The development of nonlinear science*

Alwyn Scott
Department of Mathematics
University of Arizona
Tucson, Arizona USA

August 11, 2005

Abstract

Research activities in nonlinear science over the past three centuries are reviewed, paying particular attention to the explosive growth of interest in chaos, solitons, and reaction-diffusion phenomena that occurred during the 1970s and considering whether this explosion was an example of a “scientific revolution” or Gestalt-like “paradigm shift” as proposed by Thomas Kuhn in the 1962. The broad structure of modern nonlinear science is sketched and details of developments in several areas of nonlinear research are presented, including cosmology, theories of matter, quantum theory, chemistry and biochemistry, solid-state physics, electronics, optics, hydrodynamics, geophysics, economics, biophysics and neuroscience. It is concluded that the emergence of modern nonlinear science as a collective interdisciplinary activity was a Kuhnian paradigm shift which has emerged from diverse areas of science in response to the steady growth of computing power over the past four decades and the accumulation of knowledge about nonlinear methods, which eventually broke through the barriers of balkanization. Implications of these perspectives for twenty-first-century research in biophysics and in neuroscience are discussed.

PACS: 01.65.+g, 01.70.+w, 05.45.-a

*Copyright © 2005 by Alwyn C. Scott. All rights reserved.
List of Figures

1. The annual number of articles in scientific publications that have used the term CHAOS, SOLITON, and REACTION-DIFFUSION in their titles, abstracts, or key words, and the TOTAL of these three plots. (Data from the Science Citation Index Expanded.) The VERHULST curve is calculated from Equation (1) to approximate the TOTAL curve with $N_0 = 3100$, $N(1970) = 12$ and $\lambda = 0.25$. (Note that the values of the CHAOS plot are misleadingly high before about 1975, as authors then used the term in its classical meaning.)

2. A histogram of the data in Table 1, showing that the number of journals available for publications in nonlinear science increased greatly between 1980 and 2000.

3. The number of transistors on an Intel processor vs. time (data from [256]).

4. Plots of the annual number of citations to Lorenz’s 1963 paper ([311]) and the annual number of papers using the term “chaos”. (Data from Science Citation Index Expanded.)

5. Sketch of the Lorenz attractor of a solution trajectory of Equations (3) with $\sigma = 10$, $r = 28$ and $b = 8/3$. The view is looking down from the $y$-axis onto the $(x, z)$-plane.

6. A bifurcation diagram for solutions of Equation (4). (Courtesy of Mikhail Rabinovich and Nikolai Rulkov.)

7. A hydrodynamic soliton created in a wave tank by John Scott Russell in the 1830s. (a) A column of water is accumulated at the left-hand end of the tank. (b) Release of this water by lifting the sliding panel generates a solitary wave that travels to the right with velocity $v$. (Redrawn from [448].)

8. A plot of a 2-soliton KdV solution $u_2(x, t)$ from Equation (14).

9. A single soliton (kink) on a mechanical model of Equation (17).

10. Stroboscopic photograph showing Lorentz contraction of a kink soliton on a mechanical model of Equation (17) [460]. (Dissipative effects cause the kink to slow down as it moves to the right, whereupon its size increases.)

11. A stationary breather solution of SG plotted from Equation (22) with $\omega = \pi/3$.

12. The upper figure shows a Toda lattice at rest. The lower figure shows the Toda-lattice (TL) soliton given in Equation (25).
13 (a) A computer simulation of the deadly tsunami (an example of Russell’s Great Wave) in the Indian Ocean on 26 December 2004, 1 hour and 50 minutes after it was launched by an earthquake source (green area) near Sumatra. The red colour indicates that the water surface is higher than normal, while the blue colour means that it is lower, and the darker the colour, the larger the amplitude. See http://staff.aist.go.jp/kenji.satake/animation.html for an animated version of this figure. (Courtesy of Kenji Satake, National Institute of Advanced Industrial Science and Technology, Japan.)

(b) Apollo–Soyuz photograph of the Andaman sea, showing surface evidence of internal waves [398]. [The location of (b) is shown as a small rectangle in (a).] (Courtesy of Alfred Osborne.)

14 An early cathode-ray oscilloscope photograph of a nerve impulse. The dots on the lower edge indicate time in milliseconds, increasing to the right. The solid line shows the time course of the transmembrane voltage \( V \), becoming increasingly positive inside the axon. The wide band shows the transmembrane permeability which is large (small membrane resistance) during the impulse. (Courtesy of Kenneth Cole.)

15 Diagram showing a squid axon and the measurement of internal voltage using a glass microelectrode (not to scale).

16 In a BZ reaction, two-dimensional waves spiral outward from organizing centers, intersecting to form cusps. (Courtesy of A.T. Winfree.)

17 A computer generated plot of a scroll ring. (Courtesy of A.T. Winfree.)

18 Interrelations among low-dimensional chaotic systems, reaction-diffusion (RD) systems and integrable (solitonic) systems.

19 Three-dimensional skyrmions with increasing values of topological charge. (Courtesy of Paul Sutcliffe.)

20 Annual number of articles in Physical Review D that have used the term “soliton” in the title, abstract or key words. (Data from Science Citation Index Expanded.)

21 The planar structure of a benzene molecule, showing a local mode of the CH-stretching oscillation.

22 Energy eigenvalues of the DNLS equation at the second \( (n = 2) \) quantum level with \( \epsilon = 1 \) as \( f \to \infty \). There is one soliton band for each value of \( \gamma \), which is plotted for several different values of \( \gamma \).

23 Infrared absorption spectra of crystalline acetanilide (ACN) in the region of the amide-I (CO-stretching) mode as a function of temperature. A normal (delocalized) amide-I band is at 1665 cm\(^{-1}\) and a self-localized (soliton) peak is at 1650 cm\(^{-1}\).

24 A piece of muscovite mica (overall length is 23 cm). Inside the oval marked “A”, the tracks of several moving breathers appear to be emerging from a charged-particle track. (Courtesy of Mike Russell.)

25 A Josephson oscillator in which a quantum of magnetic flux (fluxon) bounces back-and-forth between the two ends of a rectangular junction.

26 (a) Sketch of a pump-probe setup. (b) Pump-probe experiment of an excited mode.
27. A log-log plot of data from the Trinity explosion on 16 July 1945, showing the radius of expansion \( (R) \) as a function of time \( (t) \) during the first 62 ms. (Data from [505].)

28. A tidal bore (or mascaret) at Quillebeuf, France. The site is about 36 km from the mouth of the Seine, where tides up to 10 m occur. Due to dredging in the 1960s, this striking nonlinear phenomenon has disappeared. (Courtesy of Jean-Jacques Malandain.)

29. Shadowgraph showing shock waves produced by a Winchester 0.308 caliber bullet travelling through air at about 2.5 times the speed of sound. (Courtesy of Ruprecht Nennstiel, Bundeskriminalamt Wiesbaden, Germany.)

30. Rogue waves (RWs). (a) A RW approaching the stern of a merchant ship in the Bay of Biscay. (Courtesy of NOAA Photo Library.) (b) A RW produced in the Large Wave Channel in Hannover. (Courtesy of the Coastal Research Centre, Hannover, Germany.) (c) A snapshot at the peak value of a numerical RW computed from a 2 + 1 dimensional NLS system, where the parameters are chosen to correspond to the ocean and axes are labeled in metres. (Courtesy of Alfred Osborne.)

31. Natural atmospheric vortices. (a) A dust devil in Death Valley, California. (Courtesy of Dr. Sharon Johnson/GeoImages.) (b) Photo of a tornado (right) illuminated by a bolt of lightning (left). (Courtesy of National Oceanic and Atmospheric Administration.) (c) Satellite view of Hurricane Andrew, approaching the U.S. Gulf Coast, 25 August 1992. (Courtesy of National Oceanic and Atmospheric Administration.)

32. Switching action in a branching region of a squid giant axon which is excited by a double pulse [466]. (a) Preparation geometry (not to scale). (b) Blocking of the second impulse. (c) Passage of the second impulse.

33. A Neckar cube.

34. (a) A phase-plane plot from Equations (55) with \( \alpha < 0 \) (only excitatory interactions). (b) A phase-plane plot for \( \alpha > 1/3 \) (excitatory and inhibitory interactions among neurons).
1 Introduction

Research in nonlinear science underwent profound changes over the last three decades of the twentieth century, particularly during the 1970s. Before this decade, communications between researchers doing mathematically related work in different fields of nonlinear science were almost non-existent, and important concepts were either ignored or waiting to be discovered. Nowadays, conditions have changed. Several international conferences on nonlinear science are held every year, mixing participants from a variety of professional backgrounds to a degree that was not imagined in the 1960s. Nonlinear science centres have spread across the globe, bringing together diversely educated researchers to collaborate on interdisciplinary activities, combining their skills in unexpected ways. Dozens of nonlinear science journals have been launched, and hosts of textbooks and monographs are now available for introductory [1, 95, 105, 131, 134, 257, 260, 281, 285, 290, 291, 327, 388, 432, 467, 525, 551] and advanced [5, 17, 22, 25, 33, 79, 113, 164, 191, 196, 206, 205, 213, 262, 304, 383, 407, 442, 497, 514, 518, 555, 579] courses in nonlinear science. A survey of this emerging area can be found in the recently published Encyclopedia of Nonlinear Science [469].

1.1 What is nonlinear science?

Briefly, nonlinear science is the study of those phenomena for which the whole differs from the sum of its parts. Beyond this facile phrase, one can point to many dynamic effects currently being investigated under the aegis of nonlinear science, including but not limited to the following – chaos and turbulence (sensitive dependence on initial conditions or the “butterfly effect,” strange attractors, Julia sets, waterfalls, clear air turbulence), emergent structures (chemical molecules, planets, tornadoes, tsunamis, lynching mobs, optical solitons, black holes, flocks of birds and schools of fish, cities, Jupiters Great Red Spot, nerve impulses), filamentation (rivers, bolts of lightning, woodland paths, optical filaments), threshold phenomena (an electric wall switch, the trigger of a pistol, flip-flop circuits, tipping points, the all-or-nothing property of a neuron), spontaneous pattern formation (development of natural languages, fairy rings of mushrooms, the Gulf Stream, fibrillation of heart muscle, ecological domains), phase changes in condensed matter, (freezing and boiling of liquids, the onset of superconductivity in low temperature metals and superfluity in liquid helium, magnetization in ferromagnetic materials), harmonic generation (digital tuning of radio receivers, conversion of laser light from red to blue), synchronization (coupling of Huygens’s pendulum clocks, mutual entrainment of electric power generators connected to a common grid, circadian rhythms, hibernation of bears, flashing of Indonesian fireflies), shock waves (sonic booms of jet airplanes, the sound of a cannon, bow waves of a boat, sudden pileups in smoothly-flowing automobile traffic), and turbulence (wakes of ships and bullets, waterfalls, breaking waves). All of these phenomena and more comprise the subject matter of nonlinear science, which is in some sense a metascience with roots
reaching into widely diverse areas of modern research.

In the United States, an early use of the term “nonlinear science” appeared in a letter written by Joseph Ford to his colleagues in 1977, introducing his Nonlinear Science Abstracts, which evolved into the journal Physica D: Nonlinear Phenomena. In this letter, Ford defined nonlinear science as including:

nonlinear mathematical studies from completely integrable to completely stochastic systems, studies on nonlinear systems described by stochastic governing equations (differential or otherwise) as well as deterministic nonlinear systems yielding stochastic behavior, and nonlinear studies throughout the physical and social sciences from astronomy to zoology.1

Since the middle of the twentieth century, of course, the adjective “nonlinear” had been employed to modify: analysis, dynamics, mechanics, oscillations, problems, research, systems, theory, and waves, particularly in the Soviet Union [47, 408], but Ford’s use defines a broad field of related activities.

A yet deeper characterization recognizes that the definition of nonlinearity involves a statement about the nature of causality. This concept was discussed by Aristotle some twenty-three centuries ago in his classic work Physics, where it is noted that “We have to consider in how many senses because may answer the question why” [16, 73]. As a “rough classification of the causal determinants of things,” Aristotle suggested four types of cause.

- **MATERIAL CAUSE.** Material cause stems from the presence of some physical substance that is needed for a particular outcome. Following Aristotle’s suggestion that bronze is an essential factor in the making of a bronze statue, many other examples come to mind: atoms of iron are necessary to produce hemoglobin, obesity in the United States is materially caused by our overproduction of corn, water is essential for life. At a particular level of description, a material cause may be considered as a time or space average over dynamic variables at lower levels of description, entering as a slowly varying parameter at the hierarchical level of interest.

- **FORMAL CAUSE.** For some particular outcome to occur, the necessary materials must be arranged in an appropriate form. The blueprints of a house are necessary for its construction, the DNA sequence of a particular gene is required for synthesis of the corresponding protein, and a pianist needs the score to play a concerto. At a particular level of description, formal causes might arise from the more slowly varying values of dynamic variables at higher levels, which enter as boundary conditions at the level of interest.

- **EFFICIENT CAUSE.** For something to happen, according to Aristotle, there must be an “agent” that produces the effect and starts the material on its

---

1 Thanks to James Meiss for sending me a copy of this letter.
way.” Thus, a golf ball moves through the air in a certain trajectory because it was struck at a particular instant of time by the head of a club. Similarly, a radio wave is emitted into the ether in response to the current that is forced to flow through an antenna. Following Galileo, this is the standard sense in which physical scientists use the term causality [73]. Thus an efficient cause is usually represented by a stimulation-response relationship, which can be formulated as a differential equation with a dependent variable that responds to a forcing term.

- **Final Cause.** Events may come about because they are desired by some intentional organism. Thus a house is built—involving the assembly of materials, reading of plans, sawing of wood, and pounding of nails—because someone wishes to have shelter from the elements, and an economic transaction is motivated by future expectations. Purposive answers to the question “why?” are problematic in the biological sciences, and they emerge as central issues in the social sciences because they don’t conform to a general belief in reductionism [468]. In mathematical terms, final causes offer means for closed causal loops between higher and lower levels of dynamic activity which must be included in realistic models.

In more modern (yet not more precise) terms, Aristotle’s material and formal causes are grouped together as distal causes, his efficient cause is called a proximal cause, and his formal cause is either disregarded or termed a teleological cause. Although these classifications seem tidy, reality is more intricate, as Aristotle was well aware [73]. Thus causes may be difficult to sort out in particular cases, with several of them often “coalescing as joint factors in the production of a single effect” [16]. Such interactions among component causes are a typical property of nonlinear phenomena, but distinctions among Aristotle’s “joint factors” are not always easy to make.

A subtle difference between formal and efficient causes appears in the metaphor for Norbert Wiener’s cybernetics: the steering mechanism of a ship [552]. If the wheel is connected directly to the rudder (via cables), then the forces exerted by the helmsman’s arms are the efficient cause of the ship executing a change of direction. For larger vessels, however, control is established through a servomechanism in which the position of the wheel merely sets a pointer that indicates the desired position of the rudder. The forces that move the rudder are generated by a feedback control system (or servomechanism) that minimizes the difference between the actual and desired positions of the rudder. In this case, one might say that the position of the pointer is a formal cause of the ship’s turning, with the servomotor of the control system acting as the efficient cause.

Another example is provided by the conditions needed to cause the firing of a neuron in the human brain. If the synaptic weights and firing threshold are supposed to be constants, they can be viewed as formal causes of a firing event. On a longer time scale associated with learning, however, these parameters can be considered collectively as a weight vector that is governed by the
learning dynamics and might be classified as efficient causes of neuron ignition [466]. Although the switching of a real neuron is far more intricate than this simple picture suggests, the point remains valid – neural switching is a nonlinear dynamic process, melding many contributing factors.

Finally, when a particular protein molecule is constructed within a living cell, sufficient quantities of appropriate amino acids must be available to the messenger RNA as material causes. The DNA code, deciding which amino acids are to be arranged in what order, is a formal cause, and the chemical (electrostatic and valence) forces acting among the constituent atoms are efficient causes.

In the realms of the chemical and biological sciences, it is not surprising to find several different types of causes involved in a single nonlinear event – we expect that parameter values, boundary conditions, and forcing functions will all combine to influence the outcome of a given computation. How can these ideas be extended to social phenomena?

Just as a supercooled liquid, resting quietly in its fluid state, may experience the onset of a phase change during which it suddenly solidifies, collective social phenomena unexpectedly arise, sweeping away previous assumptions and introducing new perspectives. Examples of such “social phase changes” abound – the revolutions in eighteenth-century France and twentieth-century Russia, lynch mobs, the outbreak of war, collective heartbreak over the death of Princess Diana, the Copernican revolution in astronomy during the sixteenth and seventeenth centuries [278]. In the 1970s, I suggest, something similar happened in the organization and practice of nonlinear science.

Mathematical research in nonlinear science goes back at least to Isaac Newton’s successful treatment of the two-body problem of planetary motion, but until recently such activities were scattered among various professional areas, with little awareness of the common mathematical and physical principles involved. Beginning around 1970 this situation changed. Those interested in nonlinear problems became increasingly aware that dynamic concepts first observed and understood in one field (population biology, for example, or flame-front propagation or nonlinear optics or planetary motion) could be useful in others (such as chemical dynamics or neuroscience or plasma confinement or weather prediction). Thus research activities began to be driven more by a general interest in nonlinear phenomena than by specific applications.

Among possible applications, what are the general types of nonlinear problems? First, there is the phenomenon of low-dimensional chaos, comprising several unexpected behaviours that arise from the sensitive dependence of a nonlinear system on its initial conditions – popularly known as the “butterfly effect”. Second, there is the emergence of stable, particle-like entities (new “things”) in energy-conserving (Hamiltonian) systems, a phenomenon exemplified by the solitary wave, in which mass and energy form a self-consistent dynamic structure and the soliton which precisely maintains its shape and speed even after collisions with other solitons. Third, there are reaction-diffusion processes, under which a localized region of dissipative activity emerges – like a candle flame or a nerve impulse – releasing the energy that is necessarily con-
sumed by associated dynamics.

Chaos is a familiar word of Greek origin, describing, perhaps correctly, the original character of the universe, whereas the term soliton was coined in 1965 by Norman Zabusky and Martin Kruskal to indicate the particle-like properties of the solitary-wave solution of a certain nonlinear wave equation [577]. Describing processes in which some essential quantity is released by the ongoing dynamics (like the heat released by a forest fire), the term “reaction-diffusion” is widely but not universally used; thus such phenomena are sometimes referred to as “self-excited” or “self-organizing” waves. Following a coinage by Rem Khokhlov in 1974, they are widely called “autowaves” in the Russian literature [373].

Figure 1: The annual number of articles in scientific publications that have used the term CHAOS, SOLITON, and REACTION-DIFFUSION in their titles, abstracts, or key words, and the TOTAL of these three plots. (Data from the Science Citation Index Expanded.) The VERHULST curve is calculated from Equation (1) to approximate the TOTAL curve with \( N_0 = 3100, N(1970) = 12 \) and \( \lambda = 0.25 \). (Note that the values of the CHAOS plot are misleadingly high before about 1975, as authors then used the term in its classical meaning.)

1.2 An explosion of activity

Although the roots of these three components of modern nonlinear science go back at least to the nineteenth century, the frequency with which they appear

\(^2\)Interestingly, this term took a couple of decades to work its way into English dictionaries, and it was not uncommon for manuscripts published in the 1970s to have “soliton” replaced with “solution” by zealous copy editors of physics journals. The word finally entered the public mind from a Star Trek episode of the early 1990s, and now it in the official Scrabble dictionary.
in scientific publications has grown explosively since 1970, as Figure 1 shows. An aim of this review is to describe how this striking transition came about.

Broadly, these curves show an exponential rise from 1970 to 1990 (with a doubling time of about three years) followed by an apparent leveling off (or “saturation” in the jargon of electrical engineering) in the mid-1990s at a present level of more than 3000 papers per year, or about eight per day (a lower estimate). The same explosive tendency has recently been noted by Rowena Ball in her introduction to Nonlinear Dynamics: From Lasers to Butterflies [25], where she also points out that developments in nonlinear science were driven by the new ideas rather than responding to national needs like research on cancer, plasma confinement, or weapons technology. From other perspectives, these curves look like the growth of transmembrane voltage on the leading edge of a nerve impulse or the heat from a freshly lit bonfire. How can we understand these data? What is causing the initial rise of the curves? Why do they saturate? Can the tools of nonlinear science help us understand a social phenomenon – how modern nonlinear science emerged and grew?

Interestingly, the plots in Figure 1 are similar to the function \[ \frac{N(t)}{N_0 + N(t)} \]
which Pierre Verhulst derived in the mid-nineteenth century as a solution to the nonlinear ordinary differential equation (ODE)

\[ \frac{dN}{dt} = \lambda N \left( 1 - \frac{N}{N_0} \right) \]

for the growth of a biological population. Under this formulation, \( N(t_0) \) is the initial value of the population at \( t = t_0 \), and \( N_0 \) is the limiting population. Equation (2) is now known as the “logistic” or “Verhulst” equation.

In 1962, Thomas Kuhn famously proposed that the history of science is segmented into two interspersed types of activity. First are the eras of “normal science” during which the “paradigms” (or widely-accepted models) of collective understanding are agreed upon and the primary activity is “puzzle solving” by adepts of those paradigms [279] \(^3\). These times of normal science are punctuated by “revolutionary” periods of varying magnitude and theoretical importance, during which one (or more) of the previously accepted paradigms shifts rapidly, leading to abrupt changes in the ways that the scientific community (or a portion thereof) sees the world. Kuhn proposed that these changes in perspective switch in a “Gestalt-like manner” – as with a psychological change of perception in which one sees a familiar image in a new way or with the sudden development of a new species in the “punctuated evolution” of NeoDarwinism [341].

In Aristotelian terms, formal causes of this switching behaviour are pressures that build up as changes in scientific perceptions are resisted by prac-

\(^3\)The third edition of Kuhn’s classic book is particularly valuable as it includes a final chapter responding to critics of the first edition.
tioners of normal science. According to Kuhn, there are several reasons for this resistance. First, of course, most proposals of new theoretical perspectives turn out to be wrong. Second, a new idea may not be wrong (at variance with empirical observation) but it may fail to make new predictions, relegating choice between competing theories to considerations of analytic convenience or taste. (For its first century, Kuhn noted, this was so for the Copernican heliocentric theory of planetary motions vis-à-vis the Ptolemeic geocentric formulation [279], and physicists recall the competition between the seemingly different quantum theories of Werner Heisenberg and Erwin Schrödinger, which are theoretically equivalent but not in ease of application.) Third, available textbooks and university courses support the standard theoretical formulations. Finally, there is the possibility that the new theory is indeed a better representation of reality (more consistent, less ad hoc, in better agreement with experimental data), but influential professional leaders have vested interests in the established paradigm, motivating them to construct firm – if sometimes irrational – defenses of the traditional position. Examples of such resistance are the vigorous refutations by nineteenth-century religious leaders and by Harvard’s Louis Agassiz of Charles Darwin’s theory of natural selection in favor of the previously dominant paradigm of divine creation of immutable biological species [70]. (Interestingly, the anti-Darwinian position is still maintained in some benighted regions of North America.)

1.3 What caused the changes?

While considering formal causes for the exponential rise in the annual number of nonlinear science publications shown in Figure 1, it is interesting to regard Table 1 and the corresponding histogram in Figure 2, which show the growth in the number of journals committed to publishing research in nonlinear science over the past two decades. 4 This is not the whole story because many of the traditional journals of physics, applied mathematics, theoretical biology, and engineering have carried papers on nonlinear science over the past three decades and continue to do so, but the space in these new journals has largely contributed to the rising publication rate shown in Figure 1. Why were they launched at this time?

Another factor associated with the sudden growth of research activity in nonlinear science has been the worldwide emergence of interdisciplinary nonlinear science centres, of which the Center for Studies of Nonlinear Dynamics at the La Jolla Institute (founded in 1978), the Center for Nonlinear Studies at the Los Alamos National Laboratory (founded in 1980), the Institute for Nonlinear Science at UCSD (founded in 1981), and the Santa Fe Institute (founded in 1984) were among the first. A check on the World Wide Web reveals that there are now dozens of such centres of varying size around the globe, dedicated

4Although it has been around since 1962, *Physics Letters A* is included in the list of Table 1 because the publisher decided in 1986 to devote about half of their pages to a section on nonlinear science, with the aim of eventually having a full letters journal in this area.
Table 1: Recently founded journals devoted to nonlinear science.

<table>
<thead>
<tr>
<th>Journal Name</th>
<th>First Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlinear Analysis - Theory, Methods &amp; Applications</td>
<td>1976</td>
</tr>
<tr>
<td>Physica D (Nonlinear Phenomena)</td>
<td>1980</td>
</tr>
<tr>
<td>Ergodic Theory and Dynamical Systems</td>
<td>1981</td>
</tr>
<tr>
<td>Analyse Non Linéaire</td>
<td>1983</td>
</tr>
<tr>
<td>Dynamics and Stability of Systems</td>
<td>1986</td>
</tr>
<tr>
<td>Physics Letters A</td>
<td>1986</td>
</tr>
<tr>
<td>Complex Systems</td>
<td>1987</td>
</tr>
<tr>
<td>Nonlinearity</td>
<td>1988</td>
</tr>
<tr>
<td>Journal of Dynamics and Differential Equations</td>
<td>1989</td>
</tr>
<tr>
<td>Nonlinear Dynamics</td>
<td>1990</td>
</tr>
<tr>
<td>Chaos, Solitons and Fractals</td>
<td>1991</td>
</tr>
<tr>
<td>Chaos</td>
<td>1991</td>
</tr>
<tr>
<td>Journal of Nonlinear Science</td>
<td>1991</td>
</tr>
<tr>
<td>Mathematical Models and Methods in Applied Sciences</td>
<td>1991</td>
</tr>
<tr>
<td>Nonlinear Science Today</td>
<td>1991</td>
</tr>
<tr>
<td>Fractals</td>
<td>1993</td>
</tr>
<tr>
<td>Nonlinear Processes in Geophysics</td>
<td>1994</td>
</tr>
<tr>
<td>Discrete and Continuous Dynamical Systems</td>
<td>1995</td>
</tr>
<tr>
<td>Journal of Dynamical and Control Systems</td>
<td>1995</td>
</tr>
<tr>
<td>Comm. in Nonlinear Science and Numerical Simulation</td>
<td>1996</td>
</tr>
<tr>
<td>Nonlinear Studies</td>
<td>1996</td>
</tr>
<tr>
<td>Regular &amp; Chaotic Dynamics</td>
<td>1996</td>
</tr>
<tr>
<td>Discrete Dynamics in Nature and Society</td>
<td>1997</td>
</tr>
<tr>
<td>Nonlinear Analysis - Modelling and Control</td>
<td>1997</td>
</tr>
<tr>
<td>Nonlinear Dynamics, Psychology, and Life Sciences</td>
<td>1997</td>
</tr>
<tr>
<td>Open Systems &amp; Information Dynamics</td>
<td>1997</td>
</tr>
<tr>
<td>Nonlinear Phenomena in Complex Systems</td>
<td>1998</td>
</tr>
<tr>
<td>Physical Review E</td>
<td>1998</td>
</tr>
<tr>
<td>Far East Journal of Dynamical Systems</td>
<td>1999</td>
</tr>
<tr>
<td>Interfaces and Free Boundaries</td>
<td>1999</td>
</tr>
<tr>
<td>Qualitative Theory of Dynamical Systems</td>
<td>1999</td>
</tr>
<tr>
<td>International J. of Nonlinear Sciences and Numerical Simulation</td>
<td>2000</td>
</tr>
<tr>
<td>Nonlinear Oscillations</td>
<td>2001</td>
</tr>
<tr>
<td>Complexity</td>
<td>2001</td>
</tr>
<tr>
<td>International J. of Nonlinear Modelling in Science and Engineering</td>
<td>2001</td>
</tr>
<tr>
<td>SIAM Journal on Applied Dynamical Systems</td>
<td>2002</td>
</tr>
</tbody>
</table>
Figure 2: A histogram of the data in Table 1, showing that the number of journals available for publications in nonlinear science increased greatly between 1980 and 2000.

to promoting related studies of chaos, solitons and reaction-diffusion phenomena, from the fundamental perspectives of both physics and mathematics and with emphasis upon particular areas of applied research. Again one wonders why these centres were founded. What pressures were driving the changes?

As is suggested by Edward Lorenz in his recent book *The Essence of Chaos* [312], a fundamental cause for the growth of nonlinear science research in the 1970s was the recent increase in computing power. Evidence of this increase is shown in Figure 3, which plots the number of transistors on an Intel processor against time [256]. Between 1970 and 1990, the doubling time of this growth is about two years, which can be compared with Gordon Moore’s 1965 estimate estimate of a one year doubling time (curiously referred to as “Moore’s law”) [367]. Importantly, it was on one of the very first digital computers – the vacuum-tube MANIAC at the Los Alamos National Laboratory in the early 1950s – that Enrico Fermi, John Pasta and Stan Ulam carried out the now famous FPU computations, which eventually led Zabusky and Kruskal to their numerical rediscovery of the soliton in the mid-1960s. And numerical studies of mathematical models for weather prediction on an early vacuum-tube computer (the Royal-McBee LPG-30) with a 16 KB memory in the late 1950s led Lorenz to his unanticipated observation of low-dimensional chaos [311]. In the lexicon of nonlinear science, therefore, the steady increase in computing power shown in Figure 3 is a progressive change in a control parameter, much like the progressive desiccation of a forest as it prepares to burn, the gradual temperature reduction in a jar of supercooled water, or the increasing discontent of a
rebellious population.

Yet another cause of the explosion of activity shown in Figure 1 has been the accumulation of research results over previous decades and centuries, including nineteenth-century studies of planetary motions, hydrodynamics and population growth, mid-twentieth-century studies of electron beam devices (travelling-wave tubes and backward-wave oscillators), particle accelerators and plasma dynamics [303, 304], and the inventions of the laser and of various types of tunnel diodes in the early 1960s.

Thus studies of nonlinear problems have long been carried on in diverse areas of science and mathematics, but these efforts were largely balkanized, with little awareness of the unifying principles relating them. Just as progressively supercooled water becomes more and more inclined to freeze (a sudden transition that can be triggered by shaking the liquid or inserting a small crystal of ice), a drying forest becomes more flammable (ready to burn at the careless drop of a match or the strike of a lightning bolt) and a suppressed population becomes more restive, the scattered results of nonlinear research – particularly during the nineteenth century and the first half of the twentieth century – had reached an unstable level, leaving the exponential growth shown in Figure 1 ready to emerge. What shook the jar, lit the match or fired the first shot?5

---

5As an example of how “supercooled” or “ready to burn” the field of nonlinear science was in the early 1970s, consider the publication trajectory of a survey paper on solitons and solitary waves which was written by Dave McLaughlin, Flora Chu and the present author in the early 1970s and published in the Proceedings of the IEEE [479]. Although electrical engineering journals are not widely read by physicists or mathematicians, this review was unexpectedly well received, and in 1979 it was judged to be a “citation classic” by the Institute of Scientific Information [463].
1.4 Three trigger events

From Aristotle’s perspective, a trigger event is an efficient cause, several of which occurred to ignite the explosion of activity shown in Figure 1. In the summer of 1966, Zabusky and Kruskal organized a NATO-supported “International School of Nonlinear Mathematics and Physics” at the Max Planck Institute for Physics and Astrophysics in Munich [572]. Following general surveys by Werner Heisenberg on nonlinear problems in physics [230] and by Ulam on nonlinear problems in mathematics, there were more focussed talks by Nicholas Bloembergen on nonlinear optics [48], Clifford Truesdale on nonlinear field theories in mechanics, John Wheeler on cosmology, Philip Saffman on homogeneous turbulence, Ilya Prigogine on nonequilibrium statistical mechanics, and Roald Sagdeev on nonlinear processes in plasmas, among others.

In the summer of 1972, a three-week workshop on “Nonlinear Wave Motion” was organized by Alan Newell, Mark Ablowitz and Harvey Segur at Clarkson College of Technology (now Clarkson University) in Potsdam, New York [382]. At this meeting, which was attended by about 60 scientists of varied backgrounds from several different countries, Pasta described the hitherto perplexing FPU problem [170], and Kruskal and Peter Lax explained the general structure of the inverse scattering transform (IST), which had recently been formulated for constructing multi-soliton solutions of the classical Korteweg–de Vries (KdV) equation describing shallow water waves in a one-dimensional channel [192, 297]. Making connection with real-world phenomena, Joe Hammack showed that a tsunami can be viewed as a KdV soliton [219]. Kruskal and others discussed the sine-Gordon (SG) equation, which had arisen in the context of elementary particle theory and was known to have an exact two-soliton solution, showing that SG has an infinite number of independent conservation laws. During this meeting, a copy of a Russian physics journal appeared, containing an article by Vladimir Zhakov and Alexey Shabat in which an IST formulation was developed for multi-soliton solutions of the nonlinear Schrödinger (NLS) equation, which had previously arisen in studies of deep water waves, plasma waves and nonlinear optics [580]. When this paper was translated at a crowded evening session by Hermann Flaschka, excitement among the participants grew. As it became clear that several nonlinear wave systems of practical interest shared related properties, most participants left the workshop expecting that the special (IST) properties of KdV and NLS would generalize to an important class of integrable nonlinear wave equations – an expectation that was soon fulfilled [3].

In the following years there were many such workshops, bringing together researchers from many diverse backgrounds and areas of science. These new relationships and the enlarged perspectives helped to fix nonlinear paradigms in the minds of recently converted acolytes. Communications between Soviet and Western nonlinear scientists were greatly improved by a soliton conference held at the Institute of Theoretical Physics in Kiev in September of 1979 with participants comprising most of the leading researchers from both sides of the Cold War barrier. During these conferences, myriad scientific collaborations
Figure 4: Plots of the annual number of citations to Lorenz’s 1963 paper ([311]) and the annual number of papers using the term “chaos”. (Data from Science Citation Index Expanded.)

were begun and friendships formed which significantly influenced the future course of research in nonlinear science.

Also in 1972, a talk entitled “Predictability: Does the flap of a butterfly’s wings in Brazil set off a tornado in Texas?” was presented by Lorenz at the annual meeting of the American Association for the Advancement of Science in Washington, D.C. [312]. Although similar questions had been asked before and Lorenz did not answer it in this talk, his metaphor for sensitive dependence of a nonlinear system on its initial conditions caught the imagination of the scientific world, not to mention the general public, and studies of chaos began to take off.

Lorenz’s 1972 talk was based on a paper that he had written for meteorologists a decade earlier [311]. This paper presented results of numerical studies on a highly simplified model of the weather system (only three dynamic variables!) which showed that solutions evolving from closely spaced initial conditions would rapidly diverge into very different non-periodic trajectories. Although the implications for weather prediction are now obvious – computer power must grow exponentially in order to achieve a linear increase in prediction time – this paper was largely ignored, accumulating only ten citations by weathermen between 1963 and 1973, less than one per year and an order of magnitude less than the number of papers using the word “chaos” in its traditional sense. After 1975, however, this situation changed dramatically, as is shown by Figure 4. From 1975 to the end of 2003, Lorenz’s paper accumulated a total of about 3300 citations for an average of 114 per year.
As the importance of Lorenz’s results became recognized by the wider scientific community, the term “chaos” came to be used in a mathematical sense, implying sensitive dependence of solution trajectories on their initial conditions and the unanticipated non-periodic behaviour that Lorenz had reported a decade earlier. During the rapid rise of interest in deterministic chaos after 1975, references to Lorenz’s early paper exceeded the number of papers using the term “chaos” for several years, reflecting the fact that physical scientists were not yet comfortable with the new definition of this term.

Yet another trigger for the explosive growth of nonlinear science research in the early 1970s was the somewhat tardy interest of applied mathematicians in reaction-diffusion problems. Although the importance of these systems had been clearly established in the early 1950s through (i) the experimental studies of nerve impulse propagation by Alan Hodgkin and Andrew Huxley [244], *(6)* (ii) contemporary theoretical work on the problem of biological morphogenesis by Alan Turing [520] and (iii) related problems had been of interest to electrical engineers throughout the 1960s, reaction-diffusion problems were not considered by applied mathematicians until the early 1970s when several analyses appeared [86, 162, 271, 349, 435]. Around this time, Charles Conley began using reaction diffusion as an example of his geometrical approach to dynamics, and he encouraged several of his students to take up such studies. Thus by the mid-1970s, research on reaction-diffusion systems was deemed a respectable mathematical activity, leading to many of the publications counted in the lowest curve of Figure 1.

2 Fundamental phenomena of nonlinear science

In this section are presented broad descriptions of the three general areas of nonlinear research activity noted above: (i) chaos theory, (ii) solitons and solitary waves, and (iii) reaction-diffusion phenomena. In each case, attention is given to the connections among these three activities, the buildup of pressures leading to the explosion shown in Figure 1, and the difficulties experienced by the scientific community in focusing on the key issues and accepting the new concepts of nonlinear science. These observations are of more than mere historical interest as we face the scientific challenges of the twenty-first century and attempt to recognize other truly valuable new ideas.

2.1 Chaos theory

In the 1950s and 1960s, earlier studies by Henri Poincaré and George David Birkhoff on non-periodic solution trajectories of conservative dynamical systems [419, 46] influenced those studying the dynamics of plasmas and particle accelerators and observing “stochastic motions” [303, 304]. An independent contribution to the development of modern chaos theory was Lorenz’s above

---

*(6)*Hodgkin and Huxley were awarded the 1963 Nobel Prize in medicine for this research.
mentioned paper, which appeared in 1963 [311, 92]. In a recent book [312], Lorenz provides valuable insight into the circumstances surrounding the creation of this seminal work. Whereas his background and interests were in dynamical weather forecasting, Lorenz was in charge of a project devoted to linear statistical forecasting, exploring how the newly available digital computer could be put to use. As a key issue was to know how well such numerical tools can predict complex weather patterns, he sought a set of simple nonlinear differential equations that would mimic meteorological variations, thus providing a test example for the linear statistical approach. It was necessary to find a system with non-periodic behaviour, because the course of a periodic solution cannot be predicted without limit from knowledge of the trajectory over a single cycle.

After several failed attempts and some success with more complicated models, he arrived in 1961 at the following system of ordinary differential equations (ODEs):

\[
\begin{align*}
\frac{dx}{dt} &= \sigma(y - x), \\
\frac{dy}{dt} &= rx - y - xz, \\
\frac{dz}{dt} &= xy - bz,
\end{align*}
\]

which now bears his name. In these equations, \( \sigma \) is the Prandtl number (ratio of fluid viscosity to thermal conductivity), \( r \) represents a temperature difference driving the system, and \( b \) is a geometrical factor.

Figure 5: Sketch of the Lorenz attractor of a solution trajectory of Equations (3) with \( \sigma = 10 \), \( r = 28 \) and \( b = 8/3 \). The view is looking down from the \( y \)-axis onto the \((x, z)\)-plane.
With the parameter values indicated in Figure 5, Lorenz found that the region of phase-space occupied by the solution trajectory looks like a butterfly, with the solution jumping back and forth between nearly periodic oscillations on the wings at seemingly random intervals. Somewhat accidentally, he discovered that this behaviour requires neighboring trajectories to diverge exponentially over some regions of the phase space, a property that is described as sensitive dependence on initial conditions (SDIC) and can be detected by looking for one or more positive Lyapunov exponents (or logarithms of local rates of divergence) [565]. Although he avoided the term until the early 1980s (preferring Birkhoff’s “irregularity”), Lorenz eventually defined “chaos” to imply seemingly random behaviour of a determinate dynamical system [312].

The mid-1960s saw other publications in the general theory of chaos that added to the pressure released by the research explosion of Figure 1. One of these was the “horseshoe map” proposed by Stephen Smale in 1967 to study SDIC in a general dynamical system [489, 570]. In a sign of the poor interdisciplinary communications that prevailed in nonlinear science before the mid-1970s, the 132 references cited in this paper did not include Lorenz’s work. Smale used a topological construction to show that the irregular (or non-periodic) solutions previously noted by Birkhoff and Poincaré are generic – to be anticipated in a wider view of dynamics rather than shunned as impediments to finding analytic solutions. To this end, Smale proposed a method for distorting a two-dimensional Poincaré surface, which intersects a three-dimensional phase space with points on the surface, indicating a passage of the solution trajectory. Stretching the surface introduces SDIC, and folding it back upon itself (by bending it into the shape of a horseshoe) keeps the trajectories from escaping to infinity [570].

Also in 1967, another seminal contribution appeared with the title “How long is the coast of Britain? Statistical self-similarity and fractional dimension” by Benoit Mandelbrot [326]. Again there was no mention of Lorenz, perhaps understandably as this paper considers the strange structure of a static spatial curve (a coastline) rather than a dynamic trajectory. Mandelbrot showed that the nineteenth-century concept of a fractional dimension – lying between one for a smooth line and two for a smooth surface – is not a bizarre mathematical concept but something often observed in nature, including geographical edges and surfaces, Brownian motion, and structures of biological organisms, among others [327]. Mandelbrot’s seminal paper set the stage for interest in studies of fractal structures [283] which are now considered under the aegis of chaos theory.

In 1971, David Ruelle and Floris Takens introduced the term “strange attractor” for an irregular solution trajectory in a widely-cited reference on fluid turbulence [443] – a problem that Jerry Gollub and Harry Swinney studied experimentally [200] still without noticing Lorenz’s 1963 paper. This first came to the attention of the international physics community in 1974 in connection

---

For a film by Steve Strogatz showing several mechanical and electronic analogs of Lorenz attractors, go to http://dspace.library.cornell.edu/bitstream/1813/97/3/Strogatz+Demos.mov
with studies by John McLaughlin and Paul Martin on computational fluid dynamics [351, 352], followed by Hermann Haken’s analysis of nonlinear laser dynamics [210, 211], and in the seminal paper by Tien Yien Li and James Yorke entitled “Period three implies chaos” [301], which may be the first time that the term “chaos” appeared with its modern technical meaning. The pace of chaos research picked up in 1976, with a paper by Otto Rössler that introduced a chaotic dynamical system having about the same form as Equations (3) (but with only a single product nonlinearity) as a model for nonlinear chemical reactions [441], a two-variable discrete map by Michel Hénon [234], and a discrete version of the logistic (or Verhulst) equation by Robert May.

Often called the “logistic map”, this latter system has the simple form
\begin{equation}
    x_{n+1} = \alpha x_n (1 - x_n),
\end{equation}
where \( x_n \) is a normalized population, \( n \) is a discrete time variable, and the growth parameter \( \alpha \) corresponds to \( \lambda + 1 \) in Equation (2). Although Equation (1) provides a complete solution to the ODE defined in Equation (2), the situation for the corresponding difference-differential equation (DDE) defined in Equation (4) is more intricate and interesting.

For \( 0 \leq \alpha < 1 \), Equation (4) implies a steady-state population of zero, corresponding to \( \lambda < 0 \) in the ODE case of Equation (2). For \( 1 < \alpha < 3 \), there is a steady-state population of \( 1 - 1/\alpha \), much as in the ODE case, but this can also be viewed as a periodic solution with period 1. At \( \alpha = 3 \), this steady-state (or period 1) solution bifurcates into two solutions, each of period 2; thus \( x_n = x_{n+2} \). At \( \alpha = 1 + \sqrt{6} \), these two solutions bifurcate again into four solutions, each of period 4, and so on, where some values of the \( \alpha_k \) and the corresponding behaviours of \( x_n \) are given in Table 2. (All of these facts can be readily checked on your pocket calculator.)

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \alpha_k )</th>
<th>( \alpha_{k-1} &lt; \alpha &lt; \alpha_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>( x_n \rightarrow 0 )</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>( x_n \rightarrow (1 - 1/\alpha) )</td>
</tr>
<tr>
<td>2</td>
<td>( 1 + \sqrt{6} )</td>
<td>( x_n = x_{n+2} )</td>
</tr>
<tr>
<td>3</td>
<td>3.5440903596…</td>
<td>( x_n = x_{n+4} )</td>
</tr>
<tr>
<td>4</td>
<td>3.5644072661…</td>
<td>( x_n = x_{n+8} )</td>
</tr>
<tr>
<td>5</td>
<td>3.5687594195…</td>
<td>( x_n = x_{n+16} )</td>
</tr>
<tr>
<td>6</td>
<td>3.5696916098…</td>
<td>( x_n = x_{n+32} )</td>
</tr>
<tr>
<td>∞</td>
<td>3.569945672…</td>
<td>( x_n ) is chaotic</td>
</tr>
</tbody>
</table>

Interestingly, there is a critical value of \( \alpha \)
\[
    \alpha_\infty = 3.569945672…
\]
beyond which the solution becomes chaotic. The bifurcation diagram for this succession of solutions is shown in Figure 6, which displays one of several
“routes to chaos” [197]. In 1978, Mitchell Feigenbaum observed that the limiting ratio

$$\lim_{k \to \infty} \left( \frac{\alpha_{k+1} - \alpha_k}{\alpha_{k+2} - \alpha_{k+1}} \right) = 4.669201609102990671853 \ldots$$

(called the “Feigenbaum number”) is the same for all such discrete maps with quadratic maxima, a property called “universality” [167, 67]. It is a reflection on the controversial nature of chaos research during the 1970s that Feigenbaum struggled for two years to get this seminal work into print [168].

Figure 6: A bifurcation diagram for solutions of Equation (4). (Courtesy of Mikhail Rabinovich and Nikolai Rulkov.)

For several reasons the logistic map – Equation (4) – is an interesting mathematical object. First, it is motivated by the real-world problem of seasonal population growth, which is observed for many biological species. Second, it displays a surprisingly wide variety of dynamic behaviours under variations of a single parameter ($\alpha$). Finally, it is computationally simple. For these reasons, some refer to the logistic map as the “hydrogen atom” of chaotic modeling.

All of the above-mentioned chaotic systems (Lorenz, Rössler, Hénon, logistic map) are non-conservative or dissipative in the sense that an initial volume of phase space shrinks to zero under the dynamics, implying that some essential quantity ($Q$) is being supplied and consumed. In the publication records plotted in Figure 1, for example, $Q$ is related to the number of people who are writing papers on nonlinear science. For a nerve impulse model, $Q$ is the electrical charge stored in and released by the membrane capacitance [466]. In problems of chemical dynamics, $Q$ is the number of molecules in a chemical species, and in biological problems, $Q$ may be the number of organisms in a species.

For systems that conserve energy and so can be derived from a Hamiltonian ($H$), initial volumes of phase space are conserved under the dynamics [303, 304, 318]. In such cases, there is a body of related work collectively
called the KAM theorem which describes the onset of chaos as the perturbation from an integrable system is increased. This perspective was developed over the 1950s and early 1960s through contributions by Andrei Kolmogorov, his student-cum-colleague Vladimir Arnol’d, and Jürgen Moser. (An informative account of how KAM developed is given by Diacu and Holmes [129].)

Assuming an unperturbed system to be integrable, solution trajectories lie on nested tori in the phase space, defined by dynamics of the action-angle variables. If the perturbation destroys integrability and is controlled by a parameter $\varepsilon$, the main KAM results can be informally stated as follows [126].

**KAM Theorem:** *If an unperturbed Hamiltonian system is non-degenerate, then for a sufficiently small $\varepsilon$ most non-resonant invariant tori do not vanish but are only slightly deformed, so that in the phase space of the perturbed system there are invariant tori densely filled with conditionally periodic phase curves winding around them, with the number of independent frequencies equal to the number of degrees of freedom. These invariant tori form a majority in the sense that the measure of the complement to their union is small when the perturbation is small.*

An analytic method for detecting chaos in Hamiltonian systems that are periodically perturbed was developed by Alexey Mel’nikov in 1963 [359]. Based on ideas of Poincaré, Mel’nikov’s method can detect the existence of homoclinic chaos for sufficiently small perturbations [78]. In studies of planetary motion, KAM theory has applications to the weak perturbation of integrable two-body dynamics by a small third body, which is one of the classical problems of nonlinear science [129].

To perform numerical studies related to KAM, Michel Hénon and Carl Heiles proposed the simple system

$$H = (\dot{x}^2 + \dot{y}^2)/2 + (x^2 + y^2)/2 + x^2 y - y^3/3$$

in 1964 [235], and as one of the simplest non-integrable, energy-conserving systems known, this “Hénon–Heiles Hamiltonian” has become a standard model for exemplifying the KAM theorem. Numerical studies are conveniently carried out by defining a Poincaré surface (usually a plane) in the four-dimensional phase space $(x, y, \dot{x}, \dot{y})$ and marking this surface with a dot whenever the solution trajectory goes through. For small values of the total energy ($E$), these dots lie on closed curves, indicating that the corresponding trajectories lie on tori. At moderate values of $E$, regions of irregular motion appear between the closed curves and these regions grow with further increases in $E$ [208].

An important paper by Grayson Walker and Joe Ford appeared in 1969, which drew these developments together and brought them to the attention of the physics community [543]. Citing the works of Poincaré and Birkhoff, these authors described the KAM theorem, used numerical studies to show in detail how KAM applied to the Hénon–Heiles model and the FPU Hamiltonian [170], and discussed the implications for statistical mechanics. Viewed warily by physicists at the time of its publication, this paper has been highly
cited and is now accepted as a prophetic contribution to nonlinear science. At about the same time, Ilya Prigogine and his colleagues used models proposed by Henry McKean and by Mark Kac to show how the recognition of molecular chaos leads to sharper derivations of Boltzmann’s H-theorem for increasing entropy in a gas [233].

In 1979, Boris Chirikov considered the nonlinear map

\[ \theta_{n+1} - 2\theta_n + \theta_{n-1} = K \sin \theta_n, \]

which bears the same relation to the nonlinear pendulum ODE as Equation (4) does to the Verhulst ODE of Equation (2). As \( K \to 0 \), Equation (6) approaches an integrable limit, conserves mechanical energy, and has a family of periodic (Jacobi elliptic) functions as solutions [293]. If \( n \) is a spatial variable, Equation (6) is closely related to the system proposed in 1939 by Yakov Frenkel and Tatiana Kontorova to model the dynamics of crystal dislocations [182, 175].

Defining \( I_{n+1} \equiv \theta_{n+1} - \theta_n \), Equation (6) can also be written as the two-variable system

\[
\begin{bmatrix}
I_{n+1} \\
\theta_{n+1}
\end{bmatrix} =
\begin{bmatrix}
I_n + K \cos \theta_n \\
I_n + \theta_n + K \cos \theta_n
\end{bmatrix},
\]

which was introduced in the 1960s in a little noticed report by John Bryan Taylor to model the motion of a charged particle in a magnetic field [507] – during the years when such problems were of scant interest to the physics community. Taylor’s report is mentioned in a related study by John Greene in 1979 [199], which appeared at the same time as the widely-cited paper by Chirikov [90]; thus Equation (7) is now called the “Taylor–Chirikov map” or “standard map” [358].

As Equation (7) can be derived from a Lagrangian formulation [358], the KAM considerations apply [126]. The implications of this can be appreciated by noting from Equation (6) that the solutions will be regular for \( K \ll 1 \), which in turn implies that plots of \( \theta_{n+1} \) against \( \theta_n \) will be families of closed curves. As \( K \) is increased from zero but is still sufficiently small, many of these closed curves persist, interspersed with irregular (chaotic) regions, and as \( K \) is further increased the chaotic regions grow at the expense of the closed curves, a prediction that agrees with numerical studies [358].

From Figures 1 and 2, it is seen this burst of activity between 1976 and 1979 led directly into the current era of nonlinear science, in which more than five papers are published on chaos each day. In addition to many experimental studies, this work comprises analyses of additional theoretical models for chaos (both dissipative and Hamiltonian) [426], discussions of various routes to chaos [197], and considerations of the relationship between chaos and the complex phenomenon of turbulence – a phenomenon that was observed by Leonardo da Vinci in complex fluid flows [314, 386].

Interestingly, Lorenz was not the first to announce the discovery of deterministic chaos, as his studies throughout the 1960s revealed. In 1961, Yoshisuke Ueda, a electrical engineering graduate student working with Chiro Hayashi in Kyoto, observed that periodically driving a nonlinear oscillator would lead
to chaotic behaviour [224], but this publication was largely overlooked because it was a technical report written in Japanese, thus not accessible to western scientists, and also because the Japanese scientific community was not then ready to accept the concept of deterministic chaos. In 1958 Tsuneji Rikitake had observed paradoxically complex behaviour of a two-disk dynamo model [434], and the effects of periodically driving a simple oscillator had been investigated by Mary Cartwright and John Littlewood in 1945 [84], but they regarded non-periodic trajectories as “bad”, whereas Birkhoff had merely labeled them “irregular” [46].

Nor was observation of the difficulties with numerical weather prediction original with Lorenz – in the mid-1950s, Norbert Wiener had expressed concern with this problem [553], and in 1898 physicist William Suddards Franklin had written [177]: “Long range detailed weather prediction is therefore impossible, and the only detailed prediction which is possible is the inference of the ultimate trend and character of the storm from observations of its early stages; and the accuracy of this prediction is subject to the condition that the flight of a grasshopper in Montana may turn a storm aside from Philadelphia to New York!”. Thus even Lorenz’s butterfly metaphor was superseded.

Although not specifically concerned with weather, a strong statement on the problematic aspects of prediction had been made by Poincaré at the beginning of the twentieth century [419].

If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of that same universe at a succeeding moment. But even if it were the case that the natural laws had no longer any secret for us, we could still only know the initial situation approximately. If that enabled us to predict the succeeding situation with the same approximation, that is all we require, and we should say that the phenomenon had been predicted, that it is governed by laws. But it is not always so; it may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon.

Poincaré’s position on predictability was not derived from philosophical speculations, but on his unsuccessful attempt to solve the three-body problem of planetary motion, eventually bringing him to the mathematical conclusion that this classical problem cannot be solved.8 His reasoning was based on proving the SDIC of solution trajectories that Lorenz was to observe numerically for Equations (3) some six decades later.

Why was Poincaré’s admonition ignored by the scientific community? Why was it necessary for Lorenz to discover SDIC numerically, rather than learning

---

8The curious circumstances surrounding the award of a mathematical prize for this work by King Oscar II of Sweden and Norway in 1889 have been carefully researched and ably described by Florin Diacu and Philip Holmes [129] and by June Barrow-Green [32].
about it in his undergraduate mathematics lectures? Just as it was difficult for Nicolaus Copernicus, in the sixteenth century, to convince his colleagues that our Earth revolves around the Sun [278], it seems, scientists in the first seven decades of the twentieth century were unable to accept the idea that long-range predictions can be impossible in a well-defined ODE system.

Thus we recognize the explosive growth of interest in the subject of chaos shown in Figure 1 as an example of a Kuhnian paradigm shift. Prior to 1970, almost every scientist would have agreed that prediction is possible in principle for any determinate ODE system and in practice – given adequate computing power – for any sufficiently simple ODE system. The mathematical textbooks of the time (as I recall from my student days) supported this perspective by omitting to mention the dissenting views of Poincaré and Birkhoff, among other knowledgeable mathematicians. Since 1990, on the other hand, most scientists have come to accept low-dimensional chaotic phenomena as a typical feature of physical reality and dozens of textbooks are being used in university courses on nonlinear science to teach the basic ideas to undergraduate students. In addition, the nature of chaotic phenomena has been often explained to the general public, and executive toys that demonstrate irregular trajectories are readily available in boutiques.

Looking back on the decade-long neglect of his 1963 paper prior to its enthusiastic acclaim, Lorenz remarks [312]:

One may argue that the absence of an early outburst [of interest in chaos] was not caused by a prevailing lack of interest; it was the lack of interest. To some extent this is true, yet it may have been caused by the priorities of the leaders in the field. . . . Certainly Poincaré and Birkhoff and most other leaders did not suggest that the problems of the future would lie in chaos theory.

Prior to the mid-1970s, it appears, the concept of deterministic chaos was not yet accepted – even by the most clear-sighted of the scientific community.

2.2 Solitons and solitary waves

Hydrodynamic solitary waves (or solitons as they came to be called in the middle of the twentieth century) had been coursing up the fjords and firths of Europe since the dawn of time, but they were not scientifically discussed 1834. In August of that year, a young Scottish engineer named John Scott Russell was engaged in an urgent project. Britain’s horse-drawn canal boats were threatened by competition from the railroads, and he was conducting a series of experiments to measure the dependence of a boat’s speed on its propelling force, with the aim of finding design considerations for conversion from horsepower to steam. As chance would have it, a rope parted in his apparatus and

the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving
it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well defined heap of water, which continued its course along the channel without change of form or diminution of speed.

Russell did not ignore this unexpected event; instead, he “followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles and hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height” until the wave became lost in the windings of the channel. And, as is described in his now classic Report on Waves to the British Association for the Advancement of Science [448], he continued to study this phenomenon in wave tanks (see Figure 7), canals, rivers, and the Firth of Forth over the following decade.

Figure 7: A hydrodynamic soliton created in a wave tank by John Scott Russell in the 1830s. (a) A column of water is accumulated at the left-hand end of the tank. (b) Release of this water by lifting the sliding panel generates a solitary wave that travels to the right with velocity $v$. (Redrawn from [448].)

In the course of these studies, Russell found his “Great Wave of Translation” to be an independent dynamic entity that moves with constant shape and a speed given by

\[ v = \sqrt{g(d + h)}, \]

where $d$ is the depth of the water, $h$ is the height of the wave, and $g$ is the acceleration of gravity. Furthermore, he demonstrated that a sufficiently large initial mass of water will fission into two or more independent solitary waves (each moving with its own speed), and that solitary waves cross each other “without change of any kind.”

Although one might expect Russell’s careful studies to be readily accepted by the scientific community, this did not happen. Already in 1841, George Airy, the Astronomer Royal, had inferred from his theoretical investigations that a solitary wave could not be obtained from the hydrodynamic equations and so
concluded that Russell’s interpretations of his observations must be in error [9], an opinion that was soon supported by independent theoretical studies of young George Stokes [496]. In retrospect, the fact that the solitary wave travels faster than low-amplitude waves is important in these polemics, as it reinforces the idea that it is an independent dynamic entity.

Eventually, confirmation of Russell’s canal observations were presented by Henri Bazin in France in 1865 [34], and in 1872 Joseph Boussinesq published a long and detailed analysis showing that the solitary wave was indeed a possible solution of the hydrodynamic equations [65], a conclusion supported by Lord Rayleigh, Adhémar Saint-Venant and John McCowan [429, 450, 346]. (Very readable accounts of these differences of opinion in the context of nineteenth-century wave science have recently been published by Robin Bulloch [71], Alex Craik [101], Alexandre Filippov [171], and Oliver Darrigol [108].)

Finally in 1895, Diederik Kortweg and Gustav de Vries derived the following KdV equation for Russell’s shallow water waves:

\[
\frac{\partial u}{\partial t} + 
\frac{\partial u}{\partial x} + 
\kappa \frac{T}{\partial x} + 
\varepsilon \frac{\partial^{3} u}{\partial x^{3}} = 0, 
\tag{9}
\]

where \( \kappa = \sqrt{g d} \), \( \varepsilon = c(h^{2/6} - T/2\rho g) \) is a dispersion parameter, and \( \gamma = 3c/2h \) is a nonlinear parameter and \( T \) and \( \rho \) are respectively the surface tension and density of water [276]. Using the recently developed elliptic functions, they obtained a nonlinear periodic solution of Equation (9) (which they called “cnoidal” for the Jacobi cn function), and in the limit of infinite wavelength this solution becomes Russell’s solitary wave, with just the properties that he had recorded and published five decades earlier.

As the amplitude and slope of the solitary wave are decreased to zero, the last two terms on the left-hand side become small, and the wave speed reduces to \( \sqrt{g d} \) in accord with Equation (8). At larger values of \( h \), the dispersion introduced by the \( \varepsilon \)-term is balanced by the nonlinearity of the \( \gamma \)-term for an exact solution with the “smooth and well defined” shape that was observed experimentally by Russell. By this time, however, Russell was resting in his grave and interest in his solitary wave had waned, as we see from the following facts:

(i) based on the views of Stokes, the importance of the solitary wave was discounted in the Encyclopedia Britannica of 1886 [171], (ii) Horace Lamb’s opus on hydrodynamics allots a mere 3 of 730 pages to the solitary wave [286], and (iii) there were only about two dozen citations of reference [448] from its publication in 1845 to the beginning of the nonlinear science explosion in 1970 [171]. (In fairness to the nineteenth-century British scientific community, it should be noted that Russell’s case for the solitary wave was not aided by the publication of several misguided claims in a posthumous book that was prepared from his unpublished manuscripts in 1885 [449].)

---

9Contributing to this dispute may have been a subtle disdain felt by mathematically oriented wave theorists at Cambridge University for papers published in the recently organized British Association for the Advancement of Science, which was dominated by engineers.

10This equation also lies buried in Boussinesq’s tome [65].
Nonlinear science came back into the focus in the early 1950s when Fermi, Pasta, and Ulam (FPU) conducted numerical studies of thermalization on a one-dimensional mass-spring system, finding an unexpectedly frequent recurrence of the initial state.\(^{11}\) Attempts to understand and explain this anomaly led Zabusky and Kruskal to approximate the FPU mass-spring chain by KdV, discovering in 1965 that the initial conditions would fission into a spectrum of solitary waves which then reassembled themselves into (approximately) the original state after an unexpectedly short interval \([577]\). To emphasize the particle-like properties of these dynamically independent solitary waves – which confirmed the intuition of Russell in contrast to the calculations of Airy and Stokes – they were called “solitons”.

In 1967, Kruskal and his colleagues (Clifford Gardner, John Greene and Robert Miura) considered KdV written in the normalized form

\[\frac{\partial u}{\partial t} - 6u \frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0,\]

with a solitary wave solution

\[u_s(x, t) = -\left(\frac{v}{2}\right) \text{sech}^2 \left(\sqrt{\frac{v}{2}}(x - vt - x_s)\right),\]

and discovered an unexpected result. If a general solution of Equation (10) is introduced as a time-varying potential in the Schrödinger equation

\[\frac{d^2 \psi}{dx^2} + [\lambda + u(x, t)] \psi = 0,\]

the eigenvalues of Equation (12) are independent of time \([192]\), with each eigenvalue corresponding to a particular soliton in the total solution. Furthermore, the time evolution of \(\psi\) is given by

\[\frac{d \psi}{dt} = \left[-4u \frac{\partial^3 \psi}{\partial x^3} + 3u \frac{\partial \psi}{\partial x} + 3u_x \psi\right].\]

Equations (12) and (13) imply that the Schrödinger equation scattering data (eigenvalues and reflection coefficient) computed for the initial column of water in Figure 7(a) can be used to compute the time evolution \(u(x, t)\) as follows \([384, 396]\). (i) Given an initial disturbance \(u(x, 0)\), calculate the Schrödinger equation scattering data \(SD(0)\). (ii) Use the asymptotic form (large \(|x|\)) of Equation (13) to find the scattering data at a later time \(SD(t)\). (iii) Determine \(u(x, t)\) from \(SD(t)\), by using an inverse scattering calculation. As a diagram, this inverse scattering transform (IST) is:

\(^{11}\)Nick Metropolis told me that this discovery was accidental. Because on a short time scale the initial energy seemed to proceed toward a thermalized state (confirming the motivating intuition), early computations were terminated before the energy returned to its original configuration. One afternoon, they got to talking and unintentionally left MANIAC running, leading to their unexpected observation.
Writing Equations (12) and (13) respectively as $L \psi = 0$ and $\psi_t = M \psi$, Lax generalized this IST calculation to the class of operators for which $L_t = [M, L]$, where $L$ and $M$ are called a “Lax pair” [297]. Lax’s formulation is now used to obtain exact solutions for a broad class of nonlinear wave systems. For certain initial conditions on KdV, these exact solutions comprise $N$-solitons and correspond to members of a family of reflectionless potentials previously known for the Schrödinger equation [264].

Interestingly, the IST can be viewed as a nonlinear generalization of the Fourier transform (FT) in the sense that IST reduces to FT as the nonlinearity (or wave amplitude) is reduced to zero [467]. An important feature of the soliton concept was (and is) to focus attention on fully nonlinear solutions (expressed as hyperbolic, Jacobi elliptic or theta functions or numerically) rather than starting with Fourier expansions and calculating interactions among component modes. Although the FT approach is theoretically valid in principle and analytically practical in some (quasilinear) cases, a fully nonlinear approach is usually easier to implement and empirically more illuminating. This difference of perspective was at the core of Russell’s disagreement with Airy and Stokes, and it has arisen in other applications to be discussed below (fluxons on long Josephson junctions and slinky modes on reverse-field-pinch plasma confinement machines). During the 1970s, there were several experimental studies establishing the relationship between KdV solitons and shallow water waves[576, 219, 220, 500, 56].

An example of an exact 2-soliton solution to Equation (10) is

\[
\begin{align*}
&\psi \rightarrow \psi(x, t) \\
&\text{SD(0)} \rightarrow \text{SD}(t)
\end{align*}
\]

(14) $u_2(x, t) = -12 + 4 \cosh(2x - 8t) + \cosh(4x - 64t)
\]

which was published by Zabusky in 1968 based on the GGKM work [192, 573]. Figure 8 shows a plot of this function, confirming after an interval of 153 years Russell’s observation that two solitary waves pass through each without change of shape. From a careful examination of Equation (14), however, it can be seen that the trajectory of the faster soliton is displaced slightly ahead and that of the slower is pushed backward by the interaction – a feature of the phenomenon that Russell missed.

In defining the term soliton, there is a difference of attitude between mathematicians and physicists. The former tend to reserve this term for solitary-wave solutions of PDEs (like KdV) for which an IST can be formulated, leading to exact $N$-soliton solutions. Some physicists, on the other hand, prefer to call any solution of an energy-conserving (Hamiltonian) system that exhibits particle-like properties a soliton, without worrying about the existence of exact $N$-soliton solutions. Thus the physicists’ soliton might be a hydrodynamic vortex (see Figure 31) or an elementary particle that changes its character under...
collisions with other particles (see Figure 19). This question arose at a soliton conference in Gothenburg, Sweden in the summer of 1978. Speaking as a particle physicist, Tsung-Dao Lee argued for the broader usage, and when Alan Bishop (a condensed-matter physicist) objected, Lee suggested that the mathematicians’ special solitary waves could be called “aristocratic solitons” – to which Bishop replied that those without the special properties should then be called “plebian solitons”.

This was the international meeting where Alexander S. Davydov announced his theory of solitons on the alpha-helical regions of natural proteins. In a continuum approximation, these solitons are governed by the nonlinear Schrödinger (NLS) equation, which can be written in normalized form as

$$\frac{i}{4} \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2|u|^2 u = 0$$

where $u(x, t)$ is the complex amplitude of an oscillating field. A solitary wave solution of this equation is

$$u_s(x, t) = a \exp \left[ i \frac{v_e}{2} x + i \left( \frac{a^2}{4} - \frac{v_e^2}{4} \right) t \right] \sech \left[ a(x - v_c t - x_0) \right],$$

which has an envelope moving at envelope velocity $v_e$ and a carrier wave moving with velocity $v_c = \left( v_e^2/2 - a^2/4 \right)$; thus the envelope velocity and the wave amplitude are independent parameters. This is evidently more complicated than a KdV soliton, but just as in Equation (9), the second term of Equation

---

12 Organized by plasma physicist Hans Wilhelmsson, this meeting was typical many interdisciplinary nonlinear science conferences that were held in the late 1970s.
(15) introduces dispersion which is exactly balanced by the nonlinearity of the third term, resulting in a stable solitary wave with particle-like properties. As was shown by Zakharov and Shabat in 1972, the initial value problem for Equation (15) can be solved by an IST method, leading to $N$-soliton formulas corresponding to Equation (14) [580]. The NLS equation has been of interest to applied scientists since the mid-1960s as a model for deep water waves [39], nonlinear optics [261], nonlinear acoustics [217, 381, 504], and plasma waves [503]. Since the early 1970s, it has been used as the fundamental description of pulses on an optical fibre [221], among other technical applications [324]. Although both the KdV equation and the NLS equation were first considered in the context of particular engineering applications, their importance is more general. KdV, for example, arises whenever one studies a low-frequency, nonlinear wave system with lowest-order representations of dispersion and nonlinearity. Similarly, NLS is a lowest-order description of a high-frequency nonlinear wave system; thus these two equations are generic.

A physically motivated model is the sine-Gordon (SG) equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} = \sin u,$$

which was formulated by Frenkel and Kontorova in 1939 in connection with their studies of crystal dislocations [182]. Investigations of this system were carried on in the postwar years by Alfred Seeger and his collaborators, who were aware of the special properties of non-destructive solitary-wave collisions [239, 482]. In this application, $u(x, t)$ is a real variable describing the dynamics of a row of atoms in a spring-mass approximation, where the $\sin u$ term on the right-hand side of the equation accounts for the periodic forces induced by a (static) neighboring row. Another application of this equation is as a model for the simple mechanical system shown in Figure 9, where the first term accounts for the spring (elastic) forces between adjacent pendula and the other two terms describe the dynamics of the coupled nonlinear pendula [460]. The solution of Equation (17) shown in Figure 9 corresponds to one of the following functions

$$u_s(x, t) = 4\arctan \left[ \exp \left( \pm \frac{x - vt - x_0}{\sqrt{1 - v^2}} \right) \right],$$

where the “±” signs account for the fact that the solution can wind in either the clockwise or counterclockwise direction. These are solitons in the strict (mathematicians’) sense because Equation (17) was shown in 1974 to possess an IST formulation, allowing exact analytic solutions of initial value problems [2, 502]. As is demonstrated in Figure 10, SG solitons experience varying amounts of Lorentz contraction as they move with different speeds, a property that stems from the fact that Equation (17) is invariant under the Lorentz transformation: $x \rightarrow x' = (x - vt) / \sqrt{1 - v^2}$ and $t \rightarrow t' = (t - vx) / \sqrt{1 - v^2}$.

Although seemingly specific, SG also models diverse physical phenomena [72], including the dynamics of domain walls in ferromagnetic and ferroelec-
Figure 9: A single soliton (kink) on a mechanical model of Equation (17).

Figure 10: Stroboscopic photograph showing Lorentz contraction of a kink soliton on a mechanical model of Equation (17) [460]. (Dissipative effects cause the kink to slow down as it moves to the right, whereupon its size increases.)

elastic materials [133], the propagation of splay waves on biological (lipid) membranes [169], self-induced transparency of short optical pulses [287], the propagation of quantum units of magnetic flux (called fluxons) on long Josephson (superconducting) transmission lines [481], and slinky-modes in reversed-field pinch plasma confinement machines [139], and as a one dimensional model for elementary particles [35, 160, 438, 415]. If the kink is associated with an electron and an antikink with a positron, kink-antikink annihilation on the mechanical system of Figure 9 provides a simple model for electron-positron decay into cosmic (electromagnetic) radiation.

Unaware of this previous work [239] but motivated by the Korteweg–de Vries results [171], John Perring and Tony Skyrme undertook numerical studies of SG as a classical model for fermions and published the following exact kink-kink and kink-antikink solutions in 1962 [415]:

\[
\begin{align*}
    u_{k_{kk}}(x, t) &= 4 \arctan \left[ \frac{v \sinh(\frac{x}{\sqrt{1 - v^2}})}{\cosh(\frac{vt}{\sqrt{1 - v^2}})} \right], \\
    u_{k_{k\overline{k}}} (x, t) &= 4 \arctan \left[ \frac{\sinh(\frac{vt}{\sqrt{1 - v^2}})}{v \cosh(\frac{x}{\sqrt{1 - v^2}})} \right].
\end{align*}
\]
As further evidence of the poor communications among nonlinear scientists in the 1960s, Kruskal and Zabusky did not know about this work when they wrote their seminal paper on the soliton [574].

The kink-antikink solution of Equation (20) takes an interesting form if the velocity parameter \( (v) \) is allowed to be imaginary. For example, setting \( v = i\omega/\sqrt{1 - \omega^2} \) with \( \omega < 1 \), this equation becomes the stationary “breather” [2] or “bion” [88]

\[
\begin{align*}
\text{u}_{\text{bre}}(x, t) = 4\arctan \left[ \frac{\sqrt{1 - \omega^2}}{\omega} \frac{\sin \omega t}{\cosh \sqrt{1 - \omega^2} x} \right],
\end{align*}
\]

which is plotted in Figure 11 for \( \omega = \pi/5 \). Under the elementary particle analogy, this is a one-dimensional model of positronium.

Figure 11: A stationary breather solution of SG plotted from Equation (22) with \( \omega = \pi/5 \).

Taking advantage of the fact that SG is invariant under a Lorentz transformation, this stationary breather can be boosted to a breather moving at speed \( v_{\text{c}} \), which is described by the formula:

\[
\begin{align*}
\text{u}_{\text{mbr}}(x, t) = 4\arctan \left[ \frac{\sqrt{1 - \omega^2}}{\omega} \sin \left\{ \omega \left( t - v_{\text{c}} x \right) \right\} \right. \\
\left. \left[ \frac{\cosh \sqrt{1 - \omega^2} (x - v_{\text{c}} t)}{\cosh \sqrt{1 - \omega^2} (x - v_{\text{c}} t)} \right] \right],
\end{align*}
\]

In the low-amplitude limit, Equation (22) approaches the NLS soliton in Equation (16), where both expressions are independent dynamic entities that exhibit both wave and particle properties.

Although these analytic results are interesting, it is important to consider how robust they are under various perturbations, of which five types are particularly important [350].

34
1. A constant term can be added to the right-hand side (RHS) of Equation (17). This preserves the Hamiltonian structure and acts as a force, pushing a kink in one direction and an anti-kink in the other. An inhomogeneously driven SG system exhibits low-dimensional chaos [147].

2. The RHS function can be changed to a general periodic function \( f(u) = f(u + u_0) \). This preserves the Hamiltonian structure and Lorentz invariance (LI), but the IST formulation is destroyed. Thus there are no more \( N \)-soliton formulas or breathers, but kinks and antikinks survive. These may tend to move in one direction or another, depending on the nature of \( f(u) \).

3. The RHS function can be changed to a polynomial function with at least three zeros – say \( u(1 - u^2) \), the “phi-four” equation [81]. Again, the Hamiltonian structure and LI survive, and there may be kink or antikink solutions, undergoing transitions between zeros of \( f(u) \).

4. Dissipative perturbations can be added, such as LHS terms proportional to \( \partial u / \partial t \) or to \( -\partial^3 u / \partial x^3 \partial t \). In this case, the Hamiltonian structure is lost but an energy balance can be established between energy input to a kink (or antikink) in the above three cases and the effects of the dissipation, leading to a steady propagation speed that depends on the parameters of the equation [350].

5. Because Equation (17) provides only an approximate description of the mechanical system of coupled pendula shown in Figure 9, a corresponding difference-differential equation (DDE) might better be used. Under such a perturbation, the Hamiltonian structure is preserved but LI and the IST formulation are lost, although the qualitative phenomenon of Lorentz contraction is observed under weak discretization, as is seen in Figure 10. More generally, a time-independent DDE system corresponds to the standard map of Equation (6), which can exhibit spatial chaos.

Another participant in the 1978 Gothenburg meeting was Morikazu Toda who had independently developed a soliton theory for a particular form of the mass-spring lattice during the 1960s [517, 518]. This system is shown in Figure 12, comprising equal unit masses connected by nonlinear springs of potential energy \( U(r) \) where \( r \) is the change in distance between adjacent masses from its resting value at which the spring energy is a minimum. Newton’s second law thus implies the following set of DDEs:

\[
\frac{d^2 r_n}{dt^2} = \frac{dU(r_{n+1})}{dr_{n+1}} - 2 \frac{dU(r_n)}{dr_n} + \frac{dU(r_{n-1})}{dr_{n-1}},
\]

where \( n \) is an index running along the chain. For the special spring potential

\[
U(r) = \frac{a}{b} \left( e^{-br} + br - 1 \right),
\]

35
this Toda lattice (TL) system is exactly integrable through an IST method [174],
and a TL soliton has the form

\[ r_{n,s} = \frac{1}{b} \log \left[ 1 + \sinh^2 \kappa \operatorname{sech}^2 \left( \kappa n \pm t \sqrt{ab} \sinh \kappa \right) \right], \tag{25} \]

where the “±” signs allow waves to go in either direction. These are compression waves, one of which shown on the lower part of Figure 12. Interestingly, this compression wave travels at velocity

\[ v = \sqrt{ab} \sinh \kappa / \kappa \tag{26} \]
lattice points per second, which is faster than the speed \( (\sqrt{ab}) \) of small amplitude waves on the lattice as for KdV solitons in Equation (8). This again suggests that the solitary wave has escaped from the confines of a linear representation.

Figure 12: The upper figure shows a Toda lattice at rest. The lower figure shows the Toda-lattice (TL) soliton given in Equation (25).

If the nonlinear spring potential is altered from the special form of Equation (24), the property of exact integrality is lost, but stable supersonic compression waves (as shown in Figure 12) survive under a wide class of realistic interatomic potentials in solids, liquids and gasses [190]. The fact that these solitary waves can be both stable and supersonic is in accord with the 1822 observation of Captain William Edward Parry that the sound of a cannon firing can travel faster than the command to fire it – a phenomenon that had been noted by Russell [449].

Thus we have encountered four nonlinear wave systems (KdV, NLS, SG, and TL), all of which arose in applications prior to 1970 and all of which share the property of being exactly integrable under an IST method. At this point, the reader may wonder how such striking results could have escaped notice of the mathematics community. In fact they didn’t [239, 289]. Proceeding in parallel with the above described developments in physics and engineering was a line of research in differential geometry, going back to Ferdinand Minding who began the study of surfaces of constant negative Gaussian curvature (pseudospherical surfaces) in 1839 [363], and to Edmond Bour [64] and others [57, 159].
who discovered that such surfaces can be described by the PDE

\[
\frac{\partial^2 u}{\partial \xi \partial \tau} = \sin u,
\]

which is identical to SG under the independent variable transformation: \( \xi = (x + t)/2 \) and \( \tau = (x - t)/2 \). Building upon work by Luigi Bianchi [42] and Sophus Lie [305], Albert Bäcklund showed in 1883 that an infinite number of solution surfaces to Equation (27) can be generated through successive applications of the transformation [21]

\[
\frac{\partial u_1}{\partial \xi} = 2a \sin \left( \frac{u_1 + u_0}{2} \right) + \frac{\partial u_0}{\partial \xi},
\]

\[
\frac{\partial u_1}{\partial \tau} = \frac{2}{a} \sin \left( \frac{u_1 - u_0}{2} \right) - \frac{\partial u_0}{\partial \tau}.
\]

In this Bäcklund transformation (BT), \( u_0(\xi, \tau) \) is a known solution of Equation (27) and \( u_1(\xi, \tau) \) is a new solution that can be obtained by integrating a first-order ODE system \([436]\). In 1892, Bianchi showed that two successive BTs (with different values of the parameter \( a \)) commute, which leads directly to algebraic constructions of Equations (19) and (20) [43]. A decade earlier, Gaston Darboux had published a related transformation [107] that can be applied in a similar way to the linear equations of an IST formulation for KdV [18]. Although several other PDE systems were considered from these perspectives prior to 1914 – some of which were rediscovered during the flood of soliton activity during the 1970s [239] – this promising line of activity died out after 1920, possibly because many of the young researchers were killed in the First World War [288]. In 1936 a summary of this early work was published by Rudolf Steuerwald, which would have been of great interest to Frenkel, but he didn’t know about it [495].

Now it is known that in all four cases (KdV, NLS, SG, and TL) the IST formulation is equivalent to the existence of a corresponding BT and from these BTs, \( N \)-soliton formulas and countably infinite sets of conservation laws can be derived, just as for the corresponding ISTs. \( N \)-soliton formulas for these four systems can also be constructed directly using a method devised by Hirota in 1971 [242].

As previously noted, both the KdV equation and the NLS equation are generic, arising when dispersion and nonlinearity are accounted for to lowest order. The other two cases (SG and TL) are models of specific physical structures (see Figures 9 and 12) with solitary wave solutions that are robust under changes in the particular analytic functions – the “\( \sin u \)” for SG and “\( U(\tau) \)” for TL – on which the IST formulation depends. All of these solitons or solitary waves can be viewed as Newtonian particles, which are acted on by forces derived from perturbing influences. Interestingly, these physically motivated models also arise in a variety of applications. In the face of such important
empirical and theoretical results, neglect of the solitary wave concept by five
generations of the scientific community between 1845 and 1970 is a tale of lost
opportunities.

As I write, the gruesome effects of the tsunami generated by an earthquake
off the coast of Sumatra on 26 December 2004 are clearly in mind. Figure
13(a) shows a computer simulation by Kenji Satake of this wave, providing
a vivid example of Russell’s Great Wave of Translation. Thus tsunami energy
is transmitted without dispersion over long distances at several hundred kilo-
metres per hour in accord with Equation (8), and from its long, smooth shape,
a tsunami is difficult to detect on the open ocean. Upon approaching a con-
tinental shelf, tsunamis slow down, gain amplitude and fission into several
independent solitary waves, as is expected from Russell’s tank experiments.

Internal ocean waves also provide examples of oceanic solitons that are
well described by Equation (9) with parameters given by $c = \sqrt{gh_1 \Delta \rho / \rho}$,
$\varepsilon = ch_1h_2/6$, and $\gamma = -3c/2h_1$, where $\Delta \rho$ is the density difference of the
two layers, $h_1$ is the depth of the upper layer and $h_2 \gg h_1$ is the depth of
the lower layer [398, 405]. Figure 13(b) shows a satellite photograph of such
internal wave solitons in the Andaman Sea, which are visible from above due
to the surface rip phenomenon. Whereas tsunamis are sporadically initiated by
earthquakes, internal waves appear regularly every 12 hours and 26 minutes,
with greater intensities during the full and new moons; thus they are driven
by tidal flows.

With tsunamis such a familiar and frightening phenomenon, why were
Russell’s observations and conclusions ignored by the scientific community
for so many decades? One answer is that until recently scientists were con-
tioned – through their collective experiences with electromagnetism, acous-
tics, and quantum theory – to think about dynamics in terms of linear nor-
mal modes, which obey the principle of superposition. Since the discovery
of normal modes by Johann and Daniel Bernoulli in the eighteenth century
[171], there evolved a general belief that any nonlinear solution can be resolved
into its normal mode components, and then (if necessary) interactions among
these modes can be subsequently computed. Thus physics and engineering
students were taught a variety of methods (Fourier transforms, Laplace trans-
forms, Green functions) which are appropriate for analyses of linear systems,
but they learned almost nothing about nonlinear analyses. When its compo-
nent modes interact, the shape of a large amplitude wave would seem to be-
come distorted, as was asserted by Airy and Stokes, incorrectly rendering Rus-
sell’s empirical observations of a solitary wave theoretically impossible. From
this normal-mode perspective, nonlinearity came to be widely regarded as an
analytic nuisance – an inconvenience to be avoided rather than a feature of
realistic dynamics to be embraced for its unique properties.

At a deeper level, Russell’s difficulty in obtaining recognition for the impor-
tance of his solitary-wave concept is an example of what Kuhn has described
as the resistance of the scientific community to revolutionary ideas [279]. Truly
new ideas, Kuhn pointed out, are often recognized by those who are not fully
Figure 13: (a) A computer simulation of the deadly tsunami (an example of Russell’s Great Wave) in the Indian Ocean on 26 December 2004, 1 hour and 50 minutes after it was launched by an earthquake source (green area) near Sumatra. The red colour indicates that the water surface is higher than normal, while the blue colour means that it is lower, and the darker the colour, the larger the amplitude. See http://staff.aist.go.jp/kenji.satake/animation.html for an animated version of this figure. (Courtesy of Kenji Satake, National Institute of Advanced Industrial Science and Technology, Japan.) (b) Apollo–Soyuz photograph of the Andaman sea, showing surface evidence of internal waves [398]. [The location of (b) is shown as a small rectangle in (a).] (Courtesy of Alfred Osborne.)
indoctrinated into the prevailing paradigms and resisted by those who are. As an engineer, Russell undertook wave studies with fresh eyes, unclouded by details of analytic manipulations, and he became thoroughly familiar with the experimental facts. Both Airy and Stokes, on the other hand, were theorists, invested in the perspectives of their previous publications, the intricacies of which provided opportunities for them to insist that solitary waves would necessarily experience distortion. As outstanding scientists who made many important contributions to nineteenth-century research, their considerable influence probably inhibited young British scientists from pursuing Russell’s revolutionary idea of the solitary wave as an independent dynamic entity. Although Russell’s empirical observations were independently confirmed by Bazin [34] and the theoretical validity of his work was subsequently established in papers by Boussinesq [65], Rayleigh [429], Saint-Venant [450], and Korteweg and de Vries [276], it didn’t get into the textbooks. Only after his solitary-wave paradigm had been forced upon those engaged in numerical computations during the 1960s, was Russell’s work taken seriously by the scientific community.

2.3 Reaction-diffusion systems

Soon after Russell’s Report on Waves appeared, Hermann Helmholtz published measurements of the propagation speed of nerve impulses along the sciatic nerve of frog, which carries signals from the spinal cord to a leg muscle [232, 462]. Influenced by comments of Newton suggesting that the propagation of nerve activity is like that of light [385], nerve impulse speed was generally thought to be very large, but Helmholtz used a clever device to obtain a value of only 27 metres per second, which is close to recently observed speeds, and in 1868, Julius Bernstein used an even more ingenious measurement apparatus to record the pulse-like shape of an impulse [40]. Since these experiments, therefore, neuroscientists have needed to explain why the speed of an impulse was so small – a question they would puzzle over for a century.

Several answers to this question were offered as new information accumulated, corresponding to Kuhn’s “prescientific” phase of theoretical activity in which a prevailing paradigm has not yet been established. In 1902 Bernstein proposed the “membrane hypothesis” which assumed that a breakdown of the nerve’s surface resistance plays a key role in impulse conduction [41]. At about the same time, Robert Luther demonstrated a propagating chemical reaction to a meeting of the German Society for Applied Physical Chemistry, pointing out that the wave speed was of order

\[ v \sim \sqrt{D/\tau} \]  

(29)

where \( \tau \) is the reaction time for an energy releasing process and \( D \) is a diffusion constant [316]. Because both diffusion constants and reaction times can vary widely for physical systems, he suggested, such a reaction-diffusion pro-
cess provides a credible explanation for the modest speed of a nerve impulse.\textsuperscript{13} Then in 1914, young Edgar Douglas Adrian stated the \textit{all-or-nothing} principle of nerve impulse excitation \cite{adrian1914} – which implies the existence of a threshold for excitation – and went on to pioneer applications of the newly invented vacuum-tube amplifier to studies of nerve impulse dynamics, observing a \textit{refractory period} of diminished excitability following the passage of an impulse. In 1925, Ralph Lillie presented the passive-iron-wire model of impulse conduction, in which a length of wire initially rests in a nitric acid bath, prevented from dissolving by a thin surface oxide layer. Upon being disturbed (by scratching, say), the oxide layer collapses, allowing an impulse of current to flow, and in accord with Bernstein’s membrane hypothesis this disturbance propagates along the wire, after which the oxide layer reconstitutes itself to reestablish the resting state \cite{lillie1925}. Thus the iron-wire model exhibits both a threshold for excitation and a refractory period. In 1936, John Zachary (“J.Z.”) Young recognized that two long and large cylindrical structures running along the back of the common squid (\textit{Loligo vulgaris}) are in fact nerves \cite{young1936}. The diameters of these nerves are about 0.5 mm, allowing electrophysiologists of the day to directly measure both the time course of the transmembrane voltage and the membrane permeability (or electrical conductivity), as Kenneth Cole did in the classic oscilloscope photograph of a nerve impulse shown in Figure 14 \cite{cole1937}.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{nerve_impulse_figure.png}
\caption{An early cathode-ray oscilloscope photograph of a nerve impulse. The dots on the lower edge indicate time in milliseconds, increasing to the right. The solid line shows the time course of the transmembrane voltage ($V$), becoming increasingly positive inside the axon. The wide band shows the transmembrane permeability which is large (small membrane resistance) during the impulse. (Courtesy of Kenneth Cole.)}
\end{figure}

In this photograph, the solid line indicates the transmembrane voltage, which is measured inside the axon with a glass microelectrode (filled with a

\textsuperscript{13}Although Walther Nernst was both present at the meeting and interested in Luther’s ideas, his prescient observation was neglected for several decades.
saline solution) as shown in Figure 15. The wide band in Figure 14 records the transmembrane permeability (or electrical conductivity), measured by an ac bridge. Electromagnetic analysis of the axon dynamics shows that the transmembrane voltage is governed by the nonlinear diffusion equation [462]

\[
\frac{\partial^2 V}{\partial x^2} + r c \frac{\partial V}{\partial V} = r j_{\text{ion}},
\]

where \( r \) is the series resistance to longitudinal current flow, \( c \) is the membrane capacitance, and \( j_{\text{ion}} \) is the transmembrane ionic current (comprising components of both sodium and potassium ions), all per unit length of the axon. Equation (30) is nonlinear because both components of \( j_{\text{ion}} \) are nonlinear functions of the transmembrane voltage. To see how this goes, note first that throughout the dynamics the sodium-ion concentration remains large outside the axon, whereas the potassium-ion concentration is large inside (see Figure 15); thus an impulse develops as follows.

![Diagram showing a squid axon and the measurement of internal voltage using a glass microelectrode (not to scale).](image)

- At rest, the membrane resistance is very high for both sodium and potassium ions (see Figure 14) and \( j_{\text{ion}} = 0 \).
- As \( V \) increases from its resting value by about 25 mV, membrane permeability to sodium ions increases sharply, allowing positively-charged sodium ions to flow inward. This inward flowing charge further increases \( V \) in a positive feedback loop that drives \( V \) to a maximum value of about 100 mV, which is the peak value of \( V \) shown on Figure 14.
- At this peak value of \( V \), the sodium-ion permeability falls back to zero and the potassium-ion permeability rises. This allows positively-charged potassium ions to flow outward, carrying \( V \) back to its resting value. After this refractory period, the nerve is ready to carry another impulse.

Thus, along the leading edge of a squid nerve impulse, sodium ions flow into the axon, whereas potassium ions flow outward along the trailing edge.
of the impulse. In 1952, Hodgkin and Huxley (HH) quantified these nonlin-
erar dependencies of the transmembrane ionic currents and carried through
a travelling-wave analysis of Equation (30), giving values for impulse speed,
time courses of $V$ and of transmembrane conductivity that are in accord with
measurements as in Figure 14 [244]. (Done on a mechanical hand calculator in
the last of the precomputer days, this computation was a tour de force.)

Although there is a superficial resemblance between the nerve impulse of
Figure 14 and the hydrodynamic soliton shown in Figure 7(b), they are very
different entities for the following reasons. First, Equation (9) is an energy-
conserving (Hamiltonian) system, whereas Equation (30) conserves nothing.
Second, the speed of a hydrodynamic solitary wave in a uniform channel de-
pends on the initial conditions (the heap of water on the left-hand-side of Fig-
ure 7(a)), while the speed of a nerve impulse on a uniform axon depends on the
parameters of the system. Third, energy-preserving solitary waves obey New-
ton’s second law ($F = ma$) under perturbations, accelerating and decelerating
in response to applied forces. A nerve impulse, on the other hand, doesn’t
remember its previous speed and responds directly to the local properties of
the medium. Finally, a soliton or energy-conserving solitary wave expresses a
dynamic balance between the opposing effects of nonlinearity and dispersion,
whereas a nerve impulse establishes a dynamic balance between the dissipa-
tion of energy and the rate of its release through a nonlinear mechanism. It is
important to be aware of these differences, because physicists sometimes incor-
rectly refer to nerve impulses as “solitons” under their loose definition [171].

Interestingly, the shape of the leading edge of the impulse shown in Fig-
ure 14 is similar to the growth curve of Figure 1 – an initial exponential rise
followed by a leveling off (or saturation). Its speed can be computed by repre-
senting the initial sodium-ion current as the cubic polynomial

\[ \dot{J}_{Na,d} = g \frac{V(V - V_1)(V - V_2)}{V_2(V_2 - V_1)}, \]

where $g$ is the conductance per unit length of the membrane at the peak of the
impulse (the largest value of the band in Figure 14), $V_2$ is the peak value of
voltage, and $V_1$ is a threshold voltage above which inward sodium current be-
gins to flow. Although neuroscientists didn’t know how to solve this nonlinear
diffusion equation until the late 1950s, a closely related equation

\[ \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = u(u - a)(u - 1), \]

was studied in the context of flame-front propagation in 1938 – the same year
that Cole took the classic photograph of Figure 14 – by Yakov Zeldovich and
David Frank-Kamenetsky (ZF), who derived the exact travelling-wave solution
[582]

\[ u_{FW}(x - vt) = \frac{1}{1 + \exp [(x - vt)/\sqrt{2}^2]}, \]

43
with speed $v = (1 - 2a)/\sqrt{2}$. If this work had been known to Cole, he could have immediately expressed the speed of a squid nerve impulse in unnormalized form as

$$v = \sqrt{\frac{g}{re^2}} \left( \frac{V_2 - 2V_1}{\sqrt{2V_2(V_2 - V_1)}} \right),$$

where the parameter values for the “standard axon” studied by HH are given in the following table [466].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>diameter ($d$)</td>
<td>0.476</td>
<td>mm</td>
</tr>
<tr>
<td>$g$</td>
<td>0.0108</td>
<td>mhos/cm</td>
</tr>
<tr>
<td>$r$</td>
<td>$2.0 \times 10^4$</td>
<td>ohms/cm</td>
</tr>
<tr>
<td>$c$</td>
<td>$1.5 \times 10^{-7}$</td>
<td>F/cm</td>
</tr>
</tbody>
</table>

Taking $V_1$ (the voltage at which sodium current begins to flow) as 25 mV and $V_2$ (the maximum impulse voltage) as 100 mV, this formula gives an impulse velocity of 20.0 m/s, which compares favorably with the HH measured value of 21.2 m/s at 18.5 °C [244]. Furthermore, dimensional analysis shows that $\sqrt{g/re^2} \propto \sqrt{d}$, and my measurements on about two-dozen axons of *Loligo vulgaris* at 18.5 °C gave $v = 20.3\sqrt{d/0.476}$ m/s to an accuracy of about five percent [466].

Equation (34) can be viewed as the product of a parameter factor $\sqrt{g/re^2}$ times a structure factor $(V_2 - 2V_1)/\sqrt{2V_2(V_2 - V_1)}$. The structure factor was found in 1938 by ZF, and the parameter factor was proposed by Luther in 1906 (see Equation (29)), where the diffusion constant in Equation (30) is $D = 1/rc$ and the response time for the onset of sodium-ion current flow is $\tau = c/g$.

Clearly, it would have been useful for neuroscientists of the 1930s to know what was happening in contemporary studies of flame-front propagation. This lack of scientific communication is particularly distressing because in the middle of the nineteenth century Michael Faraday presented a series of Christmas Lectures to young people on “The Chemical History of a Candle”, in which he stated [165]:

> “There is no better, there is no more open door by which you can enter into the study of natural philosophy than by considering the physical phenomena of a candle.”

From Faraday’s lectures – which can be profitably read by mature scientists a century and a half later – one sees that the candle is a metaphor for nerve impulse propagation which Helmholtz vainly sought, no doubt while seated at a table under its light.

Although the breakdown in communications between the areas of neuroscience and combustion science which led to the decades-long neglect of the ZF analysis of Equation (32) was bad enough, a more egregious oversight occurred between two branches of biological science. In 1937 the famed geneticist Ronald Fisher proposed a means for wavelike advance of advantageous genes

---

14 An engraving of Faraday presenting his Christmas Lectures was on a recent British 20 pound note.
through a two-dimensional population [173], and in the same year travelling-wave propagation on this system was independently discussed in a long article by Kolmogorov and his colleagues, who changed the RHS term in Equation (32) to $u(u - 1)$, finding a minimum wave speed of 2 [274]. Evidently, this can also be viewed as adding a diffusion term ($\partial^2 u / \partial x^2$) to the Verhulst equation with which we began this review, but in spite of the obvious model and famous names involved, this work was also overlooked by the neuroscience community for several decades [462]. Why was this so?

In the early 1960s, wave propagation on reaction-diffusion systems became of interest to the engineering community for several reasons. First was the proposal by Hewitt Crane of the \textit{neuristor} as an electronic line that can carry a signal without attenuation, as does a nerve axon [102]. Such a system, it was pointed out, offers a means for realizing any Boolean logic circuit and thus a computer of arbitrary power [475]. Second, the invention of several solid-state diodes with nonlinear differential conductance [453] brought the design of novel neuristor structures within the bounds of feasibility [461]. Finally, electronics has always been an important aspect of electrophysiology, so it was natural for electrical engineers to become involved with the scientific and modeling aspects of the nerve conduction problem [379, 459]. During the 1960s, neuristor research was carried on in the electrical engineering communities of several countries, including Denmark, Japan, Russia and the United States [461].

Strangely, it was not until the early 1970s that applied mathematicians became aware of the reaction-diffusion process, their collective interest triggered by a highly-cited paper of Henry McKean which put into the mathematical literature ideas about the nerve model of Jin-ichi Nagumo that had been known to engineers for a decade [379, 349]. Why did the applied mathematics community come so late to this important area of research? Was this sudden interest part of the general activity in nonlinear dynamics that erupted in the 1970s?

Among the checkered annals of nonlinear science, there is the sad story of Boris Belousov. As a research scientist in the Soviet Ministry of Health, he sought an inorganic equivalent to the Krebs cycle and in 1950 discovered a chemical reaction (involving citric acid, bromate ions and a catalyst) that oscillates periodically between an oxidized and a reduced state as indicated by changes in colour. Attempts to publish this work over a period of six years failed because benighted editors deemed it in violation of the Second Law of Thermodynamics. Presently, these refusals are difficult to understand considering: (i) Luther’s above mentioned public demonstration of an oscillating chemical reaction to a German scientific society in 1906 [316], (ii) Alfred Lotka’s publication of a theoretical analysis of propagating chemical reactions in 1920 [313], (iii) The prompt experimental confirmation of Lotka’s theory by William Bray [66], and (iv) Implications of chemical reactions in theories of biological pattern formation by D’Arcy Wentworth Thompson [513] and by Turing [520]. Although encouraged by Simon Shnoll to continue his work, Belousov became discouraged, quit science and died a bitter man in 1970 – at the dawn of the explosion of interest in nonlinear science. At the end of this decade, he was
posthumously awarded the Lenin Prize by the Soviet Union.

In 1961, Shnoll suggested Belousov’s work to a research student, Anatol Zhabotinsky, who modified the chemistry (changing citric to malonic acid and adding ferroin sulfate as a colour indicator) to obtain the “Belousov–Zhabotinsky (BZ) reaction” [137]. In 1964, Zhabotinsky confirmed Belousov’s results [583], and in 1970 Albert Zaikin and Zhabotinsky published photographs of self-oscillating ring waves in this reaction [581]. Shortly thereafter, Arthur Winfree demonstrated that the BZ reaction supports spiral waves (shown in Figure 16) which propagate outward from organizing centers.15

![Figure 16: In a BZ reaction, two-dimensional waves spiral outward from organizing centers, intersecting to form cusps. (Courtesy of A.T. Winfree.)](image)

Examples of two-dimensional reaction-diffusion waves in nature include prairie fires, lichens [467], social amoeba (slime molds), and fairy rings of mushrooms [23]. These circular and spiral waves are characterized by two qualitative features: the formation of cusps where wave fronts meet, and smooth curves (see Figure 16). The former property is typical of wave processes having refractory zones behind the front. Stable wave fronts are typically smooth curves because the velocity of a front decreases as the local value of the curvature increases.

In quantitative terms, the velocity ($v$) of a two-dimensional front (as in Figure 16) is given by:

$$v = \frac{DA}{λ^2}$$

Where $D$ is the diffusion coefficient, $A$ is the activation energy, and $λ$ is the wavelength.

15Winfree told me that he first saw spiral waves on an iron-wire screen in Nagumo’s laboratory at the University of Tokyo during the mid-1960s. It was while trying to prove that spiral waves cannot exist in a continuous medium that he discovered them in the two-dimensional BZ reaction. For films demonstrating some oscillating chemical reactions, see http://www.williams.edu/Chemistry/epeacock/
Figure 16) is given by [467]

\[ v = v_0 - D/R, \]

where \( D \) is the diffusion constant for the process, \( v_0 \) is the zero-curvature velocity of a line wave, and \( R \) is the local radius of curvature of the wave front. In three spatial dimensions, the corresponding formula is

\[ v = v_0 - D/2R, \]

where \( v_0 \) is the zero-curvature velocity of a plane wave. In three space dimensions, solutions of reaction-diffusion systems can take the form of a scroll, which unwinds continually from a straight line, and a scroll ring, which unwinds from a circular organizing center, as is shown in Figure 17.

Figure 17: A computer generated plot of a scroll ring. (Courtesy of A.T. Winfree.)

The three-dimensional scroll ring displayed in Figure 17 is more than an analytical curiosity. The human heart is (among other things) a reaction-diffusion medium for electrochemical activity that can be disrupted by the spontaneous formation of an internal scroll ring. If this “fibrillating” behaviour is not immediately suppressed, the owner of the heart may succumb to “sudden cardiac death” [561, 562].

Generalization of the above considerations leads to a nonlinear geometrical optics for reaction-diffusion systems, which has been developed by Oleg Mornev [373]. From this perspective, a modified version of Snell’s law of refraction can be derived, in addition to conditions for blocking of a wave front at a boundary where the diffusion constant undergoes an abrupt change. This blocking condition, in turn, is useful in finding conditions for nerve impulse propagation to fail at branching regions of axons and dendrites [466].

Although Equation (36) holds for stable three-dimensional wave fronts in relatively simple reaction-diffusion processes, such fronts are always stable,
especially when gravitational or convective forces come into play. In his above-mentioned lectures for young people, Faraday noted that flames often flicker in a spatially chaotic pattern \[165, 258\], a phenomenon that was studied by Lord Rayleigh \[430\] and later by Geoffrey Taylor \[506\] and is now called the Rayleigh–Taylor instability \[337\].

After the speed of nerve impulse conduction was first measured in 1850, it was more than a century before the phenomenon was unambiguously explained as a reaction-diffusion process, because research was impeded by at least three factors. First, we have again seen the extreme balkanization, the lack of communication between areas of scientific inquiry that appear to be different but are nonetheless related by their underlying mathematical structure. Second, researchers were led astray by the qualitative character of linear diffusion – which is not at all wavelike – to anticipate corresponding properties of reaction-diffusion systems. How else can one explain that a scientist with the theoretical ability of Helmholtz did not formulate Equation (32) and find the travelling-wave solution of Equation (33)? Finally, the fact that the applied mathematicians ignored reaction-diffusion studies for two decades after the Hodgkin–Huxley research was published suggests that they were not comfortable with the concept of localized travelling waves before 1970.

2.4 Summary

In the above discussion, nonlinear science has been divided into three parts, but there are connections among the realms of low-dimensional chaos, energy-conserving solitary waves (soliton systems), and reaction-diffusion systems, as is shown in Figure 18.

- The upper oval of the diagram includes chaotic systems that can be energy-conserving or dissipative. Dissipative systems that display low-dimensional chaos include the Lorenz, Rössler, and Hénon systems, and the logistic map, among many others. Chaotic systems that conserve energy include many versions of the $N$-body problem of planetary motion, the Hénon-Heiles system, and the standard map, among many others. If the standard map is modified by adding a term in the second derivative of another independent variable, it becomes a discrete version of the sine-Gordon (SG) equation which was studied by Frenkel and Kontorova \[182\].

- At the base of this diagram are circles indicating two classes of PDE systems, one (on the left) supporting solitons and energy-conserving solitary waves and the other (on the right) supporting travelling-wave solutions of RD systems, like the nerve impulse. Solitary-waves solutions of energy-conserving systems have the properties of Newtonian particles, moving at any of a continuum of possible speeds, depending on their initial conditions. As was assumed by Aristotle on the other hand, solitary-wave solutions of RD systems move at fixed speeds that are determined by local properties of the medium. If solitonic PDEs are slightly
Figure 18: Interrelations among low-dimensional chaotic systems, reaction-diffusion (RD) systems and integrable (solitonic) systems.

perturbed by dissipation or energy inputs, their solitary-wave solutions move like Newtonian particles in weakly viscous media. If such dissipative perturbations are sufficiently large, however, the Aristotelian limit of an RD system is reached [374]. For example, an increase of such driving and dissipative terms transforms the SG equation to the Nagumo model for nerve impulse conduction [379, 466].

- The left-hand edge of the diagram interpolates between low-dimensional Hamiltonian chaos and integrable systems (including soliton systems), which sit on opposite ends of a gray scale that is governed by the KAM theorem. The structural stability of solitons under Hamiltonian perturbations is an interesting example of the strength of the KAM theorem.

- On the right-hand edge of the diagram, a similar relationship obtains between dissipative low-dimensional chaos and the solitary waves of RD systems as the degree of dispersion is increased, although the KAM theorem does not apply. In RD systems of two or three spatial dimensions, for example, spatial chaos is often observed in addition to stable solitary waves. Interestingly, both dissipative chaos and RD solitary waves take advantage of dynamic instability (positive feedback) to establish their qualitative behaviours — in the first case, the attractor is strange and in the second case a travelling-wave front emerges to balance the release and dissipation of energy.

There are several variations on these themes that are not included in Figure 18. For example, (i) Chris Eilbeck and his colleagues have shown that an
inhomogeneously driven SG system exhibits low-dimensional chaos [147], (ii) spatial chaos is not included [327], (iii) turbulence is also not included [386], (iv) the many integrable systems now known are only suggested by the circle in the lower left-hand corner [318], among many other examples of nonlinear phenomena [469]. Nonetheless, Figure 18 gives some idea of the holistic nature of the “lexicon” – to use Kuhn’s term [280] – that is currently shared by practitioners of nonlinear science. Recognition of the interrelations among these limits (low-dimensional chaos, particle-like solutions of energy-conserving PDEs, and localized waves of activity on RD systems) has implications for research in many branches of science, some of which are sketched in the following section.\footnote{A variety of interesting films demonstrating features of Figure 18 nonlinear dynamics can be found at http://www.pojman.com/NLCD-movies/NLCD-movies.html}

3 Applications of nonlinear theory

In this section, some of the research activities that have benefited and are benefiting from the insights of nonlinear science are described in greater detail. Both historical and mathematical connections among the various results are noted, paying particular attention to those related to the nonlinear science revolution described in Section 1.

3.1 Cosmology

As he lay dying in 1543, Copernicus opened an important chapter in the history of astronomy with the publication of his De Revolutionibus Orbium Caelestium, which proposed that the Earth is a planet rotating around the Sun rather than the fixed centre of the universe [278]. Although this idea had been proposed by Aristarchus of Samos in the third century BCE and seems trivial to us now – familiar as we are with the amazing photographs that space vehicles have sent back from far corners of the solar system – it was not obvious to astronomers of the sixteenth century. In addition to the fact that our Earth does not appear to move, Ptolemy’s analytic formulation of Aristotle’s cosmology predicted all observations of celestial motions over many centuries by Greek, Islamic and European astronomers, and a geocentric universe was in accord with the Christian myths that was so vividly described by Dante Alighieri in his classic La Divina Commedia.

The basic structure of the Ptolemeic universe comprised two primary spheres, an inner one being the surface of the earth and an outer (stellar) sphere carrying the stars around us every day. Within the stellar sphere were seven lesser spheres – the “orbium” of Copernicus’s title – inhabited by the seven moving bodies: Saturn, Jupiter, Mars, Sun, Venus, Mercury, and Moon (in decreasing distances), all of which were driven in their motions by the daily rotation of the outer (stellar) sphere. Beyond the stellar sphere was nothing, as the Ptolemeic universe was finite.
Copernicus had long been concerned with the “problem of the planets” – when Mercury, Venus, Mars, Jupiter and Saturn appear the brightest, their motions through the heavens seem to cease and reverse directions for a while before resuming more regular westward paths. Although the Ptolemaic astronomers could describe and predict this retrograde motion through a well-defined system of deferents, epicycles, ecliptics and equants, there was an *ad hoc* character of these explanations that can be avoided by assuming Earth to be a planet lying between Venus and Mars, rotating on its polar axis each day, and revolving around the Sun. Within a century, Johannes Kepler used the careful observations of his colleague Tycho Brahe to show that planets could be more accurately assumed to follow elliptical orbits about the Sun, with a line joining a planet to the Sun sweeping out equal areas in equal times. Galileo Galilei’s application of the telescope to celestial observations then revealed the moons of Jupiter and the phases of Venus, adding further data in accord with the heliocentric model. Thus the stage was set for Isaac Newton – born just a century after the death of Copernicus and within a year of Galileo’s death – to propose a self-consistent dynamical model of the universe in his *Principia Mathematica* (entitled to complement René Descartes *Principia Philosophiae*). Newton was truly standing “on the shoulders of giants” [223], as the work leading to the *Principia* comprised the efforts of an impressive international group, involving essential contributions from Greece (Aristotle and Ptolemy), Poland (Copernicus), Denmark (Brahe), Germany (Kepler), Italy (Galileo), France (Descartes) and Britain (Newton), among several others.

Although it has linear aspects in the vector addition of multiple forces and in the linear relationship between an effect (change in motion) and its efficient cause (a force) through the Second Law, Newton’s theory of gravity is inherently nonlinear because of the \(1/r^2\) attractions between masses. In *Principia*, Newton solved the first modern nonlinear problem – the two-body problem of planetary motion – by showing that planets must move about the Sun in elliptical orbits according to Kepler’s laws.

Successful solution of the two-body problem of planetary motion led to consideration of a three-body problem (the Sun, a planet and its moon, say) and more generally to an \(N\)-body problem (the entire solar system), which were to exercise mathematicians for some two centuries until the above mentioned proof by Poincaré that the three-body problem cannot be solved by using first integrals to reduce the dimension of the system [127]. These efforts were not wasted, however, as several exact solutions for special initial conditions were found [129], leading to a broad appreciation for the functions that can be analytically expressed, and there were significant developments in perturbation theory (PT), under which the solution to a nonlinear problem is progressively better approximated by a hierarchical set of tangential linear problems [269, 350, 493]. Successes of PT include a prediction of the return of Halley’s comet in 1758, the discovery of the planet Neptune in 1846 (from observations of unexplained variations of the orbits of known planets) [128], and ever more precise calculations of the precession of the point where Mercury’s ellipse comes closest to the Sun (its perihelion), which by the end of the
1900s was measured to rotate by 5600 seconds of arc per century (sa/c) and calculated via PT to be 5557 sa/c, a difference of 43 sa/c.

According to Aristotle’s geocentric cosmology, motion was observed to be directed toward or away from the centre of the universe (Earth), whereas Newton’s cosmology supposed that the laws of motion are independent of location in an infinite universe. The special relativity theory (SRT) – proposed independently by Albert Einstein [148] and Poincaré [420] in 1905 – and Hermann Minkowski’s space-time continuum [364] assumed that physical laws are the same on systems moving in straight lines at constant speeds. Thus two observers who are experiencing relative motion will agree on their respective measurements of the differential interval
\[ ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2. \]

Under the Lorentz group of transformations, this quadratic form can be written more generally in tensor notation as
\[ ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \]
where summation is taken over taken over repeated indices and \( \mu, \nu = 0, 1, 2, 3 \) specify the independent variables \((t, x, y, z)\) [501]. In four dimensional Euclidean space, \( g_{\mu\nu} \equiv \eta_{\mu\nu} = \text{diag} (-1, +1, +1, +1) \) – a flat metric tensor (where units have been chosen to make the speed of light equal to unity).

As formulated independently by Einstein and by David Hilbert late in 1915 [149, 216, 240], general relativity theory (GRT) follows from the further assumption that the laws of physics are the same on uniformly accelerating frames, a condition that can be satisfied if the metric tensor is curved. In tensor notation, the relevant set of ten nonlinear PDEs is
\[ R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu}, \]
where \( G \) is Newton’s gravitational constant.\(^{18}\) Called the “energy-momentum tensor,” \( T_{\mu\nu} \) on the right-hand side of this system is zero in a vacuum with non-zero components generated by the presence and motions of matter. (Specifically, the components of \( T_{\mu\nu} \) represent the flux of the \( \mu \)th component of momentum in the \( \nu \)th direction [83].) The “Einstein tensor” on the left-hand side is constructed from the “Ricci tensor” \( R_{\mu\nu} \) and its trace \( R \), which depend nonlinearly on the components of \( g_{\mu\nu} \) and the first two derivatives of these components with respect to the independent variables. Dynamics enter from the condition that small test particles will move along shortest paths (geodesics) of the curved space.

In other words, GRT says that matter causes space to be curved and motions of matter are along geodesics of the curved space. If \( T_{\mu\nu} = 0 \), the space is flat (\( g_{\mu\nu} = \eta_{\mu\nu} \)), and test particles move in straight lines with constant speed as in the empty space of Galileo and Newton. If \( \eta_{\mu\nu} \) is perturbed by a small function of the space variables \( \Phi(x, y, z) \ll 1 \), Equations (37) approximate those for a particle moving in a potential \( \Phi \) according to Newton’s Second Law [83].

\(^{17}\)Questions about the priority in formulating this equation have recently been addressed by Leo Corry, Jürgen Renn and John Stachel [99].

\(^{18}\)If the speed of light is not normalized to unity, the RHS factor in his equation is \( 8\pi G/c^4 \).
Without assuming special symmetries, fully nonlinear solutions of Equations (37) are difficult to obtain, but progress is now being made with numerical studies.19

Just as it was difficult for astronomers of the sixteenth century to accept the geocentric cosmology of Copernicus, the concept of curved space was difficult for many of Einstein’s colleagues to imagine, but a Euclidian alternative is even more problematic. If one considers four-dimensional space-time to be constraints on motions in a higher-dimensional Euclidian (flat) space-time, the dimension of that space must be ten. (This is because the number of independent elements in a symmetric $n \times n$ array is $n(n + 1)/2$, so the number of independent elements in $g_{\mu\nu}$ is ten.) Presently, GRT is widely accepted by the physics community for several reasons including these [483]: (i) GRT accounts for the above-mentioned disagreement of 43 $\text{s}a/c$ between calculated and measured values of the rotation rate of Mercury’s perihelion. (ii) GRT correctly predicts the bending of light rays from distant stars as they pass by the Sun. (iii) The predicted red shift of spectral lines in gravitational fields has been confirmed, as has the influence of gravity on clocks. (iv) Predictions of gravitational radiation are in accord with observed rates of orbital decay of neutron binary stars. (v) GRT rests on the assumption that gravitational and inertial masses are equal, which has been confirmed to an accuracy of about one part in $10^{13}$.

If $T_{\mu\nu} = 0$ in Equation (37), it reduces to $R_{\mu\nu} = 0$, which has gravity-wave solutions. The first serious attempt to detect such waves was mounted by Joseph Weber in the 1960s who designed and developed detectors comprising large (one ton) vibration-isolated cylinders [547]. Now there are five detector installations in operation around the globe (in Australia, Italy, Japan, and the United States), which are based on laser interferometry – with direct detection of a gravity wave generated by a cosmic event expected within a decade [123].

As we have seen in the previous section, the lexicon of nonlinear science includes two new concepts: localization of dynamical variables or activity and chaotic (irregular) behaviours of dynamic systems. How do these concepts fare in the context of gravitational theories?

- **LOCALIZATION.** (i) Stars, planets, moons, and comets are all examples of localized dynamic entities, with structures that can be understood from the mutual nonlinear attractive forces of Newton’s theory and the repulsive forces of compressed matter. On a larger scale, stars agglomerate into galaxies with masses of the order of $10^{11}$ to $10^{12}$ times that of our Sun and having elliptical or spiral shapes [238].

(ii) Before his untimely death in 1916, Karl Schwarzschild assumed a static and spherically symmetric solution of Equations (37) and showed that the curvature tensor near a mass $m$ becomes singular at the radius $r_s = 2Gm/c^2$ [458]. Thus if $m$ is compressed to within a “Schwarzschild
radius”, it will be impossible for a test particle leaving the surface to escape. (The Schwarzschild radius is about 1 cm for a body with the mass of Earth and about 3 km for our Sun.) In 1967, the term “black hole” was coined by John Wheeler for a cosmic entity with its mass thus concentrated, several of which have recently been observed [282].

(iii) Assuming plane-wave symmetry reduces the number of independent variables to two and greatly simplifies the structure of the Ricci tensor $R_{ij}$, allowing analytic expressions for one-dimensional black holes to be obtained [69, 330]. Also the above noted integrable property of special soliton equations (KdV, NLS, SG, etc.) permits the analytic construction of curvature tensors that satisfy reduced versions of Equation (37) [135, 136]. This work is connected with the above noted nineteenth-century studies of the relationship between nonlinear PDEs and curved surfaces [64, 57, 159, 363, 495].

(iv) Under GRT, our universe may or may not be finite, depending on the average mass-energy density $\rho$ in relation to a critical density $\rho_{\text{crit}} = 3H_0^2/(8\pi G)$, where $H_0$ is the observed rate of expansion (Hubble parameter). Some recent data suggest that the ratio $\rho/\rho_{\text{crit}}$ is close to unity [284].

- **CHAO**S. (i) Although the behaviour of our planetary system seems to have been regular over the past twenty-five hundred years, the fact that the $N$-body problem cannot be solved implies that some of these motions are chaotic. Examples are provided by Hyperion (a moon of Saturn) and Phobos and Deimos (moons of Mars), all of which tumble irregularly, and chaotic behaviour is also found in trajectories of the moons of Jupiter [142, 377, 564].

(ii) With the development of powerful computers over the past few decades, it has become possible to integrate the nonlinear equations of the solar system for several million years, leading to the discovery that the motions of the inner planets are susceptible to chaos with characteristic times of about 5 million years [142]. Thus there are difficulties in predicting the Earth’s orbit beyond a few tens of millions of years, making climate reconstruction over geological time scales problematic [296].

(iii) Finally, the orbit of Halley’s comet is chaotic with a characteristic time of about 29 returns (2200 years), placing the earliest observations of this phenomenon beyond retrodictability [91].

As Thomas Kuhn emphasized, the Copernican revolution was among the most important events in the history of Western science, involving a Gestalt-like switch in collective thought with implications far beyond those of astronomical theory [278]. In addition to the obvious religious aspects, this revolution changed the concept of motion from Aristotle’s picture of elements

---

20Note that there are three independent components of $g_{ij}$ in a curved two-dimensional space-time system which can be viewed as constraints on motion in a three-dimensional Euclidian space. This is the image that is often used to present the concept of curved space in elementary introductions to general relativity [223].
impelled to move toward or away from the centre of a finite universe to the Newtonian picture of isolated massive bodies moving in an infinite universe with constant speed along straight lines. Galileo and Newton assumed that such motion would continue until and unless acted upon by efficient causes – gravity, friction, or other mechanical forces – thereby introducing modern mechanics.

The Einsteinian revolution replaced Newton’s problematic action-at-a-distance forces by mass-dependent variations in the geometry of space-time, as described by the system of nonlinear PDEs in Equations (37). Although the Newtonian formulation of mechanics was accepted for about two centuries and led to striking advances in many realms of science, it is amusing to note that Einsteinian mechanics returns to two key Aristotelian ideas: motion determined by the local nature of space and the possibility of a finite universe [278].

At the dawn of the twenty-first century, the task of computing nonlinear solutions of Equations (37) and comparing them with astronomical data over very large intervals of space and time continues to occupy some of the best minds in science, offering hope of learning more about how the nature of our universe depends upon the matter within it. But what happens on a small scale? What is matter?

### 3.2 Nonlinear theories of matter

During the decade that Einstein was struggling to formulate GRT, Gustav Mie proposed a theory of matter in which the newly discovered electron emerges naturally from a nonlinear version of Maxwell’s electromagnetic equations [360]. In this study, Mie also sought a connection between matter and the force of gravity [273].

Following an early demonstration that Maxwell’s equations can be derived from an energy functional, Mie went on to define a world function (Φ) as a functional depending on electric field intensity (E), magnetic flux density (B) and the four components of the electromagnetic potential (A). Requiring Φ to depend only on the variables η = √|B|^2 - |E|^2 and χ = √|A|^2 - |ϕ|^2 insures Lorentz invariance, and with Φ = -η^2/2 this formulation reduces to the standard Lagrangian density for the linear electromagnetic equations. Augmenting the linear theory to Φ = -η^2/2 + aχ^6/6 leads to a static, spherically-symmetric electric potential (satisfying r^2/rE = 0) of the form

\[ \phi(r) \approx \frac{(3r_0^3/a)^{1/4}}{\sqrt{r^2 + r_0^2}}. \]

Setting (3r_0^3/a)^{1/4} = e/4π (the electronic charge) yields a spherically symmetric model for the electron with a radius of about r_0 and electric potential π(r) → e/4πr as r/r_0 → ∞. The Lorentz invariance built into the theory permits this localized solution to move with any speed up to the limiting velocity of light under appropriate Lorentz contraction.
Although Mie managed to derive a model elementary particle that emerges from a nonlinear version of Maxwell’s equations, his theory has several difficulties, including the following [273]. First, the choice of $\Phi(\eta, \chi)$ is not well motivated, as it could depend on other powers of $\eta$ and $\chi$ and products of these powers. Ideally, Mie would have had this structure determined by the analysis. Second, the presence of nonlinear terms in the analysis means that the superposition property is lost, especially near the centre of the particle, making more complicated solutions difficult to compute. Third, Mie’s method of putting gravitational attraction into the theory changes properties of the vacuum (dielectric permittivity and magnetic permeability), thereby altering the velocity of light, which disagrees with SRT. Finally, Mie’s assumption of a scalar gravitational potential allows inertial and gravitational masses to differ.

Now largely forgotten, Mie’s theory was esteemed in its day by Max Born, who showed the Göttingen mathematicians how Mie’s treatment of energy functionals for electromagnetism was related to Lagrangian mechanics [59, 273] and brought his results to the attention of Hilbert, who used the concept of a variation principle in his formulation of GRT [98, 240]. Einstein was initially unimpressed with Mie’s work, noting the inadequate treatment of gravity and so overlooking his seminal suggestions concerning variation principles and the fundamental nature of matter, but Hermann Weyl used some of Mie’s ideas in his 1921 discussion of “fields and matter” [550].

During the 1920s, theoretical physicists took two divergent paths. On one hand, Einstein and his colleagues sought a formulation that would unify GRT with the Maxwell equations (MEs), just as the MEs had unified electric and magnetic phenomena during the nineteenth century [198]. From the structure of Equation (37), this approach suggests that elementary particles be represented geometrically as little wrinkles in space. On the other hand, many physicists were involved with formulating quantum theory (QT) as a linear means for computing probabilities of finding atomic particles with given speeds and locations. Thus by the end of the 1920s several unifications were under consideration: (i) GRT with MEs, (ii) GRT with QT, (iii) MEs with QT, and (iv) a unification of GRT, MEs and QT – all of which have implications for the fundamental nature of matter. Hubert Goenner has written an excellent review of these unification efforts [198], including an extensive bibliography and many quotations from correspondence [198]. From this reference, one learns of an early suggestion by Rudolf Förster to include MEs under GRT by making $g_{\mu\nu}$ in Equation (37) nonsymmetrical, thereby adding six additional components which can represent the electromagnetic fields $\mathbf{B}$ and $\mathbf{E}$. Although Einstein was unenthusiastic about Förster’s approach, he initially liked Theodor Kaluza’s idea of assuming reality to comprise four spatial dimensions plus one of time [259]. On more physical grounds, Born began to wonder how all the energy of a particle could be contained in a small space [60], and Wolfgang Pauli claimed that it was unphysical to define electromagnetic fields within particles because the concept of a test particle is no longer valid. By the end of the 1920s, interestingly, Einstein had come to believe that particles should be represented by singularity-free solutions of field equations.
Mie’s approach to the nonlinear nature of matter was revived by the discovery of electron-positron creation by Carl Anderson in 1932. In this observation, a sufficiently energetic electromagnetic wave (cosmic ray) is transformed into an electron and its positively charged sibling, a positron, establishing an empirical link between the two manifestations of energy. Born and Leopold Infeld returned to Mie’s formulation, eliminating the $\chi$-dependence and choosing the nonlinear Lagrangian density (LD)

\begin{equation}
    \mathcal{L} = E_0^2 \sqrt{1 + \frac{(B^2 - E^2)}{E_0^2} - \frac{E^2}{E_0^2}},
\end{equation}

which reduces to the standard linear LD when the fields are small compared with $E_0$. They again found a spherically symmetric model electron with $E$ everywhere finite although the electric displacement exhibits a singularity at the origin [63]. Interestingly, a plane-wave reduction of the Born–Infeld (BE) theory (to one space dimension plus time) leads to a system with soliton-like properties [89, 479]. While it would be interesting to test a prediction of light-light scattering under the BE theory with interacting light beams, the estimated value of $E_0$ is about $10^{23}$ V/m – eight or nine orders of magnitude above the field intensities that can presently be obtained by focussing the light beams from high-power lasers.

Erwin Schrödinger became interested in Born’s nonlinear electromagnetic theory as early as 1935 [456] and continuing through the 1940s when – as founding director of the Dublin Institute of Advanced Studies – he attempted to move physics research toward nonlinear science [372]. During this period, Schrödinger had an extended transoceanic correspondence with Einstein, in which continuing attempts to connect nonlinear particle theories with GRT were unsuccessful. Having spent a decade of intense effort formulating GRT, Einstein was concerned about the difficulties of finding the correct nonlinear particle theory among myriad possibilities; yet near the end of his life he expressed agreement with Mie’s motivating ideas in the following words [152].

What appears certain to me, however, is that in the foundation of any consistent field theory, the particle concept must not appear in addition to the field concept. The whole theory must be based solely on partial differential equations and their singularity-free solutions. … (If) a field theory results in a representation of corpuscles free of singularities, then the behaviour of these corpuscles in time is determined solely by the differential equations of the field.

Research on nonlinear theories for elementary particles in the 1950s was largely that of Robert Finkelstein, who studied nonlinear spinor fields as electron models [172] and of Louis de Broglie, who had first proposed that material particles can have wave properties, which led to Schrödinger’s famed formulation of quantum mechanics. After teaching physics from conventional perspectives over three decades, de Broglie – inspired by David Bohm who had recently questioned the standard interpretation of QT [53, 54] and by the above-mentioned views of Einstein – returned to an idea he had first suggested
in the mid-1920s: the “double wave” theory. In a nonlinear version of this theory developed in the 1950s, de Broglie proposed that a particle should be more properly described by two waves: a fictitious $\psi$-wave that obeys the linear PDE of standard QT and an objectively real $\varphi$-wave that obeys a nonlinear PDE qualitatively similar to that of Mie or of Born andInfeld [119, 120]. Under de Broglie’s formulation, $\varphi = U e^{i\theta}$ and $\psi = \Psi e^{i\varphi}$, whereupon the condition that $\theta = \varphi$ (except in a small region surrounding the particle) allows the fictitious $\psi$-wave to guide the singularity of the nonlinear $\varphi$-wave which comprises the real particle. The statistical nature of standard QT enters because the guidance process is assumed to be rendered stochastic by a background noise field.

This causal interpretation of QT is related to modern soliton studies in several ways. (i) Ascribing ontological reality to a lump of energy that is localized by a nonlinear PDE is in accord with Russell’s recognition that hydodynamic waves can have particle-like solutions [448] and with other manifestations of the same idea that became evident in the nonlinear explosion of the early 1970s [479]. Interestingly, de Broglie observes that his localized $\varphi$-wave “may be compared to the theory of ‘solitary waves’ in Hydrodynamics, which exhibits certain similarities to it.” (ii) The idea of a guiding wave is related to the inverse scattering transform (IST) method of soliton theory, under which the dynamics of a localized solution of a nonlinear PDE are determined by the solution of an associated linear PDE system (see [3, 192, 297, 384]). (iii) A one-dimensional example of the nonlinear equation sought by de Broglie is the nonlinear Schrödinger (NLS) equation described above, where the soliton given in Equation (16) is governed by the solution of an associated IST and has both particle-like and wave-like properties.

In 1993, Peter Holland published a review and analysis of the de Broglie–Bohm theory [247], finding it a viable alternative to conventional QT. In accord with Kuhn’s description of the resistance by established scientists to novel ideas, the response by the Copenhagen establishment, Holland writes, “was generally unfavourable, unrestrained and at times vitriolic.” Heisenberg, for example, was disturbed by the asymmetrical treatment of position and momentum in the new theory, viewed the possibility of discerning particle orbits as “superfluous ideological superstructure” and compared proposals to test the validity of conventional interpretations of QT as “akin to the strange hope that . . . sometimes $2 \times 2 = 5$.” Such remarks may have discouraged young researchers from following this line.

In the 1960s, the Derrick–Hobart theorem was introduced [125, 243], which shows that many of the nonlinear models proposed for elementary particles [161, 229, 437] will be unstable to contraction in more than one spatial dimension; in other words, localized states can release energy by becoming smaller. Awareness of this mode of instability increased interest in nonlinear PDE systems with topological constraints like the sine-Gordon (SG) equation, which was studied (as noted above) by Skryme [415] and also by Ugo Enz [161]. From

---

21 Reading de Broglie’s book on Nonlinear Wave Mechanics in 1960 helped to awaken my interest in nonlinear wave theory.
inspection of the mechanical model shown in Figure 9, it is evident that kink solitons of SG have a finite rest energy that is permanently conserved. Skyrme originally supposed that the topological invariance of SG can be generalized to three space dimensions, and this turns out to be so for model particles called skyrmions, which are immune to the Derrick–Hobart instability. Some of these topologically stable particle models are shown in Figure 19 for increasing values of the topological charge [331, 499].

Beginning in 1975, papers on nonlinear particle theories began to appear in Physical Review D (PRD) – the flagship journal of elementary-particle theory in the United States. A quantification of this growing interest is provided by Figure 20, which plots the annual number of papers that used the term “soliton” in the title, key words or abstract. Some features of this figure can be understood by considering two complementary approaches to particle studies by the physics community.

- **POINT-PARTICLES.** In accord with Pauli’s concerns, fundamental particles are modeled as mathematical points. This approach is used in the Standard Model [100], which assumes leptons and quarks as elementary particles, unifies strong, weak and electromagnetic forces, and agrees with a substantial amount of empirical data. Disadvantages are that the masses and charges of the particles must be put into the theory as experimentally determined parameters, and unwanted infinities appear in certain analytic expressions, although these are sometimes removed using renormalization theory [523].

- **EXTENDED PARTICLES.** Following Mie, fundamental fields are assumed to be nonlinear, out of which particles emerge as localized (soliton-like) structures. As Einstein emphasized, this approach has the philosophical
advantage that particle properties arise in a natural way from the field properties. The problems are threefold: (i) There are so many possible nonlinear field theories that it is difficult to know which to use. (ii) Loss of the superposition property near the centre of the particle renders analysis problematic. (iii) Quantum analysis of a nonlinear field is substantially more difficult.

Prior to 1975 – Figure 20 shows – PRD contributors were largely unconcerned with the extended particle approach, whereas after that date publications have been increasing in a (very roughly) linear manner at a rate of about one paper per year each year. This timing suggests that the onset of the activity was part of the explosive growth of research in nonlinear science shown in Figure 1, during which several well-cited early publications on nonlinear particle models by leaders in the field made soliton studies safe for fledgling physicists [93, 183, 184, 315, 328]. Of deeper importance was a quantum analysis of the SG equation by Roger Dashen, Brosl Hasslacher and Andre Neveu in the mid-1970s, showing that the quantum value of a kink mass is fairly well approximated by the classical value; thus increasing interest in classical nonlinear field theories as useful first approximations to their quantum counterparts [111].

As reviewed in books by Vladimir Makhankov [325] and by Alexandre Filippov [171], several efforts were made during the 1980s to develop nonlinear field theories for elementary particles in three spatial dimensions. In addition to those mentioned above are the following. (i) The four-dimensional, self-
dual Yang–Mills theory [253], a system for which plane wave reductions correspond to standard soliton equations (KdV, NLS, etc.). (ii) Instantons, which are topological, solitary-wave solutions of the Yang–Mills system [252]. (iii) Twistor theory, which attempts to reach Einstein’s decades-long goal of unifying gravity and electromagnetism through the construction of a nonlinear graviton [338, 414]. (iv) String theory (ST), which hopes to unify all the interactions of physics – weak, electromagnetic, strong and gravitational forces. Constructed at the Planck scale (∼ 10⁻³⁵ m) where gravitational forces become significant, ST presently comprises a family of formulations in which the basic particles are quantized oscillations on one-dimensional structures (strings) [37]. From the perspective of this review, it is interesting to note that string dynamics may be nonlinear and therefore chaotic [36], possibly providing the stochastic underpinning needed in the de Broglie–Bohm theory.

Although it remains difficult to divine which of the many available nonlinear theories should be chosen for a particular application, the fact that classical nonlinear field theories give useful first approximations to particle mass spectra and the development of ever more powerful computers (see Figure 3) suggest that a judicious combination of the point- and extended-particle perspectives will continue to provide useful information on the fundamental nature of matter.

But what are the dynamics of matter on the atomic scale?

### 3.3 Quantum theory

Every student of physics learns in a basic course on quantum theory (QT) that the energy (E) in a mode of linear atomic-level oscillation cannot be adjusted to arbitrary precision but is given by Planck’s law: \( E_n = \hbar \omega (n + 1/2) \), where \( \omega \) is the classical frequency of the mode, the integer \( n \) is interpreted as the number of quanta (photons or phonons) in the mode, and \( \hbar \) is Planck’s constant. Furthermore, transitions are only allowed between adjacent energy levels, which implies that the linear mode can only emit or absorb energy at frequency \( \omega \), in accord with classical (non-quantum) observations.

For a single linear molecular oscillator of reduced mass \( m \) and spring constant \( K \), Schrödinger obtained Planck’s law by changing the classical expression for energy conservation \( (E = PE + KE) \) to a linear PDE, called the Schrödinger equation (SE), in which \( KE \) is equal to the classical expression \( (k/2) x^2 \) times a wave function \( \Psi(x, t) \), \( PE \) is changed to \( (-\hbar^2 / 2m) \partial^2 \Psi / \partial x^2 \) and \( E \) goes to \( i \hbar \partial \Psi / \partial t \) [455]. Finally, physicists agree with Born’s suggestion that \( |\Psi(x, t)|^2 \) should be interpreted as a time-dependent probability density for finding the classical particle [61]. Thus QT expresses conservation of energy as a linear PDE for a probability amplitude.

Although some assume that QT can be applied to all dynamical systems, this is not the case. The formulation of an SE requires conservation of energy.

---

22 According to Walter Moore, Schrödinger’s 1926 paper was cited over 100,000 times in its first 34 years, which must be an academic record.
and an associated Lagrangian or Hamiltonian structure, which is not available for reaction-diffusion (RD) systems. Thus there is no quantum theory for the flame of a candle, a nerve impulse, a living organism, or the human brain [511]. In general, this restriction does not cause problems because RD systems contain so many atoms that QT would not be needed. With these ideas in mind, let us see what QT implies in the context of nonlinear science.

**Local Modes in Molecules.** In molecular oscillations, the masses remain constant (relativistic velocities are not approached), but the spring constants are typically sublinear because bond-stretching tends to deplete interatomic electron densities. Assuming $KE = kx^2/2 - \alpha x^4/4$, the Schrödinger equation (SE) remains a linear PDE, but the energy levels are given by [467]

$$E_n = (\hbar \omega - \gamma / 2)(n + 1/2) - \gamma n^2/2,$$

where $\gamma \equiv 3\alpha \hbar^2/4m^2\omega^2$. Evidently, Equation (40) reduces to Planck’s law as $\gamma \to 0$, but in general it implies $E_n - E_{n-1} = \hbar \omega - \gamma n$, or the empirical relation

$$E_n - E_{n-1} = A - Bn,$$

which was first observed for inter-atomic vibrations by Raymond Birge and Hertha Sponer in the mid-1920s [45]. Equation (41) shows how level spacings decrease at higher energies, suggesting that the bond will break for $n \sim A/B$. The Birge–Sponer relation is closely followed for CH-stretching oscillations in benzene, where the ratio $A/B$ and measurements of heat of dissociation have provided convincing evidence since the late-1920s that higher-amplitude oscillations (i.e., larger values of $n$) are concentrated on a single bond, as shown in Figure 21 [153]. Thus nonlinear local modes of oscillation were being observed and studied in small molecules by physical chemists at the same time that the implications of Schrödinger’s 1926 papers were being sorted out by theoretical physicists.
Experimental studies of local modes in molecules were carried on during
the 1930s by Reinhard Mecke and his colleagues [354, 355, 356, 357], but his
work was neglected until the late 1970s [237, 317]. Although this curious over-
sight offers additional evidence of the poor communications among nonlin-
ear scientists prior to 1970, another explanation is as follows. The SE is linear
and a benzene molecule is symmetric under rotations by multiples of
$60^\circ$; thus
eigenstates for CH-stretching oscillations must share the same symmetry and
be spread out over the six bonds shown in Figure 21, not localized on one.
Following this reasoning, physical chemists writing papers on local modes in
the 1970s were sometimes told that their experimental results were at variance
with QT and therefore wrong [236].

True as far as it goes, this argument is incomplete. At larger values of
$n$, the physical chemist does not observe a single eigenfunction on benzene but a
wave packet of them that is spread over an energy range $\Delta E$, where [467]

$$\frac{\Delta E}{\varepsilon} \sim \frac{n(\varepsilon/\gamma)^{n-1}}{(n-1)!}$$

and $\varepsilon$ is the energy of electromagnetic interaction between adjacent CH bonds.
As $\varepsilon \approx 4 \text{ cm}^{-1}$ and $\varepsilon/\gamma \sim 1/30$ for benzene, it is clear that $\Delta E \to 0$ very
rapidly for increasing values of $n$, and for times of order $\hbar/\Delta E$, a local-mode
wave packet will remain organized. In other words, individual quantum states
comprising local mode wave packets with $n > 3$ are quasidegenerate, which
explains experimental observations of local modes without violating linear quan-
tum theory.

QUANTUM SOLITONS. Consider a chain (or one-dimensional lattice) of $f$
classical nonlinear oscillators of the form

$$\left( i \frac{d}{dt} - \omega \right) A_j + \varepsilon (A_{j+1} + A_{j-1}) + \gamma |A_j|^2 A_j = 0 ,$$

where $A_j = A_{j+1}$ is a complex mode amplitude and $j$ is an index running over
periodic boundary conditions. In this equation, $\omega$ is the oscillation frequency at
one of the lattice points, $\varepsilon$ is the interaction energy between adjacent oscillators,
and $\gamma$ is the anharmonicity of each oscillator in the rotating wave approxima-
tion. With $f = 0$, Equation (43) provides a model for the benzene oscillations
of Figure 21 [476]. In the continuum limit ($f \gg 1$ and $\gamma/\varepsilon \ll 1$), this discrete
nonlinear Schrödinger (DNLS) equation has solitary wave solutions that cor-
respond to solitons of the continuum nonlinear Schrödinger (NLS) equation
[480].

A quantum version of Equation (43) is obtained by letting $A_j (A_j^\dagger) \to b_j (b_j^\dagger)$,
where $b_j (b_j^\dagger)$ are boson lowering (raising) operators which obey the commuta-
tion relations $[b_{j,k}, b_{k,l}^\dagger] = \delta_{jk}$. Neglecting a constant term, the energy operator
becomes

$$H = - \sum [\varepsilon b_j^\dagger (b_{j+1} + b_{j-1}) + (\gamma/2) b_j^\dagger b_j b_j] ,$$

63
where the sum is over $j$. For a single quantum ($n = 1$), the nonlinear parameter ($\gamma$) falls out, and for $n \geq 2$, it is convenient to use the fact that $H$ commutes with the number operator $N = \sum b_j^\dagger b_j$. Thus in the Schrödinger picture, a general eigenfunction ($\Psi$) of $N$ can be constructed with a maximum of $(n + f - 1)!/(f - 1)!n!$ undetermined constants, which are fixed by requiring $\Psi$ to be an eigenfunction of $H$ [467].

For two quanta ($n = 2$) in the large $f$ limit, energy eigenvalues are as shown in Figure 22, where each eigenfunction changes by a factor of $e^{ik}$ under translation by one unit of $j$ (a lattice spacing); thus $k$ is the “crystal momentum” of an eigenstate [212]. This figure shows both a continuum band (the shaded area) and a soliton band given by

$$E_2(k) = \sqrt{\gamma^2 + 16\varepsilon^2 \cos^2(k/2)} = E_2(0) + k^2/2m^* + O(k^4),$$

where $m^*$ is the effective mass near the band centre. The soliton band is characterized by two features. First, it is displaced below the continuum band by a binding energy $E_b = \sqrt{\gamma^2 + 16\varepsilon^2 - 4\varepsilon}$. Second, inspection of the corresponding eigenfunctions shows that the two quanta are more likely to be on the same site for the soliton band than in the continuum band [467]. For arbitrary $n, \gamma \ll \varepsilon$ and sufficiently large $f$, the quantum binding energy is

$$E_b = \frac{\gamma^2}{48\varepsilon}n(n^2 - 1),$$

which corresponds to the binding energy of a classical NLS soliton under the identification $n = \sum |A_j|^2 \gg 1$ [325].

In the classical DNLS system with $\gamma \gg \varepsilon$, numerical studies show that the soliton becomes pinned to the lattice. Under quantum theory, this classical, nonlinear phenomenon is reflected by the fact that the effective mass, $m^* = (n - 1)!\gamma^{n-1}/2n\varepsilon^n$, becomes very large. Because the classical Ablowitz–Ladik (AL) system [4] (which also reduces to NLS in the continuum limit) is completely integrable for all parameter values, its soliton is not pinned for $\gamma \gg \varepsilon$. This classical fact is reflected by an effective mass that approaches zero under the same conditions. An equation formulated by Mario Salerno interpolates between the DNLS and AL limits [451].

Evidently, $\Psi(t)$ is a rather complicated object, and the numerical difficulty in computing it increases combinatorially with the number of quanta, but breather-breather interactions have recently been studied on the quantum DNLS system [132].\footnote{A complete quantum reconstruction of a classical soliton is even more complicated, as it is a coherent state, comprising components with different values of $n$ in addition to being a wave packet over $k$ [467].} Soliton systems that combine a local excitation with an associated distortion of a crystal lattice are yet more difficult to solve, requiring either a product approximation based on different time scales or the truncation of an infinite dimensional matrix [212, 467]. Such systems include polaron, superconducting metals, and solitons in biopolymers, all of which are discussed below.
QUANTUM INVERSE SCATTERING. An important feature of classical soliton theory is the IST method of solution, which has been mentioned above. In 1981, Ludvig Faddeev [163] and Harry Thacker [512] showed how this classical knowledge can be used to solve problems in quantum field theory. These results are of interest to high-energy physicists because particle interactions in the Standard Model are based on quantum fields.

To see how this goes, start with Equation (43). In the classical continuum limit, $A_j(t) \rightarrow u(x,t)$, and with appropriate normalizations $u$ satisfies the NLS equation given in Equation (15). Under quantization (as noted above), $A_j \rightarrow b_j$, which is a function of time in the Heisenberg picture. Thus either by taking the continuum limit of the discrete operator equation for $b_j(t)$ or quantizing the PDE for $u(x,t)$, one arrives at the following equation for the quantum field operator $\phi(x,t)$:

$$i \frac{\partial \phi}{\partial t} + \frac{\partial^2 \phi}{\partial x^2} + 2|\phi|^2 \phi = 0,$$

where $\phi(x,t)$ has the commutation relation $[\phi(x,t), \phi^\dagger(y,t)] = \delta(x - y)$. The scheme is as follows:

$$\begin{align*}
A_j(t) & \rightarrow u(x,t) \\
\downarrow & \\
B_j(t) & \rightarrow \phi(x,t)
\end{align*}$$
where downward arrows indicate quantization and right-going arrows indicate a continuum (or field) approximation.

Under the classical version of the IST method, a key function is the reflection coefficient which is constructed from \( u(x, 0) \) and depends simply on time. From the reflection coefficient, \( u(x, t) \) can be reconstructed using the Gel'fand–Levitan integral equation \([467]\). Under a quantum version of the IST method, the reflection coefficient becomes a raising operator for wave-function solutions of Equation (47) \([542]\). Among many interesting results, this leads to a binding energy \( E_B = n(n^2 - 1)/12 \) which corresponds to Equation (46) \([270]\).

Thus the set of classical PDEs that possess IST formulations (the classical soliton equations) leads directly to a set of solvable quantum field theories. Interestingly, Equation (43) serves as a model for the local modes in benzene (which were first observed in the 1920s), and it leads to the quantum NLS equation. It is also interesting to note that under quantum theory the properties of quantum solitary waves are identical to those of ordinary particles (electrons, positrons, etc.) \([467]\).

**Quantum Chaos?** According to standard quantum theory, a wave function \( \Psi(x, t) \) obeys a PDE of the form \( i\hbar \partial \Psi(x, t)/\partial t = \mathcal{H}\Psi(x, t) \), where \( \mathcal{H} \) is a linear (self-adjoint) differential operator; thus truly chaotic behaviour is not allowed for \( \Psi \). Nonetheless, a quantum system that is close to a chaotic classical limit will show behaviour reflecting the chaotic features of that limit.

As exact solutions of a quantum problem are difficult to obtain near the classical limit, one turns to semiclassical methods, but Einstein pointed out in 1917 that it is not clear how to use such methods for Poincaré’s irregular trajectories \([150]\). Although many open questions remain, progress has been made in understanding aspects of the quantum solutions in the semiclassical limit, including the following \([85, 141, 143]\). (i) The probability of finding a nearest-neighbor energy eigenvalue spacing \( S \) for an integrable system is \( \exp(-S) \) (Poisson statistics), whereas for a nonintegrable (chaotic) system governed by a Hamiltonian with real components, this probability is approximately \( \exp[-(\pi/4)S^2] \) \([335]\). A qualitative reason for this difference is that oscillation components are independent for integrable systems, allowing eigenfunctions to cross as a perturbing parameter is varied, while crossings are avoided for nearly chaotic systems because the component eigenfunctions interact. (ii) The quantum mechanical eigenfunctions of classically chaotic systems become random functions with irregular nodal patterns in the semiclassical limit; however “scarring” of a small number of these eigenfunctions near classically unstable orbits is observed \([231]\). (iii) Quantum wave packets spread out and decay more rapidly for classically chaotic systems than for classically integrable systems. It should be noted that these results are not confined to quantum systems; they can be expected for any wave system (acoustic, optical, hydrodynamic) for which the propagation of rays is chaotic in the short wavelength limit.

A perturbed hydrogen atom is a useful testbed for such ideas, because without perturbation this two-body system is classically integrable (as New-
ton showed) and thus chaos free [143]. Two examples are these. First, ion-
ization of a highly-excited hydrogen atom by a microwave field is found to
depend sharply upon the field amplitude. In the context of KAM theory, this
phenomenon is explained by noting that tori persist at low field amplitudes,
preventing small chaotic regions to coalesce, but at a critical amplitude these
tori break up, allowing trajectories to explore the entire phase space includ-
ing paths to ionization [272]. Second, the static Hamiltonian of a hydrogen
atom in a magnetic field is no longer integrable, and the corresponding energy
eigenvalue spectrum becomes randomly scrambled, in accord with theoretical
derivations of spacing statistics [185].

Finally, we recall the proposal of de Broglie and Bohm that the Schrödinger
equation (SE) is the linear approximation to a nonlinear formulation that ex-
plains the facts of atomic physics in a different way [53, 119]. To test for such
possibilities, Steven Weinberg has analyzed precise measurements of atomic
energy levels in metallic ions, finding that any nonlinear aspects of the SE must
be at energies less than \(10^{-20}\) times the Rydberg energy (13.6 eV) [55, 548].

3.4 Nonlinear chemical and biochemical phenomena

Chemistry is a vast subject, and chemical dynamics involve diverse aspects at
different levels of approximation, many of which are nonlinear. In this subsec-
tion, some salient examples of nonlinear chemical and biochemical dynamics
are described.

MOLECULAR DYNAMICS. Among the more important applications of quantum
type is a 1927 proposal by Born and J. Robert Oppenheimer for calculat-
ing the force fields binding atoms into molecules [62]. From this work, ac-
cording to Paul Dirac [130]: “The underlying physical laws for the whole of
chemistry are completely known, and the difficulty is only that the exact appli-
cation of these laws leads to equations much too complicated to be soluble.”

Based upon the fact that atomic nuclei are much heavier than electrons, the
Born–Oppenheimer (BO) approximation assumes the total wave function to be
factored into electronic and nuclear components, according to the follow-
ing computational procedure [488]: (i) assume fixed distances separating the
atoms of a molecule, (ii) use SE to compute the eigenstates of the valence elec-
trons, (iii) fill these states with the available valence electrons, and (iv) compute
the total energy of the system (valence electrons plus charged inner shells and
nuclei). Repeating these four steps for many interatomic distances allows the
construction of a multidimensional potential energy function which, in turn,
determines the interatomic forces. Although this program was numerically
daunting in the late 1920s, refinements and recent developments in computing
power make it possible to calculate energy functions for small molecules that
depend on lengths, bending and twisting of covalent bonds, electrostatic inter-
actions, and interatomic repulsions [342]. Presently, there are several molecu-
lar dynamics codes available for physical chemists to compute static structures
and vibrational modes of molecules [298, 516].
Determining the structure of a molecule requires finding a coordinate surface where the total energy is a minimum, which is a classical (i.e., not a quantum) calculation. To obtain relative intensities of vibrational spectra, on the other hand, QT must be used, where a multidimensional SE is written for the vibrating atoms (the energy operator in Equation (44) is a simple example of such an analysis). For computing the frequency and qualitative nature of a vibration, however, it is often sufficiently accurate to assume that the atomic masses follow Newton’s Second Law, where the forces are obtained by differentiating the BO potential energy with respect to interatomic distances and orientations. The local mode of CH stretching oscillation in benzene, shown in Figure 21, emerges from such a classical nonlinear calculation [467]. In general, of course, one would expect such classical dynamics to be chaotic, and this has been explored for a special case of Equation (43) with \( f = 3 \) [103]. With \( \gamma \ll \varepsilon \), the model reduces to a linear system which is integrable, and for \( \gamma \gg \varepsilon \) the model corresponds to a system of isolated nonlinear modes which is also integrable. For \( \gamma N \approx \varepsilon \), however, the dynamics of the system were found to be chaotic over about 60% of the phase space.

Energy localization in biomolecules. Throughout the 1970s, most believed that localized packets of energy had no role to play in biology. Whereas the warm and wet nature of organic tissue clearly allows for RD phenomena like the nerve impulse [244], energy-conserving processes seemed improbable; thus the paper on protein solitons by Davydov at the above-mentioned Gothenburg meeting in the summer of 1978 was an interesting surprise. Modeled in the continuum approximation by the NLS equation and so related to Equation (43), the physical nature of the nonlinearity in Davydov’s soliton is quite different. In energy localization on small molecules, \( \gamma \) is an intrinsic nonlinearity arising because the force constant of a bond decreases under stretching as the electron density is depleted. Davydov’s formulation, on the other hand, assumes that \( \gamma \) is an extrinsic nonlinearity, stemming from a local distortion of the protein structure [115, 467].

Extrinsic nonlinearity goes back to the idea of the polaron, proposed by Lev Landau in 1933 [295] and widely studied in the Soviet Union in the 1950s [410], in which an electron manages to travel through a crystal by associating itself with a moving distortion of the lattice – rather like a marble working its way through a plate of spaghetti. Although Davydov later expanded his theoretical perspectives to include biological charge transport [116], his original picture was of a localized CO stretching oscillation (amide-I mode) in the peptide units of alpha-helix protein that distorts the nearby helical structure, and this local distortion, in turn, traps the amide-I energy, preventing its dispersion and offering a means for the storage and transport of biological energy.

Closely related to Davydov’s basic formulation is the Zakharov equation, developed at about the same time (and also in the Soviet Union) to describe turbulence of plasma waves [578]. In both cases, an exact quantum analysis is more difficult than for Equation (43) because it is necessary to represent the quantum character of the lattice. Davydov dealt with this problem by using a
product wave function (similar to the BO approximation), in which the higher frequency amide-I oscillation is loosely coupled to the lower frequency lattice distortion [116, 464].

Interestingly, Davydov’s analysis is based on a well-defined Hamiltonian, the parameters of which are independently determined; thus he avoided the fudging that often characterizes theoretical studies of biological problems. On his account, Davydov was motivated by the proceedings of a conference that was organized by David Green in 1973 to consider the “crisis in bioenergetics” [203]. This crisis involved deciding whether biological energy transduction could be entirely explained by Peter Mitchell’s “chemiosmotic hypothesis” – under which energy derived from hydrolysis of adenosine triphosphate (ATP) is always stored in the electrostatic fields of charges separated by a membrane [366] – or whether there is some intermediate storage mechanism within a protein.

Of particular interest in Green’s proceedings is the summary of a series of papers by Colin McClare in which he analyzes the physics of energy transduction at the molecular level and concludes that resonant storage and transfer of energy is required to attain the efficiencies observed in biological organisms [343, 344, 345]. A useful feature of this publication is the transcription of discussions, from which it is clear that McClare’s ideas were greeted with skepticism, by critics who claimed that resonant states must decay in a small fraction of a picosecond. Discouragement by such negative responses to his ideas may have contributed to McClare’s suicide early in 1977, just as our collective understanding of energy localization from the perspectives of nonlinear science was undergoing a sea change [519]. Fortunately, McClare’s papers have been archived in King’s College, Cambridge at the direction of Maurice Wilkins, who believed that his insights would eventually be recognized.24

Throughout the 1970s, Giorgio Careri and his students at the University of Rome were experimentally investigating the infrared properties of crystalline acetanilide (ACN) (a model protein) in order to better understand the properties of natural proteins. In the course of this work, they discovered a temperature-dependent band at 1650 cm⁻¹ (see Figure 23) which they were unable to assign until they became aware of Davydov’s work in the early 1980s [82]. Although physical chemists challenged Careri’s assignment of the 1650 cm⁻¹ to nonlinear localization, no credible counter evidence was presented, and the Davydov theory correctly predicted both the Birge–Sponer coefficients prior to their measurement and the temperature dependence of the band intensity [464]. Recently, Julian Edler and Peter Hamm have used femtosecond pump-probe measurements to demonstrate conclusively that the self-localization assignment is correct [144]. These pump-probe measurements – which give null readings unless the band is nonlinear [218] – are in accord with numerical calculations by Leonor Cruziero and Shozo Takeno [104], and as discussed below, they also show room-temperature lifetimes of about 18 ps for NH stretching bands in ACN [145]. Thus McClare’s suggestion of resonant

24See: http://www.kcl.ac.uk/iss/archives/collect/1mc30-a.html.
storage and transfer of biological energy has received empirical support, some three decades after his original publications.

Chemical aggregates. Beyond the level of individual molecules, chemicals aggregate in several ways. From a commercial perspective, the most important may be polymers which comprise long chains of repeating units (monomers) [421]. These chains can be linear (like spaghetti), branched, or crosslinked into three dimensional networks as in epoxy. Nonlinear considerations enter into autocatalytic growth processes for these chains or networks which can depend strongly on temperature and pH. A recently observed phenomenon is the propagation of a thermal front through a zone of monomers, like a “liquid flame” [422]. In biochemistry a most important problem is the prediction of a protein structure from knowledge of the sequence of its constituent amino acids [353]. Although in principle a task of energy minimization, the protein folding problem is rendered difficult by the combinatorial explosion of possible structures with the length of the chain; thus a variety of heuristic procedures are presently used with modest success.

Cluster coagulation is more general than polymerization, involving coalescence of rain drops, particles of smoke and dust, and bacterial aggregation, in addition to the chemical bonding of large molecules [424]. Nonlinear dynamics enter as growth laws for all of these examples, formulated by computing growth rates as functions of current and past clusters. The Verhulst law, given in Equation (2) is a simple example, but many more complex growth phenomena are currently being studied.


Figure 23: Infrared absorption spectra of crystalline acetanilide (ACN) in the region of the amide-I (CO-stretching) mode as a function of temperature. A normal (delocalized) amide-I band is at 1665 cm⁻¹ and a self-localized (soliton) peak is at 1650 cm⁻¹.
(LB) films are interesting and potentially useful structures formed first as molecular monolayers on liquid surfaces [416]. Sophisticated techniques are currently being developed for transferring such films to crystal surfaces in connection with research in nanotechnology. Scheibe aggregates are a special type of LB film discovered independently by Günther Scheibe in Germany and Edwin Jelly in England during the mid-1930s [524]. Composed of organic dye molecules, these films are useful as efficient photon detectors, and various nonlinear models have been proposed to describe their operation, including a discrete two-dimensional version of the NLS equation [94]. Finally, liquid crystals comprise rod- or disc-shaped molecules in which the phases can be controlled by electric fields, leading to interesting nonlinear optical properties [394].

Chemical kinetics. In addition to investigating the static aspects of chemical aggregates, it is interesting and potentially important to understand the dynamics of chemical reactions. These studies go back at least to the theoretical work of Lotka in the early 1920s [313], and they have played an important role in the nonlinear science expansion of the 1970s [392, 390]. In addition to the BZ reaction described above, there is the Brusselator – a somewhat simpler and more flexible RD system sharing many of the same qualitative properties [389], and the Turing mechanism. This is an RD system in which the inhibiting component diffuses faster than the exciting component, leading to stationary patterns that may be related to those in biological organisms [58, 520]. Finally, energetic materials (explosives) provide dramatic examples of chemical kinetics [114]. These are divided into low-velocity mixtures like black powder, in which the nature of the blast depends upon the enclosure, and high explosives (nitroglycerine, dynamite, and ammonium nitrate-fuel oil mixtures) in which the high speed of the reaction makes the dynamics of a blast independent of its confinement.

3.5 Condensed-matter physics

Comprising any of the ninety-odd atomic elements bound together by highly nonlinear valence and electrostatic forces, condensed matter offers many opportunities for nonlinear phenomena to arise, some of which are mentioned in Section 2.2 on “Solitons and solitary waves”. Following the Frenkel–Kontorova formulation of the sine-Gordon (SG) equation to describe the dynamics of crystal defects [182, 239, 482], it was found that this same equation – augmented with appropriate dissipative and driving terms – provides a model for the motions of domain walls which separate regions of differing magnetization (or polarization) in ferromagnetic (or ferroelectric) materials [133, 522] and for the propagation of splay waves on biological membranes [169].

Extrinsic nonlinearity. Among the earliest proposals for nonlinear phenomena in solids were Rudolph Peierls’s 1930 suggestion of charge-density waves [533] and Landau’s above noted concept of the polaron, in which a moving charge takes advantage of extrinsic nonlinearity (interaction with and distortion of the nearby lattice structure) to move more easily through a crystal
Studied extensively in the Soviet Union in the early 1950s [410], the concept of extrinsic nonlinearity led to an understanding of the phenomenon of superconductivity, in which a pair of electrons couples through the lattice to change from two fermions into a boson-like entity which can aggregate into a macroscopic quantum state [212, 454, 515]. The dynamics of a macroscopic superconducting state, in turn, is governed by the celebrated Ginzburg–Landau (GL) equations [124], which are closely related to the nonlinear Schrödinger (NLS) equation of soliton theory [324]. At a higher level of description, one finds the formation of magnetic flux vortices which allow magnetic fields to penetrate into Type II superconductors [409], and these vortices are related to the fluxons that obey the SG equation on long Josephson-junction structures [31, 306]. Finally, if certain of its parameters are allowed to be complex, GL equations can exhibit properties of reaction-diffusion (RD) systems [323]. Interestingly, mechanisms for extrinsic nonlinearity are often omitted from molecular-dynamics models of condensed matter systems [516].

PHASE TRANSITIONS. In the introduction to this review, the concept of a phase transition was proposed as a metaphor for Kuhn’s Gestalt-like paradigm shift in the perspectives of a scientific community [121]. This idea is of importance in condensed-matter studies, providing descriptions of condensation of gasses, freezing and boiling of liquids, and melting of solids; the formation of domains of uniform magnetization (or polarization) in ferromagnets (or ferroelectrics); establishment of superconducting states in metals and superfluid states in liquid helium [277]; and the onset of coherent light output from a laser above a certain pumping level [492] – to name but a few of many applications.

Salient aspects of phase transitions can be understood by thinking about the properties of ferromagnetic materials such as a bar of iron or of nickel, which can be viewed as a collection of atomic magnets [540]. If these small magnets are assumed to be uncoupled, application of a magnetic field of an external magnetic field of \( H \) amperes/metre would result in a magnetization \( M \) that is proportional to \( H \) and inversely proportional to absolute temperature \( T \), because at high temperatures the magnetic orientations assume almost random directions. Coupling among the magnets introduces a tendency to take the same orientations as their neighbors; thus \( M \propto H/(T - T_c) \), where \( T_c \) is a Curie temperature, below which the ferromagnetic metal can remain magnetized without the action of an applied field. (For iron, \( T_c = 1043K \) and for nickel it is around 630K.) Far above the critical temperature \( T \gg T_c \), on the other hand, the coupling between neighboring atoms becomes negligible. In other words, the system undergoes a phase transition at \( T = T_c \), below which \( M \) can be finite for \( H = 0 \). Why doesn’t an ordinary piece of iron act like a magnet?

Two other contributions to the total free energy of bar magnet must be considered to understand its global behaviour. First, there is an external field energy, which increases when the atomic magnets are all aligned in one direction. This external field energy is reduced if the interior is divided into small and randomly oriented domains of uniform magnetization that are separated by domain walls, across which the orientation of \( M \) changes. Second is the
energy needed to create these internal domain walls. Thus if there are too few internal domains, the external field energy grows, and if there are too many, the total domain wall energy grows. For an ordinary bar of iron that has cooled from the melt, a balance between these two opposing tendencies is established, rendering the bar unmagnetized – little or no external field is observed. If an unmagnetized bar at room temperature is placed in a sufficiently large longitudinal magnetic field ($H_{\text{ext}}$), however, all of the internal domains are forced to become oriented in the same direction, and the internal domain walls disappear. Under these dynamics, the corresponding domain wall motions are governed by a version of the SG equation that is augmented to include both dissipative effects and the forces induced by $H_{\text{ext}}$. As $H_{\text{ext}}$ is reduced to zero, some domain walls reappear, but their positions become pinned to mesoscopic irregularities of the metal, leaving the residual external magnetic field of a standard bar magnet. Thus a simple bar magnet displays a variety of nonlinear phenomena, the social implications of which have not been lost upon despots.

Interestingly, if the natural magnetizations $M(T)$ of iron and of nickel are measured in units of their low-temperature values $M_s$, the resulting curves are universal functions of the parameter $\varepsilon = \frac{T_c - T}{T_c}$, and similar universal curves are observed for the densities of superconducting electrons [409]. More generally, a wide variety of critical phenomena (melting, vaporization, sublimation, etc.) display universal behaviours near critical points, where dependencies are upon powers of the order parameter defined in Equation (48) that are independent of particular materials [522]. These critical exponents can be computed from renormalization group theory, which assumes that macroscopic behaviour is independent of microscopic details because it can be calculated from a nested set of mesoscopic models [523]. Similar ideas are used in high-energy physics to construct field theories with properties that are independent of the length scale chosen for ultraviolet cutoff.

**Supersonic solitary waves.** Supersonic propagation of sound was first observed on February 9, 1822 by Captain William Edward during the second of his three arctic expeditions to find a northwest passage across Canada. In a description of his experiments on the velocity of sound, it is noted in the journal of that day that [157] “the officer’s word of command ‘fire’ was several times heard distinctly … about one beat of the chronometer [nearly half a second] after the report of the gun.” As the propagation distance was 1720 metres [217], these observations suggest that the high amplitude wave of the cannon had a velocity about 10% higher than that of the low amplitude command to fire it. Was Parry’s observation an isolated oddity, or does it suggest a general property of large amplitude waves in nonlinear media?

To understand Parry’s observation, consider the spring-mass lattice shown in Figure 12. If the springs are nonlinear with a certain exponential potential, the model becomes the Toda lattice (TL) having stable soliton solutions [467,
where Equation (26) shows that these solitons travel at supersonic speeds. Although the special properties of the TL require the particular exponential form given in Equation (24), supersonic solitary waves occur for a broad class of spring potentials \( U(r) \), as is shown by a theorem recently proven by Gero Friesecke and Jonathan Wattis [190].

**FW Theorem.** Assume that \( U(r) \) has a second derivative, with \( U(0) = 0 \), \( U''(0) = 0 \), and \( U(r) \geq 0 \) in some neighborhood of the origin. A sufficient condition for solitary wave solutions of the spring-mass lattice to exist is that \( U(r) \) be superquadratic on at least one side. (In other words, \( U'(r)/r^2 \) increases strictly with \( |r| \) either for \( r \) between zero and some positive value \( R_+ \), or between zero and some negative value \( -R_+ \), where \( R_+ \) and \( R_- \) could be finite or infinite.) Furthermore, these solitary waves are supersonic, with arbitrarily large and small amplitudes, and they are either all positive or all negative (entirely expansive or compressive).

Stability of these supersonic solutions has been established in a recent series of papers by Friesecke and Robert Pego [186, 187, 188, 189]. Among others, the conditions for this theorem are satisfied for the TL potential and for the potentials assumed in the FPU studies [170]. Thus an explanation for Parry’s observation of supersonic sound follows from the FW theorem and many physically reasonable models for the nonlinear compression forces of the atmosphere. As the spring-mass model of Figure 12 serves equally well for planar sound waves in a solid, supersonic solitary waves are to be expected in one-dimensional problems of condensed-matter physics [106].

**Discrete breathers.** The supersonic waves discussed in the previous section are characterized by two properties: they are one-dimensional (1D) and they have no oscillatory character. As was seen in Section 3.4, nonlinear excitations that are localized in 3D may arise if these two restrictions are lifted. Although the possibility of localized oscillatory states is suggested by Mecke’s observations of local modes in symmetric molecules during the 1930s [354, 355, 356, 357] and by Careri’s assignment of certain CO-stretching oscillations in model proteins during the early 1980s [82], most solid state physicists continued to believe throughout the 1980s that it was theoretically impossible to have localized vibrational energy in a crystal with translational symmetry.

An important contribution to this discussion was made in a 1988 paper by Albert Sievers and Takeno, who showed how the discreteness of a molecular crystal might contribute to the stability of localized nonlinear oscillations [487]. It was argued that these “intrinsic local modes” (ILMs) will be stabilized against radiation into phonons of the crystal lattice if overtones of the basic frequency lies within the stop bands of the lattice. In 1994, Robert MacKay and Serge Aubry turned this physical argument into a mathematical proof for stationary (non-moving) “discrete breathers” (DBs) under the assumption of

---

Note that the terms “intrinsic” and “extrinsic” are used in two different ways. In Section 3.4, an intrinsic nonlinearity is generated directly through changes in the valence bonds, whereas an extrinsic nonlinearity involves interactions with the crystal lattice. In the terminology of Sievers and Takeno, an intrinsic localization appears in a regular crystal, whereas extrinsic localization arises at the site of a symmetry breaking dislocation or embedded atom.
The six crystal directions

Figure 24: A piece of muscovite mica (overall length is 23 cm). Inside the oval marked “A”, the tracks of several moving breathers appear to be emerging from a charged-particle track. (Courtesy of Mike Russell.)

sufficiently strong nonlinearity [319], which generated interest within the nonlinear science community [80, 467].

Given confidence in the objective reality of stationary nonlinear oscillating states (DBs or ILMs), it is of interest to consider whether they can move. This is a more difficult question because mathematical theorems are not yet available and direct experimental observations of moving states are difficult. Nonetheless, some progress has been achieved through a combination of numerical studies and indirect empirical observations. On the numerical side, Henrik Feddersen has studied solitary waves moving through a discrete 1D lattice with extrinsic nonlinearity (Davydov’s model), finding propagating solutions to a high degree of numerical accuracy over a significant range of model parameters [166]. These results are important, as they suggest that lattice discreteness is not a barrier to the formation of moving localized breathers.

Empirically, Mike Russell has proposed that many of the dark lines appearing in large pieces of muscovite mica (KAl₃Si₃O₁₀(OH)₁₋₂F₀₋₂) provide evidence of moving breathers that were present during the cooling process [446, 447]. As shown in Figure 24, sheets of this material have $C_6$ symmetry, with the six crystal axes indicated by a star imposed on the figure. Those lines that do not correspond to these six directions have been identified as charged particles generated by cosmic rays [444, 445]. Interestingly, many of the lines along the crystal axes are seen to emerge from charged particle lines, for example, those within the oval marked “A” in Figure 24. In collaboration with Russell, Chris Eilbeck and José Marín have recently been numerically demonstrated that breathers localized on a two-dimensional plane can propagate for long distances along the directions of crystal symmetry, corresponding to many of the lines in Figure 24, [332, 334, 333].
3.6 Nonlinear electronics

Since the invention of the iron core transformer in the nineteenth century and Nikola Tesla’s use of it as a component in his alternating-current distribution system, electrical engineers have dealt with nonlinear dynamics. The power passing through a transformer, for example, is limited by the maximum magnetization \( M \) of the core, leading to a nonlinear phenomenon called hysteresis, in which \( M \) lags behind magnetic field intensity in the course of a periodic cycle [540]. At the beginning of the twentieth century, however, electrical engineers could not imagine how many additional nonlinear dynamic effects were to emerge.

**Vacuum tube electronics.** In 1906, a decade after the discovery of the electron, Lee De Forest invented the triode vacuum tube (or “audion”), a device that allowed the ratio of output to input power to substantially exceed unity. In addition to many radio and telephone applications, this amplification of an efficient cause introduced two studies of central importance in nonlinear science. The first was Adrian’s discovery of the “all-or-none” effect (or threshold phenomenon) in neurodynamics [6, 477], which opened new vistas in electrophysiology [466]. The second was Balthasar van der Pol’s seminal study of nonlinear oscillations [526]. Using a transformer to feed the resonant output voltage of a triode amplifier back to its input terminals, van der Pol arrived at an electronic system described by the equation

\[
\frac{d^2 y}{dt^2} - \varepsilon (1 - y^3) \frac{dy}{dt} + y = 0,
\]

which now bears his name.

To understand this equation, consider several values of the parameter \( \varepsilon \): (i) For \( \varepsilon = 0 \), this equation describes a sinusoidal oscillation of constant amplitude which is the behaviour of the resonant L.C. “tank circuit” of the amplifier output. (ii) For \( \varepsilon < 0 \), the oscillation amplitude decreases with time to zero. (iii) For \( 0 < \varepsilon \ll 1 \), the oscillation remains approximately sinusoidal but increases slowly with time, reflecting the fact that the amplifier is pumping energy into the tank. (iv) For \( \varepsilon \gg 1 \), the dynamics is that of a blocking oscillator (or multivibrator), switching rapidly back and forth between positive and negative solutions of \( y = \varepsilon (1 - y^3) y \). Thus under continuous changes of the parameter \( \varepsilon \) the qualitative nature of the dynamics becomes completely different [418].

In his 1926 paper – which should be read by all serious students of nonlinear science – van der Pol introduced the ideas of phase-plane analysis [267] and of averaging methods [302]. Along with piece-wise linearization, these techniques were subsequently developed to a high degree by electrical engineers and applied mathematicians in the Soviet Union [47, 408]. Thus by the early 1960s a US engineering student with an interest in general nonlinear analyses of electronic systems could consult translations of Russian books that had been written a decade or more earlier by by Alexandr Andronov and his colleagues [15] and by Nikolai Bogoliubov and Nikolai Mitropolsky [52], in addition to an important book by Nicholas Minorsky [365].
NEGATIVE AND POSITIVE FEEDBACK. With the invention of the audion, it seemed straightforward to design a system that would permit transcontinental or intercontinental telephone communications, but there was a hitch. About a hundred amplifying units would be needed at regular intervals along the telephone line to make up for resistance losses, and variations in the amplifications of the individual units (caused by deterioration of the vacuum or loss of electron-emission ability of the hot cathode) would be raised to the 100th power at the output terminal of the system. In the mid-1920s, this was serious problem, under intense study at the Bell Telephone Laboratories (BTL) by a number of electrical engineers, including young Harold S. Black.

While coming across the Hudson River on the morning of August 2, 1927, Black suddenly wrote the following formula on his copy of The New York Times

\[ G = \frac{A}{1 - \mu A}, \]

which made everything clear.26 Now known to every electrical engineer, this expression embodies one of the most important inventions of the twentieth century, making transcontinental telephone possible [68]. To understand Equation (50), note that \( A \) is the gain of three amplifier stages (audions), \( \mu \approx 1/10 \) is the loss of a negative feedback circuit which connects the output of the amplifier back to the input terminals, and \( G = 1/\mu \ (\pm 1\%) \approx 10 \) is the overall (closed-loop) gain of the amplifier. In other words, a gain ratio of 10,000 has been traded for a gain ratio of 10, with the advantage that \( G \) depends only on the passive components of the feedback circuit; thus the gain of 10 will remain stable as the amplifying properties of the audion components change under aging.

From a somewhat more general perspective, this feedback amplifier circuit can be viewed as asking that the ratio of output to input voltage be equal to \( 1/\mu \). In these terms, negative feedback was used in the steam-engine governor which was invented in the eighteenth century and studied by James Clerk Maxwell in the nineteenth century [292]. Following Black’s invention, however, Equation (50) was investigated in great detail, and the various requirements to avoid unwanted oscillations (called “singing” at BTL) were completely discussed in a classic book by Henrik Bode, which appeared as a BLT report in the early 1940s [50]. By the early 1950s, the negative feedback idea had been extended to servo-systems, in which the desired value of a variable (temperature of a room, position of a boat’s rudder, firing pattern of an anti-aircraft gun, etc.) is fixed by requiring that the difference between desired and actual values is forced to zero.

Yet more generally, negative feedback introduces closed loops of causality into electronic systems, possibly confounding relations among causes and effects. As long as Bode’s stability requirements are satisfied, the system does not oscillate, and \( G \) expresses the relationship between a simple cause and effect. But

26 Black’s copy of the Times became an important patent document, establishing the date of his invention.
sometimes – as with van der Pol’s circuit of Equation (49) – oscillation is desired, and positive feedback is employed. From the perspectives of nonlinear science, positive feedback (PFB) is more interesting than negative feedback (NFB) for three reasons. First, NFB stabilizes a system, making the relation between input and output variables more linear. Second, nonlinearity always comes into play with PFB because the initial exponential growth must eventually be limited by some nonlinear effect, as with the saturation in Figure 1. Finally, PFB is an essential element of the phenomenon of emergence, under which a qualitatively new entity (tornado, city, living organism, biological species, Kuhnian paradigm, etc.) comes into existence.

**Frequency-power formulas.** In the early days of radio broadcasting, receivers were designed to amplify the incoming signal at its carrier frequency. These “tuned-rf” receivers employed two or three stages of vacuum-tube amplification, with tuning capacitators between each stage that were mounted on a common shaft – a rather large component that was easily damaged and took up much of the volume of the set. Since the 1930s, however, receivers have employed Reginald Fessenden’s heterodyne principle, an early example of a nonlinear process in electronics [427]. Under heterodyning, the antenna signal is “mixed” with the voltage from a variable frequency oscillator to obtain an intermediate frequency signal (at the difference between oscillator and antenna signals), which is then amplified through several vacuum-tube stages of fixed frequency. Both simpler and more effective, Fessenden’s heterodyne process is now universally used in the design of radio receivers.

In 1956, Jack Manley and Harrison Rowe showed that lossless mixers obey restrictions on the ratios of input and output powers that are more severe than conservation of energy [329]. For example, if the carrier, local oscillator and intermediate frequencies of a lossless mixer are respectively $\omega_c, \omega_0$, and $\omega_i$ (with $\omega_c = \omega_0 + \omega_i$ and $P_c + P_0 + P_i = 0$), then

$$ \frac{P_c}{\omega_0} + \frac{P_i}{\omega_i} = 0 \quad \text{and} \quad \frac{P_0}{\omega_0} + \frac{P_i}{\omega_i} = 0. $$

(51)

With more frequencies, these frequency-power formulas or Manley–Rowe (MR) equations contain additional terms [413].

As optical media are lossless, MR relations govern power flows among interacting laser beams [411]. Interestingly, Equations (51) can be derived from the assumption that wave interactions in a lossless mixer involve single quanta [549].

**Synchronization.** Although scientific interest in synchronized oscillations go back to Christiaan Huygens’s seventeenth century report that two of his recently invented pendulum clocks kept exactly the same time when hanging from the same wooden beam [254], modern research on synchronization began in 1934 with van der Pol’s analysis of two coupled vacuum-tube oscillators, each described by Equation (49) [527]. If the nonlinear parameter $\varepsilon$ is small, the two oscillations are almost sinusoidal, and the oscillators must be closely tuned for frequency locking to occur. As $\varepsilon$ increases with fixed cou-
pling—electronic engineers are aware—synchronization becomes possible over a progressively wider frequency difference of the two uncoupled oscillators.

Motivated in 1958 by an interest in coupling among the brains neurons, Norbert Wiener generalized the coupled oscillator problem from van der Pol’s two to many, with natural frequencies of the uncoupled oscillations described by a Gaussian probability curve. The result of the coupling, he showed, is to collapse the oscillators near the centre of the Gaussian distribution into a single frequency (or delta-function distribution), leaving oscillators on the wings relatively uninfluenced. As Wiener noted, this phenomenon is observed in the frequency locking of generators on an electric power grid, and he suggested that it would be observed in “a great many other physical cases” including planetary motions and molecular spectra. Among examples of the latter, we have seen the emergence of local modes in benzene molecules (see Figure 21), for which the eigenfrequency spread of the six CH-stretching modes are reduced from the nearest neighbor coupling energy to the much smaller (quasidegenerate) value given in Equation (42).

A well-written survey of such physical examples has recently been published by Steven Strogatz, who describes applications to synchronized flashing of Indonesian fireflies, planetary motions, phase transitions, coherent quantum states (lasers, superconductors, superfluids, etc.), and biological oscillators. Interestingly, Strogatz also points out that the neurological data to which Wiener referred in chapter 8 of Nonlinear Problems in Random Theory[554]—entitled “Application to the study of brain waves, random time, and coupled oscillators”—was incorrect. The dynamics of alpha rhythm is the neocortex are more complicated than those of interacting sinusoidal oscillators.

NONLINEAR DIFFUSION. With the invention of the transistor at mid-century by John Bardeen, Walter Brattain and William Shockley, research in solid-state electronics surged. In 1957, Leo Esaki invented a two-terminal solid-state device with negative differential conduction—terminal current is a single-valued function of terminal voltage over some range of voltage—which allows a simple realization of Equation (49), and a corresponding superconductive device was invented by Ivar Giaever in 1960.

As an electrical-engineering graduate student at that time, these developments were of personal interest, for the aim of my research was to study large-area negative-conductance diodes from the perspectives of linear wave theory. Nonlinear effects—it was then widely believed and taught—could be included by either of two methods: (i) Piece-wise linearization, in which linear solutions are obtained over the linear branches of the nonlinear device and then matched at the boundaries between adjacent linear regions. (ii) Quasilinearization, in which a solution is first approximated by a set of linear normal modes.
after which the interactions among modes are computed. Missing was a serious search for global nonlinear solutions that cannot be well represented in either of these two approaches.\(^\text{29}\)

Under appropriate biasing, large area Esaki and Giaever diodes have volt-ampere characteristics similar to the sodium ion current of a nerve fibre membrane (see Equation (31)), leading naturally to schemes for electronic analogs of the nerve fibre [102, 379, 459, 461]. The essential idea is to build an electronic model of the simple nonlinear diffusion equation of Equation (32) with the global solution given in Equation (33), and then augment the system with a recovery variable, which returns the voltage to its original resting value.

In thinking about global solutions of PDEs that describe nerve-axon models, a key idea was to assume the existence of travelling-waves (either periodic or localized), for which all dependent variables are functions of \(x - vt\), where \(v\) is velocity parameter that is unknown at the outset of the analysis. This travelling-wave assumption connects partial derivatives as \(\frac{\partial}{\partial t} = -v\frac{\partial}{\partial x}\), thereby reducing the original PDE system to an ODE system that can be studied using phase-space techniques [222, 461]. In the course of such analyses, it turns out that the speed \((v)\) takes a fixed value at which energy is released by the global travelling wave at the rate that it is consumed by dissipative processes. Hodgkin and Huxley used a travelling-wave analysis in their classic 1952 study of nerve impulse propagation on squid axons [244].

**SHOCK WAVES AND SOLITONS.** Thinking about localized travelling-wave solutions for nonlinear dissipative systems in the early 1960s led several researchers to consider whether similar solutions can be found for nonlinear PDEs that conserve energy. Again, the cornucopia of solid-state electronics provides several possibilities, including the voltage dependence capacitance of a reverse-biased silicon diode, which have been widely used since the 1950s for the frequency-locking circuits of FM radios. It is a simple matter to construct an \(L\), \(C\) transmission line in which the series inductors are standard linear elements and the shunt capacitors are varactors. Because non-dispersive lines develop shock waves, they can be used as harmonic generators, but if varactors are employed as elements in dispersive transmission lines, one finds the Korteweg–de Vries (KdV) equation to lowest order. Investigations of these nonlinear electrical transmission lines models for KdV have long been carried on by electrical engineers, going back to the early 1960s [461], and they are useful not only for scientific investigations [308] but also for pedagogical demonstrations [432].

Another possibility of electronic solitons emerged in 1963 from Brian Josephson’s invention of a superconducting diode for which current is related to voltage as \(I \propto \sin[(2\pi/\Phi_0) \int V \, dt]\), where \(\Phi_0 = 2.068 \times 10^{-15}\) volt-seconds is a quantum of magnetic flux [31, 306, 378].\(^{30}\) From an electrical engineering per-

\(\text{\textsuperscript{29}}\)In the fall of 1960, Professor Jun-ichi Nishizawa from Tohoku University visited my laboratory, saw what I was doing, and commented: “We are working on similar problems in Japan, but we are using nonlinear wave theory.” This was the first time I heard the term that is a central subject of this review.

\(\text{\textsuperscript{30}}\)Esaki, Giaever, and Josephson shared the 1973 Nobel Prize in physics for their inventions of these three tunnelling devices.
spective, Josephson diodes are nonlinear inductors, and upon being extended in one dimension they are described by the sine-Gordon (SG) equation, for which the kink solution shown in Figure 9 represents a moving flux quantum or “fluxon” [460].

Figure 25: A Josephson oscillator in which a quantum of magnetic flux (fluxon) bounces back-and-forth between the two ends of a rectangular junction.

As shown in Figure 25, fluxons can be used in an oscillator of technical importance. If the bias current is zero and dissipative effects are neglected, the fluxon shown in this figure will bounce back and forth in its rectangular box with a frequency equal to $v/L$ and a voltage that can be expressed exactly in terms of products of Jacobi elliptic functions [467]. In general, $v$ is determined by a balance between energy input from the bias current and dissipation, leading to an oscillator that converts direct current to oscillations at frequencies of hundreds of GHz [31, 306]. For radio astronomy in the range of 100 to 500 GHz, the present standard is a superconducting heterodyne receiver with the local oscillator a long Josephson junction [409].

O(e.g.,ne of the creative tensions of nonlinear wave theory involves those who claim that the oscillator of Figure 25 can be analyzed in terms of interactions among the linear electromagnetic modes of the cavity and those who claim that this is not feasible in practice, and even if it were the quasiharmonic approach would forgo possibilities for using perturbation theory around the analytic and fully nonlinear solutions [350]. In an interesting recent paper, numerical and experimental studies of annular Josephson junctions have been conducted with sizes about equal to those of the fluxons, finding a range of parameter values where both of the pictures are appropriate [380]. More generally, studies of the Josephson effect – and the corresponding SG equations – continue to play roles in a variety of fundamental topics, including non-conventional order parameters in high $T_c$ and other classes of superconductors, macroscopic quantum tunneling, and cosmology (e.g.,[28, 29]).
In a 1927 study of frequency demultiplication, written together with Jan van der Mark, van der Pol reported that at certain driving frequencies an “irregular noise” was observed [528]. Although it was disregarded at the time, this was the first report of low-dimensional chaos in electronics.

As noted above, Ueda and his colleagues observed chaotic behaviour in the early 1960s in periodically driven nonlinear oscillator – for example, van der Pol’s equation with a non-resonant periodic driving force on the right-hand side [224]. This work was not widely distributed, perhaps because the authors were not sure that it was a correct observation. By the end of the 1970s, as we have seen, many chaotic systems had been studied, and it was clear that chaotic circuits could easily be constructed from selections of the two-terminal elements described above: varactors or tunnel diodes along with linear circuit elements and batteries.

While visiting Japan in 1983, for example, Leon Chua proposed an oscillator circuit comprising a linear inductor, a linear resistor, two linear capacitors, and a nonlinear conductance with more negative values near the origin [265]. Somewhat akin to the Rössler system of three ODEs [193], “Chua’s circuit” displays an impressive variety of behaviours, including a bifurcation sequence leading to chaos. More recently, a variety of solid-state electronic devices have been developed by Richard Taylor and his colleagues which share the chaotic properties of electronic billiards [509].

### 3.7 Nonlinear optics

Prior to the 1960 inventions of the solid-state laser by Theodore Maiman and the gas laser by Ali Javan, optics was entirely a linear science because without lasers the electric field amplitudes available from incoherent light sources are not large enough for nonlinear effects to be seen. Predicted in the late 1950s by Arthur Schawlow and Charles Townes in the US and by Nicolai Basov and Aleksandr Prokhorov in Russia as a realization of the maser (Microwave Amplification by Stimulated Emission of Radiation) at shorter wave lengths, the laser (Light Amplification by Stimulated Emission of Radiation) could have been invented at any time after 1917, when a key paper on stimulated emission of radiation was published by Einstein [151].

Lasers. In its simplest form, a laser comprises a population of quantum-mechanical entities (atoms, molecules, electronic states in crystals, etc.) with an upper (excited) energy level \(E^+\) and a lower level \(E^-\), with corresponding populations \((N^+)\) and \((N^-)\), which are coupled to an electromagnetic cavity with intensity \(I\). If \(N \equiv N^+ - N^-\), laser rate equations take the form [492]

\[
\begin{align*}
\frac{dI}{dt} &= (\alpha N - \varepsilon) I \\
\frac{dN}{dt} &= \gamma (N_0 - N) - 2\alpha NI,
\end{align*}
\]

31The 1964 Nobel Prize in physics was awarded to Townes, Basov and Prokhorov for developing the maser-laser principle.
where $\varepsilon$ is the natural decay rate for the cavity energy, $\gamma$ is the natural decay rate for the population inversion ($N$), $\alpha$ is a proportionality constant for stimulated emission, and $N_0$ is a steady-state level of inversion for $I = 0$ which is proportional to the pumping rate. With $\alpha N > \varepsilon$, the lasing threshold is attained, and the energy of the electromagnetic field within the cavity begins to grow; for $\alpha N_0 \gg \varepsilon$ this field becomes very intense in a manner that is related to Bose–Einstein condensation [571].

With $\alpha N_0 < \varepsilon$, $I = 0$, and the second of Equations (52) is related to the growth of activity in nonlinear science which is shown in Figure 1. More generally, Equations (52) are similar to those describing predator-prey dynamics of interacting biological populations.

Currently, there are many types of lasers, offering a variety of output wavelengths and performance characteristics, including the following [492]. (i) Solid state lasers, of which the first was Maiman’s ruby ($\text{Al}_2\text{O}_3$) laser, where the active entities are $\text{Cr}^{2+}$ ions and the output wavelength is 347 nm. Solid-state lasers typically employ a host crystal or glass doped with active ions and are described by three- or four-level schemes, with output wavelengths ranging from 280 to 1800 nm. Solid-state lasers are usually operated in a pumped mode, meaning that the population inversion ($N$) is raised to a high level before the lasing action is allowed to begin, which yields a pulsed output beam of modest average power but high instantaneous power. (ii) Gas lasers, of which the first was Javan’s helium-neon laser, where the active entity is an excited atomic level. These lasers are usually operated in a continuous mode and offer a variety of output wavelengths from the carbon dioxide laser ($10.6 \mu m$) to the argon laser ($409–686$ nm). (iii) Semiconductor diode lasers in which the active entities are overlapping electron and hole states of highly doped pn diodes and the pumping (pulsed or continuous) is by direct electrical current. Familiar as laser pointers, these lasers are widely used in optical communications and as optical switches. Output wavelengths range from the 0.33 $\mu m$ of $\text{ZnS}$ to 8.5 $\mu m$ of $\text{PbSe}$. (iv) Liquid lasers in which the active entities are inorganic dyes that are pumped by intense flash lamps. (v) Free-electron lasers in which the active entity is a relativistic electron beam. These large installations are operated in tunable pulse mode with output wavelengths from ir to x-ray.

Shortly after Maiman’s invention of the ruby laser, Peter Franken and his colleagues passed its high-intensity beam through quartz—converting red light (347 nm) to blue (694 nm) and introducing nonlinear optics as a new realm of research [48, 176, 369, 370, 371]. Among the many problem areas that emerged are the following. First, an aim of the laser engineer is to obtain an output beam that is coherent over the face of the cavity, which is sometimes difficult because a pure lasing mode tends to break up into independent filaments. This difficulty is related to the general problem of nonlinear distributed oscillators, in which the van der Pol equation is generalized to one-, two-, or three-space dimensions ($d = 1, 2, 3$) and independent quasiharmonic modes are allowed only for $d \geq 2$ [470]. Second, the dynamics of a laser must often be described by three or four levels, rather than two as in the simple formulation of Equations (52). In such cases, the laser rate equations may have chaotic solutions, which
are usually undesired from an engineering perspective but may be of scientific interest. Third, it was through studies of laser dynamics that Hermann Haken was led to his formulation of “synergetics” as a general description of cooperative nonlinear dynamics in physics, chemistry, biology, psychology, and sociology (which will be discussed in Section 3.12). 32 [213, 215]. Fourth, optical fibres can support solitons. Finally, advances in laser technology allow pump-probe spectroscopy in the sub-picosecond range. Let us examine these last two items.

MODULATIONAL INSTABILITY. Instability of a periodic wave in a nonlinear optical medium was first discussed by Ostrovsky in 1966 [402] and shortly thereafter by Benjamin and Feir in the context of deep water waves [38]33. This phenomenon has been analyzed in detail by Whitham who has developed an averaged Lagrangian analysis to derive nonlinear equations for the slowly-varying modulations of periodic waves – which are Jacobi elliptic functions rather than sinusoids [551]. Whitham’s formulation is interesting as an analytic example of nonlinear phenomena at two hierarchical levels of description.

In the nonlinear Schrödinger (NLS) system of Equation (15), a negative sign before the last $(2|u|^2u)$ term on the left-hand side implies modulational stability. With a positive sign before this nonlinear term, however, periodic solutions are unstable, and multi-soliton solutions will evolve from arbitrary initial conditions [400, 404]. In the context of nonlinear science, the latter case is more interesting.

SOLITONS ON OPTICAL FIBRES. Optical fibres offer a means for transmitting electromagnetic waves at the wavelengths of visible light, which is akin to the wave guide technology developed for microwaves in connection with radar during the Second World War. In 1973, Akira Hasegawa and Fred Tappert conceived of using the nonlinear dielectric constant in the core of an optical fibre as the basis for soliton transmission of information [221]. Their basic idea was that each bit would be stably carried by a soliton of the nonlinear Schrödinger (NLS) equation [268].

Although this concept has been brilliantly realized by Linn Molenauer and his colleagues over the past three decades [368], the initial reaction of the Bell Laboratories (BTL) was unenthusiastic [575]. In coming to this conclusion, BTL management was influenced by the incorrect belief that real optical fibres would have dispersion for which the sign before the $2|u|^2u$ term in Equation (15) is negative, allowing only dark solitons (blank spots in a continuous beam) to form which would require too much power.

Hasegawa and Tappert were unaware in 1973 that the NLS equation was exactly integrable; thus two deeper problems with gaining acceptance of their optical soliton proposal were: the Hasegawa–Tappert proposal came just be-

---

32 Note that Zabusky uses the term “synergetics” to mean the use of “computers as an heuristic aid to synergize mathematical progress” [575].
33 Inappropriately, this is referred to as a “Benjamin–Feir instability” rather than an “Ostrovsky instability” because of the unfortunate tendency of western scientists to avoid reading Russian journals during the Cold War.
For the explosion of interest in nonlinear science that is illustrated in Figure 1, and their proposal was in conflict with commitments by most BTL engineers to systems based on linear waves. Again we find an example of Kuhn’s claim that significant advances in science are often impeded by those with established priorities.

**Pump-Probe Spectroscopy.** From the above survey of lasers that have been developed over the past four decades, it is clear that nonlinear optics is now a sophisticated activity, offering many possibilities for novel experiments [369, 370]. Among the most important of these for nonlinear science is pump-probe spectroscopy using pulses in the femtosecond range, which allows unambiguous determination of nonlinear localization.

![Figure 26: (a) Sketch of a pump-probe setup. (b) Pump-probe experiment of an excited mode.](image)

To see how this works, consider the sketches of a pump-probe experiment in Figure 26, where an experimental question is to decide whether or not the oscillating mode giving rise to a particular spectral line is extended over the sample (linear) or localized over a small region by nonlinear effects. In pump-probe spectroscopy, the key measurements are made not from the ground state but from the first excited state, to which the sample is brought through the action of a pump pulse, as shown in Figure 26(a).

From Figure 26(b), the probe beam sees two upward transitions, \( (0 \rightarrow 1) \) and \( (1 \rightarrow 2) \), and a downward transition, \( (1 \rightarrow 0) \). If the oscillating mode is spread out over the crystal, its amplitude is small and the response is necessarily linear. In this case, the measurement with and without the pump is exactly the same [467]. In other words, the net absorption from the pumped system is the absorptions from the \( (0 \rightarrow 1) \) and \( (1 \rightarrow 2) \) transitions minus the stimulated emission from the \( (1 \rightarrow 0) \) transition (minus since stimulated emission is equivalent to a negative absorption), which is just equal to the absorption from the \( (0 \rightarrow 1) \) transition with no pumping. Thus subtraction of measurements with and without the pump gives zero.

If the oscillating mode is localized, however, this balance is disturbed because the spacing between the 0 and 1 levels is no longer equal to the spacing between the 1 and 2 levels. Thus subtraction of the readings with and without the pump is no longer zero, but a value that can be used to determine the...
nonlinear character of the mode. As recently developed by Hamm and his colleagues, this method has been used to demonstrate that the 1650 cm⁻¹ line of crystalline acetanilide (ACN) shown in Figure 23 is a localized mode [144]. The same method has been used to show that three of the room-temperature NH stretching lines in ACN are also localized [145].

Important features of pump-probe spectroscopy include the following. First, the entire setup can be assembled on an optical bench, giving the experimenter full-time access to the measurement for as long as is needed. This is important because previous time-resolved measurements on the nonlinear modes of proteins have been limited to free-electron lasers, for which beam time must be scheduled. Second, the pump pulse can be made very short (in the femtosecond range), allowing observation of the time course of the lattice oscillations that contribute to intrinsic localization; thus these lattice modes can be unambiguously identified [144, 145]. Finally, recent progress in laser technology makes it possible to perform pump-probe measurements from far infrared to soft x-rays with time resolutions much less than a picosecond [154]. With these several advantages, it is to be expected that pump-probe spectroscopy will play a central role in the future development of experimental nonlinear science.

3.8 Nonlinear fluids

The oceans have fascinated, frightened and inspired our ancestors since the dawn of human experience and sound is closely tied to our speech and music; thus dynamics of fluids have been of keen interest to Western scientists since the modern era began some four centuries ago. Fluid waves are often nonlinear so they have been considered above in this review, but more details are presented here.

Studies of water waves are especially interesting because the air-water interface is both a dependent variable and a boundary condition [204]. From the perspective of Aristotle’s categories of causality (see Section 1.1), in other words, the local wave height is both an effect of an efficient cause and a formal cause of that effect [16]. This sort of positive feedback loop is an essential feature in many areas of nonlinear science, leading to the emergence of qualitatively new phenomena in a wide variety of applications. Interestingly, Einstein’s equations of general relativity (see Equations (37)) play a related role by defining the structure of the universe.

SUPersonic waves. In Section 3.4, we noted Captain Parry’s unanticipated measurement that the “boom” of a cannon moves through the atmosphere ten percent faster than sound waves of small amplitude – an observation supported by theoretical and numerical results for nonlinear spring-mass chains [186, 187, 188, 189] and one-dimensional rods [106]. These examples are qualitatively similar to Russell’s result for solitary water waves given in Equation (8), which describes the dynamics of the recent Boxing-day tsunami shown in Figure 13.
As a well-monitored example of a supersonic wave (SSW), consider Figure 27 which plots the radius of expansion ($R$) from the first atomic explosion (Trinity) as a function of time ($t$). Dimensional analysis shows that $R^5/t^2 \propto E/\rho$ where $\rho$ is air density and $E$ is the energy released by the blast (equivalent to about 20 kilotons of TNT) [563]. At a radius of 100 m, the hemispherical wave is expanding at 2400 m/s or about seven times the speed of sound.

In the above examples – tsunamis, cannons and atomic blasts – the SSWs manage to detach themselves from their underlying linear systems. Thus SSWs differ from the bow wave of a boat and the shock waves produced by bullets and supersonic aircraft in a manner that is qualitatively similar to the difference between the near field of a radio antenna and its radiation field – bow waves and sonic booms are tied to the objects that cause them, whereas SSWs are no longer connected to their initiating earthquakes or explosions, each emerging with an independent ontological status.

Throughout his professional life, John Scott Russell remained convinced of the importance of such solitary-wave phenomena – not only in the canals and water tanks which he had so thoroughly investigated in the 1830s – but in the atmosphere and the cosmos. In 1885, three years after Russell’s death, a book bearing his name was published, which presents his general perspective in three diverse areas [449]. (i) **Shallow water waves.** Based upon his extensive empirical studies, Russell’s final thoughts in this area are well founded as has been discussed above. (ii) **Sound waves in the atmosphere.** Russell viewed the Earth’s atmosphere as an “ocean of air” through which high amplitude sound waves (the “boom” of Captain Parry’s cannon) would move at supersonic speeds, as do shallow water waves. Using the formula $v = \sqrt{g(d + h)}$ (where $v$ is 1100 ft/s, $g$ is 32 ft/s$^2$ and $d \gg h$), he computed the height of the atmosphere as $d = 7.2$ miles – an acceptable numerical value but wrong in
concept. For Parry’s SSW, an atmospheric hump \((h)\) of \(0.2 \times 7.2 = 1.4\) miles would be required, which is physically unreasonable. As we have seen in Section 3.5, a credible explanation of SSWs emerges from the theory of nonlinear compression waves. (iii) The great ocean of ether. Here Russell’s posthumous book is incorrect. Aside from several typos which he didn’t have the opportunity to amend, Russell argued by analogy with water waves and sound waves to obtain an estimate of the size of the universe that is widely at variance with present empirical evidence. Nonetheless, his idea of generalizing nonlinear phenomena from water waves and atmospheric waves to the cosmos may have merit, suggesting that astrophysicists remain open to the possibility of nonlinear gravitational phenomena exceeding the speed of light [431, 478].

SHOCK WAVES. In contrast to supersonic waves, shock waves are experienced by every child who romps on an ocean beach. The basic phenomenon is described by the seemingly simple equation

\[
\frac{\partial u}{\partial t} + (c + \beta u) \frac{\partial u}{\partial x} = 0 ,
\]

which for the initial condition \(u(x, 0) = f(x)\) has the solution \(u(x, t) = f(x - (c + \beta f)t)\). For \(\beta > 0\), higher amplitude portions of a wave travel faster than those of lower amplitude, leading to the multivalued breaking phenomenon commonly observed on beaches. In more general terms, Equation (53) is both nonlinear and non-dispersive, forcing all harmonics to travel at the same speed [202, 381]. Equation (53) also describes certain traffic waves, where \(u\) represents the local density of automobiles. In this application, \(\beta < 0\), which causes shocks to form behind regions of increased density, as many drivers have experienced [228, 551].

Figure 28: A tidal bore (or mascaret) at Quillebeuf, France. The site is about 36 km from the mouth of the Seine, where tides up to 10 m occur. Due to dredging in the 1960s, this striking nonlinear phenomenon has disappeared. (Courtesy of Jean-Jacques Malandain.)
In acoustics, multivalued breaking is not possible so shocks form, and their thickness is determined by the detailed effects of thermal conductivity, which can be approximately described by the Burgers equation \[381, 403, 412\]

\[
\frac{\partial u}{\partial t} + (c + \beta u) \frac{\partial u}{\partial x} = \delta \frac{\partial^2 u}{\partial x^2}.
\]

Interestingly, this equation can be exactly solved for arbitrary initial conditions, just as can the KdV equation discussed in Section 2.2 \[551\]. Hydraulic jumps are also described by Equation (54), a dramatic example of which is the tidal bore shown in Figure 28.

Closely related in both water waves and acoustics are the bow waves which emanate from a moving boat and the atmospheric shock waves generated by a bullet or a supersonic aircraft. In these familiar cases, which have been widely studied over the past two centuries \[551\], the boat, bullet or aircraft travels at a velocity \((v)\) that is greater than the low-amplitude wave speed \((c)\) in the medium (water or air), and a shock cone develops which – by the Huygens principle \[530\] – travels at speed \(c\) and angle \(\theta = \cos^{-1}(c/v)\) (see Figure 29) \[401\].

![Figure 29: Shadowgraph showing shock waves produced by a Winchester 0.308 caliber bullet travelling through air at about 2.5 times the speed of sound. (Courtesy of Ruprecht Nennstiel, Bundeskriminalamt Wiesbaden, Germany.)](image)

**PLASMA WAVES.** Sometimes called a “fourth state of matter,” a plasma is a fluid in which the constituent particles are charged – combining hydrodynamics with electromagnetism to form magnetohydrodynamics (MHD) \[77\]. Under MHD, motions of charged particles generate magnetic fields, which retroactively influence the charged particles; thus forming causal feedback loops that lead to many nonlinear phenomena.

In a highly-conducting plasma, magnetic field lines \((B)\) act like elastic strings, conducting waves – proposed in 1942 by Hannes Alfvén \[10\] – with low-amplitude...
(linear) velocity $v_A = \sqrt{B^2/\mu_0 \rho}$, where $\rho$ is the mass density of the plasma and $\mu_0$ is magnetic permeability [556]. Very nonlinear at higher amplitudes, Alfvén waves play roles in studies of the Earth’s magnetosphere, solar physics and quasars.

Plasma confinement devices for power generation by fusion comprise another important area of MHD research. In these plasmas, low amplitude waves may be transverse (like ordinary electromagnetic waves) or longitudinal, with dispersion relations modified by the presence of one or more plasma frequencies: $\omega_p = \sqrt{n q^2/e m \varepsilon_0}$, where $n$ is the density of charged particles, $q$ and $m$ their charge and mass, and $\varepsilon_0$ is dielectric permittivity [485]. At larger amplitudes, these waves become nonlinear and can be described by the KdV equation [255, 544], or a two-dimensional version – the Kadomtsev–Petviashvili (KP) equation [122, 309, 310] [34].

Recently, Leon Shohet has argued convincingly that “slinky modes” – which bedevil toroidal plasma-confinement machines through generation of “hot spots” – are described by the same damped-driven sine-Gordon (DDSG) equation that is successfully used to model magnetic flux dynamics on long Josephson junctions (see Figure 25) [139, 350, 485]. Although much empirical evidence supports this description of slinky modes [486], many plasma physicists prefer to use a quasiharmonic approach in which “mode locking” is a primary phenomenon rather than a fully nonlinear analysis under which mode locking arises as an artifact of DDSG kink formation [484]. Thus the insights of modern nonlinear science have not been uniformly appreciated throughout the scientific community.

A curious occurrence – often ignored, involving plasmas, and almost certainly nonlinear in nature – is “ball lightning” (BL) [428]. As BL has never been produced in a laboratory, it’s properties are derived from reports, some of which are summarized as follows: BL often follows a lightning flash, is usually spherical with a diameter of tens of centimetres, has a lifetime of a few seconds, and is bright enough to be seen in daylight. This interesting phenomenon presents a challenge to nonlinear scientists of the future [35].

**Oceanic and Atmospheric Dynamics.** Dramatic examples of nonlinear waves on the surface of the ocean are provided by “rogue waves” (RWs) (also called “freak” or “monster” waves) and nearby “holes” which create precipitous surface slopes. Long dismissed as the unreliable tales of sailors, RWs are now recognized as real phenomena that can have disastrous maritime consequences. From the statistical perspectives of linear wave theory, the sudden accumulation of localized energy in a surface wave of unusual amplitude is highly unlikely, but recent observations show that RWs occur far more often than has hitherto been expected [138]. Although RWs are not often pho-

---

[34] Karl Lonngren has suggested that the large amplitude of the tsunami near Sri Lanka in Figure 13(a) may be modelled by the KP equation.

[35] Dirk ter Haar divided people into the following classes: (1) Those who have seen BL, (2) Those who know someone who has seen BL, (3) Those who know someone who knows someone who has seen BL, and so on. The present author is a “2”.
tographed (for obvious reasons) an example is shown in Figure 30(a), and in Figure 30(b) a RW is being generated in the Large Wave Channel of Hannover [452] using a caisson model from the Admiralty Breakwater, Alderney (UK). Guided by computer modeling, these waves occur when slow waves are overtaken by faster waves, which then merge together in a typically nonlinear manner.

Two-dimensional surface waves on water are very complicated phenomena, but progress has been made in understanding their nonlinear dynamics by considering the nonlinear Schrödinger (NLS) equation of soliton theory, which serves as a model for one-dimensional surface waves [39, 399, 397]. Using results from the IST method (see Section 2.2), one finds analytic expressions for unstable nonlinear solutions that rise up out of the background field and then quickly disappear, much like the ocean wave and tank wave of Figures 30(a) and 30(b) respectively. To extend such analytic results to the more realistic problem of two spatial dimensions, Alfred Osborne and his colleagues have conducted extensive numerical studies of an NLS equation with two-space variables, showing that RWs and their nearby holes are equally characteristic phenomena. A striking example of their results is presented in Figure 30(c), where the parameters of the model have been matched to open ocean and the axes are labeled in metres. Because an accurate understanding of the statistics of RW formation is critical for the design of ships and oil platforms, this research indicates the technical importance of nonlinear science in an area where progress has long been impeded by the conceptual limitations of linear wave theory.

The atmosphere offers a wealth of nonlinear phenomena – some fascinating and some terrifying, as is the tornado shown in Figure 31(b). Formed within a cumulonimbus cloud, a tornado is driven by the temperature inversion of a cold front and can develop wind speeds of several hundred miles per hour. More benign are the “dust devils” which are driven by localized heating and convection and often appear in deserts of the American Southwest on hot summer afternoons (see Figure 31(a)). On a larger scale are hurricanes like Andrew in August of 1992, a satellite photo of which is shown in Figure 31(c). Called a typhoon in the West Pacific Ocean and a cyclone in the South Pacific, hurricane winds rotate clockwise below the equator and counterclockwise above with wind speeds of 100 to 200 mph which can last for several weeks with disastrous consequences. A 1737 hurricane in the Bay of Bengal claimed 300,000 lives, and an 1881 typhoon in China took a similar toll [406].

Thinking about tornadoes and hurricanes brings to mind the difficult problem of weather prediction, which was commented upon by both Poincaré and Wiener and led Lorenz to rediscover Birkhoff’s “irregular orbits” [46, 311]. Famously characterized by Lorenz as exhibiting the “butterfly effect”, the system of Equations (3) with dynamics shown in Figure 5 is a simple model of a weather system, which became widely recognized as an example of low-dimensional chaos during the 1970s [531].

Somewhat less problematic than weather prediction is the development of numerical models for global behaviour of the coupled ocean-atmospheric sys-
Figure 30: Rogue waves (RWs). (a) A RW approaching the stern of a merchant ship in the Bay of Biscay. (Courtesy of NOAA Photo Library.) (b) A RW produced in the Large Wave Channel in Hannover. (Courtesy of the Coastal Research Centre, Hannover, Germany.) (c) A snapshot at the peak value of a numerical RW computed from a 2 + 1 dimensional NLS system, where the parameters are chosen to correspond to the ocean and axes are labeled in metres. (Courtesy of Alfred Osborne.)
Figure 31: Natural atmospheric vortices. (a) A dust devil in Death Valley, California. (Courtesy of Dr. Sharon Johnson/GeoImages.) (b) Photo of a tornado (right) illuminated by a bolt of lightning (left). (Courtesy of National Oceanic and Atmospheric Administration.) (c) Satellite view of Hurricane Andrew, approaching the U.S. Gulf Coast, 25 August 1992. (Courtesy of National Oceanic and Atmospheric Administration.)
tem, which establishes patterns of ocean currents and atmospheric structure [8]. Such nonlinear studies lead in turn to considerations of global warming caused by human emissions of carbon dioxide [490].

**Turbulence.** In an 1839, Gotthilf Hagan described two distinct regimes of fluid flow through a pipe [209], and almost a half century later Osborne Reynolds showed that the transition between these two regimes – one smooth and the other turbulent – depends only on $Re = \frac{vd}{\nu}$ [75, 76, 433]. In this dimensionless number (which now bears Reynolds’s name), $v$ is the average flow velocity, $d$ is the pipe diameter, and $\nu$ is the fluid viscosity.

It is now known that turbulent flow is a numerical property of the Navier–Stokes equation, which describe the dynamics of viscous fluids [7]. Although turbulence is evident in the bore shown in Figure 28 and in the wake of the bullet shown in Figure 29, the difficulties associated with fully understanding the nature of this phenomenon have been amusingly characterized in 1932 by Horace Lamb, who said [24]: “When I die and go to heaven, there are two matters on which I hope enlightenment. One is quantum electrodynamics and the other is turbulence of fluids. About the former, I am really rather optimistic.”

Those insights that have been explained to Professor Lamb since his passing would include Andrej Kolmogorov’s 1941 result that the energy in a fully turbulent fluid should depend on wave number as the 5/3 power [275]. Although it has doubtless been pointed out to him that this law is invalid in systems with two or more spatial dimensions, many real systems are essentially one-dimensional for which the 5/3 law holds over several decades. Shortly after the Kolmogorov result was published, Jan Burgers proposed Equation (54) as a simple model of turbulence in a bore (see Figure 28) that can be exactly solved [74, 412, 551]. Local instabilities of the sort that generate the RWs shown in Figure 30 will certainly be important, but for a full understanding of turbulent phenomena we must wait.

### 3.9 Economic dynamics

Apart from barter, economic activities involve exchanges of goods for currencies, currency exchanges, future contracts, etc., most of which are exactly recorded. Thus the raw data available on economic activities is more complete than for fluids, but there is an additional qualitative feature that makes economic systems more difficult to understand. The basic economic agents are cognitive entities – human beings – each with her own ideas concerning the nature of the transaction. In Aristotelian terms (see Section 1.1), final causes enter the picture, offering additional possibilities for closed causal loops [16], which economists describe as expectations feedback. As Cars Hommes puts it [248]:

Decisions of economic agents are based upon their expectations and beliefs about the future state of the economy. Through these decisions, expectations feed back into the economy and affect actual realization of economic variables. These realizations lead to new
expectations, in turn affecting new realizations, implying an infinite sequence of expectational feedback. For example, in the stock market, optimistic expectations that stock prices will rise will lead to a larger demand for stocks, which will cause stock prices to rise. This process may lead to a self-fulfilling speculative bubble in the stock market. A theory of expectation formation is, therefore, a crucial part of economics, in particular for modeling dynamic asset markets.

In the jargon of electrical engineering, the gain around closed causal loops including expectations feedback can be either positive or negative. As Hommes points out, such positive feedback can lead to speculative bubbles – like the “tulip-bulb craze” of seventeenth-century Holland or the “dot-com bubble” of the late 1990s – and it may help to explain business cycles, as was assumed in the economic models of John Maynard Keynes. From a theoretical perspective, negative feedback is more problematic. On the one hand, a modest amount of negative feedback may act as a governor on the economic system, by reducing (or increasing) the money supply during inflationary (or deflationary) periods. On the other hand, a really reliable prediction of an undesired event may induce human effort to prevent it. In other words, perfect prediction of “A” may imply “NOT A” – an uncomfortable thought for those with a reductive turn of mind.

In the 1960s, this discomfort led to the development of a theory of rational expectations (RE), under which economic agents are assumed to “use all available information, including economic theory, to form optimal forecasts and that, on average, expectations coincide with realizations” [248]. This RE doctrine implies perfect foresight in deterministic economic models and no systematic bias in stochastic models. In the stochastic limit, therefore, an RE theory supposes that temporal variations in the various components of economic activity stem entirely from external events – disruptions of supply or demand, political changes, natural disasters, etc. In the deterministic limit, on the other hand, RE eliminates the influence of expectations feedback through unrealistic assumptions and views the economic system as a static balance among opposing forces of supply and demand, as had been proposed by Alfred Marshall in 1890 [336].

Why are the assumptions underlying RE unrealistic? In the jargon of electrical engineering, RE supposes that the myriad loops of expectations feedback are stabilized, so there are well-defined relationships between causes and effects as in Equation (50). But as engineers well know (or soon learn), it is all too easy for a feedback loop to change from negative to positive loop gain, leading to instabilities that confound cause-effect relations. Indeed, the point of Bode’s classic book is to show how to avoid such instabilities in the course of amplifier design, and this is a complicated task in a well-defined electronic circuit [50]. To assume Bode’s stability conditions are satisfied in the poorly defined context of an economic system is unwarranted.
Some economists – stimulated by the nonlinear science revolution of the 1970s – began to view the objects of their studies as complex dynamical systems in which unpredictable variations are assigned to deterministic (low-dimensional) chaos [117]. Others have developed the concept of bounded rational expectations (BRE), under which agents act with limited knowledge that is upgraded and adapted according to their commercial experiences. Although more realistic than RE, BRE has the disadvantage of introducing many adjustable parameters into the model – which is always problematic – and questions of stability are still ignored.

In the 1980s, the Santa Fe Institute entered the picture, supporting the complex dynamical systems formulation of economic models and introducing a perspective called “econophysics” [14], an approach that has been strongly criticized by a science writer as merely “the latest is a series of failed fads” – cybernetics, catastrophe theory, chaos, and complexity [251]. A balanced review of this occasionally intemperate discourse has been published by Barkley Rosser together with a useful list of references through the end of the twentieth century [440].

Because their basic elements are cognitive, economic systems are complicated in ways that fluids – for all of Horace Lamb’s despair – are not. While some relatively stable commodity markets may be approximately described by supply and demand curves or by BRE models, any realistic global formulation of an economic system dynamics must include both its internal nonlinearity (including positive feedback loops based on expectations) and nonlinear responses to external events (political changes, natural disasters, unanticipated acts of terrorism, etc.), which may be partly psychological in nature.

Amid all these complications, nonlinear science is expected to play an increasingly important role in the future developments of economic theory. In the following section, we will see how economic ideas developed in the late eighteenth century profoundly influenced the course of nineteenth-century biology.

3.10 Biophysics

In contrast with the physical sciences, the dynamics of life sciences are almost entirely nonlinear and living organisms possess varying degrees of cognitive ability; thus the above difficulties in formulating economic dynamics are expected to be even more severe in biology. Nevertheless as was noted in Section 1.2, Equation (2) describing population growth was studied by Verhulst in 1845 – a successful early application of nonlinear science to biology [425, 535]. With the general solution in Equation (1), Verhulst’s equation shows that as the population of a species increases, the average amount of resources supporting each individual decreases until nothing more is available to support growth and the population reaches a maximum value.\(^{36}\) His aim was to refine the widely dis-

\(^{36}\) Verhulst predicted that the population of his native Belgium would eventually limit at 9,400,000, which is remarkably close to the present population of about 10,258,000.
cussed ideas of Thomas Malthus, an English economist who argued in his 1798 *Essay on the Principle of Population* that biological populations are expected to grow exponentially, eventually running into growth limits characterized by mass starvation—leading Thomas Carlyle to describe economics as a “dismal science”.

**The Origin of Species.** It was from their mid-nineteenth-century readings of Malthus’s essay that Charles Darwin and Alfred Russel Wallace independently hit upon the idea that population pressure acts a biological force driving processes of “natural selection” (NS) and causing the emergence of new biological species without divine intervention [109]. In Kuhnian terms, the Darwin–Wallace theory of NS was—along with the heliocentric theory of Copernicus—one of the great revolutions of Western science, having equally profound theological implications [70, 278, 279].

The eventual acceptance of NS by the scientific community was by no means a foregone conclusion in November of 1859, when *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* was offered to the Victorian public [70]. Just as with the Copernican revolution, NS was resisted both on religious grounds because it implied new perspectives on the nature of God and the place of humanity in the universe, and by senior scientists whose established positions were threatened [70]—the latter motivation being particularly strong for the resistance mounted by Louis Agassiz, Harvard University’s leading biologist.37

Although a shy man who avoided the public, Darwin ably promoted his work in several ways, including the following [70]. First, *Origin* is a masterpiece of scientific exposition, which considered all known objections to NS. Second, Darwin avoided unnecessary contention by generously allowing Wallace full credit for his independent discovery of NS. Third, Darwin mailed about ninety prepublication copies of his book to scientific leaders throughout Britain, Europe and the United States along with personal letters explaining the implications of his theory and asking for comments. Fourth, he honestly thought about and responded to the many questions, criticisms and comments received from readers. Finally, Darwin was assisted in his efforts by several leading British scientists—foremost of whom was Thomas H. Huxley, popularly known as “Darwin’s bulldog”. Regarding the wide acceptance of his work, Darwin recorded the following observation [110].

> It has sometimes been said that the success of *Origin* proved “that the subject was in the air,” or “that men’s minds were prepared for it.” I do not think this is strictly true, for I occasionally sounded not a few naturalists, and never happened to come across a single one who seemed to doubt about the permanence of species. […] What I believe was strictly true is that innumerable well-observed facts were stored in the minds of naturalists ready to take their

---

37Some of Agassiz’s arguments are still used by fundamentalist Christians in the United States who continue to reject Darwin’s evolutionary perspectives.
proper places as soon as any theory which would receive them was sufficiently explained.

As noted, the concept of speciation by NS developed from theories for population growth previously formulated by Malthus and Verhulst, which were early applications of nonlinear science to biology. Equations for population growth, in turn, are closely related to those describing laser dynamics and interacting populations of chemical molecules. Thus what is now called Darwinian evolution is of central interest in nonlinear science, involving nonlinear oscillations (simple predator-prey interactions), emergence (the appearance of new species), and deterministic chaos (stemming from the unpredictable interactions among several plant and animal species).

GENETIC TRANSCRIPTION. Born less than a year apart, Charles Darwin and John Scott Russell both died in 1882, after which the theories to which they had devoted their respective lives went into decline. Seemingly superseded by the formulas of Stokes, as we know, Russell’s solitary wave was gradually forgotten until its numerical rediscovery in the 1960s, and by the 1920s a biological mechanism for Darwin’s natural selection had not been found, encouraging a leading biologist to comment that “natural selection is as extinct as the dodo” [19]. The latter oversight is even stranger than the former, as experiments on successive generations of peas by Gregor Mendel, a Czech monk with interests in biology and mathematics, had shown by the mid-1860s that biological inheritance can be of particular characteristics. Although Mendel’s work provides strong support for Darwin’s theory and he had sent a copy of his key paper to Darwin, this envelope lay unopened in Darwin’s study when he died [19].

Taken together, the work of Darwin and Mendel strongly suggested a biological means for the passing on of selected traits – leading to the “modern synthesis” of Ronald Fisher, John Haldane and Sewall Wright [341] – but what was the mechanism? In the mid-1940s, Erwin Schrödinger – who had long been interested in biophysics – published a little book entitled What Is Life?, presenting the case (based on cross-section measurements of fruit-fly mutations caused by gamma rays) that genetic traits must be stored in molecules [372]. Translated into seven languages and selling over 100,000 copies, this book encouraged young James Watson and Francis Crick to search for the molecular basis of inheritance. In a well-told story, they used x-ray diffraction data of Rosalind Franklin to find the “double-helix structure” of DNA [545].

In the double helix, genetic information is stored as patterns of four internal base pairs: thymine (T) which always shares a double hydrogen bond with adenine (A) and cytosine (C) which always shares a triple bond with guanine (G), with combinations of three base pairs needed to determine one of 20 amino acids in the primary structure of a protein. In order to impress its code on a messenger ribonucleic acid (mRNA) molecule and synthesize a protein, it

38Kuhn has compared the phenomenon of biological speciation to the development of a new scientific paradigm [280].
39Watson, Crick, and Maurice Wilkins (Franklin’s supervisor) received a Nobel Prize for this discovery in 1962. Franklin died in 1958.
is necessary for the double helix to locally uncouple, breaking an average of 2.5 hydrogen bonds in a manifestly nonlinear activity. In 1980, James Krumhansl and his colleagues proposed that this local melting could be described by the sine-Gordon equation [158], an idea that has been elaborated upon by Henry Sobell to develop the concept of a “premelton” [26, 491]. Many soliton models of excitations on DNA are discussed in a recent book by Ludmilla Yakushevich [567, 568], and an interesting Hamiltonian for DNA melting has been proposed and studied by Thierry Dauxois and Michel Peyrard [112, 417].

Individual genetic traits (or genes) are embodied in the codes for individual proteins, and the number of possible proteins is finite but very large. For proteins of average size (comprising, say, 200 amino acids), there are \(2^{600}\) possible primary structures, a finite number which Walter Elsasser has termed immense [155]. Why? Because the atomic mass of the universe times its age in picoseconds is about \(10^{118}\), so a number of possibilities greater than \(1\) cannot be individually examined.

Elsasser has pointed out that immense numbers of possible entities in biology lead to a fundamental difference between physical and life sciences. The physicist or chemist can carry out a number of experiments that is large compared with the items under investigation (hydrogen atoms, benzene molecules, etc.), whereas in biology or psychology this is not possible. There are many more possible proteins than have ever existed, and the same is so for human beings [465]. In Elsasser’s terms, physical scientists study homogeneous sets, whereas life scientists study heterogeneous sets [155]. This difference should be kept in mind by psychologists who would model their activities on those of the physicist with the aim of becoming more “scientific”. In the middle of the twentieth century, such “physics envy” motivated the behaviorists, who produced much theoretical nonsense [465].

MORPHOGENESIS. Given a gene which determines the primary structure (or amino-acid sequence) of a particular protein, it is of great interest to know what the final structure of that protein will be. In principle this seems to be a straightforward problem of energy minimization, but in practice it is difficult to carry through because the number of final shapes with approximately the same energy is immense [353]. Although some heuristic approaches to “protein-folding” have been developed for use on currently available computers, these are not satisfactory; thus IBM has recently announced its Blue Gene project to develop a “petaflop” \((10^{15}\) floating point operations per second) computer that is adapted to solving this important problem [11].

Among the most important and least understood of natural phenomena is the development of a living organism from a single cell through its embryonic stages and infancy to adult form. Although a theory of such intricate dynamics lies beyond the present scope of scientific knowledge, some contributions to be made by the nonlinear scientist. In his classic work On Growth and Form, for example, the Scottish polymath D’Arcy Thompson showed how considerations of surface energy can account for many features of the shapes of cells and

---

40 A web site on the premelton is at http://members.localnet.com/~sobell/
their aggregates, thereby contributing to the science of biological morphogenesis [513]. Although the concept of morphogenesis includes dynamics, Thompson’s discussion was almost entirely of static forms, but in 1934 Ludwig von Bertalanffy proposed a biological growth law of the form \( dL/dt = k(L_\infty - L) \) which has the solution \( L(t) = L_\infty (1 - e^{-kt}) \) [536]. Interestingly, this “von Bertalanffy growth equation” is identical to the second of Equations (52) with \( I = 0 \).

Can the study of nonlinear diffusion lead to an understanding of embryonic development? Interestingly, such a proposal was made in 1952 – the same year that the Hodgkin–Huxley (HH) paper appeared [41] – by the English mathematician Alan Turing [520]. To appreciate his ideas, assume the nonlinear diffusion of a propagator \( \psi \) and its inhibitory variable \( \phi \) in three-dimensional space, and consider various relationships between the corresponding diffusion constants \( D_\psi \) and \( D_\phi \). (i) Nerve and muscle. From the perspective of the HH equations, the propagator is a transmembrane voltage, and the inhibitory variable is related to conformational states of protein molecules that are embedded in the cell membrane, conducting ionic currents. Thus for such systems \( D_\psi = 0 \).

(ii) Chemical diffusion. In the Belousov–Zhabotinsky reaction, both the propagator and the inhibitory variable are chemical species in aqueous solution so \( D_\psi \approx D_\phi \). (iii) Turing patterns. For reaction-diffusion systems to develop stationary patterns, it is necessary that \( D_\psi \gg D_\phi \), because this allows the inhibitory variable to surround and contain the outward diffusion of the propagator; thus forming a Turing pattern (TP)

TPs are of interest as models for biological morphogenesis because they don’t move. Although TPs are readily produced in numerical simulations [376], they are difficult to find experimentally in a chemical solution because \( D_\psi \approx D_\phi \). In recent experiments that have demonstrated TPs, \( D_\psi \) has been artificially reduced by fixing the excitatory component in a gel and allowing the inhibitor to diffuse at a normal rate beyond a porous barrier [87, 300]. Such experimental problems suggest that the dynamic aspects of biological morphogenesis are more complicated than was originally anticipated by Turing. In recent books, biologists Stuart Kauffman and Brian Goodwin have emphasized the importance of nonlinear growth processes in understanding the ways that genetic information becomes realized in fully developed organisms, suggesting that Mendel’s concept of relating one gene to each characteristic is overly simplified [201, 263].

**Biological Chaos.** In 1928, van der Pol and van der Mark used a version of Equation (49) to describe the beating of a human heart, opening a new realm of mathematical biology [529]. In this application, \( \varepsilon \gg 1 \), so the oscillation of the relaxation type, which is more easily synchronized than a sinusoidal oscillation. The oscillation of a healthy heart is periodic, with aperiodic or chaotic behaviours indicating problems [194, 207]. In general, a regular heartbeat is associated with a plane RD wave travelling through the muscle, but Winfree has shown that a limited portion of a scroll ring (see Figure 17) can lead to a chaotic quivering (called fibrillation) throughout the heart muscle, causing it

---

41 Interestingly, Andrew Huxley is the grandson of Thomas Huxley.
to cease pumping blood [561, 562].

In his book on synchronism [498], Strogatz shows that the periodic activities of many biological subsystems are both nonlinear and coupled, leading to complicated global behaviors – a point also made by Winfree [562]. With nonlinear dynamics throughout, several biological phenomena are chaotic, beginning with those described by Equation (4), which models annual variations in a population. Within an organism, Leon Glass and Mike Mackey have proposed a nonlinear delay equation of the form $dx/dt = f[x(t - \tau)] - Bx$ to model dynamical disease [195, 320]. Evidently this equation is closely related to Equation (4) and can model, for example, the current flow of white cells into the blood in response to a demand at time $\tau$ in the past.42

**Biological Hierarchies.** To explain how life on Earth may have emerged from the hot Hadean soup, some three or four billion years ago, Manfred Eigen and Peter Schuster have developed the concept of a hypercycle, in which chemical cycles are components of higher-order cycles that are components of yet higher-order cycles [146]. Three or more levels of hierarchically organized dynamics – Eigen and Schuster suggest – are necessary for a dynamic system to gain the ability to reproduce itself and to undergo a rudimentary form of natural selection, thereby initiating Darwinian evolution without divine intervention. Numerical and theoretical investigations of such systems should be of considerable interest to biologically oriented nonlinear scientists over the coming decades [13, 20, 146, 393, 541].

To appreciate this feature of life science, consider the following sketch of living matter as we know it here on Earth:

- **Biosphere**
- **Species**
- **Organisms**
- **Organs**
- **Cells**
- **Processes of replication**
- **Genetic transcription**
- **Biochemical cycles**
- **Biomolecules**
- **Molecules**

Such an organization is not theoretical speculation but empirical observation, as each part of the diagram is studied by scientists who are devoted to understanding the dynamics of that particular hierarchical level.

At every level, nonlinear phenomena lead to the emergence of new entities that provide bases for the dynamics of the next higher level. Thus the biosphere is made up of interacting species of plants and animals, which, in turn, comprise individual organisms, which are composed of organs, which comprise cells, and so on, down to the proteins and nucleic acids, which are formed from chemical molecules. At each level of the diagram, there is an emergence

42See http://chaos.phy.ohiou.edu/thomas/chaos/mackey-glass.html for more details.
of some particular structures which have been selected during the course of evolution from an immense number of possibilities. In other words, the immense number of proteins that might have emerged in the course of evolution is much greater than the number that did emerge.

Although traditional branches of science are directed toward understanding the dynamics at particular levels of this hierarchy, nonlinear interactions among levels are also important, as Strogatz has clearly shown in his description of the various physiological components of circadian rhythms [498]. These interlevel interactions make use of Aristotle’s more general modes of causality sketched in Section 1.1 [16], which become particularly important in the cognitive hierarchy of neuroscience.

### 3.11 Neuroscience

As the most complicated dynamic system known to us, our brains are of central interest in many areas of modern research, including psychology, psychiatry, neurology, philosophy and cognitive science, in addition to neuroscience. Describing the brain is of particular interest for this review, as all levels of its operation – from constituent molecules to interactions with human culture – are governed by highly nonlinear dynamics. Although some of these features have been discussed in Section 2.3 (under RD systems), additional aspects are briefly noted here.

**Nerve models.** Through protein pores, active membranes of nerves are permeable to positive ions (mainly sodium, potassium, and calcium), leading to negative conductive processes much like those of the Esaki, Giaever, and Josephson junctions that are used for a variety of distributed electronic devices [241, 244, 453]. In nerve fibres, active membranes form roughly cylindrical tubes (see Figure 13), for which the transmembrane voltage is governed by Equation (30) – a general RD equation. Several nerve-fibre models are now available to nonlinear analysts, including these [466]. (i) The Hodgkin–Huxley (HH) equations, which were developed to give a good description of nerve impulse propagation on the squid giant axon shown in Figure 14 [244, 245]. (ii) The Zeldovich–Frank-Kamenetsky (ZF) equation (discussed in Section 2.3), which was first derived for flame-front propagation, can be exactly solved and well describes the leading edge of an HH impulse [375]. (iii) The Markin–Chizmadzhev (MC) model, in which the transmembrane ionic current is assumed to be a fixed time function representing the initial inflow of sodium current and subsequent outflow of potassium current [246, 466]. As the simplest representation of a complete nerve impulse, MC may be convenient for studies involving complicated membrane geometry. (iv) The FitzHugh–Nagumo (FN) equation, in which the ZF equation is augmented by a single recovery variable, leading to a nonlinear PDE that conducts a nerve impulse [466, 534]. This is the simplest model in which the impulse shape is generated by the internal dynamics of the PDE. (v) The Morris–Lecar (ML) model, which is physiologically realistic and represents inflow of calcium (rather than sodium) current
Myelinated nerve models, which represent the discrete nature of mammalian motor nerves [44, 466]. These are difference-differential equations (DDEs) which describe “saltatory” (hopping) conduction from one active node to the next.

Nerve impulses are emergent entities that propagate away from the neuron body through the branchings of axonal trees or toward it through branchings of dendritic trees [466]. In both cases, impulse propagation is influenced by fibre branchings, the effects of which must be understood for a realistic description of neuron dynamics. Propagation of impulses through fibre branchings was studied by Boris Khodorov and his colleagues in the early 1970s, who showed that dendritic branchings have the ability to act as digital logic elements [266]. At about the same time, Jerry Lettvin and his colleagues demonstrated blocking in the optic nerves (axons) of cats [96], and Steve Waxman proposed the concept of a “multiplex neuron” in which the dendritic trees process information carried by incoming impulse trains [466, 546]. Although the idea of information processing in the dendritic or axonal trees of a neuron was not taken seriously by the neuroscience community during the 1970s, it is now becoming accepted as a feature of neuron dynamics [466, 494].

Figure 32: Switching action in a branching region of a squid giant axon which is excited by a double pulse [466]. (a) Preparation geometry (not to scale). (b) Blocking of the second impulse. (c) Passage of the second impulse.
As an example of the switching phenomena that can be observed for impulse propagation through branching regions of nerves, consider Figure 32 which shows measurements of double impulse propagation on a branching squid fibre that were made by the present author at the *Stazione Zoologica* of Naples [466]. The impulse interval shown in the figure is just at the threshold above which (slightly longer interval) the second impulse makes it through the branch, and below which it doesn’t. These data are in accord with the numerical results of Khodorov and his colleagues [266], suggesting that an interesting area of research would use current concepts in nonlinear science and modern computing facilities in collaboration with electrophysiologists to further explore blocking phenomena near realistic dendritic and axonal geometries.

**Brain models.** Although it presently seems feasible to construct realistic nonlinear models of individual neurons, a corresponding formulation for the entire brain is a more formidable task. Interestingly, one of the first brain models – and still one of the best – was proposed in 1949 by Donald Hebb, by a Canadian psychologist, in his book *Organization of Behavior* [225]. This classic work offered the first credible description of how a mass of neurons with seemingly tangled interconnections can exhibit the wide variety of mental phenomena that we all experience.

Consider, for example, the Neckar cube shown in Figure 33, noting that first one face is perceived in the front, then another – switching back-and-forth every few seconds. What is happening in the brain to produce this experience? In Hebb’s view, the brain’s neurons do not act individually but as coordinated subsets of the entire population; thus [225]

> Any frequently repeated, particular stimulation will lead to the slow development of a “cell-assembly,” a diffuse structure comprising cells . . . capable of acting briefly as a closed system, delivering facilitation to other such systems and usually having a specific motor facilitation. A series of such events constitutes a “phase sequence” – the thought process. Each assembly may be aroused by a preceding

---

43This switching of perceptions of a Neckar cube is an example of the “Gestalt-like behavior” that Thomas Kuhn used as a metaphor for scientific revolutions.
assembly, by a sensory event, or – normally – by both. The central facilitation from one of these activities on the next is the prototype of “attention.”

Hebb did not include inhibition between neurons in his original formulation of the cell-assembly theory, as this had not yet been discovered by electrophysiologists. When inhibition was empirically established in the mid-1950s, it was introduced by Hebb’s colleague, Peter Milner, as a “Mark II” version of the cell assembly theory [361].

To see how the Mark II theory can be applied to ambiguous perceptions of the Neckar cube, suppose that the two perceptions are represented by two cell assemblies, composed of an approximately equal number of neurons. (Under the theory, these neurons are not localized but spread out over the optical lobes of the neocortex, like a “three-dimensional fish net” to use Hebb’s metaphor.) If \( F_1 \) and \( F_2 \) are the fractions of neurons that are firing in these two interacting assemblies, their dynamics can be modeled by the nonlinear ODE system [466]

\[
\frac{dF_1}{dt} = F_1 (1 - F_1) - \alpha F_2,
\]

\[
\frac{dF_2}{dt} = F_2 (1 - F_2) - \alpha F_1,
\]

where \( 0 \leq F_1 \leq 1 \) and \( 0 \leq F_2 \leq 1 \). These equations follow both from Hermann Haken’s concept of an order parameter (derived from condensed matter physics) [214] and also from a statistical analysis of the constituent neurons [466].

Now consider the following cases [466]. (i) With \( \alpha = 0 \), the two assemblies are uncoupled, and the firing fractions of each assembly obeys a Verhulst (logistic) equation, as in Equation (2). Thus both assemblies will eventually fire

\[ F_1(t) = \frac{1}{1 + e^{-t}}, \]

\[ F_2(t) = \frac{1}{1 + e^{-t}}, \]

Figure 34: (a) A phase-plane plot from Equations (55) with \( \alpha < 0 \) (only excitatory interactions). (b) A phase-plane plot for \( \alpha > 1/3 \) (excitatory and inhibitory interactions among neurons).
at their full rates \((F_1 = F_2 = 1)\). Evidently the model does not include a means to turn off an assembly once it starts firing. (ii) With \(\alpha < 0\), there are only excitatory interactions between the two assemblies because the \(-\alpha F_2\) term in the first equation increases \(dF_1/dt\) and similarly for the second equation. In this case, the phase-plane plot of Figure 34(a) shows that both assemblies will again eventually fire at their full rates \((F_1 = F_2 = 1)\). (iii) For \(0 < \alpha < 1/3\), there is a stable stationary state at \(F_1 = F_2 = 1 - \alpha\). (iv) For \(\alpha > 1/3\), there is an unstable stationary state at \(F_1 = F_2 = 1 - \alpha\); thus the phase-plane plot appears as in Figure 34(b), which corresponds to the switching action that we perceive when staring at the Neckar cube in Figure 33.

Interestingly, the characteristic time of the switching shown in Figure 34(b) is of order \(1/(3\alpha - 1)\), which decreases with the inhibition between assemblies. This observation is in accord with a suggestion of Hebb that we humans are more intelligent than our mammalian cousins in part because we have more inhibitory interneurons in our neocortices, allowing us to switch our attention more rapidly from one complex assembly to another [226, 227].

This analysis shows that interacting cell assemblies exhibit switching behaviour, as does an individual neuron, which is in accord with a significant amount of psychological data [226, 227]. On the modeling side, Anders Lansner, Erik Fransen and their colleagues have used currently available computers to conduct extensive numerical studies of interacting neurons with physiologically reasonable parameters. This numerical work clearly shows the “slow development” of cell-assemblies which are “capable of acting briefly as a closed system” as Hebb originally proposed a half century earlier [140, 178, 179, 181, 294].

Other formulations of collective neuronal behaviour include the following. (i) The wave theory introduced by Hugh Wilson and Jack Cowan in the early 1970s, which can be viewed as a reaction-diffusion system involving excitatory and inhibitory neurons as subsets of the cortical mass [558, 559, 557, 466]. (ii) John Hopfield’s attractor network theory which was introduced in the early 1980s. Hopfield’s approach employs a Lyapunov functional and is based on phase changes in the Ising model from condensed matter physics [249, 250]. In the limit of high average firing rate, Hopfield’s model is equivalent to Hebb’s theory [466].

Finally, it should be mentioned that there has been a regrettable tendency among some neuroscientists and biologically oriented physicists either to ignore Hebb’s classic book or to mention it only in the context of the so-called “Hebbian learning rule” which assumes that neurons will become more closely interconnected after they have experienced synchronous firings. Thus Scientific American published an entire issue on brain theories in 1992 without mentioning Hebb’s work. Upon protesting this neglect, Milner was allowed a compensatory article in a later issue [362] in which he mentioned that Hebb was amused at the emphasis on the learning rule. In fact, the learning rule that

\[\text{106}\]
Hebb used was the only component of Hebb’s theory that was *not* original, as it had been previously suggested by Sigmund Freud, among others.

**THE COGNITIVE HIERARCHY.** Upon examining our information processing systems, we find the following hierarchical structure.

```
  Human culture
     Phase sequences
     Complex assemblies
     ...

Assemblies of assemblies of assemblies
  Assemblies of assemblies
  Assemblies of neurons
    Neurons
    Nerve impulses
    Nerve membranes
    Membrane proteins
    Molecules
```

As in the biological hierarchy, the dynamics of each cognitive level are non-linear, out of which emerge entities that provide the basis for the next higher level. Again, there is an immense number of possible entities that can emerge at each level, from which a much small number are realized. In other words, the number of ideas that I *actually* have is much smaller than the immense number of ideas that I *might* have.

One difference between the biological and cognitive hierarchies is that every level of the former has been confirmed by direct inspection, whereas the cognitive levels

```
  Complex assemblies
  ...

Assemblies of assemblies of assemblies
  Assemblies of assemblies
  Assemblies of neurons
```

are theoretical constructs, based on Hebb’s cell assembly theory. Interestingly, Charles Legéndy has estimated the number of Hebbian assemblies \( C \) that can be stored in a brain of \( N \) (from a simple model that assumes that an assembly is composed of \( y \) subassemblies each comprising \( n \) neurons) as \( C = (N/n^y)^2 \) [299, 466].

Direct experimental observation of a cell assembly is a daunting task, as the constituent neurons are not localized but spread out over the brain. Although it is challenging – to say the least – to obtain recordings from a significant fraction of the neurons in one of Hebb’s three-dimensional fishnets, some progress is being made. On simpler species, it is now possible to introduce voltage-sensitive dyes into the neurons and record the dynamics of their global activity.
on arrays of silicon photodiodes. In this manner, the simultaneous activity of several hundred neurons in the abdominal ganglion of a mollusk (*Aplysia*) has been indirectly observed [566]. As an example of direct voltage measurements, neuroscientists at the University of Arizona are recording from several dozen extracellular electrodes in the hippocampus of the rat, with each electrode indicating the activity of several neurons for a total of 100 or more individual signals [560]. Interestingly, these multiple recordings can be made while the animal is undergoing psychological testing, and several other laboratories are reporting equally impressive feats [118, 348, 387]. Among the more sophisticated of such devices is the Utah electrode array (UEA), which comprises 100 external electrodes mounted in an area of one square millimetre [340, 395].

At the psychological level, Richard Taylor and his colleagues have recently considered relations between fractal patterns and those appearing in the well known paintings by Jackson Pollock [508, 510]. This interesting study argues that when Pollock moved from Manhattan’s rectilinear environment to rural Long Island in 1945, his art underwent a radical change, reflecting the influence of his spatially chaotic natural surroundings.

### 3.12 General perspectives

Although nonlinear science stems largely from nineteenth-century efforts to understand planetary motions, fluid dynamics and population growth, there were related currents in the mid-twentieth century that recognized the importance of emergent phenomena (self-organization) in biology, psychology and the social sciences. From our present perspective, it is easy to see connections among these activities, but a half-century ago they were largely balkanized and ignored by the mainstream scientific community with its predominantly reductive outlook.

**General systems theory and cybernetics.** During the 1937–1938 academic year, Ludwig von Bertalanffy – a biologist from the University of Vienna who was visiting University of Chicago biophysicist Nicolas Rashevsky on a Rockefeller Fellowship – gave one of the first North American lectures on general systems theory (GST) [537, 538]. With a long-time interest in the growth of biological form, he viewed GST as having broad potential for understanding the nature of self-organizing phenomena in the social sciences and also in the physical sciences.

In 1946 with support from the Josiah Macy, Jr. Foundation, the first of a series of yearly conferences was held under the chairmanship of psychiatrist Warren McCulloch and entitled “Feedback mechanisms and circular causal systems in biology and the social sciences.” The list of participants included anthropologist Gregory Bateson, learning theorist Lawrence Frank, psycho-analyst Lawrence Kubie, sociologist Paul Lazarsfeld, social psychologist Kurt Lewin, neurophysiologist Rafael Lorente de Nó, anthropologist Margaret Mead, physicist John von Neumann, neurophysiologist Arturo Rosenblueth, mathematician Norbert Wiener, and Walter Pitts – a brilliant young mathematician
who had recently written a seminal paper on brain theory with McCulloch [347]. This list was impressive not only for breadth and intellectual power, but because these scientists were deeply suspicious of the doctrines of logical positivism and behaviorist psychology – then at the crest of their popularity.

During the first Macy meeting, Wiener was asked to apply his mathematical ideas on feedback and control theory in a broader context, and in 1948 he published *Cybernetics: or Control and Communication in the Animal and the Machine*, which sold 21,000 copies in three months [552]. In this book, Wiener emphasized the control aspects of negative feedback, tending to view positive feedback as a source of malfunction (runaway oscillations, tremors, etc.), whereas von Bertalanffy was more concerned with self-organization and emergent phenomena which stem from the presence of positive feedback [539]. Thus both perspectives contribute in complementary ways to the new interdisciplinary field of systems theory and cybernetics.

In the autumn of 1960, Wiener visited the University of Naples where he decisively influenced the future activities of Eduardo Caianiello, who became the Founding Director of the Laboratorio di Cibernetica del CNR. Opening its doors in 1968, the Laboratory provided young researchers in philosophy, mathematics, physics, engineering, chemistry, and biology opportunities to establish professional friendships across disciplinary boundaries and organize joint research activities in a wide variety of settings. From the perspectives on nonlinear science, early results of this activity included a broad study of the sine-Gordon equation and the birth of superconductive (Josephson) device research in Italy [30, 27]. Although these researchers spent a substantial part of their creative time establishing credentials in their respective fields, this interdisciplinary environment brought many physical scientists into serious contact with the essential problems of biology and the social sciences.45

Following the visions of von Bertalanffy and Wiener, such studies are now widely accepted as an important aspect of nonlinear science. Useful surveys of the development and current status of interdisciplinary studies in complexity theory and self organization can be found in two excellent books by Klaus Mainzer [321, 322].

**NON-EQUILIBRIUM STATISTICAL DYNAMICS AND SYNERGETICS.** Behind the post-war enthusiasm for GST and cybernetics, there loomed a speculative theoretical basis, which encouraged those who would dismiss concepts of self-organization as idle dreams. Motivated by a desire to comprehend the nature of time (particularly the “arrow of time”) and by the above-mentioned ideas of Alan Turing [520], chemist Ilya Prigogine undertook studies of nonequilibrium thermodynamics in the 1960s, which led to fundamental results in the foundations of statistical mechanics and for dissipative structures in nonlinear chemical systems [423, 392].46 In addition to providing theoretical sup-

---

45I was privileged to spend the 1969–70 academic year and every summer of the 1970s at the Laboratorio, an opportunity that led to my electrophysiology research at the Stazione Zoologica shown in Figure 32.

46Prigogine was awarded the 1977 Nobel Prize in chemistry for this work.
port for the empirical observations of Boris Belousov [137, 583], Prigogine’s research has contributed to a wide range of nonlinear applications, including fluid mechanics, biology, finance, and sociology [391]; thus underpinning speculations offered at the Macy conferences.

Following his fundamental studies of laser dynamics [210, 211], physicist Hermann Haken developed a “top-down” approach to theories of self-organization, which complements the “bottom-up” theory of Prigogine [215]. Called synergetics to imply the “science of cooperation”, Haken’s formulation has two components [213]. Under microscopic synergetics, first, it is assumed that a system goes unstable at a critical value of a parameter, leading to unstable modes and to stable modes which are enslaved to the unstable modes. The nonlinear dynamics are then simplified by focusing attention on the unstable modes, allowing definitions of order parameters and including the slaves as second-order effects in subsequent calculations. When such a microscopic analysis is not possible, second, the same general structure is assumed, including the presence of an order parameter which becomes the main object of the analysis. Haken’s approach is useful in many areas of research, including physics, chemistry, biology, psychology and brain theory [214] (clearly explaining the Gestalt phenomenon of Figure 33), sociology, management, and dynamics of city growth – in addition to Thomas Kuhn’s theory of scientific revolutions [279].

4 Conclusions

One aim of this review is to show that the explosion of research activity in nonlinear science that is described in Section 1 was indeed a Kuhnian revolution. In reaching this conclusion, we can proceed in two ways: by comparing events of the 1970s with other revolutionary changes described by Kuhn and by considering the qualitative features that characterize such revolutions.

As Kuhn showed [278], the shift from Ptolemaic (geocentric) to Copernican (heliocentric) cosmology in the seventeenth century provides a clear example of a scientific revolution, leading not only to striking changes in collective attitudes about the structure of the universe and the nature of motion, but having profound implications for European religious perspectives. An equally important revolution was Darwin’s theory of biological evolution by natural selection, which altered lexicons in biology and paleontology and also had deep religious implications [109, 70]. Although the explosive growth of interest in nonlinear science during the 1970s was not as important as either of these two classic examples, it compares favorably with other revolutionary shifts proposed by Kuhn, including the following [279, 280]. (i) Recognizing oxygen as an elementary gas and the subsequent demise of the phlogiston theory, (ii) Discovery that the electricity in a Leyden jar is stored in the glass between metal foils rather than in the bottle itself. (iii) Seeing that the potential of a battery is developed between the plates rather than across an inter-metallic interface. (iv) The discovery and acceptance of X-rays. (v) Replacement of Newton’s gravita-
tional theory with the general relativity theory (GRT) at the level of planetary motions and above. (vi) Replacement of Newtonian mechanics with quantum theory (QT) at the level of atomic motions and below. According to Kuhn, scientific revolutions are characterized by three properties.

- They are holistic in the sense that they cannot be arrived at incrementally; thus they entail new perspectives – new ways of looking at the world.
- They involve meaning changes in the ways that words and phrases are used.
- They introduce new metaphors or models for thinking about reality.

Recent developments in nonlinear science satisfy these three criteria. Recall first Russell’s sudden recognition of the key properties of his solitary wave on the “happiest day” of his life [448] and Lorenz’s understanding the implications of his irregular orbits [312]. More generally, the holistic aspect of nonlinear science is displayed in Figure 18, where the various relations among low-dimensional chaos and the emergence of new dynamical structures (both dissipative and energy-conserving) are sketched. Second, nonlinear localization of dynamic activity is now widely recognized, and formerly unsatisfactory “irregular” or “stochastic” orbits are presently applauded for being “chaotic”, augmenting the meanings of an ancient word and enlarging our attitudes about the nature of causality. Finally, new metaphors and models that are now widely used throughout science include: Arnol’d diffusion, autowaves, Bäcklund transformations, breathers, butterfly effect, chaos, collective coordinates, fractals, frustration, inverse-scattering methods, Lyapunov exponents, Panlevé analysis, reaction-diffusion systems, scroll waves, self-organization, solitons, strange attractors, and universality.

In other words, the phenomena related to low-dimensional chaos and the emergence particle-like solutions of PDEs are now recognized as important aspects of physical reality, with many practical applications in addition to their implications for philosophical reevaluations of reductionism and the ontological status of emergence [467, 468]. Fully self-consistent solutions of nonlinear problems are now routinely sought (analytically or numerically) rather than relying on the former concept of quasilinear interactions among linearized modes. Even more important, nonlinearity is now viewed as a positive aspect of dynamics leading to useful behaviour rather than glumly accepted as an unwelcome feature that impedes the search for analytic solutions. Unpredictability is recognized as a generic property of low-dimensional nonlinear systems, and spatial chaos is seen as a typical aspect of nature rather than the denizen of a mathematical zoo. “How long is the coast of Britain?” has become an interesting question.

In its struggle with astrology, Newtonian science reduced Aristotle’s four meanings of the word “cause” to one – an “efficient cause” [73]. Under nonlinear science, interestingly, “material”, “formal” and “final” causes are welcomed back into in the collective lexicon, especially in formulations involving downward causation [12, 156]. On the other hand, Aristotle’s dictum that
“nothing can be its own cause” – a benchmark of mid-twentieth-century science – is rejected by nonlinear scientists, who typically study phenomena that emerge from positive feedback loops and networks [468]. Thus instabilities arising under slow parameter changes can lead to the exponential growth of small perturbations which evolve into qualitatively new phenomena. Important examples of emergence include the beginning of life from the hot chemical soup of the Hadean seas some three and a half billion years ago, the emergence of new biological species, the growth of a city, and the birth of an idea in a human brain.

Using Kuhn’s criteria, therefore, it appears that the explosive growth of research in nonlinear science was indeed a scientific revolution, having punctuated the progress of normal scientific activity in a Gestalt-like manner during the 1970s and leading – as we have seen – to new perspectives in many branches of pure and applied science. Amusingly, Kuhn’s formulation of the development of a revolutionary scientific perspective describes the explosive growth of interest in nonlinear dynamics, while this growth was itself an example of a nonlinear dynamic process. Where will the nonlinear science revolution lead?

Among the many implications of nonlinear science for future research, the most important will be in the realms of biology, psychology and social science (BPS). The reasons for making this claim are threefold. First, Low-dimensional (deterministic) chaos is now an accepted fact of dynamics, arising in a variety of contexts from meteorological models to dripping faucets and making many predictions problematic. Second, empirical studies in BPS differ fundamentally from those of physics and chemistry (PC). In PC laboratories, it is possible to perform as many experiments as needed on homogeneous sets of identical preparations (hydrogen atoms, benzene molecules, electrons, etc.), whereas BPS studies necessarily deal with heterogeneous sets in which the number of elements examined is far exceeded by the immense number of possible elements (all possible protein molecules, biochemical cycles, living creatures, human personalities, languages, social organizations, poems, etc.). Although one can study a particular human being in detail, it is difficult to generalize to all possible members of the class; thus observations of regularity in BPS do not have the same epistemological status as the laws of PC. Finally, living organisms, human brains and social systems are hierarchically organized, with the effects of downward causation acting through Aristotle’s final, formal and material causes, leading to myriad positive feedback networks [467, 468]. From networks that span several levels of these hierarchies emerge phenomena that are qualitatively different from those of physical science – a point that is yet to be fully recognized by those who study the natures of life and mind [466, 468].

Appreciating the hierarchical structure of organisms [13, 20, 146, 393, 541] and the ontological status of the higher-level emergent entities that comprise them will be central issues for future research in biological, cognitive and social sciences [471, 532]. Rather than inducing despair, nonlinear science encourages young scientists with the many new vistas that are now open for research. Based on these perspectives, the nonlinear science revolution of the 1970s will
lead us to a deeper understanding of the complicated chemical, physiological, mental and social reality in which we all live.

5 Acknowledgments

My understanding of nonlinear science has emerged from interactions over the past five decades with teachers, students, and colleagues at MIT, the University of Wisconsin, the Laboratorio di Cibernetica, the Los Alamos National Laboratory, the University of Arizona, and the Technical University of Denmark. Support during these years from the Bausch and Lomb Corporation (USA), the National Science Foundation (USA), the National Institutes of Health (USA), the Consiglio Nazionale delle Ricerche (Italy), the European Molecular Biology Organization, the Department of Energy (USA), the Technical Research Council (Denmark), the Natural Science Research Council (Denmark), the Thomas B. Thriges Foundation (Denmark), the North Atlantic Treaty Organization, and the Fetzer Foundation (USA) is gratefully acknowledged. While accepting responsibility for residual errors, I thank all who have made helpful suggestions, including Ping Ao, Henry Arbarbanel, Rowena Ball, Howard Barnes, Antonio Barone, Luc Berge, Vadim Biktashev, Friedrich Busse, Jean Guy Caputo, Giulio Casati, Peter Leth Christiansen, Leonor Cruzeiro, Florin Diacu, Chris Eilbeck, Claus Emmeche, Neville Fletcher, Beall Fowler, Leon Glass, John Guckenheimer, Hermann Haken, Peter Hamm, Arun Holden, Jim Howard, Sharon Johnson, Muthusamy Lakshmanan, Allan Lichtenberg, Karl Lonngren, Klaus Mainzer, Jean-Jacques Malandain, Boris Malomed, Jens Marklof, Jerry Marsden, James Meiss, Robert Miura, Jerry Moloney, Oleg Mornev, Nancey Murphy, Grégoire Nicolis, Alfred Osborne, Lev Ostrovsky, Serge Patlavskiy, Niels Falsig Pedersen, Michel Peyrard, John Pojman, Mikhail Rabinovich, Mike Russell, Mario Salerno, Reinold Schmidt-Koppenhagen, Leon Shohet, Henry Sobell, Steven Strogatz, Paul Sutcliffe, Richard Taylor, Michael Thompson, Arkady Tsinober, Bruce West, Ludmilla Yaksheitchev, Vyacheslav Yukalov, and Norman Zabusky. Finally, I am indebted to the entry authors, advisors and referees who contributed to the Encyclopedia of Nonlinear Science for many insights.

References


[84] CARTWRIGHT, M.L. AND LITTLEWOOD, J.E.: On nonlinear differential equations of the second order. I. The equation $\frac{dy}{dx} + k(1 - y^2) \frac{dy}{dx} + y = b \lambda k \cos(\lambda t + \alpha)$, k large, J. London Math. Soc., 20 (1945) 180–189.


126


[372] MOORE, W.: *Schrödinger: Life and Thought*, (Cambridge University Press, Cam-
bridge) 1989.


[379] NAGUMO, J., ARIMOTO, S. AND YOSHIZAWA, S.: An active impulse transmis-


