Frontiers of Science of Information: Shannon Meets Turing[1]

Wojciech Szpankowski and Ananth Grama
Center for Science of Information
Purdue University, W. Lafayette, IN 47907, USA
emails: {szpan,ayg}@purdue.edu

Abstract

Information and computation are two of the defining concepts in the modern era. Shannon laid the foundation of information (theory), demonstrating that problems of communication and compression can be precisely modeled, formulated, and analyzed. Turing, on the other hand, formalized the concept of computation, defined as the transformation of information by means of algorithms. In this article, we generalize the concept of information beyond its roots in compression and communication. We motivate a Science of Information to address key challenges in transforming *data* to *information* to *knowledge* (meaning) thought information efficient computation. We argue that a formal quantification of information should precede the development of methods in information efficient computing. In this framework, we first establish fundamental limits on information, and then seek efficient algorithms that achieve this limit. In this outlook article, we first define learnable (useful) information based on distinguishability, and highlight several novel aspects of information within the scope of Science of Information. We then discuss two important research areas: data science and privacy in our formal framework, demonstrating that our model for information efficient computation provides a powerful unifying framework for these investigations.

Keywords: Learnable information, data, knowledge, information efficient computation, data science, privacy.

1 Introduction

We describe an important and emerging area of research – the Science of Information, its relation to diverse problems in data science, life sciences, communication, economics and its applications in various domains.

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 discussed and debated in literature over the past several centuries. We define information 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, since 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 corresponds to the *Science of Information*.

We define *knowledge* (or meaning/ insight) as a combination of information and semantics associated with the domain. The notion of semantics associated with the domain enables

[1]This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01.

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 may 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 Criteria (BIC) are commonly used for selecting appropriate models for data, in conjunction with suitably formulated optimization procedures [4]. 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 [4]. We should add that there are other approaches advocated by the computational learning theory community, such as exact learning proposed by Angluin [1], probably approximately correct learning (PAC learning) proposed by Valiant [16], VC theory proposed by Vapnik and Chervonenkis [17], and Bayesian inference.

We focus here 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 address these questions, Science of Information leverages techniques from information theory, statistics, algorithms and analysis, combinatorics, numerical methods, and learning.

Claude Shannon laid the foundation of information theory, demonstrating that problems of data communication can be precisely modeled, formulated, and analyzed. He also provided basic mathematical tools for addressing these problems. While Shannon's focus was on data *recovery* in compression (source coding) and communication (channel coding), Science of Information generalizes this framework to a broad class of analyses, since information is not merely communicated but it 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 met, then no algorithm can accomplish desired tasks (e.g., 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. We propose here a new concept, which we refer to as *information efficient formulations* that simultaneously satisfy fundamental limits of information, while being computationally tractable.

We elaborate on the concept of information efficient formulations using two examples. It is well known that the problem off-line 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, achieving 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 [7]. More generally, since real data is almost inevitably corrupted, noisy, and/or incomplete, focusing on the computational hardness of deriving optimal solutions is not particularly useful in 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 those limits (e.g., there is an "information-computation gap" in the DNA assembly problem [7] and in the community recovery problem of stochastic

block model). In this sense, an analysis of limits of information must precede efforts aimed at method development (computation) for transforming data to information through algorithms. This motivates our dictum: "Shannon meets Turing".

Beyond generalizing the scope of operations, Science of Information incorporates elements of *structure, space, time, connectivity*, and *semantics* into Shannon's theory of information. 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 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 fundamental bounds on extraction of information from large data repositories?

There have been significant advances in broad areas of machine and computational learning, information theory, privacy and security, 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 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? There is pressing need for formal modeling and analyses for answering these questions. Even as we address these problems, there have been increasing concerns relating to data privacy, particularly as it relates to data analytics and querying. There are clear tradeoffs of accuracy (of querying and analysis) and disclosure, which must be analytically quantified and studied. Science of Information may provide a unifying framework, within which these problems can be formulated and studied in a principled and unified manner.

In this perspective article, we present our vision for the Science of Information (Section 2), which integrates key elements of Shannon and Turing postulates into an overarching framework for data science (Section 3), privacy, and security (Section 4).

2 What is Science of Information?

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:[2] "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" [13].

We adopt here 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",

[2]In 1948 paper [12] Shannon also 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."

that maps a dataset to an observed object. The observer function in traditional information theory may correspond to a *channel*. In a more general setting, the observer function may 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 (e.g., 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 below). Note that we implicitly associate measures of quality (e.g., 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 (minimum number of) balls B_i, such that all elements in a single ball B_i are indistinguishable w.r.t. 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 $\varepsilon \to 0$.

We first present an example of how this general definition can be specialized into a traditional framework of reliable transmission. Then we discuss another example, in which we demonstrate this function in the context of distinguishable sources in universal learning.

Example 1 [Shannon Channel Capacity] Consider a simple point to point communication channel, as originally introduced by Shannon [12]. We want to transmit a binary sequence of length n over a channel. The goal is to ensure that the probability of error $P_e := f$ on the receiving side tends to zero, as n increases. Clearly, we cannot transmit reliably all 2^n sequences. However, can we transmit reliably 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 w.r.t. 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] We now turn our attention to *computable* learnable information contained in a sequence $x^n = x_1 \dots x_n$ 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 these two models, θ and θ' are too close to each other, they are virtually indistinguishable, and they 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 (i.e., over an alphabet of size m). As discussed above, we will use the MDL (Minimum Description Length) estimator [4]. For a distance between distributions/ models we adopt the Kullback-Leibler (KL) divergence $D(\cdot||\cdot)$. Let us define $B_{KL}(\theta_0,\varepsilon) = \{\theta : D(\theta||\theta_0) \le \varepsilon\}$ 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 up the error probability at $O(1/\sqrt{n})$, we observed in [15] 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 a Mellin transform and singularity analysis of the Lambert-W function we find that [15]

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

where Γ is the Euler gamma function. 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 [4]. □

We 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, associated with it, explicit or latent structure and organization. We lack measures and meters to define and quantify information embodied in structure and organization,[3] however, see [3]. 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, and incorporate diverse dynamic observable and system state.

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

Space: In interacting systems, spatial localization often limits exchange of data – with obvious disadvantages as well as benefits. These benefits typically result from reduction in interference as well as ability of a 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, data is exchanged in space and time for decision making, thus timeliness of delivery along with 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 (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). These considerations of dynamic sources must often be incorporated into the information framework.

Limited Resources: In many scenarios, data is limited by available resources (e.g., computing devices, bandwidth of signaling channels). How much information can be extracted from

[3] In [2] F. Brooks writes: "Shannon and Weaver performed an inestimable service by giving us a definition of Information and a metric for 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."

data 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. For example, a hundred digit number sent to a recipient that can only store one digit at a time may 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. For example, imagine two texts, one in English and the other semantically identical text in Polish. These 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* may arise for a number of different reasons: the desired information may be *computationally* hard to extract; the information may be *distributed* geographically and not locally extractable; or information may 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, in the computational setting, information is not well understood qualitatively, 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 (e.g., electronic voting). All of this leads to many profound questions such as: how many bits of information does a *quantum* state contain? how does one quantify *misunderstanding*? and what is *time* in a distributed setting?

Cooperation. 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, 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.

3 Information and Data Science

Problems in data science are often confounded by issues of extreme scale ("big data") of data (typically on distributed platforms), whereas issues such as high dimensionality and sparsity patterns that manifest at multiple scales, dynamic, temporal, heterogeneous structure, complex dependencies, noise and missing data, as well as semantics are equally important. A comprehensive theory for the foundation of Data Science must pay particular attention to the *learnable and distinguishable* information contained in the data and focus on the design and analysis of algorithms that (i) extract such information from the data; (ii) have computational cost (in terms of both accuracy and running times) that depends on useful information in data and not the size of the data leading to our information efficient computation paradigm; and (iii) be robust in a statistical sense (e.g., perform well in the presence of noise and other artifacts). Our framework for information, presented in Section 2, aims to address precisely these questions. By integrating semantics into this information framework, we can abstract from data to knowledge and to actionable insights, as presented in Figure 1.

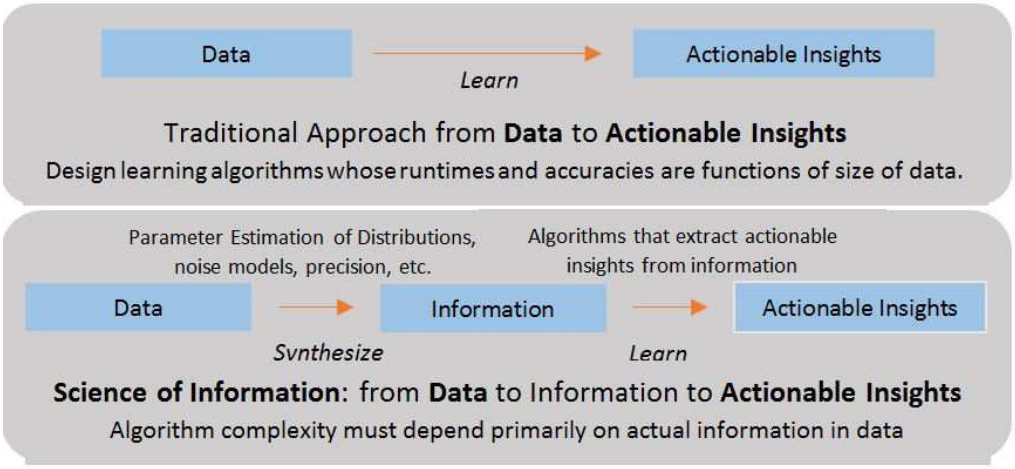

Figure 1: The Science of Information approach to extracting actionable insights from data.

We initiate and motivate our discussion using an example of a problem for which we provably cannot extract learnable information irrespective of algorithms we use.

Example 3.[Node Arrival Order in a Graph] Consider a dynamic network in which nodes are added (and perhaps deleted). Each node, has a time-stamp of its arrival. However, these time-stamps may not be available to us (e.g., Facebook users with their privacy settings, or the evolution of a protein interaction network). Given a graph structure, without time-stamps, our task is to rank all nodes from the oldest to the youngest. It turns out, there is no good solution to this problem for graphs generated by Erdős-Renyí and preferential attachment models [6]. More precisely, for these models no algorithm can recover a ranking with high probability. Interestingly, one can design algorithms that optimize a likelihood estimator, but it can be proved that with high probability these algorithms fail to yield the ground truth. □

Information and Limits of Learning. A natural and important question in data analytics applications is – what are the limits of learning 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 are sampled). A number of analytics applications operate in the model-free regime – i.e., 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 fall back 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 may 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.

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 in Section 2.

Information and Optimality of Methods. There are a number of important consequences of our information and learning framework for development of methods. While many learning techniques are posed as optimization problems (w.r.t. 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 show, analytically, that there are a large number of arrival sequences that are all optimally equiprobable. This indicates that whereas it is possible (relatively easily) to derive an optimal arrival sequences, 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 data point signify? When working with real-world data samples, we only receive information about those samples that actually exist (e.g., 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 will help guide future experiments that

directly measure new datapoints. The first fundamental challenge here, often ignored, is to assess whether there is any learnable information that can be extracted (particularly in the context of highly incomplete database, say of Netflix)?

Data Dependent Consistency Models. In order to apply rigorous information theory techniques (such as minimax, pointwise vs. uniform convergence [4]) to data science, the first question we face is related to model selection. There is often a 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, i.e., 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. Santhanam et al. [14] departs from this dichotomy, and puts forth a new analysis framework by characterizing rich model classes that may only admit pointwise guarantees, yet all information about the unknown model needed to gauge estimator accuracy can be inferred from the sample at hand.

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 may not be uniform in the weak convergence formulation. Results of [14] show how this can be clarified using the data itself by introducing the notion of *data-driven weak convergence*. 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. In order to characterize such classes of distributions, Santhanam et al. [14] introduced *deceptive distributions*. 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.

Summing up, we argue that Science of Information provides the critically-needed formal basis for Data Science – putting information at the heart of the triad 'from data to information to knowledge' as illustrated in Figure 1.

4 Information and Privacy and Security

As we build complex systems, whose operation crucially relies on information extracted from data, there is increasing focus on ensuring privacy of data and integrity of the methods used for processing data. We can leverage the foundational tools developed by the Science of Information 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 this pipeline transforming data into information. Here, we describe our formulation and findings in the context of privacy and refer the reader to Kumar et al. [10], where novel approaches to security is investigated. 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, on the one hand, with a sound theoretical measure of the statistical information preserved by algorithms manipulating databases, and on the other, 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, combined with 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 it unrecognizable. These models are largely oblivious to underlying distributions and do not account for side information that an adversary may 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 each other. 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 between databases that yield the same response to most canonical queries. Indeed, these databases possess identical information. We recognize that most machine learning and statistical queries are aimed at gleaning correlations between 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 correlation between 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 response to queries? can we permit trading of histograms to facilitate exchange of statistical information? how do we safeguard against privacy breaches? Most work in privacy study important classes of queries, such as counts and linear queries, and devise response mechanisms that prevent an adversary from reconstructing an individual's entry from query response. One drawback of this approach is that, while 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 arbitrary number of times. More concretely, in typical current methods, the information leaked is linear in the number of responses to queries. Most current-day databases are queried continuously and seldom destroyed in the interest of privacy.

Motivated by these practical considerations, we propose an architecture where the true database is *sanitized* via a mechanism that lends it impermeable to privacy breaches. In other words, the sanitized database is such that, no matter how many queries it is subject to, no adversary will be able to reconstruct the individual database entries even if he is provided responses to all queries. Recognizing that the sanitizing mechanism must be robust to post-processing, we adopt the notion of *differential privacy* to identify a sanitizing mechanism.

Dwork et. al. [5] proposed the notion of differential privacy (DP) wherein the vulnerability of an algorithm is quantified through the sensitivity of its output to individual entries. Since DP deems randomizing the output *necessary* 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 at 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 $\epsilon-$DP if for *every* pair of databases $D, \hat{D}$ that differ in a single entry and *every* output y, the

log likelihood ratio

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

A larger ϵ corresponds to more sensitivity to individual entries, i.e., less privacy preserving and vice versa. Since ϵ quantifies how much an adversary can learn about any individual's entry from the output of an $\epsilon-$DP mechanism, it must be regarded as a privacy budget. DP enjoys two desirable invariance under composition and post-processing. Composition bounds how much an adversary can learn about an individual's entry if she is provided responses of multiple DP mechanism. 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, lend it suitable for the problem at hand.

We are thus led to the following architecture. The true database is sanitized via an $\epsilon-$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 construction of *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 may 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 here a formulation of an optimization problem (linear programming) that preserves differential privacy when minimizing inaccuracy. We will follow here the notion of [9]. Consider a database with n *subjects*. Each subject is identified with a *record*. We let $\mathcal{R} = \{a_1, \cdots, a_K\}$ 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 $\underline{r} = (r_1, \cdots, r_n) \in \mathcal{R}^n$ 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 $\mathtt{h}(\underline{r})_k = \sum_{i=1}^n \mathbf{1}_{\{r_i = a_k\}}$, where $\mathbf{1}_A$ is the indicator function of A, denote number of subjects with record a_k, and $\mathtt{h}(\underline{r}) = (\mathtt{h}(\underline{r})_1, \cdots, \mathtt{h}(\underline{r})_K)$ 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. In order 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 $\theta-$DP, where $\theta = e^\epsilon$, 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:

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

subject to: (i) $\mathbb{W}(\underline{g}|\underline{h}) \geq 0$, (ii) $\sum_{\underline{g} \in \mathcal{H}^n} \mathbb{W}(\underline{g}|\underline{h}) = 1$, and (iii) $\mathbb{W}(\underline{g}|\underline{h}) - \theta\ \mathbb{W}(\underline{g}|\hat{\underline{h}}) \geq 0$ 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, leading to [9]:

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

with $S'_{K-1}(\theta)$ being the derivative of $S_{K-1}(\theta)$ and the optimal mechanism given by $\mathbb{W}^*(\underline{g}|\underline{h}) = \mathscr{E}_{\mathcal{P},f}^{-1}(\theta)(\theta)\theta^{\frac{|\underline{g}-\underline{h}|_1}{2}}$, where $\mathscr{E}_{\mathcal{P},f}(\theta)$ 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 release of D'? Thus far, we believe there is no good answer to this important question.

5 Conclusions

There are many other areas of science, engineering, economics, social sciences, agriculture, commerce, and beyond, 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 [8] that biology stands at an interesting juncture since we still lack a comprehensive understanding of important higher-level biological phenomena. 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. Flow of information in economic systems and associated control problems are vitally important, and have been recognized through recent Nobel Prizes. In social sciences, understanding flow of information, its influence on individual and collective action, its impact on overall social and economic well-being are vitally important.

Finally, we emphasize that in order 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 between academia, industry, government, and non-profits, which are all in critical need of trained individuals, from diverse backgrounds, experiences, and ideas. These individuals must have literacy in science of information, interdisciplinary capabilities, and domain-specific expertise necessary for addressing complex problems in big data and its applications.

In 2010 NSF established the Center for Science of Information (`http://www.soihub.org`) as one of the Science and Technology Centers [11]. The Center 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.

Acknowledgment

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 Profs. D. Gleich and

P. Dineas, and Drs. A. Magner and A. Padakandla for their input and contributions to this article.

References

[1] D. Angluin. Queries and concept learning. *Mach. Learning*, 2 (1988), pp. 319-342

[2] F. Brooks. Three great challenges for half-century-old computer science. *J. the ACM*, 50:25–26, 2003.

[3] Y. Choi and W. Szpankowski. Compression of graphical structures: Fundamental limits and algorithms and experiment. *IEEE Trans. on Information Theory*, 58:620–638, 2012. *IEEE Trans. on Information Theory*, 58:620–638, 2012.

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 2006.

[5] C. Dwork, Differential privacy, *Proceedings of the 33rd International Conference on Automata, Languages and Progra mming - Volume Part II*, ser. ICALP'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 1–12.

[6] A. Magner, A. Grama, J. Sreedharan, and W. Szpankowski. Recovery of vertex orderings in dynamic graphs. In *IEEE International Symposium on Information Theory*, Aachen, 2017.

[7] S. A. Motahari, G. Bresler, and D. Tse. Information theory of DNA sequencing: A basic model. *IEEE Trans. on Information Theory*, 59:6273–6289, 2013.

[8] P. Nurse. Life, logic, and information. *Nature*, 454:424–426, 2008.

[9] A. Padakandla, P. R. Kumar, and W. Szpankowski, Preservaing privacy and fidelity via Ehrhart theory. preprint 2017.

[10] J. Ponniah, Y. Hu and P.R. Kumar, A clean slate appproach to secure wirless networking, Foundation and Trends in Networking, 9, 1, 1-105, 2014.

[11] N. Savage, Information Theory After Shannon, *Commun. of the ACM*, 54, 16-18, 2011.

[12] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, (27):379–423 623–656, 1948.

[13] C. Shannon. The lattice theory of information. *IEEE Transaction on Information Theory*, 1:105–107, 1953.

[14] N. Santhanam, V. Anantharam, A. Kavcic, and W. Szpankowski. Data dependent weak universal redundancy. In *IEEE International Symposium on Information Theory*, 2014.

[15] W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34, No.2, 142-146, 1998. 34, No.2, 142-146, 1998.

[16] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27, 1984.

[17] V. Vapnik, A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16(2), 264 – 280, 2004.

Wojciech Szpankowski is Saul Rosen Distinguished Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. He received his M.S. and Ph.D. degrees in Electrical and Computer Engineering from Gdansk University of Technology. He held several Visiting Professor/Scholar positions, including McGill University, INRIA, France, Stanford, Hewlett-Packard Labs, Universite de Versailles, University of Canterbury, New Zealand, Ecole Polytechnique, France, the Newton Institute, Cambridge, UK, ETH, Zurich, and Gdańsk University of Technology, Poland. He is a Fellow of IEEE, and the Erskine Fellow. In 2010 he received the Humboldt Research Award. He published two books: "Average Case Analysis of Algorithms on Sequences", John Wiley & Sons, 2001, and "Analytic Pattern Matching: From DNA to Twitter", Cambridge, 2015. He has been a guest editor and an editor of technical journals, including Theoretical Computer Science, the ACM Transaction on Algorithms, the IEEE Transactions on Information Theory, Foundation and Trends in Communications and Information Theory, Combinatorics, Probability, and Computing, Algorithmica, and *Journal of Discrete Algorithms*. In 2008 he 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.

Ananth Grama is a Professor of Computer Science at Purdue University. His primary areas of interest are parallel and distributed computing, large scale data analytics, and applications in life sciences. He received his Ph.D. from the University of Minnesota in 1996, his M.S. in Electrical Engineering from Wayne State University in 1990, and his B. Engg. from the Indian Institute of Technology, Roorkee in 1989. He was the Director of the Computational Science and Engineering and Computational Life Sciences programs at Purdue from 2012-16, and Chaired the Biodata Management and Analysis Study Section of National Institutes of Health from 2012-14. He is a recipient of the National Science Foundation CAREER award (1998), University Faculty Scholar Award (2002-07), is a Fellow of the American Association for the Advancement of Sciences (2013), and a Distinguished Alumnus of the University of Minnesota (2015).