Frontiers of Science of Information: Shannon Meets Turing

Wojciech Szpankowski and Ananth Grama, Center for Science of Information, Purdue University

The authors' vision for a Science of Information integrates key elements of Shannon and Turing postulates into an overarching framework for data analytics, privacy, and security that provides the critically needed formal basis for data science.

Information and computation are two of the defining concepts in the modern era of computing. Claude Shannon laid the foundation of information theory, demonstrating that the problems of communication and compression can be precisely modeled, formulated, and analyzed. Alan Turing, on the other hand, formalized the concept of computation, broadly defined as the transformation of information through algorithms.

We believe that a Science of Information that integrates key elements of both Shannon and Turing postulates can address important challenges in transforming data to information to knowledge through information-efficient computation. Furthermore, a formal quantification of information should precede the development of methods in information-efficient computing. In this article, we focus on fundamental challenges in the triad from data to information to knowledge, addressing questions relating to the limits of learning, adequacy of sampling, incremental value of data, and optimality of learning methods, among others. To do this, Science of Information leverages techniques from information theory, statistics, algorithms and analysis, combinatorics, numerical methods, and learning theory.

There have been significant advances in broad areas of machine and computational learning, information theory, privacy and security, and data mining and analytics. Methodological advances have translated to significant new applications that have reshaped daily life, from autonomous vehicles to learning thermostats. At the same time, these advances have motivated important new technical challenges: What are the limits of learning from data? When is a system undersampled? When are learned models overfit? What is the incremental benefit of additional data? How do we validate learned models? How do we derive provably optimal learning techniques?

0018-9162/18/$33.00 © 2018 IEEE

THE CENTER FOR SCIENCE OF INFORMATION

In 2010, the NSF established the Center for Science of Information (www.soihub.org) as one of its Science and Technology Centers. The Center for Science of Information seeks to develop the fundamental principles underlying various aspects of information, along with their applications to diverse scientific, engineering, social, and economic domains. Its mission is to advance science and technology through new paradigms in the quantitative understanding of the representation, communication, and processing of information in biological, physical, social, and engineering systems. It aims to use tools of information theory, computer science, statistics, mathematics, and physics as a basis for extending the scope of information sciences.

There is a pressing need for formal modeling and analyses for answering these questions. Even as we address these problems, there are increasing concerns relating to data privacy, particularly regarding data analytics and querying. There are clear tradeoffs of accuracy (in querying and analysis) and disclosure, which must be analytically quantified and studied. Science of Information provides a unifying framework within which these problems can be formulated and studied in a principled and unified manner.

BACKGROUND

Science of Information and its relation to diverse problems in data science, life sciences, communication, and economics—as well as its applications in various domains—is an important emerging area of research (for more information, see the sidebar). The domain of interest in typical data-analytics applications can be modeled as a system or a process, and in this context, data is defined as the sensed or sampled values from the domain. The term "information" has been extensively debated in literature over several centuries: we define it as that which can distinguish one domain from another. Informally stated, this definition suggests that a very sparse sampling of a domain of interest does not allow us to distinguish the domain from others—consequently, it contains little information. Conversely, oversampling the same domain does not add information because it does not add to our ability to distinguish the domain from others. We use this notion of distinguishability, and the associated partitioning of the domain space, to characterize the information associated with data. The resulting formalisms, the associated analyses and methods, and their specialization to different application contexts correspond with the Science of Information.

We define "knowledge" (or meaning/insight) as a combination of information and semantics associated with a domain. The notion of semantics associated with a domain enables one to contextualize information. We define the process of abstracting data into information, and subsequently to knowledge, as "learning." When such processes are carried out by an automaton, we refer to this process as "machine learning." The related term "data science" refers to the complete pipeline associated with transforming raw data into knowledge—including such tasks as preprocessing, cleaning, analysis, and interpretation. Statisticians might view this "knowledge" as a hypothesis, a structure function, or a model (or set thereof). In this context, notions of minimum description length (MDL) and Bayesian information criterion (BIC) are commonly used for selecting appropriate models for data in conjunction with suitably formulated optimization procedures.[1] Computer scientists often view these abstractions in terms of Kolmogorov complexity (or descriptive complexity), which is the shortest program that can output the observed data.[1] There are other approaches advocated by the computational learning theory community, such as exact learning,[2] probably approximately correct learning (PAC learning),[3] VC theory,[4] and Bayesian inference.

While Shannon's focus was on data recovery in compression (source coding) and communication (channel coding), the Science of Information generalizes this framework to a broad class of analyses, because information is not merely communicated but is also acquired, represented, inferred, processed, aggregated, managed, valued, secured, and computed. In this context, Shannon's converse theorem is particularly relevant. Following Shannon's classical definition of "information," the converse theorem states that if fundamental limits on information in storage or communication are exceeded, then no algorithm can accomplish desired tasks (recovery in transmission or storage) with vanishing probability of error. In the contrapositive form, this theorem states that if we can devise an algorithm to carry out the desired task, then fundamental limits must be satisfied.

However, the converse theorem does not state anything about the computational hardness of the desired task. When generalizing beyond storage and communication, this computational hardness is a particularly important consideration. In this article, we propose information-efficient formulations, which simultaneously satisfy fundamental limits of information while being computationally tractable.

It is well known that the problem of offline optimal Lempel-Ziv (LZ) coding is NP-hard. However, millions of users rely on LZ compression daily, using linear time algorithms on typical data sequences to achieve asymptotic optimality. Likewise, the DNA assembly problem is known to be NP-hard (through the Hamiltonian path problem). However, efficient methods that achieve the information limit have been demonstrated on real DNA assembly problems.[5] More generally, because real data is almost inevitably corrupted, noisy, or incomplete, focusing on the computational hardness of deriving optimal solutions is not always useful in a data-science context. In such cases, we argue in favor of tractable information-efficient algorithms that achieve information limits. However, it should be noted that one cannot always come up with an efficient method that achieves these limits (for example, there is an information–computation gap in the DNA assembly problem[5] and in the community recovery problem of stochastic block models). In this sense, an analysis of information limits must precede efforts aimed at method development (computation) for transforming data to information through algorithms. This motivates our dictum "Shannon meets Turing."

Science of Information incorporates elements of structure, space, time, connectivity, and semantics into Shannon's theory of information. For example, a critical tool in the 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. Although methods for extracting these have been discovered (or re-discovered) in different domains, a Science of Information for complex interacting systems must answer such fundamental questions 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 the fundamental bounds on extracting information from large data repositories?

SCIENCE OF INFORMATION

Advances in information technology and the 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. In 1953, Shannon wrote: "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."[6] Prior to this in 1948, he wrote: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem."[7]

We adopt a definition of information as that which can be used to distinguish one set of data samples from another. This process of distinguishing involves an "observer function" that maps a dataset to an observed object. The observer function in traditional information theory might correspond to a channel. In a more general setting, the observer function might be an arbitrary function, such as a learning operator, formulated as an optimization procedure. For example, a given dataset can be "observed" through a classifier (such as a logistic regression function). In this case, two datasets d_i and d_j have the same information content if they yield the same classifier relative to the learning method that produces that classifier (see Example 2). Note that we implicitly associate measures of quality (for example, the goodness of the logistic regression fit over the test set), as indicated by the objective function value in this case, into the observer function.

Stated more formally, let S be a space in which each element corresponds to an input dataset. We define an observer function f on S, with respect to which we define information. The goal is to cover "almost" the entire space with a minimum number of balls B_i, such that all elements in a single ball B_i are indistinguishable with regard to the observer function f. Here, "almost" corresponds to a fraction $(1 - \varepsilon)$ of the total space or

the underlying probability space. Let the number of such balls B_i covering $(1 - \varepsilon)$ fraction of S through function f be equal to $N_\varepsilon(f)$. Then, we define the useful/learnable information as $I(f, S) = \log N_\varepsilon(f)$ as $\varepsilon \to 0$.

Example 1: Shannon channel capacity

First we present an example of how the general definition of information can be specialized into a traditional framework of reliable transmission.

Consider a simple point-to-point communication channel, as originally introduced by Shannon.[7] We wish to transmit a binary sequence of length n over a channel. The goal is to ensure that the probability of error P_e (which corresponds to the observer function f) on the receiving side tends to 0, as n increases. Clearly, we cannot reliably transmit all 2^n sequences with a nonzero probability of error in transmission. However, can we reliably transmit 2^{nR} sequences for some $R < 1$, with small probability of error? In other words, can we partition the set of "almost all" sequences into 2^{nR} distinguishable subsets with regard to P_e? The answer is yes for $R < C$, where C is the Shannon capacity, leading to reliably transmitting $I_n(P_e) = nC$ bits of information.

Example 2: Learnable sources

In this example, we demonstrate the use of our formalization of information in the context of distinguishable sources in universal learning.

We focus our attention on computable learnable information contained in a sequence $x^n = x_1 \ldots x_n$ generated by a source belonging to a class of parameterized distributions $\mathcal{M}(\Theta) = \{P_\theta : \theta \in \Theta\}$. Let $\hat{\theta}(x^n)$ be the maximum likelihood (ML) estimator, that is, $\hat{\theta}(x^n) = \arg\max_{\theta \in \Theta} P_\theta(x^n)$. Observe that for a given sequence x^n, produced either by θ or by θ', we can use $\hat{\theta}(x^n)$ to decide which model generates the data with a small error probability, provided these two parameters are sufficiently far apart by some distance measure. If the two models θ and θ' are too close to each other, they are virtually indistinguishable and do not introduce any additional useful information. Thus, learnable information about x^n is summarized in the number of distinguishable distributions (models).

As an example, we can estimate learnable information $I_n(\Theta)$ when the class of sources $\mathcal{M}$ is a set of memoryless distributions with $m - 1$ unknown parameters (in other words, over an alphabet of size m). As discussed earlier, we will use the MDL estimator.[1]

For a distance between distributions/models we adopt the Kullback–Leibler (KL) divergence $D(\cdot||\cdot)$. Let $B_{KL}(\theta_0, \varepsilon) = \{\theta : D(\theta \,||\, \theta_0) \le \varepsilon\}$ be the KL-ball of radius ε around θ_0.

The distinguishability of models depends on the error probability, which can be estimated as follows for some $\theta \in \Theta$:

$$P_\theta(\hat{\theta} \neq \theta) \approx P_\theta(\theta(X) \notin B_{KL}(\theta, \varepsilon / n)) \sim 1 - O(\varepsilon^{m/2})$$

for $\varepsilon > 0$. Setting the error probability at $O(1/\sqrt{n})$, we observed in a previous work[8] that the number of distinguishable distributions $I_n(\Theta)$ is

$$I_n(\Theta) = \sum_{k_1 + \cdots + k_m = n} \binom{n}{k_1, \ldots, k_m} \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m}.$$

The above sum is a convolution of m terms of the form $k^k/k!$, hence tools of analytic combinatorics can be used to find its asymptotics. Applying the Mellin transform and singularity analysis of the Lambert-W function, we find that

$$I_n(\Theta) = \log C_n(\Theta) = \frac{m-1}{2}\log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma\left(\frac{m}{2}\right)}\right) + \frac{\Gamma\left(\frac{m}{2}\right)m}{3\Gamma\left(\frac{m}{2} - \frac{1}{2}\right)} \cdot \frac{\sqrt{2}}{\sqrt{n}} + O(1/n),$$

where Γ is the Euler gamma function.[8] This is the learnable information that can be extracted from a class of memoryless sources. It also coincides with the minimax redundancy/regret of universal source coding.[1]

Factors Influencing Information

We now highlight some common factors associated with observer functions that significantly enhance distinguishability and therefore the ability of the function to extract information. These aspects of information are now being incorporated into a broader Science of Information.

Structure and organization. Often, data has explicit or latent structure and organization associated with it. We lack measures and meters to define and quantify information embodied in structure and organization.[9] Frederick P. Brooks Jr. wrote:[10]

> *Shannon and Weaver performed an inestimable service by giving us a definition of information and a metric for information as communicated from place to place. We have no theory however that gives us a metric for the information embodied in structure ... this is the most fundamental gap in the*

theoretical underpinning of information and computer science. A young information theory scholar willing to spend years on a deeply fundamental problem need look no further.

Examples of structure and organization in data include information coded in nanostructures, biomolecules, gene regulatory and protein interaction networks, and social networks. Ideally, these measures must account for associated context, diverse observations (for example, the same system can be measured in terms of its connectivity, geometry, and other attributes), and time-varying state.

Delay. In typical interacting systems, signals' timeliness is essential to their function. Often, timely delivery of incomplete data carries a higher premium than delayed delivery of complete data. The notion of timeliness, however, is closely related to the system state (is the system under stress?) and the receiver. Shannon's communication theory places no restrictions on complexity or delay in transmission or reception.

Space. In interacting systems, spatial localization often limits the exchange of data, with obvious disadvantages and benefits. These benefits typically result from reduction in interference as well as the ability of a system to modulate and react to stimuli.

Information and control. In addition to delay/bandwidth tradeoffs, systems often allow modifications to underlying design patterns (such as network topology or power distribution and routing in networks). Simply stated, data is exchanged in space and time for decision making, thus the timeliness of delivery, reliability, and complexity constitute basic objectives.

Dynamic information. In a complex network, data is not just communicated but also processed and even generated along the way. For example, a response to stimuli is processed at various stages—an immediate response is processed at the site of the stimulus, a higher-level response is processed in the brain, and a response to emergency events is coordinated at various levels, from first responders to command-and-control centers. These considerations of dynamic sources must be incorporated into the information framework.

Limited resources. In many scenarios, data is limited by available resources (such as computing devices and bandwidth of signaling channels). How much information can be extracted from data and processed with limited resources? This relates to complexity and information, where different representations of the same distribution might vary dramatically when complexity is taken into account. For example, a hundred-digit number sent to a recipient that can only store one digit at a time might lead to a recovery of the order of magnitude but not the exact number.

Representation-invariant information. An important question relates to whether two different data representations are information equivalent. Imagine two texts, one in English and the other semantically identical in Polish. The two texts have exactly the same information content; however, it is hard to establish this using traditional measures of information.

Information and computation. The theme of "computational information" explores those properties of information that can be feasibly extracted. Infeasibility might arise for a number of different reasons: the desired information might be computationally hard to extract, the information might be distributed geographically and not locally extractable, or the information might be encoded in (quantum) physical ways that prevent full extraction. In contrast to the classical theory of information—where precise quantitative limits can be established in most cases—information is not well understood qualitatively in the computational setting, with exponential gaps between the upper and lower bounds on the amount of feasibly extractable information in typical settings. The concept of information extractable by computationally limited sources is fundamental to cryptography, and manipulations of this notion are crucial to modern applications (such as electronic voting). All of this leads to profound questions such as: How many bits of information does a quantum state contain? How does one quantify misunderstanding? What is time in a distributed setting?

Cooperation. Often subsystems can be in conflict (for example, the problem of Byzantine generals, denial of service, or selfish attacks in computer systems) or in collusion (such as price fixing and insider trading). How one can quantify and identify information transfer in such systems, as it manifests in data sampled from different sources?

Value of information. Economic systems share many common features with complex communication networks: they consist of multiple entities

(agents) with vastly heterogeneous capabilities for acquiring, storing, sharing, and processing information and with differing degrees of authority for acting upon that information. Yet, for all the commonalities they share with communication networks, one crucial feature that distinguishes economic networks is that agents have objectives that extend beyond reliable communication. In an economic system, information has value. One major challenge in economics is to formalize the notion of information value, particularly in dynamic settings involving multiple agents. To date, there is no universally agreed-upon definition of "information value," although several reasonable alternatives have been proposed. However, we can distill one basic underlying concept: the value of information has to do with the change of the information state of one or many agents, where, broadly speaking, the information state encapsulates all payoff-relevant knowledge available to the agent(s). In general, the value of one bit of information acquired about a random variable of interest is the largest difference between expected utilities achievable with and without that additional one bit of information.

To summarize, we put forth a general notion of information that relies on an observer function that extracts information from data. The amount of information in a dataset is determined by its distinguishability, based on observer function output. Defined in this manner, information has a number of highly desirable features: it quantifies information in data, it identifies limits of learning from data, and it characterizes the under- or oversampling of data by quantifying the increase in information through additional data.

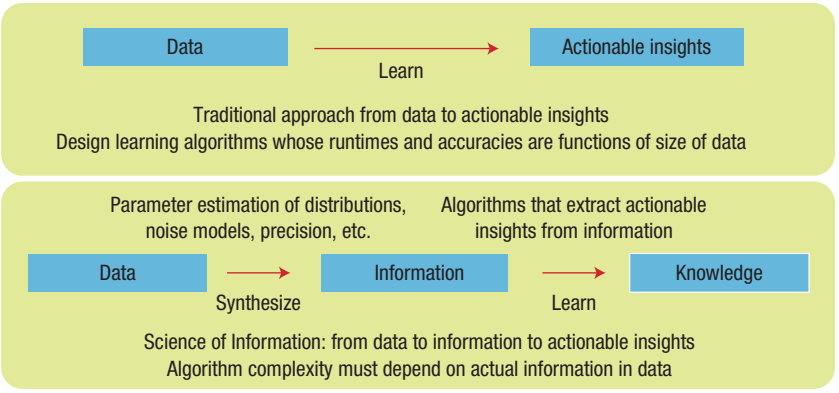

FIGURE 1. The Science of Information approach to extracting actionable insights from data.

INFORMATION AND DATA SCIENCE

Problems in data science are often compounded by issues of extreme scale (big data; typically on distributed platforms); high dimensionality and sparsity patterns that manifest at multiple scales; dynamic, temporal, and heterogeneous structures; complex dependencies; noise and missing data; and semantics. A comprehensive theory for the foundation of data science must pay particular attention to the learnable and distinguishable information contained in data and focus on the design and analysis of algorithms that

- effectively extract all learnable information from the data,
- have computational cost (in terms of both accuracy and runtimes) that depends on useful information in data and not the size of the data, leading to our information-efficient computation paradigm, and
- are robust in a statistical sense (in other words, perform well in the presence of noise and other artifacts).

Our framework for information aims to address precisely these questions. By integrating semantics into our information framework, we can extract actionable insights from data, as shown in Figure 1.

We initiate our discussion with a problem for which we (provably) cannot extract learnable information, irrespective of the algorithms used.

Example 3: Node-arrival order in a graph

Consider a dynamic network in which nodes are added (and perhaps deleted). Each node has a timestamp of its arrival. However, these timestamps might not be available to us (for example, Facebook users' privacy settings or the evolution of a protein interaction network). Given a graph structure without timestamps, our task is to rank all nodes from oldest to youngest. It turns out that there is no good solution to this problem for graphs generated by Erdős–Renyí and preferential attachment models.[11] More precisely, no algorithm can recover a ranking with high probability for these models. Interestingly, one can design algorithms that optimize a likelihood

estimator, but these algorithms fail to yield the ground truth with high probability.

Information and limits of learning

A natural and important question in data-analytics applications is what the limits of learning are and whether a given method is capable of approaching these limits. In this context, it is important to distinguish between learning from data and learning from models/distributions (from which data is sampled). A number of analytics applications operate in the model-free regime; in other words, their focus is on learning from data. In contrast, typical information-theoretic methods focus on sources (distributions)—proving limits in the context of a class of allowable distributions. The related field of statistical learning theory, on the other hand, focuses on estimating a function that maps given inputs and outputs from a collection of training data. In the model-based regime, one is faced with the problem of establishing the limits of learning from a given dataset drawn from a known model. We refer to Example 3 of inferring node-arrival orders in a dynamic network to illustrate this problem and its solution. For the sake of discussion, we assume that dynamic graphs are generated from a preferential attachment model. In this case, a data input corresponds to a snapshot of a dynamic graph, and the observer function corresponds to an arrival sequence and its likelihood of generating the input graph. The learning problem requires us to identify the arrival sequence with maximum likelihood (note that this is only one formulation and that there could be other formulations of the problem). Two input graphs are indistinguishable with respect to the observer function if they yield the same arrival sequence and likelihood. Therefore, to compute the limits of learning, we can characterize the fraction of indistinguishable inputs given an arbitrary observer function output for preferential attachment networks. We observe that in the node-age example, no matter how much computational power we have, we cannot infer the node arrivals with high probability.

SCIENCE OF INFORMATION CAN ADDRESS CHALLENGES IN TRANSFORMING DATA TO INFORMATION TO KNOWLEDGE.

Learning and data

An important question as it relates to learning and information is whether the availability of additional data significantly enhances information and associated learning. In the context of distributions, this relates to the question of undersampling. In our information framework, this question is posed as follows: Does additional input data increase information content and associated learning? If additional data increases information content, we argue that the model is undersampled. Conversely, if the addition of data does not increase distinguishability, the distribution is fully sampled. In general, additional data increases learnable information if it refines the cover of data space, as discussed earlier.

Information and method optimality

There are a number of important consequences of our information and learning framework for developing methods. Although many learning techniques are posed as optimization problems (with regard to the observer function), they do not quantify the significance of the learning outcome. For example, in our node-arrival order problem, it is possible to analytically show that there are a large number of arrival sequences that are all optimally equiprobable. This indicates that although it is possible to (relatively easily) derive an optimal arrival sequence, there are a number of other sequences that are also equally likely. Therefore, even an optimal solution to this problem is not significant. To this end, the distinguishability of inputs with respect to the optimal value also provides insights into the significance of the solution.

Noise, lossy characterizations, and missing data

Rate-distortion theory and lossy compression are important concepts in information theory. Notions of overfitting and sampling are core concepts in statistical modeling. Loss functions and optimization are core concepts in machine learning. These closely related concepts form a thread from

data to information to statistical models to learned models, yet there exists no theory that ties these concepts together in a unified framework. Similarly, concepts of missing data and sampling are closely related, yet their implications for learning are not well understood. What does the existence of a datapoint signify? When working with real-world data samples, we only receive information about those samples that actually exist (we cannot definitively say that two proteins do not interact through an experiment; likewise, we only receive information about movies users choose to rate). We need an information-theoretic characterization of the information content related to the existence of particular datapoints, which helps guide future experiments that directly measure new datapoints. The first fundamental challenge here, which is often ignored, is to assess whether there is any learnable information that can be extracted (particularly in the context of highly incomplete databases such as Netflix).

Data-dependent consistency models

To apply rigorous information-theory techniques (such as minimax and pointwise vs. uniform convergence)[1] to data science, the first question we face is related to model selection. There is often tension between the need for rich model classes to better represent data and our ability to handle these classes from a mathematical point of view. Many applications, particularly in the big data regime, force us to consider model classes that are too complex to admit estimators with traditional model-agnostic guarantees. These traditional estimators are uniformly consistent; they converge uniformly over the entire model class. In comparison, pointwise convergent estimators are often difficult to use predictively as their convergence cannot be verified. Narayana Santhanam and his colleagues departed from this dichotomy and put forth a new analysis framework by characterizing rich model classes that only admit pointwise guarantees, yet all information about the unknown model needed to gauge estimator accuracy can be inferred from the sample at hand.[12]

More formally, in the context of source coding, the problem with existing weak (or pointwise) convergence formulations is that we know a measure is a good universal sequential encoding of the unknown distribution p for sufficiently long sequences. However, the notion of "sufficiently long" depends on the unknown p, since the convergence to the limit might not be uniform in the weak convergence formulation. Santhanam's results show how this can be clarified using the data itself by introducing the notion of data-driven weak convergence.[12] Broadly speaking, it aims to find a universal estimator/encoding with a given accuracy that allows us to find an indicator that becomes small for a predetermined sequence length. To characterize such classes of distributions, Santhanam and his colleagues introduced deceptive distributions.[12] At a high level, distribution $p \in \mathcal{P}$ is deceptive if the strong redundancy of neighborhoods of p is bounded away from 0 in the limit as the neighborhood shrinks to 0. It has been shown that a class of distributions is data driven if no $p \in \mathcal{P}$ is deceptive.

SECURITY AND PRIVACY

As we build complex systems whose operation crucially relies on information extracted from data, there is increasing focus on ensuring privacy and integrity of the methods used for processing data. We can leverage the Science of Information's foundational tools in the study of other aspects of information processing to provide a fundamental and holistic view of security and privacy. Indeed, we view security and privacy as two inescapable aspects of the pipeline that transforms data into information. Here, we describe our formulation and findings in the context of privacy and refer readers to Jonathan Ponniah and his colleagues' study,[13] which investigates novel approaches to security. Guided by our unifying framework for information via distinguishability, we propose a robust architecture for extracting statistical information from databases, while providing provable guarantees on privacy. This holistic viewpoint provides us with a sound theoretical measure of the statistical information preserved by algorithms manipulating databases, and an architecture impermeable to privacy breaches, with provable guarantees.

The problem of privacy-preserving data analysis has a long history with contributions from several disciplines, including information theory, theoretical computer science, and statistics. The lack of a definitive solution and the pressing need for robust privacy-preserving algorithms have led to several notions of privacy including k-anonymity, l-diversity, and t-closeness. In essence, these notions are based on hiding an individual's data among a collection of similar entries, thereby rendering individual data unrecognizable. These models are largely oblivious to underlying distributions and do not account for side information that an adversary might have.

Motivated by a Science of Information perspective, we first identify the information content of a database. This requires us to define a basis for distinguishing databases from one another. Naturally, our basis for distinguishability depends on what queries we seek responses to. Queries are essentially functions of the database. It is natural to not distinguish among databases that yield the same response to canonical queries. Indeed, these databases possess identical information from the point of view of the query interface. We note that most machine learning and statistical queries are aimed at gleaning correlations among attributes stored in a database.

An example of a typical database query is: How many people of Caucasian ethnicity following a vegetarian diet suffer from Type-2 diabetes? Here, the query relates to a correlation among ethnicity, diet, and health. The histogram of a database captures all correlations across attributes. For this reason, we distinguish databases based on their histograms. In other words, the histogram of the database encapsulates all the information content of the database with respect to the queries (the observer function). Two databases with the same histograms are indistinguishable. We note that our framework generalizes beyond the notion of histograms. Having identified the information content of a database, the natural questions we are confronted with are the following: Can this information be provided as a response to queries? Can we permit trading of histograms to facilitate the exchange of statistical information? How do we safeguard against privacy breaches?

Most privacy research looks at important classes of queries, such as counts and linear queries, and devises response mechanisms that prevent an adversary from reconstructing an individual's entry from a query response. One drawback of this approach is that although it can bound the amount of private information leaked by responding to a small number of queries, it is unable to guard against privacy breaches if databases are queried an arbitrary number of times. More concretely, in typical current methods, the information leaked is linear in the number of responses to queries. Most current databases are queried continuously and are seldom destroyed in the interest of privacy.

Motivated by these practical considerations, we propose an architecture in which the true database is sanitized via a mechanism that makes it impermeable to privacy breaches. In other words, no matter how many queries the sanitized database is subject to, no adversary will be able to reconstruct the individual database entries even if he or she is provided with responses to all queries. Recognizing that the sanitizing mechanism must be robust to post-processing, we adopt the notion of differential privacy (DP) to identify a sanitizing mechanism. Cynthia Dwork proposed the notion of DP wherein the vulnerability of an algorithm is quantified through the sensitivity of its output to individual entries.[14] Because DP deems it necessary to randomize the output for preserving privacy, the output of an algorithm operating on a database D is characterized by the probability distribution $\mathbb{P}_D(\cdot)$ of its output. The sensitivity of the algorithm to individual entries can therefore be gleaned by the closeness of the distributions $\mathbb{P}_D(\cdot)$ and $\mathbb{P}_{\hat{D}}(\cdot)$, where D, $\hat{D}$ are databases that differ in exactly one individual's entry. A randomized algorithm, referred to in DP literature as a mechanism, is ε – DP if for every pair of databases D, $\hat{D}$ that differ in a single entry and every output y, the log likelihood ratio is

$$\left|\log \frac{\mathbb{P}_D(y)}{\mathbb{P}_{\hat{D}}(y)}\right| \leq \varepsilon =: \ln\theta . \tag{1}$$

A larger ε corresponds to more sensitivity to individual entries—less privacy preserving and vice versa. Because ε quantifies how much an adversary can learn about any individual's entry from the output of an ε – DP mechanism, it must be regarded as a privacy budget. DP enjoys two desirable invariances under composition and post-processing. Composition bounds how much an adversary can learn about an individual's entry if he or she is provided with responses of multiple DP mechanisms. Second, the output of a DP mechanism cannot be post-processed in any way that can reveal more about an individual's entry. These properties, particularly the latter, make it suitable for the problem at hand.

We are thus led to the following architecture. The true database is sanitized via an ε – DP-sanitizing mechanism, where ε is the specified privacy budget. The original database is secured and the sanitized database is employed for querying. We illustrate this approach in the context of constructing histograms over a given database. The framework extends naturally to computation of correlations. The histogram of the database preserves all correlations and the output of any transformation that preserves the histogram retains the information sought by most canonical

database queries. This suggests that the problem of designing histogram-preserving transformations might be a natural approach to preserving privacy and providing "optimal" accuracy. To address the latter issue, we resort to the rate-distortion perspective. In rate-distortion theory, the closeness of an information source to its (compressed) representation is quantified using a fidelity measure between the source and its representation. Correspondingly, we define a measure of fidelity between a pair of histograms to quantify the closeness of the information source (histogram of the original database) to its representation (histogram of the transformed database).

Example 4: Preserving privacy and fidelity

We present a formulation of an optimization problem (linear programming) that preserves DP when minimizing inaccuracy.[15] Consider a database with n subjects. Each subject is identified with a record. We let $\mathcal{R} = \{a_1, \ldots, a_k\}$ denote the set of records. We also write p_k for the probability that a subject's record is $a_k \in \mathcal{R}$. We let $\underline{r} = (r_1, \ldots, r_n) \in \mathcal{R}^n$ denote a (generic) database with n records. The histogram of a database is defined as follows: For a database $\underline{r} \in \mathcal{R}^n$ and a record $a_k \in \mathcal{R}$, we let $h(\underline{r})_k = \sum_{i=1}^{n} 1_{\{r_i = a_k\}}$, where 1_A is the indicator function of A, denote the number of subjects with record a_k, and $h(\underline{r}) = (h(\underline{r})_1, \ldots, h(\underline{r})_K)$ denote the histogram corresponding to database $\underline{r} \in \mathcal{R}^n$. Let $\mathcal{H}^n = \{(h_1, \cdots, h_K) \in \mathbb{Z}^K : h_i \geq 0, \sum_{k=1}^{K} h_k = n\}$ denote the collection of histograms. To protect privacy, we employ a DP database-sanitizing mechanism M (DSM) to output a random sanitized database. More precisely, we define DSM, $M : \mathcal{R}^n \Rightarrow \mathcal{R}^n$ to be θ – DP, where $\theta = e^{\varepsilon}$, if for every pair of neighboring databases $\underline{r}$, $\hat{\underline{r}}$ and every database $\underline{s} \in \mathcal{R}^n$ we have $\theta \mathbb{W}_M(\underline{s}|\underline{r}) \leq \mathbb{W}_M(\underline{s}|\hat{\underline{r}}) \leq \theta^{-1} \mathbb{W}_M(\underline{s}|\underline{r})$, where $\mathbb{W}_M(\underline{s}|\underline{r})$ is the probability of transforming database $\underline{r}$ into $\underline{s}$. Now, we can formulate our optimization problems that minimize inaccuracy, while preserving θ privacy:

$$D_*^n(\theta) = \min_{\mathbb{W}(\cdot|\cdot)} \sum_{\underline{h} \in H^n} \sum_{\underline{g} \in H^n} \binom{n}{h_1 \cdots h_K} \left\{ \prod_{k=1}^{K} p_k^{h_k} \right\} \mathbb{W}(\underline{g}|\underline{h}) |\underline{h} - \underline{g}|_1$$

subject to:

$$\mathbb{W}(\underline{g}|\underline{h}) \geq 0, \sum_{\underline{g} \in H^n} \mathbb{W}(\underline{g}|\underline{h}) = 1, \text{and}$$
$$\mathbb{W}(\underline{g}|\underline{h}) - \theta\, \mathbb{W}(\underline{g}|\hat{\underline{h}}) \geq 0$$

for every pair of histograms $|\underline{h} - \hat{\underline{h}}|_1 = 2$, where $|\cdot|_1$ is L_1 distance. This linear programming problem can be solved for large n, leading to:[15]

$$D_K^*(\theta) = 2\theta \left\{ \frac{K-1}{1-\theta} + \frac{S'_{K-1}(\theta)}{S_{K-1}(\theta)} \right\}, \text{where}$$
$$S_{K-1}(\theta) = \sum_{j=1}^{K-1} \theta^j \left[\binom{K-1}{j} \right]^2 \quad (2)$$

with $S'_{K\text{-}1}(\theta)$ being the derivative of $S_{K\text{-}1}(\theta)$ and the optimal mechanism is given by

$$\mathbb{W}^*(\underline{g}|\underline{h}) = \xi_{\rho,f}^{-1}(\theta)(\theta)^{\frac{|\underline{g}-\underline{h}|_1}{2}},$$

where $\xi_{\rho,f}(\theta)$ is the normalizing constant related to the Ehrhart series.

In the above example, we answer the question: Given ε – DP, how much do we lose in accuracy in terms of its histogram, when revealing a sanitized database? However, a converse problem is more interesting: When releasing an "anonymized" database D' of the true database D, what is a good measure to quantify the loss of privacy from the release of D'? Thus far, we believe there is no good answer to this important question.

There are many other areas of science, engineering, agriculture, and commerce that can benefit from a unifying Science of Information. For example, recent progress in model and method development has yielded spectacular results in areas such as life sciences. Information is an essential aspect of living systems. Paul Nurse argues that biology stands at an interesting juncture because we still lack a comprehensive understanding of important higher-level biological phenomena.[16] He further opines that the successes of disciplines such as system biology must be supplemented by deeper investigations into how living systems gather, process, store, and use information. In economics, questions of how information is valued are important, as is the flow of information in economic systems and associated control problems. In social sciences, it is critical to understand the flow of information, its influence on individual and collective action, and its impact on overall social and economic well-being.

To achieve some of these challenging goals, we must revamp and modernize our educational approach to science and engineering. It is widely acknowledged that today's students require a sophisticated set of information skills. We need education and diversity programs specifically targeted to the growing gap among academia, industry, government, and non-profits, which are all in critical need of trained individuals with diverse backgrounds,

experiences, and ideas. These individuals must have literacy in the Science of Information as well as interdisciplinary capabilities and domain-specific expertise necessary for addressing complex problems in big data and its applications.

ACKNOWLEDGMENTS

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370. We would like to thank members of the Center for Science of Information for their contributions to the scientific agenda of the Center. In particular, we are grateful to D. Gleich, P. Dineas, A. Magner, and A. Padakandla for their input and contributions to this article.

ABOUT THE AUTHORS

WOJCIECH SZPANKOWSKI is the Saul Rosen Distinguished Professor of Computer Science at Purdue University. His research interests include algorithm analysis, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. He received a PhD in electrical and computer engineering from Gdańsk University of Technology. In 2008 Szpankowski launched the interdisciplinary Institute for Science of Information, and in 2010 he became the director of the newly established NSF Science and Technology Center for Science of Information. He also received the Humboldt Research Award in 2010. Szpankowski is a Fellow of IEEE and an Erskine Fellow. Contact him at szpan@purdue.edu.

ANANTH GRAMA is a professor of computer science at Purdue University. His research interests include parallel and distributed computing, large-scale data analytics, and applications in life sciences. Grama received a PhD in computer science from the University of Minnesota. He is a recipient of the National Science Foundation CAREER award and the Purdue University Faculty Scholar Award. Grama is a Fellow of the American Association for the Advancement of Sciences and a Distinguished Alumnus of the University of Minnesota. Contact him at ayg@purdue.edu.

REFERENCES

1. T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2006.
2. D. Angluin, "Queries and Concept Learning," *Machine Learning*, vol. 2, no. 4, 1988, pp. 319–342.
3. L.G. Valiant, "A Theory of the Learnable," *Comm. ACM*, vol. 27, no. 11, 1984, vol. 1134–1142.
4. V. Vapnik and A. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, 2004, pp. 264–280.
5. S.A. Motahari, G. Bresler, and D. Tse, "Information Theory of DNA Sequencing: A Basic Model," *IEEE Int'l Symp. Information Theory Proc.* (ISIT 12), 2012; doi: 10.1109/ISIT.2012.6284020.
6. C. Shannon, "The Lattice Theory of Information," *Trans. IRE Professional Group on Information Theory*, vol. 1, no. 1, 1953, pp. 105–107.
7. C. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical J.*, vol. 27, no. 3, 1948, pp. 379–423.
8. W. Szpankowski, "On Asymptotics of Certain Recurrences Arising in Universal Coding," *Problems of Information Transmission*, vol. 34, no. 2, 1998, pp. 142–146.
9. Y. Choi and W. Szpankowski, "Compression of Graphical Structures: Fundamental Limits and Algorithms and Experiment," *IEEE Trans. Information Theory*, vol. 58, no. 2, 2012, pp. 620–638.
10. F.P. Brooks, "Three Great Challenges for Half-Century-Old Computer Science," *J. ACM*, vol. 50, no. 1, 2003, pp. 25–26.
11. A. Magner et al., "TIMES: Temporal Information Maximally Extracted from Structures," to be published in 2018; www.cs.purdue.edu/homes/spa/papers/www2018.pdf.
12. N. Santhanam et al., "Data Dependent Weak Universal Redundancy," *IEEE Int'l Symp. Information Theory*, (ISIT 14), 2014; doi; 10.1109/ISIT.2014.6875159.
13. J. Ponniah, Y.-C. Hu, and P.R. Kumar, "A Clean Slate Approach to Secure Ad Hoc Wireless Networking-Open Unsynchronized Networks," *IEEE Trans. Control of Network Systems*, vol. 4, no. 1, 2017, pp. 37–48.
14. C. Dwork, "Differential Privacy," *Proc. 33rd Int'l Colloquium Automata, Languages and Programming* (ICALP 06), 2006, Springer Verlag; www.microsoft.com/en-us/research/publication/differential-privacy.
15. A. Padakandla, P.R. Kumar, and W. Szpankowski, "Preserving Privacy and Fidelity via Ehrhart Theory," preprint, 2017; www.cs.purdue.edu/homes/spa/papers/isit18-privacy.pdf.
16. P. Nurse, "Life, Logic, and Information," *Nature*, vol. 454, 2008, pp. 424–426.

myCS Read your subscriptions through the myCS publications portal at **http://mycs.computer.org**