GEORGE DYSON [6.29.08]

For a long time we have been stuck on the idea that the brain somehow contains a "model" of reality, and that Artificial Intelligence will be achieved when we figure out how to model that model within a machine. What's a model? We presume two requirements: a) Something that works; and b) Something we understand. You can have (a) without (b). Our large, distributed, petabyte-scale creations are starting to grasp reality in ways that work just fine but that we don't necessarily understand.

Just as we may eventually take the brain apart, neuron by neuron, and never find the model, we may discover that true AI came into existence without anyone ever developing a coherent model of reality or an unambiguous theory of intelligence. Reality, with all its ambiguities, does the job just fine. It may be that our true destiny as a species is to build an intelligence that proves highly successful, whether we understand how it works or not.

The massively-distributed collective associative memory that constitutes the "Overmind" (or Kevin's OneComputer) is already forming associations, recognizing patterns, and making predictions—though this does not mean thinking the way we do, or on any scale that we can comprehend.

The sudden flood of large data sets and the opening of entirely new scientific territory promises a return to the excitement at the birth of (modern) Science in the 17th century, when, as Newton, Boyle, Hooke, Petty, and the rest of them saw it, it was "the Business of Natural Philosophy" to find things out. What Chris Anderson is hinting at is that Science will increasingly belong to a new generation of Natural Philosophers who are not only reading Nature directly, but are beginning to read the Overmind.

Will this make the scientific method obsolete? No. We are still too close to the beginning of the scientific method to say anything about its end. As Sir Robert Southwell wrote to William Petty, on 28 September 1687, shortly before being elected president of the Royal Society, "Intuition of truth may not Relish soe much as Truth that is hunted downe. "

KEVIN KELLY [6.29.08]

There's a dawning sense that extremely large databases of information, starting in the petabyte level, could change how we learn things. The traditional way of doing science entails constructing a hypothesis to match observed data or to solicit new data. Here's a bunch of observations; what theory explains the data sufficiently so that we can predict the next observation?

It may turn out that tremendously large volumes of data are sufficient to skip the theory part in order to make a predicted observation. Google was one of the first to

notice this. For instance, take Google's spell checker. When you misspell a word when googling, Google suggests the proper spelling. How does it know this? How does it predict the correctly spelled word? It is not because it has a theory of good spelling, or has mastered spelling rules. In fact Google knows nothing about spelling rules at all.

Instead Google operates a very large dataset of observations which show that for any given spelling of a word, x number of people say "yes" when asked if they meant to spell word "y. " Google's spelling engine consists entirely of these datapoints, rather than any notion of what correct English spelling is. That is why the same system can correct spelling in any language.

In fact, Google uses the same philosophy of learning via massive data for their translation programs. They can translate from English to French, or German to Chinese by matching up huge datasets of humanly translated material. For instance, Google trained their French/English translation engine by feeding it Canadian documents which are often released in both English and French versions. The Googlers have no theory of language, especially of French, no AI translator. Instead they have zillions of datapoints which in aggregate link "this to that" from one language to another.

Once you have such a translation system tweaked, it can translate from any language to another. And the translation is pretty good. Not expert level, but enough to give you the gist. You can take a Chinese web page and at least get a sense of what it means in English. Yet, as Peter Norvig, head of research at Google, once boasted to me, "Not one person who worked on the Chinese translator spoke Chinese. " There was no theory of Chinese, no understanding. Just data. (If anyone ever wanted a disproof of Searle's riddle of the Chinese Room, here it is.)

If you can learn how to spell without knowing anything about the rules or grammar of spelling, and if you can learn how to translate languages without having any theory or concepts about grammar of the languages you are translating, then what else can you learn without having a theory?

Chris Anderson is exploring the idea that perhaps you could do science without having theories.

This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

Petabytes allow us to say: "Correlation is enough. " We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can

throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

There may be something to this observation. Many sciences such as astronomy, physics, genomics, linguistics, and geology are generating extremely huge datasets and constant streams of data in the petabyte level today. They'll be in the exabyte level in a decade. Using old fashioned "machine learning, " computers can extract patterns in this ocean of data that no human could ever possibly detect. These patterns are correlations. They may or may not be causative, but we can learn new things. Therefore they accomplish what science does, although not in the traditional manner.

What Anderson is suggesting is that sometimes enough correlations are sufficient. There is a good parallel in health. A lot of doctoring works on the correlative approach. The doctor may not ever find the actual cause of an ailment, or understand it if he/she did, but he/she can correctly predict the course and treat the symptom. But is this really science? You can get things done, but if you don't have a model, is it something others can build on?

We don't know yet. The technical term for this approach in science is Data Intensive Scalable Computation (DISC). Other terms are "Grid Datafarm Architecture" or "Petascale Data Intensive Computing. " The emphasis in these techniques is the data-intensive nature of computation, rather than on the computing cluster itself. The online industry calls this approach of investigation a type of "analytics. " Cloud computing companies like Google, IBM, and Yahoo, and some universities have been holding workshops on the topic. In essence these pioneers are trying to exploit cloud computing, or the OneMachine, for large-scale science. The current tools include massively parallel software platforms like MapReduce and Hadoop (See: "A Cloudbook For The Cloud"), cheap storage, and gigantic clusters of data centers. So far, very few scientists outside of genomics are employing these new tools. The intent of the NSF's Cluster Exploratory program is to match scientists owning large databased-driven observations with computer scientists who have access and expertise with cluster/cloud computing.

My guess is that this emerging method will be one additional tool in the evolution of the scientific method. It will not replace any current methods (sorry, no end of science!) but will compliment established theory-driven science. Let's call this data intensive approach to problem solving Correlative Analytics. I think Chris Anderson squanders a unique opportunity by titling his thesis "The End of Theory" because this is a negation, the absence of something. Rather it is the beginning of something, and this is when you have a chance to accelerate that birth by giving it a positive name. A non-negative name will also help clarify the thesis. I am suggesting Correlative Analytics rather than No Theory because I am not entirely sure that these correlative systems are model-free. I think there is an emergent, unconscious, implicit model embedded in the system that generates answers. If none of the English speakers working on Google's Chinese Room have a theory of Chinese, we can still think of the Room as

having a theory. The model may be beyond the perception and understanding of the creators of the system, and since it works it is not worth trying to uncover it. But it may still be there. It just operates at a level we don't have access to.

But the models' invisibility doesn't matter because they work. It is not the end of theories, but the end of theories we understand. George Dyson says this much better in his reponse to Chris Anderson's article (see above).
What George Dyson is suggesting is that this new method of doing science—gathering a zillion data points and then having the OneMachine calculate a correlative answer—can also be thought of as a method of communicating with a new kind of scientist, one who can create models at levels of abstraction (in the zillionics realm) beyond our own powers.

So far Correlative Analytics, or the Google Way of Science, has primarily been deployed in sociological realms, like language translation, or marketing. That's where the zillionic data has been. All those zillions of data points generated by our collective life online. But as more of our observations and measurements of nature are captured 24/7, in real time, in increasing variety of sensors and probes, science too will enter the field of zillionics and be easily processed by the new tools of Correlative Analytics. In this part of science, we may get answers that work, but which we don't understand. Is this partial understanding? Or a different kind of understanding?

Perhaps understanding and answers are overrated. "The problem with computers, " Pablo Picasso is rumored to have said, "is that they only give you answers. " These huge data-driven correlative systems will give us lots of answers—good answers—but that is all they will give us. That's what the OneComputer does—gives us good answers. In the coming world of cloud computing perfectly good answers will become a commodity. The real value of the rest of science then becomes asking good questions.

[See "The Google Way of Science" on Kevin Kelly's Blog—The Technium]

STEWART BRAND [6.29.08]

Digital humanity apparently crossed from one watershed to another over the last few years. Now we are noticing. Noticing usually helps. We'll converge on one or two names for the new watershed and watch what induction tells us about how it works and what it's good for.

W. DANIEL HILLIS [6.30.08]

I am a big fan of Google, and I love looking for mathematical patterns in data, but Chris Anderson's essay "The End of Theory: Will the Data Deluge Makes the Scientific Method Obsolete?" sets up a false distinction. He claims that using a large collection of data to "view data mathematically first and establish a context for it later" is

somehow different from "the way science has worked for hundreds of years. " I disagree.

Science always begins by looking for patterns in the data, and the first simple models are always just extrapolations of what we have seen before. Astronomers were able to accurately predict the motions of planets long before Newton's theories. They did this by gathering lots of data and looking for mathematical patterns.

The "new" method that Chris Anderson describes has always been the starting point: gather lots of data, and assume it is representative of other situations. This works well as long as we do not try to extrapolate too far from what has been observed. It is a very simple kind of model, a model that says, "what we will see next will be very much what we have seen so far". This is usually a good guess.

Existing data always gives us our first hypothesis. Humans and other animals are probably hard-wired for that kind of extrapolation. Mathematical tools like differential equations and statistics were developed to help us do a better job of it. These tools of science have been used for centuries and computers have let us apply them to larger data sets. They have also allowed us to collect more data to extrapolate. The data-based methods that we apply to petabytes are the methods that we have always tried first.

The experimental method (hypothesize, model, test) is what allows science to get beyond what can be extrapolated from existing data. Hypotheses are most interesting when they predict something that is different from what we have seen so far. For instance, Newton's model could predict the paths of undiscovered planets, whereas the old-fashioned data-based models could not. Einstein's model, in turn, predicted measurements that would have surprised Newton. Models are interesting precisely because they can take us beyond the data.

Chris Anderson says that "this approach to science—hypothesize, model, test—is becoming obsolete". No doubt the statement is intended to be provocative, but I do not see even a little bit of truth in it. I share his enthusiasm for the possibilities created by petabyte datasets and parallel computing, but I do not see why large amounts of data will undermine the scientific method. We will begin, as always, by looking for simple patterns in what we have observed and use that to hypothesize what is true elsewhere. Where our extrapolations work, we will believe in them, and when they do not, we will make new models and test their consequences. We will extrapolate from the data first and then establish a context later. This is the way science has worked for hundreds of years.

Chris Anderson is correct in his intuition that something is different about these new large databases, but he has misidentified what it is. What is interesting is that for the first time we have significant quantitative data about the variation of individuals: their behavior, their interaction, even their genes. These huge new

databases give us a measure of the richness of the human condition. We can now look at ourselves with the tools we developed to study the stars.

SEAN CARROLL [6.30.08]

What Good is a Theory?

Early in the 17th century, Johannes Kepler proposed his Three Laws of Planetary Motion: planets move in ellipses, they sweep out equal areas in equal times, and their periods are proportional to the three-halves power of the semi-major axis of the ellipse. This was a major advance in the astronomical state of the art, uncovering a set of simple relations in the voluminous data on planetary motions that had been collected by his mentor Tycho Brahe.

Later in that same century, Sir Isaac Newton proposed his theory of mechanics, including both his Laws of Motion and the Law of Universal Gravitation (the force due to gravity falls as the inverse square of the distance). Within Newton's system, one could derive Kepler's laws—rather than simply positing them—and much more besides. This was generally considered to be a significant step forward. Not only did we have rules of much wider-ranging applicability than Kepler's original relations, but we could sensibly claim to understand what was going on. Understanding is a good thing, and in some sense is the primary goal of science.

Chris Anderson seems to want to undo that. He starts with a truly important and exciting development—giant new petascale datasets that resist ordinary modes of analysis, but which we can use to uncover heretofore unexpected patterns lurking within torrents of information—and draws a dramatically unsupported conclusion—that the age of theory is over. He imagines a world in which scientists sift through giant piles of numbers, looking for cool things, and don't bother trying to understand what it all means in terms of simple underlying principles.

> There is now a better way. Petabytes allow us to say: "Correlation is enough. " We can stop looking for models. We can analyze the data without hypotheses about what it might show.

Well, we can do that. But, as Richard Nixon liked to say, it would be wrong. Sometimes it will be hard, or impossible, to discover simple models explaining huge collections of messy data taken from noisy, nonlinear phenomenon. But it doesn't mean we shouldn't try. Hypotheses aren't simply useful tools in some potentially-outmoded vision of science; they are the whole point. Theory is understanding, and understanding our world is what science is all about.

JARON LANIER [6.30.08]

The point of a scientific theory is not that an angel will appreciate it. It's purpose is human comprehension. Science without a quest for theories means science without humans.

Scientists are universally thrilled with the new, big connected computing resources. I am aware of no dissent on that point.

The only idea in Chris Anderson's piece that falls outside that happy zone of consensus is that we shouldn't want to understand our own work when we use the new resources.

He finds it exciting that we could do something that works without understanding why. This is precisely what wouldn't be exciting. Some folk remedies work and we don't know why. Science is about understanding. Insight is more exciting than folk remedies.

Anderson pretends it's useless to be a human. Machines should now do the thinking, and be the heroes of discovery.

I say "pretends" because I don't believe he is being sincere. I think it's a ploy to get a certain kind of attention. Hearing anti-human rhetoric has the same sting as a movie plot about a serial killer. Some deep, moral part of each of us is so offended that we can't turn off our attention.

JOSEPH TRAUB [6.30.08]

I agree with Danny Hillis that large amounts of data will not undermine the scientific method. Indeed, scientific laws encode a hugh amount of data. Think Maxwell's equations or Kepler's laws for example. Why does Chris Anderson think that with still more data, laws (what he calls theory) will become less important?

JOHN HORGAN [7.2.08]

My first reaction to Chris Anderson's essay was, not another End-of-Something-Big prophecy. Anderson also recycles rhetoric from chaos, complexity, AI. Ever more powerful computers are going to find hidden patterns in bigger and bigger data sets and are going to revolutionize science! You don't need a computer to plot the boom-bust cycle of these claims. But the notion that computers will obviate theory and even understanding provokes a few thoughts:

Lots of groups already engage in problem-solving without understanding. Economists employ purely numerical methods for predicting markets, and mathematicians construct "computer proofs" based on massive calculations rather than comprehensible logic. This is less science than engineering. Engineers aren't looking for Truth. They're looking for a solution to a problem. Whatever works, works.

You could argue that since the advent of quantum mechanics, modern physics has also delivered prediction without understanding. Quantum theory is phenomenally successful, almost too much so for its own good, at predicting the results of accelerator experiments. But as Niels Bohr used to say, anyone who says he understands quantum theory doesn't know the first thing about it.

But I doubt that number-crunching computers will ever entirely replace human experts, as Anderson implies. The physicists at the Large Hadron Collider have to write programs to help their computers sift through the deluge of data for potentially significant events. IBM's massive number-crunching allowed Deep Blue to beat Gary Kasparov. But human chess experts also incorporated their knowledge into Deep Blue's software to make it more efficient at finding optimal moves. I bet Google's language translation programs incorporate a lot of human expertise.

Chris Anderson seems to think computers will reduce science to pure induction, predicting the future based on the past. This method of course can't predict black swans, anomalous, truly novel events. Theory-laden human experts can't foresee black swans either, but for the foreseeable future, human experts will know how to handle black swans more adeptly when they appear.

BRUCE STERLING [7.2.08]

Science Fiction Swiftly Outmoded by "Petabyte Fiction"

I'm as impressed by the prefixes "peta" and "exa" as the next guy. I'm also inclined to think that search engines are a bigger, better deal that Artificial Intelligence (even if Artificial Intelligence had ever managed to exist outside science fiction). I also love the idea of large, cloudy, yet deep relationships between seemingly unrelated phenomena—in literature, we call those gizmos "metaphors. " They're great!

Yet I do have to wonder why—after Google promptly demolished advertising—Chris Anderson wants Google to pick on scientific theory. Advertising is nothing like scientific theory. Advertising has always been complete witch-doctor hokum. After blowing down the house of straw, Google might want to work its way up to the bricks (that's a metaphor).

Surely there are other low-hanging fruit that petabytes could fruitfully harvest before aspiring to the remote, frail, towering limbs of science. (Another metaphor—I'm rolling here.)

For instance: political ideology. Everyone knows that ideology is closely akin to advertising. So why don't we have zillionics establish our political beliefs, based on some large-scale, statistically verifiable associations with other phenomena, like, say, our skin color or the place of our birth?

The practice of law. Why argue cases logically, attempting to determine the facts, guilt or innocence? Just drop the entire legal load of all known casework into the petabyte hopper, and let algorithms sift out the results of the trial. Then we can "hang all the lawyers, " as Shakespeare said. (Not a metaphor.)

Love and marriage. I can't understand why people still insist on marrying childhood playmates when a swift petabyte search of billions of potential mates worldwide is demonstrably cheaper and more effective.

Investment. Quanting the stock market has got to be job one for petabyte tech. No human being knows how the market moves—it's all "triple witching hour, " it's mere, low, dirty superstition. Yet surely petabyte owners can mechanically out-guess the (only apparent) chaos of the markets, becoming ultra-super-moguls. Then they simply buy all of science and do whatever they like with it. The skeptics won't be laughing then.

Graphic design. This one's dead easy. You just compare the entire set of pixels on a projected page of Wired to the entire set of all pixels of all paper pages ever scanned by Google. Set the creatometer on stun and generate the ultimate graphic image. Oh, and the same goes for all that music digitized in your iPod, only much more so. Why mix songs on "random" when you can tear songs down to raw wavelengths in an awesome petabyte mash-up? Then you can patent that instead of just copyrighting it.

Finally—and I'm getting a bit meta here—the ultimate issue of Edge. Rather than painfully restricting Edge responses to accredited scientists and their culturati hangers-on, the Third Culture will conquer the Earth when all Internet commentaries of any kind ever are analyzed for potentially Edgy responses, much as Google can translate Estonian to Klingon in one single step!

The result is the ultimate scientific-critical cultural thesis! It isn't a "Grand Unified Theory"—(theory is so over, because you can never print Google's databanks on a T-shirt). Still—bear with me here—metaphorically, I visualize this Petabyte Edge as a kind of infinite, Cantorian, post-human intellectual debate, a self-generating cyberculture that delicately bites its own dragon tail like a Chinese Ouroboros, chewing the nature of the distant real with a poetic crystalline clarity, while rotating and precessing on its own scaly axis, inside an Internet cloud the size of California.

DOUGLAS RUSHKOFF [7.2.08]

Yes, but.

I'm suspicious on a few levels.

First off, I don't think Google has been proven "right. " Just effective for the

moment. Once advertising itself is revealed to be a temporary business model, then Google's ability to correctly exploit the trajectory of a declining industry will itself be called into question. Without greater context, Google's success is really just a tactic. It's not an extension of human agency (or even corporate agency) but strategic stab based on the logic of a moment. It is not a guided effort, but a passive response. Does it work? For the moment. Does it lead? Not at all.

Likewise, to determine human choice or make policy or derive science from the cloud denies all of these fields the presumption of meaning.

I watched during the 2004 election as market research firms crunched data in this way for the Kerry and Bush campaigns. They would use information unrelated to politics to identify households more likely to contain "swing" voters. The predictive modeling would employ data points such as whether the voters owned a dog or cat, a two-door or four-door car, how far they traveled to work, how much they owed on their mortgage, to determine what kind of voters were inside. These techniques had no logic to them. Logic was seen as a distraction. All that mattered was the correlations, as determined by computers poring over data.

If it turned out that cat-owners with two door cars were more likely to vote a certain way or favor a certain issue, then pollsters could instruct their canvassers which telephone call to make to whom. Kids with dvd players containing ads customized for certain households would show up on the doorsteps of homes, play the computer-assembled piece, leave a flyer, and head to the next one.

Something about that process made me cynical about the whole emerging field of bottom-up, anti-taxonomy.

I'm all for a good "folksonomy, " such as when kids tag their favorite videos or blog posts. It's how we know which YouTube clip to watch; we do a search and then look for the hit with the most views. But the numbers most certainly do not speak for themselves. By forgetting taxonomy, ontology, and psychology, we forget why we're there in the first place. Maybe the video consumer can forget those disciplines, but what about the video maker?

When I read Anderson's extremely astute arguments about the direction of science, I find myself concerned that science could very well take the same course as politics or business. The techniques of mindless petabyte churn favor industry over consideration, consumption over creation, and—dare I say it—mindless fascism over thoughtful self-government. They are compatible with the ethics-agnostic agendas of corporations much more than they are the more intentional applied science of a community or civilization.

For while agnostic themselves, these techniques are not without bias. While their bias

may be less obvious than that of human scientists trained at elite institutions, their bias is nonetheless implicit in the apparently but falsely post-mechanistic and absolutely open approach to data and its implications. It is no more truly open than open markets, and ultimately biased in their favor. Just because we remove the limits and biases of human narrativity from science, does not mean other biases don't rush in to fill the vacuum.

OLIVER MORTON [7.2.08]

Chris Anderson's provocations spur much thought—I'll limit it to two specifics and two generalities. The first specific is that Anderson mischaracterises particle physics. The problem with particle physics is not data poverty—it is theoretical affluence. The Tevatron, and LEP before it, have produced amounts of data vast for their times—data is in rich supply. The problem is that the standard model explains it all. The drive beyond the standard model is not a reflection of data poverty, but of theory feeding on theory because the data are so well served.

That's not to say there's not a Googlesque angle to take here—there's a team looking at Fermilab data in what I understand to be an effectively "theory agnostic" way (see "Particle physicists hunt for the unexpected" by my Nature colleague Sarah Tomlin)—but it's not a mainstream thing. (And a minor add: a theory such as Newton's, which allows practitioners to predict with accuracy the positions of small fast-flying lumps of rock decades into the future in a solar system 10[25] larger than the rocks in question may be incomplete, but "crude" it's not.)

The second mischaracterisation is of biology. To suggest that seeing the phenotype as an interplay of genome and environment is in some way new knowledge or theoretically troubling just isn't so. But that's really all that talk of epigenetics and gene protein interactions amounts to. It's really unclear to me in what serious sense biology is "further" form a model today than it was 50 years ago. There are now models of biology that explain more of it than was explicable then, and no model for it all.

As to the general points, I don't think Feyerabend's far-from-normative account of the scientific method—"anything goes"—is the last word on the subject. But it is closer to the truth than saying that science always moves forward by models, or by any other single strategy. Science as a process of interested discovery is more than the tools it uses at any given time or disciplinary place.

And I guess my other point is "petabytes—phwaah". Sure, a petabyte is a big thing—but the number of ways one can ask questions far bigger. I'm no mathematician, and will happily take correction on this, but as I see it one way of understanding a kilobit is as a resource that can be exhausted—or maybe a space that can be collapsed—with 10 yes or no questions: that's what 2 [10] is. For a kilobyte raise the number to 13. For a petabyte raise it to 53. Now in many cases 53 is a lot of questions. But in networks of

thousands of genes, really not so much.

For understanding biology, you need to think much bigger. It's possible I described the beginnings of the way to go in "A Machine With a Mind of Its Own", an article I wrote for Wired on the robot scientist at the university of Aberystwyth, and I was pleased recently to hear that that program has started making genuine non-trivial discoveries. But maybe to do real justice to such things you will need millions or billions of experiments chosen by such algorithms—data generating data, rather than data generating knowledge; the sort of future portrayed in Vernor Vinge's Rainbows End, with its unspeakably large automated subterranean labs in San Diego.

PS Anyone who doesn't appreciate the irony in John Horgan's "not another End-of-Something-Big prophecy" should.

DANIEL EVERETT [7.3.08]

Chris Anderson's essay makes me wonder about linguistics in the age of Petabytes. In the early days of linguistic theory in the United States, linguists were, as all scientists, concerned with the discovery of regularities. Anthropologist Ruth Benedict first called regularities in human ways of giving meaning to the world "patterns in culture". Later, Edward Sapir, Kenneth Pike, and others looked for patterns in language, especially in the American Indian languages that became the focus of American linguistics and so differentiated it from inchoate linguistic studies by European scholars. Having just finished a field research guide, my own pedagogical emphasis for new field researchers is largely the same as the earliest studies of the indigenous languages of the Americas—enter a community that speaks an unstudied language and follow standard inductive procedures for finding regularities and patterns. Once the patterns have been discovered, write them up as rules, note the exceptions, and there you have it—a grammar.

But there are two ways in which linguists are becoming dissatisfied with this methodology as a result of issues that connect with Chris Anderson's thesis. First, linguists have begun to question the relevance of distinguishing rules from lists. Second, they have begun to ask whether in fact the child proceeds like a little linguist in learning their language with induction and deduction procedures built in to them genetically, or whether in fact child language learning takes place very differently than the way linguists study new languages in the field.

The difference between rules and lists, intentional vs. extensional sets, is the confrontation of the rule of law against lawlessness. As humans we are motivated by our evolution to categorize. We are deeply dissatisfied with accounts of data that look more like lists and "mere statistics" than generalizations based upon the recognition of law-like behavior. And yet as many have begun to point out, some of the most interesting facts about languages, especially the crucial facts that distinguish one

language from another, often are lists, rather than rules (or schemas). People have to learn lists in any language. Since they have to do this, is there any reason to propose a second type of learning or acquisition in the form of rules, whether these are proposed to be genetically motivated or not?

More intriguingly, do children acquire language based on a genetically limited set of hypotheses or do they treat language like the internet and function as statistical calculators, little "Googlers"? Connectionist psychologists at Carnegie Mellon, Stanford, and other universities have urged related hypotheses on us for years, though linguists have been slow to embrace them.

Linguistics has a great deal of work over the coming years to resituate itself in the age of Petabytes. Statistical generalizations over large amounts of data may be more useful in some ways, at least as we use such as parallel tools, than the armchair reflection on small amounts of data that characterizes earlier models in the human sciences. It very well may be, in fact to many of us it is looking more likely, that previous models based mainly on induction or genes are unable to explain what it is that we most want to explain—how children learn languages and how languages come to differ in interesting ways while sharing deep similarities.

GLORIA ORIGGI [7.3.08]

I agree with Daniel Hillis that Chris Anderson's point, although provocative and timely, is not exactly breakthrough news.

Science has always taken advantage of correlations in order to gain predictive power. Social science more than other sciences: we have few robust causal mechanisms that explain why people behave in such or such a way, or why wars break out, but a lot of robust correlations—for which we lack a rationale—that it is better to take into account if we want to gain insight on a phenomenon.

If the increase of child mortality rates happened to be correlated with the fall of the Soviet Empire (as it has been shown) this is indeed relevant information, even if we lack a causal explanation for it. Then, we look for a possible causal mechanism that sustains this correlation. Good social science finds causal mechanisms that are not completely ad hoc and sustain generalizations in other cases. Bad social science sticks to interpretations that often just confirm the ideological biases of the scientist.

Science depicts, predicts and explains the world: correlations may help prediction, they may also depict the world in a new way, as an entangled array of petabytes, but they do not explain anything if they aren't sustained by a causal mechanism. The explanatory function of science, that is, answering the "Why" questions, may be just a small ingredient of the whole enterprise: and indeed, I totally agree with Anderson that the techniques and methods of data gathering may be completely transformed by the

density of information available and the existence of statistical algorithms that filter this information with a tremendous computing capacity.

So, no nostalgia for the good old methods if the new techniques of data gathering are more efficient to predict events. And no nostalgia for the "bad" models if the new techniques are good enough to give us insight (take AI vs. search engines, for example). So, let's think about the Petabyte era as an era in which the "context of discovery", to use the old refrain of philosophy of science, is hugely mechanized by the algorithmic treatment of enormous amounts of data, whereas the "context of justification" still pertains to the human ambition of making sense of the world around us.

This leaves room for the "Why"-questions, that is, why are some of the statistical correlations extracted by the algorithms so damn good? We know that they are good because we have the intuition that they work, they give us the correct answer, but this "reflective equilibrium" between Google ranked answers to our queries and our intuition that the ranking is satisfying is still in need of explanation. In the case of PageRank, it seems to me that the algorithm incorporates a model of the Web as a structured social network in which each link from a node to another one is interpreted as a "vote" from that node to the other. This sounds to me as "theory", as a method of extraction of information that, even if it is realized by machines, is realized on the basis of a conceptualization of reality that aims at getting it right.

A new science may emerge in the Petabyte era, that is, a science that tries to answer the question of how the processes of collective intelligence made possible by the new, enormous amount of data that can be easily combined by powerful algorithms are sound. It may be a totally new, "softer" science, uninhibited at last by the burden of the rigor of "quantitative methods", that make scientific papers so boring to read, that leaves to algorithms this burden and lets the minds free to move around the data in the most creative way. Science may become a cheaper game from the point of view of the investment for discovering new facts: but, as a philosopher, I do not think that cheap intellectual games are less challenging or less worth playing.

LEE SMOLIN [7.3.08]

To see what to think about Anderson's hypothesis that computers storing and processing vast amounts of data will replace the need to formulate hypotheses and theories one can look at whether it has any relevance to how supercomputers are actually being used in contemporary physics.

One example that comes to mind is gravitational wave astronomy, where a large signal to noise ratio makes it impossible to simply observe gravitational waves from the outputs of the detectors. Instead, the vast data streams created by the LIGO, VIRGO and other gravitational wave antennas are scanned by computers against templates for waveforms

created by theorists modeling possible sources. These sources, such as inspiraling and merging of pairs of black holes and neutron stars, require themselves computationally intensive simulation on supercomputers to produce the needed templates.

What has been the experience after several decades of this work? While no gravitational waves have so far been identified, the detectors are up and running, as are the programs that generate the templates for the waveforms from the supercomputer simulations of sources. To reach this stage has required large amounts of computation, but that has at every stage been guided by theoretical knowledge and analytic approximations. The key issues that arose were resolved by theorists who succeeded in understanding what was going wrong and right in their simulations because they were able to formulate hypotheses and test them against analytical calculations.

While I don't work in this field, it has been clear to me over the decades I have been watching its development that progress was made by good physicists doing what good physicists always do, which is build up intuitive stories and pictures in their minds, which lead them to testable hypotheses. The fact that the hypotheses were about what was happening in their computer simulations, rather than about data coming from observations, did not change the fact that the same kinds of creativity and intuitive thinking were at work as is traditional in non-computational science.

There is a similar story in cosmology, where computer simulations of structure formation are part of an arsenal of tools, some computational, some analytic, some intuitive, which are always being challenged by and checked against each other. And there is a similar story in numerical studies of hadronic physics, where there is an interplay of results and ideas between supercomputer simulations and analytic approximations. There also, the key obstacles which arose had to do with issues of physical principle, such as how certain symmetries in the theory are broken in the numerical models. It has taken a great deal of creative, intuitive physical thinking over 30 years to overcome these obstacles, leading recently to better agreement between theory and experiment.

As a result of watching the development of these and other numerically intensive fields, it is clear to me that while numerical simulation and computation are welcome tools, they are helpful only when they are used by good scientists to enhance their powers of creative reasoning. One rarely succeeds by "throwing a problem onto a computer", instead it takes years and even decades of careful development and tuning of a simulation to get it to the point where it yields useful output, and in every case where it has done so it was because of sustained, creative theoretical work of the kind that has been traditionally at the heart of scientific progress.

JOEL GARREAU [7.6.08]

Maybe things are different in physics and biology. But in my experience studying

culture, values and society, data lags reality by definition—they are a snapshot of the past. And when human reality does not conveniently line up with established ways of thinking, the data can lag by years, if not decades.

Data are an artifact of selection, which means they reflect an underlying hypothesis or they wouldn't have been collected. For example, in my work I discovered a frightening lack of timely data to "prove" either my hypothesis that North America was behaving as if it were nine separate civilizations or economies that rarely were bounded by political jurisdictions of nation, state, or county. It was equally problematic coming up with the data to prove that places like Silicon Valley were becoming the modern iteration of "city" even when the millions of square feet of big buildings were right before your eyes. It wasn't until those "nine nations" or "edge city" models began to be seen as useful by others that people started to go through the very great deal of trouble to verify them by collecting data in a fashion that ignored all previous boundaries. Life is not obligated to follow data and it frequently does not.

Now come thinkers producing hypotheses that one can map social and cultural change onto Moore's Law. It will be interesting to see when or if the data shows up to support their predictions. Ray Kurzweil and the Singularitans see an exponential curve that eventually leads to a perfection of the human condition analogous to the Christian version of "heaven." Pessimists like Bill Joy, Francis Fukuyama, Susan Greenfield and Martin Rees see a mirror image curve leading quickly to something like "hell." These are both credible scenarios. But the data lag. It is hard to find "proof" that we are arriving at one or the other, even though they are each based on nice smooth technodeterministic curves, the likes of which rarely if ever have been a prominent artifact of human history. Lord knows how you would demonstrate through data the arrival of the "prevail" scenario described by Jaron Lanier and others. That scenario is based on the notion that a prominent aspect of future history is that the increase in our challenges is being matched quickly enough by imaginative, ornery, cussed, collective, bottom-up human responses, setting events off in unpredictable directions. If graphed, the outcome—like so much of the raw material of history—would probably appear about as organized as a plate of spaghetti.

I would love to think that data lagging behind hypotheses—much less reality—is about to change. (At last! A crystal ball!) But I eagerly await a demonstration.