IOP Publishing

EPS Grand Challenges

Physics for Society in the Horizon 2050

Mairi Sakellariadou, Claudia-Elisabeth Wulz, Kees van Der Beek, Felix Ritort, Bart van Tiggelen, Ralph Assmann, Giulio Cerullo, Luisa Cifarelli, Carlos Hidalgo, Felicia Barbato, Christian Beck, Christophe Rossel and Luc van Dyck

Chapter 7

Physics for secure and efficient societies

Felicia Barbato, Marc Barthelemy, Christian Beck, Jacob D Biamonte, J Ignacio Cirac, Daniel Malz, Antigone Marino, Zeki Can Seskir and Javier Ventura-Traveset

7.1 Introduction

Christian Beck[1] and Felicia Barbato[2]
[1]Queen Mary University of London, London, UK
[2]Gran Sasso Science Institute, L'Aquila, Italy

As we have seen in this book, physics research contains two very different aspects. There is the fundamental research driven by curiosity, with the ultimate aim of understanding very small interacting systems, very large interacting systems, and their complex behaviour on intermediate scales. There is also the applied side, where physics is applied to develop new technologies, new analysis methods, and new concepts and insights that are useful for society. Ultimately, much of what physics for interacting systems encounters is nonlinear, high-dimensional, and complex, and the final goal is to apply novel physical insights to real-world systems, providing useful applications that are helpful for society in general. The topic of this chapter, physics for secure and efficient societies, is very broad and has many different aspects, and this chapter in no way makes an attempt to treat it in full. Rather, we have selected a few topics that we find particularly interesting, with the emphasis of looking into the future—perhaps looking towards the year 2050 or towards a similar time scale.

One important aspect for the future of society is the further development of information technologies; proper communication and information processing and enhanced computer development are absolutely essential. Our world today is dominated by computers in their various shapes and sizes, from small to big, from personal to institutional, from local to worldwide. Science has made immense

doi:10.1088/978-0-7503-6342-6ch7

progress by implementing modern machine learning technologies and artificial intelligence, so there are some natural questions: Where is computing going? What is the next generation of computers made of? And what is the next generation of algorithms? Still in its infancy today, quantum computing may hold the key for outstanding novel computational developments of the future. Some problems are so complex that they cannot be solved with present conventional computers but require something that is orders of magnitudes faster, or they need algorithms and novel approaches that are very different from what is currently used in mainstream simulations.

The section of this chapter by Daniel Malz and J Ignacio Cirac summarizes the most important principles of quantum computing, exploiting quantum superpositions, and entanglement for the purpose of future quantum computers. The aim is to solve certain problems much faster than is possible with conventional computers. The section by Zeki Can Seskir and Jacob Biamonte looks at the historical development of quantum computing research and in particular quantum algorithms and new types of machine learning models, which are expected to be very relevant in the future.

How do we actually get the data that we feed into our physical models to make accurate predictions for the future, using the best computers and analytical techniques available? The problem is nontrivial, as bad data yield biased and unprecise predictions. Sensor technology has made immense progress recently. The convergence of multiple technologies, real-time analytics, machine learning, ubiquitous computing, and embedded systems gave birth to the Internet of Things (IoT), and the automation and control of industrial processes can be seen as the fourth industrial revolution, also known as the Industrial Internet of Things (IIoT). In her section, Antigone Marino describes the historical basic steps of sensor developments, arriving then at the future challenges set by Europe's climate change strategy for 2050. In this framework, smart sensors are fundamental for monitoring all services related to the automation of processes as regards to waste reduction, clean water, and environmental control and to improve the quality of life in the workplace.

The use of smart sensors is also open to the space sector, which is gaining more and more importance and is going to enter its golden age driven by the longstanding dream of humankind, the exploration of space, with many interesting new perspectives and applications for the benefit of humans on the horizon. In his section on the space sector, Javier Ventura-Traveset reviews the current status of next future missions and explores the prospects of the space sector beyond 2030–35. Many intriguing topics are covered, starting with more gnoseological problems such as space science, going through futuristic scenarios about human and robotic exploration of space, and finally touching on more practical issues such as understanding the climate change trend, its sources, its dynamics, and the major anthropogenic impacts. Javier Ventura-Traveset makes it very clear how space exploration has had, and will have even more in the future, huge impacts on our society from both economic and social points of view and how future society can benefit from this emerging sector.

Finally, another problem of utmost relevance for future societies is understanding the complexity that underlies the real-world systems that surround us and the daily aspects of our lives. Here, statistical physics, in its modern form, has a lot to say. One particular example is the science of cities. A very large part of the world population these days live in cities, but how do cities actually function, how do they evolve, and how can we improve their day-to-day structures and life quality in a sustainable and environmentally friendly way? Cities are spatially extended complex systems, and statistical physics, in its modern formulation, can be applied. In the historical Boltzmann formulation, particles are replaced by agents (companies, vehicles, people, sustainable energy sources, etc), interactions are replaced by communications (mobile phones, e-mail, Twitter, etc), phase transitions correspond to an abrupt change of relevant observables (opinions, prices, behavioural patterns, etc), and so on. In his section, Marc Barthelemy provides a state-of-the-art overview of city modeling, city growth aspects, traffic congestion, and much more, using the tools of statistical physics and complex network theory.

Overall, the example topics treated in this chapter show that often there is initially fundamental basic physical science, which then feeds into more advanced applied models relevant for future development. For example, starting from quantum physics, we proceed to modern methods and algorithms of quantum computing; starting from classical equilibrium and nonequilbrium statistical physics, we proceed to a modern science of cities; and so on. Better predictions and better models can be made if we have access to better data obtained with more powerful sensors by better satellite navigation methods, and so on. Let's hope that in 2050, when a reader looks back into this book, most of the world's population will be living in a clean, peaceful, and sustainable environment where physics helped a lot to attain this stable state.

7.2 Second quantum revolution: quantum computing and cybersecurity

7.2.1 The second quantum revolution: quantum computing and information

Daniel Malz[1] and J Ignacio Cirac[2]
[1]Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
[2]Max Planck Institute of Quantum Optics, Garching, Germany

Even though quantum physics is now more than a hundred years old, it still offers new challenges, insights, and applications. One particularly exciting possibility is to use quantum physics to process and transmit information in more efficient and secure ways. In this section we briefly summarize the principles of quantum computing (QC), which exploits quantum superpositions and entanglement to solve certain particular problems much faster than any classical device. We also review some algorithms and applications and the experimental state of the art, and we conclude with an outlook towards more powerful devices and their potential impact.

7.2.1.1 Introduction

Quantum mechanics explains all microscopic phenomena around us. As well as being of fundamental interest, it is crucial to understand materials, chemistry, and molecular biology and thus much of modern science and technology. In the past decades, it has become increasingly clear that the properties of quantum mechanics, such as superposition and entanglement, allow one to process information in a fundamentally different way, which may revolutionize communication and computation. This has led to the emergence of a new multidisciplinary field, quantum information science, which combines mathematics, computer science, and quantum physics. The ambitious key objective of this field is the construction of a quantum computer—a device that manipulates information according to the laws of quantum mechanics. Originally proposed in a seminal talk by Richard Feynman [1], quantum computers are now known to be capable of solving problems that that are too complex for classical computers. There are significant challenges that need to be overcome on the way to a fully functional quantum computer, but the journey there will provide scientists with novel insights and change our understanding of Nature.

In the second half of the 20th century, the computer revolutionized virtually all areas of society and industry. This development was enabled by electrical engineering, based on the modern understanding of classical electrodynamics and classical information theory, but also other concepts building on quantum mechanics, such as the physics of semiconductors and transistors. It is tempting to draw an analogy to the current development. Scientists around the world are making great strides towards the construction of a quantum computer, based on other aspects of quantum mechanics and quantum information theory, enabled by a technological leap in atomic physics, optics, and solid-state materials. While the potential of quantum computers is vast, there are still relatively few good algorithms available

today, and building one appears very challenging. As a result, it is very difficult to gauge the impact that quantum computers will have on society, even in the short run. Here we will nevertheless try to paint a rough picture of what awaits us.

A common misconception is that quantum computers are superior to conventional computers and that it is only a matter of time until your smartphone's processor will run quantum algorithms. In fact, tasks that classical computers are good at will very likely always run faster on classical processing devices. Instead, there are many important problems for which the best classical algorithms require exponential time in the input size, which means that adding merely one bit to the input causes the runtime of the algorithm to be multiplied by a constant factor greater than 1. Such problems quickly become infeasible to solve, even for moderate input sizes. There exist very important problems with this property, including the simulation of quantum systems such as molecules and materials, as well as optimization problems, for example, to schedule trains, route traffic, and distribute resources optimally. Scientists have invented quantum algorithms that solve a subset of these problems in polynomial time and thus remain feasible for large inputs. Which problems belong to which complexity class is subject to change as new algorithms are invented, but very likely there will always remain some that can be efficiently solved only with a quantum computer. At the risk of oversimplification, this means that there are important problems that cannot be solved at all on classical computers but become accessible with a quantum computer. For the interested reader, these fascinating concepts are explained in greater detail in the excellent texts by Nielsen and Chuang [2] and Aaronson [3].

We are now at a very exciting point in the development of quantum computers and quantum technologies (QT) in general. During the last decades, leading technology companies have announced programmes to develop quantum computers. This has boosted the pace of engineering efforts significantly. In 2019 the first experiment was performed in which a quantum processor completed a task that, according to our current knowledge, would take many years to complete on even the best supercomputers [4]. Since then, these results have been reproduced with more qubits and on other platforms [5, 6].

Naturally, these developments have attracted the attention of the media, politicians, and society as a whole. One should be careful, however, to distinguish the current generation of quantum processors from fully fault-tolerant and scalable quantum computers in order not to raise unrealistic expectations that can only be disappointed. Quantum systems are very susceptible to noise, which causes errors to accumulate during a computation. At the same time, these systems require an exceptional level of control, which makes them difficult to scale to large sizes. As a result, current devices, also referred to as noisy intermediate-scale quantum (NISQ) devices [7], can run only short computations before their output becomes completely scrambled. The previously mentioned algorithms with exponential speedup require large computers that are capable of correcting errors that arise during the computation, which places very stringent requirements on their properties. This is challenging to achieve and may take decades. Yet, as is often argued, there are important applications of NISQ devices, such as analog simulation, and even before

fault tolerance becomes a reality, there is much to be learned and gained from this research program as a whole [7, 8].

In this section, we give a very brief and basic introduction to quantum information theory, review some of the algorithms and applications of quantum computers, and discuss the most advanced platforms in which they are being implemented. We also give an outlook, trying to balance the substantial and very real promise that quantum computers hold with the equally real and formidable challenges that have to be met on the way.

As this field is changing continuously and new algorithms, platforms, and protocols, appear almost every day, we provide just few basic references here so that the interested reader has access to more detailed material. We have often given references not to original papers, but rather to reviews where those can be easily found. For a comprehensive, textbook-style account of the field, we point the reader to the book by Nielsen and Chuang [2], which is perhaps the most important reference text for the field. Another deeply insightful text in a somewhat lighter style is by Aaronson [3]. For technical reviews, which contain many further references, we point to those on quantum algorithms [9], quantum algorithms for chemistry [10], quantum machine learning [11], and specific platforms [12–14].

7.2.1.2 Quantum information theory basics

We briefly introduce the principles of quantum information theory, in which information is stored on quantum bits (qubits), computation corresponds to unitary operations, and the computational advantage relies on quantum superposition and entanglement.

7.2.1.2.1 Qubits

In quantum mechanics, energy is quantized into bits called excitations. As applied to light, which is quantized in terms of photons, this means that a system may contain zero, one, or ten of them, but not half a photon. The state of the system can thus be expressed in terms of those discrete states. Restricting our attention to just two states, which we label $|0\rangle$, $|1\rangle$, we can define a quantum bit (a qubit): a quantum system with two states. Note that the states do not have to be distinguished by photon number; they might correspond to any other degree of freedom as well, such as, for example, spin or collective excitations in a circuit. In a classical computer, this is the end of the story: Computation is done by manipulating many bits, each of which can assume either state 0 or 1 (see figure 7.1). Quantum mechanics, however, permits superpositions. The general state of a qubit is described by the following vector:

$$|\Psi\rangle = c_0\,|0\rangle + c_1\,|1\rangle, \tag{7.1}$$

where c_0 and c_1 are complex numbers that obey $|c_0|^2 + |c_1|^2 = 0$ and $|0\rangle$ and $|1\rangle$ form the basis of the vector space. It is important to note that a qubit really can be a superposition in the sense that it is distributed over both basis states. This should be contrasted with the state of a coin after it has been flipped but before the result has been recorded. The coin may be either in 0 or in 1; the qubit is in both. The reason for this subtle difference relies on the basic principles of quantum physics. For

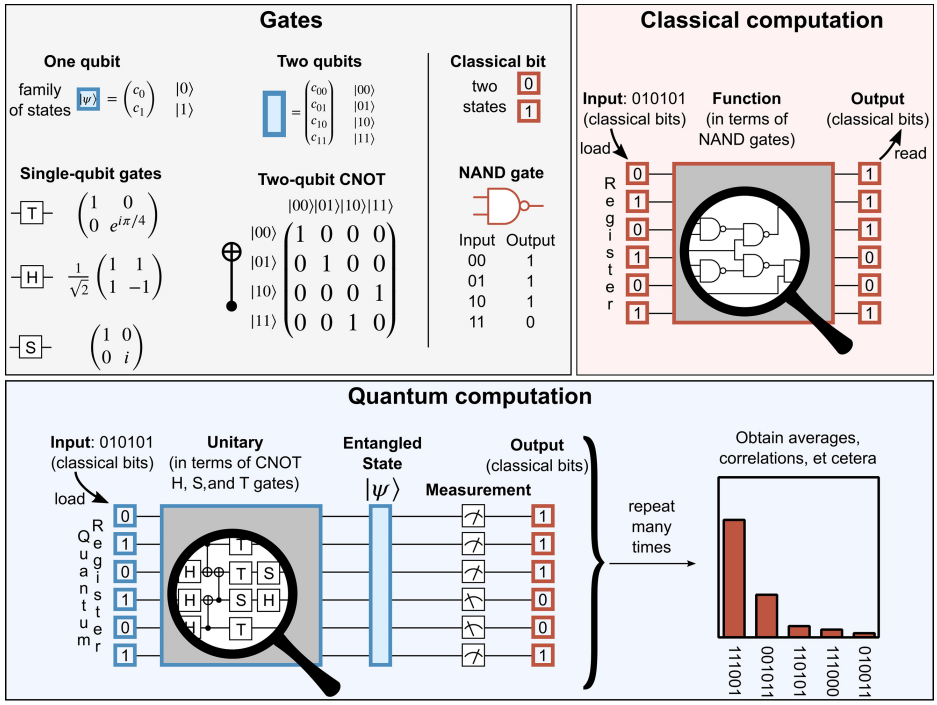

Figure 7.1. Comparison between quantum and classical information processing. Top left: A universal set of quantum gates is (in this case) composed of one two-qubit gate (CNOT) and three single-qubit gates (H, S, T). Classically, the NAND gate (and a copy operation) is universal. Top right: Classical computation can be understood as computable functions acting on an input, provided as a string of classical bits. Bottom: In quantum computation, the input is (usually) again classical data, which are loaded into a quantum register made out of qubits. A quantum algorithm is a unitary operation acting on the qubits, expressed in terms of a quantum circuit. The outcome of a computation is an entangled state. The experiment is repeated many times, which allows one to sample from the probability distribution generated by the state.

example, the state cannot be measured directly. Even if we know exactly which state the qubit is in, if we measure it, the measurement device will record either 0 or 1 at random, with probabilities $|c_0|^2$ and $|c_1|^2$. But this is different from the lack of knowledge present in the coin example. Indeed, knowing that the system is in the state equation (7.1), we could apply an (invertible, deterministic) operation that turns the system state into $|0\rangle$ without needing to perform a measurement whether the qubit was in $|0\rangle$ or $|1\rangle$. The fact that randomness comes out of deterministic operations and thus is an intrinsic property of Nature initially led to a lot of debate, but since then an incredible amount of effort has gone into making sure that this description is correct. Nowadays, such true randomness finds important application, for example, in cryptography (see section 7.2.1.8.4).

7.2.1.2.2 Entangled qubits

If we take two qubits, their possible states are $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$. For a classical computer, this is the end of the story. Quantum mechanics again permits the system

to be in a superposition but now encompassing both particles. One example is the Bell state:

$$|\Psi^+\rangle = (|01\rangle + |10\rangle)/\sqrt{2}\,. \tag{7.2}$$

This state has the fascinating property that when the constituent qubits are considered individually, there is an equal chance to find each in 0 or 1, just as a coin could come up heads or tails with equal probability. When the state of both qubits is measured, however, these random outcomes are perfectly (anti-)correlated and come up either as 0–1 or 1–0. This correlation is intrinsic in the state and has nothing to do with an interaction between them. The qubits could reside at opposite ends of the Universe and they would still exhibit this property. We call this astonishing property *entanglement*. Einstein, Podolsky, and Rosen famously questioned the reality of this 'spooky action at a distance' [15], but experiments have since confirmed this property, thereby ruling out local realistic theories, which are incompatible with quantum physics [16, 17].

Entanglement has a further important consequence: Writing down the state of N qubits requires specifying 2^N parameters, unlike the state of N bits, which is contained in a string of N 0s and 1s. Specifically, an N-qubit quantum state can be written as

$$|\Psi^+\rangle = \sum_{i_1,i_2,\ldots,i_N=0}^{1} C_{i_1,\,i_2,\,\ldots,\,i_N}|i_1,\,i_2,\,\ldots,\,i_N\rangle, \tag{7.3}$$

where the sum on the right-hand side runs over all possible configurations, which are sometimes referred to as computational states, that is, all numbers in binary from 0 to $2^N - 1$. In the worst case we thus have to keep track of 2^N numbers, which is the reason why simulating quantum systems on classical computers is difficult. A rough estimate shows that even with only 300 qubits the number of states ($2^{300} \approx 10^{90}$) exceeds the number of baryons in the observable universe. Quantum computing aims to exploit this complexity.

7.2.1.3 Classical versus quantum computing

7.2.1.3.1 Models of computation

There are many different models of computation, each describing a way in which information can be processed. Those models that are universal, or Turing complete, are all equivalent in the sense that the class of functions whose output they can compute coincides. In this sense, quantum and classical computers are equivalent. However, quantum computers disprove the common belief that all universal models of computation have the same complexity classes. Specifically, quantum computers can solve certain problems in polynomial time that classical computers cannot.

One universal model of computation is the circuit model, where the computation (the evaluation of a function on the input) is represented by a sequence of logic gates acting on the input, as illustrated in figure 7.1. Turing completeness implies that any computable function can be represented in this way. Classically, one can show that

any computation can be decomposed into a series of NAND gates (and a copy operation), which is a gate that takes two inputs and returns 1 except when both inputs are 1, in which case it returns 0 (see figure 7.1 for the truth table).

The standard model of quantum computation is also gate based and is shown in figure 7.1. The operations performed by a quantum computer on its input (a quantum state) are so-called unitary transformations (or *unitaries* for short). One can show that all unitaries can be decomposed into a quantum circuit comprising only single-qubit unitaries and one type of two-qubit gate, the CNOT gate [18], which constitutes the analog of universality for a quantum computer. Solovay and Kitaev have proven that the continuous set of single-qubit gates can be efficiently decomposed into just few gates drawn from a universal gate set. There are many such universal sets; a commonly employed one is shown in figure 7.1 and comprises the H, T, and S gates. Computational universality also exists in the quantum realm, in that different models of quantum computation are equivalent and possess the same complexity classes. In practice, this means that one can get clever in choosing the allowed operations to reduce errors or overhead, but this will not affect the computational power in a fundamental way. The key point is that the classical and quantum complexity classes are distinct, which in the circuit language means that there are quantum circuits that are exponentially shorter than the best available classical circuits tackling the same problem.

7.2.1.3.2 Accessing the outcome of the computation

The outcome of applying a given quantum circuit is a N-qubit state such as the one specified in equation (7.3). It is important to note that we cannot access the full N-body state directly. Instead, we typically measure each qubit, determining whether it is in 0 or 1. This measurement necessarily destroys the state and yields at most N (classical) bits of information. The measurement outcome is a randomly drawn bit string out of those that make up the state (cf equation (7.3)), where the string $(i_1, i_2, \ldots, i_N)$ occurs with probability $|C_{i_1, i_2, \ldots, i_N}|^2$. Repeating the same experiment many times thus yields not a single result but a probability distribution. The goal of a quantum algorithm is that this probability distribution peaks at the desired solution, such that it can be extracted in a small (polynomial in N) number of measurements. In a simple case, we might have a way to check whether a solution is correct and would like it to come up with reasonable probability during the experiment. In this case, repeating the algorithm a sufficient number of times solves the posed problem.

7.2.1.4 Quantum algorithms

The principles of quantum mechanics allow one to design algorithms that work in fundamentally different ways and thereby can be executed more efficiently. Efficiency is judged via the time and memory requirements of an algorithm given its input size, which is referred to as its complexity. The most important distinction that is made is whether algorithms run in polynomial time or in exponential time, that is, whether for large inputs N the time or memory scales as N^k or as $\exp(kN)$ for some fixed k. The most useful and important quantum algorithms offer exponential

speedup, which means that their complexity behaves like the logarithm of the complexity of the best classical algorithm. The nature of exponential scaling means that problems for which only exponential time algorithms are available are in practice unsolvable for moderately large inputs. In this sense, quantum computers can solve problems classical computers cannot.

A simple academic example, introduced by David Deutsch in 1985 [19], illustrates how quantum mechanics can help solve a problem. The problem he considered is that we are provided with a quantum system (an oracle) that implements an unknown function h: $\{0, 1\} \rightarrow \{0, 1\}$. If we send the oracle a two-qubit state $|x, y\rangle$, it returns the output two-qubit state $|x, y \oplus h(x)\rangle$, where the function h can either (1) do nothing, $h(y) = y$, (2) flip the qubit, $h(y) = \bar{y}$, (3) always return 0, $h(y) = 0$, or (4) always return 1, $h(y) = 1$. We are asked to determine whether h is balanced, as in (1) or (2), or constant, as in (3) or (4). While classically we would have to check the output on two different inputs and thus need two function calls, in quantum mechanics this can be done using just one function call. The key is to prepare the first qubit in the superposition $|+\rangle = (|0\rangle + |1\rangle)\sqrt{2}$ and the second qubit in the superposition $|-\rangle = (|0\rangle - |1\rangle)/\sqrt{2}$. One can then show that the output of the machine is $|--\rangle$ if h is balanced and $|+-\rangle$ if h in constant. These two states can be distinguished in a single measurement, which means that only one function call is required, in contrast to the classical case.

While the preceding algorithm was an academic toy problem, there are known useful algorithms, which, however, are substantially more complex. We list some of the best-known in the next subsection. This shows that there are problems for which quantum computers are required. Nevertheless, we would like to draw the analogy with the development of classical computers, in which most of the important algorithms in use today were invented only after classical computers were available to a broad community. Similarly, we expect the class of useful quantum algorithms to grow significantly as this technology becomes available.

7.2.1.4.1 List of algorithms

Two key distinguishing properties of quantum algorithms should be kept in mind, as they determine the practical use of the algorithms. One is whether they use an oracle or not. An oracle is a supposed black box with arbitrary computing power that implements a certain function or algorithm, as in the preceding example. This is an important tool in theoretical quantum information research, as it is typically easier to prove statements about the complexity of such algorithms, but they usually have no real-world applications, as implementing the oracle requires too many resources. The second is whether the quantum speedup is polynomial or exponential. Whether or not a polynomial speedup is useful will have to be decided on a case-by-case basis. The reason is that since quantum processors come with significant overhead, it may happen that the crossover beyond which a quantum computer actually has a smaller runtime occurs for very large inputs and runtimes of years or more, which makes them impractical despite the theoretical speedup. In contrast, exponential speedup means that even modestly sized quantum computers can perform calculations beyond the ability of any classical computer in existence. There are several

algorithms that offer exponential speedup, including algorithms to calculate the discrete logarithm, solve Pell's equation, evaluate Gauss sums, determine abelian hidden subgroups for a large variety of groups, or determine Jones' polynomials [9].

We also include heuristic algorithms in the list. These algorithms do not have a provable advantage, but if classical computing is any guide, heuristic algorithms are often very successful, even beyond the scope of their intended use. Prominent examples include simulated annealing [20], which is a very successful algorithm in a great variety of hard optimization problems, and machine learning, which has revolutionized many fields even beyond typical applications in computer science, the lack of theoretical guarantees notwithstanding. By their very nature, heuristic algorithms are difficult to develop if one cannot try them out. Therefore, it is reasonable to expect that the best and most impactful algorithms in this class will be found only once quantum computers have arrived.

- **Deutsch–Josza algorithm** (Deutsch and Josza [21]). This is an oracle-based algorithm similar to the one described in the previous section, but with large inputs. While the quantum algorithm requires only one query, the classical algorithm requires exponentially many. If a small probability of returning the wrong result is permitted, the classical algorithm requires only a constant number of queries. This was one of the first algorithms with provable quantum speedup.
- **Simon's algorithm** (Simon [22]). Simon's algorithm tackles a similarly academic problem, namely, to find the string s given the black box function (oracle) that obeys $f(x) = f(y)$ if and only if $x \otimes y \in \{0^N, s\}$, where N is the length of the bit string. While in the Deutsch–Josza algorithm, quantum algorithms only provide an advantage if the algorithm must always return the correct result, Simon's algorithm provides an exponential speedup even if we allow a small (<0.5) constant probability that the algorithm fails.
- **Shor's algorithm to factor large numbers and to find discrete logarithms** (Shor [23]). This algorithm was the first quantum algorithm with exponential speedup and a practical (oracle-free) application. It was inspired by Simon's algorithm and uses what is now known as the quantum Fourier transform. The technique can be adapted to solve all abelian hidden subgroup problems. Relatedly, the quantum phase estimation algorithm (Kitaev [24]) employs the quantum Fourier transform to find the eigenvalue of a unitary operator U that corresponds to a given eigenstate $|\Psi\rangle$.
- **Grover search** (Grover [25]). This is a generic algorithm to do an unstructured search. Given a desired output, it finds the correct input to a black box function (an oracle) among N choices using $O\sqrt{N}$ function calls, whereas classical algorithms require $O(N)$ calls. The central element of Grover search is amplitude amplification, which gives a quadratic speedup for generic optimization problems [26].
- **Adiabatic algorithms** [27]. This is a heuristic algorithm based on physical insight which often works well but for which no convergence guarantees exist. Many problems (notably optimization problems and many-body problems in

physics and chemistry) can be encoded by specifying a Hamiltonian H, which is a Hermitian operator that determines the system properties and its time evolution, and asking what its ground state (eigenstate with the lowest eigenvalue) is. The adiabatic theorem states that if a system starts in the ground state, and its Hamiltonian is changed very slowly, the system will remain in the ground state as long as the ground state remains unique; similar to how one should avoid jerky movements in an egg race or when carrying a full glass of water. The adiabatic algorithm proceeds by preparing the (known) ground state of a different operator H', applying the time evolution, and slowly changing the Hamiltonian to H. It fails if the energy difference between lowest and second-lowest energy eigenstate becomes too small.

- **Variational algorithm.** This is another heuristic algorithm that aims to find the ground state by defining a quantum circuit whose gates depend on some parameters and optimizing the energy (or any other cost function) with respect to the parameters. It may fail if it gets stuck in a local minimum or if the ground state is not inside the variational class.

7.2.1.5 Key applications

We now outline some known applications using the previously introduced algorithms.

- **Many-body quantum simulation.** A clear and natural use case for quantum computers is modeling quantum systems. There is a vast array of applications ranging from solid-state materials (as used in classical computers, batteries, and electric or magnetic materials) to chemistry (modeling reactions and catalysis) and high-energy physics (e.g., phases of lattice gauge theories). Numerical simulation is a key tool employed by a large fraction of researchers in these fields, but the computational complexity of simulating quantum systems has stymied progress. Thus, a quantum computer has the potential to enable breakthroughs in these fields with important indirect applications.
- **Prime factorization.** Many cryptographic protocols are based on the (presumed) hardness of prime factorization, including the ubiquitous Rivest–Shamir–Adleman (RSA) public key algorithm. While this could be regarded an antiapplication, as breaking cryptographic schemes is undesirable, it is nevertheless very important. Variations of Shor's algorithm can break other cryptographic schemes based on the hidden subgroup problem [9]. An alternative are postquantum cryptographic schemes that use functions that cannot be inverted even by a quantum computer (at least not with known quantum algorithms) or quantum cryptography, which makes it physically impossible to eavesdrop.
- **Polynomial speedup for optimization problems.** There are several optimization problems where a polynomial speedup can be shown. Finding the optimal solution to a problem (e.g., finding the fastest route) can typically be connected to some optimization problems and attacked with amplitude amplification, which gives a quadratic speedup. Other speedups exist for a

variety of problems [9]. Other potential (heuristic) algorithms may be found by mapping the optimization to finding the ground state of a Hamiltonian, which can sometimes be solved with adiabatic algorithms or quantum annealing.

- **Solving linear and nonlinear systems of equations.** A number of algorithms have been developed to solve linear and nonlinear (differential) equations with exponential speedup. These algorithms are typically formulated in terms of an oracle or access to a quantum random-access memory but may offer an advantage in some instances.
- **Inspiration for classical algorithms.** The quantum point of view has already proven to be very fruitful to understand classical problems and will continue to be important, regardless of the success of quantum computers. For instance, there are examples of 'quantum-inspired' classical algorithms that compete with or beat previously known best classical algorithms to solve certain problems, such as the Netflix recommendation problem, in which first a quantum algorithm was found that gave exponential speedup and later a classical algorithm was found that achieved the same [28].

7.2.1.6 Error correction and fault tolerance

A major challenge on the road to QC is that in a real-world device, errors can always occur. Given some error rate p per qubit per unit time, the probability that no error occurs in a single qubit ($N = 1$) and in one unit of time ($T = 1$) is $1 - p$. Thus, the probability for a large computation to succeed tends to zero exponentially fast, $(1 - p)^{NT} \to 0$, which means that we need exponentially many tries to get it right, which eliminates all speedups.

While classical error correction can be achieved simply by a repetition code (if '0' is encoded as '000', a single bit flip error can be corrected), in quantum mechanics this is much harder, since (1) the state of a qubit is continuous (see equation (7.1)), (2) quantum information cannot be copied [29], and (3) (related to (2)) a measurement of a qubit to detect whether an error that occurred alters the state uncontrollably and destroys encoded information. This problem was solved by Shor [30] and Steane [31], who independently invented codes in which a qubit is represented as the entangled state of many qubits in a way that allows one to correct all single-qubit errors.

This by itself does not suffice to show that computation can succeed, since, for example, two errors might happen at once. Worse still, the correction of the error might itself introduce further errors. For classical computers, von Neumann proved a threshold theorem showing that errors can be arbitrarily suppressed. Fault-tolerant quantum computation, introduced by Shor [32], is a scheme that removes the errors during the computation, which succeeds if the individual gate errors lie below a certain threshold. These results ensure that quantum computers are scalable in principle.

There are two main practical implications of these results:

1. There exists an error threshold of the order of 10^{-2}. The device must allow one to perform quantum gates with errors occurring with a probability below this threshold.

2. There is a large overhead. It may require thousands of physical qubits to represent a logical qubits, and logical operations may take many orders of magnitude longer to complete than the underlying physical operations.

These conditions pose substantial challenges, such that a practical fault-tolerant quantum computer is likely still decades away, as we discuss in section 7.2.1.8.

7.2.1.7 Implementations

7.2.1.7.1 Criteria

In the past two decades, several different platforms for quantum computation have been explored, and there exist working prototypes for a number of them. While these prototypes do not achieve fault tolerance, they can still be used for useful computations and simulations.

In 2000, DiVincenzo summarized the criteria that devices ought to fulfil to be viable candidates for quantum computers [33]. They are as follows (bold words correspond to the steps outlined in figure 7.1):

1. *A scalable physical system with well-characterized qubits.*
   This criterion arises from using qubits as the fundamental unit of a quantum computer. Scalable means that increasing the number of qubits should come at a moderate cost only.
2. *The ability to initialize the qubits in a simple fiducial state.*
   This is required to be able to load data into the computer (**input**).
3. *Long coherence times, much longer than the gate operation time.*
   If this is not fulfilled, one cannot apply a **unitary** of sufficient complexity, or else the **entangled state** is randomized.
4. *A universal set of quantum gates.*
   Without a universal set, one cannot implement the desired **unitary**. In figure 7.1, this was taken to be the T, H, S, and CNOT gates, but other sets can be chosen.
5. *A qubit-specific measurement capability.*
   This is needed to obtain **output** from the device.
   More recently, it has been understood that this list has to be supplemented with further conditions to ensure error correction.
6. *Scalability.*
   The error for qubit gates must lie below the error threshold for the used code, and the gate error must be independent of system size.
7. *Parallelization.*
   The number of gates that can be applied in parallel must grow linearly with system size. If this is not the case, the time taken between two gates acting on the same qubit grows with system size, which means that the intrinsic error rate increases.

7.2.1.7.2 Leading platforms

The requirements for quantum computers are steep, but several platforms are under active investigation. It is unclear which one of the following will be the platform of choice for eventual fault-tolerant quantum computers. It may even be that a new platform will be invented, in which the complications associated with fault tolerance can be overcome more naturally. This calls for parallel investigation and an open mind. We show sketches for the main platforms in figure 7.2 and briefly describe them. We provide references to only some of the theoretical proposals, leaving out many groundbreaking experimental results, which can be found in [36] or in more specialized reviews [12–14].

(a) **Trapped ions** [37]. Ions (black) are levitated in vacuum and tightly confined along two dimensions. In the third dimension they naturally form a string, stabilized by electrostatic repulsion (yellow). The ions are typically chosen to be an element with naturally two valence electrons, such that after ionization one remains. The qubit is formed by two internal states of the ion that are chosen as to have a large coherence time, such as dipole-forbidden excited (clock) states or hyperfine levels. Interaction occurs via the coupled motion of the ions and is controlled via applied lasers (red). This approach features very long coherence times and extremely high gate fidelities. Readout occurs in the transverse direction. A number of groups and companies have pursued this approach, and several tens of qubits have been fully controlled in this setup already.

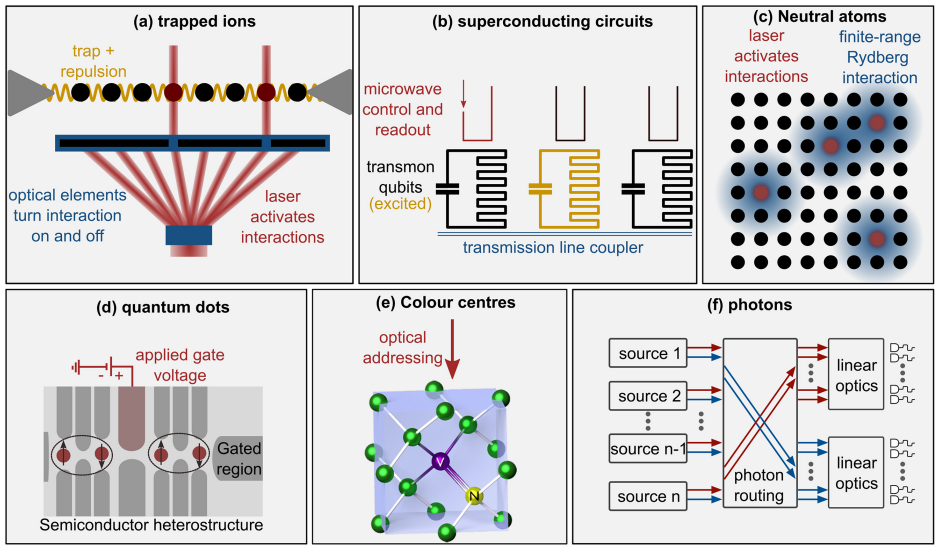

Figure 7.2. Platforms for quantum computation. Explanation in main text. Part (e) is from the NIST public domain [34]. Part (f) is inspired by reference [35].

(b) **Superconducting circuits** [38]. Nanofabricated circuits of a superconducting material are placed on a substrate. The qubit is defined as the presence or absence of a single excitation in a nonlinear oscillator. The qubits are driven with microwave tones supplied via transmission lines, which can enact both single-qubit gates and two-qubit gates; the latter use the interaction mediated through a shared resonator.

Superconducting chips are being researched by several major companies and many groups and are among the largest devices produced so far, with qubit numbers approaching the low hundreds.

(c) **Neutral atoms** [39]. Individual atoms are trapped in a lattice defined through either back-reflected laser beams or optical tweezers (not shown). A qubit is defined through the electronic ground state and another long-term stable electronic state. Interactions can (for example) be mediated by driving a transition to a Rydberg state on both atoms using a laser (shown in red). The interaction is present as long as the corresponding atoms lie within the Rydberg blockade radius (blue) of each other. Readout is done optically in the transverse direction. Hundreds of atoms have been trapped in such devices, but full control has not yet been achieved.

(d) **Quantum dots** [40]. A layered semiconductor structure is used to confine electrons to move in two dimensions (light gray). Additional gates (dark gray) are used to expel electrons from selected regions so as to define islands (those with red dots) that can host an electron. An electron is a spin-1/2 particle with two spin states that are used as the logical 0 and 1 of the qubit. Interactions can be controlled by changing the gate voltage between qubits, which brings electrons closer together or moves them further apart, depending on their spin. Being based on semiconductors, this approach can benefit from the large semiconductor industry, but progress has not moved beyond a few qubits yet.

(e) **Colour centres** [41]. Point defects (substituted atoms) in crystals such as diamond yield localized electrons in very clean environments that consequently are very stable. This makes them promising candidates for qubits, but fabrication of arrays of colour centres and control of their interactions are challenging and are only now starting to become practical.

(f) **Photons, linear optics, and detectors** [42]. Photons are mainly investigated as carriers of quantum information, as they can travel long distances before being absorbed or perturbed. Qubits can be defined through the presence or absence of photons or through their polarization. The main challenge in using them to perform computations are two-photon gates, as they hardly interact with each other. One approach uses single-photon detectors to perform gates probabilistically. This has been successfully used to entangle tens of photons.

7.2.1.7.3 Other important platforms

There are a number of platforms for which some important ingredients are unclear or have not been demonstrated experimentally as yet but that nevertheless look very promising.

(a) **Nuclear magnetic resonance quantum computer** [43]. An NMR quantum computer builds on manipulating the spin states of nuclei in molecules. It is generally understood that this platform scales poorly. Alternative proposals, based on solid-state nuclear spins, may remove this roadblock [44].

(b) **Topological quantum computer** [45]. This is a fundamentally different proposal, in which quantum gates are implemented by braiding anyonic quasiparticles in two-dimensional materials with topological order. The topological nature of the operations imply resistance to local noise and imperfections, which may make fault-tolerance much easier to achieve. Experimentally, this platform is still in its infancy, and no clear candidate material that hosts appropriate excitations has been realized.

(c) **Quantum annealers** [46]. Quantum annealing is a heuristic procedure to reach the ground state of a Hamiltonian (and thereby solve an encoded problem) by slowly lowering the temperature. These systems are not universal quantum computers but may still offer a quantum advantage for certain problems.

(d) **Measurement-based quantum computation** [47]. This is an alternative way to perform quantum computation that proceeds by (i) preparing an entangled state of many qubits and (ii) measuring the qubits sequentially in appropriate bases. It can be shown that by choosing a state with appropriate entanglement, universal quantum computation can be performed in this way. This is a very appealing model for photonic platforms.

7.2.1.8 Present situation and outlook

For several years now, the basic ingredients of quantum processors have been experimentally demonstrated in most of the platforms introduced in section 7.2.1.7. That is, several qubits have been prepared in the 0 state, manipulated with single- and two-qubit gates, and then measured. None of these operations is perfect, but each of them can be performed with errors of around 1% (or even smaller). This has also made it possible to perform small proof-of-principle quantum computations with a few qubits, which illustrate how the algorithms operate.

In some systems (most notably trapped ions, Rydberg atoms, and superconducting circuits), prototypes of several tens of qubits have been built. It is estimated that, at 30 qubits and above, it is virtually impossible to simulate the operation of a quantum computer with existing classical computers, so we are reaching a situation where the former can outperform the latter. However, there are several reasons for tempering optimism, at least as far as the range of applications of quantum computers is concerned. On the one hand, since the errors accumulate during the whole process of initialisation, execution, and measurement, the success probability

becomes exponentially small in the number of elementary operations performed. For example, even with errors at the 1% level or below, a quantum algorithm requiring 500 operations will almost certainly give a wrong result. Moreover, even if errors are reduced further, many quantum algorithms that have been developed over the years do not work correctly in the presence of errors. On the other hand, there are powerful classical algorithms for classical computers that can simulate some quantum computer operations efficiently. For example, if each qubit participates in only two (or fewer) quantum logic gates, then tensor network-based methods can be used to predict the outcome with a classical device. This sets a high bar for quantum computers to clear.

It is very difficult to use current prototypes or those to be developed in the near future to gain a real computational advantage over existing computers. Therefore, it seems that the only way to exploit computers in the future is to try to correct the errors and scale up these prototypes. This may take several years or even decades, since, as we shall see later in the chapter, it requires solving a number of major technological challenges. However, there is the possibility that these prototypes, which still produce errors, could be useful for some concrete tasks. In what follows, we will briefly describe those possibilities, as well as the road ahead for creating scalable quantum computers.

7.2.1.8.1 Noisy intermediate-scale quantum (NISQ) devices

To date, only prototypes exist in the platforms listed above. They contain several tens (or, potentially, hundreds or thousands) of qubits and operate without fault-tolerant error correction. In a recent breakthrough, researchers at Google ran an algorithm on a 53-qubit device that would take years for a classical computer to complete [4]. This and other results have opened the door to the hope that such devices will have some application in the near future.

Google's experiment, later replicated and partially improved by a team at the University of Science and Technology of China [6], consists of the following. With the $N = 53$ qubits arranged in a square geometry and initialized in 0, one defines a quantum circuit comprising logic gates between neighbouring qubits. The gates are chosen at random at the beginning of the calculation (and are then kept the same). As a result of the random circuit, measurement outcomes at the end of the whole computation are close to random (but not quite), such that the probability P_x for each outcome $x = (x_1, \ldots, x_N)$ (with $x_i = 0, 1$) is very close to $1/2^N$ for each of the 2^N possible outcomes. In the blue panel shown in figure 7.1, this would produce an essentially flat histogram of outcomes. We can then specify a computational task, namely, to sample bit strings according to this probability distribution, with a small prescribed error, chosen to be about 10^{-3} with respect to the probability. This means that the simulated (quantum or classical) probability distribution $\tilde{P}_x$ should obey $\sum_x |P_x - \tilde{P}_x| \leqslant 10^{-3}$. The quantum algorithm that performs this sampling is trivial, since it consists of executing the quantum logic gates and then measuring, a procedure that is repeated to obtain several samples (keeping the same logic gates). However, there is no known classical algorithm that can perform this task

efficiently, even if it is allowed to have the mentioned error. Note that even though each $P_x \approx 1/2^N$, replacing P with the uniform distribution is not sufficiently precise, as small deviations (e.g., by a factor of 2 or 1/2) accumulate to yield a sizeable error. In the experiment, the error per logic gate was approximately 0.3%, which allowed the task to be performed successfully. An important point is how one can verify that the NISQ device executed the task correctly, since it is not possible to verify the result with a classical computer (and, moreover, it would be necessary to obtain an exponential number of samples to be able to see that they correspond to the correct result). In the experiment, verification was done using careful analysis and extrapolation of results obtained with a smaller number of qubits (or a subset of logic gates) for which the result could be predicted with a classical computer.

The problem solved by Google is purely academic, chosen to demonstrate so-called quantum supremacy (a better term might be *quantum advantage*), although with no obvious application. It nevertheless demonstrates that NISQ devices are already capable of performing tasks that are not possible with standard computers. The question naturally arises whether there are real applications for which these devices provide a quantum advantage. This question is being studied intensively in many universities and research centres and in industry. Among the most studied candidate problems are process optimization problems. These are usually combinatorial problems, where one has to decide which configuration optimizes a variable. A standard example is the traveling salesman problem, in which, given a set of geographical points, the objective is to find the path through all of them that minimizes its length. There are many other similar problems with wide applications in industry and beyond that are difficult to solve with classical computers and that could potentially be accelerated with NISQ devices.

However, there are at least two reasons to be cautious. First, long quantum circuits lead to large errors, which on the one hand may mean that one needs to repeat the simulation an unworkable number of times, leading to a large overhead. Second, the presence of errors means that we should compare these to classical algorithms that obtain the result with a certain margin of error, which are often substantially faster. We note here that even when devices become fault tolerant, the large overhead of error correction may imply that for problems with a modest polynomial speedup, quantum advantage (i.e., the crossover between classical and quantum runtime) is achieved only for very large problem sizes and at runtimes of years, which would render them impractical. Nevertheless, it is worth insisting that it is also not clear that NISQ devices *cannot* accelerate optimization. For this reason, the current research efforts are very important. Moreover, as was mentioned earlier, research into quantum algorithms may inspire novel, more efficient classical algorithms, as has happened before.

Another area in which NISQ devices may give rise to an advantage is in the area of artificial intelligence (AI) and, more specifically, machine learning. This advantage can come from several directions. On the one hand, quantum states express certain probability distributions very efficiently. In fact, this is precisely what Google's experiment demonstrates: If a process we want to learn were given as a probability distribution obtained from a quantum circuit, we likely would not be

able to sample from it efficiently with a classical computer. Machine learning may offer problems that require this kind of probability distribution. On the other hand, an advantage may arise if the training process is more efficient on a quantum computer. Since training can be considered an optimization process, the arguments from the previous paragraph apply.

Other applications concern the solution of equations of many variables (linear, nonlinear, or differential). In this case, there are several algorithms that require an oracle or quantum random access memory (i.e., where it is assumed that there is a process that performs a subroutine whose execution time is ignored). Ignoring the runtime of the subroutine means that these algorithms are generally not useful in practice (they serve to illustrate the power of quantum computers in an abstract way), but there may exist situations in which an advantage remains even if the subroutine is taken into account. How to adapt these algorithms to NISQ devices and account for errors is also an area of active research.

In addition to the problems mentioned previously, it is to be expected that in the near future, new applications of NISQ devices will be discovered that are difficult to imagine today. The reason is that these devices have just been created, and hence algorithmic research is still in its infancy. This leads to considerable optimism and motivates further research. To mention one concrete example, it has recently been discovered that a QC device that can demonstrably execute quantum circuits can be used to generate certified random numbers. That is, we can ensure that these numbers are generated on the spot and are intrinsically random (i.e., neither precalculated nor from a quasirandom classical algorithm). This is something for which there is no classical analog and provides an unexpected new application for quantum computers.

7.2.1.8.2 Quantum simulation

The most promising application of quantum computers, both current (NISQ) and future (fault-tolerant) devices, is the simulation of complex quantum systems. Complex quantum systems appear in various incarnations in physics and chemistry and consist of a set of objects interacting with each other according to the laws of quantum physics. The objects can be atoms, electrons, spins, photons, or any other particle (elementary or not). In the case of physics, quantum many-body problems naturally appear in condensed matter physics, high-energy physics, or atomic physics, to name a few. Examples include the electrical or magnetic properties of materials at low temperatures, quantum electrodynamics, quantum chromodynamics, and strongly interacting ultracold atoms. In chemistry, they appear when studying the geometrical structure of molecules, their electronic or spin properties, their dynamics, or their reactions. On classical computers, the solution of the associated problems requires resources (memory or computing time) that grow exponentially with the number of objects (atoms, electrons, etc) or the volume of space in which they are located (cf equation (7.3) and subsequent discussion). Because of this, we usually have to use approximate methods or restrict to a few particles.

A quantum computer can circumvent this problem, as it can naturally implement quantum states. To simulate a many-body quantum system, a quantum computer must simply prepare the desired state Ψ (e.g., the ground state or the state resulting from time evolution) in N qubits and subsequently can obtain the required physical properties by making the appropriate measurements. The device naturally circumvents the memory problem, needing N qubits instead of 2^N bits. To simulate quantum dynamics on a quantum computer, one can execute the quantum logic gates that correspond to the evolution of the system, which requires a number of steps that scales polynomially with the simulation time, the number of qubits, and the inverse of the prescribed error tolerance. If one wishes to determine the ground state or equilibrium properties of a system, there are several different algorithms that perform this task. Although in the most general case, the execution time scales exponentially with N even for a quantum computer, in certain cases this time can be shortened using adiabatic and variational algorithms (see section 7.2.1.4). Either way, it is clear that quantum computers are ideally suited for quantum simulation.

Quantum simulation is also a very interesting prospect for NISQ devices. The reason is that, even with the presence of errors, it is still conceivable that useful results can be obtained in simulation problems. For example, to determine the phase of a material (e.g., is a magnet ferromagnetic or paramagnetic?), it is not necessary to obtain absolute precision in the computation of its physical properties, so even in the presence of sufficiently small errors, a quantum computer could solve problems that would be difficult to address with a classical one. An even more attractive option is that of analog quantum computation. In that case, the idea is to build a laboratory experiment in which the interactions between the qubits (or other particles) can be tuned to some degree. The goal is to engineer this system such that it emulates a certain real-world system. Apart from the obvious advantage that building one of these computers can be easier, because less precise control is needed, an analog computer is most appropriate for simulating the dynamics of quantum many-body systems, obviating the need to discretize time and avoiding the decomposition into logic gates and, therefore, errors. Platforms based on neutral atoms (in either Rydberg or hyperfine levels) are, together with trapped ions, the most advanced.

7.2.1.8.3 Fault tolerance and beyond

In the long term, the goal of QC is the development of scalable computers, that is, computers that are error-free and whose qubit number can be increased arbitrarily with reasonable effort. This is possible only with fault-tolerant quantum error correction, which, as was mentioned earlier, requires that both the initialisation error of each qubit and the error per logic gate (and, if possible, in measurement) are sufficiently small and do not grow as the system scales up. In addition, it is necessary to be able to apply the logic gates in parallel. This is an extraordinary challenge for all current platforms. First, according to current error-correction codes, many physical qubits are needed to encode a single logical qubit. This multiplies the number of required qubits by a large constant factor, which is several thousands for

errors of the order of 0.3%. This also means that many more logic gates are required for the dynamic correction of errors as they occur. It is estimated that for many applications of quantum computers, quantum advantage will require the use of hundreds of thousands if not millions of physical qubits. Engineering such large devices constitutes a formidable technological challenge and is therefore unlikely to happen in the next decade. Nevertheless, it is well worth the effort. Advancement in the field must be made not only by further developing the existing platforms, but also by looking for new ones as well as developing more efficient error correction schemes that are adapted to the specific weakness of each implementation.

How will quantum computers will affect us in the next few years? As we stated in the introduction, this question is exceedingly difficult to answer, as it heavily depends on future insights and technological breakthroughs, which, by definition, are not available to us now. We have gone to great lengths to emphasize the challenges connected to scaling up quantum computers. However, even if building fault-tolerant quantum computers takes a long time, NISQ and analog quantum computers will still be of great use in the mean time.

In the long run, the question is not whether a scalable quantum computer will be built but rather when. What seems clear, looking back at the history of classical computation, is that the most impactful applications of quantum computers are yet to be discovered. In the end, the joint research effort, uniting fundamental and applied research, academy and industry, theoreticians and experimentalists, will surely bring new and exciting avenues for research and development.

7.2.1.8.4 Other applications

In this section we concentrated on QC, as it is the quantum technology that has raised the most interest not only in the scientific world, but also in industry and society. However, quantum science gives rise to other applications as surprising as or more surprising than this first one. We conclude by mentioning some of them.

Quantum communication consists of sending messages encoded in quantum superpositions. This way, two objectives can be achieved. First, information can be encrypted in such a way that even a quantum computer cannot decrypt it. This is in contrast to many of the traditional forms of encryption we currently use, which can be attacked by quantum computers. This new form of encryption has given rise to what is known as quantum cryptography and, more specifically, quantum key distribution. Quantum cryptography is now a reality and is even being exploited commercially. The transmission is done through photons sent through optical fibres. At the moment, given that we do not have quantum computers that can attack traditional systems and that it requires specific hardware, it is not very active. Moreover, current cryptographic systems are not completely secure, as they would require hardware conditions (such as high-efficiency photon detectors) that cannot be easily incorporated into them. Furthermore, absorption of photons in the fibres limited such secure connections to a few kilometres. Two different ways of extending these ranges are currently being investigated: through quantum repeaters and on by routing them through satellites. Although there are very promising results, further developments are still needed to achieve these goals. We should mention that

alternative strategies exist to make cryptographic schemes robust against attack from quantum computers. To do so, one has to replace the current algorithm, which uses the fact that finding primes is easy but prime decomposition is difficult, by another algorithm based on a problem that is believed to be hard even for a quantum computer. One large class of such schemes are based on learning with errors. Postquantum schemes come with the caveat that one typically cannot prove that the chosen problem is hard for a quantum computer and thus may fail suddenly when such an algorithm is discovered, which may not be made public. Quantum cryptography does not have such asterisks and may thus be safer in the long run. Second, quantum communication can also be more efficient. Tasks such as voting or agreeing on a date can be done more efficiently. Similar to speedups in QC, sending qubits in specific entangled states can lead to vastly reduced resource requirements in terms of the number of photons that need to be exchanged to solve a remote problem. In addition to these applications, there are a variety of other applications including unforgible quantum money or quantum tokens.

Quantum randomness is an intrinsic property of quantum systems (see section 7.2.2) which cannot, even in principle, be predicted. Using quantum randomness thus eliminates one potential attack on protocols relying on randomness, such as encryption.

Another field in which quantum physics offers advantages is that of *sensors and metrology*. It has long been known that atomic clocks base their accuracy on the possibility of creating quantum superpositions and that the use of entangled states can further increase that accuracy. Also, these quantum properties can help build better gravitometers, accelerometers, or other sensors in general. There is a lot of research and development activity in this field, and these sensors are likely to have important applications in other fields, such as medicine.

7.2.2 Milestones of research activity in quantum computing

Zeki Can Seskir[1] and Jacob Biamonte[2]
[1]Karlsruhe Institute of Technology—ITAS Karlstraße 11, 76133 Karlsruhe, Germany
[2]Skolkovo Institute of Science and Technology Bolshoy Boulevard 30, bld. 1, Moscow 121205, Russian Federation

We argue that QC underwent an inflection point circa 2017, when long-promised funding materialised, which prompted public and private investments around the world. Techniques from machine learning suddenly influenced central aspects of the field. On one hand, machine learning was used to emulate quantum systems. On the other hand, quantum algorithms became viewed as a new type of machine learning model (creating the new model of variational quantum computation). Here, we sketch some milestones which have lead to this inflection point. We argue that the next inflection point will occur around when practical problems are first solved by quantum computers. We anticipate that by 2050 this will have become commonplace, while the world will still be adjusting to the possibilities brought by quantum computers.

7.2.2.1 General overview

What is quantum computing?
Today's computers which we know and love—whether smartphones or the mainframes behind the Internet—are all built from billions of transistors. (A transistor is an electrically controlled switch which is ultimately the power behind any electronics.) While transistors utilize quantum mechanical effects (such as tunneling, in which an electron can both penetrate and bounce off an energy barrier concurrently), the composite operation of today's computers is purely deterministic or classical. By classical, we mean classical mechanics, which is exactly the physics (also known as mechanics) that we would anticipate day to day in our lives. Quantum computers are not intended to always replace classical computers. Instead, they are expected to be a different tool to complement certain types of calculations. We will elaborate shortly.

The term *quantum mechanics* (in German, *quantenmechanik*) dates to 1925 in work by Born and Jordan [48] and comprises the physics governing atomic systems. Quantum mechanics contains principles and rules that appear to contradict the classical mechanics we are so intuitively familiar with. Such counterintuitive phenomena provide new possibilities to store and manipulate (quantum) information. This is exactly what a quantum computer should do. The information must be stored and processed in the matter. You can think of a quantum computer as providing new mechanisms to store, process, and generally manipulate data. Indeed, the ultimate limitation of computational power is given by quantum physics.

How did quantum computing begin?

QC dates back at least to 1979, when the young physicist Paul Benioff [49] at Argonne National Laboratory in the US proposed a quantum mechanical model of computation. Richard Feynman [50] and independently Yuri Manin [51] suggested that a quantum computer had the potential to simulate physical processes that a classical computer could not. Such ideas were further formulated and developed in the work of Oxford's David Deutsch [52], who formulated a quantum Turing machine and applied a sort of anthropic principle to the plausible computations allowed by the laws of physics. What we now call the Church–Turing–Deutsch principle asserts that a universal (quantum) computing device can simulate any physical process. (This hypothesis does not give an algorithm but just provides an assertion that such an algorithm exists.) Yet even the most elementary quantum systems appear impossible to fully emulate using classical computers, whereas quantum computers would readily emulate other quantum systems [50, 63].

Early insights into the computational power of quantum computers were based on the assertion that it is impossible to develop an efficient classical algorithm able to accurately emulate quantum systems [50, 53]. While this claim has not been formally proven, ample empirical evidence supports it. Yet there are many computational problems where the required computational resources are better understood than simulating physics. These problems arise in the form of, for example, the theory of numbers, groups, or properties of graphs. For decades a small number of researchers worked to understand if one might expect an exponential quantum speedup for problems long studied in computer science. The early algorithms, such as Deutsch–Jozsa [54] and Simon's [55] solved elegantly contrived problems, making prima facie practical merit difficult to envision.

A more practical breakthrough occurred in 1994 when Peter Shor proposed an efficient quantum algorithm for factoring integers. This would open the door to decrypt RSA-secured communications [56]. Shor's algorithm, if executed on a quantum computer, would require exponentially less time than the best-known classical factoring algorithms. Other seminal early findings include Grover's 1996 quantum algorithm [57], which can search through N items in a database in $\sqrt{N}$ steps. It turns out that Grover's algorithm is provably optimal: It is not possible for a quantum computer to search N unstructured elements any faster than $\sqrt{N}$ steps [58], which could be significant in practice. Can these and other quantum algorithms (see table 7.1) be realized experimentally [59]?

7.2.2.2 Where is quantum computing heading?

As of today, billions of dollars of public and private investment are being spent to build quantum computers [68]. In October 2019, Google, in partnership with NASA, performed a quantum sampling task that appears infeasible on any classical computer [69]. Late in 2020, Chinese scientists [70] reported results of like significance. At the core of these developments is the quantum mechanical bit: the *qubit*, as first coined by Benjamin Schumacher [71], who asserts that the term arose

Table 7.1. Expected quantum complexity. Theory predicts that quantum computers have the potential to rapidly execute several important algorithmic tasks. Originally developed for the gate model, variational counterparts to Grover's search [61], optimisation (QAOA [62]), linear [63] (and nonlinear [64]) systems, and quantum simulation (VQE, first proposed in [65] and first demonstrated in [66]) have been developed. Factorization and discrete log can readily be mapped to Ising optimisation problems, yet the scaling remains unclear. Polynomial time variational quantum factoring might instead be accomplished by means of the circuit to variational algorithm mapping [67].

Problem class	Quantum complexity
Factorization	Polynomial [56]
Discrete log	Polynomial [56]
Complete search	$\mathcal{O}(\sqrt{N})$ [57]
Sparse linear systems	Polynomial [60]
Quantum simulation	Polynomial [53]

Table 7.2. A brief comparison of traditional bits and quantum bits (qubits).

The quantum bit or qubit.
A traditional computer operates using bits, 0 and 1. Each classical register or classical memory must be in a single classical state, which is represented by a string of bits (e.g., 0011 or 0100000111). Quantum computers allow quantum registers and memories to be in multiple states concurrently. For this reason, quantum computers are often described as enabling parallel processing.

out of a 1996 conversation with William Wootters (for qubit, see table 7.2). Simply, a qubit is a two-state (binary) quantum system.

There are a variety of physical approaches to creating qubits. For example, a single photon of light can represent two quantum states given by vertical and horizontal polarization. Common qubit realisations include superconducting electronics [69], trapped ions [72], and various optical realisations [73]. Qubits are the building blocks of fully programmable (e.g., universal) quantum computers. The universal quantum Turing machine's abstraction from physics makes it harder to realize; consequently, this model has fallen from interest [49].

Qubits should be isolated from their surroundings yet also be made to interact. Quantum algorithms are described by quantum circuits; such circuits depict actions on and interactions between qubits by quantum gates and encompass today's *de facto* universal model of quantum computation. In practice, design imperfections and random noise can not be avoided, meaning that the ideal qubit can never exist. Such noise processes serve to restrict quantum circuit depth. Over the last several decades researchers have developed a powerful theory of quantum error correction,

proving that qubits need not be perfect to realise high-depth quantum circuits [74]. Indeed, quantum error correction proves that a small amount of noise can be tolerated: called the error tolerance threshold [75]. Hence, according to the laws of physics, nothing fundamentally prevents humans from building a quantum processor capable of executing high-depth quantum circuits.

A universal quantum computer is assumed to be error tolerant through error correction [76, 77]. Throughout the history of quantum computation, several universal models of quantum computation have been developed and shown to be computationally equivalent to the *de facto* quantum circuit model [74]. This includes adiabatic quantum computation [78, 79] both discrete and continuous quantum walks [80, 81], measurement-based quantum computation [82] as well as one of the authors installment proving universality of the variational model [67].

Assuming this idealized (universal) setting, several miraculous quantum algorithms have been developed which would offer an advantage over the best-known classical algorithms. Lower bounding the computational resources required in quantum (and classical) algorithms has, however, proven extremely difficult. How might we rule out the existence of a better algorithm when the computational power of the class of possible algorithms is not fully understood?

For example, regarding recent quantum adversarial advantage demonstrations [68, 70], who is to say that a classical algorithm will not one day be discovered which can replicate the reported sampling task(s)? This does not imply that such assumptions are not without formal footing. Elegant methods exist to compare the power of a classical computer to the power of a quantum computer [83]. It does, however, make the timeline for a practically meaningful quantum computation difficult to predict. So with all of the dramatic progress, what might we expect from NISQ era quantum processors [84]?

7.2.2.3 Design constraints

Design constraints limit manufacturing qubits. Working with current design constraints means that we must find ways to utilize imperfect qubits. The question is whether we can still build meaningful systems with imperfect devices.

Despite outperforming classical computers at an adversarial advantage [69, 70], such a demonstration was tailored to favor quantum processors and has unclear practical applications. Even with Moore's law failing [85], traditional computing resources are ever improving and increasingly accessible. Moreover, the von Neumann architecture has a first-mover advantage: The entire tool chain, the compilers, algorithms, and so on that are in use today are tailored for this architecture. To instead utilize quantum systems as a computational paradigm represents a dramatic change in how problems must be decomposed, encoded, and compiled and in how we think about computing. Perhaps gleaning lessons from sampling [69, 70], we must learn to look towards problems that are more amenable to quantum processors, with desirable criteria such as the following:

1. Problems which bootstrap physical properties of a quantum processor to reduce implementation overhead(s).

Table 7.3. The memory scaling argument asserts that even the worlds largest computer cannot store into its memory any but the simplest quantum states.

The memory scaling argument asserts that quantum states exist which cannot be stored using even the largest classical memories. Early arguments for the quantum computing advantage considered an ideal state of interacting qubits, requiring about $2^{n+1} \cdot 16$ bytes of information to store assuming 32-bit precision. This reaches 80 terabytes (TB) at just less than 43 qubits and 2.2 petabytes (PB) at just under 47, for example, the world's largest memory of the supercomputer Trinity. Hence, applications with $\geqslant$47 qubits might already outperform classical computers at certain tasks. While this argument did not account for noise and approximation/compression schemes to reduce required memory, similar arguments are considered valid lines of reasoning today.

2. Systems and/or models of computation which offer some inherent tolerance to noise and/or systematic faults.
3. Problems which utilize the ability of quantum systems to efficiently represent quantum states of matter, called the *memory scaling argument* (see table 7.3).

These constraints have led us to what is now called the variational model of quantum computation. In the absence of error correction, NISQ era quantum computation is focused on quantum circuits that are short enough and with gate fidelity high enough that these short quantum circuits can be executed without quantum error correction, as in the recent quantum supremacy experiments [69, 70]. Herein lies the heart of the variational model: by adjusting parameters in an otherwise fixed quantum circuit, low-depth noisy quantum circuits are pushed to their ultimate use case.

NISQ circuits typically bootstrap experimentally desirable regularities inline with criterion 1: The gate sequence itself is fixed, while the gate angles can be varied. A classical computer will adjust parameters of a circuit. Measurements will be used to calculate a cost function, and the process will be iterate (see table 7.4 for some examples).

The prospects of the variational model are limited by the computational overhead of outer-loop optimisation. This requires significant classical computing resources. Variational model proposes some alternatives to this. (For a further comparison of these models, see table 7.5.)

7.2.2.4 How did quantum computing develop prior to 2017?

A turning point in the development of quantum computation appears around 2017. At this point, several long-promised large funding programs began, such as the European Quantum Flagship and the American National Quantum Initiative Act (which happened around the world and was in the billions of US dollars). Most national investments appear to keep a country competitive in technological development. There are many initiatives around the world, adding up to more than

Table 7.4. Hamiltonian complexity micro zoo. Anticipated computational resources to determine ground state energy and calculate energy relative to a state. Restricted Ising denotes problems known to be in P. An asterisk denotes expected and not formally proven conjectures. Electronic structure problem instances have constant maximum size and so are assumed to be in BQP, whereas the ZZXX model admits a QMA-complete ground state energy decision problem.

Problem Hamiltonian	Finding ground energy (classical/quantum)	Calculating state energy (classical/quantum)
1-Local Hamiltonian	Polynomial	Polynomial
2-Local ising	*Exp	Polynomial
Electronic structure	*Exp	*Exp/polynomial
ZZXX model	*Exp	*Exp/*polynomial

Table 7.5. Comparison of standard gate model quantum computation versus variational quantum computation. Variational quantum computation trains short quantum circuits to reach their maximum use case yet requires significant classical coprocessing to train these quantum circuits.

Variational	Traditional
• Agnostic to systematic errors • Tightly connects hardware with software to overcome hardware constraints • Optimizes short depth circuits for optimal use • Emulates Hamiltonians by local measurements • Outer loop optimization can require significant classical computing resources • Coherence time and error rates limit circuit depth	• Intuitive and familiar, textbook quantum algorithms adhere to the circuit model • Theoretical analysis, including complexity, has largely been proven possible • Impossible to execute all but the shortest circuits (smallest examples) with current hardware • Ignores hardware constraints and susceptible to both systematic and random errors

20 billion USD committed in public funding. Many private companies also invested dramatically around this time.

Meanwhile, quantum computation was merged with machine learning in two different ways (see [86]). First, quantum circuits can be trained variationally. In other words, quantum circuits can be viewed as machine learning models. Second, machine learning can be applied to a host of problems faced in building and emulating quantum systems. These two facts encouraged the tech industry to participate in QC research and development. (In fact, a new model of computation was developed and proven to be universal [67].)

While those working in the field might readily agree that things have rapidly developed since around 2017, putting data behind this claim is the focus of this

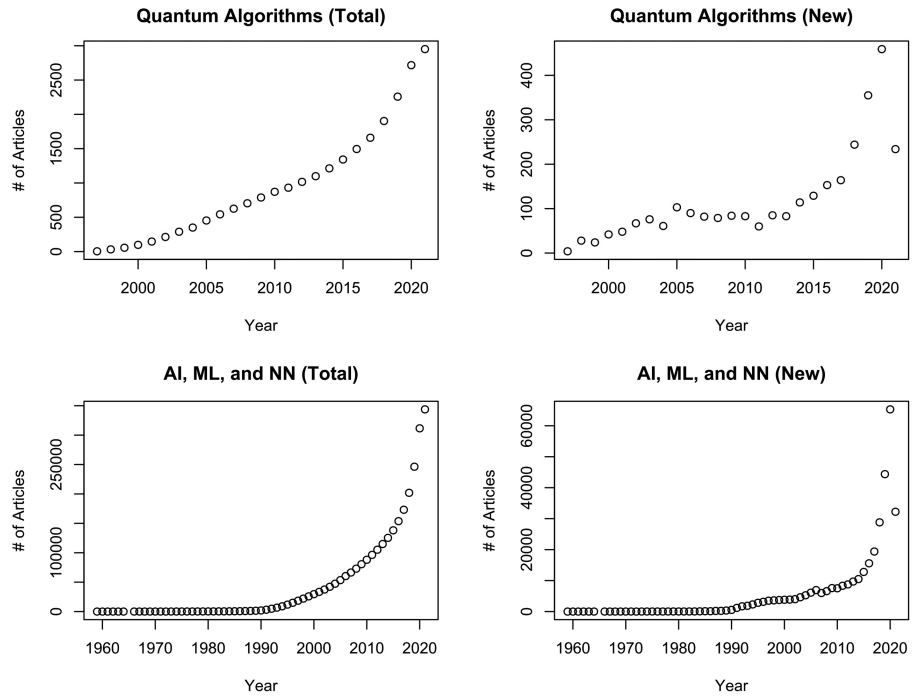

Figure 7.3. The numbers of new and total published articles in the quantum algorithm and artificial intelligence fields.

section. Hence, we will quantitatively describe the levels of activity before 2017, especially focusing on the field of quantum algorithms.

We have utilized two data sources for this section, one for academic publications and one for patent publications: Web of Science owned by Clarivate Analytics PLC and the Cipher platform owned by Aistemos Ltd. However, our readers can use the same queries[1] to reproduce our results. To generate our queries, we have used the previous ones created by previous publications in the literature [87, 88], and personal expertise.

To build our study, we first consider figure 7.3. We see in the upper left-hand panel the growth of quantum algorithms articles. We see what appears to be an increase between 2015 and 2020. By *total*, we mean the entire sum up until that

[1] For quantum software and algorithms: ((‘quantum machine learning’) OR (‘qml’ AND ‘quantum’) OR (‘quantum approximate optimization’) OR (‘vqe’ AND ‘quantum’) OR (‘variational quantum eigen*’) OR (‘quantum algorithm*’) OR (‘quantum software’) OR (‘quantum Machine Learning’) OR (‘Classical-quantum Hybrid Algorithm*’) OR (‘quantum PCA’) OR (‘quantum SVM’) OR (‘variational Circuit*’ AND ‘quantum’) OR (‘quantum Anneal*’ AND ‘algorithm*’) OR (‘quantum Enhanced Kernel Method*’) OR (‘quantum Deep Learning’) OR (‘quantum Matrix Inversion’) OR (‘quantum embed*’) OR (‘quantum neural’) OR (‘quantum perceptron’) OR (‘quantum tensor network*’)) For artificial intelligence: ((‘machine learning’) OR (‘artificial intelligence’) OR (‘neural network*’)) Date: 11 June 11 2021.

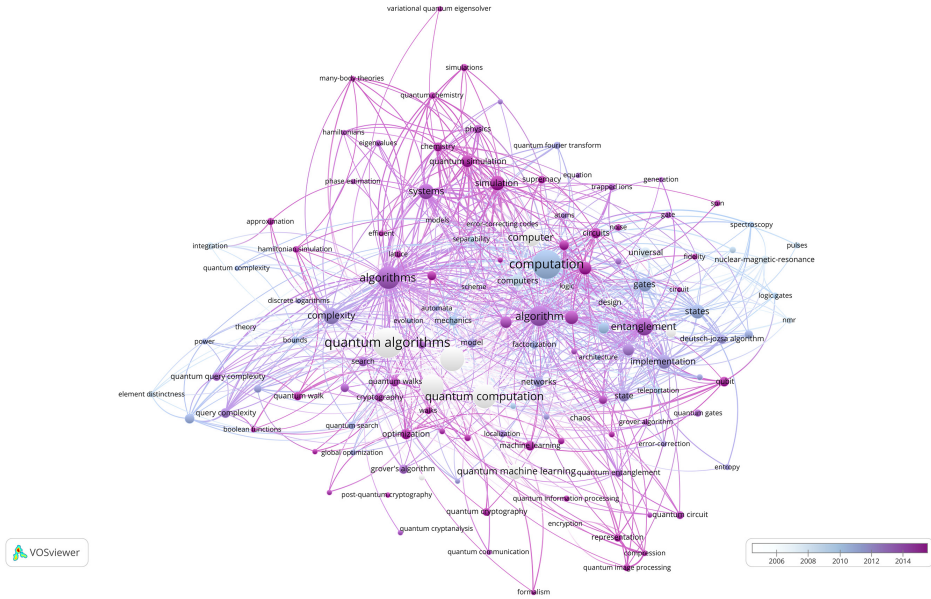

Figure 7.4. Overlay visualization of author keywords used in academic articles in time.

point. By *new*, we mean the number of articles in a given year. In the left top panel of figure 7.3, we again see a jump between 2015 and 2020.

Additionally, we ran the publication dataset through a software tool for constructing and visualizing bibliometric networks (VOSViewer) [89] to run author keyword clustering, which resulted in figure 7.4. In this figure, each node represents an author keyword, the size of the node is correlated with how many times the keyword appears in the dataset, connections between nodes represent co-occurrences of those keywords, and the color scale represents the average year of the keyword in the literature.

Here, one can notice the following point. Keywords such as *quantum algorithms*, *quantum computation*, and *quantum computer* are mainly the keywords utilized by older literature. In recent years they have been replaced with more field specific terms sch as *quantum chemistry*, *variational quantum eigensolver,* and *quantum image processing*. This indicates an evolution of the literature into partially distinct lines of research, which are more developed topics in terms of maturity, compared to earlier keywords utilized in the literature.

We ran a similar analysis for the collaboration between countries (figure 7.5). The country-level data are associated with the affiliations of authors, and this map represents only the academic literature created by cross-country collaboration. This visual reveals that, although countries such as England, Germany, China, and the US are located in the centre of the collaboration network, a considerable number of new countries joined this network in recent years (represented by dark purple in the figure).

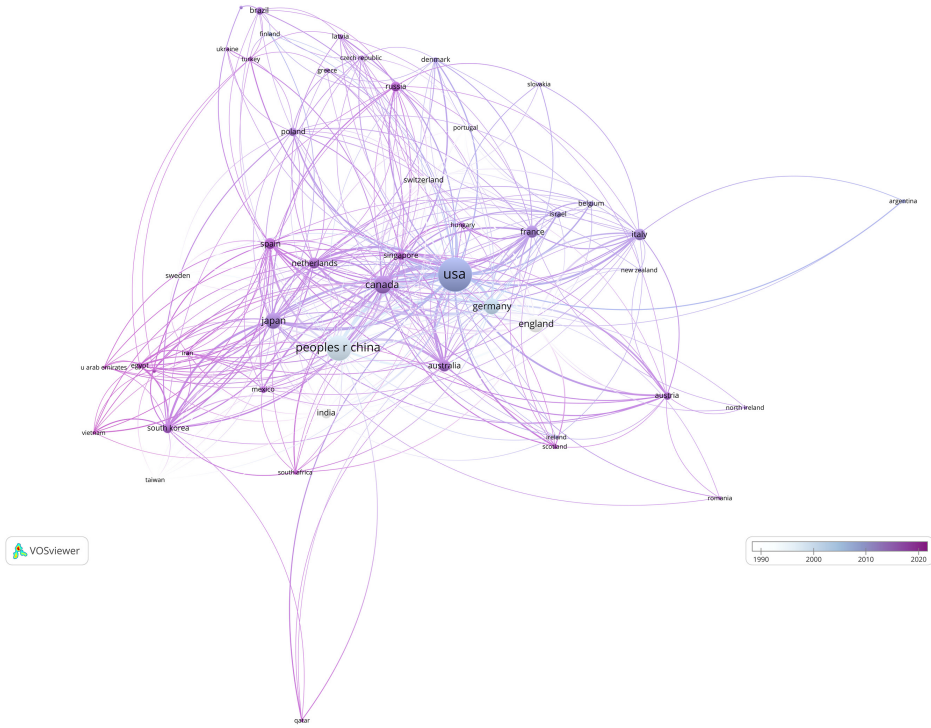

Figure 7.5. Overlay visualization of countries of origin of academic articles in time.

In terms of numbers, the patent literature reveals results similar to those of academic publications. We see a rise in the early 2000s in the top left panel in figure 7.6. We again see a jump before 2020. That sharp jump is the turning point we are discussing. The new and total numbers reflect this. We see similar jumps in the AI fields at a much larger scale (bottom two panels). When we compare the trends in figure 7.6 with those in figure 7.3, it is clear that scientific interest in both fields has been steadier than commercial interest until around 2016–17, and then both show a strongly upward trend.

This exploration of the academic and patent landscapes reveals several important insights into the current state of the field. First of all, comparison between AI and QC in terms of commercialization is clearly an overstatement in terms of patents. Using the queries given, we were able to find 122 609 patent families in the AI field versus 144 in quantum algorithms. Similarly, there were 343 808 articles in AI and 2951 in quantum algorithms. These represent differences of two to three orders of magnitude in publications and patents between these fields. In this sense, quantum algorithms can be related to the early 1990s when compared to AI, which might provide some insight into how the algorithms and QC might evolve in the coming decades.

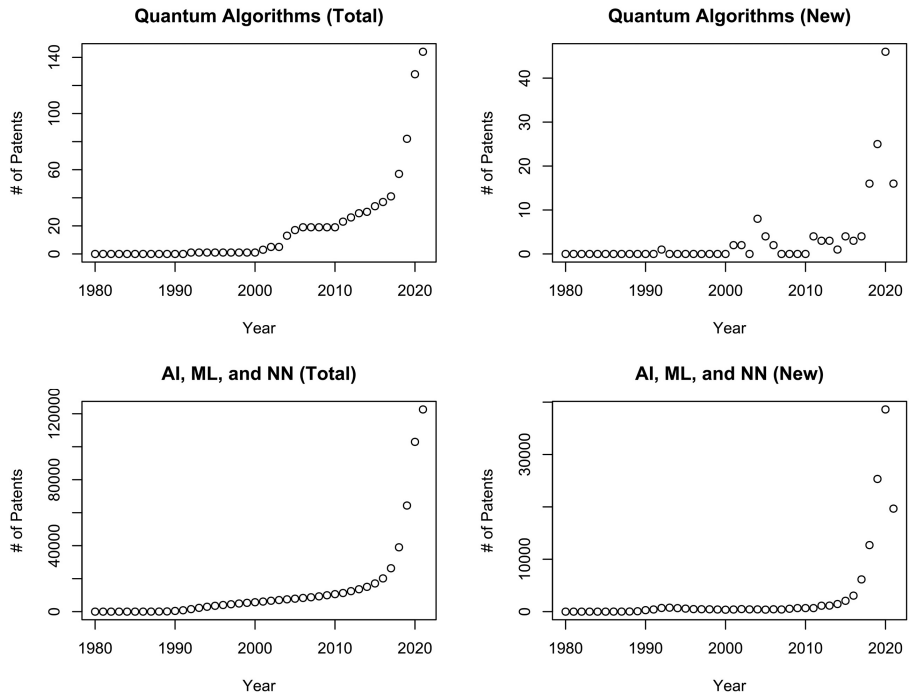

Figure 7.6. The numbers of new and total patents in the quantum algorithm and artificial intelligence fields.

Finally, to see how this activity has been translating in the entrepreneurial realm, we gathered a list of companies[2] in QT and identified the startups related to QC in them to compare them in figure 7.7. It should be noted that other fields under QT (such as quantum sensing and quantum communication/cryptography) have also been gaining popularity in the recent years. From the figure, we can see that startups in QC are a relatively new phenomenon compared to companies in QT, but both have been rapidly increasing in number during the last five years.

In summary, some historical differences and similarities between these fields can be seen in figures 7.3 and 7.6. One clear difference is of the scales, as there are orders of magnitude between the fields. The second difference is that since early 1990 there has been a steady increase in the total number of patents obtained in the field of AI compared to the almost zero activity in the field of quantum algorithms except a brief and short-lived interest in the mid-2000s. One clear similarity is the sudden spike in the late 2010s, especially after 2016–17, which is also evident in figure 7.7. This can be attributed to long-promised funding materialising, which prompted public and private investments around the world. The origins of some of

[2] This list was collected by us manually and contained 439 companies as of June 2021. We used open access resources such as Crunchbase, LinkedIn, The Quantum Insider, and Quantum Computing Report.

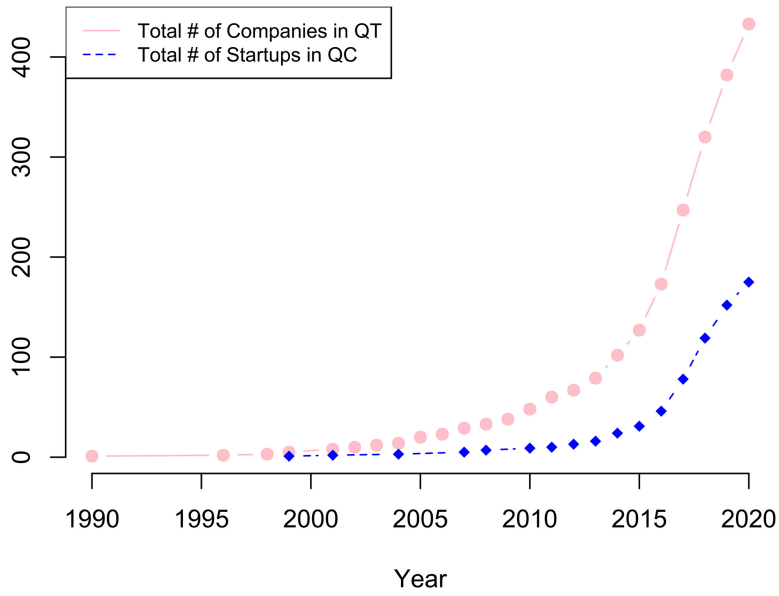

Figure 7.7. Total number of companies operating in quantum technologies and startups in quantum computing in years.

the public funding schemes can be attributed to fear of missing out for countries with existing scientific investments in the field (figure 7.5). Some can be explained by public demonstrations of IBM's and Google's superconducting quantum processors, signaling to the public (and investors) that QC is coming into the realm of calculable risk from Knightian uncertainty [90]. Regardless of the specifics, it is clear that both the number of new academic articles and new patents per year and the number of startups in QC have increased significantly, starting from 2016 to 2017.

This inflection point has not gone unnoticed in the ethical, legal, and social aspects field as well, and one of the first special issues on QT titled *The Societal Impact of the Emerging Quantum Technologies* was published in the journal *Ethics and Information Technology*[3]. Topics such as access to QT [91], the impact of QC on the future of scientific computing [92], responsible research and innovation in QT [93], and the potential impact of quantum computers on society [94] were discussed in the special issue. Since 2017, there has been growing interest in the societal impact of QT (and QC in particular). As of 2021, many researchers and commercial actors in the field have been calling the community to action regarding quantum ethics, which is a clear indication that there is a strong belief in the community that this technology will play a huge role in our future and should be developed ethically to avoid any undesired consequences while taking this quantum leap.

[3] ISSN 1388-1957 (print) 1572-8439 (web).

7.2.2.5 What is the next turning point in the development of quantum computing? It is hard to predict technology. We can assume that everything works and imagine best-case scenarios. We can assume that nothing works and imagine a sort of quantum winter. In reality, it is likely the case that the changes ahead are impossible to envision today.

QC violates no known laws of physics. It, therefore, is thought of as perhaps the world's most challenging engineering problem. So when will we engineer such devices? Currently, we have seen progress so dramatic that it would be impossible for those working in the field to have imagined even five years ago. Perhaps this means that the state of QC even five years from today is difficult to predict. We still believe that five years from now, quantum computers will be in the early stages of development and still lack error correction and other essential features to realize all their potential.

Having said that, we do think that a quantum future is inevitable. This is the natural progression of technology. This is probably the ultimate limit of computers, perhaps until we can harness new powers in the cosmos. Humankind's trajectory is set on a quantum course. Companies such as Google, IBM, and the like are in this for research and long-term prospects.

When will we see another inflection point? It's hard to tell. The saying goes that knowledge begets knowledge, so development always seems to go increasingly fast. But the next jump might have to wait until practical problems of commercial value are sufficiently solved. This should take place perhaps around 2050. We do imagine that by then, this technology will already have changed the world in ways we cannot predict now.

7.3 Sensors and their applications

Antigone Marino[1]
[1]Institute of Applied Sciences and Intelligent Systems, National Research Council (CNR), Naples, Italy

The etymology of the word *sensor* already explains its meaning in a simple way. The sensor is a device that senses. The human body is one of the most incredible devices that nature has created; it is composed of various sensors that allow us to manage the five senses, namely, sight, smell, touch, taste, and hearing. Scientific research and the development of technology have allowed the human race to develop a huge number of sensors, capable of 'feeling' where humans do not have the right sensors to do so. Sensors are among the technologies that will enable us to address many of the global challenges of the 21st century and beyond. Let's think, just to give a few examples, of food conservation, monitoring of transport, air, water, and health. The first sensors are more than a century old. However, smart sensors, with integrated Information and communications technology (ICT) capabilities, have been around for only a few decades. The 20th century saw the birth of a wide range of sensors, but it will be in the 21st century that their application, driven by the union of sensing and ICT, will affect many aspects of our life.

Due to a scholastic legacy, we are used to talking about one industrial revolution, which refers to the industrialization process of turning society from agricultural, artisanal, and commercial into a modern industrial system characterized by the generalized use of machines powered by mechanical energy and the use of new energy sources. The industrial revolution has never stopped. Since it started, there have already been four industrial revolutions. The first mainly concerned the textile and metallurgical sectors. The second industrial revolution is conventionally said to have started in 1870 with the introduction of electricity, chemicals, and fuels. Since 1970, the massive introduction of electronics, telecommunications, and information technology into industry identifies the third industrial revolution. Finally, the convergence of multiple technologies, real-time analytics, machine learning, ubiquitous computing, and embedded systems gave birth to the IoT. In parallel, the automation and control of industrial processes led to the fourth industrial revolution, also known as the Industrial Internet of Things (IIoT) (figure 7.8).

The history of sensors has gone hand in hand with that of industrial revolutions. Similar to the four phases of industrial development, we have four phases of the evolution of sensors (figure 7.9). In 1844 the French physicist Lucien Vidie invented the barograph, a device to monitor pressure, a recording aneroid barometer. This was indeed a sensor, but not a modern one, as it was a purely mechanical indicator not equipped with any electronics. The thermostat patented by Warren Seymour Johnson in 1883 [95] is considered by some to be the first modern sensor. The introduction of integrated circuits, capable of detecting a specific physical parameter and converting it into an electrical signal, led at the beginning of the 20th century to the birth of 'Sensors 2.0'. Since 1970, electronic devices have been used to measure physical quantities such as temperature, pressure, or loudness and convert them into electronic

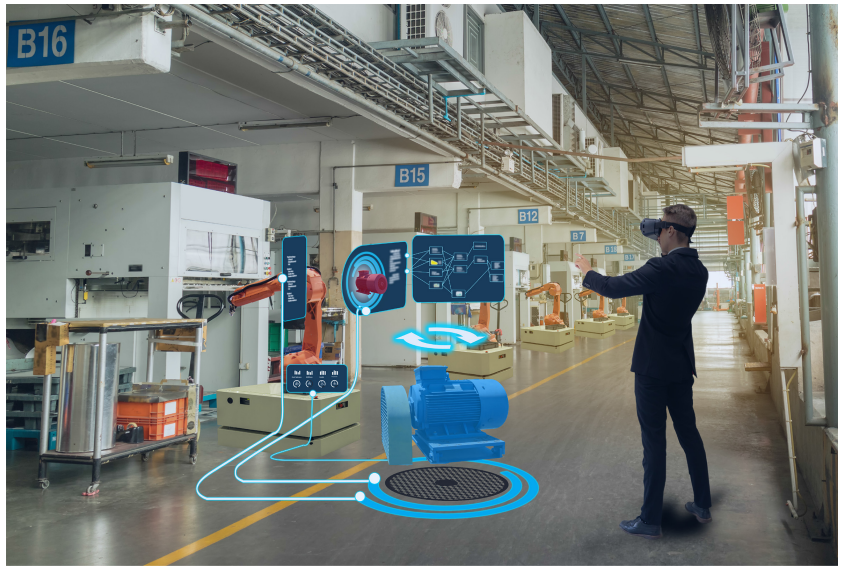

Figure 7.8. The Industrial Internet of Things has led to the development of Industry 4.0. Augmented reality and smart sensors will allow greater control of production processes.

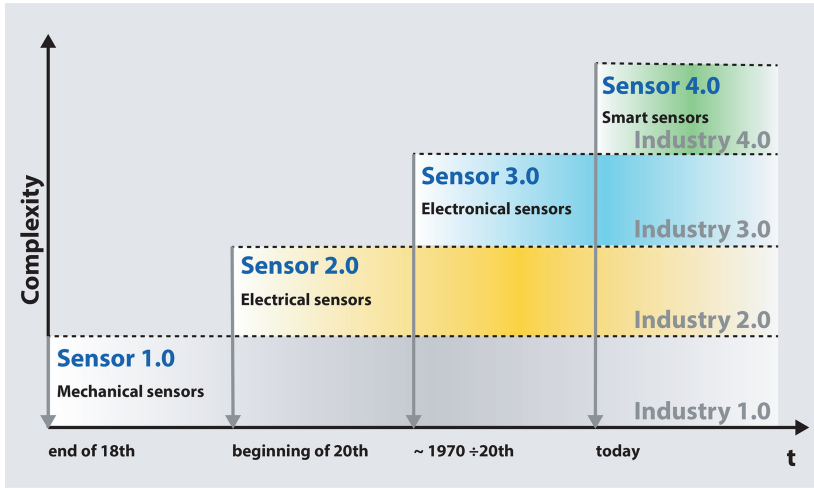

Figure 7.9. From the end of the 18th century up to today, scientific research has allowed the development of four sensor technologies, each one strongly connected to one of the four industrial revolutions: Sensors 1.0, intended as mechanical sensors, without electrical output; Sensors 2.0, thus electrical sensors; Sensors 3.0 or electronical sensor; and finally Sensors 4.0, also known as smart sensors.

signals, leading to the family of 'Sensors 3.0'. Finally, thanks to smart sensors, everything is getting clever, and the data collected during the manufacturing process are used to improve the quality of the product itself. We are now in the fourth industrial revolution thanks to 'Sensors 4.0' [96]. The importance of sensors is evident, since the industrial evolution is strongly connected to sensors and instrumentation.

In this section, we will look at how far scientific research has gone in the field of sensors and what the growth prospects are for the future.

7.3.1 Sensor characteristics

The operation process of a sensor is simple: The sensor detects a physical, chemical, or biological quantity (e.g., bacteria, proteins, waves, movement, or chemical agents). Then the measurement is processed by a transducer, which converts it into an output signal, often an electrical one (figure 7.10).

Each sensor has its own characteristics which outline its performance. These are categorized as systematic, statistical, or dynamic. Systematic characteristics are those which can be exactly quantified by mathematical means. Statistical characteristics are those which cannot be exactly quantified. Dynamic characteristics are those who describe the ways in which an element responds to sudden input changes [97].

The sensor range is a static characteristic that describes both the minimum and maximum values of the input or output [98]. The full-scale input, called span, describes the maximum and minimum input values that can be applied to a sensor without causing an unacceptable level of inaccuracy. It is also called the dynamic range. If we speak of the sensor output, it is the algebraic difference between the output signals measured at maximum and minimum input stimulus. Finally, the operating voltage range describes the minimum and maximum input voltages that can be used to operate a sensor.

The sensor transfer function describes the relationship between the measurand and the electrical output signal. If it is time independent, we can write it as $S = F(x)$, where x is the measured quantity and S is the electrical signal produced by the sensor. This function can be a very complex one. The simplest case is the one of a linear transfer function, $S = A + Bx$, where A is the sensor offset and B is the sensor slope. The sensor offset is the output value of the sensor when the input is zero. The slope of a linear transfer function is the sensor's sensitivity, which we will define shortly. Many sensors do not have a linear response; rather, it is approximated to be

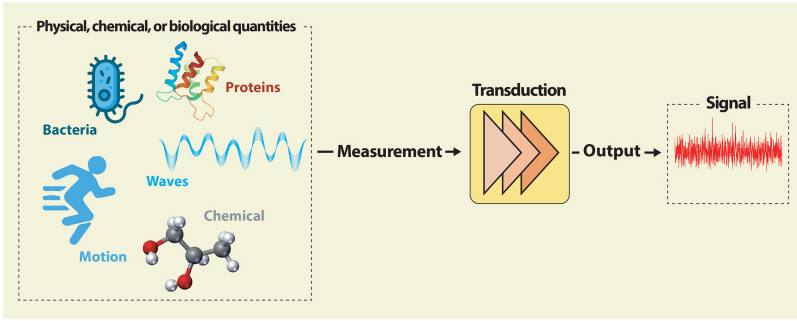

Figure 7.10. A sensor is a device that detects a physical quantity and converts its measurements into a human-readable output, such as an electric signal.

linear within certain limits. As you can imagine, having a linear or linearized function offers many advantages, such as being able to calculate the quantity we are analyzing from the output, to predict the output based on the analyzed quantity value, or even to easily obtain the offset and sensitivity. In modern sensors, such as smart sensors, the presence of the digitization of the output makes its linearization considerably easier. In the case of nonlinear transfer functions, nonlinearity is defined as the difference between the real line and the ideal straight one. As the nonlinearity can vary along the input–output graph, it is customary to indicate as characteristic of the sensor the maximum nonlinearity. Typically, this value is expressed in percentage of the span.

The sensitivity S of the sensor is the ratio between the variation of the output quantity ΔS and that of the input Δx which determined it:

$$S = \frac{\Delta S}{\Delta x} \tag{7.4}$$

A sensor is very sensitive when a small variation of the input quantity corresponds to a large variation of the output quantity. If the sensor transfer function is linear, the sensitivity will be constant over all of the sensor range. Otherwise, it will change. The range of values for which a sensor does not respond is called the dead band. In this range the sensitivity is zero. Dead band is also usually expressed as a percentage of the span. The sensor saturation point is the input value at which no more changes in the output can occur.

A sensor should be capable of following the changes of the input parameter regardless the direction in which the change is made. However, this does not always happen. The output of a sensor may be different for a given input, depending on whether the input is increasing or decreasing. The hysteresis is the measure of this property, it quantifies the presence of a 'memory' effect of the sensor whose output, with the same measurand value, could be affected by the previous operating condition. Like nonlinearity, hysteresis varies along the input–output plot; thus, maximum hysteresis is used to describe the characteristic. This value is usually expressed as a percentage of the sensor span.

The sensor resolution is the smallest change that can be detected in the quantity that is being measured. It is one of the features that most affects the cost of a sensor. Precisely for this reason it is important to understand, based on the application to be implemented, what resolution is needed. A sensor with a low resolution could cause the fail of detecting the signal, while a resolution that is too high could be unnecessarily expensive.

The sensor accuracy represents the maximum error between the real and ideal output signals. It is the sensor's ability to provide an output close to the real value of the analyzed quantity. It can be quantified as a percentage relative error using the following equation:

$$\text{Percentage Relative Error} = \frac{\text{Measured Value} - \text{True Value}}{\text{True Value}} \tag{7.5}$$

The concept of precision refers to the degree of reproducibility of a measurement. In other words, if exactly the same value were measured a number of times, an ideal sensor would output exactly the same value every time. But real sensors output a range of values distributed in some manner relative to the actual correct value. As precision relates to the reproducibility of a measure, it can be quantified as percentage standard deviation using the following equation:

$$\text{Percentage Standard Deviation} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100 \qquad (7.6)$$

Precision is often confused with accuracy. Figure 7.11 illustrates the key difference. If we repeat a measurement five times, the distribution of our data will change depending on the sensor accuracy and precision [99]. Obviously, the best sensor is the one that has greater accuracy and precision.

Like any measuring device, sensors are subject to measurement error, that is, the difference between the measured value and the true one. The errors can be either systematic or random. Systematic error always affects measurements by the same amount or the same proportion, provided that a reading is taken the same way each time. Systematic error is predictable. Random error causes one measurement to differ slightly from the next. It comes from unpredictable changes. Random errors cannot be eliminated, while most systematic errors may be reduced with compensation methods, such as feedback, filtering, and calibration [98]. Systematic errors result from a variety of factors: interfering inputs, modifying inputs, changes in

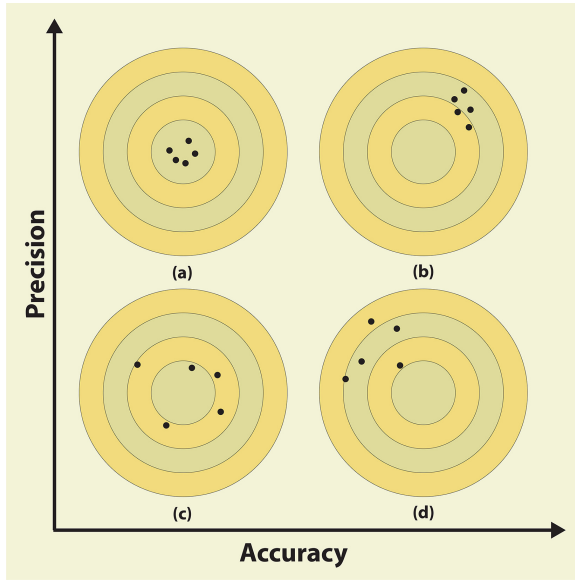

Figure 7.11. Repeatability and accuracy. If we repeat a measurement five times, the distribution of our data will change depending on the sensor accuracy and precision: (a) good accuracy and precision; (b) low accuracy and good precision; (c) good accuracy and low precision; (d) low accuracy and precision.

chemical structure or mechanical stresses, interference, signal attenuation or loss, and even human factors. Random error, also called noise, is a signal component that carries no information. The quality of a signal is expressed quantitatively as the signal-to-noise ratio, which is the ratio of the true signal amplitude to the standard deviation of the noise.

Sensors' dynamic characteristics are those that are time-dependent. These are relevant when the sensor inputs are not constant over time. The most common dynamic characteristics are response time and dynamic linearity. Sensors do not change output state immediately when an input parameter change occurs. Rather, the output will change to the new state over a period of time, called the response time. This can be defined as the time required for a sensor output to change from its previous state to a final settled value within a tolerance band of the correct new value. The tolerance band is defined based on the sensor type, the sensor application, or the preferences of the sensor designer. The dynamic linearity of the sensor is a measure of its ability to follow rapid changes in the input parameter.

7.3.2 General overview of sensors

If we stop for a moment, and observe the world around us, we will immediately realize that in our daily life we use much more sensors than we imagine. It is not trivial to classify them, or to review them all. Conventionally, they can be classified according to the measurement principle they use.

7.3.2.1 Mechanical sensors

Mechanical sensors form a class of sensors that are sensitive to changes in mechanical properties. They measure a physical quantity resulting from a stimulus that causes mechanical deformation of the sample [100] and translate it into an electric signal. The mechanical deformation can be realized with different stimuli.

The most common mechanical sensor is the strain gauge, invented by Edward E Simmons [101] and Arthur C Ruge in 1938 [102], which is used to measure strain on an object. Strain gauges are the key sensing element in a variety of sensors types, including pressure sensors, load cells, torque sensors, and position sensors. The most common type of strain gauge consists of a grid of very thin metal wire rigidly applied to a plastic material support. The gauge is attached to the object whose strain gauge is to be measured by a suitable adhesive, such as cyanoacrylate. The strain gauge wire follows the deformations of the surface to which it is glued, elongating and shortening together with it. These dimensional variations cause fluctuations of the electric wire resistance. By measuring them through a Wheatstone bridge, it is possible to trace the deformation that caused them. A key problem with strain measurements is that of thermal effects. The electric current flowing through the strain gauge causes heating by the Joule effect. Temperature compensation is required to address the problem, using a Wheatstone bridge connected to two strain gauges: the one for measurement on one side and another equal strain gauge on a piece of the same material not subjected to any stress but exposed to the same

temperature. Due to the characteristics of the Wheatstone bridge, the deformations due to temperature produce the same variation in resistance in modulus but with a different sign so as to cancel each other out. It is not always possible to know in advance the direction in which the deformation of the material will occur. In this case, it becomes necessary to use a system of strain gauges with axes oriented in different directions.

Piezoelectric sensors are a particular type of strain gauges [103]. They differ slightly in terms of physical operating principle. In fact, in piezoelectric materials the resistance change is due to resistivity, while in strain gauges the resistance varies almost exclusively because of the length and the section of the conductor that makes up the sensor. Thus, piezoelectric strain gauges have even greater sensitivity to temperature variation.

Sensors based on both mechanical and electrical variations are known as micro electromechanical systems (MEMS) [98]. These devices have been recognized as one of the promising technologies of the 21st century, capable of revolutionizing both the industrial world and that of consumer products. They are three-dimensional, miniaturized, mechanical and electrical structures, typically ranging from 1 to 100 μm, that are manufactured using standard semiconductor manufacturing techniques. MEMS sensors are widely used in the car industry and, since the early 1990s, to realize accelerometers in airbag systems, electronic stability programs, and antilock braking systems. The availability of inexpensive, ultracompact, low-power multiaxis MEMS sensors has led to rapid growth of their use in customer electronics devices. MEMS can be found in smartphones, tablets, game console controllers, portable gaming devices, digital cameras, and camcorders. They have also found application in the healthcare domain in devices such as blood pressure monitors, pacemakers, ventilators, and respirators. Two of the most important and widely used forms of MEMS are accelerometers and gyroscopes [104].

An accelerometer is a sensor capable of measuring acceleration. The use of the accelerometer has increased considerably in recent years since, alongside traditional applications in the scientific and aerospace fields, it has been adopted in numerous civil fields such as the automotive and consumer electronics industries. In most accelerometers, the operating principle is the same: It is based on detecting the inertia of a mass when subjected to acceleration. The mass is suspended from an elastic element, while some type of sensor detects its displacement with respect to the fixed structure of the device. In the presence of an acceleration the mass, which has its own inertia, moves from its rest position in proportion to the detected acceleration. The sensor transforms this displacement into an electrical signal.

MEMS are also used in gyroscopes, sensors that are able to measure the angular rate of rotation of one or more axes. As they have no rotating parts, they can be easily integrated into even very small devices. They use vibrating mechanical elements to detect rotation based on the transfer of energy between two modes of vibration caused by Coriolis acceleration.

7.3.2.2 Optical sensors

Optical sensors work by detecting light, from ultraviolet to infrared (IR). They can be designed in various ways in order to acquire the signal [105]: They can measure the light intensity change related to light emission or absorption by a quantity of interest, they can measure phase changes occurring in light beams due to interaction or interference effects, or they can measure a simple interruption of a light source. Let us now look at the most common types of optical sensors.

Photodetector sensors measure the sample photoconductivity, or the material change of conductivity when illuminated. There are many types of photodetectors, such as active pixel sensors, charge-coupled devices (CCD), light-dependent resistors, photodiodes, and phototransistors [106].

IR sensors measure and detect IR radiation in their surrounding environment [107]. There are two types of IR sensors: active and passive. Active IR sensors both emit and detect IR radiation. They are realized with a light-emitting diode (LED) and a receiver (photoelectric cells, photodiodes, or phototransistors). When an object comes close to the sensor, the IR light from the LED reflects off the object and is detected by the receiver. Active IR sensors act as proximity sensors, and they are commonly used in obstacle detection systems. Passive IR sensors detect only IR radiation. Inside, there is one or more pyroelectric materials that generate energy when exposed to heat. They are commonly used in motion-based detection, such as in-home security systems. When a moving object that generates IR radiation enters the sensing range of the detector, the difference in IR levels is measured.

Optical fiber sensors use an optical fiber as the sensing element. They are widely used in construction, such as bridge monitoring [108], and aircraft security [109], as they allow monitoring of different physical quantities. Strain can be measured as it changes the geometric properties of the fiber and so the refraction of the light passing through it; in the same way, a temperature change can be detected as it causes fiber strain; pressure sensing can be realized with an intensity sensor [110] or an interferometric sensor [111]; humidity sensing can be achieved in several ways, such as luminescent systems with fluorescent dyes that are humidity sensitive or reflective thin film-coated fibers which change their refractive index, resulting in a shift in resonance frequency [105].

Finally, interferometric sensors measure changes in a propagating light beam, such as path length or wavelength along the path of propagation [112].

Optical sensors are widespread, due to the many advantages they show: high sensitivity, high integration, suitability for remote sensing, wide dynamic range, wide range of chemical and physical parameters detection, and many others. However, they can be costly and susceptible to physical damage.

7.3.2.3 Semiconductor sensors

Semiconductor sensors owe their popularity to their low cost, high integrability, and long life. They are mostly used for gas monitoring, but they also allow the detection of temperature or optical physical quantities, as in CCDs [98].

In gas sensors, the operation of a semiconductor sensor is determined by the variation in conductivity of a semiconductor element caused by the chemical

absorption of the gases in contact with the porous surface of the semiconductor, electrically heated to a predetermined temperature. The temperature of the sensitive element (depending on the type of gas to be detected) is a determining parameter for the sensitivity and selectivity of the sector.

Semiconductor temperature sensors commonly use a bandgap element which measures variations in the forward voltage of a diode to determine temperature. They are designed with materials showing a strong thermal dependence. To achieve reasonable accuracy, these are calibrated at a single temperature point. Therefore, the highest accuracy is achieved at the calibration point, and accuracy then deteriorates for higher or lower temperatures. For higher accuracy across a wide temperature range, additional calibration points or advanced signal processing techniques can be employed.

Semiconductor magnetic field sensors exploit the galvanomagnetic effects due to the Lorentz force on charge carriers. When the sensor is placed inside a magnetic field, the voltage difference of the semiconductor depends on the intensity of the magnetic field applied perpendicular to the direction of the current flow. Electrons moving through the magnetic field are subjected to the Lorentz force at right angles to the direction of motion and the direction of the field. The Hall voltage, which is generated in response to the Lorentz force on the electrons, is directly proportional to the strength of the magnetic field passing through the semiconductor material. The output voltage is often relatively small and requires amplification of the signal. Integrated semiconductor magnetic field sensors are manufactured using integrated circuit technologies [113].

The most common optical semiconductor sensor is the photodiode, a photodetector that converts light into either current or voltage. The photodiode is a diode that works as an optical sensor by exploiting the photovoltaic effect. It is able to recognize a certain wavelength of the incident electromagnetic wave and to transform this event into an electrical signal of current by applying an appropriate electric potential to its ends. It is therefore a transducer from an optical signal to an electrical signal. Another form of photodetector is the phototransistor, which is essentially a bipolar transistor with a transparent window that allows light to hit the base–collector junction. Phototransistors have the advantage of being more sensitive than photodiodes. However, they have a slower response time.

7.3.2.4 Electrochemical sensor

Electrochemical sensors are devices that give information about the composition of a system in real time by coupling a chemically selective layer to an electrochemical transducer. In this way, the chemical energy of the selective interaction between the chemical species and the sensor is transduced into an analytically useful signal. Due to the simplicity of the procedures and instrumentation required, they are the largest and the oldest group of chemical sensors. They attract great interest nowadays because they are easy to miniaturize and integrate into automatic systems without compromising analytical characteristics.

Different families of electrochemical sensors can be recognized depending on the electrical magnitude used for transduction of the recognition event [114]:

potentiometric (change of membrane potential) [115], conductometric (change of conductance), impedimetric (change of impedance), voltammetric (change of current for an electrochemical reaction with the applied voltage), and amperometric (change of current for an electrochemical reaction with time at a fixed applied potential) [116].

Electrochemical sensors have a number of advantages, including low power consumption, high sensitivity, good accuracy, and resistance to surface-poisoning effects. However, their sensitivity, selectivity, and stability are highly influenced by environmental conditions, particularly temperature.

7.3.2.5 Biosensors

A biosensor is a device consisting of a biologically active sensitive element and an electronic part. The operating principle is simple: The biological element interacts with the substrate to be analyzed, and a transduction system converts the biochemical response into an electrical signal. A common example of a biosensor is the glucometer used by diabetics to measure the concentration of glucose in the blood. The sensitive element of this biosensor is an enzyme, glucose oxidase, which converts glucose into gluconic acid.

The analyte is a substance that we intend to measure, such as glucose, cardiac biomarkers, or tumor biomarkers. The bioreceptor is the sensitive element, it can be an enzyme (catalytic biosensor), an antibody (affinity biosensor), DNA/RNA, or an aptameter. The transducer can be electrochemical (potentiometric, amperometric), optical, electromechanical, mechanical, or even acoustic [117].

The main feature of a biosensor is the specificity, which is guaranteed by the use of biological receptors, which by their intrinsic nature are specific to particular analytes. Specificity is the ability to react only with a certain analyte and not with others that may be present in the measurement environment. In other words, the other analytes present in the measurement environment are influencing quantities with a negligible effect. The main characteristics of the biosensors are their high sensitivity, measurement speed and cost-effectiveness.

7.3.3 Sensors challenges

In 2015 the United Nations General Assembly set up a collection of 17 interlinked global goals designed to be a 'blueprint to achieve a better and more sustainable future for all'. These are called Sustainable Development Goals (SDGs). They are included in a UN Resolution called the 2030 Agenda [118], as they are intended to be achieved by the year 2030. Among the SDGs are Good Health and Well-being, Clean Water and Sanitation, Affordable and Clean Energy, Industry Innovation and Infrastructure, Sustainable Cities and Communities, Responsible Consumption and Production, Climate Action, and Life Below Water, Life On Land.

In 2020, the European Commission (EC) signed The European Green Deal [119], a set of policy initiatives with the overarching aim of making Europe climate neutral in 2050. The plan is to review each existing law on its climate merits and introduce new legislation on the circular economy, building renovation, biodiversity, farming,

and innovation. The climate change strategy is focused on a promise to make Europe a net-zero emitter of greenhouse gases by 2050 and to demonstrate that economies will develop without increasing resource usage. Also, the Green Deal has measures to ensure that nations that are reliant on fossil fuels are not left behind in the transition to renewable energy.

These community actions, at both global and European levels, require the support of scientific research in order to achieve the set goals. In the last century, scientific research has allowed the development of many types of sensors. Basical,ly more than one type of sensor is available to measure a quantity of interest. Of course, each sensor option has its advantages and disadvantages. These need to be weighed against the context in which the sensor will be used in order to determine which sensor technology is the most suitable one.

Let's examine some of the main challenges that our society is facing and how sensors can be one of the necessary tools to address them.

7.3.3.1 Environmental and earth sensing

Intensive agriculture, growing urbanization, industrialization, the increasing demand for energy, and climate change are putting planet Earth at risk. Environmental monitoring is therefore essential to reveal its state, to assess the progress that has been made to achieve certain environmental objectives, and to help detect new environmental problems.

The Joint Research Centre (JRC) is the EU's science and knowledge service which employs scientists to carry out research in order to provide independent scientific advice and support to EU policy. The results are of fundamental importance to environmental management in general, as the drafting and prioritisation of environmental policies is based on the findings of environmental monitoring. The JRC's work supports many actions, such as Copernicus (the European Earth Observation Programme), the Water Framework Directive, the Marine Strategy Framework Directive, EU Food Security Policy, European Climate Policy, EU Strategy for Sustainable Development, the Directive on Ambient Air Quality, and the Clean Air For Europe programme.

The protection of the human race and the Earth's ecosystem is the highest priority. Scientific research is trying to create scalable sensors capable of detecting pollutants, even at very low concentrations and on a widespread basis, quickly and sensitively, While many environmental monitoring solutions are already available, what sensors are being asked for is to work *in situ* and online, allowing real-time decision making. Many analyzes, such as those for bacterial contamination, still require *in situ* sampling with laboratory analyzes, which also entails a slowdown due to bureaucracy. Sensors that operate *in situ* are the new challenge for environmental monitoring, organized in a distributed network that can have an ever greater geographical distribution.

Power plants, agriculture, industrial manufacturing, and vehicle emissions are all sources of air pollutants such as sulfur dioxide, carbon monoxide, nitrogen dioxide, and benzene. Air pollution is a problem in both the developed world and the developing world. Energy production linked to fossil fuels is the main source of air

Figure 7.12. Agriculture will make use of smart sensors to remotely control the parameters that regulate crop growth, such as humidity and temperature.

pollution. The second source is urbanization and the consequent increase in vehicular emissions.

Air monitoring stations are now equipped with multiple sensors so that not only one gas can be analyzed, but also particulates, hydrocarbons, and metals, according to local regulatory requirements for air quality [120]. Air monitoring requires very sophisticated and therefore expensive instrumentation, such as spectroscopic analysis. This explains why air monitoring has a low distribution density in many countries. In the future, we will see the production of low-cost and smart sensors, which may require the most widespread and real-time air analysis (figure 7.12).

Several companies are studying the possibility of creating devices designed to transform smartphones into an environmental monitoring centre capable of recording levels of environmental humidity, carbon monoxide, fine dust, light intensity, pressure, and many other factors [121]. The spread of these integrated sensors would be a revolution in the field of air monitoring, as each citizen would constitute a data collection unit.

Regarding air monitoring, the JRC supports many actions, such as the Climate Service of Copernicus, the European Earth Observation Program relating to climate change, which among its purposes has not only monitoring, but also ensuring compliance with international standards. The JRC also maintains a worldwide database, the Emissions Database for Global Atmospheric Research (EDGAR), which allows monitoring and verification of emissions around the world, providing the information needed to determine appropriate policies.

The growing need for clean water for drinking and industrial use requires water quality monitoring. Similar to air quality, strict standards are set by national bodies and geopolitical bodies, resulting in the need for reliable sensor technologies that can monitor different water quality parameters with the required sensitivity. Again, the market requires sensors that provide real-time readings to ensure that any abnormal

changes in water quality have the least impact on human health or manufacturing operations [122].

There are three main categories of interest: physical (turbidity, temperature, conductivity), chemical (pH, dissolved oxygen, concentration of metals, nitrates, organic substances), and biological (biological oxygen demand, bacterial content). Scientific research is developing smart water networks and predictive models [123]. Sensors implemented directly in the water distribution network will contribute to the identification of leaks through pressure drops, thus allowing minimization of water losses. The development of predictive models for water quality monitoring, based on the fusion of water quality data, environmental sensors, and metrology sensors, will allow to predict potential changes in future water quality.

The JRC's work on water monitoring covers the monitoring of water quality and assessment of the impact of pollutants and chemicals, the monitoring of water and marine ecosystems, the provision of early warnings and risk management, the monitoring of floods and droughts, and the monitoring of water quantity in Europe and worldwide.

7.3.3.2 Global navigation satellite systems

The development of satellite technologies has made possible one of the most fascinating and promising projects in scientific research. The Global Navigation Satellite System (GNSS) is a constellation of satellites providing signals from space that transmit positioning and timing data to GNSS receivers [124]. The receivers then use the data to determine location. This network provides global coverage. Examples of GNSS include Europe's Galileo and the USA's NAVSTAR Global Positioning System (figure 7.13).

The satellite transmits a signal that contains the position of the satellite and the time of transmission of the signal itself, obtained from an atomic clock in order to

Figure 7.13. Satellites equipped with various sensors, organized in a sophisticated network, can transmit information about the planet and space in real time. This is the concept behind the Global Navigation Satellite System.

maintain synchronization with the other satellites in the constellation. The receiver compares the transmission time with that measured by its own internal clock, thus obtaining the time it takes for the signal to arrive from the satellite. Several measurements can be made simultaneously with different satellites, thus obtaining the positioning in real time. Each distance measurement, regardless of the system used, identifies a sphere that has a satellite as its centre; positioning is obtained from the intersection of these spheres.

GNSS systems have many applications: navigation, both with portable receivers, for example for trekking, and with devices integrated between the controls of means of transport, such as cars, trucks, ships, and aircraft; time synchronization in electronic devices; monitoring; search and rescue service; geophysics; topographic applications; machine automation applications (automatic driving of machines) for earthmoving and agriculture; wild animal tracking devices; and satellite alarm.

GNSS systems are also being studied to face a possible Kessler effect in the future. This is a scenario proposed by NASA scientist Donald J Kessler in 1978 [125]. It is a theoretical scenario in which the density of objects in low Earth orbit, due to space pollution, is high enough that collisions between objects could cause a waterfall where each collision generates space debris, which increases the likelihood of further collisions. The direct consequence of this scenario is that the increasing amount of waste in orbit would make space exploration and even the use of artificial satellites impossible for many generations. A science-fiction version of Kessler's syndrome is depicted in the 2013 movie *Gravity*, directed by Alfonso Cuarón. He imagines that the reckless downing of a spy satellite with a missile causes a chain reaction that ends in the destruction of a space shuttle and the death of its crew, the Hubble Space Telescope, the International Space Station (ISS), and the Chinese space station. The film focuses on the fate of the protagonist and does not analyze the consequences for future space travel.

The problem of space waste is very difficult to solve directly, since the small size and high speeds that characterize most of the waste make their recovery and disposal practically impossible. However, GNSS systems allow the monitoring of space waste and help institutions to eventually run emergency procedures.

7.3.3.3 Security sensors

Physical security and safety have always been critical for the welfare of individuals, families, businesses, and societies. In the past, castles were surrounded by water ditches and cities by thick walls to defend themselves from invasions.

The IoT has enabled growth in the residential and commercial security sectors, thanks to sensors that power these solutions. IR security sensors, active or passive, utilize IR light to detect motion in order to trigger an alarm. Photoelectric sensors are now more common in spy films than in real life, but you can find them at work in specific security settings. Photoelectric sensors would be helpful in, for example, an environment that contains a space that humans or objects may not enter. Also, photoelectric sensors use invisible IR light but can travel significantly farther than

IR sensors. Photoelectric beams establish an invisible barrier that, when broken, triggers a security notification. Microwave sensors are also used in order alarm systems. Like active IR sensors, microwave sensors emit and receive a signal to detect an object in motion. These sensors are generally much more sensitive than IR sensors. Moreover, they can sense motion through nonmetal materials such as wood, plastic, and drywall. One of the most recent technologies, which will see a wide diffusion in the coming years, is that of tomographic motion detection sensors. Motion tomography technology does not require a direct line of sight to trigger a safety alert. It uses a mesh network of radio emitters and receivers to detect any movement within the mesh network. This sensor technology works by detecting interruptions in signals between emitters and receivers, which it interprets as motion. The tomographic motion technology has been around for only about a decade, but it is a promising technology for the high-security commercial and industrial sectors.

Safety is a very important factor in various contexts, such as airports, transport control, detection of weapons or hazardous materials, drug detection, and even nuclear safety. In this type of application, real-time detection is required in order to be able to suddenly face dangerous situations. When very large control areas must be covered, it is necessary to use networks of smart sensors, which allow a capillary control of the situation. Sensors are an important part of the technology behind modern security systems. As processing power and software capabilities continue to increase, smart sensors will also keep pace in the security market.

7.3.3.4 Industrial IoT

Collecting and analyzing all data coming from production sensors is for the industrial market a new key to competitiveness. It means being able to analyze all the process variables, such as energy consumption, temperatures and pressures, speed, and all the measurable physical quantities within the production cycle. The IIoT is able to connect every process line, every production phase, and every single machine thanks to different types of sensors positioned at sensitive and critical points [126]. Thanks to the data generated by the objects and transmitted to the system, it is possible to obtain an accurate monitoring of the plant, the quality of the system, energy efficiency, and timely feedback criticalities. The management and integrated analysis of data coming from sensors allow successful management of real-time monitoring, alerting, energy management, quality prediction, and predictive maintenance.

7.3.4 Conclusions

Over the last decade, the term *smart sensor* has become increasingly widespread. This is the result of the development of new technologies capable of implementing communication and data processing, thanks to increasingly complex digital systems. A smart sensor is an intelligent sensor, which is a sensor that is not only capable of detecting electrical, physical, or chemical quantities, but also

capable of reprocessing the information that is collected and then transmitting it in the form of an external digital signal. A smart sensor includes an interface designed for communication and an analog-to-digital converter. These elements are essential for proper functioning.

IoT and IIoT seek to allow individuals to interact and connect with electronic tools and sensor networks. The IoT is considered a very important basis for monitoring all services related to the automation of processes as regards the sphere of waste reduction and environmental control and finally to expand the quality of life in the workplace. Conceived to be able to increase efficiency, smart sensors are designed to ensure four main features: Advanced sensitivity allows the various anomalies present during operation to be detected in a minimum time; smart tasks, on the other hand, have the function of processing data directly from the sensor, making data transmission fast and efficient; efficient communication allows a bidirectional data exchange between the sensor and the control unit; finally, diagnostics have the function of recognizing anomalies present in the system and at the same time carrying out preventive maintenance that bases interventions only on real needs and requirements. In this way, maintenance costs are halved. Errors can be identified very simply, thanks to the display modes that guarantee minimum resolution times.

In science fiction, the future of humanity has often been imagined with robots, and they will surely soon be part of our daily life, which is partially already happening. Obviously, cinema loves fiction, especially when it comes to a subject such as science. But even before robots, what has changed our lives are sensors, and in the next few years their use will become widespread in our lives. Almost all of us have occasionally checked the ambient temperature on a smartphone or used a device to check how many steps we have taken in a day. Cars available on the current market are equipped with sensors that allow us to check tire pressure without getting out of the vehicle. Medical devices allow us to monitor the oxygen level of the blood or blood pressure without having to go to a doctor. The use of sensors in medicine will allow human beings to do more prevention and to use telemedicine where health systems do not allow widespread coverage of the territory. That is why improving public understanding and perceptions of sensor science, through education and communication, is fundamental.

The seed technologies are now being developed for a long-term vision that includes smart sensors as self-monitoring, self-correcting and repairing, and self-modifying or evolving not unlike sentient beings. The ability for a system to see (photonic technology), feel (physical measurements), smell (electronic noses), hear (ultrasonics), think and communicate (smart electronics and wireless), and move (sensors integrated with actuators) is progressing rapidly and suggests an exciting future for sensors.

7.4 The space sector: current and future prospects

Javier Ventura-Traveset[1]
[1]European Space Agency, Centre Spatial de Toulouse, 18, Avenue Edouard Belin, 31401 Toulouse, France

In this section, we review the current status and future prospects of several key domains within the space sector and their associated technologies. These include space science (a golden age), human and robotic exploration, climate change and earth monitoring, and satellite navigation. We complete the section with a discussion of the problems and perspectives of space debris mitigation, the prospects of planetary defence, and the ongoing NewSpace revolution.

7.4.1 The space sector today: a quick balance

The space sector today is a mature and diversified sector without which the modern world, as we conceive it today, simply could not exist. As of 2022, and according to data from the UN Office for Outer Space Affairs, 10 countries (Europe grouped as one) have the capability to put satellites into orbit independently, and about 85 countries have (or have had) their own national satellites [127]. Currently, over 9000 operational satellites orbit our planet, and the expectation is that tens of thousands of new satellites will join them in just a few years, including the ongoing and planned megaconstellations for broadband Internet services, a revolution in the space telecommunications sector.

In 2020, the direct economic impact of the space sector was estimated at around $366 billion [128], having almost doubled in a period of 10 years. Interestingly, only around 25% of that corresponds to institutional expenditure, incurred by government agencies. The remaining 75% comes from the private sector; this is a fundamental paradigm shift experienced by the space sector in the last 30 years, with a major increasing trend, as we will discuss later.

Taken in a wider scope, space economy can be defined as 'the full range of activities and the use of resources that create and provide value and benefits to human beings in the course of exploring, understanding, managing and utilising space' [129]. If we consider this global definition, the figures are much higher and actually difficult to compute, given the high penetration of space activities in our society. For example, it is estimated that more than 10% of the gross domestic product (GDP) of the EU currently depends on the availability of satellite services, a figure that is far higher than the sector's direct turnover, and that a space blackout would imply a loss of between 500 000 and 1 000 000 jobs on the European continent [130].

A detailed analysis of the space economy also reveals an aspect of particularly high interest in this sector. Indeed, if we examine the overall space-related economic volume, we can conclude that the actual expenditure in manufacturing and launching satellites accounts for about only 6% of the total business generated by this sector. This actually means that investment in space infrastructures provokes a

major multiplication effect in the global space economy, about 15 on average. This is the case, for example, for the largest European space institutional systems, such as Galileo, Copernicus, and the meteorological Meteosat and Metop satellites, whose actual impact on the overall economy versus their development cost is even greater.

Beyond these remarkable commercial and social impacts, the contribution of space to the growth in our scientific knowledge has also been (and will continue to be) extraordinary. The data from our space science missions have contributed to remarkable breakthroughs during the last 50 years, notably in the fields of fundamental physics, solar system science, astrometry, astronomy, and astrophysics, contributing towards a much better understanding of our universe and of the fundamental physical laws governing it. Linked to space exploration, significant research advances have been attained in fields such as space medicine, space life sciences, biotechnology, space material science, microgravity fluid physics, radiation, space weather, aerodynamics, space geosciences, and so on. Moreover, linked to earth science research, we could mention the advances in geodesy, geophysics, volcanology, geochemistry, meteorology, and oceanography; the understanding of earth magnetic field dynamics, the lithosphere, and the Earth's interior; hydrology; biodiversity; and certainly, the understanding of trends in climate change trends, its sources, dynamics, and the major anthropogenic impacts.

In the following sections, we will briefly review the current status and future prospects of some of the main space sector domains, covering the topics of space science, human and robotic exploration, climate change and earth monitoring, and satellite navigation. We will complete this chapter by discussing the problematic and future prospects of space debris mitigation, the challenges of planetary defence, and the ongoing revolution in space commercialisation, often referred to as NewSpace.

7.4.2 Space science: a golden age

7.4.2.1 Current prospects (up to 2035)

We live in extraordinary times in space science with a plethora of novel and ambitious missions already planned for this decade, anticipating a scientific knowledge revolution. Three research fields are the focus of particular attention today:

- The observation of our universe through gravitational waves and multi-messenger astronomy.
- The quest for biological activity (life) beyond the Earth.
- The understanding of the dark universe: dark matter and dark energy.

Gravitational waves: a new window to observe our universe

After several decades of effort, the Laser Interferometry Gravitational Wave Observatory detected the first gravitational waves on 14 September 2015 (event GW150914), the ripples in the space–time fabric resulting from a binary black hole merger at about 410 Mpcs (about 1.3 billion light-years away). On that historic day, astronomy, as we perceived it up until then, changed forever. As is often stated, we could now 'add sound to the film of our universe', it having previously consisted only of images, the result of observation in the electromagnetic band. Just two years later,

on 17 August 2017, LIGO and Virgo detectors made another significant discovery when they detected the merging of two neutron stars, triggering a kilonova explosion. This event was subsequently observed by numerous telescopes worldwide, generating a total of 84 scientific papers in just one day. Since then, the global gravitational wave detectors network has identified over 80 mergers of black holes, two potential mergers of neutron stars, and a handful of events believed to involve black holes merging with neutron stars. Today, the quest for gravitational waves, is further intensified by the the LIGO–Virgo–KAGRA (LVK) collaboration, with enhanced instruments and various improvements in place, and will be further boosted by the future ESA's Laser Interferometer Space Antenna (LISA) Mission, recently confirmed and planned for launch in 2035. LISA will complement Earth-based gravitational wave detectors by enabling the detection of gravitational waves between 0.1 Hz and 0.1 MHz, a frequency sensitivity which should allow the capture of gravitational waves associated with events as extraordinary as the merger of supermassive black holes. LISA observations will be complemented by other ESA's planned large mission currently under study, NewAthena, scheduled for launch around 2037, and aimed at becoming the largest x-ray observatory ever built: an unprecedented temporal synergy of complementary observers from both Earth and space (please refer to section 2.6 for a detailed discussion).

The dark universe

The European Space Agency (ESA) Planck mission (put into orbit in 2009) has contributed to the most detailed knowledge to date of the cosmic microwave background (CMB) radiation, the radiation left over from the Big Bang, when the Universe was born, nearly 13.8 billion years ago. The CMB shows tiny fluctuations in temperature, anisotropies that correspond to regions that had a slightly different density in the initial moments of the history of the Universe (figure 7.14). Planck has allowed us to refine our knowledge of age, expansion, and history and to extract the most refined values yet of the Universe's ingredients [131].

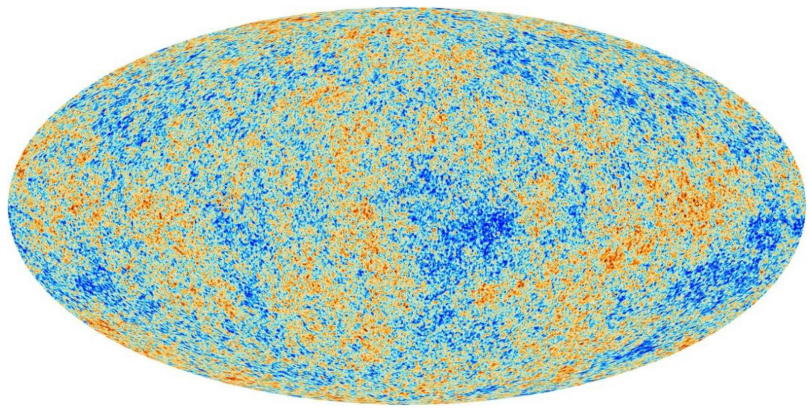

Figure 7.14. Cosmic microwave background image obtained by ESA's Planck telescope. (Credits: ESA and the Planck Collaboration.)

Thanks to data from the Planck telescope, we now know that conventional matter, of which stars, galaxies, and all baryonic matter are made, constitutes only 5% of the total mass–energy density of the Universe. Dark matter, which until now could be detected only indirectly through its gravitational interaction with galaxies or clusters of galaxies, constitutes 27%, according to our current estimates, thanks to Planck. On the other hand, dark energy, the mysterious force that we believe to be responsible for the expansion of the Universe, represents, according to our cosmological calculations, 68% [131]. In conclusion, in simple terms, we do not know the composition of 95% of our universe. Space (e.g. through the observations of the recently launched James Webb Space telescope and ESA's Eculid telescope or future planned NASA's Nancy Grace Roman Space Telescope or ESA's NewAthena missions) willprovide an unprecedented contribution towards progressing with this matter over the next 10–15 years (please refer to see section 2.5.1 for a detailed discussion).

Alongside gravitational wave observations and the understanding of the dark universe, an essential third pillar emerges strongly in space science in this decade: the search for life in the Universe.

The search for life beyond the Earth

The search for biological activity outside our planet focuses today on three main research axes: the possibilities of finding traces of life on the planet Mars, the possibility that life may have appeared in the subsurface oceans of the icy moons of Jupiter or Saturn, and the search for traces of biological activity outside the Solar System, on exoplanets orbiting other stars. The sequence of missions on-going or planned for this decade makes up the largest technological crew of seekers of life beyond the Earth ever, focusing upon those three objectives (please refer to section 4.2 for a detailed discussion).

7.4.2.2 Beyond 2035: space science priority themes

ESA has recently embarked upon the exercise of defining the scientific priorities for the 2030–50 period, a process known as 'Voyage 2050'.

Following a well-established and efficiently proven methodology, the first step consisted of a general call for ideas from the scientific community. This call, launched in 2019, generated over 100 different proposals or white papers [132], which were grouped according to specific scientific topics. Next, a defined Senior Scientific Committee was tasked with recommending the top-priority scientific topics for future ESA medium and large class missions, recommendations that were formally adopted by ESA's Science Programme Committee in June 2021 [133].

As a result, three identified top-priority scientific topics for future large missions have been proposed [133]:

- Research on giant planet moons, notably researching their potential habitability and capitalising on past Cassini-Huygens and future JUICE (Jupiter Icy Moons Explorer) missions. As noted by the science programme committee, these missions could include landers or drones for *in situ* research.

- Temperate exoplanets, proposing the detailed characterisation in the mid-IR domain of temperate exoplanet atmospheres in order to assess their habitability conditions.
- Development of new physical probes to study the early universe, which could include improving gravitational wave detection capabilities and expanding their frequency ranges, or the development of high-precision spectroscopy technologies to improve the measurement precision of the cosmic microwave background, capitalising on future LISA and past Planck missions, respectively.

Concerning medium-class missions, themes across all domains of space science have been proposed, including solar system research, fundamental physics, astrometry, astrophysics, and new astronomic observatories.

Following the same approach that was implemented for Cosmic Vision, this selection of topics was complemented by the identification of the technologies that would be needed to enable those missions. These include key technologies such as advanced x-ray interferometry for the analysis of astronomic compact objects, the development of more efficient and innovative power sources for solar system exploration, and improvement of the technologies associated with the storage of cryogenic samples for future sample return missions in the Solar System.

The identification of scientific topic priorities and the development of the associated technologies will allow ESA to launch individual calls for mission proposals during the coming years, which will detail, in turn, the new space science agenda for Europe in the 2030–50 period.

7.4.3 A new era for space exploration

7.4.3.1 Current prospects (up to 2035)

Since 19 December 1972, when the astronauts of the Apollo 17 mission returned to the Earth, no human being has left low-Earth orbit again. This is going to change during the current decade, and we may also witness a major revolution in space exploration during the next three decades, as we will discuss here.

The space exploration agenda for the next 10–15 years has been defined around four complementary elements:

1. To continue exploiting the ISS until at least 2028, the nominal date currently planned for its conclusion, maximising cooperation with private companies and facilitating the development of new business models.
2. To build a cislunar orbital station, the Gateway, which is the result of international collaboration by the main agencies of the world and is accessed through the Orion spacecraft of NASA and ESA.
3. To conduct robotic and manned lunar surface missions and the emergence of a lunar economy.
4. To perform a round-trip mission to Mars, the Mars sample return, bringing back samples of rocks from the red planet to the Earth for detailed scientific analysis.

These four phases are discussed briefly in the next part of this section.

The evolution of the ISS

The ISS is probably the most complex space project in history. A total of 15 countries have contributed to its construction, with the collaboration of the main agencies in the world. The ISS station has been inhabited without interruption since November 2000. The outcome of these 20 years of exploitation of the ISS is very positive, with important scientific advances in human physiology, molecular biology, biotechnology, materials science, fluid physics, combustion field, Earth observation, and fundamental physics.

The current proposal is to extend the useful life of the ISS until 2030, with a transition plan that enables its financial sustainability thanks to the additional contribution of private funds. Among the options that NASA is considering is the possibility of including new infrastructures, which are developed and operated privately. The ESA shares this vision and has already begun partnership activities in this regard. A good example of this are the IceCubes and Bartolomeo initiatives for experimentation marketed in microgravity conditions within the Columbus pressurised module or on the outside of it in outer space conditions, respectively (figure 7.15).

Returning to the cislunar orbit

The evolution of the ISS and its transition towards a more commercial model should allow the main space agencies to focus on new challenges beyond the Earth's orbit, with the Moon now being the natural intermediate step before a manned mission to Mars could be contemplated.

The lunar exploration roadmap for this decade is essentially defined, based on an open architecture around two main elements:

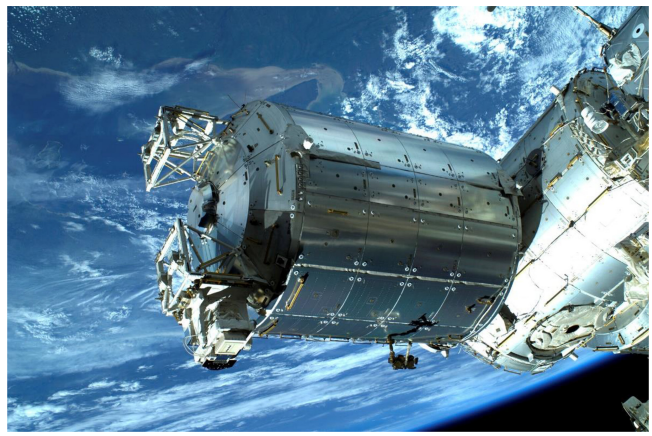

Figure 7.15. Europe's Columbus space laboratory. Image taken by ESA astronaut Luca Parmitano during his spacewalk on 9 July 2013. (Credit: ESA.)

1. A crew transport vehicle, the Orion spacecraft, capable of carrying a crew of up to six astronauts.
2. The development of a lunar station in retrograde cislunar orbit, the Gateway station, because of international collaboration, led by NASA.

The first mission of the Orion spacecraft, Artemis-1, was launched on November 16th, 2022 and had a duration of 25.5 days (figure 7.16). Passing as close as 130 km from the lunar surface, the spacecraft utilized the Moon's gravity to propel it into lunar orbit, and subsequently adjusted its trajectory to return to Earth. The first lunar flyby occurred on November 21st, with ESA's Service Module (ESM) activating its main engine to maneuver Orion behind and around the Moon. Ten days after liftoff, Orion entered the Moon's orbit on November 25th, when the ESM fired its main engine. Following a flight of 2.3 million km, the Orion capsule splashed down in the Pacific Ocean west of Baja California on December 11th, 2022. [134].

The Artemis-2 mission, an already manned spacecraft, completing a slightly different flight path, will follow the Artemis-1 mission. Artemis-2 crew members will reach a distance of 70 000 km beyond the Moon, the longest distance from the Earth that a human has ever traveled, before completing a lunar flyby and returning to the Earth. Artemis-2 is today scheduled to launch not earlier than September 2025, while Artemis III, aiming to land the first astronauts near the lunar South Pole, is today scheduled for September 2026. Artemis IV, the inaugural mission to the Gateway lunar space station, is today scheduled for 2028. The Artemis programme will establish the foundation for long-term scientific exploration at the Moon and will boost the emergence of a new lunar economy.

Figure 7.16. Orion crew transport vehicle (credits: ESA–D Ducros).

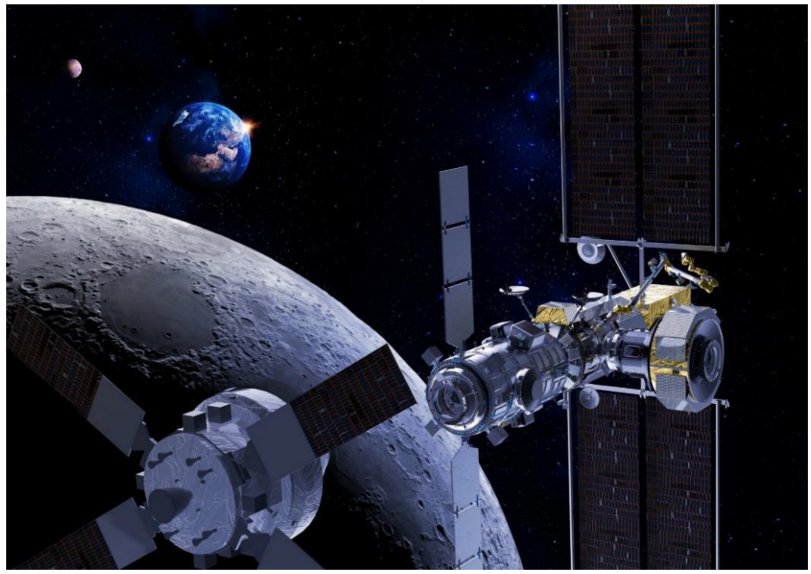

Figure 7.17. Future Gateway cislunar station. (Credits: Thales Alenia Space/Briot.)

The second element in the lunar exploration strategy, the Gateway space station, is conceived as a multimodular station, with six modules (figure 7.17). With a mass of about 40 tonnes, the Gateway station is planned to include a service module, a communications module, a connection module, a spacewalk access hatch, a habitable module, and an operations station to command the future robotic arm of the station or future lunar rovers. Europe, through the ESA, will contribute to the Gateway station with the ESPRIT (European System Providing Refuelling Infrastructure and Telecommunications) modules and the I-HAB (International Habitational module), the module that will constitute the main habitat for astronauts when they visit the Gateway station.

Astronauts are expected to inhabit the Gateway station for up to a total of 90 days consecutively. Gateway will open up enormous possibilities for future space exploration and will enable advanced technological testing and scientific advances. Gateway will also allow the remote control of robotic missions on the lunar surface and facilitate access to future manned missions. Carrying out such missions will represent the future of international collaboration and the cooperation of private industries.

Together with the contribution to the Orion spacecraft and the Gateway station, two additional axes currently form the basis of the European contribution to lunar exploration during the next 10 years:

- The Argonaut EL3 (European Large Logistic Lander) logistics vehicle, being developed as a versatile means of accessing the Moon. The ESA Argonaut programme will provide a series of recurrent landers dedicated to delivering cargo and infrastructure to support scientific operations on the lunar surface, including rovers and power stations. The first mission, Argonaut-1, is projected

to land on the Moon around 2031, with subsequent Argonaut landers planned for launch approximately every two years thereafter. Argonaut, together with the Moonlight service (referenced below), is poised to serve as a significant European contribution to NASA's Artemis lunar exploration program, providing capabilities for cargo delivery and scientific endeavours.

- The Moonlight/LCNS system, consisting of a miniconstellation around the Moon to provide a permanent satellite navigation and communication service to future lunar exploration missions [134]. Moonlight is expected to be able to offer its initial operational services in 2027–28 and its final operational capability around 2030. The Moonlight infrastructure will support both current and future generations of institutional and commercial lunar explorers, presenting a unique opportunity for European industry to assume a prominent role in the future Lunar Economy. The significance of the Moonlight program is enormous, potentially becoming the first ever extraterrestrial human infrastructure that provides commercial communication and navigation services: an extraordinary paradigm shift in the field of human exploration.

These technologies will undoubtedly contribute to the sustainable development of future recurrent robotic and manned missions to our satellite. They will, in turn, lay the foundations for the development of the necessary technologies for future manned missions to Mars, including protection against space radiation for human beings, advanced life support systems, and the possibility of exploiting lunar resources for stable human settlements. We can anticipate that this will also bring many benefits to our planet and will open up new commercial and business opportunities.

Mars sample return mission

Despite the large number of scientific missions to Mars to date, none has yet made the return trip to the Earth. Before we may consider a manned Mars return mission, it seems natural to implement that intermediate step with a robotic mission. That is the goal of the future Mars sample return mission: to make that journey to Mars and back, bringing samples from Mars to the Earth for in-depth scientific analysis. This mission, in addition to its extraordinary scientific value, should allow us to develop several of the technologies required to plan future manned missions to the Martian surface (figure 7.18).

But bringing Martian samples to the Earth is not an easy task. The current concept identifies the need to perform a minimum of three missions from the Earth and the launch of a rocket from the Martian surface, something that has never been done before.

Due to its nature, this mission also faces some key challenges concerning the environmental and biological protection of the Martian samples and Mars. For example, the perfect preservation of Martian samples is an essential element, minimising any possible organic contamination or chemical alteration of them. At the same time, all spacecraft destined for the Martian surface must incorporate unique planetary protection requirements [136].

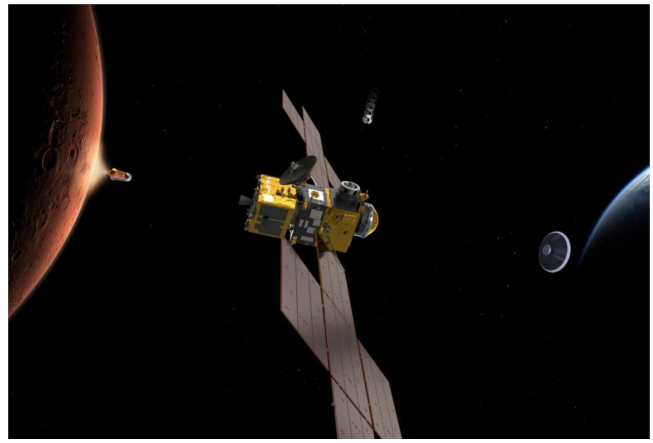

Figure 7.18. Mars sample return mission. (Credit: ESA/ATG Medialab.)

7.4.3.2 Space exploration beyond 2035

Towards a sustainable human lunar presence and research infrastructure

The idea of a permanent lunar presence on our Earth's satellite was strongly supported by ESA's former Director General, Jam Woerner, through a concept he named *Moonlight village* [137], intended to be developed through private and public cooperation, involving space and nonspace partners. This may become a reality in the next decade (figure 7.19).

During the next 10–15 years, dozens of missions, both institutional and commercial, are already envisaged. These will imply the development of much more accessible regular trips to our Earth's natural satellite and the setting up of a long-term infrastructure facilitating fundamental services such as communication and navigation services. It is also expected that these missions will enable the development of the necessary technologies to support a long-term human presence, such as the construction of radiation-shielding structures built with bulk regolith, regular water generation for life support and rocket propellant production, thermal and electrochemical processes to convert lunar regolith into oxygen and water, the consolidation of 3D printer technologies to build structures from local lunar materials, and even inflatable habitats and the development of regenerative closed-loop systems, such as those being currently studied at the ESA Melissa research infrastructure [138].

On the scientific front, the radio-quiet lunar far side, the so-called shielded zone of the Moon (SZM), may enable unique astronomical investigation, through the deployment of dedicated radio telescope infrastructures, for which specific frequency protection recommendations already exist [139]. China has initiated this research path by setting up some radio astronomy instruments as part of its recent Chang'e-4 lander [140], and several ambitious projects are being conceived today for the next decade.

Figure 7.19. Concept of a semiinflatable lunar habitat near the lunar south pole, designed by Skidmore, Owings and Merrill. (Credit: SOM.)

Future Mars human exploration: key technology needs

The experience that will be acquired in the next 10–15 years with Moon exploration, the Mars sample return, and associated robotic Mars missions should allow us to plan a crew mission to Mars in the approximate timeframe of 2035–45.

Some of the technologies needed for this human mission to Mars have been identified as part of the Global Exploration Roadmap [141]. This document is the result of the cooperative effort of 14 space agencies, with European contributions from ESA, CNES, ASI, and the UK Space Agency. Some of the identified key technologies are the following:

- Radiation protection technologies, including protecting constructions built with *in situ* regoliths and radiation-shielding technologies for human protection to sustain a long-term human presence.
- Advances in microgravity countermeasures, including the development of compact devices to limit microgravity disorders and possible medical countermeasures.
- Enhanced reliability life support systems, eliminating the dependence on Earth supply logistics and increasing systems autonomy, failure detection capabilities, and in-flight repair capabilities.
- Advances in human surface suits and EVA mobility, aimed at providing extended thermal, radiation, and vital life support.
- New propulsion technologies, based on the exploitation of *in situ* liquid oxygen–methane propellant production.
- Dust mitigation technologies to support both long-term lunar and Mars missions.
- Advances in autonomous systems enabling crew operations without the need for Earth-based support.
- Exploitation of *in situ* resources for life support, including O_2/CH_4 generation from the atmosphere and LOX/LH2 generation from soil and the

exploitation of many other chemicals and minerals to support a sustained long-term human presence.
- Advances in solar arrays and the development of fission power for surface missions' energy autonomy.

As space history has already demonstrated on multiple occasions, this focus and colossal planned technological effort towards achieving an enhanced human exploration capability in the Solar System will also bring extraordinary benefits to our life on the Earth, such as new technological capabilities with applications on the Earth, improved generation and efficiency in the use of the Earth's resources, and major advances in medical and life support technologies.

Towards a new space industry based on lunar and asteroid mining

It is known that the Moon and other celestial bodies may contain materials of great value for future space exploration or for use on the Earth. In the case of the Moon, the presence of ice in the polar regions could enable future exploration missions to produce oxygen, drinking water, or lunar rocket fuel. The Moon also contains minerals such as titanium, iron, and aluminium and an abundant presence of helium-3 from the solar wind.

Beyond the Moon, serious consideration is now also being given to the possible future mining exploitation of asteroids. Some asteroids may contain gold, silver, platinum, nickel, or cobalt, which some analysts believe could generate an extraordinary commercial opportunity if access and exploitation costs are lowered.

7.4.4 Space and climate change

The evolution of our planet and climate change are arguably recognised today as the most important global challenges. This is clearly reflected in the 2021 Annual Global Risk Report of the World Economic Forum, identifying extreme weather events, the inability to mitigate climate change trends, and anthropogenic effects on the environment as the three most important global risks [142].

The global monitoring of our planet and the continuous assessment of the effectiveness of future mitigation measures therefore becomes vital. Already today, Earth observation (EO) satellites represent more than a third of all operational satellites orbiting the Earth [143]. In the case of Europe, for example, over 22% of ESA's overall budget is today devoted to EO missions [144], including meteorological and scientific missions and the Sentinel satellites of the EU Copernicus Programme.

This proliferation of EO missions is essential for global, continuous, and long-term data relating to the so-called essential climate variables [145], defined through the Global Climate Observing System, jointly sponsored by the World Meteorological Organisation, the UNESCO Intergovernmental Oceanographic Commission, the UN Environment programme, and the International Science Council. The monitoring of these essential climate variables is key to understanding the overall health status of our planet, diagnosing and predicting its short-, mid-, and long-term evolution and to supporting government's decisions for its mitigation and

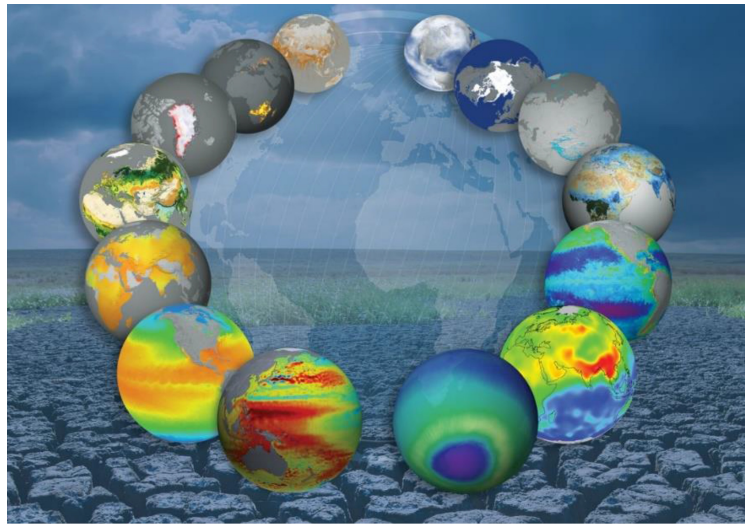

Figure 7.20. Approximately 50% of essential climate variables require satellite monitoring for global coverage. (Credit: ESA.)

risk management. In the case of Europe, this continuous monitoring will also be key in supporting the EU objective to be climate-neutral by 2050 [146], regularly assessing the current status, trends, and efficiency of the measures implemented, fully adhering to the European Green Deal strategy [147].

A total of 54 essential climate variables have been defined, grouped into three main categories: atmospheric, oceanic, and terrestrial [145]. Of these, approximately 50% require satellite measurements for their global assessment (figure 7.20).

This global and continuous observation is precisely one of the key objectives of the EU's Copernicus programme, which is currently providing the largest volume of Earth observation data worldwide. The space component of Copernicus includes the ESA-developed Sentinel satellites (the core of the programme) and other national and international contributions. All of these satellites provide essential data for the Copernicus Services, addressing challenges such as urbanisation, food security, rising sea levels, diminishing polar ice, natural disasters, and, of course, climate change [148].

There are currently eight Sentinel satellites in orbit and multiple national contributions (figure 7.21). To these we must add an extensive portfolio of more than 30 European Earth observation missions, which have already been defined for this decade, including the recently selected 7–12 Sentinel Mission families [149]:

- Sentinel 7 (**Copernicus Anthropogenic Carbon Dioxide Monitoring**, CO2M)
- Setinel 8 (Land Surface Temperature Monitoring, LTSM)
- Sentinel 9 (Copernicus Polar Ice and Snow Topography Altimeter, CRISTAL)
- Sentinel 10 (Copernicus Hyperspectral Imaging Mission for the Environment, CHIME)

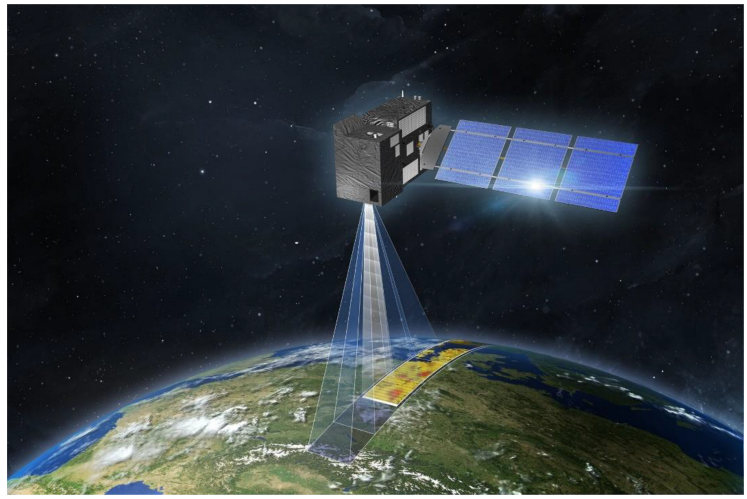

Figure 7.21. The future Sentinel 7, Copernicus Anthropogenic CO2 emissions monitoring mission. (Credit: ESA.)

- Sentinel 11 (**Copernicus Imaging Microwave Radiometer, CIMR)**
- Sentinel 12 (**Radar Observing System for Europe - L-Band**, ROSE-L).

These satellites are equipped with advanced instruments and sensors, including sea, land, and surface radiometers; altimeters; multiple-band spectrometers; synthetic aperture radars; photometers; multispectral and thermal cameras; trace gas detectors; atmospheric aerosols; very precise orbit determination devices; and so on. Their combined use enables a continuous and fine global analysis of our planet in all of its dimensions, namely, the atmosphere, hydrosphere, cryosphere, biosphere, and lithosphere, allowing us to distinguish between natural events and anthropogenic effects, which are the consequence of human activity.

The longer time period that these satellites will permit will enable the analysis of decadal trends. This is the purpose of ESA's Climate Change Initiative [150], which is aimed at generating accurate and long-term satellite-derived data to characterise the long-term evolution of the Earth's system.

7.4.4.1 Moving towards a scientific understanding of Earth dynamics

The Copernicus programme is currently the most ambitious Earth observation programme worldwide and a source of pride for Europe. Along with continuous monitoring, it is essential so as to have a deep scientific understanding of the critical variables of the Earth's system and to develop and test new technologies for potential use in future Sentinel satellites. This is precisely the objective of the ESA's Earth Explorers Programme: to contribute towards a significant improvement in our fundamental knowledge of the five main Earth science disciplines: atmosphere, cryosphere, land surface, ocean, and solid Earth. As of June 2021, this programme was composed of 11 Earth explorers, among which five were put in orbit between 2009 and 2018:

- GOCE: ESA's gravity mission (launched in 2009; ended in 2013).
- SMOS: ESA's water mission (launched in 2009; mission extended until 2025).
- CryoSat: ESA's ice mission (launched in 2010; mission extended until 2025).
- Swarm: ESA's magnetic field mission (launched in 2013; mission extended until 2025.).
- Aeolus: ESA's wind mission (launched in 2018; ended in 2023).

Six others are scheduled for launch during the next 10–15 years:

- EarthCARE: ESA's cloud and aerosol mission, planned for launch in 2024.
- Biomass: ESA's forest mission, planned for launch in 2024.
- FLEX: ESA's photosynthesis mission, expected to be lunched in 2025.
- FORUM: ESA's planet radiation budget mission.
- - Harmony: a constellation of two SAR (Synthetic Aperture Radar) satellites to monitor Earth's surface and ocean surface conditions.
- Earth Explorers 11 and 12: selection process ongoing.

7.4.4.2 Moving towards an accurate Earth digital twin

The continuous provision of Earth observation satellite data and an improvement in the understanding of the scientific principles governing the Earth's dynamic natural processes can contribute to the development, during this decade, of what has been named the Digital Twin Earth: a reliable digital replica of our planet which accurately mimics the Earth's behaviour and enables us to anticipate climate change evolution for the coming decades (figure 7.22).

A digital twin Earth should allow the monitoring and forecasting of natural and anthropogenic effects on our planet. The use of machine learning and AI technologies may prove to be necessary in order to obtain an accurate digital representation of our planet, notably for extreme weather events and accurate numerical forecasting models. This model could be extremely helpful in performing detailed simulations of the interaction between Earth's interconnected systems and human behaviour, supporting the field of sustainable development.

Figure 7.22. A digital twin Earth should allow us to reliably monitor and forecast the effects of natural and anthropogenic events on our planet. (Credit: ESA.)

7.4.4.3 ESA and NASA strategic partnership for Earth monitoring

In July 2021, ESA and NASA formalised a strategic partnership for Earth science and climate change [151]. Through this alliance, the two agencies will join forces in Earth science observation satellites, research, and applications to improve the monitoring of the Earth and its environment. This strategic agreement may also set the standard for future international collaboration, encouraging all key space stakeholders to actively share relevant satellite data in fighting against the extraordinary challenge of climate change.

7.4.5 Satellite navigation

7.4.5.1 Current prospects (up to 2030)

Knowing where we are and our time reference are two inherent measures of our human existence, linked to most of our daily activities. Not surprisingly then, satellite navigation systems have become essential in our societies, and this technology is considered by many as the 'fifth utility', along with water, electricity, gas, and telecommunications. All sectors of the economy benefit from it today: transport, energy, tourism, agriculture, fishery, livestock, civil engineering, telecommunications, the financial sector, and so on; around 40 000 different applications have been identified so far.

Because of this extraordinary economic and strategic importance, each relevant space nation in the world has or will soon have its own global or regional satellite navigation system. Altogether, over 130 operational navigation satellites are currently in orbit, of which over 100 correspond to the four GNSS constellations (figure 7.23).

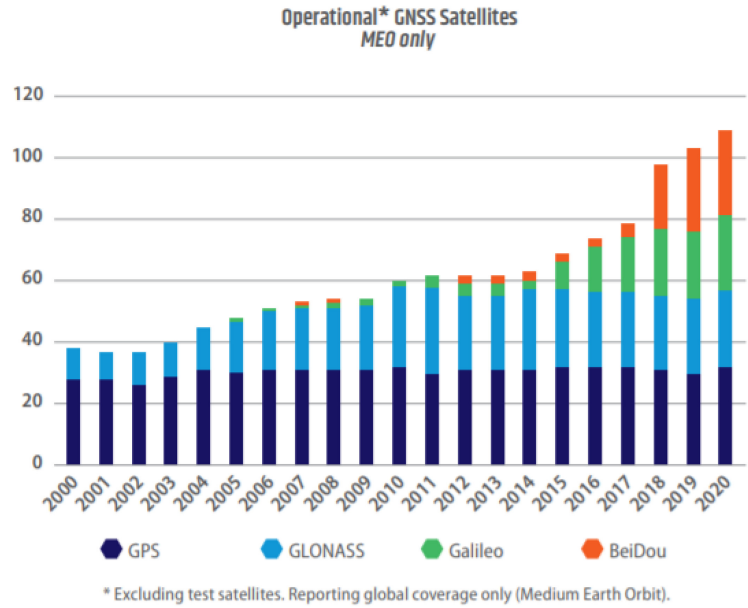

Figure 7.23. Evolution in the number of operational MEO satellite navigation satellites.

Current estimates indicate that around 10% of the European Union's GDP depends, to a greater or lesser extent, on the availability of satellite services. This economic dependence was central to Europe's decision to implement its own navigation systems, namely, EGNOS and Galileo, which are described next.

The European Geostationary Navigation Overlay Service

The European Geostationary Navigation Overlay Service (EGNOS) is Europe's regional satellite-based augmentation system. It is used to improve the performance of GPS (today) and of Galileo (in the near future) for safety of users. EGNOS provides an augmentation message to GPS L1 frequency by providing corrections and integrity information to GPS, and the ionosphere delays affecting the users. The EGNOS system began its open service in 2009 and was certified for civil aviation in 2011. The next generation of EGNOS, EGNOS V3, planned to be operational around 2026, will augment both GPS and Galileo constellations across their two operational frequencies: L1 and L5 bands, becoming the first multi-constellation and multi-frequency satellite based augmentation service.

The European Galileo system

At the end of the 1990s, in full development of the EGNOS system and thanks to the knowledge that had being acquired by the European industry in these technologies, the interest in designing a global satellite navigation system under European control arose. Several system studies were then launched, culminating in a system of 30 satellites at an altitude of 23 222 km and distributed between three orbital planes. This system, initially known as GNSS-2, was officially renamed the Galileo system in 1999. Since 15 December 2016, Galileo has officially been an operational system.

Today, Galileo is Europe's own GNSS, providing a highly accurate, guaranteed global positioning service under civilian control. By offering dual frequencies as standard, Galileo is able to deliver real-time positioning accuracy down to the metre range (figure 7.24).

Galileo features several technology and system differentiators, which allow it to reach higher levels of precision than other GNSS systems. First and foremost, Galileo satellites now include four clocks based on two different clock technologies: two rubidium atomic frequency standard (RAFS) clocks and two passive hydrogen maser (PHM) clocks. In the PHM clocks, molecules of hydrogen are dissociated into atomic hydrogen, entering a resonance cavity by passing through a collimator and a magnetic state selector, which means that only those atoms at the desired energy level enter a resonant cavity, where they tend to return to their 'fundamental' energy state, emitting a microwave frequency with very high stability. Galileo PHM clocks are about 1 order of magnitude more stable that the traditional Rb clocks, with stabilities in the range of 1 s in 3 million years. Thanks to a large and well-distributed set of Galileo sensor stations and very accurate modeling of the Galileo affecting nongravitational forces, the Galileo orbits are also computed in real time with very high accuracy, on the order of a few decimetres. Furthermore, the Galileo selected navigation signal modulation, based on what is known as the binary offset carrier family, provides more power at high frequencies away from the centre frequency,

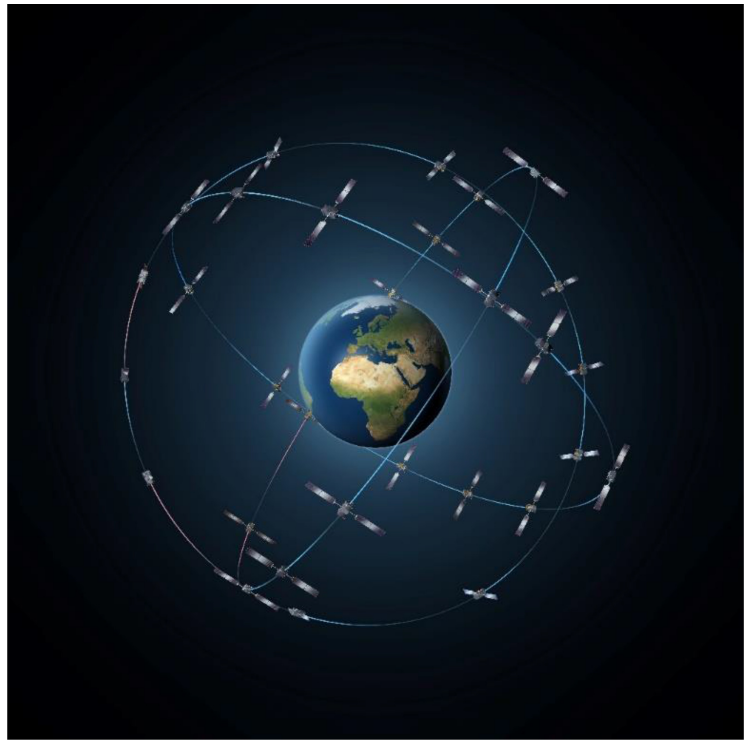

Figure 7.24. The Galileo system constellation. (Credit: ESA.)

thereby providing better ranging performances and a more robust tracking capability in multipath conditions than standard binary phase-shift keying signals [152]. Finally, Galileo satellites are also equipped with on-board laser retroreflectors, which allow periodic high calibration of the Galileo orbits with ground-based laser ranging stations and enhance its scientific utilisation potential.

With 26 satellites in orbit as of June 2021, Galileo allows Europe to have an autonomous and global system, making the continent one of the world's leading players in this field. Galileo currently offers a service of the highest quality, and all the world's leading mobile handset manufacturers include it (figure 7.25).

Beyond its open global service, Galileo also provides a robust and encrypted service for government applications (PRS, or public regulated service) and is currently also the only civilian GNSS including an operational search and rescue service, integrated as part of the COSPAR-SARSAT service [153].

Since 2023, Galileo has provided a free of charge High Accuracy Service (HAS) by offering high accuracy Precise Point Positioning (PPP) corrections through the Galileo signal (E6-B) and Internet. Galileo HAS offers today real-time improved user positioning performances down to a decimetre level [154]. This new service becomes a major differentiator for Galileo and will prove to be an enormous facilitator in the development of new applications and services that require high precision. Good

Figure 7.25. Ariane 5 (VA244) launch with Galileo satellites 23, 24, 25, and 26 on board, July 2018. (Credit: ESA.)

examples include enhanced drone navigation, augmented reality, autonomous cars, and vehicle-to-everything services, in which connected vehicles will communicate wirelessly with other vehicles and infrastructure to avoid collisions.

Another important differentiator of Galileo today is the provision of signal authentication via its Open Service Navigation Message Authentication (OSNMA) service of free access. This service provides users with assurance that the Galileo navigation message they receive originates from the system itself and has not been altered. [155]. This makes its positioning service robust against possible malicious attacks that could emit signals similar to Galileo signals with the aim of causing errors in the receivers.

Galileo second generation

The European Union and ESA are currently deeply engaged in the development of the second generation of the Galileo system, known as G2G. Galileo G2G will incorporate various new technologies, including the integration of ion propulsion in future satellites, establishment of intersatellite links between Galileo satellites, implementation of several signal enhancements for quicker acquisition and reduced energy consumption of receivers, and a heightened level of digitalization of the navigation signal generator payloads to enhance their flexibility and reconfigure their on-board capacity. These advancements will result in new service capabilities, enhancements to the existing service, a more robust and secure service, and reduced operational and maintenance costs of the system. It is envisaged that the first satellites of this Galileo second generation will be deployed into orbit around 2025, with the G2G constellation expected to achieve its initial operational capability before 2030.

7.4.5.2 Satellite navigation: technology prospects beyond 2030

Beyond the development of new generation Galileo and GPS infrastructures, Galileo second generation and GPS III, respectively, which should be in full operation by the end of this decade, there are several technology trends, currently

in the research field, which may be operational by the next decade. Some of them could include the following:

- The comprehensive implementation of a specialized low-Earth orbit position, navigation, and time system, referred to as LEO-PNT, presently undergoing thorough evaluation. The envisioned LEO-PNT constellation will facilitate a multilayered 'system of systems' navigation paradigm, wherein signals from medium-Earth orbit are complemented by those from low-Earth orbit (LEO) satellites positioned at altitudes below 2000 km—along with additional inputs from terrestrial PNT systems and user-based sensors.
- The regular operational use in future GNSS satellites of space optical clocks, based on the use of lasers with a frequency stabilised relative to an atomic transition, which could provide clock stabilities several orders of magnitude better than those of today, with a major impact on the final system performance. For example, technologies based on the Doppler-free spectroscopy of molecular iodine are under development [156].
- The complete fusion of GNSS and mobile telecommunications infrastructures in a seamless way, including the full integration of indoor and outdoor navigation services in a unique device. This, together with the full development of sensor fusion technologies with GNSS signals and the application of AI technologies, should make it possible to reach real-time accuracies at a millimetre range and global level.
- The full integration of QT, AI, and cybersecurity technologies as part of future GNSS systems enhancing security and system performances.
- The modernisation of regional augmentation systems conceived to extend the integrity services beyond aviation to all transportation services and for centimetre accuracies.
- Improvements in the Earth International Terrestrial Reference Frame references accuracy, reaching submillimetre level precision.
- The extension of advanced PNT technologies to the field of exploration, starting with the Moon (the Moonlight concept already having been planned for this decade [135]), including the deployment of local lunar differential systems, and then extended to Mars beyond 2030.
- The deployment of GNSS-like nodes in strategic space locations within the solar system, such as the Lagrangian points of the Sun, aiming to enhance deep space Positioning, Navigation, and Timing (PNT) capabilities.
- The potential extension of PNT services to deep oceans by exploiting the use of neutrino beams for navigation and time dissemination [157].

7.4.6 The problematic of space debris

Since the beginning of the space age, according to 2023 data at the ESA, it is estimated that a total of about 17000 satellites have been launched, of which about 11500 are still in orbit; of that number, about 9000 remain active [158]. In addition, the number of events that have produced ruptures, explosions, collisions, or fragmentations in orbit are estimated at more than 640. As a result, the current

balance is a total mass in orbit exceeding 11 500 tonnes, with an estimated count of about 36 500 objects measuring 10 centimetres or more, approximately 1 million objects ranging between 1 and 10 centimetres, and roughly 130 million objects sized between 1 millimetre and 1 centimetre [158]. Of all these objects, currently, we are monitoring approximately 48,000, primarily through the US catalogue of the Space Surveillance Network [159].

Nowadays, the problem is further aggravated in the LEO orbit due to the dramatic increase in the number of small satellites launched into near-Earth orbit during the last 10 years, owing to the emergence of cubesats and large satellite constellations, as illustrated in figure 7.26.

As a result, the risk of collision with functioning satellites is increasing, notably in low-Earth orbit, where a large number of Earth observation and climate change monitoring satellites operate. Unfortunately, we have already seen some real examples.

On 23 August 2016, after about two years of operations, we observed a sudden loss of power on our Copernicus Sentinel Satellite 1A. Our investigations revealed that this loss of power was preceded by a slight change in the attitude and orbit of our satellite [160]. All this made us suspect a possible collision with an object in orbit. Our analyses confirmed this, and we concluded that the collision was caused by an object of only 1 cm in length and 0.2 g of mass, which had collided at a relative speed of about 11 km s^{-1} with one of our solar panels (figure 7.27).

Therefore, we clearly see that the problem of space debris can have an impact on the availability of our critical satellite infrastructures.

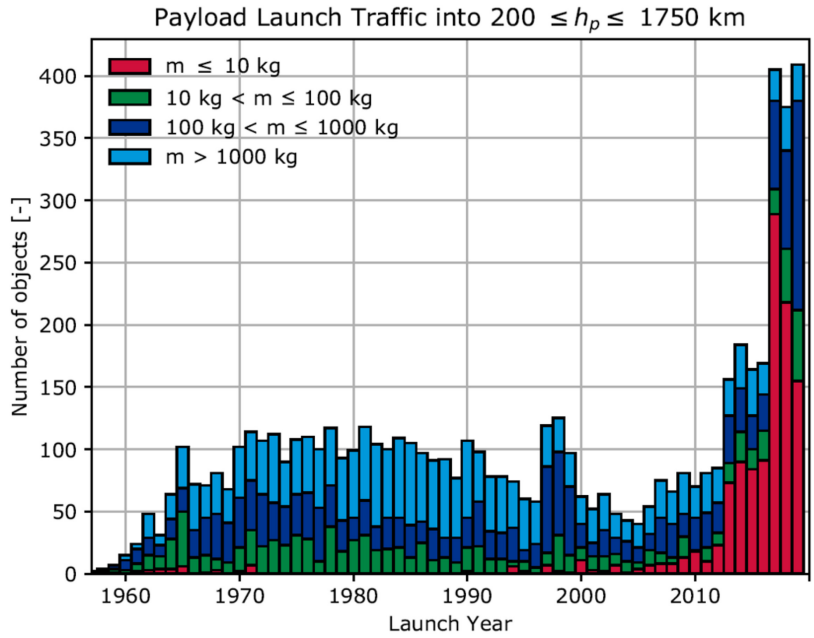

Figure 7.26. Number of satellites launched below 1750-km orbit and their associated mass [56]. (Credit: ESA.)

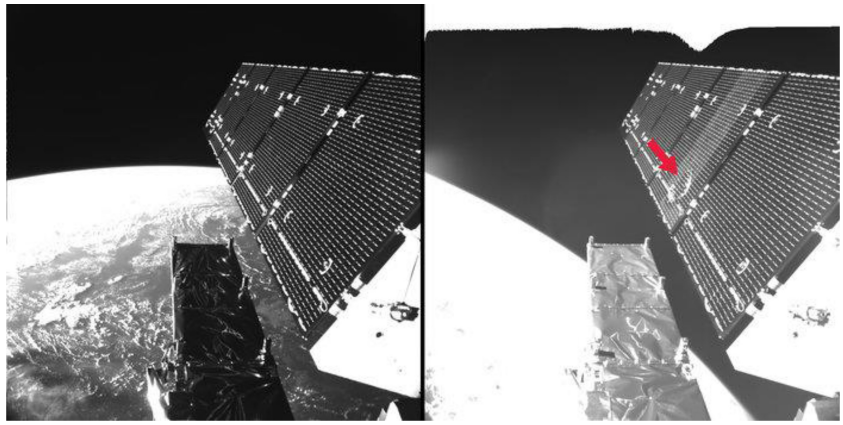

Figure 7.27. One of the sentinel 1A satellite's solar panels before and after a collision with an object of space debris in orbit. (Credit: ESA.)

The data are very worrying and require a clear sustainability policy. According to our ESA estimates, an uncontrolled long-term evolution of the space debris environment could lead to a cascading effect, rendering the orbits between 900 and 1400 km unusable for operational services. This effect, known as the Kessler syndrome [161], has been demonstrated through simulations in which fragments currently in orbit collide over time with larger objects, resulting in more fragments, causing more and more collisions. A cascading effect would have catastrophic consequences for satellites in low orbit, which could be rendered useless, and certainly for our planet (figure 7.28).

International cooperation between the different agencies and governments is, thus, essential. An important step in this direction is the effort invested by the Inter-Agency Space Debris Coordination Committee (IADC), an international governmental forum that was specifically established for coordination on this issue. Through the IADC, a specific code of conduct has been defined, and different axes of action are proposed to mitigate this problem [162].

Along with these operational, legal, and regulatory measures, it is essential to analyse strategies for reducing the number of the most dangerous passive objects that remain in orbit. In this regard, it is of interest to highlight the ESA Clean Space initiative, which, among its many proposals, includes the first demonstration mission for the active elimination of space debris: the ClearSpace 1 mission, currently planned for launch around 2026 [163].

7.4.7 The prospects of planetary defence

We can all remember the impact, on 15 February 2013, of a meteorite striking the Earth in the Russian region of Chelyabinsk, leaving more than 1500 people injured and affecting six cities in the region. The destruction, according to the best estimates,

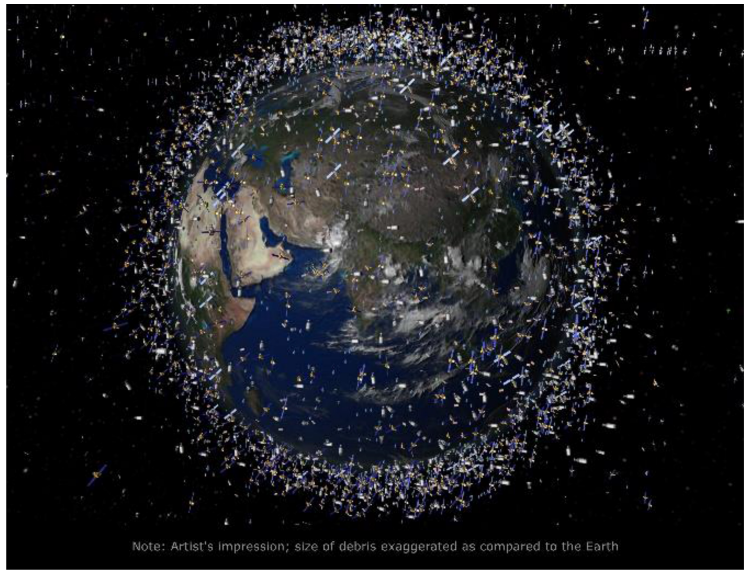

Figure 7.28. Artist's impression of space debris in low orbit. (Credit: ESA.)

was the result of the impact of a meteorite just 20 m in diameter that had not been monitored in advance [164].

Although the frequency of large asteroid impacts is not very high, the consequences may be of catastrophic dimensions. To deal with this risk in an efficient way, we need to develop technologies for our planetary defence, which we consider to be within reach over the coming decades.

The first obvious step is to be able to achieve accurate monitoring of near-earth objects, which are likely to involve a risk of collision with our planet [165]. The second planetary defence security objective, as defined by ESA, which is undoubtedly much more ambitious, would be to have the technological capacity to be able to deflect asteroids up to 1 km in diameter if they were identified more than 2 years in advance. This is certainly a great technological challenge, but it could be achieved in the next 30 years.

In preparation for this, the ESA is currently implementing the Hera mission. Hera is the European component of a full-scale planetary defence mission, complementing NASA's Double Asteroid Redirection Test (DART) mission in its analysis of the Dydimos dual asteroid system, composed of two asteroids measuring 780 m and 160 m in diameter (figure 7.29).

NASA's DART spacecraft successfully executed a kinetic impact on the smaller asteroid, known as Dimorphos, on September 26, 2022, altering its orbit by approximately 33 minutes and creating a local crater. This marked the world's first planetary defence technology demonstration. ESA's Hera mission, scheduled for launch during 2024 and estimating to reach the asteroid system in 2026, will conduct a thorough post-impact survey, becoming Europe's flagship Planetary Defender. The ultimate goal of this joint mission is to validate the basic technologies necessary

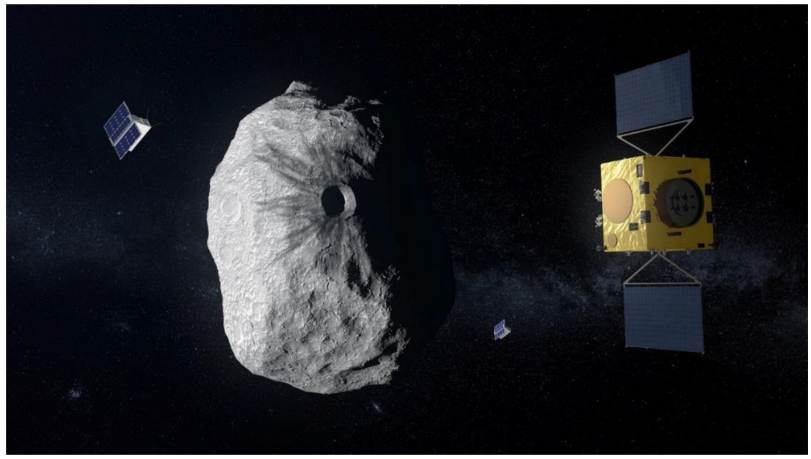

Figure 7.29. ESA's future Hera mission, aimed at contributing to the development of some of the basic technologies for future planetary defence. (Credit: ESA.)

for the development of a reliable planetary defence strategy that could be implemented in the forthcoming decades.

7.4.8 The prospects of NewSpace

Around 75% of the world's space economy is today of a private nature, with a turnover of approximately 270 billion euros (data from 2021), the downstream space business being responsible for more than 95% of that figure [166]. The current dynamic of the space industry is enormous, having doubled its turnover over the last 10 years. But this is just the beginning. Several organisations have conducted prospective studies on the future scale of the space economy, projecting figures ranging from 1 trillion to 2.7 trillion USD for the period spanning 2040–2045 [167].

An important part of this revolution is linked to the advances in digitalisation technologies and to the miniaturisation of components, which, together with an extraordinary reduction of the launch costs, makes the space sector accessible to newcomer industries: a new paradigm often referred to using the term *NewSpace*. This extraordinary transformation of the space industry is enormous, having attracted over 14.8 billion euros in private investment between 2000 and 2018 [168].

An integral aspect of this NewSpace is the emergence today of megaconstellations, comprising networks of hundreds, thousands, or even tens of thousands of small satellites. Private initiatives like SpaceX's Starlink and Amazon's upcoming Project Kuiper now offer broadband internet services, facilitating streaming, video calls, online gaming, and more. This has resulted in a completely new paradigm and revolutionized the commercial space communications sector, with a trend of a clear steady growth.

Another sector with great dynamism and from which we can expect another revolution over the coming decades is linked to space tourism.

7.4.9 Summary

The space sector is today a mature and diversified sector, present in all facets of our society and with enormous growth potential. Over 9000 operational satellites orbit our planet today and tens of thousands of new satellites are scheduled to join them in just a few years. Multiple satellite services are now simply essential for the smooth operation of our society and the global economy. In the European Union, for example, it is estimated that more than 10% of its GDP currently relies on the availability of satellite services, a figure far exceeding the sector's direct turnover. Today's space sector dynamic is enormous, and according to some recent forecasts, the space economy could grow by a factor of 10 in 2045, reaching a potential business of about 2.7 trillion dollars [167].

In this section, we have reviewed the status and future prospects of some of the main space sector domains and their associated technologies, covering the topics of space science, human and robotic exploration, climate change and earth monitoring, and satellite navigation.

Multimessenger astronomy, gravitational wave detection from space telescopes, the understanding of our dark universe, the evidence of life beyond Earth, a lunar and Martian sustainable human presence, and regular asteroid mining exploitation are just some of the examples explained here, illustrating the extraordinary prospects of space science and space exploration which could be within reach in the coming decades.

In the applications field, we explained the growing importance of earth observation in the monitoring and mitigation of climate change, our most urgent global challenge today. Scheduled space missions will allow a continuous, global, and complete monitoring of the 54 essential climate variables, providing us with a more complete understanding of climate change trends, their sources, their dynamics, and the major anthropogenic impacts. These precious data, together with a complete scientific understanding of the critical variables of the Earth system, thanks to future scientific satellite missions, and the development of a reliable digital twin Earth will provide us with the tools and knowledge to respond efficiently to our most urgent challenges in terms of food security, urbanisation, rising sea levels, diminishing polar ice, natural disasters, extreme weather events, and, of course, climate change mitigation.

We have also discussed here the current and future prospects of satellite navigations, with the planned development of second- and third-generation global navigation satellite systems, which could allow us to reach millimetre position accuracies in real time and at a global level in the coming decades. We have identified here also some of the future PNT technologies under development, which could allow us to contemplate the extension of navigation services to deep oceans or to the near Solar System, with dedicated infrastructures around the Moon and Mars, already planned today.

But the growth of the space sector also implies the need to take effective and urgent measures against space debris. Major improvements in space debris monitoring, international cooperation, the development of effective legal and regulatory

measures to avoid new debris, and new technologies for the active reduction of the most dangerous passive objects or automatic collision warnings on-board our future satellite, are some of the expected advances for these decades, which were briefly discussed here.

Protecting our planet also implies protection against possible extreme space weather events, which could endanger many of our space and critical digital Earth infrastructures, and the development of active monitoring and planetary defence technologies to protect us against near-earth objects, which could involve a risk of collision with our planet. As we have discussed here, current technology plans and some pioneering demonstration missions already undertaken in this decade could lead us to the availability of reliable early warnings in the case of dangerous asteroids measuring over 40 m in diameter and to the capacity to deflect asteroids of up to 1 km in diameter, if identified more than two years in advance.

We concluded this section by addressing some of the extraordinary prospects linked to the NewSpace paradigm and the major transformation of the space sector that this implies.

7.5 Large-scale complex sociotechnical systems and their interactions

Marc Barthelemy[1]
[1]Université Paris Saclay, Gif-sur-Yvette, France

7.5.1 Complexity and modeling

7.5.1.1 Cities are complex systems

Cities are certainly among the most complex systems built by humans. They are made of a large number of different constituents, such as individuals, various institutions, and private companies, that interact in various ways at different time and spatial scales. All the ingredients are present for giving rise to emergent collective behaviours and consequently to difficulties in understanding and modeling these systems. In particular, there is the issue of unpredictability; varying some parameters will not cause any change to the system, while varying others leads to dramatically different behaviours and outcomes. This is one of the difficulties of modeling complex systems: A naive modeling can completely fail, and it is not by adding various parameters and mechanisms that we can increase the realism of the model and the validity of its predictions.

An important thing to notice is that thanks to the variety of sensors and other devices such as mobile phones, GPS, or RFIDs, the amount of available data about urban systems increased dramatically this last decade. For example, origin–destination matrices, a fundamental ingredient that tells us where people are living and where they are working, was essentially obtained by surveys during censuses. The data were then under the form of locations and estimate of trip duration and so on. Now, thanks to the large variety of new data sources, we have access to the trips of millions of individuals at an unprecedented resolution. More generally, we have data at all scales (see table 7.6) from the minute (or less) with mobile phones and GPS to longer time scales such as months or years with socioeconomic data (e.g., taxes or real estate transactions). Even longer timescales with the digitalization of historical maps are now accessible; we can study the evolution of the road network over centuries [169]. Despite this recent availability, there is still a lot of do in terms

Table 7.6. Data sources according to their typical timescale and some phenomena occurring on these timescales.

Timescale	Data sources	Phenomena
Minutes to days	GPS, mobile phones RFID	Spatial structure mobility, urban activity
Month to years	Surveys, censuses	Social organization housing market
Decades to centuries	Historical documents and maps	Urban growth, self-organization impact of planning

of data accessibility. Some countries are better than others and the European initiative EuroStat can certainly be improved. Ideally, cities should open their own data platform that would certainly trigger the interest of many scientists.

This flow of data about many different aspects of cities allowed scientists to test their theory. While urban economics [170] mainly developed as a mathematical field with few connections to reality, we can now construct a model and test its predictions against empirical observations. This increasing availability of pervasive data opened the exciting possibility of constructing a 'new science of cities'. A new problem that we have to solve is then to extract useful information from these huge datasets and construct theoretical models for explaining empirical observations. In particular, a common difficulty shared by many complex systems such as cities is the existence of a large number of agents, acting through a variety of processes that occur over a wide range of time and spatial scales. It is then necessary to disentangle these processes and to single out the few that govern the dynamics of cities. Albeit difficult, the hierarchization of processes is of prime importance, and a failure to do so leads to models which are either too complex to give any real insight into the phenomenon or too simple to provide a satisfactory framework which can be built upon. In this context, statistical physics might bring convenient tools for both the empirical analysis and modeling, with the possibility of characterizing emergent macroscopic phenomena in terms of relevant parameters describing the basic elements of the system.

7.5.1.2 How can we model a complex system?

As became obvious in interdisciplinary studies, the notions of a model and modeling are not unique and well defined, and this ambiguity is even bigger in complex systems studies. In statistics a model is what we call in physics a fit; it is a mathematical representation of observed data. The question is then what is the best fit, but there are no mechanistic interpretations here. In contrast, in physics a model is a simplified version of the reality that is supposed to capture the essence of the phenomenon by omitting some details that we hope will be irrelevant or play the role of noise. This model usually allows a mathematical study of the system, and in some cases we need to study it numerically. The loop between the model and comparing its prediction with experiments or empirical data is what worked very well for physics these last centuries. The virtue of these models lie essentially in their simplicity and their parsimonious use of parameters. The main function of such models is to explain what happens in the system and to help us understand its functioning and identifying critical parameters.

Numerical approaches modified a bit this type of approach, especially nowadays when CPU power is large enough that implementing many details of a system is not really a problem anymore. In this sort of approach, the system contains many parameters and mechanisms, and it is tempting to simulate it directly by implementing the behaviour of its constituents (the agents) and their interactions. We usually refer to this sort of simulations as agent-based ones, and in the context of cities, the extreme version of this is called a 'digital twin,' which is supposed to model all aspects of a city. For complex systems, we immediately face several problems with

this type of approach. First, these numerical models are difficult to validate, and their sensitivity when varying their parameters can be very large and difficult to assess. Second, these models are usually very specific to a particular system (a region or a city) and sometimes cannot be easily transferred to another case. Related to this, these models act usually as black boxes and do not help us to understand what is really going on in these systems. In order to get around these criticisms, scientists use these models as a tool, not for forecasting and predicting, but to explore various scenarios and the impact of different measures or strategies. We can thus obtain projections that can be helpful for policymakers, for example. We saw this, for example, in the case of the COVID-19 pandemic. The simulations are, however, not the ground truth here, and this must be clearly stated, in particular when presented to policymakers or stakeholders. Digital twins that are exact replica (if that is even possible) of some systems and their behaviour of course diverges in general quickly from the real systems. Even if smart cities–related analyses seem to rely on this type of approach, it is unclear at this point how they will surpass other tools such as agent-based ones or more physical approaches.

Finally, there is machine learning that might revolutionize how we do science. At this point, these approaches are still limited, but we can expect to see their importance growing quickly, especially for real-world applications. So far, these algorithms that encode the complexity of data in a very large number of parameters act as black boxes and produces an output that is difficult to challenge and to understand. However, we can envision a possible future where machine learning techniques will become a tool among others and will assist researchers to make progress in our understanding. This sort of AI-assisted theoretical research could be the path to many important discoveries.

Going back to cities, an important aspect is interdisciplinarity. If we want to construct solid, scientific foundations of urban systems [171, 172], we need to produce an effort that is necessarily interdisciplinary, which is not always easy (see [173]); we have to build on early studies in urbanism to discuss morphological patterns and their evolution and on quantitative geography and spatial economics to describe the behaviour of individuals, the impact of different transportation modes, and the effect of economic variables (e.g., income, the rental market). In the following subsection, we will illustrate different aspects of this type of approach that combines ingredients coming from different fields and statistical physics using various examples ranging from the temporal evolution of cities (population and area), to congestion and CO_2 emission modeling.

7.5.2 Results and challenges in urban systems

There are obviously many aspects in cities. We will focus here on problems that can be tackled by quantitative methods. In addition, we will discuss problems that possess some duality; that is, they have a theoretical component, but also a practical one. This dual aspect is what usually make complex systems interesting, as they are rooted in reality but challenge our theoretical understanding. In each case, we will also mention interesting questions and challenges for future studies.

7.5.2.1 Searching for the equation of cities

Apparently, the simplest question about cities concerns the time evolution of urban populations. This question appeared more than a century ago with Auerbach [174], a German physicist who took some interest in the statistics of cities. Auerbach remarked that if the population of German cities are sorted in decreasing order $P_1 > P_2 \cdots > P_N$ (so that the rank $r = 1$ corresponds to the largest city), then the product $r \times P_r$ is approximately constant, implying that the population is inversely proportional to its rank (Auerbach, ahead of Zipf, also discussed that this type of relation could go well beyond cities and could be applied to other systems). This result was generalized by Zipf [175], who constructed a graphical representation of this, now known as rank–size plots. Zipf then plotted the population P_r versus its rank r and confirmed Auerbach's result for many countries. This is now called Zipf's law:

$$P_r \sim \frac{1}{r} \tag{7.7}$$

More generally (see, e.g., figure 7.30), we don't observed a strictly inverse proportional relation but something of the form $P_r \sim 1/r^{\nu}$, where the exponent ν is in general close to 1 (empirically, we also observe that this law is more accurate for smaller cities (see, e.g., figure 7.30).

This statement is equivalent to saying that the population is distributed according to the power law:

$$\rho(P) \sim \frac{1}{P^{1+1/\nu}} \tag{7.8}$$

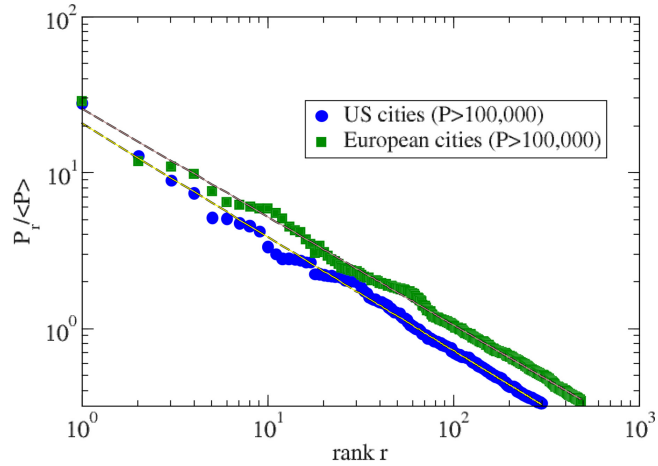

Figure 7.30. Zipf's law for European and US urban areas with population larger than 100 000 inhabitants (data from Eurostat for Europe and the S Census Bureau for the US). We normalized here the values by the average, and the plot is shown in loglog. The dashed lines represent power law fits (here the exponents are close to 0.70 in both cases).

The exponent ν is close to 1, which implies that the population distribution behaves as $\rho(P) \sim 1/P^2$, indicating a very large heterogeneity of urban systems. Quantitatively, the most fundamental problem here is to understand the hierarchical organization of city populations and the statistical occurrence of megacities described by such a distribution. On one hand, many authors have tried to explain this fact theoretically; on the other, the always growing availability of data showed that the exponent ν is not universal and can display large variations [176].

In addition, the dynamics of cities is not a smooth one; throughout history, we have observed many rises and falls of cities. A theoretical model must then be able to explain not only Zipf's law and its variation, but also this turbulent dynamics of cities. This problem was only recently answered with an approach that combined data and statistical physics tools [177]. More precisely, a stochastic equation for modeling the population growth in cities was constructed from an empirical analysis of recent datasets (for Canada, France, the UK, and the US). This equation reveals how rare but large interurban migratory shocks dominate population growth, and it predicts a complex shape for the distribution of city populations and shows that, owing to finite-time effects, Zipf's law does not hold in general, implying a more complex organization of cities. It also predicts the existence of multiple temporal variations in the city hierarchy, in agreement with observations [177, 178]. This result underlines the importance of rare events in the evolution of complex systems and, at a more practical level, in urban planning.

This long-standing problem about urban population could naively be thought as being the simplest one, but in fact it took almost a century for a more precise understanding. This problem is, however, probably much less involved than another crucial issue that mixes space and population and represents an important challenge for studies in this field: urban sprawl. In other words, how does a city evolve in space?

It is now generally accepted [178] that urban sprawl causes the loss of farmland, threathens biodiversity, and affects local climate. Despite these negative effects, urban land area increased by ~60 000 km^2 from 1970 to 2000, with China, India, and Africa having the highest urban expansion rate. Although urban sprawl can be considered a local issue, its impact in terms of biodiversity and vegetation loss is at a global scale [180]. A theoretical understanding of urban growth could be extremely helpful for proposing mitigation strategies for these problems. Growth in GDP per capita and population seem to be the important drivers for the observed urban land expansion, but much of the observed variation in urban expansion is not captured by either population, GDP or other variables [179]. This clearly shows the limits of pure econometric studies, and a more mechanistic approach, inspired by physics studies could be of great help here. The spatial evolution of a city is, however, certainly a difficult problem. For example, we show in figure 7.31 the spatial representation for the city of London (UK) for different dates from 1800 to 1978 showing how the built area spreads in space.

For a physicist, the image in figure 7.31 triggers many questions. In particular, the natural question to ask is: What is the growth process (diffusion or faster?) and can we write the corresponding growth equation? There are of course many quantities

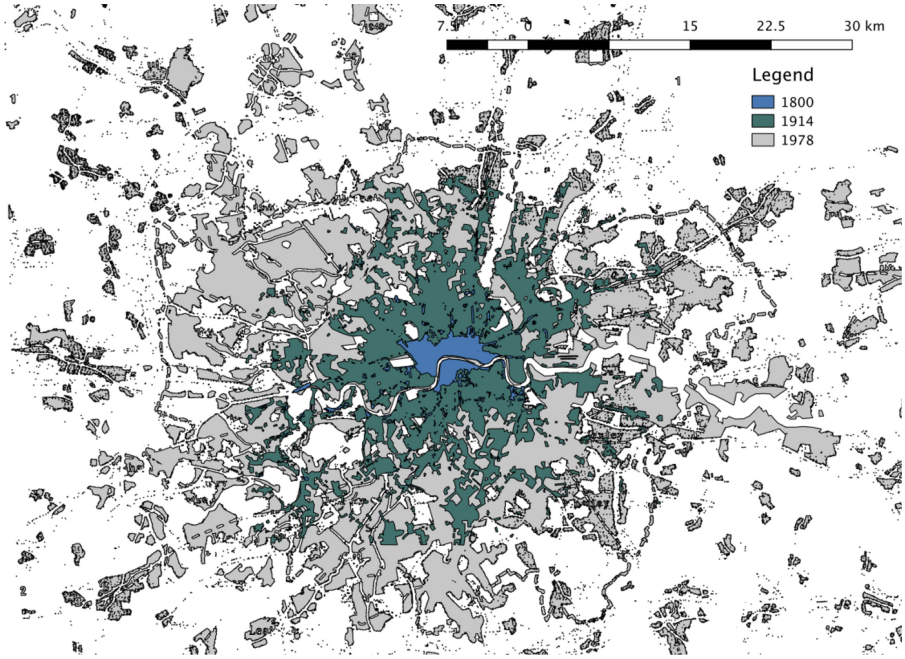

Figure 7.31. Spatial evolution of the built-up area of London for three different dates: 1800, 1914, and 1978. Data from [181].

that we can measure, but the first natural one is the surface area and how it varies with time or population. Answering this question and understanding this behaviour would represent important progress for the science of cities. Even further progress would be to describe the evolution of the frontier of the city and to find the corresponding equation. We can probably expect the effect of many factors, such as the presence of the road system, which evolves also in time and space. In addition, the city is not evolving in free space, and we expect population density and the transportation network to coevolve with the built area. This last remark actually points to the problem of the coupling between transportation network and population density: Better transportation attracts more individuals, and a large number of individuals pushes decision makers to construct better transportation. The coevolution of these quantities certainly represented an important ingredient for understanding urban sprawl. As we see with these short discussions, finding the equation that governs urban sprawl is certainly an extraordinary challenge but represents also a fantastic crossroad between the statistical physics of surface growth and the science of cities.

7.5.2.2 Congestion, CO_2 emission, and energy use

The evolution of urban sprawl triggered an increase use of cars, leading in turn to increased congestion, longer travel times, and more CO_2 emissions. More generally, understanding mobility patterns in urban areas [182] has become paramount in

reducing transport-related greenhouse gas emissions and crucial for proposing efficient environmental policies. In a seminal paper, Newman and Kenworthy correlated transport-related quantities (e.g., gasoline consumption) with a determinant spatial criterion: urban density [183]. Higher population density areas were shown to have reduced gasoline consumption per capita and thus reduced gas emissions. Their result had a significant impact on urban theories over the last decades and has become a paradigm of spatial economics [184]. This study is, however, purely empirical and has no theoretical foundation, which casts some doubts about the importance of density as the sole determinant of gasoline consumption and other car-dependent quantities. A way to test this theoretical assumption is to construct a simplified model that is validated by empirical observations and that eventually could help us to understand the effect of population density on car traffic. In order to illustrate this type of approach, we combine economic and transport ingredients into a statistical physics approach and construct a generic model [185] that predicts for different cities the share of car drivers, the CO_2 emitted by cars, and the average commuting time.

According to the classical urban economics model of Fujita and Ogawa [186], individuals choose job and dwelling places that maximize their net income after deduction of rent and commuting costs. More precisely, an agent will choose to live in x and work at location y such that the quantity

$$Z(x, y) = W(y) - C_R(x) - G(x, y) \tag{7.9}$$

is maximum. The quantity $W(y)$ is the typical wage earned at location y, $C_R(x)$ is the rent cost at x, and $G(x, y)$ is the generalized transportation cost to go from x to y. In order to simplify the discussion, we assume here that employment is located at a unique centre $y = 0$ and that wages and rent costs are of the same order for all individuals. (In fact, most large cities have many different activity centres [187, 188], but this dramatically changes our argument here.) We also assume that the residence location x is given and random; residence choice is obviously a complex problem, and replacing a complex quantity by a random one is a typical assumption made in the statistical physics of complex systems. Within these assumptions, we obtain a simplified Fujita–Ogawa model where we focus on the mode choice; individuals have already a home and work at the central business district located at 0, and the problem is about choosing a transportation mode. Within all these assumptions, the maximization of $Z(x, 0)$ implies the minimization of the transport cost $G(x, 0)$:

$$\max Z(x, 0) \Rightarrow \min G(x, 0) \tag{7.10}$$

where $G(x, 0)$ is the generalized transportation cost from home located at x and the office (located at 0). In other words, individuals will choose in this model (see figure 7.32) the transportation mode that minimizes their commuting costs to go the office. In order to discuss commuting costs, we assume that a proportion p of the population has access (meaning having to walk less than 1 km) to the subway, whereas a share $1 - p$ of the population has no choice but to commute by car. (We assume that all individuals can drive a car if needed.) Even if an individual has access

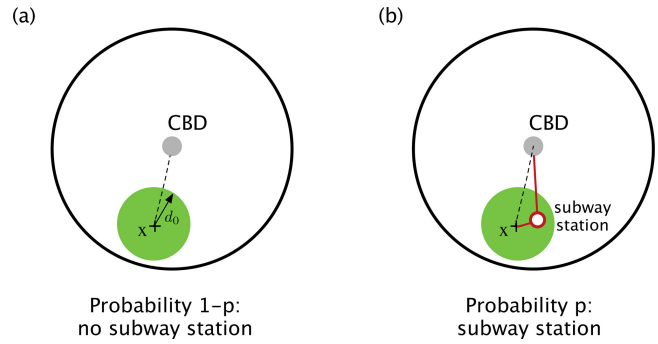

Figure 7.32. Sketch of the model. (a) For a given agent located at a random location x, there is no subway station located at a distance less than d_0 with probability $1 - p$ (in the data used, $d_0 = 1$ km). In this case, the journey to work located at the central business district (CBD) is made by car (dashed line). (b) With probability p there is a subway station in the neighbourhood of x, and the agent has to compare the cost G_{car} of car (dashed line) and the cost G_{subway} of subway (the trip is depicted by the red line) in order to choose the less costly transportation mode to go to the CBD.

to the subway, that does not necessarily mean that it is the mode chosen, and the individual needs to compare the costs G_{car} and G_{subway} in order to choose the least costly transportation mode.

The generalized cost of a mode can be written under the form $G = C_f + V\Delta t$, where C_f is the financial cost (per year, for example) of the mode, while Δt is the time needed for the trip under consideration, and V, which is in units of money per time, is called in this field the 'value of time'. It characterizes the propensity of an individual to choose a rapid transportation mode over a cheaper one. We can even include congestion effects [189] in the trip duration Δt. We can then write an expression for both modes, car and subway, that depends on the distance to go to the centre and different parameters, such as the value of time, the average velocity of car and subway, and the cost of a car. (For the subway, we neglect its monetary cost, which is small in comparison with that for cars. For details see [185].) We can then study the statistics of the minimum cost, and we find the following result [185]. For the 25 megacities in the world that are considered in this study, the subway is always more advantageous than the car. Public transportation is so economical (compared to cars) that people living near rapid transit stations are highly likely to ride them. Thus, traffic does not appear to be a determinant parameter in individual mobility choices, as it concerns mostly individuals who have no choice but to drive and who suffer from onerous commuting costs and unavoidable time-consuming trips as traffic increases. The consequence is then a simple relation between the car traffic T and population P of the form

$$\frac{T}{P} \simeq 1 - p \tag{7.11}$$

which is a nontrivial consequence of rapid transit cheapness and individual choices of mobility. We compare the empirical car modal share $\frac{T}{P}$ for these cities to this

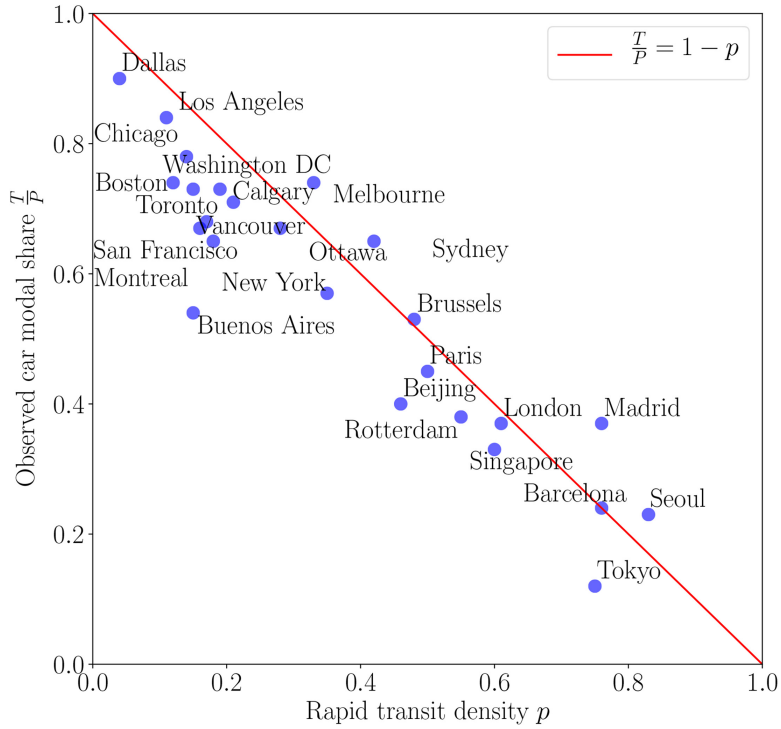

Figure 7.33. Comparison between the observed car modal share T/P and the share of population p living near rapid transit stations (less than 1 km) for 25 metropolitan areas in the world. The red line is the prediction of the model ($R^2 = 0.69$). Given the absence of any tunable parameter, the agreement is satisfactory, and discrepancies are probably mostly due to the existence of other modes of transport (walking or cycling), lower car ownership rates, a higher cost of mass rapid transit, and so on. Figure from [185] CC BY 4.0.

prediction on figure 7.33 and observe, considering the simplicity of the model, very good agreement and a relevant linear trend, highlighting the efficiency of public transportation in reducing traffic.

In particular, most of the European cities are well described by this prediction, though we observe a few deviations. These discrepancies probably have their origin in the existence of other modes of commuting, lower car ownership rates (e.g., in Buenos Aires), lower road capacities, higher cost of subway and mass rapid transit, or a high degree of polycentrism.

This model is also able to provide a prediction for the CO_2 emitted by cars. We will use the simplest assumption, where these emissions are proportional to the total time spent on roads, and we then obtain [185]

$$\frac{Q_{CO_2}}{P} \propto \sqrt{A}(1-p)(1+\tau) \tag{7.12}$$

where P is the population, τ is the average delay due to congestion and is empirically accessible from various databases (e.g., the TomTom database [190]), and A is the

surface area of the urban system. We note that Q_{CO_2} is the product of three main terms: the typical size $\sqrt{A}$ of the city $\times$ the fraction $1-p$ of car drivers $\times$ the congestion effects, which correspond indeed to the intuitive expectation about the main ingredients governing car traffic. We compare this prediction equation (7.6) to disaggregated values of urban CO_2 emissions and observe good agreement [185]. We observe some outliers, such as Buenos Aires, which has a very small car ownership rate and thus lower than expected CO_2 emissions, and New York, which appears to be one of the largest transport CO_2 emitters in the world [191].

This result obtained with a simple model and validated with empirical data illustrates the role played by public transport and traffic in modulating transport-related CO_2 emissions. Most important, we identify urban sprawl ($\sqrt{A}$) here as a major criterion for transport emissions. We note that if we introduce the average population density $\rho = \frac{P}{A}$, we can rewrite equation (7.6) approximately as $\frac{Q_{CO_2}}{P} \propto \rho^{-1/2}$, since $\sqrt{P}$ is a slowly varying function within the scope of large urban areas. We understand here how Newman and Kenworthy [183] could have obtained their result by assuming the density to be the control parameter. However, even if fitting data with a function of ρ is possible, the analysis presented here shows that it is qualitatively wrong; the area size A and the public transport density p seem to be the true parameters controlling car-related quantities such as CO_2 emissions. Mitigating the traffic is therefore not obtained by increasing the density but by reducing the area size and improving the public transport density. Increasing the population in a fixed area would increase the emission of CO_2 (due to an increase of traffic congestion, leading to an increase of τ), in contrast with the naive Newman–Kenworthy assumption where increasing the density leads to a decrease of CO_2 emissions.

The density is therefore not that pivotal, but different factors affect CO_2 emissions. In order to reduce these emissions, this model suggests increasing public transport access either by increasing the density around subway stations or by increasing the density of public transport (in contrast with the conclusions of an econometric study in the US [192]) or reducing the urban area size (impossible in most contexts). Increasing the cost of car use seems actually unable to lower car traffic in the absence of alternative transportation means. This model obviously ignores many parameters, but the main point is to fill a gap in our understanding of traffic in urban areas by using a parsimonious model with the smallest number of parameters and the largest number of predictions in agreement with data. Given the simplicity of the model, we cannot expect perfect agreement, but we may be able to capture correctly all the trends observed empirically and to identify correctly the critical factors for car traffic.

This parsimonious and generic model for the car traffic illustrates how a combination of statistical physics, economic ingredients, and empirical validation can lead to a robust understanding of systems as complex as cities. The next step would then be to understand other crucial quantities, such as the total carbon footprint of a city or the energy use of a city. Empirical studies such as [193] usually

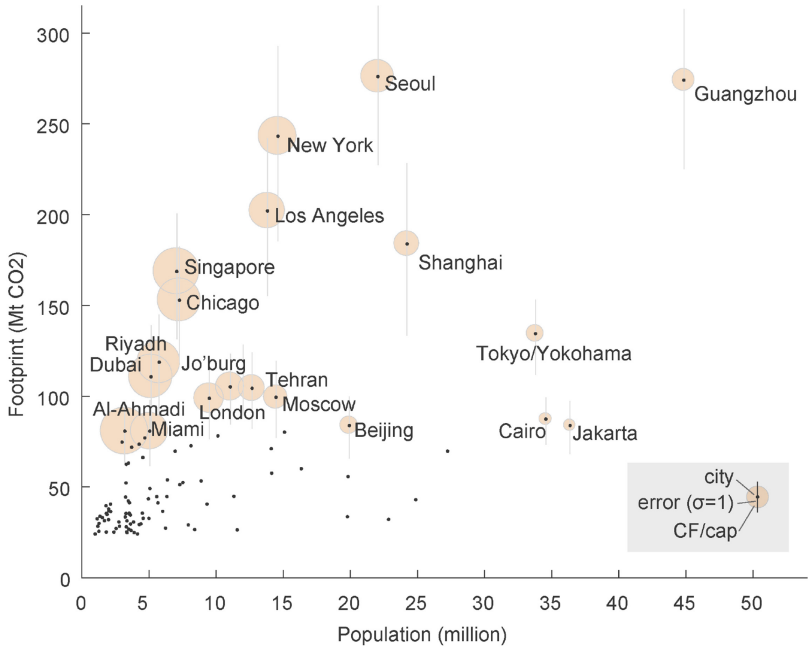

Figure 7.34. Carbon footprint of an urban area versus its population (the named cities are the top 20). The size of the disk corresponds to the carbon footprint per capita. The vertical lines correspond to one standard deviation in the carbon footprint. Figure reproduced from [193]. Copyright The Author(s). Published by IOP Publishing Ltd. CC BY 3.0.

plot the carbon footprint of an urban area (in millions of tons of CO_2) versus population (see figure 7.34).

This figure shows clearly that population is not the only determinant of the carbon footprint. The question is then how we can model this aspect of a city, and the physicist's approach could then help in identifying and understanding what are the critical parameters and their quantitative effects.

Another crucial aspect of cities is their energy use. If the current urban expansion trend continues, energy use will increase more than threefold at the horizon of 2050 [184], leading inevitably to many environmental and economic problems. Modeling this problem could help in devising urban planning and transport policities in order to limit this future increase in urban energy use. An empirical analysis over worldwide cities showed that economic activity, transport costs, geographic factors, and urban form explain 37% of urban direct energy use and 88% of urban transport energy use. As for the carbon footprint, we can illustrate the problem by plotting the energy use versus the GDP per capita (see figure 7.35).

We observe that on average, energy use increases with the economic activity measured by the GDP per capita, followed by a plateau. However, we observe large fluctuations around this average behaviour, showing that very likely the GDP per capita is not the only important factor here. In particular, the authors of [184] observed that higher population density is certainly a positive factor (and,

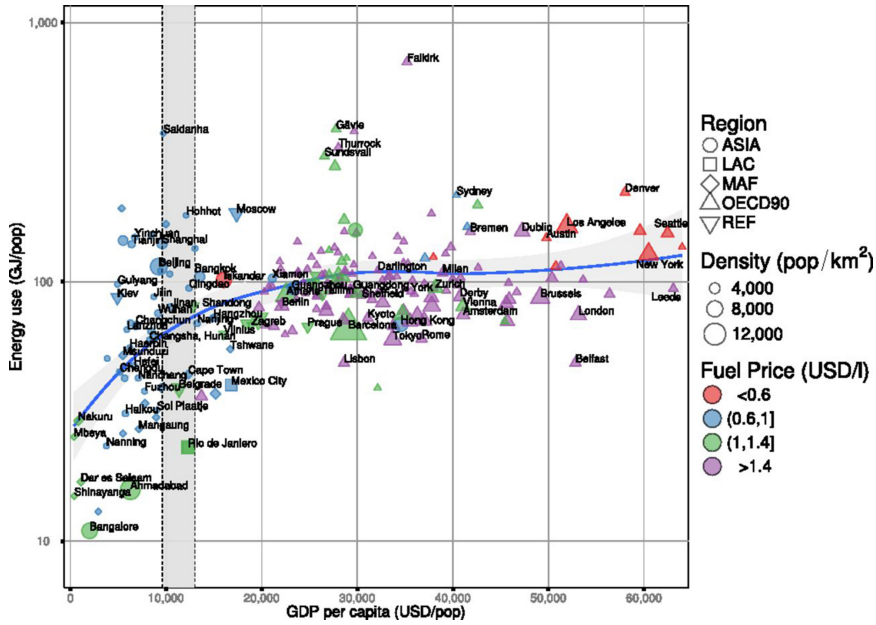

Figure 7.35. Energy use versus economic activity measured by the GDP per capita. Figure reproduced from [184], with permission from PNAS (2015), where more details can be found.

consequently, that developing countries with emerging infrastructures, urban form, and transport planning could avoid larger energy use by increasing local density). Here again, these results challenge our theoretical understanding, and a simple model, in addition to improve our fundamental understanding, could have important practical implications and effects on policies and planning.

7.5.2.3 A transition in urban traffic

The study of traffic and congestion is an old topic and has inspired physicists for a long time. The first works on the subject dealt with traffic on highways, applying various methods coming from fluid mechanics [194] or cellular automata [195]. Concepts such as the carrying capacity of a highway, the formation and progression of congestion on a highway, or the impact of multilane roads with traffic traveling at different speed on the capacity are well understood and described in the literature [196–198]. Different studies [199–201] have conducted experiments on one-lane roads, confirming that the switch from free flow to congestion is brutal and analogous to a phase transition. In [202, 203], models are proposed for the formation of congestion on a regular two-dimensional lattice, displaying a first-order phase transition between free flow and a jammed state.

In contrast, in more complex networks of roads, the mechanisms of congestion spread remain poorly understood. Urban road networks are strongly connected systems, and scaling from single-lane roads to a nonregular two-dimensional network is not an easy task. Agent-based models simulate the flow of cars on the network based on the 'microscopic' interaction between cars on the network and

between cars and the network itself (e.g., traffic lights). This type of simulations requires many assumptions and parameters, and the simulations are difficult to validate. More recent studies have analyzed the urban traffic at a macroscopic level, focusing on the description of the emergence of large-scale congestion as a phase transition. Indeed, it is known that the jamming of a node of the network which happens when the demand exceeds its capacity can have a macroscopic impact, as shown by Echenique *et al* [204] in the context of Internet protocols. However, the exact response of the network depends strongly on the network itself and on the way the information propagates on the network. Understanding the origin and the propagation of congestion on urban road networks is therefore the goal of most recent studies on the subject. At a theoretical level, studies in [205, 206] have proposed arguments to understand the emergence of congestion on networks. In [207, 208], it has been proposed that reaction–diffusion equations are relevant for traffic congestion, where the congestion spreads from a congested link to its neighbors, and Saberi *et al* [209] proposed a model inspired by epidemiological studies to describe the evolution of the number of congested links, regardless of the actual structure of the network.

At a more empirical level, a lot of attention has been given to the study of congestion in the framework of percolation [210], where the main idea is to look at the structure of the set of links at a certain level of congestion (which depends then on the hour of the day). For each date, the authors of [210] divide the network into functional and congested roads, based on a velocity threshold v^*. This leads to clusters of roads functioning with a speed $v > v^*$, separated by congested roads with speeds $v < v^*$. By varying the threshold v^*, they observe a percolation transition (which depends on the hour of the day) with a breakdown of a giant functional cluster into several small clusters. The value of the percolation threshold v_c^* can then be seen as a measure of the state of the network at that date, as it measures effectively the maximal velocity at which one can travel over the main part of the network (described by the giant component).

Generally speaking, a typical marker of a phase transition in classical statistical physics is the divergence of the correlation length close to the critical value of the control parameter (see, e.g., [211]). If there is some 'jamming' transition in urban systems, we should then observe this type of behaviour. The goal is to find an intrinsic marker of a phase transition in the data and to show that congestion in a complex road network during rush hours can be viewed as a jamming transition, going beyond one-dimensional cases. We therefore analyze the correlation between the delays on all roads of the network and identify the correlation length and its variation.

In order to do this, we use hourly traffic data for the city of Paris (France) and study the correlation function of delays [212]. More precisely, we compare roads based on the relative delay experience by users, and for each year and each hour we measure the delay $d_i(t)$ experienced on link i. We also introduce the quantity $T_i(t)$, which denotes the average travel time for that year and hour on this link. The relative delay is then given by $\tau_i(t) = d_i(t)/T_i(t)$ and indicates the importance of

congestion on this link. We consider the correlation function between links i and j of this quantity and measure it for the year y and hour t

$$C_{i,j}(y,\ t) = \langle \tau_i(t)\tau_j(t)\rangle - \langle \tau_i(t)\rangle\langle \tau_j(t)\rangle \tag{7.13}$$

The averages (denoted by the brackets $\langle\cdot\rangle$) are performed over all working days of a given year y and at a given hour t. From this definition, we construct a distance-dependent correlation function by sorting the pairs i, j according to their distance r. After having averaged over all links, we then find in the dataset that the correlation function displays a typical behaviour of the form

$$C(r,\ y,\ t) = \frac{1}{r^{\eta}} \exp\left(-\frac{r}{\xi(y,\ t)}\right) \tag{7.14}$$

where r is the distance between two roads and where $\xi(y,\ t)$ is the correlation length, depending on the hour of the day t and calculated for each year y (in this study, distances, including the correlation length, are converted in the time needed to travel them). This correlation function depends *a priori* on the year and the hour of the day, and its analysis enabled us to extract the correlation length and to examine its variation shown in figure 7.36.

We observe that during night hours, the correlation length is very small, typically of the order of the time needed to travel from a link to two or three links further. Delays appearing on a link during the night propagate to the vicinity of this link only. In sharp contrast, we observe during daytime hours a dramatic increase in the correlation length to values close to 1500 s, and during rush hours, the correlation length displays a peak with values well above 1500 s, which corresponds to a trip whose order of magnitude is the size of the system. Having $\xi \approx 1500$ s indicates a correlation at the scale of the whole system. This divergence is the sign of a transition

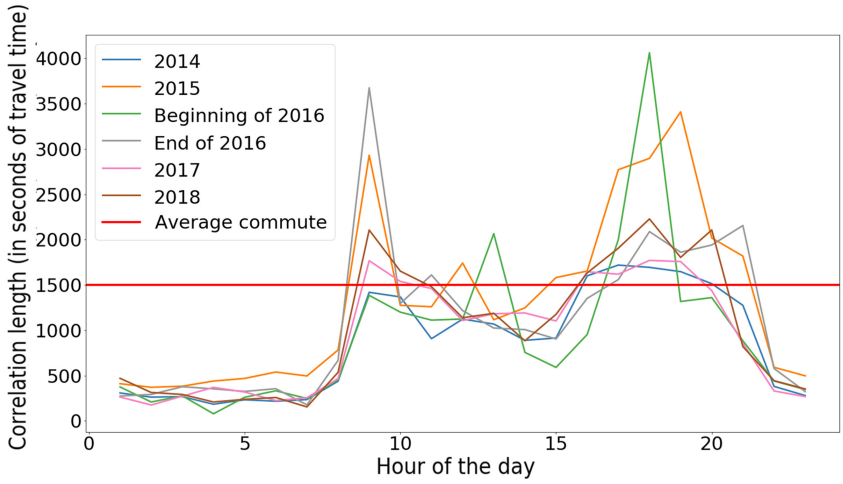

Figure 7.36. Correlation length ξ for the delays on the links of the network, as a function of time and of the average occupancy rate. Figure taken from [212] CC BY 4.0.

that occurs during rush hours with congestion expanding over the whole network. This result therefore suggests the existence of a transition between a state where congestion is localized to another state where congestion occurs over the whole network during rush hours.

This result is important empirical evidence for the existence of a jamming transition in a complex network of roads. However, this points to our lack of understanding of traffic in urban networks and opens an important research direction where the challenge is to construct a consistent theory that allows us to understand this transition but also the macroscopic behaviour of this system and how macroscopic parameters (e.g., the total traffic, the total capacity) are related to each other.

7.5.3 Discussion and perspectives

There is obviously a large number of results and of open and interesting problems about cities, and we could not discuss them all here. Among those, the evolution of infrastructures, the spatial structure of cities, and mobility patterns have made a lot of progress, thanks to new data sources such as mobile phones [213, 214] (see also [182] for a review about mobility). Important studies have also examined processes that take place in cities, such as epidemic modeling in urban areas (see, e.g., [215] and references therein) or resilience in front of flooding or other natural disasters (see, e.g., [216, 217] and references therein). Another particularly important result was obtained in [218], stating that connections between different networks increases the fragility of this multilayered network system. This is particularly relevant for cities where many networks—transportation, energy, water, and so on—are interdependent. Finally, on a larger scale, cities are not isolated but belong to a network of cities (see, e.g., [219] and references therein), and a consistent theoretical approach to cities should integrate this aspect.

We saw in this short review that the increased availability of data has already led to many results and has strengthened our knowledge of the functioning of these complex systems. There are, however, many challenges left. We saw some of them here, but more generally, the most important challenge is probably how to study a complex system such as a city in order to get reliable projections and scenarios. An important part concerns the data availability that could certainly be improved in many countries. So far, data is scattered all over the Internet, and a European or even global initiative such as a common data platform could accelerate research, facilitate reproducibility, and promote international collaborations. From a more theoretical point of view, there are challenges of very different natures. Some are purely mathematical, such as the optimal shape of a transportation network, for example [220], and some others are closer to applications, such as the traffic in urban networks. In all cases, the analysis of data is a crucial component, and we will probably assist in the future in finding hybrid approaches that combine more standard tools coming from applied mathematics or statistical physics with new algorithms, such as machine learning. It seems that we have now all the ingredients —data and tools—for constructing a robust science of cities where fundamental

understanding will be able to help urban planners mitigating the large variety of issues posed by large cities.

Acknowledgments

JB acknowledges support from the project Leading Research Center on Quantum Computing (Agreement No. 014/20).

References

[1] Feynman R P 1982 Simulating physics with computers *Int. J. Theor. Phys.* **21** 467
[2] Nielsen M A and Chuang I L 2011 *Quantum Computation and Quantum Information: 10th Anniversary Edition* 10th edn (New York: Cambridge University Press)
[3] Aaronson S 2013 *Quantum Computing Since Democritus* (New York: Cambridge University Press)
[4] Arute F *et al* 2019 Quantum supremacy using a programmable superconducting processor *Nature* **574** 505
[5] Zhong H-S *et al* 2020 Quantum computational advantage using photons *Science* **370** 1460
[6] Wu Y *et al* 2021 Strong quantum computational advantage using a superconducting quantum processor *Phys. Rev. Lett.* **127** 180501
[7] Preskill J 2018 Quantum computing in the NISQ era and beyond *Quantum* **2** 79
[8] Bharti K *et al* 2021 Noisy intermediate-scale quantum (NISQ) algorithms *Rev. Mod. Phys.* **94** 015004
[9] Montanaro A 2016 Quantum algorithms: an overview *npj Quantum Inf.* **2** 15023
[10] Bauer B, Bravyi S, Motta M and Chan G K-L 2020 Quantum algorithms for quantum chemistry and quantum materials science *Chem. Rev.* **120** 12685
[11] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 Quantum machine learning *Nature* **549** 195
[12] Blais A, Girvin S M and Oliver W D 2020 Quantum information processing and quantum optics with circuit quantum electrodynamics *Nat. Phys.* **16** 247
[13] Blatt R and Roos C F 2012 Quantum simulations with trapped ions *Nat. Phys.* **8** 277
[14] Browaeys A and Lahaye T 2020 Many-body physics with individually controlled Rydberg atoms *Nat. Phys.* **16** 132
[15] Einstein A, Podolsky B and Rosen N 1935 Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47** 777
[16] Bell J S 1964 On the Einstein Podolsky Rosen paradox *Phys. Phys. Fiz.* **1** 195
[17] Aspect A, Grangier P and Roger G 1982 Experimental realization of Einstein–Podolsky–Rosen–Bohm Gedanken experiment: a new violation of Bell's inequalities *Phys. Rev. Lett.* **49** 91
[18] DiVincenzo D P 1995 Two-bit gates are universal for quantum computation *Phys. Rev.* A **51** 1015
[19] Deutsch D 1985 Quantum theory, the Church–Turing principle and the universal quantum computer *Proc. R. Soc. A* **400** 97
[20] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 Optimization by simulated annealing *Science* **220** 671
[21] Deutsch D and Jozsa R 1992 Rapid solution of problems by quantum computation *Proc. R. Soc.* A **439** 553

[22] Simon D 1994 Proc. 35th Annu. Symp. Found. Comput. Sci. (IEEE Comput. Soc. Press) pp 116–23
[23] Shor P 1994 Proc. 35th Annu. Symp. Found. Comput. Sci. (IEEE Comput. Soc. Press) pp 124–34
[24] Kitaev A Y 1995 Quantum measurements and the Abelian Stabilizer Problem ArXiv: quant-ph/9511026
[25] Grover L K 1996 *Proc. 28th Annual ACM Symp. Theory Comput.—STOC'96* (New York: ACM Press) pp 212–9
[26] Brassard G, Hoyer P, Mosca M and Tapp A 2000 Quantum amplitude amplification and estimation ArXiv: 0005055 [quant-ph]
[27] Das A and Chakrabarti B K 2008 Colloquium: quantum annealing and analog quantum computation *Rev. Mod. Phys.* **80** 1061
[28] Tang E 2019 *Proc. 51st Annu. ACM SIGACT Symp. Theory Comput* (New York: ACM) pp 217–28
[29] Wootters W K and Zurek W H 1982 A single quantum cannot be cloned *Nature* **299** 802
[30] Shor P W 1995 Scheme for reducing decoherence in quantum computer memory *Phys. Rev.* A **52** R2493
[31] Steane A M 1996 Error correcting codes in quantum theory *Phys. Rev. Lett.* **77** 793
[32] Shor P W 1996 Fault-tolerant quantum computation ArXiv: quant-ph/9605011
[33] DiVincenzo D P 2000 The physical implementation of quantum computation *Fortschr. Phys.* **48** 771
[34] McMichael R D https://nist.gov/programs-projects/diamond-nv-center-magnetometry (Accessed 19 July 2021)
[35] Wang J *et al* 2018 Multidimensional quantum entanglement with large-scale integrated optics *Science* **360** 285
[36] Ladd T D, Jelezko F, Laflamme R, Nakamura Y, Monroe C and O'Brien J L 2010 Quantum computers *Nature* **464** 45
[37] Cirac J I and Zoller P 1995 Quantum computations with cold trapped ions *Phys. Rev. Lett.* **74** 4091
[38] Shnirman A, Schoen G and Hermon Z 1997 Quantum manipulations of small Josephson junctions *Phys. Rev. Lett.* **79** 2371
[39] Jaksch D, Cirac J I, Zoller P, Rolston S L, Coˆte´ R and Lukin M D 2000 Fast quantum gates for neutral atoms *Phys. Rev. Lett.* **85** 2208
[40] Loss D and DiVincenzo D P 1998 Quantum computation with quantum dots *Phys. Rev.* A **57** 120
[41] Nizovtsev A P 2005 A quantum computer based on NV centers in diamond: optically detected nutations of single electron and nuclear spins *Opt. Spectrosc.* **99** 233
[42] Knill E, Laflamme R and Milburn G J 2001 A scheme for efficient quantum computation with linear optics *Nature* **409** 46
[43] Gershenfeld N A and Chuang I L 1997 Bulk spin-resonance quantum computation *Science* **275** 350
[44] Kane B E 1998 A silicon-based nuclear spin quantum computer *Nature* **393** 133
[45] Kitaev A 2003 Fault-tolerant quantum computation by anyons *Ann. Phys. (N. Y).* **303** 2
[46] Brooke J 1999 Quantum annealing of a disordered magnet *Science* **284** 779
[47] Raussendorf R and Briegel H J 2001 A one-way quantum computer *Phys. Rev. Lett.* **86** 5188

[48] Born M and Jordan P 1925 Zur quantenmechanik *Z. Phys.* **34** 858–88
[49] Benioff P 1980 The computer as a physical system: a microscopic quantum mechanical hamiltonian model of computers as represented by turing machines *J. Stat. Phys.* **22** 563–91
[50] Feynman R P 1985 Quantum mechanical computers *Optics News* **11** 11–20
[51] Yuri M 1980 *Vychislimoe i nevychislimoe [Computable and Uncomputable]* (original in Russian: Soviet Radio)
[52] Deutsch D 1985 Quantum theory, the Church—turing principle and the universal quantum computer *Proc. Royal Soc. Lond. Ser. A, Math. Phys. Sci.* **400** 97–117
[53] Lloyd S 1996 Universal quantum simulators *Science* **273** 1073–8
[54] Deutsch D and Jozsa R 1992 Rapid solution of problems by quantum computation *Proc. Royal Soc. Lond. Ser. A, Math. Phys. Sci.* **439** 553–8
[55] Simon D R 1997 On the power of quantum computation *SIAM J. Comput.* **26** 1474–83
[56] Shor P W 1997 Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer *SIAM J. Comput.* **26** 1484–509
[57] Grover L K 1996 A fast quantum mechanical algorithm for database search Proc. of the 28th Annual ACM Symp. on Theory of Computing pp 212–9
[58] Bennett C H, Bernstein E, Brassard G and Vazirani U 1997 Strengths and weaknesses of quantum computing *SIAM J. Comput.* **26** 1510–23
[59] Aaronson S 2015 Read the fine print *Nat. Phys.* **11** 291–3
[60] Harrow A W, Hassidim A and Lloyd S 2009 Quantum algorithm for linear systems of equations *Phys. Rev. Lett.* **103** 150502
[61] Morales M E S, Tlyachev T and Biamonte J 2018 Variational learning of grover's quantum search algorithm *Phys. Rev.* A **98** 062333
[62] Farhi E, Goldstone J and Gutmann S 2014 A quantum approximate optimization algorithm arXiv:1411.4028
[63] Bravo-Prieto C, LaRose R, Cerezo M, Subasi Y, Cincio L and Coles P J 2020 Variational quantum linear solver arXiv:1909.05820
[64] Lubasch M, Joo J, Moinier P, Kiffner M and Jaksch D 2020 Variational quantum algorithms for nonlinear problems *Phys. Rev.* A **101** 010301
[65] Yung M-H, Casanova J, Mezzacapo A, McClean J, Lamata L, Aspuru-Guzik A and Solano E 2014 From transistor to trapped-ion computers for quantum chemistry *Sci. Rep.* **4** 3589
[66] Peruzzo A, McClean J, Shadbolt P, Yung M-H, Zhou X-Q, Love P J, Aspuru-Guzik A and O'brien J L 2014 A variational eigenvalue solver on a photonic quantum processor *Nat. Commun.* **5** 4213
[67] Biamonte J 2021 Universal variational quantum computation *Phys. Rev.* A **103** L030401
[68] Biamonte J D, Dorozhkin P and Zacharov I 2019 Keep quantum computing global and open *Nature* **573** 190–1
[69] Arute F *et al* 2019 Quantum supremacy using a programmable superconducting processor *Nature* **574** 505–10
[70] Zhong H-S *et al* 2020 Quantum computational advantage using photons *Science* **370** 1460–3
[71] Schumacher B 1995 Quantum coding *Phys. Rev.* A **51** 2738–47
[72] Bruzewicz C D, Chiaverini J, McConnell R and Sage J M 2019 Trapped-ion quantum computing: Progress and challenges *Appl. Phys. Rev.* **6** 021314

[73] Kok P, Munro W J, Nemoto K, Ralph T C, Dowling J P and Milburn G J 2007 Linear optical quantum computing with photonic qubits *Rev. Mod. Phys.* **79** 135–74
[74] Nielsen M A and Chuang I L 2009 *Quantum Computation and Quantum Information* (Cambridge University Press)
[75] Shor P W 1996 Fault-tolerant quantum computation Proc. of 37th Conf. on Foundations of Computer Science (IEEE Comput. Soc. Press)
[76] Shor P W 1995 Scheme for reducing decoherence in quantum computer memory *Phys. Rev.* A **52** R2493–6
[77] Georgescu I 2020 25 years of quantum error correction *Nat. Rev. Phys.* **2** 519 519
[78] Farhi E, Goldstone J, Gutmann S, Lapan J, Lundgren A and Preda D 2001 A quantum adiabatic evolution algorithm applied to random instances of an np-complete problem *Science* **292** 472–5
[79] Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S and Regev O 2008 Adiabatic quantum computation is equivalent to standard quantum computation *SIAM Rev.* **50** 755–87
[80] Childs A M 2009 Universal computation by quantum walk *Phys. Rev. Lett.* **102** 180501
[81] Lovett N B, Cooper S, Everitt M, Trevers M and Kendon V 2010 Universal quantum computation using the discrete-time quantum walk *Phys. Rev.* A **81** 042330
[82] Van den Nest M, Miyake A, Dür W and Briegel H J 2006 Universal resources for measurement-based quantum computation *Phys. Rev. Lett.* **97** 150504
[83] Harrow A W and Montanaro A 2017 Quantum computational supremacy *Nature* **549** 203–9
[84] Preskill J 2018 Quantum computing in the nisq era and beyond *Quantum* **2** 79
[85] Waldrop M M 2016 The chips are down for Moore's law *Nature* **530** 144–7
[86] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 Quantum machine learning *Nature* **549** 195–202
[87] Pande M and Mulay P 2020 Bibliometric survey of quantum machine learning *Sci. Technol. Libr.* **39** 369–82
[88] Seskir Z C and Aydinoglu A U 2021 The landscape of academic literature in quantum technologies *Int. J. Quantum Inform.* **19** 2150012
[89] van Eck N J and Waltman L 2009 Software survey: Vosviewer, a computer program for bibliometric mapping *Scientometrics* **84** 523–38
[90] Watkins G P and Knight F H 1922 Knight's risk, uncertainty and profit *Quart. J. Econ.* **36** 682
[91] DiVincenzo D P 2017 Scientists and citizens: getting to quantum technologies *Ethics Inform. Technol.* **19** 247–51
[92] Möller M and Vuik C 2017 On the impact of quantum computing technology on future developments in high-performance scientific computing *Ethics Inform. Technol.* **19** 253–69
[93] Coenen C and Grunwald A 2017 Responsible research and innovation (rri) in quantum technology *Ethics Inform. Technol.* **19** 277–94
[94] de Wolf R 2017 The potential impact of quantum computers on society *Ethics Inform. Technol.* **19** 271–6
[95] Johnson W S 1883 Electric tele-thermoscop *US Patent* 281,884
[96] Schütze A, Helwig N and Schneider T 2018 Sensors 4.0—smart sensors and measurement technology enable industry 4.0 *J. Sensors Sensor Syst.* **7** 359–71

[97] Bentley J P 1995 *Principles of Measurement Systems* (New York: Longman Publishing Group) 3rd edn
[98] Wilson J S 2005 *Sensor Technology Handbook* (Amsterdam: Elsevier)
[99] Taylor J R 1996 *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (Mill Valley, CA: University Science Books)
[100] Fink J K 2012 *Polymeric Sensors and Actuators* (New York: Wiley)
[101] Simmons E E 1940 Material testing apparatus *US Patent* 2,350,972
[102] Ruge A C 1939 Strain gauge *US Patent* 2,350,972
[103] Wasley R J, Hoge K G and Cast J C 1969 Combined strain gauge—quartz crystal instrumented hopkinson split bar *Rev. Sci. Instrum.* **40** 889–94
[104] Barbour N and Schmidt G 2001 Inertial sensor technology trends *IEEE Sens. J.* **1** 332–9
[105] Narayanaswamy R and Wolfbeis O S 2004 *Optical Sensors* (Berlin: Springer)
[106] Santos J and Farahi F 2014 *Handbook of Optical Sensors* (Boca Raton, FL: CRC Press)
[107] Jain Y K, Alex T K and Kalakrishnan B 1980 Ultimate ir horizon sensor *IEEE Trans. Aerospace Electron. Syst.* **16** 233–8
[108] Casas J R and Cruz P J S 2003 Fiber optic sensors for bridge monitoring *J. Bridge Eng.* **8** 362–73
[109] García I, Zubia J, Durana G, Aldabaldetreku G, Illarramendi M A and Villatoro J 2015 Optical fiber sensors for aircraft structural health monitoring *Sensors (Switzerland)* **15** 15494–519
[110] Udd E and Spillman W B 2011 *Fiber Optic Sensors: An Introduction for Engineers and Scientists* 2nd edn (New York: Wiley)
[111] Lee B H, Kim Y H, Park K S, Eom J B, Kim M J, Rho B S and Choi H Y 2012 Interferometric fiber optic sensors *Sensors* **12** 2467–86
[112] Baldini F and Mignani A G 2002 Optical-fiber medical sensors *MRS Bull.* **27** 383–7
[113] Baltes H P and Popovic R S 1986 Integrated semiconductor magnetic field sensors *Proc. IEEE* **74** 1107–32
[114] Janata J 2009 *Principles of Chemical Sensors* (New York: Springer US)
[115] Bănică F-G 2012 *Chemical Sensors and Biosensors: Fundamentals and Applications* (New York: Wiley)
[116] Baron R and Saffell J 2017 Amperometric gas sensors as a low cost emerging technology platform for air quality monitoring applications: a review *ACS Sens.* **2** 1553–66
[117] Thévenot D R, Toth K, Durst R A and Wilson G S 2001 Electrochemical biosensors: Recommended definitions and classification *Biosens. Bioelectron.* **16** 121–31
[118] United Nations 2015 *Transforming Our World: The 2030 Agenda for Sustainable Development* (New York: United Nations)
[119] A European green deal–striving to be the first climate neutral continent 2021 (https://ec.europa.eu/info/strategy/priorities-2019–2024/european-green-deal_en)
[120] Ho C K, Robinson A, Miller D R and Davis M J 2005 Overview of sensors and needs for environmental monitoring *Sensors* **5** 4–37
[121] Lewis A and Edwards P 2016 Validate personal air-pollution sensors *Nature* **535** 29–31
[122] Barabde M and Danve S 2015 Real time water quality monitoring system *Int. J. Innov. Res. Comput. Commun. Eng.* **3** 5064–9
[123] Pasika S and Gandla S T 2020 Smart water quality monitoring system with cost-effective using iot *Heliyon* **6** E04096

[124] Groves P D 2015 Principles of gnss, inertial, and multisensor integrated navigation systems *IEEE Aerosp. Electron. Syst. Mag.* **30** 26–7
[125] Kessler D J, Johnson N L, Liou J-C and Matney M 2010 The kessler syndrome: Implications to future space operations *Adv. Astron. Sci.* **137** 61
[126] Wang Q, Zhu X, Ni Y, Gu L and Zhu H 2020 Blockchain for the IoT and industrial IoT: a review *Internet of Things* **10** 100081
[127] UN Online Index of Objects Launched into Outer Space (https://unoosa.org/oosa/osoindex/search-ng.jspx?lf_id=)
[128] 2020 Global Space Economy at a Glance (https://brycetech.com/reports)
[129] Measuring the Ecomomic Impact of the Space Sector, OECD Background Paper for the G20 Space Economy Leaders' Meeting, October 7, 2020
[130] Dependence of the European Economy on Space Infrastructures, Potential Impacts of Space Assets Loss, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, European Commission
[131] Akrami Y *et al* 2020 Planck 2018 results: overview and the cosmological legacy of Planck *Astron. Astrophys.* **641** A1
[132] Voyage 2050: Long-Term Planning of the ESA Science Programme. White Papers. (https://cosmos.esa.int/web/voyage-2050/white-papers)
[133] Voyage 2050 Sets Sail: ESA Chooses Future, ESA—Voyage 2050 Sets Sail: ESA Chooses Future Science Mission Themes
[134] Exploration Mission-1 Map, NASA (https://nasa.gov/image-feature/exploration-mission-1-map)
[135] Moonlight: Connecting Earth with the Moon (https://esa.int/Applications/Telecommunications_Integrated_Applications/Lunar_satellites)
[136] Allen C *et al* 2012 Challenges of a Mars simple return mission from the samples' perspective —contamination, preservation, and planetary protection *Concepts and Approaches for Mars Exploration* June 12–14, 2012, Houston, Texas
[137] Moon Village, A Vision for Global Cooperation and Space 4.0 (https://esa.int/About_Us/Ministerial_Council_2016/Moon_Village)
[138] *Melissa, Micro Ecological Life Support System Alternative* (https://melissafoundation.org/page/melissa-pilot-plant)
[139] Protection of Frequencies for Radioastronomical Measurements in the Shielded Zone of the Moon, RECOMMENDATION ITU-R RA.479–5, ITU Radiocommunication Assembly
[140] Barhels M 2019 *Radio Telescope Unfurls 3 Antennas Beyond the Far Side of the Moon* 2 December (https://space.com/radio-telescope-beyond-moon-far-side-antennas-deploy.html)
[141] Global Exploration Roadmap, Supplement Augsut 2020, Lunar Surface Exploration Scenario Update, ISECG, August 2020 (https://globalspaceexploration.org/wp-content/uploads/2020/08/GER_2020_supplement.pdf)
[142] *The Global Risk Report 2021* 16th edn (http://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2021.pdf)
[143] *UCS (Union of Concerned Scientists) Satellite Database*, Jan 2019 (https://ucsusa.org/nuclear-weapons/space-weapons/satellite-database?_ga=2.21168895.576778038.1554626468-1092692463.1554216706#.XDZDs2l7ksc)
[144] *ESA Budget 2021* (https://esa.int/Newsroom/ESA_budget_2021)
[145] *Essential Climate Variables, World Metereological Organisation* (https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables)

[146] *EU Climate Strategies & Targets, 2050 Long-Term Strategy* (https://ec.europa.eu/clima/policies/strategies/2050_en)
[147] *The European Green Deal, Communication from the Commission to the European Parliament, The Euroepan Council, The Council, The European Economica and Social Committee and the Committee of the Regions* (https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1596443911913&uri=CELEX:52019DC0640#document2)
[148] *A Brief Outlook on Future Copernicus Missions* (https://sentinels.copernicus.eu/web/sentinel/missions/copernicus-expansion-missions)
[149] *Contracts Awarded for Development of Six New Copernicus Missions* (https://esa.int/Applications/Observing_the_Earth/Copernicus/Contracts_awarded_for_development_of_six_new_Copernicus_missions)
[150] *ESA Earth Explorers: Understanding Our Planet* (https://esa.int/Applications/Observing_the_Earth/Future_EO/Earth_Explorers/About_Earth_Explorers2)
[151] *ESA and NASA Join Forces To Understand Climate Change* (https://esa.int/Applications/Observing_the_Earth/ESA_and_NASA_join_forces_to_understand_climate_change)
[152] Avila-Rodriguez J-A *et al* 2006 The MBOC modulation: the final touch to the galileo frequency and signal plan *Proc. of the Int. Technical Meeting of the Institute of Navigation (Fort Worth, TX, 25–28 September)* ION-GNSS 2006
[153] *Galileo's Contribution to COSPAS-SARSAT Programme* (https://ec.europa.eu/growth/sectors/space/galileo/sar_en)
[154] Galileo High-Accuracy Services 2020 *Information Note* GSA (https://gsc-europa.eu/sites/default/files/sites/all/files/Galileo_HAS_Info_Note.pdf)
[155] *Assuring Authentication for all, European Global Navigation Satellite Systems Agency, GSA* (https://gsc-europa.eu/news/assuring-authentication-for-all)
[156] Schuldt T *et al* 2021 Optical clock technologies for global navigation satellite systems *GPS Solutions* **25** 83
[157] Navigation system based on neutrino detection, US. Patent, US 8,849,565 B1, 2014
[158] The Current State of Space Debris (https://esa.int/Safety_Security/Space_Debris/The_current_state_of_space_debris)
[159] Space Surveillance Network (SSN), U.S. Army, Navy and Air Force (http://au.af.mil/au/awc/awcgate/usspc-fs/space.htm)
[160] Krag H *et al* 2017 A 1 cm space debris impact onto the Sentinel-1A Solar Array *Acta Astronaut.* **137** 434–43
[161] Kessler D *et al* 1978 Collision frequency of artificial satellites: the creation of a debris belt *J. Geophys. Res.* **83** 2637–46
[162] *IADC Space Debris Mitigation Guidelines*, IADC-02-01 Revision 1 September 2007
[163] *ESA Purchases World-First Debris Removal Mission from Start-up* (https://esa.int/Safety_Security/ESA_purchases_world-first_debris_removal_mission_from_start-up)
[164] *Planetary Defence* (www.esa.int/Safety_Security/Hera/Planetary_defence)
[165] *ESA Sfatey & Security Missions: Plans for the Future* (www.esa.int/Safety_Security/Plans_for_the_future)
[166] Euroconsult December 2019 *The Space Economy Report 2019*
[167] Crane K W *et al* 2020 *Measuring the Space Economy: Estimating the Value of Economic Activities in and for Space* (Science &Technology Policy Institute)
[168] *The Future of the European Space Sector How to Leverage Europe's Technological Leadership and Boost Investments for Space Ventures* (European Investment Bank, 2019)

[169] Geohistorical data website https://geohistoricaldata.org
[170] Fujita M, Krugman P R and Venables A J 2001 *The Spatial Economy: Cities, Regions, and International Trade* (Cambridge, MA: MIT Press)
[171] Barthelemy M 2016 *The Structure and Dynamics of Cities* (Cambridge: Cambridge University Press)
[172] Batty M 2013 *The New Science of Cities* (Cambridge, MA: MIT Press)
[173] O'Sullivan D and Manson S M 2015 Do physicists have geography envy? and what can geographers learn from it? *Ann. Assoc. Am. Geogr.* **105** 704–22
[174] Auerbach F 1913 Das gesetz der bevölkerungskonzentration *Petermanns Geogr. Mitt.* **59** 74–6
[175] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
[176] Cottineau C 2017 Metazipf. a dynamic meta-analysis of city size distributions *PLoS One* **12** e0183919
[177] Verbavatz V and Barthelemy M 2020 The growth equation of cities *Nature* **587** 397–401
[178] Batty M 2006 Rank clocks *Nature* **444** 592–6
[179] Seto K C, Fragkias M, Güneralp B and Reilly M K 2011 A meta-analysis of global urban land expansion *PLoS One* **6** e23777
[180] Seto K C, Güneralp B and Hutyra L R 2012 Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools *Proc. Natl Acad. Sci.* **109** 16083–8
[181] Atlas of urban expansion 2024 http://atlasofurbanexpansion.org
[182] Barbosa H, Barthelemy M, Ghoshal G, James C R, Lenormand M, Louail T, Menezes R, Ramasco J J, Simini F and Tomasini M 2018 Human mobility: models and applications *Phys. Rep.* **734** 1 74
[183] Newman P W G and Kenworthy J R 1989 Gasoline consumption and cities: a comparison of us cities with a global survey *J. Am. Plan. Assoc.* **55** 24–37
[184] Creutzig F, Baiocchi G, Bierkandt R, Pichler P-P and Seto K C 2015 Global typology of urban energy use and potentials for an urbanization mitigation wedge *Proc. Natl Acad. Sci.* **112** 6283–8
[185] Verbavatz V and Barthelemy M 2019 Critical factors for mitigating car traffic in cities *PLoS One* **14** e0219559
[186] Fujita M and Ogawa H 1982 Multiple equilibria and structural transition of non-monocentric urban configurations *Reg. Sci. Urban Econ.* **12** 161–96
[187] Louf R and Barthelemy M 2013 Modeling the polycentric transition of cities *Phys. Rev. Lett.* **111** 198702
[188] Louf R and Barthelemy M 2014 How congestion shapes cities: from mobility patterns to scaling *Sci. Rep.* **4** 5561
[189] Branston D 1976 Link capacity functions: a review *Transp. Res.* **10** 223–36
[190] TomTom International BV (2008–2017) TomTom Traffic Index
[191] OECD 2016 The OECD Metropolitan Areas Database visualized through the Metropolitan eXplorer
[192] Duranton G and Turner M A 2011 The fundamental law of road congestion: evidence from us cities *Am. Econ. Rev.* **101** 2616–52
[193] Moran D, Kanemoto K, Jiborn M, Wood R, Többen J and Seto K C 2018 Carbon footprints of 13000 cities *Environ. Res. Lett.* **13** 064041

[194] Lighthill M J and Whitham G B 1955 On kinematic waves ii. a theory of traffic flow on long crowded roads *Proc. Royal Soc. Lond. Ser. A. Math. Phys. Sci.* **229** 317–45
[195] Nagel K and Schreckenberg M 1992 A cellular automaton model for freeway traffic *J. Phys.* I **2** 2221–9
[196] Blandin S, Argote J, Bayen A M and Work D B 2013 Phase transition model of non-stationary traffic flow: definition, properties and solution method *Transp. Res.* B **52** 55
[197] Gartner N H, Messer C J and Rathi A 2002 Traffic flow theory—a state-of-the-art report: revised monograph on traffic flow theory https://rosap.ntl.bts.gov/view/dot/35775
[198] Nagel K and Paczuski M 1995 Emergent traffic jams *Phys. Rev.* E **51** 2909
[199] Sugiyama Y, Fukui M, Kikuchi M, Hasebe K, Nakayama A, Nishinari K, Tadaki S-ichi and Yukawa S 2008 Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam *New J. Phys.* **10** 033001
[200] Tadaki S-ichi, Kikuchi M, Fukui M, Nakayama A, Nishinari K, Shibata A, Sugiyama Y, Yosida T and Yukawa S 2013 Phase transition in traffic jam experiment on a circuit *New J. Phys.* **15** 103034
[201] Tadaki S-ichi, Kikuchi M, Nakayama A, Shibata A, Sugiyama Y and Yukawa S 2016 Characterizing and distinguishing free and jammed traffic flows from the distribution and correlation of experimental speed data *New J. Phys.* **18** 083022
[202] Biham O, Middleton A A and Levine D 1992 Self-organization and a dynamical transition in traffic-flow models *Phys. Rev.* A **46** R6124
[203] Cuesta J A, Martnez F C, Molera J M and Sánchez A 1993 Phase transitions in two-dimensional traffic-flow models *Phys. Rev.* E **48** R4175
[204] Echenique P, Gómez-Gardenes J and Moreno Y 2005 Dynamics of jamming transitions in complex networks *EPL (Europhys. Lett.)* **71** 325
[205] Carmona H A, de Noronha A W T, Moreira A A, Araújo N A M and Andrade J S 2020 Cracking urban mobility *Phys. Rev. Res.* **2** 043132
[206] Lampo A, Borge-Holthoefer J, Gómez S and Solé-Ribalta A 2021 Multiple abrupt phase transitions in urban transport congestion *Phys. Rev. Res.* **3** 013267
[207] Bellocchi L and Geroliminis N 2020 Unraveling reaction-diffusion-like dynamics in urban congestion propagation: Insights from a large-scale road network *Sci. Rep.* **10** 1–11
[208] Jiang Y, Kang R, Li D, Guo S and Havlin S 2017 Spatio-temporal propagation of traffic jams in urban traffic networks *arXiv preprint* arXiv:1705.08269
[209] Saberi M, Hamedmoghadam H, Ashfaq M, Hosseini S A, Gu Z, Shafiei S, Nair D J, Dixit V, Gardner L and Waller S T *et al* 2020 A simple contagion process describes spreading of traffic jams in urban networks *Nat. Commun.* **11** 1–9
[210] Li D, Fu B, Wang Y, Lu G, Berezin Y, Stanley H E and Havlin S 2015 Percolation transition in dynamical traffic network with evolving critical bottlenecks *Proc. Natl Acad. Sci.* **112** 669–72
[211] Kadanoff L P 2000 *Statistical Physics: Statics, Dynamics and Renormalization* (Singapore: World Scientific)
[212] Taillanter E and Barthelemy M 2021 Empirical evidence for a jamming transition in urban traffic *J. R. Soc. Interface* **18** 20210391
[213] Louail T, Lenormand M, Cantu Ros O G, Picornell M, Herranz R, Frias-Martinez E, Ramasco J J and Barthelemy M 2014 From mobile phone data to the spatial structure of cities *Sci. Rep.* **4** 1–12

[214] Ratti C, Frenchman D, Pulselli R M and Williams S 2006 Mobile landscapes: using location data from cell phones for urban analysis *Environ. Plan. B: Plan. Design* **33** 727–48
[215] Chang S, Pierson E, Koh P W, Gerardin J, Redbird B, Grusky D and Leskovec J 2021 Mobility network models of Covid-19 explain inequities and inform reopening *Nature* **589** 82–7
[216] Ganin A A, Kitsak M, Marchese D, Keisler J M, Seager T and Linkov I 2017 Resilience and efficiency in transportation networks *Sci. Adv.* **3** e1701079
[217] Wang W, Yang S, Stanley H E and Gao J 2019 Local floods induce large-scale abrupt failures of road networks *Nat. Commun.* **10** 1–11
[218] Buldyrev S V, Parshani R, Paul G, Stanley H E and Havlin S 2010 Catastrophic cascade of failures in interdependent networks *Nature* **464** 1025–8
[219] Sanders L, Pumain D, Mathian H, Guérin-Pace F and Bura S 1997 Simpop: a multiagent system for the study of urbanism *Environ. Plan. B: Plan. Design* **24** 287–305
[220] Aldous D and Barthelemy M 2019 Optimal geometry of transportation networks *Phys. Rev.* E **99** 052303