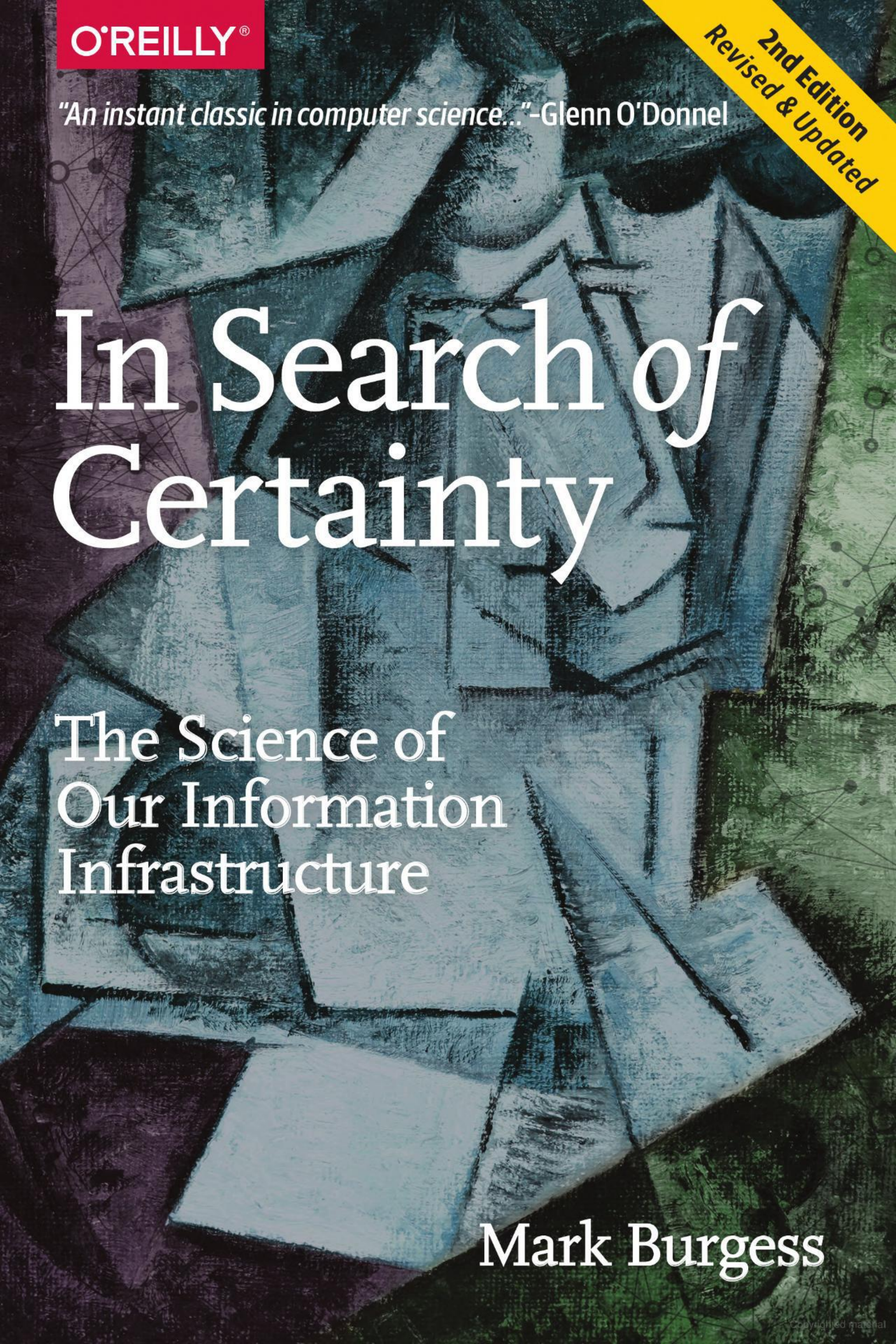

O'REILLY®

"An instant classic in computer science..."–Glenn O'Donnel

2nd Edition
Revised & Updated

In Search of Certainty

The Science of Our Information Infrastructure

Mark Burgess

In Search of Certainty

The science of our information infrastructure

Mark Burgess

In Search of Certainty, Second Edition

by Mark Burgess

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Mike Loukides and Brian Anderson

Proofreader: Jasmine Kwityn
Cover Designer: Edie Freedman

September 2013: First Edition
April 2015: Second Edition

Revision History for the First Edition

September 2013: First edition published in ebook format by χtAxis Press, Oslo, Norway
January 2014 and March 2014: Reprinted with corrections

Revision History for the Second Edition

April 2015: First Release

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights. Quotations used within this work are referenced from public sources, and pose no known commercial infringement. Figures from academic papers used by permission of the authors.

978-1-491-92307-8

[LSI]

Contents

Part III: Promises

Introduction: The Hallowed Halls

"Our great computers
Fill the hallowed halls"

– Neil Peart, Rush, 2112

Quite soon, the world's information infrastructure is going to reach a level of scale and complexity that will force scientists and engineers to think about it in an entirely new way. The familiar notions of command and control, which have held temporary dominion over our systems, are now being thwarted by the realities of a faster, denser world of communication, a world where choice, variety, and indeterminism rule. The myth of the machine, that does exactly what we tell it, has come to an end.

Many of the basic components of information infrastructure, including computers, storage devices, networks, and so forth, are by now relatively familiar; however, each generation of these devices adds new ways to adapt to changing needs, and what happens inside them during operation has influences that originate from all over the planet. This makes their behaviour far from well understood even by the engineers who build and use them. How does that affect predictability, their usefulness?

It is now fair to draw a parallel between the structures we build to deliver information, and the atomic structure of materials that we engineer for manufacturing, like metals and plastics[1]. Although these things seem far removed from one another, occupying very different scales, they have a basic similarity: physical materials are made of atoms and molecules networked through chemical bonding into structures that give them particular properties. Similarly, information infrastructures are made up of their own components wired into networks, giving them particular properties. Understanding what those properties are, and why they come about, requires a new way of thinking. In this book, I will attempt to explain why this analogy is a fair one, why it is not the full story, but what we can learn from the limited similarity.

At the engineering level, we put aside the details of *why* a structure behaves as it does, and rather make use of what properties the materials promise. We become more concerned with how to use them, as 'off-the-shelf' commodities, and describe their promises in terms of new terms like strength, elasticity, plasticity, and so on, qualities far removed from raw atoms and chemical bonds. The useful properties of materials and infrastructure lie as much in the connections between parts, as in the parts that get connected, but we prefer to think of them as continuous reliable stuff, not assemblages of tiny pieces.

The parallels between information infrastructure and the physics of matter go deeper than these superficial likenesses, and I have studied some of these parallels in my research over the past twenty years. This book will describe a few of them. Understanding this new physics of information technology is going to be vital if we are to progress beyond the current state of the art. This is not the physics of silicon wafers, or of electronics, nor is it about ideas of software engineering; rather, it's about a whole new level of description, analogous to the large scale thermodynamics that enabled the steam age, or the economic models of our financial age. Only then, will we be able to build predictably and robustly, while adapting to the accelerating challenges of the modern world.

What does this mean to you and me? Perhaps more than we might think. Every search we perform on the Internet, every music download piped to our earplugs from an embedded mobile hotspot, or more speculatively every smartphone-interfacing self-driving hybrid car purchased with a smart-chip enabled credit card, sends us careering ever onwards in the uncontrolled descent of miniaturization and high density information trawling that we cheer on as The Information Revolution.

We no longer experience information-based technology as thrilling or unusual; the developed world has accepted this trajectory, expects it, and even demands it. In the developing world, it is transforming money, communications and trade in areas where more traditional alternatives failed for a lack of physical infrastructure. The ability to harness information, to process it, and at ever greater speeds, allows the whole world to fend off threats, and exploit opportunities.

As information technology invades more and more parts of our environment, we are going to experience it in unexpected ways. We won't always see the computer in a box, or the remote control panel, but it will be information technology nevertheless. New smart materials and biologically inspired buildings already hint at what the future might look like. Yet, if we are to trust it in all of its varied forms, we have to understand it better.

For years, we've viewed information technology as a kind of icing on our cake, something wonderful that we added to the mundane fixtures of our lives, with its entertainment systems and personal communications channels; but, then these things were no longer the icing, they were the cake itself. Information systems began to invade every aspect of our environments, from the cars we drive to the locks on the front door, from the way we read books to the way we cook food. We depend on it for our very survival.

Artificial environments support more of our vital functions for each day that passes. Few in the developed world can even remember how to survive without the elaborate infrastructure of electricity, supply networks, sanitation plants, microwave ovens, cars and other utilities. So many of us rely on these for their day to day lives. Some have hearts run by pacemakers, others are reliant on technology for heating or cooling. Our very sense of value and trade is computed by technology, which some parts of our planet would consider magic. We 'outsource' increasing amounts of our survival to a 'smart' environmental infrastructure, and hence we become more and more dependent on it for each day that passes.

What makes us think we can rely on all this technology? What keeps it together today, and how might it work tomorrow? Will we even know how to build the next generation of it, or will be become lulled into a stupor of dependence brought about by its conveniences? To shape the future of technology, we need to understand how it works, else what we don't understand will end up shaping us.

As surely as we have followed the trajectory, we have come to rely on it, and thus we must know what it takes to make the next steps, and why. Some of those details are open to choice, others are constrained by the physical nature of the world we live in, so we'll need to understand much more than just a world of computers to know the technology of tomorrow.

Behold the Great Transformation, not just of technology but of society itself, adapting to its new symbiosis! It is happening now, at a datacentre near you! (Or in a test tube, or under an etching laser.) Vast halls of computing power, and laboratories of microscopic bio-chemical machinery, have supplanted the mechanisation and civil engineering of the industrial age, as the darlings of change. What will be the next thing? Nanotechnology? Human enhancement? Where will our sense of adventure set out next?

In 1997, I visited San Diego, California, for the 11th Annual Conference in Large Installation System Administration, as something of an outsider. It was a conference not for the designers of information systems, but for those who keep such systems running, on behalf of society at large. As a physicist by training, relatively new to technological research, I was excited to present some work I'd been doing on new ways to ensure the reliability of large computer systems. I presented a rather technical paper on a new kind of smart process-locking mechanism to make computers more predictable and maintainable.

To a physicist, reliability looks like a kind of stability. It is about enabling an equilibrium to come about. To me, the work I presented was just a small detail in a larger and more exciting discussion to make computer systems self-governing, as if they were as ordinary a part of our infrastructure as as the self-regulating ventilation systems. The trouble was, no one was having this discussion. I didn't get the response I was hoping for. A background in science had not prepared me for an audience with different expectations. In the world of computers, people still believed that you simply tell computers what to do, and, because they are just machines, they must obey.

I left the conference feeling somewhat misunderstood. My paper had been related to a piece of software called CFEngine that I had started developing in 1993 to configure and maintain computers without human intervention. It had become unexpectedly popular, spreading to millions of computers in datacentres and small environments, and it is still widely used today.

On the plane going home, my misery was compounded by becoming ill, and began to think about the human immune system and how smart it seemed to be in repairing a state of health. There was an answer! I became inspired to explain my work in terms of the analogy of health, and I spent much of the year thinking about how to write a paper 'Computer Immunology' which I submitted to the next conference, spelling out a manifesto for building self-healing computers[2].

The following year, 1998, the conference was in Boston, Massachusetts. This time, I was delighted to win the prize for best paper at the conference, and was immediately thrust into a world keen to know about the concepts of self-regulating, self healing computers—though it would take another ten years for the ideas to become widely recognized. The experience underlined the importance of bridging the awareness gap between cultures in different fields, even in science and technology. It underlined, perhaps, a need for books like this one.

After the conference, I was taken on a trip of honour by a former colleague Demosthenes Skipitaris from Oslo University College, to a high security, state-of-the-art datacentre facility run by a Norwegian search engine called FAST, just outside of Boston. After surrendering my passport on the way for security validation, and being frisked by over-dressed guards, we were led into a vast hall of computer racks the size of several football pitches.

Computers on top of computers, accompanied by the deafening noise of thousands of fans whirring, disks spinning and air blowing, all mounted in racks and stacked up to the ceiling. Row upon row of black boxes, separated by narrow walk-spaces, just wide enough for a person, for as far as the eye could see. We were listening to the roar of all the web searches of people from across the world being processed before our eyes, but there were no humans in sight. In the whole building I saw a single keyboard and a single screen for emergency use[3].

“All this is run by your software CFEngine,” my host told me. CFEngine is essentially a collection of software robots embedded into each machine’s operating system. I told him about my computer health analogy, and he commented that with the software running, the most common failure in the machines was that the vibration from all the computer disks would cause the removable disks to work their way out of the chassis, causing a machine to stop now and then. That was the only time a human needed to touch the machines—just push the disk back in and restart.

Then, as we passed one of the racks, he pointed to a small cable emerging from a socket. “That,” he said, “is our single point of failure. If I pull that plug, we’re offline.” We stopped there for a moments to pay our respects to the gods of fragility.

It was a telling reminder that, even with the most advanced systems at our fingertips, the smallest detail can so easily be overlooked and result in a fatal flaw.

How would we even know about a tiny error, a minute design flaw, an instability waiting to grow into a catastrophic failure? Standing amongst the anonymous array of whirring machines in that hallowed hall, it was evident that finding a needle in a haystack might be easy by comparison. In a world of software, there is nothing to even get hold of and feel a prick of a needle.

In 2012, I visited a datacentre ten times the size of the one in Boston, and was shown a room where 40 humans still sat and watched screens, in what looked like an enactment of NASA’s mission control. They were hoping to see some advance warning of signs of trouble in their global operations, still using the methods of a decade before in a pretence of holding together a system that was already far beyond their ability to comprehend—as if a few shepherds were watching over all of the wildlife in Earth’s oceans with their crooks. Those 40 humans watched with naked eye graphs of performance, somewhat like medical monitors for tens of thousands of computers spread around the globe.

I recall thinking how primitive it all was. If a single machine amongst those tens of thousands were to develop an instability, it could bring everything to a halt, in the worst case, like pulling out an essential plug. Watching those screens was like trying to locate a single malignant cell in a patient's body just by measuring a pulse. The mismatch of information was staggering. I began to wonder what role CFEngine's immune principles already played in preventing that from happening. I hope that this book can help to shed some light on what makes a system well-behaved or not.

For two decades, the world's most advanced datacentres have been run largely by automated robotic software that maintains their operational state with very little human intervention. Factories for manufacturing are manned by robot machines, and familiar services like banking, travel, and even vending machines have been automated and made more widely available than ever before. The continuity of these services has allowed us to trust them and rely on them.

How is this even possible? How is it that we can make machines that work without the need to coax and cajole them, without the need to tell them every detail of what to do? How can we trust these machines? And can we keep it up? So far, we've been lucky, but the long term answers have yet to be revealed. They can only emerge by knowing the science behind them.

If information systems are going to be mission critical in the society of today and tomorrow, then the mission controls of this increasingly 'smart' infrastructure need principles more akin to our autonomous immune systems than to nurses with heart monitors. We have to understand how continuous operation, how dependability itself, can emerge from the action of lots of individual cellular parts, and follow a path of continuous adaptation and improvement. To even grasp a knowledge of such speed, scale and complexity is beyond any human without technological assistance.

We build software systems every day, and extend the depth of this dependency on technology. We suffer sometimes from the hubris of believing that control is a matter of applying sufficient force, or a sufficiently detailed set of instructions. Or we simply hope for the best. How shall we understand the monster we are creating, when it is growing so rapidly that it can no longer be tethered by simple means, and it can no longer be outsmarted by any human?

Such a Frankensteinian vision is not as melodramatic as it sounds. The cracks in our invulnerability are already showing as tragedies emerge out of insufficiently understood systems. We have thrown ourselves into deep water with only a crude understanding of the survival equipment. Clearly, the adventure *could* go badly, in a worst case scenario. Luckily, this need not happen if we build using the best principles of science, and base technology on proper knowledge about how the world really works, adapting to its limitations.

This book is about that science, and how we may use it to build reliable infrastructure. It is about the tension between stability, witting and unwitting, and pride in a sense of control and certainty. In a sense, it is the story of how I conceived CFEngine, or if you prefer, of how to implement Computer Immunology.

The book is in three parts:

- Part I Stability: describes the fundamentals of predictability, and why we have to give up the idea of control in its classical meaning.
- Part II Certainty: describes the science of what we can know, when we don't control everything, and how we make the best of life with only imperfect information.
- Part III Promises: explains how the concepts of stability and certainty may be combined to approach information infrastructure as a new kind of virtual material, restoring a continuity to human-computer systems so that society can rely on them.

I have chosen to focus especially on the impact of computers and information on our modern infrastructure, yet the principles we need for managing technology did not emerge from computer science alone. They derive from an understanding of what is all around us, in the natural world. The answers have emerged from all those sciences that learned to deal with fundamental uncertainty about the world: physics, chemistry, economics and biology. Emergent effects, so often mysticized, are really not so mysterious once one takes the time to understand. They are inevitable consequences of information-rich systems. We must understand how to harness them for safe and creative purpose.

When civil infrastructure meant gas lamps and steam locomotives, and computers were still a ghost in the punch-card looms of the industrial revolution, the denizens of history wrestled with fundamental questions about the nature of the physical world that we still find hard to comprehend today. Without those questions and the discoveries they led to, none of what we take for granted today would have been possible.

So, to unearth the roots of this story about technological infrastructure, I want to delve into the roots of science itself, into the principles that allow us to understand system operation and design, to reveal the advances in thinking that led to modern, information rich methods of fending off *uncertainty*. Above all, this is a fascinating story of human endeavour, from a personal perspective. Just beyond reach of most of us, there is a treasure trove of understanding that has propelled humanity to the very limits of imagination. That is surely a story worth telling.

How to read this book

This book introduces concepts that will be new to a majority of readers. To do so, it builds from fundamental ideas that might initially seem to stray from the topic of information, and builds towards the more contemporary aspects of information infrastructure today. Some readers might be impatient to get to the final answers without all the basics, but science does not work that way. I have committed myself to explaining, as plausibly as I can, how this edifice of thought emerged, with cultural and historical context, in the hope that it will make sense to the committed reader. The panorama and intricacies of scientific thought are truly rewarding for those who are willing to climb the mountain.

I have provided a level of technical depth for readers who are conversant with science and technology. However, no one should feel defeated by these details. In all cases, it should be straightforward to skip across details that seem too difficult, and rejoin the train of thought later. I encourage readers to exercise personal judgement and skip over sections that seem difficult, and I cross my fingers that the book might still be enjoyed without every nuance rendered with complete fidelity.

– Mark Burgess, Oslo, 2013

Part I
Stability

Or how we base technology on science

1

King Canute and the Butterfly

How we create the illusion of being in control.

> "If the doors of perception were cleansed every thing would appear to man as it is, infinite. For man has closed himself up, till he sees all things thro' narrow chinks of his cavern."
> - William Blake, *The Marriage of Heaven and Hell*

Imagine a scene in a movie. A quiet night on the ocean, on the deck of a magnificent ship, sailing dreamily into destiny. Moonlight reflects in a calm pond that stretches off into the distance, waves lap serenely against the bow of the ship, and – had there been crickets on the ocean, we would have heard that reassuring purring of the night to calm the senses.

The captain, replete with perfectly adjusted uniform, comes up to the night helmsman and asks: "How goes it, sailor?" To which the sailor replies: "No problem. All's quiet, sir. Making a small course correction. Everything's ship-shape and under control."

At this moment, the soundtrack stirs, swelling into darker tones, because we know that those famous last words are surely a sign of trouble in any Hollywood script.

At that very moment, the camera seems to dive into the helmsman's body, swimming frantically along his arteries with his bloodstream to a cavernous opening, where we view a deadly parasite within him that will kill him within the hour. Then the camera pulls back of him and pans out, rising above the ship, up into the air to an altitude at which the clear, still pond of the ocean seems to freckle and soon becomes obscured by clouds. The calm sea, it turns out, is just the trough of a massive wave that, miles away, reaches up to ten times the height of the ship, and is racing across the planet with imminent destruction written all over it. As we rise up, and zoom out of the detail, we see the edge of a massive storm, swirling with fierce intensity and wreaking havoc on what is now a hair's breadth away on the screen. And then, pulling even farther out, just beyond the edge of the planet, is a swarm of meteorites, firing down onto the human realm, one of which is the size of Long Island (it's always Long Island), and will soon wipe out all life on Earth. Picking up speed now, the camera zooms back and we see the solar system spinning around a fiery sun, and see stars and galaxies and we return to a calm serenity, where detail is mere shades of colour on a simple black canvas. Then the entire universe is swallowed into a black hole.

Shipshape and under control?

Well, don't worry. This scene is not a prediction of anyone's future, and hopefully you recognized the smirk of irony injected for humorous effect. What the imaginary scene is supposed to convey is that our perceptions of being 'in control' always have a lot to do with the scale at which we focus our attention—and, by implication, the information that is omitted. We sometimes think we are in control because we either don't have or choose not to see the full picture.

Is this right or wrong?

That is one of the questions that I want to offer some perspectives on in this book. How much of the world can we really control and harness for our purpose? To make infrastructure, we need to make certain assurances. From the parody presented above, one has the sense that we can control some parts of the world, but that there are also forces beyond our control, 'above' and 'below' on the scale of things.

This makes eminent sense. The world, after all, is not a continuous and uniform space, it is made of bits and pieces, of enclosed regions and open expanses – and within this tangle of environments, there are many things going on about which we often have very limited knowledge. In technical language, we say that we have *incomplete information* about the world. This theme of missing information will be one of the central ideas in this book.

Some of the missing information is concealed by distance, some by obstacles standing in our way. Some is not available because it has not arrived yet, or it has since passed, and some of the information is just occurring on such a different scale that we are unable to comprehend it using the sensory apparatus that we are equipped with. Using a microscope we can see a little of what happens on a small scale, and using satellites and other remote tools we can capture imagery on a scale larger than ourselves, but when we are looking down the microscope we cannot see the clouds, and when we are looking through the satellite, we cannot perceive bacteria.

Truly, our capacity for taking in information is limited by scale in a number of ways, but we should not think for a moment that this is merely a human failing. There is more going on here.

Sometimes a lack of information doesn't matter to predicting what comes next, but sometimes it does. Phenomena have differing sensitivities to information, and that information is passed on as an *influence* between things. So this sensitivity to information is not merely a weakness that applies to humans, like saying that we need better glasses: the ability to interact with the world varies hugely with different phenomena – and those interactions themselves are often sensitive to a preferred scale. We know this from physics itself, and it is fundamental, as we shall see in the chapters ahead.

Let's consider for a moment why scale should play a role in how things interact. This will be important when we decide how to successfully build tools and systems for the world. Returning for a moment to our ship on the sea, we observe that the hull has a certain length and a certain width and a certain height, and the waves that strike it have a certain size. This is characterized by a wavelength (the distance between one wave and the next), and an amplitude (the mean height of the wave). Intuition alone should now tell us that waves that are high compared to height of the ship will have a bigger impact on it when they hit the ship than waves that are only small compared to its size. Moreover, waves that are much shorter in length compared to the length and width of the ship will strike the ship more often than a long wave: bang bang bang instead of gradually rising and falling. Indeed, if a single wave were many kilometres in wavelength, the ship would scarcely notice that anything was happening as the wave passed it, even if the wave gradually lifted the ship a distance of three times its height. Relative scales in time and in space matter.

The ship is a moving, physical object, but could we not avoid the effects of scale by just watching a scene without touching anything? In fact, we can't because we cannot observe any without interacting with it somehow. This was one of the key insights made by quantum theory founder Werner Heisenberg, at the beginning of the 20th century. He pointed out that, in order to transfer information from one object to another, we have to interact with it, and that we change it in the process. For instance, to measure the pressure of a tyre, you have to let some of the air out of the tyre into a pressure gauge, reducing the pressure slightly. There are no one-way interactions, only mutual interactions. In fact, the situation is quite similar to that of the waves on the ship, because all matter and energy in the universe has a wavelike nature.

Knowing your scales is a very practical problem for information infrastructure. To first create, and then maintain information rich services, we have to know how to extract accurate information and act on it cheaply, without making basic science errors. For example, to respond quickly to phone coverage demand in a flash crowd, you cannot merely measure the demand just twice daily and expect to capture what is going on. On the other hand, rechecking the catalogue of movies on an entertainment system every second would be a pointless excess.

One has to use the right kind of probe to see the right level of detail. When scientists take picture of atoms and microstructures, they use small wavelength waves like X-rays and electron waves, where the wavelength is comparable to the size of the atoms themselves. The situation is much like trying to feel the shape of tiny detail with your fingers. Think of a woollen sweater. Sweaters have many different patterns of stitching, but if you close your eyes and try to describe the shape of the stitching just by feeling it with your fat fingers, you would not be able to because each finger is wider than several threads. You can see the patterns with your eyes because light has a wavelength much smaller than the width of either your fingers or the wool.

Now suppose you could blow up the sweater to a much greater size (or equivalently shrink your fingers), so that the threads of wool were like ropes; then you would be able to feel the edges of each thread and sense when one thread goes under or over another. You would be able to describe a lot more information about it. Your interaction with the system of the sweater would be able to *resolve* detail and provide *sufficient information*[4] .

To develop a deep understanding of systems, we shall need to understand information in some depth, especially how it works at different scales. We'll need to discuss control, expectation and stability, and how these things are affected by the incompleteness of the information we have about the world. We need to think bigger than computers, to the world itself as a computer, and we need to ask how complete and how certain is the information we rely on.

Information we perceive is limited by our ability to probe things as we interact with them—we are trapped between the characteristic scales of the observer and the observed. Infrastructure is limited in the same way. This is more than merely a biological limitation to be overcome, or a problem of a poorly designed instrument. In truth, we work around such limitations in ingenious ways all the time, but it is not just that. This limitation in our ability to perceive is also a benefit. We also use that limitation purposely as a tool to understand things, to form the illusion of mastery and control over a limited scale of things, because by being able to isolate only a part of the world, we reduce a hopeless problem to a manageable one.

We tune into a single frequency on the radio, we focus on a particular distance, zoom in or zoom out, looking at one scale at a time. We place things in categories, or boxes, files into folders, we divide up markets into stalls, and malls into shops, cities into buildings, and countries into counties, all to make our comprehension of the world more manageable by limiting the amount of information we have to interact with at any time. Our experience of the world can be made comprehensible or incomprehensible, by design.

Analogies will be helpful to us in understanding the many technical issues about information. Thinking back to the ship, our fateful helmsman, who reported that everything was under control, was sitting inside a ship, within a calm region of ocean with sensory devices that were unable to see the bigger picture. Without being distracted by the inner workings of his body, he was able to observe the ship and the ocean around him and steer the ship appropriately within this region. By being the size that he was, he could fit inside the safety of the ship, avoiding the cold of the night, and by fitting into the calm trough of the ocean's mega-wave, the ship was able to be safely isolated from damage caused by the massive energies involved in lifting such an amount of water.

The mental model of the ship on the ocean seemed pretty stable to observers at that scale of things, with no great surprises. This allowed the ship to function in a controlled manner and for humans to perceive a sense of control in the *local* region around the ship, without being incapacitated by *global* knowledge of large scale reality. How would this experience translate into what we might expect for society's information infrastructure?

Scale is thus both a limiter and a tool. By shutting out the bigger (or smaller) picture, we create a limited arena in which our actions seem to make a difference[5]. We say that our actions *determine* the outcome, or that the outcomes are *deterministic*. On a cosmic scale, this is pure hubris—matters might be wildly out of control in the grand scheme of things, indeed we have no way of even knowing what we don't know; but that illusion of local order, free of significant threat, has a powerful effect on us. If fortune is the arrival of no unexpected surprises, then fortune is very much our ally, as humans, in surviving and manipulating the world around us.

The effect of limited information is that we perceive and build the world as a collection of containers, patches or environments, separated from one another by limited information flow. These structures define characteristic *scales*. In human society, we make countries, territories, shops, houses, families, tribes, towns, workplaces, parks, and recreation centres. They behave like *units* of organization, if not physically separated then at least de-marked from one another; and, within each, there are clusters of interaction, the molecules of human chemistry. Going beyond what humans have built, we have environments such as micro-climates, ponds, the atmosphere, the lithosphere, the magnetosphere, the atomic scale, the nuclear scale, the collision scale or mean free path, and so on. All of these features of the world that we identify can be seen as emerging from a simple principle: a finite range of influence, or limited transmission of information, relative to a certain phenomenon.

There are really two complementary issues at play here: perception and influence. We need to understand the effect that scale has upon these. The more details we can see, the less we have a sense of control. This is why layers of management in an organization tend to separate from the hands-on workers. It is not a class distinction, but a skill separation. In the semantic realm, this is called the *separation of concerns*, and it is not only a necessary consequence of loss of resolution due to scale, but also a strategy for staying sane[6]. Control seems then to be a combination of two things:

$$\text{Control} \rightarrow \text{Predictability} + \text{Interaction}$$

To profit from interactions with the world, in particular the infrastructure we build, it has to be predictable enough to use to our advantage. If the keyboard I am typing on were continuously changing or falling apart, it would not be usable to me. I have to actually be able to interact with it—to touch it.

Rather than control, we may talk about certainty. How sure we can we be of the outcomes of our actions? Later in the book, I will argue that we can say something like this:

$$\text{Certainty} \rightarrow \text{Knowledge} + \text{Information}$$

where knowledge is a relationship to the history of what we've already observed in the past (i.e. an expectation of behaviour), and information is evidence of the present: that things are proceeding as expected.

Predictability and interaction: these foundations of control lie at the very heart of physics, and are the essence of information, but can we guarantee them in sufficient measure to build a world on top? Even supposing that one were able to arrange an island of calm, in which to assert sufficient force to change and manipulate the world, are we still guaranteed absolute control? Will infrastructure succeed in its purpose?

The age of the Enlightenment was the time of figures like Galileo Galilei (1564-1642) and Isaac Newton (1642-1727), philosophers who believed strongly in the idea of reason. During these times, there emerged a predominantly machine-like view of the world. This was in contrast with the views of Eastern philosophers. The world itself, Newton believed, existed essentially as a deterministic infrastructure for God's own will. Man could merely aspire to understand and control using these perfect laws.

The concept of determinism captures the idea that cause and effect are tightly linked to bring certainty, i.e. that one action inevitably causes another action in a predictable fashion, to assure an outcome as if intended[7]. Before the upheavals of the 19th and 20th century discoveries, this seemed a reasonable enough view. After all, if you push something, it moves. If you hold it, it stops. Mechanical interaction itself was equated with control and perfectly deterministic outcome. Newton used his enormous skill and intellect to formalise these ideas.

The laws of geometry were amongst the major turning points of modern thinking that cheered on this belief in determinism. Seeing how simple geometric principles could be used to explain broad swathes of phenomena, many philosophers, including Newton, were inspired to mimic these properties to more general use. Thomas Hobbes (1588-1679) was one such man, and a figure whom we shall stumble across throughout this story of infrastructure. A secretary to Francis Bacon (1561-1626), one of the founders of scientific thinking, he attempted to codify principles to understand human cooperation, inspired by the power of such statements of truth as ‘two straight lines cannot enclose a space’. He dreamt not just of shaping technology, but society itself, by controlling it with law and reason.

Information, on the other hand, as an idea and as a commodity, played an inconspicuous role during this time of Enlightenment. It crept into science more slowly and circuitously than determinism, through bookkeeping ledgers of experimental observation, but also implicitly through the new theory for moving objects. Its presence, although incognito, was significant nonetheless in linking descriptive state with the behaviour of things. Information was the key to control, if only it could be mastered.

Laws, inspired by geometry then, began to enter science. Galileo’s law of inertia, which later became co-opted as Newton’s first law of motion, implicitly linked information and certainty with the physics of control, in a surprising way. It states that, unless acted upon by external forces, bodies continue in a state of rest or uniform motion in a straight line. This is basically a statement that bodies possess motional *stability* unless perturbed by external influences. Prior to this, it had seemed from everyday experience that one had to continually push something to make it move. The concepts of friction and dissipation were still unappreciated. Thus Newton’s insights took science from a general sense of motion being used up, like burned wood, to the idea that motion was a property that might be conserved and moved around, like money.

Newton’s first law claimed that, as long as no interactions were made with other bodies, there could be no payment, and thus motion would remain constant. Thus emerged a simple accounting principle for motion, which is what we now call *energy*. The concept of energy, as book-keeping information, was first used formally by German philosopher Gottfried Wilhelm Leibniz (1646-1716), Newton's contemporary and rival in scientific thought, though Thomas Young is recorded as the first person to use the modern terminology in lectures to the Royal Society in 1802.

Newton's second law of motion was a quantification of how these energy payments could be made between moving bodies. It provided the formula that described how transmission of a form of information altered a stable state of motion. The third law said that the transmission of influence between bodies had to result in a mutual change in both parties: what was given to one must be lost by the other. In other words, for every force given there is an equal and opposite back-reaction on the giver.

The notion of physical law emerging was that the world worked basically like a clockwork machine, with regular lawful and predictable behaviour, book-keeping its transactions as if audited by God himself. It all worked through the constant and fixed infrastructure of a *physical law*. If one could only uncover all the details of that law, Man would have total knowledge of the future. Information could become a tool, a technology. This was Newton's vision of God, and there is surely a clue in here to the search for certainty.

These thoughts were essential to our modern view of science. They painted a picture of a world happening, like a play, against a cosmic backdrop, with machine-like predictability. They still affect the way we view the separation of 'system' and 'environment', such as the activities of consumers or users from society's background infrastructure. Universality of phenomena independent of environments allowed one to reason, infer and plan ahead.

When Galileo dropped balls of different mass from the Tower of Pisa in the 16th century, predicting that they would hit the ground at the same time, he codified this as a law of nature that all masses fall at the same rate[8]. The original experiment was said to have been a hammer and a feather[9], which had not worked due to the air resistance of the feather. The experiment was repeatable and led to the promise that the same thing would happen again if the experiment were repeated. His experiment was repeated on the Moon by Apollo 15 astronaut David Scott, actually using a hammer and a feather. He proved that if one could eliminate the interfering factors, or separate the concerns, the rule remained true. (Note how the concept of promises seems to be relevant in forming expectations. This theme will return in part III of the book.)

We might choose to take this continuity of physical behaviour as an axiom, but the example illustrates something else important: that what seems to be a law often needs to be clarified for the context in which it is made. On Earth, dropping a hammer and a feather would not have the same result because of the air resistance of the feather. In outer space, the lack of uniform gravitation would prevent the objects from falling at all. Had we simply made the promise (or more boldly, the 'law') that says hammers will fall at a constant acceleration, it would have been wrong. The way we isolate effects on a local scale and use them to infer general rules is a testament to the homogeneity of physical behaviour.

But how far does this go? Still there is the issue of scale. Is the world truly clockwork in its machinations, and at all scales, and in all contexts? During Newton's time, many philosophers believed that it was[10].

The desire to control is a compelling one that has seduced many minds over the years. A significant amount of effort is used to try to predict the stock market prices, for instance, where millions of dollars in advantage can result from being ahead of competitors in the buying and selling of stocks. We desire an outcome, and we desire the ability to determine the outcome, so that we can control things. Yet this presupposes that the world is regular and predictable with the continuity of cause and effect. It presupposes that we can isolate multiple attempts to influence the behaviour of something so that the actual thing *we* do is the deciding factor for outcome —with no interference from outside. Noise and interference (radio noise, the weather, flocks of birds at airports, etc.), are constant forces to be reckoned with in building technologies for our use.

Shutting things into isolation, by separating concerns is the classic strategy to win apparent control from nature. This is reflected in the way laws of physics are formulated. For example: "A body continues in a state of rest or uniform motion in the absence of net external forces". This is a very convenient idealization. In fact, there is no known place in the universe where a body could experience exactly no external force. The law is a fiction, or – as we prefer to say – a *suitably idealized approximation* to reality. Similarly, examples of Newtonian mechanics talk about perfectly smooth spheres, with massless wires and frictionless planes. Such things do not exist, but, had they existed, they would have made the laws so much simpler, separating out only the relevant or dominant concerns for a particular scale. Physics is thus a creative work of abstraction, far from the theory of everything that is sometimes claimed.

For about a century, then, determinism was assumed to exist and to be the first requirement to be able to exert precise control over the world. This has come to dominate our cultural attitudes towards control. Today, determinism is known to be fundamentally false, and yet the illusion of determinism is still clung onto with fervour in our human world of bulk materials, artificial environments, computers and information systems.

The impact of Newton and Leibniz on our modern day science, and everyday cultural understanding of the world, can hardly be overestimated. The clockwork universe they built is surely one of the most impressive displays of theoretical analysis in history. If we need further proof that science is culture, we only have to look at the way their ideas pervade every aspect of the way that we think about the world. Words like energy, momentum, force and reaction creep into everyday speech. But they are approximations.

The problem was this: the difficult problems in physics from Newton's era were to do with the movements of celestial bodies, i.e. the physics of the very large. Seemingly deterministic methods could be developed to approximate these large scale questions, and today's modern computers make possible astounding feats of prediction, like the Martian landings and the Voyager space probe's journey through the solar system, with a very high degree of accuracy. What could be wrong with that?

The answer lay in the microscopic realm, and it would not be many years following Newton, before determinism began to show its cracks, and new methods of statistics had to be developed to deal with the uncertainties of observing even this large scale celestial world of tides and orbits. For a time science could ignore indeterminism. Two areas of physics defied this dominion of approximation, however, and shattered physicists' views about the predictability of the natural world. The first was quantum theory, and the second was the weather.

The 20th century saw the greatest shift away from determinism and its clockwork view of the universe, towards explicitly non-deterministic ideas. It began with the development of statistical mechanics to understand the errors in celestial orbits, and the thermodynamics of gases and liquids that lay behind the ingenious innovations of the industrial revolution[11]. Following that, came the discovery that the physics of the very small could only be derived from a totally new kind of physical theory, one that did not refer to actual moving objects at all, but was instead about different states of information.

That matter is made up of atoms is probably known by everyone with even the most rudimentary education today. Many also know that the different atomic elements are composed of varying numbers of protons, neutrons and electrons, and that even more particles were discovered after these, to form part of a sizable menagerie of things in the microscopic world of matter. Although commonplace, this knowledge only emerged relatively recently in the 20th century, a mere hundred years ago, and yet its impact on our culture has been immense[12]. What is less well known is that these microscopic parts of nature are goverened by the dynamics of states rather than smooth motion.

A *state* is a very peculiar idea, if you are used to thinking about particles and bodies in motion, but it is one that has, more recently, become central in information technology. A state describes something like a signal on a traffic light, i.e. its different colour combinations. We say that a traffic light can exist in a number of allowed states.

For example, in many countries, there are three allowed states: red, amber and green. In the UK and some European and Commonwealth countries, however, it can exist in one of four states: red, amber, red-amber, and green. The state of the system undergoes transitions: from green to amber to red, for instance.

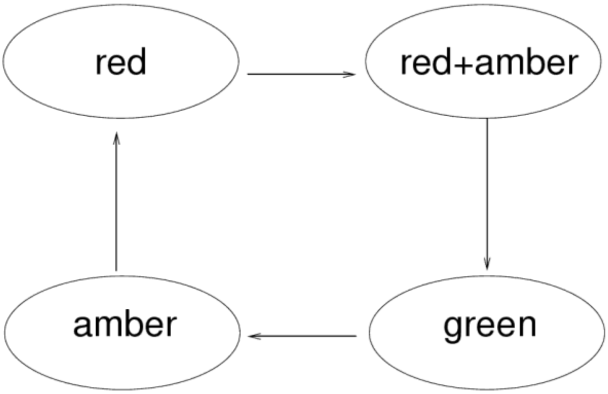

Fig. 1.1. The traffic light cycle is made up of four distinct states. It gives us a notion of orbital stability in a discrete system of states, in the next chapter.

What quantum science tells us is that subatomic particles like electrons are things that exist in certain states. Their observable attributes are influenced (but not determined) by states that are hidden somewhere from view. Even attributes like the positions of the particles are unknown *a priori*, and seem to be influenced by these states. At the quantum level, particles seem to live *everywhere* and nowhere. Only when we observe them, do we see a snapshot outcome. Moreover, instead of having just three or four states, particles might have a very large number, and the changes in them only determine the *probability* that we might be able to observe the state in the particle.

It is as if we cannot see the colour of nature's states directly, we can only infer them indirectly by watching when the traffic starts and stops. These are very strange ideas, nothing like the clockwork universe of Newton.

The history of these discoveries was of the greatest significance, gnawing at the very roots of our world view. It marked a shift from describing outcomes in terms of forces and inevitable laws, to using sampled information and chance. The classical notion of a localized point-like particle was obliterated leaving a view based just on information about observable states. The concept of a particle was still 'something observable', but we know little more than that. Amazingly, the genius of quantum mechanics has been that we also don't need to understand the details to make predictions[13]. We can get along without knowing what might be happening 'deep down'. Although a bit unnerving, knowing that we have to let go of determinism has not brought the world to a standstill.

There are lessons to be learnt from this, about how to handle the kind of uncertainties we find today in modern information infrastructure, as its complexity pushes it beyond the reach of simple methods of control.

The way 'control' emerges in a quantum mechanical sense is in the manipulation of guard-rails or constraining walls, forces called *potentials*: containers that limit the probable range of electrons to an approximately predictable region. This is not control, but loading the dice by throwing other dice at them. Similarly, when building technologies to deal with uncertainty, we must use similar ideas of constraint.

We don't experience the strangeness of the quantum world in our lives because, at the coarse scale of the human eye, of many millions of atoms, the bizarre quantum behaviour evens out—just as the waves lapping onto our ship were dwarfed by the larger features of the ocean when zooming out. We can expect the same thing to happen in information systems too. As our technology has become miniaturized to increasingly atomic dimensions, the same scaling effect is happening effectively in our information infrastructure today, thanks to increased speed and traffic density.

Determinism is thus displaced, in the modern world, from being imagined as the clockwork mechanism of Newtonian 'physical law', to merely the representing the intent to measure: the interaction with the observer[14]. This is a symptom of a much deeper shift that has shaken science from a description in terms of ballistic forces to one of message passing, i.e. an information based perspective. We'll return further to this issue and describe what this for infrastructure means in chapter 3.

Our modern theory of matter in the universe is a model that expressly disallows precise knowledge of its internal machinery. Quantum Mechanics, and its many founders[15] throughout the first half of the 20th century, showed that our earlier understanding of the smallest pieces of the world was deeply flawed, and only *seemed* to be true at the scale of the macroscopic world. Once we zoom in to look more closely, things look very different. Nevertheless, we can live with this lack of knowledge and make progress. The same is true modelling information infrastructure, if we only understand the scales in the right way. This presents a very different challenge to building technology: instead of dealing with absolute certainty, we are forced to make the best of unavoidable *uncertainty*[16].

But there is more. As if the quantum theory of the very small weren't enough, there was a further blow to determinism in the 1960s, not from quantum theory this time, but rather from the limits of information on a macroscopic scale. These limits revealed themselves during the solving of complicated hydrodynamic and fluid mechanical models, by the new generation of digital computers, for the more mundane purpose of the weather forecast.

What makes weather forecasting difficult is physics itself. The physics of the weather is considerably harder than the physics of planetary motion. Why? The simple answer is scale. In planetary motion, planets and space probes interact weakly by gravitational attraction. The force of attraction is very weak, which has the effect that planets experience a kind of calm ocean effect, like our imaginary ship. What each celestial body experiences is very predictable, and the time-scales over which major changes occur are long in most cases of interest.

The atmosphere of a planet is a thin layer of gases and water, frothing and swirling in constant motion. Gas and liquid are called fluids, because they flow (they are not solids). They experience collisions (pressure), random motion (temperature) and even internal friction and stickiness (viscosity), fluids can be like water or like treacle, and unlike planets their properties change depending on how warm they are. The whole of thermodynamics and the industrial revolution depended on these qualities. The timescales of warming and cooling and movement of the atmosphere are similar to the times over which large bodies of fluid move. All of this makes the weather a very complex fluid dynamics problem.

Now, if we zoom in with a microscope, down to the level of molecules, the world looks a little bit like a bunch of planets in motion, but only a superficially. A gas consists of point-like bodies flying around. So why is a fluid so much harder to understand? The answer is, again, the scales at work. In a gas, there are billions of times more bodies flying around than in a solar system, and they weigh basically nothing compared to a planet. Molecules in a gas are strongly impacted when sunlight shines onto them – because they are so light, they get a significant kick from a single ray of light. A planet, on the other hand, is affected only infinitesimally by the impact of light from the sun.

So when we model fluids, we have to use methods that can handle all of these complex, strong interactions. When we model planets, we can ignore all of those effects as if they were small waves lapping against a huge ship. The asteroid belt is the one feature of our solar system that one might consider modelling as a kind of fluid. It contains many bodies in constant, random motion, within a constrained gravitational field. However, even the asteroid belt is not affected by the kind of sudden changes of external force that our atmosphere experiences. All of these things conspire to make the weather an extremely difficult phenomenon to understand. We sometimes call this chaos.

The Navier-Stokes equations for fluid motion were developed in the 1840s by scientists Claude-Louis Navier (1785-1836) and George Gabriel Stokes (1819-1903), on the basis of a mechanical model for idealized incompressible fluids. They are used extensively in industry as well as in video games to simulate the flowing of simple and complex mixtures of fluids through pipes, as well as convection cells in houses and in the atmosphere. As you might imagine, the equations describing the dynamics of fluids and all of their strongly interacting parts are unlike those for planetary motion, and contain so-called non-linear parts. This makes solving problems in fluid mechanics orders of magnitude more difficult than for planetary motion, even with the help of computers. The reason for this is rather interesting.

The problem with understanding it was that its tightly coupled parts made computational accuracy extremely important. The mixing of scales in the Navier-Stokes equations was the issue. When phenomena have strongly interacting parts, understanding their behaviour is hard, because the details at one scale can directly affect the details at another scale. The result is that answers become very sensitive to the accuracy of the calculations. If couplings in systems are only weak, then a small error in one part will not affect another part very strongly and approximations work well. But in a strongly connected system, small errors get amplified. For example, planets are only weakly coupled to one another, so an earthquake in California (which is a small scale phenomenon) does not affect the orbit of the Moon (a large scale phenomenon).

Suppose our ship was not floating on the water with a propeller, but being pushed across the water by a metal shaft from the land, like a piston engine. Then the ship and the land would have been strongly coupled, and what happened to one would be immediately transmitted back to the other. Another way of putting it is that what would have remained short-range and local to the ship, suddenly attained a long-range effect. So long- and short-scale behaviour would not separate.

The coupling in a fluid is not quite as strong as a metal shaft, between every atom, but it is somewhere in between that and a collection of weakly interacting planets. This means that the separation of scales does not happen in a neat, convenient way.

What does all this have to do with the weather forecast? Well, the simple answer is that it made calculations and modelling of the weather unreliable. In fact, the effects of all of this strong coupling came very graphically into the public consciousness when another scientist in the 1960s tried to use the Navier-Stokes equations to study convection of the atmosphere. American mathematician and meteorologist Edward Lorenz (1917-2008) developed a set of equations, called naturally the Lorenz equations to model the matter of two dimensional convection, derived with the so-called Boussinesq approximation to Navier-Stokes. On the surface of the Earth, which is two dimensional, convection is responsible for driving many of the processes that lead to the weather.

The Lorenz equations, which look deceptively simple compared to the Navier-Stokes equation, are a set of mutually modulating differential equations which, through their coupling, become *non-linear*. Certain physical processes are referred to as having non-linear behaviour, which means that the graph of their response is not a straight line, giving them an amplifying effect. The response of such a process to a small disturbance is disproportionately larger than the disturbance itself, like when you whisper into a microphone that is connected to a huge public address system and blow back the hairstyles of your audience. Disproportionately means, again, not a proportional straight-line relationship, but a superlinear curve which is therefore ‘non-linear’.

Non-linearity makes amplification of effect much harder to keep track of because it amplifies different quantities by different amounts, and makes computations inaccurate in inconsistent ways. When you combine those results later, the resulting error is even harder to gauge and so the problems pile up. Everything is deterministic all the way, but neither humans nor computers can work to unlimited accuracy, and the methods of computation involve approximation that gets steadily worse. The problem is not determinism but instability.

When not much is happening in the weather, the calculational approximation is good and we can be more certain of an accurate answer, but when there is a lot of change, amplification of error gets worse and worse. That is why predicting the weather is hard. If it looks like there will be no change, that is a reliable prediction. But if it looks like change is afoot (which is what we are really interested in knowing), then our chances of getting the calculations right are much smaller.

Lorenz made popular the notion of the *butterfly effect*, which is well known these days as an illustration of chaos theory. It has appeared in popular culture in films and literature for many years[17]. The butterfly effect suggests somewhat whimsically that a delicate and tiny movement like the flapping of a butterfly wing in the Amazon rain forest could, through the non-linear amplification of the weather processes, be imagined to lead to a devastating storm or cataclysm[18] on the other side of the planet. Such is the strength of the coupling – as if an earthquake could shake the Moon. Although his point was meant as an amusing parody of non-linearity, it made a compelling image that popularized the notion of the amplification of small effects into large ones.

Strong coupling turns out to be a particular problem in computer-based infrastructure, though this point needs further explanation in the chapters to come. Chaos is easily contained, given the nature of information systems, yet systems are often pushed beyond the brink of instability. We do not escape from uncertainty so easily.

Prediction is an important aspect of control, but what about our ability to influence the world? You can't adjust the weather with a screwdriver, or tighten a screw with a fan. That tells us there are scale limitations to influence too. I would like to close this section with a story about the effect of scale on control.

The presumably apocryphal tale of King Canute, or Knut the great, who was a king of Denmark, England, Norway, and parts of Sweden at the close of the first millennium, is well known to many as a fable of futility in the face of a desire to control. Henry of Huntingdon, the 12th century chronicler, told how Canute set his throne on the beach and commanded the waves to stay back. Then, as the waves rolled relentlessly ashore, the tide unaffected by his thought processes, or the stamping of his feet, Canute declared: "Let all men know how empty and worthless is the power of kings!"

Written in this way, with a human pitting himself against a force of nature that we are all familiar with, has a certain power to capture our imaginations. It seems like an obvious lesson, and we laugh at the idea of King Canute and his audacious stunt, but we should hold our breaths. The same mistake is being made every day, somewhere in the world. Our need to control is visceral, and we go to any lengths to try it by brute force, before reason wins the day.

Many of us truly believe that brute force is the answer to such problems: that if we can just apply sufficient force, we will conquer any problem. The lessons of non-predictability count for little if you don't know them. Force can be an effective tool in certain situations, but scale plays a more important role.

Canute could not hold back the ocean because he had no influential tools on the right scale. Mere thought does not interact with anything outside of our brains, naturally, so it had little hope of holding back the water. As a man, he could certainly have held back water. He had sufficient force—after all, we are all mostly made up of water, so holding back water requires about the same amount of force as holding back other people. However, unlike other people, water is composed of much smaller pieces that can easily flow around a man. Moreover, the sheer size of the body of water would have simply run over him as a finger slides over woollen threads – the water would barely have noticed his presence. A futile mismatch of scales – he could never even have used the force he was able to exert against an army because the interaction between a solid man and liquid water is rendered so weak as to be utterly ineffective.

As humans, we forget quickly the natural disasters after they have happened and go back to believing we are in control[19]. Control is important to us. We talk about it a lot. It is ultimately connected with our notions of free will, of intent and of purpose. We associate it with our personal identity, and our very survival as a species. We have come to define ourselves as a creative species, and one that has learnt to master its environments. Much of that ingenuity gives us a sense of control.

A central theme of this chapter is that control is about information, and the scale or resolution at which we can perceive it. Surely, using today's information systems, we could build a machine to do what Canute could not? If not actually hold back the tide, then at least adapt to it so quickly as to keep dry in some smart manner? Couldn't windows clean themselves? Couldn't smart streets in developing countries clear stagnant water and deploy agents to fight off malaria and other diseases? Couldn't smart environments ward off forces on a scale that we as individuals cannot to address? These things might be very possible indeed, but would they be stable? The bounds of human ingenuity are truly great, but if such things are possible, it still begs an important question: would such smart technologies be stable to such a degree that we could rely on them to be safe?[20].

Now think one step further than this, not to a modified natural world, but to the artificial worlds of our own creation: our software and information systems. These systems are coupled to almost every aspect of human survival today. They have woven themselves into the fabric of society in the so-called developed world and we now depend on them for our survival, just like our physical environment. Are these systems stable and reliable?

In a sense, they are precisely what King Canute could not accomplish: systems that hold back a deluge of communication and data records, on which the world floats, from overwhelming us, neutralizing information processing tasks that we might have tried to accomplish by brute force in earlier generations, or could not have done at all without their help. Even if it is a problem of our own making, without computers we would truly drown in the information that our modern way of living needs to function.

What can we say about the stability of our modern IT infrastructure? Software systems do not obey the laws of physics, even though they run on machines that do. The laws of behaviour in the realm of computers are not directly related to the physical laws just discussed. Nevertheless, the principles we've been describing here, of scale and determinism, must still apply, because those principles are about information. Only the realization of the principles will be different.

Scale matters.

Will our future, information-infused world be safe and reliable? Will it be a calm moonlit ocean or a chaotic mushroom cloud? The answer to these questions is: it will be what we make it, but subject to the limitations of scale and complexity. It will depend on how stable and reliable the environment around it is. So we must ask: are the circumstances around our IT systems (i.e. the dominant forces and flows of information at play) sufficiently predictable that we can maneuver reliably within their boundaries? If so, we will feel in control, but there are always surprises in store—the freak waves, the unexpected flapping of a butterfly's wings in an unwittingly laid trap of our own making. How can we know?

Well, we go in search of certainty, and the quest begins with the key idea of the first part of this book: stability.

2

Feedback Patterns and Thresholds

How the relative sizes of things govern their behaviours

Astronomy is 1derful,
And interesting 2,
The Ear3volves around the sun,
That makes a year 4 you.
The Moon affects the sur5 heard,
By law of phy6 great,
It7 when the stars, so bright,
Do nightly scintill8.
If watchful providence be9,
With good intentions fraught,
Would not been up her watch divine,
We soon should come to 0 (nought).
– Unknown[21]

Before we can argue how stable systems ought to be, we have to be able to measure how stable they actually are[22]. That is a surprisingly difficult thing to do. To understand better, we need to cover some more of the basic science of scales, and explore how these affect stability.

There are two essential reasons for wanting to measure stability: the first is to be able to put a practical figure on the result for the sake of comparison. The other is to know how it arises so that one might diagnose failures of stability. By understanding how to measure a thing, we also hope to learn something about it in the process. From the previous chapter, it should come as no surprise that stability is closely related to scales in the system[23].

Consider some examples. A computer program crashes, or undergoes a failure, when the amount of memory it needs approaches the scale of total amount of memory available in the computer. It reaches a limit for its container. The size of the container influences the functioning of the software. Even though the container is designed by humans, and is immunized against the laws of physics by existing only in software, there is no escaping the limitations of scale.

Instability can also be benign. The division of a cell during reproduction (a process called mitosis) happens by partitioning an existing single cell, while drawing nutrients from the environment. After some time, the processes within it have duplicated the contents of the original cell and the cell reaches a point of instability at which it splits. Was the split caused by the absorption from the environment, or a result of genetic programming from within? The scales at work here involve both time and space: the rate of the process of creating internal molecular structure, by absorbing material from outside the cells takes time and uses up space. We might measure the stability of a cell by the presence of a critical level of signalling proteins called growth factors, which are believed to initiate the cell division process[24].

A radioactive atom on the other hand, like Uranium 234, decays spontaneously by alpha emission to Thorium 230. No outside influence is needed, but a certain time must elapse for quantum processes to wait for this 'random' decay event to happen. We measure the stability of a so-called 'random' decay process by its *half-life*. That is the length of time we expect to elapse for half of the material to decay[25].

Instability does not have to result in a sudden 'event', like an explosion. Such a change is called a *catastrophe* in mathematics, which is not to place any moral judgement on the change: a catastrophe is a break in the continuity of a system—a change of its structure. Other kinds of change happen more smoothly than this. We might say that the water level in a bucket is stable, if it does not change; but, if we discover a leak, it will start to sink gradually. There is no sudden catastrophe, but rather a gradual process of decay. We could even top up the bucket to keep it full, as we do with the air in our tyres.

To put it into perspective, we find the emptying of the bucket disturbing mainly because it has a finite size and we know that there will be and end to the whole process. The size of the bucket defines a lifetime for the process: a timescale. If the level could simply continue falling slowly forever, we would think this was normal and there would be no problem. We say that nitroglycerine is unstable, for instance, because the smallest nudge can cause it to explode. The revolution counter (RPM meter) on cars, has red band, above a certain level, indicating when engine revolutions per second have become higher than recommended for the engine under normal working conditions. You don't expect the car to explode at this point, but prolonged use at this level might cause wear, which in turn could precipitate a catastrophe. Given the potential for instability, we look for ways to prevent the approach of the relevant threshold by observing.

The great fire of London in 1666 precipitated the invention of the first automated threshold detector—a fire alarm, to prevent the spread of a flammable instability in wooden houses. The alarm comprised a string that 'stretched through each room of a house, and then extended to the basement where it was connected to a weight suspended over a gong'[26]. The idea was that a fire in this threshold monitored home would burn through the string and wake the household to a timely escape.

A slight improvement on this detector is the bi-metallic strip, which is two long pieces of metal with different thermal expansion properties, usually steel and copper, coiled so as to take up less space. When the temperature changes, the metals expand or contract at different rates causing the metal to bend. It is a transducer that converts temperature change into mechanical movement. Bi-metallic strips are still used in basic thermostats to regulate the switching on and off of central heating systems, based on a temperature setting. The arbitrary temperature setting on the thermostat dial defines a user-controlled scale. When the threshold is passed, it triggers an electrical switch to initiate heating, and when it cools, the circuit is broken.

In medical monitors, thresholds are used for a variety of vital signs, like heart rate and breathing rate, but because humans are not such simple machines as to be regulated by a single dial, multivariate thresholds are normal to provide early warning of patient crises.

Arbitrary thresholds are used to signify critical behaviour in many industrial monitoring systems. Such a reliance on what seems to be an arbitrary choice should prompt the obvious question: how can one know the threshold for the onset of instability? One answer would clearly be to measure it by trial and error – but that presupposes that we can reproduce the exact conditions in advance of every scenario which is unlikely. The answer is to look more deeply into how scales emerge in systems of all kinds.

Suppose someone asks how old you are, what do you say? You say, "I'm 25", of course. If you were a diligent scientist, you would immediately say: "25 what? 25 years, 25 seconds?" A smart aleck, trying to be funny, would reply: "Why 25 marriages of course".

Years, metres, seconds (and marriages) are all units of measurement. In America and parts of the Commonwealth, imperial units are used. Everywhere else today we use S.I. units from the French Systeme D'internationale, also known as the metric system (or sometimes the KMS, for kilogrammes, metres, and seconds). In physics, however, we care much less about units as the underlying *kind* of measurement.

One of the most remarkable things about measurements is that they can all be broken down into a small set of basic measures: mass, length, time. Sometimes others are used, like electric charge. These are the basic measures of the physical universe we live in. In the virtual world of software and computers, there are other measures like bits and bytes, CPU cycles, and others. If we know that two things represent the same kind of measurement, then we know that they must have the same kind of measurement scale.

We call such classes of length, time, and mass *dimensions*. This has nothing really to do with the four dimensions of space and time, or the extra dimensions written about in speculative physics books. Think rather about the dimensions of a television screen, i.e. its height and width. But now keep in mind that size is only one kind of extent in space. Oths might be age (which is time), or weight (which has to do with mass).

While all this might sound terribly prosaic, it is in fact one of the most powerful tools any scientist has in his or her bag of tricks for analysing the world, for it allows us to understand the world in scale, predict the behaviour of real objects like aircraft and buildings from scale models, test them in wind tunnels, with scaled forces. With care, the technique offers enormous insight about the world of scales—and we can use this in our quest to understand stability.

Physicist Jean Baptiste Joseph Fourier (1768-1830), interested in the dimensional analysis that Newton had made powerful use of in his work, was quick to point out that it does not make sense to add together numbers of different dimensions:

3 metres + 4 metres = 7 metres,

but

3 metres + 4 seconds = ???

doesn't make anything. Metres and seconds, he might have said, are completely separate ideas. Physicists would say the measurements were 'orthogonal', meaning literally at right angles to one another[27] – you can change length without changing time. The importance of these difference scales to science is that actual phenomena tie different scales together, and we can use them to discover when a system is going to destabilize.

For example, a moving body has a speed, which is a distance covered over a certain time. So speed has dimensions of distance per unit of time, or L/T or length over time. Acceleration is a change in speed over an interval of time, so it has dimensions of speed divided by time, or distance over time squared

$$L/(T \times T) = L/T^2.$$

Newton's second law tells us that external force is proportional to mass multiplied by acceleration, so the dimensions of force are ML/T^2, and since, by Fourier's argument, the dimensions of force must always be the same, we can infer the any force in any situation must have these dimensions.

We can go on relating measurable quantities to other measurable quantities, and always reduce a measurement to a combination of these basic dimensions: mass, length, time. At least, that is true in the physical world. In other scenarios, these measurements might not be helpful—even if they underlie some kind of basic truth about the real world. The example of bits as a unit of information is an important one. The word 'bit' is a short form of *binary digit*, the key term being *digit*. Unlike a length, which can take on a continuum of values, a bit can only have the value 1 or 0, hence it is binary in nature. From bits we get bytes (meaning 'by eight' or $\times 8$ bits), and words, and kilobits and megabytes, and so on. The digit is thus a dimension in its own right[28].

These ideas about units of measurement form the beginning of what is known as *dimensional analysis*[29], a form of analysis that allows us to see what probably ought to be obvious about system behaviour, but which might be obscured by complicated interrelationships or fancy names. As a physics student, I recall thinking that dimensional analysis seemed to be the least glamorous part of the whole undergraduate syllabus—it was certainly presented with a certain lacklustre. It was only when I attended lectures on fluid mechanics, and later quantum field theory that its significance began to emerge. Today, I think it is one of the most remarkable aspects of physics.

Dimensional analysis starts to get interesting when we consider systems with more than one scale of the same dimension, (e.g. if there are two characteristic lengths instead of just one that govern the behaviour of a system). That is the case in the bi-metallic strip, and we use that property creatively. Just as Fourier noted that it doesn't make sense to add numbers with different units, so it also makes no sense to compare them.

Comparing the size of an egg with the time it takes to boil another does not help you to compare two eggs. However, if we divide a number measured in some units by another number, measured in the same units, the result is a pure ratio that really does give us an honest comparison of the two. It is honest because it is a relative measure. Thus, for two identical eggs, the ratio of the height of one to the height of another allows us to say that they have eggsactly (sorry) the same height, similarly with weight, whiteness, or any other measure we can think of.

When you divide a length by a length, for instance, the result has no units at all. It is just a comparative scale. A ratio of any two like-dimensions is a pure number, called a *dimensionless number*. A simple example of this is the aspect ratio of a picture of computer or television screen, which is the ratio of x/y or width over height:

$$\frac{\text{Width}}{\text{Height}} = \frac{16}{9} = 1.777.$$

The height can be measured in any units of length, as long as they are the same units: 16 nanometres divided by 9 nanometres, or 16 fathoms divided by 9 fathoms—it doesn't matter. If the ratio is the same, then the shapes of the two screens are the said to be similar.

If you took a screen that was 16 metres by 9 metres and walked away from it from a safe distance, then held a 16 inch by 9 inch screen in front of your face, it would exactly cover the larger screen in the distance. It would overlap perfectly. The screens would be *similar*.

Newton, and several other scientists after him, took dimensional analysis much farther than this. They applied the idea not only to shapes, but to any pattern of behaviour in a system that could be characterized by measurements. Newton thus pioneered what, today, would be referred to as *dynamical similarity*, or in his language *similitude*. In dynamical similarity, dimensionless numbers represent invariants, or universal, unchanging qualities of systems, whereas dimensional numbers always lead to qualitative change,

For example, if a supermarket knows that 3 persons per minute enter the supermarket, then the store will begin to fill up with people unless at least 3 people per minute leave the store (the same argument applies to users arriving at a website). The queues will begin to build up at the cash registers unless these numbers are in balance. If we write down the dimensionless ratio of these rates, the result is a service ratio:

$$\frac{\text{Checkout rate}}{\text{Arrival rate}} = \frac{\text{1 person per minute}}{\text{3 persons per minute}} = 0.333.$$

Because arrival and checkout rates have the same dimensions (persons per unit time[30]), we can form a dimensionless ratio of them. This tells us that only a third of the people arriving are getting out of the shop, and soon it will be full as the queues grow. We might interpret this as the efficiency of processing the customers.

Suppose now everything speeds up, and 3 customers per second enter the shop, and only 1 per second leaves: then the efficiency is unchanged—we just scaled up the whole operation 60 times. But the shop fills up faster now, because the length of the queue is going to be 60 times bigger. Length is not a dimensionless number and so it is sensitive to the measurement scale.

Only dimensionless numbers are preserved when we scale. They are said to be *scale invariant*. Systems that have all the same dimensionless ratios are called *dynamically similar*.

In physics and engineering, analysis often begins by trying to write down all of the dimensionless ratios one can think of, as these will represent i) all the combinations of parameters that can belong together in a single relationship, and ii) identification of the fixed points or invariants of scaling. For instance, for the television we might write down at least these dimensionless ratios for length and time:

$$\frac{\text{height}}{\text{width}}, \frac{\text{height}}{\text{depth}}, \frac{\text{width}}{\text{depth}}, \frac{\text{time to switch on}}{\text{time to change channel}}$$

The mass of the television is the one-dimensional scale that has no comparison, so we would not expect the mass of television to play much of a role in its behaviour. To the casual onlooker, the significance of these ratios might not be very clear, but an engineer will soon see their relevance to the behaviour of the thing.

Dimensional analysis is common in engineering, where practical issues do not allow engineers to idealize scenarios in the same way that scientists can get away with. Thus there are often more actual scales to deal with: height, width, depth, distance from a wall, etc. All of these scales make a difference in engineering[31].

In our story of stability, the scales of dimensional analysis tell us about *thresholds* that matter in different systems, because invariant thresholds always come from dimensionless numbers. If we can define stability in terms of dimensionless quantities, then we know it will be universal. These insights might have been learnt by studying physical phenomena in the past, but they must also apply to information systems and software in the modern world.

Although the characteristics of stability are summarized by having a dimensionless number cross a threshold, the changes that probe instability must rely on the dimensionful quantities that make up those dimensionless numbers, otherwise they would be invariants (e.g., your age divided by 21 years describes the threshold for ordering drinks in a bar, in many countries). So let's ask, what kind of change would change a dimensionless ratio? What would tip a system over the edge of instability?

As always, we start with a simple model. Suppose we place a ball at the crest of a hill to that it is perfectly balanced, or balance a pencil on its end (see Figure 2.1). The smallest perturbation of the ball or pencil will destabilize it. It would cause the ball to roll to the bottom of the hill. A ball placed atop a hill is thus mechanical system that we say is *unstable* to small *perturbations* of its initial configuration. The smallest change in that configuration unleashes a much larger irreversible change in the system.

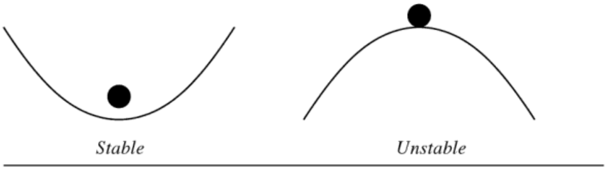

Fig. 2.1. A ball in a potential can be stable or unstable to small perturbations of ball movement. A potential well (left) is stable to movement, but unstable to filling. A potential hill (right) is unstable to movement but stable to filling.

By contrast, if we place a ball at the bottom of a valley and push it slightly, it might roll a little, and then it will roll back to where it started, preserving the original condition of the system. We would say that this system is *stable* to small perturbations[32].

The difference between the two scenarios is the shape of background they are constrained to move on. Hills, wells and valleys are regions of space that exert forces because of the force of gravity. Other forces would do just as well, but gravity is easiest for us to relate to. We call these shaped forces *potentials* in physics, because they represent the potential to convert initial location into a release of later activity. The potential stores up energy that can later be withdrawn from the bank and spent as motion. This is part of the energy accounting referred to in chapter 1.

A valley forms a potential *well* or a constraining potential for a ball, because gravity tends to pull a ball downwards so that it can't escape from the container, and the sides hinder sideways motion to escape. A hill, on the other hand, forms a potential barrier to something on either side of it because a ball would not necessarily be able to roll up hill and over the the other side without meeting the resistance of gravity wanting to pull it down again. For a ball sitting on top, it is a highly precarious, unstable potential.

This simple thought experiment is one of many notions of stability used to characterize systems, and a helpful analogy. What the definitions of stability all share in common is the idea that the state of a system is not significantly altered when we perturb systems by a small amount. Stability is therefore associated with minima (the lowest part of the wells) of a potential, and unstable points are associated with maxima (the highest part of the peaks).

Isn't the idea of perturbing a system an artificial way of deciding its stability. Isn't it cheating? The idea is that, you don't need to arrange for an external perturbation. If the system is unstable, something in the environment will eventually arrange to perturb the system for you. Designers who don't believe in Murphy's law, that which can happen will happen, are irresponsible. Many mistakes have been made by assuming such perturbations could never happen. The John Hancock Tower in Boston, Massachusetts is a telling example. The tower is a high-rise building with glass windows all the way up. During the design of the building, architects did not take into account the effect of wind on the building. Wind-generated torsion on the building[33] quickly led to one of the most embarrassing crises. Torsion is a rotational, twisting force. When windspeeds exceeded 45 miles per hour, the wind would catch the bevelled edges of the building resulting in torsional oscillations of the building's frame. This in turn caused the window frames to transmit sudden force to the glass, which then popped out and fell hundreds of feet down to the ground. Until the problem was solved, the police would have to close off the streets whenever the wind speed exceeded 45 miles per hour.

We can talk about stability that is natural or artificially maintained, stable and meta-stable. Or to put it another way, intrinsic and extrinsic stability. The valley example is an intrinsically stable system. By virtue of its configuration of parts, it is programmed to remain stable as long as the actual configuration is not obliterated by some overwhelming force. The ball atop the hill could be stabilized, for instance, by putting wedges under the ball on each side to prevent it from falling when small perturbations strike. Or we could make a small dent in the top so that it has a notch to sit in. This would be called *meta-stability* at a local minimum of the potential (see Figure 2.2). The configuration is still basically unstable to all but the very smallest perturbations.

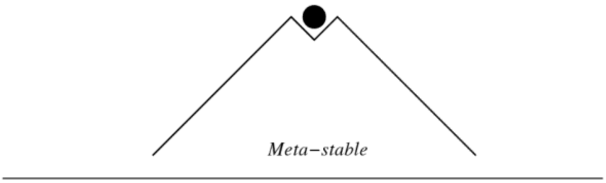

Fig. 2.2. A ball in a potential can be meta-stable, if it finds a local minimum, or dip in the potential.

Lasers are an example where meta-stability can be exploited to advantage, leading to amplification of effect. The word laser began as an acronym for Light Amplification by the Stimulated Emission of Radiation. Stimulated emission is a phenomenon in quantum theory, in which you get many electrons to balance on a kind of quantum high wire and then all jump off at the same time, to hit the ground at with combined impact, except that the high-wire is a special quantum state.

The idea of a laser is to pump atoms with photons (quanta of light) so that they end up sitting in a meta-stable state, and once they are all sitting on their high wire, to perturb them all at the same time so that all the atoms release their photons at the same time. This is called stimulated emission of light. It means that a lot of light is released in a short space of time, hence there is amplification. The idea is a bit like the less glamorous case of a toilet flush: one accumulates water in a tank, and then the flush releases all at the water at the same moment to amplify the effect of the tap feed. It only serves its purpose then the dimensionless ratio of water in the tank to the water in the basin reaches a critical threshold.

Intrinsic stability is not necessarily desirable; it might also be a nuisance. It can leave a technology unresponsive. An example of this is aircraft design. Passenger jets are designed to be as stable as possible. We basically point them in one direction and they fly in a straight line. Since they are designed to preserve human life, flight stability is a top priority. A fighter jet, on the other hand, prioritizes maneuverabilty. If a missile comes out of nowhere, it needs to be able to respond instantly. It therefore wants to operate right on the edge of stability. Fighter jets use computer-controlled flight systems that automatically compensate for minor perturbations, and the pilots do the rest. These systems correspond to someone watching the ball at the top of a hill and being ready to catch it and put it back if it should start to roll. A passenger jet can basically fly itself just as a ball can sit comfortably at the base of a valley.

The keyboard on which I am typing this text has keys that spring back and recover when I have pressed them. They spring back faster than the maximum speed of my fingers. This makes the keyboard stable and usable. Had the keys recovered much more slowly, the keyboard would have been useless to me on the timescale of my usage. If I had to wait an hour after pressing each key, I would not consider this usable under the dynamic perturbation of my fingers.

So much for the physical world, and its pitfalls. What about the unphysical worlds of our own creation: rule of law, computers and information systems all affect the functioning of complex interacting systems and they too can also be stable and unstable.

Traffic rules are a good example of this. In many countries that drive on the right hand side of the road, and before the invention of roundabouts (road circles), rules like "Wait for anyone ahead of you on your right" have been common. The idea is to allow people to enter traffic and avoid collisions, so that one of the parties knows that it has to wait for the other. Such rules are still in used many places.

A rule like this is not a physical thing, but it is easily seen to be unstable, because it tends to stop traffic rather than keep it flowing. For small side roads on non-busy trunks, this is an acceptable solution, but suppose there is a crossroads. Sooner or later four cars will arrive that this crossroads junction all together and everyone will see someone on their right, thus all of them will wait for each other. The ratio of distances from the junction have play a critical role in this instability. Assuming that the drivers actually follow the road rules, they must all stop and wait for the rest of their lives. All traffic is now dead, with no possibility to restart until someone breaks the rule.

Treated to the letter of the law, the behaviour resulting from this rule is unstable because the rule is reevaluated on a continuous basis by drivers who get stuck in a single loop of thinking. This is called negative feedback. Of course, it can easily be modified to allow the drivers to randomly wait and have a go if no one else is moving. The point is that even the careless design of rules for information systems can result in harmful behaviour. Traffic, after all, is part of the global infrastructure of our planet today.

A slightly more stable approach is the stop sign junction, as common in the United States. This uses a first-come, first-served approach: the car that arrives first gets to use the junction first. If four cars arrive at the same time, there is still a deadlock, however. The roundabout is a further improvement, designed to keep traffic moving by waiting only for traffic that is already on the roundabout itself.

What we see here is a bridge to something new: not the motion of a system running out of control, but its logic. Whenever humans are part of a system, there is the possibility of a *semantic* instability. Here, the feedback on the system is coupled by an interpretation. This will be important later in the book when we discuss economics games and the many manifestations of *reasoning*.

The semantics of a technology or system are its intended interpretation and its behavioural response to perturbations. One can imagine a sort of 'smart potential' function that adapts to different scenarios with built-in rule sets, or internal logic. Computer software behaves in this way.

The fact is that humans are not simple robotic response agents, however; we interpret and adapt and act according to a model of the world that we build on the fly, often heavily influenced by emotions. That introduces a lot of uncertainty into a feedback loop that involves humans, and there is both potential for improving and destroying stability, depending on how well placed a person is in the total system.

Here is an example of semantic instability: imagine a fearful member of tribe A, who gets caught in a trap. He sees a member of tribe B coming, who pulls out a knife to free him. The trapped tribe A member interprets this as an attack and kills the member of tribe B (the semantics of the situation are perturbed by his fear). The instability triggered by the pulling of a knife is now in motion. Tribe B declares war on tribe A and attacks, and both tribes descend into a long and protracted war over a simple misunderstanding. The escalation driven by emotional intensity creates the unstable potential for the amplification of hostilities. The threshold was perhaps fear, which could not be measured practically.

Another example of an unstable semantic or logic rule is one that was built into certain computer systems to shut down the computer if the processor began to overheat[34]. This is just like the ball atop a hill situation, waiting for a hot day to knock it off its perch—or, in fact, for hackers to perturb it.

When faced with a rule to shut something down, or deny access to a resource, in civil society[35], something will always push the wrong button, or someone will find a way to exploit it for their own benefit. That is why one usually hides such buttons out of sight. (It would be truly unfortunate if the engine stop button on a plane was located next to the intercom button.)

In several cases, hackers have been able to write computer programs to shut down systems without even having direct access to the system. This is possible if a chain of events can propagate influence indirectly, in a cascade or avalanche effect. In the case of the heat sensor, malicious software was able to perform intense calculations on the processor, generating a lot of heat which would then cause the protection system to kick in automatically—and the system would go down. Even without a malicious intent, bugs in perfectly well-intended software have also caused this to happen[36].

The most terrifying example of a feedback instability is the tragic case of Air France Flight 447, a scheduled commercial flight from Rio de Janeiro, Brazil to Paris, France. On 1 June 2009, the Airbus A330-200 airliner serving the flight crashed into the Atlantic Ocean, killing all 216 passengers and 12 aircrew[37].

Aircraft in flight are only ever meta-stable. Obviously they are poised to fall out of the sky if anything critical stops working as designed. They are extrinsically stable, held in balance by force. Even the most stable gliders that require no power can stall if they rise too steeply, or enter a dive from which they cannot recover. That was the case with the Air France Airbus flight. The aircraft crashed following an aerodynamic stall that was precipitated by a quite different failure. This precipitated failure is sometimes called an avalanche or cascade failure. It is a way of saying that, once a particular condition has occurred, the destabilization of one threshold triggers the next threshold, then the next, leading to a number of successive failures.

According to reports, ice formation in the airspeed measuring device, entering bad weather, led to an inconsistent airspeed readings. The software rules then said that the autopilot should be disengaged, as it would be unreliable. Misinterpreting the effects (a semantic instability) as a sudden loss of altitude, the pilots pulled the aircraft nose up, which further destabilized the aircraft. Despite stall warnings, this resulted in insufficient aerodynamic lift, and a complete loss of powered flight. The pilots had not received specific training in "manual airplane handling of approach to stall and stall recovery at high altitude"; this was apparently not a standard training requirement at the time of the accident. First there was the formation of ice, leading to failure of a part of the system designed to monitor airspeed. Airspeed is a controlling dimensional factor in generating lift to stay in the air, relative to angle of ascent[38]. The pilots were part of the feedback system, in charge of interpreting the semantics of the situation and responding accordingly.

From the rhyme that opens this chapter: the Moon affects the surf, I've heard, by law of physics great. Our Moon seems like the last thing to be unstable, but it too leads to a feedback system through the gravitational interaction with tides. It has been with us for four and a half billion years, yet the orbit of the Moon is also currently unstable to the tidal forces between the Earth and the Moon, on a very long timescale. Both the Earth and the Moon raise tides on each other, and this drag effect slows the Earth slightly and accelerates the Moon slightly, raising its orbit around the Earth. The Moon is thus moving away from the Earth at a measurable rate of about three centimetres per year. The system will stabilize when the Moon's orbit becomes geosynchronous.

There does not necessarily have to be physical force involved in instability. Influence requires only a transmission of information. Even messages and stories that we tell one another interact with their listeners. The concept of 'memes' was introduced by Richard Dawkins in his book *The Selfish Gene*[39]. Dawkins pointed out that certain ideas can be self-propagating. Not every idea we have is successful at being transmitted from one person to another, but certain songs, phrases of music, words, or even ideas seem to have the knack of triggering some kind of an affinity with our minds and they spread out of control. This might be called a cultural instability, though the mechanisms for memetics are not fully understood.

The foregoing examples illustrate ideas of instability using the analogy of motion around hill-valley potentials to a single perturbation. Illustrations are helpful in putting together a picture of what instability means for the short term, but for engineers to create infrastructure in our world that is stable, one needs to go beyond mere visualizations and arguments, to indisputable models that endure over time. With the discussion of scales under our belts, we can now put together a more continuous model for stability than this one-shot perturbation response.

A response to a single perturbation can be either linear or non-linear, in the sense of Chapter 1. If it is linear, the system can usually be perturbed many times and respond in a helpful, functional way. This is how we make physical tools and information-based services. A web server is a piece of computer software that receives requests and returns information in proportion to what was requested. If a web server suddenly began to return information, quite out of control based on a single request, it would be a sign of a non-linear response.

Feedback systems are part of a dialogue between something driving the input and something coming from the output. The output of the system changes and becomes part of the new input in a new iteration. This goes around and around and the system either stabilizes to a well-defined value or it blows up.

The corrective potentials are a form of what is called *feedback* in control theory. Feedback basically imagines that a dynamical system is looking at itself in a mirror and asking: are we there yet? Negative feedback tends to moderate the system, making it converge towards an equilibrium value. Positive feedback amplifies the input and feeds it back into the input again so that it grows and grows out of control. The shrieking noises from public address systems, when a microphone gets too close to the loudspeakers is an example of positive feedback.

If we are going to make infrastructure, avoiding non-linear feedback is usually an important design consideration. Non-linear systems are not predictable, as we discussed with the weather. Why do non-linear systems diverge? The answer is that they always feed back on themselves. If they feed back positively then the system is intrinsically unstable.

The idea of linear response offers a way to extract something like a signature for stability in infrastructure. To ignore non-linearity might seem over-optimistic, but to engineer a stable world, it seems clear that we must avoid non-linearity at all costs.

Consider the following symbolic representation of stability:

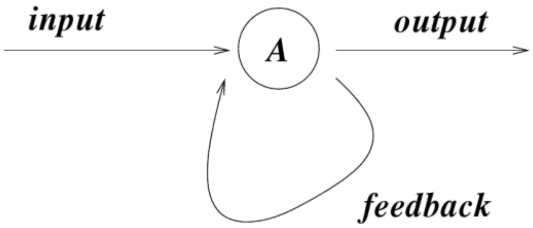

Fig. 2.3. In a feedback system, part of the response or output is fed back into the input either subtracting or adding to the input. In continuous operation this either moderates and stabilizes or amplifies.

$$\text{(Act on) thing} \longrightarrow \text{thing}$$

I am using this to mean that when we act on a thing, we get back the same thing. We call the '(Act on)' symbol an *operator*. In mathematics, we might write something like

$$Ax \longrightarrow x.$$

It's just a way of writing the act of making a change to the object concerned (in this case: "x"). As a definition of stability, it is closing in on something sufficiently formal, but the statement is too strong. What we've defined is actually called a fixed point. A system doesn't have to be fixed in exactly the same configuration at all times, otherwise we could not allow anything to happen at all. So we revise this to an equivalence:

$$\text{(Act on) before} \overset{equiv}{\longrightarrow} \text{after}$$

We propose that the thing only has to be *dynamically equivalent* in some sense before and after the operation of being acted upon. The criterion for that equivalence must be left to the context, because it depends on the dynamics of the system concerned, as well as the level of precision we are willing to tolerate[40].

For instance, if we act on a ball by rotating it 90 degrees, the ball is not strictly in the same state, but it is in a new state that is absolutely indistinguishable from its former state. Symmetry thus plays a deep role. If we can't tell the difference before and after a change, the situation must be dynamically equivalent. It is by extending and developing this idea in a more technical and precise way that we can find other ideas of stability.

Notice that the (Act on) operator has to be dimensionless (scale free) in order to lead to a scale invariant fixed point. We can see that because an equation has to have the same dimensions on both sides of the equals sign. So if we write

$$Ax = x,$$

then, because x obviously must have the same dimensions as x on left and right, A must be dimensionless.

A similar mathematical concept with the flavour of network stability is that of *eigenstates* or *eigenvectors* and *eigenvalues*, from the German 'eigen' meaning self. These are intrinsic values that can be calculated by relatively simple mathematics, and belong to the interaction between a particular operation and the body it can acts upon. Eigenvalues crop up widely in physics, engineering and mathematical disciplines, as a method of locating stable aspects of complicated systems. Statistical methods like Principal Component Analysis, and the Principal Axes of a mechanical body use eigenvalues to determine stable configurations. Recall the strange spinning of the spaceship *Discovery* in the movie *2010*, the sequel to Arthur C. Clarke's and Stanley Kubrick's famous *2001, A Space Odyssey*. The ship seemed to be twirled like a cheerleader's baton in Jupiter's gravity. In fact this was an accurate depiction of what can happen to the spacecraft design, when a long thin shape is left in a complex perturbations like the orbit of Jupiter. It will tend to rotate about its axis of maximum moment of inertia. Eigenvector stability was also effectively used as part of Google's PageRank algorithm for ranking the importance of sites on the Web.

Fixed points and principal eigenvectors are important anchor points for technology builders. If we could always build things around fixed points to all perturbations, they would be intrinsically stable.

To examine stability in information systems, we need to develop some of these ideas. By combining the idea of operators above with the picture of the perturbations in a potential, it is straightforward to come up with a ream of examples of stability and instability. The simplest idea of stability is constancy, or invariance. A thing that has no possibility to change is, by definition, immune to external perturbations.

$$\Delta v = v_{\text{after}} - v_{\text{before}} = 0$$

If the value of something (its position, its speed, its colour, or any other verifiable property) is the same before and after an operation is applied, then we can say, without much controversy, that it is stable under that operation. In physics, one calls this *invariance* under the action of the difference operator Δ.

Invariance is an important concept, but also one that has been shattered by modern ideas of physics. What was once considered invariant, is usually only apparently invariant on a certain scale. When one looks in more detail, we find that we might only have invariance of an average. That is an idea we'll return to in Chapter 4.

The next level of stability is sometime called *covariance* (meaning varying with). If we perturb the system, it changes a little in a predictable way, but retains most of its characteristics even though some details change. This is a structural stability[41]. Galileo's inertial principle is an example of covariant stability: a free and unconstrained body will continue in a state of uniform motion in a straight line (constant velocity) unless perturbed by an external force. During the perturbed, it will shift speed or direction slightly, and thereafter continue on its uniform steady state.

The situation is different if the body is not free: for instance, if it is sitting in a confining potential, like the valley in Figure 2.1. That is the case for planets in orbit around a star, or Moons in orbit around the planets. The notion of a *steady state* can be characterized in many different ways. Indeed, in physics or mechanics constant and uniform change are basically equivalent.

A planetary orbit is a simple image that everyone can understand of something dynamically stable, yet still moving. An orbit is the balance between the inertial drag of the planet which tends to move the body in a straight line, and a sun's gravity that tends to pull it towards its centre. Planetary motion happens on such a large scale that it belongs to the scale of continuum, Newtonian motion. What about an orbit of discrete states?

The traffic lights, discussed in the previous chapter, give us this a cycle of behaviour that is stable to perturbations (see Figure 1.1). The lights always move through the same pattern of red, red and amber, amber green and back to red, regardless. If someone presses a button to cross the street, the state might change sooner than expected, but the pattern is stable. It is also a kind of orbit, but in a different arena.

Fixed cycles of behaviour, like orbits and periodic sequences, are called *limit cycles*. Rather than converging onto a single end-state, like say falling to the bottom of a well, a changing system might retain some dynamical activity and end up in a forever repeating pattern, like pendulum swinging back and forth, or a satellite going into orbit instead of crashing into the planet below. For instance, our daily circadian sleep rhythm is a cycle driven by changing levels of hormones, like melatonin, whose production is influenced by daylight. When we perturb environment by changing timezones, after a long flight, the limit cycle of our sleep is disturbed and takes some time to restabilize.

The measures of stability described above are largely about individual bodies, like planets, balls or lights, yet not everything in the world is this simple. Our bodies are complex networks of interacting cells. Materials form multitudes of different networks through chemical bonding, giving each material its characteristic properties, and humans form networks through communication.

The stability of networks thus turns out to be of enormous importance throughout science and technology. Networks are everywhere, not merely connected to our computers. Food-webs, ecosystems, economic trading patterns, chemical interactions, immune systems, and of course technologies such as water and sanitation, electricity, and all of the major infrastructure items are transport networks for something vital to our life and culture. The question of how stable networks are to perturbation is thus of key interest to human technology.

The Internet itself was envisaged as a form of network stability. Consider the simple networks depicted in Figure 2.4. Diagrams like these are called *graphs* in mathematics – they are simply nodes connected together by links, or edges. The example on the left of this figure is a 'tree' (also referred to by the more technical if less poetic name of an 'acyclic graph', meaning that it has no closed loops[42]). Tree graphs have branches and leaf nodes, but they are minimal structures that contain only a single possible path from trunk to leaf, or from any node to any other. The implication of this is that, if we block or damage on of the pathways, a part of the network becomes inaccessible.

The dashed rings in the figure show places where the destruction of a node would disconnect the graph so that it would no longer be possible to cross from one part of the graph to another. Such a point is called a *single point of failure*. That is not to say that there can only be one of them, rather that the removal of just one single point can have a catastrophic structural effect on the network.

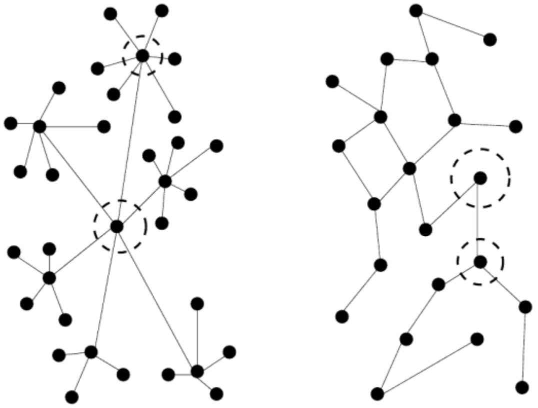

Fig. 2.4. The reachability of one part of a network from another can be stable to the perturbation of node-removal.

A chain is a graph that is not stable to node removal. It works as a very simple model of a set of dependences. If we turn it vertically, it becomes a tower, with each successive node depending on the one beneath it[43]. Chains of dependent parts make systems fragile and unstable to catastrophic failure.

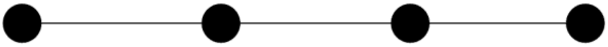

Fig. 2.5. A chain is the most fragile structure of all. The removal of any node breaks it. Thus all nodes are single points of failure.

A tree is a generalization of a chain, only with branching too. Chains and trees are the most fragile structures we can make, because they are built of many 'single points of failure'. As engineering principles go, this is not a good structure to rely on, especially if human lives are involved.

The graph or network on the right hand side of Figure 2.4 is not a tree. It contains closed loops that make it more robust. A loop means that there is more than one way around the network, i.e. there is *redundancy* or *backups* built into the multiple pathways. Networks that contain redundancy are more robust to failure, or stable to node perturbations. So, for instance, in the upper parts of the graph, the removal of a node would not disconnect the graph, we could route around an incapacitated node.

The idea that it would be possible to make networks that were robust to damaged nodes was part of the idea that spurred on the invention of the Internet as we know it today. During the 1960s, the Advanced Research Projects Agency Network (ARPANET) was the world's first operational packet switching networks, designed from conceptual studies by Paul Baran of the RAND Corporation. Baran suggested that communications could be made more stable to failures of infrastructure if messages were formed from discrete packets of information, and if the networks themselves ensured redundant pathways. The three levels of network robustness in Figure 2.6 illustrate going from a tree-like network with a critical point of failure on the left, to a *mesh* network on the right with multiple redundant pathways. The Internet was designed to be like the right hand mesh, so that, in the event of attack or other natural disaster, communications could continue uninterrupted by re-routing information. This is the system we have in use today, on an elaborate scale. We see, in action, how stability is exploited as a principle in order to make a reliant infrastructure for our society.

The foregoing examples have been continuous systems, apart from the network, which is discrete. It is also possible to generalize the initial perturbation

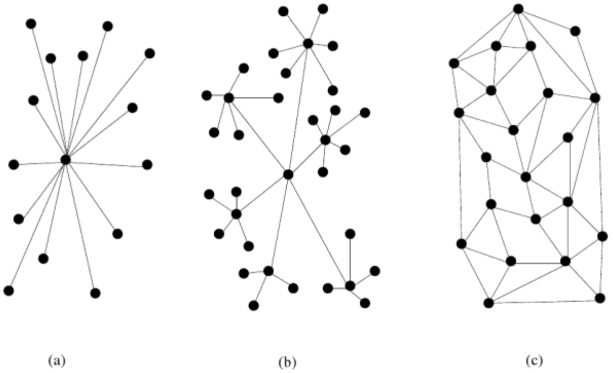

Fig. 2.6. The reachability of one part of a network from another can be stable to the perturbation of node-removal.

argument based on Figure 2.1 so that it applies to discrete systems. Why bother? The example of the Internet is one reason, and a good illustration of the basic robustness of discrete units. Another reason is *software engineering*.

The designers of our large distributed computer systems (including the well-known Internet giants) are starting to learn about the importance of stability, not only in terms of operational stability, but also data correctness and consistency. We'll return to the details in chapter 5, but we can briefly mention the key idea of converging to a known end-state. In particular, in the world of functional programming, software engineers have begun to ask some of the hard questions about why systems become semantically unstable, especially at large scale. They refer to the problem by the name *data consistency*.

There are various formulations of the idea, but the simplest visualization is in terms of networks. Potential wells can be supplemented with a result from the theory of graphs, called internal and external stability[44]. For example, in the theory of graphs, one can identify the concept of an *externally stable* region, i.e. one in which all points outside of the graph region point into a region by a single link. Thus the smallest perturbation of position around the region would be to push something by a single hop into the region, like rolling into a potential well. It is a stable place because once in the region, there is no way out. An *internally stable* region is a place in which a perturbation of position cannot take you to any new point that is stable, i.e. it says there are no links between different stable nodes, like distinct, alternative wells. Functional programmers often appeal to the language of category theory and universal algebras, where these structures are called *semi-lattices* and *semi-groups*.

External stability tells us that we can always get to the stable region from outside it. Internal stability tells us that none of the nodes in the region are connected, so if we arrive at one of these nodes we have either reached the end of the line, or we have to leave the region again. A region that is both internally and externally stable is called the *kernel* of the graph, and it represents a trap, i.e. a place such that, if we enter, we can't leave. There might be several such traps in a network, different by some criterion, but they are all structurally stable confining potentials.

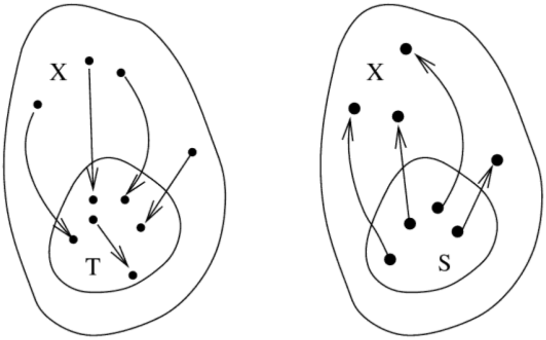

Fig. 2.7. The stability of region can be generalized from the potential well/hill view to the idea of external stability (left) and internal stability (right).

If we write the links that flow between nodes as (Flow), and the set of nodes in the region as R, then the kernel takes on the general form above:

$$(\text{Flow})R \rightarrow R$$

If you try to flow from some place in the region, you just end up back where you started, or once you're in, you're in. This is a special case of the general stability criterion: (Action) state → state, from earlier.

Each of these stable points is referred to as a *fixed point*, in mathematics. A fixed point is a place where you can end up and remain in spite of the specific perturbations that are acting. It is a self-consistent place within a system. The existence of fixed points in systems is a more or less deciding factor for the existence of a stable solution. The presence of fixed points was used to prove some of the key results in a economic game theory, for example.

When we think of networks not just as locations in a physical geography like the Internet, but as possibly abstract realms, e.g. like the map of transitions between coloured states of the traffic lights, networks become one of the most important tools in modern technology. They are used everywhere to map out the behaviour of complex interactions. They are a distinctly non-Newtonian idea, but they are a stepping stone to a deeper understanding of things that is the subject of the next chapter.

An interesting example, actually the one I came upon during my flight home from the 11th Annual Conference in Large Installation System Administration in 1997, was the human immune system. It can be regarded as having a kind of fixed-point kernel behaviour[45] that represents a state of health. Germs and viruses are perturbations to our bodies, that move us from healthy states like 'green' to poorly states like 'amber' or even deadly states like 'red'. Clearly that is a pretty simplistic view, but the principle is accurate. Viruses bring instability to our cellular processes because they re-program cells to grow and manufacture more viruses, instead of carrying out their original healthy program. They then act like cluster bombs, bursting open and releasing multiple copies of themselves to pervert neighbouring cells.

Biology has evolved stabilizing counterforces like the immune system present in vertebrates, and simpler homeostatic regulation systems in warm blooded animals[46], as a method of pseudo-intrinsic stability. It is intrinsic to our bodies, but extrinsic to the cells that are attacked, just as the CFEngine software written for computer networks was intrinsic to individual computers, but extrinsic to the files and processes it was protecting[47].

By equipping a computer with a kind of internal 'groundskeeper' or immune system, system engineers would effectively be able to architect a lasting set of *desired end-states* (like in a graph kernel) that would not immediately be attacked by the forces of decay. CFEngine accomplished this by making this groundskeeper a software robot whose specialized set of operations had to be *convergent*, meaning stable like a potential well or a graph kernel.

As long as a designer expressed all the desired system properties in terms of such fixed-point, convergent operations, the system would actually self-heal under perturbations, just by repeatedly applying the operations in a dumb manner. This is exactly what the immune system does. There is no intelligence, beyond understanding when a system is in a desired end-state. The idea was thus to create a counter-force to other, less well-behaved perturbations from outside. As long as the fixed points could be stable on the same timescale as the external perturbations, one would create an extrinsically stable state for the computer system.

These examples hint at the idea of dynamical balance, and are just cases of a more common idea: that of performing regular maintenance on structures that we build. We watch and touch up the paintwork on our houses, or dust shelves, or vacuum clean the floors. The storm drains in the streets and gutters on your house are a very simple example of a stabilizing mechanism to the perturbation of rain. All these point to a notion of dynamic stability, and are the subject of a special treatment in Chapter 4.

There is a lot to be said about stability. In one sense, it is the counterpoint of unpredictable change, and the foundation of certainty and predictability. We build on these qualities to bring reliability to the things we make. Stability takes many forms. We can approach it, as science would approach it, using abstractions and models, and we can describe it from its many exemplars. Both of these approaches will help us to understand its significance. In this chapter, I have tried to convey a little of both.

To apply these observations to technology, we need to understand the nature of the technological beast: what makes it tick, and just how clockwork is it? For that, we need to turn back a few of the pages of history, and follow its threads through the great skein of scientific understanding that led to today's technological revolution.

3

Digilogy: Cause, Effect, and Information

How transmitted information shapes the world from the bottom up.

> "We are thus obliged to be modest in our demands and content ourselves with concepts that are formal in the sense that they do not provide a visual picture of the sort one is accustomed to require ..."
> – Niels Bohr, Nobel Address

One of the running jokes, between myself and another British colleague, during the early years of founding a company around CFEngine, was my job description. "Mark is the CTO—he does *stuff* with *things*".

When two British heads come together, a wicked resonance of absurdity known as British Humour$^{\mathrm{TM}}$ typically emerges, resulting in an unstoppable if confusing force over which the remainder of the cosmopolitan workforce generally roll their eyes in pity. Doing stuff with things seemed like the most appropriate irony for the puffed up job of a Chief Technical Officer who had neither feathers nor military honours.

Still, never missing the opportunity to ruin a good joke, the recollection of these exchanges reminds me now of a crucial difference between stuff and things that forms a stepping stone on the path to control over technology; and so I must proceed now to dissect it for the good of the present story. In brief, the point is this: there might be much stuff, but there are only so many things.

If you are still with me, 'stuff' is a *continuous* measure, like 'bread' or 'gold', but a 'thing' is a *discrete* measure like 'a loaf' or 'a bar'. Hence: the streets are lined with bread (or gold), and 'tis a far far better loaf I bake' (or thing I do), and so on. All of this is highly relevant to the notion of scale, because it forces us to confront the many different ideas we have of *atoms*.

In spite of what our senses tell us, the world is not really a continuum of stuff. It is not the smooth and undulating world of Newton, with perfect spheres and light strings. It is rough, granular and random, even unpredictable, when we dig down. It just appears to appear continuous to our imperfect senses, because it is a collection of very many extremely small things, too small to resolve individually. Scale, our friend. Scale, the deceiver.

We[48], of course, are made up of cells, and each cell is made up of a rich variety of molecules, each of which is made up of any number of atoms. No one quite even knows where this story ends. The key point is that, if we look closely enough, we find a world of things and not stuff. The world seems to be fundamentally *discrete*[49]. Nonetheless, we maintain the fiction of continuity, because it is convenient, both to our senses, and to the mathematical models of our world. The so-called *continuum approximation*, of discrete systems, has led to marvellous insights and advances in the science of all kinds of 'things'.

The word 'discrete' (not to be confused with 'discreet' of discretion) means separate and distinct. In other words, it can mean the opposite of continuous, which is the usual meaning in science. In order to scratch beneath the surface of technological challenge, we need to appreciate what discreteness means, and why it is fundamental to the world. Why should we care about the nature of the physical world of the very small, if we want to build smart infrastructure on the human level? The first reason is that we have to understand the true nature of discreteness, and what makes it different from the continuous world in order to understand information systems. This is the key to fully understanding technology at every scale. The second reason is that we can learn from physics how to understand the emergence of dependable continuity from a discrete foundation. The third reason is that our technology is getting so small that the quantum nature of the physical hardware is quickly becoming the one we need to think about.

Discreteness turns out to be fundamentally important for understanding almost everything about the world of information. Moreover, our modern understanding, in terms of discrete objects, has many similarities to the quantum theory (which refers to countable or quantifiable things), and also to *information theory* (which refers to the methods of communication). This chapter is about those two developments and their importance to the modern world.

The story of how we arrived at our understanding of discreteness, and its significance to the world, is one that stretches back into ancient times, to the very earliest recorded philosophers; and it ends with our modern view of information and information-based nanotechnology. It is surely as profound as the very question of why we exist at all, and it is a link between the concepts of continuum scale, put forward in the foregoing chapters, and the digital world. For context, I want to relate some of it.

Between the 19th and 20th centuries, the entire view of matter and spacetime changed radically from what we might call a classical view to a quantum view. From the earliest systematic investigations about the nature of light and matter, going back to Christiaan Huygens (1629-1695) and Newton in the 17th century, it was evident that some kind of discrete, particle-like nature could be attributed to light, as well as a wavelike nature[50]. Newton supported a *corpuscular theory* of light, originally proposed by a French scientist Pierre Gassendi, believing light to be made of discrete particles. However, phenomena like the bending of light by refraction and diffraction could only be explained in terms of waves. Thus Huygens retained his belief in a continuous wavelike nature of light, resulting in the Huygens-Fresnel principle that is key to much of modern optics.

Matter and energy: particle or wave? Discrete or continuous? That has been the question for philosophers and scientists throughout recorded history. This apparent bicameral nature of the things and stuff that make up the world remains one of the most baffling features of modern physics, right up to our present day understanding of quantum mechanics. Both views can apparently be true at the same time, and yet waves and particles are the very opposites of one another in classical thinking—the ocean versus the ship.

Resolving this mystery has been the work of a century already, one that began with the discovery of atoms. It is hard for us to imagine now, when every child with a basic education knows about the existence of atoms, how this could not be obvious. However, by following some of the history, we shall see why we continue to make the same mistakes today. Emerging from the technology of steam (in which both steam and heat seemed to behave like fluids), the idea of microscopic particles was an abstract and difficult idea to concede, and it took up to the beginning of the 20th century to approach the modern idea of atoms.

It was Democritus (460-370 BC), of ancient Greece, who coined the Greek adjective *atomos* (ἄτομος) meaning uncuttable, from which we have today's word 'atom'. Today we know that atoms are, in fact, not uncuttable at all, but may be broken down into smaller pieces, including protons, neutrons and the electrons that make up electricity.

Atoms are not only a useful concept for matter, however. The concept is also used, metaphorically, in any number of sciences and technologies for ideas or objects that cannot or should not be broken into smaller parts. An atomic transaction in banking or database engineering, for instance, is an indivisible exchange of money or data. The identification of fundamental, elemental parts, or basic building blocks, is a key part in our understanding of the way things work.

Suppose you hand somebody a coin; that is an indivisible transaction: it either succeeds or it doesn't. The person either gets the value of the coin or they don't. The coin cannot be cut in half, it is an 'all or nothing' exchange. Yet still we see on our bank print-outs that we can have fractions of a coin in value, due to interest accrued at a rate that does not respect the discreteness of monetary value. Similarly, with databases, financial transactions and other high value information, 'stuff' is handed over as 'things' in a way that ensures this 'atomicity'. The concepts of data integrity and logic are closely related to the atomicity of bits in modern computers.

The philosopher Zeno of Elea (estimated to have been born around 480 BC) fanned the flames of indivisibility in matter by suggesting the following: if the universe were in fact infinitely divisible, it would mean that the basic building blocks of stuff would have to have zero size. That meant that nothing would have size at all, and no motion would be possible. Indeed, a thing of zero size does not even exist! If one tries to construct something from nothing, one meets the embarrassing obstacle that even the sum of an infinite number of points of zero size is merely zero[51]. Zeno followed up the argument by presenting a paradox of movement in various forms, including the famous race between the tortoise and Achilles. He argued that, if distance could actually exist independently of things, then a thing trying to move in this space would have to cover an infinite number of points between the start and end of its journey, which was no more helpful than if space itself could not exist. A moving body would first have to cover half the distance, then half of the remaining distance, and half of that again, and so on. Because there would always be half left, the process of subdivision could never end, meaning the thing could never arrive at its destination.

The full discussion of Zeno's paradox is complex, and tends to subdivide as infinitely as the journey it describes, but we can make two notes about it that are somewhat practical. The first is that the notion of distance itself is not refuted by it, as we know that the sum of distances forms a geometric series that sums to the full distance. Suppose that the total distance is 1 metre, then the distance D travelled by Zeno's adventurer is:

$$D = \frac{1}{2} + \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \ldots$$

The answer for this sum is known to be given by the formula $D = a/(1-r)$, with a being the first term and r being the multiplier: $a = r = \frac{1}{2}$, gives $D = 1$. In other words, the series does add up to the full journey. However, also note that the argument only works because we decomposed the journey into a number of discrete steps, without worrying about how the distance could exist in the first place.

What does not necessarily add up, however, is the cost associated with dividing up the distances and crossing them. It is somehow implicit in the formulation of the paradox that one sees the act of dividing up the distance each time as an additional step. Whether there is a real cost of dividing up the path, even if only a mental cost of thinking about it, there is nonetheless a penalty to this dissection (this is not an unreasonable idea, as most actions do not happen unimpeded in nature; there is usually some kind of penalty in time or effort). Moreover, the cutting cost does not halve with the distance. If we assumes that this penalty, in whatever units it is measured, is constant, then the infinite sum over the distances will cause the time needed to traverse the metre to be infinite, because an infinite sum of any fixed overhead is infinite—surely an important lesson for bureaucrats everywhere.

The expression 'a watched pot never boils' is related to this, as it suggests that the psychology of waiting makes us focus on smaller and smaller intervals of time, as we lose patience with increasing frustration. The mere act of thinking about waiting seems to increase the number of minutes required to boil water, in our perception, perhaps because we think faster and faster about the process, forcing a continuous revision of the dimensionless ratio of:

$$\frac{\text{Time for a unit of annoyance}}{\text{Time for water to boil}} \to \infty$$

to increase beyond some threshold of melancholy or madness.

A simple way to avoid the troubles with Zeno's paradox is to assume that space and time are in fact discrete to begin with, and that there is a minimum size, or a smallest unit that can be measured. Then all of these issues would melt away.

In fact, this might easily be true, though no one today has any detailed understanding of how it might actually work. The idea would solve many problems with our current understanding of the universe, however. This links back to the concept of scale that comes from dimensional analysis, as described in the previous chapter.

We would expect the nature space and time to be universal features of our universe. Where would a single unbalanced scale come from to define the minimum distance between two points? One possible answer is: from the laws of physics themselves. It turns out that one can combine three dimensional scales, believed to be universal constants into a value with the dimensions of length:

the speed of light in a vacuum c, the gravitational coupling strength between any two bodies G, and Planck's modified constant $\hbar$:

$$L = \sqrt{\frac{\hbar G}{c^3}} = 1.6 \times 10^{-35}\text{metres}.$$

This number is called the Planck scale. No one really knows whether it holds any significance, beyond the enthusiastic speculation of theorists (after all we don't know all the laws of physics, so there might be other scales to consider), however it is symbolic of a fundamental length-scale at work, and hence an indication of an intrinsic discreteness to the world we live in.

The unexplained constant $\hbar$ in the Planck scale is called Planck's (modified) constant, and it is the crucial value in the argument about the quantum nature of spacetime. It was introduced by Max Karl Ernst Ludwig Planck (1858-1947) to explain something quite unrelated to the discreteness of space: namely, an unexpected feature of the frequency spectrum of so-called blackbody radiation.

At the end of the 19th century, there was no inkling, in the world of physics, that the true nature of reality might really be discrete. The flagship theoretical achievements of the day were thermodynamics (of steam engines) and electrodynamics (of electricity, telegraph and radio). These were the industrial revolution's highest achievements, and both were continuum theories. Planck, however, almost under protest, turned science's view of the continuum of matter and energy upside down, and showed, with the help of Einstein, that there was a threshold scale beyond which matter and energy simply could not be treated continuously any longer. The failure of classical physics to predict this was known as the *ultraviolet catastrophe*, referring to the radiation scale at which the quantum effects revealed themselves.

Although it bears his name, the Planck *length* was not described by Planck himself, but it honours him as the father of quantum theory, because it is $\hbar$ that effectively inserts the scale at which the quantum mechanical character of nature appears. We assume therefore that it also indicates the scale at which spacetime itself would attain a quantum interpretation. The Planck length is so much smaller than anything we are able to experience that it is impossible to imagine just how small quantum effects are. The fact that we are able to observe any quantum effects is a testament to the ingenuity of humans.

As a landmark in the cultural heritage of the modern world, I think it's important to give a brief account of the discovery of discreteness in physics, and how it played into technology, because it has a profound effect on the way we think about everything today. More importantly, it is a stepping stone to our understanding of the technologies of the very small, and relates quite directly to observed computer behaviour.

Planck's early career is known from his short scientific autobiography[52]. It is relatively rare to have access to insights into the mind of a pioneer from first hand recollections, so we may consider ourselves fortunate to have so much of this history documented.

Like many other innovators, Planck was not encouraged to pursue physics, nor were his innovations received with much enthusiasm for some time. Even as early as the end of the 19th century, shortsighted individuals were naively claiming that there was nothing more to know about science[53]. One physics professor in Munich advised Planck against going into physics, saying, "in this field, almost everything is already discovered, and all that remains is to fill a few holes." Planck, however, was ripe for a challenge, and he showed no lack of courage and diligence in choosing topics to work on.

Planck studied under two other giants of physics Helmholtz and Kirchoff in Berlin, but in his own words he admitted: "I must confess that the lectures of these men netted me no perceptible gain. It was obvious that Helmholtz never prepared his lectures properly ... Kirchoff was the very opposite ... but he would sound like a memorised text, dry and monotonous"[54]. So he chose his research interests from the works of Clausius, another giant in the field of thermodynamics and entropy. His interests were the hard problems of the day: the thermodynamic properties of matter and radiation, the so-called black body radiation.

Coming out of the age of industrial steam power, kinetic theory, thermodynamics and electricity were the dominant problems of interest, although these subjects must seem terribly pedestrian today by the standards of what is trendy in physics. Yet it was time or enormous excitement not merely because of the pioneering inventors like Edison, Tesler, Faraday, but also other pioneering theoreticians like the Scottish physicist James Clarke Maxwell (1831–1879) who developed continuum mathematics of thermodynamics and electromagnetics, and Ludwig Boltzmann (1844-1906) who developed much of statistical mechanics. Boltzmann's kinetic theory of gases, for instance, basically presupposed the reality of atoms and molecules when almost all German philosophers and many scientists disbelieved their existence[55].

In 1894, Planck turned to the problem of black-body radiation, i.e. the expected energy profile for radiation of an idealized body in equilibrium with its surroundings. This was a problem where thermodynamics and electromagnetism met head on, and as with many other unification attempts in physics, it would lead to a major shift in thinking. Planck was inspired by a landmark paper by Wien who had calculated the spectrum of blackbody radiation in a limiting special case—at a particular scale. The failure of kinetic theory alone to explain the high energy behaviour of black body radiation was called the *ultraviolet catastrophe*, because at the scale of ultraviolet radiation and beyond, the results for energy flew off to infinity. Explaining this was the problem he took on.

Planck had no idea that he was about to revolutionize physics. His goal was a rather modest attempt to contribute to the new methods of statistical mechanics introduced by Boltzmann, but his fascination with the concept of entropy led him to pursue a different line of reasoning than many of his contemporaries, and he did not initially follow Boltzmann's prescription to the letter. In fact, his method was reminiscent of modern approaches to the theory of information.

He began using the standard approach in statistical mechanics, using a mathematical trick to make the calculation of thermodynamic quantities easier. One would start by assuming that the energies of light had to take multiples of a whole-number (integer) values: $E_n = \epsilon, 2\epsilon, 3\epsilon...$ and so on. This makes them countable. Later, one would take the limit $\epsilon \to 0$, to get an infinite resolution, a bit like Zeno's limit, and recover the continuity of allowed energies. But Planck found that he could only obtain an expression that agreed with experimental data for blackbody radiation if he did not take this continuity limit. Instead he had to assume that the step size ϵ was non-zero and proportional to the frequency, i.e. $\epsilon = 2\pi\hbar\nu$ at a given frequency ν of radiation, in lumps of size $\hbar$. The expression he obtained for the spectrum confirmed experimental data with great success. Planck published his work in 1900, but the work did not receive much acclaim.

Planck's hypothesis that energies were not fully continuous, but rather came in lumps proportional to a fixed scale $\hbar$, was remarkable. It was this discreteness that prevented the ultraviolet catastrophe from leading to infinities, just as in Zeno's paradox. The same approach would be used many times over the 20th century to avoid infinities in physical calculations.

In fact, it is easy to get the wrong impression about the quantization of energy from accounts of quantum mechanics. Planck's quantization did not say that energy quanta were universally discrete lumps of energy, analogous to fixed atoms. Rather, they had a size that depended upon the frequency of the waves they were part of. For every different frequency, there is a particular size of energy unit $h\nu$. Thus we have not abandoned the idea of continuous *scaling* completely—it is embedded in the formulation of the solution.

Astute readers might see a problem with this. Isn't the notion of quantization inconsistent? The frequency is a continuous classical variable that can take on any value. Can it make sense to have energy quantized, but based on a continuous quantity? Indeed, this is ultimately inconsistent, and in a full quantum theory, ν would also have to be treated as a discrete variable. To see why this worked anyway, think of it as the first level of a better approximation, which turned out to be enough for Planck's paricular case.

The following analogies might help. We already mentioned that the concept of energy is like money, so quanta are like energy coins, i.e. fixed-size units of energy-money. But there is not only one set of coins: there is a different set of coins at each frequency. We cannot divide any given coin up, so money is quantized in the fixed amounts, but at a different frequency the coins have a different size. At higher frequencies, above the ultraviolet, the quanta get very big and more noticeable; below that, they are not noticeably distinct (imagine a smooth powder). The frequency is included as a classical continuous quantity, but this can work for the same reason discussed in previous chapters, namely that there is only weak coupling between the natural frequency scale of a wave and the existence of a state of the system.

It's like saying that there is little connection between the weight of an apple and the number of people eating an apple at a given moment. The number of apples is a discrete quantity, but we can get away with treating the weight as continuous because the quanta that make up the people and the apple are so small that they are not relevant to the eating process. People on the other hand are most definitely quantized as individuals. Apples' weights are really quantized, atom by atom, but no one cares about those microscopic differences in apple weights.

The significance of Planck's discovery lay in demonstrating that experiment had probed reality in way that revealed a deeper truth about its nature. What had seemed continuous and deterministic, was in fact discrete and non-deterministic, when examined closely. Only by zooming out to a sufficiently large scale could one recover an apparent continuum. Thus, determinism is not an adequate model of discrete systems. This is a lesson we are now learning in reverse, while trying to assure predictability in information infrastructure.

Planck's result alone might not have been enough to convince anyone of the reality of quanta. It was rather technical, and the crucial insight was hard to understand. Confirmation came from Albert Einstein (1879–1955), however, in 1905 with his explanation of the *photoelectric effect* in metals.

The photoelectric effect is a phenomenon where electrons are emitted from a metal when the metal is irradiated by light. Experiments showed that electron energies in the metal depended on the frequency of incident light and not its intensity. This was another mysterious result to a classical physicist.

If the classical continuum theory of light had been true, increasing the intensity should have allowed more light to be absorbed in a continuous fashion until an electron had enough energy to escape from the metal surface. However, this was not the case. It depended only on the frequency. Einstein showed that the apparent paradox could be solved if one assumed that the light was quantized, as Planck had proposed, because then only quanta or ‘coins’ of sufficient size could be absorbed as a single lump. The amount of energy could not be gradually accumulated over time, as with a continuous quantity. Einstein won the Nobel prize in 1921 for his analysis of this work.

Despite the success of his discovery, Planck had much trouble accepting the physical reality of a quantum hypothesis. His own words about the struggle to reconcile these world views give us a sense of the ‘quantum leap’ he had to make to give up on continuity. Most likely, he worked backwards from the result, then he spent years trying to find an error in his reasoning that would allow him to refute his own hypothesis. Einstein had been less worried, because he saw that Planck had not followed Boltzmann’s prescription properly, meaning that he was not able to see the truth of it. Einstein corrected the error in 1906 and showed that the result was indeed right and consistent with Boltzmann. Max Born wrote about Planck:

> “He was by nature and by the tradition of his family conservative, averse to revolutionary novelties and skeptical towards speculations. But his belief in the imperative power of logical thinking based on facts was so strong that he did not hesitate to express a claim contradicting to all tradition, because he had convinced himself that no other resort was possible.”

Later Planck himself wrote:

> “My futile attempts to fit the elementary quantum of action somehow into the classical theory continued for a number of years and they cost me a great deal of effort. Many of my colleagues saw in this something bordering on a tragedy. But I feel differently about it, for the thorough enlightenment I thus received was all the more valuable.”

In 1911, in epilogue of this discovery, the Belgian industrialist Ernest Solvay invited the main thinkers of the day to the first of a series of conferences on fundamental physics, which have later become enshrined in the history of quantum theory. At the first of these conferences, Planck had finally committed to the idea of quantization. Of all the physicists attending, only Planck, Einstein, and three others took the quantum hypothesis seriously.

The significance of Einstein's work did not stop there, however. It went on to be confirmed further by the developments emerging from the understanding of atomic structure. From experiments, it was known that electrons surround the nucleus of atoms somehow, but classically this was hard to understand. The only stable option for classical mechanics was to assume that electrons were in orbit about the nucleus, like the familiar planetary motion of Newton. Classical electromagnetism said that an orbiting point-like electron would have to radiate energy continuously as electromagnetic waves, and this would cause the electron to be instantly drained of energy and spiral down into the nucleus itself, causing all of matter to collapse instantly. Since this did not happen, something unknown had to be preventing it.

The answer was almost the reverse of the process imagined for the photoelectric effect. Surely enough, electrons would surround the nucleus, but not in any visual way we understand. They were not accelerating in an orbit, so they would not radiate continuously as predicted by the classical model. Instead, they would take on only very specific discrete traffic-light states with corresponding discrete energies, and these could be calculated from quantum mechanics. Radiation could only be emitted as a discrete unit of energy called a 'photon' or quantum of light, with a certain probability, making atoms stable to a very good approximation.

Planck and Einstein's work show how we owe our entire existence to the discreteness of the basic building blocks of our universe. The stability of the electron energy states, and our entire atomic infrastructure are propped up by the discreteness of energy states in matter. This shift in understanding had a profound effect on modern thinking, as the discreteness of our material world invaded scientific culture repeatedly in the years that followed.

While physicists were unravelling the connection between discrete state information and physical law during the first half of the 20th century, mathematicians were developing their own ideas, spurred on by the development of the telegraph, radio, and digital signalling. The need for creating and deciphering secret messages (cryptography) sent over radio and telegraph during the world wars, no doubt accelerated the investment of time and money that went into researching information, and ultimately led to the invention of the digital computer.

Communication was something that everyman could understand the need for. The telegraph had already taken the first steps towards globalization of trade and economy, and its development was stimulating commerce and leading to all manner of new inventions. The telegraph system was one of the technological wonders of the British Empire. It transformed communications in matters of both war and commerce[56], but no one could have imagined that so practical a concern as the signalling of intent across geographical distances would bring about such a revolution. Indeed, it led directly to the very idea that would unify our view of the most fundamental physical processes with the way we relieve ourselves of daily gossip, and then go on to propel humanity into an age of information technology. That deep and subtle unification is still being unravelled today.

The study of digital communication began, in its own right, during the 1930s and 1940s, with mathematicians such as Alonzo Church (1903-1995), Alan Turing (1912-1954)), John Von Neumann (1903-1957) and Claude Shannon (1916-2001) in the lead.

The effect of the quantum mechanics on these thinkers was likely profound. Turing and Von Neumann, two of the most well-known figures in the development of modern computation, were certainly both well versed in the modern theoretical physics of the preceding years[57], and could hardly have avoided being influenced by the deep role played by information in these theories. From statistical mechanics of Boltzmann and Einstein to the quantum mechanical formulations of Schrödinger and Dirac, discrete units of what could only be described as information (about *states* of matter) played a central role in reinterpreting our view of reality. Turing, in particular, was captivated by the meaning of reality itself and many of his arguments about computers and artificial intelligence would reflect his grasp of these ideas.

Representing information as symbolic states, and exchanging state information between participants, is the simplest of ideas, but it proves also to be the basic unifying idea to understand information systems. Claude Shannon referred to this as a communications channel, and studied it in depth. The details of his work were published under the title *The Mathematical Theory Of Communication* in 1949. Today it is also called Information Theory.

Information theory begins with the *representation*, *interpretation* and *transmission* of patterns of *data*, i.e. patterns made up of different kinds of 'things'. We attach meanings to these patterns and call the result information. Patterns of data are mainly of interest when they are transmitted from a *source* to a *receiver*, for instance:

- Morse code sent by telegraph or by lantern.
- Speech transmitted acoustically or by telephony.
- Text read from a page and transmitted to our brain.

- Data copied from hard-disk to memory.
- Data copied from memory to screen.
- DNA copied using RNA within a cell.

In each of these cases, a pattern represented in some medium, like paper, a brain, or computer memory, is transferred from one representation to another. Information is fundamentally about pattern representation.

Shannon began by defining a set of symbols, traditionally called an alphabet in the literature of coding. This might be something like the Latin alphabet A-Z: the basis of most Western writing systems. An alphabet is really just a set of reusable 'atoms' for making patterns.

Suppose we have the simple alphabet: { A,T,G,C }, recognizable as the short symbols used for nucleotides Adenine, Cytosine, Guanine, and Thymine, that make up genes. We can form any number of different messages based on these. In biology, clusters of three such nucleotides represent codons, e.g. "GUG" or "UUG", which represent different amino acids. In the context of the cellular machinery of DNA and RNA, these symbols, written in molecular form, can be copied, interpreted and treated as symbolic operators that change the states of an elaborate chemical feedback system.

If the symbols were received by a human, written on paper, we might imagine that they represented acronyms, or words in some language or other. We could equally make the same patterns out of images of fruit, as in one-armed bandit gambling machines. These sequences of symbols are just patterns that we can recognize and attach interpretation to. The symbols are often called digits, for, although the symbols represented in a channel encoding do not have to be numerical, it turns out to be sufficient to represent everything as a countable set of things. Indeed, the simplest possible representation is binary digits, consisting of only 1s and 0s.

We can convert any alphabet into any other by numbering the symbols from 1 to the maximum number. For example, in the well-known ASCII encoding of the Latin alphabet used in most computers, the association between numbers was as follows:

Sym	→	Dec	→	Bin
A	→	65	→	1000001
B	→	66	→	1000010
C	→	67	→	1000011
...				

Now I have sneaked in three different set of coding here: i) 'Sym', 'Dec', 'Bin' for symbolic, decimal and binary counting systems, ii) 65, 66, 67 for the ASCII values represented in decimal, and iii) the alphabet 1, 0, representing binary numbers. We are so used to handling codes today that you probably hardly noticed all of these symbolic transformations. However, technologists cannot take these things for granted when making machinery to represent and transmit information.

When actually realizing communications technology, it is much easier to represent only 1 and 0, rather than having to distinguish many distinct symbols. There is no fundamental reason for this choice, however—after all, DNA gets along just fine with four.

A reasonable question to ask is: what is the difference between information *about* something, and the thing itself? The answer is: not much. There is a direct correspondence between information about something, and the physical thing itself. In fact, we could say that the thing itself is just one possible representation of the information, but there can also be others.

For instance, we can think of a factory that makes tins of soup. We can model this using the principles of engineering and design and build the actual factory. The information is now represented both in someone's brain, and in building materials and people. Alternatively, we can represent tins of soup by symbols and numbers and model the whole process on paper, or in a computer or scraped in the mud, or indeed any other physical medium. We would call that an information system, but it is not fundamentally different from the real thing, except in what we consider to be the original. We could, after all, make a scale model where the symbols were three dimensional, but, if we are going to do that, then why not use the factory itself to track the information? In that case, how is the information different from the thing? It is just a representation.

We see that *representation* of information is a key issue. Information about something is not fundamentally different from a real thing itself, because both need a physical representation. It is only a question of how we interpret the things. Warren Weaver a contemporary of Shannon summarized three levels of information in 1949:

1. How accurately symbols can be transmitted (copying fidelity)

2. How precisely symbols convey the intended information (semantics)

3. How well understood messages are by humans.

Shannon's work pointed out that patterns of information are not always perfect, but can become distorted over time or during transmission over distance (space). His model of communication builds on the model of the so-called noisy channel (see figure 3.1). The model comes right out of the telegraph age, but it is surprisingly adaptable to almost any other situation, as it has the basic structure of all communication.

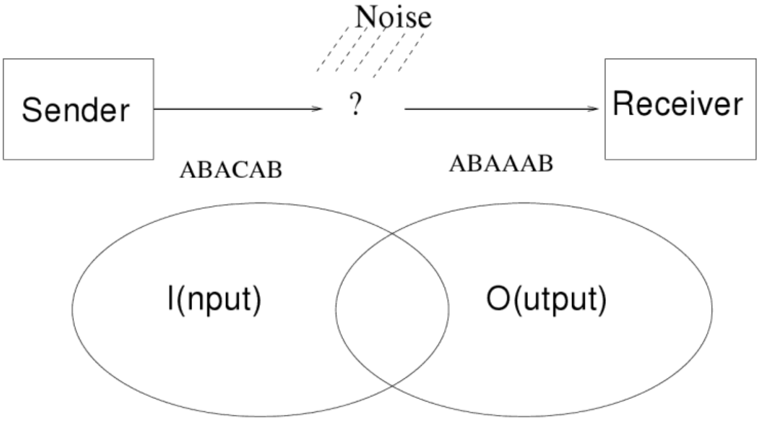

Fig. 3.1. Shannon's model of symbol transmission from sender to receiver, or source to sink. Transmission of the input message overlapped with the output message, or *mutual information* was exchanged.

The idea of the model is that some kind of generic *sender* creates a message as a stream of symbols, which are then propagated to the *receiver* at some rate. The receiver interprets the patterns and, assuming that it understands the alphabet, interprets the same message from the same alphabet. The sender might be a person, a sensor, or your fingertip when touching something. The receiver might be a microphone or a harddisk, or your brain when registering the signal.

Many signals in the physical world are really multi-leg journeys: elaborate chains of signal relays, passing messages through a variety of media. Visual signals are passed as light, then eye to brain by electro-chemical transfer. From sensor to computer is electrical, from computer to harddisk involves magnetic flux. At each stage, some kind of *transducer* is needed to convert the symbols from one representation into another. Every possible interaction between things involves the transmission of some kind of signal from one place and time to another place and time.

Even with a transducer that can transduce one medium for another, it does not guarantee communication of intent. Pouring DNA proteins into the disk tray of your laptop, for instance, would not result in a baby being born from your new 3D printer. That translation software does not yet exist at the time of writing, though it is not impossible to imagine something like it. There is thus, also a need for alphabet transformation, also known as *decryption* or *decipherment* of the stream. Ciphers or encryption methods represent a huge subject of great importance to the idea of certainty[58]. We'll revisit them in part II of the book.

A crucial principle embodied in Shannon's model is that a receiver interprets information completely independently of how the sender intends it. The sender cannot force the receiver to mirror its own intention. The receiver is not even guaranteed to receive the message. There is no determinism involved, no intrinsic, inevitable meaning. The receiver might not recognize the symbols correctly, or the symbols could have become distorted by noise in the channel, flipping bits or distorting one symbol into another. Think of message symbols as change operations, acting on the states of the receiver, changing them. Ideally, this communication would converge to the desired outcome set by the sender, but that simple deterministic idea might be spoiled by the presence of noise from the surrounding environment.

What do these messages have to do with our story of stability? The answer lies in a rephrasing of the question: what might decay look like in a discrete system? What corresponds to fading paintwork, or weeds in the flowerbed of a digital message? How could we *maintain* state over time in a discrete system? How can we make a state stable to its environment?

Shannon's answer was that noise, in a communications channel, led to symbol errors that needed correcting. The intrusion of a rogue message, during the propagation of an intended message, is what led to decay, and the subsequent need for maintenance. This, for example, is how genetic mutation takes place: copying errors in DNA cause the intended genetic message to be altered so that the interpretation of the message at the receiver leads to incorrect response, and this leads to a mutation when the genes are activated. Clearly noise is a threat to the stability of digital information.

Think of a state that is unchanging as time passes. It is just a repeated message that reiterates the same state again and again. Each time the symbol repeats is like a tick of the clock. Suppose we start with an A, then a stable unchanging state would look like this:

AAAAAAAAAAAAAAAAAAAAA ... etc.

The passage of time is marked by a new symbol in the message. A limit cycle might look like this:

ABCABCABCABCABCABCABC ... etc.

The system cycles through a fixed pattern. A noisy system with decay would look like this:

AAAAAAABBBBBBBBBBBBBBBBCCCDDDEEEEE ...

Each time noise flips one of the symbols, it changes to a new state and remains there until more noise changes it again. Had we stabilized the state with an immunity operation, then we could correct errors in the transmission of the message. So then we might have

AAAAAAABAAAAAAAAAAAAAA ... etc.

The sudden occurrence of a B is now immediately corrected back to its desired state. This is sometimes possible, as long as we have a description of what the intended state actually is. This is the case for an immune system, and it is the case for the software CFEngine mentioned earlier. We often call that description a design or a *policy*. In DNA replication, most mutations lead to cells that 'crash' and die. Only if the mutation leads to a message that makes sense does the mutation survive and continue to propagate.

Now consider: what if we could not only allow a state to change, but actually control the way it changed. Then we could transform one message into another. That would be computation! The fact that puts us in charge of manipulating the world around us is this ability to identify controlled evolution of states. We call it computation or we call it medicine, or carpentry, depending on what states we are talking about. The suggestion here is a slightly incredible one: that the universe itself is a computer, whose very evolution in time has the character of a carefully programmed message.

This begins to sound like mysticism, and we should be very cautious in swallowing such a simplistic account. In fact, the universe is much more complex, much more interesting than this. Why? Because the quantum states that Planck discovered are not governed by just one message, but by a vast number of possible messages that we cannot predict. To close the loop on this story, and make this connection, we have to see how the natural world actually behaves like a highly complex computer—and how it doesn't.

You probably know the terms 'analogue' and 'digital'. Modern culture refers to information as being one or the other of these, but few of us now remember the origins of these terms.

Analogue tends to bring up vague connotations of old-fashionedness, or of being that good-old wholesome organic stuff of a pre-digital age, like vinyl records and valve amplifiers[59]. Digital, on the other hand, tends to represent cold and soulless machinery, but also high quality. But, all prejudice aside, the terms actually have very specific meanings that are quite unrelated to the images they now conjure.

Before the invention of modern computers[60], there were so-called analogue computers. These did not calculate according to the rules of arithmetic, but by analogy. The rules of modern arithmetic as we calculate today, descend from the mathematics of Chinese, Indian, and Arabic scholars. We calculate with digits (from the Latin *digitus* for finger or toe), arranged in patterns of 10 or 20 in most cultures, because that is the usual number of fingers and toes. Our methods are a variation on the Chinese abacus approach, and are also based on the method of counting.

Digital computers simulate the arithmetic approach that we humans use to count with, using symbolic manipulation. Analogue computers, on the other hand, do not use counting to compute, but work by *analogy* to certain systems of equations, using the principle of dynamical equivalence discussed in the previous chapter.

How could this work? Suppose we could construct a physical system that mimicked the behaviour of something we wanted to calculate? It might then be possible to construct a device that, by virtue of its very existence, would be a simulation of an equation, and could therefore calculate the answer to a problem. All we would have to do would be to sit back and watch the answer unfold. The laws of physics would do the work, and we could simply read off the answer at some appropriate moment. This is the idea of the analogue computer.

All measurable physical phenomena behave essentially like calculating systems, but some are more convenient to interact with than others. In electrical circuitry, there are components like resistors, capacitors, and inductors that behave like processes in a calculation. In practice, you need to be able to add and subtract, and from this anything is possible. However, electronic circuits can also perform calculus operations like integrate and differentiate. Capacitors accumulate charge and act like adders or integrators. The accumulated charge in a capacitor is proportional to the applied voltage $Q = CV$, and the charge is the integral of a current.

$$V = \frac{1}{C}\int_0^t I dt$$

Voltage and current are easily measurable, so with a multimeter, one could imagine solving integrals. Inductors, such as one finds in radios, behave like differentiators, because electric currents respond to rates of change of magnetic flux, which in turn depend on the current.

$$V = L\frac{dI}{dt}$$

By applying principles like this, and other components like transistors or operational amplifiers, clever minds were able to construct electrical circuits which would compute answers to specific problems.

Unlike digital computers, there was never a generic, multi-purpose analogue computer that ran a version of an office spreadsheet and other programs. Analogue computers had to be specially built for one specific problem at a time. At best one could hope for a 'breadboard' of components with patch-bay of connections to make the assembly easier. The user would have to build the right circuit and then set a voltage on the input contacts, and measure the outgoing voltage or current from the circuit, which would translate directly into the answer. The answer was literally computed by the analogy between the laws of electricity and mathematical operations.

Analogue computers were never a good solution because physical systems are prone to noise and non-linear effects, causing instabilities that lead to incorrect results. Their results were not discrete values, but were interpreted as continuous voltages, thus today we use the term 'analogue' to mean the opposite of 'digital'. The science has become culture.

History has produced many examples of analogue computers, using different technologies through the ages, but a particularly interesting form of analogue computation that has been explored in recent years is biological computation using DNA, because it is both an analogy engine and a digital system. This is an interesting kind of digital-analogy hybrid computer which uses DNA in bacterial colonies to perform computational operations.

Every cell has a program coded implicitly in its DNA, which receives chemical inputs from around the cell wall, and expresses proteins the same way. Cells communicate using these chemical channels in our bodies. Thus even our bodies are computational systems. The program for carrying out cellular logic is written in DNA. Unlike a digital computer, it does not have a regular clock, so time is not predictable in the same sense as it is in a digital computer. Rather, it is quantized in 'events'. i.e. the arrival of a messenger protein. Its power lies in massive parallelism. Cells are easy to copy, thanks to the mechanisms of DNA itself. Genes may be switched on and off by messenger proteins which bind to a particular site, which becomes the input. The result is a the manufacture of another protein. These proteins form a vast high-level code, written in an alphabet of amino acids. This process happens in our bodies all the time, and in that sense we can claim to be emergent effects of a physical simulation executing a program written in DNA. Progress has been made in using this process for artificial computations too.

Today, there does not seem to be a future for human-built analogue computers in the world of human technology. Digital computers offer a much more predictable approach to computation, thanks to the concept of digital error correction; they are easier to build upon to make reliable information driven systems for our human infrastructure. Analogue computers were in use up until the end of the 1980s in a variety of applications, like car and aircraft control systems, but today they are entirely replaced by more reliable digital computers.

There are two things to be taken from this. The first is that the physical world we live in is, in fact, a computer—an analogue computer. The laws of physics are literally computing our future at every moment, all over the universe. We just don't know what problem it is computing[61]. The second, and more important point, is that—no matter what approach we use to try to compute the answer to a question—every answer we come up with is an approximation in one way or another. Either the computation is performed using an indirect method, or it is carried out with finite accuracy due to the discrete nature of information in the world.

Finite resolution also means finite accuracy, and we never even get close to the fundamental limits of accuracy due to the quantum nature of matter with today's technology. However, we have chosen digital computers as the basis of modern information technology, and information technology as the extrinsic control system for almost everything else around us. Even when the applications seem to be continuous, it is hard to avoid the effects of the finiteness of the world. This is but one possible representation of information, but it is a natural one.

From the analogy between physics and computation, we have a picture of a computer as a general device which takes an input stream, and evaluates an output stream according to a set of rules. This describes pretty much any linear, physical system.

In the previous chapter, we looked at the linear rules for driving the evolution of a physical system. They had the form of

$$\text{(Act on) thing} \longrightarrow \text{new thing}$$

It is also the transmission of a message:

$$\text{(Copy) local state} \longrightarrow \text{remote state}$$

And, in physics, this kind of evolution crops up all the time. It was not the usual way of formulating physical problems, at least before quantum mechanics, but it is a viable representation of physics in linear cases. It represents an information viewpoint of system evolution in terms of discrete transitions, and it can usually be derived from any system of variables and constraints that represent a model of reality.

In physics, if we try to cast the evolution of the system in this form it takes the form of a 'linear response' function, also called a Green's function[62] of the previous chapter. In physics it often looks something like this[63]:

$$q(t) = \int (dt')\; G_r(t,t')\, F(t')$$

but the rather beautiful calligraphic notation is just window-dressing. This is just special notation for:

(Evolve and propagate) 'edge state F' → 'new state $q(t)$'

or

(Evolve and propagate) 'sender state' → 'receiver state'

Propagate means effectively 'follow over time'. In this form, physics recovers the operational view of states from the previous chapter, and we also see that it is just a generating loop for writing a message by repeating this propagation over and over again, marking out time. The response to the message at the source is information about the new state:

Data $G_r(t,t')$ from 'edge state F' → 'new state $q(t)$'

or, even more simply:

(Develop) Cause → Effect.

This is basically a computation:

(Compute) Input → Output.

This generalized system thus has the behaviour of a rather specific computer[64]. It evaluates only one very specific problem, i.e. how to obey this law of physics, in the form of a message of states that evolve over time. The only thing that distinguishes it from a general computer is that it is stuck on one program: integrating a simple differential equation which is physics itself. This is the basic template from which we humans are able to turn knowledge of this science into a technology to make computers of our own.

Think simply of states that can take whole number values, and imagine creating transformation operators that can perform addition and subtraction. This is possible, and indeed this is how modern computers are constructed using electronic circuits. If we let (Inc) mean increment (add one), and (Dec) mean decrement (subtract one) from a state, then here is a simple operator behaviour for elementary digital computation:

(Inc) 4 ⟶ 5

(Dec) 4 ⟶ 3

(Dec)(Dec) 4 ⟶ 2

(Zero) 4 ⟶ 0

(Zero) 0 ⟶ 0

(Zero)(Zero) 4 ⟶ 0

These are two of the basic operations in a modern digital computer. If we iterate a constant message of (Inc), the state will count upwards.

What jumps out from the simple formulations above is not the technical details, which are tragically simplified here, though certainly have their own beauty, but rather the underlying information structure of the relations themselves. It suggests that the world is in fact a kind if information pump, evolving states (which are measures of information) by causing transitions from one state to another. Just as the absorption of a photon in a laser would cause a change of state from one level to another, so the absorption of a message (another measure of information) from a sender would cause a change of knowledge state in the receiver. This is computation not by analogy, but by 'digilogy'[65]!

Now we might close the loop on this story of the discreteness of things and what they mean to us. Because a message is a collection of symbols, we can represent the message as a journey through a state chart, like the traffic light example in Figure 1.1, we can represent any message as a journey through a network of the different states. Such a network is called a state machine. If there is a finite number of the states, then it is called a *finite state machine*.

We can then separate the symbols into two types: operators and states.

The bold letters represent the different states of the machine: empty, stopped, playing, spooling, etc. The arrows in between show the *operations* that initiate a transition from one state to another. Thus, an empty player moves from state 'empty' to 'stopped' by inserting a tape, disk or other source medium. In the operator language, this would be written like this:

(insert) 'empty' → 'stopped'

(eject) 'stopped' → 'empty'

(play) 'stopped' → 'playing'

(stop) 'playing' → 'stopped'

(fwd) 'stopped' → 'spooling fwd'

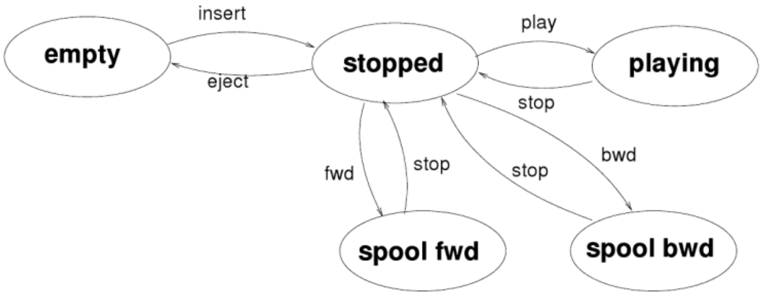

Fig. 3.2. A finite state machine model of a video player. The bold symbols are states and the symbols on the links are operators.

(stop) ‘spooling fwd’ → ‘stopped’
(bwd) ‘stopped’ → ‘spooling bwd’
(stop) ‘spooling bwd’ → ‘stopped’

Thus, when we act on the ‘empty’ state with an ‘insert’ operation, the state of the machine moved from being ‘empty’ to ‘stopped’. Then when we press play (act on the stopped state with the ‘play’ operation), the state moves from ‘stopped’ to ‘playing’.

These are called state transitions, and we can write them as a *transition matrix*, which is simply a table of all possibilities mapping all the states to all the others. If there is an arrow from one state to another, we write a 1; if there is no arrow, we write a 0. We read the state machine’s graph by reading from a sender state on the left to a receiver state across the top.

(row, column)	empty	stopped	playing	spoolfwd	spoolbwd
empty	0	insert	0	0	0
stopped	eject	0	play	fwd	bwd
playing	0	stop	0	0	0
spoolfwd	0	stop	0	0	0
spoolbwd	0	stop	0	0	0

The rows and columns of the this matrix (rows go along the floor, and columns hold up the ceiling) show the transitions in and out of the states. Or, we can strip this down to bare bones and write only the transitions between in coming (row) states and outgoing (column) states:

$$T(\text{in}, \text{out}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.3 & 0 & 0.5 & 0.1 & 0.1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

where 1 means ‘allowed transition’ and 0 means none. This state machine is a non-deterministic state machine. There is only a certain probability of a particular transition being made.

The numbers in the table are made up, for the sake of example. One could understand them as follows. Once the tape is in the player, the probability of pressing play is slightly higher than pressing eject, but we know that we have to eject something at least once. It is not usual to feed tapes and disks into players just to eject them without playing them, but we might stop play during a movie several times before ejecting it. There is nothing programmed about these probabilities, nothing inevitable or deterministic. They represent only behavioural experience.

The signature of message passing and information propagation is evident in both the technology of communication and in many examples of physical law described above[66]. The question remains, however, how closely are these related? Are they merely inspired by one another, or is there something deeper at work? Certainly there are parallels, but there are also differences. As in the case of the ship on the ocean, a formulation of a quantum system is sensitive to the way in which we separate the big picture from the small. The alphabet of states (as we understand it) in the quantum theory is not a constant fixture for instance: it depends on a complex interplay of boundary conditions. However, before leaving this chapter, I want to return to the quantum story to reconnect the loose ends, with the benefit Shannon’s perspective. This takes us back to the years following Planck and Einstein’s discovery.

We left the pioneers of quantum theory at the cusp of the growing acceptance of a quantum view of the world, but little was yet understood about the structure of atoms and their interactions with light. The black-body radiation signature of Planck had been a first, semi-classical step in the analysis of the interaction between matter and radiation, where the quantum nature of energy for the first time could not be avoided.

As the idea that matter and energy were discrete sank in, a variety of models appeared to describe the behaviour of electrons confined in potential wells, or scattering through slits. Calculations of these, based on generalizations of the accounting principles for energy that Newton and Leibniz had taught us, yielded a variety of promising results to explain experimental observations, but the real difficulty lay in understanding what the theory *meant*.

The remark by Niels Bohr at the start of this chapter summarizes the frustration physicists felt with possessing an effective description of the phenomena, but having no effective understanding of why it worked. The equations of Schrödinger and Heisenberg were akin to having a slot machine that produced fortune cookies predicting experimenters' likely success in measuring results. It predicted accurately the alphabets of states that *could* be measured, but not what states *would* be measured. Newton's clockwork determinism was gone; cause and effect were no longer simply related.

During the 1930s and 1940s, developing the theoretical understanding of how quantum theory and Einstein's Special Theory of Relativity could be reconciled became a priority, as several measurable effects were related to this. Explaining the rapidly improving experiments became the chief goal of physicists. It was an effort that took many years to fully gestate, interrupted by the second world war. It began with British physicist Paul Dirac (1902-1984) and ended with the work of two Americans, and a Japanese physicist, each working independently using what seemed to be quite different approaches.

Quantum Electrodynamics (or QED for short) was pieced together by three physicists: Richard Feynman (1918-1988), Julian Schwinger (1918-1994), and Sin Itiro Tomonaga (1906-1979) between 1948 and 1949. However, their formulations of the subject seemed so radically different from one another that it took the efforts of a fourth, Freeman Dyson, to see the connection between them. All three, without Dyson, later won the Nobel prize for their efforts in 1965. The approaches used in QED to describe quantum processes are interesting because they reveal a deeper connection between physics of the small and the theory of digital communication due to Shannon.

Feynman's approach to the problem has attained the greatest popular appeal, mainly due to his greater personal flamboyance, and legendary status as a teacher. In Feynman's version of QED, he used diagrams to visualize what he believed were quantum processes, now referred to as Feynman diagrams (see Figure 3.3). Initially Feynman saw his diagrams as representing actual particle trajectories, though this became more a formality as more viewpoints emerged (Schwinger, for instance, was reported as being less than impressed), but they helped him to get past the philosophical issues with the interpretation of quantum mechanics that Bohr referred to, by not having to mystify the existence of the particles. The diagrams were not arbitrary, but came out of Feynman's own personal reformulation of quantum theory based on a principle of least action[67]. Feynman was intrigued by the experiments that showed electrons interfering, which suggested that they followed multiple paths at the same time, somewhat like a wave. This quantum interference is one of the strange and mysterious aspects of quantum theory that defies satisfactory explanation, even to this day. In his formulation, he set up a so-called 'path integral' or sum over paths that electrons might take, as if not just taking one fortune cookie but the whole machine at once.

Feynman thus did not fully give up the idea of the electron as a point object, but instead allowed it to behave in apparently unphysical ways—more like a collection of messages than a point particle. Dyson later wrote:

> Dick Feynman told me about his "sum over histories" version of quantum mechanics. "The electron does anything it likes," he said. "It just goes in any direction at any speed, forward or backward in time, however it likes, and then you add up the amplitudes and it gives you the wave-function." I said to him, "You're crazy." But he wasn't.

Whether one believes the picture he used to assemble the results, Feynman's 'sum over histories' approach turned out to give good answers. Each of his diagrams represented different kinds of processes in the sum over all possibilities—and each had the form of a message sent from a sender to a receiver.

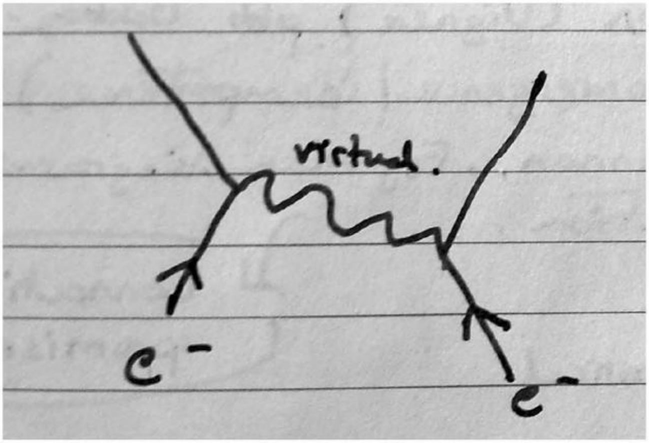

Fig. 3.3. Feynman diagram. The solid lines represent electrons, and the wavy line a photon. Electrons repel electrically by a message passed via a photon, the carrier of electromagnetic force.

Feynman was adept at intuitively writing down the processes that could go into the sums. He drew diagrams like that in figure 3.3 that seem to represent particle processes. However, in fact, each diagram did not represent a single particle process but was merely a template for a sum over all possible variations of that configuration. The lines that seemed to indicate particle trajectories to Feynman, perhaps only needed to represent the structure of causal influence, i.e. a message of information propagated from one interaction site to another, unwittingly resembling Shannon's model of communication. Being only a template message, a Feynman diagram was not like a single message, but more like a collection of similar messages, each representing a different possible version of the message machine. One couldn't know exactly what symbol is going to arrive at a particular moment, but you could add up the expected frequencies to see what was the probable outcome.

Using Tomonaga's work as a bridge, Dyson was then able to show how this picture related to an explicit and very elegant construction by Schwinger[68] that ignored the particle concept altogether. In Schwinger's formulation, the idea was to emphasize not particles, but elements of spacetime itself as the dynamical quantities. It was a quantum *field theory*, not just quantum mechanics of particles[69]. The sum over possibilities in Schwinger's work was by appeal to generating functionals, and his original formulation looked daunting and opaque to physical insight. But even though Schwinger's theory was superior in scope to Feynman's in 1948, he was aware of its deficiencies, and began to strip it down and reformulate it over the years that followed. During this time Shannon's influence became clearer. The notion of particles was quite possibly just a particular representation for a transmission of information.

One of the most prized possessions in my home is an original hardbound 1949 edition of Shannon and Weaver's book, *The Mathematical Theory of Communication*, given to me by Clarice Schwinger, the wife of Nobel Laureate Julian Schwinger in 1997, just after his unfortunate and sudden death by cancer in 1994. I became friends with the Schwingers a year before Julian's death by a lucky chance encounter[70]. While visiting their home in Bel Air, with its beautiful view overlooking Los Angeles, I perused the bookshelves during the early morning. It was the first time I had seen Shannon's book, and I was intrigued. In it, I saw a clear precursor to Schwinger's own reformulation of Quantum Theory, which he called *source theory*. Although Schwinger never acknowledged any link between the two, to my knowledge, the model of source and sink, sender and receiver, came just at the right moment and fed directly into Schwinger's formalism.

In source theory, Schwinger tried to strip away the formal details that he'd erected to prove mathematical consistency, and he reduced interactions to simple messages written in the alphabet particle processes, between a point of creation and a point of detection: a source and a sink, or a sender and a receiver. Without the actual diagrams that Feynman drew, there is an essential similarity in the formulation of a process of information transfer. In both their approaches, the word particle is empty of any real meaning. Schwinger wrote in 1970

> "Theory thus affirms, and experiment abundantly confirms, that the concept of a particle as an immutable object is untenable ..."[71].

What Schwinger's formulation, along with Feynman's alternative view, shows is the communication of information is the underlying mechanism for transmission of influence, even in the physics of the very small. The Newtonian idea of particles was not compatible with reality on a quantum level. It was an illusion of scale.

During the 20th century, our investigations into the fundamental nature of our world have brought about a major shift in our understanding of the nature of reality. They have influenced the way we view natural processes around us, and also led to technological innovations inspired by those processes.

The discreteness of matter and energy is a gift of understanding, that links the physical world to the virtual world of information in unexpected ways. Rather than closing a door on a dream of infinite resolution, it opens our minds to new ways of understanding the stability on which our world is built. Those insights have been crucial to the design of real and imagined technologies in the post war period. Atoms are not deterministic planetary systems, with absolute containment, they are something we do not fully understand, kept stable by strong but not absolute constraints.

Advances in the miniaturization and increases in transmission speeds of data now force us to confront the quantum nature of physics that previously had little impact on human time scales. This means that statistics and probability must play a role in our models of the way the world works. Discreteness, representing units of information, whether on the smallest scales or the largest, is an essential trait of both physical reality and of the artificial systems we build. We have to understand it, model it, and tolerate the way it interacts with important scales in order to exercise a sense of technological control over the world. As scientists, it allows us to know nature's secrets; as technologists, it offers us the stability on which to build for survival.

Whether by analogy or by 'digilogy', we employ the phenomena of stability to creative purpose.

4

All the Roads to Nowhere

How keeping things in balance is the essence of control.

"Plus ça change, plus c'est la même chose."
(The more everything changes, the more it all stays the same)
– Jean-Baptiste Alphonse Karr

Don't move! Stay where you are! Don't leave town! Don't go anywhere!

Let's hope these are not phrases you have had occasion to hear very often, at least outside of fiction. They are arresting exclamations, challenging someone to stay put, i.e. maintain a stable location, with varying degrees of accuracy—which is a turgid way of saying that they describe a status quo. One could also say that they are different ways of expressing an absence of change, at different scales: they represent different approximations to staying put. The sequence starts with millimetre movements of your muscles, then relaxes to your immediate surroundings, falls back to a geographic region, and finally gives up altogether being specific about location.

We understand these vague concepts intuitively. Our brains' semantic analyzers regularly decode such patterns of words, and attach meaning to them in the context of a scenario. That is the power of the human mind. If someone says, "I didn't move for 20 years", this does not mean that they were frozen in liquid nitrogen for two decades. It probably means that they settled in the same home, in the same town, i.e. that their average position remained within some general threshold radius, even allowing for one or two excursions to holiday destinations.

We are so used to using such concepts that we don't think much about what they actually mean, and there is good reason for that. Roughness and approximation play an important role in keeping logic and reasoning manageable and cheap to process. Even if someone moves around quite a bit, we can feel happy that they are basically in the same place. A hierarchy of approximation is a useful tool for both understanding and making use of scale as we build things. As we'll see in part II of the book, we handle approximation effortlessly when we model the world, in order to avoid overloading our brains with pointless detail. In the context of stability, however, this semantic interpretation probably emerges from a key survival need: to comprehend shifts in average behaviour.

This chapter is about what change means to stability, on a *macroscopic* scale—which is the scale we humans experience. It will introduce new concepts of dynamical stability and statistical stability, which form the principles on which a public infrastructure can prosper. It takes us from the realm of singular, atomic things and asks how we build up from these atoms to perceive the broader material of stuff. In the last chapter, we examined the world from the bottom up and discovered patterns of isolated microscopic behaviour; here, those patterns of behaviour will combine into a *less* detailed picture of a world that we can actually comprehend. This is the key principle behind most of modern technology.

The four exclamations in the opening of this chapter may all be viewed as expressions about required stability. If we are willing to overlook some inexactness, then we can characterize the location of a person as being sufficiently similar over some interval of time to be able to ignore a little local variation for all intents and purposes. This then begs the question we've met earlier: at exactly what threshold is something sufficiently similar for these intents and purposes.

By now, you will surely get the picture—of course, it's all about scale again, and there are no doubt some dimensionless ratios that govern the determination of that essential threshold. We understand the difference between 'don't move' and 'stay in town' because it is easy to separate the scales of an individual human being from a town. The ratio of length of my arm to the radius of the city gives a pretty clean measure of the certainty with which I can be located within the city. Had we all been giant blobs from outer space, this might not have been such an obvious distinction. I, at least, am not a giant blob (from outer space), so I can easily tell the difference between the boundaries of my city and the end of my fingertips. Similarly, if my average journey to and from work each day is much less than the radius of the city, then it is fair to say that my movements do not challenge the concept of the city as a coarse grain to which I belong.

Separability of scale is again an important issue. However, there is another aspect of stability present in this discussion that has fallen through the cracks so far. It is the idea of dynamical activity, i.e. that there can be continual movement, e.g. within the cells of our bodies that doesn't matter at all. This simple observation leads us to step away from the idea of exactness and consider the idea of *average* patterns of behaviour[72].

Suppose we return to the concept of information, from the previous chapter, for a moment, and insert it into the description of behavioural patterns at different scales. A new twist then becomes apparent in the way we interpret scale. As we tune in and out of different scales, by adjusting the zoom on our imaginary microscope, we see different categories of information, some of which are interested to us, and others that are uninteresting. We elevate trends and we gloss over details; we listen for signals and we reject noise. We focus and we avoid distractions.

This focus is an artifact of our human concerns. Nature itself does not care about these distinctions[73], but they are useful to us. This is perhaps the first time in the book that we meet the idea of a *policy* for categorizing what is important. We are explicitly saying that part of the story (the large scale trend) is more interesting than another part (the fluctuation). In technology, we use this ability to separate out effects from one another to identify those that can be used to make tools. For example, I consider (*ad hoc*) the up-and-down movement of keys on my keyboard when I type is significant, but the side to side wobble by a fraction of a millimetre isn't, because one triggers a useful result and the other doesn't. But that is a human prejudice: the universe is not particularly affected by my typing (regardless of what my deflated ego might yearn for). There is nothing natural or intrinsic in the world that separates out these scales. On the other hand, there is a fairly clear separation, from the existing structure of matter, between the movement of the few atoms in a key on my keyboard and the separation of planets and stars in the galaxy.

Sometimes policies about the significance of information are *ad hoc*: for example, a journey may be considered local to the city if it is of no more than three blocks, but four blocks is too far to overlook. This judgement has no immediate basis in the scale of the city. Other times, policies are directly related to the impact of the information on a process we have selected as important to our purposes. For making technology, we often choose effects that *persist* for long enough to be useful, or show greater average stability so that we can rely on their existence.

We can visualize the importance of 'persistence stability' with the help of an analogy. Before digital television, television pictures were transmitted as 'analogue', pseudo-continuous signals by modulating high frequency radio waves. When there was no signal, the radio receiver would pick up all of the stray radiation flying around the universe and show it as a pattern of fuzzy dots, like the background in Figure 4.1. This was accompanied by the rushing sound of

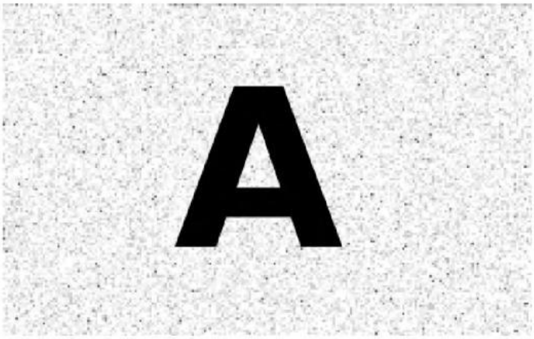

Fig. 4.1. Analogue noise on a television set contains a large amount of information.

the sea on the audio channel (from which the concept of noise comes). The persistent pattern in between the random dots was an example of a stable pattern from which one could perceive information. After digital signals, loss of signal results in nothing at all, or sometimes missing frames or partial frames, which one of the advantages of digital.

Years ago, when I used to teach basic information theory to students at the University, I would switch on the old pre-digital television without the antenna and show the fuzzy dots[74]. Then I would plug in the antenna and show them a simple channel logo BBC or NRK, like the letter 'A' in Figure 4.1. Then I would ask the students which of the two images they thought contained the most information. Of course, everyone said that the letter 'A' contained more information than the fuzzy dots. Then I would tell them: imagine now that you have to write a letter to a friend to describe the exact image you see, so that they could reproduce every details of the picture at their end. Think about the length of the letter you would need to write. Now, tell me, which of the two pictures do you think contains more information?

The point was then clear. Noise is not too little information, it is so much that we don't find any persistent pattern in it. The simple letter 'A' has large areas (large compared to the dots) all of the same colour, so the total amount of information can be reduced, by factoring out all the similar pixels. In a digital encoding of the picture, the letter 'A' is just a single alphabetic character (one symbol), but in an old analogue encoding, we have to send all the variations of light and dark broken down into lines in a raster display, and this is much more sensitive to noise. To write down a description of every fuzzy dot and every change in every split second, would require a very long letter, not to mention a very boring one that might make writing to relatives seem gleeful by comparison.

Despite the presence of noise, we can find a signal, like the letter ‘A’, by choosing to ignore variations in the data smaller than a certain scale threshold. Recall the discussion of thresholds in Chapter 2. Try half-closing your eyes when looking at Figure 4.1, and watch the noise disappear as you impair the ability of your eyes to see detail. This same principle of separating transmission as part signal and part noise can be applied to any kind of system in which there is variation[75]. It doesn’t matter whether the variation is in time (as in Figure 4.2) or whether it is pattern in space, like the letter ‘A’ (as in Figure 4.1).

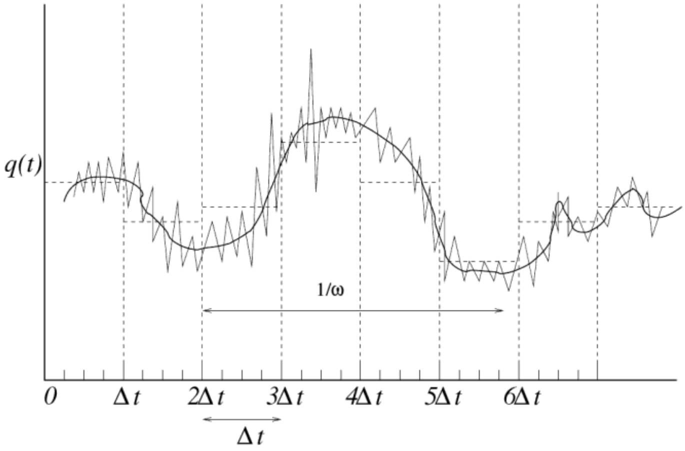

Fig. 4.2. Coarse graining of different representations of a pattern changing in time: the actual detailed signal (jagged line), smoothed into a trend (smooth curve), and then re-digitized into coarse blocks of width Δt.

This idea of separating out noise, is not really different from the way we handle any kind of detail, whether perceiving or building something. Imagine how you would depict a scene for someone: if you visually sketched or otherwise described the object, you would probably begin by drawing the main lines, the broad strokes of the image, and only then start to fill in details at lower levels. You would start from the top down in the scale hierarchy. However, if you tried to draw something like a lawn of grass, with thousands of detailed parts all the same, you would probably just start drawing the details one by one, because there would be no trend to identify. There would be nothing to gain from starting anywhere else, because there is not identifiable trend to extract. The half-closed eye test allows us to see major trends and features in an image.

Trends are part of the way we perceive change, whether it is change in time or in space. It is common to decompose variation into something that is *slowly varying* that we want to see, and something else that is *fluctuating* on top of it that we don't (see Figure 4.2). It is a common technique in mathematics, for instance: it is the basis of Fourier analysis, and it is the basic method of *perturbation theory*. It is a property of weakly coupled (linear) systems that this decomposition of scales leads to a clear description of trends. In fact it is a property of the interference of waves. We can express it in a number of suggestive ways:

Change = Stable Variation + Perturbation

Variable data = Trend + Variation

Transmission = Signal + Noise

In each case, the aim is to make the contributions on the right of the plus sign as small as possible.

Recall, for example, our love boat from chapter 1, sailing on the calm ocean. At any particular zoom level, the picture of this boat had some general structural lines and some details that we would consider superfluous to the description. Viewed from the air, the picture was dominated by a giant storm, and then the ship was but a blip on the ocean. At the level of the giant waves, the ripples close to the ship were irrelevant details. At the level of the ripples, the atomic structure of the water was irrelevant, and so on.

Remember too that a small perturbation can unleash a self-amplifying effect in a non-linear system—the *butterfly effect*, from Chapter 1. If the small variations in Figure 4.2 did not average out to keep the smooth trend, but in fact sent it spinning out of control, then we would be looking at useless instability.

That is not to say that details are always unwelcome in our minds. Occasionally we are willing to invest the brute force to try to be King Canute in the face of an onslaught of detail. Perhaps the most pervasive trend in modern times is that we are including faster and faster processes, i.e. shorter and shorter timescales, in our reckoning of systems. We used to let those timescales wash over us as noise, but now we are trying to engage with them, and control them, because the modern world has put meaning there.

The fighter jet's aerodynamic stability was one such example. Stock market trading prices are another where people actually care about the detailed fluctuations in the data. Traders on the markets make (and lose) millions of dollars in a split second by predicting (or failing to predict) the detailed movements of markets a split second ahead of time. Complex trading software carries out automated trading. This is analogous to the tactics of the fighter jet, to live on the brink of stability, assisted by very fast computers that keep it just about in the air, in a fine detailed balance at all times, modulo weekly variations. In most cases, however, we are not trying to surf on the turbulence of risk, we are looking for a safe and predictable outcome: something we can rely on to be our trusted infrastructure.

We now have a representation of change based on scales in space and time, and of rates of change in space and time. This is going to be important for understanding something as dynamic as information infrastructure. These two aspects (position and rate of change) are known as the canonical variables of a dynamical system. Let's pursue them further.

As we zoom into any picture, we may divide it up into grains[76], as in figure 4.2, and imagine that the trend is approximately constant, at least on average, over each granular interval. The variations within the grain can thus be replaced by the flat line average of the values in the interval, for most purposes. If the grains are small enough (so that the size of $\Delta t \to 0$), the result is that one smooths out the jagged fluctuations, leaving the smooth curve in the figure.

One may also do the same thing with the difference between the actual value and the average value. This gives us the average size of fluctuations, relative to that average base value in the grain. We can separate clarity and fuzz. The method does not work unless there is a separation of scale, however. If fluctuations are no smaller than the variations of a trend, then we can't tell the difference between them, and the line is simply irreducibly jagged at all scales. This is called *scale-free*, *self-similar*, or even *fractal* behaviour. Dividing up a model into grains is also a strategy for computational modelling, but it does not work very well in non-linear systems, which is why it is hard to predict the weather (see chapter 1).

Now back to information. Smoothing out fluctuations into a continuous curve is one approach to finding a stable representation of a process, but it is still varying continuously (or pseudo-continuously). What if we don't make the grain size vanishingly small, but keep discrete grains or 'buckets'? Then the value of the variation stays at the same fixed value for longer. This is a kind of quantization of the horizontal time axis. By removing detail, removing information, we actually find more meaning.

This idea can be used to exploit another approach called re-digitizing the signal. This is a way of sorting a signal into fixed symbol categories on the vertical axis (see Figure 4.3). This is where digital signal transmission comes back to us as a technology rather than as a discovery.

Recall that the telegraph used the clarity of digital Morse code to avoid the noisy transmission lines of the electrical wiring. By standardizing discrete units of time for dots and dashes, i.e. by digitizing the time for a dot, we can distinguish a dot from a dash in terms of a simple digital time scale. By setting a threshold between signal and no signal, a message of dots and the dashes could be distinguished as symbols from no message at all.

This same principle was applied both to the time axis and to the signal value, to digitize many other kinds of signal, starting from around the 1980s. What had previously been treated as continuous was cut up into a discrete template, or *sampled* (see Figure 4.3) so that it could be encoded as digits or symbols. Digital music resulted in the MiniDisc in Japan, and the Compact Disc (CD) all over the world. Video discs, DVDs and Blue-Ray followed. Now digital television and radio are the norm.

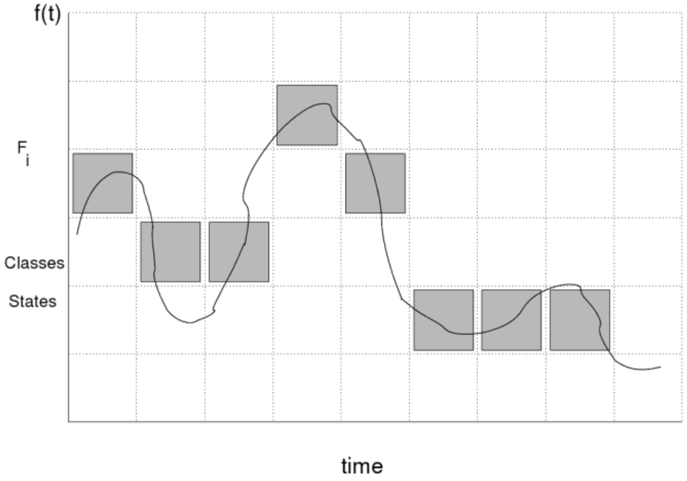

Fig. 4.3. Re-digitization of the vertical axis an analogue signal is called A/D conversion. The rendition is not 'perfect', but we can make the symbol sensitivity as large or small as we like, so that variations within a digit do not matter.

Re-digitization brings stability to our perception of location on these maps of sound and picture: as long as a signal stays within the bounds of certain thresholds we have assigned, the digital representation records a specific symbol. Thus, if we divide up the 'continuous' signal values into coarse enough ranges, then the effect of noise will be almost unnoticeable, because it will be like moving a couple of blocks within the radius of the city. Our ears are not sensitive enough to detect the difference. A perturbation of the average signal will still be within the bounds of a category. This is the method of analogue to digital (A/D) conversion.

Harry Nyquist (1889-1976) made famous a theorem which tells us that, in order to capture the detail of a change in time, we have to sample the signal into buckets at least twice as fast as the fastest expected change. That is why digital music is represented in sampling rates of 44kHz (16 bits) and 48-192kHz (24 bits), as the highest frequencies available to human hearing are reckoned to be at most 20kHz. The higher quality sampling used by audiophiles is designed to capture more of the harmonics and interactions that happen during the reproduction of the music, even if the actual frequencies cannot be perceived directly.

Average behaviour seems to be a key to the way humans perceive change. I'll return to this topic in the second part of the book. We can grasp trends over a limited range, and we ignore the fluctuations on a smaller scale. Thus our interaction and comprehension of a scene depends completely on how we set our scope. These matters are important when designing technology for human interaction, and they tell us about the sensitivity of machinery we build to environmental noise.

So let's delve deeper into the balance of fluctuations, and consider what this new notion of stability means, based on the more sophisticated concept of averages and statistical measures. A change of setting might help.

Global warming and climate change are two topics that rose swiftly into public consciousness a few years ago, to remind us of the more tawdry concept of accounting and balancing the books. It took what had previously been a relatively abstract issue for meteorologists about the feedback in the global weather systems and turned it into a more tangible threat to human society. The alleged symptoms became daily news items, and pictures of floods, storms, and drowning polar bears became closely associated with the idea. Climate change is a new slogan, but it all boils down to yet another example of scaling and stability, but this time of the average, statistical variety.

The weather system of our planet is a motor that is fuelled primarily by sunlight. During the daytime, the Earth is blasted continuously by radiation from the Sun. The pressure of that radiation bends the magnetic field of the planet and even exerts a pressure on the Earth, pushing it slightly outwards into space. Much of that radiation is absorbed by the planet and gets turned into heat. Some of it, however, gets reflected back into space by snow and cloud cover. During the night, i.e. on the dark side of the globe, the heat is radiated back into space, cooling one half of the planet. All of these effects perturb the dynamical system that is the Earth's climate.

As chemical changes occur in the atmosphere, due to pollution and particles in the atmosphere, the rates of absorption and reflection of light and heat change too. This leads to changes in the details that drive the motor, but these changes are typically local variations. They happen on the scales of continents and days.

Usually, there is some sense in which the average weather on the planet is more or less constant, or is at least slowly varying. We call the slow variation of the average patterns the *climate* and we call the local fluctuations the *weather*. By analogy with the other cases above, we could imagine writing:

$$\text{Atmospheric evolution} = \text{Climate} + \text{Weather}$$

and, even though the weather is far from being a linear thing, we might even hope that there is a sufficient stability in the average patterns to allow us to separate the daily weather from the climate's long term variation.

It makes sense then that, in the debates of global warming, the average temperature of the planet is often referred to. This is a slightly bizarre idea. We all know that the temperature in Hawaii and the temperature in Alaska are very different most of the time, so temperature is not a uniform thing. Moreover, the retention of heat in the atmosphere is non-linear, with weather systems creating all manner of feedback loops and transport mechanisms to move heat around, store it and release it. Ocean currents, like the gulf stream, transport heat from a hot place to a cold place. The planetary climate system is as intricate as biology itself.

The average temperature somewhere on the planet is thus in a continuous state of re-evaluation. It changes, on the fly, on a timescale that depends on the scale of the region we look at. For the average temperature of the whole Earth to be unchanging, the various absorptions, reflections, and re-radiations would have to balance out, in a truly Byzantine feat of cosmic accounting. A net effect of zero means that for all the energy that comes in from the Sun, or is generated on the planet, the same amount of heat must leak away. Such a balance is called a state of *equilibrium*.

The word equilibrium (plural equilibria) is from the Latin *aequilibrium*, from aequus 'equal', and *libra* meaning a scale or balance (as in the zodiac sign). Its meaning is quite self-explanatory, at least in intent. However, as a phenomenon, the number of different ways that exist in the world for weighing measures is vast, and it is this that makes equilibrium perhaps the most important concept in science.

Once again, equilibrium is about scale in a number of ways. As any accountant knows, yearly profits and monthly cashflow are two very different things.

You might make a profit over the year, and still not be able to pay your bills on a monthly basis. Similarly, planetary heat input and dissipation both on the short term and in the long term are two very different issues. Of global warming, one sometimes hears the argument that all the fuel that is burned on the Earth is really energy from the Sun, because, over centuries, it is the conversion of raw materials into living systems that make trees and oil and combustible hydrocarbons. Thus nothing Man does with hydrocarbons can affect the long-term warming of the planet, since all the energy that went into them came from the Sun. On the short term, however, this makes little sense, since the energy stored in these vast memory bank oil reserves can be released in just a few years by little men, just as the chemical energy it took to make a stick of dynamite can be released in a split second. The balance of energy in and out thus depends on your bank account's savings and your spending. Timescales matter.

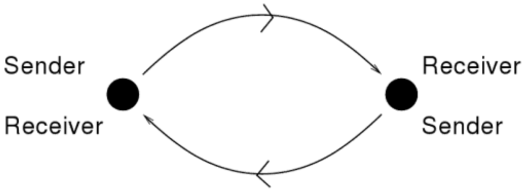

Fig. 4.4. The detailed balance between two parts of a system showing the flows that maintain equilibrium. The balance sheet always shows zero at equilibrium, but if we look on shorter timescales, we must see small imbalances arise and be corrected.

The basic idea of an equilibrium can be depicted as an interaction between two parties, as in Figure 4.4. The sender on the left sends some transaction of something (energy, data, money) to the receiver on the right at some kind of predictable average rate, and the receiver absorbs these transactions. Then, to maintain the balance, the roles are also reversed and the sender on the right transmits at an equivalent rate to the receiver on the left. The net result is that there is no build up of this transacted stuff on either side. It is like a ball game, throwing catch back and forth, except that there are usually far more transactions to deal with. In the case of global climate patterns there are staggering numbers of molecules flying back and forth in the atmospheric gases, transferring heat and energy in a variety of forms, and there are clouds and ocean currents transporting hot and cold water.

Equilibrium is what we observe when the slowly varying part of information grinds to a halt and becomes basically constant, so that one may see balance in the fluctuations. Equilibrium can be mechanical or seemingly *static*[77], like when a lamp hangs from the ceiling, held in balance between the cord it is attached to and the force of gravity, or it can be more *dynamic*, such as the heating of the Earth, or a monthly cashflow, with a continuous input and output of income and expenses. (Imagine putting a group of people on a very large balance and weighing them first standing still, and then jumping up and down.) When an equilibrium is disturbed, the shift perturbs a system and this can lead to instability. Sudden destabilization catastrophes can occur, leading to new metastable regimes, like ice ages, microclimates, and other things. Equilibrium of discrete informatic values is often called the *consensus* problem in information science[78].

Equilibrium might be the single most important idea in science. It plays a role everywhere in 'the accounting of things'. It is crucial to our understanding of heat, chemistry, and economics—and to the Internet.

Although the simple depiction of equilibrium in Figure 4.4 shows only two parties, equilibrium usually involves vast numbers of atomic parts, taken in bulk. That is because equilibrium is a statistical phenomenon—it is about averages, taken over millions of atoms, molecules, data, money, or whatever currency of transaction we are observing. Staying in town, or quivering in your shoes are also statistical equilibria.

The accounting that leads to equilibrium is called a situation of *detailed balance*. It is detailed, because we can, if we insist, burrow down into the details of each individual exchange of atoms or data or money, but it is the balance over a much larger scale that is the key. The principle of detailed balance was introduced explicitly for collisions by Boltzmann. In 1872, he proved his H-theorem using this principle. The notion of temperature itself comes from a detailed balance condition.

A simple example of detailed balance is what happens in a queue. Recall the supermarket checkout example in chapter 2. Customers arrive at a store, and they have to leave again, passing through the checkout. It seems reasonable to expect that this would lead to an equilibrium. But what if it doesn't? What if more customers arrive than leave? When a sudden change in this equilibrium happens, the balance of flows is affected and the detailed balance goes awry until a new stable state is reached. This is a problem for non-equilibrium dynamics. However, often we can get away without understanding those details, by dealing only with the local epochs of equilibrium, in the pseudo-constant grains of Figure 4.2.

Queueing theory begins with a simple model of this equilibrium that represents the arrival and processing of customers to a service handling entity, like a store, a call centre, or form-processing software on the Internet. In the simplest queueing theory, one imagines that customers arrive at random. At any moment, the arrival of a new customer is an independent happening, i.e. it has nothing to do with what customers are already in the queue. We say that the arrival process has no *memory* of what happened before. This kind of process is called a Markov process after Russian mathematician Andrey Markov (1856-1922), and it is a characteristic signature of statistical equilibrium. An equilibrium has no memory of the past, and no concept of time. It is a *steady state*.

Queueing theory can be applied by setting up the detailed balance condition and solving the mathematics. In normal terminology, we say that there is a rate of customers arriving, written λ requests per unit time, and there is a rate at which customers are served, called μ requests per unit time. From dimensional analysis, we would expect the behaviour of the queue to depend on the dimensionless ratio called the *traffic intensity*,

$$\rho = \frac{\lambda}{\mu}.$$

as indeed it does. To see that, we write the detailed balance as follows. A queue of length n persons will not be expected to grow or shrink if:

Expected arrivals = Expected departures

Or, as flow rates:

$\lambda\times$ (Probability of $n-1$ in queue) = $\mu\times$ (Probability of n in queue)

This says that, if a new person arrives when there are already $n-1$ persons in line, it had better balance the rate at which someone leaves once there are now n persons in line. This detailed balance condition can be solved to work out an average queue length that can be supported, based on the dimensionless ratio ρ:

$$n_{\text{average}} = \frac{\rho}{1-\rho}.$$

A picture of the results of this average length can be seen in Figure 4.5, and they are quite intuitive. As long as the arrival rate of customers is a bit less than the rate at which customers can be processed, ρ is less than the dimensionless number 1 and the queue length is small and under control. However, there is a very critical turnaround as one scale approaches the other, sending the queue into an unstable mode of growth. The suddenness of the turning point is perhaps surprising. The queue has a major instability, indicating that we should try to keep queues small at all times to avoid a total breakdown. Indeed, after a half

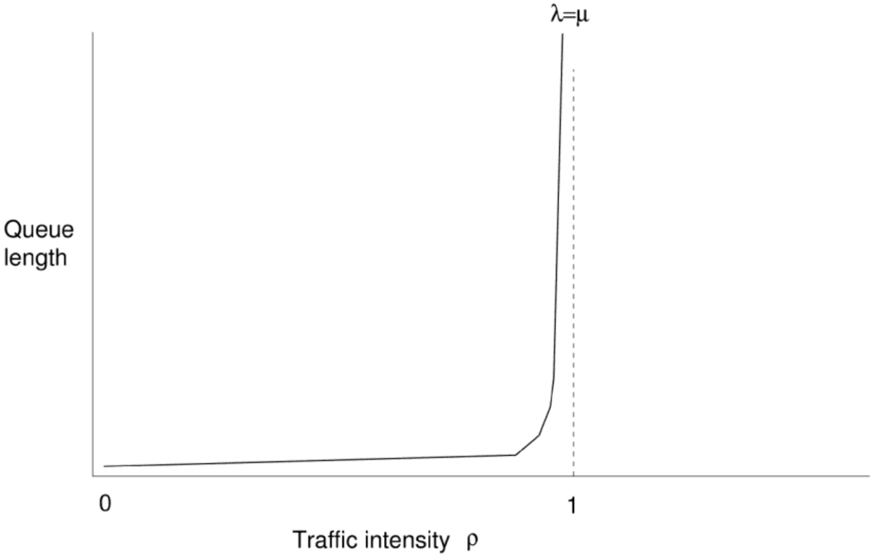

Fig. 4.5. The behaviour of the average queue length predicted by a detailed balance relation in queueing theory. The queue goes unstable and grows out of control when $\rho \to 1$. theory

century of queueing theory, the crux of what we know is very simple. A queue has basically two modes of operation. Either the expected queue length is small and stable, or it is growing wildly out of control—and the distance between these two modes is very small.

The model above is completely theoretical, but since it is based mainly on simple scaling assumptions, we would expect it to represent reality at least qualitatively, if not quantitatively. For computer processing, queueing is a serious issue, that affects companies financially when they make legal agreements called Service Level Agreements, claiming how fast they can process data requests. Companies need to invest in a sufficient number of servers at the ‘data checkout’ of their infrastructure to process arrivals in time.

In 2007, together with students Jon Henrik Bjørnstad, Sven Ulland and Gard Undheim, I studied how well this very real issue is represented by the simple model above, using real computer servers. For all the wonders of mathematics and modelling, simplicity is an asset when building technology, so we wanted to know if a simplest of arguments would provide good-enough guidance to datacentre infrastructure designers. The results can be seen in Figures 4.6 and 4.7.

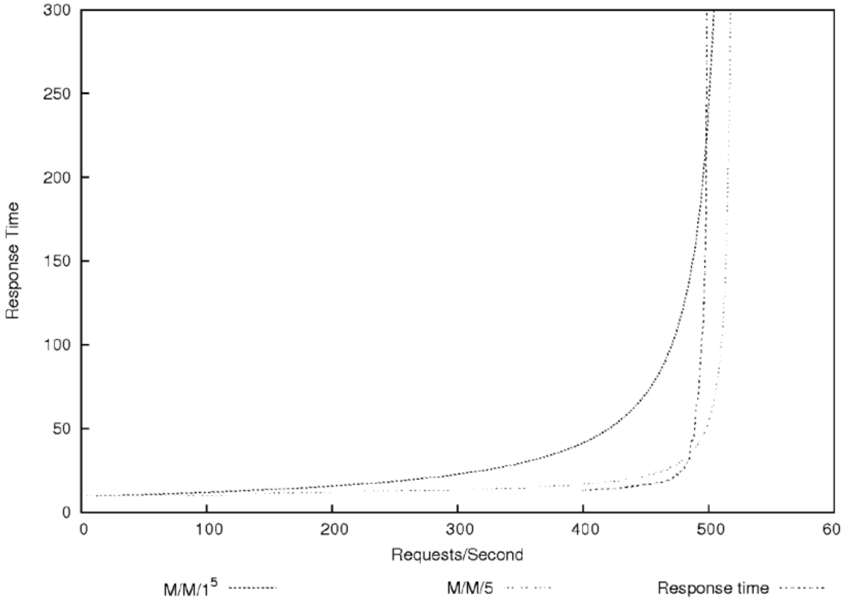

Fig. 4.6. Queue behaviour of five servers in different configurations, comparing the two models with measurements.

The different lines in Figure 4.6 show the same basic form as the theoretical graph in Figure 4.5, indicating that the basic qualitative form of the behaviour is quite well described by the simple model, in spite of many specific details that differ[79]. This is an excellent proof of the universality of dimensional arguments based on scaling arguments.

Figure 4.6 also shows the different ways of handling requests. The experiment was set up with five servers to process incoming online requests. Queueing theory predicts that it makes a difference how you organize your queue. The optimum way is to make all customers stand in one line and take the first available server from a battery of five as they become available (written M/M/5). This is the approach you will see at airports and other busy centres for this reason. The alternative, is to have a separate line for each sever (written $M/M/1^5$). This approach performs worse, because if the lines are not balanced, more customers can end up in a busy line, while another line stands empty. So we see, from the graph, that the single line keeps faster response times much closer to the critical turning point than the multiple-line queue, which starts going bad earlier—but both queue types fail at the same limit.

Unlike the simple theory, we can also see how the finite capacity of the com-