Frontiers of Science of Information

Wojciech Szpankowski [1,]

[1] Department of Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A.
E-Mails: szpan@purdue.edu

In today's knowledge-based society, the central role of data, data-driven insights, and their use in scientific, commercial, and social enterprise is widely recognized. *Information* is acquired, stored, communicated, curated, organized, aggregated, analyzed, valued, secured, and used to understand, optimize, and control complex processes. In the context of real systems, these tasks pose significant challenges stemming from incomplete or noisy data, varying levels of abstraction and heterogeneity, requirements on scalability, constraints on resources, considerations of privacy and security, and real-time performance. These problems lie at the core of a large class of applications, ranging from analysis of biochemical processes in living cells to building robust wide-area communication systems, and understanding global-scale economic and social systems. Claude Shannon laid the foundation of information theory, demonstrating that problems of data transmission and compression can be precisely modeled, formulated, and analyzed. He also provided basic mathematical tools for addressing these problems. Motivated by Shannon's focus on the fundamental, and his precise quantitative analysis, the *Center for Science of Information* aims to develop rigorous principles guiding all aspects of information, integrating elements of *space*, *time*, *structure*, *semantics*, *cooperation*, and *value*, in different application contexts. Led by Purdue University, Center member institutions include Bryn Mawr, Howard, MIT, Princeton University, Stanford University, Texas A&M University, University of California at Berkeley and San Diego, University of Hawaii at Manoa, and University of Illinois at Urbana-Champaign. Center mission is to advance science and technology through a new paradigm in the quantitative understanding of the representation, communication, and processing of information.

The research theme of the Center focuses on transforming *data* to *information* to *knowledge*. It targets fundamental principles, formal methodologies and frameworks, and a rich set of algorithms and computational tools that are validated on applications in life sciences, communication systems, and economics. The Center targets the following fundamental problems: (i) modeling complex systems and developing analytical methods for quantifying information representation and flow in such systems; (ii) methods for quantification and extraction of informative substructures; (iii) understanding cooperation, competition, and adversaries in complex systems; (iv) developing information-theoretic models for managing, querying, and analyzing data with real-world constraints of incompleteness, noise, distribution, and resource constraints.

Following the principles of Shannon and Turing - who engaged themselves with practical systems before arriving at their theoretical abstractions, the Center focuses on specific applications with the goal of obtaining a broader and more general understanding of information. For instance, timeliness of information is extremely important when information is used in cyber-physical systems – leading one to investigate the issue of delay, which Shannon's theory largely ignored. The use of information also brings to focus the issue of semantics -- the meaning of the message is integral to performance of the consequent task, leading one to investigate goal-oriented communication and control under constrained communication. These problems are helping define information-semantics in fundamentally new and relevant ways by the Center scientists. Cyber-physical systems bring together processing and communication of information which is explored in the context of various applications ranging from vehicle information systems to (human) body sensor networks.

Investigation of biological systems at the Center motivates understanding of representation, inference, and aggregation of data. Since the time of Shannon, biology has undergone a major revolution, giving rise to significant challenges in interpreting data and understanding biological mechanisms. From a historical perspective, Henry Quastler first introduced information theory in biology in 1949, just a year after the landmark paper of Shannon, and four years before the inception of molecular biology (shaped by the work of Crick and Watson). Continuing this effort, Quastler organized two symposia on ``Information Theory in Biology''. These attempts were rather unsuccessful, as argued by Henry Linschitz who pointed out that there are ``difficulties in defining information of a system composed of functionally interdependent units and channel information (entropy) to ``produce a functioning cell''. The advent of high-throughput technologies for data collection from living systems, coupled with our refined understanding of biological processes provides new impetus for efforts aimed at understanding how biological systems (from biomolecules to tissues) represent and communicate information. How can one infer this information optimally (from genome sequencing to functional image analysis)? How can one control specific functional and structural aspects of processes based on this understanding?

In yet other applications such as economics, questions of how information is valued are important. Flow of Information in economic systems and associated control problems are vitally important, and have been recognized through recent Nobel Prizes. More recently, with the ability to collect large amounts of data from diverse systems such as sensed environments and business processes, problems in `big and small data' have gained importance. Data analytics at scale is critically reliant on models and methods whose performance can be quantified. Issues of privacy and security pose problems for data management, obfuscation, querying, and secure computations. The Center is at the cutting edge of research in knowledge extraction from massive and structural data.