The Eureka Effect

The Art and Logic of Breakthrough Thinking

DAVID PERKINS

"Perkins' style is engaging–not for eggheads only–and the brainteasers are entertaining and surprisingly fresh."
–CHICAGO TRIBUNE

tions, and criteria for progress and success are not usually as neatly or overtly defined as in chess. But after all, chess and many other games are deliberately constructed in lean, logical ways to remove the fuzziness characteristic of real problem-solving situations. Whether the world at hand is the world of chess, cats, apple pickers, poems, paintings, scientific theories, bridge architecture, or recipes for bouillabaisse, the language of spaces of possible states, operations, and gauges of progress and success provides a conceptual system for examining the process of thought and contrasting the demands of different kinds of problems.

When Smart Is Reasonable

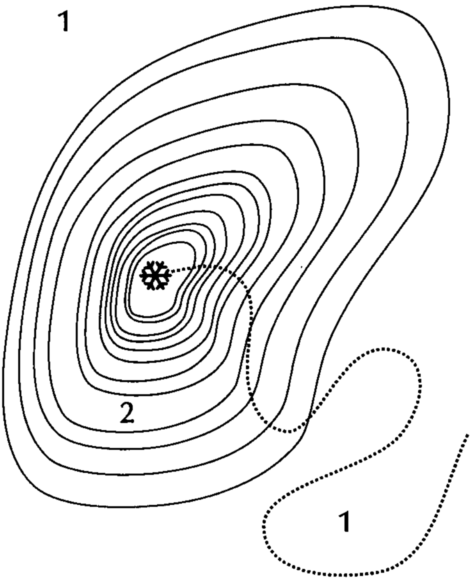
With all this as background, we turn to the contrast between reasonable and unreasonable problems. Both are a matter of search. In the one case, reasoning offers a smart way of searching, but in the other it does not.

The simplest search strategy is to examine all the possibilities. In some situations this is the right thing to do. Most people at some point master the familiar game of tic-tac-toe, either learning the ropes from someone older or figuring it out. I remember getting annoyed when my father kept beating me, so I sat down with paper and pencil and figured it out. When one takes into account the symmetries, there are really only three places to start—corner, side, or middle. For the next play, there are only a few other genuinely different possibilities. From then on, I could hold my own.

Examining all the possibilities is not always a trivial strategy. It requires attentive and systematic action to be sure not to miss possibilities. Nonetheless, for most problems examining all the possibilities is not smart strategy because there are too many to investigate. The DONALD + GERALD = ROBERT cryptarithmetic puzzle has ten different letters, standing for the ten digits 0 through 9. It's stated that D = 5, so there are nine remaining letters to assign to the nine remaining digits. Mathematically, this might be done in 362,880 different ways. A problem solver certainly would not want to consider all those possibilities. In general, in problems that are at all challenging, there are too many states and too few solution states to find solutions by searching thoroughly. What is needed is another style of smart search. Cognitive scientists speak of heuristic search, the term *heuristic* referring to strategies that increase the chances of success without guaranteeing success.

A basic strategy of smart search is to follow the measure of promise, tracking its increases through the fitness landscape to a solution. In the case of a cryptarithmetic problem, this means proceeding incrementally, starting with one letter-digit assignment consistent with the information given, then adding another, then adding another until all are in place. The key to this strategy is to use the logic of the situation. A smart search would try to find a digit-letter assignment forced by the given information—that is, one that could not be any other way—and then another and another, drawing out any immediate implications and exploring alternative branches only when necessary. When there are many possibilities, smart search examines far less than the entire space.

The strategy of following promise applies not only to formal problems but to fuzzy ones too. Consider the woman searching through different possibilities for an apple picker and how one step led to the next. The search was progressive: first a hook, but the apples would bruise on the ground; then a bag, but it would get unbalanced; then a counterweight, but that would be heavy too; then a cloth tunnel. In a loose sense, this progression is like solving a cryptarithmetic problem. The first idea contributed part of the solution, leading on to other adjustments and extensions that provided a more complete solution.



Search in a Homing Space: 1. Clueless regions.
2. Large clued regions leading to the target.

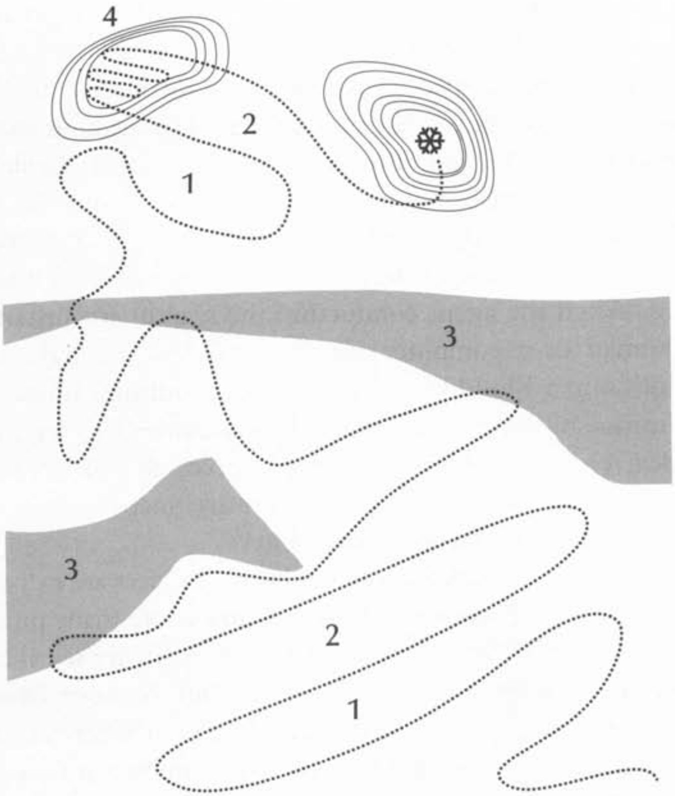
The accompanying figure illustrates in a general way the fitness landscape of a reasonable possibility space—one that lends itself to following promise. As the picture shows, such a possibility space has a relatively simple structure. The contour lines show regions where there is a discernible slope to the measure of promise. In those regions, the problem solver can follow the slope of increasing promise to the solution, homing in on it systematically. Such possibility spaces might be called homing spaces because of this convenient structure.

Is homing easy? Not necessarily. Although following promise is a reasonable approach, it can be a very challenging one. Following promise requires systematic attention and meticulous logic. Following promise means making the most of the given information. Solving a difficult problem of this sort is like climbing a cliff: The climber needs to capitalize on each little niche for a handhold or toehold.

When Smart Is Unreasonable

Challenging though the world of homing spaces can be, it is not the world of Leonardo da Vinci's or the Wright brothers' insights about flight, nor the world of Gutenberg's invention of the printing press, nor Darwin's discovery of natural selection, nor the Sufi tales and their hidden meanings, nor The Nine Dots problem. These problems and puzzles have unreasonable possibility spaces—Klondike spaces. A visualization of a Klondike space appears in the next figure.

Think of the space as huge with, in this case, only one solution in the top right. The size and rarity of solutions reflects a wilderness of possibilities. There are large regions without contours, where the measure of promise points in no particular direction—a clueless plateau. The lower part is boxed off, keeping the search away from the upper regions until it finally breaks through at a thin spot—the lower part is a narrow canyon of exploration. And there is a tempting contoured region that contains no solution—an oasis of false promise where a problem solver might



Search in a Klondike Space: 1. A large space wtih few solutions (a wilderness trap). 2. Regions with no clues pointing direction (plateau traps). 3. A barrier isolates the solution (creating a canyon trap). 4. An area of high promise but no solution (an oasis trap).

linger in hopes of finally finding a solution. Smart search in such a possibility space is a matter of thinking in ways that cope with the wilderness, plateau, canyon, and oasis traps. It's a matter of setting sequential reasoning aside and being unreasonable in a smart way, along the lines of the four operations introduced earlier: roving around flexibly, detecting hidden clues, reframing the situation, and decentering from false promise.

Since Klondike spaces have their own distinctive difficulties, one might hope that the challenges of homing spaces could be left behind. But not so. Notice that Klondike spaces have small homing spaces inside them. In the Klondike figure, the small contoured region in the upper right that contains a solution is a homing space in miniature. When the agent conducting the search (for instance, a human or a computer) finally gets close enough to a solution in a Klondike space to detect promising signs, the solution still has to be extracted. Sometimes this happens reflexively, as the human mind assembles all the information spontaneously in a quick cognitive snap. Sometimes the process of extraction takes longer.

Of course, Klondike spaces and homing spaces are extreme types. Most real problems are mixed, and so are many puzzle problems. Smart search means having a sensitivity to what a problem requires and a readiness to shift between breakthrough thinking and sequential reasoning as the terrain suggests. Klondike spaces and homing spaces are worth thinking about not because they sort the world of problems into two extremes but because they anchor the two ends of a continuum, revealing the continuum more clearly.

The Structure of Breakthrough Thinking

Although wilderness, plateau, canyon, and oasis are metaphors, they have rigorous interpretations as formal features of possibility spaces. They can be described in terms of the state space of possibilities, the available operators, and the indicators of promise and success.

Wilderness of possibilities. In the language of possibility spaces, a wilderness has a large number of possible states, only a few of which are solution states. An effective search process somehow has to cope with the sheer magnitude of the state space and the rarity of solutions.

Clueless plateau. In a possibility space, a plateau is a large region of neighboring possible states where the measure of promise does not vary much, or perhaps varies erratically from state to state around an average for the whole plateau, so there is no trend. On such a plateau, a search process cannot progress from possibility to possibility with a steady improvement in the measure of promise.

Narrow canyon of exploration. In a possibility space, a canyon trap is a solutionless region of many neighboring possible states with a boundary around it that tends to trap the search process. Such a boundary can arise from the available operations. Perhaps only a very few operations from a very few states in the region take the search process outside the region. Metaphorically, there are very few paths out of the canyon. Alternatively, the boundary can arise from the measure of promise. The measure may drop to very low values all the way around the region. Although the search process can penetrate those areas in principle, it tends not to because of the low promise. Also, these two kinds of boundaries can work in combination to create a canyon.

Oasis of false promise. In a possibility space, an oasis is a state where the measure of promise has a relatively high peak that is not quite a solution state. The search process tends to circulate near the deceptive peak in hopes of finding a full solution nearby, rather than venturing off to possibilities of lower promise.

The five-phase pattern of breakthrough thinking introduced in Chapter 1 also finds an explanation in the language of possibility spaces.

1. *Long search.* Why are insights preceded by long searches? Because the space is large with few solutions (wilderness), the measure of promise does not point a clear and systematic direction (plateau), the measure of promise and available operations tend to constrain the search to limited regions (canyon), and the measure of promise yields high points with no real solution that cause the search to linger fruitlessly in their neighborhood (oasis).

2. *Little apparent progress.* Why is the apparent progress minimal for most of the search? The same reasons apply. In particular, only close to a solution does the measure of promise offer a clear guide to homing in on it.

3. *Precipitating event.* What causes precipitating events? A precipitating event can take different forms. It may simply be the arrival of the search process at a small homing subspace within the larger space. The search process then relatively quickly converges on a solution. Alternatively, the precipitating event can be some cue, internal or external, that leads the search process to escape from an oasis or canyon into another region that, relatively quickly searched, leads to a homing region and a solution.

4. *Cognitive snap.* What is the cognitive snap? The rapid homing process that occurs when the search process

finally arrives at a solution-containing homing space within the larger Klondike possibility space. The homing space allows quick convergence to a solution.

5. *Transformation.* Why the sense of transformation? Solutions tend to surprise because they often involve escape from an oasis of false promise or from a narrow canyon of exploration to a solution of a very different kind than expected.

An unusual and important feature of these explanations is that they do not specifically concern the human mind. They have little to do with a creative knack or a peculiarly flexible cerebral cortex. A Klondike possibility space is a breakthrough waiting to happen!

The search process that achieves the breakthrough may unfold in the human mind, as an artist, scientist, inventor, or businessperson explores alternative perspectives. It may occur in a computer, as automated processes of heuristic search examine a large set of possibilities. It may happen in the course of the long blind search of biological evolution, as the random shuffling of genes tries this and that new prototype for survival and reproduction. (This theme is revisited in the last chapters of the book.) Whatever the setting, most fundamentally the phenomena of breakthrough thinking derive from the underlying structure of a Klondike space.

Jack London's Klondike

One of the many outdoor tales of Jack London is a short story called "All Gold Canyon." In a few pages, London relates the struggles of a prospector to find gold. The prospector comes with the basics: a pick, a shovel, a gold pan, and most of all a hungry spirit. He chooses to start his dig in this particular canyon because there is "wood an' water an' grass an' a side-hill! A pocket hunter's delight . . . a secret pasture for prospectors and a resting place for tired burros, by damn!"

London's prospector begins with a shovelful of dirt from the edge of the stream below the side hill. He pours it into his gold pan. He partially immerses the pan in the stream and with a circular motion sluices out most of the dirt until only fine dirt and the smallest bits of gravel remain. Now comes the slow and deliberate work, the prospector washing more and more delicately until the pan seems empty of all but water. But with a quick semicircular motion that sends water flying over the rim into the stream, he reveals a thin layer of black sand on the bottom of the pan. A close look discloses a tiny gold speck. He drains more water over the black grains. A second speck of gold appears.

He pursues the painstaking process, working a small portion of the black sand at a time up the shallow rim of the pan. His efforts yield a count of seven gold specks. Not enough to keep, but enough to charge up his hopes. He continues down the stream repeating the same tedious procedure—a pan of gravel, the careful washing, the meticulous teasing out of tiny specks of gold. As he works his way downstream, his "golden herds" diminish. A pan yields one speck, another none. So he returns to where he began and starts panning upstream. His tally of gold specks mounts to thirty, then pan by pan dwindles to nothing. He has homed in on the richest point in the stream, but still nothing worth keeping. The real treasure lies above, somewhere on the face of the side hill.

A few feet up from his first line of test pans he begins digging a second row of holes, crosscutting the hillside. Fill the pan, carry it to the stream, pan out the gravel, count the flecks—each tedious cycle gathers more information. He works his way up the side hill in rows of holes. The center of each row yields the richest pans, and each row ends where no gold specks appear. The rows grow shorter as he mounts the hill, forming an inverted V. The converging sides of the V mark the boundaries of the gold-bearing dirt.

The apex of the inverted V is the prospector's goal, where "Mr. Pocket" resides. As the prospector mounts the hill, the pans get rich enough for their yield to be worth saving. But the work grows harder. As the sides of the V converge, the gold retreats underground. The gold at the edge of the stream was right at the roots of the grasses. Then it lies 30 inches down, then 35. Then 4 feet, then 5.

Finally the sides of the V come together at one point. He digs his way 6 feet down into the earth. His pick grates on rotten quartz. He digs the pick in deeper, fracturing the rock with every stroke. He holds a fragment of the rotten quartz in hand and rubs away the dirt. Half the rock is virgin gold. More scrabbling about yields nuggets of pure gold. Eventually the prospector draws 400 pounds of gold from the find.

Can London's tale of the real Klondike have meaning in the esoteric conceptual world of possibility spaces and smart search? Indeed it can. London's prospector carries out a smart homing search for gold, progressing systematically up from the creek bed toward the source. Each hole he digs probes a possibility, yielding some gold and helping him to focus his search better. In Klondike terms, London's prospector searches a homing space within the larger Klondike wilderness.

SCIENCE

Breakthrough thinking comes as a sudden, seemingly un
countable moment of inspiration: From Archimedes' disc
ery in the bathtub of the principle of water displacement
Einstein's Theory of Relativity, from Brunelleschi's develo
ment of perspective drawing to the Impressionist revoluti
from the taming of fire to the creation of the laser, it
shaped and advanced civilization.

David Perkins explores the common logic behind breakthroughs in ma
fields, historical periods, and evolutionary epochs. Drawing on a rich know
edge of both artificial intelligence and cognitive psychology, he sets forth
uniquely integrative theory of how insights occur. Along the way he off
dozens of often playful puzzles and illustrations that reveal the four k
processes behind breakthrough thinking.

> *"[An] absorbing introduction to cognitive theory."*
> —PUBLISHERS WEEKLY

David Perkins is a founding member of the think tank "Proje
Zero" at the Harvard Graduate School of Education and has published boo
on mind, intelligence, creativity, and learning.

ORIGINALLY PUBLISHED UNDER THE TITLE
Archimedes' Bathtub

COVER ART & DESIGN: MARK MELNICK

W. W. NORTON
NEW YORK · LONDON

WWW.WWNORTON.CO

ISBN 0-393-32255-6

51495

9 780393 322552

$14.95 USA · $21.99 CAN

How to Solve It

a new aspect of mathematical method

With a new foreword by John H. Conway

G. POLYA

HOW TO SOLVE IT

UNDERSTANDING THE PROBLEM

First.

You have to *understand* the problem.

What is the unknown? What are the data? What is the condition?

Is it possible to satisfy the condition? Is the condition sufficient to determine the unknown? Or is it insufficient? Or redundant? Or contradictory?

Draw a figure. Introduce suitable notation.

Separate the various parts of the condition. Can you write them down?

DEVISING A PLAN

Second.

Find the connection between the data and the unknown. You may be obliged to consider auxiliary problems if an immediate connection cannot be found. You should obtain eventually a *plan* of the solution.

Have you seen it before? Or have you seen the same problem in a slightly different form?

Do you know a related problem? Do you know a theorem that could be useful?

Look at the unknown! And try to think of a familiar problem having the same or a similar unknown.

Here is a problem related to yours and solved before. Could you use it? Could you use its result? Could you use its method? Should you introduce some auxiliary element in order to make its use possible?

Could you restate the problem? Could you restate it still differently? Go back to definitions.

If you cannot solve the proposed problem try to solve first some related problem. Could you imagine a more accessible related problem? A more general problem? A more special problem? An analogous problem? Could you solve a part of the problem? Keep only a part of the condition, drop the other part; how far is the unknown then determined, how can it vary? Could you derive something useful from the data? Could you think of other data appropriate to determine the unknown? Could you change the unknown or the data, or both if necessary, so that the new unknown and the new data are nearer to each other?

Did you use all the data? Did you use the whole condition? Have you taken into account all essential notions involved in the problem?

CARRYING OUT THE PLAN

Third.

Carry out your plan.

Carrying out your plan of the solution, *check each step.* Can you see clearly that the step is correct? Can you prove that it is correct?

LOOKING BACK

Fourth.

Examine the solution obtained.

Can you *check the result?* Can you check the argument?

Can you derive the result differently? Can you see it at a glance?

Can you use the result, or the method, for some other problem?

How to Solve It

A New Aspect of Mathematical Method

G. Polya

With a new foreword by John H. Conway

A perennial bestseller by eminent mathematician G. Polya, *How to Solve It* will show anyone in any field how to think straight.

In lucid and appealing prose, Polya reveals how the mathematical method of demonstrating a proof or finding an unknown can be of help in attacking any problem that can be "reasoned" out—from building a bridge to winning a game of anagrams. Generations of readers have relished Polya's deft—indeed, brilliant—instructions on stripping away irrelevancies and going straight to the heart of the problem.

From reviews of the original edition:

"Every prospective teacher should read it. In particular, graduate students will find it invaluable. The traditional mathematics professor who reads a paper before one of the Mathematical Societies might also learn something from the book: 'He writes *a*, he says *b*, he means *c*; but it should be *d*.'"
—E. T. Bell, *Mathematical Monthly*, December 1945

"[This] elementary textbook on heuristic reasoning, shows anew how keen its author is on questions of method and the formulation of methodological principles. Exposition and illustrative material are of a disarmingly elementary character, but very carefully thought out and selected."
—Herman Weyl, *Mathematical Review*, October 1948

"Any young person seeking a career in the sciences would do well to ponder this important contribution to the teacher's art."
—A. C. Schaeffer, *American Journal of Psychology*, April 1946

GEORGE POLYA (1887–1985) was one of the most influential mathematicians of the twentieth century. His basic research contributions span complex analysis, mathematical physics, probability theory, geometry, and combinatorics. He was a teacher par excellence who maintained a strong interest in pedagogical matters throughout his long career. Even after his retirement from Stanford University in 1953, he continued to lead an active mathematical life. He taught his final course, on combinatorics, at the age of ninety. JOHN H. CONWAY is the John von Neumann Distinguished Professor of Mathematics at Princeton University. He was awarded the London Mathematical Society's Polya Prize in 1987. Like Polya, he is interested in many branches of mathematics, and in particular, has invented a successor to Polya's notation for crystallographic groups.

PRINCETON SCIENCE LIBRARY
psl.princeton.edu

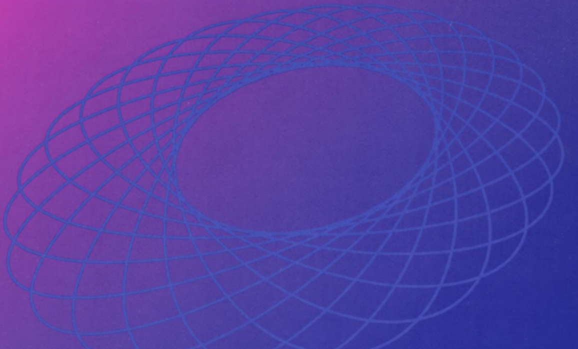
From its publication of Albert Einstein's *The Meaning of Relativity* in 1922, Princeton University Press has been committed to bringing the writings of leading scientists to a broad professional and general audience. The Princeton Science Library now makes the best of these works available in inexpensive paperback editions for a new generation of scientists and teachers, students and nonspecialists.

Cover design by Carmina Alvarez-Gaffin

ISBN 0-691-11966-X
90000
9 780691 119663

PROBLEM-SOLVING STRATEGIES FOR EFFICIENT AND *Elegant* SOLUTIONS

A Resource for the Mathematics Teacher



Alfred S. Posamentier
Stephen Krulik

Foreword by Nobel Laureate Herbert A. Hauptman

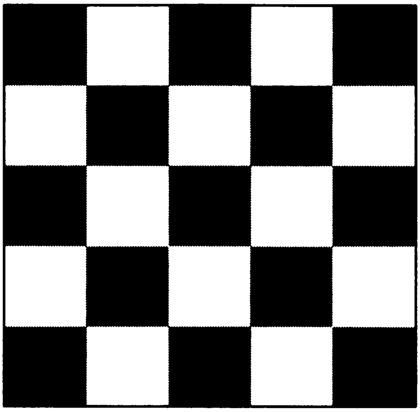


Figure 7.17

Now, to change seats as Mr. Strauss instructed, each student must move from a black square (or "seat") into a white square. Because there are 13 black squares and only 12 white squares, the students cannot follow Mr. Strauss's directions.

Problem 7.11

A jeweler makes silver earrings from silver blanks. Each blank makes 1 earring. The shavings left over from 6 blanks are then melted down and recast to form another blank. The jeweler orders 36 blanks to fill an order. How many earrings can be made from the 36 blanks?

Solution. Students usually assume that 36 blanks will yield 36 earrings. They are quite surprised when they find that this is not the correct answer. Some astute students may recognize that the leftover silver can form 6 additional blanks, thus yielding 42 earrings. This, too, is incorrect.

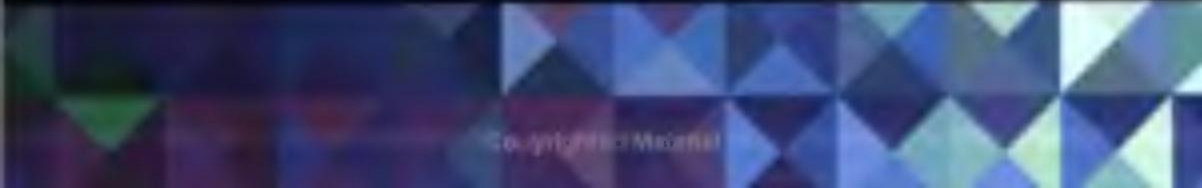
Let's use the **visual representation (make a drawing)** strategy to see what happens as the jeweler works. From the original 36 blanks, we do obtain 36 earrings. However, notice that the shavings left over from these 36 blanks are melted down and form 6 new blanks, yielding 6 additional earrings. However, we don't stop here—the shavings from these 6 blanks are then melted down and recast to form 1 new blank from which we obtain 1 additional earring. Thus, Figure 7.18, 43 earrings are possible.

OXFORD
MATHEMATICS

solving mathematical problems

a personal perspective

TERENCE TAO



Preface to the first edition

Proclus, an ancient Greek philosopher, said:

> This therefore, is mathematics: she reminds you of the invisible forms of the soul; she gives life to her own discoveries; she awakens the mind and purifies the intellect; she brings to light our intrinsic ideas; she abolishes oblivion and ignorance which are ours by birth ...

But I just like mathematics because it is fun.

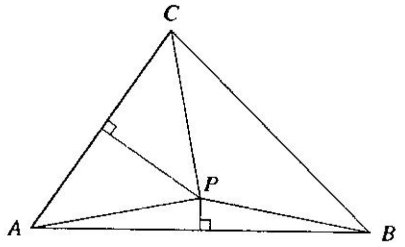
Mathematical problems, or puzzles, are important to real mathematics (like solving real-life problems), just as fables, stories, and anecdotes are important to the young in understanding real life. Mathematical problems are 'sanitized' mathematics, where an elegant solution has already been found (by someone else, of course), the question is stripped of all superfluousness and posed in an interesting and (hopefully) thought-provoking way. If mathematics is likened to prospecting for gold, solving a good mathematical problem is akin to a 'hide-and-seek' course in gold-prospecting: you are given a nugget to find, and you know what it looks like, that it is out there somewhere, that it is not too hard to reach, that it is unearthing within your capabilities, and you have conveniently been given the right equipment (i.e. data) to get it. It may be hidden in a cunning place, but it will require ingenuity rather than digging to reach it.

In this book I shall solve selected problems from various levels and branches of mathematics. Starred problems (*) indicate an additional level of difficulty, either because some higher mathematics or some clever thinking are required; double-starred questions (**) are similar, but to a greater degree. Some problems have additional exercises at the end that can be solved in a similar manner or involve a similar piece of mathematics. While solving these problems, I will try to demonstrate some tricks of the trade when problem-solving. Two of the main weapons—experience and knowledge—are not easy to put into a book: they have to be acquired over time. But there are many simpler tricks that take less time to learn. There are ways of looking at a problem that make it easier to find a feasible attack plan. There are systematic ways of reducing a problem into successively simpler sub-problems. But, on the other hand, solving the problem is not everything. To return to the gold nugget analogy, strip-mining the neighbourhood with bulldozers is clumsier than doing a careful survey, a bit of geology, and a small amount of digging. A solution should be relatively short, understandable, and hopefully have a touch of elegance. It should also be fun to discover. Transforming a nice, short little geometry question into a ravening monster of an equation by textbook coordinate geometry does not have the same taste of victory as a two-line vector solution.

As an example of elegance, here is a standard result in Euclidean geometry:

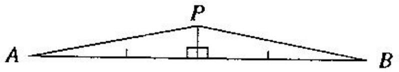
Show that the perpendicular bisectors of a triangle are concurrent.

This neat little one-liner *could* be attacked by coordinate geometry. Try to do so for a few minutes (hours?), then look at this solution:



Proof. Call the triangle ABC. Now let P be the intersection of the perpendicular bisectors of AB and AC. Because P is on the AB bisector, $|AP| = |PB|$. Because P is on the AC bisector, $|AP| = |PC|$. Combining the two, $|BP| = |PC|$. But this means that P has to be on the BC bisector. Hence all three bisectors are concurrent. (Incidentally, P is the circumcentre of ABC.) □

The following reduced diagram shows why $|AP| = |PB|$ if P is on the AB perpendicular bisector: congruent triangles will pull it off nicely.



This kind of solution—and the strange way that obvious facts mesh to form a not-so-obvious fact—is part of the beauty of mathematics. I hope that you too will appreciate this beauty.

Acknowledgements

Thanks to Peter O'Halloran, Vern Treilibs, and Lenny Ng for their contributions of problems and advice.

Special thanks to Basil Rennie for his corrections and ingenious shortcuts in solutions, and finally thanks to my family for their support, encouragement, spelling corrections, and put-downs when I was behind schedule.

Almost all of the problems in this book come from published collections of problem sets for mathematics competitions. These are sourced in the texts, with full details given in the reference section of the book. I also used a small handful of problems from friends or from various mathematical publications; these have no source listed.

Preface to the second edition

This book was written 15 years ago; literally half a lifetime ago, for me. In the intervening years, I have left home, moved to a different country, gone to graduate school, taught classes, written research papers, advised graduate students, married my wife, and had a son. Clearly, my perspective on life and on mathematics is different now than it was when I was 15. I have not been involved in problem-solving competitions for a very long time now, and if I were to write a book now on the subject it would be very different from the one you are reading here.

Mathematics is a multifaceted subject, and our experience and appreciation of it changes with time and experience. As a primary school student, I was drawn to mathematics by the abstract beauty of formal manipulation, and the remarkable ability to repeatedly use simple rules to achieve non-trivial answers. As a high-school student, competing in mathematics competitions, I enjoyed mathematics as a sport, taking cleverly designed mathematical puzzle problems (such as those in this book) and searching for the right 'trick' that would unlock each one. As an undergraduate, I was awed by my first glimpses of the rich, deep, and fascinating theories and structures which lie at the core of modern mathematics today. As a graduate student, I learnt the pride of having one's own research project, and the unique satisfaction that comes from creating an original argument that resolved a previously open question. Upon starting my career as a professional research mathematician, I began to see the intuition and motivation that lay behind the theories and problems of modern mathematics, and was delighted when realizing how even very complex and deep results are often at heart be guided by very simple, even common-sensical, principles. The 'Aha!' experience of grasping one of these principles, and suddenly seeing how it illuminates and informs a large body of mathematics, is a truly remarkable one. And there are yet more aspects of mathematics to discover; it is only recently for me that I have grasped enough fields of mathematics to begin to get a sense of the endeavour of modern mathematics as a unified subject, and how it connects to the sciences and other disciplines.

As I wrote this book before my professional mathematics career, many of these insights and experiences were not available to me, and so in many places the exposition has a certain innocence, or even naivety. I have been reluctant to tamper too much with this, as my younger self was almost certainly more attuned to the world of the high-school problem solver than I am now. However, I have made a number of organizational changes: formatting the text into LaTeX, arranging the material into what I believe is a more logical order, and editing those parts of the text which were inaccurate, badly worded, confusing, or unfocused. I have also added some more exercises. In some places, the text is a bit dated (Fermat's last theorem, for instance, has now been proved rigorously), and I now realize that several of the problems here could be handled more quickly and cleanly by more 'high-tech' mathematical tools; but the point of this text is not to present the slickest solution to a problem or to provide the most up-to-date survey of results, but rather to show how one approaches a mathematical problem for the first time, and how the painstaking, systematic experience of trying some ideas, eliminating others, and steadily manipulating the problem can lead, ultimately, to a satisfying solution.

I am greatly indebted to Tony Gardiner for encouraging and supporting the reprinting of this book, and to my parents for all their support over the years. I am also touched by all the friends and acquaintances I have met over the years who had read the first edition of the book. Last, but not least, I owe a special debt to my parents and the Flinders Medical Centre computer support unit for retrieving a 15-year old electronic copy of this book from our venerable Macintosh Plus computer!

Terence Tao
Department of Mathematics,
University of California, Los Angeles
December 2005

1 Strategies in problem solving

> The journey of a thousand miles begins with one step.
> Lao Tzu

Like and unlike the proverb above, the solution to a problem begins (and continues, and ends) with simple, logical steps. But as long as one steps in a firm, clear direction, with long strides and sharp vision, one would need far, far less than the millions of steps needed to journey a thousand miles. And mathematics, being abstract, has no physical constraints; one can always restart from scratch, try new avenues of attack, or backtrack at an instant's notice. One does not always have these luxuries in other forms of problem-solving (e.g. trying to go home if you are lost).

Of course, this does not necessarily make it easy; if it was easy, then this book would be substantially shorter. But it makes it possible.

There are several general strategies and perspectives to solve a problem correctly; (Polya 1957) is a classic reference for many of these. Some of these strategies are discussed below, together with a brief illustration of how each strategy can be used on the following problem:

> PROBLEM 1.1. A triangle has its lengths in an arithmetic progression, with difference d. The area of the triangle is t. Find the lengths and angles of the triangle.

Understand the problem. What kind of problem is it? There are three main types of problems:

- 'Show that ...' or 'Evaluate ...' questions, in which a certain statement has to be proved true, or a certain expression has to be worked out;
- 'Find a ...' or 'Find all ...' questions, which requires one to find something (or everything) that satisfies certain requirements;
- 'Is there a ...' questions, which either require you to prove a statement or provide a counterexample (and thus is one of the previous two types of problem).

The type of problem is important because it determines the basic method of approach. 'Show that ...' or 'Evaluate ...' problems start with given data and the objective is to deduce some statement or find the value of an expression; this type of problem is generally easier than the other two types because there is a clearly visible objective, one that can be deliberately approached. 'Find a . . .' questions are more hit-and-miss; generally one has to guess one answer that nearly works, and then tweak it a bit to make it more correct; or alternatively one can alter the requirements that the object-to-find must satisfy, so that they are easier to satisfy. 'Is there a . . .' problems are typically the hardest, because one must first make a decision on whether an object exists or not, and provide a proof on one hand, or a counter-example on the other.

Of course, not all questions fall into these neat categories; but the general format of any question will still indicate the basic strategy to pursue when solving a problem. For example, if one tries to solve the problem 'find a hotel in this city to sleep in for the night', one should alter the requirements to, say 'find a vacant hotel within 5 kilometres with a room that costs less than 100$ a night' and then use pure elimination. This is a better strategy than proving that such a hotel does or does not exist, and is probably a better strategy than picking any handy hotel and trying to prove that one can sleep in it.

In Problem 1.1 question, we have an 'Evaluate . . .' type of problem. We need to find several unknowns, given other variables. This suggests an algebraic solution rather than a geometric one, with a lot of equations connecting d, t, and the sides and angles of the triangle, and eventually solving for our unknowns.

Understand the data. What is given in the problem? Usually, a question talks about a number of objects which satisfy some special requirements. To understand the data, one needs to see how the objects and requirements react to each other. This is important in focusing attention on the proper techniques and notation to handle the problem. For example, in our sample question, our data are a triangle, the area of the triangle, and the fact that the sides are in an arithmetic progression with separation d. Because we have a triangle, and are considering the sides and area of it, we would need theorems relating sides, angles, and areas to tackle the question: the sine rule, cosine rule, and the area formulas, for example. Also, we are dealing with an arithmetic progression, so we would need some notation to account for that; for example, the side lengths could be a, $a+d$, and $a+2d$.

Understand the objective. What do we want? One may need to find an object, prove a statement, determine the existence of an object with special properties, or whatever. Like the flip side of this strategy, 'understand the data', knowing the objective helps focus attention on the best weapons to use. Knowing the objective also helps in creating tactical goals which we know will bring us closer to solving the question. Our example question has the objective of 'find all the sides and angles of the triangle'. This means, as mentioned before, that we will need theorems and results concerning sides and angles. It also gives us the tactical goal of 'find equations involving the sides and angles of the triangle'.

Select good notation. Now that we have our data and objective, we must represent it in an efficient way, so that the data and objective are both represented as simply as possible. This usually involves the thoughts of the past two strategies. In our sample question, we are already thinking of equations involving d, t, and the sides and angles of the triangle. We need to express the sides and angles in terms of variables: one could choose the sides to be a, b, and c, while the angles could be denoted α, β, γ. But we can use the data to simplify the notation: we know that the sides are in arithmetic progression, so instead of a, b, and c, we can have a, $a+d$, and $a+2d$ instead. But the notation can be even better if we make it more symmetrical, by making the side lengths $b-d$, b, and $b+d$. The only slight drawback to this notation is that b is forced to be larger than d. But on further thought, we see that this is actually not a restriction; in fact the knowledge that $b > d$ is an extra piece of data for us. We can also trim the notation more, by labelling the angles α, β, and $180^\circ - \alpha - \beta$, but this is ugly and unsymmetrical—it is probably better to keep the old notation, but bearing in mind that $\alpha + \beta + \gamma = 180^\circ$.

Write down what you know in the notation selected; draw a diagram. Putting everything down on paper helps in three ways:

(a) you have an easy reference later on;

(b) the paper is a good thing to stare at when you are stuck;

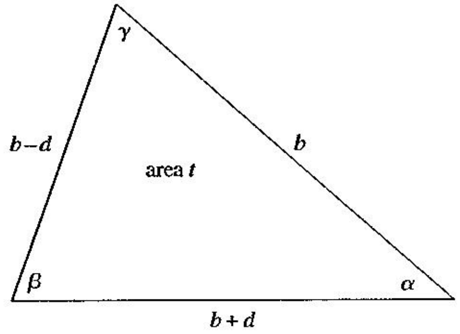
(c) the physical act of writing down of what you know can trigger new inspirations and connections.

Be careful, though, of writing superfluous material, and do not overload your paper with minutiae; one compromise is to highlight those facts which you think will be most useful, and put more questionable, redundant, or crazy ideas in another part of your scratch paper. Here are some equations and inequalities one can extract from our example question:

- (physical constraints) $\alpha, \beta, \gamma, t > 0$, and $b \geq d$; we can also assume $d \geq 0$ without loss of generality;
- (sum of angles in a triangle) $\alpha + \beta + \gamma = 180^\circ$;
- (sine rule) $(b-d)/\sin\alpha = b/\sin\beta = (b+d)/\sin\gamma$;
- (cosine rule) $b^2 = (b-d)^2 + (b+d)^2 - 2(b-d)(b+d)\cos\beta$, etc.;
- (area formula) $t = (1/2)(b-d)b\sin\gamma = (1/2)(b-d)(b+d)\sin\beta = (1/2)b(b+d)\sin\alpha$;
- (Heron's formula) $t^2 = s(s-b+d)(s-b)(s-b-d)$, where $s = ((b-d)+b+(b+d))/2$ is the semiperimeter;
- (triangle inequality) $b+d \leq b+(b-d)$.

Many of these facts may prove to be useless or distracting. But we can use some judgement to separate the valuable facts from the unhelpful ones. The equalities are likely to be more useful than the inequalities, since our objective and data come in the form of equalities. And Heron's formula looks especially promising, because the semiperimeter simplifies to $s = 3b/2$. So we can highlight 'Heron's formula' as being likely to be useful.

We can of course also draw a picture. This is often quite helpful for geometry questions, though in this case the picture does not seem to add much:



Modify the problem slightly. There are many ways to vary a problem into one which may be easier to deal with:

(a) Consider a special case of the problem, such as extreme or degenerate cases.

(b) Solve a simplified version of the problem.

(c) Formulate a conjecture which would imply the problem, and try to prove that first.

(d) Derive some consequence of the problem, and try to prove that first.

(e) Reformulate the problem (e.g. take the contrapositive, prove by contradiction, or try some substitution).

(f) Examine solutions of similar problems.

(g) Generalize the problem.

This is useful when you cannot even get started on a problem, because solving for a simpler related problem sometimes reveals the way to go on the main problem. Similarly, considering extreme cases and solving the problem with additional assumptions can also shed light on the general solution. But be warned that special cases are, by their nature, special, and some elegant technique could conceivably apply to them and yet have absolutely no utility in solving the general case. This tends to happen when the special case is *too* special. Start with modest assumptions first, because then you are sticking as closely as possible to the spirit of the problem.

In Problem **1.1**, we can try a special case such as $d = 0$. In this case we need to find the side length of an equilateral triangle of area t. In this case, it is a standard matter to compute the answer, which is $b = 2t^{1/2}/3^{1/4}$. This indicates that the general answer should also involve square roots or fourth roots, but does not otherwise suggest how to go about the problem. Consideration of similar problems draws little as well, except one gets further evidence that a gung-ho algebraic attack is what is needed.

Modify the problem significantly. In this more aggressive type of strategy, we perform major modifications to a problem such as removing data, swapping the data with the objective, or negating the objective (e.g. trying to disprove a statement rather than prove it). Basically, we try to push the problem until it breaks, and then try to identify where the breakdown occurred; this identifies what the key components of the data are, as well as where the main difficulty will lie. These exercises can also help in getting an instinctive feel of what strategies are likely to work, and which ones are likely to fail.

In regard to our particular question, one could replace the triangle with a quadrilateral, circle, etc. Not much help there: the problem just gets more complicated. But on the other hand, one can see that one does not really need a triangle in the question, but just the dimensions of the triangle. We do not really need to know the position of the triangle. So here is further confirmation that we should concentrate on the sides and angles (i.e. $a, b, c, \alpha, \beta, \gamma$) and not on coordinate geometry or similar approaches.

We could omit some objectives; for example, instead of working out all the sides and angles we could work out just the sides, for example. But then one can notice that by the cosine and sine rules, the angles of the triangle will be determined anyway. So it is only neccesary to solve for the sides. But we know that the sides have lengths $b - d$, b, and $b + d$, so we only need to find what b is to finish the problem.

We can also omit some data, like the arithmetic difference d, but then we seem to have several possible solutions, and not enough data to solve the problem. Similarly, omitting the area t will not leave enough data to clinch a solution. (Sometimes one can *partially* omit data, for instance, by only specifying that the area is larger or smaller than some threshold t_0; but this is getting complicated. Stick with the simple options first.) Reversal of the problem (swapping data with objective) leads to some interesting ideas though. Suppose you had a triangle with arithmetic difference d, and you wanted to shrink it (or whatever) until the area becomes t. One could imagine our triangle shrinking and deforming, while preserving the arithmetic difference of the sides. Similarly, one could consider all triangles with a fixed area, and mold the triangle into one with the sides in the correct arithmetic progression. These ideas could work in the long run: but I will solve this question by another approach. Do not forget, though, that a question can be solved in more than one way, and no particular way can really be judged the absolute best.

Prove results about our question. Data is there to be used, so one should pick up the data and play with it. Can it produce more meaningful data? Also, proving small results could be beneficial later on, when trying to prove the main result or to find the answer. However small the result, do not forget it—it could have bearing later on. Besides, it gives you something to do if you are stuck.

In a 'Evaluate . . .' problem like the triangle question, this tactic is not as useful. But one can try. For example, our tactical goal is to solve for b. This depends on the two parameters d and t. In other words, b is really a function: $b = b(d, t)$. (If this notation looks out of place in a geometry question, then that is only because geometry tends to ignore the functional dependence of objects. For example, Heron's formula gives an explicit form for the area A in terms of the sides a, b, and c: in other words, it expresses the function $A(a, b, c)$.) Now we can prove some mini-results about this function $b(d, t)$, such as $b(d, t) = b(-d, t)$ (because an arithmetic progression has an equivalent arithmetic progression with inverted arithmetic difference), or $b(kd, k^2t) = kb(d, t)$ (this is done by dilating the triangle that satisfies $b = b(d, t)$ by k). We could even try differentiate b with respect to d or t. For this particular problem, these tactics allow us to perform some normalizations, for instance setting $t = 1$ or $d = 1$, and also provide a way to check the final answer. However, in this problem these tricks turn out to only give minor advantages and we will not use them here.

Simplify, exploit data, and reach tactical goals. Now we have set up notation and have a few equations, we should seriously look at attaining our tactical goals that we have established. In simple problems, there are usually standard ways of doing this: (For example, algebraic simplification is usually discussed thoroughly in high-school level textbooks.) Generally, this part is the longest and most difficult part of the problem: however, once can avoid getting lost if one remembers the relevant theorems, the data and how they can be used, and most importantly the objective. It is also a good idea to not apply any given technique or method blindly, but to think ahead and see where one could hope such a technique to take one; this can allow one to save enormous amounts of time by eliminating unprofitable directions of inquiry before sinking lots of effort into them, and conversely to give the most promising directions priority.

In Problem 1.1, we are already concentrating on Heron's formula. We can use this to attain our tactical goal of solving for b. After all, we have already noted that the sine and cosine rules can determine α, β, γ once b is known. As further evidence that this is going to be a step forward, note that Herons formula involves d and t—in essence, it uses all our data (we have already incorporated the fact about the sides being in arithmetic progression into our notation). Anyway, Herons formula in terms of d, t, b becomes

$$t^2 = \frac{3b}{2}\left(\frac{3b}{2} - b + d\right)\left(\frac{3b}{2} - b\right)\left(\frac{3b}{2} - b - d\right)$$

which we can simplify to

$$t^2 = \frac{3b^2(b - 2d)(b + 2d)}{16} = \frac{3b^2(b^2 - 4d^2)}{16}.$$

Now we have to solve for b. The right-hand side is a polynomial in b (treating d and t as constants), and in fact it is a quadratic in b^2. Now quadratics can be solved easily: if we put clear denominators and put everything on the left-hand side we get

$$3b^4 - 12d^2b^2 - 16t^2 = 0$$

so, using the quadratic formula,

$$b^2 = \frac{12d^2 \pm \sqrt{144d^4 + 196t^2}}{6} = 2d^2 \pm \sqrt{4d^2 + \frac{16}{3}t^2}.$$

Because b has to be positive, we get

$$b = \sqrt{2d^2 + \sqrt{4d^4 + \frac{16}{3}t^2}},$$

as a check, we can verify that when $d = 0$ this agrees with our previous computation of $b = 2t^{1/2}/3^{1/4}$. Once we ompute the sides $b - d, b, b + d$, the evaluation of the angles α, β, γ then follows from the cosine laws, and we are done!

Authored by a leading name in mathematics, this engaging and clearly presented text leads the reader through the various tactics involved in solving mathematical problems at the Mathematical Olympiad level. Covering number theory, algebra, analysis, Euclidean geometry, and analytic geometry, *Solving Mathematical Problems* includes numerous exercises and model solutions throughout. Assuming only basic high-school mathematics, the text is ideal for general readers and students of 14 years and above with an interest in pure mathematics.

Terence Tao was born in Adelaide, Australia, in 1975. In 1987, 1988, and 1989 he competed in the International Mathematical Olympiad for the Australian team, winning a bronze, silver, and gold medal respectively, and being the youngest competitor ever to win a gold medal at this event. Since 2000, Terence has been a full professor of mathematics at the University of California, Los Angeles. He now lives in Los Angeles with his wife and son.

CONTENTS

"I must say I find it delightful . . . I have absolutely no doubt that this book will be popular."

Professor W T Gowers, University of Cambridge

Cover illustration: © [illegible]

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 0-19-920560-4

9 780199 205608