Robert Williamson

E1 Reconceiving Machine Learning

E2 Aims and Background

Beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method oriented rather than problem oriented. The method-oriented man is shackled: the problem-oriented man is at least reaching freely toward what is most important. [52]

Context Machine Learning is a sub-discipline of Information and Communication Technology (ICT) that develops the technologies for machines to recognise and learn patterns in data. It is distinct from, although related to, statistics. It can be differentiated by its focus on creating technology rather than the human-centred analysis of data. It is the science and engineering behind Data Mining.

Machine learning is *pervasive*: it plays a key role in all stages of the scientific process and across diverse fields including bioinformatics, engineering and finance. It is widely accepted that ICT plays an enabling role across almost all technological disciplines. Analogously, Machine Learning plays an enabling role across most parts of ICT, from embedded to enterprise systems, and consequently is a crucial enabler of the Digital Economy [16]. Vast quantities of data are now routinely collected and stored because it is affordable to do so. Machine learning makes sense of this data flood.

The Problem The massive reduction in the cost of collecting, storing, transporting and processing data has meant an increasing need for tools to make sense of it. Unfortunately, the deployment of modern machine learning tools is more akin to a craft than an engineering discipline: the inference problems to be solved are often under-specified or ill-posed and the available tools are often *ad hoc* — lacking generality, transparency, usability and interoperability. Our premise is that the root cause of these difficulties is a lack of a clear conceptual basis for machine learning as an information engineering discipline.

Research in machine learning is currently organised by technique (e.g. kernel methods, neural networks, graphical models), by domain (e.g. KDD, Bioinformatics), or by some central philosophical or analytic technique (e.g. Bayesianism, Minimum Description Length, Empirical Risk Minimisation). The last of these organisational approaches tend to be monistic — proposing that all problems be framed according to their principles. The first two approaches make no attempt to be comprehensive.

There are two key symptoms arising from this lack of conceptual foundations: A *lack of direction* leading to no clear research agenda and a plethora of incremental advances that do not help solve real problems [31] leading to "errors of the third kind (giving the right answer to the wrong question)" [32]. Furthermore there is a *lack of usability* of the tools created. There is no simple set of principles that can be imparted to lay-people in specific domains so they can identify the problem they need to solve, choose the appropriate techniques and correctly interpret the results.

What is required is a principled method for identifying, organising and relating key concepts in machine learning: one that is *pluralistic* — making use of the insights already gained in the existing philosophic and analytic appraisals of the field without subscribing to any single view — and *problem focussed* so as to provide an alternative to a research culture of incremental advancements.

Aims The aim of this proposal is to build these conceptual foundations and reconceive machine learning as an engineering discipline that can address the needs of an increasingly data-saturated world. We will do this using Locke's categorisation of how the mind exerts its power over simple ideas [49] Chapter 12, para 1 (combining several simple ideas into a compound one; relating two seperate ideas; and abstraction). Using this categorisation, we propose to pursue three broad aims:

Language Development (Abstraction) The development of a general and formal vocabulary and language for *defining the multitude of problems* and relations between them. The language needs to be *descriptive* (capturing all of the problems in details) and *prescriptive* (allowing the development of conceptual infrastructure to relate problems);

Cataloging of Problems (Combining) A comprehensive and modular cataloging of all possible *problems* and *relations* (taking account of the numerous combinatorial factors involved) and a consequent mapping of known results into this catalog;

Relations between Problems (Relations) The development of new relations between problems: a systematic study of which problems can be related to others (or sets of problems to sets of problems). The development of primitives, representations and approximations.

The nature of the proposed work could be considered building *infrastructure*[11]. Whilst there is a large element of Science (theoretical analysis) the ultimate goals are Engineering knowledge.

The synthesis sought is (in Foucault's terminology) a *general* rather than *total* history: it seeks to understand relations rather than develop a single grand unified theory. It would lay the foundations for a design-by-composition approach to solving real problems. The development of suitable modularity can be expected to bring huge efficiency gains as it does in software construction and Engineering[3].

The intent is to assimilate different learning problems (that is precisely define them and catalog them). By "learning problem" we mean a statement of what it is to be solved. A complete solution needs a model and a learning algorithm.

Some (terse and imperfect) analogies may help explain the proposal's goal to non-specialists:

Computer Science Computational complexity (especially NP-completeness) through Garey & Johnson's book and Johnson's ongoing *catalog*[38] has lead to a detailed and structured understanding of the inter-relationships between many fundamental problems in computer science.

Mathematics Consider machine learning problems as functions. In the 19th century, each function was considered separately. Functional Analysis reconceived them by considering *sets* of functions and *relations* (mappings) between them and subsequently developed many new and powerful tools. The increasing abstraction and focus on relations has remained a powerful force in mathematics.

Biology A systematic *cataloging* (taxonomy) resonates with Biology's Linnean past. But this proposal is not merely taxonomic — it seeks to understand the relationships analogously to Systems Biology (itself a paradigm shift or reconception of a field[1]) which "is about putting together rather than taking apart, integration rather than reduction.... Successful integration at the systems level must be built on successful reduction, but reduction alone is far from sufficient."[51].

Background There are many different problems in machine learning. Consider the following (unstructured!) list, which illustrates the confused state at present (confused because these are at all different levels of abstraction, and are in fact often the same problem renamed). Citations for each are available but omitted for space reasons: batch, online, transductive, off-training set, semi-supervised, noisy (label, attribute, constant noise / variable noise, data of variable quality), data of different costs, weighted loss functions, active, distributed, classification (binary weighted binary multi-class), structured output, probabilistic concepts / scoring rules, class probability estimation, learning with statistical queries, Neyman-Pearson classification, regression, ordinal regression, ranked regression, ranking, ranking the best, optimising the ROC curve, optimising the AUC, regression, selection, novelty detection, multi-instance learning, minimum volume sets, density level sets, regression level sets, sets of quantiles, quantile regression, density estimation, data segmentation, clustering, co-training, co-validation, learning with constraints, conditional estimators, estimated loss, confidence / hedging estimators, hypothesis testing, distributional distance estimation, learning relations, learning total orders, learning causal relationships, and estimating performance (cross validation).

Hand[31] has recently argued that the mere pursuit of even more methods for standard problems alone adds little. What is needed is to understand the relationships between problems.

The majority of the literature comprises specific contributions to particular problems such as listed above. Below I summarise work concerning how different problems *relate* to each other. The general idea of relating different problems is ancient — it can be traced through Leibniz, Comte, Linneaus etc, as is the notion of taking a relative perspective as foundational[7]. The need for a thorough reconception has been articulated in the statistics literature by Bradley Efron (he was talking of statistics, but it applies more generally to Machine Learning too)

> History seems to be repeating itself: we've returned to an era of ragtag heuristics, propelled with energy but with no guiding direction. Maybe we can hope that history really will repeat itself and that some brand-new Fishers and Neymans will succeed in rationalizing all this activity over a solid theoretical foundation[22].

A recent NSF workshop concluded that "More attention must be given to consolidation of knowledge and the development of new theories and methods with broad applicability."[48] Within the Machine

Learning community, Bousquet has stated the need for an agreed vocabulary, a clear statement of the main problems, and to "revisit what has been done or discovered so far with a fresh look"[10].

There are different approaches to unification. One distinction is between *Monistic* and *Pluralistic* approaches. Monistic approaches aim for a single all encompassing theory. A problem with most monistic approaches is that you have to accept them "all or nothing." They include

Low level data interchange: There is a small amount of work on developing standards for interchanging data sets[28]. There are also some limited higher level attempts such as ontologies[62].

Modelling frameworks: To solve a machine learning problem, one needs models. There is a rich literature on graphical models or factor graphs which have allowed the unification of sets of problems[72], with a focus on the modelling and computational techniques for particular problems.

Comparison of frameworks: There are several philosophical frameworks/approaches to designing inference and learning algorithms. There are several works[4] that compare and contrast these. They are effectively comparing different monistic frameworks, not comparing problems.

Overarching frameworks: These include Bayesian[57], information-theoretic[37], game-theoretic[30], MDL[29], regularised distance minimisation[2], and more narrowly focussed "unifying frameworks"[20] such as information geometry, exponential families and the information bottleneck.

Pluralistic approaches are closer to what is proposed here, and resonate with Kiefer's assertion that "Statistics is too complex to be codified in terms of a simple prescripton that is a panacea for all settings, and ... one must look as carefully as possible at a variety of possible procedures..."[39] They combine existing results without the need to discard existing frameworks. Examples include:

Problem Catalogs: There are few attempts to catalog a range of problems. They typically just consider certain aspects such as different notions of cost[65] or a restricted set of problems[54].

Wald's Decision Theoretic Framework: Wald's[70] decision theoretic approach to statistics is central to the proposed work. It makes precise what the goal of the learning problem is (via the *loss* function) and is present in the majority of the machine learning frameworks. Even when the goal is to just "gather information" a loss function is effectively being utilised[18].

Representation results: There are results concerning *representations* of problems in terms of more elementary ones, including classification[7] and scoring rules[12]. The insights these representational results bring allow much better understanding of the behaviour of popular techniques[50].

Comparison of experiments: A significant precedent is the notion of comparison of experiments[64] developed by Blackwell[8] and extended by LeCam[46]. An experiment $E = \{(X, S_X); (P_\theta, \Omega)\}$, where X is a random variable on sample space S_X generated by a distribution P_θ parametrised by $\theta \in \Omega$; i.e. a *parametric model*. (A *problem* is an experiment with a particular loss function, without a fixed model.) If $F = \{(Y, S_Y); (Q_\theta, \Omega)\}$ is another experiment, E is *always better than* F if for *every* decision problem (i.e. loss) involving θ and for *every* decision rule $\delta(Y)$ based on F, there is a decision rule $\delta_*(X)$ based on E such that the risk (expected loss) of $\delta_*(X)$ is less than or equal to the risk of $\delta(Y)$. This is a powerful notion since it holds for *all* losses. LeCam's theory is formulated in an abstract way to make its theorems elegant. Its abstract formulation has made it inaccessible[53].

Connections between classification and distances: There are deep connections between classification problems and distances between probability distributions. We have extended and unified a number of such results[55] by exploiting integral representations of divergences and risks in terms of primitives[12](see progress report in section D2). The integral representation of scoring rules has recently been extended to a much wider class of problems[41] and we expect this will enable the concomitant extension of the relation results[55] to a much wider class of problems.

Reductions: The most significant development in the machine learning literature for this proposal is the notion of a *reduction* between learning problems[5]. Although apparently developed in ignorance of the literature on comparison of experiments, the setup is similar, but differs in a crucial aspect: rather than considering *all* loss functions, effort is focussed on a *particular* loss function. (It is a relationship between *problems* rather than *experiments*.)

There is no satisfactory formal definition of a reduction. In order to formally define a reduction, one needs a formal definition of a learning *problem*. Taking that for granted here, and restricting consideration to a simple reduction from a single problem P_1 to another single problem P_2, a reduction comprises 1) a transformation of inputs from P_1 to a form suitable for algorithm A_2 which solves P_2;

2) a transformation of the outputs of A_2 back into the form needed for P_1; and 3) a statement relating the performance of A_2 on P_2 to the performance of the synthesised algorithm on P_1.

E3 Significance and Innovation

The significance and innovation in the proposal are addressed under four key headings.

How the Knowledge Base of the Discipline is Advanced The knowledge base is advanced by synthesising it and drawing numerous connections. The anticipated outcomes will mean that it will be possible to effectively comprehend all machine learning problems. This will be a clear advance because such a comprehensive understanding is hardly possible at present. It will allow the formulation of new problems (for example "in-between" others) and the development of solutions for them.

We also expect to develop new tools and learning algorithms. There is already evidence that indirect techniques[43] can outperform (statistically and computationally) direct ones for non-trivial machine learning problems. Thus we expect concrete advances in knowledge through new algorithms.

Novelty and Innovation in the Aims and Concepts The *aim* of the proposal is certainly novel: no-one has tried such a reconception of the field on such a scale, and in such a manner. As far as we can tell, there is no trace of the idea in the refereed literature. The closest is Bousquet's[10] posting on his blog or a NSF workshop[48]. Most forward looking views on the future of machine learning focus on incorporating knowledge[21] or complex data types, temporality, pre-processing and usability[40]. Whilst there have been numerous monistic attempts, they all have severe deficiencies; e.g.:

- Even the general *frameworks* (such as expressing every learning problem as a regularised distance minimisation problem[2]) are limited in terms of the set of problems that can be captured, and they give little insight into the overall *structure* of the sets of problems and do not help the end-user much. Other monistic frameworks are similarly limited: whilst they have served to make significant advances in unifying sets of problems, they do so from a particular viewpoint, and remain limited in scope.
- Attacking the problem via *principles* (e.g. MDL, Bayesian, Frequentist, etc.) is making a category error: these frameworks are ways to solve subsets of problems, not to relate them to each other. And since some of these are nearly a century old, and often vigorously defended[36,68] by their proponents, it seems unlikely much progress will be made in comparing them.

Comparison of Experiments (CoE) is partially similar to our goal but differs in five crucial respects: 1) The CoE setup compares experiments for *all* losses: a much stronger notion of comparison than needed which makes obtaining results difficult. (There is a brief mention of the idea of comparison for *fixed* losses on page 633 of Torgersen's book[64], but he does not pursue the idea far.) 2) The CoE setup uses a model. Machine Learning researchers tend to agree with Kempthorne that "It is a truism, I believe, that there is never an adequate mathematical statistical model for any actual situation,"[59] and thus "model-free" results are preferred. They are obtainable via Reductions. 3) There is a *limited range* of concrete problems addressed — nothing approaching the richness of the list in section E3. 4) There is no notion of a *protocol* — all the experiments are by default batch (although there are results on the "information contained in additional observations")[45]. 5) No attention is paid to issues of *computational complexity*.

Nevertheless there *are* CoE results we believe we can exploit — for example Liese and Vajda's equivalence between Le Cam deficiency and weighted integrals of f-divergences[47].

The work closest to that proposed is the existing reduction results. There are clear differences though: there is no formal statement of all of the problems; most of the reductions are not clearly stated in a way that actually enables their fully modular use; and the very notion of a reduction has only been defined rather informally to date. There is no connection to representation questions or comparison of experiments. Whilst there is the beginning of a big picture[44], it is limited in scope, and only covers a small subset of learning problems. Nevertheless we expect a component of the proposed research to involve the development of new reductions, relating reductions to representational results (for example viewing reductions as "quantised approximations of representation results"), formalising, codifying and extending existing results, and generating "higher-order" results about reductions (for example considering reductions and representations in terms of embeddings between metric spaces, along the lines of the classical embedding theorems).

Significance for the Discipline The significance of the research is threefold: 1) It will *set a research agenda* that will exert influence well beyond the duration of the grant and, we expect, will change the way basic machine learning is taught. 2) It will *reduce the amount of re-invention* and focus research on areas that need it; it will guide the development of machine learning software by indicating where to put the most effort; it will cross-fertilise different parts of the field with ideas from other parts and from other disciplines; it will develop new tools; and it will formulate new problems and thus avoid Hand's "errors of the third kind"[32] thus making it easier for users to solve their real problems. 3) It will provide a conceptual basis to allow for the development of compositional inference techniques for the "web of data"[33] that are easier to use.

New Methodologies and Technologies The main new methodology will be the language and techniques for relating problems to each other, and consequently being able to pose new problems, and to develop new solutions to old and new problems. This will enable a compositional machine learning technology in contrast to the current "start-from-scratch." By building a "map" of all existing machine learning problems, it will create a new research methodology for people studying new problems — first position their problem relative to others on the map.

E4 Approach and Methodology

This is clearly an ambitious proposal. How can it actually be achieved?

Crucially we are *not* trying to subsume every solution to every problem in one massive unified theory. That is both impossible and unnecessary. Instead we are focussing on the problems, and in particular how they relate to each other. By developing an understanding of that it will be possible to bring order to the field and to develop new solutions (by a modularised approach). The key outputs of the project will be the uniform cataloging of problems (and the language to do so), the development of relations between them, and the exploitation of these in the form of new solutions to new (and old) problems. We explain our approach at two levels: an *overall workplan* and *technical details* (the conceptual framework) of the proposed work.

The proposal will exploit the investigators' deep and broad experience in a range of aspects of machine learning. The work summarised in section D2 gives us confidence that the approach will yield significant results.

Overall Workplan / Strategy / Timeline There is a logical dependence between the four aims of the proposal: 1) *language*; 2) *cataloging* of all problems; 3) *relations* between problems; and 4) *tools* — one needs the language in order to define the problems, and one needs the tools to develop the relations. But there is a reverse dependence as well: the language needs to be developed in a manner that ensures all the problems can be efficiently represented and reasoned with, and the tools need to be developed knowing the type of relations sought. Getting the language right can help enormously[35].

This suggests an *iterative high-level work plan*. A small set of problems will be examined first. We will start with some of those listed in section E3 and develop a way of expressing them as clearly as possible. Then work on the various relations and necessary tools will proceed. The language can be adapted as difficulties are encountered. Then a larger set of problems will be considered etc. This staged approach mitigates risk and delivers impact early on. Obtaining some early successes (solving real problems) will be essential to convince the community of the value of the proposed approach.

Conceptual Framework / Technical Details The framework builds upon the work explained in D2 but goes significantly beyond it.

Language and Cataloging: The *process* of cataloging problems is straight-forward but substantial. It will entail a systematic reading of the machine learning literature, and, crucially, will involve a broader study of the use of machine learning (in applications literature) as well as "field work" involving interacting directly with end-users of machine learning. The purpose of this is to try and understand the real problems that need solving, rather than pre-digested versions (which tend to involve the force-fitting of existing solutions to the actual use-inspired problem).

When considering just one problem, the choice of language or notation is usually trivial. The challenge arises from wanting to express a large set of problems consistently. One needs the right conceptualisations / componentisation, notation and definition of spaces, and a way of expressing

protocols. For the relatively simple problems enumerated in E3, the componentisation will entail (at least) *Inputs* (e.g. an input space $\mathcal{I}$ in which inputs i live); *Labels* $\mathsf{l} \in \mathcal{L}$ (e.g. the labels in a traditional binary classification, where $\mathcal{L} = \{0, 1\}$); *Queries* $\mathsf{q} \in \mathcal{Q}$ on which the output of the learner is tested on (e.g. for classical induction $\mathcal{Q} = \mathcal{O}$); *Outputs* of the learned hypothesis $\mathsf{o} \in \mathcal{O}$; *Auxiliary Information* (e.g. noise models); *Evaluator* (for example, binary classification $\mathbb{E}_{(\mathsf{i},\mathsf{l})\sim\mathbb{P}_{\mathcal{I}\times\mathcal{L}}}\ell_{0-1}(\mathsf{l}, h(\mathsf{i}))$, where h is the *hypothesis* produced by the learner, and ℓ_{0-1} is the 0-1 loss); and *Protocol* (a procedural specification of how the learning algorithm and the world interact). Additionally there is the underlying *Model* M (logically distinct from the problem but needed for the use of a problem in practice). These components may be parameterised (e.g. number of labelled and unlabelled examples). With a suitable language and notation, the tabulation of all the problems will be a straight-forward but substantial undertaking, although we expect to continue to encounter multiple variations of the "same" problem, and new problems in-between existing ones.

Relations Getting the language right has logical primacy, but the majority of the research will involve developing new relations and tools, and understanding new problems that arise as a consequence of the systematic program. There is no silver bullet — no single technique that will allow us to achieve the overall aim. We will be eclectic in *tools* and *techniques*. A starting point will be the search for different representations of problems in terms of primitives (confer D2). We also expect to make use of techniques from the comparison of experiments[64], standard methods of analysis of machine learning problems (uniform law of large numbers techniques[68], online analysis methods[13]), approximation theory, functional analysis, and convex analysis.

Starting Points Amongst the diverse set of issues to consider, the following are some of the natural starting points where we are confident it will be possible to make early progress towards the goals of the project. Most of these are "orthogonal dimensions" that can be effectively investigated in parallel.

Classifying Types of Relations: An early task will be to classify the different *types* of relations to be considered. A starting point will be to look at variants of comparison of experiments (say with fixed losses) for binary problems[63], representations, reductions (between single problems, single to sets, sets to sets, etc), conditional reductions, reductions with side information, higher-order relations (e.g. between De Groot statistical information and Le Cam deficiency), distances and embeddings, and relations between distances[26]. Such a taxonomy will be a fecund source of new questions to ask.

Problems that utilise conditioning, loss estimation and ancillarity: A crucial aspect of machine learning algorithms (and their performance guarantees) is whether they "condition on the data." That Bayesian techniques automatically do condition is an oft-used argument for their superiority. But the issue is far more subtle than Bayesian philosophy or not[56]. This is actually a difference in *problems*, not just *techniques*. There is a rich literature on conditional procedures within a frequentist decision theoretic setting[25]. The CI jointly re-developed a more general notion of this independently[34]. Related to this are algorithms (estimators) that in addition to estimating the quantity of interest, provide an estimate of its quality (and even the quality of *that* estimate)[24]. There is a (seemingly forgotten) statistical literature[58] and a re-discovery of the notion in the Machine Learning literature under the guise of "self-bounding" learning algorithms[42] or "confidence procedures" and multiple predictions sets[69]. These methods are effectively utilising the notion[25] of "*ancillary statistics*, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it"[23]. This provides a broad starting point to consider all the existing standard machine learning problems and explore them from a conditioning point of view; that is to develop new problems that take account of conditioning, and relate them to those that do not.

Unifying Representation results: A powerful device to bring order to a disparate set of problems is to represent them all as some form of combinations of primitives. As explained in section D2 this is now well understood for simple binary experiments where one is interested in class probability estimation. Recent work by Lambert et al[41] has shown that the integral representations for proper scoring rules extend to a wider range of problems where one is eliciting information. We plan to explore the implications of this in detail. For example, we will study the representations for a much richer range of machine learning problems (a subset of the long list in section E2), identifying the primitives, relating weight functions, constructing concrete reductions and exploring implications (analogous to the surrogate regret bounds and Pinsker bounds mentioned in D2). We will endeavour

to view many existing reductions as approximate versions of exact representations. This will open the door to the utilisation of approximation theoretic machinery for their analysis. Based on the success to date, we expect to develop a deep understanding of the relationship between these problems by exploiting their representations in terms of primitives. Starting points will be multi-class classification, Neyman-Pearson classification, semi-supervised classification, quantile regression and clustering.

Classifying and Relating Types of Noise: Traditional wisdom is that there are only a few different types of noise in learning problems[61]. A closer investigation reveals the situation is more complex, with only some of the possibilities examined to date[17]. For example, semi-supervised[14] binary classification is a limiting case of binary classification with known variable label noise (unlabelled points corresponding to probability of a label flip of $\frac{1}{2}$). We will investigate problems associated with label noise, attribute noise, distribution noise with uniform noise rate, variable noise rate (either known or unknown, depending on the sample index or sample value) for standard or cost-sensitive problems. We will consider the connections to models such as agnostic learning and to theoretical models of when unlabelled examples help[9]. We will also consider the "duality" between the Statistical Queries learning model (where the labels are probabilities and predictions $\{0, 1\}$-valued) and scoring rules[27] (where the labels are $\{0, 1\}$-valued and predictions are probabilities) in the context of noise.

Formalising Inductive principles: An "inductive principle" means at least two things in the literature. Vapnik[66] means a recipe by which one turns a problem that is insoluble empirically (e.g. minimise an expected loss) into one which can be solved (minimise an empirical loss) with a provable relationship between the solutions. Elsewhere there are notions such as the conditionality principle[15] and the likelihood principle[6]. One can study these principles from an "external" perspective — such as frequentist analyses of Bayesian procedures[60], or one can try to understand "principles" from an engineering perspective. If Vapnik's inductive principles are recipes, then it should be possible to codify them formally as functions mapping between mathematical objects. Are such principles simply "higher order reductions or relations"?

Timeline Because of its broad scope, we envisage this project will take longer than a typical discovery grant. We have sought 5 years of funding because this is (deliberately!) a large scale endeavour which requires both the development of new frameworks and their exploitation, across a diverse problem base. Although we expect the project will take 5 years, it will deliver results and impact early on. The starting points mentioned above will be commenced in the first year and we expect results by the second. The language framework should be cemented by the middle of the second year. In years 3 and 4 we will be extending the work to a wider range of problems and by year 4 we expect to have enthused other researchers around the world to be contributing to the high level goals.

E5 National Benefit

Significance The significance of the proposed research is that it will reconceive the field of machine learning, turning it from a craft to an Engineering discipline. It avoids a single-minded "one true way" and embraces a diversity of principles and solutions.

Expected Outcomes The expected outcomes include a view of the field as a whole which will lead to understanding the best way to set up problems that are more relevant and valuable in practice. It is expected that new solution techniques will be developed by approaching problems from different perspectives and seeing how they relate to each other. It is expected that by translating results from one approach / problem to the other, significant specific technical advances can also be made.

Likely Impact The likely impact is that it will influence the way Machine Learning is researched, taught and used. Instead of a grab-bag of techniques that are developed in many cases without really understanding the problem, the new framework will shift the focus to understanding how the different techniques relate to each other. This should significantly assist students and practioners wishing to learn the discipline, and thus create substantial impact outside of the research community. It is likely to influence the way Machine Learning software is constructed — for example, by understanding which problems can be efficiently reduced to others, one can determine where effort is best expended in creating ultra-efficient implementations. And by understanding the semantics of problems clearly and their relationships, a more compositional machine learning technology which is suitable for the

web of data can be developed. Finally it can serve to focus national research efforts in Machine Learning by providing a framework for groups with varied approaches to work together.

Priority Goals The proposal contributes to two Priority Goals in the National Research Priority of Frontier Technologies for Building and Transforming Australian Industries: 1) *Frontier Technologies*, specifically in assisting in advancing ICT. The proposal resonates with the recognition that "Also important are advanced frameworks such as complex systems"[19] (One could view the proposal as a new framework.) 2) *Smart Information Use*, which "involv[es] improved data management ... [which can] provide huge opportunities to improve the performance of key Australian industries"[19].

Economic Benefit Machine Learning is pervasively used in industry and underpins the digital economy[16], and the proposed reconception can provide economic benefit by aiding industry solve the right problems and have access to internationally competitive techniques. A pathway to this will be the collaboration on NICTA's Elefant platform, which is being developed in a manner to facilitate its use by Australian industry. The "field-work" will ensure connection with real industry problems.

E6 Communication of Results

The broad reconception proposed demands a concomitant communication strategy. The goal is both "inreach" and outreach: to influence the discipline (change the way Machine Learning is researched and taught), and to provide a better way for outsiders to use the results of Machine Learning.

Scientific Journals and Conferences: We will publish the key technical results obtained in tier 1 machine learning journals and conferences such as *Journal of Machine Learning Research*, *IEEE Transactions on Information Theory*, NIPS, COLT and ICML.

Workshops: The subject suits a workshop well. We plan to organise several workshops at NIPS analogous to the *(Ab)use of Bounds* workshop the CI co-organized in 2004, and to organize a Dagstuhl seminar which will create impact by influencing leaders in the field.

Book: Having criticised existing texts, we are motivated to write a better one based on the proposed reconception. This would influence teaching at the undergraduate, postgraduate and professional level, and have a wide impact beyond the particular discipline of machine learning.

Collaborations: The nature of the proposal makes collaborative work with a range of specialists easy. This will promulgate the "meme" of reconception and is thus a dissemination and impact strategy. In addition to various international machine learning researchers, we expect a fruitful collaboration with the NICTA team developing a significant open-source machine learning platform.

Fieldwork: Directly interacting with users of machine learning will also communicate the results. The current proposal provides a theoretical complement to the Linkage Project proposal *Structures and Protocols for Inference* which provides a condit to CISRA. Further conduits will be sought.

Direct publishing to web: We expect there will be a demand and opportunity to provide an up-to-date repository of results within the new framework (like Johnson's[38] ongoing catalog of complexity results). We will maintain a website to do this. It will commence in year 1.

Posters: The proposal will enable the distillation of the whole discipline of Machine Learning to an A0 poster that would be attractive enough for most machine learning researchers and users to want to stick it on their office wall. It would distill the main problems and represent their fundamental relationships. We will prepare such posters as way of promulgating the new way of looking at the field and distribute them at conferences. The first version will be produced after 2 years.

E7 Role of Personnel

Williamson will lead, manage and coordinate the project and contribute to all topics mentioned. Herbrich will contribute to the taxonomy, language and representation results. Von Luxburg will concentrate on problems relating to clustering. Grünwald will focus on ancilliarity, conditioning and conditional relations. The exact role of the Research associate will depend on the particular expertise, but we expect a core role will be codifying problems within the language and building and maintaining catalogues of representations and relations. The PhD students will work on more narrowly focussed projects centered on subsets of problems.

E8 References

[1] M. Allarakhia and A. Wensley. Systems biology: A disruptive biopharmaceutical research paradigm. *Technological Forecasting & Social Change*, 74(9):1643–1660, 2007.

[2] Y. Altun and A. Smola. Unifying Divergence Minimization and Statistical Inference via Convex Duality. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2006.

[3] Carliss Y. Baldwin and Kim B. Clark. Modularity in the design of complex engineering systems. In Dan Braha Ali Minai and Yaneer Bar Yam, editors, *Complex Engineered Systems: Science Meets Technology*. Springer, 2006.

[4] Vic Barnett. *Comparative Statistical Inference*. John Wiley and Sons, Chichester, 3rd edition, 1999.

[5] Alina Beygelzimer, John Langford, and Bianca Zadrozny. Machine learning techniques — reductions between prediction quality metrics. Preprint, March 2008.

[6] A. Birnbaum. On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

[7] Allan Birnbaum. On the foundations of statistical inference: Binary experiments. *The Annals of Mathematical Statistics*, 32(2):414–435, June 1961.

[8] D. Blackwell. Comparison of Experiments. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102, Berkeley and Los Angeles, 31 July – 12 August 1951. University of California Press.

[9] A. Blum and M.F. Balcan. An augmented PAC model for semi-supervised learning. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-supervised learning*. MIT Press, 2006.

[10] Olivier Bousquet. Making machine learning more scientific. http://ml.typepad.com/machine_learning_thoughts/ 2006/06/making_machine_.html, 2006.

[11] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, Mass., 1999.

[12] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications. Technical report, University of Pennsylvania, November 2005.

[13] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[14] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. MIT Press, 2006.

[15] D. R. Cox. Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics*, 29(2):357–372, 1958.

[16] DBCDE. Digital economy future directions consultation paper. 18 December 2008.

[17] S.E. Decatur. *Efficient Learning from Faulty Data*. PhD thesis, Harvard University, Cambridge, MA, July 1995.

[18] M.H. DeGroot. Uncertainty, Information, and Sequential Experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

[19] DEST. Frontier technologies for building and transforming australian industries. November 2003.

[20] P. Domingos and M. Richardson. Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields*, pages 49–54, 2004.

[21] Pedro Domingos. Toward knowledge-rich data mining. *Data Mining and Knowledge Discovery*, 15:21–28, 2007.

[22] Bradley Efron. The Future of Statistics. http://www-stat.stanford.edu/∼brad/talks/future.pdf, July 2007.

[23] R.A. Fisher. The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, 98:39–82, 1935.

[24] RA Fisher. The underworld of probability. *Sankhyã*, 18:201–10, 1957.

[25] D.A.S. Fraser. Ancillaries and conditional inference. *Statistical Science*, 19(2):333–369, 2004.

[26] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

[27] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

[28] Robert L. Grossman, Mark F. Hornick, and Gregor Meyer. Data mining standards initiatives. *Commun. ACM*, 45(8):59–61, 2002.

[29] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

[30] Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

[31] D. J. Hand. Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1):1–14, 2006.

[32] D.J. Hand. Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):317–356, 1994.

[33] Tom Heath. How will we interact with the web of data? *IEEE Internet Computing*, 12(5):88–91, 2008.

[34] R. Herbrich and R.C. Williamson. Algorithmic Luckiness. *Journal of Machine Learning Research*, 3(2):175–212, 2002.

[35] K.E. Iverson. Notation as a Tool of Thought, 1979 ACM Turing Award Lecture. *Communications of the ACM*, 23(8):444–465, 1980.

[36] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
[37] R. Jenssen. *An Information Theoretic Approach to Machine Learning*. PhD thesis, Department of Physics, University of Tromsø, Norway, May 2005.
[38] David S. Johnson. NP-Completeness Columns. *Journal of Algorithms; ACM Transactions on Algorithms*, 2–13; 1–3, 1982–1992; 2005–2007. http://www.research.att.com/∼dsj/columns/.
[39] J. Kiefer. The foundations of statistics — are there any? *Synthese*, 36:161–176, 1977.
[40] Hans-Peter Kriegel, Karsten M. Borgwardt, Peer Kröger, Alexey Pryakhin, Matthias Schubert, and Arthur Zimek. Future trends in data mining. *Data Mining and Knowledge Discovery*, 15:87–97, 2007.
[41] Nicolas Lambert, David Pennock, and Yoav Shoham. Elicitability. In *Proceedings of the ACM Conference on Electronic Commerce*, 2008.
[42] J. Langford and A. Blum. Microchoice Bounds and Self Bounding Learning Algorithms. *Machine Learning*, 51(2):165–179, 2003.
[43] J. Langford, R. Oliveira, and B. Zadrozny. Predicting Conditional Quantiles via Reduction to Classification. In *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence*. AUAI Press, 2006.
[44] John Langford. Machine learning reductions tutorial. Slides presented at the *Machine Learning Summer School*, July 2006.
[45] L. LeCam. On the information contained in additional observations. *Ann. Statist*, 2(4):630–649, 1974.
[46] Lucien LeCam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.
[47] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.
[48] Bruce G. Lindsay, Jon Kettenring, and David. O. Siegmund. A Report on the Future of Statistics. *Statistical Science*, 19(3):387–413, 2004.
[49] John Locke. *An Essay Concerning Human Understanding*. Thomas Basset, London, 1690.
[50] David Mease, Abraham J. Wyner, and Andreas Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8:409–439, 2007.
[51] Denis Noble. *The Music of Life: Biology Beyond the Genome*. Oxford University Press, 2006.
[52] J.R. Platt. Strong inference. *Science*, 146(3642):347–353, October 1962.
[53] David Pollard. Some Thoughts on LeCam's Statistical Decision Theory. Preprint, Statistics Department, Yale University, http://www.stat.yale.edu/∼pollard/Papers/thoughts.pdf, May 2000.
[54] Sarunas Raudys. *Statistical and Neural Classifiers: An integrated approach to design*, chapter Taxonomy of Pattern Classification Algorithms. Springer, London, 2001.
[55] Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. arXiv preprint arXiv:0901.0356v1, 89 pages, January 2009.
[56] N. Reid. The Roles of Conditioning in Inference. *Statistical Science*, 10(2):138–157, 1995.
[57] Christian P. Robert. *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer, New York, 1994.
[58] Andrew L. Rukhin. Estimated loss and admissible loss estimators. In *Statistical Decision Theory and Related Topics IV*, volume 1, pages 409–418, 1988.
[59] L. J. Savage, George Barnard, Jerome Cornfield, Irwin Bross, George E.P. Box, I. J. Good, D. V. Lindley, C. W. Clunies-Ross, John W. Pratt, Howard Levene, Thomas Goldman, A.P. Dempster, Oscar Kempthorne, and Allan Birnbaum. On the Foundations of Statistical Inference: Discussion. *Journal of the American Statistical Association*, 57(298):307–326, 1962.
[60] John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *COLT '97: Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, New York, NY, USA, 1997. ACM Press.
[61] Robert H. Sloan. Four Types of noise in data for PAC Learning. *Information Processing Letters*, 54:157–162, 1995.
[62] L.N. Soldatova and R.D. King. An ontology of scientific experiments. *Interface: The Journal of The Royal Society*, 3(11):795–803, 2006.
[63] E.N. Torgersen. Comparison of experiments when the parameter space is finite. *Probability Theory and Related Fields*, 16(3):219–249, 1970.
[64] E.N. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.
[65] P. Turney. Types of cost in inductive concept learning. In *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, pages 15–21, 2000.
[66] V. Vapnik. *Empirical Inference Science*. Springer, 2006. Afterword of 2006 to [67].
[67] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 2006. 2nd edition.
[68] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
[69] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.
[70] Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, New York, 1950.
[71] Alfred North Whitehead. *Process and Reality*. MacMillan, New York, 1929.
[72] A.P. Worthen and W.E. Stark. Unified design of iterative receivers using factor graphs. *IEEE Transactions on Information Theory*, 47(2):843–849, 2001.