Complexity and Diversity

When Rob Boyd and I started thinking about the evolution of human behavior back in the 1970s, we decided that the complexity and diversity of human societies and cultures was one of the main challenges. For those of us with a background in biology—I came from entomology and ecology—an appreciation for the complexity of organisms and their diversity was part of the atmosphere. Entomologists are fond of the reply that the great early 20th Century evolutionist JBS Haldane is reputed to have given to a clergyman who asked him what he had learned of the Creator by studying his work: "He must have an inordinate fondness for beetles." Beetles are the most diverse order of the most diverse class of animals, the insects. Regarding complexity, Haldane was equally pithy: "My suspicion is that the world is not only queerer than we suppose, but queerer than we *can* suppose."

One of the paradoxes of ecology and evolutionary biology was, and still is, that theoretical biologists were fond of creating simple models the world. Doesn't a complex real world suggest that our thought about that world have to be complex? Biological theorists made three different kinds of arguments for applying simple models to complex, diverse phenomena:

> 1) Surface diversity and complexity concealed a lot of simplicity that proper theory accurately depicts. This position is often called reductionism.
>
> 2) Complex models are not worth the effort it takes to construct them. This used to be inescapable, but with advent of digital computers that could manage complexity it became less true. Ecologists could imagine making quite complex models and many did. But in practice complex models had only slightly or no greater predictive power than rather simpler models.
>
> 3) Simple models are hard enough to understand and complex ones are impossible. Science is about using the human mind to think about the world. Since there are limits to our ability to understand complexity we either have to use rather simple models or to blindly trust the robotic "thinking" of computers.

Even if we grant that the reductionist argument is too simplistic, or just plain wrong, the second two arguments remain valid. Besides, if complex realistic models are ultimately useful, they will have to be built up from simpler components.

The philosopher of biology Bill Wimsatt argued for a toolkit view of theory in complex diverse fields. Evolutionists and ecologists make many models for diverse problems. In the face of particular problems, biologists rummage through their toolkit for appropriate models likely to explain the case at hand. It often turns out that the behavior of the system they are studying is dominated by a manageable number of processes. Its behavior can be understood and predicted and understand with a relatively simple model. Of course, there are no guarantees, but as their toolkit expands biologists can hope to explain more and more phenomena. The toolkit itself becomes a highly organized system of submodels with known properties so that the analyst can readily guess that, for example, a frequency dependent model of selection is likely to apply if strategic interactions between individuals are important.

Humans are certainly diverse, complex organisms. Through culture, people adapt to their environments in historically contingent ways leading cultures to become diverse. In ecological terms humans resemble an adaptive radiation of species rather than a single species. The diversity of individual behavior within complex societies is comparable to that of the diversity of tissues and organs that make up multicellular organisms.

Many anthropologists and historians celebrate the complexity and diversity of human life much as natural historians do biological complexity and diversity. Other social scientists take a more reductionist view, arguing that much of the complexity and diversity of human behavior is superficial. Most famously, Noam Chomsky argued that most of linguistic diversity was superficial. Underneath the surface, languages were underlain by a few innate principles each with a relatively few parameter settings that differed from language to language. Perhaps much apparent human complexity and diversity will collapse in this way.

For Rob Boyd and I, the issue of whether human complexity and diversity turned out to be superficial or deep didn't matter in the first instance. Simple models of cultural evolution would be useful no matter how the reductionism issue came out. So we and others built up a toolkit of simple models that typically have variables and parameters that are open to interpretation. People are biased in favor of adapting some cultural variants as opposed to others in such models. Are the kinds of biases few or many? Are the bias parameters controlled by genes, culture, or a mix of the two? Are individuals or cultures variable with respect to the biases they apply or is there a lot of inter-individual and cross-population similarity? The models themselves were silent about such empirical issues.

After decades of studying ecosystems, ecologists and evolutionists still find that nature being "queerer than we can suppose" resonates. In my guise as an ecologist, I conducted long term studies of three lakes, all of which turned up major surprises even after years of study. As the applied ecologists Kenneth Burnham and David Anderson put it in their extremely influential 2002 statistics text, *Model selection and multimodel inference: a practical information-theoretic approach* (10,000+ sites according to Google Scholar) put it "we believe that 'truth' (full reality) in the biological sciences has essentially infinite dimension." At the same time, the data available about the complex real world is always strictly limited. In practice, our data limits us to only a glimpse of reality. The statistical theory Burnham and Anderson describe is designed to help us choose models that extract all of the information in our data (avoid "underfitting") while not imagining that our data contains more information than it does (avoid "overfitting").

Scientists have long worried about the problems of overfitting and underfitting data. Consider Einstein's famous aphorism "models should be as simple as possible, but no simpler." Information theoretic statisticians aim to put Einstein's intuition on a sound theoretical footing.

Economists, policy analysts and the like are fitting models to human societies all the time, trying to understand, trying to make sense of the complexity and diversity based on very little data – trying to predict outcomes. We want simple models but we want to choose the models that account for the maximum amount of the available data. Models that "underfit" leave real information in our data unused. As Nassim Taleb explained in his book *Black Swan*, risk analysts inspired by too-simple models "underfit" the problem of uncertainty contributed to recent economic policy disasters. They ignored the fact that real economic time series data had more extreme values than predicted by the too-simple normal distribution, leading them to create masses of securities that were far riskier than they believed them to be.

By the same token, we don't want to accept a simple genetic explanation for a bit of human behavior if culture also plays an important role, or vice-versa. The attempt of linguists to find a compact set of principles and parameters underlying the superficial diversity of languages failed as Fredrick Newmeyer, one who tried, explained to me recently. Languages have a lot of real diversity. Joe Henrich, Richard Nisbett and their colleagues have used experimental methods to buttress classical ethnographic accounts of cultural diversity. At the same time increasingly abundant gene sequence data is telling an increasingly complex story of rapid organic evolution in the late Pleistocene and early Holocene as modern humans spread out of Africa and adopted agricultural subsistence.

On the other hand, modern statistical packages tempt us to plug lots of explanatory variables into a multiple regression model to try to explain our data. But if we're not careful, we'll fit the noise as well as the information in the data. If we "overfit" our data, any truth it contains will be confounded with error. Chocolate consumption will appear to cause cancer in one study and protect against it in the next, even though we have no real grounds for thinking that chocolate has much to do with cancer one way or the other.

Information theoretic methods have a neat twist that reflects the fundamental uncertainty of the scientific enterprise. "Full reality" in all its complexity appears on both sides of a key equation in the derivation the information theoretic measure of "goodness of fit" of models. It therefore drops out of the analysis. We are left with an estimate how much closer or further away from full reality models in our analysis are compared to each other, but we remain completely in the dark about how close the best model is to the truth in an absolute sense. Full reality attracts our models, if we do our science right, but is itself never visible. This is an elegant way of restating the basic falsificationist or falliblist philosophy of science. Some theories are decisively poorer than others given the data we currently have. But more or better data and new candidate theories are likely to change things in the future in ways we cannot predict today.

Underfitting, overfitting and kindred problems are not rare in scientific practice. John Ioannidis wrote an influential paper in PLoS Medicine in 2005 entitled *Why most published findings are false*. Journalist David Freedman's new book *Wrong: Why experts keep failing us* rehearses these problems for a general audience. If we have a favorite explanation, and most of us do, and compare its ability to explain a set of data against a null hypothesis we risk underfitting. Any number of other hypotheses, alone or in combination with our favorite hypothesis, may fit the data better than our favorite. Others often approach data with no preconceived idea of what might explain a particular phenomenon. This invites "data dredging." Data dredging is an example of overfitting. Survey researchers and government statisticians sometimes produce huge data sets. In such cases, we are tempted to use all of the possible independent variablesin a blind hunt for a plausible explanation. If the list of independent variables is long, one of them is quite likely to fit the data quite well for spurious reasons.

In a complex, diverse world where full reality is always hidden, Burnham and Anderson recommend first thinking carefully about the explanations that are plausible candidates for the problem you are interested in. Not knowing full reality you are always prisoners of existing science and what you can imagination. No help for that! To the best of your ability, you put all plausible explanations on the table lest you underfit our data. But you want to rule out implausible explanations that invite overfitting before you address the data. Then, if you are lucky, your toolkit already contains good formal models of our candidate explanations. If not, you may have some theoretical work to do. With a considered collection of models in hand you set out to design an experiment or collect observational data to fit your models to. Finally, you to draw conclusions about the best model(s) given your data.

Even if we have followed the best practices possible, scientists of the complex and diverse should have a large dose of humility. Sooner or later, better models and better data will almost certainly come along. If our conclusion is lucky enough to be on the main trend of future findings, it may become part of the currently accepted wisdom. If we're unlucky, our findings may shortly be shown to be probably false. Even then, you can never be sure! Perhaps the disconfirming data is false, not your model. Darwin believed that the inheritance of acquired variation was an important evolutionary process. Throughout most of the 20th Century, this was held to be one of his greatest errors. Now, new data, reviewed by Eva Jablonka in a recent number of the *Quarterly Review of Biology*, suggests that trans-generational epigenetic inheritance provides a mechanism for the inheritance of acquired variation after all.

If you are tempted to enter the policy arena and give expert advice based on your science you are under a special obligation to be careful. Evolutionary scientists face a potential paradox. Our science suggests that behavior is very often adaptive and adaptations can be quite subtle. Of course, maladaptations also exist. But giving advice presumes that we know the difference!

No matter how much you know, you never know it for certain!