Knowledge-based Data Science

Lawrence E. Hunter

Computational manipulation of knowledge is an important, and often underappreciated, aspect of biomedical Data Science. The first Data Science initiative from the US National Institutes of Health was entitled "Big Data to Knowledge (BD2K)." The main emphasis of the more than $200M allocated to that program has been on "Big Data;" the "Knowledge" component has largely been the implicit assumption that the work will lead to new biomedical knowledge. However, there is long-standing and highly productive work in *computational* knowledge representation and reasoning, and computational processing of knowledge has a role in the world of Data Science.

Knowledge-based biomedical Data Science involves the design and implementation of computer systems that *act as if they knew* about biomedicine. There are many ways in which a computational approach might act as if it knew something: for example, it might be able to answer a natural language question about a biomedical topic, or pass an exam; it might be able to use existing biomedical knowledge to rank or evaluate hypotheses; it might explain or interpret data in light of prior knowledge, either in a Bayesian or other sort of framework. After a brief survey of existing approaches to knowledge-based data science, this position paper argues that such research is ripe for expansion, and expanded application.

1 Representations of Biomedical Knowledge

All computational approaches to knowledge require specification of how the computer system *represents* knowledge internally, and how it might compute with those representations to produce outputs (often called, perhaps metaphorically, *reasoning*). Classic descriptions of knowledge representation and reasoning systems, e.g. (Davis, Shrobe, & Szolovits, 1993) focus on what ontological commitments a knowledge representation makes, what inferences are possible with it, and, sometimes, which of those inferences can be made efficiently. These issues remain useful in thinking about how knowledge representation and reasoning play a role in today's data science environment. Much more contemporary work also addresses these issues.

As (Davis et al., 1993) pointed out, knowledge representations entail ontological commitments. Adoption of existing ontologies, rather than creating idiosyncratic or single-use ontologies provides significant advantages for reproducibility in scientific research, for inter-operability, and in avoiding pitfalls in the modeling of knowledge. A great deal of work has been done in biomedical ontology (e.g. (T. G. O. Consortium, 2017),(Ong et al., 2017),(Sharp, 2017),(Smedley, 2017),(Bandrowski et al., 2016) and many others), and these increasingly mature ontological resources form an important basis for knowledge-based data science. Community-curated ontologies (such as those meeting the Open Biomedical Ontologies (OBO) Foundry criteria (Smith et al., 2007)) capture a consensus view of the entities and processes involved in biology, medicine and biomedical research, analogous to how nomenclature committees systematize naming conventions. While not meeting all of the criteria of the OBO Foundry, terminological resources such as UMLS (Lindberg, 1990), Snomed-CT (Bhattacharyya, 2015) and the NCI thesaurus (Fragoso, Coronado, Haber, Hartel, & Wright, 2004) have also been used to provide useful pseudo-ontological foundations for knowledge representations.

While ontologies identify the basic elements from which a knowledge representation is constructed, they are agnostic about the mechanisms by which ontological units are assembled into representations of knowledge. Building on decades of work in artificial intelligence research, the W3C produced an international standard for assembling ontological entities into assertions and managing collections of assertions, together referred to as the Semantic Web. The focus of the Semantic Web standard is to make it possible to link web elements with shared meaning, and is sometimes described as the *Linked Data* paradigm. The Semantic Web builds on the standard Resource Description Framework (RDF), which provides a way to link three uniform resource identifiers (URIs) to specify a pair of entities and a relationship between them (forming an RDF "triple"). Collections of triples form a graph, and a computational mechanism for managing such collections is called a triple store. The Semantic Web standards also define RDF Schemas (RDFS) and a Web Ontology Language (OWL) which facilitate richer knowledge representations, SPARQL, which provides a query language for interrogating RDF graphs or triple stores, and the Simple Knowledge Organization System (SKOS), which provides a basic ontology, including simple semantic relationships. While the Semantic Web standards are intended to be general representation tools for all knowledge, the combination of Semantic Web standards and biomedical ontologies are the basis of most current biomedical knowledge representation systems.

2 Knowledge-based inference

Representations of knowledge are sterile without use. Although human visualization of computationally represented knowledge (e.g. (Lohmann, Negru, Haag, & Ertl, 2014)) can be useful, the primary use of computationally represented knowledge is inference. There are many forms of inference, and thousands of publications describing computational methods of reasoning. Although too broad to survey here, a brief introduction to the types of knowledge-based inference common in biomedical applications gives some idea of its potential.

2.1 Logical inference

Computational logical inference is a mapping from a base set of assertions to create additional assertions that are entailed by the base. While deductive reasoning is the classic form of logical inference, it is, in general, computationally intractable. Various restricted forms of deductive inference, such as those based on description logics, have better computational performance, at the cost of greatly restricting the utility of the inferences. Description logics, for example, are limited to inferring subsumption relationships based on necessary and sufficient class definitions. Contemporary applications of description logic inference in biomedical knowledge representation have been successful primarily in checking for modeling errors (e.g. (Bodenreider, Smith, Kumar, & Burgun, 2007), (Jansen, Kim, Coenen, Saba, & Hardiker, 2016)), although some other applications have been attempted (e.g. (Boeker, França, Bronsert, & Schulz, 2016; Hochheiser, Castine, Harris, Savova, & Jacobson, 2016; Holford & Krauthammer, 2015)).

Deductive retrieval is a special case of deductive inference, where the inference is to compute whether a set of logical axioms and base assertions can be combined to satisfy a query; the programming language Prolog and the W3C standard for the Semantic Web Rule Language (SWRL) are examples of approaches to deductive retrieval. Triple stores extended with deductive retrieval are much more valuable than those that can retrieve only queries that match exactly. Several knowledge-bases of biomedicine based on these technologies have been developed (e.g. (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008; Willighagen et al., 2013; Barros & Couto, 2016; Livingston, Bada, Baumgartner, & Hunter, 2015), and their uses extend beyond deductive retrieval alone.

2.2 Inference from ontology annotation

In addition to the creation of biomedical ontologies, a great deal of effort has gone into annotating genes and other biological entities to ontological categories. Gene Ontology annotations of genes and gene products

figure prominently in major databases such as Uniprot and the Mouse Genome Informatics database(U. Consortium, 2017; Blake et al., 2017). These annotations provide a quick summary of knowledge about gene function, subcellular localization and biological processes. By far the most common application of computational representations of knowledge to problems in biomedicine is enrichment analysis, see e.g. .(Soldatos, Perdigão, Brown, Sabir, & O'Donoghue, 2015; Tipney & Hunter, 2010; Huang, Sherman, & Lempicki, 2009). Enrichment analysis generates hypotheses about the concerted functions of collections of genes by testing for annotations that occur more frequently in the collection than would be expected by chance. Ontology annotation directly supports other sorts of knowledge-based inference as well. For example, phenotype annotations play a major role in mapping between human disease and animal models (e.g. (Mungall et al., 2017, 2015; Kibbe et al., 2015). Formal representations of metabolic pathways (eg. (Keseler et al., 2017; Fabregat et al., 2016)) have been used to analyze metabolomic data and support metabolic engineering.

2.3 Inference from the biomedical literature

Despite the rapid growth of databases with ontological annotation, the main and by far the largest repository of biomedical knowledge remains the published literature. An important domain of knowledge-based data science involves natural language processing with the goal of producing computational representations of the knowledge in the literature. The most basic of these approaches involves tagging passages in the literature with ontological terms (e.g. EuroPMC's SciLite annotations, or (Funk et al., 2014)). Computational methods to identify semantically well-defined entities in the literature support further analysis that identifies links both among different documents in the literature (e.g. (Zheng et al., 2015)) and between entities in the literature and database entries about them (e.g. (Pafilis et al., 2009)). More ambitious literature mining goals involve producing more complex knowledge representations directly by processing natural language documents, e.g. (Xia, Fang, & Zhang, 2014; Cohen et al., 2011), although significant improvements in performance are likely to be necessary before the results of such processing find widespread use in biomedical research. Text mining approaches applied to clinical records and social media, e.g. for pharmacovigilance applications, have also made significant strides recently (Demner-Fushman & Elhadad, 2016). The best performing text mining systems themselves often use representations of prior knowledge to drive understanding of text.

Natural language processing systems have also been used to support automated question answering. Perhaps the most well known of these efforts is IBM's Watson system (Chen, Elenee, & Weber, 2016), which has found significant biomedical application. Many other computational systems for question answering, targeted to biomedical researchers and clinicians, have been fielded, e.g. as reviewed in (Athenikos & Han, 2010; Bauer & Berleant, 2012). Computational approaches to building systems that can answer biomedical exam questions have also been developed, e.g. (Clark et al., 2016).

2.4 Hypothesis generation, evaluation and modification

Perhaps the oldest method of computing with knowledge is Bayesian inference (Heckerman, 1998). By providing a quantitative framework for the idea that observations consistent with prior knowledge are more likely than ones that contradict it, Bayesian reasoning has provided a framework for knowledge-based computation long before computation was automated. Contemporary computers provide the power necessary to support more elaborate Bayesian inference, including model selection as well as estimating model parameters (Chipman et al., 2001).

Network-based inference, such as link prediction or community finding, have been successfully applied to generate significant biomedical hypotheses. Systems that compute over representations of knowledge of biomedicine have been used to propose as yet unobserved relationships among biological entities, e.g. for drugs (Lu, Guo, & Korhonen, 2017), microRNAs (Zeng, Zhang, Liao, & Pan, 2016), diseases (Suthram et al., 2010) and proteins (Tripathi, Moutari, Dehmer, & Emmert-Streib, 2016); some of these predictions have

been empirically validated, e.g. (Jahchan et al., 2013).

Perhaps the most exciting potential for knowledge-based computational systems is in the development and refinement of mechanistic explanations of biomedical phenomena. The vast scope and rapid evolution of the biomedical literature, combined with the breakdown of disciplinary boundaries driven by genome-scale research has made it increasingly difficult for researchers to effectively assimilate all the knowledge potentially relevant to interpreting the results of their own experiments. Although most computational approaches aim to provide material for the *Results* section of a paper, a few are beginning to target the *Discussion* section as well. While no knowledge-based computer system has repeatedly generated important biomedical hypotheses *de novo*, promising proof-of-concept systems include systems to generate hypotheses from the literature (Smalheiser, Torvik, & Zhou, 2009) and those aimed at hypothesis generation or refinement from data (Racunas, Shah, Albert, & Fedoroff, 2004; Callahan, Dumontier, & Shah, 2011), as well as mixed initiative human-computer hypothesis generation (Leach et al., 2009). Although it remains aspirational, the synthesis of computational simulation with knowledge-based generation and refinement of hypotheses has received substantial interest from funding agencies (You, 2015).

3 Open Challenges in Knowledge-based Data Science

As is clear from the NIH BD2K experience, computation over knowledge is a less widespread research focus than analysis of big data, and to date has had less impact in biomedicine. Certain applications, such as enrichment analysis and link prediction, have found widespread use in biomedical research. Text mining systems are increasingly deployed in areas such as helping clinicians keep up with rapidly changing clinical data ("Bringing Precision Medicine to Community Oncologists.", 2017) and pharmacovigilance. However, there are significant challenges to realizing the potential for knowledge-based data science. Perhaps the foremost among these is the knowledge acquisition bottleneck: human curation, even for the relatively simple task of annotation of genes to gene ontology terms is difficult to scale (Baumgartner, Cohen, Fox, Acquaah-Mensah, & Hunter, 2007). Alternatives to manual curation, including applications of text mining and machine learning, have shown promise, but have yet to scale to the entire biomedical literature. Another important understudied question is how to represent what is *not* known: any scientist can describe gaps, ambiguities and uncertainties in existing knowledge, yet there are few computational methods capable of representing, let alone reasoning about, such ignorance.

Even more challenging than developing representations of what is already known is the application of that knowledge to the pressing problems of biomedical research. Existing inference methods are far short of the range and creativity of human experts in developing potential explanations, generating significant hypotheses, and generally interpreting results in light of previous knowledge. Many promising inference methods scale poorly, and are constrained in their ability to harness large knowledge-bases by the extremely large computational loads involved. Even deductive retrieval systems can be computationally intractable over large knowledge-bases; more complex forms of inference hit the limits of current hardware with even smaller knowledge-bases. The Semantic Web standard was developed largely with description logic inference in mind; while it provides a solid foundation for knowledge representation systems, representational transformations may improve the efficiency of other sorts of inference.

Perhaps the biggest challenges in knowledge-based data science are in developing the vision for what such a system could effectively contribute to biomedical research. Is it possible to build computational systems that bring to bear disparate yet relevant facts from across all biomedical disciplines and scales, exploiting their ability to process far more information than any individual human being? Could such a system make sound judgements ranking alternative hypotheses based on an exhaustive comprehension of the literature? Is it possible for computational systems to generate significant and novel mechanistic and pathomechanistic hypotheses about open questions in biomedicine? It is positive answers to questions like these that will drive knowledge-based data science into the mainstream of biomedical research.

References

Athenikos, S., & Han, H. (2010, Jul). Biomedical question answering: a survey. *Comput Methods Programs Biomed*, *99*, 1-24.

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M., Bug, B., Chibucos, M., et al. (2016, Apr). The Ontology for Biomedical Investigations. *PLoS One*, *11*, e0154556.

Barros, M., & Couto, F. (2016, Nov). Knowledge Representation and Management: a Linked Data Perspective. *Yearb Med Inform*, 178-183.

Bauer, M., & Berleant, D. (2012, Sep). Usability survey of biomedical question answering systems. *Hum Genomics*, *6*, 17.

Baumgartner, W. J., Cohen, K., Fox, L., Acquaah-Mensah, G., & Hunter, L. (2007, Jul). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, *23*, i41-8.

Belleau, F., Nolin, M., Tourigny, N., Rigault, P., & Morissette, J. (2008, Oct). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*, *41*, 706-16.

Bhattacharyya, S. B. (2015, dec). Overview of SNOMED CT. In *Introduction to SNOMED CT* (pp. 1–2). Springer Nature.

Blake, J., Eppig, J., Kadin, J., Richardson, J., Smith, C., & Bult, C. (2017, Jan). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res*, *45*, D723-D729.

Bodenreider, O., Smith, B., Kumar, A., & Burgun, A. (2007, mar). Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *Artificial Intelligence in Medicine*, *39*(3), 183–195.

Boeker, M., França, F., Bronsert, P., & Schulz, S. (2016, Nov). TNM-O: ontology support for staging of malignant tumours. *J Biomed Semantics*, *7*, 64.

Bringing Precision Medicine to Community Oncologists. (2017, Jan). *Cancer Discov*, *7*, 6-7.

Callahan, A., Dumontier, M., & Shah, N. (2011, May). HyQue: evaluating hypotheses using Semantic Web technologies. *J Biomed Semantics*, *2 Suppl 2*, S3.

Chen, Y., Elenee, A. J., & Weber, G. (2016, Apr). IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin Ther*, *38*, 688-701.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., & Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, 65–134.

Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P. D., et al. (2016). Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions. In *Proceedings of the thirtieth AAAI conference on artificial intelligence february 12-17, 2016, phoenix, arizona, USA.* (pp. 2580–2586).

Cohen, K., Verspoor, K., Johnson, H., Roeder, C., Ogren, P., Baumgartner, W. J., et al. (2011, Nov). HIGH-PRECISION BIOLOGICAL EVENT EXTRACTION: EFFECTS OF SYSTEM AND OF DATA. *Comput Intell*, *27*, 681-701.

Consortium, T. G. O. (2017, Jan). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*, *45*, D331-D338.

Consortium, U. (2017, Jan). UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, *45*, D158-D169.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine*, *14*(1), 17-33.

Demner-Fushman, D., & Elhadad, N. (2016, Nov). Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. *Yearb Med Inform*, 224-233.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., et al. (2016, Jan). The Reactome pathway Knowledgebase. *Nucleic Acids Res*, *44*, D481-7.

Fragoso, G., Coronado, S. de, Haber, M., Hartel, F., & Wright, L. (2004). Overview and Utilization of the NCI Thesaurus. *Comparative and Functional Genomics*, *5*(8), 648–654.

Funk, C., Baumgartner, W. J., Garcia, B., Roeder, C., Bada, M., Cohen, K., et al. (2014, Feb). Large-scale

biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, *15*, 59.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In *Learning in graphical models* (pp. 301–354). Springer.

Hochheiser, H., Castine, M., Harris, D., Savova, G., & Jacobson, R. (2016, Sep). An information model for computable cancer phenotypes. *BMC Med Inform Decis Mak*, *16*, 121.

Holford, M., & Krauthammer, M. (2015, Dec). Mutadelic: mutation analysis using description logic inferencing capabilities. *Bioinformatics*, *31*, 3742-7.

Huang, d. W., Sherman, B., & Lempicki, R. (2009, Jan). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, *37*, 1-13.

Jahchan, N., Dudley, J., Mazur, P., Flores, N., Yang, D., Palmerton, A., et al. (2013, Dec). A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov*, *3*, 1364-77.

Jansen, K., Kim, T., Coenen, A., Saba, V., & Hardiker, N. (2016). Harmonising Nursing Terminologies Using a Conceptual Framework. *Stud Health Technol Inform*, *225*, 471-5.

Keseler, I., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., et al. (2017, Jan). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res*, *45*, D543-D550.

Kibbe, W., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., et al. (2015, Jan). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*, *43*, D1071-8.

Leach, S., Tipney, H., Feng, W., Baumgartner, W., Kasliwal, P., Schuyler, R., et al. (2009, Mar). Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol*, *5*, e1000215.

Lindberg, C. (1990, May). The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc*, *61*, 40-2.

Livingston, K., Bada, M., Baumgartner, W. J., & Hunter, L. (2015, Apr). KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*, *16*, 126.

Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2014). VOWL 2: User-Oriented Visualization of Ontologies. In *Lecture notes in computer science* (pp. 266–281). Springer Nature.

Lu, Y., Guo, Y., & Korhonen, A. (2017, Jan). Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics*, *18*, 39.

Mungall, C., McMurry, J., Köhler, S., Balhoff, J., Borromeo, C., Brush, M., et al. (2017, Jan). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*, *45*, D712-D722.

Mungall, C., Washington, N., Nguyen-Xuan, J., Condit, C., Smedley, D., Köhler, S., et al. (2015, Oct). Use of model organism and disease databases to support matchmaking for human disease gene discovery. *Hum Mutat*, *36*, 979-84.

Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., et al. (2017, Jan). Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res*, *45*, D347-D352.

Pafilis, E., O'Donoghue, S. I., Jensen, L. J., Horn, H., Kuhn, M., Brown, N. P., et al. (2009, jun). Reflect: augmented browsing for the life scientist. *Nature Biotechnology*, *27*(6), 508–510.

Racunas, S., Shah, N., Albert, I., & Fedoroff, N. (2004, Aug). HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, *20 Suppl 1*, i257-64.

Sharp, M. (2017, Jan). Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *J Biomed Semantics*, *8*, 2.

Smalheiser, N., Torvik, V., & Zhou, W. (2009, May). Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput Methods Programs Biomed*, *94*, 190-7.

Smedley, D. (2017, feb). *Faculty of 1000 evaluation for The human phenotype ontology: semantic unification of common and rare disease.* Faculty of 1000 Ltd.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007, Nov). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, *25*, 1251-5.

Soldatos, T., Perdigão, N., Brown, N., Sabir, K., & O'Donoghue, S. (2015, Mar). How to learn about gene function: text-mining or ontologies? *Methods*, *74*, 3-15.

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., & Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*, *6*(2), e1000662.

Tipney, H., & Hunter, L. (2010, Feb). An introduction to effective use of enrichment analysis software. *Hum Genomics*, *4*, 202-6.

Tripathi, S., Moutari, S., Dehmer, M., & Emmert-Streib, F. (2016, Mar). Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC Bioinformatics*, *17*, 129.

Willighagen, E., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A., Tkachenko, V., et al. (2013, May). The ChEMBL database as linked open data. *J Cheminform*, *5*, 23.

Xia, J., Fang, A., & Zhang, X. (2014). A novel feature selection strategy for enhanced biomedical event extraction using the Turku system. *Biomed Res Int*, *2014*, 205239.

You, J. (2015, Jan). Artificial intelligence. DARPA sets out to automate research. *Science*, *347*, 465.

Zeng, X., Zhang, X., Liao, Y., & Pan, L. (2016, Nov). Prediction and validation of association between microRNAs and diseases by multipath methods. *Biochim Biophys Acta*, *1860*, 2735-9.

Zheng, J., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., et al. (2015). Entity linking for biomedical literature. *BMC Med Inform Decis Mak*, *15 Suppl 1*, S4.