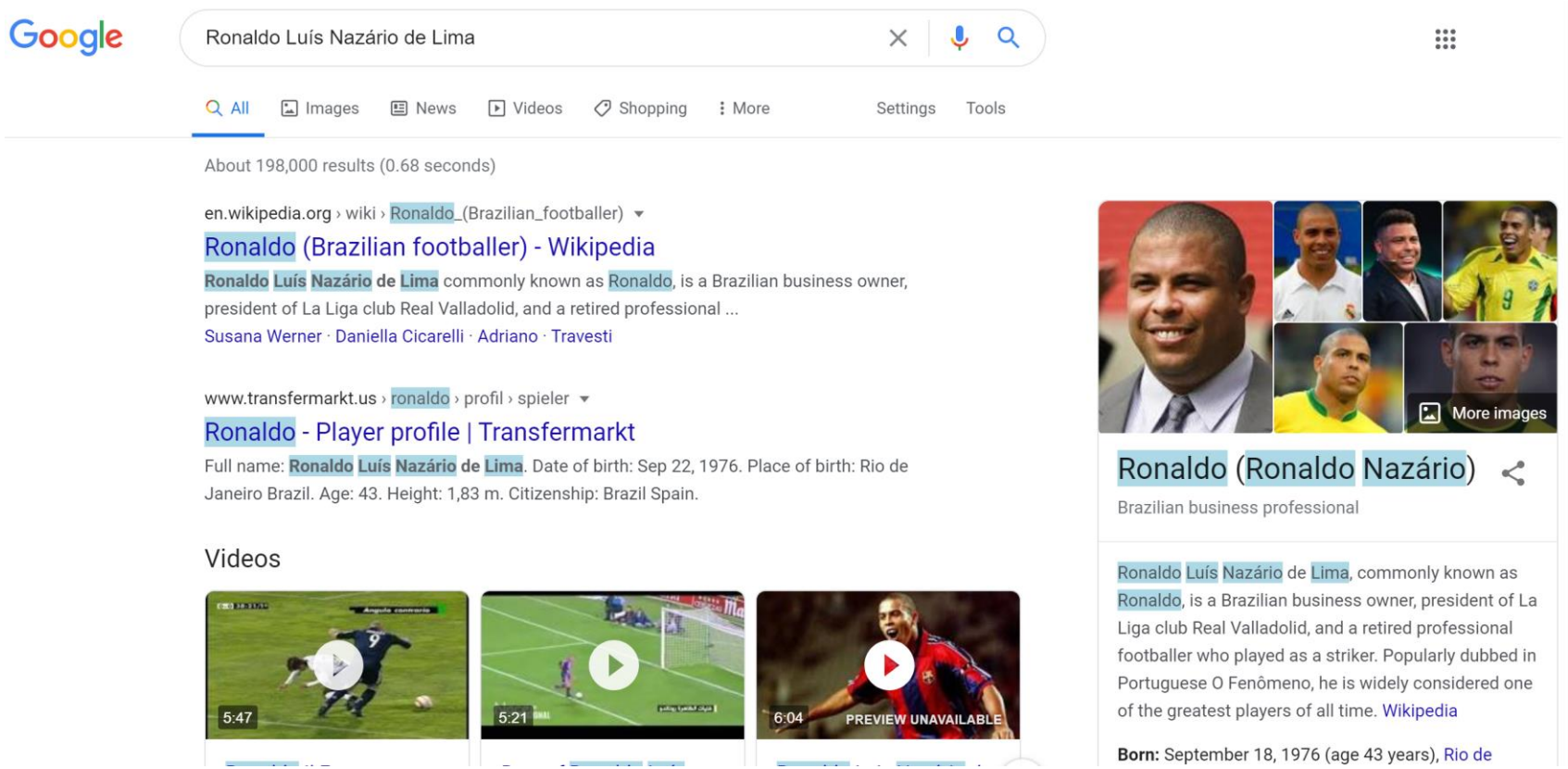
# KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences

导读人：付小康
武汉大学社会地理计算联合研究中心
whistle_ant@whu.edu.cn
主持人：叶信岳
德州农工大学

北京时间：2020年9月12日上午9:00-10:00点
U.S. Time: 9:00-10:00 PM, Sep 11, 2020 (Eastern Time)

# 研究背景

## Why knowledge graph



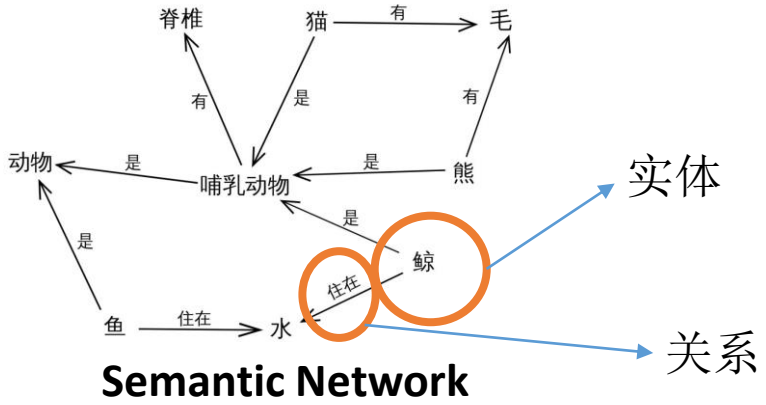

Google search and google knowledge graph

*Large knowledge bases (or graph) (KB’s) about entities, their properties, and the relationships between entities, have become an important asset for* ***semantic search, analytics, and smart recommendations over Web contents and other kinds of Big Data.***

# 研究背景

## What is knowledge graph

A **knowledge graph** is a knowledge base that uses a graph-structured data model or topology to integrate knowledge and data.

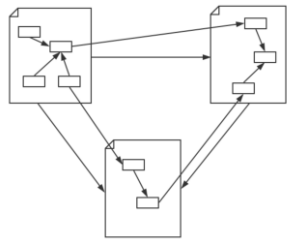



Semantic Network

## Other related concepts

**语义网络（Semantic Network）：**Quillian于上世纪60年代提出的知识表达模式，用相互连接的节点和边来表示知识。节点表示对象、概念，边表示节点之间的关系。

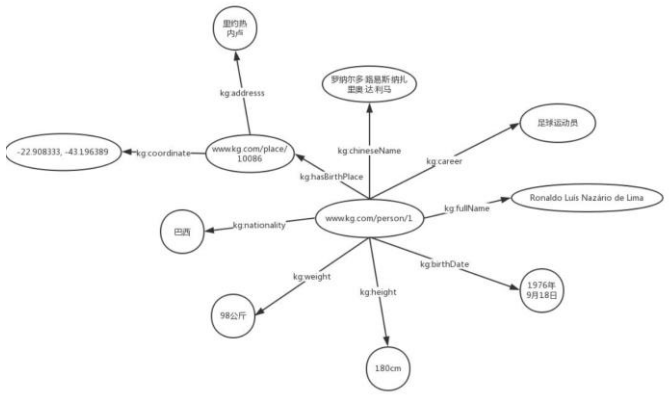
**语义网（ Semantic Web）和链接数据（ linked data ）：**语义网和链接数据是万维网之父Tim Berners Lee分别在1998年和2006提出的。相对于语义网络，语义网和链接数据倾向于描述万维网中资源、数据之间的关系。其实，本质上，语义网、链接数据还有Web 3.0都是同一个概念，只是在不同的时间节点和环境中，它们各自描述的角度不同。**它们都是指W3C制定的用于描述和关联万维网数据的一系列技术标准，即语义网技术栈**。



Semantic Web and linked data

**知识图谱(Knowledge Graph)：**知识图谱是对链接数据这个概念的进一步包装。两个区别在于:

1. 链接数据更强调不同RDF数据集（知识图谱）的相互链接。
2. 知识图谱不一定要链接到外部的知识图谱。另外，知识图谱数据质量要求比较高且要容易访问，要能够提供面向终端用户的信息服务（查询、问答等等）。




Knowledge Graph

# 研究背景

## Related work

**Universal knowledge bases**

- Dbpedia
- Yago
- Freebase（freebase.com）
- Google Knowledge Graph

**Domain knowledge bases**

- the Gene Ontology
- the Disease Ontology
- the National Drug File Reference Terminology
- the Foundational Model of Anatomy（解剖学）

## limitation

These biomedical domain KB covers only a relative **narrow topic** within the life sciences, and there is very **little interlinkage between the KB's**

# 研究背景

*In order to build **a comprehensive biomedical KB**, the following **three bottlenecks** must be addressed*

- **Beyond manual curation**

生物医学知识的发展速度远远超过任何一个人类所能吸收的速度。 因此，依靠手动管理创建知识图谱必将成为瓶颈。 为了充分利用所有已发布的知识，必须从输入文本中自动提取信息（ information extraction , IE）。

- **Beyond scientific literature**

以前有关生物医学IE的工作仅关注科学文献，而完全忽略了卫生门户网站和社区网站。

- **Beyond molecular (分子水平，主要指蛋白质) entities**

来自生物医学文献的IE多关注分子水平上的实体和关系，典型任务是提取蛋白质之间的相互作用。关于将各种实体类型，跨越基因，疾病，症状，解剖部位，药物，药物作用等联系起来的综合方法的工作很少。特别是，以前关于知识库构建的工作都没有涉及环境和生活方式等风险因素方面。

# 研究背景

## Knowlife Contributions

- **Beyond manual curation**

通过模式匹配结合逻辑一致性约束的方式进行实体的自动抽取

- **Beyond scientific literature**

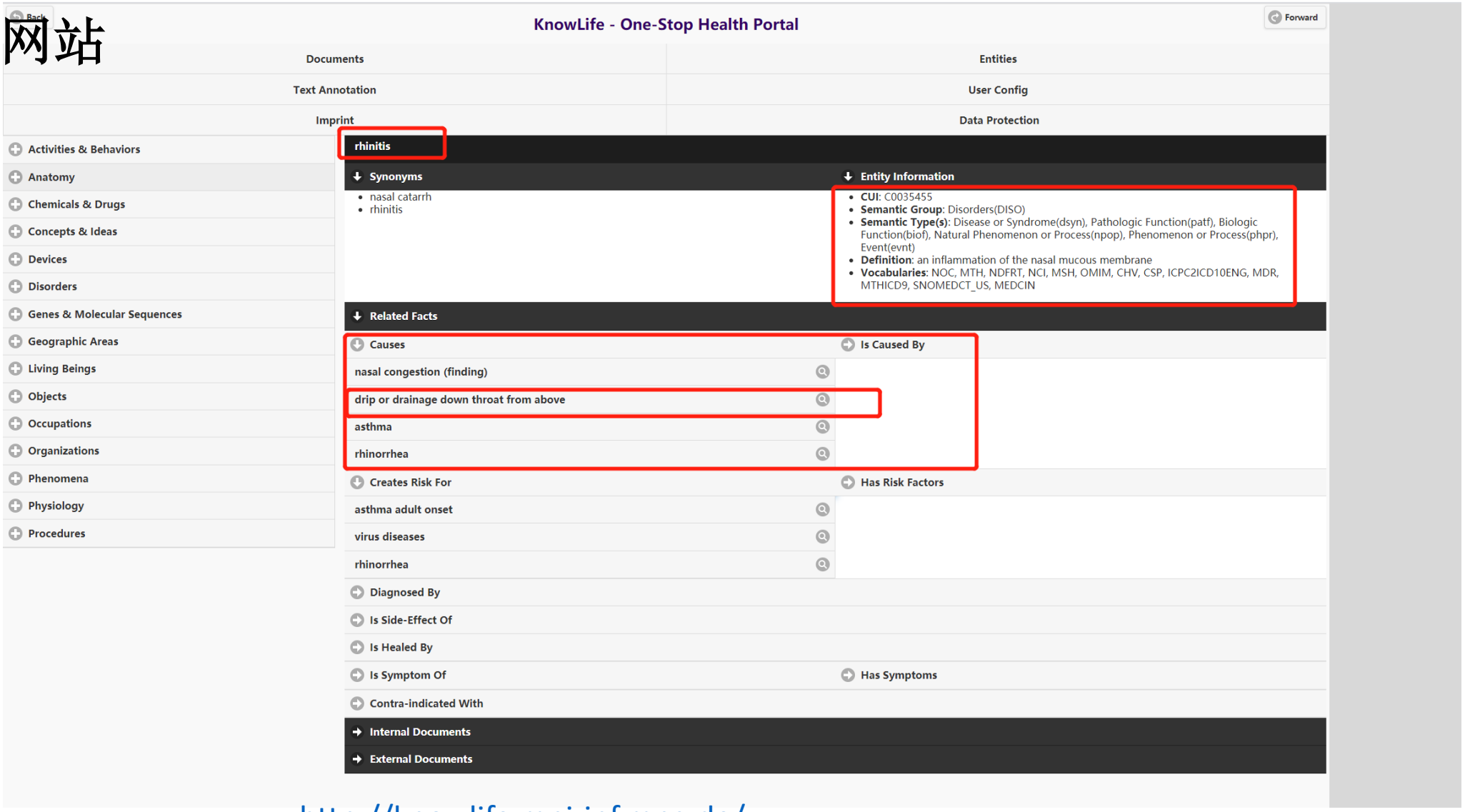
不只是考虑了科学文献，也加入了相关门户网站和社区网站，对不同的语料库所得到知识图谱的精度进行评估和讨论。

- **Beyond molecular (分子水平，主要指蛋白质) entities**

不只是包含基因和蛋白质类型的实体，还包含疾病，症状，解剖部位，药物，药物作用，环境和生活方式等。

# 研究背景

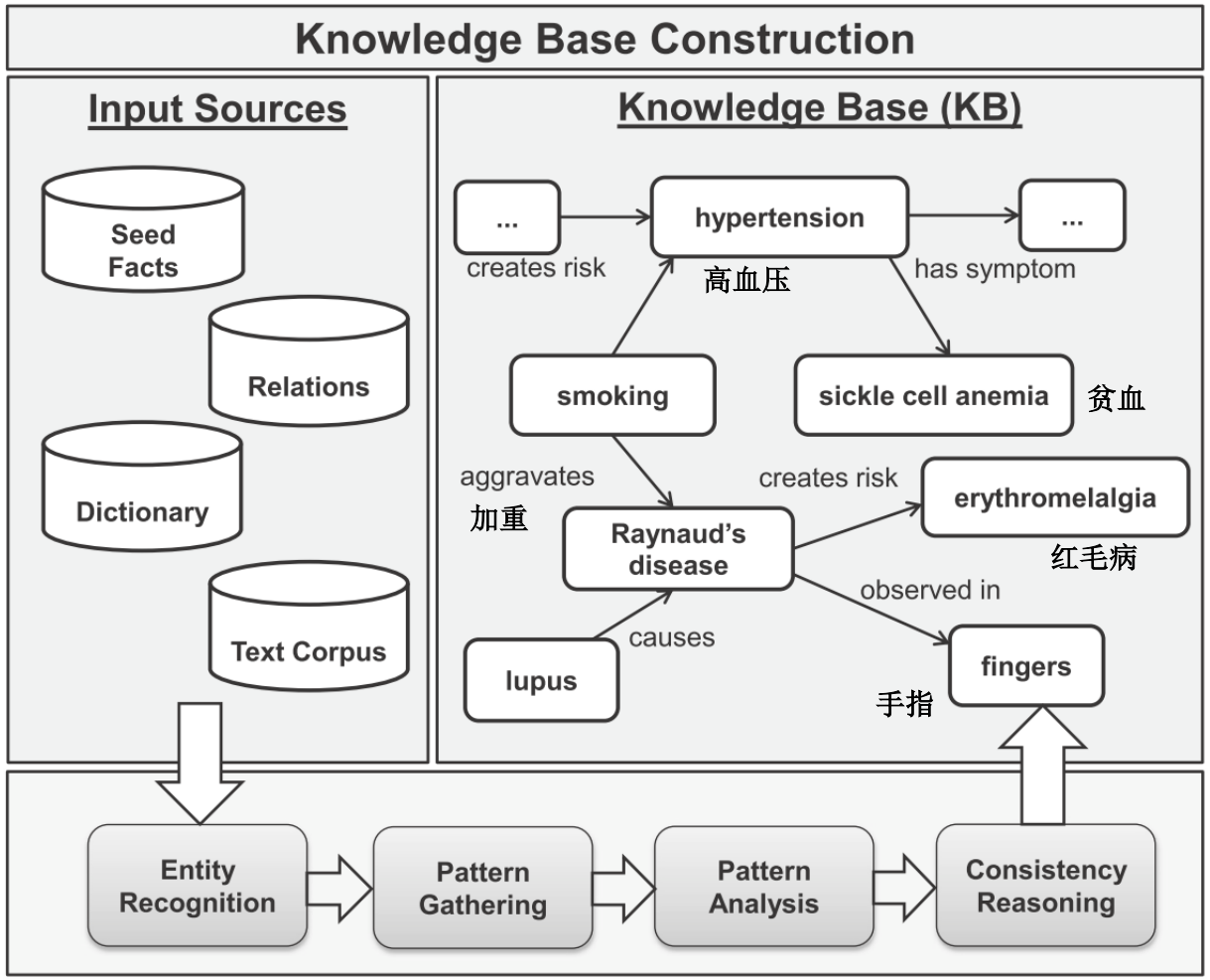
## Knowlife门户网站




http://knowlife.mpi-inf.mpg.de/

# 研究方法

## Overview of the KnowLife KB and processing pipeline




Ernst, P., Siu, A., Weikum, G., 2015. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 16, 157. https://doi.org/10/gb8w8d

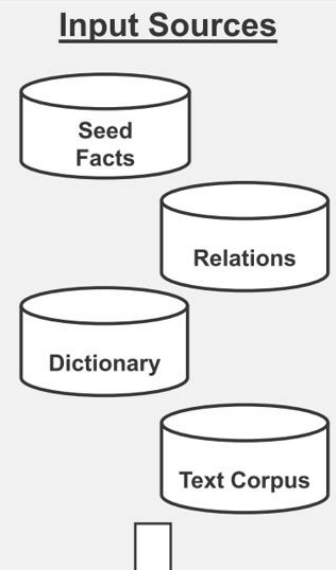
Pipline target: 从各种数据中提取事实（facts），即可以构建知识图谱，事实可以表示为：

R(e1, e2 )

其中，e1, e1表示实体，R表示实体e1和实体e2之间的关系

# 研究方法

## Input sources




### Dictionary

美国国立医学图书馆（英文：National Library of Medicine，NLM）设计了并负责维护着UMLS。UMLS每季度更新一次，且可以免费使用。该项目最初是由Donald Lindberg博士于1986年发起的（Donald Lindberg后来担任了国立医学图书馆馆长）

我们使用**UMLS（统一医学语言系统）**作为生物医学实体的词典。UMLS是一个词库，是生物医学词典的最大集合，包含290万个实体以及1,140万个实体名称和同义词。UMLS词典使KnowLife能够检测文本中的实体，这些实体不仅仅涉及基因和蛋白质，还涉及有关解剖，生理学和疗法的实体。

### Relations

**Table 1 KnowLife relations, their type signatures, and number of seeds**

| Relation | Domain | Range | Seed facts |
|---|---|---|---|
| Affects | Disease | Organ | 23 |
| Aggravates | Ecofactor | Disease | 21 |
| Alleviates | Drug | Disease | 18 |
| Causes | Disease | Disease | 70 |
| ComplicationOf | Disease | Disease | 5 |
| Contraindicates | Drug | Disease | 26 |
| CreatesRisk | Ecofactor | Disease | 103 |
| Diagnoses | Device | Disease | 29 |
| Interacts | Drug | Drug | 9 |
| IsSymptom | Symptom or Disease | Disease | 69 |
| ReducesRisk | Drug or Behavior | Disease | 24 |
| SideEffect | Symptom or Disease | Drug | 12 |
| Treats | Drug | Disease | 58 |

### Seed facts

A seed fact R(e1, e2 ) for relation R is a triple presumed to be true based on expert statements. We collected 467 seed facts from the medical online portal uptodate.com

### Text Corpus

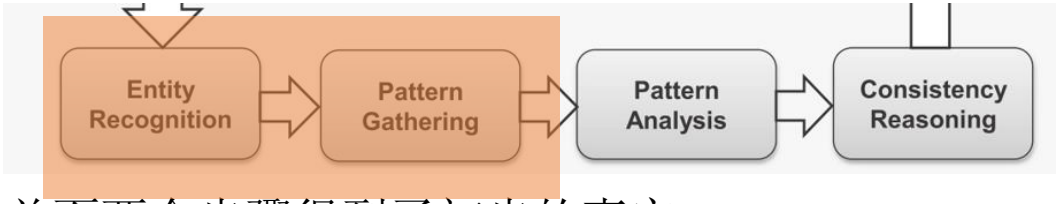
**Table 2 Overview of KnowLife's input corpus**

| Genre | Source | Documents | Sentences |
|---|---|---|---|
| Scientific Publications | PubMed Medline | 580,892 | 5,875,006 |
| | PubMed Central | 12,532 | 2,765,580 |
| Encyclopedic Articles | Drugs.com | 31,837 | 7,586,236 |
| | Mayo Clinic | 2,166 | 570,325 |
| | Medline Plus | 3,076 | 197,055 |
| | RxList | 2,515 | 1,102,791 |
| | Wikipedia Health | 20,893 | 787,148 |
| Social Sources | Healthboards.com | 752,778 | 37,270,371 |
| | Patient.co.uk | 44,610 | 1,081,420 |
| | **Total** | **1,451,299** | **57,235,932** |

We use **the Stanford CoreNLP software** to preprocess all texts, such that they are tokenized, split into sentences, tagged with parts-of-speech, lemmatized, and parsed into syntactic dependency graphs.

Ernst, P., Siu, A., Weikum, G., 2015. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 16, 157. https://doi.org/10/gb8w8d

# 研究方法




Ernst, P., Siu, A., Weikum, G., 2015. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 16, 157. https://doi.org/10/gb8w8d

前面两个步骤得到了初步的事实

## Entity recognition

KnowLife管道中的第一阶段**识别可以表达关系事实的句子**。

对句子中的实体进行识别，主要是和UMLS词典中的字符串进行匹配。（快速匹配算法）

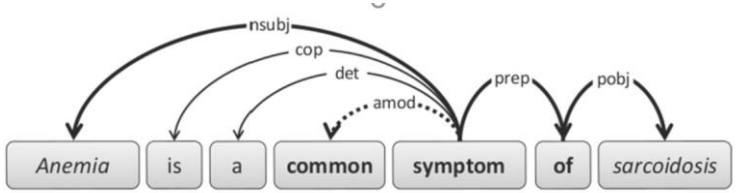
- *Anemia* is a common symptom of *sarcoidosis*.
- Eventually, a *heart attack* leads to *arrythmias*.
- Ironically, a *myocardial infarction* can also lead to *pericarditis*.

心律失常

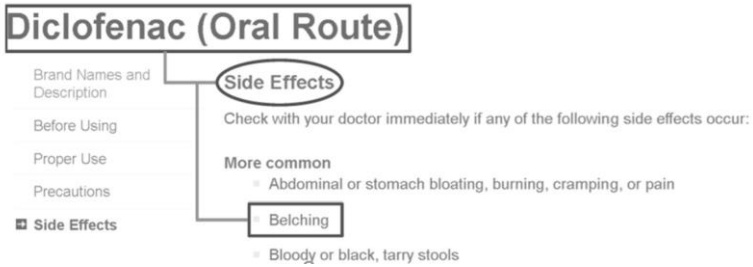
实体

## Pattern gathering

knowlife方法通过句子的句法结构或网页DOM（文档对象模型）树中的路径来提取连接两个已识别实体的文本模式。 其中包含两种类型的模式：

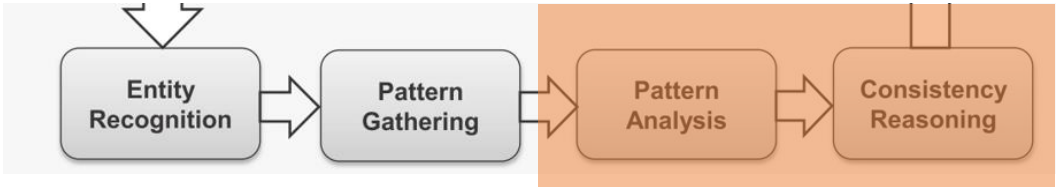



Sentence-level **Patterns**




Document-structure **Patterns**

# 研究方法




Ernst, P., Siu, A., Weikum, G., 2015. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 16, 157. https://doi.org/10/gb8w8d

后面两个步骤对得到的事实进行进一步的筛选

## Pattern analysis

模式分析的目的是从到目前为止收集的所有模式候选中识别出最有用的**种子模式**。我们利用Prospera工具中开发的技术进行该步骤。

confidence, such that the confidence for a pattern $q$ in a set of sentences $S$ is defined as

$$confidence(q) = \frac{|\{s \in S \mid \exists (e_1, e_2) \in SX(R_i)\ \ q, e_1, e_2\ occur\ in\ s\}|}{|\{s \in S \mid \exists (e_1, e_2) \in SX(R_i) \cup CX(R_i)\ \ q, e_1, e_2\ occur\ in\ s\}|}$$

模式与特定关系的种子事实实体之间的关联越强，我们就越有信心该模式表达该关系。 选择置信度大于阈值（在我们的实验中设置为0.3）的模式作为种子模式。

## Consistency reasoning

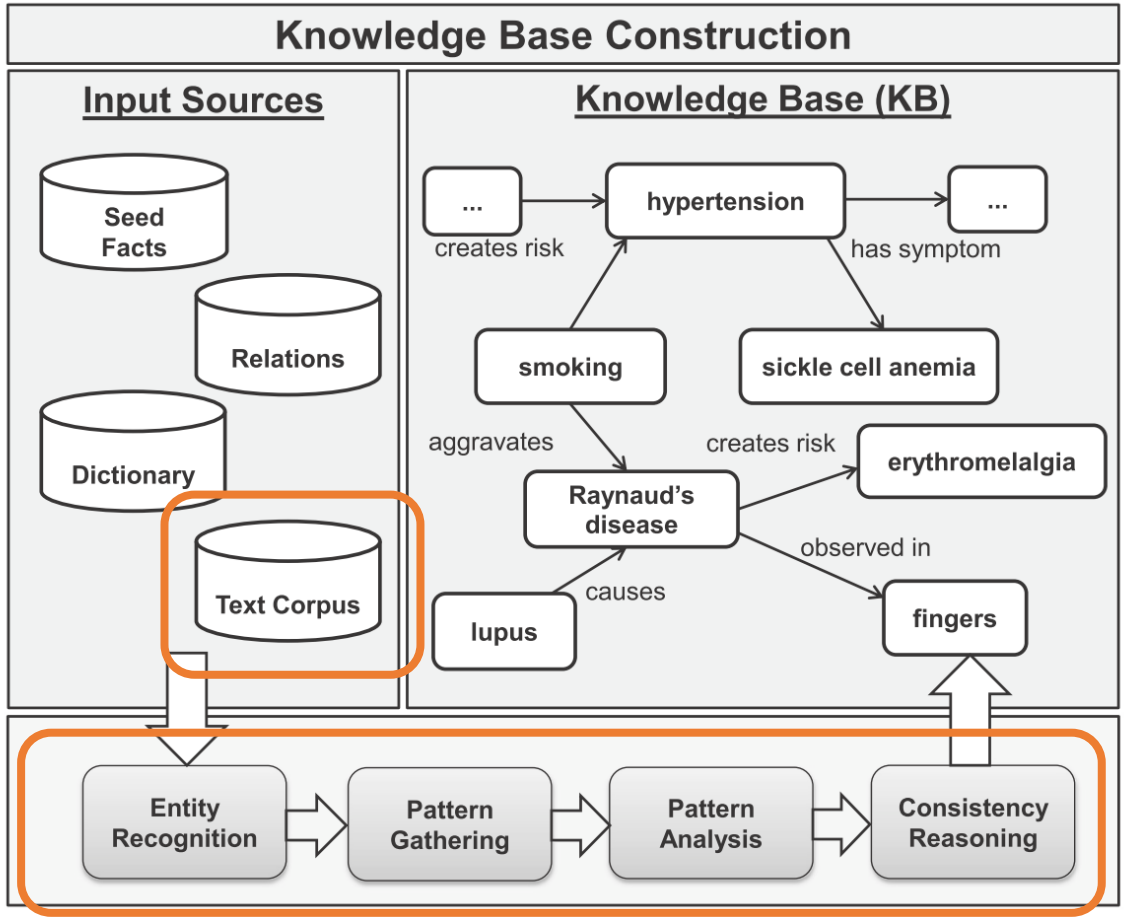
为了减少这些错误并提高准确性，KnowLife的最后阶段将逻辑一致性约束应用于事实候选者，并仅接受它们的一致子集。

We leverage two kinds of manually defined semantic constraints: i) the type signatures of relations (see Table 1) for type checking of fact candidates, and ii) mutual exclusion constraints between certain pairs of relations.

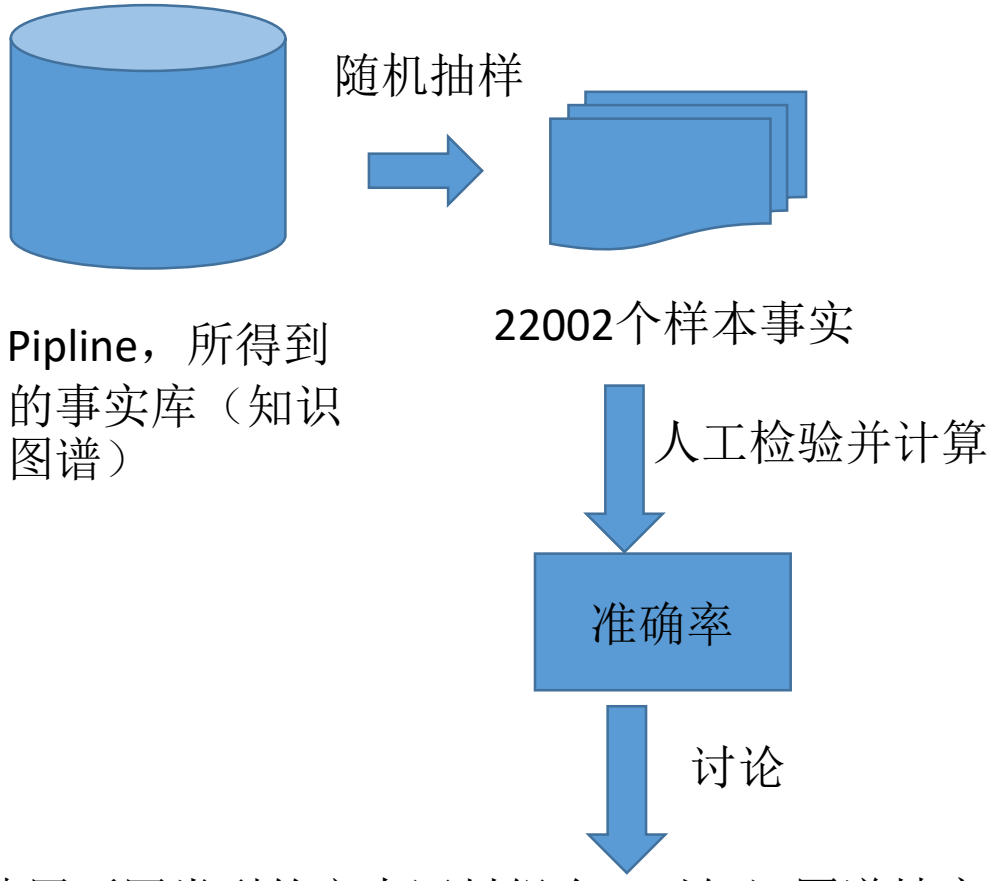
如果药物有某种症状作为副作用，则不能同时治疗该症状。

# 结果解读

## 知识图谱评估方法




Overview of the KnowLife KB and processing pipeline




1. 使用不同类型的文本语料组合，对知识图谱精度有什么影响
2. 缺失某个数据处理步骤，对知识图谱精度有什么影响

# 结果解读

1. 使用不同类型的文本语料组合，对知识图谱精度有什么影响

**Table 2 Overview of KnowLife's input corpus**

| Genre | Source | Documents | Sentences |
|---|---|---|---|
| Scientific Publications | PubMed Medline | 580,892 | 5,875,006 |
| | PubMed Central | 12,532 | 2,765,580 |
| Encyclopedic Articles | Drugs.com | 31,837 | 7,586,236 |
| | Mayo Clinic | 2,166 | 570,325 |
| | Medline Plus | 3,076 | 197,055 |
| | RxList | 2,515 | 1,102,791 |
| | Wikipedia Health | 20,893 | 787,148 |
| Social Sources | Healthboards.com | 752,778 | 37,270,371 |
| | Patient.co.uk | 44,610 | 1,081,420 |
| | **Total** | **1,451,299** | **57,235,932** |

所使用语料类型

使用不同类型的文本语料组合所得到的知识图谱精度

**Table 4 Evaluation of different text genres**

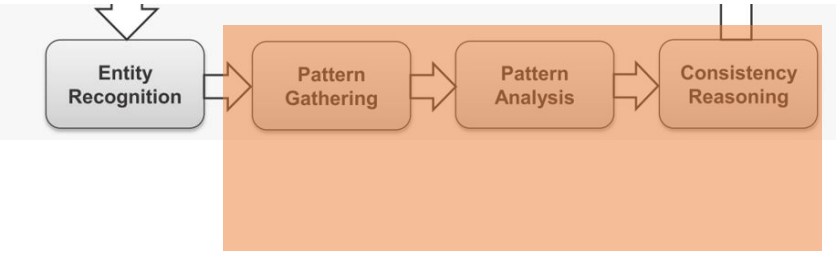
| Relation | Precision | | | | Harvested facts | | | |
|---|---|---|---|---|---|---|---|---|
| | Encyclopedic sources | Scientific sources | Encyclopedic + scientific sources | Encyclopedic + scientific + social sources | Encyclopedic sources | Scientific sources | Encyclopedic + scientific sources | Encyclopedic + scientific + social sources |
| Affects | 0.855±0.047 | 0.762±0.049 | **0.825±0.047** | 0.767±0.048 | 1,278 | 450 | **2,388** | 5,053 |
| Aggravates | 0.810±0.041 | 0.459±0.044 | **0.829±0.049** | 0.785±0.049 | 130 | 371 | **432** | 708 |
| Alleviates | 0.953±0.039 | 0.735±0.048 | **0.786±0.046** | 0.736±0.048 | 903 | 4,433 | **4,530** | 6,790 |
| Causes | 0.904±0.039 | 0.674±0.049 | **0.801±0.049** | 0.792±0.049 | 28,119 | 19,203 | **47,463** | 62,407 |
| Complication | 0.917±0.039 | 0.397±0.049 | **0.897±0.041** | 0.869±0.046 | 1,011 | 1,475 | **1,524** | 1,566 |
| Contraindicates | 0.874±0.048 | 0.710±0.000 | **0.961±0.030** | 0.908±0.048 | 512 | 49 | **1,808** | 1,831 |
| CreatesRisk | 0.878±0.047 | 0.569±0.049 | **0.720±0.040** | 0.620±0.049 | 4,407 | 24,695 | **18,508** | 32,211 |
| Diagnoses | 0.964±0.035 | 0.839±0.049 | **0.860±0.048** | 0.840±0.047 | 813 | 5,920 | **4,832** | 9,743 |
| Interacts | 0.964±0.035 | 0.709±0.000 | **0.965±0.034** | 0.957±0.034 | 164,912 | 103 | **164,912** | 164,912 |
| IsSymptom | 0.891±0.042 | 0.482±0.050 | **0.858±0.048** | 0.694±0.048 | 4,878 | 2,320 | **6,395** | 11,017 |
| ReducesRisk | 0.797±0.045 | 0.637±0.046 | **0.762±0.048** | 0.751±0.049 | 1,712 | 4,684 | **4,489** | 5,865 |
| SideEffect | 0.956±0.038 | 0.826±0.000 | **0.964±0.035** | 0.971±0.026 | 270,600 | 139 | **270,709** | 271,416 |
| Treats | 0.850±0.048 | 0.581±0.045 | **0.898±0.041** | 0.566±0.048 | 11,915 | 9,318 | **14,699** | 35,803 |
| Aggregated* | 0.951 | 0.630 | **0.933** | 0.892 | 491,190 | 73,160 | **542,689** | 609,322 |

*Precision values are averaged and numbers of harvested facts are summed.

# 结果解读

2. 缺失某个数据处理步骤，对知识图谱精度有什么影响

数据处理步骤




缺失某个数据处理步骤所得到的知识图谱精度

**Table 5 Evaluation of the impact of different components**

| Relation | Precision | | | | Harvested facts | | | |
|---|---|---|---|---|---|---|---|---|
| | Full pipeline encyclopedic + scientific sources | Without document structure | Without statistical analysis | Without consistency reasoning | Full pipeline encyclopedic + scientific sources | Without document structure | Without statistical analysis | Without consistency reasoning |
| Affects | 0.825±0.047 | 0.882±0.044 | 0.821±0.048 | 0.171±0.051 | 2,388 | 2,350 | 4,088 | 29,477 |
| Aggravates | 0.829±0.049 | 0.833±0.036 | 0.598±0.049 | 0.592±0.053 | 432 | 431 | 592 | 1,730 |
| Alleviates | 0.786±0.046 | 0.778±0.050 | 0.320±0.049 | 0.289±0.062 | 4,530 | 4,387 | 18,142 | 16,943 |
| Causes | 0.801±0.049 | 0.800±0.046 | 0.631±0.048 | 0.490±0.069 | 47,463 | 30,563 | 66,833 | 91,784 |
| Complication | 0.897±0.041 | 0.781±0.048 | 0.376±0.050 | 0.739±0.050 | 1,524 | 700 | 4,812 | 2,955 |
| Contraindicates | 0.961±0.030 | 0.914±0.043 | 0.122±0.049 | 0.630±0.059 | 1,808 | 365 | 26,298 | 15,279 |
| CreatesRisk | 0.720±0.040 | 0.750±0.044 | 0.386±0.047 | 0.406±0.067 | 18,508 | 17,282 | 77,158 | 48,159 |
| Diagnoses | 0.860±0.048 | 0.887±0.044 | 0.802±0.049 | 0.303±0.063 | 4,832 | 4,002 | 7,467 | 35,326 |
| Interacts | 0.965±0.034 | 0.858±0.046 | 0.953±0.047 | 0.941±0.049 | 164,912 | 392 | 200,935 | 187,201 |
| IsSymptom | 0.858±0.048 | 0.691±0.050 | 0.625±0.049 | 0.328±0.064 | 6,395 | 2,920 | 9,543 | 29,776 |
| ReducesRisk | 0.762±0.048 | 0.729±0.050 | 0.228±0.046 | 0.406±0.067 | 4,489 | 4,043 | 11,023 | 14,729 |
| SideEffect | 0.964±0.035 | 0.938±0.048 | 0.941±0.046 | 0.879±0.050 | 270,709 | 924 | 270,427 | 338,645 |
| Treats | 0.898±0.041 | 0.784±0.050 | 0.549±0.050 | 0.402±0.067 | 14,699 | 14,057 | 23,473 | 45,439 |
| Aggregated* | 0.933 | 0.784 | 0.777 | 0.707 | 542,689 | 82,416 | 720,791 | 857,443 |

*Precision values are averaged and numbers of harvested facts are summed.

# 总结与讨论

- 论文的主要创新在于：（1）Beyond manual curation；（2）Beyond scientific literature；（3）Beyond molecular (分子水平，主要指蛋白质) entities
- 论文指出了清晰的pipline和相关工具，可借鉴性、可复制性强
- 论文局限之处在和未来可能研究方向：（1）实体依赖于UMLS（质量依赖、涵盖范围依赖）（2）二元关系不能表达全部医学知识

*发烧*是怀孕期间*狼疮耀斑*的症状

# Pipline复制工作

- Python
- Knime

# 参考文献


- Ernst, P., Meng, C., Siu, A., Weikum, G., 2014. KnowLife: A knowledge graph for health and life sciences, in: 2014 IEEE 30th International Conference on Data Engineering. Presented at the 2014 IEEE 30th International Conference on Data Engineering, pp. 1254–1257. https://doi.org/10/gg8678
- Ernst, P., Siu, A., Weikum, G., 2015. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 16, 157. https://doi.org/10/gb8w8d
- https://zhuanlan.zhihu.com/p/31864048
- http://knowlife.mpi-inf.mpg.de/