Bayesian Surprise Attracts Human Attention

Laurent Itti, Pierre Baldi

The concept of surprise is central to sensory processing, adaptation, learning, and attention. Yet, no widely-accepted mathematical theory currently exists to quantitatively characterize surprise elicited by a stimulus or event, for observers that range from single neurons to complex natural or engineered systems. We describe a formal Bayesian definition of surprise that is the only consistent formulation under minimal axiomatic assumptions. Surprise quantifies how data affects a natural or artificial observer, by measuring the difference between posterior and prior beliefs of the observer. Using this framework we measure the extent to which humans direct their gaze towards surprising items while watching television and video games. We find that subjects are strongly attracted towards surprising locations, with 72% of all human gaze shifts directed towards locations more surprising than the average, a figure which rises to 84% when considering only gaze targets simultaneously selected by all subjects. The resulting theory of surprise is applicable across different spatio-temporal scales, modalities, and levels of abstraction. Life is full of surprises, ranging from a great christmas gift or a new magic trick, to wardrobe malfunctions, reckless drivers, terrorist attacks, and tsunami waves. Key to survival is our ability to rapidly attend to, identify, and learn from surprising events, to decide on present and future courses of action [1]. Yet, little theoretical and computational understanding exists of the very essence of surprise, as evidenced by the absence from our everyday vocabulary of a quantitative unit of surprise: Qualities such as the "wow factor" have remained vague and elusive to mathematical analysis. Informal correlates of surprise exist at nearly all stages of neural processing. In sensory neuroscience, it has been suggested that only the unexpected at one stage is transmitted to the next stage [2]. Hence, sensory cortex may have evolved to adapt to, to predict, and to quiet down the expected statistical regularities of the world [3, 4, 5, 6], focusing instead on events that are unpredictable or surprising. Electrophysiological evidence for this early sensory emphasis onto surprising stimuli exists from studies of adaptation in visual [7, 8, 4, 9], olfactory [10, 11], and auditory cortices [12], subcortical structures like the LGN [13], and even retinal ganglion cells [14, 15] and cochlear hair cells [16]: neural response greatly attenuates with repeated or prolonged exposure to an initially novel stimulus. Surprise and novelty are also central to learning and memory formation [1], to the point that surprise is believed to be a necessary trigger for associative learning [17, 18],

as supported by mounting evidence for a role of the hippocampus as a novelty detector [19, 20, 21]. Finally, seeking novelty is a well-identified human character trait, with possible association with the dopamine D4 receptor gene [22, 23, 24]. In the Bayesian framework, we develop the only consistent theory of surprise, in terms of the difference between the posterior and prior distributions of beliefs of an observer over the available class of models or hypotheses about the world. We show that this definition derived from first principles presents key advantages over more ad-hoc formulations, typically relying on detecting outlier stimuli. Armed with this new framework, we provide direct experimental evidence that surprise best characterizes what attracts human gaze in large amounts of natural video stimuli. We here extend a recent pilot study [25], adding more comprehensive theory, large-scale human data collection, and additional analysis.

****

On Local Rewards and Scaling Distributed Reinforcement Learning

Drew Bagnell, Andrew Ng

We consider the scaling of the number of examples necessary to achieve good performance in distributed, cooperative, multi-agent reinforcement learning, as a function of the the number of agents n. We prove a worstcase lower bound showing that algorithms that rely solely on a global reward signal to learn policies confront a fundamental limit: They require a number of real-world examples that scales roughly linearly in the number of agents. For settings of interest with a very large number of agents, this is impractical. We demonstrate, however, that there is a class of algorithms that, by taking advantage of local reward signals in

large distributed Markov Decision Processes, are able to ensure good performance with a number of samples that scales as O(log n). This makes them applicable even in settings with a very large number of agents n.
****

Learning Shared Latent Structure for Image Synthesis and Robotic Imitation
Aaron Shon, Keith Grochow, Aaron Hertzmann, Rajesh PN Rao
We propose an algorithm that uses Gaussian process regression to learn common hidden structure shared between corresponding sets of heterogenous observations. The observation spaces are linked via a single, reduced-dimensionality latent variable space. We present results from two datasets demonstrating the algorithms's ability to synthesize novel data from learned correspondences. We first show that the method can learn the nonlinear mapping between corresponding views of objects, filling in missing data as needed to synthesize novel views. We then show that the method can learn a mapping between human degrees of freedom and robotic degrees of freedom for a humanoid robot, allowing robotic imitation of human poses from motion capture data.
****

Recovery of Jointly Sparse Signals from Few Random Projections
Michael B. Wakin, Marco Duarte, Shriram Sarvotham, Dror Baron, Richard G. Baraniuk
Compressed sensing is an emerging ■eld based on the revelation that a small group of linear projections of a sparse signal contains enough information for reconstruc- tion. In this paper we introduce a new theory for distributed compressed sensing (DCS) that enables new distributed coding algorithms for multi-signal ensembles that exploit both intra- and inter-signal correlation structures. The DCS theory rests on a new concept that we term the joint sparsity of a signal ensemble. We study three simple models for jointly sparse signals, propose algorithms for joint recov- ery of multiple signals from incoherent projections, and characterize theoretically and empirically the number of measurements per sensor required for accurate re- construction. In some sense DCS is a framework for distributed compression of sources with memory, which has remained a challenging problem in information theory for some time. DCS is immediately applicable to a range of problems in sensor networks and arrays.
****

Fixing two weaknesses of the Spectral Method
Kevin Lang
We discuss two intrinsic weaknesses of the spectral graph partitioning method, both of which have practical consequences. The first is that spectral embeddings tend to hide the best cuts from the commonly used hyperplane rounding method. Rather than cleaning up the resulting suboptimal cuts with local search, we recommend the adoption of flow-based rounding. The second weakness is that for many "power law" graphs, the spectral method produces cuts that are highly unbalanced, thus decreasing the usefulness of the method for visualization (see figure 4(b)) or as a basis for divide-and-conquer algorithms. These balance problems, which occur even though the spectral method's quotient-style objective function does encourage balance, can be fixed with a stricter balance constraint that turns the spectral mathematical program into an SDP that can be solved for million-node graphs by a method of Burer and Monteiro.
****

Saliency Based on Information Maximization
Neil Bruce, John Tsotsos
A model of bottom-up overt attention is proposed based on the principle of maximizing information sampled from a scene. The proposed opera(cid:173) tion is based on Shannon's self-information measure and is achieved in a neural circuit, which is demonstrated as having close ties with the cir(cid:173) cuitry existent in the primate visual cortex. It is further shown that the proposed saliency measure may be extended to address issues that cur(cid:173) rently elude explanation in the domain of saliency based models. Results on natural images are compared with experimental eye tracking data re(cid:173) vealing the efficacy of the model in predicting the deployment of overt attention

as compared with existing efforts.
****
Learning vehicular dynamics, with application to modeling helicopters
Pieter Abbeel, Varun Ganapathi, Andrew Ng
We consider the problem of modeling a helicopter's dynamics based on state-actio n trajectories collected from it. The contribution of this paper is two-fold. Fi rst, we consider the linear models such as learned by C I F E R (the industry st andard in helicopter identification), and show that the linear parameterization makes certain properties of dynamical systems, such as inertia, fundamentally di fficult to capture. We propose an alternative, acceleration based, parameterizat ion that does not suffer from this deficiency, and that can be learned as effici ently from data. Second, a Markov decision process model of a helicopter's dynam ics would explicitly model only the one-step transitions, but we are often inter ested in a model's predictive performance over longer timescales. In this paper, we present an efficient algorithm for (approximately) minimizing the prediction error over long time scales. We present empirical results on two different heli copters. Although this work was motivated by the problem of modeling helicopters , the ideas presented here are general, and can be applied to modeling large cla sses of vehicular dynamics.
****
Active Learning for Misspecified Models
Masashi Sugiyama
expressed as
****
Identifying Distributed Object Representations in Human Extrastriate Visual Cort ex
Rory Sayres, David Ress, Kalanit Grill-spector
The category of visual stimuli has been reliably decoded from patterns of neural activity in extrastriate visual cortex [1]. It has yet to be seen whether objec t identity can be inferred from this activity. We present fMRI data measuring re sponses in human extrastriate cortex to a set of 12 distinct object images. We u se a simple winner-take-all classifier, using half the data from each recording session as a training set, to evaluate encoding of object identity across fMRI v oxels. Since this approach is sensitive to the inclusion of noisy voxels, we des cribe two methods for identifying subsets of voxels in the data which optimally distinguish object identity. One method characterizes the reliability of each vo xel within subsets of the data, while another estimates the mutual information o f each voxel with the stimulus set. We find that both metrics can identify subse ts of the data which reliably encode object identity, even when noisy measuremen ts are artificially added to the data. The mutual information metric is less eff icient at this task, likely due to constraints in fMRI data.
****
Convex Neural Networks
Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, Patrice Marcotte
Convexity has recently received a lot of attention in the machine learning commu nity, and the lack of convexity has been seen as a major disadvantage of many le arning algorithms, such as multi-layer artificial neural networks. We show that training multi-layer neural networks in which the number of hidden units is lear ned can be viewed as a convex optimization problem. This problem involves an inf inite number of variables, but can be solved by incrementally inserting a hidden unit at a time, each time finding a linear classifier that minimizes a weighted sum of errors.
****
Temporal Abstraction in Temporal-difference Networks
Eddie Rafols, Anna Koop, Richard S. Sutton
We present a generalization of temporal-difference networks to include temporall y abstract options on the links of the question network. Temporal-difference (TD) networks have been proposed as a way of representing and learning a wide varie ty of predictions about the interaction between an agent and its environment. Th ese predictions are compositional in that their targets are defined in terms of

other predictions, and subjunctive in that that they are about what would happen if an action or sequence of actions were taken. In conventional TD networks, the inter-related predictions are at successive time steps and contingent on a single action; here we generalize them to accommodate extended time intervals and contingency on whole ways of behaving. Our generalization is based on the options framework for temporal abstraction. The primary contribution of this paper is to introduce a new algorithm for intra-option learning in TD networks with function approximation and eligibility traces. We present empirical examples of our algorithm's effectiveness and of the greater representational expressiveness of temporallyabstract TD networks. The primary distinguishing feature of temporal-difference (TD) networks (Sutton & Tanner, 2005) is that they permit a general compositional specification of the goals of learning. The goals of learning are thought of as predictive questions being asked by the agent in the learning problem, such as "What will I see if I step forward and look right?" or "If I open the fridge, will I see a bottle of beer?" Seeing a bottle of beer is of course a complicated perceptual act. It might be thought of as obtaining a set of predictions about what would happen if certain reaching and grasping actions were taken, about what would happen if the bottle were opened and turned upside down, and of what the bottle would look like if viewed from various angles. To predict seeing a bottle of beer is thus to make a prediction about a set of other predictions. The target for the overall prediction is a composition in the mathematical sense of the first prediction with each of the other predictions. TD networks are the first framework for representing the goals of predictive learning in a compositional, machine-accessible form. Each node of a TD network represents an individual question--something to be predicted--and has associated with it a value representing an answer to the question--a prediction of that something. The questions are represented by a set of directed links between nodes. If node 1 is linked to node 2, then node 1 rep-
****
Robust Fisher Discriminant Analysis
Seung-jean Kim, Alessandro Magnani, Stephen Boyd
Fisher linear discriminant analysis (LDA) can be sensitive to the problem data. Robust Fisher LDA can systematically alleviate the sensitivity problem by explicitly incorporating a model of data uncertainty in a classification problem and optimizing for the worst-case scenario under this model. The main contribution of this paper is show that with general convex uncertainty models on the problem data, robust Fisher LDA can be carried out using convex optimization. For a certain type of product form uncertainty model, robust Fisher LDA can be carried out at a cost comparable to standard Fisher LDA. The method is demonstrated with some numerical examples. Finally, we show how to extend these results to robust kernel Fisher discriminant analysis, i.e., robust Fisher LDA in a high dimensional feature space.
****
Measuring Shared Information and Coordinated Activity in Neuronal Networks
Kristina Klinkner, Cosma Shalizi, Marcelo Camperi
Most nervous systems encode information about stimuli in the responding activity of large neuronal networks. This activity often manifests itself as dynamically coordinated sequences of action potentials. Since multiple electrode recordings are now a standard tool in neuroscience research, it is important to have a measure of such network-wide behavioral coordination and information sharing, applicable to multiple neural spike train data. We propose a new statistic, informational coherence, which measures how much better one unit can be predicted by knowing the dynamical state of another. We argue informational coherence is a measure of association and shared information which is superior to traditional pairwise measures of synchronization and correlation. To find the dynamical states, we use a recently-introduced algorithm which reconstructs effective state spaces from stochastic time series. We then extend the pairwise measure to a multivariate analysis of the network by estimating the network multi-information. We illustrate our method by testing it on a detailed model of the transition from gamma to beta rhythms. Much of the most important information in neural systems is share

d over multiple neurons or cortical areas, in such forms as population codes and distributed representations [1]. On behavioral time scales, neural information is stored in temporal patterns of activity as opposed to static markers; therefore, as information is shared between neurons or brain regions, it is physically instantiated as coordination between entire sequences of neural spikes. Furthermore, neural systems and regions of the brain often require coordinated neural activity to perform important functions; acting in concert requires multiple neurons or cortical areas to share information [2]. Thus, if we want to measure the dynamic network-wide behavior of neurons and test hypotheses about them, we need reliable, practical methods to detect and quantify behavioral coordination and the associated information sharing across multiple neural units. These would be especially useful in testing ideas about how particular forms of coordination relate to distributed coding (e.g., that of [3]). Current techniques to analyze relations among spike trains handle only pairs of neurons, so we further need a method which is extendible to analyze the coordination in the network, system, or region as a whole. Here we propose a new measure of behavioral coordination and information sharing, informational coherence, based on the notion of dynamical state. Section 1 argues that coordinated behavior in neural systems is often not captured by exist-
****
Computing the Solution Path for the Regularized Support Vector Regression
Lacey Gunter, Ji Zhu
In this paper we derive an algorithm that computes the entire solu- tion path of the support vector regression, with essentially the same computational cost as ■tting one SVR model. We also propose an unbiased estimate for the degrees of freedom of the SVR model, which allows convenient selection of the regularization parameter.
****
Soft Clustering on Graphs
Kai Yu, Shipeng Yu, Volker Tresp
We propose a simple clustering framework on graphs encoding pairwise data similarities. Unlike usual similarity-based methods, the approach softly assigns data to clusters in a probabilistic way. More importantly, a hierarchical clustering is naturally derived in this framework to gradually merge lower-level clusters into higher-level ones. A random walk analysis indicates that the algorithm exposes clustering structures in various resolutions, i.e., a higher level statistically models a longer-term diffusion on graphs and thus discovers a more global clustering structure. Finally we provide very encouraging experimental results.
****
Efficient estimation of hidden state dynamics from spike trains
Marton G. Danoczy, Richard Hahnloser
Neurons can have rapidly changing spike train statistics dictated by the underlying network excitability or behavioural state of an animal. To estimate the time course of such state dynamics from single- or multi- ple neuron recordings, we have developed an algorithm that maximizes the likelihood of observed spike trains by optimizing the state lifetimes and the state-conditional interspike-interval (ISI) distributions. Our non- parametric algorithm is free of time-binning and spike-counting prob- lems and has the computational complexity of a Mixed-state Markov Model operating on a state sequence of length equal to the total num- ber of recorded spikes. As an example, we ■t a two-state model to paired recordings of premotor neurons in the sleeping songbird. We ■nd that the two state-conditional ISI functions are highly similar to the ones mea- sured during waking and singing, respectively.
****
From Batch to Transductive Online Learning
Sham Kakade, Adam Tauman Kalai
It is well-known that everything that is learnable in the dif■cult online setting, where an arbitrary sequences of examples must be labeled one at a time, is also learnable in the batch setting, where examples are drawn independently from a distribution. We show a result in the opposite di- rection. We give an ef■cient

conversion algorithm from batch to online that is transductive: it uses future unlabeled data. This demonstrates the equivalence between what is properly and e f■ciently learnable in a batch model and a transductive online model.
****

Learning Depth from Single Monocular Images
Ashutosh Saxena, Sung Chung, Andrew Ng
We consider the task of depth estimation from a single monocular image. We take a supervised learning approach to this problem, in which we begin by collecting a training set of monocular images (of unstructured outdoor environments which i nclude forests, trees, buildings, etc.) and their corresponding ground-truth dep thmaps. Then, we apply supervised learning to predict the depthmap as a function of the image. Depth estimation is a challenging problem, since local features a lone are insufficient to estimate depth at a point, and one needs to consider th e global context of the image. Our model uses a discriminatively-trained Markov Random Field (MRF) that incorporates multiscale local- and global-image features , and models both depths at individual points as well as the relation between de pths at different points. We show that, even on unstructured scenes, our algorit hm is frequently able to recover fairly accurate depthmaps.
****

Non-Local Manifold Parzen Windows
Yoshua Bengio, Hugo Larochelle, Pascal Vincent
To escape from the curse of dimensionality, we claim that one can learn non-loca l functions, in the sense that the value and shape of the learned function at x must be inferred using examples that may be far from x. With this objective, we present a non-local non-parametric density esti- mator. It builds upon previousl y proposed Gaussian mixture models with regularized covariance matrices to take into account the local shape of the manifold. It also builds upon recent work on non-local estimators of the tangent plane of a manifold, which are able to gene ralize in places with little training data, unlike traditional, local, non-param etric models.
****

Bayesian Sets
Zoubin Ghahramani, Katherine A. Heller
Inspired by “Google™ Sets”, we consider the problem of retrieving items from a c oncept or cluster, given a query consisting of a few items from that cluster. We formulate this as a Bayesian inference problem and de- scribe a very simple alg orithm for solving it. Our algorithm uses a model- based concept of a cluster an d ranks items using a score which evaluates the marginal probability that each i tem belongs to a cluster containing the query items. For exponential family mode ls with conjugate priors this marginal probability is a simple function of suf■c ient statistics. We focus on sparse binary data and show that our score can be e valuated ex- actly using a single sparse matrix multiplication, making it possib le to apply our algorithm to very large datasets. We evaluate our algorithm on t hree datasets: retrieving movies from EachMovie, ■nding completions of author se ts from the NIPS dataset, and ■nding completions of sets of words appearing in t he Grolier encyclopedia. We compare to Google™ Sets and show that Bayesian Sets gives very reasonable set completions.
****

Modeling Neuronal Interactivity using Dynamic Bayesian Networks
Lei Zhang, Dimitris Samaras, Nelly Alia-klein, Nora Volkow, Rita Goldstein
Functional Magnetic Resonance Imaging (fMRI) has enabled scientists to look into the active brain. However, interactivity between functional brain regions, is s till little studied. In this paper, we contribute a novel framework for modeling the interactions between multiple active brain regions, using Dynamic Bayesian Networks (DBNs) as generative mod- els for brain activation patterns. This frame work is applied to modeling of neuronal circuits associated with reward. The nov elty of our frame- work from a Machine Learning perspective lies in the use of D BNs to reveal the brain connectivity and interactivity. Such interactivity mod- els which are derived from fMRI data are then validated through a group classi■c ation task. We employ and compare four different types of DBNs: Parallel Hidden

Markov Models, Coupled Hidden Markov Models, Fully-linked Hidden Markov Models and Dynamically Multi- Linked HMMs (DML-HMM). Moreover, we propose and compare two schemes of learning DML-HMMs. Experimental results show that by using DBNs, group classi■cation can be performed even if the DBNs are constructed from as few as 5 brain regions. We also demonstrate that, by using the proposed learning algorithms, different DBN structures charac- terize drug addicted subjects vs. control subjects. This ■nding provides an independent test for the effect of psychopathology on brain function. In general, we demonstrate that incorporation of computer science prin- ciples into functional neuroimaging clinical studies provides a novel ap- proach for probing human brain function.
****
A Theoretical Analysis of Robust Coding over Noisy Overcomplete Channels
Eizaburo Doi, Doru Balcan, Michael Lewicki
Biological sensory systems are faced with the problem of encoding a high-fidelity sensory signal with a population of noisy, low-fidelity neurons. This problem can be expressed in information theoretic terms as coding and transmitting a multi-dimensional, analog signal over a set of noisy channels. Previously, we have shown that robust, overcomplete codes can be learned by minimizing the reconstruction error with a constraint on the channel capacity. Here, we present a theoretical analysis that characterizes the optimal linear coder and decoder for one- and twodimensional data. The analysis allows for an arbitrary number of coding units, thus including both under- and over-complete representations, and provides a number of important insights into optimal coding strategies. In particular, we show how the form of the code adapts to the number of coding units and to different data and noise conditions to achieve robustness. We also report numerical solutions for robust coding of highdimensional image data and show that these codes are substantially more robust compared against other image codes such as ICA and wavelets.
****
A Probabilistic Approach for Optimizing Spectral Clustering
Rong Jin, Feng Kang, Chris Ding
Spectral clustering enjoys its success in both data clustering and semisupervised learning. But, most spectral clustering algorithms cannot handle multi-class clustering problems directly. Additional strategies are needed to extend spectral clustering algorithms to multi-class clustering problems. Furthermore, most spectral clustering algorithms employ hard cluster membership, which is likely to be trapped by the local optimum. In this paper, we present a new spectral clustering algorithm, named "Soft Cut". It improves the normalized cut algorithm by introducing soft membership, and can be efficiently computed using a bound optimization algorithm. Our experiments with a variety of datasets have shown the promising performance of the proposed clustering algorithm.
****
Learning Multiple Related Tasks using Latent Independent Component Analysis
Jian Zhang, Zoubin Ghahramani, Yiming Yang
We propose a probabilistic model based on Independent Component Analysis for learning multiple related tasks. In our model the task parameters are assumed to be generated from independent sources which account for the relatedness of the tasks. We use Laplace distributions to model hidden sources which makes it possible to identify the hidden, independent components instead of just modeling correlations. Furthermore, our model enjoys a sparsity property which makes it both parsimonious and robust. We also propose efficient algorithms for both empirical Bayes method and point estimation. Our experimental results on two multi-label text classification data sets show that the proposed approach is promising.
****
Context as Filtering
Daichi Mochihashi, Yuji Matsumoto
Long-distance language modeling is important not only in speech recognition and machine translation, but also in high-dimensional discrete sequence modeling in general. However, the problem of context length has almost been neglected so far and a nave bag-of-words history has been i employed in natural language process

ing. In contrast, in this paper we view topic shifts within a text as a latent stochastic process to give an explicit probabilistic generative model that has partial exchangeability. We propose an online inference algorithm using particle filters to recognize topic shifts to employ the most appropriate length of context automatically. Experiments on the BNC corpus showed consistent improvement over previous methods involving no chronological order.
****
A matching pursuit approach to sparse Gaussian process regression
Sathiya Keerthi, Wei Chu
In this paper we propose a new basis selection criterion for building sparse GP regression models that provides promising gains in accuracy as well as efficiency over previous methods. Our algorithm is much faster than that of Smola and Bartlett, while, in generalization it greatly outperforms the information gain approach proposed by Seeger et al, especially on the quality of predictive distributions.
****
Phase Synchrony Rate for the Recognition of Motor Imagery in Brain-Computer Interface
Le Song, Evian Gordon, Elly Gysels
Motor imagery attenuates EEG and rhythms over sensorimotor cortices. These amplitude changes are most successfully captured by the method of Common Spatial Patterns (CSP) and widely used in braincomputer interfaces (BCI). BCI methods based on amplitude information, however, have not incoporated the rich phase dynamics in the EEG rhythm. This study reports on a BCI method based on phase synchrony rate (SR). SR, computed from binarized phase locking value, describes the number of discrete synchronization events within a window. Statistical nonparametric tests show that SRs contain significant differences between 2 types of motor imageries. Classifiers trained on SRs consistently demonstrate satisfactory results for all 5 subjects. It is further observed that, for 3 subjects, phase is more discriminative than amplitude in the first 1.5-2.0 s, which suggests that phase has the potential to boost the information transfer rate in BCIs.
****
Inference with Minimal Communication: a Decision-Theoretic Variational Approach
O. Kreidl, Alan Willsky
Given a directed graphical model with binary-valued hidden nodes and real-valued noisy observations, consider deciding upon the maximum a-posteriori (MAP) or the maximum posterior-marginal (MPM) assignment under the restriction that each node broadcasts only to its children exactly one single-bit message. We present a variational formulation, viewing the processing rules local to all nodes as degrees-of-freedom, that minimizes the loss in expected (MAP or MPM) performance subject to such online communication constraints. The approach leads to a novel message-passing algorithm to be executed offline, or before observations are realized, which mitigates the performance loss by iteratively coupling all rules in a manner implicitly driven by global statistics. We also provide (i) illustrative examples, (ii) assumptions that guarantee convergence and efficiency and (iii) connections to active research areas.
****
Variational EM Algorithms for Non-Gaussian Latent Variable Models
Jason Palmer, Kenneth Kreutz-Delgado, Bhaskar Rao, David Wipf
We consider criteria for variational representations of non-Gaussian latent variables, and derive variational EM algorithms in general form. We establish a general equivalence among convex bounding methods, evidence based methods, and ensemble learning/Variational Bayes methods, which has previously been demonstrated only for particular cases.
****
Fast Information Value for Graphical Models
Brigham S. Anderson, Andrew Moore
Calculations that quantify the dependencies between variables are vital to many operations with graphical models, e.g., active learning and sen- sitivity analysis. Previously, pairwise information gain calculation has involved a cost quadra

tic in network size. In this work, we show how to perform a similar computation with cost linear in network size. The loss function that allows this is of a for m amenable to computation by dynamic programming. The message-passing algorithm that results is described and empirical results demonstrate large speedups witho ut de- crease in accuracy. In the cost-sensitive domains examined, superior ac- curacy is achieved.
****
Value Function Approximation with Diffusion Wavelets and Laplacian Eigenfunction s
Sridhar Mahadevan, Mauro Maggioni
We investigate the problem of automatically constructing ef■cient rep- resentati ons or basis functions for approximating value functions based on analyzing the structure and topology of the state space. In particu- lar, two novel approaches to value function approximation are explored based on automatically constructin g basis functions on state spaces that can be represented as graphs or manifolds : one approach uses the eigen- functions of the Laplacian, in effect performing a global Fourier analysis on the graph; the second approach is based on diffusio n wavelets, which generalize classical wavelets to graphs using multiscale dilat ions induced by powers of a diffusion operator or random walk on the graph. Toge ther, these approaches form the foundation of a new generation of methods for so lving large Markov decision processes, in which the underlying repre- sentation and policies are simultaneously learned.
****
A Bayesian Spatial Scan Statistic
Daniel Neill, Andrew Moore, Gregory Cooper
We propose a new Bayesian method for spatial cluster detection, the “Bayesian sp atial scan statistic,” and compare this method to the standard (frequentist) sca n statistic approach. We demonstrate that the Bayesian statistic has several adv antages over the frequentist approach, including increased power to detect clust ers and (since randomization testing is unnecessary) much faster runtime. We eva luate the Bayesian and fre- quentist methods on the task of prospective disease surveillance: detect- ing spatial clusters of disease cases resulting from emerg ing disease out- breaks. We demonstrate that our Bayesian methods are successful in rapidly detecting outbreaks while keeping number of false positives low.
****
Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operator s
Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, Ronald Coifman
This paper presents a diffusion based probabilistic interpretation of spectral c lustering and dimensionality reduction algorithms that use the eigenvectors of t he normalized graph Laplacian. Given the pairwise adjacency matrix of all points , we define a diffusion distance between any two data points and show that the l ow dimensional representation of the data by the first few eigenvectors of the c orresponding Markov matrix is optimal under a certain mean squared error criteri on. Furthermore, assuming that data points are random samples from a density p(x) = e-U (x) we identify these eigenvectors as discrete approximations of eigenfu nctions of a Fokker-Planck operator in a potential 2U (x) with reflecting bounda ry conditions. Finally, applying known results regarding the eigenvalues and eig enfunctions of the continuous Fokker-Planck operator, we provide a mathematical justification for the success of spectral clustering and dimensional reduction a lgorithms based on these first few eigenvectors. This analysis elucidates, in te rms of the characteristics of diffusion processes, many empirical findings regar ding spectral clustering algorithms. Keywords: Algorithms and architectures, lea rning theory.
****
Scaling Laws in Natural Scenes and the Inference of 3D Shape
Tai-sing Lee, Brian Potetz
This paper explores the statistical relationship between natural images and thei r underlying range (depth) images. We look at how this relationship changes over scale, and how this information can be used to enhance low resolution range dat

a using a full resolution intensity image. Based on our findings, we propose an extension to an existing technique known as shape recipes [3], and the success o f the two methods are compared using images and laser scans of real scenes. Our extension is shown to provide a two-fold improvement over the current method. Fu rthermore, we demonstrate that ideal linear shape-from-shading filters, when lea rned from natural scenes, may derive even more strength from shadow cues than fr om the traditional linear-Lambertian shading cues.
****
On the Convergence of Eigenspaces in Kernel Principal Component Analysis
Laurent Zwald, Gilles Blanchard
This paper presents a non-asymptotic statistical analysis of Kernel-PCA with a f ocus different from the one proposed in previous work on this topic. Here instea d of considering the reconstruction error of KPCA we are interested in approxima tion error bounds for the eigenspaces themselves. We prove an upper bound depend ing on the spacing between eigenvalues but not on the dimensionality of the eige nspace. As a consequence this allows to infer stability results for these estima ted spaces.
****
Layered Dynamic Textures
Antoni Chan, Nuno Vasconcelos
A dynamic texture is a video model that treats a video as a sample from a spatio -temporal stochastic process, speci■cally a linear dynamical sys- tem. One probl em associated with the dynamic texture is that it cannot model video where there are multiple regions of distinct motion. In this work, we introduce the layered dynamic texture model, which addresses this problem. We also introduce a varian t of the model, and present the EM algorithm for learning each of the models. Fi nally, we demonstrate the ef■cacy of the proposed model for the tasks of segment ation and syn- thesis of video.
****
Subsequence Kernels for Relation Extraction
Raymond Mooney, Razvan Bunescu
We present a new kernel method for extracting semantic relations between entitie s in natural language text, based on a generalization of subsequence kernels. Th is kernel uses three types of subsequence patterns that are typically employed i n natural language to assert relationships between two entities. Experiments on extracting protein interactions from biomedical corpora and top-level relations from newspaper corpora demonstrate the advantages of this approach.
****
Optimal cue selection strategy
Vidhya Navalpakkam, Laurent Itti
Survival in the natural world demands the selection of relevant visual cues to r apidly and reliably guide attention towards prey an d predators in cluttered env ironments. We investigate whether our visu al system selects cues that guide sea rch in an optimal manner. We formall y obtain the optimal cue selection strategy by maximizing the signal to noise ratio (S N R) between a search target and sur rounding distractors. This optimal strategy successfully accounts for several ph enom ena in visual search behavior, including the effect of target-distracto r d iscriminability, uncertainty in target's features, distractor heterogenei ty, an d linear separability. Furthermore, the theory generates a new predict ion, whic h we verify through psychophysical experiments with human subj ects. Our results provide direct experimental evidence that humans sel ect visual cues so as to m aximize S N R between the targets and surrounding clutter.
****
Infinite latent feature models and the Indian buffet process
Zoubin Ghahramani, Thomas Griffiths
We define a probability distribution over equivalence classes of binary matrices with a finite number of rows and an unbounded number of columns. This distribut ion is suitable for use as a prior in probabilistic models that represent object s using a potentially infinite array of features. We identify a simple generativ e process that results in the same distribution over equivalence classes, which

we call the Indian buffet process. We illustrate the use of this distribution as a prior in an infinite latent feature model, deriving a Markov chain Monte Carlo algorithm for inference in this model and applying the algorithm to an image dataset.
****
Non-Gaussian Component Analysis: a Semi-parametric Framework for Linear Dimension Reduction
Gilles Blanchard, Masashi Sugiyama, Motoaki Kawanabe, Vladimir Spokoiny, Klaus-Robert Müller
We propose a new linear method for dimension reduction to identify nonGaussian components in high dimensional data. Our method, NGCA (non-Gaussian component analysis), uses a very general semi-parametric framework. In contrast to existing projection methods we define what is uninteresting (Gaussian): by projecting out uninterestingness, we can estimate the relevant non-Gaussian subspace. We show that the estimation error of finding the non-Gaussian components tends to zero at a parametric rate. Once NGCA components are identified and extracted, various tasks can be applied in the data analysis process, like data visualization, clustering, denoising or classification. A numerical study demonstrates the usefulness of our method.
****
Divergences, surrogate loss functions and experimental design
XuanLong Nguyen, Martin J. Wainwright, Michael Jordan
In this paper, we provide a general theorem that establishes a correspon- dence between surrogate loss functions in classi■cation and the family of f-divergences. Moreover, we provide constructive procedures for determining the f-divergence induced by a given surrogate loss, and conversely for ■nding all surrogate loss functions that realize a given f-divergence. Next we introduce the notion of universal equivalence among loss functions and corresponding f-divergences, and provide nec- essary and suf■cient conditions for universal equivalence to hold. These ideas have applications to classi■cation problems that also involve a com- ponent of experiment design; in particular, we leverage our results to prove consistency of a procedure for learning a classi■er under decen- tralization requirements.
****
Mixture Modeling by Affinity Propagation
Brendan J. Frey, Delbert Dueck
Clustering is a fundamental problem in machine learning and has been approached in many ways. Two general and quite different approaches include iteratively ■tting a mixture model (e.g., using EM) and linking to- gether pairs of training cases that have high af■nity (e.g., using spectral methods). Pair-wise clustering algorithms need not compute suf■cient statistics and avoid poor solutions by directly placing similar examples in the same cluster. However, many applications require that each cluster of data be accurately described by a prototype or model, so af■nity-based clustering – and its bene■ts – cannot be directly realized. We describe a technique called “af■nity propagation”, which combines the advantages of both approaches. The method learns a mixture model of the data by recursively propagating af■nity messages. We demonstrate af■nity prop- agation on the problems of clustering image patches for image segmen- tation and learning mixtures of gene expression models from microar- ray data. We ■nd that af■nity propagation obtains better solutions than mixtures of Gaussians, the K-medoids algorithm, spectral clustering and hierarchical clustering, and is both able to ■nd a pre-speci■ed number of clusters and is able to automatically determine the number of clusters. Interestingly, af■nity propagation can be viewed as belief propagation in a graphical model that accounts for pairwise training case likelihood functions and the identi■cation of cluster centers.
****
Tensor Subspace Analysis
Xiaofei He, Deng Cai, Partha Niyogi
Previous work has demonstrated that the image variations of many ob- jects (human faces in particular) under variable lighting can be effec- tively modeled by l

ow dimensional linear spaces. The typical linear sub- space learning algorithms include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Locality Preserving Projec- tion (LPP). All of these methods consider an n1 × n2 image as a high dimensional vector in Rn1×n2, while an image represented in the plane is intrinsically a matrix. In this paper, we propose a new algorithm called Tensor Subspace Analysis (TSA). TSA considers an image as the sec- ond order tensor in Rn1 ⊗ Rn2, where Rn1 and Rn2 are two vector spaces. The relationship between the column vectors of the image ma- trix and that between the row vectors can be naturally characterized by TSA. TSA detects the intrinsic local geometrical structure of the tensor space by learning a lower dimensional tensor subspace. We compare our proposed approach with PCA, LDA and LPP methods on two standard databases. Experimental results demonstrate that TSA achieves better recognition rate, while being much more ef■cient.
****
Query by Committee Made Real
Ran Gilad-bachrach, Amir Navot, Naftali Tishby
Training a learning algorithm is a costly task. A major goal of active learning is to reduce this cost. In this paper we introduce a new algorithm, KQBC, which is capable of actively learning large scale problems by using selective sampling. The algorithm overcomes the costly sampling step of the well known Query By Committee (QBC) algorithm by projecting onto a low dimensional space. KQBC also enables the use of kernels, providing a simple way of extending QBC to the non-linear scenario. Sampling the low dimension space is done using the hit and run random walk. We demonstrate the success of this novel algorithm by applying it to both artificial and a real world problems.
****
An Application of Markov Random Fields to Range Sensing
James Diebel, Sebastian Thrun
This paper describes a highly successful application of MRFs to the problem of generating high-resolution range images. A new generation of range sensors combines the capture of low-resolution range images with the acquisition of registered high-resolution camera images. The MRF in this paper exploits the fact that discontinuities in range and coloring tend to co-align. This enables it to generate high-resolution, low-noise range images by integrating regular camera images into the range data. We show that by using such an MRF, we can substantially improve over existing range imaging technology.
****
Analysis of Spectral Kernel Design based Semi-supervised Learning
Tong Zhang, Rie Kubota Ando
We consider a framework for semi-supervised learning using spectral decomposition based un-supervised kernel design. This approach sub- sumes a class of previously proposed semi-supervised learning methods on data graphs. We examine various theoretical properties of such meth- ods. In particular, we derive a generalization performance bound, and obtain the optimal kernel design by minimizing the bound. Based on the theoretical analysis, we are able to demonstrate why spectral kernel design based methods can often improve the predictive performance. Ex- periments are used to illustrate the main consequences of our analysis.
****
Representing Part-Whole Relationships in Recurrent Neural Networks
Viren Jain, Valentin Zhigulin, H. Seung
There is little consensus about the computational function of top-down synaptic connections in the visual system. Here we explore the hypothesis that top-down connections, like bottom-up connections, reflect partwhole relationships. We analyze a recurrent network with bidirectional synaptic interactions between a layer of neurons representing parts and a layer of neurons representing wholes. Within each layer, there is lateral inhibition. When the network detects a whole, it can rigorously enforce part-whole relationships by ignoring parts that do not belong. The network can complete the whole by filling in missing parts. The network can refuse to recognize a whole, if the activated parts do not conform to a stored part-whole relationship. Parameter regimes in which these behaviors happen

are identified using the theory of permitted and forbidden sets [3, 4]. The netw ork behaviors are illustrated by recreating Rumelhart and McClelland's "interact ive activation" model [7]. In neural network models of visual object recognition [2, 6, 8], patterns of synaptic connectivity often reflect part-whole relations hips between the features that are represented by neurons. For example, the conn ections of Figure 1 reflect the fact that feature B both contains simpler featur es A1, A2, and A3, and is contained in more complex features C1, C2, and C3. Suc h connectivity allows neurons to follow the rule that existence of the part is e vidence for existence of the whole. By combining synaptic input from multiple so urces of evidence for a feature, a neuron can "decide" whether that feature is p resent. 1 The synapses shown in Figure 1 are purely bottom-up, directed from sim ple to complex features. However, there are also top-down connections in the vis ual system, and there is little consensus about their function. One possibility is that top-down connections also reflect part-whole relationships. They allow f eature detectors to make decisions using the rule that existence of the whole is evidence for existence of its parts. In this paper, we analyze the dynamics of a recurrent network in which part-whole relationships are stored as bidirectiona l synaptic interactions, rather than the unidirectional interactions of Figure 1 . The network has a number of interesting computational capabilities. When the n etwork detects a whole, it can rigorously enforce part-whole relationships Synap tic connectivity may reflect other relationships besides part-whole. For example , invariances can be implemented by connecting detectors of several instances of the same feature to the same target, which is consequently an invariant detecto r of the feature. 1
****
Size Regularized Cut for Data Clustering
Yixin Chen, Ya Zhang, Xiang Ji
We present a novel spectral clustering method that enables users to incorporate prior knowledge of the size of clusters into the clustering process. The cost fu nction, which is named size regularized cut (SRcut), is defined as the sum of th e inter-cluster similarity and a regularization term measuring the relative size of two clusters. Finding a partition of the data set to minimize SRcut is prove d to be NP-complete. An approximation algorithm is proposed to solve a relaxed v ersion of the optimization problem as an eigenvalue problem. Evaluations over di fferent data sets demonstrate that the method is not sensitive to outliers and p erforms better than normalized cut.
****
How fast to work: Response vigor, motivation and tonic dopamine
Yael Niv, Nathaniel Daw, Peter Dayan
Reinforcement learning models have long promised to unify computa- tional, psych ological and neural accounts of appetitively conditioned be- havior. However, th e bulk of data on animal conditioning comes from free-operant experiments measur ing how fast animals will work for rein- forcement. Existing reinforcement learn ing (RL) models are silent about these tasks, because they lack any notion of vi gor. They thus fail to ad- dress the simple observation that hungrier animals wi ll work harder for food, as well as stranger facts such as their sometimes great er produc- tivity even when working for irrelevant outcomes such as water. Here, we develop an RL framework for free-operant behavior, suggesting that subjects choose how vigorously to perform selected actions by optimally balancing the cos ts and bene■ts of quick responding. Motivational states such as hunger shift the se factors, skewing the tradeoff. This accounts normatively for the effects of m otivation on response rates, as well as many other classic ■ndings. Finally, we suggest that tonic levels of dopamine may be involved in the computation linking motivational state to optimal responding, thereby explaining the complex vigor-related ef- fects of pharmacological manipulation of dopamine.
****
Robust design of biological experiments
Patrick Flaherty, Adam Arkin, Michael Jordan
We address the problem of robust, computationally-efficient design of biological experiments. Classical optimal experiment design methods have not been widely a

dopted in biological practice, in part because the resulting designs can be very brittle if the nominal parameter estimates for the model are poor, and in part because of computational constraints. We present a method for robust experiment design based on a semidefinite programming relaxation. We present an application of this method to the design of experiments for a complex calcium signal transduction pathway, where we have found that the parameter estimates obtained from the robust design are better than those obtained from an "optimal" design.
****

On the Accuracy of Bounded Rationality: How Far from Optimal Is Fast and Frugal?
Michael Schmitt, Laura Martignon
Fast and frugal heuristics are well studied models of bounded rationality. Psychological research has proposed the take-the-best heuristic as a successful strategy in decision making with limited resources. Take-thebest searches for a sufficiently good ordering of cues (features) in a task where objects are to be compared lexicographically. We investigate the complexity of the problem of approximating optimal cue permutations for lexicographic strategies. We show that no efficient algorithm can approximate the optimum to within any constant factor, if P = NP. We further consider a greedy approach for building lexicographic strategies and derive tight bounds for the performance ratio of a new and simple algorithm. This algorithm is proven to perform better than take-the-best.
****

Assessing Approximations for Gaussian Process Classification
Malte Kuss, Carl Rasmussen
Gaussian processes are attractive models for probabilistic classification but unfortunately exact inference is analytically intractable. We compare Laplace's method and Expectation Propagation (EP) focusing on marginal likelihood estimates and predictive performance. We explain theoretically and corroborate empirically that EP is superior to Laplace. We also compare to a sophisticated MCMC scheme and show that EP is surprisingly accurate. In recent years models based on Gaussian process (GP) priors have attracted much attention in the machine learning community. Whereas inference in the GP regression model with Gaussian noise can be done analytically, probabilistic classification using GPs is analytically intractable. Several approaches to approximate Bayesian inference have been suggested, including Laplace's approximation, Expectation Propagation (EP), variational approximations and Markov chain Monte Carlo (MCMC) sampling, some of these in conjunction with generalisation bounds, online learning schemes and sparse approximations. Despite the abundance of recent work on probabilistic GP classifiers, most experimental studies provide only anecdotal evidence, and no clear picture has yet emerged, as to when and why which algorithm should be preferred. Thus, from a practitioners point of view probabilistic GP classification remains a jungle. In this paper, we set out to understand and compare two of the most wide-spread approximations: Laplace's method and Expectation Propagation (EP). We also compare to a sophisticated, but computationally demanding MCMC scheme to examine how close the approximations are to ground truth. We examine two aspects of the approximation schemes: Firstly the accuracy of approximations to the marginal likelihood which is of central importance for model selection and model comparison. In any practical application of GPs in classification (usually multiple) parameters of the covariance function (hyperparameters) have to be handled. Bayesian model selection provides a consistent framework for setting such parameters. Therefore, it is essential to evaluate the accuracy of the marginal likelihood approximations as a function of the hyperparameters, in order to assess the practical usefulness of the approach Secondly, we need to assess the quality of the approximate probabilistic predictions. In the past, the probabilistic nature of the GP predictions have not received much attention, the focus being mostly on classification error rates. This unfortunate state of affairs is caused primarily by typical benchmarking problems being considered outside of a realistic context. The ability of a classifier to produce class probabilities or confidences, have obvious
****

Fast Online Policy Gradient Learning with SMD Gain Vector Adaptation

Jin Yu, Douglas Aberdeen, Nicol Schraudolph
Reinforcement learning by direct policy gradient estimation is attractive in theory but in practice leads to notoriously ill-behaved optimization problems. We improve its robustness and speed of convergence with stochastic meta-descent, a gain vector adaptation method that employs fast Hessian-vector products. In our experiments the resulting algorithms outperform previously employed online stochastic, of■ine conjugate, and natural policy gradient methods.
****
A Bayes Rule for Density Matrices
Manfred K. K. Warmuth
The classical Bayes rule computes the posterior model probability from the prior probability and the data likelihood. We generalize this rule to the case when the prior is a density matrix (symmetric positive definite and trace one) and the data likelihood a covariance matrix. The classical Bayes rule is retained as the special case when the matrices are diagonal. In the classical setting, the calculation of the probability of the data is an expected likelihood, where the expectation is over the prior distribution. In the generalized setting, this is replaced by an expected variance calculation where the variance is computed along the eigenvectors of the prior density matrix and the expectation is over the eigenvalues of the density matrix (which form a probability vector). The variances along any direction is determined by the covariance matrix. Curiously enough this expected variance calculation is a quantum measurement where the co-variance matrix specifies the instrument and the prior density matrix the mixture state of the particle. We motivate both the classical and the generalized Bayes rule with a minimum relative entropy principle, where the Kullbach-Leibler version gives the classical Bayes rule and Umegaki's quantum relative entropy the new Bayes rule for density matrices.
****
Combining Graph Laplacians for Semi--Supervised Learning
Andreas Argyriou, Mark Herbster, Massimiliano Pontil
A foundational problem in semi-supervised learning is the construction of a graph underlying the data. We propose to use a method which optimally combines a number of differently constructed graphs. For each of these graphs we associate a basic graph kernel. We then compute an optimal combined kernel. This kernel solves an extended regularization problem which requires a joint minimization over both the data and the set of graph kernels. We present encouraging results on different OCR tasks where the optimal combined kernel is computed from graphs constructed with a variety of distances functions and the `k ' in nearest neighbors.
****
Integrate-and-Fire models with adaptation are good enough
Renaud Jolivet, Alexander Rauch, Hans-rudolf Lüscher, Wulfram Gerstner
Integrate-and-Fire-type models are usually criticized because of their simplicity. On the other hand, the Integrate-and-Fire model is the basis of most of the theoretical studies on spiking neuron models. Here, we develop a sequential procedure to quantitatively evaluate an equivalent Integrate-and-Fire-type model based on intracellular recordings of cortical pyramidal neurons. We find that the resulting effective model is sufficient to predict the spike train of the real pyramidal neuron with high accuracy. In in vivo-like regimes, predicted and recorded traces are almost indistinguishable and a significant part of the spikes can be predicted at the correct timing. Slow processes like spike-frequency adaptation are shown to be a key feature in this context since they are necessary for the model to connect between different driving regimes.
****
Large-Scale Multiclass Transduction
Thomas Gärtner, Quoc Le, Simon Burton, Alex J. Smola, Vishy Vishwanathan
We present a method for performing transductive inference on very large datasets. Our algorithm is based on multiclass Gaussian processes and is effective whenever the multiplication of the kernel matrix or its inverse with a vector can be computed suf■ciently fast. This holds, for instance, for certain graph and string kernels. Transduction is achieved by varia- tional inference over the unlabele

d data subject to a balancing constraint.
****
Sparse Gaussian Processes using Pseudo-inputs
Edward Snelson, Zoubin Ghahramani
We present a new Gaussian process (GP) regression model whose covariance is parameterized by the the locations of M pseudo-input points, which we learn by a gradient based optimization. We take M N, where N is the number of real data points, and hence obtain a sparse regression method which has O(M 2 N) training cost and O(M 2) prediction cost per test case. We also find hyperparameters of the covariance function in the same joint optimization. The method can be viewed as a Bayesian regression model with particular input dependent noise. The method turns out to be closely related to several other sparse GP approaches, and we discuss the relation in detail. We finally demonstrate its performance on some large data sets, and make a direct comparison to other sparse GP methods. We show that our method can match full GP performance with small M , i.e. very sparse solutions, and it significantly outperforms other approaches in this regime.
****
Fast biped walking with a reflexive controller and real-time policy searching
Tao Geng, Bernd Porr, Florentin Wörgötter
In this paper, we present our design and experiments of a planar biped robot ("RunBot") under pure reflexive neuronal control. The goal of this study is to combine neuronal mechanisms with biomechanics to obtain very fast speed and the on-line learning of circuit parameters. Our controller is built with biologically inspired sensor- and motor-neuron models, including local reflexes and not employing any kind of position or trajectory-tracking control algorithm. Instead, this reflexive controller allows RunBot to exploit its own natural dynamics during critical stages of its walking gait cycle. To our knowledge, this is the first time that dynamic biped walking is achieved using only a pure reflexive controller. In addition, this structure allows using a policy gradient reinforcement learning algorithm to tune the parameters of the reflexive controller in real-time during walking. This way RunBot can reach a relative speed of 3.5 leg-lengths per second after a few minutes of online learning, which is faster than that of any other biped robot, and is also comparable to the fastest relative speed of human walking. In addition, the stability domain of stable walking is quite large supporting this design strategy.
****
Learning from Data of Variable Quality
Koby Crammer, Michael Kearns, Jennifer Wortman
We initiate the study of learning from multiple sources of limited data, each of which may be corrupted at a different rate. We develop a com- plete theory of which data sources should be used for two fundamental problems: estimating the bias of a coin, and learning a classi■er in the presence of label noise. In both cases, ef■cient algorithms are provided for computing the optimal subset of data.
****
Two view learning: SVM-2K, Theory and Practice
Jason Farquhar, David Hardoon, Hongying Meng, John Shawe-taylor, Sándor Szedmák
Kernel methods make it relatively easy to define complex highdimensional feature spaces. This raises the question of how we can identify the relevant subspaces for a particular learning task. When two views of the same phenomenon are available kernel Canonical Correlation Analysis (KCCA) has been shown to be an effective preprocessing step that can improve the performance of classification algorithms such as the Support Vector Machine (SVM). This paper takes this observation to its logical conclusion and proposes a method that combines this two stage learning (KCCA followed by SVM) into a single optimisation termed SVM-2K. We present both experimental and theoretical analysis of the approach showing encouraging results and insights.
****
Extracting Dynamical Structure Embedded in Neural Activity
Byron M. Yu, Afsheen Afshar, Gopal Santhanam, Stephen Ryu, Krishna V. Shenoy, Maneesh Sahani

Spiking activity from neurophysiological experiments often exhibits dy- namics beyond that driven by external stimulation, presumably re■ect- ing the extensive recurrence of neural circuitry. Characterizing these dynamics may reveal important features of neural computation, par- ticularly during internally-driven cognitive operations. For example, the activity of premotor cortex (PMd) neurons during an instructed de- lay period separating movement-target speci■cation and a movement- initiation cue is believed to be involved in motor planning. We show that the dynamics underlying this activity can be captured by a low- dimensional non-linear dynamical systems model, with underlying re- current structure and stochastic point-process output. We present and validate latent variable methods that simultaneously estimate the system parameters and the trial-by-trial dynamical trajectories. These meth- ods are applied to characterize the dynamics in PMd data recorded from a chronically-implanted 96-electrode array while monkeys perform delayed-reach tasks.
****
Large-scale biophysical parameter estimation in single neurons via constrained linear regression
Misha Ahrens, Liam Paninski, Quentin Huys
Our understanding of the input-output function of single cells has been substantially advanced by biophysically accurate multi-compartmental models. The large number of parameters needing hand tuning in these models has, however, somewhat hampered their applicability and interpretability. Here we propose a simple and well-founded method for automatic estimation of many of these key parameters: 1) the spatial distribution of channel densities on the cell's membrane; 2) the spatiotemporal pattern of synaptic input; 3) the channels' reversal potentials; 4) the intercompartmental conductances; and 5) the noise level in each compartment. We assume experimental access to: a) the spatiotemporal voltage signal in the dendrite (or some contiguous subpart thereof, e.g. via voltage sensitive imaging techniques), b) an approximate kinetic description of the channels and synapses present in each compartment, and c) the morphology of the part of the neuron under investigation. The key observation is that, given data a)-c), all of the parameters 1)-4) may be simultaneously inferred by a version of constrained linear regression; this regression, in turn, is efficiently solved using standard algorithms, without any "local minima" problems despite the large number of parameters and complex dynamics. The noise level 5) may also be estimated by standard techniques. We demonstrate the method's accuracy on several model datasets, and describe techniques for quantifying the uncertainty in our estimates.
****
Data-Driven Online to Batch Conversions
Ofer Dekel, Yoram Singer
Online learning algorithms are typically fast, memory efficient, and simple to implement. However, many common learning problems fit more naturally in the batch learning setting. The power of online learning algorithms can be exploited in batch settings by using online-to-batch conversions techniques which build a new batch algorithm from an existing online algorithm. We first give a unified overview of three existing online-to-batch conversion techniques which do not use training data in the conversion process. We then build upon these data-independent conversions to derive and analyze data-driven conversions. Our conversions find hypotheses with a small risk by explicitly minimizing datadependent generalization bounds. We experimentally demonstrate the usefulness of our approach and in particular show that the data-driven conversions consistently outperform the data-independent conversions.
****
Learning Rankings via Convex Hull Separation
Glenn Fung, Rómer Rosales, Balaji Krishnapuram
We propose ef■cient algorithms for learning ranking functions from or- der constraints between sets—i.e. classes—of training samples. Our al- gorithms may be used for maximizing the generalized Wilcoxon Mann Whitney statistic that accounts for the partial ordering of the classes: spe- cial cases include maximizing the area under the ROC curve for binary classi■cation and its generalization for ord

inal regression. Experiments on public benchmarks indicate that: (a) the propose
d algorithm is at least as accurate as the current state-of-the-art; (b) computa
tionally, it is sev- eral orders of magnitude faster and—unlike current methods—
it is easily able to handle even large datasets with over 20,000 samples.
****
Interpolating between types and tokens by estimating power-law generators
Sharon Goldwater, Mark Johnson, Thomas Griffiths
Standard statistical models of language fail to capture one of the most striking
 properties of natural languages: the power-law distribution in the frequencies
of word tokens. We present a framework for developing statistical models that ge
nerically produce power-laws, augmenting stan- dard generative models with an ad
aptor that produces the appropriate pattern of token frequencies. We show that t
aking a particular stochastic process – the Pitman-Yor process – as an adaptor j
usti■es the appearance of type frequencies in formal analyses of natural languag
e, and improves the performance of a model for unsupervised learning of morpholo
gy.
****
Response Analysis of Neuronal Population with Synaptic Depression
Wentao Huang, Licheng Jiao, Shan Tan, Maoguo Gong
In this paper, we aim at analyzing the characteristic of neuronal population res
ponses to instantaneous or time-dependent inputs and the role of synapses in neu
ral information processing. We have derived an evolution equation of the membran
e potential density function with synaptic depression, and obtain the formulas f
or analytic computing the response of instantaneous re rate. Through a technical
 analysis, we arrive at several signi cant conclusions: The background inputs pl
ay an important role in information processing and act as a switch betwee tempor
al integration and coincidence detection. the role of synapses can be regarded a
s a spatio-temporal lter; it is important in neural information processing for t
he spatial distribution of synapses and the spatial and temporal relation of inp
uts. The instantaneous input frequency can affect the response amplitude and pha
se delay.
****
Sequence and Tree Kernels with Statistical Feature Mining
Jun Suzuki, Hideki Isozaki
This paper proposes a new approach to feature selection based on a sta- tistical
 feature mining technique for sequence and tree kernels. Since natural language
data take discrete structures, convolution kernels, such as sequence and tree ke
rnels, are advantageous for both the concept and accuracy of many natural langua
ge processing tasks. However, experi- ments have shown that the best results can
 only be achieved when lim- ited small sub-structures are dealt with by these ke
rnels. This paper dis- cusses this issue of convolution kernels and then propose
s a statistical feature selection that enable us to use larger sub-structures ef
fectively. The proposed method, in order to execute ef■ciently, can be embedded
into an original kernel calculation process by using sub-structure min- ing algo
rithms. Experiments on real NLP tasks con■rm the problem in the conventional met
hod and compare the performance of a conventional method to that of the proposed
 method.
****
Faster Rates in Regression via Active Learning
Rebecca Willett, Robert Nowak, Rui Castro
This paper presents a rigorous statistical analysis characterizing regimes in wh
ich active learning significantly outperforms classical passive learning. Active
 learning algorithms are able to make queries or select sample locations in an o
nline fashion, depending on the results of the previous queries. In some regimes
, this extra flexibility leads to significantly faster rates of error decay than
 those possible in classical passive learning settings. The nature of these regi
mes is explored by studying fundamental performance limits of active and passive
 learning in two illustrative nonparametric function classes. In addition to exa
mining the theoretical potential of active learning, this paper describes a prac
tical algorithm capable of exploiting the extra flexibility of the active settin

g and provably improving upon the classical passive techniques. Our active learn ing theory and methods show promise in a number of applications, including field estimation using wireless sensor networks and fault line detection.
****

Top-Down Control of Visual Attention: A Rational Account
Michael Shettel, Shaun Vecera, Michael C. Mozer
Theories of visual attention commonly posit that early parallel processes extrac t con- spicuous features such as color contrast and motion from the visual field . These features are then combined into a saliency map, and attention is directe d to the most salient regions first. Top-down attentional control is achieved by modulating the contribution of different feature types to the saliency map. A k ey source of data concerning attentional control comes from behavioral studies i n which the effect of recent experience is exam- ined as individuals repeatedly perform a perceptual discrimination task (e.g., “what shape is the odd-colored o bject?”). The robust finding is that repetition of features of recent trials (e. g., target color) facilitates performance. We view this facilitation as an adapt ation to the statistical structure of the environment. We propose a probabilisti c model of the environment that is updated after each trial. Under the assumptio n that attentional control operates so as to make performance more efficient for more likely environmental states, we obtain parsimonious explanations for data from four different experiments. Further, our model provides a rational explanat ion for why the influence of past experience on attentional control is short liv ed.
****

The Role of Top-down and Bottom-up Processes in Guiding Eye Movements during Vis ual Search
Gregory Zelinsky, Wei Zhang, Bing Yu, Xin Chen, Dimitris Samaras
To investigate how top-down (TD) and bottom-up (BU) information is weighted in t he guidance of human search behavior, we manipulated the proportions of BU and T D components in a saliency-based model. The model is biologically plausible and implements an arti■cial retina and a neuronal population code. The BU component is based on feature- contrast. The TD component is de■ned by a feature-template match to a stored target representation. We compared the model’s behavior at dif fer- ent mixtures of TD and BU components to the eye movement behavior of human observers performing the identical search task. We found that a purely TD model provides a much closer match to human behavior than any mixture model using BU i nformation. Only when biological con- straints are removed (e.g., eliminating th e retina) did a BU/TD mixture model begin to approximate human behavior.
****

A Bayesian Framework for Tilt Perception and Confidence
Odelia Schwartz, Peter Dayan, Terrence J. Sejnowski
The misjudgement of tilt in images lies at the heart of entertaining visual illu sions and rigorous perceptual psychophysics. A wealth of findings has attracted many mechanistic models, but few clear computational principles. We adopt a Baye sian approach to perceptual tilt estimation, showing how a smoothness prior offe rs a powerful way of addressing much confusing data. In particular, we faithfull y model recent results showing that confidence in estimation can be systematical ly affected by the same aspects of images that affect bias. Confidence is centra l to Bayesian modeling approaches, and is applicable in many other perceptual do mains. Perceptual anomalies and illusions, such as the misjudgements of motion a nd tilt evident in so many psychophysical experiments, have intrigued researcher s for decades.13 A Bayesian view48 has been particularly influential in models o f motion processing, treating such anomalies as the normative product of prior i nformation (often statistically codifying Gestalt laws) with likelihood informat ion from the actual scenes presented. Here, we expand the range of statistically normative accounts to tilt estimation, for which there are classes of results (on estimation confidence) that are so far not available for motion. The tilt ill usion arises when the perceived tilt of a center target is misjudged (ie bias) i n the presence of flankers. Another phenomenon, called Crowding, refers to a los s in the confidence (ie sensitivity) of perceived target tilt in the presence of

flankers. Attempts have been made to formalize these phenomena quantitatively. Crowding has been modeled as compulsory feature pooling (ie averaging of orientations), ignoring spatial positions.9, 10 The tilt illusion has been explained by lateral interactions11, 12 in populations of orientationtuned units; and by calibration.13 However, most models of this form cannot explain a number of crucial aspects of the data. First, the geometry of the positional arrangement of the stimuli affects attraction versus repulsion in bias, as emphasized by Kapadia et al14 (figure 1A), and others.15, 16 Second, Solomon et al. recently measured bias and sensitivity simultaneously.11 The rich and surprising range of sensitivities, far from flat as a function of flanker angles (figure 1B), are outside the reach of standard models. Moreover, current explanations do not offer a computational account of tilt perception as the outcome of a normative inference process. Here, we demonstrate that a Bayesian framework for orientation estimation, with a prior favoring smoothness, can naturally explain a range of seemingly puzzling tilt data. We explicitly consider both the geometry of the stimuli, and the issue of confidence in the esti-
****
Multiple Instance Boosting for Object Detection
Cha Zhang, John Platt, Paul Viola
A good image object detection algorithm is accurate, fast, and does not require exact locations of objects in a training set. We can create such an object detector by taking the architecture of the Viola-Jones detector cascade and training it with a new variant of boosting that we call MILBoost. MILBoost uses cost functions from the Multiple Instance Learning literature combined with the AnyBoost framework. We adapt the feature selection criterion of MILBoost to optimize the performance of the Viola-Jones cascade. Experiments show that the detection rate is up to 1.6 times better using MILBoost. This increased detection rate shows the advantage of simultaneously learning the locations and scales of the objects in the training set along with the parameters of the classifier.
****
Generalization to Unseen Cases
Teemu Roos, Peter Grünwald, Petri Myllymäki, Henry Tirri
We analyze classification error on unseen cases, i.e. cases that are different from those in the training set. Unlike standard generalization error, this off-training-set error may differ significantly from the empirical error with high probability even with large sample sizes. We derive a datadependent bound on the difference between off-training-set and standard generalization error. Our result is based on a new bound on the missing mass, which for small samples is stronger than existing bounds based on Good-Turing estimators. As we demonstrate on UCI data-sets, our bound gives nontrivial generalization guarantees in many practical cases. In light of these results, we show that certain claims made in the No Free Lunch literature are overly pessimistic.
****
Nearest Neighbor Based Feature Selection for Regression and its Application to Neural Activity
Amir Navot, Lavi Shpigelman, Naftali Tishby, Eilon Vaadia
We present a non-linear, simple, yet effective, feature subset selection method for regression and use it in analyzing cortical neural activity. Our algorithm involves a feature-weighted version of the k-nearest-neighbor algorithm. It is able to capture complex dependency of the target func- tion on its input and makes use of the leave-one-out error as a natural regularization. We explain the characteristics of our algorithm on syn- thetic problems and use it in the context of predicting hand velocity from spikes recorded in motor cortex of a behaving monkey. By applying fea- ture selection we are able to improve prediction quality and suggest a novel way of exploring neural data.
****
Visual Encoding with Jittering Eyes
Michele Rucci
Under natural viewing conditions, small movements of the eye and body prevent the maintenance of a steady direction of gaze. It is known that stimuli tend to fa

de when they are stabilized on the retina for several sec- onds. However, it is unclear whether the physiological self-motion of the retinal image serves a visual purpose during the brief periods of natural visual ■xation. This study examines the impact of ■xational instability on the statistics of visual input to the retina and on the structure of neural activity in the early visual system. Fixational instability introduces ■uc- tuations in the retinal input signals that, in the presence of natural images, lack spatial correlations. These input ■uctuations strongly in■uence neu- ral activity in a model of the LGN. They decorrelate cell responses, even if the contrast sensitivity functions of simulated cells are not perfectly tuned to counter-balance the power-law spectrum of natural images. A decorrelation of neural activity has been proposed to be bene■cial for discarding statistical redundancies in the input signals. Fixational insta- bility might, therefore, contribute to establishing ef■cient representations of natural stimuli.
****
Learning to Control an Octopus Arm with Gaussian Process Temporal Difference Methods
Yaakov Engel, Peter Szabo, Dmitry Volkinshtein
The Octopus arm is a highly versatile and complex limb. How the Octo- pus controls such a hyper-redundant arm (not to mention eight of them!) is as yet unknown. Robotic arms based on the same mechanical prin- ciples may render present day robotic arms obsolete. In this paper, we tackle this control problem using an online reinforcement learning al- gorithm, based on a Bayesian approach to policy evaluation known as Gaussian process temporal difference (GPTD) learning. Our substitute for the real arm is a computer simulation of a 2-dimensional model of an Octopus arm. Even with the simpli■cations inherent to this model, the state space we face is a high-dimensional one. We apply a GPTD- based algorithm to this domain, and demonstrate its operation on several learning tasks of varying degrees of dif■culty.
****
Improved risk tail bounds for on-line algorithms
Nicolò Cesa-bianchi, Claudio Gentile
We prove the strongest known bound for the risk of hypotheses selected from the ensemble generated by running a learning algorithm incremen(cid:173) tally on the training data. Our result is based on proof techniques that are remarkably different from the standard risk analysis based on uniform convergence arguments.
****
Kernelized Infomax Clustering
David Barber, Felix Agakov
We propose a simple information-theoretic approach to soft clus- tering based on maximizing the mutual information I(x, y) between the unknown cluster labels y and the training patterns x with re- spect to parameters of speci■cally constrained encoding distribu- tions. The constraints are chosen such that patterns are likely to be clustered similarly if they lie close to speci■c unknown vectors in the feature space. The method may be conveniently applied to learning the optimal a■nity matrix, which corresponds to learn- ing parameters of the kernelized encoder. The procedure does not require computations of eigenvalues of the Gram matrices, which makes it potentially attractive for clustering large data sets.
****
Temporally changing synaptic plasticity
Minija Tamosiunaite, Bernd Porr, Florentin Wörgötter
Recent experimental results suggest that dendritic and back-propagating spikes can influence synaptic plasticity in different ways [1]. In this study we investigate how these signals could temporally interact at dendrites leading to changing plasticity properties at local synapse clusters. Similar to a previous study [2], we employ a differential Hebbian plasticity rule to emulate spike-timing dependent plasticity. We use dendritic (D-) and back-propagating (BP-) spikes as post-synaptic signals in the learning rule and investigate how their interaction will influence plasticity. We will analyze a situation where synapse plasticity c

haracteristics change in the course of time, depending on the type of post-synap tic activity momentarily elicited. Starting with weak synapses, which only elici t local D-spikes, a slow, unspecific growth process is induced. As soon as the s oma begins to spike this process is replaced by fast synaptic changes as the con sequence of the much stronger and sharper BP-spike, which now dominates the plas ticity rule. This way a winner-take-all-mechanism emerges in a two-stage process , enhancing the best-correlated inputs. These results suggest that synaptic plas ticity is a temporal changing process by which the computational properties of d endrites or complete neurons can be substantially augmented.
****
Modeling Neural Population Spiking Activity with Gibbs Distributions
Frank Wood, Stefan Roth, Michael Black
Probabilistic modeling of correlated neural population firing activity is centra l to understanding the neural code and building practical decoding algorithms. N o parametric models currently exist for modeling multivariate correlated neural data and the high dimensional nature of the data makes fully non-parametric meth ods impractical. To address these problems we propose an energy-based model in w hich the joint probability of neural activity is represented using learned funct ions of the 1D marginal histograms of the data. The parameters of the model are learned using contrastive divergence and an optimization procedure for finding a ppropriate marginal directions. We evaluate the method using real data recorded from a population of motor cortical neurons. In particular, we model the joint p robability of population spiking times and 2D hand position and show that the li kelihood of test data under our model is significantly higher than under other m odels. These results suggest that our model captures correlations in the firing activity. Our rich probabilistic model of neural population activity is a step t owards both measurement of the importance of correlations in neural coding and i mproved decoding of population activity.
****
Searching for Character Models
Jaety Edwards, David Forsyth
We introduce a method to automatically improve character models for a handwritte n script without the use of transcriptions and using a minimum of document speci ■c training data. We show that we can use searches for the words in a dictionary to identify portions of the document whose transcriptions are unambiguous. Usin g templates extracted from those regions, we retrain our character prediction mo del to drastically improve our search retrieval performance for words in the doc ument.
****
An aVLSI Cricket Ear Model
Andre Schaik, Richard Reeve, Craig Jin, Tara Hamilton
Female crickets can locate males by phonotaxis to the mating song they produce. The behaviour and underlying physiology has been studied in some depth showing t hat the cricket auditory system solves this complex problem in a unique manner. We present an analogue very large scale integrated (aVLSI) circuit model of this process and show that results from testing the circuit agree with simulation an d what is known from the behaviour and physiology of the cricket auditory system . The aVLSI circuitry is now being extended to use on a robot along with previou sly modelled neural circuitry to better understand the complete sensorimotor pat hway.
****
An Analog Visual Pre-Processing Processor Employing Cyclic Line Access in Only-N earest-Neighbor-Interconnects Architecture
Yusuke Nakashita, Yoshio Mita, Tadashi Shibata
An analog focal-plane processor having a 128128 photodiode array has been develo ped for directional edge ■ltering. It can perform 44-pixel kernel convolution fo r entire pixels only with 256 steps of simple ana- log processing. Newly develop ed cyclic line access and row-parallel processing scheme in conjunction with the “only-nearest-neighbor in- terconnects” architecture has enabled a very simple implementation. A proof-of-concept chip was fabricated in a 0.35-(cid:0)m 2-poly

3-metal CMOS technology and the edge ■ltering at a rate of 200 frames/sec. has been experimentally demonstrated.
****
The Curse of Highly Variable Functions for Local Kernel Machines
Yoshua Bengio, Olivier Delalleau, Nicolas Roux
We present a series of theoretical arguments supporting the claim that a large class of modern learning algorithms that rely solely on the smoothness prior with similarity between examples expressed with a local kernel are sensitive to the curse of dimensionality, or more precisely to the variability of the target. Our discussion covers supervised, semisupervised and unsupervised learning algorithms. These algorithms are found to be local in the sense that crucial properties of the learned function at x depend mostly on the neighbors of x in the training set. This makes them sensitive to the curse of dimensionality, well studied for classical non-parametric statistical learning. We show in the case of the Gaussian kernel that when the function to be learned has many variations, these algorithms require a number of training examples proportional to the number of variations, which could be large even though there may exist short descriptions of the target function, i.e. their Kolmogorov complexity may be low. This suggests that there exist non-local learning algorithms that at least have the potential to learn about such structured but apparently complex functions (because locally they have many variations), while not using very specific prior domain knowledge.
****
Fast Gaussian Process Regression using KD-Trees
Yirong Shen, Matthias Seeger, Andrew Ng
1 Introduction
****
Using ``epitomes'' to model genetic diversity: Rational design of HIV vaccine cocktails
Nebojsa Jojic, Vladimir Jojic, Christopher Meek, David Heckerman, Brendan J. Frey
We introduce a new model of genetic diversity which summarizes a large input dataset into an epitome, a short sequence or a small set of short sequences of probability distributions capturing many overlapping subsequences from the dataset. The epitome as a representation has already been used in modeling real-valued signals, such as images and audio. The discrete sequence model we introduce in this paper targets applications in genetics, from multiple alignment to recombination and mutation inference. In our experiments, we concentrate on modeling the diversity of HIV where the epitome emerges as a natural model for producing relatively small vaccines covering a large number of immune system targets known as epitopes. Our experiments show that the epitome includes more epitopes than other vaccine designs of similar length, including cocktails of consensus strains, phylogenetic tree centers, and observed strains. We also discuss epitome designs that take into account uncertainty about Tcell cross reactivity and epitope presentation. In our experiments, we find that vaccine optimization is fairly robust to these uncertainties.
****
Learning Cue-Invariant Visual Responses
Jarmo Hurri
Multiple visual cues are used by the visual system to analyze a scene; achromatic cues include luminance, texture, contrast and motion. Singlecell recordings have shown that the mammalian visual cortex contains neurons that respond similarly to scene structure (e.g., orientation of a boundary), regardless of the cue type conveying this information. This paper shows that cue-invariant response properties of simple- and complex-type cells can be learned from natural image data in an unsupervised manner. In order to do this, we also extend a previous conceptual model of cue invariance so that it can be applied to model simple- and complex-cell responses. Our results relate cue-invariant response properties to natural image statistics, thereby showing how the statistical modeling approach can be used to model processing beyond the elemental response properties visual neur

ons. This work also demonstrates how to learn, from natural image data, more sophisticated feature detectors than those based on changes in mean luminance, thereby paving the way for new data-driven approaches to image processing and computer vision.
****
Learning Topology with the Generative Gaussian Graph and the EM Algorithm
Michaël Aupetit
Given a set of points and a set of prototypes representing them, how to create a graph of the prototypes whose topology accounts for that of the points? This problem had not yet been explored in the framework of statistical learning theory. In this work, we propose a generative model based on the Delaunay graph of the prototypes and the ExpectationMaximization algorithm to learn the parameters. This work is a first step towards the construction of a topological model of a set of points grounded on statistics.
****
Optimizing spatio-temporal filters for improving Brain-Computer Interfacing
Guido Dornhege, Benjamin Blankertz, Matthias Krauledat, Florian Losch, Gabriel Curio, Klaus-Robert Müller
Brain-Computer Interface (BCI) systems create a novel communication channel from the brain to an output device by bypassing conventional motor output pathways of nerves and muscles. Therefore they could provide a new communication and control option for paralyzed patients. Modern BCI technology is essentially based on techniques for the clas- si■cation of single-trial brain signals. Here we present a novel technique that allows the simultaneous optimization of a spatial and a spectral ■lter enhancing discriminability of multi-channel EEG single-trials. The eval- uation of 60 experiments involving 22 different subjects demonstrates the superiority of the proposed algorithm. Apart from the enhanced clas- si■cation, the spatial and/or the spectral ■lter that are determined by the algorithm can also be used for further analysis of the data, e.g., for source localization of the respective brain rhythms.
****
Unbiased Estimator of Shape Parameter for Spiking Irregularities under Changing Environments
Keiji Miura, Masato Okada, Shun-ichi Amari
We considered a gamma distribution of interspike intervals as a statisti- cal model for neuronal spike generation. The model parameters consist of a time-dependent ■ring rate and a shape parameter that characterizes spiking irregularities of individual neurons. Because the environment changes with time, observed data are generated from the time-dependent ■ring rate, which is an unknown function. A statistical model with an unknown function is called a semiparametric model, which is one of the unsolved problem in statistics and is generally very dif■cult to solve. We used a novel method of estimating functions in information geometry to estimate the shape parameter without estimating the unknown function. We analytically obtained an optimal estimating function for the shape parameter independent of the functional form of the ■ring rate. This estimation is ef■cient without Fisher information loss and better than maximum likelihood estimation.
****
Coarse sample complexity bounds for active learning
Sanjoy Dasgupta
We characterize the sample complexity of active learning problems in terms of a parameter which takes into account the distribution over the input space, the specific target hypothesis, and the desired accuracy.
****
A PAC-Bayes approach to the Set Covering Machine
François Laviolette, Mario Marchand, Mohak Shah
We design a new learning algorithm for the Set Covering Ma- chine from a PAC-Bayes perspective and propose a PAC-Bayes risk bound which is minimized for classi■ers achieving a non trivial margin-sparsity trade-o■.
****
Logic and MRF Circuitry for Labeling Occluding and Thinline Visual Contours

Eric Saund

This paper presents representation and logic for labeling contrast edges and ridges in visual scenes in terms of both surface occlusion (border ownership) and thinline objects. In natural scenes, thinline objects in- clude sticks and wires, while in human graphical communication thin- lines include connectors, dividers, and other abstract devices. Our analy- sis is directed at both natural and graphical domains. The basic problem is to formulate the logic of the interactions among local image events, speci■cally contrast edges, ridges, junctions, and alignment relations, such as to encode the natural constraints among these events in visual scenes. In a sparse heterogeneous Markov Random Field framework, we de■ne a set of interpretation nodes and energy/potential functions among them. The minimum energy con■guration found by Loopy Belief Prop- agation is shown to correspond to preferred human interpretation across a wide range of prototypical examples including important illusory con- tour ■gures such as the Kanizsa Triangle, as well as more dif■cult ex- amples. In practical terms, the approach delivers correct interpretations of inherently ambiguous hand-drawn box-and-connector diagrams at low computational cost.

****

Non-iterative Estimation with Perturbed Gaussian Markov Processes

Yunsong Huang, B. Keith Jenkins

We develop an approach for estimation with Gaussian Markov processes that imposes a smoothness prior while allowing for discontinuities. In- stead of propagating information laterally between neighboring nodes in a graph, we study the posterior distribution of the hidden nodes as a whole—how it is perturbed by invoking discontinuities, or weakening the edges, in the graph. We show that the resulting computation amounts to feed-forward fan-in operations reminiscent of V1 neurons. Moreover, using suitable matrix preconditioners, the incurred matrix inverse and determinant can be approximated, without iteration, in the same compu- tational style. Simulation results illustrate the merits of this approach.

****

Products of ``Edge-perts

Max Welling, Peter Gehler

Images represent an important and abundant source of data. Understanding their statistical structure has important applications such as image compression and restoration. In this paper we propose a particular kind of probabilistic model, dubbed the "products of edge-perts model" to describe the structure of wavelet transformed images. We develop a practical denoising algorithm based on a single edge-pert and show state-ofthe-art denoising performance on benchmark images.

****

Ideal Observers for Detecting Motion: Correspondence Noise

Hongjing Lu, Alan L. Yuille

We derive a Bayesian Ideal Observer (BIO) for detecting motion and solving the correspondence problem. We obtain Barlow and Tripathy’s classic model as an approximation. Our psychophysical experiments show that the trends of human performance are similar to the Bayesian Ideal, but overall human performance is far worse. We investigate ways to degrade the Bayesian Ideal but show that even extreme degradations do not approach human performance. Instead we propose that humans perform motion tasks using generic, general purpose, models of motion. We perform more psychophysical experiments which are consistent with humans using a Slow-and-Smooth model and which rule out an alterna- tive model using Slowness.

****

Selecting Landmark Points for Sparse Manifold Learning

Jorge Silva, Jorge Marques, João Lemos

There has been a surge of interest in learning non-linear manifold models to approximate high-dimensional data. Both for computational complexity reasons and for generalization capability, sparsity is a desired feature in such models. This usually means dimensionality reduction, which naturally implies estimating the intrinsic dimension, but it can also mean selecting a subset of the data to use as landmarks, which is especially important because many existing algorithms have quadratic complexity in the number of observations. This paper presents an algo

rithm for selecting landmarks, based on LASSO regression, which is well known to favor sparse approximations because it uses regularization with an l1 norm. As an added benefit, a continuous manifold parameterization, based on the landmarks, is also found. Experimental results with synthetic and real data illustrate the algorithm.
****

Statistical Convergence of Kernel CCA

Kenji Fukumizu, Arthur Gretton, Francis Bach

While kernel canonical correlation analysis (kernel CCA) has been applied in many problems, the asymptotic convergence of the functions estimated from a finite sample to the true functions has not yet been established. This paper gives a rigorous proof of the statistical convergence of kernel CCA and a related method (NOCCO), which provides a theoretical justification for these methods. The result also gives a sufficient condition on the decay of the regularization coefficient in the methods to ensure convergence.
****

Efficient Estimation of OOMs

Herbert Jaeger, Mingjie Zhao, Andreas Kolling

A standard method to obtain stochastic models for symbolic time series is to train state-emitting hidden Markov models (SE-HMMs) with the Baum-Welch algorithm. Based on observable operator models (OOMs), in the last few months a number of novel learning algorithms for similar purposes have been developed: (1,2) two versions of an "efficiency sharpening" (ES) algorithm, which iteratively improves the statistical efficiency of a sequence of OOM estimators, (3) a constrained gradient descent ML estimator for transition-emitting HMMs (TE-HMMs). We give an overview on these algorithms and compare them with SE-HMM/EM learning on synthetic and real-life data.
****

AER Building Blocks for Multi-Layer Multi-Chip Neuromorphic Vision Systems

R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, H. Kolle Riis, T. Delbruck, S. C. Liu, S. Zahnd, A. M. Whatley, R. Douglas, P. Hafliger, G. Jimenez-Moreno, A. Civit, T. Serrano-Gotarredona, A. Acosta-Jimenez, B. Linares-Barranco

A 5-layer neuromorphic vision processor whose components communicate spike events asychronously using the address-event- representation (AER) is demonstrated. The system includes a retina chip, two convolution chips, a 2D winner-take-all chip, a delay line chip, a learning classi■er chip, and a set of PCBs for computer interfacing and address space remappings. The components use a mixture of analog and digital computation and will learn to classify trajectories of a moving object. A complete experimental setup and measurements results are shown.
****

A Hierarchical Compositional System for Rapid Object Detection

Long Zhu, Alan L. Yuille

We describe a hierarchical compositional system for detecting deformable objects in images. Objects are represented by graphical models. The algorithm uses a hierarchical tree where the root of the tree corresponds to the full object and lower-level elements of the tree correspond to simpler features. The algorithm proceeds by passing simple messages up and down the tree. The method works rapidly, in under a second, on 320 240 images. We demonstrate the approach on detecting cats, horses, and hands. The method works in the presence of background clutter and occlusions. Our approach is contrasted with more traditional methods such as dynamic programming and belief propagation.
****

Conditional Visual Tracking in Kernel Space

Cristian Sminchisescu, Atul Kanujia, Zhiguo Li, Dimitris Metaxas

We present a conditional temporal probabilistic framework for recon- structing 3D human motion in monocular video based on descriptors en- coding image silhouette observations. For computational ef■ciency we restrict visual inference to low-dimensional kernel induced non-linear state spaces. Our methodology (kBME) combines kernel PCA-based non-linear dimensionality reduction (kPCA) and Conditional

Bayesian Mixture of Experts (BME) in order to learn complex multivalued pre- di ctors between observations and model hidden states. This is necessary for accura te, inverse, visual perception inferences, where several proba- ble, distant 3D solutions exist due to noise or the uncertainty of monoc- ular perspective proje ction. Low-dimensional models are appropriate because many visual processes exhi bit strong non-linear correlations in both the image observations and the target , hidden state variables. The learned predictors are temporally combined within a conditional graphi- cal model in order to allow a principled propagation of un certainty. We study several predictors and empirically show that the proposed al go- rithm positively compares with techniques based on regression, Kernel Depend ency Estimation (KDE) or PCA alone, and gives results competi- tive to those of high-dimensional mixture predictors at a fraction of their computational cost. W e show that the method successfully reconstructs the complex 3D motion of humans in real monocular video sequences.
****

Fast Krylov Methods for N-Body Learning
Nando Freitas, Yang Wang, Maryam Mahdaviani, Dustin Lang
This paper addresses the issue of numerical computation in machine learning doma ins based on similarity metrics, such as kernel methods, spectral techniques and Gaussian processes. It presents a general solution strategy based on Krylov sub space iteration and fast N-body learning methods. The experiments show significa nt gains in computation and storage on datasets arising in image segmentation, o bject detection and dimensionality reduction. The paper also presents theoretica l bounds on the stability of these methods.
****

A Connectionist Model for Constructive Modal Reasoning
Artur Garcez, Luis Lamb, Dov M. Gabbay
We present a new connectionist model for constructive, intuitionistic modal reas oning. We use ensembles of neural networks to represent in- tuitionistic modal t heories, and show that for each intuitionistic modal program there exists a corr esponding neural network ensemble that com- putes the program. This provides a m assively parallel model for intu- itionistic modal reasoning, and sets the scene for integrated reasoning, knowledge representation, and learning of intuitionis tic theories in neural networks, since the networks in the ensemble can be train ed by examples using standard neural learning algorithms.
****

Spiking Inputs to a Winner-take-all Network
Matthias Oster, Shih-Chii Liu
Recurrent networks that perform a winner-take-all computation have been studied extensively. Although some of these studies include spik- ing networks, they con sider only analog input rates. We present results of this winner-take-all comput ation on a network of integrate-and-■re neurons which receives spike trains as i nputs. We show how we can con- ■gure the connectivity in the network so that the winner is selected after a pre-determined number of input spikes. We discuss sp iking inputs with both regular frequencies and Poisson-distributed rates. The ro bustness of the computation was tested by implementing the winner-take-all netwo rk on an analog VLSI array of 64 integrate-and-■re neurons which have an innate variance in their operating parameters.
****

Estimation of Intrinsic Dimensionality Using High-Rate Vector Quantization
Maxim Raginsky, Svetlana Lazebnik
We introduce a technique for dimensionality estimation based on the notion of qu antization dimension, which connects the asymptotic optimal quantization error f or a probability distribution on a manifold to its intrinsic dimension. The defi nition of quantization dimension yields a family of estimation algorithms, whose limiting case is equivalent to a recent method based on packing numbers. Using the formalism of high-rate vector quantization, we address issues of statistical consistency and analyze the behavior of our scheme in the presence of noise.
****

Generalization error bounds for classifiers trained with interdependent data

Nicolas Usunier, Massih R. Amini, Patrick Gallinari
In this paper we propose a general framework to study the generalization properties of binary classi■ers trained with data which may be depen- dent, but are deterministically generated upon a sample of independent examples. It provides generalization bounds for binary classi■cation and some cases of ranking problems, and clari■es the relationship between these learning tasks.
****
Bayesian model learning in human visual perception
Gerg■ Orbán, Jozsef Fiser, Richard N. Aslin, Máté Lengyel
Humans make optimal perceptual decisions in noisy and ambiguous conditions. Computations underlying such optimal behavior have been shown to rely on probabilistic inference according to generative models whose structure is usually taken to be known a priori. We argue that Bayesian model selection is ideal for inferring similar and even more complex model structures from experience. We ■nd in experiments that humans learn subtle statistical properties of visual scenes in a completely unsupervised manner. We show that these ■ndings are well captured by Bayesian model learning within a class of models that seek to explain observed variables by independent hidden causes.
****
Active Learning For Identifying Function Threshold Boundaries
Brent Bryan, Robert C. Nichol, Christopher R. Genovese, Jeff Schneider, Christopher J. Miller, Larry Wasserman
We present an ef■cient algorithm to actively select queries for learning the boundaries separating a function domain into regions where the func- tion is above and below a given threshold. We develop experiment selec- tion methods based on entropy, misclassi■cation rate, variance, and their combinations, and show how they perform on a number of data sets. We then show how these algorithms are used to determine simultaneously valid $1 - \alpha$ con■dence intervals for seven cosmological parameters. Ex- perimentation shows that the algorithm reduces the computation neces- sary for the parameter estimation problem by an order of magnitude.
****
Cue Integration for Figure/Ground Labeling
Xiaofeng Ren, Jitendra Malik, Charless Fowlkes
We present a model of edge and region grouping using a conditional random field built over a scale-invariant representation of images to integrate multiple cues. Our model includes potentials that capture low-level similarity, mid-level curvilinear continuity and high-level object shape. Maximum likelihood parameters for the model are learned from human labeled groundtruth on a large collection of horse images using belief propagation. Using held out test data, we quantify the information gained by incorporating generic mid-level cues and high-level shape.
****
Inferring Motor Programs from Images of Handwritten Digits
Vinod Nair, Geoffrey E. Hinton
We describe a generative model for handwritten digits that uses two pairs of opposing springs whose stiffnesses are controlled by a motor program. We show how neural networks can be trained to infer the motor programs required to accurately reconstruct the MNIST digits. The inferred motor programs can be used directly for digit classi■cation, but they can also be used in other ways. By adding noise to the motor program inferred from an MNIST image we can generate a large set of very different images of the same class, thus enlarging the training set available to other methods. We can also use the motor programs as additional, highly informative outputs which reduce over■tting when training a feed-forward classi■er.
****
Dual-Tree Fast Gauss Transforms
Dongryeol Lee, Andrew Moore, Alexander Gray
In previous work we presented an ef■cient approach to computing ker- nel summations which arise in many machine learning methods such as kernel density estimation. This approach, dual-tree recursion with ■nite- difference approximation, gen

eralized existing methods for similar prob- lems arising in computational physics in two ways appropriate for sta- tistical problems: toward distribution sensitivity and general dimension, partly by avoiding series expansions. While this proved to be the fastest practical method for multivariate kernel density estimation at the optimal bandwidth, it is much less ef■cient at larger-than-optimal bandwidths. In this work, we explore the extent to which the dual-tree approach can be integrated with multipole-like Hermite expansions in order to achieve reasonable ef■ciency across all bandwidth scales, though only for low di- mensionalities. In the process, we derive and demonstrate the ■rst truly hierarchical fast Gauss transforms, effectively combining the best tools from discrete algorithms and continuous approximation theory.
****
Describing Visual Scenes using Transformed Dirichlet Processes
Antonio Torralba, Alan Willsky, Erik Sudderth, William Freeman
Motivated by the problem of learning to detect and recognize objects with minimal supervision, we develop a hierarchical probabilistic model for the spatial structure of visual scenes. In contrast with most existing models, our approach explicitly captures uncertainty in the number of object instances depicted in a given image. Our scene model is based on the transformed Dirichlet process (TDP), a novel extension of the hierarchical DP in which a set of stochastically transformed mixture components are shared between multiple groups of data. For visual scenes, mixture components describe the spatial structure of visual features in an objectcentered coordinate frame, while transformations model the object positions in a particular image. Learning and inference in the TDP, which has many potential applications beyond computer vision, is based on an empirically effective Gibbs sampler. Applied to a dataset of partially labeled street scenes, we show that the TDP's inclusion of spatial structure improves detection performance, flexibly exploiting partially labeled training images.
****
Worst-Case Bounds for Gaussian Process Models
Sham M. Kakade, Matthias W. Seeger, Dean P. Foster
We present a competitive analysis of some non-parametric Bayesian al- gorithms in a worst-case online learning setting, where no probabilistic assumptions about the generation of the data are made. We consider models which use a Gaussian process prior (over the space of all func- tions) and provide bounds on the regret (under the log loss) for com- monly used non-parametric Bayesian algorithms — including Gaussian regression and logistic regression — which show how these algorithms can perform favorably under rather general conditions. These bounds ex- plicitly handle the in■nite dimensionality of these non-parametric classes in a natural way. We also make formal connections to the minimax and minimum description length (MDL) framework. Here, we show precisely how Bayesian Gaussian regression is a minimax strategy.
****
Analyzing Auditory Neurons by Learning Distance Functions
Inna Weiner, Tomer Hertz, Israel Nelken, Daphna Weinshall
We present a novel approach to the characterization of complex sensory neurons. One of the main goals of characterizing sensory neurons is to characterize dimensions in stimulus space to which the neurons are highly sensitive (causing large gradients in the neural responses) or al- ternatively dimensions in stimulus space to which the neuronal response are invariant (de■ning iso-response manifolds). We formulate this prob- lem as that of learning a geometry on stimulus space that is compatible with the neural responses: the distance between stimuli should be large when the responses they evoke are very different, and small when the re- sponses they evoke are similar. Here we show how to successfully train such distance functions using rather limited amount of information. The data consisted of the responses of neurons in primary auditory cortex (A1) of anesthetized cats to 32 stimuli derived from natural sounds. For each neuron, a subset of all pairs of stimuli was selected such that the responses of the two stimuli in a pair were either very similar or very dissimilar. The distance function was trained to ■t these constraints. The resulting distance functions generalized to predic

t the distances between the responses of a test stimulus and the trained stimuli .
****
Cyclic Equilibria in Markov Games
Martin Zinkevich, Amy Greenwald, Michael Littman
Although variants of value iteration have been proposed for ■nding Nash or correlated equilibria in general-sum Markov games, these variants have not been shown to be effective in general. In this paper, we demon- strate by construction that existing variants of value iteration cannot ■nd stationary equilibrium policies in arbitrary general-sum Markov games. Instead, we propose an alternative interpretation of the output of value it- eration based on a new (non-stationary) equilibrium concept that we call “cyclic equilibria.” We prove that value iteration identi■es cyclic equi- libria in a class of games in which it fails to ■nd stationary equilibria. We also demonstrate empirically that value iteration ■nds cyclic equilibria in nearly all examples drawn from a random distribution of Markov games.
****
The Information-Form Data Association Filter
Brad Schumitsch, Sebastian Thrun, Gary Bradski, Kunle Olukotun
This paper presents a new ■lter for online data association problems in high-dimensional spaces. The key innovation is a representation of the data association posterior in information form, in which the “proxim- ity” of objects and tracks are expressed by numerical links. Updating these links requires linear time, compared to exponential time required for computing the exact posterior probabilities. The paper derives the algorithm formally and provides comparative results using data obtained by a real-world camera array and by a large-scale sensor network simu- lation.
****
Consistency of one-class SVM and related algorithms
Régis Vert, Jean-philippe Vert
We determine the asymptotic limit of the function computed by support vector machines (SVM) and related algorithms that minimize a regularized empirical convex loss function in the reproducing kernel Hilbert space of the Gaussian RBF kernel, in the situation where the number of examples tends to infinity, the bandwidth of the Gaussian kernel tends to 0, and the regularization parameter is held fixed. Non-asymptotic convergence bounds to this limit in the L2 sense are provided, together with upper bounds on the classification error that is shown to converge to the Bayes risk, therefore proving the Bayes-consistency of a variety of methods although the regularization term does not vanish. These results are particularly relevant to the one-class SVM, for which the regularization can not vanish by construction, and which is shown for the first time to be a consistent density level set estimator.
****
Nested sampling for Potts models
Iain Murray, David MacKay, Zoubin Ghahramani, John Skilling
Nested sampling is a new Monte Carlo method by Skilling [1] intended for general Bayesian computation. Nested sampling provides a robust alternative to annealing-based methods for computing normalizing constants. It can also generate estimates of other quantities such as posterior expectations. The key technical requirement is an ability to draw samples uniformly from the prior sub ject to a constraint on the likelihood. We provide a demonstration with the Potts model, an undirected graphical model.
****
Correlated Topic Models
John Lafferty, David Blei
Topic models, such as latent Dirichlet allocation (LDA), can be useful tools for the statistical analysis of document collections and other discrete data. The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary. A limitation of LDA is the inability to model topic correlation even though, for example, a document about

genetics is more likely to also be about disease than x-ray astronomy. This lim
itation stems from the use of the Dirichlet distribution to model the variabilit
y among the topic proportions. In this paper we develop the correlated topic mod
el (CTM), where the topic proportions exhibit correlation via the logistic norma
l distribution [1]. We derive a mean-field variational inference algorithm for a
pproximate posterior inference in this model, which is complicated by the fact t
hat the logistic normal is not conjugate to the multinomial. The CTM gives a bet
ter fit than LDA on a collection of OCRed articles from the journal Science. Fur
thermore, the CTM provides a natural way of visualizing and exploring this and o
ther unstructured data sets.
****
Learning in Silicon: Timing is Everything
John V. Arthur, Kwabena Boahen
We describe a neuromorphic chip that uses binary synapses with spike timing-depe
ndent plasticity (STDP) to learn stimulated patterns of activ- ity and to compen
sate for variability in excitability. Speci■cally, STDP preferentially potentiat
es (turns on) synapses that project from excitable neurons, which spike early, t
o lethargic neurons, which spike late. The additional excitatory synaptic curren
t makes lethargic neurons spike ear- lier, thereby causing neurons that belong t
o the same pattern to spike in synchrony. Once learned, an entire pattern can be
 recalled by stimulating a subset.
****
Correcting sample selection bias in maximum entropy density estimation
Miroslav Dudík, Steven Phillips, Robert E. Schapire
We study the problem of maximum entropy density estimation in the presence of kn
own sample selection bias. We propose three bias cor- rection approaches. The ■r
st one takes advantage of unbiased suf■cient statistics which can be obtained fr
om biased samples. The second one es- timates the biased distribution and then f
actors the bias out. The third one approximates the second by only using samples
 from the sampling distri- bution. We provide guarantees for the ■rst two approa
ches and evaluate the performance of all three approaches in synthetic experimen
ts and on real data from species habitat modeling, where maxent has been success
- fully applied and where sample selection bias is a signi■cant problem.
****
Rate Distortion Codes in Sensor Networks: A System-level Analysis
Tatsuto Murayama, Peter Davis
This paper provides a system-level analysis of a scalable distributed sens- ing
model for networked sensors. In our system model, a data center ac- quires data
from a bunch of L sensors which each independently encode their noisy observatio
ns of an original binary sequence, and transmit their encoded data sequences to
the data center at a combined rate R, which is limited. Supposing that the senso
rs use independent LDGM rate dis- tortion codes, we show that the system perform
ance can be evaluated for any given ■nite R when the number of sensors L goes to
 in■nity. The analysis shows how the optimal strategy for the distributed sensin
g prob- lem changes at critical values of the data rate R or the noise level.
****
Beyond Pair-Based STDP: a Phenomenological Rule for Spike Triplet and Frequency
Effects
Jean-pascal Pfister, Wulfram Gerstner
While classical experiments on spike-timing dependent plasticity analyzed synapt
ic changes as a function of the timing of pairs of pre- and postsynaptic spikes,
 more recent experiments also point to the effect of spike triplets. Here we dev
elop a mathematical framework that allows us to characterize timing based learni
ng rules. Moreover, we identify a candidate learning rule with five variables (a
nd 5 free parameters) that captures a variety of experimental data, including th
e dependence of potentiation and depression upon pre- and postsynaptic firing fr
equencies. The relation to the Bienenstock-Cooper-Munro rule as well as to some
timing-based rules is discussed.
****
Is Early Vision Optimized for Extracting Higher-order Dependencies?

Yan Karklin, Michael Lewicki
Linear implementations of the efficient coding hypothesis, such as independent c
omponent analysis (ICA) and sparse coding models, have provided functional expla
nations for properties of simple cells in V1 [1, 2]. These models, however, igno
re the non-linear behavior of neurons and fail to match individual and populatio
n properties of neural receptive fields in subtle but important ways. Hierarchic
al models, including Gaussian Scale Mixtures [3, 4] and other generative statist
ical models [5, 6], can capture higher-order regularities in natural images and
explain nonlinear aspects of neural processing such as normalization and context
 effects [6, 7]. Previously, it had been assumed that the lower level representa
tion is independent of the hierarchy, and had been fixed when training these mod
els. Here we examine the optimal lower-level representations derived in the cont
ext of a hierarchical model and find that the resulting representations are stri
kingly different from those based on linear models. Unlike the the basis functio
ns and filters learned by ICA or sparse coding, these functions individually mor
e closely resemble simple cell receptive fields and collectively span a broad ra
nge of spatial scales. Our work unifies several related approaches and observati
ons about natural image structure and suggests that hierarchical models might yi
eld better representations of image structure throughout the hierarchy.
****
An Approximate Inference Approach for the PCA Reconstruction Error
Manfred Opper
The problem of computing a resample estimate for the reconstruction error in PCA
 is reformulated as an inference problem with the help of the replica method. Us
ing the expectation consistent (EC) approximation, the intractable inference pro
blem can be solved efficiently using only two variational parameters. A perturba
tive correction to the result is computed and an alternative simplified derivati
on is also presented.
****
Active Bidirectional Coupling in a Cochlear Chip
Bo Wen, Kwabena A. Boahen
We present a novel cochlear model implemented in analog very large scale integra
tion (VLSI) technology that emulates nonlinear active cochlear behavior. This si
licon cochlea includes outer hair cell (OHC) electromotility through active bidi
rectional coupling (ABC), a mech- anism we proposed in which OHC motile forces,
through the mi- croanatomical organization of the organ of Corti, realize the co
chlear ampli■er. Our chip measurements demonstrate that frequency responses beco
me larger and more sharply tuned when ABC is turned on; the de- gree of the enha
ncement decreases with input intensity as ABC includes saturation of OHC forces.
****
Affine Structure From Sound
Sebastian Thrun
We consider the problem of localizing a set of microphones together with a set o
f external acoustic events (e.g., hand claps), emitted at un- known times and un
known locations. We propose a solution that ap- proximates this problem under a
far ■eld approximation de■ned in the calculus of af■ne geometry, and that relies
 on singular value decompo- sition (SVD) to recover the af■ne structure of the p
roblem. We then de■ne low-dimensional optimization techniques for embedding the
solu- tion into Euclidean geometry, and further techniques for recovering the lo
cations and emission times of the acoustic events. The approach is use- ful for
the calibration of ad-hoc microphone arrays and sensor networks.
****
Distance Metric Learning for Large Margin Nearest Neighbor Classification
Kilian Q. Weinberger, John Blitzer, Lawrence Saul
We show how to learn a Mahanalobis distance metric for k -nearest neighbor (kNN)
 classification by semidefinite programming. The metric is trained with the goal
 that the k -nearest neighbors always belong to the same class while examples fr
om different classes are separated by a large margin. On seven data sets of vary
ing size and difficulty, we find that metrics trained in this way lead to signif
icant improvements in kNN classification--for example, achieving a test error ra

te of 1.3% on the MNIST handwritten digits. As in support vector machines (SVMs)
, the learning problem reduces to a convex optimization based on the hinge loss.
 Unlike learning in SVMs, however, our framework requires no modification or ext
ension for problems in multiway (as opposed to binary) classification.
****
Estimating the wrong Markov random field: Benefits in the computation-limited se
tting
Martin J. Wainwright
Consider the problem of joint parameter estimation and prediction in a Markov ra
ndom field: i.e., the model parameters are estimated on the basis of an initial
set of data, and then the fitted model is used to perform prediction (e.g., smoo
thing, denoising, interpolation) on a new noisy observation. Working in the comp
utation-limited setting, we analyze a joint method in which the same convex vari
ational relaxation is used to construct an M-estimator for fitting parameters, a
nd to perform approximate marginalization for the prediction step. The key resul
t of this paper is that in the computation-limited setting, using an inconsisten
t parameter estimator (i.e., an estimator that returns the "wrong" model even in
 the infinite data limit) is provably beneficial, since the resulting errors can
 partially compensate for errors made by using an approximate prediction techniq
ue. En route to this result, we analyze the asymptotic properties of M-estimator
s based on convex variational relaxations, and establish a Lipschitz stability p
roperty that holds for a broad class of variational methods. We show that joint
estimation/prediction based on the reweighted sum-product algorithm substantiall
y outperforms a commonly used heuristic based on ordinary sum-product. 1 Keyword
s: Markov random fields; variational method; message-passing algorithms; sum-pro
duct; belief propagation; parameter estimation; learning.
****
Gaussian Processes for Multiuser Detection in CDMA receivers
Juan Murillo-fuentes, Sebastian Caro, Fernando Pérez-Cruz
In this paper we propose a new receiver for digital communications. We focus on
the application of Gaussian Processes (GPs) to the multiuser detection (MUD) in
code division multiple access (CDMA) systems to solve the near-far problem. Henc
e, we aim to reduce the interference from other users sharing the same frequency
 band. While usual approaches minimize the mean square error (MMSE) to linearly
retrieve the user of interest, we exploit the same criteria but in the design of
 a nonlinear MUD. Since the optimal solution is known to be nonlinear, the perfo
rmance of this novel method clearly improves that of the MMSE detectors. Further
more, the GP based MUD achieves excellent interference suppression even for shor
t training sequences. We also include some experiments to illustrate that other
nonlinear detectors such as those based on Support Vector Machines (SVMs) exhibi
t a worse performance.
****
Metric Learning by Collapsing Classes
Amir Globerson, Sam Roweis
We present an algorithm for learning a quadratic Gaussian metric (Maha- lanobis
distance) for use in classi■cation tasks. Our method relies on the simple geomet
ric intuition that a good metric is one under which points in the same class are
 simultaneously near each other and far from points in the other classes. We con
struct a convex optimization problem whose solution generates such a metric by t
rying to collapse all examples in the same class to a single point and push exam
ples in other classes in■nitely far away. We show that when the metric we learn
is used in simple clas- si■ers, it yields substantial improvements over standard
 alternatives on a variety of problems. We also discuss how the learned metric m
ay be used to obtain a compact low dimensional feature representation of the ori
ginal input space, allowing more ef■cient classi■cation with very little reducti
on in performance.
****
Walk-Sum Interpretation and Analysis of Gaussian Belief Propagation
Dmitry Malioutov, Alan Willsky, Jason Johnson
This paper presents a new framework based on walks in a graph for analysis and i

nference in Gaussian graphical models. The key idea is to decompose correlations between variables as a sum over all walks between those variables in the graph. The weight of each walk is given by a product of edgewise partial correlations. We provide a walk-sum interpretation of Gaussian belief propagation in trees and of the approximate method of loopy belief propagation in graphs with cycles. This perspective leads to a better understanding of Gaussian belief propagation and of its convergence in loopy graphs.
****
Analyzing Coupled Brain Sources: Distinguishing True from Spurious Interaction
Guido Nolte, Andreas Ziehe, Frank Meinecke, Klaus-Robert Müller
When trying to understand the brain, it is of fundamental importance to analyse (e.g. from EEG/MEG measurements) what parts of the cortex interact with each other in order to infer more accurate models of brain activity. Common techniques like Blind Source Separation (BSS) can estimate brain sources and single out artifacts by using the underlying assumption of source signal independence. However, physiologically interesting brain sources typically interact, so BSS will--by construction-- fail to characterize them properly. Noting that there are truly interacting sources and signals that only seemingly interact due to effects of volume conduction, this work aims to contribute by distinguishing these effects. For this a new BSS technique is proposed that uses anti-symmetrized cross-correlation matrices and subsequent diagonalization. The resulting decomposition consists of the truly interacting brain sources and suppresses any spurious interaction stemming from volume conduction. Our new concept of interacting source analysis (ISA) is successfully demonstrated on MEG data.
****
Group and Topic Discovery from Relations and Their Attributes
Xuerui Wang, Natasha Mohanty, Andrew McCallum
We present a probabilistic generative model of entity relationships and their attributes that simultaneously discovers groups among the entities and topics among the corresponding textual attributes. Block-models of relationship data have been studied in social network analysis for some time. Here we simultaneously cluster in several modalities at once, incor- porating the attributes (here, words) associated with certain relationships. Signi■cantly, joint inference allows the discovery of topics to be guided by the emerging groups, and vice-versa. We present experimental results on two large data sets: sixteen years of bills put before the U.S. Sen- ate, comprising their corresponding text and voting records, and thirteen years of similar data from the United Nations. We show that in compari- son with traditional, separate latent-variable models for words, or Block- structures for votes, the Group-Topic model’s joint inference discovers more cohesive groups and improved topics.
****
Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care
Christopher Williams, John Quinn, Neil Mcintosh
The observed physiological dynamics of an infant receiving intensive care are affected by many possible factors, including interventions to the baby, the operation of the monitoring equipment and the state of health. The Factorial Switching Kalman Filter can be used to infer the presence of such factors from a sequence of observations, and to estimate the true values where these observations have been corrupted. We apply this model to clinical time series data and show it to be effective in identifying a number of artifactual and physiological patterns.
****
A Criterion for the Convergence of Learning with Spike Timing Dependent Plasticity
Robert Legenstein, Wolfgang Maass
We investigate under what conditions a neuron can learn by experimen- tally supported rules for spike timing dependent plasticity (STDP) to pre- dict the arrival times of strong “teacher inputs” to the same neuron. It turns out that in contrast to the famous Perceptron Convergence Theo- rem, which predicts convergence of the perceptron learning rule for a simpli■ed neuron model whenever a stable s

olution exists, no equally strong convergence guarantee can be given for spiking
 neurons with STDP. But we derive a criterion on the statistical dependency stru
cture of input spike trains which characterizes exactly when learning with STDP
will converge on average for a simple model of a spiking neuron. This criterion
is reminiscent of the linear separability criterion of the Percep- tron Converge
nce Theorem, but it applies here to the rows of a correlation matrix related to
the spike inputs. In addition we show through computer simulations for more real
istic neuron models that the resulting analyti- cally predicted positive learnin
g results not only hold for the common interpretation of STDP where STDP changes
 the weights of synapses, but also for a more realistic interpretation suggested
 by experimental data where STDP modulates the initial release probability of dy
namic synapses.
****
Q-Clustering
Mukund Narasimhan, Nebojsa Jojic, Jeff A. Bilmes
We show that Queyranne's algorithm for minimizing symmetric submodular functions
 can be used for clustering with a variety of different objective functions. Two
 specific criteria that we consider in this paper are the single linkage and the
 minimum description length criteria. The first criterion tries to maximize the
minimum distance between elements of different clusters, and is inherently "disc
riminative". It is known that optimal clusterings into k clusters for any given
k in polynomial time for this criterion can be computed. The second criterion se
eks to minimize the description length of the clusters given a probabilistic gen
erative model. We show that the optimal partitioning into 2 clusters, and approx
imate partitioning (guaranteed to be within a factor of 2 of the the optimal) fo
r more clusters can be computed. To the best of our knowledge, this is the first
 time that a tractable algorithm for finding the optimal clustering with respect
 to the MDL criterion for 2 clusters has been given. Besides the optimality resu
lt for the MDL criterion, the chief contribution of this paper is to show that t
he same algorithm can be used to optimize a broad class of criteria, and hence c
an be used for many application specific criterion for which efficient algorithm
 are not known.
****
Generalization Error Bounds for Aggregation by Mirror Descent with Averaging
Anatoli Juditsky, Alexander Nazin, Alexandre Tsybakov, Nicolas Vayatis
We consider the problem of constructing an aggregated estimator from a ■nite cla
ss of base functions which approximately minimizes a con- vex risk functional un
der the ■1 constraint. For this purpose, we propose a stochastic procedure, the
mirror descent, which performs gradient de- scent in the dual space. The generat
ed estimates are additionally aver- aged in a recursive fashion with speci■c wei
ghts. Mirror descent algo- rithms have been developed in different contexts and
they are known to be particularly ef■cient in high dimensional problems. Moreove
r their implementation is adapted to the online setting. The main result of the
paper is the upper bound on the convergence rate for the generalization error.
****
Asymptotics of Gaussian Regularized Least Squares
Ross Lippert, Ryan Rifkin
We consider regularized least-squares (RLS) with a Gaussian kernel. We prove tha
t if we let the Gaussian bandwidth $\sigma \to \infty$ while letting the regularization parame
ter $\lambda \to 0$, the RLS solution tends to a polynomial whose order is controlled by t
he rielative rates of decay of 1 σ2 and λ: if $\lambda = \sigma-(2k+1)$, then, as $\sigma \to \infty$, the
RLS solution tends to the kth order polynomial with minimal empirical error. We
illustrate the result with an example.
****
Large scale networks fingerprinting and visualization using the k-core decomposi
tion
J. Alvarez-hamelin, Luca Dall'asta, Alain Barrat, Alessandro Vespignani
We use the k-core decomposition to develop algorithms for the analysis of large
scale complex networks. This decomposition, based on a re- cursive pruning of th
e least connected vertices, allows to disentangle the hierarchical structure of

networks by progressively focusing on their cen- tral cores. By using this strat egy we develop a general visualization algo- rithm that can be used to compare t he structural properties of various net- works and highlight their hierarchical structure. The low computational complexity of the algorithm, O(n + e), where n is the size of the net- work, and e is the number of edges, makes it suitable fo r the visualization of very large sparse networks. We show how the proposed visu alization tool allows to ■nd speci■c structural ■ngerprints of networks.
****
Norepinephrine and Neural Interrupts
Peter Dayan, Angela J. Yu
Angela J. Yu
****
Consensus Propagation
Benjamin Roy, Ciamac C. Moallemi
We propose consensus propagation, an asynchronous distributed protocol for avera ging numbers across a network. We establish convergence, characterize the conver gence rate for regular graphs, and demonstrate that the protocol exhibits better scaling properties than pairwise averaging, an alternative that has received mu ch recent attention. Consensus propagation can be viewed as a special case of be lief propagation, and our results contribute to the belief propagation literatur e. In particular, beyond singly-connected graphs, there are very few classes of relevant problems for which belief propagation is known to converge.
****
CMOL CrossNets: Possible Neuromorphic Nanoelectronic Circuits
Jung Hoon Lee, Xiaolong Ma, Konstantin K. Likharev
Hybrid “CMOL” integrated circuits, combining CMOS subsystem with nanowir e crossbars and simple two-terminal nanodevices, promise to extend the exponential Moore-Law development of microelectronics into the sub-10-nm range. We are developing neuromorphic network (“CrossNet”) architecture s for this future technology, in which neural cell bodies are implemented in CMOS, nanowires are used as axons and dendrites, while nanodevices (b istable latching switches) are used as elementary synapses. We have sh own how CrossNets may be trained to perform pattern recovery and clas sification despite the limitations imposed by the CMOL hardware. Prel iminary estimates have shown that CMOL CrossNets may be extremely dense (~ 107 cells per cm2) and operate approximately a million times faster than biolog ical neural networks, at manageable power consumption. In Conclusion, we discuss in brief possible short-term and long-term applications of the emergi ng technology.
****
A General and Efficient Multiple Kernel Learning Algorithm
Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer
While classical kernel-based learning algorithms are based on a single kernel, i n practice it is often desirable to use multiple kernels. Lankriet et al. (2004) considered conic combinations of kernel matrices for classi- ■cation, leading t o a convex quadratically constraint quadratic program. We show that it can be re written as a semi-in■nite linear program that can be ef■ciently solved by recycl ing the standard SVM implementa- tions. Moreover, we generalize the formulation and our method to a larger class of problems, including regression and one-class classi■ca- tion. Experimental results show that the proposed algorithm helps fo r automatic model selection, improving the interpretability of the learn- ing re sult and works for hundred thousands of examples or hundreds of kernels to be co mbined.
****
Stimulus Evoked Independent Factor Analysis of MEG Data with Large Background Ac tivity
Kenneth Hild, Kensuke Sekihara, Hagai Attias, Srikantan Nagarajan
This paper presents a novel technique for analyzing electromagnetic imaging data obtained using the stimulus evoked experimental paradigm. The technique is base d on a probabilistic graphical model, which describes the data in terms of under

lying evoked and interference sources, and explicitly models the stimulus evoked paradigm. A variational Bayesian EM algorithm infers the model from data, suppresses interference sources, and reconstructs the activity of separated individual brain sources. The new algorithm outperforms existing techniques on two real datasets, as well as on simulated data.
****
Augmented Rescorla-Wagner and Maximum Likelihood Estimation
Alan L. Yuille
We show that linear generalizations of Rescorla-Wagner can perform Maximum Likelihood estimation of the parameters of all generative models for causal reasoning. Our approach involves augmenting variables to deal with conjunctions of causes, similar to the agumented model of Rescorla. Our results involve genericity assumptions on the distributions of causes. If these assumptions are violated, for example for the Cheng causal power theory, then we show that a linear Rescorla-Wagner can estimate the parameters of the model up to a nonlinear transformtion. Moreover, a nonlinear Rescorla-Wagner is able to estimate the parameters directly to within arbitrary accuracy. Previous results can be used to determine convergence and to estimate convergence rates.
****
Laplacian Score for Feature Selection
Xiaofei He, Deng Cai, Partha Niyogi
In supervised learning scenarios, feature selection has been studied widely in the literature. Selecting features in unsupervised learning scenarios is a much harder problem, due to the absence of class labels that would guide the search for relevant information. And, almost all of previous unsupervised feature selection methods are "wrapper" techniques that require a learning algorithm to evaluate the candidate feature subsets. In this paper, we propose a "filter" method for feature selection which is independent of any learning algorithm. Our method can be performed in either supervised or unsupervised fashion. The proposed method is based on the observation that, in many real world classification problems, data from the same class are often close to each other. The importance of a feature is evaluated by its power of locality preserving, or, Laplacian Score. We compare our method with data variance (unsupervised) and Fisher score (supervised) on two data sets. Experimental results demonstrate the effectiveness and efficiency of our algorithm.
****
Variational Bayesian Stochastic Complexity of Mixture Models
Kazuho Watanabe, Sumio Watanabe
The Variational Bayesian framework has been widely used to approximate the Bayesian learning. In various applications, it has provided computational tractability and good generalization performance. In this paper, we discuss the Variational Bayesian learning of the mixture of exponential families and provide some additional theoretical support by deriving the asymptotic form of the stochastic complexity. The stochastic complexity, which corresponds to the minimum free energy and a lower bound of the marginal likelihood, is a key quantity for model selection. It also enables us to discuss the effect of hyperparameters and the accuracy of the Variational Bayesian approach as an approximation of the true Bayesian learning.
****
Hierarchical Linear/Constant Time SLAM Using Particle Filters for Dense Maps
Austin I. Eliazar, Ronald Parr
We present an improvement to the DP-SLAM algorithm for simultane- ous localization and mapping (SLAM) that maintains multiple hypothe- ses about densely populated maps (one full map per particle in a par- ticle ■lter) in time that is linear in all signi■cant algorithm parameters and takes constant (amortized) time per iteration. This means that the asymptotic complexity of the algorithm is no greater than that of a pure localization algorithm using a single map and the same number of parti- cles. We also present a hierarchical extension of DP-SLAM that uses a two level particle ■lter which models drift in the particle ■ltering process itself. The hierarchical approach enables recovery from the inevitable drift

that results from using a ■nite number of particles in a particle ■lter and perm its the use of DP-SLAM in more challenging domains, while maintaining linear tim e asymptotic complexity.
****
An exploration-exploitation model based on norepinepherine and dopamine activity
Samuel M. McClure, Mark S. Gilzenrat, Jonathan D. Cohen
We propose a model by which dopamine (DA) and norepinepherine (NE) combine to a lternate behavior between relatively exploratory and exploitative modes. The model is developed for a target detection task for which there is extant single neuron recording data available from locus coeruleus (LC) NE neurons. An exploration-exploitation trade-off is elicited by regularly switching which of the two stimuli are rewarded. DA functions within the model to change synaptic weights according to a reinforcement lear ning algorithm. Exploration is mediated by the state of LC firing, wi th higher tonic and lower phasic activity producing greater response v ariability. The opposite state of LC function, with lower baseline firing rate and greater phasic responses, favors exploitative behavior. Changes in LC firing mode result from combined measures of response conflict an d reward rate, where response conflict is monitored using models of a nterior cingulate cortex (ACC). Increased long-term response conflict and decr eased reward rate, which occurs following reward contingency switch, fa vors the higher tonic state of LC function and NE release. This incr eases exploration, and facilitates discovery of the new target.
****
Kernels for gene regulatory regions
Jean-philippe Vert, Robert Thurman, William Noble
We describe a hierarchy of motif-based kernels for multiple alignments of biolog ical sequences, particularly suitable to process regulatory regions of genes. Th e kernels incorporate progressively more information, with the most complex kern el accounting for a multiple alignment of orthologous regions, the phylogenetic tree relating the species, and the prior knowledge that relevant sequence patter ns occur in conserved motif blocks. These kernels can be used in the presence of a library of known transcription factor binding sites, or de novo by iterating over all k -mers of a given length. In the latter mode, a discriminative classif ier built from such a kernel not only recognizes a given class of promoter regio ns, but as a side effect simultaneously identifies a collection of relevant, dis criminative sequence motifs. We demonstrate the utility of the motif-based multi ple alignment kernels by using a collection of aligned promoter regions from fiv e yeast species to recognize classes of cell-cycle regulated genes. Supplementar y data is available at http://noble.gs.washington.edu/proj/pkernel.
****
Transfer learning for text classification
Chuong B. Do, Andrew Y. Ng
Linear text classi■cation algorithms work by computing an inner prod- uct betwee n a test document vector and a parameter vector. In many such algorithms, includ ing naive Bayes and most TFIDF variants, the parame- ters are determined by some simple, closed-form, function of training set statistics; we call this mapping mapping from statistics to parameters, the parameter function. Much research in text classi■cation over the last few decades has consisted of manual efforts to identify better parameter func- tions. In this paper, we propose an algorithm fo r automatically learning this function from related classi■cation problems. The parameter func- tion found by our algorithm then de■nes a new learning algorithm for text classi■cation, which we can apply to novel classi■cation tasks. We ■nd that our learned classi■er outperforms existing methods on a variety of multicl ass text classi■cation tasks.
****
The Forgetron: A Kernel-Based Perceptron on a Fixed Budget
Ofer Dekel, Shai Shalev-shwartz, Yoram Singer
The Perceptron algorithm, despite its simplicity, often performs well on online classification tasks. The Perceptron becomes especially effective when it is use

d in conjunction with kernels. However, a common difficulty encountered when implementing kernel-based online algorithms is the amount of memory required to store the online hypothesis, which may grow unboundedly. In this paper we present and analyze the Forgetron algorithm for kernel-based online learning on a fixed memory budget. To our knowledge, this is the first online learning algorithm which, on one hand, maintains a strict limit on the number of examples it stores while, on the other hand, entertains a relative mistake bound. In addition to the formal results, we also present experiments with real datasets which underscore the merits of our approach.
****
Online Discovery and Learning of Predictive State Representations
Peter Mccracken, Michael Bowling
Predictive state representations (PSRs) are a method of modeling dynamical systems using only observable data, such as actions and observations, to describe their model. PSRs use predictions about the outcome of future tests to summarize the system state. The best existing techniques for discovery and learning of PSRs use a Monte Carlo approach to explicitly estimate these outcome probabilities. In this paper, we present a new algorithm for discovery and learning of PSRs that uses a gradient descent approach to compute the predictions for the current state. The algorithm takes advantage of the large amount of structure inherent in a valid prediction matrix to constrain its predictions. Furthermore, the algorithm can be used online by an agent to constantly improve its prediction quality; something that current state of the art discovery and learning algorithms are unable to do. We give empirical results to show that our constrained gradient algorithm is able to discover core tests using very small amounts of data, and with larger amounts of data can compute accurate predictions of the system dynamics.
****
Hyperparameter and Kernel Learning for Graph Based Semi-Supervised Classification
Ashish Kapoor, Hyungil Ahn, Yuan Qi, Rosalind Picard
There have been many graph-based approaches for semi-supervised clas- si■cation. One problem is that of hyperparameter learning: performance depends greatly on the hyperparameters of the similarity graph, trans- formation of the graph Laplacian and the noise model. We present a Bayesian framework for learning hyperparameters for graph-based semi- supervised classi■cation. Given some labeled data, which can contain inaccurate labels, we pose the semi-supervised classi■cation as an in- ference problem over the unknown labels. Expectation Propagation is used for approximate inference and the mean of the posterior is used for classi■cation. The hyperparameters are learned using EM for evidence maximization. We also show that the posterior mean can be written in terms of the kernel matrix, providing a Bayesian classi■er to classify new points. Tests on synthetic and real datasets show cases where there are signi■cant improvements in performance over the existing approaches.
****
Fusion of Similarity Data in Clustering
Tilman Lange, Joachim Buhmann
Fusing multiple information sources can yield signi■cant bene■ts to suc- cessfully accomplish learning tasks. Many studies have focussed on fus- ing information in supervised learning contexts. We present an approach to utilize multiple information sources in the form of similarity data for unsupervised learning. Based on similarity information, the clustering task is phrased as a non-negative matrix factorization problem of a mix- ture of similarity measurements. The tradeoff between the informative- ness of data sources and the sparseness of their mixture is controlled by an entropy-based weighting mechanism. For the purpose of model se- lection, a stability-based approach is employed to ensure the selection of the most self-consistent hypothesis. The experiments demonstrate the performance of the method on toy as well as real world data sets.
****
Sensory Adaptation within a Bayesian Framework for Perception
Alan A. Stocker, Eero Simoncelli

We extend a previously developed Bayesian framework for perception to account fo
r sensory adaptation. We ■rst note that the perceptual ef- fects of adaptation s
eems inconsistent with an adjustment of the inter- nally represented prior distr
ibution. Instead, we postulate that adaptation increases the signal-to-noise rat
io of the measurements by adapting the operational range of the measurement stag
e to the input range. We show that this changes the likelihood function in such
a way that the Bayesian estimator model can account for reported perceptual beha
vior. In particu- lar, we compare the model's predictions to human motion discri
mination data and demonstrate that the model accounts for the commonly observed
perceptual adaptation effects of repulsion and enhanced discriminability.
****
Radial Basis Function Network for Multi-task Learning
Xuejun Liao, Lawrence Carin
We extend radial basis function (RBF) networks to the scenario in which multiple
 correlated tasks are learned simultaneously, and present the cor- responding le
arning algorithms. We develop the algorithms for learn- ing the network structur
e, in either a supervised or unsupervised manner. Training data may also be acti
vely selected to improve the network's gen- eralization to test data. Experiment
al results based on real data demon- strate the advantage of the proposed algori
thms and support our conclu- sions.
****
Prediction and Change Detection
Mark Steyvers, Scott Brown
We measure the ability of human observers to predict the next datum in a sequen
ce that is generated by a simple statistical process undergoing change at rando
m points in time. Accurate performance in this task requires the identification
 of changepoints. We assess individual differences between observers both empir
ically, and using two kinds of models: a Bayesian approach for change detection
  and a family of cognitively plausible fast and frugal models. Some individual
s detect too many changes and hence perform sub-optimally due to excess variab
ility. Other individuals do not detect enough changes, and perform sub-optimall
y because they fail to notice short-term temporal trends.
****
Gaussian Process Dynamical Models
Jack Wang, Aaron Hertzmann, David J. Fleet
This paper introduces Gaussian Process Dynamical Models (GPDM) for nonlinear tim
e series analysis. A GPDM comprises a low-dimensional latent space with associat
ed dynamics, and a map from the latent space to an observation space. We margina
lize out the model parameters in closed-form, using Gaussian Process (GP) priors
 for both the dynamics and the observation mappings. This results in a nonparame
tric model for dynamical systems that accounts for uncertainty in the model. We
demonstrate the approach on human motion capture data in which each pose is 62-d
imensional. Despite the use of small data sets, the GPDM learns an effective rep
resentation of the nonlinear dynamics in these spaces. Webpage: http://www.dgp.t
oronto.edu/
****
Rodeo: Sparse Nonparametric Regression in High Dimensions
Larry Wasserman, John Lafferty
We present a method for nonparametric regression that performs bandwidth selecti
on and variable selection simultaneously. The approach is based on the technique
 of incrementally decreasing the bandwidth in directions where the gradient of t
he estimator with respect to bandwidth is large. When the unknown function satis
fies a sparsity condition, our approach avoids the curse of dimensionality, achi
eving the optimal minimax rate of convergence, up to logarithmic factors, as if
the relevant variables were known in advance. The method--called rodeo (regulari
zation of derivative expectation operator)--conducts a sequence of hypothesis te
sts, and is easy to implement. A modified version that replaces hard with soft t
hresholding effectively solves a sequence of lasso problems.
****
Pattern Recognition from One Example by Chopping

Francois Fleuret, Gilles Blanchard
We investigate the learning of the appearance of an object from a single image o
f it. Instead of using a large number of pictures of the object to recognize, we
 use a labeled reference database of pictures of other ob- jects to learn invari
ance to noise and variations in pose and illumination. This acquired knowledge i
s then used to predict if two pictures of new objects, which do not appear on th
e training pictures, actually display the same object. We propose a generic sche
me called chopping to address this task. It relies on hundreds of random binary
splits of the training set chosen to keep together the images of any given objec
t. Those splits are extended to the complete image space with a simple learning
algorithm. Given two images, the responses of the split predictors are combined
with a Bayesian rule into a posterior probability of similarity. Experiments wit
h the COIL-100 database and with a database of 150 de- graded LATEX symbols comp
are our method to a classical learning with several examples of the positive cla
ss and to a direct learning of the sim- ilarity.
****
Hot Coupling: A Particle Approach to Inference and Normalization on Pairwise Und
irected Graphs
Firas Hamze, Nando de Freitas
This paper presents a new sampling algorithm for approximating func- tions of va
riables representable as undirected graphical models of arbi- trary connectivity
 with pairwise potentials, as well as for estimating the notoriously dif(cid:2)c
ult partition function of the graph. The algorithm (cid:2)ts into the framework
of sequential Monte Carlo methods rather than the more widely used MCMC, and rel
ies on constructing a sequence of in- termediate distributions which get closer
to the desired one. While the idea of using (cid:147)tempered(cid:148) proposals
 is known, we construct a novel se- quence of target distributions where, rather
 than dropping a global tem- perature parameter, we sequentially couple individu
al pairs of variables that are, initially, sampled exactly from a spanning tree
of the variables. We present experimental results on inference and estimation of
 the parti- tion function for sparse and densely-connected graphs.
****
Separation of Music Signals by Harmonic Structure Modeling
Yun-gang Zhang, Chang-shui Zhang
Separation of music signals is an interesting but difficult problem. It is helpf
ul for many other music researches such as audio content analysis. In this paper
, a new music signal separation method is proposed, which is based on harmonic s
tructure modeling. The main idea of harmonic structure modeling is that the harm
onic structure of a music signal is stable, so a music signal can be represented
 by a harmonic structure model. Accordingly, a corresponding separation algorith
m is proposed. The main idea is to learn a harmonic structure model for each mus
ic signal in the mixture, and then separate signals by using these models to dis
tinguish harmonic structures of different signals. Experimental results show tha
t the algorithm can separate signals and obtain not only a very high Signalto-No
ise Ratio (SNR) but also a rather good subjective audio quality.
****
Learning Minimum Volume Sets
Clayton Scott, Robert Nowak
Given a probability measure P and a reference measure μ, one is often interested
 in the minimum μ-measure set with P -measure at least α. Minimum volume sets of
 this type summarize the regions of greatest probability mass of P , and are use
ful for detecting anoma- lies and constructing con■dence regions. This paper add
resses the problem of estimating minimum volume sets based on independent sample
s distributed according to P . Other than these samples, no other information is
 available regarding P , but the reference mea- sure μ is assumed to be known. W
e introduce rules for estimating minimum volume sets that parallel the empirical
 risk minimization and structural risk minimization principles in classi■cation.
 As in classi■cation, we show that the performances of our estimators are contro
lled by the rate of uniform convergence of empirical to true probabilities over
the class from which the estimator is drawn. Thus we obtain ■nite sample size pe

rformance bounds in terms of VC dimension and related quantities. We also demonstrate strong universal consistency and an oracle inequality. Estimators based on histograms and dyadic partitions illustrate the proposed rules.
****
Spectral Bounds for Sparse PCA: Exact and Greedy Algorithms
Baback Moghaddam, Yair Weiss, Shai Avidan
Sparse PCA seeks approximate sparse "eigenvectors" whose projections capture the maximal variance of data. As a cardinality-constrained and non-convex optimization problem, it is NP-hard and is encountered in a wide range of applied fields, from bio-informatics to finance. Recent progress has focused mainly on continuous approximation and convex relaxation of the hard cardinality constraint. In contrast, we consider an alternative discrete spectral formulation based on variational eigenvalue bounds and provide an effective greedy strategy as well as provably optimal solutions using branch-and-bound search. Moreover, the exact methodology used reveals a simple renormalization step that improves approximate solutions obtained by any continuous method. The resulting performance gain of discrete algorithms is demonstrated on real-world benchmark data and in extensive Monte Carlo evaluation trials.
****
A Domain Decomposition Method for Fast Manifold Learning
Zhenyue Zhang, Hongyuan Zha
We propose a fast manifold learning algorithm based on the methodology of domain decomposition. Starting with the set of sample points partitioned into two subdomains, we develop the solution of the interface problem that can glue the embeddings on the two subdomains into an embedding on the whole domain. We provide a detailed analysis to assess the errors produced by the gluing process using matrix perturbation theory. Numerical examples are given to illustrate the efficiency and effectiveness of the proposed methods.
****
Generalized Nonnegative Matrix Approximations with Bregman Divergences
Suvrit Sra, Inderjit Dhillon
Nonnegative matrix approximation (NNMA) is a recent technique for dimensionality reduction and data analysis that yields a parts based, sparse nonnegative representation for nonnegative input data. NNMA has found a wide variety of applications, including text analysis, document clustering, face/image recognition, language modeling, speech processing and many others. Despite these numerous applications, the algorithmic development for computing the NNMA factors has been relatively deficient. This paper makes algorithmic progress by modeling and solving (using multiplicative updates) new generalized NNMA problems that minimize Bregman divergences between the input matrix and its lowrank approximation. The multiplicative update formulae in the pioneering work by Lee and Seung [11] arise as a special case of our algorithms. In addition, the paper shows how to use penalty functions for incorporating constraints other than nonnegativity into the problem. Further, some interesting extensions to the use of "link" functions for modeling nonlinear relationships are also discussed.
****
Predicting EMG Data from M1 Neurons with Variational Bayesian Least Squares
Jo-anne Ting, Aaron D'souza, Kenji Yamamoto, Toshinori Yoshioka, Donna Hoffman, Shinji Kakei, Lauren Sergio, John Kalaska, Mitsuo Kawato
An increasing number of projects in neuroscience requires the sta- tistical analysis of high dimensional data sets, as, for instance, in predicting behavior from neural ■ring or in operating arti■cial de- vices from brain recordings in brain-machine interfaces. Linear analysis techniques remain prevalent in such cases, but classical linear regression approaches are often numerically too fragile in high dimensions. In this paper, we address the question of whether EMG data collected from arm movements of monkeys can be faith- fully reconstructed with linear approaches from neural activity in primary motor cortex (M1). To achieve robust data analysis, we develop a full Bayesian approach to linear regression that auto- matically detects and excludes irrelevant features in the data, reg- ularizing against over■tting. In comparison with ordinary least squares, stepwise reg

ression, partial least squares, LASSO regres- sion and a brute force combinatorial search for the most predictive input features in the data, we demonstrate that the new Bayesian method o■ers a superior mixture of characteristics in terms of reg- ularization against over■tting, computational e■ciency and ease of use, demonstrating its potential as a drop-in replacement for other linear regression techniques. As neuroscienti■c results, our anal- yses demonstrate that EMG data can be well predicted from M1 neurons, further opening the path for possible real-time interfaces between brains and machines.
****
Comparing the Effects of Different Weight Distributions on Finding Sparse Representations
Bhaskar Rao, David Wipf
Given a redundant dictionary of basis vectors (or atoms), our goal is to ■nd maximally sparse representations of signals. Previously, we have argued that a sparse Bayesian learning (SBL) framework is particularly well-suited for this task, showing that it has far fewer local minima than other Bayesian-inspired strategies. In this paper, we provide further evi- dence for this claim by proving a restricted equivalence condition, based on the distribution of the nonzero generating model weights, whereby the SBL solution will equal the maximally sparse representation. We also prove that if these nonzero weights are drawn from an approximate Jef- freys prior, then with probability approaching one, our equivalence con- dition is satis■ed. Finally, we motivate the worst-case scenario for SBL and demonstrate that it is still better than the most widely used sparse rep- resentation algorithms. These include Basis Pursuit (BP), which is based on a convex relaxation of the ■0 (quasi)-norm, and Orthogonal Match- ing Pursuit (OMP), a simple greedy strategy that iteratively selects basis vectors most aligned with the current residual.
****
Goal-Based Imitation as Probabilistic Inference over Graphical Models
Deepak Verma, Rajesh PN Rao
Humans are extremely adept at learning new skills by imitating the actions of others. A progression of imitative abilities has been observed in children, ranging from imitation of simple body movements to goalbased imitation based on inferring intent. In this paper, we show that the problem of goal-based imitation can be formulated as one of inferring goals and selecting actions using a learned probabilistic graphical model of the environment. We first describe algorithms for planning actions to achieve a goal state using probabilistic inference. We then describe how planning can be used to bootstrap the learning of goal-dependent policies by utilizing feedback from the environment. The resulting graphical model is then shown to be powerful enough to allow goal-based imitation. Using a simple maze navigation task, we illustrate how an agent can infer the goals of an observed teacher and imitate the teacher even when the goals are uncertain and the demonstration is incomplete.
****
Policy-Gradient Methods for Planning
Douglas Aberdeen
Probabilistic temporal planning attempts to ■nd good policies for acting in domains with concurrent durative tasks, multiple uncertain outcomes, and limited resources. These domains are typically modelled as Markov decision problems and solved using dynamic programming methods. This paper demonstrates the application of reinforcement learning — in the form of a policy-gradient method — to these domains. Our emphasis is large domains that are infeasible for dynamic programming. Our ap- proach is to construct simple policies, or agents, for each planning task. The result is a general probabilistic temporal planner, named the Factored Policy-Gradient Planner (FPG-Planner), which can handle hundreds of tasks, optimising for probability of success, duration, and resource use.
****
Message passing for task redistribution on sparse graphs
K. Y. Michael Wong, David Saad, Zhuo Gao
The problem of resource allocation in sparse graphs with real variables is studi

ed using methods of statistical physics. An ef■cient distributed algorithm is de vised on the basis of insight gained from the analysis and is examined using num erical simulations, showing excellent performance and full agreement with the th eoretical results.
****
Neuronal Fiber Delineation in Area of Edema from Diffusion Weighted MRI
Ofer Pasternak, Nathan Intrator, Nir Sochen, Yaniv Assaf
Diffusion Tensor Magnetic Resonance Imaging (DT-MRI) is a non inva- sive method for brain neuronal ■bers delineation. Here we show a mod- i■cation for DT-MRI th at allows delineation of neuronal ■bers which are in■ltrated by edema. We use th e Muliple Tensor Variational (MTV) framework which replaces the diffusion model of DT-MRI with a mul- tiple component model and ■ts it to the signal attenuation with a vari- ational regularization mechanism. In order to reduce free water co n- tamination we estimate the free water compartment volume fraction in each vox el, remove it, and then calculate the anisotropy of the remaining compartment. T he variational framework was applied on data collected with conventional clinica l parameters, containing only six diffusion di- rections. By using the variation al framework we were able to overcome the highly ill posed ■tting. The results s how that we were able to ■nd ■bers that were not found by DT-MRI.
****
A Computational Model of Eye Movements during Object Class Detection
Wei Zhang, Hyejin Yang, Dimitris Samaras, Gregory Zelinsky
We present a computational model of human eye movements in an ob- ject class det ection task. The model combines state-of-the-art computer vision object class de tection methods (SIFT features trained using Ad- aBoost) with a biologically pla usible model of human eye movement to produce a sequence of simulated ■xations, culminating with the acqui- sition of a target. We validated the model by compar ing its behavior to the behavior of human observers performing the identical obj ect class detection task (looking for a teddy bear among visually complex non- t arget objects). We found considerable agreement between the model and human data in multiple eye movement measures, including number of ■xations, cumulative pro bability of ■xating the target, and scanpath distance.
****
Efficient Unsupervised Learning for Localization and Detection in Object Categor ies
Nicolas Loeff, Himanshu Arora, Alexander Sorokin, David Forsyth
We describe a novel method for learning templates for recognition and localizati on of objects drawn from categories. A generative model repre- sents the con■gur ation of multiple object parts with respect to an object coordinate system; thes e parts in turn generate image features. The com- plexity of the model in the nu mber of features is low, meaning our model is much more ef■cient to train than c omparative methods. Moreover, a variational approximation is introduced that all ows learning to be or- ders of magnitude faster than previous approaches while i ncorporating many more features. This results in both accuracy and localization im- provements. Our model has been carefully tested on standard datasets; we com pare with a number of recent template models. In particular, we demonstrate stat e-of-the-art results for detection and localization.
****
Beyond Gaussian Processes: On the Distributions of Infinite Networks
Ricky Der, Daniel Lee
A general analysis of the limiting distribution of neural network functions is p erformed, with emphasis on non-Gaussian limits. We show that with i.i.d. symmetr ic stable output weights, and more generally with weights distributed from the n ormal domain of attraction of a stable variable, that the neural functions conve rge in distribution to stable processes. Conditions are also investigated under which Gaussian limits do occur when the weights are independent but not identica lly distributed. Some particularly tractable classes of stable distributions are examined, and the possibility of learning with such processes.
****
Preconditioner Approximations for Probabilistic Graphical Models

John Lafferty, Pradeep Ravikumar
We present a family of approximation techniques for probabilistic graphical mode
ls, based on the use of graphical preconditioners developed in the scientific co
mputing literature. Our framework yields rigorous upper and lower bounds on even
t probabilities and the log partition function of undirected graphical models, u
sing non-iterative procedures that have low time complexity. As in mean field ap
proaches, the approximations are built upon tractable subgraphs; however, we rec
ast the problem of optimizing the tractable distribution parameters and approxim
ate inference in terms of the well-studied linear systems problem of obtaining a
 good matrix preconditioner. Experiments are presented that compare the new appr
oximation schemes to variational methods.
****
Structured Prediction via the Extragradient Method
Ben Taskar, Simon Lacoste-Julien, Michael Jordan
We present a simple and scalable algorithm for large-margin estima- tion of stru
ctured models, including an important class of Markov net- works and combinatori
al models. We formulate the estimation problem as a convex-concave saddle-point
problem and apply the extragradient method, yielding an algorithm with linear co
nvergence using simple gra- dient and projection calculations. The projection st
ep can be solved us- ing combinatorial algorithms for min-cost quadratic ■ow. Th
is makes the approach an ef■cient alternative to formulations based on reduction
s to a quadratic program (QP). We present experiments on two very different stru
ctured prediction tasks: 3D image segmentation and word alignment, illustrating
the favorable scaling properties of our algorithm.
****
Noise and the two-thirds power Law
Uri Maoz, Elon Portugaly, Tamar Flash, Yair Weiss
The two-thirds power law, an empirical law stating an inverse non-linear relatio
nship between the tangential hand speed and the curvature of its trajectory duri
ng curved motion, is widely acknowledged to be an invariant of upper-limb moveme
nt. It has also been shown to exist in eyemotion, locomotion and was even demons
trated in motion perception and prediction. This ubiquity has fostered various a
ttempts to uncover the origins of this empirical relationship. In these it was g
enerally attributed either to smoothness in hand- or joint-space or to the resul
t of mechanisms that damp noise inherent in the motor system to produce the smoo
th trajectories evident in healthy human motion. We show here that white Gaussia
n noise also obeys this power-law. Analysis of signal and noise combinations sho
ws that trajectories that were synthetically created not to comply with the powe
r-law are transformed to power-law compliant ones after combination with low lev
els of noise. Furthermore, there exist colored noise types that drive non-power-
law trajectories to power-law compliance and are not affected by smoothing. Thes
e results suggest caution when running experiments aimed at verifying the power-
law or assuming its underlying existence without proper analysis of the noise. O
ur results could also suggest that the power-law might be derived not from smoot
hness or smoothness-inducing mechanisms operating on the noise inherent in our m
otor system but rather from the correlated noise which is inherent in this motor
 system.
****
From Lasso regression to Feature vector machine
Fan Li, Yiming Yang, Eric Xing
Lasso regression tends to assign zero weights to most irrelevant or redundant fe
atures, and hence is a promising technique for feature selection. Its limitation
, however, is that it only offers solutions to linear models. Kernel machines wi
th feature scaling techniques have been studied for feature selection with non-l
inear models. However, such approaches require to solve hard non-convex optimiza
tion problems. This paper proposes a new approach named the Feature Vector Machi
ne (FVM). It reformulates the standard Lasso regression into a form isomorphic t
o SVM, and this form can be easily extended for feature selection with non-linea
r models by introducing kernels defined on feature vectors. FVM generates sparse
 solutions in the nonlinear feature space and it is much more tractable compared

to feature scaling kernel machines. Our experiments with FVM on simulated data
show encouraging results in identifying the small number of dominating features
that are non-linearly correlated to the response, a task the standard Lasso fail
s to complete.
****
Maximum Margin Semi-Supervised Learning for Structured Variables
Y. Altun, D. McAllester, M. Belkin
Many real-world classi■cation problems involve the prediction of multiple inter-
dependent variables forming some structural depen- dency. Recent progress in mac
hine learning has mainly focused on supervised classi■cation of such structured
variables. In this paper, we investigate structured classi■cation in a semi-supe
rvised setting. We present a discriminative approach that utilizes the intrinsic
ge- ometry of input patterns revealed by unlabeled data points and we derive a
maximum-margin formulation of semi-supervised learning for structured variables.
Unlike transductive algorithms, our for- mulation naturally extends to new test
points.
****
Generalization in Clustering with Unobserved Features
Eyal Krupka, Naftali Tishby
We argue that when objects are characterized by many attributes, clustering them
on the basis of a relatively small random subset of these attributes can captur
e information on the unobserved attributes as well. Moreover, we show that under
mild technical conditions, clustering the objects on the basis of such a random
subset performs almost as well as clustering with the full attribute set. We pr
ove a finite sample generalization theorems for this novel learning scheme that
extends analogous results from the supervised learning setting. The scheme is de
monstrated for collaborative filtering of users with movies rating as attributes
.
****
Variable KD-Tree Algorithms for Spatial Pattern Search and Discovery
Jeremy Kubica, Joseph Masiero, Robert Jedicke, Andrew Connolly, Andrew Moore
In this paper we consider the problem of finding sets of points that conform to
a given underlying model from within a dense, noisy set of observations. This pr
oblem is motivated by the task of efficiently linking faint asteroid detections,
but is applicable to a range of spatial queries. We survey current tree-based a
pproaches, showing a trade-off exists between single tree and multiple tree algo
rithms. To this end, we present a new type of multiple tree algorithm that uses
a variable number of trees to exploit the advantages of both approaches. We empi
rically show that this algorithm performs well using both simulated and astronom
ical data.
****
Neural mechanisms of contrast dependent receptive field size in V1
Jim Wielaard, Paul Sajda
Based on a large scale spiking neuron model of the input layers 4C and of macaq
ue, we identify neural mechanisms for the observed contrast dependent receptive
field size of V1 cells. We observe a rich variety of mechanisms for the phenomen
on and analyze them based on the relative gain of excitatory and inhibitory syna
ptic inputs. We observe an average growth in the spatial extent of excitation an
d inhibition for low contrast, as predicted from phenomenological models. Howeve
r, contrary to phenomenological models, our simulation results suggest this is n
either sufficient nor necessary to explain the phenomenon.
****
Benchmarking Non-Parametric Statistical Tests
Mikaela Keller, Samy Bengio, Siew Wong
Although non-parametric tests have already been proposed for that purpose, stati
stical significance tests for non-standard measures (different from the classifi
cation error) are less often used in the literature. This paper is an attempt at
empirically verifying how these tests compare with more classical tests, on var
ious conditions. More precisely, using a very large dataset to estimate the whol
e "population", we analyzed the behavior of several statistical test, varying th

e class unbalance, the compared models, the performance measure, and the sample size. The main result is that providing big enough evaluation sets non-parametric tests are relatively reliable in all conditions.
****
Convergence and Consistency of Regularized Boosting Algorithms with Stationary B-Mixing Observations
Aurelie C. Lozano, Sanjeev Kulkarni, Robert E. Schapire
We study the statistical convergence and consistency of regularized Boosting methods, where the samples are not independent and identi- cally distributed (i.i.d.) but come from empirical processes of stationary β-mixing sequences. Utilizing a technique that constructs a sequence of independent blocks close in distribution to the original samples, we prove the consistency of the composite classi■ers resulting from a regulariza- tion achieved by restricting the 1-norm of the base classi■ers' weights. When compared to the i.i.d. case, the nature of sampling manifests in the consistency result only through generalization of the original condition on the growth of the regularization parameter.
****
A Cortically-Plausible Inverse Problem Solving Method Applied to Recognizing Static and Kinematic 3D Objects
David Arathorn
Recent neurophysiological evidence suggests the ability to interpret biological motion is facilitated by a neuronal "mirror system" which maps visual inputs to the pre-motor cortex. If the common architecture and circuitry of the cortices is taken to imply a common computation across multiple perceptual and cognitive modalities, this visual-motor interaction might be expected to have a unified computational basis. Two essential tasks underlying such visual-motor cooperation are shown here to be simply expressed and directly solved as transformation-discovery inverse problems: (a) discriminating and determining the pose of a primed 3D object in a real-world scene, and (b) interpreting the 3D configuration of an articulated kinematic object in an image. The recently developed map-seeking method provides tractable, cortically-plausible solution to these and a variety of other inverse problems which can be posed as the discovery of a composition of transformations between two patterns. The method relies on an ordering property of superpositions and on decomposition of the transformation spaces inherent in the generating processes of the problem.
****
Dynamic Social Network Analysis using Latent Space Models
Purnamrita Sarkar, Andrew Moore
This paper explores two aspects of social network modeling. First, we generalize a successful static model of relationships into a dynamic model that accounts for friendships drifting over time. Second, we show how to make it tractable to learn such models from data, even as the number of entities n gets large. The generalized model associates each entity with a point in p-dimensional Euclidian latent space. The points can move as time progresses but large moves in latent space are improb- able. Observed links between entities are more likely if the entities are close in latent space. We show how to make such a model tractable (sub-quadratic in the number of entities) by the use of appropriate kernel func- tions for similarity in latent space; the use of low dimensional kd-trees; a new ef(cid:2)cient dynamic adaptation of multidimensional scaling for a (cid:2)rst pass of approximate projection of entities into latent space; and an ef(cid:2)- cient conjugate gradient update rule for non-linear local optimization in which amortized time per entity during an update is O(log n). We use both synthetic and real-world data on upto 11,000 entities which indicate linear scaling in computation time and improved performance over four alternative approaches. We also illustrate the system operating on twelve years of NIPS co-publication data. We present a detailed version of this work in [1].
****
Principles of real-time computing with feedback applied to cortical microcircuit

models
Wolfgang Maass, Prashant Joshi, Eduardo Sontag
The network topology of neurons in the brain exhibits an abundance of feedback c onnections, but the computational function of these feedback connections is larg ely unknown. We present a computational theory that characterizes the gain in co mputational power achieved through feedback in dynamical systems with fading mem ory. It implies that many such systems acquire through feedback universal comput ational capabilities for analog computing with a non-fading memory. In particula r, we show that feedback enables such systems to process time-varying input stre ams in diverse ways according to rules that are implemented through internal sta tes of the dynamical system. In contrast to previous attractor-based computation al models for neural networks, these ■exible internal states are high-dimensiona l attractors of the circuit dynamics, that still allow the circuit state to abso rb new information from online input streams. In this way one arrives at novel m odels for working memory, integration of evidence, and reward expectation in cor tical circuits. We show that they are applicable to circuits of conductance-base d Hodgkin-Huxley (HH) neurons with high levels of noise that re■ect experimental data on in- vivo conditions.
****
Location-based activity recognition
Lin Liao, Dieter Fox, Henry Kautz
Learning patterns of human behavior from sensor data is extremely important for high-level activity inference. We show how to extract and label a person's activ ities and significant places from traces of GPS data. In contrast to existing te chniques, our approach simultaneously detects and classifies the significant loc ations of a person and takes the highlevel context into account. Our system uses relational Markov networks to represent the hierarchical activity model that en codes the complex relations among GPS readings, activities and significant place s. We apply FFT-based message passing to perform efficient summation over large numbers of nodes in the networks. We present experiments that show significant i mprovements over existing techniques.
****
Modeling Memory Transfer and Saving in Cerebellar Motor Learning
Naoki Masuda, Shun-ichi Amari
There is a long-standing controversy on the site of the cerebellar motor learnin g. Different theories and experimental results suggest that either the cerebella r flocculus or the brainstem learns the task and stores the memory. With a dynam ical system approach, we clarify the mechanism of transferring the memory genera ted in the flocculus to the brainstem and that of so-called savings phenomena. T he brainstem learning must comply with a sort of Hebbian rule depending on Purki nje-cell activities. In contrast to earlier numerical models, our model is simpl e but it accommodates explanations and predictions of experimental situations as qualitative features of trajectories in the phase space of synaptic weights, wi thout fine parameter tuning.
****
TD(0) Leads to Better Policies than Approximate Value Iteration
Benjamin Roy
We consider approximate value iteration with a parameterized approximator in whi ch the state space is partitioned and the optimal cost-to-go function over each partition is approximated by a constant. We establish performance loss bounds fo r policies derived from approximations associated with fixed points. These bound s identify benefits to having projection weights equal to the invariant distribu tion of the resulting policy. Such projection weighting leads to the same fixed points as TD(0). Our analysis also leads to the first performance loss bound for approximate value iteration with an average cost objective.
****
Gradient Flow Independent Component Analysis in Micropower VLSI
Abdullah Celik, Milutin Stanacevic, Gert Cauwenberghs
We present micropower mixed-signal VLSI hardware for real-time blind separation and localization of acoustic sources. Gradient flow representation of the travel

ing wave signals acquired over a miniature (1cm diameter) array of four micropho nes yields linearly mixed instantaneous observations of the time-differentiated sources, separated and localized by independent component analysis (ICA). The gr adient flow and ICA processors each measure 3mm 3mm in 0.5 m CMOS, and consume 54 W and 180 W power, respectively, from a 3 V supply at 16 ks/s sampling rate. Experiments demonstrate perceptually clear (12dB) separation and precise localiz ation of two speech sources presented through speakers positioned at 1.5m from t he array on a conference room table. Analysis of the multipath residuals shows t hat they are spectrally diffuse, and void of the direct path.
****
An Alternative Infinite Mixture Of Gaussian Process Experts
Edward Meeds, Simon Osindero
We present an in■nite mixture model in which each component com- prises a multiv ariate Gaussian distribution over an input space, and a Gaussian Process model o ver an output space. Our model is neatly able to deal with non-stationary covari ance functions, discontinuities, multi- modality and overlapping output signals. The work is similar to that by Rasmussen and Ghahramani [1]; however, we use a full generative model over input and output space rather than just a conditional model. This al- lows us to deal with incomplete data, to perform inference over inverse functional mappings as well as for regression, and also leads to a more powerful and consistent Bayesian speci■cation of the effective ‘gating network’ for the different experts.
****
Silicon growth cones map silicon retina
Brian Taba, Kwabena Boahen
We demonstrate the ■rst fully hardware implementation of retinotopic self-organi zation, from photon transduction to neural map formation. A silicon retina trans duces patterned illumination into correlated spike trains that drive a populatio n of silicon growth cones to automatically wire a topographic mapping by migrati ng toward sources of a diffusible guidance cue that is released by postsynaptic spikes. We varied the pat- tern of illumination to steer growth cones projected by different retinal ganglion cell types to self-organize segregated or coordina ted retinotopic maps.
****
Bayesian models of human action understanding
Chris Baker, Rebecca Saxe, Joshua Tenenbaum
We present a Bayesian framework for explaining how people reason about and predi ct the actions of an intentional agent, based on observ- ing its behavior. Actio n-understanding is cast as a problem of inverting a probabilistic generative mod el, which assumes that agents tend to act rationally in order to achieve their g oals given the constraints of their en- vironment. Working in a simple sprite-wo rld domain, we show how this model can be used to infer the goal of an agent and predict how the agent will act in novel situations or when environmental constr aints change. The model provides a qualitative account of several kinds of infer ences that preverbal infants have been shown to perform, and also ■ts quantita- tive predictions that adult observers make in a new experiment.
****
Learning Influence among Interacting Markov Chains
Dong Zhang, Daniel Gatica-perez, Samy Bengio, Deb Roy
We present a model that learns the influence of interacting Markov chains within a team. The proposed model is a dynamic Bayesian network (DBN) with a two-level structure: individual-level and group-level. Individual level models actions of each player, and the group-level models actions of the team as a whole. Experim ents on synthetic multi-player games and a multi-party meeting corpus show the e ffectiveness of the proposed model.
****
Off-policy Learning with Options and Recognizers
Doina Precup, Cosmin Paduraru, Anna Koop, Richard S. Sutton, Satinder Singh
We introduce a new algorithm for off-policy temporal-difference learning with fu nction approximation that has lower variance and requires less knowledge of the

behavior policy than prior methods. We develop the notion of a recognizer, a fil ter on actions that distorts the behavior policy to produce a related target pol icy with low-variance importance-sampling corrections. We also consider target p olicies that are deviations from the state distribution of the behavior policy, such as potential temporally abstract options, which further reduces variance. T his paper introduces recognizers and their potential advantages, then develops a full algorithm for linear function approximation and proves that its updates ar e in the same direction as on-policy TD updates, which implies asymptotic conver gence. Even though our algorithm is based on importance sampling, we prove that it requires absolutely no knowledge of the behavior policy for the case of state -aggregation function approximators.
****
A Probabilistic Interpretation of SVMs with an Application to Unbalanced Classif ication
Yves Grandvalet, Johnny Mariethoz, Samy Bengio
In this paper, we show that the hinge loss can be interpreted as the neg-log-lik elihood of a semi-parametric model of posterior probabilities. From this point o f view, SVMs represent the parametric component of a semi-parametric model ■tted by a maximum a posteriori estimation pro- cedure. This connection enables to de rive a mapping from SVM scores to estimated posterior probabilities. Unlike prev ious proposals, the sug- gested mapping is interval-valued, providing a set of p osterior probabil- ities compatible with each SVM score. This framework offers a new way to adapt the SVM optimization problem to unbalanced classi■ca- tion, wh en decisions result in unequal (asymmetric) losses. Experiments show improvement s over state-of-the-art procedures.
****
Nonparametric inference of prior probabilities from Bayes-optimal behavior
Liam Paninski
We discuss a method for obtaining a subject's a priori beliefs from his/her beha vior in a psychophysics context, under the assumption that the behavior is (near ly) optimal from a Bayesian perspective. The method is nonparametric in the sens e that we do not assume that the prior belongs to any ■xed class of distribution s (e.g., Gaussian). Despite this increased generality, the method is relatively simple to implement, being based in the simplest case on a linear programming al gorithm, and more generally on a straightforward maximum likelihood or maximum a posteriori formulation, which turns out to be a convex optimization problem (wi th no non-global local maxima) in many important cases. In addition, we develop methods for analyzing the uncertainty of these esti- mates. We demonstrate the a ccuracy of the method in a simple simulated coin-■ipping setting; in particular, the method is able to precisely track the evolution of the subject's posterior distribution as more and more data are observed. We close by brie■y discussing a n interesting connection to recent models of neural population coding.
****
Oblivious Equilibrium: A Mean Field Approximation for Large-Scale Dynamic Games
Gabriel Y. Weintraub, Lanier Benkard, Benjamin Van Roy
We propose a mean-■eld approximation that dramatically reduces the computational complexity of solving stochastic dynamic games. We pro- vide conditions that gu arantee our method approximates an equilibrium as the number of agents grow. We then derive a performance bound to assess how well the approximation performs fo r any given number of agents. We apply our method to an important class of probl ems in ap- plied microeconomics. We show with numerical experiments that we are able to greatly expand the set of economic problems that can be analyzed computa tionally.
****
Dynamical Synapses Give Rise to a Power-Law Distribution of Neuronal Avalanches
Anna Levina, Michael Herrmann
There is experimental evidence that cortical neurons show avalanche activity wit h the intensity of firing events being distributed as a power-law. We present a biologically plausible extension of a neural network which exhibits a power-law avalanche distribution for a wide range of connectivity parameters.

****
From Weighted Classification to Policy Search
Doron Blatt, Alfred Hero
This paper proposes an algorithm to convert a T -stage stochastic decision problem with a continuous state space to a sequence of supervised learning problems. The optimization problem associated with the trajectory tree and random trajectory methods of Kearns, Mansour, and Ng, 2000, is solved using the Gauss-Seidel method. The algorithm breaks a multistage reinforcement learning problem into a sequence of single-stage reinforcement learning subproblems, each of which is solved via an exact reduction to a weighted-classification problem that can be solved using off-the-self methods. Thus the algorithm converts a reinforcement learning problem into simpler supervised learning subproblems. It is shown that the method converges in a finite number of steps to a solution that cannot be further improved by componentwise optimization. The implication of the proposed algorithm is that a plethora of classification methods can be applied to find policies in the reinforcement learning problem.
****
Off-Road Obstacle Avoidance through End-to-End Learning
Urs Muller, Jan Ben, Eric Cosatto, Beat Flepp, Yann Cun
We describe a vision-based obstacle avoidance system for off-road mobile robots. The system is trained from end to end to map raw in put images to steering angles. It is trained in supervised mode to predict the steering angles provided by a human driver during training r uns collected in a wide variety of terrains, weather conditions, lighting conditions, and obstacle types. The robot is a 50cm off-road truck, with two f orwardpointing wireless color cameras. A remote computer process es the video and controls the robot via radio. The learning system is a lar ge 6-layer convolutional network whose input is a single left/right pair of unprocessed low-resolution images. The robot exhibits an excellent ability to detect obstacles and navigate around them in real time at speeds of 2 m/s.
****