# Principles and Applications of Science of Information

By THOMAS COURTADE, *MEMBER IEEE*
*Guest Editor*
ANANTH GRAMA
*Guest Editor*
MICHAEL W. MAHONEY
*Guest Editor*
TSACHY WEISSMAN, *FELLOW IEEE*
*Guest Editor*

## I. INTRODUCTION

Information is the basic commodity of our era, permeating every facet of our lives. Our current understanding of the formal underpinnings of information dates back to Claude Shannon's seminal work in 1948 resulting in a general mathematical theory for reliable communication in the presence of noise. This theory enabled much of the current day storage and communication infrastructure. Shannon's notion of information quantified the extent to which a recipient of data can reduce its statistical uncertainty. For example, a single yes/no answer to a (suitably selected) question is capable of eliminating half of the space of possible states in a system. On the other hand, the answer to an irrelevant question might not eliminate any of the possible states. In this case, the first answer is considered more "informative" than the second. An important aspect of Shannon's theory is the assumption that the "semantic aspects of communication are irrelevant" (Shannon, 1948).

This special issue contains papers on models and methods in the science of information, along with their applications in diverse domains.

While information theory has profound impact, its application to other abstractions and applications, beyond storage and communication of digital data, poses foundational challenges. Major open questions relating to extraction, comprehension, and manipulation of structural, spatiotemporal, and semantic information in scientific and social domains remain unresolved. We believe that answers to these questions, grounded in formal information-theoretic concepts, hold the potential for tremendous advances. This motivates development of novel concepts integrating formal modeling, methods, analyses, algorithm, and infrastructure development.

New developments in information theory are strongly motivated by and contribute to diverse application domains. For example, generalizing from Shannon's information theory, we can argue that the "information content" of a crystalline structure is less than that of a "polycrystaline" material, which in turn is less than an "amorphous" material. How do we formally quantify this notion of information in structure and what implication does this have for associated material properties (cf. K. Chung "The Nature of Glass Remains Anything but Clear," NYT, July 2008)? Similarly, we know of oft-repeated structures (structural motifs) in biomolecules, which are known to be associated with similar biological functions. Can we quantify the informative content of these repeated structures? Can we then use these metrics to relate structure and function in formal ways?

A large and increasing repository of data relates to interactions between entities ranging from biomolecules (cell signaling, binding, regulation, etc.), biosystems (pluripotency of stem cells, immune response), individuals (social interaction), ecosystems, and beyond. How can we capture the informative content of such networks of interactions? How can we then use this information to perform application-specific analyses? We now have the ability to collect signaling pathways of large numbers of healthy and diseased (e.g., diabetes, cancer) cells. Can we examine these pathways and precisely identify parts of the network that inform us of specific

Digital Object Identifier: 10.1109/JPROC.2016.2646778

diseases and phenotype? Similarly, we have extensive data on social networks, recommender networks, networks of transactions, and computer communications. Developing a formal methodology for analyses based on a unifying theory of information is important.

A critical aspect of many interactions is the timeliness of signals. A 24-h weather forecast that arrives after the fact has no informative content. A signaling event in a living cell, for example, a lytic or lysogenic response in Lambda Phage bacteria to external stress, is of little consequence if the stress results in cell death. Clearly, the notion of timeliness of data is dependent on the system, the data, and its context. How do we incorporate this notion of timeliness into the information content of data? The complexity of this problem is rooted in the fact that any notion of timeliness must be in the context of the eventual use (or function) associated with the data.

In typical scientific applications, information is associated with its consequence, i.e., it is a property of the recipient's ability to understand and act on it. For example, an intracellular signaling mechanism works only through appropriate (transmembrane) signal receptors. A message in a communication system is informative only if the recipient can decode it. This emphasizes the notion of semantics associated with information. Indeed, virtually all systems are currently modeled in a constrained environment where all parties are assumed to have the ability to "understand" messages through a common language. In a number of important applications, for example, information obfuscation, deciphering informative aspects of financial transactions, statistically significant (dominant or deviant) patterns of social interactions, etc., underlying semantics play important roles. How do we quantify the semantics of information? How do we infer the semantic content? How do we identify statistically significant semantic artefacts in large data sets?

Connections between information and energy have piqued intellectual curiosity for over a century. Correspondence between physical entropy (Boltzman/Gibbs/Maxwell, circa 1870), quantum mechanical entropy (von Neumann, circa 1927), and information entropy (Shannon/Hartley, circa 1945) have been topics of several studies. Researchers in other disciplines have proposed related measures of entropy to characterize, in specific contexts, underlying information content. Examples of these measures include psychological entropy, social entropy, and economic entropy. The original thought experiments of Maxwell (Maxwell's deamon) and Szilard quantify the energy content of a single bit of information ($k_B T \ln 2$). Can we use these approaches to quantifying information in different contexts to explore deep problems in quantum confinement/entanglement, generalizing to quantum information theory?

Various aspects of information represented in structure, space, time, connectivity, and semantics have been explored in different scientific disciplines in somewhat *ad hoc* ways. For example, a critical tool in analysis of genomic (or proteomic) sequences is the notion of alignment. The underlying hypothesis is that conserved (aligned) subsequences are "informative" with respect to their structure and function. In molecular biology, structural motifs inform us of the associated function of the molecule. In social networks, repeated patterns of interaction are studied as canonical mechanisms of information flow. While methods for extracting these have been discovered (or rediscovered) in different domains, a comprehensive theory of information for complex interacting systems remains elusive. We have yet to answer (or even suitably formulate) fundamental questions such as: How do we quantify, represent, and extract information in commonly used abstractions for diverse systems? How is information created and in what ways can it be transferred? What is the value of information as represented in various abstractions? What are fundamental bounds on extraction of information from large data repositories? The mismatch between our ability to describe and design complex systems and our ability to understand them is rooted in the notion of abstraction. While we can design systems with trillions of components and more (e.g., recent Petascale to Exascale Computing platforms) through suitable hierarchical abstractions and design paradigms, we cannot even model, comprehensively, the function of a single living cell. This gap in understanding, in our opinion, is a fundamental impediment to continued scientific discovery.

These complex questions motivate a new science of information. This issue highlights some of the recent advances in the science of information.

## II. EMERGING DIRECTIONS AND OUTSTANDING CHALLENGES

Advances in information technology and widespread availability of information systems and services have largely obscured the fact that information remains undefined in its generality, though considerable collective effort has been invested into its understanding. Shannon wrote in 1953: "The word "information" has been given many different meanings … it is likely that at least a number of these will prove sufficiently useful in certain applications and deserve further study and permanent recognition." As we have discussed, our intuitive understanding of information cannot be formalized without taking into account the structure, context, timeliness, the objective its recipient wants to achieve, and knowledge of the recipient's internal rules of conduct or its protocol. One may also argue, similarly, that information extraction depends on available resources. It follows, therefore, that in its generality, information is that which can impact a recipient's ability to achieve the objective of some activity in a given context using limited available resources.

It can be argued that many of the advances in storage and communication may have been accomplished without grounding in Shannon's information theory. However, what Shannon's theory provided was fundamental bounds on such quantities as information content, compression, transmission, error resilience and recovery, etc. These bounds were essential for subsequent algorithms that form the core of our information infrastructure. In a similar fashion, we aim to establish fundamental bounds on information content, extraction, tolerance to error (or lack of data), etc., use these bounds to develop models and methods with formally quantifiable performance, and to demonstrate their use in important and diverse scientific applications.

- Structure and organization: We lack measures and meters to define and quantify information embodied in structure and organization (e.g., information in nanostructures, biomolecules, gene regulatory and protein interaction networks, social networks, networks of financial transactions, etc.). Ideally, these measures must account for associated context, and incorporate diverse (physical, social, economic) dynamic observables and system state.
- Delay: In typical interacting systems, timeliness of signals is essential to function. Often timely delivery of partial information carries higher premium than delayed delivery of complete information. The notion of timeliness, however, is closely related to the semantics (is the signal critical?), system state (is the system under stress?), and the receiver.
- Space: In interacting systems, spatial localization often limits information exchange, with obvious disadvantages as well as benefits. These benefits typically result from reduction in interference as well as ability of system to modulate and react to stimulus.
- Information and control: In addition to delay-bandwidth tradeoffs, systems often allow modifications to underlying design patterns (e.g., network topology, power distribution and routing in networks). Simply stated, information is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute basic objectives.
- Semantics: In many scientific contexts, one is interested in signals, without knowing precisely what these signals represent but little more than that can be assumed *a priori*. Is there a general way to account for the "meaning" of signals in a given context? How to quantify the information content of natural language?
- Dynamic information: In a complex network, information is not just communicated but also processed and even generated along the way (e.g., response to stimuli is processed at various stages, with immediate response processed at site of stimulus, higher level response processed in the brain; response to emergency events is coordinated at various levels, ranging from first responders to command and control centers). How can such considerations of dynamic sources be incorporated into an information-theoretic model?
- Learnable information: Data-driven science (as opposed to hypothesis-driven modeling) focused on extracting information from data has received considerable recent research attention. How much information can actually be extracted from a given data repository? Is there a general theory that provides natural model classes for data at hand? What is the cost of learning the model, and how does it compare with the cost of actually describing the data? Is there a principled way to approach the problem of extracting relevant information?
- Limited resources: In many scenarios, information is limited by available resources (e.g., computing devices, bandwidth of signaling channels). How much information can be extracted and processed with limited resources? This relates to complexity and information, where different representations of the same distribution may vary dramatically when complexity is taken into account.
- Cooperation: How does cooperation impact information? Often subsystems may be in conflict (e.g., the problem of Byzantine generals, denial of service, or selfish attacks in computer systems) or in collusion (e.g., price fixing, insider trades). How one can quantify and identify information transfer in such systems?

These questions pose fundamental intellectual challenges, which have significant potential for impacting diverse application domains.

## III. OVERVIEW OF ISSUE

This issue contains papers on models and methods in science of information, along with their applications in diverse domains.

In "High-probability guarantees in repeated games: Theory and applications in information theory," Delgosha *et al.* fuse ideas from both information theory and game theory to study repeated games with incomplete information in a "high-probability framework." Under this framework, users attempt to guarantee a certain payoff with high probability, in contrast to traditional game theory where players aim to maximize the expected payoff. Examples of this framework naturally arise in information theory, where a

transmitter aims to communicate a message to a receiver with high probability through repeated channel uses. Analysis under this framework requires the development of several novel techniques, which may be applicable in information and game theory, more broadly.

In "Information-theoretic approach to strategic communication as a hierarchical game," Akyol *et al.* study information disclosure problems of economics through an information-theoretic lens. Such problems differ from those in communications, involving different objectives for the encoder and the decoder, which are aware of mismatch and act accordingly. A hierarchical communication game is considered, where the transmitter announces an encoding strategy with full commitment, and its distortion measure depends on a private information sequence whose realization is available at the transmitter. The receiver employs a decoding strategy minimizing its own distortion based on the announced encoding map and the statistics. Equilibrium strategies and their associated costs are characterized for various canonical scenarios. The results are also extended to the broader context of decentralized stochastic control.

In "Dynamic watermarking: Active defense of networked cyber–physical systems," Satchidanandan and Kumar investigate the problem of secure control of networked cyber–physical systems. In particular, the authors introduce a general watermarking technique, whereby malicious tampering with signal can be discovered under various criteria. This is a particularly timely and interesting topic in light of recent events, which have shown that critical infrastructure systems, such as energy, healthcare, and transportation, are increasingly vulnerable to cyber attacks.

In "Decision making with quantized priors leads to discrimination," Varshney *et al.* introduce an information-based model of signal detection motivated by the question of racial discrimination in decision-making scenarios such as police arrests. The model incorporates the likelihood ratio test for maximizing expected utility but constrains the threshold to a small discrete set. This precision constraint is argued to follow from both bounded rationality in human recollection and finite training data for estimating priors. When combined with social aspects of human decision making and precautionary cost settings, the model predicts the own-race bias that has been widely observed in econometrics.

In "An optimization approach to locally-biased graph algorithms," Fountoulakis *et al.* investigate a class of locally-biased graph algorithms for finding local or small-scale structures in large graphs. The focus of the paper is on algorithms that compute relevant answers, while only looking at a fraction of the graph; in this sense, they are sublinear in their runtime. The paper discusses various issues relating to algorithmic complexity and statistical considerations, and highlights common threads across different approaches that have been proposed in literature, along with important applications and avenues for future research.

In "An information and control framework for optimizing user-compliant human–computer interfaces," Tantiongloc *et al.* present a framework for a human–computer interface. In this model, a human's knowledge is represented as a point in Euclidean space, the intention of the human is signaled to the computer over a noisy channel, and the computer queries the human in a manner that is amenable to human operation. This model is used to characterize a class of systems that are information theoretically optimal in the sense that they enable computers to infer human intent rapidly. The proposed framework provides a simplified method based on optimal transport theory to generate optimal feedback signals between the computer and human in high dimension, while still preserving communication optimality. The framework also lends itself to the integration of a human user by attempting to moderate the difficulty of the task presented to the user.

In "A study of the Boltzmann sequence-structure channel," Magner *et al.* present a channel that maps sequences (e.g., a sequence of amino acids) from a finite alphabet to self-avoiding walks in a 2-D grid. The channel, which is inspired by a model of protein folding, is called the Boltzmann sequence-structure channel. It is characterized by a Boltzmann/Gibbs distribution with a free parameter corresponding to temperature under which folding occurs. The authors estimate the conditional entropy between the input sequence and the output fold, giving an upper bound that exhibits a phase transition with respect to temperature. They then formulate a class of parameter settings under which the dependence between random walk energies is governed by their number of shared contacts. This setting is used to derive a lower bound on the conditional entropy. Using these bounds, the paper concludes that the mutual information tends to zero in a nontrivial regime of high temperature.

In "Fundamentals of molecular information and communication science," Akan *et al.* consider molecular communication (MC) as a communication paradigm for nanonetwork realization. Since MC significantly differs from classical communication systems, this suggests reinvestigation of information and communication theory fundamentals. The authors review the existing literature on intrabody nanonetworks, and they highlight future research directions and open issues that need to be addressed for revealing the fundamental limits of molecular information science, both in general and in the specific case of molecular information science in life sciences.

In "Doubly penalized LASSO for reconstruction of biological networks," Asadi *et al.* consider reconstruction of biological and biochemical networks. This is a crucial step in extracting knowledge and causal information from large biological data sets, and this task is particularly challenging when dealing with time-series data from dynamic networks. The authors present a new method for

the reconstruction of dynamic biological networks, and they present two case studies to compare the relative performance of their method with previous methods, illustrating that their method does particularly well for networks with low and moderate density, a common situation in many applications.

In "Addressing the need for a model selection framework in systems biology using information theory," DeVilbiss and Ramkrishna note that models are used in systems biology to organize biological knowledge and make predictions of complex processes that are hard to measure, and they consider the generation and evaluation of such models. Attempting to go beyond subjective arguments to choose one model framework over another, they develop the argument that information-theoretic model selection metrics should be extended to nonnested model comparison applications in systems biology. They also make a novel comparison of kinetic, constraint-based, and cybernetic models of metabolism based not only on model accuracy, but also on model complexity.

In "A critical survey of deconvolution methods for separating cell types in complex tissues," Mohammadi *et al.* focus on *in silico* deconvolution of signals associated with complex tissues into their constitutive cell-type components. The authors survey a variety of models, methods, and assumptions underlying deconvolution techniques. They investigate the choice of the different loss functions for evaluating estimation error, constraints on solutions, preprocessing and data filtering, feature selection, and regularization to enhance the quality of solutions, along with the impact of these choices on the performance of commonly used regression-based methods for deconvolution. Various combinations of these elements are studied: some that have been proposed in the literature, and others that represent novel algorithmic choices for deconvolution. Shortcomings of current methods are identified, along with new approaches to their remediation. The findings are effectively summarized in a prescriptive step-by-step format, applicable to a wide range of deconvolution problems.

In "An information-theoretic view of EEG sensing," Grover and Venkatesh explore potential advantages of high-density EEG systems for high-resolution imaging of the brain. In particular, they challenge the application of the conventional Nyquist sampling paradigm to this problem, and argue that high-density sensing systems are necessary for accurate EEG scalp signal recovery, as well as for estimation of spatiotemporal dynamics of activity inside the brain. The authors also propose a hierarchical sensing technique and analyze its performance.

In "The information content of glutamine-rich sequences define protein functional characteristics," Sen *et al.* investigate the relation of abnormally expanded glutamine *(Q)* repeats within specific proteins and their function. The paper notes that regions with low complexity (low information content) can display exquisite functional specificity. Using detailed statistical analysis, the authors identify *Q*-rich *(QR)* regions in coils of yeast transcription factors and endocytic proteins. This analysis shows that when the non-*Q* amino acids from an endocytic protein were exchanged by the ones enriched in *QR* from transcription factors, the resulting protein was unable to localize to the plasma membrane and was instead found in the nucleus. These results indicate that while *QR* repeats can efficiently engage in binding, non-*Q* amino acids provide essential specificity information. The paper hypothesizes that coupling low complexity regions with information-intensive determinants might be a strategy used in many protein systems involved in different biological processes. ■

## ACKNOWLEDGMENT

The editors would like to acknowledge the National Science Foundation Center for Science of Information (http:\\www.soihub.org) for providing the intellectual environment and resources to conduct research in the science of information. The editors would also like to thank various reviewers, as well as the editorial staff at IEEE for their superlative effort.

## ABOUT THE AUTHORS

**Thomas Courtade** (Member, IEEE) received the B.Sc. degree (*summa cum laude*) in electrical engineering from Michigan Technological University, Houghton, MI, USA, in 2007 and the M.S. and Ph.D. degrees from the University of California Los Angeles (UCLA), Los Angeles, CA, USA, in 2008 and 2012, respectively.

He is an Assistant Professor with the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, USA. Prior to joining UC Berkeley in 2014, he was a Postdoctoral Fellow supported by the National Science Foundation (NSF) Center for Science of Information.

Prof. Courtade received a Distinguished Ph.D. Dissertation Award and an Excellence in Teaching Award from the UCLA Department of Electrical Engineering, and a Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory. He also received a Hellman Fellowship in 2016.

**Ananth Grama** received the B.Engg. degree in computer science from the Indian Institute of Technology, Roorkee, India, in 1989, the M.S. degree in computer engineering from the Wayne State University, Detroit, MI, USA, in 1990, and the Ph.D. degree in computer science from the University of Minnesota, Minneapolis, MN, USA, in 1996.

He is a Professor of Computer Science at Purdue University, West Lafayette, IN, USA and the Associate Director of the Center for Science of Information, a Science and Technology Center of the National

Science Foundation (NSF). He was Director of the Computational Science and Engineering and Computational Life Sciences programs at Purdue (2012–2016) and Chaired the Biodata Management and Analysis Study Section of the National Institutes of Health (2012–2014). His research interests span the areas of parallel and distributed computing algorithms and applications, including modeling, simulation, and control of diverse scientific processes and phenomena.

Prof. Grama is a recipient of the National Science Foundation CAREER Award (1998), and the University Faculty Scholar Award at Purdue (2002–2007). He is a Fellow of the American Association for the Advancement of Sciences (2013) and Distinguished Alumnus of the University of Minnesota (2015).

**Michael W. Mahoney** received the Ph.D. degree from Yale University, New Haven, CT, USA, in 2000, with a dissertation in computational statistical mechanics.

He is in the Department of Statistics and at the International Computer Science Institute (ICSI), University of California at Berkeley, Berkeley, CA, USA. He works on algorithmic and statistical aspects of modern large-scale data analysis. Much of his recent research has focused on large-scale machine learning, including randomized matrix algorithms and randomized numerical linear algebra; geometric network analysis tools for structure extraction in large informatics graphs; scalable implicit regularization methods; and applications in genetics, astronomy, medical imaging, social network analysis, and internet data analysis. He has worked and taught in the Mathematics Department, Yale University; at Yahoo Research, Sunnyvale, CA, USA; and in the Mathematics Department, Stanford University, Stanford, CA, USA.

Prof. Mahoney is on the National Advisory Committee of the Statistical and Applied Mathematical Sciences Institute (SAMSI), was on the National Research Council's Committee on the Analysis of Massive Data, coorganized the Simons Institute's Fall 2013 program on the Theoretical Foundations of Big Data Analysis, and runs the biennial MMDS Workshops on Algorithms for Modern Massive Data Sets.

**Tsachy Weissman** (Fellow, IEEE) received the B.Sc. degree in electrical engineering (*summa cum laude*) and the Ph.D. degree from the Technion—Israel Institute of Technology, Haifa, Israel, in 1997 and 2001, respectively.

He then worked at Hewlett Packard Laboratories with the information theory group until 2003, when he joined Stanford University, Stanford, CA, USA, where he is currently Professor of Electrical Engineering and incumbent of the STMicroelectronics Chair in the School of Engineering. He has spent leaves at the Technion, and at ETH Zurich. His research is focused on information theory, compression, communication, statistical signal processing, the interplay between them, and their applications.

Prof. Weissman is recipient of several best paper awards, and prizes for excellence in research and teaching. He served on the editorial board of the IEEE Transactions on Information Theory from September 2010 to August 2013, and currently serves on the editorial board of *Foundations and Trends in Communications and Information Theory*. He is Founding Director of the Stanford Compression Forum.