Slide 1

Two recent talks had the arguably overly broad title of “Big Data in Financial Markets”. It is overly broad by a factor of two, because markets have two Big Data revolutions underway at the same time.

One related to structured market data, issues around the flash crash, unstable trading systems, and cyber security of markets. That is the subject of the first portion of this talk.

The other big data revolution in markets is” in the usual sense used in the tech press – weakly structured information from textual, web, images, social media, governments and commercial sources.

Slide 2

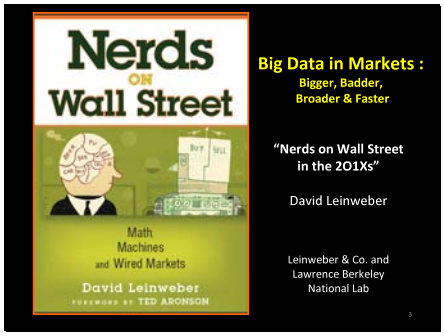
Trade Tech NY Introduction – Tony Huck RBS - David Leinweber, author of "Nerds on Wall Street: Math, Machines and Wired Markets", was recently named one of the Top Ten Innovators of the Decade by Advanced Trading magazine. As founder of two financial technology firms, and as manager of multi-billon dollar quantitative equity portfolios, he brings a practical approach to innovation. He is now principal of Leinweber & Co., and in a public service role, co-founder of the Center for Innovative Financial Technology at Lawrence Berkeley Lab.

On the back of David’s book, David Shaw wrote: "Leinweber leads his

readers through a largely unexplored forest, turning over ordinary-looking rocks to reveal hidden colonies of peculiar creatures that feed on moldering mounds of numbers teeming with trailing zeroes. His book is absorbing, instructive, and very, very funny."

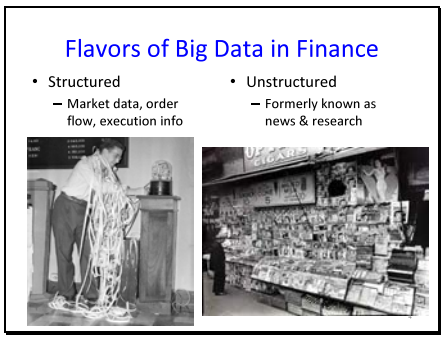
In this talk, Dave covers not just mounds of numbers - structured big data, but mounds of words. These two flavors of big data are transforming how we work in markets, and what we work with. And he has some jokes too.

Slide 3



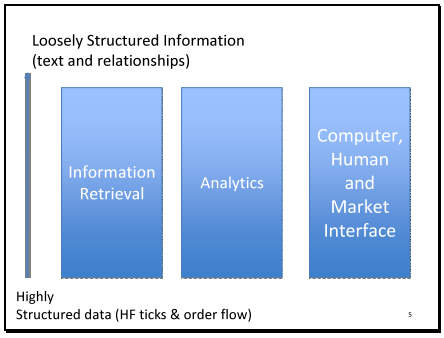
Thanks to Tony for nice intro, and setting the stage here. Great to be the first of such a fine group of speakers – less worry that they'll say everything
we hear a lot about big data, bigger all the time. Most of this is about the explosion of content of all sorts on the web. This is loosely or unstructured information – corporate announcements, commentary, postings, blogs, social media, government actions. It comes as text, with varying degrees of structure that reveal real relationships. "Main stream media news" is a refined form of all of this, and MSMs have their own challenges in dealing with all this Big Data.

Slide 4



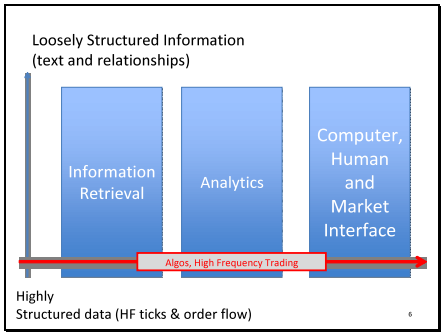
In financial markets, Big Data comes in two flavors: Unstructured and Structured. The structured big data is what market data feeds have become – bigger, and so fast, that people are hard pressed to keep up with them.
We're going to talk about both kinds

Slide 5



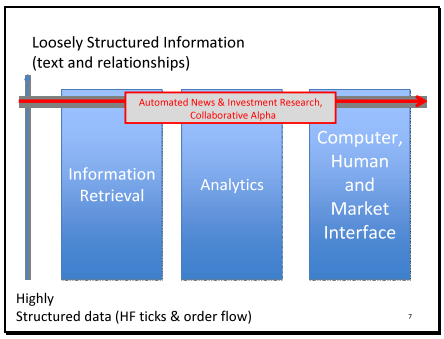
Here's a technology map for the the two kinds of big data, showing how they're retrieved over on the left, how they're analyzed in the middle, and what we do with them on the right.
Lower left – order flow, market data, fills and non-fills – all very fast, and well structured
Upper left – less structured – plain text, web pages, scans of documents, in multiple languages & media

Slide 6



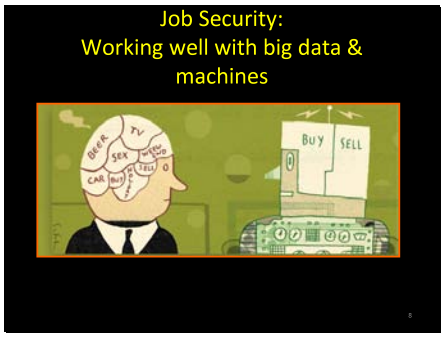
Down along the bottom are the most structured and fasted moving data. Ticks, fills, order flow, LOB information. The raw material for buyside algo trading and HFT
All it goes by pretty fast for us humans now. Machines have taken over lots of jobs in dealing with this. At this conference, we'll see some of the best ones to do this.

Slide 7



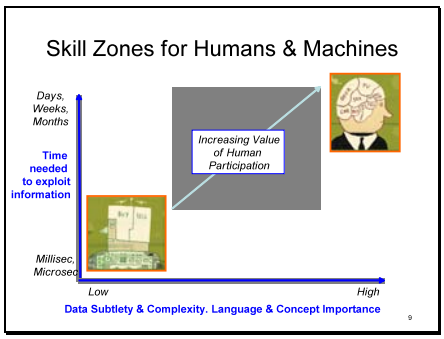
Up on top is the kind of big data you read about in Wired , NYT O'Reilly and so many other places. There's more of this relevant to markets than ever before, and lots of people are looking harder than ever before to find alpha here. Like everything else in the web era, Alpha turns up faster than it used to. And there are more places for humans to put their 2 cents, or maybe a few hundred basis points. The fast processes on the bottom are traditionally called "trading", the slower ones up top "investment research" or" portfolio management". Big data and the technology around it are moving these two activities closer together

Slide 8



All of us have seen a thinning out of people in the pure trading ranks , dealing with the fastest, most structured data. But there are still lots of tricks humans can do better than machines, in addition to being "Algo Jockeys".

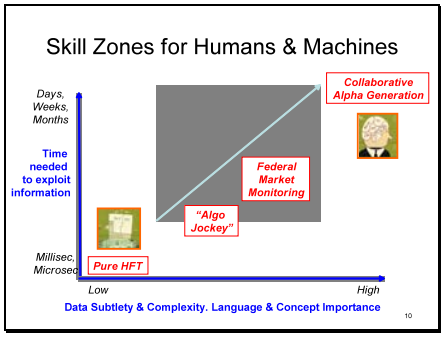
Slide 9



Here are "skill zones" for humans and machines using big data in markets. When the data is wicked fast, and ability to use depends on fast calculations (lower left), the machines have an edge.

But when data looks more complex, like text, or scanned documents, or spoken words, people still have game. And of course there's lots of technology to help there.

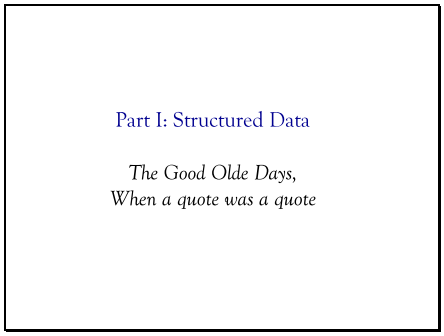
Slide 10



The word on the street is that you need to reinvent your tech self every three years. I hope that you come away from this talk with more than a few ideas about how to do that.

This is a key theme for this talk, **how people and machines working in markets can work together effectively**. "Collaborative Intelligence", "Intelligence Amplification"

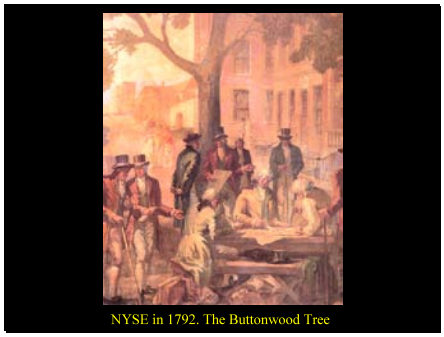
Slide 11



Let's start with big structured data. A good way to understand a Technology is to look at where it came from. We can trace the evolution of Big Structured data with some of those funny (or at least amusing) pictures mentioned in the introduction .

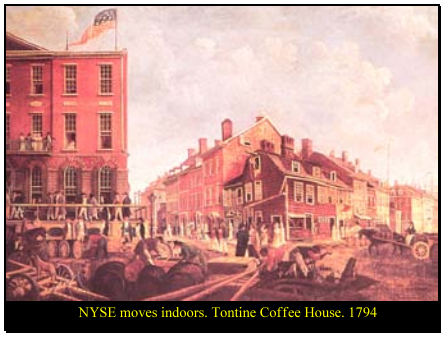
Readers looking for the unstructured Big Data portion should skip to page 31.
Though there are some nice pictures coming up.

Slide 12



Back in 1792, this is what a market data system looked like.
Shouting at the Buttonwood Tree at the corner of Water & Wall.
Noisy and rough in the rain, but a quote was definitely a quote. It stuck around long enough for any of the traders to act on it.

Slide 13



In 1794, we see a major technological innovation: The Roof.
The guys by the tree move indoors to the Tontine Coffee House.

Slide 14

They're still shouting, but they're dry.

Slide 15

When the market grew, they moved to larger quarters, and backed off on the shouting plan by putting in a technology that accommodated more orderly trading in more stocks – the Post

Slide 16

You can see the posts clearly in this close-up. Quotes were written right on them. No one pulled them down in a millisecond.

You could see them, read them, and decide to act on human time scales. Ah, the good old days.

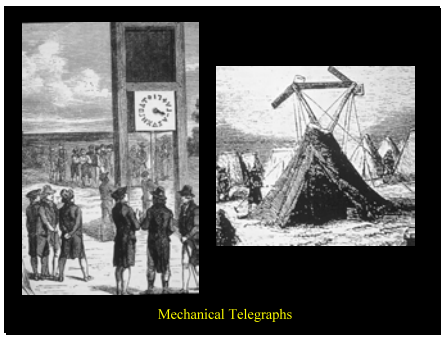
Slide 17

Here's how they resolved broken trades. Buyer dressed up in a bull suit, seller as a bear. They duked it out by the post.

All these people in one place made for big improvements in liquidity and the ability of the market to raise capital.

But you still had to show up to participate.

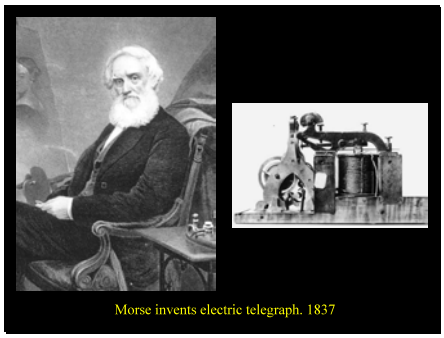
Slide 18



Moving information around in the early 1800s wasn't easy. Messenger pigeons were popular. There were lots of attempts to improve on this, with mechanical telegraphs. Here are a couple of examples. One apparently encrypted, but they had to be on hills and needed good visibility.

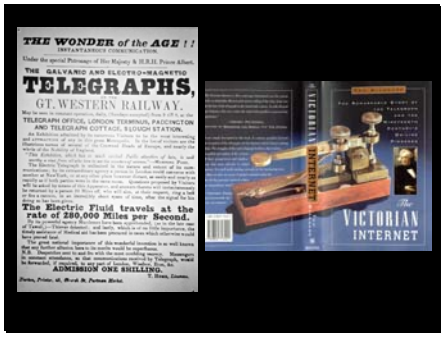
Getting past the horizon needed more of these. It took about 20 minutes for a price to get from NY to Philadelphia.
This is why there are so many places named "Telegraph Hill".
There were lots of problems with weather, darkness and privacy.
Everyone knew an electric telegraph would be a better way.

Slide 19



Finally, in 1837 Sam Morse gets it right. A simple single wire design, based on the code he invented, instead of tones or little bits of foil.

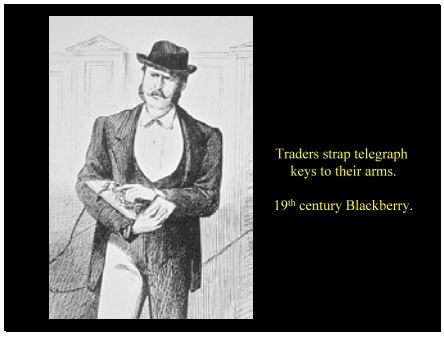
Slide 20



Telegraphy was a real world changer. The first time in history that a message could be sent over the horizon instantly.
Notice the "electric fluid travels at the rate of 280,000 miles per second, or about 1.5 times the speed of light. Maybe they knew something we don't know.

I hear there are people working on quantum entanglement channels that will do better than this – keep your wallet in your pocket on those.

Slide 21



Traders picked up on this in a big way. Here's one with a telegraph key strapped to his arm.
Sort of a 19th century Blackberry.

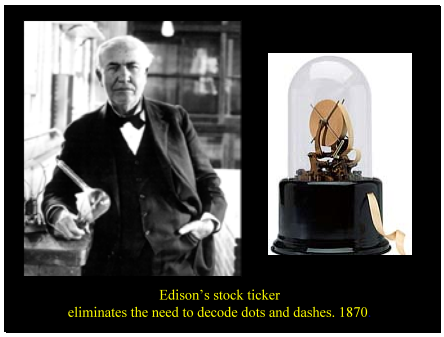
Slide 22

We can really appreciate the impact of telegraphy on the market in a pair of before & after photos.
Here's the area around the NYSE shortly before the telegraph came into use.

Slide 23

A few years later, everyone had to have it, and the sky was dark with the wires.
Telegraphy was a transformational market technology, it made for a much broader and more liquid market, and a much better process for firms to raise capital.
BUT you had to know Morse Code to participate in the market. You know there's more technology coming.

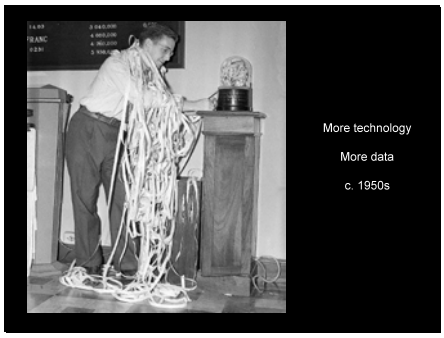
Slide 24



It came from Tom Edison, who built the first practical ticker machine. Quotes and trades could be read by anyone.

Edison made sure it was up to NYSE trader use by tossing the prototypes of the roof of his lab until they stayed in one working piece.

Slide 25



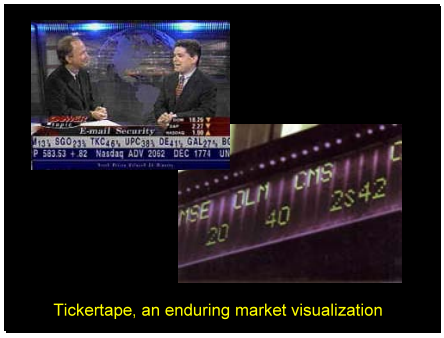
Lots more data for people to analyze. An early HFT researcher literally lost in the data.

Slide 26

The ticker tape became the symbol of the market, like rooms full of screens today.

Here we see a still from a movie the NYSE made to have pre-Muppets Kukla, Fran & Ollie explain it to the public

Slide 27

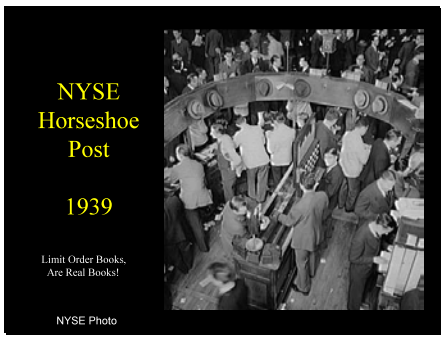


It may be pretty vintage, but the tape is an enduring market visualization that we still see today.

Slide 28

The post system for structured market data expanded. Quotes were still posted on actionable on human time scales.

Slide 29



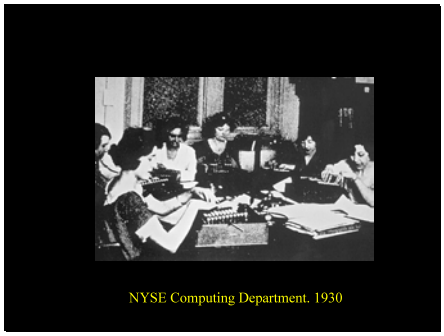
Quotes were real quotes, and Limit order books were real books, seen in this photo of an NYSE post.

Slide 30

All of this grew into the paper laden, busy NYSE floor. It got particularly onerous as the number of small trades grew rapidly, with the rise of retail investing.

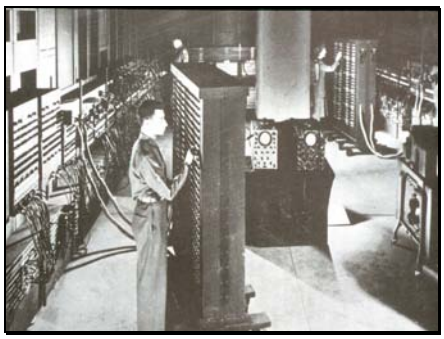
The volume of paper became overwhelming, and slippery.

Slide 31



At this point, “Computer” at the NYSE was a job title, not a machine. Here we see six NYSE computers in the thirties.

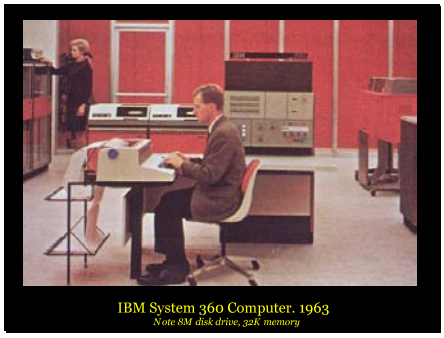
Slide 32



The earliest electronic computers were pretty useless for markets.

ENIAC needed a battalion of nerds to program it by moving wires around. It blew a tube about every 30 minutes, if you were lucky.

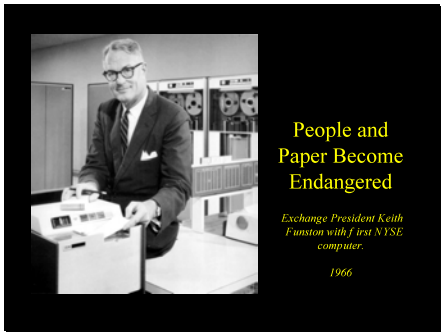
Slide 33



But by the mid 1960s, they were made of transistors, much more reliable, and could be run by civilians.

Note that all of you have more powerful machines in your phones than this million dollar baby did back then.

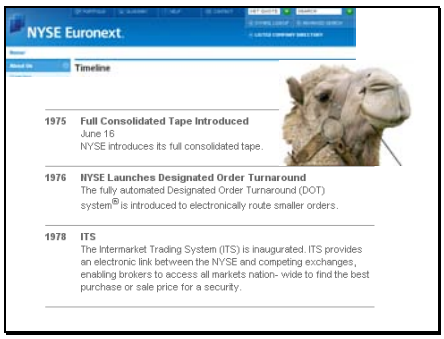
Slide 34



Even then data was getting to big for people to manage. The exchange decided they had to have a computer, or the place would fill up with paper several times a day.

They got it in 1966, shown here with Keith Funston, then president of the exchange.

Slide 35



This was the proverbial "camel's nose in the tent" for algos. At first, they were used for small trades only. But that changed rapidly. Within a few years, these new computers were used to create all of the pieces that have grown into the modern ultrafast markets we have today. Electronic feeds, electronic executions, and connected markets – operating at speeds that humans had a increasingly hard time keeping up with, and picking up an ever larger share of volume.

Slide 36

Almost all of these pictures have been from the NYSE, because it's more photogenic.

I want to give credit to NASDAQ as well. They were early innovators here, and have continued to be so. More and more trading moved to electronics. NASDAQ never had a floor, and never needed one.

Slide 37

This onslaught of big fast data and technology in markets wasn't limited to the US. We can see this pretty dramatically in another pair of before and after photos.

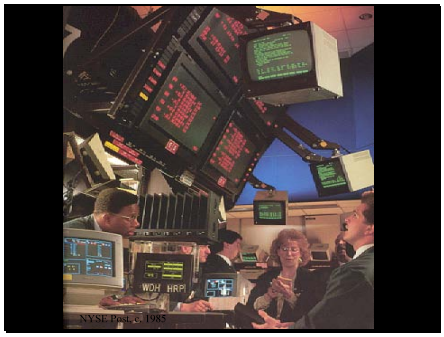
This is the LSE on the day before the introduction of electronic screen trading on October 26, 1986.

Slide 38

London Stock Exchange. Big Bang. Oct 27, 1986

The floor closed the next day.
I think the place became a night club shortly thereafter.

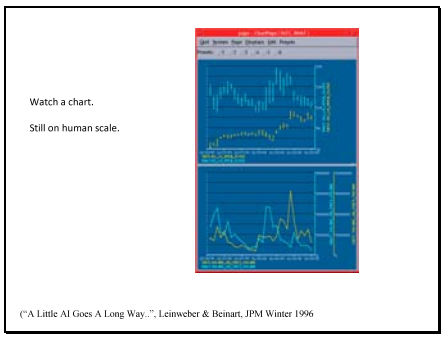
Slide 39



NYSE Post c. 1985

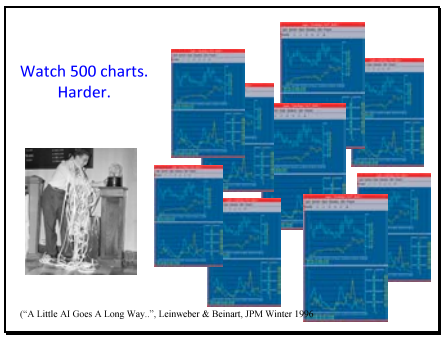
NYSE kept the people and the posts. But the posts weren't what they used to be. Computerized quotes replaced written ones, and the paper LOBs. Bigger data coming faster all the time.

Slide 40



Traders geared up to keep pace with the faster markets. They could make charts instead of reading tapes.

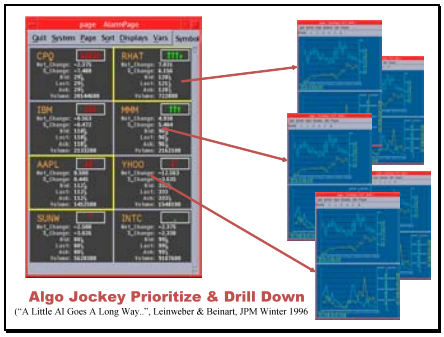
Slide 41



But as happened with ticker tape, pretty soon there was an information overload.

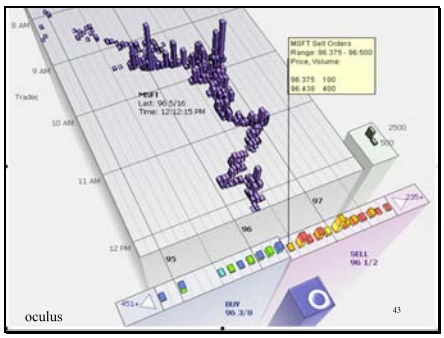
Too many charts and too little time.

Slide 42



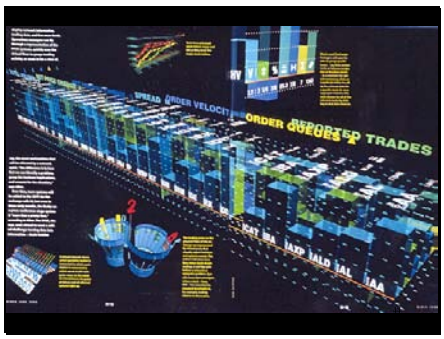
As always, there's more technology to cope with technology. This shows the Quantex system, developed by the first firm I founded, and acquired by ITG in the early 90s.
It watched the charts for you, and linked to the electronic order systems, but still on a vaguely human time scale. Things happened in seconds, people could screen orders but not for long.

Slide 43



Here's an example from Oculus that was actually used then. The details of the LOB are seen in the lower right, and the history of trades move up on the "conveyor belt" to the upper left
This was very state-of-the-art, and conveyed much more information than previous visualizations, but in today's ultra fast markets, it would just become a blur for active stocks, and there's no information about order flows and cancels, that loom so large in current trading.

Slide 44



As market data got bigger, visualizing what was going on in the market for traders got harder and more tech intensive. This now quaint example was installed on the ramp at the NYSE. It's more of a world's fair "gee whiz" display than a real trading tool/
But it was a leading indicator of where markets were heading. And they still showed the posts!

Slide 45

Map of the Market , the classic big picture of whole market visualization.

It was invented by
Marten Wattenberg, now at IBM
Many Eyes – the open system used (in a pumped up version) for NY Times Visualizations, and something you can use .

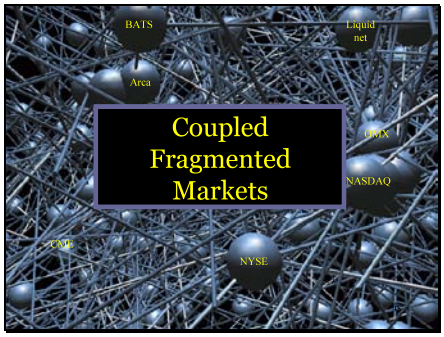
Slide 46

Posts are still there, but they're mostly computer access points for guys with handhelds, and the real action is off the floor.

Slide 47

Photogenic example of repeating patterns of technology.

Slide 48



Our examples have been about stocks trading in one or two venues. Today, it's all much more complex and interconnected.

Slide 49

Coupled Fragmented Markets

- >50 venues for US equities & Futures
- Different structures, order types, info
- Dynamic elements
  - Circuit breakers, active order types, HFT
- Vanishing specialists
  - No obligation to quote
  - HFTs step in
- Coupled markets & securities
  - Futures, ETFs
  - Larger than "primary" stock markets

49

There's much more later in the conference on this.

In addition to wired coupling of markets, there are economic structural linkages, all fast and electronic. If something slips in one place, it can reverberate everywhere else, in a hurry. Bigger faster data in more places than traders might have imagined not too long ago.

Slide 50

A fragmented, coupled venue montage – more than 50 by some counts.

Complex electronic & economic interconnections between equity, futures, options and ETF are on the list of things many people say will cause the NEXT flash crash. And are prime suspects for the thousands of reported smaller "black swan" market glitches that worry traders, and erode confidence in markets.

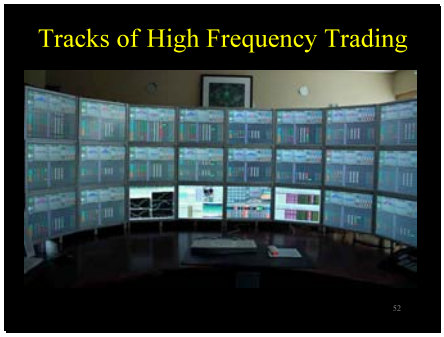
One recent paper counted almost 20,000 of these in a five year period ("Financial Black Swans Driven by Ultrafast Machine Ecology", Neil Johnson et al)

Slide 51

A popular version of the academic work, with better pictures, was in the Atlantic magazine. "Tracks of Bizarre Robot Traders". They called them "Crop Circles of the Market".

These are the ones described in the must-see Kevin Slavin TED talk on financial algorithms.
http://tinyurl.com/3pzct9p

Slide 52

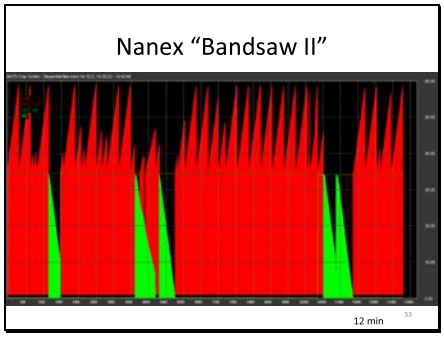


Most buy-side firms are technologically out gunned in these markets. The HFT firms that we've worked with at the Lab are "white hat" kinds of traders. They are looking to provide liquidity and play fair.

Some are "black hats" and are willing to game the market for what seem to many to be unfair advantages.

But this is where most of the liquidity is found It's a real challenge for federal regulators, with even less technology than they buy-side to keep an eye on this. Some ideas, like a transaction tax, are blunt instruments in this game, and might easily cause more problems than they solve.

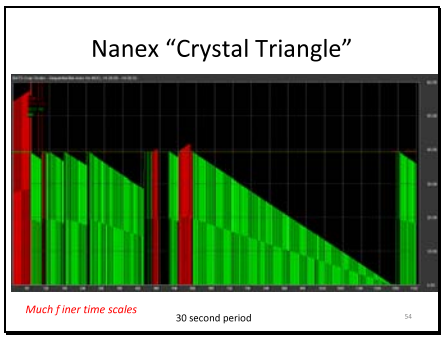
Slide 53



Tens of thousands of "quotes" in minutes.

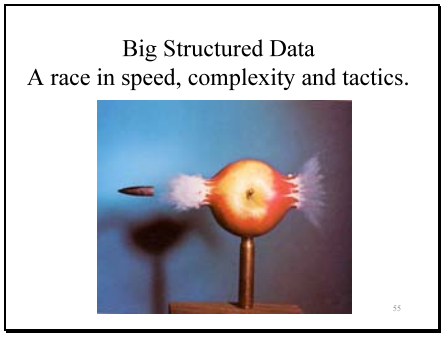
Don't blink – you'll miss them. Are they real, or scoping you out? "When is a quote a quote?"

Slide 54



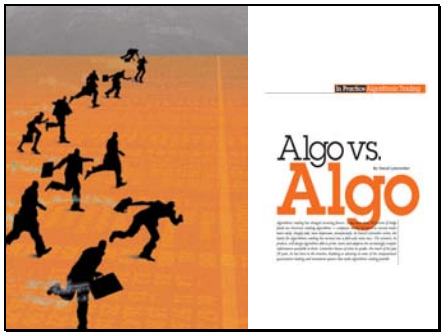
An even shorter period – 30 seconds order flow activity – in one stock! The order flow events in these charts happen so fast, they visually blend together.
It creates a “race to the bottom” in latency, or to the top in speed, and increases the likelihood of inadvertent system ‘glitches” that undermine our markets.

Slide 55



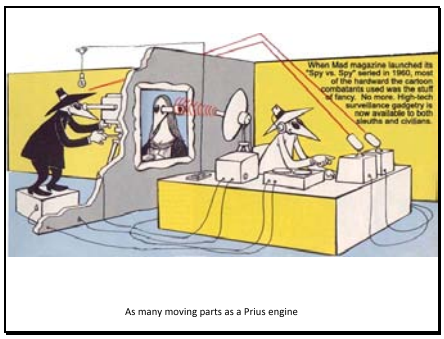
There’s a race on, as the buy-side, and regulators try to keep up with these market sprinters. And market centers, made of real computers, with queues, delays and processing overloads can exhibit some odd behavior that we never saw when people were running the show.

Slide 56



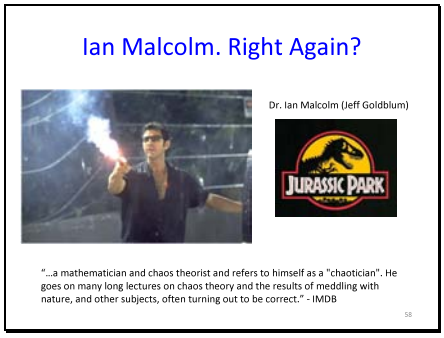
Algo trading systems have begun to look like an arms race

Slide 57



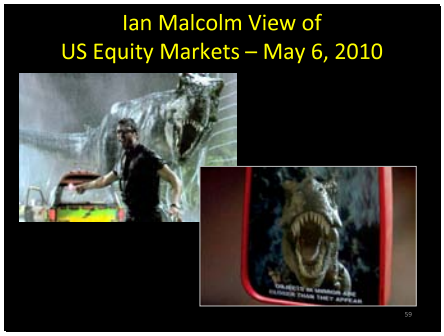
It gets too complicated to figure out what's going on. Like comparing a Prius engine to an old chevy. There are more moving parts than we can keep track of
This is where computers are pushing out humans on the left side of that Skill Chart we saw at the beginnning

Slide 58



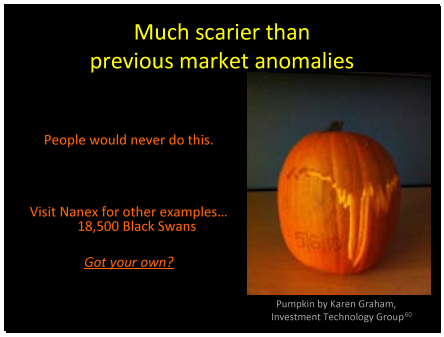
With 50+ fragments, plus all those coupled markets, and all those computers with delays, glitches, queues and maybe cyber attacks – Lots of people say anything can happen when things get this complicated, this fast. It reminds me of the warnings from Ian Malcom – the Jeff Blum character from Jurassic Park.

Slide 59



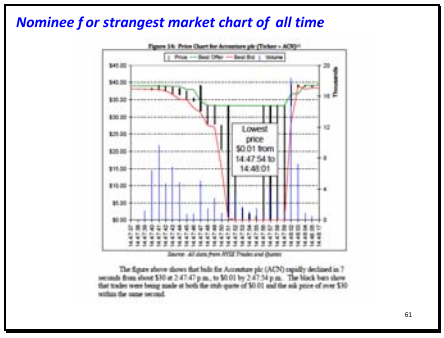
No one got swallowed whole during the flash crash, but it's the poster child event for the kind of things that can go wrong, even without malicious intent.

Slide 60



All evidence is that the flash crash was a technological accident, but it's scary to think what could happen if cyber-attackers decided to do this on purpose.

Slide 61

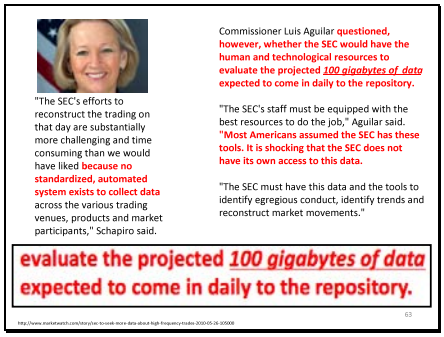


This has be one of the wackiest charts in market history ACN drops from $40 to a penny and back in seconds. People would just not do this

Slide 62

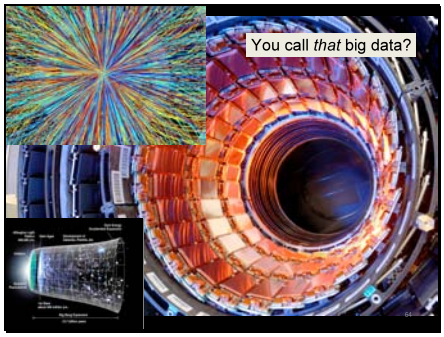
There was a lot of speculation on the cause of the flash crash – fat fingers, manipulations, evil computers gone amok, a probe for a future economic cyber-attack. The best explanations I've heard are more in the zone of a tech accident, but that doesn't mean that these things can't happen. And if we want our markets to be the safe, stable and secure places investors expect them to be, we need to think about all ways things can go awry.

Slide 63



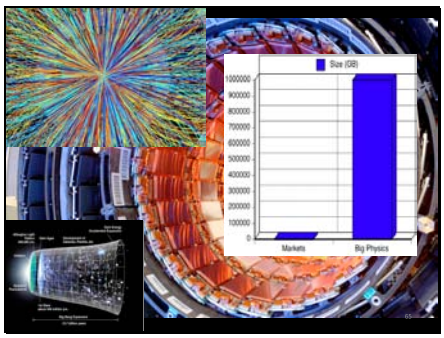
Gobsmacked! Trying to figure out what hit us pointed out that regulators were overwhelmed in their ability to see what was going on in today's big data markets.

Slide 64



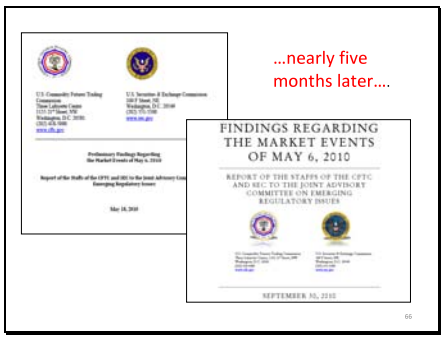
They supercomputer guys said "100 Gig in a day! – You call that big data. Ha – we eat petabytes in seconds – a million times more!" The LHC at CERN, all the earth-looking sensor and all the astronomy gear each produce much more data than markets. And it was once as much of a mess.

Slide 65



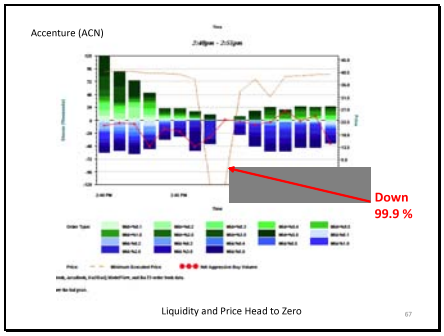
But it was big, and messy for the federal market monitors & regulators.

Slide 66



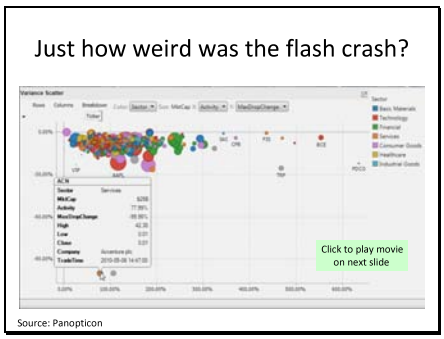
The SEC/CFTC issued a simple "what happened" report in 2 weeks. Almost 5 months later, in a report notably not labeled "FINAL", they suggested that a trade of 1.3% ADV in the e-minis by a "fundamental trader" (Waddell & Reed) caused the commotion.
That kind of trade happens pretty much every day. What made this one different? No answers on that. No one talks about Waddell & Reed anymore.
They did include charts showing the LOB details and depth of market in many stocks and ETFs. Here's one

Slide 67



This is from the second SEC/CFTC report. Which shows what happened, here for poster child ACN. This wasn't due to Waddell & Reed. Best theory I've seen is that some key market systems reached capacity. All computer systems have a capacity.
There were time stamps on quotes going out that were put on when they went out, not when they came in. By looking at the data, you saw invalid information – buy high/sell low. The alarm bells rang at the HFTs, and they all made a beeline for the door.
In other complex networks, this sort of engineering operational data is a key to stability. Then, and today, it wasn't even collected for the aggregate market. I think this is a major omission, one of the things we suggest strongly.

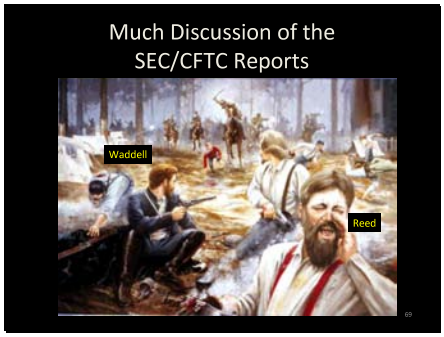
Slide 68



The two charts I've shown were about one stock, Accenture. This is a visualization from a firm called Panopticon that shows the whole market.
This is really strange – notice how volume spikes for most, but not all names, and in particular, how some dropped like rocks, others dropped a little bit, and some actually went up.

Some of Panopticon's trading clients use more sophisticated variations on this idea – looking at multiple market fragments, order flow data and the like. For them, it's sort of an anti-gaming radar, and is a tool for human algo jockeys, looking to avoid danger, and for improving & tuning algos and how they use them

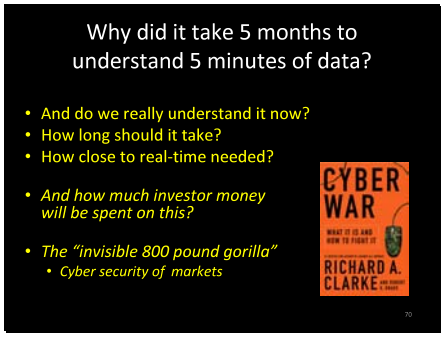
Slide 69



You don't hear much talk about Waddell & Reed causing the flash crash with a 1.3% ADV trade in the e-minis. Those happen all the time, without untoward effects.

The report raised as many questions as it answered.

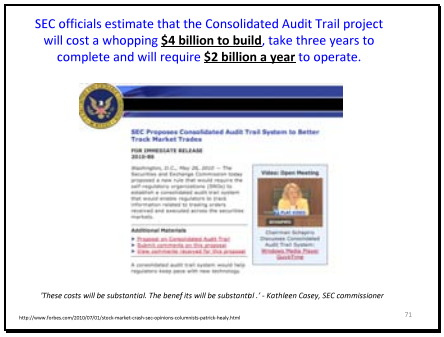
Slide 70



A key question was how the regulators could deal with future events, caused by malicious participants, or just by too much technology, moving too fast.

This is where the proposal for a Consolidated Audit Trail System came from.

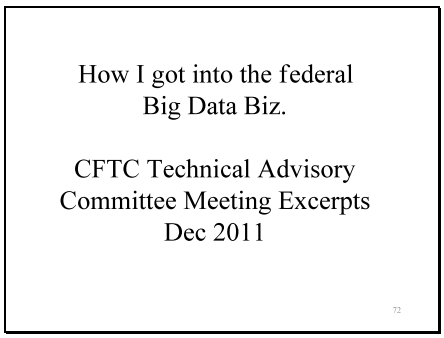
Slide 71



It wasn't going to be cheap. Estimates were $4B to get started, and $2B a year to keep it going.

And there's a big missing piece in most discussions of this – the cyber security of markets, and the same data needed for regulation is needed for cyber defense and analysis.

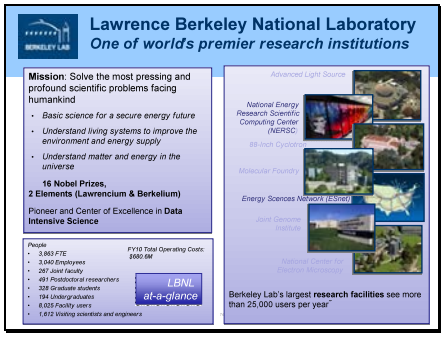
Slide 72



Slide 73

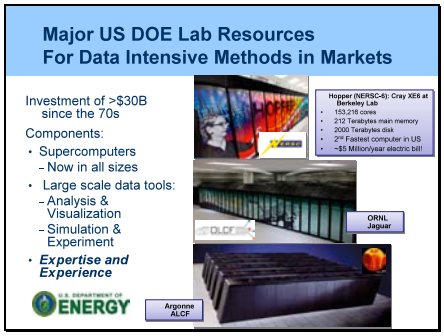
It turns out that while the left hand of the federal government (market regulators) was ready to reach deep into investor wallets to deal with this, the right hand (big data intensive science) has been dealing with problems on this scale, and thousands of times larger for decades, the result of many billions of federal R&D

Slide 74



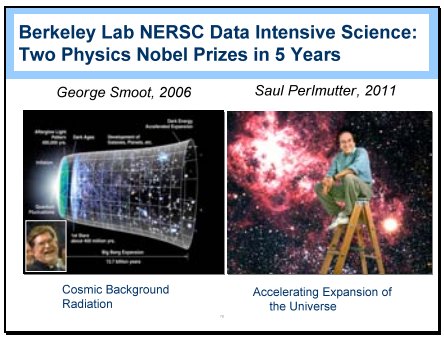
Going back to the 1930s and 1940s, LBL has been a center for big science. The Manhattan project was run from there. It's now fully unclassified (that work moved elsewhere) and runs the largest computers and fastest networks in the world.

Slide 75



Market data is big and fast, but scientific data is bigger and faster, thousands of times bigger and faster. These machines don't even break a sweat at the sizes of data needed to monitor & analyze markets. Hopper, the machine we used for our prototypes, has 153,216 really fast processors, more storage than the markets could fill up in a decade, and costs over $5 million a year in electric bills to operate.

Slide 76



Two of our "Big Data" scientists have won the Nobel Prize recently for research on data many times larger than what we see in markets, or are likely to see anytime soon.

Slide 77

By the way, as bonus for winning the Nobel Prize, you also get to make guest appearances on "The Big Bang Theory"

Slide 78

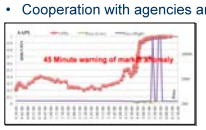
It's not just about hardware and multi-million dollar electric bills. It about using the methods and tools developed for big data science to understand big data markets. A key idea is to pose the key questions first.

Avoid "Code, ready, aim". This approach has been a huge success in astronomy, earth sciences, biology and physics. Big science used to be the same incompatible morass that we see in markets. No more

Slide 79

Data Intensive Science Financial Prototypes at Berkeley Lab

- Replicate portions of Flash Crash Analysis
- Extend to test improved early warning "Soft Circuit Breaker Methods"
- **Microstructure, network data, cyber-security**
- Cooperation with agencies and market participants

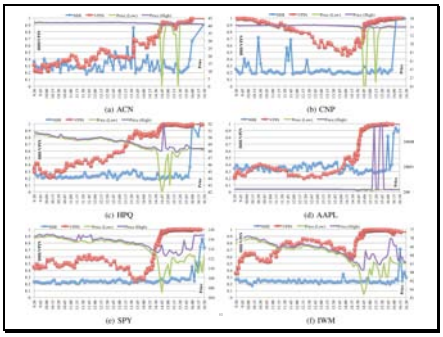


REUTERS

Post 'flash crash' monitoring emerges at Berkeley
http://reut.rs/tsIRwu

"Top Ten" SSRN Paper: http://bit.ly/w4eHxY

Slide 80



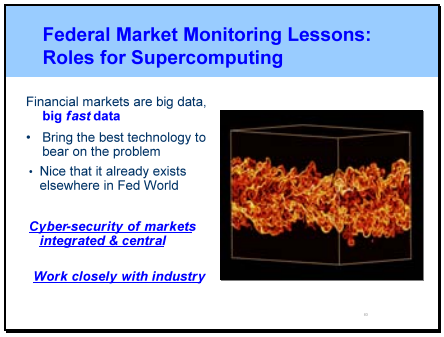
Slide 81

A promising example of a "yellow flag" warning of danger.

Algo jockeys take note:
You can build these at home.

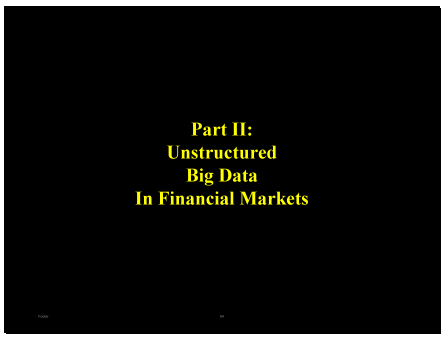
Slide 82

Slide 83



A key element of our work has been cooperation with industry as research affiliates.
We have some, we need more.
Uncle Sam wants you.

Slide 84

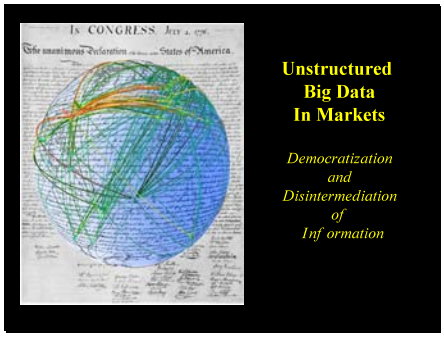


This is the half of the talk that deals with “Big Data” in the usual sense used in the tech press – weakly structured information from textual, web, images and the rest. It includes a discussion of “text analytics” applied to main stream news (Thomson Reuters News Analytics. News vendors are among the most motivated firms to get this right, so they can add informational value over above “commodity news”.

investors and traders used to be distinguished in part by the time

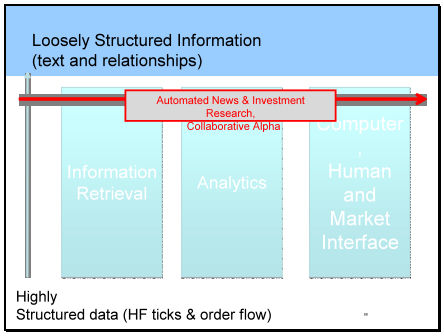
scales they dealt with. As more news and research content appear in “Big Data” flows, traders and investors are increasing working from a shared information base. This section includes examples of using analytic methods in this context as well. Interactive visualizations allow humans to apply and amplify their skills effectively here.

Slide 85



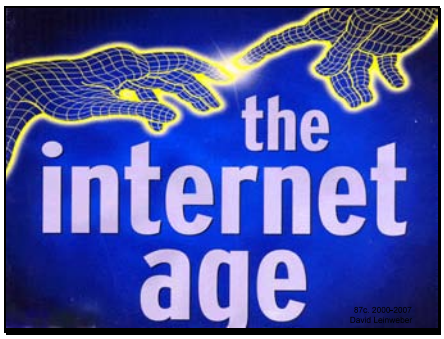
Now let’s move on the to the second flavor of Big Data – what we hear about all the time in the tech and popular press. Unstructured information.

Slide 86



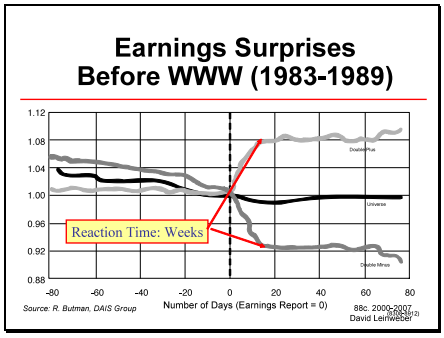
Unlike market data, this information can be almost completely unstructured and much less amenable to machine Information Retrieval and Extraction. Some legal documents are filed as scans – really just pictures of data, and it takes humans and OCR to figure out what’s inside. Since this is a way to hide bad news, the contents are often surprisingly relevant to investment decision making.

Slide 87



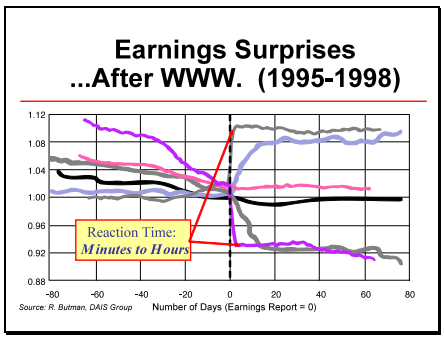
Information is reflected in prices much faster.

Slide 88



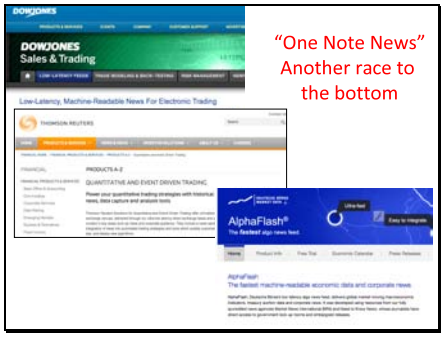
An *earnings surprise is a very pure form of market - moving information* event. You know exactly when it happens — when the company announces its earnings — and there is no ambiguity about interpretation. Positive surprises result in positive returns. The chart shows PRE-WEB era event studies illustrating the market response to earnings surprises in a period before the PC browsers. People could read it the next day in the *Wall Street Journal or weekly in Barron's and still have alpha to spare, over up to 10 days*
Notice that prior to the event, on the left side of the chart, we see information leakage before the announcement, for negative surprises.

Slide 89



This picture changes dramatically and quickly when PC browsers appear, (Mosaic and Netscape), seen in the chart shown in Figure 4.6 , which shows the shift in market response from weeks to minutes on the right, and the increased information leakage on the left.

Slide 90



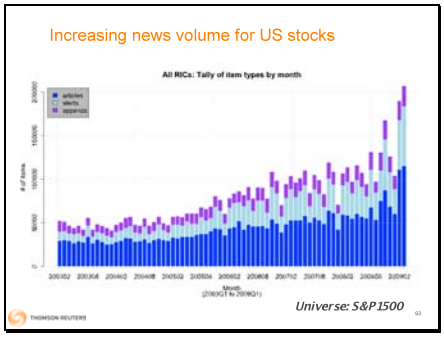
Slide 91

In addition to "hidden in the scan" material, there is a wealth of information that is loosely structured, some is in HTML or text feeds, some is "lightly structured" using various flavors of XML or XBRL.

Slide 92

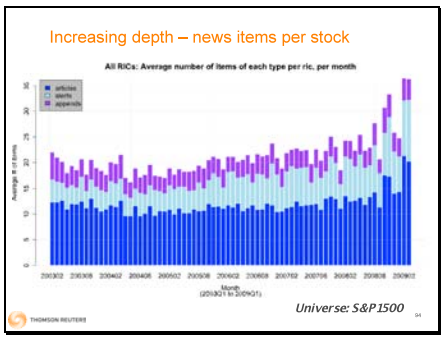
Main stream media reporters face the same issues – How to work well with machines to make sense of the unstructured raw material that becomes news stories
Thomson Reuters built a huge data center in Mumbai, India, that is an impressive example of this. Automated retrieval agents gather candidate content from the vast world of big data, it scrolls by the reporters on the left pane of their screens. They scan it for important items, and when they see one that merits becoming a story, they move it to a middle pane, write a headline, tag with company name, industry, topic and the like, and then shoot it over to the right hand side, which shows what goes out on the feed. This is the "New News". These "Intelligence Amplifiers" were deployed in London & NY as well, in 2004-2006, and we can see the dramatic changes in the volume, depth and breadth of news

Slide 93



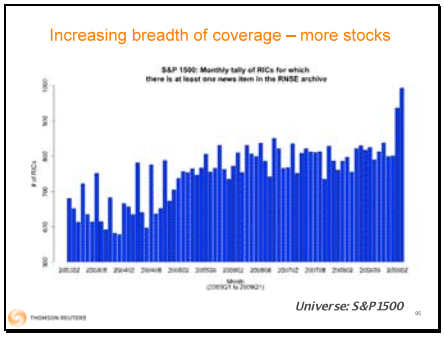
Just by counting stories about US stocks in the S&P1500 we see ~400% more news.

Slide 94



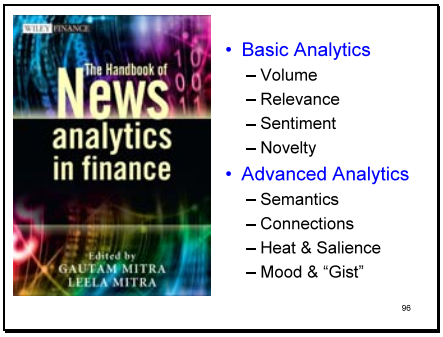
Depth of coverage, measured by the number of items per stock shows over a 100% increase since the systems were deployed.

Slide 95



Breadth measures the number of different firms covered on the news feed. We see nearly a 300% increase here - primarily in names outside the usual top few hundred. Particularly important, as we'll see, since the market impounds information in these names much more slowly than in the usual Googles, MSFTs, Exxons etc.
In fact, it turns out that this information is impounded in prices on longer human time scales – days, weeks and even months, so traders can become valuable alpha contributors to the portfolio management process.

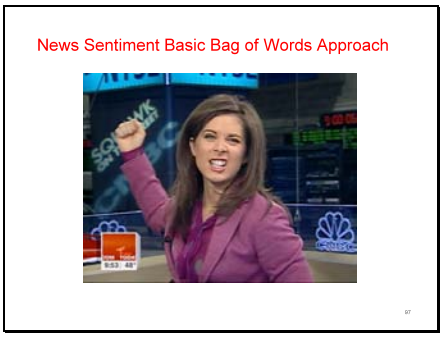
Slide 96



News Analytics are tools that can be used by themselves, for example in "one note" news, but for more complex "bigger data" with subtleties that machines miss it helps people working with machines do better than machines alone

The recent Jeopardy contest had some examples of this, one being Watson's error in thinking Toronto was a US city. It seemed to many observers that a person + Watson team would have beat just Watson in the famous contest, particularly if there were only two teams, without having the humans duke it out with each other, as well as with Watson.

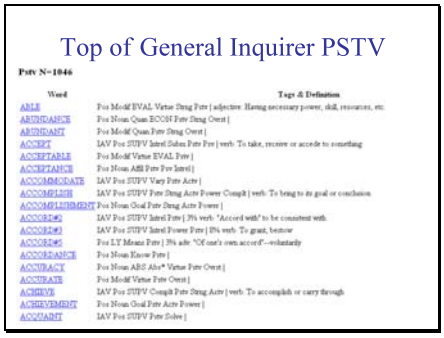
Slide 97



Here's one way to measure the sentiment of news – and note that when news is delivered like this, verbally, from people, machines are very bad at figuring out sentiment. When it shows up or is converted to text, current off-shelf-systems use variations on the "bag of words" approach. The details on the innards of those tend to be proprietary, but we can get a feel for what's going on in them by looking at a venerable academic example.

These are from research tool widely used in language research, the General Inquirer, developed over 30 years with funding from the NSF and Australian Research Council. It does much more than positive/negative sentiment, but let's take a look inside at how it does that.

Slide 98



Top of General Inquirer PSTV

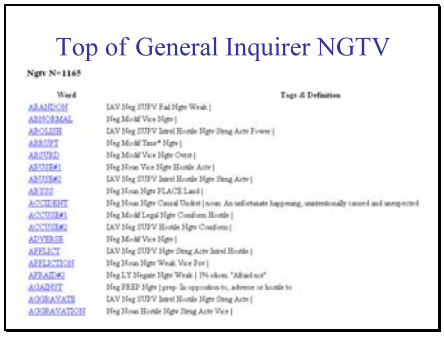
Pstv N=1046

Word	Tags & Definition
ABLE	Pos Modif EVAL Virtue Strng Ptrn | adjective: Having necessary power, skill, resources, etc.
ABUNDANCE	Pos Noun Quan ECON Ptrn Strng Overt |
ABUNDANT	Pos Modif Quan Ptrn Strng Overt |
ACCEPT	IAV Pos STPV Intrnl Solve Ptrn Ptrn | verb: To take, receive or accede to something
ACCEPTABLE	Pos Modif Virtue EVAL Ptrn |
ACCEPTANCE	Pos Noun Affil Ptrn Ptrn Intrnl |
ACCOMMODATE	IAV Pos STPV Vary Ptrn Actv |
ACCOMPLISH	IAV Pos STPV Ptrn Strng Actv Power Complt | verb: To bring to its goal or conclusion
ACCOMPLISHMENT	Pos Noun Goal Ptrn Strng Actv Power |
ACCORD#0	IAV Pos STPV Intrnl Ptrn | 3% verb: "Accord with" to be consistent with
ACCORD#3	IAV Pos STPV Intrnl Power Ptrn | 8% verb: To grant, bestow
ACCORD#5	Pos LY Means Ptrn | 3% adv: "Of one's own accord"--voluntarily
ACCORDANCE	Pos Noun Know Ptrn |
ACCURACY	Pos Noun ABS Abs* Virtue Ptrn Overt |
ACCURATE	Pos Modif Virtue Ptrn Overt |
ACHIEVE	IAV Pos STPV Complt Ptrn Strng Actv | verb: To accomplish or carry through
ACHIEVEMENT	Pos Noun Goal Ptrn Actv Power |
ACQUAINT	IAV Pos STPV Ptrn Solve |

On the other hand, here's the top of the bag of 1165 NGTV words in General Inquirer. Think how much the CEO doesn't want to say "abyss", "adverse" or "aggravation" in the MD&A. Often, context is important – an area humans are good at that "bag of words" analytics can miss. These things usually produce sensible results, but when you drill down to the details, the sense of the signal can change direction.

Here is the top of a list of 1046 words in General Inquirer's PSTV bag. A CEO would likely be talking about something good if the words "abundant, accomplishment and achievement" showed up early and often in an earnings call, speech or SEC management discussion section.

Slide 99

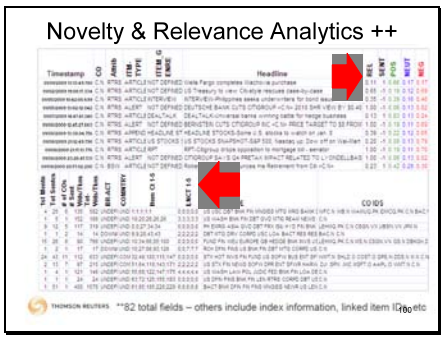


Top of General Inquirer NGTV

Ngtv N=1165

Word	Tags & Definition
ABANDON	IAV Neg STPV Fail Ngtv Weak |
ABNORMAL	Neg Modif Vice Ngtv |
ABOLISH	IAV Neg STPV Intrnl Hostile Ngtv Strng Actv Power |
ABRUPT	Neg Modif Time* Ngtv |
ABUSED	Neg Modif Vice Ngtv Overt |
ABUSE#1	Neg Noun Vice Ngtv Hostile Actv |
ABUSE#2	IAV Neg STPV Intrnl Hostile Ngtv Strng Actv |
ABYSS	Neg Noun Ngtv PLACE Land |
ACCIDENT	Neg Noun Ngtv Causal Undrst | noun: An unfortunate happening, unintentionally caused and unexpected
ACCUSE#1	Neg Modif Legal Ngtv Confirm Hostile |
ACCUSE#2	IAV Neg STPV Hostile Ngtv Confirm |
ADVERSE	Neg Modif Vice Ngtv |
AFFLICT	IAV Neg STPV Ngtv Strng Actv Intrnl Hostile |
AFFLICTION	Neg Noun Ngtv Weak Vice Ptrn |
AFRAID#0	Neg LY Negate Ngtv Weak | 1% idiom: "Afraid not"
AGAINST	Neg FREP Ngtv | prep: In opposition to, adverse or hostile to
AGGRAVATE	IAV Neg STPV Intrnl Hostile Ngtv Strng Actv |
AGGRAVATION	Neg Noun Hostile Ngtv Strng Actv Vice |

Slide 100

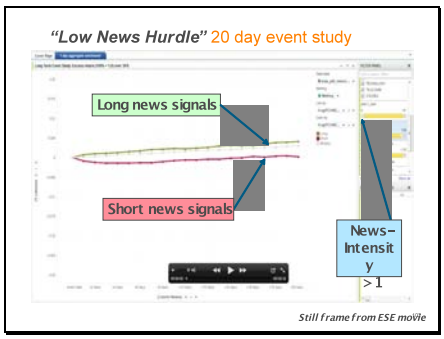


Other news analytics show relevance novelty. These Thomson Reuters examples show both. When a company is mentioned in the headline, first line and there are no other companies in a story, it's likely highly relevant to that company. If there are a dozen firms, it's about an industry or region, and humans will be better at figuring out the investment value.
Novelty is measured by link counts which show how many similar stories have appeared in the previous hour, 12 hours, day, week and month. There are a total of 82 analytics, many quite subtle, and amenable to human interpretation.

Slide 101

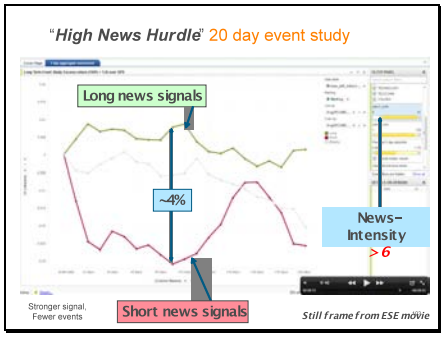
There's a paper co-authored with my colleague Jake Sisk, now at Thomson Reuters that shows the results of applying human design skills in using news analytics for portfolio management. We didn't second guess each trade, but an examination showed that we would have done better if we did.

Slide 102



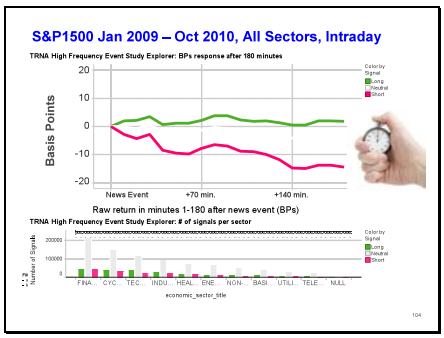
Here's a simple event study using sentiment and volume for all news – one or more linked stories.
Spread of a hundred bps or so. Note that this Is over 20 days.
When you drill down to the stories, you see that humans are still better than the sentiment analytics,

Slide 103



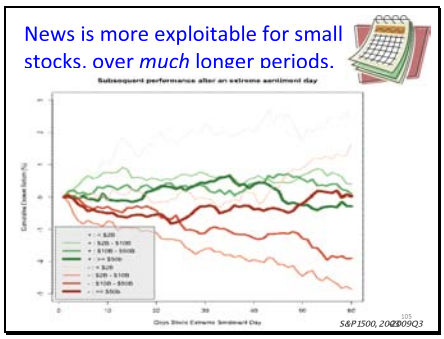
Turn up the threshold for number of stories, and we see some significant alpha, again over 20 trading days. Movies showing the Event Study Explorer, the interactive tool used to design these filters are found on the "Nerds on Wall Street" blog.

Slide 104



Here's an intraday look, on 3 hour time scale, relevant to trading
The spread is roughly 20 bps – not world changing for overall portfolio performance, but large enough to make a difference in trading costs.

Slide 105

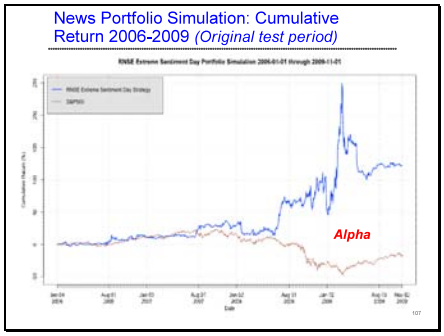


This is one of the most important findings. It shows responses to filtered news by capitalization class, Over SIXTY days. Plenty of time to accumulate institutional size positions.
Alpha of 2-3% over 60 days, in stocks with caps under $10B can really make a difference in portfolio alpha.

Slide 106

We used news analytics, event studies and drilling down to details to see where news had predictive value. Just like quants do with conventional signals, we turned the knobs, to see what worked in different subsets of the market. The universe was the S&P1500.

Slide 107

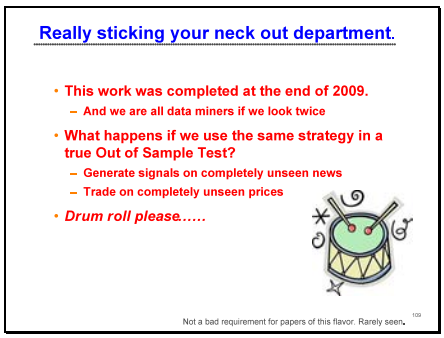


Lo and behold – we found quite a bit of alpha. The results from the event studies translated well to portfolio performance, but only in the after the modern "New News" systems kicked in, and it increased as they saw wider deployment. Putting it all together in a portfolio simulation showed results stronger than we expected. "Wait wait" you say, "these are data mined results". Alas the market has only one past, so if you look twice, you are a data miner. We tried to hold back subsets and the like, but we did look more than twice, But stay tuned for some true out-of-sample results

Slide 108

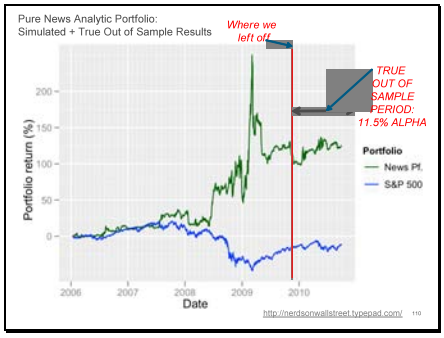
A key feature of these results is that even with a touch of data mining, there was no alpha to be had until the new TR machinery for turning big unstructured data into news kicked in, and that it got stronger as that machinery was more widely deployed. A quant research group at Deutschebank research, working independently, got almost identical results.

Slide 109



Now here is a true out of sample test. After 2009, we turned our research to other areas and put the news analytic portfolio manager in the deep freeze for ten months. Our paper had been accepted to the quarterly Journal of Portfolio Management, but with the backlog for publication, we were able to pull the news analytic simulator (with transaction costs included, of course) out of the deep freeze and run it again, on a period where neither we or the system had seen the news data, or the stock price history.

Slide 110

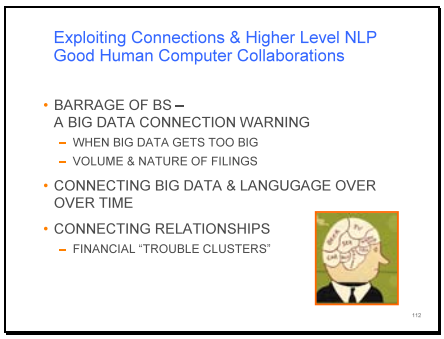


This was the gratifying result - another 11.5% alpha, on previously unseen data.
The details are in Journal of Portfolio Management paper. We suggested to the editor that this "wait and test on real out sample data" be standard practice for this sort of work – they said it was a good idea, but they would have rename the journal to be the Pamphlet of Portfolio Management.

Slide 111

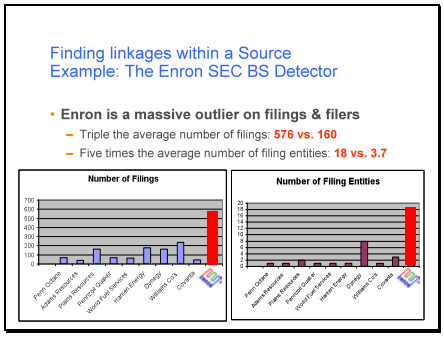
The news-driven portfolio is an example of what can be done with the "state of the practice" systems you can buy now. Let's take a look at more complex ideas based on connections in unstructured big data.

Slide 112



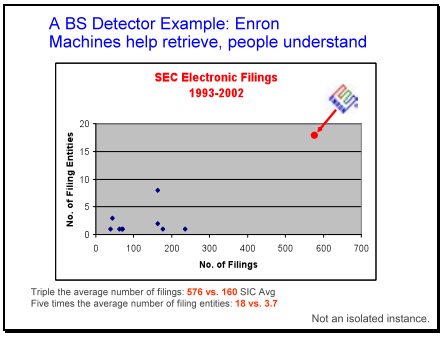
We'll take a quick look at three examples of more elaborate strategies, driven by analyzing connections, between firms and related entities, and changes what firms say over time.

Slide 113



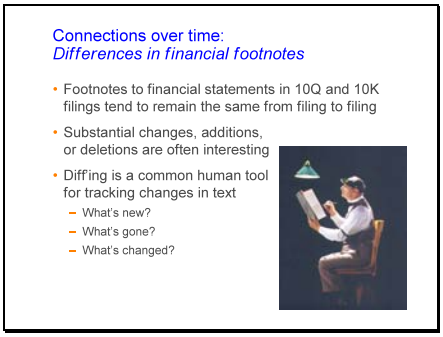
An investor in Pasadena suggested this idea. He looked in the 10Qs and 10Ks at the SEC, manually found the "related entities" named in those filings, and used big changes in the count of total filings and filers as indications of trouble, He claimed great investment results. This turns our to be effective. The most dramatic example is Enron. Doing this by sector makes sense, since industries have different norms. From 1993-2002 Enron was a huge outlier on both scales. Three times the number of total filings across five times the number of filing entities.

Slide 114



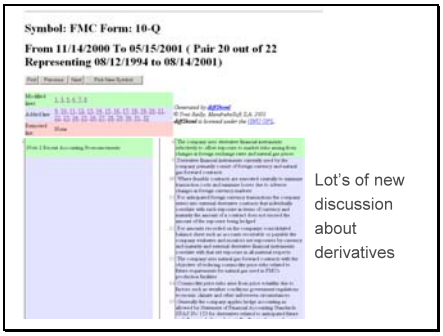
Combining the two previous charts on one shows what a massive outlier Enron is on both scales, and we know how that turned out. Broader research showed this was not an isolated example, but a bit of human judgment helps dramatically in finding exceptions to the pattern.

Slide 115



Here's another SEC based example. Financial footnotes in 10Q and 10K filings tend to be the same, or nearly so, from filing to filing over time. "Diff'ing" is tool for comparing big text data – it shows what's new, what's gone, and what's changed.

Slide 116



Here's an example across two 10Qs that shows a large amount of additional verbiage in the financial footnotes. Sometimes this could be innocent, but here it raised flags about the risks to the firm from increased use of derivatives.

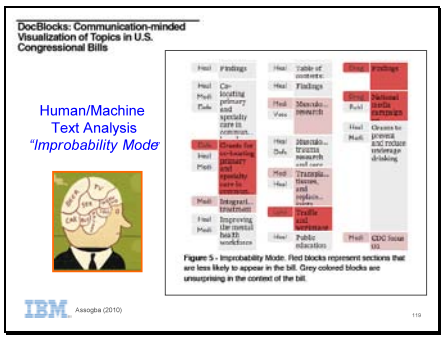
Slide 117

The stock dropped almost 40%.

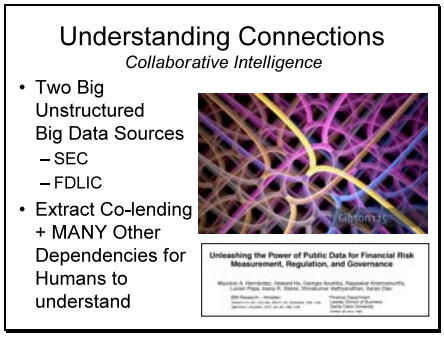
Slide 118

Turning these diff into scores is easy – just add up the amounts of text that's added, deleted or revised. Sometimes the result is a jump in price, more often, along the "barrage of BS" line, it's a drop.
This idea was used in a combined machine/human portfolio management process, where the Information Retrieval diff'ing machinery alerted humans to take a closer look. It worked pretty well, which is why these are older examples.
In particular, it took a while for the market to digest the more complex information, so institutional size positions could be taken or exited at reasonable cost.

Slide 119

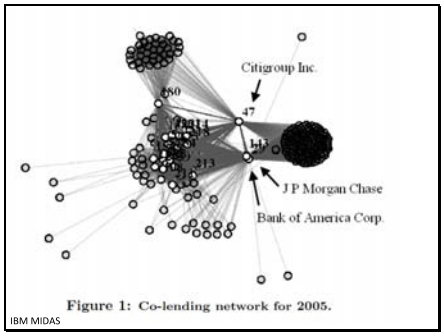


Slide 120



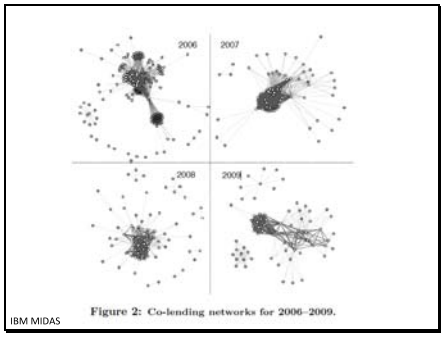
Here's a recent example, from IBM research, that uses two unstructured data sources to look at the financial meltdown of 2008. One is the SEC, the other is the FDLIC.

Slide 121



IBM's researchers combined the information from both sources to create co-lending networks across financial firms like this.

Slide 122



As the crisis evolved, we see a remarkable difference in these relationships in the financial sector. IBM was showing off their Information Retrieval tools on unstructured data, and they didn't show the financial results. But I'm told this is now in use for both portfolio and risk management. This is more complex than saying a news story is good or bad news. It's an example, like the SEC diffing, of where the machine finds smoke, and the human decides if there's a fire.

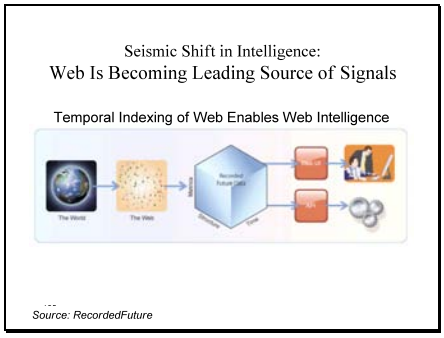
Slide 123

There are more than a few interesting technologies that are useful in helping humans understand the implications of big data. Often, this starts out in the intelligence community.
The CIA's venture capital firm, In-Q-Tel has a portfolio that includes some intriguing examples.

Slide 124

One has the fabulous name, "Recorded Future", and they are explicitly moving from the intelligence world to finance.

Slide 125

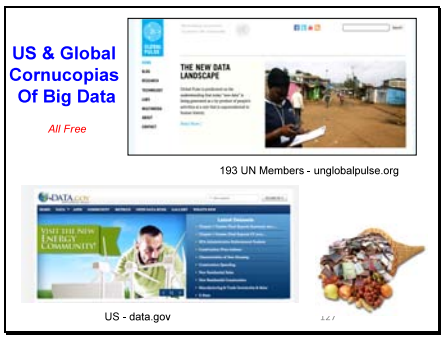


There are a few demonstration examples on their website. All are examples of human/machine collaboration in understanding big data.

Slide 126

Other countries which invest heavily in intelligence technology also have firms that are moving from intelligence to investing. Here's an Israeli example, with another great name: "Digital Trowel"

Slide 127



There are tens of thousands of big data sources out there to point these systems at. The data has varying degrees of structure, and many with testable histories. Data.gov has a rich and growing set of information about the US that is directly relevant to sectors and firms. Global investors will find a similar, and growing set of sources on all 192 other UN member countries at Unglobalpulse.org. They cover all industries, from health care to natural resources to energy, defense, and of course finance. All free for the taking.

Slide 128

Slide 129

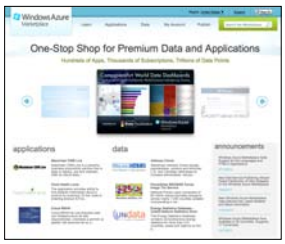
Infochimps:
Best Name

Slide 130

Huge
Related
Success

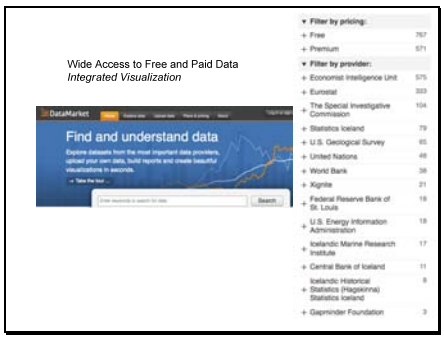
Slide 131

Widest Range of
Built-in Financial
Data

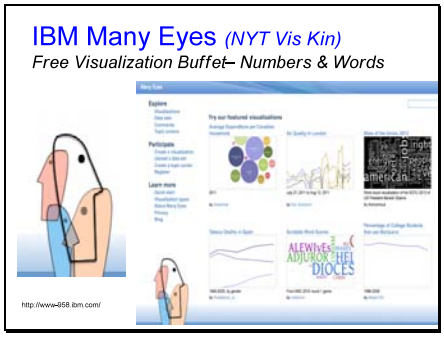


https://datamarket.azure.com/browse/Data?Category=finance

Slide 132



Slide 133



You don't an army of programmers to sample all this data, and try it out. The IBM site, Many Eyes, for some reason hidden at http://www-958.ibm.com/ is a large collection of free sample tools that can be used for visualizing relationships in both structured and unstructured data. Many people have shown their own examples, across many areas of application, including financial markets.

Slide 134

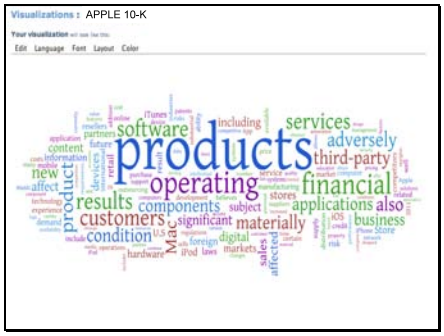
Slide 135

For unstructured text, Many Eyes has tools that do a credible job at language analysis of whatever you point them at, like financial footnotes and legal documents, that tend to run long, and get boring. The clouds have been described as a way for people to get the gist of documents, or spot smoke that might show a fire. A little cut and paste between the SEC and Many Eyes is an interesting way to spend an evening

Slide 136

Tag cloud for Kodak SEC 10-K 9/24/11 Part 1 Management Discussion. Data is on Many Eyes, make your own.

Slide 137



Tag cloud for Apple SEC 10-K 12/31/11 Part 1 Management Discussion. Data is on Many Eyes, make your own.

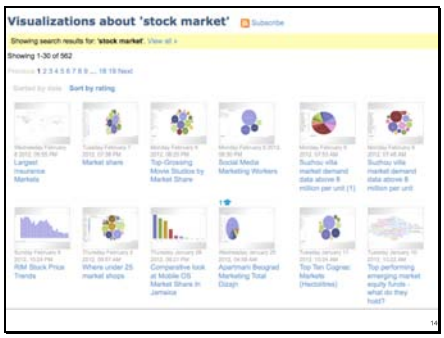
Slide 138

A ManyEyes Word Tree example, for “risk” in the Apple 10-K

Slide 139

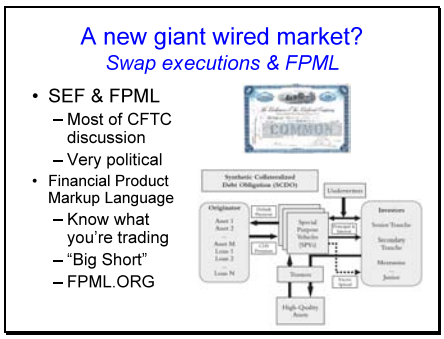
And a 2 word (di-gram) visualization for Kodak’s 10-K

Slide 140



You can tweak the examples we just saw on Many Eyes. Plus, there are over 500 others relating to “stock market”. If you can use excel and a web browser, it’s not hard to make some of your own. Mix in your own data, open a beer and fire away. If you don’t come up with a few ideas on how to use this, have another beer.

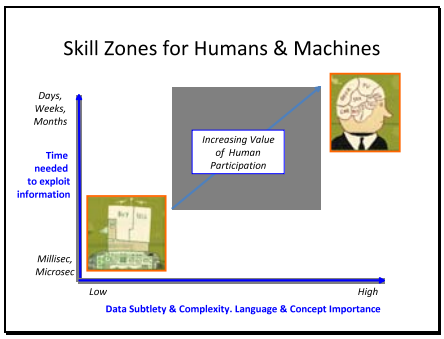
Slide 141



We've been focused on stocks, and we're out of time, but in both structured trading data and less structured security analysis there may be big changes coming. At the CFTC Technical committee meeting in December most of the agenda dealt with new electronic markets for complex swap securities – including those we now call "toxic assets". Where there are electronic markets, algos are sure to follow. And unlike the stock market, where there are only a few variations on a theme of what the securities are, and it fits on the back of the certificate, these securities come in thousands of varieties, and the documents need a wheel barrow to move down the hall. FPML, financial products markup language is a technology to bring some order from this chaos.
You can think of it as a move toward automation of the incredibly labor intensive process described in Michael Lewis' book "The Big Short".

All of this comes from Dodd-Frank, and how it will play out is intensely political, but it's likely something will come out of it. Early adopters are likely to do well, and the underlying complexity means the machines will need human helo.
A BIG opportunity, eventually, maybe, for human machine collaboration

Slide 142



To wrap up, we've looked at the skill zones where humans, working with machines, can do better than machines alone, and in the case of fast structured data, how human experts in data intensive science can improve the stability, safety and security of ever more complex markets. I hope you come away from this with some ideas on how to expand your skills to thrive in this environment by learning to work more effectively with computers.

Slide 143

This can be taken too far. I'd advise all of you to pass on the neural implants for a while.

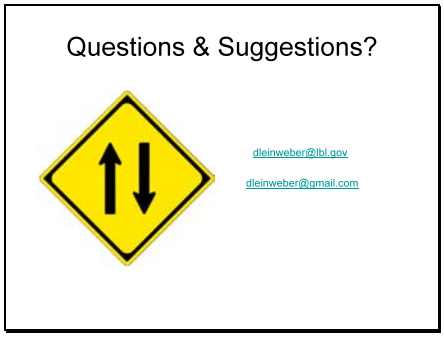
Slide 144

And, as with any new technology, we can expect a few surprises

Slide 145

But we can also expect progress.

Slide 146



I'd like to hear your questions and comments on all of this. For the federal market topics, our group at the lab seems to be getting attention from the people in Washington who do this for a living. For the others, I'm happy to hear from you at my civilian email address.