# Form 101, Part II: Research Proposal

## 1 Framework

Many many fields have massive amounts of data — ranging from log files from web browsers, to telemetry data from equipment, to hospital records. The patterns within this data can be very valuable; *e.g.*, these patterns may identify which machine needs prventative maintanence, which sites a user may want to visit and which products s/he is likely to purchase, or identify which patients will respond well to a given treatment. Unfortunately, these patterns may be difficult to extract, due to both the quantity, and complexity, of the data.

The field of machine learning (ML) provides many technologies, and tools, for this task. Much of this work involves understanding, developing and using tools that take as input a "labeled dataset" of instances, and returns a classifier, which can map novel instances to labels. To be more concrete, imagine our goal is to determine whether a given patient will respond well to some treatment. To do this, we run a microarray of a biopsy taken from this patient, which provides the "gene expression level" (a real number) of thousands of genes, $\{g_1, \ldots, g_G\}$. We want a classifier that can map these thousands of values to a label: "1" if the treatment is effective for the current patient (whose biopsy was microarray-ed) and "0" otherwise.

Unfortunately, no one knows this microarray-to-bit ($\Re^G \mapsto \{0, 1\}$) mapping. Fortunately, however, we often have a "labeleld dataset": here a matrix whose rows correspond to the patients (the $i^{th}$ row to the $i^{th}$ patient, for $P$ different patients) and whose first $G$ columns each correspond to a specific gene. Hence, the $m_{p,g}$ value of this matrix is the expression level of gene $g$ for patient $p$. The $G + 1^{st}$ column records the $\{0, 1\}$ labels for the patients. There are many ML tools that take such $(G + 1) \times P$ matrices as input and produce as output a $\Re^G \mapsto \{0, 1\}$ mapping.

Much of my recent research has focused on ways to apply such ML systems to challenges in bio- and medical-informatics, covering the gambit from automated tools that learn patterns in Magnetic Resonance images of brains that correspond to brain tumors, to systems that correlate patterns of Single Nucleotide Polyomorphisms with breast cancer and other patient states, learning properties of proteins (eg, general function, subcellular location, and appearance within specific metabolic pathways) and recently relating diseases to metabolomic profiles and microarray data. I plan to continue applying ML techniques to such bio- and medical-informatics tasks. This proposal summarizes several of these planned activities, as well as some of the foundational issues I will explore over the next 5 years, focussing on the tasks that this NSERC grant will fund.

## 2 Bio- and Medical- Informatics Tasks

**A. Learning within the Human Metabolome Project:** One set of projects is related to the human "metabolome" — the set of all small endogamous and exogenous chemicals that appear in a non-trivial quantity in people. Our team will continue to address a variety of ML challenges here, such as learning properties (like solubility or melting point) of a chemical based on its formula. We plan to use NSERC funds to address the following, more foundational questions:

(1) Finding ways to quickly and autonomously obtain a person's "metabolic profile" (the concentrations of many metabolites) from a single NMR or MassSpec scan; this will involve novel combinations of signal processing and signal interpretation, to correctly interpret the complex mixtures of metabolites that appear in the patient's urine or blood.

Our preliminary work here (with former MSc student J Newton, incorporated in the algorithm initially used by the ChenomX software **[J12]**) naively assumed that each metabolite had a single NMR "signature" that, for a fixed temperature and pH, was invariant; this led to an algorithm, based on linear least squares, for quickly estimating the concentrations of dozens to hundreds of metabolites in a mixture. Experiments have since shown that each metabolite's signature in

fact changes with the concentrations of non-organic compounds (ions and metals) in the mixture. We (with both UofA students and ChenomX employees) plan to explore ways to use the overall spectrum to estimate the concentrations of these inorganic compounds, which will then help us to assess the metabolite concentrations.

This will require first building a dataset that shows the position of each compound's peaks, for various ion concentration level. The obvious way to do this involve collecting spectra of thousands of metabolites, at each of dozens of parameter settings for the various non-organics. Instead, we plan to use ideas from *budgeted learning* (see **(E)** below) to determine which subset of the samples to collect.

(2) How to learn a classifier that can predict a patient's state (eg, specific disease) based on such metabolic profiles. Preliminary results (with V Baracos [Alberta Cancer Board, ACB], members of the UofAlberta Hospital and summer student J Wagner) demonstrated that one can use a patient's profile to reliably predict whether s/he will exhibit cachexia, or bacterial pneumonia. We plan to extend this body of techniques to other diseases. We will also consider ways to enhance the profile information, by exploiting the observation that various metabolites are connected within metabolomic pathways. For example, imagine a disease will typically block a metabolomic step that changes metabolite A to B. This suggests we should expect an overabundance of A and reduction in B, and means that the ratio of A/B should be much higher in patients with this disease, compared to the base population. We have observed this in some studies, based on ratios of compounds determined by hand. We will explore ways to *discover* the connections by analysing such metabolic pathways, and also go the other way: use the connections found to propose possible metabolic associations with diseases. This requires addressing the several foundational question, such as how best to use this graphical "pathway" information.

**B. PolyomX Project:** The PolyomX project extends this theme of finding biological markers that help to identify a patient's state. This project, however, considers a larger array of possible markers, including genetic and proteomic indicators as well as metabolomic.

N Asgarian (MSc student) and I have begun to explore algorithms that classify a patient based on a microarray of a biopsy taken from his/her tumor. One of our current algorithms is based on a novel bi-clustering technique. As above, we let $m_{p,g}$ be the observed expression level of gene $g$ for patient $p$, and then $\bar{p} = \langle m_{p,g_1}, m_{p,g_2}, \ldots, m_{p,g_m} \rangle$ be the set of expression values of all $m$ genes for patient $p$. Standard *clustering* algorithms will identify various sets of patients whose gene expression values are *all* correlated with one another — *i.e.*, for all patients $p_i$, $p_j$ within cluster $c_1$, the values of $\bar{p_i}$ are correlated with $\bar{p_j}$. By contrast, a *bi-clustering* algorithm will identify a specific subset of the patients for which a specified subset of the genes are correlated — *e.g.*, for patients $\{p_1, \ldots, p_7\}$, (only) the genes $\{g_{100}, \ldots, g_{117}\}$ are correlated. This is easy to motivate from biological perspective — *e.g.*, if something happening to these patients that caused the expression levels for this small subset of their respective genes to become correlated. Our initial studies suggest that this algorithm is very effective: it has been able to achieve over 90% accuracy on several standard databases where the previously best known results was only 70-80%. Moreover, this accurate classifier is based on one or two bi-clusters, that collectively involve only a handful of genes. We plan to use NSERC funds to better understand when this (class of) bi-clustering algorithms, to help determine whether it will effective on other domains — *e.g.*, for SNP analysis, metabolic profiles, and perhaps for machine learning tasks in general.

In the future, we will design yet other algorithms for predicting many other diseases, from these modalities (genomic/SNP, microarray, and metabolomic) and others, as well as ways to combine information from multiple diverse modalities. This will require further advances in dimensionality reduction, even to deal with a single modality, and techniques like probabilistic graphical models

(see **(F)** below) to combine the results of the different modalities.

**C. Proteome Analyst Project:** A third related research project is exploring ways to learn properties of proteins. Our team has already developed world-class predictors of general function **[J6]**,**[J7]** and subcellular location **[J9]**, and have begun to explore ways to use related techniques to predict which proteins participate in which metabolic pathways **[J1]**. We plan to further improve these related systems by both incorporating other factors (such as hierarchical information **[C12]**, as well as natural language information that appears in abstracts) and by improving the underlying learning and classification systems.

The database already produced is one of the most accurate, and comprehensive known **[J6]**. We plan to explore this data itself. For example, we can quickly determine "organism" statistics: *e.g.*, what percentage of a *human*'s proteome is involved with transport, and what percentage is located in the nucleus; and get similar information about other model organisms, such as *rat* or *E.coli*. We can use this list of "subcellular-location + general-funtion" percentage-values to characterize each species, then use this information to learn effective phylogenetic trees, or perhaps to help identify yet other protein properties, based on the gross organism characteristics. We will use NSERC funds to help understand ways to incorporate these organism-level properties, when making individual predictions.

**D. Brain Tumor Analysis Project:** The classifiers (and regressors) used in the previous projects have all dealt with "independent and identically distributed" instances: *e.g.*, the chance that one patient will be cachexic is independent of whether a second patient is. Now consider applying a similar "classification" technology to an imaging task; *e.g.*, given a Magnetic Resonance (MR) image of a patient's brain, we would like to identify which voxels correspond to tumor cells, versus healthy brain cells — *i.e.*, we want a classifier that maps (a description of) each voxel to a label that is either $\{$tumor, healthy$\}$. Here, however, the labels of adjacent voxels are strongly correlated as they tend to have the same label, which means that finding one voxel is tumorous (resp., healthy) increases the chance that its neighbor will be a tumor (resp., healthy). We therefore need to use other techniques to accomodate this task, such as Random Fields (RFs), which involve two potentials: one that tries to identify which voxels, in isolation, correspond to tumor cells, and the other, to help model the "connections" between pairs of voxels.

Our team (including Prof J Sander from CS at UofAlberta and Dr A Murtha, Radiation Oncology [ACB], as well as several former, and curent, students) is using this technology for the general task of attempting to locate tumorous cells within a patient's brain — including both cells in regions that are detectable in a MR image, and also "radiographically occult" tumour cells. To date, most of our efforts have focused on the first task. Most Conditional RFs [LMP01] use potentials based on logistic function; we (with M Schmidt (MSc) and C-H Lee (PhD)) improved this by instead using a variant of Support Vector Machines **[C13]**. We also explored ways to produce similar results *significantly faster* **[C4]**, and of using some unlabeled data **[C2]**. We will use the NSERC funds to continue to address such foundational issues, with the expectation that results here will help with a variety of tasks, in addition to segmenting brain tumors.

Another factor that contributed to our success was our use of appropriate features **[C11]**; here, we augmented the three input MR modalities (T1, T2, T1-contrast) with "alignment-based features" — that is, information obtained after registering the brain images against appropriate templates. To be yet more accurate here, we plan to incorporate additional imaging modalities, including Diffusion Tensor Imaging and MR spectroscopy. These modalities are significantly different, as they produce (respectively) a vector field (*i.e.*, 3 real values at each voxel) and an entire spectrum at each voxel. We anticipate needing new ML and imaging techniques to effectively exploit these new types of information.

We will also contine work designing a system that predicts the "invisible" regions, by exploring many growth+diffusion modeling techniques, both traditional and novel, using techniques that range from CRFs (but trained on a different "labels" than the ones used for the simple segmentation) to differential equations. We have already begun discussing this with Prof T Hillen (UofA Math/Stats), who already has some relevant results here. We will use NSERC funds to help develop and understand general techniques for integrating differential equations with CRFs.

## 3 Theoretical Foundation

While each of the projects listed above has a strong theoretical component, the work itself is heavily motivated by some specific application. In addition to these "application pull" projects, I will use NSERC funds to pursue several "technology push" investigations.

**E. Budgeted Learning:** I will continue to explore a fundamental question in experimental design: Given a limited budget of funds to spend acquiring training information, which specific "datapoints" should we acquire — *e.g.*, which specific tests should we run on which specified training patients? Our preliminary results (with D Lizotte (MSc) and O Madani (PDF) **[C21]**) here formally framed this task and provided preliminary results, both theoretical (showing the task is NP-hard and that many standard algorithms are not approximation algorithms) and empirical — providing several novel algorithms, and demonstrating that they are more effective than the traditional ones **[C24]**. A Kapoor (MSc) and I then examined an obvious extension: seeking the best "fixed-cost" classifier — *i.e.*, how to spend at most $10,000 to learn a $30/patient classifier **[C14]**.

We plan to explore many extensions: One goal is a better understanding of this task in general. For example, our NP-hardness result applies to a certain class of distributions and cost values; we do not know that this task is NP-hard for the simple standard case of unimodal prior distributions and unit-cost tests. We also want to explore the "sample complexity" of budgeted learning: what is the minimum number of probes (each of the form: what is the value of feature $f$ for patient $p$) required to guarantee we have the information required to learn a good classifier. Notice a standard PAC-learning [Val84] result of $O(k(\epsilon, \delta))$ would translate to a lower bound of $O(n \times k(\epsilon, \delta))$ probes, as each "unit" in the PAC-analysis corresponds to acquiring all $n$ features of an instance. We expect that our model will require significantly fewer probes, as we give the learner the option of sequentially deciding which specific individual probe to purchase.

We will also seek more effective, and efficient, algorithms for this task. For example, given a distribution over features and class labels, we can use a dynamic program to produce the optimal "fixed-cost" classifier. Given "naive bayes assumptions", A Farhangfar (PhD), M Zinkevich (PDF) and I recently found a much more efficient algorithm for finding this classifier. We plan to explore ways to use this as a foundation for a budgeted learning system: *i.e.*, for deciding which features to probe, to obtain the distribution that leads to an optimally accurate classifier. Finally, we anticipate the study suggested in **(A)** above will suggest yet other research directions.

**F. Probabilistic Graphical Models:** There are now several types of graphical models commonly used to represent information. One type are Random Fields, including CRFs, which are often used to deal with images; **(D)** above already listed some of our advances related to this technology. We plan to continue exploring ways to improve such systems, to be more accurate, more efficient and more robust to missing data.

Bayesian belief nets [Pea88] are another type of graphical model. We will continue to explore ways to learn such systems, extending our current results **[C35]**,**[J11]**,**[C33]**,**[C18]**,**[C29]**,**[C23]**,**[J5]**.

My work with Math/Statistics Professor P Hooper and MSc students T van Allen and A Singh led to an effective way to estimate the variance around a belief net response **[C31]**; we were able

to use this as part of a tool for combining different belief-net based classifiers **[C6]** (with C-H Lee (PhD) and S Wang (PDF)), and for discriminatively learning belief net structures (with Y Guo (PhD)). Our current system is designed for dealing with discrete values, where the parameters are represented as explicit Conditional Probability tables, which were learned from complete training data. We will explore ways to use a similar analysis for dealing with other types of parameters for discrete variables (such as Noisy-Or), for dealing with *continuous* variables, and for coping with missing training data. We will also explore ways to use this variance information for other tasks, such as Active Learning **[C1]**, and cost sensitive classification [Tur00].

## 4 Expected Significance of the Research

Our research has been cutting edge in machine learning, as evidenced by the large number of publications in the very top ML venues. Other publications demonstrate that we have also made significant contributions in the application areas — notably cancer research **[J2]**,**[J8]**, proteomics **[J1]**,**[J6]**,**[J7]**,**[J9]**, genomics **[J4]** and clinical chemistry **[J12]**. I anticipate that my research plans, outlined above, will continue to contribute to both core ML techology and our various application areas.

## 5 Training of Highly Qualified Personnel

Over the last 6 years, our team has been very sucessful in training — producing several undergraduate students, as well as 18 graduate students, and 3 PDFs, with more en route. Most of these have continued to do research, in either academia (at UofAlberta or elsewhere), or in industry. In addition, many of the larger projects have also involved numerous research programmers, several of whom have gone to graduate school, or to industry, where they are continuing to apply the skills they developed while assisting these projects.

We will continue these policies of using undergraduates, graduates, postdoctoral fellows and research programmers, and fully anticipate this will lead to a steady stream of HQP.