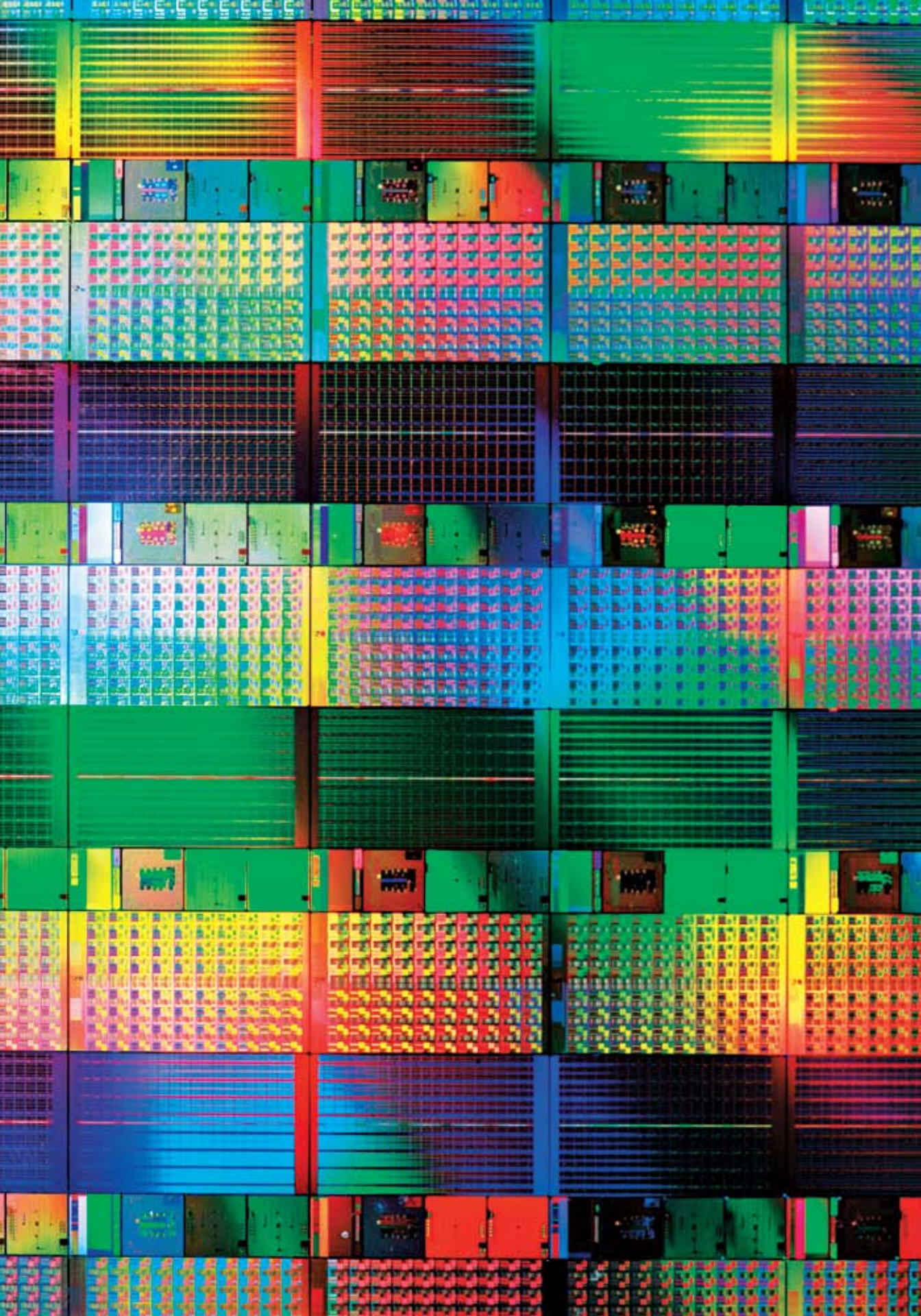




3. SCIENTIFIC INFRASTRUCTURE





Introduction

DARON GREEN | Microsoft Research

WARNING! The articles in Part 3 of this book use a range of dramatic metaphors, such as “explosion,” “tsunami,” and even the “big bang,” to strikingly illustrate how scientific research will be transformed by the ongoing creation and availability of high volumes of scientific data. Although the imagery may vary, these authors share a common intent by addressing how we must adjust our approach to computational science to handle this new proliferation of data. Their choice of words is motivated by the opportunity for research breakthroughs afforded by these large and rich datasets, but it also implies the magnitude of our culture’s loss if our research infrastructure is not up to the task.

Abbott’s perspective across all of scientific research challenges us with a fundamental question: whether, in light of the proliferation of data and its increasing availability, the need for sharing and collaboration, and the changing role of computational science, there should be a “new path for science.” He takes a pragmatic view of how the scientific community will evolve, and he is skeptical about just how eager researchers will be to embrace techniques such as ontologies and other semantic technologies. While avoiding dire portents, Abbott is nonetheless vivid in characterizing a disconnect between the supply of scientific knowledge and the demands of the private and government sectors.



To bring the issues into focus, Southan and Cameron explore the “tsunami” of data growing in the EMBL-Bank database—a nucleotide sequencing information service. Throughout Part 3 of this book, the field of genetic sequencing serves as a reasonable proxy for a number of scientific domains in which the rate of data production is brisk (in this case, a 200% increase per annum), leading to major challenges in data aggregation, workflow, backup, archiving, quality, and retention, to name just a few areas.

Larus and Gannon inject optimism by noting that the data volumes are tractable through the application of multicore technologies—provided, of course, that we can devise the programming models and abstractions to make this technical innovation effective in general-purpose scientific research applications.

Next, we revisit the metaphor of a calamity induced by a data tidal wave as Gannon and Reed discuss how parallelism and the cloud can help with scalability issues for certain classes of computational problems.

From there, we move to the role of computational workflow tools in helping to orchestrate key tasks in managing the data deluge. Goble and De Roure identify the benefits and issues associated with applying computational workflow to scientific research and collaboration. Ultimately, they argue that workflows illustrate primacy of method as a crucial technology in data-centric research.

Fox and Hendler see “semantic eScience” as vital in helping to interpret interrelationships of complex concepts, terms, and data. After explaining the potential benefits of semantic tools in data-centric research, they explore some of the challenges to their smooth adoption. They note the inadequate participation of the scientific community in developing requirements as well as a lack of coherent discussion about the applicability of Web-based semantic technologies to the scientific process.

Next, Hansen et al. provide a lucid description of the hurdles to visualizing large and complex datasets. They wrestle with the familiar topics of workflow, scalability, application performance, provenance, and user interactions, but from a visualization standpoint. They highlight that current analysis and visualization methods lag far behind our ability to create data, and they conclude that multidisciplinary skills are needed to handle diverse issues such as automatic data interpretation, uncertainty, summary visualizations, verification, and validation.

Completing our journey through these perils and opportunities, Parastatidis considers how we can realize a comprehensive knowledge-based research infrastructure for science. He envisions this happening through a confluence of traditional scientific computing tools, Web-based tools, and select semantic methods.