In the preceding five parts of this series of pedagogical articles on big bang cosmology, in an approximately historical chronology we have reviewed motivations, principal concepts, and crucial observations that support the model.[1-5,6] Here we conclude our survey by revisiting what we might call Einstein’s enigma: the cosmological constant.

**The Original Cosmological Constant**

As we have seen, in 1917 Einstein boldly launched modern cosmology by applying general relativity to the *entire universe.* He disposed of the Newtonian problem at infinity by invoking a closed geometry.[2] For his model to be consistent with contemporary astronomical data that suggested a static universe, Einstein reluctantly introduced an extra term that opposed gravity, a “cosmological constant” $\Lambda > 0$. It had to be small because its presence was not needed at galactic and smaller scales; in SI units, $\Lambda_{obs} \lesssim 10^{-35}$ s$^{-2}$.[7]

To oppose gravity it exerts negative pressure,

$$P_\Lambda = -\frac{\kappa}{3} \Lambda \sim -5 \times 10^{-5} \text{ N/m}^2$$

where $\kappa = 8\pi G/3c^2$, with $G$, Newton’s gravitation constant and $c$, the speed of light.

In the 1920s the time-dependent cosmologies of Alexander Friedmann and Georges Lemaître predicted the “velocity–distance relation”.[4] This stretching of space at the cosmic scale was first empirically documented by Edwin Hubble in 1929, and we speak of “Hubble’s law”,

$$H_0 s = v$$

where $H_0$ denotes the present value of “Hubble’s parameter” that relates the relative velocity $v$ between galaxies to the distance $s$ between them. In the early 1930s Einstein visited Hubble at the Mt. Wilson Observatory. By then Einstein had almost urged that the cosmological constant be dropped, for in 1923 Hermann Weyl and Arthur Eddington showed that test particles recede from one another.

In a letter of May 23, Einstein wrote to Weyl, “If there is no quasi-static world, then away with the cosmological term.”[8]

But $\Lambda$ could not be so easily dismissed. A cosmological constant term is *allowed* by the general covariance of Einstein’s field equations. However, no known principle exists that determines the value of $\Lambda$, zero or otherwise. Eddington argued through the 1920s and 1930s that the expansion is *driven* by $\Lambda$.[9] But the universe could expand with or without $\Lambda$, and it’s small anyway, so as a practical matter most investigators took $\Lambda = 0$. Not until the mid-1990s was there any compelling *empirical* reason to think otherwise.

Then the question of $\Lambda$ has re-emerged, and sharpens a significant dissonance between the two most successful theories of physics: general relativity and quantum mechanics.

**Standard Model Observables**

We have seen how the unperturbed standard model of big bang cosmology finds expression in the Friedmann equations.[4] In SI units they read

$$H^2 + c^2 k/R^2 = \kappa \rho$$

and

$$R = -\frac{1}{2} \kappa (\rho + 3P)R$$

where $R$ denotes the cosmic scale factor ($R/R_1$ gives the factor by which the universe expands from time $t_1$ to $t$); $k = 0$ or $\pm 1$ (see below); $\rho = $ energy density; $P = $ pressure, $H = R/R$, and overdots denote time derivatives. The energy density includes contributions from radiation, matter, and $\Lambda$,

$$\rho = \rho_r + \rho_m + \rho_\Lambda.$$  

The energy density and partial pressure of constituent $i$ are related by an equation of state,

$$w_i = P_i/\rho_i. $$

For nonrelativistic matter $w_i = 0$; $w_i = ½$ for radiation; and $w_\Lambda = -1$.

The Friedmann equations with thermodynamics shows that $\rho_m \sim R^{-3}$, which in Eq. (3) (neglecting $k$) leads to $\rho_r \sim R^4$ for radiation, $\rho_m \sim R^{-3}$ for matter, while $\rho_\Lambda \approx \text{const.}$ In the expanding universe, early on when $R$ is very small, radiation and ultrarelativistic particles dominate the energy budget. As $R$ grows, matter dominates next as the energy density of radiation thins out faster. However, if $\Lambda \neq 0$ then $\rho_\Lambda$ eventually dominates. In the Friedmann equations these $R$-dependent densities give $\dot R(t) \sim R^{1/2}$ when radiation dominates, $R(t) \sim t^{2/3}$ during matter dominance, and $R(t) \sim e^{t/\sqrt{\Lambda k}}$ when $\rho_m$ dominates. When radiation and matter dominate, gravity slows the expansion, but when $\rho_\Lambda$ dominates, its negative pressure eventually changes deceleration into positive acceleration (see Fig. 1).

To compare any model to reality one must express the model’s relationships in terms of observables. Using Eq. (3) we define a “critical energy density” by setting $\Lambda = 0$ and $k = 0$, giving $\rho_{cr} \equiv H^2/\kappa$. Let us also define $\Omega = \rho/\rho_{cr}$, the ratio of actual to critical energy density, doing likewise for each constituent species, $\Omega_i = \rho_i/\rho_{cr}$. Thus Eq. (3), for nonzero $\Lambda$ and $k$ may be elegantly written

$$1 + c^2 k/R^2 H^2 = \Omega_m + \Omega_r + \Omega_\Lambda = \Omega$$
If \( k = +1 \) then \( \Omega > 1 \), and gravity wins over speed. The expansion will someday halt, and then the universe will begin to collapse (“closed” universe). If \( k = -1 \) then \( \Omega < 1 \), and speed triumphs to exceed over gravity (“open” universe). If \( k = 0 \) and \( \Omega = 1 \), the universe is “flat” (Euclidean) and will continue to expand with a rate that (if \( A = 0 \)) asymptotically approaches zero.

If only matter and radiation fill the universe, then the expansion continuously decelerates. We define the “deceleration parameter” \( q \equiv -\frac{R}{RH^2} \), with which Eq. (4) neatly becomes

\[
q = \frac{1}{2} \Omega_m + \Omega_r - \Omega_\Lambda.
\]

Because of the finite speed of light, the signals we receive from galaxies or supernovas show us what they were when the light was emitted, not what they are when we receive those signals now. The contemporary values of the \( \Omega \) may be expressed as rescalings from earlier epochs, that depend upon the redshift of a source’s light. Recall that the fractional change in wavelength between a moving emitter and a stationary receiver defines the redshift \( z \equiv \left( \frac{\lambda_{\text{rec}}}{\lambda_{\text{emit}}} \right) - 1 \). The stretching of space follows the same pattern,

\[
\frac{R_{\text{now}}}{R_{\text{then}}} = 1 + z.
\]

Thus, \( \Omega \) (for instance) as a function of \( z \) can be written \( \Omega(z) = \Omega_m H_o^2 (1+z)^{\gamma} \) (sub-o means present value), because \( \rho_i \sim T^4 \) by Stefan’s law, and \( R \sim 1/T \) from thermodynamics.[4]

Through such rescalings the present values of observables can be related to various models of the universe’s history and composition.

Measuring \( \Omega \)

By the early 1990s, data was beginning to constrain \( \Omega \). In those days one assumed \( \Omega = \Omega_m + \Omega_r \). Sky surveys such as those done by Robert Kirshner et al. measured \( \Omega_m \) as follows. In several deep-space fields, each about a cubic megaparsec in volume, an inventory was made of the mass of luminous matter. Using the best estimate for the mass ratio of dark to luminous matter, the total mass in each field could be inferred. These methods yielded \( \Omega_m \approx 0.26 \).[10] suggesting an open universe. But this answer, like all answers, raised other questions.

For instance, it was well known from primordial nucleosynthesis constraints that baryonic matter (protons and neutrons) could contribute only \(-0.05\) to \( \Omega \).[4] This puzzle fed into another one, the “dark matter” problem that had been anticipated for galaxy clusters by Fritz Zwicky in the 1930s, with significant evidence for dark matter within individual galaxies provided by Vera Rubin in the early 1950s.[11] They showed us that the rotational kinetic energy of galaxies exceeds the gravitational binding energy of their luminous matter. In addition, to be consistent with theories of galaxy formation, most of this nonbaryonic dark matter would have to be heavy and slow to seed galaxy formation; baryons would not get around to coalescing into stars until a billion years or so after the big bang. However, nobody knew (or yet knows) what this nonbaryonic dark matter really is, although hypothetical candidates exist.

There was a long-standing puzzle about \( \Omega \) from the theoretical side too, namely, why should \( \Omega \), be of order unity, and not, say, 500 or 10,000? For if you write \( \Omega = 1 + \varepsilon(t) \), then the Friedmann equations say that a nonzero \( \varepsilon \) grows rapidly with time, such that to have \( \Omega = 1 \) today requires fine-tuning the initial conditions.

As discussed in Ref. 5, such fine-tuning problems share a common theoretical solution called “inflation”, first articulated in 1981 by Alan Guth.[5,12] Inflation took the universe through a brief but violent epoch of exponential growth (from \( t \sim 10^{-34} \) s to \( t \sim 10^{-32} \) s). This episode was driven not by \( \Lambda \) (\( \rho_\Lambda \) was too small), but by a hypothetical scalar field that would have filled all space. Its unstable but nonzero expectation value would provide a large but temporary effective \( \rho_i \) that briefly dominated the energy density. The ensuing exponential growth of \( R \) drove \( kR^2 \rightarrow 0 \), and thus \( \Omega \rightarrow 1 \), independent of initial conditions. Then the scalar field decayed into its stable ground state, releasing latent energy that reheated the universe with new particle–antiparticle production. After inflation was over, the universe expanded normally.

If the inflationary phase drove the total \( \Omega \) to 1, and if \( \Omega_m \approx 0.26 \) and \( \Omega_r = 0 \), then one must seek another \( \frac{1}{3} \) from somewhere—\( \Lambda \) and/or something that mimics it.[13] On this point Kirshner mused, “When I want to tease Jerry [Ostriker, a theorist], I say that he and/or something that mimics it!"[13] On this point Kirshner mused, “When I want to tease Jerry [Ostriker, a theorist], I say that he and/or something that mimics it!

But Kirshner politely added, “Like all effective teasing, this is a little unfair, because there is another cosmological fact that [Michael] Turner and Ostriker and [Paul] Steinhardt could match with \( \Lambda \), but could not match without it. That is the age of the universe.”[14]

The value of \( H_o \) is about 70 (km/s)/Mpc.[15] Had the universe been expanding at a constant rate throughout its history (a “coasting universe”), the time since any two points now at a distance \( s \) apart began separating, and thus the time since the big bang itself, would be \( t = s/H_o = 14 \) Gyr, the “Hubble time”. If you allow gravitational braking due to matter, you get \( \frac{1}{2} H_o^{-1} \) for the age of the universe, younger than the oldest stars in it! Indeed, one can estimate a star’s lifetime from its observed mass, luminosity, and the rate energy is released by the fusion reactions powering it. But if gravitational braking were approximately cancelled by the effects of \( \Lambda \), then the actual age of the universe and the Hubble time could be in closer agreement.

Recall that in 1993 the Cosmic Background Explorer (COBE) satellite found the temperature fluctuations in the cosmic microwave background radiation.[16] At smaller angular scales than
COBE could resolve, the power spectrum of these fluctuations were expected to exhibit a fine structure of peaks and valleys. Their heights, widths, and locations can be predicted in various models, to be compared with measured values of the observables.[5] Data from COBE and its successors[17-19] have so far affirmed a flat inflationary universe with cold dark matter and positive $\Lambda$ (or its equivalent). Jumping ahead in the chronology, in 2001 the Degree Angular Scale Interferometer group, for instance, reported $\Omega_\text{mantle} = 0.90 \pm 0.04$. [18] A 2007 survey from the Wilkinson Microwave Anisotropy Probe increased the precision by two more orders of magnitude:[19]

$$
\begin{align*}
\Omega_\text{m} & = 1.0052 \pm 0.0064, \\
\Omega_\Lambda & = 0.721 \pm 0.015, \\
\Omega_\text{baryon,0} & = 0.0462 \pm 0.0015, \\
\Omega_\text{dark matter,0} & = 0.233 \pm 0.013 \\
H_0 & = 70.1 \pm 1.3 \text{ (km/s)/Mpc,} \\
t_\text{e} & = 13.73 \pm 0.12 \text{ Gyr.}
\end{align*}
$$

If $\Omega_\Lambda = 0.721$ is the signature of the cosmological constant, then the universe had better be accelerating!

### The Accelerating Universe

Intergalactic distances can be measured with confidence if “standard candles”, light sources of known luminosity, are available. The first of these were the Cepheid variable stars. Recall that Henrietta Leavitt found a correlation between a Cepheid’s average brightness and its period, which enabled Edwin Hubble to measure distances to other galaxies and discover the immensity of the universe.[1] However, a Cepheid cannot be seen across the universe. Supernovas, which outshine an entire galaxy for a short time, are visible across cosmic distances.

Supernovas of so-called Types 1b, 1c, and II result from a catastrophic core collapse and rebound that explodes the star, leaving behind a neutron star or black hole, depending on the compact object’s mass. The wide range of stellar masses that lead to these events precludes their utility as standard candles.

In contrast, the progenitor stars of Type Ia supernovas (SN1a) are only massive enough to fuse hydrogen to helium, then helium to carbon and oxygen. After this, the star swells into the red giant phase and puffs off its outer layers, and eventually the core settles down as a carbon–oxygen white dwarf.

To become an SN1a the white dwarf must also belong to a multiple star system. The dwarf accretes matter from a companion star until the captured material raises the dwarf’s mass toward the Chandrasekhar limit (1.4 solar masses), the maximum mass for a white dwarf to be supported against further collapse, thanks to the Pauli exclusion principle and its electron degeneracy pressure.[20] When this limit is approached the star undergoes explosive carbon fusion, obliterating the white dwarf. Thus, all these supernovas explode from nearly the same state, making them plausible standard candles. The light from the explosion lasts about a month and is so bright that the SN1a light curve (intensity vs. time) can be measured across intergalactic distances (see Fig. 2).

Although these stars exhibit some variance in their peak luminosities, a tight correlation exists between the maximum luminosity and the rate of its decline.[21] With this calibration, SN1a events serve well for gauging cosmic distances.

Let the calibrated peak luminosity be $L$. When we receive the light of flux $I$ at some distance $r$ away,[22] in a static universe we would write $I = L/4\pi r^2$. In an expanding universe the distance-at-emission $r$ stretches during the light’s journey (recall Eq. 9), so that

$$
I = L / [4\pi (v+1)^2 \left(\frac{r}{cz}\right)]
$$

We measure $I$ with our photometer, $z$ from the redshift of the spectral lines, and $L$ from the calibration, to infer $r$.

What can such data tell us about cosmic kinematics? Toward this end consider a spacetime diagram that shows the world lines of three light sources whose light we receive tonight (event “now” in Fig. 3). World line C shows the fiducial “coasting universe”. World line D describes deceleration as gravity applies the brakes. World line A shows a universe that expands with positive acceleration. If the universe decelerates, tonight’s received intensity from the supernova is brighter than it would be in the coasting universe. If the universe accelerates, the light will be dimmer.

Just as Ohm’s law holds with accuracy only over a limited range of applied voltages, similarly does the linear Hubble’s law, Eq. (2), hold accurately only for galaxies at low redshift. As the redshift increases, from the Friedmann equations, without assuming a value for $w_\Lambda$, neglecting $\Omega_\Lambda$, and noting for both Doppler and cosmological redshifts that $z = v/c$ for $v/c << 1$, one may derive nonlinear corrections to Hubble’s law:[23]

$$
H_0 s = c z - \frac{1}{2} \left[1 + q_0 \right] c z^2 + ...
$$

$$
= v - \frac{1}{2} \left[1 + \frac{1}{2} \Omega_\text{m} + 3 w_\Lambda \Omega_\Lambda / 2 \right] v^2/c + ... \quad (11)
$$

In the 1990s two groups geared up to carry out systematic searches for SN1a candidates at high redshift. They aimed to test the implications of Eq. (11), and others like it, to measure such quantities as $\Omega_\text{m}$, $\Omega_\Lambda$, and $w_\Lambda$. Near a new moon, deep-space fields were scanned with CCD cameras mounted on moderately large telescopes. At the next new moon the same fields were imaged again and the consecutive images subtracted, revealing supernovas. When a candidate SN1a was found, its light curve and redshift were measured. The first such group to get started, the Supernova Cosmology Project that was up and running by the mid-1990s, was led by Saul Perlmutter of the Lawrence Berkeley Lab.[24] By the mid-’90s another supernova...
that the universe carries zero energy density, but quantum physics teaches us that we cannot know the vacuum energy to be zero to an infinite number of decimal places. A realistic zero means “0 ± ΔE”.

Quantum theory sees space as seething with virtual particle–antiparticle pairs that pop into and out of existence. For instance, to make an electron–positron pair for which 2mc² ~ 1 MeV, the needed energy can be “borrowed” for ~10⁻²¹ s, within the quantum tolerance of the uncertainty principle, ΔE Δt ~ ℏ.

(The universe itself may have begun with such a fluctuation. A small ΔE squeezed within the Planck volume ~10⁻¹⁰⁰ cm³ would be energy density enough, with the kinetic energy we see today offset by negative gravitational potential energy.) These evanescent particles that populate the vacuum can be “polarized”, as measured in phenomena such as the Casimir effect.

So perhaps ρₘ and ρvac are two names for the same thing. Very well then; merely calculate the vacuum energy density using quantum field theory. Depending on how one does it (summing quantum oscillator modes and choosing a cutoff), one obtains a result that exceeds the empirical value of ρₘ by about 30 to 120 orders of magnitude.[28]

Alternatively, perhaps the density of dark energy is not really constant! One can model this possibility with a hypothetical scalar field Φ(t), playfully called “quintessence”, having kinetic energy density \( \frac{1}{2} \dot{\Phi}^2 \) and interactions described by some potential \( V(\Phi) \). As a relativistic fluid, its pressure and density yield the equation of state

\[
w = \frac{P}{\rho} = -\frac{\frac{1}{2} \dot{\Phi}^2 - V(\Phi)}{\frac{1}{2} \dot{\Phi}^2 + V(\Phi)}
\]

In the hot early universe, kinetic energy dominates and \( w = +1 \). In the old cold universe, the kinetic energy vanishes and \( w = -1 \). This possibility, and the present values of \( \Omega_{\text{m}} \) and \( \Omega_{\Lambda} \), suggest a coincidence puzzle.[29] For if \( \Omega_{\Lambda} \) is const. across all time, then in a matter-dominated universe \( \Omega_{\text{m}} / \Omega_{\Lambda} \) evolves as \( 1/R^3 \). When such a universe is young then \( R \) is small and \( \Omega_{\text{m}} / \Omega_{\Lambda} \rightarrow \infty \), but matter density would become negligible as \( R \) grows, and ultimately \( \Omega_{\text{m}} / \Omega_{\Lambda} \rightarrow 0 \). Today, \( \Omega_{\text{m}} / \Omega_{\Lambda} \approx 0.25 \). What should we make of the fact that life and intelligence appear in the universe when \( \Omega_{\text{m}} / \Omega_{\Lambda} \) are comparable? Extremes in temperature are inhospitable to life as we know it; could \( \Omega_{\text{m}} / \Omega_{\Lambda} \) be a thermometer, with \( \Omega_{\text{m}} / \Omega_{\Lambda} \sim 1 \) necessary to support life? Cosmology raises strange and interesting questions!

**Dark Energy:** Our “Ether”?

In the 19th century, the anomalous 43°/century precession of Mercury’s perihelion was well known—and baffling. Some new stuff was proposed that, within the Newtonian paradigm, might account for the anomaly. For instance, a hypothetical planet named Vulcan was suggested. Working backward from the anomaly, its trajectory was calculated, but observations revealed no Vulcan. (This is how Neptune was discovered, by working from perturbations in the orbit of Uranus.) It was also suggested that Newtonian gravitation might need to be modified. This turned out to be the remedy, the general theory of relativity.

General relativity without \( \Lambda \) allows the clean image of gravitation as the “curvature of spacetime”. Introducing \( \Lambda \) complicates that ele-
gant metaphor, but the rigorous logic of general covariance allows the A term, and anything not explicitly forbidden (perhaps by some symmetry principle and its conservation law) could exist. Einstein’s mistake was not in taking the cosmological constant seriously when he thought he had to, but in not taking it seriously enough when he thought he did not have to. But sometimes being technically right is less important than raising deep questions.[30]

Today the accelerating universe and dark energy are as astonishing to us[31] as the null result of the 1889 Michelson–Morley experiment was to physicists of that generation. They conceived of the ‘luminiferous ether’ as the medium necessary for the propagation of light waves. As “stuff” this ether had to carry contradictory properties. Its density had to be so low that we breathe through it, yet it had to support a wave velocity 10^10 times faster than the speed of sound in steel!

As you know, these paradoxes were resolved in 1905 when Einstein showed that the ‘ether as stuff’ concept missed the point. One wonders if we find ourselves in a similar predicament today over dark energy. The enigma of dark energy becomes acute when we think if it as a thing. Perhaps that’s why it’s called dark energy and not “dark matter type II”. But the accelerating universe and dark energy has put to the forefront the long-standing dissonance between gravitation and quantum theory. The conflict began with the realization that quantum gravity is not renormalizable (which means that certain infinities do not cancel out). It shows up acutely now as our inability to understand the vacuum. Maybe general relativity forms the limiting case of a deeper theory, just as it contains Newtonian gravitation as a limiting case. Perhaps quantum theory has reached the limits of what it can do and its assumptions need to be challenged. As Steven Weinberg noted, “Until it is solved, the problem of the dark energy will be a roadblock on our path to a comprehensive fundamental physical theory.”[32] Einstein’s cosmological constant may have the last laugh on us.

“This tale grew in the telling...” — J.R.R. Tolkien

Acknowledgments

I am grateful to David Branch for clarifying many points of supernova physics for me; to Art Hobson for a stimulating conversation; and to Thomas Olsen for carefully reading the draft and making thoughtful corrections.

***

[22] To astronomers “flux” means power/area, and “intensity” means flux/steradian.
[23] One writes \( R(t) = R(t_\alpha - \delta) \) where \( t_\alpha \) denotes the present. Perform a Taylor series expansion about \( \delta = 0 \) and use Eqs. (3), (4), and (9). To rewrite the distance \( s \), from the Friedmann metric we integrally \( s = c D(t) \). For more details see M. Hobson, G. Efstathiou, and A. Lasenby, General Relativity: An Introduction for Physicists, §14.10 (Cambridge Univ. Press, 2006).
[31] The Dec. 18, 1998 cover of Science announced “The Accelerating Universe” to be the “Breakthrough of the Year”.