On the value of efficient search*

Christopher J. Costello† and Michael B. Ward

November 1, 2004

Abstract

Costly search for rare items pervades society. We model R&D as a process of uncertain search and discovery. A collection of research leads can be searched at random or in a more efficient manner informed by additional information. We examine the value of information that facilitates efficient search. We find that an inherent tension exists that renders this value very small for a large class of problems. For example, when applied to the current controversy over biological prospecting, optimally ordering research leads improves the value of the collection only 2% above random search. Results contradict widely-held views on the value of the research process.

1 Introduction

What do a cure for cancer, the incandescent light bulb, and the proof of Fermat's Last Theorem have in common? These are just a few examples of research and development (R&D) applications that require costly search of a potentially large collection of research leads. Information acquired through, for example, preliminary testing, indigenous knowledge, or developing a scientific framework, can improve search efficiency and enhance the value of the R&D project. However, acquiring this information may be costly. When determining whether to pursue additional information, the pivotal question, and the focus of this paper is: how large are the benefits of efficient search? And while a large theoretical literature addresses the question of how to conduct maximally efficient R&D (Granot and Zuckerman 1991; Gallini and Kotowitz 1985; Roberts and Weitzman 1981; Lucas 1971; Ross 1969; Charnes and Stedry 1966), the value of R&D efficiency has not been examined.

Across actual R&D applications, the amount of preliminary testing can vary significantly. Take for example two famous R&D projects by the American inventor, Thomas Edison. In the development of the incandescent light bulb, Edison's technicians laboriously waded, essentially at random, through thousands of different filament materials before finding that carbonized sewing thread met his durability and brightness criteria. In contrast to this largely unordered search, during the search for a domestic source of rubber, Edison carefully ranked hundreds of plant species according to their potential before extensively testing and cross-breeding varieties of the most promising species.

Why might some R&D applications call for a meticulously researched scientific framework to inform discovery, while others require no more than a random search of even large collections of research leads? In this paper we examine theoretically the value of efficient

*__DRAFT DATE November 1, 2004__. PLEASE DO NOT CITE

†Corresponding Author: Donald Bren School of Environmental Science & Management, 4410 Bren Hall, Santa Barbara, CA 93117, Costello@bren.ucsb.edu

R&D. We apply the theoretical model to the pharmaceutical R&D problem of bioprospecting.

Intuitively, it would seem that efficient search must have high value. Indeed, Rausser and Small (2000, p. 174) argue "It is a powerfully general rule that no one ever searches for anything by examining large collections of objects in random order". Ordered search generates an efficiency gain equal to the number of searches saved times the cost of an individual search. If search cost is significant, this efficiency gain can be substantial, ceteris paribus. However, our analysis reveals a counteracting effect that suppresses the value of efficient search. This second effect is driven by the initial identification of the collection of leads worth searching. As search cost increases, the collection of worthy leads shrinks. This effect always acts to decrease the value of efficient search.

When the effects are combined, we find overall that for a large class of search problems, the value of information facilitating efficient search is trivially small; this is the value of Phase II R&D. This result is shown to be insensitive to particular features of the probability distribution of leads. Moreover, we determine what properties tend to make efficient search more, or less, valuable. Even in some seemingly favorable cases, the value is still small.

The paper is organized as follows. We develop in section 2 a model in which a collection of leads is ordered and searched sequentially. We derive the expected value of efficient search. The inherent tension between the intensive and extensive margins of search is explored in section 3.

In section 4 we apply our theoretical model empirically to the case of bioprospecting, where we find, contrary to previous studies, that the value of efficient search is only about 2%. We find that features of the bioprospecting problem (e.g. a very large number of highly diverse research leads) should lead to a higher, rather than lower, value of efficient search. Viewed in that light, we conclude that search problems with less extreme characteristics may benefit even less from efficient search. We conclude in section 5.

2 R&D as a search

In this section, we lay out a stylized model of R&D, with three distinct phases: identifying the collection of leads (Phase I), optionally refining information on quality of leads (Phase II), and searching the leads for an eventual success (Phase III).

2.1 Overview

At the beginning of a project, the researcher faces a large collection of leads of potentially varying quality (probability of success). Searching each of these leads entails a cost, c. The search terminates on the first success, and a gross payoff of R is realized. The researcher is risk neutral and so maximizes the expected payoff from the R&D project. In Phase I, the researcher coarsely sorts the collection into two piles: those worthy of further investigation, and those that should be discarded. Phase I sorting is conducted on the basis of crude prior knowledge of the leads' potentials. In the context of the light bulb example, one might think of Phase I as the choice to search among carbonized filament materials.[1] In the search for a domestic rubber source, Phase I might be thought of as the limitation of leads to endemic plant species with some known rubber content.

[1] In fact Edison purchased a patent on the use of carbonized filament from another inventor. The preferred embodiment listed in the patent, carbonized bamboo filament, was not sufficiently durable for commercial use.

Once the collection of worthy leads has been identified, the researcher faces a choice that is central to our analysis: whether to collect additional information about lead qualities in order to more efficiently search among them. This is Phase II of R&D. Alternatively, the researcher can skip this phase, and simply search randomly among the collection of leads identified in Phase I as worthy.[2] Of course the value of the project under ordered search (with the information acquired in Phase II of research) will exceed the value of the project under random search (without Phase II). But acquiring the information in Phase II may carry a cost that exceeds this efficiency gain. In the search for a domestic source of rubber, Edison's team planted thousands of species for preliminary testing. After assessing the rubber content of these initial plantings, the team ranked the species in terms of promise for more thorough investigation with the top one being goldenrod. In contrast, in the light bulb search, testing of thousands of candidate filament materials proceeded in a fairly haphazard way. In our stylized view, we think of the light bulb search as skipping Phase II.

The actual search, Phase III, involves the full testing of leads for a success. In this phase, the researcher sequentially tests leads. These tests resolve all uncertainty. In the light bulb search, this involved checking whether a candidate filament in light bulbs met minimal durability and brightness criteria. In the case of rubber, this involved testing whether by cross-breeding of varieties a sufficiently low-cost, high-yield specimen could be developed.

This stylized view of R&D as a multi-stage process of lead identification, investigation, and uncertain discovery of patentable products is consistent with approaches taken in the literature (see, e.g. Gallini and Kotowitz (1985), Fudenberg et al. (1983), Roberts and Weitzman (1981), Charnes and Stedry (1966)).[3] For example, Fudenberg et al. (1983) model two stages of R&D: a preliminary invention phase followed by the development phase; our analyis maps these to Phase II and Phase III, respectively. Similarly Gallini and Kotowitz (1985) distinguish between the basic research and the development phases. As in Phase II here, the purpose of the basic research step is to reduce the number of research avenues so that in the development phase research can focus in on the few most promising projects.

As in Rausser and Small (2000), Granot and Zuckerman (1991), Gallini and Kotowitz (1985), Lucas (1971) and others we will focus initially on an R&D project that is already under way. Here, the researchers have already identified in Phase I which leads from a potentially infinite initial set are worthy of further investigation. Most of this literature focuses on optimally engaging in Phase II, the second development stage of a project. While our focus is also on this phase, we find that the interplay between the basic research and the development phases of research is central to the analysis of the value of efficient search.

2.2 Model of the research process

In this section we formalize the model of efficient search of those leads. Following Rausser and Small (2000), Simpson et al. (1996), Ross (1969), and others, N research leads iden-

[2]Of course, it is likely that a researcher would have *some* priors on lead quality. Here, we take a stylized view that no prior information is available to rank lead quality. Then, if Phase II is omitted, search must proceed in random order. This stylized view is conservative in that we tend to understate the real information available to a researcher, and thus overstate the benefits of acquiring more information.

[3]We analyze the problem from the perspective of a single researcher or firm attempting to identify a success among a pool of research leads. This is distinct from the "patent race" literature in which multiple firms are competing to achieve the same success (see, e.g. Gallini and Kotowitz (1985) and Fudenberg et al. (1983)).

tified in Phase I as worty of search are to be sequentially tested for a success. The search terminates upon the first success yielding gross payoff R. Each test involves search cost c, and lead i has success probability p_i, which is independent of the success of all other leads and, from the researcher's perspective, may be uncertain. As a consequence of Phase I research, leads are retained only if their expected benefits exceeded the cost of their search: $p_i R > c$. We assume that while the researcher may not have well-formed priors over the probabilities of success, he can distinguish between those leads worthy of search ($p_i > c/R$) and those that are not ($p_i < c/R$). The presence of sequential search with positive search cost is what distinguishes this model from that in Polasky and Solow (1995), who also consider the value of a collection of leads. Because they treat search cost as zero, search order is irrelevant, and therefore the type of information considered here could not have value.

The researcher may refine his beliefs about success probabilities by collecting additional information, perhaps at some cost. The details of this updating process will depend on the R&D application, but may involve things like preliminary testing, developing a scientific framework, or incorporating indigenous knowledge. While this information will improve search efficiency, its acquisition may be costly. To calculate the value of the information, we take as a benchmark the case in which the researcher knows the distribution of probabilities of success, but does not know the probability of success for any particular lead. Under this assumption the researcher has no way to distinguish the quality of leads and so the collection is examined in random order.

Undertaking Phase II of the research process would provide further refining information that identifies the probabilities precisely and would thus facilitate a more efficient search. We use the phrase "value of efficient search" to denote the difference in the *ex ante* expected value of the collection of leads under fully optimal and random search; that is the difference between the expected value of the project with the information provided in Phase II and the expected value of the project when the researcher skips immediately from Phase I to Phase III. Because less precise information would confer lower value, our calculations represent an upper bound on the value of efficient search.

2.3 Value of a collection

Let S be a collection of N research leads with an associated set of probabilities $\{p_i\}$. By the virtue of Phase I, each lead in this collection is of sufficiently high quality to justify its search; we relax this assumption in a later section. In this section, we calculate the value of optimally *ordering* those leads in the search queue.

Denote a specific ordering of the N research leads in set S by $S^\wedge$ with associated probabilities $\{p_1^\wedge, p_2^\wedge, ..., p_N^\wedge\}$, respectively. The level of certainty the researcher has about these probabilities depends on the information he has at the time of the search. In this arbitrary ordering, lead $S_i^\wedge$ is the i^{th} lead to be tested. The expected value of the ordered collection $S^\wedge$ is given by

$$V(S^\wedge) = R(1 - a_{N+1}(S^\wedge)) - c\sum_{n=1}^{N} a_n(S^\wedge) \tag{1}$$

where $a_n(S^\wedge) = \prod_{i=1}^{n-1}(1 - p_i^\wedge)$ is the probability that every preceding lead has been tested unsuccessfully. The formula (1) has an intuitive explanation. The first term on the right hand side is the expected revenue of the search: the probability that at least one lead contains a success times the revenue upon success. The total value of the ordered collection

is that revenue less the expected cost – the expected number of trials until the first success multiplied by the cost of each trial.

2.4 Value of improving search efficiency

We consider two different queues consisting of the same collection of leads, each represented by a researcher who searches that queue. Consider first Researcher 1 who acquires the efficiency generating information in Phase II and thus knows precisely each of the probabilities of success; this information facilitates efficient search of the collection. In other words, he has a well-developed subjective belief about the success probabilities of her research leads, and can therefore order them optimally. The problem of efficiently ordering the search queue is then a special case of the "Pandora's box" problem analyzed by Weitzman (1979). In our case, an efficient search queue orders the leads in descending order of the probability of success[4]. We denote by S^* the optimally ordered queue. The *ex ante* expected value of the collection by Researcher 1 is then $V_1 \equiv V(S^*)$.

How much would Researcher 1 be willing to pay, *ex ante*, for the queue S^* rather than some alternative ordering of the same leads, $S^\wedge$? This value is given by:

$$V_1 - V(S^\wedge) = c\left(\sum_{n=1}^{N} a_n(S^\wedge) - \sum_{n=1}^{N} a_n(S^*)\right) \tag{2}$$

Note that expected revenue plays no role here because the probability of conducting a full search without any successes is independent of the search order: $a_{N+1}(S^*) = a_{N+1}(S^\wedge)$ for the ordering, $S^\wedge$. Equation 2 holds *for any* sequence $S^\wedge$.

At the other extreme is Researcher 2 who skips Phase II and thus lacks any information distinguishing one lead from another, and so searches the collection in random order. Denote by V_2 the *ex ante* expected value of the collection for Researcher 2. It is given by the sum of the values of all possible search queues divided by the number of such queues, as follows:

$$V_2 = E\left[V(S^\wedge)\right] = \frac{\sum_{S^\wedge} V(S^\wedge)}{N!}. \tag{3}$$

Here we would like to emphasize that search efficiency has value because it can lead to success with fewer searches; to the extent that search cost is a factor, this efficiency can generate significant savings, ceteris paribus. Because equation 2 holds for any alternative queue $S^\wedge$, it also holds for a random search. This implies that the value of efficient search is a linear (increasing) function of the coast of an individual search, c. In the extreme case in which search is costless ($c = 0$, see Polasky and Solow (1995)), efficient search has no additional value relative to random investigation of the same collection.

We denote the difference in value between Researcher 1 and Researcher 2 by $\Theta_{12} \equiv V_1 - V_2$, which gives the value of efficient search of the collection of leads. While the remainder of the paper explores the properties and magnitude of this nominal measure of the value of efficient search, an alternative is in percentage terms. We define this measure $\Pi_{12} \equiv (V_1 - V_2)/V_1 = \Theta_{12}/V_1$ as the percent difference between the value of a collection when it is searched optimally and the value of the same collection when it is searched at random. This measure may be useful as it captures the value of efficient search relative to the scale of the optimal search problem.

[4] A more general treatment would allow cost to vary across leads. Applying Weitzman's (1979) formula, it can be shown in that case that leads should be ordered in descending order of p_i/c_i.

2.5 Implications for information gathering

The value of efficient search, Θ_{12} provides a measure of the gross economic benefit of the information acquired in Phase II of R&D. Ultimately, the question faced by the researcher is whether acquiring the information exceeds the cost of the information. While the focus of this paper is on calculating the benefits, we provide the benefits vs. cost comparison for completeness. Suppose the cost of acquiring information on each lead is k. Then acquiring information on the entire pool of leads would be kN. Denoting by ΔS the number of searches saved by ordered, relative to random search, we have $\Theta_{12} = c(\Delta S)$. So the benefits of the information exceed the costs if and only if

$$k < c\frac{\Delta S}{N} \tag{4}$$

Equation 4 reveals that information should be acquired if and only if the cost of information is less than a certain fraction of the full search cost for a lead. The fraction ($\frac{\Delta S}{N}$) can be interpreted as the fraction of searches saved relative to an exhaustive search. For example if the info will be expected to save 10 searches, and there are 1000 leads in the collection, then the researcher should acquire the additional information if and only if acquiring information about the lead is less than 1% as costly as fully testing the lead for a success.

3 Analyzing the theoretical model

In this section we analyze the theoretical model of Section 2 to derive general properties of the value of efficient search. We begin by identifying and exploring an inherent tension that we find suppresses the value of efficient search. We also examine the features of an R&D problem that would tend to make gathering information more, or less, valuable and use those results to derive an upper bound on the value of efficient search. We continue to assume that the researcher has completed Phase I, and is faced with a collection of leads already known to be *individually* worthy of search, but whose precise probabilities may be unknown.

3.1 Exploring a tension: intensive vs. extensive margin

Equation 2 emphasizes the importance of search cost in the value of efficient search. If search cost is low, efficiency in search confers little value to the R&D program. Intuitively, then, it would seem that larger values of c should generate larger values of efficient search. Indeed this intuition is correct on the intensive margin where other parameters of the problem do not adjust.

While casual intuition and equation 2 may suggest an unambiguous effect of search cost on the value of efficient search, Lemma 1 below reveals a second, countervailing effect on the extensive margin. Larger values of c also raise the Phase I cutoff for the minimum acceptable probability (recall that the researcher retains only leads for which $p_i > c/R$). In general then, the pool of leads shrinks for larger values of c. We will demonstrate that this effect always works in a predictable direction - to decrease Θ_{12}. This extensive margin result, which decreases the value of efficient search, is in tension with intensive margin effect, which increases the value.

Lemma 1 *: The value of efficient search, Θ_{12} decreases if the lowest probability lead is dropped from a collection of leads.*

The proof is in the Appendix.

The tension between the intensive and extensive margin effects is made explicit in Proposition 1 below.

Proposition 1 *: As c increases, the value of efficient search, Θ_{12} (a) increases on the intensive margin, (b) decreases on the extensive margin (c) is 0 for sufficiently high c.*

Proof. (a) Equation 2 shows that on the intensive margin Θ_{12} is linear in c, with a positive slope.

(b) Lemma 1 shows that on the extensive margin Θ_{12} decreases when a lead is dropped.

(c) Let c be large enough that only leads with the highest probability level survive Phase I. Since all remaining leads have the same (high) probability value, they are homogenous from a search perspective, and no value accrues to distinguishing among them. ■

Intuitively, the tension illuminated in Proposition 1 can be illustrated with a continuous representation of research leads. If we define by $D(c)$ the pool of leads that are discarded as a function of c, then we can think of Θ_{12} as a function of search cost (c) and the pool of discarded leads: $\Theta_{12}(c, D(c))$. Then the effect of search cost on the value of efficient search can be intuited by taking the derivative of this expression with respect to c:

$$\frac{d\Theta_{12}}{dc} = \frac{\partial\Theta_{12}}{\partial c} + \frac{\partial\Theta_{12}}{\partial D}\frac{\partial D}{\partial c}. \tag{5}$$

The first term is the direct effect of c on Θ_{12}, holding the pool of leads constant. This effect is always positive, and we think of this as the intensive margin because no other features are adjusting to the search cost. The second term is always negative. It consists first of the term $\frac{\partial\Theta_{12}}{\partial D}$ which is the cost of losing a lead at the cutoff c/R, which by Lemma 1 is always negative. The term $\frac{\partial D}{\partial c}$ is the rate of loss of leads as c increases, which is always positive or zero. This tension means that the effect of increasing c on Θ_{12} is not always positive. In fact, as c gets larger, the extensive margin effect dominates, and eventually, for high enough c, the value $\Theta_{12} = 0$. In other words, for large enough c, the value of efficient search is zero.

We illustrate the intensive vs. extensive margin of search with a simple example with $N = 100$ leads that are uniformly distributed between 0 and 1 ($p_1^* = 1.0$, $p_2^* = .99$,...,$p_N^* = .01$) and revenue on success $R = 1000$. Figure 1 depicts Θ_{12} as a function of search cost, c. Although higher costs are associated with higher values of Θ_{12} on the intensive margin, the figure illustrates the importance of the extensive margin which reduces the size of the pool of leads in Phase I of R&D, thus driving Θ_{12} down. As shown in Proposition **??**, the extensive margin effect eventually dominates, driving Θ_{12} to 0.

3.2 Bounding the value of efficient search

Proposition 1 shows that while $\frac{d\Theta}{dc}$ may be positive for very low values of c, the extensive margin eventually dominates, thus suppressing the value of efficient search. But the size of Θ is still an empirical question. How large could Θ get? And in particular, do certain characteristics of the distribution of p's tend to generate a larger value of efficient search? In general, the value of efficient search will depend in a complicated way on the probabilities of all the leads, as shown in equation (2). It thus seems intuitively plausible that for some favorable probability distributions the value of efficient search could be enormous. In this section we examine this question by deriving the non-parametric distribution of probabilities that gives the theoretically maximal value of efficient search.

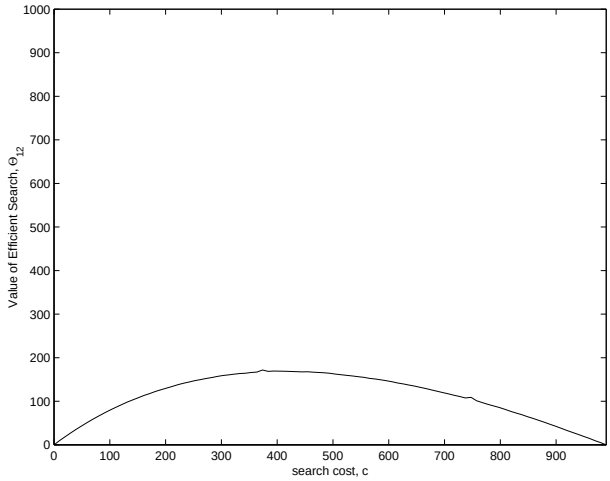

Figure 1: Value of efficient search, Θ_{12} as a function of search cost for $N = 100$ uniformly distributed leads ($\in [0, 1]$) and $R = 1000$.

Dependence on probability of success

We now explore the role of the overall probability of success of the search in determining the value of efficient search. To do so we need to be specific about how the overall probability of success is increased. One natural way to do so is to increase by a small amount all of the probabilities. Intuition may suggest that the less likely is an eventual success, the less important is information that facilitates optimal search. This intuition turns out to be precisely backwards. Careful analysis reveals that the savings in the expected number of searches as the probabilities are increased is smaller under the optimal ordering than under *any* alternative ordering. Therefore, the value of efficient search declines as the probabilities are increased. This result is summarized in Lemma 2 below.

Lemma 2 *The value of efficient search, Θ_{12} decreases if all lead probabilities are additively increased by the same small amount.*

A proof is in the Appendix.

Dependence on the spread of the probabilities

Lemma 2 explores how the value of efficient search depends on the overall probability of success. But how does it depend on the spread of probabilities? To begin, note that if all probabilities are equal, random search is optimal, so $\Theta_{12} = \Pi_{12} = 0$. In contrast, the greater the dispersion of lead qualities, the greater is the opportunity to exploit information to re-order the search. This is essentially the phenomenon that Rausser and Small (2000) intended their data to illustrate.

To analyze this effect, we need a working definition of "spread". Suppose we replace two leads p_i, p_j in a pool with q_i, q_j such that $q_i > p_i > p_j > q_j$ and $(1 - p_i)(1 - p_j) = (1 - q_i)(1 - q_j)$. Then we say that the second pool has greater spread, as suggested by the values of p and q. The second condition is a normalization, which preserves the overall probability of success in the search. The effect of spread is summarized in Lemma 3 below:

Lemma 3 *The value of efficient search, Θ_{12} increases if the spread of lead qualities is increased.*

A proof is in the Appendix.

An upper bound

What would the most favorable non-parametric probability distribution look like? Under optimal search, the worst lead must have probability of at least c/R to justify search. This is because the marginal value of the worst lead under optimal search is $p_iR - c$ and so if $p_i < c/R$ the lead would be dropped.[5] Holding constant the overall success probability, P, the probability distribution that yields the largest value of efficient search would, by Lemma 3, have maximal spread. In this case, we would have $N-1$ leads of probability c/R. The remaining lead $\bar{p}$ would then be set large enough that the overall success probability equals P, as required; this value is $\bar{p} = 1 - \frac{(1-P)}{\delta^{N-1}}$. Under this probability distribution, Researcher 2 (who searches at random) must wade through a large number of marginal leads before eventually succeeding but Researcher 1 (who searches the best lead first) is most likely to find success on the first trial. Proposition 2 below calculates the value of efficient search under this probability distribution. Because this value would be smaller under any other probability distribution, we interpret it as an upper bound on the value of efficient search. Defining by $\delta = 1 - c/R$ the probability of failure when searching a marginal lead, we have:

Proposition 2 *The nominal value of Phase II R&D, Θ_{12}, cannot exceed:*

$$\bar{\Theta}_{12} = c(\frac{\bar{p}}{1-\delta} - 1)(1 - \frac{1-\delta^N}{N(1-\delta)}). \tag{6}$$

A Proof is in the Appendix.

For reference, note that rewriting Proposition 2 in percentage terms, we find that the percentage value of Phase II R&D, Π_{12}, cannot exceed[6] :

$$\bar{\Pi}_{12} = 100\left[1 - \frac{1-\delta^N}{N(1-\delta)}\right] \tag{7}$$

The calculations in Proposition 2 are important for two reasons. First, they will support our empirical investigation in the next section. Second, they provide rules of thumb that could help guide the direction of research. Because $\bar{\Theta}_{12}$ and $\bar{\Pi}_{12}$ do not require any knowledge of the probability distributions of leads, the bounding argument could be used in practice to provide an upper bound on the value of the research process even when very little is known about the nature of the leads themselves. $\bar{\Pi}_{12}$ is particularly useful in this regard because it requires estimates only of the size of the collection of leads (N) and the ration of search cost to revenue ($c/R = 1 - \delta$). In particular, it is independent of the overall success probability, P so little information is required to assess whether further refining information would be

[5]Note that under random search the "cutoff" value could be slightly larger than c/R, so leads with probability $< c/R$ would not ever be searched. Calculating analytically this precise cutoff value is extremely complex. Simulations reveal that the cutoff is numerically extremely close to c/R (typically within less than 1%). Using c/R as the cutoff only overstates the value of efficient search. This is a second reason for calling the derived value an upper bound.

[6]Formally, it can be shown that the maximal spread distribution maximized Π as well as Θ.

useful in the R&D process. It is straightforward to show using the bounding argument that the attractiveness of efficient search increases in the pool of leads, N, and the relative size of c/R; $\frac{\partial \bar{\Pi}_{12}}{\partial N} > 0$ and $\frac{\partial \bar{\Pi}_{12}}{\partial \delta} < 0$. The empirical magnitude of $\bar{\Pi}_{12}$ remains an open question which we address in section 4.

4 Bioprospecting as an R&D search

In this section we apply the theoretical results from sections 2 and 3 to the pharmaceutical R&D problem of bioprospecting - the search for valuable products in nature. This is a problem with some pedigree within the economics literature. Simpson, Sedjo, and Reid (1996) are the first to consider the search aspect of the bioprospecting problem. They are primarily interested in the conservation implications from bioprospecting and thus focus attention on the marginal values of research leads - endemic plant species present in biodiversity hotspots around the world. In that model, the researcher has no prior information on how to order the search, and so all leads are treated as having the same probability of success, $\tilde{p}$, and they are searched at random. The search terminates upon the identification of a natural compound with high pharmaceutical value. They find surprisingly that the marginal value of a species is likely to be small regardless of $\tilde{p}$. When $\tilde{p}$ is small, the marginal value of a species is small because any one species is unlikely to produce a success. When $\tilde{p}$ is large, the marginal value is small because species are close substitutes (any one species is likely to be redundant). Using a species area curve relationship, they translate the value per species into a value per hectare. The maximum value per hectare is about $21 and is in Western Ecuador, the most biodiverse region on earth.

In sharp contrast is the result derived by Rausser and Small (2000) in which leads are kilo-hectares of biologically diverse land (they consider the same 18 biodiversity hotspots considered by Simpson et al.). In that model, success probabilities are heterogeneous, and the researcher benefits from ordering her search to improve efficiency. Rausser and Small (2000) conclude that the most biodiverse hectare of land on earth (again in Western Ecuador) has a bioprospecting value of $9,177 given an efficient search queue. Rausser and Small attribute this dramatic increase (from $21/hectare to $9,177/hectare) to efficient search. The large discrepancy in marginal values between these two studies is suggestive that efficient search may have high value in this case. We will examine this possibility below, beginning with the value of Phase I R&D.

4.1 Phase I sorting

We begin our empirical inquiry with an emphasis of the importance of Phase I of the research process: deciding which leads are worthy of search. We provide the following example. Suppose that in addition to the 18 biodiversity regions originally considered, the collection had also contained a less promising biodiversity region. To make the example concrete, suppose the biodiversity-rich country of Argentina was included in the original collection. Argentina contains about 1100 endemic plant species over 273,669 kilo-hectares for a success probability of p_A =4.8E-8.[7] Since $p_A R < c$, Argentina would be eliminated from the collection in Phase I. But before Phase I the researcher has insufficient information to eliminate Argentina and all leads (including Argentina) are searched in random order,

[7]We assume here the Rausser and Small (2000) formula for the probability of success, which is $\tilde{p} *$ 1100/kilo-hectares.

the *ex ante* expected value of the collection is only $40 million. This can be compared with a value of $141 million when Argentina is removed but the remaining leads are searched in random order (Phase I only), and a value of $144 million when Argentina is removed and the remaining leads are searched in fully efficient order (Phases I and II).

Even more striking is the case in which the collection has a sufficiently large number of low probability leads to render the expected value of random search negative. In the absence of further refining information, such collections would never be searched. For example, suppose in addition to the California Floristic Province (hotspot # 18), the remainder of the United States was included in the original collection of biodiversity hotspots. This region contains 1900 endemic plant species over 891,296 kilo-hectares for a probability of success of p_{US} =2.5E-8. If no information distinguishing lead quality is available, adding the US renders the total collection of leads valueless (searching this entire collection at random would entail an expected *loss* of $194 million), so the collection would never be searched. In such cases, a coarse-grained sorting of leads can eliminate the low quality leads from the collection thereby saving search costs and substantially increasing the value of the remaining collection. Even when the remaining collection of leads is searched in random order the value of the collection is vastly improved.

4.2 Phase II ordering

We now apply the value of efficient search model derived in section 2. While the Simpson et al. and Rausser and Small models did differ slightly, the central message of Rausser and Small (2000) is that information that facilitates optimal search significantly improves value, and therefore raises the likelihood of private sector conservation. They refer to the random search assumption in Simpson et al. (1996) as a "nearly cost maximizing approach to discovery" (p. 175). In that debate, search efficiency appears to have an important effect; under efficient search the marginal values increase 440-fold (from $21/hectare to $9,177/hectare). Here we carefully examine that result.

As a point of departure, we conduct the following experiment. We employ the setup and parameter values from Rausser and Small (2000) to determine the value of the collection of 18 biodiversity hotspots under three different assumptions about the information available to the researcher (and therefore about the search order). The first case is analogous to our Researcher 1, who has already conducted Phase II of R&D and therefore searches the pool of leads in the most efficient order. The second case is analogous to our Researcher 2, who searches the leads at random. Finally, for illustrative purposes, we present a third case in which the leads are searched in the maximally inefficient order - in ascending order of the likelihood of success.Table 1 provides the results of this experiment.

Search Order	Value of Collection	% Loss relative to Optimal
Fully Efficient	$144 million	–
Random	$141 million	2.3%
Fully Inefficient	$138 million	4.9%

Surprisingly, efficient search has relatively little value in the bioprospecting example. The value of the collection of the 18 biodiversity hotspots is reduced by only about $3 million (2.3%) if the researcher is forced to search for a success at random. Even in the worst possible case in which the researcher searches in the maximally inefficient order (backwards), the value of the collection is relatively unchanged ($138 million vs. $144 million).

To what extent is this result an artifact of specific features of the bioprospecting problem? The bioprospecting data have several salient features. There is a very large pool of leads (74,600 kilo-hectares). Although success probabilities vary dramatically among the leads, the probabilities are all quite small. The greatest probability is 0.000105, and the smallest is about two orders of magnitude smaller. Given the empirically small value of efficient search in the bioprospecting problem, it is tempting to conclude that at least some of these features are responsible for the low value of efficient search. In fact careful inspection of equation 2 reveals the opposite; each of these factors tends to increase the value efficient search.

Ultimately the question faced by the researcher is whether additional information should be pursued. In the bioprospecting example, this information may take the form of, for example, ethnobotanical knowledge. Implementing equation 4, we find that this information should be acquired if and only if its cost is less than \$46 per lead (about 9% of the cost of fully searching the lead for a success).

We have shown using the data from Rausser and Small (2000) that the value of efficient search is small and that gross features of the bioprospecting model are conducive to a relatively high, rather than low, value. To determine whether some peculiar feature of the probability distribution itself is responsible for the low value of efficient search we implement Proposition 2 and find that even under the most favorable nonparametric probability distribution of leads, the value is still small ($\bar{\Theta}_{12}$ =\$5.2 million; in percentage terms $\bar{\Pi}_{12}$=3.9%).

4.3 Interplay between Phase I and Phase II

We have argued that while information facilitating efficient search has low value (Phase II), there may be significant value derived from eliminating inferior leads from the collection (Phase I). In the Rausser and Small bioprospecting model, c and R were chosen specifically so the marginal value of the worst lead would be zero (see p. 192). In that example if c were higher than \$485, an efficient search would entail eliminating some leads from the queue. How well will the researchers fare under different values of c?

Figure 2 illustrates the empirical effect of search cost on the value of efficient search. The top panel does so for the nominal value (Θ_{12}) and the bottom panel presents results as a function of the total value of the collection (Π_{12}).

Inspection of Figure 2 reveals that for the bioprospecting data the loss from inefficient search never exceeds about \$6 million (about 6%), regardless of the search cost. This suppression of the value of efficient search, even for favorable choices of c is a direct consequence of the intensive vs. extensive margin tension captured in Proposition 1.

5 Discussion

One purpose of research is to resolve uncertainty and thus inform better decisions. Whether searching for WMD, the cure for cancer, or the meaning of life, efficient search of a collection of leads must have higher value than random investigation of the same collection. Previous work suggested that dramatic dividends typically accrue from the use of an organizing scientific framework to better sequence search (Rausser and Small 2000). However, because such organizing information is often costly to acquire, the pivotal question is: how large is the value of information that facilitates efficient search?

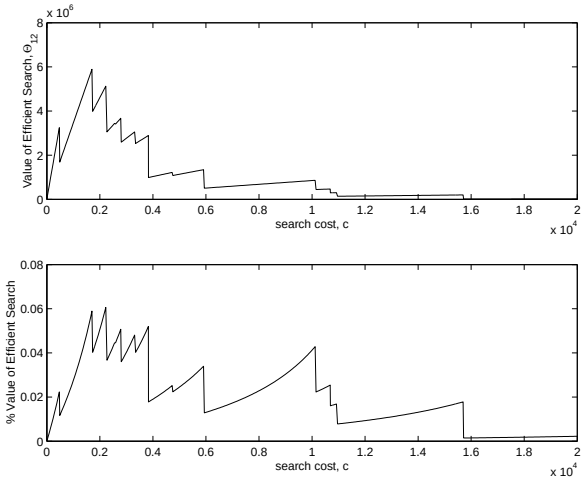

Figure 2: Value of efficient search (top) and percentage value relative to V_1 (bottom) as a function of search cost, c increases.

We identified an inherent tension between the intensive and extensive margins of search. When search cost is low, search efficiency is less important, because the searches it saves are inexpensive. On the other hand, when search cost is high, the collection of leads worthy of search is diminished. This effect always acts to decrease the value of efficient search. And for sufficiently high search cost, this effect always dominates. This effect tends to suppress the value of efficient search.

But in general, the empirical magnitude of this value will depend in a complicated way on the probability distribution of leads. We derived an expression for the value of efficient search, and calculated its theoretical upper bound as a function of the key parameters of any R&D problem. In the bioprospecting problem, we found this upper bound to be exceedingly small; about 4% of the total value of the collection. We argued that features of the bioprospecting problem were particularly well positioned to generate a high value of efficient search. For search problems with less extreme characteristics (e.g. smaller N or less heterogeneous probabilities) efficient search would be even less valuable.

Do the results of this analysis imply that basic scientific research and development has only negligible value? No. The result that even under favorable conditions the value of efficient search is often small applies only to a collection of leads that has already been identified as worthy of search. But as the tension in Proposition 1 reveals, leads for which $pR < c$ should never be searched. Information that allows the researcher to eliminate very low quality leads from the collection may have tremendous value, even when the good leads are subsequently searched at random. In many cases, information enabling coarse sorting can provide incentive to search a collection that was previously valueless; this was empirically illustrated with the example of adding Argentina or the United States to the collection of biodiversity hotspots. Again, in that case, the additional value of optimally ordering the good leads may be negligible.

The principles of this kind of search problem apply to any research inquiry. In all search applications of this type, the researcher faces the conceptually separate questions of which leads to search and in what order to search them. Although the R&D literature to date has focused on search efficiency, our analysis suggests that the former question is much more

crucial than the latter.

6 Appendix

Proof of Lemma 1. Let S^* denote the optimally ordered collection of leads and let $S^\wedge$ be an alternative ordering. We require notation for these ordered collections for the case in which the worst lead is omitted from the collection. We denote these by $\tilde{S}^*$ and $\tilde{S}^\wedge$, respectively.

Let $p_t^\wedge \equiv p_N^*$ be the value of the lowest lead.

Let $a_{i,j} = \prod_{k=i}^{j-1}(1-p_k^*)$. For $i \geq j, a_{i,j} = 1$

Let $b_{i,j} = \prod_{k=i}^{j-1}(1-p_k^\wedge)$. For $i \geq j, b_{i,j} = 1$

Then,

$$\Delta\Theta = (V(S^*) - V(S^\wedge)) - (V(\tilde{S}^*) - V(\tilde{S}^\wedge)).$$

The expected revenue for the pairs $V(S^*), V(S^\wedge)$ and $V(\tilde{S}^*), V(\tilde{S}^\wedge)$ are the same, because the same set of leads is being searched (in different order). So, Θ is determined only by search costs.

Let EC be the expected search cost for each case. Then rearranging,

$$\Delta\Theta = (EC(S^\wedge) - EC(\tilde{S}^\wedge)) - (EC(S^*)) - EC(\tilde{S}^*)).$$

Note that the marginal value of lead $p_t^\wedge$ in the alternative queue is:

$$(p_t^\wedge R - c)b_{1,t} - \frac{p_t^\wedge}{1-p_t^\wedge}\sum_{i=t+1}^{N}(p_i^\wedge R - c)b_{1,i}.$$

So, $EC(S^\wedge) - EC(\tilde{S}^\wedge) = c\,(b_{1,t} - p_t^\wedge \sum_{i=t+1}^{N} b_{1,i})$. Similarly, $EC(S^*) - EC(\tilde{S}^*) = c\,a_{1,N}$.

Substituting in these expressions yields

$$\Delta\Theta = c\,(b_{1,t} - \frac{p_t^\wedge}{1-p_t^\wedge}\sum_{i=t+1}^{N} b_{1,i} - a_{1,N}).$$

Since the worst lead goes last in the optimal queue, $a_{1,N} = b_{1,t}\ b_{t+1,N+1}$. Using this fact, then factoring out $b_{1,t}$ yields

$$\Delta\Theta = c\,b_{1,t}(1 - \frac{p_t^\wedge}{1-p_t^\wedge}\sum_{i=t+1}^{N} b_{t,i} - b_{t+1,N+1}).$$

We now bound the summation term. Since $p_t^\wedge$ is the worst lead, $b_{t,i} \leq (1-p_t^\wedge)^{i-t}$. So,

$$\sum_{i=t+1}^{N} b_{t,i} \leq \sum_{i=t+1}^{N}(1-p_t^\wedge)^{i-t} = (1-(1-p_t^\wedge)^{N-t})(1-p_t^\wedge)/p_t^\wedge.$$

Substituting in this bound,

$$\Delta\Theta \geq c\,b_{1,t}((1-p_t^\wedge)^{N-t} - b_{t+1,N+1}).$$

By the same argument, $b_{t+1,N+1} \leq (1-p_t^\wedge)^{N-t}$. Substituting in this bound,

$$\Delta\Theta \geq 0.$$

Since this is true for any alternative queue, it must also be true in expectation for a random search queue; dropping the worst lead decreases (weakly) Θ_{12}. ■

Proof of Lemma 2. Consider an alternative researcher A who searches the collection in order $S^{\wedge}$. Then let $\Theta_{1A} = V(S^*) - V(S^{\wedge})$. The optimally ordered sequence is S^*; the alternative sequence is $S^{\wedge}$. Letting $P_i^* \equiv (1 - p_i^*)$ and $P_i^{\wedge} \equiv (1 - p_i^{\wedge})$, the value of ordered search is:

$$\Theta_{1A} = c\left[\left(1 + P_1^{\wedge} + P_1^{\wedge}P_2^{\wedge} + ...\right) - \left(1 + P_1^* + P_1^*P_2^* + ...\right)\right]$$

. To prove our result, consider increasing each p_i $(\forall i)$ by some amount, k. We now have

$$\begin{aligned}\frac{\Theta_{1A}}{c} = &\left((P_1^{\wedge} - k) + (P_1^{\wedge} - k)(P_2^{\wedge} - k) + (P_1^{\wedge} - k)(P_2^{\wedge} - k)(P_3^{\wedge} - k) + ...\right) \\ &- \left((P_1^* - k) + (P_1^* - k)(P_2^* - k) + (P_1^* - k)(P_2^* - k)(P_3^* - k) + ...\right)\end{aligned} \tag{8}$$

Taking the derivative, and evaluating at $k = 0$ gives:

$$\begin{aligned}\frac{1}{c}\frac{d\Theta_{1A}}{dk}|_{k=0} = &\left[P_1^*\left(\frac{1}{P_1^*}\right) + P_1^*P_2^*\left(\frac{1}{P_1^*} + \frac{1}{P_2^*}\right) + P_1^*P_2^*P_3^*\left(\frac{1}{P_1^*} + \frac{1}{P_2^*} + \frac{1}{P_3^*}\right) + ...\right] \\ &- \left[P_1^{\wedge}\left(\frac{1}{P_1^{\wedge}}\right) + P_1^{\wedge}P_2^{\wedge}\left(\frac{1}{P_1^{\wedge}} + \frac{1}{P_2^{\wedge}}\right) + P_1^{\wedge}P_2^{\wedge}P_3^{\wedge}\left(\frac{1}{P_1^{\wedge}} + \frac{1}{P_2^{\wedge}} + \frac{1}{P_3^{\wedge}}\right) + ...\right]\end{aligned} \tag{9}$$

which is term-by-term non-positive, meaning that $\frac{d\Theta_{1A}}{dk} \leq 0$ with equality only if $S^{\wedge} = S^*$. Since this result holds for any sequence $S^{\wedge}$, it holds for the randomized sequence. ■

Proof of Lemma 3. The expected value of a random search can be evaluated by taking the average value of all search order permutations. Consider any specific permutation $\{..., i, ..., j, ...\}$ and pair it with the corresponding permutation which reverses the position of i and j. For example, let these permutations be $\{1, i, 2, j, 3, 4\}$ and $\{1, j, 2, i, 3, 4\}$. Let $\{p_i\}$ be the set of lead probabilities, and let $P_k = 1 - p_k$. The expected number of searches under the first permutation is $1+P_1+P_1P_i+P_1P_iP_2+P_1P_iP_2P_j+P_1P_iP_2P_jP_3$. The average search duration over both permutations is thus $1+P_1+(P_1+P_1P_2)(P_i+P_j)/2+P_1P_iP_2P_j+P_1P_iP_2P_jP_3$. Now consider an increase in spread that increases p_i to q_i and decreases p_j to q_j such that $P_iP_j = Q_iQ_j$, where $Q_k = 1 - q_k$. For the new set of probabilities, the only terms that change in the average search expression are those with the element $(P_i + P_j)$. The leading terms do not depend on i or j. The trailing terms are unchanged because of the normalization assumption. So, the increase in search duration attributed to the spread is $(P_1 + P_1P_2)(Q_i + Q_j - P_i - P_j)/2$. This term is positive since $P_iP_j = Q_iQ_j$ and $Q_j > P_j > P_i > Q_i$. This same approach applies to *any* pair of permutations which switches the positions of lead i and j. Averaging over all such pairs gives the expected value of a random search. ■

Proof of Proposition 2. Let P be overall probability of success. Let δ be $1 - c/R$. By Lemma 3 the maximal spread distribution will give an upper bound for Θ_{12}. Given N leads, the maximal spread distribution satisfying P has $N-1$ leads at $1-\delta$. The high-value lead has probability $\bar{p}$ satisfying

$$1 - \delta^{N-1}(1 - \bar{p}) = P.$$

$$\bar{p} = 1 - (1 - P)/\delta^{N-1}.$$

Given this distribution, the expected number of searches can be calculated by averaging over all N distinct positions for $\bar{p}$. If we have T initial low leads, the expected number of searches in that queue is (for example)

$$1 + \delta + \delta^2 + \ldots \delta^T + \delta^T(1 - \bar{p})(1 + \delta + \delta^2 + \ldots \delta^{N-T-2}).$$

$$= (1 - \delta^{T+1})/(1 - \delta) + \delta^T(1 - \bar{p})(1 - \delta^{N-T-1})/(1 - \delta)$$

$$= (1 - \delta^{T+1})/(1 - \delta) + (1 - \bar{p})(\delta^T - \delta^{N-1})/(1 - \delta)$$

Averaging over all positions and reducing the sums yields the expected number of searches in random order (Researcher 2):

$$ES_2 = (1 - \delta)^{-1}(1 - (1 - \bar{p})\delta^{N-1} + \frac{1 - \delta^N}{N(1 - \delta)}(1 - \bar{p} - \delta)).$$

Now the optimal search queue orders $\bar{p}$ first. In that case, the expected number of searches in optimal order (Researcher 1) is

$$ES_1 = 1 + (1 - \bar{p})(1 - \delta^{N-1})/(1 - \delta)$$

So the maximal Θ_{12} is

$$\bar{\Theta}_{12} = c(ES_2 - ES_1)$$

Collecting terms in the expected search expressions and simplifying yields

$$\bar{\Theta}_{12} = c(\frac{\bar{p}}{1 - \delta} - 1)(1 - \frac{1 - \delta^N}{N(1 - \delta)}).$$

■

References

Charnes, A. and A. Stedry (1966). A chance-constrained model for real-time control in research and development management. *Management Science 12*(8), B353–B361.

Fudenberg, D., R. Gilbert, J. Stiglitz, and J. Tirole (1983). Preemption, leapfrogging and competition in patent races. *European Economic Review 22*, 3–31.

Gallini, N. and Y. Kotowitz (1985). Optimal R and D processes and competition. *Economica 52*, 321–334.

Granot, D. and D. Zuckerman (1991). Optimal sequencing and resource allocation in research and development projects. *Management Science 37*(2), 140–156.

Lucas, R. (1971). Optimal management of a research and development project. *Management Science 17*(11), 679–697.

Polasky, S. and A. Solow (1995). On the value of a collection of species. *Journal of Environmental Economics and Management 29*(3), 298–303.

Rausser, G. and A. Small (2000). Valuing research leads:bioprospecting and the conservation of genetic resources. *Journal of Political Economy 108*(1), 173–206.

Roberts, K. and M. Weitzman (1981). Funding criteria for research, development, and exploration projects. *Econometrica 49*(5), 1261–1288.

Ross, S. (1969). A problem in optimal search and stop. *Operations Research 17*, 984–992.

Simpson, D., R. Sedjo, and J. Reid (1996). Valuing biodiversity for use in pharmaceutical research. *Journal of Political Economy 104*(1), 163–185.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica 47*(3), 641–654.