Analytis, P. P., Moussaïd, M., Artinger, F., Kämmer, J. E., & Gigerenzer, G. (2014). **"Big data" needs an analysis of decision processes [Open peer commentary]**. *Behavioral and Brain Sciences*, *37*(1), 76–78.
https://doi.org/10.1017/S0140525X13001659

In these early days of big data, however, many studies seem to show us the obvious. In the best case, big-data studies will not compete with more traditional behavioral science but instead will allow us to see better how known behavioral patterns apply in novel contexts. In fact, they may even validate the most basic Bayesian analysis of human behavior there is, which is human experience. Humans sample the actions of their peers just by living among them for a lifetime. This takes us back to the northwest: Popularity does not guarantee quality. As long as people trust their own individual experiences, even in observing the behavior of others, a collective wisdom is possible.

ACKNOWLEDGMENTS

We thank Barbara Finlay, Herbert Gintis, David Geary, and our anonymous reviewers for excellent comments and advice on earlier drafts. We also thank Daniel Nettle for his many insights and suggestions for strengthening the argument, and Jonathan Geffner and Sumitra Mukerji for their editorial and production assistance.

NOTES

1. We can be more precise in the context of empirical statistical work by specifying (b_t, J_t) as functions of covariates and parameters of interest to estimate – for example, $b(x_t, \theta_b)$, $J_t = J(x_t, \theta_J)$, where x_t is a vector of potentially relevant covariates, which can include past values of the same covariates as well as past average choices over potentially relevant reference groups (for potential "contagion" effects) and average choices over potentially relevant reference groups (for potential "contextual" effects) (Manski 1993). Here, θ_b and θ_J are vectors of parameters that can be estimated. Once estimation is done, hypotheses can be proposed and tested using statistical methods. The era of "big data" opens up new possibilities for empirical work, formulation of hypotheses, and formal statistical testing of these hypotheses versus plausible alternatives.

2. The intense interest in these distributions, such as power laws, has led to a productive debate such that multiple alternative right-skewed distributions are now critically compared, with recognition that subtle differences in distributions can be informative as to the processes that produce them (Frank 2009; Laherrère & Sornette 1998; Venditti et al. 2010).

3. Care must be taken with the assumption that patterns in the northwest will always be Gaussian. Here is an example to the contrary. Consider the discrete-choice model with two choices {–1, +1} and with b_t and J_t being anywhere from zero to infinity. Let $h_t = u_{+,t} - u_{-,t}$, which is just the payoff difference of the two options. Then, the probabilities of choice at date t are given by

$$P_{+,t} = \frac{e^{b_t h_t}}{1 + e^{b_t h_t}}, \quad P_{-,t} = \frac{1}{1 + e^{b_t h_t}}. \tag{3}$$

Suppose that h_t exhibits a Gaussian distribution with zero mean and finite variance. As b_t approaches infinity, we observe only $P_{+,t} = 0$ *or* 1. Here we see that a Gaussian distribution of h_t is turned into a bimodal distribution with all mass at 0 or 1, even though we are in the northwest quadrant of the map. In the southwest, b_t is small. At the extreme south it is zero, and $P_{+,t} = 1/2$, $P_{-,t} = 1/2$, no matter the value of h_t. Now, given that h_t exhibits a Gaussian distribution with mean zero and finite variance, we can see that a small value of b_t will produce a unimodal, hence "Gaussian looking," distribution of P_+ when the system is in the southwest quadrant. As we move north by increasing b_t, we expect eventual bimodality of the distribution of choice probabilities, $\{P_{+,t}\}$, as we sample this choice process over time with increasing b_t.

4. One established model assumes the popularity, n_t, of a choice at time t as proportional to its popularity at time $t-1$:

$$n_t = (1 + g_t)n_{t-1},$$

where g_t, normally distributed over time, expresses the fluctuating rate at which agents in the population make new decisions. The result is a log-normal distribution of the accumulated popularity that spreads outward through time on a logarithmic scale. The model holds that the probability of a behavioral option accumulating popularity n at time t is given by

$$P(n) = \frac{1}{n\sqrt{t2\pi\sigma^2}} \exp\left[-\frac{(\ln n - g_0 t)^2}{2\sigma^2 t}\right],$$

where g_0 is the mean of g over time, with standard deviation σ. The position $g_0 t$ and width $\sigma^2 t$ of the log-normal peak increases with time t, such that the accumulated popularity distribution for options of the same age (e.g., citations of journal articles published in the same year) spreads outward through time on a logarithmic scale. This model can also be fit dynamically, as described by Wu and Huberman (2007): For each behavioral choice at time t, one calculates the logarithm of its popularity minus the logarithm of its initial popularity when the sampling started. Then to represent time t, the mean versus the variance of these logged values is plotted. Repeating this for all time slices in the sample, the resulting cluster of points will yield a linear correlation between the means and variances of the logged values (i.e., for this area of the northeast quadrant).

5. In the negative-binomial theorem, the probability of k choices of specific option x, given that there have been k+r total choices overall, is as follows:

$$\Pr(x = k) = \binom{k + r - 1}{k} p^k (1 - p)^r$$

6. A group size of 150 often is quoted as an average, but Dunbar never used either an average or a range in his original paper (Dunbar 1992), which had to do with neocortex size and group size in nonhuman primates. Group size in humans was addressed in later papers (Dunbar 1993; 1998). Often misunderstood is that Dunbar was referring to "meaningful" relationships, not simply the number of people one remembers: "The social brain hypothesis is about the ability to manipulate information, not simply to remember it" (Dunbar 1998, p. 184).

Open Peer Commentary

"Big data" needs an analysis of decision processes

doi:10.1017/S0140525X13001659

Pantelis P. Analytis,[a] Mehdi Moussaïd,[a] Florian Artinger,[b] Juliane E. Kämmer,[a] and Gerd Gigerenzer[a]
[a]*Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany;* [b]*Behavioural Science Group, Warwick Business School, The University of Warwick, Coventry CV4 7AL, United Kingdom.*
analytis@mpib-berlin.mpg.de moussaid@mpib-berlin.mpg.de
florian.artinger@wbs.ac.uk kaemmer@mpib-berlin.mpg.de
gigerenzer@mpib-berlin.mpg.de
http://www.mehdimoussaid.com/

Abstract: We demonstrate by means of a simulation that the conceptual map presented by Bentley et al. is incomplete without taking into account people's decision processes. Within the same environment, two decision processes can generate strikingly different collective behavior; in two environments that fundamentally differ in transparency, a single process gives rise to virtually identical behavior.

We applaud Bentley et al. for postulating a map of the environment in the form of a two-by-two classification. Too much of psychological theorizing still focuses on internal factors alone, such as traits, preferences, and mental systems – a kind of theorizing that social psychologists once labeled the "fundamental attribution error." Yet Bentley et al. are in danger of committing the opposite error: to theorize without regard of the cognitive processes. Herbert Simon (1956) noted more than 50 years ago that decision making is rather akin to the two blades of a scissors: the one blade is the decision strategy or heuristic; the other, the environment. Decision strategies have evolved and adapted to a given environment, and their rationality is ecological: therefore, one needs to analyze both the cognitive processes and the structure of the environment (Gigerenzer et al. 1999).

To apply this argument to Bentley et al., we show that the distributions in the four quadrants do not simply depend on the two environmental features, but, in addition, on the decision processes people rely on. We illustrate this point by a demonstration: We define two decision processes that have a strong social component but differ in whether social influence comes in the first step (the construction of the consideration set) or in the second step (the choice from this set). We then demonstrate that (1) within the same environmental structure (quadrant), the two decision processes can generate different distributions, and that (2) when the same decision process is used in two quadrants that differ in terms of transparency, the resulting distributions are almost identical.

We simulate 10,000 agents who sequentially choose from a set of 100 items (e.g., cameras, wines) drawn from a multivariate normal distribution. The agents aim for the item with the highest possible quality. The agents cannot evaluate the quality of the item with certainty but have to infer it from three attributes. They learn about each attribute's validity, that is, the relative frequency with which the attribute correctly predicts the item with the highest quality in pair-comparisons, in an initial training phase where they randomly sample 10 items.

Endowed with this knowledge, agents use a two-step decision process to select the best item. In the first step, each agent forms a subset of the available items, the consideration set ($n = 3$; Hauser & Wernerfelt 1990). In the second step, the agents decide which item to choose from this set by using a lexicographic decision rule akin to the take-the-best heuristic (Gigerenzer & Goldstein 1996). It ranks the attributes according to their validity; the item with the highest value on the most valid attribute is then chosen. In case two or more items have the same value on the first attribute, the second, and then the third attribute, respectively, is examined; if neither discriminates, the agents choose randomly.

Social influence is introduced in both processes but in different steps of the process. In the *popularity-set heuristic,* social influence is introduced in the first step. The probability that an item is part of the consideration set is proportional to how often the item has been selected by other agents in the past. In the *popularity-cue heuristic,* social influence enters in the second step as an additional, equally treated, fourth attribute corresponding to how often the item had been selected by others.

In the low-transparency environment, the correlations between quality and the three attributes are weak (0.11, 0.10, 0.10). In the high-transparency environment, the correlations between quality and attributes are strong (0.90, 0.69, 0.61). These environments correspond to the very north and south of the map by Bentley et al. just short of the border. In both scenarios, the inter-correlations between the attributes are held constant at 0.4. [The details of the simulation can be found at: http://www.mehdimoussaid.com/.]

As shown in Figure 1, the *popularity-set heuristic* generates collective herding, regardless of the environmental features. Here, a feedback loop operates: The more people choose an item, the more this item becomes attractive for subsequent decision-makers.

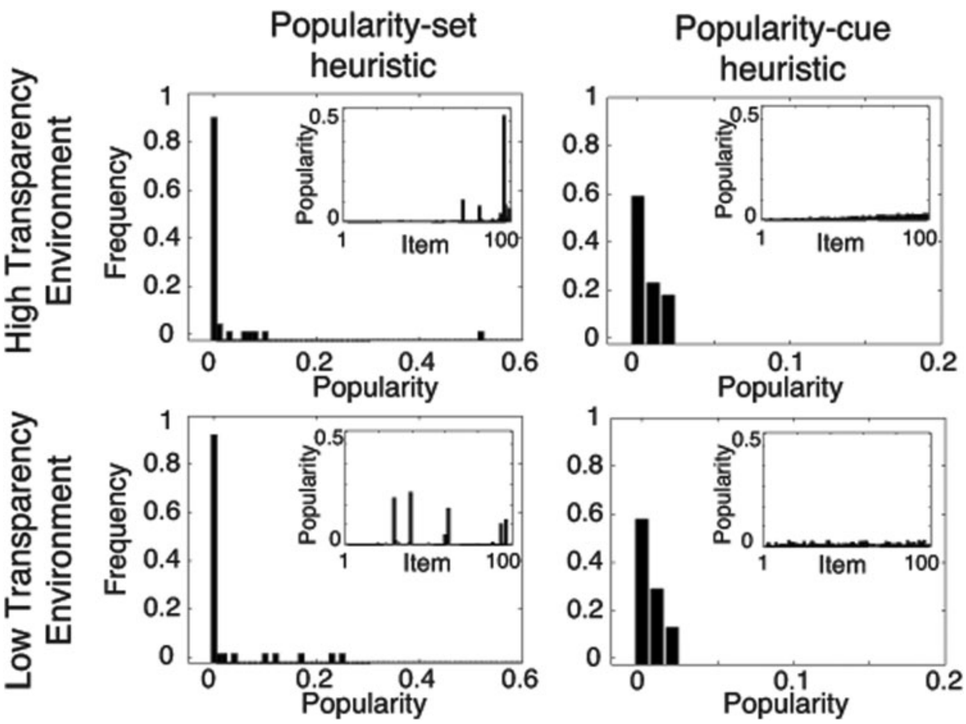

Figure 1 (Analytis et al.). The popularity distributions do not simply depend on Bentley et al.'s distinction between low- and high-transparency environments but also on the decision processes (heuristics). Within the same environmental structure (rows), the distributions change with the decision processes. When the same process (heuristic) is applied to different environments (columns), the result can be the same distribution. The outside graphs show the distribution of items' popularity at the end of the simulation. The popularity of an item is measured as the fraction of agents who have chosen that item. The inner graphs show the popularity reached by each of the 100 items separately, which are ordered according to their quality, from low quality (left) to high quality (right).

This leads to a few products becoming very popular, while most others are ignored. In contrast, the *popularity-cue heuristic* generates a homogeneous distribution of popularity, where each item receives a similar, low amount of choices, in high- and low-transparency environments alike. Here, social influence applies to a reduced subset of the items only, which prevents the feedback loop from setting up. These results show that *different* processes generate strikingly different results in the *same* environment. Likewise, the *same* process generates similar results when it is applied to *different* environments.

The conceptual map by Bentley et al. neatly reduces the environment to two variables. However, whether decisions are arrived at independently or not, and whether the information is transparent or not, is only half of the story. To fully understand what patterns emerge, one needs to account for the decision process. The relevance of such an approach is heightened by the interdependence of the social context where both mind and environment adapt to each other (see, e.g., Artinger & Gigerenzer, in preparation; Moussaïd et al., 2013). The advent of big data and the combination of experimental and simulation methods provide ample opportunities to study adaptive decision processes, stepping outside the black box of "as-if" decision theories.

"The map is not the territory"

doi:10.1017/S0140525X13001660

Fred L. Bookstein

Department of Anthropology, University of Vienna, A-1091 Vienna, Austria, and Department of Statistics, University of Washington, Seattle, WA 98195.
fred.bookstein@univie.ac.at

Abstract: Bentley et al.'s claim that their "map … captures the essence of decision making" (target article, Abstract) is deconstructed and shown to originate in a serious misunderstanding of the role of principal components and statistical graphics in the generation of pattern claims and hypotheses from profile data. Three alternative maps are offered, each with its radiation of further investigations.

My title, "The map is not the territory," is a famous sentence from a 1931 lecture by the general semanticist Alfred Korzybski (1933, p. 750). His meaning is that the graphical structure of a map need not be the structure of the territory (here, the scientific field) that it purports to represent. Statistical graphics, the discipline to which I have migrated his aperçu, is a field in which this insight has particular force. The target article's authors, Bentley et al., declare in their Abstract that they have "create[d] a multiscale comparative 'map' that, like a principal-components representation, captures the essence of decision making," that "each quadrant … features a signature behavioral pattern," and that "the map will lead to many new testable hypotheses concerning human behavior." My critique would have been Korzybski's: this map of theirs is not the territory, and cannot be trusted to capture any "essence." In particular, its topology, which is the conventional topology of principal component score plots, is all wrong for this subject-matter, which is preference profile patterns.

The objection is not so much to Bentley et al.'s quadrants per se as to the geometry of their diagram, which, in keeping with the conventional interpretation of principal components, is flat with a particularly depauperate connectivity. In the upper left panel of my Figure 1, I have redrawn Bentley et al.'s Figure 1 to emphasize the contrasts that concern the authors; these are along the edges of this square. The corresponding philosophical anthropology is a pure distillation of Levi-Strauss: two verbal oppositions ("independent" vs. "social" and "transparent" vs. "opaque") are treated as if these are as obvious and fundamental as water versus land, male versus female, light versus darkness, alive versus dead, or raw versus cooked, then converted into abstract propositions. We have no information about relationships along the diagonals, or relationships of the periphery to the center; in fact, the center is no construct at all. Nor is there any argument that the right axes are the north–south and east–west here; the phrases remain mere words. Nor do we know if the meaning of "independent versus social," for instance, is the same in "transparent" as in "opaque" domains of decision-making. There may be more than two types of edges here.

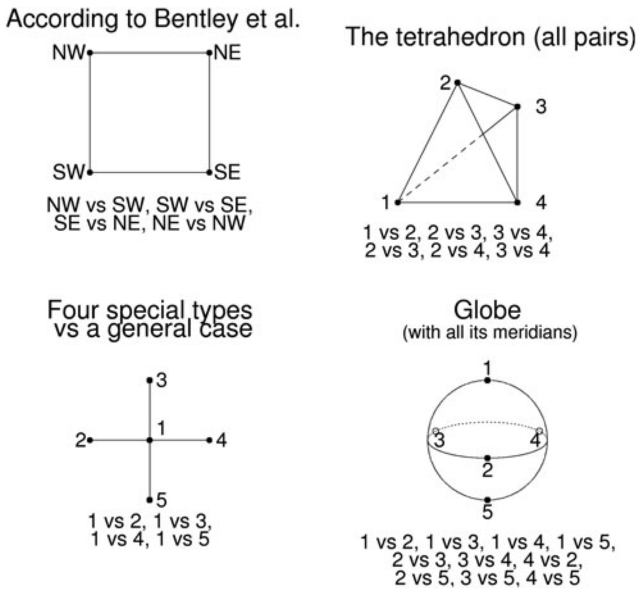

Figure 1 (Bookstein). The "map" according to Figure 1 of Bentley et al., together with three other graphics (tetrahedron, contrast with a general type, globe) based on the same data resources but leading to quite different lists of contrasts, rhetorics of interpretation, and suggested hypotheses for subsequent testing. See text.

"The map," which is Bentley et al.'s Figure 1, is certainly not "the territory," which is the actual information content of the data resources. For data arising from samples of time-series, as shown in Bentley et al.'s Equation (1) and Figures 2 and 3, there are many other ways to organize a diagram that lead to quite different reporting languages and quite different "testable hypotheses." Here are three other possibilities.

Upper right: Four types symmetrically connected. This is the general situation of four types. No evidence is given that the configuration of the authors' data reduces to the two dimensions Bentley et al. show or, indeed, any two dimensions. Then there need to be six contrasts, not four.

Lower left: Four specialized types out of a common center. This is a commonly encountered topology in studies of biological evolution, wherein multiple descendant species are characterized by derived features that all descend from the same original feature. To the extent that preferences are developmental, they may embody the same central focus.

Lower right: Globe with an axis. Imagine Bentley et al.'s map as a local expansion of coordinate possibilities along an axis that shrinks these possibilities toward zero at either of two extremes, "everything popular" (without further profile) and "everything unpopular" (without further profile). The topology is now a globe. On it, the north and south poles stand for the pure configurations, while an equatorial band offers space for additional parameters corresponding to the decision profiles that incorporate behavioral modifiers. Such data structures are commonly encountered in the compositional sciences, such as mineralogy or personality profiles. The authors' map is the equatorial plane of this construction (but it still has the wrong topology).