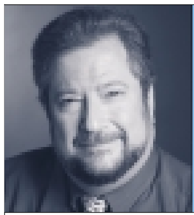
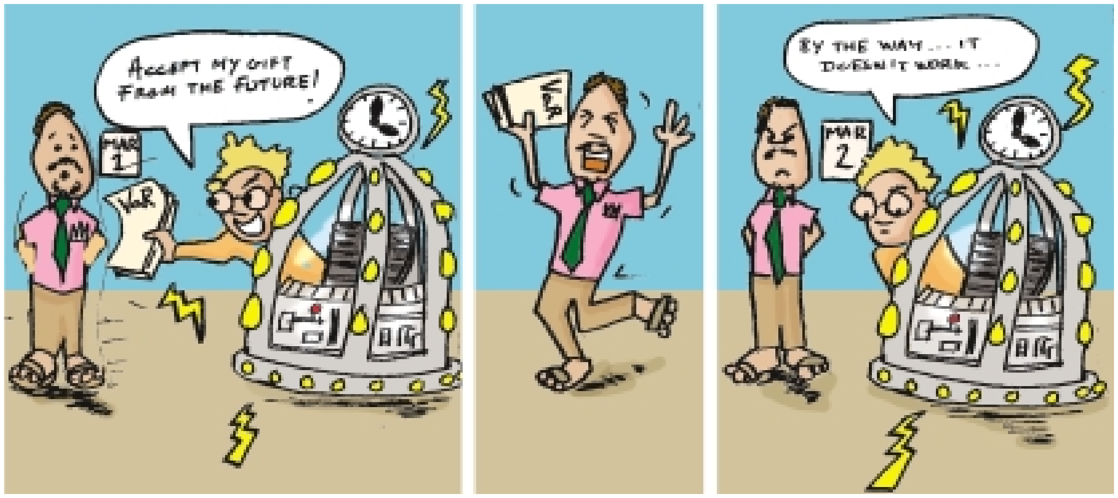


Aaron Brown

Close, But No See GHAR

A look into the black hole between Pre-VaR and Post-VaR worlds



A long time ago in a galaxy VaR, VaR away...

The history of quantitative finance is being written at about twice the speed we lived it, at least judging by questions I get from researchers. They're now up to the late 1980s. I got a call about a month ago from someone wanting to know if I had discussed VaR at a seminar I gave at Yale on November 19, 1987.

The answer was easy: it was much too early for VaR. Like a number of others, I was fooling around with proto-VaR ideas, but I recognize no claim to a modern VaR within four years of that seminar. Till Guldimann and his team made an extensive presentation on a fully developed VaR, using that name, in the JPMorgan 1993 Fixed Income conference, and the next year JPMorgan gave it away free to the world, with data. People I respect claim credit for having it before Guldimann or Morgan, in 1991 or 1992, and some of those claimants were around for the entire process, beginning October 21, 1987. But I don't take seriously any claim to have had all the pieces put together right before 1991.

At the other end, I don't recognize any VaR contribution before October 19, 1987. People knew all about percentiles long before that date, and had all kinds of thoughts about risk measurement, some of which are related conceptually or mathematically to aspects of VaR. But VaR is a very specific set of ideas that was worked out *and implemented* at a specific place and time, for specific reasons. It's not an abstraction, it's a way of managing risk.

Where I was

With some prodding, I was induced to look around for my notes for that seminar. At the time I gave it, I was head of mortgage securities. I had presented my dissertation at the University of Chicago for a finance Ph.D. in 1983, *Resampling Financial Confidence Intervals*, but still had not completed the additional work my committee had imposed, much of which I felt was researching and summarizing unrelated prior work. I felt that the Crash of 1987 validated my idea and distinguished it from prior work, which might incline my committee to accept an additional chapter incorporating the Crash in lieu of some of the earlier requests. The trip to Yale was to try out the new material on an academic audience.

It was the Thursday before the Harvard–Yale game (Harvard won the game, 14–10, and thereby the Ivy Championship), in New Haven that year, so a lot of people were in town. I was particularly anxious to have Jon Ingersoll attend, as I had asked him to substitute for another committee member who had left the country and was hard to reach. He had moved from Chicago to Yale the previous year. It was minus 10 degrees Celsius, which is below the official lowest November temperature in New Haven, but I was there and will swear that the record book is wrong.

I found the transparencies for the talk, but nothing else. For those under 40, transparencies were a primitive communication technology. I printed my talk using a daisy-wheel printer (don't ask), then Xeroxed it onto clear paper. The seminar room had an overhead projector that shone a bright light through the clear paper, onto a mirror, and thence onto a screen (or would have, if we had had a screen; in this case we used a more or less white wall because the screen was broken or missing).

Much to my surprise, there was "VAR" staring me right in the face, alongside the familiar drawing of a bell curve with a line marking a small area of the left tail! Maybe I had invented it five years before Guldimann, and had just forgotten about it. Unfortunately for that theory, there was also SHAR, GHAR, and CAR, not to mention FERL and other acronyms. The "V" in "VAR" was for "Value," all right, but the “AR” was “Analysis Region,” not “at-Risk.” A closer look at the diagram showed that “VAR” referred to the region to the right of the line, not the line itself. The “AR” in “CAR” was for “at-risk” but the “C” was “capital,” not “value.”

Still, this was closer than I remembered. Unfortunately, the pages had fused together over the years. When separated, half the ink transferred to the sheet in front. That made it tedious to reconstruct the talk, but rewarding in the end. The title was “Kelly Betting with Potential Disaster.” The main problem I considered was how a Kelly bettor should modify her strategy if large bets carried a small probability of highly unfavorable outcomes.

Inconstant object of inconstant cause

I began by relating the problem to my dissertation, which concerned statistical analysis of sudden periods of intense volatility and unusual behavior in the markets. These periods seem to happen about every four years on average, although most do not get extensive attention. They are unexpected and affect many seemingly unrelated markets at once. They are bad for almost everyone, partly because they usually mean declines in the nonzero-sum securities (equities and commodities), partly because there are almost no good long-term strategies that are long volatility, but mostly because trading conditions force people to unwind positions at unfavorable times. A price might soar, wiping out the shorts, then plunge, wiping out everyone else, and end up close to where it started. Only those who had the courage and capital to take risks in the aftermath made a lot of money.

In the early 1980s, a number of people were working on the general problem of analyzing time series with nonconstant volatility. In San Diego, Robert Engel took a classical time series approach, which developed into ARCH and earned him a Nobel Prize. I was working for Craig Ainsley at Chicago; we were racing to come up with a Bayesian solution first (we finished later, and generated much less interest). In addition, on my own, I was looking for a data analysis version, inspired by teachers Frederick Mosteller, John Tukey, and Bradley Efron.

I felt that both Engel and Ainsley were too parametric – that is, they made too many assumptions for mathematical tractability and elegance. Moreover, they were implicitly assuming that you were equally concerned with predictions during volatile and nonvolatile periods. I felt that, in finance, the outcomes during volatile periods were unpredictable, except that they were probably big and bad, because you couldn’t count on trading, or on normal price relationships persisting. If you used data from the volatile periods, they dominated your parameter estimates and discouraged you from taking any risk at all. But if you ignored those periods, you went broke. Finally, I felt that both the classical and Bayesian approaches observed the phenomenon rather than controlled it – that is, they were asking what a security price might do tomorrow, rather than what would be the P&L of a dynamic trading strategy which used the analysis results as inputs.

My approach split the data into SHAR and GHAR, the specific history analysis region and the general history analysis region. SHAR is the center of the multivariate distribution – what happens on normal days. You simulated the future by resampling from the recent past (say, three months) in low-dimensional vectors (say, three market factors). The tails are GHAR. I claimed that you often couldn’t define P&L in the tails, that markets might be closed or illiquid, or that different ways of decomposing your position might result in different values. Creditors, exchanges, or governments might change the rules of measurement (I was thinking here of things like the Hunt brothers’ problems with silver, which were recent at the time). Even if you could somehow define a price, it would be a firesale price, rather than something that brought together willing buyers and sellers, and was a close to unbiased predictor of future prices. These things were particularly important, as I was thinking about strategies that required trading to execute, rather than fixed portfolios.

In GHAR you simulated high-dimensional vectors from models based on long-term historical tail movements. My dissertation consisted of a systematic way to do separate SHAR and GHAR and simulate each with minimal assumptions, along with some theorems and empirical data to support the technique. My main claim was not that the method offered superior predictions to alternative methods in traditional measures, like mean-squared error, but that it produced better outcomes when used to guide dynamic trading strategies.

The talk

I began my November 1987 talk by reviewing the main points of my dissertation, then discussing the Crash. On October 19, 1987, stock markets around the world declined by around 25 percent, for no apparent reason. But that was the least of the shocks. For a trader, there were all kinds of disruptions that would have been considered inconceivable a week earlier. There were unprecedented volumes in some markets, and no liquidity in others. Spreads moved to economically irrational levels. Some markets broke, others became far more efficient. Most incredible of all, when the dust cleared, prices were much more rational than before (e.g., consistent skews and smiles had appeared in all liquid option markets and some long-standing arbitrages had disappeared from the mortgage markets).

Next, I gave one historical example from my dissertation. In July 1946, the US inflation rate shot up eight standard deviations above its historical average. Data from the previous 25 years showed a slight kurtosis, but nothing that would invalidate the use of Gaussian-based projections. From the beginning of 1921, when the consumer price index (CPI) was introduced, to June 1946 there was net deflation in the country; July 1946 on its own was enough to offset the deflation of the previous quarter century and move the CPI to net inflation, from which it never recovered.

Immediately after the fact, people published a variety of explanations: pent-up spending, post-war euphoria, measurement changes, good weather, government spending, relaxation of price controls – all of which missed the point completely. They may or may not have hit on some of the immediate, specific causes of the price increase, but they treated it as an anomaly to be explained away, rather than a harbinger of increased price kurtosis that remains in effect 60 years later and a regime shift from systematic deflation to persistent inflation. The monthly standard deviation of inflation fell from 0.75 percent to 0.45 percent pre-July 1946 to post-July 1946, but kurtosis jumped from two to 32. The mean went from negative to positive.

One transparency is titled, "Disasters Come from the Future, Not the Past." People were obsessed with explaining disasters based on the past, so they could predict the next one, and so they could pretend they understood the economy. But disasters were far more clearly tied to future events than past ones. Prediction seemed impossible and explanations never generated consensus. The trick, I claimed, was first to survive them and then figure out what they meant, not what caused them. In 1946, it would have been far more valuable to know that future inflation would be positive, less volatile, but kurtotic, than to know that the July inflation increase had been caused by good weather (eight standard deviations good?). Not only that, it was easier to guess the future – it was staring you in the face – than it was to explain the past.

You might think I was working toward Nassim Taleb's conception of a Black Swan. While there are similarities, such as I thought the crises were unpredictable and had extreme effects, there were fundamental differences as well. I had been raised on ideas like punctuated equilibrium and catastrophe theory. Unlike Taleb, I thought of a simple, smooth, rational underlying process. It was experienced as chaos because we didn't understand the full dimensionality and nonlinearity, and we didn't even try to understand those things because measurement error and random noise made them useless for practical prediction. The regime shifts were sudden changes in human understanding, not sudden changes in reality. Thus, to me at that time, the Black Swan was the illusion; to Taleb it is the reality. On a more practical level, Taleb emphasizes living and investing to take maximum advantage of Black Swans, while I was interested only in surviving them. Taleb downplays efforts to control everyday events, as he says they make insignificant contributions to outcomes. My strategy was to exploit everyday events to the maximum, subject to constraints imposed to survive disasters. Finally, my version was restricted to dynamic trading, while Black Swans are a much broader philosophic concept.

I'd like to say my next slide was "Implications of Skew and Smile for Option Trading," but it wasn't. I'd like to say that's because I was protecting a secret trading strategy. But I wasn't. I had noticed the change in option pricing and trading, but I misinterpreted it. I said it was easier to guess the future than to explain the past; I didn't say it was easy to guess the future. Still, I'm proud that I was in the game, betting on the future, rather than explaining to everyone that the crash was not an extraordinary event but a simple consequence of past actions, whose main implication for the future is that I'm smart. And as a result, I discovered I was wrong and tried new things. I didn't spend the next few years arguing about the cause of the crash. Also, I did a better job of interpreting the changes to the mortgage market, perhaps because that was my professional specialty at the time. That was not in the talk, however.

I'd like to say my next slide was "Implications of Skew and Smile for Option Trading," but it wasn't

Wherever there's Kellys, there's trouble

Instead, I moved on to discuss my Kelly problem. The Kelly criterion of maximizing long-term growth is a very useful approach to studying investment and trading questions. Like risk-neutral pricing, it always gives you a consistent answer, and one that can't be too far (in some hand-waving sense) from optimal. Unlike risk-neutral pricing, it factors in risk as well as expected return. That makes it a two-dimensional consistent criterion instead of a one-dimensional one, which in turn means it gives better approximations to reality when market incompleteness is an important issue. Just as no investor need be risk-neutral for risk-neutral pricing to be correct, no investor needs to use Kelly for the optimal Kelly strategy to be right.

The simplest case I looked at assumed you had an investment opportunity with payoff cumulative density function $F()$; that is, if you bet \$1, then the probability that your payout is less than c is $F(c)$. You could choose any bet size, positive or negative. If you bet B, the chance that your payout is less than c is $F(c/B)$. It is well known that a Kelly investor will bet approximately:

$$K = \frac{w\mu}{\mu^2 + \sigma^2}$$

where w is the investor's wealth, μ is the expected value of F and σ is the standard deviation of F.

Now suppose there is a small probability p that instead of getting the payoff above, you will get $-SB^2$. Squaring B means you lose for large bets, regardless of direction, and big bets cause more than proportionate harm. One possible intuition for this situation is that you think in a market crash you will be forced to unwind, and the larger your position, the greater price impact it will have. Another is that larger positions require more leverage, and the cost of leverage goes up in a crisis.

Note that this is not the same as assuming a fat-tailed distribution. The disaster is not an extreme market move; it's a nonlinear response to your bet size. This reflected my view that the really unusual thing about crises was not the large movements in underlying prices, but the large losses in hedged strategies that both history and economics argued should have been safe, and the unstable feedback loop from positions to prices and back. It's not the fat tails that kill smart traders, it's the deviations from pattern. These usually occur during large market moves, or at least during times of high market volatility, but that's not assumed in this particular simple example. For this problem, if the disaster occurs there is only one possible loss amount, given your bet size – it doesn't matter what the market does.

It turns out to be convenient to define:

$$S = \frac{-mF^{-1}(p)}{K}$$

This is no constraint on S, since we can make m anything we choose. m measures how bad the disaster is, relative to the original Kelly investor. She will lose:

$$F^{-1}(p)\,K$$

or more, with probability p. If the bet size K is maintained, the disaster payoff will be:

$$-SK^2 = -\frac{-mF^{-1}(p)}{K}K^2 = mF^{-1}(p)\,K$$

which is m times the original Kelly loss. In modern language (obviously not something I could have said at the talk), m is the ratio of stress test loss to VaR.

Some very simple calculus and algebra get us the result that the Kelly investor should bet:

$$\frac{1-p}{1-p+2pm\frac{F^{-1}(p)}{\mu}}K$$

This is a simple fractional Kelly strategy – although at the time, I'd never heard of fractional Kelly. It's very likely someone came up with that idea before me, but mine was an independent development. It fell naturally out of the analysis. I'm pretty sure I was the first (and only) person to come up with this formula for determining the fraction.

Let's go to the numbers

Next, I spent some time on empirical examples. Today, we can do better and look at a real firm. Taking Goldman Sachs more or less at random, we see the firm reported an average daily VaR of about \$200 million for fiscal 2008, and made about \$30 million trading profit on an average day. Goldman uses a 95 percent VaR, so we get:

$$p = 0.05$$

$$F^{-1}(0.05) = -\frac{200}{G}$$

$$\mu = \frac{30}{G}$$

where G is Goldman's position size (it cancels out for our purposes, so we can define it in any arbitrary way). Of course, I didn't have those numbers at the time for a typical firm, since VaR hadn't been invented, Goldman Sachs was a private partnership, and even public companies didn't disclose this type of information.

Sticking Goldman's values into the formula, we have everything we need but m. If $m = 1$, there is no disaster tail. Sure, you lose \$200 million on the 5 percent of days, but the original Kelly bettor lost \$200 million *or more* on 5 percent of the days. In that case, Goldman Sachs should trade at 59 percent of full Kelly. That's a dramatic reduction, given that it doesn't seem as if we've increased risk much. The key is that it's the derivative that matters, the power to which we raise B, far more than the probability or size of the disaster loss.

At $m = 2$, our disaster losses are \$400 million, and we should bet 42 percent Kelly. At $m = 10$, we should bet 12 percent Kelly. I don't consider 10 an unreasonable value of m; positions with a VaR of \$200 million could lose \$2 billion in abnormal markets over several trading days. Granted, this is a highly simplified example, but it seems to indicate that your notion of what might happen in a rare market disruption has a significant effect on the amount you should risk on trades.

I was surprised to discover that my next example was the crowded trade problem, although not under that name. I did a model in which the disaster loss depended not only on your trade size, but on the total size of traders following similar strategies. I have no recollection of this at all, and don't know if it was an original idea or something floating around at the time. I didn't do anything about it, and when I studied crowded trades in earnest about ten years later, I didn't revisit this work. It did show that as the trade got more crowded it could become profitable to reverse it, even if it had a good Sharpe Ratio in SHAR. That may sound unsurprising today, but it was very much out of step with quant thinking in 1987. I did not offer a practical way to measure the crowdedness of a trade, so the result was purely theoretical.

Three wise men of Gotham

The next thing I did was discuss the "V" in "VaR." There were three risk management camps in the quant world which had already come to be known as Value, Earnings, and Capital. I was in the Value camp, which held that risk should be defined by mark-to-market P&L change on normal market days. That's a trader's perspective; every day is a fresh start.

The Earnings camp thought in longer terms, like making a loan. They said risk should be defined by effect on earnings. For trading positions, this amounts to the same thing, although there are differences in detail. But for a loan you intended to hold to maturity, only P&L moves due to changing credit counted for risk. P&L moves due to interest rates were irrelevant, since they didn't affect your cash flows. We Value types said that a long-term fixed-rate bond was riskier than a floating rate bond with the same credit, because its price was more volatile. The Earnings people said that the fixed-rate bond was less risky, because you had less uncertainty about your cash flows.

The Capital gang dismissed us as controllers and accountants, respectively. They defined risk by the cost of the economic capital required to support a strategy. What you held now wasn't as important as what you might hold in the future. On the one hand, if your strategy had a big limit, but small current positions, they said you had a lot of risk, because you would need to hold a lot of capital to support potential future positions. On the other hand, if you had large positions with a stop loss, you could have a relatively small risk.

By the time of my talk, all three groups had settled on formal definitions of the random variable underlying risk. We Value people defined it as the P&L move of your current positions frozen until the next close of business, assuming normal markets at that time. Positions were frozen because we didn't believe in relying on the ability to trade, and also to remove ambiguities in the measurement. The next close of business was used because that was the only practical time to combine positions from different markets and

>

An awful lot of it (VaR) had been invented in a month by people with full-time financial jobs in a market crisis

time zones (as a practical matter, it was done anyway and we believed in using the data we had). Normal markets were specified to remove ambiguities from closed or illiquid markets.

The Earnings camp defined the underlying variable as the change in the present value of the expected cash flows of your current holdings over their projected holding periods, discounted at the average risk-free rate over the period. Capitalists said that it was the undiscounted minimum value of your total portfolio over a period appropriate for your business. All three of us did Monte Carlo simulations of potential futures to estimate the distribution of our favored risk variable.

This is how "VAR" got into my talk. I discussed how your definition of risk affected your choice of resampling parameters for your simulation. There was a Value Analysis Region, an Earnings Analysis Region, and a Capital Analysis Region. In this discussion, I sketched out how different risk measures could explain some of the institutional structure of the financial industry. Commercial banks were judged on capital, nonbank public companies by earnings, and private entities and fund managers by value; each of which led to different optimal decisions.

Words, words, mere words, no matter from the heart

People sometimes ask me what the term "Value-at-Risk" means. It's not obvious from the words. "Value" comes from the camps above; it uses daily P&L. But Value people tended to summarize risk using moments of the distribution. We were looking at one-day measurements we thought were reasonably uncorrelated from day to day. So, we had Value Standard Deviation, Value Skewness, Value Kurtosis, and so on. We didn't call them that; once you were in the Value camp you only had to say "standard deviation" for people to know what you meant.

The Earnings camp thought in terms of quarterly earnings. They used statistics like the probability of wiping out a full quarter's earnings, which they called the Full Earnings Risk Level. They also had Half Earnings Risk Level and Twice Earnings Risk Level, and so on.

It was the Capital types who thought of things "at risk." Your capital at risk was the amount of capital you were using, and might be more or less than the actual capital you held. Of course, you tried to match the two; holding capital that was not at risk wasted money and not having enough capital to cover what was at risk meant you could blow up. I don't recall ever seeing it abbreviated "CaR," it was usually all capitals. That was often true of "VaR" as well, until the late 1990s, when eBay, eTrade, and the like accustomed people to nonstandard capitalization.

By 1990, some of the more aggressive trading banks noticed that big losses were coming not from extreme market movements, as in the crash of 1987, but from several different parts of the firm making the same economic bet. A search began for a way to aggregate risk across different types of trading desks. Value moments failed, because you couldn't get good estimates of the required covariances, not to mention higher-order dependences like coskewnesses. Capital at risk used percentiles, which are easy to aggregate because they can be estimated without distributional assumptions. But Capital was too business specific to be defined the same way across the firm. Earnings suffered from both problems; it wasn't defined consistently across businesses and it couldn't be aggregated. Someone, probably several people independently, got the idea to combine the Value risk measure with the Capital risk metric, and Value at Risk was born. But it wasn't just getting the idea; teams of people had to do a lot of painstaking theoretical and data work to make it a reality. VaR was developed and implemented by group effort of quants over a six-year period. Also, the synergies are essential to what VaR became. It was more than the sum of a risk measure for fat tails plus a risk metric for aggregation; it solved not only the two distinct problems for which it was designed, it led to a deep reappraisal of the meaning of risk.

At the risk of sounding immodest, I was impressed by my 1987 talk. We had a lot more of VaR, and modern risk management in general, in 1987 than I had remembered, and an awful lot of it had been invented in a month by people with full-time financial jobs in a market crisis. It's not always quiet time for deep reflection that leads to fundamental advances; sometimes it's noisy events that shake minds out of ruts. The main thing that was missing was the experience of bringing risk measures out of the seminar room and into real financial institutions. During the presentations and arguments of that process, the ideas got refined and polished. We got input from physical risk managers and financial controllers that improved the process enormously.

I have trouble talking about this period. It was a tremendously exciting time. We young quants had nothing above us: no older people who knew more about things, no obvious limit on how much change or money we could make, no rules to guide us. It seemed entirely possible that one minor advance could restructure finance or create huge wealth, or both. Many people in their 20s find some enthralling ideas and an energizing group of like-minded talented people. Those days seem to burn brighter than any time before or since. You can never explain it to people afterwards; if you weren't there, you can't know. I'm happier now then I was then, and I wouldn't want to go back, even for a visit; but the times mean something important to me, something I will take to my grave.

Richard Winters was quoted by Steven Ambrose in *Band of Brothers* as telling his grandson, "I wasn't a hero, but I served in a Company of heroes." While I admit that being a financial quant in the 1980s is quite different from being a paratrooper in World War II, I have a similar feeling about that first generation of Wall Street quants. What we did collectively was important, what we did individually was not. I wasn't a revolutionary, but I worked with a gang that revolutionized finance. W