Public-Access Provocations: An Informal Column

A funny thing happened in mid-January 1994. I was updating the weekly usage graph for Eureka, a manual operation (in Quattro Pro) based on a sampled weekly statistical summary. After a couple of weeks in which usage was growing back from low holiday levels, suddenly usage was about half of the preceding week. (No, this isn't another "Eureka column." Bear with me.) What happened? Where did all the users go? We were expecting to see continued growth as more libraries implement Eureka. Even though the weekly graph is based on partial sampling and is an informal measure, the sharp decline was unexpected and startling. After some discussion, we concluded that the horrendous weather in the eastern United States could be responsible. As it turned out, that wasn't the cause. Instead, an unexpected data condition caused the data analysis routines to misbehave. Closer examination showed that there was perhaps a 10% dip in usage, almost certainly because of weather, followed by new record highs in each of the next two weeks. As my message to those looking at weekly figures noted: Remember that drop in usage last week? Well, there's a new explanation: It didn't happen. What did happen is a little embarrassing: after 26 years making my living using computers, I actually believed something that didn't make sense, because it emerged from a computer. I should have known better.

If It Doesn't Make Sense, It's Probably Wrong If there's one rule every experienced computer user should know, it's this one. When "the computer" says something that violates your expectations, your first assumption should be that "the computer" is wrong. Check the raw data, check intermediate calculations, check the algorithms. Chances are, something went wrong along the way. Did the computer actually make a mistake? Probably not. Computers rarely suffer internal processing failures that they don't catch. Well-written programs rarely fail to calculate properly. But calculations are no better than the algorithms used to code them, and algorithms are no better than the designs used to prepare them. More to the point, "GIGO" is as true now as ever: if the raw data has been corrupted, the output is useless. Is That Calculation Really Calculated? Spreadsheets and other similar programs may represent the worst case. With most spreadsheet software, there's nothing at all to prevent a user from keying a number into a slot that should be a calculation, thus disrupting not only that particular cell but any other cells that depend on it. Any spreadsheet should be regarded with some suspicion, particularly if any of the calculated figures appear extraordinary: maybe they're simply wrong. Those of us who have been programmers should know this, of course, but there's a powerful temptation to assume that

computers never lie. Putting the most nonsensical assumptions and erroneous data into nicely-formatted spreadsheet form gives it validity in many eyes, even though the data may be flawed. Better yet: make a chart out of it--and if the chart isn't impressive enough, use a non-zero baseline. But then, we all know better than to fall for misleading graphics and statistics, don't we? How about this statistic? The use of Zyzix, the hot new Internet tool, has increased 19,000% over the past six months. Which could mean that six months ago two people used it and this month 380 people used it: that's a 19,000% growth rate.

+ Page 23 +

Heuristics and Skepticism To use statistics and computer-generated numbers well, you must be able to do mental approximations. You must have the heuristics handy to see whether the computer's output is reasonable. Most of the time, of course, the output will be perfectly sensible--but you should always be ready to look twice at something that's sharply out of line. What does this have to do with public access? Quite a bit. If you're looking at access versus collections, you need to look closely at the economic arguments, and look at them in totality. When you find usage of a new system has jumped by an order of magnitude (i.e., 1000%) over the past year, be aware that such a jump probably will not be repeated: percentages without numbers are essentially meaningless. Not only do you need to be skeptical when looking at analyses and projections, you need to find ways to encourage patrons to be skeptical. More on that in the next two Public-Access Provocations columns. About the Author Walt Crawford, Senior Analyst, The Research Libraries Group, Inc., 1200 Villa Street, Mountain View, CA 94041-1100. Internet: br.wcc@rlg.stanford.edu.