The Potential Value of Information and Data – An SFI Working Group

April 2 – April 3, 2024

Participants Short Summaries

Working Group Summary – Amos Golan

The goal of this working Group (WG) was to study and explore the basic notion of value of information and the potential value of data. We concentrated on the following Four Questions:

1. Can information theory be used for developing new tools for evaluating the full potential, and potential value, of datasets and the information stored in these data?
2. Is it possible to extend information theory to account for the meaning of the information embedded in the data?
3. Is the value of information independent of the inferential approach used?
4. Is it possible to measure the value (or potential value) of a model or a theory?

To explore these questions, the fundamental question of 'what is information' arose naturally.

In my opening talk I discussed the basic background related to value and potential value of information, as well as the fundamental question of what is 'information.' Rather than provide my thoughts (and definitions), I expressed the different ideas as questions so as not to bias the discussion before we start.

The group was inherently interdisciplinary (math, ecology and evolution, complex systems, physics, information theory, information-theoretic inference, computer science, visualization, political science, international development, ecosystem ecology, biodiversity, biochemistry and personalized cancer therapy, economic statistics, agricultural economics, electrical engineering, geosciences and, social sciences, philosophy, law, statistics, data science, computational biology, economics and econometric) which triggers a nice discussion, though no basic agreements on the concept of value or potential value.

Though much of the debate is to do with the notion of 'value' and subjectivity, some of the focus was on trying to think about the problem from a more practical point of view. Among the most debated issues were the (i) relationship and dependence (or independence) between a model, the answer that model is designed to solve and the information in the data used, and (ii) the relationship between Shannon information (or 'communication') theory, meaning and value.

Overall, this set of questions is tough to sort out and demands much work and research. However – and possibly due to the interdisciplinary nature of the group – many interesting, and at times, new ideas came up in the conversations that may open the door for further study of that topic.

There is much more to do, but I see it as a very productive start.

Summary of the SFI Working Group on the value of data and information

John Harte

My initial hoped-for outcome of this Working Group was a quantifiable rule for estimating the value of data and/or information. This anticipated rule would resolve questions such as: Does value scale additively or multiplicatively with some measure of the quantity or cost of data? Is value most naturally expressed on a logarithmic or a linear scale? Does value derive from an axiomatic foundation, the way Shannon Information Entropy does? Clearly such an outcome would be of great "value", if for no other reason than providing a rationale for funding the collection of data.

Yet I rather dogmatically titled my opening presentation: "The value of data depends on what you do with it". In part, this was to incite others to prove me wrong, to come up with a foundational basis for assigning value to data and information irrespective of the hypothesis, model, or theory that the data are used to test, irrespective of the outcomes of such tests.

In my talk I discussed, in the context of my Maximum Entropy Theory of Ecology, what constitutes useful versus useless ecosystem census data. Two data sets, with the same Information Entropy, can provide very different value; moreover, the data set that is useless in the context of my theory can be extremely useful for initializing or testing a different theory.

From the Working Group discussions and presentations, the contextuality of value seems to be the conclusion across disciplines and across the various analytical frameworks used to make better use of data.

The value of data or information is, I had concluded in my talk, much like the value of time. Without question, the time spent at SFI with this Group was extremely valuable. It would, of course, have been easy to waste those two days doing something else. While my initial hope for a quantifiable foundational approach to the value of data and information was not to be, the group process of arriving at that conclusion was fun and fulfilling. Whether the conclusion is final may require another meeting.

The Value of Data and Information — The Four Professors Problem

Min Chen, University of Oxford, United Kingdom

The *four professors problem* was introduced to exemplify the relative merits of, and the characteristic trade-offs among, four major types of processes in data intelligence workflows, namely, *statistics*, *algorithms*, *visualization*, and *interaction*.

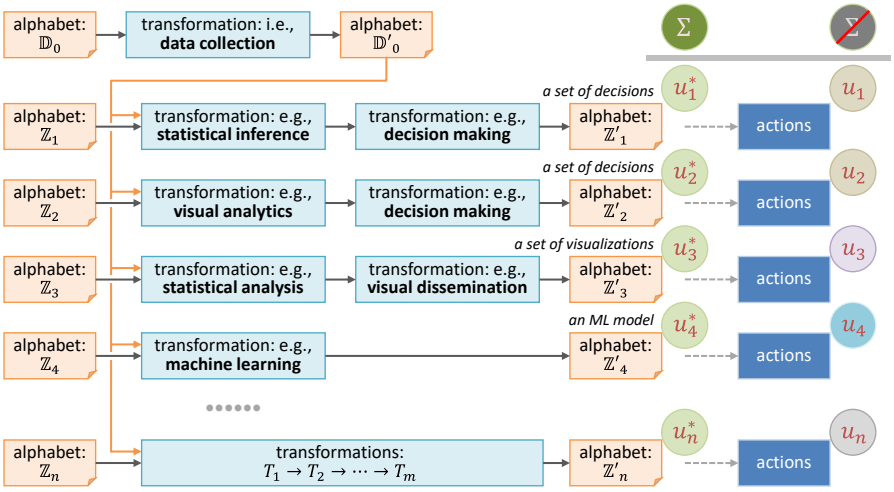
Four professors sit down to discuss the examination marks of N students.

- Professor S (who teaches statistics) argues that we should just consider a few statistical measures such as the mean value of the N marks and the standard deviation.
- Professor A (who teaches algorithms) argues that we should just use algorithms to make the decision, e.g., first determining the minimum and maximum marks, and then dividing the N marks into four equal-sized groups.
- Professor V (who teaches data visualization) argues that we should just visualize the N marks, e.g., using a bar chart and then make decisions.
- Professor I (who teaches human-computer interaction) argues that whenever we need the information about a specific student, we should just type the name of the student into the computer to search the database where the N marks and other relevant information are stored and make decisions based on the search results.

Of course, we all know that the four professors should really work together. This raises a fundamental question as to how to optimise the uses of machine-centric processes (e.g., statistics and algorithms) and human-centric processes (e.g., visualization and interaction) in a data intelligence workflow, where data is transformed to decisions. Information-theoretically, a dataset is a collection of instances of valid letters in a data alphabet $\mathbb{D}_0$, and a decision is an instance of a valid letter in a decision alphabet $\mathbb{Z}_i (i = 1,2,\cdots,n)$. Here "decision" is a generalized term that encompasses the outcomes of a data intelligence workflow, such as a decision, a mathematical expression, a machine-learned model, a piece of representable knowledge, or so on. Chen and Golan proposed a cost-benefit ratio [1],

$$\frac{\text{benefit}}{\text{cost}} = \frac{\text{alphabet compression} - \text{potential distortion}}{\text{cost}} = \frac{\mathcal{H}(\mathbb{A}_j) - \mathcal{H}(\mathbb{A}_{j+1}) - \mathcal{D}(\mathbb{A}'_j||\mathbb{A}_j)}{\text{cost}}$$

for assessing any process P_j in a data intelligence workflow. Here $\mathbb{A}_j$ and $\mathbb{A}_{j+1}$ are the input and output alphabets of P_j respectively. $\mathbb{A}'_j$ is an alphabet that has the same letters as $\mathbb{A}_j$ but likely has a different distribution. $\mathbb{A}'_j$ is reconstructed from $\mathbb{A}_{j+1}$. $\mathcal{H}()$ is the Shannon entropy and $\mathcal{D}()$ is a divergence measure bounded by $[0, \mathcal{H}_{\max}(\mathbb{A}_j)]$ [2]. The unit of the benefit term is *bit* if binary logarithm is used. The fundamental cost is energy, and its unit is *joule*, which can be approximated using time, a monetary quantity, etc. This cost-benefit ratio can be used to compare different processes as well as different workflows. Building on this concept, Chen and Ebert formulated a systematic methodology for optimizing a data intelligence workflow that can be used in practice in a coarse-grain manner [3].



This working group meeting (at SFI on April 2-3, 2024) provides an opportunity to consider the feasibility of using the benefit term as an *informative measure* of the value of data. During the working group meeting, this proposal was shown to have met several critical criteria suggested by the participants, such as (i) a value measurement must reflect the quality of data and decisions; (ii) because a dataset may be used by different workflows, the values obtained for these workflows must be additive; and (iii) the value measurement must capture the essence of usefulness while maintaining a reasonable level of abstraction and generality. As shown in the above figure, we can consider using the benefit term as an informative value $u_i^* (i = 1,2,\cdots,n)$ with a common unit *bit*, while defining the application-specific utility functions as $u_i = f(u_i^*, \alpha_1, \alpha_2, \cdots)$ where $\alpha_1, \alpha_2, \cdots$ are application-specific variables.

[1] M. Chen and A. Golan, What may visualization processes optimize? 2017. doi: 10.1109/TVCG.2015.2513410
[2] M. Chen and M. Sbert. A bounded measure for estimating the benefit of visualization (Part I). 2022. doi:10.3390/e24020228
[3] M. Chen and D. S. Ebert. An ontological framework for supporting the design and evaluation of visual analytics systems. 2019. doi:10.1111/cgf.13677

On the Intrinsic Value of Information

Radu Balan

Abstract (Draft)

The objective of this project is to investigate whether it is possible to construct an intrinsic (objective/absolute) value of information. We take the notions of "intrinsic value", "absolute value", and "objective value" to be exactly the same here. There are two ways to construct such a measure. The first is that such an intrinsic (absolute) notion of information is based on the information's intrinsic value; it is free of the observer, the agent or any other reference set. The second is to construct the absolute value of information based on subjective preferences. We show two basic results. The first is that an absolute value of information cannot be subjectively based. The second is that there exists a value function that satisfies basic principles but it has two major flaws: it grows with the number of elements in the information set (but remains finite), and it is not free of subjectivity – it a function of some inherent, and non-unique weights. We conclude that an intrinsic value of information function does not exist.

More specifically, we first introduce the axioms that any value of information function should satisfy. Let X denote an information set, e.g., actions, decisions, objects. Let $\mathcal{P}$(X) denote the power set of X, that is, the set of all subsets of X, $\mathcal{P}(X) = \{A\ ,\ A \subset X\}$. In the following we assume the set X is finite of cardinal N. A map $v: \mathcal{P}(X) \to \mathbb{R}$ is called a *value of information* if it satisfies four properties: positivity, normalization, monotonicity, and concavity:

P1. (positivity) For any $A \subset X$, $v(A) \geq 0$;

P2. (normalization) $v(\emptyset) = 0$, $v(X) = 1$.

P3. (monotonicity) For any $A, B \subset X$ if $B \subset A$ then $v(B) \leq v(A)$;

P4. (concavity) For any , $B \subset X$, $v(A) + v(B) \leq v(A \cup B) + v(A \cap B)$.

Our results so far show that the collection of value of information functions form a compact convex set with a finite number of extreme points. The number of extreme points is finite but very large (double exponential in N). Different hierarchical levels produce different super additive measures: first order extreme points correspond to ordinary (additive) measures; second order extreme points correspond to the Dempster-Shafer evidence theory. Higher order extreme points remain to be found and classified.

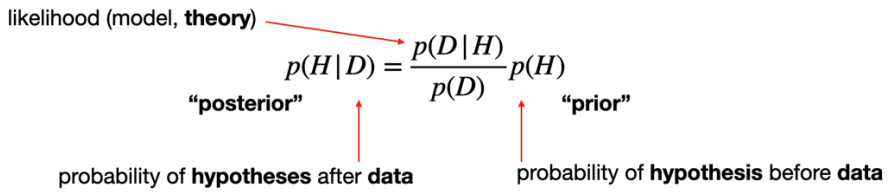
The Potential Value of Information and Data

Luís M. A. Bettencourt, University of Chicago, April 2024

The potential value of information and data is a natural question in complex systems, but it does not arise in traditional disciplines such as Physics, with its focus on energy, or even in information theory by itself, because of its principal focus on information transmission and inference. It is by connecting issues of energy flow and information that the potential value of data can be defined and assessed.

Specifically, all agent in complex system are dissipative and consequently are required to obtain free energy (resources) from their incompletely known environments. This defines the most essential objective for any agent, to which others can be added if necessary. Obtaining energy from an environment requires predicting its states, often from existing signals, and acting accordingly, thus "investing" energy. Doing this optimally requires incorporating such signals and data from past experience optimally, which is done via Bayesian updating of the conditional probability of environmental states given "data".

This structure of energy investment, prediction of the environment, acting accordingly, and reaping benefits (probabilistically) is general and can be applied to any situation by changing the objective, the environment, and the signals that the agent can access. The potential value of data then follows from the resources that can be returned by this process, relative to present resources. The figure below summarizes this scheme:



optimal way of learning+predicting: choose H that maximizes $p(H|D)$ — Bayes classifier

information gain from datum: $D(p(H|D)||p(H)) = \sum_H p(H|D) \log \frac{p(H|D)}{p(H)} = f(D)$ — relative entropy

information about H from all data: $I(H,D) = \sum_D p(D) D(p(H|D)||p(H))$ — mutual information

This formalization also clarifies that the value of data depends on the theory of the environment adopted (likelihood), which is a model for what should be observed (data), if a specific hypothesis (state of the environment) is realized. An observer with a 'bad' theory will interpret the data as gibberish and find no value in it, whereas one with a good theory can realize a high value. This means that it is often worth it to store data as a "public good", in the hope that better theories can be found among a larger population of agents or discovered at later times. This also means that for data to be created and stored (which is costly) it will likely have an intended primal use, with some tangible expected value. However, in general that value is almost certainly a *lower bound* to the potential value of data and the information it conveys on a wider range of problems not initially considered or even imagined. Thus, it is in general very difficult to place an upper bound on the value of data and the information it can convey on a range of problems.

Basic Questions:

1. Can information theory be used for developing new tools for evaluating the full potential, and potential value, of datasets and the information stored in these data?

Yes: Information Theory + (resource) value of Prediction

2. Is it possible to extend (if possible) information theory to account for the meaning of the information embedded in the data?

Yes: include an objective, prediction and its probable returns in energy or resources.

3. Is the value of information independent of the inferential approach used?

No, the value of data, and information are always defined via the “theory”.

4. Is it possible to measure the value (or potential value) of a model or a theory?

Yes, it is the “likelihood”. Its value depends on data, prediction objective, and its returns

As these elements are not all foreseeble and are cumulative, the potential value remains open and likely higher

Summary of presentation

My works focusses on the value of qualitative information for inference to best explanation. Many scholars in my field (political science) are skeptical of the contributions that qualitative information can make to causal inference (Don Green has gone as far as asserting that nothing can be learned from observational evidence), and qualitative information tends to be neglected relative to quantitative data. That tendency also exists in other fields. Many doctors seem to equate scientific diagnosis with running tests that give hard quantitative information. But qualitative information can be invaluable for figuring out what is going on, whether it comes from listening to a patient describe their symptoms and family history,[1] or interviewing key informants about a policy process to understand how the govt was able to enact a reform.

When qualitative information is considered, the reasoning is often sloppy, biased, or even illogical. And much of the guidance in political science for how to handle qualitative evidence is informed by frequentist statistics, which is highly problematic. According to its own foundation principles, frequentism can only be used to analyze stochastic data. But it does not make sense to pretend that qualitative information from expert interviews or archival records or news reports is in any way analogous to a random sample. King, Koehane & Verba's book, *Designing Social Inquiry,* is a prominent example of frequentist-inspired prescriptions for qualitative research. Many qualitative scholars thought their approach was misguided, but leading efforts to push back have not managed to articulate a rigorous and cogent alternative. So despite a lot of effort, scholars have not been doing a good job of extracting the inferential value of qualitative information.

Andy Charman and I have been trying to convince social scientists that Bayesianism provides a far better inferential framework. Our book in essence takes E.T. Jaynes' work on probability as extended logic and translates it into guidance for rational reasoning with qualitative evidence (Fairfield & Charman 2022). In plain language, the five basic steps are as follows:

(i) Consider a pair of rival hypotheses, H_1 and H_2
(ii) Use any relevant background knowledge (I) to assess how plausible H_1 is relative to H_2 [*prior odds:* $P(H_1|I)/P(H_2|I)$]
(iii) Find some evidence, E
(iv) Evaluate how strongly E favors H_1 relative to H_2 [*likelihood ratio:* $P(E|H_1 I)/P(E|H_1 I)$] by mentally inhabiting the world of each hypothesis and asking which one makes E more expected
(v) Update your prior odds to obtain your *posterior odds* [$P(H_1|E I)/P(H_2|E I)$]*:* whatever our prior odds were, we gain confidence in whichever hypothesis makes the evidence more expected.

In my context, hypotheses are plain language propositions that aim to explain how and why something happened. Evidence can be any salient and well-documented observation that

[1] In a recent interview with Terry Gross, Dr. Elizabeth Comen (author of *All in Her Head*) commented: "I don't think additional testing is necessarily the way to go. Most of the time you can discern what's going on with a patient by listening to them."

bears on the truth of the hypotheses, with essentially no restrictions. Consider the example below:

> H_Z = The covid-19 pandemic originated via zoonosis, whereby SARS-CoV-2 or a closely related progenitor virus was transmitted from bats to an intermediary animal that subsequently infected one or more people and then spread within the human population.
>
> H_L = The covid-19 pandemic originated from laboratory research. Staff at a research institute became infected while conducting genetic engineering for gain of function research, and subsequently transmitted the virus into the human population.
>
> *E* = Dr. Shi Zhengli, a lead WIV virologist at the center of lab-leak theories, conveyed the following to *Scientific American*: "If coronaviruses were the culprit, she remembers thinking, 'Could they have come from our lab?' ... Shi frantically went through her lab's records from the past few years to check for any mishandling of experimental materials... Shi breathed a sigh of relief when the results came back: none of the sequences matched those of the viruses her team had sampled from bat caves. 'That really took a load off my mind. I had not slept a wink for days.'" *(Qiu 2020)*

These hypotheses cannot be easily reduced to mathematical models. And the evidence cannot naturally be quantified. Here one could collapse the evidence into a binary clue, where the possible values are either "yes" or "no" when Shi Zhengli is asked whether her lab had the virus. But then we would lose all of the nuanced details of what she said that matter for evaluating the relative likelihood of the evidence.

But we can quantify our degrees of belief, if we work with the log-odds form of Bayes' rule: *posterior log-odds = prior log-odds + weight of evidence*, where the weight of evidence is proportional to the logarithm of the likelihood ratio. We recommend using decibels (dB), which is a familiar log scale (Fairfield & Charman, Chap. 4). Then when we assess the weight of evidence, we can use an analogy to sound, where we ask how loudly the evidence is talking—e.g., does it whisper, or does it shout in favor of one hypothesis over the other? This approach also helps leverage intuition, because as per the Weber-Fechner law, perception works on a log scale, not a linear scale.

Working with log-odds and quantifying in decibels helps us express our probability assessments more precisely. That helps pinpoint sources of disagreement between scholars more effectively. It also helps us aggregate the net inferential weight more carefully when the evidence does not all stack up neatly in favor of the same hypothesis. There is one important caveat: probabilities cannot be unambiguously specified in many social science contexts. Especially (but not exclusively) in qualitative research, there will always be some arbitrariness and subjectivity that creeps in when we are working with complex, real-world evidence. Our approach promotes more systematic and transparent reasoning, but it cannot eliminate all subjectivity or automatically produce consensus among scholars.

Example of Bayesian reasoning

For the covid example above, we have testimonial evidence, where we need to reason about the likelihood that Dr. Shi would give the reporter this account of events under each of

the rival hypotheses. I will invoke as background information that on the one hand, Dr. Shi is highly respected among her international colleagues, while on the other hand, she lives under and effectively works for the highly authoritarian Chinese government.

In the world of H_Z, the lab is not the source of the outbreak, and the most likely scenario is that Shi is telling the truth when reporting that the WIV did not have a SARS-CoV-2 progenitor virus. This kind of response would be expected. Shi is known as a responsible and respected scientist, she collects and studies bat viruses, and the outbreak occurred in very city where her lab is located. She would check her records just to make sure, and she would have every incentive to report the findings that clear up her concerns and absolve the lab.

Under H_L, we would reason that Shi is lying—it seems implausible for her to make an honest mistake on this matter in a world where her lab was conducting genetic engineering on a SARS-CoV-2 progenitor virus. This particular lie is a good cover story, because it mimics how one would expect her to behave in the H_Z world. But it seems at least a bit less likely for Shi to make this kind of statement under H_L. Assuming that Shi would be under great pressure from the government to cover up a lab leak, we would expect a less detailed, more adamant denial of responsibility that definitively dismisses the possibility that the virus came from her lab—something like: "I was highly confident in our safety protocols, but out of an abundance of caution, I double checked and found nothing."

Accordingly, *E* weakly favors H_Z over H_L. But the inferential weight depends on one's background information. The more confidence one has in Shi's probity and the more room to speak honestly one believes she would have regardless of what the truth entails, the more *E* favors zoonosis, whereas the greater the incentives one thinks she would have to defend her reputation even if the virus leaked from her lab and the more serious the consequences one believes the government would impose on her for disclosing a lab leak, the less informative the evidence becomes. We cautiously attribute only very weak weight to the evidence, about 4dB, assuming it would be very costly for Shi to disclose the truth in a lab leak world.

Applications of information theory

Our book includes a few applications of information theory for understanding test strength and guiding case selection (Fairfield & Charman, Chaps. 11, 12), which are big issues in political science that have been subject to a lot of debate and confusion. (Case selection can be thought of as a qualitative analog of experimental design.) Before gathering data, we might want a measure of anticipated test strength that takes into account our uncertainty about what evidence we will find. Relative entropy does that by averaging the weights of evidence (W) over the possible evidentiary outcomes, weighted by their respective likelihoods. For binary evidence that can take one of two possible values, K or ~K we have:

$$D(H_1;H_0) = P(K \mid H_1\mathcal{I})\, W_K + P(\overline{K} \mid H_1\mathcal{I})\, W_{\overline{K}}$$

$$D(H_0;H_1) = -P(K \mid H_0\mathcal{I})\, W_K - P(\overline{K} \mid H_0\mathcal{I})\, W_{\overline{K}}$$

Expected information gain in turn averages the relative entropies over our uncertainty about which H is true, which gives a single measure of anticipated test strength:

$$\overline{D} = P(H_1 \mid \mathcal{I})\; D(H_1;H_0) + P(H_0 \mid \mathcal{I})\; D(H_0;H_1)$$

In essence, it tells us how loudly we expect whatever evidence turns up to speak in favor of the best hypothesis.

For inference to best explanation, the value of information depends on the inferential approach. For Bayesian inference to best explanation (the optimal framework), the value of information is best defined as the weight of evidence, which tells us how much we have learned about which hypothesis is more plausible. The value of information would be identical to the quality of evidence: high quality evidence provides a large weight of evidence. Prospectively, the value of information that we plan to collect would be expected information gain, which takes the more general form:

$$\overline{\mathcal{D}}(S_{\mathrm{C}}, \mathcal{I}) = \sum_j P(H_j \,|\, \mathcal{I}) \sum_k P(E_k \,|\, H_j \, S_{\mathrm{C}} \, \mathcal{I}) \, \log\Big[\frac{P(E_k \,|\, H_j \, S_{\mathrm{C}} \, \mathcal{I})}{P(E_k \,|\, \overline{H_j} \, S_{\mathrm{C}} \, \mathcal{I})}\Big]$$

The value of information is accordingly relative to the inferential problem: it is conditional on the hypotheses under consideration; it is conditional on our background information; prospectively, it is also conditional on the details of how and where we plan to look for the information (S_c). The weight of evidence and expected information gain are ideally or approximately objective, in that rational thinkers with identical background information should assign the same values (or hopefully similar values in practice).

In contrast, the value of information for solving problems is inherently subjective. It depends on the uses and implications of knowledge. And it depends on the decision maker's utility function, which may vary dramatically for different actors. For the covid origins question, relevant actors have motivations that are very much at odds—to the extent that some would like to continue investigating whereas others would prefer not to learn anything more at all. Another important point here is that making decisions should be considered logically distinct and subsequent to inference—the optimal framework is Bayesian decision theory (as discussed by A.E. Charman).

Thoughts on moving forward

I am dubious about quantifying the full potential value of data, considering that we cannot know or even begin to evaluate the uses to which data might be put in the future, the (as yet unknown) hypotheses that might tested, the lines of inquiry that might be pursued, or the implications of the knowledge generated. As such, I would propose finding some more pragmatic considerations for deciding what data to collect and justifying its potential (albeit unknowable) value. For the USDA, one route might involve inviting data-collection proposals from academics and users more broadly, to find out what range of projects and questions might draw on data that falls under the agency's purview. One might in essence solicit something akin to a grant proposal, where authors justify the inferential value of the data they are interested in and elaborate the potential relevance of their findings for public policy or other activities that the agency or its superiors would like to foster. Data that could be used for multiple project proposals or for projects deemed to be more important or more relevant could be considered higher priority, and the proposals themselves could be used to justify the agency's decisions about what data to collect. Presumably the goal would be to collect data that is not otherwise available through other databases or public agencies.

Using thermodynamics and information theory for interpreting molecular tumor data and predicting cellular responses and treatments

Nataly Kravchenko-Balasha, HUJI, Israel.

Following the workshop sessions, I present my summary and conclusions on my understanding of the "value of information" and its application to biology and cancer research.

The value of information can be described as the amount of knowledge a scientist obtains at the end of the analytical method. This knowledge is a "power" that enables scientists to alter the examined system, possibly by changing its course. The dynamics and recent development of information theory toward many scientific fields and interpretations enables for its ongoing progress and expansion into new applications.

I'll give an example of how to apply information theory and determine the value of information in the field of personalized medicine.

I propose computing free energy changes in each cancer tissue to translate biological experimental information into knowledge. We calculate free energy change by integrating the patient-specific constraints that lead the tissue to deviate from its steady state[1].

Why should results from clinical or biological tests be understood in terms of free energy or thermodynamics? Because this informs us whether our system is stable or unstable. If a system is unstable, then, as in chemistry and physics, a spontaneous shift from higher to lower free energy states, known as steady state, will always occur when the system is unconstrained. To induce a transition towards steady state in unstable systems, we must first identify and understand the constraints before determining how to remove them. In our latest study, which included over 800 leukemia patients, we found that diseased states are often less stable[1]. Thus, to restore a tissue's steady state, we proposed targeting central proteins involved in each constraint that cause the system to shift from its steady state.

We experimentally demonstrated that the "amount of knowledge" we obtained at the end of the thermodynamic-based approach, namely the computation of patient-specific constraints reflecting ongoing, tumor-specific molecular processes, allowed us to effectively construct tailored therapies that focused on patient-specific constraints[2,3]. Thus, in the case of precision medicine, the value of information can be evaluated by our capacity to stop/ reduce tumor growth.

Ongoing research in our lab aims to use the knowledge embedded in patient-specific constraints to modify tumor tissue in a number of additional parameters, such as inhibiting tumor cell spreading, stimulating the immune system, or preventing cancer cells from communicating with the non-cancer environment.

References

1 Uechi L, Vasudevan S, Vilenski D, Branciamore S, Frankhouser D, O'Meally D *et al.* Transcriptome free energy can serve as a dynamic patient-specific biomarker in acute myeloid leukemia. *npj Syst Biol Appl 2024 101* 2024; **10**: 1–10.

2 Alkhatib H, Conage-Pough J, Roy Chowdhury S, Shian D, Zaid D, Rubinstein AM *et al.* Patient-specific signaling signatures predict optimal therapeutic combinations for triple negative breast cancer. *Mol Cancer* 2024; **23**: 1–7.

3 Vasudevan S, Flashner-Abramson E, Alkhatib H, Roy Chowdhury S, Adejumobi IA, Vilenski D *et al.* Overcoming resistance to BRAFV600E inhibition in melanoma by deciphering and targeting personalized protein network alterations. *npj Precis Oncol* 2021; **5**. doi:10.1038/s41698-021-00190-3.

The Value of Data and Information: the role of uncertainty and view perspective

Rossella Bernardini Papalia, Department of Statistical Sciences, University of Bologna, Italy

The value of data and information is closely connected to the concept of uncertainty of data and information. Data is information with many types, forms, origins, and content. Each form of data may be characterized by different quality dimensions as: reproducibility and replication, volume, velocity, variety, veracity. However, information is not a statistical data, it is a form of substantial knowledge when processed and validated. The statistical data arises from a set of conditions: a cognitive objective, a system of hypotheses, a definition and classification criteria, a strategy and a controlled measurement, a validation process that allow reliability to be measured through clear, objective, and understandable indicators. A-critical use of information can generate of which cognitive significance may not be clear or even misleading or incorrect. The Value of the data and information is strictly connected with their use, and it depends on the view perspective, so it is necessary to distinguish between the producer or user view. For a producer of statistical data, the actual challenges for evaluating the full potential of available sources of data, datasets and related information strictly depend on the integration and linking of registers and the new database stemming from IT or AI based measurements. Understanding uncertainties related to these new forms of data is very important and crucial. Evaluating data/information uncertainty is then necessary to reveal the value of the information and its nexus with: (i) the potential complex system that generated information; (ii) the data generation process when data is analyzed; (iii) the meaning and statistical properties of the data generated by the AI measurement tools. By processing information: uncertainty can be reduced at both input and output levels. At a data Input level data quality can be measured in accordance with known dimensions while, at an output data level, the quality of results can be improved in terms of predictive power and estimate accuracy with reference to the specified relationship and related statistical models. Since Shannon's Entropy represents a lower-bound on the number of actual bits required to store or transmit information, the entropy can show the real amount of information that a field is providing in each database. The maximum Entropy can show the potential amount of information that a field or a set of data can provide (some available fields potentially could not provide any information). Both are useful to understand the nature of the dataset in terms of information messages stemming from different sited or sources and may represent suitable data quality measures at an input level. From a user's perspective that operates at an output level, more focus on extracting the information content of data, it could be useful to adopt Information Theoretic based methods less affected by the constraints of sampling and or non-sampling errors in data with the advantages of evaluating the informative power of constraints and/or theoretic basic assumptions. Finally, some specific considerations are required to overcome the challenges in order to evaluate the full potential of data/information when new IT/AI based tools are used: experimental work is needed to interpret the information content of this the new AI form of data. The working group meeting provides an opportunity to analyze critical points from different perspectives often connected to case studies and theories. Qualitative and theory-based aspects have been shown as major drivers of information value.

Notes on Generic Value of Information

Wojciech Szpankowski

1 Generic Definition

Let us assume that there is unknown *hidden* random variable $Y \in \mathcal{Y}$ upon which one makes a decision (e.g., in medical diagnostics). In order to learn Y we perform a series of tests $t \in \mathcal{T} = \{1, \ldots, n\}$. We shall assume that we can observe $x_{\mathcal{T}} = \{x_t\}_{t \in \mathcal{T}}$, however, we also may have access only to partial observables denoted as $x_{\mathcal{A}} = \{x\}_{t \in \mathcal{A} \subset \mathcal{T}}$. The cost of a test t is denoted as $c(t)$.

Unfortunately, Y is usually unknown, so a test t reveals some outcome $X_t \in \mathcal{X}$ that is correlated to Y. We would like to estimate the posterior probability $P(Y|X)$.

Finally, we have set of decisions $\mathcal{D}$ to choose from based on the outcome X. In principle, after performing a set of tests and observing X about hidden Y, we make a decision $d \in \mathcal{D}$. To quantify its benefit we introduce a utility function $u(y, d) : \mathcal{Y} \times \mathcal{D} \to \mathbf{R}_{\geq 0}$.

Now we are ready to state a general definition of *value of information*.

Definition 1 *Upon observing $x_{\mathcal{T}}$ (or partial observation $x_{\mathcal{A}}$) we define the value of information (VoI) as*

$$\mathrm{VoI}(x_{\mathcal{T}}) := \max_{d \in \mathcal{D}} \mathbf{E}_Y[u(Y, d)|x_{\mathcal{T}}]$$

where the expectation is with respect to Y. We can replace in the above $x_{\mathcal{T}}$ by $x_{\mathcal{A}}$, and then $\mathrm{VoI}(u, x_{\mathcal{A}})$ *also depends on $x_{\mathcal{A}}$.*

The above definition is in the spirit of Howard [2] and was inspired by [1]. One can ask, however, whether we can formulate an axiomatic definition of VoI based on [3].

2 Optimal Strategy

Following [1] we can define an interesting optimiztion problem.

Define first *regret* as

$$R(d|x_{\mathcal{A}}) := \max_{x_{\mathcal{T}}: P(x_{\mathcal{T}}|x_{\mathcal{A}}) > 0} [\mathrm{VoI}(x_{\mathcal{T}}) - \mathbf{E}_Y[u(Y, d)|x_{\mathcal{T}}]].$$

Our goal is to find the best policy π that minimizes the cost with regret not exceeding some ε. More formally, let us denote by $\mathcal{S}(\pi, x_{\mathcal{T}}) \subset \mathcal{T} \times \mathcal{X}$ a set of observations by running policy π. Then, we can define our optimization problem as

$$\pi^* \in \arg_\pi \min c(\pi) \;\; \text{s.t.} \;\; \forall_{x_{\mathcal{T}}} \exists_d : \;\; R(d|\mathcal{S}(\pi, x_{\mathcal{T}}) \leq \varepsilon$$

for some small ε. This formulation is solved in [1] using submodular functions.

References

[1] Y. Chen, S. Javdani, A. Karbasi, J. Bagnell, S. Srinivasa, and A. Krause, Submodular Surrogates for Value of Information, *Proc, Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3511a3518, 2015.

[2] R. Howard, Information value theory, *IEEE Transactions on Systems Science and Cybernetics*, 2(1): 22-26, 1966.

[3] B. Steudel, D. Janzing, and B. Scholkopf, Causal Markov Condition for Submodular Information Measures, rerXiv:1002.4020, 2010.

Brief summary of my experience at the Value of Information working group meeting
Gideon Yaffe
4/9/24

I found the Value of Information working group meeting in April 2024 to be very helpful. Much of my current research concerns normative questions about the law of evidence. By "normative questions" I mean questions about how the law of evidence *should* be. Since sometimes how the law should be is how it is already, my research sometimes concerns crafting convincing arguments in favor of particular laws of evidence as they are currently formulated.

One research problem with which I have been concerned recently involves explaining and rationalizing a particular entrenched practice on the part of legal professionals tasked with applying the law of evidence. The practice of interests concerns the inadmissibility of many forms of so called "statistical evidence"—evidence about the statistical properties of sets of data which one side or another wants to offer to fact-finders to allow them to make inferences about the properties or behaviors, not of groups, but of particular individuals that are relevant to the legal decision at hand. Information, for instance, about the percentage of individuals who have a particular set of characteristics that the perpetrator of the crime possessed, for instance, is often inadmissible—attorneys on both sides are barred from presenting such information to the court. If, for instance, the defendant is charged with entering the baseball stadium without a ticket, the prosecution is barred from explaining to the jury that while 100,000 people attended the event, only 1000 bought tickets, which implies that the probability is 99% that any person who attended the event (which the defendant admits having done) did not purchase a ticket. However, very similar evidence is sometimes admissible; attorneys are not barred from showing DNA evidence—evidence concerning the percentage of people who have DNA that matches the perpetrator of the crime. The challenge for a researcher like myself is to explain what the difference is, if any, between statistical evidence that ought not to be presented to the fact-finder and statistical evidence that can be. Although there have been efforts made to solve this problem, they have not been successful; the challenge remains unmet.

I think it is fair to say that the most important thing that I learned at the conference was that this problem—along with others of the sort with which I have been concerned—might be tackled with help from information theory. Systematic and rigorous study of the law of evidence is in many ways in its infancy, and information theory has simply never been used for this purpose. A couple of times in my career I have encountered tools where there are sociological, rather than substantive, reasons why they have not been used to solve the problems of greatest interest to me. Neuroscience, for instance, has been used less often to tackle question about criminal responsibility, or how mental illness bears on it, than it could be. This is largely because doing so requires both neuroscientific and legal expertise, and people very rarely have both. I suspect that the reason information theory has not been used to query the law of evidence is for similar reasons. I am hoping to remedy this by taking steps to master information theory, and then use what I learn to answer fundamental normative questions about the law of evidence. It was the experience at the working group meeting that convinced me to take this step.

Information and Data in veridical data science

Bin Yu
Statistics, EECS, Comp. Bio., UC Berkeley

According to the New Oxford American Dictionary, information is "facts provided or learned about something or someone". Information about facts in my data science world is obtained through a process called data science life cycle (DSLC) to answer domain questions (whose answers are the facts to seek) such as:
What genes drive a heart disease (HCM)? Whether a patient has prostate cancer?
How to predict Madden-Julian-Oscillation (MJO)?

DSLC is not deterministic hence information about facts comes with uncertainty. Reliable uncertainty quantification is key to trustworthy information about facts. DSLC starts with domain question(s) and proceeds with data collection or access, data cleaning and curation, exploratory data analysis, model or algorithm development, validation, and communication of data-driven results in the context of domain question(s). Every step of a DSLC is a source of uncertainty due to data collection process and human judgment calls. The Predictability-Computability-Stability (PCS) framework and documentation have been introduced for veridical (truthful) data science ([1,2,3,4]) to synthesize, unify, streamline, and expand on ideas and best practices in both ML and Stats to arrive at scientifically reproducible results. Specifically, PCS considers sources of uncertainty from data cleaning and model choices in addition to the traditional statistical uncertainty from sample-to-sample variability under a well vetted probabilistic data generation model. The uncertainty from data cleaning is well-known among practitioners who clean data themselves, and the uncertainty from algorithm or model choices by different teams of data scientists have gotten attention lately (e.g. [5]). It is high time to formally address these additional sources of uncertainty and PCS is developed exactly for this purpose through "S" and to enhance the requirement for reality check through "P" in every step of a DSLC. "C" is a necessity for both "P" and "S". PCS has enjoyed successes in many research areas including cancer prediction and seeking genetic drivers for HCM. However, addressing "C" for "S" is a frontier research topic for big data AI research including LLM developments and for entropy or mutual information estimation via multiple methods and possibly based on differently cleaned data sets. PCS is a practical philosophy and a research program for trustworthy information extraction from data. It is evolving as more and more researchers use it to meet challenges in their own research fields.

[1] B. Yu (2013). Stability. Bernoulli, 19 (4), 1484-1500.

[2] B. Yu and K. Kumbier (2020). Veridical data science. PNAS, 117 (8), 3920-3929.

[3] B. Yu (2023). What is uncertainty in today's practice of data science? Journal of Econometrics. 237. 105519.

[4] B. Yu and R. Barter (2024). "Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making", MIT Press (forthcoming). On-line free version: vdsbook.com

[5] N. Breznau et al (2023) Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. PNAS. 119 (44) e2203150119

Spiro Stefanou
Administrator, USDA/Economic Research Service

The Economic Research Service (ERS) is one of the 13 Federal Statistical Agencies comprising the US Federal Statistical System. We have a $90 million budget with full discretion on how it is allocated. As a Federal Statistical Agency, we are a data factory as well as a research and analysis agency.

Data play an important role in maintaining and advancing a civil society. Data and facts provide evidence to policy making. A stakeholder will take facts and align them with their interests and then overlay their values to advance policy. A competing policy should be based on the same facts aligned with different interests and overlaying different values. And then the policy debate ensues. Quality data will lead to good statistics that can lead to good policy. Poor data will lead to bad statistics, that will lead to bad policy.

Data are foundational to our work and constitute approximately 20% of the budget for both in-house surveys and proprietary data purchases. This begs the question: What is the value of publicly provided data and information? We have engaged in the background work to start documenting the corpus of literature using our data assets using machine learning algorithms, so can assess how are data products are informing the literature. The next steps are to assess how our data are feeding to policy documents at the Congressional level, think tanks and related policy paper series. ERS hosts over 80 publicly available data products and 8 confidential data products.

ERS has a broad mission to anticipate trends and emerging issues in agriculture, food, the environment, and rural America and conducts high-quality, objective economic research to inform and enhance public and private decision making. We accomplish this accomplish this by producing timely economic statistics, building, and strengthening research data on agriculture and rural America, and providing the public with trusted sources of information and secure means of data dissemination.

We are organized into three research divisions, all heavily dependent on public and proprietary data products. The Food Economics Division conducts economic research and analysis on policy-relevant issues related to the food sector (food safety, food prices, and markets); consumer behavior related to food choices (food consumption, diet quality, and nutrition); and food and nutrition assistance programs. This division also provides data and statistics on food prices, food expenditures, and the food supply chain.

The Market and Trade Economics Division monitors, evaluates, and conducts research on domestic and foreign economic and policy factors affecting agricultural markets and trade. Research focuses on policy and program alternatives, domestic and international markets, commodity analysis and forecasts, international food security, and development of analytical tools and data.

The Resources and Rural Economics Division studies linkages between agricultural, energy, climate, and environmental policies; ecosystem services and land use; research and development of agricultural technologies and agricultural productivity; dynamics of farming; rural development; and the well-being of farm and rural households. This division also collaborates with USDA's National Agricultural Statistics Service to plan and implement an annual national survey of farm enterprises and farm households.

Brief Summaries (Q2 and Q3) of the Breakout Groups

Basic Questions for the Groups:

1. Can we develop concepts of potential value based on meaning and semantics?
2. Are there fundamental differences in the informational content of information generated from a complex system/process and one that is a result of a simple process?
3. What is the relationship between value of a model and the value of Information used to construct that model?

Q2: Are there fundamental differences in the informational content of information generated from a complex system/process and one that is a result of a simple process?

Radu Balan (moderator, note taker)

In information is Shannon's information, then more complex systems, that have a larger n (number of states/outcomes) then the larger n the larger the entropy hence information.

Several points of view were discussed.

A.

If we deal with data compression, a memoryless system,

If the system is a sum of collection of subsystems,

simple system: x->x' , y->y'

a complex system: (x(t+1),y(t+1)) = F(x(t),y(t)) where transitions can happen between x->x' or x->y' and y->x' and y->y' with certain probabilities.

B.

Biophotons can be produced by plants, or simple systems, or humans, or more complex systems.

Measurements would be able to differentiate the steady-state radiation, and distinguish between these types of radiation..

C.

A probabilistic modeling of the distinction between simple and complex system:

Complex system (C) has a larger observation vector (x,y) drawn from:

(x,y) ~ p(x,y;a,b)

A simple system/process (S) has instead a smaller observation vector distributed according to

x ~ q(x;a)

If y is not measured/known or y is missing, then how to deal with this problem?

Options to consider are:

1) Need to consider the marginal distribution:

(C') : x ~ r(x;a,b) = int p(x,y;a,b) dy (marginal over y)

2) Use the nuisance parameter approach, where y is estimated using the maximum likelihood approach:

(C'') : x ~ s(x;a,b) , where s(x;a,b) = p(x,yhat;a,b) , yhat(x) = argmax_y p(x,y;a,b)

Then:

Information extracted from one of models (C'), (C'') is to be compared to information extracted from S.

D.

Model Selection/decision between a simpler or a more complex system is performed

by optimization of an objective function that performs a trade-off of two components: (1) how well the model explains the data x; (2) complexity of the model.

AIC, BIC, MDL are just variations of the penalty term, but, fundamentally, any and all these approaches are the same. Differences are the degree of belief.

Conclusion: there are no fundamental differences between simpler/more complex system from information point of view. There are only differences in terms of degree of computational complexity.

3. What is the relationship between value of a model and the value of information used to construct that model?

John Harte, Tasha Fairfield, Spiro Stefanou. Luis Bettencourt (moderator, note taker)

We discussed many things, largely using USDA examples on data collections and goals.

We agreed that – based on previous working group discussions – the value of a model and the value of information (data) used to construct it are not independent. Each definition of value needs the other: the model needs the data, and the data needs the model in order to acquire value.

That said, sometimes one starts with data collections, and some other times with a model. Subsequently the two must iterate to realize (and even potentially realize) value.

It is possible that once a model or theory are codified from data, that the original data cease to be essential and in that sense loose (some of) their value. The theory should be able to both generate new data, or assess the consistency of new data with its predictions. This can be seen clearly in a Bayesian estimation scheme, where P(D|H) is used for inference (as a likelihood), but can also be used later as a posterior, a generative model to produce data.

We feel that an important concept that came from the discussions is the idea of a minimum (expected) value for the data. This is usually defined by the most proximate motivations to collect the data; without such motivation the data will not be collected. However, this is a minimum value in the sense we discussed that many other uses and users, including in the context additional theories are possible and even likely in the future. This should add motivation to data collections but is hard to calculate.

A few additional threads in the conversation are reported loosely below:

Spiro: Acquiring data at a public scale is a massive undertaking, you should have a model for when and how to do it.

An example of a pilot at the USDA on snap program (food stamps): data collection of food acquired and consumed … did a pilot, analyzed by the census, they concluded that the data collections via cell phones were inappropriate as they excluded many sectors of the population… the drive to collect location of food acquisition and consumption was driven by the question of inaccessibility of nutritious food.

Do federal agencies create hypothesis and questions before they collect data? Spiro though so, but we were not sure that there was a formal way in which this is incorporated into their process.

Tasha: What are we trying to learn? What are we trying to explain? What evidence to collect?

John : There is information to construct the model versus information to test the model.