Advance Access publication on June 18, 2008
doi:10.1093/comjnl/bxm117

Foreword re C. S. Wallace

David L. Dowe

*School of Computer Science and Software Engineering, Clayton School of I.T., Monash University,
Clayton, Vic 3800, Australia*
Email: dld@bruce.csse.monash.edu.au
http://www.csse.monash.edu.au/~dld

0.1 CHRIS WALLACE'S OWN WORDS (ESSENTIALLY)

One of the second generation of computer scientists, Chris Wallace completed his tertiary education in 1959 with a Ph.D. in nuclear physics, on cosmic ray showers, under Dr Paul George at Sydney University. Needless to say, computer science was not, at that stage, an established academic discipline.

With Max Brennan[1] and John Malos he had designed and built a large automatic data logging system for recording cosmic ray air shower events and with Max Brennan also developed a complex computer programme for Bayesian analysis of cosmic ray events on the recently installed SILLIAC computer.

Appointed lecturer in Physics at Sydney in 1960 he was sent almost immediately to the University of Illinois to copy the design of ILLIAC II, a duplicate of which was to be built at Sydney. ILLIAC II was not in fact completed at that stage and, after an initial less than warm welcome by a department who seemed unsure exactly what this Australian was doing in their midst, his talents were recognized and he was invited to join their staff (under very generous conditions) to assist in ILLIAC II design[2]. He remained there for two years helping in particular to design the input output channels and aspects of the advanced control unit (first stage pipeline).

In the event, Sydney decided it would be too expensive to build a copy of ILLIAC II, although a successful copy (the Golem) was built in Israel using circuit designs developed by Wallace and Ken Smith. In spite of the considerable financial and academic inducements to remain in America, Wallace returned to Australia after three months spent in England familiarizing himself with the KDF9 computer being purchased by Sydney University to replace SILLIAC.

Returning to the School of Physics he joined the Basser Computing Laboratory under Professor J.M. Bennett[3]. Between the years 1964 to 1968, among other tasks, he wrote utility software for the KDF9 computer and, with Brian Rowswell, redesigned and rebuilt the Direct Memory Access subsection of the KDF9, increasing its peak channel performance while halving the hardware. Also with Brian Rowswell he designed and constructed a high speed data link between the KDF9 and a Control Data Computer. It was during this period that he published a suggested design for a fast multiplier/divide unit, now known as the Wallace Tree [256, 254][4]. This design eventually formed the basis of multiply units in most modern computers. Other achievements during this fruitful period included the development of the hardware component of the undergraduate course in computing and, in 1967, of the first Honours level Computing Course in Australia. He also developed a program for the analysis of urban water reticulation networks[5] using automatic block relaxation which was widely used by New South Wales water authorities.

In 1968, with David Boulton[6], Wallace developed the 'SNOB' program for unsupervised classification which was the first application of Minimal Message Length (MML) inference [290]. That same year he was appointed Foundation Professor (Chair) of the Department of Computer Science (initially called Information Science)[7] at Monash University. The following years were not easy. Computer Science in the early years was (and some would say still is) widely either misunderstood or an altogether unknown quantity in Science as well as the Humanities. The struggle to establish it as an academic discipline rather than simply a 'trade skill' consumed much of his energy for many years in this post, and depleted his health[8]. Nevertheless, he found the time and energy actively to supervise PhD students[9] in subjects as

[1] see, e.g., [40] and sec. 0.3.1 (Brennan)

[2] see also [73, pp85–86] and fig. 4

[3] see [104] [40, sec. 1] and, e.g., [27, 28]

[4] see also [63] and sec. 0.3.1

[5] see [258] (and [288]) and perhaps also [247] [73, Appendix 3, pp145–146 and p154]

[6] see footnote 213 and text around it

[7] in more recent times, some years after Chris stepped down as Head, said group (moved from the Faculty of Science to the Faculty of Computing and I.T. and then later) became part of the School of Computer Science and Software Engineering and subsequently (since Chris's death) now part of the Clayton School of Information Technology. The word 'Information' has returned, although its meaning might possibly have changed since the 1960s. See also [213] and perhaps also fig. 4.

[8] perhaps see also pointers from footnote 36

[9] and Master's students (e.g., [32, 134, 120, 132, 173, 210, 205, 51, 325, 162, 149])

diverse as the geometry of megalithic stone circles [195], the efficient scheduling of multi-tasking operating systems[10], forecasting the development of cyclones [215], and several others[11]. Fortunate in attracting a small but outstanding staff, his department for some years was producing about half the PhD computing graduates in Australia.

Severe resource constraints drove his development of home-made hardware such as a colour graphics display and process-stack and indexing enhancements to the department's small HP2100A computer. With Gopal Gupta, he developed a novel formalism for multi-step ODE algorithms leading to improved high-order stiffly stable methods. His main efforts in these later years aimed at secure operating systems and, very slowly, the development of Minimum Message Length (MML) as a general principle of statistical and inductive inference and machine learning.

Although the fundamental theory of MML was published in 1975 [291][12], exposure in the statistical literature was delayed by referees[13] who, to quote one, found the idea 'repugnant'. It did not occur until, with Peter Freeman's help, a paper in 'proper' statistical jargon appeared in 1987 [305][14] [15]. Subsequent progress in MML led to many successful applications in machine learning, model selection and estimation[16], and this progress continues[17], although there remains widespread misunderstanding of the principle[18].

0.1.1. After-note

Apart from my insertion of the references and the footnotes (of which Chris gave neither), the above is autobiographical text of Chris Wallace's given to Judy Wallace[19] and (in turn) me in June 2004. Chris died about 2 months later in August 2004, with his polished, deep and insightful manuscript on MML [287] essentially complete and to appear the following year[20].

0.2. OTHER - MY WORDS

I suspect that most people reading these words will be doing so within a few metres of at least one Wallace multiplier chip[21]. And yet many broadly-read researchers still express surprise (as possibly the reader does now) to learn that the Wallace of the Wallace multiplier [254, 256] is the same Wallace who developed Minimum Message Length (MML) [290, 305, 300, 287][22].

Indeed, without being too facetious, it is no great stretch to imagine hardware to whose design Wallace contributed (e.g., the University of Illinois's ILLIAC [250, 251, 252, 253, 254][23], the University of Sydney's SILLIAC [247][24], or something more recent) - including several Wallace multiplier chips [254, 256] - cooled thermoelectrically or thermomagnetically by a thermocouple (if the esky[25] machine were big enough) to whose design Wallace contributed [182, 181] running a secure operating system to whose design Wallace contributed testing any number of Wallace-designed pieces of MML software (dating back as far as 1968, and which are far from unknown to out-perform all other [contemporaneously] known methods)[26] using any number of Wallace-designed physically or pseudo-random number

[10] this is perhaps [324]. For more on Chris's and his students' contributions to operating systems, see [49].

[11] e.g., [170, 323, 211, 236, 126, 33, 60, 23, 200, 55, 48, 150], including being associate (second) supervisor of [25], being involved officially and actually with [4], and perhaps confidentially examining or probably choosing the examiners of (or both) [46] as a Doctor of Science (D.Sc.) (which was a collection of published papers).

[12] while this humble assertion might strictly be technically true, the reader should not forget (their earlier) [290, 34, 35, 32, 38, 37, 39, 33]. And, re the justification of Strict MML in the 1975 paper, see especially [291, sec. 3.3].

[13] as it perhaps still is today - as evidenced by, e.g., the curious rejection of [277]

[14] which is largely a polished version of [267]

[15] see also [287, p v]

[16] and statistical inference, econometrics, philosophy of science and inductive inference, reasoning under uncertainty, knowledge discovery and "data mining". For further relevance to "intelligent systems", see also text surrounding footnote 179 (and perhaps also footnote 203 and text leading to it).

[17] see especially Chris Wallace's (posthumous) MML book [287] and also, e.g., [102, 64, 65, 233, 114, 240] and text in and near footnotes 62 to 65

[18] see, e.g., [300, secs. 5.1 and 7] [302, sec. 2] [287, sec. 7.3] [243] [65, secs. 11.3 and 11.4.3] [287, sec. 5.1.2], etc. for corrections of some earlier misunderstandings of the principle (and see, e.g., [176] for a correction using MML to what was otherwise one of the *few* more viable and coherent purported criticisms of Occam's razor - perhaps see also footnote 182).

Perhaps see also footnote 223 (and surrounding text) and footnote 108 re the difference(s) between inference and prediction. And perhaps see footnote 158 re the difference(s) between Maximum A Posteriori (MAP) and MML.

[19] who possibly paraphrased it slightly in places. I have elected rightly or wrongly (in the middle of p. 1 col. 1) to treat "department" as a collective noun and not to change "department who" to "department which". Also, I have seen 'SNOB' variously spelt as: SNOB, 'SNOB' and Snob. Perhaps because it's a proper noun and likewise because (as per footnote 113) it is not an acronym, I usually opt for Snob - although I have no evidence against Judy Wallace's claim that Chris always wrote SNOB in capitals.

[20] Despite the delays caused by Chris's polishing, re-writing and polishing, this work - like much of Chris's work - still seems ahead of its time. Possibly see also footnote 82 (and possibly also see just after half-way into footnote 218).

[21] I acknowledge here partial support from Australian Research Council (ARC) Discovery Grant DP0343650.

[22] see also footnote 136

[23] with [254] becoming [256]. Had he stayed longer at Illinois, I have little doubt he would have been an author of several ILLIAC papers - with presumably at least some of [250, 251, 252, 253] being both published and with Chris as at least an author.

[24] I am grateful to John Deane for drawing my attention to his very recent book on SILLIAC [73] and also for drawing my attention (quite accurately) to parts [73, p2, p64, p70, p76, pp85–86, p87, p90, p96, also p111, pp136–138 (Appendix 1), (p146 and) p154 (Appendix 3), p158 (and p159) (Appendix 4)] which specifically mention Chris Wallace - see footnotes 97, 98, 101 and 100 and see also [27] (which is ref. 85 in [73]) and [104] (which is ref. 116A in [73]). Perhaps see also fig. 4.

I have to also mention that this book [73] contains a photo of Chris [73, p96] (see also [40]) and quite a few of Judy [73, front cover, p4, p125, p138], along with mentions of Judy [73, p4, p45, p50 (and p62), p96, p125, p136] and other pieces of history.

[25] "*Esky*" is a word used in Australia for a portable container which keeps drinks and food cold. New Zealanders refer to this as a "chilly bin".

[26] see, e.g., [290, 291, 305, 300, 301, 302], or, e.g., [307, 306, 319, 189, 188, 294, 277, 96, 100, 241, 243], etc.

generators [263, 269, 272, 274, 279]. However, it is a shortcoming in the current author that this immediate past sentence fails to mention Chris's work on data processing in the early cosmic ray experiments in Sydney [45, 40], one of his papers that can cause some mirth [167], his analysis of nasal support [59], his direct contributions to computer science course structure, syllabi and education[27], and vastly many other works. I have tried to cite them below - from[28] [29] his first works [182, 181] helping a fellow (1954) Honours student (Brian O'Brien[30]) to get his project to work, to his last work published while he was alive [217] to his posthumously published book on MML [287] and the only other two works of his to date that have appeared posthumously [151, 293] - and[31] I have also endeavoured to list them at www.csse.monash.edu.au/~dld/CSWallacePublications.

0.2.1. Chris as a colleague

Some[32] of the main things I miss about Chris Wallace are his gifted intellect[33], his unpretentious humility[34] [35], his sense of integrity[36] and the security I felt in sharing my best (and probably also some of my less good) ideas with him that he would respect the ideas as having been mine and treat them accordingly.

FIGURE 1. A young Chris (4th from left) with friends.

Before I first started working with Chris in May 1991, those rumours that I had heard about his incredible brilliance I attributed understandably but incorrectly to a combination of the shortcomings of those telling the stories and of hype.

Chris was the brilliant brain-box genius, humble and unpretentious. Although an atheist[37], Chris seemed to have more honesty, integrity and tolerance than many I have met who would proclaim these virtues.

Chris sometimes got frustrated and angry working with mortal people like me of finite intelligence. I recall the mirth[38] in the room during one of the speeches at the large gathering for his 60th birthday in 1993 when one of the speech-makers mentioned having been afraid of Chris (and I felt great relief to know I was not alone, and I'm sure I wasn't the only one who could relate to this). Chris's temper did flare a few times (and a few too many for my liking) in my early years with him (and undoubtedly on

[27] see also fig. 4

[28] I am grateful to Dr Carlo Kopp for passing me a scan of a newspaper article which begins with the words: 'A boy, described by his physics master as "the most gifted pupil" he had ever taught, came equal third in the state in Mathematics I, equal first in Mathematics II, and third in Physics in the Leaving Certificate Honours list. The boy, Christopher Wallace, of the Scots College, . . .'.

And I am grateful to Caroline Yeh, Librarian, Univ. of Sydney, for telling me that "Wallace appears to have won three prizes in first year - Barker Scholarship no. II (1st Year Exam for proficiency in mathematics), K. K. Saxby Prize (Proficiency in Mathematics) and the Smith Prize (1st Year experimental physics)".

[29] This last footnote, footnote 28, reminds me of a story Chris once told me about a (3-valve or 4-valve) radio that he once had (well before he left school, I believe) that wasn't working. He found that one of the valves was not working and he didn't have a spare valve, but he then realised that the same radio could be made to work with one less valve, and so he subsequently fixed it to work with one less valve.

Despite this radio (hardware) story immediately above, I do recall Chris's once saying to me of himself: "I was no Seymour Cray.". (Possibly see [73, p162] re Cray's work at age 25.) Bearing in mind Chris's achievements in a wide range of areas, the reader can make of Chris's comment what the reader will.

[30] who subsequently wrote [180]

[31] after finishing this sentence, the reader is at liberty to skip to sec. 0.3.1 re this issue's authors and their contributed articles [40, 228, 143, 47, 63, 49].

[32] I will say that there have been quite a few people who have been anything from sceptical to downright categorically dismissive [and not always politely] upon first hearing about MML, and yet at least two such people that I can immediately think of have gone on to embrace MML (not always with the most complete memory). Suffice it also to say here and now to the young, the inexperienced, the naive, the "uninitiated" and/or the idly curious that (academic) life is by no means always just and to add (the [perhaps] obvious) that Chris, others and I were - or have been - (sometimes very) short-changed at times.

Perhaps also see the start of sec. 0.2.2, and perhaps also see footnote 93 and the text leading up to it.

This all said, Chris was particularly fortunate in being able to enjoy much loyalty and much fair attribution from his Monash (Computer Science) colleagues.

[33] I recall Peter Freeman, one time Prof. of Statistics at Univ. of Leicester and Chris's co-author on [305, 306], referring to Chris as a "wonderbrain" (if not also "brain box")

[34] and a cherishable gargantuan ratio of ability divided by self-promotion

[35] see also, e.g., the title of [256], the sentence leading to footnote 67, footnote 113, [278, p315, sec. 9, last sentence], and/or [287, p vi, Disclaimer and sec. 4.9]

[36] perhaps see the start of sec. 0.2.2 and possibly also see part of footnote 32.

[37] but quietly so, as he was in many matters, but he seemed to feel that introducing the notion of God didn't do much to shorten the length of the second part (which encodes the observed data) of an MML message [287, secs. 9.6.2 and 9.4].

Possibly cf. footnote 206

[38] this is probably as good a place as any to mention some of those who appreciated Chris's sense of humour. Both Judy Wallace and Julie Austin have variously made special mention of it. I recall Gopal K. Gupta's telling a story of how (many many years ago) Chris had been appointed as President of some society (most probably against his will) and how he had to deliver the Presidential Address. Story has it that he went to the blackboard, wrote up the words "Presidential Address" and then wrote down where he lived. Everyone laughed, and then he went on to give an excellent address. Some years after I had been working with Chris, I recall seeing for the first time, fairly prominently displayed in his office, this piece of metal which had been shaped into something like a hybrid of a knot and a three-dimensional figure 8 and mounted on a piece of wood. Whether or not I have the first two words in the correct order, I remember being at first puzzled when I saw the thing and read the three words on the attached plaque: "International Standard Bit". Some time later, I gradually grew to see the funny side of this. See also footnote 113.

occasions with some others or possibly if he had lost an unsaved computer file), and there is no doubt that there were years that I found largely harrowing. But then I began to adapt better and surely also to improve technically[39]. Curiously, from around about the time in 1995ish when he was no longer required to give lectures :-), I can't recall any such outburst. And, looking back upon his outbursts of frustration and on all of my time with him, I don't think I can ever recall his being rude[40].

0.2.2. Collegiality, generosity and perhaps naivety

On matters of level of contribution and authorship, Chris was reasonable to generous to (I believe) often exceedingly generous. There is a story I heard from Chris and which I have heard from two other colleagues (including one who was a colleague of Chris's at the time) of Chris's coming across a work in progress of someone with a similar and related idea to work Chris had published more than once some years previously. Chris found a fundamental mistake in the work, corrected it and wrote to the author(s) with the correction. Some time later, the original work in progress was withdrawn and vanished, and a paper by that author appeared in print with Chris's correction on board but no acknowledgement.

There is another story of Chris's being invited to a well-funded well-known organisation outside of Australia, going and arriving there, making the insightful breakthrough in the relevant work and then having the work written up without his name on it.

I know of several works where Chris appeared in print as last author, despite having done both the mathematics[41] and written the software.

This above all said, I don't think that Chris was bitter about any of this (or indeed about anything in particular) - although he probably had several reasons to be so - and I'm not even convinced that it upset him. He would possibly be embarrassed and think it irrelevant that I raise this. I suspect that he had a quiet purpose about advancing knowledge and understanding, and that his getting due credit was - to him - pretty much secondary[42].

Chris had been a pilot in early (perhaps even teenage) life[43], he did *not* like split infinitives, he had very broad (and deep) general knowledge[44], and the first of only two occasions (so far) I have ever been on a winning team at a trivia night was with Chris and Judy Wallace on my team (although I find it pleasantly amusing that Judy claims to have contributed nothing).

I recall hearing that Brendan Macmillan approached Chris once in the early 1990s and asked Chris whether he knew any good random number generation software. Chris[45] went off and produced [274, 279][46] [47].

When Gopal K. Gupta took over from what had been my initial efforts to get Chris an Association for Computing Machinery (ACM) Turing Award, we had been using Julie Austin (then Chris's personal assistant) to get information from Chris. At one point, after one of Julie's attempts to get information, Chris replied: "Who wants to know?". Gopal finally persuaded a very reluctant Chris to accept the honour of being nominated for an ACM Fellowship [2, pp5–6][48].

In my own experience, perhaps I should have gone as first author on [294], but Chris did at least most of the mathematics – as well as helping de-bug my software and giving feedback on my write-up - but I put him down first as a measure of respect. The Neyman-Scott[49] work [99, 100, 101] was something Chris told me he had done in his head[50]. When I wrote it up with his name first, he reversed the order.

Wallace [283] (1998) kindly cites Wallace and Dowe (1997) [298] the incorrect but generous way around as Dowe and Wallace (1997).

Some time after I gave the talk (on 1 September 1994) which ultimately led to [139], Chris commented that we should focus on trying to infer a (chess) player's search strategy rather than their evaluation function[51]. Some years later,

[39] perhaps cf. footnote 164 from a few years later, in 1996

[40] although I do recall one or possibly two (one professional, one personal) occasions when a person conveyed to me something to the effect that Chris had (essentially) been rude. (I guess these things are subjective, but I can say that I was nearby on both occasions and didn't hear anything.) And I also recall a private occasion with Chris where he said to me of one person that said person "would take candy from a baby". But I would weigh all of this up against Chris's long-tested patience enduring people who not only didn't follow - but often dismissed - his ideas (as per sec. 0.2.3 and last para. of sec. 0.1, and cf. footnotes 13 and 69) and also against his "(not always totally voluntary) generosity" (for want of a much better way to say that some of his works and/or ideas were sometimes written up by others without due acknowledgement and not always with his consent).

[41] which, in some of these cases, would have been and presumably surely still would be beyond the relevant co-authors

[42] perhaps also see footnote 93 and the text leading up to it

[43] see fig. 2

[44] I recall a former colleague's telling me some time ago that, during his Hons year, he and fellow Hons students steered a conversation with Chris in several directions without being able to find a topic where Chris was out of his depth. That story is not inconsistent with my personal experience. More recently, Julie Austin (who is also mentioned in sec. 0.2.2 and in footnotes 82 and 185) wrote shortly after Chris's death: "I feel very privileged to have had the opportunity to assist Chris in his work - he had an amazing brain and memory for facts and figures (one member of staff said to me once "I will ask Chris - he knows everything.")."

[45] already the author of [263, 269, 272]

[46] for a discussion of this work, see Richard P. Brent's paper in this issue [47]

[47] Dr Peter E. Tischer tells me that he once asked Chris about applying MML to images (and compression thereof) and that Chris seemed to respond to such comments as (invitations or) challenges. Peter says that his comment is what led to [283].

[48] perhaps see also middle of footnote 229. I have been reliably informed by someone that, in similar vein, Chris would not agree to said someone's offer to nominate Chris to the Australian Academy of Science.

[49] see also [287, secs. 4.2–4.5 and 4.8]

[50] and, from the context in which he spoke, this would have been no later than the mid-1980s. Mind you, using Chris's own Wallace-Freeman (1987) approximation [267, 305], this "only" requires collecting powers of σ from the likelihood (and the Bayesian prior) and the expected Fisher information.

[51] I am reminded that Chris once told me some years ago of a (chess-like) game that he had invented, although I didn't write down the details. The board was rectangular, perhaps square, perhaps 8 × 8. The game was symmetrical and (even more symmetrical because) both sides move(d) simultaneously. I am fairly sure that a piece could not move onto a square currently occupied by a piece of its own side/colour. A piece captures in the same way that it moves,

I gradually came to agree with him. (Of course, one could - and ideally should - try to infer both.)

Re [79], Chris explained the notion of minimising the expected Kullback - Leibler distance (as a similar but alternative version of invariant Bayesian point estimation to MML) [79] [287, secs. 4.7–4.9] in his office to just me and then again within a few days in his office to Rohan Baxter, Jon Oliver and me. When we expressed a desire to write this up, we offered Chris first authorship, he said he didn't need to be an author, we persisted and he finally gave in and let us include him as (a compromise) last author.

Vapnik discusses MDL SVMs [238, sec. 4.6]. Of course, using multinomial distributions with MML enables us to create MML SVMs with two or more classes without requiring "one against the rest" [233] (see also [154] [155, sec. 12] [76]). Chris had maybe two but possibly three different ways of using MML to encode SVMs. I got to see at least one or two or conceivably three of these in probably different sessions of sitting with Chris, but I'm not sure that Chris or I or anyone else wrote much - if anything - down[52]. One of these schemes was to encode the support vectors and their classes (or regions they represent) as part of the first part (hypothesis)[53]of the two-part MML message, with the first part also encoding - for each region - the probabilities (100% or less) of the classes. Another, alternative, coding scheme of Chris's was to encode the coefficients of the separating hyperplane (rather than the support vectors) to the appropriate precision (as dictated by MML [287, secs. 4.12.1 and 5.2.14]) in the first part of the message. Peter J. Tan's MML SVM coding scheme [233] gives a probability for the SVM in the first part of the message as the amount of (what I shall call) 'wiggle volume' the hyperplane has without changing the classification of the data. Whether or not Chris might have had this idea some years earlier, I have little doubt Peter Tan's work [233] was independent[54] [55]. Chris was well aware of formulae saying that the VC (Vapnik-Chervonenkis) dimension was the minimum of (e.g.) two expressions. MML advocates (e.g. [300]) the minimum of alternative forms (schemes) of message length as giving the best (MML) model. Chris expressed a belief that (I paraphrase) putting SVMs in an MML framework and - given a data set - choosing the SVM coding scheme (from the above alternatives) that resulted in the briefest encoding would give an almost identical expression

capturing an opponent's piece by moving on to the square occupied by that opponent's piece. If two pieces move simultaneously onto the same vacant square, both pieces are deemed captured and removed from the board. The equivalent piece (in Chris's game) of the (chess) pawn moved forward one square at a time, and captured the same way (forward one square). There was some sort of notion of a (chess-like) pin. As in chess, at the start of the game, each player's pawns line up side by side on and filling the same rank - it might have been the first rank and it might have been the second rank (but it was definitely one of these, and the issue of whether 1st or 2nd rank can typically be side-stepped by just adding or removing a row/rank at the end of the rectangular board). I am not quite 100% sure what happened if two pawns faced one another on adjacent squares and both wanted to move on to the other's square, but I'm pretty sure that both were deemed captured (and removed from the game). Perhaps pawns were permitted to move sideways, or at least when the square in front was occupied (by a piece of the other side). I am not sure whether there were only pawns, but I think so. If there was another piece, then there was only one of them and it was fairly immobile (perhaps it could not move at all, or perhaps it could only move sideways one square at a time along the back rank) and played the same role as the King in chess. The objective might have been (or was something like) to get a pawn through to the back (row or) rank.

If forced to (remember hard and) commit, I think the following. There were only pawns (and no King). The initial set-up was symmetrical with all one's pawns filling one rank, namely (I think) the back rank. Pawns could move forward one square at a time or sideways one square at a time (but possibly only sideways if there is an opponent's piece directly in front - although I'll guess that there was no requirement that an opponent's pawn be directly in front), but not onto a square currently occupied by one of one's own pawns. Two pieces moving simultaneously on to the same (vacant) square or onto one another's squares were deemed captured. The objective was to be the first player to get a pawn through to the back rank.

[52] except probably in pen and it might not have been preserved or might have been misplaced

[53] Assuming that the sender and receiver both know the positions of the data points, the sender can send a message to the receiver encoding/conveying which of the data points will have their classes described (and these can be thought of as the support vectors) followed by the class label for each such support vector. With N data points and dimensionality D, there would appear to be at most $min\ \{D + 1, N\}$ support vectors, appearing to cost $\log(min\ \{D + 1, N\} - 1)$ to encode this value, N_{SV}. (We could assume that $N_{SV} \leq 1$ when there is no separating hyperplane, but from hereon we shall presume that there are two regions and that $N_{SV} \geq 2$. We might also argue that the upper bound on N_{SV} is too tight.) With dimensionality D, N data points, N_{SV} support vectors, and N_{SV^+} and $N_{SV^-} = N_{SV} - N_{SV^+}$ support vectors for the positive (+) and negative (−) classes respectively (for some N_{SV^+} s.t. $1 \leq N_{SV^+} \leq N_{SV} - 1$) and proceeding as above, this should cost (approximately)

$\log(min\ \{D + 1, N\} - 1) + \log(N \text{ choose } N_{SV}) + \log(N_{SV} - 1) +$
$\log(N_{SV} \text{ choose } N_{SV^+})$, or (equivalently)
$\log((min\ \{D, N - 1\})\ (N_{SV} - 1)\ N!)$
$-\log((N - N_{SV})!\ N_{SV^+}!\ N_{SV^-}!)$.

I think that that, with at least the terms

$\log(N!\ /\ ((N - N_{SV})!\ N_{SV^+}!\ N_{SV^-}!))$, is where Chris left it without implementing it, but there are (at least) two (or three [or four]) ways to refine this scheme - many of which should similarly apply for MML encoding of (partitioning) DNA microarray data. (Possibly cf. [235] or footnote 196 re microarrays.)

The first, and perhaps simplest, of which Chris surely would have been aware, is to use the geometry of the data points to avoid (allocating code-words to) particular allocations of (N_{SV}) points that could not possibly be support vectors - such as three points in triangular extremeties in a plane of points.

The next, and perhaps next simplest, way is to use latent factor analysis [306, 277, 105] [287, sec. 6.9] or principal components analysis (PCA) to modify the axes.

The third - and arguably most important or most subtle - way is to use the notions of I_{1D} [287, sec. 4.10] and Ideal Group (DI_1) [287, sec. 4.1], as per the example(s) in [287, sec. 3.2.3 and sec. 8.8.2, p360], to group together similar hypotheses - i.e., similar sets of support vectors - while the shortening in the (expected) length of the first part of the message outweighs the damage to the (expected) length of the second part. (Possibly see also material surrounding footnote 62.) We should note here that, in combining two (or more) SVMs with different support vectors into one, there is the potential for two (or more) boundaries and three or more regions. While having more than two regions would have predictive merit (cf. footnote 223 and text surrounding it), for our purposes of inference our SVM should be standard in having two regions - even though the grouping might necessitate taking support vectors from the two (or more) (grouped) models (or possibly new support vectors not in those models) and possibly increasing the dimensionality. (Possibly cf. footnote 153.)

A fourth way would be to use inverse learning, implicit learning, generalised Bayesian nets or generative learning from sec. 0.2.5. In a nutshell, the idea here is of VC dimension as the conventional minimum of two or perhaps three or so other expressions[56].

And, of course, where one wants to create new ([non-linear] kernel) functions on which to attempt to discriminate (e.g., $x_1^2 + x_2^2$ in the case that perhaps one class lies predominantly inside a circle and another class lies predominantly outside this circle), we can use MML to encode any such functions.

Staying with VC dimension but moving from SVMs to Structural Risk Minimisation (SRM), in other (probably unpublished) work of Chris's, he drew at least my attention[57] to a subtle flaw in the purported guaranteed error bound in SRM. Although each individual model in a family (e.g., each polynomial of degree d for varying d in the family of polynomials) might have a "*guaranteed*" error bound, the subtle trap is that the best performing model is unlikely to honour its guarantee[58].

My MMLD (or I_{1D}) message length approximation [113, 112, 6, 109, 5] [287, secs. 4.10, 4.12.2 & 8.8.2, p360] $-\log(\int_R h(\vec{\theta})\, d\vec{\theta}) - (\int_R h(\vec{\theta}) \log f(x|\vec{\theta})\, d\vec{\theta})/(\int_R h(\vec{\theta})\, d\vec{\theta})$ was something I first developed (after many false starts) in the second half (approx. September) of 1999 and which I told Chris about at the next available opportunity a few days later. He conceded after a few seconds that it all looked kosher (which was welcome news) and when I told him I didn't know how to find the boundary of the region, R, he remarked a few seconds later that the message length on the boundary of R would be 1 nit longer than the (prior-weighted) average on the interior of R. He left to go home and would have been there by the time about half an hour later when I realised that he was right. MMLD was motivated partly by a desire to retain the statistical invariance of the Wallace-Freeman approximation [305] [287, secs. 5.1 (I1B) and 5.2.6] while ideally being more robust[59] [60].

Although R gives us a message length, I was initially unsure as to how to get the (invariant) point estimate. I opted for the posterior-weighted minimum expected Kullback-Leibler distance (MEKLD) [79] [287, sec. 4.7] but restricting the posterior to be over R, but the inference-minded Chris preferred the *prior*-weighted MEKLD restricting the prior again to be over R. But, later on in [287, sec. 4.10], Chris opts not to choose a "region centre" at all but rather to use an appropriately encoded randomly chosen point estimate[61] from within R.

essentially that we are not interested in only encoding the values of the target attribute(s) [the class], but rather we are interested in encoding all the attribute values.

[54] and, of course, whichever of these above schemes we might use to encode the SVM's separating hyperplane(s), we can also use MML to encode any possible (computable) choice of SVM kernel function.

[55] For what it's worth, compared to *both* C4.5 and C5 [204], the MML (oblique) decision trees with SVMs in the leaf nodes [233] have *both* a higher "right"/"wrong" predictive accuracy and a substantially better log-loss probabilistic score [see text in sec. 0.2.5 from footnote 170 maybe as far as to footnote 175 or even to footnote 176] while also having less leaf nodes. (Digressing, there is the potential to possibly improve the oblique tree coding scheme from [233] as follows: each time a split is being considered to take us a level deeper in the tree, see whether the breadth (of "wiggle") in a higher split could be increased without affecting the allocation of data to the leaves.)

Perhaps see also start of footnote 153, and possibly cf. near the end of footnote 135.

[56] a recent and potentially important paper here is [214]

[57] and I understand that he earlier, when visiting Royal Holloway (RHUL) in 1997, drew it to the attention of an audience including at least some of V. Vapnik, A. Chervonenkis, R. J. Solomonoff and J. J. Rissanen. (Chris had been working on univariate polynomial regression comparing SRM, MML and other techniques at about that time [281].)

[58] To highlight the point with a somewhat extreme case, imagine analysing the heights of TS botanical trees, with T trees grown from each of S species. From height observations $\{h_{s,t} : s = 1,\ldots, S;\ t = 1,\ldots,T\}$, we wish to infer the mean height μ_s for each species and especially the largest mean, $\max_{\{s=1,\ldots,S\}} \mu_s$. If we do this by choosing $\hat{\mu}_s = (\sum_{t=1}^{T} h_{s,t})/T$, then both extreme cases (maximum and minimum) will typically be exaggerated, especially in the pronounced case when all species are identical and for some μ for all s $\mu_s = \mu$. In similar vein, many quantitative finance experts are at least generally aware that the top-performing funds of recent years will not be expected to perform as well in the subsequent year.

[59] The Wallace-Freeman approximation [267, 305] (cf. Phillips and Ploberger PIC [198]) assumes that the prior is constant (or varying linearly) over (at least) the uncertainty region, and it then takes the Taylor expansion of the log-likelihood function only as far as the 2nd order (quadratic) term - from which the Fisher information comes. Three other ideas I had for retaining invariance but incorporating more terms were:

(i) to get more terms in the Taylor expansion by (simply) transforming to a space where the prior is uniform and then taking as many terms in the Taylor expansion as necessary. [In more than one dimension, terms beyond the second order (quadratic) would probably involve tensors.] This is unique in one dimension (and my 4*th* year Hons student, Edmund Lam, briefly investigated it on my behalf in 2000, in 1 dimension). Chris had concerns about lack of uniqueness of the transformation in more than one dimension (and I think he might have also at least at one stage had some general concerns about the over-dependence upon the prior).

(ii) again with prior transformed to be uniform, to modify the assumption of a symmetric uncertainty region $[-\delta/2,\ \delta/2]$ and corresponding integral $(1/\delta)\ \int_{-\delta/2}^{\delta/2} \cdots$ to have uncertainty region $[-\delta_l,\ \delta_u]$ and corresponding integral $(1/(\delta_l + \delta_u))\ \int_{-\delta_l}^{\delta_u} \cdots$.

(iii) both (i) and (ii).

[60] If we restrict ourselves to one dimension and if our uncertainty region (R) is invariant, then by virtue of our being in one dimension, we could choose our point estimate to be the median of the prior or the posterior - or possibly some other function - over the uncertainty region. Where the choice of function (e.g., the square root of the posterior) does not give an invariant median, then as in the previous footnote (footnote 59), we can get invariance in one dimension by transforming to a space where the prior is uniform.

[61] This raises (or that raised) the issue of *random coding*, something Chris told me about in the early-to-mid 1990s but which he never seems to have written up (other than to brush past it in [287, sec. 4.10.1]). In the MML framework, the sender and receiver have the same likelihood function and the same prior over the entire parameter space. Let us give sender and receiver the same deterministic pseudo-random number generator with the same seed (vector) so that they sample the same values of $\vec{\theta}_1, \vec{\theta}_2, \ldots, \vec{\theta}_i, \ldots$. Using a code over the natural numbers such as (Rissanen's) log* (or the Wallace Tree Code [287, Fig. 2.13 and sec. 2.1.14]), for every positive integer i we can encode data x with a code of length $\log^*(i) - \log f(x|\vec{\theta}_i)$, where $\log^*(i)$ is the cost of encoding the hypothesis given by $\vec{\theta}_i$ and $-\log f(x|\vec{\theta}_i)$ is the length of the second part of the message encoding data x given $\vec{\theta}_i$. Because the second part of the message must be of length at least $-\log f(x|\vec{\theta}_{\text{Maximum Likelihood}})$ and to a

At least some months or years later, I modified the parameter-space MMLD (or I_{1D} or $I1D$) described above to a similar form in data space, which I called SMMLD. SMMLD was motivated by a desire to be invariant like Strict MML (SMML) [291, 305, 278, 79, 300, 301, 108, 112, 109, 5] [287, chap. 3] [65, sec. 11.2] [82] while (like MMLD and unlike SMML) requiring only the local part of the code book to be constructed and (relatedly) being appropriately sensitive to changes in the data. I gladly and proudly shared SMMLD with Chris, who perhaps understood it a bit *too* well, and came back with a corrected (or perhaps generalised) version, which he called the Ideal Group (or IG) estimator [287, secs. 4.1, 4.3 and 4.9] [5, sec. 3.3.3, pp60–62] [109, sec. 5.2, p70, ftn 1]. At one point, Chris offered to attribute his corrected version [287, sec. 4.1.2, p199][62] to me, but I declined. I think it took me until 2006 to appreciate that his corrected version is indeed (most probably) correct[63] [64] [65].

When Peter Grünwald first invited me in about 2001 to write the one solitary chapter he wanted on MML for the book[66] on Minimum Description Length (MDL) that he was to be first editor of, I asked him for two chapters (so that I could get one for Chris). Peter declined, I offered Chris my chapter, and Chris declined (perhaps to focus on his book [287]).

Chris pretty much did his work for its own sake, content to leave his light under the proverbial bushel, just sharing with those around him, and not seeking fame but rather often obstructing any attempts from others to gain him formal recognition[67]. Yet there were very infrequent remarks - seemingly more observations than (clear) expressions of frustration - that he felt his work was unduly slow to catch on[68]. He was well aware of the early receptions[69] given to Galileo Galilei and to Wegener (of continental drift) [287, sec. 9.1, p387] and of Thomas Kuhn's curious observation [287, Chap. 9 p385] that part of scientific revolution seems to be waiting for the

lesser degree because the length of the first part of the message is non-decreasing as i increases, a sufficiently (possibly *very*) long search will (eventually) find the i and the $\vec{\theta}_i$ resulting in the MML inference under this scheme. At least at one stage, Chris thought that this coding scheme would be a reasonable alternative when all else seemed intractable.

The above is random coding as Chris described it to me, although I note that when he brushes past random coding in [287, sec. 4.10.1] he elects to stop with the first $\vec{\theta}_i$ in the uncertainty region (A_x, or R) - because, in this case, he wishes to sample from the prior over R. I am not sure of the pros (and cons) of continuing the search beyond the first $\vec{\theta}_i$ inside R.

Possibly see also sec. 0.3.1 (Solomonoff) on "resource limited ALP" (or Resource Bounded Probability [RBP]).

[62] which essentially properly accounts for the fact that some potentially observable data might be measured to vastly different accuracies to other observable data, but (given parameter vector $\vec{\theta}$) the ratio of probability $f(y|\vec{\theta})$ divided by marginal probability $r(y)$ is more or less constant as the measurement accuracy of y varies (and decreases down towards and into tiny).

[63] This last footnote, footnote 62, raises a point about a highly-related estimator which Chris never (quite) wrote down but which (see, e.g., [287, sec. 4.3]) I can only presume he would have thought of. For want of a better name, I'll call it either Tiny Accuracy Ideal Group (TAIG) or Infinite(simal) Accuracy Ideal Group (IAIG). Here, we modify the set of observable data, X [287, sec. 3.2, p153], to X_∞, where X_∞ contains the observed data (also in X) but any other values in X (not actually observed) which were discretised from the continuum (because of finite measurement accuracy [294, pp1–3] [296, p38, secs. 2 and 2.1] [78, sec. 2] [146, p651] [303, sec. 2 p74, col. 2] [64, sec. 9] [114, eqn (19)] [287, secs. 3.1.1 and 3.3] [65, sec. 11.3.3 p270] [82]) are now ("undiscretised" and) allowed to take any value from the continuum in the pre-image to the discretisation - i.e., any value from the continuum which would have discretised (or rounded) to the relevant value in X. Perhaps slightly more formally, we (successively) let $X_1, X_2, \ldots, X_n, \ldots$ (in turn) be X modified so as to keep the actually observed data but then - for the unobserved data - have every continuous-valued attribute (which was measured in X to accuracy ϵ) measured to the finer accuracy of ϵ/n. So, $X_1 = X$ and, in some crude sense, $X_\infty = \lim_{n\to\infty} X_n$. We can then replace the averages in the boundary rule [287, sec. 4.1.2] (over subsets t of X) by integrals (over subsets of X_∞). In Ideal Group (IG) estimation, the observed data is necessarily a member of the Ideal Group. In (TAIG or) IAIG, the observed data (or, equivalently, that part of the data space in the continuum corresponding to its pre-image under discretisation) is again necessarily contained in the Ideal Group.

The point of (TAIG or) IAIG is that we probably shouldn't care about the measurement accuracy of any potentially observable data in X that we didn't actually observe, so why not make all such data of infinite(simal) accuracy and - in turn - enable ourselves to replace our sums by integrals? Of course, when all attributes are discrete (or categorical) and no attributes are continuous, then nothing changes and (TAIG or) IAIG is identical to Ideal Group. The other - perhaps slightly ironic - thing to mention is that the integrals (in X_∞) that we were keen to replace the summations (in X) with will almost certainly in practice end up being replaced (numerically) by approximating summations (even if, of course, different to the summations in X).

I do seem to recall Chris's telling me that, given observed data, x, for any estimator $\hat{\theta}$ one could give a message length. He went on to say that one chose the $\hat{\theta}$ giving the minimum such message length, but I couldn't keep up with him at the time. Certainly, looking back, given observed data, x, and an estimator $\hat{\theta}$, one can use the boundary rule [287, sec. 4.1.2] to construct an Ideal-like Group giving $\hat{\theta}$ (or something close to it) as the estimator. Using (TAIG or) IAIG, I think one should be able to construct an Ideal-like Group giving $\hat{\theta}$ as the estimator. One then needs to search through the parameter space for that value of $\hat{\theta}$ minimising the message length and giving *the* Ideal Group.

Construction of the I1D (or MMLD) region, R, and in turn the approximated message length [287, sec. 4.10], entails a search in parameter space. Construction of the Ideal Group (Tiny Accuracy or otherwise) seems to be more CPU-intensive, appearing to require an embedded search with an outer search in parameter space and - for each purported parameter estimate, $\hat{\theta}$ - an inner search in data space.

(Note that in the construction just given of TAIG if we had also divided the actually observed data to finer accuracy ϵ/n in each iteration, X_n [let's call this X'_n rather than X_n], and then insisted that all divisions of the actually observed data be placed in the same (ideal) group (of X'_n) then we would get a nearly [and possibly] identical grouping and estimate. However, something (probably) more CPU-intensive but perhaps at least worth exploring in principle would be Tiny Accuracy Strict MML [TASMML], in which we perform Strict MML on X'_n - possibly [or possibly not] insisting that all divisions of the actually observed data be placed in the same group. [This might be fairly stable as n increases.] When all attributes are [categorical or] discrete and none are continuous, then this will clearly be SMML.)

[64] While dealing with the technical issues from the recent footnotes 62 and 63, I will mention (again) that Strict MML (SMML) partitions in data space. Chris once proposed an SMML-like estimator which instead partitions in *parameter* space (or hypothesis space) [287, sec. 4.11 and perhaps sec. 4.12], and this has variously been called Fairly Strict MML (FSMML) [79, p88 and sec. 4, p90] [112] and MMLA [111]. Chris mentioned to me late in his life that [287, sec. 4.11, p214] such a "code construction … has flaws which vitiate the approach".

Possibly relevant here is some text from Daniel F. Schmidt of 28 May 2007 (revised on 13 June 2007) proposing a related new estimator. Daniel Schmidt's old guard to die. Having developed the notion of a (non-universal) Educated Turing Machine [300, sec. 4 (C4 and C5)] [287, sec. 2.3.11] and having thought about the evolution of priors [287, sec. 2.3.13], Chris was well aware of how input to a (until then) Universal TM (UTM) could cause it to be "Educated" and to lose its universality [300, sec. 4 (C4 and C5)] [287, sec. 2.3.6 p119 and sec. 9.6.1]. In one of his not uncommon deep mixes of philosophy and matters computational, he came to me once and muttered that (and I paraphrase this as best I can) for any given UTM given random input, the probability that it will lose its universality is 1 (or close to 1) - and that this explains why people lose the ability to learn[70]. I think that the comment was a combination of the facetious, the serious and the philosophically enticing[71].

One of his last technical comments in May or June 2004 was something partly sceptical about the potential of quantum computing [and therefore maybe quantum MML], seemingly comparing the Heisenberg uncertainty (or indeterminability) principle to some quite well-known property in Fourier analysis[72] [73] [74].

Although the MDL principle was first published a decade after MML [300, sec. 1 p271, col. 1] [65, sec. 11.1], in Chris's papers he would regularly cite several different papers of various different versions of MDL[75] as they appeared and as he became aware of them - and he was politely constructive in his criticisms[76] [77].

Chris showed much generosity (sometimes I would say naively so) with his ideas and, at times, his co-authors[78]. One colleague in particular often asks whether Chris, a full professor and department head (foundation professor and chair - see fig. 4) at age 34 (and clearly more than deservedly so), would have ever made it to professor under the later modern regime[79].

0.2.3. Some more quiet achievements

Suppose we set a task of doing the mathematics for some inference software, writing the software, debugging it, testing it with some random number generations and also giving a

text (which is perhaps best read in parallel with [287, sec. 4.10]) follows: "One approach to resolve the issue of selecting a point estimate within the I1D message length framework is to reformulate the message length expression so that the coding quantum and round-off terms are independent of the data; this is termed the I1F message length. The I1F estimator is derived from the FSMML estimator in a similar way to the Ideal Group estimator's derivation from the SMML estimator: it finds the optimal coding region for a given point estimate, ignoring the effects of all other regions in the parameter space (hence *ideal*). For a given uncertainty region $\Omega \subset \Theta$, the assertion cost and excess round-off cost (also deemed the *model cost*) is then

$$\mathcal{D}(\Omega, \theta^*) = -\log q(\Omega) + \frac{1}{q(\Omega)} \int_{\theta \in \Omega} h(\theta) \Delta(\theta \| \theta^*) \, d\theta \qquad (1)$$

where

$$q(\Omega) = \int_{\theta \in \Omega} h(\theta) \, d\theta \qquad (2)$$

is the prior mass assigned to the uncertainty region, θ^* is the point estimate under consideration, and $\Delta(\theta \| \theta^*)$ is the KL-divergence between the 'true' generating model θ^* and 'approximating' model θ. Inference within this framework is then performed by seeking the solution of

$$\hat{\theta} = \arg \min_{\theta^* \in \Theta} \left\{ L(y|\theta^*) + \min_{\Omega \subset \Theta} \{ \mathcal{D}(\Omega, \theta^*) \} \right\}, \qquad (3)$$

where y is the data under consideration. The model cost is invariant under one-on-one model reparameterisations and satisfies the boundary rule. This new message length expression may be reconciled with the usual I1D formula by simply replacing $\Delta(\cdot \| \cdot)$ with the empirical Kullback-Leibler divergence $\Delta_N (\theta \| \theta^*) = L(y|\theta) - L(y|\theta^*)$ and choosing θ^* to coincide with the Maximum Likelihood estimates. The I1F approximation also serves as a basis for the well-known Wallace-Freeman estimator. By assuming that the prior density is roughly flat within the uncertainty region, that the KL-divergence may be adequately captured by a second order Taylor series expansion, and that the uncertainty region is congruent to a suitably scaled, translated and rotated cell of an optimal quantising lattice, the integrals in (3) may be solved explicitly, yielding the Wallace-Freeman estimator. Preliminary investigations indicate that the I1F estimator should yield a consistent estimate of scale for the Neyman-Scott problem. More details may be found in [220]."

[65] While I can't claim to have totally grasped Daniel F. Schmidt's ideas from footnote 64 immediately above, my immediate intuition - correct or otherwise - is that we should replace 1 by (the amount of data) N in his equation (1). Although, then again, perhaps this just comes down to his and my defining the probability distributions differently, whereupon my immediate intuition would then be to agree with him. (As such, depending upon one's notion of probability distribution, please read the term N in equations (4) and (5) as though it might be 1.) Venturing on boldly and perhaps blindly, my further intuition is that his equation (3) should possibly be something like

$$\hat{\theta} = \arg \min_{\theta^* \in \Theta} \left\{ -\log q(\Omega) + (L(y|\theta^*) + \frac{N}{q(\Omega)} \int_{\theta \in \Omega} h(\theta) \Delta(\theta \| \theta^*) \, d\theta) \right\} \qquad (4)$$

The other observation is that Daniel is using $\Delta\ (\theta \parallel \theta^*)$, the Kullback-Leibler distance from θ^* to θ. Of course, one could argue instead that the expansion of the second part of the message in the Wallace-Freeman (1987) approximation [305] is to do with how much θ^* (the chosen estimator) differs on average from the various θ it is meant to represent - rather than the other way around. As such, if I am following and if my intuition is sound, then I would be tempted to reverse the Kullback-Leibler distance so that we instead have $\Delta\ (\theta^* \| \theta)$. Be this as it may, my stretched intuition is that using Daniel's choice of $\Delta\ (\theta \| \theta^*)$ is going to take us closer to Maximum Likelihood than is the alternative choice of $\Delta\ (\theta^* \| \theta)$. I say this because of Maximum Likelihood's propensity to over-fit (and give overly small probability estimates to many events), and so the Kullback-Leibler distance from the Maximum Likelihood estimate to a neighbouring estimate will typically be smaller than the Kullback-Leibler distance from the (typically more conservative) neighbouring estimator to the (over-fitting) Maximum Likelihood estimate.

To throw up a possible alternative formula engendered from Daniel's excellent idea, perhaps his equation (3) should possibly then be the modification to equation (4) to give something like

$$\hat{\theta} = \arg \min_{\theta^* \in \Theta} \left\{ -\log q(\Omega) + (L(y|\theta^*) + \frac{N}{q(\Omega)} \int_{\theta \in \Omega} h(\theta) \Delta(\theta^* \| \theta) \, d\theta) \right\} \qquad (5)$$

Regardless of the accuracy or otherwise of any part(s) of my note here, certainly Daniel's idea (from footnote 64 and apparently also from [220]) is excellent and likewise merits closer inspection and further study.

Having said the above, I will now digress and add in passing that I think that the data-space boundary rule [287, sec. 3.3.2] is worth bearing in mind when contemplating (such) MML approximations - even approximations in parameter space.

philosophical justification for this methodology. I do not exaggerate when I say that if we were to pit Chris on his own against a team including every other employee of Monash University's Faculty of I.T. (past or present, going back to the very early 1990s) and augment this team by every co-author (apart from Chris) that I have had so far[80] [81], then I have little doubt that Chris would have completed the task both more thoroughly and earlier than the assembled team[82] - and I certainly mean no disrespect to the assembled team, myself or any of its other members. The downsides of Chris's humble genius were that he did not have a reputation that came anywhere near matching his ability and the depth and breadth of his contributions.

One result of this is that we see cases where terminology he used (e.g., *nit* [35, p63] [326, p211] [146, p651] [140] [3, p1426] [287, p78] [65, sec. 11.4.1]) has been overlooked for later terminology and some of his MML ideas undercited or indeed even uncited in subsequent writings by other authors[83]. Such ideas include the Snob program for MML finite mixture modelling [290, 268, 271, 270, 296, 297, 299, 298, 303] [287, sec. 6.8][84] [85], MML hierarchical clustering [37] (applied in [293]), MML as a form of invariant Bayesian point estimation [291, 305] [278, sec. 3.5 and elsewhere] [79, 300, 65, 287, 303, 96] (which might be regarded as Maximum A Posteriori (MAP) done properly [303, secs. 2 and 6.1] [65, sec. 11.3.1][86]), MML inference of probabilistic finite state automata (machines) and grammars [307] [287, sec. 7.1], mixtures of line segments [307] [117, sec. 5 and fig. 2] and outliers (as part of what some might call robust regression)[87], statistical consistency and efficiency of MML in general [305, sec. 2, p241] [24, 278] [287, sec. 3.4.5] and on the Neyman-Scott problem [100, 101] [287, secs. 4.2–4.5][88], the use of both the Fisher information (determinant of expected Fisher information matrix) and lattice constants (κ_n) in MML [305]

[66] with M.I.T. Press, and the chapter ultimately became [65], with the final camera-ready copy submitted in October 2003

[67] see footnote 48 and nearby about nomination for ACM Award and non-nomination to the Australian Academy of Science.

[68] see also end of footnote 79 and text leading to footnote 70

[69] cf. footnote 32. Recalling also the text of footnote 13 and the text immediately following it, although I can't speak directly for Chris, I'd have to say that from my experiences in explaining MML to some fellow researchers (over the years and decades) I sometimes empathise with the frustrations of the main character in "Groundhog Day". (Possibly see middle of footnote 79.)

[70] Informally, to calculate a Chaitin Omega (Ω) number [54], one feeds however many random bit strings to the UTM. Once the machine halts, it would halt and has halted regardless of any artificial suffix we might theoretically pretend to append to the string. Similarly, to calculate the probability of a UTM's becoming non-universal, one feeds however many random bit strings to the UTM. Once the machine becomes non-universal, it will remain non-universal regardless of whatever bit string suffices (or suffixes) we might append.

Putting it another way, the Chaitin Ω number of a UTM, U, is the probability that, given a particular infinitely long random bit string as input, U will halt at some point. What we might call the Wallace non-universality probability (or whatever) of a UTM, U, is the probability that, given a particular infinitely long random bit string as input, U will become non-universal at some point.

For those wanting something a bit more mathematically formal looking, I asked Chris (on Tue. 8 June 2004) whether he was okay with the suggestion that I had made to others earlier that, for each natural number c, we look at the 2^c bit strings of length c and ask for what proportion of such strings, s, does $U(s)$ remain universal and then look at the limit of this ratio as c tends to infinity. Chris sounded fine with that, but seemed to think that the intuitive stuff (above) should be sufficient.

Chris said on Mon. 31 May 2004 that he felt that (paraphrase) in most cases (of input strings), U would end up becoming non-universal. On Tue. 8 June 2004, Chris said on the phone that he thought the probability (measure) was 1 (or *unity*). I was then and still am fairly inclined to agree. From the small sample of people I've asked who've ventured a response, this seems the majority view. Clearly, this "Wallace non-universality" probability will be 1 for a given UTM iff it is 1 for all UTMs.

However relevant, I'll seize the opportunity to take this further. For each of the 2^c inputs, s, of length c to UTM, U, we asked above about the proportion that left $U(s)$ non-universal. I'll further ask what proportion are provably non-universal (including, e.g., the ones that have halted) and what proportion are non-provably non-universal. We are mainly interested in the limits of these proportions as $c \to \infty$.

[71] for more on simplest UTMs and their use in inference and prediction, see, e.g., [287, sec. 2.3.12], [115, 116], [168] and [76].

[72] Dr David M. Paganin tells me that Chris was probably referring to reciprocal spreading [230, sec. 4.4], which is apparently also well-known in optics. I am grateful to Fionn Murtagh for further pointing me to [229, pp67–68 and p432], where Chris's comparison seems to be made.

[73] as presumably unrelated and as surely independent as the comments are, I feel compelled to also point to the views of a one-time associate of A.N. Kolmogorov, namely Leonid Levin, at www.cs.bu.edu/fac/lnd/expo/qc.htm from [160, sec. 2]

[74] and, re Heisenberg (which gives us lower bounds on the likes of $\Delta E\,.\,\Delta t$ and $\Delta x\,.\,\Delta p_x$) and Chris's work - from the start of sec. 0.2.5, footnote 233 and [287, Chap. 8] - about entropy not being the arrow of time (and about entropy's changes being apparently bi-directional in time), I'd like to have asked Chris what Heisenberg says about ΔEntropy.

[75] e.g., $k \log N$ penalty to Maximum Likelihood [207] [300, sec. 7 p280 col. 2] (motivated [207, p465] by algorithmic information theory), Fisher information [208] (and Jeffreys's "prior" [287, sec. 1.15.3] [96, p217] [301, sec. 2.3]), complete coding (or completing the code) [208, 74] [301, secs. 1 and 2.2] [300, sec. 6.2], and (versions seemingly further removed from algorithmic information theory) 'model classes' [209] [301, sec. 2.1–2.2] and Normalised Maximum Likelihood (NML) [209] [301, sec. 2.3] [287, sec. 10.2].

Indeed, for those who advocate the use of a one-part non-explanation code (of length $I_0(x) = -\log\, r(x)$) [287, sec. 3.2 p154] rather than an at most slightly longer [300, sec. 6.2] [301, sec. 1.1 p332, col. 1] [302, sec. 2] [287, secs. 3.2.4 and 3.3] [291, 82] two-part explanation code, it is probably worth mentioning that Boulton and Wallace were aware in 1969 [34, 290] that a one-part non-explanation code [34] is typically slightly shorter than a two-part explanation code [290]. It is perhaps also worth mentioning that between my "*inverse learning*" specification of 25th June 1996 [102] [64, secs. 4.3 and 11] and the subsequent fleshed out publication of general(ised) (and hybrid) Bayesian nets [64, 65], I had originally incorrectly had in mind that [64, sec. 4.3] for a data-set D we needed to compare $r(D_1) \cdot Pr(H_2|D_1) \cdot Pr(D_2|D_1, H_2)$ with $r(D_2) \cdot Pr(H_1|D_2) \cdot Pr(D_1|D_2, H_1)$, where (for $i = 1, 2$) $r(D_i)$ is the marginal probability of data-set D_i (obtained by summing/integrating over hypotheses, H_i) and $I_0(D_i) = -\log\, r(D_i)$ is the one-part (MDL) code-length of data-set D_i. However, as Josh Comley pointed out to me, this MDL-like form is wrong but rather [64, sec. 4.3] what should be compared is $Pr(H_1) \cdot Pr(D_1|H_1) \cdot Pr(H_2|H_1, D_1) \cdot Pr(D_2|H_1, D_1, H_2)$

$= Pr(H_1) \cdot Pr(H_2|H_1) \cdot Pr(D_1|H_1) \cdot Pr(D_2|H_1, H_2, D_1)$

$= Pr(H_1 \& H_2)\ Pr(D_1 \& D_2|H_1 \& H_2)$ versus

$Pr(H_2) \cdot Pr(D_2|H_2) \cdot Pr(H_1|H_2, D_2) \cdot Pr(D_1|H_2, D_2, H_1)$

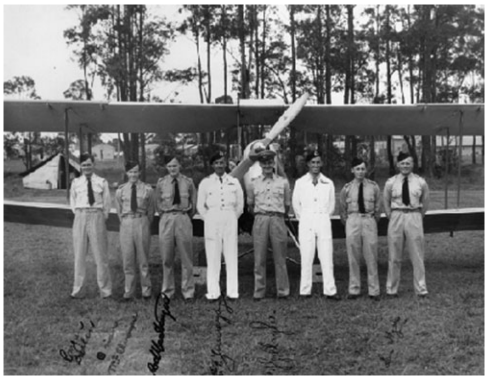


FIGURE 2. Chris (2nd from right) in front of a De Havilland Tiger Moth during his national service in the 1950s.

(this has since become the most commonly used approximation) [287, secs. 5.1 and 10.2.1][89], MML modelling of single [306, 105] and multiple [277, 284] [287, sec. 6.9] factor analysis, MML modelling of directional data [294, 295, 93, 94, 96, 97] [287, sec. 6.5] and mixture modelling of circular data [296, 297, 299, 298, 303], MML mixture modelling of Poisson distributions [296, sec. 2.3] [193, p224] [297, 299, 298, 303][90], MML mixture modelling of (Gaussian) Markov fields [283][91], MML univariate [281, 285, 113] and second-order multivariate [217] polynomial regression, minimising the expected Kullback-Leibler distance (MEKLD) as an alternative (but related) form of invariant Bayesian point estimation [79] [287, sec. 4.7], correcting an inefficient and ineffective MDL-like coding scheme [144] for a binary cut-point problem using MML so that it works properly [243][92] [287, sec. 7.3], and MML itself [290, p185, sec. 2] [34] [35, p64, col. 1] [32] [37, sec. 1, col. 1] [36] [39, sec. 1, col. 1] [291, sec. 3] [33].

Some of these ideas have been picked up on years later - e.g., at least decades later [37] or 7 or 9 or 14 years later [305], or 2 or 10 years later [268] or 13 years later [307] or 11 years later [296, sec. 2.3] - without citation, or

$= Pr(H_2) \cdot Pr(H_1|H_2) \cdot Pr(D_2|H_2) \cdot Pr(D_1|H_2, H_1, D_2)$

$= Pr(H_2 \& H_1) \cdot Pr(D_2 \& D_1|H_2 \& H_1)$, where all is now in the two-part MML form of hypothesis followed by data given hypothesis.

Whether or not he might have been one of the reviewers, Chris was well aware in the early 1990s that the Barron and Cover [24] convergence results require(d) a message of two-part MML form.

[76] and I like(d) his notion in [301, sec. 3, p335] of "using whatever 'code' or representation was used in the presentation of the raw data" (to essentially give a *de facto* 'prior' for MDL to use)

[77] also on MML and MDL, possibly see also [26]

[78] and I think he also showed, at times, a naive gullibility about some people's claims and stated intentions and undertakings

[79] Indeed, not unrelatedly, another colleague once sought to draw my attention to the relative merits of their work compared to Chris's by endeavouring to demonstrate that their work had been better funded than Chris's. When I think of slick blowhards and of Chris's not getting his deserved recognition in his life-time, I am reminded of the following [71, p21]: "But because of his . . . attire, nobody believed him. Grown-ups are like that. Fortunately . . . repeated his demonstration . . . , dressed in an elegant suit. And this time, everybody was convinced."

For those from the text around footnote 13, some of those alluded to in footnote 32 and others who wish to dismiss either MML or Chris's work in the area out of hand (possibly cf. text leading to footnote 70), please feel welcome and invited to bolster my coffers by offering a large bet at odds matching your stated conviction.

I forget whether Chris gave a time-frame for when he thought MML would complete the transition from different (and "repugnant") to having fully caught on and become mainstream, but he did once say that people would resist it (for a while) and then (and I quote to the best of my recollection), almost suddenly, ". . . and then everybody'll be doing it.".

[80] which includes more than one person who has aspired for one of Australia's prestigious Federation Fellowships

[81] and I don't think it would really matter overly much if we were to organise these people in advance into some sort of optimal arrangement

[82] Chris had what looked like a very slow style of working, but it was slow and deliberate, getting things right pretty much the first time. He told me once in our early days together that he didn't (have to) spend much time de-bugging his programs. I forget the exact amount, but it was at most a day or two and possibly much (much) less. Julie Austin (who is also mentioned in sec. 0.2.2 and in footnotes 44 and 185) had about as close to a "fly on the wall" view of watching Chris work as one could. Julie wrote shortly after Chris's death: "Chris used to sit at his desk some days and write a paper as though it was imprinted on a computer screen in his brain - there was no draft, this was the final version.".

Despite Chris's ability to write pretty much in one (uni-directional) go with at most little back-tracking, I have heard a story from one of his close colleagues that no later than the early 1990s or even earlier Chris apparently threw out a draft of well over 100 pages in length of his MML book to start again from scratch. As might be clear from footnote 20, he was polishing and refining his MML book [287] until he no longer could.

[83] Given that David Boulton [290, 34, 35, 32, 37, 36, 39, 291, 33] (see also sec. 0.3 and footnote 213) doesn't seem to have published on MML since the mid 1970s, the fact that two papers appearing in 2005 referred to [296] as a non-existent Wallace and Boulton (1994) perhaps highlights this point.

[84] this includes Chris's independently re-inventing the EM (Expectation Maximisation/Optimisation) algorithm in [290] and using it in subsequent Snob MML mixture modelling writings, deriving the general Wallace-Freeman message length approximation (with associated uncertainty in parameter estimates in terms of expected Fisher information) [305, 267] and using it in [268] and subsequent Snob MML mixture modelling writings (e.g., [296, 298, 299, 303])

[85] and from a fairly routine application of which I was able to see evidence [103, sec. 6] [78, sec. 5 (p253)] [283, sec. 4.2] that proteins apparently fold with the Helices (and Extendeds) forming first and then the "Other" turn classes forming subsequently to accommodate these structures

[86] possibly see footnote 158

[87] an unpublished manuscript of Chris Wallace's on outliers from circa 1982 - well after [290] and before [267, 305] and [268] - was discovered after Chris's death, possibly in 2007. This work assumed two classes, using the partial assignment ideas published in [268]. (My understanding from [143] is that an earlier version of [268] appeared as a technical report [266].) Much later work on outliers using total assignment (as per [290]) but discussing partial assignment [268, 296, 298, 299, 303, 7, 6] is given in [164].

[88] including on an econometric panel data problem (which I began working on with Manuel Arellano some time ago but which is as yet unpublished) which also has the amount of data per parameter bounded above and for which Maximum Likelihood and AIC are again statistically inconsistent and for which MML is again statistically consistent. More specifically, the relevant problem was and is a Gaussian stationary autoregressive panel model with a unit specific intercept (with likelihood function in [157]). The common parameters are the autoregressive coefficient and the error scale, while the nuisance parameters are the intercepts. There are a large number of units observed over a small number of time periods. I acknowledge here support from Australian Research Council (ARC) Discovery Grant DP0343650.

[89] see also [300, sec. 7, p280, col. 2]

FIGURE 3. Chris as a student.

years later (e.g., 5 years later [243]) with inaccurate and/or far less than due attribution[93].

0.2.4. Chris's (published) works

H. Akaike deservedly gained Japan's prestigious Kyoto prize for important and very influential work [8, 9] which preceded Minimum Description Length (MDL) [207][94] but came after the Wallace and Boulton (1968) MML work [290], with the relative merits of AIC and MML possibly best summarised in [82][95].

We now gloss through Chris's achievements, the sum total of which would seem to me to be at least worthy of one Turing award[96]. (I hope I don't get too much wrong and that I present things at least mainly in the correct order.)

In a nutshell (and often hidden under a bushel), Chris worked on, published in and/or observed, etc. at least the following:

- thermocouples [182, 181],
- (cosmic ray) air showers [292][97] [41, 43, 45, 42, 44, 246, 322, 248] [249][98] [257, pp236–237] (and presumably also) [166][99],
- papers in *Nature* [45, 42, 44][100] [322],
- Illiac (ILLInois Automatic Computer) [250, 251, 252, 253, 254], hardware [27][101] [317, 28, 142, 260, 310, 311, 52, 57, 56, 320, 58, 50], operating systems [27, 321, 133, 21, 201, 202, 22, 304] and organisation of a computer [257, pp227–230] [171, 261, 174, 212, 265],
- his popular computerised robotic waving-arm device [1][102],
- fast multiplier [254, 256],

[90] a version using a Poisson model and an outlier (noise) class but only total assignment is given in [164].

It is perhaps worth also mentioning here that some preliminary Gamma mixture modelling using MML has been done in [7]. And we also mention that some preliminary work on extending MML mixture modelling of uncorrelated Gaussian distributions to include correlation models for both Gaussian and t distributions is given in [6]. See also [31] for work on Dirichlet distributions.

Work in progress includes extending the current clustering of time series [172].

It is possibly also worth mentioning here that it would be nice to develop work on the modelling of circles (and ellipses) from data so that this could be extended to using mixture modelling to infer (e.g.) Olympic rings. We could proceed as follows. Letting $\mathrm{Norm}(r, n) = 1/(\pi r^2 n!/(n^{n+1})) = n^{n+1}/(\pi r^2 n!) = n^n/(\pi r^2 (n-1)!) = n^n/(\pi r^2 \Gamma(n))$, then $f((x, y)|x_0, y_0, r, n) = \mathrm{Norm}(r, n) \cdot (((x - x_0)^2 + (y - y_0)^2)/r^2)^n \ e^{-n((x-x_0)^2+(y-y_0)^2)/r^2}$ gives a likelihood function which appears to have the following desirable properties:

(i) it has a minimum at $(x, y) = (x_0, y_0)$
(ii) it peaks for x and y such that $(x - x_0)^2 + (y - y_0)^2 = r^2$, on the circumference of that circle
(iii) the value at the minimum is 0 and at the peak is $\mathrm{Norm}(r, n).\ \mathrm{e}^{-n}$, which for large n seemingly increases as $\sqrt{n}/(\sqrt{2\pi^3}\ r^2) \propto \sqrt{n}$
(iv) the normalisation constant appears to be correct, and
(v) the peak gets tighter for larger n.

We would then try to get a message length. Whether or not the above is probably a variant of a likelihood function used for a circular Hough transform, it would be nice to (re-)visit this using MML.

[91] perhaps see also Gerhard Visser's and my attempt to extend this in [240]

[92] I recall once hearing a criticism (of this work) - which I relayed to Chris - that the MML model sometimes under-estimated the number of cuts. I remember Chris pointed out that MML was beating all the rival methods quite clearly in Kullback-Leibler distance [243, sec. 8 and pp413–415], so anything inferring more cuts wasn't gaining anything from them. And then one of Chris and I pointed out that, given the nature of the problem, one could arbitrarily add an even number of cut-points arbitrarily close to one another without changing anything. Indeed, one could likewise add arbitrarily many cuts ultra-close to one of the ends again without changing anything. Bearing this in mind, one can keep the MML model but cosmetically augment it (to deal with this criticism) by doing a full Bayesian integration to infer the number of cut-points which is most probable (maximum a posteriori) given that it must (almost certainly) be at least as great as the number in the MML model (of [243]). (Possibly cf. footnote 152 and start of footnote 153.)

[93] I asked Chris once in later years how he endured the likes of this, bearing in mind my own frustrations with the sorts of events vaguely alluded to in footnote 32 (and footnote 69). He replied dryly: "I take it out on the cat", although - and one can smirk at this - after Chris's death, Judy Wallace told me that they don't (or didn't) have a cat. Gopal K. Gupta tells me that Chris didn't let this stuff get him down, although I sometimes suspect Gopal isn't always sure of this. Perhaps also see the start of sec. 0.2.2 re Chris's lack of bitterness.

[94] some readers would also recommend the slightly earlier [206]

[95] perhaps see also start of sec. 0.2.7

[96] an opinion presumably shared by others involved in supporting this earlier-mentioned nomination (from footnote 48 and text leading to it) - perhaps see also footnote 144

[97] see [73, p64], where [292] is apparently referred to as "Harry Messel's baby"

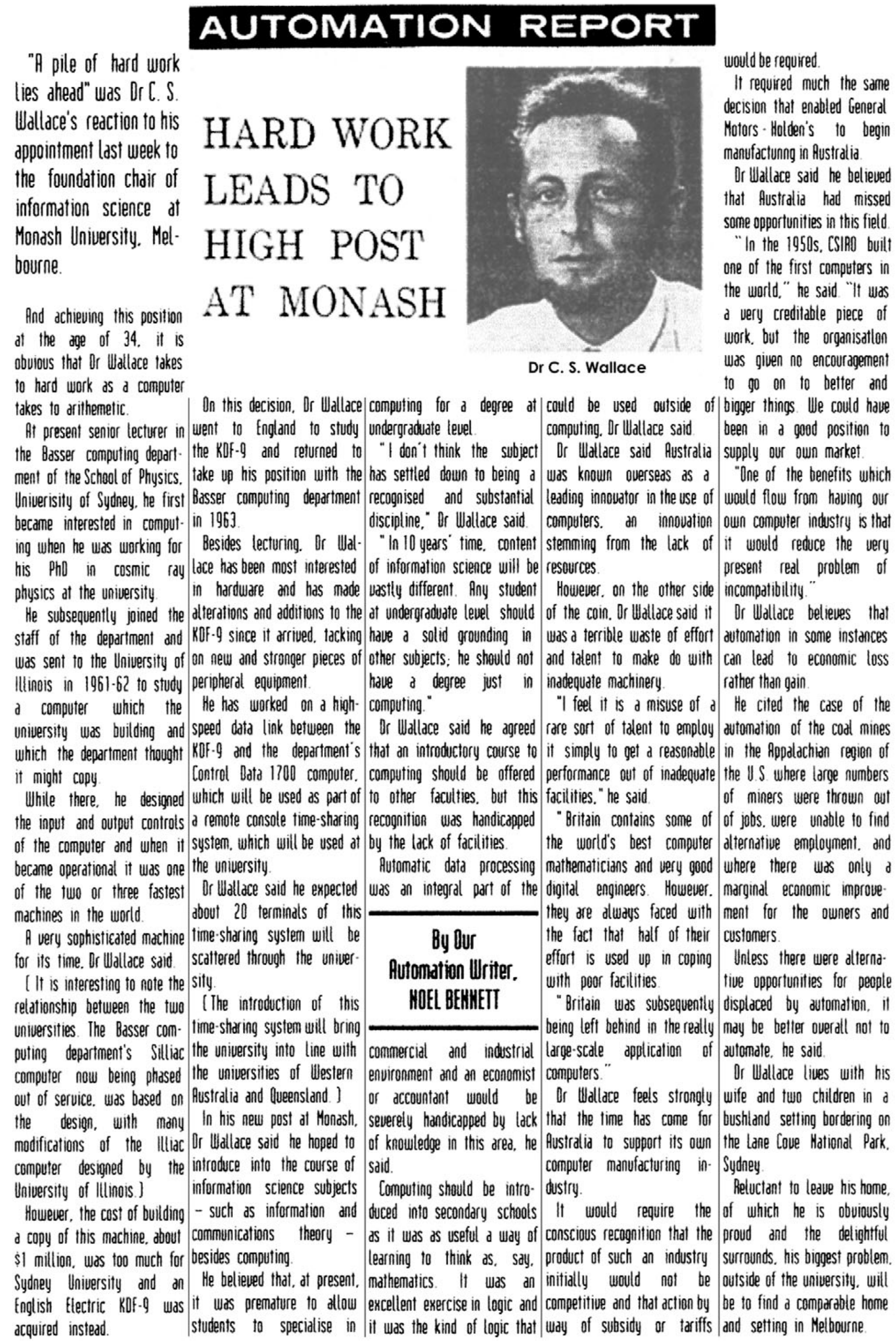
[98] see [73, p87]

[99] see also both end of sec. 1 and [40]

[100] see [73, p70] regarding background to these three Wallace papers in *Nature* and also a fourth [159]

[101] see [73, p96 and p90] and fig. 4

[102] I don't think that there was ever a publication *per se* about the waving arm. I have to confess to being one of the very last converts to appreciate it. It was (and is) a simple electronic device which kept a top-heavy "waving arm" balanced upright with the weight at the top [most of the time], something akin to balancing a hammer vertically with the base of its handle on your finger-tip. Before I met Chris, I remember seeing it at at least one Monash University Open Day and trying (not without success) to bump it so that it had to swing around to get the weight back to the top. But I largely missed the point. It (apparently) had



AUTOMATION REPORT

HARD WORK LEADS TO HIGH POST AT MONASH

Dr C. S. Wallace

"A pile of hard work lies ahead" was Dr C. S. Wallace's reaction to his appointment last week to the foundation chair of information science at Monash University, Melbourne.

And achieving this position at the age of 34, it is obvious that Dr Wallace takes to hard work as a computer takes to arithemetic.

At present senior lecturer in the Basser computing department of the School of Physics, Univerisity of Sydney, he first became interested in computing when he was working for his PhD in cosmic ray physics at the university.

He subsequently joined the staff of the department and was sent to the University of Illinois in 1961-62 to study a computer which the university was building and which the department thought it might copy.

While there, he designed the input and output controls of the computer and when it became operational it was one of the two or three fastest machines in the world.

A very sophisticated machine for its time, Dr Wallace said.

[It is interesting to note the relationship between the two universities. The Basser computing department's Silliac computer now being phased out of service, was based on the design, with many modifications of the Illiac computer designed by the University of Illinois.]

However, the cost of building a copy of this machine, about $1 million, was too much for Sydney University and an English Electric KDF-9 was acquired instead.

On this decision, Dr Wallace went to England to study the KDF-9 and returned to take up his position with the Basser computing department in 1963.

Besides lecturing, Dr Wallace has been most interested in hardware and has made alterations and additions to the KDF-9 since it arrived, tacking on new and stronger pieces of peripheral equipment.

He has worked on a high-speed data link between the KDF-9 and the department's Control Data 1700 computer, which will be used as part of a remote console time-sharing system, which will be used at the university.

Dr Wallace said he expected about 20 terminals of this time-sharing system will be scattered through the university.

[The introduction of this time-sharing system will bring the university into line with the universities of Western Australia and Queensland.]

In his new post at Monash, Dr Wallace said he hoped to introduce into the course of information science subjects – such as information and communications theory – besides computing.

He believed that, at present, it was premature to allow students to specialise in computing for a degree at undergraduate level.

"I don't think the subject has settled down to being a recognised and substantial discipline," Dr Wallace said.

"In 10 years' time, content of information science will be vastly different. Any student at undergraduate level should have a solid grounding in other subjects; he should not have a degree just in computing."

Dr Wallace said he agreed that an introductory course to computing should be offered to other faculties, but this recognition was handicapped by the lack of facilities.

Automatic data processing was an integral part of the commercial and industrial environment and an economist or accountant would be severely handicapped by lack of knowledge in this area, he said.

By Our Automation Writer, NOEL BENNETT

Computing should be introduced into secondary schools as it was as useful a way of learning to think as, say, mathematics. It was an excellent exercise in logic and it was the kind of logic that could be used outside of computing, Dr Wallace said.

Dr Wallace said Australia was known overseas as a leading innovator in the use of computers, an innovation stemming from the lack of resources.

However, on the other side of the coin, Dr Wallace said it was a terrible waste of effort and talent to make do with inadequate machinery.

"I feel it is a misuse of a rare sort of talent to employ it simply to get a reasonable performance out of inadequate facilities," he said.

"Britain contains some of the world's best computer mathematicians and very good digital engineers. However, they are always faced with the fact that half of their effort is used up in coping with poor facilities.

"Britain was subsequently being left behind in the really large-scale application of computers."

Dr Wallace feels strongly that the time has come for Australia to support its own computer manufacturing industry.

It would require the conscious recognition that the product of such an industry initially would not be competitive and that action by way of subsidy or tariffs would be required.

It required much the same decision that enabled General Motors - Holden's to begin manufactunng in Australia.

Dr Wallace said he believed that Australia had missed some opportunities in this field.

"In the 1950s, CSIRO built one of the first computers in the world," he said. "It was a very creditable piece of work, but the organisatlon was given no encouragement to go on to better and bigger things. We could have been in a good position to supply our own market.

"One of the benefits which would flow from having our own computer industry is that it would reduce the very present real problem of incompatibility."

Dr Wallace believes that automation in some instances can lead to economic loss rather than gain.

He cited the case of the automation of the coal mines in the Appalachian region of the U.S. where large numbers of miners were thrown out of jobs, were unable to find alternative employment, and where there was only a marginal economic improvement for the owners and customers.

Unless there were alternative opportunities for people displaced by automation, it may be better overall not to automate, he said.

Dr Wallace lives with his wife and two children in a bushland setting bordering on the Lane Cove National Park, Sydney.

Reluctant to leave his home, of which he is obviously proud and the delightful surrounds, his biggest problem, outside of the university, will be to find a comparable home and setting in Melbourne.

FIGURE 4. Article in "*The Australian*" newspaper (re-constructed verbatim from original) from the 1960s (ca. 1967) re Chris's appointment to Monash, also showing his views on education and social impact of automation.

- correlated round-off [255],
- ram semen [167],
- prediction from the mid-1960s of how computers might be used en masse (possibly foretelling the

very simple electronics, and - as has recently been explained to me - it was far better at balancing a top-heavy object than most humans are. See also fig. 5.

Chris also did some work with remotely (radio) controlled model aircraft, perhaps inspired by his period as a pilot (see fig. 2). At least one period in which he did this was roughly around 1980.

notion of client-server) [257, pp244–245, one-sentence final paragraph],

- research overviews [257, 264], edited volumes [289, 275, 276], and education [29, 259, 282][103],
- water (fluid) flow and reticulation (possibly [288] became [258]) [247] [73, Appendix 3, pp145–146 and p154],
- MML theory[104] [105] (starting in 1968) [290, p185, sec. 2] [34] [35, p64, col. 1] [32] [37, sec. 1 col. 1] [36] [39, sec. 1 col. 1] [291, sec. 3] [33, 196, 197, 307, 117, 118, 267] [119][106] [268, 305, 17, 16, 15, 271, 270, 18, 189, 318, 20, 19, 188, 190, 273, 306, 12, 294, 319, 13, 14, 295, 296, 312, 93, 94, 277, 66, 67, 96, 97, 186, 280, 278, 297, 99, 316, 315, 69, 68, 298, 100, 299, 101, 106, 107, 313, 152, 79, 86, 102, 284, 285, 185, 283, 153, 177, 178, 314, 241, 243, 286, 300, 301, 302, 303, 216, 108, 217, 242][107] [287][108] and applications (including) [199, 91, 92, 169, 95, 145, 77, 78, 146, 147, 293],
- Chris's notions of Bayesian bias [278, sec. 4.1][109] and false oracles [278, sec. 3] [82], his intuitive but nonetheless important proof that sampling from the Bayesian posterior is a false oracle [278, sec. 3.4] and his arguments that the Strict MML estimator[110] (which is deterministic) approximates a false oracle [278, secs. 5–7],
- the relationship between MML and Ed Jaynes's notion of maximum entropy (or MaxEnt) priors [140][111], whose idea is to be as uninformative - by assuming as little - as possible. Chris would comment that this depends upon the parameterisation [287, secs. 1.15.5 and 2.1.11][112],
- the Snob[113] program for MML mixture modelling and the various related notions of insignificant [296, sec. 5 p41 and sec. 6] [326, p211 and Conclusion] (also [203, p896]), missing data [303, sec. 2.5] [203, p896], the minimum value of the variance depending upon the measurement accuracy [296, sec. 2.1] [78, sec. 2] [146, p651] [64, sec. 9] [65, sec. 11.3.3, p270] and Expectation-Maximisation (EM)[114] [115],
- the (statistical) likelihood principle [287, sec. 5.8] [301, sec. 2.3.5], for which MML's violation is "innocent enough - a misdemeanour rather than a crime" [287, sec. 5.8, p254][116],
- nasal support [59][117],

FIGURE 5. Chris and "waving arm", ca. 1977.

[103] see [73, p76], fig. 4 and perhaps also [127]

[104] including his observations [257, pp233–234, pp237–238 and surrounds] on computers in scientific inference

[105] see notes in sec. 1 from Chris's talk of 20 Nov. 2003 going back to his early thoughts as an Hons student

[106] I don't seem to be able to locate [119]

[107] see sec. 1

[108] including explaining however many times the difference between inference/induction and prediction [225] [300, sec. 8] [302, sec. 4] [287, sec. 4.8, p208] [79, sec. 4.5, p92]

[109] possibly cf. end of sec. 1, possibly also noting Chris's "somewhat controversial" Bayesian approach in his mid-20s in the 1950s [40, sec. 4] [45, Appendix]

[110] recall Strict MML from footnote 12 and from sec. 0.2.2, and possibly also see footnotes 153, 158 and 196

[111] which is something I still get regularly asked about

[112] and, even if we can supposedly settle upon this choice of (parameterisation and) prior, I would then rhetorically ask whether we are supposed to use it to do MML, Maximum A Posteriori (MAP) or some other Bayesian approach. (I think I would generally advocate MML.)

Possibly see also [47, sec. 5] for how Chris uses Maximum Entropy in the different context of random number generation; and possibly see also [283, 240] for use of entropy in numerical approximations to the message length in problems involving Markov Random Fields (equivalently, Gibbs Random Fields); and see [287, Chap. 8], the start of sec. 0.2.5 and footnotes 74 and 233 re Chris's views on entropy not being time's arrow.

[113] my first job with Chris in May 1991 was to work on the Snob clustering program. I ventured one day early on to ask why the program was called *Snob*, probably expecting some sort of acronym. I didn't quite know what to say when he asked me in response what a snob was. With his dry (and self-effacing) humour he explained that it is so called because it makes "arbitrary class distinctions" (as also per [143, sec. 1]). (Perhaps see also [49, sec. 1] for how the name 'Walnut' came to be, and see footnote 38 for other accounts of his humour.)

[114] see footnote 84

[115] see also [143]

[116] see footnote 192 and text surrounding it.

Also, I am grateful to Claire Leslie for telling me about the conceivably related *Bus number problem*: Suppose we arrive in a new town, see only one bus and observe its number, x_{obs}, and we then wish to estimate the number of buses, θ, assuming that the θ buses are numbered $1, \ldots, \theta$. The likelihood function is $f(x|\theta) = 1/\theta$ (for $\theta \geq x$), and so Maximum Likelihood gives

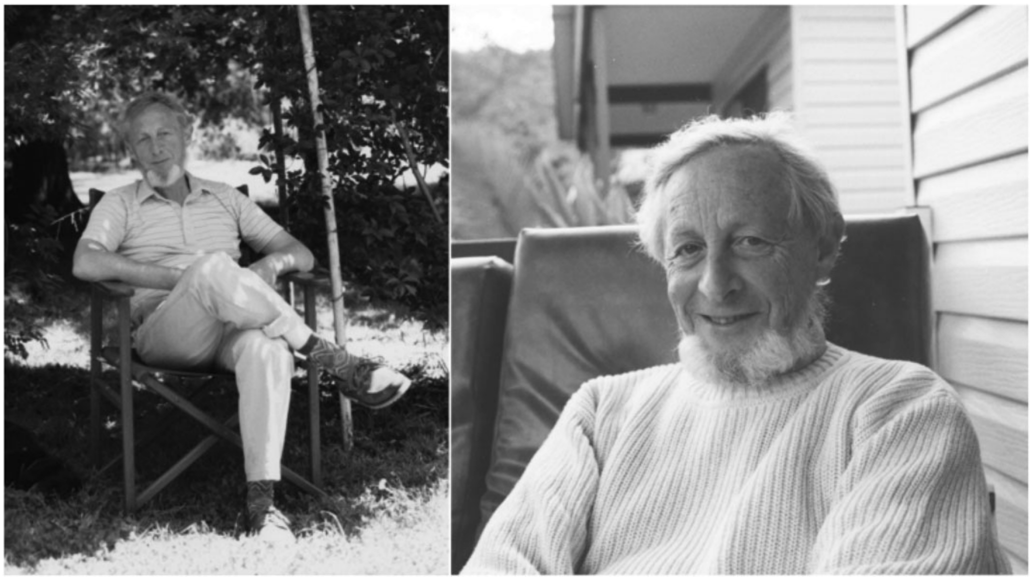


FIGURE 6. 6(a) Chris at home 6(b) Chris smiling.

- simulation of a 2-d gas [262] [287, Chap. 8] and discrete system [309] and tropospheric simulation [151],
- combinatorial (statistical) work [38] pertaining to MML,
- multi-step ODEs [128, 308, 129, 130],
- random number generation [263, 269, 272, 274, 279, 47],
- (probably not published under his name) work on sub-pixel resolution (or super-resolution),
- Chris's graduation ceremony address on (the prisoners' dilemma and) the tragedy of the commons - and the importance for society of (honesty and) co-operation [282][118] [119] [120],
- Chris's response [287, sec. 5.1.2][121] [122] to a worthy idea in Grünwald *et al.* (1998) [123] which at least on the surface appears to overly strain the Wallace-Freeman (1987) approximation, and Grünwald *et al.*'s polite private "collective" note of concession in response[123] to that[124] [125],
- Chris used to say (at least twice or) every now and then that if the statistician and the information-theorist talked for 15 minutes, they would discover MML - and if they spoke for 30 minutes then they would discover [291] Strict MML[126],
- perhaps because MML does point estimation (and invariant [291], too), some people have sought a loss function

$\hat{\theta}_{\text{MaximumLikelihood}} = x_{obs}$, which in turn hints at problems with Maximum Likelihood. It would be interesting to know what SMML, I_{1D}, and Ideal Group (IG) give. It appears to me that (for reasonable priors), given a bus number, x_{obs}, the region R for I_{1D} will have x_{obs} as its smallest member, that $\hat{\theta}_{MMLD} \geq x_{obs}$ and often $\hat{\theta}_{MMLD} > x_{obs}$. For Ideal Group (IG), I think x_{obs} will again be the smallest but typically neither the only nor the largest member of the Ideal Group, and again typically not the estimate. With SMML, we expect to see several groups of more than one member, and the largest member of any group should be the estimate for all members of that group.

Possibly see also footnote 196 and text immediately preceding it.

[117] co-authored with a now very prominent Australian

[118] a version of which appeared within the following weeks in Melbourne's *Herald Sun* newspaper

[119] I have to mention here that in a conversation - presumably after 11 Sept. 2001 - I was mentioning concerns I had for our planet (including terrorism). In his often concise (MML-like) way, Chris said that our main concern was "over-population", and he seemed pessimistic. I wish I had asked him to elaborate.

[120] In not dissimilar vein, I believe that Chris and Judy Wallace were strong supporters of the Australian Labor Party (ALP), with Judy's having obtained state pre-selection and Chris's having been a vote-counting scrutineer. In greenie mode, I also understand that Judy used to drag Chris along to tree-planting sessions before it became overly fashionable.

I will digress here and mention that, starting in 1987 and into the early 1990s (when I began working with Chris), the Australian Government began implementing its (new) policy of blurring and eliminating the distinction between Universities and Institutes of Technology. I quote Chris's comment in the staff room from the early-to-mid-1990s as accurately as I can remember it (possibly verbatim): "It'll take them 5 years to implement it, 5 years to realise that it doesn't work and 5 years to do anything about it.". While 5 years was probably an under-estimate, at least up to the change of Australian Government on 24 November 2007, Chris's prediction otherwise has seemed very accurate. Chris's social conscience was combined with insight.

[121] originally in e-mail of no later than Mon. 22 Feb. 1999 (and probably no earlier than Mon. 15 Feb. 1999) containing a 14-page postscript file entitled "Improved approximations in MML inference" dated 15 Feb. 1999. Chris also left behind a later 15-page postscript file version of this "revised 20-5-2000" with LaTeX .dvi timestamped 26 May 2000.

[122] possibly see also [301, sec. 1, p331, col. 2] and [65, sec. 11.4.3, p273]

[123] of no later than Fri 22 Feb 2002

[124] see also [122, sec. 17.4, An Apologetic Remark]

[125] and, of course, breaking an approximation is not the same as breaking a method proper (but, on this occasion, Chris manages to save the approximation anyway)

[126] recall Strict MML from (footnote 12 and) sec. 0.2.2 (and possibly also footnotes 153, 158 and 196), and then cf. [287, sec. 5.1 and p287] and possibly cf. [302, sec. 1, col. 2, last para.].

which MML (supposedly) seeks to minimise. Chris's related writings on loss functions seem to be [287, sec. 3.4.2 p189] [302, sec. 1] [300, sec. 8] - although I would also recommend [278]. My current best understanding is something along the lines that [82] MML seeks the truth[127],

- philosophy of natural language(s) [287, secs. 9.3–9.4 and 2.1.9] and of scientific language(s) [287, sec. 9.5][128] [129],
- mathematical detail [290, 291, 305, 306, 277][130], multinomial Fisher [287, sec. 5.4.1], rigour [278], etc.

Asked to give his top 20 publications, in (or no later than) June 2004 Chris gave his top 20 as [182] [42][131][256, 321, 290, 308, 291, 268, 305, 21, 272, 306, 279] [278][132] [96, 241] [300][133] [301][134] [314, 243] - and, in fact, also an additional five [189][135] [283, 258, 303, 319][136].

And then we come to (some of) Chris's unpublished work(s).

Chris (also) left behind software without corresponding publications[137] [138]. As well as Chris's work on mixture modelling and clustering under the assumption that variables within classes/components were not correlated [290, 268, 270, 296, 298, 299, 303], the early (preliminary) work on hierarchical mixture modelling [37][139] [140], his work on mixture modelling with spatial correlation [283][141] and his work on single [306] and multiple [277, 284] [287, sec. 6.9][142] latent factor analysis, Chris has (also) left behind some software to do hierarchical mixture modelling with single latent factor analysis[143] - but without an accompanying publication.

I'll add that Chris's unpublished comments to other people to assist them in their work would more than constitute many academic careers. As another case in point, he often asked salient questions when he went to conferences (e.g., Australian Statistical Congress, July 1998); and, no matter what the topic, I can't remember Chris being at a seminar - and he went to almost all of them - where he didn't ask a salient question. (I think we all came to know to wait for Chris's question(s).)

[127] which I would also say was one of Chris's hallmarks

[128] I recall discussing Goodman's "grue" paradox with Chris and (I think) some philosophers (Alan Hajek and perhaps also Bas van Fraassen) in the early 1990s, maybe 1992 or 1993. The idea is essentially that, for some time t_0, if we define *grue* to be green before t_0 and blue after t_0 and if we likewise define *bleen* to be blue before t_0 and green after t_0, then one could hypothetically argue that the terms grue and bleen are just as appropriate as blue and green (and that - after all - green is just grue before t_0 and bleen after t_0, and blue is just bleen before t_0 and grue after t_0). Here is a paraphrase of my best recollection of Chris's solution, as indicated in [65, sec. 11.4.4]. Suppose someone is growing and harvesting crops, commencing (much) before t_0 and finishing (much) after t_0. We expect the grass and certain moulds to be green, and we expect the sky and certain weeds to be blue. The notions of grue and bleen here offer at most little in return other than sometimes to require (time-based) qualification and to make the language sometimes unnecessarily cumbersome. That said, I will add that if t_0 were the time of the next expected reversal of the earth's magnetic field, then in talking on such a time-scale we have reason to disambiguate between magnetic north and geographic north in our language - as these notions are approximately equal before t_0 and approximately antipodal (for at least some time) after t_0. But the terms 'grue' and 'bleen' cost us but seem to gain us nothing. By and large, languages will develop, import, qualify and/or abbreviate terms when these terms warrant (sufficient) use.

Ray Solomonoff gives a very similar argument independently and briefly in [225], which he expands and elaborates upon in [226, sec. 5].

[129] possibly relatedly, in response to a question in Nov. 2005 by Stephen Muggleton, [192, 191] can be extended to infer (otherwise unknown) ancestral proto languages. It appears that from n observed languages we can infer no more than $n - 1$ or $n - 2$ "new" (previously unknown) proto languages. To gloss with a concise example, if we observe 4 languages - W denoted by 0101101010101010, X: 1010010110101010, Y: 1010101001011010 and Z: 1010101010100101, then we can infer an ancestral proto language A: 1010101010101010 from which we can hypothesise these 4 languages all descended.

As far as I know, and possibly inspired by the notion of decision graph(s) (see footnote 135), [192, 191] is perhaps the first work permitting many-many (evolutionary) models with many parents and many children. Re ongoing research and the future, we know how to do many-many (directed acyclic) graphs with copies and changes (but no inserts or deletes [indels]) and we know how to do one parent many children (one-many) trees with indels, but as yet (without an innocuous assumption by Peter Tan from work in progress) we have not unified/generalised to give general many-to-many graphs with indels.

[130] MML requires the minimisation of the length (namely, $-\log Pr(H) - \log Pr(D|H)$) of a two-part message the length of whose first part is $-\log Pr(H)$ and the length of whose second part is $-\log Pr(D|H)$. Some authors have instead advocated minimising $-\alpha \log Pr(H) - \log Pr(D|H)$ for various constants $\alpha > 0$. We note that $\alpha = 0$ gives Maximum Likelihood and $\alpha = 1$ gives MML. I don't know where Chris got these following ranges from (and I'm not totally and utterly sure that I've got the ranges right), but my recollection is that he said to me at least once or twice in the 1990s many years ago that $\alpha < 1$ can give statistical inconsistency (with too much weight on the likelihood), $\alpha = 1$ gives MML (and therefore statistical consistency), $1 < \alpha < 2$ gives statistical consistency (but with slower convergence - he might have said smilingly with quiet amusement that it "just gets there") and $\alpha > 2$ (with too little weight on the likelihood) is prone to statistical inconsistency.

See also not unrelated text on statistical (in)consistency in sec. 0.2.5 leading to and around footnote 163.

[131] In his notes, Chris had this paper [42] listed with M. H. Rathgeber as an additional (fourth and second last) author. But Max Brennan (from sec. 0.3.1 footnote 222, text leading to it, and [40]) has assured me that it is correct as cited.

[132] this foundational paper is one of my favourites

[133] The paper puts MML in the formal context of Kolmogorov complexity (or algorithmic information theory) [222, 223, 148, 53]. I had presented informal talks (on "Strict MML and Kolmogorov Complexity" on 14/ Nov./1996 and on "Strict MML and conditional Kolmogorov Complexity" on 6/Dec./1996) on this stuff in late 1996, largely motivated by my not seeing eye to eye with [244] and (especially) the seemingly liberal use there of the $O(1)$ terms (cf. [300, sec. 7] [65, sec. 11.3.2]). When I came to write this up (originally in 1997 for, I think, the Australian Artificial Intelligence conference), Chris identified at least one or two mistakes in what I had. Of the three papers [300, 301, 302], there is no doubt that Chris produced at least 70% or 80% of [300] and perhaps as much again of [301] (while I probably wrote most of the less significant [302]). I recall working with Chris on [300] and thinking at the time that this would be the highlight of our time together, and I guess it was. Over the years that Chris and I worked at Monash, Chris would usually leave shortly after 6pm (or very rarely as late as 6:45pm if he lost track of time), but one night we worked back until about 10:00pm to finish off [300]. This paper and our work on it is one of my fondest memories of Chris. Chris died in August 2004. When the news came through from Fionn Murtagh in April/May 2005 that [300] was at that time the *Computer Journal*'s most downloaded "full text as .pdf" article [175, sec. 3], it was the most pleasant of sweet surprises, but tinged by the fact that I couldn't share the news with Chris.

0.2.5. Some of Chris's (other) ongoing works, ideas, discussions and influences

Chris died in August 2004. Although his book [287] was not quite fully complete at his death, sources close to him say that he felt it was very near complete.

- Judy Wallace told me after Chris's death that he would have liked to have continued on from [287, Chapter 8][144] [145] and written another two papers on the arrow of time.
- I recall once probably in the mid- (or late) 1990s Chris was drawing something on a piece of paper. He was looking at designing a lens which could focus from a range of directions. He went on to say that it would have to be spherically symmetric but that the refractive index could and would vary as a function of distance from the centre. At least some days or weeks later, he told me that others had already been working on this[146].
- I recall Chris's once telling me that he felt that operating systems should do some things that they certainly weren't doing at the time and possibly still aren't. One of these was to check when files were last used and, if indications were that they wouldn't be used for a while, to compress them.
- Chris once proposed replacing the logistic likelihood regression function by a trigonometric function, such as $\sin^2$ or $\cos^2$ (of course, on a restricted range). This ends up being quite friendly, as the 2^{nd} derivative of $-\log(\sin^2)$ can be seen to be $2/(\sin^2)$ and so the expected Fisher information is a constant. (If my notes are correct, then introducing displacement/shift terms still leaves the expected Fisher information matrix both diagonal and friendly.) This makes the Wallace–Freeman (1987) approximation [305] relatively easy.
- Chris once commented that (e.g., medical) journals and other publication outlets which will not publish negative

[134] In case anyone is scouring this tribute looking for evidence of (my admitting) a technical failing in Chris, then I'll follow the above not unemotional note of tribute and appreciation from footnote 133 with about the only example I can think of of anything (even remotely?) approaching any sort of technical shortcoming. While we were working on [301], Chris discussed with me the issue of normalising the Jeffreys prior (which ultimately ended up in [301, sec. 2.3]) and the fact that he knew that there were cases when it wouldn't normalise. He came to me with what is now the infinite summation in [301, sec. 2.3.4] and asked me whether or not I could show that this summed to ∞. I showed it fairly easily using my first year undergraduate mathematics. I still have a sneaking suspicion that perhaps he just wanted to delegate that to me to make me feel more included - but, as with some other matters, I might never know.

[135] My brother tells me that he (clearly) recalls my telling him that Chris Wallace told me either that decision graphs [189, 188, 183] would make Jon Oliver famous or that (I paraphrase) decision graphs would one day be high profile - and I'll add here that it was a kind gesture of Jon Oliver's to include my name on [188], which he kindly did because [188, Fig. 3, sec. 4.2, p366] effectively came from [91]. (Decision trees (e.g., [319]) form splits, but decision graphs enable us to (re-)join, which is the meaning of one of the notions of the term "*pooling*" used by econometricians.) More recently, in the 2000s, MML decision graphs were re-visited independently and at about the same time by Peter Tan [231, 232] and by Chris Wallace. Chris did not publish his work, whereas [231] was published and later incorporated an excellent idea of William Uther's [237] to become [232]. (For what it's worth, compared to *both* C4.5 and C5 [204] on *both* real-world and artificial data-sets, *both* [231] and especially [232] have *both* a higher "right"/"wrong" predictive accuracy and a substantially better log-loss probabilistic score [see text in sec. 0.2.5 from footnote 170 maybe as far as to footnote 175 or even to footnote 176] while also having less leaf nodes. Possibly see also start of footnote 153.)

Possibly see also [233] and footnote 55 on oblique decision trees and possibly also [234] and [235] on weighted decision forests.

[136] Monash University started in (or arguably not many years before) 1961, and in 2005 it listed amongst its top 10 discoveries [194] in-vitro fertilization (IVF), work on stem cells and no less than 2 discoveries involving Chris Wallace - the Wallace multiplier [256] and Minimum Message Length inference [290]. Interestingly, the University of Melbourne lists high on its list of discoveries the high-profile and successful work on the cochlear implant by Graeme Clark, a one-time co-author of Chris's [59] as per footnote 117.

[137] such as the cut-point scheme and software for continuous-valued attributes associated with [319] which is not mentioned in the paper but which is described roughly a decade later in [156, sec. 4.1, pp123–124]. And see also part of footnote 135. Another example is *fastnorm3* from (sec. 0.3.1 [Brent] and) [47].

[138] Following on from footnote 137 immediately above, I am fairly regularly asked about schemes for (MDL or) MML cut-points. So, for those readers interested in general (MDL or) MML schemes for cut-points, here are some possible ways to proceed:

(i) the scheme described in the [319] software which is described in [156, sec. 4.1, pp123–124],

(ii) the separation ratio (in 1 dimension) [156, sec. 4.1, p124], generalised to be the amount of 'wiggle space' in [233, sec. 2.1] through which the cut can move without affecting any classification (see text around footnote 54). Note that if we have several layers of cut (on differing attributes) branching deep into the tree, then at that point we can possibly broaden the 'wiggle space' of one of the higher cuts without affecting any classification,

(iii) we could re-visit (I_{1D} or) MMLD from sec. 0.2.2, expanding the 'wiggle volume' most probably beyond the point where the classification changes, up to where the log-likelihood of the data from the boundary of the 'wiggle volume' is 1 nit more than the (prior-)weighted log-likelihood from the interior. (This will be quite CPU-intensive in 1 dimension, and will only get worse if we follow (ii) above and simultaneously consider several layers of cut.) As well as I_{1D} (or MMLD), we could possibly examine some of the other approximations from sec. 0.2.2,

(iv) possibly see the related scheme in [243] or the rather basic related scheme in [164].

Possibly see also [233, sec. 5, 1st para.] and the text in and following footnote 164 (re generalised Bayesian nets, inverse learning or implicit learning [of generative models]).

[139] which, apart from a typo, should have been cited in [98, p663, col. 1]

[140] from at least two separate conversations with researchers in Information Systems, it seems clear that being thus able to infer a hierarchical taxonomy would be of great interest and use to them

[141] possibly also see the more recent [240]

[142] I have also obtained his written notes on this, labelled "Multifactor Part 2" spanning at least the range 18/Dec./1989 to 10/10/1990.

[143] my work with Russell Edwards [105] did not do hierarchical mixture modelling, and its single latent factor analysis only used total assignment rather than the (more) correct partial assignment

[144] I am far from alone in commending this work. Another version of it was to appear in this current special issue, but has been omitted for space reasons. If the work could be experimentally tested and if Chris were alive, then

results (except in response to an earlier claim of a positive result) are introducing a bias[147].

- On areas such as "data mining" and "terabyte science", Chris once made the point that if one needs an extremely large data-set to find a pattern then the pattern probably isn't very strong. I think he then went on to advocate taking a sub-sample.
- Some time in what I think was most probably the mid-to-late 1990s, Chris commented that all the available MML software (Snob [for mixture modelling and clustering], DTreeProg [for inferring (probabilistic) decision trees], etc.) was like a (I quote) "dog's breakfast"[148]. I think that he would have been very pleased with the efforts led by Lloyd Allison to design CDMS ("Core Data Mining System") and endeavour - among other things - to integrate the existing MML software[110][149].
- Chris's work on univariate polynomial regression begun in the unpublished 1997 Royal Holloway Technical Report [281] was possibly substantially tied up in [241, 216]. (Perhaps see also [113].)
- Chris and I did sometimes discuss MML hypothesis testing - as an alternative to classical non-Bayesian tests, such as t, F and χ^2-tests[150]. I recall pretty well that Chris felt that many such classical hypothesis tests were somewhat silly - see also sec. 1. I am sure he told me once that the classical t test could be regarded as a Bayesian hypothesis test but with a rather strange prior.

In the Snob program for MML mixture modelling [270, 296, 298, 299, 303], the significance of an attribute in a component (class) is given by comparing the difference between the (negative log-likelihood) cost of encoding the values of the attribute for the things in that class using the population distribution (of that attribute) versus the length of a (more conventional) two-part MML message encoding the distribution parameters of the values of that attribute for the things in that class followed by the encoding of the (attribute values) data. Where it is cheaper to encode the attribute values given the population distribution, the attribute is deemed insignificant for the class [296, sec. 5 p41 and sec. 6] [326, p211 and Conclusion] (also [203, p896]). In fact, this is a form of hypothesis test. Where it is, say, 10 bits cheaper to encode by first inferring class distribution parameters and doing the two-part coding, this is tantamount to saying that the attribute data is $2^{10} = 1024 \approx 1000$ times more probable to have come from an inferred class distribution than from the population distribution, which would in turn correspond (in classical language) to significance at the 0.1% level[151].

And, of course, MML has been used in problems relating to model order selection - e.g., univariate polynomial regression (above) [281, 241, 216, 113], econometric time series autoregression [114], etc.[152] [153].

- Chris was well aware of the limits that undecidability gives to optimal inference even using a Universal Turing Machine [98] (and that human induction[154] and societal induction can be less than universal) and that financial markets are in general at very best non-provably efficient and typically inefficient [87] [61, sec. 1] [62, sec. 1] [287, sec. 9.1 p387][155].
- I remember first telling Chris in 1997 my conjecture [79, p93] [105, sec. 5.3] [300, p282] [303, p78] [65, sec. 11.3.1, p269] [82, sec. 8] that (in its various forms) in

(despite his most probable protestations) I would advocate it for the Nobel Prize in Physics.

[145] see [287, p vii] for acknowledgements for feedback that was to be used in this current special issue but which was used instead for [287, Chapter 8]

[146] Dr David M. Paganin tells me that this might have been Maxwell's 'fish-eye' [30, sec. 4.2.2]

[147] while this is presumably no secret, the bias nonetheless remains

[148] as perhaps we found in producing [187]

[149] including the re-implementation [10, 11] of work [64, 65] on MML Bayesian nets with decision trees in their internal nodes - see also footnote 164 and possibly part of footnote 75

[150] and I even had an Hons student work on this in 2002

[151] One could similarly test the null hypothesis (and ditto significance) that a coin is balanced and unbiased with $p_{\mathrm{Head}} = 0.5 = 1 - p_{\mathrm{Head}}$ versus the non-null hypothesis by comparing the cost of encoding the data as coming from a binomial distribution (in the case of the non-null) with that (in the case of the null) of using $p_{\mathrm{Head}} = 0.5$, where every datum will cost 1 bit and conveying results of N coin tosses will cost N bits.

[152] The notion of model order selection raises another issue. If we truly want the model order, we should do the necessary (Bayesian) integration, integrating out parameters that might otherwise have been of interest. The order of the MML model might not necessarily be the most probable model order. As an analogy, suppose we have a class of 12 students: 11 male and 1 female. Suppose that the probability the male student i wins a prize is $i/100$ and the probability that the female student wins is $1 - \sum_{i=1}^{11} i/100 = 1 - 0.66 = 0.34$. Corresponding to the question of model order, it is more probable that the prize-winner is male (0.66) than female (0.34). But the individual student most likely to win is the female. As Lloyd Allison once said (to the best of my recollection): "To get right answer, must ask right question."

[153] In such studies [281, 114], not only do we note MML's tendency to be the method most accurate in determining model order, but we also note that when MML gets the model order wrong, it has a strong tendency to err on the side of the simple(r) lower-order model (possibly cf. footnote 55 and near end of footnote 135).

I should perhaps qualify the above (very) slightly. The Strict MML estimator (recall sec. 0.2.2 and also footnotes 12, 158 and 196) typically prefers the model of highest available order [287, sec. 3.4.6] *but* - and the rest of this footnote is from my strong recollections of at least one conversation with Chris, and this also at least largely remains my intuition as I write - it does this with insignificantly small near-zero terms on the irrelevant (otherwise zero) higher-order coefficients. These higher-order near-zero terms are such that their inclusion or non-inclusion makes well less than a standard deviation's difference in the estimated function(s) over the range of the data; and changing the estimates of the higher-order coefficients to be zero would do relatively little damage to the SMML code-book and its message lengths. Furthermore (for anyone still concerned), one can go a little further and modify the MML message format so that the model order must be stated first, and this will remove the (perceived) problem while resulting in only a negligible increase in expected message length [82, sec. 7.1].

[154] see, e.g., footnote 70 and text leading up to it

[155] this is a point which countless economists without a background in computability theory have difficulty grasping, but I will add that it can certainly happen that the degree of inefficiency can be less than the transaction costs

order to get both invariance and consistency one needs either MML or a closely-related Bayesian method[156]. My conjecture was intended partly for well-specified models (where there is a true underlying model within the space of models being considered). But I know from conversations with him that Chris felt that, even when the true model was outside the space of models being considered, even then MML would converge to the model(s) within the model space as close as possible in Kullback-Leibler distance to the true model [305, p241] [287, sec. 3.4.5, p190, note use of "*agrees*"] [82, sec. 5.3.4]. So, with that influence from Chris, my conjecture was and has been always also intended for the (misspecified) case when the true model was outside the space of models being considered. Work by Grünwald and Langford [124, 125] shows some inconsistencies for some Bayesian approaches and some versions of MDL on some misspecified models, but [124, *caveat(s)* in Introduction] [125, sec. (7.4 and) 7.5] there is currently no evidence against MML. Meanwhile, we see some comparatively simple well-specified problems (such as [301, sec. 1.2 (and 1.3)] [291, sec. 6] [287, sec. 3.3.6] and Wallace's "gappy" problem [82, sec. 6.2.4]) where MML is statistically consistent but it appears that certain other methods will often be statistically inconsistent (or undefined). And, of course, as we know, there are much harder well-specified problems than these, where the amount of data per parameter to be estimated is bounded above (e.g., Neyman-Scott problem [100, 101][157], single and multiple factor analysis [306, 277] [65, sec. 11.3.1, p269] and fully-parameterised [or fully-specified] mixture modelling [303, sec. 4.3] [287, sec. 6.8] and for which MML remains consistent but those rival methods examined are not)[158]. (Note the fact that the statistical consistency is coming from the information-theoretic properties of MML and[159] the invariance is actually coming from the Bayesian prior.) But, of course, harder still will be the possibly as yet unexplored misspecified problems of a Neyman-Scott nature - i.e., where the amount of data per parameter to be estimated is bounded above[160] - but I remain optimistic for MML in these cases. Chris and I also had an interest in trying out MML and other methods on a version of the Neyman-Scott problem involving model selection between (e.g.) von Mises circular distributions and wrapped Normal distributions[161]. Last, Chris told me once probably in the late 1990s that he thought my conjecture (presumably for the well-specified case) was wrong, not because of his perceiving any case where MML would trip up but rather he was intimating that some (non-Bayesian) classical method (such as the method of moments) could possibly be tuned to give statistical consistency[162] - but I may never know.[163]

- Very related to the stuff immediately above on consistency of MML and my conjecture is my development and motivation of "implicit learning" [64, sec. 4.3 and elsewhere] [65, sec. 11.4.6] [102][164] [303, sec. 9.2

[156] although not, as originally thought [79, p93], Minimum Expected Kullback-Leibler Distance [287, sec. 4.8] [82, sec. 6.1.4]

[157] see also footnote 88

[158] For those who might - still - think that MML is the same as Maximum A Posteriori (MAP), please first see [96] [301, secs. 1.2–1.3] [302, sec. 2, col. 1] [303, secs. 2 and 6.1] [65, sec. 11.3.1] [82, sec. 5.1, coding prior] (and perhaps also [184]). Now, to highlight the difference, imagine the original Neyman-Scott problem or one of these other (Neyman-Scott-like) problems just discussed, and let us transform into a parameter space where the priors are uniform. (By the invariance of both Maximum Likelihood and MML, there is no issue here.) Now, let our continuous-valued prior be replaced by a discrete prior which is fairly dense in the continuum, made up of M Dirac delta spikes of height $1/M$ each for some large M (and we can let $M \to \infty$), and which closely resembles the continuous prior. In this case, because the prior is uniform (in this parameterisation), MAP will choose the spike of highest likelihood. By setting M sufficiently large, we can get this MAP estimate as close as we like to the continuous-parameterisation Maximum Likelihood estimate, which we know to be inconsistent. But, re Strict MML (SMML), for large M the calculations of the marginal probabilities $r(x)$ will closely approximate the values that would be obtained by integrating using the original continuous-valued prior [303, secs. 2 and 6.1] [65, sec. 11.3.1]. Hence, for sufficiently large M, we should be able to make the problem for SMML arbitrarily close to the original continuous-valued prior problem for SMML. And we believe SMML to have been consistent for this problem. In short, MAP is inconsistent (as is Maximum Likelihood), (Strict) MML is consistent (or at worst as close as we like with sufficiently large M), and so MAP and (Strict) MML are different.

[159] This is interesting for those who would criticise Bayesianism for supposedly not being able to give invariant estimates. For those wanting to see more detail of examples of invariant Bayesian point estimates, please see, e.g., [287, Chap. 3–4 and parts of Chap. 5] and footnotes 59, 61, 63 and probably also 64 (and 65).

(Wanting neither to digress overly here nor to add fuel to a fire already overly inflamed, I am grateful to Claire Leslie for pointing out that "the" classical approach is not objective, at least insofar as - as I understood her - there doesn't currently seem to be anything like a uniformly agreed upon consensus classical non-Bayesian way of analysing a data set.)

[160] e.g., a Neyman-Scott problem of a wrapped Normal nature which is misspecified as being of a von Mises nature or (the other way around) a Neyman-Scott problem of a von Mises nature which is misspecified as being of a wrapped Normal nature

[161] we both suspected that this could cause problems for Normalised Maximum Likelihood (NML)

[162] although we suspect [82, last sentence of sec. 8 Conclusion] that, even if Chris were right, such a (non-Bayesian) classical method would at best show relatively slow convergence

[163] see also footnote 130

[164] [102] was based on a talk I gave on 25th June 1996. At the time, Chris poked fun at me in public, challenging what I was doing, which probably wasn't as bad as it felt at least at the time. Over the course of the next month, (I think) I gradually and finally won him over to my way of thinking - which for me was a rare experience of which I (understandably) felt proud.

After he had been (more or less) won over, Chris often said that he'd like to see an example with (at least) one numeric (continuous) attribute and (at least) one symbolic (discrete, categorical) attribute, and eventually Josh Comley and I provided this [64, sec. 4.1 and Fig. 2] [65, Fig. 11.3], with Josh's also giving several of each in [64, Table 1 and Fig. 3]. (Mind you, while [64] was in progress, Chris weighed in with the 'sup-net' suggestion attributed to him in [64, sections 7 and 8]. I think a proper analysis of this is an important area for ongoing research.) Chris also liked my example which went on to appear in [233, sec. 5, 1st para.], where we take the notion of a decision graph [189, col. 2] [233, sec. 5][165] [166], where some (e.g., Jebara [141]) refer to "inverse learning" or "implicit learning" as generative learning.

The whole point of (generalised) Bayesian nets (or "inverse learning" or "implicit learning") was and is to generalise the model space (from *discriminative* to also include *generative*) and, paramountly and crucially[167], to use the *consistency* [102] [64, sec. 4.3, foot of p8] [65, sec. 11.3 and sec. 11.4.4, p274] of MML to ensure that the model will be converged upon. (If we use a method that can be inconsistent - and it would appear that MML is not a case in point - then we will indeed be able to get cases where [163] generative learning fails.) An example I give in [102] says that if we try to infer y as a function of x and the actual underlying source is a mixture model, then MML will infer this and will then correspondingly reveal y given x to be the corresponding cross-section of the (inferred) mixture model.

As such, the (apparent) statistical consistency of MML and the relationship between MML and Kolmogorov complexity [300] gives us that MML inverse learning (or MML generative learning, or MML implicit learning) enables us (at least in theory)

- to (further) generalise Bayesian nets to include Inductive Logic Programming (ILP), and
- when doing linear regression, to know when to regress y on x, when to regress x on y, when to do total least squares, etc.
- etc.[168] [169]
- From my early days with him - if not probably earlier - Chris was very much an advocate of log-loss probabilistic scoring, or what I think we originally referred to as probabilistic scoring.[170] It is not too hard to show that the optimal strategy in a log-loss probabilistic scoring system in the long run is to use the true probability; and Chris at least seemed to convey the impression that he thought (within a multiplicative and an additive constant) that log-loss was unique in having this property of the true probability as the long-term optimal strategy (cf. [79, sec. 3]). (As I would find out some time later, I. J. Good advocated binomial log-loss scoring in 1952 [121].) This association with Chris inspired me to use log-loss probabilistic scoring for multinomial distributions (with scores in bits) in 1993 [88, p4, Table 3][171]. Early (just before Round 3) in the 1995 Australian Football League (AFL) season, Jon Oliver pointed out to me in passing that we could have a (log-loss) probabilistic AFL competition. I somehow got Kevin Lentin to write some software to take e-mail inputs, and so this competition began [89]. The scoring system was and still is $1 + \log_2 p = \log_2(2p)$ if you're right and $1 + \log_2\ (1 - p) = \log_2\ (2(1 - p))$ if you're wrong[172]. For the 1996 season, I decided that we could have a log-loss probabilistic Gaussian scoring system on the margin of the game[173]. An entrant in the competition would choose a μ and a σ, and if the team won by x points (a loss of x can be regarded as a "win" by $-x$), then, letting $f(y|\mu,\ \sigma^2) = (1/(\sqrt{2\pi}\sigma))\ e^{-((y-\mu)^2)/(2\sigma^2)}$, their reward would be a constant[174] plus $\log_2(\int_{x-1/2}^{x+1/2} f(y)\ dy)$.

Mike Deakin came along later [72] and showed that log p (plus or minus multiplicative and/or additive constants) is by no means unique in having the property alluded to

183] (see also footnote 135) and invert it so that we can join together *output* classes to potentially reduce the number of classes by finding out that two or more classes are actually essentially parts of the same class. Current MML work has extended MML Bayesian networks to include MML decision trees in the internal nodes (to give hybrid Bayesian nets) [64, 65], but my idea which Chris liked as just described would be one of the potential gains if we generalise further to include MML decision *graphs* in the internal nodes.

[165] part of which is described in footnote 75

[166] although I'll add that Josh Comley and I had and presumably still have slightly different recollections as to which one of us first realised that our work seemed to be a variant of Bayesian networks

[167] and I think many people have failed to recognise this

[168] and, as such, it would appear that anyone advocating a "meta-learning" principle about (e.g.) when to use a decision tree versus when to use a neural network or some such, etc., would do well to familiarise themselves also with MML

[169] on the notion of "causality" (and those who like ascribing it), I rhetorically ask whether one should ascribe causality along the direction of an arrow in one of these inverse learning (or implicit learning, or generative learning) generalised Bayesian nets.

The reader should note that evolutionary tree (or graph) models (of languages) (as per footnote 129) which are insufficiently general can - and often will - lead to models where there is an arrow from a language to an ancestor; and this presumably should not be seen as causal. Indeed, for cases of families of nested model classes of increasing complexity (such as univariate polynomials of increasing degree), I have little doubt that there are cases where, given some $M \geq 3$ and given some N arbitrarily large, for some underlying model with no more than M nodes and for sufficient data generated from this underlying model, $\forall d : 2 \leq d \leq N$, the best model inferred for degree not exceeding $(d - 1)$ has at least one arrow in a different direction to the best model inferred for degree not exceeding d. Care should be taken here in attributing the notion of causality.

[170] possibly also see footnote 92

[171] and we have continued to do so [79, sec. 3] [176, Figs. 3–5] [231, sec. 4] [156, Table 2] [64, sec. 9] [232, sec. 5.1] [65, sec. 11.4.2] [233, sec. 3.1] [154, Tables 2–3] [155] [234, secs. 4.2–4.3] (and possibly also [235, sec. 4.3])

[172] For the rare but occasional draw, Chris advocated log $p(1 - p)$, but I think Graham Farr might have agreed with me on $\frac{1}{2}\log p(1 - p)$, and - rightly or wrongly - that's what we've been using all along

[173] These log-loss compression-based competitions have been running since the mid-1990s - the probabilistic competition since Round 3, 1995 and my idea of the Gaussian competition since the start of the 1996 season. John Hurst put them on the WWW in 1997, and since 1998 until no earlier than 2007 they are in the current format at www.csse.monash.edu.au/~footy. (See footnote 217.) Scores have always been and still are given in bits.

See also text of and leading up to footnote 182 and possibly see text preceding footnote 200.

[174] Chris said that he thought that this constant could have been chosen (to depend upon one of the entrants' low scores) so that the lowest score in the competition - I forget whether for each individual round or progressively throughout the season - was 0. (I presume that he meant for each individual round.) But we opted instead (rightly or wrongly) for a "real" constant (independent of any of the entrants' scores) [90, 80, 81, 85] [79, sec. 3].

above of the true probability as the long-term optimal strategy. Chris and I remained unfazed, perhaps partly because of the sorts of reasons from [232, sec. 5.1]. But I think I (might) have since uncovered the desired uniqueness property of log-loss (probabilistic bit cost) scoring that perhaps Chris might have sensed intuitively[175] [176].

- Chris's fondness for Kullback-Leibler (or K-L) distance [287, secs. 4.6–4.9 and p287] and for log-loss probabilistic scoring (as per text around footnotes 170 and 171, and surely sometimes as a numerical approximation to K-L distance) led to our reporting K-L distances in many papers [294, 96], including those involving partitions of multinomial distributions [156, sec. 4.3 and tables 1–2] [154, sec. 3 and table 1] [155]. The extension to Bayesian networks via taking a weighted sum of distances between states is trivial [234, sec. 4.2] and a numerical approximation of this with log-loss probabilistic scoring is discussed in [64, sec. 9].
- Inductive inference is part of intelligence[177], and MML tells us that inductive inference (or inductive learning)[178] is about (two-part) compression [83, sec. 2] [84, sec. 2] [218, sec. 5.2][179] [180]. Chris offered a note of support (which was arguably based on Occam's razor[181]) for the (predictive) merit of this approach of relating MML to intelligence (and predictive success)[182].
- In trying to communicate with an alien intelligence (as per the book and film "Contact"), Chris contended[183] that we would send a message which (eventually) described addition. We would endeavour to describe arithmetic (and presumably move on to Turing machines). To move on to talking about the world around us, Chris then proposed that we start sending the Lyman series, which at first would apparently look like something from elementary arithmetic (possibly like reciprocals) but which eventually draw attention to a sufficiently advanced receiver that we were talking about nuclei[184].

[175] Suppose we have a set of questions, such as on a quiz show. These questions could be multiple choice (and they could possibly also even involve (e.g.) the estimation of a Gaussian distribution). Some of the multiple choice questions could have 2 possible answers (classes), and some could have more. We could turn all of these questions into one big question, whose correct answer is the conjunction of the correct answer to all the other (sub-)questions. It only seems appropriate that a scoring system should have the property that the total score obtained from answering the (sub-)questions should be the same as the score obtained from alternatively answering the one big question. As a case in point, imagine we are trying to infer the gender and height (and age and possibly even dexterity) of one or more people. We can have one big four-valued question (Short Female, Tall Female, Short Male or Tall Male?) or possibly two two-valued questions (e.g., Short or Tall? and Female or Male?). Clearly the number of questions correctly answered will not work, as there is a score of 1 from the big question if and only if all the (sub-)questions are answered correctly. The fact that the log of a product is equal to the sum of the logs would suggest that, within a multiplicative and/or an additive constant, only log loss (probabilistic bit) scoring will work. And, of course, if we later return to try to determine (a new question of) whether young or old, there will be no problem for the log-loss (probabilistic bit cost) scoring system.

We have shown that the log loss (probabilistic bit cost) scoring system is immune to re-framing the level of detail in the classes in the specification of an individual question, and probably unique (within a multiplicative and/or an additive constant) in being so. But we can take this a little further. If question 1 was whether person 1 was short or tall, question 2 was whether person 2 was short or tall, and question 3 was whether person 3 was left-handed, right-handed or ambidexterous, then this could be framed as 2 2-class questions and a 3-class question (as it currently reads), or alternatively (re-)framed as 1 12-class question. With the multiplicative/additive constants, it would appear that log-loss is unique in giving the same answer regardless of how the problem is framed. Note that this immunity and invariance to re-framing holds regardless of how the probabilities of the answers to one question might be correlated with or conditionally dependent upon the answers to another.

In the exchange [115, sec. 2.1] [76] [116, pp173–174, Interpretation and Packaging], I am not sure how robust the system of [115, sec. 2.1] outputting the 'confidence' in a prediction as being 1 minus the second largest randomness level detected is going to be with respect to our above-discussed re-framing of problems. Nor am I sure how robust ROC and AUC (Area Under Curve) will be with respect to this re-framing.

There will perhaps always be both quiz shows and "right"/"wrong" accuracy, but in light of the above I would contend that those doing boosting and other methods focussed on "right"/"wrong" accuracy should pay more attention to log-loss probabilistic score. After all, given a 50%-50% target attribute and enough unrelated input attributes, many forms of boosting will fit spuriously resulting in no damage to the "right"/"wrong" accuracy but an abysmal (log-loss) probabilistic score. Bearing in mind that gamblers, people with major (medical) decisions and others often care at least as much about the difference between 55% and 95% than about the fact that they're both greater than 50%, probabilities are important and log-loss gives us a seemingly unique way of scoring probabilistic predictions while being invariant to re-framing. (Possibly see sec. 0.2.6 for an MML perspective on boosting.)

[176] In Dec. 2002, Luke Hope presented an innovative if not excellent idea that Bayesian priors could be incorporated into log-loss probabilistic scoring [137]. After hearing this, I was able to correct and did correct the fundamental mathematical flaw in [137] - and before the conference was over, I privately clearly and explicitly explained and showed that the correct thing to do was to keep the log-loss scoring system but to add (or subtract) a term corresponding to the entropy of the Bayesian prior [234, sec. 4.2]. (The point of doing this is that it then becomes just as hard/easy to get a good score on a hard/easy question as on an easy/hard question.) (Slightly more formally, we would take logs of ratios of probabilities, which will equal this unless both distributions have infinite entropy.)

Further to footnote 175 and the additive (and multiplicative) constants that we can add in log-loss scoring, if we now add (or subtract) the entropy of the prior (or a multiple thereof) into the scoring for our questions then, again, whether we use the original (sub-)questions or the one big question, we get the same answer. And, again, this immunity to the re-framing of the problem holds true regardless of any correlations or conditional dependencies in the prior(s).

This means that not only does log-loss probabilistic scoring appear to be unique in being immune and invariant to the detail of the (re-)framing of the problem, but that it also appears to remain so if we incorporate the Bayesian prior as per the entropy term from [234, sec. 4.2].

[177] I once put it to Chris some years ago that - not unlike Maxwell's daemon in thermodynamics - it takes 'intelligence' not just to compress (as per text around footnote 178) but also to (expend energy and) reduce (local) entropy, such as sorting using a sorting algorithm. Although my memory is hazy, I am pretty sure that Chris disagreed. (He might possibly have said that lowering the temperature will reduce entropy without requiring intelligence.) For a possibly related discussion of 'intelligence' and entropy, see sec. 0.2.7 near footnote 204.

[178] at least that and possibly also other parts of 'intelligence'

- As trivial as it both sounds and perhaps is, I do recall Chris's once telling me in passing that the (blackish) residue on the inside of his tea-cup would have been about 1 micron (or $10^{-6}\,m$) thick[185].

0.2.6. Some other stuff that Chris at least inspired

Below are some topics that Chris surely at least inspired:

- MML neural networks (e.g., [165])
- Following from near the end of footnote 175, it would be nice to (further) reconcile boosting's desire for "right"/ "wrong" accuracy (but often poor (log-loss) probabilistic performance) with MML's (statistically consistent) search for the truth[186] and its consequent good probabilistic scores. Rather than fix our Beta/Dirichlet prior [287, p47 and sec. 5.4] to have $\alpha = 1$, one possibility here is what I shall call some 'boosting priors', whose rough form [234, sec. 3.4, p598] on α could be, e.g., $3/(2\sqrt{\alpha}(1+\sqrt{\alpha})^4)$ or $(e^{-\alpha/\pi})/(\pi\sqrt{\alpha})$. The idea is simply to retain a mean of (approximately) 1 but to have a large spike near $\alpha = 0$, which in turn increases our propensity to have pure classes.

  A simpler but highly related idea is to have some weight of Dirac delta spike[187] at $\alpha = 0$. And another more elaborate option is to have a prior which is a (possibly weighted) average of those listed above.
- As well as the MML applications in sec. 0.2.4 and elsewhere, some other areas of MML applications with ongoing research include search and rescue, engineering [164], linguistics[188], climate modelling, Infrared Astronomical Satellite (IRAS) Low Resolution Spectroscopy (LRS) analysis [105], directional data [294, 96], preferred direction in universe [96, sec. 7, p225], etc. - to name but a few.

And perhaps I should also give a pointer here to sec. 0.3.1 and this issue's papers [40, 228, 143, 47, 63, 49].

0.2.7. Some other things I'd like to discuss with Chris

Here are some of the many other things[189] I would like to discuss with Chris to which I know he would have given a polite (although conceivably cheeky) but useful response:

- Recall Chris's notion of Bayesian bias [278] from sec. 0.2.4, and also recall assertions along the lines that (e.g.) MML is a better estimator than AIC (e.g., [82, 31, 294, 96, 281, 114]). Consider two estimators, A and B. Consider a space of likelihood functions, a space of parameterisations, a space of priors (to be used by Bayesian estimators and ignored by non-Bayesian estimators, or possibly instead the distribution from which test data

[179] I originally spoke to an audience of Chris Wallace and others on this topic in a talk entitled "MML and the Turing Test" on 27 Feb. 1996. I originally intended to publish it in the proceedings of a Cognitive Science conference (in Newcastle, NSW, Australia) at which I presented it in September 1997, but the proceedings were never published. See also highly related independent work in [136, 135] and a more recent survey in [158].

[180] on the issue(s) of one-part compressions and *prediction* versus two-part (MML) compressions and *inference* [300, sec. 8] [287, sec. 10.1] as per sec. 0.3.1 (Solomonoff), I put it to the reader that as interesting as it is to compress part of Wikipedia or something else (large), is it not at least as interesting or even more interesting to do a two-part compression and actually inductively infer a theory from the data? (For those curious as to how much longer the two-part code might be than the one-part code, see (expressions for $I_1 - I_0$ in), e.g., [301, sec. 1.1, p332] [287, secs. 3.2.4, 3.3 and 5.2].)

[181] possibly see (part of) footnote 18

[182] I talked to Chris Wallace on the morning on Mon 9 Mar 1998 re my Turing Test conjecture about having two hypotheses H_1 and H_2 with equal likelihoods but one being more likely a priori (and therefore a posteriori) but asking whether we should prefer the simpler (or more probable one a priori) predictively [83, sec. 5.1] [84, p105, sec. 5].

Which do we prefer?

Intuitively, we prefer the simpler (or more probable a priori) one, but it is well known that inference is not the same as prediction [300, sec. 8] [302, sec. 4] [287, secs. 4.7–4.8] (see also footnotes 108, 180 and 223), and a more complicated theory could more than conceivably have a smaller expected Kullback-Leibler (K-L) distance (from the truth) than a less complicated theory.

In looking at the (so-called) "Turing Test" (or Turing's Imitation Game), programs will probably be inferring parameter values to monitor the conversation taking place. Let us presume that the more complicated (i.e., less likely a priori) program is the one which more closely monitors/fits the conversational data.

According to Chris, Akaike argues (e.g., [8, 9]) that Expected (L of future data) = E (L of future data) = current L − const.(no. of free parameters).

So, if both programs (hypotheses) H_1 and H_2 are performing equally well so far, i.e. they have the same current L, then Akaike's argument gives us that the more likely (less complicated) one should have a higher E (L of future data) and is therefore to be preferred.

Chris surely suggested AIC because of its focus on prediction (rather than inference). Mind you, given that Chris was well aware of the merits of MML over Akaike's Information Criterion (AIC) - as later summarised in [82] - it still seems slightly ironic that Chris would use AIC to support what essentially seems like an MML argument.

Possibly see also footnote 173.

[183] in the not late 1990s, and my memory is hazy

[184] after which we could talk about matter, etc. Perhaps see text around footnote 200 and possibly recall footnote 128 (on grue).

[185] Of this same cup, Julie Austin (who is also mentioned in sec. 0.2.2 and in footnotes 44 and 82) wrote shortly after Chris's death: "I also recall that he had an English bone china mug which he used every day for his tea. It was stained brown inside but I was instructed not to scrub it clean as it added to the flavour. He used that cup to the end."

[186] see also footnote 127 and perhaps surrounding text

[187] proceeding down this path, we can see that, given a noiseless database, as the amount of data increases, database normalisation (to 1st, 2nd, 3rd, 4th and 5th Normal Form, etc.) (and the creation of new tables) will result out of applying the MML principle (somewhat akin to Inductive Logic Programming (ILP) using MML). (Where the database has some noise in it and won't give a conventional [noiseless] normalisation, possibly there might be some merit in giving some sort of probabilistic normalisation in the first part of the message, with the data then encoded in the second part of the message using these probabilities.)

[188] see footnote 129, [192, 191] and perhaps also [20]

[189] as well as, e.g., footnotes 53, 62–65, 74, 88, 90, 116, 129, 175-176 and 187

will be generated), a space of (non-negative integer) sample sizes and a space of penalty functions. Let's abbreviate this combined space to LNPPP space[190] [191], although[192] some would argue that we should also include the protocol. For each likelihood, sample size, parameterisation, prior and penalty (LNPPP), we can compare the expected penalty error(s) of A and B[193]. We can define a measure over (the permissible parts of) LNPPP-space. Using a Turing machine to generate the measure, we can generate a countably dense subset of LNPPP-space. We can then take either the measure of the region over which A has a smaller penalty than B[194] or a weighted value of the penalty over the space[195]. In published comparisons, it would seem that people sample - however they do so - from LNPPP-space.

- (Bayesian) inference when the amount of data is less than (or equal to) the number of parameters[196].
- particle physics and studying 2nd order decay (signature) data (in search of the Higgs boson)[197]
- Some experiments have random (clinical) trials. But designing an experiment for which the best MML model of the set of input variables over an appropriate model space is simply the null model[198] tells us to regard this set as random[199].
- Re MML, Kolmogorov complexity, the Turing test [83, 84, 136, 135, 218, 158][200], artificial life (ALife)[201] and even machine consciousness, I wonder about the merit of (what I shall call) (universal) redundant[202] Turing Machines - a simple case in point being a machine T_{sim} which reads in pairs, and simulates another machine, T, in the sense that T_{sim} given input 00 behaves as T given 0 and T_{sim} given input 11 behaves as T given 1. I think I'd then have wanted to go further and try to discuss with Chris relating this to Chaitin's mathematical theory of life and Adleman's DNA computing.

[190] where L is from likelihood, N is from sample size, and the three Ps are parameterisation, prior and penalty

[191] Restrictions can be made, such as, e.g., one or more of

- we only consider statistically invariant estimators,
- we only consider estimators satisfying certain forms of statistical consistency,
- we don't permit Bayesian estimators,
- our penalty is 0 at the true estimate,
- between any value in parameter space and the true estimate, there is a path in parameter space along which the penalty changes monotonically,
- the penalty function must be invariant under re-parameterisation (e.g., Kullback-Leibler distance),
- and/or etc.

[192] as per issues pertaining to the Likelihood Principle (see text preceding footnote 116), the differences between the Binomial and Negative Binomial distributions, and (stopping) protocol issues in (e.g.) [301, secs. 2.3.3–2.3.5].

[193] where test data of given sample size is generated from the likelihood - whose parameter values are sampled from the prior - over the parameterisation. The goodness of fit of A and B is respectively given by the penalty between the true value(s) and the estimated value(s).

[194] in the improbable event that A beats B everywhere in LNPPP-space, then this measure would be 1

[195] I am grateful to Jonathan Manton for discussion(s) and questions leading me to make this suggestion - as preliminary and rough as it admittedly is.

[196] such as often happens with DNA microarrays - cf. part of footnote 53 and also footnote 116. Whereas Maximum Likelihood will often have at least one way of making the likelihood as large as possible, here are some observations about the apparent behaviour of Strict MML (recall footnotes 12, 153 and 158 and sec. 0.2.2) in some cases. When there is no data, the length of the second part of the message will be 0 and so any estimator should be apparently fine for the first part of the message (which will also have 0 length). For the binomial with a uniform prior and one observation, there appear to be two ways both of equal expected message length. One way is to have only one hypothesis, $p = 0.5$. As this is the only hypothesis, it has (coding) prior probability 1 and so costs 0 bits to transmit. The second part of the message will cost 1 bit, and the total (expected) message length will be 1 bit. Alternatively, a second way is to have two hypotheses: $p = 0$ and $p = 1$, both of which have (coding) prior probability 0.5. In this way, the first part of the message will cost 1 bit and the second part of the message (using a correct probability of 100%) will have length 0, and so again the total (expected) message length will be 1 bit. For a trinomial with a uniform prior and one observation, it seems that we can get an expected code-length of log 3 by having any one of the following code-books:

- {(1/3, 1/3, 1/3)},
- {(1/2, 1/2, 0), (0, 0, 1)} and its two variations [namely, {(1/2, 0, 1/2), (0, 1, 0)} and {(0, 1/2, 1/2), (1, 0, 0)}], and
- {(1, 0 , 0), (0, 1, 0), (0, 0, 1)}.

For a multinomial with uniform prior and one observation, it would appear that we can have several equally valid code-books all of equal (expected) message length, log n. As with the trinomial, one code-book will have 1 solitary hypothesis corresponding to the probability of all states being $1/n$, and (at the other extreme) another code-book will have n hypotheses, each corresponding to a probability of 1 for that state. There will be other code-books whose entries essentially correspond to entries from code-books of multinomials of lower dimensionality.

In general, in many cases where the likelihood can always be made arbitrarily large (unlike the Bus number problem from footnote 116), choosing Maximum Likelihood estimates gives us $\forall x \; I_1(x) = I_0(x) = -\log r(x)$ with 0 length to the second part of the message, and so $I_1 = I_0$, and therefore SMML would advocate Maximum Likelihood. And, of course, given that these problems already seem harder than Neyman-Scott problems (from secs. 0.2.3 and 0.2.5 and text in and around footnote 49), adding in the misspecifications discussed in and around footnote 160 should make them harder than misspecified Neyman-Scott problems.

[197] From talking to Dr Csaba Balazs, the analysis of such data sounds like it is at least partly a problem of mixture modelling. (Possibly see sec. 0.3.1 (Brennan) and [40] re times when Chris himself was doing not totally unrelated statistical data analysis in physics.)

[198] or, more generally but less tractably, a set (given in some appropriate format) which some given Universal Turing Machine (UTM) cannot compress

[199] Digressing, Chris once told me that the best - albeit very slow - file compressor is to loop through bit strings in order of length and lexicographic order until one of these input to the given UTM re-produces the data-set. Such a file compressor is not really a secret, but it might be more interesting to insist that the compression is two-part - as per [300, sec. 4] [287, sec. 2.3.6 and onwards].

[200] see text leading up to footnote 182, and also text of and surrounding footnote 173 re our compression-based competition on Australian Football League (AFL) football going back to 1995

[201] Perhaps see text around footnote 184

[202] However new or possibly otherwise this notion might be, it gives us a way of doing effectively equivalent calculations at different speeds and a sense

- Given his [136, 135] and my [83, 84, 218] independent ideas of relating MML and Kolmogorov complexity to (at least the inductive learning part of) 'intelligence' and a quantification thereof, I have discussed with J. Hernandez Orallo the notion of quantifying the intelligence of a system of agents and endeavouring to quantify how much of this comes from the individual agents (in isolation) and how much comes from their communication[203].
- Re notions of (i) 'intelligence', (ii) laws of physics and (iii) entropy[204], can we say anything at all about roughly how much universe we might need for intelligence to happen (over time, seemingly from nowhere)? On the not unrelated notions of intelligent design (ID)[205] and miracles [287, sec. 1.2, p7], we could use MML to encode the hypothesis of an asserted "miracle"[206].
- One would presume that we can use information theory (and perhaps MML along the way) to quantify the originality of an idea or the degree of creativity of an act or design - or humour - in terms of how much information it takes to describe this in terms of what is already known[207].
- I would like to use MML - and most probably approximations from sec. 0.2.2 - to infer systems of one or more probabilistic/stochastic (partial or) ordinary (difference or) differential equations (plus at least one noise term) from (presumably noisy) data. (Some might call this "complex systems" or perhaps "computational modelling".)
- How do we best combine MML image analysis [283, 240] with MML (autoregressive) time series [114] to do sequential analysis of (changing) images - such as brain, other medical or weather images?[208]
- Chris was kindly encouraging (recalling section 0.2.2) re my trivial observation [75] that there are probability distributions bounded below with finite mean and infinite variance - and so any economic investment with such a return distribution would be valued at $-\infty$ by a mean-variance trade-off model even though it might be guaranteed to at least double in value. The Pasadena game [179] challenges decision theory by questioning the (expected) value of a game whose return is a conditionally convergent series. One can show that one can organise things - by jiggling probabilities and returns in the events of those outcomes - to get any expected pay-off one wants. If we take a distribution which is instead convergent (with no negative terms, and also bounded below) with finite mean but infinite variance [75], any finite sample averaged from this distribution would still have convergent (bounded below with no negative terms) finite mean and infinite variance.
- I don't know what Chris would have made of the seemingly artificial or contrived two-envelope paradox or "exchange paradox" [221], but on matters economic[209] I would have liked his views on my financial trading (or foreign trading) paradox, which seems to bring the otherwise arguably contrived two-envelope paradox into the real world[210].
- My "add 1" (or "elusive model") paradox[211].

in which two essentially equivalent Turing Machines might be argued to be of different ([two-part] compressive) intelligence

[203] this is perhaps where 'complex systems' meets 'intelligent systems'

[204] see (i) text around footnote 182 and possibly text around footnotes 184 and 200, (ii) [287, secs. 8.2 and 8.6] and (iii) [287, Chapter 8] respectively, and possibly also footnotes 74 and 177

[205] possibly see footnote 37

[206] Roughly, in addition to encoding data by (more) conventional means, we would need an initial bit string of whatever length to prefix the assertion of a "miracle". (The longer this initial string, the lower our prior probability of a "miracle".) Statements of (countably many) "miracles" could then be enumerated with a prefix code (possibly encoding the index/number of the "miracle") appended to this initial bit string. The question would then be asked whether any (reliably?) observed data would be encoded more concisely as coming from (more) conventional means or by the two-part encoding of (i) a particular "miracle" followed by (ii) the encoding of the observed data given said "miracle".

Not unrelatedly, some critics of Intelligent Design (ID) criticise it on the grounds that it fails to make a prediction (not already offered by conventional theories). In MML terms (and in fairness to ID), this objection could be weakened to instead request that ID (possibly encompassing a weighted mixture of conventional models and "miracles") be able to give sufficient change in probability distributions (e.g., by possibly adding some perhaps small weight to the possibility of [enumerable] "miracles") to improve probabilistic predictive performance (on new data) - see, e.g., sec. 0.2.5 in the approximate ranges of footnotes 170 to 176. For example, for some ϵ, we could have the probability distributions from various theories (or models or hypotheses), Th, given as $Th_{ID} = (1-\epsilon)Th_{Conventional} + \epsilon Th_{Miracles}$. I do not wish to be misconstrued as taking a side in this debate, argument or discussion here. I merely wish to address the terms of the discussion - saying that rather than require a new prediction (contrary to conventional theories), we might simply request a tangible improvement in predicted probabilities.

[207] the reader is welcome to inspect not unrelated ideas in [224, 219] in order to determine the originality of this idea

[208] Chris's work on (sub-pixel or) super-resolution might not be irrelevant here.

Possibly also related or relevant here is Chris's comment to me of several years ago of how, given the sizes of sampling errors, supposed weekly changes in opinion polls reported in the media are often comparable to (or even smaller than) the sampling error(s). (I have seen this very recently discussed in a statistics e-mail list.)

[209] see also text around footnote 155

[210] It goes as follows: Under reasonable assumptions, the movements of an exchange rate will move according to a distribution which is symmetric in logarithmic scale, just as likely to go from \$1 to \$1.25c as to fall to \$0.80c, just as likely to rise to \$2 as to fall to \$0.50c. Since the arithmetic mean is at least as large as the geometric mean (of \$1), it appears on the surface that buying the other currency has an expected positive return - and that this is true no matter which currency one holds. In more detail, let $x \geq 0$ and $0 \leq y < 1$ be such that $x = y/(1-y)$ and $y = x/(1+x)$. Then if the currency is equally likely to be $(1+x)$ as it is to be $1/(1+x)$ or, equivalently, equally likely to be $1/(1-y)$ as it is to be $(1-y)$, then the arithmetic mean is $1 + x^2/(2(1+x)) = 1 + y^2/(2(1-y))$, which exceeds the geometric mean of 1, only equalling it when $x = y = 0$. This apparent paradox suggests at least on the surface that, when there are only two currencies, in terms of expected return in one's own currency, everyone is best off holding as much as possible of the other currency. (This surely generalises to more than two currencies.)

- On probabilities of conditionals and conditional probabilities [161] [138, sec. 4], once we (choose our Turing machine or) specify our likelihoods and Bayesian priors, Strict MML[212] gives us a code-book with a countable set of permissible estimates and a ("coding") prior (probability) [82] of using each of these - and, as such, a (["coding"] prior) probability of conditional (probability [estimates]).

0.3. LEGACY AND FUTURE WORK

Chris's legacy perhaps begins with his works - which we have just reviewed on the previous pages - including discussions, influences and things he has inspired. It includes the ongoing works of Peter J. Tan, Daniel Schmidt, Gerhard Visser, Graham Farr, David Albrecht, Peter Tischer, Lloyd Allison and ever more (emerging) others. And it includes at least the completed works of at least Jon Oliver, Rohan Baxter, Julian Neil, Josh Comley and Leigh Fitzgibbon. And, of course, both these lists are incomplete and could already have more names on them, and both these lists are in the process of being gradually but surely enhanced by ever more emerging others. Whether or not I should mention many able Minimum Description Length (MDL) researchers in addition to the above MML researchers, certainly Chris's legacy also continues in many other areas.

If Chris is unsung (and he still is as this goes to print), then let us not forget David Boulton [290, 34, 35, 32, 38, 37, 36, 39, 291, 33] (and, e.g., [199]). I think I recall from both David Boulton and Chris words to the effect that in their early discussions leading to [290], David Boulton put forward the argument that the periodic table of elements is an example of rules followed by exceptions - at least a germ towards MML. Chris was already a Bayesian from his mid-20s in the 1950s [40, sec. 4] [45, Appendix] while David Boulton was clearly talking here in the spirit of (en)coding. Story has it that they had their separate approaches, went away and did their mathematics separately, re-convened about 6 weeks later and found out that they were doing essentially the same thing[213]. I would very much like to see one day at least some notes by David Boulton recalling some of the history of his times at Sydney and at Monash.

Thinking both of his clear impact[214] and especially the impact of the Wallace multiplier [254, 256], but weighing

[211] This is very much like a two-part MML version of the Berry paradox, using the relationship between MML and Kolmogorov complexity [300] [287, secs. 2.2–2.3] (and perhaps see also [245, 239]). Given a (Bayesian) prior specified by some distribution or (possibly universal) Turing machine, let $H_{x_1,...,x_i}$ be the MML hypothesis (or function) inferred from data $x_1,..., x_i$ and let $x_{i+1} = H_{x_1,...,x_i}\ (i+1) + 1$. (More forcefully, we could let $x'_{i+1} = 1 + \sum_{j=1}^{i} |H'_{x'_1,...,x'_j}\ (j+1)|$.) This definition by mathematical induction depends upon the choice of (Turing machine) prior and x_1. Let us abbreviate $H_{x_1,...,x_i}$ to H_i (and $H'_{x'_1,...,x'_i}$ to H'_i). It would appear that, in general, by construction, $H_i \neq H_{i+1}$ and $H'_i \neq H'_{i+1}$.

However, the thing crucial to this paradox is that, as per footnote 130, and as per the text in sec. 0.2.5 both immediately before footnote 156 and following on from it (up to about footnote 163), etc., the choice of MML (as two-part Kolmogorov complexity) as our inference method should guarantee statistical consistency. So, in general, we should find

$\exists i\ \forall j\ (j \geq i) \rightarrow (H_j = H_i)\ \&\ (H'_j = H'_i)$.

The resolution of the paradox would appear to be that the hypotheses H and H' get increasingly complicated until either

(i) if the model space is not sufficiently general, then it will never be able to converge on H or H', or

(ii) if the model space is sufficiently general (e.g., with a UTM as a prior), then H and H' are not total functions - i.e., there will be values i and j such that $H_{x_1,...,x_i}\ (i+1)$ and $H'_{x'_1,...,x'_j}\ (j+1)$ are not defined (because the relevant calculations do not terminate) and so x_{i+1} and x'_{j+1} are not defined. Or perhaps, as in (i) above, we are again forced out of the model space, this time (out of the space of computable functions) to non-computable functions.

Note that this paradox only seems to require statistical consistency in the inference method, and yet this seems to have been sufficient to force us either into the domain of algorithmic information theory (or Kolmogorov complexity) so that we can get partial (non-total) functions or indeed - as it might seem - beyond this domain and into the realm of non-computable functions.

[212] see references in text near footnote 62 and also footnotes 12, 153, 158 and 196

[213] with regard to the early part of David Boulton's and Chris's joint MML work, text from Chris Wallace's talk of 20/Nov./2003 after that in sec. 1 (which mainly described Chris's Hons year) includes: "Now David was an engineer with a good decent science background and he also knew something about information theory. And he had the bright idea that he had always thought of theories as a way of concisely representing data and his classic example was the periodic table of the elements, something which some of you may know about. [. . .] It is a way of arranging the 92 elements (or as it was originally done [when] only about 70 elements [were] known), in a table in a way which reveals a lot of relationships and similarities among the chemical elements. Ones in the same column of the table have very similar chemistries. Ones in the same row of the table have rather similar masses. And so if you know the table, you already know a lot about these elements, just represented in a simple geometric arrangement of where you have listed the element names.

So his cut on doing classification was that if you do a decent classification and describe what each class looks like and then what class each thing in your sample belongs to, you have already said a lot about the thing because if you've said it belongs to a particular class then you know it's got to look pretty much like the description of the things in the class. And so his take on it was that it was all a matter of getting the data and using some classification in some way to encode the data more concisely – data compression. Now I thought well yeah, well maybe, but look mate this is statistics, what one really wants to do is to do a proper Bayesian statistical analysis here, but taking account of how precisely you have to specify the properties of each class and thing. And the more classes you bung in, the more properties you'll have to describe, so the more precise, or the more complex, the whole description of the population becomes. Let's go and do this properly with Bayesian statistics, putting on proper prior distributions on numbers of classes and prior probabilities on parameters and so forth.

[. . .] We argued about this for some weeks, waving arms and shouting at each other, and then we decided well look this was no good, we will have to prove our points by going away and doing the maths and writing the algorithms and seeing what happens. So peace fell on the subterranean rooms where we were working for a week or two, then we came together again and I looked at his maths and he looked at my maths and we discovered that we had ended up at the same place. There really wasn't any difference between doing Bayesian analyses the way I thought one ought to do Bayesian analyses and doing data compression the way he thought you ought to do data compression. Great light dawned . . .".

[214] see, e.g., footnotes 96 and 144, but cf. footnote 79

up his personality, I once asked him whether it was due to altruism[215] that he didn't patent it. He (denied altruism and) said of himself "I was a fool". The reader can decide that one, but we loved him [131].

If you're ever getting a machine to do a calculation for you, if you're ever calling a random number generator, if you're ever looking at (one or) two or more photos of the same scene that you'd like to restore to better quality, or maybe if you're just cooling your drinks in the esky, then[216] maybe spare a thought for Chris.

Before proceeding to secs. 0.3.1 and 1, there are people to thank[217]. And, of course, I thank Chris[218].

0.3.1. This issue's authors and their papers

While Chris was alive, he and I sat and discussed possible topics and possible authors[219], but some topics in Chris's vast and diverse array[220] have been left unaddressed - at least here - for want of an author (if not also space and time). Chris personally reviewed at least most of the submissions.

A common thread throughout these papers is the (sheer) innovation and originality of Chris's approaches. Sometimes it's so elegantly novel that it's retrospectively "I wish I'd thought of that (but I didn't)" obvious. And maybe other times it's somewhat "out there" - or, to borrow from Phil Dawid [70, sec. 1] upon first learning about MML, "weird and wonderful".

I could have ordered the papers in a variety of ways. Here is a reasonable grouping that I didn't use but should at least point out. Chronologically, Chris worked on cosmic ray statistical data processing in the 1950s, as surveyed by *Brennan* [40]. Chris developed his fast tree-based multiplier in the early 1960s, as discussed by *Colon-Bonet* and *Winterrowd* [63], who take it up to Very Large Scale Integration (VLSI). Chris developed MML (an information-theoretic principle) and (the Snob program for) MML mixture modelling later in the 1960s, with *Solomonoff* [228] and *Jorgensen* and *McLachlan* [143] respectively discussing these. Chris first wrote on random number generation in the 1970s [263], and *Richard Brent* discusses [47] Chris's broader contributions to the area. *Castro*, *Pose* and *Kopp* [49] discuss the Walnut Kernel as a password-capability (based) secure operating system. But below is the grouping that I have opted for.

Max Brennan [40] worked with Chris (and others, including Mike Rathgeber) on data processing in the early cosmic ray experiments in Sydney [41, 43, 45, 42, 44, 292][221], and was likewise involved with SILLIAC [73, p64, p70][222]. Max talks of the statistical distributions, Chris's (early [Bayesian]) data analysis and the contemporary set-up and equipment. This work described here was amongst Chris's very first published work and appears to be almost certainly Chris's first published statistical work, showing the young Bayesian (but pre-MML) in his mid-20s [40, sec. 4] [45, Appendix].

Staying with (Bayesian) data analysis, **Ray Solomonoff** [228] is a pioneer (with relevant work from the early 1960s) [222, 223] now in his 80s. Chris and Ray only met a few times, and their interactions were limited but (in a word) riveting. Some references are [227, p259, col. 1], [287, sec. 10.1]. A general gist (with disclaimers) is that Ray wants to use (algorithmic) information theory (or Kolmogorov complexity) to *predict* (using many theories), but Chris wants to use (Bayesian) information theory to *infer* (using one, MML, theory [290])[223] and is concerned that Ray's use of many theories is cumbersome [300, sec. 8], that he wants to predict without necessarily understanding why [287, sec. 10.1] and that if asked to re-construct a

[215] cf. sec. 0.2.4, footnote 119, part of footnote 120, text immediately before these on the tragedy of the commons [282], and comments from fig. 4 on the social impact of automation

[216] recalling [63], [47], sec. 0.2.4 and [182, 181] respectively; and as per footnote 220

[217] I would like to thank C"Bluey"B for an early piece of encouragement very early on; and I thank TAW and also LJB for many many more. I thank FCB for always being so uplifting. I thank family, friends, those listed in sec. 0.3 and others. I thank Gopal Gupta and Jeanette Niehus for helping me compile Chris's publication list, and especially Peter Tan with both the BibTeX and the photos. I thank Torsten Seemann for doing all the unsung work behind the scenes in administering the probabilistic footy-tipping competition described around footnotes 172 and 174 in sec. 0.2.5. I thank Max Brennan for at least two photos [40], and Carlo Kopp both for assembling a photo gallery (of Chris) and for assisting me with selecting (and processing) from said gallery. I thank Chris's widow, Judy Wallace, for both her patience and her input. I am grateful for partial support from Australian Research Council (ARC) Discovery Grant DP0343650. I thank anyone I've forgotten. I especially thank the authors for their patience (not just in revising their papers), even if (perhaps sometimes) reluctant. I thank Florence Leroy of the *Computer Journal*. And perhaps most especially I thank Fionn Murtagh, the ever courteous and polite but possibly long-suffering editor-in-chief.

[218] and I thank him (partly) for his faith in me. In a signed hand-written letter to me of 12 Jan. 2004, Chris asked me to finish his book should anything happen to him (in surgery that was then planned but did not eventuate). When Chris died almost 7 months later in August 2004, his book was far nearer completion - although, from conversations with him, I do believe he would have endeavoured to include stuff on generalised (hybrid) Bayesian nets [64, 65, 102] with particular regard to my example of joining together *output* classes from footnote 164 (and see also footnote 75 - or part thereof - and text surrounding footnote 164). However, speaking as a perfectionist who would have been obliged to try as much as possible to honour Chris's higher-than-atmosphere standards, in hindsight we can be grateful that David Albrecht and Ingrid Zukerman oversaw the completion of the task (as per [287, p vii]) and promptly so. And, as I write this, trying to recount conversations, unpublished works and some ideas known only to Chris and me, not to mention dotting *i*s, crossing *t*s and etc., I realise that this foreword is now my completion (to the best of my ability and more than available time) to Chris's book.

[219] see also footnote 144 (re what is now [287, Chapter 8]) and text immediately preceding it

[220] recall start of sec. 0.2 and much of sec. 0.2.4

[221] where, as per footnote 100, they published [45, 42, 44] in *Nature*.

[222] as per sec. 0.1. More recently, Max was Chair of the Australian Research Council (ARC) for several years in the 1990s and his current positions include being Chief Scientist of South Australia.

[223] see footnote 108 (and [82, sec. 6.1.4]) for more pointers to this distinction, which many either misunderstand or seem oblivious to

damaged image Ray might offer many alternatives. Ray's response might be along the lines that the single best theory often gets overwhelming weight anyway and, if not, then we want to see the rival theories (and alternative re-constructed images) [225]. (I repeat my disclaimers[224].) As per [300, sec. 8] [287, sec. 10.1.2], I don't think Chris felt that Ray used the word "induction" as Chris would (although, then again, see [225]). The approaches are perhaps also very indirectly contrasted in [287, secs. 4.8 and 4.9] - and I'm content to be educated while they presumably agree to disagree[225].

I add that Ray raised the issue of doing his algorithmic probability (ALP) [and prediction] when resources are limited - "resource limited ALP" [225] (or Resource Bounded Probability (RBP) [226]). I am pretty sure that Chris endorsed this approach (which might seem to require modifying our Bayesian prior to reward/penalise quickly/slowly inferred functions[226] [and possibly also quickly/slowly executed functions]) [302, sec. 3] but seemed to indicate that MML was (or is) hard enough as it currently is without yet bringing in this additional (but valid) issue. I also add that Chris commented to me at least once that, even though Ray was doing prediction combining many theories, (some of) Ray's convergence results essentially came from the main, dominant (or MML) theory.[227] Possibly this [227, 228] might help with the statistical consistency discussions, questions and conjecture of sec. 0.2.5.

The seminal Wallace and Boulton (1968) [290] paper - advocating MML data analysis and statistical inference - was focussed on MML mixture modelling. **Murray Jorgensen** and **Geoff McLachlan** [143] are well-known and well-regarded within the mixture modelling community. Murray, Geoff, Chris and I also once ran a slightly animated conference panel session together [86]. Use of MML in mixture modelling is increasing [31],[228] and Murray has found some affinity with MML, while Geoff is a welcoming guy who is sometimes strident but always informed and coherent in his challenges.

Staying with things mathematical, **Richard Brent** [46, 47] was one of the people in the relatively early days at Monash with Chris.[229] Much can be found about Richard's contributions to random number generation at wwwmaths.anu.edu.au/~brent/software.html, and here [47] he also comments on Chris's. I know that Richard's correspondence with Chris [47, sec. 5] led Chris to refine his *fastnorm*, and by the time of [his *fastnorm2* (possibly) or certainly] his *fastnorm3* I know that Chris felt (chuffed) that he had ironed out all the known criticisms. Given my faith in Chris and his faith in *fastnorm3*, publication of testing and analysis of *fastnorm3* would be a welcome advance.

Chris's random number generation work also included a hardware aspect [272] [47, sec. 2] (and the Password-Capability System [21, 49] included this hardware, which is also applicable to other password-capability based operating systems [49, sec. 3.2]).

In both the abovementioned early cosmic ray data processing [40] and subsequent random number generation work [47], we see both mathematics and hardware. **Glenn Colon-Bonet** and **Paul Winterrowd** [63] were suggested to me by Chris as people who could write about computer arithmetic (and the Wallace tree-based multiplier [254, 256]) from Chris's recollection of having met a very interested Glenn at the ARITH14 conference in Adelaide back in April 1999. In this issue, Colon-Bonet and Winterrowd discuss multiplier VLSI implementations.

We can observe the progression - the evolution - of Chris's work circa 1960 on the SILLIAC with its "massive storage memory for the time of 1024 words with a word length of 40 bits" [40, sec. 1] to Chris's 1964 fast multiplier [254, 256] to (e.g.) the recent 800 × 550 μm^2 Montecito multiplier [63].

Maurice Castro, **Ronald Pose** and **Carlo Kopp** [49] all did their PhDs (and Carlo also his Master's) with Chris in related areas [48, 200, 149, 150], and all had some involvement with the Walnut Kernel [50, 48] – a secure password-capability based operating system.[230]

As with a great deal of Chris's work, the implications and applications go well beyond the initial problem he set out to solve. For the Password-Capability System and its successor, the Walnut Kernel [49], the motivation was the emerging domain of multiprocessing and distributed computer

[224] but I do have to mention here that the relationship between MML and Kolmogorov complexity (or algorithmic information theory) is discussed in [300]

[225] but one thing on which they (overtly) agree is the Bayesianism in their approaches. If we wish to compress a bit string denoting the results of coin tosses (Head = 0, Tail = 1) and we get a Head on the 1st toss, then the probability of a Head (or Tail) and its respective code length will depend upon the choice of Bayesian prior (or of UTM). See also [287, sec. 2.1.9] and, more generally, [227, secs. 1 and 2.1], [228, secs. 3.2–3.3], [300, secs. 5, 7 and 9], [302, sec. 2] and [65, sec. 11.3.2].

Possibly also see footnote 180 and text leading to it.

[226] a reasonable example of this might be *random coding* [287, sec. 4.10.1] from footnote 61. (Digressing, for those interested in "anytime"/online learning (or inference) on a growing data-set using MML, one way to proceed would be to have two or more parallel processes, including one which is using the current data set to find the MML inference however quickly and a second one which is using the last viably inferred structure from a smaller sample size and fitting parameters from the new data since then.)

[227] which perhaps at least partly accounts for the similarity in the (Wallace) inference-based and (Solomonoff) prediction-based responses to Goodman's "grue" paradox in footnote 128

[228] and footnote 197 suggests that perhaps use of mixture modelling (and related techniques) in general might increase

[229] from where Richard went on to get his Doctor of Science (D.Sc.) [46] as per footnote 11. Richard was also at least once Monash chess champion (placing equal third at about that time in the Australian championships in Hobart). (Perhaps see also [213].) Australia has had 3 Association for Computing Machinery (ACM) Fellows that I know of: Richard was the 1st and (recalling text leading to footnote 48) Chris the 2nd. In quite recent times, Richard was at Oxford; and in more recent times, Richard has deservedly gained one of Australia's prestigious Federation Fellowships.

[230] possibly also see early in sec. 0.2.5 re operating systems

systems, their control, management and security[231]. Capabilities were invented in the 1970s but [49, sec. 1] Password-Capabilities was a conceptual advance that simplified the implementation and improved the flexibility of such systems - providing simple, elegant means of authentication and type security, with great potential in preventing computer virus propagation. (The elegantly novel and retrospectively obvious notions of "money" and charging [320] [49, secs. 3.2–3.3] started as a mechanism for garbage collection, but have found application in reducing e-mail spam.[232])

And, although it's not included, at one stage we intended to include one of Chris's works here. Some eligible candidates included his strong case that entropy isn't the arrow of time [287, chap. 8],[233] his 1998 graduation ceremony address on competition, the tragedy of the commons and co-operation [282], his 1997 work on univariate polynomial regression [281] and his outstanding - but rejected - work (now a technical report) on multiple latent factor analysis [277].

Before proceeding to sec. 1, Chris said in an e-mail (of Sat 3 July 2004 15:41:40) that he didn't know that the talk would be recorded and (perhaps due to the impromptu or spontaneous nature of some of his remarks) he would like the transcript to be junked.[234] I present below an excerpt of that talk - as transcribed by Lloyd Allison, apparently alone without administrative support, listening to the tape about three times over and typing. If you want the rest[235] (bearing in mind the above), I have little doubt you can find it.

I should and do also mention Chris's taped public talk of 25th November 1992 at the Department of Computer Science, Monash University, Clayton, Australia, entitled "A model of inductive inference" [273].

Inviting you one more time to re-read the sentence leading to footnote 67, I now put the authors' papers to you. But the first words (from sec. 0.1) and the last words (from sec. 1, below, recollections of his Hons year from his talk of 20/Nov. /2003) here go to Chris.

1. MORE FROM CHRIS

"For me the whole business I suppose started to present itself to me as a bit of a problem when I was an honours student studying physics. An experimental group in the school had got very excited by a recent result that they had got. They had been measuring the intensity of a certain type of cosmic ray and discovered that the intensity seemed to go up and down according to the sidereal time. That is to say according to the attitude of the earth with respect to the distant stars. Now if that was so, it would have indicated a preferred direction for the cosmic rays to be coming from, and they were of a type which, it was thought, would not have a preferred direction. So this was quite an exciting result. And the statistical analysis of their intensity records showed that the results they had were, according to the standard statistical tests, significant to about the 1% level. That is to say, you would have expected to get such results by chance only 1% of the time. Now that is a reasonably good significance level. Most studies in medicine and psychology and the more difficult sciences jump up and down if they get a result which is even 5% significant. But it occurred to me listening to their presentation, at a seminar rather similar to this one, that the fit which their presumed sidereal-time variation gave to their data required them to say where the peak of this wave was coming to within about one hour, plus or minus one hour, say two hours. But if you fitted a curve with its peak outside this range the significance level dropped like a stone. So what?

Well then I said to myself, well they would have got just as excited had they found the peak at a different time of the day. And if the peak was sort of about two hours wide, well that would mean anywhere in twelve different positions round the clock (24-hour clock) would have been just as exciting. [...]

So the chances of their getting a result by pure chance that would have made them excited wasn't one in a hundred, it was one in a hundred divided by twelve which is only about one in eight. And then it further occurred to me, well they would have got just as excited had they discovered something which showed two peaks every day, or a peak every year, or whatever. So there were really rather a lot of hypotheses that they would have embraced if they found that they gave a fit to the data with a significance of 1%. And when you throw that in, the real significance of the result wasn't one in a hundred, or even one in eight, but something probably more like one in five. In other words there would be at least a one in five chance that purely random data would have thrown up something that would have got them excited.

Well, while accidents of the order of one in a hundred don't happen often, accidents of the order of chances one in five happen at least once a week. So I didn't think they had a decent result. I raised my little voice in the seminar and tried to explain my reservations but, as usual, was howled down; remember I was only an honours student and that's the way you've got to treat honours students. And they went on collecting data. But over the next year the effect disappeared. I didn't say I told you so because by that time I had become a research student and they have got to be even more careful about what they say.

But it got me thinking, [...] if you are talking about the fit of a hypothesis to some data, and the fit looks good, you really

[231] part of the security involved a physically random number generator [272, 47], [49, sec. 3.2].

[232] and I also like the idea [49, sec. 8.2] of reducing *tragedy of the commons* problems in this setting. Possibly cf. [282] and pointers from footnote 215.

[233] from amongst his last works. Recall start of sec. 0.2.5 as also per footnote 145 (and possibly also see footnote 112).

[234] but another colleague says Chris told that person that he (Chris) wanted the talk kept

[235] including the material from footnote 213

have to discount this by how many hypotheses you would have contemplated before choosing the one that fitted the data best. Put it another way, the more precisely you have to specify your hypothesis out of a range of possibilities in order to get a good fit to the data, the less plausible that hypothesis really becomes given the data.

So anyway, I put that thought to the back of my mind and forgot all about it. Later on as a research student I got involved in some rather complex statistical analyses of data from a different sort of cosmic ray and got introduced to Bayesian statistics because that seemed to be necessary to get more or less unbiased results."

REFERENCES

[1] Tribute to IT pioneer Chris Wallace (2004) *Monash memo*. www.monash.edu.au/news/monashmemo/stories/20041013/wallace.html, October 13.

[2] New Australian ACM Fellow (1995) *ACS Victorian Bulletin*, January/February, pp. 5–6. Australian Computer Society (ACS), Victoria, Australia.

[3] *McGraw-Hill Dictionary of Scientific and Technical Terms* (2003) (6th edn); (1st edn, 1974). McGraw-Hill, USA. http://www.accessscience.com www.accessscience.com/Dictionary/dictionary.htm.

[4] Abramson, D.A. (1982) Computer hardware to support capability based addressing in a large virtual memory. PhD Thesis, Department of Computer Science, Monash University, Australia.

[5] Agusta, Y. (2005) Minimum message length mixture modelling for uncorrelated and correlated continuous data applied to mutual funds classification. PhD Thesis, School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Australia.

[6] Agusta, Y. and Dowe, D.L. (2003) Unsupervised Learning of Correlated Multivariate Gaussian Mixture Models Using MML. *Lecture Notes in Artificial Intelligence (LNAI) 2903, Proc. 16th Australian Joint Conf. Artificial Intelligence*, pp. 477–489. Springer.

[7] Agusta, Y. and Dowe, D.L. (2003) Unsupervised Learning of Gamma Mixture Models Using Minimum Message Length. In Hamza, M.H. (ed.) *Proc. 3rd IASTED Conf. Artificial Intelligence and Applications*, Benalmadena, Spain, September, pp. 457–462. ACTA Press.

[8] Akaike, H. (1970) Statistical prediction information. *Ann. Inst. Statist. Math.*, **22**, 203–217.

[9] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov, B.N. and Csaki, F. (eds) *Proc. 2nd Int. Symp. Information Theory*, pp. 267–281.

[10] Allison, L. (2005) Models for machine learning and data mining in functional programming. *J. Funct. Program.*, **15**, 15–32, January.

[11] Allison, L. (2006) A Programming Paradigm for Machine Learning With a Case Study of Bayesian Networks. *Proc. Australian Comp. Sci. Conf. (ACSC) 2006*, January, pp. 103–111.

[12] Allison, L. and Wallace, C.S. (1993) The Posterior Probability Distribution of Alignments and its Application to Parameter Estimation of Evolutionary Trees and to Optimisation of Multiple Alignments. Technical Report CS 93/188, Department of Computer Science, Monash University, Melbourne, Australia.

[13] Allison, L. and Wallace, C.S. (1994) An Information Measure for the String to String Correction Problem with Applications. *17th Australian Comp. Sci. Conf.*, January, pp. 659–668. *Aust. Comp. Sci. Comm.*, **16**.

[14] Allison, L. and Wallace, C.S. (1994) The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *J. Mol. Evol.*, **39**, 418–430, October. Early version: 1993, TR 93/188, July, Department of Computer Science, Monash University.

[15] Allison, L., Wallace, C.S. and Yee, C.N. (1990) Induction Inference Over Macro-molecules. Technical Report 90/148, Monash University, Clayton, Victoria, Australia.

[16] Allison, L., Wallace, C.S. and Yee, C.N. (1990) Induction Inference Over Macro-molecules. *Working Notes AAAI Spring Symposium Series*, pp. 50–54. Stanford University, California, USA.

[17] Allison, L., Wallace, C.S. and Yee, C.N. (1990) When is a String Like a String? *Int. Symp. Artificial Intelligence and Mathematics*, January.

[18] Allison, L., Wallace, C.S. and Yee, C.N. (1991) Minimum Message Length Encoding, Evolutionary Trees and Multiple-Alignment. Technical Report CS 91/155, Department of Computer Science, Monash University, Melbourne, Australia.

[19] Allison, L., Wallace, C.S. and Yee, C.N. (1992) Finite-state models in the alignment of macro-molecules. *J. Mol. Evol.*, **35**, 77–89, July. Extended abstract titled: Inductive inference over macro-molecules in joint sessions at AAAI Symposium, Stanford, March 1990 on (i) Artificial Intelligence and Molecular Biology, pp. 5–9 and (ii) Theory and Application of Minimal-Length Encoding, pp. 50–54.

[20] Allison, L., Wallace, C.S. and Yee, C.N. (1992) Minimum Message Length Encoding, Evolutionary Trees and Multiple Alignment. *25th Hawaii Int. Conf. Sys. Sci.*, January, Vol. **1**, pp. 663–674. Another version in 1991: TR 91/155, Department of Computer Science, Monash University, Clayton, Victoria Australia.

[21] Anderson, M., Pose, R.D. and Wallace, C.S. (1986) A password-capability system. *Comp. J.*, **29**, 1–8.

[22] Anderson, M. and Wallace, C.S. (1988) Some comments on the implementation of capabilities. *Aust. Comp. J.*, **20**, 122–133.

[23] Anderson, M.S. (1987) Password capability system. PhD Thesis, Department of Computer Science, Monash University, Australia.

[24] Barron, A.R. and Cover, T.M. (1991) Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, **37**, 1034–1054.

[25] Baxter, R.A. (1997) Minimum message length inference: Theory and applications. PhD Thesis, Department of Computer Science, Monash University, Australia.

[26] Baxter, R.A. and Oliver, J.J. (1995) MDL and MML: Similarities and Differences. Technical Report TR 94/207, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

[27] Bennett, J.M., Wallace, C.S. and Winings, J.W. (1968) Software and Hardware Contributions to Efficient Operating on a Limited Budget. *Proc. 3rd Australian Comp. Conf.*, pp. 127–129.

[28] Bennett, J.M., Wallace, C.S. and Winings, J.W. (1968) A Grafted Multi-access Network. *IFIP Congress*, Vol. 2, pp. 917–922.

[29] Boden, A., Branagan, Davies, Gould, Graham, Kelly, Mercer, Ross and Wallace, C.S. Advancing with Science. Science Press, 1966. This was at least once a textbook for secondary schools, of which Chris Wallace wrote one chapter and contributed substantially to four others.

[30] Born, M. and Wolf, E. (1999) *Principles of Optics – Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (7th edn). Cambridge University Press.

[31] Bouguila, N. and Ziou, D. (2007) High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29**, 1716–1731, October.

[32] Boulton, D.M. (1970) Numerical classification based on an information measure. MSc Thesis, Basser Computing Department, University of Sydney, Sydney, Australia.

[33] Boulton, D.M. (1975) The information measure criterion for intrinsic classification. PhD Thesis, August, Department of Computer Science, Monash University, Clayton, Australia.

[34] Boulton, D.M. and Wallace, C.S. (1969) The information content of a multistate distribution. *J. Theor. Biol.*, **23**, 269–278.

[35] Boulton, D.M. and Wallace, C.S. (1970) A program for numerical classification. *Comp. J.*, **13**, 63–69, February.

[36] Boulton, D.M. and Wallace, C.S. (1973) A Comparison Between Information Measure Classification. *Proc. Australian & New Zealand Association for the Advancement of Science (ANZAAS) Congress*, August. Abstract.

[37] Boulton, D.M. and Wallace, C.S. (1973) An information measure for hierarchic classification. *Comp. J.*, **16**, 254–261.

[38] Boulton, D.M. and Wallace, C.S. (1973) Occupancy of a rectangular array. *Comp. J.*, **16**, 57–63.

[39] Boulton, D.M. and Wallace, C.S. (1975) An information measure for single link classification. *Comp. J.*, **18**, 236–238.

[40] Brennan, M.H. Data processing in the early cosmic ray experiments in Sydney. *Comp. J.*, **51**, 561–565.

[41] Brennan, M.H., Lehane, J.A., Malos, J., Millar, D.D., Wallace, C.S. and Winn, M.M. (1958) The Sydney Air Shower Experiment. *N. 2 del Supplemento al*, Vol. **8**, Serie X, del Nuovo Cimento, pp. 653–661.

[42] Brennan, M.H., Malos, J., Millar, D.D. and Wallace, C.S. (1958) Air showers of size greater than 10^5 particles – (2) Cerenkov radiation accompanying the showers. *Nature*, **182**, 973–977, October 11.

[43] Brennan, M.H., Malos, J., Millar, D.D. and Wallace, C.S. (1958) Cerenkov light from air showers. *N. 2 del Supplemento al*, Vol. **8**, Serie X, del Nuovo Cimento, pp. 662–664.

[44] Brennan, M.H., Millar, D.D., Rathgeber, M.H. and Wallace, C.S. (1958) Air showers of size greater than 10^5 particles – (3) comparison between the response of Geiger counters and scintillation counters. *Nature*, **182**, 1053–1054, October 18.

[45] Brennan, M.H., Millar, D.D. and Wallace, C.S. (1958) Air showers of size greater than 10^5 particles – (1) core location and shower size determination. *Nature*, **182**, 905–911, October 4.

[46] Brent, R.P. (1980) Topics in computational complexity and the analysis of algorithms. Doctor of Science (DSc) Thesis, Department of Computer Science, Monash University, Australia.

[47] Brent, R.P. Some comments on C.S. Wallace's random number generators. *Comp. J.*, **51**, 579–584.

[48] Castro, M.D. (1996) The Walnut Kernel: a password-capability based operating system. PhD Thesis, Department of Computer Science, Monash University, Australia.

[49] Castro, M.D., Pose, R.D. and Kopp, C. Password-capabilities and the Walnut Kernel. *Comp. J.*, **51**, 595–607.

[50] Castro, M.D., Pringle, G. and Wallace, C.S. (1995) The Walnut Kernel: Program Implementation Under the Walnut Kernel. Technical Report 95/230, Department of Computer Science, Monash University, Australia, August.

[51] Cathro, D. (1988) An I/O subsystem for a multiprocessor. MSc Thesis, Department of Computer Science, Monash University, Australia, 1988.

[52] Cathro, E.D., Wallace, C.S. and Anderson, M.S. (1989) An I/O Subsystem for a Multiprocessor. *Proc. 12th Australian Comp. Sci. Conf.*, February, pp. 275–285.

[53] Chaitin, G.J. (1966) On the length of programs for computing finite sequences. *J. Assoc. Comp. Mach.*, **13**, 547–569.

[54] Chaitin, G.J. (2005) *Meta Math! The Quest for Omega.* Pantheon. ISBN 0-375-42313-3 (978-0-375-42313-0).

[55] Chew, C.E. (1992) An interbus connection for a capability based multiprocessor. PhD Thesis, Department of Computer Science, Monash University, Australia.

[56] Chew, C.E. and Wallace, C.S. (1990) De-Bruijn Interconnection Networks. Technical Report CS 90/139, Department of Computer Science, Monash University, Melbourne, Australia.

[57] Chew, C.E. and Wallace, C.S. (1990) An Inter-Bus Connection for a Capability-Based Multiprocessor. Technical Report CS 90/140, Department of Computer Science, Monash University, Melbourne, Australia.

[58] Chew, C.E. and Wallace, C.S. (1991) An Inter-Bus Connection for a Capability Based Multiprocessor. *Proc. 14th Australian Comp. Sci. Conf.*

[59] Clark, G.M. and Wallace, C.S. (1970) Analysis of nasal support. *Arch. Otolaryngol.*, **92**, 118–129, August.

[60] Coles, C.J. (1978) Global register allocation in high-level microprogramming language translation. PhD Thesis, Department of Computer Science, Monash University, Australia.

[61] Collie, M.J., Dowe, D.L. and Fitzgibbon, L.J. (2005) Stock Market Simulation and Inference Technique. *5th Int. Conf. Hybrid Intelligent Systems (HIS'05)*, Rio de Janeiro, Brazil, November.

[62] Collie, M.J., Dowe, D.L. and Fitzgibbon, L.J. (2005) Trading Rule Search with Autoregressive Inference Agents. Technical Report CS 2005/174, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia.

[63] Colon-Bonet, G. and Winterrowd, P. Multiplier evolution – a family of multiplier VLSI implementations. *Comp. J.*, **51**, 585–594.

[64] Comley, J.W. and Dowe, D.L. (2003) General Bayesian Networks and Asymmetric Languages. *Proc. Hawaii Int. Conf. Statistics and Related Fields*, June 5–8.

[65] Comley, J.W. and Dowe, D.L. (2005) Minimum Message Length and Generalized Bayesian Nets with Asymmetric Languages. In Grünwald, P., Pitt, M.A. and Myung, I.J. (eds) *Advances in Minimum Description Length: Theory and Applications (MDL Handbook)*. pp. 265–294. M.I.T. Press, April, Chapter 11, ISBN 0-262-07262-9. Final camera-ready copy submitted in October 2003 (originally submitted with title: 'Minimum message length, MDL and generalised Bayesian networks with asymmetric languages').

[66] Dai, H., Korb, K.B. and Wallace, C.S. (1996) The Discovery of Causal Models with Small Samples. *Australian New Zealand Conf. Intelligent Information Systems Proc. ANZIIS96*, IEEE, Piscataway, NJ, USA, 27–30.

[67] Dai, H., Korb, K.B. and Wallace, C.S. (1996) A Study of Causal Discovery with Weak Links and Small Samples. Technical Report CS 96/298, Department of Computer Science, Monash University, Melbourne, Australia.

[68] Dai, H., Korb, K.B., Wallace, C.S. and Wu, X. (1997) A Study of Causal Discovery with Weak Links and Small Samples. Technical Report SD TR97-5, Department of Computer Science, Monash University, Melbourne, Australia.

[69] Dai, H., Korb, K.B., Wallace, C.S. and Wu, X. (1997) A Study of Causal Discovery with Weak Links and Small Samples. *Proc. 15th Int. Joint Conf. Artificial Intelligence, IJCAI 97*, pp. 1304–1309.

[70] Dawid, A.P. (1999) Discussion of the papers by Rissanen and by Wallace and Dowe. *Comp. J.*, **42**, 323–326.

[71] de Saint-Exupery, A. (1995) *The Little Prince*. Wordsworth, May. Originally in French (1944), transl. by Irene Testot-Ferry; ISBN 1-85326-158-0.

[72] Deakin, M.A.B. (2001) The characterisation of scoring functions. *J. Aust. Math. Soc.*, **71**, 135–147.

[73] Deane, J. (2006) *SILLIAC Vacuum Tube Supercomputer* (1st edn). The Science Foundation for Physics, University of Sydney and The Australian Computer Museum Society Inc., Sydney. ISBN: 1-86487-844-4.

[74] Dom, B.E. (1996) MDL Estimation for Small Sample Sizes and Its Application to Linear Regression. Technical Report RJ 10030 (90526). IBM Almaden Research Division, CA, USA.

[75] Dowe, D.L. (1997) The Contribution of Variance to Utility Functions. *Australasian Meeting of the Econometric Society (ESAM'97)*, Vol. **3** (Macroeconometrics and Finance), Melbourne, Australia, July, pp. 1–6.

[76] Dowe, D.L. (2007) Discussion following 'Hedging predictions in machine learning, A. Gammerman and V. Vovk'. *Comp. J.* **2**, 167–168.

[77] Dowe, D.L., Allison, L., Dix, T.I., Hunter, L., Wallace, C.S. and Edgoose, T. (1995) Circular Clustering by Minimum Message Length of Protein Dihedral Angles. Technical Report CS 95/237, Department of Computer Science, Monash University, Melbourne, Australia.

[78] Dowe, D.L., Allison, L., Dix, T.I., Hunter, L., Wallace, C.S. and Edgoose, T. (1996) Circular Clustering of Protein Dihedral Angles by Minimum Message Length. *Pacific Symp. Biocomputing '96*, January, pp. 242–255. World Scientific.

[79] Dowe, D.L., Baxter, R.A., Oliver, J.J. and Wallace, C.S. (1998) Point Estimation Using the Kullback–Leibler Loss Function and MML. In Wu, X., Kotagiri, R. and Korb, K. (eds) *Proc. 2nd Pacific-Asia Conf. Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, Berlin, April 15–17, LNAI, Vol. 1394, pp. 87–95. Springer.

[80] Dowe, D.L., Farr, G.E., Hurst, A.J. and Lentin, K.L. (1996) Information-Theoretic Football Tipping. *3rd Conf. Maths and Computers in Sport*, pp. 233–241.

[81] Dowe, D.L., Farr, G.E., Hurst, A.J. and Lentin, K.L. (1996) Information-Theoretic Football Tipping. Technical Report TR 96/297, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

[82] Dowe, D.L., Gardner, S. and Oppy, G.R. (2007) Bayes not bust! Why simplicity is no problem for Bayesians. *Br. J. Philos. Sci.*, **58**, 709–754, December.

[83] Dowe, D.L. and Hajek, A.R. (1997) A Computational Extension to the Turing Test. Technical Report 97/322, Depatment of Computer Science, Monash University, Australia, October.

[84] Dowe, D.L. and Hajek, A.R. (1998) A Non-Behavioural, Computational Extension to the Turing Test. *Proc. Int. Conf. Computational Intelligence & Multimedia Applications (ICCIMA'98)*, Gippsland, Australia, February, pp. 101–106.

[85] Dowe, D.L., Hurst, A.J., Lentin, K.L., Farr, G. and Oliver, J.J. (1996) Probabilistic and Gaussian Football Prediction Competitions – Monash. Artificial Intelligence in Australia Research Report, June.

[86] Dowe, D.L., Jorgensen, M., McLachlan, G. and Wallace, C.S. (1998) Information-Theoretic Estimation. In Robb, W. (ed.) *Proc. 14th Biennial Australian Statistical Conf. (ASC-14)*, Queensland, Australia, July, p. 125.

[87] Dowe, D.L. and Korb, K.B. (1996) Conceptual Difficulties with the Efficient Market Hypothesis: Towards a Naturalized Economics. *Proc. Information, Statistics and Induction in Science (ISIS)*, pp. 212–223; see also Technical Report TR 94/215, Department of Computer Science, Monash University, Australia, 1994.

[88] Dowe, D.L. and Krusel, N. (1993) A Decision Tree Model of Bushfire Activity. Technical Report TR 93/190, Department of Computer Science, Monash University, Clayton, Victoria, Australia, September.

[89] Dowe, D.L. and Lentin, K.L. (1995) Information-Theoretic Footy-Tipping Competition – Monash. Computer Science Association Newsletter (Australia), pp. 55–57, December.

[90] Dowe, D.L., Lentin, K.L., Oliver, J.J. and Hurst, A.J. (1996) An Information-Theoretic and a Gaussian Footy-Tipping Competition. FCIT Faculty Newsletter, Monash University, Australia, pp. 2–6, June.

[91] Dowe, D.L., Oliver, J.J., Allison, L., Dix, T.I. and Wallace, C.S. (1992) Learning Rules for Protein Secondary Structure Prediction. In McDonald, C., Rohl, J. and Owens, R. (eds) *Proc. 1992 Department Research Conf.* Department of Computer Science, University of Western Australia, July.

[92] Dowe, D.L., Oliver, J.J., Allison, L., Wallace, C.S. and Dix, T.I. (1992) A Decision Graph Explanation of Protein Secondary Structure Prediction. Technical Report CS 92/163, Department of Computer Science, Monash University, Melbourne, Australia.

[93] Dowe, D.L., Oliver, J.J., Baxter, R.A. and Wallace, C.S. (1995) Bayesian Estimation of the von Mises Concentration Parameter. *Proc. 15th Int. Workshop on Maximum Entropy and Bayesian Methods*, Santa Fe, July.

[94] Dowe, D.L., Oliver, J.J., Baxter, R.A. and Wallace, C.S. (1995) Bayesian Estimation of the von Mises Concentration Parameter. Technical Report CS 95/236, Department of Computer Science, Monash University, Melbourne, Australia.

[95] Dowe, D.L., Oliver, J.J., Dix, T.I., Allison, L. and Wallace, C.S. (1993) A Decision Graph Explanation of Protein Secondary Structure Prediction. *26th Hawaii Int. Conf. Sys. Sci.*, January, Vol. **1**, 669–678.

[96] Dowe, D.L., Oliver, J.J. and Wallace, C.S. (1996) MML Estimation of the Parameters of the Spherical Fisher Distribution. *Algorithmic Learning Theory, 7th International Workshop, ALT '96*, Sydney, Australia, October, *Proceedings, Lecture Notes in Artificial Intelligence*, Vol. 1160, pp. 213–227. Springer.

[97] Dowe, D.L., Oliver, J.J. and Wallace, C.S. (1996) MML Estimation of the Parameters of the Spherical Fisher Distribution. Technical Report CS 96/272, Department of Computer Science, Monash University, Melbourne, Australia.

[98] Dowe, D.L. and Oppy, G.R. (2001) Universal Bayesian inference? *Behav. Brain Sci. (BBS)*, **24**, pp. 662–663, August.

[99] Dowe, D.L. and Wallace, C.S. (1996) Resolving the Neyman–Scott Problem by Minimum Message Length (Abstract). *Proc. Sydney Int. Stat. Congress*, pp. 197–198.

[100] Dowe, D.L. and Wallace, C.S. (1997) Resolving the Neyman–Scott Problem by Minimum Message Length. *Proc. Computing Science and Statistics – 28th Symp. Interface*, Vol. 28, 614–618.

[101] Dowe, D.L. and Wallace, C.S. (1997). Resolving the Neyman–Scott Problem by Minimum Message Length. Technical Report TR No. 97/307, Department of Computer Science, Monash University, Clayton, Victoria, Australia, February; see also *Proc. Sydney Int. Stat. Congress (SISC-96)*, Sydney, pp. 197–198; *IMS Bulletin* (1996), **25**, pp. 410–411.

[102] Dowe, D.L. and Wallace, C.S. (1998) Kolmogorov Complexity, Minimum Message Length and Inverse Learning. In Robb, W.(ed.) *Proc. 14th Biennial Australian Statistical Conf. (ASC-14)*, Queensland, Australia, July, p. 144.

[103] Edgoose, T., Allison, L. and Dowe, D.L. (1998) An MML Classification of Protein Structure That Knows About Angles and Sequences. *Pacific Symp. Biocomputing '98*, January, pp. 585–596. World Scientific.

[104] Edwards, J. (2001) John Makepeace Bennett – An inspiration. *J. Res. Pract. Inf. Technol.*, **33**, 273–279, November. Special issue of JRPIT: A tribute to John Bennett.

[105] Edwards, R.T. and Dowe, D.L. (1998) Single Factor Analysis in MML Mixture Modelling. In Wu, X., Kotagiri, R. and Korb, K.B. (eds) *Proc. 2nd Pacific-Asia Conf. Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, Berlin, April 15–17, *Lecture Notes in Artificial Intelligence (LNAI)*, Vol. 1394, pp. 96–109. Springer.

[106] Farr, G.E. and Wallace, C.S. (1997) Algorithmic and Combinatorial Problems in Strict Minimum Message Length Inference. *Res. Combinatorial Algorithms, Queensland University of Technology*, Brisbane, Australia, pp. 50–58.

[107] Farr, G.E. and Wallace, C.S. (1997) The Complexity of Strict Minimum Message Length Inference. Technical Report CS 97/321, Department of Computer Science, Monash University, Melbourne, Australia.

[108] Farr, G.E. and Wallace, C.S. (2002) The Complexity of Strict Minimum Message Length Inference. *Comp. J.*, **45**, 285–292.

[109] Fitzgibbon, L.J. (2004) Message from Monte Carlo: A framework for minimum message length inference using Markov chain Monte Carlo methods. PhD Thesis, School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Australia.

[110] Fitzgibbon, L.J., Allison, L. and Comley, J.W. (2003) Probability model type sufficiency. *Proc. 4th Int. Conf. Intelligent Data Engineering and Automated Learning (IDEAL-2003)*, Hong Kong, March, *LNCS* 2690, pp. 530–534. Springer.

[111] Fitzgibbon, L.J., Dowe, D.L. and Allison, L. (2003) Bayesian Posterior Comprehension via Message from Monte Carlo. *2nd Hawaii Int. Conf. Statistics and Related Fields (HICS)*, June.

[112] Fitzgibbon, L.J., Dowe, D.L. and Allison, L. (2002) Change-Point Estimation Using New Minimum Message Length Approximations. *Lecture Notes in Artificial Intelligence (LNAI)*, 2417, *7th Pacific Rim Int. Conf. Artificial Intelligence (PRICAI)*, pp. 244–254. Springer.

[113] Fitzgibbon, L.J., Dowe, D.L. and Allison, L. (2002) Univariate Polynomial Inference by Monte Carlo Message Length Approximation. *19th Int. Conf. Machine Learning (ICML)*, pp. 147–154.

[114] Fitzgibbon, L.J., Dowe, D.L. and Vahid, F. (2004) Minimum Message Length Autoregressive Model Order Selection. *Proc. Int. Conf. Intelligent Sensors and Information Processing*, Chennai, India, January, pp. 439–444.

[115] Gammerman, A. and Vovk, V. (2007) Hedging predictions in machine learning. *Comp. J.*, **2**, 151–163.

[116] Gammerman, A. and Vovk, V. (2007) Rejoinder: Hedging predictions in machine learning. *Comp. J.*, **2**, 173–177.

[117] Georgeff, M.P. and Wallace, C.S. (1984) A General Selection Criterion for Inductive Inference. In O'Shea, T. (ed.) *Advances in Artificial Intelligence: Proc. 6th European Conf. Artificial Intelligence (ECAI-84)*, Amsterdam, September, pp. 473–482. Elsevier Science Publishers B.V., North Holland.

[118] Georgeff, M.P. and Wallace, C.S. (1984) A General Selection Criterion for Inductive Inference. Technical Report TR 44, Department of Computer Science, Monash University, Clayton, Victoria, Australia, June.

[119] Georgeff, M.P. and Wallace, C.S. (1985) Minimum information estimation of structure. In O'Shea, T. (ed.) *Advances in Artificial Intelligence*, pp. 219–228. Elsevier.

[120] Goldschlager, L.M. (1976) A microprogram debugging, tracing and timing system. MSc Thesis, Department of Computer Science, Monash University, Australia.

[121] Good, I.J. (1952) Rational decisions. *J. Roy. Statist. Soc. B*, **B 14**, 107–114.

[122] Grünwald, P.D. (2007) *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. M.I.T. Press.

[123] Grünwald, P.D., Kontkanen, P., Myllymaki, P., Silander, T. and Tirri, H. (1998) Minimum Encoding Approaches for Predictive Modeling. *Proc. 14th Int. Conf. Uncertainty in Artificial Intelligence (UAI'98)*, pp. 183–192.

[124] Grünwald, P.D. and Langford, J. (2004) Suboptimal Behavior of Bayes and MDL in Classification Under Misspecification. *COLT*, pp. 331–347.

[125] Grünwald, P.D. and Langford, J. (2007) Suboptimal behavior of Bayes and MDL in classification under misspecification. *Mach. Learn.*, **66**, 119–149, March.

[126] Gupta, G.K. (1975) New multistep methods for the solution of ordinary differential equations. PhD Thesis, Department of Computer Science, Monash University, Australia.

[127] Gupta, G.K. (2007) Computer science curriculum developments in the 1960s. *IEEE Ann. Hist. Comput.*, **29**, 40–54.

[128] Gupta, G.K. and Wallace, C.S. (1973) Some New Multistep Methods for Solving Ordinary Differential Equations. *Proc. Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conf.*, Abstract.

[129] Gupta, G.K. and Wallace, C.S. (1975) Some new multistep methods for solving ordinary differential equations. *Math. Comput.*, **29**, 489–500, April. ISSN 0025-5718.

[130] Gupta, G.K. and Wallace, C.S. (1979) A new step-size changing technique for multistep methods. *Math. Comput.*, **33**, 125–138, January.

[131] Gupta, G.K., Zukerman, I. and Albrecht, D.W. (2004) Obituaries: Australia's Inspiring Leader in the Computing Revolution – Christopher Wallace Computer Scientist. *The Age*, Melbourne, Australia, October 1, Friday, p. 9.

[132] Hagan, R.A. (1977) Virtual memory hardware for a HP2100A minicomputer. MSc Thesis, Department of Computer Science, Monash University, Australia.

[133] Hagan, R.A. and Wallace, C.S. (1977) A virtual memory system for the HP2100A. *Comput. Archit. News*, **6**, 5–13.

[134] Herman, P.M. (1975) Data flow analysis in program testing. Master's Thesis, Department of Computer Science, Monash University, Australia.

[135] Hernández-Orallo, J. (2000) Beyond the Turing test. *J. Logic Lang. Inf.*, **9**, 447–466.

[136] Hernández-Orallo, J. and Minaya, N. (1998) A Formal Definition of Intelligence Based on an Intensional Variant of Kolmogorov Complexity. *Proc. Int. Symp. Engineering of Intelligent Systems*, pp. 146–163, ICSC Press.

[137] Hope, L.R. and Korb, K. (2002) Bayesian Information Reward. *Lecture Notes in Artificial Intelligence*, Vol. 2557, pp. 272–283, ISSN: 0302-9743. Springer, Berlin, Germany.

[138] Hutter, M. (2006) On the Foundations of Universal Sequence Prediction. In Cai, J.-Y., Cooper, S.B. and Li, A. (eds) *Proc. 3rd Annual Conf. Theory and Applications of Models of Computation (TAMC'06), LNCS*, Vol. 3959, 408–420. Springer.

[139] Jansen, A.R., Dowe, D.L. and Farr, G.E. (2000) Inductive Inference of Chess Player Strategy. *6th Pacific Rim Int. Conf. Artificial Intelligence (PRICAI), Lecture Notes in Artificial Intelligence*, Vol. 1886, pp. 61–71.

[140] Jaynes, E.T. (2003) *Probability Theory: The Logic of Science*. Cambridge University Press.

[141] Jebara, A. (2003) *Machine Learning: Discriminative and Generative*. Kluwer Academic Publishers, Norwell, MA, USA.

[142] Jessup, A.M. and Wallace, C.S. (1968) A cheap graphic input device. *Aust. Comp. J.*, **1**, 95–96.

[143] Jorgensen, M.A. and McLachlan, G.J. Wallace's approach to unsupervised learning: the Snob program. *Comp. J.*, **51**, 571–578.

[144] Kearns, M., Mansour, Y., Ng, A.Y. and Ron, D. (1997) An experimental and theoretical comparison of model selection methods. *Mach. Learn.*, **27**, 7–50.

[145] Kissane, D.W., Bloch, S., Burns, W.I., Patrick, J.D., Wallace, C.S. and McKenzie, D.P. (1994) Perceptions of family functioning and cancer. *Psycho-oncology*, **3**, 259–269.

[146] Kissane, D.W., Bloch, S., Dowe, D.L., Snyder, R.D., Onghena, P., McKenzie, D.P. and Wallace, C.S. (1996) The Melbourne family grief study, I: perceptions of family functioning in bereavement. *Am. J. Psychiatry*, **153**, 650–658, May.

[147] Kissane, D.W., Bloch, S., Onghena, P., McKenzie, D.P., Snyder, R.D. and Dowe, D.L. (1996) The Melbourne family

grief study, II: psychosocial morbidity and grief in bereaved families. *Am. J. Psychiatry*, **153**, 659–666, May.

[148] Kolmogorov, A.N. (1965) Three approaches to the quantitative definition of information. *Probl. Inf. Transm.*, **1**, 4–7.

[149] Kopp, C. (1997) An I/O and stream inter-process communications library for a password capability system. Master's Thesis, Department of Computer Science, Monash University, Australia, August.

[150] Kopp, C. (2000) The properties of high capacity microwave airborne ad hoc networks. PhD Thesis, School of Computer Science and Software Engineering, Monash University, Australia, October.

[151] Kopp, C. and Wallace, C.S. (2004) TROPPO – A Tropospheric Propagation Simulator. Technical Report 2004/161, School of Computer Science and Software Engineering, Monash University, Australia, November.

[152] Korb, K.B. and Wallace, C.S. (1997) In Search of the Philosopher's Stone: Remarks on Humphreys and Freedman's Critique of Causal Discovery. Technical Report CS 97/315, Department of Computer Science, Monash University, Melbourne, Australia.

[153] Korb, K.B. and Wallace, C.S. (1999) In search of the philosopher's stone: Remarks on Humphreys and Freedman's critique of causal discovery. *Br. J. Philos. Sci.*, **48**, 543–553. TR 97/315, March 1997, Department of Computer Science, Monash University, Australia.

[154] Kornienko, L., Albrecht, D.W. and Dowe, D.L. (2005) A Preliminary MML Linear Classifier Using Principal Components for Multiple Classes. *Proc. 18th Aust. Joint Conf. Artificial Intelligence (AI'2005), Lecture Notes in Artificial Intelligence (LNAI)*, Sydney, Australia, December, Vol. 3809, 922–926. Springer.

[155] Kornienko, L., Albrecht, D.W. and Dowe, D.L. (2005) A Preliminary MML Linear Classifier Using Principal Components for Multiple Classes. Technical Report CS 2005/179, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia.

[156] Kornienko, L., Dowe, D.L. and Albrecht, D.W. (2002) Message Length Formulation of Support Vector Machines for Binary Classification – A Preliminary Scheme. *Lecture Notes in Artificial Intelligence (LNAI)*, *Proc. 15th Aust. Joint Conf. Artificial Intelligence*, Vol. 2557, pp. 119–130. Springer-Verlag.

[157] Lancaster, A. (2002) Orthogonal parameters and panel data. *Rev. Econom. Stud.*, **69**, 647–666.

[158] Legg, S. and Hutter, M. (2007) Universal intelligence: a definition of machine intelligence. *Minds Mach.*, **17**, 391–444.

[159] Lehane, J.A., Millar, D.D. and Rathgeber, M.H. (1958) The distribution of nuclear-active particles and of mu-mesons. *Nature*, **182**, 1699–1704, December 20.

[160] Levin, L.A. (2003) The tale of one-way functions. *Probl. Inf. Transm. (Problemy Peredachi Informatsii)*, **39**, 92–103.

[161] Lewis, D.K. (1976) Probabilities of conditionals and conditional probabilities. *Philos. Rev.*, **85**, 297–315, July.

[162] Liang, S. (1990) The design of an intelligent memory module. MSc Thesis, Department of Computer Science, Monash University, Australia.

[163] Long, P. and Servedio, R. (2006) Discriminative Learning can Succeed Where Generative Learning Fails. *19th Annual Conf. Learning Theory*, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

[164] Maheswaran, T., Sanjayan, J.G., Dowe, D.L. and Tan, P.J. (2006) MML Mixture Models of Heterogeneous Poisson Processes With Uniform Outliers for Bridge Deterioration. *Lecture Notes in Artificial Intelligence (LNAI)*, *Proc. 19th Aust. Joint Conf. Artificial Intelligence*, December, pp. 322–331. Springer.

[165] Makalic, E., Allison, L. and Dowe, D.L. (2003) MML Inference of Single-Layer Neural Networks. *Proc. 3rd IASTED Int. Conf. Artificial Intelligence and Applications*, September. pp. 636–642. see also Technical Report TR 2003/142, CSSE, Monash University, Australia, October.

[166] Malos, J., Millar, D.D. and Wallace, C.S. (1962) Cerenkov radiation from E.A.S. *J. Phys. Soc. Japan*, **17** (Suppl. A-III).

[167] Martin, I.C.A. and Wallace, C.S. (1965) Impedance change frequency of diluted ram semen recorded on a digital scaler. *J. Reprod. Fertil.*, **10**, 425–437.

[168] Martin-Löf, P. (1966) The definition of random sequences. *Inf. Control*, **9**, 602–619.

[169] McKenzie, D.P., McGorry, P.D., Wallace, C.S., Low, L.H., Copolov, D.L. and Singh, B.S. (1993) Constructing a minimal diagnostic decision tree. *Methods Inf. Med.*, **32**, 161–166.

[170] McQuade, T.J. (1971) Storage and retrieval of medical information. PhD Thesis, Department of Computer Science, Monash University, Australia.

[171] McQuade, T.J., Race, D. and Wallace, C.S. (1969) Storage and Retrieval of Medical Information. *Proc. 4th Aust. Computer Conf.*, pp. 539–544.

[172] Molloy, S.B., Albrecht, D.W., Dowe, D.L. and Ting, K.M. (2006) Model-based Clustering of Sequential Data. *Proc. 5th Annual Hawaii Int. Conf. Statistics, Mathematics and Related Fields*, January.

[173] Moneypenny, W.H. (1977) Cobol for an educational computing system. MSc Thesis, Department of Computer Science, Monash University, Australia.

[174] Montgomery, A.Y. and Wallace, C.S. (1972) Evaluation and Design of Random Access Files. *Proc. 5th Aust. Computer Conf.*, May, pp. 142–150.

[175] Murtagh, F. (2005) Editorial. *Comp. J.*, **48**, 381.

[176] Needham, S.L. and Dowe, D.L. (2001) Message Length as an Effective Ockham's Razor in Decision Tree Induction. *Proc. 8th Int. Workshop on Artif. Intelligence and Statistics (AI + STATS 2001)*, January, pp. 253–260.

[177] Neil, J.R., Wallace, C.S. and Korb, K.B. (1999) Learning Bayesian Networks With Restricted Causal Interactions. In Laskey, K.B. and Prade, H. (eds) *Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI-99)*, San Francisco, California, July 30–August 1, pp. 486–493. Morgan Kaufmann.

[178] Neil, J.R., Wallace, C.S. and Korb, K.B. (1999) Bayesian Networks with Non-interacting Causes. Technical Report 1999/28, School of Computer Science & Software Engineering, Monash University, Australia, September.

[179] Nover, H. and Hajek, A.R. (2004) Vexing expectations. *Mind*, **113**, 237–249.

[180] O'Brien, B.J. (1959) Nuclear emulsion studies of the heavy nuclei of the primary cosmic radiation. PhD Thesis, Department of Physics, University of Sydney, Australia.

[181] O'Brien, B.J. and Wallace, C.S. (1958) Ettingshausen effect and thermomagnetic cooling. *J. Appl. Phys.*, **29**, 1010–1012, July.

[182] O'Brien, B.J., Wallace, C.S. and Landecker, K. (1956) The cascading of Peltier-couples for thermo-electric cooling. *J. Appl. Phys.*, **27**, 820–823, July.

[183] Oliver, J.J. (1993) Decision Graphs – An Extension of Decision Trees. *Proc. 4th Int. Workshop on Artificial Intelligence and Statistics*, pp. 343–350. Extended version available as TR 173, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

[184] Oliver, J.J. and Baxter, R.A. (1994) MML and Bayesianism: Similarities and Differences. Technical Report TR 94/206, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

[185] Oliver, J.J., Baxter, R.A. and Wallace, C.S. (1998) Minimum Message Length Segmentation. In Wu, X., Kotagiri, R. and Korb, K.B. (eds) *Proc. 2nd Pacific-Asia Conf. Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, Berlin, April 15–17, *LNAI*, Vol. 1394, pp. 222–233. Springer.

[186] Oliver, J.J., Baxter, R.A. and Wallace, C.S. (1996) Unsupervised Learning Using MML. *Proc. 13th Int. Conf. Machine Learning*, pp. 364–372. Morgan Kaufmann.

[187] Oliver, J.J. and Dowe, D.L. (1995) Using Unsupervised Learning to Assist Supervised Learning. *Proc. 8th Australian Joint Conf. Artificial Intelligence*, November, pp. 275–282. TR 95/235, Department of Computer Science, Monash University, Australia, September.

[188] Oliver, J.J., Dowe, D.L. and Wallace, C.S. (1992) Inferring Decision Graphs Using the Minimum Message Length Principle. *Proc. 1992 Aust. Joint Conf. Artificial Intelligence*, September, pp. 361–367.

[189] Oliver, J.J. and Wallace, C.S. (1991) Inferring Decision Graphs. *Proc. 12th Int. Joint Conf. Artificial Intelligence (IJCAI-91), Workshop 8*, January.

[190] Oliver, J.J. and Wallace, C.S. (1992) Inferring Decision Graphs. Technical Report CS 92/170, Department of Computer Science, Monash University, Melbourne, Australia.

[191] Ooi, J.N. and Dowe, D.L. (2005) Inferring Phylogenetic Graphs for Natural Languages Using MML. Technical Report TR 2005/178, School of Computer Science and Software Engineering, Monash University, Clayton, Victoria, Australia.

[192] Ooi, J.N. and Dowe, D.L. (2005) Inferring Phylogenetic Graphs of Natural Languages Using Minimum Message Length. *CAEPIA 2005 (11th Conf. Spanish Association for Artificial Intelligence)*, November, Vol. **1**, pp. 143–152.

[193] Papp, E., Dowe, D.L. and Cox, S.J.D. (1993) Spectral Classification of Radiometric Data Using an Information Theory Approach. *Proc. Advanced Remote Sensing Conf.*, UNSW, Sydney, Australia, July, pp. 223–232.

[194] Parry, L. (2005) Midas touch. *The Age*, Melbourne, Australia, Education section, p. 6, Monday 20 June. www.TheAge.com.au; www.monash.edu.au/policy/midas.htm.

[195] Patrick, J.D. (1978) An information measure comparative analysis of megalithic geometries. PhD Thesis, Department of Computer Science, Monash University, Australia.

[196] Patrick, J.D. and Wallace, C.S. (1977) Stone Circles: A Comparative Analysis of Megalithic Geometry. *Proc. 48th Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conf.*, Abstract.

[197] Patrick, J.D. and Wallace, C.S. (1982) Stone Circle Geometries: An Information Theory Approach. In Heggie, D. (ed.) *Archaeoastronomy in the New World*, pp. 231–264. Cambridge University Press.

[198] Phillips, P.C.B. and Ploberger, W. (1996) An asymptotic theory of Bayesian inference for time series. *Econometrica*, **64**, 240–252.

[199] Pilowsky, I., Levine, S. and Boulton, D.M. (1969) The classification of depression by numerical taxonomy. *Br. J. Psychiatry*, **115**, 937–945.

[200] Pose, R.D. (1992) A capability-based tightly-coupled multiprocessor. PhD Thesis, Department of Computer Science, Monash University, Australia.

[201] Pose, R.D., Anderson, M.S. and Wallace, C.S. (1986) Aspects of a Multiprocessor Architecture. *Proc. Workshop on Future Directions in Computer Architecture and Software*, May, pp. 293–295.

[202] Pose, R.D., Anderson, M.S. and Wallace, C.S. (1987) Implementation of a Tightly-Coupled Multiprocessor. *Proc. 10th Aust. Computer Science Conf. (ACSC-10)*, February, pp. 330–342. Published as *Proc. ACSC-10*, Vol. 9, No. 1.

[203] Prior, M., Eisenmajer, R., Leekam, S., Wing, L., Gould, J., Ong, B. and Dowe, D.L. (1998) Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. *J. Child Psychol. Psychiatry*, **39**, 893–902.

[204] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, USA. C5 is available from http://www.rulequest.com.

[205] Reed, K. (1983) The short address problem on digital computers. Master's Thesis, Department of Computer Science, Monash University, Clayton, Australia.

[206] Rissanen, J.J. (1976) Generalized Kraft inequality and arithmetic coding. *IBM J. Res. Develop.*, **20**, 198–203, May.

[207] Rissanen, J.J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–471.

[208] Rissanen, J.J. (1996) Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory*, **42**, 40–47, January.

[209] Rissanen, J.J. (1999) Hypothesis selection and testing by the MDL principle. *Comp. J.*, **42**, 260–269.

[210] Roberts, A.F.W. (1981) Formatting concepts of computer-generated operational reports. Master's Thesis, Department of Computer Science, Monash University, Australia.

[211] Roberts, P.S. (1974) Hardware design for compiling & other non-numeric processes. PhD Thesis, Department of Computer Science, Monash University, Australia.

[212] Roberts, P.S. and Wallace, C.S. (1972) A microprogrammed lexical processor. *Inf. Process.*, **71**, 577–581.

[213] Rood, S.J. (2008) *On silicon foundations: The formation of the Faculty of Information Technology*. Above is provisional title at this stage. www.waybackwhen.com.au/work-in-progress.html.

[214] Rubinstein, B., Bartlett, P. and Rubinstein, J.H. (2007) Shifting, One-inclusion Mistake Bounds and Tight Multiclass Expected Risk Bounds. In Schölkopf, B., Platt, J. and Hoffman, T. (eds) *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. MIT Press, Cambridge, MA, USA.

[215] Rumantir, G.W. (2004) Minimum message length criterion for second-order polynomial model selection applied to tropical cyclone intensity forecasting. PhD Thesis, School of Computer Science and Software Engineering, Monash University, Australia.

[216] Rumantir, G.W. and Wallace, C.S. (2001) Sampling of Highly Correlated Data for Polynomial Regression and Model Discovery. *4th Int. Symp. Intelligent Data Analysis (IDA)*, pp. 370–377.

[217] Rumantir, G.W. and Wallace, C.S. (2003) Minimum Message Length Criterion for Second-Order Polynomial Model Selection Applied to Tropical Cyclone Intensity Forecasting. *5th Int. Symp. Intelligent Data Analysis (IDA)*, pp. 486–496.

[218] Sanghi, P. and Dowe, D.L. (2003) A Computer Program Capable of Passing I.Q. Tests. *4th Int. Conf. Cognitive Science (and 7th Australasian Society for Cognitive Science Conf.)*, University of NSW, Sydney, Australia, July, Vol. **2**, pp. 570–575.

[219] Schmidhuber, J. (2007) Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity & Creativity. *Lecture Notes in Computer Science (LNCS),* 4755, pp. 26–38. Springer.

[220] Schmidt, D.F. (2008) Minimum message length inference of autoregressive moving average models. PhD Thesis, Faculty of Information Technology, Monash University, to appear.

[221] Scott, A.D. and Scott, M. (1997) What's in the two envelope paradox? *Analysis*, **57**, 34–41, January.

[222] Solomonoff, R.J. (1960) A Preliminary Report on a General Theory of Inductive Inference. Report V-131, Zator Co., Cambridge, MA, USA, 4 February.

[223] Solomonoff, R.J. (1964) A formal theory of inductive inference. *Inf. Control*, **7**, 1–22, 224–254.

[224] Solomonoff, R.J. (1995) The discovery of algorithmic probability: a guide for the programming of true creativity. In Vitanyi, P. (ed.), *Computational Learning Theory: EuroCOLT'95*, pp. 1–22. Springer.

[225] Solomonoff, R.J. (1996) Does Algorithmic Probability Solve the Problem of Induction? In Dowe, D.L., Korb, K.B. and Oliver, J.J. (eds), *Proc. Information, Statistics and Induction in Science (ISIS) Conf.*, Melbourne, Australia, August, pp. 7–8. World Scientific. ISBN 981-02-2824-4.

[226] Solomonoff, R.J. (1997) Does Algorithmic Probability Solve the Problem of Induction? Report, Oxbridge Research, Cambridge, MA, USA. http://world.std.com/~rjs/isis96.pdf.

[227] Solomonoff, R.J. (1999) Two kinds of probabilistic induction. *Comp. J.*, **42**, 256–259. Special issue on Kolmogorov Complexity.

[228] Solomonoff, R.J. Three kinds of probabilistic induction: universal and convergence theorems. *Comp. J.*, **51**, 566–570.

[229] Strang, G. and Nguyen, T. (1997) *Wavelets and Filter Banks* (Revised edn), Wellesley-Cambridge. 1st edn, 1996.

[230] Stuart, R.D. (1961) An Introduction to Fourier Analysis. In *Methuen's Monographs on Physical Subjects*. Methuen, London; Wiley, New York.

[231] Tan, P.J. and Dowe, D.L. (2002) MML Inference of Decision Graphs with Multi-Way Joins. In McKay, R. and Slaney, J. (eds) *Proc. 15th Aust. Joint Conf. Artificial Intelligence – Lecture Notes in Artificial Intelligence,* 2557, December, pp. 131–142. Springer.

[232] Tan, P.J. and Dowe, D.L. (2003) MML Inference of Decision Graphs with Multi-way Joins and Dynamic Attributes. *Lecture Notes in Artificial Intelligence (LNAI),* 2903, *Proc. 16th Aust. Joint Conf. Artificial Intelligence*, Perth, Australia, December, pp. 269–281. Springer.

[233] Tan, P.J. and Dowe, D.L. (2004) MML Inference of Oblique Decision Trees. *Lecture Notes in Artificial Intelligence (LNAI),* 3339, *Proc. 17th Aust. Joint Conf. Artificial Intelligence*, Cairns, Australia, December, pp. 1082–1088. Springer.

[234] Tan, P.J. and Dowe, D.L. (2006) Decision Forests with Oblique Decision Trees. *Lecture Notes in Artificial Intelligence (LNAI),* 4293, *Proc. 5th Mexican Int. Conf. Artificial Intelligence*, Apizaco, Mexico, November, pp. 593–603. Springer.

[235] Tan, P.J., Dowe, D.L. and Dix, T.I. (2007) Building Classification Models from Microarray Data with Tree-Based Classification Algorithms. *Lecture Notes in Artificial Intelligence (LNAI),* 4293, Proc. 20th Aust. Joint Conf. Artificial Intelligence, December. Springer.

[236] Tan, W.P.H. (1974) System analysis and simulation studies of the multiprogrammed, closed-queueing computer system. PhD Thesis, Department of Computer Science, Monash University, Australia.

[237] Uther, W.T.B. and Veloso, M.M. (2000) The Lumberjack Algorithm for Learning Linked Decision Forests. *Proc. 6th Pacific Rim Int. Conf. Artificial Intelligence (PRICAI'2000), Lecture Notes in Artificial Intelligence (LNAI),* 1886, pp. 156–166. Springer.

[238] Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*. Springer.

[239] Vereshchagin, N.K. and Vitányi, P.M.B. (2004) Kolmogorov's structure functions and model selection. *IEEE Trans. Inf. Theory*, **50**, 3265–3290.

[240] Visser, G. and Dowe, D.L. (2007) Minimum Message Length Clustering of Spatially-Correlated Data with Varying Inter-Class Penalties. *Proc. 6th IEEE Int. Conf. Computer and Information Science (ICIS) 2007*, July, pp. 17–22.

[241] Viswanathan, M. and Wallace, C.S. (1999) A Note on the Comparison of Polynomial Selection Methods. In Heckerman, D. and Whittaker, J. (eds), *Proc. Uncertainty 99: The 7th Int. Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, Florida, USA, January, pp. 169–177. Morgan Kaufmann, San Francisco, CA, USA.

[242] Viswanathan, M. and Wallace, C.S. (2003) An Optimal Approach to Mining Boolean Functions from Noisy Data. *4th Int. Conf. Intelligent Data Engineering and Automated Learning (IDEAL)*, pp. 717–724.

[243] Viswanathan, M., Wallace, C.S., Dowe, D.L. and Korb, K.B. (1999) Finding Cutpoints in Noisy Binary Sequences – A Revised Empirical Evaluation. *Proc. 12th Aust. Joint Conf. Artificial Intelligence, Lecture Notes in Artificial Intelligence*, Vol. 1747, pp. 405–416. Springer.

[244] Vitányi, P.M.B. and Li, Ming (1996) Ideal MDL and Its Relation to Bayesianism. In Dowe, D.L., Korb, K.B. and Oliver, J.J. (eds) *Proc. Information, Statistics and Induction in Science (ISIS) Conf.*, Melbourne, pp. 282–291. World Scientific.

[245] Vitányi, P.M.B. and Li, Ming (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. Inf. Theory*, **46**, 446–464.

[246] Wallace, C.S. (1958) The Determination of the Radial Distribution of Electrons, and the Size Spectrum of Extensive Air Showers. *Proc. A.A.E.C. Symp.*, Sydney.

[247] Wallace, C.S. (1959) SILLIAC Library Program "W3" – Water Pipe Network Analysis. Technical Report, University of Sydney, Australia, July. W3 was a program (rather than a technical report) to run on the SILLIAC computer.

[248] Wallace, C.S. (1960) Comparison Between the Response of Geiger and Scintillation Counters to the Air Shower Flux. *Proc. Moscow Cosmic Ray Conf.*, Vol. **II**, p. 316.

[249] Wallace, C.S. (1960) Counter experiments on extensive cosmic ray air showers. PhD Thesis, Department of Physics, University of Sydney, Australia.

[250] Wallace, C.S. (1960) Specifications for Paper Tape Editing Equipment. Technical Report 352, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[251] Wallace, C.S. (1961) Proposal for an Input–Output Buffering System. Technical Report 351, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[252] Wallace, C.S. (1961) Transfer Interruption Routine (TIR). Technical Report 370, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[253] Wallace, C.S. (1962) The Input–Output System of the Illiac II. Technical Report 487, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[254] Wallace, C.S. (1963) Suggested Design for a Very Fast Multiplier. Technical Report UIUCDCS-R-63-133, Department of Computer Science, University of Illinois at Urbana-Champaign.

[255] Wallace, C.S. (1964) Correlated round-off errors in digital integrating differential analyzers. *Comp. J.*, **7**, 131–134, July.

[256] Wallace, C.S. (1964) A suggestion for a fast multiplier. *IEEE Trans. Electron. Comp.*, 14–17, February. Reprinted in E.E. Swartzlander (1990) *Computer Arithmetic*, Vol. 1, IEEE Computer Society Press Tutorial, Los Alamitos, CA, USA.

[257] Wallace, C.S. (1966) Digital computers. In Butler, S.T. and Messel, H. (eds) *Atoms to Andromeda*, pp. 215–245. Shakespeare-Head, Sydney.

[258] Wallace, C.S. (1967) Accelerated relaxation of the non-linear equations of reticulation networks. *J. Inst. Eng.*, October, pp. 185–188

[259] Wallace, C.S. (1969) A Digital Logic Simulator for Teaching Purposes. *Proc. 4th Aust. Computer Conf.*, pp. 215–219.

[260] Wallace, C.S. (ed.) (1969) *4th Aust. Computer Conf. Control Systems*, August, Vol. 2.

[261] Wallace, C.S. (1971) Simulation Studies in File Structures. *A.C.S. Victorian Branch Professional Development Workshop*, February, p. 16.

[262] Wallace, C.S. (1973) Simulation of a Two-dimensional Gas. *Proc. Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conf.*, August, p. 19. Abstract.

[263] Wallace, C.S. (1976) Transformed rejection generators for Gamma and Normal pseudorandom variables. *Aust. Comp. J.*, **8**, 103–105, November.

[264] Wallace, C.S. (1977) Computing research in Australia. *Aust. Comp. J.*, **9**, 21–24.

[265] Wallace, C.S. (1978) Memory and Addressing Extensions to a HP2100A. *Proc. 8th Aust. Computer Conf.*, August–September, Vol. 4, pp. 1796–1811.

[266] Wallace, C.S. (1984) An Improved Program for Classification. Technical Report 47, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

[267] Wallace, C.S. (1984) Inference and Estimation by Compact Coding. Technical Report 84/46, Department of Computer Science, Monash University, Clayton, Victoria, Australia, August.

[268] Wallace, C.S. (1986) An Improved Program for Classification. *Proc. 9th Australian Computer Science Conf. (ACSC-9)*, February, pp. 357–366. Published as *Proc. ACSC-9*, Vol. 8.

[269] Wallace, C.S. (1989) A Long-period Pseudo-random Generator. Technical Report 89/123, Department of Computer Science, Monash University, Clayton, Australia, February.

[270] Wallace, C.S. (1990) Classification by Minimum-Message-Length Encoding. In Akl, S.G. *et al.* (ed.), *Advances in Computing and Information – ICCI '90, Lecture Notes*

in Computer Science (LNCS), Vol. 468, pp. 72–81. Springer, May.

[271] Wallace, C.S. (1990) Classification by Minimum-Message-Length Inference. *Working Notes AAAI Spring Symposium Series*, pp. 65–69. Stanford University, California, USA.

[272] Wallace, C.S. (1990) Physically random generator. *Comput. Syst. Sci. Eng.*, **5**, 82–88.

[273] Wallace, C.S. (1992) A Model of Inductive Inference. *Seminar*, Wednesday, November 25. Also on video, Department of Computer Science, Monash University, Clayton, Australia.

[274] Wallace, C.S. (1994) Fast Pseudo-random Generators for Normal and Exponential Variates. Technical Report CS 94/197, Department of Computer Science, Monash University, Melbourne, Australia.

[275] Wallace, C.S. (ed.) (1994) *Proc. Research Conf.*, Melbourne, Australia. Faculty of Computing and Information Technology, Monash University, 1994.

[276] Wallace, C.S. (ed.) (1995) Machine Learning. *Proc. 2nd Theory Day*. University of NSW.

[277] Wallace, C.S. (1995) Multiple Factor Analysis by MML Estimation. Technical Report CS TR 95/218, Department of Computer Science, Monash University, Clayton, Melbourne, Victoria, Australia.

[278] Wallace, C.S. (1996) False Oracles and SMML Estimators. In Dowe, D.L., Korb, K.B. and Oliver, J.J. (eds) *Proc. Information, Statistics and Induction in Science (ISIS) Conf.*, Melbourne, Australia, August, pp. 304–316. World Scientific. ISBN 981-02-2824-4; was Technical Report 89/128, Department of Computer Science, Monash University, Australia, June 1989.

[279] Wallace, C.S. (1996) Fast pseudorandom generators for Normal and exponential variates. *ACM Trans. Math. Softw.*, **22**, 119–127, March.

[280] Wallace, C.S. (1996) MML Inference of Predictive Trees, Graphs and Nets. In Gammerman, A. (ed.) *Computational Learning and Probabilistic Reasoning*, pp. 43–66. Wiley.

[281] Wallace, C.S. (1997) On the Selection of the Order of a Polynomial Model. Technical Report, Royal Holloway College, England, UK. but it is not clear that it was ever released there. Soft copy certainly does exist, though. Perhaps see www.csse.monash.edu.au/~dld/CSWallace Publications.

[282] Wallace, C.S. (1998) *Competition isn't the only way to go, a Monash FIT graduation address*, April. Perhaps see www.csse.monash.edu.au/~dld/CSWallacePublications.

[283] Wallace, C.S. (1998) Intrinsic classification of spatially correlated data. *Comp. J.*, **41**, 602–611.

[284] Wallace, C.S. (1998) Multiple Factor Analysis by MML Estimation. *Proc. 14th Biennial Australian Statistical Conf. (ASC-14)*, Queensland, Australia, July, p. 144.

[285] Wallace, C.S. (1998) On the Selection of the Order of a Polynomial Model. In Robb, W. (ed.) *Proc. 14th Biennial Australian Statistical Conf.*, Queensland, Australia, July, 145.

[286] Wallace, C.S. (1999) Minimum Description Length (Major Review). *The MIT Encyclopedia of the Cognitive Sciences (MITECS)*, pp. 550–551. The MIT Press, London, England, ISBN 0-262-73124-X.

[287] Wallace, C.S. (2005) Statistical and Inductive Inference by Minimum Message Length. *Information Science and Statistics*. Springer, May. ISBN 0-387-23795-X.

[288] Wallace, C.S. Block relaxation of fluid flow networks. *J. Inst. Eng. Aust.*, to appear (circa 1967).

[289] Wallace, C.S., Allison, L. and Dix, T.I. (eds) (1990) *Australian Computer Science Communications*, Australian Computer Science Assoc., Melbourne, Australia.

[290] Wallace, C.S. and Boulton, D.M. (1968) An information measure for classification. *Comp. J.*, **11**, 185–194.

[291] Wallace, C.S. and Boulton, D.M. (1975) An invariant Bayes method for point estimation. *Classif. Soc. Bull.*, **3**, 11–34.

[292] Wallace, C.S. and Brennan, M.H. (1958) The Automatic Digital Recording of Information from Cosmic Ray Air Showers. *Proc. Conf. Data Processing and Automatic Computing Machines*, Weapons Research Establishment, Salisbury, South Australia, June 3–8, 1957. Session II, Engineering, December.

[293] Wallace, C.S. and Dale, M.B. (2005) Hierarchical clusters of vegetation types. *Community Ecol.*, **6**, 57–74. ISSN 1585-8553.

[294] Wallace, C.S. and Dowe, D.L. (1993) MML Estimation of the von Mises Concentration Parameter. Technical Report 93/193, Department of Computer Science, Monash University, Clayton, Australia, December.

[295] Wallace, C.S. and Dowe, D.L. (1994) Estimation of the von Mises Concentration Parameter Using Minimum Message Length. *Proc. 12th. Aust. Stat. Soc. Conf.*, 1-page abstract.

[296] Wallace, C.S. and Dowe, D.L. (1994) Intrinsic Classification by MML – The Snob Program. *Proc. 7th Aust. Joint Conf. Artificial Intelligence*, November, pp. 37–44. World Scientific.

[297] Wallace, C.S. and Dowe, D.L. (1996) MML Mixture Modelling of Multi-State, Poisson, von Mises Circular and Gaussian Distributions. *Proc. Sydney Int. Statistical Congress (SISC-96)*, Sydney, Australia, p. 197.

[298] Wallace, C.S. and Dowe, D.L. (1997) MML Mixture Modelling of Multi-State, Poisson, von Mises Circular and Gaussian Distributions. *Proc 28th Symp. the Interface*, 608–613.

[299] Wallace, C.S. and Dowe, D.L. (1997) MML Mixture Modelling of Multi-State, Poisson, von Mises Circular and Gaussian Distributions. *6th Int. Workshop on Artificial Intelligence and Statistics*, Society for AI and Statistics, San Francisco, USA, pp. 529–536.

[300] Wallace, C.S. and Dowe, D.L. (1999) Minimum message length and Kolmogorov complexity. *Comp. J.*, **42**, 270–283.

[301] Wallace, C.S. and Dowe, D.L. (1999) Refinements of MDL and MML coding. *Comp. J.*, **42**, 330–337.

[302] Wallace, C.S. and Dowe, D.L. (1999) Rejoinder. *Comp. J.*, **42**, 345–347.

[303] Wallace, C.S. and Dowe, D.L. (2000) MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Stat. Comput.*, **10**, 73–83, January.

[304] Wallace, C.S. *et al*. (Multi Group) (1988). Using a Unix system as a multiprocessor frontend. The Organisation for Unix, Linux and Open Source Professionals (AUUGN), Vol. **7**, pp. 15–21.

[305] Wallace, C.S. and Freeman, P.R. (1987) Estimation and inference by compact coding. *J. Roy. Stat. Soc. Ser. B*, **49**, 240–252. Discussion, pp. 252–265.

[306] Wallace, C.S. and Freeman, P.R. (1992) Single-factor analysis by minimum message length estimation. *J. Roy. Stat. Soc. Ser. B*, **54**, 195–209.

[307] Wallace, C.S. and Georgeff, M.P. (1983) A General Objective for Inductive Inference. Technical Report 83/32, Department of Computer Science, Monash University, Clayton, Australia, March. Reissued in June 1984 as TR No. 44.

[308] Wallace, C.S. and Gupta, G.K. (1973) General linear multistep methods to solve ordinary differential equations. *Aust. Comp. J.*, **5**, 62–69.

[309] Wallace, C.S., Harper, D. and Hagan, R.A. (1976) A discrete system simulation package for a mini computer. *ACM Simuletter*, **7**, 9–13.

[310] Wallace, C.S. and Jessup, A.M. (1970) A simple graphic I/O device. *Aust. Comp. J.*, **2**, 39–40.

[311] Wallace, C.S. and Koch, D. (1985) TTL-compatible multiport bus. *Comput. Syst. Sci. Eng.*, **1**, 47–52.

[312] Wallace, C.S. and Korb, K.B. (1994) A Bayesian Learning Agent. In Wallace, C.S. (ed.) *Research Conf.*, Faculty of Computing and Information Technology, Monash University, Melbourne, 19.

[313] Wallace, C.S. and Korb, K.B. (1997) Learning Linear Causal Models by MML Sampling. Technical Report CS 97/310, Department of Computer Science, Monash University, Melbourne, Australia.

[314] Wallace, C.S. and Korb, K.B. (1999) Learning Linear Causal Models by MML Sampling. In Gammerman, A. (ed.) *Causal Models and Intelligent Data Management*, pp. 89–111. Springer. see also TR 97/310, Department of Computer Science, Monash University, Australia, June 1997.

[315] Wallace, C.S., Korb, K.B. and Dai, H. (1996) Causal Discovery via MML. Technical Report CS 96/254, Department of Computer Science, Monash University, Melbourne, Australia.

[316] Wallace, C.S., Korb, K.B. and Dai, H. (1996) Causal Discovery via MML. *13th Int. Conf. Machine Learning (ICML-96)*, pp. 516–524.

[317] Wallace, C.S. and Longe, O. (1967) Reading gapless tapes. *IEEE Trans. Elec. Comp.*, **EC-16**, 517–518, August.

[318] Wallace, C.S. and Patrick, J.D. (1991) Coding Decision Trees. Technical Report CS 91/153, Department of Computer Science, Monash University, Melbourne, Australia.

[319] Wallace, C.S. and Patrick, J.D. (1993) Coding decision trees. *Mach. Learn.*, **11**, 7–22.

[320] Wallace, C.S. and Pose, R.D. (1990) Charging in a Secure Environment. In Rosenberg, J. and Keedy, J.L. (eds) *Proc. Int. Workshop on Computer Architectures to Support Security and Persistence of Information, Workshops in Computing*, London, May 8–11, pp. 85–96. Springer.

[321] Wallace, C.S. and Rowswell, B.G. (1967) Competition for memory access in the KDF 9. *Comp. J.*, **10**, 64–68, May.

[322] Wallace, C.S., Winn, M.M. and Ogilvie, K.K. (1958) Dependence of the nucleonic component of air showers on radius. *Nature*, **182**, 1653–1654, December.

[323] Whitehouse, L.G. (1973) Compiling techniques for higher level languages. PhD Thesis, Department of Computer Science, Monash University, Australia.

[324] Wood, R.H. (1975) Scheduling multiprogrammed virtual memory computers. PhD Thesis, Department of Computer Science, Monash University, Australia.

[325] Young, W. (1989) Parking systems modelling. MSc Thesis, Department of Computer Science, Monash University, Australia.

[326] Zakis, J.D., Cosic, I. and Dowe, D.L. (1994) Classification of Protein Spectra Derived for the Resonant Recognition Model Using the Minimum Message Length Principle. *Aust. Comp. Sci. Conf. (ACSC-17)*, January, pp. 209–216.