SANTA FE INSTITUTE

Santa Fe Institute Workshop and Working Group Summary Report

Meeting Title: The Potential Value of Information and Data

In-Person Working Group (with one participant via zoom due to injury)

Dates: April 2 – April 3, 2024

Organizers & Affiliations:
Amos Golan (American U.; SFI)
John Harte (UC Berkeley; SFI)
Min Chen (Oxford; Pembroke College)

I. Please provide the following content:

A. Required: A narrative summary of the meeting (1-3 pages)

This can be partially based on the approved SFI meeting proposal but should reflect how the meeting went. Include any notable feedback you received from participants and please also note what worked well and what could have been improved. The summary can include a short description of outcomes that might emerge from the meeting [e.g., new perspectives/questions/theory, new methods/algorithms/data, new projects/collaborations, future meetings, manuscripts, grant proposals, etc.]. Feel free to assign the summary report to a participant – for example, it can be a good exercise for an early career colleague, but should be vetted by the organizers.

Basic Summary. Information and data emerging from complex systems are transformed, via different human- and machine-centric processes (e.g., theory-based analysis, analytical models, machine learned models, etc.), to knowledge about these systems. But can we evaluate the potential (relative) value of these data and information, and its interrelationship with the modeling and inferential approaches used to transform the information into knowledge? In this working group we studied these issues within the context of information theory and information-theoretic inference.

Overall, this WG deals with a few interdependent questions – all to do with our understanding of the information and data in complex systems and the potential value of that information and data. *The first* question is whether information theory can be used for developing new tools for evaluating the full potential, in a quantifiable manner, of datasets and the potential value (such as full theoretical potential or estimated attainable potential) of datasets, as well as the information stored in these data? *The second* basic question is how to extend (if possible) information theory to account for the meaning of the information embedded in the data? A fundamental issue that arises in both cases is to do with the relationship between the potential value measures and the inferential procedure used to transform data/information to knowledge. *The third* question is then, should the value be independent of the inferential/modeling approach used? *The fourth* question is how can we measure the value (or potential value) of a model? One can hypothesize that it is conditional on the information needed for constructing and developing that model, but it is an open question that demands more thinking.

Though value is always a relative concept, interest in the philosophy, including meaning and value of information as well as other aspects in the philosophy of information, goes back half a

century but has rapidly increased recently with the availability of more complex data, as well as many new directions of research into the meaning, measures, and quantification of incomplete, large, blurry and complex data. Theoretical advances in these directions will have a substantial impact on a wide range of real-world applications. Methodological and technical advances will assist policy and decision makers and will have a direct impact on private and public agencies that produce data for public and private use and research. For scientists, the formulation of such a (relative) potential value will provide an additional tool to evaluate the information used – and needed – in modeling and inference. (A more detailed summary and broader context is provided in Appendix A.)

Structure and Outcomes.

Structure: The basic structure was a (i) plenary group discussion, each period devoted to a subset of the questions, including short presentations and Q&A with intermittent joint discussions on the different questions and the interrelationships among these questions, and (ii) a one-time breakout session where each group discussed one of the questions. Details on the breakout session are below. The complete agenda appears on page 4 of this document.

The three groups in the breakout session discussed these questions:

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?

Summaries of the group discussions are provided in the 'Value SFI WG 4 2024 Participants Summaries.pdf' document.

Outcomes*:*

Part 1. As one would expect from such a fundamental discussion, there is no single outcome that the group agreed on. Such an outcome may emerge in the future, but not yet. However, we can think of a 'softer' definition of outcome we referred to as 'implicit outcome' which is expressed as the approximate intersection of definitions of most participants. That outcome relates potential value of data to models and theories, and to prediction and testing. We express it as follows: 'The value of a data is determined by the model (likelihood) or theory that uses (or may use) the data. By 'use' we mean all types of uses, such as initializing a model using data, extracting key indicators or features from data, making prediction based on data, testing, and validating the components of a transformation (e.g., theories, algorithms, and models) using data, and evaluating decision making and derived policies using data.

Under that loose argument, it means that in at times, possibly more emphasis should be placed on constructing models and theories, as compared with simply gathering more and more data in the hope it will be useful. However, at other times, many models are dated quickly. Covid-19 and cybersecurity models are typical examples.

Lacking a satisfactory model or theory, we can express our objectives – the set of questions (or policies) we want to solve and understand – and attempt to use them as a guide for gathering additional data. A new dataset, accompanied by all the needed metadata (documentation of definitions, procedures used to gather the data, etc.), can also lead to formulation of new

1399 Hyde Park Road | Santa Fe, New Mexico 87501 | www.santafe.edu | 505.944.8800

questions and guide the development of new theory, independent of inference procedures, thereby freeing the potential value of the data from the model. Ultimately, however, satisfactory models/theories are needed to obtain confidence in our predictions and policy recommendations.

In all approaches described above we may be surprised by finding new and unexpected knowledge hidden in the data. Therefore, the above two arguments provide a minimum bound on the potential value of data. Last, in all cases, we agreed in the WG discussion, that the quality of the data can be measured (at least relatively).

With the above in mind, we must acknowledge that "… truly complex systems, such as ecosystems, economies, and climate, appear especially refractory to analysis by merely processing big datasets with these advanced tools and new mathematical discoveries." (Golan and Harte, PNAS 2022).

Part 2. Overall, the discussion, especially during the second day, brought up a number of new ideas that emerged from the groups' discussions. Some ideas (outcomes) were more surprising than others. Some were more theoretical while others were more practical. As an example, there was (almost) an agreement that the value of a model and the value of data used to create the model and to investigate that model, must be connected (model and its related information are not independent). Two examples of questions where the opinions did not converge include the role of semantic (and meaning) in evaluating the potential value of information/data and whether information theory by itself can do the job.

Part 3. We plan to follow-up on that WG in Fall, 2024 during an Info-Metrics Workshop at American University.

Part 4. A one-page subjective summary by each participant. That summary includes the participant's view on the potential value of information and data, as well as thoughts about the WG discussions. Some of these ideas were expressed above.

We will assemble it and put is in the WG webpage at American University: https://www.american.edu/cas/economics/info-metrics/workshop/working-group-potential-value.cfm

Additional Comments

Overall, our two-day WG plan worked fine. The interdisciplinary aspect was essential, and participants were all involved. However, a number of things can be improved. First, a full third day would have contributed significantly (but it is hard to design, especially for a relatively large group). It would be the day when some 'things' may have fallen into place and some research ideas may have converged. Second, a stronger emphasis on the material presented by the participants on the first day (though it would be hard to impose), especially on the connection to the working group's objectives. Third, possibly devoting more time in the beginning of the first day to discussing the basic questions (and possibly to let each one express her/his view on that for about 2-3 minutes).

To sum up, it was very productive. Many ideas and new directions were proposed. Of course, much more is needed, but that WG helped in planting the seeds for further developments.

B. Optional: Short answers to the following questions:

1. *What was the big question or idea the meeting was designed to explore?*

Understanding of the information and data in complex systems and the potential value of that information and the inferential models used for evaluating and studying this information.

2. *What was the goal of the meeting?*

To have an inherently interdisciplinary discussion and flows of ideas among mathematical abstraction, philosophical discourse, and practical consideration about the above questions.

3. What was the single most important outcome of the meeting?

First, though it was expected, defining, modeling, and estimating the potential value of data (which most researchers agree is relative) is a complex problem, and incorporating the semantic and meaning within that value (function) needs much more work. Second, developing the potential value of a model needs much more work as well.

4. What disciplines/fields were represented at the meeting?

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.

5. What one or two new research direction(s) did the meeting suggest?

6. What was the most interesting thing said during the meeting, and who said it?

A number of interesting ideas came up, but we cannot think of 'most interesting.'

7. What idea from the meeting is likely to be most impactful for science?

Combining 'meaning and semantics' within information theory. Possibly also new ways of evaluating the potential value of a model.

8. What idea from the meeting is potentially translatable into an application—i.e., something useful to businesses, policy makers, government agencies, or other non-academic audiences?

The practical, and relative, potential value of data as is described in Golan's technical report (prior to the WG meeting) probably captures the most applicable (current) approach for evaluating the potential value (that relies on IT to quantify different aspects of the data quality), but it is far from perfect. Chen's discussion of the cost-benefit idea is also applicable, but unlike the potential value of data, it probably captures the potential value of the inferential model/s and approach/es used to convert the data into knowledge and decisions. Wojtek formulation also provides a special case for evaluating decisions based solely on information theory, but it does not take the direct meaning into account, and it is based on current (or past decisions). Other ideas of value came up but like all the other ideas discussed at the WG, each one had some missing parts – they only evaluated partial potential value. The only axiomatic approach for deriving the value function was discussed by Balan (based on a joint work with Golan). Other different approaches, even if not discussed in detail during the WG, appear in the participants' brief summaries.

9. What new method/technology/algorithm/etc. resulted (or may result) from this meeting?

Nothing yet.

II. SFI Events staff will append four additional items, as applicable, to your summary:

1. The final participant list including affiliations.

Co-Organizers:

Amos Golan, Economics and Info-Metrics, American University; External faculty SFI
John Harte, Department of Environmental Science, Policy, & Management UC Berkeley; External faculty SFI
Min Chen, Professor of Scientific Visualization, Department of Engineering Science, U Oxford (Fellow, Pembroke College)

Participants:

Radu Balan, Professor of Applied Mathematics, department of Mathematics and Center for Scientific Computation and Mathematical Modeling, U Maryland, College Park)
Luis Bettencourt, Professor of Ecology and Evolution, U Chicago; External faculty SFI
Ariel Caticha, Department of Physics, University at Albany, State University of New York
Andrew Charman, Department of Physics, UC Berkeley
Tasha Fairfield, Development Studies, Department of International Development, London School of Economics, UK
Nataly Kravchenko-Balasha, The Faculty of Dental Medicine, Hebrew University - Hadassah Medical Campus, Faculty of Dental Medicine, Hebrew University of Jerusalem, Israel **Rossella Bernardini Papalia**, Department of Statistical Sciences, University of Bologna, Italy
Spiro Stefanou, Administrator, Economic Research Service, USDA
Wojciech Szpankowski (Wojtek), Saul Rosen Professor of Computer Science Computer Science, Director of Center for the Science of Information, Purdue
Trina Weilert, Supervisory Geographer, Branch Chief, Economic Research Service, USDA
Gideon Yaffe, Wesley Newcomb Hohfeld Professor of Jurisprudence, Professor of Philosophy & Psychology Yale Law School, Yale
Bin Yu, Chancellor's Distinguished Professor and Class of 1936 Second Chair, Departments of Statistics and Electrical Engineering and Computer Sciences, UC Berkeley

2. The meeting agenda, if applicable (if you had an agenda that was not provided to event staff, please include that with your summary).

The Potential Value of Information and Data

April 2-3, 2024

Agenda

Day 1: Tuesday April 2, 2024

8:15 Hotel Santa Fe Shuttle departs hotel to Santa Fe Institute

8:30 – 9:00 Breakfast at Santa Fe Institute

9:00AM – 9:20AM

Opening Session: Information and Potential Information

Amos Golan, John Harte, Min Chen

9:20AM – 12:00PM (with 30 minutes break in the middle).

Morning: Individual Presentations (At most 10 Minutes/person).

The Presentation should be an 'introduction' and summary of the individual's research with an emphasis of how it is related to the Working Group main focus. Please use the second part of your presentation to provide your thoughts on some (or all) of the questions posed in the 'Working Group Objectives' and the way you will go about attacking it.

(Power Point is fine, but keep in mind that you have no more than 10 minutes.)

Note: Remember this working group is inherently interdisciplinary. The connecting thread is information and its potential value in conjunction with data, inference and modeling of problems and complex systems. Therefore, prepare your presentation to an interdisciplinary group. (you probably want to avoid symbols and equations to save time.)

Morning Speakers (in order)

Moderator: John Harte

Bin Yu
Nataly Kravchenko-Balasha
Andrew Charman
Essie Maasoumi
Tasha Fairfield

Coffee Break (30 Minutes)

Spiro Stefanou
Rossella Bernardini Papalia
Radu Balan
Gideon Yaffe
John Harte

12:00PM – 1:00PM Lunch

1:00PM – 2:45PM

Individual Presentations Continues (in order)

Moderator: Min Chen

Wojtek Szpankowski
Luis Bettencourt
Trina Wellert
Ariel Caticha
Amos Golan
Min Chen

2:45PM – 3:15PM Coffee Break

3:15PM – 5:00PM Group Discussion

Moderator: Spiro Stefanou

We will concentrate on the questions:

1. What is value, or potential value, of information?
2. 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?

1399 Hyde Park Road | Santa Fe, New Mexico 87501 | www.santafe.edu | 505.944.8800

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

5:10 **Hotel Santa Fe Shuttle departs Santa Fe Institute**

7:00PM **Group Dinner (Location TBD)**

1399 Hyde Park Road | Santa Fe, New Mexico 87501 | www.santafe.edu | 505.944.8800

Day 2: Wednesday, April 3, 2024

8:15 Hotel Santa Fe Shuttle departs hotel to Santa Fe Institute

8:30 – 9:00 Breakfast at Santa Fe Institute

9:00AM – 10:30AM

Moderator: Ariel Caticha

Open Discussion: Thoughts and New Ideas Following Last Session of Day 1

10:30AM – 11:00AM Coffee Break

11:00AM – 12:00PM

Breakout Groups (Four Groups)

Basic Questions for the Groups:

4. Can we develop concepts of potential value based on meaning and semantics?
5. 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?
6. What is the relationship between value of a model and the value of Information used to construct that model?

Rules: *Each group discusses these questions for about an hour. Group leader summarizes and then presents to all.*

12:0PM – 1:00PM Lunch

1:00PM – 1:30PM
Welcoming Remarks and Thoughts on Information and Value
David Krakauer (SFI, President)
1:30PM – 2:00PM
Moderator: Tasha Fairfield

Summary of Group Breakout Discussions

2:00PM – 2:45PM
Moderator: Essie

Open Discussion: What new ideas emerged in the Groups' discussions?

To prepare for that discussion (which is based on the previous sessions) we will try to reach the group's goal by outlining and sketching together a journal article (or a major talk) on the topic of potential value of information and data and its implications for complex data and systems.

2:45PM – 3:15PM Coffee Break

3:15PM – 5:00PM

Moderators: Min, John, Amos

Part 1 (~1 Hours). What new questions and perspectives have resulted from the first day and a half? Are the big questions about the potential value of information, data, modeling, and inference came into focus?

Part 2 (~45 minutes). Future work (Fall workshop, Journal issue, other)

Each person provides her/his conclusions – about 3-5 minutes. Then, open Discussions (including suggestions for specific actions and future activities/research). John/Amos/Min will conclude.

Conclusions and Plans for Future

5:00 Hotel Santa Fe Shuttle departs Santa Fe Institute

4. Any communication/media produced related to the meeting (e.g., any news item produced for Parallax or other SFI media). If you have any non-SFI media related to the meeting, please include that with your summary.

Workshop on the potential value of data and information to be held in April (From Parallax, January 2024):

This workshop will bring together researchers from a range of disciplines to work on quantifying the value of data.

In an age of abundant information, one of the major questions to answer is how to quantify the value of this data. "The potential value of data is not just about the quality of the information contained in a dataset, but about what types of questions we can answer with it," says SFI External Professor Amos Golan (American University). "On one hand, this is a very philosophical question, on the other hand, we want to make it very empirical."

In April, researchers from around the world will meet at SFI for a two-day workshop, titled "The Potential Value of Data," to discuss methods for quantifying the potential value of specific datasets. "Information can appear to be useless until a model is constructed that renders it useful," says SFI External Professor John Harte (UC Berkeley), who is co-organizing the workshop with Golan and Min Chen, a professor at Oxford University.

The effort to quantify the value of data was prompted in part by recent efforts by government agencies to compile and maintain publicly available datasets. Given the enormous cost requirements of creating and maintaining these datasets, this creates a need to quantify the value of existing datasets and predict the value of future datasets. "Usually when people talk about the value of data, they look at what people have already gotten out of a dataset, such as the number of papers published, but a more important issue is to think about the potential value" Golan says.

In addition to discussing ways of quantifying the value of existing datasets, they also plan to discuss methods for optimizing future datasets, which includes identifying specific high-value information that will increase the number of questions it can answer. "Thinking about the questions that you can answer with the data can help us in the practical sense, because then we can also evaluate what is the data that we wish we had, and what would be the cost of acquiring it," Golan says. "Sometimes the answer is very simple."

4. Participant survey responses.

The final summary report will be provided to the organizers and relevant SFI staff. Feel free to share with your participants.

1399 Hyde Park Road | Santa Fe, New Mexico 87501 | www.santafe.edu | 505.944.8800

SANTA FE INSTITUTE

Appendix A. The Broader Context and Detailed Objectives the WG

The Broader Context, Background and Specific Objectives

A Brief Background

Individuals, researchers, and policy makers need information to make informed and educated decisions. These decisions are improved if the information used is of high quality, and if the inferential approach used to transform the information into knowledge and decisions is efficient and logical. In this working group we concentrate on studying the first issue: the quality of information and its value and its relationship with the inferential approach used. Within that topic our emphasis is on observed information: data. But unlike much of the literature dealing with the quality – and access to – data, we are interested in studying measures to evaluate the *potential value* of information and data. It is somewhat like the 'option value' of the data. We define the potential value as the overall value that society may obtain from a certain data set, assuming all the information and knowledge embedded in that data are extracted. It is not a value based on past use of the data, but rather the complete potential of that data, if indeed it will materialize. It is based on the quality of the data and meaning of the information under different contexts.

The motivation for this study is both philosophical and practical. The philosophical one deals with a special case of the more abstract approach for dealing with information and its value (see Dunn and Golan, 2021 and the references provided there) that concentrates on the value of observable information used for inference and decision making. The practical motivation stems from the need for a simple and applicable way for evaluating the complete potential of datasets and the models used for the analysis (where models may be mathematically defined, a simulation program, or a machine-learned model). This need has been underlined recently by policy makers who require federal agencies to collect, develop and produce data for public use. The tools, however, for such evaluations are yet to be defined and developed.

Complex Systems, Information and Value

This Working Group deals with a few interdependent questions – all related to our understanding of the information and data in complex systems and the potential value of that information and the inferential models used for evaluating and studying this information.

The first basic question: 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?

The second basic question: Is it possible to extend (if possible) information theory to account for the meaning of the information embedded in the data?

Information, Inference and Value

A fundamental issue that arises in studying the above questions is the possible relationship between the potential value measures and the inferential procedure used to transform data/information to knowledge.

The third basic question: Is the value of information independent of the inferential approach used?

Model, Theory and Value

Here a model may be a mathematically-defined model, a simulation program, or a machine-learned model.

Information and data have value. That value is relative and may be conditional on the inferential approach. But is it possible to evaluate the informational value of a model or a theory, and will information theory provide us with the necessary tools for doing so? That takes us to the next basic question:

The fourth basic question: Is it possible to measure the value (or potential value) of a model or a theory?

Working Group Objectives

In this workshop we want to explore the above four questions. Within these questions we are also interested in the following special topics:

1. Can we quantify the notion of quality and value of information beyond the tools of information theory?
2. Can we develop concepts of value based on meaning and semantics?
3. 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?
4. What is the relationship between value of a model and the value of Information used to construct that model.
5. What is the possible impact (if at all) of AI on the value of information?
6. Is the value of acquiring data for solving problems and modeling different than the value of the information collected?
7. What are the fundamental differences between the quality and value of Information?
8. Is there a difference between the value of information/data and the value of actually being able to process and understand these data?

1399 Hyde Park Road | Santa Fe, New Mexico 87501 | www.santafe.edu | 505.944.8800