Recent Advances on Graph Analytics and Its Applications in Healthcare

Knowledge Graph Construction and Inference

KDD 2020 Tutorial

August 23, morning

Fei Wang, Peng Cui, Jian Pei, ***Yangqiu Song***, Chengxi Zang,

http://www.calvinzang.com/kdd2020_tutorial_medical_graph_analytics.html

Outline

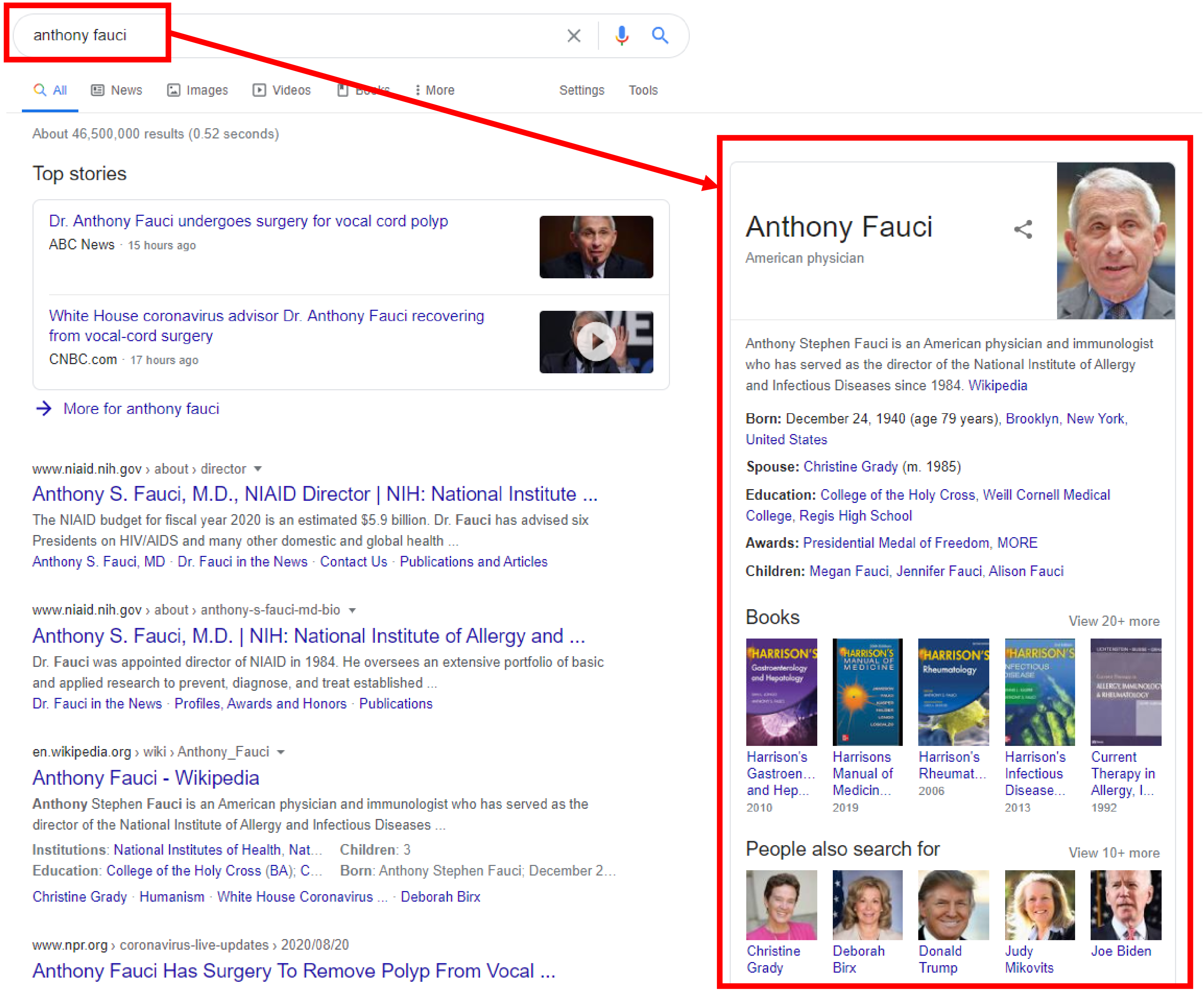
- Introduction
  - General knowledge graphs
  - Knowledge graphs in healthcare
- Knowledge Graph Construction
- Knowledge Graph Inference

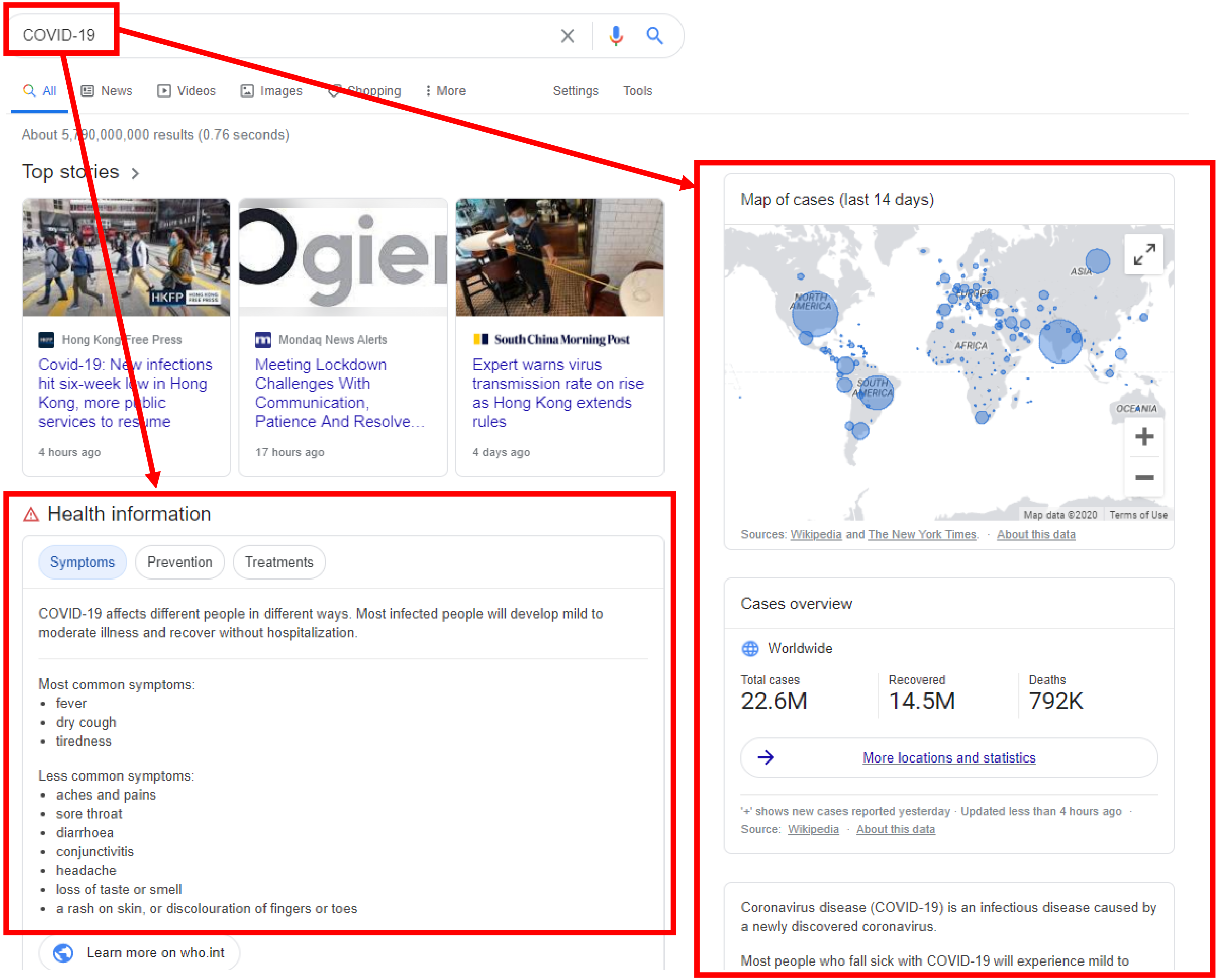
What's a Knowledge Graph?

- A knowledge graph has many names in the history
  - Semantic networks, knowledge base, ontology, …

- In 2012, Google released its project "Google Knowledge Graph"
  - A graph-based knowledge representation connecting real-world entities to support search
    - Landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art and more
  - Get information instantly relevant to a query

https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html


anthony fauci
All
News
Images
Videos
Books
More
Settings
Tools
About 46,500,000 results (0.52 seconds)
Top stories
Dr. Anthony Fauci undergoes surgery for vocal cord polyp
ABC News · 15 hours ago
White House coronavirus advisor Dr. Anthony Fauci recovering from vocal-cord surgery
CNBC.com · 17 hours ago
More for anthony fauci
www.niaid.nih.gov › about › director
Anthony S. Fauci, M.D., NIAID Director | NIH: National Institute ...
The NIAID budget for fiscal year 2020 is an estimated $5.9 billion. Dr. Fauci has advised six Presidents on HIV/AIDS and many other domestic and global health ...
Anthony S. Fauci, MD · Dr. Fauci in the News · Contact Us · Publications and Articles
www.niaid.nih.gov › about › anthony-s-fauci-md-bio
Anthony S. Fauci, M.D. | NIH: National Institute of Allergy and ...
Dr. Fauci was appointed director of NIAID in 1984. He oversees an extensive portfolio of basic and applied research to prevent, diagnose, and treat established ...
Dr. Fauci in the News · Profiles, Awards and Honors · Publications
en.wikipedia.org › wiki › Anthony_Fauci
Anthony Fauci - Wikipedia
Anthony Stephen Fauci is an American physician and immunologist who has served as the director of the National Institute of Allergy and Infectious Diseases ...
Institutions: National Institutes of Health, Nat...
Children: 3
Education: College of the Holy Cross (BA); C...
Born: Anthony Stephen Fauci; December 2...
Christine Grady · Humanism · White House Coronavirus ... · Deborah Birx
www.npr.org › coronavirus-live-updates › 2020/08/20
Anthony Fauci Has Surgery To Remove Polyp From Vocal ...
Anthony Fauci
American physician
Anthony Stephen Fauci is an American physician and immunologist who has served as the director of the National Institute of Allergy and Infectious Diseases since 1984. Wikipedia
Born: December 24, 1940 (age 79 years), Brooklyn, New York, United States
Spouse: Christine Grady (m. 1985)
Education: College of the Holy Cross, Weill Cornell Medical College, Regis High School
Awards: Presidential Medal of Freedom, MORE
Children: Megan Fauci, Jennifer Fauci, Alison Fauci
Books
View 20+ more
Harrison's Gastroen... and Hep... 2010
Harrisons Manual of Medicin... 2019
Harrison's Rheumat... 2006
Harrison's Infectious Disease... 2013
Current Therapy in Allergy, I... 1992
People also search for
View 10+ more
Christine Grady
Deborah Birx
Donald Trump
Judy Mikovits
Joe Biden


COVID-19
All
News
Videos
Images
Shopping
More
Settings
Tools
About 5,790,000,000 results (0.76 seconds)
Top stories
Hong Kong Free Press
Covid-19: New infections hit six-week low in Hong Kong, more public services to resume
4 hours ago
Mondaq News Alerts
Meeting Lockdown Challenges With Communication, Patience And Resolve...
17 hours ago
South China Morning Post
Expert warns virus transmission rate on rise as Hong Kong extends rules
4 days ago
Health information
Symptoms
Prevention
Treatments
COVID-19 affects different people in different ways. Most infected people will develop mild to moderate illness and recover without hospitalization.
Most common symptoms:
fever
dry cough
tiredness
Less common symptoms:
aches and pains
sore throat
diarrhoea
conjunctivitis
headache
loss of taste or smell
a rash on skin, or discolouration of fingers or toes
Learn more on who.int
Map of cases (last 14 days)
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
OCEANIA
Map data ©2020
Terms of Use
Sources: Wikipedia and The New York Times. · About this data
Cases overview
Worldwide
Total cases
22.6M
Recovered
14.5M
Deaths
792K
More locations and statistics
'+' shows new cases reported yesterday · Updated less than 4 hours ago ·
Source: Wikipedia · About this data
Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.
Most people who fall sick with COVID-19 will experience mild to

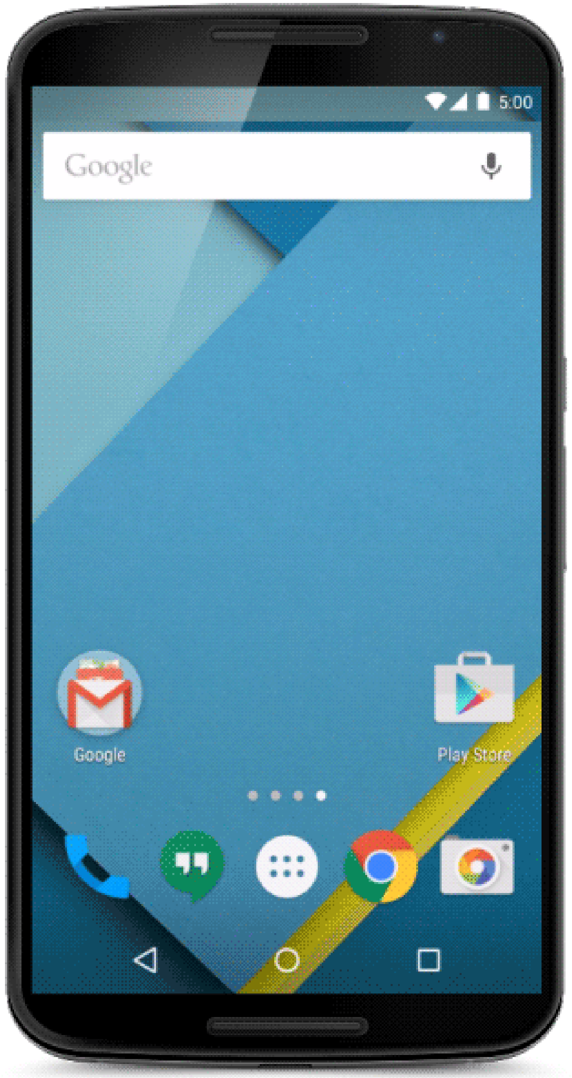
State-of-the-art Enterprise-level KGs

	Data Model	Size of Nodes	Size of Edges	Development Stage
Google	Strongly typed entities, relations with domain and range inference	~1 Billion	~70 Billions	Actively used in products
Microsoft	The types of entities, relations, and attributes in the graph are defined in an ontology	~2 Billions	~55 Billions	Actively used in products
Facebook	All of the attributes and relations are structured and strongly typed, and optionally indexed to enable efficient retrieval, search, and traversal.	~50 Millions	~500 Millions	Actively used in products
eBay	Entities and relation, well-structured and strongly typed	~100 Million	~1 Billion	Early stages of development and deployment
IBM	Entities and relations with evidence information associated with them	~100 Millions	~5 Billions	Actively used in products and by clients

Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, Jamie Taylor: Industry-scale knowledge graphs: lessons and challenges. Commun. ACM 62(8): 36-43 (2019)

Google Knowledge Graph in Healthcare

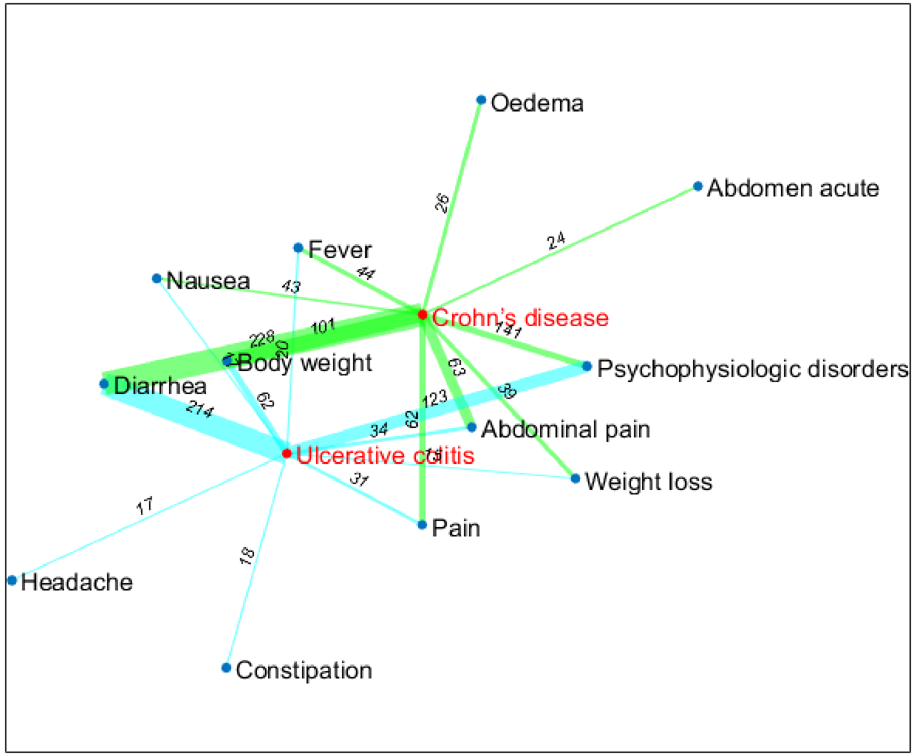
- Knowledge graph is useful for clinical decision support systems and self-diagnostic symptom checkers
  - For example, major symptoms of a heart attack are pain or discomfort chest, arms or shoulder, jaw, neck, or back, feeling weak, lightheaded or faint and shortness of breath
- Answers to common medical questions for searches related to health conditions based on
  - Aggregated information from search results
  - Multi-faceted medical facts
    - Google worked with a "team of doctors" to "carefully compile, curate, and review this information"
    - "All of the gathered facts represent real-life clinical knowledge from these doctors and high-quality medical sources across the web, and the information has been checked by medical doctors at Google and the Mayo Clinic for accuracy."



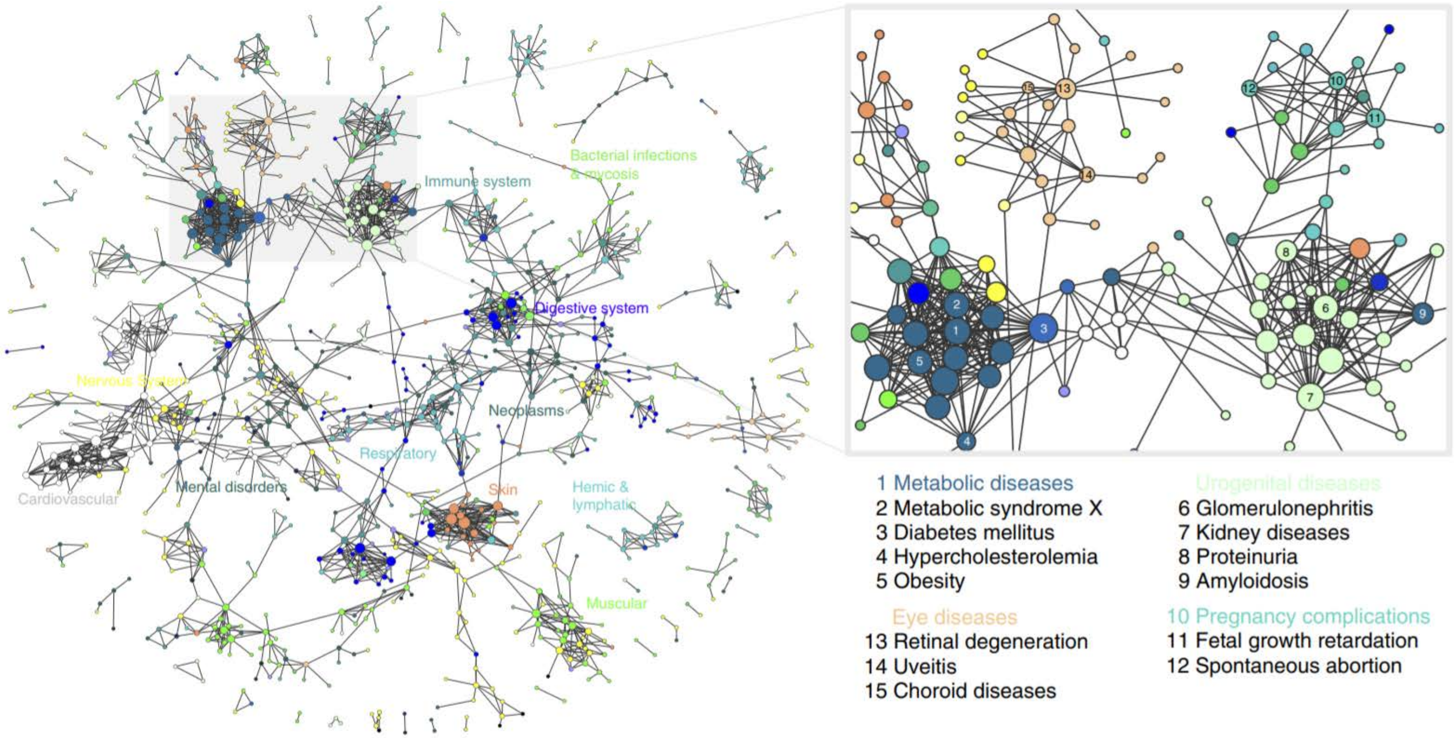
https://techcrunch.com/2015/03/11/google-knowledge-graph-meets-healthcare/
Figure from: https://googleblog.blogspot.com/2015/02/health-info-knowledge-graph.html

Example of Knowledge Graphs in Healthcare

Symptom and disease linking from biomedical literature database: PubMed



A symptom-disease network extracted from PubMed

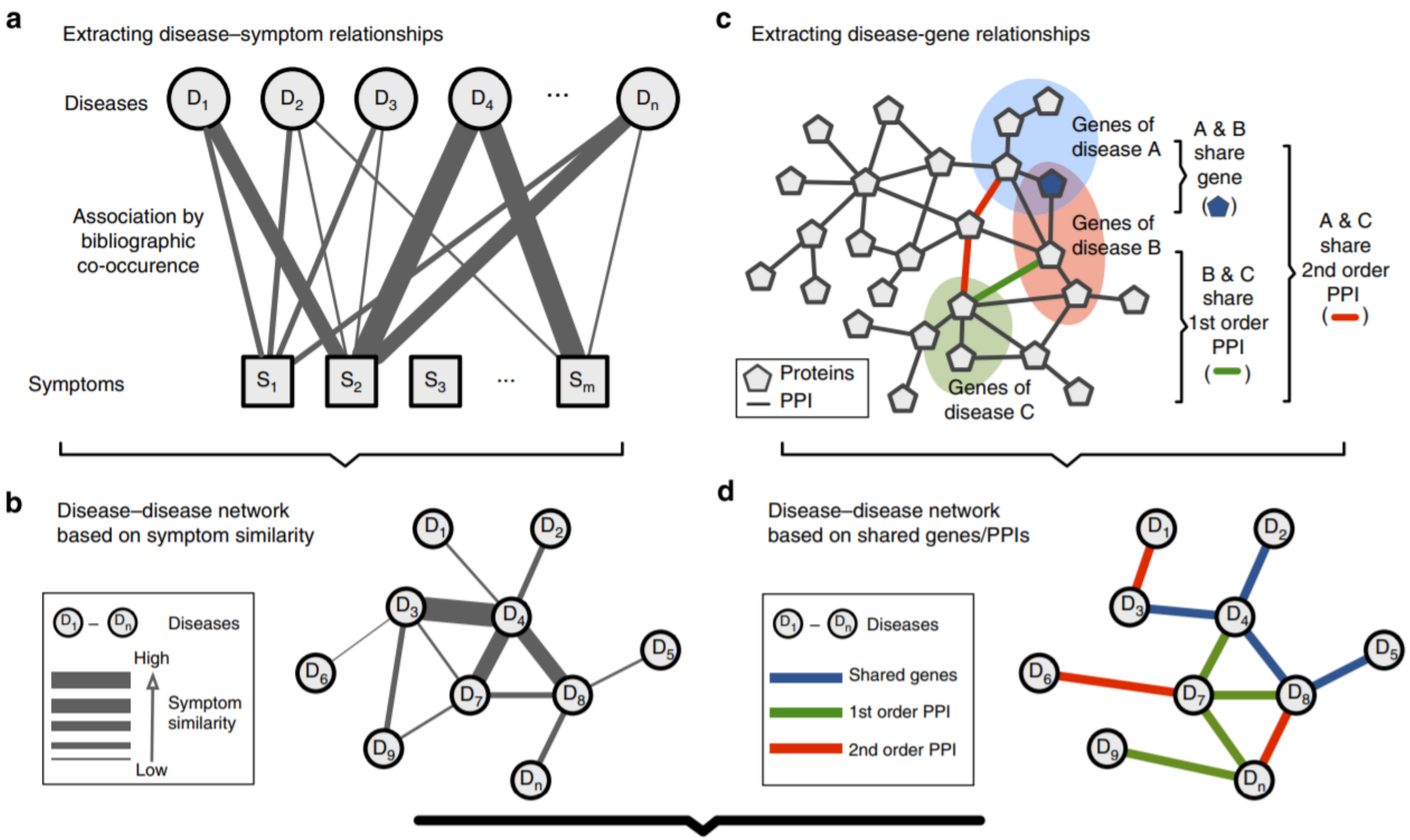


A disease network shows similarities based on symptoms reveals disease clusters

Data and figure from: XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma. Human symptoms–disease network. Nature Communications. 2014

Example of Knowledge Graphs in Healthcare

- The disease-symptom network can be further integrated to analyze disease similarities
  - A disease network where link weight between two diseases quantifies the similarity of their respective symptoms
  - A disease network where disease–gene association and protein–protein interaction (PPI) are used



- Shared symptoms indicate shared genes between diseases
- Shared symptoms indicate shared protein interactions.

Figure from: XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma. Human symptoms–disease network. Nature Communications. 2014

Example of Knowledge Graphs in Healthcare

Symptom and disease linking from large-scale EHR data

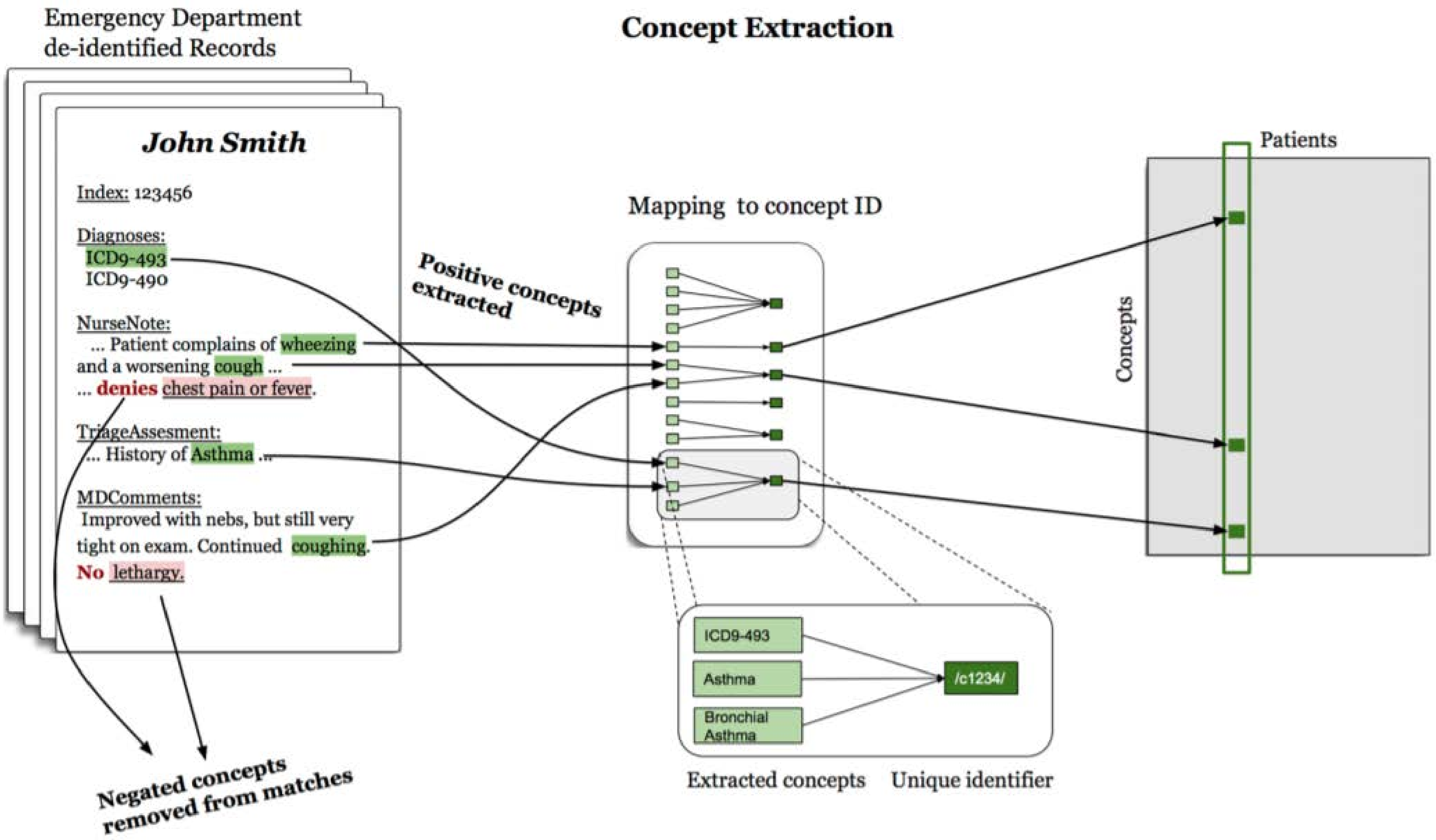
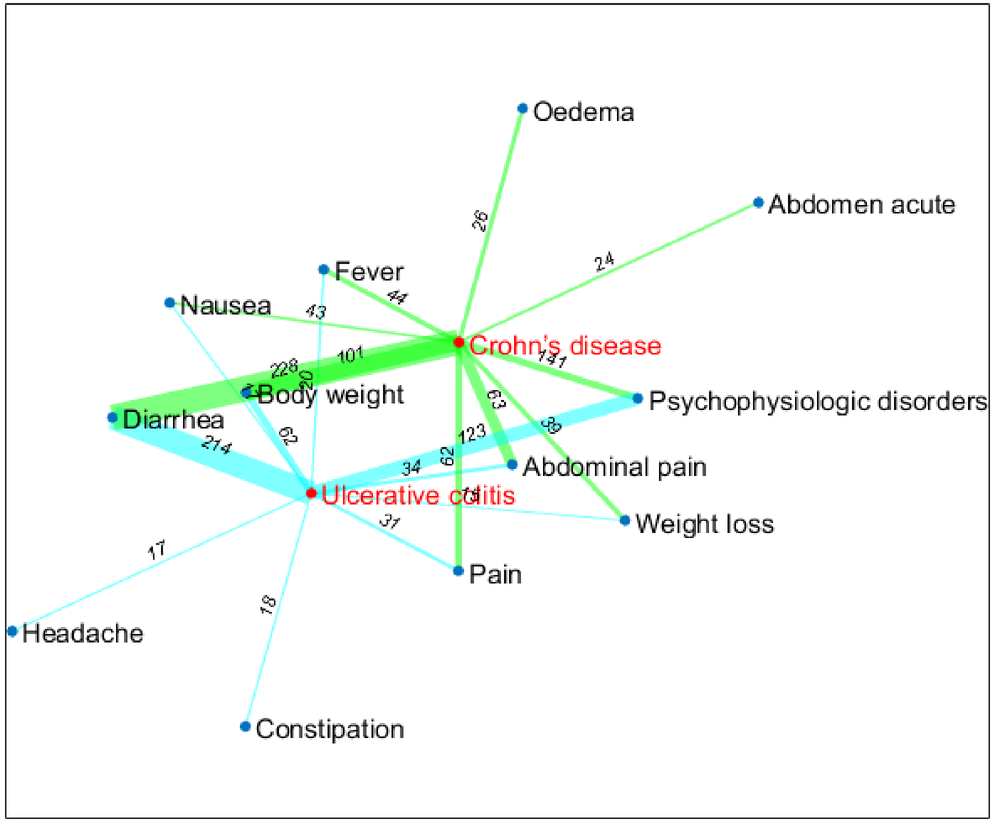


Figure from: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports. 2017

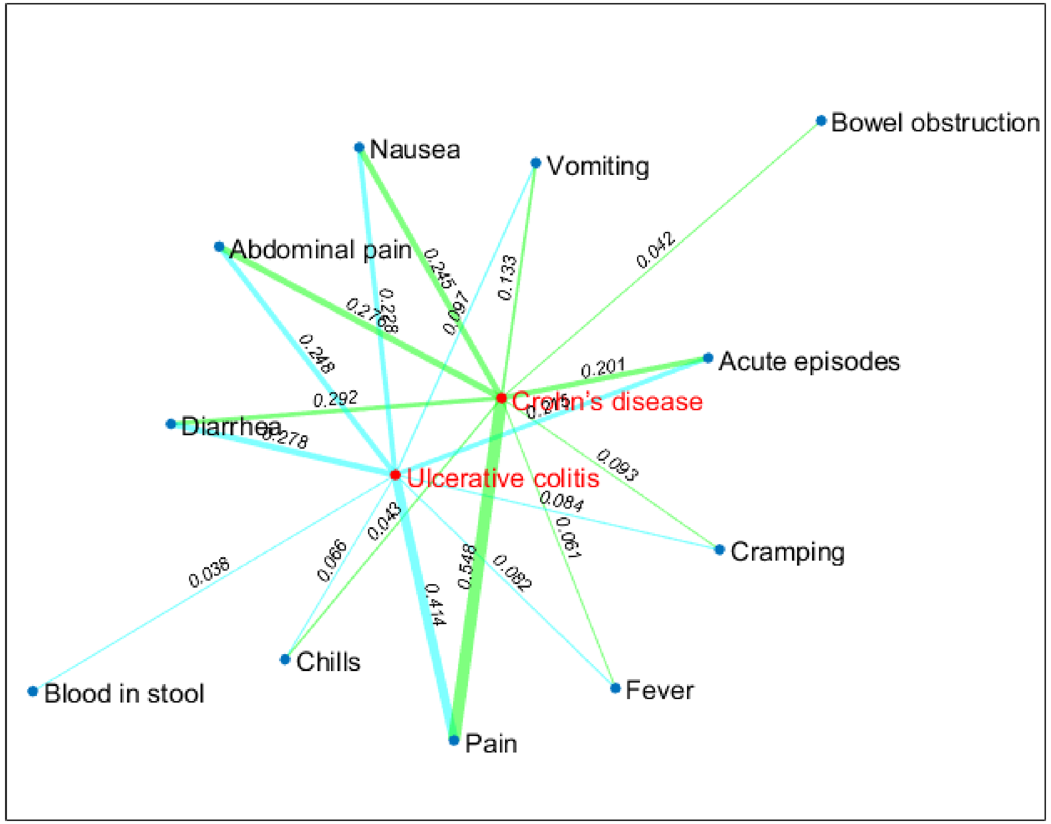
Comparison Between PubMed and EHR

PubMed: more declarative, more formal language and regular texts, more typical symptoms to promote learning



Data from: XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma. Human symptoms–disease network. Nature Communications. 2014

EHR: more statistical, noisier text but recording the practical medicine use



Data from: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports. 2017

Quality of Such Kind of Knowledge Graphs

- Precision-recall curve rated according to physicians' expert opinion
- Google Health Knowledge Graph has two tags "always" and "frequent"

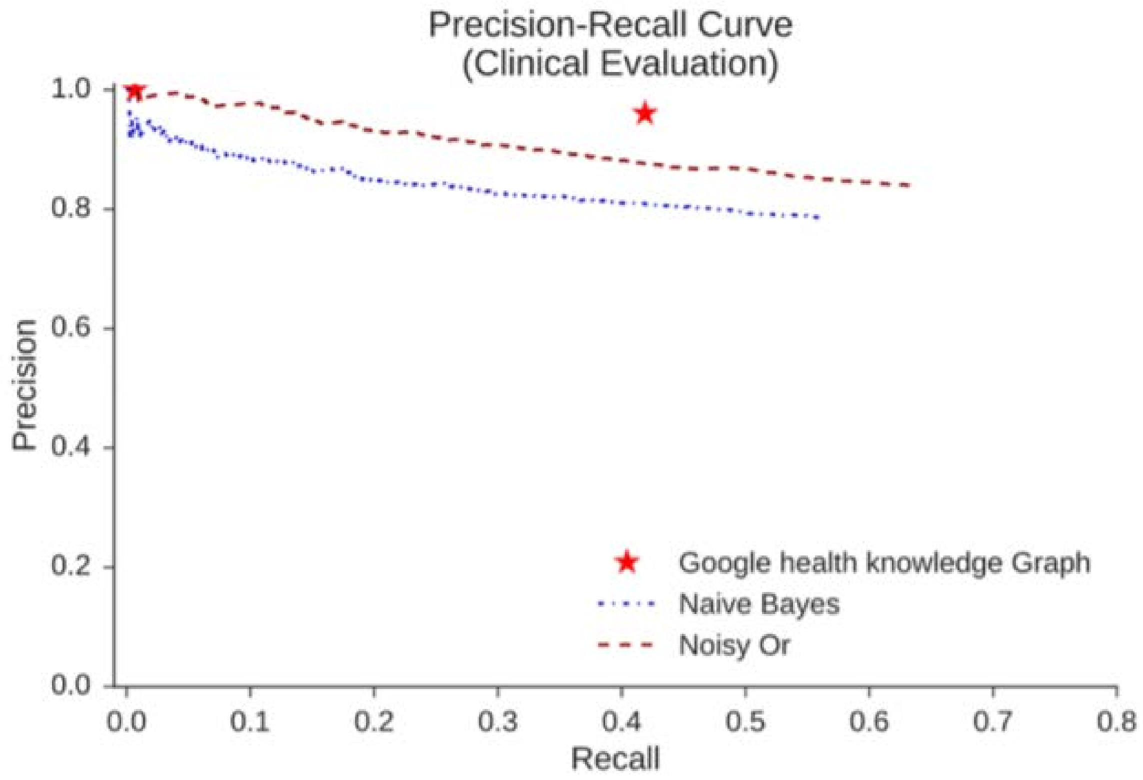


Figure from: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports. 2017

Information Extraction and Text Mining

- Extracting information from PubMed/HER is useful, however, there is a lot of variability and ambiguity of language and terminology use. For example,
  - Amyotrophic lateral sclerosis, motor neurone disease, and Lou Gehrig's Disease refer to the same disease
  - According to Medical Subject Headings (MeSH), obesity belongs to
    - Nutritional and Metabolic Diseases
    - Diagnosis
    - Physiological Phenomena
    - Pathological Conditions, Signs and Symptoms
- There are many (~200) existing expert annotated knowledge bases in Unified Medical Language System (UMLS)
  - 127 semantic types organized as a hierarchy
  - 3.2 million unique concepts
    - A primary name and a set of aliases
    - 8% are linked to more than one types

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. 32 (Database issue):D267–D270, 2004.

Why Knowledge Graphs in Healthcare?

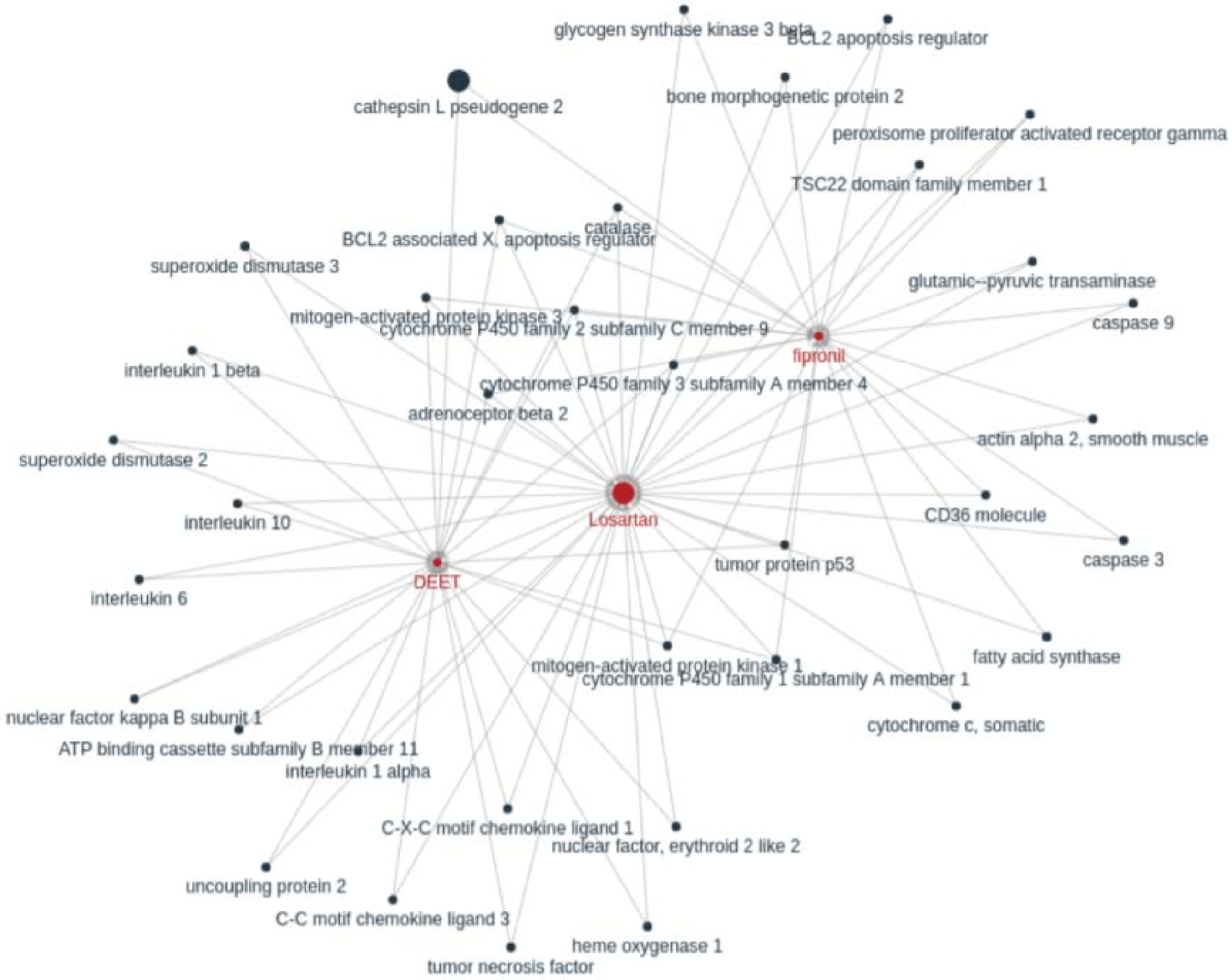
- Medical doctors are overwhelmed by the huge amount of data, e.g.,
  - Electronic Health Records (EHRs)
  - Research articles from PubMed
  - Providing complementary information to existing knowledge bases

- We need an efficient tool to “connect the dots”
  - Humans can only process a few objects/variables at a time
  - Human’s summarization of concepts can be vague
  - Types of concepts are heterogeneous, e.g., patient, disease, symptom, gene, chemical, etc.
  - Demands to reason multi-hop relations
  - Traditional logic inferences tools may not be able to handle
    - Large amount of heterogeneous knowledge sources
    - Ambiguity of entities and relations
    - Processes (activities), states, events, and their relations etc.

Learning a Health Knowledge Graph from Electronic Medical Records. Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; and Sontag, D. Nature Scientific Reports, 7(1): 5994. 2017.

UIUC COVID-19 Literature Knowledge Graph

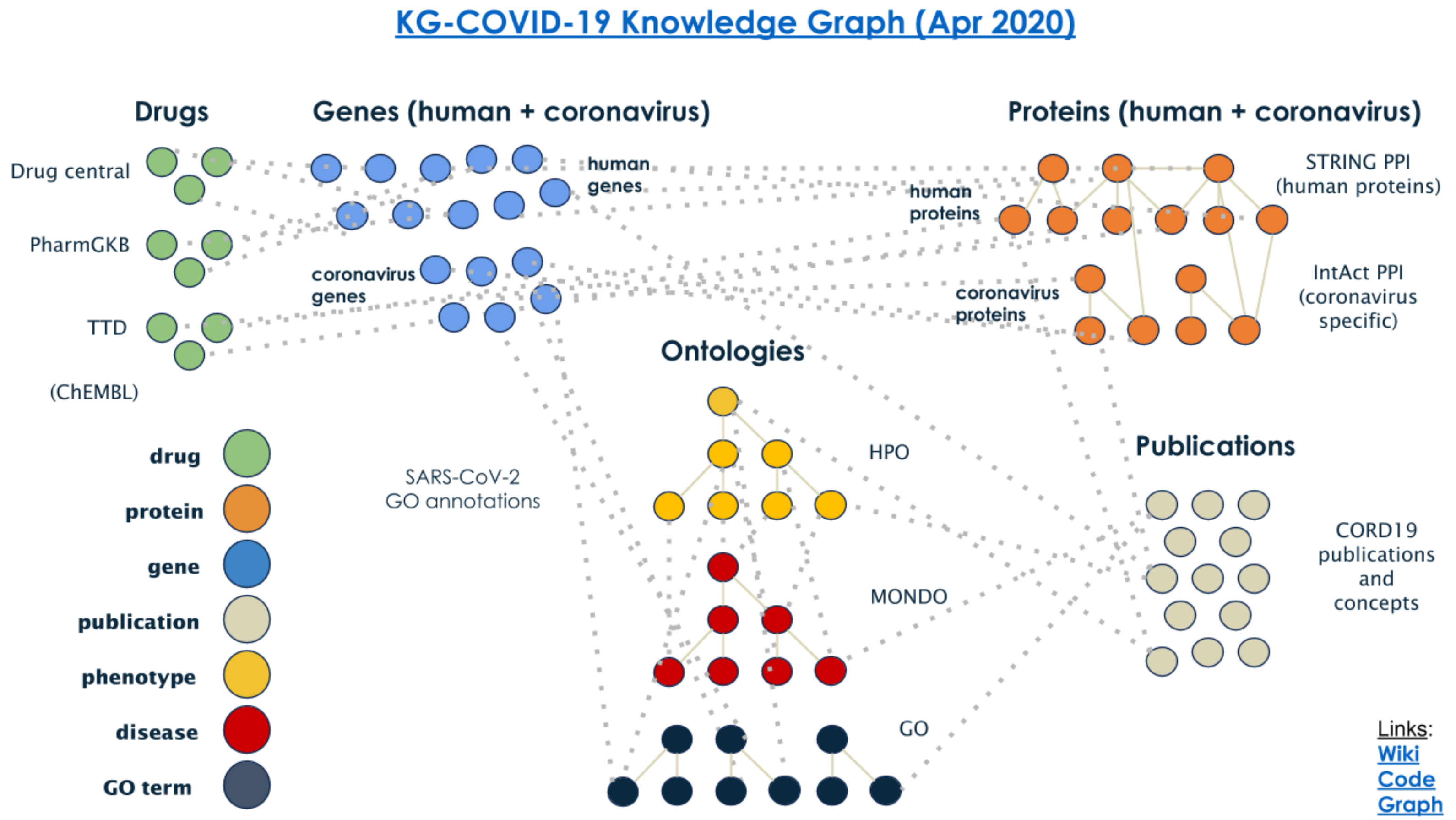
- http://blender.cs.illinois.edu/covid19/
- Extract entities, relations and events from tex
  - 50,752 Gene nodes
  - 10,781 Disease nodes
  - 5,738 Chemical nodes
  - 535 Organism nodes
  - 133 relation types
  - 13 Event types
- Knowledge extraction from images, and do cross-media fusion and inference with entitie and events



Q. Wang et al., COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. Arxiv, 2020

Berkeley Lab COVID-19 Knowledge Graph

32,000 drugs, 21,000 human and 272 viral proteins plus roughly the same number of genes, and more than 50,000 scientific studies and clinical trials.



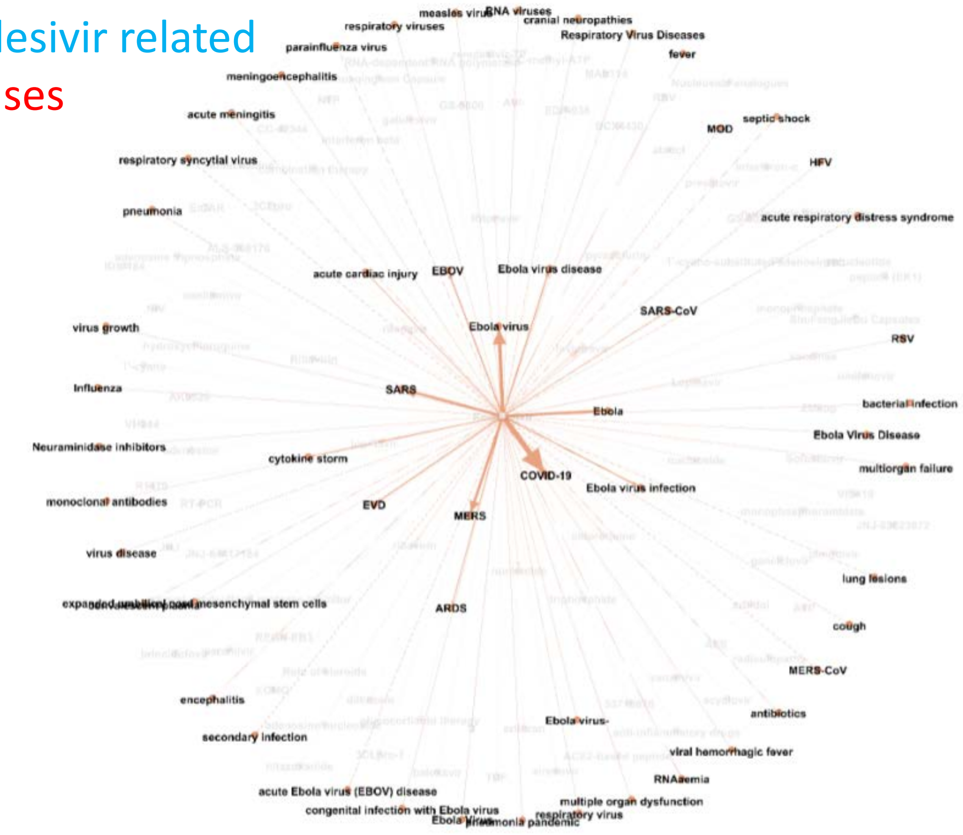
https://federallabs.org/news/berkeley-lab-creates-knowledge-graph-to-make-covid-19-drug-predictions

Image from: https://github.com/Knowledge-Graph-Hub/kg-covid-19/wiki

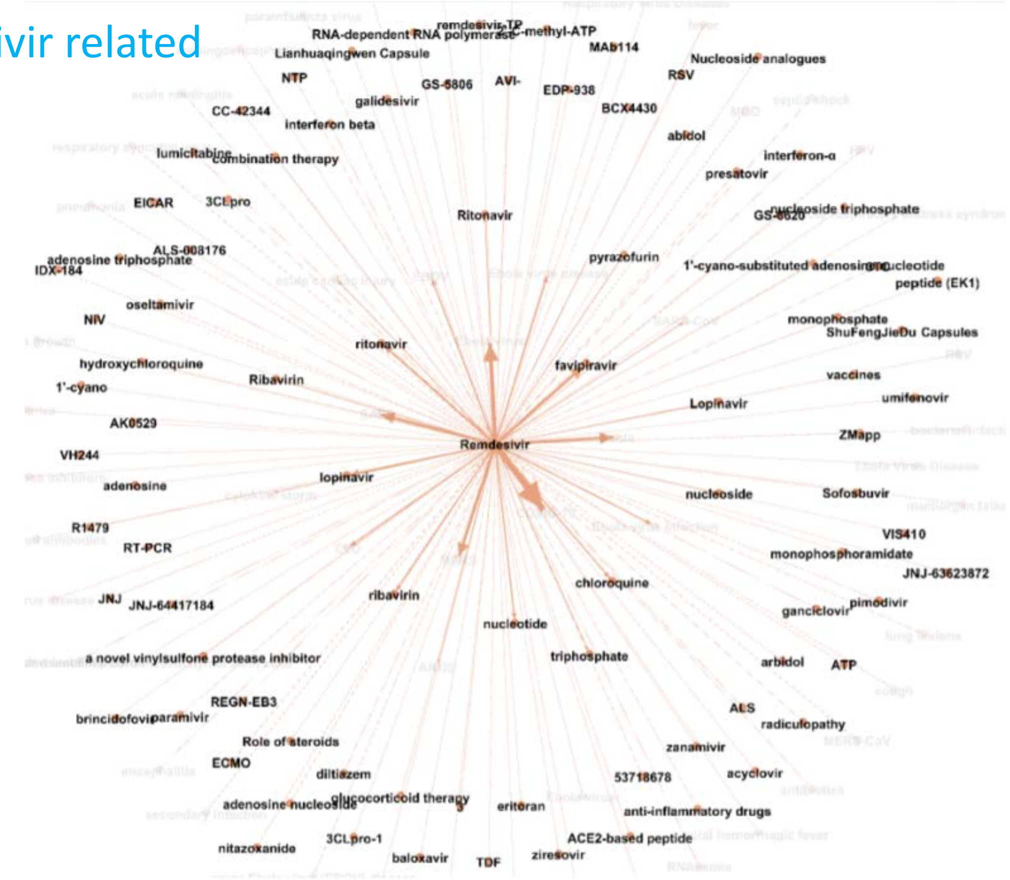
UT Austin COVID-19 Knowledge Graph

- 53,523 Drugs, 12,077 Diseases, 15,519 Species, 18,678 Genes, Gene mutations extracted from CORD-19 dataset

remdesivir related diseases



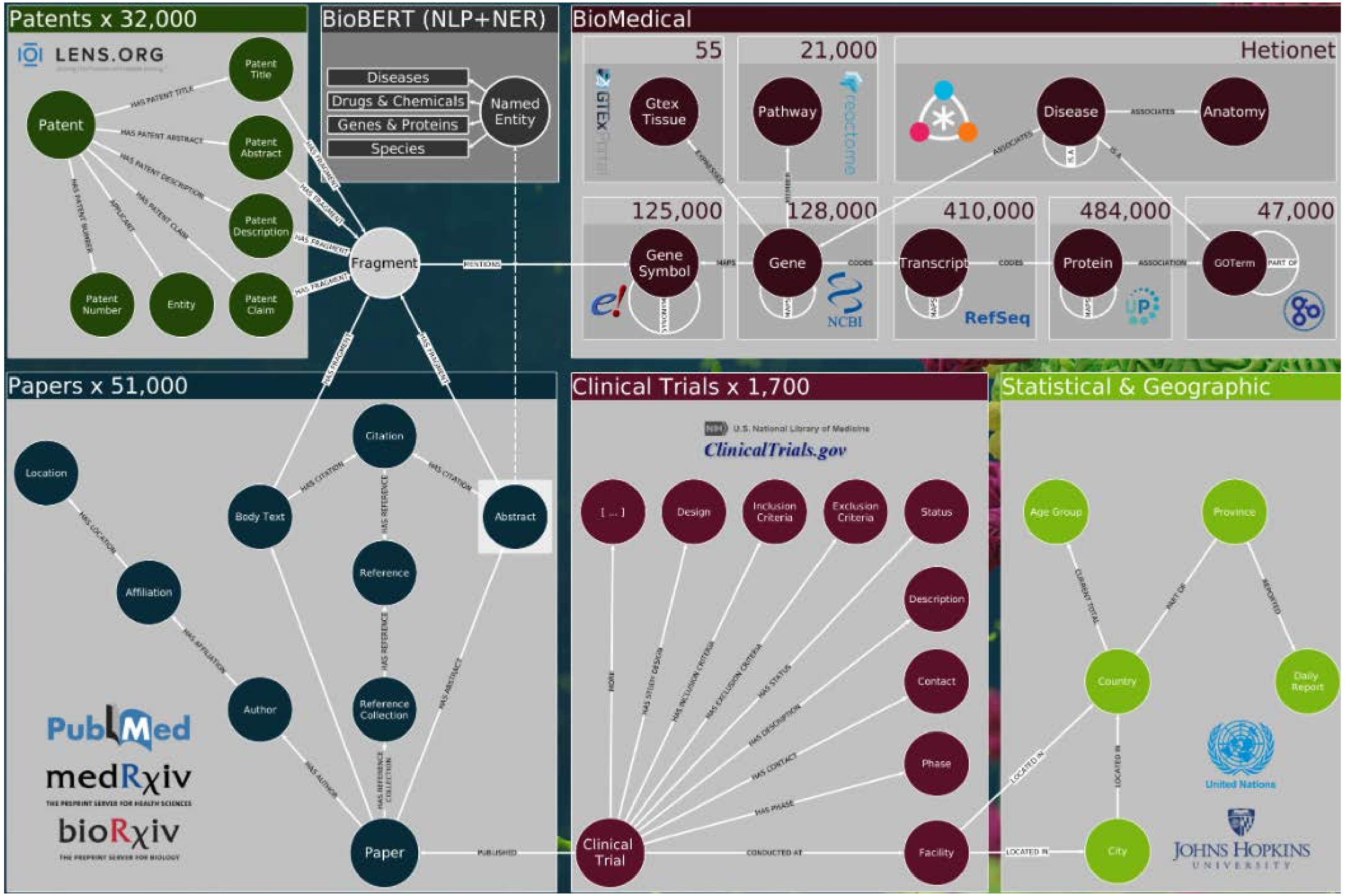
remdesivir related drugs



Chen, C., Ebeid, I.A., Bu, Y., & Ding, Y. (2020). Coronavirus knowledge graph: A case study. KDD Workshop on Knowledge Graph, 2020.
https://www.semanticscholar.org/cord19

COVIDGraph: https://covidgraph.org/

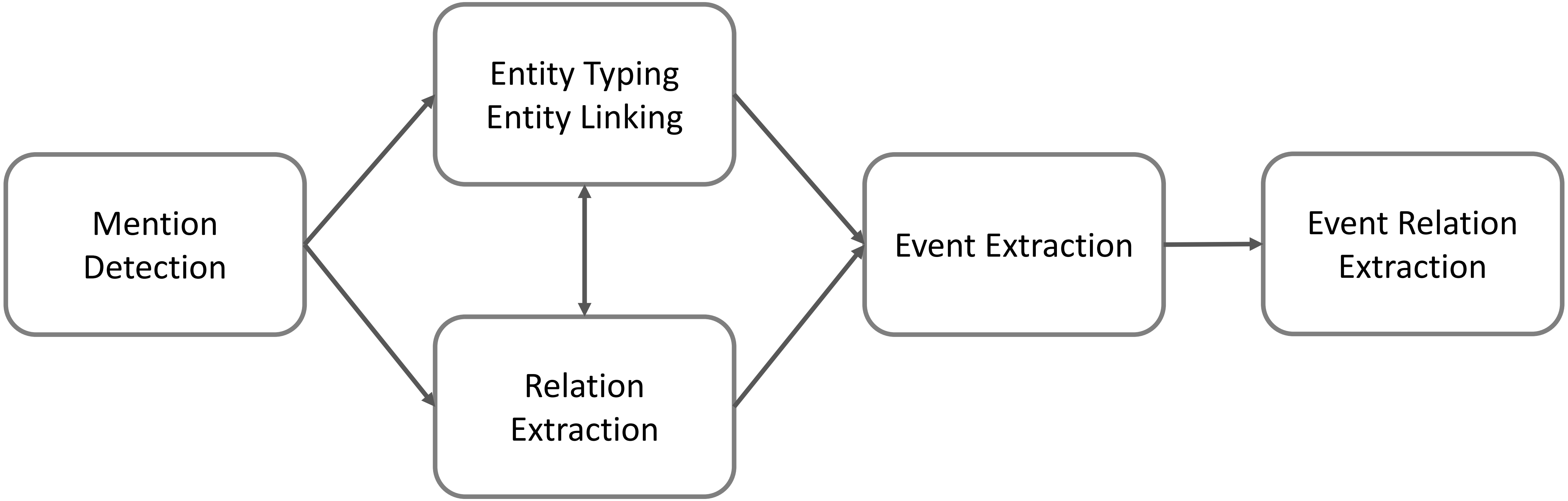
- “CovidGraph is a non-profit collaboration of researchers, software developers, data scientists and medical professionals.”



Outline

- Introduction
- Knowledge Graph Construction
  - Entities, typing and linking
  - Entity Relations
  - Events
  - Event relations
- Knowledge Graph Inference

General Procedure of Information Extraction for Knowledge Graph Construction



Information Extraction

- Information extraction can be pattern based or learning based
- Pattern based detection
  - Reflecting human's knowledge of language
  - Perform well when there are less annotated data
  - Especially good for open information extraction (Open-IE)
    - No pre-defined domains and relation (predicate) types among mentions
- Learning based detection
  - Better performance when there are many annotated training examples

Pattern Based Information Extraction

- Patterns can be extracted from dependency parsing results (Wang et al., 2018)

Pattern	Type	Example	Derived Clauses
		Basic Patterns	
SV_i	SV	Colectomy works.	(Colectomy, works)
SV_eA	SVA	Pulmonary toxicity is in lungs.	(Pulmonary toxicity, is, in lungs)
SV_cC	SVC	Pulmonary toxicity is vital.	(Pulmonary toxicity, is, vital)
$SV_{mt}O$	SVO	Nitrofurantoin causes pulmonary toxicity.	(Nitrofurantoin, causes, pulmonary toxicity)
$SV_{dt}O_iO$	SVOO	Colectomy gives the patient a chance.	(Colectomy, gives, the patient, a chance)
$SV_{ct}OA$	SVOA	Colectomy removes the colon away.	(Colectomy, removes, the colon, away)
$SV_{ct}OC$	SVOC	Pulmonary toxicity causes rats to die.	(Pulmonary toxicity, causes, rats, to die)

  - Extended pattern: Pulmonary toxicity often appears in lungs in rats. (SVAA)

- But as there is no restricted relations/predicates, relations should be disambiguated

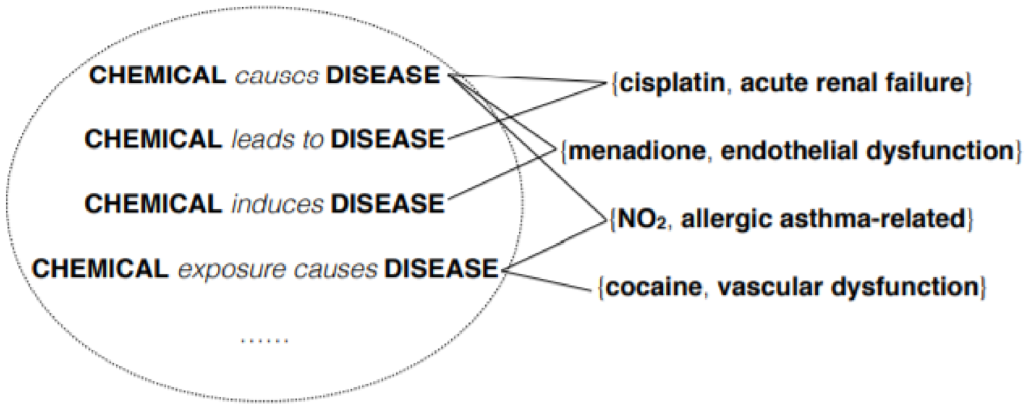


Table and Figure from: Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, Jiawei Han: Open Information Extraction with Meta-pattern Discovery in Biomedical Literature. BCB 2018

Learning Based Information Extraction

- New trend: pre-training and fine-tuning
- Pre-training with large amount of data, e.g.,

Model	Data	Method
BioBERT	PubMed abstracts/full texts	continual pre-training from BERT
Clinical-BERT	MINIC clinical notes	continual pre-training from BERT
SciBERT	Scientific papers from Semantic Scholar	from scratch
BlueBERT:	PubMed abstracts + MIMIC clinical notes	continual pre-training from BERT
PubMedBERT	PubMed abstracts/full texts	from scratch

- Fine-tuning: tasks specific datasets, e.g.,
  - Named entity recognition
  - Relation extraction

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020
Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew B. A. McDermott: Publicly Available Clinical BERT Embeddings. CoRR abs/1904.03323 (2019)
Iz Beltagy, Kyle Lo, Arman Cohan: SciBERT: A Pretrained Language Model for Scientific Text. EMNLP/IJCNLP (1) 2019

BioBERT

- Shows 0.62% Average F1 score improvement on 9 biomedical named entity recognition datasets and 2.80% Average F1 score improvement on 3 biomedical relation extraction datasets

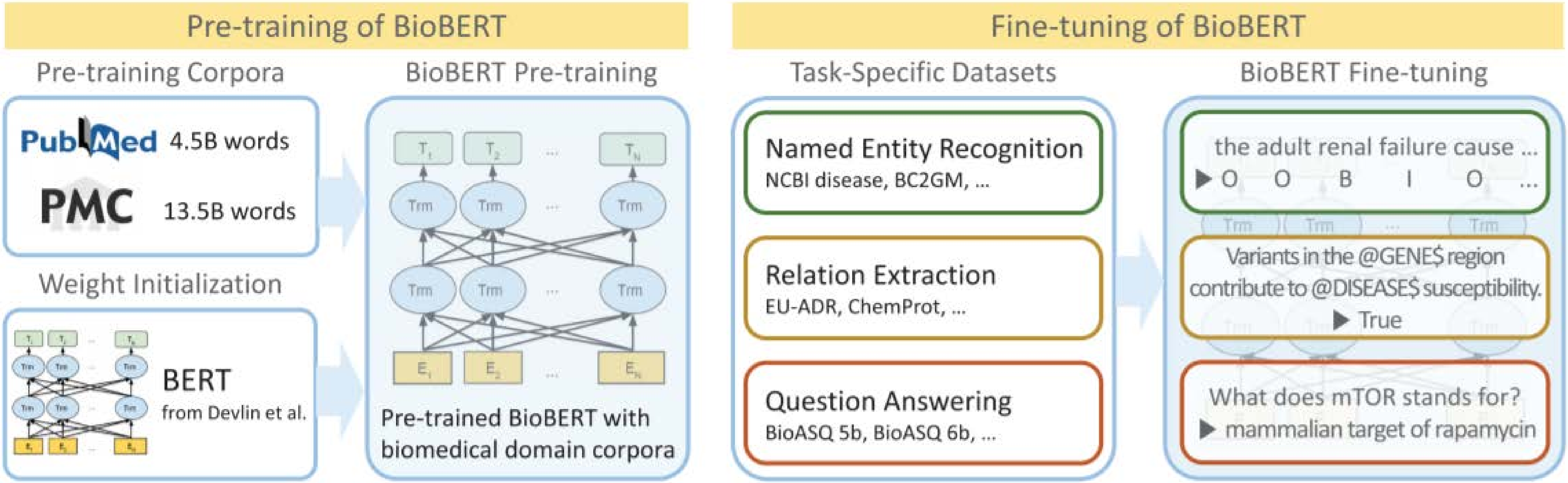
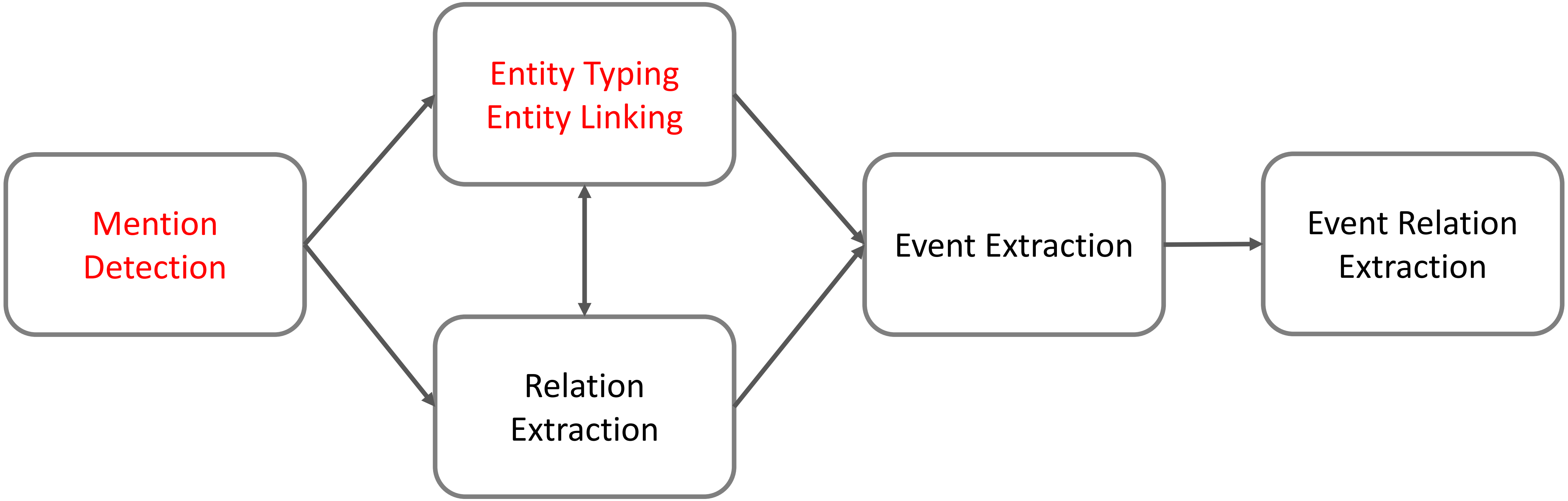


Figure from: Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020

General Procedure of Information Extraction for Knowledge Graph Construction



Entity Mention Detection and Typing

- Treat mention detection and typing as a word classification problem
- A typical label encoding is the BIO encoding
  - B-NP: beginning of a named entity chunk
  - I-NP: inside of a named entity chunk
  - O: outside of a named entity chunk

The	most	common	fatal	bacterial	diseases	are	respiratory	infections	.
O	O	O	O	O	O	O	B-Disease	I-Disease	O

- Then we can build a multi-class classifier or CRF model over certain features
  - E.g., representations provided by deep models

Entity Mention Detection and Typing

- Can also treat detection and typing as two separate tasks

- First, we use a tagger to extraction all mentions of interests
  - In this case, we can collect more training examples for each label

The	most	common	fatal	bacterial	diseases	are	respiratory	infections	.
O	O	O	O	O	O	O	B	I	O

- Second, we apply a multi-class (sometimes multi-label) classifier over the mentions

Entity Linking can Improve Entity Typing

- Especially for fine-grained entity typing
  - Most of the tasks are based weak/distant supervision
    - E.g., generating training data using the anchor links in Wikipedia
  - Entity linking to existing knowledge bases provide useful information

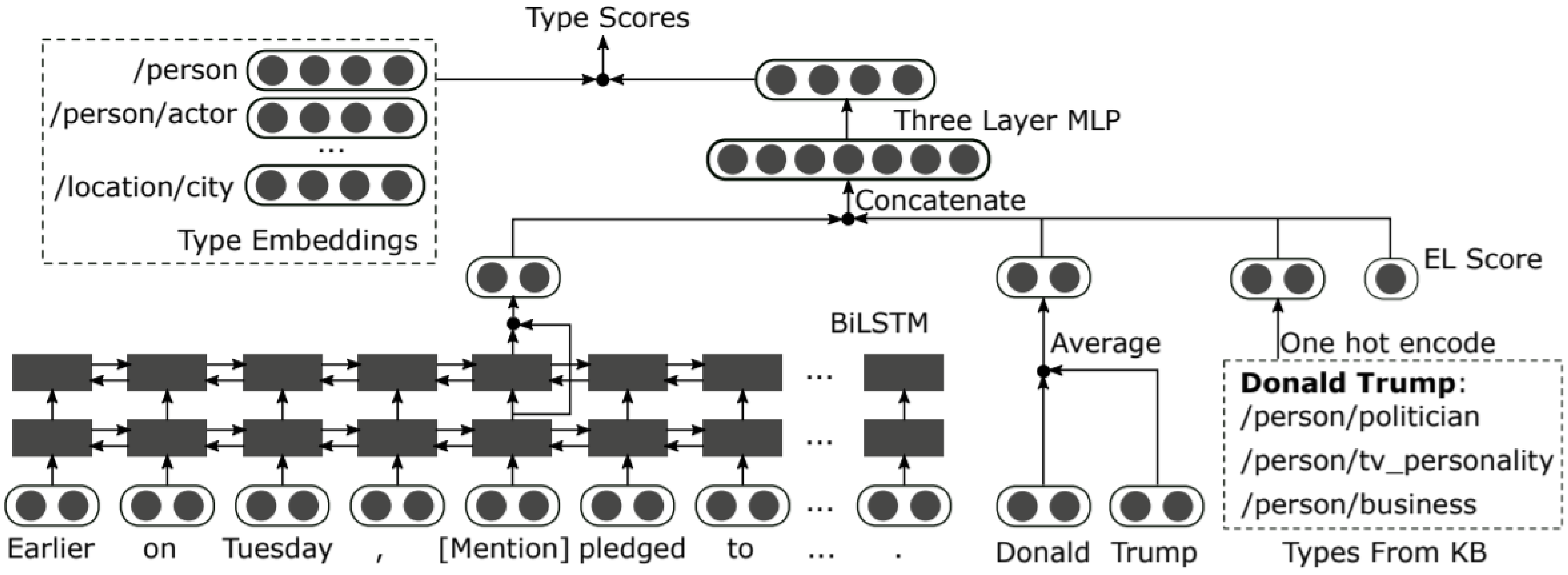


Figure from: Hongliang Dai, Donghong Du, Xin Li, Yangqiu Song: Improving Fine-grained Entity Typing with Entity Linking. EMNLP/IJCNLP (1) 2019
Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, Carolyn Penstein Rosé: MedType: Improving Medical Entity Linking with Semantic Type Prediction. Arxiv, 2020

An Example of COVID-19 NER

- 75 entity types including
  - Common biomedical entity types (e.g., genes, chemicals, and diseases),
  - New types related to COVID-19 (e.g., coronaviruses, viral proteins, evolution, materials, substrates and immune responses)

Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target. SARS-CoV-2 **CORONAVIRUS** has been sequenced [3 **CARDINAL**] . A phylogenetic **EVOLUTION** analysis [3 **CARDINAL** , 4 **CARDINAL**] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS** . There is a diversity of possible intermediate hosts for SARS-CoV-2 **CORONAVIRUS** , including pangolins **WILDLIFE** , but not mice **EUKARYOTE** and rats **EUKARYOTE** [5 **CARDINAL**] . There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS** . Using computer modeling , Xu et al . [6 **CARDINAL**] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE** . SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity to human ACE2 **GENE_OR_GENOME** , based on biochemical interaction studies and crystal structure analysis [7 **CARDINAL**] . SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and ……

Figure from: Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, Jiawei Han: Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision. CoRR abs/2003.12218 (2020)

Entity Linking

- Link a mention to an existing knowledge graph
- Can be done with a two-step approach
  - Mention detection
  - Disambiguation
- Disambiguation can be helped with relatively coarse-grained entity typing
  - E.g., aggregated 24 super types from 127 fine-grained UMLS types

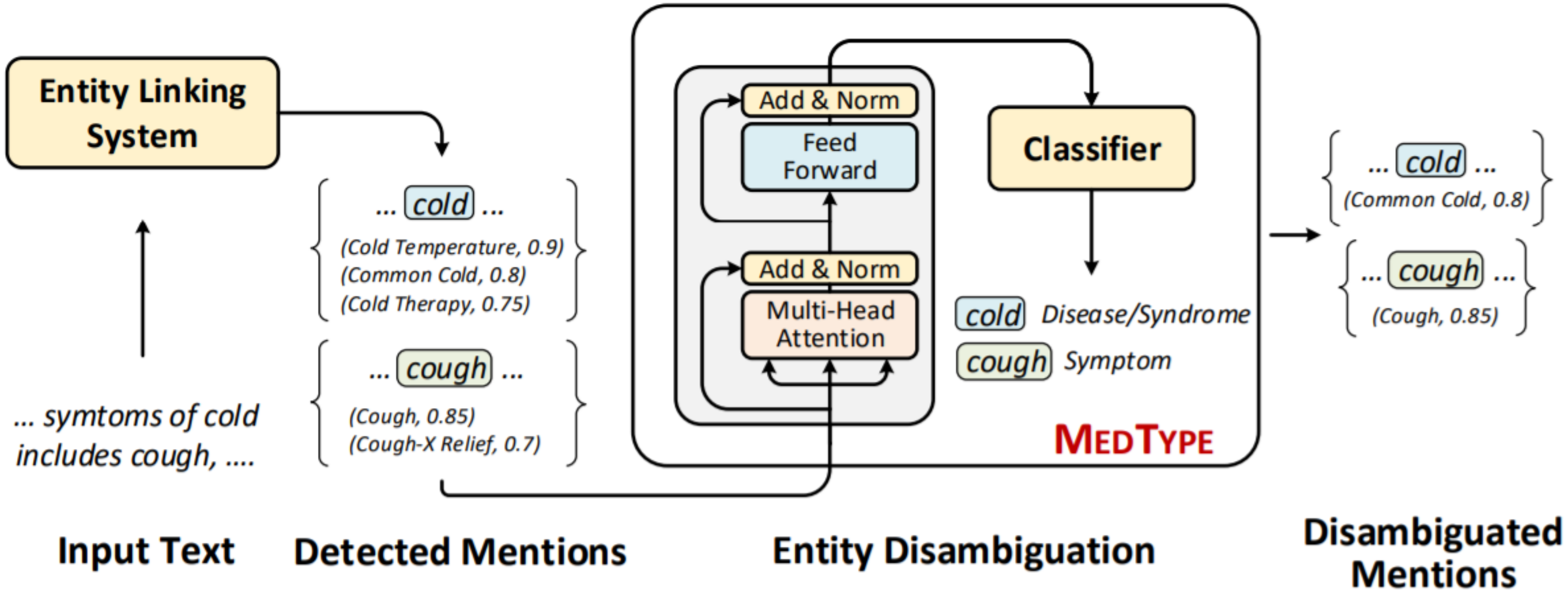


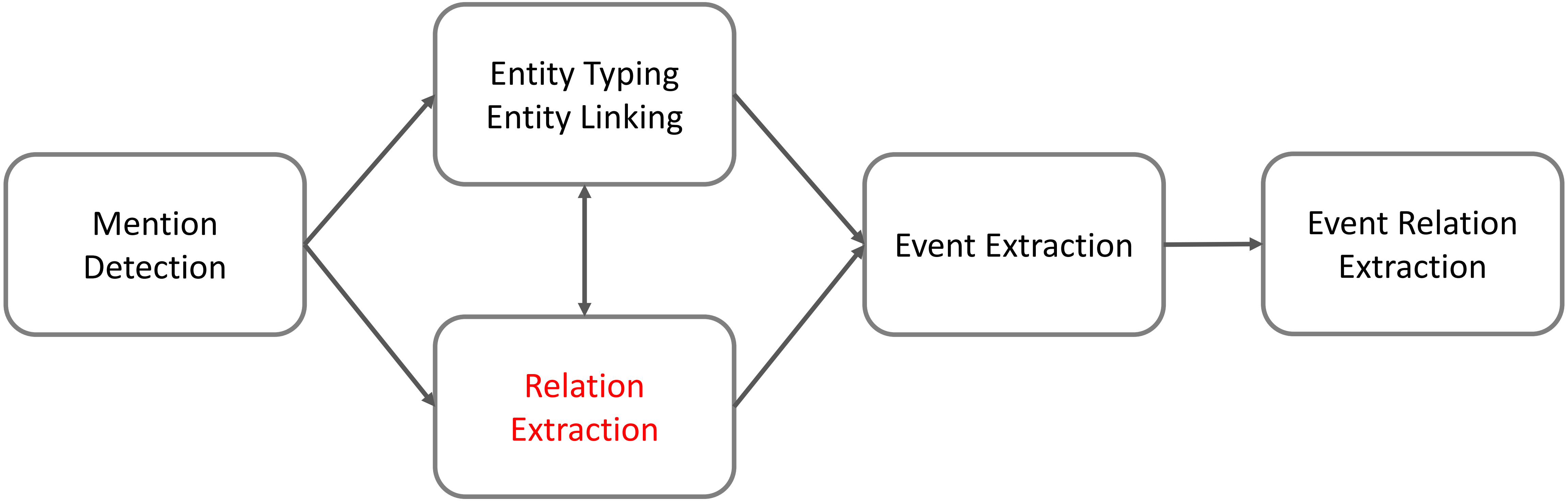
Figure from: Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, Carolyn Penstein Rosé: MedType: Improving Medical Entity Linking with Semantic Type Prediction, Arxiv, 2020

Mention Detection, Typing, and Linking

- Three tasks can be similar in designing models
  - With different annotation scheme and loss functions
- When typing is more difficult
  - E.g., there is lack of annotation
  - Adding linking results into features is helpful
- When linking is more difficult
  - E.g., lots of out-of-sample examples in test set
  - Adding typing can improve the results
- Can also perform joint learning when both types of annotation are available (Leaman and Lu, 2016; Zhao et al., 2019; Mohan and Li, 2019)

Robert Leaman, Zhiyong Lu: TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinform. 32(18): 2839-2846 (2016)
Sunil Mohan and Donghui Li: MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. AKBC, 2019
Sendong Zhao, Ting Liu, Sicheng Zhao, Fei Wang: A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. AAAI 2019

General Procedure of Information Extraction for Knowledge Graph Construction



Relation Extraction

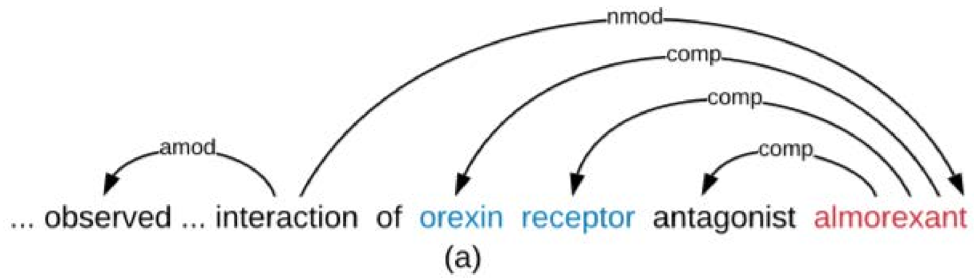
- Many related tasks for relation extraction in biomedical domain
  - Chemical (drug) induced diseases [Jiao Li et al., 2016]
  - Chemical (drug)-protein (gene) interaction [Martin Krallinger et al. 2017]
  - Phenotype (e.g., disease)-gene relations [Diana Sousa et al., 2019]
  - …
- For example, "observed ... interaction of orexin receptor antagonist almorexant"
  - Identity entities, e.g., protein and chemical
  - Classify relations

Group	*Eval.*	*CHEMPROT relations belonging to this group*
CPR:1	N	PART_OF
CPR:2	N	REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR|MODULATOR-ACTIVATOR|MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT

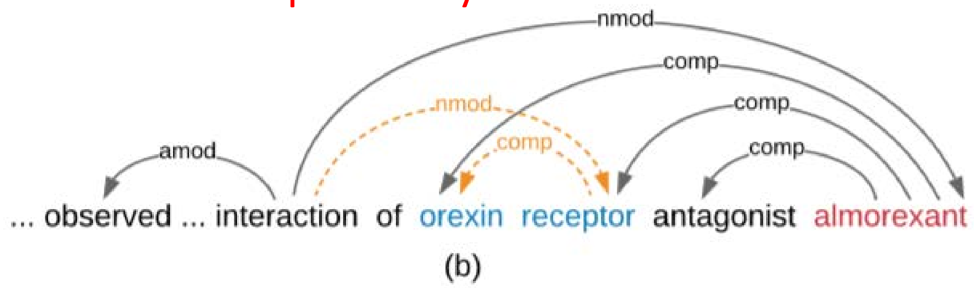
Jiao Li et al., BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database. 2016
Figure from: Martin Krallinger et al., Overview of the BioCreative VI chemical-protein interaction Track, BioCreative Challenge Evaluation Workshop. 2017
Diana Sousa, Andre Lamurias, Francisco M. Couto: A Silver Standard Corpus of Human Phenotype-Gene Relations. NAACL-HLT (1) 2019

Learning for Relation Extraction

- A recent work show that using dependency parse forest than the 1-best tree with a GNN can improve the performance



Dependency Parse Tree



Dependency Parse Forest

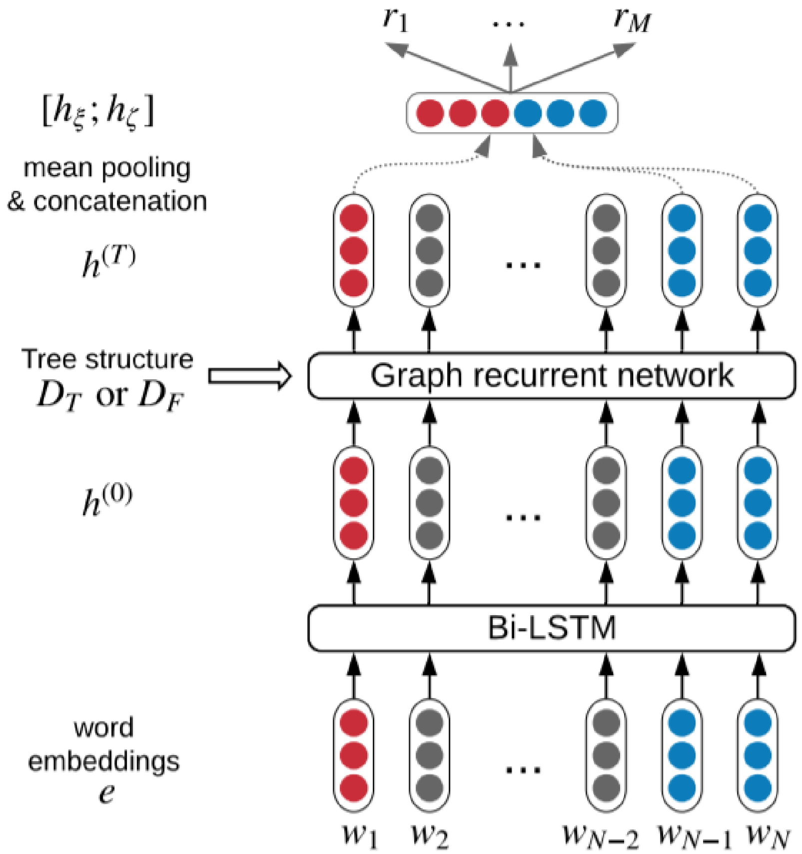


Figure from: Linfeng Song, Yue Zhang, Daniel Gildea, Mo Yu, Zhiguo Wang, Jinsong Su: Leveraging Dependency Forest for Neural Medical Relation Extraction. EMNLP/IJCNLP (1) 2019

N-Ary Relation Extraction

- Many relations (facts) involve more than two arguments, not always possible to decompose them into binary facts without losing information

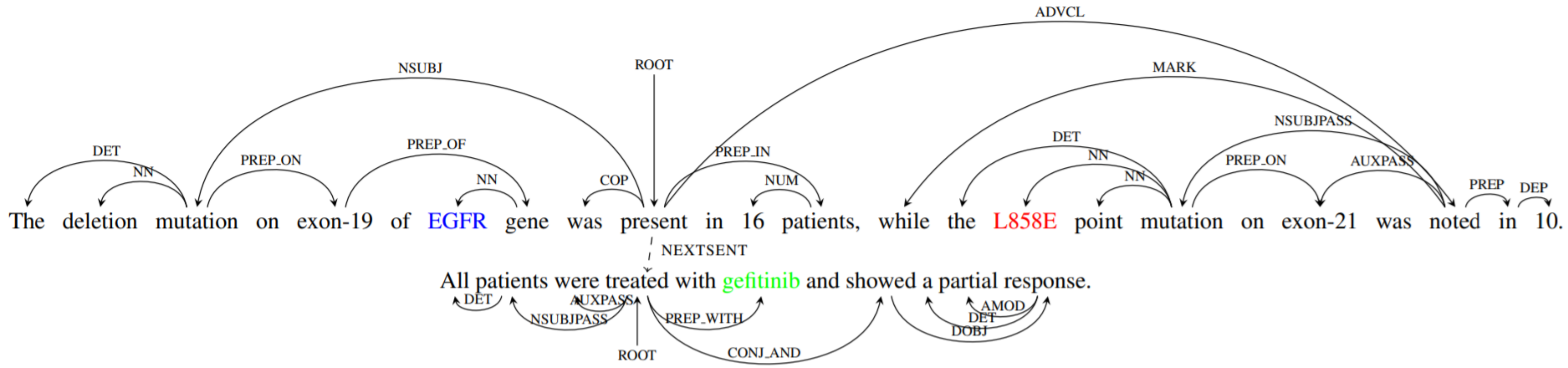
➢ 2.5 mg Albuterol may be used to treat acute exacerbations, particularly in children.
➢ Salmonella infection is a common cause of bacteremia in Africa.

Relation	Arity	Signature
Treats	5	$Drug \times Disease \times Dosage \times DosageForm \times Targetgroup$
ReducesRisk	4	$(Drug \cup Behavior \cup Ecofactor) \times Disease \times Targetgroup \times Condition$
Causes	4	$Disease \times Disease \times Targetgroup \times Condition$
Diagnoses	3	$DiagnosticProcedure \times Disease \times (BodyPart \cup Organ)$

Table from: Patrick Ernst, Amy Siu, Gerhard Weikum: HighLife: Higher-arity Fact Harvesting. WWW 2018

Learning for N-Ary Relation Extraction

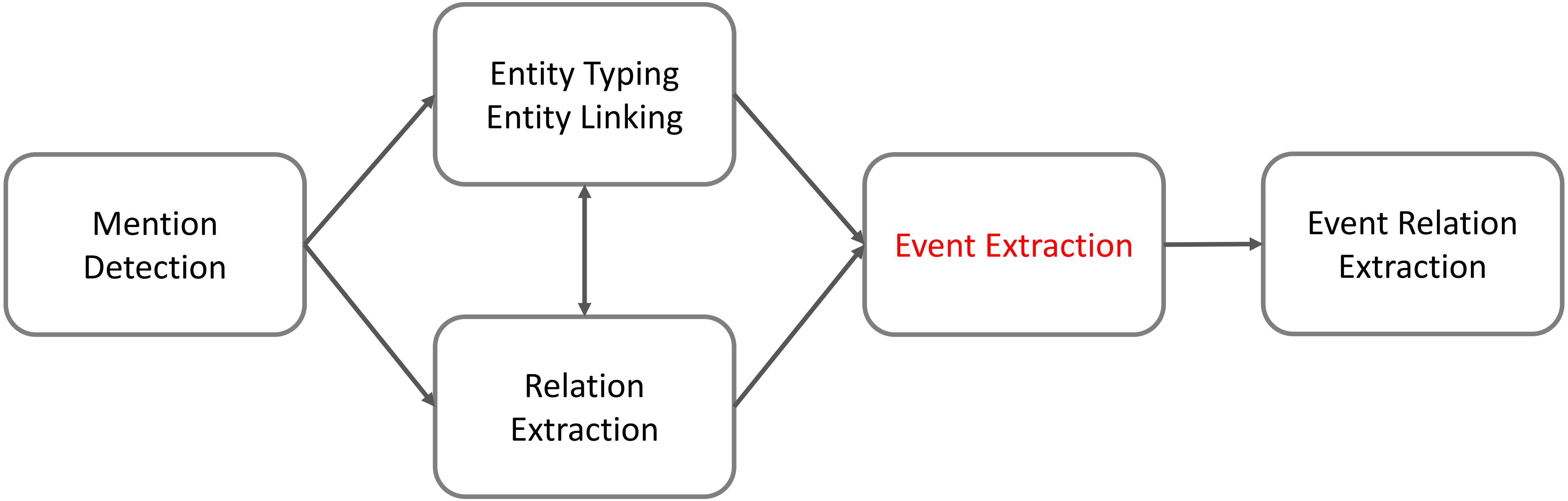
- N-Ary relation extraction with graph LSTM
  - Compared to event extraction, there is no simplification based on Davidsonian semantics (trigger-argument relations)
- Cross-sentence extraction
- Assisted with many distant supervision examples



Fact: tumors with L858E mutation in EGFR gene can be treated with gefitinib.

Figure from: Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih: Cross-Sentence N-ary Relation Extraction with Graph LSTMs. TACL, 2017

General Procedure of Information Extraction for Knowledge Graph Construction



Event Extraction

- According to BioNLP Shared Task 2013, there are several domain-dependent event extraction tasks
  - Genia Event Extraction
  - Cancer Genetics
  - Pathway Curation
  - Corpus Annotation with Gene Regulation Ontology
  - Gene Regulation Network in Bacteria
  - Bacteria Biotopes
- They share similar annotation types
  - Text-bound annotation (entity/event trigger)
  - Equiv: entity aliases
  - E: event
  - M: event modification
  - R: relation
  - N: normalization (external reference)

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, Pierre Zweigenbaum: Overview of BioNLP Shared Task 2013. BioNLP@ACL (Shared Task) 2013

Event Extraction Example

- Cancer genetics
  - 40 event types
  - 18 types of entities
  - 600 PubMed abstracts
  - 17,000 events

- Example Roles:
  - Theme: Entity or event that undergoes the primary effects of the event.
  - Cause: Entity or event that is causally active in the event.

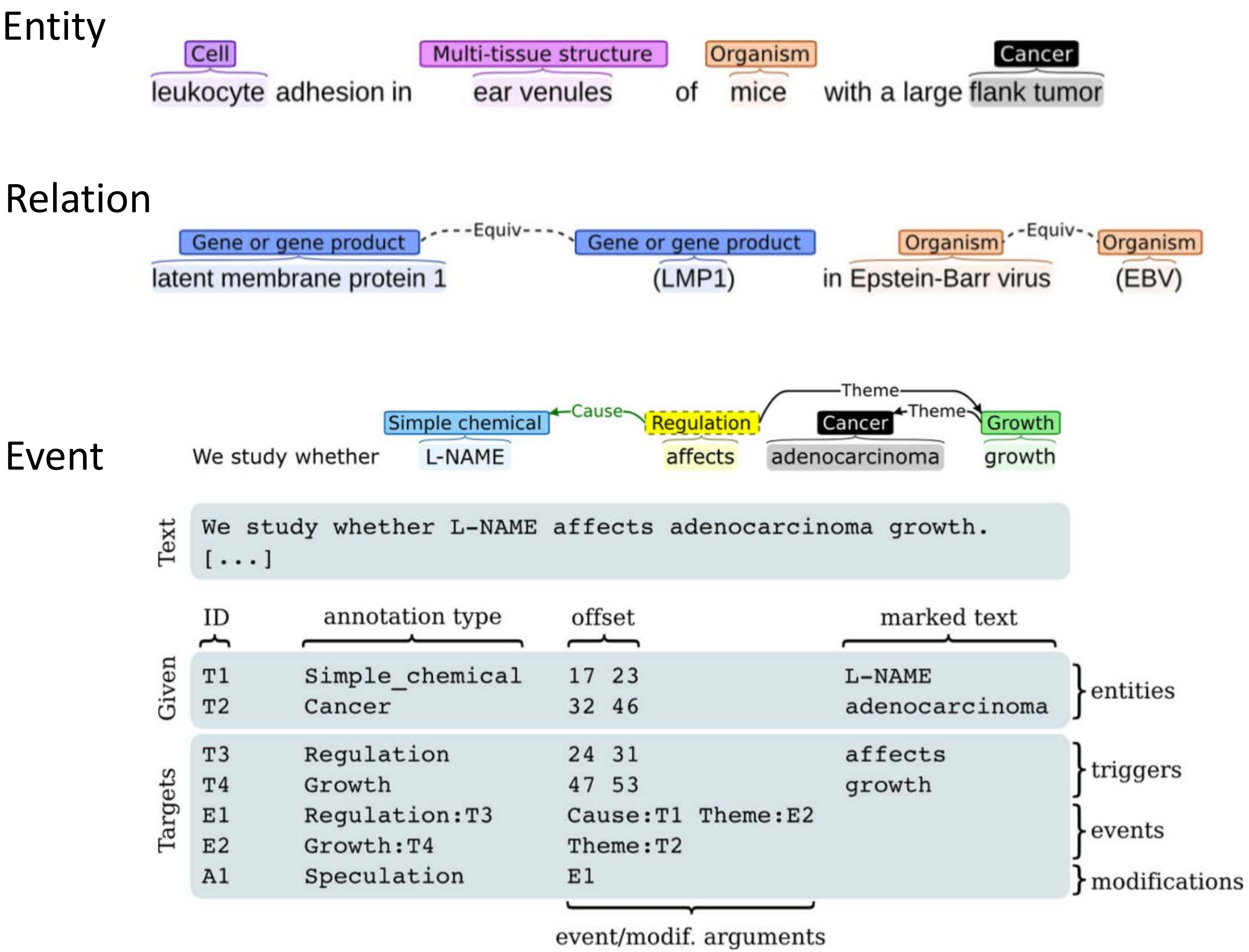


Figure from: Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, Sophia Ananiadou. BMC Bioinformatics, 2015.

Learning for Event Extraction

- Usually done with a pipeline based approach
  - Step 1: Trigger identification and classification
  - Step 2: Argument identification and role classification
    - Sometimes need to consider entity type constraints

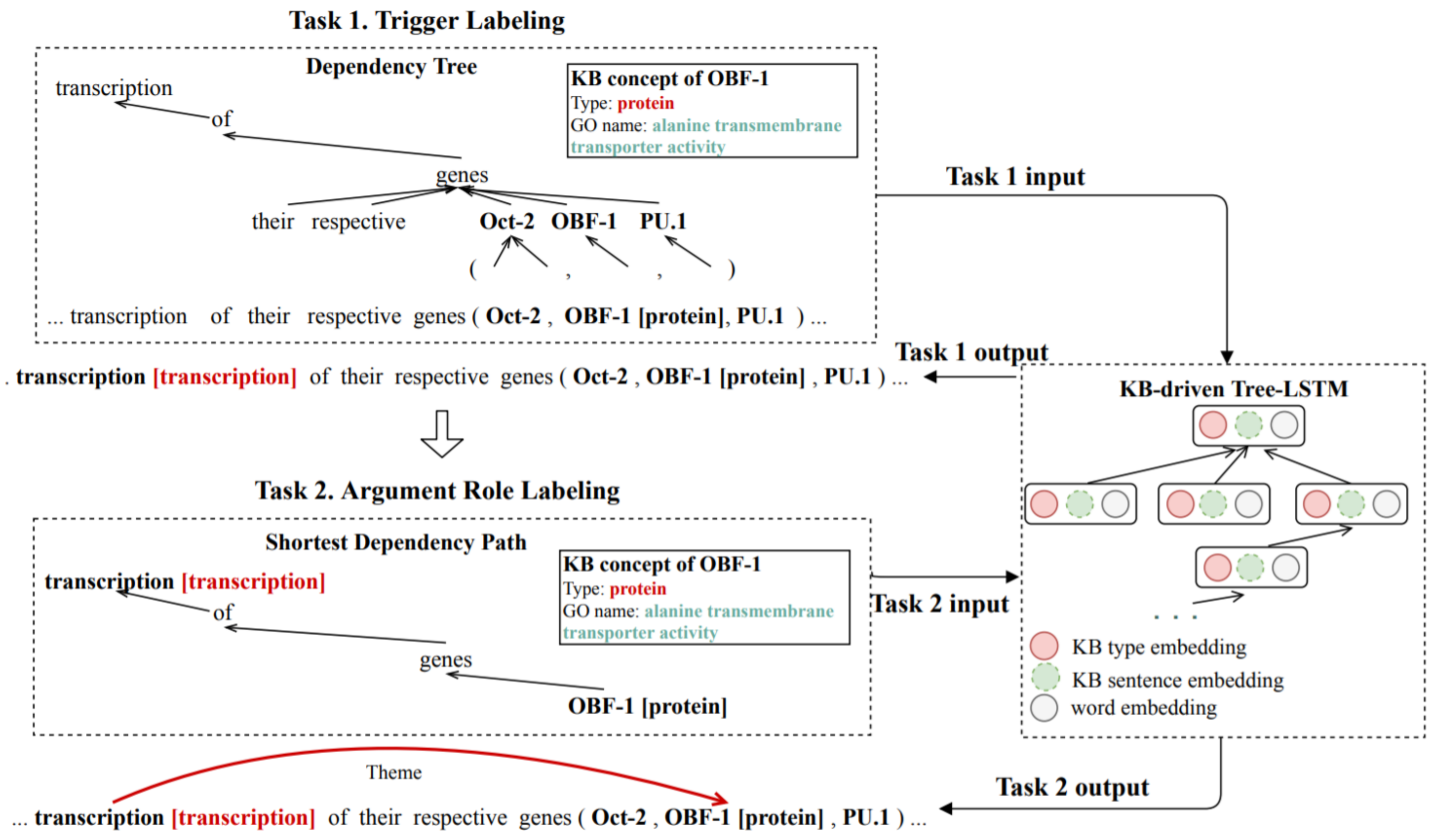
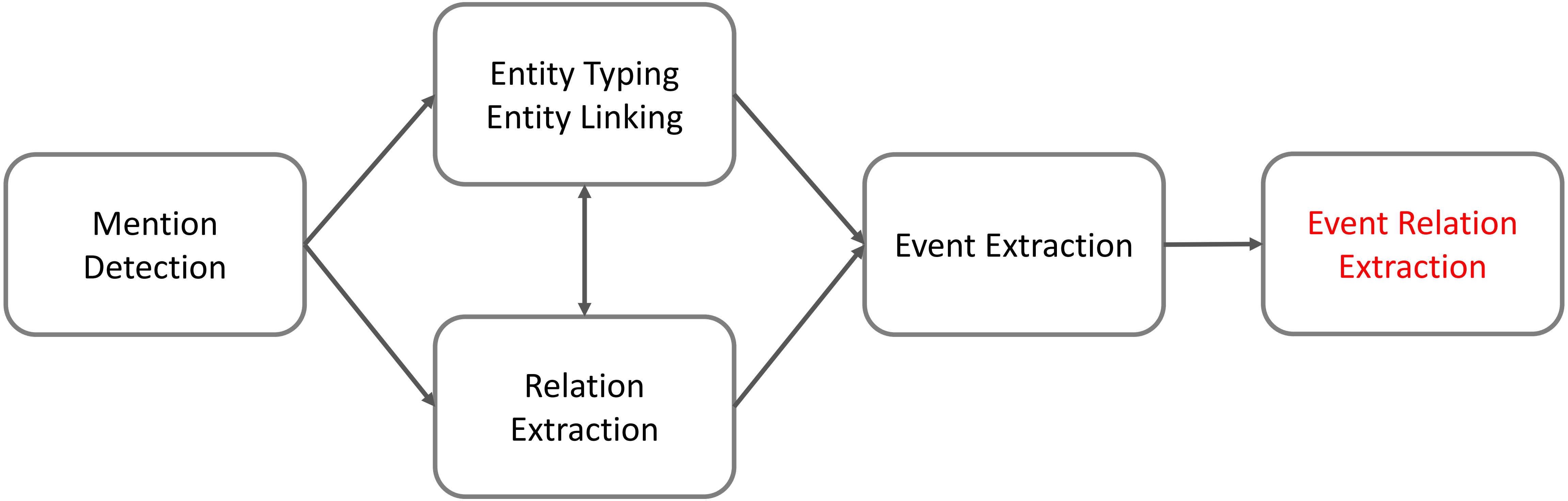


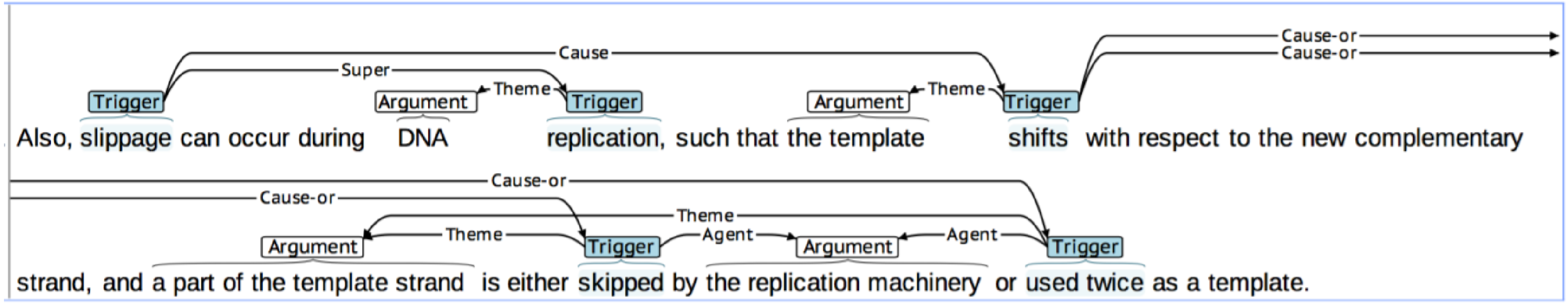
Figure from: Diya Li, Lifu Huang, Heng Ji, Jiawei Han: Biomedical Event Extraction based on Knowledge-driven Tree-LSTM. NAACL-HLT (1) 2019

General Procedure of Information Extraction for Knowledge Graph Construction



Event Relations

- Usually appear in temporal relation extraction, narrative schema, discourse analysis in NLP
- Annotation Scheme
  - Triggers and arguments
  - Semantic roles
  - Event-event relations
    - Cause, Enable, Prevent, Super



Aju Thalappillil Scaria, Jonathan Berant, Mengqiu Wang, Peter Clark, Justin Lewis, Brittany Harding, Christopher D. Manning: Learning Biological Processes with Global Constraints. EMNLP 2013
Figure from: Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, Christopher D. Manning: Modeling Biological Processes for Reading Comprehension. EMNLP 2014

Event Relations

- Answering questions

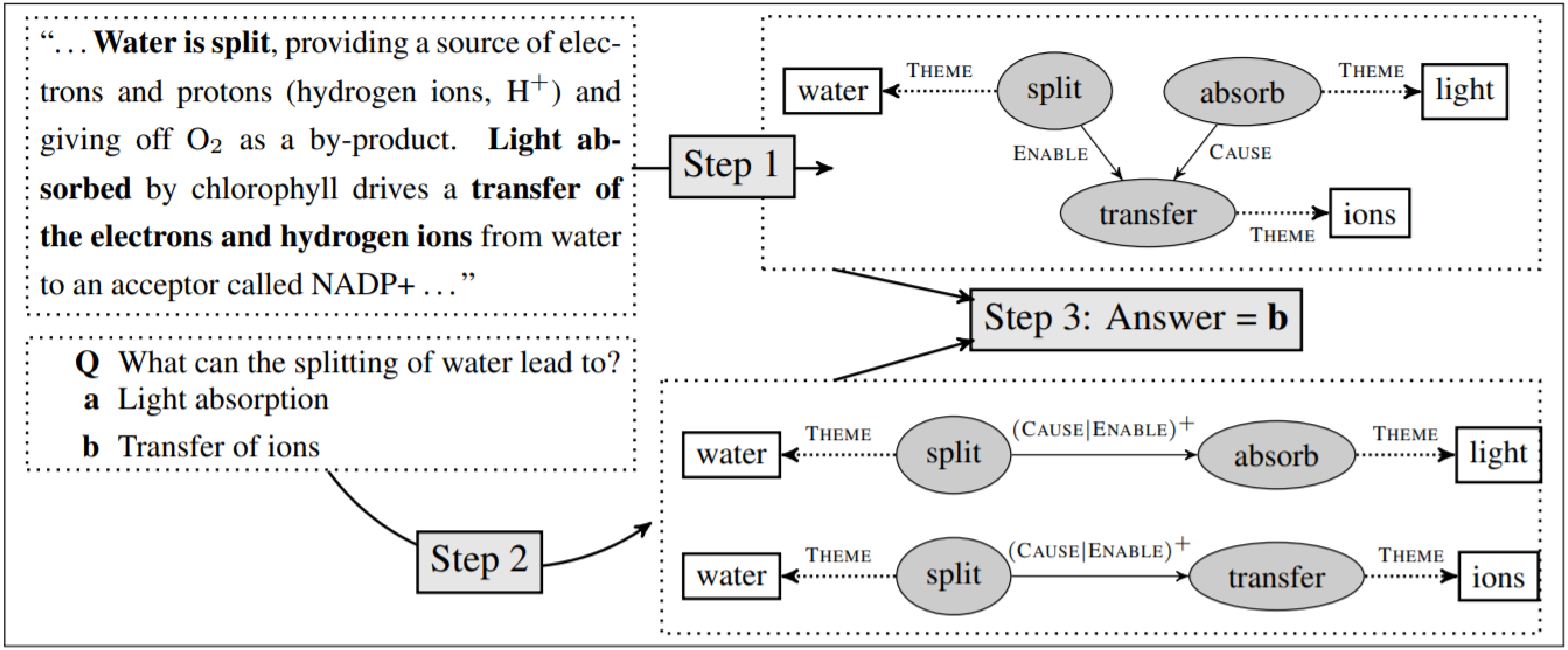


Figure from: Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, Christopher D. Manning: Modeling Biological Processes for Reading Comprehension. EMNLP 2014

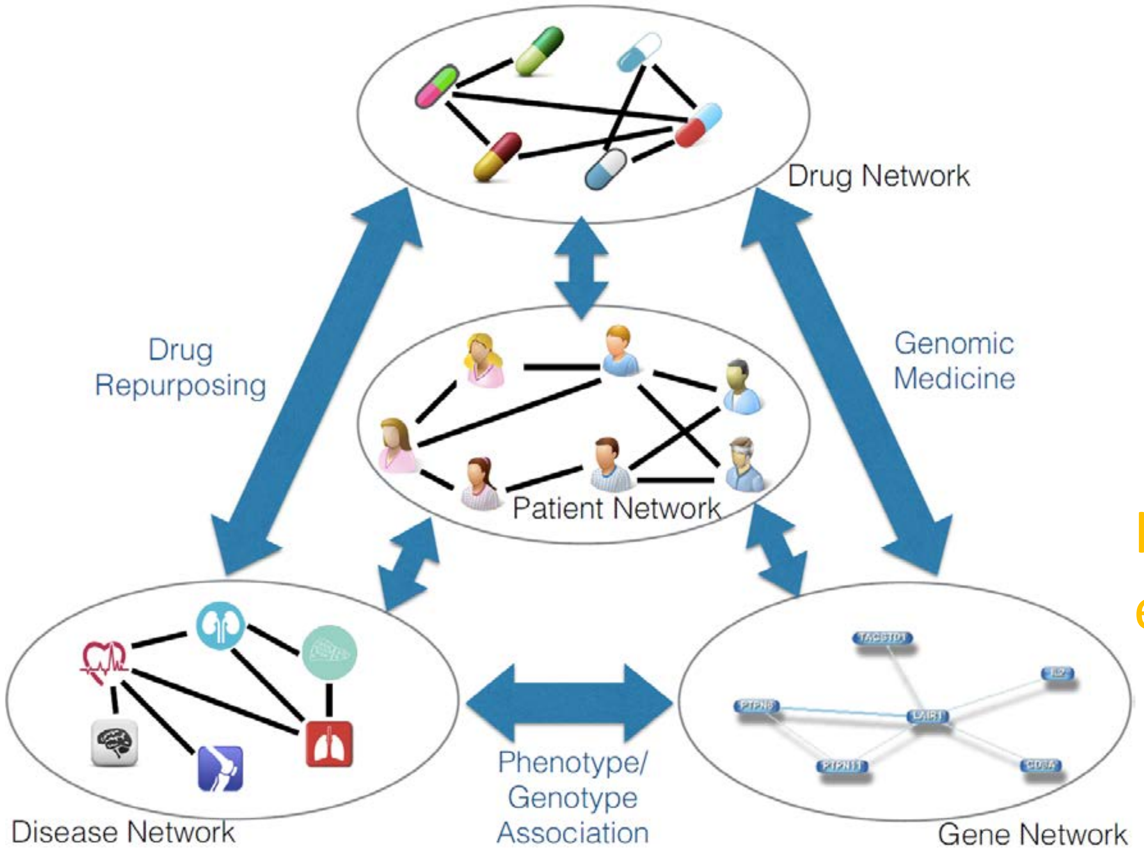
Outline

- Introduction
- Knowledge Graph Construction
- Knowledge Graph Inference
  - Knowledge graph representations
  - Application: QA with knowledge graph

What is the difference..

- between a knowledge graph and traditional graphs, such as social networks?
  - Knowledge graphs are information networks with semantic meanings
    - Entities/instances can be typed/conceptualized
  - Knowledge graph structures show different properties
    - Types/concepts can be organized as a hierarchical tree.

KG as a Multi-relational Graph



- Nodes are treated as the same type
- Relations are distinguished in triplets (head, relation, tail)

$$\text{Score}(\boldsymbol{h}_i + \boldsymbol{r}_k - \boldsymbol{t}_j) = \|\boldsymbol{h}_i + \boldsymbol{r}_k - \boldsymbol{t}_j\|$$

Head entity ID i
e.g., Crohn's disease

Relation type k
e.g., SymptomOfDisease

Tail entity ID j
e.g., Diarrhea

$$\text{Score}(\boldsymbol{h}_i + \boldsymbol{r}_k - \boldsymbol{t}_j) > \text{Score}(\boldsymbol{h}_i + \boldsymbol{r}_k - \boldsymbol{t}_l) + \delta$$

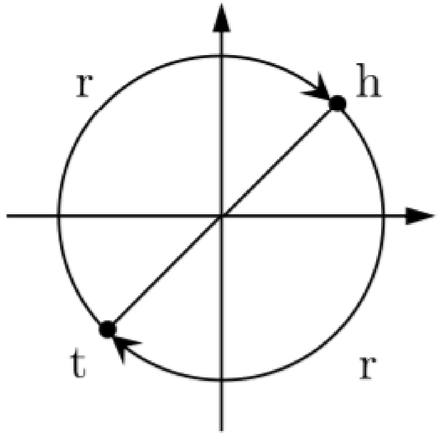
Figure Credit: Fei Wang

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko: Translating Embeddings for Modeling Multi-relational Data. NIPS 2013
David Chang, et al., Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. BioNLP 2020

Many Variants of Multi-relational Embedding

Model	Score Function	Symmetry	Antisymmetry	Inversion	Composition
SE	$-\|\boldsymbol{W}_{r,1}\mathbf{h} - \boldsymbol{W}_{r,2}\mathbf{t}\|$	✗	✗	✗	✗
TransE	$-\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$	✗	✓	✓	✓
TransX	$-\|g_{r,1}(\mathbf{h}) + \mathbf{r} - g_{r,2}(\mathbf{t})\|$	✓	✓	✗	✗
DistMult	$\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle$	✓	✗	✗	✗
ComplEx	$\mathrm{Re}(\langle \mathbf{h}, \mathbf{r}, \overline{\mathbf{t}} \rangle)$	✓	✓	✓	✗
RotatE	$-\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$	✓	✓	✓	✓

- Symmetry
  - $\forall x, y \;\; r(x, y) \Rightarrow r(y, x)$
- Anitsymmetry
  - $\forall x, y \;\; r(x, y) \Rightarrow \neg r(y, x)$
- Inversion
  - $\forall x, y \;\; r_2(x, y) \Rightarrow r_1(y, x)$
- Composition
  - $\forall x, y, z \;\; r_1(x, y) \wedge r_2(y, z) \Rightarrow r_3(x, z)$



An example of modeling symmetric relations with RotatE model. The relation corresponds to a counterclockwise rotation by θ_r radians about the origin of the complex plane

Figure and Table from: Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, “Rotate: Knowledge graph embedding by relational rotation in complex space,” in ICLR, 2019.

KG as a Heterogeneous Information Network

- A Heterogeneous Information Network
  - Typed entities
  - Typed relations

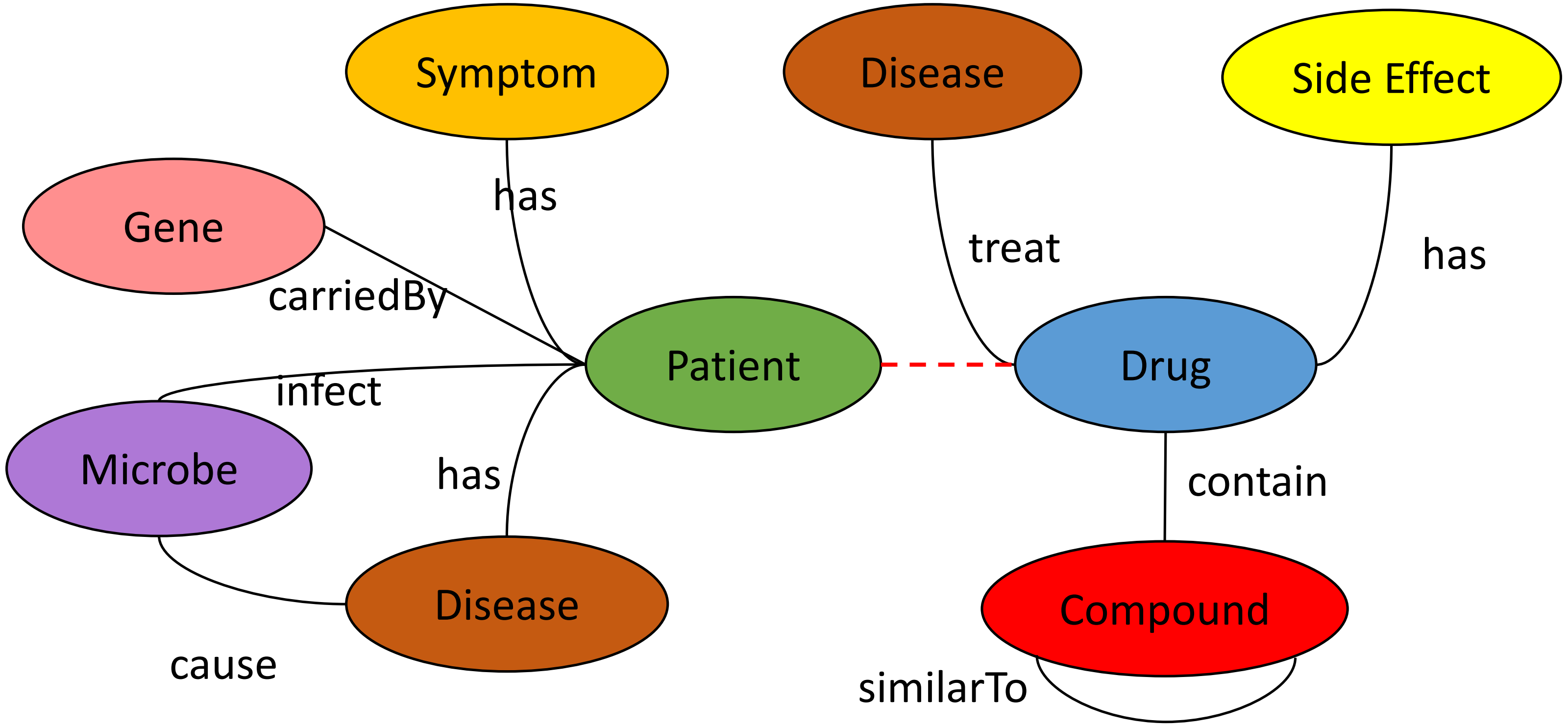
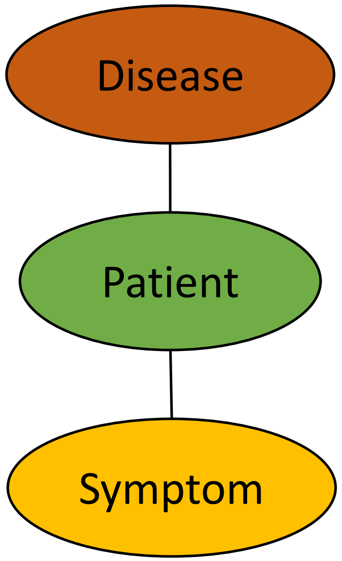


Figure Credit: Yizhou Sun, WWW'17 Tutorial

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu: PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. Proc. VLDB Endow. (2011)

Heterogeneous Information Network (HIN)

- A powerful tool provided by HIN representation is the semantic similarities computed based on metapaths or metagraphs



	p_1	p_2	p_3
d_1		1	
d_2	1		1
d_3		1	

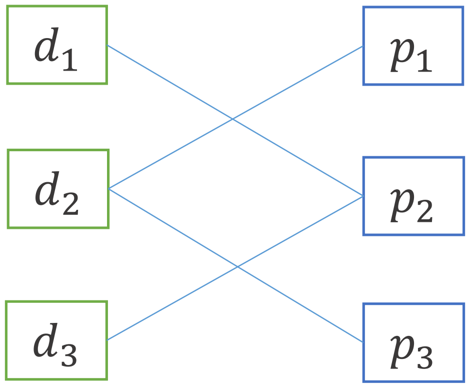
W_{DP}

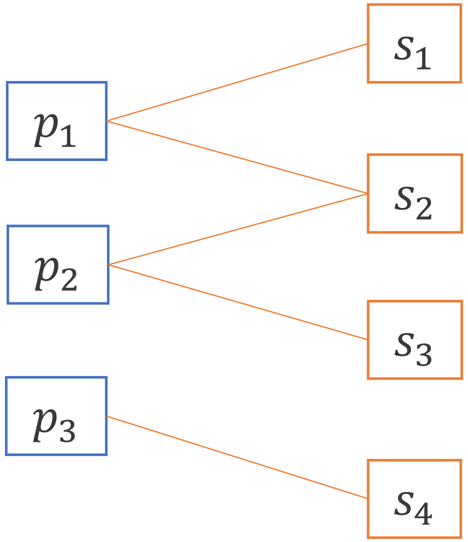
	s_1	s_2	s_3	s_4
p_1	1	1		
p_2		1	1	
p_3				1

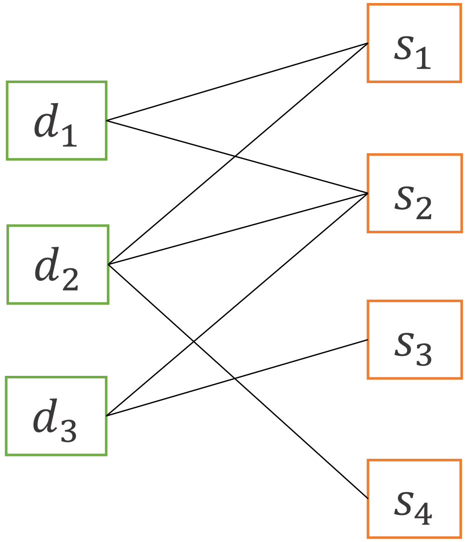
W_{PS}

	s_1	s_2	s_3	s_4
d_1		1	1	
d_2	1	1		1
d_3		1	1	

$W_{DS} = W_{DP}\, W_{PS}$

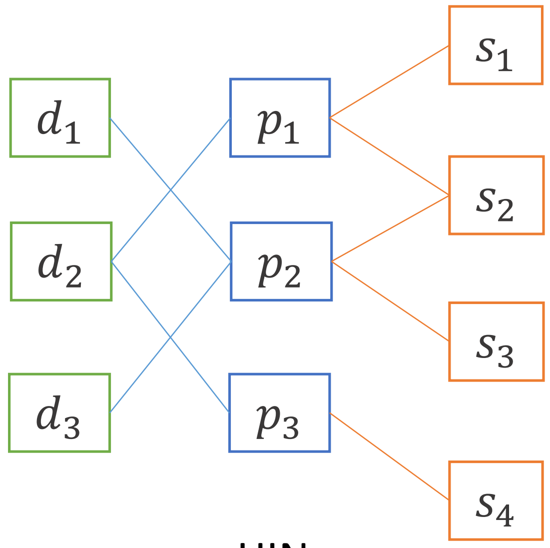




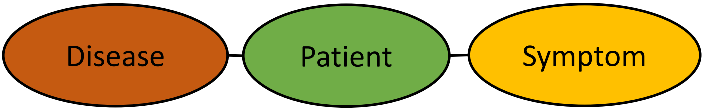


HIN Embeddings—MetaPath2Vec [Dong et al., 2017]

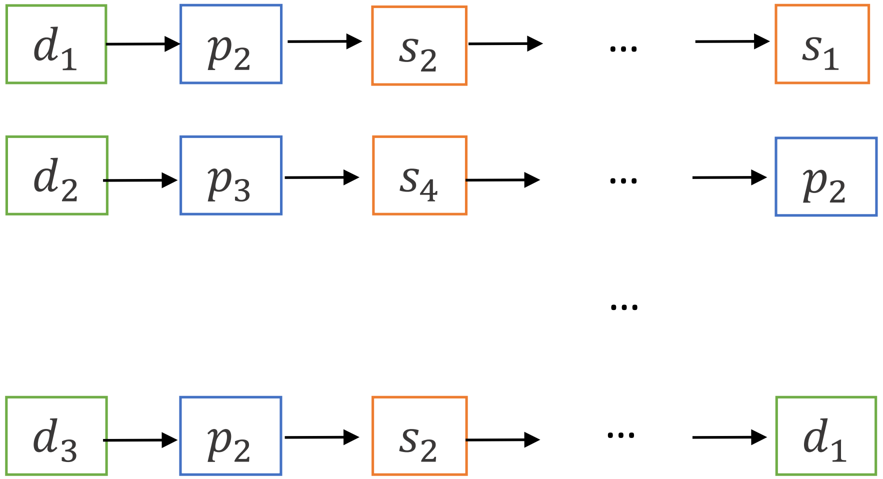
- Embedding nodes based on DeepWalk [Perozzi et al., 2014] guided by metapaths



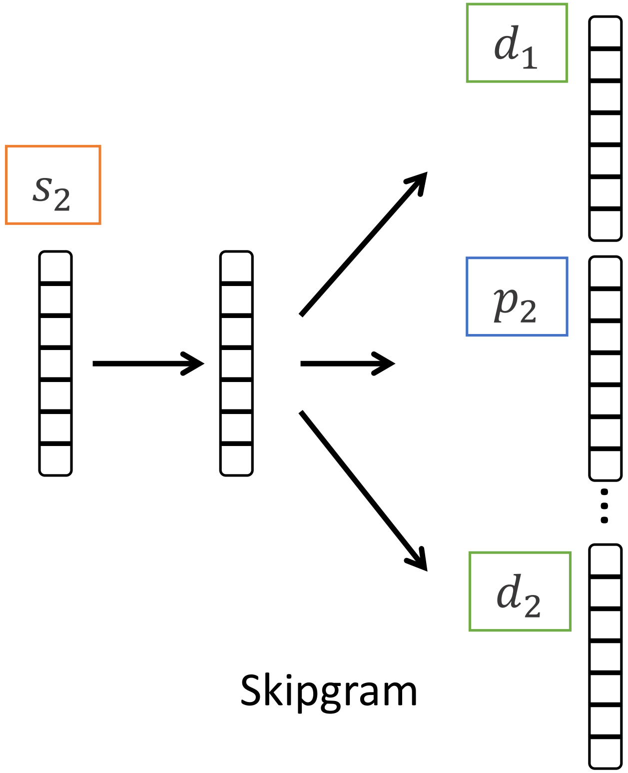
HIN



Metapath



Guided random walk
following the given meta-path



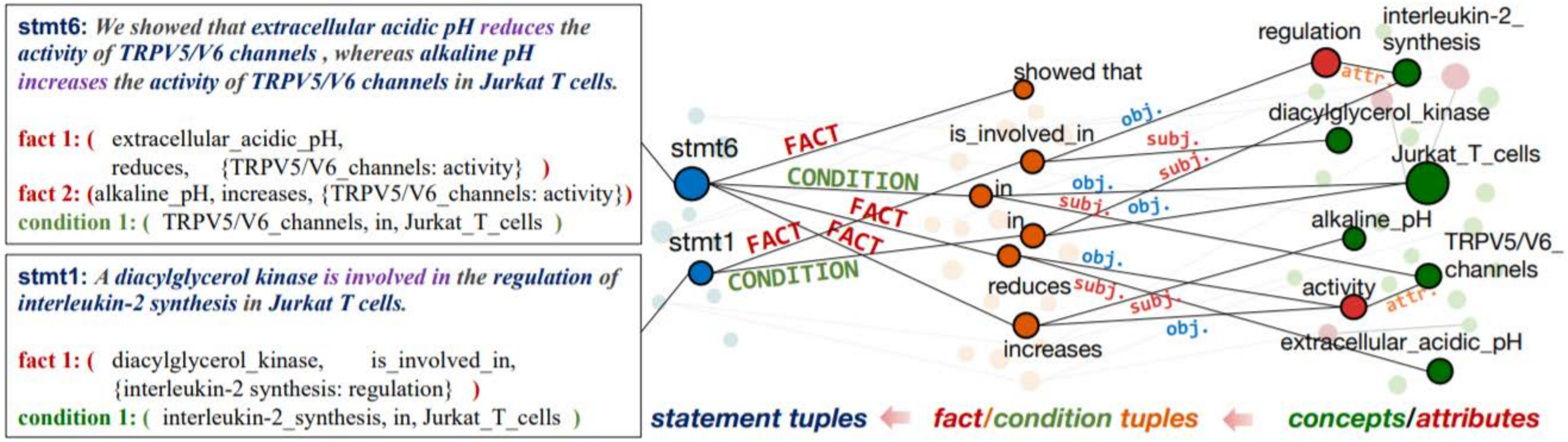
Skipgram

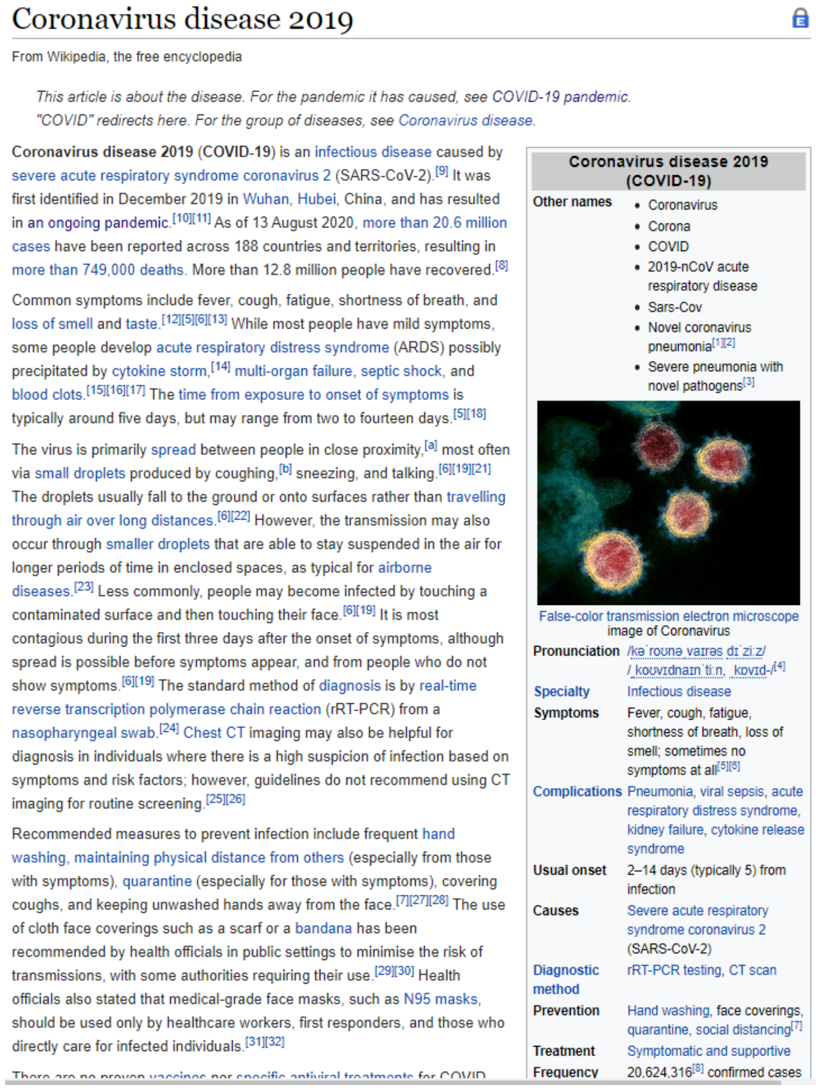
Also a lot of extensions to this approach

Bryan Perozzi, Rami Al-Rfou, Steven Skiena: DeepWalk: online learning of social representations. KDD 2014
Yuxiao Dong, Nitesh V. Chawla, Ananthram Swami: metapath2vec: Scalable Representation Learning for Heterogeneous Networks. KDD 2017
Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, Majid Sarrafzadeh: HeteroMed: Heterogeneous Information Network for Medical Diagnosis. CIKM 2018
Tingyi Wanyan et al., Heterogenous Graph Embeddings of Electronic Health Records Improve Critical Care Disease Predictions. AIME 2020

KG as an Attributed Network

- Attributes are important features of a knowledge graph
  - Wikipedia info box
  - Attribute knowledge extraction





Coronavirus disease 2019

From Wikipedia, the free encyclopedia

This article is about the disease. For the pandemic it has caused, see COVID-19 pandemic.
"COVID" redirects here. For the group of diseases, see Coronavirus disease.

Coronavirus disease 2019 (**COVID-19**) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[9] It was first identified in December 2019 in Wuhan, Hubei, China, and has resulted in an ongoing pandemic.[10][11] As of 13 August 2020, more than 20.6 million cases have been reported across 188 countries and territories, resulting in more than 749,000 deaths. More than 12.8 million people have recovered.[8]

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of smell and taste.[12][5][6][13] While most people have mild symptoms, some people develop acute respiratory distress syndrome (ARDS) possibly precipitated by cytokine storm,[14] multi-organ failure, septic shock, and blood clots.[15][16][17] The time from exposure to onset of symptoms is typically around five days, but may range from two to fourteen days.[5][18]

The virus is primarily spread between people in close proximity,[a] most often via small droplets produced by coughing,[b] sneezing, and talking.[6][19][21] The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances.[6][22] However, the transmission may also occur through smaller droplets that are able to stay suspended in the air for longer periods of time in enclosed spaces, as typical for airborne diseases.[23] Less commonly, people may become infected by touching a contaminated surface and then touching their face.[6][19] It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms.[6][19] The standard method of diagnosis is by real-time reverse transcription polymerase chain reaction (rRT-PCR) from a nasopharyngeal swab.[24] Chest CT imaging may also be helpful for diagnosis in individuals where there is a high suspicion of infection based on symptoms and risk factors; however, guidelines do not recommend using CT imaging for routine screening.[25][26]

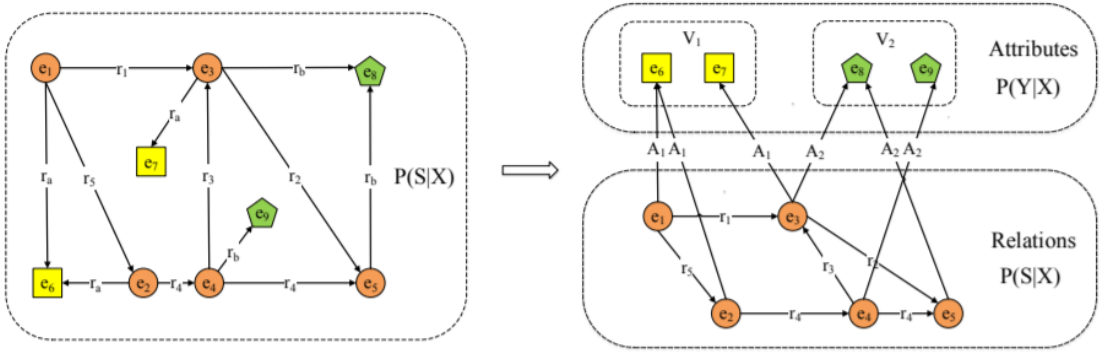
Recommended measures to prevent infection include frequent hand washing, maintaining physical distance from others (especially from those with symptoms), quarantine (especially for those with symptoms), covering coughs, and keeping unwashed hands away from the face.[7][27][28] The use of cloth face coverings such as a scarf or a bandana has been recommended by health officials in public settings to minimise the risk of transmissions, with some authorities requiring their use.[29][30] Health officials also stated that medical-grade face masks, such as N95 masks, should be used only by healthcare workers, first responders, and those who directly care for infected individuals.[31][32]

Coronavirus disease 2019 (COVID-19)	
Other names	Coronavirus; Corona; COVID; 2019-nCoV acute respiratory disease; Sars-Cov; Novel coronavirus pneumonia[1][2]; Severe pneumonia with novel pathogens[3]
False-color transmission electron microscope image of Coronavirus	
Pronunciation	/kəˈroʊnəˌvaɪrəs dɪˈziːz/ /ˌkoʊvɪdnaɪnˈtiːn, ˈkɒvɪd-/[4]
Specialty	Infectious disease
Symptoms	Fever, cough, fatigue, shortness of breath, loss of smell; sometimes no symptoms at all[5][6]
Complications	Pneumonia, viral sepsis, acute respiratory distress syndrome, kidney failure, cytokine release syndrome
Usual onset	2–14 days (typically 5) from infection
Causes	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
Diagnostic method	rRT-PCR testing, CT scan
Prevention	Hand washing, face coverings, quarantine, social distancing[7]
Treatment	Symptomatic and supportive
Frequency	20,624,316[8] confirmed cases

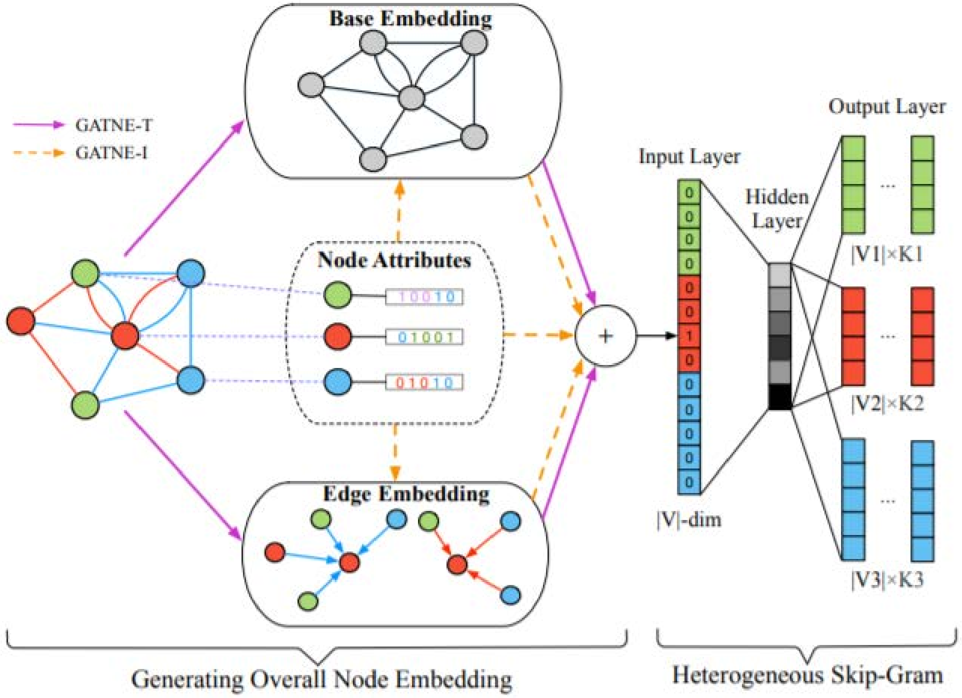
Figure from: Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V. Chawla, Meng Jiang: The Role of: A Novel Scientific Knowledge Graph Representation and Construction Model. KDD 2019

KG as a Attributed Network

- Attribute network embedding
  - Has been studied in data mining community for a few years
  - Traditional knowledge graph representation can be reformulated to attributed networks [Lin et al., 2016]
  - Little has been done for biomedical knowledge graphs but future work has been mentioned by existing embedding work [Yue et al., 2016]



[Lin et al., 2016]

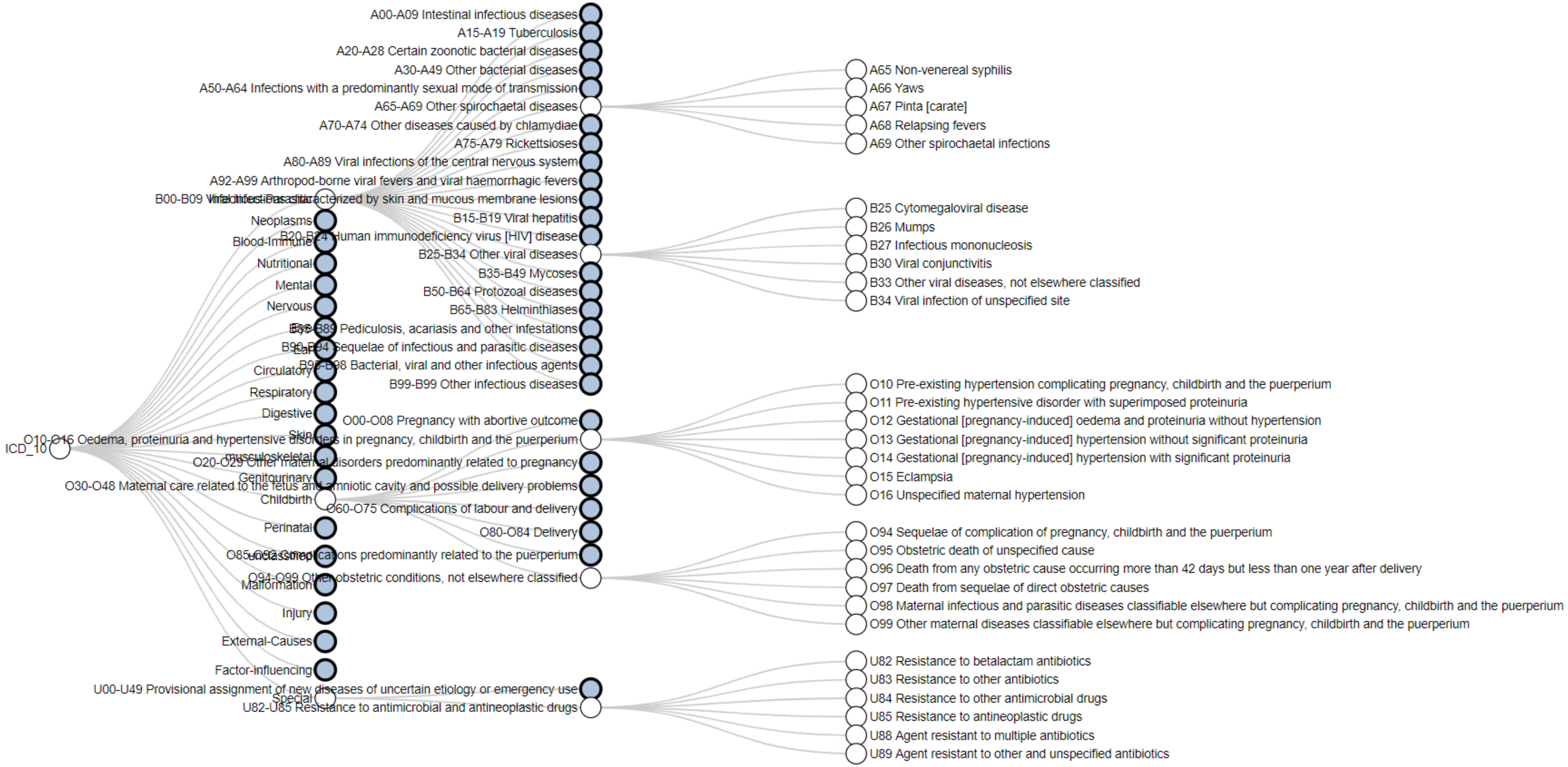


[Cen et al., 2016]

Figure from: Yankai Lin, Zhiyuan Liu, Maosong Sun: Knowledge Representation Learning with Entities, Attributes and Relations. IJCAI 2016
Figure from: Yukuo Cen, Xu Zou, Jianwei Zhang, Hongxia Yang, Jingren Zhou, Jie Tang: Representation Learning for Attributed Multiplex Heterogeneous Network. KDD 2019
Xiang Yue, et al.: Graph embedding on biomedical networks: methods, applications and evaluations. Bioinform. 36(4): 1241-1251 (2020)

KG as a Hierarchical Network

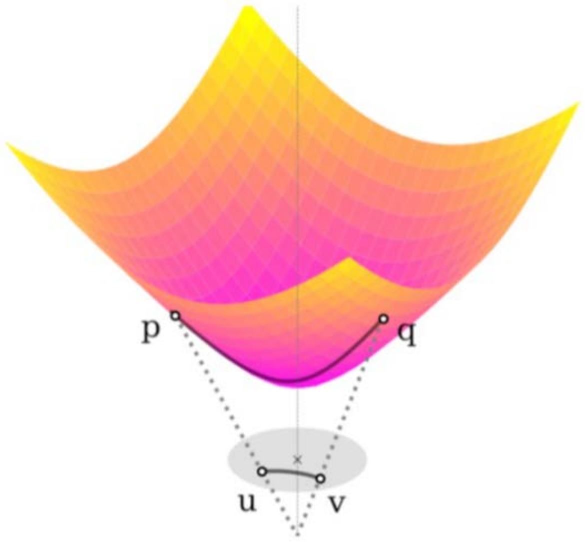
- Many knowledge graphs has tree-based categorizations
- An example of human diseases: the ICD10 Classification

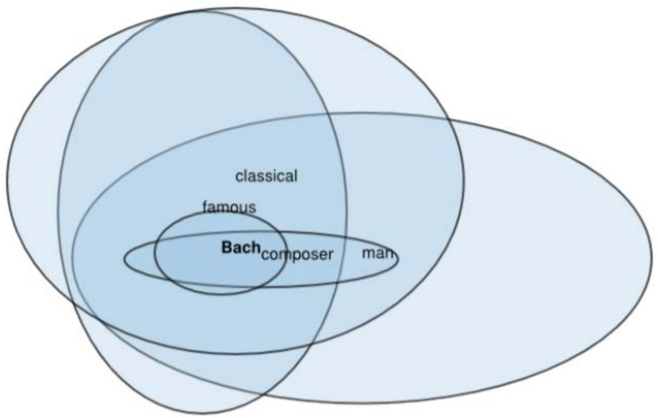


https://holtzy.github.io/Visualizing-the-ICD10-Classification/

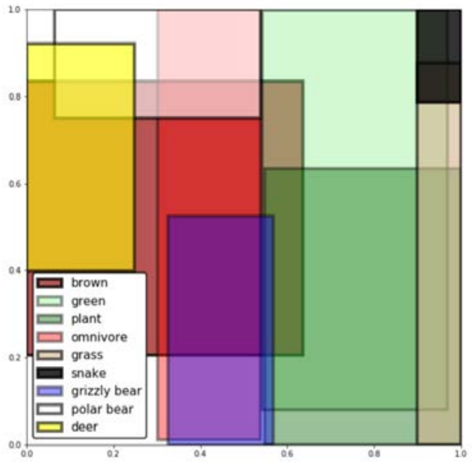
Hierarchical Network Embeddings

- Gaussian Embedding [Luke Vilnis, Andrew McCallum, 2015]
- Box Embedding [Luke Vilnis, et al., 2018]
- Hyperbolic Embedding [Nickel and Kiela, 2017]
  - Application in biomedical KG [Cao et al., 2020]
- …

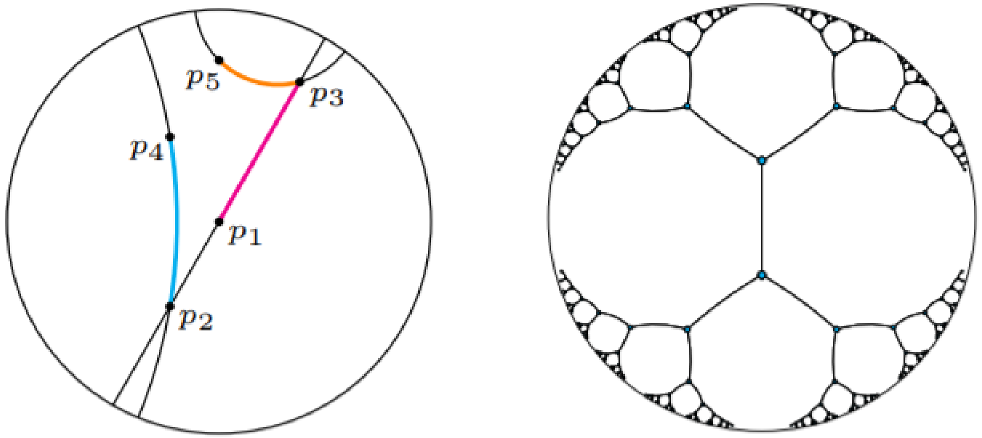




[Vilnis and McCallum, 2015]



[Vilnis et al., 2018]



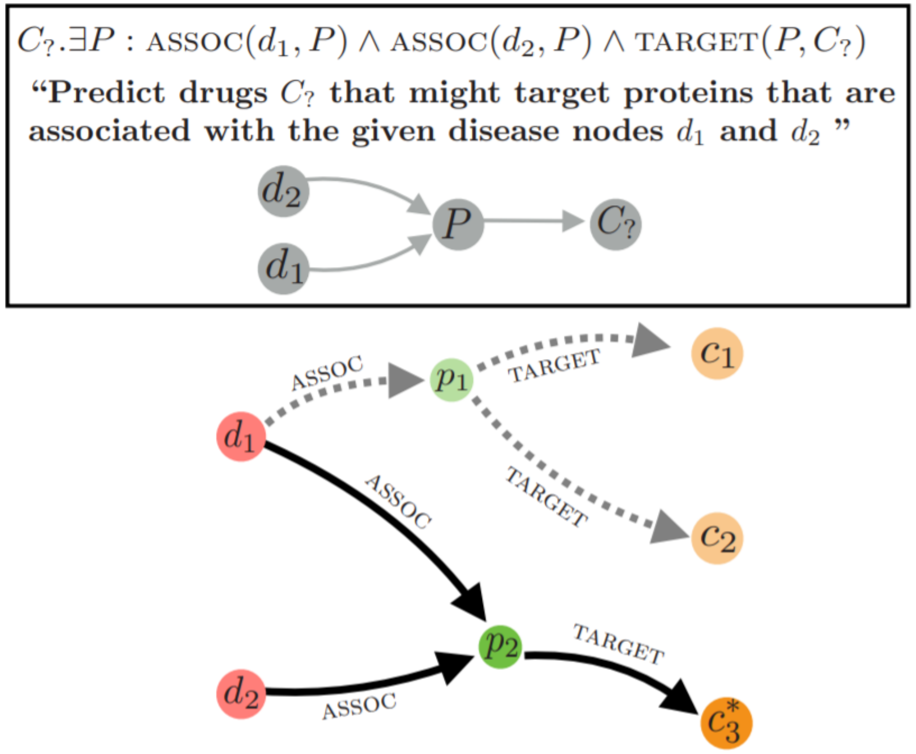
(a) Geodesics of the Poincaré disk (b) Embedding of a tree in $\mathcal{B}^2$

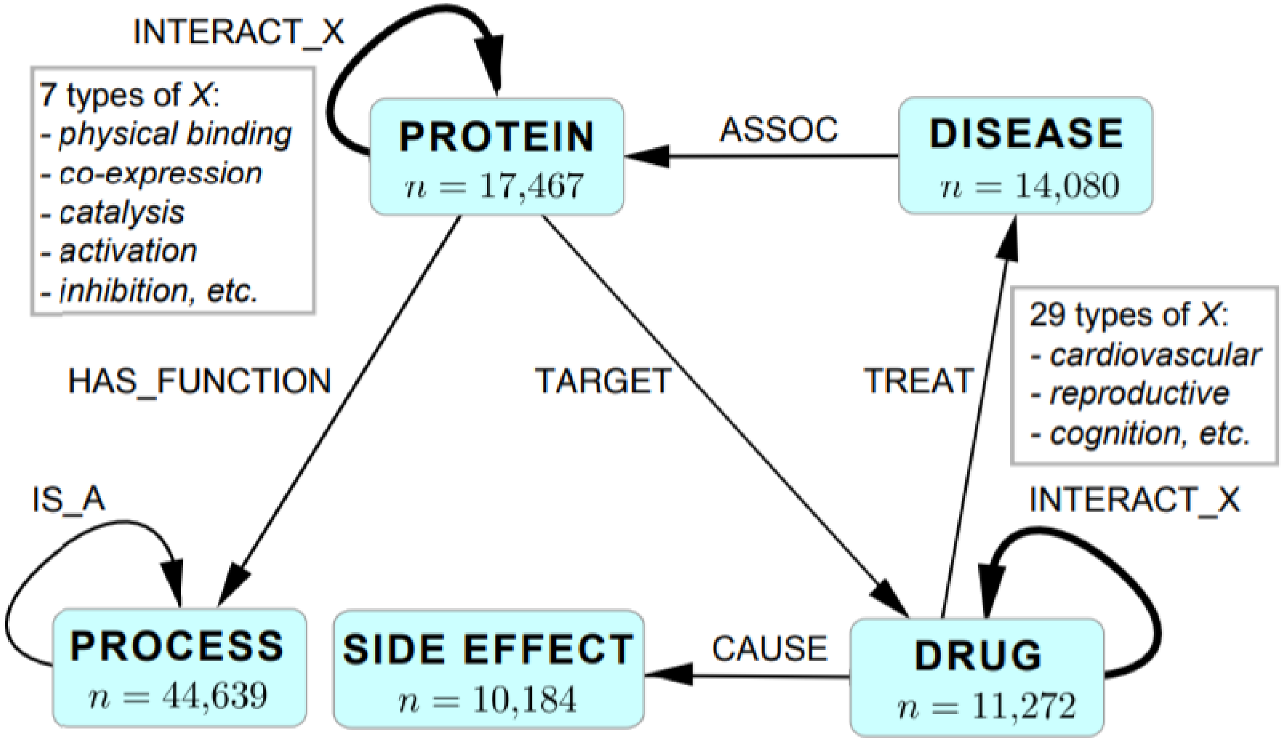
Poincaré Ball Model
[Nickel and Kiela, 2017]

Luke Vilnis, Andrew McCallum: Word Representations via Gaussian Embedding. ICLR 2015
Luke Vilnis, Xiang Li, Shikhar Murty, Andrew McCallum: Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures. ACL (1) 2018
Maximilian Nickel, Douwe Kiela: Poincaré Embeddings for Learning Hierarchical Representations. NIPS 2017: 6338-6347
Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, Weifeng Chong: HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. ACL 2020

New Applications: Answering Logical Queries

- With a KG, predict *what drugs are likely to target proteins involved with both diseases X and Y*?
  - Requires reasoning about all possible proteins that might interact with diseases X and Y

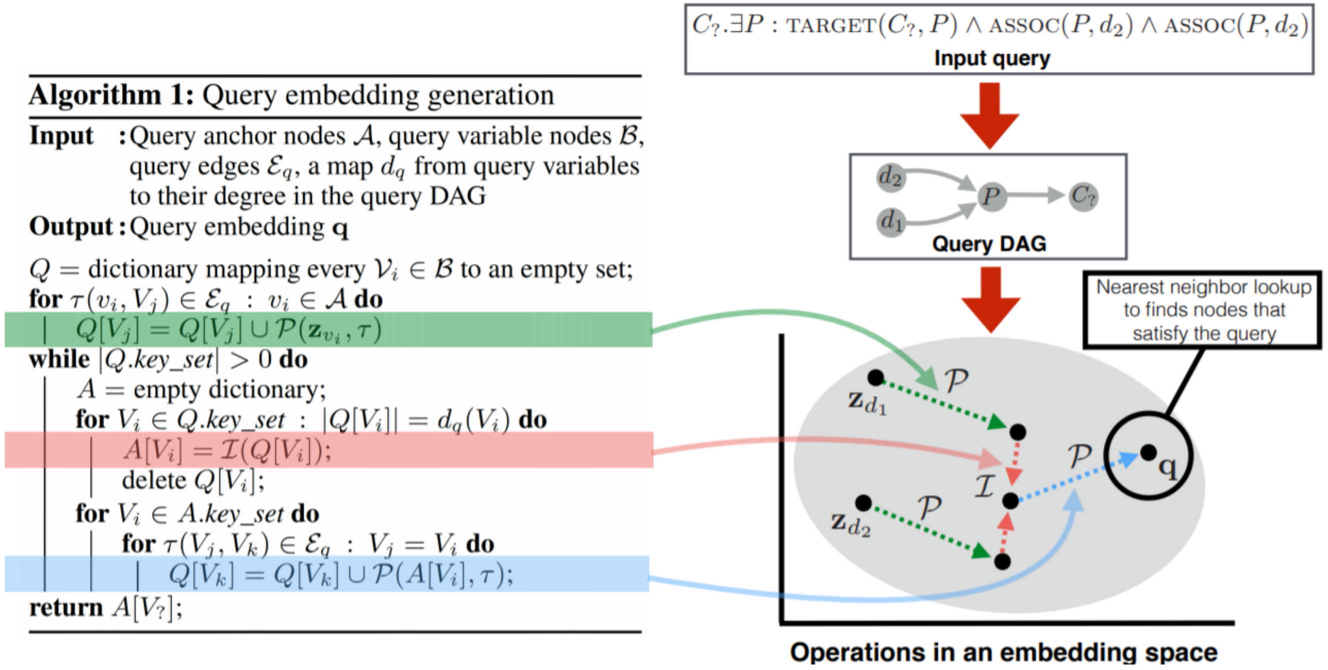




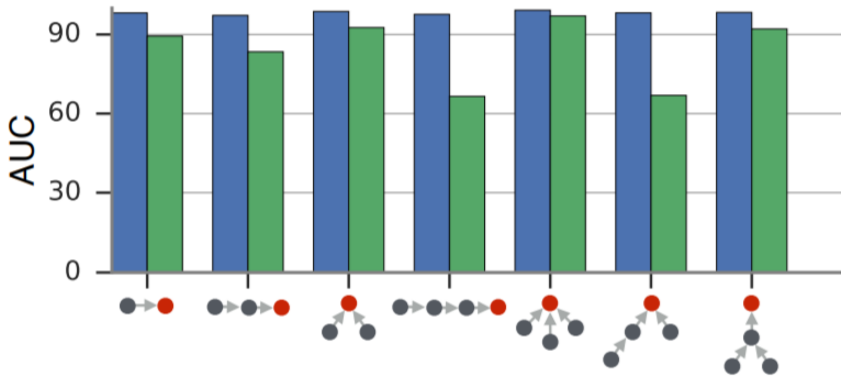
Figures from: William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, Jure Leskovec: Embedding Logical Queries on Knowledge Graphs. NeurIPS 2018

New Applications: Answering Logical Queries

- With a KG, predict *what drugs are likely to target proteins involved with both diseases X and Y*?
  - Encoding edge and path relationships in knowledge graphs: $\mathcal{P}$
  - Geometric intersection, $\mathcal{I}$, takes this set of query embeddings and produces a new embedding



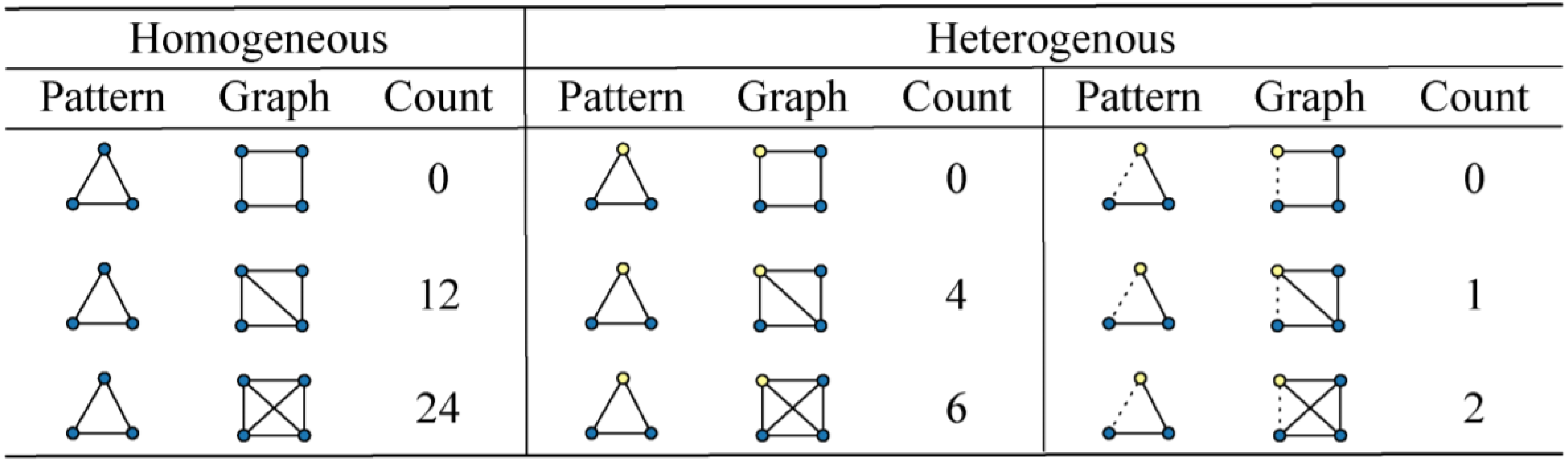
		Bio data		
		Bilinear	DistMult	TransE
GQE training	AUC	**91.0**	90.7	88.7
	APR	**91.5**	91.3	89.9
Edge training	AUC	79.2	86.7	78.3
	APR	78.6	87.5	81.6



Figures from: William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, Jure Leskovec: Embedding Logical Queries on Knowledge Graphs. NeurIPS 2018

New Applications: Subgraph Isomorphism Counting

- With a KG, predict *how many drugs are likely to target proteins involved with two coronavirus diseases in the knowledge graph*?
    - Requires scanning the whole graph and memorize all possible subgraphs involves one drug and two coronavirus diseases

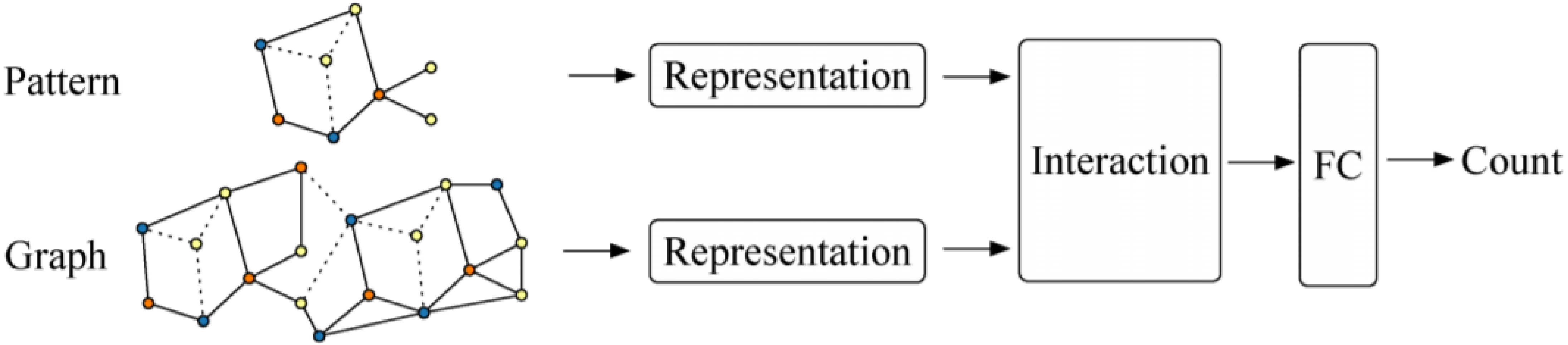


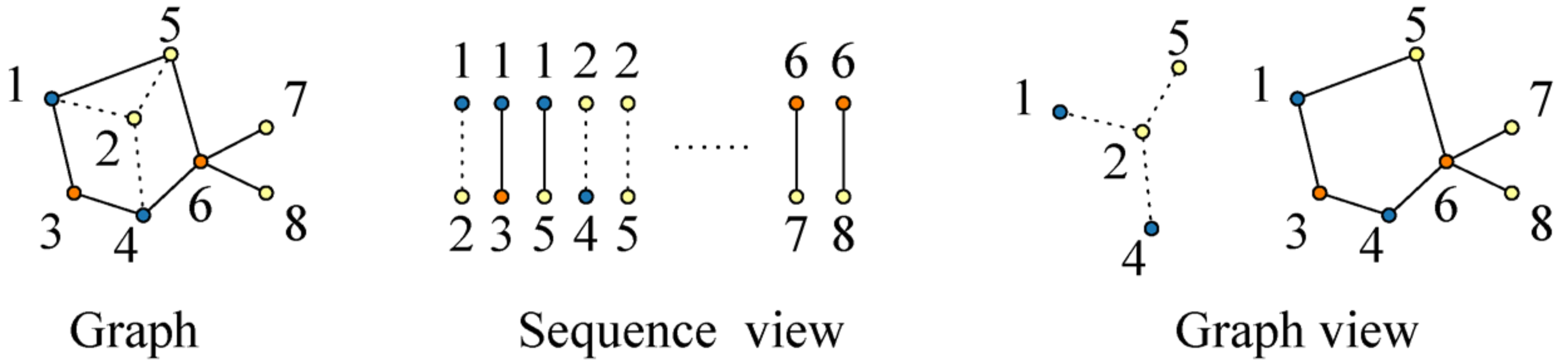
Homogeneous			Heterogenous					
Pattern	Graph	Count	Pattern	Graph	Count	Pattern	Graph	Count
		0			0			0
		12			4			1
		24			6			2

Figures from: Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, Lifeng Shang: Neural Subgraph Isomorphism Counting. KDD 2020

New Applications: Subgraph Isomorphism Counting

- A general QA model is applied to graph

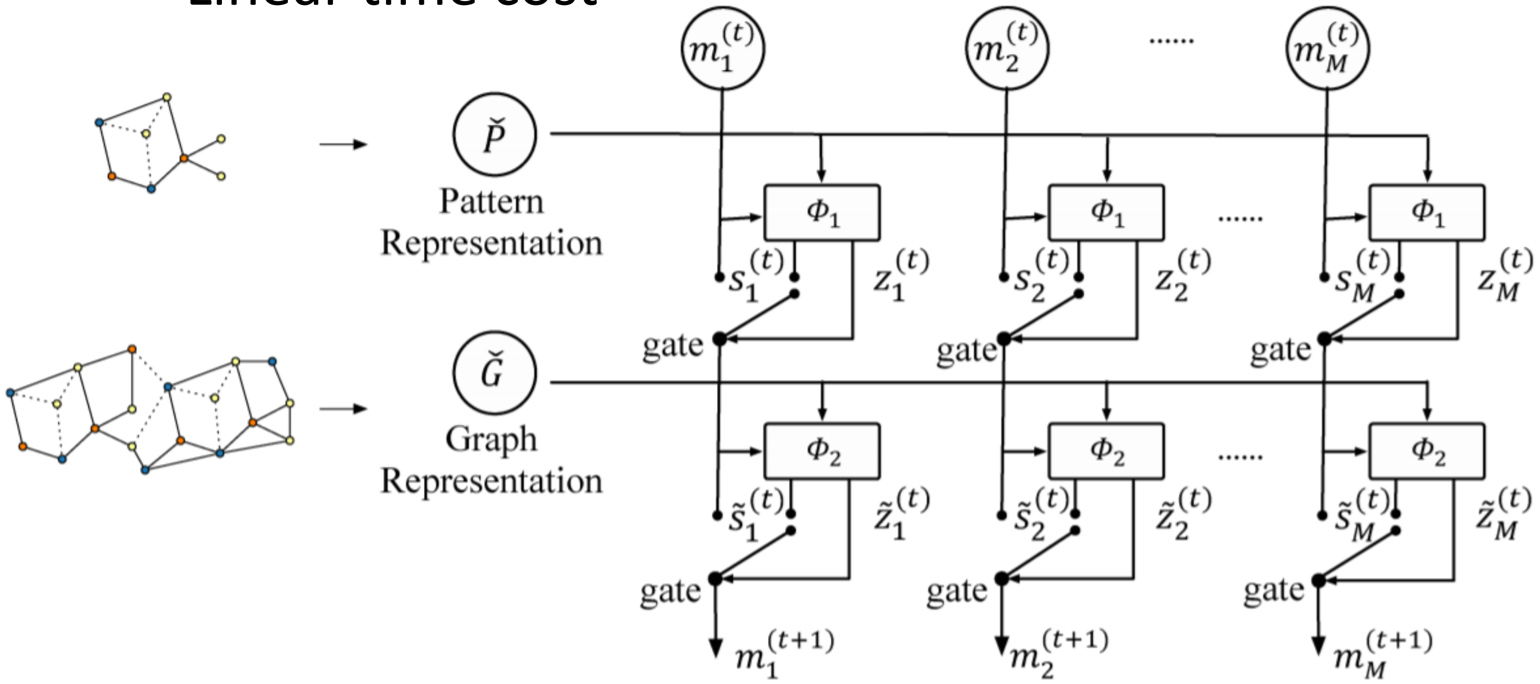




Figures from: Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, Lifeng Shang: Neural Subgraph Isomorphism Counting. KDD 2020

New Applications: Subgraph Isomorphism Counting

- Interactions between the query graph and target graph can be computationally costly
- An intermediate memory is used to iteratively attend query and target
  - Linear time cost



More details at: Session 6
Presentation - Thu 10 AM-12 PM
Poster Q&A - Wed 5-6 PM
Repeat for CET - Thu 5-6 AM
Graph Mining 3

Figures from: Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, Lifeng Shang: Neural Subgraph Isomorphism Counting. KDD 2020

Conclusions

- We covered knowledge graph related topics in the healthcare related domains
  - Knowledge graph construction
  - Knowledge graph learning and inference
- Many new technologies have been developed for
  - Natural language processing
    - Generation
    - Dialogue
    - Personalization
  - Recommendation
  - Drug discovery
  - …

Thank you for your attention! ☺