Temporal Coherence, Natural Image Sequences, and the Visual Cortex
Jarmo Hurri, Aapo Hyvärinen
We show that two important properties of the primary visual cortex emerge when t
he principle of temporal coherence is applied to natural image sequences. The pr
operties are simple-cell-like receptive ■elds and complex-cell-like pooling of s
imple cell outputs, which emerge when we apply two different approaches to tempo
ral coherence. In the ■rst approach we extract receptive ■elds whose outputs are
 as temporally co- herent as possible. This approach yields simple-cell-like rec
eptive ■elds (oriented, localized, multiscale). Thus, temporal coherence is an a
lterna- tive to sparse coding in modeling the emergence of simple cell receptive
 ■elds. The second approach is based on a two-layer statistical generative model
 of natural image sequences. In addition to modeling the temporal coherence of i
ndividual simple cells, this model includes inter-cell tem- poral dependencies.
Estimation of this model from natural data yields both simple-cell-like receptiv
e ■elds, and complex-cell-like pooling of simple cell outputs. In this completel
y unsupervised learning, both lay- ers of the generative model are estimated sim
ultaneously from scratch. This is a signi■cant improvement on earlier statistica
l models of early vision, where only one layer has been learned, and others have
 been ■xed a priori.
****
Nonparametric Representation of Policies and Value Functions: A Trajectory-Based
 Approach
Christopher Atkeson, Jun Morimoto
A longstanding goal of reinforcement learning is to develop non- parametric repr
esentations of policies and value functions that support rapid learning without
suffering from interference or the curse of di- mensionality. We have developed
a trajectory-based approach, in which policies and value functions are represent
ed nonparametrically along tra- jectories. These trajectories, policies, and val
ue functions are updated as the value function becomes more accurate or as a mod
el of the task is up- dated. We have applied this approach to periodic tasks suc
h as hopping and walking, which required handling discount factors and discontin
u- ities in the task dynamics, and using function approximation to represent val
ue functions at discontinuities. We also describe extensions of the ap- proach t
o make the policies more robust to modeling error and sensor noise.
****
A Differential Semantics for Jointree Algorithms
James Park, Adnan Darwiche
A new approach to inference in belief networks has been recently proposed, which
 is based on an algebraic representation of belief networks using multi{linear f
unctions. According to this approach, the key computational question is that of
representing multi{linear functions compactly, since inference reduces to a simp
le process of ev aluating and di■erentiating such functions. W e show here that
mainstream inference algorithms based on jointrees are a special case of this ap
proach in a v ery precise sense. W e use this result to prov e new properties of
 jointree algorithms, and then discuss some of its practical and theoretical imp
lications.
****
"Name That Song!" A Probabilistic Approach to Querying on Music and Text
Brochu Eric, Nando de Freitas
We present a novel, ■exible statistical approach for modelling music and text jo
intly. The approach is based on multi-modal mixture models and maximum a posteri
ori estimation using EM. The learned models can be used to browse databases with
 documents containing music and text, to search for music using queries consisti
ng of music and text (lyrics and other contextual information), to annotate text
 documents with music, and to automatically recommend or identify similar songs.
****
Automatic Derivation of Statistical Algorithms: The EM Family and Beyond
Bernd Fischer, Johann Schumann, Wray Buntine, Alexander Gray
Machine learning has reached a point where many probabilistic meth- ods can be u
nderstood as variations, extensions and combinations of a much smaller set of ab

stract themes, e.g., as different instances of the EM algorithm. This enables the systematic derivation of algorithms cus- tomized for different models. Here, we describe the AUTO BAYES sys- tem which takes a high-level statistical model speci■cation, uses power- ful symbolic techniques based on schema-based program synthesis and computer algebra to derive an ef■cient specialized algorithm for learning that model, and generates executable code implementing that algorithm. This capability is far beyond that of code collections such as Matlab tool- boxes or even tools for model-independent optimization such as BUGS for Gibbs sampling: complex new algorithms can be generated with- out new programming, algorithms can be highly specialized and tightly crafted for the exact structure of the model and data, and ef■cient and commented code can be generated for different languages or systems. We present automatically-derived algorithms ranging from closed-form solutions of Bayesian textbook problems to recently-proposed EM algo- rithms for clustering, regression, and a multinomial form of PCA.
****

Going Metric: Denoising Pairwise Data
Volker Roth, Julian Laub, Klaus-Robert Müller, Joachim Buhmann
Pairwise data in empirical sciences typically violate metricity, ei(cid:173) ther due to noise or due to fallible estimates, and therefore are hard to analyze by conventional machine learning technology. In this paper we therefore study ways to work around this problem. First, we present an alternative embedding to multi-dimensional scaling (MDS) that allows us to apply a variety of classical ma(cid:173) chine learning and signal processing algorithms. The class of pair(cid:173) wise grouping algorithms which share the shift-invariance property is statistically invariant under this embedding procedure, leading to identical assignments of objects to clusters. Based on this new vectorial representation, denoising methods are applied in a sec(cid:173) ond step. Both steps provide a theoretically well controlled setup to translate from pairwise data to the respective denoised met(cid:173) ric representation. We demonstrate the practical usefulness of our theoretical reasoning by discovering structure in protein sequence data bases, visibly improving performance upon existing automatic methods.
****

String Kernels, Fisher Kernels and Finite State Automata
Craig Saunders, Alexei Vinokourov, John Shawe-taylor
In this paper we show how the generation of documents can be thought of as a k-stage Markov process, which leads to a Fisher ker(cid:173) nel from which the n-gram and string kernels can be re-constructed. The Fisher kernel view gives a more flexible insight into the string kernel and suggests how it can be parametrised in a way that re(cid:173) flects the statistics of the training corpus. Furthermore, the prob(cid:173) abilistic modelling approach suggests extending the Markov pro(cid:173) cess to consider sub-sequences of varying length, rather than the standard fixed-length approach used in the string kernel. We give a procedure for determining which sub-sequences are informative features and hence generate a Finite State Machine model, which can again be used to obtain a Fisher kernel. By adjusting the parametrisation we can also influence the weighting received by the features . In this way we are able to obtain a logarithmic weighting in a Fisher kernel. Finally, experiments are reported comparing the different kernels using the standard Bag of Words kernel as a baseline.
****

Extracting Relevant Structures with Side Information
Gal Chechik, Naftali Tishby
The problem of extracting the relevant aspects of data, in face of multiple con■icting structures, is inherent to modeling of complex data. Extract- ing structure in one random variable that is relevant for another variable has been principally addressed recently via the information bottleneck method [15]. However, such auxiliary variables often contain more in- formation than is actually required

due to structures that are irrelevant for the task. In many other cases it is i
n fact easier to specify what is irrelevant than what is, for the task at hand.
Identifying the relevant structures, however, can thus be considerably improved
by also mini- mizing the information about another, irrelevant, variable. In thi
s paper we give a general formulation of this problem and derive its formal, as
well as algorithmic, solution. Its operation is demonstrated in a synthetic exam
ple and in two real world problems in the context of text categoriza- tion and f
ace images. While the original information bottleneck problem is related to rate
 distortion theory, with the distortion measure replaced by the relevant informa
tion, extracting relevant features while removing irrelevant ones is related to
rate distortion with side information.
****
Classifying Patterns of Visual Motion - a Neuromorphic Approach
Jakob Heinzle, Alan A. Stocker
We report a system that classi■es and can learn to classify patterns of visual m
otion on-line. The complete system is described by the dynam- ics of its physica
l network architectures. The combination of the fol- lowing properties makes the
 system novel: Firstly, the front-end of the system consists of an aVLSI optical
 ■ow chip that collectively computes 2-D global visual motion in real-time [1].
Secondly, the complexity of the classi■cation task is signi■cantly reduced by ma
pping the continu- ous motion trajectories to sequences of ’motion events’. And
thirdly, all the network structures are simple and with the exception of the opt
ical ■ow chip based on a Winner-Take-All (WTA) architecture. We demon- strate th
e application of the proposed generic system for a contactless man-machine inter
face that allows to write letters by visual motion. Re- garding the low complexi
ty of the system, its robustness and the already existing front-end, a complete
aVLSI system-on-chip implementation is realistic, allowing various applications
in mobile electronic devices.
****
How the Poverty of the Stimulus Solves the Poverty of the Stimulus
Willem Zuidema
Language acquisition is a special kind of learning problem because the outcom
e of learning of one generation is the input for the next. That makes it possi
ble for languages to adapt to the particularities of the learner. In this
 paper, I show that this type of language change has important consequen
ces for models of the evolution and acquisition of syntax.
****
Fast Transformation-Invariant Factor Analysis
Anitha Kannan, Nebojsa Jojic, Brendan Frey
Dimensionality reduction techniques such as principal component analy- sis and f
actor analysis are used to discover a linear mapping between high dimensional da
ta samples and points in a lower dimensional subspace. In [6], Jojic and Frey in
troduced mixture of transformation-invariant component analyzers (MTCA) that can
 account for global transforma- tions such as translations and rotations, perfor
m clustering and learn lo- cal appearance deformations by dimensionality reducti
on. However, due to enormous computational requirements of the EM algorithm for
learn- ing the model, O(is the dimensionality of a data sample, MTCA was not pr
actical for most applications. In this paper, we demon- strate how fast Fourier
transforms can reduce the computation to the or- . With this speedup, we show th
e effectiveness of MTCA der of in various applications - tracking, video texture
s, clustering video se- quences, object recognition, and object detection in ima
ges.
****
Spikernels: Embedding Spiking Neurons in Inner-Product Spaces
Lavi Shpigelman, Yoram Singer, Rony Paz, Eilon Vaadia
Inner-product operators, often referred to as kernels in statistical learning, d
e- ■ne a mapping from some input space into a feature space. The focus of this p
aper is the construction of biologically-motivated kernels for cortical ac- tivi
ties. The kernels we derive, termed Spikernels, map spike count sequences into a
n abstract vector space in which we can perform various prediction tasks. We dis

cuss in detail the derivation of Spikernels and describe an ef■cient al- gorithm for computing their value on any two sequences of neural population spike counts. We demonstrate the merits of our modeling approach using the Spikernel and various standard kernels for the task of predicting hand move- ment velocities from cortical recordings. In all of our experiments all the ker- nels we tested outperform the standard scalar product used in regression with the Spikernel consistently achieving the best performance.
****

Graph-Driven Feature Extraction From Microarray Data Using Diffusion Kernels and Kernel CCA

Jean-philippe Vert, Minoru Kanehisa

We present an algorithm to extract features from high-dimensional gene expression pro■les, based on the knowledge of a graph which links to- gether genes known to participate to successive reactions in metabolic pathways. Motivated by the intuition that biologically relevant features are likely to exhibit smoothness with respect to the graph topology, the algorithm involves encoding the graph and the set of expression pro- ■les into kernel functions, and performing a generalized form of canoni- cal correlation analysis in the corresponding reproducible kernel Hilbert spaces. Function prediction experiments for the genes of the yeast S. Cerevisiae validate this approach by showing a consistent increase in performance when a state-of-the-art classi■er uses the vector of features instead of the original expression pro■le to predict the functional class of a gene.
****

Spike Timing-Dependent Plasticity in the Address Domain

R. Vogelstein, Francesco Tenore, Ralf Philipp, Miriam Adlerstein, David Goldberg, Gert Cauwenberghs

Address-event representation (AER), originally proposed as a means to communicate sparse neural events between neuromorphic chips, has proven ef■cient in implementing large-scale networks with arbitrary, con■gurable synaptic connectivity. In this work, we further extend the functionality of AER to implement arbitrary, con■gurable synaptic plas- ticity in the address domain. As proof of concept, we implement a bi- ologically inspired form of spike timing-dependent plasticity (STDP) based on relative timing of events in an AER framework. Experimen- tal results from an analog VLSI integrate-and-■re network demonstrate address domain learning in a task that requires neurons to group corre- lated inputs.
****

Efficient Learning Equilibrium

Ronen Brafman, Moshe Tennenholtz

We introduce efficient learning equilibrium (ELE), a normative ap(cid:173) proach to learning in non cooperative settings. In ELE, the learn(cid:173) ing algorithms themselves are required to be in equilibrium. In addition, the learning algorithms arrive at a desired value after polynomial time, and deviations from a prescribed ELE become ir(cid:173) rational after polynomial time. We prove the existence of an ELE in the perfect monitoring setting, where the desired value is the expected payoff in a Nash equilibrium. We also show that an ELE does not always exist in the imperfect monitoring case. Yet, it exists in the special case of common-interest games. Finally, we extend our results to general stochastic games.
****

Using Tarjan's Red Rule for Fast Dependency Tree Construction

Dan Pelleg, Andrew Moore

We focus on the problem of ef■cient learning of dependency trees. It is well-known that given the pairwise mutual information coef■cients, a minimum-weight spanning tree algorithm solves this problem exactly and in polynomial time. However, for large data-sets it is the construc- tion of the correlation matrix that dominates the running time. We have developed a new spanning-tree algorithm which is capable of exploiting partial knowledge about edge weights. The partial knowledge we main- tain is a probabilistic con■dence interval on the coef■cients, which we derive by examining just a small sample of the data. The algorithm is able to ■ag the need to shrink an interval, which translates to inspec- tion of more

data for the particular attribute pair. Experimental results show running time t hat is near-constant in the number of records, with- out signi■cant loss in accu racy of the generated trees. Interestingly, our spanning-tree algorithm is based solely on Tarjan’s red-edge rule, which is generally considered a guaranteed re cipe for bad performance.
****
Approximate Inference and Protein-Folding
Chen Yanover, Yair Weiss
Side-chain prediction is an important subtask in the protein-folding problem. We show that finding a minimal energy side-chain con(cid:173) figuration is equivalent to performing inference in an undirected graphical model . The graphical model is relatively sparse yet has many cycles. We used t his equivalence to assess the performance of approximate inference algorithms in a real-world setting. Specifi(cid:173) cally we compared belief propaga tion (BP), generalized BP (GBP) and naive mean field (MF). In cases where exact inference was possible, max-product BP al(cid:173) ways found the gl obal minimum of the energy (except in few cases where it failed to converge), while other approximation algorithms of similar complexity did not. In the full protein data set, max(cid:173) product BP always found a lower energy configuration than the other algorithms, including a widely used protein-folding software (SCWRL).
****
Learning a Forward Model of a Reflex
Bernd Porr, Florentin Wörgötter
We develop a systems theoretical treatment of a behavioural system that interact s with its environment in a closed loop situation such that its mo- tor actions in■uence its sensor inputs. The simplest form of a feedback is a re■ex. Re■exes occur always “too late”; i.e., only after a (unpleas- ant, painful, dangerous) r e■ex-eliciting sensor event has occurred. This de■nes an objective problem which can be solved if another sensor input exists which can predict the primary re■e x and can generate an earlier reaction. In contrast to previous approaches, our linear learning algo- rithm allows for an analytical proof that this system lear ns to apply feed- forward control with the result that slow feedback loops are r eplaced by their equivalent feed-forward controller creating a forward model. In other words, learning turns the reactive system into a pro-active system. By me ans of a robot implementation we demonstrate the applicability of the theoretica l results which can be used in a variety of different areas in physics and engin eering.
****
Information Regularization with Partially Labeled Data
Martin Szummer, Tommi Jaakkola
Classi■cation with partially labeled data requires using a large number of unlab eled examples (or an estimated marginal P (x)), to further con- strain the condi tional P (yjx) beyond a few available labeled examples. We formulate a regulariz ation approach to linking the marginal and the conditional in a general way. The regularization penalty measures the information that is implied about the label s over covering regions. No parametric assumptions are required and the approach remains tractable even for continuous marginal densities P (x). We develop algo rithms for solving the regularization problem for ■nite covers, establish a limi ting differential equation, and exemplify the behavior of the new regulariza- ti on approach in simple cases.
****
Intrinsic Dimension Estimation Using Packing Numbers
Balázs Kégl
We propose a new algorithm to estimate the intrinsic dimension of data sets. The method is based on geometric properties of the data and re- quires neither para metric assumptions on the data generating model nor input parameters to set. The method is compared to a similar, widely- used algorithm from the same family of geometric techniques. Experi- ments show that our method is more robust in term s of the data generating distribution and more reliable in the presence of noise

.
****

Mismatch String Kernels for SVM Protein Classification
Eleazar Eskin, Jason Weston, William Noble, Christina Leslie
We introduce a class of string kernels, called mismatch kernels, for use with support vector machines (SVMs) in a discriminative approach to the protein classi■cation problem. These kernels measure sequence sim- ilarity based on shared occurrences of -length subsequences, counted with up to mismatches, and do not rely on any generative model for the positive training sequences. We compute the kernels ef■ciently using a mismatch tree data structure and report experiments on a benchmark SCOP dataset, where we show that the mismatch kernel used with an SVM classi■er performs as well as the Fisher kernel, the most success- ful method for remote homology detection, while achieving considerable computational savings.
****

Adaptive Caching by Refetching
Robert B. Gramacy, Manfred K. K. Warmuth, Scott Brandt, Ismail Ari
We are constructing caching policies that have 13-20% lower miss rates than the best of twelve baseline policies over a large variety of request streams. This represents an improvement of 49–63% over Least Recently Used, the most commonly implemented policy. We achieve this not by designing a speci■c new policy but by using on-line Machine Learning algorithms to dynamically shift between the standard policies based on their observed miss rates. A thorough experimental evaluation of our techniques is given, as well as a discussion of what makes caching an interesting on-line learning problem.
****

Timing and Partial Observability in the Dopamine System
Nathaniel Daw, Aaron C. Courville, David Touretzky
According to a series of in■uential models, dopamine (DA) neurons sig- nal reward prediction error using a temporal-difference (TD) algorithm. We address a problem not convincingly solved in these accounts: how to maintain a representation of cues that predict delayed consequences. Our new model uses a TD rule grounded in partially observable semi-Markov processes, a formalism that captures two largely neglected features of DA experiments: hidden state and temporal variability. Previous models pre- dicted rewards using a tapped delay line representation of sensory inputs; we replace this with a more active process of inference about the under- lying state of the world. The DA system can then learn to map these inferred states to reward predictions using TD. The new model can ex- plain previously vexing data on the responses of DA neurons in the face of temporal variability. By combining statistical model-based learning with a physiologically grounded TD theory, it also brings into contact with physiology some insights about behavior that had previously been con■ned to more abstract psychological models.
****

Multiple Cause Vector Quantization
David Ross, Richard Zemel
We propose a model that can learn parts-based representations of high- dimensional data. Our key assumption is that the dimensions of the data can be separated into several disjoint subsets, or factors, which take on values independently of each other. We assume each factor has a small number of discrete states, and model it using a vector quantizer. The selected states of each factor represent the multiple causes of the input. Given a set of training examples, our model learns the association of data dimensions with factors, as well as the states of each VQ. Inference and learning are carried out ef■ciently via variational algorithms. We present applications of this model to problems in image decomposition, collaborative ■ltering, and text classi■cation.
****

Unsupervised Color Constancy
Kinh Tieu, Erik Miller
In [1] we introduced a linear statistical model of joint color changes in images due to variation in lighting and certain non-geometric camera pa- rameters. We

did this by measuring the mappings of colors in one image of a scene to colors i n another image of the same scene under different lighting conditions. Here we i ncrease the ■exibility of this color ■ow model by allowing ■ow coef■cients to va ry according to a low order polynomial over the image. This allows us to better ■t smoothly vary- ing lighting conditions as well as curved surfaces without end owing our model with too much capacity. We show results on image matching and sh adow removal and detection.
****

Value-Directed Compression of POMDPs
Pascal Poupart, Craig Boutilier
We examine the problem of generating state-space compressions of POMDPs in a way that minimally impacts decision quality. We analyze the impact of compres- sion s on decision quality, observing that compressions that allow accurate policy ev aluation (prediction of expected future reward) will not affect decision qual- i ty. We derive a set of suf■cient conditions that ensure accurate prediction in t his respect, illustrate interesting mathematical properties these confer on loss less lin- ear compressions, and use these to derive an iterative procedure for ■ nding good linear lossy compressions. We also elaborate on how structured repres entations of a POMDP can be used to ■nd such compressions.
****

Constraint Classification for Multiclass Classification and Ranking
Sariel Har-Peled, Dan Roth, Dav Zimak
The constraint classi■cation framework captures many ■avors of mul- ticlass clas si■cation including winner-take-all multiclass classi■cation, multilabel classi■ cation and ranking. We present a meta-algorithm for learning in this framework t hat learns via a single linear classi■er in high dimension. We discuss distribut ion independent as well as margin-based generalization bounds and present empiri cal and theoretical evidence showing that constraint classi■cation bene■ts over existing methods of multiclass classi■cation.
****

Neural Decoding of Cursor Motion Using a Kalman Filter
W Wu, M. Black, Y. Gao, M. Serruya, A. Shaikhouni, J. Donoghue, Elie Bienenstock
The direct neural control of external devices such as computer displays or prost hetic limbs requires the accurate decoding of neural activity rep- resenting con tinuous movement. We develop a real-time control system using the spiking activi ty of approximately 40 neurons recorded with an electrode array implanted in the arm area of primary motor cortex. In contrast to previous work, we develop a co ntrol-theoretic approach that explicitly models the motion of the hand and the p robabilistic re- lationship between this motion and the mean ■ring rates of the cells in 70 bins. We focus on a realistic cursor control task in which the sub- ject must move a cursor to “hit” randomly placed targets on a computer monitor. Encoding and decoding of the neural data is achieved with a Kalman ■lter which h as a number of advantages over previous linear ■ltering techniques. In particula r, the Kalman ■lter reconstructions of hand trajectories in off-line experiments are more accurate than previ- ously reported results and the model provides ins ights into the nature of the neural coding of movement.
****

On the Dirichlet Prior and Bayesian Regularization
Harald Steck, Tommi Jaakkola
A common objective in learning a model from data is to recover its ne twork structure, while the model parameters are of minor in(cid:173) terest. F or example, we may wish to recover regulatory networks from high-through put data sources. In this paper we examine how Bayesian regularization usi ng a product of independent Dirichlet priors over the model parameters af fects the learned model struc(cid:173) ture in a domain with discrete v ariables. We show that a small scale parameter - often interpreted as "e quivalent sample size" or "prior strength" - leads to a strong regulariz ation of the model structure (sparse graph) given a sufficiently large dat a set. In par(cid:173) ticular, the empty graph is obtained in the limit of a vanishing scale parameter. This is diametrically opposite to what one may exp

ect in this limit, namely the complete graph from an (unregularized) maximum likelihood estimate. Since the prior affects the parame(cid:173) ters as expected, the scale parameter balances a trade-off between regularizing the parameters vs. the structure of the model. We demonstrate the benefits of optimizing this trade-off in the sense of predictive accuracy.
****
Scaling of Probability-Based Optimization Algorithms
J. Shapiro
Population-based Incremental Learning is shown require very sen(cid:173) sitive scaling of its learning rate. The learning rate must scale with the system size in a problem-dependent way. This is shown in two problems: the needle-in-a haystack, in which the learning rate must vanish exponentially in the system size, and in a smooth function in which the learning rate must vanish like the square root of the system size. Two methods are proposed for removing this sensitiv(cid:173) ity. A learning dynamics which obeys detailed balance is shown to give consistent performance over the entire range of learning rates. An analog of mutation is shown to require a learning rate which scales as the inverse system size, but is problem independent.
****
Forward-Decoding Kernel-Based Phone Recognition
Shantanu Chakrabartty, Gert Cauwenberghs
Forward decoding kernel machines (FDKM) combine large-margin clas(cid:173) sifiers with hidden Markov models (HMM) for maximum a posteriori (MAP) adaptive sequence estimation. State transitions in the sequence are conditioned on observed data using a kernel-based probability model trained with a recursive scheme that deals effectively with noisy and par(cid:173) tially labeled data. Training over very large data sets is accomplished us(cid:173) ing a sparse probabilistic support vector machine (SVM) model based on quadratic entropy, and an on-line stochastic steepest descent algorithm. For speaker-independent continuous phone recognition, FDKM trained over 177 ,080 samples of the TlMIT database achieves 80.6% recognition accuracy over the full test set, without use of a prior phonetic language model.
****
Optoelectronic Implementation of a FitzHugh-Nagumo Neural Model
Alexandre Romariz, Kelvin Wagner
An optoelectronic implementation of a spiking neuron model based on the FitzHugh-Nagumo equations is presented. A tunable semiconduc- tor laser source and a spectral ■lter provide a nonlinear mapping from driver voltage to detected signal. Linear electronic feedback completes the implementation, which allows either electronic or optical input sig- nals. Experimental results for a single system and numeric results of model interaction con■rm that important features of spiking neural mod- els can be implemented through this approach.
****
Margin-Based Algorithms for Information Filtering
Nicolò Cesa-bianchi, Alex Conconi, Claudio Gentile
In this work, we study an information ■ltering model where the relevance labels associated to a sequence of feature vectors are realizations of an unknown probabilistic linear function. Building on the analysis of a re- stricted version of our model, we derive a general ■ltering rule based on the margin of a ridge regression estimator. While our rule may observe the label of a vector only by classfying the vector as relevant, experiments on a real-world document ■ltering problem show that the performance of our rule is close to that of the on-line classi■er which is allowed to observe all labels. These empirical results are complemented by a theo- retical analysis where we consider a randomized variant of our rule and prove that its expected number of mistakes is never much larger than that of the optimal ■ltering rule which knows the hidden linear model.
****
Half-Lives of EigenFlows for Spectral Clustering
Chakra Chennubhotla, Allan Jepson
Using a Markov chain perspective of spectral clustering we present an algorithm

to automatically ■nd the number of stable clusters in a dataset. The Markov chain's behaviour is characterized by the spectral properties of the matrix of transition probabilities, from which we derive eigen■ows along with their hal■ives. An eigen■ow describes the ■ow of probabil- ity mass due to the Markov chain, and it is characterized by its eigen- value, or equivalently, by the hal■ife of its decay as the Markov chain is iterated. A ideal stable cluster is one with zero eigen■ow and in■- nite half-life. The key insight in this paper is that bottlenecks between weakly coupled clusters can be identi■ed by computing the sensitivity of the eigen■ow's hal■ife to variations in the edge weights. We propose a novel EIGENCUTS algorithm to perform clustering that removes these identi■ed bottlenecks in an iterative fashion.
****
The RA Scanner: Prediction of Rheumatoid Joint Inflammation Based on Laser Imaging
Anton Schwaighofer, Volker Tresp, Peter Mayer, Alexander Scheel, Gerhard Müller
We describe the RA scanner, a novel system for the examination of pa- tients suffering from rheumatoid arthritis. The RA scanner is based on a novel laser-based imaging technique which is sensitive to the optical characteristics of ■nger joint tissue. Based on the laser images, ■nger joints are classi■ed according to whether the in■ammatory status has improved or worsened. To perform the classi■cation task, various lin- ear and kernel-based systems were implemented and their performances were compared. Special emphasis was put on measures to reliably per- form parameter tuning and evaluation, since only a very small data set was available. Based on the results presented in this paper, it was con- cluded that the RA scanner permits a reliable classi■cation of patholog- ical ■nger joints, thus paving the way for a further development from prototype to product stage.
****
Optimality of Reinforcement Learning Algorithms with Linear Function Approximation
Ralf Schoknecht
There are several reinforcement learning algorithms that yield ap(cid:173) proximate solutions for the problem of policy evaluation when the value function is represented with a linear function approximator. In this paper we show that each of the solutions is optimal with respect to a specific objective function. Moreover, we characterise the different solutions as images of the optimal exact value func(cid:173) tion under different projection operations. The results presented here will be useful for comparing the algorithms in terms of the error they achieve relative to the error of the optimal approximate solution.
****
Evidence Optimization Techniques for Estimating Stimulus-Response Functions
Maneesh Sahani, Jennifer Linden
An essential step in understanding the function of sensory nervous sys- tems is to characterize as accurately as possible the stimulus-response function (SRF) of the neurons that relay and process sensory informa- tion. One increasingly common experimental approach is to present a rapidly varying complex stimulus to the animal while recording the re- sponses of one or more neurons, and then to directly estimate a func- tional transformation of the input that accounts for the neuronal ■ring. The estimation techniques usually employed, such as Wiener ■ltering or other correlation-based estimation of the Wiener or Volterra kernels, are equivalent to maximum likelihood estimation in a Gaussian-output-noise regression model. We explore the use of Bayesian evidence-optimization techniques to condition these estimates. We show that by learning hyper- parameters that control the smoothness and sparsity of the transfer func- tion it is possible to improve dramatically the quality of SRF estimates, as measured by their success in predicting responses to novel input.
****
Binary Coding in Auditory Cortex
Michael Deweese, Anthony Zador
Cortical neurons have been reported to use both rate and temporal codes. Here we describe a novel mode in which each neuron generates exactly 0 or 1 action po

tentials, but not more, in response to a stimulus. We used cell-attached record ing, which ensured single-unit isolation, to record responses in rat auditory c ortex to brief tone pips. Surprisingly, the majority of neurons exhibited bina ry behavior with few multi-spike responses; several dramatic examples consisted of exactly one spike on 100% of trials, with no trial-to-trial variability in spike count. Many neurons were tuned to stimulus frequency. Since individual tr ials yielded at most one spike for most neurons, the information about stimulus frequency was encoded in the population, and would not have been accessible t o later stages of processing that only had access to the activity of a single u nit. These binary units allow a more efficient population code than is possible with conventional rate coding units, and are consistent with a model of cortic al processing in which synchronous packets of spikes propagate stably from one neuronal population to the next.
****
Learning Attractor Landscapes for Learning Motor Primitives
Auke Ijspeert, Jun Nakanishi, Stefan Schaal
Many control problems take place in continuous state-action spaces, e.g., as in manipulator robotics, where the control objective is of- ten de■ned as ■nding a desired trajectory that reaches a particular goal state. While reinforcement lea rning o■ers a theoretical frame- work to learn such control policies from scratc h, its applicability to higher dimensional continuous state-action spaces remain s rather limited to date. Instead of learning from scratch, in this paper we sug gest to learn a desired complex control policy by transforming an existing simpl e canonical control policy. For this purpose, we represent canonical policies in terms of di■erential equations with well-de■ned attractor properties. By nonlin early transforming the canonical attractor dynamics using techniques from nonpar ametric regression, almost arbitrary new nonlinear policies can be gener- ated w ithout losing the stability properties of the canonical sys- tem. We demonstrate our techniques in the context of learning a set of movement skills for a humano id robot from demonstrations of a human teacher. Policies are acquired rapidly, and, due to the properties of well formulated di■erential equations, can be re-u sed and modi■ed on-line under dynamic changes of the environment. The linear par ameterization of nonparametric regression moreover lends itself to recognize and classify previously learned movement skills. Evaluations in simulations and on an actual 30 degree-of- freedom humanoid robot exemplify the feasibility and rob ustness of our approach.
****
Combining Dimensions and Features in Similarity-Based Representations
Daniel Navarro, Michael Lee
This paper develops a new representational model of similarity data that combine s continuous dimensions with discrete features. An al- gorithm capable of learni ng these representations is described, and a Bayesian model selection approach f or choosing the appropriate number of dimensions and features is developed. The approach is demonstrated on a classic data set that considers the similarities b etween the numbers 0 through 9.
****
Bayesian Monte Carlo
Zoubin Ghahramani, Carl Rasmussen
We investigate Bayesian alternatives to classical Monte Carlo methods for evalua ting integrals. Bayesian Monte Carlo (BMC) allows the in- corporation of prior k nowledge, such as smoothness of the integrand, into the estimation. In a simple problem we show that this outperforms any classical importance sampling method. We also attempt more chal- lenging multidimensional integrals involved in comput ing marginal like- lihoods of statistical models (a.k.a. partition functions and model evi- dences). We ■nd that Bayesian Monte Carlo outperformed Annealed Impo rtance Sampling, although for very high dimensional problems or problems with ma ssive multimodality BMC may be less adequate. One advantage of the Bayesian appr oach to Monte Carlo is that samples can be drawn from any distribution. This all ows for the possibility of active design of sample points so as to maximise info rmation gain.

****
A Model for Learning Variance Components of Natural Images
Yan Karklin, Michael Lewicki
We present a hierarchical Bayesian model for learning ef■cient codes of higher-order structure in natural images. The model, a non-linear gen- eralization of independent component analysis, replaces the standard as- sumption of independence for the joint distribution of coef■cients with a distribution that is adapted to the variance structure of the coef■cients of an ef■cient image basis. This offers a novel description of higher- order image structure and provides a way to learn coarse-coded, sparse- distributed representations of abstract image properties such as object location, scale, and texture.
****
Effective Dimension and Generalization of Kernel Learning
Tong Zhang
We investigate the generalization performance of some learning prob- lems in Hilbert function Spaces. We introduce a concept of scale- sensitive effective data dimension, and show that it characterizes the con- vergence rate of the underlying learning problem. Using this concept, we can naturally extend results for parametric estimation problems in ■nite dimensional spaces to non-parametric kernel learning methods. We de- rive upper bounds on the generalization performance and show that the resulting convergent rates are optimal under various circumstances.
****
Nash Propagation for Loopy Graphical Games
Luis E. Ortiz, Michael Kearns
We introduce NashProp, an iterative and local message-passing algo- rithm for computing Nash equilibria in multi-player games represented by arbitrary undirected graphs. We provide a formal analysis and exper- imental evidence demonstrating that NashProp performs well on large graphical games with many loops, often converging in just a dozen itera- tions on graphs with hundreds of nodes. NashProp generalizes the tree algorithm of (Kearns et al. 2001), and can be viewed as similar in spirit to belief propagation in probabilis- tic inference, and thus complements the recent work of (Vickrey and Koller 2002), who explored a junction tree approach. Thus, as for prob- abilistic inference, we have at least two promising general-purpose ap- proaches to equilibria computation in graphs.
****
Neuromorphic Bisable VLSI Synapses with Spike-Timing-Dependent Plasticity
Giacomo Indiveri
We present analog neuromorphic circuits for implementing bistable syn- apses with spike-timing-dependent plasticity (STDP) properties. In these types of synapses, the short-term dynamics of the synaptic ef■cacies are governed by the relative timing of the pre- and post-synaptic spikes, while on long time scales the ef■cacies tend asymptotically to either a potentiated state or to a depressed one. We fabricated a prototype VLSI chip containing a network of integrate and ■re neurons interconnected via bistable STDP synapses. Test results from this chip demonstrate the synapse’s STDP learning properties, and its long-term bistable charac- teristics.
****
Source Separation with a Sensor Array using Graphical Models and Subband Filtering
Hagai Attias
Source separation is an important problem at the intersection of several ■elds, including machine learning, signal processing, and speech tech- nology. Here we describe new separation algorithms which are based on probabilistic graphical models with latent variables. In contrast with existing methods, these algorithms exploit detailed models to describe source properties. They also use subband ■ltering ideas to model the reverberant environment, and employ an explicit model for background and sensor noise. We leverage variational techniques to keep the compu- tational complexity per EM iteration linear in the number of frames.
****

Dopamine Induced Bistability Enhances Signal Processing in Spiny Neurons
Aaron Gruber, Sara Solla, James Houk
Single unit activity in the striatum of awake monkeys shows a marked dependence on the expected reward that a behavior will elicit. We pr esent a computational model of spiny neurons, the principal neurons of t he striatum, to assess the hypothesis that di(cid:173) rect neuromodulatory ef fects of dopamine through the activation of D 1 receptors mediate the r eward dependency of spiny neuron activity. Dopamine release results in the amplification of key ion currents, leading to the emergence of bis tability, which not only modulates the peak firing rate but also introduces a temporal and state dependence of the model's response, thus improving t he de(cid:173) tectability of temporally correlated inputs.
****
An Information Theoretic Approach to the Functional Classification of Neurons
Elad Schneidman, William Bialek, Michael Ii
A population of neurons typically exhibits a broad diversity of responses to sen sory inputs. The intuitive notion of functional classi■cation is that cells can be clustered so that most of the diversity is captured by the iden- tity of the clusters rather than by individuals within clusters. We show how this intuition can be made precise using information theory, with- out any need to introduce a metric on the space of stimuli or responses. Applied to the retinal ganglion cel ls of the salamander, this approach re- covers classical results, but also provi des clear evidence for subclasses beyond those identi■ed previously. Further, we ■nd that each of the gan- glion cells is functionally unique, and that even wit hin the same subclass only a few spikes are needed to reliably distinguish betwe en cells.
****
An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition
Samy Bengio
This paper presents a novel Hidden Markov Model architecture to model the jo int probability of pairs of asynchronous sequences de(cid:173) scribing the same event. It is based on two other Markovian models, namely Asynchronous Input / Output Hidden Markov Models and Pair Hidden Markov Models. An EM algorit hm to train the model is presented, as well as a Viterbi decoder that c an be used to ob(cid:173) tain the optimal state sequence as well as the alignment between the two sequences. The model has been tested on an a udio-visual speech recognition task using the M2VTS database and yielde d robust performances under various noise conditions.
****
Shape Recipes: Scene Representations that Refer to the Image
William Freeman, Antonio Torralba
The goal of low-level vision is to estimate an underlying scene, given an observ ed image. Real-world scenes (eg, albedos or shapes) can be very complex, convent ionally requiring high dimensional representations which are hard to estimate an d store. We propose a low-dimensional rep- resentation, called a scene recipe, t hat relies on the image itself to de- scribe the complex scene con■gurations. Sh ape recipes are an example: these are the regression coef■cients that predict th e bandpassed shape from image data. We describe the bene■ts of this representati on, and show two uses illustrating their properties: (1) we improve stereo shape estimates by learning shape recipes at low resolution and applying them at full resolution; (2) Shape recipes implicitly contain information about lighting and materials and we use them for material segmentation.
****
Real-Time Particle Filters
Cody Kwok, Dieter Fox, Marina Meila
Particle ■lters estimate the state of dynamical systems from sensor infor- matio n. In many real time applications of particle ■lters, however, sensor informatio n arrives at a signi■cantly higher rate than the update rate of the ■lter. The p revalent approach to dealing with such situations is to update the particle ■lte r as often as possible and to discard sensor information that cannot be processe

d in time. In this paper we present real-time particle ■l- ters, which make use of all sensor information even when the ■lter update rate is below the update ra te of the sensors. This is achieved by represent- ing posteriors as mixtures of sample sets, where each mixture component integrates one observation arriving du ring a ■lter update. The weights of the mixture components are set so as to mini mize the approximation error introduced by the mixture representation. Thereby, our approach focuses computational resources (samples) on valuable sensor inform ation. Exper- iments using data collected with a mobile robot show that our appr oach yields strong improvements over other approaches.
****
Critical Lines in Symmetry of Mixture Models and its Application to Component Sp litting
Kenji Fukumizu, Shotaro Akaho, Shun-ichi Amari
We show the existence of critical points as lines for the likelihood func- tion of mixture-type models. They are given by embedding of a critical point for mode ls with less components. A suf■cient condition that the critical line gives loca l maxima or saddle points is also derived. Based on this fact, a component-split method is proposed for a mixture of Gaus- sian components, and its effectivenes s is veri■ed through experiments.
****
Manifold Parzen Windows
Pascal Vincent, Yoshua Bengio
The similarity between objects is a fundamental element of many learn- ing algor ithms. Most non-parametric methods take this similarity to be ■xed, but much rec ent work has shown the advantages of learning it, in particular to exploit the l ocal invariances in the data or to capture the possibly non-linear manifold on w hich most of the data lies. We propose a new non-parametric kernel density estim ation method which captures the local structure of an underlying manifold throug h the leading eigen- vectors of regularized local covariance matrices. Experimen ts in density estimation show signi■cant improvements with respect to Parzen den sity estimators. The density estimators can also be used within Bayes classi- ■e rs, yielding classi■cation rates similar to SVMs and much superior to the Parzen classi■er.
****
Parametric Mixture Models for Multi-Labeled Text
Naonori Ueda, Kazumi Saito
We propose probabilistic generative models, called parametric mix- ture models (PMMs), for multiclass, multi-labeled text categoriza- tion problem. Conventional ly, the binary classi(cid:12)cation approach has been employed, in which whether or not text belongs to a cat- egory is judged by the binary classi(cid:12)er fo r every category. In con- trast, our approach can simultaneously detect multiple categories of text using PMMs. We derive e(cid:14)cient learning and prediction algo- rithms for PMMs. We also empirically show that our method could signi(cid :12)cantly outperform the conventional binary methods when ap- plied to multi-la beled text categorization using real World Wide Web pages.
****
Transductive and Inductive Methods for Approximate Gaussian Process Regression
Anton Schwaighofer, Volker Tresp
Gaussian process regression allows a simple analytical treatment of ex- act Baye sian inference and has been found to provide good performance, yet scales badly with the number of training data. In this paper we com- pare several approaches towards scaling Gaussian processes regression to large data sets: the subset of representers method, the reduced rank approximation, online Gaussian processes, and the Bayesian commit- tee machine. Furthermore we provide theoretical insight into some of our experimental results. We found that subset of representers met hods can give good and particularly fast predictions for data sets with high and medium noise levels. On complex low noise data sets, the Bayesian committee mac hine achieves signi■cantly better accuracy, yet at a higher computational cost.
****
Adapting Codes and Embeddings for Polychotomies

Gunnar Rätsch, Sebastian Mika, Alex Smola
In this paper we consider formulations of multi-class problems based on a genera lized notion of a margin and using output coding. This includes, but is not rest ricted to, standard multi-class SVM formulations. Differ- ently from many previo us approaches we learn the code as well as the embedding function. We illustrate how this can lead to a formulation that allows for solving a wider range of pro blems with for instance many classes or even “missing classes”. To keep our opti mization problems tractable we propose an algorithm capable of solving them usin g two- class classi■ers, similar in spirit to Boosting.
****
Topographic Map Formation by Silicon Growth Cones
Brian Taba, Kwabena A. Boahen
We describe a self-configuring neuromorphic chip that uses a model of a ctivity-dependent axon remodeling to automatically wire topographic maps bas ed solely on input correlations. Axons are guided by growth cones, which are modeled in analog VLSI for the first time. Growth cones migrate up ne urotropin gradients, which are represented by charge diffusing in trans istor channels. Virtual axons move by rerouting address-events. We refi ned an initially gross topographic projection by simulating retinal wave inpu t.
****
Analysis of Information in Speech Based on MANOVA
Sachin Kajarekar, Hynek Hermansky
We propose analysis of information in speech using three sources - lang uage (phone), speaker and channeL Information in speech is measured as mutua l information between the source and the set of features extracted from spee ch signaL We assume that distribu(cid:173) tion of features can be modeled using Gaussian distribution. The mutual information is computed using the results of analysis of variability in speech. We observe similarity i n the results of phone variability and phone information, and show that the res ults of the proposed analysis have more meaningful interpretations than the analysis of variability.
****
Discriminative Binaural Sound Localization
Ehud Ben-reuven, Yoram Singer
Time difference of arrival (TDOA) is commonly used to estimate the az- imuth of a source in a microphone array. The most common methods to estimate TDOA are bas ed on ■nding extrema in generalized cross- correlation waveforms. In this paper we apply microphone array tech- niques to a manikin head. By considering the ent ire cross-correlation waveform we achieve azimuth prediction accuracy that excee ds extrema locating methods. We do so by quantizing the azimuthal angle and trea ting the prediction problem as a multiclass categorization task. We demonstrate the merits of our approach by evaluating the various ap- proaches on Sony’s AIBO robot.
****
Fractional Belief Propagation
Wim Wiegerinck, Tom Heskes
We consider loopy belief propagation for approximate inference in prob- abilisti c graphical models. A limitation of the standard algorithm is that clique margin als are computed as if there were no loops in the graph. To overcome this limita tion, we introduce fractional belief propagation. Fractional belief propagation is formulated in terms of a family of ap- proximate free energies, which include s the Bethe free energy and the naive mean-■eld free as special cases. Using the linear response correc- tion of the clique marginals, the scale parameters can be tuned. Simula- tion results illustrate the potential merits of the approach.
****
Stability-Based Model Selection
Tilman Lange, Mikio Braun, Volker Roth, Joachim Buhmann
Model selection is linked to model assessment, which is the problem of comparing different models, or model parameters, for a speci■c learning task. For supervi

sed learning, the standard practical technique is cross- validation, which is no
t applicable for semi-supervised and unsupervised settings. In this paper, a new
 model assessment scheme is introduced which is based on a notion of stability.
The stability measure yields an upper bound to cross-validation in the supervise
d case, but extends to semi-supervised and unsupervised problems. In the experim
ental part, the performance of the stability measure is studied for model order
se- lection in comparison to standard techniques in this area.
****
Hidden Markov Model of Cortical Synaptic Plasticity: Derivation of the Learning
Rule
Michael Eisele, Kenneth Miller
Cortical synaptic plasticity depends on the relative timing of pre- and postsyna
ptic spikes and also on the temporal pattern of presynaptic spikes and of postsy
naptic spikes. We study the hypothesis that cortical synap- tic plasticity does
not associate individual spikes, but rather whole ■r- ing episodes, and depends
only on when these episodes start and how long they last, but as little as possi
ble on the timing of individual spikes. Here we present the mathematical backgro
und for such a study. Stan- dard methods from hidden Markov models are used to d
e■ne what “■r- ing episodes” are. Estimating the probability of being in such an
 episode requires not only the knowledge of past spikes, but also of future spik
es. We show how to construct a causal learning rule, which depends only on past
spikes, but associates pre- and postsynaptic ■ring episodes as if it also knew f
uture spikes. We also show that this learning rule agrees with some features of
synaptic plasticity in super■cial layers of rat visual cortex (Froemke and Dan,
Nature 416:433, 2002).
****
Automatic Alignment of Local Representations
Yee Teh, Sam Roweis
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Field-Programmable Learning Arrays
Seth Bridges, Miguel Figueroa, Chris Diorio, David Hsu
This paper introduces the Field-Programmable Learning Array, a new paradigm for
rapid prototyping of learning primitives and machine- learning algorithms in sil
icon. The FPLA is a mixed-signal counterpart to the all-digital Field-Programmab
le Gate Array in that it enables rapid prototyping of algorithms in hardware. Un
like the FPGA, the FPLA is targeted directly for machine learning by providing l
ocal, parallel, on- line analog learning using ■oating-gate MOS synapse transist
ors. We present a prototype FPLA chip comprising an array of recon■gurable compu
tational blocks and local interconnect. We demonstrate the via- bility of this a
rchitecture by mapping several learning circuits onto the prototype chip.
****
Boosting Density Estimation
Saharon Rosset, Eran Segal
Several authors have suggested viewing boosting as a gradient descent search for
 a good ■t in function space. We apply gradient-based boosting methodology to th
e unsupervised learning problem of density estimation. We show convergence prope
rties of the algorithm and prove that a strength of weak learnability prop- erty
 applies to this problem as well. We illustrate the potential of this approach t
hrough experiments with boosting Bayesian networks to learn density models.
****
Support Vector Machines for Multiple-Instance Learning
Stuart Andrews, Ioannis Tsochantaridis, Thomas Hofmann
This paper presents two new formulations of multiple-instance learning a
s a maximum margin problem. The proposed extensions of the Support Vector
Machine (SVM) learning approach lead to mixed integer quadratic programs th
at can be solved heuristically. Our generalization of SVMs makes a state-of

-the-art classification technique, including non-linear classification via k
ernels, available to an area that up to now has been largely dominated
by special purpose methods. We present experimental results on a pharm
a(cid:173) ceutical data set and on applications in automated image indexing an
d document categorization.
****
Bias-Optimal Incremental Problem Solving
Jürgen Schmidhuber
Given is a problem sequence and a probability distribution (the bias) on program
s computing solution candidates. We present an optimally fast way of incremental
ly solving each task in the sequence. Bias shifts are computed by program pre■xe
s that modify the distribution on their suf- ■xes by reusing successful code for
previous tasks (stored in non-modi■- able memory). No tested program gets more
runtime than its probability times the total search time. In illustrative experi
ments, ours becomes the ■rst general system to learn a universal solver for arbi
trary disk Tow- ers of Hanoi tasks (minimal solution size). It demonstrates t
he advantages of incremental learning by pro■ting from previously solved, simple
r tasks involving samples of a simple context free language.
****
Rate Distortion Function in the Spin Glass State: A Toy Model
Tatsuto Murayama, Masato Okada
We applied statistical mechanics to an inverse problem of linear mapping to inve
stigate the physics of optimal lossy compressions. We used the replica symmetry
breaking technique with a toy model to demonstrate Shannon’s result. The rate di
stortion function, which is widely known as the theoretical limit of the compres
sion with a ■delity criterion, is derived. Numerical study shows that sparse con
structions of the model provide suboptimal compressions.
****
Adaptive Nonlinear System Identification with Echo State Networks
Herbert Jaeger
Echo state networks (ESN) are a novel approach to recurrent neu(cid:173) ral
network training. An ESN consists of a large, fixed, recurrent "reservo
ir" network, from which the desired output is obtained by training suitable
output connection weights. Determination of op(cid:173) timal output weights
becomes a linear, uniquely solvable task of MSE minimization. This ar
ticle reviews the basic ideas and de(cid:173) scribes an online adaptat
ion scheme based on the RLS algorithm known from adaptive linear system
s. As an example, a 10-th or(cid:173) der NARMA system is adaptively id
entified. The known benefits of the RLS algorithms carryover from linear sy
stems to nonlinear ones; specifically, the convergence rate and misadjustme
nt can be determined at design time.
****
Real Time Voice Processing with Audiovisual Feedback: Toward Autonomous Agents w
ith Perfect Pitch
Lawrence Saul, Daniel Lee, Charles Isbell, Yann Cun
We have implemented a real time front end for detecting voiced speech and estima
ting its fundamental frequency. The front end performs the signal processing for
voice-driven agents that attend to the pitch contours of human speech and provi
de continuous audiovisual feedback. The al- gorithm we use for pitch tracking ha
s several distinguishing features: it makes no use of FFTs or autocorrelation at
the pitch period; it updates the pitch incrementally on a sample-by-sample basi
s; it avoids peak picking and does not require interpolation in time or frequenc
y to obtain high res- olution estimates; and it works reliably over a four octav
e range, in real time, without the need for postprocessing to produce smooth con
tours. The algorithm is based on two simple ideas in neural computation: the int
roduction of a purposeful nonlinearity, and the error signal of a least squares
■t. The pitch tracker is used in two real time multimedia applica- tions: a voic
e-to-MIDI player that synthesizes electronic music from vo- calized melodies, an
d an audiovisual Karaoke machine with multimodal feedback. Both applications run
on a laptop and display the user’s pitch scrolling across the screen as he or s

he sings into the computer.
****
An Impossibility Theorem for Clustering
Jon Kleinberg
Although the study of clustering is centered around an intuitively compelling go
al, it has been very di(cid:14)cult to develop a uni(cid:12)ed framework for rea
soning about it at a technical level, and pro- foundly diverse approaches to clu
stering abound in the research community. Here we suggest a formal perspective o
n the di(cid:14)culty in (cid:12)nding such a uni(cid:12)cation, in the form of
an impossibility theo- rem: for a set of three simple properties, we show that t
here is no clustering function satisfying all three. Relaxations of these prop-
erties expose some of the interesting (and unavoidable) trade-o(cid:11)s at work
 in well-studied clustering techniques such as single-linkage, sum-of-pairs, k-m
eans, and k-median.
****
Visual Development Aids the Acquisition of Motion Velocity Sensitivities
Robert Jacobs, Melissa Dominguez
We consider the hypothesis that systems learning aspects of visual per- ception
may bene■t from the use of suitably designed developmental pro- gressions during
 training. Four models were trained to estimate motion velocities in sequences o
f visual images. Three of the models were “de- velopmental models” in the sense
that the nature of their input changed during the course of training. They recei
ved a relatively impoverished visual input early in training, and the quality of
 this input improved as training progressed. One model used a coarse-to-multisca
le develop- mental progression (i.e. it received coarse-scale motion features ea
rly in training and ■ner-scale features were added to its input as training prog
ressed), another model used a ■ne-to-multiscale progression, and the third model
 used a random progression. The ■nal model was non- developmental in the sense t
hat the nature of its input remained the same throughout the training period. Th
e simulation results show that the coarse-to-multiscale model performed best. Hy
potheses are offered to account for this model’s superior performance. We conclu
de that suit- ably designed developmental sequences can be useful to systems lea
rn- ing to estimate motion velocities. The idea that visual development can aid
visual learning is a viable hypothesis in need of further study.
****
On the Complexity of Learning the Kernel Matrix
Olivier Bousquet, Daniel Herrmann
We investigate data based procedures for selecting the kernel when learn- ing wi
th Support Vector Machines. We provide generalization error bounds by estimating
 the Rademacher complexities of the corresponding function classes. In particula
r we obtain a complexity bound for function classes induced by kernels with give
n eigenvectors, i.e., we allow to vary the spectrum and keep the eigenvectors ■x
. This bound is only a loga- rithmic factor bigger than the complexity of the fu
nction class induced by a single kernel. However, optimizing the margin over suc
h classes leads to over■tting. We thus propose a suitable way of constraining th
e class. We use an ef■cient algorithm to solve the resulting optimization proble
m, present preliminary experimental results, and compare them to an alignment-ba
sed approach.
****
Clustering with the Fisher Score
Koji Tsuda, Motoaki Kawanabe, Klaus-Robert Müller
Recently the Fisher score (or the Fisher kernel) is increasingly used as a featu
re extractor for classi■cation problems. The Fisher score is a vector of paramet
er derivatives of loglikelihood of a probabilistic model. This paper gives a the
oretical analysis about how class information is pre- served in the space of the
 Fisher score, which turns out that the Fisher score consists of a few important
 dimensions with class information and many nuisance dimensions. When we perform
 clustering with the Fisher score, K-Means type methods are obviously inappropri
ate because they make use of all dimensions. So we will develop a novel but simp
le clus- tering algorithm specialized for the Fisher score, which can exploit im

- portant dimensions. This algorithm is successfully tested in experiments with artiￚcial data and real data (amino acid sequences).
****
Artefactual Structure from Least-Squares Multidimensional Scaling
Nicholas Hughes, David Lowe
We consider the problem of illusory or artefactual structure from the vi- sualis ation of high-dimensional structureless data. In particular we ex- amine the rol e of the distance metric in the use of topographic mappings based on the statist ical ￚeld of multidimensional scaling. We show that the use of a squared Euclide an metric (i.e. the SS TRESS measure) gives rise to an annular structure when th e input data is drawn from a high- dimensional isotropic distribution, and we pr ovide a theoretical justiￚca- tion for this observation.
****
Adaptation and Unsupervised Learning
Peter Dayan, Maneesh Sahani, Gregoire Deback
Adaptation is a ubiquitous neural and psychological phenomenon, with a wealth of instantiations and implications. Although a basic form of plasticity, it has, b ar some notable exceptions, attracted computational theory of only one main vari ety. In this paper, we study adaptation from the perspective of factor analysis, a paradigmatic technique of unsuper- vised learning. We use factor analysis to re-interpret a standard view of adaptation, and apply our new model to some rece nt data on adaptation in the domain of face discrimination.
****
Learning in Zero-Sum Team Markov Games Using Factored Value Functions
Michail G. Lagoudakis, Ronald Parr
We present a new method for learning good strategies in zero-sum Markov games in which each side is composed of multiple agents col- laborating against an oppos ing team of agents. Our method requires full observability and communication dur ing learning, but the learned poli- cies can be executed in a distributed manner . The value function is rep- resented as a factored linear architecture and its structure determines the necessary computational resources and communication ban dwidth. This approach permits a tradeoff between simple representations with lit tle or no communication between agents and complex, computationally inten- sive representations with extensive coordination between agents. Thus, we provide a p rincipled means of using approximation to combat the exponential blowup in the j oint action space of the participants. The ap- proach is demonstrated with an ex ample that shows the efￚciency gains over naive enumeration.
****
Convergence Properties of Some Spike-Triggered Analysis Techniques
Liam Paninski
vVe analyze the convergence properties of three spike-triggered data analysis te chniques. All of our results are obtained in the set(cid:173) ting of a (possibl y multidimensional) linear-nonlinear (LN) cascade model for stimulus-driven neur al activity. We start by giving exact rate of convergence results for the common spike-triggered average (STA) technique. Next, we analyze a spike-triggered cov ariance method, variants of which have been recently exploited successfully by B ialek, Simoncelli, and colleagues. These first two methods suf(cid:173) fer from extraneous conditions on their convergence; therefore, we introduce an estimato r for the LN model parameters which is de(cid:173) signed to be consistent under general conditions. We provide an algorithm for the computation of this estimat or and derive its rate of convergence. We close with a brief discussion of the e fficiency of these estimators and an application to data recorded from the prima ry motor cortex of awake, behaving primates.
****
Learning to Take Concurrent Actions
Khashayar Rohanimanesh, Sridhar Mahadevan
We investigate a general semi-Markov Decision Process (SMDP) framework for model ing concurrent decision making, where agents learn optimal plans over concurrent temporally extended actions. We introduce three types of parallel termination s chemes { all, any and continue { and theoretically and experimentally compare th

em.
****
Handling Missing Data with Variational Bayesian Learning of ICA
Kwokleung Chan, Te-Won Lee, Terrence J. Sejnowski
Missing data is common in real-world datasets and is a problem for many estimati on techniques. We have developed a variational Bayesian method to perform Indepe ndent Component Analysis (ICA) on high-dimensional data containing missing entri es. Missing data are handled naturally in the Bayesian framework by integrating the generative density model. Mod- eling the distributions of the independent so urces with mixture of Gaus- sians allows sources to be estimated with different kurtosis and skewness. The variational Bayesian method automatically determines the dimen- sionality of the data and yields an accurate density model for the ob - served data without over■tting problems. This allows direct probability estima tion of missing values in the high dimensional space and avoids dimension reduct ion preprocessing which is not feasible with missing data.
****
Ranking with Large Margin Principle: Two Approaches
Amnon Shashua, Anat Levin
We discuss the problem of ranking k instances with the use of a "large margin" principle. We introduce two main approaches: the first is the "fixed margin" po licy in which the margin of the closest neighboring classes is being maximized - which turns out to be a direct generaliza(cid:173) tion of SVM to ranking lear ning. The second approach allows for k - 1 different margins where the sum of m argins is maximized. This approach is shown to reduce to lI-SVM when the number of classes k = 2. Both approaches are optimal in size of 21 where I is the tot al number of training examples. Experiments performed on visual classification and "collab(cid:173) orative filtering" show that both approaches outperform exi sting ordinal regression algorithms applied for ranking and multi-class SVM app lied to general multi-class classification.
****
A Bilinear Model for Sparse Coding
David Grimes, Rajesh P. N. Rao
Recent algorithms for sparse coding and independent component analy- sis (ICA) h ave demonstrated how localized features can be learned from natural images. Howe ver, these approaches do not take image transfor- mations into account. As a res ult, they produce image codes that are redundant because the same feature is lea rned at multiple locations. We describe an algorithm for sparse coding based on a bilinear generative model of images. By explicitly modeling the interaction be tween im- age features and their transformations, the bilinear approach helps re duce redundancy in the image code and provides a basis for transformation- invar iant vision. We present results demonstrating bilinear sparse coding of natural images. We also explore an extension of the model that can capture spatial relat ionships between the independent features of an ob- ject, thereby providing a ne w framework for parts-based object recogni- tion.
****
Data-Dependent Bounds for Bayesian Mixture Methods
Ron Meir, Tong Zhang
We consider Bayesian mixture approaches, where a predictor is constructed by for ming a weighted average of hypotheses from some space of functions. While such p rocedures are known to lead to optimal predictors in several cases, where su–cie ntly accurate prior information is available, it has not been clear how they per form when some of the prior assumptions are violated. In this paper we establish data-dependent bounds for such procedures, extending previous randomized approa ches such as the Gibbs algorithm to a fully Bayesian setting. The ■nite-sample g uarantees established in this work enable the utilization of Bayesian mixture ap proaches in agnostic settings, where the usual assumptions of the Bayesian parad igm fail to hold. Moreover, the bounds derived can be directly applied to non-Ba yesian mixture approaches such as Bagging and Boosting.
****
Information Diffusion Kernels

Guy Lebanon, John Lafferty
A new family of kernels for statistical learning is introduced that ex- ploits t
he geometric structure of statistical models. Based on the heat equation on the
Riemannian manifold de■ned by the Fisher informa- tion metric, information diffu
sion kernels generalize the Gaussian kernel of Euclidean space, and provide a na
tural way of combining generative statistical modeling with non-parametric discr
iminative learning. As a special case, the kernels give a new approach to applyi
ng kernel-based learning algorithms to discrete data. Bounds on covering numbers
 for the new kernels are proved using spectral theory in differential geometry,
and experimental results are presented for text classi■cation.
****
Boosted Dyadic Kernel Discriminants
Baback Moghaddam, Gregory Shakhnarovich
We introduce a novel learning algorithm for binary classi(cid:12)cation with hyp
erplane discriminants based on pairs of training points from opposite classes (d
yadic hypercuts). This algorithm is further extended to nonlinear discriminants
using kernel functions satisfy- ing Mercer’s conditions. An ensemble of simple d
yadic hypercuts is learned incrementally by means of a con(cid:12)dence-rated ve
rsion of Ad- aBoost, which provides a sound strategy for searching through the (
cid:12)nite set of hypercut hypotheses. In experiments with real-world datasets
from the UCI repository, the generalization performance of the hypercut classi(c
id:12)ers was found to be comparable to that of SVMs and k-NN classi(cid:12)ers.
 Furthermore, the computational cost of classi(cid:12)cation (at run time) was f
ound to be similar to, or bet- ter than, that of SVM. Similarly to SVMs, boosted
 dyadic kernel discriminants tend to maximize the margin (via AdaBoost). In cont
rast to SVMs, however, we o(cid:11)er an on-line and incremental learning machin
e for building kernel discriminants whose complex- ity (number of kernel evaluat
ions) can be directly controlled (traded o(cid:11) for accuracy).
****
Derivative Observations in Gaussian Process Models of Dynamic Systems
E. Solak, R. Murray-smith, W. Leithead, D. Leith, Carl Rasmussen
Gaussian processes provide an approach to nonparametric modelling which allows a
 straightforward combination of function and derivative observations in an empir
ical model. This is of particular importance in identi■cation of nonlinear dynam
ic systems from experimental data. 1) It allows us to combine derivative informa
tion, and associated uncertainty with normal function observations into the lear
ning and inference pro- cess. This derivative information can be in the form of
priors speci■ed by an expert or identi■ed from perturbation data close to equili
brium. 2) It allows a seamless fusion of multiple local linear models in a consi
s- tent manner, inferring consistent models and ensuring that integrability cons
traints are met. 3) It improves dramatically the computational ef- ■ciency of Ga
ussian process models for dynamic system identi■cation, by summarising large qua
ntities of near-equilibrium data by a handful of linearisations, reducing the tr
aining set size – traditionally a problem for Gaussian process models.
****
Global Versus Local Methods in Nonlinear Dimensionality Reduction
Vin Silva, Joshua Tenenbaum
Recently proposed algorithms for nonlinear dimensionality reduction fall broadly
 into two categories which have different advantages and disad- vantages: global
 (Isomap [1]), and local (Locally Linear Embedding [2], Laplacian Eigenmaps [3])
. We present two variants of Isomap which combine the advantages of the global a
pproach with what have previ- ously been exclusive advantages of local methods:
computational spar- sity and the ability to invert conformal maps.
****
Learning Graphical Models with Mercer Kernels
Francis Bach, Michael Jordan
We present a class of algorithms for learning the structure of graphical models
from data. The algorithms are based on a measure known as the kernel generalized
 variance (KGV), which essentially allows us to treat all variables on an equal
footing as Gaussians in a feature space obtained from Mercer kernels. Thus we ar

e able to learn hybrid graphs involving discrete and continuous variables of arbitrary type. We explore the computational properties of our approach, showing how to use the kernel trick to compute the relevant statistics in linear time. We illustrate our framework with experiments involving discrete and continuous data.
****
Stochastic Neighbor Embedding
Geoffrey E. Hinton, Sam Roweis
We describe a probabilistic approach to the task of placing objects, de- scribed by high-dimensional vectors or by pairwise dissimilarities, in a low-dimensional space in a way that preserves neighbor identities. A Gaussian is centered on each object in the high-dimensional space and the densities under this Gaussian (or the given dissimilarities) are used to de■ne a probability distribution over all the potential neighbors of the object. The aim of the embedding is to approximate this distribu- tion as well as possible when the same operation is performed on the low-dimensional “images” of the objects. A natural cost function is a sum of Kullback-Leibler divergences, one per object, which leads to a simple gradient for adjusting the positions of the low-dimensional im- ages. Unlike other dimensionality reduction methods, this probabilistic framework makes it easy to represent each object by a mixture of widely separated low-dimensional images. This allows ambiguous objects, like the document count vector for the word “bank”, to have versions close to the images of both “river” and “■nance” without forcing the images of outdoor concepts to be located close to those of corporate concepts.
****
Learning in Spiking Neural Assemblies
David Barber
We consider a statistical framework for learning in a class of net- works of spiking neurons. Our aim is to show how optimal local learning rules can be readily derived once the neural dynamics and desired functionality of the neural assembly have been speci■ed, in contrast to other models which assume (sub-optimal) learning rules. Within this framework we derive local rules for learning tem- poral sequences in a model of spiking neurons and demonstrate its superior performance to correlation (Hebbian) based approaches. We further show how to include mechanisms such as synaptic de- pression and outline how the framework is readily extensible to learning in networks of highly complex spiking neurons. A stochas- tic quantal vesicle release mechanism is considered and implications on the complexity of learning discussed.
****
A Model for Real-Time Computation in Generic Neural Microcircuits
Wolfgang Maass, Thomas Natschläger, Henry Markram
Henry Markram Brain Mind Institute
****
Categorization Under Complexity: A Unified MDL Account of Human Learning of Regular and Irregular Categories
David Fass, Jacob Feldman
We present an account of human concept learning-that is, learning of categories from examples-based on the principle of minimum descrip(cid:173) tion length (MDL). In support of this theory, we tested a wide range of two-dimensional concept types, including both regular (simple) and highly irregular (complex) structures, and found the MDL theory to give a good account of subjects' performance. This suggests that the intrin(cid:173) sic complexity of a concept (that is, its description -length) systematically influences its leamability.
****
Learning with Multiple Labels
Rong Jin, Zoubin Ghahramani
In this paper, we study a special kind of learning problem in which each training instance is given a set of (or distribution over) candidate class labels and only one of the candidate labels is the correct one. Such a problem can occur, e.g., in an information retrieval setting w

here a set of words is associated with an image, or if classes label s are organized hierarchically. We propose a novel discriminative appro ach for handling the ambiguity of class labels in the training examples . The experiments with the proposed approach over five different UCI datase ts show that our approach is able to find the correct label among the se t of candidate labels and actually achieve performance close to the cas e when each training instance is given a single correct label. In co ntrast, naIve methods degrade rapidly as more ambiguity is introduced into t he labels.
****
Branching Law for Axons
Dmitri Chklovskii, Armen Stepanyants
What determines the caliber of axonal branches? We pursue the hypothes is that the axonal caliber has evolved to minimize signal propagation delays, while keeping arbor volume to a minimum. We show that for a gene ral cost function the optimal diameters of mother (do) and daughter (d], d2) branches at a bifurcation obey h a ranc mg aw: e envatIOn re l es on t e fact that the conduction speed scales with the axon diamete r to the power V (v = 1 for myelinated axons and V = 0.5 myelinated axons). We test the branching law on the available experimental data an d find a reasonable agreement.
****
Dyadic Classification Trees via Structural Risk Minimization
Clayton Scott, Robert Nowak
Classi■cation trees are one of the most popular types of classi■ers, with ease o f implementation and interpretation being among their attractive features. Despi te the widespread use of classi■cation trees, theoretical analysis of their perf ormance is scarce. In this paper, we show that a new family of classi■cation tre es, called dyadic classi■cation trees (DCTs), are near optimal (in a minimax sen se) for a very broad range of clas- si■cation problems. This demonstrates that o ther schemes (e.g., neural networks, support vector machines) cannot perform sig ni■cantly better than DCTs in many cases. We also show that this near optimal pe rfor- mance is attained with linear (in the number of training data) complexity growing and pruning algorithms. Moreover, the performance of DCTs on benchmark d atasets compares favorably to that of standard CART, which is generally more com putationally intensive and which does not possess similar near optimality proper ties. Our analysis stems from the- oretical results on structural risk minimizat ion, on which the pruning rule for DCTs is based.
****
PAC-Bayes & Margins
John Langford, John Shawe-Taylor
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Dynamical Causal Learning
David Danks, Thomas Griffiths, Joshua Tenenbaum
theories of human causal
****
Maximally Informative Dimensions: Analyzing Neural Responses to Natural Signals
Tatyana Sharpee, Nicole Rust, William Bialek
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****
Fast Exact Inference with a Factored Model for Natural Language Parsing
Dan Klein, Christopher D. Manning
We present a novel generative model for natural language tree structures in whic

h semantic (lexical dependency) and syntactic (PCFG) structures are scored with separate models. This factorization provides concep- tual simplicity, straightforward opportunities for separately improving the component models, and a level of performance comparable to simi- lar, non-factored models. Most importantly, unlike other modern parsing models, the factored model admits an extremely effective A* parsing al- gorithm, which enables ef■cient, exact inference.
****

Morton-Style Factorial Coding of Color in Primary Visual Cortex
Javier Movellan, Thomas Wachtler, Thomas D. Albright, Terrence Sejnowski
We introduce the notion of Morton-style factorial coding and illustrate how it may help understand information integration and perceptual cod- ing in the brain. We show that by focusing on average responses one may miss the existence of factorial coding mechanisms that become only apparent when analyzing spike count histograms. We show evidence suggesting that the classical/non-classical receptive ■eld organization in the cortex effectively enforces the development of Morton-style factorial codes. This may provide some cues to help understand perceptual cod- ing in the brain and to develop new unsupervised learning algorithms. While methods like ICA (Bell & Sejnowski, 1997) develop independent codes, in Morton-style coding the goal is to make two or more external aspects of the world become independent when conditioning on internal representations.
****

One-Class LP Classifiers for Dissimilarity Representations
Elzbieta Pekalska, David M.J. Tax, Robert Duin
Problems in which abnormal or novel situations should be detected can be approached by describing the domain of the class of typical exam- ples. These applications come from the areas of machine diagnostics, fault detection, illness identi■cation or, in principle, refer to any prob- lem where little knowledge is available outside the typical class. In this paper we explain why proximities are natural representations for domain descriptors and we propose a simple one-class classi■er for dissimilarity representations. By the use of linear programming an ef■cient one-class description can be found, based on a small number of prototype objects. This classi■er can be made (1) more robust by transforming the dissimi- larities and (2) cheaper to compute by using a reduced representation set. Finally, a comparison to a comparable one-class classi■er by Campbell and Bennett is given.
****

Regularized Greedy Importance Sampling
Finnegan Southey, Dale Schuurmans, Ali Ghodsi
Greedy importance sampling is an unbiased estimation technique that re- duces the variance of standard importance sampling by explicitly search- ing for modes in the estimation objective. Previous work has demon- strated the feasibility of implementing this method and proved that the technique is unbiased in both discrete and continuous domains. In this paper we present a reformulation of greedy importance sampling that eliminates the free parameters from the original estimator, and introduces a new regularization strategy that further reduces variance without com- promising unbiasedness. The resulting estimator is shown to be effective for dif■cult estimation problems arising in Markov random ■eld infer- ence. In particular, improvements are achieved over standard MCMC estimators when the distribution has multiple peaked modes.
****

Modeling Midazolam's Effect on the Hippocampus and Recognition Memory
Kenneth Malmberg, René Zeelenberg, Richard Shiffrin
'~1idazolam
****

Fast Kernels for String and Tree Matching
Alex Smola, S.v.n. Vishwanathan
In this paper we present a new algorithm suitable for matching discrete objects such as strings and trees in linear time, thus obviating dynarrtic programming with quadratic time complexity. Furthermore, prediction cost in many cases can be reduced to linear cost in the length of the se(cid:173) quence to be classif

ied, regardless of the number of support vectors. This improvement on the curre
ntly available algorithms makes string kernels a viable alternative for the pra
ctitioner.
****
Expected and Unexpected Uncertainty: ACh and NE in the Neocortex
Peter Dayan, Angela J. Yu
Inference and adaptation in noisy and changing, rich sensory environ- ments are
rife with a variety of speci■c sorts of variability. Experimental and theoretica
l studies suggest that these different forms of variability play different behav
ioral, neural and computational roles, and may be reported by different (notably
 neuromodulatory) systems. Here, we re- ■ne our previous theory of acetylcholine
's role in cortical inference in the (oxymoronic) terms of expected uncertainty,
 and advocate a theory for norepinephrine in terms of unexpected uncertainty. We
 suggest that norepinephrine reports the radical divergence of bottom-up inputs
from prevailing top-down interpretations, to in■uence inference and plasticity.
We illustrate this proposal using an adaptive factor analysis model.
****
Feature Selection and Classification on Matrix Data: From Large Margins to Small
 Covering Numbers
Sepp Hochreiter, Klaus Obermayer
We investigate the problem of learning a classi■cation task for datasets which a
re described by matrices. Rows and columns of these matrices correspond to objec
ts, where row and column ob- jects may belong to di■erent sets, and the entries
in the matrix express the relationships between them. We interpret the matrix el
- ements as being produced by an unknown kernel which operates on object pairs a
nd we show that - under mild assumptions - these ker- nels correspond to dot pro
ducts in some (unknown) feature space. Minimizing a bound for the generalization
 error of a linear classi- ■er which has been obtained using covering numbers we
 derive an objective function for model selection according to the principle of
structural risk minimization. The new objective function has the advantage that
it allows the analysis of matrices which are not pos- itive de■nite, and not eve
n symmetric or square. We then consider the case that row objects are interprete
d as features. We suggest an additional constraint, which imposes sparseness on
the row objects and show, that the method can then be used for feature selection
. Finally, we apply this method to data obtained from DNA microar- rays, where \
column" objects correspond to samples, \row" objects correspond to genes and mat
rix elements correspond to expression levels. Benchmarks are conducted using sta
ndard one-gene classi■- cation and support vector machines and K-nearest neighbo
rs after standard feature selection. Our new method extracts a sparse set of gen
es and provides superior classi■cation results.
****
Learning Semantic Similarity
Jaz Kandola, Nello Cristianini, John Shawe-taylor
The standard representation of text documents as bags of words suffers from wel
l known limitations, mostly due to its inability to exploit semantic similarity
 between terms. Attempts to incorpo(cid:173) rate some notion of term similarity
 include latent semantic index(cid:173) ing [8], the use of semantic networks [9
], and probabilistic methods [5]. In this paper we propose two methods for infe
rring such sim(cid:173) ilarity from a corpus. The first one defines word-simila
rity based on document-similarity and viceversa, giving rise to a system of eq
uations whose equilibrium point we use to obtain a semantic similarity measure.
 The second method models semantic relations by means of a diffusion process on
 a graph defined by lexicon and co-occurrence information. Both approaches prod
uce valid kernel functions parametrised by a real number. The paper shows how
the alignment measure can be used to successfully perform model selection over
this parameter. Combined with the use of support vector machines we obtain posi
tive results.
****
How Linear are Auditory Cortical Responses?
Maneesh Sahani, Jennifer Linden

By comparison to some other sensory cortices, the functional proper- ties of cells in the primary auditory cortex are not yet well understood. Recent attempts to obtain a generalized description of auditory cortical responses have often relied upon characterization of the spectrotempo- ral receptive ■eld (STRF), which amounts to a model of the stimulus- response function (SRF) that is linear in the spectrogram of the stimulus. How well can such a model account for neural responses at the very ■rst stages of auditory cortical processing? To answer this question, we de- velop a novel methodology for evaluating the fraction of stimulus-related response power in a population that can be captured by a given type of SRF model. We use this technique to show that, in the thalamo-recipient layers of primary auditory cortex, STRF models account for no more than 40% of the stimulus-related power in neural responses.
****
Incremental Gaussian Processes
Joaquin Quiñonero Candela, Ole Winther
In this paper, we consider Tipping’s relevance vector machine (RVM) [1] and formalize an incremental training strategy as a variant of the expectation-maximization (EM) algorithm that we call Subspace EM (SSEM). Working with a subset of active basis functions, the sparsity of the RVM solution will ensure that the number of basis functions and thereby the computational complexity is kept low. We also introduce a mean ■eld approach to the intractable classi■cation model that is ex- pected to give a very good approximation to exact Bayesian inference and contains the Laplace approximation as a special case. We test the algorithms on two large data sets with O(103 (cid:0) 104) examples. The re- sults indicate that Bayesian learning of large data sets, e.g. the MNIST database is realistic.
****
Dynamic Bayesian Networks with Deterministic Latent Tables
David Barber
The application of latent/hidden variable Dynamic Bayesian Net- works is constrained by the complexity of marginalising over latent variables. For this reason either small latent dimensions or Gaus- sian latent conditional tables linearly dependent on past states are typically considered in order that inference is tractable. We suggest an alternative approach in which the latent variables are modelled using deterministic conditional probability tables. This specialisa- tion has the advantage of tractable inference even for highly com- plex non-linear/non-Gaussian visible conditional probability tables. This approach enables the consideration of highly complex latent dynamics whilst retaining the bene(cid:12)ts of a tractable probabilistic model.
****
Generalized^2 Linear^2 Models
Geoffrey J. Gordon
We introduce the Generalized2 Linear2 Model, a statistical estima(cid:173) tor which combines features of nonlinear regression and factor anal(cid:173) ysis. A (GL)2M approximately decomposes a rectangular matrix X into a simpler representation j(g(A)h(B)). Here A and Bare low-rank matrices, while j, g, and h are link functions. (GL)2Ms include many useful models as special cases, including principal components analysis, exponential-family peA, the infomax formu(cid:173) lation of independent components analysis, linear regression, and generalized linear models. They also include new and interesting special cases, one of which we describe below. We also present an iterative procedure which optimizes the parameters of a (GL)2M. This procedure reduces to well-known algorithms for some of the special cases listed above; for other special cases, it is new.
****
Spectro-Temporal Receptive Fields of Subthreshold Responses in Auditory Cortex
Christian K. Machens, Michael Wehr, Anthony Zador
How do cortical neurons represent the acoustic environment? This ques- tion is often addressed by probing with simple stimuli such as clicks or tone pips. Such stimuli have the advantage of yielding easily interpreted answers, but have the disadvantage that they may fail to uncover complex or higher-order neuronal resp

onse properties. Here we adopt an alternative approach, probing neuronal respons
es with complex acoustic stimuli, including animal vocalizations and music. We h
ave used in vivo whole cell methods in the rat auditory cortex to record subthre
shold membrane potential ■uctuations elicited by these stimuli. Whole cell recor
ding reveals the total synaptic input to a neuron from all the other neurons in
the circuit, instead of just its output—a sparse bi- nary spike train—as in conv
entional single unit physiological recordings. Whole cell recording thus provide
s a much richer source of information about the neuron's response. Many neurons
responded robustly and reliably to the complex stimuli in our ensemble. Here we
analyze the linear component—the spectro- temporal receptive ■eld (STRF)—of the
transformation from the sound (as represented by its time-varying spectrogram) t
o the neuron's mem- brane potential. We ■nd that the STRF has a rich dynamical s
tructure, including excitatory regions positioned in general accord with the pre
dic- tion of the simple tuning curve. We also ■nd that in many cases, much of th
e neuron's response, although deterministically related to the stimulus, cannot
be predicted by the linear component, indicating the presence of as-yet-uncharac
terized nonlinear response properties.
****
Bayesian Models of Inductive Generalization
Neville Sanjana, Joshua Tenenbaum
We argue that human inductive generalization is best explained in a Bayesian fra
mework, rather than by traditional models based on simi- larity computations. We
 go beyond previous work on Bayesian concept learning by introducing an unsuperv
ised method for constructing ■ex- ible hypothesis spaces, and we propose a versi
on of the Bayesian Oc- cam's razor that trades off priors and likelihoods to pre
vent under- or over-generalization in these ■exible spaces. We analyze two publi
shed data sets on inductive reasoning as well as the results of a new behavioral
 study that we have carried out.
****
Dynamical Constraints on Computing with Spike Timing in the Cortex
Arunava Banerjee, Alexandre Pouget
If the cortex uses spike timing to compute, the timing of the spikes must
 be robust to perturbations. Based on a recent framework that provides
 a simple criterion to determine whether a spike sequence produced by a
 generic network is sensitive to initial conditions, and numerical simulatio
ns of a variety of network architectures, we argue within the limits
 set by our model of the neuron, that it is unlikely that precise se
quences of spike timings are used for computation under conditions typica
lly found in the cortex.
****
Bayesian Estimation of Time-Frequency Coefficients for Audio Signal Enhancement
Patrick Wolfe, Simon Godsill
The Bayesian paradigm provides a natural and effective means of exploit- ing pri
or knowledge concerning the time-frequency structure of sound signals such as sp
eech and music—something which has often been over- looked in traditional audio
signal processing approaches. Here, after con- structing a Bayesian model and pr
ior distributions capable of taking into account the time-frequency characterist
ics of typical audio waveforms, we apply Markov chain Monte Carlo methods in ord
er to sample from the resultant posterior distribution of interest. We present s
peech enhance- ment results which compare favourably in objective terms with sta
ndard time-varying ■ltering techniques (and in several cases yield superior per-
 formance, both objectively and subjectively); moreover, in contrast to such met
hods, our results are obtained without an assumption of prior knowledge of the n
oise power.
****
Real-Time Monitoring of Complex Industrial Processes with Particle Filters
Rubén Morales-Menéndez, Nando de Freitas, David Poole
This paper discusses the application of particle ■ltering algorithms to fault di
agnosis in complex industrial processes. We consider two ubiq- uitous processes:
 an industrial dryer and a level tank. For these appli- cations, we compared thr

ee particle ■ltering variants: standard parti- cle ■ltering, Rao-Blackwellised particle ■ltering and a version of Rao- Blackwellised particle ■ltering that does one-step look-ahead to select good sampling regions. We show that the overhead of the extra process- ing per particle of the more sophisticated methods is more than compen- sated by the decrease in error and variance.
****

Retinal Processing Emulation in a Programmable 2-Layer Analog Array Processor CMOS Chip

R. Carmona, F. Jiménez-garrido, R. Dominguez-castro, S. Espejo, A. Rodriguez-vázquez

A bio-inspired model for an analog programmable array processor (APAP), based on studies on the vertebrate retina, has permitted the realization of complex programmable spatio-temporal dynam- ics in VLSI. This model mimics the way in which images are pro- cessed in the visual pathway, rendering a feasible alternative for the implementation of early vision applications in standard tech- nologies. A prototype chip has been designed and fabricated in a 0.5(cid:22)m standard CMOS process. Computing power per area and power consumption is amongst the highest reported for a single chip. Design challenges, trade-o(cid:11)s and some experimental results are presented in this paper.
****

Bayesian Image Super-Resolution

Michael Tipping, Christopher Bishop

The extraction of a single high-quality image from a set of low(cid:173) resolution images is an important problem which arises in fields such as remote sensing, surveillance, medical imaging and the ex(cid:173) traction of still images from video. Typical approaches are based on the use of cross-correlation to register the images followed by the inversion of the transformation from the unknown high reso(cid:173) lution image to the observed low resolution images, using regular(cid:173) ization to resolve the ill-posed nature of the inversion process. In this paper we develop a Bayesian treatment of the super-resolution problem in which the likelihood function for the image registra(cid:173) tion parameters is based on a marginalization over the unknown high-resolution image. This approach allows us to estimate the unknown point spread function, and is rendered tractable through the introduction of a Gaussian process prior over images. Results indicate a significant improvement over techniques based on MAP (maximum a-posteriori) point optimization of the high resolution image and associated registration parameters.
****

Charting a Manifold

Matthew Brand

We construct a nonlinear mapping from a high-dimensional sample space to a low-dimensional vector space, effectively recovering a Cartesian coordinate system for the manifold from which the data is sampled. The mapping preserves local geometric relations in the manifold and is pseudo-invertible. We show how to estimate the intrinsic dimensionality of the manifold from samples, decompose the sample data into locally linear low-dimensional patches, merge these patches into a single low- dimensional coordinate system, and compute forward and reverse map- pings between the sample and coordinate spaces. The objective functions are convex and their solutions are given in closed form.
****

A Minimal Intervention Principle for Coordinated Movement

Emanuel Todorov, Michael Jordan

Behavioral goals are achieved reliably and repeatedly with movements rarely reproducible in their detail. Here we offer an explanation: we show that not only are variability and goal achievement compatible, but indeed that allowing variability in redundant dimensions is the optimal control strategy in the face of uncertainty. The optimal feedback control laws for typical motor tasks obey a “minimal intervention” principle: deviations from the average trajectory are only corrected when they interfere with the task goals. The resulting behavior exhibits ta

sk-constrained variabil- ity, as well as synergetic coupling among actuators—whi
ch is another unexplained empirical phenomenon.
****
A Probabilistic Approach to Single Channel Blind Signal Separation
Gil-jin Jang, Te-Won Lee
We present a new technique for achieving source separation when given only a sin
gle channel recording. The main idea is based on exploiting the inherent time st
ructure of sound sources by learning a priori sets of basis ■lters in time domai
n that encode the sources in a statistically ef■cient manner. We derive a learni
ng algorithm using a maximum likelihood approach given the observed single chann
el data and sets of basis ■lters. For each time point we infer the source signal
s and their contribution factors. This inference is possible due to the prior kn
owledge of the basis ■lters and the associated coef■cient densities. A ■exible m
odel for density estimation allows accurate modeling of the observation and our
experimental results exhibit a high level of separation performance for mixtures
of two music signals as well as the separation of two voice signals.
****
A Digital Antennal Lobe for Pattern Equalization: Analysis and Design
Alex Holub, Gilles Laurent, Pietro Perona
Re-mapping patterns in order to equalize their distribution may greatly
simplify both the structure and the training of classifiers. Here, the
properties of one such map obtained by running a few steps of discrete
-time dynamical system are explored. The system is called 'Digital Antennal
Lobe' (DAL) because it is inspired by recent studies of the antennallob
e, a structure in the olfactory sys(cid:173) tem of the grasshopper. The pat
tern-spreading properties of the DAL as well as its average behavior as
a function of its (few) de(cid:173) sign parameters are analyzed by exten
ding previous results of Van Vreeswijk and Sompolinsky. Furthermore, a techniq
ue for adapting the parameters of the initial design in order to obtain
opportune noise-rejection behavior is suggested. Our results are demonstrate
d with a number of simulations.
****
Knowledge-Based Support Vector Machine Classifiers
Glenn Fung, Olvi Mangasarian, Jude Shavlik
Prior knowledge in the form of multiple polyhedral sets, each be(cid:173)
longing to one of two categories, is introduced into a reformulation of a lin
ear support vector machine classifier. The resulting formu(cid:173) lation lead
s to a linear program that can be solved efficiently. Real world examples, fro
m DNA sequencing and breast cancer prognosis, demonstrate the effectiveness of
the proposed method. Numerical results show improvement in test set acc
uracy after the incorpo(cid:173) ration of prior knowledge into ordinary, dat
a-based linear support vector machine classifiers. One experiment also sh
ows that a lin(cid:173) ear classifier, based solely on prior knowledge, fa
r outperforms the direct application of prior knowledge rules to classify d
ata. Keywords: use and refinement of prior knowledge, sup(cid:173) port
vector machines, linear programming
****
Conditional Models on the Ranking Poset
Guy Lebanon, John Lafferty
A distance-based conditional model on the ranking poset is presented for use in
classi■cation and ranking. The model is an extension of the Mallows model, and
generalizes the classi■er combination methods used by several ensemble learning
algorithms, including error correcting output codes, discrete AdaBoost, logistic
regression and cranking. The algebraic structure of the ranking poset leads to
a simple Bayesian inter- pretation of the conditional model and its special case
s. In addition to a unifying view, the framework suggests a probabilistic interp
retation for error correcting output codes and an extension beyond the binary co
ding scheme.
****
A Formulation for Minimax Probability Machine Regression

Thomas Strohmann, Gregory Grudic
We formulate the regression problem as one of maximizing the mini- mum probabili
ty, symbolized by (cid:10), that future predicted outputs of the regression mode
l will be within some (cid:6)" bound of the true regression function. Our formul
ation is unique in that we obtain a direct estimate of this lower probability bo
und (cid:10). The proposed framework, minimax probability machine regression (MP
MR), is based on the recently de- scribed minimax probability machine classi■cat
ion algorithm [Lanckriet et al.] and uses Mercer Kernels to obtain nonlinear reg
ression models. MPMR is tested on both toy and real world data, verifying the ac
curacy of the (cid:10) bound, and the ef■cacy of the regression models.
****
Multiclass Learning by Probabilistic Embeddings
Ofer Dekel, Yoram Singer
We describe a new algorithmic framework for learning multiclass catego- rization
 problems. In this framework a multiclass predictor is composed of a pair of emb
eddings that map both instances and labels into a common space. In this space ea
ch instance is assigned the label it is nearest to. We outline and analyze an al
gorithm, termed Bunching, for learning the pair of embeddings from labeled data.
 A key construction in the analysis of the algorithm is the notion of probabilis
tic output codes, a generaliza- tion of error correcting output codes (ECOC). Fu
rthermore, the method of multiclass categorization using ECOC is shown to be an
instance of Bunching. We demonstrate the advantage of Bunching over ECOC by comp
aring their performance on numerous categorization problems.
****
Improving a Page Classifier with Anchor Extraction and Link Analysis
William W. Cohen
Most text categorization systems use simple models of documents and document col
lections. In this paper we describe a technique that im- proves a simple web pag
e classi■er's performance on pages from a new, unseen web site, by exploiting li
nk structure within a site as well as page structure within hub pages. On real-w
orld test cases, this technique signi■cantly and substantially improves the accu
racy of a bag-of-words classi■er, reducing error rate by about half, on average.
 The system uses a variant of co-training to exploit unlabeled data from a new s
ite. Pages are labeled using the base classi■er; the results are used by a restr
icted wrapper-learner to propose potential “main-category anchor wrappers”; and
■nally, these wrappers are used as features by a third learner to ■nd a categori
zation of the site that implies a simple hub structure, but which also largely a
grees with the original bag-of-words classi■er.
****
VIBES: A Variational Inference Engine for Bayesian Networks
Christopher Bishop, David Spiegelhalter, John Winn
In recent years variational methods have become a popular tool for approximate i
nference and learning in a wide variety of proba- bilistic models. For each new
application, however, it is currently necessary (cid:12)rst to derive the variat
ional update equations, and then to implement them in application-speci(cid:12)c
 code. Each of these steps is both time consuming and error prone. In this paper
 we describe a general purpose inference engine called VIBES (‘Variational Infer
- ence for Bayesian Networks’) which allows a wide variety of proba- bilistic mo
dels to be implemented and solved variationally without recourse to coding. New
models are speci(cid:12)ed either through a simple script or via a graphical int
erface analogous to a drawing package. VIBES then automatically generates and so
lves the vari- ational equations. We illustrate the power and (cid:13)exibility
of VIBES using examples from Bayesian mixture modelling.
****
A Convergent Form of Approximate Policy Iteration
Theodore Perkins, Doina Precup
We study a new, model-free form of approximate policy iteration which uses Sarsa
 updates with linear state-action value function approximation for policy evalua
tion, and a “policy improvement operator” to generate a new policy based on the
learned state-action values. We prove that if the policy improvement operator pr

oduces -soft policies and is Lipschitz continuous in the action values, with a constant that is not too large, then the approximate policy iteration algorithm converges to a unique solu- tion from any initial policy. To our knowledge, this is the ■rst conver- gence result for any form of approximate policy iteration under similar computational-resource assumptions.
****
Mean Field Approach to a Probabilistic Model in Information Retrieval
Bin Wu, K. Wong, David Bodoff
We study an explicit parametric model of documents, queries, and rel- evancy assessment for Information Retrieval (IR). Mean-■eld methods are applied to analyze the model and derive ef■cient practical algorithms to estimate the parameters in the problem. The hyperparameters are es- timated by a fast approximate leave-one-out cross-validation procedure based on the cavity method. The algorithm is further evaluated on several benchmark databases by comparing with standard algorithms in IR.
****
Exponential Family PCA for Belief Compression in POMDPs
Nicholas Roy, Geoffrey J. Gordon
Geoffrey Gordon
****
Dynamic Structure Super-Resolution
Amos J. Storkey
The problem of super-resolution involves generating feasible higher resolution images, which are pleasing to the eye and realistic, from a given low resolution image. This might be attempted by us- ing simple (cid:12)lters for smoothing out the high resolution blocks or through applications where substantial prior information is used to imply the textures and shapes which will occur in the images. In this paper we describe an approach which lies between the two extremes. It is a generic unsupervised method which is usable in all domains, but goes beyond simple smoothing methods in what it achieves. We use a dynamic tree-like architecture to model the high resolution data. Approximate conditioning on the low resolution image is achieved through a mean (cid:12)eld approach.
****
Independent Components Analysis through Product Density Estimation
Trevor Hastie, Rob Tibshirani
We present a simple direct approach for solving the ICA problem, using density estimation and maximum likelihood. Given a candi(cid:173) date orthogonal frame, we model each of the coordinates using a semi-parametric density estimate based on cubic splines. Since our estimates have two continuous derivatives, we can easily run a sec(cid:173) ond order search for the frame parameters. Our method performs very favorably when compared to state-of-the-art techniques.
****
Discriminative Densities from Maximum Contrast Estimation
Peter Meinicke, Thorsten Twellmann, Helge Ritter
We propose a framework for classi■er design based on discriminative densities for representation of the differences of the class-conditional dis- tributions in a way that is optimal for classi■cation. The densities are selected from a parametrized set by constrained maximization of some objective function which measures the average (bounded) difference, i.e. the contrast between discriminative densities. We show that maximiza- tion of the contrast is equivalent to minimization of an approximation of the Bayes risk. Therefore using suitable classes of probability den- sity functions, the resulting maximum contrast classi■ers (MCCs) can approximate the Bayes rule for the general multiclass case. In particular for a certain parametrization of the density functions we obtain MCCs which have the same functional form as the well-known Support Vec- tor Machines (SVMs). We show that MCC-training in general requires some nonlinear optimization but under certain conditions the problem is concave and can be tackled by a single linear program. We indicate the close relation between SVM- and MCC-training and in particular we show that Linear Programming Machines can be viewed as an approxi- mat

e realization of MCCs. In the experiments on benchmark data sets, the MCC shows a competitive classi■cation performance.
****
Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterio
r
Patrik Hoyer, Aapo Hyvärinen
The responses of cortical sensory neurons are notoriously variable, with the num
ber of spikes evoked by identical stimuli varying signi■cantly from trial to tri
al. This variability is most often interpreted as ‘noise’, purely detrimental to
 the sensory system. In this paper, we propose an al- ternative view in which th
e variability is related to the uncertainty, about world parameters, which is in
herent in the sensory stimulus. Speci■- cally, the responses of a population of
neurons are interpreted as stochas- tic samples from the posterior distribution
in a latent variable model. In addition to giving theoretical arguments supporti
ng such a representa- tional scheme, we provide simulations suggesting how some
aspects of response variability might be understood in this framework.
****
Combining Features for BCI
Guido Dornhege, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Müller
Recently, interest is growing to develop an effective communication in- terface
connecting the human brain to a computer, the ’Brain-Computer Interface’ (BCI).
One motivation of BCI research is to provide a new communication channel substit
uting normal motor output in patients with severe neuromuscular disabilities. In
 the last decade, various neuro- physiological cortical processes, such as slow
potential shifts, movement related potentials (MRPs) or event-related desynchron
ization (ERD) of spontaneous EEG rhythms, were shown to be suitable for BCI, and
, con- sequently, different independent approaches of extracting BCI-relevant EE
G-features for single-trial analysis are under investigation. Here, we present a
nd systematically compare several concepts for combining such EEG-features to im
prove the single-trial classi■cation. Feature combi- nations are evaluated on mo
vement imagination experiments with 3 sub- jects where EEG-features are based on
 either MRPs or ERD, or both. Those combination methods that incorporate the ass
umption that the sin- gle EEG-features are physiologically mutually independent
outperform the plain method of ’adding’ evidence where the single-feature vector
s are simply concatenated. These results strengthen the hypothesis that MRP and
ERD re■ect at least partially independent aspects of cortical processes and open
 a new perspective to boost BCI effectiveness.
****
A Probabilistic Model for Learning Concatenative Morphology
Matthew Snover, Michael Brent
This paper describes a system for the unsupervised learning of morpho- logical s
uf■xes and stems from word lists. The system is composed of a generative probabi
lity model and hill-climbing and directed search algo- rithms. By extracting and
 examining morphologically rich subsets of an input lexicon, the directed search
 identi■es highly productive paradigms. The hill-climbing algorithm then further
 maximizes the probability of the hypothesis. Quantitative results are shown by
measuring the accuracy of the morphological relations identi■ed. Experiments in
English and Pol- ish, as well as comparisons with another recent unsupervised mo
rphol- ogy learning algorithm demonstrate the effectiveness of this technique.
****
The Decision List Machine
Marina Sokolova, Mario Marchand, Nathalie Japkowicz, John Shawe-taylor
We introduce a new learning algorithm for decision lists to allow features that
are constructed from the data and to allow a trade- o■ between accuracy and comp
lexity. We bound its generalization error in terms of the number of errors and t
he size of the classi■er it ■nds on the training data. We also compare its perfo
rmance on some natural data sets with the set covering machine and the support v
ector machine.
****
How to Combine Color and Shape Information for 3D Object Recognition: Kernels do

the Trick
B. Caputo, Gy. Dorkó
This paper presents a kernel method that allows to combine color and shape in
formation for appearance-based object recognition. It doesn't require to defi
ne a new common representation, but use the power of kernels to combine differe
nt representations together in an effective manner. These results are achieved
 using results of statis(cid:173) tical mechanics of spin glasses combined with
Markov random fields via kernel functions. Experiments show an increase in re
cognition rate up to 5.92% with respect to conventional strategies.
****
Reconstructing Stimulus-Driven Neural Networks from Spike Times
Duane Nykamp
We present a method to distinguish direct connections between two neu- rons from
 common input originating from other, unmeasured neurons. The distinction is com
puted from the spike times of the two neurons in response to a white noise stimu
lus. Although the method is based on a highly idealized linear-nonlinear approxi
mation of neural response, we demonstrate via simulation that the approach can w
ork with a more re- alistic, integrate-and-■re neuron model. We propose that the
 approach exempli■ed by this analysis may yield viable tools for reconstructing
stimulus-driven neural networks from data gathered in neurophysiology experiment
s.
****
Selectivity and Metaplasticity in a Unified Calcium-Dependent Model
Luk Chong Yeung, Brian Blais, Leon Cooper, Harel Shouval
A uni(cid:12)ed, biophysically motivated Calcium-Dependent Learning model has be
en shown to account for various rate-based and spike time-dependent paradigms fo
r inducing synaptic plasticity. Here, we investigate the properties of this mode
l for a multi-synapse neuron that receives inputs with di(cid:11)erent spike-tra
in statistics. In addition, we present a physiological form of metaplasticity, a
n activity-driven regulation mechanism, that is essential for the ro- bustness o
f the model. A neuron thus implemented develops stable and selective receptive (
cid:12)elds, given various input statistics
****
Learning to Detect Natural Image Boundaries Using Brightness and Texture
David Martin, Charless Fowlkes, Jitendra Malik
The goal of this work is to accurately detect and localize boundaries in natural
 scenes using local image measurements. We formulate features that respond to ch
aracteristic changes in brightness and texture associated with natural boundarie
s. In order to combine the information from these features in an optimal way, a
classi■er is trained using human labeled images as ground truth. We present prec
ision-recall curves showing that the resulting detector outperforms existing app
roaches.
****
Concurrent Object Recognition and Segmentation by Graph Partitioning
Stella X. Yu, Ralph Gross, Jianbo Shi
Segmentation and recognition have long been treated as two separate pro(cid:173)
 cesses. We propose a mechanism based on spectral graph partitioning th
at readily combine the two processes into one. A part-based recogni(cid:173) t
ion system detects object patches, supplies their partial segmentations as well
 as knowledge about the spatial configurations of the object. The goal of patc
h grouping is to find a set of patches that conform best to the object configur
ation, while the goal of pixel grouping is to find a set of pixels that ha
ve the best low-level feature similarity. Through pixel-patch in(cid:173)
teractions and between-patch competition encoded in the solution space, these t
wo processes are realized in one joint optimization problem. The globally opt
imal partition is obtained by solving a constrained eigenvalue problem. We de
monstrate that the resulting object segmentation elimi(cid:173) nates false p
ositives for the part detection, while overcoming occlusion and weak contours
 for the low-level edge detection.
****

Approximate Linear Programming for Average-Cost Dynamic Programming
Benjamin Roy, Daniela Farias
This paper extends our earlier analysis on approximate linear program- ming as a
n approach to approximating the cost-to-go function in a discounted-cost dynamic
 program [6]. In this paper, we consider the average-cost criterion and a versio
n of approximate linear programming that generates approximations to the optimal
 average cost and differential cost function. We demonstrate that a naive versio
n of approximate linear programming prioritizes approximation of the optimal ave
rage cost and that this may not be well-aligned with the objective of deriving a
 policy with low average cost. For that, the algorithm should aim at producing a
 good approximation of the differential cost function. We propose a two- phase v
ariant of approximate linear programming that allows for external control of the
 relative accuracy of the approximation of the differential cost function over d
ifferent portions of the state space via state-relevance weights. Performance bo
unds suggest that the new algorithm is compat- ible with the objective of optimi
zing performance and provide guidance on appropriate choices for state-relevance
 weights.
****
Learning Sparse Topographic Representations with Products of Student-t Distribut
ions
Max Welling, Simon Osindero, Geoffrey E. Hinton
We propose a model for natural images in which the probability of an im- age is
proportional to the product of the probabilities of some ■lter out- puts. We enc
ourage the system to ■nd sparse features by using a Student- t distribution to m
odel each ■lter output. If the t-distribution is used to model the combined outp
uts of sets of neurally adjacent ■lters, the sys- tem learns a topographic map i
n which the orientation, spatial frequency and location of the ■lters change smo
othly across the map. Even though maximum likelihood learning is intractable in
our model, the product form allows a relatively ef■cient learning procedure that
 works well even for highly overcomplete sets of ■lters. Once the model has been
 learned it can be used as a prior to derive the “iterated Wiener ■lter” for the
 pur- pose of denoising images.
****
Margin Analysis of the LVQ Algorithm
Koby Crammer, Ran Gilad-bachrach, Amir Navot, Naftali Tishby
Prototypes based algorithms are commonly used to reduce the computa- tional comp
lexity of Nearest-Neighbour (NN) classi■ers. In this paper we discuss theoretica
l and algorithmical aspects of such algorithms. On the theory side, we present m
argin based generalization bounds that sug- gest that these kinds of classi■ers
can be more accurate then the 1-NN rule. Furthermore, we derived a training algo
rithm that selects a good set of prototypes using large margin principles. We al
so show that the 20 years old Learning Vector Quantization (LVQ) algorithm emerg
es natu- rally from our framework.
****
Feature Selection by Maximum Marginal Diversity
Nuno Vasconcelos
We address the question of feature selection in the context of visual recognitio
n. It is shown that, besides ef■cient from a computational standpoint, the infom
ax principle is nearly optimal in the minimum Bayes error sense. The concept of
marginal diversity is introduced, lead- ing to a generic principle for feature s
election (the principle of maximum marginal diversity) of extreme computational
simplicity. The relation- ships between infomax and the maximization of marginal
 diversity are identi■ed, uncovering the existence of a family of classi■cation
proce- dures for which near optimal (in the Bayes error sense) feature selection
 does not require combinatorial search. Examination of this family in light of r
ecent studies on the statistics of natural images suggests that visual recogniti
on problems are a subset of it.
****
A Neural Edge-Detection Model for Enhanced Auditory Sensitivity in Modulated Noi
se

Alon Fishbach, Bradford May
Psychophysical data suggest that temporal modulations of stimulus amplitude env elopes play a prominent role in the perceptual segregation of concurrent sounds . In particular, the detection of an unmodulated signal can be significantly im proved by adding amplitude modulation to the spectral envelope of a competing masking noise. This perceptual phenomenon is known as “Comodulation Masking Re lease” (CMR). Despite the obvious influence of temporal structure on the percep tion of complex auditory scenes, the physiological mechanisms that contribute t o CMR and auditory streaming are not well known. A recent physiological study by Nelken and colleagues has demonstrated an enhanced cortical representation o f auditory signals in modulated noise. Our study evaluates these CMR-like respo nse patterns from the perspective of a hypothetical auditory edge-detection neu ron. It is shown that this simple neural model for the detection of amplitude transients can reproduce not only the physiological data of Nelken et al., but also, in light of previous results, a variety of physiological and psychoacoust ical phenomena that are related to the perceptual segregation of concurrent sou nds.
****
Concentration Inequalities for the Missing Mass and for Histogram Rule Error
Luis E. Ortiz, David McAllester
This paper gives distribution-free concentration inequalities for the miss- ing mass and the error rate of histogram rules. Negative association meth- ods can b e used to reduce these concentration problems to concentration questions about i ndependent sums. Although the sums are independent, they are highly heterogeneou s. Such highly heterogeneous independent sums cannot be analyzed using standard concentration inequalities such as Hoeffding’s inequality, the Angluin-Valiant b ound, Bernstein’s in- equality, Bennett’s inequality, or McDiarmid’s theorem.
****
Learning Sparse Multiscale Image Representations
Phil Sallee, Bruno Olshausen
We describe a method for learning sparse multiscale image repre- sentations usin g a sparse prior distribution over the basis function coe(cid:14)cients. The pri or consists of a mixture of a Gaussian and a Dirac delta function, and thus enco urages coe(cid:14)cients to have exact zero values. Coe(cid:14)cients for an ima ge are computed by sampling from the resulting posterior distribution with a Gib bs sampler. The learned basis is similar to the Steerable Pyramid basis, and yie lds slightly higher SNR for the same number of active coe(cid:14)cients. De- noi sing using the learned image model is demonstrated for some standard test images , with results that compare favorably with other denoising methods.
****
The Effect of Singularities in a Learning Machine when the True Parameters Do No t Lie on such Singularities
Sumio Watanabe, Shun-ichi Amari
A lot of learning machines with hidden variables used in infor- mation science h ave singularities in their parameter spaces. At singularities, the Fisher inform ation matrix becomes degenerate, resulting that the learning theory of regular s tatistical models does not hold. Recently, it was proven that, if the true param eter is contained in singularities, then the coe(cid:14)cient of the Bayes gen- eralization error is equal to the pole of the zeta function of the Kullback info rmation. In this paper, under the condition that the true parameter is almost bu t not contained in singularities, we show two results. (1) If the dimension of t he parameter from in- puts to hidden units is not larger than three, then there exits a region of true parameters where the generalization error is larger than those of regular models, however, if otherwise, then for any true parameter, the generalization error is smaller than those of regular models. (2) The symmetry of the generalization error and the training error does not hold in singular mod els in general.
****
Learning About Multiple Objects in Images: Factorial Learning without Factorial Search

Christopher K. I. Williams, Michalis Titsias
We consider data which are images containing views of multiple objects. Our task is to learn about each of the objects present in the images. This task can be approached as a factorial learning problem, where each image must be explained by instantiating a model for each of the objects present with the correct instantiation parameters. A major problem with learning a factorial model is that as the number of objects increases, there is a combinatorial explosion of the number of con■gurations that need to be considered. We develop a method to extract object models sequentially from the data by making use of a robust statistical method, thus avoid- ing the combinatorial explosion, and present results showing successful extraction of objects from real images.
****
Distance Metric Learning with Application to Clustering with Side-Information
Eric Xing, Michael Jordan, Stuart J. Russell, Andrew Ng
@cs.berkeley.edu
****
Location Estimation with a Differential Update Network
Ali Rahimi, Trevor Darrell
Given a set of hidden variables with an a-priori Markov structure, we derive an online algorithm which approximately updates the posterior as pairwise measurements between the hidden variables become available. The update is performed using Assumed Density Filtering: to incorporate each pairwise measurement, we compute the optimal Markov structure which represents the true posterior and use it as a prior for incorporating the next measurement. We demonstrate the resulting algorithm by cal- culating globally consistent trajectories of a robot as it navigates along a 2D trajectory. To update a trajectory of length t, the update takes O(t). When all conditional distributions are linear-Gaussian, the algorithm can be thought of as a Kalman Filter which simpli■es the state covariance matrix after incorporating each measurement.
****
Hyperkernels
Cheng Ong, Robert C. Williamson, Alex Smola
We consider the problem of choosing a kernel suitable for estimation using a Gaussian Process estimator or a Support Vector Machine. A novel solution is presented which involves de■ning a Reproducing Ker- nel Hilbert Space on the space of kernels itself. By utilizing an analog of the classical representer theorem, the problem of choosing a kernel from a parameterized family of kernels (e.g. of varying width) is reduced to a statistical estimation problem akin to the problem of minimizing a regularized risk functional. Various classical settings for model or kernel selection are special cases of our framework.
****
Prediction and Semantic Association
Thomas Griffiths, Mark Steyvers
We explore the consequences of viewing semantic association as the result of attempting to predict the concepts likely to arise in a particular context. We argue that the success of existing accounts of semantic representation comes as a result of indirectly addressing this problem, and show that a closer correspondence to human data can be obtained by taking a probabilistic approach that explicitly models the generative structure of language.
****
A Maximum Entropy Approach to Collaborative Filtering in Dynamic, Sparse, High-Dimensional Domains
Dmitry Pavlov, David Pennock
We develop a maximum entropy (maxent) approach to generating recom- mendations in the context of a user’s current navigation stream, suitable for environments where data is sparse, high-dimensional, and dynamic— conditions typical of many recommendation applications. We address sparsity and dimensionality reduction by ■rst clustering items based on user access patterns so as to attempt to minimize the apriori probabil- ity that recommendations will cross cluster boundaries and then recom- mending only within clusters. We address the inherent dynamic natu

re of the problem by explicitly modeling the data as a time series; we show how this representational expressivity ■ts naturally into a maxent frame- work. We c onduct experiments on data from ResearchIndex, a popu- lar online repository of over 470,000 computer science documents. We show that our maxent formulation out performs several competing algo- rithms in of■ine tests simulating the recommend ation of documents to ResearchIndex users.
****
Annealing and the Rate Distortion Problem
Albert Parker, Tomá\v Gedeon, Alexander Dimitrov
In this paper we introduce methodology to determine the bifurcation structure of optima for a class of similar cost functions from Rate Distortion Theory, Deter min- istic Annealing, Information Distortion and the Information Bottleneck Meth od. We also introduce a numerical algorithm which uses the explicit form of the bifur- cating branches to ■nd optima at a bifurcation point.
****
Self Supervised Boosting
Max Welling, Richard Zemel, Geoffrey E. Hinton
Boosting algorithms and successful applications thereof abound for clas- si■cati on and regression learning problems, but not for unsupervised learning. We propo se a sequential approach to adding features to a ran- dom ■eld model by training them to improve classi■cation performance between the data and an equal-sized s ample of “negative examples” gen- erated from the model’s current estimate of th e data density. Training in each boosting round proceeds in three stages: ■rst w e sample negative examples from the model’s current Boltzmann distribution. Next , a fea- ture is trained to improve classi■cation performance between data and n egative examples. Finally, a coef■cient is learned which determines the importan ce of this feature relative to ones already in the pool. Negative examples only need to be generated once to learn each new feature. The validity of the approac h is demonstrated on binary digits and continuous synthetic data.
****
An Estimation-Theoretic Framework for the Presentation of Multiple Stimuli
Christian Eurich
A framework is introduced for assessing the encoding accuracy and the discrimina tional ability of a population of neurons upon simul- taneous presentation of mu ltiple stimuli. Minimal square estima- tion errors are obtained from a Fisher in formation analysis in an abstract compound space comprising the features of all stimuli. Even for the simplest case of linear superposition of responses and Gau ssian tuning, the symmetries in the compound space are very di(cid:11)erent from those in the case of a single stimulus. The analysis allows for a quantitative description of attentional e(cid:11)ects and can be extended to include neural n onlinearities such as nonclassical receptive (cid:12)elds.
****
Convergent Combinations of Reinforcement Learning with Linear Function Approxima tion
Ralf Schoknecht, Artur Merke
Convergence for iterative reinforcement learning algorithms like TD(O) d epends on the sampling strategy for the transitions. How(cid:173) ever, in p ractical applications it is convenient to take transition data from ar bitrary sources without losing convergence. In this paper we investigate the problem of repeated synchronous updates based on a fixed set of transiti ons. Our main theorem yields suffi(cid:173) cient conditions of convergence for combinations of reinforcement learning algorithms and linear function app roximation. This allows to analyse if a certain reinforcement learning algori thm and a cer(cid:173) tain function approximator are compatible. For the com bination of the residual gradient algorithm with grid-based linear interpolatio n we show that there exists a universal constant learning rate such that the iteration converges independently of the concrete transi(cid:173) tio n data.
****
Informed Projections

David Cohn
Low rank approximation techniques are widespread in pattern recogni- tion resear
ch — they include Latent Semantic Analysis (LSA), Proba- bilistic LSA, Principal
 Components Analysus (PCA), the Generative As- pect Model, and many forms of bib
liometric analysis. All make use of a low-dimensional manifold onto which data a
re projected. Such techniques are generally “unsupervised,” which allows them to
 model data in the absence of labels or categories. With many practi- cal proble
ms, however, some prior knowledge is available in the form of context. In this p
aper, I describe a principled approach to incorpo- rating such information, and
demonstrate its application to PCA-based approximations of several data sets.
****
Automatic Acquisition and Efficient Representation of Syntactic Structures
Zach Solan, Eytan Ruppin, David Horn, Shimon Edelman
The distributional principle according to which morphemes that occur in identica
l contexts belong, in some sense, to the same category [1] has been advanced as
a means for extracting syntactic structures from corpus data. We extend this pri
nciple by applying it recursively, and by us- ing mutual information for estimat
ing category coherence. The resulting model learns, in an unsupervised fashion,
highly structured, distributed representations of syntactic knowledge from corpo
ra. It also exhibits promising behavior in tasks usually thought to require repr
esentations anchored in a grammar, such as systematicity.
****
Application of Variational Bayesian Approach to Speech Recognition
Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura, Naonori Ueda
In this paper, we propose a Bayesian framework, which constructs shared-state tr
iphone HMMs based on a variational Bayesian approach, and recognizes speech base
d on the Bayesian prediction classi(cid:2)cation; variational Bayesian estimatio
n and clustering for speech recognition (VBEC). An appropriate model structure w
ith high recognition perfor- mance can be found within a VBEC framework. Unlike
conventional methods, including BIC or MDL criterion based on the maximum likeli
- hood approach, the proposed model selection is valid in principle, even when t
here are insuf(cid:2)cient amounts of data, because it does not use an asymptoti
c assumption. In isolated word recognition experiments, we show the advantage of
 VBEC over conventional methods, especially when dealing with small amounts of d
ata.
****
Stable Fixed Points of Loopy Belief Propagation Are Local Minima of the Bethe Fr
ee Energy
Tom Heskes
We extend recent work on the connection between loopy belief propagation and the
 Bethe free energy. Constrained minimization of the Bethe free energy can be tur
ned into an unconstrained saddle-point problem. Both converging double-loop algo
rithms and standard loopy belief propagation can be inter- preted as attempts to
 solve this saddle-point problem. Stability analysis then leads us to conclude t
hat stable (cid:12)xed points of loopy belief propagation must be (local) minima
 of the Bethe free energy. Perhaps surprisingly, the converse need not be the ca
se: minima can be unstable (cid:12)xed points. We illustrate this with an exampl
e and discuss implications.
****
Identity Uncertainty and Citation Matching
Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart J. Russell, Ilya Shpitser
Identity uncertainty is a pervasive problem in real-world data analysis. It aris
es whenever objects are not labeled with unique identi■ers or when those identi■
ers may not be perceived perfectly. In such cases, two ob- servations may or may
 not correspond to the same object. In this paper, we consider the problem in th
e context of citation matching—the prob- lem of deciding which citations corresp
ond to the same publication. Our approach is based on the use of a relational pr
obability model to de■ne a generative model for the domain, including models of
author and title corruption and a probabilistic citation grammar. Identity uncer
tainty is handled by extending standard models to incorporate probabilities over

the possible mappings between terms in the language and objects in the domain. Inference is based on Markov chain Monte Carlo, augmented with speci■c methods for generating ef■cient proposals when the domain contains many objects. Results on several citation data sets show that the method outperforms current algorithms for citation matching. The declarative, relational nature of the model also means that our algorithm can determine object characteristics such as author names by combining multiple citations of multiple papers.
****
Fast Sparse Gaussian Process Methods: The Informative Vector Machine
Neil Lawrence, Matthias Seeger, Ralf Herbrich
We present a framework for sparse Gaussian process (GP) methods which uses forward selection with criteria based on information- theoretic principles, previously suggested for active learning. Our goal is not only to learn d{sparse predictors (which can be evalu- ated in O(d) rather than O(n), d (cid:28) n, n the number of training points), but also to perform training under strong restrictions on time and memory requirements. The scaling of our method is at most O(n (cid:1) d2), and in large real-world classi(cid:12)cation experiments we show that it can match prediction performance of the popular support vector machine (SVM), yet can be signi(cid:12)cantly faster in training. In contrast to the SVM, our approximation produces esti- mates of predictive probabilities (‘error bars’), allows for Bayesian model selection and is less complex in implementation.
****
Recovering Articulated Model Topology from Observed Rigid Motion
Leonid Taycher, John Iii, Trevor Darrell
Accurate representation of articulated motion is a challenging problem for machine perception. Several successful tracking algorithms have been developed that model human body as an articulated tree. We pro- pose a learning-based method for creating such articulated models from observations of multiple rigid motions. This paper is concerned with recovering topology of the articulated model, when the rigid motion of constituent segments is known. Our approach is based on ■nding the Maximum Likelihood tree shaped factorization of the joint probability density function (PDF) of rigid segment motions. The topology of graph- ical model formed from this factorization corresponds to topology of the underlying articulated body. We demonstrate the performance of our al- gorithm on both synthetic and real motion capture data.
****
Learning to Perceive Transparency from the Statistics of Natural Scenes
Anat Levin, Assaf Zomet, Yair Weiss
Certain simple images are known to trigger a percept of trans- parency: the input image I is perceived as the sum of two images I(x; y) = I1(x; y) + I2(x; y). This percept is puzzling. First, why do we choose the \more complicated" description with two images rather than the \simpler" explanation I(x; y) = I1(x; y) + 0 ? Sec- ond, given the in■nite number of ways to express I as a sum of two images, how do we compute the \best" decomposition ? Here we suggest that transparency is the rational percept of a sys- tem that is adapted to the statistics of natural scenes. We present a probabilistic model of images based on the qualitative statistics of derivative ■lters and \corner detectors" in natural scenes and use this model to ■nd the most probable decomposition of a novel image. The optimization is performed using loopy belief propa- gation. We show that our model computes perceptually \correct" decompositions on synthetic images and discuss its application to real images.
****
Replay, Repair and Consolidation
Szabolcs Káli, Peter Dayan
A standard view of memory consolidation is that episodes are stored tem- porarily in the hippocampus, and are transferred to the neocortex through replay. Various recent experimental challenges to the idea of transfer, particularly for human memory, are forcing its re-evaluation. However, although there is independent neurophysiological evidence for replay, short of transfer, there are few theoretical ideas for what it might be doing. We suggest and demonstrate two important

computational roles associated with neocortical indices.
****
Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis
Alexei Vinokourov, Nello Cristianini, John Shawe-Taylor
The problem of learning a semantic representation of a text document from data is addressed, in the situation where a corpus of unlabeled paired documents is available, each pair being formed by a short En- glish document and its French translation. This representation can then be used for any retrieval, categorization or clustering task, both in a stan- dard and in a cross-lingual setting. By using kernel functions, in this case simple bag-of-words inner products, each part of the corpus is mapped to a high-dimensional space. The correlations between the two spaces are then learnt by using kernel Canonical Correlation Analysis. A set of directions is found in the ■rst and in the second space that are max- imally correlated. Since we assume the two representations are com- pletely independent apart from the semantic content, any correlation be- tween them should re■ect some semantic similarity. Certain patterns of English words that relate to a speci■c meaning should correlate with cer- tain patterns of French words corresponding to the same meaning, across the corpus. Using the semantic representation obtained in this way we ■rst demonstrate that the correlations detected between the two versions of the corpus are signi■cantly higher than random, and hence that a rep- resentation based on such features does capture statistical patterns that should re■ect semantic information. Then we use such representation both in cross-language and in single-language retrieval tasks, observing performance that is consistently and signi■cantly superior to LSI on the same data.
****
Cluster Kernels for Semi-Supervised Learning
Olivier Chapelle, Jason Weston, Bernhard Schölkopf
We propose a framework to incorporate unlabeled data in kernel classifier, based on the idea that two points in the same cluster are more likely to have the same label. This is achieved by modifying the eigenspectrum of the kernel matrix. Experimental results assess the validity of this approach.
****
A Hierarchical Bayesian Markovian Model for Motifs in Biopolymer Sequences
Eric Xing, Michael Jordan, Richard Karp, Stuart J. Russell
We propose a dynamic Bayesian model for motifs in biopolymer se- quences which captures rich biological prior knowledge and positional dependencies in motif structure in a principled way. Our model posits that the position-speci■c multinomial parameters for monomer distribu- tion are distributed as a latent Dirichlet-mixture random variable, and the position-speci■c Dirichlet component is determined by a hidden Markov process. Model parameters can be ■t on training motifs using a vari- ational EM algorithm within an empirical Bayesian framework. Varia- tional inference is also used for detecting hidden motifs. Our model im- proves over previous models that ignore biological priors and positional dependence. It has much higher sensitivity to motifs during detection and a notable ability to distinguish genuine motifs from false recurring patterns.
****
Speeding up the Parti-Game Algorithm
Maxim Likhachev, Sven Koenig
In this paper, we introduce an ef■cient replanning algorithm for nonde- terministic domains, namely what we believe to be the ■rst incremental heuristic minimax search algorithm. We apply it to the dynamic dis- cretization of continuous domains, resulting in an ef■cient implemen- tation of the parti-game reinforcement-learning algorithm for control in high-dimensional domains.
****
Kernel-Based Extraction of Slow Features: Complex Cells Learn Disparity and Translation Invariance from Natural Images
Alistair Bray, Dominique Martinez
In Slow Feature Analysis (SFA [1]), it has been demonstrated that high-order invariant properties can be extracted by projecting in(cid:173) puts into a

nonlinear space and computing the slowest changing features in this s pace; this has been proposed as a simple general model for learning non linear invariances in the visual system. How(cid:173) ever, this method is h ighly constrained by the curse of dimension(cid:173) ality which limits it t o simple theoretical simulations. This paper demonstrates that by using a d ifferent but closely-related objective function for extracting slowly varying f eatures ([2, 3]), and then ex(cid:173) ploiting the kernel trick, this curse can be avoided. Using this new method we show that both the complex cell p roperties of transla(cid:173) tion invariance and disparity coding can be learnt simultaneously from natural images when complex cells are driven by s imple cells also learnt from the image.
****
Monaural Speech Separation
Guoning Hu, Deliang Wang
that deals with
****
Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector M achines
Fei Sha, Lawrence Saul, Daniel Lee
We derive multiplicative updates for solving the nonnegative quadratic programmi ng problem in support vector machines (SVMs). The updates have a simple closed f orm, and we prove that they converge monotoni- cally to the solution of the maxi mum margin hyperplane. The updates optimize the traditionally proposed objective function for SVMs. They do not involve any heuristics such as choosing a learni ng rate or deciding which variables to update at each iteration. They can be use d to adjust all the quadratic programming variables in parallel with a guarantee of im- provement at each iteration. We analyze the asymptotic convergence of th e updates and show that the coef■cients of non-support vectors decay geometrical ly to zero at a rate that depends on their margins. In practice, the updates con verge very rapidly to good classi■ers.
****
Kernel Design Using Boosting
Koby Crammer, Joseph Keshet, Yoram Singer
The focus of the paper is the problem of learning kernel operators from empirica l data. We cast the kernel design problem as the construction of an accurate ker nel from simple (and less accurate) base kernels. We use the boosting paradigm t o perform the kernel construction process. To do so, we modify the booster so as to accommodate kernel operators. We also devise an ef■cient weak-learner for si mple kernels that is based on generalized eigen vector decomposition. We demonst rate the effective- ness of our approach on synthetic data and on the USPS datas et. On the USPS dataset, the performance of the Perceptron algorithm with learne d kernels is systematically better than a ■xed RBF kernel.
****
Rational Kernels
Corinna Cortes, Patrick Haffner, Mehryar Mohri
We introduce a general family of kernels based on weighted transduc- ers or rati onal relations, rational kernels, that can be used for analysis of variable-leng th sequences or more generally weighted automata, in appli- cations such as comp utational biology or speech recognition. We show that rational kernels can be co mputed ef■ciently using a general algo- rithm of composition of weighted transdu cers and a general single-source shortest-distance algorithm. We also describe s everal general families of positive de■nite symmetric rational kernels. These ge neral kernels can be combined with Support Vector Machines to form ef■cient and power- ful techniques for spoken-dialog classi■cation: highly complex kernels be come easy to design and implement and lead to substantial improve- ments in the classi■cation accuracy. We also show that the string kernels considered in appli cations to computational biology are all speci■c in- stances of rational kernels .
****
Linear Combinations of Optic Flow Vectors for Estimating Self-Motion - a Real-Wo

rld Test of a Neural Model
Matthias Franz, Javaan Chahl
The tangential neurons in the ■y brain are sensitive to the typical optic ■ow pa
tterns generated during self-motion. In this study, we examine whether a simpli■
ed linear model of these neurons can be used to esti- mate self-motion from the
optic ■ow. We present a theory for the con- struction of an estimator consisting
 of a linear combination of optic ■ow vectors that incorporates prior knowledge
both about the distance distri- bution of the environment, and about the noise a
nd self-motion statistics of the sensor. The estimator is tested on a gantry car
rying an omnidirec- tional vision sensor. The experiments show that the proposed
 approach leads to accurate and robust estimates of rotation rates, whereas tran
sla- tion estimates turn out to be less reliable.
****
Kernel Dependency Estimation
Jason Weston, Olivier Chapelle, Vladimir Vapnik, André Elisseeff, Bernhard Schöl
kopf
We consider the learning problem of finding a dependency between a general cl
ass of objects and another, possibly different, general class of objects.
  The objects can be for example: vectors, images, strings, trees or graphs.
  Such a task is made possible by employing similarity measures in both inp
ut and output spaces using ker(cid:173) nel functions, thus embedding t
he objects into vector spaces. We experimentally validate our approach
 on several tasks: mapping strings to strings, pattern recognition, and reco
nstruction from par(cid:173) tial images.
****
Circuit Model of Short-Term Synaptic Dynamics
Shih-Chii Liu, Malte Boegershausen, Pascal Suter
We describe a model of short-term synaptic depression that is derived from a sil
icon circuit implementation. The dynamics of this circuit model are similar to t
he dynamics of some present theoretical models of short- term depression except
that the recovery dynamics of the variable de- scribing the depression is nonlin
ear and it also depends on the presynap- tic frequency. The equations describing
 the steady-state and transient re- sponses of this synaptic model ■t the experi
mental results obtained from a fabricated silicon network consisting of leaky in
tegrate-and-■re neu- rons and different types of synapses. We also show experime
ntal data demonstrating the possible computational roles of depression. One pos-
 sible role of a depressing synapse is that the input can quickly bring the neur
on up to threshold when the membrane potential is close to the rest- ing potenti
al.
****
Coulomb Classifiers: Generalizing Support Vector Machines via an Analogy to Elec
trostatic Systems
Sepp Hochreiter, Michael C. Mozer, Klaus Obermayer
We introduce a family of classi■ers based on a physical analogy to an electrosta
tic system of charged conductors. The family, called Coulomb classi■ers, include
s the two best-known support-vector machines (SVMs), the ”{SVM and the C{SVM. In
 the electrostat- ics analogy, a training example corresponds to a charged condu
ctor at a given location in space, the classi■cation function corresponds to the
 electrostatic potential function, and the training objective function correspon
ds to the Coulomb energy. The electrostatic framework provides not only a novel
interpretation of existing algo- rithms and their interrelationships, but it sug
gests a variety of new methods for SVMs including kernels that bridge the gap be
tween polynomial and radial-basis functions, objective functions that do not req
uire positive-de■nite kernels, regularization techniques that allow for the cons
truction of an optimal classi■er in Minkowski space. Based on the framework, we
propose novel SVMs and per- form simulation studies to show that they are compar
able or su- perior to standard SVMs. The experiments include classi■cation tasks
 on data which are represented in terms of their pairwise prox- imities, where a
 Coulomb Classi■er outperformed standard SVMs.
****

Robust Novelty Detection with Single-Class MPM
Laurent Ghaoui, Michael Jordan, Gert Lanckriet
the "single-class minimax probabil(cid:173)
****
Binary Tuning is Optimal for Neural Rate Coding with High Temporal Resolution
Matthias Bethge, David Rotermund, Klaus Pawelzik
Here we derive optimal gain functions for minimum mean square re(cid:173) constr uction from neural rate responses subjected to Poisson noise. The shape of these functions strongly depends on the length T of the time window within which spik es are counted in order to estimate the under(cid:173) lying firing rate. A phas e transition towards pure binary encoding occurs if the maximum mean spike count becomes smaller than approximately three provided the minimum firing rate is ze ro. For a particular function class, we were able to prove the existence of a se cond-order phase tran(cid:173) sition analytically. The critical decoding time w indow length obtained from the analytical derivation is in precise agreement wit h the numerical results. We conclude that under most circumstances relevant to i nforma(cid:173) tion processing in the brain, rate coding can be better ascribed to a binary (low-entropy) code than to the other extreme of rich analog coding.
****
Feature Selection in Mixture-Based Clustering
Martin Law, Anil Jain, Mário Figueiredo
There exist many approaches to clustering, but the important issue of feature se lection, i.e., selecting the data attributes that are relevant for clustering, i s rarely addressed. Feature selection for clustering is dif■cult due to the abse nce of class labels. We propose two approaches to feature selection in the conte xt of Gaussian mixture-based clustering. In the ■rst one, instead of making hard selections, we estimate feature saliencies. An expectation-maximization (EM) al gorithm is derived for this task. The second approach extends Koller and Sahami' s mutual-information- based feature relevance criterion to the unsupervised case . Feature selec- tion is then carried out by a backward search scheme. This sche me can be classi■ed as a “wrapper”, since it wraps mixture estimation in an oute r layer that performs feature selection. Experimental results on synthetic and r eal data show that both methods have promising performance.
****
Minimax Differential Dynamic Programming: An Application to Robust Biped Walking
Jun Morimoto, Christopher Atkeson
We developed a robust control policy design method in high-dimensional state spa ce by using differential dynamic programming with a minimax criterion. As an exa mple, we applied our method to a simulated ■ve link biped robot. The results sho w lower joint torques from the optimal con- trol policy compared to a hand-tuned PD servo controller. Results also show that the simulated biped robot can succe ssfully walk with unknown disturbances that cause controllers generated by stand ard differential dy- namic programming and the hand-tuned PD servo to fail. Lear ning to compensate for modeling error and previously unknown disturbances in con junction with robust control design is also demonstrated.
****
Theory-Based Causal Inference
Joshua Tenenbaum, Thomas Griffiths
People routinely make sophisticated causal inferences unconsciously, ef- fortles sly, and from very little data – often from just one or a few ob- servations. We argue that these inferences can be explained as Bayesian computations over a hy pothesis space of causal graphical models, shaped by strong top-down prior knowl edge in the form of intuitive theories. We present two case studies of our appro ach, including quantitative mod- els of human causal judgments and brief compari sons with traditional bottom-up models of inference.
****
Learning to Classify Galaxy Shapes Using the EM Algorithm
Sergey Kirshner, Igor Cadez, Padhraic Smyth, Chandrika Kamath
We describe the application of probabilistic model-based learning to the problem of automatically identifying classes of galaxies, based on both morphological a

nd pixel intensity characteristics. The EM algorithm can be used to learn how to spatially orient a set of galaxies so that they are geometrically aligned. We augment this “ordering-model” with a mixture model on objects, and demonstrate how classes of galaxies can be learned in an unsupervised manner using a two-level EM algorithm. The resulting models provide highly accurate classi£cation of galaxies in cross-validation experiments.
****
FloatBoost Learning for Classification
Stan Li, Zhenqiu Zhang, Heung-yeung Shum, Hongjiang Zhang
AdaBoost [3] minimizes an upper error bound which is an exponential function of the margin on the training set [14]. However, the ultimate goal in applications of pattern classi■cation is always minimum error rate. On the other hand, AdaBoost needs an effective procedure for learning weak classi■ers, which by itself is dif■cult especially for high dimensional data. In this paper, we present a novel procedure, called FloatBoost, for learning a better boosted classi■er. FloatBoost uses a backtrack mechanism after each iteration of AdaBoost to remove weak classi■ers which cause higher error rates. The resulting ■oat-boosted classi■er consists of fewer weak classi■ers yet achieves lower error rates than AdaBoost in both training and test. We also propose a statistical model for learning weak classi■ers, based on a stagewise approximation of the posterior using an overcomplete set of scalar features. Experi- mental comparisons of FloatBoost and AdaBoost are provided through a dif■cult classi■cation problem, face detection, where the goal is to learn from training examples a highly nonlinear classi■er to differentiate be- tween face and nonface patterns in a high dimensional space. The results clearly demonstrate the promises made by FloatBoost over AdaBoost.
****
Adaptive Scaling for Feature Selection in SVMs
Yves Grandvalet, Stéphane Canu
This paper introduces an algorithm for the automatic relevance determi- nation of input variables in kernelized Support Vector Machines. Rele- vance is measured by scale factors de■ning the input space metric, and feature selection is performed by assigning zero weights to irrelevant variables. The metric is automatically tuned by the minimization of the standard SVM empirical risk, where scale factors are added to the usual set of parameters de■ning the classi■er. Feature selection is achieved by constraints encouraging the sparsity of scale factors. The resulting algorithm compares favorably to state-of-the-art feature selection proce- dures and demonstrates its effectiveness on a demanding facial expres- sion recognition problem.
****
Adaptive Classification by Variational Kalman Filtering
Peter Sykacek, Stephen J. Roberts
We propose in this paper a probabilistic approach for adaptive inference of generalized nonlinear classi■cation that combines the computational advantage of a parametric solution with the ■exibility of sequential sam- pling techniques. We regard the parameters of the classi■er as latent states in a ■rst order Markov process and propose an algorithm which can be regarded as variational generalization of standard Kalman ■lter- ing. The variational Kalman ■lter is based on two novel lower bounds that enable us to use a non-degenerate distribution over the adaptation rate. An extensive empirical evaluation demonstrates that the proposed method is capable of infering competitive classi■ers both in stationary and non-stationary environments. Although we focus on classi■cation, the algorithm is easily extended to other generalized nonlinear models.
****
A Statistical Mechanics Approach to Approximate Analytical Bootstrap Averages
Dörthe Malzahn, Manfred Opper
We apply the replica method of Statistical Physics combined with a vari- ational method to the approximate analytical computation of bootstrap averages for estimating the generalization error. We demonstrate our ap- proach on regression with Gaussian processes and compare our results with averages obtained by Monte-Carlo sampling.

****
Developing Topography and Ocular Dominance Using Two aVLSI Vision Sensors and a Neurotrophic Model of Plasticity
Terry Elliott, Jörg Kramer
A neurotrophic model for the co-development of topography and ocular dominance c olumns in the primary visual cortex has recently been pro- posed. In the present work, we test this model by driving it with the output of a pair of neuronal vi sion sensors stimulated by disparate mov- ing patterns. We show that the tempora l correlations in the spike trains generated by the two sensors elicit the devel opment of re■ned topogra- phy and ocular dominance columns, even in the presence of signi■cant amounts of spontaneous activity and ■xed-pattern noise in the sen sors.
****
Gaussian Process Priors with Uncertain Inputs Application to Multiple-Step Ahead Time Series Forecasting
Agathe Girard, Carl Rasmussen, Joaquin Quiñonero Candela, Roderick Murray-Smith
We consider the problem of multi-step ahead prediction in time series analysis u sing the non-parametric Gaussian process model. -step ahead forecasting of a di screte-time non-linear dynamic system can be per- formed by doing repeated one-s tep ahead predictions. For a state-space at time model of the form is based on t he point estimates of the previous outputs. In this pa- per, we show how, usin g an analytical Gaussian approximation, we can formally incorporate the uncertai nty about intermediate regressor values, thus updating the uncertainty on the cu rrent prediction.
****
Maximum Likelihood and the Information Bottleneck
Noam Slonim, Yair Weiss

****
The Stability of Kernel Principal Components Analysis and its Relation to the Pr ocess Eigenspectrum
Christopher Williams, John Shawe-taylor
In this paper we analyze the relationships between the eigenvalues of the m x m Gram matrix K for a kernel k(·, .) corresponding to a sample Xl, ... , Xm drawn from a density p(x) and the eigenvalues of the corresponding conti nuous eigenproblem. We bound the dif(cid:173) ferences between the two spectra and provide a performance bound on kernel peA.
****
Adaptive Quantization and Density Estimation in Silicon
David Hsu, Seth Bridges, Miguel Figueroa, Chris Diorio
We present the bump mixture model, a statistical model for analog data where th e probabilistic semantics, inference, and learning rules derive from low-level transistor behavior. The bump mixture model relies on translinear circuits to p erform probabilistic infer- ence, and floating-gate devices to perform adaptatio n. This system is low power, asynchronous, and fully parallel, and supports var i- ous on-chip learning algorithms. In addition, the mixture model can perform several tasks such as probability estimation, vector quanti- zation, classificat ion, and clustering. We tested a fabricated system on clustering, quantization, and classification of handwritten digits and show performance comparable to th e E-M algorithm on mix- tures of Gaussians.
****
Improving Transfer Rates in Brain Computer Interfacing: A Case Study
Peter Meinicke, Matthias Kaper, Florian Hoppe, Manfred Heumann, Helge Ritter
In this paper we present results of a study on brain computer interfacing. We ad opted an approach of Farwell & Donchin [4], which we tried to improve in several aspects. The main objective was to improve the trans- fer rates based on of■ine analysis of EEG-data but within a more realistic setup closer to an online real ization than in the original studies. The ob- jective was achieved along two dif ferent tracks: on the one hand we used state-of-the-art machine learning techniq ues for signal classi■cation and on the other hand we augmented the data space b

y using more electrodes for the interface. For the classi■cation task we utilize d SVMs and, as mo- tivated by recent ■ndings on the learning of discriminative d ensities, we accumulated the values of the classi■cation function in order to co mbine several classi■cations, which ■nally lead to signi■cantly improved rates a s compared with techniques applied in the original work. In combina- tion with t he data space augmentation, we achieved competitive transfer rates at an average of 50.5 bits/min and with a maximum of 84.7 bits/min.
****
Discriminative Learning for Label Sequences via Boosting
Yasemin Altun, Thomas Hofmann, Mark Johnson
This paper investigates a boosting approach to discriminative learning o f label sequences based on a sequence rank loss function. The proposed method c ombines many of the advantages of boost(cid:173) ing schemes with the efficie ncy of dynamic programming methods and is attractive both, conceptually and c omputationally. In addi(cid:173) tion, we also discuss alternative approaches based on the Hamming loss for label sequences. The sequence boosting algorith m offers an interesting alternative to methods based on HMMs and the mor e recently proposed Conditional Random Fields. Applications areas for the p resented technique range from natural language processing and information extr action to computational biology. We include experiments on named entity recognition and part-of-speech tag(cid:173) ging which demonstrate the va lidity and competitiveness of our approach.
****
Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games
Xiaofeng Wang, Tuomas Sandholm
Multiagent learning is a key problem in AI. In the presence of multi- ple Nash e quilibria, even agents with non-con■icting interests may not be able to learn an optimal coordination policy. The problem is exac- cerbated if the agents do not know the game and independently receive noisy payoffs. So, multiagent reinforfc ement learning involves two inter- related problems: identifying the game and le arning to play. In this paper, we present optimal adaptive learning, the ■rst al gorithm that converges to an optimal Nash equilibrium with probability 1 in any team Markov game. We provide a convergence proof, and show that the algorithm’s parameters are easy to set to meet the convergence conditions.
****
Using Manifold Stucture for Partially Labeled Classification
Mikhail Belkin, Partha Niyogi
We consider the general problem of utilizing both labeled and un(cid:173) labele d data to improve classification accuracy. Under t he assump(cid:173) tion that the data lie on a submanifold in a high dimensional space, we develop an algori thmic framework to classify a partially labeled data set in a principled manner . The central idea of our approach is that classification functions are natura lly defined only on t he sub(cid:173) manifold in question rather than the total ambient space. Using the Laplace Beltrami operator one produces a basis for a Hilbert space of square integrable functions on the submanifold. To recover suc h a basis , only unlab eled examples are required. Once a basis is ob(cid:173) tained , training can be performed using the labeled data set. Our algorithm mo dels the manifold using the adjacency graph for the data and approximates the L aplace Beltrami operator by the graph Laplacian. Practical applications to imag e and text classification are considered.
****
Recovering Intrinsic Images from a Single Image
Marshall Tappen, William Freeman, Edward Adelson
We present an algorithm that uses multiple cues to recover shading and re■ectanc e intrinsic images from a single image. Using both color in- formation and a cla ssi■er trained to recognize gray-scale patterns, each image derivative is classi ■ed as being caused by shading or a change in the surface’s re■ectance. Generali zed Belief Propagation is then used to propagate information from areas where th e correct classi■cation is clear to areas where it is ambiguous. We also show re sults on real images.

****
A Note on the Representational Incompatibility of Function Approximation and Factored Dynamics
Eric Allender, Sanjeev Arora, Michael Kearns, Cristopher Moore, Alexander Russell
We establish a new hardness result that shows that the dif■culty of plan- ning in factored Markov decision processes is representational rather than just computational. More precisely, we give a ■xed family of fac- tored MDPs with linear rewards whose optimal policies and value func- tions simply cannot be represented succinctly in any standard parametric form. Previous hardness results indicated that computing good policies from the MDP parameters was dif■cult, but left open the possibility of succinct function approximation for any ■xed factored MDP. Our result applies even to policies which yield a polynomially poor approximation to the optimal value, and highlights interesting connectionswith the com- plexity class of Arthur-Merlin games.
****
A Prototype for Automatic Recognition of Spontaneous Facial Actions
M.S. Bartlett, G.C. Littlewort, T.J. Sejnowski, J.R. Movellan
We present ongoing work on a project for automatic recognition of spon- taneous facial actions. Spontaneous facial expressions differ substan- tially from posed expressions, similar to how continuous, spontaneous speech differs from isolated words produced on command. Previous methods for automatic facial expression recognition assumed images were collected in controlled environments in which the subjects delib- erately faced the camera. Since people often nod or turn their heads, automatic recognition of spontaneous facial behavior requires methods for handling out-of-image-plane head rotations. Here we explore an ap- proach based on 3-D warping of images into canonical views. We eval- uated the performance of the approach as a front-end for a spontaneous expression recognition system using support vector machines and hidden Markov models. This system employed general purpose learning mech- anisms that can be applied to recognition of any facial movement. The system was tested for recognition of a set of facial actions de■ned by the Facial Action Coding System (FACS). We showed that 3D tracking and warping followed by machine learning techniques directly applied to the warped images, is a viable and promising technology for automatic facial expression recognition. One exciting aspect of the approach pre- sented here is that information about movement dynamics emerged out of ■lters which were derived from the statistics of images.
****
Exact MAP Estimates by (Hyper)tree Agreement
Martin J. Wainwright, Tommi Jaakkola, Alan Willsky
We describe a method for computing provably exact maximum a poste- riori (MAP) estimates for a subclass of problems on graphs with cycles. The basic idea is to represent the original problem on the graph with cy- cles as a convex combination of tree-structured problems. A convexity argument then guarantees that the optimal value of the original problem (i.e., the log probability of the MAP assignment) is upper bounded by the combined optimal values of the tree problems. We prove that this upper bound is met with equality if and only if the tree problems share an opti- mal con■guration in common. An important implication is that any such shared con■guration must also be the MAP con■guration for the original problem. Next we develop a tree-reweighted max-product algorithm for attempting to ■nd convex combinations of tree-structured problems that share a common optimum. We give necessary and suf■cient conditions for a ■xed point to yield the exact MAP estimate. An attractive feature of our analysis is that it generalizes naturally to convex combinations of hypertree-structured distributions.
****
Prediction of Protein Topologies Using Generalized IOHMMs and RNNs
Gianluca Pollastri, Pierre Baldi, Alessandro Vullo, Paolo Frasconi
We develop and test new machine learning methods for the predic- tion of topological representations of protein structures in the form of coarse- or (cid:12)ne-grained contact or distance maps that are transla- tion and rotation invariant.

The methods are based on generalized input-output hidden Markov models (GIOHMMs) and generalized recursive neural networks (GRNNs). The methods are used to pre- dict topology directly in the (cid:12)ne-grained case and, in the coarse- grained case, indirectly by (cid:12)rst learning how to score candidate graphs and then using the scoring function to search the space of possible con(cid:12)gurations. Computer simulations show that the pre- dictors achieve state-of-the-art performance.
****