Suggested Readings:

1. Shannon Information and **Kolomogorov Complexity**, by Peter Grunwald and Paul Vitanyi, July 22, 2010.

   The elementary theories of Shannon information and Kolmogorov complexity are compared, the extent to which they have a common purpose, and where they are fundamentally different. The focus is on: Shannon entropy versus Kolmogorov complexity, the relation of both to universal coding, Shannon mutual information versus Kolmogorov ('algorithmic') mutual information, **probabilistic sufficient statistic** versus **algorithmic sufficient statistic** (related to lossy compression in the Shannon theory versus **meaningful information** in the Kolmogorov theory), and **rate distortion theory** versus **Kolmogorov's structure function**.

2. An Information Theory Account of **Cognitive Control**, by Jin Fan (http://www.frontiersin.org/people/u/2093)

   Our ability to efficiently **process information** and **generate appropriate responses** depends on the processes collectively called ***cognitive control.*** Despite a considerable focus in the literature on the cognitive control of information processing, **neural mechanisms** underlying control are still unclear, and have not been characterized by considering the quantity of information to be processed. A novel and comprehensive account of cognitive control is proposed using concepts from information theory, which is concerned with communication system analysis and the quantification of information. This account treats the **brain** as an **information-processing entity** where cognitive control and its underlying brain networks play a pivotal role in dealing with conditions of uncertainty. This hypothesis and theory article justifies the validity and properties of such an account and relates experimental findings to the **frontoparietal network** under the framework of information theory.

3. Information Theoretic-based **Interpretation of a Deep Neural Network Approach** in Diagnosing Psychogenic Non-Epileptic Seisures, by Sara Gasparini, et al., Entropy, 2018.

   The use of a deep neural network scheme is proposed to help clinicians solve a difficult diagnosis problem in neurology. The proposed multilayer architecture includes a feature engineering step (from time-frequency transformation), a double compressing stage trained by unsupervised learning, and a classification stage trained by supervised learning. After fine-tuning, the deep network is able to discriminate well the class of patients from controls with around 90% sensitivity and specificity. This deep model gives better classification performance than some other standard discriminative learning algorithms. As **in clinical** problems **there is a need for explaining decisions**, an effort has been carried out to **qualitatively justify the classification results**. The main novelty of this paper is indeed to give an **entropic interpretation of how the deep scheme works and reach the final decision**.

4. On Lower Bounds for **Statistical Learning Theory**, by Po-Ling Loh, Entropy 2017.

In recent years, tools from information theory have played an increasingly prevalent role in statistical machine learning. In addition to developing efficient, computationally feasible algorithms for analyzing complex datasets, it is of theoretical importance to determine **whether such algorithms are "optimal"** in the sense that no other algorithm can lead to **smaller statistical error**. This paper provides a survey of various techniques used to derive **information-theoretic lower bounds for estimation and learning**. We focus on the settings of parameter and function estimation, community recovery, and online learning for multi-armed bandits. A common theme is that **lower bounds** are established by **relating the statistical learning problem to a channel decoding problem**, for which lower bounds may be derived involving information-theoretic quantities such as the mutual information, total variation distance, and **Kullback–Leibler divergence.** We close by discussing the use of information-theoretic quantities to measure independence in machine learning applications ranging from causality to medical imaging, and mention techniques for estimating these quantities efficiently in a data-driven manner.

5. **Secret Sharing** and **Shared Information**, by Johannes Rauh, Entropy 2017.

   Secret sharing is a cryptographic discipline in which the goal is to **distribute information about a secret over a set of participants** in such a way that **only specific authorized combinations of participants together can reconstruct the secret**. Thus, secret sharing schemes are systems of variables in which it is very clearly specified **which subsets have information about the secret**. As such, they provide perfect model systems for **information decompositions**. However, following this intuition too far leads to an information decomposition with **negative partial information** terms, which are **difficult to interpret**. One possible explanation is that the partial information lattice proposed by Williams and Beer is incomplete and has to be extended to incorporate terms corresponding to higher-order redundancy. These results put bounds on information decompositions that follow the partial information framework, and they hint at where the partial information lattice needs to be improved.

6. Digital Image **Stabilization** Method Based on **Variational Mode Decomposition** and **Relative Entropy**, by Duo Ho, et al., Entropy 2017.
   Cameras mounted on vehicles frequently suffer from image shake due to the vehicles' motions. To remove jitter motions and preserve intentional motions, a hybrid digital image stabilization method is proposed that uses variational mode decomposition (VMD) and relative entropy (RE). In this paper, the **global motion vector** (GMV) is initially decomposed into several narrow-banded modes by VMD. REs, which exhibit the difference of probability distribution between two modes, are then calculated to identify the **intentional** and **jitter motion modes.** Finally, the summation of the jitter motion modes constitutes jitter motions, whereas the subtraction of the resulting sum from the GMV represents the intentional motions. The proposed stabilization method is compared with several known methods, namely, medium filter (MF), Kalman filter (KF), wavelet decomposition (MD) method, empirical mode decomposition (EMD)-based method, and enhanced EMD-

based method, to evaluate stabilization performance. Experimental results show that the proposed method outperforms the other stabilization methods.

7. On **Normalized Mutual Information**: Measure Derivations and Properties, by Tarald O. Kvalseth, Entropy 2017.

   Starting with a new formulation for the mutual information (MI) between a pair of events, this paper derives alternative upper bounds and extends those to the case of two discrete random variables. Normalized mutual information (NMI) measures are then obtained from those bounds, emphasizing the use of least upper bounds. Conditional NMI measures are also derived for **three different events and three different random variables.** Since the MI formulation for a pair of events is always nonnegative, it can properly be extended to include weighted MI and NMI measures for pairs of events or for random variables that are analogous to the well-known weighted entropy. This weighted MI is generalized to the case of continuous random variables. Such weighted measures have the advantage over previously proposed measures of always being nonnegative. A simple transformation is derived for the NMI, such that the transformed measures have the value-validity property necessary for making various appropriate comparisons between values of those measures. A numerical example is provided.

8. Functional Analysis of the **Euclidean Distance** between **Probability Distributions**, by Namyong Kim, Entropy 2018.

   Minimization of the Euclidean distance between output distribution and **Dirac delta functions** as a performance criterion is known to match the distribution of system output with delta functions. In the analysis of the algorithm developed based on that criterion and **recursive gradient estimation**, it is revealed in this paper that the minimization process of the cost function has two gradients with different functions; one that forces **spreading of output samples** and the other one that **compels output samples to move close to symbol points**. For investigation the two functions, each gradient is controlled separately through individual normalization of each gradient with their related input. From the analysis and experimental results, it is verified that **one gradient is associated with the role of accelerating initial convergence speed** by **spreading output samples** and **the other gradient is related with lowering the minimum mean squared error (MSE)** by **pulling error samples close together**.

9. A new **Classifier** based on Information Theoretic **Learning with Unlabeled Data**, by Kyu-Hwa Jeong, Neural Networks, 2005.

   Supervised learning is conventionally performed with pairwise input–output labeled data. After the training procedure, the adaptive system's weights are fixed while the testing procedure with unlabeled data is performed. Recently, in an attempt to improve classification performance unlabeled data has been exploited in the machine learning community. In this paper, we present an **information theoretic learning** (ITL) approach based on density divergence minimization to obtain an extended training algorithm using unlabeled data during the testing. The method uses a boosting-like algorithm with an ITL based cost function. Preliminary

simulations suggest that the method has the potential to improve the performance of classifiers in the application phase.

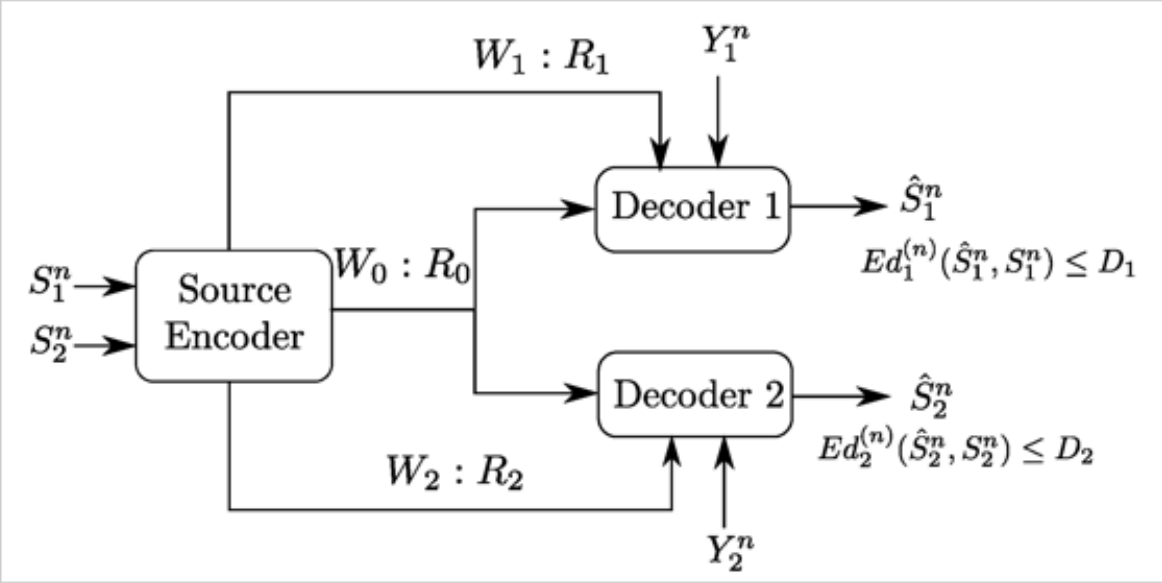
10. **Cosine Similarity Entropy**: Self-Correlation-Based Complexity Analysis of Dynamical Systems, by Theerasak Chanwimalueang, Entropy 2017.

    The nonparametric Sample Entropy (SE) estimator has become a standard for the quantification of structural complexity of nonstationary time series, even in critical cases of unfavorable noise levels. The SE has proven very successful for signals that exhibit a certain degree of the underlying structure, but do not obey standard probability distributions, a typical case in real-world scenarios such as with physiological signals. However, the SE estimates structural complexity based on uncertainty rather than on (self-) correlation, so that, for reliable estimation, the SE requires long data segments, is sensitive to spikes and erratic peaks in data, and owing to its amplitude dependence it exhibits lack of precision for signals with long-term correlations. To this end, we propose a class of new entropy estimators based on the similarity of embedding vectors, evaluated through the angular distance, the Shannon entropy and the coarse-grained scale. Analysis of the effects of embedding dimension, sample size and tolerance shows that the so introduced Cosine Similarity Entropy (CSE) and the enhanced Multiscale Cosine Similarity Entropy (MCSE) are amplitude-independent and therefore superior to the SE when applied to short time series. Unlike the SE, the CSE is shown to yield valid entropy values over a broad range of embedding dimensions. By evaluating the CSE and the MCSE over a variety of benchmark synthetic signals as well as for real-world data (heart rate variability of three different cardiovascular pathologies), the proposed algorithms are demonstrated to be able to quantify degrees of structural complexity in the context of self-correlation over small to large temporal scales, thus offering physically meaningful interpretations and rigor in the understanding the intrinsic properties of the structural complexity of a system, such as the number of its degrees of freedom.

11. **Rate-Distortion** Region of a Gary-Wyner Model with Side Information, by Meryem Benammar, et al., Entropy 2018.

    In this work, we establish a full single-letter characterization of the rate-distortion region of an instance of the Gray–Wyner model with side information at the decoders. Specifically, in this model, an encoder observes a pair of memoryless, arbitrarily correlated, sources (Sn1, Sn2) and communicates with two receivers over an error-free rate-limited link of capacity R0 , as well as error-free rate-limited individual links of capacities R1 to the first receiver and R2 to the second receiver. Both receivers reproduce the source component Sn2 losslessly; and Receiver 1 also reproduces the source component Sn1 lossily, to within some prescribed fidelity level D1. In addition, Receiver 1 and Receiver 2 are equipped, respectively, with memoryless side information sequences Yn1 and Yn2. Important in this setup, the side information sequences are arbitrarily correlated among them, and with the source pair (Sn1, Sn2) ; and are not assumed to exhibit any particular ordering. Furthermore, by specializing the main result to two Heegard–Berger models with *successive refinement* and *scalable coding*, we shed light on the roles of the common and private descriptions that the encoder should produce and the role of

each of the common and private links. We develop intuitions by analyzing the developed single-letter rate-distortion regions of these models, and discuss some insightful binary examples.



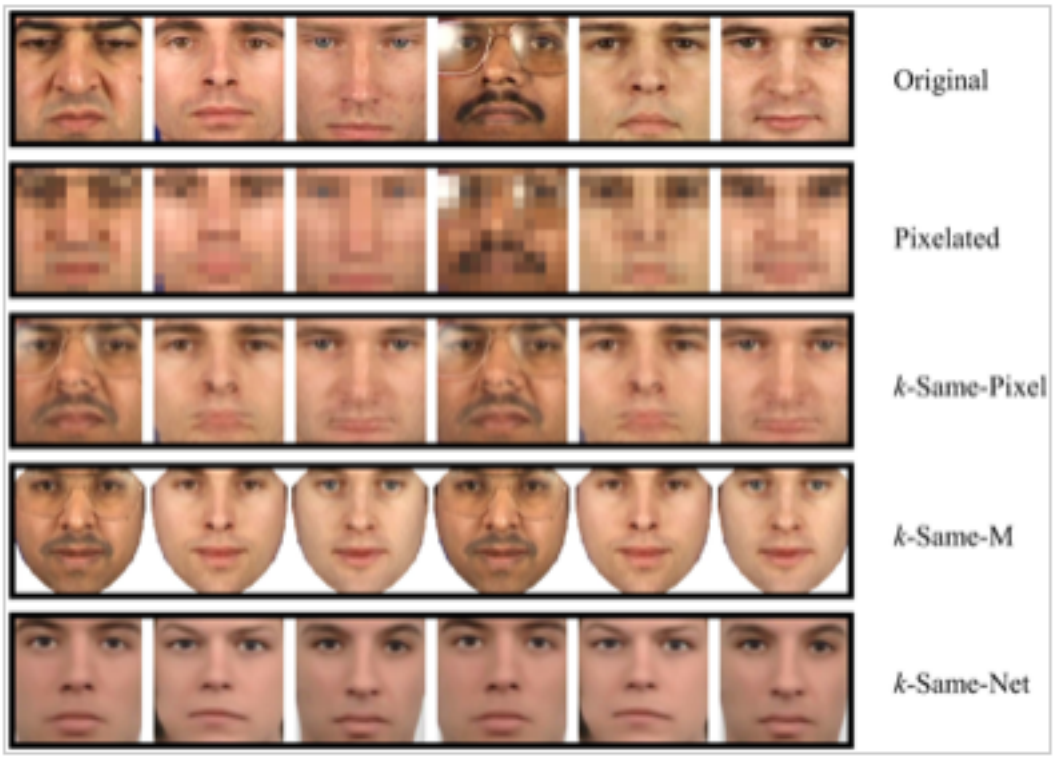
12. **Information Theoretic Learning : Renyi's Entropy and Kernel Perspectives**, by Jose´ C. Principe
ISSN 1613-9011, ISBN 978-1-4419-1569-6
DOI 10.1007/978-1-4419-1570-2
Springer New York Dordrecht Heidelberg London

This book is an outgrowth of ten years of research at the **University of Florida Computational Neuro-Engineering Laboratory (**CNEL) in the general area of statistical signal processing and machine learning. One of the goals of writing the book is exactly to bridge the two fields that share so many common problems and techniques but are not yet effectively collaborating.
Unlike other books that cover the state of the art in a given field, this book cuts across engineering (signal processing) and statistics (machine learning) with a common theme: learning seen from the point of view of information theory with an emphasis on **Renyi's definition of information**. The basic approach is to utilize the information theory descriptors of entropy and divergence as nonparametric cost functions for the design of adaptive systems in unsupervised or supervised training modes. Hence the title: **Information-Theoretic Learning** (ITL). In the course of these studies, we discovered that the main idea enabling a synergistic view as well as algorithmic implementations, does not involve the conventional central moments of the data (mean and covariance). Rather, the core concept is the α-norm of the PDF, in particular its expected value (<=2), which we call the information potential. This operator and related nonparametric estimators link information theory, optimization of adaptive systems, and reproducing kernel Hilbert spaces in a simple and unconventional way.
Due to the pervasive nature of learning, the reading of the material re- quires prior basic knowledge on a broad set of subjects such as information theory, density estimation, adaptive filtering, pattern recognition, reproducing kernel Hilbert spaces (RKHS), and kernel machines. Because there are few researchers with such broad

interests, the first chapter provides, in simple terms, the minimal foundations of information theory, adaptive filtering, and RKHS, while the appendix reviews density estimation. Once the reader is able to grasp these fundamentals, the book develops a nonparametric frame- work that is rich in understanding, setting the stage for the evolution of a new generation of algorithms of varying complexity. This book is therefore useful for professionals who are interested in improving the performance of traditional algorithms as well as researchers who are interested in exploring new approaches to machine learning.

13. K-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification, by Blaz Meden, et al., Entropy 2018.

    Image and video data are today being shared between government entities and other relevant stakeholders on a regular basis and require careful handling of the personal information contained therein. A popular approach to ensure **privacy protection** in such data is the use of **deidentification** techniques, which aim at concealing the identity of individuals in the imagery while still preserving certain aspects of the data after deidentification. In this work, we propose a novel approach towards **face deidentification**, called *k*-Same-Net, which combines recent **Generative Neural Networks (GNNs)** with the well-known ***k*-Anonymity mechanism** and provides formal guarantees regarding privacy protection on a closed set of identities. Our GNN is able to generate synthetic surrogate face images for deidentification by seamlessly combining features of identities used to train the GNN model. Furthermore, it allows us to control the image-generation process with a small set of appearance-related parameters that can be used to alter specific aspects (e.g., facial expressions, age, gender) of the synthesized surrogate images. We demonstrate the feasibility of *k*-Same-Net in comprehensive experiments on the XM2VTS and CK+ datasets. We evaluate the efficacy of the proposed approach through reidentification experiments with recent recognition models and compare our results with competing deidentification techniques from the literature. We also present **facial expression recognition experiments** to demonstrate the utility-preservation capabilities of *k*-Same-Net. Our experimental results suggest that *k*-Same-Net is a viable option for facial deidentification that exhibits several desirable characteristics when compared to existing solutions in this area



14. **ON THE INFORMATION BOTTLENECK THEORY** OF **DEEP LEARNING**, by Andrew M. Saxe, et al., ICLR 2018

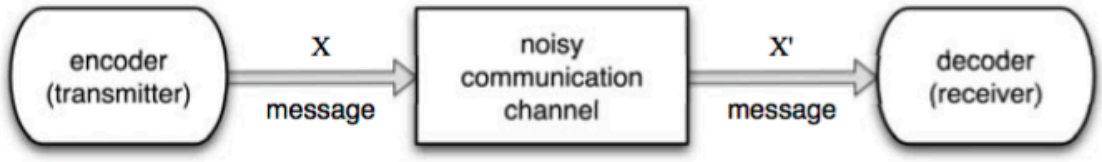
The practical successes of **deep neural networks** have not been matched by theoretical progress that satisfyingly explains their behavior. In this work, we study the **information bottleneck (IB) theory of deep learning**, which makes three specific claims: first, that deep networks undergo **two distinct phases** consisting of an **initial fitting phase** and **a subsequent compression phase;** second, that **the compression phase is causally related to the excellent generalization performance of deep networks**; and third, that **the compression phase occurs due to the diffusion-like behavior of stochastic gradient descent**. Here we show that **none of these claims hold true in the general case.** Through a combination of analytical results and simulation, we demonstrate that the information plane trajectory is predominantly a function of the neural nonlinearity employed: **double-sided saturating nonlinearities like tanh** yield a compression phase as neural activations enter the saturation regime, but linear activation functions and single-sided saturating nonlinearities like the widely used **ReLU in fact do not**. Moreover, we find that there is no evident causal connection between compression and generalization: networks that do not compress are still capable of generalization, and vice versa. Next, we show that the compression phase, when it exists, does not arise from stochasticity in training by demonstrating that we can replicate the IB findings using full batch gradient descent rather than stochastic gradient descent. Finally, we show that when an input domain consists of a subset of task-relevant and task-irrelevant information, hidden representations do compress the task-irrelevant information, although the overall information about the input may monotonically increase with training time, and that this compression happens concurrently with the fitting process rather than during a subsequent compression period.

15. Applications of Entropy in **Finance**: A Review, by Rongxi Zhou, Entropy 2013.

    Although the concept of entropy is originated from thermodynamics, its concepts and relevant principles, especially the principles of maximum entropy and minimum cross-entropy, have been extensively applied in finance. In this paper, we review the concepts and principles of entropy, as well as their applications in the field of finance, especially in **portfolio selection** and **asset pricing**. Furthermore, we review the effects of the applications of entropy and compare them with other traditional and new methods.

16. Information Theory in **Scientific Visualization**, by Prof. **Han-Wei Shen**, Entropy 2011.

    In recent years, there is an emerging direction that leverages information theory to solve many challenging problems in **scientific data analysis and visualization**. In this article, we review the key concepts in information theory, discuss how the principles of **information theory can be useful for visualization**, and provide specific examples to draw connections between data communication and data visualization in terms of how information can be measured quantitatively. As the amount of digital data available to us increases at an astounding speed, the goal of this article is to introduce the interested readers to this new direction of data analysis research, and to inspire them to identify new applications and seek solutions using information theory.



(a) message transmission

(b) data visualization