**What Evidence is Total?**

**1. Introduction**

A maxim that seems to originate with Bernoulli, but is often attributed to Carnap as "the Principle of Total Evidence" goes like this (Carnap, 1947; p. 138-139):

"A principle which seems generally recognized, [J.M. Keynes, A Treatise on Probability (1921)], p.313] refers to `Bernoulli's maxim, that in reckoning a probability, we must take into account all the information which we have'. although not always obeyed, says that if we wish to apply such a theorem of the theory of probability to a given knowledge situation, then we have to take as evidence the *total evidence* available to the person in question at the time in question, that is to say, his total knowledge of the results of his observations."

The Principle has been cited by Ayer (1956) in asking why a scientist should seek new evidence, to which Good (1967) answered using the maximum utility principle. Carnap's Principle was endorsed by Hempel (2009), invoked Sober (2007) against hypothesis testing, and analyzed by Autzen (2016) in refutation of Sober; Epstein (2017) employs a familiar idea about what is *not* in the total evidence bearing on the hypothesis--any data that provoked consideration of the hypothesis.

Depending on how it is taken, Carnap's remark that the Principle of Total Evidence is "not always obeyed" is an understatement. If "the total evidence" is all of the data collected with the original intention of being used to measure something that bears on some scientific goal, the Principle of Total Evidence is routinely defied. Franklin (1989) vividly describes Millikan's rejection of measurements in the conduct of his oil drop experiments; every cognitive neuroscientist is familiar with deleting features of an EEG time series that are due to artifacts, or with deliberate alterations to BOLD time series in fMRI measurements, for example eliminating low frequency components of the signal--"highpass filtering." Both experimental trials and features of trials are routinely deleted or altered. Of course, if "evidence" in the Principle of Total Evidence refers just to the data, or aspects of the data, that are actually used in inference, the Principle is virtually tautological. We take Carnap to have intended that there is some standard of relevance of data to the goals of an inquiry, and within a body of data a subset or sub-aspect (or functions thereof) should only be taken as all the evidence the data provides only if no proper superset or functions would provide a better basis for inference. This is indeed vague, on three counts: "relevant" and "better" and "basis for inference" Relevant to what and better for what with what methods of inference? In considering some of the answers, we are interested in criteria of total evidence that are scientifically feasible, and for that reason we consider as well some aspects of computational complexity for various conceptions of total evidence. Our examples are chiefly but not exclusively from cognitive neuroscience.

It seems to us that the very idea of total evidence is not univocal and has more complexity than philosophical discussions have entertained. Autzen maintains that the appropriate notion of total evidence depends on the methodological framework. We agree, but think there is considerable texture involved: the idea of total evidence is fraught with complexities and ambiguities. Total evidence depends on the goals and the methodology used to attempt to achieve them; given the goals not all methodologies, and so not all conceptions of total evidence, are equally adequate--indeed some methods and uses of data are inconsistent with the very goals they are announced to serve. Some goals are local, and so is the total evidence for them, and some are more global and aspirational, and what is then total evidence may differ from local total evidence. Some goals, especially aspirational goals such as convergence to true information in the large sample limit, are combined with preferences for a certain kind of speed, and that preference may influence what should count as total evidence. In some contexts, the goals and methods, and thus the relevant total evidence, trade off accuracy against informativeness.

**2. New Kinds of Data**

We understand the topic of the tradition of discussions of total evidence this way: given an inference goal and a body of data, when does every proper subset of the data, or of its features, or functions of them, provide a worse basis for attaining the goal than some larger subset? But the topic can be framed more broadly. Sometimes there are measurements of a kind that could be made but were not, and the question arises as to whether the goal would have been better served by adding measurements of that kind. What kinds of data are to be taken into account, and how the data are segmented of summarized, is a serious matter in the sciences. In biological systematics, for example, there was (at least at one time) a debate as to whether cases should be divided into classes that are analyzed separately and a "consensus" obtained from them ("congruence"), or whether all the cases should be analyzed jointly ("total evidence") (Ermisse and Kluge, 1993; Fitzhugh, 2006), and whether both molecular and morphological data should be used, as against one or the other alone. Pylyshyn (1986) argued that data on active brain physiology should not be used in psychology; psychology has ignored his proscription. Methods of identifying regulator genes and the network of gene expressions they affect relied for decades on measures of concentration of RNA expression aggregated over many, many cells. For several reasons, the very aggregation distorted the results of inference (Chu, et al., 2003; Wimberly, et al., 2009), and yet the information wanted was there in the individual cells if only it could be obtained. In some sense, the data was not total. More recent methods of single cell cytometry now allow the relevant measurements on single cells.

It may not be known whether a body of evidence is total. A new kind of data may have to wait patiently for a method that can put it to good use. Consider a psychologist in 1995 interested in the causes of autism. Informed of brain imaging devices she may very well have no idea whether the data they can produce can provide evidence that bears on her problems. She might reasonably have the same uncertainty about genomic data. To get into a "total evidence" club for an existing goal, a new kind of data must be shown to be worthy. Worthy or not for an existing goal, a new kind of data may literally create its own goals by making them seem feasible. Thus determining neural connectivity became a goal of its own with fMRI data. Often, the new data can only prove its worth with the help of new methodologies. Magnetic resonance has a long history as a technique for chemical identification; it was only with the development of methods for tuning the instrument to blood oxygenation levels and extracting images from the data that it proved useful for psychology, and then only as useful as methods could be found to make inferences of scientific interest from those images.

**3. The Bayesian Formulation**

It would seem that the Bayesian framework provides a straightforward account of total evidence: Given a data set **D** and a set **H** of hypotheses, the total evidence that **D** provides relevant to **H** is any subset or marginal distribution **D'** of **D** such that for all supersets **D\*** of **D'** and for all h in **H**, Pr(h | **D\***) = Pr(h | **D'**). The total evidence may not be a unique subset of **D**. This works very well in some contexts. For example, if **H** is a set of hypotheses about variable X (note that any function of the given sample is also a random variable), the Markov blanket of X is the smallest set of variables in **D** conditional on which all other variables in **D** are independent of X. (Here we have assumed that the Markov blanket of X is unique.) Then the Markov blanket of X provides the total evidence about **H** in the Bayesian sense. For various classes of distributions, the Markov Blanket of a variable is easy to find. But practical considerations weigh against the Bayesian formulation in circumstances that we will discuss in section 5.

**4. Misfitting: Data, Goals, Methods and Evidence**

In several subjects--climate research, genomics, and neuroscience for example--the acquisition of huge data sets with large numbers of variables potentially standing in a still larger number of relationships has prompted the development and application of algorithms capable of searching for multiple explanatory relationships among the variables. These procedures can use multiple hypothesis tests or, alternatively, a score of some kind that is a surrogate for posterior probability. The goal in such inquiries is to extract from the data as much true information as possible about the underlying relationships. Which collection of functions of the data serve as the total evidence is a complicated matter because there are trade-offs between how much information is extracted and how accurate that information is. We refer to the probability that a relationship inferred by a method is true as "precision" and the probability that true relationships are discovered as "recall." Ideally, one wants both high precision and high recall, but in practice the cost of greater precision is lower recall, and vice-versa. One does not want to apply to data methods that are assured of low precision--i.e., false positives--no matter how good the measurements or how large the sample.

*Correlation and partial correlation in fMRI*

Consider correlations. Brain imaging used to estimate neural "functional" connectivity illustrates how a goal, a conception of total evidence, and a methodology can be so misaligned that save by sheer luck the kind of evidence used as the method prescribes could not possibly realize the goal even with ideal evidence of that kind and unlimited sample sizes. Functional magnetic resonance produces time series of 2 to 3 cm3 "voxels" of sources of radio frequency signals emitted by brain tissue when subjected to a combination of magnetic fields and an input radio frequency signal. The original use of the technology in neuropsychology was to identify which regions of the brain "light up" when subjects are given various tasks, in the hope of identifying a region or regions essential to a kind of cognitive task. Ambitions for the method soon evolved into attempts to identify "functional connectivity"--an ambiguous term that was often understood to mean something like the causal pathways between brain regions that contribute to some cognitive function. The (still) most common method is to identify such connections with thresholded correlations of averages of time series of clusters of voxels, sometimes called "Regions of Interest." The result is an undirected graph supposed to represent something about neural connectivity beyond anatomical connections--the edges in the graph are supposed to represent relatively (to the other variables) direct influences with unspecified directions. But simple correlations are the wrong functions of the data to identify direct or indirect influences between brain regions. For example, X can cause Y and Y can cause Z and X and Z will be correlated although there is no direct connection between X and Z. If there are multiple intermediates like Y, the X, Z correlation can be stronger than the X, Y or Y, Z correlations. Thresholding correlations will not find true causal connections except by luck. Neuropsychologists have also recommended identifying causal mechanisms from imaging data using partial correlations, that is, inferring that two variables have a causal connection if they are dependent conditional on all other variables measured. This produces an error when X -> Y <- Z, finding a false X - Z connection.

It would be a mistake to regard the correlation matrix of the data, or its inverse, the partial correlation matrix, as the "total evidence" that imaging data provides for influences of neural regions on one another. When the underlying distribution is Gaussian, more complex functions of the correlations are needed to identify the causal relations. When the underlying distributions are non-Gaussian, the correlation matrix is not the total evidence in the straightforward sense that more accurate and more informative inferences can be made using higher moments of the probability distribution estimated from the sample.

*Consistency*

Various inference procedures have associated criteria for convergence to the truth, or in the odd parlance of statistics "consistency." In hypothesis testing consistency might require that in the large sample limit the probabilities of Type II errors specified by the alternative hypothesis converge to zero. Part of what counts as total evidence in hypothesis testing depends on whether the test used and the data are consistent for the hypothesis to be tested against one or more alternatives.

Addressing Sober's criticism of hypothesis testing, Autzen shows that (assuming the sample size is fixed) the p values for one-sided hypothesis tests are a sufficient statistic for the mean of a distribution because there is a bijective map from the sufficient statistic values (e.g. the sample average in tests of the mean) to the p values ; there is no such correspondence for two-sided tests because different values of the sample average can correspond to the same p value, e.g. averages of .5 or of -.5 where the null hypothesis is that the mean of a symmetric distribution is zero. (Kuffner and Walker (2017) say the correspondence extends to two-sided tests, but the argument they give does not hold for a two-sided test.) The larger point is that sufficient statistics for a test may only capture the total evidence indirectly. For example, suppose a variable can take positive and negative values, the hypothesis is that the mean is zero, and the alternatives are mean values greater than or less than zero. A one-sided test would evidently not be consistent; clearly we need a two-sided test, but a statistical decision for a two-sided test of a symmetric distribution is immediate from a one-sided p value and the significance level used for it: reject if p < alpha/2 or (1- p).< alpha/2.

**5. A Case Study: Causal Structure Search**

In the literature of causal structure search, the “total evidence” for the task is sensitive to the methodologies used. For example, the PC algorithm (Spirtes et al., 2001) uses only the covariance matrix of the data and the sample size to derive conditional independence estimates and outputs an equivalence class of causal Directed Acyclic Graphs (DAGs); however, methods based on Linear, Non-Gaussian, Acyclic Models (LinGAMs) (Shimizu et al., 2006) further exploit certain non-Gaussianity aspects of the data distribution as part of the total evidence, leading to the ability to estimate a unique causal DAG.

Not only are different kinds of total evidence required by different methods, methods affect accuracy and scalability., For example, when searching for causal relationships, the Two-Step algorithm of Zhang (2017) has excellent precision, but for computational reasons is limited to a thousand or so variables. In contrast, the FGES algorithm (Ramsey, et al. 2017) can recover causal relations from a million variables in a sparse network with a sample size of only a thousand, but it is less accurate than Two-Step even on small problems. FGES gives the wrong results when there are feedback relations or unmeasured common causes. The aspects of the data used by the methods are different. The two-step algorithm uses functions of higher moments of the joint distribution of the sampled variables; FGES uses only a function of the second moment, the correlation. Which one represents the total evidence depends on what is wanted, or at least what one will settle for: accurate inferences among a smaller set of variables, or less frequently correct inferences among a much larger set of variables. It also depends on the hypothesis space of the search. If one were confident that there are no significant unmeasured causes, and no feedback, FGES might satisfy both for recall and precision and one might be satisfied that functions of the correlations provide total evidence. But if not, then not.

In functional connectivity studies, the time series of some variables ("voxels") are usually clustered into a handful of regions of interest and connections among those regions are estimated from the average of the signals of the voxels in the respective clusters. The reasons are sometimes because the regions are thought to be specifically relevant to a question, but the neglect of voxels other than those clustered raises the possibility that the associations between clusters are confounded by the processes in the omitted voxels. The data for all voxels in the cortex or the entire brain is available in the fMRI scans, but it is rarely used. Why not analyze relationships between voxels themselves rather than between selected clusters? The reason has been computational complexity: until very recently (Ramsey, 2016) available search methods could only process a small number of variables. If potentially relevant features of the data are neglected because available methodology cannot use them, or use them all, is the Principle of Total Evidence violated? That is perhaps the wrong question--the right one is how to find methods that can make use of the neglected data--but it illustrates one ambiguity of the Principle. One may have to include available methodology as a “condition” or “conditional variable” when defining total evidence—making it clear that such total evidence is methodology-specific.

A related issue of computational complexity arises with regard to practical applications of the Bayesian conception of total evidence. Bayesian inference is computationally demanding. Until the development of modern digital computers, Bayesian analysis was only feasible for very small, simple problems because of the difficulty of computing posterior probabilities. Digital computation and algorithms for approximating posterior probabilities (e.g., the Gibbs sampler) extended the range of Bayesian methods, but they are still computationally demanding. Strictly Bayesian searches for causal relations among variables is an NP hard problem, and currently limited to about 20 variables. In consequence, scoring methods tend to use surrogate criteria that are faster to compute, such as the Bayes Information Criterion (BIC) or Akaike scores, both of which are penalized functions of the likelihoods. These scores, which assess models by a function of the data, constitute the evidence used by various search architectures. What is total evidence depends on what can be computed.

**6. ­­Speed**

Autzen offers the following version of the Principle of Total Evidence for all frameworks:

"Suppose data *d*1 are strictly logically stronger than data *d*2, then an inference about hypothesis *H* should be based on *d*1 if changing between *d*1 and *d*2 changes the evidential assessment."

Practice aside, Autzen's condition seems unexceptionable, but it does not say whether if d1 does *not* change the evidential assessment from that of d2, which data set (or encoding) should be used; in other words, whether the evidence is total if there are available further data that would not change the evidential assessment. It would seem that such data can be censored and need not be included for the evidence to be total. But the matter is not so simple. One of the things one wants of statistical estimates and decisions is speed of convergence to the truth, and that aim can influence what counts as total evidence. Rather than examine details of rates of convergence of various estimators, consider a simple, contrived example.

*Colors and Engravings*: Suppose there is a collection of machines that each produce black and silver coins. Each machine can produce coins in an unlimited sequence, with independent outcomes--the colors of the coins. Nothing is known about the distribution over the various machines of probabilities for producing black coins. Suppose further that all coins of all machines have an engraving. Some unknown positive proportion of the machines will eventually produce a coin on which is engraved the true probability of black coins for that machine; all other coins will have "Hello" engraved on them. To be estimated is the probability that a randomly selected machine produces black coins. That probability can be estimated using the frequency of black coins in a finite sequence of data from the machine, and the frequency is a consistent estimator of that probability and a sufficient statistic. The question is whether the colors of the observed coins provide the total evidence, or whether instead the engraving should be included. Suppose all of the observed coins are engraved "Hello." That aspect of the data does nothing to change the assessment of the probability of black coins. But one would not want the engravings data censored *a priori*, for if the censored data had included a coin engraved with, say "0.7" the estimation problem would have been solved. For any method M that does not include the engravings as data there is a method that does include the engravings as evidence and that converges faster for some machines. The total evidence one wants is the count of observed black balls (or silver balls) *and* the engravings of all of the balls seen, even if the actual values of the engravings are uninformative.

**7. Conclusion**

The very idea of total evidence takes multiple forms in various settings. The details of the settings, and the respects in which evidence can fail to be "total" are manifold. It may be that the idea is too ambiguous, too broad in scope and too varying in detail, to be a useful methodological guide, and philosophy of science would do better to mind the details of how what evidence is used by what methods to what purposes.

**References**

Ayer, A. J., & Marić, S. (1956). *The problem of knowledge* (pp. 173-175). Harmondsworth:

Penguin books.

Carnap, Rudolf (1947). "On the application of inductive logic." *Philosophy and phenomenological research* 8, no. 1 133-148.

Chu, T., Glymour, C., Scheines, R., & Spirtes, P. (2003). A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, *19*(9), 1147-1152.

Eernisse, D. J., & Kluge, A. G. (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molecular Biology and Evolution*, *10*(6), 1170-1195.

Epstein, Peter Fisher (2017). "The Fine-Tuning Argument and the Requirement of Total Evidence." *Philosophy of Science* 84, no. 4 639-658.

Fitzhugh, Kirk (2006). "The ‘requirement of total evidence’ and its role in phylogenetic systematics." *Biology and Philosophy* 21, no. 3): 309-351.

Franklin, Allen (1989). *The Neglect of Experiment*. Cambridge University Press.

Good, I. J. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science*, *17*(4), 319-321.

Hempel, C. (2009). Scientiﬁc Explanation. *Philosophy of Science: An Historical Anthology*, 344.

Pylyshyn, Z. W. (1986). *Computation and cognition: Toward a foundation for cognitive science*. The MIT Press.

Ramsey, J., Glymour, M., Sanchez-Romero, R. (2017). A Million Variables and More. *International Journal of Data Science and Analytics*, 3: 121. https://doi.org/10.1007/s41060-016-0032-z.

Shimizu, S., Hoyer, P.O., Hyvärinen, A., & Kerminen, A.K. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003-2030.

Sober, Elliott (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition.

Wimberly, F., Danks, D., Glymour, C., & Chu, T. (2009). Problems for Structure Learning: Aggregation and Computational. *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*, 310.

Zhang, K., Glymour, C., Sanchez-Romero, R., Huang, B., & Ramsey, J. (2017) A Two-Step Method for Large-Scale Cyclic Causal Discovery. Available on arXiv.