CONTENTS

CHAPTER ONE

Introduction

THE STORY OF how cosmology grew is fairly simple, compared to what people have been doing in other branches of science, but still complicated enough that sorting it out requires a better plan than the common practice in science. Papers reporting research in cosmology and other parts of physics usually begin with an outline of what came before. Abandoned ideas and roads not taken are seldom mentioned, and there is the natural human tendency to follow patterns of attributions found in introductions in other recent papers. This builds evolving creation stories that efficiently set the current context for the research to be described. We tell these creation stories in the classroom for a quick introduction to what we are really interested in: the nature of the science. But the stories tend to be at best only vaguely related to what actually happened. Their gross incompleteness may not be a problem for ongoing research, except of course when good ideas have been overlooked or abandoned and lost. But the creation stories leave a woefully incomplete and inaccurate impression of how science is done.

To do better, we have to look further back in time, and we certainly have to consider the ideas that seemed interesting but were falsified or otherwise found not to be so interesting after all. A closer account of how cosmology grew presented in chronological order would be awkward, because different parts of what became the established theory were making progress at different rates following different methods and motivations until they started to come together. This account accordingly presents histories of six lines of research that were developing more or less separately. They are reviewed in Chapters 2 to 7. The advantage is a modest degree of continuity within each chapter. The disadvantage is the need to refer back and forth in time to what was happening in different lines of research. The arrangement is explained in more detail in Section 1.2 in this chapter, in the form of an outline and guide to the story to come. But first let us consider our traditions of research in the natural sciences, with particular attention to the operating conditions in cosmology.

1.1 The Science and Philosophy of Cosmology

The starting assumption for cosmology, as in all branches of natural science, is that nature operates by kinds of logic and rules that we can discover by careful examination of what is observed, informed by past experience of what has worked. The results are impressive; I urge any who might disagree to consider the rich fundamental physics employed in the construction and operation of their cellphones. But despite the many demonstrations of its power, physics, along with all the rest of natural science, is incomplete. Maybe discoveries to come will make the physical basis for science complete, revealing the final rules by which nature operates. Or maybe it's successive approximations all the way down.

The standard and accepted methods of science must be adapted to what can be done, of course. In physical cosmology and extragalactic astronomy, we can look but never touch. In cosmology, we cannot run the experiment again; we must instead resort to what can be inferred from fossils of times past. We find some fossils relatively nearby, as in the rocks on Earth and the stars in our galaxy and others, all of which have their own creation stories. Our past light cone offers us views of times past, because radiation detected here has been approaching us at the speed of light: the greater the distance of an object, the earlier in the evolution of the universe it is observed. Our light cone integrated through human history captures an exceedingly thin slice of what has been happening, but it reveals the way things were over a long range of time in a large universe that offers a lot to see and to seek to interpret.

The research path to where we are now in cosmology is marked by debates on open questions, as is usual in natural science. But the issues in cosmology have been defended and criticized with considerably more vigor than might have been expected from the modest weight of the evidence at the time. This was in part because observations that might settle questions in cosmology have tended to seem just out of reach or perhaps just barely possible. And I think an important factor has been the tendency to take a personal interest in the nature of our world. Is the universe really evolving, or might it be in a steady state? If evolving, how might it all end, in a big crunch or a big freeze? And where did it all come from? Such debates are quieter now, because we at last have a theory that passes an abundance of tests, but they continue.

Research in cosmology in the twentieth century usually was done in small groups, often an individual working alone or maybe with a colleague or a student or two. In the twenty-first century, ongoing research in cosmology grew richer and called for larger groups to develop special-purpose equipment for data acquisition, which in turn called for groups of comparable size to reduce the data and interpret it. Big Science has become important to this subject: We have to get used to gathering data in vast amounts, analyzing these data, and employing massive numerical simulations that help bridge the gap between theory and observation. But Big Science best takes aim at well-motivated and sharply defined questions. The main considerations in this book are about how small groups working on seemingly independent lines of research found their results coming together in a cosmology that looked good enough to call for the demanding tests afforded by Big Science. I date this revolutionary convergence to a credible theory to the half decade from 1998 to 2003.

Research certainly continued to be active and productive after the revolution; the difference is that the community had agreed on a paradigm, in Kuhn's (1962) terms. (This is what the majority was thinking, of course; not all agreed.) An example of the adherence to the normal science of cosmology is the study of how the galaxies formed and evolved, which builds theories of galaxy formation on the standard and accepted theory of the evolution of the universe. Normal scientific research of this sort may uncover anomalies that point to a still better underlying theory. This is a point of particular interest in cosmology, because the theory is at the same time well and persuasively tested and particularly incomplete.

Our present normal science of cosmology includes an excellent case for the presence of dark matter that interacts weakly if at all with ordinary matter. There are tight constraints on the properties of dark matter, but no clear evidence exists of detection of this substance other than the inference from the effects of its gravitational attraction. Some argue that dark matter will remain only hypothetical until there is more evidence of it than that: maybe detection in the laboratory, maybe indications of what it is doing to galaxies apart from holding them together. Others argue that the case for dark matter already is so tight that it is abundantly clear that the dark matter really exists. The same applies to Einstein's cosmological constant, Λ. It has gained a new name: dark energy. But that is a poor disguise for a fudge factor that we accept because it serves to unify theory and observations so well. There are other fudge factors, hypotheses to allow the theory to save the phenomena, in the present standard science of cosmology and in all the other branches of natural science. Research in the sciences continues to improve tests of our theories that, whether intended or not, may lead to better theories that inspire new tests. And they might on occasion replace fudge factors with unified theories in paradigms that bring parts of this enterprise closer together. It happens.

The physical cosmology that is the subject of this history is an empirical science, that is, it is based on and tested by what can be observed or measured by detectors, such as microscopes and telescopes and people. But we must pay attention to the role of theory, and intuition, and what Richard Dawid (2013 and 2017) terms "nonempirical theory assessment." The prime example in this history is that during most of the past century of research in cosmology, the community majority implicitly accepted Einstein's general theory of relativity. Few pointed out that this is an enormous extrapolation from the few meager tests of general relativity that we had in the 1960s. By the 1990s, as research in cosmology was starting to converge on a well-tested theory, there were demanding checks of the predictions of general relativity on scales ranging from the laboratory to the solar system, probing out to length scales of about 10^{13} cm. But the application to cosmology on the scale of the Hubble length, about 10^{28} cm, extrapolates from the precision tests by some fifteen orders of magnitude in length scale. This was not often mentioned, in my experience, and when mentioned, it tended to make some scientists a little uneasy, at least temporarily. In the first decades of the twenty-first century, the parts of general relativity that are relevant to the standard cosmology have passed an abundance of demanding tests. In short, the theory Einstein built on laboratory experiments was seriously tested only by the orbit of the planet Mercury. (The test of the prediction of the gravitational deflection of light by the mass of the sun, led by the people pictured in Plate III, was heavily cried up but in retrospect, their evidence seems marginal.) We find that this theory successfully extrapolates to applications on the immense scales of the observable universe. It is a remarkable result.

General relativity is an elegant extension of electromagnetism in flat spacetime; it has been said that it is a theory waiting to be found (though that is easier to say in hindsight). The faith in its extrapolation exemplifies the powerful influence and very real successes of nonempirical theory assessment. Of course, influential nonempirical assessments can mislead: Consider that in the 1930s through the 1990s, few objected to the assertions by respected experts that Einstein's cosmological constant, Λ, surely may be discarded. The evidence now is that Λ, under its new name—dark energy—is an essential part of our well-tested cosmology.

The practice of nonempirical assessments is sometimes termed "post-empiricism," but I have not found this term in Dawid's writing. Dawid (in a personal communication, 2018) states instead that

> non-empirical assessment as I understand it crucially depends on the ongoing collection of empirical data elsewhere in the research field and on the continued search for empirical confirmation of the theory under scrutiny. In a "post-empirical" phase where no substantially new data comes in any more, non-empirical assessment would get increasingly questionable and eventually would come to a halt as well.

This is consistent with what I understand to be normal practice in the physical sciences. That is, I have in mind the kind of nonempirical assessments we have been practicing all along without thinking much about it.

I take account of three other kinds of assessments: personal; community, though some may disagree; and pragmatic. The first two speak for themselves. I take examples of the third from cosmology. The usual practice has been to analyze data and observations in terms of general relativity. This surely has been due in part to the beauty of the theory, and in part to respect for Albert Einstein's magnificent intuition. But it was important also that the use of a common theory allowed comparisons of conclusions from independent analyses of the same or different data on a common fundamental ground. I do not imagine much thought has been given to this point, but I believe the implicitly pragmatic approach in cosmology (and I suppose in other branches of natural science) has helped reduce the chaos of multiple theories.

The pragmatic approach to science, if carried too far, could waste time and resources by directing research along a path as it grows increasingly clear that something is wrong. And even if the popular and pragmatically chosen path proves to be leading us in a useful direction, it can be important to have well-defended alternatives to standard ideas to motivate careful evaluations of approved ideas and observations. It may reveal corrections large or small that point toward a more profitable path. For example, a stimulating proposal in the mid-twentieth century was that textbook physics may have to be adjusted to include continual spontaneous creation of matter. The brave souls who argued for this steady-state cosmology were not always gently treated, but from what I saw, they gave as good as they got in debates over the relative merits of the general relativity and steady-state world views, arguments that were more intense than warranted by the evidence for or against either side. The idea of continual creation in the universe as it is now is no longer seriously considered in cosmology, but it had a healthy effect. New ideas can inspire defense and attacks that stimulate research, while a pragmatic defense of the old ways may help keep research from degenerating into confusion.

An important example of an implicitly pragmatic assessment is the general acceptance of Einstein's proposal that the universe is homogeneous in the average over local irregularities. Prior to the 1960s, there was scant evidence of this. Maps of distributions of the galaxies across the sky suggested instead that the galaxies are moving away from one another into space that is asymptotically empty or close to it, as in a fractal galaxy distribution. But whether by accident or design, this quite pertinent thought was put aside for the most part, and the main debate kept more sharply focused on the concepts of evolution or else a steady state of a nearly homogeneous universe. The first serious evidence for homogeneity came a half century after Einstein, from research for other purposes in the 1960s, as will be discussed in Chapter 2. Whether by good luck or good taste, the community was not much distracted by the elegant but wrong idea of a fractal universe.

It is not always easy to see why some issues receive much more attention than others; I suppose such things are to be considered eventualities. We do have reasonably clear standards for rejecting an apparently interesting idea. For example, the steady-state cosmology introduced in 1948 is elegant, but its predictions clearly violate the later accumulation of empirical tests. I do not know of a clear prescription for a move in the other direction, namely, the promotion of a working model to a standard theory. We might use the term "community opinion" to describe such decisions.

In 1990, general relativity usually was taken to be the appropriate basis for the study of the large-scale nature of the universe, but as argued above, it was an implicitly pragmatic assessment that the theory was serving well as a working basis for research. In 2003, after the revolution, the cosmological tests gave weight to the community opinion that the universe actually is well described by general relativity applied to the set of assumptions in what became known as the ΛCDM cosmological model. The introduction of these assumptions, including Einstein's cosmological constant Λ and the hypothetical cold dark matter, is reviewed in Section 8.2. Some disagreed, to be sure, but to most the accumulation of evidence (reviewed in Chapter 9) had become tight enough to have emboldened talk of what "really happened" far away and in the remote past, based on the ΛCDM theory. The notion of reality is complicated, so a more secure statement would be that whatever happened—and we assume something did happen—left traces that closely resemble those predicted by ΛCDM. And the traces are abundant and well enough cross-checked that the community opinion, including mine, is that this theory almost certainly is a useful though incomplete approximation to what actually happened.

1.2 An Overview

I have sorted this history of cosmology into lines of research that operated more or less independently of one another through stretches of time in the twentieth century. I consider the developments in each of the lines of research roughly in chronological order, but because different lines of research were at best only loosely coordinated, there have to be references back and forth in time as different lines of research started to interact. This outline is meant to explain how I have arranged the presentation of the research and how it all fits together, at least roughly, apart from the wrong turns taken.

I begin in Chapter 2 with considerations of Albert Einstein's (1917) proposal, from pure thought, that a philosophically sensible universe is homogeneous and isotropic: no preferred center or direction, no observable edges to the universe as we see it around us. That of course is apart from the minor irregularities of matter concentrated in people and planets and stars. Einstein's homogeneity is essential to the thought that we might be able to find a theory of the universe as a whole rather than of one or another of its parts. It was an inspired intuitive vision or maybe just a lucky guess; Einstein certainly had no observational evidence that suggested it. The history of how Einstein's thought was received and tested exemplifies the interplay in science between theory and practice, sometimes reinforcing each other; sometimes in serious tension; and, as in this case, sometimes aided by unexpected developments. Because I have not found a full discussion elsewhere, I consider in some detail the development of the evidence that supports what became known as Einstein's cosmological principle.

Einstein's general theory of relativity predicts that a close-to-homogeneous universe has to expand or contract. Expansion was indicated by astronomers' observations that starlight from galaxies of stars is shifted to the red, as if Doppler shifted, because the galaxies are moving away from us. Chapter 3 reviews the importance of the discovery that the Doppler shift, or redshift, is larger for galaxies that are farther away. This is the expected behavior if the universe is expanding in a nearly homogeneous way. The big bang cosmology discussed in Sections 3.1 and 3.2 uses the general theory of relativity to describe the evolution of a near-homogeneous expanding universe.

We should pause here to note that the name, "big bang," is inappropriate, because a bang connotes an event in spacetime. Unlike a familiar bang, this cosmology has nothing to do with a special position or time. The theory is instead a description of cosmic evolution of a universe that is homogeneous on average, and it attempts to follow cosmic evolution to the present from the earliest time of formation of fossils that can be observed and interpreted. That has come to include the epoch of light-element formation, when the temperature of the universe was some nine orders of magnitude larger than it is now. This is a spectacular extrapolation back in time, but not to a bang, and not to a singular start of things: We must assume that something different happened before the singularity. Simon Mitton (2005) concludes that Fred Hoyle coined the term "big bang" for a lecture on BBC radio in March 1949. It was meant as a pejorative; Hoyle favored the steady-state picture. Though unfortunate, the name "big bang" is commonly accepted. I have not encountered a better term, and the pragmatic assessment is that it is to be used in this book.

It was important that there were testable alternatives to the big bang picture; these alternatives inspired the search for tests. The leading idea, the steady-state model, is discussed in Section 3.3. It will be termed the "1948 steady-state model" to distinguish it from variants introduced later. In contrast to the prominence of the steady-state alternative to the big bang model through the mid-1960s, the leading alternative to Einstein's idea of homogeneity—a fractal distribution of matter—only became widely discussed after we at last had reasonably clear evidence of homogeneity (Section 2.6).

Hermann Bondi's (1952, 1960) book *Cosmology* in two editions, gives a valuable picture of thinking at the time. Which if either of the big bang or 1948 steady-state models, or perhaps some other model then still being considered, is the most reasonable and sensible, and on what grounds, empirical or nonempirical? Helge Kragh (1996) presents a historian's perspective of this mainstream research in cosmology up to the 1960s. Sections 3.4–3.7 augment these sources with my thoughts about the similarities and differences of assessments of the two cosmologies. I take it that in the 1950s and early 1960s, nonempirical issues account for the lack of popularity of the steady-state model in many quarters, despite its greater predictive power for observers.

The weaker predictive power of the big bang model may help account for the abundance of nonemipirical assessments discussed in Section 3.5.

The greatest effort devoted to the empirical study of the big bang cosmological model in the years around 1990 was the measurement of the mean mass density. Sections 3.6.3 and 3.6.4 review the considerable variety of these probes, and Section 3.6.5 offers an overview of what was learned. The motivation for this large effort was in part to see whether the mass density is large enough that its gravity will cause the expansion to stop and the universe to collapse, and the results were important for the empirical establishment of cosmology. But I think in large part the motivation became simply that this is a fascinating problem whose resolution is difficult but maybe not quite impossible.

The topic of Chapter 4 is the informative fossils left from a time when the universe was very different from now, dense and hot enough to produce the light elements and the sea of thermal radiation that nearly uniformly fills space. Since it was (and is) exceedingly difficult to imagine how the light elements and the radiation with its thermal spectrum could have originated in the universe as it is now, these fossils were a valuable addition to the evidence that our universe is evolving, not in a steady state. The book *Finding the Big Bang* (Peebles, Page, and Partrige 2009) recalls how these fossils were recognized in the mid-1960s, with recollections from those involved of how the recognition led to the research that produced the first good evidence that our universe really did evolve from a hot early state at about the rate of expansion predicted by general relativity. The tangled story of how Gamow and colleagues anticipated these fossils a decade before they were recognized is presented in the paper, "Discovery of the Hot Big Bang: What Happened in 1948" (Peebles 2014). Section 4.2 presents a shorter version of the main points. The sea of thermal radiation has become known as the cosmic microwave background, or CMB. The later developments leading to its central place in the revolution that established the ΛCDM cosmology are reviewed in Chapter 9. This theory of the expanding universe assumes the general theory of relativity applied to a close-to-homogeneous universe (Chapter 2), the presence of Einstein's cosmological constant Λ (Section 3.5), dark matter (Chapter 7), and particular choices of initial conditions (Section 5.2.6).

It was natural to explore how the very evident departures from Einstein's homogeneity—stars in galaxies in groups and clusters of galaxies—might have formed in an expanding universe. In the established cosmology, cosmic structure formed by the gravitational instability of the relativistic expanding universe. The early confusion about the physical meaning of this instability is an important part of the history. These considerations are reviewed in Chapter 5, along with assessments of early scenarios of how cosmic structure might have formed. The importance of these considerations for the convergence to the standard cosmology is a recurring topic throughout the rest of this book.

The subject of Chapter 6 is the astronomers' discoveries of apparent anomalies in the measurements of masses of galaxies and concentrations of galaxies. Other accounts of the exploration of these phenomena are in Courteau *et al.* (2014) and de Swart, Bertone, and van Dongen (2017). Fritz Zwicky was the first to recognize the phenomenon: He saw that the galaxies in the rich Coma Cluster of galaxies seem to be moving relative to one another too rapidly to be held together by the gravitational attraction of the mass seen in the stars in the galaxies in the cluster. One way to put it is that the mass required to hold this concentration of galaxies together by gravity seemed to be missing, always assuming the gravitational inverse square law of gravity (in the nonrelativistic Newtonian limit of general relativity). It was later seen that mass also seemed to be missing from the outer parts of spiral galaxies, based on the measurements discussed in Section 6.3 of circular motions of stars and gas in the discs of spiral galaxies. Much the same conclusion came from the studies described in Section 6.4 of how galaxies with prominent discs acquired their elegant spiral patterns. By the mid-1970s, it had become clear that understanding this is much easier if the seen mass is gravitationally held in near-circular motion in the disc with the help of the gravitational attraction of less-luminous matter that is more securely stabilized by more nearly random orientations of the orbits.

These observations pointed to a key idea for the establishment of cosmology: the existence of "dark matter," the new name for what was variously known as "missing," "hidden," or "subluminal" mass. The idea came almost entirely out of pursuits in astronomy, not cosmology, and for this purpose, the subluminal component need not be very exotic: low-mass stars would do, though they would have to be present in surprising abundance relative to counts of the more luminous observed stars. But in the 1970s, another key idea for cosmology was growing out of particle physicists' growing interest in the possible forms of nonbaryonic matter. Gas and plasma, people, planets, and normal stars are all forms of what is termed "baryonic matter." Most of the mass of baryonic matter is in atomic nuclei; the accompanying electrons are termed "leptons," but they are also counted in the mass of baryonic matter. The neutrinos are leptons that we now know have small but nonzero rest masses. Thus they act as nonbaryonic dark matter that contributes to the masses of galaxies, but in the standard cosmology, this contribution is much smaller than the total indicated by the astronomical evidence. We need a new kind of nonbaryonic matter.

The thought that the astronomers' subluminal matter is the particle physicists' nonbaryonic matter and the cosmologists' dark matter was and remains a conjecture at the time of writing. The only empirical evidence of the new nonbaryonic dark matter is the effect of its gravity. It has been a productive idea, however, that passes demanding checks. The particle physicists' considerations of nonbaryonic matter reviewed in Chapter 7 takes into account the condition that if this nonbaryonic matter were produced in the hot early stages of expansion of the universe, then its remnant mass density must not exceed that allowed by the relativistic big bang cosmological model (again, assuming the relativistic theory). But it is notable that cosmologists took over the notion of nonbaryonic dark matter before the particle physics community had taken much interest in the astronomers' evidence of the presence of subluminal matter.

The nonbaryonic dark matter most broadly discussed in the 1980s came in two varieties, cold and hot. The latter would be one of the known class of neutrinos with rest mass of a few tens of electron volts (Sections 5.2.7 and 7.1). The initially hot (meaning rapidly streaming) neutrinos in the early universe would have smoothed the mass distribution, and that smoothing would have tended to cause the first generation of structure to be massive systems that must have fragmented to form galaxies. The spurious indication in 1980 of a laboratory detection of a neutrino mass appropriate for the hot dark matter picture certainly enhanced interest in the indicated formation of galaxies by fragmentation. This model was considered but had to be rejected: the observations show hierarchical growth of structure, from smaller to larger mass distributions.

The prototype for the nonbaryonic matter that is an essential component of the established cosmology was introduced by particle physicists in 1977. The idea occurred to five groups who published in the space of 2 months. These papers do not exhibit much interest in the astronomers' subluminal mass phenomena, but the considerations certainly were relevant to subluminal matter. Was this a curious coincidence or an idea that somehow was "in the air?" This is considered a little further in Sections 7.2.1 and 10.4.

Sections 8.1 and 8.2 review why in the early 1980s cosmologists co-opted the astronomers' subluminal mass and the particle physicists' nonbaryonic matter in what became known as the standard cold dark matter, or sCDM, cosmological model. The letter "s" might be taken to mean that the model was designed to be simple (as it was) but it instead signified "standard," not because it was established but because it came first. It was meant to distinguish this version from the many variants to be considered in Section 8.4. A large part of the cosmology community soon adopted variants of the sCDM model as bases for exploration of how galaxies might have formed in the observed patterns of their space distribution and motions (Section 8.3), and for analyses of the effect of galaxy formation on the angular distribution of the sea of thermal radiation. This widespread adoption was arguably overenthusiastic, because it was easy to devise other models, less simple to be sure, that fit what we knew at the time. And it was complicated by the nonempirical feeling that space sections surely are flat. In general relativity that could be because the mass density is large enough to produce flat space sections, or because Einstein's cosmological constant, Λ, makes it so. The nonempirical reasons for preferring flat space sections, preferably without resorting to Λ, are discussed in Section 3.5. These reasons were influential and long-lasting enough to have played a significant role in the confusion of variants and alternatives to the sCDM idea considered in the 1990s.

The reduction of confusion in the years 1998–2003 was great enough to be termed a revolution. It was driven by the two great experimental advances discussed in Chapter 9. The first is the measurement of the relation between the redshift of the spectrum of an object and its brightness in the sky, given its luminosity: the cosmological redshift–magnitude relation. Its detection had been a goal for cosmology since the 1930s; it was at last accomplished by two independent groups at the turn of the century (Section 9.1). The second is the detailed mapping of the angular distribution of the CMB radiation. Work on this began in the mid-1960s, and coincidently also produced demanding constraints on cosmological models at the turn of the century. These results from the two sets of measurements, together with what was already known, made a tight case for the presence of Einstein's cosmological constant Λ and the nonbaryonic CDM in the relativistic hot big bang ΛCDM theory. It was a dramatic development.

It was proper to have asked whether the introduction of two very significant hypothetical components, CDM and Λ, along with all the other assumptions that go into the choice of a cosmological model, might only amount to adjusting the theory to fit the measurements. That line of debate did not become very prominent, because the ΛCDM cosmology that fit the two critical measurements brought together so many other lines of evidence in a tight network of empirical tests. This is the topic of Section 9.3.

By the year 2003, the community had at last settled on a respectably well-supported theory of the large-scale nature of the universe. Skeptics remained, as is appropriate, for this theory is an immense extension of the reach of established physics. Indeed, the 2003 theory has been modified to fit later measurements, but these changes amount to fine adjustments of parameters, not challenges to the basic framework of the theory. It is the nature of science to advance by successive approximations, and it would not be at all surprising to find that there is a still better theory than ΛCDM. But we have excellent reason to expect that a better theory will describe a universe that behaves much like ΛCDM, because ΛCDM passes an abundance of empirical tests that probe the universe in so many different ways.

I cannot think of any lesson to be drawn from this story of how cosmology has extended the boundaries of established science that cannot be drawn from other branches of natural science. This is no surprise, because cosmology operates by the methods of natural science. But I think there are lessons to be drawn with greater clarity in the relatively uncluttered historical development of this subject. My offerings are given in Chapter 10.

INDEX