Contents

Foreword xi

Acknowledgments xv

Introduction 1
  Who Should Read This Book? 3
  What’s in This Book? 4
  How to Contact Us 6

Chapter 1 Healthcare, History, and Heartbreak 7
  Top Issues in Healthcare 9
  Data Management 16
  Biosimilars, Drug Pricing, and Pharmaceutical Compounding 18
  Promising Areas of Innovation 19
  Conclusion 25
  Notes 25

Chapter 2 Genome Sequencing: Know Thyself, One Base Pair at a Time 27
  Content contributed by Sheetal Shetty and Jacob Brill
  Challenges of Genomic Analysis 29
  The Language of Life 30
  A Brief History of DNA Sequencing 31
  DNA Sequencing and the Human Genome Project 35
  Select Tools for Genomic Analysis 38
  Conclusion 47
  Notes 48

Chapter 3 Data Management 53
  Content contributed by Joe Arnold
  Bits about Data 54
  Data Types 56
  Data Security and Compliance 59
  Data Storage 66
  SwiftStack 70
  OpenStack Swift Architecture 78
  Conclusion 94
  Notes 94
Chapter 4  Designing a Data-Ready Network Infrastructure  105
Research Networks: A Primer  108
ESnet at 30: Evolving toward Exascale and Raising Expectations  109
Internet2 Innovation Platform  111
Advances in Networking  113
InfiniBand and Microsecond Latency  114
The Future of High-Performance Fabrics  117
Network Function Virtualization  119
Software-Defined Networking  121
OpenDaylight  122
Conclusion  157
Notes  157

Chapter 5  Data-Intensive Compute Infrastructures  163
Content contributed by Dijiang Huang, Yuli Deng, Jay Etchings, Zhiyuan Ma, and Guangchun Luo
Big Data Applications in Health Informatics  166
Sources of Big Data in Health Informatics  168
Infrastructure for Big Data Analytics  171
Fundamental System Properties  186
GPU-Accelerated Computing and Biomedical Informatics  187
Conclusion  190
Notes  191

Chapter 6  Cloud Computing and Emerging Architectures  211
Cloud Basics  213
Challenges Facing Cloud Computing Applications in Biomedicine  215
Hybrid Campus Clouds  216
Research as a Service  217
Federated Access Web Portals  219
Cluster Homogeneity  220
Emerging Architectures (Zeta Architecture)  221
Conclusion  229
Notes  229

Chapter 7  Data Science  235
NoSQL Approaches to Biomedical Data Science  237
Using Splunk for Data Analytics  244
Statistical Analysis of Genomic Data with Hadoop  250
Extracting and Transforming Genomic Data  253
Processing eQTL Data  256
Generating Master SNP Files for Cases and Controls  259
Generating Gene Expression Files for Cases and Controls  260
Cleaning Raw Data Using MapReduce  261
Transpose Data Using Python  263
Statistical Analysis Using Spark  264
Hive Tables with Partitions  268
Conclusion  270
Notes  270