Information Science and Statistics

Series Editors
M. Jordan, J. Kleinberg, B. Schölkopf

C.S. Wallace

Statistical and Inductive Inference by Minimum Message Length

With 22 Figures

Springer

C.S. Wallace
(deceased)

Series Editors

Michael Jordan
Department of Computer Science
University of California, Berkeley
Berkeley, CA 94720
USA
jordan@stat.berkeley.edu

Professor Jon Kleinberg
Department of Computer Science
Cornell University
Ithaca, NY 14853
USA

Professor Bernhard Schölkopf
Max Planck Institute for Biological
Cybernetics
Spemannstrasse 38
72076 Tübingen
Germany

Library of Congress Cataloging-in-Publication Data
Wallace, C. S. (Christopher S.), d. 2004.
Statistical and inductive inference by minimum message length / C.S. Wallace.
p. cm. — (Information science and statistics)
Includes bibliographical references and index.
ISBN 0-387-23795-X (alk. paper)
1. Minimum message length (Information theory) 2. Mathematical
statistics—Methodology. 3. Induction (Mathematics) I. Title. II. Series.
QA276.9.W35 2005
519.5—dc22 2004059195

ISBN-10: 0-387-23795-X Printed on acid-free paper.
ISBN-13: 978-0387-23795-4

© 2005 Springer Science+Business Media, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America. (EB)

9 8 7 6 5 4 3 2 1

springeronline.com

Preface

My thanks are due to the many people who have assisted in the work reported here and in the preparation of this book. The work is incomplete and this account of it rougher than it might be. Such virtues as it has owe much to others; the faults are all mine.

My work leading to this book began when David Boulton and I attempted to develop a method for intrinsic classification. Given data on a sample from some population, we aimed to discover whether the population should be considered to be a mixture of different types, classes or species of thing, and, if so, how many classes were present, what each class looked like, and which things in the sample belonged to which class. I saw the problem as one of Bayesian inference, but with prior probability densities replaced by discrete probabilities reflecting the precision to which the data would allow parameters to be estimated. Boulton, however, proposed that a classification of the sample was a way of briefly encoding the data: once each class was described and each thing assigned to a class, the data for a thing would be partially implied by the characteristics of its class, and hence require little further description. After some weeks' arguing our cases, we decided on the maths for each approach, and soon discovered they gave essentially the same results. Without Boulton's insight, we may never have made the connection between inference and brief encoding, which is the heart of this work.

Jon Patrick recognized in the classification work a possible means of analysing the geometry of megalithic stone circles and began a PhD on the problem. As it progressed, it became clear that the message-length tools used in the classification method could be generalized to apply to many model-selection and statistical inference problems, leading to our first attempts to formalize the "Minimum Message Length" method. However, these attempts seemed to be incomprehensible or repugnant to the referees of statistical journals. Fortunately, Peter Freeman, a proper statistician who had looked at the stone circle problem, saw some virtue in the approach and very kindly spent a year's sabbatical helping to frame the idea in acceptable statistical terms, leading to the first publication of MML in a statistical journal [55]. Acceptance was probably assisted by the simultaneous publication of the independent but related work of Jorma Rissanen [35].

Over the 35-year gestation of this book, I have benefited greatly from the suggestions, comments and criticisms of many colleagues and anonymous referees. The list includes Mike Georgeff, Peter Cheeseman, Ray Solomonoff, Phil Dawid, David Hand, Paul Vitanyi, Alex Gammerman, Ross Quinlan, Peter Tischer, Lloyd Allison, Trevor Dix, Kevin Korb, Murray Jorgenson, Mike Dale, Charles Twardy, Jon Oliver, Rohan Baxter and especially David Dowe, who has contributed significantly both to the range of applications of MML and to the development of new approximations for message lengths and MML estimators.

I must also thank Julie Austin, who typed and proofread the early chapters, and Steve Gardner and Torsten Seeman, who helped convert the original draft into LaTeX.

Finally, without the constant support of my wife Judith, I would never have managed to complete the work.

Victoria, Australia, August 2004 C.S. Wallace

Disclaimer

The reader should be warned that I make no claim to be an authority on statistical inference, information theory, inductive reasoning or the philosophy of science. I have not read widely in any of these fields, so my discussions of others' work should be treated with some suspicion. The ideas in this book are those of a one-time physicist who drifted into computing via work on computer hardware and arithmetic. In this uncertain progress towards enlightenment, I encountered a succession of problems in analysing and understanding data for which I could find no very satisfactory solution in standard texts. Over the years, the MML approach was developed from rather ad hoc beginnings, but the development was driven mostly by the challenge of new problems and informal argument with colleagues, rather than by a proper study of existing work. This casual, indeed almost accidental, evolution partly excuses my paucity of citations.

Editorial Notes

This book is essentially the manuscript left behind by Christopher Wallace when he died on August 7, 2004.

We wanted to publish a book that was as close as possible to the original manuscript. We have therefore made only minimal changes to the manuscript. We have corrected typing and spelling errors. We have also attempted as best as we could to include all the references that the author intended to include. Where the author made it clear that he wanted to add citations, but did not indicate to what they referred, we have included our best guesses of what these references might be.

Colleagues in the School of Computer Science and Software Engineering at Monash University are in the process of developing a web site that will provide additional material and references on Minimum Message Length to assist the readers of this book.

We acknowledge the contributions of Craig Callender, H. Dieter Zeh, Douglas Kutach and Huw Price for their helpful comments on Chapter 8, the editorial assistance of Sarah George and Yuval Marom in preparing this chapter, and that of Jeanette Niehus in preparing the index. We are also indebted to all the people who helped produce this manuscript after Chris' death. In particular, we thank the following people who assisted us in proofreading the final version of this book: Lloyd Allison, David Dowe, Graham Farr, Steven Gardner, Les Goldschlager, Kevin Korb, Peter Tischer, Charles Twardy and Judy Wallace.

Victoria, Australia, February 2005 D.W. Albrecht and I. Zukerman

Acknowledgment

Judith, Christopher Wallace's wife and executor, gratefully acknowledges the tireless efforts of his colleagues at Computer Science and Software Engineering, Monash University in the final preparation of his manuscript for publication. Thanks are due to Professor Gopal Gupta for his advice and support, and in particular to Dr Albrecht and Associate Professor Zukerman for the very many hours spent editing the final text.

Contents

1. Inductive Inference

1.1 Introduction

The best explanation of the facts is the shortest.

This is scarcely a new idea. In various forms, it has been proposed for centuries, at least from Occam's Razor on. Like many aphorisms, it seems to express a notion which is generally accepted to be more or less true, but so vague and imprecise, so subject to qualifications and exceptions, as to be useless as a rule in serious scientific enquiry. But, beginning around 1965, a small number of workers in different parts of the world and in different disciplines began to examine the consequences of taking the statement seriously and giving it a precise, quantitative meaning. The results have been surprising. At least three related but distinct lines of work have been developed, with somewhat different aims and techniques. This book concentrates on just the one line in which the author has worked, but two other important lines are briefly surveyed in Chapter 10. The major claims of this line refer to several fields.

- Bayesian Inference: The new method unifies *model selection* and *estimation*, usually treated as separate exercises. In many cases, the results obtained by treating both questions at once are superior to previous methods. While closely related to existing Bayesian statistical theory, it provides a sound basis for point estimation of parameters which, unlike "MAP" and "mean of posterior" estimates, do not depend on how the assumed data distribution is parameterized.
- Best Explanation of the Data: For the engineer, scientist or clinician who needs to work with a single, well-defined "best guess" hypothesis, the new result is more useable than methods which provide only "confidence intervals" or posterior densities over a sometimes complex hypothesis space.
- Induction: The work gives a new insight into the nature of inductive reasoning, i.e., reasoning from a body of specific facts and observations to general theories. Hitherto, there has been no accepted logical basis for inductive reasoning despite its great importance.
- Philosophy of Science: The discovery, refinement (and sometimes wholesale replacement) of scientific theories is essentially inductive, and the philosophy of science has been hampered by a lack of a logic for induction. The

new insight is at least a step towards a theory of scientific enquiry which is both normative and descriptive.
- Machine Learning: As a branch of Artificial Intelligence research, machine learning is an attempt to automate the discovery of patterns in data, which amounts to the formation of a theory about the data. One result of the new work has been a sound criterion for assessing what has been "learnt", leading to successful new algorithms for machine learning applications.

These claims may seem rather dry, of interest only to the specialist statistician, machine learning expert, or logician. There are much wider implications.

If the basis of the new approach is sound, it seems to lead to a clearer understanding of the role and methods of science and the validity of its claim to be a search for objective truth about the world. It also places scientific enquiry in the same conceptual basket as the development of human language, traditional techniques of navigation, tool-use, agriculture, hunting, animal husbandry, and all the other skills our species has learnt. They all, including science as we practice it, are based on inductive reasoning from real-world observations to general theories about how the world behaves. In the earliest developments of human culture, these "theories" were possibly not consciously formulated: the emergence of vocal signals for danger, food, enemy, friend, come-here, etc. and the earliest skills for finding food, more likely came from many generations of gradual refinement of simple instinctive actions, but in logical terms they are theories indeed: recognition of similarity or more subtle regularity among many things or happenings. In this light, science and engineering are not wholly revolutionary initiatives of recent centuries, but just the gradual systemization of what humans have always done.

Viewing science as common sense used carefully, we find that the new insight gives strong theoretical support for the belief that, given the same physical environment, any sufficiently long-lived, large and motivated community of intelligent beings will eventually come to the same, or at least equivalent, theories about the world, and use more or less equivalent languages to express them. This conclusion is not unqualified: the development of theories about different aspects of the world may well proceed at different rates and not follow the same paths of refinement and replacement in different communities, but, if our account of induction is right, convergence will occur. The idea that there may be different, correct, but incompatible views of reality seems untenable. If two sets of belief are incompatible but equally valid, it can only be that they are equally wrong. Note that we are not asserting that scientific communities will inevitably converge to a finite set of fundamental theories which then express everything which can be learnt. Our account of inductive reasoning admits the possibility that complete and ultimate "theories of everything" may never be reached in a finite time, and perhaps may not even be expressible, as will be explained later.

It has been said that to a man with a hammer, all problems look like nails. Having perhaps acquired a new and shiny hammer in the shape of a theory of induction, we will of course fall to temptation and swing it at a couple of “nails” which it may miss. These attempts have been left to the end of this work, but may have some value.

This book presents the basic theory of the new approach and shows in numerous examples its application to problems in statistical inference and automated inductive inference, which is usually called “learning” in the Artificial Intelligence literature. I emphasize statistical and machine-learning applications because in these limited arenas, the new approach can be applied with sufficient rigour to allow its performance to be properly assessed. In less well-understood and wider arenas, the approach can arguably be shown to have some merit, but the arguments cannot at this stage be made compelling and must involve some arm-waving. By contrast, statistical inference is relatively simple and its language of probability well defined in operational terms, even if it rests on somewhat ambiguous conceptual foundations. If our approach cannot at least handle problems of some difficulty in this relatively simple field, it cannot be credible. We therefore think it important to show that it performs well in this field, and believe the examples given in this book demonstrate that it does. Moreover, at least some of the examples seem to show its performance to better that of previous general principles of statistical inference, and we have so far found no problems where its performance is notably inferior.

The formal arguments in-principle for the approach, as opposed to specific demonstrations of performance on particular problems, are mainly confined to statistical inference, but are extended to a less-restricted formal treatment of inductive inference. The extensions are based on the theory of Universal Turing Machines, which deals with the capabilities of digital computers, and as far as is currently known, also covers the capabilities of all sufficiently general reasoning systems, including human reasoning. The extensions draw on the work of Turing, Solomonoff, Chaitin and others and provide formal arguments for supposing that our approach is applicable to the inductive inference of any theories whose implications are computable. According to some theorists, this range includes all theories which can be explicitly communicated from one person to another and then applied by the recipient, but we will not pursue that argument. We will, however, argue that the approach provides the basis for a partial account of scientific inference which is both normative and descriptive. It says how science ought to choose its theories, and fairly well describes how it actually does, but has little to contribute to an account of what drives science as a social activity, i.e., what determines the direction of social investment in different areas of enquiry.

This first chapter continues with an informal introduction to inductive inference and our approach to it. It outlines the more obvious problems associated with inductive inference and mentions a couple of well-known approaches to these problems. The critique of classical approaches is neither comprehensive nor fair.

To advance any further with the argument, the informal discussion must be followed by a formal treatment which is inevitably quantitative and mathematical. We assume some familiarity with elementary statistics and simple distributions such as the Normal and Binomial forms. Where less familiar statistical models are used, the models will be briefly introduced and described. The mathematics required is restricted to elementary calculus and matrix algebra.

The chapter concludes with a brief introduction to probability and statistics, essentially just to establish the notations and assumptions used later. The nature of the inferences which can be made using conventional non-Bayesian and Bayesian reasoning are outlined, and certain criticisms made. We do not pretend this critique is comprehensive or impartial. It is intended merely to clarify the distinctions between conventional statistical inference and the method developed in this work. Not all of the workers who have contributed to the new approach would necessarily agree with the critique, and the results obtained with the new approach do not depend on its validity. These sections might well be merely skimmed by readers with a statistical background, but the criticisms may be of interest.

The second chapter introduces the elementary results of Information Theory and Turing Machine theory which will be needed in the sequel. These results are needed to define the notion of the "length" of an explanation and to sharpen the concepts of "pattern" and "randomness". There is nothing really novel in this treatment, and it could well be skipped by readers who are familiar with Shannon information and Kolmogorov-Chaitin complexity. However, at the time of writing (2004) it seems some of this material is still not as well understood as one might expect. In recent years several papers have been published on applications of Minimum Message Length or Rissanen's related Minimum Description Length, which have made significant and in some cases serious errors in estimating the length of "explanation" messages. The commonest errors have arisen from a failure to realize that for the statement of a hypothesis about a given body of data, the shortest useable code need not in general allow the encoding of all hypotheses initially contemplated, and should never state a parameter of the asserted hypothesis more precisely than is warranted by the volume and nature of the data. A couple of examples are discussed in later chapters. Other errors have arisen because the authors have applied approximations described in early MML papers in which the limitations of the approximations were not emphasized.

The third, fourth and fifth chapters formally develop the new approach to statistical inference. In Chapter 3, the development is exact, but leads to a method which is computationally infeasible except in the simplest problems. Chapters 4 and 5 introduce approximations to the treatment, leading to useable results. A number of simple inference problems are used in these chapters to illustrate the nature and limitations of the exact and approximate treatments. Chapter 6 looks in more detail at a variety of fairly simple problems, and introduces a couple of techniques needed for some rather more difficult problems. Chapter 7 gives examples of the use of MML in problems where the possible models include models of different order, number of free parameters or logical structure. In these, MML is shown to perform well in selecting a model of appropriate complexity while simultaneously estimating its parameters. Chapter 8 is speculative, presenting an argument that, while deductive (probabilistic) logic is properly applied in predicting the future state of a system whose present state is partly known, useful assertions about the past state of the system require inductive reasoning for which MML appears well suited. Chapter 9 considers whether scientific enquiry can be seen as conforming to the MML principle, at least over the long term. Chapter 10 briefly discusses two bodies of work, Solomonoff's predictive process and Rissanen's Normalized Maximum Likelihood, both of which embody the same "brief encoding" notion as Minimum Message Length but apply it to different ends.

1.2 Inductive Inference

The term "inductive" is sometimes used in the literature to apply to any reasoning other than deductive, i.e., any reasoning where the conclusions are not provably correct given the premises. We will use the term only in the above narrower (and more common) sense. Deductive reasoning, from general theories and axioms to specific conclusions about particular cases, has been studied and systematized since Aristotle and is now fairly well understood, but induction has been much more difficult to master. The new results give an account of inductive reasoning which avoids many of the difficulties in previous accounts, and which has allowed some limited forms of inductive reasoning to be successfully automated.

With the exception of "knowledge" that we are born with or which comes through extraordinary routes such as divine inspiration, our knowledge of the external world is limited to what our senses tell us and to the inferences we may draw from this data. To the extent that a language like English allows, the information gained about our surroundings can be framed as very specific propositions, for instance "I feel warm", "I hear a loud rhythmic sound of varying pitch", "I see blue up high, green-brown lower down, brown at the bottom". Groups of such elementary sensory propositions can be interpreted by most people to arrive at propositions whose terms involve some abstraction from the immediate sense data. "That motor car has rust in its doors", "Joe is lying down", "Percy said 'Crows are black' ". Let us accept that at least this degree of abstraction may be taken for granted, to save us the trouble of having to treat all our information in purely sensory terms. Attempts such as Carnap's [7] to base all reasoning on natural-language sentences relating to uninterpreted sense data have not been fruitful. Then virtually all of our knowledge comes from simple observational propositions like those above, which I will call directly available. Note that even when we are taught by our mothers or learn from books, the propositions directly available to us are not “crossing the road is dangerous” or “France is a major wine producer”, but the observational propositions “I heard Mother say crossing the road is dangerous”, “I read in this book maps and tables implying that France is a major wine producer”.

Each observation tells us something about a specific object or event at a specific time, and when framed as a proposition has no generality. That is, the propositions do not concern classes of things or events, only single isolated things or events. Yet somehow we work from collections of such specific propositions to very general ones, which make assertions about wide classes of objects and events most of which we have never observed, and never will. “Apples and pears both contain malic acid”, “1960 V-8 Roadrats are a bad buy”, “All power corrupts”, “The universe became transparent to electromagnetic radiation one year after the big bang”. The process(es) used to obtain such general propositions from masses of specific ones is called “inductive” reasoning or “induction”.

This inductive process is fundamental to our culture, our technology, and our everyday survival. We are perhaps more used to regarding deductive reasoning, and in particular the formalized quantitative deduction of the sciences, as being the hallmark of rational activity, but deduction must be based on premises. The premises of scientific deductions include many general propositions, the “natural laws” of the physical world. Except for deductions based on hypotheses accepted for the sake of argument, deductive reasoning requires the fruits of induction before it can start, and our deduced conclusions are no better than the inductively derived premises on which they are based. Thus, induction is at least as important a mode of reasoning as deduction.

The priority of induction is even stronger than these arguments have shown. When we look at the specific propositions illustrated above, we find that their very expression relies on previous inductive steps. Before I can frame an observation as an assertion like “This car has rust in its doors”, I and my cultural forebears must somehow form the belief that there is a class of observable phenomena which share so many features correlating usefully with features of other phenomena that the class deserves a name, say, “rust”. This is an inductive conclusion, no doubt based on many thousands of observations of pieces of iron. Inductive conclusions of this type must lie behind the invention of all the common nouns and verbs of natural languages. Indeed, without induction, language could use only proper nouns: the subject of every sentence could be named only as itself, with its own unique grunt. Induction is needed for us to invent and accept the general proposition that all observed phenomena satisfying certain criteria are likely to share certain unobserved but interesting properties, and hence are worth a common name. Thus, we can claim that every generally used common noun and verb in a natural language is the product of inductive inference. We do not claim those inductions were all necessarily sound. The empirical justification for some of them, such as those leading to the nouns “dragon”, “miracle” and the compound “free will”, may be quite unsound.

With every such word are associated two clusters of propositions. The first cluster one might call the defining propositions — those which allow us to recognize an instance of the class named by the word. For example:

> Cows tend to be between 1.5 and 3 m long.
> Cows usually have 4 legs.
> Cows move against the background.
> Cows have a head at one end, often bearing horns, etc., etc.

Then there is a second cluster of propositions which are not needed for recognition, but which allow useful inferences. We may call them “consequential”.

> Cows are warm.
> Cows can (sometimes) be induced to give milk.
> Those that can’t are often dangerous.
> Cows need vegetation to eat.
> etc., etc.

The two clusters often overlap: some propositions may be used as defining in some instances, but treated as consequential if not directly observed. The concept of “cow” is accepted because we can recognize an instance of “cow” using a subset of the propositions, and can then infer the probable truth of the remaining propositions in this instance, even though we have not observed them. The induction embodied in the “cow” concept is thus the general proposition that any phenomenon observed to satisfy a certain defining cluster of propositions will also (usually) satisfy the associated consequential propositions.

Some common nouns result from conscious, systematic, “scientific” reasoning. Terms such as “electron”, “quark”, “cyclonic depression” and “catalyse” label clusters of propositions whose association was unobvious and discovered only after much directed effort. Others, like “man” and “fire” (in whatever language) predate history. We suggest that inductive processes are not necessarily a matter of conscious rational thought, or even of any sort of reasoning.

Any biological organism can be regarded as, at least in a metaphorical sense, embodying such clusters of propositions. Organisms have means, not always neural, of detecting properties of their environment, and many have means to detect with some reliability when several properties are present simultaneously or in a specific sequence. That is, many organisms are equipped to detect when a cluster of propositions is true of their environment. Detection of such a defining cluster instance may trigger behaviour expected to be advantageous to the organism if the environment has other properties not directly detectable by the organism. That is to say, the organism may behave in a way whose benefit depends on certain consequential propositions being true of its environment, although the truth of these propositions cannot at the time be detected by the organism. For instance, seedlings of some species, when grown in a closed dark box with a tiny hole admitting some light, will grow towards the hole even though the light admitted by it is far too weak to support photosynthesis. This behaviour is beneficial to the species because, in the environments naturally encountered by its seedlings, it is indeed usually the case that instances of weak light coming from some direction are associated with useful light being available in a region located in that direction. We do not suggest that such a seedling "knows" or has "inferred" a concept "light source" as a cluster of defining and consequential propositions. However, it is not unreasonable to suggest that the genetic endowment of the species incorporates in some way an association among a cluster of possible properties of its environment, and that other species which grow in environments where such clustering is not evident will not show such behaviour. Further, we suggest that the incorporation of such a cluster of environmental properties differs from a "concept" formed by a reasoning agent only in that the latter is expressible in language-centred terms such as "proposition" and "assertion".

It seems to us proper to regard the genetic makeup of organisms as incorporating many powerful theories about the natural environment. These have not been induced by any reasoning agent. Rather, mutation and other mechanisms result in the creation of many organisms each incorporating a different set of theories. The inductive process which infers good theories is the natural selection of those organisms which carry them. Our aim in this work is to characterize what are good inductions, regardless of how the inductions have been made. The phototropism of a growing seedling, the alarm cry of a seagull, the concept of momentum, and wave equations of quantum mechanics are all the result of the inductive inference of general propositions from hosts of specific "observations". The methods of induction may differ greatly among these examples, but any satisfactory account of how specific data can lead to general assertions should cover them all.

What then are the problems in giving an account of, or logic for, inductive inference? Clearly, one negative requirement which an inferred general proposition must satisfy is that it should not be contradicted by the accepted data. Thus, we cannot conclude that all crows are black if our observations record a large number of black crows, but a few white ones as well. In strictly logical terms, a single counter-example is sufficient to show a general proposition to be false. In practice, the matter is not so clear-cut. Two qualifications are (almost always) implicit in our assertion of a general proposition. First, we are well aware that a general proposition cannot be proved true by any number of conforming observations, so when we assert "All crows are black", most of us mean something a little less. We expect our audience to understand the assertion as an acceptable abbreviation for something like “My best guess is that all crows are black. However, my evidence is incomplete, and I will be prepared to modify my assertion in the light of conflicting evidence”. If the assertion is so read, the discovery of one or two white crows among millions of black ones would not cause us to apologize abjectly for misleading. Outside of deductive argument from unquestioned premises, an assertion of universal scope is either meant to be understood in this less than literal way or else is almost certainly unjustified. How can anyone possibly know that all crows, past, present and future, are black?

The second qualification is logically more of a problem. Even if we read the assertion “All crows are black” in its absolute sense, will we really abandon it if we see a single white crow fly past? Not necessarily. The counter-example proposition “The crow that just flew past was white” may itself be suspect. Was the bird indeed a crow? Had it been bleached white by some joker? Was it really white or was it a trick of the light or that last drink over lunch? In the real world, a general assertion which has been long and widely held by people knowledgeable in the field is not rejected on the grounds of a single reported contrary observation. There is always the possibility of error in the observation, perhaps even of malicious misrepresentation or delusion.

Even a steady trickle of contrary reports may not suffice to discredit the proposition. If one observer can make an error, so can others, and perhaps one or two mistaken observations per year must be expected. We also tend to be skeptical of observations, no matter how frequent, which cannot be confirmed by others. Joe may report, frequently and consistently, seeing crows which seem to others black, but which to Joe are distinguished from their fellows by a colour which he cannot otherwise describe. Even if Joe shows he is able consistently to distinguish one crow from another by “colour”, we will suspect that the exceptional nature of the observations lies in Joe rather than in what everyone else calls the crows’ colour. Similar doubts may arise with respect to observational apparatus: does it really measure what we think it measures?

Before we reject a well-regarded general assertion, it seems we need to be satisfied that, perhaps only under certain specified but achievable conditions, any competent observer can obtain as many counter-examples as are wished. If the bird can be caught, and we find that any competent ornithologist will confirm that it is indeed a crow and indeed white, we will abandon the proposition at least in its absolute sense. If anyone willing to visit lower Slobovia with a pair of binoculars can count a dozen white crows within a day of arrival, we will abandon it completely. But note what needs to be established before we regard the proposition as false: we need to establish that any competent person can consistently accumulate contrary data. This is itself a general proposition. It asserts something about a possibly large class of phenomena: the unlimited observations of any number of competent observers. Note that the large class of observations might all relate to the one white crow.

To summarize, except in special domains, a well-supported general proposition is not regarded as disproven by a single contrary observation or even a limited number of contrary observations. It is usually rejected only when we come to accept a new general proposition, namely that we can get as much credible contrary evidence as we like. The evidence may come from one or many counter-instances (white crows), but the evidential observations are in principle unlimited. However, the requirement that the inductive inference should not conflict with the data, while valid and necessary, does not much advance our understanding of how the inference can be formed from and be supported by the data.

A third qualification may be implicit in a general assertion. It may be true only in an approximate sense, its context implying that it is meant as an approximation. For instance, Boyle's Law that the pressure of a confined gas rises in proportion to its absolute temperature is still taught at school, but is only approximately true of most gasses.

A second necessary requirement of an inductive inference is that the generalization should be *falsifiable*. The importance of this requirement was first clearly stated by Popper (1934), whose writings have influenced much modern discussion of induction. The first requirement we asked was that the inference not be falsified by the data we have. Now we require also that it be possible for future data to falsify the inference. That is, we require it to be at least conceivable, given all we know of the source and nature of the data, that we might find data sufficient to make us reject the proposition. Essentially, this requirement is equivalent to requiring the inferred proposition to have empirical content. If we can deduce, from what is already known about the nature and source of the data, that it is impossible for future data to meet our criteria for falsifying the proposition, then the proposition is telling us nothing of interest. To accept the proposition is to exclude no repeatable observation except those already excluded as impossible.

Popper rightly criticizes theories, advanced as inductive inferences from known data, which are so phrased and hedged with qualifications that no conceivable new data can be considered as damning. He finds most of his bad examples among social, political and economic theories, but examples are not unknown in other domains. His requirement places on any proposed account of inductive inference the duty of showing that any inference regarded as acceptable in the proposed framework must ipso facto be falsifiable. This duty we hope to fulfill in the present work.

The two requirements above, that an inductively derived general proposition be falsifiable but not yet falsified, are far from sufficient to describe inductive inference. They also perhaps place undue emphasis on the notion of disproof. Many useful inductive inferences (let us call them theories for brevity) are known to be false in the form originally inferred, yet are still regarded as useful premises in further reasoning, provided we are careful. A classic example are Newton's laws of motion and gravitation. Reproducible sources of data in apparent conflict with these laws have been known at least since the early 1900s. The reaction to this "falsification" was, first, a series of quite successful attempts to modify the interpretation of these laws, and then the inference of the new theories of Relativity. The new theories were rapidly recognized as superior to the old Newtonian theories, explaining simply all of the results which appeared to falsify the old theories, at least in their original form. Yet the old unreconstructed Newtonian "laws" continue to be used for the great majority of engineering calculations. Although known to be wrong, they in fact fit and explain vast bodies of data with errors that are negligible compared with the measurement errors and uncertainties of the data. In fact, one of the early concerns of the exponents of the Relativistic theories was to show that the new theories did *not* contradict Newtonian theories except under extreme conditions. Similarly, the new quantum theories which replaced Newtonian mechanics under other extreme conditions had to be shown not to contradict the old theory to any measurable degree outside these extreme conditions.

We are forced to conclude that an account of inductive inference must accommodate the fact that theories which have been conclusively falsified can remain acceptable (albeit within a circumscribed domain of phenomena) even though their basic concepts have been shown to be mistaken. The account must also accommodate the fact that two theories can both command general acceptance even though their formulations appear mutually inconsistent. We accept the present situation that Relativistic theory is basically concerned with the relationships among "events" regarded as having precise locations in space and time, while quantum theory denies the possibility of precisely locating an event involving only finite energy. We accept that Relativistic theory describes gravitational effects in geometric terms, while quantum theory, insofar as it can treat gravity at all, must invoke as yet undiscovered particles.

Any satisfactory account of induction must therefore not be overly concerned with absolute notions of truth and falsification. In practice, we do not expect our inferences to be true. We tolerate falsifying data provided it relates to conditions outside our immediate arena, and we tolerate the co-existence (and even the joint application to the one problem) of theories whose conceptual bases seem to belong to different universes.

1.3 The Demise of Theories

The strictly logical requirement that a theory not be falsified cannot be accepted at face value. Merely being shown to be wrong is not sufficient to damn an inference. We can accept the requirement that a theory be falsifiable, i.e., that we can conceive of data which would falsify a theory, as otherwise the theory is empirically vacuous, but we cannot accept that such a falsification will necessarily lead us to reject the inference, because history shows otherwise. How, then, do we ever come to reject theories?

One possible route to rejection is the accumulation of falsifying data. When a theory is falsified, we may not reject it but we must at least qualify it by restricting its application to arenas not including such data, and/or weakening its assertions to approximations. If falsifying data is found in a sufficient range of conditions, the theory may become so qualified and restricted that it ceases to be falsifiable, i.e., becomes empty. The cases of data which do appear to conform to the theory may be found to be no more than might be expected to arise by chance in the absence of the theory, in which case we may decide that the amended theory explains nothing and should be abandoned.

Another route to rejection is that the theory is never decisively falsified, but is supplemented by a theory of greater accuracy or wider applicability. That is, it is found that all the data explained by the theory is explained as well or better by a new theory, which may in addition explain data not covered by the old theory.

A third route is the usual fate of most hypotheses proposed in a scientific investigation. The theory may be compatible with known data, but not regarded as adding much to our understanding of that data compared with other possible theories about the same data. A new experiment or observation is designed such that its expected outcome, if the theory is valid, is one which would not be expected without the theory. The observation is performed, and does not conform with the prediction of the theory. The theory is then rejected as having little explanatory power for the old data, and not fulfilling the hope of explaining the new data.

A fourth, less common, route is that the theory is supplanted by a new theory which is not (at least initially) in better conformity with the known data either in accuracy or scope, but which is in some way simpler or more "elegant" than the old. The criteria of simplicity and elegance are not obviously quantifiable, especially the latter, and people may legitimately disagree in their assessments of theories on these criteria. However, there might be general agreement that, for instance, of two theories otherwise similar in structure, the one needing fewer numeric values to be assumed in order to explain a set of data is the simpler. Similarly, of two theories requiring the same number of assumed quantities, we might assess as the simpler the theory having the shorter mathematical or logical description.

An example may serve to clarify these notions. Observation of the apparent positions of the planets, sun and moon gave rise to the "Ptolemaic" theory, which supposed the motions of the bodies to be composed of simple circular motions with constant radii and speeds. To fit the observations, it was necessary to assume that the motions of most of the heavenly bodies were epicyclic. That is, a body moved round a circle whose centre was moving round another circle whose centre might be moving round yet another circle. This theory could be made to fit the observations quite well, to predict future movements with fair accuracy, and to predict events such as eclipses. It is structurally a simple theory: the circle is one of the simplest geometric shapes by any criterion. However, it required the assumption of a rather large number of numeric values, the radii and speeds of rotation of all the circles, of which there were two, three or more for each body. These quantities had to be assumed: the theory gave no explanation of their values and asserted no useful relationships among them.

The later Keplerian theory was in marked contrast. Structurally, it might be considered more complex or less elegant, since it assumed the motions to be elliptical rather than circular, and to take place with varying rather than constant speed. Each ellipse requires both a major and minor axis to be specified rather than just a radius. In these respects, the new theory seems messier and more complex than the Ptolemaian. However, only one ellipse is needed per body, rather than several circles. The speeds of motion, while not constant, have a fixed and simple relationship to the position of the body round its elliptic path, and the one number for each body required to describe this relationship was shown to have a fixed relation to the size of the ellipse. Thus, the number of values which had to be assumed dropped from half a dozen per body to essentially two. (We are deliberately oversimplifying here: the descriptions of the orbital planes of the bodies involve more numbers but these are essentially the same in both theories.)

The smaller number of arbitrary constants required by Kepler's laws could be held to outweigh his use of more complex geometry, but the issue was not clear-cut on this score. Of course, as observational accuracy increased, it was found that Kepler's theory required only minor refinement to maintain its agreement with observation, whereas more and more circles had to be added to the epicycles of the Ptolemaian model, each with its new inexplicable numbers. The "simplicity" argument in favour of Kepler's model became overwhelming.

This sketch of how theories may be rejected, usually but not always in favour of a new theory, has argued that rejection is not a simple matter of falsification. Rather, it involves factors such as the scope of data explained, the accuracy of explanation, the number of inexplicable or arbitrary values which must be assumed, and some notion of structural simplicity or elegance. Together, we may call these factors the *explanatory power* of the theory. The explanatory power increases with the volume and diversity of the data explained, and with the accuracy of the explanation. It decreases with the structural complexity of the theory, and with the number and precision of the parameters, initial conditions and unobservable quantities which must be assumed in the explanation.

In the above, we have used the terms *explanation* and *accuracy* without elaboration. While it is our intent to use these words in accordance with normal usage, both are sufficiently loosely used in everyday speech as to demand some definition.

1.4 Approximate Theories

The notion of the accuracy of a theory, as applied to some data, rests on belief that a theory is rarely intended or taken in an absolute sense. If we assert the theory that a floating ship displaces its own weight of water, we do not intend to claim that careful measurement of the ship and the displaced water will show them to be equal within a milligram. Rather, we are claiming that they will be equal within a small margin due to measurement error, the effects of wind and wave, motion of the ship, etc. The theory does not attempt to explain the causes of this margin of error, which in practice might be of the order of 0.1%. We might then say the theory is "accurate" within 0.1%. It could be argued that the theory does indeed claim exact equality, at least under certain ideal and probably unattainable conditions such as zero wind, zero motion, no surface tension, etc., and that it is unfair to regard measurement errors and deviations caused by inevitable disturbances as inaccuracies of the theory. But a theory which asserts a conclusion only under forever unattainable conditions is empty, since it can never apply to real data. It may alternatively be suggested that if careful measurement is made, it will be found that the weight of the ship, plus any downwards pull of surface tension, minus any hydrodynamic wave force, etc., etc., will exactly equal the weight of water displaced within measurement error. But this is not the same theory — it is a more elaborate and perhaps more accurate theory.

In a slightly different vein, the Newtonian equation for the kinetic energy of a moving mass, $E = \frac{1}{2}mv^2$, can be said to have inaccuracies due not only to the kind of error and unobserved effects described above, but also an error of order $\frac{1}{2}v^2/c^2$ because it ignores relativistic effects. If the speed v is less than 1000 km/sec, this inaccuracy is less than 0.01%. Whatever the sources of error, it seems plausible that all theories will fail to match our data exactly, but that some will be more accurate than others. Exactly how we can most usefully quantify the inaccuracy of a theory will be discussed later.

1.5 Explanation

A dictionary definition of the word "explain" is "to make plain or understandable". We take an explanation of a body of data to be a demonstration that the data are not unexpected given a relatively small set of premises. By "not unexpected" we mean that the premises either imply the data propositions, or, more commonly, imply close approximations to the data. Two cases need to be distinguished. In some explanations, the necessary premises are already known and accepted by the reader of the explanation. In this case, the explanation is purely a deductive demonstration that the data should be expected to be more or less as they are, given what is already known. We will not be interested in such explanations. In other explanations, not all of the necessary premises are known *a priori*. Rather, the explanation proposes one or more new premises, and then goes on to show that the new premises, combined with ones already known and accepted, imply or approximately imply the data. In forming such explanations, the new premises are an inductive inference from the data. Typically, they are general propositions which cannot be deduced from the data and premises already known. However, if they are assumed to be true, the data is found to be unsurprising.

Two imaginary examples may clarify the distinction we wish to draw.

First, suppose there is an amateur carpenter who knows and is familiar with concepts of length, area and angle, is competent at arithmetic and elementary algebra, but who has never studied geometry. The carpenter notices that it is possible to make right-angled triangular frames whose sides are integral numbers of decimetres, but only if the numbers are multiples of a few sets such as {3, 4, 5}, {5, 12, 13} and {15, 8, 17}. In seeking an explanation of these observations, he might, given time, deduce Pythagoras's theorem from the premises he knows and accepts about lines, areas and angles, then deduce that these sets of integers satisfy the theorem but most others do not. He might even be able to deduce that any such integer set must have the form $\{(a^2 - b^2), 2ab, (a^2 + b^2)\}$, where a and b are any unequal positive integers. This would be an explanation of the first kind: a demonstration that what has been observed is not surprising given what the carpenter already believed. No new premise is required and nothing is inductively inferred from the data.

Now imagine an ancient Egyptian surveyor who was a competent user of geometry and knew many of the simple properties of triangles, but otherwise knew no more than his fellows. In particular, he knew that the sum of the three angles of a triangle equals two right angles (180°). As he rose in the surveyors' hierarchy, he noticed minor inconsistencies appearing in the data and ordered the large-scale resurveying of the kingdom. He was surprised to find that in the largest triangles covered by the survey, the sum of the angles consistently exceeded 180° by a small amount, which seemed to be proportional to the area of the triangle. After much reflection, he finds that the data can be explained if he supposes that the world is not flat, as everyone had thought, but spherical, with a diameter of about 7000 miles. If he adds this premise to what he knows of geometry, he can deduce that the sum of the angles of a triangular piece of land should exceed 180° by about five thousandths of a degree for every thousand square miles of area. This deduction agrees well with the survey data, so he accepts the explanation.

This explanation requires a new premise in addition to what the surveyor knew. The new premise, that the world is a sphere of 7000 miles diameter, is more complex than the old implicit assumption of a flat earth, and involves a number. It could not be deduced from the data. Rather, it was derived by induction from the data, and the diameter estimated from the data. The new premise is falsifiable — new data could conceivably show it to be untrue — and is actually false. The world is not quite spherical, nor is its diameter 7000 miles. However, the explanation is good. At the expense of one inferred premise of no great complexity, the deviations of the data from what is expected are greatly reduced. Henceforth, we will restrict the term *explanation* to this second sort, which involves the inductive inference of a new premise or theory and/or the estimation of unknown quantities.

1.5.1 Explanatory Power

Our view is that an inductive inference from a body of data is a premise which, if assumed, allows the data to be explained. Other propositions already known and accepted may be involved in the explanation, but are by themselves insufficient to allow a satisfactory explanation of a purely deductive kind.

To develop this view into an account of when an inductive inference can be regarded as satisfactory and how competing inferences may be compared, it is necessary to develop a quantitative measure of the merit of an explanation, or at least of the relative merits of competing explanations. We have suggested that the explanatory power of an inductive inference or theory increases with the volume of data explained and the accuracy of the explanation. It decreases with the complexity of the theory, the number of inexplicable parameter values appearing in the theory, and (we will see later) the precision with which these quantities must be specified in order to achieve an accurate explanation. In short, a good inductive inference is one which explains much by assuming little. Other considerations, such as causal structure, have been proposed as contributing to or necessary for explanatory power. At least for now, we will not consider them, and rather discuss only what follows from the criteria above. We now propose a step towards quantifying these considerations.

First, we will simplify the problem by assuming that all the inductive inferences to be assessed apply to the same body of data. The extension to situations where one theory explains more data than another is easy but is best treated later.

For a given fixed body of data, we propose to recast all competing explanations into the same canonical form.

1.5.2 The Explanation Message

An *explanation message* of a body of data comprises two parts. The first is a statement of all the inductively derived premises used in the explanation, including numeric values for quantities assumed in these premises (the diameter of the earth, for example). The second part of the explanation is a statement of all those details of the data which cannot be deduced from the combination of the induced premises and such other premises as are already known, accepted, and not in question. Let us call the already-known premises the *prior premises*. Being already known to the receiver, no statement of the prior premises need appear in the message.

First, note that a person knowing only the prior premises and not the data is able to recover the original body of data exactly from this message. The first part tells him to assume the truth of the new premises — the "theory". From these and the prior premises, he can then deduce much about the data. The second part completes his knowledge of the data by telling him all the details which he could not so deduce. Thus, the explanation message may be regarded as a restatement of the data without loss of any information or detail. The restatement is in a "coded" form which can be "decoded" by anyone with knowledge of the prior premises. Another way of regarding the explanation message is that it states a theory about the data, then states the data itself in a condensed form which assumes the truth of the theory.

We will argue that the best explanation of the data is the one leading to the shortest explanation message, and that the best inductive inference which can be drawn from the data is the inference used in the shortest explanation message. That is, we claim the shortness of the explanation message using an inferred theory is a measure of its explanatory power. Henceforth, we will not distinguish between an explanation message and an explanation expressed in other forms. When we refer to an explanation or its length, we mean the explanation message or its length.

Even from the above informal account, it is clear that the length of an explanation takes into account all the factors affecting explanatory power. The length of the first part, which states the inductively inferred premises or theory, will be longer for a complex theory than for a short one. Its length increases with every quantity assumed, as the first part must state its assumed numeric value. Its length increases with the precision to which these values need be specified.

On the other hand, the length of the second part decreases with the scope and accuracy of the theory. Data falling outside the scope of the theory must be stated in full, since nothing about such data can be deduced from the theory. Hence, the greater the scope of the theory, the less data need be stated fully. Typically, the theory, together with the prior premises, will not allow exact deduction of the data as observed. For quantitative data, the best that we can hope is that values may be deduced close to but not exactly equaling the measured values. The observed value may be corrupted by measurement error, and the deduced value will often be deduced in part from other values in the data, and hence itself be corrupted by error. And of course the theory and its parameters may only be approximate. Thus, for quantitative data within the scope of the theory, the second part of the message must at least record the differences between deduced and measured values. The more accurate