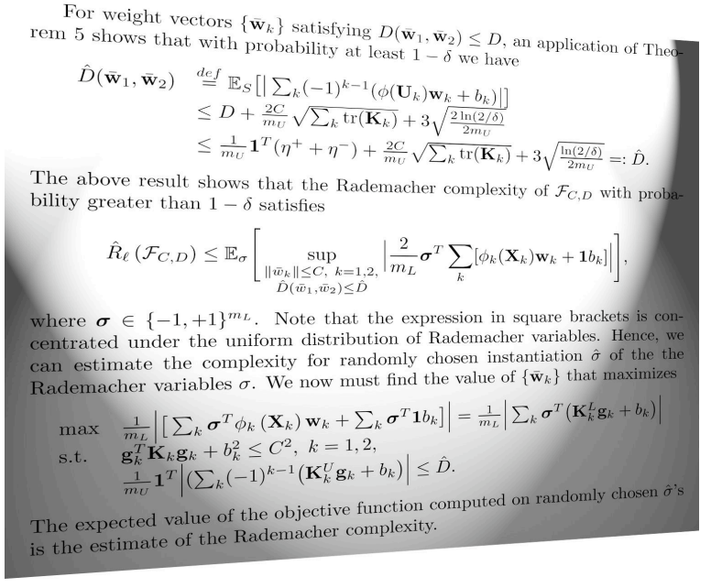
## Normalised Compression Distance Measures and Their Application in Unsupervised and Supervised Analysis of Polymorphic Data

The project researchers have built a generic open-source software package (Complearn) for building tree structured representations from normalised compression distance (NCD) or normalised Google distance (NGD) distance matrix. Preliminary empirical tests with the software have indicated that the data representation is crucial for the performance of the algorithms, which led us to further study applications of lossy compression algorithms (audio stream to midi and lossy image compression through wavelet transformation).

In the second part of the project, the methods and tools developed in the project were applied in a challenging and exciting real-world problem. The goal was to recover the relations among different variants of a text that has been gradually altered as a result of imperfectly copying the text over and over again. In addition to using the currently available methods in Complearn, we also developed a new compression-based method that is specifically designed for stemmatic analysis of text variants.

The various methods developed in the project were applied and tested using the tradition of the legend of St. Henry of Finland, which forms a collection of the oldest written texts found in Finland. The results were quite encouraging: the obtained family tree of the variants, the stemma, corresponds to a large extent with results obtained with more traditional methods (as verified by the leading domain expert, Tuomas Heikkilä Ph.D., Department of History, University of Helsinki). Moreover, some of the identified groups of manuscripts are previously unrecognised ones. Due to the impossibility of manually exploring all plausible alternatives among the vast number of possible trees, this work is the first attempt at a complete stemma for the legend of St. Henry. The new compression-based methods developed specifically for the stemmatology domain will be released in the future as part of the open-source Complearn package. We are also considering the possibility of creating a Pascal challenge using this type of data.



For weight vectors $\{\bar{\mathbf{w}}_k\}$ satisfying $D(\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2) \leq D$, an application of Theorem 5 shows that with probability at least $1-\delta$ we have

$$\begin{aligned}\hat{D}(\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2) &\stackrel{def}{=} \mathbb{E}_S\left[\left|\sum_k (-1)^{k-1}(\phi(\mathbf{U}_k)\mathbf{w}_k + b_k)\right|\right] \\ &\leq D + \tfrac{2C}{m_U}\sqrt{\textstyle\sum_k \mathrm{tr}(\mathbf{K}_k)} + 3\sqrt{\tfrac{2\ln(2/\delta)}{2m_U}} \\ &\leq \tfrac{1}{m_U}\mathbf{1}^T(\eta^+ + \eta^-) + \tfrac{2C}{m_U}\sqrt{\textstyle\sum_k \mathrm{tr}(\mathbf{K}_k)} + 3\sqrt{\tfrac{\ln(2/\delta)}{2m_U}} =: \hat{D}.\end{aligned}$$

The above result shows that the Rademacher complexity of $\mathcal{F}_{C,D}$ with probability greater than $1-\delta$ satisfies

$$\hat{R}_\ell\left(\mathcal{F}_{C,D}\right) \leq \mathbb{E}_\sigma\left[\sup_{\substack{\|\bar{w}_k\|\leq C,\ k=1,2,\\ \hat{D}(\bar{w}_1,\bar{w}_2)\leq \hat{D}}} \left|\frac{2}{m_L}\boldsymbol{\sigma}^T \sum_k [\phi_k(\mathbf{X}_k)\mathbf{w}_k + \mathbf{1}b_k]\right|\right],$$

where $\boldsymbol{\sigma} \in \{-1,+1\}^{m_L}$. Note that the expression in square brackets is concentrated under the uniform distribution of Rademacher variables. Hence, we can estimate the complexity for randomly chosen instantiation $\hat{\sigma}$ of the the Rademacher variables $\sigma$. We now must find the value of $\{\bar{\mathbf{w}}_k\}$ that maximizes

$$\begin{aligned}\max\quad & \tfrac{1}{m_L}\left|\left[\textstyle\sum_k \boldsymbol{\sigma}^T \phi_k\left(\mathbf{X}_k\right)\mathbf{w}_k + \sum_k \boldsymbol{\sigma}^T \mathbf{1}b_k\right]\right| = \tfrac{1}{m_L}\left|\textstyle\sum_k \boldsymbol{\sigma}^T(\mathbf{K}_k^L\mathbf{g}_k + b_k)\right| \\ \text{s.t.}\quad & \mathbf{g}_k^T\mathbf{K}_k\mathbf{g}_k + b_k^2 \leq C^2,\ k=1,2, \\ & \tfrac{1}{m_U}\mathbf{1}^T\left|(\textstyle\sum_k (-1)^{k-1}(\mathbf{K}_k^U\mathbf{g}_k + b_k)\right| \leq \hat{D}.\end{aligned}$$

The expected value of the objective function computed on randomly chosen $\hat{\sigma}$'s is the estimate of the Rademacher complexity.

## Learning with Labeled and Unlabeled Data

Semi-supervised learning belongs to the main directions of the recent machine learning research. The exploitation of the unlabeled data is an attractive approach either to extend the capability of the known methods or to derive novel learning devices. Learning a rule from a finite sample is the fundamental problem of machine learning. For this purpose two resources are needed: a big enough sample and enough computational power. While the computational power has been growing rapidly, the cost of collecting a large sample remains high since it is labour intensive.

The unlabeled data can be used to find a compact representation of the data which preserves as much as possible its original structure.

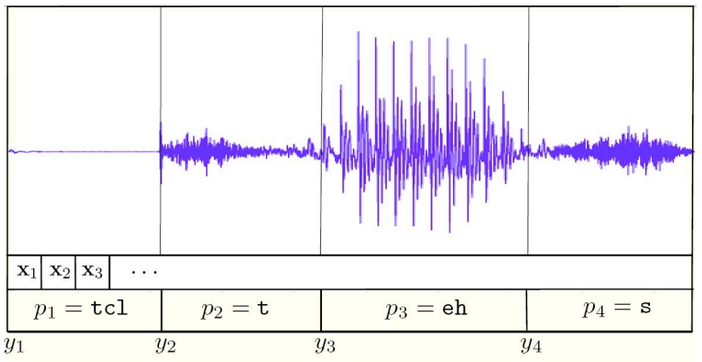
## Dynamics and Uncertainty in Interaction

This PASCAL funded pump-priming project aims to bring continuous and uncertain interaction methods into both brain-computer interfaces, and to the interactive exploration of song spaces. By treating the interaction problem as a continuous control process, a range of novel techniques can be brought to bear on the BCI and song exploration problem.

EEG brain-computer interfaces suffer from high noise levels and heavily-lagged dynamics. Existing user interface models are inefficient and frustrating for interaction. By explictly taking the noise and dynamical properties of the BCI control signals into account, more suitable interfaces can be devised.

The song-exploration problem involves navigation of very high-dimesional feature spaces. The mapping from these spaces to user intention is uncertain. Uncertain and predictive displays, combined with intelligent navigation controls, can aid users in intuitively navigating musical spaces.

This work is in collaboration with the IDA group at Fraunhofer First and the Intelligent Signal Processsing Group at DTU.




## Large Margin Algorithms and Kernel Methods for Speech Recognition

Research on large margin algorithms in conjunctions with kernel methods has been both exciting and successful. While there have been quite a few preliminary successes in applying kernel methods for speech applications, most research efforts have focused on non-temporal problems such as text classification and optical character recognition (OCR). We propose to design, analyze, and implement learning algorithms and kernels for hierarchical-temporal speech utterances. Our first and primary end-goal is to build and test thoroughly a full-blown speech phoneme classifier that will be trained on millions of examples and will achieve the best results in this domain. This project is a joint reseach effort between The Hebrew University and IDIAP.

## Multi-Task Learning: Optimisation Methods and Applications

This project is focused on multi-task learning (MTL) for the purposes of developing optimisation methods, statistical analysis and applications. On the theoretical side, we propose to develop a new generation of MTL algorithms; on the practical side, we will explore applications of these algorithms in the areas of marketing science, bioinformatics and robot learning. As an increasing number of data analysis problems require learning from multiple data sources, MTL should receive more attention in Machine Learning and we expect that more researchers will work on this topic in the coming years.

We are particularly interested in optimisation approaches to MTL. In particular, our proposed approach will: 1) allow one to model constraints among the tasks; 2) allow semi-supervised learning -- only some of the tasks have available data but we still wish to learn all tasks; 3) lead to efficient optimisation algorithms; 4) subsume related frameworks such as collaborative filtering and learning vector fields.

## Online Performance of Reinforcement Learning with Internal Reward Functions

We consider reinforcement learning under the paradigm of online learning where the objective is good performance during the whole learning process. This is in contrast to the typical analysis of reinforcement learning where one is interested in learning a finally near-optimal strategy. We will conduct a mathematically rigorous analysis of reinforcement learning under this alternate paradigm and expect as a result novel and efficient learning algorithms.

We believe that for intelligent interfaces the proposed online paradigm provides significant benefits as such an interface would deliver reasonable performance even early in the training process.

The starting point for our analysis will be the method of upper confidence bounds which has already been very effective for simplified versions of reinforcement learning. To carry the analysis to realistic problems with large or continuous state spaces we will estimate the utility of states by value function approximation through kernel regression. Kernel regression is a well founded function approximation method related to support vector machines and holds significant promise for reinforcement learning.

Finally we are interested in methods for reinforcement learning where no or only little external reinforcement is provided for the learning agent. Since useful external rewards are often hard to come by, we will investigate the creation of internal reward functions which drive the consolidation and the extension of learned knowledge, mimicking cognitive behaviour.