Chapter 6
Complex Systems: A New Epistemological Crisis

6.1 The Twenty-first Century: Starved for Data

The preceding chapter discussed the manner in which the modern scientific epistemology originating with Galileo reached a deep understanding in the first half of the Twentieth Century; however, the book on epistemology is far from closed. The epistemological challenges confronting the Twenty-first Century are the most severe since the dawning of the Seventeenth Century. They arise from a desire to model complex systems that exceed human conceptualization ability. As a consequence, people attempt to use highly flexible mathematical structures with large numbers of parameters that can be adjusted to fit the data, the result often being models that fit the data well but lack structural representation of the phenomena and thus are not predictive outside the range of the data. The situation is exacerbated by uncertainty regarding model parameters on account of insufficient data relative to model complexity, which in fact means uncertainty regarding the models themselves. More importantly from the standpoint of epistemology, the amount of available data is often miniscule in comparison to the amount needed for validation. The desire for knowledge has far outstripped experimental/observational capability. We are starved for data.

With all the talk these days of "Big Data," one must remember that bigness is relative to need. While the current amount of data may be big relative to small systems, it is paltry compared to the data required for large complex systems, especially if it is not collected with a sharp eye to the intended use, which often it is not. We need only recall the warnings of Bacon and Kant about groping in the dark. With complex systems, experimental design is even more imperative. Still, with or without experimental design, in many cases it is virtually impossible to obtain the data required for model validation.

6.2 Gene Regulatory Networks

The Twenty-first Century is sometimes viewed as the century of biology; yet in biology complexity reaches heights undreamed of until very recently. A human body consists of trillions of cells containing about 100,000 different types of proteins and 30,000 genes interconnected in a myriad of signaling pathways, and let us not forget that each gene consists of a region of DNA, and the genome is subject to an immense number of single nucleotide polymorphisms, which are variations in a single nucleotide. We will discuss complexity in the context of modeling gene regulation in a single cell, which, although it represents only a small portion of the full system, presents unworkable levels of complexity even when only a relatively small number of genes are involved.

The regulatory system in a cell is mainly based in its genetic structure. The basic paradigm has two parts. *Transcription* refers to the process by which the genetic information in a gene is copied into messenger RNA (mRNA). When this process is occurring the gene is said to be *expressing* (or activated). Expression is governed by signaling proteins attaching themselves (binding) to the gene's *promoter region*. In essence, each gene is controlled by the states of a set of genes, so that its activation or non-activation depends on a combination of the expression levels in its regulating genes. *Translation*, which occurs subsequent to transcription, refers to the production of protein based on the code carried by the mRNA. The resulting protein can either be involved in maintaining the cell structure or function as a signal (*transcription factor*) to instigate or prohibit further gene expression by binding to the promoter region of a gene and forming a complex with other transcription factors to regulate the gene. This process goes on continuously across the genome to produce signaling pathways that regulate gene activity dynamically. Other factors affect gene activity, but we will focus solely on this basic transcriptional system.

A *gene regulatory network* (GRN) is a mathematical model comprised of a set of entities called "genes" and a regulatory structure that governs their behavior over time. GRNs can be finely detailed, as with differential-equation models, or coarse-grained, with discrete expression levels transitioning over discrete time. There is no expectation that coarse models closely represent actual molecular structure; rather, their purpose is to model interaction at the gene level in order to serve as a framework for studying regulation and provide rough models that can be used to develop strategies for controlling aberrant cell behavior, such as finding optimal drug treatments. While it might appear that gene-level modeling mistakenly ignores the molecular interaction constituting genetic activity, as well as the myriad of other molecular activity in a cell, it needs to be recognized that, while biological function requires chemistry, biology is not chemistry. Although there is no clear dividing line, biology concerns the operation of the cell at the level of genes, proteins, and other macromolecules involved in the life functions of the cell, not the physiochemical infrastructure of these macromolecules.

6.2.1 Deterministic Boolean networks

In the late 1960s, Stuart Kauffman introduced a discrete model known as a *Boolean network* [Kauffman, 1993]. Each gene can have logical values 1 or 0, corresponding to expressing or not expressing, respectively, and regulation is specified by logical operations among genes. Thus, the functional relationships between genes can be specified by a truth table. While the Boolean model is very coarse, it does model the thinking of cancer biologists, who speak of a gene being on or off under different conditions. Moreover, although the original formulation is two-valued, 0 or 1, the concept applies to any number of discrete gene values.

Formally, a Boolean network is defined by k binary variables, $x_1, x_2, \ldots, x_k$, where the value x_i of gene g_i at time $t + 1$ is determined by the values of some regulator genes at time t via a Boolean function f_i operating on the regulator genes. A typical function would be of the form $x_3 = f_3(x_2, x_4) = x_2 \wedge x_4$, where $\wedge$ means "and." This means that gene g_3 is on (expressing) at time $t + 1$ if and only if genes g_2 and g_4 are on (expressing) at time t. There are k such Boolean functions, one for each gene, and together they determine the deterministic dynamic evolution of the system over time. If there are four genes, then a typical dynamic trajectory over three time points would look like 0101 → 1100 → 1101. Given an initial state, a Boolean network will eventually reach a set of states, called an *attractor cycle*, through which it will cycle endlessly. Each initial state corresponds to a unique attractor cycle and the set of initial states leading to a specific attractor cycle is known as the *basin of attraction* of the attractor cycle.

We consider a small network involving the tumor suppressor gene p53. In mammalian genomes p53 is a transcription factor for hundreds of downstream genes that modulate cell cycle progression, repair damaged DNA, and induce senescence and apoptosis (cell self-destruction). Figure 6.1 shows some major pathways involving p53 that are activated in the presence of DNA double strand breaks. Adapted from [Batchelor, et al., 2009], it is not meant to be inclusive. An arrow indicates an activation signal, and a blunt end indicates suppression. Note that p53 activates Mdm2 and activated Mdm2 has a suppressing effect on p53. Even in this small network one can see the complicating effect of feedback.

Given this kind of pathway diagram, which is inherently logical, one would like to find Boolean networks whose state transitions generate the pathways [Layek et al., 2011]. The problem is ill-posed because there may be numerous networks that realize the pathways and there may be logical inconsistencies among the pathways since they have been found under various conditions in different studies. These kinds of issues are common with complex systems.

We consider two Boolean networks having states [ATM, p53, Wip1, Mdm2] generated for the pathways in Fig. 6.1. An external input signal, denoted dna_dsb, takes on the value 1 or 0, depending on whether there is or is not DNA damage. This leads to two 4-gene Boolean networks determined by the following logical rules [Imani and Braga-Neto, 2016]:

$$\mathrm{ATM}_{\mathrm{next}} = \overline{\mathrm{Wip1}} \wedge \mathrm{dna\_dsb}$$

$$\mathrm{p53}_{\mathrm{next}} = \overline{\mathrm{Mdm2}} \wedge \mathrm{ATM} \wedge \overline{\mathrm{Wip1}}$$

$$\mathrm{Wip1}_{\mathrm{next}} = \mathrm{p53}$$

$$\mathrm{Mdm2}_{\mathrm{next}} = (\overline{\mathrm{ATM}} \wedge (\mathrm{p53} \vee \mathrm{Wip1})) \vee (\mathrm{p53} \wedge \mathrm{Wip1})$$

The symbols $\wedge$, $\vee$, and $\overline{\ }$ represent logical "and", "or", and "not", respectively.

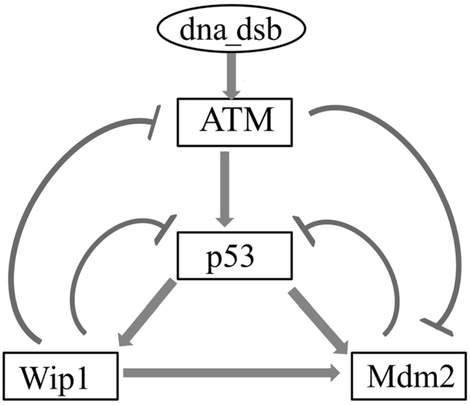

Figure 6.1 p53 pathways (adapted from [Imani and Braga-Neto, 2016]).

The state transition diagrams for these networks are shown in Fig. 6.2: (a) dna_dsb = 0; (b) dna_dsb = 1. Absent damage, from any initial state the network evolves into the single attractor state 0000; with damage, the network evolves into a 5-state attractor cycle in which p53 (state number 2) oscillates between expressing and not expressing. If one were to observe the network without knowing the damage status, then network behavior would appear stochastic, for instance, 0001 → 0000 when dna_dsb = 0 and 0001 → 1000 when dna_dsb = 1.

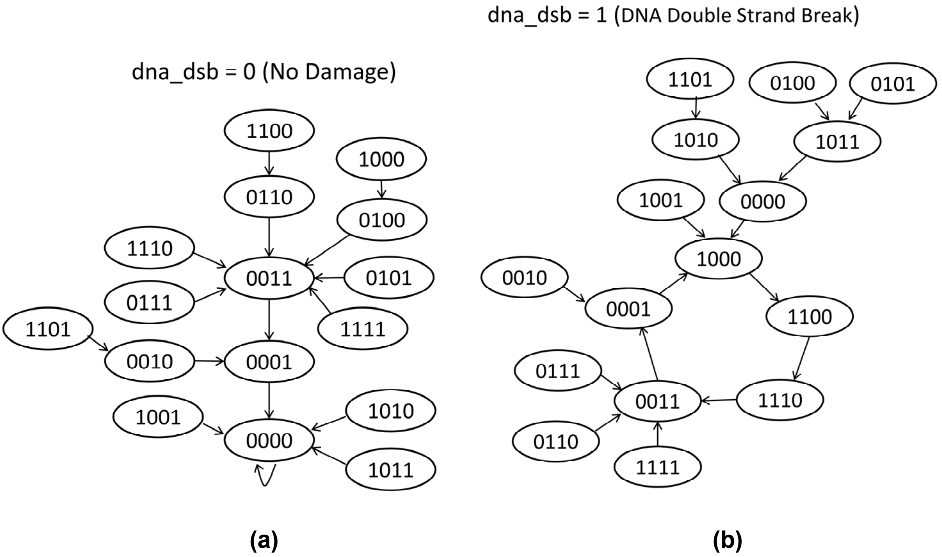

Figure 6.2 State transition diagrams for p53 networks: (a) no damage—single attractor state; (b) damage—five-state attractor cycle.

6.2.2 Probabilistic Boolean networks

From the perspective of each of the two p53 networks, the damage signal is a *latent variable* exterior to the network. Given the value of the latent variable, the network is deterministic; however, latency means that the damage signal is not part of the network and is not observed. Hence, when observed the network is stochastic. One might argue that the problem would be solved by including dna_dab in the network. That would just push the latency further out because dna_dab is being influenced by other unobserved physical events. The central point is that the model system cannot be isolated from interaction with its environment, so, recalling Russell, even if the universe is deterministic, one would have to include all events not totally disconnected from the network, a practical impossibility.

One can incorporate both p53 Boolean networks into a single network by viewing each individual Boolean network as a *context* (constituent) of a network whose regulatory structure is defined at a given time point by setting the damage signal to either 0 or 1. The new network maintains that regulatory structure until it randomly switches to the other Boolean regulation, say dna_dsb = 0 to dna_dsb = 1, with some *switching probability*. The resulting network is called a *probabilistic Boolean network* (PBN) [Shmulevich and Dougherty, 2010]. The PBN inherits the attractor structures of the constituent Boolean networks, the difference being that context switching can result in the network jumping out of an attractor cycle into a different basin of attraction and then transitioning into a different attractor cycle. While the p53 PBN has two contexts, the general definition of a PBN allows any number of context Boolean networks. It also does not require binary-valued genes.

To illustrate network (and biological pathway) switching, suppose there is no damage and the network has settled into the attractor state 0000, as shown in Fig. 6.2(a). Since the role of the p53 network is to respond to DNA damage and since there is no damage, this dormant state is what one might expect. Suppose DNA damage is detected. Then dna_dsb flips to 1, the Boolean network of Fig. 6.2(b) becomes operative, and the state changes from 0000 to 1000 in the next time step, so that almost immediately the 5-state cyclic attractor is entered and p53 oscillates between 0 and 1 on each cycle.

The PBN model incorporates randomness in a structured manner. Should this uncertainty be considered intrinsic, as in the case of quantum mechanics? One could certainly argue that there are hidden variables and that, if we could observe all of them, then the uncertainty would be eliminated. The debate is academic because the physical system is too complex and consists of tens of thousands of variables—genes, proteins, and other macromolecules within a single cell plus all elements pertaining to extra-cellular signaling. Forming a sufficiently extensive model to eliminate latency is impossible. There are two choices: use a deterministic model if the latency is very small, or include the latency-induced stochasticity in the model, as with PBNs.

Further randomness can be introduced to a Boolean network via perturbations. Specifically, for each gene there is some small *perturbation* *probability* that it will randomly switch values. This is practical because there is random variation in the amount of mRNA and protein produced. Perturbations allow a network to jump out of an attractor cycle and, as with context switching, eventually transition to a new attractor. A probabilistic Boolean network is usually assumed to have perturbation randomness in addition to context-switching randomness.

6.3 Validation of Complex Systems

In the classical deterministic scenario, a model consists of a few variables and physical constants. The relational structure of the model is conceptualized by the scientist via intuition gained from thinking about the physical world. Intuition means that the scientist has some mental construct regarding the interactions beyond positing a skeletal mathematical system he believes is sufficiently rich to capture the interactions and then depending upon data to infer the relational structure and estimate a large number of parameters. Classically, there are few parameters to estimate and they are estimated from a handful of experiments. Owing to the deterministic character of the model, it can be tested with a few numerical predictions whose disagreement with future observations is due to either experimental error or model failure, with the former being mitigated by careful experimentation. The theory is contingently accepted if predictions are deemed to be concordant with observations.

As model complexity grows to tens, then hundreds, and then thousands of variables and parameters, the classical procedures become increasingly difficult to carry out. The problem is exacerbated by stochasticity because prediction then includes testing the accuracy of probability distributions in the model. Systems with thousands of variables are virtually unvalidatable.

6.3.1 Validation of deterministic models

For a deterministic model, initial conditions can be set and, in principle, the state at some future time determined exactly, although in practice there will be some experimental variability. If the initial conditions of a test experiment are aligned with those of the model and the experiment run to some future time, then agreement between the final model and experimental states can be checked. Large-scale deterministic systems have high-dimensional state vectors, so that test experiments are more demanding; nevertheless, the ultimate comparison is still between model and experimental state vectors. It is prudent to run tests using a variety of initial conditions so that a large portion of the state space is tested.

Consider validating a Boolean network with k genes. Initializing the state vector at $\mathbf{x}_0$, one determines the state vectors $\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_b$ at times $t = 1, 2,\ldots, b$ via the regulatory logic, initializes the experimental set-up at $\mathbf{z}_0$, runs the experiment taking measurements at each step to compute $\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_b$, and checks for agreement between $\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_b$ and $\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_b$, or perhaps just at some subset of time points.

To see why it is prudent to consider various initial conditions, suppose the Boolean network has two attractor cycles A_1 and A_2, with corresponding basins B_1 and B_2. If the initial state lies in basin B_1, then after some number of steps the network will arrive in attractor cycle A_1. If A_1 and A_2 correspond to modeling two different phenotypes, since the regulatory pathways in the different phenotypes are different, the model might be a good fit for one phenotype but not the other, and this would never be tested by a single initial condition in basin B_1. One would at least need to test initial conditions in the two basins. Beyond that, a single initial condition in a basin will lead to a specific state trajectory, so that disagreements on other pathways might not show up. Thus, prudence dictates testing a number of initial conditions. As network complexity increases, so does the number of tests.

As an example, suppose there is a mutation and the p53 network of Fig. 6.2(b) is altered so that state 0000 becomes an attractor; that is, the network stays in 0000 when there is DNA damage. This can happen with a single alteration in the regulatory logic: when the network is in state 0000, instead of $\text{ATM}_{\text{next}} = 1$, $\text{ATM}_{\text{next}} = 0$. This is a serious mutation because p53 remains off when there is DNA damage so that the downstream effects that it should actuate are not actuated. Regarding validation, the mutated network has two attractors, the singleton 0000 and the original 5-state attractor cycle. If one proceeds to validate the network starting from initial state 1101, then the experiment should end with state 0000. Is this sufficient validation? All that has been tested is the path 1101 → 1010 → 0000. What about initialization at 0101 or, more importantly, at 0111 or 1111, where the cyclic attractor would be tested? It is clear that testing must involve more than a single initial state.

6.3.2 Validation of stochastic models

With a stochastic model, the situation is more challenging. Given an initial state, the final state will not be determined exactly; rather, there will a probability distribution of possible final states. Hence, comparison must be between the state distribution, which is generally multivariate, and a state histogram generated by many experimental runs, the number of required runs growing exponentially with the number of variables. Distributional statistical tests are required. For instance, with hypothesis testing one decides between two hypotheses—the distributions match or they do not match. A decision to accept the theory depends on the acceptance threshold. The theory and test are inter-subjective, but the decision to accept or reject depends on subjective considerations, as with a hypothesis test, where the acceptance region depends on a chosen level of significance. The overall procedure can be onerous (or impossible) depending on the number of experimental runs required, especially with complex systems, where distributions are high-dimensional. Validation of a wave pattern in the double-slit experiment constitutes a low-dimensional example of the method: compare the electron distribution on the detector with the pattern predicted by the wave model.

To illustrate the problem, consider the p53 network in Fig. 6.2(b). State 0000 is important because, if there is no damage, then the ground state is 0000, but now there is damage. At once the cyclic attractor is entered, so that oscillation of p53 takes place. Now suppose ATM is unstable and is subject to perturbation with some probability. While cycling through the attractor states, suppose at state 0011 ATM flips to 1 so that the network is in state 1011. It will then transition through 0000 into the attractor cycle. After several cycles suppose the network arrives at state 1100 and ATM flips to 0 so that the network is in state 0100. Then it will transition through 1011 and 0000 to again be in the attractor cycle. The point is that starting at the initial state 0000 the network will not reach a determined state after a given amount of time; instead, there will be probabilities of being in many, or all, states. To check the model this probability distribution must be tested against living cells, which is extremely difficult even for modest sized networks. This is for one initial state among 16 possible initial states. For a Boolean network with k genes there are 2^k possible initial states.

6.4 Model Uncertainty

Parameter estimation is a basic aspect of model construction and historically it has been assumed that data are sufficient to estimate the parameters, for instance, correlations that are part of the model; however, when the number of parameters is too large for the amount of data, accurate parameter estimation becomes impossible. The result is model uncertainty.

Insufficient data for accurate estimation is an old problem in statistics. For a simple illustration, consider a one-dimensional variable governed by a normal distribution with known standard deviation and unknown mean μ. The standard method of estimating μ is to take a random sample of points $x_1, x_2, \ldots, x_n$ and form the *sample mean* $(x_1 + x_2 + \ldots + x_n)/n$. The sample mean provides a good estimate of the mean if the sample size n is sufficiently large. The precision of the estimate can be quantified in terms of the sample size. If the sample size is small, then rather than a point estimate it may be preferable to provide an interval estimate of the form $[a, b]$, so that there is no specific estimate of the mean. In effect, this means that one is assuming that the "true" model is among the infinite number of possible models compatible with the interval estimate.

For a situation in which model complexity plays a role, consider the p53 network for no damage and suppose that the regulatory function for ATM is unknown. The truth table defining the regulatory structure for the network has 64 $= 4\times2^4$ rows because there are 2^4 possible input states for each of the four genes: 0000, 0001,…, 1111. This means that there are 64 parameters taking values 0 or 1. If there is no existing knowledge concerning the regulation of ATM, then there are 16 unknown parameters: $f_1(0000), f_1(0001), \ldots, f_1(1111)$. Since each of these can have two possible values, 0 or 1, there are 2^{16} possible networks, one for each combination. Owing to uncertainty, instead of one network there is an *uncertainty class* of 65,536 possible networks. Each is represented by a parameter vector θ_k of length 16, so that the uncertainty class takes the form $\Theta = \{\theta_1, \theta_2, \ldots, \theta_{65,536}\}$. This is for a single unknown regulatory function in a single 4-gene binary network!

If there is prior knowledge that can be applied, then the uncertainty class can be reduced. For example, suppose it is known that $\text{ATM}_{\text{next}} = 0$ if Wip1 = 1. This knowledge would result from a scenario in which the presence of the Wip1 transcription factor on the promoter region of ATM blocks the binding of activating proteins. In this case, there are only 8 unknown parameters, $f_1(0000)$, $f_1(0001),\ldots, f_1(0111)$, and 2^8 networks in the uncertainty class. This kind of huge complexity reduction puts a premium on prior (existing) knowledge in model construction. The effect of prior knowledge will be seen in the next chapter when we discuss model-based operator design.

6.5 Data Mining

The classical approach to model design is to construct a mathematical structure satisfying the scientist's conceptualization of phenomenal behavior and then estimate model parameters. As models become more complex, in addition to increasing numbers of parameters to estimate, conceptualizing interacting phenomena becomes more taxing. Thus, it has become popular to posit a very general mathematical structure and then, instead of using some statistically best estimate such as maximum likelihood to estimate individual parameters, the parameters are manipulated as a group until the model fits the data to some desired degree. Data-fitting algorithms can be ingenious and may take advantage of high-performance computing to employ models with thousands of parameters.

6.5.1 Overfitting

At first glance, this approach, known as *data mining*, may seem attractive and appear to circumvent the need for conceptualization; however, fitting the data without a satisfactory conceptualization of the interactions (law) underlying the behavior of the phenomena can easily lead to a model that *overfits* the data. The model fits the data but does not model the relevant physical processes, the result being that it poorly predicts future observations and may not even successfully predict existing data not used in model construction. Indeed, the mathematical structure (neural network, graph, etc.) may not be of a suitable form to model the physical processes but is sufficiently flexible on account of its complexity and high dimensionality that it can be fit to the data. To add to the predicament, even if the fitted structure should happen to provide a good model for the underlying processes, there often is no method for precisely estimating its accuracy. Hence, if it is accurate, there is no way to know so.

Climate scientists Tibaldi and Knutti articulate the problem as manifested in their discipline:

> Most models agree reasonably well with observations of the present-day mean climate and simulate a realistic warming over the Twentieth Century (of course, the specific performance depends on each model/metric combination), yet their predictions diverge substantially for

> the Twenty-First century, even when forced with the same boundary conditions. [Tibaldi and Knutti, 2007]

Recall Reichenbach: "Observation informs us about the past and the present, reason foretells the future." Perhaps some reason has been used in constructing climate models, but not enough. Faced with the complexity of climate systems, is it reasonable to believe that there can ever be enough reason?

To illustrate overfitting, consider the problem of finding a *regression* function $y = g(x)$ that best estimates the value of Y given a value of X, where X and Y possess a joint distribution. We denote the random value of Y given a fixed value $X = x$ by $Y|x$. The best estimate in the mean-square sense is the one that minimizes the average value of $|Y|x - \gamma_x|^2$ among all possible estimates γ_x. This average value is known as the expected value and is denoted by E, so the aim is to minimize $E[|Y|x - \gamma_x|^2]$. The minimum mean-square estimate is the mean of $Y|x$, which is denoted by $\mu_{Y|x}$.

In the case of a bivariate normal distribution, if the means of X and Y are μ_X and μ_Y, respectively, their standard deviations are σ_X and σ_Y, respectively, and the correlation coefficient is ρ, then the regression function is given by

$$\mu_{Y|x} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \tag{6.1}$$

which is a straight line with slope $\rho\sigma_Y/\sigma_X$.

A basic problem in statistics is to estimate the regression function from points randomly drawn from the joint distribution. Since for normal distributions the regression function is a straight line, given a joint normal distribution the estimated regression function is taken to be a straight line also. This *sample regression line* constructed from the data is the line $y = a + bx$ that minimizes the error sum of squares

$$\text{SSE} = |y_1 - (a + bx_1)|^2 + |y_2 - (a + bx_2)|^2 + \ldots + |y_n - (a + bx_n)|^2, \tag{6.2}$$

where the sample points are (x_1, y_1), (x_2, y_2),…, (x_n, y_n). As the number of data points grows, the sample regression line becomes closer to the true regression line (in a probabilistic sense).

Suppose one does not know that the joint distribution is normal. Then the true regression line can take almost any form. Should the regression line be highly nonlinear, then assuming a straight line, in particular, using the best-fit regression line for a normal distribution would constrain the estimation to one that is far from accurate regardless of the number of data points. To avoid this kind of constraint, instead of assuming a linear regression, one can assume a polynomial regression. But what order polynomial should be chosen? Should it be high order to better fit the data? Such a choice may provide excellent data fitting on account of complexity and the large number of parameters to be adjusted, but this may result in overfitting if the assumed regression model is overly complex relative to the true regression equation.

Figure 6.3 provides an example involving a joint normal distribution with means $\mu_X = \mu_Y = 3$, standard deviations $\sigma_X = \sigma_Y = 1$, and correlation coefficient $\rho = 0.5$. Each part of the figure shows ten randomly generated data points, the true regression line, and a sample regression line found via the error sum of squares for the assumed form of the line: linear, cubic, fifth-order polynomial, seventh-order polynomial, ninth-order polynomial, and eleventh-order polynomial. As the order of the polynomial grows, the sample regression line fits the data better but gets further away from the true regression line. This is classic overfitting.

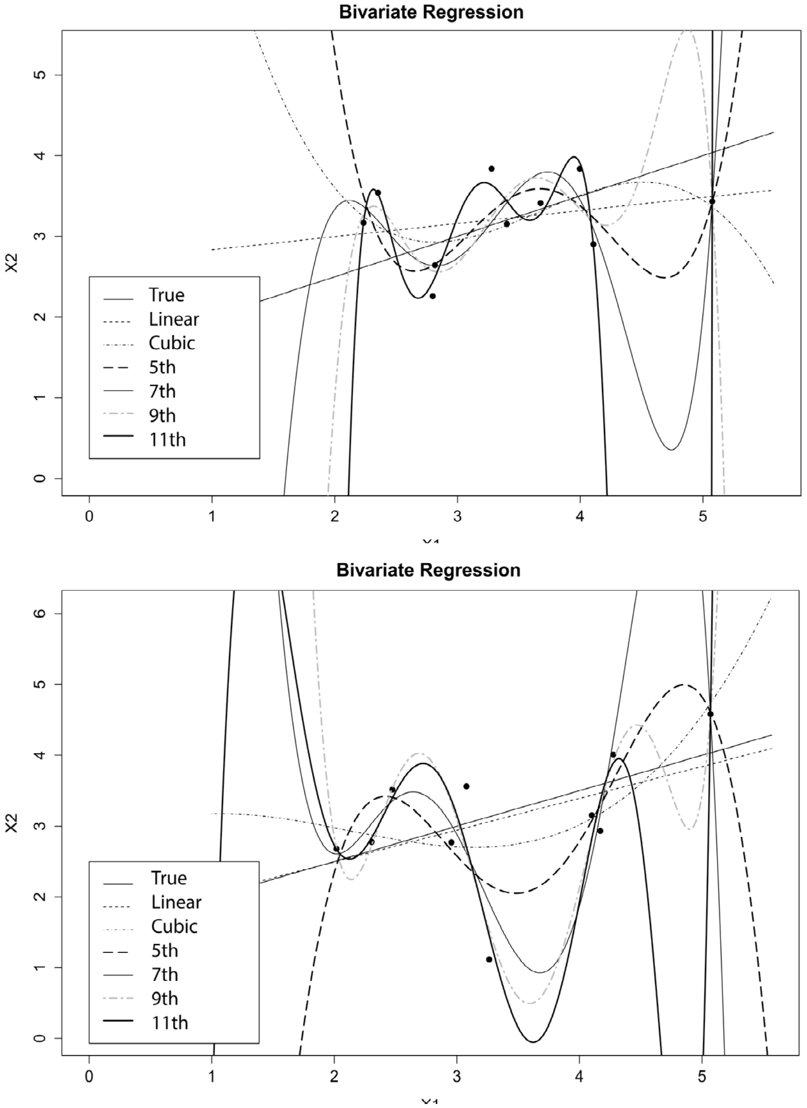

Figure 6.3 Polynomial regression lines of increasing degree fit to two different sets of 10 randomly generated points from a bivariate normal distribution.

If fitting the data is the sole aim, then having enough computing power to fit a complex model, say, one with tens of thousands of parameters in an equally vast dimensional space, is typically the sole issue; however, scientifically, fitting the data is not a sensible aim. A hundred points are lost in thousand-dimensional space and are easily overfit. Think of modeling the approximately 30,000 genes in the human genome. The bigness of data depends on its relation to model dimension, not simply the number of gigabytes.

The complexity dilemma—choosing low model complexity and not capturing behavioral complexity versus choosing high model complexity and overfitting the data—is caused by ignorance. One is trying to model phenomena without sufficient knowledge to do so.

Maxwell addresses the issue squarely:

> As students of physics we observe phenomena under varied circumstances and endeavor to deduce the laws of their relations. Every natural phenomenon is, to our minds, the result of an infinitely complex system of conditions. What we set ourselves to do is to unravel these conditions, and by viewing the phenomenon in a way which is in itself partial and imperfect, to piece out its features one by one, beginning with that which strikes us first, and thus gradually learning how to look at the whole phenomenon so as to obtain a continually greater degree of clearness and distinctness. In this process, the feature which presents itself most forcibly to the untrained inquirer may not be that which is considered most fundamental by the experienced man of science; for the success of any physical investigation depends on the judicious selection of what is to be observed as of primary importance, combined with a voluntary abstraction of the mind from those features which, however attractive they appear, we are not yet sufficiently advanced in science to investigate with profit. [Maxwell, 2003]

In Maxwell's phraseology, an "untrained inquirer" throwing together a huge number of features in the hope that some data-mining algorithm in conjunction with massive computational machinery will discover a nugget is "not yet sufficiently advanced in science." Or, as stated by William Barrett, "The absence of an intelligent idea in the grasp of a problem cannot be redeemed by the elaborateness of the machinery one subsequently employs." [Barrett, 1979]

6.5.2 Asymptotic theory

The complexity dilemma arises from insufficient knowledge to make sufficient assumptions to render principled model design feasible. Modeling assumptions carry risk in the sense that the phenomena may not satisfy them; in fact, they will almost certainly not satisfy them. Nevertheless, absent assumptions there can be no propositions. Omitting distributional assumptions might seem desirable so as not to limit the scope of the theory; however, as seen with regression, the absence of distributional assumptions easily leads to meaningless results.

Can we appeal to asymptotic (sample size → ∞) statistical theory to guarantee model accuracy? Theorems concerning the convergence to zero of the difference between a parameter estimate and the parameter as sample size goes to infinity go back to Jacob Bernoulli (1655–1705). At best, asymptotic results may say something about estimation accuracy for large samples but they say virtually nothing about small samples—and small samples are the problem for complex systems. Even if data are abundant, unless there are distributional assumptions, an asymptotic theorem usually does not specify how large the sample must be and assumptions have to be imposed to obtain propositions concerning required sample size.

In 1925, Ronald Fisher commented on the limitations of asymptotic theory:

> Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a canon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data. [Fisher, 1925]

Twenty years later, Harald Cramér strongly supported Fisher's position:

> It is clear that a knowledge of the *exact* form of a sampling distribution would be of a far greater value than the knowledge of a number of moment characteristics or a limiting expression for large values of n. Especially when we are dealing with *small samples*, as is often the case in the applications, the asymptotic expressions are sometimes grossly inadequate, and a knowledge of the exact form of the distribution would then be highly desirable. [Cramér, 1945]

Fisher and Cramér, two giants of statistics, make it very clear that real-world problems are often small-sample problems and, for these, asymptotic theory will not do—and they never witnessed today's complexity. Small-sample theory is necessary for statistics to play a major role in acquiring scientific knowledge. For the most part, data mining, which is void of small-sample theory, is high-performance pre-Baconian groping in the dark.

6.6 Limitations of Science

While post-Galilean science has from the outset been restricted to mathematical representation and the ability to perform confirming experiments, the strong limitations of science, as a form of knowledge, implied by these restrictions has become clearer with the desire to apply scientific method to complex stochastic systems. The stumbling block is that the predominant problems in the Twenty-first Century are very different from Einstein's $E = hf$, which only requires estimating Planck's constant. Even modest-sized models in biology contain large numbers of parameters, dwarfing the complexity of the p53 network considered herein. Model uncertainty together with stochasticity precludes the possibility of full-model validation. Partial validation via prediction of some characteristics (features or properties) of the model may be feasible; however, even accepting Einstein's stipulation that "it is only necessary that enough propositions of the conceptual system be firmly enough connected with sensory experiences," this proviso must be applied to such a degree that validation can, at best, be only fragmentary.

Beyond the impediment of mathematical and computational complexity, limitations on measurement accuracy and the inability to perform the large number of experiments required to validate large stochastic systems limit the degree of validation, and therefore the knowledge carried by a model.

A salient example of experimental limitation on scientific knowledge occurs in climate science, where model validation can involve various characteristics, such as mean global temperature and the amount of atmospheric CO_2. While these may be weak compared to full-model validation, application-wise they are important. Because the system is stochastic, prediction involves distributions and data must be obtained for constructing empirical distributions. Is this possible? If a prediction involves the earth and takes place over a long time period, then it may be hard to draw a sufficient number of points. For a time period of ten years, even without random initialization and using successive ten-year periods, it would take a millennium to generate a decent histogram. Reducing validation to model characteristics does not help; the impediment is that sufficient observation of the system is impossible.

Tibaldi and Knutti state the problem:

> The predictive skill of a model is usually measured by comparing the predicted outcome with the observed one. Note that any forecast produced in the form of a confidence interval, or as a probability distribution, cannot be verified or disproved by a single observation or realization since there is always a non-zero probability for a single realization to be within or outside the forecast range just by chance. Skill and reliability are assessed by repeatedly comparing many independent realizations of the true system with the model predictions through some metric that quantifies agreement between model forecasts and observations (e.g. rank histograms). For projections of future climate change over decades and longer, there is no verification period, and in a strict sense there will never be any, even if we wait for a century. The reason is that the emission scenario assumed as a boundary condition is very likely not followed in detail, so the observations from the single climate realizations will never be fully compatible with the boundary conditions and scenario assumptions made by the models. And even if the scenario were to be followed, waiting decades for a single verification dataset is clearly not an effective verification strategy. This

> might sound obvious, but it is important to note that climate projections, decades or longer in the future by definition, cannot be validated directly through observed changes. Our confidence in climate models must therefore come from other sources. [Tibaldi and Knutti, 2007]

Tibaldi and Knutti confront the epistemological crisis of the Twenty-first Century: the desire for valid scientific knowledge and the inability to get it on account complexity or experimental limitations. They state that "climate projections, decades or longer in the future by definition, cannot be validated directly through observed changes." Combine this with Schrödinger's statement that "there does not seem to be much sense in inquiring about the real existence of something, if one is convinced that the effect through which the thing would manifest itself, in case it existed, is certainly not observable." One might argue that climate projections are not theoretically impossible, only pragmatically impossible. But does this matter in practice? Tibaldi and Knutti say that confidence must come from "other sources," but this does not produce a validated scientific theory. There is no scientific truth.

Confronting the limits of verifiability in evolutionary theory, Kauffman calls for a new scientific epistemology:

> What we think of as natural law may not suffice to explain Nature. We now know for example, that evolution includes Darwinian pre-adaptations—unused features of organisms that may become useful in a different environment and thus emerge as novel functionalities, such as our middle ear bones, which arose from the jaw bones of an early fish. Could we pre-state all the possible Darwinian pre-adaptations even for humans, let alone predict them? It would seem unlikely. And if not, the evolution of the biosphere, the economy and civilization are beyond natural law. If this view holds, then we will undergo a major transformation of science. [Kauffman, 2007]

Kauffman is expressing a desire for knowledge that lies outside the bounds of science but he wants it to be scientific in character. This can only be achieved if the requirements for scientific knowledge are weakened.

Regarding the inability to make predictions, in his essay, "Breaking the Galilean Spell," Kauffman writes,

> This incapacity to foresee has profound implications. In the physicist Murray Gell-Mann's definition, a 'natural law' is a compact description beforehand of the regularities of a process. But if we cannot even prestate the possibilities, then no compact descriptions of these processes beforehand can exist. These phenomena, then, appear to be partially beyond natural law itself. This means something astonishing and powerfully liberating. We live in a universe, biosphere, and human

> culture that are not only emergent but radically creative. We live in a world whose unfoldings we often cannot prevision, prestate, or predict—a world of explosive creativity on all sides. This is a central part of the new scientific worldview. [Kauffman, 2008]

Standing in opposition to Kauffman's new scientific worldview is physicist Lee Smolin, who, in reference to string theory, writes,

> A theory has failed to make any predictions by which it can be tested, and some of its proponents, rather than admitting that, are seeking leave to change the rules so that their theory will not need to pass the usual tests we impose on scientific ideas. It seems rational to deny this request and insist that we should not change the rules of science just to save a theory that has failed to fulfill the expectations we originally had for it. [Smolin, 2006]

The conflict between the desire for knowledge concerning complex systems and the impossibility of testing a model by observing future behavior lies at the center of the epistemological crisis of the Twenty-first Century. There appear to be four basic options for science:

1. Dispense with modeling complex systems that cannot be validated.
2. Model complex systems and pretend they are validated.
3. Model complex systems, admit that the models are not validated, utilize them pragmatically where possible, and be extremely prudent when interpreting them.
4. Strive to develop a new and perhaps weaker scientific epistemology.

Option three carries the risk of eviscerating science as a result of laziness; however, option one leaves major problems in medicine, engineering, economics, etc. that have substantial impact on the human condition outside of systematic investigation. Option three is certainly better than option two, which appears to be widespread. Recall Woodcock's estimate that as much as 75% of published biomarker associations are not replicable—and although these may be high dimensional, their complexity is low compared to other systems being investigated. Pretending that theories are scientifically valid when they are not inevitably leads to poor policy decisions by political leaders who must put their faith in science, while at the same time rendering the scientific literature suspect. Pursuing option three may motivate a serious effort in regard to option four, which could lead to a multi-level epistemology that would support meaningful scientific theories at different levels of validation.

If the requirements of science are to be weakened, this needs to be done with great care, deep philosophic reflection, and in a manner that maintains a rigorous formal relationship between theory and phenomena. Given the substantial obstacles confronting the pursuit of scientific knowledge in complex systems, a satisfactory resolution could easily be a century or more away, if at all. Human beings are limited in their capacity for knowledge. It took three centuries from the birth of modern science until quantum theory to fully clarify the epistemological revolution of Galileo, during which time the greatest minds took up the challenge. Perhaps we have reached our limit and the rules of the game cannot be relaxed without collapsing the entire enterprise into a Tower of Babel. Whatever the case, the issue is too important to ignore and let science aimlessly become "primitive and muddled."