

# Constructing Knowledge Graphs and Their Biomedical Applications

*This manuscript ([permalink](#)) was automatically generated from [greenelab/knowledge-graph-review@d1ae256](#) on September 3, 2019.*

## Authors

---

- **David Nicholson**

 [0000-0003-0002-5761](#) ·  [danich1](#)

Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552 and T32 HG000046

- **Jane Roe**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [janeroe](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

## Abstract

---

1. Give high level description of review as it pertains to knowledge graphs (creation and application)

## Introduction

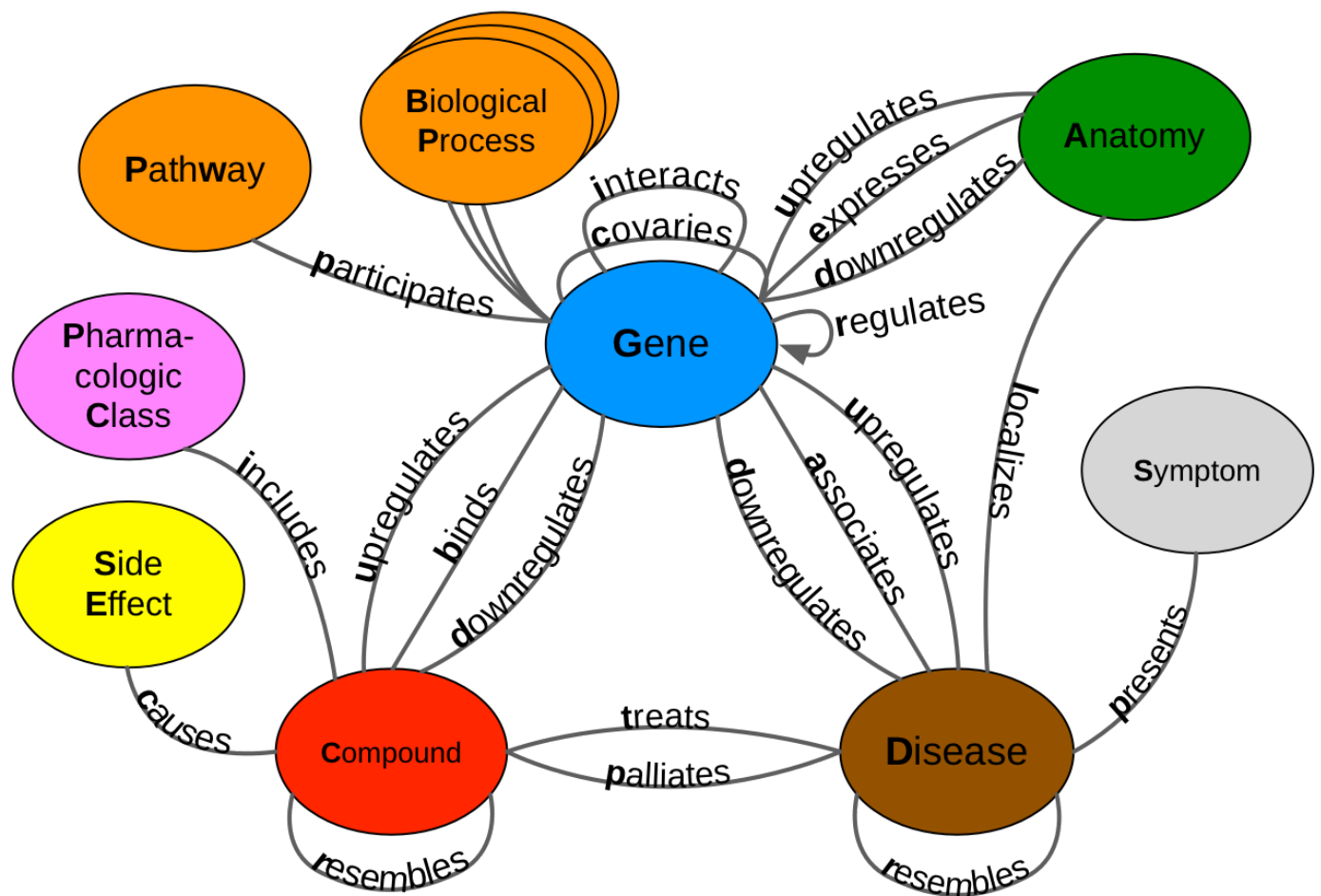
---

Knowledge graphs are a practical resource for many real world applications. They have been used in social medial mining to classify nodes [1] or to create a recommendation system [2]. Knowledge graphs have also been used to understand natural language via interpreting simple questions and using relational information to provide answers [3,4]. In a biomedical setting these graphs have been used to prioritize genes relevant to disease [5,6,7,8], perform drug repurposing [9] and identify drug-target interactions [10].

Despite their utility, precisely defining a knowledge graph is a difficult task because there are multiple conflicting definitions [11]. For this review, we define a knowledge graph as the following: a resource that integrates single or multiple sources of information into the form of a graph. This graph allows for the capacity to make semantic interpretation, continuously incorporate new information and uncover novel hidden knowledge through computational techniques and algorithms. Based on this definition resources like Hetionet [9] would be considered a knowledge graph. Hetionet integrates multiple sources of information into the form of a graph (example shown in Figure 1) and was used to derive novel information concerning unique drug treatments [9]. We do not consider databases like DISEASES [12] and DrugBank [13] to be knowledge graphs. These resources contain essential information, but do not represent their data in graph form.

Knowledge graphs are often constructed from manually curated databases [9,14,15,16]. These sources provide previously established information that can be incorporated into a graph. For example, a graph using DISEASES [12] as a resource would have genes and diseases as nodes, while edges would be added between nodes that have an association. This example shows a single type of relationship; however, there are graphs that use databases with multiple relationships. Other approaches have used natural language processing techniques to build knowledge graphs [17,18]. One example used a text mining system to extract sentences that indicated a protein interacting with another protein [19]. Once these sentences have been identified, they are incorporated as evidence for establishing edges in a knowledge graph.

In this review we describe various approaches for constructing and applying knowledge graphs in a biomedical setting. We discuss the pros and cons of constructing a knowledge graph via manually curated databases and via text mining systems. We also compare assorted approaches for applying knowledge graphs to solve biomedical problems. Lastly, we conclude on the practicality of knowledge graphs and point out future applications that have yet to be explored.



**Figure 1:** A metagraph (schema) of the heterogeneous network used in the Rephetio project [9]. This undirected network depicts pharmacological and biomedical information. The nodes (circles) represent entities and edges (lines) depict relational information between two entities.

## Building Biomedical Knowledge Graphs

1. Set up the context for relationship extraction
  1. Define the problem
  2. Talk about the importance of the problem (filling knowledge bases -> point researchers to relevant papers)
2. Give overview towards taxonomy of approaches (hand written rules, unsupervised machine learning, supervised machine learning etc.)

## Constructing Databases and Manual Curation

1. Talk about papers that construct knowledge graphs without text mining approaches
2. Discuss the positives and negatives for these methods

## Text Mining for Relationship Extraction

### Rule-Based Natural Language Processing

1. Mention papers on hand written rules and expressions

### Unsupervised Machine Learning

1. Mention Clustering Analysis
2. Mention Co-Occurrence approaches

## Supervised Machine Learning

1. Mention the availability of publically available data
  1. PPI - 5 datasets
    1. 10.1016/j.artmed.2004.07.016
    2. 10.1186/1471-2105-8-50
    3. Learning language in logic - genic interaction extraction challenge
    4. 10.1093/bioinformatics/btl616
    5. <http://helix-web.stanford.edu/psb02/ding.pdf>
  2. DaG - 3 datasets
    1. 10.1016/j.jbi.2012.04.004
    2. 10.1186/s12859-015-0472-9
    3. 10.1186/1471-2105-14-323
    4. 10.1186/1471-2105-13-161
  3. CiD
  4. 10.1093/database/baw068
  5. CbG
  6. Biocreative VI track 5 - raw citation
  7. more if exists talk about deep learning methods
2. Mention the use of Support Vector Machines and other non deep learning classifiers
  1. Will have to mention that field has moved to deep learning.
  2. 10.1186/s13326-017-0168-3
  3. 10.1371/journal.pcbi.1004630
3. Mention deep learning methods
  1. 1901.06103v1
  2. 10.1016/j.knosys.2018.11.020
  3. 10.1177/0165551516673485
  4. 1706.01556v2
  5. ^^ A few papers here but a lot more will be put into place
  6. Mention caveat which is the need for large annotated datasets
  7. Mention a direction the field is moving to which is weak supervision and more that info that will come in time.

## Applying Knowledge Graphs to Biomedical Challenges

---

1. Mention that these graphs can be used for discovery
2. Mention representation learning (aka representing a graph as dense vectors for nodes and/or edges)
- 3.

## Unifying Techniques

1. Set up the problem that maps a knowledge graph into a low dimensional space

## Matrix Factorization

1. Mention techniques for these with some papers

## Deep Learning

1. Define node neighborhoods
2. Talk about random walks
3. Talk about auto encoders random walk independent approaches

## Unifying Applications

1. Mention how the previous section is used in a biomedical setting

### **Disease and Gene Interactions**

1. Mention disease gene prioritization
2. Mention Disease gene associations

### **Protein Protein Interactions**

1. Mention predicting genes interacting genes

### **Drug Interactions**

1. Talk about drug side effects
2. Drug repurposing
3. Drug-Disease Interactions

### **Clinical applications**

1. Can mention EHR use and other related applications
2. Mention Tiffany's work on private data embeddings

### **Conclusion**

---

1. Summarize discussed positives and pitfalls
2. Leave some open ended questions yet to be explored
3. Will come into play as I write this review paper

## References

---

### 1. Node Classification in Social Networks

Smriti Bhagat, Graham Cormode, S. Muthukrishnan  
*arXiv* (2011-01-17) <https://arxiv.org/abs/1101.3291v1>  
DOI: [10.1007/978-1-4419-8462-3\\_5](https://doi.org/10.1007/978-1-4419-8462-3_5)

### 2. Network Embedding Based Recommendation Method in Social Networks

Yufei Wen, Lei Guo, Zhumin Chen, Jun Ma  
*Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (2018)  
<https://doi.org/gf6rtt>  
DOI: [10.1145/3184558.3186904](https://doi.org/10.1145/3184558.3186904)

### 3. Open Question Answering with Weakly Supervised Embedding Models

Antoine Bordes, Jason Weston, Nicolas Usunier  
*arXiv* (2014-04-16) <https://arxiv.org/abs/1404.4326v1>

### 4. Neural Network-based Question Answering over Knowledge Graphs on Word and Character Level

Denis Lukovnikov, Asja Fischer, Jens Lehmann, Sören Auer  
*Proceedings of the 26th International Conference on World Wide Web - WWW '17* (2017)  
<https://doi.org/gfv8hp>  
DOI: [10.1145/3038912.3052675](https://doi.org/10.1145/3038912.3052675)

### 5. Towards integrative gene prioritization in Alzheimer's disease.

Jang H Lee, Graciela H Gonzalez  
*Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2011)  
<https://www.ncbi.nlm.nih.gov/pubmed/21121028>  
PMID: [21121028](https://pubmed.ncbi.nlm.nih.gov/21121028/)

### 6. PhenoGeneRanker: A Tool for Gene Prioritization Using Complete Multiplex Heterogeneous Networks

Cagatay Dursun, Naoki Shimoyama, Mary Shimoyama, Michael Schläppi, Serdar Bozdog  
*Cold Spring Harbor Laboratory* (2019-05-27) <https://doi.org/gf6rtr>  
DOI: [10.1101/651000](https://doi.org/10.1101/651000)

### 7. Biological Random Walks: Integrating heterogeneous data in disease gene prioritization

Michele Gentili, Leonardo Martini, Manuela Petti, Lorenzo Farina, Luca Becchetti  
*2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (2019-07) <https://doi.org/gf6rts>  
DOI: [10.1109/cibcb.2019.8791472](https://doi.org/10.1109/cibcb.2019.8791472)

### 8. Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes

Mona Alshahrani, Robert Hoehndorf  
*Bioinformatics* (2018-09-01) <https://doi.org/gd9k8n>  
DOI: [10.1093/bioinformatics/bty559](https://doi.org/10.1093/bioinformatics/bty559) · PMID: [30423077](https://pubmed.ncbi.nlm.nih.gov/30423077/) · PMCID: [PMC6129260](https://pubmed.ncbi.nlm.nih.gov/PMC6129260/)

### 9. Systematic integration of biomedical knowledge prioritizes drugs for repurposing

Daniel Scott Himmelstein, Antoine Lizée, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini

eLife (2017-09-22) <https://doi.org/cdfk>  
DOI: [10.7554/elife.26726](https://doi.org/10.7554/elife.26726) · PMID: [28936969](https://pubmed.ncbi.nlm.nih.gov/28936969/) · PMCID: [PMC5640425](https://pubmed.ncbi.nlm.nih.gov/PMC5640425/)

## 10. Assessing Drug Target Association Using Semantic Linked Data

Bin Chen, Ying Ding, David J. Wild

*PLoS Computational Biology* (2012-07-05) <https://doi.org/rn6>

DOI: [10.1371/journal.pcbi.1002574](https://doi.org/10.1371/journal.pcbi.1002574) · PMID: [22859915](https://pubmed.ncbi.nlm.nih.gov/22859915/) · PMCID: [PMC3390390](https://pubmed.ncbi.nlm.nih.gov/PMC3390390/)

## 11. Towards a definition of knowledge graphs

Lisa Ehrlinger, Wolfram Wöß

*SEMANTiCS* (2016)

## 12. DISEASES: Text mining and data integration of disease-gene associations

Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafo, Janos X. Binder, Lars Juhl Jensen

*Methods* (2015-03) <https://doi.org/f3mn6s>

DOI: [10.1016/j.ymeth.2014.11.020](https://doi.org/10.1016/j.ymeth.2014.11.020) · PMID: [25484339](https://pubmed.ncbi.nlm.nih.gov/25484339/)

## 13. DrugBank 5.0: a major update to the DrugBank database for 2018

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, ... Michael Wilson

*Nucleic Acids Research* (2017-11-08) <https://doi.org/gcwtzk>

DOI: [10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037) · PMID: [29126136](https://pubmed.ncbi.nlm.nih.gov/29126136/) · PMCID: [PMC5753335](https://pubmed.ncbi.nlm.nih.gov/PMC5753335/)

## 14. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information

Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, Jianyang Zeng

*Nature Communications* (2017-09-18) <https://doi.org/gbxwrc>

DOI: [10.1038/s41467-017-00680-8](https://doi.org/10.1038/s41467-017-00680-8) · PMID: [28924171](https://pubmed.ncbi.nlm.nih.gov/28924171/) · PMCID: [PMC5603535](https://pubmed.ncbi.nlm.nih.gov/PMC5603535/)

## 15. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks

Hui Liu, Yinglong Song, Jihong Guan, Libo Luo, Ziheng Zhuang

*BMC Bioinformatics* (2016-12) <https://doi.org/gf6v27>

DOI: [10.1186/s12859-016-1336-7](https://doi.org/10.1186/s12859-016-1336-7) · PMID: [28155639](https://pubmed.ncbi.nlm.nih.gov/28155639/) · PMCID: [PMC5259862](https://pubmed.ncbi.nlm.nih.gov/PMC5259862/)

## 16. Finding disease similarity based on implicit semantic similarity

Sachin Mathur, Deendayal Dinakarpandian

*Journal of Biomedical Informatics* (2012-04) <https://doi.org/b7b3tw>

DOI: [10.1016/j.jbi.2011.11.017](https://doi.org/10.1016/j.jbi.2011.11.017) · PMID: [22166490](https://pubmed.ncbi.nlm.nih.gov/22166490/)

## 17. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences

Patrick Ernst, Amy Siu, Gerhard Weikum

*BMC Bioinformatics* (2015-05-14) <https://doi.org/gb8w8d>

DOI: [10.1186/s12859-015-0549-5](https://doi.org/10.1186/s12859-015-0549-5) · PMID: [25971816](https://pubmed.ncbi.nlm.nih.gov/25971816/) · PMCID: [PMC4448285](https://pubmed.ncbi.nlm.nih.gov/PMC4448285/)

## 18. Constructing biomedical domain-specific knowledge graph with minimum supervision

Jianbo Yuan, Zhiwei Jin, Han Guo, Hongxia Jin, Xianchao Zhang, Tristram Smith, Jiebo Luo

*Knowledge and Information Systems* (2019-03-23) <https://doi.org/gf6v26>

DOI: [10.1007/s10115-019-01351-4](https://doi.org/10.1007/s10115-019-01351-4)

**19. Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein-protein interaction**

Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, Pushpak Bhattacharyya

*Knowledge-Based Systems* (2019-02) <https://doi.org/gf4788>

DOI: [10.1016/j.knosys.2018.11.020](https://doi.org/10.1016/j.knosys.2018.11.020)