Storytelling with data

04 | numeracy in narratives — composition and layout

Scott Spencer | Columbia University

Ideate a **data analytics project** addressing problem or opportunity for **your audience**

FINAL

Write **memo** pitching idea to **Chief Analytics Officer**

YOU ARE HERE

DRAFT
FINAL

Write a **project proposal** to **Chief Analytics Officer**

FINAL

Present **critique** of exemplary **information graphic**

DRAFT
FINAL

Create **information graphics** of project and insights for an **External Audience**

DRAFT
FINAL

Build **interactive graphics** for **Chief Marketing Officer** to explore project and insights

FINAL

Write **critique** of **peer presentation**

DRAFT
FINAL

Present **project** and **insights** to **Chief Executive Officer**

CONDUCT DATA ANALYSIS

contextualize numbers — *who* *what* *when* *where* — and *compare*

25

context for numbers, the w's, example — suppose you want to include some mortality statistics in the introductory section of a paper about the Black Plague in fourteenth-century Europe:

worse

There were 25 million deaths.

¯_(ツ)_/¯

During the fourteenth century, 25 million people died in Europe.

better

When the Black Plague hit Europe in the latter half of the fourteenth century, it took the lives of 25 million people, young and old, city dwellers and those living in the countryside. The disease killed about one-quarter of Europe's total population at the time (Mack, n.d.).

context for numbers, effective examples and comparisons — for choosing, aim for simplicity and plausibility

worse

In 2001, the average temperature in the New York City area was 56.3 degrees Fahrenheit.

¯_(ツ)_/¯

In 2001, the average temperature in the New York City area was 56.3 degrees Fahrenheit, 1.5 degrees above normal.

better

In 2001, the average temperature in the New York City area was 56.3 degrees Fahrenheit, 1.5 degrees above normal, making it the seventh warmest year on record.

context for numbers, interpret, don't just report (recall Doumont's "messages, not just information"?)

worse

In 1998, total expenditures on health care in the United States were estimated to be more than $1.1 trillion (Centers for Medicare and Medicaid 2004).

¯_(ツ)_/¯

In 1998, total expenditures on health care in the United States were estimated to be more than $1.1 trillion, equivalent to $4,178 for every man, woman, and child in the nation (Centers for Medicare and Medicaid 2004).

better (for context)

Health care costs in other countries suggest per capita costs in the United States is too high, averaging $4,108 in the 1990s, 13.0% of gross domestic product. That was higher than in any other country. In comparison, Switzerland—with the second highest per capita health costs—spent approximately $3,835 per person, or 10.4% of GDP. No other country exceeded $3,000 per capita (World Bank 2001).

Human body
Animals
Plants
Buildings and constructions
Machines and tools
Games and Sport
Money
Cooking and food
Heat and cold
Light and darkness
Movement and direction

source domain > **target domain**

The thing you are trying to explain

 effective metaphors, analogies — example

> To bring [Rembrandt] back, we distilled the artistic DNA from his work and used it to create *The Next Rembrandt*. . . . To create new artwork using data from Rembrandt's paintings, we had to maximize the data pool from which to pull information. . . . We created a height map using two different algorithms that found texture patterns of canvas surfaces and layers of paint. That information was transformed into height data, allowing us to mimic the brushstrokes used by Rembrandt.

— Ing. *The Next Rembrandt*, https://www.nextrembrandt.com. April 2016.

 effective metaphors, analogies — example

setting up the metaphor

How do we think about the albums we love? A lonely microphone in a smoky recording studio? A needle's press into hot wax? A rotating can of magnetic tape? A button that clicks before the first note drops? No!

The mechanical ephemera of music's recording, storage, and playback may cue nostalgia, but they are not where the magic lies. The magic is in the music. The magic is in the information that the apparatuses capture, preserve, and make accessible. It is the same with all information.

— Andrews, R J. *Info We Trust: How to Inspire the World with Data*. Wiley, 2019.

referring back

When you envision data, do not get stuck in encoding and storage. Instead, try to see the music.

…

Looking at tables of any substantial size is a little like looking at the grooves of a record with a magnifying glass. You can see the data but you will not hear the music.

...

Then, we can see data for what it is, whispers from a past world waiting for its music to be heard again.

context for numbers, **for relationships between numbers — compare with *direction* and *magnitude***

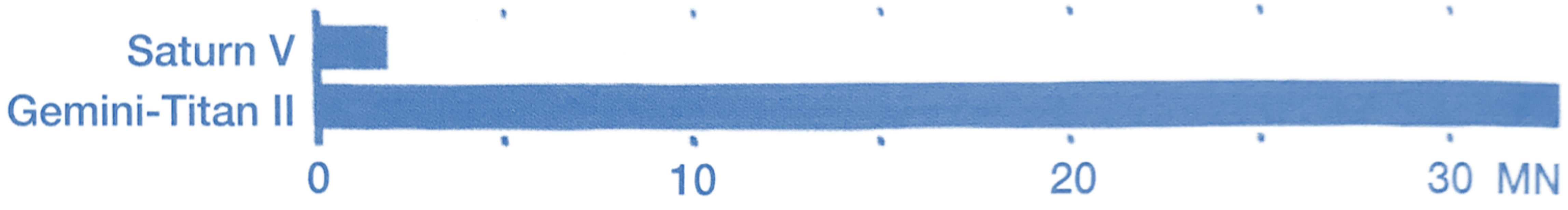
worse	¯_(ツ)_/¯	better
Mortality and age are correlated.	As age increases, mortality increases.	Among the elderly, mortality roughly doubles for each successive five-year age group.

The Apollo program crew had **one more** astronaut than Project Gemini. Apollo's Saturn V rocket had about **seventeen times more** thrust than the Gemini-Titan II.

"Seventeen times more"

"1,700 percent more"

"33 versus 1.9"



context for numbers, summarizing numeric patterns — generalizations, examples, exceptions

generalizations

For a generalization, come up with a description that characterizes a relationship among most, if not all, of the numbers.

examples

Illustrate your generalization with numbers from your table or chart. This step anchors your generalization to the specific numbers upon which it is based.

It ties the prose and table or chart together. By reporting a few illustrative numbers, you implicitly show your readers where in the table or chart those numbers came from as well as the comparison involved.

exceptions

When portraying an exception, explain its overall shape and how it differs from the generalization you described and illustrated. *Is it higher or lower? By how much? If a trend, is it moving toward or away from the pattern you are contrasting it against?* Finally, provide numeric examples from the table or chart to illustrate the exception.

organizing numbers — tables and semi-graphic displays

organizing numbers, **tables for comparing exact numbers**

"The conventional sentence is a poor way to show more than two numbers because it prevents comparisons within the data.

The linearly organized flow of words, folded over at arbitrary points (decided not by content but by the happenstance of column width), offers less than one effective dimension for organizing the data."

— Edward Tufte, *The Visual Display of Quantitative Information*

Instead of:

> Nearly 53 percent of the type A group did something or other compared to 46 percent of B and slightly more than 57 percent of C.

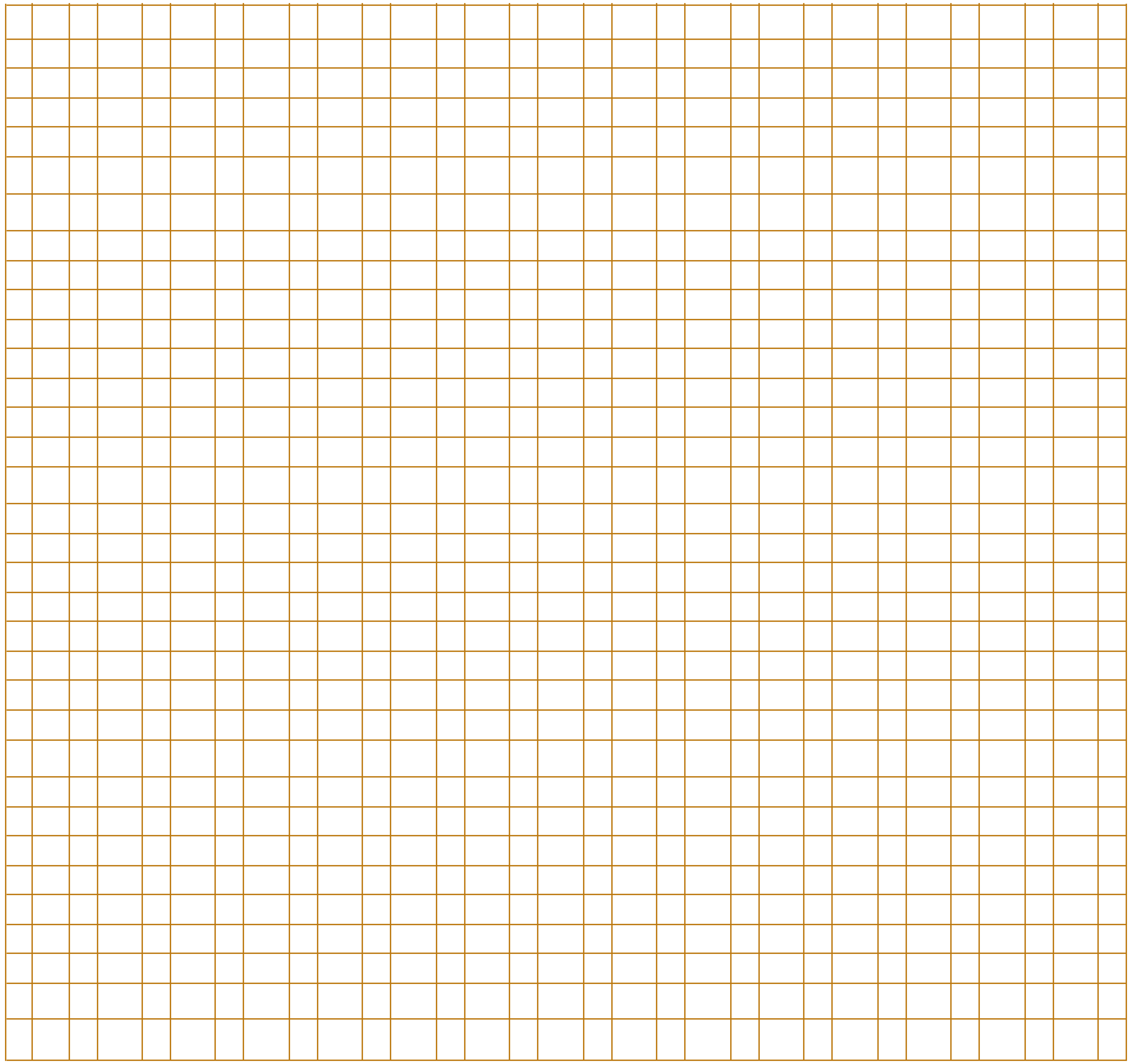
Arrange the type to facilitate comparisons, as in this *text-table*:

> The three groups differed in how they did something or other:
>
> Group A 53%
> Group B 46%
> Group C 57%

There are nearly always better sequences than alphabetical-for example, ordering by content or by data values:

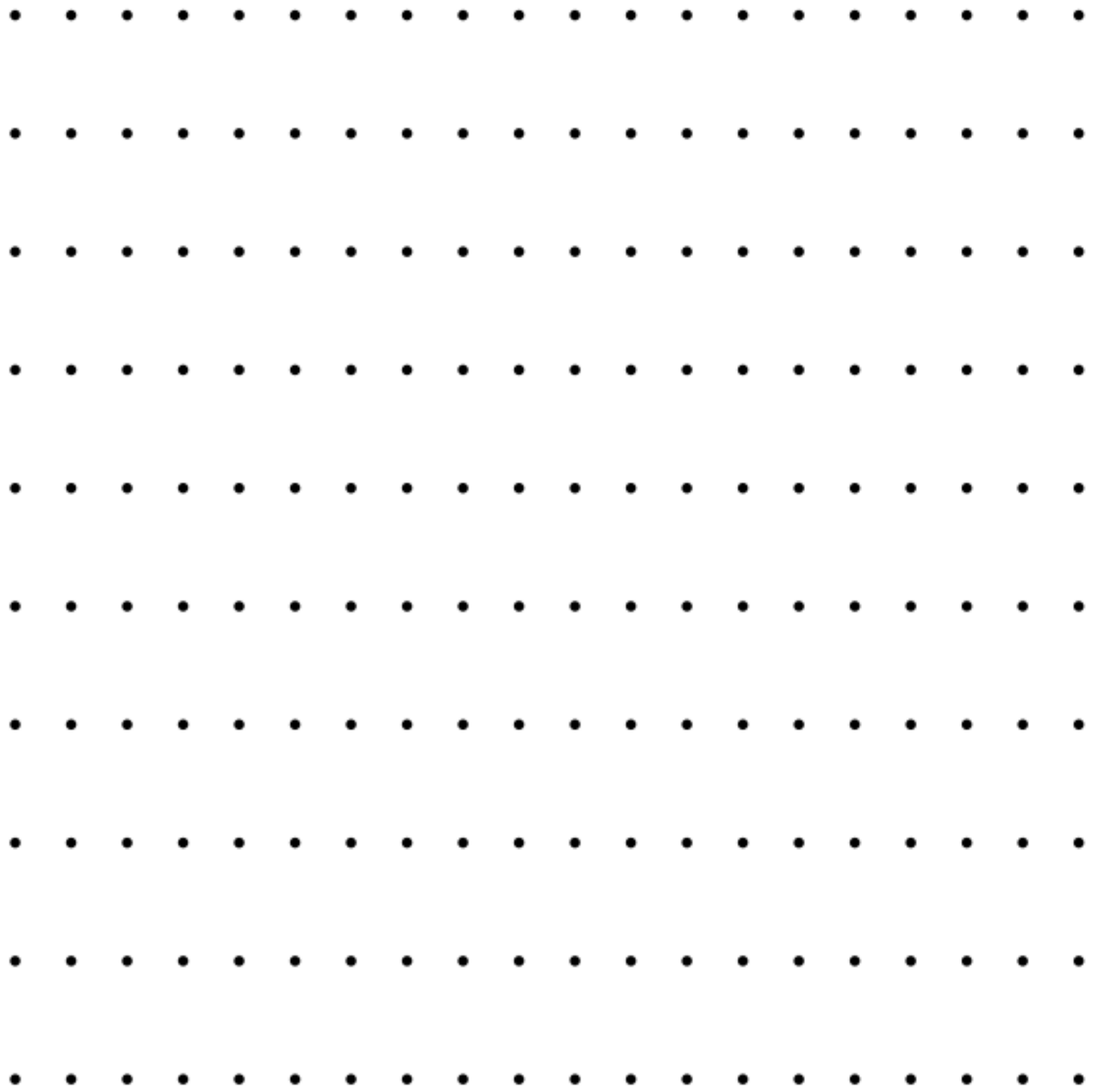
> Group B 46%
> Group A 53%
> Group C 57%

organizing numbers, using |g|r|i|d|s| for arranging (a table of) numbers

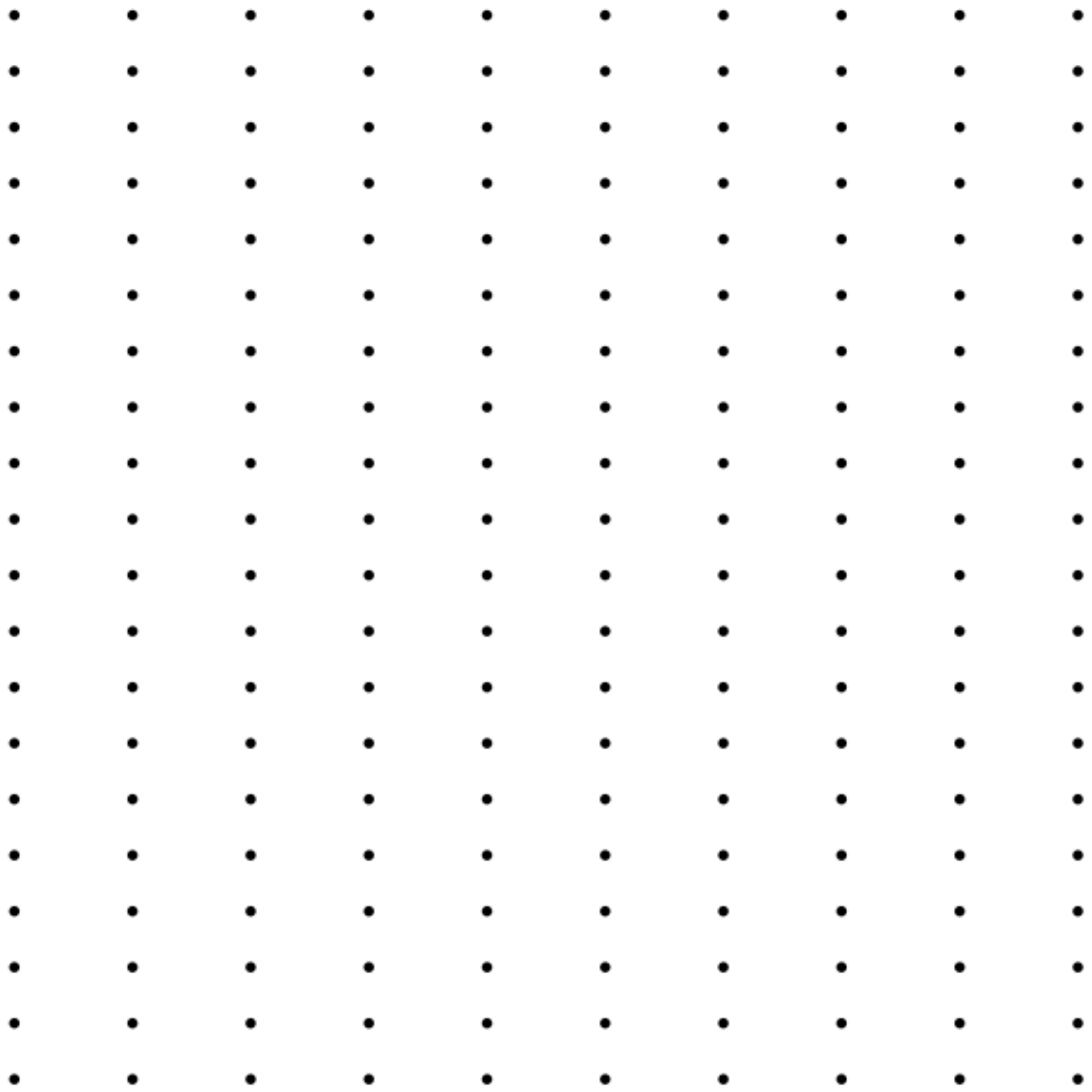


organizing numbers, **placement in grid? reduce cognitive load — Gestalt principle of *proximity***

spacing—horizontal ***narrower*** than vertical



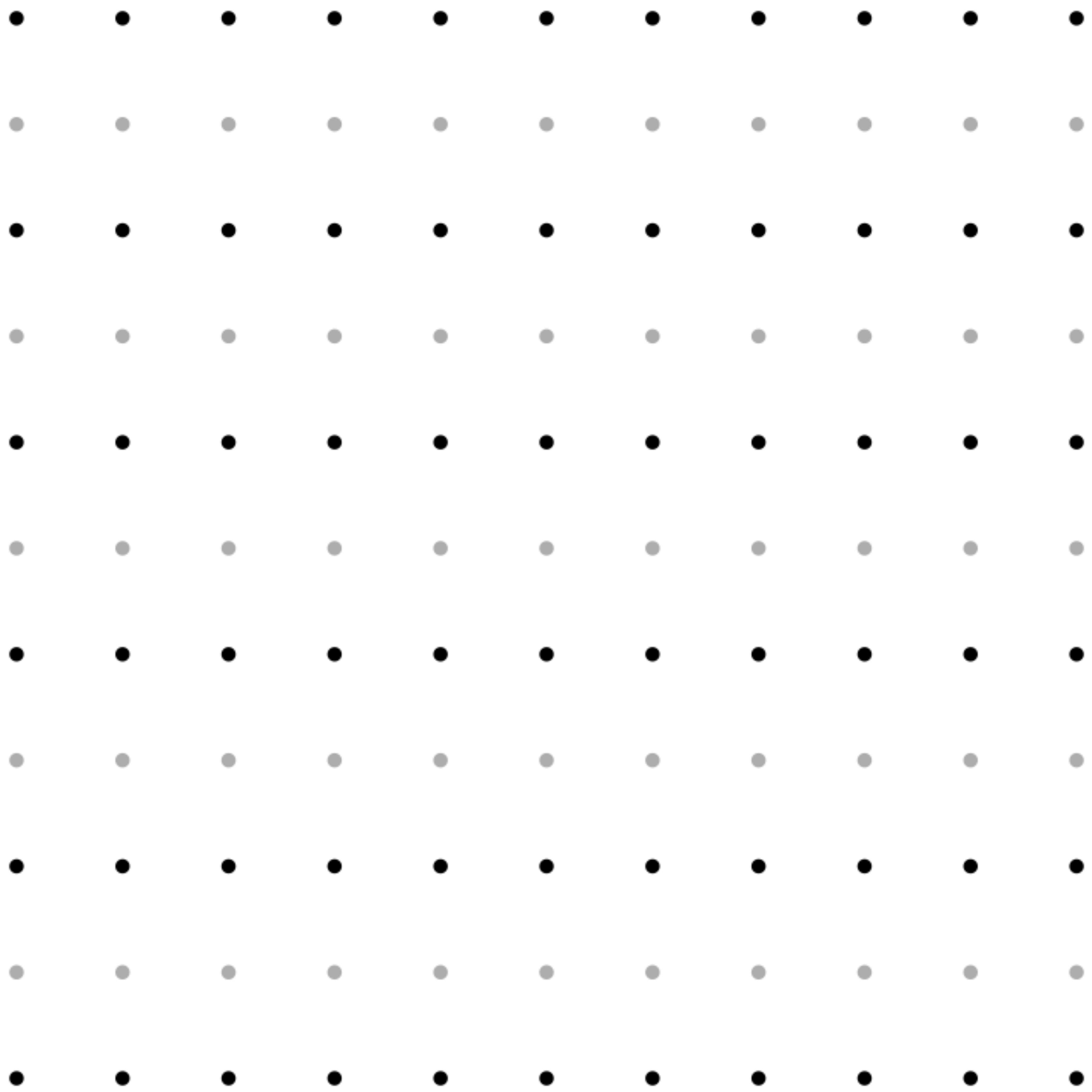
spacing—horizontal ***wider*** than vertical



organizing numbers, example placement in grid — proximity for perceived column groups, groups of columns

0.54	0.50	0.44	0.59	0.54	0.47	0.66	0.61	0.54	0.77	0.74	0.69
0.96	0.89	0.79	1.04	0.96	0.85	1.16	1.08	0.96	1.34	1.29	1.22
1.17	1.10	1.00	1.25	1.18	1.06	1.37	1.29	1.17	1.55	1.50	1.43
1.54	1.47	1.35	1.65	1.55	1.42	1.79	1.69	1.55	2.01	1.95	1.86
1.40	1.36	1.29	1.45	1.40	1.33	1.53	1.48	1.40	1.65	1.61	1.57
1.80	1.71	1.57	1.91	1.81	1.65	2.08	1.96	1.80	2.33	2.26	2.16
1.93	1.86	1.75	2.02	1.94	1.82	2.15	2.06	1.94	2.34	2.29	2.21
2.28	2.19	2.03	2.40	2.29	2.12	2.58	2.46	2.29	2.84	2.77	2.67
0.29	0.26	0.22	0.32	0.29	0.24	0.37	0.34	0.29	0.44	0.42	0.39
0.56	0.51	0.43	0.62	0.56	0.48	0.71	0.65	0.56	0.84	0.81	0.75
0.71	0.66	0.58	0.77	0.71	0.63	0.86	0.80	0.71	1.00	0.96	0.91
0.98	0.91	0.79	1.07	0.99	0.86	1.21	1.11	0.98	1.41	1.35	1.27
0.97	0.89	0.77	1.07	0.98	0.85	1.21	1.11	0.97	1.42	1.36	1.28
1.21	1.13	0.98	1.33	1.22	1.07	1.50	1.38	1.22	1.75	1.68	1.58
1.33	1.24	1.09	1.45	1.34	1.18	1.62	1.51	1.34	1.88	1.81	1.71
1.62	1.50	1.31	1.78	1.63	1.43	1.99	1.84	1.63	2.33	2.24	2.10
0.11	0.10	0.07	0.13	0.11	0.09	0.16	0.14	0.11	0.20	0.19	0.17
0.23	0.20	0.15	0.27	0.24	0.18	0.33	0.29	0.23	0.42	0.40	0.36
0.30	0.26	0.20	0.36	0.31	0.24	0.44	0.38	0.31	0.55	0.52	0.48
0.45	0.39	0.29	0.53	0.46	0.35	0.65	0.57	0.46	0.82	0.77	0.70
0.35	0.30	0.22	0.41	0.36	0.27	0.50	0.44	0.35	0.64	0.60	0.55
0.50	0.43	0.32	0.59	0.50	0.38	0.71	0.62	0.50	0.90	0.85	0.77
0.52	0.45	0.33	0.62	0.53	0.40	0.75	0.66	0.53	0.95	0.89	0.81
0.82	0.71	0.53	0.97	0.83	0.64	1.18	1.03	0.83	1.49	1.41	1.28

organizing numbers, separating information types — Gestalt principle of *similarity* (e.g., by color)



organizing numbers, **example placement in grid — with labels and annotations (and color encoding)**

Table 1: For takes, our Bayesian model expects batting-team runs to increase by these amounts during the half-inning, given game state and count.

Gamestate	Count											
—:-	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2	3-0	3-1	3-2
No outs												
—:0	0.54	0.50	0.44	0.59	0.54	0.47	0.66	0.61	0.54	0.77	0.74	0.69
1–:0	0.96	0.89	0.79	1.04	0.96	0.85	1.16	1.08	0.96	1.34	1.29	1.22
-2-:0	1.17	1.10	1.00	1.25	1.18	1.06	1.37	1.29	1.17	1.55	1.50	1.43
12-:0	1.54	1.47	1.35	1.65	1.55	1.42	1.79	1.69	1.55	2.01	1.95	1.86
–3:0	1.40	1.36	1.29	1.45	1.40	1.33	1.53	1.48	1.40	1.65	1.61	1.57
1-3:0	1.80	1.71	1.57	1.91	1.81	1.65	2.08	1.96	1.80	2.33	2.26	2.16
-23:0	1.93	1.86	1.75	2.02	1.94	1.82	2.15	2.06	1.94	2.34	2.29	2.21
123:0	2.28	2.19	2.03	2.40	2.29	2.12	2.58	2.46	2.29	2.84	2.77	2.67
One out												
—:1	0.29	0.26	0.22	0.32	0.29	0.24	0.37	0.34	0.29	0.44	0.42	0.39
1–:1	0.56	0.51	0.43	0.62	0.56	0.48	0.71	0.65	0.56	0.84	0.81	0.75
-2-:1	0.71	0.66	0.58	0.77	0.71	0.63	0.86	0.80	0.71	1.00	0.96	0.91
12-:1	0.98	0.91	0.79	1.07	0.99	0.86	1.21	1.11	0.98	1.41	1.35	1.27
–3:1	0.97	0.89	0.77	1.07	0.98	0.85	1.21	1.11	0.97	1.42	1.36	1.28
1-3:1	1.21	1.13	0.98	1.33	1.22	1.07	1.50	1.38	1.22	1.75	1.68	1.58
-23:1	1.33	1.24	1.09	1.45	1.34	1.18	1.62	1.51	1.34	1.88	1.81	1.71
123:1	1.62	1.50	1.31	1.78	1.63	1.43	1.99	1.84	1.63	2.33	2.24	2.10
Two outs												
—:2	0.11	0.10	0.07	0.13	0.11	0.09	0.16	0.14	0.11	0.20	0.19	0.17
1–:2	0.23	0.20	0.15	0.27	0.24	0.18	0.33	0.29	0.23	0.42	0.40	0.36
-2-:2	0.30	0.26	0.20	0.36	0.31	0.24	0.44	0.38	0.31	0.55	0.52	0.48
12-:2	0.45	0.39	0.29	0.53	0.46	0.35	0.65	0.57	0.46	0.82	0.77	0.70
–3:2	0.35	0.30	0.22	0.41	0.36	0.27	0.50	0.44	0.35	0.64	0.60	0.55
1-3:2	0.50	0.43	0.32	0.59	0.50	0.38	0.71	0.62	0.50	0.90	0.85	0.77
-23:2	0.52	0.45	0.33	0.62	0.53	0.40	0.75	0.66	0.53	0.95	0.89	0.81
123:2	0.82	0.71	0.53	0.97	0.83	0.64	1.18	1.03	0.83	1.49	1.41	1.28

Note:

Expectations from model fit to Statcast data, 2017-2019.

organizing numbers, example placement in grids — gridlines are invisible

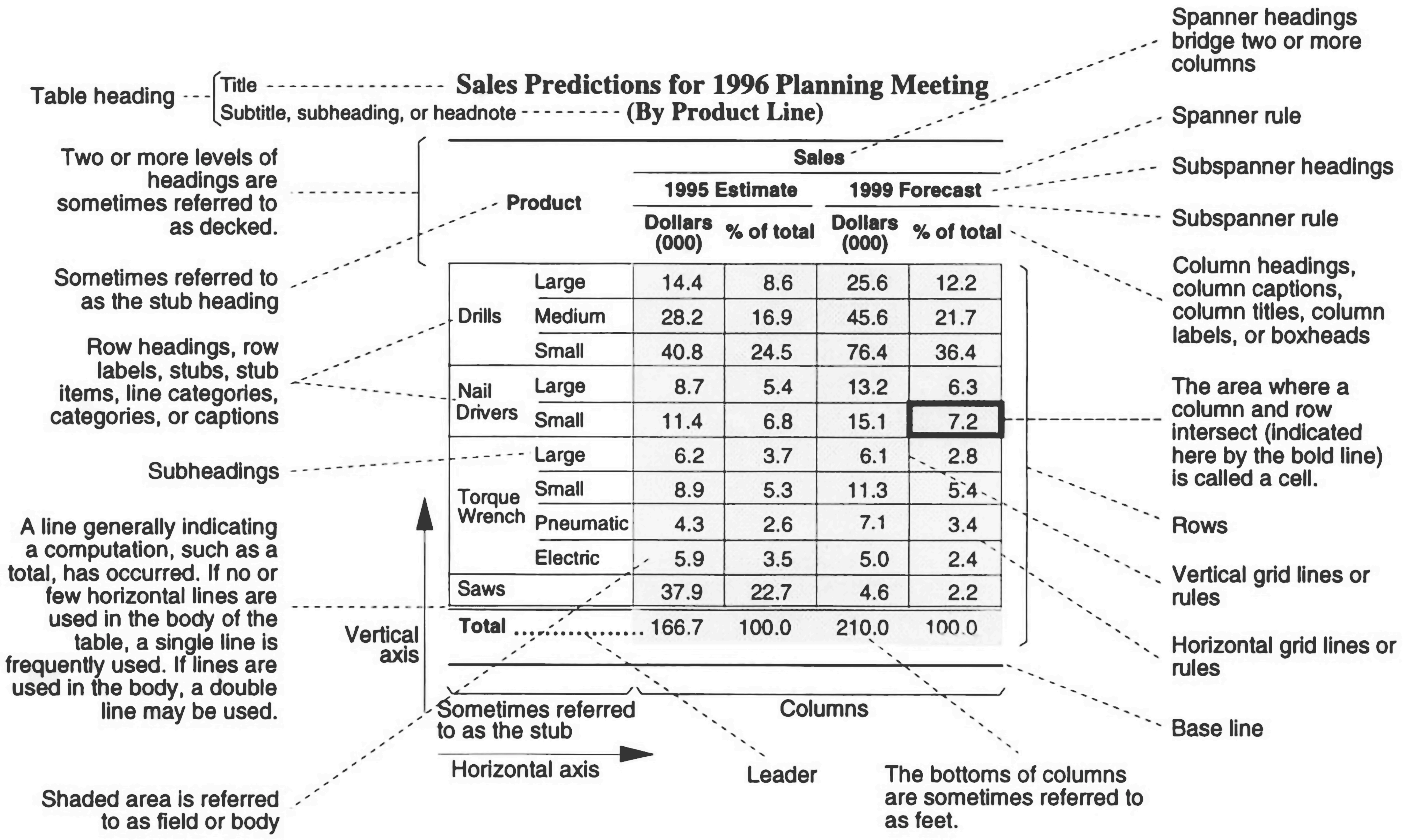
Table 1: For takes, our Bayesian model expects batting-team runs to increase by these amounts during the half-inning, given game state and count.

Game State	Count											
—:-	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2	3-0	3-1	3-2
No outs												
—:0	0.54	0.50	0.44	0.59	0.54	0.47	0.66	0.61	0.54	0.77	0.74	0.69
1–:0	0.96	0.89	0.79	1.04	0.96	0.85	1.16	1.08	0.96	1.34	1.29	1.22
-2-:0	1.17	1.10	1.00	1.25	1.18	1.06	1.37	1.29	1.17	1.55	1.50	1.43
–3:0	1.40	1.36	1.29	1.45	1.40	1.33	1.53	1.48	1.40	1.65	1.61	1.57
12-:0	1.54	1.47	1.35	1.65	1.55	1.42	1.79	1.69	1.55	2.01	1.95	1.86
1-3:0	1.80	1.71	1.57	1.91	1.81	1.65	2.08	1.96	1.80	2.33	2.26	2.16
-23:0	1.93	1.86	1.75	2.02	1.94	1.82	2.15	2.06	1.94	2.34	2.29	2.21
123:0	2.28	2.19	2.03	2.40	2.29	2.12	2.58	2.46	2.29	2.84	2.77	2.67
One out												
—:1	0.29	0.26	0.22	0.32	0.29	0.24	0.37	0.34	0.29	0.44	0.42	0.39
1–:1	0.56	0.51	0.43	0.62	0.56	0.48	0.71	0.65	0.56	0.84	0.81	0.75
-2-:1	0.71	0.66	0.58	0.77	0.71	0.63	0.86	0.80	0.71	1.00	0.96	0.91
–3:1	0.97	0.89	0.77	1.07	0.98	0.85	1.21	1.11	0.97	1.42	1.36	1.28
12-:1	0.98	0.91	0.79	1.07	0.99	0.86	1.21	1.11	0.98	1.41	1.35	1.27
1-3:1	1.21	1.13	0.98	1.33	1.22	1.07	1.50	1.38	1.22	1.75	1.68	1.58
-23:1	1.33	1.24	1.09	1.45	1.34	1.18	1.62	1.51	1.34	1.88	1.81	1.71
123:1	1.62	1.50	1.31	1.78	1.63	1.43	1.99	1.84	1.63	2.33	2.24	2.10
Two outs												
—:2	0.11	0.10	0.07	0.13	0.11	0.09	0.16	0.14	0.11	0.20	0.19	0.17
1–:2	0.23	0.20	0.15	0.27	0.24	0.18	0.33	0.29	0.23	0.42	0.40	0.36
-2-:2	0.30	0.26	0.20	0.36	0.31	0.24	0.44	0.38	0.31	0.55	0.52	0.48
–3:2	0.35	0.30	0.22	0.41	0.36	0.27	0.50	0.44	0.35	0.64	0.60	0.55
12-:2	0.45	0.39	0.29	0.53	0.46	0.35	0.65	0.57	0.46	0.82	0.77	0.70
1-3:2	0.50	0.43	0.32	0.59	0.50	0.38	0.71	0.62	0.50	0.90	0.85	0.77
-23:2	0.52	0.45	0.33	0.62	0.53	0.40	0.75	0.66	0.53	0.95	0.89	0.81
123:2	0.82	0.71	0.53	0.97	0.83	0.64	1.18	1.03	0.83	1.49	1.41	1.28

Note:

Expectations from model fit to Statcast data, 2017-2019.

organizing numbers, names and descriptions of common table components



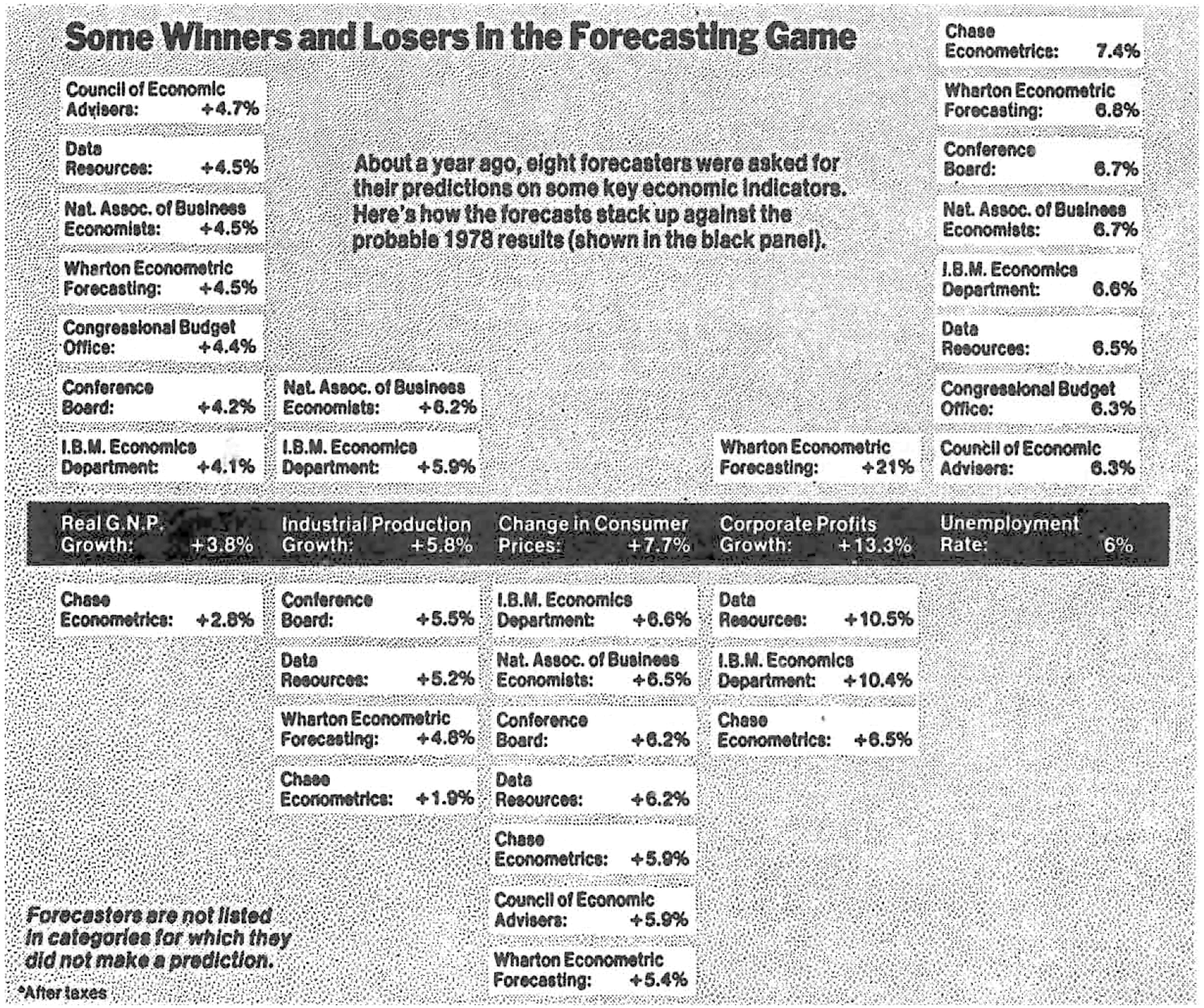
Sales Predictions for 1996 Planning Meeting
(By Product Line)

Product		Sales			
		1995 Estimate		1999 Forecast	
		Dollars (000)	% of total	Dollars (000)	% of total
Drills	Large	14.4	8.6	25.6	12.2
	Medium	28.2	16.9	45.6	21.7
	Small	40.8	24.5	76.4	36.4
Nail Drivers	Large	8.7	5.4	13.2	6.3
	Small	11.4	6.8	15.1	7.2
Torque Wrench	Large	6.2	3.7	6.1	2.8
	Small	8.9	5.3	11.3	5.4
	Pneumatic	4.3	2.6	7.1	3.4
	Electric	5.9	3.5	5.0	2.4
Saws		37.9	22.7	4.6	2.2
Total		166.7	100.0	210.0	100.0

non-rectangular, tabular data and semi-graphic displays (*e.g.*, stem-and-leaf)

non-rectangular and semi-graphic, tabular variations, example — stem-and-leaf diagram

0	1	2	6	7									
1	3	3	4	4	7	8							
2	0	0	1	1	2	2	3	4	5	8	8	9	
3	1	2	3	3	4	4	5	7	7	8	9	9	
4	0	0	2	2	3	4	4	4	5	5	8	8	
5	1	1	4	9									
6	5	8											
7	0	0	0	0	0	3	4	4	5	8	9	9	
8	1	2	4	4	5	5	7	7	7	9	9	9	9
9	3	7											
10	0												



Some Winners and Losers in the Forecasting Game

About a year ago, eight forecasters were asked for their predictions on some key economic indicators. Here's how the forecasts stack up against the probable 1978 results (shown in the black panel).

				Chase Econometrics: 7.4%
Council of Economic Advisers: +4.7%				Wharton Econometric Forecasting: 6.8%
Data Resources: +4.5%				Conference Board: 6.7%
Nat. Assoc. of Business Economists: +4.5%				Nat. Assoc. of Business Economists: 6.7%
Wharton Econometric Forecasting: +4.5%				I.B.M. Economics Department: 6.6%
Congressional Budget Office: +4.4%				Data Resources: 6.5%
Conference Board: +4.2%	Nat. Assoc. of Business Economists: +6.2%			Congressional Budget Office: 6.3%
I.B.M. Economics Department: +4.1%	I.B.M. Economics Department: +5.9%		Wharton Econometric Forecasting: +21%	Council of Economic Advisers: 6.3%
Real G.N.P. Growth: +3.8%	**Industrial Production Growth: +5.8%**	**Change in Consumer Prices: +7.7%**	**Corporate Profits Growth: +13.3%**	**Unemployment Rate: 6%**
Chase Econometrics: +2.8%	Conference Board: +5.5%	I.B.M. Economics Department: +6.6%	Data Resources: +10.5%	
	Data Resources: +5.2%	Nat. Assoc. of Business Economists: +6.5%	I.B.M. Economics Department: +10.4%	
	Wharton Econometric Forecasting: +4.8%	Conference Board: +6.2%	Chase Econometrics: +6.5%	
	Chase Econometrics: +1.9%	Data Resources: +6.2%		
		Chase Econometrics: +5.9%		
		Council of Economic Advisers: +5.9%		
		Wharton Econometric Forecasting: +5.4%		

Forecasters are not listed in categories for which they did not make a prediction.

*After taxes

— NY Times, *Last Year's Forecasts: Why So Many Erred.* 1979 Jan 2.

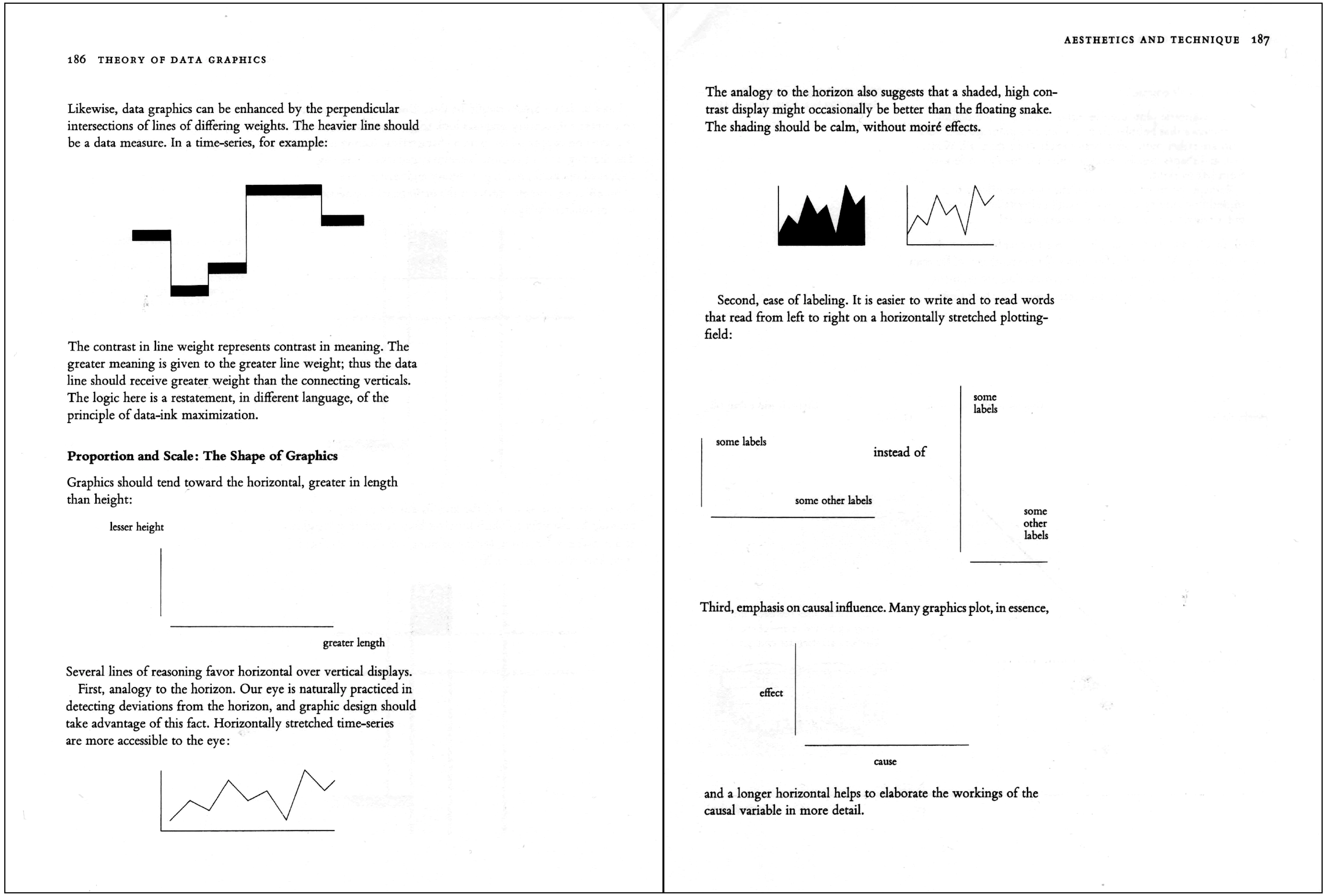
integrating data with text, text with data

text-data integration, **integrate data tables and graphics into narrative (principle of proximity)**

“The principle of *data/text integration* is:

> data graphics are paragraphs about data and should be treated as such.”

— Edward Tufte, *The Visual Display of Quantitative Information*



186 THEORY OF DATA GRAPHICS

Likewise, data graphics can be enhanced by the perpendicular intersections of lines of differing weights. The heavier line should be a data measure. In a time-series, for example:

The contrast in line weight represents contrast in meaning. The greater meaning is given to the greater line weight; thus the data line should receive greater weight than the connecting verticals. The logic here is a restatement, in different language, of the principle of data-ink maximization.

Proportion and Scale: The Shape of Graphics

Graphics should tend toward the horizontal, greater in length than height:

lesser height

greater length

Several lines of reasoning favor horizontal over vertical displays. First, analogy to the horizon. Our eye is naturally practiced in detecting deviations from the horizon, and graphic design should take advantage of this fact. Horizontally stretched time-series are more accessible to the eye:

AESTHETICS AND TECHNIQUE 187

The analogy to the horizon also suggests that a shaded, high contrast display might occasionally be better than the floating snake. The shading should be calm, without moiré effects.

Second, ease of labeling. It is easier to write and to read words that read from left to right on a horizontally stretched plotting-field:

some labels

some other labels

instead of

some labels

some other labels

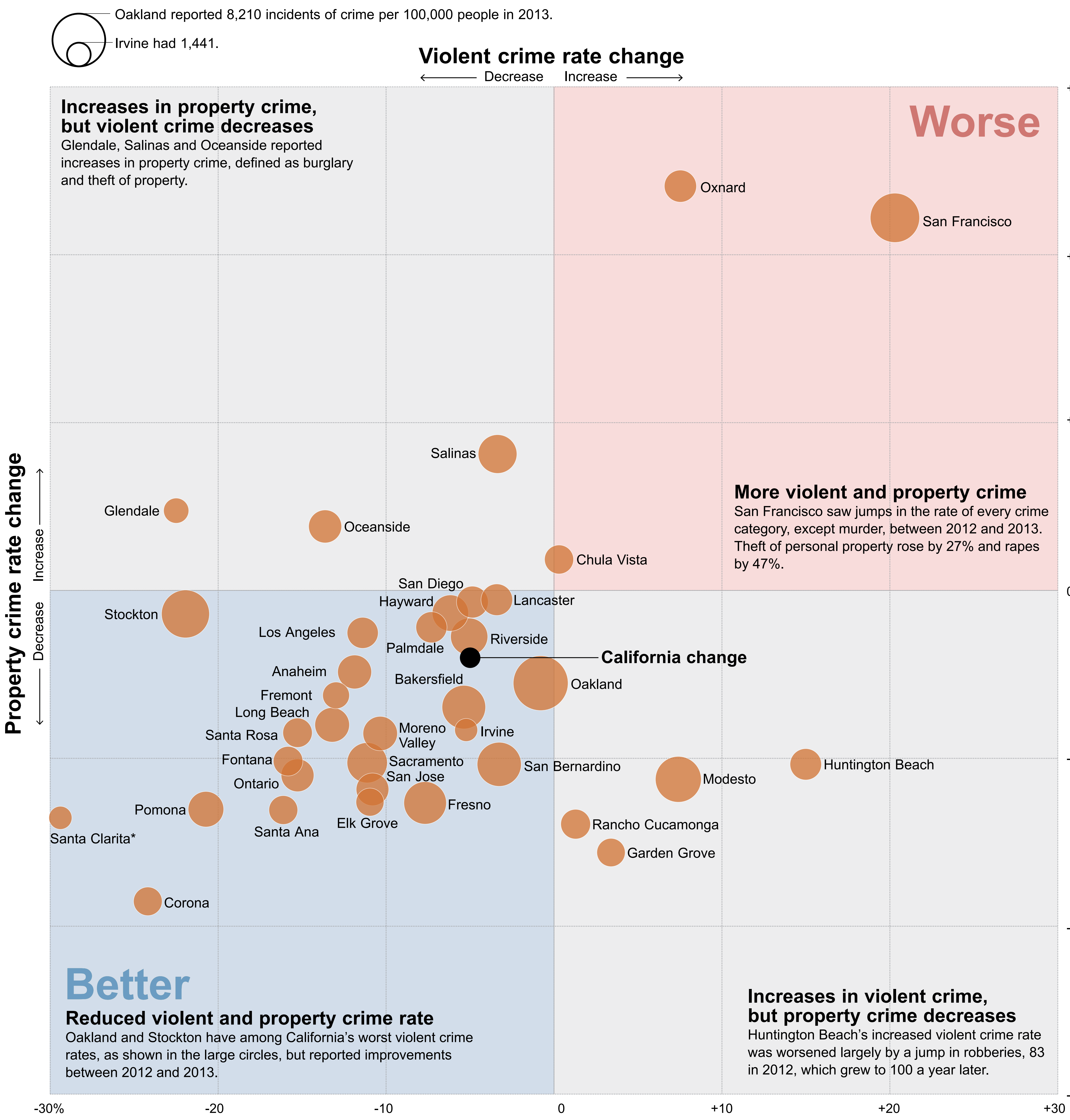
Third, emphasis on causal influence. Many graphics plot, in essence,

effect

cause

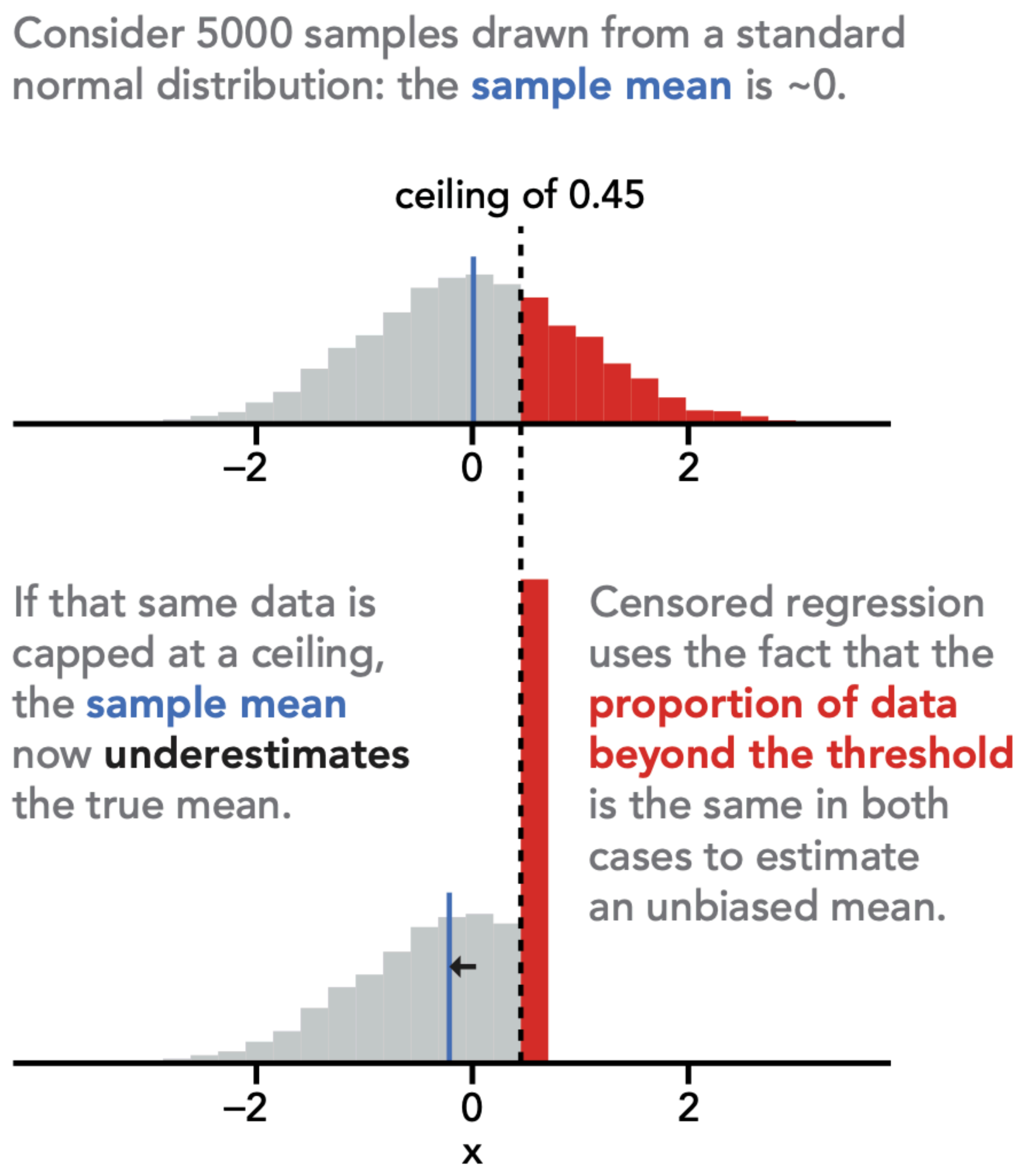
and a longer horizontal helps to elaborate the workings of the causal variable in more detail.

text-data integration, annotate data graphics with descriptions (principle of proximity) — example



— Schleuss, Jon, and Rong-Cong Lin II. 2013. “*California Crime 2013.*” Los Angeles Times.

text-data integration, **linking data and narrative with color (principle of similarity) — example**



— Kay, Matthew, and Jeffrey Heer. *Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation*. IEEE Transactions on Visualization and Computer Graphics 22, no. 1 (January 31, 2016): 469–78.

text-data integration, *linking language*—successive sentences or phrases similar in length, parallel in structure

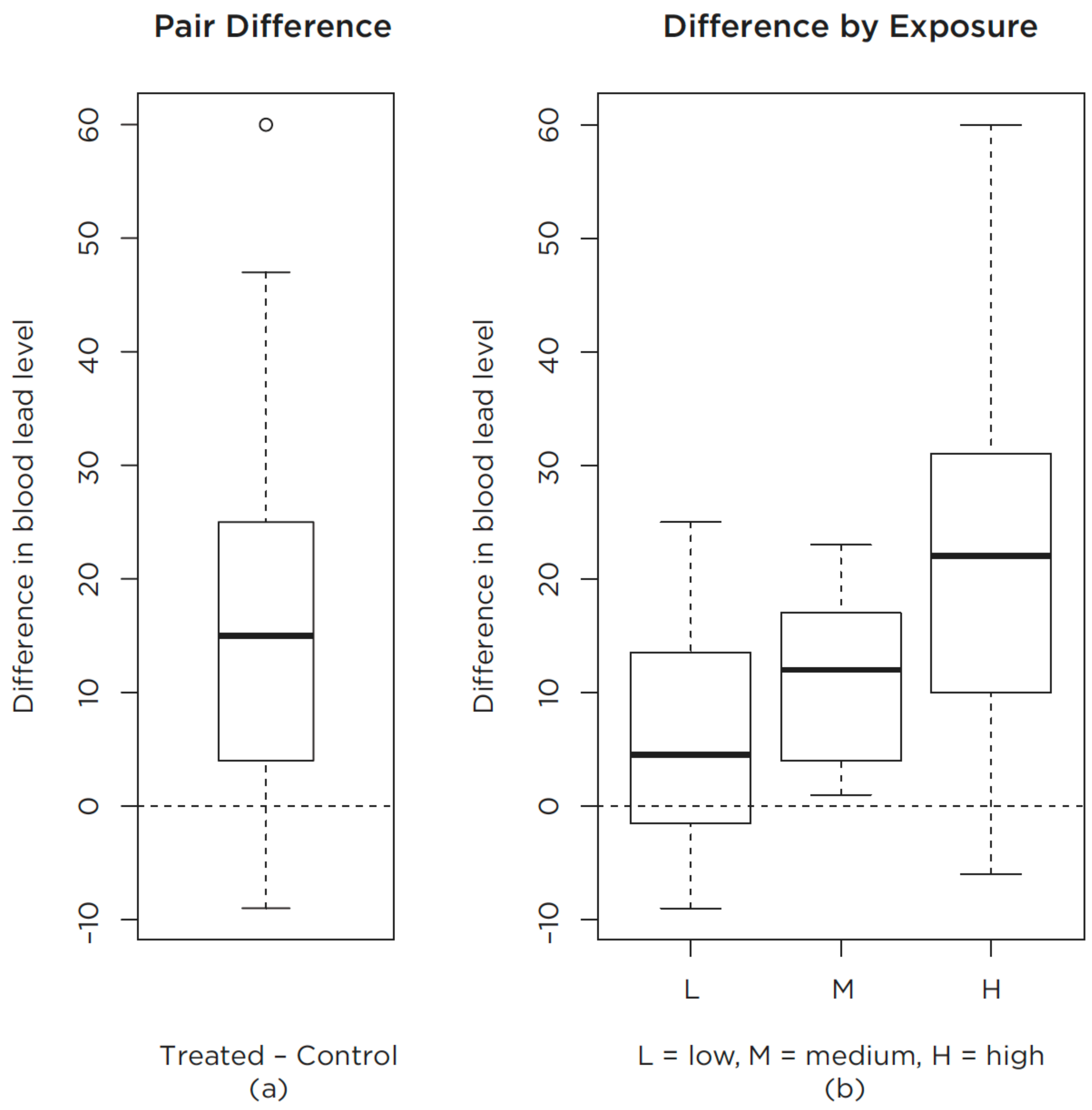


Figure 7.3. Matched pair differences, treated-minus-control, in levels of lead in children's blood, μg/dl. In each figure there is a horizontal line at zero. Panel (a) shows the differences, while panel (b) separates the differences into three groups based on the level of exposure to lead of the exposed father.

"**One might hope** that panel (a) of Figure 7.3 is analogous to a simple randomized experiment in which one child in each of 33 matched pairs was picked at random for exposure. **One might hope** that panel (b) of Figure 7.3 is analogous to a different simple randomized experiment in which levels of exposure were assigned to pairs at random. **One might hope** that panels (a) and (b) are jointly analogous to a randomized experiment in which both randomizations were done, within and among pairs. **All three of these hopes** may fail to be realized: there might be bias in treatment assignment within pairs or bias in assignment of levels of exposure to pairs."

— Rosenbaum, Paul. *Observation and Experiment*

text-data integration, readability improves with parallel structure between narrative and *sorted* table or graphic

empirical ordering, theoretical grouping

Decide on the main point you want to make about the data and arrange the rows and columns accordingly.

Ordering: for many tables or charts presenting distributions or associations, an important aim is to show which items have the highest and the lowest values and where other categories fall relative to those extremes.

Grouping: consider arranging items into conceptually related sets.

Alphabetical: ordering alphabetically is *rarely* the best approach but it is the default setting in many software tools. Take control over your displays.

parallel structure in narrative

When writing about the patterns shown in tables or charts, proceed systematically, describing the numbers in the same order as in those displays.

Another tip: if possible, use the same organizing principles in all the tables within a document, such as tables reporting descriptive statistics and multivariate results for the same set of variables.

text-data integration, parallel structure between narrative and *sorted* table or graphic — example

example empirical ordering

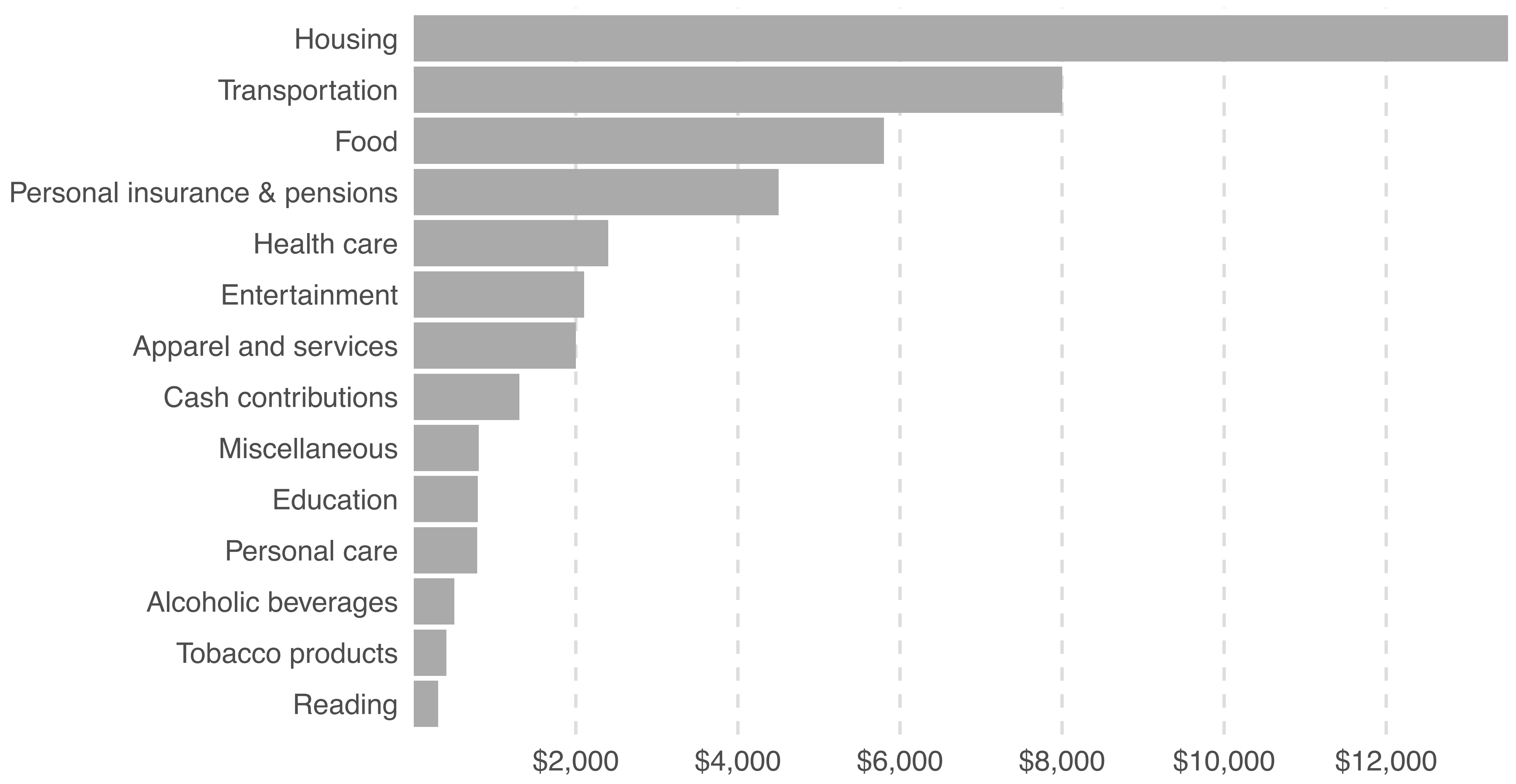


Fig. 3. Major categories of expenditures, descending dollar value, 2002 U.S. Consumer Expenditure Survey

example parallel structure in narrative

Figure 3 presents average consumer expenditures for the United States in 2002 in descending order of dollar value. Housing **was the highest** expenditure category, **followed by** transportation, food, and personal expenditures . . .

 narrative for our example table

Using Table 1, we can calculate the value of a strike by subtracting the expected run value of a strike, given the game state and count, from the value of a ball, starting from the same game state and count. Let's say there is a runner on first and second with one out, and the count is 1 ball, 1 strike, giving us 0.99 expected runs the rest of the inning. Assuming the batter doesn't swing on the next pitch, a strike lowers expected runs to 0.86 while a ball raises it to 1.11. Thus, in this scenario, the expected value of a strike would be 0.86 - 1.11, or 0.25 runs.

Using Table 1, we can calculate the value of a strike by subtracting the expected run value of a strike, given the game state and count, from the value of a ball, starting from the same game state and count. Let's say there is a runner on first and second with one out, and the count is 1 ball, 1 strike, suggesting we should expect **0.99** more runs this inning:

Table 1: For takes, our Bayesian model expects batting-team runs to increase by these amounts during the half-inning, given game state and count.

Game State	Count											
—:-	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2	3-0	3-1	3-2
No outs												
—:0	0.54	0.50	0.44	0.59	0.54	0.47	0.66	0.61	0.54	0.77	0.74	0.69
1–:0	0.96	0.89	0.79	1.04	0.96	0.85	1.16	1.08	0.96	1.34	1.29	1.22
-2-:0	1.17	1.10	1.00	1.25	1.18	1.06	1.37	1.29	1.17	1.55	1.50	1.43
–3:0	1.40	1.36	1.29	1.45	1.40	1.33	1.53	1.48	1.40	1.65	1.61	1.57
12-:0	1.54	1.47	1.35	1.65	1.55	1.42	1.79	1.69	1.55	2.01	1.95	1.86
1-3:0	1.80	1.71	1.57	1.91	1.81	1.65	2.08	1.96	1.80	2.33	2.26	2.16
-23:0	1.93	1.86	1.75	2.02	1.94	1.82	2.15	2.06	1.94	2.34	2.29	2.21
123:0	2.28	2.19	2.03	2.40	2.29	2.12	2.58	2.46	2.29	2.84	2.77	2.67
One out												
—:1	0.29	0.26	0.22	0.32	0.29	0.24	0.37	0.34	0.29	0.44	0.42	0.39
1–:1	0.56	0.51	0.43	0.62	0.56	0.48	0.71	0.65	0.56	0.84	0.81	0.75
-2-:1	0.71	0.66	0.58	0.77	0.71	0.63	0.86	0.80	0.71	1.00	0.96	0.91
–3:1	0.97	0.89	0.77	1.07	0.98	0.85	1.21	1.11	0.97	1.42	1.36	1.28
12-:1	0.98	0.91	0.79	1.07	**0.99**	0.86	1.21	1.11	0.98	1.41	1.35	1.27
1-3:1	1.21	1.13	0.98	1.33	1.22	1.07	1.50	1.38	1.22	1.75	1.68	1.58
-23:1	1.33	1.24	1.09	1.45	1.34	1.18	1.62	1.51	1.34	1.88	1.81	1.71
123:1	1.62	1.50	1.31	1.78	1.63	1.43	1.99	1.84	1.63	2.33	2.24	2.10
Two outs												
—:2	0.11	0.10	0.07	0.13	0.11	0.09	0.16	0.14	0.11	0.20	0.19	0.17
1–:2	0.23	0.20	0.15	0.27	0.24	0.18	0.33	0.29	0.23	0.42	0.40	0.36
-2-:2	0.30	0.26	0.20	0.36	0.31	0.24	0.44	0.38	0.31	0.55	0.52	0.48
–3:2	0.35	0.30	0.22	0.41	0.36	0.27	0.50	0.44	0.35	0.64	0.60	0.55
12-:2	0.45	0.39	0.29	0.53	0.46	0.35	0.65	0.57	0.46	0.82	0.77	0.70
1-3:2	0.50	0.43	0.32	0.59	0.50	0.38	0.71	0.62	0.50	0.90	0.85	0.77
-23:2	0.52	0.45	0.33	0.62	0.53	0.40	0.75	0.66	0.53	0.95	0.89	0.81
123:2	0.82	0.71	0.53	0.97	0.83	0.64	1.18	1.03	0.83	1.49	1.41	1.28

Note:
Expectations from model fit to Statcast data, 2017-2019.

Assuming the batter doesn't swing on the next pitch, a strike lowers expected runs to 0.86 while a ball raises it to 1.11. Thus, in this scenario, the expected value of a strike would be 0.86 - 1.11, or -0.25 runs.

bringing teachings together — *draft* proposal as example

data in narrative, **proposal as a multi-level narrative — title, headings, body, captions**

"Orderliness **adds credibility** to the information and **induces confidence.** Information presented with clear and logically set out titles, subtitles, texts, illustrations and captions will not only be read more quickly and easily but the information will also be **better understood.**"

— Müller-Brockmann, *Grid systems in graphic design*

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Senior Baseball Analyst, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on *expectations* of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

In a game against New York Yankees, should Milwaukee Brewers's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the *expectation* that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

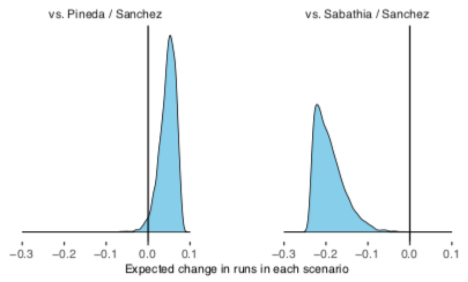


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

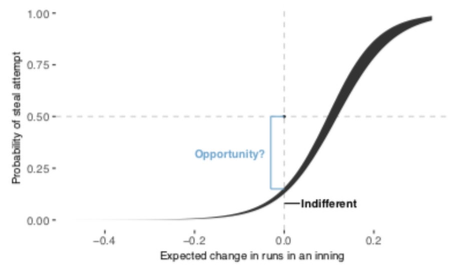


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

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 above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a *replacement-level* player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky: *we only see the outcome of our decisions.* But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et. al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." McKinsey Quarterly 13 June 2018: 1–9.

———. 2018b. "A view from the front lines of baseball's data-analytics revolution." McKinsey Quarterly 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/.

Readability Statistics

Counts	
Words	720
Characters	3,997
Paragraphs	16
Sentences	35
Averages	
Sentences per Paragraph	4.3
Words per Sentence	18.1
Characters per Word	5.3
Readability	
Flesch Reading Ease	33.2
Flesch-Kincaid Grade Level	13
Passive Sentences	0%

data in narrative, messages first, details follow

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Senior Baseball Analyst, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on *expectations* of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JO

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a *replacement-level* player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky: *we only see the outcome of our decisions.* But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et. al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." McKinsey Quarterly 13 June 2018: 1–9.

———. 2018b. "A view from the front lines of baseball's data-analytics revolution." McKinsey Quarterly 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

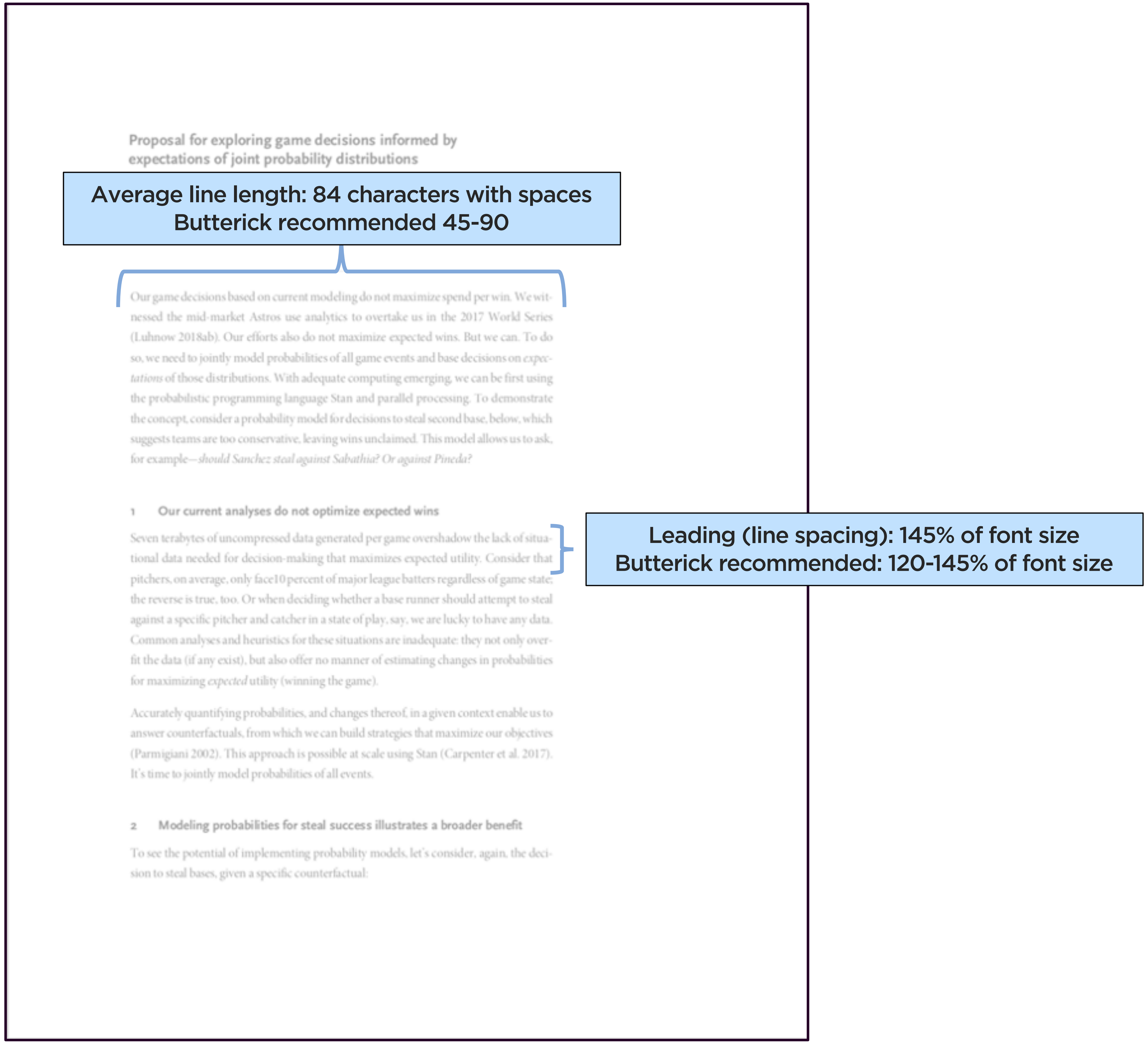
Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/.

Get our audience(s) to **pay attention** to,
understand,
(be able to) **act upon**

a maximum of **messages,**
given **constraints.**

— Doumont, *Trees, Maps, Theorems*

data in narrative, **best practices in typography**



Proposal for exploring game decisions informed by expectations of joint probability distributions

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on *expectations* of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

"Most readers are looking for reasons to stop reading. . . . Readers have other demands on their time. . . . The goal of most professional writing is persuasion, and attention is a prerequisite for persuasion. Good typography can help your reader devote less attention to the mechanics of reading and more attention to your message."

— Butterick, Matthew, *Practical Typography*

data in narrative, data graphics as paragraphs about data — linking narrative and data

“Words, graphics, and tables are different mechanisms with but a *single purpose*—the presentation of information. Why should the flow of information be broken up into different places on the page…?”

— Edward Tufte, *The Visual Display of Quantitative Information*

ness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

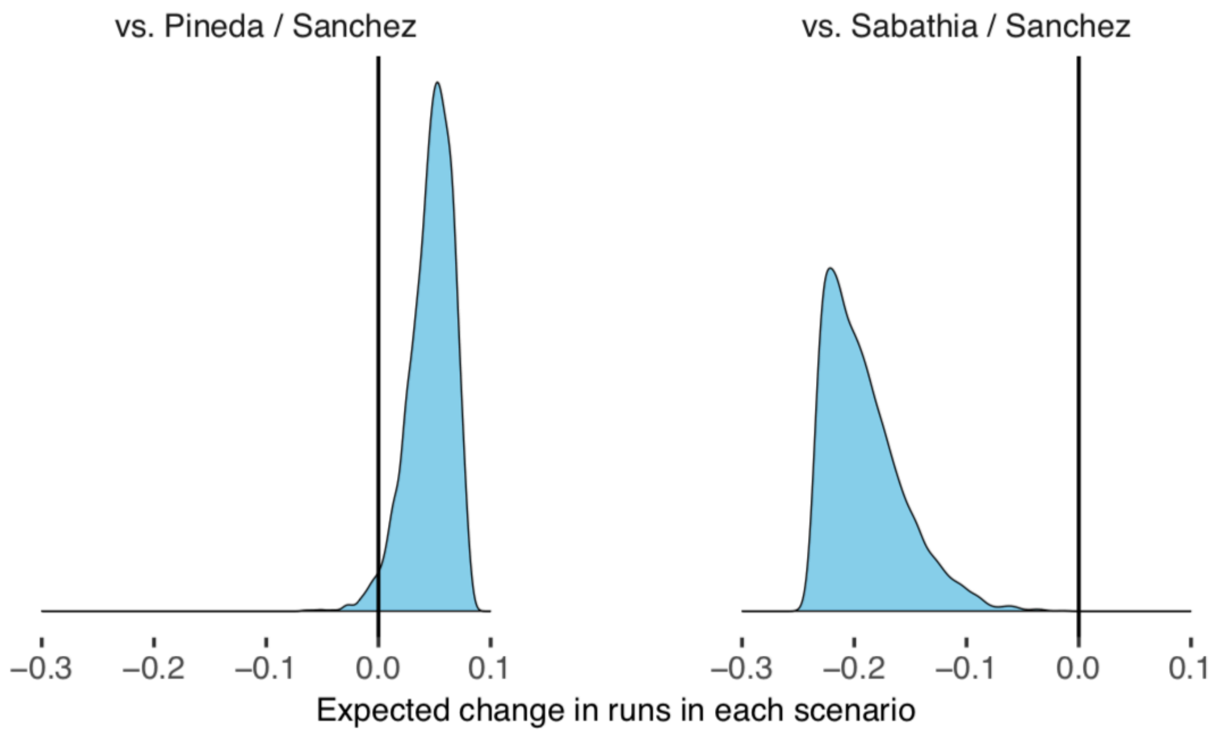


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

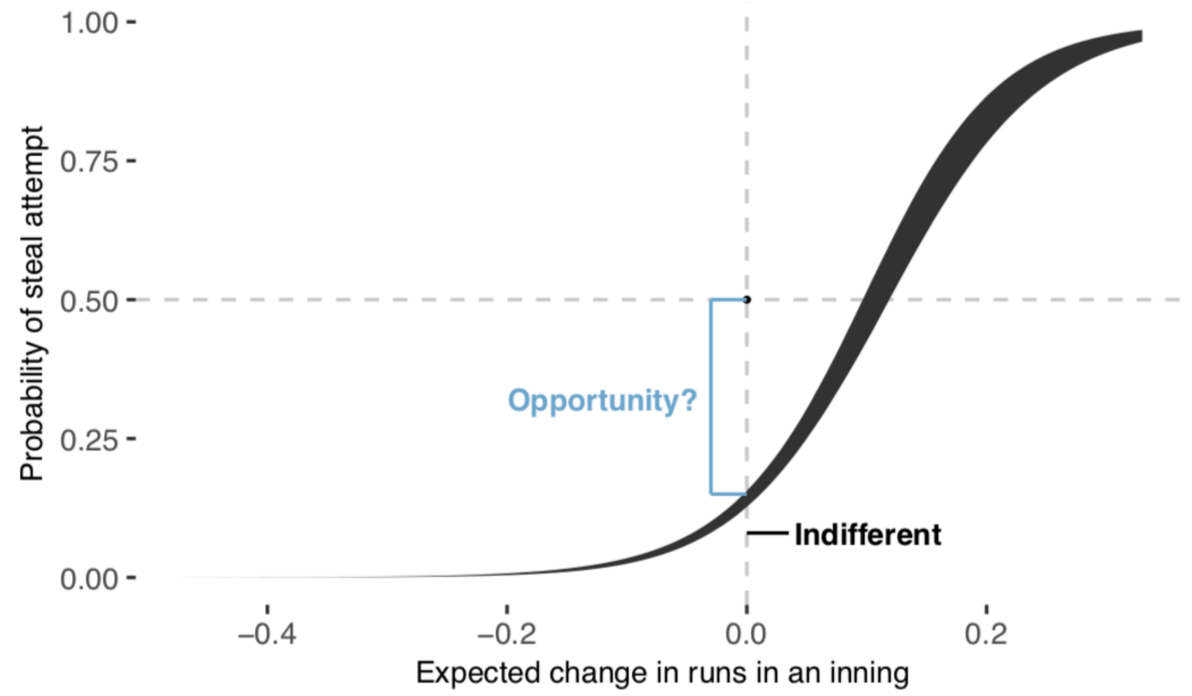


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

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 above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly

next deliverable, *your* proposal. content?

resources

References

Spencer, Scott. "Integrating text and data." In *Data in Wonderland*. 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Doumont, Jean-Luc. "Effective Written Documents." In *Trees, Maps, and Theorems. Effective Communication for Rational Minds*. Principiæ, 2009.

Farnsworth, Ward. *Classical English Rhetoric*. David R. Godine Publisher, 2011.

Few, Stephen. "Table Design." In *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Second edition. Burlingame, Calif: Analytics Press, 2012.

Harris, Robert L. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press, 1999.

Lakoff, George, and Mark Johnson. *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.

Miller, Jane E. "Organizing Data in Tables and Charts: Different Criteria for Different Tasks." Teaching Statistics 29, no. 3 (August 2007): 98–101. https://doi.org/10.1111/j.1467-9639.2007.00275.x.

———. "Seven Basic Principles" and "Creating Effective Tables." In *The Chicago Guide to Writing about Multivariate Analysis*, Second edition., 13–33. Chicago Guides to Writing, Editing, and Publishing. Chicago: University of Chicago Press, 2013.

Müller-Brockmann, Josef. *Grid Systems in Graphic Design. A Visual Communication Manual for Graphic Designers, Typographers, and Three Dimensional Designers.* ARTHUR NIGGLI LTD., 1996.

Rutter, Richard. "Arrangement and Composition". In *Web Typography. A Handbook for Designing Beautiful and Effective Responsive Typography*. 177-192. Ampersand Type, 2017.

Spencer, Scott. (Draft) Proposal to Scott Powers. "*Proposal for Exploring Game Decisions Informed by Expectations of Joint Probability Distributions*." February 14, 2019.

Tufte, Edward R. "Aesthetics and Technique in Data Graphical Design." In *The Visual Display of Quantitative Information*, 176–90. Graphics Press, 2001.

Ware, Colin. *Information Visualization: Perception for Design*. Fourth. Philadelphia: Elsevier, Inc, 2020.

Zhu, Hao. "*KableExtra: Construct Complex Table with 'kable' and Pipe Syntax.*" Manual, 2020. https://CRAN.R-project.org/package=kableExtra.