Edward R. Tufte

Visual Explanations

Images and Quantities, Evidence and Narrative

Graphics Press · Cheshire, Connecticut

Published by Graphics Press
Post Office Box 430, Cheshire, Connecticut 06410

Printed in the United States of America

Contents

2 *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*

WHEN we reason about quantitative evidence, certain methods for displaying and analyzing data are better than others. Superior methods are more likely to produce truthful, credible, and precise findings. The difference between an excellent analysis and a faulty one can sometimes have momentous consequences.

This chapter examines the statistical and graphical reasoning used in making two life-and-death decisions: how to stop a cholera epidemic in London during September 1854; and whether to launch the space shuttle Challenger on January 28, 1986. By creating statistical graphics that revealed the data, Dr. John Snow was able to discover the cause of the epidemic and bring it to an end. In contrast, by fooling around with displays that obscured the data, those who decided to launch the space shuttle got it wrong, terribly wrong. For both cases, the consequences resulted directly from the *quality* of methods used in displaying and assessing quantitative evidence.

The Cholera Epidemic in London, 1854

In a classic of medical detective work, *On the Mode of Communication of Cholera*,[1] John Snow described—with an eloquent and precise language of evidence, number, comparison—the severe epidemic:

> The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there were upwards of five hundred fatal attacks of cholera in ten days. The mortality in this limited area probably equals any that was ever caused in this country, even by the plague; and it was much more sudden, as the greater number of cases terminated in a few hours. The mortality would undoubtedly have been much greater had it not been for the flight of the population. Persons in furnished lodgings left first, then other lodgers went away, leaving their furniture to be sent for. . . . Many houses were closed altogether owing to the death of the proprietors; and, in a great number of instances, the tradesmen who remained had sent away their families; so that in less than six days from the commencement of the outbreak, the most afflicted streets were deserted by more than three-quarters of their inhabitants.[2]

[1] John Snow, *On the Mode of Communication of Cholera* (London, 1855). An acute disease of the small intestine, with severe watery diarrhea, vomiting, and rapid dehydration, cholera has a fatality rate of 50 percent or more when untreated. With the rehydration therapy developed in the 1960s, mortality can be reduced to less than one percent. Epidemics still occur in poor countries, as the bacterium *Vibrio cholerae* is distributed mainly by water and food contaminated with sewage. See Dhiman Barua and William B. Greenough III, eds., *Cholera* (New York, 1992); and S. N. De, *Cholera: Its Pathology and Pathogenesis* (Edinburgh, 1961).

[2] Snow, *Cholera*, p. 38. See also *Report on the Cholera Outbreak in the Parish of St. James's, Westminster, during the Autumn of 1854*, presented to the Vestry by The Cholera Inquiry Committee (London, 1855); and H. Harold Scott, *Some Notable Epidemics* (London, 1934).

Cholera broke out in the Broad Street area of central London on the evening of August 31, 1854. John Snow, who had investigated earlier epidemics, suspected that the water from a community pump-well at Broad and Cambridge Streets was contaminated. Testing the water from the well on the evening of September 3, Snow saw no suspicious impurities, and thus he hesitated to come to a conclusion. This absence of evidence, however, was not evidence of absence:

> Further inquiry . . . showed me that there was no other circumstance or agent common to the circumscribed locality in which this sudden increase of cholera occurred, and not extending beyond it, except the water of the above mentioned pump. I found, moreover, that the water varied, during the next two days, in the amount of organic impurity, visible to the naked eye, on close inspection, in the form of small white, flocculent [loosely clustered] particles. . . .[3]

From the General Register Office, Snow obtained a list of 83 deaths from cholera. When plotted on a map, these data showed a close link between cholera and the Broad Street pump. Persistent house-by-house, case-by-case detective work had yielded quite detailed evidence about a possible cause-effect relationship, as Snow made a kind of streetcorner correlation:

> On proceeding to the spot, I found that nearly all of the deaths had taken place within a short distance of the pump. There were only ten deaths in houses situated decidedly nearer to another street pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pump which was nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street. Two of them were known to drink the water; and the parents of the third think it probable that it did so. The other two deaths, beyond the district which this pump supplies, represent only the amount of mortality from cholera that was occurring before the irruption took place.
>
> With regard to the deaths occurring in the locality belonging to the pump, there were sixty-one instances in which I was informed that the deceased persons used to drink the pump-water from Broad Street, either constantly or occasionally. In six instances I could get no information, owing to the death or departure of every one connected with the deceased individuals; and in six cases I was informed that the deceased persons did not drink the pump-water before their illness.[4]

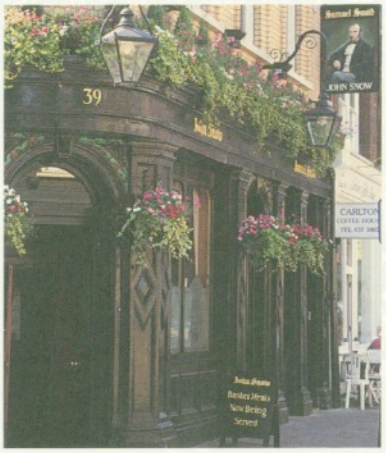
Thus the theory implicating the particular pump was confirmed by the observed covariation: in this area of London, there were few occurrences of cholera exceeding the normal low level, except among those people who drank water from the Broad Street pump. It was now time to act; after all, the reason we seek causal explanations is in order to *intervene*, to govern the cause so as to govern the effect: "Policy-thinking is and must be causality-thinking."[5] Snow described his findings to the authorities responsible for the community water supply, the Board of Guardians of St. James's Parish, on the evening of September 7, 1854. The Board ordered that the pump-handle on the Broad Street well be removed immediately. The epidemic soon ended.

[3] Snow, *Cholera*, p. 39. Writing a few weeks after the epidemic, Snow reported his results in a first-person narrative, more like a laboratory notebook or a personal journal than a modern research paper with its pristine, reconstructed science.

[4] Snow, *Cholera*, pp. 39–40.

[5] Robert A. Dahl, "Cause and Effect in the Study of Politics," in Daniel Lerner, ed., *Cause and Effect* (New York, 1965), p. 88. Wold writes "A frequent situation is that description serves to maintain some *modus vivendi* (the control of an established production process, the tolerance of a limited number of epidemic cases), whereas explanation serves the purpose of *reform* (raising the agricultural yield, reducing the mortality rates, improving a production process). In other words, description is employed as an aid in the human *adjustment* to conditions, while explanation is a vehicle for ascendancy over the environment." Herman Wold, "Causal Inference from Observational Data," *Journal of the Royal Statistical Society*, A, 119 (1956), p. 29.

Moreover, the result of this intervention (a before/after experiment of sorts) was consistent with the idea that cholera was transmitted by impure water. Snow's explanation replaced previously held beliefs that cholera spread through the air or by some other means. In those times many years before the discovery of bacteria, one fantastic theory speculated that cholera vaporously rose out of the burying grounds of plague victims from two centuries earlier.[6] In 1886 the discovery of the bacterium *Vibrio cholerae* confirmed Snow's theory. He is still celebrated for establishing the mode of cholera transmission *and* consequently the method of prevention: keep drinking water, food, and hands clear of infected sewage. Today at the old site of the Broad Street pump there stands a public house (a bar) named after John Snow, where one can presumably drink more safely than 140 years ago.



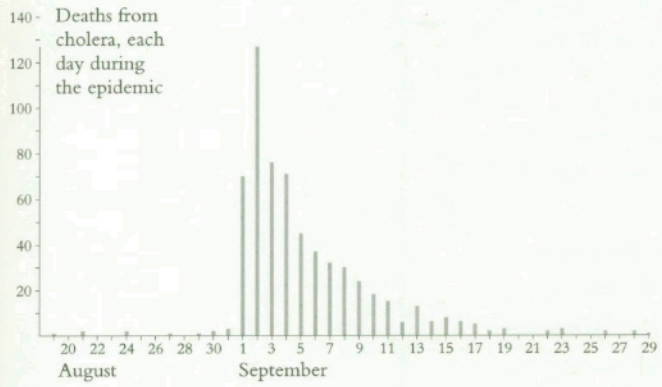
WHY was the centuries-old mystery of cholera finally solved? Most importantly, Snow had a *good idea*—a causal theory about how the disease spread—that guided the gathering and assessment of evidence. This theory developed from medical analysis and empirical observation; by mapping earlier epidemics, Snow detected a link between different water supplies and varying rates of cholera (to the consternation of private water companies who anonymously denounced Snow's work). By the 1854 epidemic, then, the intellectual framework was in place, and the problem of how cholera spread was ripe for solution.[7]

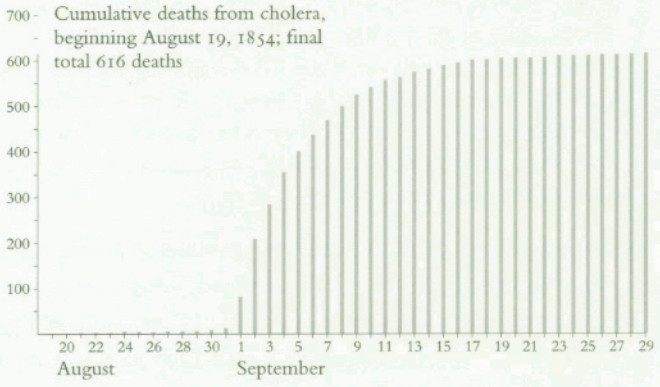
Along with a good idea and a timely problem, there was a *good method*. Snow's scientific detective work exhibits a shrewd intelligence about evidence, a clear logic of data display and analysis:

1. *Placing the data in an appropriate context for assessing cause and effect.* The original data listed the victims' names and described their circumstances, all in order by date of death. Such a stack of death certificates naturally lends itself to time-series displays, chronologies of the epidemic as shown below. *But descriptive narration is not causal explanation;* the passage of time is a poor explanatory variable, practically useless in discovering a strategy of how to intervene and stop the epidemic.

[6] H. Harold Scott, *Some Notable Epidemics* (London, 1934), pp. 3–4.

[7] Scientists are not "admired for failing in the attempt to solve problems that lie beyond [their] competence. . . . If politics is the art of the possible, research is surely the art of the soluble. Both are immensely practical-minded affairs. . . . The art of research [is] the art of making difficult problems soluble by devising means of getting at them. Certainly good scientists study the most important problems they think they can solve. It is, after all, their professional business to solve problems, not merely to grapple with them. The spectacle of a scientist locked in combat with the forces of ignorance is not an inspiring one if, in the outcome, the scientist is routed. That is why so many of the most important biological problems have not yet appeared on the agenda of practical research." Peter Medawar, *Pluto's Republic* (New York, 1984), pp. 253–254; 2–3.





Instead of plotting a time-series, which would simply report each day's bad news, Snow constructed a graphical display that provided direct and powerful testimony about a possible cause-effect relationship. Recasting the original data from their one-dimensional temporal ordering into a two-dimensional spatial comparison, Snow marked deaths from cholera (▮▮▮▮▮) on this map, along with locations of the area's 11 community water pump-wells (◉). The notorious well is located amid an intense cluster of deaths, near the D in BROAD STREET. This map reveals a strong association between cholera and proximity to the Broad Street pump, in a context of simultaneous comparison with other local water sources and the surrounding neighborhoods without cholera.

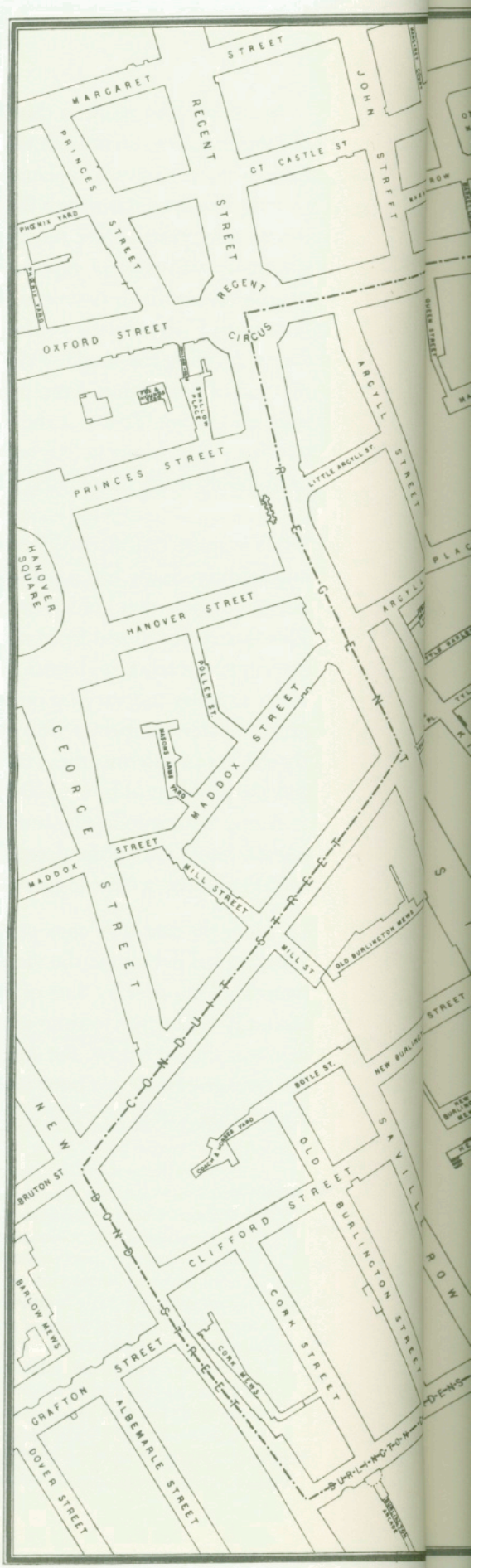
2. *Making quantitative comparisons.* The deep, fundamental question in statistical analysis is *Compared with what?* Therefore, investigating the experiences of the victims of cholera is only part of the search for credible evidence; to understand fully the cause of the epidemic also requires an analysis of those who *escaped* the disease. With great clarity, the map presented several intriguing clues for comparisons between the living and the dead, clues strikingly visible at a brewery and a workhouse (tinted yellow here). Snow wrote in his report:

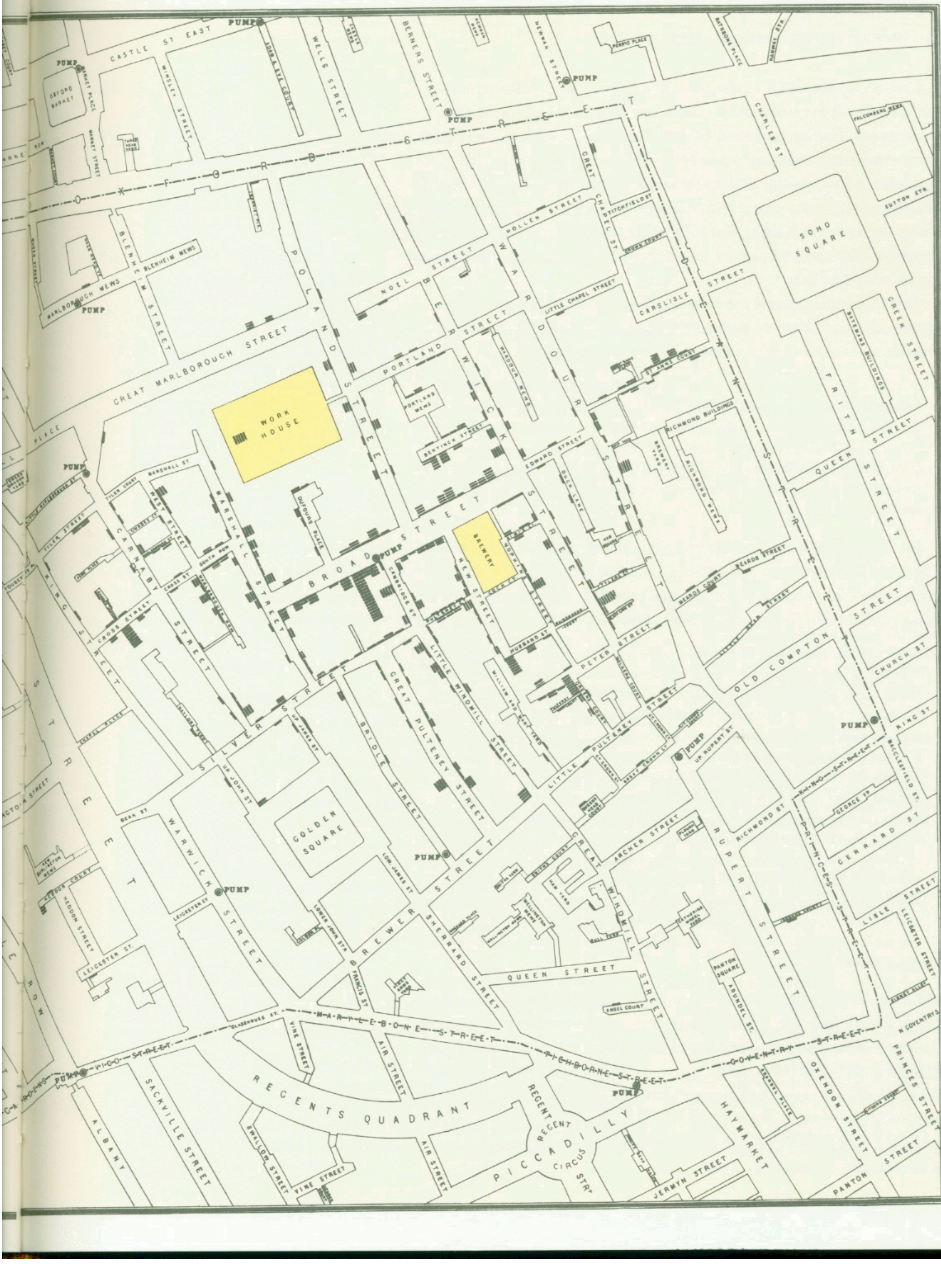
> There is a brewery in Broad Street, near to the pump, and on perceiving that no brewer's men were registered as having died of cholera, I called on Mr. Huggins, the proprietor. He informed me that there were above seventy workmen employed in the brewery, and that none of them had suffered from cholera—at least in severe form—only two having been indisposed, and that not seriously, at the time the disease prevailed. The men are allowed a certain quantity of malt liquor, and Mr. Huggins believes they do not drink water at all; and he is quite certain that the workmen never obtained water from the pump in the street. There is a deep well in the brewery, in addition to the New River water. (p. 42)

Saved by the beer! And at a nearby workhouse, the circumstances of non-victims of the epidemic provided important and credible evidence about the cause of the disease, as well as a quantitative calculation of an expected rate of cholera compared with the actual observed rate:

> The Workhouse in Poland Street is more than three-fourths surrounded by houses in which deaths from cholera occurred, yet out of five-hundred-thirty-five inmates only five died of cholera, the other deaths which took place being those of persons admitted after they were attacked. The workhouse has a pump-well on the premises, in addition to the supply from the Grand Junction Water Works, and the inmates never sent to Broad Street for water. If the mortality in the workhouse had been equal to that in the streets immediately surrounding it on three sides, upwards of one hundred persons would have died. (p. 42)

Such clear, lucid reasoning may seem commonsensical, obvious, insufficiently technical. Yet we will soon see a tragic instance, the decision to launch the space shuttle, when this straightforward logic of statistical (and visual) comparison was abandoned by many engineers, managers, and government officials.




CASTLE ST EAST
PUMP
OXFORD MARKET
MARKET PLACE
WELLS STREET
BERNERS STREET
NEWMAN STREET
PERRYS PLACE
RATHBONE PLACE
HANWAY STR.
O X F O R D S T R E E T
CHARLES ST
FALCONBERG MEWS
SUTTON STR.
SOHO SQUARE
HOLLEN STREET
GREAT CHAPEL ST
STITCHFIELD ST
CROWN COURT
BLENHEIM MEWS
BLENHEIM STREET
MARLBOROUGH MEWS
POLAND STREET
NOEL STREET
BERWICK STREET
WARDOUR STREET
LITTLE CHAPEL STREET
CARSLISLE STREET
DEAN STREET
CREEK STREET
BATEMANS BUILDINGS
FRITH STREET
GREAT MARLBOROUGH STREET
PORTLAND STREET
PORTLAND MEWS
ST ANNS COURT
RICHMOND BUILDINGS
RICHMOND MEWS
WORK HOUSE
BENTINCK STREET
EDWARD STREET
QUEEN STREET
MARSHALL STREET
BROAD STREET
BREWERY
DUFOURS PLACE
CARNABY STREET
KING STREET
CROSS STREET
SOUTH ROW
NEW STREET
CAMBRIDGE ST
HUSBAND ST
PETER STREET
OLD COMPTON STREET
BEARDS COURT
CHURCH ST
KING ST
MACCLESFIELD ST
GERRARD ST
GEORGE YD
PRINCES STREET
LISLE STREET
LEICESTER STREET
SILVER STREET
LITTLE WINDMILL STREET
GREAT PULTENEY STREET
BRIDLE STREET
UP JAMES ST
UP JOHN ST
GOLDEN SQUARE
BEAK ST
WARWICK STREET
REGENT STREET
LITTLE PULTENEY STREET
UP RUPERT ST
RUPERT STREET
RICHMOND ST
ARCHER STREET
GREAT WINDMILL STREET
BREWER STREET
SHERRARD STREET
LOW JAMES ST
LOWER JOHN STR
LEICESTER ST
HEDDON STREET
HEDDON COURT
QUEEN STREET
PANTON SQUARE
ARUNDEL ST
ANGEL COURT
MARYLEBONE STREET
GLASSHOUSE ST
VINE STREET
AIR STREET
FRANCIS ST
TICHBORNE STREET
COVENTRY STREET
N COVENTRY ST
PRINCES STREET
OXENDON STREET
HAYMARKET
VIGO STREET
REGENTS QUADRANT
REGENT CIRCUS
PICCADILLY
ALBANY
SACKVILLE STREET
SWALLOW STREET
JERMYN STREET
PANTON STREET

3. *Considering alternative explanations and contrary cases.* Sometimes it can be difficult for researchers—who both report *and* advocate their findings—to face up to threats to their conclusions, such as alternative explanations and contrary cases. Nonetheless, the credibility of a report is enhanced by a careful assessment of *all* relevant evidence, not just the evidence overtly consistent with explanations advanced by the report. The point is to get it right, not to win the case, not to sweep under the rug all the assorted puzzles and inconsistencies that frequently occur in collections of data.[8]

Both Snow's map and the time-sequence of deaths show several apparently contradictory instances, a number of deaths from cholera with no obvious link to the Broad Street pump. And yet . . .

> In some of the instances, where the deaths are scattered a little further from the rest on the map, the malady was probably contracted at a nearer point to the pump. A cabinet-maker who resided on Noel Street [some distance from Broad Street] worked in Broad Street. . . . A little girl, who died in Ham Yard, and another who died in Angel Court, Great Windmill Street, went to the school in Dufour's Place, Broad Street, and were in the habit of drinking the pump-water. . . .[9]

In a particularly unfortunate episode, one London resident made a special effort to obtain Broad Street well-water, a delicacy of taste with a side-effect that unwittingly cost two lives. Snow's report is one of careful description and precise logic:

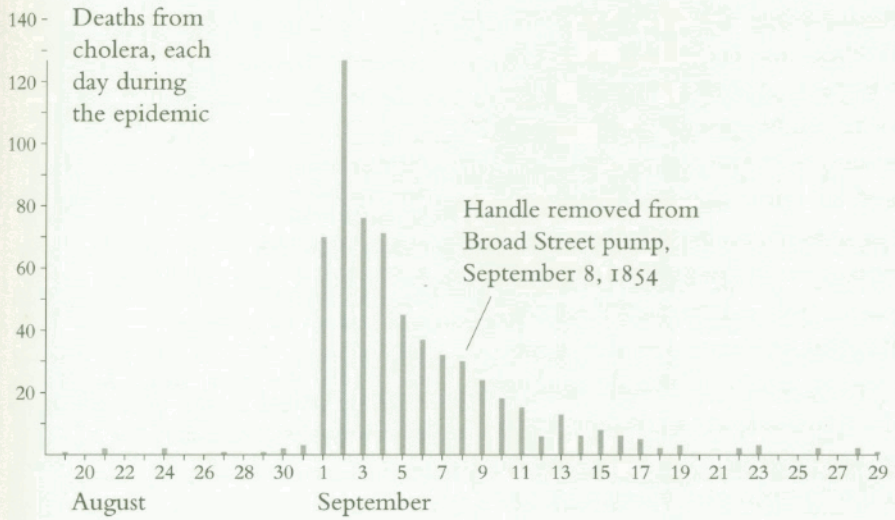
> Dr. Fraser also first called my attention to the following circumstances, which are perhaps the most conclusive of all in proving the connexion between the Broad Street pump and the outbreak of cholera. In the 'Weekly Return of Births and Deaths' of September 9th, the following death is recorded: 'At West End, on 2nd September, the widow of a percussion-cap maker, aged 59 years, diarrhea two hours, *cholera epidemica* sixteen hours.' I was informed by this lady's son that she had not been in the neighbourhood of Broad Street for many months. A cart went from Broad Street to West End every day, and it was the custom to take out a large bottle of the water from the pump in Broad Street, as she preferred it. The water was taken on Thursday, 31st August, and she drank of it in the evening, and also on Friday. She was seized with cholera on the evening of the latter day, and died on Saturday. . . . A niece, who was on a visit to this lady, also drank of the water; she returned to her residence, in a high and healthy part of Islington, was attacked with cholera, and died also. There was no cholera at the time, either at West End or in the neighbourhood where the niece died.[10]

Although at first glance these deaths appear unrelated to the Broad Street pump, they are, upon examination, strong evidence pointing to that well. There is here a clarity and undeniability to the link between cholera and the Broad Street pump; only such a link can account for what would otherwise be a mystery, this seemingly random and unusual occurrence of cholera. And the saintly Snow, unlike some researchers, gives full credit to the person, Dr. Fraser, who actually found this crucial case.

[8] The distinction between science and advocacy is poignantly posed when statisticians serve as consultants and witnesses for lawyers. See Paul Meier, "Damned Liars and Expert Witnesses," and Franklin M. Fisher, "Statisticians, Econometricians, and Adversary Proceedings," *Journal of the American Statistical Association*, 81 (1986), pp. 269–276 and 277–286.

[9] Snow, *Cholera*, p. 47.

[10] Snow, *Cholera*, pp. 44–45.



Data source: plotted from the table in Snow, *Cholera*, p. 49.

Ironically, the most famous aspect of Snow's work is also the most uncertain part of his evidence: it is not at all clear that the removal of the handle of the Broad Street pump had much to do with ending the epidemic. As shown by this time-series above, the epidemic was already in rapid decline by the time the handle was removed. Yet, in many retellings of the story of the epidemic, the pump-handle removal is *the* decisive event, the unmistakable symbol of Snow's contribution. Here is the dramatic account of Benjamin Ward Richardson:

> On the evening of Thursday, September 7th, the vestrymen of St. James's were sitting in solemn consultation on the causes of the [cholera epidemic]. They might well be solemn, for such a panic possibly never existed in London since the days of the great plague. People fled from their homes as from instant death, leaving behind them, in their haste, all the mere matter which before they valued most. While, then, the vestrymen were in solemn deliberation, they were called to consider a new suggestion. A stranger had asked, in modest speech, for a brief hearing. Dr. Snow, the stranger in question, was admitted and in few words explained his view of the 'head and front of the offending.' He had fixed his attention on the Broad Street pump as the source and centre of the calamity. He advised removal of the pump-handle as the grand prescription. The vestry was incredulous, but had the good sense to carry out the advice. The pump-handle was removed, and the plague was stayed.[11]

Note the final sentence, a declaration of cause and effect.[12] Modern epidemiologists, however, are distinctly skeptical about the evidence that links this intervention to the epidemic's end:

> John Snow, in the seminal act of modern public health epidemiology, performed an intervention that was non-randomized, that was appraised with historical controls, and that had major ambiguities in the equivocal time relationship between his removal of the handle of the Broad Street pump and the end of the associated epidemic of cholera—but he correctly demonstrated that the disease was transmitted through water, not air.[13]

[11] Benjamin W. Richardson, "The Life of Jonn Snow, M.D.," foreword to John Snow, *On Chloroform and Other Anaesthetics: Their Action and Administration* (London, 1858), pp. xx–xxi.

[12] Another example of the causal claim: "On September 8, at Snow's urgent request, the handle of the Broad Street pump was removed and the incidence of new cases ceased almost at once," E. W. Gilbert, "Pioneer Maps of Health and Disease in England," *The Geographical Journal*, 124 (1958), p. 174. Gilbert's assertion was repeated in Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut, 1983), p. 24.

[13] Alvan R. Feinstein, *Clinical Epidemiology: The Architecture of Clinical Research* (Philadelphia, 1985), pp. 409–410. And A. Bradford Hill ["Snow—An Appreciation," *Proceedings of the Royal Society of Medicine*, 48 (1955), p. 1010] writes: "Though conceivably there might have been a second peak in the curve, and though almost certainly some more deaths would have occurred if the pump handle had remained in situ, it is clear that the end of the epidemic was not dramatically determined by its removal."

At a minimum, removing the pump-handle prevented a recurrence of cholera. Snow recognized several difficulties in evaluating the effect of his intervention; since most people living in central London had fled, the disease ran out of possible victims—which happened simultaneously with shutting down the infected water supply.[14] The case against the Broad Street pump, however, was based on a diversity of additional evidence: the cholera map, studies of unusual instances, comparisons of the living and dead with their consumption of well-water, and an idea about a mechanism of contamination (a nearby underground sewer had probably leaked into the infected well). Also, the finding that cholera was carried by water—a life-saving scientific discovery that showed how to intervene and prevent the spread of cholera—derived not only from study of the Broad Street epidemic but also from Snow's mappings of several other cholera outbreaks in relation to the purity of community water supplies.

[14] "There is no doubt that the mortality was much diminished, as I said before, by the flight of the population, which commenced soon after the outbreak; but the attacks had so far diminished before the use of the water was stopped, that it is impossible to decide whether the well still contained the cholera poison in an active state, or whether, from some cause, the water had become free from it." Snow, *Cholera*, pp. 51–52.

4. *Assessment of possible errors in the numbers reported in graphics.* Snow's analysis attends to the sources and consequences of errors in gathering the data. In particular, the credibility of the cholera map grows out of supplemental details in the text—as image, word, and number combine to present the evidence and make the argument. Detailed comments on possible errors annotate both the map and the table, reassuring readers about the care and integrity of the statistical detective work that produced the data graphics:

> The deaths which occurred during this fatal outbreak of cholera are indicated in the accompanying map, as far as I could ascertain them. There are necessarily some deficiencies, for in a few of the instances of persons who died in the hospitals after their removal from the neighbourhood of Broad Street, the number of the house from which they had been removed was not registered. The address of those who died after their removal to St. James's Workhouse was not registered; and I was only able to obtain it, in a part of the cases, on application at the Master's Office, for many of the persons were too ill, when admitted, to give any account of themselves. In the case also of some of the workpeople and others who contracted the cholera in this neighbourhood, and died in different parts of London, the precise house from which they had removed is not stated in the return of deaths. I have heard of some persons who died in the country shortly after removing from the neighbourhood of Broad Street; and there must, no doubt, be several cases of this kind that I have not heard of. Indeed, the full extent of the calamity will probably never be known. The deficiencies I have mentioned, however, probably do not detract from the correctness of the map as a diagram of the topography of the outbreak; for, if the locality of the few additional cases could be ascertained, they would probably be distributed over the district of the outbreak in the same proportion as the large number which are known.[15]

> The deaths in the above table [the time-series of daily deaths] are compiled from the sources mentioned above in describing the map; but some deaths which were omitted from the map on account of the number of the house not being known, are included in the table. . . .[16]

[15] Snow, *Cholera*, pp. 45–46.

[16] Snow, *Cholera*, p. 50.

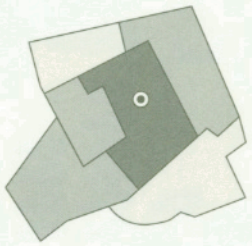
Snow drew a *dot map*, marking each individual death. This design has statistical costs and benefits: death *rates* are not shown, and such maps may become cluttered with excessive detail; on the other hand, the sometimes deceptive effects of aggregation are avoided. And of course dot maps aid in the identification and analysis of individual cases, evidence essential to Snow's argument.

The big problem is that dot maps fail to take into account the number of people living in an area and at risk to get a disease: "an area of the map may be free of cases merely because it is not populated."[17] Snow's map does not fully answer the question *Compared with what?* For example, if the population as a whole in central London had been distributed just as the deaths were, then the cholera map would have merely repeated the unimportant fact that more people lived near the Broad Street pump than elsewhere. This was not the case; the entire area shown on the map—with and without cholera—was thickly populated. Still, Snow's dot map does not assess varying densities of population in the area around the pump. Ideally, the cholera data should be displayed on both a dot and a rate map, with population-based rates calculated for rather small and homogeneous geographic units. In the text of his report, however, Snow did present rates for a few different areas surrounding the pump.

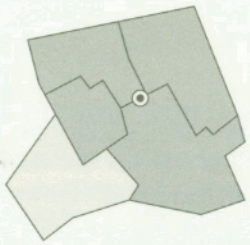
Aggregations by area can sometimes mask and even distort the true story of the data. For two of the three examples at right, constructed by Mark Monmonier from Snow's individual-level data, the intense cluster around the Broad Street pump entirely vanishes in the process of geographically aggregating the data (the greater the number of cholera deaths, the darker the area).[18]

In describing the discovery of how cholera is transmitted, various histories of medicine discuss the famous map and Snow's analysis. The cholera map, as Snow drew it, is difficult to reproduce on a single page; the full size of the original is awkward (a square, 40 cm or 16 inches on the side), and if reduced in size, the cholera symbols become murky and the type too small. Some facsimile editions of *On the Mode of Communication of Cholera* have given up, reprinting only Snow's text and not the crucial visual evidence of the map. Redrawings of the map for textbooks in medicine and in geography fail to reproduce key elements of Snow's original. The workhouse and brewery, those essential compared-with-what cases, are left unlabeled and unidentified, showing up only as mysterious cholera-free zones close to the infected well. Standards of quality may slip when it comes to visual displays; imprecise and undocumented work that would be unacceptable for words or tables of data too often shows up in graphics. Since it is *all* evidence—regardless of the method of presentation—the highest standards of statistical integrity and statistical thinking should apply to *every* data representation, including visual displays.

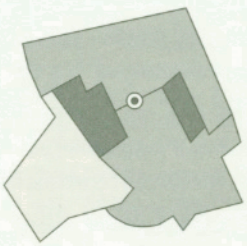
[17] Brian MacMahon and Thomas F. Pugh, *Epidemiology: Principles and Methods* (Boston, 1970), p. 150.



In this aggregation of individual deaths into six areas, the greatest number is concentrated at the Broad Street pump.

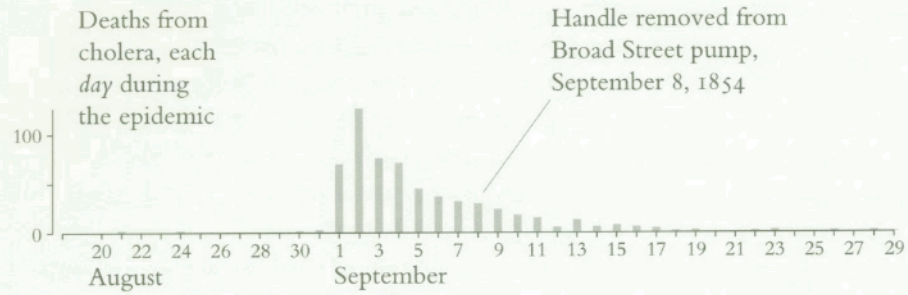


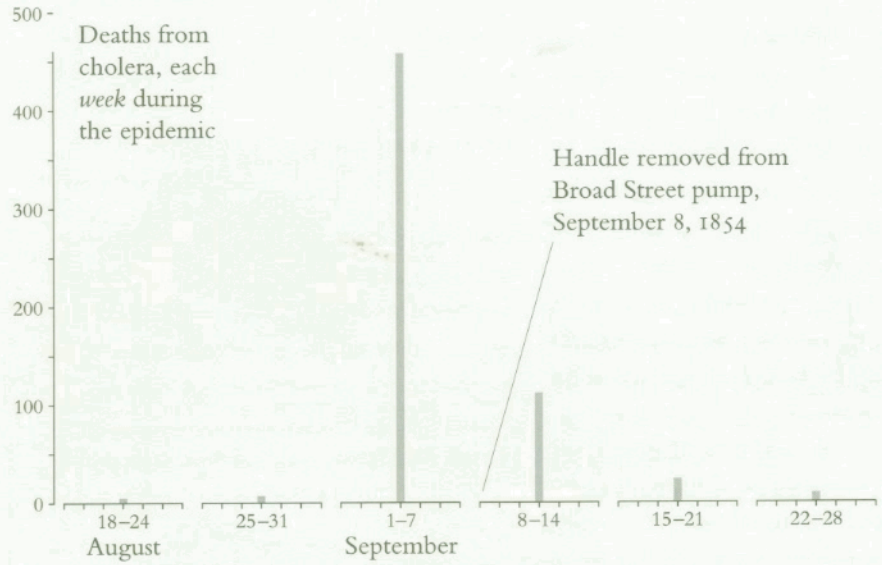
Using different geographic subdivisions, the cholera numbers are nearly the same in four of the five areas.



In this aggregation of the deaths, the two areas with the most deaths do not even include the infected pump!

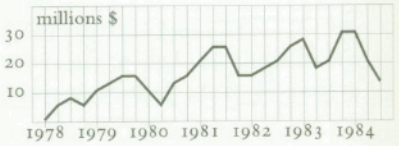
[18] Mark Monmonier, *How to Lie with Maps* (Chicago, 1991), pp. 142–143.



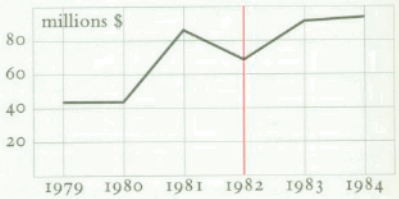


Aggregations over time may also mask relevant detail and generate misleading signals, similar to the problems of spatial aggregation in the three cholera maps. Shown at top is the familiar *daily* time-series of deaths from cholera, with its smooth decline in deaths unchanged by the removal of the pump-handle. When the daily data are added up into *weekly* intervals, however, a different picture emerges: the removal had the apparent consequence of reducing the weekly death toll from 458 to 112! But this result comes purely from the aggregation, for the daily data show no such effect.[19] Conveniently, the handle was removed in early morning of September 8; hence the plausible weekly intervals of September 1–7, 8–14, and so on. Imagine if we had read the story of John Snow as reported in the first few pages here, and if our account showed the weekly instead of daily deaths—then it would all appear perfectly convincing although quite misleading.

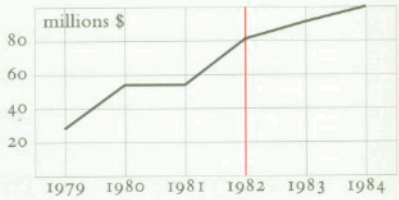
Some other weekly intervals would further aggravate the distortion. Since two or more days typically pass between consumption of the infected water and deaths from cholera, the removal date might properly be *lagged* in relation to the deaths (for example, by starting to count post-removal deaths on the 10th of September, 2 days *after* the pump



Above, this chart shows *quarterly* revenue data in a financial graphic for a legal case. Several dips in revenue are visible.

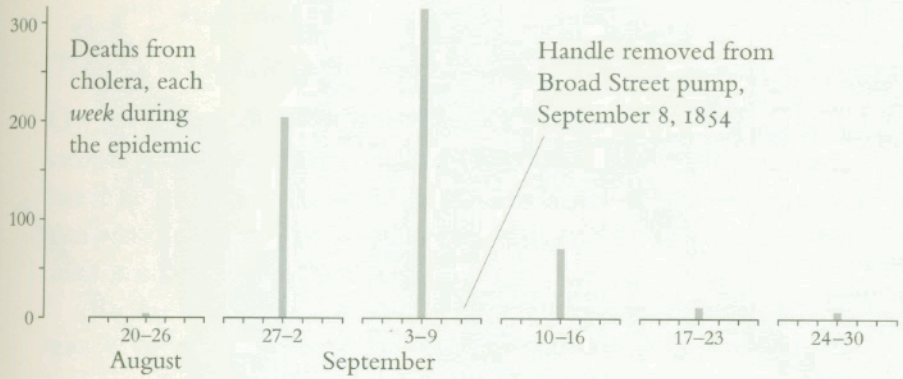


Aggregating the quarterly data into years, this chart above shows revenue by *fiscal year* (beginning July 1, ending June 30). Note the dip in 1982, the basis of a claim for damages.



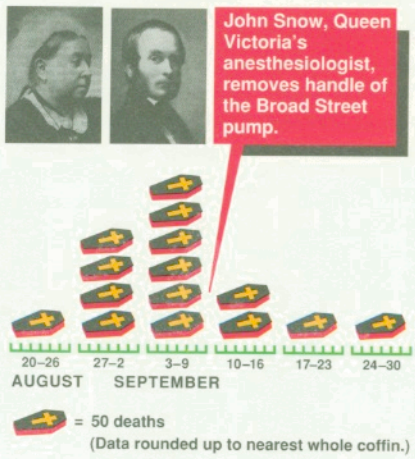
Shown above are the same quarterly revenue data added up into *calendar years*. The 1982 dip has vanished.

[19] Reading from the top, these clever examples reveal the effects of temporal aggregation in economic data; from Gregory Joseph, *Modern Visual Evidence* (New York, 1992), pp. A42–A43.



handle was taken off). These lagged weekly clusters are shown above. The pseudo-effect of handle removal is now even stronger: after three weeks of increasing deaths, the weekly toll plummets when the handle is gone. A change of merely two days in weekly intervals has radically shifted the shape of the data representation. As a comparison between the two weekly charts shows, the results depend on the arbitrary choice of time periods—a sign that we are seeing method not reality.

These conjectural weekly aggregations are as condensed as news reports; missing are only the decorative clichés of "info-graphics" (the language is as ghastly as the charts). At right is how pop journalism might depict Snow's work, complete with celebrity factoids, over-compressed data, and the isotype styling of those little coffins.



Time-series are exquisitely sensitive to choice of intervals and end points. Nonetheless, many aggregations are perfectly sensible, reducing the tedious redundancy and uninteresting complexity of large data files; for example, the *daily* data amalgamate times of death originally recorded to the hour and even minute. If in doubt, graph the detailed underlying data to assess the effects of aggregation.

A further difficulty arises, a result of fast computing. It is easy now to sort through thousands of plausible varieties of graphical and statistical aggregations—and then to select for publication only those findings strongly favorable to the point of view being advocated. Such searches are described as *data mining, multiplicity*, or *specification searching*.[20] Thus a prudent judge of evidence might well presume that those *graphs, tables, and calculations revealed in a presentation are the best of all possible results chosen expressly for advancing the advocate's case.*

EVEN in the face of issues raised by a modern statistical critique, it remains wonderfully true that John Snow did, after all, show exactly how cholera was transmitted and therefore prevented. In 1955, the *Proceedings of the Royal Society of Medicine* commemorated Snow's discovery. A renowned epidemiologist, Bradford Hill, wrote: "For close upon 100 years we have been free in this country from epidemic cholera, and it is a freedom which, basically, we owe to the logical thinking, acute observations and simple sums of Dr. John Snow."[21]

[20] John W. Tukey, "Some Thoughts on Clinical Trials, Especially Problems of Multiplicity," *Science*, 198 (1977), pp. 679–684; Edward E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (New York, 1978). On the other hand, "enough exploration must be done so that the results are shown to be relatively insensitive to plausible alternative specifications and data choices. Only in that way can the statistician protect himself or herself from the temptation to favor the client and from the ensuing cross-examination." Franklin M. Fisher, "Statisticians, Econometricians, and Adversary Proceedings," *Journal of the American Statistical Association*, 81 (1986), p. 279. Another reason to explore the data thoroughly is to find out what is going on! See John W. Tukey, *Exploratory Data Analysis* (Reading, Massachusetts, 1977).

[21] A. Bradford Hill, "Snow—An Appreciation," *Proceedings of the Royal Society of Medicine*, 48 (1955), p. 1012.