Natural Sciences and Engineering Research Council of Canada | Conseil de recherches en sciences naturelles et en génie du Canada

Committee / Panel 1507

1507

Institutional Identifier

System-ID (for NSERC use only) 145608430

FORM 101
Application for a Grant
PART I

Date 2010/10/25

Family name of applicant	Given name	Initial(s) of all given names	Personal identification no. (PIN)
Poole	David	DL	**Valid** 16877

Institution that will administer the grant: British Columbia

Language of application: [X] English [ ] French

Time (in hours per month) to be devoted to the proposed research / activity: 40

Type of grant applied for: Discovery Grants - Individual

For Strategic Projects, indicate the Target Area and the Research Topic; for Strategic Networks indicate the Target Area.

Title of proposal: Large Scale Reasoning Under Uncertainty

Provide a maximum of 10 key words that describe this proposal. Use commas to separate them.
Probabilistic Reasoning, Semantic Web, Machine Learning, Decision Making, Relational Probabilistic Models, Statistical Relational AI

Research subject code(s)		Area of application code(s)	
Primary	Secondary	Primary	Secondary
2800	2805	1200	802

CERTIFICATION/REQUIREMENTS

If this proposal involves any of the following, check the box(es) and submit the protocol to the university or college's certification committee.

Research involving : Humans [ ] Human pluripotent stem cells [ ] Animals [ ] Biohazards [ ]

Does any phase of the research described in this proposal a) take place outside an office or laboratory, or b) involve an undertaking as described in Part 1 of Appendix B?

[X] NO [ ] If YES to either question a) or b) – Appendices A and B must be completed

TOTAL AMOUNT REQUESTED FROM NSERC

Year 1	Year 2	Year 3	Year 4	Year 5
106,280	106,280	106,280	106,280	106,280

SIGNATURES (Refer to instructions "What do signatures mean?")

It is agreed that the general conditions governing grants as outlined in the NSERC *Program Guide for Professors* apply to any grant made pursuant to this application and are hereby accepted by the applicant and the applicant's employing institution.

Applicant

Applicant's department, institution, tel. and fax nos., and e-mail

Computer Science
British Columbia
Tel.: (604) 822 6254
FAX: (604) 822 5485
poole@cs.ubc.ca

Head of department

Dean of faculty

President of institution (or representative)

The information collected on this form and appendices will be stored in the Personal Information Bank for the appropriate program.

PROTECTED WHEN COMPLETED

Version française disponible

Canada

Personal identification no. (PIN)	Family name of applicant
Valid 16877	Poole

SUMMARY OF PROPOSAL FOR PUBLIC RELEASE (Use plain language.)

This plain language summary will be available to the public if your proposal is funded. Although it is not mandatory, you may choose to include your business telephone number and/or your e-mail address to facilitate contact with the public and the media about your research.

Business telephone no. (optional):

E-mail address (optional): poole@cs.ubc.ca

Individuals and society are increasingly faced with making decisions based on enormous amounts of information and they need to have a principled way to judge that information. Inspired by applications in geology, the first thrust of this proposal is to build the foundations of what we have called "semantic science" (in analogy with the semantic web, but for scientific data and hypotheses), to allow for ontologies that enable semantic interoperability, observational data that provides evidence, and hypotheses that make (probabilistic) predictions on data and can be used to form models to make predictions for specific cases. Making these all work together is a major research challenge.

Spatial planning under uncertainty, such as arise in forestry applications, where effects of actions and utilities are not local is the second thrust. We are investigating policy gradient methods that iteratively improve relational representations of policies.

A third thrust is in efficient inference for probabilistic relational models. In 2003 I first proposed lifted inference, where we do not distinguish individuals about which we have the same information. There has been considerable advances, but we still have not reached the "holy grail" where we can do inference exponentially faster than grounding in the number of indistinguishable individuals, although it seems plausible that there is such a method. This promises to form a foundation of the next generation of probabilistic modelling and programming languages.

A fourth area is in reasoning about existence and identity uncertainty. Typically models refer to roles of individuals, but the observations of the world do not specify which individuals, if any, fulfill the roles of the models. This work requires advances in representations as well as reasoning and learning algorithms.

Other Language Version of Summary (optional).

Personal identification no. (PIN)	Family name of applicant
Valid 16877	Poole

Before completing this section, **read the instructions** and consult the *Use of Grant Funds* section of the NSERC Program Guide for Professors concerning the eligibility of expenditures for the direct costs of research and the regulations governing the use of grant funds.

TOTAL PROPOSED EXPENDITURES (Include cash expenditures only)

	Year 1	Year 2	Year 3	Year 4	Year 5
1) Salaries and benefits					
a) Students	87,880	87,880	87,880	87,880	87,880
b) Postdoctoral fellows	0	0	0	0	0
c) Technical/professional assistants	0	0	0	0	0
d)	0	0	0	0	0
2) Equipment or facility					
a) Purchase or rental	5,000	5,000	5,000	5,000	5,000
b) Operation and maintenance costs	1,500	1,500	1,500	1,500	1,500
c) User fees	2,100	2,100	2,100	2,100	2,100
3) Materials and supplies	800	800	800	800	800
4) Travel					
a) Conferences	9,000	9,000	9,000	9,000	9,000
b) Field work	0	0	0	0	0
c) Collaboration/consultation	0	0	0	0	0
5) Dissemination costs					
a) Publication costs	0	0	0	0	0
b)	0	0	0	0	0
6) Other (specify)					
a)	0	0	0	0	0
b)	0	0	0	0	0
TOTAL PROPOSED EXPENDITURES	106,280	106,280	106,280	106,280	106,280
Total cash contribution from industry (if applicable)					
Total cash contribution from university (if applicable)					
Total cash contribution from other sources (if applicable)	0	0	0	0	0
TOTAL AMOUNT REQUESTED FROM NSERC (transfer to page 1)	106,280	106,280	106,280	106,280	106,280

PROTECTED WHEN COMPLETED

Justification for proposed expenditure

My most pressing needs are in support of graduate students. The students supported by this grant are doing foundational research, and as such, the NSERC discovery grants are the appropriate funding source. We also need funding for travel to expose our research to the wider comminity and for appropriate tools that allow us to carry out our research.

1 (a) Students: support for three Ph.D. students ($19,000 each). I expect to support Mark Crowley, David Buchman and one other Ph.D. student.

Support for taking on one M.Sc. student each year at $16,500 per student per year. I have to pay each student for 4-terms, as the department pays for their first two terms. There are always M.Sc. students who are interested in working for me, and it will not be difficult to find a student.

Support for 1 Summer student, at $5,500 as a top-up for a USRA. I have supported USRA students over a number of years. We have had good experience with undergraduate students in the past; they usually do very good programming work and we can get them excited about research.

We also need to pay 4% benefits.

Item	Rate	Cost
Ph.D. students	3 @ $19,000	$57,000
M.Sc. student	$\frac{4}{3}$ @ $16,500	$22,000
Undergraduate summer student	1 @ $5,500	$5,500
Benefits	4%	$3,380
Total		$87,880

2 (a) Purchase or rental: $6,000 will allow me to maintain (reasonably) modern computers for myself and my students associated with this project.

2(b) This is for repairs of computers and for Internet connection.

2 (c) User fees: The UBC Department of Computer Science charges for the direct costs of technical support to research grants. This includes installation and support of equipment, technical support for researchers, file servers, network servers, printing and other similar direct costs.

3 Materials and supplies: paper, postage, photocopying, research books.

4(a) Travel Conferences: including IJCAI, AAAI, UAI, KR conferences, and other more specialized workshops. This includes graduate student travel.

Large-Scale Reasoning Under Uncertainty

Objectives of the Research Program

Fundamentally, I am interested in decision making: what should an agent do as a function of its abilities, its values, its prior knowledge and its observations? This could be at an individual level (e.g., what medicine should I take or what house should I spend my time considering buying) or at a societal level (e.g., where should we extract minerals with the lowest environmental impact or how can we exploit resources in a sustainable way). To make good decisions, we should use the best information available in the world. Current search engines used to find information base their rankings using measures of popularity, such as pagerank, and often return sources that are supposedly authoritative. As a scientist, I am not content with using popularity and appeal to authority as the basis for determining what is true. If a person or web site makes a prediction, it is reasonable to ask them for the evidence for such a prediction. Ultimately predictions should be grounded in observations.

Building on the foundations of Bayesian decision theory, we want to make probabilistic predictions that can be combined with utilities to decide on the best actions. To make predictions, we create hypotheses and condition on all available data. To find all of the relevant data, ontologies, which specify the meaning of symbols in knowledge sources, are required to ensure that the hypotheses and the data can interoperate at a semantic level.

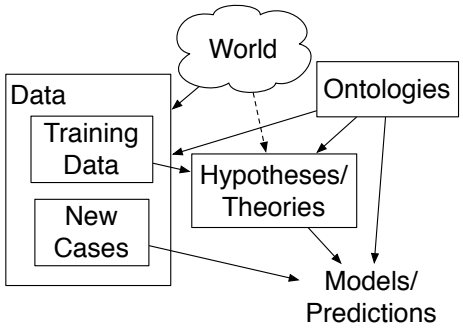

Figure 1: Semantic Science Architecture

We have called this technology to support decisions "semantic science", based on the semantic web [1], which is an endeavor to make all of the world's knowledge accessible to computers, and using scientific methodology to make predictions. Figure 1 shows the main components of semantic science. Ontologies are used to define the vocabulary of data and hypotheses. Observational data, which depends on the world and the ontologies, are published. Such data sets can be very heterogenous, at widely varying levels of abstraction and detail. Hypotheses that make probabilistic predictions are also published. Hypotheses are not created in isolation, but depend on training data. Hypotheses can be judged by their prior plausibility and how well they predict the data; the best hypotheses are called theories. Given a new case, various hypotheses are combined to form models that can be used to make predictions on that case. Given a prediction, users can ask what hypotheses were used to make that prediction, and for each hypothesis, they can ask what data was used to support such a hypothesis. In this way all decisions can be based on all of the available evidence. This is, of course, a much bigger project than can be carried out in an NSERC discovery grant, but there remain many fundamental research problems that need to be solved before the vision can be brought to reality.

Typically data is not just a set of mappings from features into a fixed set of values, as is often assumed by traditional machine learning, but often refers to individuals that are only referred to by name; it is the properties of these individuals and the relationships among these individuals

that is important for prediction. Following my publication of what could be argued was the first probabilistic relational language [12], [13], (pdf-33)[1], the combination of logical and probabilistic reasoning has blossomed. There are still many outstanding fundamental problems for representations, inference and learning. One of the main techniques that I proposed [14] and I am still working on is the problem of lifted inference: carrying out probabilistic inference reasoning about classes of individuals as a unit, without reasoning about them individually. Another problem (pdf-12) is where models refer to individuals in terms of the roles they fill, but the data does not label the observed individuals with roles.

Given probabilistic predictions, to make decisions we also need utilities and the ability to plan actions. I am also interested in sequential decision making, particularly for large spatial domains.

In summary, I am interested in representations and algorithms that enable making decisions based on the best available information. As part of working on the foundations, I try to ensure that the representations and algorithms for the various parts can fit together into a coherent whole.

Recent progress in research activities related to proposal

Over the last 6 years, I have been working both on applications and on fundamental theory. Both are essential for carrying out science in this field. The theory needs to be driven by how it can be used, and the applications need to be built on solid foundations.

I have been working with geologists to develop practical instances of the semantic science vision. We started developing qualitative probabilistic models (pdf-17, pdf-20) of landslide susceptibility and mineral occurrences, concentrating on the problems of interacting with rich ontologies and reasoning at multiple levels of abstraction and detail. Interacting with geologists who have been developing ontologies, we have developed what we call Aristotelian definitions (pdf-3) which are both natural for the geologists but also provide the hooks needed for probabilistic reasoning. Existing qualitative models were not expressive enough, so we have started doing Bayesian probabilistic reasoning with ontologies and heterogenous relational data sets. Paper pdf-2 deals with interacting with ontologies, with reasoning at multiple levels of abstraction and detail and with modelling of roles. It uses a naive-Bayesian approach (modelling $P(\mathit{description}|\mathit{model})$ and $P(\mathit{model})$), which we discovered may not be the best approach because we want the probability of the model to depend on the description of the model.

As well as the practical applications, which use single models for predictions, and are not learned from data (but are evaluated using data), we have also been working on the big picture (pdf-32), including how to make predictions on specific cases using multiple hypotheses (pdf-21, pdf-8). These foundations are essential for future applications.

I have also been working on probabilistic inference, which is an essential component of decision making. Rita Sharma and I developed techniques for variables with large hierarchically structured domains (pdf-15). One particularly exciting development is lifted probabilistic inference for relational domains, which I first proposed in 2003 [14]. It still is not solved yet despite a number of Ph.D. theses on the topic. Jacek Kisyński and I have made progress for directed models (pdf-9) and in understanding the constraint processing involved (pdf-10).

[1] References of the form (pdf-*nn*) refer to item *nn* in my personal data form. References of the form [*nn*] refer to item *nn* in the References section at the end of the proposal.

I have also been working on representations of complex phenomena. One area that has recently resurfaced is in probabilistic programming languages. I have shown how many problems that have been solved by other researchers can be represented using the earlier developed probabilistic programming language ICL (pdf-33). I have developed a unifying framework for understanding probabilistic programming languages (pdf-31). I have also been working on the problem of existence uncertainty; uncertainty about the existence of an object that fills a role (pdf-12) which turns out to be a very tricky problem. Michael Chiang and I have been working on learning for these and other relational models, particularly with unobserved relations (pdf-31).

Mark Crowley and I have been working on planning in large spatial domains inspired by applications in forestry. We have developed a policy gradient method that works in the extremely large domains characterized by actions at each spatial location (pdf-11).

Alan Mackworth and I have been working on understanding and explaining a coherent picture of AI. This has resulted in an AI textbook (pdf-29), and a suite of learning tools known as AIspace (aispace.org). This has resulted in a number of publications (pdf-18, pdf-4) that show the interactive tools are useful pedagogically. We believe that the book is a research contribution because it shows a set of techniques that all fit together, and is not just a collection of ad hoc techniques.

Literature Pertinent to the Research

The idea of semantic science is partly based on the semantic web [1], from which we take the idea of publishing information with reference to formal ontologies. However, we do not need the semantic web to be fully functioning before semantic science can work. There has been work based on publishing scientific ontologies [16] and publishing scientific data referring to these formal ontologies [5, 8], which form part of the semantic science vision. None of these deal with making probabilistic predictions. The closest work with this would be the work on PR-OWL [2], which defines ontologies for uncertainty.

There have been a number of relational probabilistic representations and probabilistic programming languages that have been proposed, most noticeably ProbLog [3], Blog [9] and Church [6] (see pdf-31 for an overview). None of these are at the level where we could apply them to our semantic science domains, mainly because they do not interact with ontologies, do not have good solutions to the problem of existence uncertainty, and also because their inference algorithms are not as good as special purpose algorithms (although that may change soon).

For the problem of lifted inference, following my initial proposal in 1993 [14], de Salvo Braz et al. [4] invented counting elimination, Milch et al. [10] proposed a representation of counting formulae, and Singla and Domingos [15] proposed a way to use lifted approximate inference for learning. This body of research has not fulfilled the promise of lifted inference, as the algorithms still need to ground in some cases.

The spatial planning work is based on relational policy gradient methods [7], which rely on adjusting the parameters of a parametrized representation of a stochastic policy to improve the expected utility. Our policy is specified in terms of a causal calculus [11], where the probability distribution of the actions at each location depends on what is done at neighbouring locations.

Methods and Proposed Approach

My methodology is to be inspired by real problems and to develop the foundations upon which the solutions to these and related problems can be built. By considering real problems (problems that someone is interested in the solution to), we first see what can be solved by existing technologies. Either existing technologies can be directly applied to solve the problems, or there is research required to develop the appropriate technology. When we build the solutions, I like to make them as generally applicable as possible. For example, the work on semantic science began with the problems of predicting mineral occurrences and predicting landslides. It became clear that we needed standardized ontologies to describe the world, but we also wanted to make probabilistic predictions, as these are what is required for making decisions under uncertainty. To anticipate future domains, we have generalized the problem to include training data and the need to combine multiple hypotheses to make a prediction. This vision of semantic science goes beyond the applications in geology (because there is typically little data from which to learn in those problems), and is inspiring research on the bigger picture that can be applied to the specific applications.

The following paragraphs give some more examples of this methodology for research.

In the spatial planning application (pdf-11), we need a representation of a spatial policy. Such a policy specifies a distribution over actions for each location that depends on the actions at neighbouring locations. Conditional probabilities are not the appropriate representation because the specification is inherently cyclic. We tried to apply the theory of causal modelling [11], but found it inadequate for the sorts of cyclic models we were considering. We are working on developing a theory of causal modelling based on the idea that the probability distribution after all interventions is the equilibrium distribution of a Markov chain. This is being developed in a very general setting, but should be able to be applied for the specific case of stochastic spatial policies.

When building relational models, we can really only check the models against ground truth for small models. However, there are problems with developing and learning models at one population size and applying them at other population sizes. Most of the current representations do not allow the flexible specification of how the predictions change as the population changes. We need new representations that allow the modelling of how the predictions depend on population sizes.

Traditionally, in Bayesian modelling we represent $P(m)$ and $P(d|m)$, where m is a specific model, and d describes what the model predicts. When we have rich relational models, it is difficult to determine the prior probability of a model, as there is not enough data to estimate it from, and so we must compute it from the model's description. We are exploring a number of solutions that are mathematically well-founded and fit the constraints of real applications.

When carrying out inference with relational domains, we would like to carry out inference in a lifted manner, treating all undistinguished individuals as a group and counting them. None of the existing solutions fully solve this, because they are all based on variable elimination, a dynamic programming approach that relies on representing the intermediate results, however current lifted representations are not closed under the operations of multiplication and summing out (parametrized) variables. While it may be possible to find more complex representations of the intermediate results, Jacek Kisyński, Fahiem Bacchus (University of Toronto) and I are investigating search based methods, which we have reason to believe can be exponentially faster (as the population size increases for a fixed model) than grounding: when the grounding is polynomial, lifted search can be logarithmic and when the grounding is exponential, lifted search can be polynomial.

Current representations of discrete probability distributions are usually in terms of tables that do not make explicit the internal structure of the distribution. David Buchman, Nando de Freitas and I are working on a new spectral representation that has no redundant parameters and can represent the interactions between the variables, at multiple frequencies. We believe this has great potential for learning and approximate inference, where we may only need to represent high frequency interactions when supported by the data or when it makes a difference in prediction.

I measure progress by the usual scientific measures of publications and impact on future research. I would prefer to struggle with difficult problems, propose novel techniques for these problems, and inspire others to work on them, rather than play the game of incremental improvements of existing techniques. While this strategy may not maximize my rate of publication, I believe it maximizes my impact.

Anticipated Significance of the Work

I believe that semantic science has the potential to change the world. When it comes to fruition, people will wonder why in 2010 people were content with isolated non-interacting data sets, data and hypotheses that cannot be easily found and tested against each other, and predictions that cannot be justified in terms of all of the relevant data.

In the shorter term, by building the foundations of large-scale reasoning under uncertainty, inspired by real applications, we can help develop the science of AI as both useful and intellectually stimulating. Many of the problems tackled by AI are too difficult for people to solve unaided, and we need the tools of AI to hope to solve them.

Training of Highly Qualified Personel

The research outlined in this proposal provides a tremendous opportunity for training highly qualified personnel, from undergraduates doing summer programming jobs to Ph.D. students. We need people who are trained in the scientific methodology, and the work presented here integrates theory, empirical studies and applications that should serve the students well whether they continue with university research or move to industry.

I have concentrated on supervising Ph.D. students, not necessarily by design. When I have supervised M.Sc. student they have tended to want to stay with me to do a Ph.D. (Sharma and Crowley), have come to UBC to do a Ph.D. (Kisyński and Chiang) or are transferring from an M.Sc. to a Ph.D. (Buchman). I have regularly supervised undergraduate students as USRAs.

Funding is my main limit to supervising more students. I cannot afford to support more than 3 graduate students at a time on my current grant.

I have mostly relied on NSERC funding, mainly because I am working on novel techniques. Semantic science, for example, is not on other funders' radar, as it is too new. There is still a lot of fundamental research, the sort that is appropriate for NSERC funding, that needs to be carried out before the full impact of this work can be applied. I have been able to find funding from MITACS to support two of my students (Sharma and Kisyński) to apply their work in an industrial setting for short periods after their graduation. Unfortunately the companies and government departments we interact with are too small to provide ongoing research support for the fundamental research.

References

[1] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, May:28–37, 2001.

[2] Paulo C. G. da Costa, Kathryn B. Laskey, and Kenneth J. Laskey. PR-OWL: A Bayesian ontology language for the semantic web. In *Proceedings of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web*, Galway, Ireland, Nov 2005.

[3] L. De Raedt, A. Kimmig, and H. Toivonen. ProbLog: A probabilistic Prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, pages 2462–2467, 2007.

[4] R. de Salvo Braz, E. Amir, and D. Roth. Lifted first-order probabilistic inference. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. M.I.T. Press, 2007.

[5] P. Fox, D.L. McGuinness, D. Middleton, L. Cinquini, J.A. Darnell, J. Garcia, P. West, J. Benedict, and S. Solomon. Semantically-enabled large-scale science data repositories. In *5th International Semantic Web Conference (ISWC06)*, volume 4273 of *Lecture Notes in Computer Science*, pages 792–805. Springer-Verlag, 2006.

[6] N. Goodman, V. Mansinghka, D.M. Roy, K. Bonawitz, and J. Tenenbaum. Church: a language for generative models. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2008.

[7] K. Kersting and K. Driessens. Non-parametric policy gradients: A unified treatment of propositional and relational domains. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

[8] D. McGuinness, P. Fox, L. Cinquini, P. West, J. Garcia, J.L. Benedict, and D. Middleton. The virtual solar-terrestrial observatory: A deployed semantic web application case study for scientific research. In *Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07)*, Vancouver, BC, Canada, July 2007.

[9] Brian Milch, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L. Ong, and Andrey Kolobov. BLOG: Probabilistic models with unknown objects. In *IJCAI-05*, Edinburgh, 2005.

[10] Brian Milch, Luke S. Zettlemoyer, Kristian Kersting, Michael Haimes, and Leslie Pack Kaelbling. Lifted probabilistic inference with counting formulas. In *Proceedings of the Twenty Third Conference on Artificial Intelligence (AAAI)*, 2008.

[11] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[12] D. Poole. Representing diagnostic knowledge for probabilistic Horn abduction. In *IJCAI-91*, pages 1129–1135, Sydney, August 1991.

[13] D. Poole. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64(1):81–129, 1993.

[14] David Poole. First-order probabilistic inference. In *Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 985–991, Acapulco, Mexico, 2003.

[15] Parag Singla and Pedro Domingos. Lifted first-order belief propagation. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1094–1099, 2008.

[16] Barry Smith. Ontology. In L. Floridi, editor, *Blackwell Guide to the Philosophy of Computing and Information*, pages 155—166. Oxford: Blackwell, 2003.