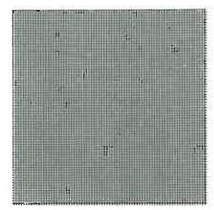
8 *High-Resolution Data Graphics*

OUR eyes can make a remarkable number of distinctions within a small area. With the use of very light grid lines, it is easy to locate 625 points in one square inch or, equivalently, 100 points in one square centimeter. Or consider how an 80 by 80 grid over a square inch—about 30 by 30 over a square centimeter—divides the space to make 25,281 distinct locations:[1]



[1] A square grid formed on each side by n parallel black and n-1 parallel white lines contains n^2 intersections of two black lines (corners of squares), $(n-1)^2$ intersections of two white lines (white squares), and 2n(n-1) intersections of a black and white line (sides of squares), for a total of $(2n-1)^2$ line intersections or distinct locations.

With the help of considerable redundancy and context, our eyes make fine distinctions of this sort all the time. Some measurement instruments used in engineering, architectural, and machine work are engraved with scales of 20 increments to the centimeter and 50 to the inch. Or consider the reading of fine print. Below, this type in the U.S. *Statistical Abstract* is set at 12 lines per vertical inch, with each line running at about 23 characters per inch for a maximum density of 276 characters per square inch. The overall density, given the white space, is 185 characters per square inch or 28 per square centimeter.

For serious data analysis, as in sports statistics, the data densities of tables are intense—as shown in this baseball box score at right. Millions of people routinely read high-resolution sports, weather, and financial data tables. Such tables provide an excellent model for all tables, even those in corporate presentations.

NO. 1450. STEEL PRODUCTS—NET SHIPMENTS, BY MARKET CLASSES: 1960 TO 1978

[In thousands of short tons. Comprises carbon, alloy, and stainless steel. "N.e.c." means not elsewhere classified]

MARKET CLASS	1960	1965	1970	1973	1974	1975	1976	1977	1978
Total [1]	**71,149**	**92,666**	**90,798**	**111,430**	**109,472**	**79,957**	**89,447**	**91,147**	**97,935**
Steel for converting and processing	2,928	3,932	3,443	4,714	4,486	3,255	4,036	3,679	4,612
Independent forgers, n.e.c.	841	1,250	1,048	1,213	1,339	1,098	952	998	1,192
Industrial fasteners [2]	1,071	1,234	1,005	1,278	1,331	675	912	848	870
Steel service centers, distributors	11,125	14,813	16,025	20,383	20,400	12,700	14,615	15,346	17,333
Construction, incl. maintenance	9,664	11,836	8,913	10,731	11,360	8,119	7,508	7,553	9,612
Contractors' products	3,602	5,018	4,440	6,459	6,249	3,927	4,502	4,500	3,480
Automotive	14,610	20,123	14,475	23,217	18,928	15,214	21,351	21,490	21,253
Rail transportation	2,525	3,805	3,098	3,228	3,417	3,152	3,056	3,238	3,549
Freight cars, passenger cars, locomotives	1,763	2,875	2,005	1,997	2,097	1,794	1,428	1,709	2,188
Rails and all other [3]	762	930	1,093	1,231	1,320	1,358	1,628	1,529	1,361
Shipbuilding and marine equip.	622	1,051	859	1,019	1,339	1,413	969	869	845
Aircraft and aerospace	78	94	56	69	79	69	59	63	60
Oil and gas industries	1,759	1,936	3,550	3,405	4,210	4,171	2,653	3,650	4,140
Mining, quarrying, and lumbering	288	392	497	534	644	596	536	486	508
Agricultural, incl. machinery	1,003	1,483	1,126	1,772	1,859	1,429	1,784	1,743	1,805
Machinery, industrial equip., tools	3,958	5,873	5,169	6,351	6,440	5,173	5,180	5,566	5,992
Electrical equipment	2,078	2,985	2,694	3,348	3,242	2,173	2,671	2,639	2,811
Appliances, utensils, and cutlery	1,760	2,179	2,160	2,747	2,412	1,653	1,950	2,129	2,094
Other domestic commercial equip.	1,959	2,179	1,778	1,990	1,941	1,390	1,813	1,846	1,889
Containers, packaging, shipping	6,429	7,331	7,775	7,811	8,218	6,053	6,914	6,714	6,595
Cans and closures	4,976	5,867	6,239	6,070	6,349	4,859	5,290	5,173	4,950
Ordnance and other military	165	289	1,222	918	654	405	219	193	207
Exports (reporting companies only)	2,563	2,078	5,985	3,138	3,961	1,755	1,839	1,076	1,224

NATIONALS 8, METS 1

Washington	**ab**	**r**	**h**	**bi**	**bb**	**so**	**avg.**
Maxwell cf	5	1	2	1	1	0	.429
Belliard 2b	4	0	2	1	1	0	.211
Zimmerman 3b	4	0	1	0	1	0	.284
Dunn 1b	4	0	0	0	0	1	.281
N.Johnson 1b	1	0	1	0	0	0	.373
Willingham lf	3	1	0	0	1	1	.133
Kearns rf	4	3	2	1	1	1	.256
Flores c	4	2	3	3	1	1	.327
Alb.Gonzalez ss	5	1	2	1	0	0	.262
Zimmermann p	3	0	0	0	0	2	.000
Hinckley p	1	0	0	0	0	1	.000
Mock p	0	0	0	0	0	0	—
Cintron ph	1	0	0	0	0	1	.000
K.Wells p	0	0	0	0	0	0	—
Totals	**39**	**8**	**13**	**7**	**6**	**8**	
Mets	**ab**	**r**	**h**	**bi**	**bb**	**so**	**avg.**
Jos.Reyes ss	3	0	1	0	1	0	.316
Dan.Murphy lf	3	0	0	0	1	0	.302
Beltran cf	4	1	1	0	0	1	.406
Delgado 1b	4	0	1	1	0	2	.250
D.Wright 3b	4	0	0	0	0	3	.271
Church rf	3	0	1	0	0	0	.357
Santos c	4	0	1	0	0	1	.294
Castillo 2b	2	0	0	0	0	0	.365
Fossum p	0	0	0	0	0	0	—
Sheffield ph	1	0	0	0	0	1	.136
Feliciano p	0	0	0	0	0	0	—
Stokes p	0	0	0	0	0	0	—
O.Perez p	1	0	1	0	0	0	.333
Cora 2b	1	0	0	0	1	0	.167
Totals	**30**	**1**	**6**	**1**	**3**	**8**	

Washington	**022**	**040**	**000—8**	**13**	**1**
Mets	**100**	**000**	**000—1**	**6**	**1**

E—Zimmermann (1), D.Wright (2). **LOB**—Washington 11, Mets 6. **2B**—Kearns (3), Alb.Gonzalez (3). **3B**—Delgado (1). **HR**—Flores (2), off O.Perez; Kearns (3), off O.Perez. **RBIs**—Maxwell (1), Belliard (1), Kearns (11), Flores 3 (10), Alb.Gonzalez (4), Delgado (16). **SB**—Maxwell 2 (3), Jos.Reyes (4). **CS**—Beltran (1). **DP**—Washington 2; Mets 1

Washington	**ip**	**h**	**r**	**er**	**bb**	**so**	**np**	**era**
Zmmmmn W2-0	5 1/3	6	1	1	2	5	103	2.38
Hinckley	1 2/3	0	0	0	0	1	13	2.08
Mock	1	0	0	0	1	1	13	3.00
K.Wells	1	0	0	0	0	1	12	2.70
Mets	**ip**	**h**	**r**	**er**	**bb**	**so**	**np**	**era**
O.Perez L1-2	4 1/3	9	7	7	3	3	92	9.31
Fossum	2 2/3	3	1	1	3	2	46	2.25
Feliciano	1	1	0	0	0	2	20	3.86
Stokes	1	0	0	0	0	1	14	0.00

T—2:52. **A**—40,023 (41,800).

Maps often present even finer detail. A cartographer writes that "the resolving power of the eye enables it to differentiate to 0.1 mm where provoked to do so. Clearly, therefore, conciseness is of the essence and high resolution graphics are a common denominator of cartography."[2] Distinctions at 0.1 mm mean 254 per inch.

[2] D. P. Bickmore, "The Relevance of Cartography," in John C. Davis and Michael J. McCullagh, eds., *Display and Analysis of Spatial Data* (London, 1975), 331.

How many statistical graphics take advantage of the ability of the eye to detect large amounts of information in small spaces? And how much information should graphics show? Let us begin by considering an empirical measure of graphical performance, the data density.

Data Density in Graphical Practice

The numbers that go into a graphic can be organized into a data matrix of observations by variables. Taking into account the size of the graphic in relation to the amount of data displayed yields the *data density*:

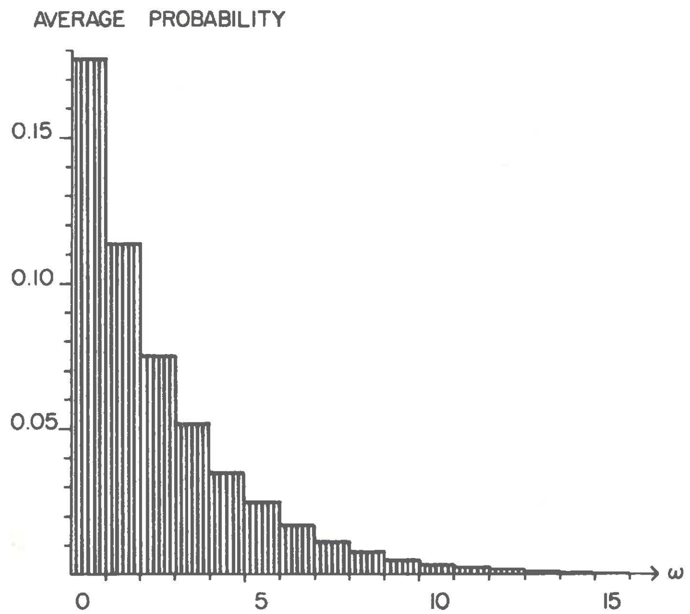
$$\text{data density of a display} = \frac{\text{number of entries in data matrix}}{\text{area of data display}}$$

Data matrices and data densities vary enormously in practice. At one extreme, this overwrought display (originally in five colors) presents a data matrix of four entries, the names and the numbers for the two bars on the right. The left bar is merely the total of the other two. The original graph covered 26.5 square inches (171 square centimeters), resulting in a data density of .15 numbers per square inch (.02 numbers per square centimeter), which is thin indeed. In contrast, the workday tables used in reporting sports data, which millions of people happily read every day, have data densities 100 times greater than the clunky display at right. And first-rate scientific graphics have data densities 1,000 times greater.

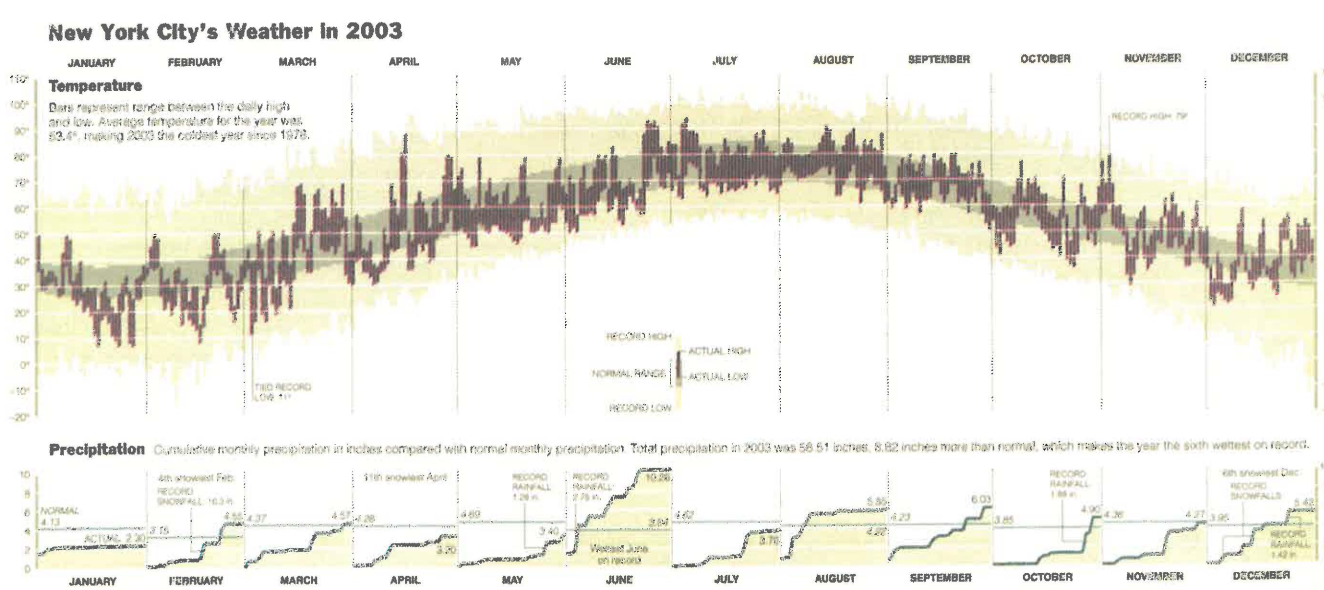
Executive Office of the President, Office of Management and Budget, *Social Indicators, 1973* (Washington, DC, 1973), 86.

High-density displays can be genuinely interactive, allowing viewers to select, to narrate, to recast and personalize data for their own thinking. Thus control of information is given over to viewers, not to editors, decorators, and badly designed computer graphics (such as PowerPoint and Excel). Data-thin displays move viewers toward ignorance and passivity, and at times diminish the credibility of the source. Thin data displays rightly prompt suspicions that the display-makers have cherry-picked their data: "What are they leaving out? Is that all they know? Is that all the analytical work they did? Do they think we're fools? Why are we having this meeting?"

This alleged exemplar from the style sheet of the *Journal of the American Statistical Association* comes in at a lightweight and unscientific 4 numbers per square inch, or 0.6 numbers per square centimeter, and a small data matrix of 32 entries:



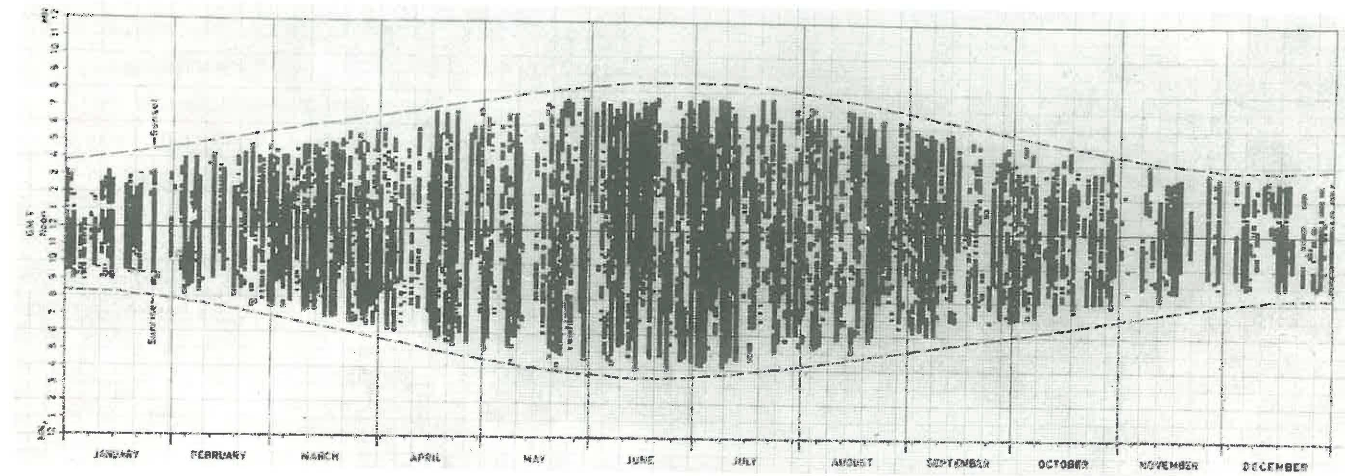
In luscious contrast, the New York weather history does very well at better than 300 numbers per square inch, or 45 per square centimeter:



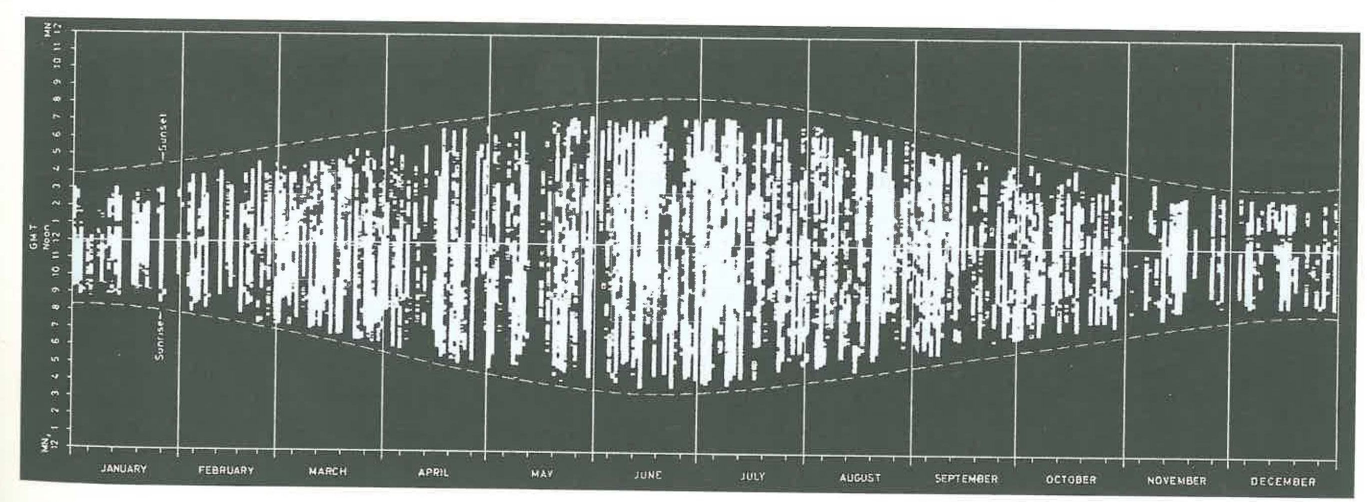
New York Times, January 4, 2004, A15.

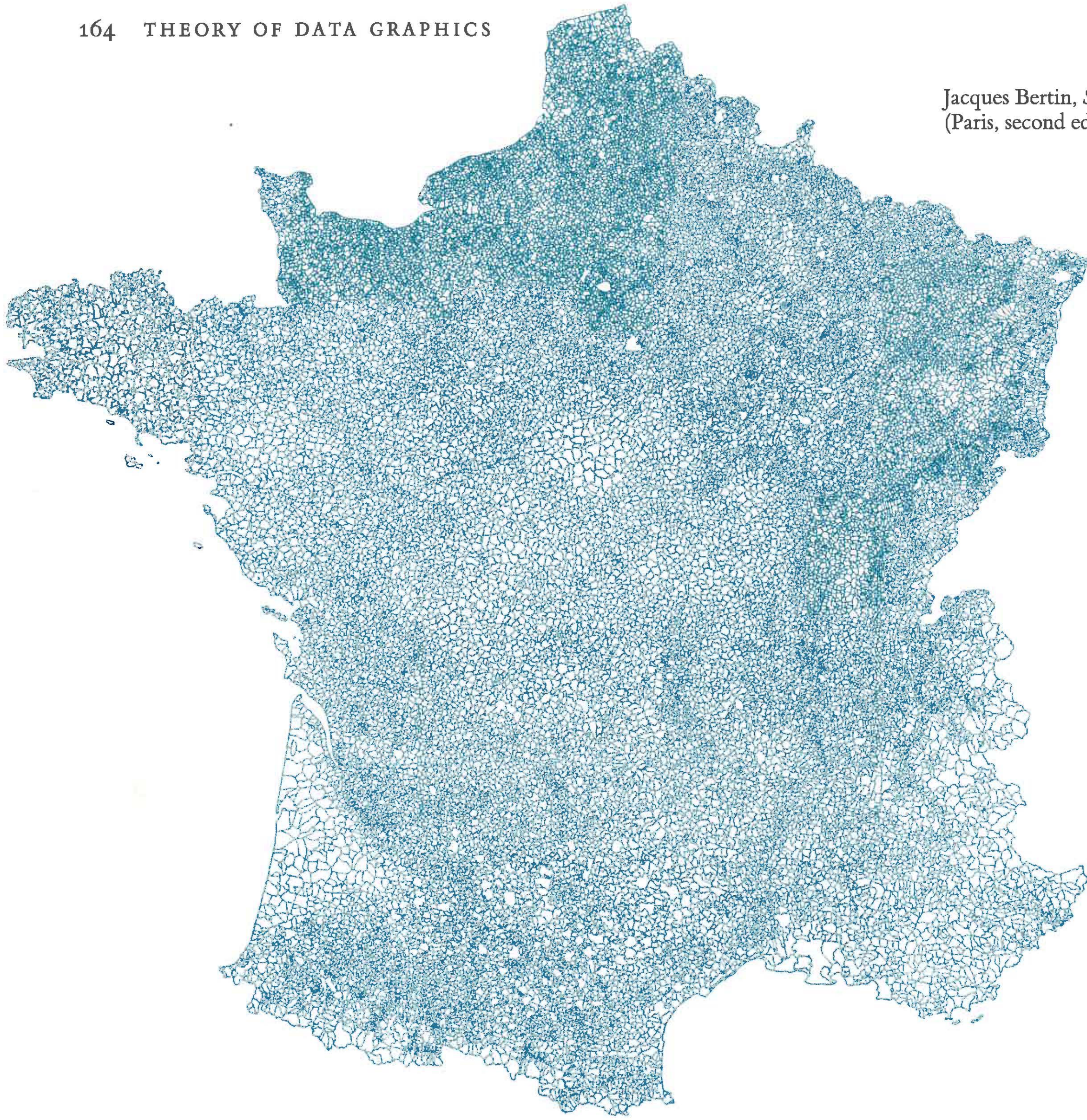
An annual sunshine record reports about 1,000 numbers per square inch, or 160 per square centimeter:

F. J. Monkhouse and H. R. Wilkinson, *Maps and Diagrams* (London, third edition, 1971), 242–243.



The visual metaphor corresponds more appropriately to the data if the image is reversed, so that the light areas are the times when the sun shines:

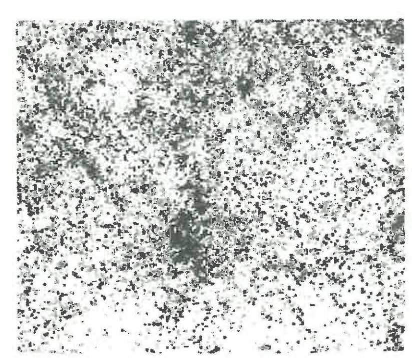




Jacques Bertin, *Semiologie Graphique* (Paris, second edition, 1973), 152.

This map (27 square inches, 175 square centimeters) shows the location and boundaries of 30,000 communes of France. It would require at least 240,000 numbers to recreate the data of the map (30,000 latitudes, 30,000 longitudes, and perhaps six numbers describing the shape of each commune). Thus that data density is nearly 9,000 numbers per square inch, or 1,400 numbers per square centimeter.

The new map of the galaxies (that we saw in Chapter 1) locates 2,275,328 encoded rectangles on a two-dimensional surface of 61 square inches (390 square centimeters). Each rectangle represents three numbers (two by its location, one by its shading), yielding a data density of 110,000 numbers per square inch or 17,000 numbers per square centimeter. That might be a world record.



Data Density and the Size of the Data Matrix: Publication Practices

The table shows the data density and the size of the data matrix for graphics sampled from scientific and news publications. At least 20 graphics from each publication were examined.

The table records an enormous diversity of graphical performances both within and between publications. A few data-rich designs appear in nearly every publication. The opportunity is there but it is rarely exploited: the average published graphic is rather thin,

Data Density and Size of Data Matrix, Statistical Graphics in Selected Publications, Circa 1979–1980

	Data Density (numbers per square inch)			Size of Data Matrix		
	median	minimum	maximum	median	minimum	maximum
Nature	48	3	362	177	15	3780
Journal of the Royal Statistical Society, B	27	4	115	200	10	1460
Science	21	5	44	109	26	316
Wall Street Journal	19	3	154	135	28	788
Fortune	18	5	31	96	42	156
The Times (London)	18	2	122	50	14	440
Journal of the American Statistical Association	17	4	167	150	46	1600
Asahi	13	2	113	29	15	472
New England Journal of Medicine	12	3	923	84	8	3600
The Economist	9	1	51	36	3	192
Le Monde	8	1	17	66	11	312
Psychological Bulletin	8	1	74	46	8	420
Journal of the American Medical Association	7	1	39	53	14	735
New York Times	7	1	13	35	6	580
Business Week	6	2	12	32	14	96
Newsweek	6	1	13	23	2	96
Annuaire Statistique de la France	6	1	25	96	12	540
Scientific American	5	1	69	46	14	652
Statistical Abstract of the United States	5	2	23	38	8	164
American Political Science Review	2	1	10	16	9	40
Pravda	0.2	0.1	1	5	4	20

based on about 50 numbers shown at the rate of 10 per square inch. Among the world's newspapers, the *Wall Street Journal, The Times* (London), and *Asahi* publish data-rich graphics, with data densities equal to those of the *Journal of the American Statistical Association*. Most of the American papers and magazines, along with *Pravda*, publish less data per graphic than the major papers of other industrialized countries.

Very few statistical graphics achieve the information display rates found in maps. Highly detailed maps portray 100,000 to 150,000 bits per square inch. For example, the average U.S. Geological Survey topographic quadrangle (measuring 17 by 23 inches) is estimated to contain over 100 million bits of information, or about 250,000 per square inch (40,000 per square centimeter).[3] Perhaps some day statistical graphics will perform as successfully as maps in carrying information.

High-Information Graphics

Data graphics should often be based on large rather than small data matrices and have a high rather than low data density. More information is better than less information, especially when the marginal costs of handling and interpreting additional information are low, as they are for most graphics. The simple things belong in tables or in the text; graphics can give a sense of large and complex data sets that cannot be managed in any other way. If the graphic becomes overcrowded (although several thousand numbers represented may be just fine), a variety of data-reduction techniques—averaging, clustering, smoothing—can thin the numbers out before plotting.[4] Summary graphics can emerge from high-information displays, but there is nowhere to go if we begin with a low-information design.

Data-rich designs provide a context and credibility for statistical evidence. High-density graphics help us to compare parts of the data by displaying much information within the common eyespan: we look at one display at a time and the more in the display, the more effective and comparative our eye can be.[5] The principle:

> Maximize data density and the size of the data matrix, within reason (but at the same time exploiting the *maximum resolution* of the available data-display technology).

High-information graphics must be designed with special care. As the quantity of data increases, data measures must shrink (smaller dots for scatters, thinner lines for busy time-series).

[3] Morris M. Thompson, *Maps for America* (Washington, DC, 1979), 187.

[4] Paul A. Tukey and John W. Tukey, "Summarization: Smoothing; Supplemented Views," in Vic Barnett, ed., *Interpreting Multivariate Data* (Chichester, England, 1982), ch. 12; and William S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74 (1979), 829–836.

[5] It is suggested in the analysis of x-ray films to "search a reduced image so that the whole display can be perceived on at least one occasion without large eye movement." Edward Llewellyn Thomas, "Advice to the Searcher or What Do We Tell Them?" in Richard A. Monty and John W. Senders, eds., *Eye Movements and Psychological Processes* (Hillsdale, New Jersey, 1976), 349.

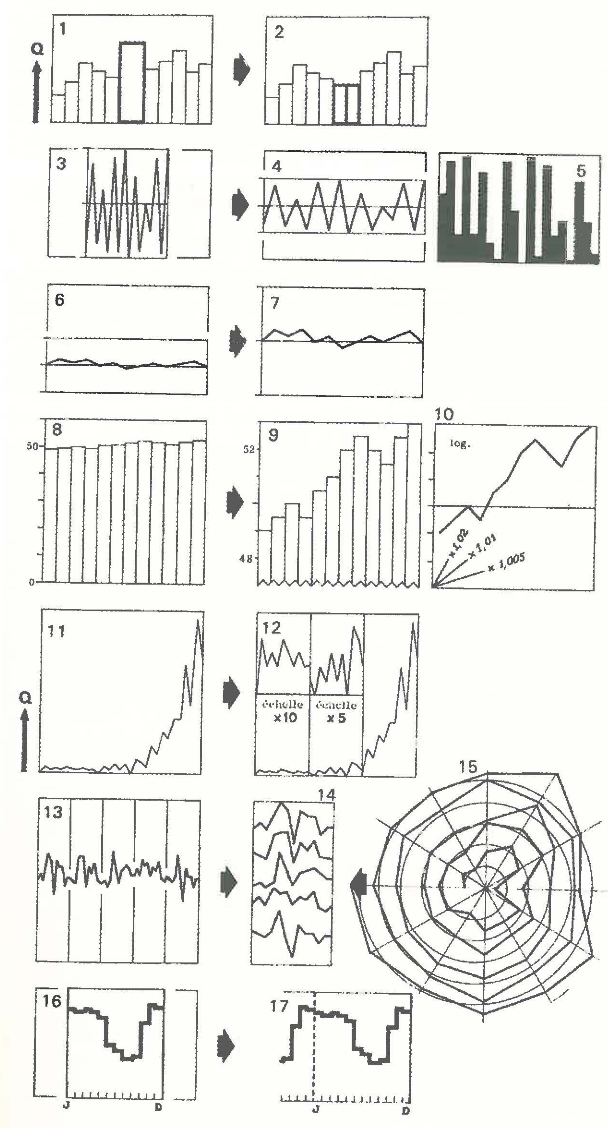
The clutter of chartjunk, wasted space, non-data-ink, and redundant data-ink is even more costly than usual in data-rich designs.

The way to increase data density other than by enlarging the data matrix is to reduce the area of a graphic. The Shrink Principle has surprisingly wide application:

> Graphics can be shrunk way down.

Many data graphics can be reduced in area by more than half their currently produced size with virtually no loss in legibility. For example, Bertin's crisp and elegant line allows the display of 17 small-scale graphics on a single page along with extensive text. Repeated application of the Shrink Principle leads to a powerful and effective graphical design, the small multiple.

Jacques Bertin, *Semiologie Graphique* (Paris, second edition, 1973), 214.



PROBLEMES GRAPHIQUES POSES PAR LES CHRONIQUES

Un total sur deux cases (sur deux ans) doit être divisé par deux (1).
Un total pour six mois sera multiplié par deux dans des cases annuelles.

Courbes trop pointues, réduire l'échelle des Q; la sensibilité angulaire s'inscrit dans une zone moyenne autour de 70°.
Si la courbe n'est pas réductible (grandes et petites variations) employer les colonnes remplies (5).
Courbes trop plates : augmenter l'échelle des Q.

Variations très faibles par rapport au total. Celui-ci perd de l'importance et le zéro peut être supprimé, à condition que le lecteur voit sa suppression (9). Le graphique peut être interprété comme une accélération si l'étude fine des variations est nécessaire (échelle logarithmique (10) (v. p. 240).

Très grande amplitude entre les valeurs extrêmes. Il faut admettre :
1°) Soit de ne pas percevoir les plus petites variations.
2°) Soit de ne s'intéresser qu'aux différences relatives (échelle logarithmique) sans connaître la quantité absolue.
3°) Soit admettre des périodes différentes dans la composante ordonnée et les traiter à des échelles différentes au-dessus de l'échelle commune (12).

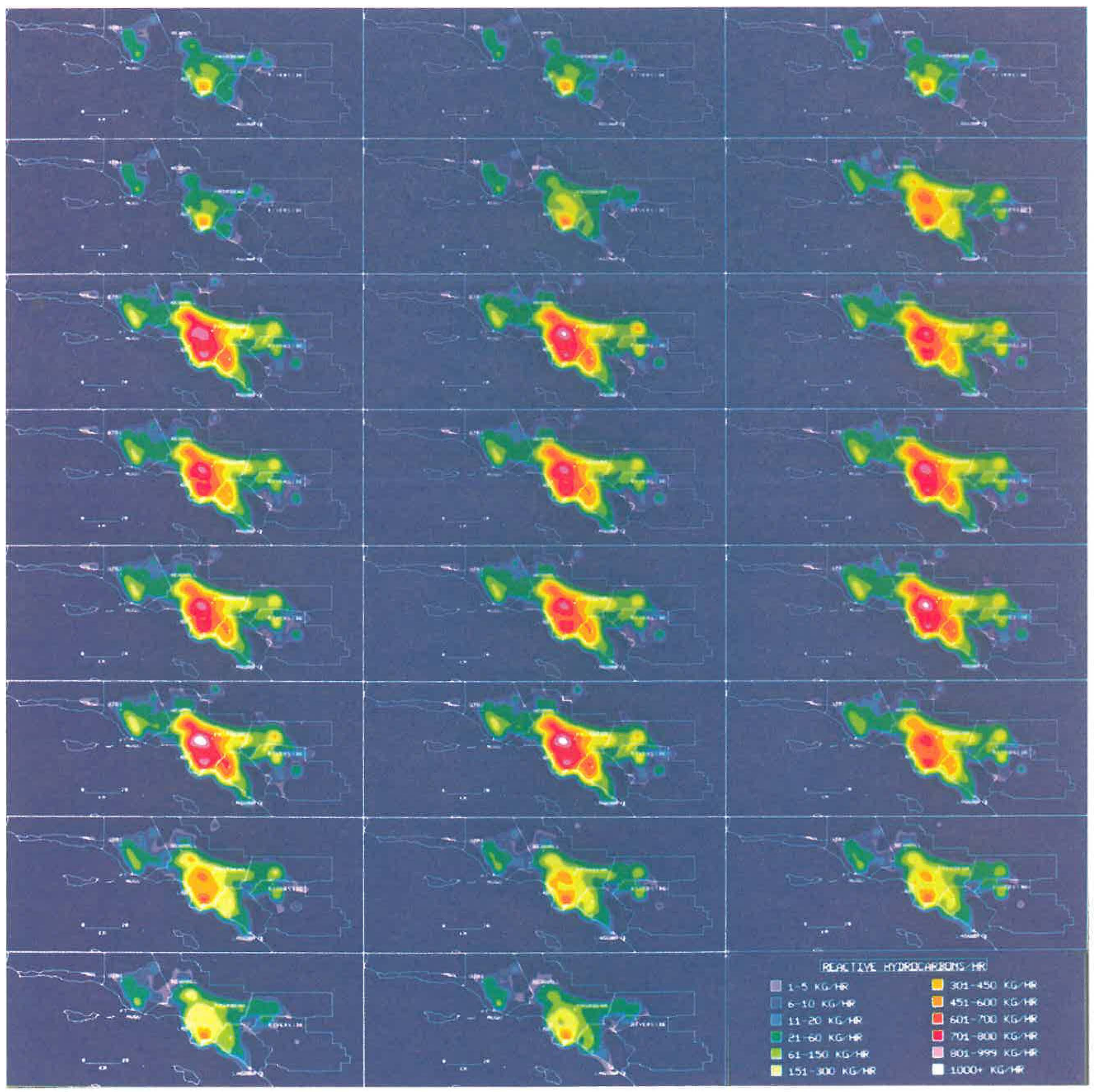
Cycles très marqués.
Si l'étude porte sur la comparaison des phases de chaque cycle, il est préférable de décomposer (13) de manière à superposer les cycles (14). La construction polaire peut être employée, de préférence dans une forme spirale (15) (ne pas commencer par un trop petit cercle); pour spectaculaire qu'elle soit, elle est moins efficace que la construction orthogonale.

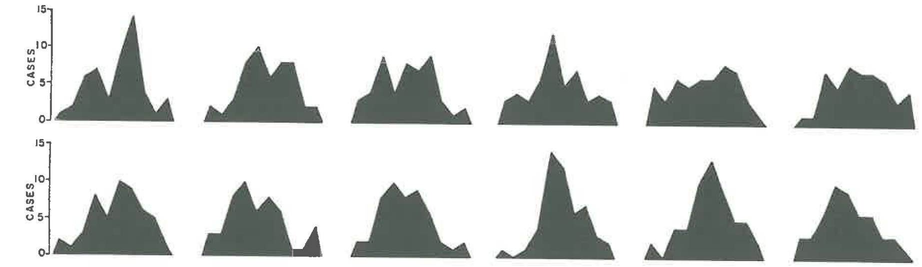
Courbes annuelles de pluie ou de température. Un cycle possède deux phases (17), pourquoi n'en offrir qu'une à la perception du spectateur? (16).

Small Multiples

Small multiples resemble the frames of a movie: a series of graphics, showing the same combination of variables, indexed by changes in another variable. Twenty-three hours of Los Angeles air pollution are organized into this display, based on a computer generated video tape. Shown is the hourly average distribution of reactive hydrocarbon emissions. The design remains constant through all the frames, so that attention is devoted to shifts in the data:

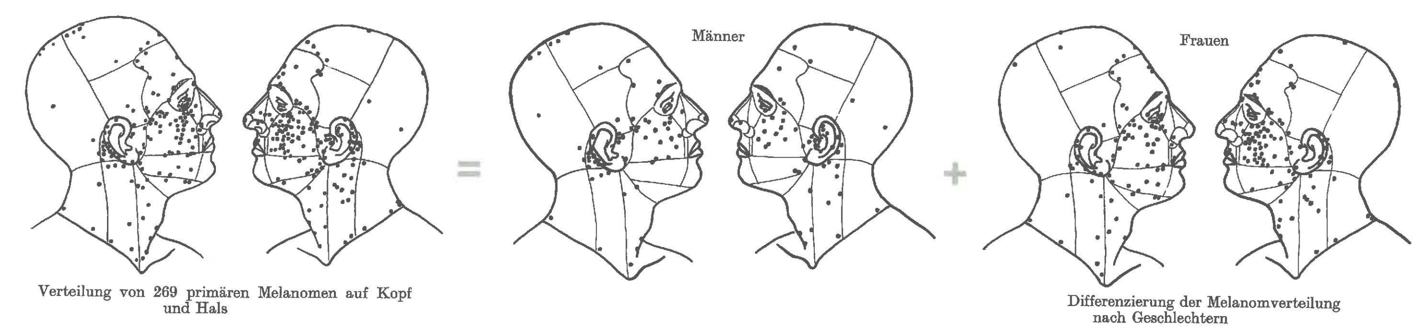
From video tape by Gregory J. McRae, California Institute of Technology. The model is described in G. J. McRae, W. R. Goodin, and J. H. Seinfeld, "Development of a Second-Generation Mathematical Model for Urban Air Pollution. I. Model Formulation," *Atmospheric Environment*, 16 (1982), 679–696.





Edmond A. Murphy, "One Cause? Many Causes? The Argument from the Bimodal Distribution," *Journal of Chronic Diseases*, 17 (1964), 309.

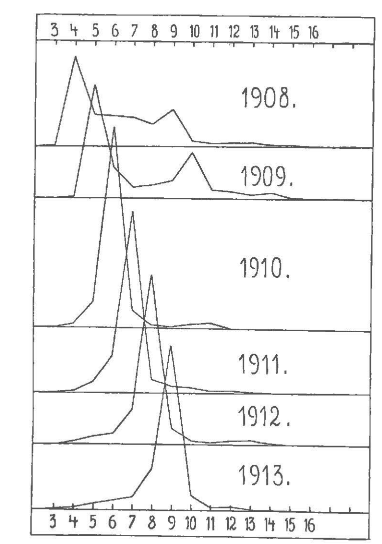
The 12 distributions above each result from 12 samples of 50 random normal deviates. Several of these utterly random distributions show patterns that might lead gullible researchers to jump to conclusions about bimodal distributions and, in turn, about multiple causes.



This grim small multiple above shows the spatial distribution of the cancer melanoma. Sites of all 269 primary melanomas are indicated, which are then split into separate distributions for men and women. Four variables are depicted: the cancer's three-dimensional location and gender. Small multiples are inherently multivariate, like nearly all interesting problems and solutions in data analysis.

Arthur Wiskemann, "Zur Melanomentstehung durch chronische Lichteinwirkung," *Der Hautarzt*, 25 (1974), 21.

Six distributions at right show the age composition of herring catches for each year from 1908 to 1913 in the great fisheries of Northern Europe. Notice the shifting age composition as the years go by. An enormous number of herring were spawned in 1904, and that class and cohort dominate the 1908 catch as four-year-olds, then the 1909 catch as five-year-olds, and so on. Aren't data wonderful?

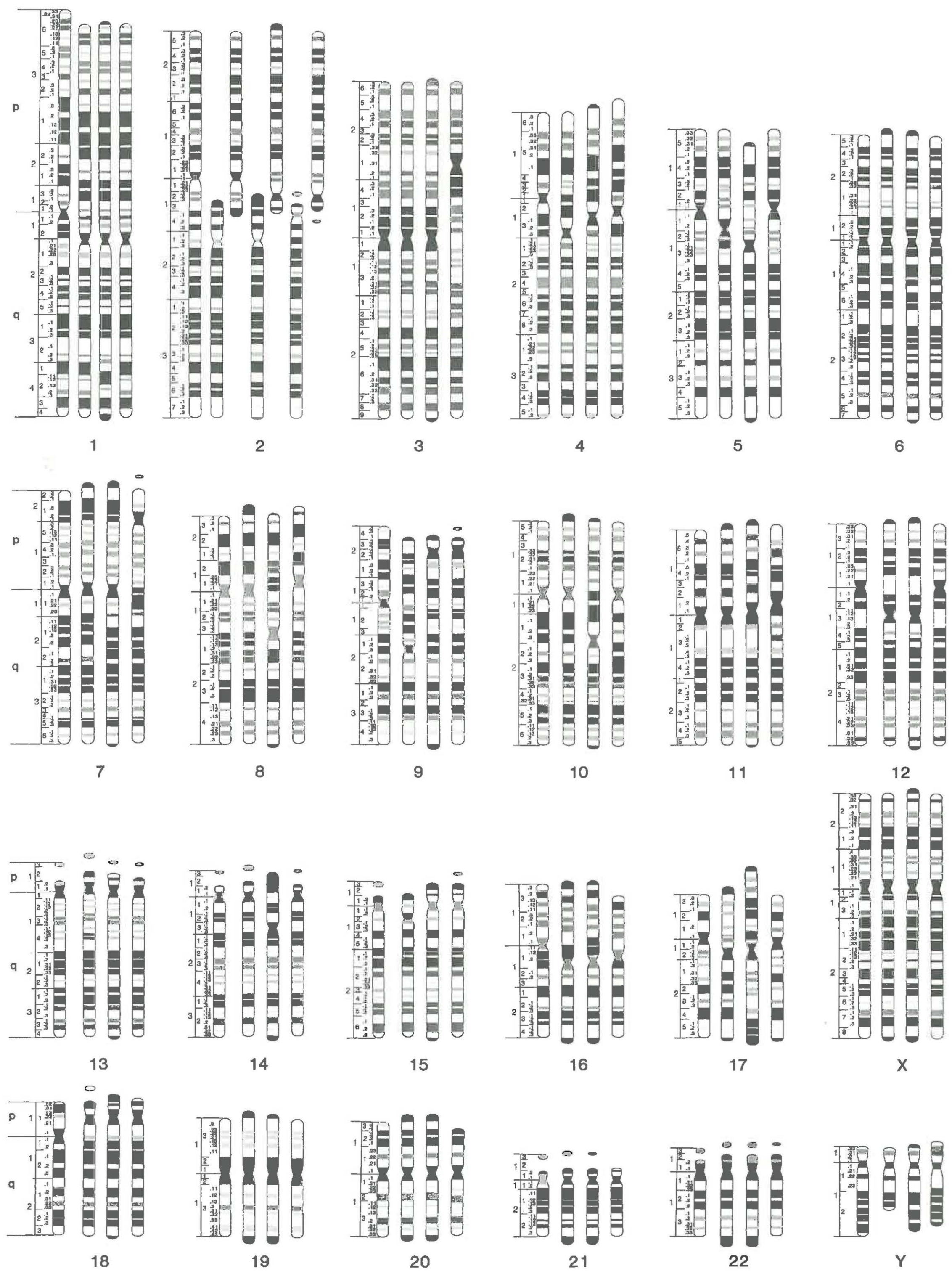


Johan Hjort, "Fluctuations in the Great Fisheries of Northern Europe," *Rapports et Proces-Verbaux*, 20 (1914), in Susan Schlee, *The Edge of an Unfamiliar World* (New York, 1973), 226.

Well-designed small multiples are inevitably comparative, deftly multivariate, shrunken, high-density graphics, usually based on a large data matrix drawn almost entirely with data-ink, efficient in interpretation, and often narrative in content–thereby showing shifts in the relationship between variables as the index variable changes (in turn revealing interaction or multiplicative effects). Small multiples are an excellent architecture for showing large quantities of multivariate data.

This notable display compares a complex set of data: shown are the chromosomes of (from left to right) human, chimpanzee, gorilla, and orangutan. The similarities between humans and the great apes are to be noted.

Jorge J. Yunis and Om Prakash, "The Origin of Man: A Chromosomal Pictorial Legacy," *Science*, 215 (March 19, 1982), 1527.



Sparklines: Intense, Simple, Word-Sized Graphics[6]

THE most common data display is a noun accompanied by a number. For example, a medical patient's current level of glucose is reported in a clinical record as a word and number:

glucose 6.6

Placed in the relevant context, a single number gains meaning. Thus the most recent measurement of glucose should be compared with earlier measurements for the patient. This data-line shows the path of the last 80 readings of glucose:

 glucose 6.6

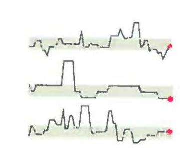
Lacking a scale of measurement, this free-floating line is dequantified. At least we do know the value of the line's right-most data point, which corresponds to the most recent value of glucose, the number recorded at far right. Both representations of the most recent reading are tied together with a color accent:

 glucose 6.6

Some useful context is provided by showing the *normal range* of glucose, here as a gray band. Compared to normal limits, readings above the band horizon are elevated, those below reduced:

 glucose 6.6 or glucose  6.6

For clinical analysis, the task is to detect quickly and assess wayward deviations from normal limits, shown here by visual deviations outside the gray band. Multiplying this format brings in additional data from the medical record; a stack, which can show hundreds of variables and thousands of measurements, allows fast effective parallel comparisons:

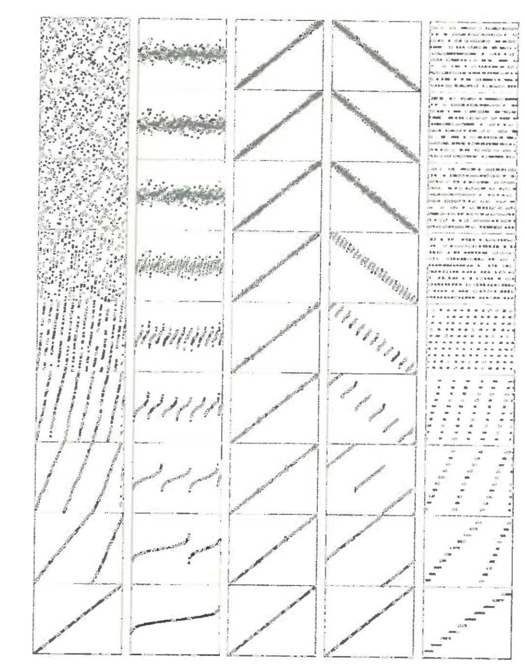


These little data lines, because of their active quality over time, are named *sparklines*—small, high-resolution graphics usually embedded in a full context of words, numbers, images. Sparklines are *datawords*: data-intense, design-simple, word-sized graphics.

Sparklines and sparkline-like graphs can also move within complex multivariate spaces, as in these 9-step sequential results (reading down the columns) in merge-sorting 5 different types of input files. Four variables and 18,000 numbers are depicted in these small multiples.

[6] See Edward Tufte, *Beautiful Evidence* (2006), 46–63. Recent work on sparklines and their open-source computer coding are posted at www.tufte.com.

Below, Robert Sedgewick, *Algorithms in C* (Reading, Massachusetts, 1998), 353.



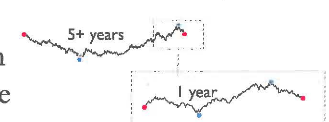
SPARKLINES have obvious applications for financial and economic data—by tracking and comparing changes over time, by showing overall trend along with local detail. Embedded in a data table, this sparkline depicts an exchange rate (dollar cost of one euro) for every day for one year:

		2003.4.28	12 months	2004.4.28	low	high
Euro foreign exchange	$	1.1025		1.1907	1.0783	1.2858

Colors help link the sparkline with the numbers: red = the oldest and newest rates in the series; blue = yearly low and high for daily exchange rates. Extending this graphic table is straightforward; here, the price of the euro versus 3 other currencies for 65 months and for 12 months:

		1999.1.1	65 months	2004.4.28	low	high		2003.4.28	12 months	2004.4.28	low	high
Euro foreign exchange	$	1.1608		1.1907	.8252	1.2858	$	1.1025		1.1907	1.0783	1.2858
Euro foreign exchange	¥	121.32		130.17	89.30	140.31	¥	132.54		130.17	124.80	140.31
Euro foreign exchange	£	0.7111		0.6665	.5711	0.7235	£	0.6914		0.6665	0.6556	0.7235

Daily sparkline data can be standardized and scaled in all sorts of ways depending on the content: by the range of the price, inflation-adjusted price, percent change, percent change off of a market baseline. Thus *multiple sparklines* can describe the same noun, just as multiple columns of numbers report various measures of performance. These sparklines reveal the details of the most recent 12 months in the context of a 65-month daily sequence (shown in the fractal-like structure at right).



Consuming a horizontal length of only 14 letterspaces, each sparkline in the big table above provides a look at the price and the changes in price for every day for years, and the overall time pattern. *This financial table reports 24 numbers accurate to 5 significant digits; the accompanying sparklines show about 14,000 numbers readable from 1 to 2 significant digits. The idea is to be approximately right rather than exactly wrong.*[7]

By showing recent changes in relation to many past changes, sparklines provide a context for nuanced analysis—and, one hopes, better decisions. Moreover, the year-long daily history reduces *recency bias,* the persistent and widespread over-weighting of recent events in making decisions. Tables sometimes reinforce recency bias by showing only current levels or recent changes; sparklines improve the attention span of tables.

Tables of numbers attain maximum densities of only 300 characters per square inch or 50 characters per square centimeter. In contrast, graphical displays have far greater resolutions; recall that "the resolving power of the eye enables it to differentiate to 0.1 mm where provoked to do so."[8] Distinctions at 0.1 mm mean 250 per linear inch, which implies 60,000 per square inch or 10,000 per square centimeter, which is plenty.

[7] On being "approximately right rather than exactly wrong," see John W. Tukey, "The Technical Tools of Statistics," *American Statistician,* 19 (1965), 23–28.

[8] D. P. Bickmore, "The Relevance of Cartography," in J. C. Davis and M. J. McCullagh, eds., *Display and Analysis of Spatial Data* (London, 1975), 331.

Here is a conventional financial table comparing various return rates of 10 popular mutual funds:

"Favorite Funds," *The New York Times,* August 10, 2003, 3-1.

Popular mutual funds, based on assets under management.

ASSETS (MIL.)	FUND	RETURN 4 WKS.	2003	3-YR.	5-YR.
$64,368	Vanguard Index 500 Index	- 2.0%	+12.2%	- 11.7%	- 0.8%
62,510	Fidelity Magellan	- 2.1	+11.3	- 12.9	- 0.2
50,329	Amer A Invest Co of Am	- 1.2	+09.4	- 3.9	+ 4.0
47,355	Amer A WA Mutual Inv	- 1.5	+09.9	+ 00.8	+ 3.0
40,500	PIMCO Instl Tot Return	- 2.3	+02.4	+ 09.4	+ 7.6
37,641	Amer A Grow Fd of Amer	- 2.9	+14.1	- 11.0	+ 7.4
31,161	Fidelity Contrafund	- 1.0	+10.7	- 6.5	+ 3.0
28,296	Fidelity Growth & Inc	- 1.8	+ 8.2	- 8.7	- 0.1
25,314	Amer A Inc Fund of Amer	- 0.5	+ 9.9	+ 05.5	+ 5.4
24,155	Vanguard Instl Index	- 2.0	+12.3	- 11.6	- 0.7

This is a common display in data analysis: a list of nouns (mutual funds, for example) along with some numbers (assets, changes) that accompany the nouns. The analyst's job is to look over the data matrix and then decide whether or not to go crazy—or at least to make a decision (buy, sell, hold) about the noun based on the data. But along with the summary clumps of tabular data, let us also look at the day-to-day path of prices and their changes for the entire last year. Here is the sparkline table:[9]

	$64,368	Vanguard 500 Index	-2.0%	+12.2%	-11.7%	-0.8%
	62,510	Fidelity Magellan	-2.1	+11.3	-12.9	-0.2
	50,329	Amer A Invest Co Am	-1.2	+09.4	-03.9	+4.0
	47,355	Amer A WA Mutual Inv	-1.5	+09.9	+00.8	+3.0
	40,500	PIMCO Instl Tot Return	-2.3	+02.4	+09.4	+7.6
	37,641	Amer A Grow Fd Amer	-2.9	+14.1	-11.0	+7.4
	31,161	Fidelity Contrafund	-1.0	+10.7	-06.5	+3.0
	28,296	Fidelity Growth & Inc	-1.8	+08.2	-08.7	-0.1
	25,314	Amer A Inc Fund Amer	-0.5	+09.9	+05.5	+5.4
	24,155	Vanguard Instl Index	-2.0	+12.3	-11.6	-0.7

[9] In our redesigned table, the typeface Gill Sans does quite well compared to the Helvetica in the original *Times* table. Smaller than the Helvetica, the Gill Sans appears sturdier and more readable, in part because of the increased white space that results from its smaller x-height and reduced size. The data area (without column labels) for our sparkline table is only 21% larger than the original's data area, and yet the sparklines provide an approximate look at 5,000 more numbers.

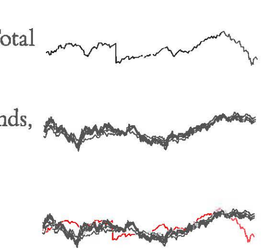
Astonishing and disconcerting, the finely detailed similarities of these daily sparkline histories are not all that surprising, after the fact anyway. Several funds use market index-tracking or other copycat strategies, and all the funds are driven daily by the same amalgam of external forces (news, fads, economic policies, panics, bubbles). Of the 10 funds, only the unfortunately named PIMCO, the sole bond fund in the table, diverges from the common pattern of the 9 stock funds, as seen by comparing PIMCO's sparkline with the stacked pile of 9 other sparklines at right.

In paper financial tables, down the deep columns of numbers, sparklines can be added to tables set at 8 lines per inch (as in our example above). This yields about 160 sparklines per column, or *400,000 additional daily graphical prices and their changes* on an A3 or 11 by 17 inch paper. Readers can scan sparkline-tables, making simultaneous multiple comparisons, searching for nonrandom patterns in the random walks of prices.

PIMCO Instl Total Return fund

All 9 other funds, overlapped

All 10 funds, overlapped, PIMCO in red



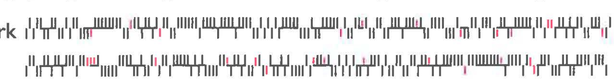
Sparklines efficiently display and narrate binary data (presence/absence, occurrence/non-occurrence, win/loss). For baseball, here is the season-long history of wins (upward whisker) and losses (downward) for one team, . This 162-game graphical sequence enables sports fans to recall the poor start, the first few scattered wins, another losing streak, and so on through the entire season.

Sparklines that depict on going sequences can help report all sorts of daily news. This sparkline, embedded in a sentence, is a *dataword* defined as "the win/loss sequence for the season's first 39 games." ⟶ Most sports fans will quickly derive the sparkline's meaning from the context; nearly all readers will understand the meaning the second time a sparkline is published. That's pretty good for a brand new word.

ORIOLES EXTEND WIN STREAK

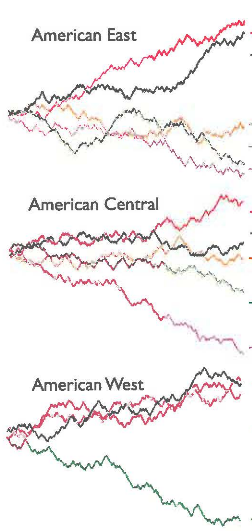
Yesterday the surging Baltimore Orioles won their fourth exciting game in a row. Still the team has hardly recovered from a poor start,with a season record so far of 8 wins and 31 losses ,
currently the worst in baseball.

Sparklines can simultaneously accommodate several variables. A team plays on its own field and on opposition fields, a distinction indicated graphically by a horizontal line for home games and no line for away games, . The apparent serial correlation (short waves of greater or lesser success) in this time-series is the result of an alternating schedule of home and away games, as teams are more successful at home. A useful strategy for data displays is *to multiply a good design*. These stacked sparklines compare seasons:

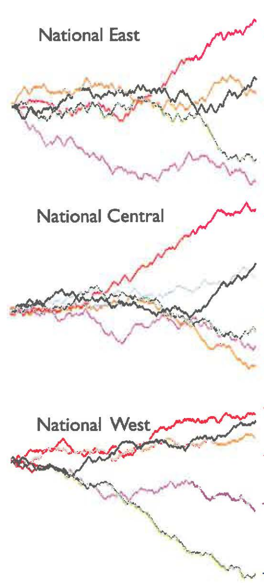


A short red whisker indicates the losing team was held scoreless. These 2 sparklines depict and compare 5 variables (ordered sequence of games, win/loss, home/away, no shutout/shutout, and team) for 162 games.

Below, 6 *paragraphs of sparklines* tell the story of the 2004 season for 6 divisions, showing competitive paths (wins - losses = net games over .500) for 30 baseball teams for all 162 games played by each team. The tables report 500 digits; these sparklines trace out 4,856 win/loss outcomes:



				home	road	pct
American East	+40	NY Yankees	101-61	57-24	44-37	.623
	+34	Boston	98-64	55-26	43-38	.605
	− 6	Baltimore	78-84	38-43	40-41	.481
	−22	Tampa Bay	70-91	41-39	29-52	.435
	−27	Toronto	67-94	40-41	27-53	.416
American Central	+23	Minnesota	92-70	49-32	43-38	.568
	+ 4	Chicago Sox	83-79	46-35	37-44	.512
	− 3	Cleveland	80-82	44-37	36-45	.494
	−18	Detroit	72-90	38-43	34-47	.444
	−46	Kansas City	58-104	33-47	25-57	.358
American West	+22	LA Angels	92-70	45-36	47-34	.568
	+20	Oakland	91-71	52-29	39-42	.562
	+16	Texas	89-73	51-30	38-43	.549
	−36	Seattle	63-99	38-44	25-55	.389



				home	road	pct
National East	+30	Atlanta	96-66	49-32	47-34	.593
	+10	Philadelphia	86-76	42-39	44-37	.531
	+ 4	Florida	83-79	42-38	41-41	.512
	−20	NY Mets	71-91	38-43	33-48	.438
	−28	Washington	67-95	35-45	32-50	.414
National Central	+48	St. Louis	105-57	53-28	52-29	.648
	+22	Houston	92-70	48-33	44-37	.568
	+16	Chicago Cubs	89-73	45-37	44-36	.549
	−10	Cincinnati	76-86	40-41	36-45	.469
	−17	Pittsburgh	72-89	39-41	33-48	.447
	−27	Milwaukee	67-94	36-45	31-49	.416
National West	+24	LA Dodgers	93-69	49-32	44-37	.574
	+20	San Francisco	91-71	47-35	44-36	.562
	+12	San Diego	87-75	42-39	45-36	.537
	−26	Colorado	68-94	38-43	30-51	.420
	−60	Arizona	51-111	29-52	22-59	.315

THE DATAWORDS of sparklines vastly increase the amount of data within our eyespan. Operating at the resolution of good typography, sparklines can be everywhere a number or word can be. With resolutions 5 to 100 times greater than conventional graphics and tables, sparklines give us a better chance to learn from the vast flood of numbers produced by modern measurement, monitoring, and surveillance. By providing a straightforward and contextual look at intense evidence, sparklines may help us, in John Tukey's words, to find an approximate answer to the right question (rather than an exact answer to the wrong question).

Approximately 5% of the graphics published in major scientific journals depict data with sparkline-like resolutions, as in chromosome sequencing, event monitoring, acoustical analysis, and many sorts of mappings. High-resolution graphics (200 numbers per square centimeter, or 1200 per square inch) help describe, explore, present, and understand the huge data sets of scientific research. By 2009, the median data-graphic published in *Nature* and *Science* presented >1200 numbers. For these major scientific journals during the past 10 years, the data densities of published graphics have doubled as scientific measurement and resolution have greatly increased. Serious amounts of data (as in science, engineering, medicine, finance, sports, weather) require high-resolution statistical graphics and tables. How else can we see the data?

For non-data-ink, less is more.

For data-ink, less is a bore.[10]

[10] The two aphorisms on the meaning of "less" are usually credited to the architects Ludwig Mies van der Rohe and Robert Venturi.