If You Had Everything Computationally...

Where would you put it, financially?

David Leinweber

The *Journal of Portfolio Management*'s 30th anniversary issue touches on the changes brought about in investment management by progress in computation, and some areas where we can expect significant progress in the future (Focardi, Kolm, and Fabozzi [2004]). I'd like to thank Frank Fabozzi for inviting me to expand on this topic.

It's illuminating to look back on past so-called technological breakthroughs of the century to see which of these ideas have proven valuable, which have slipped into the "further study is warranted" zone, and which will get you in trouble if mentioned in polite financial company. It is also illuminating, if highly speculative, to look at the current state of technology and consider the changes in investment and trading that will be brought about in the future.

I have seen a number of technological charges up the hill, and a number of retreats right back down. This perspective is valuable in assessing our past progress, and perhaps a realistic filter in extrapolating into the future.

Moore's Law is the well-known doubling of computation power every 18 months. Metcalfe's Law is the less well-known maxim that the utility of a network grows with the square of the number of users. These are as close as we come to laws of nature for technology.[1]

The extent of the progress as we move out to the further reaches of Moore and Metcalfe space can't be understated. There's more computation power on your desk than existed in the world 30 or 40 years ago. This may even be true for your wrist watch if you have a sufficiently fancy model.

David Leinweber
is a principal at Leinweber & Co., in Pasadena, CA.
dleinweber@post.harvard.edu

There's no need to belabor analogies about computer progress comparing automobiles and airplanes. A novel way of looking at what we have at our disposal for technological innovation in finance is provided by Hans Moravec, of the Carnegie Mellon Robotics Institute and author of *Mind Children: The Future of Human and Robot Intelligence* [1988].

He compares artificial and natural "computing machines" by storage capacity and complexity. His remarkable chart is seen in Exhibit 1, showing the computers we've relied on for years as comparable to common household pests.[2]

IS IT ARTIFICIAL INTELLIGENCE OR JUST COMPUTATION?

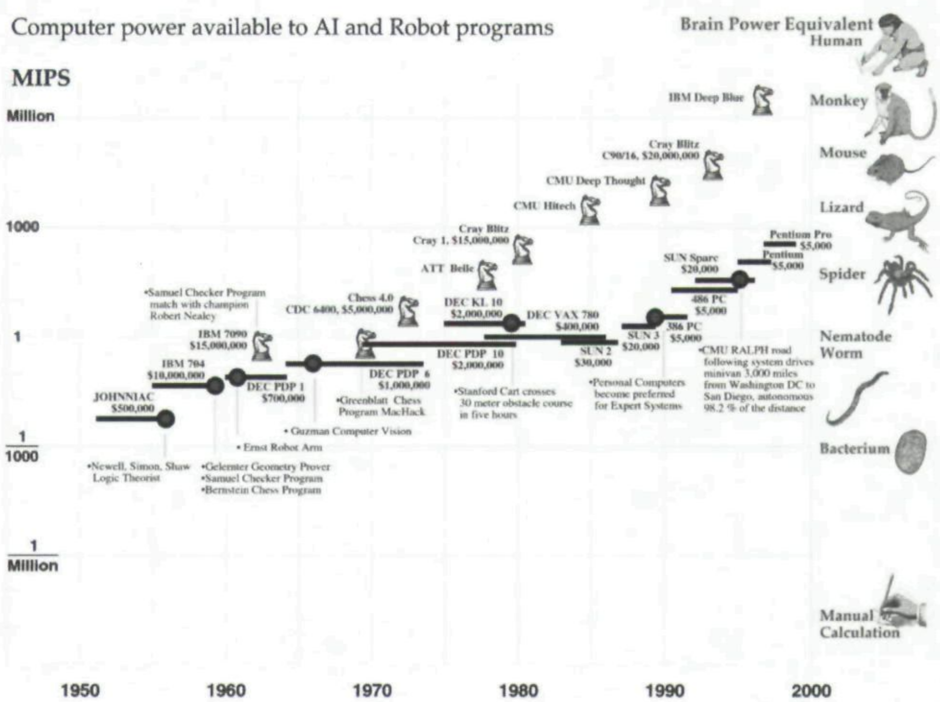
Deep Blue, famous for achieving one of the long time goals of the artificial intelligentsia, is shown in Exhibit 1 as an AI program. Remarkably, the developers of Deep Blue disagree. "It's not AI—it's just search," say Deep Blue's makers.

Even more remarkably, Kasparov disagrees:[3]

> People in future generations would look back and say this was the moment when for the first time a machine was superior to all human beings in a purely intellectual field. . . .
>
> The match draws attention to very important questions that will confront us in many different areas in the not-so-distant future. Deep Blue shows us that machines can use very different strategies from those of the human brain and still produce intelligent behaviors.
>
> If you watch the machine play—and especially when you play against it—it is very difficult not to think of it as being intelligent. Man will have to accept that using the specific faculties of the human brain is not the only way to solve intellectual problems. . . .
>
> Often you get the feeling that the computer is trying to trick you, that it is "enjoying" the position, or even that it is laughing at you. Of course it is not, it is just processing billions of numbers. But its prodigious calculations translate into a behavior that is very hard to distinguish from the activities of a human chess master.

EXHIBIT 1

Hans Moravec's March of Progress—Computer Power Available to AI and Robot Programs



The point is that it is getting hard to distinguish between simple computation and intelligent behavior. The Massachusetts Institute of Technology recently merged its Artificial Intelligence and Computer Science laboratories to reflect this truth.

Along the path from adding machines to Deep Blue and the World Wide Web, we've seen some remarkable market applications of (take your choice) AI and computer technology. We can certainly expect to see more. I describe in this article some of the greatest hits and misses of the past, and engage in the risky practice of picking tomorrow's tech stars.

UNDISPUTED CHAMPS

There are some clear nominees (and one honorable mention) for the Financial Technology Hall of Fame that have brought widespread, nearly universal, changes in the way people view and participate in markets. Topping my list are:

- Electronic market access.
- Market data graphics.
- Spreadsheets.
- Databases and Internet information.

Electronic Market Access

Thirty years ago, the DOT and Nasdaq electronic systems were concepts on their way to notions. A program trade involved literally wheelbarrows of paper trade tickets physically distributed to floor traders.

Today, trading floors from London to Tokyo have been replaced by machinery. Purely electronic markets, the ECNs have come from nowhere to claim a significant portion of volume. Increasingly, the discussion of the future of exchanges is a discussion of technology.

In August 1999, the cover of *The Industry Standard*, an Internet trade publication, proclaimed "Stock Exchanges: R.I.P." That was premature, but the terms of the Archipelago-NYSE merger show a remarkable shift in the relationship of exchanges and technology. The market centers that were once customers for technology are becoming technology firms themselves. (*The Industry Standard* itself rested in peace following the 2001 market "adjustment.")

This ongoing refinement of the techniques of electronic market access has also been one of the stealth success stories of the past decades, and is a likely prospect for future innovation.

Market Data Graphics

Through the early 1980s, the market data terminals, almost all called *quotrons*, whatever their corporate origins, looked like computer screens, but they were really keyboard-controlled cable television. They provided a picture of the data, the last trade, and the quote. It was the same information one would find by grabbing the end of a paper ticker tape and running it back to find the newest information for the stock of interest. This was literally what people did for a long time to get current quotes.

With the advent of true machine-readable market data, it was not long before these basic pictures of data, showing only simple text, were replaced by screens with an increasing variety of market charts, showing both common and customized analytics. The tedious work once performed by analysts with pencils, calculators, and rulers essentially disappeared.

Market data graphics are ubiquitous. The ticker tape that was everywhere on Wall Street gave way to primitive text-based electronic displays, typified by Quotron. Those early efforts have been displaced by the multiscreen color graphics seen in such abundance today.

When we consider the advances in graphic technology and scientific visualization tools in other fields, it is actually somewhat surprising that the systems in wide use today remain as simple as they are, showing tables and two-dimensional minimally interactive charts. While this is perhaps testament to the difficulty of conveying abstract visualization well, it is also a clue on where we may expect significant progress in the future.

Spreadsheets

In the mid-1970s Bob Franston and Dan Bricklin, students at MIT (and later Harvard Business School), were truly the defining example of the nerd. They tended to punctuate their sentences by saying "semicolon," which made them outliers even in the nerd-rich environment at MIT. But they were smart nerds, and were working on an interesting problem: Could a non-technical person "teach" a computer to write a program by showing it an example?

This was a real problem. Getting a computer to do anything useful requires jumping through many software hoops: compilers, loaders, linkers, debuggers, and the like. Franston and Bricklin had the excellent idea first, of restricting the problem to the domain of two-dimensional tables, and, second, using a light-pen (grandfather of a mouse) to lay out the grid and the relations on a then-rare cathode ray terminal. After many iterations, they had a system that they called *visual programming*. When it ran on million dollar PDP-10s, it was a clever idea. When it moved over to the Apple II, it was called Visicalc, later supplanted by Lotus 1-2-3, and later still by Excel.[4]

Spreadsheets revolutionized Wall Street and the rest of global business. Spreadsheets made all of us programmers, and we hardly noticed. Spreadsheets moved customized computation from the purview of wizards in white coats to your desk.

Databases, the Internet, and the World Wide Web

No list of breakthroughs of the century would be complete without a mention of the Internet. Before wide use of the net, private databases such as Compustat saw wide use among technically adept investors. Now these and thousands of additional sources are accessible to all on the Web, another example of the migration of complex technology from early adopters to general use.

We often think of the Web and the Internet as the same thing, but the Internet is more than the Web. It is the technology that has expanded and transformed markets by making computer communication as simple as turning on the lights.

Many readers will recall the complexity and the flashing red lights of the early modems, and our efforts to master the cryptic rites to connect to other machines. All that is gone. Internet protocols, from the simplest connection layers (TCP/IP) through e-mail and the Web, have leveled all but deliberate barriers and many of those as well.

No aspect of financial life is untouched: research, risk management, trading, and investor communication.

We are much more capable at using structured and quantitative information on the Net than textual qualitative information. We are just starting to learn how to effectively use this kind of information driven by advances in language and new Net technologies, such as XML.[5]

Honorable Mention: Portfolio Optimization

For quantitative investors, I'd add portfolio optimization. One surprising point noted several times in the JPM anniversary articles is the barriers to wider acceptance of this technology—sensitivity to errors, and unintuitive results unless they are constrained. What impressed the Nobel Prize committee has been less impressive to financial professionals, but resampling and other computationally intensive approaches made possible by Moore's Law are reducing these barriers.

HITS AND MISSES

It's instructive to look back at technological ideas in finance that have succeeded, some spectacularly, and those that have been marginalized. This approach avoids actually predicting what will work in the future. With the speed of technological progress, it's hard to avoid extrapolating about the new and the novel. If these extrapolations prove incorrect, you can send me a nasty note in 30 years.

The Crackpot as Visionary

Peter Bernstein's popular book, *Capital Ideas*, tells the story of Bill Sharpe wandering Wall Street looking for enough computer time to run a simple CAPM optimization, when he was regarded as something of a crackpot for doing so. Now these computations are routine, although not without the problems arising from sensitivity to errors and noise.

Yesterday's crackpot becomes today's visionary Nobel laureate. Several articles in the 30th anniversary JPM describe the changes brought about by the sheer rise in calculating power.

The Crackpot as Crackpot

Some crackpots are just crackpots. The past 30 years have also seen their share of wacky ideas that remain wacky. In the 1980s, expert systems produced some impressive results in diagnosing diseases, designing networks, and running factories. Business magazines called them "the breakthrough of the century." Of course, Wall Street wanted in.

A cover of *Wall Street Computer Review* magazine showed Socrates standing on the steps of the NYSE, surrounded by CRTs, captioned "Artificial Intelligence: Teaching computers to emulate the great thinkers." Another featured a wired-up Rodin "Thinker," captioned "Computers that think like pros." This proved to be an earlier incarnation of irrational technology exuberance.

It didn't work out the way they expected. This year's breakthrough of the century turned into next year's R&D write-off. These ultra-high concept notions of building robotic market wizards were (and remain) overly ambitious. Strong believers in efficient markets might maintain that the market wizards were just lucky in the first place, and even if they were using a set of rules that could be emulated by these technologies.

Expert systems were not a fundamentally flawed idea. They worked by chaining sets of if-then rules, and there were many successes—but there were also many overblown promises.

Marvin Minsky, MIT's AI uber-maven, declared an "AI Winter" at the 1987 International Joint Conference on AI, the premier academic, industrial, and venture capital conclave for the field. It got so bad that the must-have souvenir of the event was a poster showing a bull concluding his digestive process labeled "AI" inside the international circle with the slashmark "not allowed" symbol.

Grand strategies crashed, but simpler applications, managing aggregates of increasingly complex market data and trading systems, did prove successful. One such effort led to the widely adopted and copied Quantex trading system, described in Leinweber and Beinart [1996]. This style of algorithmic trading is perhaps the fastest growing segment of the equity execution market.

I don't mean to pick on expert systems or any particular technology here. Neural nets, wavelets, chaos, genetic algorithms, fuzzy logic, and any number of others could be tarred with the same brush used here. Nor do I mean to imply that all these ideas are utterly without merit, although many of them do fall into the category of too-easy-to-abuse, becoming dangerous power tools for reckless data mining,

There is good reason for pursuing these approaches. A manager of multibillion dollar institutional portfolios at Fidelity made the point to me in private correspondence

> classical methods of data analysis assume that there is a stable process generating the data, so it is valid to attempt to deduce the underlying process from a selection of known observations. These new techniques don't require that assumption, so they may be useful in especially noisy systems like stock forecasting.

Emanuel Derman, former head of quantitative strategies at Goldman, Sachs, describes this overeagerness to blindly apply quantitative and computerized methods in Derman [2004, p. 3].

> The best quantitative finance brings real insight into the relation between value and uncertainty, and it approaches the quality of real science; the worst is a pseudoscientific hodgepodge of complex mathematics used with obscure justification.

This excellent book focuses primarily on the use of quantitative methods applied to fixed-income securities, but many of Derman's insights have wider applicability. The dangers of rampant data mining using many of these approaches cannot be overemphasized. I deliver an extensive sermonette on this topic in Leinweber [2003].

The Crackpot as Billionaire

Not everyone who decides to burn CPU cycles on Wall Street ends up writing a lessons-learned essay. There are more than a few examples of computer scientists and economists with a gleam in their eye who publish that last technical paper, and then surface some years later at the top of the hedge fund rankings, employing more programmers and Ph.D.s than a medium-sized university.

This illustrious and distinguished group includes David Shaw, once employed at the Columbia computer science department. After a stint at Morgan Stanley's Automated Trading Group, he started D.E. Shaw & Co. in 1988 with $28 million; current assets are reported in the billions.

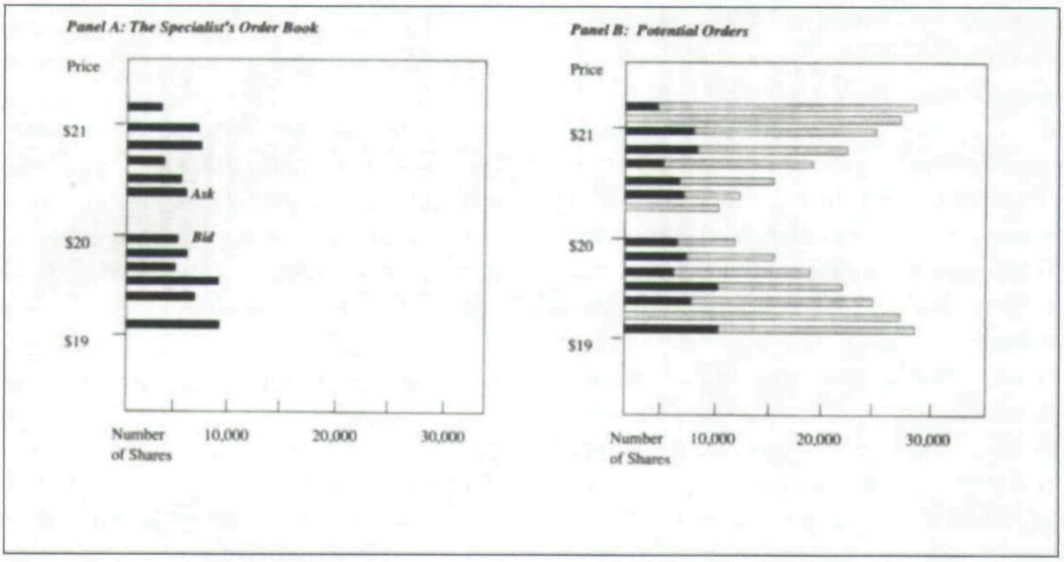
None of these firms are particularly forthcoming. In an interview with *Wall Street & Technology* magazine, Shaw once answered a question about the firm's use of neural nets by saying, "I would tell you but then I'd have to kill you." No foul play occurred, and the editor remains healthy.

One of Shaw's last publications on trading dealt with the mechanics of interfacing UNIX systems with the current generation of electronic trading systems. Subsequent in-house developments made the firm a leader (reportedly) in electronic market making, statistical arbitrage, and other fast electronic trading strategies.

David Whitcomb, a market microstructure economist at Rutgers, coauthored an early monograph on electronic trading strategies (Schwartz and Whitcomb [1988]). According to a May 2000 *Trader's Magazine* profile, he faced extreme skepticism selling his ideas to Wall Street. Finding no institutional backing, he joined forces with a computer scientist colleague to found Automated Trading Desk in the proverbial garage (in this case in Charleston, South Carolina). The firm reportedly grew from its first trade in 1990 to one of the leading electronic market participants, with over 80 million trades executed in 2002.

EXHIBIT 2

Example of Liquidity-Providing Array of Limit Orders: ATD



Source: Schwartz and Whitcomb [1988].

Exhibit 2, from Schwartz and Whitcomb [1988], shows the core idea behind these trading systems: the dynamic array of limit orders. Decimalization has made the use of these arrays even more important as liquidity is distributed deeper into the book.

Automated Trade Desk and others exploit the ability of fast machine-driven trading to find hidden liquidity, enabling them to provide better than expected (e.g., VWAP) executions for their clients, and to profit from these better executions in the same way a traditional capital commitment broker can profit by using stock in inventory to fill client orders at favorable prices.

A dynamic array of limit orders probes the market, and finds liquidity not visible in the best bid and offer information in the quotes. When spreads were wider, this could be a very lucrative activity. The move to sixteenths and later to decimals has doubtlessly made these activities more challenging.

What was a niche business open to technological innovators in the 1990s is becoming more accessible with the availability of off-the-shelf electronic market access systems from a growing number of mainstream brokers. The increasing transparency of the new National Market System, with ECNs competing with the exchanges, and the various initiatives to make more of the combined book visible lessen the information advantage to participants who deduce it using advanced technology.

A Trip in the Wayback Machine

How these firms have achieved their success is not something you read in the library, or on the Web. Their websites are limited and cryptic. In one case, Jim Simon's Renaissance Technologies (reputed to be the world's most successful hedge fund), the website is gone completely.

We can infer from the appearance of these firms at the top of electronic trade volume lists that they are keeping their machinery very active in the market.

Using the remarkable Internet Archive, the record of the Web in the past, we can see what they were saying, when they were saying more. The Archive is the world's largest history of the Web, with over 1,000 terabytes of content. For comparison, the printed collection of the Library of Congress is estimated at 10 terabytes.[6]

The picture that emerges is actually not all that surprising. The technologically innovative firms describe increasingly sophisticated trading strategies. They show expansion into ever more markets. Recruiting information and old press releases show a staff of Ph.D.s, largely from areas outside finance; computer science, AI, mathematics, and physics are popular. More surprising is the presence of computational linguists, people whose expertise lies in the area of textual qualitative information.

With these clues, we can move on to areas where technology will be more important to the broader investment community in the future.

FUTURE TECHNOLOGICAL STARS?

Prediction is very difficult, especially about the future. Nevertheless, I will venture to describe two technological areas that I believe will be important to investors and traders in the coming decade.

- Advanced electronic market tools—All investors trade, on their own or through brokers. Half of all trading is now automated. As markets become more transparent and more electronic, the growth of algorithmic trading will become more important to both buy- and sell-side traders. Simple strategies are widely used today; more sophisticated strategies are used by technology leaders. As seen in many other contexts, the complex versions will become more accessible in the future.
- Understanding textual information—Investment decisions are made on the basis of a mix of textual and quantitative information. The key elements of understanding text are collection, extraction, aggregation, and categorization. These are useful in both automated and semiautomated settings, for different investing styles. Search engines and other language-based tools will progress dramatically, with specialized systems to extract relevant portions, and combine and categorize information from text.

Electronic Markets, Electronic Trading, and Transparency

The business of "Stock Exchanges: RIP" is serious. In developed markets, virtually all orders are delivered to the point of sale electronically. Somewhere between all and most of these are executed electronically, and all are cleared electronically.

Brokers look back wistfully at twenty-five cent fixed commissions, when they have dropped to tenths of pennies. Most of the investment managers paying these commissions are using direct electronic market access. The innovators I have described have been using electronic trading as a source of profit; most typical investors use them as a means of execution.

The overwhelming majority of these executions look to transact at the volume-weighted average price. This goal, of average performance, can be thought of as the indexing of the trading world.

At the same time, very sophisticated and very fast computerized traders represent a growing share of the growing electronic execution volume. Many of these firms are considered extraordinary money managers.

What is it that the few are seeing and exploiting that evades the many that has allowed them to learn skill here? One clue comes in conversations with electronic traders who are stymied by human-speed perception in a market where limit orders come and are canceled in milliseconds. Just staying away from a buzzing stream of limit orders, or getting on the right side of it, is a good idea. But no current trading system is adept at measuring, displaying, and using this information.

Computational progress has made higher goals, profitable trading, and faster strategies accessible to more and more investors. Furthermore, the changes being made in the National Market System, which exposes the depth of orders within and across market segments, provide more raw information material to drive those strategies.

In real financial markets, we have experience only with the less-than-full degree of transparency we have had in the past. This is not true for experimental markets, and we may be able to discern some future developments by looking there. Vernon Smith shared the 2002 Nobel Prize in Economics for pioneering experimental economics. Smith's colleagues can create (and have created) markets that have any degree of transparency they want. They have created automated and semiautomated systems that may give us insight into how we will approach markets technologically in the future

Dynamic Market Visualization

Applying scientific visualization techniques in areas that map directly to physical phenomena has become standard in meteorology, protein manipulation, medical imaging, and hundreds of other areas. The full electronic equity market doesn't have such a neat physical representation to translate graphically. But the NYSE does, and it has set up a web graphic, seen in Exhibit 3.

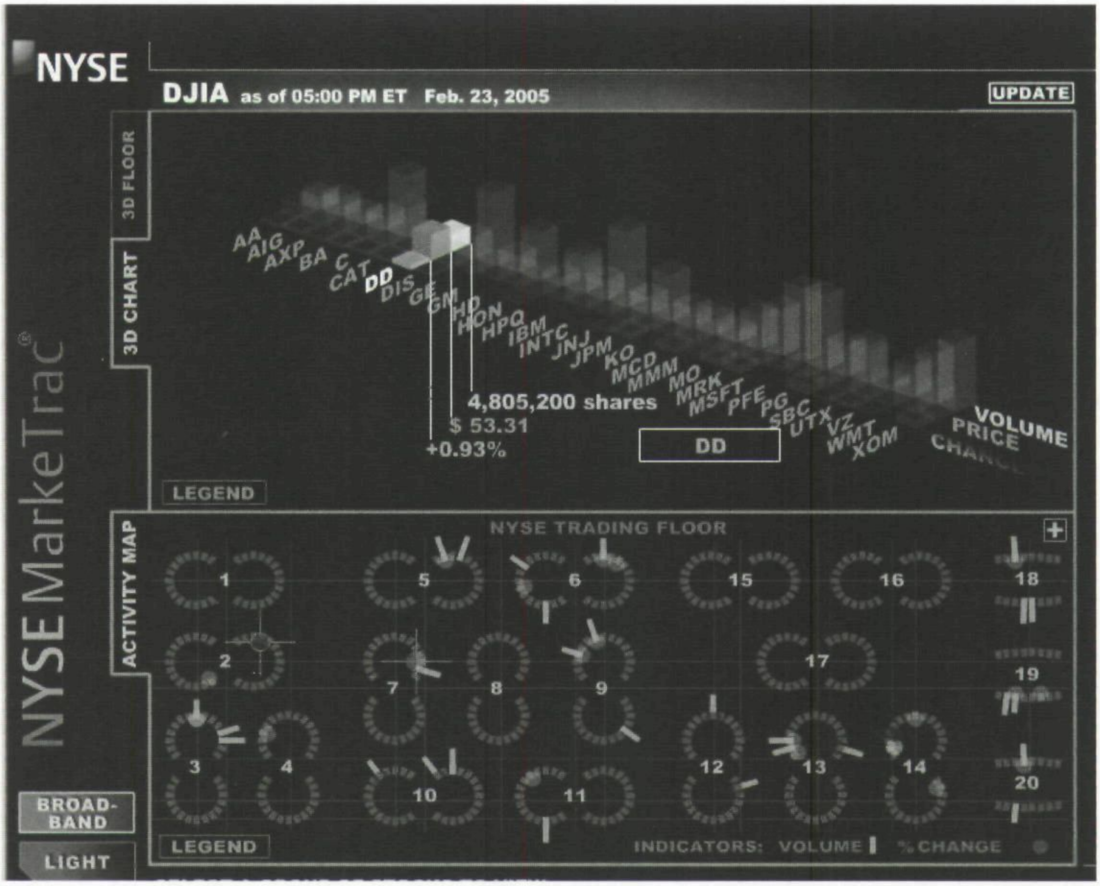
As with all these examples, still pictures are what we see in this Journal, but the dynamic updates and multiple views in the real things should be experienced.

Innovative Abstract Visualizations

Charles Plott, at Caltech, has developed a novel visualization that allows participants to look deeply into the workings of the market he created. His invention, which he calls Jaws, can be seen in living color and full

EXHIBIT 3

Part of NYSE's Marketrac Display, Mapped to the Physical Floor



The live version is found at http://marketrac.nyse.com/.

animation at his website at the Caltech Experimental Economics Lab. The basic representation of a limit order market is seen in Exhibit 4.

The real news is what you see when you run the movies, conveyed weakly by the frame shown in Exhibit 5, where the "tongue" computed from the weighted information in the book is remarkably accurate at forecasting subsequent price changes. Traders in this artificial market seem to have the ability to predict future prices by examining its current state.

People can do this, and so can machines. Other researchers, at MIT and IBM laboratories, have created automated systems using this type of information that can outperform human traders participating in the same markets.

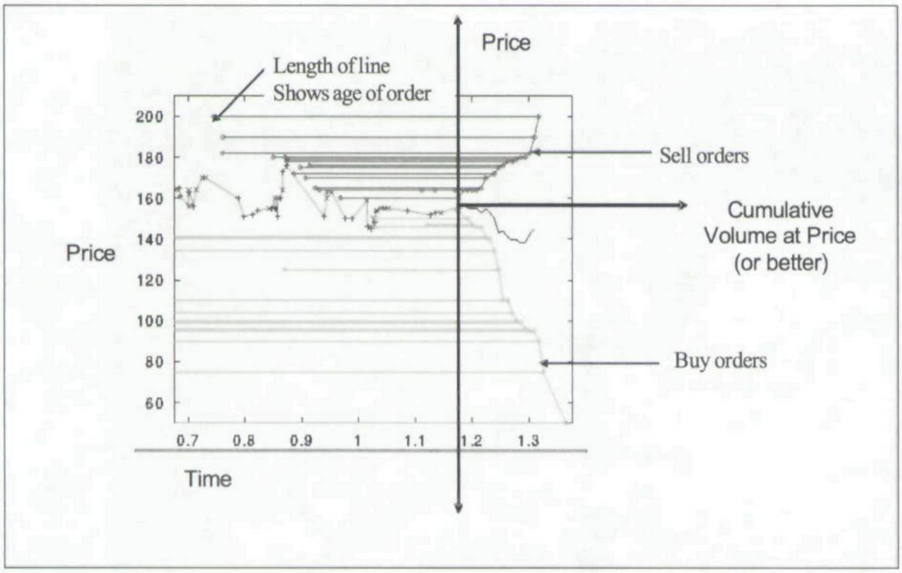
Visualization of markets, and increasing sophistication of algorithmic trading, allows traders to apply their expertise in these venues, beyond looking for VWAP executions.

Trading in the Dark

In conversations with algorithmic traders and the providers of the systems they use, we learn there are far more strategies available than people use. Yet we see from some of the fastest, most adept, trading firms that there can be substantial benefit to using them.

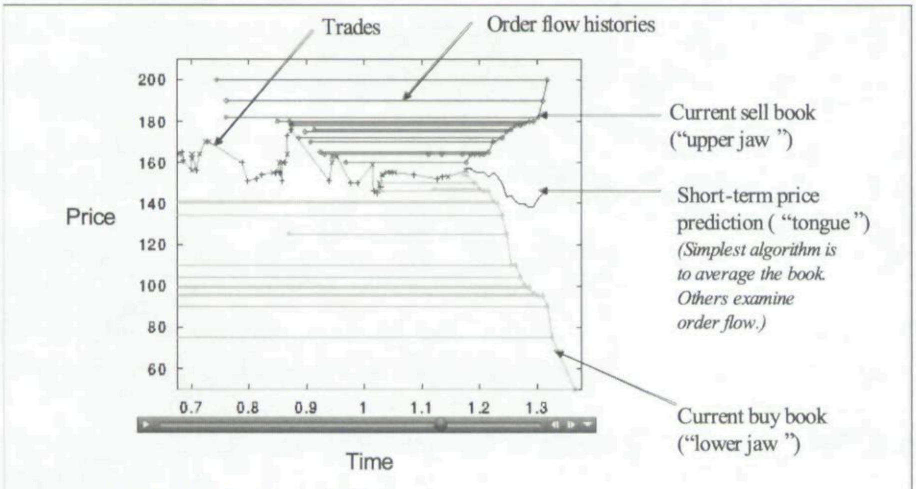
Part of the reason these have been adopted slowly by general traders is that the traders have an insufficient view of the market to decide when and how to use these strategies and to automate this process so they can handle a realistic number of orders simultaneously. With the greater transparency in the new National Market System, there is more to see for traders, but with the half-life of limit orders now measured in seconds, they lack the tools to see and understand the information available to them. Jaws is an academic example of innovative views of a market.

EXHIBIT 4
"Jaws" View of Limit Order Markets



The description is found at http://eeps.caltech.edu/. Jaws animations are at http://eeps.caltech.edu/mov/jaws.html.

EXHIBIT 5
"Jaws" Prediction Calculation



Source: Charles Plott, Caltech Econ/HSS.

High Information Bandwidth Market Visualization

Seeing the market in the relevant dimensions greatly facilitates trading effectively and intelligent automation of algorithmic trading beyond one-size-fits-all strategies. Flexible views, animation, and drill-downs convey far more information than the flat static market displays in general use.

Exhibit 6 provides a current example of a high-end viewer.

THE TEXT FRONTIER

Successful investors and traders produce value by superior access to and analysis of both quantitative and non-quantitative information. The occasional shot of blind dumb luck is helpful too, but nothing we can rely on. Firms that get a firm grip on the tools for exploiting information will build a better business than those that don't.

This has been true for a long time. In 1789 George Washington decided it was a poor start for a new country to default on its debt. He dispatched messengers from New York to sail the coast and spread the word. New Yorkers, who have changed little in 200 years, saddled up the fastest horses they could find, and set out to front-run the messengers and buy up the suddenly valuable bonds from owners who were happy to see a nickel on the dollar.

Being first in line for information has always been a good place to be, and it isn't necessary to do anything vaguely shady to get there. What it does require, in a world that becomes more wired by the day, is an ability to gather and exploit information from electronic sources.

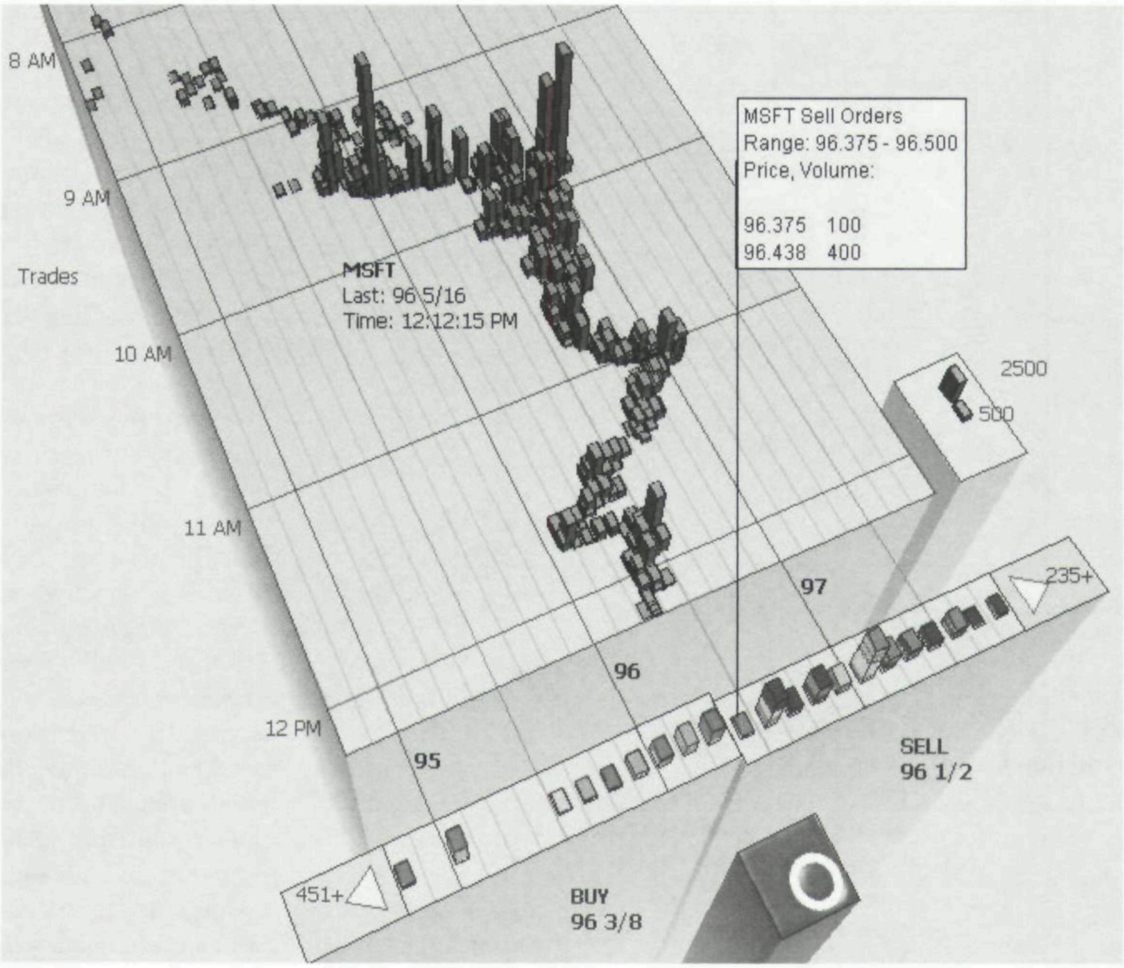
An excellent and fairly current text on technologies for machine understanding of text is Manning and Shutze [1999].

Mining the Deep Web

Popular search engines like Google index over 3 billion web pages, but that is just the tip of the information iceberg. Much, if not most, of the information relevant to

EXHIBIT 6

Dynamic View of a Market with Limit Order Book Transparency



Source: Oculus Information's Visible Marketplace (© 2004 Oculus Information). Layers on the book show grouped sizes at each price. An animated version can be seen at www.oculusinfo.com.

investment decisions is in the "deep web," below the levels that are scanned by these one-size-fits-all search tools. No one really knows just how big the deep web is, but estimates range from 500 to 1,000 times larger than the surface web.

It's easy to do a simple experiment to convince yourself. Type the name of any public company into a search engine. You'll find thousands of links, but nothing pointing you to the SEC, even though the SEC has thousands of investment-relevant pieces of information about every public company. The same is true for legal actions, regulatory decisions, and specialized research and news.

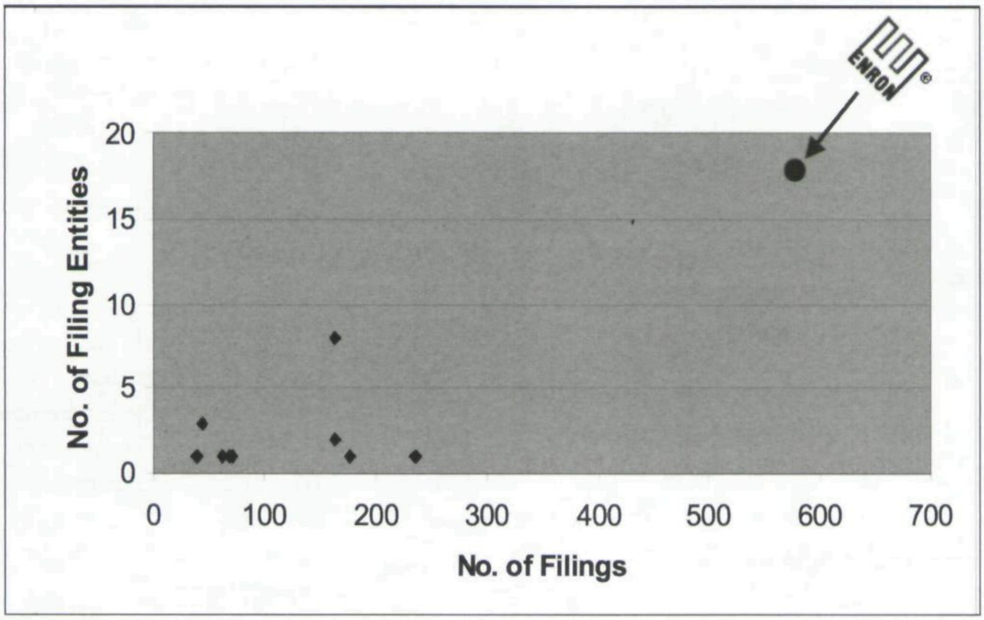
The reason is that search engines just don't look in the deep web. They send out web crawler programs appropriately called spiders that retrieve the information you see in your search results. They typically look at home pages and links off home pages. They never log on to proprietary sites, or make queries to databases hosted on the web. They don't aggregate, compare, or combine data in ways that make it useful for investors.

Lessons from the Wise Man in the SEC Reading Room

I recently learned a bit about this kind of aggregation from a prosperous retired investor. He used to spend much of his time in the SEC reading room, looking for signs of corporate trouble and getting rich by going short

EXHIBIT 7

Enron as an Outlier—SEC Electronic Filings (1993-2002)



when he found them.

So what did this man look for back in the SEC reading room, when there was an SEC reading room? "A barrage of BS," he explained, semipolitely:

> Firms are obligated to file only a relatively few documents with the SEC, 10Ks, 10Qs, proxy statements, and insider trading reports. When they start filling up the shelves, it means that the lawyers are nervous, or that something unusual is going on. Especially when there are all sorts of revisions and amendments.

I decided to see how his theory works out for Enron, the modern-day poster child for corporate BS. With a little additional tutoring, I learned that I should be looking for filings from Enron-related entities. Armed with this lesson, I dove into the deep web at SEC.gov and retrieved the gruesome details, turning up 576 filings from 18 filers.

But was this anything unusual? For comparison purposes, I did the same thing for the other top ten companies in Enron's SIC code. The eye-opening result is seen in Exhibit 7.

This kind of aggregation and comparison is admittedly pretty simple, but it's nothing that you can do with a search engine. The process can be automated, using specialized web spiders designed for this task.

This example illustrates a modern investment application of one of the oldest forms of electronic surveillance: traffic analysis. It goes back to as far as World War I, when battlefield commanders monitored traffic on the other side's radio frequencies to see when the other side was getting busy, even though they often didn't know the codes. These methods include no consideration of content whatsoever.

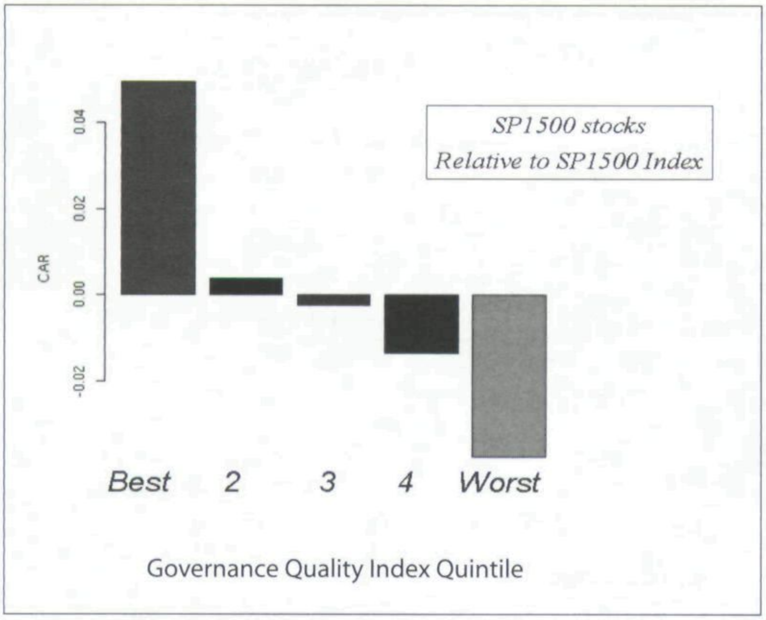
Another useful application of traffic analysis in investment is to keep an eye on court actions. Bankruptcy news often shows up first in court records. With nearly 10,000 publicly traded companies, and the reduction in research coverage for companies outside the large indexes, the collective set of financial analysts and reporters is hard-pressed to follow them all on a timely basis.

Quality of Governance and SEC Text

Of course, it is possible, and generally a good idea, to consider content as well. The SEC provides a rich source for this approach. 8K filings are for "extraordinary events," including resignations of officers and directors, mergers, acquisitions, and changes in capital structure. Financial research has shown that a better quality of corporate governance is associated with better returns.

Automated extraction tools are capable of monitoring directors' activities, audit reports, and legal trou-

Exhibit 8
Cumulative Return over 2002 CAR



Advanced proprietary trading firms have added these techniques to the sophisticated quantitative analytics in their technological arsenals.

Many of these financial applications are more secret than their Cold War cousins, but there are examples from academia that give a glimpse of how they work.

bles. When a specialized spider goes beyond retrieving information and scans it for content, it acts as an early warning system for events that may show up in tomorrow's paper, or next week's. For thinly followed firms, the events may not show up in the news at all—but they do show up in investment performance.

Textual information extracted from SEC filings can be combined to rank quality of corporate governance, by board stability and number of boards directors serve on. When this is related to stock returns for the firms in the S&P 1500, the picture is remarkable. Companies ranked in the top 20% by measures of board stability outperformed companies in the worst 20% by over 7 percentage points in 2002 alone. Firms with intermediate rankings produce intermediate excess returns, as seen in Exhibit 8.

Language Technology: Your Tax Dollars at Work

The U.S. government has been busy spending our money on technologies to do this kind of content extraction and analysis for years. DARPA (the Defense Advanced Research Projects Agency) are the folks who paid for the development of the Internet, which was called the ARPAnet in the early days. Many additional millions have gone into programs with names like Machine Understanding and Classification and Text Retrieval and Categorization. A great deal of this research has now found its way into tools and products for mining the deep web.

"The Words of the Prophets are Written on the Subway Walls"?

A number of studies of stock message boards have found that spikes in traffic, both with and without characterization of content, are predictive of subsequent spikes in volatility and volume. These results are summarized and extended in Antweiler and Frank [2002].

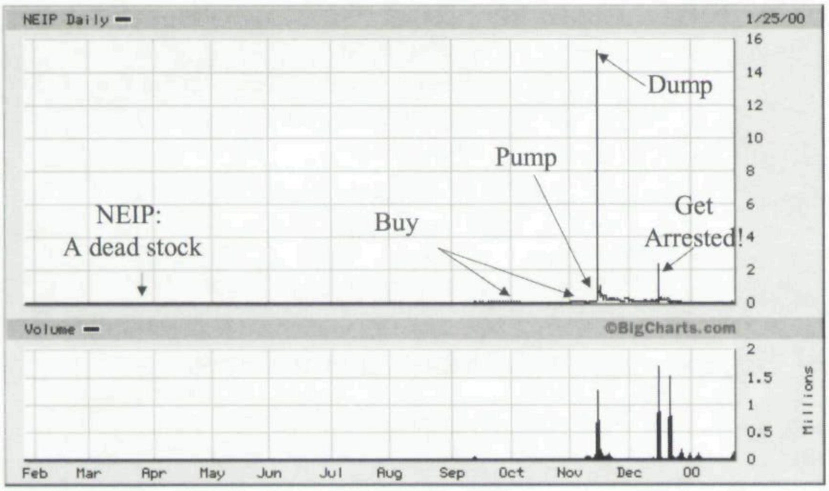
Anybody with the inclination can post messages about stocks, with no quality control anywhere in sight. This will come as no surprise to those who follow the SEC's indictments for market manipulation. In one notorious example, 130 well-placed messages sent a penny stock up 160,000% in a single day, and a month later sent the message posters to jail. Details, including the messages later used at the trial, are found in Leinweber and Madhavan [2001]. The remarkable story is told in the stock chart shown in Exhibit 9.

Outright falsification will bring the sheriff to the door but enthusiastic promotion of opinions has been going on legally for hundreds of years. It continues electronically. Jonathan Lebed, a New Jersey high school student accused of manipulation using message boards, was not charged with any crime. His fine amounted to a fraction of his profits.

It should come as no surprise that when a thinly traded stock is suddenly brought to the attention of millions of people trading electronically, there is likely to be more action than on typical days when it remains well below everyone's radar. Many options traders and market makers track net message activity as a textual indicator of risk, opportunity, and volatility.

In a recent survey of volatility prediction, Poon and Granger [2005] conclude that the best and most elaborate quantitative models do not rival the predictions of implied volatilities. In their conclusion, they write: "A potentially useful area for future research is whether forecasting can be enhanced by using exogenous variables" [2005, p. 55]. Many market makers, specialists, and options traders have chosen to use textually based exogenous

EXHIBIT 9
Market Manipulation with Messages



Source: NEI Webworld data from www.bigcharts.com.

variables for this purpose.

Can Computerized Language Analysis Predict the Market?

In a *Barron's* story, Alpert [2001] asks whether computerized language analysis can predict the market. He reports on a University of Massachusetts academic project that uses the idea of language models to look into the content of news stories, and runs a simulated portfolio based on the machine's ability to distinguish between stories that move a stock's price up and those that move it down.

Despite an admittedly bone-headed trading strategy that churned the portfolio, and without any information other than the news classifications, the system made a statistically significant profit. A simulation that ignored the news signals but made random trades at the same times they occurred incurred a statistically significant loss when trading costs were included. So even this simple approach added value.

APPROACHES TO LANGUAGE

Language models are easy to understand. A handmade language model is just a set of keywords and phrases that characterize a collection of documents. It would be easy to build one that could distinguish, for example, stories in French from stories in English. Similarly, it would be easy to distinguish stories about finance from stories about horses. Would it be possible to distinguish, say, SEC filings from companies that subsequently do well from the filings of companies that do poorly?

The statistical language models used in eAnalyst to distinguish good news from bad news depend on some of the fruits of all those DOD research projects to let the computer extract root words, and learn large sets of keywords and phrases, using that relentless way computers have of approaching human tasks, like playing chess. It finds many of the same words humans would find. "Indictment" is rarely good news, but the model also finds hundreds of others, many of them particular to an industry, or a single firm.

Here's an example of a simple experiment with language models that do this. These are the most frequently occurring words, word pairs, and word triples that occur in particular sections of SEC 10-Q filings in 1997 and 1998, segregated by firms that did well and poorly in the six months following the filing.

Many of the same words and groupings occur in filings of both high- and low-return filers. The language models include many similar terms, but the differences are striking. Looking at the top of the models, these n-grams are in the high-return group, but not in the low returns:

Research; development.
Technology.
Charge; million.
Patents.
Marketable securities.

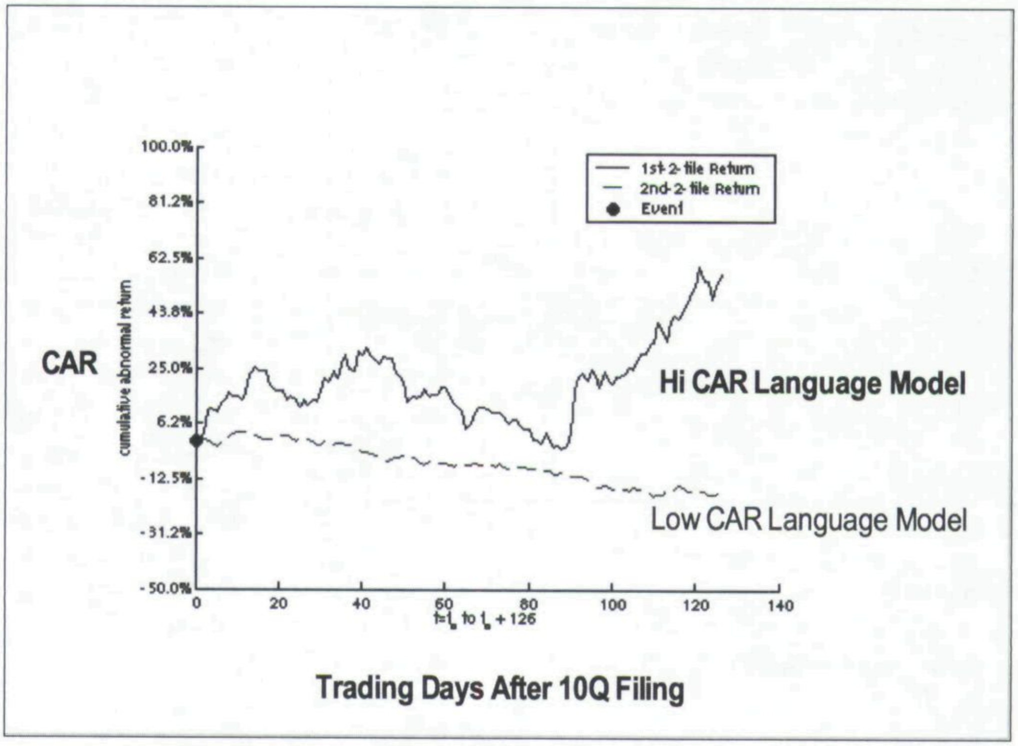
These n-grams are in the low-return group, but not in the high returns:

Regulations; securities; exchange.
Note; earnings.
Note; earnings/share.
Antidilutive.
Condensed, omitted, pursuant.

Results of an out-of-sample test, seen in Exhibit 10, that examined subsequent six-month abnormal returns (cumulative abnormal returns relative to the S&P 1500 index) for the next two years for firms whose SEC filings

EXHIBIT 10

CAR Results for Firms Differentiated by Language



matched the two language models, are striking. The graph shows a 60 percentage point spread in returns, in the expected direction, of each group after 140 trading days.

Technological Content Extraction: XML

An emerging web standard, XML, or the extensible markup language, is designed to automate the process of extracting the relevant features and portions of electronic material.

The original language of the web, hypertext markup language (HTML), was designed to present pictures of information. The "tags" used to create web pages reflected this. They described how information looked on the screen – "blue," "centered," "underlined." We are now in the early stages of a shift from pictures of data on the Web to real information. The new extended language of the web, XML, conveys the meaning of the underlying information, with tags like "ticker," "CUSIP," "news type," "revision direction," "court filing," "auditor opinion," and literally thousands of others.

XML has transformed many aspects of modern business. We are just beginning to see this transformation in the movement of information. But even at this early stage, the technology is appearing in government filings, news, and financial research.[7]

Future Uses of Language Systems

Numerical information is the bedrock of financial analysis, fundamental, quantitative, and technical. But people look at the words too. Sometimes, the footnotes to a table change the meaning of the numbers. Sometimes the words tell a story that doesn't show up in the numbers. The challenge of the text frontier is to find ways to use both types of information to improve investment and trading.

SUMMARY

I have attempted to highlight the most important impacts of technology on investment management in the past, and prognosticate on some of the innovations that may have a similar impact in the future. Of course, the former are on much firmer ground than the latter. Think how unlikely it would have been to have described the spreadsheet or the Web before they were invented.

Everybody trades; everybody reads. Highlighting electronic market access and the text frontier as promising areas for broad impact the future, I rely on the fact that they provide support for trading and reading, two very common and very central activities in investment management.

ENDNOTES

The author thanks his colleagues in both finance and technology for their helpful discussions and contributions to this work: especially Jacob Sisk, who worked on much of this article, Ronald Liesching, Andrew Lo, André Perold, Charles Plott, David Porter, Evan Schulman, Steven Snider, and others too numerous to mention.

[1]Metcalfe is the inventor of Ethernet, the ubiquitous wiring of the Internet, and the founder of 3Com. His law was demonstrated before the Net era in the lower-tech case of fax machines.

[2]Moravec expands on these and other proactive notions at his CMU website: http://www.frc.ri.cmu.edu/~hpm/.

[3]For the full interview, and more, see IBM's Deep Blue site: http://www.research.ibm.com/deepblue/.

[4]For tech history buffs, see http://www.bricklin.com/history/sai.htm.

[5]Credit for development of the web goes to Timothy Berners-Lee, then at CERN, now at the WWW consortium at MIT, www.w3c.org , which is the mother lode for information on new web technologies. Credit for the original Net, the ARPANet, lies with the U.S. Department of Defense's Advanced Research Projects Agency, www.darpa.mil.

[6]Known as the Wayback Machine, the archive can be used, for free, at http://www.archive.org/. Like Sherman & Peabody, the cartoon characters with the original "Wayback Machine," you set the time dial, and where you want to go.

[7]See the websites http://www.sec.gov, http://www.newsml.org, and http://www.rixml.org for the latest details. The SEC's version of XML is called SGML.

REFERENCES

Alpert, Bill. "Can Computerized Language Analysis Predict the Market?" *Barron's*, November 19, 2001. Available at http://online.wsj.com/article/0,,SB1005962163718064080,00.htm (for *WSJ* or *Barron's* subscribers).

Antweiler, Werner, and Murray Frank. "Is All That Talk Just Noise? The Information Content of Stock Message Boards." Working paper, University of British Columbia, 2002.

Derman, Emanuel. *My Life as a Quant: Reflections on Physics and Finance*. New York: Wiley & Sons, 2004.

Focardi, Sergio, Peter Kolm, and Frank Fabozzi. "New Kids on the Block." *The Journal of Portfolio Management*, September 2004, pp. 42-53.

Leinweber, D. "Perils and Pitfalls of Evolutionary Computation." *The Journal of Investing*, Fall 2003.

Leinweber, D., and Y. Beinart. "A Little Artificial Intelligence Goes A Long Way on Wall Street." *The Journal of Portfolio Management*, Winter 1996.

Leinweber, D., and A. Madhavan. "Three Hundred Years of Market Manipulation." *The Journal of Investing*, Summer 2001.

Manning, Christopher D., and Hinrich Shutze. *Foundations of Statistical Language Processing*. Cambridge: MIT Press, 1999.

Moravec, Hans. *Mind Children: The Future of Human and Robot Intelligence*. Cambridge: Harvard University Press, 1988.

Poon, S., and C. Granger. "Practical Issues in Forecasting Volatility." *Financial Analysts Journal*, January/February 2005, pp. 45-55.

Schwartz, Robert, and David Whitcomb. "Transaction Costs and Institutional Investor Trading Strategies." Monograph 1988-23, NYU Stern School of Business, 1988.