*Female:* Welcome to a podcast from the Strategic and Emerging Technology section of the American Accounting Association. Visit the section at <aaahq.org/set>. This podcast is sponsored by CaseWare-IDEA, a leading supplier of software solutions to accountants and auditors worldwide, including the IDEA data analysis software. On the web at <caseware.com>.

Interviewer: Welcome to another podcast from the Strategic and Emerging Technologies section of the American Accounting Association, and this is the first in an occasional series of discussions with recent winners of the outstanding dissertation award from the section. And the 2006 award winner was Mark Cecchini now at the University of South Carolina, so welcome Mark.

*Mark Cecchini:* Oh thank you. Thank you, Roger.

Interviewer: So when I look back at the title of your dissertation, “Quantifying the Risk of Financial Events Using Kernel Methods and Informational Retrieval”. I see that you fail on one really key dimension of this title. There is *no* colon in the title!

*Mark Cecchini:* You’re right. There is no colon.

Interviewer: That is the first rule of PhD dissertation titles -- to at least have a colon somewhere in the title!

*Mark Cecchini:* That’s true. Well, you know I go through a board that actually looks at the way that you format your dissertation. The board supposed to look at things like did you such as is there enough room between your tables and all that kind of stuff. And the guy at the board commented on my title and said he didn’t think my title was very good, so actually that has – I said, “That’s not inside of your scope of service.” So he didn’t like my title either.

Interviewer: Well – but what does it *mean* actually? Tell me what it means.

*Mark Cecchini:* Well, it is a little bit wordy. When I look back at it now, it is kind of funny now to think about it a couple years later. However, basically I think that the kernel of our ideas was to come up with things that would help people to detect financial events, and by financial events we mean things that markedly affect the value of a firm. So we tended to focus on the negative; although, this should be able to generalize the positive as well, but we’re thinking of things like bankruptcy, fraud, restatements even for reasons other than fraud, but maybe any kind of material restatement.  
  
You know, those are the kind of things that we were thinking of, trying to detect, and we’re thinking of is there a way to do this without being able to have close knowledge of the firm. So we’re thinking about things that are sort of like just being able to use publicly available data, are there certain things that we can figure out, certain ways that we can tweak the data to be able to predict that.

Interviewer: So you concentrated here on using text in reports. So how can text be used to predict financial events?

*Mark Cecchini:* Well, we have a couple different main focuses of the dissertation. Since we were focusing on trying to find ways to detect fraud and bankruptcy and stuff, we noticed that we were using quantitative variables, basically financial statement variables, for part of it. And as we were looking at it, we noticed that huge area of the financial statement wasn’t being used at all, and that was the text. Happened to be that there was an expert information retrieval guy that was on my committee, so he was real helpful with this.  
  
So he was saying, “Well, we could probably do some of the same things you do with looking at numbers, looking at text because now there’s all kinds of automatic methods to look at the text.” So that’s kind of where it started. We thought about using text in the same way people use numbers where you make functions, think of the Altmans Z for bankruptcy. It has the score that gives you about four or five ratios of financial variables, and if it comes up higher than some number, you’re safe. If it comes down lower than another, you’re not. You’re a bankruptcy risk.  
  
Well, we were thinking could you add another variable in there, some kind of text information? So that was kind of how it started and then of course it gets more complicated than that. We find out that there’s lots of research in that area, and so what we did was we actually applied some of the research that had been used in the information retrieval area in the past to this dichotomous variable problem, which was let’s see if we can figure out how to tell is this one going to be a fraud and this one a non-fraud type of thing.

Interviewer: Now, when you look at – we’ve been using the word text loosely, but how widely do you spread your net? Obviously, the face of the financial statements, but what the notes, the MD&A, you know, other classes of text?

*Mark Cecchini:* Well, that’s a good question. When I came up with this idea originally, in 2003, it was the same time maybe two or three other people were coming up with the idea ‘cause at that time nothing was happening with it, and now It has actually getting a little bit popular. So from the perspective that most people tend to use the management discussion and analysis (MD&A), and I guess the main reason is is that we see it as more of a fertile ground because It has open-ended. There’s a little bit of a chance of judgment in future-looking statements.  
  
Whereas a lot of the other material in the financial statements you feel might not be as productive, but that doesn’t mean that’s true necessarily. I think there is a lot of interesting stuff in the footnotes as well. I’ve also noticed that some people in more of the financial arena have taken a look at things in the popular press about a company. They’ve gone a different direction and looked around that way, but I mean we – focusing on the financial statement, I’d say 90 percent of the research tends to focus on the management discussion and analysis.

Interviewer: So what did you find?

*Mark Cecchini:* What we found was that you can actually determine whether a company is going to be fraudulent or non-fraudulent a certain percentage of the time by using the text of the financial statement. So, if you can actually get yourself a data set and you separate in some way, say fraud for this example, but you could use bankruptcy as well. If you have a set that’s fraudulent and a set that’s non-fraudulent you then run them through this methodology we created, which is a big part of the dissertation. The methodology basically comes up with a dictionary, and the dictionary is at of the center.  
  
The dictionary is supposed to be the terms that are most different between the two groups. So the goal is if this dictionary has a term and then It has going to have a weight for that term. Then It has going to have another term and a weight for that term. All the terms that are weighted the highest are the ones that separate fraud from non-fraud. If I was dreaming up the perfect research scenario, I would have put in MD&As of companies with fraud, and the MD&As of non-fraud companies. What’s going to come out is going to be something like, number one, offshore accounts or something to that effect, mark-to-mark accounting or something like Enron.   
  
Something to that effect, but – whereas the true answer isn’t always like that necessarily because with text a lot of what you’re doing is trial and error and there’s a lot of subjectivity involved. So your number – out of the top 100 text variables, there’s going to be – some of ‘em are going to be wonderful and relevant and some of ‘em aren’t, but you do find that a number of ‘em are relevant to what you intuitively would think of as being sort of interesting.

Interviewer: Now, do you couple the text with financial statement numbers as well? Are these working in tandem or just separately?

*Mark Cecchini:* That’s a good question. We wanted to be able to take these and put them all together. We hoped to have text and numbers alongside each other and see how that worked. Well, as of the dissertation, what we did was we tried it and were not finding all that much success putting the two together.   
  
Since then I have continued to work on what comes out of the text side of the dissertation. We’ve done this for both the fraud and the bankruptcy case. For fraud, we took a Bineach model, which is kind of a well-known model of predicting fraud, and we took the Altman Z score for bankruptcy.   
  
What was interesting was that we found that – so if you took the text by itself, we got pretty good results. If you took these quantitative things, the Altman and the Bineach, by themselves, the results were pretty poor. But then when you added Bineach to the text for fraud, the results improved upon the text, so basically what we found is that there was some kind of complementarity going on with that, and we found the same thing with bankruptcy.

Interviewer: Now, one of the tools you use are SVMs, support vector machines. How do they work and how do they support your research?

*Mark Cecchini:* Whenever I first mention support vector machines to people, they often get a little quizzical – it sounds a little strange like I’m science fiction. I think term machine really just means algorithm, and I think It has just a funny way of saying it. So yes, support vector machine is a –a machine learning method sometimes called computational intelligence or, like our old name for this group, artificial intelligence. So what it says is it just learns from examples, so the idea what makes support vector machines unusual is that It has based on statistical learning theory, so it has statistical backing.  
  
It has got two goals.– one goal is to minimize errors, which is obviously one of the main goals, but that tends to be the only goal in a lot of different machine learning methods. Its second goal is to minimize the obstructural components because it kind of goes with a little bit of the theme of Occam’s Razor, which is that if you can find the smallest set that can give you the best results that’s always better because you want to avoid over fitting. So those are some of the main reasons why it is interesting. But the reason it is called the support vector machines is that – and one of the ways that it does avoid over fitting -- is that you basically want to take two different classes and you want to separate them by a line or more generally like a hyperplane. And so when you separate the ones that are closest to the line, are the support vectors.  
  
So if you’re training on, say, 10,000 or 100,000 examples, instead of having to put all those examples and all the computational intensity it would take, you only have to look at the examples that are the support vectors. So that tends to minimize the computational complexity of the problem.

Interviewer: I want to come back in a moment and talk about how this work on text can be extended in the forward looking area in the bankruptcy prevention, but first I want to just attack one question, which is about the meaning of the text itself.

*Mark Cecchini:* Okay.

Interviewer: So at the moment the text that you’ve got is just plain text. There’s no semantic meaning to it, but once we get into the XBRL world and we can have a semantic overlay to it, so we actually have some idea what the text actually means even if It has just a block of text. The text declares that it has such and such semantic meaning. How would that impact on the work that you do?

*Mark Cecchini:* That would be pretty big in the sense that more than half the work in the dissertation is about trying to add context to the text. So text by itself can be hugely confusing. So when you start adding context then you make the text smarter and you can do more with it. The idea of actually having XBRL tags on different parts of the text of financial statements would actually enable you to do a lot more. So yes, I think that would open up a wide range of research in that area.

Interviewer: And what about some of the extensions to your work in the text area?

*Mark Cecchini:* Well, I’ll explain, by going backwards a little bit. Let me saythat what we did with the financial events prediction using text is we would create this large vector, basically, of – for simplicity we’ll call ‘em word counts basically. And each part of this vector would have a count of different words that show up in that financial statement, and each financial statement would have its own vector. The good side of that is that you cover everything. The downside is that you get these large sparse vectors where that’s lots of zeros in them.  
  
So one of the possible extensions that we’ve been working on is to try to actually cull this down into one or two dimensions. When we collapse all the dimensions in such a way that we can actually put this into regressions and do some other things that people more commonly use in accounting. So that’s one extension. Then the other thing is that we’re trying to think about is multi-class classification. Right now we look at things – we look kind of simplistically it has to be a yes or a no situation. Is that fraud? Is that not fraud? Can we do more subtle analysis by looking at things that are more maybe continuous variables by looking at multi-class classifications? So those are couple of things I’m working on now.

Interviewer: And how about publishing this material? What are the sort of challenges that you face? You know, you’re between two disciplines, which is always difficult I think.

*Mark Cecchini:* Oh yes. You hit the nail on the head. I mean as an assistant professor who’s going to be up for tenure in a couple years, I can say that that’s been a major challenge in the sense that – so if I go to a very good journal with one of my papers, what ends up happening is is that usually what happens is I get a reviewer that’s an accounting reviewer type reviewer. And I’ll get a reviewer that would be a technical type of reviewer, and both of them of course want completely different things. And I – so one of my simple answers to that I’ve said before is I come from an information systems program where we did quantitative information systems work, and I would say that people focus all on the neatness and the novelty of a model and then aren’t – and then will just say, “Okay, we’ll take any data set and show that it works.”  
  
And then I moved into accounting as an accounting professor, where in accounting people could mostly care less about the about the model. They really care about the data and details about whether you have the right data set? Do you have enough data? Have you cleaned the data right? And everything’s focused on that area. So in one sense I’d say it makes you a very good researcher because you gotta focus really hard on both ends of it/ On the other hand it makes it a little bit more complicated as far as getting things published quickly.

Interviewer: And what are the sorts of things that you’re going to work on in the next year or two?

*Mark Cecchini:* Well, I’ve a couple things I’m – one of the other parts of the dissertation – the other big part of the dissertation was actually in using – just using the basic quantitative variables of financial statements. And that came up from the fact that we’re – again, we – basically if we’re thinking about a difference in the way we did things versus the way things have been done in the past, I was thinking about how can you actually use methodology to support things like fraud prediction and bankruptcy prediction. So instead of just saying, “Okay, well what if we just tried this one variable and this variable and put those together and saw if this was good at predicting fraud?” I wanted to figure out is there a way that I could actually develop something that could do a bunch of different variables and try a bunch of different combinations.  
  
And so what we did – or vector machines. There’s a thing that’s for vector machines called a kernel, and basically what it is is It has a mathematical mapping and it allows you to take things from basic linear space and put ‘em into non-linear space. So you can actually – so the cool thing about the support vector machines is you could actually have an instant dimensional feature space, so in other words you can actually get a bunch of different – you can compare on a bunch of different dimensions and not have any trouble computing that computationally.  
  
So what we did was think about, “Well, is there a way that we can actually develop a kernel that would actually be really good for these particular issues?” And what we found was, of course, if you look at almost all the research in fraud bankruptcy, everyone’s using some kind of ratios and some kind of year-over-year changes in ratios and those kind of things. And so went back and found out – and actually developed this kernel that would actually take any number of attributes and then it would actually blow them out into all possible combinations of those ratios and year-over-year changes in those ratios, so it hopefully would try to cover everything.  
  
That was one part of the dissertation that was kind of important, and so I’m continuing to be working on that now and trying to kind of possibly improve that model. I’m also thinking about trying to actually use – there’s been some recent research on strategic support vector machines, which actually look at the idea of actually taking – trying to kind of go back and use sort of a utility function of the person who’s going to maybe commit the fraud and think about, from their perspective, what’s the easiest things to change on a financial statement. What are the hardest things, and how can you actually improve your predictability of fraud by starting to understand the kind of moves that somebody might make if they’re close to the line on committing fraud or not fraud. So those are some of the things.

Interviewer: Well Mark, thank you very much. This is really very interesting and important work, and I look forward to seeing the results of this work at section meetings and in journals. Congratulations on your award and thank you very much for your time.

*Mark Cecchini:* Thank you. I appreciate it, Roger.

*Female:* This podcast was produced by Roger Debreceny of the University of Hawaii at Manoa and Stephanie Farewell at the University of Arkansas at Little Rock on behalf of the Strategic and Emerging Technology section of the American Accounting Association. The SET section strives to stimulate and improve the research, teaching, and application of emerging technologies, methods, and techniques in accounting. Visit the section online at <aaahq.org/set>.   
  
The SET section thanks our sponsor CaseWare IDEA. On the web at caseware.com. The views expressed in this podcast are those of the participants and do not necessarily represent the views of the SET section, the American Accounting Association, or CaseWare IDEA. SET podcast can be used for teaching purposes and are available from aaahq.org/set, the AAA Commons at aaacommons.org, and on iTunes.

[End of Audio]