

Introduction and Motivation

BEFORE YOU START working your way through this book, you may ask yourself—Why analyze data? This is an important, basic question, and it has several compelling answers.

The simplest need for data analysis arises most naturally in disciplines addressing phenomena that are, in all likelihood, inherently nondeterministic (e.g., feelings and psychology or stock market behavior). Since such fields of knowledge are not governed by known fundamental equations, the only way to generalize disparate observations into expanded knowledge is to analyze those observations. In addition, in such fields predictions are entirely dependent on empirical models of the types discussed in chapter 9 that contain parameters not fundamentally constrained by theory. Finding these models' numerical values most suitable for a particular application is another important role of data analysis.

A more general rationale for analyzing data stems from the complementary relationship of empirical and theoretical science and dominates contexts and disciplines in which the studied phenomena have, at least in principle, fully knowable and usable fundamental governing dynamics (see chapter 7). In these contexts, best exemplified by physics, theory and observations both vie for the helm. Indeed, throughout the history of physics, theoretical predictions of yet unobserved phenomena and empirical observations of yet theoretically unexplained ones have alternately fixed physics' ropes.¹ When theory leads, its predictions must be tested against experimental or observational data. When empiricism is at the helm, coherent, reproducible knowledge is systematically and carefully gleaned from noisy, messy observations. At the core of both, of course, is data analysis.

Empiricism's biggest triumph, affording it (ever so fleetingly) the leadership role, arises when novel data analysis-based knowledge—fully acquired and processed—proves at odds with relevant existing theories (i.e., equations previously thought to govern the studied phenomenon fail to explain and reproduce the new observations). In such cases, relatively rare but game changing,

¹ As beautifully described in Feuer, L. S. (1989) *Einstein and the Generations of Science*, 2nd ed., Transaction, 390 pp., ISBN-10: 0878558993, ISBN-13: 978-0878558995, and also, with different emphasis, in Kragh, H. (2002) *Quantum Generations: A History of Physics in the Twentieth Century*, Princeton University Press, Princeton, NJ, 512 pp., ISBN13: 978-0-691-09552-3.

the need for a new theory becomes apparent.² When a new theory emerges, it either generalizes existing ones (rendering previously reigning equations a limiting special case, as in, e.g., Newtonian vs. relativistic gravity), or introduces an entirely new set of equations. In either case, at the root of the progress thus achieved is data analysis.

Once a new theory matures and its equation set becomes complete and closed, one of its uses is model-mediated predictions. In this application of theory, another rationale for data analysis sometimes emerges. It involves phenomena (e.g., fluid turbulence) for which governing equations may exist in principle, but their applications to most realistic situations is impossibly complex and high-dimensional. Such phenomena can thus be reasonably characterized as fundamentally deterministic yet practically stochastic. As such, practical research and modeling of such phenomena fall into the first category above, that addressing inherently nondeterministic phenomena, in which better mechanistic understanding requires better data and better data analysis.

Data analysis is thus essential for scientific progress. But is the level of algebraic rigor characteristic of some of this book's chapters necessary? After all, in some cases we can use some off-the-shelf spreadsheet-type black box for some rudimentary data analysis without any algebraic foundation. How you answer this question is a subjective matter. My view is that while in a few cases some progress can be made without substantial understanding of the underlying algebraic machinery and assumptions, such analyses are inherently dead ends in that they can be neither generalized nor extended beyond the very narrow, specific question they address. To seriously contribute to any of the progress routes described above, in the modular, expandable manner required for your work to potentially serve as the foundation of subsequent analyses, there is no alternative to thorough, deep knowledge of the underlying linear algebra.

² Possibly the most prominent examples of this route (see Feuer's book) are the early development of relativity partly in an effort to explain the Michelson-Morley experiment, and the emergence of quantum mechanics for explaining blackbody radiation observations.