"Fintech Codgers Look Back 25 Years"

Working Draft
Journal of Investing 25th Anniversary Issue (March 2017)

David Leinweber
dleinweber@post.harvard.edu


“Who am I and why am I here?” is likely the only utterance by Admiral James Stockdale, Ross Perot’s 1992 running mate that anyone remembers. It is also a reasonable question about my invitation to contribute to this 25th anniversary issue. I was a frequent contributor on technology to the Journal of Investing, its sister publication, the Journal of Portfolio Management, and others starting in the late 1980s and going through the early 2010s. Many of those vintage financial tech publications are in references [15-27] at the end of the article, to give a sense of “why I’m here”.

The best of the bunch were updated and included in the book “Nerds on Wall Street: Math, Machines and Wired Markets” [15], available for purchase at Amazon, or as a bootlegged download, some virus-free, at warez sites around the world. In the NOWS book, and most of the articles, I tried to keep the presentation light, accessible, and hopefully, funny. The occasional successes in this regard led to giving lots of dinner talks to audiences with non-financial spouses and guests, who generally had been enjoying open bars through out the evening. I was often introduced as the “class clown of quantitative investing and trading”, which I took as a compliment. Some live on still on YouTube. [32-34]

Over the past thirty years, I was a founder or (co-founder) of two financial technology firms, an advisor to many others, a multi-billion institutional quant equity manager, and a low form of faculty life (visitor, adjunct, fellow) at Caltech, USC, and Berkeley. Given my absence from finance (except for a number of surreal expert witness experiences) I was flattered get Brian Bruce’s invitation to contribute to this issue.

The initial plan for this article was to have an “e-roundtable” discussion with a small group of other “fintech codgers”, passing around a series of drafts for informed 25-year insights/pontification/commentary/flames on these topics:

- High speed markets and High Frequency Trading
- Cyber-risk and TEOTWAWKI ("The End of the World As We Know It")
- What happens when everything is indexed?
- Impacts of Artificial Intelligence.
- Age-ism in the investment industry.
Doing an asynchronous e-roundtable isn’t easy. A search on “theory and practice joke” produces 11 million results. Most of them are applicable to the experience of trying to organize the electronic round-table discussion on these topics. It sounded good in theory, but was problematic in practice. Some people sent aggregated tomes larger than a whole issue of the Jol. Some sent marketing slogans. From “Nah!” to “I’m too busy running my bar on Maui.” to “I have condos to build in Switzerland.” to “Let’s do lunch.” There were some useful and timely contributions, which are included. I have drawn on many of the verbal comments and I appreciate the efforts that went into all of them.

Elaborations, distillations and highlights of the 25-year retrospective financial technology codger views on these topics follow.

**High speed markets and HFT**

Back in the 1980s, Dale Prouty and I left the mainstream AI world to develop a real-time market analysis product called “MarketMind” ([22] and chapter 3 in NOWS). It was acquired by Jefferies, extended to do actual trading, and renamed QuantEx, for “quantitative execution”, then spun off, with the POSIT cross, as the Investment Technology Group [ITG:NYSE]. QuantEx was one of the first commercial AI systems for the buy-side to use NYSE and NASDAQ’s electronic stock order systems. We didn’t call it HFT then – the characteristic times were in full seconds. For years, there was a “safety stop” feature that allowed human traders to review orders and cancels. It was still enormously faster than phone and fax communication between institutional investors and brokers. Automation led to the precipitous drop in commissions (fixed at 25 cents per share for years prior) and trading costs. Systems in today’s much more complex markets have characteristic times in micro- and even nano-seconds. There is a range of opinions on whether this is a positive development.

The role of equity markets is, in theory, to allow for efficient capital allocation and investment in companies that produce jobs, wealth, and economic growth. Technology was a great facilitator in this from the beginning. Telegraphy, ticker tape, market data feeds, and electronic execution systems expanded the pool of potential investors from a bunch of Dutchmen under a tree to the whole world. Markets were liquid, robust, and mostly efficient. See chapter 1 of the NOWS book for many entertaining pictures on this topic.

There have always been intermediaries in markets – including market makers, and floor specialists and upstairs brokers. Technology (and regulation) has transformed these roles dramatically. Moving from trading in eighths, to sixteenths to pennies to sub-pennies and from time scales of minutes to nanoseconds may have adverse consequences when taken to extremes.
The Flash Crash

By some accounts, the Flash Crash of May 6, 2010 was an example of an adverse consequence. Highly fragmented markets, with a more than a dozen electronic trading venues, with thousands of participants feeding orders and cancels at megahertz rates can saturate communication systems. Delays from the electronic market analog of the dreaded “spinning wheel” can lead to apparent invalid market states, which send all the trade-bots to the sidelines, resulting in an evaporation of liquidity in the limit order book, followed by flash crashes (the big one in 2010 was just the most impressive example). Figure 1, from the SEC/CFTC report on the Flash Crash [8] illustrates this.

Figure 1a. Vanishing liquidity on both sides of the aggregate limit order book for S&P 500 stocks at 2:45pm on May 6, 2010. Source: SEC/CFTC report [8].

Figure 1b. For some stocks, like Accenture (ACN), limit orders near the day’s range vanished entirely, leaving only stubs. ACN dropped from $40 to one cent in seconds. Let’s not do this again.
One early official explanation for the 2010 Flash Crash was that it was all due to some ill-timed futures trading by Waddell and Reid, a medium sized investment firm. Later theories laid some of the blame on a lone trader operating out of his parents’ basement outside London. These may have elements of truth, but shed no light on the many other widely reported (non Flash Crash) lesser anomalous or unstable market events, including an embarrassing failure of the BATS market system on the day of the BATS IPO in 2012. [38]

**Synchronization of market systems**

On closer inspection, it really is no surprise that we see instabilities in fragmented financial systems. These are such a technical and regulatory mashup that even their clocks don’t have to match. In 2010, it was alarming to find that systems trading NMS stocks on microsecond time scales only required clock synchronization between systems of plus or minus 3 whole seconds (i.e. 3,000,000 microseconds) relative to the NIST standard clock! This is indeed like timing a swim meet with a calendar, and makes no sense other than to accommodate the large number of outdated legacy systems (some called “clocks”) at many small firms. Member firms at exchanges get one governing vote irrespective of size. Electoral College fans take note.

In the intervening years, the synchronization standard was reduced to a full second, and now, five years later, there is progress [4]. “*Given the increasing speed of trading in today’s automated markets, FINRA believes the current one second tolerance is no longer appropriate for computer system clocks recording events in NMS securities and OTC Equity Securities, thus FINRA proposed to tighten the synchronization requirement for computer system clocks that record events in NMS securities and OTC Equity Securities by reducing the drift tolerance from one second to 50 milliseconds.*”

At one point, MIFID, the European financial governing body wanted to address this by setting a one nanosecond standard. Researchers at CERN and the LHC pointed out that this was a high bar even for multibillion dollar world-beating physics labs to achieve, so MIFID backed off to 50 nanoseconds. The whole discussion, here and in Europe was remarkably uninformed and silly. Standard GPS clocks costing under $100 are accurate and synchronized to under 100 nanoseconds, and have been available for decades [9]. The lack of synchronization between financial technology regulators and technologists, as described later, goes well beyond clocks. It remains a wholly unnecessary obstacle to safer and more stable electronic markets.

**Regulators and the Flash Crash.**

The SEC and CFTC were clearly “gobsmacked” by the Flash Crash and the task of analyzing what had happened. Chairperson Mary Shapiro wondered aloud if there were any computers that could possibly handle the huge datasets of billions of orders, cancels and trades occurring in a single day. The SEC staff, working with a handful of PCs, were similarly gobsmacked. They learned quickly that office computers were not the right tools for the job. Plans to spend billions on a future
Consolidated Audit Trail (CAT) system were announced. Wagons were circled. Hearings were held. Committees and advisory panels were assembled.

At the time, I had accepted an invitation to move across the road in Berkeley, from the business school to the Lawrence Berkeley National Lab (LBL), home to some of the world’s largest supercomputers, operated for the Department of Energy. These are huge machines [11], costing up to $100 million, with annual electric bills in the vicinity of $5 million. Horst Simon, a supercomputing pioneer, now deputy director of LBL, and I thought it would be worthwhile to point out to the aforementioned gobsmacked securities regulators that the Federal government already had the computational capacity to handle events thousands of times more data intensive than the Flash Crash.

LBL worked with financial firms to demonstrate the capabilities of supercomputers in analyzing high frequency financial data [28]. The Journal of Portfolio Management published a page 1 guest editorial on “Avoiding a Billion Dollar Federal Financial Technology Rat Hole” [29]. We went to Washington and spoke in front of the CFTC. One CFTC commissioner, Scott O’Malia, who as congressional staffer had worked with the DoE supercomputer centers, sent enthusiastically supportive fan mail.

In the end, nothing happened. The Lab went back to science. O’Malia went into the private sector. Regulators and the “usual suspects” organized more committees, subcommittees and panels. Deadlines were extended, revised, and extended again. Eventually a website was set up to store the rapidly growing document trail. A recent report found there [13] summarizes the status. Six years later, there is a CAT advisory group with 37 members, none of them from the parts of the Federal system that have been successfully dealing with far larger computational and data tasks for many years. Along with the ill-informed efforts on something as simple as clock synchronization, the bureaucratic response to integrating markets and technology is not encouraging.

Electronic markets clearly fill a necessary role. We are not going back to shouting under the Buttonwood tree. Rapid execution at reliable posted prices is essential to healthy markets. But there are issues of safety, security, stability and fairness that remain troubling. Michael Lewis is one of the most readable and entertaining business authors, going back to “Liar’s Poker”. His recent book, “Flash Boys: A Wall Street Revolt” [10] was an entertaining, but arguably heavy-handed whack on the fairness of modern electronic markets. It is also a very accessible and mostly accurate description of the technology. Still, a number of surprising errors slipped past the editors of a book destined from birth for the best-seller list. The Modern Markets Initiative [2], an HFT trade group advocates for faster increasingly wired markets. Their site includes much discussion of the topic, including “Flash Boys: Not So Fast: An Insider’s Perspective on High-Frequency Trading” [35] a detailed critique of “Flash Boys”, which follows the organization of Lewis’ book.

Steve Wunsch is no fan of the current state of electronic markets. He is a pioneer innovator in the field, founding Wunsch Auction Systems in the late 80s, which later
became the all-electronic Arizona Stock Exchange, and he holds two patents relating to the electronic International Stock Exchange. In his book [3] he is highly critical of modern markets and market regulators for losing sight of the original economic role of markets for investing. He writes: “Today, the evident economic malaise of the wealthy western countries is traceable to their having abandoned the principles of freedom that made them wealthy, in particular with regard to how they design and run their capital markets. The collapse of the U.S. IPO market’s ability to fund new technology companies and industries is examined in detail and seen as the clear consequence of this abandonment. And the rest of the west is following the U.S.’s lead. If markets had been run on the principles of modern regulators when they were forming, markets would not have formed in the first place.”

The good news.

Not all aspects of electronic market evolution are discouraging. New market safety features – circuit breakers, price move triggered pauses and halts, have helped avoid repeats of the dreadful system-wide meltdown of 2010 and the erosion of market confidence that would follow. Critics call them “Band-aids”, but for now, band-aids work.

Regulators and market participants also seem to be rethinking the wisdom of driving to extremes on speed, tick sizes and effective spreads. There are pilot projects to go back to tick sizes larger than pennies. Even though there would no technological impediment to trading in micro-penny ticks at gigahertz rates, this doesn’t seem to be a widely accepted goal.

Those keen on following this ongoing financial tech saga will find a fount of current discussion at the Tabb Forum [5].

**Cyber-risk and TEOTWAWKI**

("The End of the World As We Know It").

Network security was not something that anyone was worried about back in the early days of both the Internet and of electronic trading. Trading was initially over private lines, and the Internet was a very exclusive club, with under a thousand sites, none doing anything financial. There was no WWW, just e-mail, telnet and file transfer. No porn! No social media! There were even paper directories with everyone’s email. There’s one somewhere in my garage, and it’s a pretty small book. Only scientists used the first browsers in the early 90s, and everyone was considered generally trustworthy. Security was not a concern or a central design issue. This may be the Internet community’s biggest regret.

Now there are billions of computing machines on the net. They are not just computers. The “Internet of Things” surprise package was malware and viruses hiding in connected things like thermostats, doorbells, and light bulbs, where no one
runs any security measures at all. It sounds crazy that some guy in Lithuania can use your toaster to hack your bank account, but welcome to the 21st century.

There are all too many dreadful tales of financial cyber-crime. Many combine “social hacking”, such as phishing emails and official sounding phone calls with pure software attacks. Details are closely held to avoid telegraphing plans for successful attacks to potential attackers. Bruce Schneier is one of the most informed reputable practitioners and commentators on information security. His site [6] is a good place to stay current and BS-free.

The fact that we have not suffered a massive financial cyber attack (“The End of the World As We Know It” to conspiracy buffs) may be due to good defenses, and to the fact that there may be more of an incentive to reap on-going profits, likely well into the billions, from smaller exploits.

For those keen on seeing what a large-scale disaster might look like, I can heartily recommend the unfortunately named television series “Mr. Robot” [12]. From the title, it sounds like a kid’s show. It isn’t. It is a fictionalized, but not utterly implausible exposition of a successful attack on a financial conglomerate even bigger than the ones we actually have.

Mr. Robot went from an obscure show for nerds to an Emmy winner and a front-page cyber security story, and has spawned a cottage industry of commentary. With the details being fiction, informed security experts can speak more freely on the subject. The increasing use of cyber-money – bitcoin and blockchain systems - make all this even more complex and worrisome. Many fintech startups are active in this area. Hope they do a better and more coordinated job on security than we have seen in the past.

You might want to rebalance your portfolio into canned food, liquor and ammunition after watching.

**What happens when everything is indexed?**

One of the most interesting annual statistics in finance is the number of active investment managers who beat passive index strategies. People didn’t even keep score until the 1970s when computers nosed their way onto greater Wall Street sufficiently to make that possible. Malkiel’s 1974 classic, “A Random Walk Down Wall Street”, now in the ninth edition, broke the bad news and helped spawn the first index funds, and eventually, Vanguard, iShares, and the ever expanding world of passive management.

Recent numbers are particularly grim. The Financial Times examined 25,000 active funds, after fees, over 2005-15 [1]. Their headline in March 2016 “86% of active equity funds underperform” referred to European equity managers. Further on, the
numbers were even worse: “An in-depth study by S&P Dow Jones Indices found that 100 per cent of actively managed equity funds sold in the Netherlands have failed to beat their benchmark over the past five years. 98.9 per cent of US equity funds underperformed over the past 10 years, 97 per cent of emerging market funds and 97.8 per cent of global equity funds.”

Not a pretty picture, especially considering that “reverse survivor bias” would mean that the situation was even worse, which seems barely possible. These numbers are dramatically worse than those reported 25 years back, in the early days of the Journal of Investing.

Technology may be part of the reason. The raw material for alpha (beating the index) is information, and there is a lot more information out there for the taking with every passing day. Information is reflected in prices ever faster in the Internet era. An early striking example comes from Bob Butman. Seen in figure 2a and 2b, it compares the speed of earnings surprise information in stock prices before and after the introduction of web browsers for PCs. Details are in NOWS, chapter 4.

Figure 2a. Earnings surprises and stock price reaction before the World Wide Web. It took weeks for information to be fully reflected in prices. You could use a two-week-old copy of “Barron’s” as an information source and still make money. Source: Bob Butman, DAIS Group.
Figure 2b. Earnings surprises and stock price reaction in the first years after investors and traders could access the World Wide Web using PC browsers. The time for information to be impounded in prices dropped from weeks to minutes almost immediately. Today, it is in milli- and microseconds. Active management got harder. This effect has expanded to many more sources of information, and continues to accelerate. Source: Bob Butman, DAIS Group.

This becomes faster and more pronounced with more people looking at more information. When the status of every loading dock, mine, well, farm and laboratory is kept on net connected computers by someone, along with every economic statistic from every agency in the world, it really isn’t much of a stretch to imagine that some of this information may be found just a few seconds early, and not a stretch at all to realize that millions of investors and traders can legitimately see it and react simultaneously.

Folding in the idea of cybersecurity breaches with the role of information in producing alpha is not all that implausible. Supposedly “private” information could be exposed to profit motivated “black hat hackers”. If the email from multiple directors of national intelligence agencies, heads of state, major multinational corporations and Andrew Weiner can be compromised, the idea that active managers can be front-run seems plausible. I have no specific facts to back this up, but “facts” are so 20th century!

The motivation for more indexing is clear, but the consequences are not. A clever paper by financial codger Keith Quinton, “The Last Active Investor: A Work Of Fiction” [7] describes the travails of a fictional tournament poker player and active investor whose vanishingly small percentage of equity orders using “an archaic technique was referred to in the economic history books as ‘fundamental research’” result in “alarming price swings and circuit breaker triggers”.

Source: R. Butman, DAIS Group. Number of Days (Earnings Report = 0)
His activities result in a massive big data investigation by the SEC and FBI. "Once he was found he was handcuffed and arrested on charges of market manipulation and insider trading as his price estimates were labeled "material non-public information" and his trading caused wide swings in prices. Then some senior researchers realized he was the solution rather than a problem. He struck a plea deal with the SEC and was released and agreed to, on a monthly basis, preannounce his trades and release his pricing sheet. Companies agreed to issue him stock directly for his buys and buy back his sells."

In the non-fiction world, we can certainly count on a population of more than one active professional or individual investors who feel lucky and believe (correctly, some of the time) that they can beat the markets, and facilitate price discovery.

Jack Bogle is the founder of index giant Vanguard, and a charter member of the Fintech Codger Hall of Fame. Bogle was a principal player in the taking the investing world on the long “random walk down Wall Street” over the last 50 years. He recently sounded an alarm about indexing and ETFs [39]. Writing in the Financial Times, he points out that while traditional index funds are low turnover, their newer relations, Exchange Traded Funds, generate substantial, and potentially destabilizing turnover for a variety of reasons. He writes: “ETFs’ impact on stock trading has reached mammoth proportions. They account for nearly one-half of all trading in US stocks.” With over 6000 ETFs trading, he sees growing risks to returns, from turnover costs, and has concerns about the volume of arbitrage and speculative trading.

**Impact of Artificial Intelligence**

The term “artificial intelligence” was first used in 1956, but what it means has shifted often since then. An early AI success was “visual programming by example” which allowed non-programmers to lay out data elements on a screen and define relationships between them. Results were automatically recalculated when the data changed. Today we call this technology “spreadsheets”.

Other goals of AI were speech recognition, which now finally seems to have reached a usable level, and capturing of “intelligent behavior”, often in the form of rules. Those technologies, called expert systems, were the basis for the QuantEx automated execution systems described earlier, and many similar systems. In the late 80s and 90s, the financial world was deeply enamored with expert systems, both as investments, and as tools for investing and trading. Expectations were high, as evidenced by the *Wall Street Computer Review* magazine cover from June 1987 shown in figure 3, below.
Figure 3. High expectations for AI in trading. Socrates on the steps of the NYSE, with CRT disciples. The technology worked within limits, but the hype was extreme. This year’s breakthrough of the century often became next year’s R&D write-off. An advertisement in the magazine touted a system to “Forecast Any Market, With Any Inputs”, this is generally not a good idea. Complex AI and statistical methods, combined with access to ever more computation following Moore’s Law, make it very easy to over-fit models. One generation’s anti-data mining methods are often overwhelmed by the next generation’s computing capacity. This can be hazardous to your financial health.

Expert systems could emulate complex behaviors, described by rules or logic, but they didn’t learn. If you had somehow devised a golden goose of a trading system, but made the small error of buying when you should sell and selling when you should buy, the expert system wouldn’t notice. There was no learning. Later waves of AI technology addressed this. These included neural nets, support vector machines, genetic and evolutionary computation. The current “breakthrough of the century” is machine learning on big data. This was driven by web analytics to sell you more stuff, but has found wider application, including many in finance.
Neural Net Pioneers

Brad Lewis and his quantitative group at Fidelity were among the earliest adopters of neural nets for investing, starting in the late 1980s. In 1992 Steve Snider joined Brad’s team. Of that time, Steve writes:

“At the time I joined the quantitative group at Fidelity in the early 1990s, it was already considerably far advanced in using non-linear stock prediction models. The group leader, Brad Lewis, had replaced his multivariate linear regression models with some of the earliest neural networks for stock selection, and within a month of joining the company I was on my first business trip, to NeuralWare’s headquarters, to learn how to use and interpret their early software. Brad was already managing several billion dollars in three quantitative funds, and his newly-launched small cap fund had grown to $500 million in under two months and was taking in $13 million a day. ... The technology was growing as fast as the limited computing power could handle. Some of the early neural network software providers had their own PCI cards to install in our computers to handle the computations!”

“We had good success with our early (generic) neural networks, and eventually moved on to programming our own nets. They’d earned a reputation as indecipherable “black boxes” and some investors were scared off by that, but we knew that by perturbing the inputs systematically we could identify a response curve of outputs and “see through” the model to understand its non-linear relationships, if admittedly imperfectly. Neural networks thrive on large data sets and we learned that our cross-sectional models across thousands of stocks were more robust than time series models. The results were reproducible, logical, and consistent with less complex models (albeit more nuanced), giving us confidence that we’d not fallen prey to “GIGO” modeling. Overfitting was by far the larger concern, and one that persists to this day. That, and of course that computing power was limited in the early days and a complex network might take days to train! As computing power improved we added other techniques (genetic algorithms, rule-based expert systems, SVMs, and others) to complement the original neural networks and provide a more robust ensemble of learning models. When I retired some 20 years later we still had a few “legacy” NNs running behind the scenes to help forecast stocks.”

Other flavors of AI

Over the years, I’ve been on the call list for people with AI based investment strategies. Most want investors, some want advice. Almost all are overfit – they have extensive back-testing, often amazingly comprehensive back-testing, with nods to bootstrapping and multi-level hold-backs. Then they collapse in the only truly out-of-sample test: The Future.

An exception is Cerebellum Capital [36], founded by David Andre and Astro Teller, who may be familiar as Google’s high profile “Head of Moonshots” at Google X.
Cerebellum is an ongoing proprietary firm, and the founders are some years away from qualifying as "codgers", so they are somewhat circumspect in their description: "Cerebellum’s machine learning platform automates much of the role of a data scientist, including the processes of discovering, evaluating, and improving strategies in risk-managed portfolios. The system utilizes a wide variety of publicly available and licensed data sources. By replicating the activities of a data scientist, Cerebellum’s automated strategy discovery process optimizes for a proprietary mix of expected return maximization, risk/volatility reduction, and portfolio independence from major markets. The system has rediscovered many strategies that used to work (simple stat- arb, the January effect, simple value strategies, simple momentum strategies). We have pushed the technology to find and combine strategies directly, rather than just tuning human-built models."

In a mainstream investing vote of confidence in the AI methods, Cerebellum announced in December 2016, as this piece was being written, that Victory Capital Management, a multi-strategy firm with over $51 billion under management, had taken a minority stake in the firm. David Brown, chairman of Victory Capital explained the motivation: “We believe artificial intelligence represents a tremendous opportunity to augment and evolve active investment management. We are excited to take part in this evolution of learning and look forward to leveraging Cerebellum’s platform to create and deliver innovative investment solutions for our clients.” [37]

**Over-fitting and butter in Bangladesh**

The astute reader will note that the word “over-fitting” occurs frequently in the discussion of AI. It’s data mining gone bad. The vast majority of attempts to apply AI to trading I’ve seen suffer from it. In the 90s, at First Quadrant, we wanted to illustrate how things could go wrong with an eye-catching example. With my colleague Dave Krider, then a recent Caltech grad, we undertook to find blatant over-fitting example – a totally bogus application of data mining in finance.

In total disregard of common sense, we demonstrated the strong statistical association between the annual changes in the S&P 500 index and butter production in Bangladesh, along with other farm products. Reporters picked up on it, and it has found its way into the curriculum at the Stanford Business School and elsewhere. People I didn’t know said “Oh Yeah! You’re the Butter in Bangladesh Guy!” We didn’t publish at the time, since it was supposed to be a joke, and it existed only as a set of charts. Subsequent versions appeared as a Caltech working paper, and all over the business press. A video discussion with Jason Zweig lives on [32]. An actual paper, “Stupid Data Miner Tricks: Over-fitting the S&P 500”, eventually appeared in the *Journal of Investing* in 2007 [30]. Mark Twain (or Disraeli, or both) spoke of “lies, damn lies, and statistics” this example has all three. So, in honor of the JoI’s 25th anniversary, let’s dive into the data mine again.
Strip Mining the S&P 500

Before AI and machine learning, regression was the main statistical technique used to quantify the relationship between two or more variables. It was invented by Adrien-Marie Legendre in 1805. He used a pencil. Neural nets and all flavors of modern machine learning can find the same relationships and make the same mistakes, but regression is a good place to start. Regression analysis would show a positive relationship between height and weight. If we threw in waistline along with height, we’d get an even better regression to predict weight.

The most common measure of the accuracy of a regression is called $R^2$. A perfect relationship, with no error, would have an $R^2$ of 1.00 or 100 percent. Strong relationships, like height and weight, would have an $R^2$ of around 70 percent. A meaningless relationship, like zip code and weight, would have an $R^2$ of zero.

With this background, we can get down to some serious data mining. First, we need some data to mine. We’ll use the annual closing price of the S&P 500 index for the 10 years from 1983 to 1993, shown in Figure 4.

![The S&P 500 1983-1993](image)

Figure 4. The Unmined Data.

This is the raw data, the S&P 500 for the period, what we are going to predict. Now, we want to go into the data mine and find some data to use to predict the stock index. If we included other U.S. stock market indexes such as the Dow Jones Industrial Average or the Russell 1000, we would see very good fits, with $R^2$s close to 1.00. That would be an uninspired choice, though—and useless for making the point about the hazards of data mining.

Now we need some more data to mine in which to fit the S&P data; that is, make a correlation. We found some on a CD-ROM published by the United Nations. There are all kinds of data series from 140 member countries. If we were trying to do this S&P 500 fit for real, we might look at things like changes in interest rates, economic
growth, unemployment, and the like, but we'll stay away from those. We'll eliminate that kind of data and screen the rest to see what does the best job as a predictor in a regression.

Figure 5. The best single predictor, out of thousands. Butter production in Bangladesh. Time to start your hedge fund!

We have a winner! Butter production in Bangladesh! Yes, there it is: a simple, single dairy product that explains 75 percent of the variation in the S&P 500 over 10 years. R\textsuperscript{2} is 0.75; not bad at all. (See Figure 5)

Why stop here? Maybe we can do better. Let's go global on this and expand our selection of dairy products. We'll put in cheese and include U.S. production as well. This works remarkably well. We're up to 95 percent accuracy here. (See Figure 6)

Figure 6. We are on an over-fitting roll now! Including US Cheese takes our R-squared to 95%
How much better can we do? All this dairy is getting boring. Let’s poke around in the UN data to see what looks strongest for boosting R-squared. How about 99 percent with our third variable: US Sheep Population. This is an awesome fit (see Figure 7).

![Overfitting the S & P 500](image)

Figure 7. The three variable model with butter, cheese and sheep fits with an accuracy of 99%. We are headed for the Forbes 400 list!

It seems too good to be true, and it is. That is the point. It is utterly useless for anything outside the particular period here.

You can keep increasing the accuracy as long as you can find data not perfectly correlated with butter, cheese, sheep, and so on. There is no shortage of that.

Outside the fitted period of 1983-93, this is a total crock. This kind of chance association will inevitably show up if you look at enough data. The butter fit was the result of a lucky fishing expedition. The rest comes from throwing in a few other series that were uncorrelated to the first one. Pretty much anything would have worked, but we liked sheep. They are more photogenic than dairy products, and make for a great slide when this stuff is shown to an audience at one of those open-bar financial conference dinners.

If someone showed up in your office with a model relating stock prices to interest rates, gross domestic product (GDP), trade, housing starts, and the like, it might have statistics that looked as good as this nonsense, and it might make as much sense (i.e., none), even though it sounded much more plausible.

“Stupid Data Miner Tricks” first appeared in 1994. Since then, we gone through roughly 15 iterations of Moore’s Law, with computational capacity (measured by number of transistors per unit of chip area) up by a factor of $2^{15}$, or 32,768. We can get much fancier than simple regression in how we choose to predict. We can use thousands of parallel processors to try all sort of fancy methods. These are always pretty impressive in sample, and in the right applications, but it is hard not to data-
mine when predicting markets or prices – they have only one past. It’s not like learning to read x-rays (something IBM’s Watson has taken up lately) where there’s always a new guy in the waiting room. One example is seen in another JoI paper, “Perils and Pitfalls of Evolutionary Computation in Finance” [31].

It’s not impossible to avoid fooling yourself with all the petaflops and terabytes, but it’s not easy. The NOWS book, and the “Perils” and “Stupid Data Mining” papers have fire and brimstone sermonettes about data mining, but I keep getting calls to sign an NDA and have lunch to see that latest and greatest efforts. At least I get lunch.

That is pretty much a wrap on the 25th anniversary codger comments and rants on financial technology. I’d like to close with a non-technical rant.

**Ageism in the investment industry**

Now that I’m in the fusion business, I don’t get to as many financial events as I used to, but when I do, it’s easy to spot old friends. All I have to do is stand up and look for the heads with gray (or no) hair. Often, there are none. When some of the early work described in this paper was done, I used to fend off dozens of head-hunter calls a month. They thinned out when I hit the age where I was supposed to buy a red sports car (but only did the test drive). Then, the questions turned from "Would you be interested in.....?" to "Do you know anyone who would be interested in....?"

Millenials and Gen-Xers take note: Fifty or fifty-five is pretty much your “sell by date” on greater Wall Street, unless your name is on the door. Most of the people I worked with were gone by then. Some of it was due to the admittedly good pay in the field. Those snarky lines about people being too busy “running the bar in Maui” or “building condos in Switzerland” aren’t pure fiction. There are more than a few yachts and jets around. But people seem to head for, or be directed to, the doors earlier than necessary. It’s part of the culture. Some firms have high stress Darwinian cultures and getting out fast is a healthy thing to do.

**When’s your LFD?**

Fans of Amy Schumer will recall a segment that one reviewer justifiably called “one of the greatest sketches in television history” [14] about actresses in Hollywood disappearing after turning 50. In order to maintain the dignity of the Journal of Investing, and sound vaguely financial to boot, let’s spell-check it to be called the “Last Fundable Day”, widely known as LFD.

Given that my LFD in finance is now behind me, I was particularly pleased to be asked to do this article for the JoI 25th anniversary. I have no doubt that someone reading this will pass their own LFD, and be asked to do one for the 50th anniversary issue. Say “yes”.
Until then: Buy low, sell high, and don't confuse brilliance with a bull market.

Acknowledgements

References:


[11] Top500 is a site showing the 500 fastest non-classified scientific supercomputers in the world. https://www.top500.org/


[19] Leinweber, D. “If you had everything computationally, where would you put it, financially?” Journal of Portfolio Management, Winter 2005


[27] Leinweber, D and Madhavan, A., “Three Hundred Years of Market Manipulations”, Journal of Investing, Summer 2001


[32] Video of WSJ Interview with Jason Zweig on **Butter in Bangladesh as a Market Predictor**


[33] Videos - **“Markets & Technology”** Light 15 minute talk for Institutional Investor, with many fine pictures of livestock. Nicely broken into four YouTube size snippets.

http://www.youtube.com/watch?v=R_OdzQrBvGE
1-Introduction

http://www.youtube.com/watch?v=OQt_YsoQ6k4&feature=channel
2-markets

http://www.youtube.com/watch?v=gLxwxIfXrz4&feature=channel
3-technology

http://www.youtube.com/watch?v=fmDB9-7XjOQ
4-computers

[34] **Video - Google Tech Talk: D. Leinweber, 2008**
"If you had everything computationally, where would you put it financially?"

http://www.youtube.com/watch?v=HJqtqNl5G4E


[37] “**Victory Capital Announces Investment in Cerebellum Capital**”, BusinessWire, December 5, 2016,

[38] “Most embarrassing IPO failure wants a redo”, CNN Money, April 4, 2016