News from New Zealand

by

C. S. Calude

Department of Computer Science, University of Auckland
Auckland, New Zealand
cristian@cs.auckland.ac.nz

1 Scientific and Community News

0. The latest CDMTCS research reports are (http://www.cs.auckland.ac.nz/staff-cgi-bin/mjd/secondcgi.pl):

425. S. Hartmann and S. Link. The Implication Problem of Data Dependencies over SQL Table Definitions: Axiomatic, Algorithmic and Logical Characterizations, 10/2012

426. M. Kirchberg, S. Hartmann and S. Link. Design by Example for SQL Table Definitions with Functional Dependencies, 10/2012

427. F. Ferrarotti, S. Hartmann and S. Link. Efficiency Frontiers of XML Cardinality Constraints, 10/2012

428. S. Link. Sound Approximate Reasoning about Saturated Conditional Probabilistic Independence under Controlled Uncertainty, 10/2012

429. C.S. Calude, E. Calude and M.S. Queen. Inductive Complexity of the P Versus NP Problem, 10/2012

430. M.J. Dinneen and R. Versteegen. Obstructions for the Graphs of Vertex Cover Seven, 12/2012

431. H. ElGindy, R. Nicolescu and H. Wu. Fast Distributed DFS Solutions for Edge-disjoint Paths in Digraphs, 03/2012

2 A Dialogue with Ian H. Witten: My story so far: a stroll through the gardens of computer science

Ian Witten, `http://www.cs.waikato.ac.nz/~ihw/`*, is a Professor in the Department of Computer Science of the University of Waikato in New Zealand. His many research interests include information retrieval, machine learning, text compression, and programming by demonstration. As head of the New Zealand Digital Library Research Group* `http://www.nzdl.org/cgi-bin/library.cgi`*, Professor Witten oversaw the development of Greenstone Digital Library software, used by the BBC, New York Botanical Gardens and UNESCO. He has written several books, the latest being* Data Mining *(2000),* How to Build a Digital Library *(2002),* Web Dragons: Inside the Myths of Search Engine Technology *(2007). His Birthday Book, 2007 (*`http://www.nzdl.org/Books/Birthday/index.html`*) includes contributions from collaborators, friends, and students, from all around the world. Professor Witten is a fellow of the ACM and the Royal Society of New Zealand. He was awarded the IFIP Namur Prize (2004) for "contributions to the awareness of social implications of information technology ... and the need for an holistic approach in the use of information technology that takes account of social implications", and one of 2010 World Class New Zealand Award. In 2012 he was appointed Chief Scientific Advisor at Pingar* `http://www.pingar.com` *to provide advice on developing technologies and solutions to augment enterprise ability to manage unstructured data.*

Cristian Calude: How was computer science during your studies at Cambridge University (MA in Mathematics), the University of Calgary (MSc in Computer Science) and Essex University (PhD in Electrical Engineering)?

Ian H. Witten: Very different from today! At Cambridge in 1968 I took a course on numerical analysis from the famous computer science pioneer Maurice Wilkes, but I'm sorry to say I found it very boring. I wrote a couple of programs in a language called ÒFocalÓ, a precursor to Basic I think. At Calgary in 1969–70 I met both Fortran (punched cards) and a PDP-12 (teletype and paper tape), which I used interactively—very cool, because I could actually see the computer! At

Essex in the early 1970s I was bequeathed an ancient PDP-1, and had a lab of several PDP-8s, each with 4 KB RAM—one even had a massive 10 KB disk. Around 1976 I installed the second UNIX installation in the UK. I loved UNIX because of its openness (still do). All of these computers had less power than a digital watch does today.

CC: Did the education in three subjects—Mathematics, Computer Science and Electrical Engineering—help you well in your career?

IHW: It's a great selling point to have degrees that cover these three fields; everyone's terrifically impressed. But in actuality my Computer Science MSc and Electrical Engineering PhD were in very similar fields, which would be more accurately called applied statistics. I do think that mathematics is a great foundation for thinking in general, though I have no regrets about abandoning my early aspirations to become an actuary in favour of moving to computer science.

CC: You lived in three corners of the world and eventually settled in New Zealand in 1991. What was the motivation?

IHW: When others ask me that I tell them that if they had ever been to New Zealand the answer would be obvious, but you know that already, Cris. It's like asking someone in heaven why they chose that over the alternative. I wanted a more relaxed lifestyle with a greater emphasis on a balanced life and more outdoor recreational opportunities, but also with a good hi-tech environment. New Zealanders are very quick to pick up on new technologies, particularly networking, which is obviously very important over here.

CC: As you said, I know and I fully agree. Tells us about Greenstone, the suite of software for building and distributing digital library collections.

IHW: Greenstone emerged from a desire to apply the *Managing Gigabytes* indexing and searching techniques on a grander scale. A key event was our involvement with UNESCO beginning around 1999/2000, which happened completely by accident. It made me aware of the enormous potential of end-user-built collections for disseminating information in developing countries. Of course, Greenstone is not just for developing countries—it's widely used almost everywhere, including the US—but I became passionate about its use in the developing world. We take libraries, and the internet, and the ready availability of information, for granted; but life is very different in other places—and the degree of western cultural hegemony is awful. Greenstone enables people to build and disseminate collections of their own information, in their own language—the interface has been translated, by volunteers, into over 50 languages, some of which you've probably never heard of. And it's been a tremendous personal opportunity for me as well: Greenstone has taken me to meet new friends in places like Cuba, Fiji, Micronesia, Nepal, Trinidad, as well as several African countries and all over India.

CC: Fascinating, I think you have a myriad of stories from these trips.

IHW: Stories and stories. I'll never forget a week-long workshop in a packed computer lab in Havana with ancient Windows computers and no air conditioning. It was glorious bedlam!—my first real encounter with the Hispanic temperament, I guess. People chatting and singing and flirting on the side, and me fighting to retain control. And then the computers started acting up. "Over here, Ian, Greenstone's stopped working on *this* one." "And here too!" I'd take a look: garish windows popping up autonomously all over the screen. Viruses. One by one the lab computers succumbed, until at the end only a couple were still usable. I since learned that viruses, essentially invisible in my own, obviously well-protected, computer environment, are the curse of the developing world, a productivity-sapping "white man's disease" created by affluent westerners that cripples poor countries that lack the technical support needed to fight them. Once (and only once) I passed my memory stick around a class in Africa to distribute sample files: it took just minutes to pick up hundreds of viruses! Spam email is a similar issue: virtually non-existent for me at home, but a plague in these people's lives. Imagine deleting hundreds of spam messages on a painfully slow internet connection, every day. I have learned to respect other things that we take for granted, like electricity. Just minutes before starting a hands-on workshop in a computer lab in Kathmandu I was told of the scheduled rotating 8-hour electricity cuts ("didn't you know?"), and today's began ... ("let me check the schedule") ... now! And different cultural mores: after a conspiratorially whispered question "do you like alcohol?" (the answer is obvious to those who know me) I was hijacked from a lively and eagerly-anticipated student party in India by smart-suited top officials who insisted I join them in a clandestine, prolonged, and rather dismal drinking session.

CC: Please explain the importance of "keyphrase indexing" and your work in this subject.

IHW: Here's an area where computers outperform people at a task that is obviously human! When you choose keyphrases, or index terms or whatever you want to call them, for a document, the aim is to be consistent with what other people are likely to choose, because ultimately the terms are going to be used for searching and browsing by others in order to find the information they seek. The aim is consistency with others, doing what others would do—to be boring, if you like. And success can be measured in terms of the degree of agreement with other people who independently index the same documents. In terms of this measure—agreement with humans—our experiments have shown that computer indexing technology can outperform "ordinary" people and even rival specialists, including professional indexers. That sounded outrageous when my student speculated that that was possible, but she turned out to be right and I was wrong. This

seems to happen a lot with me and my students.

CC: Your book *Managing Gigabytes* (co-authored with Alistair Moffat and Timothy C. Bell) published by Morgan Kaufmann Publishing, San Francisco in two editions, is about compressing and indexing documents and images. "This book is the Bible for anyone who needs to manage large data collections", wrote Steve Kirsch, cofounder of Infoseek. How do you see this subject after 18 years since the first edition and 13 years since the second one?

IHW: The Bible?—I know I live in heaven, but did Steve really say that?

CC: Yes, he did.

IHW: I also heard that the book was required reading for early employees at Google. We made a big mistake by saying in the first edition's Preface that maybe the second edition would be called *Managing Terabytes*, and we had to eat our words when we came to the second edition. I'd love to write Managing Terabytes, but in truth the problems are completely different when you move upscale. What amazes me about today's search engines is not so much that they can answer queries so quickly but that they can keep going in the face of continuous failure. With a hundred thousand disks, a dozen must fail every day; with a million, one must fail every ten mins. And it's not just disks...

CC: Data deluge is not a danger? Sciences are shifting to engineering by using statistical techniques to sniff through huge databases to find patterns, and, amazingly, with good and very good predictive results. This is fine unless this paradigm "kills" one of the important scientific quests, the effort to *understand*. This issue was discussed in a very interesting conversation between Noam Chomsky and Yarden Katz titled *On Where Artificial Intelligence Went Wrong*, recently published in *The Atlantic*.

IHW: Yep, understanding is indeed degenerating in favour of publishable "results", and I regret to say that I have personally contributed to the problem. I cringe when I read papers that compare this machine learning technique to that one on half a dozen standard datasets and present nothing but a nicely-formatted table of statistical results (to 5 significant digits), with no insights at all. This is pointless research, Weka-enabled. The pressure to publish has become all-consuming. (The only upside, and it's not a very big one, is that such "research" does contribute to my citation count. (That was a joke, by the way.))

As the years pass I've become less interested in philosophical issues surrounding artificial intelligence, brain theory, the mind/body problem, consciousness, and so on. However, one strand of my current work is stimulated by the need for computers to apply knowledge rather than, or as well as, statistics. I think that Wikipedia, although embarrassingly primitive and limited at present, signals a sea change in how our society deals with knowledge. A few hundred years ago, control of society's knowledge was wrested from the Church and relocated

in academic institutions. Now, to the great chagrin of us university professors, our monopoly is under threat: society can collaboratively create, edit, and refine knowledge artefacts without even asking us! (And we are threatened on the teaching side too by the rise of instant internet universities and MOOCs, but that's another story.) An important side effect is that computers can now peruse these knowledge artefacts and benefit from them too. So I'm interested in knowledge mining from Wikipedia and other public information sources. Also—remember Cyc, Doug Lenat's common-sense knowledge project from the 1980s? The project's still going, and although it may seem as though time (and crowdsourcing) has passed it by, Cyc does contain a wealth of core knowledge (about disjunctive concepts, for example, and argument restrictions) carefully handcoded by professionals. Some of our current work is aimed at reaping that and using it for ontological quality control of information garnered from the Linked Data movement and inferred from less reliable sources such as Wikipedia.

CC: Your book *Web Dragons* was described as "not a resource on how search engines work, but rather what ideas and ideals have been realised in the development of search engines, the political and human challenges they face and problems and opportunities they present to humans and to the nature of knowledge and information." How do you see the future of search engines? What about their social role?

IHW: Boy, if only I could answer questions like that! *Web Dragons* was written in 2006, before the astounding rise of Facebook and Twitter, and predicted a social dimension to information retrieval that has now become commonplace. But the future? I think people are less interested in information than I realised, and more interested in phatic communication. Obviously I think social interaction is important, but so is real information—and perhaps it's getting lost in our obsession with trivia. Personally I prefer my social interactions to be face to face. Several followers of my Twitter account are still eagerly awaiting my first post!

CC: Machine learning is another area in which your group has excelled over the years. You have written state-of-the-art software for developing machine learning techniques and then applied it to real-world data mining problems. Please describe the software WEKA and MOA.

IHW: When I came to New Zealand in 1992 I wanted to initiate a project that enabled people in the rather obscure little computer science department I had joined to work together and develop new lines of research. We hit upon machine learning as a technology that was interesting, futuristic-sounding, and had potential relevance to agriculture, the staple regional and national industry. We began work on a C++ machine learning workbench, which at the time was in stiff competition with Stanford's MLC++, and moved to Java early on, which was risky because of performance issues at the time but proved to be an excellent decision.

WEKA has really taken off—which is astonishing considering that it denotes a small flightless bird. But meanwhile I have moved on to other projects, and have not been involved with MOA, a stream-oriented ML project that stands for Massive Online Analysis and denotes a huge extinct New Zealand bird twice the size of an ostrich.

CC: What industrial applications have you developed?

IHW: I'm not very good at working directly with industry. Everything my students and I do is issued as open source software, which is picked up by academics and industry alike. As well as the better known Weka, Greenstone, and Wikipedia Miner, we have Kea (keyphrase assignment), MAUI (multi-purpose automatic topic indexing), Realistic Books, Katoa (knowledge assisted text organisation algorithm), FLAX (flexible language acquisition), and, very recently, FFTS (the fastest FFT in the south, and considerably faster than FFTW, the fastest in the west). All done by my students, I should emphasise; not by me.

CC: Programming and proving are very similar mental activities. It took me a long time to understand that programming is more demanding than proving: the difference comes from the agent validating the product—code or proof, a computer or a human expert. What is programming by demonstration?

IHW: Programming by demonstration involves showing a computer what to do rather than instructing it in some programming language. People often have to do boring, repetitive tasks on computers—reformatting references or addresses; processing lists; drawing sequences of boxes. Given such a chore, perhaps you should write a program—or perhaps it's quicker to just go ahead and do the job manually. Some of my early research was on calculators that inferred iterative computations from the beginning of a sequence of key-presses, and a predictive typing interface for the disabled that set the scene for predictive text entry on today's cellphones. My students and I created a "smart mouse" that automated repetitive graphical editing tasks, an instructible interface that acquired data descriptions and procedures by being taught rather than programmed, and a programming-by-demonstration agent that worked with a set of common, unmodified applications on a popular computer platform. This was amongst my best and most creative research: highly interactive interfaces that incorporate a learning component, with enormous potential to expedite many human-computer dialogues. But I became discouraged. Reviewers rejected our papers, demanding more tightly controlled human evaluation, which we could not deliver because interactions between a user and a "learning agent" evolve over time.

CC: Even top researchers can get discouraged...

IHW: I don't know about top researchers, but I certainly can. Discouraged and rejected. My Dean once said "it's all right for you, Ian; no-one ever rejects *your* papers". But nothing could be further from the truth! I'm sure there can't be

many people who've had more rejections than me. Our book *Computer Science Unplugged* (with Tim Bell, principal author, and Mike Fellows) was rejected by 27 publishers before we gave up. Sixteen years later it has spawned a major movement in the teaching of computer science to school children all over the world, including the US and UK, and has been translated (with our permission) into Arabic, Chinese, French, German, Hebrew, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, and Swedish. This has been possible because, since it's unpublished, we still own the copyright. The silver lining.

CC: If your Dean thinks that *you* (read: as a top researcher) are spared the misery of rejection (sometimes without real base, using unnecessary harsh arguments like "my weakest student would have done it better", "there is no subject for the paper"), imagine how the younger colleagues imagine the "status" of the academic establishment . . .

IHW: Well, this was just one offhand remark of one Dean: other Deans I have worked with are far better informed about the realities of academic publishing. However, I agree with you entirely: the academic establishment is often really tough on younger colleagues. And I have learned that university administrators tend to present different persona to senior professors than to junior staff (which I think is reprehensible), so that in many cases I experience entirely different, and more humane, personalities than colleagues do. But neither do I think university academics should complain too much: life is far tougher, I believe, on most other working people, whether they are digging ditches, teaching school kids, or staring into people's mouths all day. And when I look at referee reports, many of the really harsh ones come from younger colleagues. We all need to do better and be more understanding when evaluating each others' work.

CC: The story of your unpublished book is fascinating: please tell us more.

IHW: Unplugged and unpublished. The book, as the title implies, is about the teaching of computer *science* without using computers, in contrast with IT skills such as the use of Microsoft products, which is what many schoolchildren have experienced up until recently (and they think university computer science will be about advanced word processing). So it's intentionally revolutionary, or at least runs counter to the established culture in the teaching world. The publishers' responses were hilarious: hilarious, that is, if you hadn't invested a large chunk of your life in what they were rejecting. One wrote that they "will not pursue the idea of publishing the book", yet described it as "your wonderful volume ..." and said it "would be a real pity not to have this book released". A children's publisher said it "may be too academic for children", while an academic publisher referred us to a children's publisher. One publishers' educational arm referred it to their computing department, who responded that they couldn't publish a book if it wasn't about how to do things on a computer. Yet now, if you google "unplugged",

we come in above Eric Clapton!

CC: What does it mean “unplugged"?

IHW: You know, stop fooling around, unplug your computer, and start to learn some real computer *science*. As you know, probably better than anyone else, there are fundamental ideas about computation that do not depend on computers at all. *Computer Science Unplugged* has games and activities that teach kids about computation but do not involve computers. My favourite is a kind of formation dance that expresses a parallel sorting algorithm. Kids follow lines chalked on the playground that represent a sorting network and end up sorting N numbers in O(N) time. The `http://csunplugged.org/` website has a video of 21 kids sorting 21 numbers in . . . um . . . about 7 seconds (admittedly the video is played in fast-forward mode).

CC: I have rushed to `http://csunplugged.org/` to download it as I will soon become grandfather . . .

IHW: Congratulations! The activities are suitable for kids of all ages, 8–80. You’ll have to wait a while to try them on your grandchildren, but you might be able to start with your parents!

CC: Recently you have been appointed Chief Scientific Advisor at Pingar.

IHW: Yes. Pingar is a small NZ company that is developing really interesting technology for document analysis and organisation. My ex-student Alyona Medelyan is their Chief Research Officer and it’s great to keep in touch and involved with what they’re doing.

CC: You play jazz . . .

IHW: Now you’re talking. Yes. Live music is the really big thing in life, far more important than computers and technology. A couple of weeks ago at the SPIRE conference in Cartagena, Colombia, I had the great pleasure of jamming with a duo from Spain in their open-air concert and in the bar afterwards into the wee small hours of the morning. But I play classical music as well as jazz. I play second clarinet in the Trust Waikato Symphony Orchestra—this weekend we have a concert called the Waikato Proms, modelled after the famous BBC promenade concerts in London. And a clarinet group meets at my house every week. Often it’s quartets or quintets, but last night it was trios—an excellent evening sight-reading 19th and 20th Century music; a real challenge. I’m lucky enough to play with musicians who are better than me, so it’s a constant learning experience. We play everything: classical, modern, light, jazz, and they say I have the most comprehensive library of clarinet ensemble music in the country. Currently, having spent two months in Buenos Aires last year, I’m obsessed with Piazzolla tangos.

CC: Your yacht Beulah is a 28 foot Nova, New Zealand designed, launched in the early 80s, built of double diagonal kauri wood, fibreglassed over . . .

IHW: Ah yes. Sailing is my other passion, and part of the reason for moving to New Zealand from Calgary where the sea is a couple of day's drive away. I began sailing as a kid. Indeed I once raced internationally, picked to represent Northern Ireland against the South in a youth championships. We lost. And I was never chosen again. Moving quickly on, NZ's Hauraki Gulf is the best water in the world for the kind of weekend cruising that I enjoy now; hundreds of beautiful islands with lovely anchorages. Beulah, my pride and joy, is where I get away from it all: no computer, no internet, no phone even—and we rarely use the engine. There's hardly any electricity, but of course there's live music wherever you go in Beulah. The simple life. We sail and swim and play with dolphins; watch the sun set, drink wine, eat well, and sleep. And we have adventures. You can read about them in our annual family Christmas letters, which Google will find for you if you ask it nicely.

CC: Many thanks.

IHW: My pleasure. I could go on for hours. It's always nice to talk about oneself.