The value of efficient search

Christopher J. Costello* and Michael B. Ward

May 16, 2005

Abstract

Conventional economic wisdom holds that information begets efficiency. On this premise a large literature on R&D (and more generally, search) is devoted to identifying efficient search algorithms. We examine theoretically the value of information that enables more efficient search. We model research and development as a process of uncertain search and discovery. A collection of research leads can be queued for search based on existing information. Alternatively, the priors can be updated to facilitate efficient search, through, for example, the development of an organizing scientific framework. Results indicate that this value is usually small, even in favorably engineered cases. For example, in the development of pharmaceuticals from biological sources optimally ordering research leads improves the value of the collection only 2% above random search. Results contradict widely-held views on the value of information in the research process.

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

1 Introduction

Costly search of a collection of leads is a fundamental characteristic of all research and development problems, from the light bulb to a cure for cancer. More broadly, many economic problems center on search such as finding a job, workers, housing, low prices, natural resources, or even a spouse. A large literature examines the question of how to search a collection of leads efficiently *given* prior information (Granot and Zuckerman 1991; Gallini and Kotowitz 1985; Roberts and Weitzman 1981; Lucas 1971; Ross 1969; Charnes and Stedry 1966). And conventional economic wisdom holds that information begets efficiency; when applied to R&D, the information rents generated can be enormous (Rausser and Small 2000). We challenge this widely-held conclusion by examining theoretically the value of information that facilitates efficient search. We then apply this theory to a well-known pharmaceutical R&D problem. Results contradict widely held views on the value of information in search.

We consider perhaps the most basic case of the general "Pandora's Box" search problem solved by Weitzman (1979). In this problem, as in the broader search literature, the researcher faces the conceptually distinct questions of which leads deserve investigation, and in what order those leads should be examined. Of course, search would never be undertaken at all without the first step, since the set of leads would be effectively infinite. The interesting question, then, is whether rents from information allowing an efficient queue of reasonable leads are significant. Would a researcher be much worse off proceeding randomly through a collection, without guidance from search theory? If not, then devoting significant effort to fine-tuning search beyond merely identifying a promising set of leads is an inefficient allocation of resources. This may, in turn, inform the level of more basic scientific research in which a firm invests to guide R&D.

Intuition suggests that information rents from such efficient ordering should often be large, and particularly so if the cost of testing each lead is high. However, we identify a counteracting effect that suppresses the value of efficient search. This effect is driven by the initial identification of the collection of leads worth searching. For higher costs, the collection of leads worth searching shrinks. This effect always decreases the value of efficiently ordering search over the worthy collection. As a

consequence, information rents are usually small, whether testing costs are high or low. This occurs even in cases that are favorably engineered. This result is insensitive to particular features of the probability distribution of leads.

While this result might conflict with basic intuition, it is consistent with some notable historical examples of R&D applications. In one famous example, the development of a practical incandescent light bulb, search was undertaken essentially at random, guided without "science or system" (Friedel and Israel 1986). First a basic collection of leads worthy of search was identified: carbonized filaments.[1] Then, Thomas Edison's technicians laboriously waded through thousands of different filament materials, with little organizing principle, before finding that carbonized sewing thread met basic durability and brightness criteria.

The paper is organized as follows. In section 2 we examine a model in which a collection of leads is ordered and searched sequentially.[2] We derive the expected value of efficient search. In section 3 we explore the dependence of the value of efficient search on features of R&D problems. Further we identify a crucial inherent tension between the intensive and extensive margins of search. In section 4 we apply our theoretical model empirically to the case of pharmaceutical development from biological sources. We find, consistent with our theoretical results, that the value of efficient search in this case is only about 2%. In section 5 we present a concluding discussion. All proofs are presented in the appendix.

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).

---

[1] Edison purchased an 1875 patent on the general use of carbonized filaments from another inventor. The patent's preferred embodiment was not suitable for commercial use.

[2] For a discussion of parallel search see Vishwanath (1988).

2.1 Overview

The researcher begins by identifying which leads in the collection are sufficiently promising to justify their search. We call this stage of R&D Phase I, in which the researcher coarsely sorts the collection into two piles: those worthy of further investigation, and those that should be discarded. In the light bulb example one might think of Phase I as the choice to search among carbonized filaments. Phase I sorting may be conducted on the basis of either prior knowledge of the leads' potentials or some crude information that indicates whether a lead is sufficiently promising to justify its further investigation. Regardless of the exact mechanism, our stylized approach assumes the researcher can identify whether any particular lead is of sufficiently high quality to justify its further investigation.

This kind of sorting can have tremendous value. Taken to an extreme, in the absence of *any* information about lead quality, the pool of research leads would effectively include an infinite number of leads with no chance of success. Searches over such pools would never occur because they would in expectation generate net losses. Information that extracts only the promising, or *worthy*, leads would in such cases have high value. Importantly, we assume that, while Phase I facilitates sorting the leads into two piles, it does not inform the order of investigation of the remaining leads. This assumption allows us to obtain sharp results concerning the value of information in each stage.

Once a collection of worthy leads has been identified, the researcher faces the central choice of our analysis: whether to collect additional information about lead qualities in order to more efficiently search among them. We call this Phase II of R&D. Alternatively, the researcher can skip this phase, and search the collection of leads identified as worthy in Phase I based only on prior information. While it is, of course, likely that a researcher would have *some* priors on lead quality within the worthy collection, our analysis assumes 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 quite conservative in that we tend to understate the real information available to a researcher, and thus overstate the benefits of acquiring more information.

The actual search, Phase III, involves the sequential testing of leads for a success. These tests reveal with certainty whether a lead is successful. In the light bulb example, this involved testing

prototypes to determine whether a candidate filament met durability and brightness criteria. Again, the search terminates upon the first success.

This stylized view of R&D as a multi-stage process of lead identification, investigation, and uncertain discovery 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.

2.2 Model of the research process

We adopt a standard model of a sequential search process used by, for example, Ross (1969), Weitzman (1979), and others.[4] At the beginning of a project, the researcher faces a large collection of leads of varying quality. N of these research leads are identified in Phase I as worthy investiganion, and these are to be sequentially tested for a success. Each test entails a cost of c; a successful test yields a gross revenue of R.

We assume that in Phase I the researcher is able to determine whether a lead's quality exceeds some threshold, but he is unable to assess the quality precisely. A reasonable value of this threshold is the point at which he would be indifferent between keeping and culling the lead. This point is where expected benefits equal expected costs. Thus, the researcher keeps those leads for which he believes the probability of success exceeds the threshold c/R.

Once this collection of worthy leads has been identified, the researcher may optionally gather additional information to permit more efficient ordering of the worthy collection. The details will

---

[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 strategic, or "patent race", literature in which multiple firms are competing to achieve the same success (see, e.g. Reinganum (1982), Fudenberg et al. (1983), and Gallini and Kotowitz (1985)).

[4] Our basic setup is a special case of Weitzman's "Pandora's box" problem. Weitzman does not analyze the value of efficient search.

depend on the specific R&D application, but may involve preliminary testing, developing a scientific framework, or acquiring indigenous knowledge. For simplicity, we assume this choice is dichotomous: either the researcher collects a fixed amount of information, or he collects none.

The difference between the information sets available to a researcher who collects the information and a researcher who collects none is formalized as follows. We assume that the uninformed researcher knows the distribution of probabilities of success, but does not know the probability of success for any particular lead. To be precise, the researcher knows how many leads of each probability there are, but does not know which lead maps to which probability.[5] Under this assumption, the uninformed researcher has no way to distinguish the quality of leads and so the collection must be examined at random. On the other hand, by gathering this information, the researcher would learn precisely the success probabilities of all leads, $\{p_i\}$, $i = 1, ..., N$.

While such information about $\{p_i\}$ improves search efficiency, its acquisition will be costly. Hence, the costs of information must be balanced against expected benefits. The presence of 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 "value of efficient search" will then be the difference in the *ex ante* expected value of the collection of leads under optimal (informed) search and random (uninformed) search. In other words, this measure 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 chooses not to acquire this information, and thus conducts the search (over the same collection of research leads) at random. Our calculations represent an upper bound on the value of efficient search because we conservatively assume that an uninformed researcher, who skips Phase II, has no prior information to distinguish between leads in the worthy collection.

---

[5] There are, of course, other ways to account for uncertainty that differ in the details. The selected approach is particularly tractable.

2.3 Value of a collection

In this section, we calculate the value of optimally ordering a set of leads in the search queue. Let S be the collection of N research leads with an associated set of probabilities $\{p\}$. By 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.

Denote a specific ordering of the N research leads in set S by S' with associated probabilities $\{p'_1, p'_2, ..., p'_N\}$. In this arbitrary ordering, a subscript of n indicates the lead will be the nth tested. The expected value of the ordered collection S' is given by

$$V(S') = \sum_{n=1}^{N}(p'_n R - c)a_n(S') \tag{1}$$

where $a_n(S') = \prod_{i=1}^{n-1}(1 - p'_i)$ is the probability that every preceding lead has been tested unsuccessfully (thus $a_1(\cdot) = 1$). Equation (1) has an intuitive explanation. Conditional on the search continuing to the nth lead, the direct profit from the lead is expected revenues $p'_n R$ minus cost c. The probability of getting to the nth lead is a_n. This formula can also be written as the mathematically equivalent expression

$$V(S') = (1 - a_{N+1})R - c\sum_{n=1}^{N}a_n(S'). \tag{2}$$

Here, the term $(1 - a_{N+1})$ is the overall probability of success because a_{N+1} is the probability of reaching the end of the search queue without success. The term $\sum_{n=1}^{N}a_n(S')$ is the expected search duration — that is, the expected number of tests until success.

2.4 Value of improving search efficiency

We next use the basic framework from the previous section to characterize the value of enhancing search efficiency. Consider two researchers who search the same collection of leads, but who queue the leads in different order.

Researcher 1 completes Phase I but skips Phase II and thus has no supplemental information distinguishing one lead in the collection from another. This researcher's search is thus more haphazard, and, we assume as a benchmark, fully random. We denote the *ex ante* expected value of the

collection for Researcher 1 by V_1. This is simply the sum of the values of all possible search queues divided by the number of such queues, as follows:

$$V_1 = \mathrm{E}V(S) = \frac{\sum V(S')}{N!}, \tag{3}$$

where the expectation and summation are across all $N!$ distinct permutations of search order for set S.

Researcher 2 completes both Phase I and Phase II, and thus knows precisely which lead maps to which probability of success; this information allows him to more efficiently search the collection. In other words, he has a well-developed subjective belief about the success probabilities of the research leads, and can therefore order them optimally. Applying Weitzman's (1979), this informed researcher should queue leads in descending order of p_i/c. We denote Researcher 2's optimally ordered queue by S^*. The *ex ante* expected value of the collection is then $V_2 = V(S^*)$.

How much would a researcher be willing to pay, *ex ante*, for the queue S^* rather than some alternative ordering of the same leads, S'? This value is given by:

$$V(S^*) - V(S') = c \left(\sum_{n=1}^{N} a_n(S') - \sum_{n=1}^{N} a_n(S^*) \right) \tag{4}$$

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: $a_{N+1}(S^*) = a_{N+1}(S')$ for *any* alternative ordering, S'. Rather, the value depends on expected search durations and on the test cost, c. This emphasizes that efficient ordering has value because it can lead to success with fewer tests.

Returning to the comparison between an optimally ordered collection (S^*) and a randomly ordered collection, we denote the difference in value between Researcher 2 and Researcher 1 by

$$\Theta_{12} \equiv V_2 - V_1, \tag{5}$$

which gives the value of efficient search of the collection of leads. The remainder of the paper explores the properties, magnitude, and empirical significance of this nominal measure of the value of efficient search.

A supplemental measure in percentage terms is

$$\Pi_{12} \equiv (V_2 - V_1)/V_2 = \Theta_{12}/V_2.$$

This latter expression is simply 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. This measure indicates, for example, whether the possibility of ordering information changes the fundamental nature of a search problem.

2.5 Implications for information gathering

Ultimately, the question faced by the researcher is whether the benefits from acquiring the information in Phase II (that is, identifying the p_i's) exceeds the cost of the information. The main focus of this paper is analyzing the benefits, Θ_{12}. However, for completeness, we briefly formalize here the benefits versus cost comparison. Suppose the cost of acquiring Phase II information on each lead is k. Then, acquiring information on the entire pool of leads would be kN. Letting ΔS be the expected number of tests saved by Phase II information, the value of efficient search $\Theta_{12} = c\,\Delta S$. So, the benefits of the information exceed the costs if and only if

$$k < c\frac{\Delta S}{N} \tag{6}$$

Equation 6 reveals that the researcher should engage in Phase II if the cost of information is less than a certain fraction of the full search cost for a lead. This fraction ($\frac{\Delta S}{N}$) can be interpreted as the proportion of tests saved relative to an exhaustive search. For example if the information is expected to save 10 tests, and there are 1000 leads in the collection, then acquiring the additional information would be profitable if the per-lead information cost were less than 1% of the cost of fully testing a lead for a success.

3 Analyzing the theoretical model

In this section, we analyze the theoretical model laid out in Section 2 to derive general properties of the value of efficient search. We begin by examining the features of an R&D problem that tend to make gathering information more, or less, valuable. We then use those results to derive an upper

bound on the value of efficient search. Further, we identify and explore an inherent tension that suppresses the value of efficient search.

3.1   Exploring the value of efficient search

The value of efficient search Θ_{12} is naturally determined by the parameters of the model: R, c, and the characteristics of the collection of leads. For a given collection of worthy leads, it is straightforward by equation 4 to observe that Θ_{12} is independent of R and linearly increasing in c. However, the dependence of Θ_{12} on the characteristics of the leads can be quite complex in general, reflecting the interplay of the success probabilities of the various leads. Here we examine how the characteristics of distribution of lead qualities impact Θ_{12}.

The most intuitive result is that Θ_{12} should become larger if the leads are more diverse, or more *spread*. In contrast, in a collection of homogenous leads it makes no difference in which order one searches the leads. On the other hand, if one lead is superior (higher success probability) to the remainder, then the value of placing this lead first must be greater the more superior the lead.

To formalize this notion, 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 first condition requires that the new probabilities q are more diverse than the old set p. The second condition is a normalization which preserves the overall probability of success in the search, $1 - \prod_{k=1}^{N}(1 - p_k)$. Under this definition increased spread decreases the value of random search and increases the value of ordered search.

Proposition 1 *The value of efficient search Θ_{12} increases if the* spread *of success probabilities increases.*

The magnitude of the probabilities themselves is another important determinant of the value of efficient search. In a search over leads of generally poor quality, the value of the collection, whether searched in an efficient or random order, is generally low. But the effect on the value of efficient search is less obvious. It turns out that if the leads are of generally poor quality, the value

of efficient search will tend to be high. As overall lead qualities improve, the expected number of searches declines under both optimal and random orderings; however, the decline is greater under random search. Similarly as lead qualities improve, the expected value of search increases under both optimal and random orderings. Thus, there is a tendency for efficient search to be most valuable when the potential value of search is low. This notion is formalized as an additive decrease in the success probabilities, as follows:

Proposition 2 *The value of efficient search Θ_{12} increases if the lead probabilities are decreased by an additive factor.*

We have shown that more diverse but generally low probabilities tend to increase the value of efficient search. The size of the pool of acceptable leads (i.e. N) is also of interest. Is information that facilitates efficient search more valuable when the researcher faces a large pool of leads or a small one? For a small pool of leads, even the worst-case search duration is small. As the pool of leads grows, so does the number of possible tests, and thus the possibility of avoiding some of those tests by exploiting Phase II information. Intuitively, then it would seem that the value of efficient search would grow as the pool of leads expands.

To show this formally requires a precise notion of what it means to expand the pool size. If a given pool is expanded by adding very high quality leads, a different result about the value of efficient search might obtain than by adding low quality leads. So we would like to construct an approach to changing the pool size that would not change the underlying lead qualities in a fundamental way. This would allow us to disentangle the effect of size from the effect of changing probabilities. One reasonable approach is to assess the impact of dropping a randomly selected lead from the pool. For this measure, the pool characteristics do not change on average except for the number of leads searched.

Proposition 3 *The value of efficient search Θ_{12} decreases in expectation if the pool of acceptable leads is shrunk by omitting a randomly selected lead.*

A different approach to changing pool size, with a similar outcome, is addressed in Proposition 5

below.

3.2 Bounding the value of efficient search

The results in the previous section allow us to bound the value of efficient search. At one extreme, if all leads are of identical quality, the search order is irrelevant. In that case, $\Theta_{12} = 0$. If the leads are instead heterogeneous, the value of efficient search will depend in a complicated way on the probabilities of all the leads, as shown in equation 4. 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 issue by deriving the non-parametric distribution of probabilities that gives the theoretically maximal value of efficient search.

What non-parametric probability distribution gives rise to the largest value of Θ_{12}? By virtue of Phase I, 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 $p_n R - c$ and so if $p_n < c/R$ the lead is dropped.[6] By Proposition 1, a maximally *spread* distribution yields the highest value of efficient search. The maximally spread distribution is obtained by setting $N-1$ leads to the lowest possible quality for a collection individually worthy of search, $p_2, \ldots, p_N = 1-\delta$, where $\delta = 1-c/R$. Then the remaining lead contains a certain success: $p_1 = 1$. Under this probability distribution, Researcher 1 must wade at random through a large number of almost worthless leads before eventually succeeding, while Researcher 2, conducting an optimal search, is guaranteed to find success on the first trial. Proposition 4 below calculates the value of efficient search under this probability distribution. Because Θ_{12} would be smaller under any other probability distribution, this is an upper bound on the value of efficient search.

Proposition 4 *The nominal value of Phase II R&D, Θ_{12}, cannot exceed*

$$\bar{\Theta}_{12} = c\delta \frac{N(1-\delta) + \delta^N - 1}{N(1-\delta)^2}. \tag{7}$$

---

[6] Note that under random search the "cutoff" value could be slightly larger than c/R, so the pool of acceptable leads could me marginally smaller. 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. Calculating analytically the precise cutoff value is extremely complex. However, simulations reveal that the cutoff is numerically extremely close to c/R (typically within less than 1%).

Calculated in percentage terms, the upper bound becomes

$$\bar{\Pi}_{12} = 100 \left[1 - \frac{1 - \delta^N}{N(1 - \delta)} \right]. \tag{8}$$

The calculations in Proposition 4 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, these provide a practical upper bound on the value of information in 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 ratio of search cost to revenue ($c/R = 1 - \delta$). We elaborate on this point in section 4.

What intuition about the value of efficient search can we glean from this proposition? First consider the case where the individual search cost c is low. Recall, that the maximally spread distribution sets the $N - 1$ worst leads at probabilities $p_2, \ldots, p_N = c/R$. If c is low, then even relatively poor quality leads are retained, and these leads are likely to result in failure. Since random search entails wading through a large number of very poor quality leads, ordered search is likely to dramatically reduce the number of searches. In the limiting case of $c = 0$, the efficient search saves an expected $\frac{N-1}{2}$ searches. However, this produces no value, $\bar{\Theta}_{12} = 0$, exactly because there is no search cost.[7]

On the other hand, suppose c is large. Then, the poor lead qualities in the maximal spread distribution have relatively high probability. In this case, a random search is likely to be successful early on because even the poor leads are relatively good. So, an efficient ordering only slightly diminishes the expected number of tests relative to random search. Consequently, the maximal value of efficient search is also low in the case of high c. In the limiting case of $c = R$, $\bar{\Theta}_{12} = 0$. At the two extremes of c, then, even the theoretically maximal value of efficient search is zero.

---

[7] Repeated application of L'Hopital's rule obtains: $\lim_{\delta \to 1} \bar{\Theta}_{12} = c \frac{N-1}{2} = 0$.

3.3 Exploring a tension: intensive vs. extensive margin

In the maximally spread probability distribution considered in section 3.2, the pivotal parameter was search cost, c. Returning to the case with a general, as opposed to maximally spread, probability distribution of leads, equation 4 emphasizes the importance of search cost in the value of efficient search. If search cost is low, an efficient ordering 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.

However, there is a second and countervailing effect on the extensive margin. Larger values of c raise the Phase I cutoff for the minimum acceptable probability. So, the pool of acceptable leads shrinks as c grows larger. We 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.

Proposition 5 : *The value of efficient search, Θ_{12} decreases if the lowest probability lead is dropped from a collection of leads.*

The tension between the intensive and extensive margin effects implicit in Proposition 5 is made explicit in Proposition 6 below.

Proposition 6 : *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.*

The tension in Proposition 6 can be illuminated by analogy with a continuous representation of research leads.[8] We define by $N(c)$ the number of leads that remain after discarding poor leads as a function of search cost c. Then, we can think of the value of efficient search as a function,

---

[8] While the point here is to build intuition, this notion of continuity can be made concrete. Suppose that, rather than atoms of success probability $p_1, \ldots, p_N$, we used a smoothed approximation to discrete leads $p(n)$, such that $\int_{n-1}^{n} p(t)dt = p_n$. Under this continuous approximation, the value of ordering information works out to be $\Theta_{12} = \frac{c}{\bar{p}} \int_0^N (p(n) - \bar{p}) e^{\int_0^n \log 1 - p(m)dm} dn$, where $\bar{p}$ is the average of lead probabilities.

$\Theta_{12}(c, N(c))$. The effect of search cost on the value of efficient search can be then be seen intuitively by taking the derivative of this expression with respect to c:

$$\frac{d\Theta_{12}}{dc} = \frac{\partial \Theta_{12}}{\partial c} + \frac{\partial \Theta_{12}}{\partial N}\frac{\partial N}{\partial c}. \tag{9}$$

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 represents the extensive margin, and is always negative. It consists first of the term $\frac{\partial \Theta_{12}}{\partial N}$, which is positive by Proposition 5. The term $\frac{\partial N}{\partial c}$ is negative, reflecting that increasing c decreases the pool of leads worthy of search. This tension means that the effect of increasing c on Θ_{12} may be negative over some range. In fact, as c gets larger the extensive margin effect dominates; for large enough c the value of efficient search must be zero, $\Theta_{12} = 0$.

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 when revenue on success is $R = 1000$. 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 6, the extensive margin effect eventually dominates, driving Θ_{12} to 0.

4   Bioprospecting as an R&D search

In this section we apply the theoretical results from sections 2 and 3 to a problem specifically chosen because its features are favorable to a high value of efficient search. This is the R&D problem of bioprospecting — the search for valuable pharmaceutical products (such as a cure for cancer) in nature. This is a subject with some pedigree in the economics literature. Below we consider two competing analyses from this literature. A key difference between these analyses is the problem of ordered versus unordered search — precisely the focus of this paper.

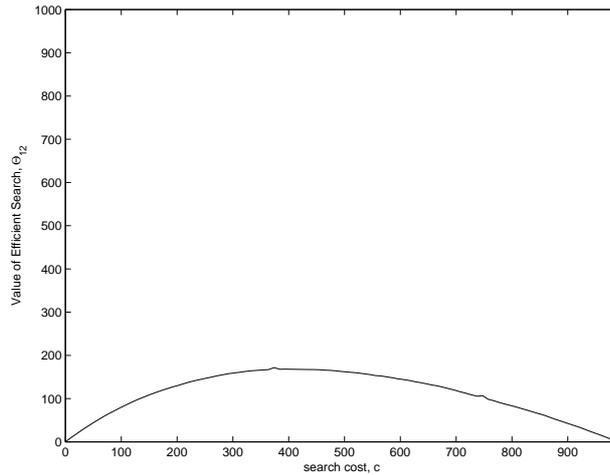Simpson et al. (1996) were the first to consider the search aspect of the bioprospecting problem.

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

They were primarily interested in the conservation implications from bioprospecting and thus focused on the marginal values of research leads — endemic plant species in biodiversity hotspots around the world. In this model, the researcher has no prior information on how to order the search, and so all leads are subjectively viewed as exchangeable and search order as irrelevant. The search terminates upon the identification of a natural compound with high pharmaceutical value. They find that the marginal value of a species is likely to be small regardless of the (common) probability of success $\tilde{p} = .000012$. When $\tilde{p}$ is low, the marginal value of a species is small because any one species is unlikely to produce a success. When $\tilde{p}$ is high, the marginal value is small because species are close substitutes, so any one species is likely to be redundant. They concluded that the maximum value per hectare across all hotspots was about \$21; this is for Western Ecuador — the most biodiverse region in their study.

The conclusion of Rausser and Small (2000) stands in sharp contrast. Analyzing the same 18 biodiversity hotspots considered by Simpson et al., they found substantial private-sector incentives for biodiversity conservation. In this model, success probabilities are heterogeneous (they span the range [.0000011, .00011]), and so the researcher benefits from ordering his search to improve efficiency. The average probability of success across all species is, however, nearly the same as $\tilde{p}$

16

from the homogenous model. Under ordered search Rausser and Small concluded that the most valuable area (Western Ecuador again) has a bioprospecting value of \$9,177 given an efficient search queue. They attributed this dramatic increase (from \$21/hectare to \$9,177/hectare) to more efficient search from exploiting additional knowledge about lead qualities. The large discrepancy in value estimates between these two studies is suggestive that efficient search may indeed have high value in this case.[9] We examine this conjecture 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. Suppose that in addition to the 18 biodiversity regions originally considered in both papers cited above, the collection had also contained a less promising biodiversity region such as Argentina. There are about 1100 endemic plant species over 273,669 kilo-hectares in Argentina, implying a success probability per kilo-hectare of p_A =4.8E-8.[10] Since $p_A R < c$, Argentina should be eliminated from the collection in Phase I. However, a researcher with no information whatsoever would randomly search all hectares, including those in Argentina. Given such random search, we apply equation 3 and find that 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 are cases 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 that the entire United States was included in the original collection of biodiversity zones.[11] This region contains 1900 endemic plant species over 891,296 kilo-hectares for a probability of success of p_{US} =2.5E-8.

[9] Another possible explanation is different choices of parameters.

[10] We use the Rausser and Small parameters: $p_i = .000012 * e_i$ (where e_i is the density of plant species), $c = 485$, and R =\$450 million.

[11] This addition excludes the California Floristic Province hotspot, which was already part of the analysis.

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, the coarse-grained sorting of leads in Phase I substantially increases the value of the collection by eliminating poor quality leads entirely. Even if the remaining collection of leads is searched in random order, the value of the collection is still vastly improved.

4.2   Phase II ordering

We next apply the value of efficient search model derived in section 2. The central message of Rausser and Small is that a more efficient approach to bioprospecting can dramatically increase the marginal value of the most biodiverse land. This increase (from the \$21/hectare to \$9,177/hectare) is a 440-fold increase over the results of Simpson et al. Therefore, exploiting a scientific framework that enables an efficient search order should provide significant private sector incentives for conservation. Here we carefully examine that result.

As a point of departure, we conduct the following experiment. Using the model and parameter values of Rausser and Small, we 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 2, who has already conducted Phase II of R&D and therefore searches the pool of leads in the most efficient order — in descending order of the probability of success. The second case is analogous to our Researcher 1, 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 probability of success. For each case we calculate the *ex ante* expected value of the collection in the search for a single success. The table below 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 74,640 leads (kilo-hectares in biodiversity hotspots) is reduced by only about $3 million (2.3%) if the researcher is forced to search for a success at random rather than in fully efficient order. 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, a 4.9% loss).

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 (74,640 kilo-hectares). The success probabilities are all quite low, with the largest being 0.000105. And, the probabilities vary dramatically in magnitude, with the smallest probability about two orders of magnitude smaller than the largest. 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 the comparative statics results in section 3 (Propositions 1, 2, and 3) reveal the opposite; each of these factors tends to increase the value efficient search over a given collection of leads. This is, therefore, an example favorable to a high value of 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, indigenous knowledge that identifies plants that are likely to contain compounds of pharmaceutical value. Implementing equation 6, 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 that, for this example, 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 rule out the possibility that some peculiar feature of the probability distribution itself is responsible for the
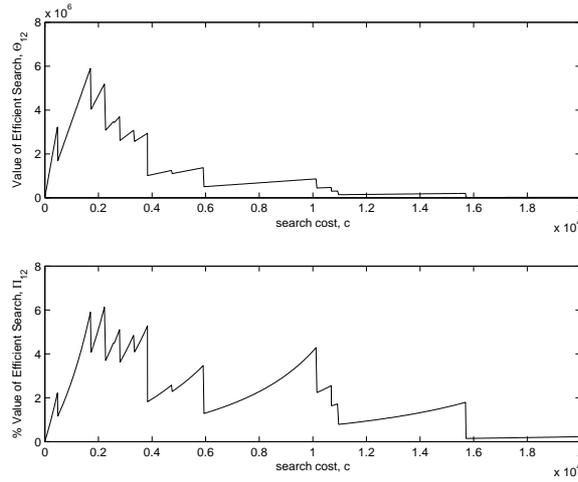
Figure 2: Value of efficient search (top) and percentage value relative to V_2 (bottom) as a function of search cost, c increases.

low value of efficient search we implement Proposition 4 which derives the probability distribution that yields the theoretically maximal value of efficient search. But even with the theoretically most favorable set of probabilities (74,639 leads at c/R and 1 lead at 1), the value is still small (in percentage terms $\bar{\Pi}_{12}$=3.9%, and in nominal terms $\bar{\Theta}_{12}$ =$17.6 million).

4.3   Importance of search cost

Why, even for this favorably engineered example, is the value of efficient search so low? Careful inspection of the parameters used in this application reveals that the search cost is too low to ever yield a high value of efficient search. The cost of even an exhaustive search is small ($36 million = $485*74,640) relative to the revenue on success ($450 million). If c were higher, the relevance of search efficiency would increase on the intensive margin. On the other hand, if c were higher, an efficient search would entail eliminating some leads from the queue. Overall, how well will the researchers fare under different values of c?

Figure 2 illustrates the empirical consequence of search cost on the value of efficient search. The top panel presents results for the nominal value (Θ_{12}) and the bottom panel presents results for the percentage value (Π_{12}). Again, we find that even favorable choices of c cannot produce a high

value of efficient search. 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 6.

We next turn to a concluding discussion which reviews these specific results and examines implications for the R&D process more broadly.

5 Discussion

One purpose of research is to resolve uncertainty and thus inform better decisions. Whether searching for a new medicine, a marketable invention, or an elegant proof of a theorem, efficient search must have higher value than random investigation of the same collection. Of course, information identifying a collection of leads worthy of search is of paramount importance. However, because acquiring information is usually costly, the pivotal question given a collection of leads is: what is the value of information facilitating more efficient search?

We identified the inherent tension between the intensive and extensive margins of search. The value of information increases on the intensive margin. However, the value decreases on the extensive margin, tending to suppress the value. When search cost is low, search efficiency is less important because searching lower probability leads is 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. For sufficiently high search cost, this effect always dominates.

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. Even if the lead probabilities were chosen to maximize the value of efficient search in the bioprospecting problem, the upper bound would be 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,

higher overall probabilities, or more homogenous leads) 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. While the previous literature has focused on that aspect of search, information that allows the researcher to eliminate very low quality leads from the collection (Phase I) may have tremendous value, even when the good leads are subsequently searched at random. In many cases, information enabling coarse sorting can provide the 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 basic principles of this search problem apply to any research inquiry. In all search applications of this kind, the researcher faces the conceptually separate questions of *what* leads to search and *how* to order the search queue of the worthy collection. Although the R&D literature to date has focused on *how to search efficiently*, our analysis suggests that the former question of *what to search* is much more crucial.

6 Appendix

Proof of Proposition 1. 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. As a concrete example, and without loss of generality, 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 search duration under the first permutation is $1 + P_1 + P_1 P_i + P_1 P_i P_2 + P_1 P_i P_2 P_j + P_1 P_i P_2 P_j P_3$. The average search duration over both permutations is thus $1 + P_1 + (P_1 + P_1 P_2)(P_i + P_j)/2 + P_1 P_i P_2 P_j + P_1 P_i P_2 P_j P_3$. Now consider an increase in spread that increases p_i to q_i and decreases p_j to q_j such that $P_i P_j = Q_i Q_j$, where $Q_k = 1 - q_k$. For the new set of probabilities, the only terms that change in the average duration 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 duration attributed to the spread is $(P_1 + P_1 P_2)(Q_i + Q_j - P_i - P_j)/2$. This term is positive since $P_i P_j = Q_i Q_j$ and $Q_j > P_j > P_i > Q_i$, which can be seen from simple algebraic manipulation. This same approach applies to *any* pair of permutations which switches the positions of leads i and j. Averaging over all such pairs gives the expected random search duration, which increases since each component of the average does. So the value of a random search *decreases* with greater spread.

In contrast the value of ordered search *increases* with greater spread. Consider again the expression for search duration in terms of P (see above). Terms in this expression containing P_i, but not P_j, decrease with spread because $P_i < P_j$ for ordered search. Terms containing both P_i and P_j are unchanged with spread because of the normalization. This implies that duration in an optimally ordered sequence decreases with spread; therefore, the search value increases with greater spread. Combining this with the result above on random search value, it follows that the value of information Θ_{12} increases with greater spread. ■

23

Proof of Proposition 2. Recall from equation 2 that the expected search duration for sequence S' is $\sum_{i=1}^{N} a_i'$. Since we are evaluating the effects of a small additive perturbation to the probabilities, we define $a_i'(\epsilon)$ analogously to a_i' by $a_i'(\epsilon) = \prod_{j=1}^{i-1}(1 - (p_j' + \epsilon))$. Note that $a_i'(0) \equiv a_i'$. Then, $\sum_{i=1}^{N} a_i'(\epsilon)$ is the expected search duration for sequence S' when each probability is increased by ϵ.

Let S^* be the optimally ordered search. The value of efficient search, as a function of the perturbation ϵ, is given by:

$$\Theta(\epsilon) = c \sum_{i=1}^{N} \big(a_i'(\epsilon) - a_i^*(\epsilon) \big). \tag{10}$$

The derivative of $\Theta(\epsilon)$, evaluated at $\epsilon = 0$, is given by

$$\left. \frac{d\Theta(\epsilon)}{d\epsilon} \right|_{\epsilon=0} = c \sum_{i=1}^{N} \left[\sum_{j=1}^{i-1} \prod_{k \neq j}^{i-1}(1 - p_k^*) - \sum_{j=1}^{i-1} \prod_{k \neq j}^{i-1}(1 - p_k') \right] \tag{11}$$

The first sum of products (over p^*) is the minimum possible because S^* is sequenced in descending order of the probabilities. Therefore, the term in brackets is negative for all i, and so the entire derivative is negative. Thus, *increasing* probabilities a common amount *decreases* the value of an optimal search sequence S^* relative to any alternative S'. ∎

Proof of Proposition 3. First, we examine the change in duration for random search, then we turn to ordered search. Let S be the set of all leads, and let S_t be the same set omitting only lead t.

The expected duration of a random search is found by averaging over all $N!$ search permutations of S. Similarly, for a search where lead t is omitted, there are $(N-1)!$ search permutations of S_t. Since there are N leads which may be omitted, there are $N(N-1)! = N!$ distinct permutations of search with one lead dropped. Each search permutation of S can be uniquely paired with a permutation from S_t for some t. Suppose lead t is placed last in the search permutation S'. Then, let S_t' (which omits lead t) be the search sequence which is identical in the first $N-1$ positions with S'. The difference in expected search duration $D(\cdot)$ between S' and S_t' is then simply the probability that the search queue S_t' is exhausted without a success, or equivalently, that the search of S' reaches the last lead without prior success. The probability of this event is $\prod_{i \neq t}^{N}(1 - p_i)$. Averaging this expression over all permutations of S' and over all omitted leads t, we find that the expected change

in the duration of a random search from dropping a randomly selected lead is

$$\mathrm{E}\frac{1}{N}\sum_{t=1}^{N}D(S') - D(S'_t) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j\neq i}^{N}(1-p_j),\tag{12}$$

where the expectation operator is over search sequences.

Turning to ordered search, the expected change from dropping a random lead can be assessed by simple subtraction of the formula in equation 2, yielding

$$D(S^*) - D(S_t^*) = a_t^* - \frac{p_t^*}{1-p_t^*}\sum_{i=t+1}^{N}a_i^*\tag{13}$$

Summing that expression over t, where each lead is dropped in turn, yields

$$\sum_{t=1}^{N}D(S^*) - D(S_t^*) = \sum_{t=1}^{N}\Big(a_t^* - \frac{p_t^*}{(1-p_t^*)}\sum_{i=t+1}^{N}a_i^*\Big) = \sum_{t=1}^{N}a_t^*\Big(1 - \sum_{i=1}^{t-1}\frac{p_i^*}{(1-p_i^*)}\Big),\tag{14}$$

where the third expression comes from switching the order of summation. By algebraic manipulation, this expression expands to

$$\frac{1}{N}\sum_{t=1}^{N}D(S^*) - D(S_t^*) = \frac{1}{N}\sum_{t=1}^{N}a_t^*\Big(t - \sum_{j=1}^{t-1}\frac{1}{(1-p_j^*)}\Big) \leq \frac{1}{N}\sum_{t=1}^{N}a_t^*\Big(t - \frac{t-1}{(1-p_{t-1}^*)}\Big) = a_N^*.\tag{15}$$

Here, the inequality between the second and third expression results from replacing the terms p_j^* with the (smaller) term p_{t-1}^* in the inner summation. The equality between the third and last expression results from simple cancelation of terms.

Now compare the change in duration for random search in (12) with the upper bound on the change in duration for ordered search in (15). Because $a_N^* = \prod_{i=1}^{N-1}(1-p_i^*)$ is the smallest possible product over $N-1$ leads, the exact formula for the change in random search duration exceeds the upper bound for the change in ordered search duration. Because the value of information itself does not depend on expected revenues, the change in the value of information from dropping a lead is only a function of the change in the expected search durations. Therefore, the proposition follows immediately from the result on durations. ∎

Proof of Proposition 4. Let $\delta = 1 - c/R$. By Proposition 1 the maximally spread distribution will give an upper bound for Θ_{12}. Given N leads, the maximally spread distribution has $N-1$ leads with probability $1 - \delta$. The remaining lead has probability 1.

Given this distribution, the expected search duration can be calculated by averaging over all N distinct positions for the high-probability lead. If the high-probability lead is in the nth position, then the expected search duration is $\frac{1-\delta^n}{1-\delta}$. Under random search, the high-probability lead is equally likely to be positioned in any of the N positions. The expected search duration D_1 is then given by averaging over the N possible positions

$$D_1 = \frac{\sum_{i=1}^{N} 1 - \delta^i}{N(1-\delta)} = \frac{N(1-\delta) + \delta^{N+1} - \delta}{N(1-\delta)^2} \tag{16}$$

The expected search duration under an optimal queue is $D_2 = 1$ (because a sure success is searched first). The value of efficient search is simply $\bar{\Theta}_{12} = c(D_1 - D_2)$, which simplifies to

$$\bar{\Theta}_{12} = c\delta \frac{N(1-\delta) + \delta^N - 1}{N(1-\delta)^2} \tag{17}$$

■

Proof of Proposition 5.   Let the success probability of the lowest quality lead in a given sequence S' be p'_t. In the optimally ordered sequence S^* the same lead will be in last place, so $p'_t \equiv p^*_N$. Let S'_t and S^*_N be the corresponding sequences with the lowest lead dropped. Let the change in the value of efficient search from dropping a lead be $\Delta\Theta = \big(V(S^*_N) - V(S'_t)\big) - \big(V(S^*) - V(S')\big)$. Expanding out this expression (see equation 4) we have

$$\Delta\Theta = c \left(a'_t - \frac{p'_t}{1 - p'_t} \sum_{i=t+1}^{N} a'_i - a^*_N \right). \tag{18}$$

It is convenient here to generalize the definition of a. Let $a_{i,j} = \prod_{k=i}^{j-1}(1 - p_k)$. Note that $a_i \equiv a_{1,i}$. Since the worst lead goes last in the optimal queue, $a^*_{1,N} = a'_{1,t} \, a'_{t+1,N+1}$. Using this fact, then factoring out $a'_{1,t}$ yields

$$\Delta\Theta = c \, a'_{1,t} \Big(1 - \frac{p'_t}{1 - p'_t} \sum_{i=t+1}^{N} a'_{t,i} - a'_{t+1,N+1} \Big). \tag{19}$$

We next bound these terms. Since p'_t is the worst lead, $a'_{t,i} \le (1 - p'_t)^{i-t}$. So,

$$\sum_{i=t+1}^{N} a'_{t,i} \le \sum_{i=t+1}^{N} (1 - p'_t)^{i-t} = \Big(1 - (1 - p'_t)^{N-t}\Big) \frac{(1 - p'_t)}{p'_t}. \tag{20}$$

Substituting this bound into the previous equation,

$$\Delta\Theta \ge c \, a'_{1,t} \Big((1 - p'_t)^{N-t} - a'_{t+1,N+1} \Big). \tag{21}$$

By the same argument, $a'_{t+1,N+1} \leq (1 - p'_t)^{N-t}$. 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 Proposition 6.

(a) Equation 4 shows that on the intensive margin Θ_{12} is linear in c, with a positive slope.

(b) Proposition 5 shows that on the extensive margin Θ_{12} decreases when a lead is dropped.

(c) Let c be large enough that only lead(s) 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 between them.

∎

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.

Friedel, R. and P. Israel (1986). *Edison's Electric Light: Biography of an Invention*. Rutgers University Press.

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.

Reinganum, J. (1982). Stategic search theory. *International Economic Review 23*(1), 1–17.

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.

Vishwanath, T. (1988). Parallel search and information gathering. *The American Economic Review 78*(2), 110–116.

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