparing**. Another part is finding what's **possible**. Look at more ideas than you can use.



See, Think, Design, Produce

understand

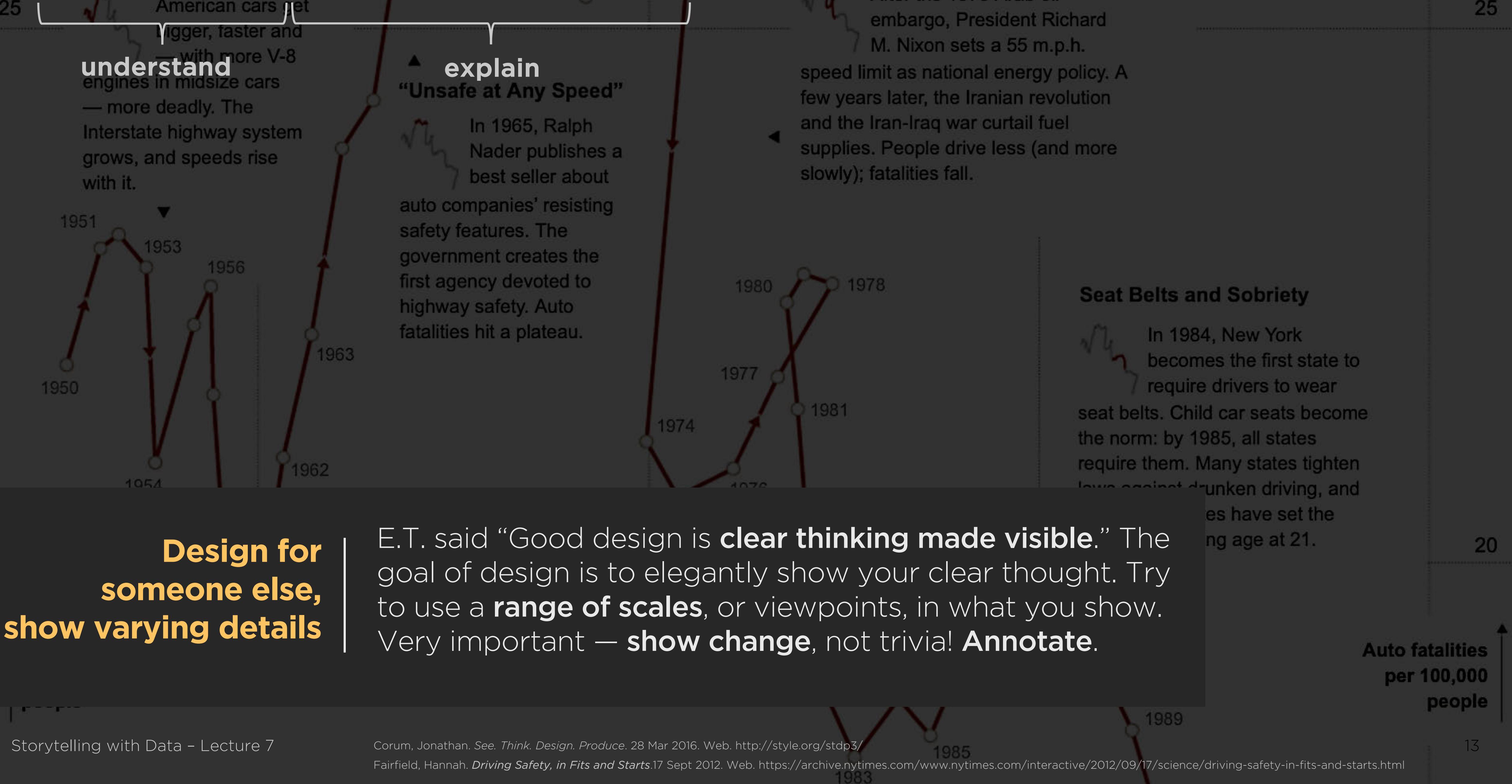
explain



Sketch
until your
aha! moment

Finding a **clear thought** through visualization can begin with **sketching**, on either paper or screen. Sketching is visual problem solving, not a commitment. It's much easier to begin with an ugly sketch and make it prettier as you work on design.

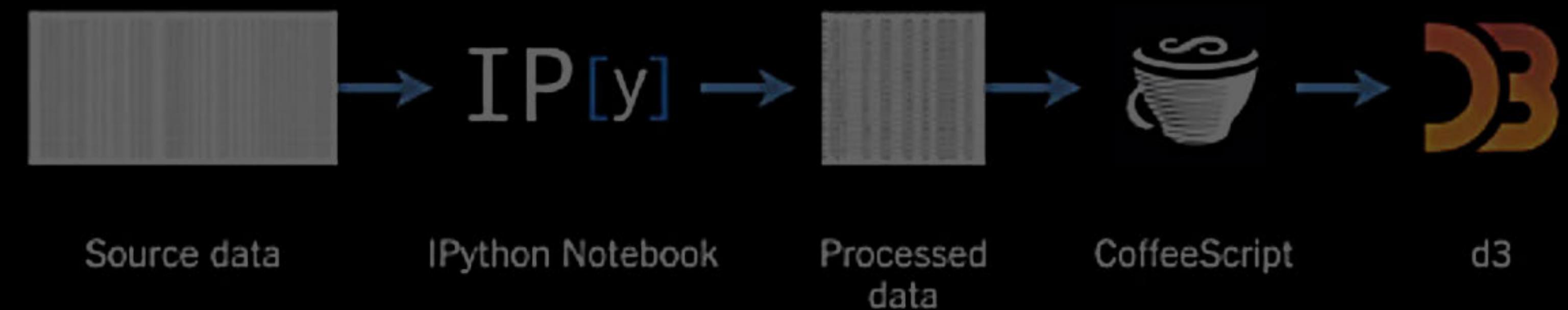
See, Think, Design, Produce



See, Think, Design, Produce

understand

explain



Hone ideas Within limitations

Embrace limitations; use them to **hone your ideas.** Understand every step—leave nothing to magic—in your production. Design is **cumulative decision making.**

Questions for discussion.

What did you find interesting or helpful in the way Jonathan described his process of visual storytelling?

Which of his examples were most vivid, most memorable, to you? What made them so?

Why review data graphically?

Classic example, data from Anscombe

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Classic example, data from Anscombe

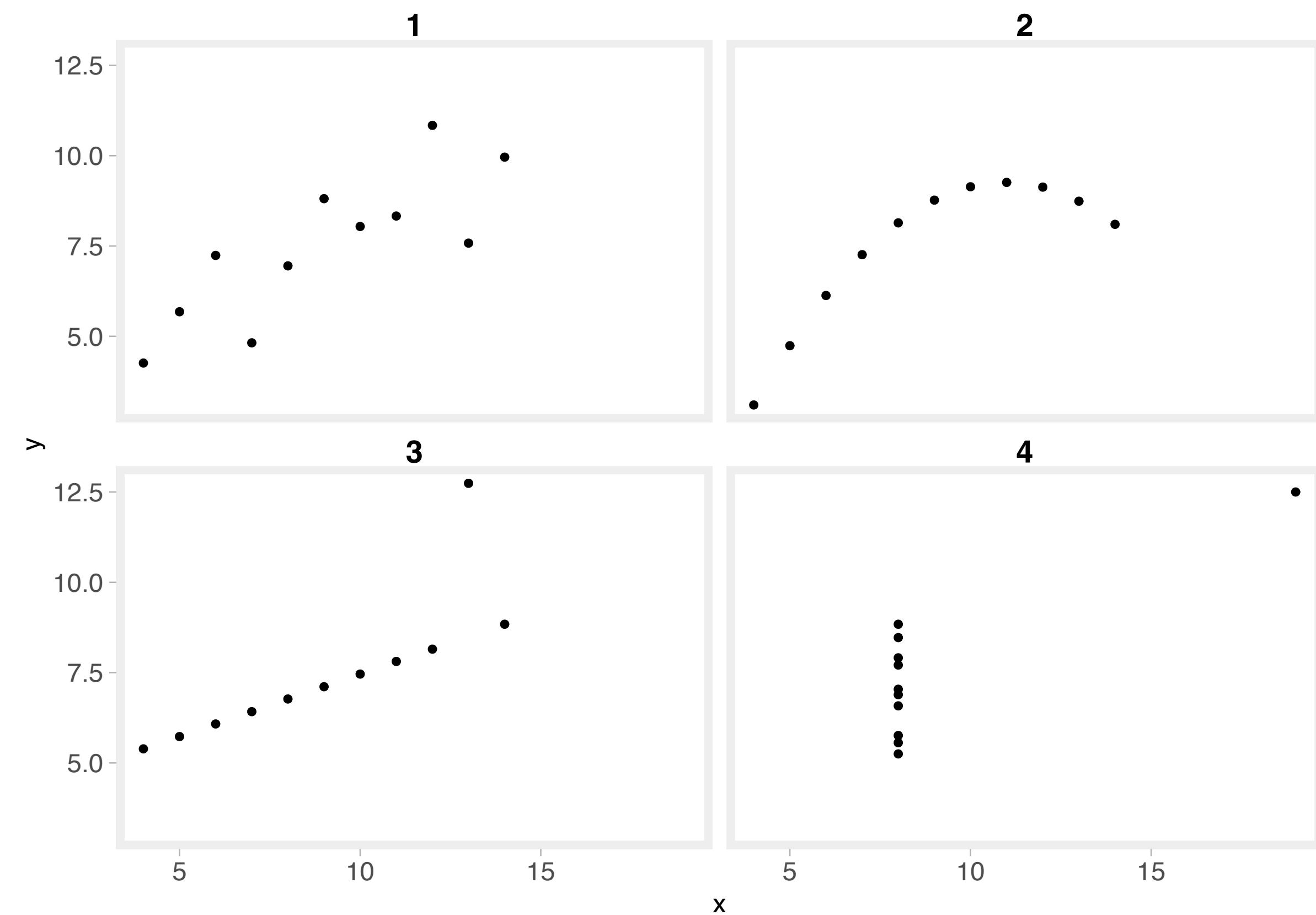
1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary statistics

	1		2		3		4	
	x	y	x	y	x	y	x	y
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
sd	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
<hr/>								
Parameter	Mean	Std Err	t-val	p-val				
Dataset 1								
(Intercept)	3.000	1.125	2.667	0.026				
x	0.500	0.118	4.241	0.002				
Dataset 2								
(Intercept)	3.001	1.125	2.667	0.026				
x	0.500	0.118	4.239	0.002				
Dataset 3								
(Intercept)	3.002	1.124	2.670	0.026				
x	0.500	0.118	4.239	0.002				
Dataset 4								
(Intercept)	3.002	1.124	2.671	0.026				
x	0.500	0.118	4.243	0.002				

Classic example, data from Anscombe

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



Encoding data graphically

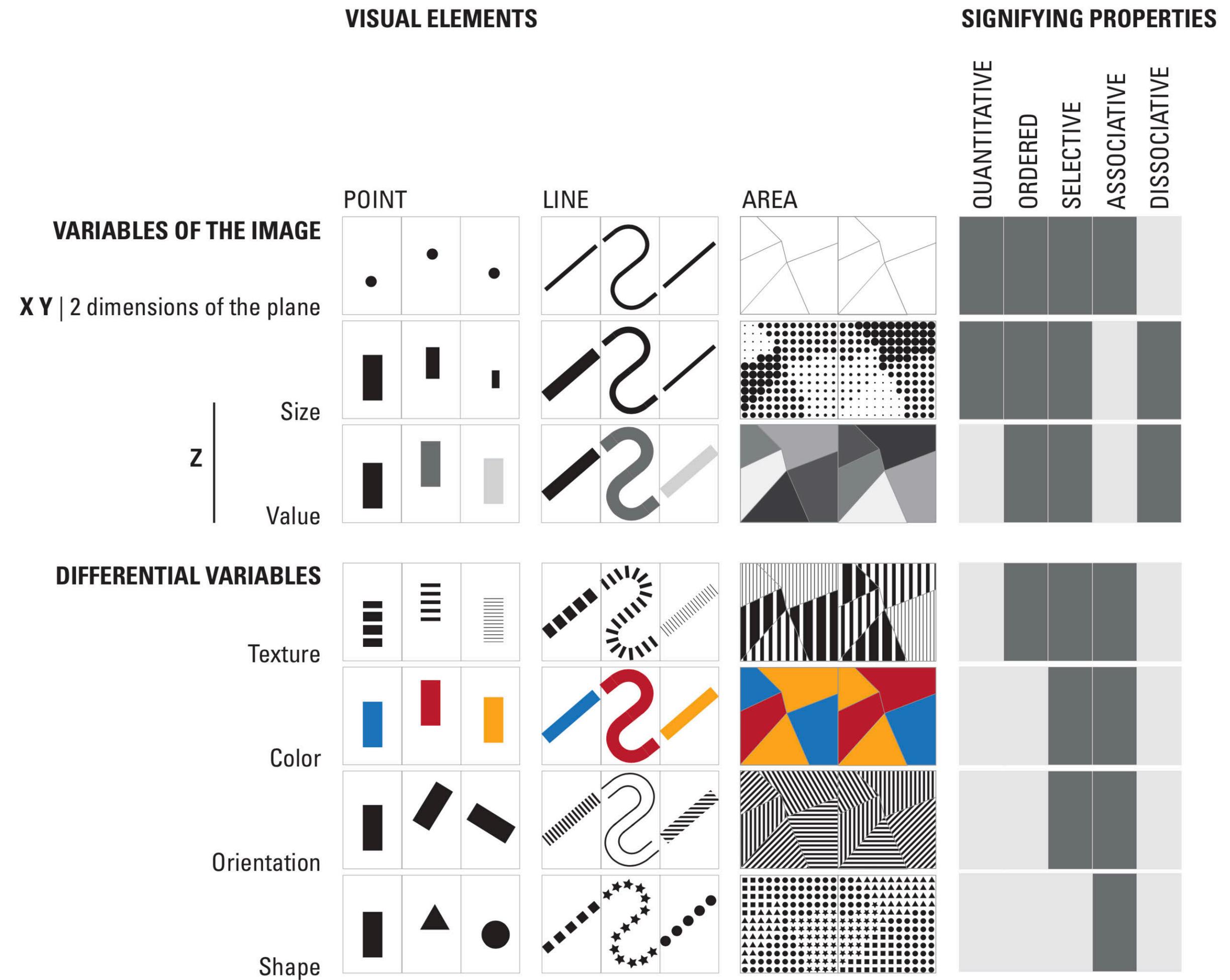
After fixing a point on a plane, we encode or mark information in six possible ways



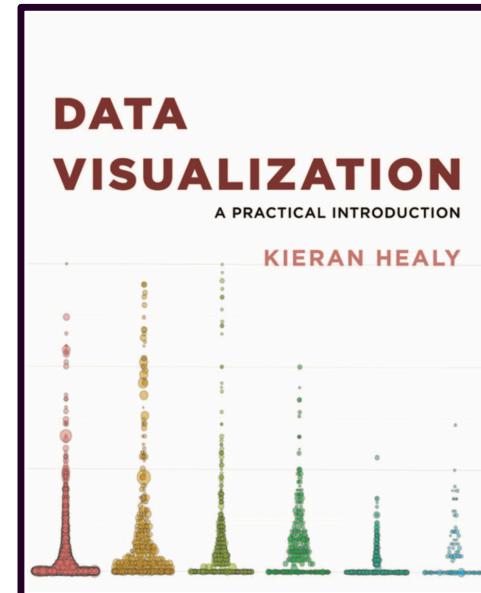
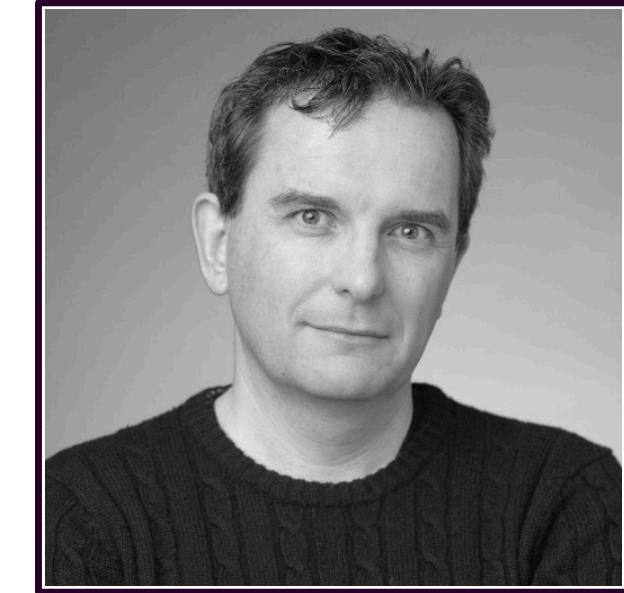
Semiology of Graphics diagrams, networks, maps

Bertin

Jacques was a world-renown authority of the subject of information visualization. His text, *Semiology of Graphics*, is foundational in the fields of design and cartography.



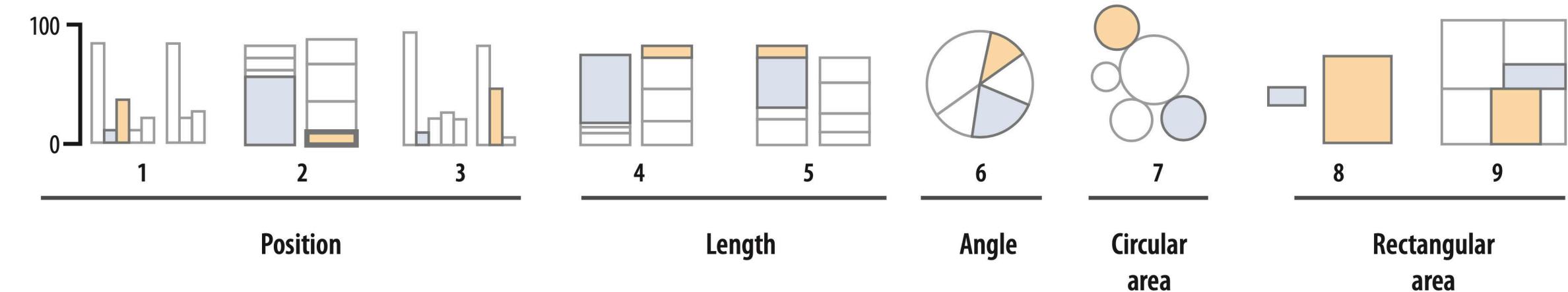
Decoding graphic representations of data



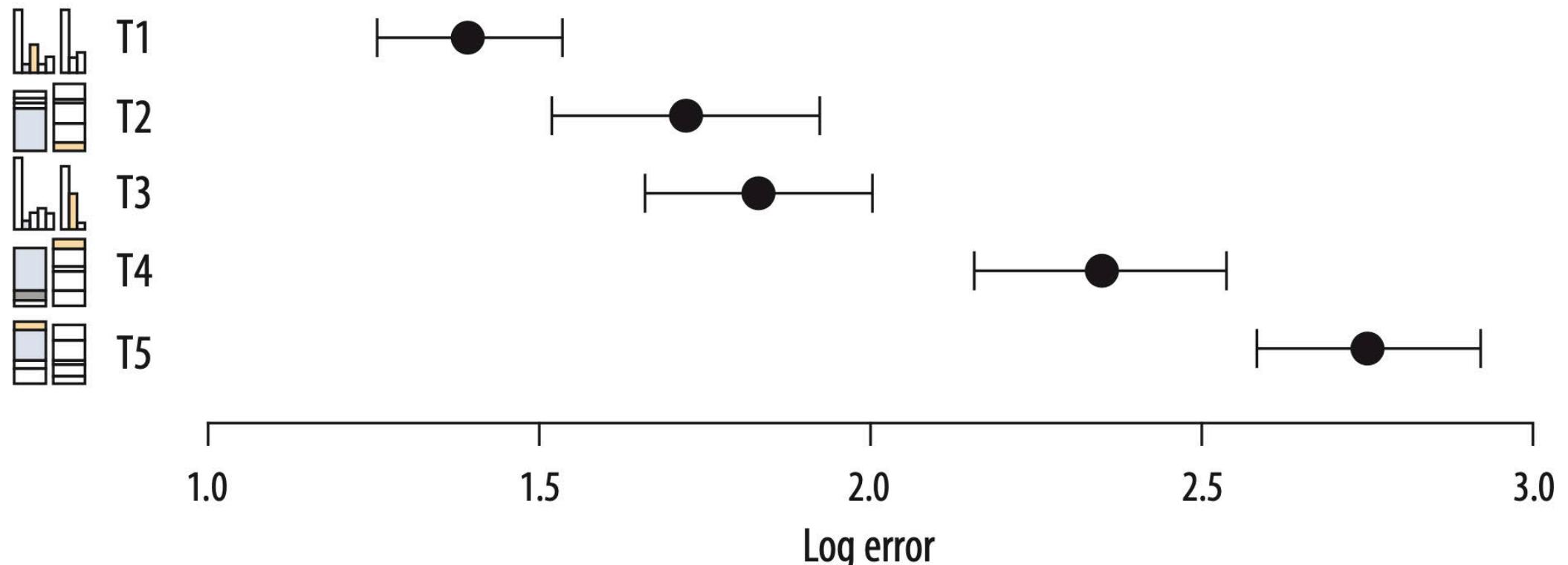
Data Visualization: a practical introduction

Healy

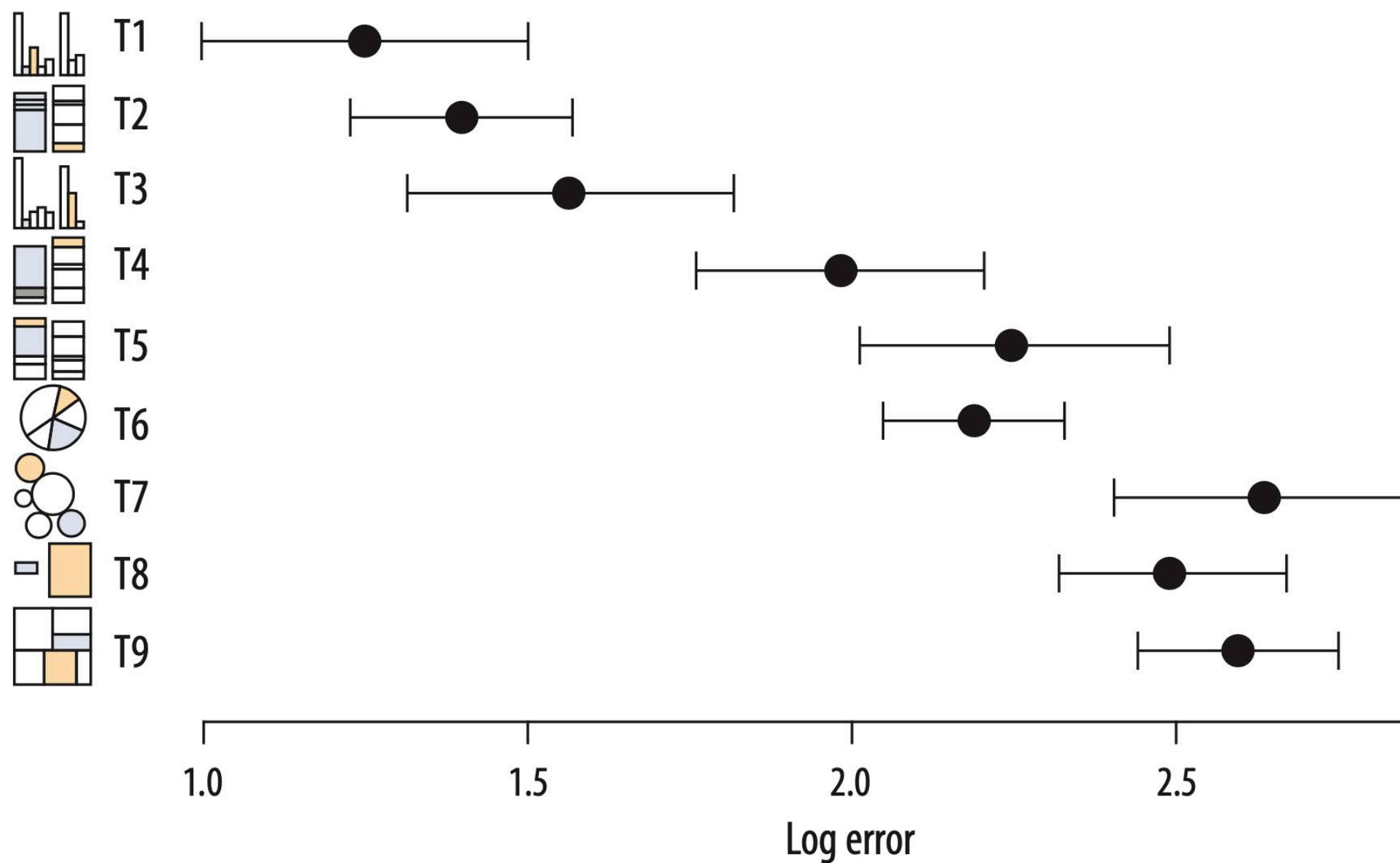
A PhD graduate from Princeton, Kieran is associate professor of sociology at Duke University. His book has been described as “covering the ‘why do’ as well as the ‘how to’ of data visualization.” — Andrew Gelman



Cleveland & McGill's results



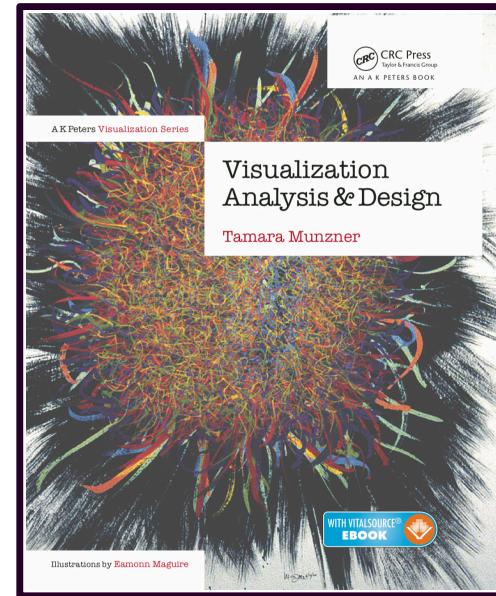
Crowdsourced results



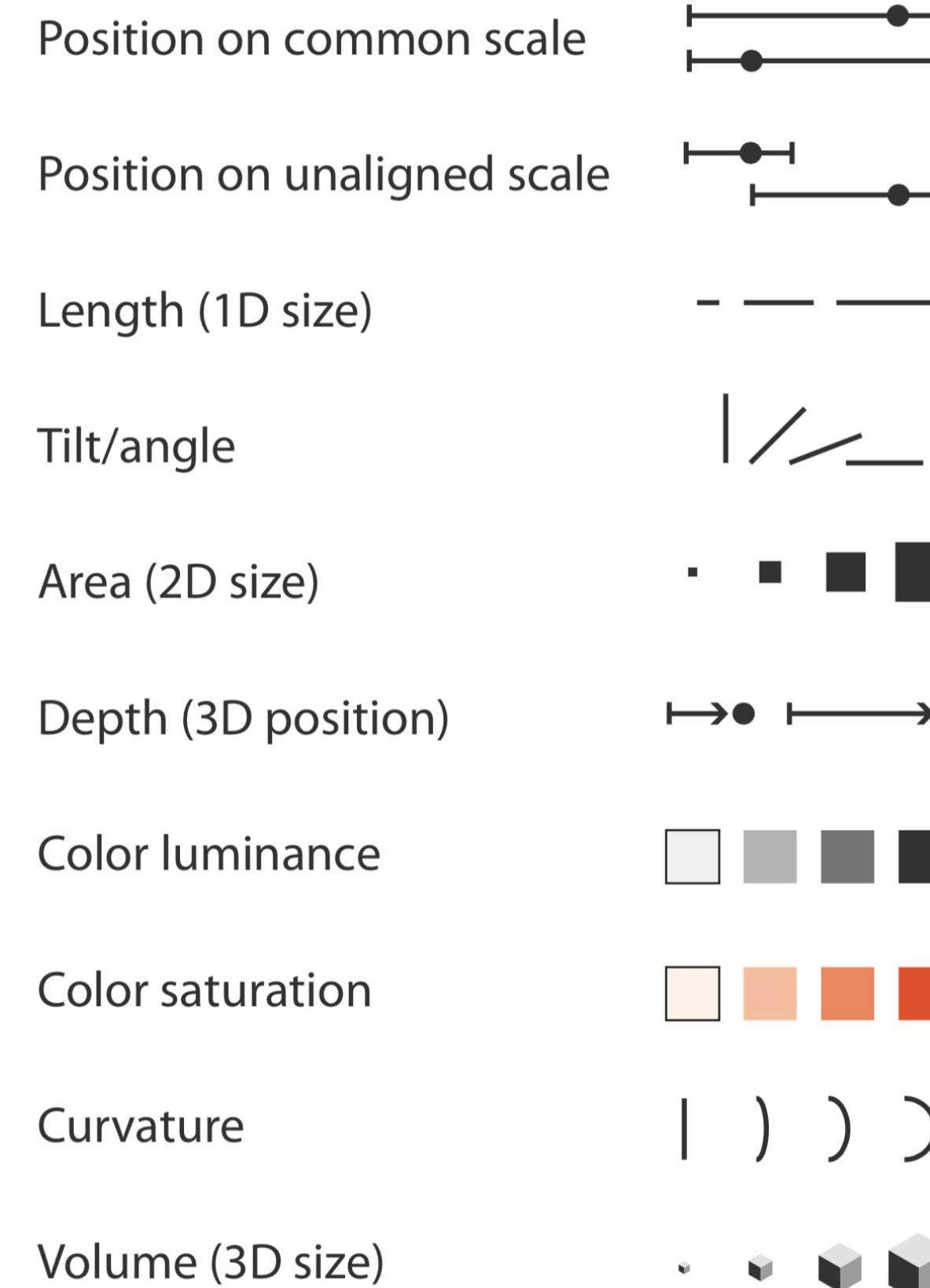
Visualization Analysis and Design

Munzner

Tamara is a professor at the University of British Columbia Department of Computer Science, and holds a PhD from Stanford. Her work in information design is well-known.



→ Magnitude Channels: Ordered Attributes



→ Identity Channels: Categorical Attributes

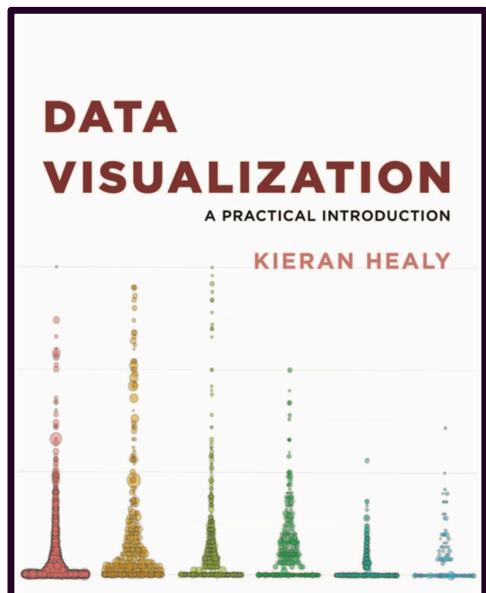


Audience (mis)perceptions decoding data graphics

Data Visualization: a practical introduction

Healy

A PhD graduate from Princeton, Kieran is associate professor of sociology at Duke University. His book has been described as “covering the ‘why do’ as well as the ‘how to’ of data visualization.” — Andrew Gelman



Issues when visually encoding data

Aesthetic

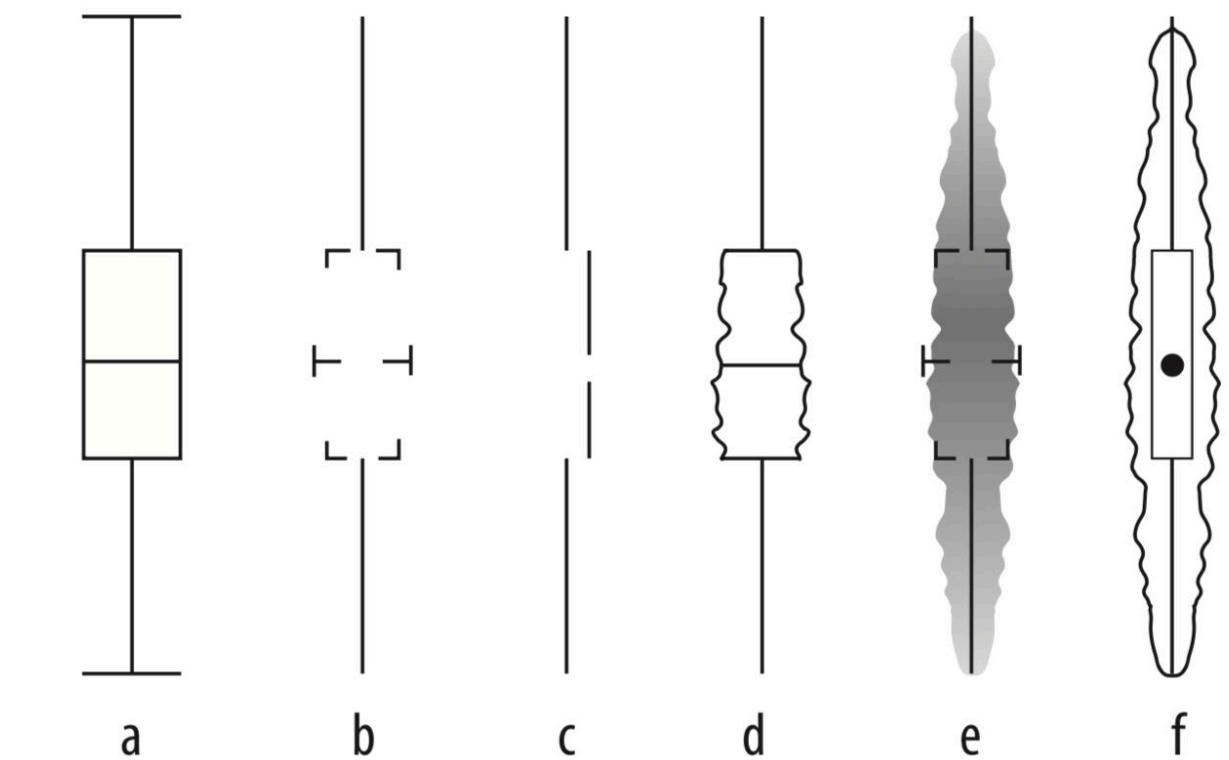
Aesthetic
Substantive
Perceptual

Consider whether every mark, color, and luminance on a statistical graph conveys information and supports messaging.

Memorable,
but hard to read.
“Junk chart”



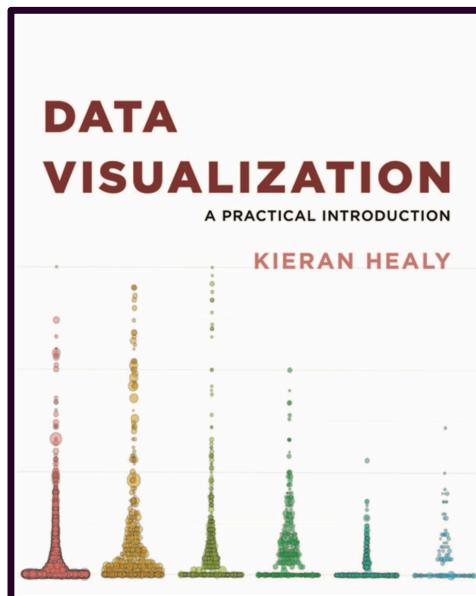
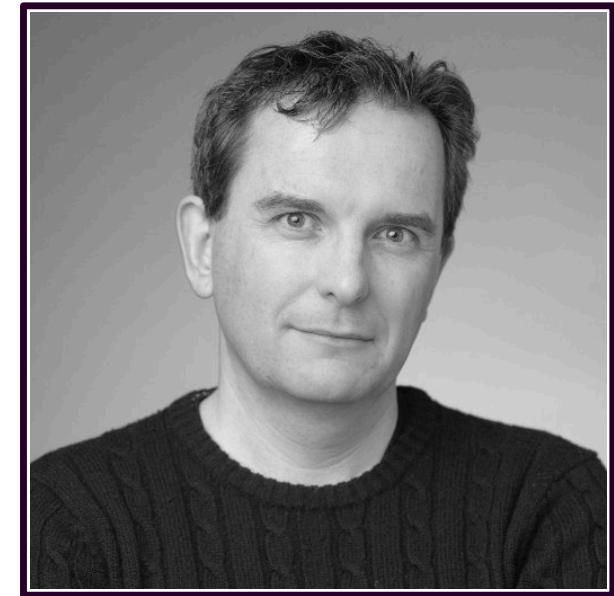
Consider the audience when erasing non- or redundant-data ink



Data Visualization: a practical introduction

Healy

A PhD graduate from Princeton, Kieran is associate professor of sociology at Duke University. His book has been described as “covering the ‘why do’ as well as the ‘how to’ of data visualization.” — Andrew Gelman



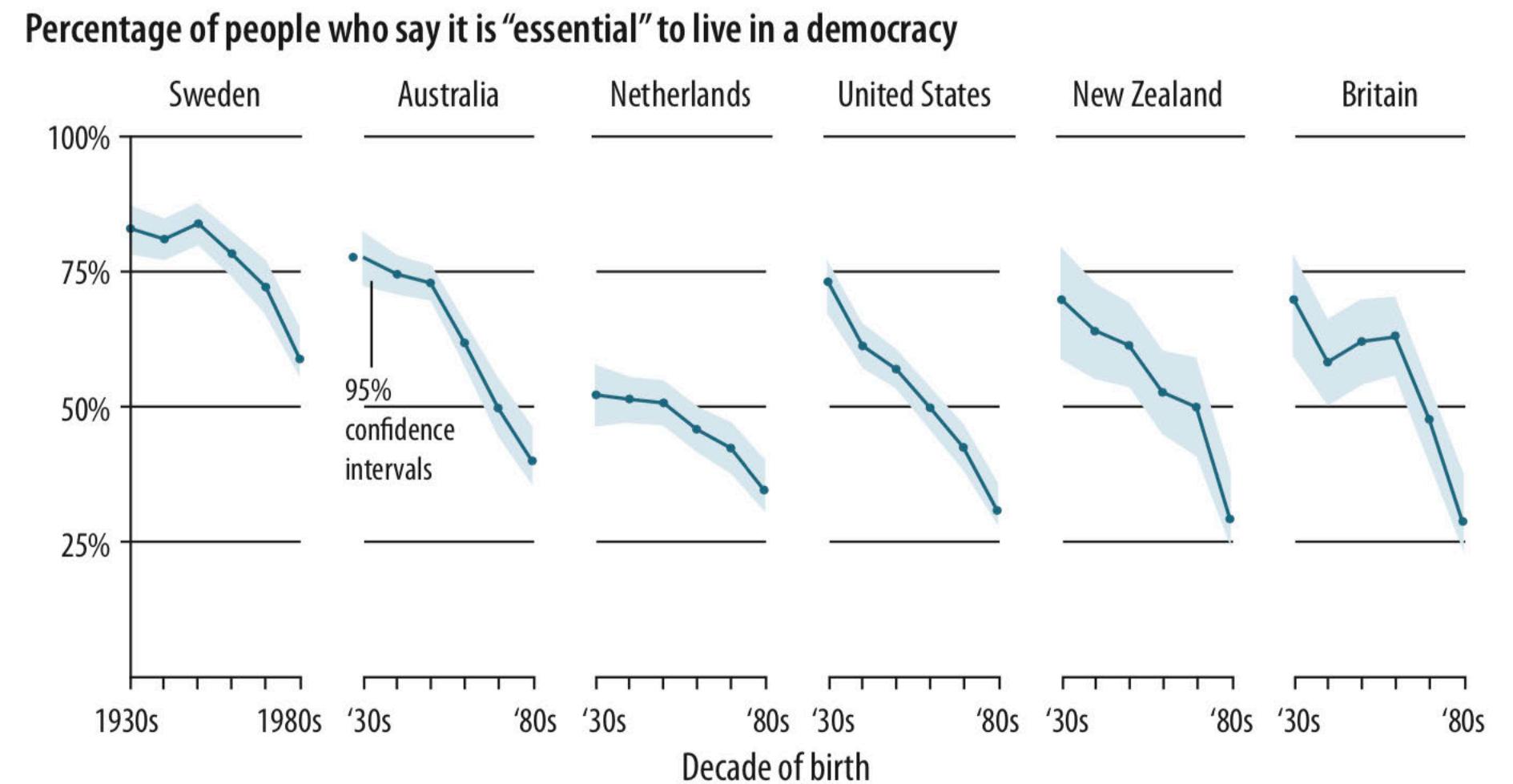
Issues when visually encoding data

Aesthetic
Substantive
Perceptual

Substantive

Consider whether your data honestly and fairly represent your message.

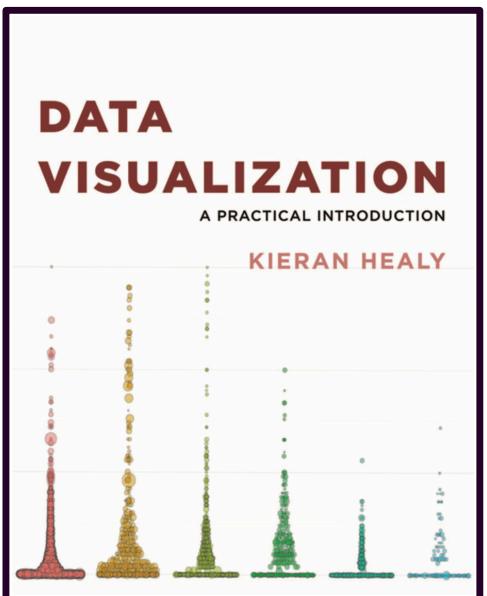
The original chart below graphed the relative change in response of respondents who selected 10 on a 10 point scale.



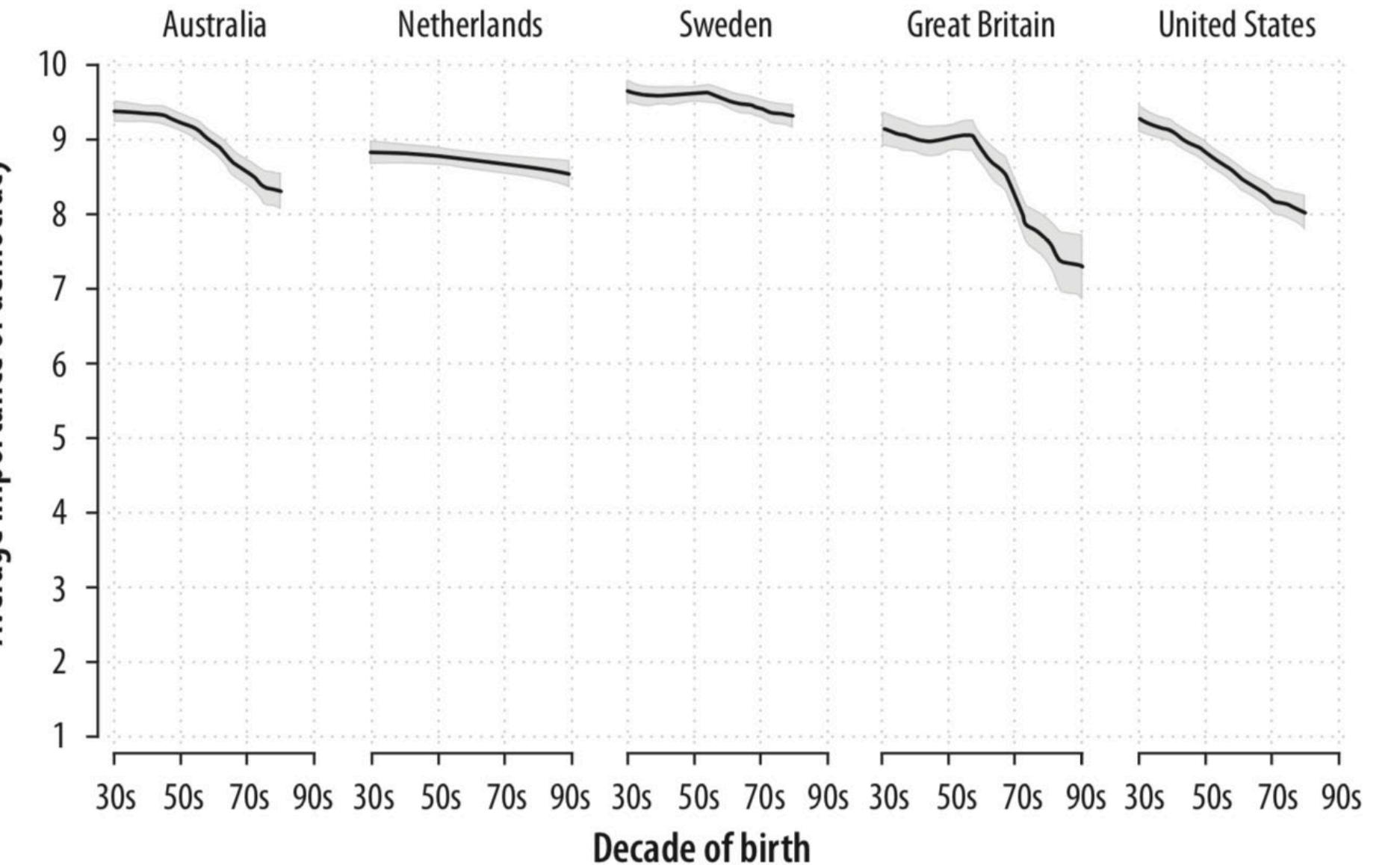
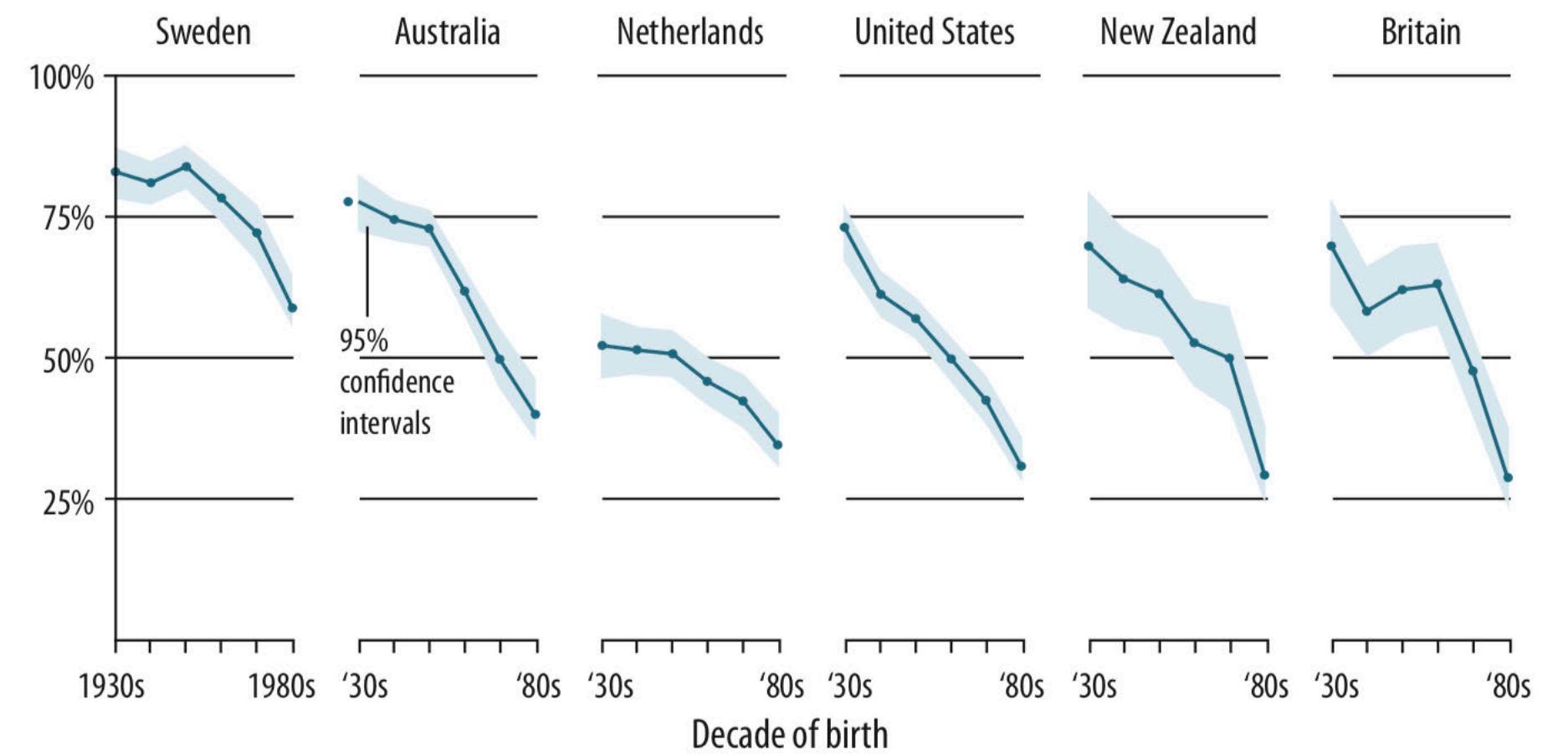
Data Visualization: a practical introduction

Healy

A PhD graduate from Princeton, Kieran is associate professor of sociology at Duke University. His book has been described as “covering the ‘why do’ as well as the ‘how to’ of data visualization.” — Andrew Gelman



Percentage of people who say it is “essential” to live in a democracy

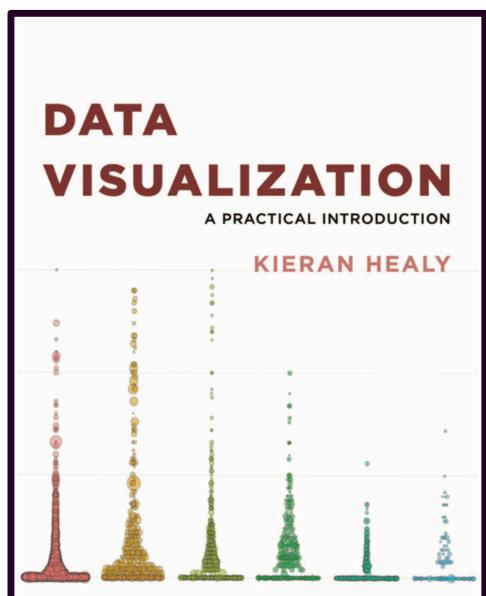
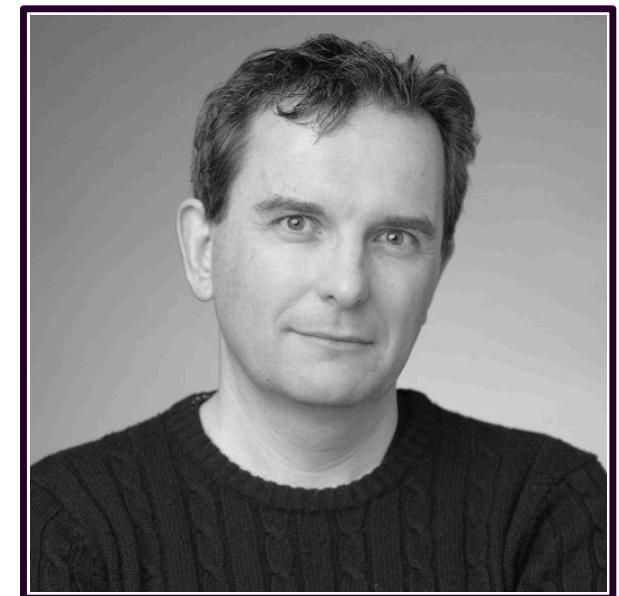


Graph by Erik Voeten, based on WVS 5

Data Visualization: a practical introduction

Healy

A PhD graduate from Princeton, Kieran is associate professor of sociology at Duke University. His book has been described as “covering the ‘why do’ as well as the ‘how to’ of data visualization.” — Andrew Gelman

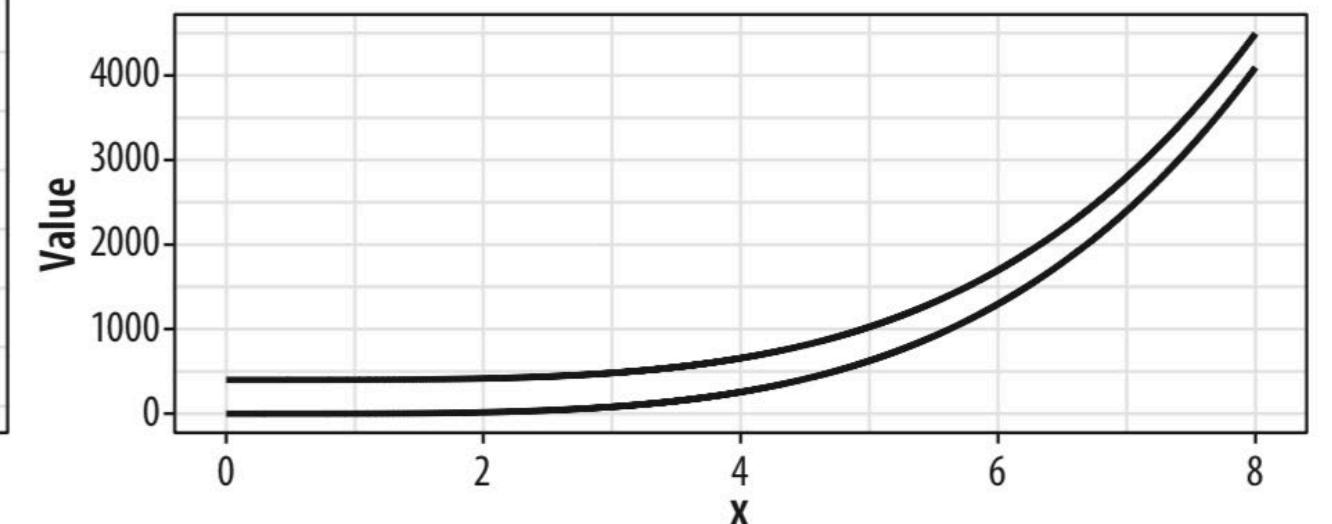
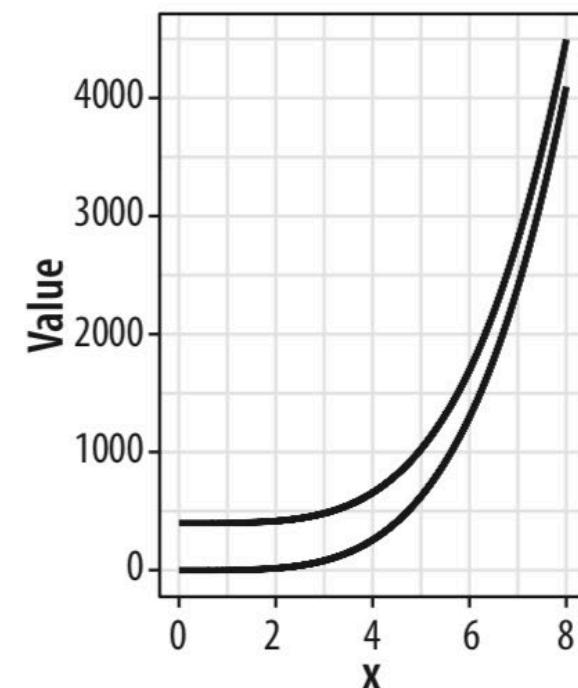
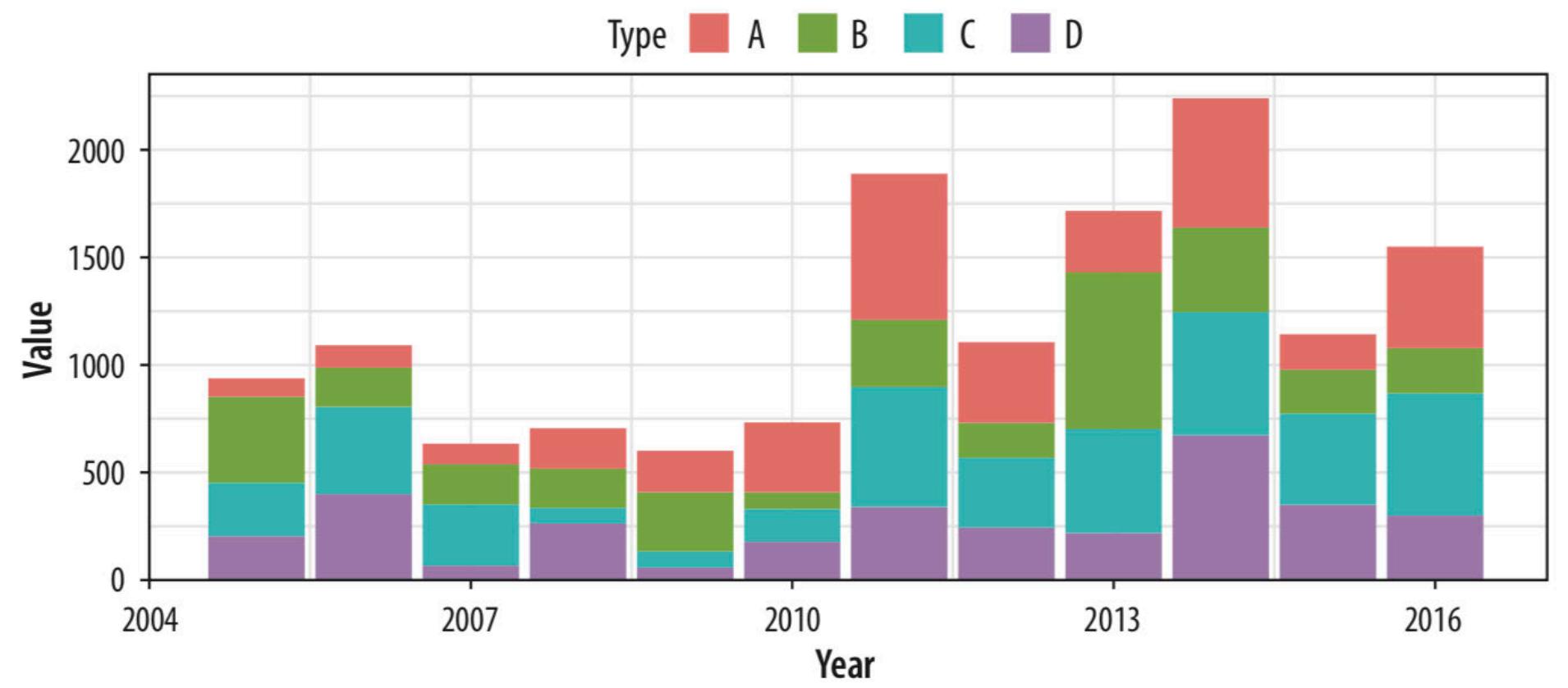


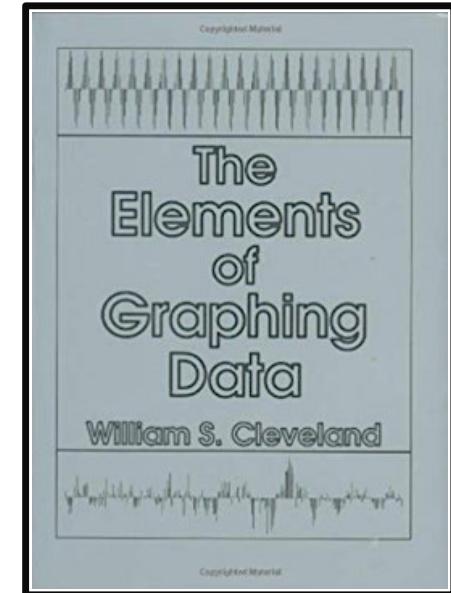
Issues when visually encoding data

Aesthetic
Substantive
Perceptual

Perceptual

Even with a reasonably-high data-ink ratio, you must choose an encoding that most naturally guides the audience to understand and compare the data. The graphs below have perceptual issues.





The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.

Superposed curves have a decoding problem

Decoding differences between two lines or curves on a graph can be inaccurate because we naturally compare the shortest distance between the lines instead of the vertical distance between the lines.





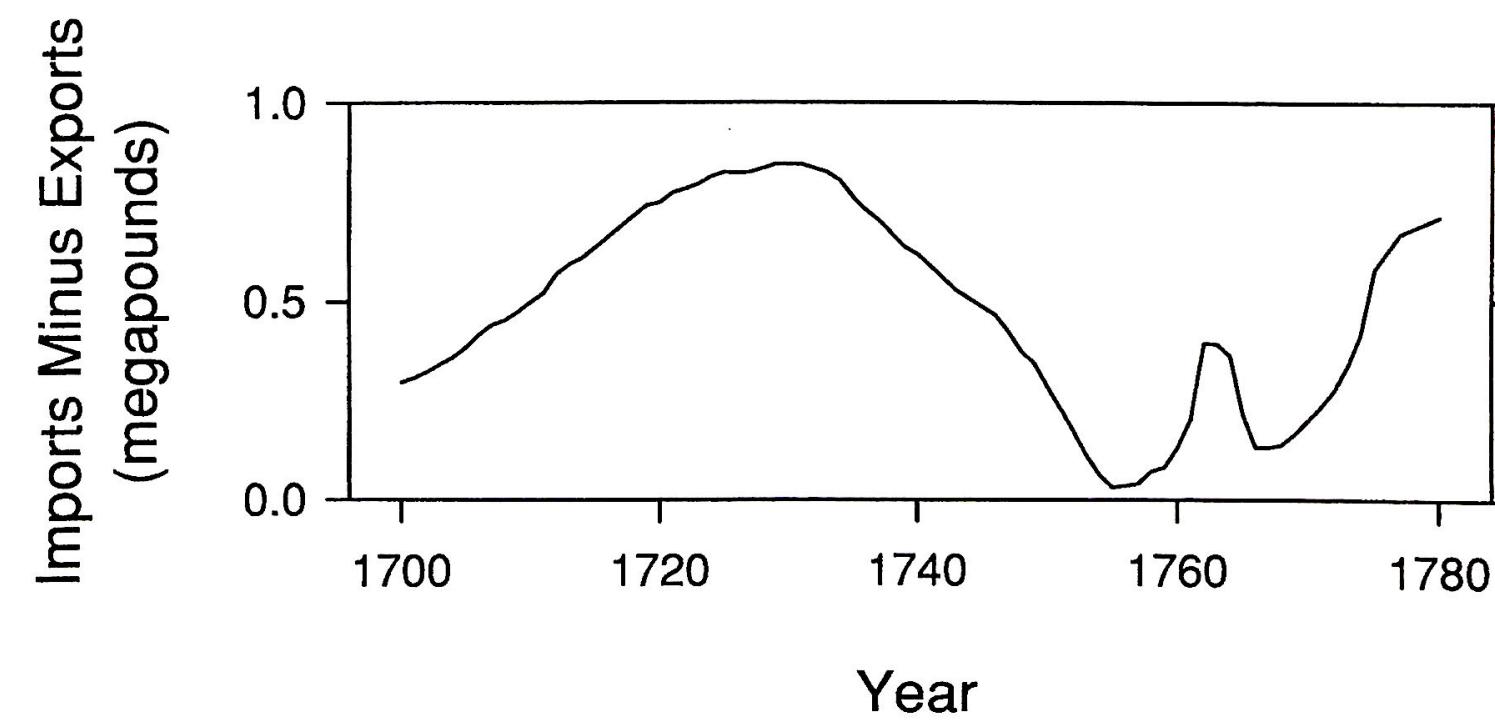
The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.

Superposed curves have a decoding problem

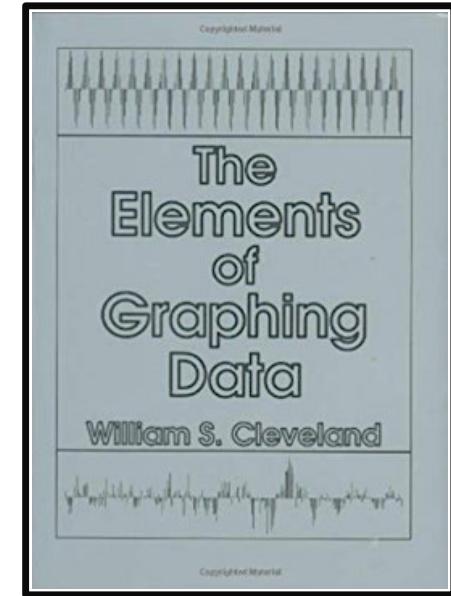
Decoding differences between two lines or curves on a graph can be inaccurate because we naturally compare the shortest distance between the lines instead of the vertical distance between the lines.



The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



We need a fixed percentage change in something to detect a difference

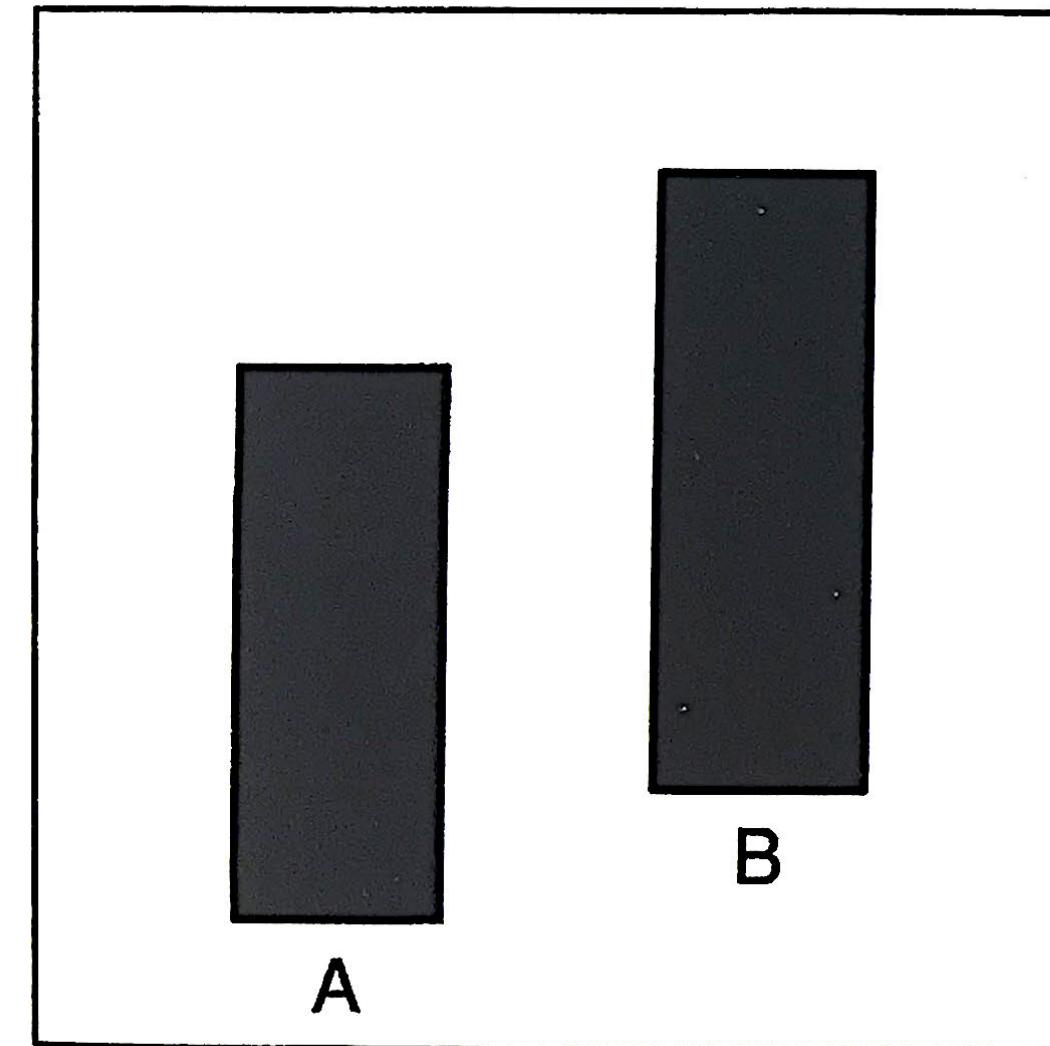
Our visual perception of differences can be stated mathematically, as Weber's Law.

If x is the magnitude of a physical attribute, say, length of a line segment, and $w_p(x)$ is a positive number such that a line of length

$$x + w_p(x)$$

is discriminated with probability p to be longer than the line of length x then,

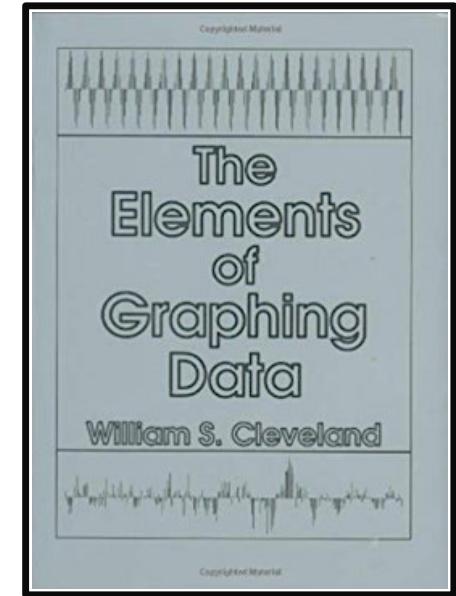
$$w_p(x) = k_p x$$



The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



We need a fixed percentage change in something to detect a difference

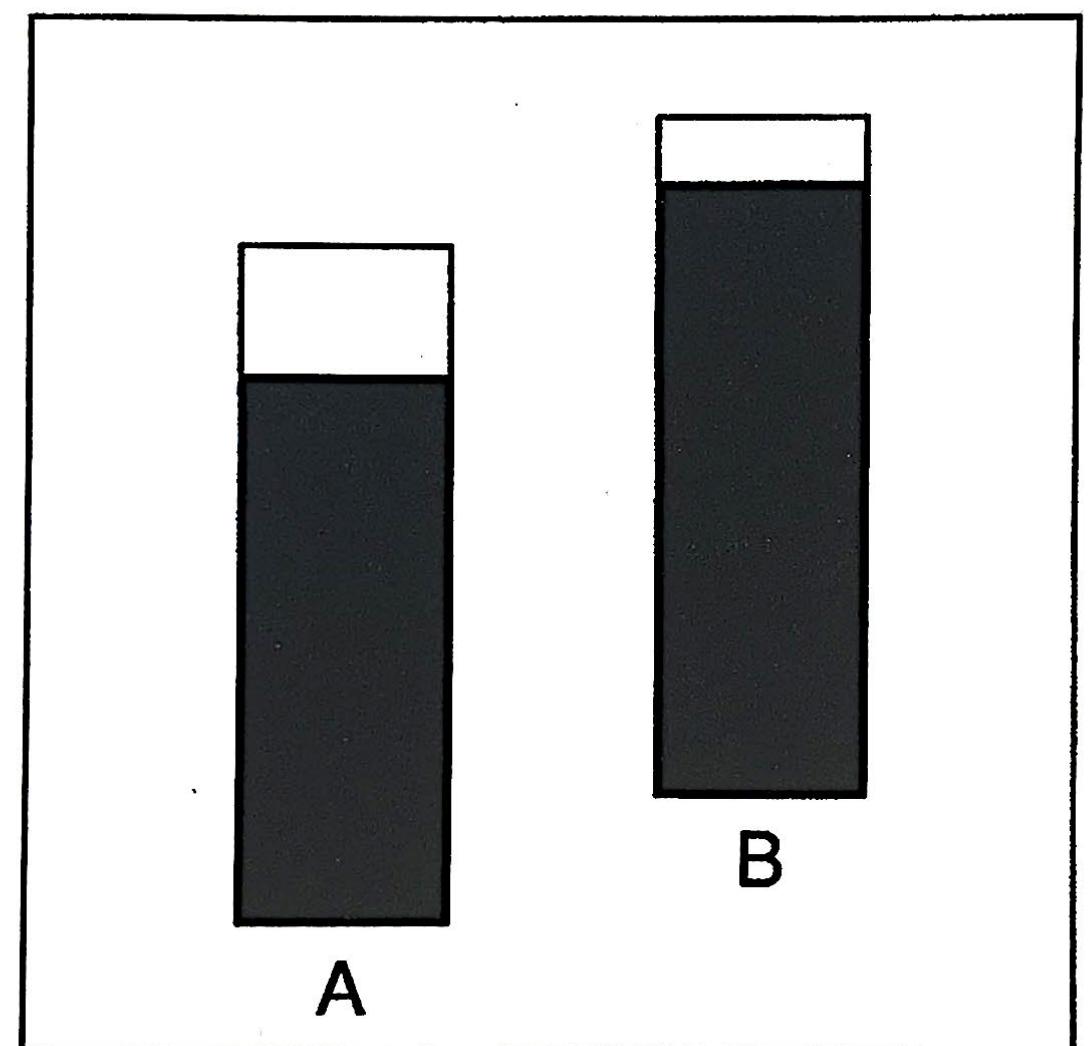
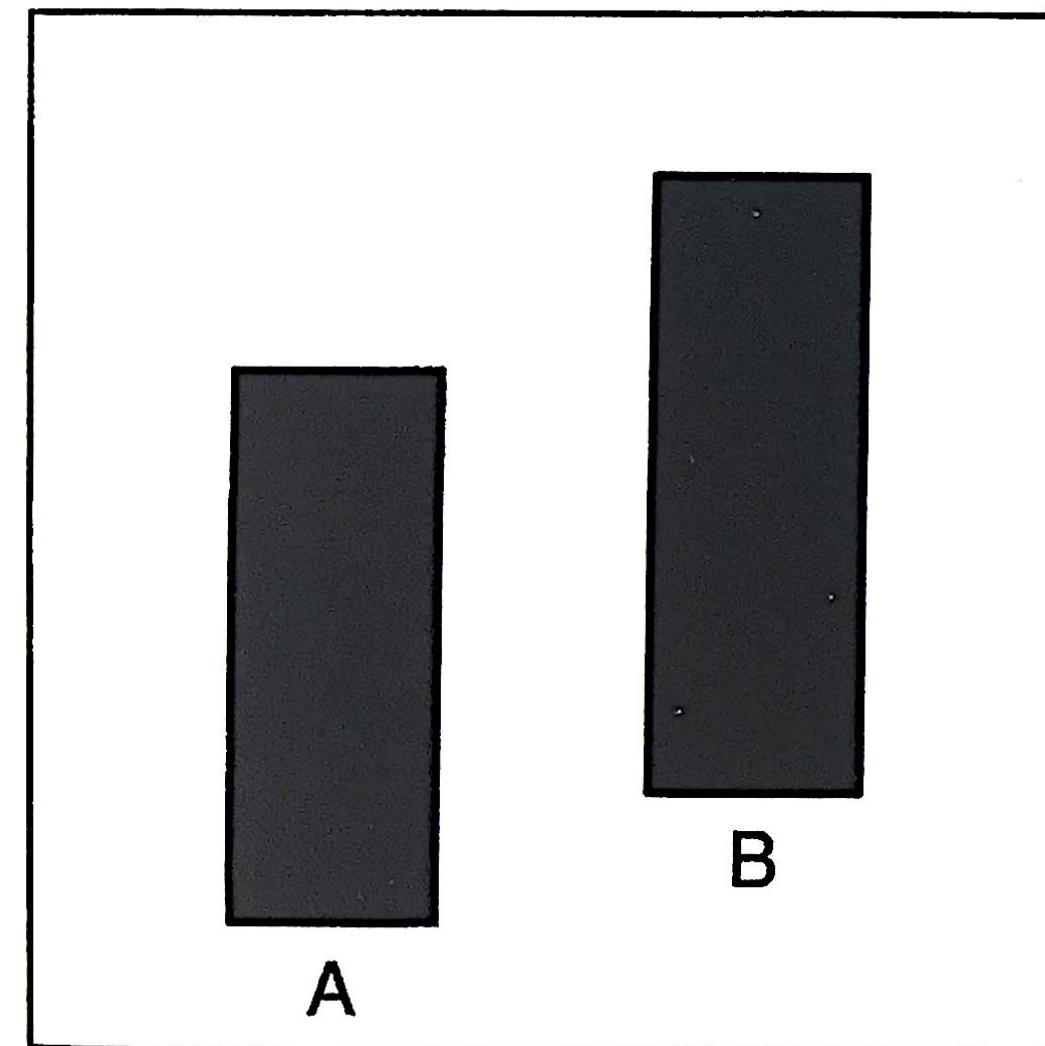
Our visual perception of differences can be stated mathematically, as Weber's Law.

If x is the magnitude of a physical attribute, say, length of a line segment, and $w_p(x)$ is a positive number such that a line of length

$$x + w_p(x)$$

is discriminated with probability p to be longer than the line of length x then,

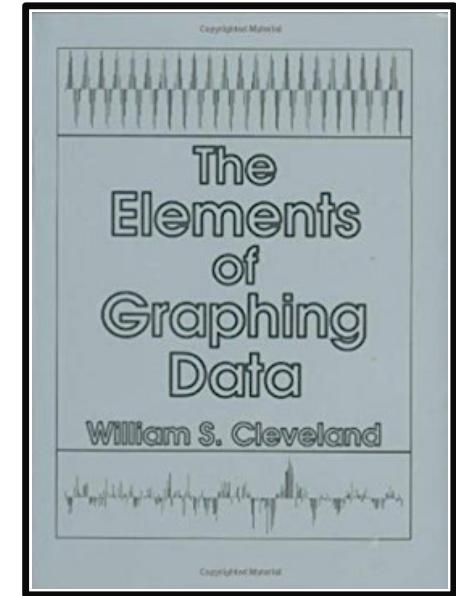
$$w_p(x) = k_p x$$



The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



Consider (only) using reference grids in aid of Weber's Law

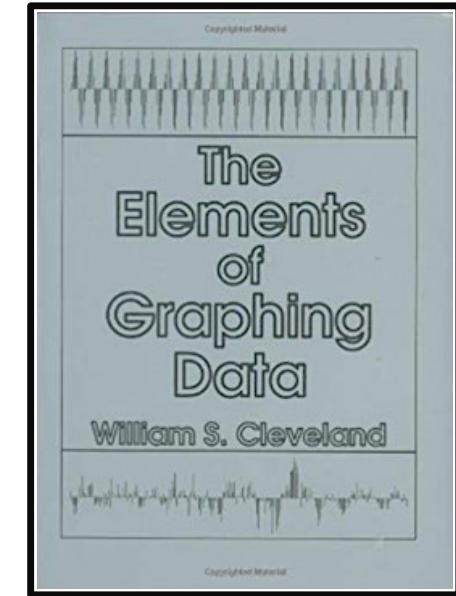
Reference grids can help us perceive the differences between two measurements, but they can also distract. Consider only using reference lines that guide your audience to the relevant point of your message.

And a little goes a long way: for any needed reference lines, consider making them, say, very faint gray or otherwise distinguish them from the data.

The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



Consider (only) using reference grids in aid of Weber's Law

Reference grids can help us perceive the differences between two measurements, but they can also distract. Consider only using reference lines that guide your audience to the relevant point of your message.

And a little goes a long way: for any needed reference lines, consider making them, say, very faint gray or otherwise distinguish them from the data.

How did we do in the example proposal for the Dodgers?

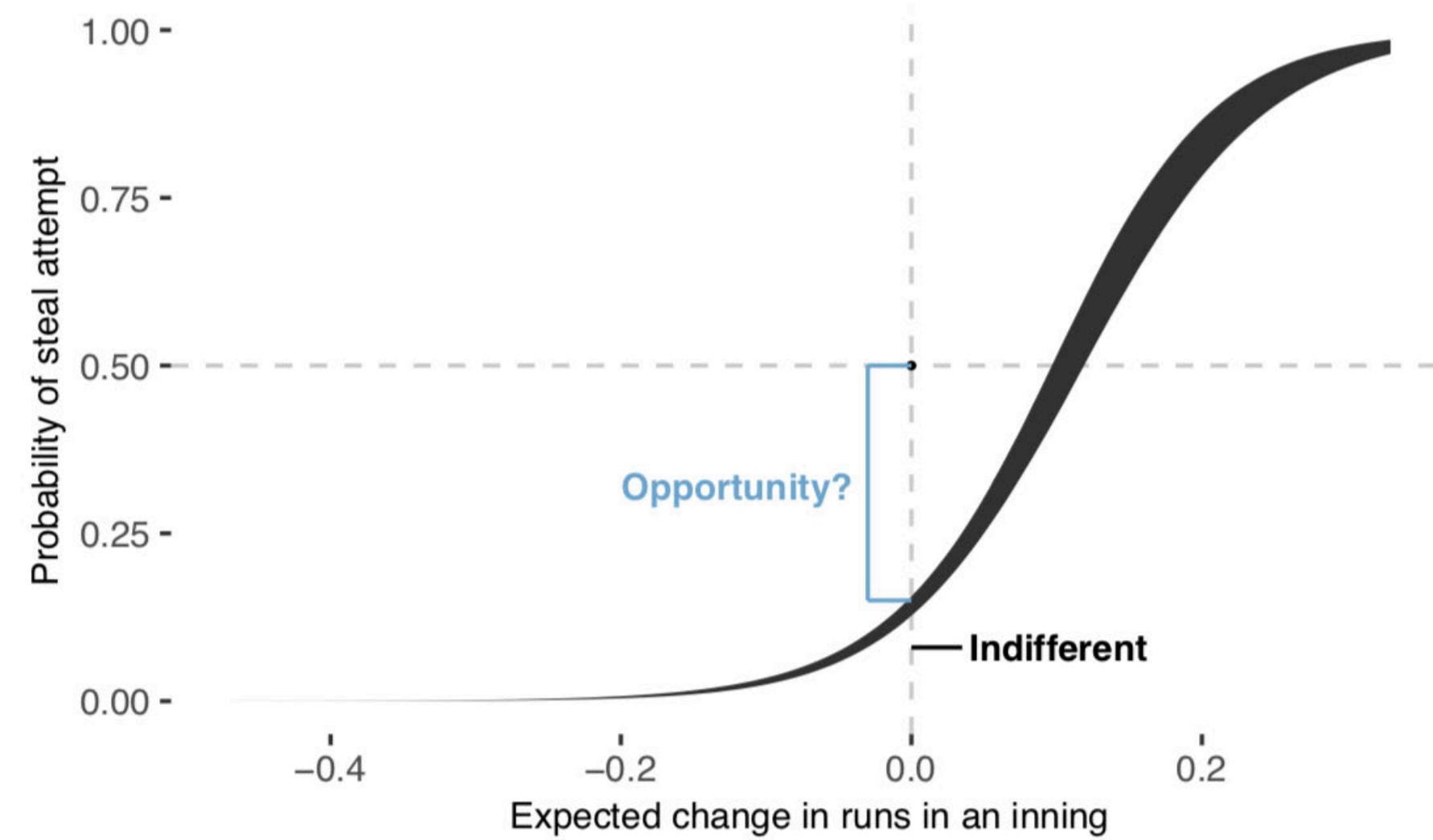


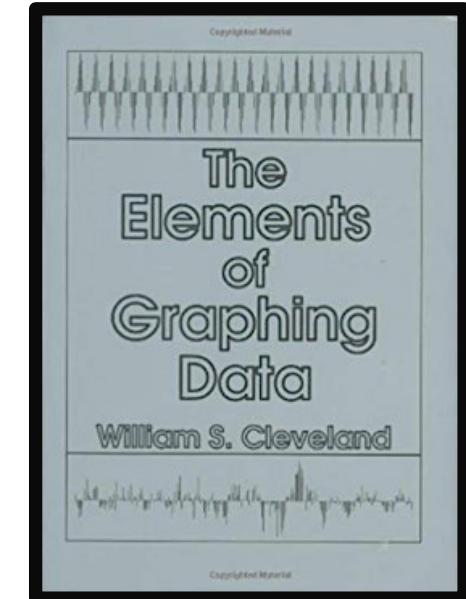
Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only **10 percent** of the time, leaving opportunity.

The Elements of Graphing Data

Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



Consider (only) using reference grids in aid of Weber's Law

Reference grids can help us perceive the differences between two measurements, but they can also distract. Consider only using reference lines that guide your audience to the relevant point of your message.

And a little goes a long way: for any needed reference lines, consider making them, say, very faint gray or otherwise distinguish them from the data.

How did we do in the example proposal for the Dodgers?

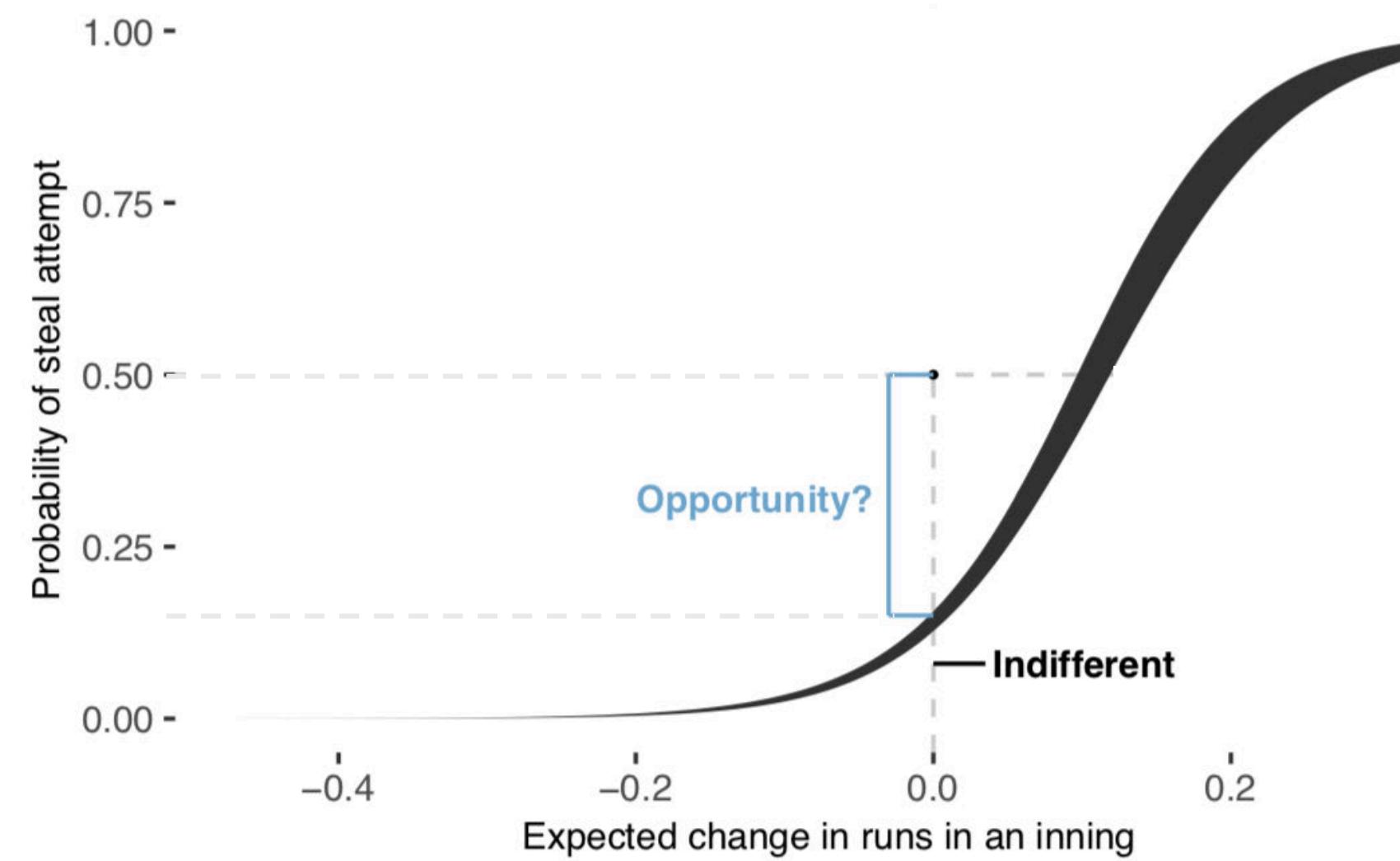


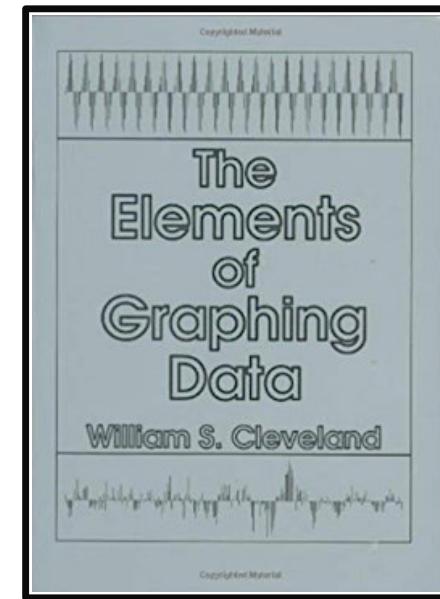
Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only **10 percent** of the time, leaving opportunity.

The Elements of Graphing Data

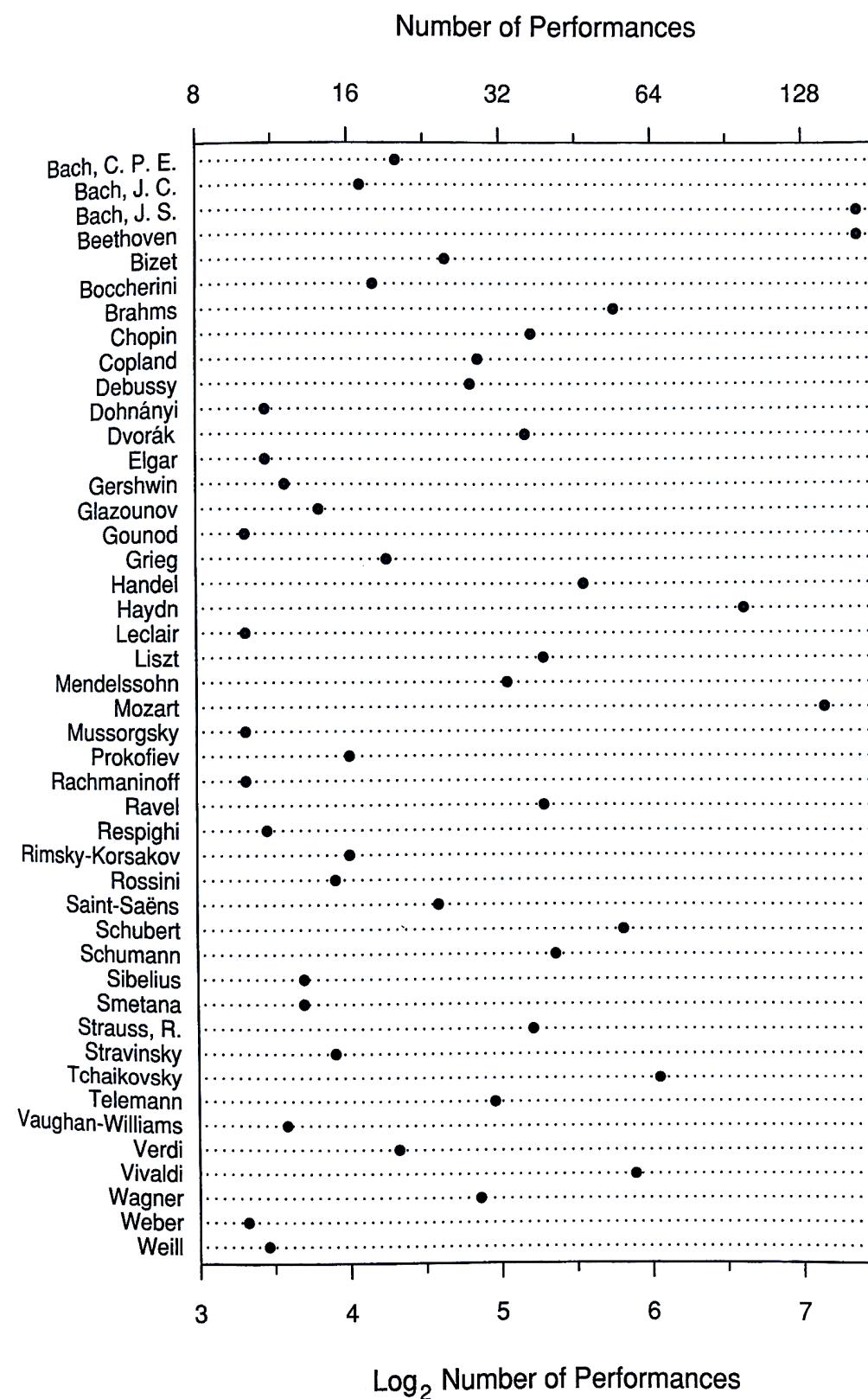
Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



Ordering for categorical variables substantially affects our visual decoding

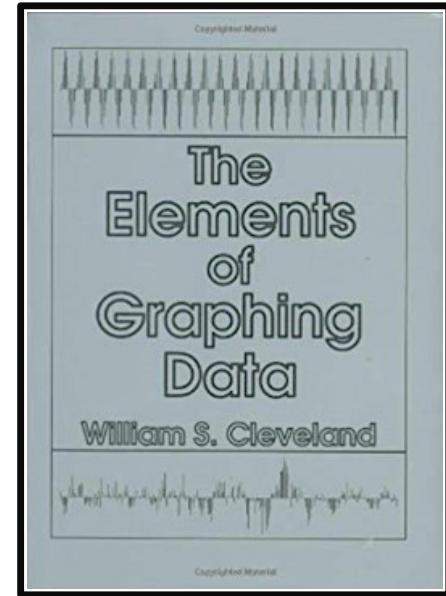
Ordering from, say, smallest to largest enhances our visual decoding of the distribution of values along the measurement scale.



The Elements of Graphing Data

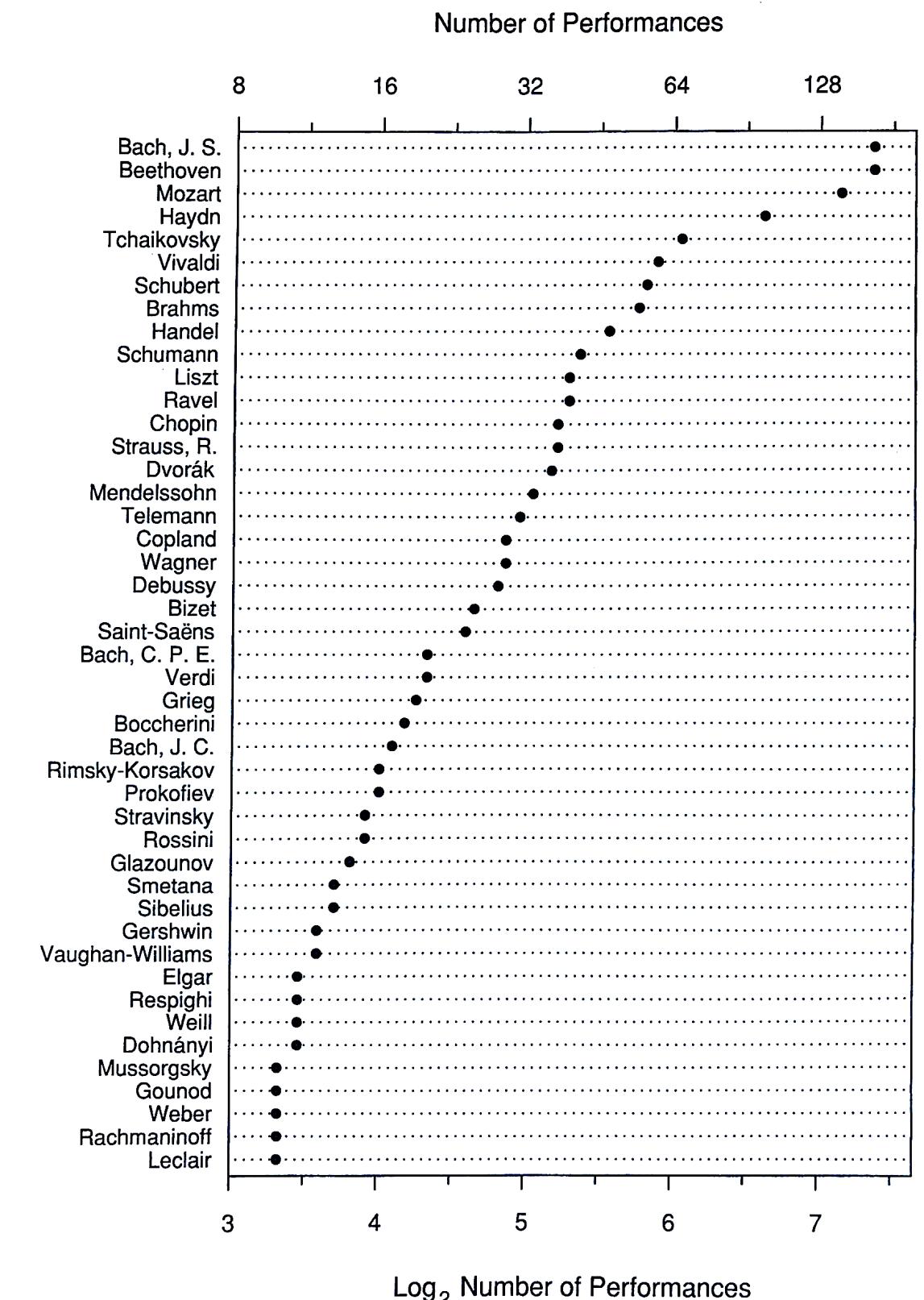
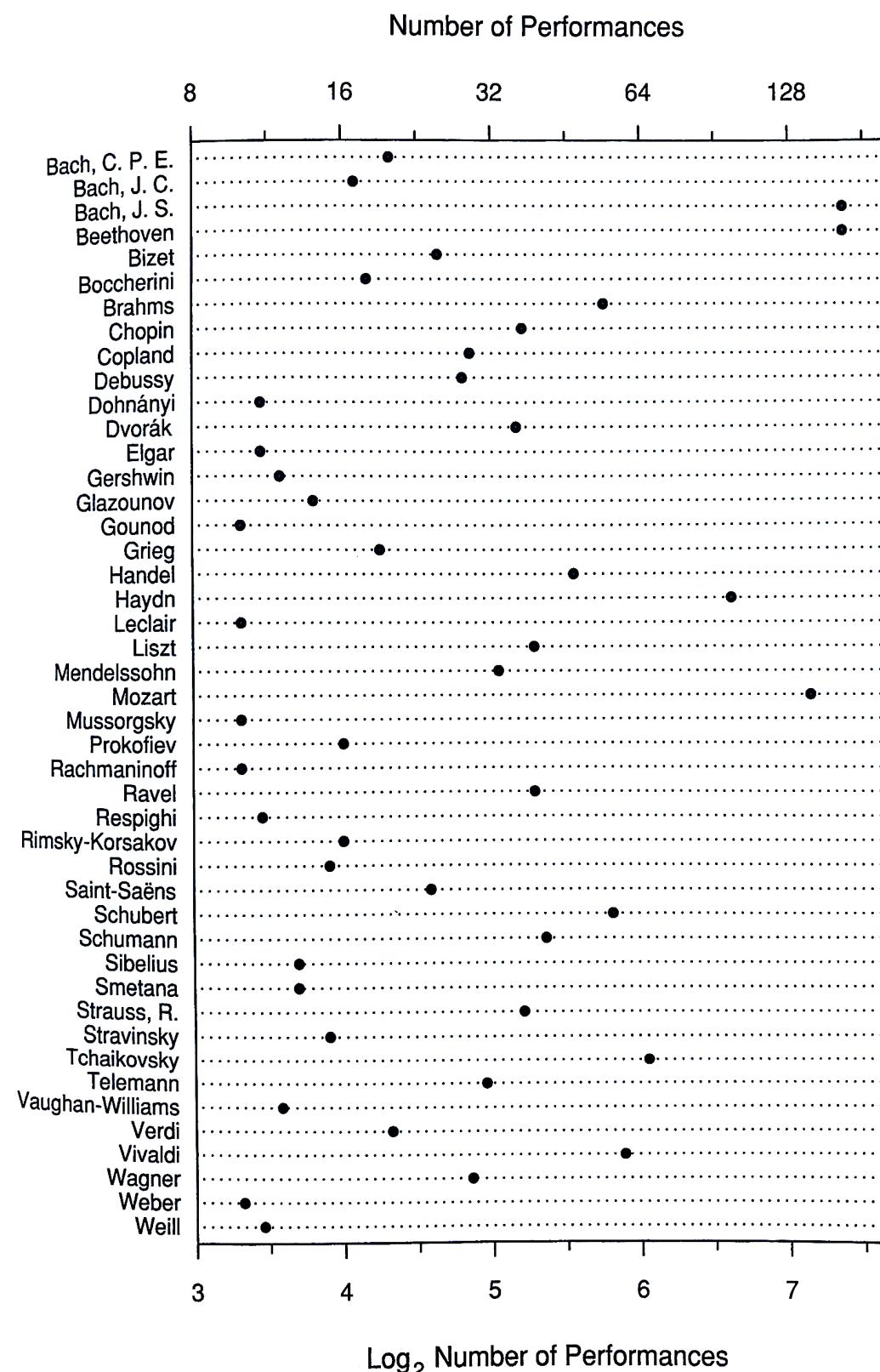
Cleveland

A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



Ordering for categorical variables substantially affects our visual decoding

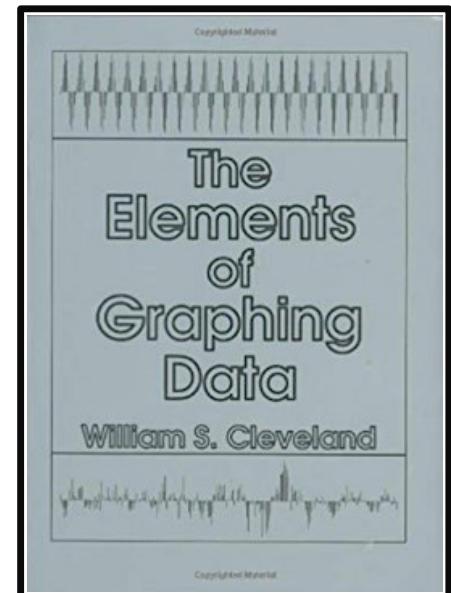
Ordering from, say, smallest to largest enhances our visual decoding of the distribution of values along the measurement scale.



The Elements of Graphing Data

Cleveland

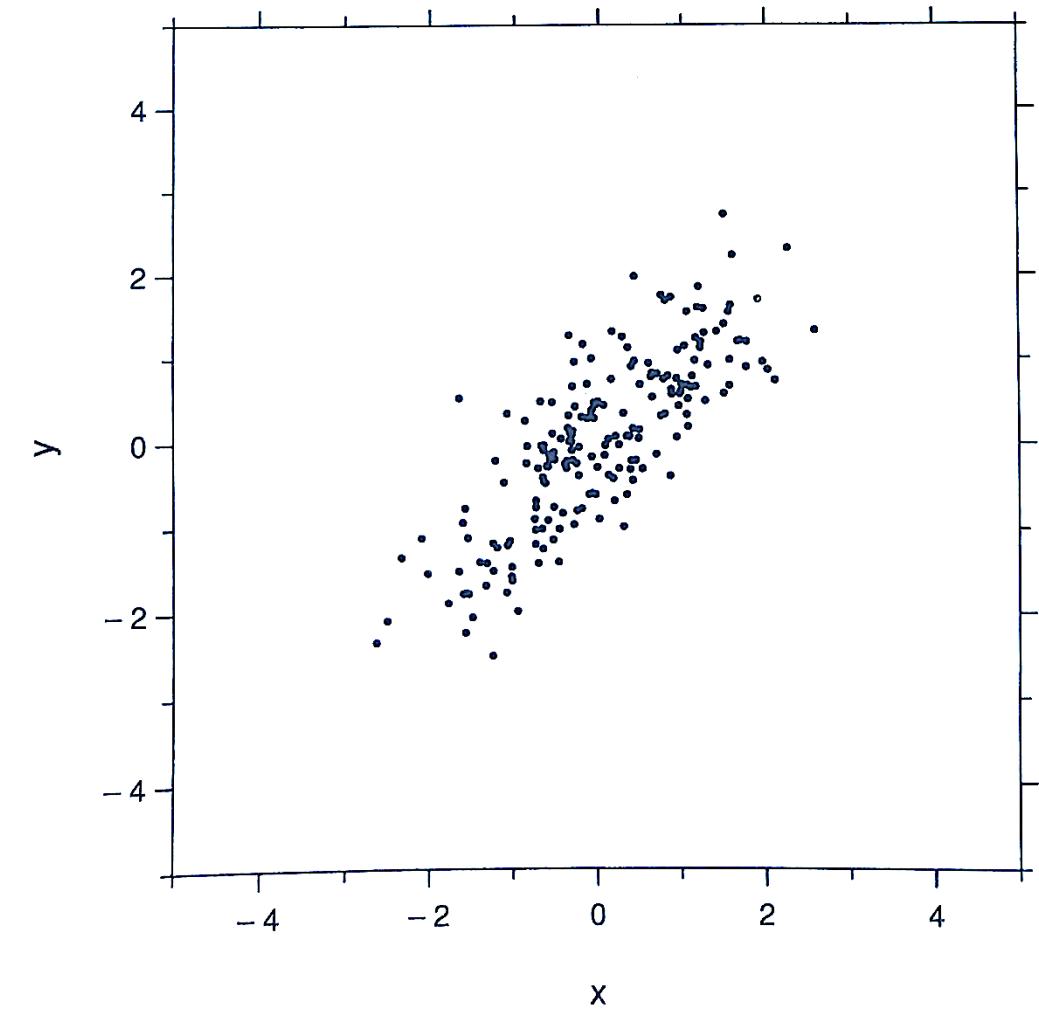
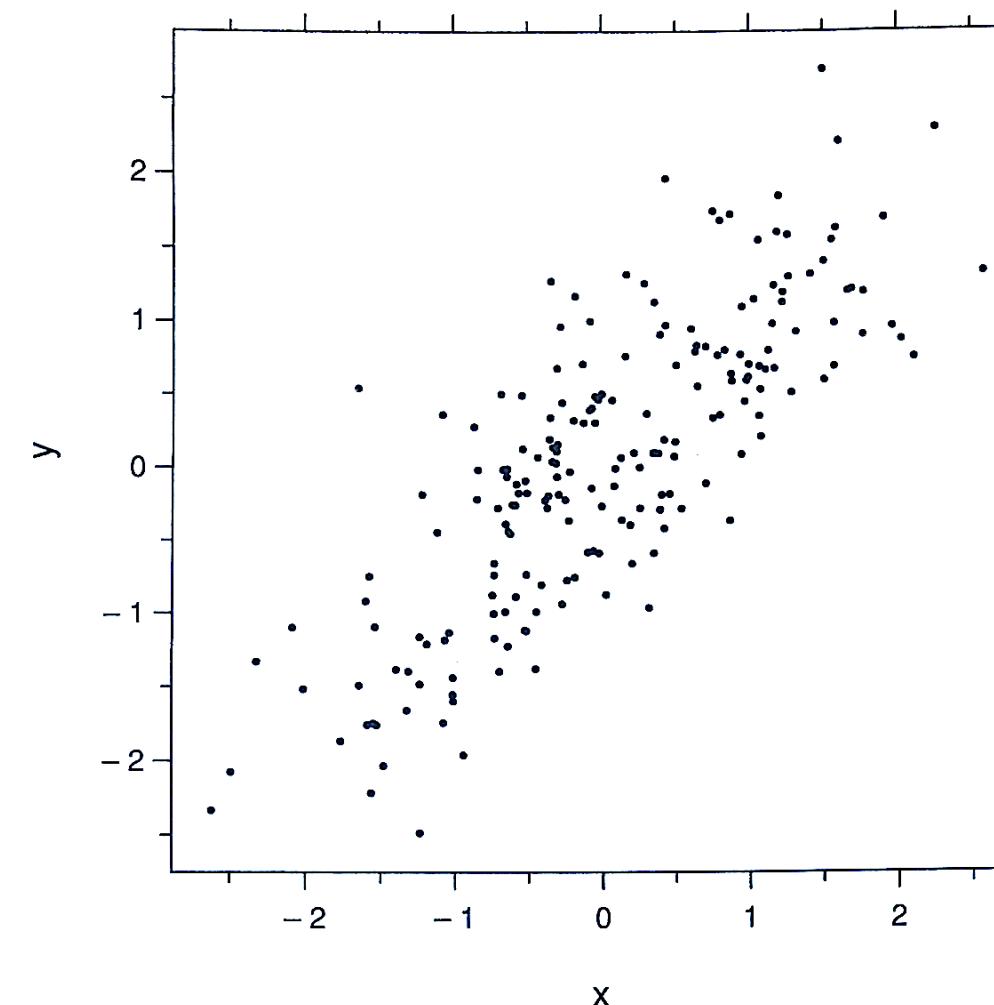
A graduate of Princeton, William is a computer scientist and Professor of Statistics and Professor of Computer Science at Purdue University, known for his work on data visualization, particularly on nonparametric regression and local regression.



Perceived correlation depends on ratio of plot area to data area

Our estimation of correlation is affected by the area of the data rectangle divided by the area of the scale-line rectangle.

Showing the same data, the left panel displays a 1 to 1 ratio, while in the right panel displays the data rectangle as much smaller than the scale-line rectangle.

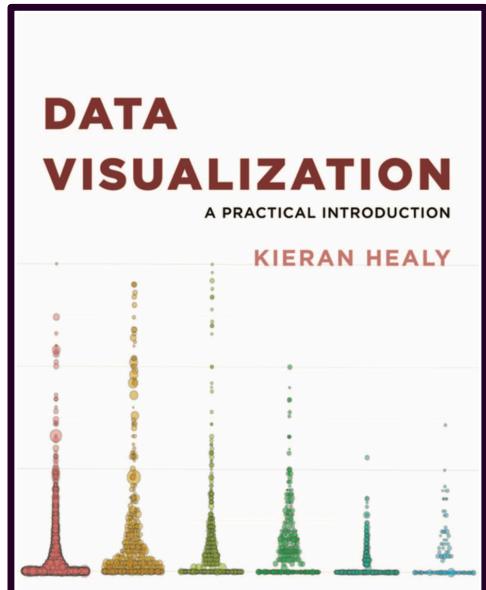


Focusing visual attention

Data Visualization: a practical introduction

Healy

A PhD graduate from Princeton, Kieran is associate professor of sociology at Duke University. His book has been described as “covering the ‘why do’ as well as the ‘how to’ of data visualization.” — Andrew Gelman



Help your audience with Gestault principles

Our eyes automatically search for (grouping), difference and change.

Proximity: Things that are spatially near to one another seem to be related.

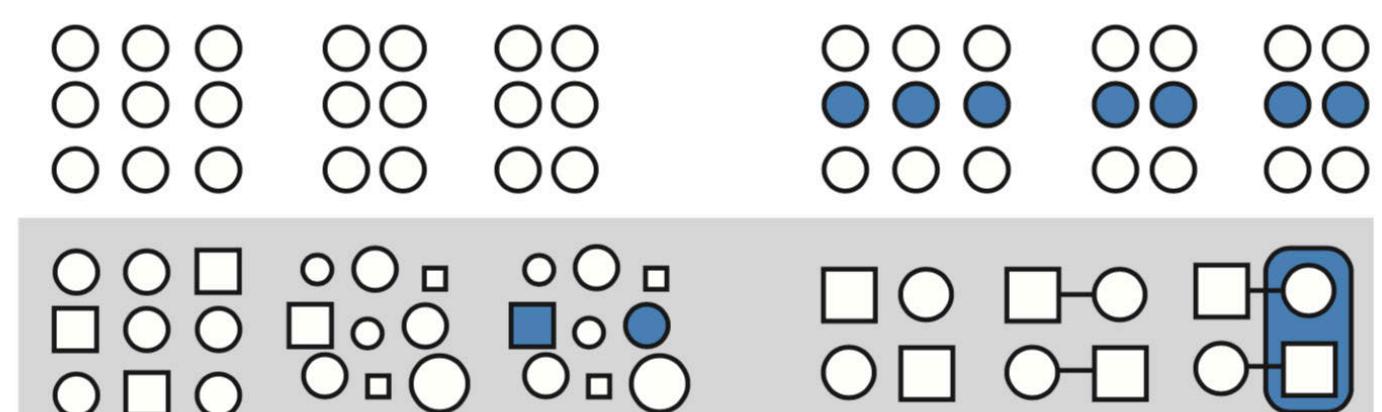
Similarity: Things that look alike seem to be related.

Connection: Things visually tied together seem to be related.

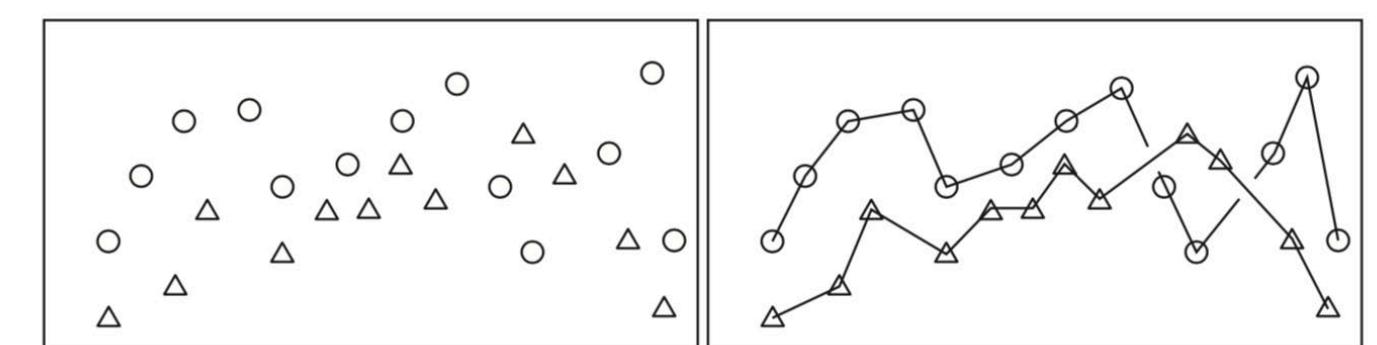
Continuity: Partially hidden objects are completed into familiar shapes.

Closure: Incomplete shapes are perceived as complete.

Figure and ground: Visual elements are taken to be either in the foreground or in the background.

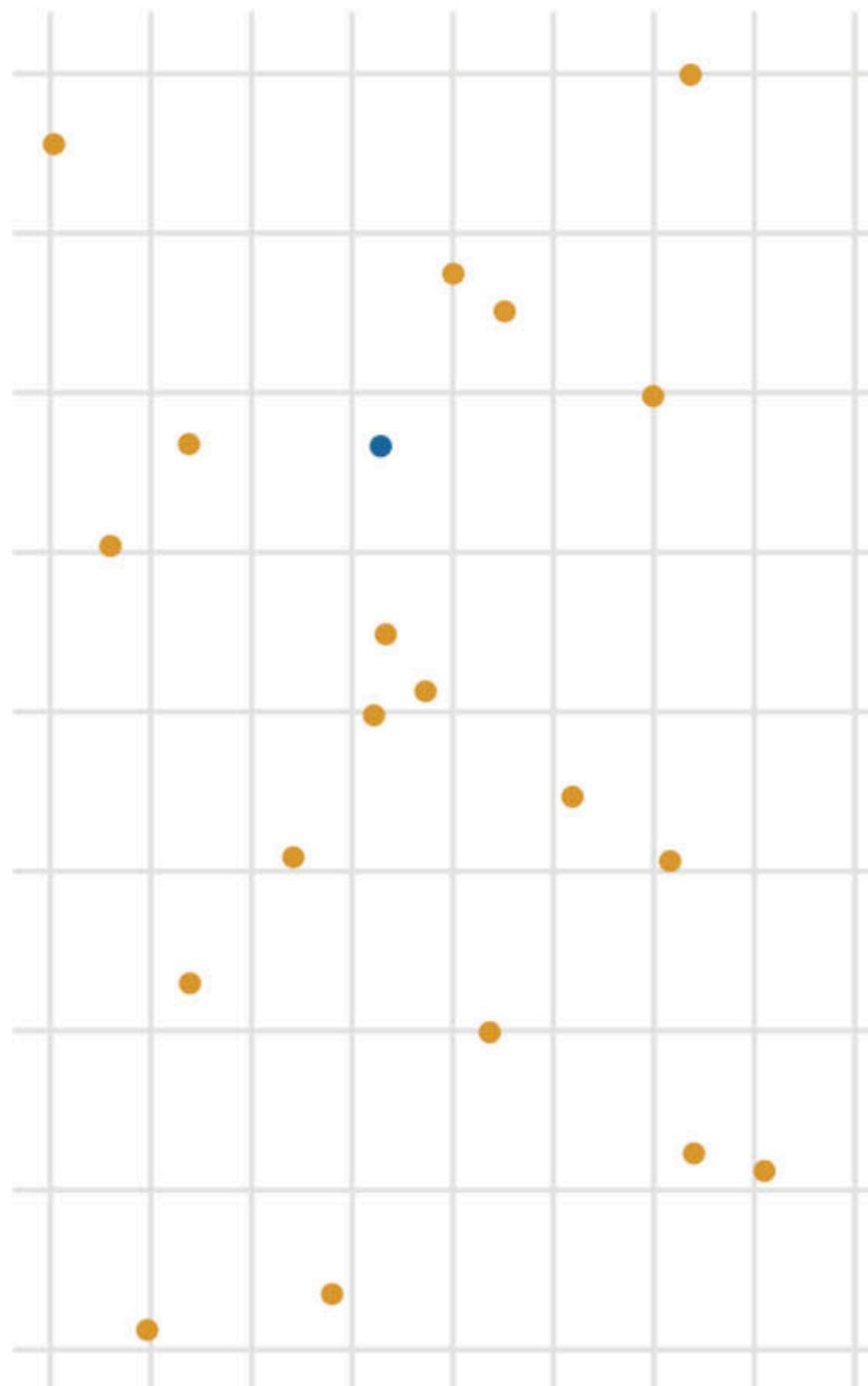


Common fate: Elements sharing a direction of movement are perceived as a unit.

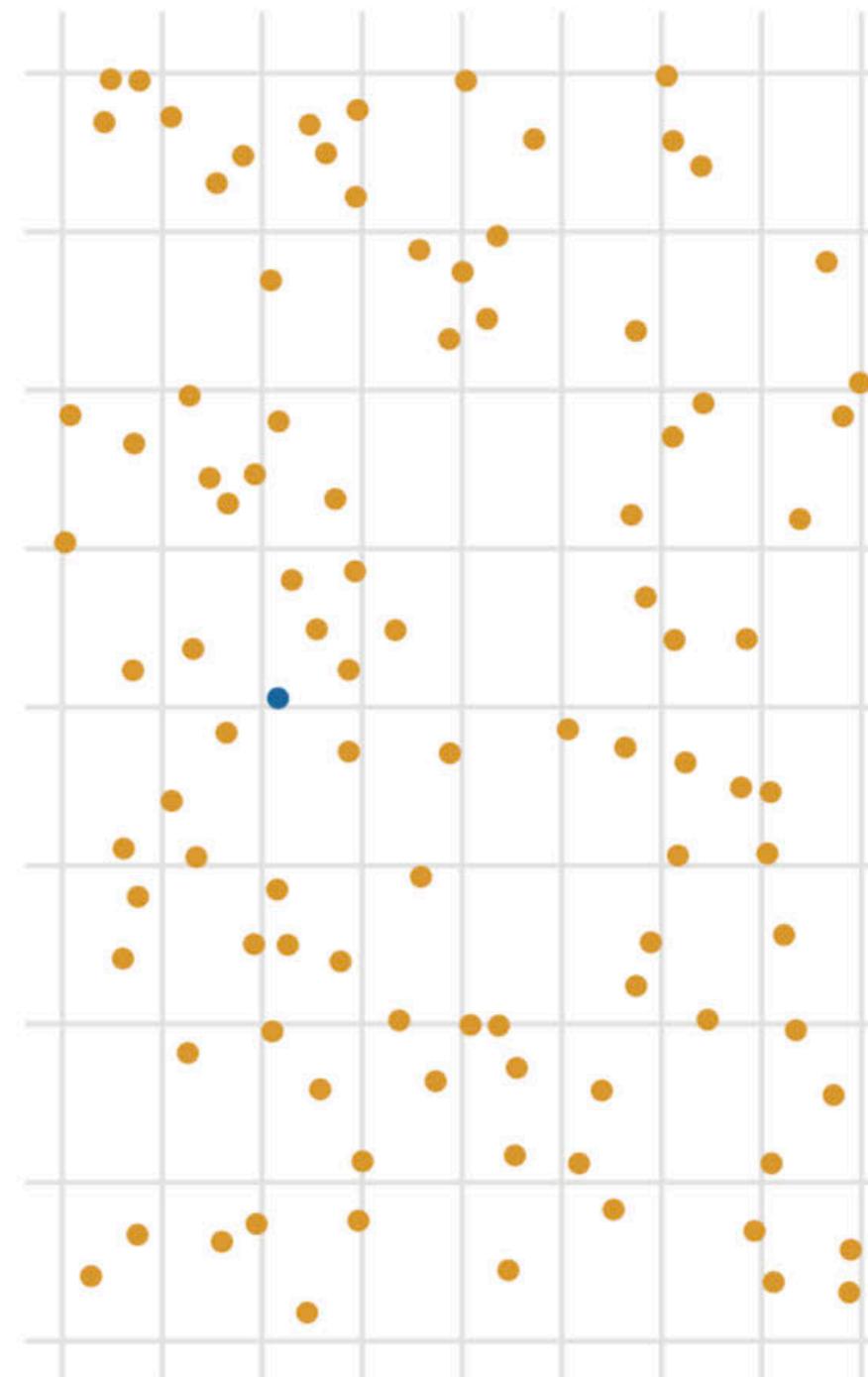


Consider Gestault principles when trying—as an audience—to find the blue circle.

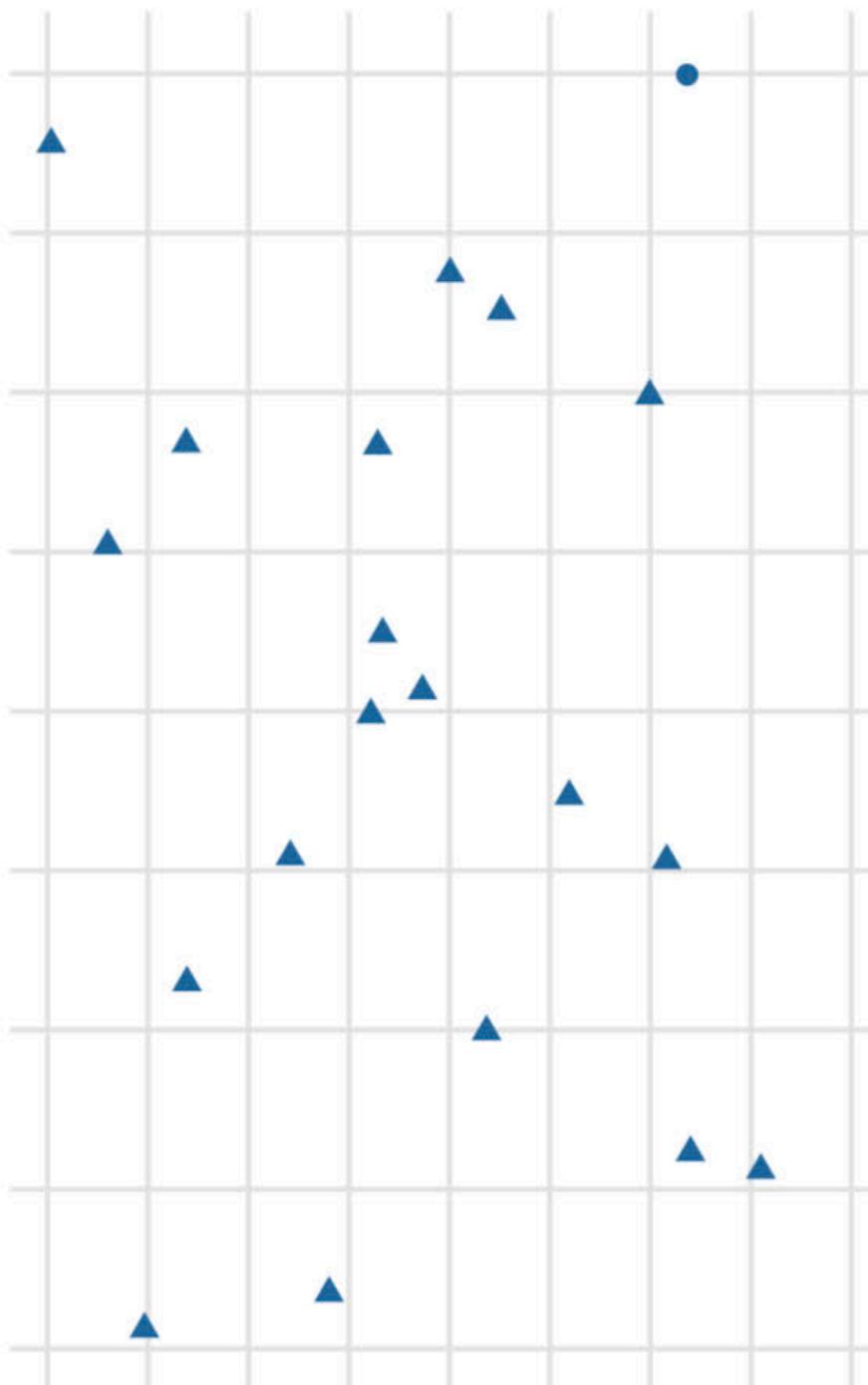
Color only, $N = 20$



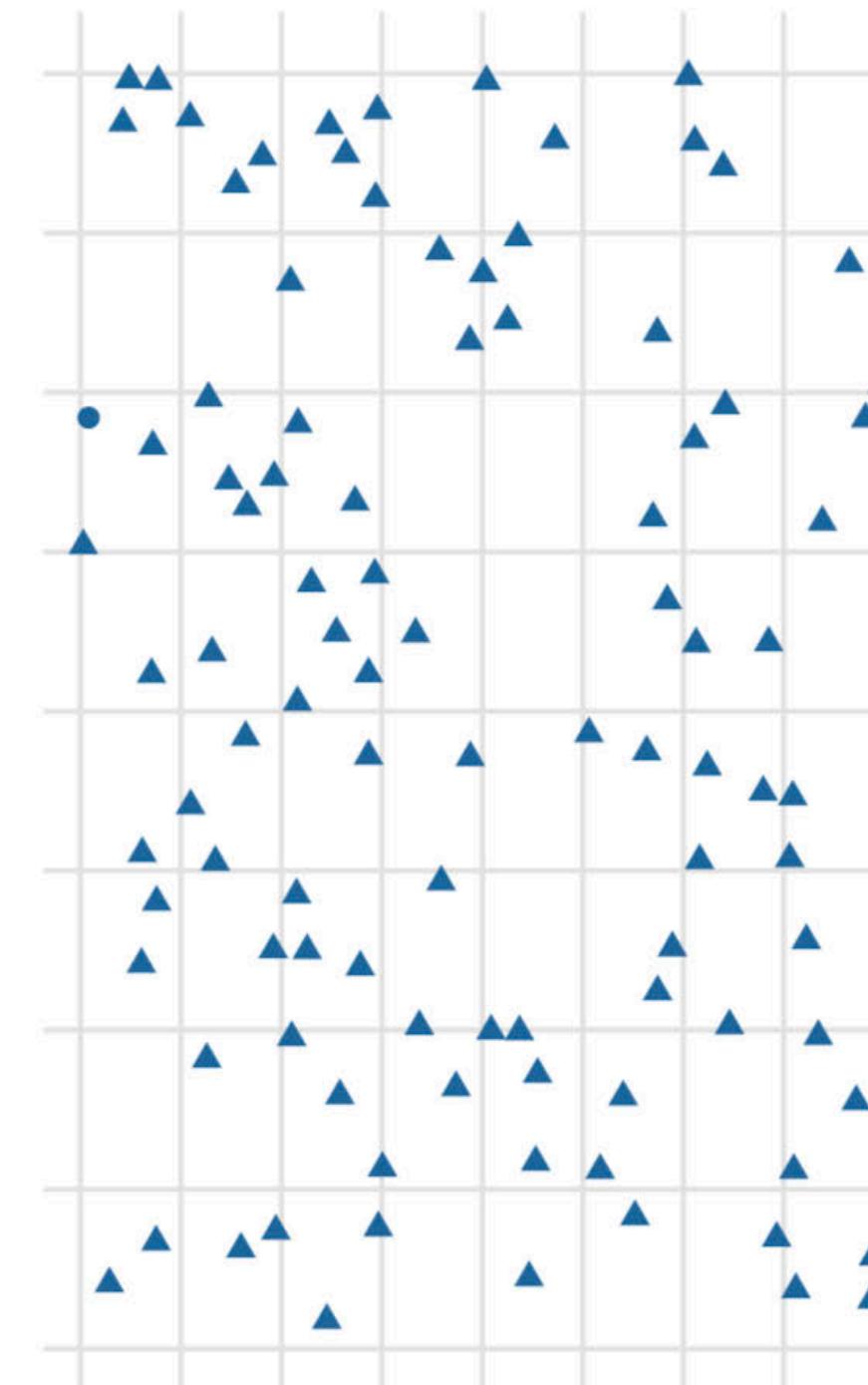
Color only, $N = 100$



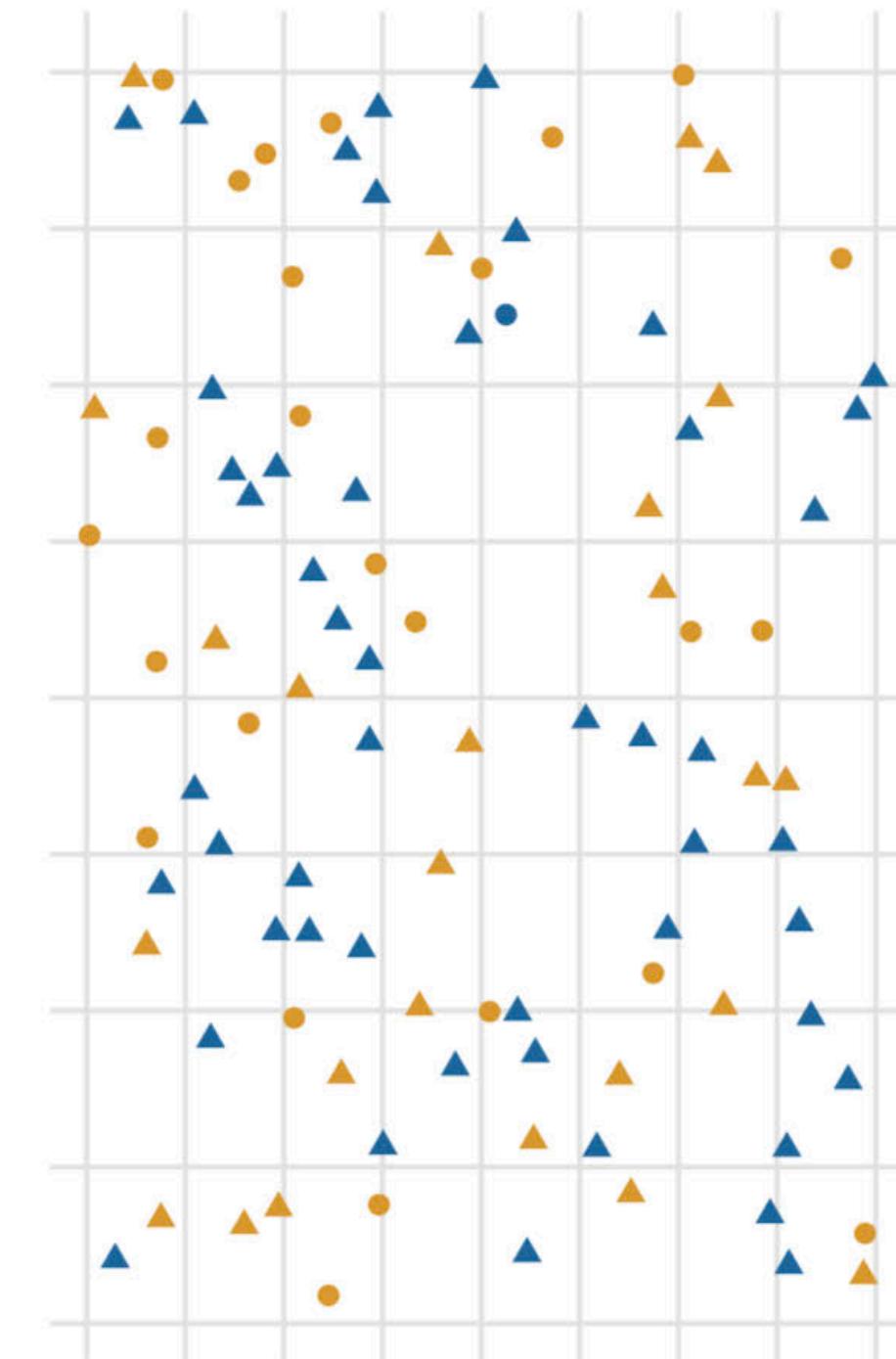
Shape only, $N = 20$



Shape only, $N = 100$

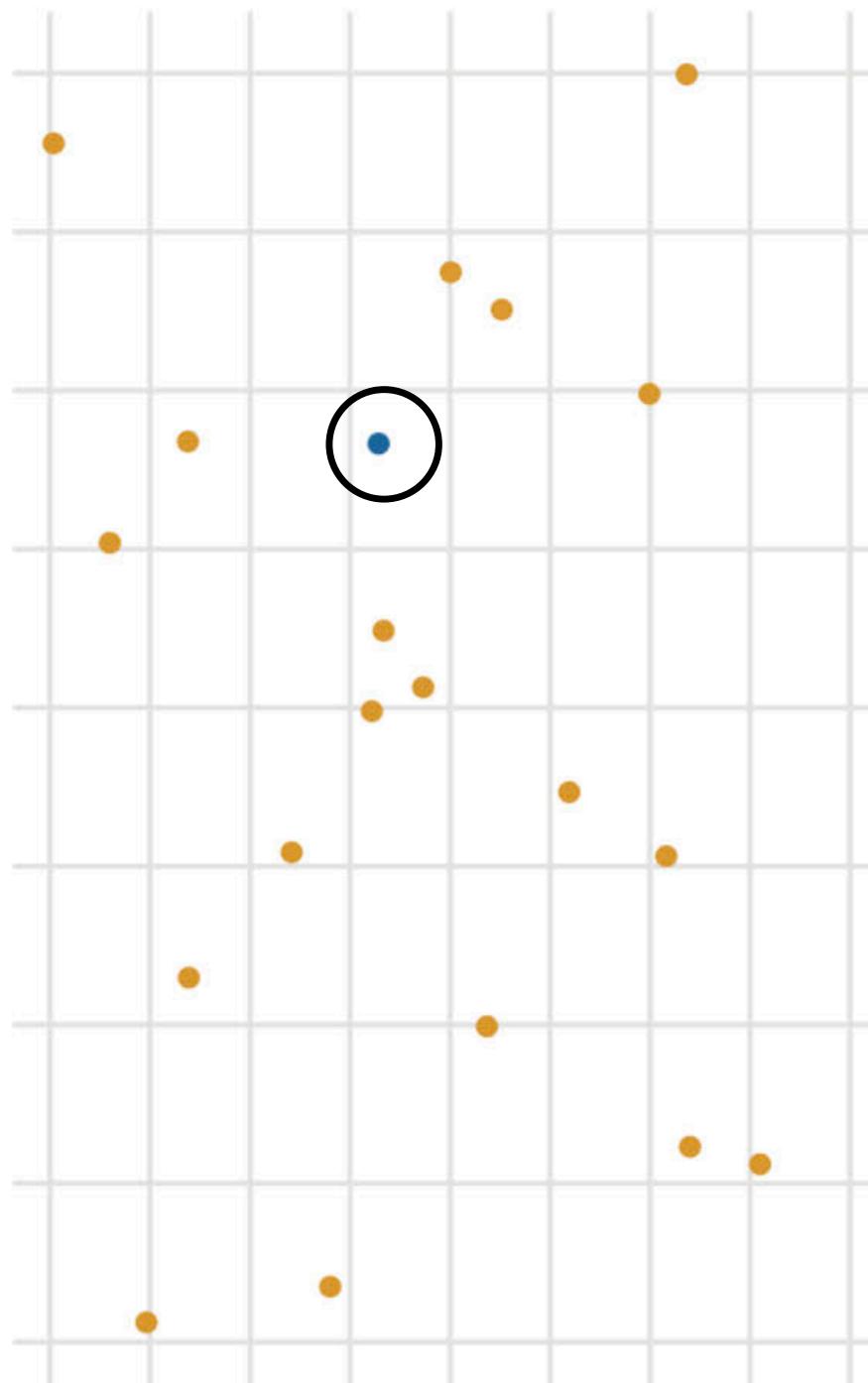


Color & shape, $N = 100$

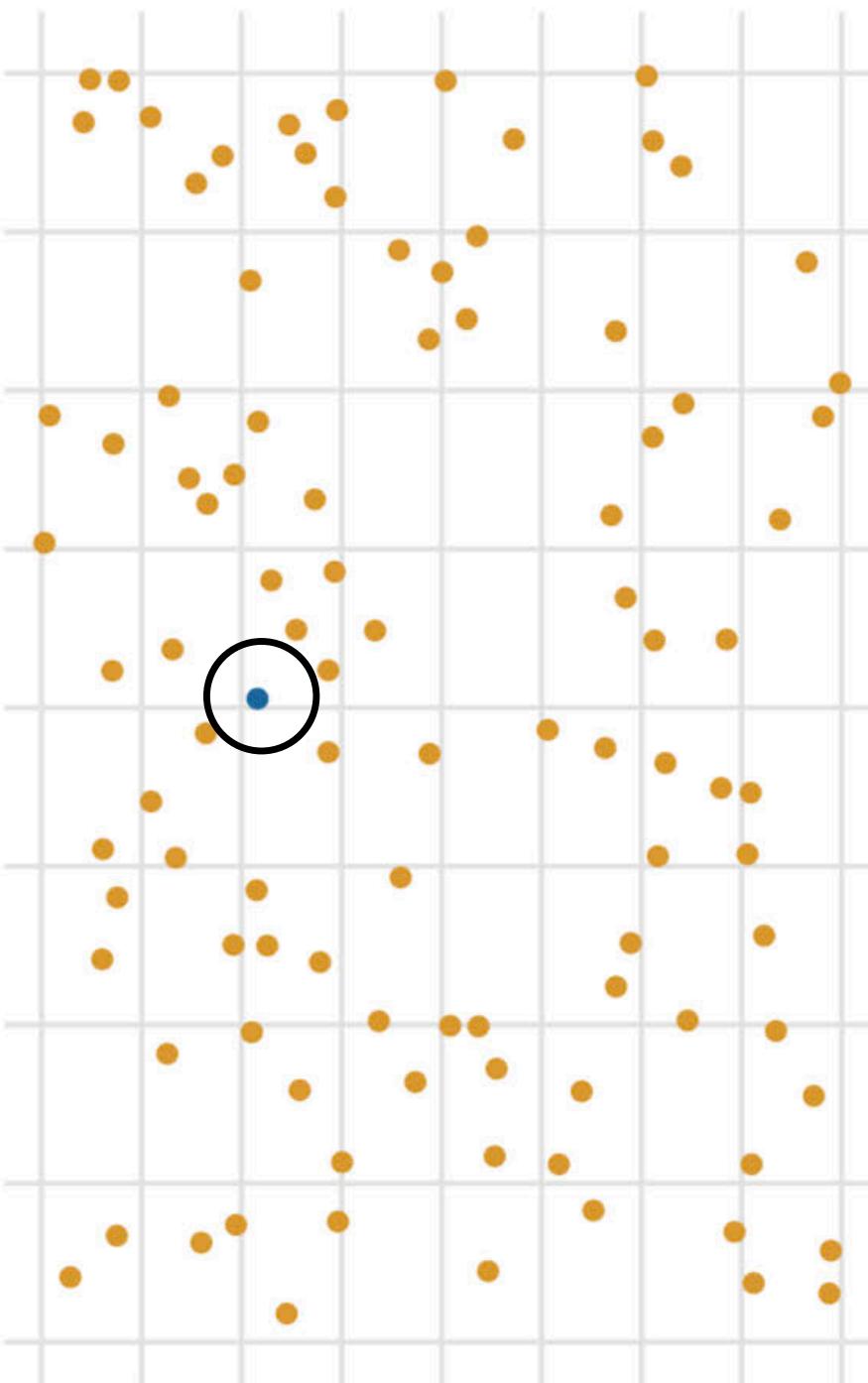


For these, another choice (enclosure) is more effective, which draw the eye instantly.

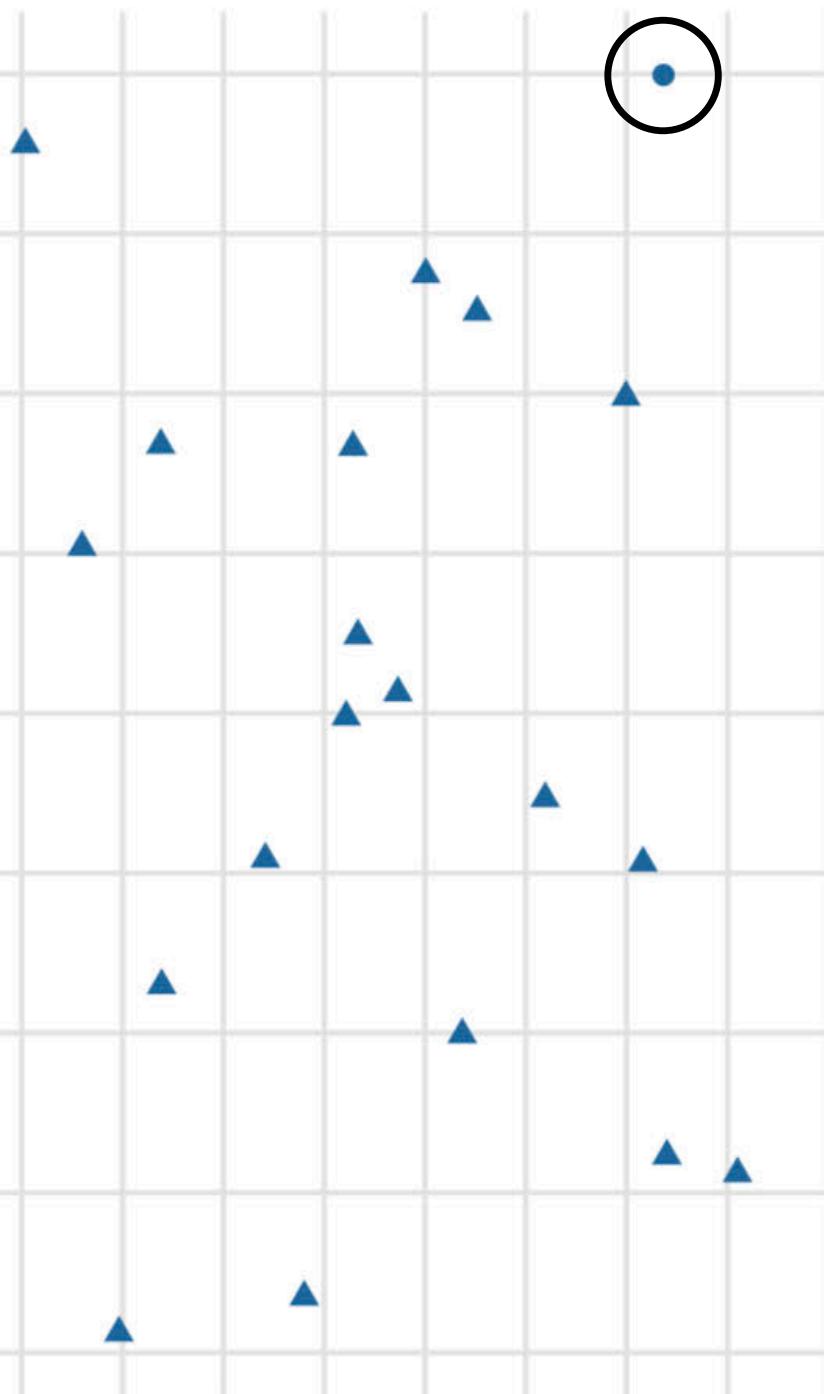
Color only, $N = 20$



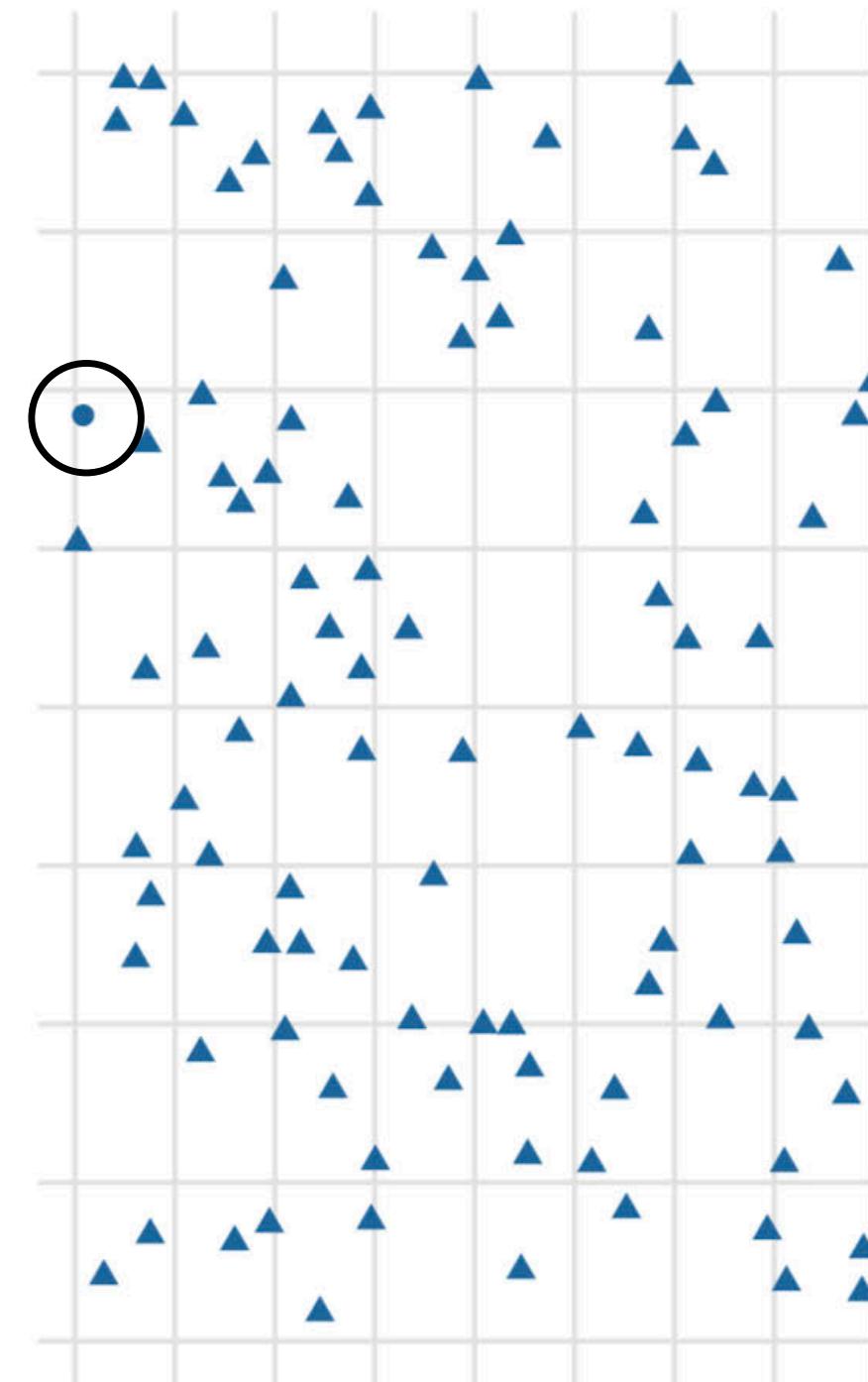
Color only, $N = 100$



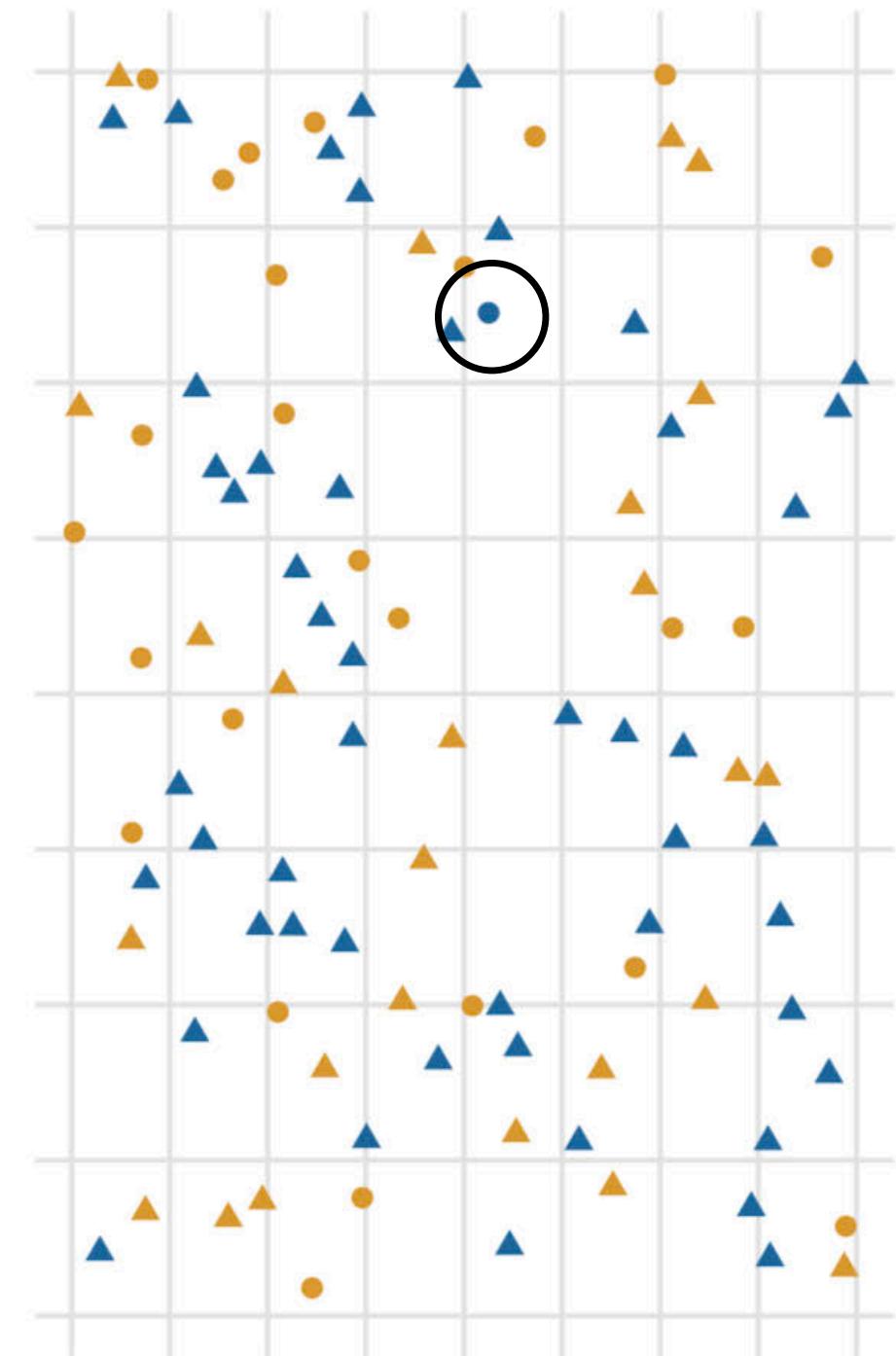
Shape only, $N = 20$



Shape only, $N = 100$



Color & shape, $N = 100$

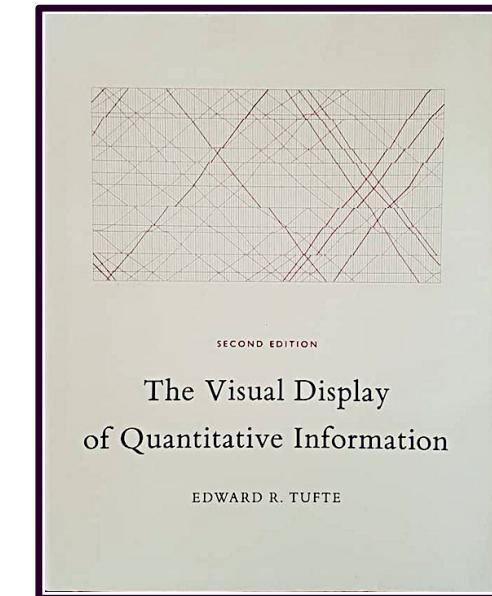
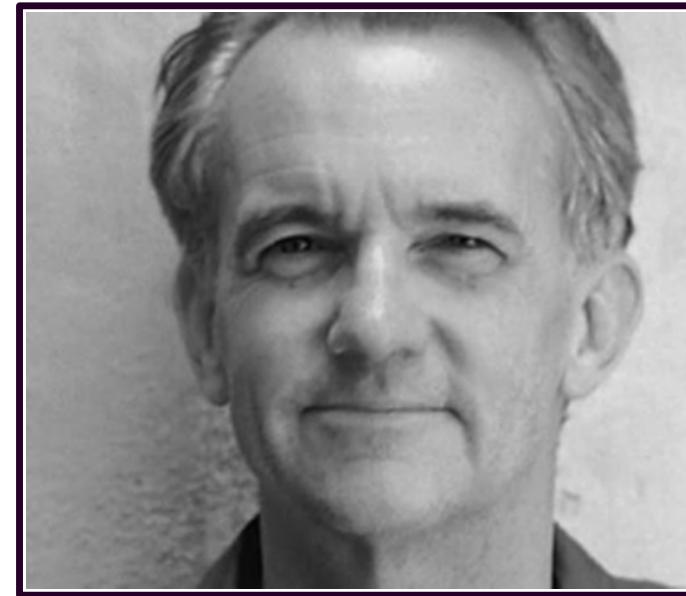


From default visual to visual narrative

The Visual Display of Quantitative Information

Tufte

Hailed "The Leonardo da Vinci of data" by the New York Times. He is professor emeritus of Political Science, Statistics, and Computer Science at Yale University.



Simplicity of design, complexity of data

Words and pictures belong together

Proportion and scale: the shape of graphics

Graphical excellence is often found in simplicity of design and complexity of data.

Viewers need the help that words can provide. **Words on graphics are data-ink**, making effective use of the space freed up by erasing redundant and non-data-ink.

Note, the **size of type** on and around graphics can be quite **small**, since the phrases and sentences are usually not too long.

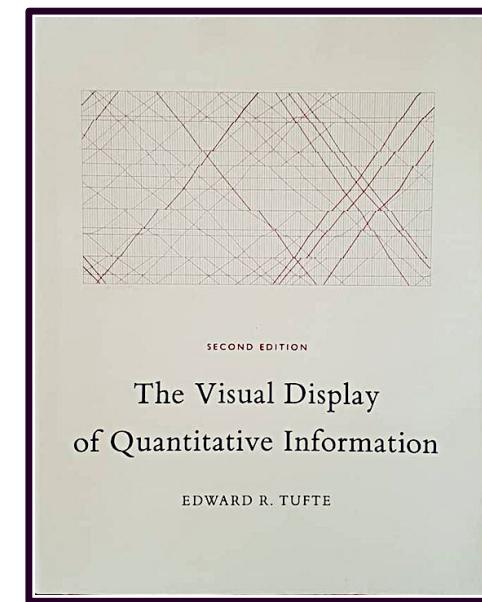
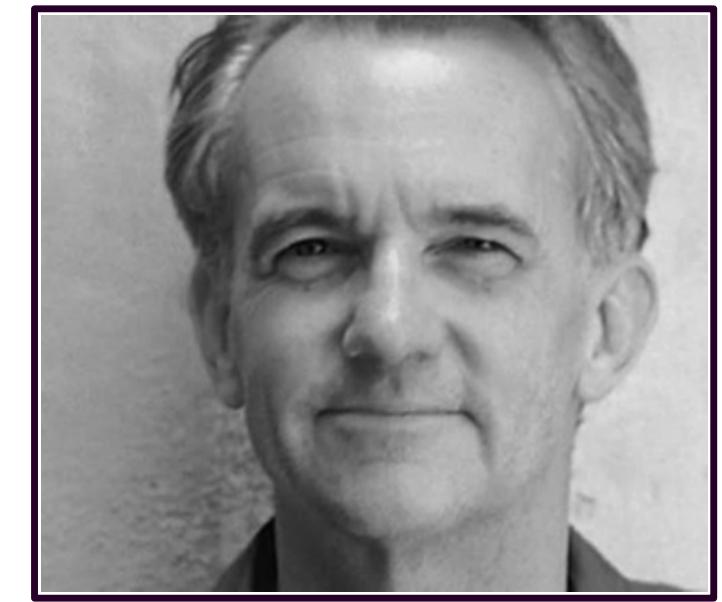
The principle of data/text integration is: data **graphics are paragraphs about data** and should be treated as such.

Our eye is naturally practiced in detecting **deviations from the horizon**.

Horizontally shaped plots tend to make it **easier to directly label** and explain the data.

Tradition places the effect vertically, the cause horizontally.

The empirically studied **Golden Rectangle**, a 1.0×1.618 ratio is aesthetically pleasing.



The Visual Display of Quantitative Information

Tufte

Hailed "The Leonardo da Vinci of data" by the New York Times. He is professor emeritus of Political Science, Statistics, and Computer Science at Yale University.

Data-ink

The non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented. Annotations are part of the data-ink.

$$\text{data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

= proportion of a graphic's ink devoted to the non-redundant display of data-information

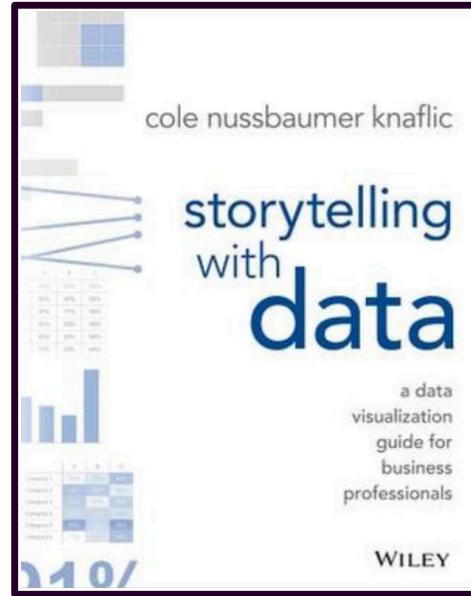
= $1.0 - \text{proportion of a graphic that can be erased without loss of data-information}$

Maximize within reason

Maximize the data-ink ratio, within reason.
Erase non-data-ink, within reason.
Erase redundant data-ink, within reason.

Size graphics for legibility

As the quantity of data increases, data measures must shrink. . . . The way to increase data density other than by enlarging the data matrix is to reduce the area of a graphic.
Graphics can be shrunk way down.



Storytelling with data

Knafllic

The author is a consultant focused on visual displays. Her experience arose from human resources in Google where she applied theory learned as a student of Yale's Edward Tufte.

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

Discussion: what differences do you see? What advice has she applied?

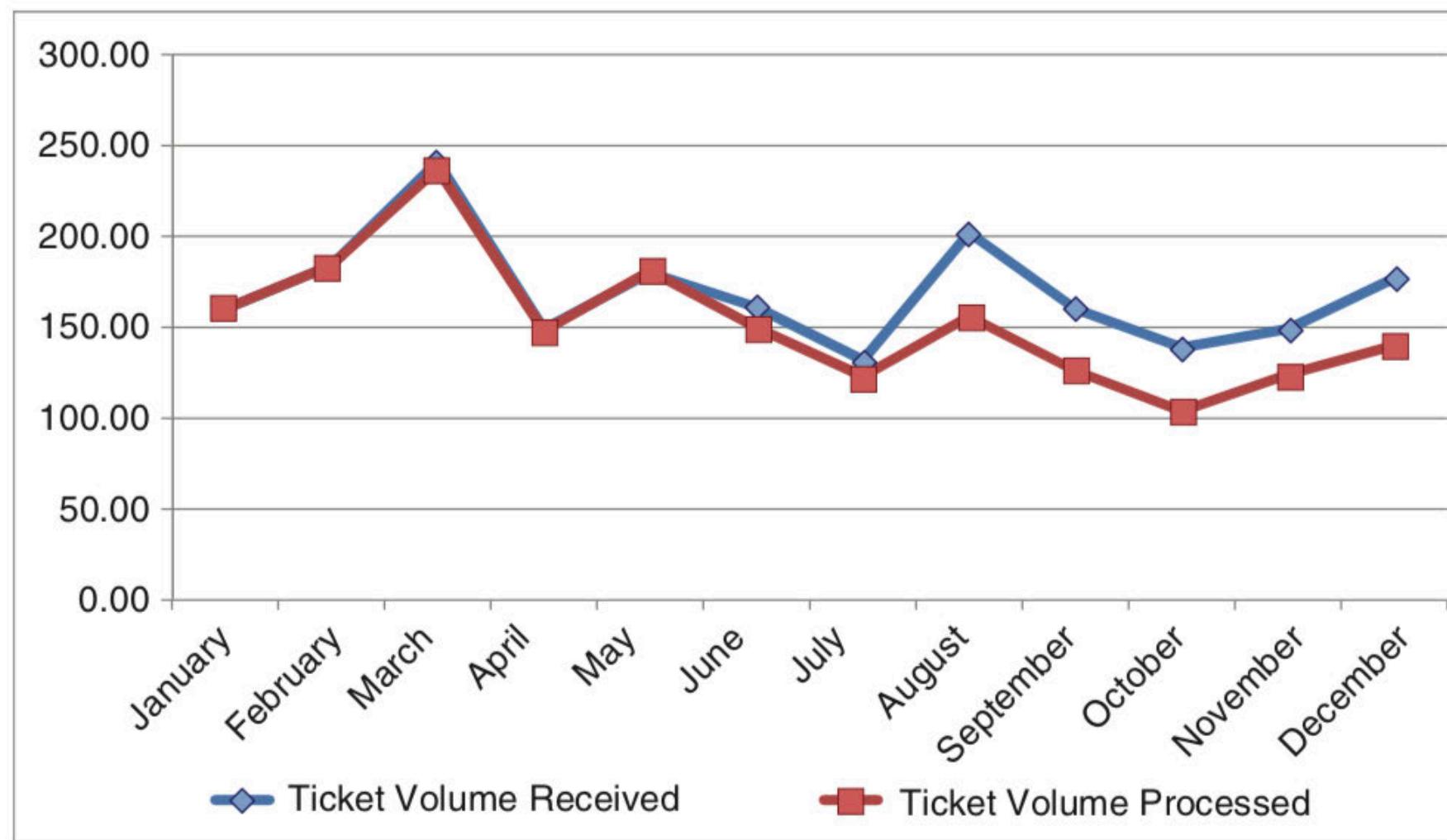
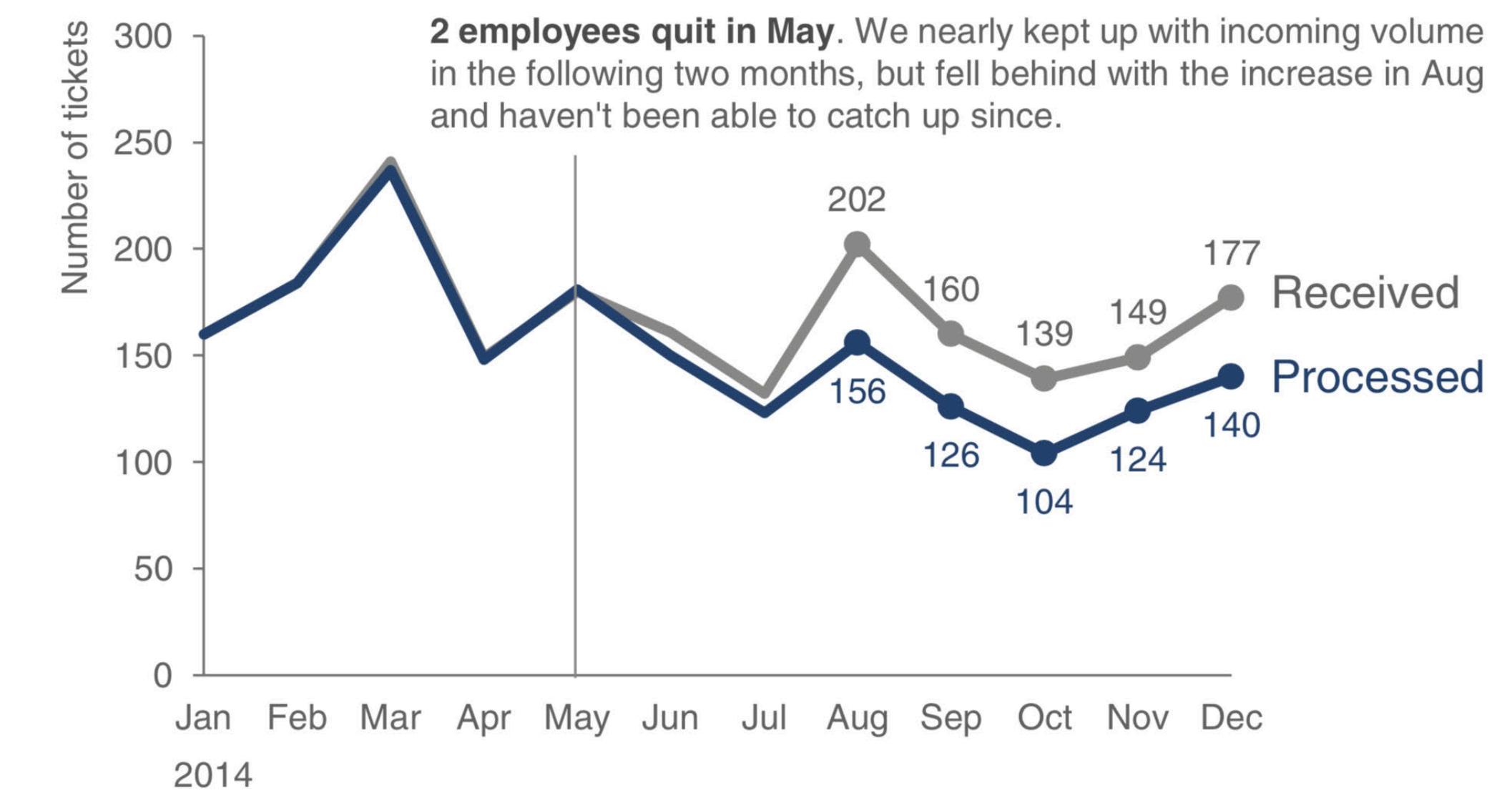


FIGURE 3.17 Original graph

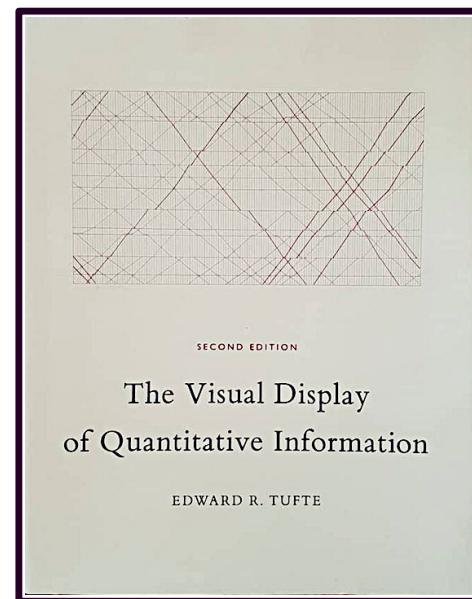
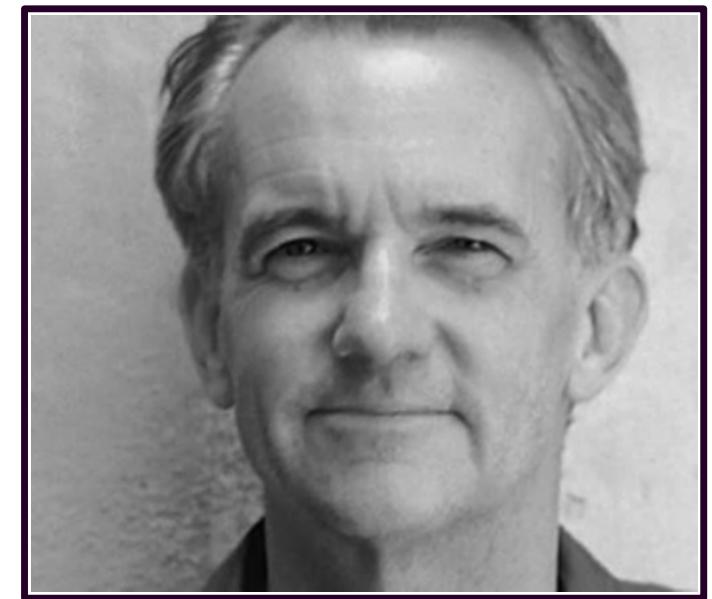
Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.



The Visual Display of Quantitative Information

Tufte

Hailed "The Leonardo da Vinci of data" by the New York Times. He is professor emeritus of Political Science, Statistics, and Computer Science at Yale University.

“

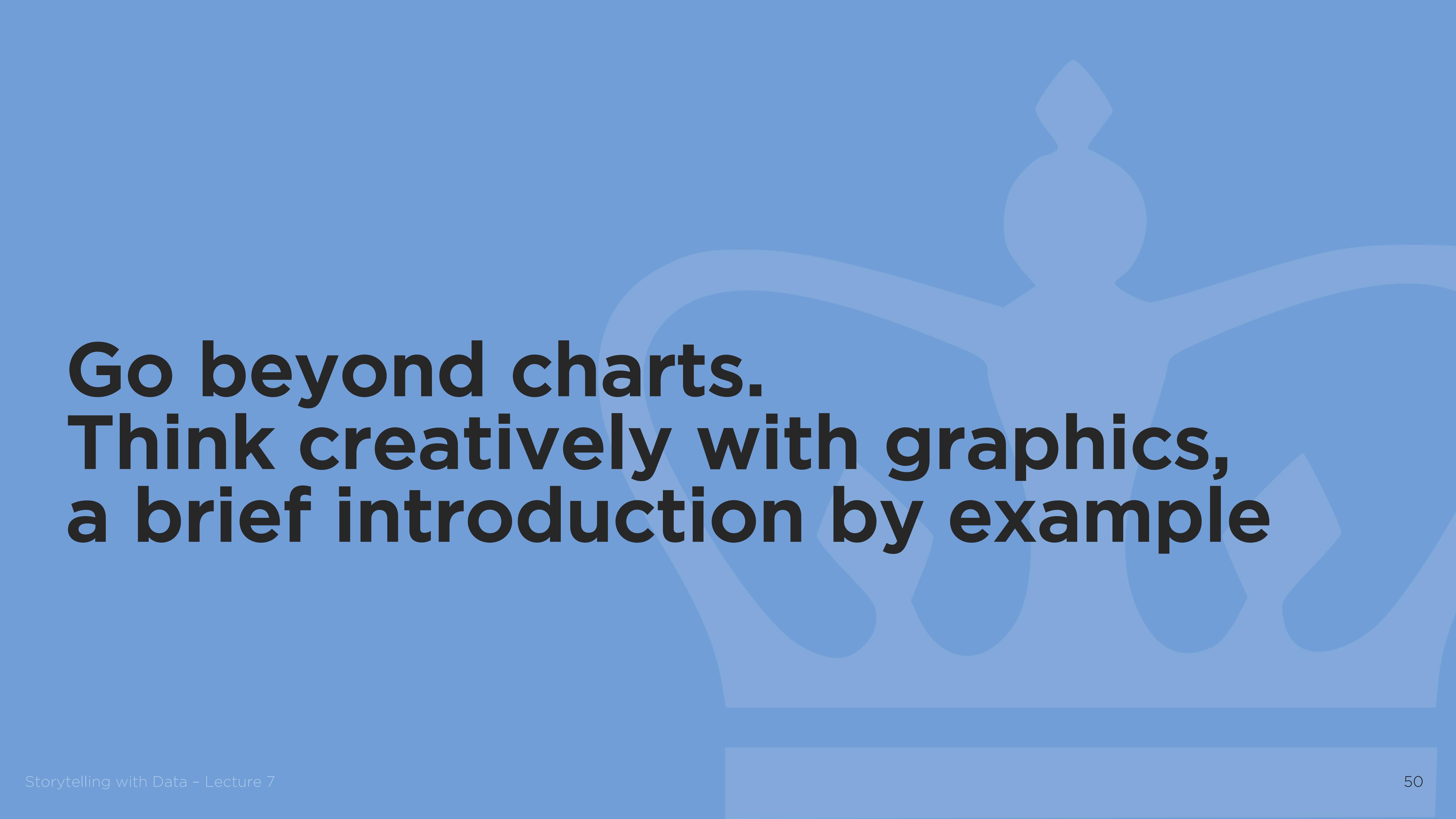
Maximizing data ink (within reason) is but a single dimension of a complex and multivariate design task.

The principle **helps conduct experiments** in graphical design.

Some of those experiments will succeed.

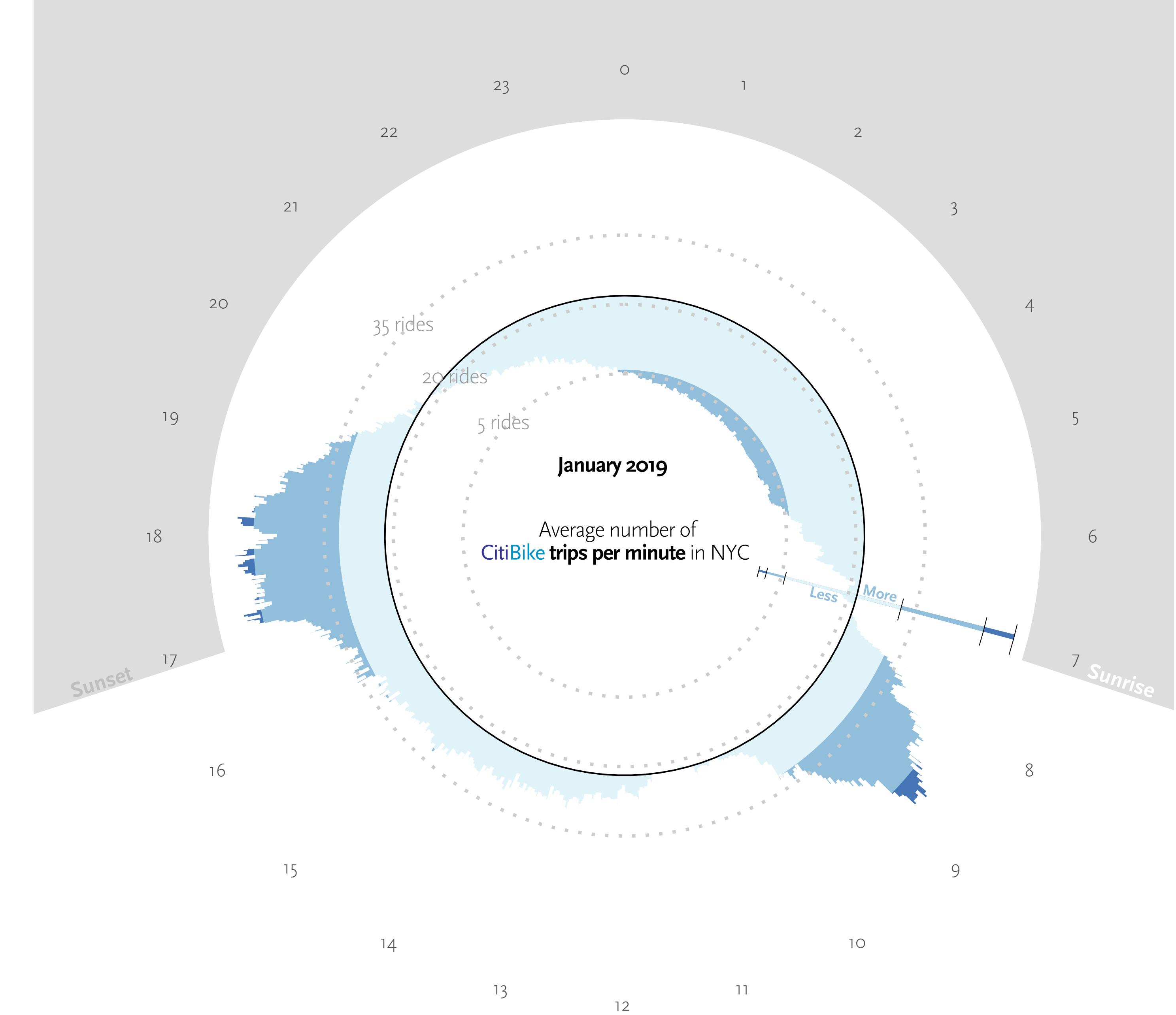
There remain, however, **many other considerations** in the design of statistical graphics – not only of efficiency, but also of complexity, structure, density, and even beauty.

”



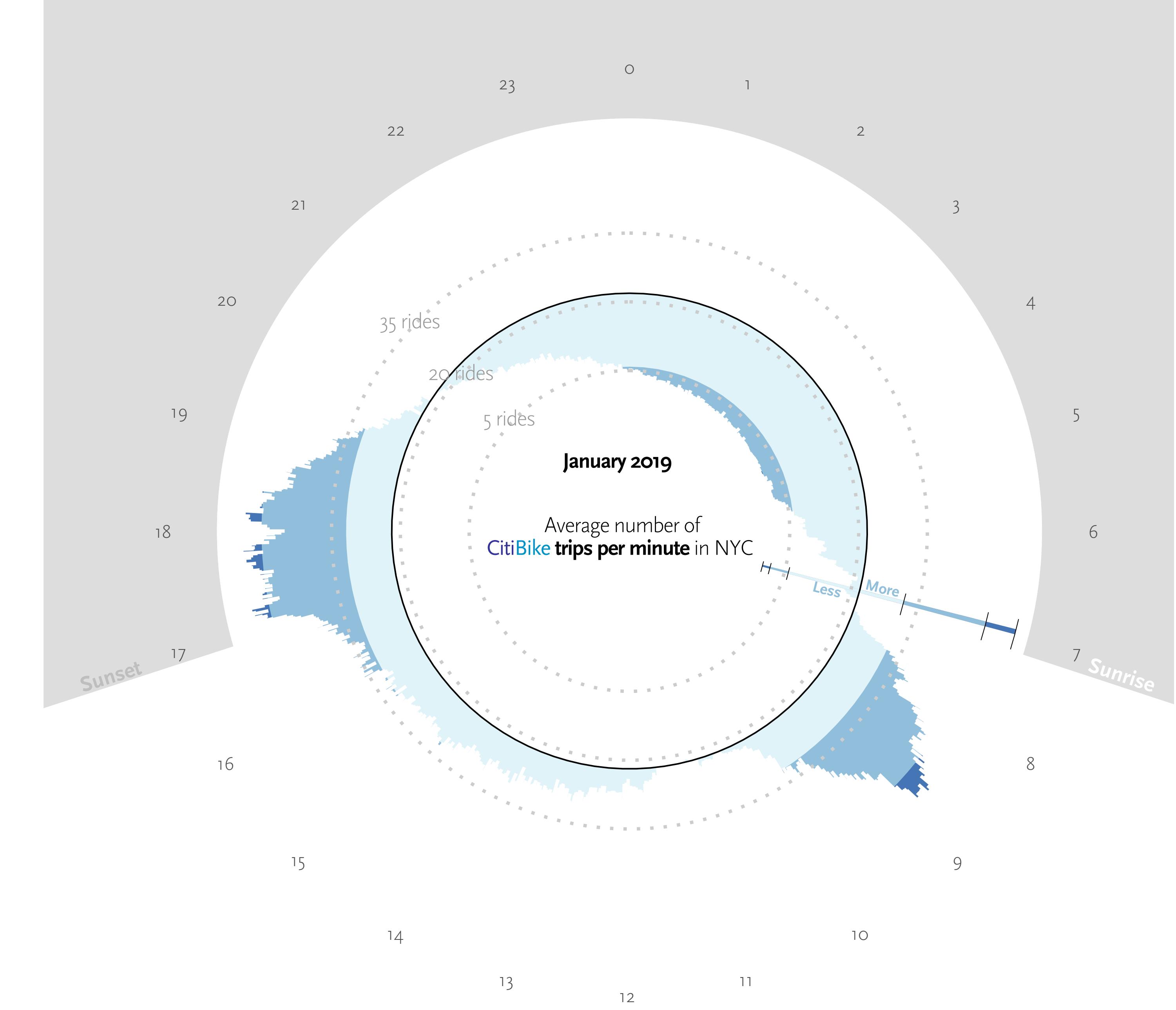
**Go beyond charts.
Think creatively with graphics,
a brief introduction by example**

Think about graphics as l/a/y/e/r/s:

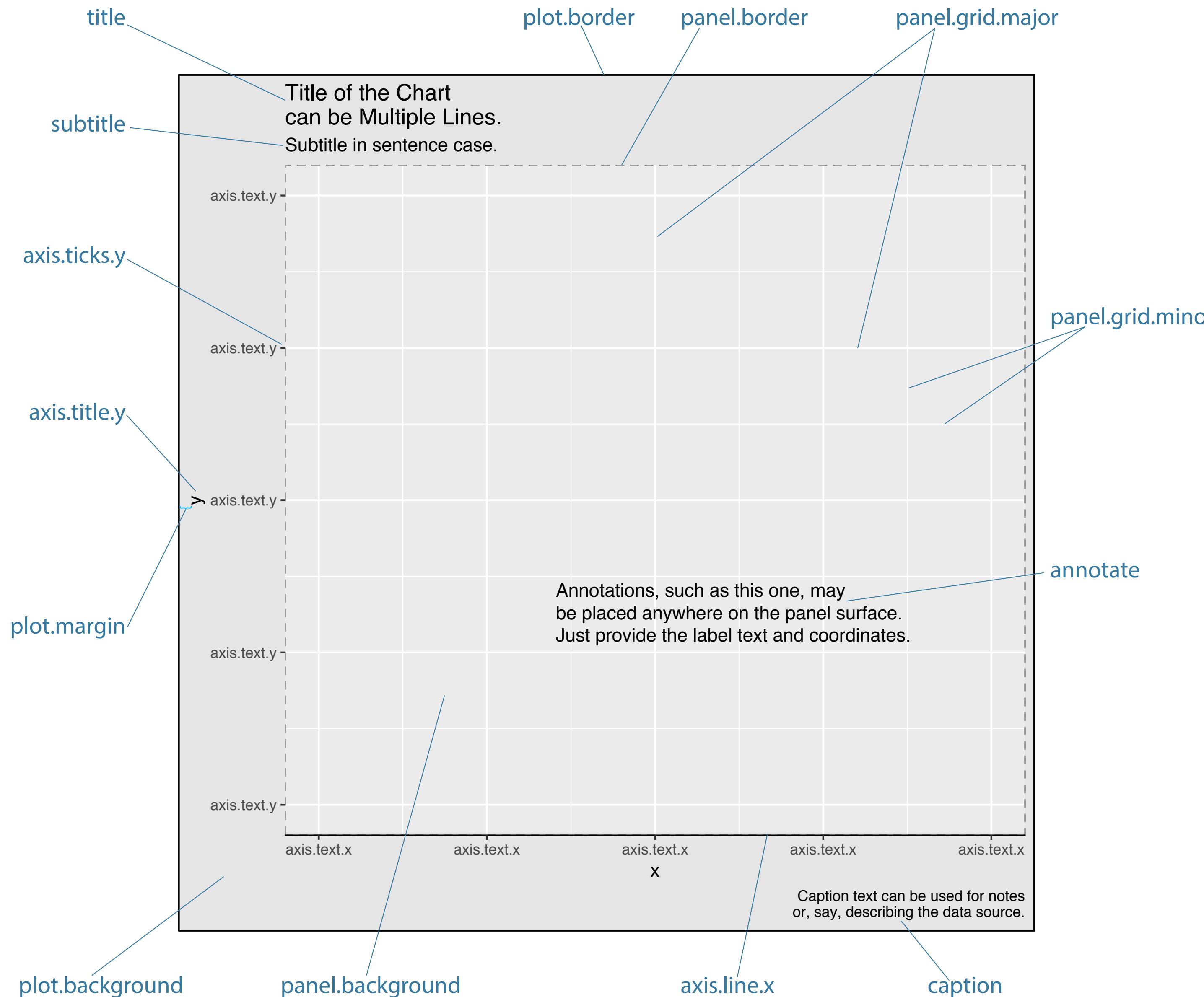


Think about
graphics as l/a/y/e/r/s:

Here, a white layer
partly masks a
rectangular band of blues



First, think visually in describing the graphic. Then create the layers.



```
# load grammar of graphics  
library(ggplot2)
```

```
p <-
```

```
# functions for data ink
```

```
ggplot(data = <data>,  
       mapping = aes(<aesthetic> = <variable>,  
                     <aesthetic> = <variable>,  
                     <...> = <...>) +  
       geom_<type>(<...>) +  
       scale_<mapping>_<type>(<...>) +  
       coord_<type>(<...>) +  
       facet_<type>(<...>) +  
       <...> +
```

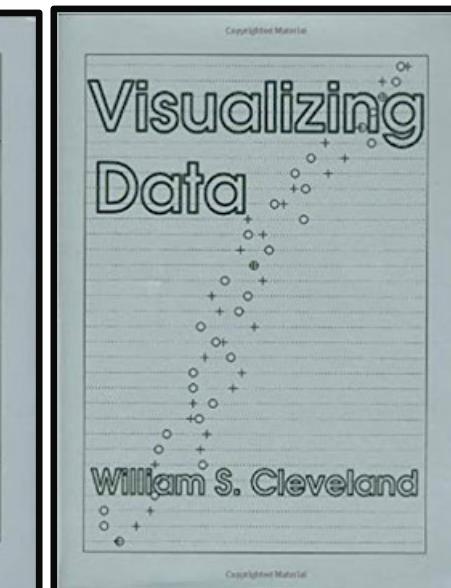
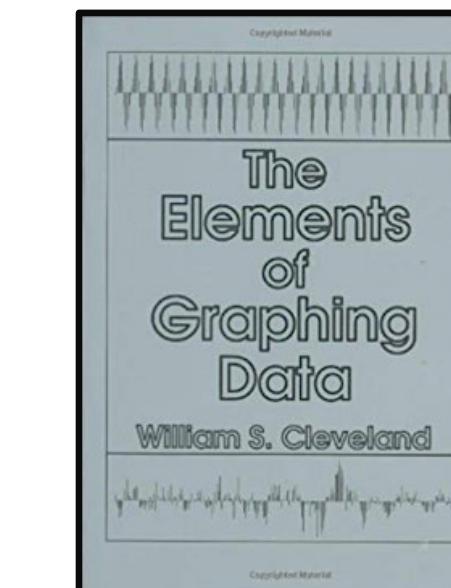
```
# functions for non-data ink
```

```
labs(<...>) +  
theme(<...> = <...>) +  
annotate(<...>) +  
<...>
```

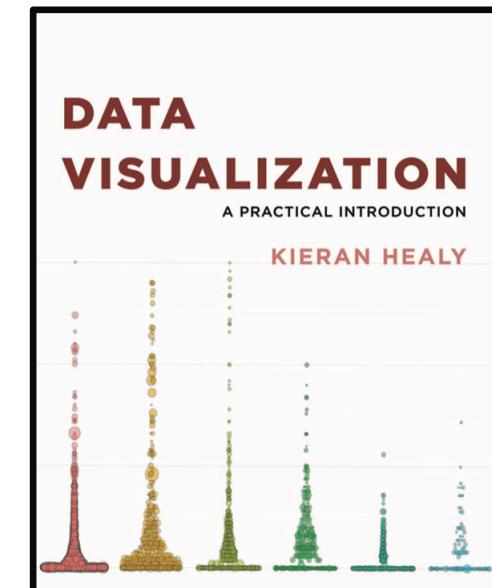
```
element_blank()  
element_line(<...> = <...>)  
element_rect(<...> = <...>)  
element_text(<...> = <...>)
```

Learning & References

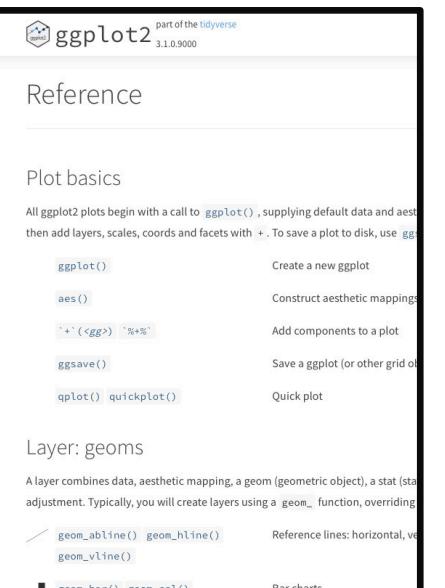
---- visual relationships in data ----



Implementation in R



ggplot reference



socviz.co

ggplot2.tidyverse.org

Between now and next class

For Next Week, Module 8:

Agenda next week

The minimum

Work on your data analysis
Theory and best practices for the visual components of your analytics project, continued

Wickham, Hadley. “*A Layered Grammar of Graphics.*” *Journal of Computational and Graphical Statistics* 19.1 (2010): 3–28. Web.

Consider thinking about graphics as layered components, and the role of each of those layers.

Lupi, Giorgia. *The New Aesthetic of Data Narrative* in Chapter 3 of Bihanic, David. *New Challenges for Data Design*. Springer, 2015. Print.

Read to get a sense of how Lupi thinks through making a graphic.

Liu, Zhicheng et al. “*Data Illustrator.*” New York, New York, USA: ACM Press, 2018. 1–13. Print.

Consider how our choice of tools may affect the way we communicate visually about data.

Example use of Data Illustrator: Nobel Laureates and Prizes. <http://data-illustrator.com/example.php?v=nobel-no-org>

Watch the video to see how Lupi’s award-winning graphic can be created with this tool.

For online discussion

Uncertain?

What types of uncertainty have you identified in your project and data, and how might identifying that uncertainty be important for our different audiences?

Visualize it

Consider a few ways you might visually represent the uncertainty you identified.

See you
next week!

