2 Explore/Exploit

The Latest vs. the Greatest

Your stomach rumbles. Do you go to the Italian restaurant that you know and love, or the new Thai place that just opened up? Do you take your best friend, or reach out to a new acquaintance you'd like to get to know better? This is too hard—maybe you'll just stay home. Do you cook a recipe that you know is going to work, or scour the Internet for new inspiration? Never mind, how about you just order a pizza? Do you get your "usual," or ask about the specials? You're already exhausted before you get to the first bite. And the thought of putting on a record, watching a movie, or reading a book —*which one?*—no longer seems quite so relaxing.

Every day we are constantly forced to make decisions between options that differ in a very specific dimension: do we try new things or stick with our favorite ones? We intuitively understand that life is a balance between novelty and tradition, between the latest and the greatest, between taking risks and savoring what we know and love. But just as with the look-or-leap dilemma of the apartment hunt, the unanswered question is: what balance?

In the 1974 classic *Zen and the Art of Motorcycle Maintenance*, Robert Pirsig decries the conversational opener "What's new?"—arguing that the question, "if pursued exclusively, results only in an endless parade of trivia and fashion, the silt of tomorrow." He endorses an alternative as vastly superior: "What's best?"

But the reality is not so simple. Remembering that every "best" song and restaurant among your favorites began humbly as something merely "new" to you is a reminder that there may be yet-unknown bests still out there—and thus that the new is indeed worthy of at least some of our attention.

Age-worn aphorisms acknowledge this tension but don't solve it. "Make new friends, but keep the old / Those are silver, these are gold," and "There is no life so rich and rare / But one more friend could enter there" are true enough; certainly their scansion is unimpeachable. But they fail to tell us anything useful about the *ratio* of, say, "silver" and "gold" that makes the best alloy of a life well lived.

Computer scientists have been working on finding this balance for more than fifty years. They even have a name for it: the explore/exploit tradeoff.

Explore/Exploit

In English, the words "explore" and "exploit" come loaded with completely opposite connotations. But to a computer scientist, these words have much more specific and neutral meanings. Simply put, exploration is *gathering* information, and exploitation is *using* the information you have to get a known good result.

It's fairly intuitive that never exploring is no way to live. But it's also worth mentioning that never exploiting can be every bit as bad. In the computer science definition, exploitation actually comes to characterize many of what we consider to be life's best moments. A family gathering together on the holidays is exploitation. So is a bookworm settling into a reading chair with a hot cup of coffee and a beloved favorite, or a band playing their greatest hits to a crowd of adoring fans, or a couple that has stood the test of time dancing to "their song."

What's more, exploration can be a curse.

Part of what's nice about music, for instance, is that there are constantly new things to listen to. Or, if you're a music journalist, part of what's terrible about music is that there are *constantly* new things to listen to. Being a music journalist means turning the exploration dial all the way to 11, where it's nothing but new things all the time. Music lovers might imagine working in music journalism to be paradise, but when you constantly have to explore the new you can never enjoy the fruits of your connoisseurship—a particular kind of hell. Few people know this experience as deeply as Scott Plagenhoef, the former editor in chief of *Pitchfork*. "You try to find spaces when you're working to listen to something that you just want to listen to," he says of a critic's life. His desperate urges to stop wading through unheard tunes of dubious quality and just listen to what he loved were so strong that Plagenhoef would put only new music on his iPod, to make himself physically incapable of abandoning his duties in those moments when he just really, really, really wanted to listen to the Smiths. Journalists are martyrs, exploring so that others may exploit.

In computer science, the tension between exploration and exploitation takes its most concrete form in a scenario called the "multi-armed bandit problem." The odd name comes from the colloquial term for a casino slot machine, the "one-armed bandit." Imagine walking into a casino full of different slot machines, each one with its own odds of a payoff. The rub, of course, is that you aren't told those odds in advance: until you start playing, you won't have any idea which machines are the most lucrative ("loose," as slot-machine aficionados call it) and which ones are just money sinks.

Naturally, you're interested in maximizing your total winnings. And it's clear that this is going to involve some combination of pulling the arms on different machines to test them out (exploring), and favoring the most promising machines you've found (exploiting).

To get a sense for the problem's subtleties, imagine being faced with only two machines. One you've played a total of 15 times; 9 times it paid out, and 6 times it didn't. The other you've played only twice, and it once paid out and once did not. Which is more promising?

Simply dividing the wins by the total number of pulls will give you the machine's "expected value," and by this method the first machine clearly comes out ahead. Its 9–6 record makes for an expected value of 60%, whereas the second machine's 1–1 record yields an expected value of only 50%. But there's more to it than that. After all, just two pulls aren't really very many. So there's a sense in which we just don't yet *know* how good the second machine might actually be.

Choosing a restaurant or an album is, in effect, a matter of deciding which arm to pull in life's casino. But understanding the explore/exploit tradeoff isn't just a way to improve decisions about where to eat or what to listen to. It also provides fundamental insights into how our goals should change as we age, and why the most rational course of action isn't always trying to choose the best. And it turns out to be at the heart of, among other things, web design and clinical trials—two topics that normally aren't mentioned in the same sentence.

People tend to treat decisions in isolation, to focus on finding each time the outcome with the highest expected value. But decisions are almost never isolated, and expected value isn't the end of the story. If you're thinking not just about the *next* decision, but about *all* the decisions you are going to make about the same options in the future, the explore/exploit tradeoff is crucial to the process. In this way, writes mathematician Peter Whittle, the bandit problem "embodies in essential form a conflict evident in all human action."

So which of those two arms should you pull? It's a trick question. It completely depends on something we haven't discussed yet: how long you plan to be in the casino.

Seize the Interval

"Carpe diem," urges Robin Williams in one of the most memorable scenes of the 1989 film *Dead Poets Society*. "Seize the day, boys. Make your lives extraordinary."

It's incredibly important advice. It's also somewhat self-contradictory. Seizing a day and seizing a lifetime are two entirely different endeavors. We have the expression "Eat, drink, and be merry, for tomorrow we die," but perhaps we should also have its inverse: "Start learning a new language or an instrument, and make small talk with a stranger, because life is long, and who knows what joy could blossom over many years' time." When balancing favorite experiences and new ones, nothing matters as much as the interval over which we plan to enjoy them.

"I'm more likely to try a new restaurant when I move to a city than when I'm leaving it," explains data scientist and blogger Chris Stucchio, a veteran of grappling with the explore/exploit tradeoff in both his work and his life. "I mostly go to restaurants I know and love now, because I know I'm going to be leaving New York fairly soon. Whereas a couple years ago I moved to Pune, India, and I just would eat friggin' everywhere that didn't look like it was gonna kill me. And as I was leaving the city I went back to all my old favorites, rather than trying out new stuff.… Even if I find a slightly better place, I'm only going to go there once or twice, so why take the risk?"

A sobering property of trying new things is that the value of exploration, of finding a new favorite, can only go down over time, as the remaining opportunities to savor it dwindle. Discovering an enchanting café on your last night in town doesn't give you the opportunity to return.

The flip side is that the value of exploitation can only go *up* over time. The loveliest café that you know about today is, by definition, at least as lovely as the loveliest café you knew about last month. (And if you've found another favorite since then, it might just be more so.) So explore when you will have time to use the resulting knowledge, exploit when you're ready to cash in. The interval makes

the strategy.

Interestingly, since the interval makes the strategy, then by observing the strategy we can also infer the interval. Take Hollywood, for instance: Among the ten highest-grossing movies of 1981, only two were sequels. In 1991, it was three. In 2001, it was five. And in 2011, *eight* of the top ten highest-grossing films were sequels. In fact, 2011 set a record for the greatest percentage of sequels among major studio releases. Then 2012 immediately broke that record; the next year would break it again. In December 2012, journalist Nick Allen looked ahead with palpable fatigue to the year to come:

> Audiences will be given a sixth helping of X-Men plus *Fast and Furious 6*, *Die Hard 5*, *Scary Movie 5* and *Paranormal Activity 5*. There will also be *Iron Man 3*, *The Hangover 3*, and second outings for *The Muppets*, *The Smurfs*, *GI Joe* and *Bad Santa*.

From a studio's perspective, a sequel is a movie with a guaranteed fan base: a cash cow, a sure thing, an exploit. And an overload of sure things signals a short-termist approach, as with Stucchio on his way out of town. The sequels are more likely than brand-new movies to be hits this year, but where will the beloved franchises of the future come from? Such a sequel deluge is not only lamentable (certainly critics think so); it's also somewhat poignant. By entering an almost purely exploit-focused phase, the film industry seems to be signaling a belief that it is near the end of its interval.

A look into the economics of Hollywood confirms this hunch. Profits of the largest film studios declined by 40% between 2007 and 2011, and ticket sales have declined in seven of the past ten years. As the *Economist* puts it, "Squeezed between rising costs and falling revenues, the big studios have responded by trying to make more films they think will be hits: usually sequels, prequels, or anything featuring characters with name recognition." In other words, they're pulling the arms of the best machines they've got before the casino turns them out.

Win-Stay

Finding optimal algorithms that tell us exactly how to handle the multi-armed bandit problem has proven incredibly challenging. Indeed, as Peter Whittle recounts, during World War II efforts to solve the question "so sapped the energies and minds of Allied analysts … that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage."

The first steps toward a solution were taken in the years after the war, when Columbia mathematician Herbert Robbins showed that there's a simple strategy that, while not perfect, comes with some nice guarantees.

Robbins specifically considered the case where there are exactly two slot machines, and proposed a solution called the **Win-Stay, Lose-Shift** algorithm: choose an arm at random, and keep pulling it as long as it keeps paying off. If the arm doesn't pay off after a particular pull, then switch to the other one. Although this simple strategy is far from a complete solution, Robbins proved in 1952 that it performs reliably better than chance.

Following Robbins, a series of papers examined the "stay on a winner" principle further.

Intuitively, if you were already willing to pull an arm, and it has just paid off, that should only increase your estimate of its value, and you should be only more willing to pull it again. And indeed, win-stay turns out to be an element of the optimal strategy for balancing exploration and exploitation under a wide range of conditions.

But lose-shift is another story. Changing arms each time one fails is a pretty rash move. Imagine going to a restaurant a hundred times, each time having a wonderful meal. Would one disappointment be enough to induce you to give up on it? Good options shouldn't be penalized too strongly for being imperfect.

More significantly, Win-Stay, Lose-Shift doesn't have any notion of the interval over which you are optimizing. If your favorite restaurant disappointed you the last time you ate there, that algorithm always says you should go to another place—even if it's your last night in town.

Still, Robbins's initial work on the multi-armed bandit problem kicked off a substantial literature, and researchers made significant progress over the next few years. Richard Bellman, a mathematician at the RAND Corporation, found an exact solution to the problem for cases where we know in advance exactly how many options and opportunities we'll have in total. As with the full-information secretary problem, Bellman's trick was essentially to work backward, starting by imagining the final pull and considering which slot machine to choose given all the possible outcomes of the previous decisions. Having figured that out, you'd then turn to the second-to-last opportunity, then the previous one, and the one before that, all the way back to the start.

The answers that emerge from Bellman's method are ironclad, but with many options and a long casino visit it can require a dizzying—or impossible—amount of work. What's more, even if we are able to calculate all possible futures, we of course don't always know exactly how many opportunities (or even how many options) we'll have. For these reasons, the multi-armed bandit problem effectively stayed unsolved. In Whittle's words, "it quickly became a classic, and a byword for intransigence."

The Gittins Index

As so often happens in mathematics, though, the particular is the gateway to the universal. In the 1970s, the Unilever corporation asked a young mathematician named John Gittins to help them optimize some of their drug trials. Unexpectedly, what they got was the answer to a mathematical riddle that had gone unsolved for a generation.

Gittins, who is now a professor of statistics at Oxford, pondered the question posed by Unilever. Given several different chemical compounds, what is the quickest way to determine which compound is likely to be effective against a disease? Gittins tried to cast the problem in the most general form he could: multiple options to pursue, a different probability of reward for each option, and a certain amount of effort (or money, or time) to be allocated among them. It was, of course, another incarnation of the multi-armed bandit problem.

Both the for-profit drug companies and the medical profession they serve are constantly faced

with the competing demands of the explore/exploit tradeoff. Companies want to invest R & D money into the discovery of new drugs, but also want to make sure their profitable current product lines are flourishing. Doctors want to prescribe the best existing treatments so that patients get the care they need, but also want to encourage experimental studies that may turn up even better ones.

In both cases, notably, it's not entirely clear what the relevant interval ought to be. In a sense, drug companies and doctors alike are interested in the *indefinite* future. Companies want to be around theoretically forever, and on the medical side a breakthrough could go on to help people who haven't even been born yet. Nonetheless, the present has a higher priority: a cured patient today is taken to be more valuable than one cured a week or a year from now, and certainly the same holds true of profits. Economists refer to this idea, of valuing the present more highly than the future, as "discounting."

Unlike previous researchers, Gittins approached the multi-armed bandit problem in those terms. He conceived the goal as maximizing payoffs not for a fixed interval of time, but for a future that is endless yet discounted.

Such discounting is not unfamiliar to us from our own lives. After all, if you visit a town for a ten-day vacation, then you should be making your restaurant decisions with a fixed interval in mind; but if you live in the town, this doesn't make as much sense. Instead, you might imagine the value of payoffs decreasing the further into the future they are: you care more about the meal you're going to eat tonight than the meal you're going to eat tomorrow, and more about tomorrow's meal than one a year from now, with the specifics of how much more depending on your particular "discount function." Gittins, for his part, made the assumption that the value assigned to payoffs decreases geometrically: that is, each restaurant visit you make is worth a constant fraction of the last one. If, let's say, you believe there is a 1% chance you'll get hit by a bus on any given day, then you should value tomorrow's dinner at 99% of the value of tonight's, if only because you might never get to eat it.

Working with this geometric-discounting assumption, Gittins investigated a strategy that he thought "at least would be a pretty good approximation": to think about each arm of the multi-armed bandit separately from the others, and try to work out the value of that arm on its own. He did this by imagining something rather ingenious: a bribe.

In the popular television game show *Deal or No Deal*, a contestant chooses one of twenty-six briefcases, which contain prizes ranging from a penny to a million dollars. As the game progresses, a mysterious character called the Banker will periodically call in and offer the contestant various sums of money to *not* open the chosen briefcase. It's up to the contestant to decide at what price they're willing to take a sure thing over the uncertainty of the briefcase prize.

Gittins (albeit many years before the first episode of *Deal or No Deal* aired) realized that the multi-armed bandit problem is no different. For every slot machine we know little or nothing about, there is some guaranteed payout rate which, if offered to us in lieu of that machine, will make us quite content never to pull its handle again. This number—which Gittins called the "dynamic allocation index," and which the world now knows as the **Gittins index**—suggests an obvious strategy on the casino floor: always play the arm with the highest index.*

In fact, the index strategy turned out to be more than a good approximation. It completely solves the multi-armed bandit with geometrically discounted payoffs. The tension between exploration and

exploitation resolves into the simpler task of maximizing a single quantity that accounts for both. Gittins is modest about the achievement—"It's not quite Fermat's Last Theorem," he says with a chuckle—but it's a theorem that put to rest a significant set of questions about the explore/exploit dilemma.

Now, actually calculating the Gittins index for a specific machine, given its track record and our discounting rate, is still fairly involved. But once the Gittins index for a particular set of assumptions is known, it can be used for any problem of that form. Crucially, it doesn't even matter how many arms are involved, since the index for each arm is calculated separately.

In the table on the next page we provide the Gittins index values for up to nine successes and failures, assuming that a payoff on our next pull is worth 90% of a payoff now. These values can be used to resolve a variety of everyday multi-armed bandit problems. For example, under these assumptions you should, in fact, choose the slot machine that has a track record of 1–1 (and an expected value of 50%) over the one with a track record of 9–6 (and an expected value of 60%). Looking up the relevant coordinates in the table shows that the lesser-known machine has an index of 0.6346, while the more-played machine scores only a 0.6300. Problem solved: try your luck this time, and explore.

Looking at the Gittins index values in the table, there are a few other interesting observations. First, you can see the win-stay principle at work: as you go from left to right in any row, the index scores always increase. So if an arm is ever the correct one to pull, and that pull is a winner, then (following the chart to the right) it can only make more sense to pull the same arm again. Second, you can see where lose-shift would get you into trouble. Having nine initial wins followed by a loss gets you an index of 0.8695, which is still higher than most of the other values in the table—so you should probably stay with that arm for at least another pull.

	Wins									
Losses	0	1	2	3	4	5	6	7	8	9
0	.7029	.8001	.8452	.8723	.8905	.9039	.9141	.9221	.9287	.9342
1	.5001	.6346	.7072	.7539	.7869	.8115	.8307	.8461	.8588	.8695
2	.3796	.5163	.6010	.6579	.6996	.7318	.7573	.7782	.7956	.8103
3	.3021	.4342	.5184	.5809	.6276	.6642	.6940	.7187	.7396	.7573
4	.2488	.3720	.4561	.5179	.5676	.6071	.6395	.6666	.6899	.7101
5	.2103	.3245	.4058	.4677	.5168	.5581	.5923	.6212	.6461	.6677
6	.1815	.2871	.3647	.4257	.4748	.5156	.5510	.5811	.6071	.6300
7	.1591	.2569	.3308	.3900	.4387	.4795	.5144	.5454	.5723	.5960
8	.1413	.2323	.3025	.3595	.4073	.4479	.4828	.5134	.5409	.5652
9	.1269	.2116	.2784	.3332	.3799	.4200	.4548	.4853	.5125	.5373

Gittins index values as a function of wins and losses, assuming that a payoff next time is worth 90% of a payoff now.

But perhaps the most interesting part of the table is the top-left entry. A record of 0–0—an arm that's a complete unknown—has an expected value of 0.5000 but a Gittins index of 0.7029. In other

words, something you have no experience with whatsoever is more attractive than a machine that you know pays out seven times out of ten! As you go down the diagonal, notice that a record of 1–1 yields an index of 0.6346, a record of 2–2 yields 0.6010, and so on. If such 50%-successful performance persists, the index does ultimately converge on 0.5000, as experience confirms that the machine is indeed nothing special and takes away the "bonus" that spurs further exploration. But the convergence happens fairly slowly; the exploration bonus is a powerful force. Indeed, note that even a failure on the very first pull, producing a record of 0–1, makes for a Gittins index that's still above 50%.

We can also see how the explore/exploit tradeoff changes as we change the way we're discounting the future. The following table presents exactly the same information as the preceding one, but assumes that a payoff next time is worth 99% of one now, rather than 90%. With the future weighted nearly as heavily as the present, the value of making a chance discovery, relative to taking a sure thing, goes up even more. Here, a totally untested machine with a 0–0 record is worth a guaranteed 86.99% chance of a payout!

Losses \ Wins	0	1	2	3	4	5	6	7	8	9
0	.8699	.9102	.9285	.9395	.9470	.9525	.9568	.9603	.9631	.9655
1	.7005	.7844	.8268	.8533	.8719	.8857	.8964	.9051	.9122	.9183
2	.5671	.6726	.7308	.7696	.7973	.8184	.8350	.8485	.8598	.8693
3	.4701	.5806	.6490	.6952	.7295	.7561	.7773	.7949	.8097	.8222
4	.3969	.5093	.5798	.6311	.6697	.6998	.7249	.7456	.7631	.7781
5	.3415	.4509	.5225	.5756	.6172	.6504	.6776	.7004	.7203	.7373
6	.2979	.4029	.4747	.5277	.5710	.6061	.6352	.6599	.6811	.6997
7	.2632	.3633	.4337	.4876	.5300	.5665	.5970	.6230	.6456	.6653
8	.2350	.3303	.3986	.4520	.4952	.5308	.5625	.5895	.6130	.6337
9	.2117	.3020	.3679	.4208	.4640	.5002	.5310	.5589	.5831	.6045

Gittins index values as a function of wins and losses, assuming that a payoff next time is worth 99% of a payoff now.

The Gittins index, then, provides a formal, rigorous justification for preferring the unknown, provided we have some opportunity to exploit the results of what we learn from exploring. The old adage tells us that "the grass is always greener on the other side of the fence," but the math tells us why: the unknown has a chance of being better, even if we actually expect it to be no different, or if it's just as likely to be worse. The untested rookie is worth more (early in the season, anyway) than the veteran of seemingly equal ability, precisely because we know less about him. Exploration in itself has value, since trying new things increases our chances of finding the best. So taking the future into account, rather than focusing just on the present, drives us toward novelty.

The Gittins index thus provides an amazingly straightforward solution to the multi-armed bandit problem. But it doesn't necessarily close the book on the puzzle, or help us navigate *all* the explore/exploit tradeoffs of everyday life. For one, the Gittins index is optimal only under some strong assumptions. It's based on geometric discounting of future reward, valuing each pull at a

constant fraction of the previous one, which is something that a variety of experiments in behavioral economics and psychology suggest people don't do. And if there's a cost to switching among options, the Gittins strategy is no longer optimal either. (The grass on the other side of the fence may look a bit greener, but that doesn't necessarily warrant climbing the fence—let alone taking out a second mortgage.) Perhaps even more importantly, it's hard to compute the Gittins index on the fly. If you carry around a table of index values you can optimize your dining choices, but the time and effort involved might not be worth it. ("Wait, I can resolve this argument. That restaurant was good 29 times out of 35, but this other one has been good 13 times out of 16, so the Gittins indices are … Hey, where did everybody go?")

In the time since the development of the Gittins index, such concerns have sent computer scientists and statisticians searching for simpler and more flexible strategies for dealing with multi-armed bandits. These strategies are easier for humans (and machines) to apply in a range of situations than crunching the optimal Gittins index, while still providing comparably good performance. They also engage with one of our biggest human fears regarding decisions about which chances to take.

Regret and Optimism

> *Regrets, I've had a few. But then again, too few to mention.*
>
> —FRANK SINATRA

> *For myself I am an optimist. It does not seem to be much use being anything else.*
>
> —WINSTON CHURCHILL

If the Gittins index is too complicated, or if you're not in a situation well characterized by geometric discounting, then you have another option: focus on *regret*. When we choose what to eat, who to spend time with, or what city to live in, regret looms large—presented with a set of good options, it is easy to torture ourselves with the consequences of making the wrong choice. These regrets are often about the things we failed to do, the options we never tried. In the memorable words of management theorist Chester Barnard, "To try and fail is at least to learn; to fail to try is to suffer the inestimable loss of what might have been."

Regret can also be highly motivating. Before he decided to start Amazon.com, Jeff Bezos had a secure and well-paid position at the investment company D. E. Shaw & Co. in New York. Starting an online bookstore in Seattle was going to be a big leap—something that his boss (that's D. E. Shaw) advised him to think about carefully. Says Bezos:

> The framework I found, which made the decision incredibly easy, was what I called—which only a nerd would call—a "regret minimization framework." So I wanted to project myself forward to age 80 and say, "Okay, now I'm looking back on my life. I want to have minimized the number of regrets I have." I knew that when I was 80 I was not going to regret having tried this. I was not going to regret trying to participate in this thing called the Internet that I thought was going to be a really big deal. I knew that if I failed I wouldn't regret that, but I knew the one thing I might regret is not ever having tried. I knew that that would haunt me every day, and so, when I thought about it that way it was an incredibly easy decision.

Computer science can't offer you a life with no regret. But it can, potentially, offer you just what Bezos was looking for: a life with *minimal* regret.

Regret is the result of comparing what we actually did with what would have been best in hindsight. In a multi-armed bandit, Barnard's "inestimable loss" can in fact be measured precisely, and regret assigned a number: it's the difference between the total payoff obtained by following a particular strategy and the total payoff that theoretically could have been obtained by just pulling the best arm every single time (had we only known from the start which one it was). We can calculate this number for different strategies, and search for those that minimize it.

In 1985, Herbert Robbins took a second shot at the multi-armed bandit problem, some thirty years after his initial work on Win-Stay, Lose-Shift. He and fellow Columbia mathematician Tze Leung Lai were able to prove several key points about regret. First, assuming you're not omniscient, your total amount of regret will probably never stop increasing, even if you pick the best possible strategy—because even the best strategy isn't perfect every time. Second, regret will increase at a slower rate if you pick the best strategy than if you pick others; what's more, with a good strategy regret's rate of growth will go down over time, as you learn more about the problem and are able to make better choices. Third, and most specifically, the minimum possible regret—again assuming non-omniscience—is regret that increases at a *logarithmic* rate with every pull of the handle.

Logarithmically increasing regret means that we'll make as many mistakes in our first ten pulls as in the following ninety, and as many in our first year as in the rest of the decade combined. (The first decade's mistakes, in turn, are as many as we'll make for the rest of the century.) That's some measure of consolation. In general we can't realistically expect someday to never have any more regrets. But if we're following a regret-minimizing algorithm, every year we can expect to have fewer new regrets than we did the year before.

Starting with Lai and Robbins, researchers in recent decades have set about looking for algorithms that offer the guarantee of minimal regret. Of the ones they've discovered, the most popular are known as **Upper Confidence Bound** algorithms.

Visual displays of statistics often include so-called error bars that extend above and below any data point, indicating uncertainty in the measurement; the error bars show the range of plausible values that the quantity being measured could actually have. This range is known as the "confidence interval," and as we gain more data about something the confidence interval will shrink, reflecting an increasingly accurate assessment. (For instance, a slot machine that has paid out once out of two pulls will have a wider confidence interval, though the same expected value, as a machine that has paid out five times on ten pulls.) In a multi-armed bandit problem, an Upper Confidence Bound algorithm says, quite simply, to pick the option for which the top of the confidence interval is highest.

Like the Gittins index, therefore, Upper Confidence Bound algorithms assign a single number to each arm of the multi-armed bandit. And that number is set to the highest value that the arm could reasonably have, based on the information available so far. So an Upper Confidence Bound algorithm doesn't care which arm *has* performed best so far; instead, it chooses the arm that *could* reasonably perform best in the future. If you have never been to a restaurant before, for example, then for all you know it could be great. Even if you have gone there once or twice, and tried a couple of their dishes,

you might not have enough information to rule out the possibility that it could yet prove better than your regular favorite. Like the Gittins index, the Upper Confidence Bound is always greater than the expected value, but by less and less as we gain more experience with a particular option. (A restaurant with a single mediocre review still retains a *potential* for greatness that's absent in a restaurant with hundreds of such reviews.) The recommendations given by Upper Confidence Bound algorithms will be similar to those provided by the Gittins index, but they are significantly easier to compute, and they don't require the assumption of geometric discounting.

Upper Confidence Bound algorithms implement a principle that has been dubbed "optimism in the face of uncertainty." Optimism, they show, can be perfectly rational. By focusing on the best that an option *could* be, given the evidence obtained so far, these algorithms give a boost to possibilities we know less about. As a consequence, they naturally inject a dose of exploration into the decision-making process, leaping at new options with enthusiasm because any one of them could be the next big thing. The same principle has been used, for instance, by MIT's Leslie Kaelbling, who builds "optimistic robots" that explore the space around them by boosting the value of uncharted terrain. And it clearly has implications for human lives as well.

The success of Upper Confidence Bound algorithms offers a formal justification for the benefit of the doubt. Following the advice of these algorithms, you should be excited to meet new people and try new things—to assume the best about them, in the absence of evidence to the contrary. In the long run, optimism is the best prevention for regret.

Bandits Online

In 2007, Google product manager Dan Siroker took a leave of absence to join the presidential campaign of then senator Barack Obama in Chicago. Heading the "New Media Analytics" team, Siroker brought one of Google's web practices to bear on the campaign's bright-red DONATE button. The result was nothing short of astonishing: $57 million of additional donations were raised as a direct result of his work.

What exactly did he do to that button?

He A/B tested it.

A/B testing works as follows: a company drafts several different versions of a particular webpage. Perhaps they try different colors or images, or different headlines for a news article, or different arrangements of items on the screen. Then they randomly assign incoming users to these various pages, usually in equal numbers. One user may see a red button, while another user may see a blue one; one may see DONATE and another may see CONTRIBUTE. The relevant metrics (e.g., click-through rate or average revenue per visitor) are then monitored. After a period of time, if statistically significant effects are observed, the "winning" version is typically locked into place—or becomes the control for another round of experiments.

In the case of Obama's donation page, Siroker's A/B tests were revealing. For first-time visitors to the campaign site, a DONATE AND GET A GIFT button turned out to be the best performer, even after

the cost of sending the gifts was taken into account. For longtime newsletter subscribers who had never given money, PLEASE DONATE worked the best, perhaps appealing to their guilt. For visitors who had already donated in the past, CONTRIBUTE worked best at securing follow-up donations—the logic being perhaps that the person had already "donated" but could always "contribute" more. And in all cases, to the astonishment of the campaign team, a simple black-and-white photo of the Obama family outperformed any other photo or video the team could come up with. The net effect of all these independent optimizations was gigantic.

If you've used the Internet basically at all over the past decade, then you've been a part of someone else's explore/exploit problem. Companies want to discover the things that make them the most money while simultaneously making as much of it as they can—explore, exploit. Big tech firms such as Amazon and Google began carrying out live A/B tests on their users starting in about 2000, and over the following years the Internet has become the world's largest controlled experiment. What are these companies exploring and exploiting? In a word, you: whatever it is that makes you move your mouse and open your wallet.

Companies A/B test their site navigation, the subject lines and timing of their marketing emails, and sometimes even their actual features and pricing. Instead of "the" Google search algorithm and "the" Amazon checkout flow, there are now untold and unfathomably subtle permutations. (Google infamously tested forty-one shades of blue for one of its toolbars in 2009.) In some cases, it's unlikely that any pair of users will have the exact same experience.

Data scientist Jeff Hammerbacher, former manager of the Data group at Facebook, once told *Bloomberg Businessweek* that "the best minds of my generation are thinking about how to make people click ads." Consider it the millennials' *Howl*—what Allen Ginsberg's immortal "I saw the best minds of my generation destroyed by madness" was to the Beat Generation. Hammerbacher's take on the situation was that this state of affairs "sucks." But regardless of what one makes of it, the web is allowing for an experimental science of the click the likes of which had never even been dreamed of by marketers of the past.

We know what happened to Obama in the 2008 election, of course. But what happened to his director of analytics, Dan Siroker? After the inauguration, Siroker returned west, to California, and with fellow Googler Pete Koomen co-founded the website optimization firm Optimizely. By the 2012 presidential election cycle, their company counted among its clients both the Obama re-election campaign *and* the campaign of Republican challenger Mitt Romney.

Within a decade or so after its first tentative use, A/B testing was no longer a secret weapon. It has become such a deeply embedded part of how business and politics are conducted online as to be effectively taken for granted. The next time you open your browser, you can be sure that the colors, images, text, perhaps even the prices you see—and certainly the ads—have come from an explore/exploit algorithm, tuning itself to your clicks. In this particular multi-armed bandit problem, you're not the gambler; you're the jackpot.

The process of A/B testing itself has become increasingly refined over time. The most canonical A/B setup—splitting the traffic evenly between two options, running the test for a set period of time, and thereafter giving all the traffic to the winner—might not necessarily be the best algorithm for

solving the problem, since it means half the users are stuck getting the inferior option as long as the test continues. And the rewards for finding a better approach are potentially very high. More than 90% of Google's approximately $50 billion in annual revenue currently comes from paid advertising, and online commerce comprises hundreds of billions of dollars a year. This means that explore/exploit algorithms effectively power, both economically and technologically, a significant fraction of the Internet itself. The best algorithms to use remain hotly contested, with rival statisticians, engineers, and bloggers endlessly sparring about the optimal way to balance exploration and exploitation in every possible business scenario.

Debating the precise distinctions among various takes on the explore/exploit problem may seem hopelessly arcane. In fact, these distinctions turn out to matter immensely—and it's not just presidential elections and the Internet economy that are at stake.

It's also human lives.

Clinical Trials on Trial

Between 1932 and 1972, several hundred African-American men with syphilis in Macon County, Alabama, went deliberately untreated by medical professionals, as part of a forty-year experiment by the US Public Health Service known as the Tuskegee Syphilis Study. In 1966, Public Health Service employee Peter Buxtun filed a protest. He filed a second protest in 1968. But it was not until he broke the story to the press—it appeared in the *Washington Star* on July 25, 1972, and was the front-page story in the *New York Times* the next day—that the US government finally halted the study.

What followed the public outcry, and the subsequent congressional hearing, was an initiative to formalize the principles and standards of medical ethics. A commission held at the pastoral Belmont Conference Center in Maryland resulted in a 1979 document known as the Belmont Report. The Belmont Report lays out a foundation for the ethical practice of medical experiments, so that the Tuskegee experiment—an egregious, unambiguously inappropriate breach of the health profession's duty to its patients—might never be repeated. But it also notes the difficulty, in many other cases, of determining exactly where the line should be drawn.

"The Hippocratic maxim 'do no harm' has long been a fundamental principle of medical ethics," the report points out. "[The physiologist] Claude Bernard extended it to the realm of research, saying that one should not injure one person regardless of the benefits that might come to others. However, even avoiding harm requires learning what is harmful; and, in the process of obtaining this information, persons may be exposed to risk of harm."

The Belmont Report thus acknowledges, but does not resolve, the tension that exists between acting on one's best knowledge and gathering more. It also makes it clear that gathering knowledge can be so valuable that some aspects of normal medical ethics can be suspended. Clinical testing of new drugs and treatments, the report notes, often requires risking harm to some patients, even if steps are taken to minimize that risk.

> The principle of beneficence is not always so unambiguous. A difficult ethical problem remains, for example, about research [on childhood diseases] that presents more than minimal risk without immediate prospect of direct benefit to the children involved. Some have argued that such research is inadmissible, while others have pointed out that this limit would rule out much research promising great benefit to children in the future. Here again, as with all hard cases, the different claims covered by the principle of beneficence may come into conflict and force difficult choices.

One of the fundamental questions that has arisen in the decades since the Belmont Report is whether the standard approach to conducting clinical trials really does minimize risk to patients. In a conventional clinical trial, patients are split into groups, and each group is assigned to receive a different treatment for the duration of the study. (Only in exceptional cases does a trial get stopped early.) This procedure focuses on decisively resolving the question of which treatment is better, rather than on providing the best treatment to each patient in the trial itself. In this way it operates exactly like a website's A/B test, with a certain fraction of people receiving an experience during the experiment that will eventually be proven inferior. But doctors, like tech companies, are gaining some information about which option is better *while* the trial proceeds—information that could be used to improve outcomes not only for future patients beyond the trial, but for the patients currently in it.

Millions of dollars are at stake in experiments to find the optimal configuration of a website, but in clinical trials, experimenting to find optimal treatments has direct life-or-death consequences. And a growing community of doctors and statisticians think that we're doing it wrong: that we should be treating the selection of treatments as a multi-armed bandit problem, and trying to get the better treatments to people even while an experiment is in progress.

In 1969, Marvin Zelen, a biostatistician who is now at Harvard, proposed conducting "adaptive" trials. One of the ideas he suggested was a randomized "play the winner" algorithm—a version of Win-Stay, Lose-Shift, in which the chance of using a given treatment is increased by each win and decreased by each loss. In Zelen's procedure, you start with a hat that contains one ball for each of the two treatment options being studied. The treatment for the first patient is selected by drawing a ball at random from the hat (the ball is put back afterward). If the chosen treatment is a success, you put another ball for that treatment into the hat—now you have three balls, two of which are for the successful treatment. If it fails, then you put another ball for the *other* treatment into the hat, making it more likely you'll choose the alternative.

Zelen's algorithm was first used in a clinical trial sixteen years later, for a study of extracorporeal membrane oxygenation, or "ECMO"—an audacious approach to treating respiratory failure in infants. Developed in the 1970s by Robert Bartlett of the University of Michigan, ECMO takes blood that's heading for the lungs and routes it instead out of the body, where it is oxygenated by a machine and returned to the heart. It is a drastic measure, with risks of its own (including the possibility of embolism), but it offered a possible approach in situations where no other options remained. In 1975 ECMO saved the life of a newborn girl in Orange County, California, for whom even a ventilator was not providing enough oxygen. That girl has now celebrated her fortieth birthday and is married with children of her own. But in its early days the ECMO technology and procedure were considered highly experimental, and early studies in adults showed no benefit compared to conventional treatments.

From 1982 to 1984, Bartlett and his colleagues at the University of Michigan performed a study on newborns with respiratory failure. The team was clear that they wanted to address, as they put it, "the ethical issue of withholding an unproven but potentially lifesaving treatment," and were "reluctant to withhold a lifesaving treatment from alternate patients simply to meet conventional random assignment technique." Hence they turned to Zelen's algorithm. The strategy resulted in one infant being assigned the "conventional" treatment and dying, and eleven infants in a row being assigned the experimental ECMO treatment, all of them surviving. Between April and November of 1984, after the end of the official study, ten additional infants met the criteria for ECMO treatment. Eight were treated with ECMO, and all eight survived. Two were treated conventionally, and both died.

These are eye-catching numbers, yet shortly after the University of Michigan study on ECMO was completed it became mired in controversy. Having so few patients in a trial receive the conventional treatment deviated significantly from standard methodology, and the procedure itself was highly invasive and potentially risky. After the publication of the paper, Jim Ware, professor of biostatistics at the Harvard School of Public Health, and his medical colleagues examined the data carefully and concluded that they "did not justify routine use of ECMO without further study." So Ware and his colleagues designed a second clinical trial, still trying to balance the acquisition of knowledge with the effective treatment of patients but using a less radical design. They would randomly assign patients to either ECMO or the conventional treatment until a prespecified number of deaths was observed in one of the groups. Then they would switch all the patients in the study to the more effective treatment of the two.

In the first phase of Ware's study, four of ten infants receiving conventional treatment died, and all nine of nine infants receiving ECMO survived. The four deaths were enough to trigger a transition to the second phase, where all twenty patients were treated with ECMO and nineteen survived. Ware and colleagues were convinced, concluding that "it is difficult to defend further randomization ethically."

But some had already concluded this *before* the Ware study, and were vocal about it. The critics included Don Berry, one of the world's leading experts on multi-armed bandits. In a comment that was published alongside the Ware study in *Statistical Science*, Berry wrote that "randomizing patients to non-ECMO therapy as in the Ware study was unethical.… In my view, the Ware study should not have been conducted."

And yet even the Ware study was not conclusive for all in the medical community. In the 1990s yet another study on ECMO was conducted, enrolling nearly two hundred infants in the United Kingdom. Instead of using adaptive algorithms, this study followed the traditional methods, splitting the infants randomly into two equal groups. The researchers justified the experiment by saying that ECMO's usefulness "is controversial because of varying interpretation of the available evidence." As it turned out, the difference between the treatments wasn't as pronounced in the United Kingdom as it had been in the two American studies, but the results were nonetheless declared "in accord with the earlier preliminary findings that a policy of ECMO support reduces the risk of death." The cost of that knowledge? Twenty-four more infants died in the "conventional" group than in the group receiving ECMO treatment.

The widespread difficulty with accepting results from adaptive clinical trials might seem incomprehensible. But consider that part of what the advent of statistics did for medicine, at the start of the twentieth century, was to transform it from a field in which doctors had to persuade each other in ad hoc ways about every new treatment into one where they had clear guidelines about what sorts of evidence were and were not persuasive. Changes to accepted standard statistical practice have the potential to upset this balance, at least temporarily.

After the controversy over ECMO, Don Berry moved from the statistics department at the University of Minnesota to the MD Anderson Cancer Center in Houston, where he has used methods developed by studying multi-armed bandits to design clinical trials for a variety of cancer treatments. While he remains one of the more vocal critics of randomized clinical trials, he is by no means the only one. In recent years, the ideas he's been fighting for are finally beginning to come into the mainstream. In February 2010, the FDA released a "guidance" document, "Adaptive Design Clinical Trials for Drugs and Biologics," which suggests—despite a long history of sticking to an option they trust—that they might at last be willing to explore alternatives.

The Restless World

Once you become familiar with them, it's easy to see multi-armed bandits just about everywhere we turn. It's rare that we make an isolated decision, where the outcome doesn't provide us with any information that we'll use to make other decisions in the future. So it's natural to ask, as we did with optimal stopping, how well people generally tend to solve these problems—a question that has been extensively explored in the laboratory by psychologists and behavioral economists.

In general, it seems that people tend to over-explore—to favor the new disproportionately over the best. In a simple demonstration of this phenomenon, published in 1966, Amos Tversky and Ward Edwards conducted experiments where people were shown a box with two lights on it and told that each light would turn on a fixed (but unknown) percentage of the time. They were then given 1,000 opportunities either to observe which light came on, or to place a bet on the outcome without getting to observe it. (Unlike a more traditional bandit problem setup, here one could not make a "pull" that was both wager and observation at once; participants would not learn whether their bets had paid off until the end.) This is pure exploration vs. exploitation, pitting the gaining of information squarely against the use of it. For the most part, people adopted a sensible strategy of observing for a while, then placing bets on what seemed like the best outcome—but they consistently spent a lot more time observing than they should have. How much more time? In one experiment, one light came on 60% of the time and the other 40% of the time, a difference neither particularly blatant nor particularly subtle. In that case, people chose to observe 505 times, on average, placing bets the other 495 times. But the math says they should have started to bet after just 38 observations—leaving 962 chances to cash in.

Other studies have produced similar conclusions. In the 1990s, Robert Meyer and Yong Shi, researchers at Wharton, ran a study where people were given a choice between two options, one with a known payoff chance and one unknown—specifically two airlines, an established carrier with a

known on-time rate and a new company without a track record yet. Given the goal of maximizing the number of on-time arrivals over some period of time, the mathematically optimal strategy is to initially only fly the new airline, as long as the established one isn't clearly better. If at any point it's apparent that the well-known carrier is better—that is, if the Gittins index of the new option falls below the on-time rate of the familiar carrier—then you should switch hard to the familiar one and never look back. (Since in this setup you can't get any more information about the new company once you stop flying it, there is no opportunity for it to redeem itself.) But in the experiment, people tended to use the untried airline too little when it was good and too much when it was bad. They also didn't make clean breaks away from it, often continuing to alternate, particularly when neither airline was departing on time. All of this is consistent with tending to over-explore.

Finally, psychologists Mark Steyvers, Michael Lee, and E.-J. Wagenmakers have run an experiment with a four-armed bandit, asking a group of people to choose which arm to play over a sequence of fifteen opportunities. They then classified the strategies that participants seemed to use. The results suggested that 30% were closest to the optimal strategy, 47% most resembled Win-Stay, Lose-Shift, and 22% seemed to move at random between selecting a new arm and playing the best arm found so far. Again, this is consistent with over-exploring, as Win-Stay, Lose-Shift and occasionally trying an arm at random are both going to lead people to try things other than the best option late in the game, when they should be purely exploiting.

So, while we tend to commit to a new secretary too soon, it seems like we tend to stop trying new airlines too late. But just as there's a cost to not having a secretary, there's a cost to committing too soon to a particular airline: the world might change.

The standard multi-armed bandit problem assumes that the probabilities with which the arms pay off are fixed over time. But that's not necessarily true of airlines, restaurants, or other contexts in which people have to make repeated choices. If the probabilities of a payoff on the different arms change over time—what has been termed a "restless bandit"—the problem becomes much harder. (So much harder, in fact, that there's no tractable algorithm for completely solving it, and it's believed there never will be.) Part of this difficulty is that it is no longer simply a matter of exploring for a while and then exploiting: when the world can change, continuing to explore can be the right choice. It might be worth going back to that disappointing restaurant you haven't visited for a few years, just in case it's under new management.

In his celebrated essay "Walking," Henry David Thoreau reflected on how he preferred to do his traveling close to home, how he never tired of his surroundings and always found something new or surprising in the Massachusetts landscape. "There is in fact a sort of harmony discoverable between the capabilities of the landscape within a circle of ten miles' radius, or the limits of an afternoon walk, and the threescore years and ten of human life," he wrote. "It will never become quite familiar to you."

To live in a restless world requires a certain restlessness in oneself. So long as things continue to change, you must never fully cease exploring.

Still, the algorithmic techniques honed for the standard version of the multi-armed bandit problem are useful even in a restless world. Strategies like the Gittins index and Upper Confidence Bound

provide reasonably good approximate solutions and rules of thumb, particularly if payoffs don't change very much over time. And many of the world's payoffs are arguably more static today than they've ever been. A berry patch might be ripe one week and rotten the next, but as Andy Warhol put it, "A Coke is a Coke." Having instincts tuned by evolution for a world in constant flux isn't necessarily helpful in an era of industrial standardization.

Perhaps most importantly, thinking about versions of the multi-armed bandit problem that do have optimal solutions doesn't just offer algorithms, it also offers insights. The conceptual vocabulary derived from the classical form of the problem—the tension of explore/exploit, the importance of the interval, the high value of the 0–0 option, the minimization of regret—gives us a new way of making sense not only of specific problems that come before us, but of the entire arc of human life.

Explore …

While laboratory studies can be illuminating, the interval of many of the most important problems people face is far too long to be studied in the lab. Learning the structure of the world around us and forming lasting social relationships are both lifelong tasks. So it's instructive to see how the general pattern of early exploration and late exploitation appears over the course of a lifetime.

One of the curious things about human beings, which any developmental psychologist aspires to understand and explain, is that we take years to become competent and autonomous. Caribou and gazelles must be prepared to run from predators the day they're born, but humans take more than a year to make their first steps. Alison Gopnik, professor of developmental psychology at UC Berkeley and author of *The Scientist in the Crib*, has an explanation for why human beings have such an extended period of dependence: "it gives you a developmental way of solving the exploration/exploitation tradeoff." As we have seen, good algorithms for playing multi-armed bandits tend to explore more early on, exploiting the resulting knowledge later. But as Gopnik points out, "the disadvantage of that is that you don't get good payoffs when you are in the exploration stage." Hence childhood: "Childhood gives you a period in which you can just explore possibilities, and you don't have to worry about payoffs because payoffs are being taken care of by the mamas and the papas and the grandmas and the babysitters."

Thinking about children as simply being at the transitory exploration stage of a lifelong algorithm might provide some solace for parents of preschoolers. (Tom has two highly exploratory preschool-age daughters, and hopes they are following an algorithm that has minimal regret.) But it also provides new insights about the rationality of children. Gopnik points out that "if you look at the history of the way that people have thought about children, they have typically argued that children are cognitively deficient in various ways—because if you look at their exploit capacities, they look terrible. They can't tie their shoes, they're not good at long-term planning, they're not good at focused attention. Those are all things that kids are really awful at." But pressing buttons at random, being very interested in new toys, and jumping quickly from one thing to another are all things that kids are really great at. And those are exactly what they should be doing if their goal is exploration. If you're a

baby, putting every object in the house into your mouth is like studiously pulling all the handles at the casino.

More generally, our intuitions about rationality are too often informed by exploitation rather than exploration. When we talk about decision-making, we usually focus just on the immediate payoff of a single decision—and if you treat every decision as if it were your last, then indeed only exploitation makes sense. But over a lifetime, you're going to make a lot of decisions. And it's actually rational to emphasize exploration—the new rather than the best, the exciting rather than the safe, the random rather than the considered—for many of those choices, particularly earlier in life.

What we take to be the caprice of children may be wiser than we know.

… And Exploit

> *I had reached a juncture in my reading life that is familiar to those who have been there: in the allotted time left to me on earth, should I read more and more new books, or should I cease with that vain consumption—vain because it is endless—and begin to reread those books that had given me the intensest pleasure in my past.*
>
> —LYDIA DAVIS

At the other extreme from toddlers we have the elderly. And thinking about aging from the perspective of the explore/exploit dilemma also provides some surprising insights into how we should expect our lives to change as time goes on.

Laura Carstensen, a professor of psychology at Stanford, has spent her career challenging our preconceptions about getting older. Particularly, she has investigated exactly how, and why, people's social relationships change as they age. The basic pattern is clear: the size of people's social networks (that is, the number of social relationships they engage in) almost invariably decreases over time. But Carstensen's research has transformed how we should think about this phenomenon.

The traditional explanation for the elderly having smaller social networks is that it's just one example of the decrease in quality of life that comes with aging—the result of diminished ability to contribute to social relationships, greater fragility, and general disengagement from society. But Carstensen has argued that, in fact, the elderly have fewer social relationships by choice. As she puts it, these decreases are "the result of lifelong selection processes by which people strategically and adaptively cultivate their social networks to maximize social and emotional gains and minimize social and emotional risks."

What Carstensen and her colleagues found is that the shrinking of social networks with aging is due primarily to "pruning" peripheral relationships and focusing attention instead on a core of close friends and family members. This process seems to be a deliberate choice: as people approach the end of their lives, they want to focus more on the connections that are the most meaningful.

In an experiment testing this hypothesis, Carstensen and her collaborator Barbara Fredrickson asked people to choose who they'd rather spend thirty minutes with: an immediate family member, the

author of a book they'd recently read, or somebody they had met recently who seemed to share their interests. Older people preferred the family member; young people were just as excited to meet the author or make a new friend. But in a critical twist, if the young people were asked to imagine that they were about to move across the country, they preferred the family member too. In another study, Carstensen and her colleagues found the same result in the other direction as well: if older people were asked to imagine that a medical breakthrough would allow them to live twenty years longer, their preferences became indistinguishable from those of young people. The point is that these differences in social preference are not about age as such—they're about where people perceive themselves to be on the *interval* relevant to their decision.

Being sensitive to how much time you have left is exactly what the computer science of the explore/exploit dilemma suggests. We think of the young as stereotypically fickle; the old, stereotypically set in their ways. In fact, both are behaving completely appropriately with respect to their intervals. The deliberate honing of a social network down to the most meaningful relationships is the rational response to having less time to enjoy them.

Recognizing that old age is a time of exploitation helps provide new perspectives on some of the classic phenomena of aging. For example, while going to college—a new social environment filled with people you haven't met—is typically a positive, exciting time, going to a retirement home—a new social environment filled with people you haven't met—can be painful. And that difference is partly the result of where we are on the explore/exploit continuum at those stages of our lives.

The explore/exploit tradeoff also tells us how to think about advice from our elders. When your grandfather tells you which restaurants are good, you should listen—these are pearls gleaned from decades of searching. But when he only goes to the same restaurant at 5:00 p.m. every day, you should feel free to explore other options, even though they'll likely be worse.

Perhaps the deepest insight that comes from thinking about later life as a chance to exploit knowledge acquired over decades is this: life should get better over time. What an explorer trades off for knowledge is pleasure. The Gittins index and the Upper Confidence Bound, as we've seen, inflate the appeal of lesser-known options beyond what we actually expect, since pleasant surprises can pay off many times over. But at the same time, this means that exploration *necessarily* leads to being let down on most occasions. Shifting the bulk of one's attention to one's favorite things should increase quality of life. And it seems like it does: Carstensen has found that older people are generally more satisfied with their social networks, and often report levels of emotional well-being that are higher than those of younger adults.

So there's a lot to look forward to in being that late-afternoon restaurant regular, savoring the fruits of a life's explorations.