1
Introduction

This book is concerned with the physical processes related to the formation and evolution of galaxies. Simply put, a galaxy is a dynamically bound system that consists of many stars. A typical bright galaxy, such as our own Milky Way, contains a few times 10^{10} stars and has a diameter ($\sim 20\,\mathrm{kpc}$) that is several hundred times smaller than the mean separation between bright galaxies. Since most of the visible stars in the Universe belong to a galaxy, the number density of stars within a galaxy is about 10^7 times higher than the mean number density of stars in the Universe as a whole. In this sense, galaxies are well-defined, astronomical identities. They are also extraordinarily beautiful and diverse objects whose nature, structure and origin have intrigued astronomers ever since the first galaxy images were taken in the mid-nineteenth century.

The goal of this book is to show how physical principles can be used to understand the formation and evolution of galaxies. Viewed as a physical process, galaxy formation and evolution involve two different aspects: (i) initial and boundary conditions; and (ii) physical processes which drive evolution. Thus, in very broad terms, our study will consist of the following parts:

- Cosmology: Since we are dealing with events on cosmological time and length scales, we need to understand the space-time structure on large scales. One can think of the cosmological framework as the stage on which galaxy formation and evolution take place.
- Initial conditions: These were set by physical processes in the early Universe which are beyond our direct view, and which took place under conditions far different from those we can reproduce in earth-bound laboratories.
- Physical processes: As we will show in this book, the basic physics required to study galaxy formation and evolution includes general relativity, hydrodynamics, dynamics of collisionless systems, plasma physics, thermodynamics, electrodynamics, atomic, nuclear and particle physics, and the theory of radiation processes.

In a sense, galaxy formation and evolution can therefore be thought of as an application of (relatively) well-known physics with cosmological initial and boundary conditions. As in many other branches of applied physics, the phenomena to be studied are diverse and interact in many different ways. Furthermore, the physical processes involved in galaxy formation cover some 23 orders of magnitude in physical size, from the scale of the Universe itself down to the scale of individual stars, and about four orders of magnitude in time scales, from the age of the Universe to that of the lifetime of individual, massive stars. Put together, it makes the formation and evolution of galaxies a subject of great complexity.

From an empirical point of view, the study of galaxy formation and evolution is very different from most other areas of experimental physics. This is due mainly to the fact that even the shortest timescales involved are much longer than that of a human being. Consequently, we cannot witness the actual evolution of individual galaxies. However, because the speed of light is finite, looking at galaxies at larger distances from us is equivalent to looking at galaxies when

the Universe was younger. Therefore, we may hope to infer how galaxies form and evolve by comparing their properties, in a statistical sense, at different epochs. In addition, at each epoch we can try to identify regularities and correspondences among the galaxy population. Although galaxies span a wide range in masses, sizes and morphologies, to the extent that no two galaxies are alike, the structural parameters of galaxies also obey various scaling relations, some of which are remarkably tight. These relations must hold important information regarding the physical processes that underlie them, and any successful theory of galaxy formation has to be able to explain their origin.

Galaxies are not only interesting in their own right, they also play a pivotal role in our study of the structure and evolution of the Universe. They are bright, long-lived and abundant, and so can be observed in large numbers over cosmological distances and time scales. This makes them unique tracers of the evolution of the Universe as a whole, and detailed studies of their large scale distribution can provide important constraints on cosmological parameters. In this book we therefore also describe the large scale distribution of galaxies, and discuss how it can be used to test cosmological models.

In Chapter 2 we start by describing the observational properties of stars, galaxies and the large scale structure of the Universe as a whole. Chapters 3 through 10 describe the various physical ingredients needed for a self-consistent model of galaxy formation, ranging from the cosmological framework to the formation and evolution of individual stars. Finally, in Chapters 11 to 16 we combine these physical ingredients to examine how galaxies form and evolve in a cosmological context, using the observational data as constraints.

The purpose of this introductory chapter is to sketch our current ideas about galaxies and their formation process, without going into any detail. After a brief overview of some observed properties of galaxies, we list the various physical processes that play a role in galaxy formation and outline how they are connected. We also give a brief historical overview of how our current views of galaxy formation have been shaped.

1.1 The Diversity of the Galaxy Population

Galaxies are a diverse class of objects. This means that a large number of parameters is required in order to characterize any given galaxy. One of the main goals of any theory of galaxy formation is to explain the full probability distribution function of all these parameters. In particular, as we will see in Chapter 2, many of these parameters are correlated with each other, a fact which any successful theory of galaxy formation should also be able to reproduce.

Here we list briefly the most salient parameters that characterize a galaxy. This overview is necessarily brief and certainly not complete. However, it serves to stress the diversity of the galaxy population, and to highlight some of the most important observational aspects that galaxy formation theories need to address. A more thorough description of the observational properties of galaxies is given in Chapter 2.

(a) Morphology One of the most noticeable properties of the galaxy population is the existence of two basic galaxy types: spirals and ellipticals. Elliptical galaxies are mildly flattened, ellipsoidal systems that are mainly supported by the random motions of their stars. Spiral galaxies, on the other hand, have highly flattened disks that are mainly supported by rotation. Consequently, they are also often referred to as disk galaxies. The name 'spiral' comes from the fact that the gas and stars in the disk often reveal a clear spiral pattern. Finally, for historical reasons, ellipticals and spirals are also called early- and late-type galaxies, respectively.

Most galaxies, however, are neither a perfect ellipsoid nor a perfect disk, but rather a combination of both. When the disk is the dominant component, its ellipsoidal component is generally

called the bulge. In the opposite case, of a large ellipsoidal system with a small disk, one typically talks about a disky elliptical. One of the earliest classification schemes for galaxies, which is still heavily used, is the Hubble sequence. Roughly speaking, the Hubble sequence is a sequence in the admixture of the disk and ellipsoidal components in a galaxy, which ranges from early-type ellipticals that are pure ellipsoids to late-type spirals that are pure disks. As we will see in Chapter 2, the important aspect of the Hubble sequence is that many intrinsic properties of galaxies, such as luminosity, color, and gas content, change systematically along this sequence. In addition, disks and ellipsoids most likely have very different formation mechanisms. Therefore, the morphology of a galaxy, or its location along the Hubble sequence, is directly related to its formation history.

For completeness, we stress that not all galaxies fall in this spiral vs. elliptical classification. The faintest galaxies, called dwarf galaxies, typically do not fall on the Hubble sequence. Dwarf galaxies with significant amounts of gas and ongoing star formation typically have a very irregular structure, and are consequently called (dwarf) irregulars. Dwarf galaxies without gas and young stars are often very diffuse, and are called dwarf spheroidals. In addition to these dwarf galaxies, there is also a class of brighter galaxies whose morphology neither resembles a disk nor a smooth ellipsoid. These are called peculiar galaxies and include, among others, galaxies with double or multiple subcomponents linked by filamentary structure and highly-distorted galaxies with extended tails. As we will see, they are usually associated with recent mergers or tidal interactions. Although peculiar galaxies only constitute a small fraction of the entire galaxy population, their existence conveys important information about how galaxies may have changed their morphologies during their evolutionary history.

(b) Luminosity and Stellar Mass Galaxies span a wide range in luminosity. The brightest galaxies have luminosities of $\sim 10^{12}\,\mathrm{L}_\odot$, where $\mathrm{L}_\odot$ indicates the luminosity of the Sun. The exact lower limit of the luminosity distribution is less well defined, and is subject to regular changes, as fainter and fainter galaxies are constantly being discovered. In 2007 the faintest galaxy known was a newly discovered dwarf spheroidal Willman I, with a total luminosity somewhat below $1000\,\mathrm{L}_\odot$.

Obviously, the total luminosity of a galaxy is related to its total number of stars, and thus to its total stellar mass. However, the relation between luminosity and stellar mass reveals a significant amount of scatter, because different galaxies have different stellar populations. As we will see in Chapter 10, galaxies with a younger stellar population have a higher luminosity per unit stellar mass than galaxies with an older stellar population.

An important statistic of the galaxy population is its luminosity probability distribution function, also known as the luminosity function. As we will see in Chapter 2, there are many more faint galaxies than bright galaxies, so that the faint ones clearly dominate the number density. However, in terms of the contribution to the total luminosity density, neither the faintest nor the brightest galaxies dominate. Instead, it is the galaxies with a characteristic luminosity similar to that of our Milky Way that contribute most to the total luminosity density in the present-day Universe. This indicates that there is a characteristic scale in galaxy formation, which is accentuated by the fact that most galaxies that are brighter than this characteristic scale are ellipticals, while those that are fainter are mainly spirals (at the very faint end dwarf irregulars and dwarf spheroidals dominate). Understanding the physical origin of this characteristic scale has turned out to be one of the most challenging problems in contemporary galaxy formation modeling.

(c) Size and Surface Brightness As we will see in Chapter 2, galaxies do not have well defined boundaries. Consequently, several different definitions for the size of a galaxy can be found in the literature. One measure often used is the radius enclosing a certain fraction (e.g., half) of the total luminosity. In general, as one might expect, brighter galaxies are bigger. However, even for

a fixed luminosity, there is a considerable scatter in sizes, or in surface brightness, defined as the luminosity per unit area.

The size of a galaxy has an important physical meaning. In disk galaxies, which are rotation supported, the sizes are a measure of their specific angular momenta (see Chapter 11). In the case of elliptical galaxies, which are supported by random motions, the sizes are a measure of the amount of dissipation during their formation (see Chapter 13). Therefore, the observed distribution of galaxy sizes is an important constraint for galaxy formation models.

(d) Gas Mass Fraction Another useful parameter to describe galaxies is their cold gas mass fraction, defined as $f_{\rm gas} = M_{\rm cold}/[M_{\rm cold} + M_\star]$, with $M_{\rm cold}$ and $M_\star$ the masses of cold gas and stars, respectively. This ratio expresses the efficiency with which cold gas has been turned into stars. Typically, the gas mass fractions of ellipticals are negligibly small, while those of disk galaxies increase systematically with decreasing surface brightness. Indeed, the lowest surface brightness disk galaxies can have gas mass fractions in excess of 90 percent, in contrast to our Milky Way which has $f_{\rm gas} \sim 0.1$.

(e) Color Galaxies also come in different colors. The color of a galaxy reflects the ratio of its luminosity in two photometric passbands. A galaxy is said to be red if its luminosity in the redder passband is relatively high compared to that in the bluer passband. Ellipticals and dwarf spheroidals generally have redder colors than spirals and dwarf irregulars. As we will see in Chapter 10, the color of a galaxy is related to the characteristic age and metallicity of its stellar population. In general, redder galaxies are either older or more metal rich (or both). Therefore, the color of a galaxy holds important information regarding its stellar population. However, extinction by dust, either in the galaxy itself, or along the line-of-sight between the source and the observer, also tends to make a galaxy appear red. As we will see, separating age, metallicity and dust effects is one of the most daunting tasks in observational astronomy.

(f) Environment As we will see in §§2.5-2.7, galaxies are not randomly distributed throughout space, but show a variety of structures. Some galaxies are located in high density clusters containing several hundreds of galaxies, some in smaller groups containing a few to tens of galaxies, while yet others are distributed in low-density filamentary or sheet-like structures. Many of these structures are gravitationally bound, and may have played an important role in the formation and evolution of the galaxies. This is evident from the fact that elliptical galaxies seem to prefer cluster environments, whereas spiral galaxies are mainly found in relative isolation (sometimes called the field). As briefly discussed in §1.2.8 below, it is believed that this morphology-density relation reflects enhanced dynamical interaction in denser environments, although we still lack a detailed understanding of its origin.

(g) Nuclear Activity For the majority of galaxies, the observed light is consistent with what we expect from a collection of stars and gas. However, a small fraction of all galaxies, called active galaxies, show an additional non-stellar component in their spectral energy distribution. As we will see in Chapter 14, this emission originates from a small region in the centers of these galaxies, called the active galactic nucleus (AGN), and is associated with matter accretion onto a supermassive black hole. According to the relative importance of such non-stellar emission, one can separate active galaxies from normal (or non-active) galaxies.

(h) Redshift Because of the expansion of the Universe, an object that is farther away will have a larger receding velocity, and thus a larger redshift. Since the light from high-redshift galaxies was emitted when the Universe was younger, we can study galaxy evolution by observing the galaxy population at different redshifts. In fact, in a statistical sense the high-redshift galaxies are the progenitors of present-day galaxies, and any changes in the number density or intrinsic properties of galaxies with redshift give us a direct window on the formation and evolution of the

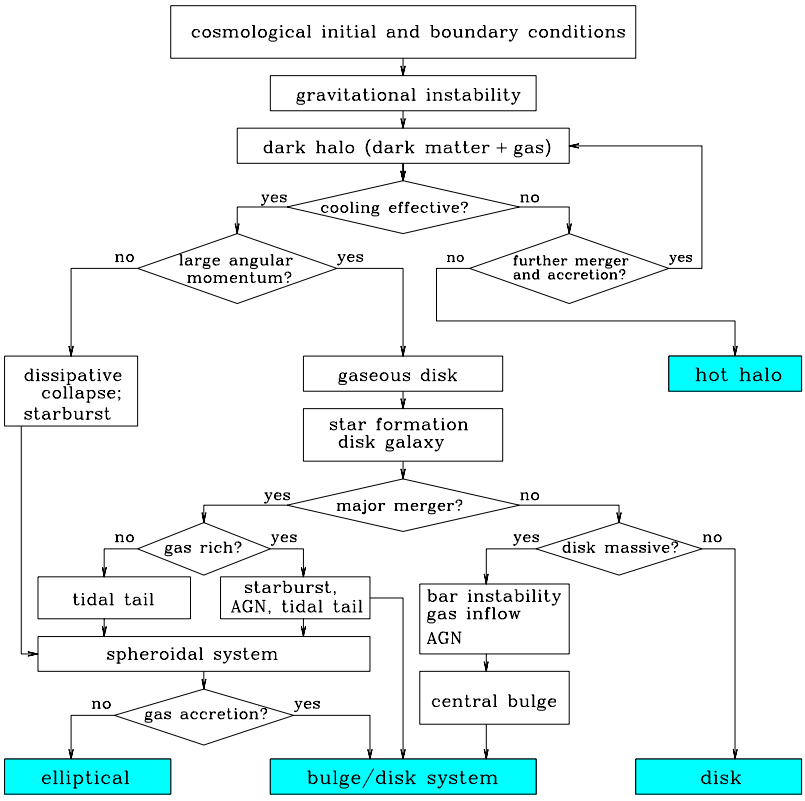

Fig. 1.1. A logic-flow chart for galaxy formation. In the standard scenario, the initial and boundary conditions for galaxy formation are set by the cosmological framework. The paths leading to the formation of various galaxies are shown along with the relevant physical processes. Note, however, that processes do not separate as neatly as this figure suggests. For example, cold gas may not have the time to settle into a gaseous disk before a major merger takes place.

galaxy population. With modern, large telescopes we can now observe galaxies out to redshifts beyond six, making possible for us to probe the galaxy population back to a time when the Universe was only about 10 percent of its current age.

1.2 Basic Elements of Galaxy Formation

Before diving into details, it is useful to have an overview of the basic theoretical framework within which our current ideas about galaxy formation and evolution have been developed. In this section we give a brief overview of the various physical processes that play a role during the formation and evolution of galaxies. The goal is to provide the reader with a picture of the relationships among the various aspects of galaxy formation to be addressed in greater detail in the chapters to come. To guide the reader, Fig. 1.1 shows a flow-chart of galaxy formation, which illustrates how the various processes to be discussed below are intertwined. It is important to stress, though, that this particular flow-chart reflects our current, undoubtedly incomplete view of galaxy formation. Future improvements in our understanding of galaxy formation and evolution may add new links to the flow-chart, or may render some of the links shown obsolete.

1.2.1 The Standard Model of Cosmology

Since galaxies are observed over cosmological length and time scales, the description of their formation and evolution must involve cosmology, the study of the properties of space-time on large scales. Modern cosmology is based upon the Cosmological Principle, the hypothesis that the Universe is spatially homogeneous and isotropic, and Einstein's theory of General Relativity, according to which the structure of space-time is determined by the mass distribution in the Universe. As we will see in Chapter 3, these two assumptions together lead to a cosmology (the standard model) that is completely specified by the curvature of the Universe, K, and the scale factor, $a(t)$, describing the change of the length scale of the Universe with time. One of the basic tasks in cosmology is to determine the value of K and the form of $a(t)$ (hence the spacetime geometry of the Universe on large scales), and to show how observables are related to physical quantities in such a universe.

Modern cosmology not only specifies the large-scale geometry of the Universe, but also has the potential to predict its thermal history and matter content. Because the Universe is expanding and filled with microwave photons at the present time, it must have been smaller, denser and hotter at earlier times. The hot and dense medium in the early Universe provides conditions under which various reactions among elementary particles, nuclei and atoms occur. Therefore, the application of particle, nuclear and atomic physics to the thermal history of the Universe in principle allows us to predict the abundances of all species of elementary particles, nuclei and atoms at different epochs. Clearly, this is an important part of the problem to be addressed in this book, because the formation of galaxies depends crucially on the matter/energy content of the Universe.

In currently popular cosmologies we usually consider a Universe consisting of three main components. In addition to the 'baryonic' matter, the protons, neutrons and electrons† that make up the *visible* Universe, astronomers have found various indications for the presence of dark matter and dark energy (see Chapter 2 for a detailed discussion of the observational evidence). Although the nature of both dark matter and dark energy is still unknown, we believe that they are responsible for more than 95 percent of the energy density of the Universe. Different cosmological models differ mainly in (i) the relative contributions of baryonic matter, dark matter, and dark energy, and (ii) the nature of dark matter and dark energy. At the time of writing, the most popular model is the so-called ΛCDM model, a flat universe in which ~ 75 percent of the energy density is due to a cosmological constant, ~ 21 percent is due to 'cold' dark matter (CDM), and the remaining 4 percent is due to the baryonic matter out of which stars and galaxies are made. Chapter 3 gives a detailed description of these various components, and describes how they influence the expansion history of the Universe.

1.2.2 Initial Conditions

If the cosmological principle held perfectly and the distribution of matter in the Universe were perfectly uniform and isotropic, there would be no structure formation. In order to explain the presence of structure, in particular galaxies, we clearly need some deviations from perfect uniformity. Unfortunately, the standard cosmology does not in itself provide us with an explanation for the origin of these perturbations. We have to go beyond it to search for an answer.

A classical, General Relativistic description of cosmology is expected to break down at very early times when the Universe is so dense that quantum effects are expected to be important. As we will see in §3.6, the standard cosmology has a number of conceptual problems when applied to the early Universe, and the solutions to these problems require an extension of the standard

† Although an electron is a lepton, and not a baryon, in cosmology it is standard practice to include electrons when talking of baryonic matter

cosmology to incorporate quantum processes. One generic consequence of such an extension is the generation of density perturbations by quantum fluctuations at early times. It is believed that these perturbations are responsible for the formation of the structures observed in today's Universe.

As we will see in §3.6, one particularly successful extension of the standard cosmology is the inflationary theory, in which the Universe is assumed to have gone through a phase of rapid, exponential expansion (called inflation) driven by the vacuum energy of one or more quantum fields. In many, but not all, inflationary models, quantum fluctuations in this vacuum energy can produce density perturbations with properties consistent with the observed large-scale structure. Inflation thus offers a promising explanation for the physical origin of the initial perturbations. Unfortunately, our understanding of the very early Universe is still far from complete, and we are currently unable to predict the initial conditions for structure formation entirely from first principles. Consequently, even this part of galaxy formation theory is still partly phenomenological: typically initial conditions are specified by a set of parameters that are constrained by observational data, such as the pattern of fluctuations in the microwave background or the present-day abundance of galaxy clusters.

1.2.3 Gravitational Instability and Structure Formation

Having specified the initial conditions and the cosmological framework, one can compute how small perturbations in the density field evolve. As we will see in Chapter 4, in an expanding universe dominated by non-relativistic matter, perturbations grow with time. This is easy to understand. A region whose initial density is slightly higher than the mean will attract its surroundings slightly more strongly than average. Consequently, over-dense regions pull matter towards them and become even more over-dense. On the other hand, under-dense regions become even more rarefied as matter flows away from them. This amplification of density perturbations is referred to as gravitational instability and plays an important role in modern theories of structure formation. In a static universe, the amplification is a run-away process, and the density contrast $\delta\rho/\rho$ grows exponentially with time. In an expanding universe, however, the cosmic expansion damps accretion flows, and the growth rate is usually a power law of time, $\delta\rho/\rho \propto t^{\alpha}$, with $\alpha > 0$. As we will see in Chapter 4, the exact rate at which the perturbations grow depends on the cosmological model.

At early times, when the perturbations are still in what we call the linear regime ($\delta\rho/\rho \ll 1$), the physical size of an overdense region increases with time due to the overall expansion of the Universe. Once the perturbation reaches overdensity $\delta\rho/\rho \sim 1$, it breaks away from the expansion and starts to collapse. This moment of 'turn-around', when the physical size of the perturbation is at its maximum, signals the transition from the mildly non-linear regime to the strongly non-linear regime.

The outcome of the subsequent non-linear, gravitational collapse depends on the matter content of the perturbation. If the perturbation consists of ordinary baryonic gas, the collapse creates strong shocks that raise the entropy of the material. If radiative cooling is inefficient, the system relaxes to hydrostatic equilibrium, with its self-gravity balanced by pressure gradients. If the perturbation consists of collisionless matter (e.g., cold dark matter), no shocks develop, but the system still relaxes to a quasi-equilibrium state with a more-or-less universal structure. This process is called violent relaxation and will be discussed in Chapter 5. Non-linear, quasi-equilibrium dark matter objects are called dark matter halos. Their predicted structure has been thoroughly explored using numerical simulations, and they play a pivotal role in modern theories of galaxy formation. Chapter 7 therefore presents a detailed discussion of the structure and formation of dark matter halos. As we shall see, halo density profiles, shapes, spins and internal substructure

all depend very weakly on mass and on cosmology, but the abundance and characteristic density of halos depend sensitively on both of these.

In cosmologies with both dark matter and baryonic matter, such as the currently favored CDM models, each initial perturbation contains baryonic gas and collisionless dark matter in roughly their universal proportions. When an object collapses, the dark matter relaxes violently to form a dark matter halo, while the gas shocks to the virial temperature, $T_{\rm vir}$ (see §8.2.3 for a definition) and may settle into hydrostatic equilibrium in the potential well of the dark matter halo if cooling is slow.

1.2.4 Gas Cooling

Cooling is a crucial ingredient of galaxy formation. Depending on temperature and density, a variety of cooling processes can affect gas. In massive halos, where the virial temperature $T_{\rm vir} \gtrsim 10^7\,{\rm K}$, gas is fully collisionally ionized and cools mainly through Bremsstrahlung emission from free electrons. In the temperature range $10^4\,{\rm K} < T_{\rm vir} < 10^6\,{\rm K}$, a number of excitation and de-excitation mechanisms can play a role. Electrons can recombine with ions, emitting a photon, or atoms (neutral or partially ionized) can be excited by a collision with another particle, thereafter decaying radiatively to the ground state. Since different atomic species have different excitation energies, the cooling rates depend strongly on the chemical composition of the gas. In halos with $T_{\rm vir} < 10^4\,{\rm K}$, gas is predicted to be almost completely neutral. This strongly suppresses the cooling processes mentioned above. However, if heavy elements and/or molecules are present, cooling is still possible through the collisional excitation/de-excitation of fine and hyperfine structure lines (for heavy elements) or rotational and/or vibrational lines (for molecules). Finally, at high redshifts ($z \gtrsim 6$), inverse Compton scattering of cosmic microwave background photons by electrons in hot halo gas can also be an effective cooling channel. Chapter 8 will discuss these cooling processes in more detail.

Except for inverse Compton scattering, all these cooling mechanisms involve two particles. Consequently, cooling is generally more effective in higher density regions. After non-linear gravitational collapse, the shocked gas in virialized halos may be dense enough for cooling to be effective. If cooling times are short, the gas never comes to hydrostatic equilibrium, but rather accretes directly onto the central protogalaxy. Even if cooling is slow enough for a hydrostatic atmosphere to develop, it may still cause the denser inner regions of the atmosphere to lose pressure support and to flow onto the central object. The net effect of cooling is thus that the baryonic material segregates from the dark matter, and accumulates as dense, cold gas in a protogalaxy at the center of the dark matter halo.

As we will see in Chapter 7, dark matter halos, as well as the baryonic material associated with them, typically have a small amount of angular momentum. If this angular momentum is conserved during cooling, the gas will spin up as it flows inwards, settling in a cold disk in centrifugal equilibrium at the center of the halo. This is the standard paradigm for the formation of disk galaxies, which we will discuss in detail in Chapter 11.

1.2.5 Star Formation

As the gas in a dark matter halo cools and flows inwards, its self-gravity will eventually dominate over the gravity of the dark matter. Thereafter it collapses under its own gravity, and in the presence of effective cooling, this collapse becomes catastrophic. Collapse increases the density and temperature of the gas, which generally reduces the cooling time more rapidly than it reduces the collapse time. During such runaway collapse the gas cloud may fragment into small, high-density cores that may eventually form stars (see Chapter 9), thus giving rise to a visible galaxy.

Unfortunately, many details of these processes are still unclear. In particular, we are still

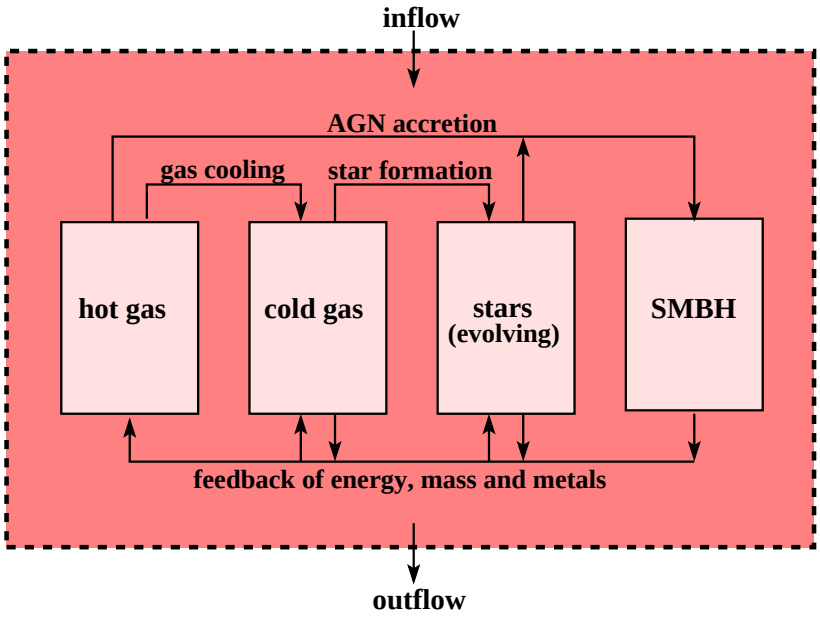

Fig. 1.2. A flow chart of the evolution of an individual galaxy. The galaxy is represented by the dashed box which contains hot gas, cold gas, stars and a supermassive black hole (SMBH). Gas cooling converts hot gas into cold gas, star formation converts cold gas into stars, and dying stars inject energy, metals and gas into the gas components. In addition, the SMBH can accrete gas (both hot and cold) as well as stars, producing AGN activity which can release vast amounts of energy which affect primarily the gaseous components of the galaxy. Note that in general the box will not be closed: gas can be added to the system through accretion from the intergalactic medium and can escape the galaxy through outflows driven by feedback from the stars and/or the SMBH. Finally, a galaxy may merge or interact with another galaxy, causing a significant boost or suppression of all these processes.

unable to predict the mass fraction of, and the time-scale for, a self-gravitating cloud to be transformed into stars. Another important and yet poorly-understood issue is concerned with the mass distribution with which stars are formed, i.e. the initial mass function (IMF). As we will see in Chapter 10, the evolution of a star, in particular its luminosity as function of time and its eventual fate, is largely determined by its mass at birth. Predictions of observable quantities for model galaxies thus require not only the birth rate of stars as a function of time, but also their IMF. In principle, it should be possible to derive the IMF from first principles, but the theory of star formation has not yet matured to this level. At present one has to assume an IMF *ad hoc* and check its validity by comparing model predictions to observations.

Based on observations, we will often distinguish two modes of star formation: quiescent star formation in rotationally supported gas disks, and starbursts. The latter are characterized by much higher star formation rates, and are typically confined to relatively small regions (often the nucleus) of galaxies. Starbursts require the accumulation of large amounts of gas in a small volume, and appear to be triggered by strong dynamical interactions or instabilities. These processes will be discussed in more detail in §1.2.8 below and in Chapter 12. At the moment, there are still many open questions related to these different modes of star formation. What fraction of stars formed in the quiescent mode? Do both modes produce stellar populations with the same IMF? How does the relative importance of starbursts scale with time? As we will see, these and related questions play an important role in contemporary models of galaxy formation.

1.2.6 Feedback Processes

When astronomers began to develop the first dynamical models for galaxy formation in a CDM dominated universe, it immediately became clear that most baryonic material is predicted to cool and form stars. This is because in these 'hierarchical' structure formation models, small dense halos form at high redshift and cooling within them is predicted to be very efficient. This disagrees badly with observations, which show that only a relatively small fraction of all baryons are in cold gas or stars (see Chapter 2). Apparently, some physical process must either prevent the gas from cooling, or reheat it after it has become cold.

Even the very first models suggested that the solution to this problem might lie in feedback from supernovae, a class of exploding stars that can produce enormous amounts of energy (see §10.5). The radiation and the blastwaves from these supernovae may heat (or reheat) surrounding gas, blowing it out of the galaxy in what is called a galactic wind. These processes are described in more detail in §8.6 and §10.5.

Another important feedback source for galaxy formation is provided by Active Galactic Nuclei (AGN), the active accretion phase of supermassive black holes (SMBH) lurking at the centers of almost all massive galaxies (see Chapter 14). This process releases vast amounts of energy – this is why AGN are bright and can be seen out to large distances, which can be tapped by surrounding gas. Although only a relatively small fraction of present-day galaxies contain an AGN, observations indicate that virtually all massive spheroids contain a nuclear SMBH (see Chapter 2). Therefore, it is believed that virtually all galaxies with a significant spheroidal component have gone through one or more AGN phases during their life.

Although it has become clear over the years that feedback processes play an important role in galaxy formation, we are still far from understanding which processes dominate, and when and how exactly they operate. Furthermore, to make accurate predictions for their effects, one also needs to know how often they occur. For supernovae this requires a prior understanding of the star formation rates and the IMF. For AGN it requires understanding how, when and where supermassive black holes form, and how they accrete mass.

It should be clear from the above discussion that galaxy formation is a subject of great complexity, involving many strongly intertwined processes. This is illustrated in Fig. 1.2, which shows the relations between the four main baryonic components of a galaxy, hot gas, cold gas, stars, and a supermassive black hole. Cooling, star formation, AGN accretion and feedback processes can all shift baryons from one of these components to another, thereby altering the efficiency of all the processes. For example, increased cooling of hot gas will produce more cold gas. This in turn will increases the star formation rate, hence the supernova rate. The additional energy injection from supernovae can reheat cold gas, thereby suppressing further star formation (negative feedback). On the other hand, supernova blastwaves may also compress the surrounding cold gas, so as to boost the star formation rate (positive feedback). Understanding these various feedback loops is one of the most important and intractable issues in contemporary models for the formation and evolution of galaxies.

1.2.7 Mergers

So far we have considered what happens to a single, isolated system of dark matter, gas and stars. However, galaxies and dark matter halos are not isolated. For example, as illustrated in Fig. 1.2, systems can accrete new material (both dark and baryonic matter) from the intergalactic medium, and can lose material through outflows driven by feedback from stars and/or AGN. In addition, two (or more) systems may merge to form a new system with very different properties from its progenitors. In the currently popular CDM cosmologies, the initial density fluctuations have larger amplitudes on smaller scales. Consequently, dark matter halos grow hierarchically,

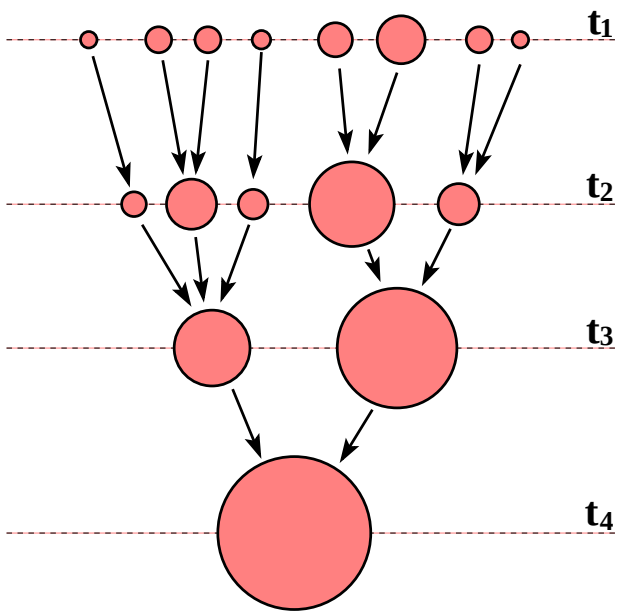

Fig. 1.3. A schematic merger tree, illustrating the merger history of a dark matter halo. It shows, at three different epochs, the progenitor halos that at time t_4 have merged to form a single halo. The size of each circle represents the mass of the halo. Merger histories of dark matter halos play an important role in hierarchical theories of galaxy formation.

in the sense that larger halos are formed by the coalescence (merging) of smaller progenitors. Such a formation process is usually called a hierarchical or 'bottom-up' scenario.

The formation history of a dark matter halo can be described by a 'merger tree' that traces all its progenitors, as illustrated in Fig. 1.3. Such merger trees play an important role in modern galaxy formation theory. Note, however, that illustrations such as Fig. 1.3 can be misleading. In CDM models part of the growth of a massive halo is due to merging with a large number of much smaller halos, and to a good approximation, such mergers can be thought of as smooth accretion. When two similar mass dark matter halos merge, violent relaxation rapidly transforms the orbital energy of the progenitors into the internal binding energy of the quasi-equilibrium remnant. Any hot gas associated with the progenitors is shock-heated during the merger and settles back into hydrostatic equilibrium in the new halo. If the progenitor halos contained central galaxies, the galaxies also merge as part of the violent relaxation process, producing a new central galaxy in the final system. Such a merger may be accompanied by strong star formation or AGN activity if the merging galaxies contained significant amounts of cold gas. If two merging halos have very different mass, the dynamical processes are less violent. The smaller system orbits within the main halo for an extended period of time during which two processes compete to determine its eventual fate. Dynamical friction transfers energy from its orbit to the main halo, causing it to spiral inwards, while tidal effects remove mass from its outer regions and may eventually dissolve it completely (see Chapter 12). Dynamical friction is more effective for more massive satellites, but if the mass ratio of the initial halos is large enough, the smaller object (and any galaxy associated with it) can maintain its identity for a long time. This is the process for the build-up of clusters of galaxies: a cluster may be considered as a massive dark matter halo hosting a relatively massive galaxy near its center and many satellites that have not yet dissolved or merged with the central galaxy.

As we will see in Chapters 12 and 13, numerical simulations show that the merger of two

galaxies of roughly equal mass produces an object reminiscent of an elliptical galaxy, and the result is largely independent of whether the progenitors are spirals or ellipticals. Indeed, current hierarchical models of galaxy formation assume that most, if not all, elliptical galaxies are merger remnants. If gas cools onto this merger remnant with significant angular momentum, a new disk may form, producing a disk-bulge system like that in an early-type spiral galaxy.

It should be obvious from the above discussion that mergers play a crucial role in galaxy formation. Detailed descriptions of halo mergers and galaxy mergers are presented in Chapter 7 and Chapter 12, respectively.

1.2.8 Dynamical Evolution

When satellite galaxies orbit within dark matter halos, they experience tidal forces due to the central galaxy, due to other satellite galaxies, and due to the potential of the halo itself. These tidal interactions can remove dark matter, gas and stars from the galaxy, a process called tidal stripping (see §12.2), and may also perturb its structure. In addition, if the halo contains a hot gas component, any gas associated with the satellite galaxy will experience a drag force due to the relative motion of the two fluids. If the drag force exceeds the restoring force due to the satellite's own gravity, its gas will be ablated, a process called ram-pressure stripping. These dynamical processes are thought to play an important role in driving galaxy evolution within clusters and groups of galaxies. In particular, they are thought to be partially responsible for the observed environmental dependence of galaxy morphology (see Chapter 15).

Internal dynamical effects can also reshape galaxies. For example, a galaxy may form in a configuration which becomes unstable at some later time. Large-scale instabilities may then redistribute mass and angular momentum within the galaxy, thereby changing its morphology. A well-known and important example is the bar-instability within disk galaxies. As we shall see in §11.5, a thin disk with too high a surface density is susceptible to a non-axisymmetric instability, which produces a bar-like structure similar to that seen in barred spiral galaxies. These bars may then buckle out of the disk to produce a central ellipsoidal component, a so-called 'pseudo-bulge'. Instabilities may also be triggered in otherwise stable galaxies by interactions. Thus, an important question is whether the sizes and morphologies of galaxies were set at formation, or are the result of later dynamical process ('secular evolution', as it is termed). Bulges are particularly interesting in this context. They may be a remnant of the first stage of galaxy formation, or as mentioned in §1.2.7, may reflect an early merger which has grown a new disk, or may result from buckling of a bar. It is likely that all these processes are important for at least some bulges.

1.2.9 Chemical Evolution

In astronomy, all chemical elements heavier than helium are collectively termed 'metals'. The mass fraction of a baryonic component (e.g. hot gas, cold gas, stars) in metals is then referred to as its metallicity. As we will see in §3.4, the nuclear reactions during the first three minutes of the Universe (the epoch of primordial nucleosynthesis) produced primarily hydrogen ($\sim 75\%$) and helium ($\sim 25\%$), with a very small admixture of metals dominated by lithium. All other metals in the Universe were formed at later times as a consequence of nuclear reactions in stars. When stars expel mass in stellar winds, or in supernova explosions, they enrich the interstellar medium (ISM) with newly synthesized metals.

Evolution of the chemical composition of the gas and stars in galaxies is important for several reasons. First of all, the luminosity and color of a stellar population depend not only on its age and IMF, but also on the metallicity of the stars (see Chapter 10). Secondly, the cooling efficiency of gas depends strongly on its metallicity, in the sense that more metal-enriched gas cools faster (see §8.1). Thirdly, small particles of heavy elements known as dust grains, which are mixed with

the interstellar gas in galaxies, can absorb significant amounts of the starlight and re-radiate it in infrared wavelengths. Depending on the amount of the dust in the ISM, which scales roughly linearly with its metallicity (see §10.3.7), this interstellar extinction can significantly reduce the brightness of a galaxy.

As we will see in Chapter 10, the mass and detailed chemical composition of the material ejected by a stellar population as it evolves depend both on the IMF and on its initial metallicity. In principle, observations of the metallicity and abundance ratios of a galaxy can therefore be used to constrain its star formation history and IMF. In practice, however, the interpretation of the observations is complicated by the fact that galaxies can accrete new material of different metallicity, that feedback processes can blow out gas, perhaps preferentially metals, and that mergers can mix the chemical compositions of different systems.

1.2.10 Stellar Population Synthesis

The light we receive from a given galaxy is emitted by a large number of stars that may have different masses, ages, and metallicities. In order to interpret the observed spectral energy distribution, we need to predict how each of these stars contributes to the total spectrum. Unlike many of the ingredients in galaxy formation, the theory of stellar evolution, to be discussed in Chapter 10, is reasonably well understood. This allows us to compute not only the evolution of the luminosity, color and spectrum of a star of given initial mass and chemical composition, but also the rates at which it ejects mass, energy and metals into the interstellar medium. If we know the star formation history (i.e., the star formation rate as a function of time) and IMF of a galaxy, we can then synthesize its spectrum at any given time by adding together the spectra of all the stars, after evolving each to the time under consideration. In addition, this also yields the rates at which mass, energy and metals are ejected into the interstellar medium, providing important ingredients for modeling the chemical evolution of galaxies.

Most of the energy of a stellar population is emitted in the optical, or, if the stellar population is very young ($\lesssim 10\,\mathrm{Myr}$), in the ultraviolet (see §10.3). However, if the galaxy contains a lot of dust, a significant fraction of this optical and UV light may get absorbed and re-emitted in the infrared. Unfortunately, predicting the final emergent spectrum is extremely complicated. Not only does it depend on the amount of the radiation absorbed, it also depends strongly on the properties of the dust, such as its geometry, its chemical composition, and (the distribution of) the sizes of the dust grains (see §10.3.7).

Finally, to complete the spectral energy distribution emitted by a galaxy, we also need to add the contribution from a possible AGN. Chapter 14 discusses various emission mechanisms associated with accreting SMBHs. Unfortunately, as we will see, we are still far from being able to predict the detailed spectra for AGN.

1.2.11 The Intergalactic Medium

The intergalactic medium (IGM) is the baryonic material lying between galaxies. This is and has always been the dominant baryonic component of the Universe and it is the material from which galaxies form. Detailed studies of the IGM can therefore give insight into the properties of the pregalactic matter before it condensed into galaxies. As illustrated in Fig. 1.2, galaxies do not evolve as closed boxes, but can affect the properties of the IGM through exchanges of mass, energy and heavy elements. The study of the IGM is thus an integral part of understanding how galaxies form and evolve. As we will see in Chapter 16, the properties of the IGM can be probed most effectively through the absorption it produces in the spectra of distant quasars (a certain class of active galaxies, see Chapter 14). Since quasars are now observed out to redshifts beyond

6, their absorption line spectra can be used to study the properties of the IGM back to a time when the Universe was only a few percent of its present age.

1.3 Time Scales

As discussed above, and as illustrated in Fig. 1.1, the formation of an individual galaxy in the standard, hierarchical formation scenario involves the following processes: the collapse and virialization of dark matter halos, the cooling and condensation of gas within the halo, and the conversion of cold gas into stars and a central supermassive black hole. Evolving stars and active AGN eject energy, mass and heavy elements into the interstellar medium, thereby determining its structure and chemical composition and perhaps driving winds into the intergalactic medium. Finally, galaxies can merge and interact, re-shaping their morphology and triggering further starbursts and AGN activity. In general, the properties of galaxies are determined by the competition among all these processes, and a simple way to characterize the relative importance of these processes is to use the time scales associated with them. Here we give a brief summary of the most important time scales in this context.

- **Hubble time:** This is an estimate of the time scale on which the Universe as a whole evolves. It is defined as the inverse of the Hubble constant (see §3.2), which specifies the current cosmic expansion rate. It would be equal to the time since the Big Bang if the Universe had always expanded at its current rate. Roughly speaking, this is the timescale on which substantial evolution of the galaxy population is expected.
- **Dynamical time:** This is the time required to orbit across an equilibrium dynamical system. For a system with mass M and radius R, we define it as $t_{\rm dyn} = \sqrt{3\pi/16G\overline{\rho}}$, where $\overline{\rho} = 3M/4\pi R^3$. This is related to the free-fall time, defined as the time required for a uniform, pressure-free sphere to collapse to a point, as $t_{\rm ff} = t_{\rm dyn}/\sqrt{2}$.
- **Cooling time:** This time scale is the ratio between the thermal energy content and the energy loss rate (through radiative or conductive cooling) for a gas component.
- **Star-formation time:** This time scale is the ratio of the cold gas content of a galaxy to its star-formation rate. It is thus an indication of how long it would take for the galaxy to run out of gas if the fuel for star formation is not replenished.
- **Chemical enrichment time:** This is a measure for the time scale on which the gas is enriched in heavy elements. This enrichment time is generally different for different elements, depending on the lifetimes of the stars responsible for the bulk of the production of each element (see §10.1).
- **Merging time:** This is the typical time that a halo or galaxy must wait before experiencing a merger with an object of similar mass, and is directly related to the major merger frequency.
- **Dynamical friction time:** This is the time scale on which a satellite object in a large halo loses its orbital energy and spirals to the center. As we will see in §12.3, this time scale is proportional to $M_{\rm sat}/M_{\rm main}$, where $M_{\rm sat}$ is the mass of the satellite object and $M_{\rm main}$ is that of the main halo. Thus, more massive galaxies will merge with the central galaxy in a halo more quickly than smaller ones.

These time scales can provide guidelines for incorporating the underlying physical processes in models of galaxy formation and evolution, as we describe in later chapters. In particular, comparing time scales can give useful insights. As an illustration, consider the following examples:

- Processes whose time scale is longer than the Hubble time can usually be ignored. For example, satellite galaxies with mass less than a few percent of their parent halo normally have dynamical friction times exceeding the Hubble time (see §12.3). Consequently, their orbits do

not decay significantly. This explains why clusters of galaxies have so many 'satellite' galaxies – the main halos are so much more massive than a typical galaxy that dynamical friction is ineffective.

- If the cooling time is longer than the dynamical time, hot gas will typically be in hydrostatic equilibrium. In the opposite case, however, the gas cools rapidly, losing pressure support, and collapsing to the halo center on a free-fall time without establishing any hydrostatic equilibrium.
- If the star formation time is comparable to the dynamical time, gas will turn into stars during its initial collapse, a situation which may lead to the formation of something resembling an elliptical galaxy. On the other hand, if the star formation time is much longer than the cooling and dynamical times, the gas will settle into a centrifugally supported disk before forming stars, thus producing a disk galaxy (see §1.4.5).
- If the relevant chemical evolution time is longer than the star formation time, little metal enrichment will occur during star formation and all stars will end up with the same, initial metallicity. In the opposite case, the star-forming gas is continuously enriched, so that stars formed at different times will have different metallicities and abundance patterns (see §10.4).

So far we have avoided one obvious question, namely, what is the time scale for galaxy formation itself? Unfortunately, there is no single useful definition for such a time scale. Galaxy formation is a process, not an event, and as we have seen, this process is an amalgam of many different elements, each with its own time scale. If, for example, we are concerned with its stellar population, we might define the formation time of a galaxy as the epoch when a fixed fraction (e.g. 1% or 50%) of its stars had formed. If, on the other hand, we are concerned with its structure, we might want to define the galaxy's formation time as the epoch when a fixed fraction (e.g. 50% or 90%) of its mass was first assembled into a single object. These two 'formation' times can differ greatly for a given galaxy, and even their ordering can change from one galaxy to another. Thus it is important to be precise about definition when talking about the formation times of galaxies.

1.4 A Brief History of Galaxy Formation

The picture of galaxy formation sketched above is largely based on the hierarchical cold dark matter model for structure formation, which has been the standard paradigm since the beginning of the 1980s. In the following, we give an historical overview of the development of ideas and concepts about galaxy formation up to the present time. This is not intended as a complete historical account, but rather as a summary for young researchers of how our current ideas about galaxy formation were developed. Readers interested in a more extensive historical review can find some relevant material in the book 'The Cosmic Century: A History of Astrophysics and Cosmology' by Malcolm Longair.

1.4.1 Galaxies as Extragalactic Objects

By the end of the 19th century, astronomers had discovered a large number of astronomical objects that differ from stars in that they are fuzzy rather than point-like. These objects were collectively referred to as 'nebulae'. During the period 1771 to 1784 the French astronomer Charles Messier cataloged more than 100 of these objects in order to avoid confusing them with the comets he was searching for. Today the Messier numbers are still used to designate a number of bright galaxies. For example, the Andromeda galaxy is also known as M31, because it is the 31st nebula in Messier's catalog. A more systematic search for nebulae was carried

out by the Herschels, and in 1864 John Herschel published his *General Catalogue of Galaxies* which contains 5079 nebular objects. In 1888, Dreyer published an expanded version as his *New General Catalogue of Nebulae and Clusters of Stars*. Together with its two supplementary *Index Catalogues*, Dreyer's catalogue contained about 15,000 objects. Today, NGC and IC numbers are still widely used to refer to galaxies.

For many years after their discovery, the nature of the nebular objects was controversial. There were two competing ideas, one assumed that all nebulae are objects within our Milky Way, the other that some might be extragalactic objects, individual 'island universes' like the Milky Way. In 1920 the National Academy of Sciences in Washington invited two leading astronomers, Harlow Shapley and Heber Curtis, to debate this issue, an event which has passed into astronomical folklore as 'The Great Debate'. The controversy remained unresolved until 1925, when Edwin Hubble used distances estimated from Cepheid variables to demonstrate conclusively that some nebulae are extragalactic, individual galaxies comparable to our Milky Way in size and luminosity. Hubble's discovery marked the beginning of extragalactic astronomy. During the 1930s, high-quality photographic images of galaxies enabled him to classify galaxies into a broad sequence according to their morphology. Today Hubble's sequence is still widely adopted to classify galaxies.

Since Hubble's time, astronomers have made tremendous progress in systematically searching the skies for galaxies. At present deep CCD imaging and high-quality spectroscopy are available for about a million galaxies.

1.4.2 Cosmology

Only four years after his discovery that galaxies truly are extragalactic, Hubble made his second fundamental breakthrough: he showed that the recession velocities of galaxies are linearly related to their distances (Hubble, 1929, see also Hubble & Humason 1931), thus demonstrating that our Universe is expanding. This is undoubtedly the greatest single discovery in the history of cosmology. It revolutionized our picture of the Universe we live in.

The construction of mathematical models for the Universe actually started somewhat earlier. As soon as Albert Einstein completed his theory of General Relativity in 1916, it was realized that this theory allowed, for the first time, the construction of self-consistent models for the Universe as a whole. Einstein himself was among the first to explore such solutions of his field equations. To his dismay, he found that all solutions require the Universe either to expand or to contract, in contrast with his belief at that time that the Universe should be static. In order to obtain a static solution, he introduced a cosmological constant into his field equations. This additional constant of gravity can oppose the standard gravitational attraction and so make possible a static (though unstable) solution. In 1922 Alexander Friedmann published two papers exploring both static and expanding solutions. These models are today known as Friedmann models, although this work drew little attention until Georges Lemaitre independently rediscovered the same solutions in 1927.

An expanding universe is a natural consequence of General Relativity, so it is not surprising that Einstein considered his introduction of a cosmological constant as 'the biggest blunder of my life' once he learned of Hubble's discovery. History has many ironies, however. As we will see later, the cosmological constant is now back with us. In 1998 two teams independently used the distance-redshift relation of Type Ia supernovae to show that the expansion of the Universe is accelerating at the present time. Within General Relativity this requires an additional mass/energy component with properties very similar to those of Einstein's cosmological constant. Rather than just counterbalancing the attractive effects of 'normal' gravity, the cosmological constant today overwhelms them to drive an ever more rapid expansion.

Since the Universe is expanding, it must have been denser and perhaps also hotter at earlier

times. In the late 1940's this prompted George Gamow to suggest that the chemical elements may have been created by thermonuclear reactions in the early Universe, a process known as primordial nucleosynthesis. Gamow's model was not considered a success, because it was unable to explain the existence of elements heavier than lithium due to the lack of stable elements with atomic mass numbers 5 and 8. We now know that this was not a failure at all; all heavier elements are a result of nucleosynthesis within stars, as first shown convincingly by Fred Hoyle and collaborators in the 1950s. For Gamow's model to be correct, the Universe would have to be hot as well as dense at early times, and Gamow realized that the residual heat should still be visible in today's Universe as a background of thermal radiation with a temperature of a few degrees Kelvin, thus with a peak at microwave wavelengths. This was a remarkable prediction of the cosmic microwave background radiation (CMB), which was finally discovered in 1965. The thermal history suggested by Gamow, in which the Universe expands from a dense and hot initial state, was derisively referred to as the Hot Big Bang by Fred Hoyle, who preferred an unchanging Steady State Cosmology. Hoyle's cosmological theory was wrong, but his name for the correct model has stuck.

The Hot Big Bang model developed gradually during the 1950s and 1960s. By 1964, it had been noticed that the abundance of helium by mass is everywhere about one third that of hydrogen, a result which is difficult to explain by nucleosynthesis in stars. In 1964, Hoyle and Tayler published calculations that demonstrated how the observed helium abundance could emerge from the Hot Big Bang. Three years later, Wagoner et al. (1967) made detailed calculations of a complete network of nuclear reactions, confirming the earlier result and suggesting that the abundances of other light isotopes, such as helium-3, deuterium and lithium could also be explained by primordial nucleosynthesis. This success provided strong support for the Hot Big Bang. The 1965 discovery of the cosmic microwave background showed it to be isotropic and to have a temperature (2.7K) exactly in the range expected in the Hot Big Bang model (Penzias & Wilson, 1965; Dicke et al., 1965). This firmly established the Hot Big Bang as the standard model of cosmology, a status which it has kept up to the present day. Although there have been changes over the years, these have affected only the exact matter/energy content of the model and the exact values of its characteristic parameters.

Despite its success, during the 1960s and 1970s it was realized that the standard cosmology had several serious shortcomings. Its structure implies that the different parts of the Universe we see today were never in causal contact at early times (e.g., Misner, 1968). How then can these regions have contrived to be so similar, as required by the isotropy of the CMB? A second shortcoming is connected with the spatial flatness of the Universe (e.g. Dicke & Peebles, 1979). It was known by the 1960s that the matter density in the Universe is not very different from the critical density for closure, i.e., the density for which the spatial geometry of the Universe is flat. However, in the standard model any tiny deviation from flatness in the early Universe is amplified enormously by later evolution. Thus, extreme fine tuning of the initial curvature is required to explain why so little curvature is observed today. A closely related formulation is to ask how our Universe has managed to survive and to evolve for billions of years, when the timescales of all physical processes in its earliest phases were measured in tiny fractions of a nanosecond. The standard cosmology provides no explanations for these puzzles.

A conceptual breakthrough came in 1981 when Alan Guth proposed that the Universe may have gone through an early period of exponential expansion (inflation) driven by the vacuum energy of some quantum field. His original model had some problems and was revised in 1982 by Linde and by Albrecht & Steinhardt. In this scenario, the different parts of the Universe we see today were indeed in causal contact *before* inflation took place, thereby allowing physical processes to establish homogeneity and isotropy. Inflation also solves the flatness/timescale problem, because the Universe expanded so much during inflation that its curvature radius grew

to be much larger than the presently observable Universe. Thus, a generic prediction of the inflation scenario is that today's Universe should appear flat.

1.4.3 Structure Formation

(a) Gravitational Instability In the standard model of cosmology, structures form from small initial perturbations in an otherwise homogeneous and isotropic universe. The idea that structures can form via gravitational instability in this way originates from Jeans (1902), who showed that the stability of a perturbation depends on the competition between gravity and pressure. Density perturbations grow only if they are larger (heavier) than a characteristic length (mass) scale [now referred to as the Jeans' length (mass)] beyond which gravity is able to overcome the pressure gradients. The application of this Jeans criterion to an expanding background was worked out by, among others, Gamow & Teller (1939) and Lifshitz (1946), with the result that perturbation growth is power-law in time, rather than exponential as for a static background.

(b) Initial Perturbations Most of the early models of structure formation assumed the Universe to contain two energy components, ordinary baryonic matter and radiation (CMB photons and relativistic neutrinos). In the absence of any theory for the origin of perturbations, two distinct models were considered, usually referred to as adiabatic and isothermal initial conditions. In adiabatic initial conditions all matter and radiation fields are perturbed in the same way, so that the total density (or local curvature) varies, but the ratio of photons to baryons, for example, is spatially invariant. Isothermal initial conditions, on the other hand, correspond to initial perturbations in the ratio of components, but with no associated spatial variation in the total density or curvature.†

In the adiabatic case, the perturbations can be considered as applying to a single fluid with a constant specific entropy as long as the radiation and matter remain tightly coupled. At such times, the Jeans' mass is very large and small-scale perturbations execute acoustic oscillations driven by the pressure gradients associated with the density fluctuations. Silk (1968) showed that towards the end of recombination, as radiation decouples from matter, small-scale oscillations are damped by photon diffusion, a process now called Silk damping. Depending on the matter density and the expansion rate of the Universe, the characteristic scale of Silk damping falls in the range of $10^{12} - 10^{14}\,\mathrm{M}_{\odot}$. After radiation/matter decoupling the Jeans' mass drops precipitously to $\simeq 10^{6}\,\mathrm{M}_{\odot}$ and perturbations above this mass scale can start to grow,‡ but there are no perturbations left on the scale of galaxies at this time. Consequently, galaxies must form 'top-down', via the collapse and fragmentation of perturbations larger than the damping scale, an idea championed by Zel'dovich and colleagues.

In the case of isothermal initial conditions, the spatial variation in the ratio of baryons to photons remains fixed before recombination because of the tight coupling between the two fluids. The pressure is spatially uniform, so that there is no acoustic oscillation, and perturbations are not influenced by Silk damping. If the initial perturbations include small-scale structure, this survives until after the recombination epoch, when baryon fluctuations are no longer supported by photon pressure and so can collapse. Structure can then form 'bottom-up' through hierarchical clustering. This scenario of structure formation was originally proposed by Peebles (1965).

By the beginning of the 1970s, the linear evolution of both adiabatic and isothermal perturbations had been worked out in great detail (e.g., Lifshitz, 1946; Silk, 1968; Peebles & Yu, 1970; Sato, 1971; Weinberg, 1971). At that time, it was generally accepted that observed structures must have formed from finite amplitude perturbations which were somehow part of the initial

† Note that the nomenclature 'isothermal', which is largely historical, is somewhat confusing; the term 'isocurvature' would be more appropriate.

‡ Actually, as we will see in Chapter 4, depending on the gauge adopted, perturbations can also grow before they enter the horizon.

conditions set up at the Big Bang. Harrison (1970) and Zeldovich (1972) independently argued that only one scaling of the amplitude of initial fluctuations with their wavelength could be consistent with the formation of galaxies from fluctuations imposed at very early times. Their suggestion, now known as the Harrison-Zel'dovich initial fluctuation spectrum, has the property that structure on every scale has the same dimensionless amplitude, corresponding to fluctuations in the equivalent Newtonian gravitational potential, $\delta\Phi/c^2 \sim 10^{-4}$.

In the early 1980s, immediately after the inflationary scenario was proposed, a number of authors realized almost simultaneously that quantum fluctuations of the scalar field (called the inflaton) that drives inflation can generate density perturbations with a spectrum that is close to the Harrison-Zeldovich form (Hawking, 1982; Guth & Pi, 1982; Starobinsky, 1982; Bardeen et al., 1983). In the simplest models, inflation also predicts that the perturbations are adiabatic and that the initial density field is Gaussian. When parameters take their natural values, however, these models generically predict fluctuation amplitudes that are much too large, of order unity. This apparent fine-tuning problem is still unresolved.

In 1992 anisotropy in the cosmic microwave background was detected convincingly for the first time by the Cosmic Background Explorer (COBE) (Smoot et al., 1992). These anisotropies provide an image of the structure present at the time of radiation/matter decoupling, $\sim$400,000 years after the Big Bang. The resolved structures are all of very low amplitude and so can be used to probe the properties of the initial density perturbations. In agreement with the inflationary paradigm, the COBE maps were consistent with Gaussian initial perturbations with the Harrison-Zel'dovich spectrum. The fluctuation amplitudes are comparable to those inferred by Harrison and Zel'dovich. The COBE results have since been confirmed and dramatically refined by subsequent observations, most notably by the Wilkinson Microwave Anisotropy Probe (WMAP) (Bennett et al., 2003; Hinshaw et al., 2007). The agreement with simple inflationary predictions remains excellent.

(c) Non-Linear Evolution In order to connect the initial perturbations to the non-linear structures we see today, one has to understand the outcome of non-linear evolution. In 1970 Zel'dovich published an analytical approximation (now referred to as the Zel'dovich approximation) which describes the initial non-linear collapse of a coherent perturbation of the cosmic density field. This model shows that the collapse generically occurs first along one direction, producing a sheet-like structure, often referred to as a 'pancake'. Zeldovich imagined further evolution to take place via fragmentation of such pancakes. At about the same time, Gunn & Gott (1972) developed a simple spherically symmetric model to describe the growth, turn-around (from the general expansion), collapse and virialization of a perturbation. In particular, they showed that dissipationless collapse results in a quasi-equilibrium system with a characteristic radius that is about half the radius at turn-around. Although the non-linear collapse described by the Zel'dovich approximation is more realistic, since it does not assume any symmetry, the spherical collapse model of Gunn & Gott has the virtue that it links the initial perturbation directly to the final quasi-equilibrium state. By applying this model to a Gaussian initial density field, Press & Schechter (1974) developed a very useful formalism (now referred to as Press-Schechter theory) that allows one to estimate the mass function of collapsed objects (i.e., their abundance as a function of mass) produced by hierarchical clustering.

Hoyle (1949) was the first to suggest that perturbations (and the associated proto-galaxies) might gain angular momentum through the tidal torques from their neighbors. A linear perturbation analysis of this process was first carried out correctly and in full generality by Doroshkevich (1970), and was later tested with the help of numerical simulations (Peebles, 1971; Efstathiou & Jones, 1979). The study of Efstathiou and Jones showed that clumps formed through gravitational collapse in a cosmological context typically acquire about 15% of the angular momentum needed for full rotational support. Better simulations in more recent years have shown that the

correct value is closer to 10%. In the case of 'top-down' models, it was suggested that objects could acquire angular momentum not only through gravitational torques as pancakes fragment, but also via oblique shocks generated by their collapse (Doroshkevich, 1973).

1.4.4 The Emergence of the Cold Dark Matter Paradigm

The first evidence that the Universe may contain dark matter (undetected through electromagnetic emission or absorption) can be traced back to 1933, when Zwicky studied the velocities of galaxies in the Coma cluster and concluded that the total mass required to hold the cluster together is about 400 times larger than the luminous mass in stars. In 1937 he reinforced this analysis and noted that galaxies associated with such large amounts of mass should be detectable as gravitational lenses producing multiple images of background galaxies. These conclusions were substantially correct, but remarkably it took more than 40 years for the existence of dark matter to be generally accepted. The tide turned in the mid-1970s with papers by Ostriker et al. (1974) and Einasto et al. (1974) extending Zwicky's analysis and noting that massive halos are required around our Milky Way and other nearby galaxies in order to explain the motions of their satellites. These arguments were supported by continually improving 21cm and optical measurements of spiral galaxy rotation curves which showed no sign of the fall-off at large radius expected if the visible stars and gas were the only mass in the system (Roberts & Rots, 1973; Rubin et al., 1978, 1980). During the same period, numerous suggestions were made regarding the possible nature of this dark matter component, ranging from baryonic objects such as brown-dwarfs, white dwarfs and black holes (e.g., White & Rees, 1978; Carr et al., 1984), to more exotic, elementary particles such as massive neutrinos (Gershtein & Zel'Dovich, 1966; Cowsik & McClelland, 1972).

The suggestion that neutrinos might be the unseen mass was partly motivated by particle physics. In the 1960s and 1970s, it was noticed that Grand Unified Theories (GUTs) permit the existence of massive neutrinos, and various attempts to measure neutrino masses in laboratory experiments were initiated. In the late 1970s, Lyubimov et al. (1980) and Reines et al. (1980) announced the detection of a mass for the electron neutrino at a level of cosmological interest (about 30 eV). Although the results were not conclusive, they caused a surge in studies investigating neutrinos as dark matter candidates (e.g., Bond et al., 1980; Sato & Takahara, 1980; Schramm & Steigman, 1981; Klinkhamer & Norman, 1981), and structure formation in a neutrino-dominated universe was soon worked out in detail. Since neutrinos decouple from other matter and radiation fields while still relativistic, their abundance is very similar to that of CMB photons. Thus, they must have become nonrelativistic at the time the Universe became matter-dominated, implying thermal motions sufficient to smooth out all structure on scales smaller than a few tens of Mpc. The first non-linear structures are then Zel'dovich pancakes of this scale, which must fragment to make smaller structures such as galaxies. Such a picture conflicts directly with observation, however. An argument by Tremaine & Gunn (1979), based on the Pauli exclusion principle, showed that individual galaxy halos could not be made of neutrinos with masses as small as 30 eV, and simulations of structure formation in neutrino-dominated universes by White et al. (1984) demonstrated that they could not produce galaxies without at the same time producing much stronger galaxy clustering than is observed. Together with the failure to confirm the claimed neutrino mass measurements, these problems caused a precipitous decline in interest in neutrino dark matter by the end of the 1980s.

In the early 1980s, alternative models were suggested, in which dark matter is a different kind of weakly interacting massive particle. There were several motivations for this. The amount of baryonic matter allowed by cosmic nucleosynthesis calculations is far too little to provide the flat universe preferred by inflationary models, suggesting that non-baryonic dark matter may be present. In addition, strengthening upper limits on temperature anisotropies in the CMB made it

increasingly difficult to construct self-consistent, purely baryonic models for structure formation; there is simply not enough time between the recombination epoch and the present day to grow the structures we see in the nearby Universe from those present in the high-redshift photon-baryon fluid. Finally, by the early 1980s, particle physics models based on the idea of supersymmetry had provided a plethora of dark matter candidates, such as neutralinos, photinos and gravitinos, that could dominate the mass density of the Universe. Because of their much larger mass, such particles would initially have much smaller velocities than a 30 eV neutrino, and so they were generically referred to as Warm or Cold Dark Matter (WDM or CDM, the former corresponding to a particle mass of order 1 keV, the latter to much more massive particles) in contrast to neutrino-like Hot Dark Matter (HDM). The shortcomings of HDM motivated consideration of a variety of such scenarios (e.g., Peebles, 1982; Blumenthal et al., 1982; Bond et al., 1982; Bond & Szalay, 1983).

Lower thermal velocities result in the survival of fluctuations of galactic scale (for WDM and CDM) or below (for CDM). The particles decouple from the radiation field long before recombination, so perturbations in their density can grow at early times to be substantially larger than the fluctuations visible in the CMB. After the baryons decouple from the radiation, they quickly fall in these dark matter potential wells, causing structure formation to occur sufficiently fast to be consistent with observed structure in today's Universe. Davis et al. (1985) used simulations of the CDM model to show that it could provide a good match to the observed clustering of galaxies provided either the mass density of dark matter is well below the critical value, or (their preferred model) that galaxies are biased tracers of the CDM density field, as expected if they form at the centers of the deepest dark matter potential wells (e.g. Kaiser, 1984). By the mid 1980s, the 'standard' CDM model, in which dark matter provides the critical density, Hubble's constant has a value $\sim 50\,\mathrm{km\,s^{-1}Mpc^{-1}}$, and the initial density field was Gaussian with a Harrison-Zel'dovich spectrum, had established itself as the 'best bet' model for structure formation.

In the early 1990s, measurements of galaxy clustering, notably from the APM galaxy survey (Maddox et al., 1990a; Efstathiou et al., 1990) showed that the standard CDM model predicts less clustering on large scales than is observed. Several alternatives were proposed to remedy this. One was a mixed dark matter (MDM) model, in which the universe is flat, with $\sim 30\%$ of the cosmic mass density in HDM and $\sim 70\%$ in CDM and baryons. Another flat model assumed all dark matter to be CDM, but adopted an enhanced radiation background in relativistic neutrinos (τCDM). A third possibility was an open model, in which today's Universe is dominated by CDM and baryons, but has only about 30% of the critical density (OCDM). A final model assumed the same amounts of CDM and baryons as OCDM but added a cosmological constant in order to make the universe flat (ΛCDM).

Although all these models match observed galaxy clustering on large scales, it was soon realized that galaxy formation occurs too late in the MDM and τCDM models, and that the open model has problems in matching the perturbation amplitudes measured by COBE. ΛCDM then became the default 'concordance' model, although it was not generally accepted until Garnavich et al. (1998) and Perlmutter et al. (1999) used the distance-redshift relation of Type Ia supernovae to show that the cosmic expansion is accelerating, and measurements of small-scale CMB fluctuations showed that our Universe is flat (de Bernardis et al., 2000). It seems that the present-day Universe is dominated by a dark energy component with properties very similar to those of Einstein's cosmological constant.

At the beginning of this century, a number of ground-based and balloon-borne experiments measured CMB anisotropies, notably Boomerang (de Bernardis et al., 2000), MAXIMA (Hanany et al., 2000), DASI (Halverson et al., 2002) and CBI (Sievers et al., 2003). They successfully detected features, known as acoustic peaks, in the CMB power spectrum, and showed their wavelengths and amplitudes to be in perfect agreement with expectations for a ΛCDM cosmology. In 2003, the first year data from WMAP not only confirmed these results, but also allowed much

more precise determinations of cosmological parameters. The values obtained were in remarkably good agreement with independent measurements; the baryon density matched that estimated from cosmic nucleosynthesis, the Hubble constant matched that found by direct measurement, the dark-energy density matched that inferred from Type Ia supernovae, and the implied large-scale clustering in today's Universe matched that measured using large galaxy surveys and weak gravitational lensing (see Spergel et al., 2003, and references therein). Consequently, the ΛCDM model has now established itself firmly as the standard paradigm for structure formation. With further data from WMAP and from other sources, the parameters of this new paradigm are now well constrained (Spergel et al., 2007; Komatsu et al., 2009).

1.4.5 Galaxy Formation

(a) Monolithic Collapse and Merging Although it was well established in the 1930s that there are two basic types of galaxies, ellipticals and spirals, it would take some 30 years before detailed models for their formation were proposed. In 1962, Eggen, Lynden-Bell & Sandage considered a model in which galaxies form from the collapse of gas clouds, and suggested that the difference between ellipticals and spirals reflects the rapidity of star formation during the collapse. If most of the gas turns into stars as it falls in, the collapse is effectively dissipationless and infall motions are converted into the random motion of stars, resulting in a system which might resemble an elliptical galaxy. If, on the other hand, the cloud remains gaseous during collapse, the gravitational energy can be effectively dissipated via shocks and radiative cooling. In this case, the cloud will shrink until it is supported by angular momentum, leading to the formation of a rotationally-supported disk. Gott & Thuan (1976) took this picture one step further and suggested that the amount of dissipation during collapse depends on the amplitude of the initial perturbation. Based on the empirical fact that star formation efficiency appears to scale as ρ^2 (Schmidt, 1959), they argued that protogalaxies associated with the highest initial density perturbations would complete star formation more rapidly as they collapse, and so might produce an elliptical. On the other hand, protogalaxies associated with lower initial density perturbations would form stars more slowly and so might make spirals.

Larson (1974a,b, 1975, 1976) carried out the first numerical simulations of galaxy formation, showing how these ideas might work in detail. Starting from near-spherical rotating gas clouds, he found that it is indeed the ratio of the star-formation time to the dissipation/cooling time which determines whether the system turns into an elliptical or a spiral. He also noted the importance of feedback effects during galaxy formation, arguing that in low mass galaxies, supernovae would drive winds that could remove most of the gas and heavy elements from a system before they could turn into stars. He argued that this mechanism might explain the low surface brightnesses and low metallicities of dwarf galaxies. However, he was unable to obtain the high observed surface brightnesses of bright elliptical galaxies without requiring his gas clouds to be much more slowly rotating than predicted by the tidal torque theory; otherwise they would spin up and make a disk long before they became as compact as the observed galaxies. The absence of highly flattened ellipticals and the fact that many bright ellipticals show little or no rotation (Bertola & Capaccioli, 1975; Illingworth, 1977) therefore posed a serious problem for this scenario. As we now know, its main defect was that it left out the effects of the dark matter.

In a famous 1972 paper, Toomre & Toomre used simple numerical simulations to demonstrate convincingly that some of the extraordinary structures seen in peculiar galaxies, such as long tails, could be produced by tidal interactions between two normal spirals. Based on the observed frequency of galaxies with such signatures of interactions, and on their estimate of the time scale over which tidal tails might be visible, Toomre & Toomre (1972) argued that most elliptical galaxies could be merger remnants. In an extreme version of this picture, all galaxies initially form as disks, while all ellipticals are produced by mergers between pre-existing galaxies. A

virtue of this idea was that almost all known star formation occurs in disk gas. Early simulations showed that the merging of two spheroids produces remnants with density profiles that agree with observed ellipticals (e.g., White, 1978). The more relevant (but also the more difficult) simulations of mergers between disk galaxies were not carried out until the early 1980s (Gerhard, 1981; Farouki & Shapiro, 1982; Negroponte & White, 1983; Barnes, 1988). These again showed merger remnants to have properties similar to those of observed ellipticals.

Although the merging scenario fits nicely into a hierarchical formation scheme, where larger structures grow by mergers of smaller ones, the extreme picture outlined above has some problems. Ostriker (1980) pointed out that observed giant ellipticals, which are dense and can have velocity dispersions as high as $\sim 300\,\mathrm{km\,s^{-1}}$, could not be formed by mergers of present-day spirals, which are more diffuse and almost never have rotation velocities higher than $300\,\mathrm{km\,s^{-1}}$. As we will see below, this problem may be resolved by considering the dark halos of the galaxies, and by recognizing that the high redshift progenitors of ellipticals were more compact than present-day spirals. The merging scenario remains a popular scenario for the formation of (bright) elliptical galaxies.

(b) The Role of Radiative Cooling An important question for galaxy formation theory is why galaxies with stellar masses larger $\sim 10^{12}\,\mathrm{M_\odot}$ are absent or extremely rare. In the adiabatic model, this mass scale is close to the Silk damping scale and could plausibly set a *lower* limit to galaxy masses. However, in the presence of dark matter Silk damping leaves no imprint on the properties of galaxies, simply because the dark matter perturbations are not damped. Press & Schechter (1974) showed that there is a characteristic mass also in the hierarchical model, corresponding to the mass scale of the typical non-linear object at the present time. However, this mass scale is relatively large, and many objects with mass above $10^{12}\,\mathrm{M_\odot}$ are predicted, and indeed are observed as virialized groups and clusters of galaxies. Apparently, the mass scale of galaxies is not set by gravitational physics alone.

In the late 1970s, Silk (1977), Rees & Ostriker (1977) and Binney (1977) suggested that radiative cooling might play an important role in limiting the mass of galaxies. They argued that galaxies can form effectively only in systems where the cooling time is comparable to or shorter than the collapse time, which leads to a characteristic scale of $\sim 10^{12}\,\mathrm{M_\odot}$, similar to the mass scale of massive galaxies. They did not explain why a typical galaxy should form with a mass near this limit, nor did they explicitly consider the effects of dark matter. Although radiative cooling plays an important role in all current galaxy formation theories, it is still unclear if it alone can explain the characteristic mass scale of galaxies, or whether various feedback processes must also be invoked.

(c) Galaxy Formation in Dark Matter Halos By the end of the 1970s, several lines of argument had led to the conclusion that dark matter must play an important role in galaxy formation. In particular, observations of rotation curves of spiral galaxies indicated that these galaxies are embedded in dark halos which are much more extended than the galaxies themselves. This motivated White & Rees (1978) to propose a two-stage theory for galaxy formation; dark halos form first through hierarchical clustering, the luminous content of galaxies then results from cooling and condensation of gas within the potential wells provided by these dark halos. The mass function of galaxies was calculated by applying these ideas within the Press & Schechter model for the growth of non-linear structure. The model of White and Rees contains many of the basic ideas of the modern theory of galaxy formation. They noticed that feedback is required to explain the low overall efficiency of galaxy formation, and invoked Larson's (1974a) model for supernova feedback in dwarf galaxies to explain this. They also noted, but did not emphasize, that even with strong feedback, their hierarchical model predicts a galaxy luminosity function with far too many faint galaxies. This problem is alleviated but not solved by adopting CDM initial conditions rather than the simple power-law initial conditions they adopted. In 1980, Fall

& Efstathiou developed a model of disk formation in dark matter halos, incorporating the angular momentum expected from tidal torques, and showed that many properties of observed disk galaxies can be understood in this way.

Many of the basic elements of galaxy formation in the CDM scenario were already in place in the early 1980s, and were summarized nicely by Efstathiou & Silk (1983) and in Blumenthal et al. (1984). Blumenthal et al. invoked the idea of biased galaxy formation, suggesting that disk galaxies may be associated with density peaks of typical heights in the CDM density field, while giant ellipticals may be associated with higher density peaks. Efstathiou & Silk (1983) discussed in some detail how the two-stage theory of White & Rees (1978) can solve some of the problems in earlier models based on the collapse of gas clouds. In particular, they argued that, within an extended halo, cooled gas can settle into a rotation-supported disk of the observed scale in a fraction of the Hubble time, whereas without a dark matter halo it would take too long for a perturbation to turn around and shrink to form a disk (see Chapter 11 for details). They also argued that extended dark matter halos around galaxies make mergers of galaxies more likely, a precondition for Toomre & Toomre's merger scenario of elliptical galaxy formation to be viable.

Since the early 1990s many studies have investigated the properties of CDM halos using both analytical and N-body methods. Properties studied include the progenitor mass distributions (Bond et al., 1991), merger histories (Lacey & Cole, 1993), spatial clustering (Mo & White, 1996), density profiles (Navarro et al., 1997), halo shapes (e.g., Jing & Suto, 2002), substructure (e.g., Moore et al., 1998a; Klypin et al., 1999), and angular-momentum distributions (e.g., Warren et al., 1992; Bullock et al., 2001a). These results have paved the way for more detailed models for galaxy formation within the CDM paradigm. In particular, two complementary approaches have been developed: semi-analytical models and hydrodynamical simulations. The semi-analytical approach, originally developed by White & Frenk (1991) and subsequently refined in a number of studies (e.g., Kauffmann et al., 1993; Cole et al., 1994; Dalcanton et al., 1997; Mo et al., 1998; Somerville & Primack, 1999), uses knowledge about the structure and assembly history of CDM halos to model the gravitational potential wells within which galaxies form and evolve, treating all the relevant physical processes (cooling, star formation, feedback, dynamical friction, etc.) in a semi-analytical fashion. The first three-dimensional, hydrodynamical simulations of galaxy formation including dark matter were carried out by Katz in the beginning of the 1990s (Katz & Gunn, 1991; Katz, 1992) and focused on the collapse of a homogeneous, uniformly rotating sphere. The first simulation of galaxy formation by hierarchical clustering from proper cosmological initial conditions was that of Navarro & Benz (1991), while the first simulation of galaxy formation from CDM initial conditions was that of Navarro & White (1994). Since then, numerical simulations of galaxy formation with increasing numerical resolution have been carried out by many authors.

It is clear that the CDM scenario has become the preferred scenario for galaxy formation, and we have made a great deal of progress in our quest towards understanding the structure and formation of galaxies within it. However, as we will see later in this book, there are still many important unsolved problems. It is precisely the existence of these outstanding problems that makes galaxy formation such an interesting subject. It is our hope that this book will help you to equip yourself for your own explorations in this area.