Aos meus pais, Francisco de Oliveira Couto e Maria Fernanda dos Santos Moreira Couto.
Preface

During the last decades, I witnessed the growing importance of computer science skills for career advancement in Health and Life Sciences. However, not everyone has the skill, inclination, or time to learn computer programming. The learning process is usually time-consuming and requires constant practice, since software frameworks and programming languages change substantially overtime. This is the main motivation for writing this book about using shell scripting to address common Health and Life data and text processing tasks. Shell scripting has the advantages of being: i) nowadays available in almost all personal computers; ii) almost immutable for more than four decades; iii) relatively easy to learn as a sequence of independent commands; iv) an incremental and direct way to solve many of the data problems that Health and Life professionals face.

During the last decades, I had the pleasure to teach introductory computer science classes to Health and Life Sciences undergraduates. I used programming languages, such as Perl and Python, to address data and text processing tasks, but I always felt to loose a substantial amount of the time teaching the technicalities of these languages, which will probably change over time and are uninteresting for the majority of the students that do not intend to pursue advanced bioinformatics courses. Thus the purpose of this book is to motivate and help specialists to automate common data and text processing tasks after a short learning period. If they become interested (and I hope some do), the book presents pointers to where they can acquire more advanced computer science skills.

This book does not intend to be a comprehensive compendium of shell scripting commands, but instead a introductory guide for Health and Life specialists. This book introduces the commands as they are required to automate data and text processing tasks. The selected tasks have a strong focus on text mining and biomedical ontologies given my research experience and their growing relevance for Health and Life studies. Nevertheless, the same type of solutions presented in the book are also applicable to many other research fields and data sources.

Lisboa, January 2019

Francisco Couto
Acknowledgments

I am grateful to all the people who helped and encouraged me along this journey, specially to Rita Ferreira for all the insightful discussions about shell scripting.

I am also grateful for all the suggestions and corrections given by my colleague Prof. José Baptista Coelho, and by my college students: Alice Veiros, Ana Ferreira, Carlota Silva, Catarina Raimundo, Daniela Matias, Inês Justo, João Andrade, João Leitão, João Pedro Pais, Konil Solanki, Mariana Custódio, Marta Cunha, Manuel Fialho, Miguel Silva, Rafaela Marques, Raquel Chora and Sofia Morais.

This work was supported by FCT through funding of DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017 (http://dest.rd.cienclas.ulisboa.pt/), and LASIGE Research Unit, ref. UID/CEC/00408/2019.
# Acronyms

| Acronym | Description |
|---------|-------------|
| ChEBI   | Chemical Entities of Biological Interest |
| CSV     | Comma-Separated Values |
| cURL    | Client Uniform Resource Locator |
| DAG     | Directed Acyclic Graph |
| DBMS    | Database Management System |
| DiShIn  | Semantic Similarity Measures using Disjunctive Shared Information |
| DO      | Disease Ontology |
| EBI     | European Bioinformatics Institute |
| GO      | Gene Ontology |
| HTTP    | Hypertext Transfer Protocol |
| HTTPS   | HTTP Secure |
| ICD     | International Classification of Diseases |
| MER     | Minimal Named-Entity Recognizer |
| MeSH    | Medical Subject Headings |
| NCBI    | National Center for Biotechnology Information |
| NER     | Named-Entity Recognition |
| OBO     | Open Biological and Biomedical Ontology |
| OWL     | Web Ontology Language |
| PMC     | PubMed Central |
| RDFS    | RDF Schema |
| SNOMED CT | Systematized Nomenclature of Medicine - Clinical Terms |
| SQL     | Structured Query Language |
| TSV     | Tab-Separated Values |
| UMLS    | Unified Medical Language System |
| UniProt | Universal Protein Resource |
| URI     | Uniform Resource Identifier |
| URL     | Uniform Resource Locator |
| XLS     | Microsoft Excel file format |
| XML     | Extensible Markup Language |
| XPath   | XML Path Language |
References

Allen and Owens, 2011. Allen, G. and Owens, M. (2011). *The Definitive Guide to SQLite*. Books for professionals by professionals. Apress.

Angermueller et al., 2016. Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, 12(7):878.

Aramaki et al., 2011. Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.

Aras et al., 2014. Aras, H., Hackl-Sommer, R., Schwantner, M., and Sofean, M. (2014). Applications and challenges of text mining with patents. In *IPuMin*.

Ashburner et al., 2000. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Baker and Milligan, 2014. Baker, J. and Milligan, I. (2014). Counting and mining research data with Unix. Technical report, The Editorial Board of the Programming Historian.

Barros and Couto, 2016. Barros, M. and Couto, F. M. (2016). Knowledge representation and management: a linked data perspective. *Yearbook of medical informatics*, 25(01):178–183.

Blumenthal and Tavenner, 2010. Blumenthal, D. and Tavenner, M. (2010). The “meaningful use” regulation for electronic health records. *New England Journal of Medicine*, 363(6):501–504.

Borst and Borst, 1997. Borst, W. and Borst, W. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, University of Twente, Netherlands.

Campos et al., 2018. Campos, L., Pedro, V., and Couto, F. M. (2018). Impact of translation on named-entity recognition in radiology texts. *Database*, 2018.

Canese, 2006. Canese, K. (2006). Pubmed celebrates its 10th anniversary. *NLM Tech Bull*, 352:e5.

Ching et al., 2018. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387.

Cock et al., 2009. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Cook et al., 2017. Cook, C. E., Bergman, M. T., Cochrane, G., Apweiler, R., and Birney, E. (2017). The european bioinformatics institute in 2017: data coordination and integration. *Nucleic acids research*, 46(D1):D21–D29.

Couto and Lamurias, 2018. Couto, F. and Lamurias, A. (2018). MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(58).

Couto and Lamurias, 2019. Couto, F. and Lamurias, A. (2019). Semantic similarity definition. In Ranganathan, S., Nakai, K., Schönbach, C., and Grishkov, M., editors, *Encyclopaedia of Bioinformatics and Computational Biology*, volume 1. Oxford: Elsevier.

Couto et al., 2017. Couto, F. M., Campos, L. F., and Lamurias, A. (2017). MER: a minimal named-entity recognition tagger and annotation server. *Proc BioCreative*, 5:130–7.

Couto et al., 2006. Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., and Rebholz-Schuhmann, D. (2006). GOAnnotator: linking protein go annotations to evidence text. *Journal of biomedical discovery and collaboration*, 1(1):19.
References

Leonelli, 2016. Leonelli, S. (2016). Data-Centric Biology: A Philosophical Study. University of Chicago Press.

Lesk, 2014. Lesk, A. (2014). Introduction to bioinformatics. Oxford University Press.

Li et al., 2015. Li, W., Cawley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y. M., Buso, N., and Lopez, R. (2015). The embl-ebi bioinformatics web and programmatic tools framework. Nucleic acids research, 43(W1):W580–W584.

Lin et al., 1998. Lin, D. et al. (1998). An information-theoretic definition of similarity. In lncf, volume 98, pages 296–304. Citeseer.

Lu, 2011. Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. Database, 2011.

McGuinness et al., 2004. McGuinness, D. L., Van Harmelen, F., et al. (2004). OWL web ontology language overview. W3C recommendation, 10(10):2004.

Nosek et al., 2015. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., et al. (2015). Promoting an open research culture. Science, 348(6242):1422–1425.

Ong et al., 2016. Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., and He, Y. (2016). Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic acids research, 45(D1):D347–D352.

Rawat and Meena, 2014. Rawat, S. and Meena, S. (2014). Publish or perish: Where are we heading? Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, 19(2):87.

Rebholz-Schuhmann et al., 2005. Rebholz-Schuhmann, D., Kirsch, H., and Couto, F. (2005). Facts from text—is text mining ready to deliver? PLoS biology, 3(2):e65.

Resnik, 1995. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1, pages 448–453. Morgan Kaufmann Publishers Inc.

Richardson and Ruby, 2008. Richardson, L. and Ruby, S. (2008). RESTful web services. " O'Reilly Media, Inc."

Ritchie, 1971. Ritchie, D. M. (1971). Unix programmer’s manual. Technical report, Tech. report. Bell.

Robinson and Bauer, 2011. Robinson, P. N. and Bauer, S. (2011). Introduction to bio-ontologies. Chapman and Hall/CRC.

Schriml et al., 2018. Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2018). Human disease ontology 2018 update: classification, content and workflow expansion. Nucleic acids research.

Schuemie et al., 2004. Schuemie, M. J., Weeber, M., Schijvenaars, B. J., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. Bioinformatics, 20(16):2597–2604.

Shah et al., 2003. Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords? BMC bioinformatics, 4(1):20.

Shotts Jr, 2012. Shotts Jr, W. E. (2012). The Linux command line: a complete introduction. No Starch Press.

Singhal, 2012. Singhal, A. (2012). Introducing the knowledge graph: things, not strings. Official google blog, 5.

Smith et al., 2007. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11):1251.

Spasic et al., 2005. Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. Briefings in bioinformatics, 6(3):239–251.
Acronyms

Stajich et al., 2002. Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., et al. (2002). The bioperl toolkit: Perl modules for the life sciences. *Genome research*, 12(10):1611–1618.

Stephens et al., 2015. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195.

Studer et al., 1998. Studer, R., Benjamins, V. R., Fensel, D., et al. (1998). Knowledge engineering: principles and methods. *Data and knowledge engineering*, 25(1):161–198.

Styler IV et al., 2014. Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143.

Tomczak et al., 2018. Tomczak, A., Mortensen, J. M., Winnenburg, R., Liu, C., Alessi, D. T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N. H., et al. (2018). Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. *Scientific reports*, 8(1):5115.

Wei et al., 2013. Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522.

Wu and Fung, 1994. Wu, D. and Fung, P. (1994). Improving chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proc. of the 4th Conference on Applied Natural Language Processing*.

Yeh et al., 2003. Yeh, A., Hirschman, L., and Morgan, A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics*, 19(1):i331–i339.