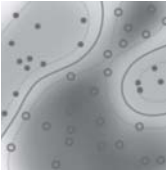
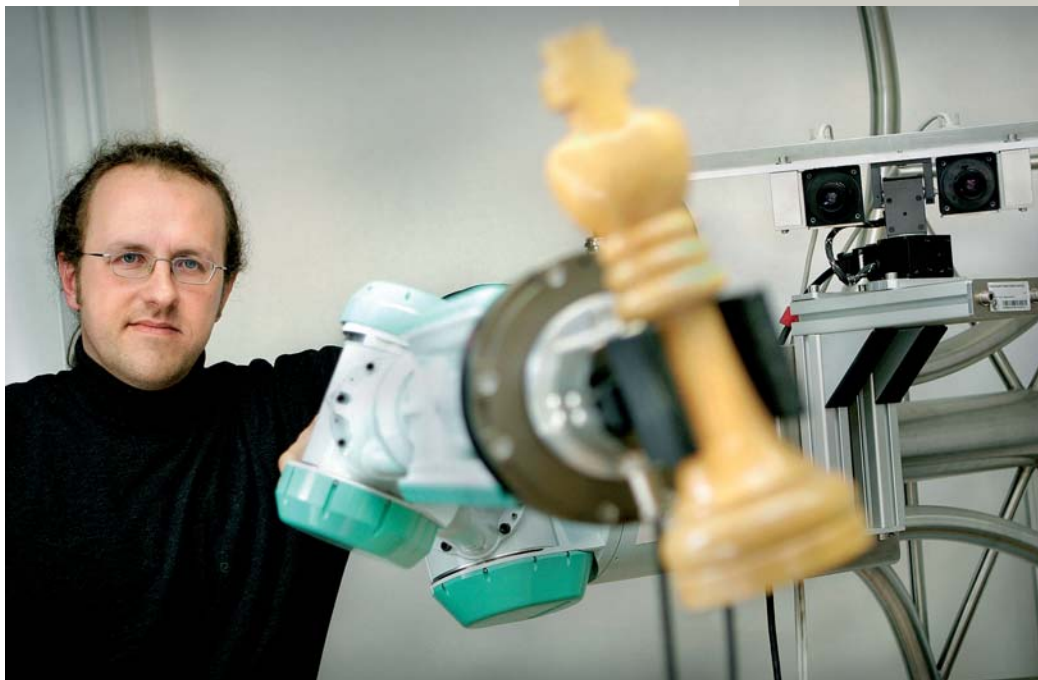




Max-Planck-Institut  
für biologische Kybernetik



# Empirical Inference



Anne Faden

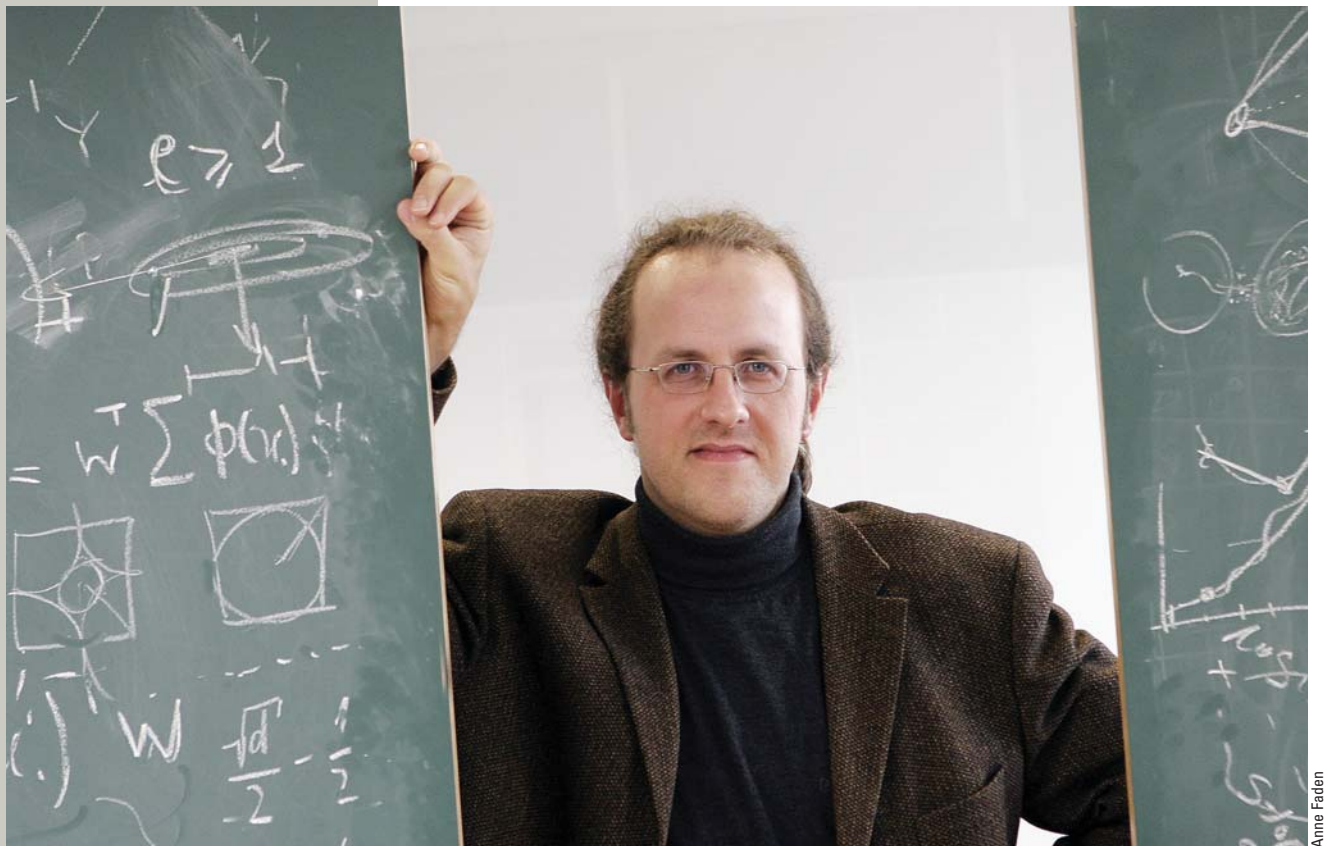
## Max Planck Institute for Biological Cybernetics

### Report 2006 – 2008

Bernhard Schölkopf .....	2
Empirical Inference .....	6
Research Program .....	14
Publications .....	62
Awards .....	74
Provision of Materials, Equipment and Working Space .....	76
Activities Regarding the Transfer of Knowledge .....	78



MAX-PLANCK-GESELLSCHAFT



Anne Faden

## Bernhard Schölkopf

**B**ernhard Schölkopf received degrees in mathematics (University of London, 1992) and physics (Eberhard-Karls-Universität Tübingen, 1994), and a doctorate in computer science (TU Berlin, 1997). He has worked at AT&T Bell Labs, at GMD FIRST, Berlin, and at Microsoft Research Cambridge, UK. He has held visiting fellowships at various labs including the Australian National University, MIT, and RIKEN. In July 2001, he was elected scientific member of the Max Planck Society and director at the MPI for Biological Cybernetics; in October 2002, he was appointed Honorarprofessor for Machine Learning at the Technical University Berlin.

Bernhard Schölkopf's work has been crucial for the inception of the field of kernel machines. Among his main contributions are the insights that every method that can be cast in terms of inner products can be generalized to a corresponding nonlinear method using positive definite kernels, and that kernels can also be applied to data which do not live in a vector space. His work led to record performances on a variety of real, world problems ranging from handwritten digit recognition to gene regulation prediction tasks.

## Philosophy

Empirical Inference, the title of the department, describes the task of drawing conclusions from the observation of empirical data. In its full generality, this arguably subsumes the task that all empirical sciences are facing. It is generally tackled by modeling: based on empirical data and human ingenuity, scientists come up with models that guide further experiments and are continuously refined. At the same time, processes of empirical inference also lie at the heart of what brains accomplish. In Helmholtz's view, perception is subconscious inductive inference, and according to Barlow, the brain is nothing but a statistical decision organ.

The specific strength of many machine learning methods for empirical inference lies in the fact that they only require rather weak model assumptions. The world contains many regularities that present-day science cannot hope to model explicitly, yet they are certainly non-random in that we can build predictors that generalize to future data. Such predictors can often be useful for scientists interested in the regularity being analyzed; however, the main scientific motivation of the field of empirical inference is not the solution of such individual problems, but rather the methodology underlying the inference process: how can we encode weak models, or weak forms of prior knowledge, capturing aspects that are likely to be true for real world data? Moreover, given such knowledge, supplemented with empirical data, how can we arrive at the best prediction or generalization?

Our main interest is to understand and further develop this methodology. This endeavor necessarily entails not only theoretical studies on learning and generalization, but also empirical studies in a variety of problem domains that are characterized by the conditions under which empirical inference can make a difference, i.e., the analysis of regularities that are difficult or impossible to understand mechanistically using present day science. Biology is full of such problems, ranging from neuroscience to bioinformatics.

Our main mode of dissemination is the publication of results at the leading machine learning conferences. These are NIPS (Neural Information Processing Systems), ICML (International Conference on Machine Learning), and for theoretical work, COLT (Conference on Learning Theory). Our presence at these rather competitive conferences makes us one of the top few machine learning labs worldwide. In addition, we sometimes submit our work to the leading application oriented conferences in fields including computer vision (ICCV, ECCV, CVPR), data mining (KDD, ICDM), and computational biology (ISMB, RECOMB).

Our work has earned us best paper prizes at major conferences (COLT 2003, NIPS 2004, COLT 2005, COLT 2006, ICML 2006, ALT 2007, CVPR 2008, ISMB 2008).

## Memberships and Awards

Bernhard Schölkopf is a member of the Institute of Electrical and Electronics Engineers, the Association for Computing Machinery, and the Deutsche Mathematiker-Vereinigung. He is the recipient of the Lionel Cooper Memorial Prize of the University of London (1993), the dissertation award of the German society of Computer science (1998), and the J.K. Aggarwal Prize of the International Association for Pattern Recognition (2006). He is a member of the boards of the NIPS foundation, the International Machine Learning Society, and of Kernel-Machines.org. He serves or has served on committees and boards of the leading scholarly dissemination media in machine learning and computer vision:

- Journal of Machine Learning Research, Machine Learning Journal – the two flagship journals of machine learning
- International Journal of Computer Vision, IEEE Transactions on Pattern Analysis and Machine Intelligence – the two flagship journals in computer vision
- Journal of Artificial Intelligence Research
- SIAM Journal on Imaging Sciences
- Foundations and Trends in Machine Learning
- Information Science and Statistics – a monograph series published by Springer
- Co-chair of the first two kernel workshops in Breckenridge, Colorado (1998, 1999)
- Co-founder of a series of summer schools on Machine Learning ([www.mlss.cc](http://www.mlss.cc))
- Program (co-)chair of COLT'03, DAGM'04, NIPS'05
- General chair of NIPS'06
- Program-Committee member of most major conferences in the field (NIPS, COLT, ICML, UAI, Snowbird Learning Workshop, CVPR, ICDM)

Bernhard Schölkopf has been invited to present talks at a number of major conferences in machine learning, statistics and mathematics, including the American Association for Artificial Intelligence, Madison, Wisconsin, USA (1998), Bernoulli Society, Tokyo (2000), Interface, Orange County (2001), Neural Information Processing Systems, Vancouver (2001), European Conference on Machine Learning & European Conference on Principles and Practice of Knowledge Discovery in Databases, Helsinki (2002), International Statistical Institute, Berlin (2003), Mathematics and Image Analysis, Paris (2004), International Work-Conference on Artificial Neural Networks, Barcelona (2005), International Conference on Pattern Recognition, Hong Kong (2006), International Conference on Machine Learning, Corvallis (2007), and the International Conference on Artificial Neural Networks, Porto (2007).

## Talks

---

2006

### Statistical Learning Theory and Compression

Workshop on Model Selection and Data Fitting

Research School of Information Sciences and Engineering (RSISE), ANU  
Canberra, Australia  
February 24, 2006

---

### Kernel Methods and Applications

Johannes Gutenberg-Universität Mainz  
Institute for Computer Sciences  
Artificial Intelligence Public Group  
Mainz, Germany  
April 25, 2006

---

### Warum sind Computer dumm?

“Kinder-Uni 2006”

Universität Tübingen, Germany  
May 02, 2006

---

### RKHS Methods for Estimation using Nonlinear Function Classes

International Conference for Pattern Recognition (ICPR) 2006 (J.K. Aggarwal Prize Keynote)  
Hong Kong, China  
August 21, 2006

---

### Kernel Machines for Computer Graphics

International Association for Pattern Recognition Conference (ICPR) 2006  
Hong Kong, China  
August 21, 2006

---

### A New Method for Solving the Two-sample Problem

The Hong Kong University of Science and Technology (HKUST)  
Hong Kong, China  
August 22, 2006

---

### Applications of Kernel Methods

Symposium “Positive Definite Functions and Applications”  
Annual Conference of the German Mathematicians’ Association (DMV) 2006  
Bonn, Germany  
September 18, 2006

---

### Kernel Methods for Machine Learning

Mathematical Colloquium  
Universität Tübingen, Germany  
November 06, 2006

---

### The Two-sample Problem and Covariate Shifts

Kick-off meeting to the Fraunhofer Society and Max Planck Society cooperation  
Stuttgart, Germany  
November 20, 2006

---

### Thoughts on Kernels

International Workshop on Current Challenges in Kernel Methods (Keynote)  
Brussels, Belgium  
November 27, 2006

---

2007

### Thoughts on Kernels

24<sup>th</sup> International Conference on Machine Learning (ICML) 2007 (Keynote)  
Oregon, USA  
June 21, 2007

---

### Kernel Tricks, Means and Ends

International Conference on Artificial Neural Networks (ICANN) 2007 (Keynote)  
Porto, Portugal  
September 10, 2007

---

### MPI für biologische Kybernetik

Board of Trustees Meeting  
Max-Planck-Haus, Tübingen, Germany  
October 26, 2007

---

2008

### Kernel Tricks, Means and Ends

16<sup>th</sup> Reconnaissance des Formes et Intelligence Artificielle - Pattern Recognition and Artificial Intelligence (RFIA) 2008 (Keynote)  
Amiens, France  
January 23, 2008

---

### Machine Learning using Positive Definite Kernels

Internationales Graduiertenkolleg Basel-Graz-Tübingen: Hadronen im Vakuum, in Kernen und Sternen  
Universität Tübingen, Germany  
March 27, 2008

---

### Maschinelles Lernen und Empirische Inferenz

Forum Scientarium  
Universität Tübingen, Germany  
April 17, 2008

---

### Positive Definite Similarity Measures and Applications

Cognitive Information Processing Workshop  
Technische Universität München, Germany  
May 30, 2008

---

### Kernel Tricks, Means and Ends

1<sup>st</sup> Workshop on Cognitive Information Processing (CIP) 2008 (IAPR Distinguished Speaker Keynote)  
Santorini, Greece  
June 09, 2008

---

### Kernel Methods for Computational Algebraic Topology

Learning Theory and Approximation Symposium Oberwolfach, Germany  
July 04, 2008

---

### Machine Learning Applications of Positive Definite Kernels

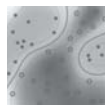
GfKI 2008 – The 32<sup>nd</sup> Annual Conference of the German Classification Society – Gesellschaft für Klassifikation (GfKI) (Keynote) Hamburg, Germany  
July 16, 2008

---

### Kernel Tricks, Means and Ends

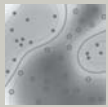
16<sup>th</sup> European Signal Processing Conference (EUSIPCO 2008) (Keynote)  
Lausanne, Switzerland  
August 29, 2008

---





# Empirical Inference



## Research Interests

The problems studied in our department can be subsumed under the heading of empirical inference. This term refers to inference performed on the basis of empirical data. The type of inference can vary, including for instance induction (estimation of models such as functional dependencies) and transduction (predictions or decisions for individual points). Likewise, the type of empirical data can vary, ranging from sparse experimental measurements (e.g., microarray data) to visual patterns. Our department is conducting theoretical, algorithmic and experimental studies to try and understand the problem of empirical inference. The department was started around statistical learning theory and certain recent developments in the field of machine learning, in particular support vector machines (SVMs). It has since broadened its set of inference tools to include a stronger component of Bayesian methods, including graphical models with a recent focus on issues of causality. In terms of the inference tasks being studied, we have moved towards tasks that go beyond the relatively well-studied problem of supervised learning, such as semi-supervised learning or, more recently, structured estimation. Finally, we have continuously striven to analyze challenging datasets from biology and other domains, leading to the inclusion of several application areas in our portfolio. When performed in collaboration with domain experts, such work can be rewarding for both sides and it provides us with additional insights into tasks and methods for empirical inference pertaining to our department's core interests. In cases where the application areas are close to our own expertise, we also carry out application-oriented research on our own. An example thereof is robotics, where we have substantially expanded our activities. No matter whether the applications are done in the department or in collaboration with external partners, considering a whole range of applications helps us study *principles and methods of inference*, rather than inference applied to one specific problem domain.

On the following pages, we briefly describe our research directions in more detail. We begin with a high level overview of the broad directions, followed by short descriptions of individual projects. The overview includes page references to the project descriptions, including some which appear in the sections of this report devoted to other departments and groups. In most cases, the project descriptions contain pointers to our publications in the area, many of which are available electronically from <http://www.kyb.mpg.de/bs/publication.html>.

Theoretical studies, algorithms and applications often go hand in hand. For instance, it may be the case that someone working on a specific application will develop a customized algorithm that turns out to be of independent theoretical interest. Likewise, cross-fertilization and serendipity are desired side effects caused by interaction across groups and research areas, for instance during our frequent departmental talks. The linear organization of the text does not permit an adequate representation of all these connections. Below, we have opted for an organization of the material that devotes individual sections to our largest application areas (vision and image processing, robot learning, neuroscience, and bioinformatics), and that comprises three methodological sections, on kernel algorithms, causal and probabilistic inference, and statistical learning theory. We begin with the latter.

Statistical learning theory analyzes learning algorithms within a statistical framework. In the supervised learning case, which has been the focus of most studies, the question is how to estimate a functional dependency between inputs and outputs based on empirical data, i.e., a set of observed input-output pairs. The only assumption that is made about the data-generating process is that all the examples in the dataset are generated independently from the same source (characterized by a fixed probability distribution). This means that the only relationship that exists between the training data given to the algorithm and the data it will later be tested on is that they are generated from the same distribution. In contrast to classical statistics, no (or only weak) assumptions are made on the nature of this probability distribution.

The goal is to identify the best possible model from a set of competing ones. The criterion of interest is the degree to which the model succeeds in capturing the regularities in the data generating process. This is measured in terms of the generalization error on previously unseen examples. As the generalization error cannot be assessed on the training data, one of the main goals of learning theory is to develop methods with which we can derive upper bounds on it. These bounds only depend on the class of functions that are used to describe the inherent relationships and on quantities that can be measured on the training data. They express a trade-off between the fit of the model to the data and the richness of the model. Loosely speaking, if we can explain the data using a complex model, then this is not surprising, since complex models can explain many possible datasets. If, however, we explain the data using a model which *a priori* would only be able to explain few possible datasets, then it is likely that we have discovered some of the true underlying structure, and there is reason to believe that our explanation will generalize. The richness or complexity of a model class can be quantified by various types of capacity measure, which essentially capture the richness of datasets a given model class could possibly explain. The analysis of these measures then provides further insight into which quantities it is important that the learning algorithm controls. An example of such a quantity is the *margin*: It can be shown that the performance of linear classifiers in certain spaces depends directly on the width of the margin of separation it induces on the training data.

Over the last decade or so, statistical learning theory has made significant progress in the analysis of supervised learning problems. We have contributed towards this progress by analyzing large-margin algorithms using methods of statistics and functional analysis. In particular, we have been able to show how to generalize these notions from reproducing kernel Hilbert spaces (i.e., the SVM setting) to general metric spaces, leading to a class of distance-based Lipschitz classifiers which contain SVMs and nearest neighbor classifiers as special cases. However, in the past few years we have increasingly shifted our attention to areas of machine learning where statistical learning theory is still in its infancy. This is in particular true for many unsupervised learning techniques such as manifold methods or clustering, but also for newly emerging areas. In our department, we try to contribute to those exciting new challenges where pioneering work can be done, rather than pursue technical refinements for the better-established branches of statistical learning theory.

One focus in our group is to establish *theoretical foundations for clustering*. Even though data clustering techniques have already been employed for many decades in various scientific disciplines, surprisingly little is known about their theoretical aspects. We already started to work on this topic in the previous reporting

period (2003-2005), investigating convergence and consistency properties of spectral clustering. In the current period (2006-2008) we looked at several foundational questions related to clustering.

Continuing our work on clustering consistency, we considered the question how clustering objective functions can be minimized in a statistically consistent way. We found out that in order to achieve consistency, the general requirements are similar to the case of supervised learning: the clusterings have to come from some “small” function class. We developed an algorithm called nearest neighbor clustering, which is a generic scheme to minimize arbitrary clustering objective functions. We managed to prove under which conditions this scheme leads to statistical consistency, and showed that the performance of our algorithm is rather good (see [p. 16]).

A second major question concerns clustering stability. This concept is often used to select the number of clusters: among several choices of the number of clusters, one picks the one where the clustering results tend to be most stable. However, while this method is popular among practitioners and seems to follow a fundamental principle, in fact it is rather unclear what it really does. We published several papers on this topic (one of them won the COLT 2006 best student paper award). All those papers have a slightly negative connotation: clustering stability does not always do what people think it does (see [p. 17] for more details). Those papers initiated a strong debate in the stability community. As a consequence, we organized a workshop *Stability and resampling methods for clustering* in July 2007 in Tübingen, bringing together researchers from either side of the debate about clustering stability. This workshop was very fruitful and led to several insights and conjectures.

Another field of machine learning that has drawn our interest is the one of *graph-based methods*. Many machine learning algorithms use the input data to construct a suitable similarity graph which models local properties of the data. The algorithm then is supposed to discover certain global properties. Applications are manifold methods, semi-supervised learning, or graph-based clustering algorithms. In the previous reporting period we already investigated convergence of the graph Laplace operator on nearest neighbor graphs to the Laplace-Beltrami operator of an underlying manifold. These studies were pursued further [p. 18], and led to a method for manifold denoising: we assume that the data come from some low-dimensional manifold in a high-dimensional ambient space, but they are corrupted by (high-dimensional) noise. The goal is then to discover the original manifold structure (see [p. 18]). A second question under consideration is the one of regression between manifolds. That is, not only the input domain is a curved manifold, but also the output space is no longer Euclidean and a curved space as well. We were able to develop a regularization approach which generalizes various classical methods and takes the manifold structure into account in an appropriate parameterization independent way (see [p. 19]).

Another series of papers deals with cluster identification on neighborhood graphs. Here the question is how we should construct neighborhood graphs on given similarity data in order to obtain the best clustering results (that is, to find the correct clusters with high probability). Using methods from random geometric graphs, we compared different types of graphs (for example,  $k$ -nearest neighbor graphs or epsilon-graphs) and different rules to choose their connectivity parameters ( $k$  or epsilon) in various clustering scenarios (see [p. 20]). We observed different behavior of different graphs, and could establish guidelines for choosing the connectivity parameters. For parts of this work, Markus Maier received the E.M. Gold award at the ALT conference (see [p. 20]).

We are studying the principled design of learning algorithms that are able to identify regularities in data. Subjects of research in this area are thus not only the development of improved algorithms for generic learning problems, but also the design of new algorithms for specific applications.

Technically, many of the approaches in the department fall into the category of *kernel algorithms*. They are based on the notion of positive-definite kernels. These kernels can be shown to play three roles in learning. First, the kernel can be thought of as a (nonlinear) *similarity measure* that is used to compare the data (e.g., visual images). Second, the kernel can be shown to correspond to an inner product in an associated linear space with the mathematical structure of a reproducing kernel Hilbert space (RKHS). In this way, the kernel induces a linear *representation* of the data. Third, it can be shown that a large class of kernel algorithms leads to solutions that can be expanded in terms of kernel functions centered on the training data. In this sense, the kernel also determines the *function class* used for learning, i.e., the hypotheses that are used in examining the dataset for regularities. All three issues lie at the heart of empirical inference, rendering kernel methods an elegant mathematical framework for studying learning and designing learning algorithms.

During their relatively short history in machine learning, kernel methods have already undergone several conceptual changes. Initially, kernels were viewed as a way of “kernelizing” algorithms, i.e., constructing nonlinear variants of existing linear algorithms. The next step was the use of kernels to induce a linear representation of data that did not come from a vector space to begin with, thus allowing the use of a number of linear methods for data types such as strings or graphs. The third change happened only recently. It was observed that kernels sometimes let us rewrite optimization problems over large classes of nonlinear functions as linear problems in RKHSs. In a statistical context, this usually amounts to transforming certain higher order statistics into first-order (linear) ones, and handling them using convenient tools from linear algebra and functional analysis. An example that we have co-developed is a class of methods for *distribution embeddings* in RKHSs [p. 21].

As well as providing a measure of distance on probability distributions (namely, the RKHS distance between embeddings), these mappings directly imply a measure of dependence between random variables, consisting of the RKHS distance between the embedding of the joint distribution and that of the product of marginals.

When the embeddings are computed on the basis of finite samples, a question of particular interest is whether the distance between embeddings is large enough to be statistically significant (and thus, whether the distributions are deemed to be different on the basis of the observations). We have provided means for verifying this significance, and associated non-parametric hypothesis tests for homogeneity, independence, and conditional independence.

The behavior and performance of any kernel algorithm using these distribution embeddings hinges upon properties of the kernel used. This led us to a detailed study of the class of kernels that induce injective RKHS embeddings, i.e., embeddings that do not lose information and uniquely characterize all probability distributions from a given set [p. 22].

Kernel dependence measures based on distribution embeddings may be used not only to detect whether significant dependence exists, but can also be optimized to reveal underlying structure in

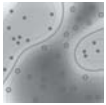
the data [p. 23]. Thus, data can be clustered more effectively when the resulting clusters are given structure using side information, by maximizing a kernel dependence measure with respect to this side information. Such information may take the form of additional descriptions of the data, such as captions for images, or might involve imposing a structure on the clusters using prior knowledge about their mutual relations. In the first case (additional descriptions of the data), we have developed a novel clustering algorithm, Correlational Spectral Clustering, which uses the kernel canonical correlations (closely related to the dependence measures in [p. 21]) between the data and the side information to improve spectral clustering. In the example considered, images were clustered more consistently with human labeling when side information in the associated descriptions was used to guide the clustering. In the second case (prior cluster structure known), the clusters were assumed to follow a tree structure, leading to the Numerical Taxonomy Clustering algorithm.

A second set of projects is concerned with use of non-standard inference principles in machine learning. We have already in the past devoted substantial efforts to such inference principles, including in particular semi-supervised learning. In a series of recent papers, we investigated the use of local inference in several learning problems [p. 24]. We have also contributed to algorithms implementing a novel approach for regularization termed the “Universum,” and linked it up with known methods of learning and data analysis [p. 25] (one of these works received the best paper prize at ICML 2007).

In another focus started during the previous reporting period, we have continued and expanded our work on structured-output learning, dealing with learning algorithms that generalize classification and regression to a situation where the goal is to learn an input-output mapping between arbitrary sets of objects. Canonical examples of an output in this framework are sequences, trees and strings. For such problems, the large size of the output space renders standard learning methods ineffective (e.g., the size of the output space for sequences scales exponentially with the length of the sequences). A series of papers has developed new supervised and semi-supervised learning methods that combine advantages of kernel methods with the efficiency of dynamic programming algorithms [p. 26]. In recent work, we have provided a unifying analysis of existing supervised inference methods for structured prediction using convex duality techniques.

Our analysis has shown that these methods can be cast as duals of various divergence minimization problems with respect to *structured* data constraints. By extending these constraints to employ unlabeled data, we developed a class of semi-supervised learning methods for structured output learning [p. 27]. Another direction we pursued in this framework is extending supervised learning methods to complex tasks, which consist of multiple structured output problems. This is particularly challenging since exact inference of such problems is intractable. We used multi-task learning techniques and devised an efficient approximation algorithm to learn multiple structured output predictors jointly [p. 26].

The two last projects in the area of kernel algorithms draw their motivation from the needs of practical problems that we encountered in application domains. In kernel machines, the solution is usually written as an expansion in terms of an a-priori chosen kernel function. The choice of the kernel function is nontrivial yet important in practice. Sometimes a linear combination from a large library of kernels works best (see [p. 28; p. 59] for methods to compute it automatically), or a multi-scale approach, with aggressive sparsification to keep the runtime complexity under control ([p. 29]). This last work improves upon our earlier work



on sparsification of kernel machines as it turns out that with the additional degree of freedom introduced by the multi-scale approach, a higher degree of sparsification can be achieved.

## Causal and Probabilistic Inference

Uncertainties are present at many levels in biological and artificial adaptive systems. Exemplars from which the system learns are frequently noisy, mislabeled, or atypical. Even with data of high quality, gauging and combining a multitude of data sources and constraints in usually imperfect models of the world requires us to represent and process uncertain knowledge in order to take viable decisions. In the increasingly complicated settings of modern science, model structure or causal relationships may not be known *a-priori*. The probabilistic framework for working with uncertain information is at the heart of the Causal and Probabilistic Inference group. In the Bayesian setting, probabilities are used to represent beliefs about variables of interest, and their relationships. The calculus is based on simple rules from probability theory, which are used to refine or update beliefs in the light of new data. The theory of Bayesian inference is well developed, and its fundamental role for decision theory has long been established. A central difference between the Bayesian and other mainstream traditions in statistics lies in the treatment of unobserved variables: these are summed over all possible instances in the former, while often estimated or otherwise conveniently imputed in many of the latter. While justified, the Bayesian insistence on summations leads to two major problems, which need to be addressed. First, the prior distributions, or equivalently the weighting of instances in the summations need to be specified or assessed. Second, the summations are typically computationally hard due to nonlinear dependencies between many variables, and approximations are typically unavoidable. These problems are usually related, in that some classes of priors may lead to more tractable computations than others, and modeling versus approximation errors need to be traded off against each other. Large-scale Bayesian applications in machine learning or statistics are conceptually guided by, but often diverge considerably from, the pure theory. This leaves much room and high demand for novel concepts, algorithms, and theoretical as well as empirical analyses, in what is sometimes misleadingly presented in textbooks as a closed and solved framework.

The focus of our research directly addresses these questions, both at the level of algorithms for machine learning, and of individual applications of inference. We develop new algorithms and approximation techniques, assess their quality both in theory and in empirical evaluations, and apply them to challenging problems. Below we exemplify some of these developments.

*Gaussian process* (GP) models are probabilistic kernel-based learning algorithms. Through parameterization of the kernel (or covariance function) a flexible family of models can be treated, and inference over kernels can be approximated. We are addressing several issues of fundamental importance for GPs, such as the use of sparse approximations and the evaluation of approximation techniques for classification models [p. 30; p. 29], as well as theoretical aspects including generalization error bounds or online sequence prediction [p. 30]. We apply GP models to problems in reinforcement learning and control [p. 48], and to real-time inverse kinematics in autonomous robots [p. 45].

When observations are measured as time series, for example neuronal recordings, or the monitoring of a machine or factory, it is important to capture the temporal dynamics. This allows us to group time series according to similar dynamics, for instance to deduce common functions in biology, or to segment a time

course into regimes of distinct dynamical behavior, for example to detect anomalies. We have developed Bayesian approaches for time-series clustering and segmentation, using deterministic variational approximations [p. 32], which work without supervision and benefit from classical approaches in terms of accuracy and stability. We applied these models to spike sorting of neuronal recordings, and to clustering and segmentation of human-movement trajectories for robot imitation learning.

Bayesian methods are especially valuable for higher order problems, such as the optimization of hyperparameters in hierarchical models (an example is the optimization of the kernel in a GP model [p. 30]). In this context, it is possible to support even non-Bayesian methods (which may be preferable due to faster running time) by higher order Bayesian learning, if the only alternative lies in time-consuming cross-validation procedures.

In *experimental design*, the goal is to optimize data sampling or measurement architectures, such that inference or estimation can be realized faster or at lower costs. We develop Bayesian experimental design and efficient approximate inference algorithms for GP and sparse generalized linear models [p. 31], and address applications in systems biology, natural image reconstruction, and magnetic resonance imaging [p. 55]. The expertise spanned by our group allows us to combine Bayesian techniques, such as variational inference, message passing, and MCMC with ideas coming from adjacent fields, such as sparse estimation and convex optimization.

While Bayesian practitioners can rely on robust and consistent inference principles, they need to master a large and rapidly growing number of approximations and computational techniques, and be prepared to find a good trade-off specifically for their problem at hand. Many of these techniques and concepts come from or have been developed in the context of classical statistics. While an unfortunate divide still exists between many Bayesian and classical statisticians, we try to build on both these traditions to get closer to the common goal of efficient, reliable Bayesian machine learning.

Being probabilistic methods, Bayesian approaches make statements about distributions of parameters, variables, or predictions. Such statements can also concern dependencies between observed quantities (e.g., about the correlation between the size of the population of storks and the human birth rate). A convenient way of visualizing the dependency structure implied by a joint distribution is a graphical model, or Bayesian network. The joint density then factorizes into conditional densities of each variable given its parents for every graph that renders the joint distribution Markovian. One should, however, not mistake the links in such networks for *causal links* indicating that a parent variable exerts a causal influence on its child nodes. To infer the *causal graph* that visualizes the causal structure of the data generating process is a highly non-trivial task requiring additional assumptions. Indeed, there used to be a substantial controversy in the statistics community as to whether this is possible at all when using data only from passive observations.

A small but growing community studying causal inference, however, develops methods to construct causal graphs from such data. These are directed acyclic graphs whose links represent causal relationships, and for which the joint distribution is Markovian and faithful, meaning that the observed pattern of conditional independences is generic in the sense that it is common to almost all choices of the conditional densities assigned to each node. State-of-the-art causal inference algorithms work with these assumptions but share the following drawbacks: first, conditional statistical independence tests often rely on strong assumptions like multivariate Gaussianity and

linearity of causal links and, second, the Markov and the faithfulness condition are often consistent with a large set of causal graphs (“Markov equivalent graphs”) rather than selecting a unique structure.

To address the first problem we have modified the algorithms by using kernel independence tests [p. 33] which indeed improved the performance. To tackle the second problem, i.e., to distinguish between Markov equivalent causal graphs, additional assumptions are needed. Implicit or explicit priors that favor “simple” conditionals can render one of several competing causal hypotheses more likely. We have explored novel approaches postulating additional statistical asymmetries between cause and effect that we sketch now. Apart from an entropy-based criterion for formally selecting hypotheses that yield simple conditionals, we have defined complexity of conditional probability densities by semi-norms in reproducing kernel Hilbert spaces and developed causal inference algorithms based upon this definition.

Another method, which uses priors on conditionals in a more implicit way, has recently been proposed in the literature. It exploits the fact that linear structural equations between non-Gaussian variables induce joint distributions that do not admit any causal graph other than the original one, which is then assumed to be the true causal structure. We have extended this approach (called “LINGAM”) to deal with time series, where the ground truth is known because the time order excludes causal links influencing the past. Our work confirmed that empirical time series are significantly more likely to admit a linear autoregressive moving average model in forward than in backward direction [p. 34].

We have also transferred the LINGAM idea to *non-linear structural equation models*, showing that functional relations with additive noise between two variables induce, in the generic case, joint distributions that do not admit an additive noise model in backward direction [p. 35]. Although non-linearity makes the analysis significantly harder, this setting has the advantage that it can also work when the noise is Gaussian, a case where the linear approach fails. To test whether such an additive noise model exists, one checks whether the predictor variable and the residual errors of a non-linear regression are statistically independent. Kernel independence tests (see [p. 21]) provided a helpful tool for this task.

To justify causal inference rules of the above type, we have developed graphical models that represent algorithmic dependences among single objects instead of statistical dependences among random variables [p. 33]. We have proved that three versions of the Markov condition that are known to coincide for statistical dependences essentially also coincide for algorithmic dependences. Based on these results, we have postulated an algorithmic Markov condition relating the observed algorithmic dependences to causality and shown that it can be justified in strong analogy to the statistical causal Markov condition. This adds another level to causal inference because a graph may explain the statistical but not the algorithmic dependences. The algorithmic Markov condition imposes, for instance, that the shortest description of the joint density  $P$  (cause, effect) is given by separate descriptions of  $P$  (cause) and  $P$  (effect | cause). Causal graphs for which the conditional densities assigned to the different nodes share algorithmic information must be rejected as “non-generic” for the same reason as one does not accept unfaithful distributions (because they would require specific adjustments of the network parameters).

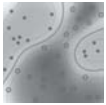
The analysis and synthesis of images is a particularly challenging problem for automatic systems. It typically involves large amounts of high dimensional data with significant between-feature correlations that are corrupted by non-uniform sensor noise. The fact that, nevertheless, humans and animals have developed very efficient visual systems makes the analysis and synthesis of images an important field of research that both inspires new machine learning research and allows testing existing techniques for their practical applicability. Our own research in this area can be subdivided into three main directions: (classical) computer vision, image processing and – more recently – computational photography. The boundaries between these areas are not sharply defined, as they all deal with processing image data and often rely on similar techniques.

As *computer vision* we consider the task of extracting high-level information from images, e.g. the presence of objects or the classification of events. In the time since we started our own work in this area in 2002, computer vision has become a showcase example of how machine learning can all but take over a field of research previously dominated by hand crafted techniques. Today, the use of machine learning methods has become common practice for many computer vision researchers. Our own research in this field therefore follows a dual agenda: on the one hand, we develop new methods for computer vision problems based on recent machine learning techniques. In particular this includes the use of structured input and output spaces, a field that has not yet become mainstream in computer vision research, and thus enables us to establish new grounds ourselves rather than following existing trends. On the other hand, we focus on finding methods that are not limited to the solution of specific problems, but that will be applicable to other areas as well.

Our work on learning structured features is an example of this dual strategy: we have generalized frequent item set mining and graph mining, two retrieval techniques that are currently successful in data mining research. By defining the quantity of interest for a pattern (an item set or graph) to measure how discriminative it is for a prediction task, instead of its frequency in the data corpus, we obtained *discriminative item set mining and graph mining*. We demonstrated that applying these techniques to regions of natural images yields powerful structured features that can be used for the detection of objects in images and for the classification of actions in videos [p. 36]. At the same time, the methodology is applicable to other domains where structured information is processed, e.g., bioinformatics.

An important focus in our recent research has been the question how automatic systems can learn to *localize objects in images*. We developed a method for this that replaces the usual and sub-optimal two-step procedure by a joint formulation of structured output regression that allows consistent end-to-end training [p. 37]. Part of this work won the best paper award at CVPR 2008. Subsequently, we developed a method that allows the use of image and object context when training such detection systems for multiple classes. It determines relevant dependencies between predictors for different classes automatically using multiple kernel learning. This work was awarded the Main Prize at the 2008 Symposium of the German Association for Pattern Recognition (DAGM).

Further projects that originated in computer-vision-related problems are in *clustering and taxonomy discovery* [p. 23], *learning of optimal kernel combinations* [p. 28] and *learning of image interest operators from eye movements* [p. 53]. These are discussed in the corresponding sections on *Kernel Algorithms* and *Machine Learning in Neuroscience*.



Taking a medium-term perspective, we expect that machine learning may have as strong an influence on other image-related domains as it had on computer vision. We try to make this happen by devoting an increasing amount of our attention to *computer graphics*, *image processing*, and the newly established field of *computational photography*. Central themes of our research in these areas are the exploitation of image statistics, often in conjunction with their use for image reconstruction.

*Steganalysis* has the goal to detect hidden signals that have been embedded invisibly into image data. Most previous approaches in this area relied on explicit models of the suspected steganographic method, which severely limits their applicability. In contrast, we developed a model-free method that relies only on relatively universal image statistics, characterized by how well image pixels can be predicted from their neighbors. This helps detect unknown forms of steganography [p. 38].

Medical imaging applications, such as positron emission tomography (PET), are of special interest to our research, as images cannot be taken directly in this field, but they have to be reconstructed numerically from noisy sensor measurements. This can benefit from the use of statistical image models, as they should allow inferring higher quality images from fewer measurements, thereby saving time and cost. In this line of research, we have recently developed a new non-monotonic method for maximum-likelihood PET image reconstruction that improves reconstruction speed and quality over previous approaches [Bülthoff report, p. 33].

In computer graphics we have also continued to study reconstruction problems, where the task is to infer a three-dimensional object shape from two-dimensional views. This again is an area where the integration of statistical knowledge about shape and surface smoothness can improve an algorithms' performance. Specifically, we developed a method for *3D face reconstruction* that creates a face model of a specific person from a monocular video, or even from a single image, without user interaction [p. 40]. Progress was also made in the area of mesh tracking, where we developed a fast technique that allows realistic animation of time-varying, flexible objects [p. 29].

*Computational photography* is an area of research that aims at enhancing photographic imaging processes beyond the capabilities of traditional film-based cameras. It combines many of the aspects of computer vision, image processing and computer graphics. As an entry into this field we have conducted work on *color constancy*, the problem of making photos of natural scenes look consistent independent of the lighting conditions in which the photo was taken. A Bayesian approach for automatic white balancing improves the visual impression of scenes by means of a prior function that is learned empirically from a reference dataset [p. 41].

Another focus of our research has been *cross-modal image prediction*, i.e., the prediction of images in one modality from another. An important application area for this is PET-MR imaging, where attenuation correction of PET scans requires the estimation of a synthetic X-ray image (see also [p. 54]). Using statistical techniques and the recent structured-output support vector machine framework, we developed methods for such cross-modal prediction tasks. As a test application, we considered the task to predict color from gray scale images [p. 42].

While many of the above projects exploit structure in the joint distribution of pixel values arising from corresponding structure in the objects being imaged, one can also exploit dependences in the joint distribution of sensor noise. A common problem in long exposure or high ISO photography is thermal noise accumulated during exposures. This noise is non-stationary, e.g., due to changes in the ambient temperature; however, it is highly structured

and it can significantly be reduced by a method that uses a sample of a camera's thermal and readout noise distribution combined with an image prior to generated plausible low noise images [p. 43].

We have also recently initiated work in the area of *blind image deconvolution*. In particular, we have developed a method that removes image blur caused by an unknown point spread function if multiple exposures of the same object or scene are available. This has important applications, e.g., in high-resolution astronomical imaging [p. 44].

## Robot Learning

Creating autonomous robots that can learn to assist humans in situations of daily life is a fascinating challenge for machine learning. While this aim has been a long-standing vision of artificial intelligence and the cognitive sciences, we have yet to achieve the first step of creating robots that can learn to accomplish many different tasks triggered by environmental context or higher-level instruction. The goal of our robot learning laboratory is the investigation of the ingredients for such a general approach to motor skill learning, to get closer towards human-like performance in robotics. We thus focus on the solution of basic problems in robotics while developing domain-appropriate machine-learning methods.

Starting from theoretically well-founded approaches to representing the required control structures for task representation and execution, we replace the analytically derived modules by more flexible, learned ones.

An essential problem in robotics is the accurate execution of desired movements using only low-gain controls such that the robot will accomplish the desired task while not harming human beings in its environment. Following a trajectory with little feedback requires the accurate prediction of the needed torques, which cannot be achieved using classical methods for sufficiently complex robots. However, learning such models is hard as the joint-space can never be fully explored and the learning algorithm has to cope with a never-ending data stream in real time. We have developed learning methods both for accomplishing tasks represented in operational space [p. 45] as well as in joint-space [p. 46].

While learning to execute tasks is a component essential to a framework for motor skill learning, learning the actual task is of even higher importance as discussed in [p. 47]. Here, we focus on the learning of elementary tasks or movement primitives, which are parameterized task representations based on nonlinear differential equations with desired attractor properties. We mimic how children learn new motor tasks using imitation learning for initializing these movement primitives while employing reinforcement learning to subsequently improve the task performance. We have learned tasks such as Ball-in-a-Cup or bouncing a ball on a string using this approach.

Efficient reinforcement learning for continuous states and actions is essential for robotics and control. We follow two approaches depending on the dimensionality of the domain. For high-dimensional state and action spaces, it is often easier to directly learn policies without estimating accurate system models. The resulting algorithms are parametric policy search algorithms inspired by expectation-maximization methods and can be employed for motor primitive learning [p. 47]. For lower-dimensional systems, Bayesian approaches to control can be shown to be able to cope with the optimization bias introduced by the model errors in model-based reinforcement learning. As a result, these methods can learn good policies at a rapid pace based on only little interaction of the system [p. 48].

Currently, we are moving towards learning complex tasks, requiring the solution of a variety of hard problems. Among these are the decomposition of large tasks into movement primitives (MP), the acquisition and self-improvement of MPs, the determination of the number of MPs in a data set (see [p. 32] for some early steps in this direction), the determination of the relevant task-space, perceptual context estimation and goal learning for MPs, as well as the composition of MPs for new complex tasks. These questions are tackled in order to make progress towards fast and general motor skill learning for robotics.

## Machine Learning in Neuroscience

The neurosciences present some of the steepest challenges to machine learning. Among diverse problem settings and approaches, certain commonalities can be identified. Nearly always there is a very high-dimensional input structure—particularly relative to the number of exemplars, since each data point is usually gathered at a high cost in time and money. To avoid overfitted solutions, inference must therefore make considerable use of domain knowledge from physics, neurophysiology and anatomy. Solutions typically occupy a relatively small subspace of the input representation, the rest being made up of noise that may be of much larger magnitude (often composed largely of the manifestations of other neurophysiological processes, besides the ones of interest). In finding generalizable solutions, one usually has to contend with a high degree of variability, both between individuals and across time, leading to problems of covariate shift and non-stationarity. In all cases, even the high-dimensional raw input is a vastly simplified reflection of the underlying processes and structures. New ways of measuring relevant information, and new ways of transforming the data, are therefore still waiting to be found, leading to feature representations that are more relevant, less noisy, or more transferrable between experimental sessions and subjects.

One specific neuroscientific application area in which we have an active interest is that of brain-computer interfacing, or BCI (see [p. 49]). This research aims to construct systems that could allow a paralyzed person to communicate, by decoding the user's intentions from measured brain signals. Currently, BCI systems based on electroencephalogram (EEG) signals do allow communication, and can be used even by people with very little remaining motor control, but they are still slow and difficult enough to use that they are not an attractive alternative to other communication methods. Among people who have absolutely no voluntary motor control, and who therefore stand to benefit most from BCI since they have no alternative, a convincing demonstration of successful communication has yet to be published. We have been working hands-on in collaboration with two Tübingen University departments, the Institute of Medical Psychology and Behavioral Neurobiology, and the Department of Neurosurgery at the University Clinics, in order to develop and test BCI systems with paralyzed patients. In addition, we pursue several laboratory-based lines of research. The contribution of machine learning to BCI is in developing, refining and applying algorithms to improve the accuracy with which neural signals are decoded, to interpret users' communication intentions more reliably and to reduce training times for the user. Methods with good generalization properties may especially improve performance in the hardest cases, where data sets are small or particularly noisy, as is typically the case with patient data. We regard algorithmic development ([p. 49], [p. 51]) and improvements in experimental methodology (for example, as described in [p. 50]) as significant contributions towards making clinical BCI systems a reality.

In contrast to the engineering approach of BCI, another role that machine learning can play in neuroscientific research is that of analysis-by-synthesis. In [p. 52] and [p. 53], we see perceptual problems that neural systems solve in the natural world being resolved by simplified machine-learning systems. This provides insight into the mechanisms of perception, and allows computational theories of these mechanisms to be advanced.

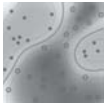
System identification methods also play an important role for computational neuroscience. In order to get better ideas about coding and computation in the early visual systems, we are collaborating with the Bethge Group in developing a model for the prediction of multi-cell responses capable of inferring excitatory and inhibitory interactions between neurons [Bethge report, p. 18]. The model is convex and parameters can be fitted efficiently. We obtain an estimate of the uncertainty of the parameters by using approximate Bayesian inference methods.

We have further applied machine learning techniques as interpretive statistical tools in neural data analysis, in collaboration with the Neurophysiology Department. In short, a “learnable” classification or regression problem can often reveal a property of the underlying system being studied. In one example, we have developed a classification protocol for high field (7T) fMRI BOLD responses to various object categories, to reveal the locations in the brain where patterns of activity allow discrimination of these classes [Logothetis report, p. 44]. In this case, by employing multivariate analysis techniques (including SVMs), we were able to recognize distributed activity patterns across multiple voxels, rather than considering the responses of individual voxels in isolation, as done in classical general linear model studies.

In a second study with the Neurophysiology department [Logothetis report, p. 75], we established the relation between local field potentials (LFPs; relatively slow fluctuations in the electric field measured at the electrode, thought to reflect the input of a given cortical area as well as its local intracortical processing) and spikes from electrode recordings in the primary visual cortex, during presentation of naturalistic movie stimuli. We used a support vector machine to predict spiking behavior on the basis of LFP features. This revealed both the LFP features that were most predictive of spikes (the amplitude and phase of the LFP at frequencies lower than 10Hz, and power fluctuations in the Gamma (40-90Hz) range), and the properties of spike trains that were predictable (low frequency structure in the 100ms range). We subsequently employed information-theoretic techniques to determine the stimulus information jointly encoded by LFPs and spikes, and the redundancy in information encoded at different LFP frequencies [Logothetis report, p. 74].

In a third study [Logothetis report, p. 93], we investigated the modulatory effects of pharmacological agents on the brain through simultaneous electrode recordings and fMRI measurements, where the dependencies between the electrophysiological and fMRI signals were revealed through kernel canonical correlation analysis (KCCA; a close relation to the kernel dependence measures described elsewhere [p. 21]). KCCA is of particular interest since the results are interpretable: it reveals the features of both signals that are most closely related. This addresses the important open question of how fMRI relates to neurophysiological activity in the presence of externally administered neuro-modulatory compounds, and is thus relevant to the field of Pharmacological Magnetic Resonance Imaging (phMRI).

Imaging techniques play a central role in modern neuroscience, and we thus also include two projects in this overview that do not address neuroscientific issues directly, but instead focus on methodological improvements of imaging techniques. The first



one contributes towards the development of combined PET/MR scanners, by providing a method to predict an attenuation correction map for the PET signal based on the MR information [p. 54]. In the second project, machine learning could help in the neurosciences by optimizing the way in which questions are asked. On page [p. 55] we see an automatic experimental design approach to making the acquisition of MRI images more efficient, which in turn has the potential to improve the temporal resolution of functional MRI.

## Bioinformatics

Genomics and proteomics are currently producing massive datasets containing a wealth of information about underlying biological mechanisms. However, many interesting properties of biomolecules, such as drugs, DNA and proteins, cannot be easily determined experimentally and thus form worthwhile targets for computational prediction. In our bioinformatics group, we are interested in developing new machine learning methods to find regularities or interpretable knowledge in biological data and, in turn, exploit the knowledge to predict important properties of biomolecules.

Biological data, drawn from the observation of the complex mechanisms of the cell, are represented in various types of data structures. For example, gene expression profiles can be represented as vectors, genome sequences as symbol sequences, phylogenetic profiles as trees, chemical compounds as graphs and relationships among proteins as networks. One of our main goals is to develop effective and theoretically well-founded machine learning methods for such *structured data*. Current projects employ diverse approaches such as graph mining, network module enumeration, structured output learning and Bayesian inference for dealing with structured data.

*Graph mining* is a recently emerging approach to finding frequently appearing subgraphs in large databases of graphs. It has been used, for example, for finding abundant phrases in a text corpus, but has a large potential for computational biology as well. In particular, graph mining has been proven effective in cheminformatics, where the task is to predict biochemical properties such as toxicity, mutagenicity, or solubility of drug candidates. Such properties can be screened by experimental means, however, given a large amount of drug candidates, computational prediction is useful for the drug discovery processes by providing a form of virtual screening.

We have been working on developing efficient and accurate algorithms for learning from graph data. Our goal is to derive prediction rules that depend on only a few, interpretable subgraph features. Among several approaches we have proposed so far, the flagship method is *graph boosting*, where graph mining is repeatedly called to collect necessary subgraph descriptors progressively [p. 56]. In comparison with conventional methods that create all subgraph descriptors at once, graph boosting is faster and requires less memory by avoiding the enumeration of useless descriptors. So far, graph mining methods have been mostly heuristic and lacking solid theoretical basis. In contrast, graph boosting has guaranteed convergence properties, and the maximum-margin property to warrant good generalization. The impact of this work was partly evidenced by the best paper award at the MLG (mining and learning with graphs) workshop in 2006. Along this line of research, we also published graph least angle regression (gLARS), graph partial least squares regression (gPLS) and graph principal component analysis (gPCA) at leading machine learning and data mining conferences. Recently, we developed a graph boosting method that can predict the chemical activity with error bars [p. 58]. This

Bayesian approach allows us to evaluate the uncertainty of predictions, which is often important in applications. This can eventually lead to novel methods for active learning and automated drug discovery algorithms.

As biological networks such as protein interaction networks and gene regulatory networks have recently become available in a large number of species, bioinformatics groups worldwide are working on novel methods for the discovery of useful knowledge from the networks. In particular, discovery of densely connected protein groups (i.e., modules) is crucial, as biological functions tend to be achieved not by individual proteins, but by groups or complexes of proteins. Most methods partition a network into disjoint modules; however, this is not appropriate for discovering protein complexes, because one protein often participates in multiple complexes. We have proposed a dense module enumeration method that can find all dense modules [p. 57]. It is based on a technique called reverse search, which was developed in the 90s and can solve difficult problems where conventional algorithms such as branch-and-bound become inefficient. This work received the best poster award at the student council symposium of ISMB 2007.

In addition to novel data mining techniques, we continue to apply kernel methods to problems of computational biology. In collaboration with the Friedrich Miescher Lab (FML) group led by Gunnar Rätsch, we applied SVMs to the prediction of *subcellular localization of proteins* from their amino acid sequences [p. 59]. The SVM was extended to choose automatically an optimal combination from a large set of possibly useful kernels; this way, competitive accuracy was achieved without tedious handcrafting of models. This work led to a best student paper prize at ISMB 2008. A second project done with the FML group concerns domain adaptation methods in the field of genome annotation. For model organisms such as *C. elegans*, a substantial amount of labeled data is available, but there is only little data available for newly sequenced organisms. In a recent project, we evaluated a number of domain adaptation methods (including some newly developed ones) to transfer knowledge from one organism to another [p. 60], showing that significant improvements can be achieved. It is planned to include such methods into the genome annotation system developed at the FML over the last years with significant involvement of members of our department [Rätsch report, p. 12/13].

In another campus collaboration we have continued a joint project with the Department of Protein Evolution at the Max Planck Institute for Developmental Biology, working on the development of Bayesian three-dimensional protein structure elucidation methods based on *Inferential Structure Determination (ISD)* [p. 61]. In contrast to optimization-based methods, ISD can take the uncertainty into account that is inherent in NMR data. It allows sampling the structures from the full posterior distribution with Markov chain Monte Carlo methods, and assessing the precision and quality of the obtained structures. Recently, our main focus has been the integration of heterogeneous data sources such as NMR, X-ray crystallography, and cryo-electron microscopy (cryo-EM). The ISD method and software were extended to integrate low-resolution density maps from cryo-EM and structure prediction by homology. These developments contributed to the determination of the first structure of a mitochondrial membrane protein (VDAC).