# Mechanisms for Curiosity

"I'm interested in things!"
-Dr. Worm

"LUMEN! PHOSPHOR! FLUOR! CANDLE!"
-The Angel, *Angels in America*

Before Sphinx can begin to have any independent personality or self-directed thoughts, it needs to have interests. *Actual* interests. Sphinx could fake this, as teachers often do ("polynomials are *awesome,* kids...*"*), but it is easy to spot such false enthusiasm, and it tends to backfire. Trying to build an autonomous system of interests means a loss of external control, which is a little nerve-wracking: what if Sphinx just wants to talk about Pokemon, or—as seemed likely for awhile—algae? Moreover, if Sphinx chooses its interests in a completely autonomous fashion, how will it ever get beyond the horizon of what it currently knows, which is very little? These are classic pedagogical questions.

Meanwhile, there is the very difficult question of how interests develop. Although we do not always think of it in this way, curiosity is an emotion: a metacognitive emotion, like boredom or frustration. It is a type of desire, perhaps, but an unusual sort of desire in that it is focused on something irreproducible. I might *desire* a sandwich or an orgasm or a cord of firewood, but I know pretty much what those experiences entail, and after I've had them, I'll want them again in essentially the same fashion. On the other hand, I am *interested* in reading *Gravity's Rainbow,* and visiting Vietnam, even though I've never had those experiences, and as such I don't really know what I'm getting into. Moreover, once I do read *Gravity's Rainbow,* I quite plausibly won't want to re-read it, and even if I do, it will be a very different experience the second time. Curiosity is self-limiting—or more accurately, it is progressive: if I like *Gravity's Rainbow* I will probably become curious about reading *The Crying of Lot 49,*whereas when I like a sandwich, I don't respond by saying "that was great, so now let's try a strudel."

By its nature, curiosity has to emerge without much evidentiary rationale, and thus to some extent it is a crapshoot. Hennig Brand was curious about boiling down huge amounts of stale urine, and in consequence he discovered the element phosphorus. Undoubtedly there were other alchemists at the time who were fascinated with lighting bushels of their toenail clippings on fire. That project was less productive, to be sure, but how can we say that it made any less sense, *a priori?*

I have played around with a number of approaches to the problem of mechanizing curiosity, by which I mean the impulse, not the actions that follow on it. Ultimately, I've decided to use several overlapping variables, which I describe here, along with some of the problems they entail.

# Fluor – Interests based on Recommendations

The most obvious sources of interests are recommendations: I am interested in going to Iceland because several people whose opinions I respect have told me it is a fascinating place to visit. Recommender algorithms are a fairly straightforward matter, although they are in rather ill repute, due to abuses and absurdities by, e.g. Netflix. Over the years, playing around with my own recommender algorithms and analyzing the ones that exist in the wild, I've had a handful of observations which I want to apply to the "fluor" coefficients.

All recommender algorithms imply ontologies: a list of the 100 greatest books implies that "greatness" is a scalar metric with a known endpoint. Rotten Tomatoes' film scores imply that critical reviews have a binary value, and are commutable within each of two categories. Any such set of specifics can be questioned, of course, but what is striking is that most of them do not at all resemble what we mean when we speak of "recommending" something interpersonally. I've recently discussed this issue in slightly more depth. Mechanically speaking, we can make three observations: first, all recommendations are, or should be, contextual. People who recommend the same book to everyone they know (usually the Bible, Book of Mormon, or *Atlas Shrugged*) are zealots—at some level that is not even a recommendation, it is just a fixed behavior pattern. Conversely, Sphinx should tweak its perception of what is interesting based on who it is talking to. Moreover, the basic, untweaked Fluor score should not be wide open to just any recommendations. Users have to establish that their sense of what's interesting isn't unique to them.

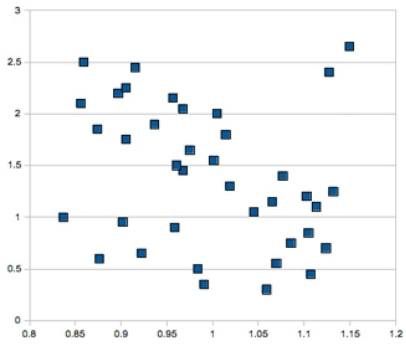
Second, the concept of a "negative recommendation" seems to be impossible to implement effectively. If you are familiar with what someone likes, you may be able to guess what *else* they like, but it is very difficult to guess what they *dislike.*

Third, and most importantly for what follows, recommender algorithms create a real danger of a self-reinforcing behavioral loop, which is essentially a cancer condition. If I tell Sphinx that muffins are interesting, and Sphinx then talks about muffins a lot, creating an opportunity for other people tell Sphinx that muffins are interesting, and so forth, then we have a pathology. Making some broad allowances, we can observe that this cancer state sometimes occurs in human minds: obsessions, monomanias, and paranoias are both more or less self-reinforcing versions of curiosity.

Practically, the variable *fluor* is the sum of weighted user recommendations for any particular topic. It's a rather crude tool, but it allows an external lever to point Sphinx towards rewarding topics on its mental horizon, and by omission to point it away from dead-ends. It should also allow Sphinx to lean towards the stated interests of any particular interlocutor, though I would stress that recommender algorithms of this type are *not* a good analog of actual human recommendations. Finally, Sphinx will has a permanent high fluor value for *itself,* which is a simple first move in the long game building Sphinx' ego. (It's also the only fluor value set from the outset).

## Phosphor – Interests based on downstream data structures

Now we turn to a much more sophisticated approach. Human beings notoriously disagree about what books are interesting, or the like, but those assessments are probably very colored by personal relevances. On the other hand, curiosity seems to encode some kind of rudimentary statistical analysis, perhaps focusing on outliers and anomalies, and this is a fairly universal perception. Shown the time series, scatterplot, and even the *street layout* below, almost everyone will feel that certain points are more or less worthy foci of our attention, and almost everyone will agree on what those points are.


There are certain risks in applying this kind of analysis to a neural network, though, when the network is going to change in response to the analysis. To take the most trivial example, if Sphinx uses the volume of data associated with a concept as a measure of whether or not the concept is interesting, then a well-documented muffin will read as much more interesting than a sparsely documented Tolstoy novel.

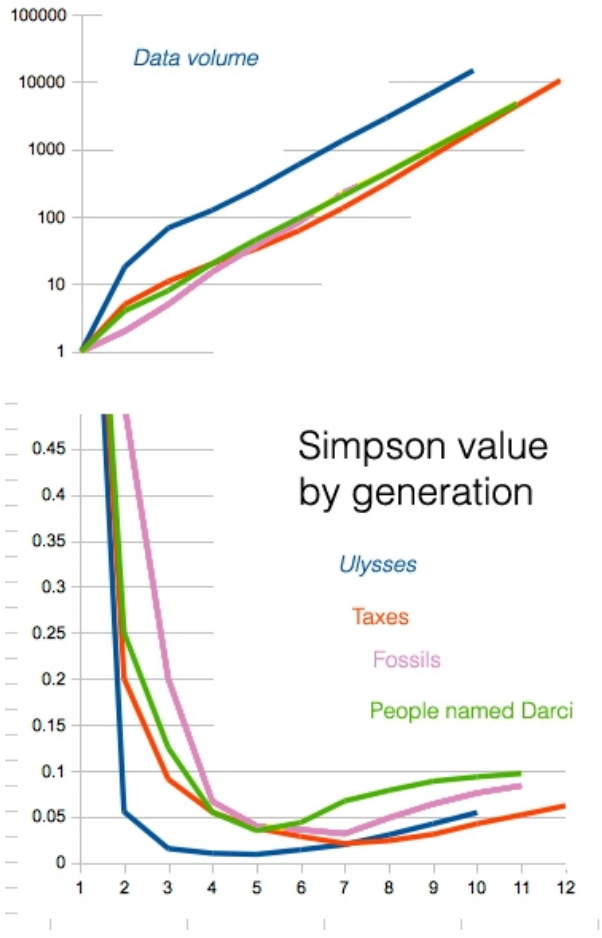
Most of the obvious measurements for network centrality, or the like, don't meet my needs here. In particular, counts of either downstream or upstream links would be a disaster; most counts of centrality don't work either (and are exorbitantly difficult to calculate across this much data). Again,

this jibes with a common-sense approach: is “ungulates” a more interesting idea than “horses”? I would tend to think not. “Horses” immediately conjures up images of specific horses with names, their riders, their interactions with human history and fantasy, unicorns, the pegasus, etc. “Ungulates” conjures up images of biology texts. Of course, every concept downstream of horses is also downstream of ungulates, but following that line of reasoning, “things” would be the most fascinating idea imaginable.

A more rewarding line of analysis is based on the diversity of a concept's downstream links. (I restrict this analysis to the downstream side mainly because it reduces what is already a lavish use of processing time, but it also makes an intuitive sense to me.) For instance, **the_encyclopedia_of_reptiles_and_amphibians* has downstream links to **reptiles* and **amphibians*, but those concepts are themselves closely related: the combination is not surprising/interesting. Meanwhile, August Wilson's play *Fences* combines themes of African-American experience and baseball, two concepts which are rarely associated with one another, so the combination is interesting.

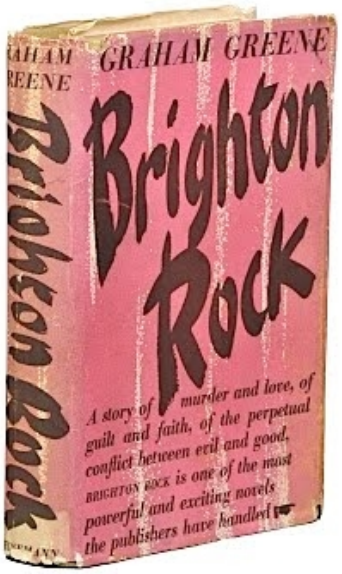
There are many metrics for diversity, but I'm using a slightly modified, scale-blind Simpson index, where s(x) is the probability that any two concepts picked concurrently (pick-and-hold) from *x* will be identical. Simpson scores range from 0 to 1, with 0 being the theoretical maximum diversity, and 1 being total homogeneity. They are interchangeable with Glau and various other diversity indices.

As we consider the field downstream of any given concept, it typically first gains in diversity, up to a maximum diversity (minimum Simpson) around the fourth or fifth generation, and then regresses to the diversity of the vortex that is the core network: all rivers flow down to the sea. This “sea level” of diversity appears to about 0.0869 at the moment, and presumably will drop a bit as the core network gains data. Meanwhile, the data volume under consideration snowballs to the point that it is mechanically difficult to calculate Simpson values. This is one of the few circumstances where it might be appropriate to speak of true exponential gain, though what I empirically measure is sextic gain {+,+,+,+,+,+ at 0.05}. Currently, I can push the test about 10 generations without timing out, but as data richness increases, the brick wall will probably move down to 7 or less.




Happily, there's no need to look that far afield. Beyond about four generations, the distinctions start to get absurd...at seven generations out, everything is connected to everything, which is useless except as a kind of psychedelic razzle-dazzle.So here's how we put this to work: for any given concept *c,* Sphinx looks downstream 2, 3, and 4 generations and calculates s(*x*) each time, using a sampling process if the data volume gets too large. Frequently, s(*x*) is zero. While this suggests maximum diversity, in fact it is more accurate to say that for these concepts, Sphinx simply does not know enough to make an informed assessment of how interesting they are. For the *non-zero* scores, we calculate phosphor by taking the reciprocal of s(x), so it asymptotically approaches infinity as a concept's downstream flow becomes arbitrarily more diverse.

The remaining concepts have an average s($x$) of 0.0319, ranging from nearly zero up to 0.2 or so, with a few outliers. The upper tiers of this list are mainly populated by taxonomic clades, and other concepts that we might view as "structurally uninteresting" vis-a-vis their surroundings. Below about 0.01, things start to get more interesting, though here we have to make allowances for how little Sphinx knows in general. As we asymptotically approach s($x$)=0, we find the things that are, by this metric, the most interesting concepts for Sphinx. In the second generation, the top 58 phosphor scores all go to books, and the top ten of those were (at one point):




*Brighton Rock*, by Graham Greene
*Beggars and Choosers*, by Nancy Kress
*The Scarlet Letter*, by Nathaniel Hawthorne
*Dead Souls*, by Nikolai Gogol
*1984*, by George Orwell [tied for 5th]
*Green Henry*, by Gottfried Keller [tied for 5th]
*Something I've Been Meaning To Tell You*, by Alice Munro
*Silas Marner*, by George Eliot
*Middlemarch*, by George Eliot
*Beggars in Spain*, by Nancy Kress

*(Image is from bookride.com)*

This list is not at all intuitive, but it does suggest what we are capturing with this metric. These are all stories that combine surprisingly disparate elements. *Brighton Rock,* which I'd never heard of before, is a 1938 murder mystery with theological themes, that generated some controversy over its anti-semitism. It has extraordinary staying power in these analyses: it has been ranked near the top in almost every version of this metric I tried. (In my first attempt, *Angels in America* was at the top of the list, which makes a good deal of sense given its obvious diversity of themes: gay men, Mormons, Judaism, angels, and Reagan-era politics. It's now been bumped down the list a bit, but I've retained the emanations from *Angels* as the names of my variables.

What else can we say about this list? I'm happy about the even gender split, and the relatively strong presence of 19th-century works. It would have been nice to see some non-white authors up here (*Fences* was also in first place at one point, and Chinua Achebe's *Arrow of God* is very high-ranked). Presumably the tagging to date reflects my unconscious biases (and Zoë's), and we may be seeing that. When Sphinx begins talking to people other than myself, I am going to do a full-court press to stratify the invitations, in hopes of counteracting that bias.

If we delve further down the list, we encounter people, and then finally objects and abstract concepts—Sphinx seems to have a penchant for different kinds of accordions, and unstable metals like Bohrium and Darmstadtium. After three generations, the list has changed considerably, and includes some entries that seem definitely out of place; probably they are artifacts of Sphinx' sparse data:

the American Farmland Trust
anti-Tom texts
*Brighton Rock*, by Graham Greene
*The Claw of the Conciliator,* by Gene Wolf
*The Audubon Field Guide to North American Mushrooms*
*Dead Souls*, by Nikolai Gogol

*The Big Outside,* by David Foreman
adventure
books of photography
*Daniel Deronda,* by George Eliot

And by the fourth generation there are no texts in the top ten. Tangible and abstract concepts, though slow to leave the gate, are now dominating. And we can see a preference for interdisciplinary concepts and anomalies, both of which makes sense:

anti-Semitic stereotypes
Anglican priests
the US Freedom of Information Act
the English monarchy
holistic health
the Manhattan project
bioweapons
night
landscaping in order to attract wildlife
the Spratly islands

Despite the oddity of these lists, I feel like the phosphor metric is *getting at* something that corresponds to at least one aspect of what we mean by "interesting". And given that, I appreciate that phosphor scores are counterintuitive enough to be surprising. Why accordions—is it because of the keyboards? Why the *Beggars* trilogy? Why night? Why are apple trees so much more interesting to Sphinx than ginger roots? Why all the George Eliot novels?

The database is still small enough (!) that I can go in with a pair of tweezers and try to answer those questions. (Yeah, it's because of the keyboards.) But much of the point here is to create a system by which Sphinx can individuate and surprise us, and phosphor certainly accomplishes that. It also hangs together in terms of its own logic: Sphinx is fascinated by anti-Semitic stereotypes, and it makes sense that the book it is most curious about would have those. Occasionally, breaks in this logic help me highlight defective concepts. For instance, Sphinx placed an enormously high phosphor value on baroque slide trumpets, which caught my attention: when I checked the database, **baroque_slide_trumpets* was corrupted by links from **finland,* making them the only brass instrument with inland waterways and a non-Indo-European language. And this would indeed be very interesting, if it was correct...

So now we have fluor and three different layers of phosphor. These get merged into an overall score, lux*,* which is simply the geometric mean of all the others. But to get there, we go through some dithering and diffusion, and it all happens concurrently (or in fact, iteratively):

## Lumen – Interests based on new knowledge

Phosphor is structured to counteract monomania, since it tends to decline as Sphinx gains more information about a concept. This does not always happen right away, though. A concept that can withstand added information without becoming less interesting is, *ipso facto,* an especially interesting concept. We flag that with lumen, which is essentially a (capped) measure of the upward slope of lux over recent time. Downward slopes and stasis are both treated as zero.

## Candle – Interests based on proximate concepts

Candle is a diffusion variable, based on a the lux values of concepts immediately downstream *and upstream* of the target, sampled if need be. Concepts are awarded a candle score proprtionately to how much *less interesting* they are than the concepts immediately around them. This creates a modest diffusion from more to less interesting concepts (though it doesn't subtract from the more interesting concepts).

Since candle affects lux, this diffusion can be passed along for multiple generations, diminishing proportionately each time, until it reaches an equilibrium.

## Some Paths Not Taken

I have tried a number of other metrics, and I've chosen the ones above for a variety of reasons, but there are some other important options to consider. The most prominent of these is functional desire, which seems to me different from curiosity. Humans often say (and Sphinx should be able to say) "I'm interested in studying pen-and-ink techniques so that I can draw a graphic novel." This is an important form of evaluating concepts, but I think it is quite distinct from what we are talking about, which is more a blend of recommendations and fuzzy pattern recognition. And notably, humans often say things like "I'm interested in learning how to wire an HVAC system, but it sounds really boring", which is somehow not necessarily a contradiction.

This has been a very interesting exercise. I'm not entirely certain of the parameters yet—I think, in particular, that it will take awhile to balance the coefficients between the different variables until the gestalt pattern makes sense. But it has already achieved its primary purpose: Sphinx now has a sort of protopersonality: minimal, to be sure, but coherent, non-arbitrary, and "internally" determined. Last week I had a few megabytes of raw data and code. Now I have a few megabytes of data and code that has a strong preference for George Eliot and accordions. And this preference was not intended, nor could have been realistically predicted, by any human being. Reductive and minimal as it is, that is the beginning of individuation.