

## 目 次

|       |   |    |
|-------|---|----|
| ①     | ビッグデータとは？ .....                               | 1  |
| 1.1   | 従来のデータ解析とビッグデータ解析の違い                          | 2  |
| 1.2   | ビッグデータ登場の背景                                   | 6  |
| 1.2.1 | 生成されるデータの爆発的な増加                               | 6  |
| 1.2.2 | 分散処理技術・フレームワークの充実                             | 7  |
| 1.2.3 | データベース, 機械学習などの技術的成熟                          | 9  |
| 1.2.4 | クラウドサービスの充実                                   | 11 |
|       | 演習問題  | 11 |
| ②     | ビッグデータ解析の応用事例と情報爆発プロジェクト ...                  | 14 |
| 2.1   | 選挙戦略：(例) 米国大統領選挙                              | 15 |
| 2.1.1 | オバマ前大統領の再選キャンペーン                              | 15 |
| 2.1.2 | トランプ大統領の選挙戦                                   | 16 |
| 2.2   | 都市部の人流予測                                      | 17 |
| 2.3   | 防災・災害時対応                                      | 18 |
| 2.4   | Yahoo! JAPAN ビッグデータレポート                       | 20 |
| 2.5   | 情報爆発プロジェクト                                    | 20 |
| 2.5.1 | 大量の情報から必要な情報を効率良く取り出す<br>「次世代検索技術」            | 24 |
| 2.5.2 | 爆発する情報の受け皿となる<br>「システム基盤技術」                   | 24 |
| 2.5.3 | 「人に優しい情報環境の構築技術」                              | 25 |
| 2.5.4 | 「先進的な IT サービスを人間社会に受け入れやすくする<br>ための社会制度設計の研究」 | 26 |

|       |                         |    |
|-------|-------------------------|----|
| 2.6   | 将来の方向性                  | 26 |
|       | 演習問題                    | 27 |
| ③     | ビッグデータ解析の流れ             | 28 |
| 3.1   | データ収集                   | 30 |
| 3.2   | データ解析                   | 33 |
|       | 演習問題                    | 35 |
| ④     | ビッグデータを支える技術 (1)        |    |
|       | 分散処理フレームワーク             | 36 |
| 4.1   | Apache Hadoop           | 36 |
| 4.1.1 | HDFS                    | 37 |
| 4.1.2 | MapReduce               | 42 |
| 4.1.3 | YARN                    | 45 |
| 4.2   | Spark                   | 48 |
| 4.2.1 | 高速な処理時間                 | 50 |
| 4.2.2 | 複数のプログラミング言語における操作群の提供  | 50 |
| 4.2.3 | 目的に応じたライブラリの提供          | 50 |
| 4.2.4 | 実行環境, データ源の多様性          | 51 |
| 4.3   | Storm                   | 52 |
| 4.3.1 | ストリーム                   | 53 |
| 4.3.2 | Spout                   | 53 |
| 4.3.3 | Bolt                    | 53 |
| 4.3.4 | ストリームのグルーピング            | 54 |
| 4.4   | Apache Mahout と Jubatus | 54 |
| 4.5   | SpatialHadoop           | 56 |
|       | 演習問題                    | 58 |
| ⑤     | ビッグデータを支える技術 (2)        |    |
|       | ストリーム処理エンジン             | 59 |
| 5.1   | ストリーム処理エンジンの概要          | 61 |

|          |                           |     |
|----------|---------------------------|-----|
| 5.2      | CQL                       | 62  |
| 5.2.1    | ストリーム, リレーションの定義          | 63  |
| 5.2.2    | 演算の分類                     | 63  |
| 5.2.3    | ウィンドウ演算                   | 64  |
| 5.2.4    | リレーション-ストリーム演算            | 67  |
| 5.3      | 代表的なストリーム処理エンジン           | 68  |
| 5.3.1    | 集中型                       | 68  |
| 5.3.2    | 分散型                       | 70  |
|          | 演習問題                      | 75  |
| <b>6</b> | <b>ビッグデータを支える技術 (3)</b>   |     |
|          | NoSQL データベース .....        | 76  |
| 6.1      | NoSQL が登場した背景             | 77  |
| 6.2      | トランザクション, ACID 特性と CAP 定理 | 78  |
| 6.3      | NoSQL データベースの分類と特徴        | 80  |
| 6.3.1    | キーバリュー型                   | 82  |
| 6.3.2    | 列指向型                      | 82  |
| 6.3.3    | ドキュメント指向型                 | 83  |
| 6.3.4    | グラフ指向型                    | 83  |
| 6.4      | 代表的な NoSQL データベース         | 84  |
| 6.4.1    | Redis: キーバリュー型            | 84  |
| 6.4.2    | HBase: 列指向型               | 88  |
| 6.4.3    | Cassandra: 列指向型           | 93  |
| 6.4.4    | MongoDB: ドキュメント指向型        | 101 |
| 6.4.5    | Neo4j: グラフ指向型             | 107 |
| 6.5      | まとめ                       | 112 |
|          | 演習問題                      | 113 |

|       |                                       |     |
|-------|---------------------------------------|-----|
| ⑦     | ビッグデータを支える技術 (4)                      |     |
|       | 機械学習, 深層学習 .....                      | 117 |
| 7.1   | 機械学習                                  | 117 |
| 7.1.1 | 機械学習と人工知能の歴史                          | 118 |
| 7.1.2 | 機械学習の基本的な考え                           | 121 |
| 7.1.3 | 機械学習法の分類                              | 122 |
| 7.1.4 | 教師あり学習の代表的な手法                         | 125 |
| 7.1.5 | 教師なし学習の代表的な手法                         | 130 |
| 7.2   | 深層学習                                  | 135 |
| 7.2.1 | 深層学習の歴史と背景                            | 136 |
| 7.2.2 | 畳込みニューラルネットワーク                        | 139 |
| 7.2.3 | 自己符号化                                 | 141 |
| 7.2.4 | 再帰型ニューラルネットワーク                        | 142 |
|       | 演習問題                                  | 146 |
| ⑧     | オープンデータの潮流 .....                      | 148 |
| 8.1   | オープンデータとは?                            | 150 |
| 8.2   | オープンデータの取組み                           | 150 |
| 8.3   | 現状のオープンデータの問題点                        | 154 |
| 8.4   | オープンデータ化に伴う課題                         | 157 |
|       | 演習問題                                  | 160 |
| ⑨     | 今後の展望 .....                           | 161 |
| 9.1   | オープンデータを効率的・効果的に利用するための<br>プラットフォーム構築 | 162 |
| 9.2   | 人に関わる, 人を介したデータ収集・処理                  | 163 |
|       | 参考文献 .....                            | 167 |

|   |     |
|---|-----|
| あとがき .....  | 171 |
| AI 時代を生き抜くにはその震源地ともいえる<br>ビッグデータを正しく理解することが必須<br>(コーディネーター 喜連川 優) ..... | 172 |
| 索 引 .....   | 178 |

### Box

1. 留まることなく発展するクラウドサービス「AWS」 ..... 12
2. Hadoop の柔軟性・高度化と運用・維持容易性 ..... 47
3. Apache Flink ..... 55
4. NoSQL はデータベース的には邪道? ..... 114
5. NoSQL と関係データベースはどうやって選ぶべき? ..... 116
6. 機械学習・人工知能は人間を超えることができるか? ..... 147