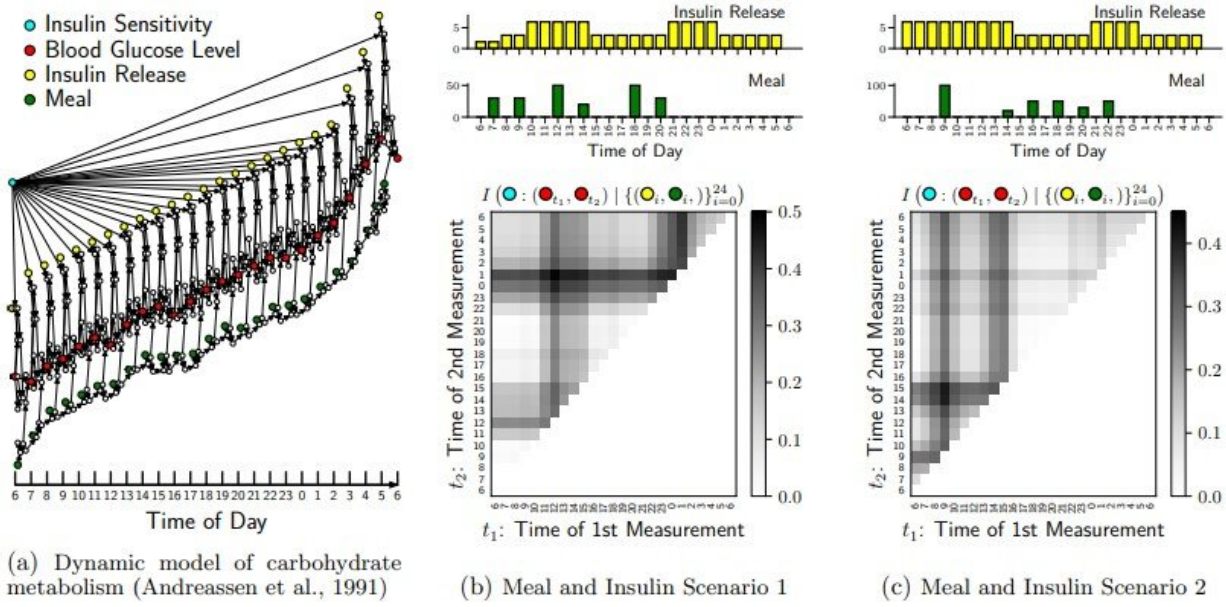
Estimating the informativeness of data

April 26 2022, by Rachel Paiste



(a) Dynamic model of carbohydrate metabolism (Andreassen et al., 1991)

(b) Meal and Insulin Scenario 1

(c) Meal and Insulin Scenario 2

Inferring optimal pairs of times to measure blood glucose level (red) that maximize information about a patient's latent insulin sensitivity (blue). Each heatmap in (b)–(c) shows estimates from EEVI of the conditional mutual information of insulin sensitivity with a pair of blood glucose measurements for all pairs of times, under a certain scenario of the patient's insulin release (yellow) and meal (green) schedule. Each scenario has a different optimal pair of times to measure blood glucose: 12pm/1am and 9am/3pm, respectively. Credit: https://arxiv.org/pdf/2202.12363.pdf

Not all data are created equal. But how much information is any piece of data likely to contain? This question is central to medical testing, designing scientific experiments, and even to everyday human learning

and thinking. MIT researchers have developed a new way to solve this problem, opening up new applications in medicine, scientific discovery, cognitive science, and artificial intelligence.

In theory, the 1948 paper, "A Mathematical Theory of Communication," by the late MIT Professor Emeritus Claude Shannon answered this question definitively. One of Shannon's breakthrough results is the idea of entropy, which lets us quantify the amount of information inherent in any random object, including random variables that model observed data. Shannon's results created the foundations of information theory and modern telecommunications. The concept of entropy has also proven central to computer science and machine learning.

The challenge of estimating entropy

Unfortunately, the use of Shannon's formula can quickly become computationally intractable. It requires precisely calculating the probability of the data, which in turn requires calculating every possible way the data could have arisen under a probabilistic model. If the data-generating process is very simple—for example, a single toss of a coin or roll of a loaded die—then calculating entropies is straightforward. But consider the problem of medical testing, where a positive test result is the result of hundreds of interacting variables, all unknown. With just 10 unknowns, there are already 1,000 possible explanations for the data. With a few hundred, there are more possible explanations than atoms in the known universe, which makes calculating the entropy exactly an unmanageable problem.

MIT researchers have developed a new method to estimate good approximations to many information quantities such as Shannon entropy by using probabilistic inference. The work appears in a paper presented at AISTATS 2022 by authors Feras Saad, a Ph.D. candidate in electrical engineering and computer science; Marco-Cusumano Towner, Ph.D.;

and Vikash Mansinghka, Ph.D., a principal research scientist in the Department of Brain and Cognitive Sciences. The key insight is, rather than enumerate all explanations, to instead use probabilistic inference algorithms to first infer which explanations are probable and then use these probable explanations to construct high-quality entropy estimates. The paper shows that this inference-based approach can be much faster and more accurate than previous approaches.

Estimating entropy and information in a probabilistic model is fundamentally hard because it often requires solving a high-dimensional integration problem. Many previous works have developed estimators of these quantities for certain special cases, but the new estimators of entropy via inference (EEVI) offer the first approach that can deliver sharp upper and lower bounds on a broad set of information-theoretic quantities. An upper and lower bound means that although we don't know the true entropy, we can get a number that is smaller than it and a number that is higher than it.

"The upper and lower bounds on entropy delivered by our method are particularly useful for three reasons," says Saad. "First, the difference between the upper and lower bounds gives a quantitative sense of how confident we should be about the estimates. Second, by using more computational effort we can drive the difference between the two bounds to zero, which 'squeezes' the true value with a high degree of accuracy. Third, we can compose these bounds to form estimates of many other quantities that tell us how informative different variables in a model are of one another."

Solving fundamental problems with data-driven expert systems

Saad says he is most excited about the possibility that this method gives

for querying probabilistic models in areas like machine-assisted medical diagnoses. He says one goal of the EEVI method is to be able to solve new queries using rich generative models for things like liver disease and diabetes that have already been developed by experts in the medical domain. For example, suppose we have a patient with a set of observed attributes (height, weight, age, etc.) and observed symptoms (nausea, blood pressure, etc.). Given these attributes and symptoms, EEVI can be used to help determine which medical tests for symptoms the physician should conduct to maximize information about the absence or presence of a given liver disease (like cirrhosis or primary biliary cholangitis).

For insulin diagnosis, the authors showed how to use the method for computing optimal times to take blood glucose measurements that maximize information about a patient's insulin sensitivity, given an expert-built probabilistic model of insulin metabolism and the patient's personalized meal and medication schedule. As routine medical tracking like glucose monitoring moves away from doctor's offices and toward wearable devices, there are even more opportunities to improve data acquisition, if the value of the data can be estimated accurately in advance.

Vikash Mansinghka, senior author on the paper, adds, "We've shown that probabilistic inference algorithms can be used to estimate rigorous bounds on information measures that AI engineers often think of as intractable to calculate. This opens up many new applications. It also shows that inference may be more computationally fundamental than we thought. It also helps to explain how human minds might be able to estimate the value of information so pervasively, as a central building block of everyday cognition, and help us engineer AI expert systems that have these capabilities."

The paper, "Estimators of Entropy and Information via Inference in Probabilistic Models," was presented at AISTATS 2022.

More information: Feras A. Saad, Marco Cusumano-Towner, Vikash K. Mansinghka, Estimators of Entropy and Information via Inference in Probabilistic Models. arXiv:2202.12363v2 [stat.ML], arxiv.org/abs/2202.12363

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology