Early in January in a Chicago hotel, Campbell Harvey gave a rip-snorting presidential address to the American Finance Association, the world’s leading society for research on financial economics. To get published in journals, he said, there’s a powerful temptation to torture the data until it confesses—that is, to conduct round after round of tests in search of a finding that can be claimed to be statistically significant. Said Harvey, a professor at Duke University’s Fuqua School of Business: “Unfortunately, our standard testing methods are often ill-equipped to answer the questions that we pose.” He exhorted the group: “We are not salespeople. We are scientists!”

The problems Harvey identified in academia are as bad or worse in the investing world. Mass-market products such as exchange-traded funds are being concocted using the same flawed statistical techniques you find in scholarly journals. Most of the empirical research in finance is likely false, Harvey wrote in a paper with a Duke colleague, Yan Liu, in 2014. “This implies that half the financial products (promising outperformance) that companies are selling to clients are false.”
Most of us have a vague sense that we’re being ripped off by investment firms that charge hefty fees while producing results that are no better than you’d get throwing darts at a page of stock listings. It’s troubling nonetheless to find out we’re correct. And it’s important to understand the mechanics of what has gone wrong.

The core of the problem is that it’s hard to beat the market, but people keep trying anyway. An abundance of computing power makes it possible to test thousands, even millions, of trading strategies. The standard method is to see how the strategy would have done if it had been used during the ups and downs of the market over, say, the past 20 years. This is called backtesting. As a quality check, the technique is then tested on a separate set of “out-of-sample” data—i.e., market history that wasn’t used to create the technique.

In the wrong hands, though, backtesting can go horribly wrong. It once found that the best predictor of the S&P 500, out of all the series in a batch of United Nations data, was butter production in Bangladesh. The nerd webcomic *xkcd* by Randall Munroe captures the ethos perfectly: It features a woman claiming jelly beans cause acne. When a statistical test shows no evidence of an effect, she revises her claim—it must depend on the flavor of jelly bean. So the statistician tests 20 flavors. Nineteen show nothing. By chance there’s a high correlation between jelly bean consumption and acne breakouts for one flavor. The final panel of the cartoon is the front page of a newspaper: “Green Jelly Beans Linked to Acne! 95% Confidence. Only 5% Chance of Coincidence!”

It’s worse for financial data because researchers have more knobs to twist in search of a prized “anomaly”—a subtle pattern in the data that looks like it could be a moneymaker. They can vary the period, the set of securities under consideration, or even the statistical method. Negative findings go in a file drawer; positive ones get submitted to a journal (tenure!) or made into an ETF whose performance we rely on for retirement. Testing out-of-sample data to keep yourself honest helps, but it doesn’t cure the problem. With enough tests, eventually by chance even your safety check will show the effect you want.

Harvey’s term for torturing the data until it confesses is “p-hacking,” a reference to the p-value, a measure of statistical significance. P-hacking is also known as overfitting, data-mining—or data-snooping, the coinage of Andrew Lo, director of MIT’s Laboratory of Financial Engineering. Says Lo: “The more you search over the past, the more likely it is you are going to find exotic patterns that you happen to like or focus on. Those patterns are least likely to repeat.”

Such tricks weren’t necessary when Wall Streeters could make a good living charging for their stock-selection skills. That gig became scarcer when it became clear that few could consistently beat a low-cost index fund that tracks, say, the S&P 500.

Index funds are cheap because their sponsors don’t need to hire expensive stockpickers, but they aren’t perfect. Stocks in the index are weighted by market value, so Apple is 3.7 percent of the S&P 500 while Rupert Murdoch’s News Corp. is just 0.008 percent. When a stock gets hot for whatever reason, the index fund has to buy even more of it, which may not be the wisest choice.

Wall Street’s answer is today’s most stylish investment style: smart beta. At the end of February, more than $500 billion [https://www.bloomberg.com/news/videos/2017-01-17/smart-beta-ets-hits-500-billion](https://www.bloomberg.com/news/videos/2017-01-17/smart-beta-ets-hits-500-billion) was invested in equity exchange-traded funds in the U.S. that use smart beta strategies [https://www.bloomberg.com/news/articles/2017-02-07/peak-quants-threaten-bet-on-smart-beta?source=shr-cption-rss](https://www.bloomberg.com/news/articles/2017-02-07/peak-quants-threaten-bet-on-smart-beta?source=shr-cption-rss), according to data compiled by Bloomberg. “Beta” is lingo for the return on investment you get from owning a slice of the entire stock market, as in a conventional index fund; the “smart” part refers to breaking the link with market value. Stocks in a smart beta index may be weighted by anything from company sales or book value to special-sauce ingredients such as “quality” (on the theory that well-managed companies tend to outperform in the stock market).

Trouble is, fund managers have gotten too creative in the competition for investors’ dollars. To quote the burlesque strippers in *Gypsy*, you gotta get a gimmick: There are ETFs for more than a thousand new indexes. Creativity, alas, doesn’t equal success. Vanguard, the big
investment manager, calculated in 2012 that ETFs did great on their backtests, outperforming the market by 10 percentage points a year on average in the five years before they went live, but then underperformed the market by 1 percentage point a year in the five years afterward. The most complex strategies suffer the biggest drop-off from their backtests, according to an article in the Journal of Portfolio Management.

Fights have broken out over who is or isn’t coming up with spurious investment concepts. AQR Capital Management of Greenwich, Conn., which started as a quant hedge fund, has grown rapidly by managing other people’s money in smart beta funds. It focuses on “factors” such as quality and momentum that it says lead to reliable outperformance. AQR’s founder and chief investment officer, Clifford Asness, is a billionaire. Rob Arnott, a rival who’s CEO and founder of Research Affiliates in Newport Beach, Calif., says, “I think Cliff has done some outstanding work over the years,” but adds that he’s “insufficiently skeptical about the pervasiveness of data-mining and its impact even in the factors he uses.”

Asness responds by email that the two firms “largely believe in a very similar set of factors like value, low risk, and momentum, to which we think we’ve both applied a lot of a priori skepticism.” He adds that “there is little evidence” so far that the quality-based factor in AQR’s research is performing differently now from how it did in backtesting. But that’s not much of a claim, says Robert Novy-Marx, a professor at the University of Rochester’s Simon Business School who once consulted for AQR and now consults for another group, Dimensional Fund Advisors. “Even if it performed really poorly, you wouldn’t know. There’s just not enough out-of-sample time to make any claim one way or another.”

The old adage applies: If asset managers and finance professors are super-smart, why ain’t they super-rich? The big money is being made by firms that ignore finance theory. Renaissance Technologies on Long Island is dripping with mathematicians and physicists but will not hire a finance Ph.D. Two Sigma Investments is run by computer scientists and mathematicians. D.E. Shaw was founded by a computational biologist. And so on. Reflecting mathematicians’ disdain for sloppiness in finance, a 2014 essay in the Journal of the American Mathematical Society referred to backtest overfitting as “pseudo-mathematics and financial charlatanism.”

Harvey, who’s since completed his term as president of the American Finance Association, has written that finance lags other fields, including genetics, in making certain that its findings are statistically valid. “Many in our profession, including me,” have subjected data to inadequate tests in the past, he said in Chicago. That speech disgruntled some people, he says now, but that’s OK. “To push the field further, you sometimes have to be willing to be very unpopular.”

—With Salil Kishan and Dani Burger

(Updates Clifford Asness’s statement in 13th paragraph. Corrects quality-based fund to quality-based factor in 13th paragraph.)